IntechOpen

Satellite Communications

Edited by Nazzareno Diodato

Satellite Communications

edited by
Nazzareno Diodato

Satellite Communications
http://dx.doi.org/10.5772/253
Edited by Nazzareno Diodato

Notice

Statements and opinions expressed in the chapters are these of the individual contributors and not necessarily those of the editors or publisher. No responsibility is accepted for the accuracy of information contained in the published chapters. The publisher assumes no responsibility for any damage or injury to persons or property arising out of the use of any materials, instructions, methods or ideas contained in the book.

First published in Croatia, 2010 by INTECH d.o.o.
eBook (PDF) Published by IN TECH d.o.o.
Place and year of publication of eBook (PDF): Rijeka, 2019.
IntechOpen is the global imprint of IN TECH d.o.o.
Printed in Croatia

Legal deposit, Croatia: National and University Library in Zagreb

Additional hard and PDF copies can be obtained from orders@intechopen.com

Satellite Communications
Edited by Nazzareno Diodato

p. cm.

ISBN 978-953-307-135-0

eBook (PDF) ISBN 978-953-51-5932-2

Meet the editor

Nazzareno Diodato was born in Benevento, Italy in 1966 and was educated at the University of Naples \"Federico II\" where he received five years of education in Urban and Landscape Planning. Thereafter he conducted graduate studies at the several International Schools on GIS and Earth Sciences. Between 2008 and 2010 he was titled to the Fellow degree by Royal Meteorological Society. On 2009 Mr Diodato was regarded Member of the Scientific and Technical Committees-Editorial Review Boards at the World Academy of Science, Engineering and Technology. Currently Mr Diodato is a Geoscientist at the Met European Research Observatory (MetEROBS). Established in 1986, MetEROBS is an accredited observational supersite, awarded in 2011 as Laureate by the ComputerWorld Honors Program Computer-World Honors Program.

Contents

Preface

The title of this study incorporates two currently very popular concepts: *satellite* and *communication*. In an era where geoinformation demand are growing at worldwide scale, remote sensing and telecommunication around digital and electronic networks are seriously challenged by the diverse applications of novel digital and information communication technologies.
The omnipresence of personal computers and, especially, the introduction of the World Wide Web and the enormous growth of Internet use that has followed, has concretized the possibility to exchange information.
The massive amount of writing on the information society has been very much a result of sociologic, philosophic, and political discussion. It is generally accepted that the currently ongoing transition, resulting in a knowledge-based information society, can bring enormous benefits to societies.
Decisions made and actions taken today may affect the evolution considerably. Therefore individuals active in various fields need to have a basic understanding of the topics, their relation to each other and the factors influencing their combined evolution.

This study is motivated by the need to give the reader a broad view of the developments, key concepts, and technologies related to information society evolution, with a focus on the wireless communications and geoinformation technologies and their role in the environment. Giving perspective, it aims at assisting people active in the industry, the public sector, and Earth science fields as well, by providing a base for their continued work and thinking.
Questions addressed in the study are developed within 23 chapters which can be enclosed in four main thematic areas:

♣ System and Models for Satellite Access and Communications
♣ Emergency Communications in social areas
♣ Mapping Communication in Earth Sciences
♣ Optimization Problems and Develop of New Technologies

In the first issue System and models for satellite access and communications are presented five chapters: 1) **About-QoS in DVB-S2/RCS systems** that deals with design and evaluation of Quality of Service Architecture to be implemented in DVB-S2/RCS systems. In this way, geostationary satellite networks are expected to play a decisive role in bridging the existing digital divide through providing broadband access to multimedia services in low terrestrial infrastructure areas. 2) **Antenna system for land mobile satellite communications** describes an antenna system for land mobile satellite communications particularly in design, development and chamber measurement as well as realization in field experiment using the geostationary satellite. The developed antenna system is simple, compact and promising in low cost production. 3) **Cooperative strategies for satellite access** can be seen as a new form of spatial

diversity in which the diversity gain can be achieved through the cooperation of different users, opportunely grouped in clusters. In this way a new class of communications, called cooperative communications, has been proposed as a valuable alternative to spatial diversity techniques which require the deployment of additional antennas in order to mitigate fading effects. 4) In **MIMO channel modelling** is presented an overview and analytical analysis of standard MIMO channel models for satellite communication systems. In order to further test the channel models and the effect of the propagation environment, was investigated a novel application of using satellite diversity in conjunction with compact MIMO antenna array configurations (exploiting the polarization and pattern diversity features provided by these compact MIMO antennas) in order to enhance the capacity of satellite communication links.

In the second issue Emergency Communications in social areas are presented four chapters: 1) **Experiment in telecommunications by satellite and wireless network solution for rural areas**, where the TWISTER project (2005/2007) responses to the European Commission's call for tenders, dedicated to satellite solutions. This analysis has enabled to confirm the relevance such as the 'applications-territories' coupling, certain territorial stakeholders want to develop more particularly structuring applications, such as, for example, e-administration services (for citizens), telemedicine (for retirement homers, ski resorts, etc.). 2) **Design and implementation of satellite-based networks and services for ubiquitous access to healthcare** was explicited by different projects that provided satellite-based trans-European competence network for telemedicine, telemedical support in cases of disaster emergencies, support on-board of cruise ships, and the EMISPHER project that enabled an equal access for most of the Euro-Mediterranean countries to online services for healthcare in the required quality of service. 3) In **rate emergency communications satellite link in Ku/Ka/Q/V bands** is expounded the link budget dimensioning and a customized, enhanced DVB-S2-like air interface proposed to support minimum emergency communications in a severely impaired channel environment in high frequencies. The paper then outlines the combined Excel//Juzzle/Matlab high-level transmission link software simulation platform that was developed in order to assess the performance of theproposed transmission scheme.

In the third issue Mapping in Earth Sciences are presented six chapters: 1) **Combining satellite and geospatial technologies for exploring rainstorm hazard** where a set of sequential GIScience rules was utilized for converting coding data of a Rainstorm Hazard Index (*RHI*) from point record to spatial information using TRMM–NASA satellite rain data as covariate. Examples of probability estimation for different precipitation durations, ranging from 3 to 48 hours and the quantification of hydrological hazard fields were used with probability maps of damaging rainstorms prone-areas for the test-region of Southern Italy. 2) In **design and simulation of a DVB-S2-like adaptive air interface designed for low bit**, a DVB-S2-based adaptive air interface was proposed and simulated to meet the performance constraints of bidirectional satellite communication links to be established in a post-disaster emergency situation in Ku/Ka and Q/V bands where strong channel impairments occur. So that this chapter expounds the link budget dimensioning and a customized, enhanced DVB-S2-like air interface proposed to support minimum emergency communications in a severely impaired channel environment in high frequencies. 3) **Mapping and estimation of chemical concentrations in surface soils** has used LANDSAT TM satellite imagery as an alternative method for determining and mapping the physical and chemical characteristics of the soil. Consequently, the focus of this study was to determine the use of remote sensing

to map chemical variability in bare soils in elemental concentrations of soils amended with biosolids, and in to use LANDSAT TM data to map these elemental concentrations of the soils when they are not covered by vegetation. 4) In **paperless solution for the collection, management and dissemination of marine data** were addressed many of the problems related to the complexity of the logical linkages between the different types of data in fisher management, Olrac (www.olrac.com) a South African company, that has developed a data collection and management system it has named Olfish (www.olfish.com) for the specific use of operators and managers in the marine environment with a special focus on the commercial fishing industry. 5) In **Vegetation mapping and management** was studied the vegetation types in the semiarid to arid region of Mond Protected Area, south-west Iran, based on unsupervised classification of the Spot XS bands and then produced updated maps. Based on field observation and supported by satellite maps, three major habitat zones in the study area, namely, the coastal, riverine and inland zones, were recognized. Twelve vegetation types were recognized in the field that showed a good compatibility with the satellite image. 6) **Earth to space link** describes how to calculate link-power budget in order to relate to two quantities, the transmitted power and the received power, and show in detail how the difference between these two powers is accounted. To arrive at an accurate answer, factors such as the uplink power amplifier gain and noise factors, transmit antenna gain, slant angles and corresponding atmospheric loss over distance and climatic attenuation factors must be taken into account.

In the fourth issue Optimization problems and develop of new technologies are presented with ten chapters: 1) **Guidelines for satellite tracking** raise the need to know where to look for the artificial and natural satellites in the evening sky has been the obsession for many. So that mathematical models and positioning mechanisms were explained in this chaper to paint the picture of satellite tracking and give the brief insight in when to use what mathematical model to pinpoint the object in sky, even if the final goal is just a view through a telescope. 2) **Interference studies in cellular satellite systems** accounts to reduce the array antenna aperture on board the satellite for multiple-spot-beam cellular coverage. The reduction in the antenna aperture with this approach translates into significant reductions in number of array elements, RF components, and A/D and D/A converters. Analysis has shown that in spite of the smaller aperture and the broader beams of the sub-beams, the co-channel interference between sub-beams using the same frequency segment is not adversely affected. 3) In **Ultra-life cycle utilization of GEO** have proposed and verified practical development plan for satellites navigation based on GEO satellites. Particularly was introduced one kind of satellite communication techniques with micro-terminal low-information rate developed in this section. This technique can satisfy some communications and positioning requirements such as unattended measurement and short-message emergency communications. 4) **Planar antennas for satellite communications** was devoted to planar antennas not only for that already in use but proposing other kind that could be used for satellite applications. In this way was described actual planar antennas used in satellite communications systems and finished with a proposal of new developments of planar antennas that authors think could be used in the near future. 5) **Power and spectral efficient multiuser broadband wireless communication system** focuses on different aspects of the communication system, namely PAPR reduction and power control in transmitter, channel estimation for design of adaptive and optimized system, multiuser detection at the receiver for increase in user capacity. *The newly proposed CI/MC-CDMA system discussed in this chapter may become an efficient multiple access scheme for*

Satellite communication used for long distance broadband signal transmission in conjunction with cellular system. 6) In **Quantum based information transfer** was introduced the quantum communication to help to establish a secure communication link, and present solutions with zero redundancy error correction. 7) **Satellite coverage optimization problems with shaped reflector antennas** was devoted to optimization formulations of satellite coverage problems and their solution which are specially useful in satellite broadcasting applications, and where the information goes from one to many. 8) In **Satellite Laser Communication** is used a cloud attenuation model derived from key Italsat results to find laser cloud attenuation for satellite-ground links with gaseous attenuation included for 2-10 micron attenuation. Link availability is raised with suitable Northern Latitude satellites, such as Earth Observation Satellites. Soviet cloud correlation results indicate that link availability would be raised to acceptable levels with ground sites separated by 100-290 km. 9) In **Satellite motion** is described a little history on the satellite motion and its mathematical formulation. This includes the dimensions of orbit and the position of satellite on its orbits, and what velocity for a satellite is requested.
10) **System aspects of active phased arrays** concluded this book by reviewing early developments of phased arrays for multiple-beam satellite communications applications. A key component in these developments is the modular monolithic microwave integrated circuit (MMIC) beam-forming matrices that generate a number of simultaneous and independently digitally controlled beams.

Editor

Nazzareno Diodato
MetEROBS – Met European Research Observatory, GEWEX-CEOP Network, World Climate Research Programme, via Monte Pino snc, 82100 Benevento Italy

1

About QoS in DVB-S2/RCS Systems

Baptiste Jacquemin[1,2], Pascal Berthou[1,2],
Thierry Gayraud[1,2] and Lionel Bertaux[1,2]
[1]*CNRS; LAAS; 7 avenue du Colonel Roche, F-31077 Toulouse, France*
[2]*Université de Toulouse; UPS, INSA, INP, ISAE; LAAS; F-31077 Toulouse, France*

1. Introduction

The standardization of a Return Channel via Satellite and the satellite community efforts in term of interoperability over the last few years stands for major milestones in the development of reliable, efficient and low cost satellite equipments. It leads to quite a positive outcome: geostationary satellite networks are expected to play a decisive role in bridging the existing digital divide through providing broadband access to multimedia services in low terrestrial infrastructure areas.
However, unlike cable or 3GPP access networks, a lot of work on IP over satellite has been needed, especially about Quality of Service (QoS). The QoS architecture takes benefits from DVB-RCS dynamic allocation schemes and IP QoS architecture to cope with the satellite delay and the scarce uplink resources. This chapter deals with design and evaluation of Quality of Service Architecture to be implemented in DVB-S2/RCS systems.

The first section of this chapter aims at introducing DVB-S2/RCS Systems.
The long term efforts to optimize the DVB-S standard to lower the price of satellite access networks led to a new evolution of the standard: DVB-S2. A better encapsulation mechanism of IP packets and a new adaptative transmission scheme are the main concerns for the QoS architecture.
The encapsulation of IP packets in DVB-S has always been a complex problem. This section presents the evolution of the standard from the Multiprotocol Encapsulation (MPE) and the Ultra Lightweight Protocol (ULE) to the Generic Stream Encapsulation (GSE).
The Adaptive Coding and Modulation (ACM) technique that increase the network efficiency according to the weather conditions is a major evolution. The variable transmission rate impacts the QoS management and offers new perspectives for future system evolution.

DVB-S Satellite Terminals can only receive frames from the satellite. The need for a return link rapidly becomes essential so as to support emerging Internet services via satellite.
The return link access scheme in DVB-S/RCS systems is MF-TDMA. The return link is segmented into portions of time and frequency (“superframes. A Network Control Center (NCC) performs the entire satellite system control, especially Satellite Terminals synchronization and resource allocation. It periodically broadcasts a signaling frame, the TBTP (Terminal Burst Time Plan), which updates the timeslot allocation within a

superframe between every competing ST. This allocation can be dynamically modified on STs demand thanks to a bandwidth on demand protocol called Demand Assignment Multiple Access (DAMA). This system is presented here.

The next section rapidly overviews the concepts and mechanisms of Quality of Service management in basic architectures such as IETF Intserv and IETF / Diffserv. Others mechanisms such as Traffic shaping / conditioning, SLA, Scheduling and Admission control that have a main impact on the QoS are also described.

The next part aims at describing what means QoS in satellite networks thanks to the DVB-S2/RCS QoS Architecture example.
From the very first system only based on MPLS, a first architecture based on Diffserv was proposed. It was then enhanced to better fit to the DVB-RCS system in the IST project Satsix.

The next part will answer a main question related to the satellite networking systems that is: How to develop new services with Satellite Systems?
Based on our research work and results in the field, we'll explain how to use Simulation (using NS-2 or NS-3), Basic Emulation (using Linux TC/Simnet) and Advanced Emulation testbed like the one that was developed in various projects we were involved in. And we'll conclude that part with our skill on Real Deployed Systems.

The last part deals with Performance Evaluation of the described proposals. We first evaluated DVB-S/RCS NS-2 emulation model with QoS. The next way used to evaluate the proposed architecture was done through the PLATINE emulation testbed coming as the main result of the Satsix project. Our last experiment was done in the OURSES project, labeled in the Aerospace Valley research center with the following main devices from Astra (satellite), Thales Alenia Space (Gateway) and Advantech (Satellite Terminal).

To conclude this chapter, results summary and lessons learned will be given and future work will be described.

2. DVB-S2/RCS main features

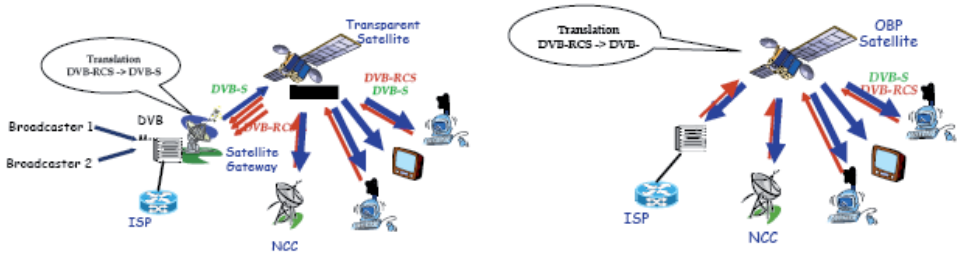
2.1 First overview of DVB-S/RCS systems

Started in 1993, the international European DVB Project published, in the end-nineties, a family of digital transmission specifications, based upon MPEG-2 (Motion Picture Expert Group) video compression and transmission techniques. Data are transported within MPEG-2 transport streams (MPEG2-TS) which are identified through DVB Service Information Tables. Adapted for satellite systems, DVB-S defines one of the most widespread formats used for Digital TV over the last years and still nowadays. However, DVB-S Satellite Terminals can only receive frames from the satellite. The need for a return link rapidly becomes essential so as to support emerging Internet services via satellite, leading to 3 solutions:

- UDLR (UniDirectional Link Routing) which emulates a cheap bidirectional solution through a terrestrial return link,
- DVB-S system with low speed terrestrial return link,

– DVB-RCS, which provides a full bidirectional satellite architecture [Fig. 1].

A good overview of DVB-S/RCS satellite networks architecture is given in Fig. 1, compliant with the architecture adopted within the ETSI BSM [3] group and the DVB-RCS standards. It consists in a geostationary satellite network with Ka MF-TDMA (Multiple Frequency Time Division Multiple Access) uplinks and Ku TDM (Time Division Multiplexed) downlinks.



a) transparent satellite b) regenerative satellite

Fig. 1. DVB-S/RCS architecture

Satellite Terminals (RCST) provide single PC or LANs with the access to the network, while Gateways (GWs) allow the connection with Internet core networks. The uplink access from each RCST is managed through DVB-RCS interfaces.

On the 2 topologies, the end-user side of the platform is on the right. On the left is shown the provider/enterprise/Internet side of the platform. It can be distinguished also between the satellite network side (in the middle) and the IP network sides (on left and right ends), interconnected by RCSTs. So, the 3 main components in the satellite network side (middle) are the Satellite, the Return Channel by Satellite Terminals (RCST) and the Network Control Center (NCC).

In Fig. 1–a, the architecture relies on a transparent satellite offering a star topology. All the forward links from GW to the RCSTs are DVB-S links and all the return links are DVB-RCS links. This allows the satellite payload to work in a simple transparent way without any computation to be made on the received frames before resending. Such a payload is easier to design and was the first implemented in such GEO satellites. But the main constraints of such architecture are due to the mandatory double hop to be done to go from one RCST to another one as it is needed to go through the GW to access to an RCST.
On the opposite, this problem is solved in the second kind of architecture shown in Fig. 1–b. In this topology, the uplinks (to the satellite) are DVB-RCS links only and the downlinks (to the RCST) are DVB-S. The complexity of this solution is located in the satellite where the payload has to be regenerative to translate incoming frames in DVB-RCS in outgoing frames in DVB-S. More complex to implement, the regenerative payload was designed later than the transparent one.

It has to be noticed that it is now time to implement hybrid payload including two parts one transparent and the other one regenerative inducing more complexity of the payload, but nothing new in the architecture components where the two kinds of network components coexist, but in separated configurations.

2.2 Specific DVB-RCS features

DVB-RCS systems involve lots of specific techniques, but only a few of them impact the QoS of such a satellite network. So this section is dedicated to the 2 main ones that are DAMA and Encapsulation.

2.2.1 DVB-RCS Demand Assignment Multiple Access (DAMA)

Furthermore, DVB-RCS requires a Medium Access Control (MAC) protocol because Satellite Terminals (ST) is able to simultaneously access the return channel capacity. The standard method relies on a Multi-Frequency Time Division Multiple Access (MF-TDMA). It basically relies on the availability of several TDMA channels (corresponding to different carrier frequencies), each subdivided into frames and further into timeslots of fixed length (bursts) during which the STs are able to transmit data through MPEG2-TS or ATM traffic bursts.

The entity responsible for this timeslot allocation within the Superframe shared by competing STs is the **NCC (Network Control Center)** that centralizes the satellite resources management. Thus it periodically broadcasts a signaling frame, the TBTP (Terminal Burst Time Plan) that contains the information on which STs relies to know when to transmit their bursts. This allocation can be dynamically modified by STs requests so as to prevent from wasting satellite resources that would be otherwise statically allocated. The implementation of such a mechanism is generally known as bandwidth on demand (BoD) algorithm.

In order to dynamically manage the bandwidth allocation, a bandwidth on demand protocol called Demand Assignment Multiple Access (DAMA) is defined. It relies on the STs ability to request frequently "capacities" to the NCC which enables a regular update of the TBTP to fit to the STs respective traffic load [Fig. 2]. The latter provides signaling schemes as well as MAC QoS Classes and their mapping on capacity types.

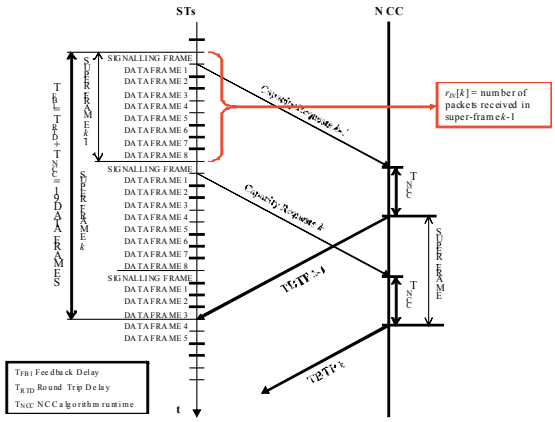


Fig. 2. DAMA algorithm: TBTP computed from RCST requests

Thus, the norm defines 5 Capacity Categories to fit the applications needs that will be detailed further in this paper. Capacity types are vital to return path QoS support at MAC layer; therefore they will be described in more details in the following. Any given ST can be

assigned one or a mix of the four capacity types. Generally, higher priority classes of service are associated with guaranteed capacity (CRA, RBDC), while lower priority classes are predominantly given best effort capacity (VBDC, FCA).
The DVB-RCS standard has left many issues open, e.g. how capacity requests are triggered, how and when certain parameters are negotiated (CRA), and if they can be re-negotiated, etc. It defines that when the NCC assigns timeslots to a certain RCST through the TBTP table, it can indicate a "channel" to which the timeslots are assigned.
It is obvious that this DAMA mechanism has great impact on what we will discuss later in this paper.

2.2.2 Encapsulation: from MPE to ULE

The multiprotocol encapsulation (MPE) provides a mechanism for transporting data networks on top of the MPEG2 TS in DVB networks. It has been adapted for carriage of IP packets, both IPv4 and IPv6. The encapsulation shall be done in accordance with the "Multiprotocol Encapsulation" technique described in the ETSI/DVB standard EN 302 192 and TR 101 202]. MPE includes methods for addressing the receivers of the data in the MPE header, which is necessary when many users have access to the same data channel. This feature allows several logical networks to be established without assigning PID values to each service.
IP datagrams are encapsulated in "datagram_sections" as defined in ISO 13818-6. The section_number and last_section _number must be "0" when carrying the IP protocol. The section format provides a format for mapping the datagram to the MPEG2 TS, and support filtering of datagram based on the MAC address using hardware or software demultiplexers.
The mapping of the datagram_section into MPEG2-TS is defined in ISO 13818-1. The sections are inserted into the payload of the MPEG2 packets (only first packet). The MPEG2 header is added to each packet. The resulting stream is the output of the data encapsulator/multiplexer, which is fed to the DVB modulator and satellite uplink equipment.
Many network operators and manufacturers of electronic equipment have adapted the MPE standard. That means that the standard is already in use and working well. Even though it is not the most efficient scheme for IP data transport.

An alternative encapsulation method has been defined by the IETF, RFC 4259. This directly places packet data into a Stream. This is called the Uni-Directional Link Encapsulation (ULE) defined in RFC 4326. The design of ULE simplifies processing, by reducing the number of header bytes and by significantly reducing the number of protocol fields that a receiver needs to parse. ULE also uses a Type field that resembles that adopted by the IEEE Ethernet standard, permitting easy interfacing to a wide range of network service (including IEEE 802.1pQ; MPLS; IPv4; IPv6).

ULE allows transmission of SNDUs up to 32 KB (compared to a maximum of 4KB in MPE). ULE also provides an extension header format (with an associated IANA registry), which allows future addition of new protocol fields to an encapsulated PDU, while providing backwards compatibility with existing implementations. This method is used to provide an efficient bridging method, but in future could also be used for encryption, compression, etc.

ULE is still old fashioned and solutions better fitted to Internet communications for instance have led to other proposal. The most promising called GSE will be presented later in this paper.

2.3 DVB-S enhancement: DVB-S2 standard and its new mechanisms

This section deals with the presentation of the new standard DVB-S2 and will be dedicated to the presentation of the main new features of such a satelite network, that are the DRA/ACM and GSE encapsulation scheme.

2.3.1 DRA/ACM

2.3.1.1 Return link

For the return link, different physical layers for individual terminals and for collective terminals have to be considered. The combination of adaptive coding, adaptive modulation and variable symbol rate can lead to different trade-offs depending on the type of scenario and of terminal.

The return link physical layer is based on the DVB-RCS standard with an adaptive waveform. DRA is considered, this means it is possible to change the coding rate, the modulation scheme and/or the transmission symbol rate. DRA is not included in the current DVB-RCS standard, but can be implemented using the standard capabilities without any additional information or changes (except for adaptive modulation using 8PSK and BPSK, which are not currently available in the DVB-RCS standard).

A DRA scheme is defined as an association of coding rate, modulation scheme and carrier symbol rate. The coding rate may be chosen among the set of DVB-RCS coding rates. Modulation can be either QPSK or 8PSK (with BPSK if required by link budgets). There is no specific constraint on the transmission symbol rate. This must be adapted to the terminal requirements in terms of the peak data rate range. However, due to equipment constraints (demodulation in particular), frequency plan constraints (in particular if the frequency plan needs to be reconfigured dynamically), the only transmitted symbol rates that are considered are multiples of each other.
A total of 70 combinations are thus possible.

The choice of DRA schemes to be retained is a trade-off between:

- the spectral efficiency (and thus the system capacity, the total number of users in the system and the average bit rate per user),
- the useful peak data rate (maximum data rate seen by a user)
- and the required SNR for each DRA scheme (the increase in required SNR should be reasonable with respect to the increase in peak rate or spectral efficiency).

When going from lowest DRA schemes to higher ones (i.e. from low SNR to a better SNR), we should:

- increase the required C/N0 with at least 2 or 3 dB steps between schemes to ensure the stability of the control loop,

- increase the spectral efficiency or the symbol peak data rate or both (that is increase the useful data rate).

When considering all possible combinations, we have 14 possible combinations of coding rate and modulation scheme, and 70 possible DRA schemes (combination of coding rate, modulation scheme and symbol rate). To select the set of DRA schemes required for the system, we consider some additional constraints resulting from implementation issues, as well as some specificity from the scenarios. We target a residential scenario with individual terminals, where we will try to optimise the spectral efficiency (to get a higher system capacity and a large number of users) rather than the peak data rate (that will be more a priority for collective terminals). However a reasonable peak data rate should still be offered to remain competitive with terrestrial solutions.
At the end, the selection process leads operational real systems to focus on around ten schemes only.

2.3.1.2 Forward link

The forward link physical layer is based on the DVB-S2 standard that supports an adaptive physical layer thanks to ACM (Adaptive Coding and Modulation). It addresses different kinds of terminals and different link conditions by allowing a large set of possible MODCOD schemes. With ACM, the coding rate and Modulation scheme can be chosen depending on the link quality within the set of available MODCOD (MODulation and CODing) schemes. The link can also be adapted dynamically by the implementation of a control loop for each station and dynamic measurement of the channel quality based on the SNIR estimation (with pilot symbols transmitted within the DVB-S2 frame).

For the forward link, there is no specific trade-off in the physical layer definition for the individual versus the collective terminals. They will all use a subset of the overall DVB-S2 possible MODCOD schemes. The MODCOD schemes retained for such systems only depend on the performances of the MODCODs proposed by the DVB-S2 standard. We select the best schemes in terms of spectral efficiency versus required Es/N0. Then the MODCODs that are actually used by individual or collective terminals will only depend on the Es/N0 range that they can reach.

This means the same DVB-S2 downlink carrier could be used to address both individual and collective terminals, the only constraint being that all the MODCOD schemes shall be supported at the same time. However, for the computations hereafter we will separate the MODCOD schemes distribution computation for individual and collective terminals.
And, as before on the return link, the selection process leads operational real systems to focus on around ten schemes only.

2.3.2 Encapsulation in DVB-S2: GSE

S2 introduces a new physical layer that supports a set of transmission waveforms that use a combination of higher-order modulation and powerful FEC coding. Although backwards-compatibility with existing DVB-S is supported, the main advantage arises when S2 is used with a range of terminal capabilities, particularly when the waveform is dynamically chosen

to match the prevailing conditions of the receivers. A transmission frame consists of a 90-bit physical layer header providing a preamble and identifying the ModCod used. The payload of a physical layer frame is known as a baseband frame (BBframe) and includes a 10 byte signalling header, which is followed by the BBframe payload. The size of this payload depends on the ModCod that was selected and can be up to 8 KB, significantly larger than an MPEG-2 TS Packet. The BBframe payload may carry a sequence of TS Packets, as in other MPEG-2 networks. Therefore, PDUs can be sent by encapsulating them using MPE or ULE. In S2, the transmission is optimised by omitting any Null TS Packets from the BBframe, and re-inserting these at the Receiver (preserving end-to-end timing).

The requirements for GSE differ to those of MPEG-2 TS, in that the GS uses a strong FEC code, and a much larger frame compared to that of a TS Packet. The requirement to check the integrity of a received SNDU therefore differs from that for the MPEG-2 TS. The current proposal is not to check every SNDU, but instead to place a 32-bit CRC at the end of each BBframe. Fig. 3 shows a general picture of the different encapsulation mechanisms involved in IP communication over DVB-S2 through GSE.

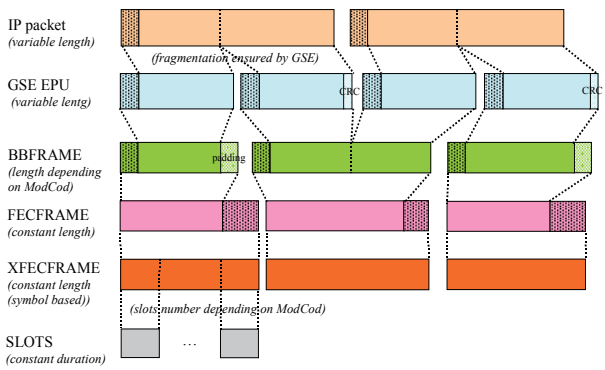


Fig. 3. GSE encapsulation in DVB-S2

The BBFRAME is the container that adapts GSE to fit the DVB-S2 properties to a data stream. DVB-S2 allows the use of a different ModCod (Modulation and Coding scheme) for each BBFRAME. This permits to take into account (while sending DVB-S2 data) the reception conditions, if available (e.g. through DVB-RCS).
A scheduling algorithm needs to take such variations into account.

In the DVB-S2 architecture, both modulation and coding can change from one BBFRAME to another, leading to variable length and variable duration BBFRAMEs. These two parameters are, however, tackled independently with the help of FECFRAMEs and XFECFRAMEs.
However, once the ModCod (more precisely the coding scheme) has been chosen, the BBFRAME length is implicitly fixed so that a padding mechanism is needed.

The FECFRAME is a constant size container that is used to encapsulate the BBFRAME with its associated coding. There are two possible values for a FECFRAME length (16200 bits or 64 800 bits) only one of which is implemented in a given system.

We will not describe more detail about FECFRAMEs, as the only pertinent information regarding a scheduler implementation is the constant length of the FECFRAME and the "variable" length of its payload.

The GSE protocol uses the BBframe header to determine the frame size (and presence of padding). The encapsulated PDUs also known as EPUs (Encapsulated Payload Units) are carried in DVB-S2 BBFRAMEs.
Not all receivers necessarily support the same set of ModCods (e.g. implementation of different sets of ModCods, different locations within the satellite footprint, local propagation conditions, or other reasons). Care therefore needs to be taken to ensure that all SNDUs or SNDU fragments are sent using GSUs with a ModCod that can be received by the intended recipient(s). Sending a SNDU or a SNDU fragment with a higher ModCod than required consumes unnecessary satellite capacity, while sending with a lower ModCod could result in loss.
At the time of writing, issues remain open such as flow security, support for signalling/control, the effect of small packets (concatenation) and these are the overall performance of the different features when used in an environment where traffic and terminal reception conditions are both varying.

3. QoS in terrestrial networks

The concept of quality of service (QoS) has gradually emerged to meet the various requirements demanded by certain types of applications, primarily in the world of interactive. Indeed, initially, the majority of Internet traffic consisted of textual data having no specific requirements but progressively, tools involving simultaneously files transfer, instant messaging, audio and video appeared. Consequently, guarantees on bandwidth, delay or jitter should be provided to users to ensure proper functioning.
However, the architecture of the Internet, based on TCP/IP, is not designed to differentiate the types of traffic and is currently dominated by a single model of service: best effort. This architecture can only guarantee correct functioning of all types of applications by offering the oversizing of the network that consists in configuring a network with a capacity that far exceeds the requirements. But this approach only postpones the problem and can hardly be applied to wireless access technologies with limited bandwidth.
Thus, effective management of resources is necessary to provide to Internet users QoS guarantees adapted to their needs. In this part, we will therefore firstly introduce the fundamentals of QoS. Then, the precursor models of QoS, IntServ and DiffServ will be detailed and finally we will show how the two signaling protocols, COPS and SIP, can participate in the implementation of QoS.

3.1 The basis of QoS

According to the standard ISO 8402 (ISO, 2000), quality of service is defined as "the totality of characteristics of an entity that bear on its ability to satisfy stated and implied needs". The E800 recommendation ITU-T (ITU-T, 1993) defines it as "the collective effect of service performance which determines the degree of satisfaction of a user of the service".
Thus, different types of QoS can be defined from these definitions and can be separated into three categories as defined in (Hardy, 2001):

- The intrinsic quality of service that is provided directly by the network itself and described by objective parameters such as, for example, the delay or the losses. It is the primary IETF focus.
- The perceived quality of service that matches the quality perceived by the user (also called QoE, Quality of Exeprience). It depends heavily on network performance but is measured by an average of users' opinions. The most common method is the MOS (Mean Opinion Score) in which a group of users separately evaluate perceived quality of an application between 1 and 5, an average score being then performed. The MOS is typically used for audio or video quality of an application, but may also involve perceived QoS connection time, the perceived security of user, service availability, etc… In addition, there is not necessarily a correspondence between intrinsic QoS and perceived QoS, the latter being very subjective. ETSI and ITU primarily use the term QoS as perceived QoS and prefer the term network performance to describe the technical part of QoS.
- The assessed quality of service that refers to the willingness of a user to continue using a particular service. It depends on the QoS perceived but also the price, service support offered by the supplier and other commercial aspects.

The main parameters that describe the intrinsic QoS in IP networks are:

- The transfer delay of packets, in milliseconds. It is usually measured from end to end but may be calculated on a portion of the network.
- The jitter or delay variation of packet forwarding, expressed in milliseconds.
- The bit rate, expressed in bits per second (bit/s or bps) or bytes per second.
- The rate of packet loss, defined as the percentage of packets lost in relation to the total number of packets transmitted.

3.2 The main existing models to ensure QoS

Two major proposals were made by the IETF to ensure QoS and proper functioning of IP services, real time or not: IntServ (Integrated Services) and DiffServ (Differenciated Services) each associated with a working group of the same name.

3.2.1 The IntServ model

The work of the IntServ working group led in 1994 to the definition of an integrated services architecture (Braden et al., 1994) composed essentially of two parts: an extended service model, called the IS model and a reference implementation framework for the development of this model.

This architecture, allowing to support QoS without changing the IP, is based on a per flow resource reservation. Each router must maintain the state of the flows going through it, which fundamentally modifies the principle of the Internet, which, on the contrary, was till now based on a conservation of the state of flows at the user terminals. That's why routers are then equipped with four additional features:

- The packet scheduler, responsible for the delivery of packets streams, using a set of queues as well as other mechanisms such as timers.
- The classifier that realizes the correspondence between an incoming packet and the class of service to which it is associated. The level of QoS provided by each service class is programmable for each stream.

- The admission control that implements the decision algorithm used by the router to determine whether a new flow may or may not obtain the requested QoS without degrading the guarantees offered previously.
- The reservation establishment protocol, needed to create and maintain the state of flow on routers. The protocol chosen for this function is RSVP (ReSerVation Protocol), defined later as Resource reSerVation Protocol in (Braden et al., 1997).

Two new classes of service are then defined in addition to best-effort that receives no special treatment at routers:

- The Guaranteed Service (GS) (Shenker et al., 1997) provides guarantees in terms of bandwidth and maximum transfer delay of packets, expressible quantitatively. If the stream respects the reserved parameters, this service ensures that all packets will arrive with a maximum delay and that they will not be lost in the queues in case of congestion. This service is suitable for real-time applications with strong time constraints such as videoconferencing or VoIP. However, no average delay is guaranteed, so it's the application itself which has to manage the delay variations at the receiver side by using buffering mechanisms.
- The Controlled-Load service (LC) (Wroclawski, 1997) is a service expressible qualitatively in terms of bandwidth, which ensures the user that its data stream will be transmitted with a QoS close to the one obtained in a network not overloaded (not congested).

The guarantees are obtained from end-to-end by the concatenation of such assurances, offered separately by each router crossed along the path. Furthermore, as stated previously, the protocol used to configure these routers is RSVP and we will now explain its functioning.

Initially, a PATH message is propagated from the transmitter (sending application) to the receiver. This message contains the traffic specification (TSPEC) that will be generated by the application. This specification can not be modified along the path but other information can be added through additional specification (AdSpecs) to precise specific resources constraints. Once the message arrived, the receiver responds with a RESV message that contains the description of traffic flow to which the resource reservation should apply (TSPEC Receiver) and the parameters demanded to implement the required service (RSpec). These descriptions may also change along the path.

RESV messages must follow the reverse path of PATH messages and trigger an effective reservation of resources (state have to be stored at each router) if the admission control mechanisms of each router validate the request. If a router validates the request, it creates and maintains a state corresponding to this flow. However, the reservations lifespan is limited and PATH/RESV messages must be exchanged periodically so that the reservation remains valid. It allows the system to be robust to changes in routing for exemple.

Once QoS is configured, when a reserved flow is going through a router, the classifier identifies it by its 5-tuple (Source IP address, destination IP address, Protocol, TCP/UDP source port, TCP/UDP destination port). The scheduler then handles the queue management.

RSVP is so a signaling protocol that allows to reserve dynamically bandwidth and ensure a maximum delay from end-to-end. This reservation, initiated by the receiver, can prevent that some applications monopolize resources unnecessarily and allows in the case of multicast communication to differentiate the reservation (and billing) for each receiver.

Moreover, its dynamic functioning can adapt to changing communications (changes in the number of participants, route changes, etc…). Finally, this protocol also has the advantages of being adaptable to both IPv4 and IPv6 and seamlessly passing non-RSVP routers.
However, RSVP requires maintaining state information for each flow at each node or router along a path connecting a transmitter to its receiver. So, when the number of users and flows increases, the number of states becomes significant and the traffic is all the more saturated as refreshments between RSVP routers become important and create overhead. The main shortcoming of the IntServ model and the associated RSVP protocol remains their lack of adaptation to the scale factor, especially as the reservation in RSVP is unidirectional. So, for a bidirectional application requiring QoS in both directions, the amount of messages is twice as high. The IntServ/RSVP model is therefore more adapted to small networks such as LAN.
It is one of the reasons that explain that another model of architecture has been proposed: the DiffServ architecture.

3.2.2 The DiffServ model

To solve the problem of scalability posed by the per flow management of QoS in the core network routers, the DiffServ working group has therefore proposed ((Blake et al., 1998) and (Grossman, 2002)) to separate traffic by classes of service. Thus, per flow QoS treatment is realized only at edge routers that aggregate flows by traffic class. The number of state maintained by core routers is so reduced to the number of classes and not anymore to the number of flows, which greatly reduces the complexity.
Each service class is identified by a value encoded in an existing field of IP header, redefined by the DiffServ group and named DSCP (DiffServ Code Point), which presents the advantage of not requiring the use of an additional signaling protocol. This is the TOS (Type Of Service) for IPv4 and TC (Traffic Class) for IPv6.

The DiffServ architecture is primarily based on the concept of DiffServ domain that consists in the grouping of one or more networks under a single administrative authority. This domain is composed of nodes or core routers that are only connected to nodes within the same DiffServ domain and edge routers that interconnect the DiffServ domain to other domains, DiffServ or not. The edge routers can also acts as well an incoming router when the traffic enters into a DiffServ domain as an outgoing router when the traffic leaves the domain.
A client of a DiffServ domain (which can be either a user or another domain DiffServ) must negotiate a contract with the service provider responsible for this area that specifies the terms and conditions of use of the concerned services: this contract is called an SLA (Service Level Agreement) and the technical part is specified by different SLS (Service Level Specification). An SLA contains the following information:

- The traffic that the customer is likely to generate in terms of data volume, rate, number of users, etc...
- The QoS that the service provider has to provide the customer in terms of availability, security, reliability or performance (delay, bandwidth, etc…).
- The policy followed by the service provider in case of overflow traffic (rejected, accepted but surcharge, etc...).

Finally, (Nichols, 1999) defines an entity called bandwidth broker with knowledge of the resources availability and policies of the associated domain. One of its main tasks is the admission control. In addition, to allow an end-to-end allocation of resources across all the domains taking into account the different SLAs negotiated between them, this entity must communicate with the bandwidth broker of the neighbouring domains.
To manage a DiffServ domain, the service provider in charge of it first starts by sizing its network according to the contracted SLAs with all of its customers.
The processing of packets entering the DiffServ domain is then realized at the edge routers in charge of flows classification by class of service, and traffic conditioning. To do this, they are composed of a classifier, a meter, a marker, a regulator and finally a dropper of non-conforming traffic. The classifier identifies a flow from either the DSCP field only, or a combination of one or more fields such as source IP address, destination IP address, DSCP, protocol ID, source and destination ports or other information as the input interface. Then, mechanisms for profiling and traffic measurement allow on the one hand marking or re-marking the packets and, on the other hand, to shape flows or drop them totally or partially to meet the negotiated traffic profiles.
During the marking, the DSCP field is updated using one of the different classes of service or PHB (Per Hop Behaviour). In addition to the Best Effort class which is not subject to any special treatment, two PHBs have been defined by IETF:

- Expedited Forwarding (EF) (Jacobson et al., 1999) which corresponds to the highest priority and ensures the transfer of flows with high temporal constraints (eg VoIP, videoconferencing) guaranteeing a certain bandwidth and low delays, jitter and loss rate.
- Assured Forwarding (AF) (Heinanen et al., 1999) which defines four levels of priority (AF1, AF2, AF3, AF4) on the delivery of certain packets in case of congestion.

Once packets have been marked by the edge routers, they are treated within the DiffServ domain by the core routers depending on the PHB deducted from the DSCP field which describes the forwarding behavior of these routers. The priorities are then performed by scheduling algorithms such as PQ (Priority Queuing), WFQ (Weighted Fair Queuing) or CBQ (Class Based Queuing).
The DiffServ model therefore provides a QoS management more adapted and more realistic than the IntServ one. The per class management (or per aggregate) can indeed be much more resistant to the scale factor. Moreover, Diffserv does not require signaling protocol like RSVP by adapting the header of IP packets, which saves bandwidth.
However, the sizing of a DiffServ domain taking into account the SLAs contracted with neighboring domains and users is a heavy and complex task to implement that does not dynamically adapt to rapid changes in traffic. Similarly, not using a signaling protocol at the application level implies that the user is unable to dynamically change the resources according to its needs.

3.3 QoS Signaling protocol

The previous paragraphs show that, although IntServ is unsuitable for scaling up, the use of a resource reservation protocol (RSVP) allows a more precise (adapted to each stream) and dynamic QoS management. The idea of using one or more signaling protocols to negotiate and establish end-to-end QoS has consequently been the subject of a number of research among the scientific community.

The objective of this part is then to present some of these protocols that enable the implementation of QoS, the other concepts (security, supervision, etc.) being not addressed.

3.3.1 COPS and the notion of QoS management by policy

The RAP (Resource Allocation Protocol) working group of the IETF defined in 2000 an architecture based on the notion of policy to improve the admission control mechanisms of a network. A policy is defined as a set of rules for the administration, management and access control to network resources. Each rule is then associated with a set of conditions which corresponds to a series of actions to do in case of these conditions are met. To enable the exchange of such policies on a client/server model, a protocol is also defined by that working group, the protocol COPS (Common Open Policy Service) (Durham et al., 2000).
This model of by policy management is composed of two core elements: PEP (Policy Enforcement Point), responsible for implementing policy decisions and the PDP (Policy Decision Point), in charge of making decisions based on defined policy. These two components communicate via the COPS protocol. To take its decisions, the PDP communicates with a database of policies through the LDAP protocol (Leightweight Directory Access Protocol) and can query other entities such as an authentication server or a bandwidth broker using SNMP (Simple Network Management Protocol) for example.
This generic model of network control by policy allows for two distinct modes of control: outsourcing and configuration (also known as provisioning).
In the outsourcing model, a router (including PEP) which has to make a decision on acceptance of a reservation sends a COPS request to the PDP and the latter responds with the decision taken in accordance with the policy rules. It can especially be used in relation with the IntServ model when a router receives a RESV message and must decide to accept or decline the resource reservation. Indeed, that decision may require other information than the resources locally available at the router and the use of a PDP can thus be judicious.
In the configuration model, when external events require a change in the configuration of routers (and thus PEPs), the PDP may communicate to them new rules to apply through the COPS protocol. They no longer have to seek the PDP before making a decision. This model can overcome the main weakness of DiffServ model in which classes of service configuration is static. Indeed, a network administrator can define, for example, two types of policies, one appropriate to the day where many VoIP calls are held and one for the night which suits rather backup's servers or data downloads. In this case, edge routers act as PEPs, while the bandwidth broker plays the role of PDP.

3.3.2 SIP: the session control and the QoS

SIP (Session Initiation Protocol) is a signaling protocol standardized by the IETF (Rosenberg et al., 2002) designed to establish, modify and terminate sessions with one or more participants. Its main use is now the session management of voice over IP (VoIP), for which it is currently the most common open standard, but the fact that it is independent of the type of data transmitted and the type of protocol used also allows him to develop many other applications such as instant messaging, video conferencing, distance learning, video games or virtual reality. The parameters of these sessions are described by the Session Description Protocol or SDP (Handley et al., 2006) and, in this part, we will see how these parameters can be helpful for the dynamic reservation and release of resources.

These parameters are negotiated during the establishment of the session but also during the session modifications. This negotiation can be conducted directly by the SIP clients located on users' terminals but, although they are aware of some important parameters, such as codecs they can use, the latter have no real means to know the status of available resources along the path their communications will cross. Moreover, it would imply the integration of QoS mechanisms such as RSVP or COPS on SIP clients, which has the double disadvantage of making the SIP clients more complex and not allowing SIP clients which are unaware of QoS to use this option.
That's why these new features must be rather implemented at the SIP proxy: the intermediate entities where SIP clients are registered and that intercept SIP messages. Indeed, they may allow operators to know the duration of the session, the number of involved media, the codecs used and their associated characteristics (bandwidth, etc..). From this information, automatic management of the resources reservation/release and of admission control according to the policy adopted by the operator becomes feasible and no changes are required at the SIP clients.
Two modes of session establishment with QoS reservation can then be distinguished:

- The "enabled" mode in which the session establishment and resource reservation are performed in parallel but do not depend on each other. The session will start regardless of the outcome of the reservation. For example in the case of DiffServ networks, QoS will be BE if the reservation has failed and EF if it worked.
- The "assured" mode in which the establishment of the session is done only if the QoS reservation has been realized.

A standard (Camarillo et al., 2002) detailing the operation of these two modes has been proposed. It is based on new messages such as Session Progress, UPDATE or PRACK and a set of preconditions added to the session descriptors. The SIP clients can then specify whether the implementation of QoS is "mandatory" ("assured" mode) or "optional" ("enabled" mode) for each direction of communication (reception, transmission) and if the QoS must be e2e (end-to-end) or local (at the access network). However, RFC 3312 (Camarillo et al., 2002) considers that QoS is implemented by the participants of the SIP session while we will consider that it is implemented by the SIP proxy. The Fig. 4 illustrates an example of a SIP session integrating resource reservation/release in "assured" mode.

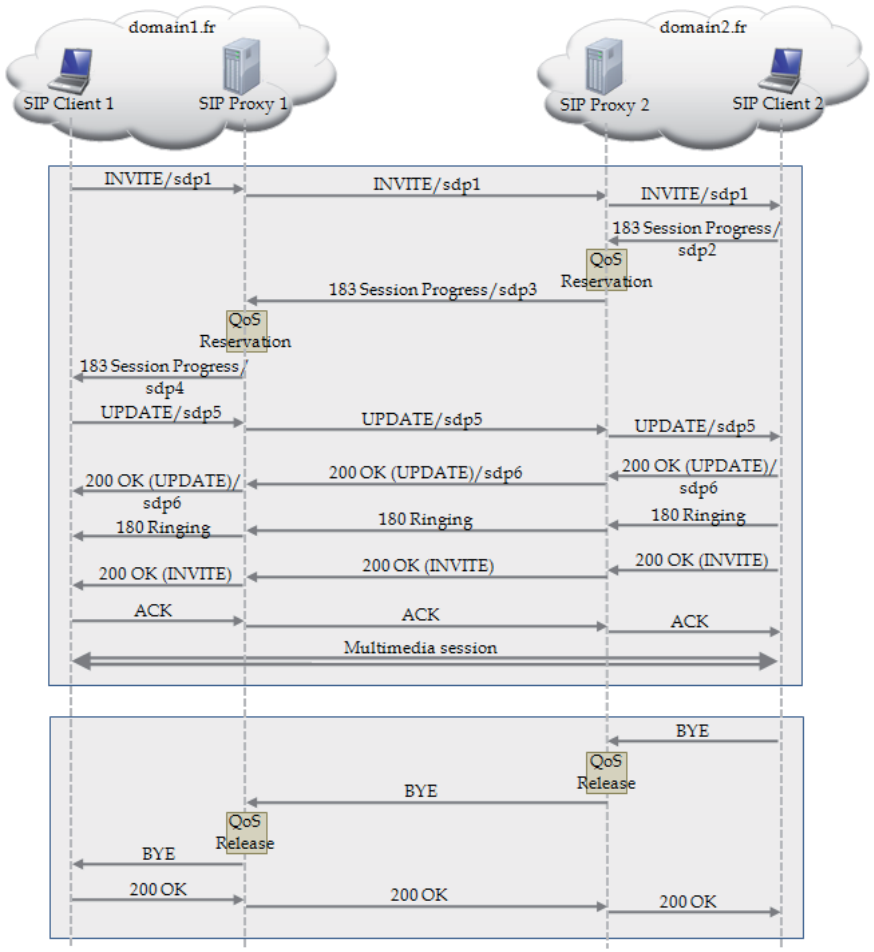


Fig. 4. SIP session with resources reservation/release

PRACK and 200 OK (PRACK) messages are not presented to make the figure more comprehensible but they are normally exchanged at the reception of a 1xx message to make it reliable. It can also be noticed that, in this example, all messages are exchanged via all the SIP proxies to manage the QoS along the whole path. Finally, QoS reservation/release in Fig. 4 may for example represent the exchange of messages (COPS, RSVP, etc..) allowing SIP proxies to communicate with the entities responsible for resources management within the concerned domains. For example, if we consider a DiffServ domain between domain1.fr and domaine.fr, the SIP proxies will be able to exchange COPS messages with the Bandwidth Broker in charge of the DiffServ domain. The latter can then configure the edge routers of this domain to prioritize the flows of the future SIP session.
Moreover, whatever the session establishment mode is, if a change of session is triggered by a re-INVITE message, SIP proxies can analyze the new parameters and automatically warn the entities responsible for resource management. The SIP signaling thus allows much more dynamic management of QoS for the applications it controls.

SIP, although it was designed originally to allow the session control, can be used very efficiently in the dynamic management of QoS.

4. QoS in satellite networks: DVB-S2/RCS QoS Architecture

The quality of service in DVB-S2/RCS networks is basically managed at layer 2. The first section gives an overview of classical layer 2 QoS architecture. The QoS offer should satisfy the application QoS requirements then this relationship is covered by the upper layers. Three QoS management strategies that can be considered as three different architecture maturity degrees are presented in this section. The first one, corresponding to a short term technique to introduce QoS in satellite system is based on MPLS over satellite systems. Then a fully IP based solution is presented as a medium term solution and the complete integration of satellite networks in Next Generation Networks (NGN) consists in the long term evolution of theses architectures.

4.1 Layer 2 QoS management

- In star architectures, the gateway (GW) centralizes the traffic and the signaling path and uses all of the offered bandwidth. Then, the quality of service management could be simply ensured by a correct scheduling algorithm, according to the packets destination and their service classes.
  But, when considering meshed topologies, the return link is shared among multiples satellite terminals. As one of the first things that impact the quality of service in the satellite network is the network access, packets should be sent on the air interface as soon as possible or even compliantly with the delay or the bandwidth required by the application. Then, in theses architectures, the quality of service management is essentially assured for the return link, i.e the DVB-RCS part.
  As said previously, the DAMA request/assignment cycle exhibits a non negligible latency and additional delays that cannot always match interactivity requirements of multimedia services. In order to both maximize satellite resource use and meet multimedia requirements, the DVB-RCS norm discriminates RCST capacity requests into 5 categories:
- Continuous Rate Assignment (CRA): Fixed slots are assigned in each MF-TDMA frame for the whole duration of a RCST connection
- Rate-Based Dynamic Capacity (RBDC): a dynamic rate capacity (in slots/frame) granted in response to explicit RCST requests
- Volume-Based Dynamic Capacity (VBDC & AVBDC): a dynamic cumulative volume capacity (in slots), granted in response to explicit RCST requests
- Free Capacity Allocation (FCA), which is assigned to STs on an “as available” basis from unused capacity

The standard defines separate MAC traffic priority queues and suggests a requesting strategy for each of them, that is to say a relevant mapping between traffic and request categories. Any given RCST can be assigned one or a mix of the four capacity types. In general, higher priority classes of service are associated with guaranteed capacity (CRA, RBDC), while lower priority classes are predominantly given best effort capacity (VBDC, FCA).
The generally recommended MAC queues are the following:

- Real Time (RT) queue used by temporally constrained applications as VoIP or visioconference. A CRA allocation mechanism is often used to feed it.
- Critical Data (CD) queue used by critical applications. It could be file transfer applications as well as low importance visioconference. RBDC requests are generally used.
- Best Effort (BE) queue used by non critical applications, as emailing or web browsing. RBDC, VBDC, or AVBDC are recommended.
- Network Management (NM) queue is recommended to offer a guaranteed path to signaling protocols as those presented in the following sections.

The CRA capacity allocated to the RCST should be consumed by the Real Time queue, but could be shared with other classes if unused.
However those suggestions are not yet sufficient to seamlessly integrate satellite networks into an end-to-end NGN QoS Architecture. The next sections propose three solutions, from the simple one to the NGN compatible one.

4.2 QoS with MPLS based architectures

One of the challenges the Internet is facing since the middle of 1990's is the exponential increased level of the IP traffic, sometime beyond what routers are able to handle. One solution to overcome the router bottlenecks it to delegate more of the IP forwarding to the layer 2. This allows the router to shortcut the heavy layer 3 processing based on routing table and implemented in software, while using instead a lightweight layer 2 switching table implemented in dedicated ASICs in the forwarding process.
MPLS is being defined by the Internet Engineering Task Force (IETF) as a standards-based approach to apply label-switching technology to large-scale networks. Although MPLS, as its name suggests, can conceptually support multiple protocols, the current work is focused on the integration of IPv4 and IPv6 with ATM, Ethernet, Wavelength Division Multiplexing (WDM) and Frame Relay. Thus, MPLS could be seen as the simplest solution for the seamless IPv4 to IPv6 migration of a heterogeneous network.
MPLS forwards data using labels that are attached to each data packet. Intermediate MPLS nodes do not need to look at the content of the data in each packet. In particular the destination IP addresses in the packets are not examined, which enables MPLS to offer an efficient encapsulation mechanism for private data traffic traversing the SP backbone. Service providers are therefore offering MPLS as a VPN technology.

4.2.1 Intserv / Diffserv QoS support in MPLS

MPLS acts as a level 2/3 protocol, its role is to transport a native network level traffic. IP includes in its header ToS field in IPv4 or Traffic Class field in IPv6, which provides information on the kind of traffic to transport. The traffic level quality information that should be borne by MPLS is fully dependent on its upper layer protocol or the resources reservation provided by RSVP. Both IntServ and DiffServ classes are supported over the MPLS architecture. IntServ and DiffServ classes are mapped onto DVB-S/RCS classes. Bandwidth allocations and flow isolation is efficiently supported by means of the MPLS VPN mechanisms.
The **IntServ** service architecture incorporates one main component to support the IntServ QoS model: RSVP-TE (Awduche, 2001). This IntServ Signalling protocol for MPLS is an

extension of RSVP for MPLS. It allows traffic engineering via label distribution. RSVP-TE works on the Downstream-on-Demand principle, i.e. the upstream router requests a label from the downstream router. This allows an ingress router to specify a route for a given flow. Resources reservation is optional with RSVP-TE; this allows the label distribution even if the paths do not need reserved network resources. Unlike RSVP, RSVP-TE works only between ingress and egress routers, and not between the sources and the destinations of the flows. RSVP-TE also allows the traffic policing working with th‹e LSP instead of the destination IP address.

Unlike IntServ, **DiffServ** does not require the reservation of resources end-to-end. It actually runs on a Per-Hop-Behaviour (PHB) principle. This means that each router filters the incoming packets. In MPLS, LSRs do not read the IP headers, thus they do not have any direct information regarding the PHB for arriving packets. MPLS encapsulates IP packets with a shim header that includes the label and the Exp field. This last one is composed of 3 bits intended to support marking of packets for DiffServ, which on its side defines its classes of service with the 6-bit DSCP field. The problem is that DiffServ can have 64 possible DSCPs whereas MPLS/Exp bits can only address up to 8 possible types of Per-Hob Behaviors (PHBs). In order to support for instance 1 EF class, 4 AF classes with 3 dropping priorities and a default Best Effort class larger field containing QoS control information is needed. Two possible labeling configurations exists (Le Faucheur, 2002):

- A configuration called E-LSP (Exp LSP) is able to treat most of the QoS requirements. E-LSP allows one label, thus 8 different PHBs. This is enough to support a minimal configuration: EF, BE, and four AF classes with two drops precedence level.
- If a network has more than 8 PHBs, then 3 Exp bits will not be able to convey all the PHBs to LSPs. One way is to use the label itself to convey PHBs, therefore, an LSR uses the label to determine the PHB. For L-LSP PHB is determined from label and AF drop precedence is determined from Exp bits.

Both E-LSP and L-LSP can be used, but it depends on which kind of traffic is going to be forwarded. E-LSP is simpler to implement and is close to DiffServ. However, L-LSP is more sophisticated and could offer more possibilities when the QoS over MPLS is more mature. It depends on the requirements.

4.2.2 IntServ/Diffserv Mapping over DVB-S/RCS

Performances of MPLS are also determined by the capabilities of the satellite architecture lower layers, based on DVB-S/RCS technologies, especially, in terms of bandwidth allocation capabilities. MAC/DAMA concerns the satellite segment procedures for bandwidth assignments, which, on one hand, aim at an efficient use of the valuable satellite capacity while, on the other hand, should allow the respect of QoS requirements for IP traffic streams.

Two algorithms of bandwidth allocation exist:

- A Connection Admission Control (CAC) provides the RCST with a statically assigned capacity, which is always available to the ST.
- While a Bandwidth-on-Demand (BoD) mechanism provides the RCST with an additional amount of capacity which dynamically adapts to the current RCST traffic conditions.

Due to the long propagation delay associated with the satellite links, the BoD schemes implicitly introduce an additional delay. This additional delay, which is given by the

queuing delay in the RCST buffers, is in the range of 0.5-1 sec., depending on the BoD control policy.
Taking into consideration the **CoS contract** characterizing the MPLS traffic classes, the crucial point is to supply a proper mapping; possibly, the traffic classes might be supported by a mix of the capacity request categories.
The **CoS requirements** can be classified in bandwidth, delay and jitter constraints:

- Bandwidth constraints might impose that a defined amount of capacity is granted to the connection for the connection lifetime;
- Delay constraints enforce an upper limit on the queuing delay in the buffers;
- Jitter constraints require a limited delay variation in the transmission of the packets.

Due to the additional queuing delay introduced by BoD mechanisms, MPLS connections with stringent delay and jitter requirements – i.e., real time (RT) connections – cannot rely on dynamic access to the network: the proper amount of capacity, computed on the basis of the bandwidth requirements, must be statically reserved via Connection Admission Control (CAC) procedures.
On the other hand, Best Effort (BE) connections can avail of dynamic capacity assignments (via the BoD mechanism), since no particularly stringent delay requirements have to be satisfied. Furthermore, since no bandwidth requirements are defined, also the CAC procedure is not triggered; this means that BE connections are the most subject to congestions.
Finally, the connections with no jitter requirements, no (or loose) delay requirements but with bandwidth constraints – i.e., Critical Data (CD)– must be subject to CAC admission decisions but they may access a portion of reserved capacity via the BoD mechanism. In this way, these connections are not subject to congestions but, when the reserved capacity is not fully requested, the leftover part can be distributed among the best effort connections.
For each identified MPLS traffic category, the following table shows the CoS requirements, the suitable DVB priority class, the QoS guarantees and the functional modules, which allow the QoS requirements to be met.

IP classes	*CoS requirements*	*DVB priority class*	*QoS guarantees*	*CAC*	*BoD*
GS - IntServ, EF – DiffServ	Bandwidth Delay Jitter	RT	Bandwidth Delay Jitter	yes	no
CS-IntServ, AF-DiffServ	Bandwidth	CD	Bandwidth	yes	optional
Best Effort	-	BE	-	no	yes

Table 1. IntServ/DiffServ QoS classes mapping onto DVB-S/RCS

4.3 IP oriented architecture

Motivated by the need of more efficient and simplier to manage architectures, the next evolution of DVB-S2/RCS satellite network resides in IP over DVB, or at least as closed as it can. In fact, several encapsulation layers are still needed (MPE, ULE, GSE...) according to the referred architectures, but MPLS could be avoided to save its encapsulation overhead.
The satellite terminal becomes an **IP/Diffserv edge router**, and embeds all the mechanisms required to manage several PHB. The QoS is managed in three steps (Fig. 5).

- The first step take place at Medium Access Control level where the DAMA allocates the bandwidth on a fixed basis for the real time queue and on an on-demand basis for other queues (critical data and best efforts queues).
- The second step is at IP layer as described above. A dedicated IP level module implements a queue management system aiming at providing a differentiated service with regards to several service classes.
- The third step is at user layer level. The user classifies its own flows and provides the classification to a QoS server, located in the RCST, that configures its QoS architecture. The goal is to exploit the capabilities offered by the IP-level QoS.

In other words each step involves a service offering to the above.

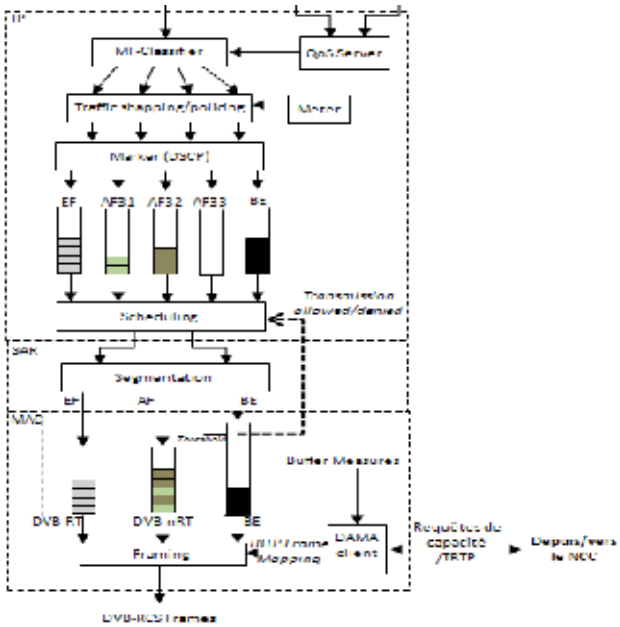


Fig. 5. The QoS Architecture

4.3.1 RCST QoS architecture components

A *Service Class* (SC) is given by the underlying applications behaviour that uses the service. In our example there exist five Service Classes for IP layer and three for MAC layer. Service classes for MAC layer are: Real-Time, Critical Data and Best Effort. Service classes for IP layer are: Real-Time (EF), three Assured Forwarding level (AF1, AF2, AF3) and one Best-Effort (BE).

- Real time (RT) service class: this is the service class for non-jitter, low delay flows typically filled with small packets. The underlying satellite bandwidth allocation mechanism (the DAMA at MAC level) should have reserved sufficient static bandwidth. This could be done at the terminal logon.
- AF service classes: these service classes deserve a special treatment with regards to best effort but with more priority. For exemple, signalling could use the AF1 queue, streaming the AF2, and web browsing the AF3.

- Best effort (BE) service Class: all the traffic that does not belong to real-time or critical data service class belongs to best effort service class.

A *scheduler* is used to differentiate IP queues and to map IP queues over MAC queues. This is a crucial component of the architecture. The EF class is generally mapped with the DVB-RT class and the AF queues are mapped over the CD queue that could use RBDC and VBDC allocation schemes. For this purpose, the traffic categories could be served by a scheduler using a simple priority queuing (PQ) discipline or a Weight Fair Queuing (WFQ). In fact, MAC queues should be as small as possible (only the necessary space to save the fragment of several IP packets) to reduce the end-to-end delay. To avoid MAC queues overflooding, interaction mechanisms (cross layering) between IP and MAC layer should be setup. For instance, when the BE queue (at MAC level) exceed a given threshold, a "backpressure" algorithm stops the scheduler to feed the MAC queue. This threshold could be set up according to the DAMA algorithm.

A *Traffic Flow* (TF) is a given data flow generated by protocol entities of a given application. A TF is uniquely recognized by a TF ID designing the originating network address, the destination network address, the originating transport service access point and the destination transport service access point. Classically in (TCP, UDP)/IP networks, this is given by the 4-tuple <IP-source, IP-destination, port source, port destination>. The *MF-Classifier* sets up the DSCP packet field according to the QoS policy.

A *shaper* is a specific algorithmic that enforces the emission profile of a flow according to its specifications (rate, delay, etc). A dual leaky bucket (DLB) is a simple solution to implement it but a Hierarchical Dual Leacky Bucket (HDLB) offers better flexibility in the queue management.

A *dropper* is a more stringent version of a shaper. In place of attempting to shape the traffic, it simply drops the non-conforming traffic. It does this by dropping each packet of the flow until the traffic goes back to a normal state. This can be implemented by reducing the input queue size of a dual leaky bucket to one.

4.3.2 Network components

The *Connection Admission Control* (CAC) algorithm in the NCC performs preventive congestion control by deciding the admission of new connections in the network. CAC decisions are taken on the basis of the QoS requirements and guarantee the availability of bandwidth and/or the fulfillment of delay requirements.

The elements in this QoS architecture that perform or can perform CAC are the following: the QoS server on the RCST side, which is informed of the user QoS requests; the ARC, Access Resource Controller (in IP-oriented architecture), which is a separate component located at the NCC side that takes the final decision according to the SLA, the network capacity, etc. To ensure the CAC procedure, a dedicated protocol has to be set up between RCST, NCC and ARC.

These components are mandatory to set up a Diffserv compliant QoS architecture in a DVB-S2/RCS system. However, this is not sufficient to have a perfect integration of satellite systems in NGN networks. The section 0 reviews the lack components.

4.3.3 Satlabs

To reach a better interoperability of DVB-RCS products, after several year of extensive work and discussions, both "Quality of Service" and "Management and Control Harmonization" Satlabs Working Groups have proposed recommendations (Combes & Pirio, 2008). With the incorporation of the recommendations into the SatLabs certification program, DVB-RCS certified that terminals will be able to interoperate in different systems requiring only changes to their configuration and to provide real-time services efficiently.
In line with the latest revision of the DVB-RCS standard, these recommendations detail how the following enhanced features operate in a DVB-RCS system:

- Harmonized support for IP Classes of Service according to the IETF DiffServ recommendations, and their mapping on DVB-RCS specific MAC layer Capacity Categories and Dynamic Requests,
- Standardized management procedures based on the IETF SNMP recommendations,
- A Common SNMP Management Information Base (MIB) for all DVB-RCS terminals,
- Standardized procedures and protocols for terminal's Configuration file and log files management,
- A Common protocol for terminal's software updates based on Multicast distribution.

As part of the short term QoS harmonization, SatLabs defines that the RCST shall support at least the following PHBs:

- Expedited Forwarding (EF)
- At least one Assured Forwarding PHB Class (AF3), as defined in RFC2597 , with at least two drop precedences
- Best effost PHB, as defined in RFC 2474.

4.4 Next Generation Satellite Networks

The following definition is an excerpt of ETSI's NGN-Starter Group conclusions: « The term NGN is commonly used to give a name to the changes to the service provision infrastructures that have already started in the telecom and IT industry. As such it is not a term that can be precisely defined but is rather an umbrella term to describe developments following PSTN/ISDN/GSM phase 2+ era.
One of the main characteristics of NGN is the *uncoupling of services and networks*, allowing them to be offered separately and to evolve independently. Therefore, in the NGN architectures proposals, there is a clear separation between the services functions and the transport functions. An *open interface* is provided between both. NGN allow the provisioning of both existing and new services independently of the network and the access type used. In NGN the *functional entities* controlling policy, sessions, media, resources, service delivery, security, etc. may be distributed over the infrastructure, including both existing and new networks. When they are physically distributed, they communicate over open interfaces. That is why the NGN architectures proposed in standards bodies and fora consist of layers and planes and show a lot of reference points. New protocols are being standardised to provide the communication between those functional entities. *Interworking* between NGN and existing networks such as PSTN, ISDN and GSM is provided by means of Gateways.
The following deals with the IP Multimedia Subsystem (IMS) architecture that could be considered as the most NGN mature architecture. The integration of DVB-RCS system in an IMS global network is presented.

4.4.1 IMS architecture

The QoS management in IMS architecture relies on several functional components:

- The Call Session Control Function (CSCF) is the standardized access point for an IMS user. It is in charge of the service policy management. According to the network architecture it could be a proxy (P-CSCF), a server (S-CSCF) or an edge entity (I-CSCF).
- The Policy Decision Function (PDF) is in charge of the service policy decisions.
- The Policy Enforcement Function (PEF) is an IP packet conditioner (traffic shaping) and a classifier.
- The Application Functions (AF) offers the services to the user.

IMS doesn't implement directly the QoS management but relies on the underlying networks. IMS standardizes the interface for the QoS management of these networks.
When considering access networks, the interface between PDF and the access link is ensured by COPS and the interface between the user terminal and the P-CSCF is based on SIP protocol.

4.4.2 Resource provisioning using SIP in DVB-RCS networks

In the IP oriented satellite architecture, the QoS functions are triggered in the RCST by the QoS Server as explain before, but this could directly be done by application servers to avoid user configuration. A SIP Proxy (Alphand, 2005) with extended functionalities has been developed to automate QoS reservations with multimedia applications. SIP (Session Initiation Protocol) should be used by multimedia applications for the session management and codec negotiation. The main functionality of a SIP proxy is to route SIP messages (session establishment for instance) from a SIP entity (an application) to another SIP entity. To remain the QoS reservation transparent for the application, the enhanced SIP proxy has to extract SDP (Session Description Protocol) descriptions from SIP messages exchanged between caller and callee, translate the parameters of each media in the session in QoS characteristics and send QoS reservations with the QoS Server on behalf of the application. The QoS-Aware SIP proxy could be centralized in the operator network part, but it could be also implemented in each RCST to reduce the connection setup time.

4.4.3 Integration of DVB-S2/RCS networks

The IP oriented architecture presented before is compatible with the implementation of the QoS in the IMS architecture (Baudoin, 2009) according to several modifications.
Compared to this medium term architecture, the P-CSCF functions are managed by a QoS-Aware SIP proxy collocated with the Gateway. The ARC should act as a PDF. It has ”enforcement” functionalities when it configures at the same time the DAMA. The QoS Server has the role of QoS enforcement function (PEF) and could be triggered by the ARC using the COPS protocol. The SIP protocol is used continuously during the QoS configuration.
However, the IMS concept is well suited to a deployment in mobile networks, where end-users are directly linked across the wireless link. Conversely, in the DVB-RCS model, end users and satellite networks are separate entities, which is the major obstacle for IMS integration in DVB-based networks. DNS or DHCP could partially provide solutions. Moreover, accounting in satellite systems is generally private and centralized. Efforts have to be done to adapt it to a larger and compatible accounting architecture.

5. Evaluation of a DVB-S/RCS satellite system

5.1 Methodologies

In this section, we will describe the different methodologies allowing evaluating the performance of a DVB-S/RCS satellite system.

5.1.1 Simulation: NS-2

Development and implementation of new services on satellite networks take time and money, so before being implemented on a real system, an implementation on a test-bed has to be done. The downside of emulation is the time needed to build and configure it, moreover it needs consequent resources and it is not easily scalable. Simulations allow overcoming some of these issues by providing a scalable and effortlessly reconfigurable network at lower cost, the major disadvantage being the realness of the results obtained since they are ruled by the model accuracy.

Network Simulator 2 (NS-2) is a discrete event simulator for networks research. Supported by DARPA and NSF, its modularity made it one of the most famous network simulators: researchers use it all around the world to implement new components or to check theoretical works. NS-2 provides protocols and communication mediums that can be settled depending on simulations needs. Updates from the development team and patches from research groups allow NS-2 to simulate current networks and latest protocols. NS-2 handles simulations of satellite networks but several components are missing, dynamic bandwidth allocation is provided by patch and there is neither QoS architecture nor encapsulation scheme. The TDMA-DAMA mechanism has been implemented and tested in (Gotta et al., 2006); the MAC layer introduced is used by our model to represent the dynamic bandwidth allocation (see section 0).

5.1.2 Advanced Emulation testbed: PLATINE

Simulation and emulation both provide the opportunity to evaluate performances, at low cost, on more or less realistic systems. When simulation needs a complete modeling of the systems from applications to physical network and operates in virtual time, emulation is more demonstrative since real applications can be deployed over the model describing transfer characteristics, delay and error behavior for instance.

The (IST SATSIX, 2004) project relies on these different methods of evaluation and validation, with the design and development of simulations, emulations and trials. The emulation testbed relies on the satellite emulation platform (PLATINE) (Baudoin et al., 2007) formerly developed within the frame of the (IST SATIP6, 2001) project. It aims to demonstrate the network and application services integration on next generation satellite systems and the possibility to interoperate with terrestrial networks. With regards to the previous release, PLATINE includes a DVB-RCS and DVB-S2 emulation, with ULE/MPEG2-TS and AAL5/ATM stacks together with the adaptive physical layer simulation and the associated radio resources management (RRM). A complete QoS architecture mixes ehanced SIP proxies (as described in 0), IP/MAC scheduling and cross layer techniques is available.

5.1.3 Real experiments: OURSES

During the (OURSES, 2006) project, we had he opportunity to use a DVB-S2/RCS system. A platform compliant with the IP oriented architecture was setup during the project. The gateway and the terminals are compliant with the Satlabs recommendations. The four (VoIP, ViC, Critical Data, Best Effort) Diffserv class of service are offered on the STM satlink 1000 terminals we used. A Service Level Agreement (SLA) is setup on the gateway side (Thales A9780 model) for each customer. It fixes the limits in terms of bandwidth with each MAC service classes. The tests were done using a satellite channel emulator and the Ka band.

5.2 Performance evaluations

5.2.1 DVB-S/RCS NS-2 simulation model with QoS

This section briefly describes the DVB-S/RCS NS-2 simulation model with QoS architecture that have been developed at LAAS/CNRS, further details on implementation and simulations can be found in (Gayraud et al., 2009). Such model can be used to simulate new protocols or to compare results with measurements obtained through emulation or experimentations done on a real link. To be efficient, architecture and behaviour have to be as closed as possible from the chosen satellite network (the one from OURSES project in our case). The model is using a TDMA-DAMA MAC layer above the physical layer defined by NS-2. The simulation made on the model without our contribution shows a really efficient behaviour with a dynamic bandwidth allocation and a fast establishment of connections. However, to be closer from the real system and to improve performances, some features need to be added:

- Dynamic encapsulation of IP packets,
- Substitution of the single queue at MAC layer by two distinct queues,
- Addition of queues at IP layer (inspired from DiffServ architecture).

Since packets fragmentation is not possible with Network Simulator 2, the MAC layer adjusts the sending time to the available bandwidth based on the assigned slots. The dynamic encapsulation (from IP to ATM frames) doesn't fragment the packets either, but their size is settled according to AAL5 protocol, resulting in a consistent overhead (around 10 percent).

QoS architecture is implemented by duplicating the queue at MAC layer and by adding buffers at IP layer: flows are aggregated, differentiated and stocked according to the DIFFSERV architecture from terrestrial network; management is done by a packet scheduler below the queues. To study the model behaviour, two kind of traffic with specific constraints were generated:

- Constant Bit Rate (CBR) needing low delay and jitter, associated to Real Time flows (RT),
- File Transfer Protocol (FTP) needing large bandwidth regardless to delay, associated to non Real Time flows (n-RT).

The chosen transport protocols are respectively UDP and TCP, the most commons for such flows. During experiments, the available bandwidth is settled to 128kbps per slot (this ratio depending on weather conditions and chosen coding scheme), since one satellite terminal can have at most two slots (256kbps), CBR rate has been settled to 128 kbps (without encapsulation overhead). Flows will compete with each other, the main point of simulations being to show the efficiency of the QoS architecture added to the model: CBR flows should

get the lowest delay possible while FTP flows would still be able to establish communication and transfer data.
The rest of this section will focus on the QoS architecture by taking a look at the model's behavior. To illustrate the competition between the two types of flows, delay suffered by the communications and the throughput they can achieved are shown on Fig. 6.

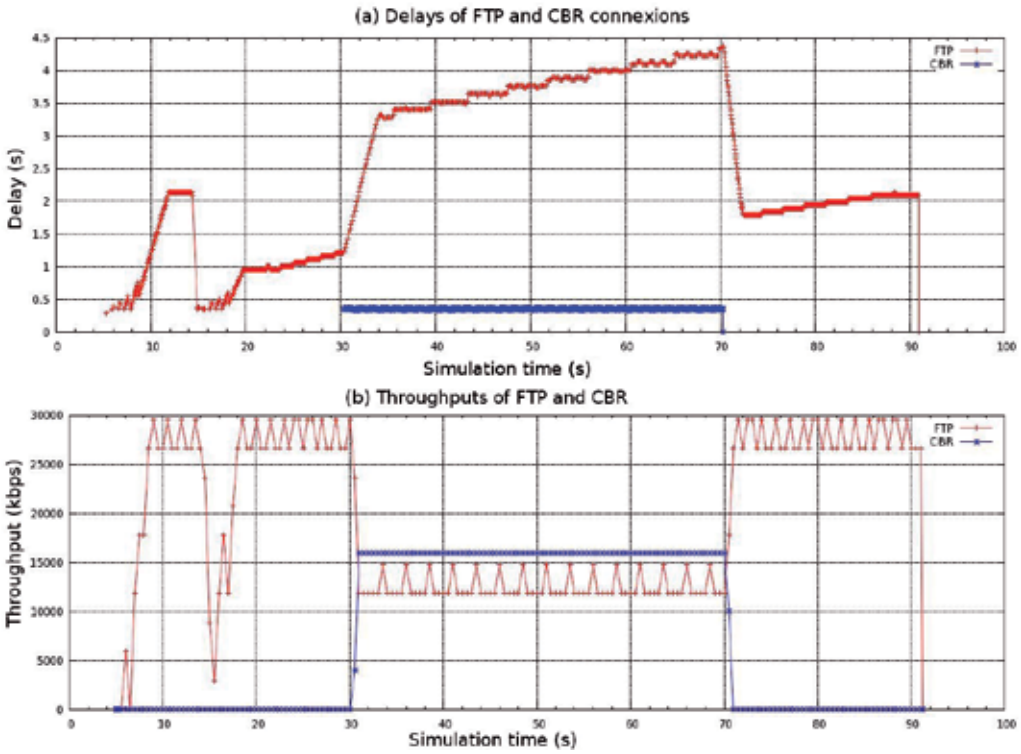


Fig. 6. a) Delay suffered by connections. b) Throughput of connections.

Differences between RT and n-RT flows are clearly visible: delay suffered by CBR is stable and below 500ms while FTP delay fluctuates and is above 3s (Fig. 6a). On Fig. 6.b, it is noticeable that FTP throughput is restricted by bandwidth taken by CBR. These two results illustrate the model's behavior by showing the differentiation done on those flows; the one with more constraints is getting the lower delay possible and enough bandwidth so no loss occurs. For n-RT flow, the throughput and the delay are fluctuating depending on network load and bandwidth allocated to the satellite terminal.
The model behaves properly and reacts as we expect: indeed; it provides an efficient QoS architecture to the basic satellite network from NS-2. But some improvements can still be done on the model: using a more efficient manager below the MAC buffers and providing a thinner encapsulation mechanism. There are also some features needing to be tested: using RED instead of DropTail policy in satellite terminal buffers or using a more realistic error model (already implemented but not used during simulations). The latest experiments were done to study SCTP (Stream Control Transport Protocol) behavior on a QoS satellite network and compare it with TCP; results can be found in (Bertaux et al., 2010).

5.2.2 PLATINE performances evaluations

In the following parts, we will show two exemples of QoS management in DVB-S2/RCS satellite systems, using the VisioSIP client (a SIP videoconferencing tool) and a QoS-aware SIP proxy (located behind each ST and based on the NIST-SIP Proxy) that send reservation or release messages to a QoS Server, located on RCSTs and able to reconfigure DiffServ queues to prioritize flows with strong time-constraints (VoIP, videoconferencing, etc...). Moreover, we consider that each ST has a total bandwidth of 1000kbps.

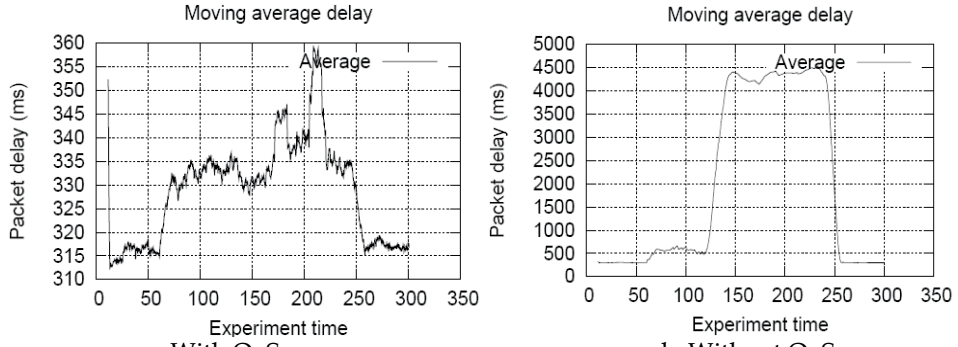
5.2.2.1 Impact of the queue management: BE vs EF

We consider here that a SIP videoconferencing session is initiated between two SIP clients located behind two separate STs. The SIP session starts at t=t0 + 10s and then 3 concurrent UDP flows (500 kbps) start respectively at t=t0+60s, t=t0+120s and t=t0+180s and terminate at t=t0+240s. Finally the SIP session ends at t=t0+300s. Moreover, 150 kbps of CRA is allocated to the studied ST to support, in terms of bandwidth, the video and audio flows.

We will make the analysis on the audio delays graphs presented on Fig. 7, but the same analysis will apply to the video delays graphs that are similar.

These two series of delays' graphs show a real benefit of the IPv6 QoS usage and a fair separation of the classes of service can be observed on the first graphs. Detail analysis of those graphs is now provided.

First, concerning the comparison of the graphs with and without QoS, it can be observed a real improvement when the QoS architecture is running especially when background traffic is high: The "moving average delay" graphs show that when two or three concurrent flows are running (between 120 and 240 ms) a very high increase of average delay is experienced by the audio flow when the QoS is not set (above 4 seconds delay) while the average delay remains below 360 ms when the QoS is set, which is compatible with audio conference requirements. In the case of high load on the satellite return link, the impact of the QoS architecture is clearly shown here.



a. With QoS b. Without QoS

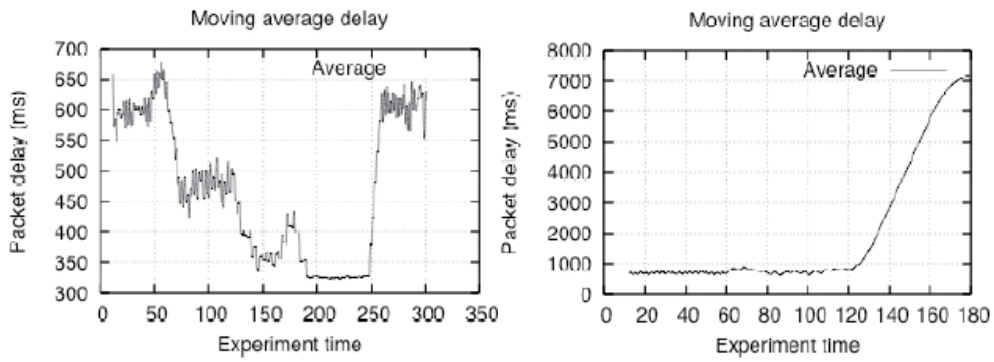
Fig. 7. Moving average delay for the audio flow

When no concurrent flows are running, delay for the audio flow is around 300ms in both cases (with and without QoS), cf. graphs between 0 and 60 seconds. This can be explained by the fact that all CRA resources, in this case, are used by the multimedia flows and no on-demand capacity is needed. When just one concurrent UDP flow of 500kbps is running the delay of

VoIP application is increasing in both cases but very slightly when QoS is set (from 315 ms up to 330 ms on average) while it's increasing up to 500 ms on average in the case no QoS is set. The capacity of the channel should be enough for both flows but the CRA capacity is not enough and on-demand RBDC bandwidth is required. So the audio flow experiences more delay when QoS is not set; this is due to the fact that all flows (audio, video and best-effort flows) are using the same MAC buffer and PVC, and so the same delay is experienced by all packets in this buffer, implied by the capacity allocation scheme. When the QoS is set, a different MAC buffer and PVC is used for high priority traffic (audio and video packets) and is served first compared to the low priority MAC buffer. Consequently, the audio flow is protected and the delay is increasing very slightly: it's experiencing an end-to-end delay compatible with audio conference application requirements (under 400ms).
Secondly, concerning the classes of service separation, we can notice on the first graph (a) with QoS that the impact on high priority classes of service of concurrent flows is rather low, and does not degrade the overall quality for end-to-end users: the delay remains below 400ms which is acceptable for interactive audio conference applications. The delay increases from 315 ms up to 360 ms, which can be explained by the sending time for large low priority packets.

5.2.2.2 Impact of the RBDC mechanism on interactive applications

The following experiments show the impact of the DAMA algorithm on interactive applications. The teleconferencing application takes the EF service class and the background traffic the Best Effort service class, but, unlike the previous experiment, there is no CRA allocated to this Satellite Terminal, all the capacity is given with RBDC requests. The teleconferencing application is first started, then 3 concurrent UDP flows of 400 kbps start and terminate at the same time than the previous experiment.
On Fig. 8.a., the delay experienced by the audio stream is less than 700 ms (this is the same values for video stream) and decreases to very low delay. The first noticeable thing is that the DAMA algorithm works fine with audio and video streams. The delay stays stable, around 650 ms, even with the throughput variation. The second noticeable thing is the delay diminution that occurs during the experiments. This can be explained by the fact that the teleconferencing application takes benefit from the RBDC requests made for background traffic as this traffic has a better priority.



a. Audio stream delay b. First UDP flow delay

Fig. 8. Moving average delay for the audio flow

On Fig. 8.b, as the link capacity is not reached, the packet delay is stable, below one second but, of course, when there is no more capacity, the delay increase, but only for the Best Effort Class.
The main problem, in this case, is that the delay of audio and video flows is often higher than what is advised in ITU-T recommandations (ITU-T, 2001), namely a value inferior to 400 ms. Consequently, to provide a solution to lower the delay given with the RBDC mechanism when no CRA (or no sufficient CRA) are allocated to a specific ST, a new extension of the SIP Proxy has been proposed to allow it to communicate with an entity located at the NCC side: the Access Resource Controller (ARC). When a SIP session is initiated, the SIP proxy can intercept the SDP, deduct the codec bitrate and ask to the ARC to increase the quantity of CRA allocated to the concerned ST corresponding to the sum of codec bitrates. The ARC checks if the SIP clients are authorized to use this service and decides to accept or reject the resource reservation.

6. Conclusion

This chapter has explained the way that can be used to provide such a satellite network client with the QoS he requested. It was proven that these QoS archtectures are feasible, that their performances are good enough by several actions like simulation, emulation, and real systems.
The work on QoS architecture is still ongoing and heterogeneous access networks mixing satellite and other radio techniques such as Wimax, and wireless systems in general. This work will lead in the very next future to the implementation of some of ours. It seems that the first network ensuring QoS may be the satellite systems that were described, designed and evaluated in the work as described in this paper.

7. References

Baudoin, C.; Dervin, M.; Berthou, P.; Gayraud, T.; Nivor, F.; Jacquemin, B.; Barvaux, D. & Nicol, J. (2007). PLATINE: DVB-S2/RCS enhanced testbed for next generation satellite networks. *Proceedings of International Workshop on IP Networking over Next-generation Satellite Systems (INNSS'07)*, pp. 251-267, ISBN: 978-0-387-75427-7, Budapest, July 2007, Springer New-York.

Bertaux, L.; Gayraud, T. & Berthou, P. (2010). How is SCTP Able to Compete with TCP on QoS Satellite Networks ? *The Second International Conference on Advances in Satellite and Space Communications (SPACOMM'10)*, Greece, June 2010.

Blake, S.; Black, D.; Carlson, M.; Davies, E.; Wang, Z. & Weiss, W. (1998). *An Architecture for Differentiated Service*, IETF RFC 2475.

Braden, R.; Clark, D. & Shenker, S. (1994). *Integrated Services in the Internet Architecture : an Overview*, IETF RFC 1633.

Braden, R.; Zhang, L.; Berson, S.; Herzog, S. & Jamin, S. (1997). *Resource ReSerVation Protocol (RSVP) – Version 1 Functional Specification*, IETF RFC 2205.

Camarillo, G.; Marshall, W. & Rosenberg, J. (2002). *Integration of Resource Management and Session Initiation Protocol (SIP)*, IETF RFC 3312.

Durham, D.; Boyle, J.; Cohen, R.; Herzog, S.; Rajan, R. & Sastry, A. (2000). *The COPS (Common Open Policy Service) Protocol*, IETF RFC 2748.

Gayraud, T.; Bertaux, L. & Berthou, P. (2009). A NS-2 Simulation model of DVB-S2/RCS Satellite network. *Proceedings of the 15th Ka and Broadband Communications – KaBand'09),* pp.663-670, Italia, September 2009.

Gotta, A.; Potorti, F. & Secchi, R (2006). Simulating Dynamic Bandwidth Allocation on Satellite Links. *Proceeding from the 2006 workshop on ns-2: the IP network simulator (WNS2)*, ISBN:1-59593-508-8, Italia, October 2006, ACM New York.

Grossman, D. (2002). *New Terminology and Clarifications for DiffServ*, IETF RFC 3260.

Handley, M.; Jacobson, V. & Perkins, C. (2006). *SDP : Session Description Protocol*, IETF RFC 4566.

Hardy, W. C. (2001). *QoS Measurements and Evaluation of Telecommunications Quality of Service*, ISBN : 0-471-49957-9, Wiley.

Heinanen, J.; Baker, F.; Weiss, W. & Wroclawski, J. (1999). *Assured Forwarding PHB Group*, IETF RFC 2597.

ISO8402 (2000). *Quality Management and Quality Assurance Vocabulary*. Technical Report, International Organization for Standardization.

ITU-T-Rec. E.800 (1993). *Terms and Definitions Related to Quality of Service and Network Performance Including Dependability*, Technical Report, International Telecommunication Union.

ITU-T-Rec. G.1010 (2001). *End-user Multimedia QoS Categories*, Technical Report, International Telecommunication Union.

Jacobson, V.; Nichols, K. & Poduri, K. (1999). *An Expedited Forwarding PHB*, IETF RFC 2598.

Nichols, K.; Jacobson, V. & Zhang, L. (1999). *A Two-bit Differentiated Services Architecture for the Internet*, IETF RFC 2638.

Rosenberg, J.; Schulzrinne, H.; Camarillo, G.; Johnston, A.; Peterson, J.; Sparks, R.; Handley, M. & Schooler, E. (2002). *SIP: Session Initiation Protocol*, IETF RFC 3261.

Shenker, S.; Partridge, C. & Guerin, R. (1997). Specification of Guaranteed Quality of Service, IETF RFC 2212.

Wroclawski, J. (1997). Specification of the Controlled-Load Network Element Service, IETF RFC 2211.

D. Awduche and al., (2001), RFC 3209: RSVP-TE: Extensions to RSVP for LSP Tunnels.

F. Le Faucheur and al. (2002), RFC 3270: Multi-Protocol Label Switching (MPLS) Support of Differentiated Services.

S. Combes, S. Pirio, (2008), ESA/ESTEC, SatLabs System Recommendations – Quality of Service Specifications.

C. Baudoin and al., (2009), On DVB Satellite Network Integration in IMS, IWSSC, Sienna, Italy.

O. Alphand, and al, (2005), QoS Architecture over DVB-RCS satellite networks in a NGN framework, Globecom, St Louis, United States.

IST SATIP6 Project, (2001), (Contract IST-2001-34344)

IST SATSIX Project (2004), (Contract IST-2004-26950)

OURSES project, (2006), http://www.ourses-project.fr

2

Antenna System for Land Mobile Satellite Communications

Basari, Kazuyuki Saito, Masaharu Takahashi and Koichi Ito
Chiba University
Japan

1. Introduction

Personal wireless communications is a true success story and has become part of people's everyday lives around the world. Whereas in the early days of mobile communications Quality of Service (QoS) was often poor, nowadays it is assumed the service will be ubiquitous, of high speech quality and the ability to watch and share streaming video or even broadcast television programs for example is driving operators to offer even higher uplink and downlink data-rates, while maintaining appropriate QoS.

Terrestrial mobile communications infrastructure has made deep inroads around the world. Even rural areas are obtaining good coverage in many countries. However, there are still geographically remote and isolated areas without good coverage, and several countries do not yet have coverage in towns and cities. On the other hand, satellite mobile communications offers the benefits of true global coverage, reaching into remote areas as well as populated areas. This has made them popular for niche markets like news reporting, marine, military and disaster relief services. However, until now there has been no wide-ranging adoption of mobile satellite communications to the mass market.

Current terrestrial mobile communication systems are inefficient in the delivery of multicast and broadcast traffic, due to network resource duplication (i.e. multiple base stations transmitting the same traffic). Satellite based mobile communications offers great advantages in delivering multicast and broadcast traffic because of their intrinsic broadcast nature. The utilization of satellites to complement terrestrial mobile communications for bringing this type of traffic to the mass market is gaining increasing support in the standards groups, as it may well be the cheapest and most efficient method of doing so.

In order to challenge the great advantages of mobile satellite communications, the Japan Aerospace Exploration Agency (JAXA) has developed and launched the largest geostationary S-band satellite called Engineering Test Satellite-VIII (ETS-VIII) to meet future requirements of mobile communications. The ETS-VIII conducted various orbital experiments in Japan and surrounding areas to verify mobile satellite communications functions, making use of a small satellite handset similar to a mobile phone. The mobile

communication technologies adopted by ETS-VIII are expected to benefit our daily life in the field of communications, broadcasting, and global positioning. Quick and accurate directions for example, can be given to emergency vehicles by means of traffic control information via satellite in the event of a disaster (JAXA, 2003).

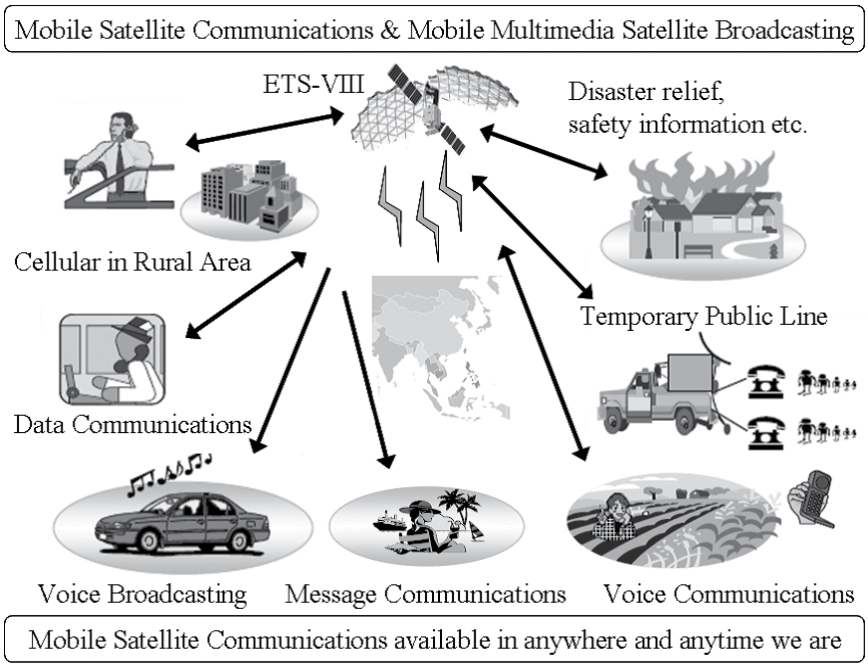


Fig. 1. Conceptual chart of mobile satellite communications and broadcasting system (JAXA & i-Space, 2003)

Figure 1 shows some of services made possible through the technological developments with the ETS-VIII. The mission of ETS-VIII is not only to improve the environment for mobile-phone based communications, but also to contribute to the development of technologies for a satellite-based multimedia broadcasting system for mobile devices. It will play as important role in the provision of services and information, such as the transmission of CD-quality audio and video; more reliable voice and data communications; global positioning of moving objects such as cars, broadcasting; faster disaster relief, etc (JAXA & i-Space, 2003).

In addition, nowadays as can be seen with the spreading of the GPS or the Electronic Toll Collection (ETC), the vehicular communications systematization is remarkable. From this phenomenon, in the near future, system for mobile satellite communications using the Internet environment will be generalized and the demand for on-board mobile satellite communications system as well as antenna is expected to increase. So far, we are enrolled in the experimental use of ETS-VIII and develop an onboard antenna system for mobile satellite communications, in particular for land vehicle applications.

In this chapter, we will figure out realization of an antenna system and establishing a mobile communication through a geostationary satellite by designing smaller and more compact antenna, developing a satellite-tracking program which utilizes Global Positioning System (GPS) receiver or gyroscope sensor, and data acquisition program which utilizes spectrum analyzer for outdoor measurement using the signal from the satellite. First, in order to minimize the bulky antenna system, a new structure of active integrated patch array antenna is proposed and developed without phase shifter circuit, to realize a light and low profile antenna system with more in reliability and high-speed beam scanning possibility. Then, the antenna system is built by the proposed antenna which its beam-tracking characteristics is determined by the control unit as the vehicle's bearing from a navigation system (either gyroscope or GPS receiver). Here, the antenna system will be installed in a vehicle and communicate with the satellite by tracking it during travelling as a concept of the antenna system.

This chapter will be divided by several sections from the research background, antenna design, numerical results, chamber measurement verification, realization on overall antenna system design, and finally antenna system verification by conducting measurement campaign using the satellite.

This chapter is organized as follows. Section 2 will provide review of mobile satellite communications systems in particular its design parameters. An example of a link budget for a mobile satellite application is given. Section 3 describes designing issues on vehicle antennas for mobile satellite system from their mechanical and electrical requirements, and also their tracking functions. In this section, we also describe our proposed antenna system, especially aimed at ETS-VIII applications. Section 4 will focus on the planar antenna design for compactness and integrated construction. It provides details about the measurement results of some basic antenna performances, such as S_{11}, axial ratio and radiation pattern characteristics that compared with the numerical results which are calculated by use of moment method. Section 5 will describe about verification of all antenna system in laboratory test and experimentally confirm in outdoor immobile-state measurement to verify the satellite-tracking performances using gyro sensor system under pre-test for field measurement campaign. The effect of radome and ground plate also will be discussed. Section 6 will show various field experiments results by utilizing the satellite to verify the validity of our developed antenna system. Overall system is tested for its performance validity not only propagation characteristics but also bit error rate performance. Finally, the last section draws conclusions on the work, and provides scope and direction for promotion in the future applications.

2. Mobile Satellite System Communication

2.1 Mobile Satellite System Architecture

Figure 2 describes a typical design for mobile satellite communication system. Three basic segments: satellite, fixed and mobile earth station are included. A propagation path is added as another fourth segment owing to its importance factor that mainly affects the channel quality of the communication system. In land mobile satellite system, the most serious propagation problem is the effect of blocking caused by buildings and surroundings objects,

which cause losing the satellite signal completely. The second problem is shadowing caused by tree and foliage, resulting the signal attenuation. The other is multipath fading, which is mainly caused by surrounding buildings, poles and trees. However, such an effect can usually be ignored when the directional antenna is used since less reflected signals approach the receiver.

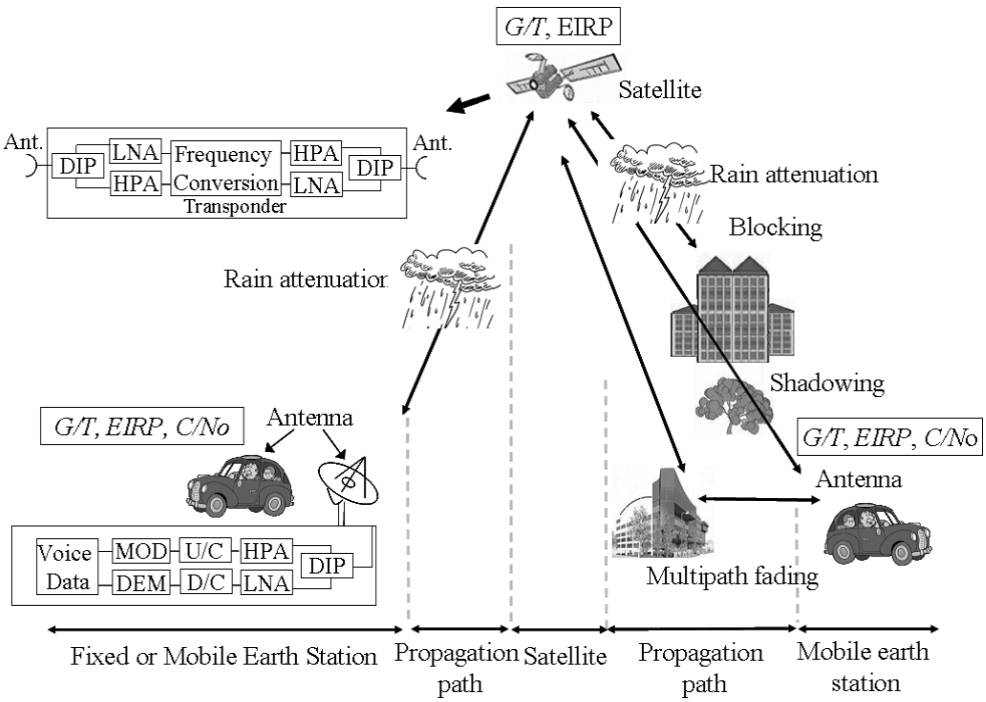


Fig. 2. Typical configuration of mobile satellite communications

Fixed or mobile earth station system consists of antenna, diplexer (DIP), up-converter and down-converter (U/C and D/C), high power amplifier (HPA) and low noise amplifier (LNA), as well as modulator (MOD) and demodulator (DEM). The satellite system is almost similar either for fixed or mobile earth station which can be constructed by antenna and up-converter and down-converter, called a transponder. Most of commercial satellites do not have modulator and demodulator. They only transmit a signal after converting its frequency and amplify the received weak signals, or usually called a bent pipe transponder or a transparent transponder.

2.2 Mobile Satellite System Link Parameters

Performance of mobile satellite system is characterized by three main parameters for link budget. Those parameters indicate the performance of three segments –namely satellite, fixed and mobile earth station– are *G/T* (ratio of antenna gain to system noise temperature or usually called figure of merit), effective isotropically radiated power (*EIRP*) and C/N_0 (ratio of carrier power to noise power density). The *G/T* and *EIRP* denote the receiving and transmitting capabilities, respectively, of satellite, fixed earth station and mobile terminal. The C/N_0 indicates the quality of the communication channel.

The *G/T* is calculated from a value of *G* which means system gain at the input port to the Low Noise Amplifier (LNA). Consequently, the ratio of antenna gain to noise temperature at the input port to the LNA can be written as:

$$\frac{G}{T} = \frac{G_{\mathrm{R}}}{T_{\mathrm{a}} + T_0\left(L_{\mathrm{f}} - 1\right) + T_{\mathrm{R}} L_{\mathrm{f}}} \ \text{(dBK)} \tag{1}$$

where *G/T*: figure of merit, G_{R}: gain of receiving antenna, T_{a}: antenna noise temperature, T_0: physical temperature when the circuit immersed, T_{R}: receiver noise temperature, L_{f}: total loss of feed lines and components such as diplexers, cables, and phase shifters (if used). As for the transmitter, the *EIRP* is one of the important parameters to describe the capabilities of transmission. The *EIRP* can be expressed as:

$$\begin{aligned} EIRP &= G_{\mathrm{T}} \cdot P_{\mathrm{T}} \ \ \text{(watts)} \\ \text{or} \ \ \left[EIRP\right] &= \left[G_{\mathrm{T}}\right] + \left[P_{\mathrm{T}}\right] \ \ \text{(dBW)} \end{aligned} \tag{2}$$

where G_{T}: gain of transmitting antenna and P_{T}: transmitted power.

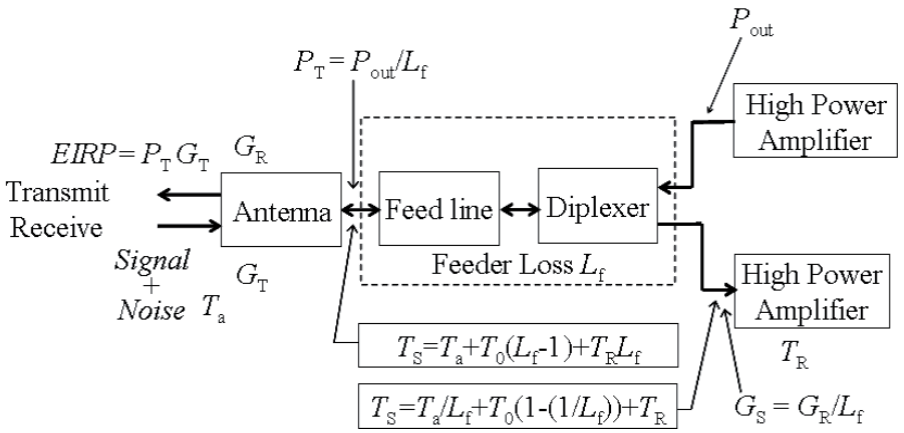


Fig. 3. Typical RF stage at earth station and satellite

In general, the radio frequency stages of earth station and satellite consist of antenna, feed line, diplexer, high power amplifier (HPA), and low noise amplifier (LNA), as shown in Fig. 3. From the figure, the ratio of input signal power (*C*) to noise power density (N_0) or simply called carrier to noise density ratio (C/N_0) at the input point to the antenna can be written as follows:

$$\begin{aligned} \frac{C}{N_0} &= \frac{EIRP}{L_{\mathrm{P}}}\left(\frac{G_{\mathrm{R}}}{T_{\mathrm{S}}}\right)\frac{1}{\kappa} \\ \left[\frac{C}{N_0}\right] &= \left[EIRP\right] - \left[L_{\mathrm{P}}\right] + \left[\frac{G_{\mathrm{R}}}{T_{\mathrm{S}}}\right] + 228.6 \end{aligned} \tag{3}$$

where L_P: free space propagation, G_R/T_S: figure of merit, and κ : Boltzman's constant (1.38×10^{-23} watt/sec/K).

The total channel quality in the system is calculated by including the uplink and downlink channels, given by:

$$\left(\frac{C}{N_0}\right)_{Total} = \left(\frac{1}{\left(\frac{C}{N_0}\right)_{Uplink}} + \frac{1}{\left(\frac{C}{N_0}\right)_{Downlink}}\right)^{-1} \quad \text{(dBHz)} \tag{4}$$

In land mobile satellite communications, the gain of mobile station is quite smaller than the satellite has, allowing the total quality is dominated by the poor uplink and the total channel quality will never exceed the uplink quality no matter how much downlink quality is increased.

Once the system channel quality is calculated, the next is what kind of modulation scheme is suitable for communication. Mobile satellite communication system currently uses digital modulation schemes, such as π/4-QPSK, OQPSK, or MSK (Lodge, 1991); low-bit rate digital voice encoder-decoder of about 4.8 to 6.7 kbps, such as vector sum excited linear prediction (VSELP), low-delay code excited linear prediction (LD-CELP), adaptive differential pulse code modulation (ADPCM), regular pulse excited linear prediction code with long-term prediction (RPE-LTP); and powerful forward error correction (FEC) technologies.

Next, another most important performance parameter for mobile satellite modulation schemes is efficiency, which includes both power and bandwidth efficiency, since mobile satellite communication systems usually have limited availability of both power and bandwidth. The power efficiency is defined as the ratio of required signal energy per bit to noise density (E_b/N_0) required to achieve a given bit error rate (BER) over an additive white Gaussian noise (AWGN) channel, although in fact, mobile satellite communication channel is Ricean fading channel. However, let we deal with performance over AWGN for simplicity in design. The bandwidth efficiency is defined as ratio of information rate R [bit/s] and the required channel bandwidth B.

For analog signals (passband signals) C/N_0 is used in the same way as E_b/N_0 for digital (baseband signals). C and E_b are related by the bit rate by:

$$C = E_b \cdot R_b \tag{5}$$

So, C/N_0 is:

$$C / N_0 = \left(E_b / N_0\right) \cdot R_b \tag{6}$$

Due C/N_0 does not relate with bit rate, as increasing bit rate does not affect C/N_0 value, but E_b/N_0 get decreasing.

Some parameters besides efficiency as described earlier, such as immunity to nonlinearity, simplicity for implementation are also be considered for designing a simple and small mobile satellite modem. Table 1 shows comparison of several modulation schemes in terms of bandwidth, E_b/N_0 for *BER* 10^{-5}, non linearity immunity, and implementation simplicity for BPSK, QPSK, OQPSK, π/4-PSK, and MSK.

* in normalized frequency offset from the center frequency
** A is highest value

Modulation scheme	Half-power Bandwidth*	Noise Bandwidth*	E_b/N_0 *for* *BER* = 10^{-5}	Nonlinearity Immunity**	Implementation Simplicity**
BPSK	0.88	1.00	9.6 dB	D	A
QPSK	0.44	0.50	9.6 dB	C	B
OQPSK and π/4-PSK	0.44	0.50	9.6 dB	B	C
MSK	0.59	0.62	9.6 dB	A	D

Table 1. Several digital modulation schemes for mobile satellite communications (Xiong, 1994)

Finally, we calculate the above mentioned parameters for link budget. Here, we design vehicle to vehicle communication via satellite for low rate data and voice communications. We assume that the data rate is 8 kbps by using a convolutional code with viterbi decoder, and 5.6 kbps PSI-CELP voice codec are suitable for conveying low data rate and voice signal from ground vehicle to another vehicle through the geostationary satellite. An example of link budget calculation is depicted in Table 2 below. From this example, by use of a large satellite antenna (i.e. very high gain antenna), we can design a small mobile earth station to enable development of a compact land vehicle communication.

Link parameter	Forward link	
Uplink		
Uplink frequency (GHz)	Vehicle	2.6575
T_x power (Watt)		1.00
Feed loss (dB)		1.70
Antenna gain (dBi)		5.00
T_x EIRP (dBW)		3.20
Tracking loss (dB)		3.00
Propagation loss (dB)		192.35
Received level (dBW)		- 190.25

Satellite antenna gain (dBi)	Satellite	43.80
Feed loss (dB)		2.60
Satellite G/T (dB/K)		14.04
System noise temperature (K)		520.00
Uplink C (dBW)		- 151.95
N_0 (dB/Hz)		- 201.44
Uplink C/N_0 (dBHz)		49.49
Downlink		
Downlink frequency (GHz)	Satellite	2.5025
T_x power (Watt)		40.00
Feed loss (dB)		2.60
Satellite gain (dBi)		43.80
Pointing loss (dB)		3.00
T_x EIRP (dBW)		54.22
Propagation loss (dB)		191.83
Received level (dBW)		- 138.71
Antenna gain (dBi)	Vehicle	5.00
Feed loss (dB)		1.70
Tracking loss (dB)		3.00
Satellite G/T (dB/K)		- 22.92
System noise temperature (K)		418.60
Downlink C (dBW)		- 138.41
N_0 (dB/Hz)		- 202.38
Downlink C/N_0 (dBHz)		63.97
Calculation Results		
Total C/N_0 (dBHz)		49.34
Bit rate (kbps)		8.00
E_b/N_0 (dB)		10.31
Coding gain (Convolutional code R=1/2, K=5, with Viterbi decoder and without interleaver) for BER=10^{-5}		5.00
Required C/N_0 (dBHz)		43.63
Margin (dB)		5.71

Table 2. Link budget calculation for land mobile satellite communication

3. Antenna System Design for Vehicle Application

3.1 Vehicle Antennas

In order to develop an antenna system for land vehicle application in mobile satellite communication system, considering the requirements of antenna properties both in mechanical and electrical characteristics is required. Thus, their characteristics are briefly explained in the following subsection.

3.1.1 Mechanical Characteristics

3.1.1.1 Compactness and Lightweight

Design of mobile antennas is required as compact and lightweight as possible to minimize the space and easy installation (Ohmori et al., 1998 & Rabinovich et al., 2010). However, a compact antenna has two major disadvantages in electrical characteristics such as low gain and wide beamwidth. Due to its low gain and limited electric power supply, it is quite difficult for mobile antennas to have enough receiving capability (i.e. G/T) and transmission power (i.e. *EIRP*). Nonetheless, such disadvantages of mobile terminals can be compensated by providing a satellite has a large antenna and huge power amplifier with enough electric power.

The second demerit is that a wide beam antenna is likely to transmit undesired signals to, and receive them from, undesired directions, which will cause interference in and from other systems. The wide beam is also suffered from multipath fading in land mobile satellite communication. Therefore, a directive sufficient-gain antenna is expected to prevent fading and interference.

3.1.1.2 Installation

Easy installation and appropriate physical shape are worthwhile requirements besides compactness and lightweight (Ohmori et al., 1998 & Rabinovich et al., 2010). The requirements antennas for cars or aircraft are different from shipborne which has enough space for antenna system installation. In case of cars, low profile and lightweight equipment is required. Aircraft antenna is required more stringent to satisfy aerodynamics standard such as low air drag (Ohmori et al., 1998). Our research concern on designing and developing a compact, lightweight and easy installation.

3.1.2 Electrical Characteristics

3.1.2.1 Frequency and Bandwidth

The Radio Regulations of International Telecommunication Union (ITU) regulates the satellite services including allocated frequency according to each region (three regions, i.e. Region 1: Europe, Russia, & Africa; Region 2: North & South America and Region 3: Asia). The typical frequencies allocated to mobile satellite communications are the L (1.6/1.5 GHz) and S (2.6/2.4) bands which being operated in the present, Ka (30/20 GHz) band and millimeter wave for future systems (ITU-R Radio Regulation, 2004).

The required frequency bandwidth is about 7% for L band, 10% for S band, and 40% for Ka band. This chapter provides an antenna for S band application with wide frequency bandwidth and will be discussed in the next subsection.

3.1.2.2 Polarization and Axial Ratio

In mobile satellite communications, circular polarized waves are used to avoid polarization tracking and Faraday rotation. When both satellite and mobile earth stations use linearly (vertical or horizontal) polarized waves, the mobile earth stations have to keep the antenna coinciding with the polarization. If the direction of the mobile antenna rotates 90^{o}, the antenna cannot receive signals from the satellite. Even if circular polarization waves are

used, the polarization mismatch loss caused by the axial ratio has to be taken into account to link budget. Generally, we design a circular polarized antenna below 3 dB axial ratio (Sri Sumantyo et al., 2005).

3.1.2.3 Gain and Beam Coverage

Required antenna gain is determined by a link budget, which is calculated by taking into account the satellite capability and the required channel quality. The channel quality (C/N_0) depends on the G/T and the EIRP values of the satellite and mobile earth stations. Typical gains are shown in Table 3 according to their application at L band satellite communications.

Antenna	Typical Gain (dBi)	Typical G/T (dBK)	Typical antenna (dimension)	Typical service
Directional	20–24	-4	Dish (1mφ)	Voice, high speed data
	17–20	-8 to -6	Dish (0.8mφ)	Ship (Inmarsat-A,B)
Semi	8–16	-18 to -10	SBF (0.4mφ)	Voice/high speed data
directional			Phased array	Aircraft (Inmarsat-Aero)
	4–8	-23 to -18	Array (2-4 elements)	Ship (Inmarsat-M)
			Helical, patch	Land mobile
Omni	0–4	-27 to -23	Quadrifilar,	Low speed data (message)
directional			Drooping-	Ship (Inmarsat-C)
			dipole	Aircraft
			Patch	Land mobile

Table 3. Typical gain for L band satellite communications (Ohmori et al., 1998 & Ilcev, 2005)

The beams of mobile antennas are required to cover the upper hemisphere independent of mobile motions. Low gain antennas have advantages in terms of establishing communication channel without tracking the satellite because of their omnidirectional beam patterns. In contrary, high gain antennas have to track satellites owing to their narrow directional beam patterns. We design a medium gain antenna owing to the use of large dimension and high gain satellite antenna in this application.

3.1.2.4 Satellite Tracking

Unlike omnidirectional antennas, medium and high gain antennas need a tracking function. Tracking capabilities depend on the beamwidth of the antennas and the speed of mobile motions, where the directional antennas with narrow beams have to track the satellite both in elevation and azimuth directions. In general, the required accuracy of tracking is considered to be within 1 dB (Ohmori et al., 1998), which is an angular accuracy within about a half of half power beam width (HPBW). However, directional antennas with relatively narrow beams

should track the satellite only in the azimuth directions because the elevation angles to the satellite are almost constant, especially in land mobile satellite communications.

Figure 4 classifies satellite-tracking functions. Tracking function divide into two function groups, namely beam steering and tracking control method. There are two types of satellite tracking systems: mechanical and electrical. A mechanical tracking system uses mechanical structures to keep the antenna in the satellite direction by utilizing a motor or mechanical drive system. An electrical tracking system tracks the satellite by electrical beam scanning.

There are two tracking algorithm, namely an opened-loop method and closed-loop method. The difference between them is whether the satellite signal is considered or not. The opened-loop uses information of mobile position and its bearing from one or several sensors regardless the satellite signal. In contrary, the closed-loop method utilizes the satellite signal to track it. To use this method, received signals from the satellite must be stable without severe fading. It is adopted in aeronautical and maritime mobile communications but quite difficult to apply in land mobile satellite communication due to its stability is predominantly affected by shadowing and fading.

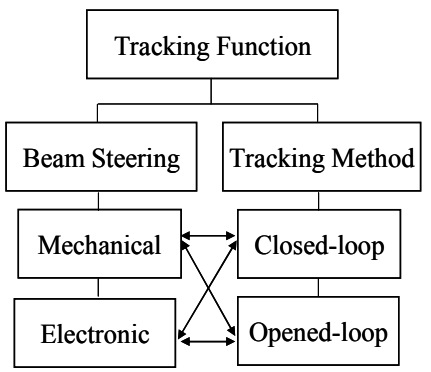


Fig. 4. Classification of satellite tracking function

3.2 Design of Vehicle-Mounted Antenna System

In mobile satellite communications, an antenna model is expected to be able to respond to changes in the direction of a mobile object. Several antennas were able to meet mobile satellite antenna requirements have been extensively investigated, are widely available in the literature include the conical beam antennas by using wire antennas such as quadrifilar or bifilar helix (Kilgus, 1975, Terada & Kagoshima, 1991, Nakano et al., 1991, Yamaguchi & Ebine, 1997), drooping dipole (Gatti & Nybakken, 1990) or even patch antenna in higher mode operation (Nakano et al., 1990, & Ohmine et al., 1996) and the satellite-tracking antennas (Ito et al., 1988). As described in the previous subsections, an attractive feature of the former antenna design is that, as the radiation is omnidirectional in the conical-cut direction and also their beam is broad in the elevation plane, satellite-tracking is not necessary. However, such antennas offer typical gain about 0 - 4 dBi (Ohmori et al., 1998, Ilcev, 2005, Fujimoto & James, 2008) because of their isotropic energy in the conical-cut direction. Further, owing to our application target for ETS-VIII, which is described by a link budget calculation in Table 2, the gain is designed by more than 5 dBi in the overall azimuth

coverage area at specified elevation angles (Table 4). Therefore, in case of omnidirectional antenna, their typical gain will not satisfy the specification. Therefore, a beam-tracking antenna is selected to suit the target. By utilizing a beam-tracking antenna, owing to its directional beam property, the beam can be deflected towards the satellite direction when the vehicle moves. Although such antenna type needs a tracking function, owing to the generation of directional beam and smaller fading effects from surrounding terrain, higher transmission rate are possible.

Most recent decades of the developed antenna system for vehicle-based applications are impractical since their design, based on mechanical steering that makes them extremely bulky. This type of antenna system is heavyweight and high power consumption as well as low tracking-speed owing to the use of electric motors responsible for mechanical steering (Kuramoto et al., 1988, Huang & Densmore, 1991, Jongejans et al., 1993, Strickland, 1995). An alternative solution is a planar phased array antenna which performs beam steering by electronic means (Nishikawa et al., 1989, Ohmori et al., 1990, Alonso et al., 1996, Konishi, 2003). However, the use of phase shifters for beam forming is quite expensive owing to their large quantities requirement. Such phase shifters, need to be properly designed in order to avoid the beam squinting in which the beam direction may considerably differ at receive and transmit frequencies. Moreover, nonlinear effects from electronic phase shifter and switches generate the noise problem in phased array antenna (Ohmori, 1999). Design of antenna system requires as compact and lightweight as possible to minimize the space and easy installation. In the following subsection, we describe specifications, targets and structure of our vehicle antenna system.

3.2.1 Specifications and Targets

The specifications and targets of the antenna are shown in Table 4. The ETS-VIII is providing voice/data communications with satellite mobile terminals in the S-band frequency (2.5025 GHz and 2.6575 GHz for reception and transmission, respectively). The polarization is left-handed circular (LHCP) for both transmission and reception units. As this antenna is assumed to be used in Tokyo and its vicinity, the targeted elevation angle is set to 48°. In the system, the antenna beam is expected to be steered towards the satellite and cover the whole azimuth space by more than 5 dBic and less than 3 dB for the gain and the axial ratio, respectively.

Specifications		
Frequency bands	Transmission (T_x)	2655.5–2658.0 MHz
	Reception (R_x)	2500.5–2503.0 MHz
Polarization	Left-handed circular polarization for both T_x and R_x	
Targets		
Angular ranges	Elevation angle *(El)*	48° (Tokyo) ±10°
	Azimuth angle *(Az)*	0° to 360°
Minimum gain	5 dBic	
Maximum axial ratio	3 dB	

Table 4. Specifications and targets for ETS-VIII applications

3.2.2 Antenna System Architecture

Figure 5 depicts a satellite-tracking system built with the beam switching method. As shown in this figure, the localization of the satellite is determined, based on the location and travelling direction of the mobile station by use of currently available car navigation systems and gyroscope, and the appropriate beam direction is selected. Then the signal emitted by the tracking unit is received and by appropriately controlling the activation of the feedings of each element, through the switching circuit used to control the feeding of the antenna, the beam is switched in three directions in the azimuth plane and the satellite can be followed. Since the antenna system utilizes the gyroscope, the satellite-tracking can be kept as the GPS satellite is out of sight.

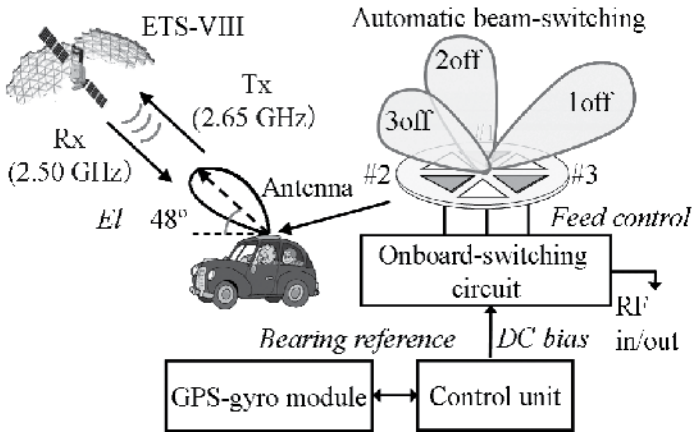
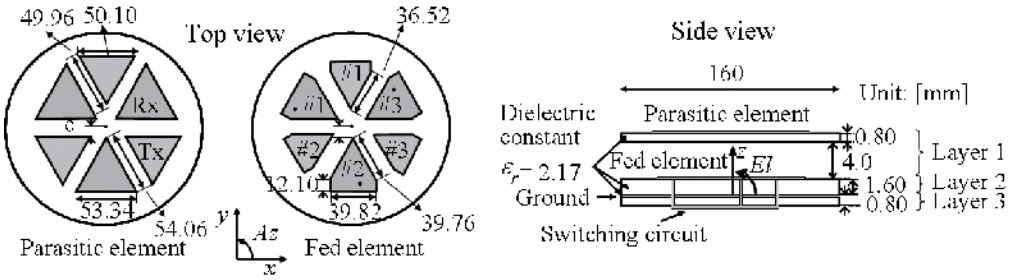


Fig. 5. Antenna system architecture

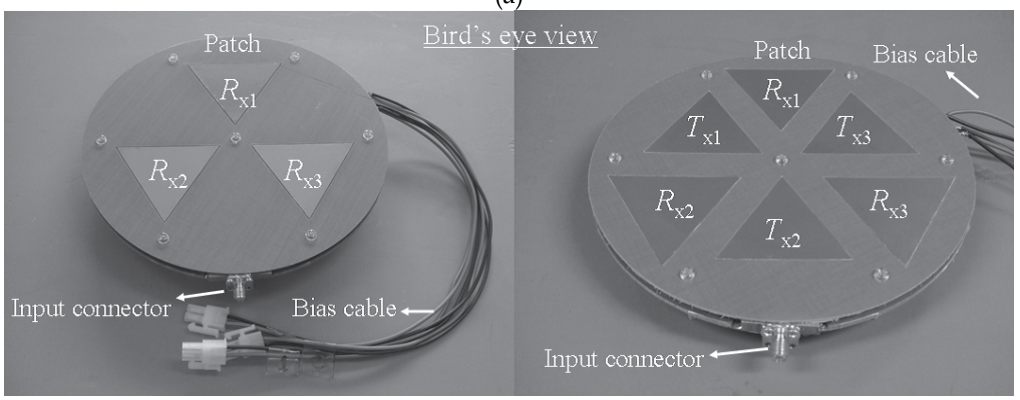
The array antenna configuration has a beam switching capability, where in principle, n circularly polarized elements are ($360°/n$) sequentially rotated (Teshirogi et al., 1983 & Hall et al., 1989) and set with an equal distance between each elements following a circular path. Our antenna is a 120° sequentially physical rotated and set with an equal distance between each element following a circular path. With such alignment, each element is fed in-phase allowing their relative phase is physically shifted. The feeding of each antenna element is successively turned off by controlling the onboard-switching circuit and thus the whole azimuth range can be scanned by step of 120°. Three beams can be generated to cover all of the azimuth angles. The beam is generated in the azimuth plane at -90° from the element that is turned off. As a result, if each element i.e. element no. 1, 2 and 3 is turned off, the beam is generated in the direction Az = 0, 120° and 240° (Sri Sumantyo et al., 2005), respectively as shown in Fig. 1. In addition, the satellite-tracking is conducted in the azimuth plane regardless considering the elevation direction owing to the antenna gain is predicted quite enough to communicate with the geostationary satellite as earlier listed in Table 2, Section 2.2.

The antenna system is operated by a control unit allowing the antenna beam is automatically steered. The beam-forming of array antenna is generated by providing bias voltages to switch on and off the P-I-N diodes on the onboard-switching circuit and thus two elements of the array can be correctly fed and afterwards a desired beam is created

among three selectable-beams. As the ETS VIII satellite lies at southern from the Japanese archipelago, by considering position of array antenna installation on car's roof, the antenna beam can be invariably controlled in south direction. By such a way, antenna tracks the satellite in good reliability and high-speed beam scanning.



(a)



(b) (c)

Fig. 6. Structure and fabricated antenna: (a) proposed construction, (b) receive-use antenna, (c) receive and transmit-use antenna

3.2.3 Array Antenna Structure

We basically design an antenna for both of transmit and receive use by arranging them on the same layer for achieving a compactness. We develop the antenna started from a single element, a receive-use antenna up to a final structure namely a receive–transmit-use antenna. Examples of antenna structures and their developed antenna are depicted in Fig. 6. The antenna is composed of three layers, i.e. parasitic elements with air gap (layer 1), fed elements (layer 2) and switching circuit (layer 3). The fed elements are three pentagonal patch antennas that directly excited from the feeding line on layer 3. In the top of the construction is put three isosceles triangular patches as parasitic elements to enhance bandwidth and gain of the antenna. In order to match on 50-ohm feeding, air gap is inserted in the layer between fed and parasitic elements. This design excites two near-degenerate orthogonal modes of equal amplitudes and 90° phase difference for left-handed circular polarization (LHCP) operation. Good axial ratio performance can be obtained by adjusting position of feeding point, air gap height, and parasitic patch size.

The antenna structure is integrated with a power divider and a switching circuit, which mounted on the backside of the structure. The circuit is functioned as a feeding control of the antenna. The mounted circuit composed of a simple power divider and a Double Pole Triple Throw (DP3T) switching circuit developed by (Kaneko et al., 2006). This integrated antenna allows its compactness and low loss because no additional cable required.

4. Laboratory Test Performance

In earlier section, configuration of array antenna structure is proposed for receive-use and receive–transmit-use as well. Nevertheless, due to satellite problem (NICT, 2007), we utilize the antenna for receive-use only in our measurements. Basically, the transmit-use antenna performance and its characteristics are similar to the receive-use one. Once we fabricate the antenna, we test in laboratory for some basic antenna measurements. Following is the measurement results such as S_{11}, axial ratio, and radiation pattern characteristics.

4.1 Input Characteristics

Owing to the stacked-parasitic patch structure, two generated-resonant modes enhance impedance bandwidth as described in Fig. 7(a). Impedance bandwidth of the antenna i.e. ($|S_{11}|$ < -10 dB) is about 8.50%, quite enough for ETS-VIII applications. Moreover, the measured antenna gives a good input matching at the target frequency 2.5025 GHz by (50.28 – j24.86) ohm.

4.2 Axial Ratio Characteristics

The array antenna gives good performance at El = 48° in the target frequency. Good axial ratio is required to eliminate polarization tracking because of circular polarization. The measured result shows the axial ratio is 1.0 dB at center frequency 2.5025 GHz for each of three generated-beams. In addition, the 3 dB axial ratio bandwidth gives about 1.8% as shown in Fig. 7(b).

4.3 Radiation Characteristics

4.3.1 Radiation Pattern in Elevation-Cut Plane

Figure 7(c) shows the radiation characteristics of the array antenna in the elevation-cut plane when element no. 1 is switched off. The antenna main beam is generated at Az = 0 which is shown at the right side of the figure. Same manner is obtained in case of no. 2 is off and no.3 is off where each beam occurs at Az = 120^o and Az = 240^o, respectively. Noted that the gain more than 5.2 dBic and the axial ratio less than 1.7 dB meets the requirements for elevation angle El = 38^o–58^o where this is an approximation of elevation range of the satellite by considering the Japanese archipelago from northern to southern. We confirmed that the gain is 6.6 dBic and the axial ratio 1.2 dB at El = 48^o. The measurement is taken at center frequency 2.5025 GHz. These results surely allow us to track the satellite regardless its elevation with regard to vehicle.

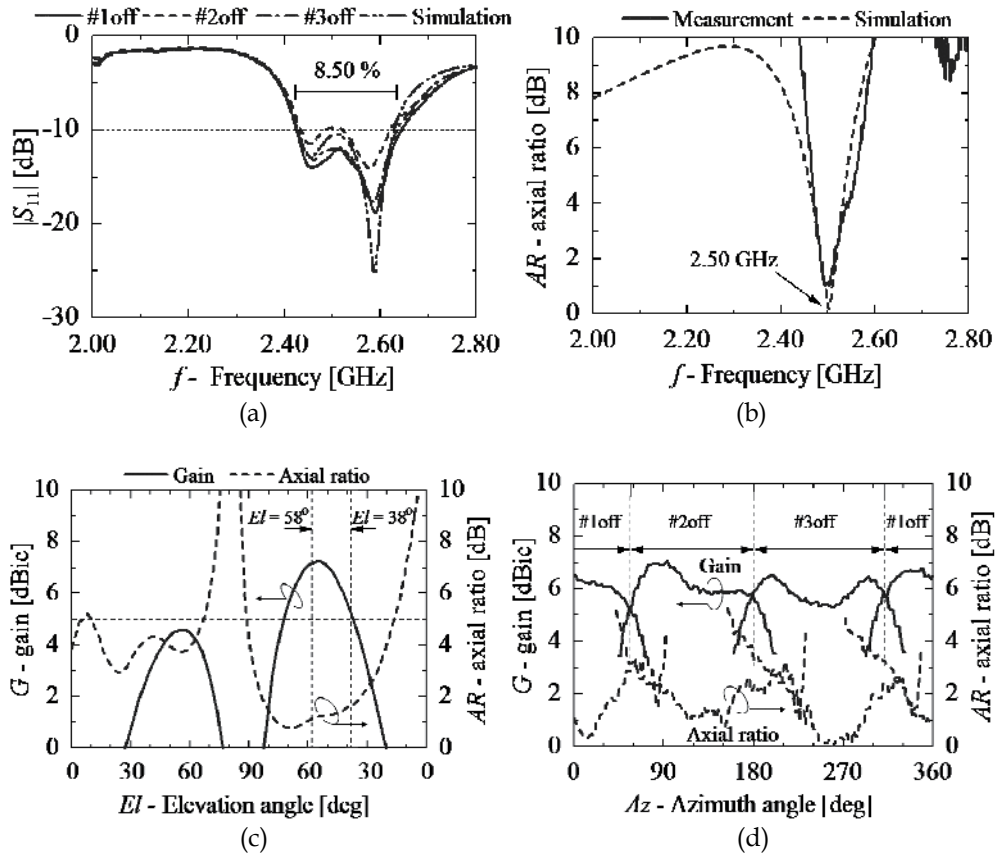


Fig. 7. Antenna performance in laboratory test measurement: (a) S_{11} (b) axial ratio characteristics, (c) radiation pattern in *El*-plane and (d) radiation pattern in *Az*-plane

4.3.2 Radiation Pattern for Beam-Switching

Figure 7(d) shows the measurement results of gain and axial ratio for each of three antenna beams in the azimuth plane at $El = 48^{o}$. Small difference among each antenna beam-shape is observed. Such unsymmetrical property is considered due to the phase difference effect of the switching circuit and dimensionally discrepancy antenna fabrication. However, the 5 dBic-coverage in the 360^{o} of the conical direction is satisfied enough. In addition, the beam is possibly switched at minimum gain 5.2 dBic and the axial ratio below 3 dB is possibly to be obtained for automatic satellite-tracking.

5. Verification of Antenna System

Following all the required components of the system are developed and individually examined, testing of the completed array antenna is performed in anechoic chamber. Here, we test the tracking system performance of the antenna with regard to the conical-cut radiation pattern at specified elevation angle. Figure 8(a) illustrates an antenna system measurement set up where a transmitting (*T*x) antenna is employed as though a satellite and

the array antenna as a receiving antenna. The antenna is covered by a radome and put on a ground plane because in mobile measurement campaign the antenna will be located on car roof as well as to avoid mechanical restriction or weather hindrance like snow, wind, and rainfall. The antenna is put 8 mm upper from the ground plate to avoid the switching circuit of the antenna on the rear-side touches the ground plane which may cause a shorted-circuit due to the bias cables connections. This measurement evaluates capability of the array antenna allowing the beam automatically switched pursuing the transmitting antenna.

In order to realize such a measurement we developed a simple control program on PC to control the antenna beam by use of a gyro sensor inside of GPS module. The measurement is performed without and with ground plane–radome. Such measurement is carried out to grasp the influence of ground plane and radome on the antenna performance specifically the gain and the axial ratio in azimuth scanning. In fact, we confirmed that a hemisphere radome is less effect on gain and axial ratio performance as well.

The beam of the antenna is generated by a mechanism that consists of switching off one of the radiating elements as earlier mentioned in Section 3.2.2 where the tested results is depicted in Fig. 7(d). Fig. 7(d) represents the beam-switching characteristics in manual operation that each beam is separately measured. Having performed a manual beam measurement in Fig. 7(d), we decide at which point we should to switch the beam automatically with regard to the gain value at the coincide point of each beam. Since the beam is possibly to be switched at minimum gain more than 5 dBic, then the antenna beam is switched at specific azimuth angle (this measurement is at Az = 56^o, 180^o and 312^o). With such decision, we confirmed that the gain can be switched automatically at the given azimuth angles and the axial ratio for each beam satisfies below 3 dB to cover 360^o conical-plane. This tracking result is shown in Fig. 8(b). In this case, the beam is switched by itself as the azimuth angle changes to track the transmitting antenna.

In order to figure out the differences of the antenna characteristics caused by ground plane and radome installed, Fig. 8(c) depicts the automatic tracking when ground plane and radome installed. It is shown that the characteristics of each beam does not change drastically when the radome and ground plate are employed (Fig. 8(c)) compared with the only antenna structure (Fig. 8(b)), except the axial ratio owing to the scattering from the ground plane affects the incident phase onto the antenna. The effect of radome is considerably neglected since its hemisphere shape gives minimum scattering and its thin material provided less loss. In general, our antenna system is able to automatically control and correctly select the beam in the azimuth direction whose gain more than 5 dBic and axial ratio less than 3 dB, although when the ground plane is mounted, its axial ratio rises by 0.5 dB at the coincide point.

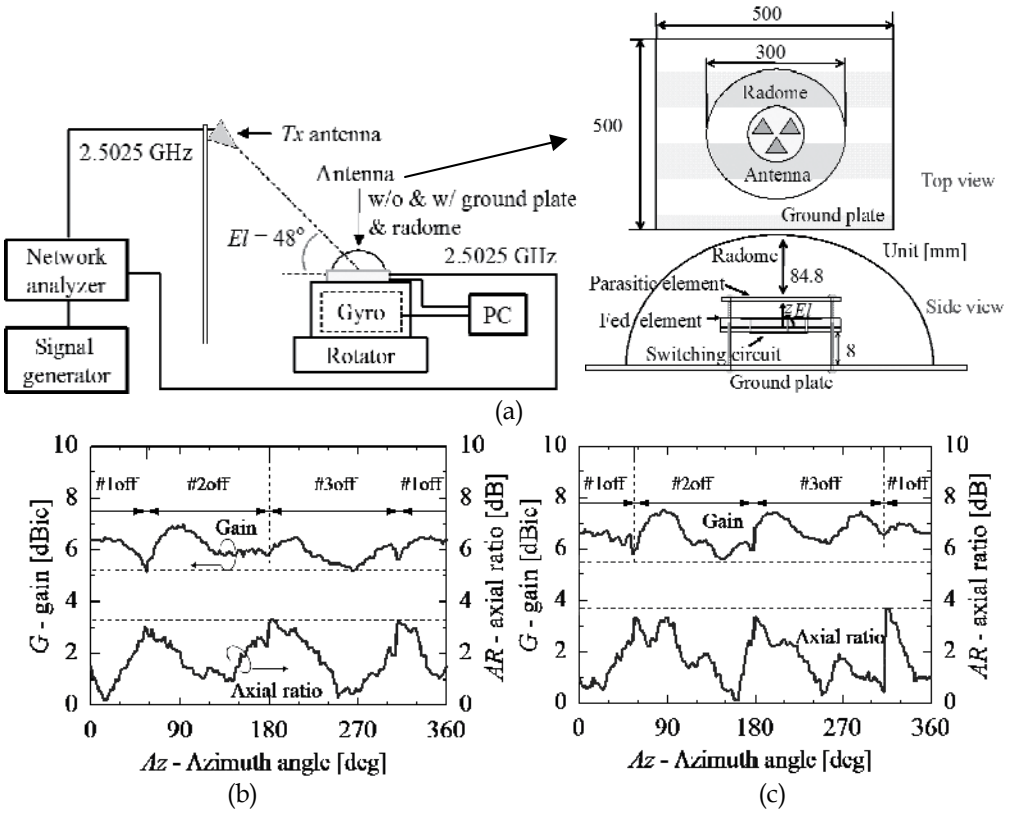


Fig. 8. Verification of antenna system in laboratory anechoic chamber: (a) Measurement set up (b) antenna performance without radome and ground, (c) antenna performance with radome and ground

Additionally, before we perform a measurement campaign, we also confirm the validity of the overall antenna system by carrying out an outdoor measurement using signal from the ETS-VIII satellite. The antenna system is tested by a fixed-state testing rig in Chiba Prefecture area (El = 48°) without obstacle present (direct signal) as illustrated in Fig. 9(a). This measurement utilizes a spectrum analyzer (Agilent E4403B) to measure the received signal power from the satellite. In order to compensate the weak satellite signal, an amplifier (Agilent 83017A) is associated with the antenna and thus the signal level has enough C/N_0. Measured result showed that C/N_0 is about 47.30 dBHz with link margin 1.45 dB, sufficiently to make a satellite-tracking measurement.

The measurement of satellite-tracking for array antenna is performed as same as the antenna system test in the anechoic chamber. For this purpose, we take the received signal power while the antenna is rotating. A gyro sensor inside of GPS module is put on the beneath of the antenna and connected to PC for automatic satellite-tracking. As a result, the satellite-tracking is well operated with good level as depicted in Fig. 9(b). Moreover, three selectable beams are smoothly switched to cover all the azimuth direction.

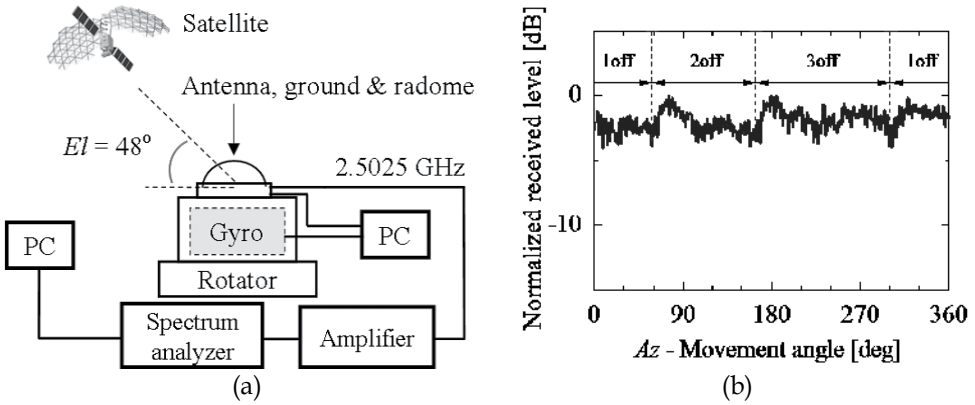


Fig. 9. Verification of antenna system in outdoor test: (a) Measurement set up (b) satellite tracking performance

6. Measurement Campaign using Geostationary Satellite

We have validated the antenna system performance by immobile-state measurements in Section 5. Afterwards, we carry out a measurement campaign to verify possibility for application in real environment. Reported by (NICT, 2007), however, large deployable reflector (LDR) antenna that installed on ETS-VIII satellite cannot be used because of improper situation at power supply of low noise amplifier (PS-LNA), thus the measurement campaign is conducted by using high accuracy clock (HAC) receiving antenna with lower gain 25 dBi instead of 43.80 dBi of the LDR antenna. Because of such circumstance, the measurement is also assigned only for forward link namely from ground fixed-station (transmitter) to vehicle (receiver) through the ETS-VIII satellite (Basari et al., 2009). In this case, we modify the calculated-link budget instead of listed-parameters in Table 2, to deal with the situation, for 8 kbps of transmission rate and BER = 1.0×10^{-4} . We use a parabolic antenna (gain 22.4 dBi) at the transmitter to boost a transmitted signal. Afterwards, we measure received signal power and average bit error rate (BER) at the receiver. The received signal power is retrieved from intermediate frequency (IF) signal of handset terminal. Configuration of measurement is represented by a block diagram in Fig. 10(a) and an example of measurement is viewed in Fig. 10(b).

6.1 Verification of Satellite-Tracking

At first, verifying a single antenna beam is performed in line of sight area without obstacles present. We utilize a spectrum analyzer (Agilent E4403B) to retrieve the IF signal amplitude. By setting it at zero-span in specified frequency, received signal power can be recorded. One of three selectable beams of the antenna is viewed in Fig. 11(a). The experiment result tends to agree with the calculation even though small difference is observed. We employ an aluminium ground plane on the antenna under which is mounted (Fig. 10(b)). Such plate affects the antenna performance especially its beam. Small difference exists, particularly the beam-shape and the beam-width. However, such differences do not significantly worsen the received signal of the antenna.

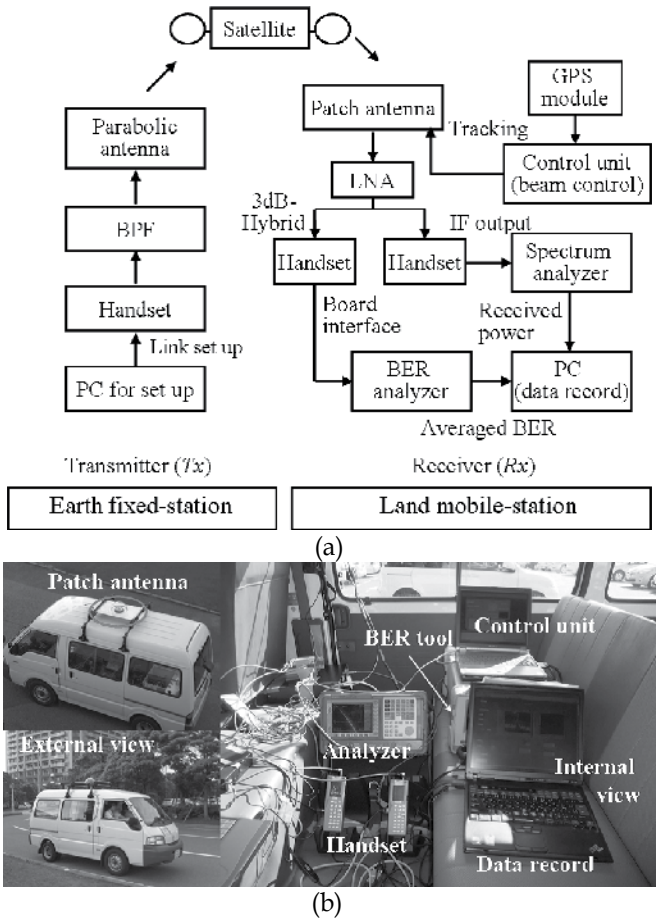
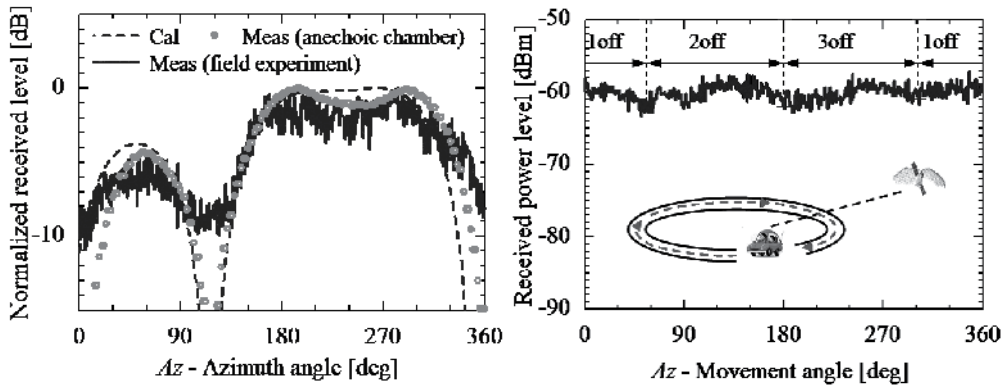


Fig. 10. Experimental set-up for measurement campaign: (a) block diagram of configuration (b) measurement view

The measurement also verifies the satellite-tracking of antenna system. While vehicle moves around at the rotary, the antenna beam is electronically steered pursuing the satellite associated with vehicle's orientation. Three selectable beams are smoothly switched towards the satellite in azimuth direction as described in Fig. 11(b). In addition, due to we use spectrum analyzer for measuring signal power, the C/N_0 can be calculated by the following formula (Hranac & Currivan, 2006 and Agilent, 2003):

$$C/N_0 = (\text{signal power}) - (\text{noise floor}) + (10 \log \text{RBW}) \quad (7)$$

where, RBW is resolution bandwidth of measurement. By considering the received signal power is averaged by -60 dBm and noise floor -83 dBm, the C/N_0 is approximately 47.77 dBHz and link margin 1.94 dB is obtained.



(a) (b)

Fig. 11. Verification of antenna system measurement using satellite in line of sight places: (a) Single-beam verification (b) automatic satellite tracking verification

6.2 Evaluation in Blockage Areas

In this measurement, we thoroughly evaluate the effect of obstacle objects namely buildings, foliages of trees, utility poles such as electricity power poles, and highway overpasses or bridge roof, to the received signal qualities of the antenna. Fig. 12(a) shows an experiment at buildings area. There are two high buildings on the test field which expected to block the signal from the satellite. While vehicle moves on straight path, the signal from the satellite is blocked becoming same level with the noise floor and thus it cannot be detected well as a received signal. If it happens for too long period, communication link will be disconnected.

We also examine the blockage due to roadside-trees. The measurement is carried out in summer season since dense foliages exist and will affect the performance of the received signal power. As shown in Fig. 12(b), the measurement is performed on straight path where the satellite is situated at the left side of the measurement path. Unlike the blockage due to buildings, roadside-trees are less effect to the received signal even though they barred by an average 3–5 dB, decreasing well-received signal. It is obvious shown in Fig. 12(b) in comparison with the direct wave signal. In addition, almost 75% of the satellite signal is approximately attenuated by 2–6 dB at this dense foliages area.

In order to confirm the antenna system usability in common land mobile applications, we test the system in expressway and common road. The field experiment of expressway is carried out in Chiba prefecture area (El = 48°) between Kisarazu-kita and Soga interchange by speed 70–80 km/h. With almost no obstacles present, well received signals are obtained and yet the beam-switch happens at the beginning and the last measurement path. We confirmed that the signal drastically dropped in very short time (0.2–0.4 s) when the vehicle passes through overpasses or bridges. This result is shown in Fig. 13(a).

Our antenna has radiation characteristics in the elevation direction, so thus it is required to confirm with real environment for vehicle application. For that purpose, we test the antenna system at inclined-road. The vehicle passes on the 5° inclined-road with speed of motion is constantly kept by 40 km/h without any beam-switch happens. Fig. 13(b) shows an example

of the measured satellite signal for downward-motion. The received signal tends to increase in 0.5 dB. In this case, downward-motion means increasing elevation allowing increasing the antenna gain.

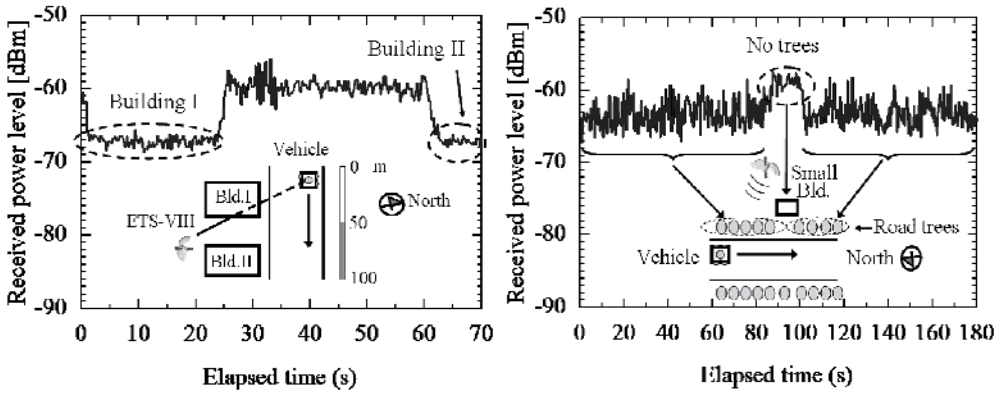


Fig. 12. Evaluation on blockage signal measurements: (a) due to buildings (b) due to roadside trees

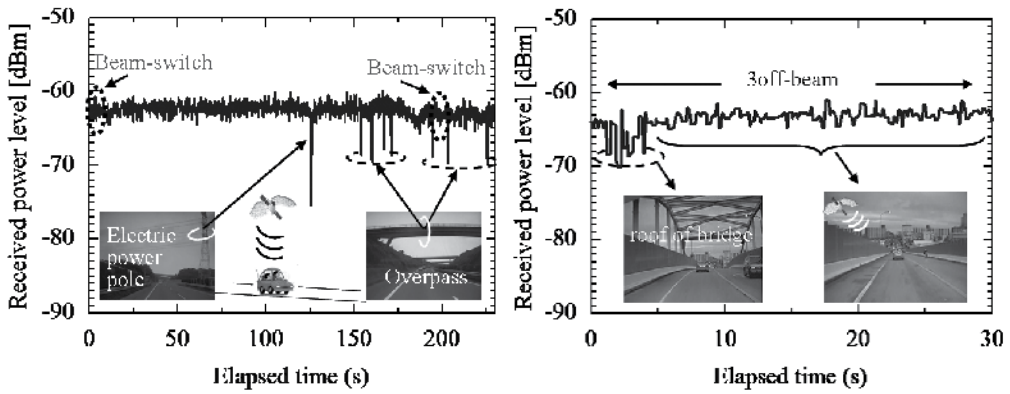


Fig. 13. Evaluation on blockage signal measurements: (a) due to overpasses and pole (b) due to roof of bridge and inclined-road constraint

6.3 Bit Error Rate Verification

As we earlier designed the antenna system, BER is set by 1.0×10^{-4}. Therefore, we confirm it by performing a measurement on BER characteristics besides the propagation measurements. Here, the measurement is carried out at rotary in line of sight area to verify the satellite-tracking. As previously reported, binary-phase shift keying (BPSK)-modulated pseudo-noise sequences (PN-code) is transmitted from the ground-station through the satellite and received at our vehicle-station. We use a BER analyzer (Anritsu MD6420A) to retrieve the BER value that measured from interface board to which connected with the analyzer from handset terminal. The modulation rate is set to be 8 kbps. The measurement circumstance is similar to measurement path in Subsection 6.1 for satellite-tracking. As

shown in Fig. 14, the BER performance is kept stable in range 5–7×10-4 while beam-switching occurs. Thus, validity of our antenna system design in the beginning Section is confirmed.

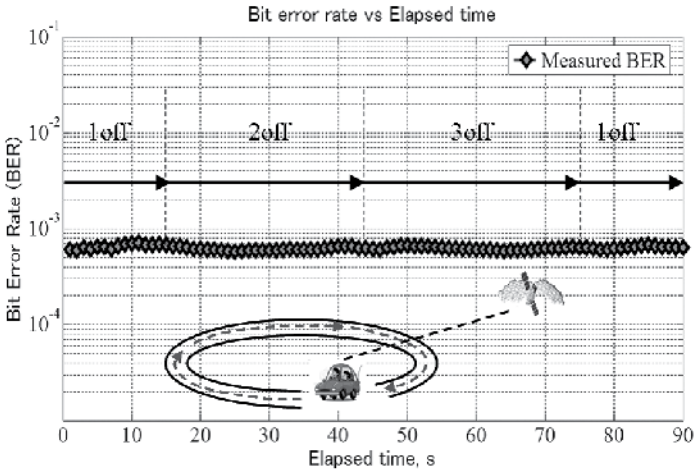


Fig. 14. Bit error rate characteristics on satellite-tracking

7. Conclusion

Simple selectable-beam antenna system for vehicle-mounted mobile satellite communication applications was presented. We have discussed it from design, laboratory test as well as outdoor test and verification in measurement campaign. System components discussed include array antenna includes its switching system to activate the array, and satellite-tracking function as well. The measurement in anechoic chamber was thoroughly examined with satisfactory performances, well suited with the calculations. The basic measurement included S_{11}, axial ratio, and radiation pattern characteristics. Overall antenna system performances in anechoic chamber test gave satisfactory results. Further, the antenna system was also confirmed by using signal from satellite for immobile-state outdoor measurement on the testing-rig. Without any obstacles present, the system was able to correctly select the beam for tracking the satellite with regard to the rotation of the antenna.

Following the satisfactory performances in anechoic chamber measurement and fixed-state outdoor measurement, the antenna system was also examined in measurement campaign using the satellite signal to verify the validity of the developed system for possibility use in land mobile satellite applications. We built a connection with the satellite and tested the antenna system in particular its tracking capability. As a result, the system was correctly track the satellite while vehicle was moving by considering its position and bearing information that retrieved from GPS module.

In addition, we also considered and tested the antenna system under environment constraints that affecting its received signal power, such as buildings, roadside-trees, utility poles, highway overpasses and inclined-road present. The results showed us the blockage and shadowing happened and significantly attenuated the received signal owing to its

dependence on direct wave signal. Moreover, the BER performance of the system has been verified with satisfactory result by about 10^{-4}.

In addition, this vehicle-mounted antenna system design and its measurement results will help engineers and practitioners to design a costly-effective land-mobile satellite communication system. Ultimately, our developed antenna system is simple, compact and promising in low cost, for contribution in the future mobile satellite communication applications.

8. References

Agilent Technologies (2003). Spectrum Analyzer Measurements and Noise: Measuring Noise and Noise like Digital Communications Signals with a Spectrum Analyzer. *Application Note 1303*.

Alonso, J.I.; Blas, J.M.; Garcia, L.E.; Ramos, J.; de Pablos, J.; Grajal, J.; Gentili, G.G.; Gismero, J. & Perez, F. (1996). Low cost electronically steered antenna and receiver system for mobile satellite communications. *IEEE Trans. on Microwave Theory and Techniques,* Vol. 44, No. 12, (Dec. 1996) pp. 2438–2449, ISSN: 0018-9480

Basari; Purnomo, M.F.E; Saito, K.; Takahashi, M. & Ito, K. (2009). Realization of simple antenna system using ETS-VIII satellite for land vehicle communications. *IEICE Trans. Communications,* Vol. E92-B, No. 11, (Nov. 2009) pp. 3375–3383, ISSN: 0916-8516

Basari; Purnomo, M.F.E; Saito, K.; Takahashi, M. & Ito, K. (2009). Simple switched-beam array antenna system for mobile satellite communications. *IEICE Trans. Communications,* Vol. E92-B, No. 12, (Dec. 2009) pp. 3861–3868, ISSN: 0916-8516

Fujimoto, K. & James, J.R. (2008). *Mobile Antenna Systems Handbook,* Artech House, ISBN: 978-1-59693-126-8, Norwood, MA, USA

Gatti, M.S. & Nybakken, D.J. (1990). A circularly polarized crossed drooping dipole antenna, *Proceedings of IEEE Antennas and Propagation Society Intl. Symposium*, pp. 254-257, ISBN: 978-0780370708, Dallas, TX, USA, May 1990, IEEE, Piscataway, NJ

Hall, P.S.; Dahele, J.S. & James, J.R. (1989). Design principles of sequentially fed, wide bandwidth, circularly polarised microstrip antennas. *IEE Proceedings H Microwaves Antennas Propagation,* Vol. 136, No. 5, (Oct. 1989) pp. 381– 389, ISSN: 0950-107X

Hranac, R. & Currivan, B. (2006). Digital Transmission: Carrier-to-Noise Ratio, Signal-to-Noise Ratio, and Modulation Error Ratio. *Broadcom Corporation and Cisco Systems, Inc.*, white paper.

Huang, J. & Densmore, A.C. (1991). Microstrip yagi array antenna for mobile satellite vehicle application. *IEEE Trans. Antennas Propagation,* Vol. 39, No. 7, (July 1991) pp. 1024–1030, ISSN: 0018-926X

i-Space (2003). *Engineering Test Satellite (ETS-VIII)*. http://i-space.jaxa.jp/experiments20-02/01_ETS8ver6.pdf

Ilcev, S.D. (2005). Global Mobile Satellite Communications: For Maritime, Land and Aeronautical Applications. Springer, ISBN: 1-4020-7767-X, Dordrecht, the Netherlands

International Telecommunication Union Radiocommunication Sector (ITU-R) Radio Regulations (2004). ICT Regulation Toolkit, Module 5 Radio Spectrum Management

Ito, K.; Ohmaru, K. & Konishi, Y. (1988). Planar antennas for satellite reception. *IEEE Trans. Broadcasting,* Vol. 34, No. 4, (Dec. 1988) pp. 457–464, ISSN: 0018-9316

Japan Aerospace Exploration Agency–JAXA (2003). *Engineering Test Satellite (ETS-VIII)*. http://www.jaxa.jp/projects/sat/ets8/index_e.html

Jongejans, A.W.; Rinous, P.J. & Roederer, A.G. (1993). Review of ESA mobile antenna developments for satellite communications, *Proceedings of IEEE Antennas and Propagation Society Intl. Symposium*, pp. 1838–1841, ISBN: 978-0780370708, Ann Arbor, MI, USA, June–July 1993, IEEE, Piscataway, NJ

Kaneko, K.; Tanaka, T.; Takahashi, M. & Ito, K. (2006). Electronic scanning characteristics of simple satellite-tracking array antenna. *IEICE Trans. Communications (Japanese Edition),* Vol. J89-B, No. 9, (Sept. 2006) pp. 1696–1704, ISSN: 0915-1877

Kilgus, C.C. (1975). Shaped-conical radiation pattern performance of the backfire quadrifilar helix. *IEEE Trans. Antennas Propagation,* Vol. 23, No. 3, (May 1975) pp. 392–397, ISSN: 0018-926X

Konishi, Y. (2003). Phased Array Antennas. *IEICE Trans. Communications,* Vol. E86-B, No. 3, (Mar. 2003) pp. 954–967, ISSN: 0916-8516

Kuramoto, T.; Yamane, T. & Endo, N. (1988). Mechanically steered tracking antenna for land mobile satellite communications, *Proceedings of IEEE Antennas and Propagation Society Intl. Symposium*, pp. 1314–1317, ISBN: 978-0780370708, Syracuse, NY, USA, June 1988, IEEE, Piscataway, NJ

Lodge, J.H. (1991). Mobile satellite communication systems: toward global personal communications. *IEEE Communications Magazine,* Vol. 29, No. 11, (Nov. 1991) pp. 24–30, ISSN: 0163-6804

Nakano, H.; Vichien, K.; Sugiura, T. & Yamauchi, J. (1990). A singly-fed patch antenna radiating a circularly polarised conical beam. *Electronic Letters,* Vol. 26, No. 10, (May 1990) pp. 638-640, ISSN: 0013-5194

Nakano, H.; Saura, Y.; Mimaki, H. & Yamauchi, J. (1991). Circularly polarized conical beam formation by backfire helical antennas. *IEICE Trans. Communications,* Vol. E74-B, No. 10, (Oct. 1991) pp. 3246–3252, ISSN: 0916-8516

National Institute of Information and Communications Technology–NICT (2007). *Engineering Test Satellite (ETS-VIII) Project*. http://www2.nict.go.jp/p/p463/ETS8/ETS8.html

Nishikawa, K.; Sato, K. & Fujimoto, M. (1989). Phased array antenna for land mobile satellite communications. *IEICE Trans. Communications (Japanese Edition),* Vol. J72-B, No. 7, (July 1989) pp. 323–329, ISSN: 0915-1877

Ohmine, H.; Sunahara, Y. & Matsunaga, M. (1996). A TM21 mode annular ring microstrip antenna for personal satellite communication use. *IEICE Trans. Communications,* Vol. E79-B, No. 9, (Sept. 1996) pp. 1227–1233, ISSN: 0916-8516

Ohmori, S.; Wakana, H. & Kawase, S. (1998). *Mobile Satellite Communications,* Artech House, ISBN: 0-89006-843-7, Norwood, MA, USA

Ohmori, S.; Tanaka, K.; Yamamoto, S. & Tsuchiya, M. (1990). A phased array tracking antenna for vehicles, *Technical Report of the IEICE (Japanese Edition)*, AP90-75, pp. 33–40, ISSN: 09135685, Oct. 1990, IEICE, Tokyo

Ohmori, S. (1999). Phased array antennas for mobile communications. *Springer Journal Annals of Telecommunications,* Vol. 54, No. 1–2, (Jan. 1999) pp. 92–102, ISSN: 1958-9395

Rabinovich, V.; Alexandrov, N. & Alkhateeb, B. (2010). *Automotive Antenna Design and Applications,* CRC-Press, ISBN: 978-1-4398-0407-0, Boca Raton, FL, USA

Sri Sumantyo, J.T.; Ito, K. & Takahashi, M. (2005). Dual band circularly polarized equilateral triangular patch array antenna for mobile satellite communications. *IEEE Trans. Antennas Propagation,* Vol. 39, No. 7, (July 1991) pp. 1024–1030, ISSN: 0018-926X

Strickland, P.C. (1995). Planar arrays for MSAT and INMARSAT land mobile satellite communications, *Proceedings of IEEE Antennas and Propagation Society Intl. Symposium*, pp. 1838–1841, ISBN: 978-0780370708, Newport Beach, CA, USA, June 1995, IEEE, Piscataway, NJ

Terada, N. & Kagoshima, K. (1991). Compatible mobile antenna for mobile satellite and cellular communication systems, *Technical Report of the IEICE (Japanese Edition)*, AP 91-65, pp. 29–34, ISSN: 09135685, Sept. 1991, IEICE, Tokyo

Teshirogi, T.; Chujo, W.; Tanaka, M.; Itoh, T. & Komuro, H. (1983). Circular polarization array antenna with sequential rotation and phase shifts, *Technical Report of the IEICE (Japanese Edition)*, AP83-57, pp. 49–54, ISSN: 09135685, Aug. 1983, IEICE, Tokyo

Xiong, F. (1994). Modem technique in satellite communications. *IEEE Communications Magazine,* Vol. 32, No. 8, (Aug. 1994) pp. 84–98, ISSN: 0163-6804

Yamaguchi, R. & Ebine, Y. (1997). Design of vehicle antenna using helical array for mobile satellite communication system, *Technical Report of the IEICE (Japanese Edition)*, AP 97-64, pp. 97–64, ISSN: 09135685, July 1997, IEICE, Tokyo

3

Cooperative Strategies for Satellite Access

Luca Simone Ronga and Rosalba Suffritti
CNIT
Italy

Enrico Del Re
University of Florence
Italy

1. Introduction

Satellite communications have become an important node of the global telecommunication infrastructure. Satellite capacity request is growing quickly, driven not only by broadcast applications but, mainly, by broadband services, in particular by the expectation of "always-on" broadband services available everywhere. Thus, new "killer" applications such as HDTV (High Definition Digital Television) and broadband Internet access, provided through satellites, can help to face the growth of capacity demand foreseen in the near future. Moreover, in addition to the provision of satellite multimedia services to fixed terminals, there is an increasing demand for broadband communications on the move (i.e. on ships, trains, aircrafts, vans, cars).
Analysing such an increasing demand of satellite communications, the work reported in this chapter is focused on the study of different techniques which allow the improvement of the performance of satellite users displaced in severe environments. The analysis of this context, in fact, has revealed the need to adopt adequate advanced techniques to achieve a sufficient quality in satellite links, especially in those scenarios where the link budget is tighter, such as, for example, the mobile satellite one. These considerations have motivated the study of *cooperative strategies* which allow the mitigation of the deleterious effects of fading. This is obtained thanks to a new form of spatial diversity in which the diversity gain can be achieved through the cooperation of different users which generate a virtual MIMO (Multiple-Input Multiple-Output) system. The adoption of these methodologies can be very helpful in those scenarios characterised by continuous occurrences of NLOS (Non-line-of-sight) and LOS (Line-of-sight) channel conditions and, therefore, it is interesting to assess their implementation in critical satellite contexts. Considering such a context, the chapter will investigate the adoption of different cooperative techniques in some satellite access scenarios, pointing out its advantages and drawbacks. The chapter is organised as follows. Starting from the identification of critical issues in different satellite access scenarios, reported in Section 2, a general overview on cooperative strategies and, in particular, on the selected cooperative approaches, is provided in Section 3. Then, Section 4 and Section 5 report some different satellite case studies in order to show the advantages of using this kind of approach in uplink and in downlink satellite access, respectively. Finally, Section 6 provides some concluding remarks.

2. Satellite Access: scenarios and critical issues

Satellite communications have developed a global success in the field of digital audio/TV broadcasting because they offer a wide coverage area and, therefore, they are suitable for the distribution of multimedia contents to a large number of potential users, also in rural environments. Moreover, they allow the extension of the coverage area of terrestrial, fixed and mobile, networks. One of the most interesting example concerning this capability, is provided by Inmarsat which has developed a broadband global area network service for mobile terminals on land, at sea and in the air. Users can send and receive voice and data services nearly everywhere on Earth. In particular, in some specific cases as the transoceanic maritime and aeronautical communications, satellites are the only practical solution to telecommunications requirements.
Broadband satellite systems can also help to bridge the digital divide because they can provide a rapid deployment compared with other terrestrial infrastructures, without gigantic investments. For example, continents (e.g. Africa) and large countries which, currently, lack in infrastructures could satisfy their needs (mobile phones, Internet access, etc.) and create new opportunities for human development. Applications like telemedicine, e-learning or simply an easy access to information can allow economic activities to grow and develop.
Satellite systems can allow a multitude of valuable services and applications to emerge. Besides for commercial services such as broadcasting, multimedia transmission and broadband services, the use of satellite for telecommunication is also considered for other application scenarios such as public services, emergency services, data relay services, etc. For example, the monitoring and the protection of critical infrastructures such as pipelines and oil platforms, depend on data transmission via satellite. And also coastal and maritime security has increased thanks to the use of new satellite technologies suitable for tracking the position and the state of goods transported by sea. In fact, vessels are required to carry satellite terminals that transmit their identity and position. The benefits of satellite communications are well visible also in emergency applications wherein the world-wide Civil Protection is involved in order to guarantee safety to population. In case of floods, earthquakes, volcanic eruptions and other major disasters, terrestrial communication networks could be damaged and not be able anymore to provide the services required by first responder teams, such as, for example, a robust voice communication system. Rescue teams terminals should be also compatible with other different kinds of terminals if the disaster involves more than one country and so multinational rescue operations are needed. In such a situation, satellites can flexibly connect different first responder team clusters over large distance across incompatible standards. In fact, for large disasters, only satellites are actually able to cover the whole scene and provide broadband services. A satellite communication component is considered in the Air Traffic Management scenario, as well. Also in this application, the main satellite communication strengths are the large coverage area and the rapid deployment. Thanks to the use of satellites, a seamless service between air traffic controllers and pilots could be provided in Europe, including not only areas of dense traffic but also remote areas such as Mediterranean sea, transatlantic routes, deserts, etc.
However, analysing all these scenarios, some critical issues in the use of satellite systems, common to many contexts, can be highlighted. In particular, the presence of link impairments and fading conditions (multipath, long periods of shadowing and blockage) or the mobility effects (occurrence of visibility and not visibility conditions) require the adoption of solutions in order not to reduce system performance and capabilities. Moreover, power constraints have to be taken into account, as well, especially in case mobile terminals are considered.

3. Overview on Cooperative Communications

Some years ago, a new class of techniques, called *cooperative communications*, has been proposed as a valuable alternative to the spatial diversity techniques which require the deployment of additional antennas in order to mitigate the fading effects.
Cooperative communications are based on the concept that a group of mobile terminals can share their single antennas in order to generate a "virtual" multiple antenna, obtaining the same effects than a MIMO system, (Nosratinia et al., 2004; Ribeiro & Giannakis, 2006). This approach can be seen as a new form of spatial diversity in which, however, the diversity gain can be achieved through the cooperation of different users, opportunely grouped in clusters, which can assume the double role of *active user*, i.e. the user which transmits its own information data and *cooperator*, i.e. the user which "helps" the active user in its transmission, (Sendonaris et al., 2003a;b).
The key concept is that each user sees an independent fading process and that spatial diversity can be generated by transmitting each user's data through different paths, as shown in Fig. 1.

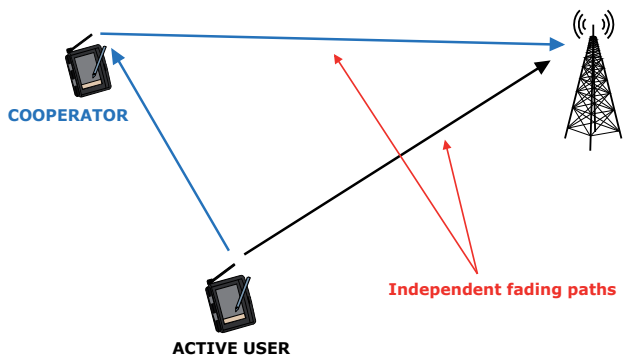


Fig. 1. Example of cooperative communications

An effective way to mitigate fading is to supply the receiver with multiple replicas of the same information-bearing signal transmitted over independent channels. Because of this independence, the probability that all the considered signals are simultaneously vanishing due to fading, is considerably reduced.
If p, $(0 \leq p \leq 1)$, is the probability that any signal is faded below a threshold value, the probability that all L independent fading channels, containing the same signal, are faded below the threshold value, is given by:

$$p_{tot} = \prod_{i=1}^{L} p = p^L \tag{1}$$

and, therefore, it is lower than p, (Lee & Chugg, 2006).
The cooperative approach turns to be useful for mobile terminals which, because of their size constraints, cannot support multiple antennas and it allows them to increase their performance in terms of Bit Error Rate, Packet Error Rate and Outage probability.
The scenarios wherein the idea of cooperation has been applied so far are, mainly, the cellular networks, the wireless sensor networks and the ad hoc networks, but it can be very interesting to consider the adoption of such strategies also in mobile satellite scenarios which are characterised by the continuous occurrence of LOS and NLOS conditions.

There are several cooperative methods which have been proposed in literature (Nosratinia et al., 2004; Ribeiro & Giannakis, 2006; Sendonaris et al., 2003a;b). However, the main cooperative strategies can be summarised in:

- *Amplify and Forward* (*AF*)
- *Decode and Forward* (*DF*)
- *Selective Forwarding* (*SF*)
- *Coded-Cooperation*

3.1 Amplify and Forward

The *Amplify and Forward* is the simplest cooperative method. In this scheme cooperators receive a noisy version of the signal transmitted by active users which, then, amplify and retransmit towards the final destination. Thus, in this case, also the noise component is amplified and retransmitted by cooperators.
Considering the case of one active user and one cooperator, the amplification factor A can be written as follows, (Darmawan et al., 2007; Ribeiro & Giannakis, 2006):

$$A^2 = \frac{P_c}{P_u |h(u,c)|^2 + N} \tag{2}$$

being P_c the power of the signal transmitted by the cooperator, P_u the power of the signal transmitted by the active user, $|h(u,c)|^2$ is the coefficient of the channel between active user and cooperator, and N is the noise power.
The *Amplify and Forward* strategy requires minimal processing at cooperator terminals but needs a consistent storage capability of the received signal consuming, therefore, memory resources. This method is particularly efficient when the cooperator is close to final destination, as shown in Fig. 2, so that the link from the cooperator to the destination, d_2, is characterized by high signal-to-noise ratios and, hence, the link between the active user and the cooperator, d_1, becomes comparable to the link between the active user and the destination, d_3.

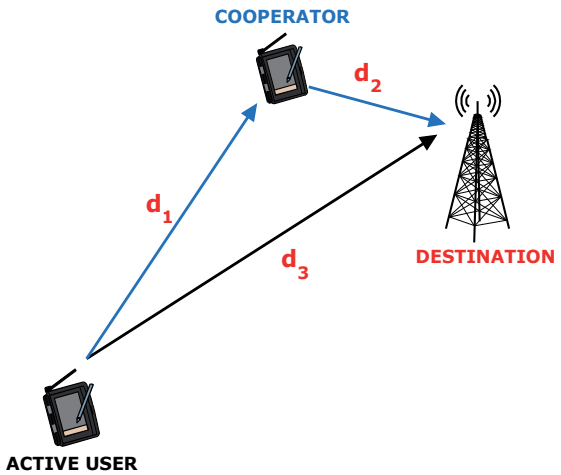


Fig. 2. *Amplify and Forward*: efficient terminals displacement

3.2 Decode and Forward

In the traditional *Decode and Forward* scheme, instead, each cooperator always decodes signal coming from the active users, $u(i)$ (with $i = 1 \dots N_u$, where N_u is total of active users), obtaining an estimate of transmitted signal, $\hat{u}(i)$. Then, it retransmits the signal, $c(i)$:

$$c(i) = \hat{u}(i) \qquad i = 1 \dots N_u \tag{3}$$

after a re-encoding generally with a repetition-coded scheme.

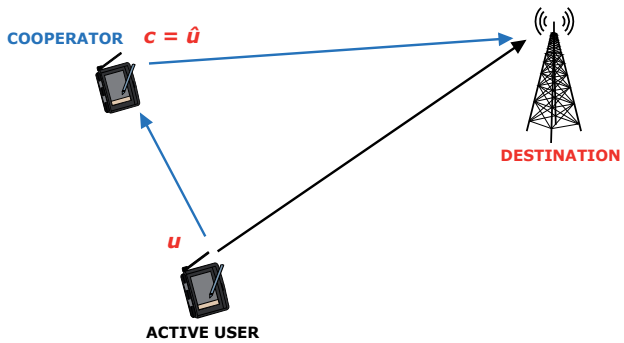


Fig. 3. *Decode and Forward* scheme

Although it has the advantage to be a simple scheme, this cooperative method does not achieve diversity gain. In fact, considering the case of one active user and one cooperator, it is proven that the diversity order is only one, because the overall error probability over two links is dominated by the error probability in the link between the active user and the cooperator, (Laneman et al., 2004; Ribeiro & Giannakis, 2006).

3.3 Selective Forwarding Cooperation

The *Selective Forwarding* strategy derives from the Decode and Forward technique and it is based on the concept that cooperators repeat active users' packets by transmitting them through different channel paths with the condition that only the successfully decoded packets received from active users, are sent toward the final destination.
This strategy is more complex than the Decode and Forward method, (Nosratinia et al., 2004; Ribeiro & Giannakis, 2006), because it requires FEC (Forward Error Correction) decoding followed by a CRC (Cyclic Redundancy Check) check to detect possible errors in the packets sent from the active users to the cooperators, but it has some important advantages.
First of all, *Selective Forwarding* is the simplest cooperative method from the perspective of the destination even though it overworks the digital processor at cooperating terminals. Moreover, differently from the Decode and Forward, it allows to achieve diversity and, therefore, to increase the diversity order. Assuming that wireless links between active users and cooperators (d_1), are much better than links between active users and their final destinations, (d_3), as shown in Fig. 4, and that all users in the considered cluster see uncorrelated channels, the diversity order can be considered equal to the number of users involved in a transmission (active user and its cooperators), (Alamouti, 1998). In this case, *Selective Forwarding* turns to be the best choice for implementing a cooperation process.
Since, for example, in a return link satellite scenario the previous assumptions can be considered valid, the *Selective Forwarding* scheme can be selected as a right cooperative strategy to be implemented in such kind of environments.

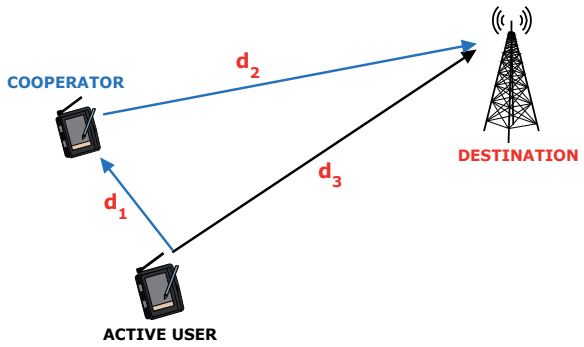


Fig. 4. *Selective Forwarding*: best implementation scenario

3.4 Coded-Cooperation

In the *Coded-Cooperation*, the cooperative strategy is integrated with channel coding techniques. In this case, instead of producing more replicas of the active user's signal, as it happens in other cooperative methods, the codewords produced by each user belonging to a determined cluster, are divided in different portions which are transmitted through different independent fading channels, by the considered user and by a selected group of users, called *partners*, which are involved in the cooperation process, (Hunter & Nosratinia, 2002; 2006; Janani et al., 2004).

The basic idea is that each user tries to transmit an incremental redundancy of its partners data, besides its own data. Considering, for example, the case of two users, they cooperate by dividing their own codewords of length N, in two successive segments, as shown in Fig. 5.

In the first segment, each user transmits a codeword of length N_1 containing its own data,

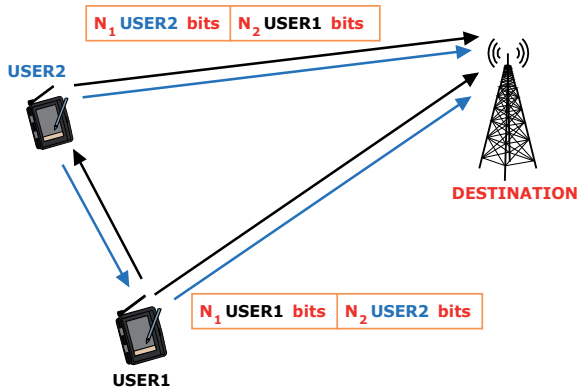


Fig. 5. *Coded-Cooperation* scheme

obtained by its original codeword. Then, each user receives and decodes its partner's first segment. If this is correctly decoded, each user can compute the additional parity bits of the partner's data and transmit the new codeword of length N_2 containing the partner's data, in the second segment. If the partner's info cannot be correctly decoded, the user reverts to the non-cooperative mode and it transmits its own data. In fact, if a certain terminal is unable to cooperate, because of the wrong reception of the partner's data, it can always use the available capacity to transmit its own data.

The idea of *Coded-Cooperation* is to use the same overall code rate and power for transmission as in a comparable non-cooperative system, i.e. the same system resources are used. Moreover, this cooperation methodology can provide a higher degree of flexibility with respect to other cooperation methods and a higher adaptability to channel conditions, by allowing the use of different channel coding and partitions schemes. For example, the overall code can be a block code or a convolutional code or a combination of both and, then, coded bits to put into the different segments, can be selected through puncturing, product codes, etc., (Hunter & Nosratinia, 2006).

4. Cooperation Techniques for Uplink Satellite Access

Considering what said above, the *Selective Forwarding* and the *Coded-Cooperation* turn to be two cooperative strategies which are suitable to be used in critical satellite scenarios, in particular in the return link suffering from a tighter link budget especially if the involved users are mobile terminals. Therefore, in the following, a specific uplink satellite scenario which presents some tricky issues, is proposed as "case study", in order to show the advantages deriving from the adoption of such cooperative strategies.
The considered model is composed of a set of N_u vehicular users which are interconnected through reliable wireless links and connected to a terrestrial gateway through a geostationary satellite, as shown in Fig. 6.

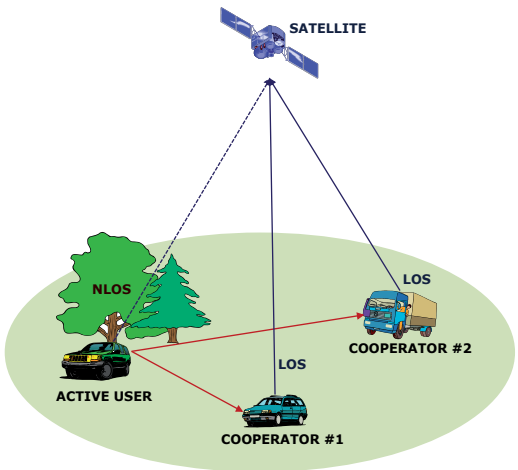


Fig. 6. Satellite cooperative scenario

The forward link is based on the DVB-S2 (Digital Video Broadcasting - Satellite second generation) standard, (DVB-S2 standard, 2009), while the return link (on which this analysis is focused) is based on DVB-RCS (Digital Video Broadcasting - Return Channel Satellite), (DVB-RCS standard, 2005). According to the MF-TDMA (Multi Frequency - Time Division Multiple Access) scheme employed by such a standard, a certain number of frequency/time slots are assigned to users within a superframe depending on their specific demand. The adopted propagation satellite channel model is mainly taken from (Ernst et al., 2008), and it is summarised here for the sake of completeness. The model considers a frequency non-selective

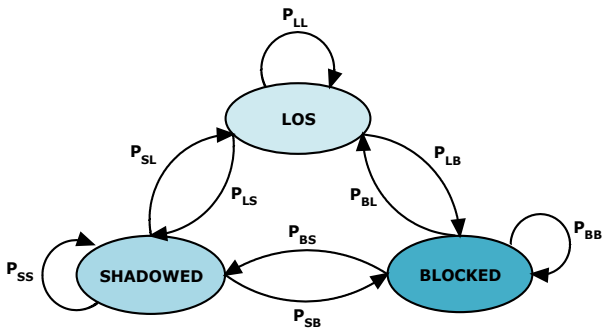


Fig. 7. 3-states channel model

channel at *Ku* band. In these conditions, a generic passband received signal, $r(t)$, can be written as:

$$r(t) = \mathrm{Re}\{A(t) \cdot \widetilde{s}(t - t_0)e^{j2\pi f_0 t}\} + n(t) \tag{4}$$

where $A(t)$ is the multiplicative time-varying channel coefficient, $\widetilde{s}(t)$ the complex-envelope of the transmitted signal, t_0 the propagation delay, f_0 the carrier frequency and $n(t)$ the additive thermal noise.
The channel coefficient is a complex term and, therefore, it can be expressed through its absolute value (also called modulus), $|A(t)|$, and its phase $\phi(t)$:

$$A(t) = |A(t)|e^{\phi(t)} \tag{5}$$

The amplitude of the channel coefficient, $|A(t)|$, represents the amplitude of the fading term which, according to this class of models, can be divided into fast and slow fading. Slow fading events, commonly referred to as shadowing, model the attenuation caused by the orography and large obstacles, such as hills, buildings, trees, etc., through absorption and diffraction mechanisms, and they are normally modelled as a finite state machine. Fast fading events, instead, due to the irregularity of the obstacles (e.g. vegetative shadowing) and to the multipath propagation phenomena caused by reflections over surrounding surfaces, can be additionally modelled as superimposed random variations that follow a given Probability Density Function (PDF) for each state.
At an arbitrary time instant t and assuming that the transmitted signal $\widetilde{s}(t)$ has unitary amplitude[1], the overall PDF describing the received signal amplitude, called below $R(t)$, can be written as:

$$p_R(r) = \sum_{k=1}^{N} P_k \cdot p_{R,k}(r) \tag{6}$$

being N the number of states, P_k the absolute probability of being in the state k (that can be easily obtained from the *State Transition Matrix* $S = [p_{ij}]$, containing in each element the probability of transition from the state i to the state j) and $p_{R,k}(r)$ the PDF associated to the fast fading within state k.
Following this approach, a three states (LOS, Shadowed and Blocked) Markov-chain based model is assumed for the fading process, as shown in Fig. 7.

[1] Under this hypothesis, the received signal amplitude, $R(t)$ corresponds to the amplitude of the fading term, i.e. $R(t) = |A(t)|$.

The *LOS* state is characterised by a Rician PDF of the following form:

$$p_R(r) = \frac{r}{\sigma^2} \cdot \exp\left(-\frac{r^2+z^2}{2\sigma^2}\right) \cdot I_0\left(\frac{r \cdot z}{\sigma^2}\right), \qquad r \geq 0 \tag{7}$$

being I_0 the zero-order modified Bessel function of the first kind, z the amplitude of the line-of-sight component and σ^2 the power of the real part or the imaginary part of the scattered component.
The *Shadowed* state is characterised by a Suzuki PDF, (Suzuki, 1977). The Suzuki process is a product process of a Rayleigh process and a Lognormal (LN) process, (Finn & Flemming, 1977; Pätzold, 2002). The slow signal fading is, in this case, modelled by the Lognormal process taking the slow time variation of the average local received power into account. The Rayleigh process models, instead, the fast fading. The Suzuki PDF can be expressed as follows, (Lin et al., 2005):

$$\begin{aligned} p_R(r) \quad = \quad & \int_0^{+\infty} \left[\frac{r}{\sigma_{ray}^2 L^2} \cdot \exp\left(-\frac{r^2}{2\sigma_{ray}^2 L^2}\right)\right] \cdot \\ & \left[\frac{1}{\sqrt{2\pi}\phi\sigma_{ln}L} \cdot \exp\left\{-\frac{1}{2}\left(\frac{\ln(L)-\phi\mu_{ln}}{\phi\sigma_{ln}}\right)^2\right\}\right] dL \end{aligned} \tag{8}$$

wherein the first term represents the conditional joint Lognormal and Rayleigh PDF while the second term is the Lognormal PDF which characterises the random variable L. Moreover, $\phi = \ln 10/20$ while μ_{ln} and σ_{ln} are the mean and standard deviation, respectively, of the associated Gaussian distribution in dB unit.
Finally, the *Blocked* state is characterised by no signal availability. The set of considered parameters is provided in Table 1 for the environment considered next, namely *highway*. The average state transition period is equal to 0.0417 s, corresponding to blocks of 1000 samples at the sampling frequency of 24 kHz. The above mentioned state duration refers to average speed v of 100 Km/h.

Environment	*State Transition Matrix*	*P (LOS, SH, BL)*	*Rice z*	*Rice σ*	*Rice Factor*	σ_{ln}	μ_{ln}
Highway	0.9862 0.0138 0.0000	0.8922	0.9892	0.0947	17 dB	1.5 dB	-8 dB
	0.1499 0.8378 0.0123	0.0823					
	0.0008 0.0396 0.9596	0.0255					

Table 1. *Ku*-band land-vehicular channel parameters

Doppler Spectrum is estimated as proposed in (Dubey & Wee Teck Ng, 2002; Law et al., 2001), taking into account a realistic antenna beamwidth and the angle between satellite position and terminal direction by means of the following equation:

$$S(f) = \begin{cases} \dfrac{A}{f_d\sqrt{1-\left(\dfrac{f}{f_d}\right)^2}} & \text{if} \quad f_d\cos(\phi+\alpha) < f < f_d\cos(\phi-\alpha) \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

The following values have been considered:

- $\alpha = \pi/2$

- $f_d = v \cdot f_0 / c$
- $2\phi = \theta_{3\,dB} = 70\lambda / D$
- $D = 65$ cm

being D the antenna diameter, v the terminal speed defined above and $f_0 = c/\lambda$, the carrier frequency at *Ku* band equal to 14 GHz.

4.1 Selective Forwarding Cooperation for Critical Satellite Scenarios

The analysis considers the adoption, in the scenario described above, of a cooperative strategy which allows the users to share the uplink effort according to the *Selective Forwarding* cooperation scheme. Fig. 8 shows an example of the used procedure which describes how the resources are allocated and managed in the TDMA scheme. Groups of timeslots, named *frames*, are assigned to active users and cooperators in order that they can transmit their *traffic bursts* (in the following named simply "packets").

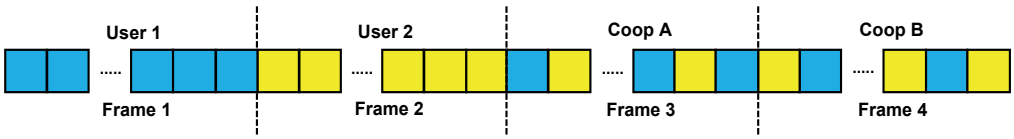
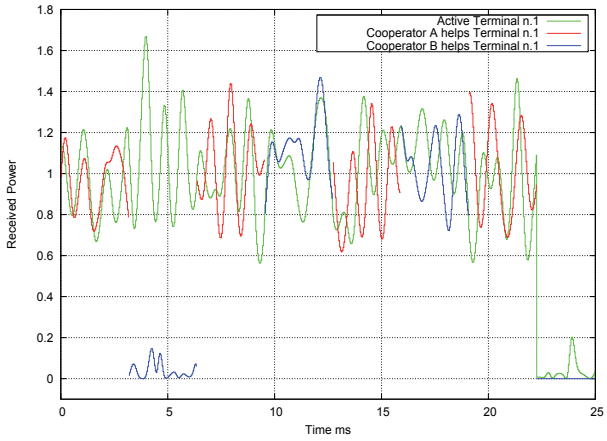
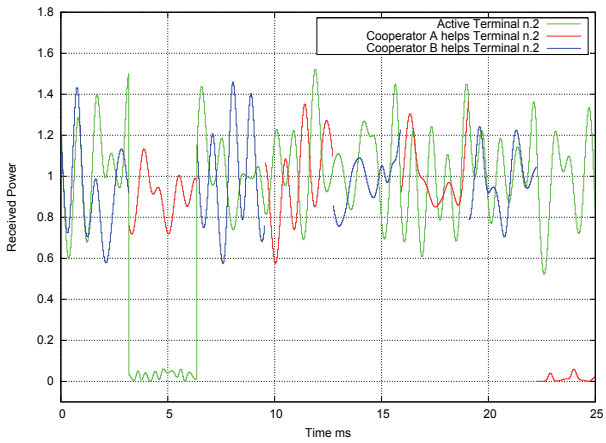


Fig. 8. Example of timeslot assignation in a superframe: 2 active users and 2 cooperators

Within each superframe, the active users (*User1* and *User2*) convey their informative packets while the cooperators (*Coop A* and *Coop B*) repeat each one half *User1*'s packets and half *User2*'s packets in an alternate way. In particular, *Coop A* retransmits before a *User1*'s packet and then a *User2*'s packet, whereas, vice versa, *Coop B* starts repeating before a *User2*'s packet and then a *User1*'s packet. Hence, in this case, two replicas of the same packet for each active user are sent through the satellite and the receiver can apply a CRC mechanism in order to detect the correct packets among those received. Such a method can be simply extended to a different number of active users and cooperators.
The benefits of this procedure can be assessed observing Fig. 9 wherein the received signal power of each active user and its cooperators, is reported. In some time portions, in fact, the cooperators can experiment better satellite channel conditions than the active users and their retransmission of packets becomes fundamental in order to not to lose some pieces of information sent by the active users. The receiver can process differently corrupted replicas of the same packet and the probability to detect packets successfully increases considerably.
In the model, the terrestrial wireless links between active users and cooperators, used to share packets, are characterized by error-free conditions in order to evaluate the efficiency of the cooperative strategy in the satellite land-vehicular scenario.
In the following, some results achieved through computer simulations are presented. First of all, it is shown how the number of involved cooperators affects the system performance. In particular, in Fig. 10, the performance comparison in terms of average PER (Packet Error Rate) between the no cooperation and cooperation (with 2 cooperators and 4 cooperators) cases in the highway environment is reported. The number of active users is considered equal to 2 in all simulated cases. Focusing mainly on this Figure, it can be seen that as the number of cooperators increases, the PER values decrease considerably for fixed E_b/N_0 values and, in particular, it can be noted that, the case considering 4 cooperators has a PER floor at about $2 \cdot 10^{-3}$ for E_b/N_0 values starting from 2 dB with respect to the no cooperation case which



(a) *Active user*: *User1*



(b) *Active user*: *User2*

Fig. 9. Received signal power of *Active user*, *Cooperator A* and *Cooperator B*

has, instead, a PER floor at $1.1 \cdot 10^{-1}$. The presence of PER floors is due to the occurrence, with the given probabilities already shown in Table 1, of Shadowed and Blocked state channel conditions. However, the context taken into account for satellite broadband communications is, mainly, that of elastic IP traffic generated by applications like e-mail, web browsing, FTP and TELNET services, which are not completely compromised by a delay, loss or bandwidth limitations, due also to the occurrence of NLOS channel conditions. For these reasons, it is worth analysing how the cooperation strategy affects the system performance when the satellite channel is only in LOS or in NLOS conditions in order to evaluate the realistic behaviour of the system which works for the most part of the time in LOS conditions. The LOS state is, as a matter of facts, the state with the highest absolute probability (89.22% in the considered highway environment).

Fig. 11 shows, therefore, a comparison in terms of PER between no cooperation and cooperation (4 cooperators) cases considering the satellite channel being only in the LOS state or

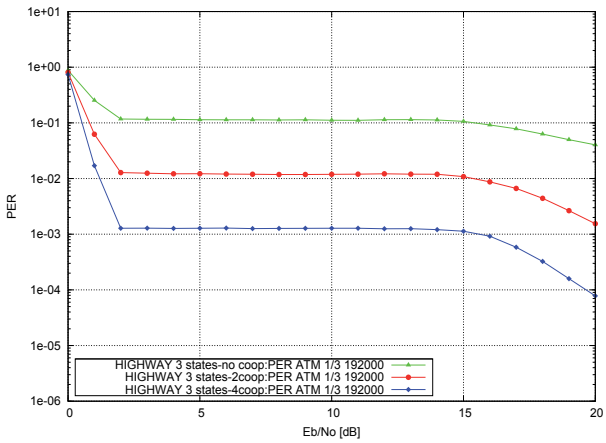


Fig. 10. PER performance for ATM cell, code rate 1/3, data rate 192 kbit/s, HIGHWAY environment: 3 states - *Ideal case* 4 cooperators, 2 cooperators and no cooperation cases

only in the Shadowed state. The Blocked state, as already said, is characterised by no signal availability so the achieved BER (Bit Error Rate) values are equal to 0.5.
The results concerning the LOS state are encouraging because they show that the adoption of the cooperation (4 cooperators) allows improving the system performance achieving the PER value 10^{-6} with a gain equal to 1.4 dB with respect to the case of absence of cooperation.

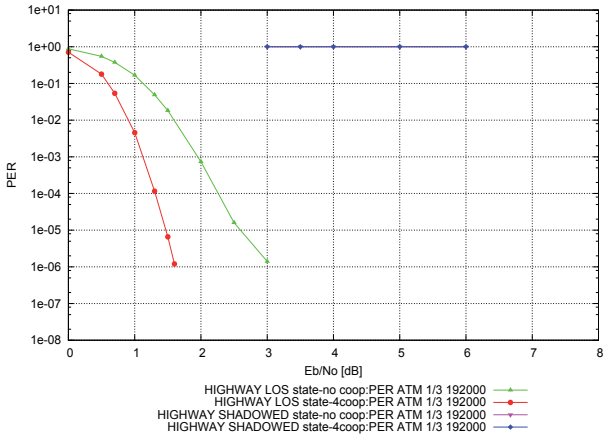


Fig. 11. PER performance for ATM cell, code rate 1/3, data rate 192 kbit/s, HIGHWAY environment: LOS state and Shadowed state - *Ideal case* 4 cooperators and no cooperation cases

4.2 Coded-Cooperation in Mobile Satellite Systems

In the following, the adoption of *Coded-Cooperation* in the same return link scenario previously described, is taken into account. In this case, the analysis starts considering the i-th user (with $i = 1 \dots N_u$) which aims at transmitting a message of size k bits. The message is first encoded by the physical layer encoder, obtaining the codeword $\mathbf{c}(\mathbf{i})$ of size n bits. Once all

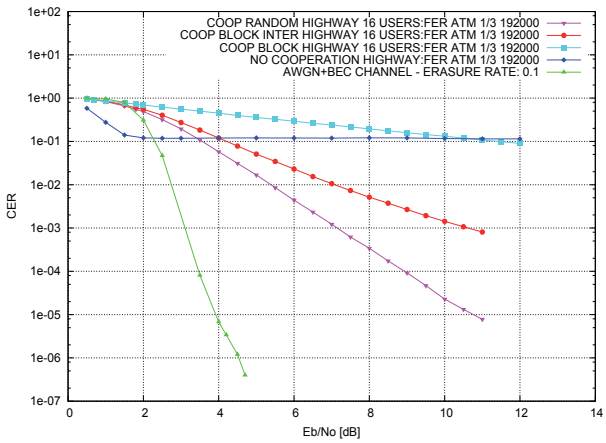


Fig. 12. Performance comparison in terms of CER between cooperative (16 users) and non-cooperative schemes for ATM cell, code rate 1/3, data rate 192 kbit/s: HIGHWAY environment

codewords $\mathbf{c}(\mathbf{i})$ are ready, they are exchanged through terrestrial links among the N_u users. At each user i, each generic message $\mathbf{c}(\mathbf{j})$ coming from the other users, is divided in N_u sub-blocks, $\mathbf{c}(\mathbf{j}) = [c_1(j), c_2(j), \ldots, c_{N_u}(j)]$. A new vector bit $\mathbf{x}(\mathbf{i})$, hereafter referred to as *combined codeword*[2], is then produced by the generic i-th user by combining N_u sub-blocks belonging to different users' codewords. The vector $\mathbf{x}(\mathbf{i})$ is, then, sent by the i-th user through the satellite link. The selection of the sub-blocks involved in the combined codewords can be based on predefined or random patterns depending on the considered *Coded-Cooperation* scheme, under the constraint that all the sub-blocks of a codeword $\mathbf{c}(\mathbf{i})$ are sent through different combined codewords.

Some results which prove the effectiveness of such a procedure are presented in the following. Performance has been analysed in terms of CER (Codeword Error Rate) vs. E_b/N_0 at the output of the FEC decoder in the gateway. In the plot in Fig. 12, a comparison among three different coded-cooperative schemes considering sixteen users, and the non-cooperative case is reported. In the first two schemes, named *cooperation block* and *cooperation block inter*, the codeword of the i-th user, constituted by a systematic part and a parity part, is divided in as many portions as the number of cooperative users and each of them transmits a combined codeword, as previously explained. The difference between these two schemes is in the rule that assigns each portion of the original codeword to each user. In the first scheme, a simple rule is used: the first user transmits the first portion of the systematic part and the first portion of the parity part of all codewords, the second one transmits the second portion of both parts and so on for all users. In the second scheme, instead, the portions sent by each user are assigned pseudo-randomly bearing however in mind that all sub-blocks of each codeword $\mathbf{c}(\mathbf{i})$ shall be transmitted. So, for instance, the first user transmits the first portion of systematic part but not the first one of the parity part. In the third scheme, named *cooperation random*, the partitioning of the codeword between systematic part and parity part is not considered

[2] Note that a combined codeword does not belong to a specific code book, i.e. it is not a result of an encoding procedure. It represents a concatenation of portions belonging to different actual codewords.

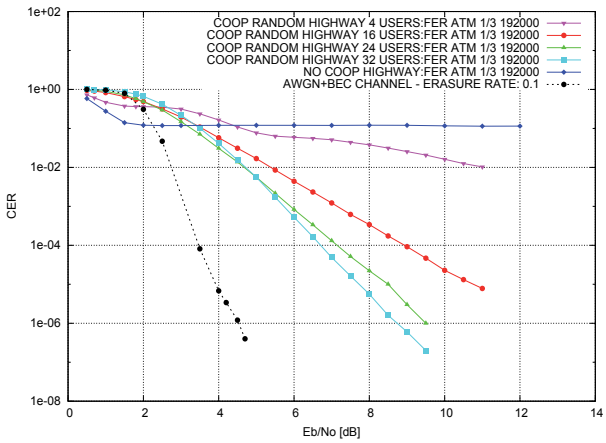


Fig. 13. Performance in terms of CER of the cooperation random scheme for different number of users, for ATM cell, code rate 1/3, data rate 192 kbit/s: HIGHWAY environment

anymore. In this case, the codeword portions composing the combined codeword are constituted by the bits of the original codeword of each user, which are assigned to each user using a random rule. Thus, the i-th user can transmit a portion composed by as many systematic bits as parity bits depending on the distribution of the bits that the random rule has generated.
Using this last scheme the highest randomization level is guaranteed and, as it can be seen in Fig. 12, the deleterious effects of fading can be more effectively counteracted. Also the performance over the AWGN (Additive White Gaussian Noise) channel with erasures, in the following named *AWGN+BEC*, is reported. This curve represents a reasonable reference which, for high E_b/N_0 values, could be taken as an acceptable lower bound to the system performance: under the assumption that only the LOS state can be successfully decoded, and in case the diversity introduced by cooperation could break any channel correlation effect, each codeword would in fact virtually face an uncorrelated channel with an erasure rate equal to the NLOS share, given by the sum of $P_{SHADOWED}$ and $P_{BLOCKED}$.
In Fig. 13, the *cooperation random* scheme is further investigated and it is shown how the number of users affects the system performance. It can be seen how, as the number of users increases, the CER values decrease for a fixed E_b/N_0 value. The performance improvement is more remarkable for increasing E_b/N_0 values. Using this scheme it is possible to achieve CER values performing a feasible system which does not present anymore a high floor value as it is, instead, for the non-cooperative case which has a CER floor at 10^{-1}. In particular, it can be noted that the CER value 10^{-5} is achieved for E_b/N_0 equal to 7.7 dB. This results is encouraging also because, if the channel state information were introduced in the simulation model, the achieved improving could be more relevant.

5. Cooperation Techniques for Downlink Satellite Access

Generally, in a downlink scenario, the link from the satellite to the active terminal is comparable with the links from the satellite to cooperating devices and, therefore, the *Amplify and Forward* strategy can be particularly efficient in this kind of scenarios. For this reason, a particular downlink satellite scenario is taken into account in order to show how the use of such a strategy can led to improvements in the system performance.

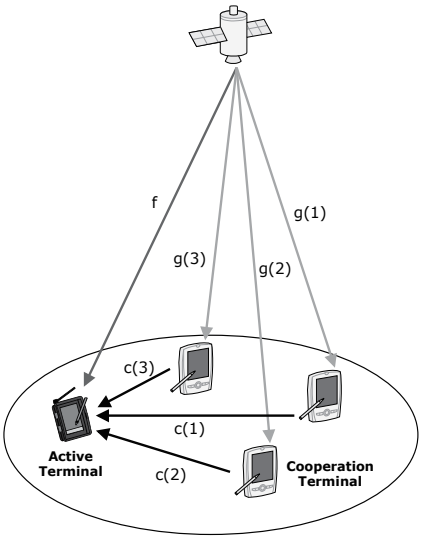


Fig. 14. Downlink Satellite Cooperation Scenario

d_{sat}	36000	[Km]	satellite terminal distance
d_{coop}	10	[Km]	cooperative terminal
L_{sat}	-205.34	[dB]	satellite terminal path loss
L_{coop}	-118.5	[dB]	cooperative terminal path loss
B_{sat}	36	[MHz]	transpoder bandwidth
P_{sat}	70	[dBW]	satellite power
P_{max}	250	[mW]	cooperative terminal maximum power
G/T_{Rx}	-24	[dB/K]	handheld receiver G/T
T_{sys}	290	[K]	system temperature
F_c	2000	[MHz]	cooperation channel frequency
F_d	11750	[MHz]	downlink channel frequency

Table 2. Main operational parameters

The adopted downlink cooperation scenario is depicted in Fig. 14. A DVB-S2 hub processes and sends digital signals to some users grouped in a cluster, through the satellite. A potential mobile DVB-S2 receiver (the active terminal) combines the signals coming from the satellite and from several mobile cooperators belonging to the same cluster. The satellite-to-earth link is modelled with a Corazza-Vatalaro process, (Corazza & Vatalaro, 1994), while the cooperator-to-active user link is represented only by an AWGN channel. The Corazza-Vatalaro channel model is a combination of a Rice and a Log-normal factors, with shadowing affecting both direct and diffused components. The cooperative path-loss value of 118 dB, reported in Table 2, derives from the choice of a cooperation frequency $F_c = 2$ GHz and a cooperator distance $d_{coop} = 10$ Km.
The fading effect on the cooperative links is not considered, as expected in environments characterized by limited distances (within 10 Km) and good visibility among terminals. The model considers a time resolution equal to:

$$\frac{1}{2B_{sgn}} = \frac{1}{14.8}\ \mu s \tag{10}$$

being B_{sgn} the bandwidth of the modulated QPSK signal ($FEC = 1/2$) considering an useful data rate of 7.2 Mbaud.

5.1 Amplify and Forward Cooperation for Mobile Satellite Terminals

The basic idea of *Amplify and Forward* strategy is that around a given terminal, there can be other single-antenna terminals which can be used to enhance diversity by forming a virtual (or distributed) multiantenna system where the satellite signal is received from the active terminal and a number of cooperating relays. Cooperating terminals retransmit the received signal after amplification. As said before, the AF strategy is particularly efficient when cooperating terminals are located close to the active one so that the cooperative links ($c(1)$,$c(2)$,$c(3)$ in Fig. 14) are characterized by high signal-to-noise ratios and the link from the satellite to the active terminal (f) is comparable with the links from the satellite to cooperating devices ($g(1)$,$g(2)$,$g(3)$ in Fig. 14). Starting from Eq. (2), the considered amplification factor A is given by:

$$A_i^2 = \frac{P_{max}}{P_{sat}|g(i)|^2 + N} \tag{11}$$

where P_{sat} is the satellite downlink power and P_{max} the cooperative terminal maximum power, $g(i)$ the i-th link pathloss and $N = KT_{sys}B_{sat}$ the noise power at the earth terminals.
With this choice, the resulting C/N on the active terminal is given by the following expression, assuming that all of the cooperating terminals, M, have the same characteristics and the cooperative channels, c, are similar:

$$\frac{C}{N} = \frac{P_{sat}|f|^2}{N}\left(1 + M\frac{A^2|c|^2}{1 + A^2|c|^2}\right) \tag{12}$$

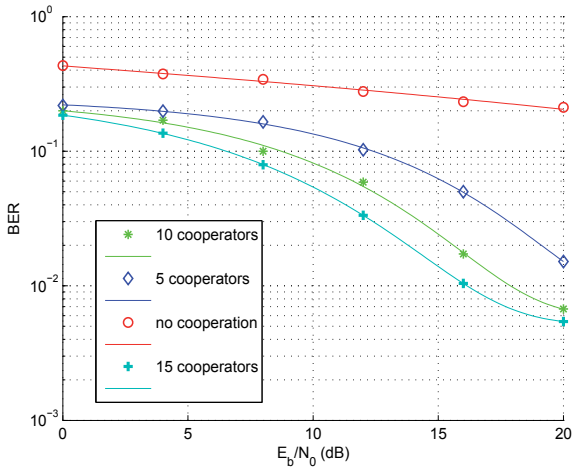


Fig. 15. BER performance: QPSK, $5 - 10 - 15$ cooperators, $R = 1$

System performance has been analysed in terms of BER and the resulting BER versus E_b/N_0 curves for different configurations have been plotted. The curves of Fig. 15 show the advantages deriving from the use of the cooperation AF with a QPSK modulation for various number of cooperators (5, 10 and 15). All the handsets share the same Rice factor $R = 1$ (medium

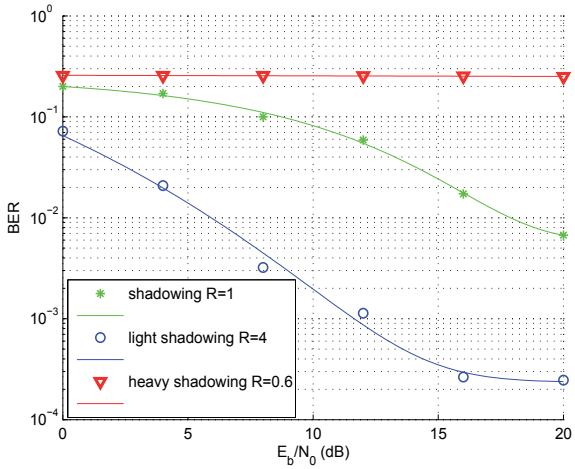


Fig. 16. BER performance: QPSK for variable Rice Factor $R = 0.6 - 1 - 4$ and 10 cooperators

shadowing), modeling the situation where the consumers cooperators all work under homogeneous operational conditions. Fig. 16 shows QPSK performances obtained by varying the Rice factor R. The case of heavy shadowing ($R = 0.6$), medium shadowing ($R = 1$) and light shadowing ($R = 4$) are compared. For $R = 4$ the performance is close to the target ($BER = 10^{-4}$), while for $R = 0.6$ the BER values are higher than target, resulting unacceptable for the DVB-S2 system.

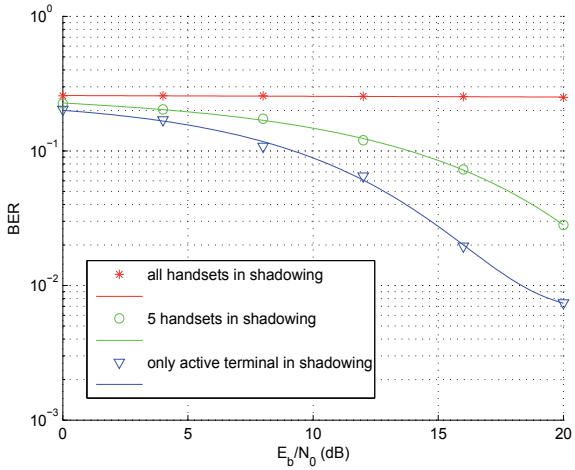


Fig. 17. BER performance: QPSK, varying handset number in heavy shadowing for $R = 0.6$

Finally, Fig. 17 shows the BER performance in the case a varying number of handsets are in heavy shadowing ($R = 0.6$) while the remaining ones have $R = 1$. By considering such a less critical situation, where only a subset of cooperating terminals are subject to heavy shadowing, it can be seen that the system performance improves.

6. Conclusion

This chapter has presented the possible adoption of cooperation strategies in satellite access, focusing on two case studies showing an uplink and downlink mobile satellite scenario. The use of these different techniques and methodologies in various applications scenarios, can led to the achievement of improvement of the system performance in terms of Bit Error Rate and Packet Error Rate.
In particular, in the uplink scenario, the introduction of the *Coded-Cooperation* for DVB-RCS terminals working in a land vehicular scenario, allows improving considerably, for increasing E_b/N_0 values, the system performance compared with the non-cooperative system, especially if a codeword partitioning scheme maximising the level of randomness in the distribution of the sub-blocks among different users is adopted. In the best simulated scenario, if it is considered a Codeword Error Rate value of 10^{-5}, the system performance is, however, still roughly 3.8 dB away from the reference (AWGN with erasure rate equal to NLOS share) case, leaving significant room for further optimisation of the system. However, a trade-off between the number of cooperative users, the resulting system complexity and the achievable performance is necessary. Moreover, also the adoption of a *Selective Forwarding* cooperation in a DVB-RCS land-vehicular scenario, allows improving sensibly the system performance in the considered environments, depending on the number of users involved in the cooperation process. The simulation results have shown that, considering 4 cooperators which cooperate with 2 active users, a cooperation gain equal to 1.4 dB can be achieved with respect to the case of absence of cooperation.
As what concerns, instead, the downlink scenario the idea was to build a cooperation among a set of mobile terminals, in a way that the signal received by each single device is the result of the composition of more replicas of the same signal sent by other cooperating devices. Link cooperation, in this case, enables the reception of satellite services from handheld terminals when a cluster of cooperating users is present.

7. References

Alamouti, S. M. (1998). A Simple Transmit Diversity Technique for Wireless Communications, *IEEE Journal on Selected areas in Communications*, Vol. 16, pp. 1451-1458, October 1998.

Corazza, G. & Vatalaro, F. (1994). A Statistical Model for Land Mobile Satellite Channels and Its Applications to Nongeostationary Orbit Systems, *IEEE Transactions on Vehicular Technology*, vol. 43, pp. 738-741, August 1994.

Darmawan, A. & Kim, S.W. & Morikawa, H. (2007). Amplify-and-Forward Scheme in Cooperative Spatial Multiplexing, *16th IST Mobile and Wireless Communications Summit*, July 2007, Budapest, Hungary.

Dubey, V.K. & Wee Teck Ng, (2002). Comments on On the Doppler spectrum at the mobile unit employing a directional antenna, *IEEE Communication Letters*, Vol. 6, No. 11, pp. 472-474, November 2002.

Ernst, H. & Harles, G. & Scalise, S. (2008). Measurement and Modelling of the Land Mobile Satellite Channel at Ku-Band, *IEEE Transactions on Vehicular Technology*, Vol. 57, No. 2, pp. 693-703, March 2008.

ETSI EN 301 790 v 1.4.1 (2005). Digital Video Broadcasting (DVB): Interaction channel for satellite distribution systems, September 2005.

ETSI EN 302 307 v 1.2.1 (2009). Digital Video Broadcasting (DVB): Second generation framing structure, channel coding and modulation system for Broadcasting, Interactive Ser-

vices, News Gathering and other broadband satellite applications (DVB-S2), August 2009.

Finn, M.I. & Flemming, H. (1977). Mobile Fading-Rayleigh and Lognormal Superimposed, *IEEE Transactions on Vehicular Technology*, Vol. 26, No. 4, pp. 332-335, November 1977.

Hunter, T.E. & Nosratinia, A. (2002). Cooperation Diversity through Coding, *IEEE International Symposium on Information Theory (ISIT)*, July 2002, Lausanne, Switzerland.

Hunter, T.E. & Nosratinia, A. (2006). Diversity through Coded Cooperation, *IEEE Transaction on Wireless Communications*, Vol. 5, No. 2, pp. 283-289, February 2006.

Janani, M. & Hedayat, A. & Hunter, T.E. & Nosratinia, A. (2004). Coded Cooperation in Wireless Communications: Space-Time Transmission and Iterative Decoding, *IEEE Transaction on Signal Processing*, Vol. 52, No. 2, pp. 362-371, February 2004.

Laneman, J.N. & Tse, D.N.C. & Wornell, G.W. (2004). Cooperative Diversity in Wireless Networks: Efficient Protocols and Outage Behavior, *IEEE Transaction on Information Theory*, Vol. 50, No. 12, pp. 3062-3080, December 2004.

Law, C.L. & Yoshida, S. & Xu, C.Q. (2001). On the Doppler power spectrum at the mobile unit employing a directional antenna, *IEEE Communication Letters*, Vol. 5, No. 1, pp. 13-15, January 2001.

Lee, D.K. & Chugg, K.M. (2006). A Pragmatic Approach to Cooperative Communication, *IEEE Military Communications Conference (MILCOM)*, October 2006, Washington, D.C.

Lin, D.B. & Lin, H.P. & Tseng, M.C. (2005). Performance analysis of M-ary PSK adaptive modulation system over Rayleigh-lognormal fading channel, *IEEE Vehicular Technology Conference Spring (VTC2005Spring)*, May 2005, Stockholm, Sweden.

Nosratinia, A. & Hunter, T.E. & Hedayat, A. (2004). Cooperative Communication in Wireless Networks, *IEEE Communications Magazine*, Vol. 42, No. 10, pp. 74-80, October 2004.

Pätzold, M. (2002). *Mobile Fading Channels*, Wiley, January 2002.

Ribeiro, A. & Giannakis, G.B. (2006). Fixed and Random Access Cooperative Networks, *EURASIP Newsletter*, pp. 3-24, March 2006.

Sendonaris, A. & Erkip, E. & Aazhang, B. (2003). User cooperation diversity - part I: System Description, *IEEE Transactions on Communications*, Vol. 51, No. 11, pp. 1927-1938, November 2003.

Sendonaris, A. & Erkip, E. & Aazhang, B. (2003). User cooperation diversity - part II: Implementation Aspects and Performance Analysis, *IEEE Transactions on Communications*, Vol. 51, No. 11, pp. 1939-1948, November 2003.

Suzuki, H. (1977). A Statistical Model for Urban Radio Propagation, *IEEE Transactions on Communications*, Vol. 25, No. 7, pp. 673-680, July 1977.

4

MIMO Channel Models for Satellite Communications

Abbas Mohammed and Asad Mehmood
Blekinge Institute of Technology
Sweden

1. Introduction

In recent years considerable attention has been drawn to the systems with multiple element transmitter and receiver arrays. It has been demonstrated both in theory and practice that multiple-input multiple-output (MIMO) systems offer the promise of increased capacity, high spectral efficiency and high gains by exploiting space-time processing techniques in particular space-time coding (STC) techniques under different propagation environments. This superior performance is achieved by the decomposition of the radio channel into a set of independent spatially uncorrelated channels. The basic method to achieve diverse channels, in addition to the exploitation of time and frequency dimensions, is to use the additional dimension of space, i.e., antennas are spatially separated from each other resulting in space-time processing. The advantages obtained by MIMO channels are highly dependent on the orientations of scatterers and the correlations among signal carriers which limit the performance of MIMO systems. The antenna separation, in terms of wavelength of the operating frequency, has significant impact on the spatial correlation. To achieve uncorrelated fading paths adequate antenna spacing along with rich scattering environment is necessary. An alternative solution to achieve low fading correlation is to use antenna arrays with cross polarizations, i.e., antenna arrays with polarizations in orthogonal or near orthogonal orientations without increasing the bandwidth and in particular, the concept of three-dimensional (3D) polarization has a significant role in achieving diversity by polarization. It has been shown in recent research that it is possible to attain even more channels by using benefits of the combined spatial and polarization diversity in rich scattering environments.

Satellite communication systems are not immune from this wave of innovation. However, due to difference in the propagation conditions in satellite and terrestrial links, the applicability and designs of MIMO systems are different as well. Due to very large path lengths, transmit and/or receive antennas must be placed at appropriate distances from each other to realize diverse paths. To achieve this, the possible diversity sources, i.e., satellite diversity and site diversity can be exploited in forming the MIMO channels for satellites. In the case of satellite diversity, the satellites are far apart from each other to achieve diversity and as a result the path lengths and the time of arrival of signals can vary

vastly between the satellites and the ground terminals resulting in synchronization problem. These issues can be dealt with by employing cooperative satellite diversity concept or the use of compact antennas in which the problem of synchronization does not exist. The frame work for the most recent developments in satellite communications includes satellite land mobile and fixed communication systems, satellite navigation systems, Earth Observation systems and the state of art propagation models and evaluation tools for these systems. The influence of radio channel is a critical issue for the design, performance assessment and real-time operation of these highly reconfigurable hybrid (satellite and terrestrial) radio networks providing voice, text and multimedia services operating at RF frequencies ranging from 100 MHz to 100 GHz and optical frequencies.

The organization of the chapter is as follows: in section 2, a brief introduction to directional channel modelling including MIMO channel modelling is presented. In section 3 we present an overview of MIMO channel models for satellite communications. Section 4 describes MIMO channel models for satellites based on polarization concept. Finally, conclusions and suggestions are given in section 5.

2. Directional Channel Modelling

The difficulties in modelling a radio channel are due to complex and varying propagation environments. In case of land mobile satellite (LMS) communications the signal travelling between a land mobile and a satellite suffers from different propagation impairments (chapter 1). The diverse nature of propagation media (e.g., ionosphere, troposphere and local effects) adds further a dimension of complexity in predicting the affects of propagation impairments on radio signals. It is important to note that the level of information obtained from a channel model about an environment is highly dependent on the type of system under assessment. In order to predict the performance of single sensor narrowband system it may be sufficient to consider the time varying amplitude distribution and the received signal power. Thus, classical channel models which provide information about signal power level distributions and Doppler shifts may be adequate for narrowband systems. Broadband communication systems build on the classical understanding of the received signal power distributions and Doppler spread also exploit spatial processing to operate in highly complex and diverse propagation environments. Thus, it is necessary to incorporate new concepts such as adaptive antenna arrays, angles of arrival (AoA), angles of departure (AoD), delay spread and multiple antennas at both the ends of a communication link, the so-called MIMO systems.

When investigating MIMO channels, the additional dimension that comes into play is space which needs to be modelled in a similar way as frequency and time variations have been modelled for the wideband single-input single-output (SISO) systems. In contrast to the systems which deal with only temporal spreading, the MIMO channels require the angular distribution of energy at both the ends of the communication link. The impulse response of the double directional channel between a transmitter positioned at P_t and a receiver positioned at P_r with n paths in 3D space can be written as (Steinbauer et al., 2001):

$$h(P_t, P_r, \tau, \Omega_t, \Omega_r) = \sum_{i=1}^{n} h_i(p_t, p_r, \tau, \Omega_t, \Omega_r) \tag{1}$$

where τ, Ω_t and Ω_r denote the delay, AoD and AoA, both in 3D space, respectively and h_i is the contribution of the i^{th} component, i.e.,

$$h_i(p_t, p_r, \tau, \Omega_t, \Omega_r) = \alpha_i e^{j\phi i} \delta(\tau - \tau_i)\delta(\Omega_t - \Omega_{t,i})\delta(\Omega_r - \Omega_{r,i}) \tag{2}$$

where α_i, φ_i, τ_i, $\Omega_{t,i}$ and $\Omega_{r,i}$ represent the amplitude, phase, time delay, AoD and AoA of the i^{th} multipath contribution, respectively.

These parameters are determined by the relative location of the transmitter and the receiver. When either the transmitter or the receiver moves, these variables become a function of time and can change drastically over large periods of time (long distances). Therefore a more compact representation of time variant double-directional channel impulse response is given by,

$$h(t, \tau, \Omega_t, \Omega_r) = \sum_{i=1}^{n} h_k(t, \tau, \Omega_t, \Omega_r) \tag{3}$$

In addition to these parameters, the double directional channel impulse response is also dependent on the antenna patterns and the modelling bandwidth.

2.1 The MIMO Propagation Channel

If multiple antennas are deployed at both the ends of a communication link, a MIMO system is obtained as shown in Fig. 1. The key idea underlying MIMO theory is that signals sampled in the spatial domain at both the ends are combined in such a way that multiple parallel channels are created. The double-direction description of the channel can be extended to MIMO channel by considering n_t transmit and n_r receive spatially separated antennas at both the ends. The corresponding MIMO channel matrix can be defined as:

$$H(t,\tau) = \begin{bmatrix} h_{11}(t,\tau) & h_{12}(t,\tau) & \cdots & h_{1m}(t,\tau) \\ h_{21}(t,\tau) & h_{22}(t,\tau) & \cdots & h_{2m}(t,\tau) \\ \vdots & \vdots & \ddots & \vdots \\ h_{n1}(t,\tau) & h_{n2}(t,\tau) & \cdots & h_{nm}(t,\tau) \end{bmatrix} \tag{4}$$

where $h_{nm}(.)$ denotes the narrowband channel (wide sense stationary uncorrelated scattering process) between the m^{th} transmit and n^{th} receive antenna. The elements of the channel matrix (i.e., individual narrowband SISO channels) are taken as independent circularly symmetric complex Gaussian variables with equal variance.

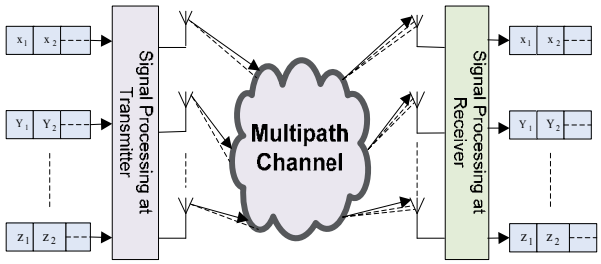


Fig. 1. The MIMO propagation channel.

The representation of a MIMO channel by individual SISO channels in not complete description of the multiple antenna systems. In order to get full benefit from MIMO systems (i.e., diversity gain, spatial multiplexing gain and array gain) certain trade-offs exist between these gains to achieve an adequate bit error rate at all times in an interference and noise limited system and at the same time maximizing throughput. The performance of MIMO techniques requires the exploitation of spatial correlation between all channel matrix elements. In (3) the elements of the channel matrix are assumed to be independent and therefore any two elements are uncorrelated. Practically, there is always some correlation between the channel matrix elements. These correlations are owing to small antenna array separation, antenna geometry and small amount of angular spread at the transmitter or the receiver side or both. Different studies have been found in literature to investigate the effects of correlation on the performance of systems employing multiple antenna techniques. In (Lokya, 2001) n equal power and equal rate parallel sub-channels are considered with 'r' as correlation coefficient between any two sub-channels. The capacity of such a channel can be written as:

$$C(r) = n.\log_2\left[1+\frac{\rho}{n}(1-r)\right] + \log_2\left[1+\frac{n\rho r}{n+\rho(1-r)}\right] \tag{5}$$

where 'ρ' represents the signal-to-noise ratio (SNR). When $r = 0$ the above equation reduces to the well-known formula for capacity:

$$C(r) = n.\log_2\left[1+\frac{\rho}{n}(1-r)\right] \quad \text{bps/Hz} \tag{6}$$

Comparison of (5) and (6) illustrate that the SNR decreases inversely with increase in the value of correlation coefficient (e.g., $r = 0.7$ results in 3 dB reduction in SNR). The MIMO channel capacity versus correlation coefficient for different values of 'n' at SNR value of 30 dB is shown in Fig. 2.

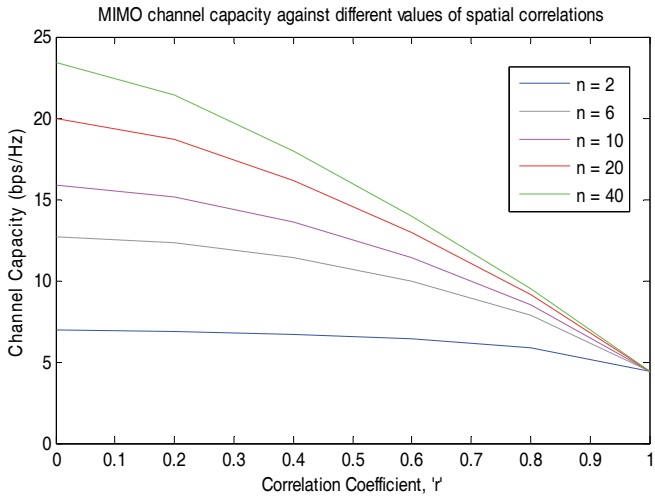


Fig. 2. MIMO channel capacity for 'n' channels against different values of correlation coefficient 'r'.

The other factors that limit the performance of MIMO channels are: number of scatterers, presence of line-of-sight (LOS) path, correlation between antenna elements at the transmitter or receiver side or both, keyholes, antenna patterns and geometry etc.

3. MIMO Channel Models for Satellite Communications

The wireless world has seen substantial increase in the demand of high quality broadband wireless services during recent years which results in the evolution of MIMO communication systems. The satellite communication systems are not immune to this change. In order to evaluate the performance of these systems employing MIMO technology, radio propagation models with high accuracy are required which can capture all those effects that affect different aspects of these systems whilst remaining computationally simple and require less simulation times.

Different approaches can be used to build a channel model with certain trade-offs (Almer et al., 2007). For example, physical or deterministic channel models based on ray-tracing algorithms can provide accurate results for a particular scenario, however due to global coverage of satellites this approach is not often used. On the other hand statistical models are built around measurement data and provide reasonable accuracy for the environments similar to that for which the model was built. However, they provide little insight into the propagation mechanisms and depend on the accuracy of measured data. An intermediate approach between these models is physical-statistical model. The physical-statistical modelling approach is the most appropriate in predicting the 'ON/OFF' nature and finding the small scale fading effects over large coverage areas applicable to LMS communication systems (Saunders et al., 2007). In this section some of the recently developed MIMO channel models based on different channel modelling approaches are presented.

3.1 The Physical-Statistical MIMO Channel Model

This physical-statistical MIMO channel (King et al., 2005) for LMS communications is based on 'clusters' concept and uses the same methodology given in (Correia, 2001; Molisch, 2004). The ray-tracing algorithm has been exercised to find different propagation effects like roof-top diffraction, specular reflection, shadowing caused by trees and foliage and blockage by buildings in LMS communication links in urban and high-way environments. The model can generate high resolution time series data and power delay profile for communication links between the satellite and mobile terminal antennas and can also predict the correlation between these links. In this model, obstacles (e.g., buildings, trees etc) are grouped into clusters of spherical shapes where clusters centers are randomly positioned. The building heights follow lognormal distribution. Twenty scatterers are placed randomly around the cluster centre each with dimension following Laplacian distribution (Correia, 2001). The building densities are assumed to be 90% in clusters representing urban environment, whereas trees are considered 90% of time in clusters for high-way environment. In order to validate the model, the parameters used are obtained from measurements data collected in Munich, Germany at L-band (1.54 GHz), for urban and high-way environments.

When signals reflected by distant clusters are blocked by buildings in the local clusters, these contributions are rejected. The signals from the satellite antennas to the mobile antennas are selected through appropriate clusters of scatters. Each scatterer in the cluster is assigned randomly the same reflection coefficient from a uniform magnitude distribution between 0 and 1 and phase 0 to 2π. This channel model considers three paths between a satellite and moving mobile terminal: a line-of-sight (LOS) path, a blocked LOS path and an attenuated path by trees. The time series data 'I' for a satellite or a high altitude platform (HAP) antenna M and each moving mobile antenna N can be written as:

$$\alpha_{M,N} = \begin{cases} P_{M,N} + b\sum_{i=1}^{n} T_i F_i P_{M,N,i} exp(jkd_{M,N,i}) & Clear\ \ Path \\ D_{M,N} P_{M,N} + b\sum_{i=1}^{n} T_i \Gamma_i P_{M,N,i} exp(jkd_{M,N,i}) & Blocked\ \ Path \\ T_{M,N} P_{M,N} + b\sum_{i=1}^{n} T_i \Gamma_i P_{M,N,i} exp(jkd_{M,N,i}) & LOS\ \ Path \end{cases} \quad (7)$$

where $P_{M,N}$, $D_{M,N}$ and $T_{M,N}$ denote the LOS path loss, the diffraction loss and the LOS trees loss, respectively, between satellite antenna M and moving mobile terminal antenna N. The term 'b' is the clutter factor derived from measurements in each environment, T_i is the tree attenuation from scatter i, Γ_i is the reflection coefficient at scatterer i, $P_{M,N,i}$ is the path loss and $D_{M,N,i}$ is the distance between satellite antenna M and mobile terminal antenna N via scatterer i. The small scale fading parameters such as AoA distribution, shadowing depth and wideband parameters like root mean square (RMS) delay spread or coherence bandwidth can be approximated in each environment using the output time series and spatial power delay profile data of the model.

3.2 Analytical MIMO LMS channel model at Ku-Band and above

The physical-statistical channel model described earlier is designed for L (1-2 GHz) and S (2-4 GHz) frequency bands for LMS communication systems. The application of MIMO systems for satellite communications at Ku frequency band (12-18 GHz) and above has been discussed in (Liolis et al., 2007). In this model, two features of MIMO technology are presented: (i) a 2x2 MIMO spatial multiplexing system is used to achieve capacity improvements and a closed form expression for the outage capacity is derived (ii) MIMO spatial diversity scheme with receive antenna selection is applied in order to reduce interference in LMS communication links. In addition, an analytical closed form expression for interference mitigation on forward link of a satellite 2x2 MIMO diversity system with antenna selection is also obtained. In order to discuss the features of MIMO techniques, the model assumes propagation phenomena such as clear LOS, high antenna directivity, rain fading and rainfall spatial homogeneity. The propagation delay offset (synchronization problem) in LMS communications is also considered and a practical solution to this problem is found where matched filters are applied, first to the received signals for the detection of propagation delay offset and then the resulting signals are fed to the timing aligner. Subsequently, the delay offsets are eliminated by adjusting the timing of a signal serial-to-parallel converter.

The figure of merit for the analysis of MIMO fading channel is the outage capacity. A dual-satellite MIMO communication channel at Ku-band and above is shown in Fig. 1 in (Liolis et al., 2007). The terminal station is equipped with two highly directive collocated antennas to communicate with two satellites S1 and S2. The separation, $\Delta\theta$, between the antennas is kept large enough so that the spatial correlation due to rain along the relevant paths is as low as possible. Considering the clear LOS between the terminal station and each satellite and that each terminal station antenna is at bore-sight with the corresponding satellite, the total path loss along each link can be written as follows (Liolis et al., 2007):

$$A_i = FSL_i + A_{Ri} \qquad i = 1, 2 \tag{8}$$

where $FSL_i = 10\log_{10}(4\pi d_i f / c)^2$ is the free space loss along each link, f is the operating frequency and c is the speed of light. The term A_{Ri} represents the rain induced attenuation. The convective raincell model using Crane's assumptions is employed for the description of vertical variation of the rain fall structure using the same approach as in (Panagopoulos & Kanellopoulis, 2002). Based on these suppositions if $\Delta\theta$ is sufficiently large, the spatial correlation between random variables A_{Ri} representing channel coefficients is low and a decorrelated (ideal) MIMO satellite channel model is obtained.

To find out the capacity of dual satellite MIMO channel, it is assumed that the two satellites transmit different and independent data streams and the channel is perfectly known at the terminal side while the transmitting satellites have no information about the channel. Equal powers are allocated to the two satellites owing to distributive nature of the system in the absence of channel state information (CSI). The capacity of the dual satellite MIMO channel based on the standard MIMO theory, taking into the account above assumptions, can be written as follows:

$$C = \sum_{i=1}^{2} log_2(1 + 0.5SNR_{CS_i} 10^{-A_{Ri}/10}) \tag{9}$$

where SNR_{CS_i} are the nominal SNR ratio values under clear sky conditions. The above equation gives the instantaneous capacity of deterministic 2x2 MIMO channel. However due to rain induced attenuation and stochastic behavior of the channel, the appropriate metric to characterize the resulting fading channel is the outage capacity which can be written in the following form:

$$(C \le C_{out,q}) = \frac{1}{2} \int_{u_A}^{+\infty} du_1 fv_1(u_1) erfc\left(\frac{u_B \rho_{n12} u_1}{\sqrt{2(1-\rho_{n12}^2)}} \right) = q \tag{10}$$

where $C_{out,q}$ is the information rate guaranteed for (*1-q*) *100%* of channel realizations. The term ρ_{n12} is logarithmic correlation coefficient between normal random variables u_i and u_A, u_B . Complete mathematical details about these variables can be found in (Liolis et al., 2007) This model is applicable over a large range of SNR values and shows significant capacity gains of the MIMO systems over the SISO systems for moderate and high SNR levels.

The major factor limiting the capacity of LMS system is the adjacent satellite cochannel interference on its forward link. A satellite 2x2 MIMO diversity system based on receive antenna selection to mitigate the cochannel interference problem on the forward link due to differential rain attenuation from adjacent satellites is presented in (Liolis et al., 2007). In addition an analytical model is provided for interference mitigation based on satellite 2x2 MIMO diversity systems. To ease the complexity and cost associated with multiple RF chains only one RF chain is used at the receiver that performs antenna selection, i.e., it utilizes receive selection diversity to detect the signal related to the path with the highest SNR (Sanayei & Nosratinia, 2004). For optimal selection of a signal, the receiver scans the two antennas for a training signal transmitted along with data and the signal with the highest SNR is selected for the reception of the next data burst. This model uses a distinct approach for interference mitigation as compared to the conventional approaches found in MIMO theory. The effect of rainfall on interference analysis, differential rain attenuation related to an adjacent satellite and spatial inhomogeneity of rainfall medium are taken into account especially for congested urban areas where the increased demand for link capacity and radio coverage imposes the coexistence of many satellite radio links over the same geographical and spectral area. The satellite 2x2 MIMO diversity model using antenna selection and operating at Ku-band and above is illustrated in Fig. 4 in (Liolis et al., 2007). This model does not require channel knowledge at the transmitter side and only the information about wanted signals' channels at the receiver is desired. Mathematical derivation of this model and influence of various geometrical and operational system parameters on system performance can be found in (Liolis et al., 2007).

4. Application of Polarized MIMO Channels to Satellite Communications

The large MIMO gains can be achieved with low correlation between antenna elements at both ends of MIMO communication system. A fundamental way of achieving low antenna correlation is to use antenna elements with significant separation such that the relative phases of multipath contributions arriving at the receiving antennas are significantly different. However, owing to size restriction it is difficult to deploy multiple antennas with large separation at the mobile terminal. In addition, in communications where LOS is dominant (e.g., satellite communications), the MIMO systems offer reduced performance since LOS components overpower multipath components in the received signal. An alternative solution is to use polarized arrays for multiple antenna systems (Mohammed & Hult, 2008; Hult et al., 2010). With spatially separated cross-polarized antenna arrays, both the polarization diversity and polarization multiplexing can be achieved (Jiang et al., 2004; Moriatis et al., 2009) (e.g., two dual polarized spatially separated arrays form 4-antenna arrays).

In order to get benefit from polarization dimension, the cross-polar transmissions (e.g., transmission from vertically polarized antenna to horizontally polarized antenna) should be zero. However, in real scenarios there is always some polarization mismatch since linearly polarized antenna arrays have non-zero patterns for cross-polar fields (Hult & Mohammed, 2008). In addition, multipath effects (e.g., diffraction, scattering, reflection etc.) may change the plane of polarization of incident electromagnetic waves at the receiver.

4.1 Physical-Statistical Polarized MIMO Channel

In order to achieve additional diversity in satellite communication links, an extension of the physical-statistical MIMO model to 2x2 dual polarized MIMO model is presented in (Horvath et al., 2007) for a single satellite serving land mobile. A single satellite containing two antennas with right-hand circular polarization (RHCP) and left-hand circular polarization (LHCP), respectively, communicates with mobile terminal using the same antenna configuration. It is assumed in this model that LOS paths between co- and cross-polar channels are fully correlated and diffused multipaths are fully uncorrelated between co- and cross-polar components. The polarization characteristics are described by Stoke's theorem and the related concepts. The channel model construction is similar to (King, 2005) with additional polarization features are included as follows: In the case of unobstructed LOS signal, path loss is applied to the co-polar channels while cross-polar channels are neglected. When the LOS signal is blocked by buildings, rooftop diffraction loss is applied to both the co- and cross-polar channels. In the case of LOS path through vegetation, attenuation is applied depending upon the path length and using an attenuation factor of -1.3 dB. A mathematical representation of data generated by this model can be found in (Horvath et al, 2007). Extensive measurements were carried out along tree-lined/highway and urban environments using an artificial platform in order to optimize the model by fitting its parameters to the measured data. The model is capable of producing statistically accurate wideband channel first and second order characteristics and can be used to evaluate the capacity and diversity benefits of MIMO applications in LMS communication systems. The use of dual polarized system results in 2-fold increase in capacity and 4-fold boost in diversity gain (Horvath et al., 2007). The diversity gain can be further increased by

employing the concept of compact 3D-polarized antennas (Horvath & Friyges, 2006). In the case of synchronization problem in multi-satellite to ground links, one solution can be that MIMO antennas have to be collocated onboard a single satellite. This case is similar but not identical to handheld devices where the available space is limited. Here the available space is significant and antennas with large aperture and high gain can be applied. The channel statistics and power delay profiles are dependent on size and positioning of scatterers and reflection coefficients. The details of the experiments conducted to fine-tune the model can be found in (King, 2007).

4.2 Empirical-Statistical MIMO Channel Models for Satellites

The empirical-statistical channel models are built statistically around field measurements and can produce appropriate channel statistics for similar environments. An empirical-statistical 2x2 dual polarized LMS-MIMO channel model based on measurement campaigns is presented in (King, 2007). This MIMO channel model characterizes two models for LMS communication systems: the narrowband and the wideband channels. In order to describe 'ON/OFF' phenomenon occurring in LMS communication links, Markov chain approach is used. In narrowband model, single or dual lognormal fading, which is correlated over MIMO domain, is applied across each Markov state. In the case of wideband model lognormal fading is correlated over the delay domain as well and Rician factor is dependent on the large scale fading level.

To design a narrowband channel model, large scale data (lognormal fading) for each of the four channels in each Markov state is generated using a Gaussian random number generator with zero mean and unit standard deviation. The memoryless data streams (data samples) are then filtered using first order recursive linear time invariant digital filter in order to obtain correct temporal fading. The cross-correlation between large scale fading channels can be obtained in the following way:

$$vec(Y) = (\chi)^{1/2} vec(Y) \tag{11}$$

where χ is the correlation matrix obtained from measurements. Two large scale time series vectors are defined to represent dual lognormal fading which produces 8 time series correlated vectors: four vectors for each Markov state (each state represents a MIMO channel). Markov chains are illustrated by their state frame lengths, state vector W and transition matrix P. A minimum state frame of length 1 meter is found appropriate to capture the rapid changes in the level by observations. The probability of changing from state-i to state-j (both for co- and cross polar) can be calculated as follows:

$$P_{i,j} = \frac{M_{i,j}}{M_i} \tag{12}$$

where $M_{i,j}$ is the number of transitions from state-i to state-j and M_i is the number of state frames corresponding to state-i. A Markov chain Monte Carlo method can create random walk through the states for co- and cross polar channels using the Metropolis-Hastings

algorithm (Chib & Greenberg, 1995). Markov states are associated with each of the two channels obtained from 4 MIMO channels (2 co-polar and 2 cross-polar) resulting in 4 channels for each MIMO channel. Now Rician factor is defined using polynomial approach for large scale fading and Rician samples are generated by Rice's sum of sinusoids method (Pätzold et al., 1998). The average correlation coefficient value can be calculated on the basis of small scale MIMO channel correlation measurement data. The small scale and large scale fading time series data for each MIMO channel are added at each sampling instant to get the narrowband channel model.

A wideband channel model is constructed based on measurement campaigns and the schematic diagram of such a model is shown in Fig. 7.4 in (King, 2007). The 2x2 LMS MIMO channel matrix for this model can be written as:

$$Y_{R,L} = \begin{bmatrix} h(t,\tau)_{R,R} & h(t,\tau)_{R,L} \\ h(t,\tau)_{L,R} & h(t,\tau)_{L,L} \end{bmatrix} X_{R,L} + N_{R,L} \tag{13}$$

where $Y_{R,L}$, $X_{R,L}$ and $N_{R,L}$ are the received, transmitted and noise vectors, respectively. The subscripts, *L* and *R*, are used to denote the antennas polarizations at each end of the radio link. The amplitudes, propagation delays and phases of each path are randomly time varying parameters due to motion of the vehicle or the satellite. The phases of the paths are assumed to be mutually independent random variables with uniform distribution in 0 to 2π interval. All taps are considered to maintain frequency and delay domain resolution spaced *10 ns* and maximum resolvable delay is *400 ns* for the tree-line roads and urban environments and *200 ns* for the suburban environment. Large scale fading samples with correct autocorrelation and cross-correlation properties are generated for each tap using parameters obtained from the measurement data by the method similar to the narrowband model. In the case of wide band, the lognormal fading is correlated across both the MIMO domain and the delay domain. All taps in the wideband case are represented by the Markov state probability and state transition probability matrices with all taps in state 1 or in state 2 simultaneously. Each state in Markov chain is assigned a set of lognormal fading generators obtained from Gaussian random number generator with zero mean and unit standard deviation. Rician samples are generated in a similar way as for the narrowband case. The first tap of the delay line model is correlated with large scale fading level using polynomial fit. The value of K factor for remaining taps, derived from the measurement data, is lognormally distributed and independent of large scale fading level. The small scale fading generators are uncorrelated for all taps due to uncorrelated scattering effects. The small scale MIMO matrix samples are partially correlated for each delay tap. Now the small scale and large scale fading generators are combined by summing the outputs of the taps to obtain the channel impulse response time series generator.

The models described above can generate the narrowband and wideband channel models with well conserved first order, second order statistics and MIMO channel cross-correlations. However the parameters used to generate these models are based on specific measurement campaigns for particular environments. Thus, these models can be used to

generate fading effects for similar types of environments which limit the application of these models for different environments.

5. Conclusion

This chapter presented an overview of standard MIMO channel models for LMS radio communication systems. Standard channel models play a vital role in the design and performance assessment of advanced transceivers techniques and smart antennas employed to establish reliable communication links in LMS communication systems.

The quality of service and spectral efficiency in a LMS communication system suffer from limited transmit power, high path loss, blockage, shadowing and high link delay. In order to overcome these effects by employing MIMO techniques to enhance LMS performance, physical-statistical MIMO channel models including 3D polarization concept are developed. The ray-tracing algorithm is used which can model small and large fading effects in an efficient way and can cover large coverage areas for satellites systems. The physical-statistical channel models with multiple satellites and/or dual polarized antenna configurations can be simulated under different propagation environments and satellite elevations and show significant improvements in capacity and link reliability. Analytical channel models are developed to investigate the effects of MIMO techniques on capacity improvement and interference mitigation in LMS systems. The empirical-statistical MIMO channel models (narrowband and wideband) are constructed around experimental data. These channel models can generate the narrowband and wideband channel models with well conserved first order, second order statistics and MIMO channel cross-correlations. However, these models have some limitations since they are built from experimental data obtained from specific measurement campaigns for particular environments, they can be used to generate fading effects for similar types of environments which limit the application of these models for all environments.

The MIMO channel models for LMS systems are built and validated using some available experimental data and the data obtained from measurement campaigns. These channel models need to be refined to discover realistic hybrid physical-statistical channel models and satellite to indoor channel models using the data acquired from experimental campaigns that have been done or still going on in urban, sub-urban, rural, forest or indoor areas at L, S and C band and also the data gathered from Earth Observation systems. The spatial characteristics in multi-antenna channel modelling including polarization effects (especially 3D polarization) are expected to be crucial in future LMS communications systems. Thus, new and improved hybrid physical-statistical channel models, satellite-to-indoor propagation models and propagation impairment mitigation techniques based on multiple antennas techniques are necessary to assess the parameters and performance of satellite communication systems. The MIMO channel models for LMS/HAPs and indoor propagation environments employing compact antennas including 3D polarization concept based on different measurement campaigns and using the multiple antenna features similar to terrestrial communications systems will be the topic of interest for future satellite communication systems.

6. References

Almers, P., Bonek, E., Burr, A., Czink, N., Debbah, M., Degli-Esposti, Hofstetter, H., Kyosti, P., Laurenson, D., Matz, G., Molisch, A., F., Oestages, C., & Ozcelik, H. (2007). Survey of Channel and Radio Propagation Models for Wireless MIMO Systems. EURASIP Journal of Wireless Communication and Networking, volume 2007.

Chib, S., & Greenberg. (1995). Understanding the Metropolis-Hastings Algorithm. The American Statistician, 49(4), 327-335.

Correia, L. M. (2001). Wireless Flexible Personalized Communications. Cost 259: European Co-operation in Mobile Radio Research, 148-200.

Horvath, P., & Friyges, I. (2006). Application of 3D-Polarization Concept in Satellite MIMO Systems. In proceedings of 49th Annual IEEE Global Telecommunications Conference, San Francisco, Calif, USA.

Horvath, P., Karagiannidis, K. G., King, P. R., Stavrou, S., & Frigyes, I. (2007). Investigation in MIMO Satellite Channel Modelling: Accent on Polarization. EURASIP Journal of Wireless Communications and Networking.

Hult, T., & Mohammed, A. (2008). Evaluation of Depolarization Effects on the Performance of High Altitude Platforms (HAPs). IEEE 67th Vehicular Technology Conference, VTC08-Spring, Singapore.

Hult, T., Mohammed, A. Yang, Z., & Grace, D. (2010). Performance of a Multiple HAP System Employing Multiple Polarization. Invited Paper, Special Issue, Springer Wireless Personal Communications Journal, 52(1), 105-117.

Jiang. L., Thiele, L., & Jungnickel, V. (2004). On the Modelling of Polarized MIMO Channel. Fraunhofer Institute for Telecommunications Heinrich-Hertz-Institute Einsteinufer 37, D-10587, Berlin, Germany.

King, P. R., Evans, B. G., & Stavrou, S. (2005). Physical-Statistical Model for the Land Mobile-satellite Channel Applied to Satellite/HAP MIMO. 11th European Wireless Conference.

King, P. R. (2007). Modelling and Measurement of the Land Mobile Satellite MIMO Radio Propagation Channel. Ph. D Thesis, Centre for Communication Systems Research, University of Surrey, Guildford, UK.

Liolis, P.K., Panagopoulos, D. A., & Cottis, G. P. (2007). Multi-Satellite MIMO Communications at Ku-Band and above: Investigation on Spatial Multiplexing for Capacity Improvement and Selection Diversity for Interference Mitigation. EURASIP Journal of Wireless Communication and Networking, volume 2007.

Lokya, L. S. (2001). Channel Capacity of MIMO Architecture Using the Exponential Correlation Matrix. IEEE Communication Letters, 5(9), 369-370.

Mohammed, A., & Hult, T. (2008). Performance Evolution of a MIMO Satellite Diversity System. European Space Agency 10th International Workshop on Signal Processing for Space Communications, (SPSC 2008), Rhodes Island, Greece.

Molisch, A. F. (2004). A Generic Model for MIMO Wireless Propagation Channels in Macro- and Microcells. IEEE Transactions on Signal Processing, 52(1), 61-71.

Moraitis, N., Horvath, P., Constantinou, P., & Friyges, I. (2009). On the Capacity Evaluation of a Land Mobile Satellite System using Multiple Antennas at the Receiver. 3rd European Conference on Antennas and Propagation, Berlin, Germany.

Panagopoulos, A. D., & Kanellopoulos, J. D. (2002). Prediction of Triple-Orbital Diversity Performance in Earth-Space Communications. International Journal of Satellite Communications, 20(3), 187-200.

Pätzold, M., Killat, U., & Laue, F. (1998). On the Statistical Properties of Deterministic Simulation Models for Mobile Fading Channels. IEEE Transactions on Vehicular Technology, 47(1), 254-269.

Sanayei, S., & Nosratinia, A. (2004). Antenna Selection in MIMO Systems. IEEE Communication Magazine, 42(10), 68-73.

Saunders, S., & Zavala, A. (2007). Antennas and Propagation for Wireless Communication Systems. John Wiley & Sons Ltd.

Steinbauer, M., Molisch, A. F., & Bonek, E. (2001). The Double Directional Radio Channel. IEEE Antennas and Propagation Magazine, 43(4), 51-63.

5

Analysis of Uses and Metrology: an Experiment in Telecommunications by Satellite and Wireless Network Solution for Rural Areas

Fautrero Valérie, Fernandez Valérie and Puel Gilles
Télécom ParisTech/Université de Toulouse
France

1. Introduction

The TWISTER "Telecommunications by Satellite and Wireless Network Solution for Rural Areas" project (2005/2007) is a response to the European Commission's call for tenders, dedicated to satellite solutions. It brings together a consortium of satellite operators (EADS Astrium, Aramisca, Eutelsat and the CNES) and demonstrates not only the determination of Europe to combat the "digital divide" but also its refusal to abandon an important market, which is essentially in the hands of American operators. The project involves numerous experimental sites in seven European countries (notably France, Spain and Poland). It was based on a wide-scale test. Twister offers high-speed access via satellite, often coupled with Wi-Fi, as well as related applications specific to rural needs, concerning agriculture, health, e-business, e-learning, etc.

2. The research model

2.1 The research problem(s)

The research problem revolves around one question: that of analysing the *process of adoption and the prospects of wider deployment of technical systems*. The question of the wider deployment of the technical systems concerns not only the analysis of learning mechanisms and the appropriation of technology, but also that of the impact/effects on the territories used in the experiment and of the interplay of the stakeholders.

The analysis of the couplings referred to below is based on the means chosen to achieve the objective. Four "variables" interact here: technical systems; applications; territories; and the level of expertise of users.

Technical systems are understood to mean the mixed "alternative" systems of high-speed Internet access proposed in the Twister project, namely satellite (collection and coverage) and Wi-Fi (coverage).

Applications are understood to mean the protocols (and therefore the related uses) used by the people taking part in the experiment during the period of the project. These include among others the Web, transfers, mail, instantaneous messaging and VoIP.

Territories are understood to mean the experimental sites benefiting from the Twister project, and more specifically Tibiran-Jaunac, Lesponne/La Mongie (Haute-Pyrénées) for the qualitative part of the study, then the communes of Alpartir, Longas, Luesia, (13 in total, located in France, Spain and Poland).
Users are understood to mean the people benefiting from the service proposed within the framework of the Twister project. They have varying levels of ICT expertise: novices, regular users, experts.
These variables can be considered from the point of view of various couplings:

- focusing the analysis on the "applications/territories" coupling, in order to identify, for example, the variables explaining the adoption and appropriation of an application by a territory and the conditions necessary to ensure its effective transposition to other territories. A strong coupling between these two variables could reflect the wish of a local authority, responsible for territorial development, to put in place a structuring application: for example the deployment of health services as in Greece;
- the "applications/technical systems" coupling, in order to identify, for example, the attributes of the alternative technologies used, which might be an obstacle to the use of an application: problems of latency, sharing bandwidth in the case of use by communities, etc.). This coupling reflects whether or not the technical system is capable of supporting certain applications; for example, the satellite network system as deployed in Twister does not support applications such as VoIP, videoconferences, etc.;
- the "systems and territories" coupling: where the application plays the role of an intermediary variable in analysing the success factors of a technology, etc. A strong coupling reflects the dynamic created when a local authority decides to structure a territory around a technical system, such as covering non-ADSL enabled areas;
- the "technical systems and users" coupling: in order to identify, for example, the process of the adoption of a system by users. A strong coupling between these two variables could reflect the wish of users to appropriate technologies in order to have high-speed access.

2.2 Research premises

This research is based on the examination of pilot projects deploying alternative technologies which ran in France between 2005 and 2007, observed longitudinally using a qualitative approach (Yin, 2003). The "alternative" technologies pilot projects involved a large number of participants with differing rationalities: the State, local authorities, operators, components manufacturers, users, researchers, etc. How can we understand the complex dynamic interactions which occurred between these various participants in the pilot projects?
In economic literature, experimentation can be defined as a scientific research method, used to test products and services (in a laboratory or in the field), with a view to obtaining strategic information (technical, social or economic), before competitors. Its main interest is its force of persuasion in comparison with other types of studies, but its high cost means that only a small sample is used[1]. The demand for information increases when a new, untested product is

[1] *Encyclopaedia of sociology*, 2000, p.887–890

launched on a market (Grossman and others, 1977). Thus, companies that enter a new, unfamiliar market want information. They find themselves facing a situation where the level of demand is unknown and experimentation can help them to improve their market knowledge. For Bolton and Harris (1999), experimentation can lead to a strategic game between players, as a result of the information garnered. The player that puts in place such a system will obtain data which will, over time (this is variable), finish in the public domain, where it will be accessible to its competitors. Leach and Madhavan (1993) explain that in the short term, experimentation is expensive. But it can be cost-effective in a dynamic context where a large amount of information can be profitable for future actions.

In the telecommunications sector, the dual nature of experimentation makes it an important strategic tool, since experimentation is at the same time a research method, a test method (yielding information) and a defined territory. The opportunities for action, in an uncertain environment, are thus twofold: the deployment of such systems helps to shape an overall strategy with utilisation, service and product tests; the geographic proximity to potential customers enables the creation of demand and a market, as well as greater notoriety. In France, the Act on Confidence in the Digital Economy (of 21 June 2004) now allows local authorities to establish, in their territory, both passive and active infrastructures and to make them available to telecoms operators or users of independent networks. The first actions of the authorities involved local experiments. The persistence of areas not covered by ADSL (the citizens concerned demanded solutions) and the arrival of "new[2]" high-speed technologies on a "liberalised" market, led the State to develop these technologies. Thus an "alternative technologies" support fund was made available between 2003 and 2005 which presents, through an "organising" vision (Swanson and Ramiller, 1997), the latter as "credible". This credibility, which can be broken down into three aspects: operability, quality of service (performance, security and bandwidth) and cost, seem to vary considerably according to the territorial and time dimensions.

Wi-Fi technology is sensitive to the environment in which it is deployed: the presence of natural obstacles (trees, climate or topography) or imposing human constructions (buildings) can disrupt the propagation of waves (outdoors) and necessitate an investment in additional equipment to facilitate the latter or to fell trees that are sometimes over a hundred years old, etc. Coupled with satellite, it conveys the limits intrinsic in the latter technology: latency period, sharing bandwidth.

The market, as a social construction, is shaped by the environmental context and the political, economic and technological issues of the stakeholders involved. Therefore, the emergence, commercialisation and adoption of a technology are the fruit of negotiations involving private-sector stakeholders (operators, associations, components manufacturers, ISPs, etc.) and public-sector stakeholders (the State, local authorities, etc.): for example, the development of a technology or a technical standard supposes that the public authorities (through their standardisation bodies and their legislators) have authorised their use. The use of these alternative technologies, not always equivalent to the standard, seems consequently to have been the result of lobbying actions and the work of pressure groups. In the case of Wi-Fi, the presence of geeks grouped together in communities in a region, and their enthusiasm for this technology and its ideology of sharing, is a significant factor in the roll-out of wireless networks. In the case of "satellite" technology, the partnership between

[2] Some technologies have existed for many years, such as Wi-Fi (with different uses) and PLC (kept on the fringes of the market for some time – according to some observers for political reasons).

public-sector stakeholders (in particular locally elected representatives in municipalities located in white spaces), satellite and wireless operators led to the roll-out of "satellite/Wi-Fi" type high-speed solutions in mountainous or isolated areas.
The question of alternative technologies is therefore linked to the surrounding territorial and environmental context. Alternative technologies, as well as the interaction of pressure groups (relevant groups) are regionalised and bring together several levels (national, local, etc.) of stakeholders and negotiations. The "spirit of technology" (Swanson and Ramiller, 1997), which makes it possible to give a meaning and a collective image to the latter, is a joint venture in a specific socio-institutional context, the subject of negotiations, where the public authorities, users and industrialists-operators interact. From this analytical standpoint, technology is at the same time both a material artefact and a social object. "Users" in the same way as the other stakeholders participate in the social construction of technologies which have a "flexible nature" (Orlikowski, 1992). The interpretative flexibility varies according to the groups involved which may have a different perception of the same technology. Bijker (1995) mobilises the concept of "relevant social groups", categories of stakeholders that have an influence on the development of a technology.
The territories differ (socially, culturally and economically) and do not have the same social capital. We can assume that the types of usage and technology habits differ according to the interactions between the elements composing the territorial systems. However, each case is not unique and falls within the scope of a typical ideal. Therefore, we have observed territorial models (see above), which comply with a series of recurring types of behaviour and which, accordingly, may prove essential not only for understanding the world but also for the implementation of technical systems and public or private actions within the framework of territorial development.
The quantitative and qualitative studies of these questions may provide further ideas to be explored or responses on which appropriate action can be based: which jobs are "lost" as a result of the deployment of technology (which intermediaries?) and which jobs are developed or created? Where are they located?
Different replies to these questions may admittedly correspond to models of territorial contexts, but also possibly to territorial development choices. The nature and quality of the services deployed by the technical system will be an important parameter for analysing the results: developing a local granite production system or agritourism gives rise to different practices and territorial links.
However, the historical process will be equally pertinent for our analysis: there are moments in the deployment of technology (phase of appropriation, intensification of types of usage per level, etc.), which can explain differences.

3. Methodology and metrics

This survey relies more particularly on two partners of the Twister project (mentioned in the continuation of the document): EADS Astrium, co-ordinator and responsible for the technical supervision of the satellite/Wi-Fi networks and Aramiska, then Ouranos (from September 2006), the Internet satellite access service provider.
Our study covers the period from June 2005 to December 2006. Initially, preliminary work was necessary to delimit the territories covered by the Twister project. This first phase led us to examine a selection of 13 sites deployed in France, Spain and Poland. Geographical,

demographic and economic data were collected by way of Internet searches, supplemented by data provided by INSEE and its Spanish counterpart (INE) and finally the site dossiers provided by the local project leaders or co-ordinators.
At this stage, an initial typology comprising two categories of territories was presented at a meeting (end of August 2005) to EADS Astrium:
- the territories constructed by the services within the framework of the Twister project. Certain territories benefit from specific services and applications, in the areas of agriculture (in France, in the Allier, Dordogne and Charente), e-business (in Poland, in Suwalki, Hanzca, Lajski and Warsaw, etc.). The Twister typology of the sites for certain applications (e-business, education, agriculture, health, etc.) is the somewhat artificial result of the construction of the project. However, a more precise analysis would undoubtedly reveal similarities between the territories in Poland, Spain, France and Greece (territories with geographically isolated communities);
- community territories, that is to say territories benefiting from Internet access shared between the people taking part in the experiment in the territory; these may be "open", having a certain attractiveness in terms of population, tourism, culture, or "closed", that is to say suffering from their geographical isolation.
This typological interpretation enabled us to lay the foundation for a more detailed analysis of certain sites. The objective of the second phase, based essentially on quantitative data (provided by EADS Astrium), was to detect the usage processes of the people taking part in the experiment and/or of the sites, so as to be able to propose a more precise typology.
The data provided by EADS Astrium can be broken down as follows:

- monthly EADS Astrium data, in the form of charts:
- In and Out traffic by site (Upload and Download),
- breakdown of traffic consumed by each category of users ;
- the most used protocols for each site,
- the most used protocols for each category of users ;
- statistical reports sent to the co-ordinators of the first six DPZ sites, Cemagref Sites; University of Warsaw, Montaignac, Luesia, Francescas, Alpartir;
- Aramiska data: monthly aggregated In + Out traffic per site (namely, the aggregation of Up- and Downloads).

This data enabled us to test certain assumptions:

- the territory (through its characteristics) may be a variable, which explains certain specific means of appropriation;
- the type of user (institution, residential, etc.) may explain the development of certain types of use (e.g. *peer to peer*);
- the consumption of users may be linked to the days of the week.

The data were obtained from several sources (Aramiska, EADS Astrium), each using different collection methods, and relate to different periods (the projects were not put in place at the same time). For reasons of accuracy and in order to avoid approximations as much as possible, EADS Astrium provided us with raw data relating to the aforementioned charts (see Excel listing). We have therefore reprocessed the statistics in Excel and produced charts, in order to obtain results that are as objective as possible and overcome the difficulty of comparing projects, which were initiated at different periods. Thus, to supplement the charts representing the experiments according to the month of the year (e.g. May 2005, June

2005, etc.), but not taking into consideration the temporality of the projects, our charts represent the projects at the same stage of their deployment (Month 1, Month 2, etc.).
At the end of this phase, we selected the sites for the qualitative survey in order to cross-refer the results with the quantitative data.
We selected the communes of Lesponne, La Mongie and Tibiran-Jaunac (Hautes-Pyrénées). Our selection was based, on the one hand, on the deployment schedule for the experiments (it involved examining projects, which have been in place for a sufficient period of time for the people taking part in the experiment to have had time to use the system) and, on the other hand, the opinion of our partner (we had to be sure that the technical installation period had been completed and that there were no specific technical problems affecting the correct use of the system).
In order to delimit more closely the experimental sites, and to get to know and understand as closely as possible the people taking part in the experiment, and their types of usage, our research was based on a series of face-to-face interviews with users, local administrators and administrative managers. The interviews were conducted in April 2006 in Lesponne, La Mongie and in December 2006 in Tibiran-Jaunac.
The interview guide used, adapted according to the three aforementioned categories of interviewees, comprised three main parts: presentation of the interviewee, presentation of the territory and opinions on the experimentation (see appendix).

3.1 The limits of the work

However, several factors have to be taken into consideration in any analysis of the results obtained:

- the occasional lack of data: the (traffic) monitoring data was affected by several technical failures. These failures affected almost all the sites studied, which had an adverse effect on the "correct" interpretation of the statistics;
- the *a priori* selection of the sites given that our work is based only on the information provided. The sites studied were therefore selected *de facto*;
- the recent implementation of the experiment and therefore limited basis of assessment. Our assessment of changes in use needs to be seen in the context of the recent deployment of the systems. They were deployed less than six months earlier;
- the data is collected by volume (MB). Our analysis is therefore based on the actual traffic consumed by the people taking part in the experiment. Accordingly, types of usage requiring little traffic (for example, e-mails) are under-represented;
- finally, we lack information on the territorial ecology of the various sites and on the people taking part in the experiment (for example, their professional and personal activities, etc.).[3]

Questionnaires and face-to-face interviews are the only way of gaining a better understanding of actual uses.

[3] Information on the profile of users (professional or residential), as well as the age profiles for certain users (DPZ) was available.

4. The quantitative part

4.1 Bandwidth consumption

The aim of the chart 1 is to correct the over- or under-representation of certain sites, which have a large number or few people taking part in the experiment. The aim, therefore, is to look at the average consumption per user of each of the sites on a monthly basis.

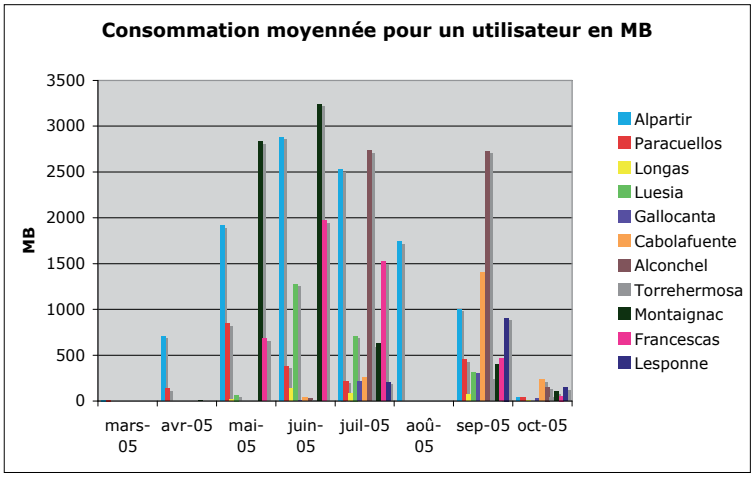


Fig. 1. Average consumption of bandwidth per user

Consumption seems to follow in time an inverted "U" curve. We can therefore hypothesise that it is a normal law, which reoccurs often in the learning process.

Several sites stand out: Alpartir, Francescas, Montaignac and Alconchel. The sample of people taking part in the experiment in Alpartir includes a significant number of students (the site is near a college offering ICT courses[4]), and in Francescas and Montaignac a not insignificant percentage of companies. However, at this stage of our research, there are no specific reasons that explain the consumption of Alconchel.

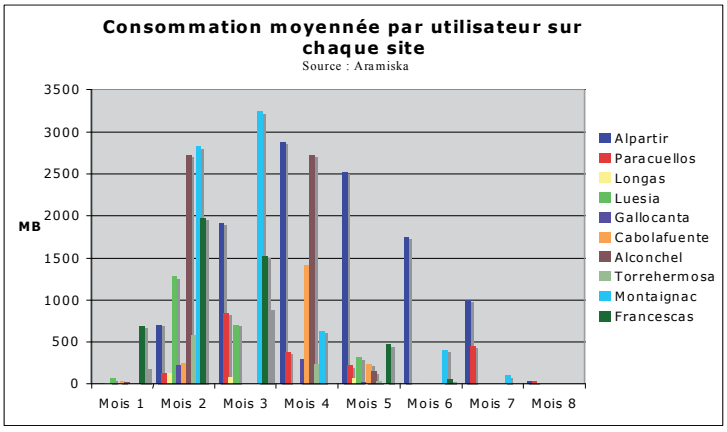


Fig. 2. Average consumption and stage of deployment

[4] Which seems to confirm the importance of the sociological variable on the profile of users in our analysis.

Although the graph allows the different sites to be viewed simultaneously, it is to be noted that the sites are not at the same stage of deployment. For example, the experiment started in Cabolafuente in June 2005, consumption increased in July, was at its highest in September and then declined in October. On the other hand, Luesia started earlier and peaked earlier. The "time" dimension, as it is represented, makes any comparison difficult. The four sites that stand out must be treated with circumspection. However, we can consider that traffic has not yet stabilised.
The aim of the chart 2 is to compare the average consumption of one user on each of the sites, at the same stage of deployment of the system. For these sites, the duration of the experiment is between three and six months. The inverted "U" curve (or normal law) is confirmed. In the first month(s), consumption is low. This period is followed by a strong increase in traffic then, more or less stable, consumption declines at differing rates. Alpartir illustrates perfectly this curve.
We can therefore put forward the hypothesis that the people taking part in the experiment need time to adapt and familiarise themselves with the connection, and that their consumption is important during the initial part of the experiment and that their interest seems to decline subsequently. Or is it their use of the connection which changes? Greater use of the Web? More e-mails? It must be borne in mind that the data are in MB and that the representation of types of usage requiring little traffic is difficult[5].

4.2 Types of usage

During the period of the experiment, EADS Astrium analysed the most used network protocols, by category of users per site and for all the sites. The traffic consumed for each of the protocols was then aggregated by type of usage. Based on that analysis, 10 main types of usage were determined via the related protocols. The following table shows the types of usage and their related protocols:

Web Browsing	HTTP, HTTPS, NNTP-TCP
File Transfer	FTP, TFTP, NETBIOS-IP, NFS, SYSLOG, PRINTER, PRINT-SRV, RCP, SUNRCP, CMD
Mail	POP, SMTP, MS Exchange, IMAPS, CCMAIL, LOTUS-NOTES
Online Messaging	MSN Messenger, AOL/ICQ, Yahoo, IRC
Admin & Supervision	SNMP, TELNET, etc.
Online Gaming	COUNTERSTRIKE, MSN GAME, NEED FOR SPEED, etc.
Peer To Peer	KAZAA, EDONKEY, GNUTELLA, BitTorrent, WINMX, OVERNET, Soulseek, etc.
Real Time Stream Video & Audio	RTSP, Msplayer, REALAUDIO, NETSHOW, Winamp, etc.
VoIP	SKYPE, MGCP, H323, T120, VOCALTEC-IPHONE, PHILIPS-VC-TCP
All Others	All Others: Network Infrastructure (ARP, RIP, DNS etc.)+Unidentified protocols ports

Table 1. Types of usage and their related protocols

[5] A comparison of these data with data relating similar systems (ADSL type) would undoubtedly be interesting.

As a result of the important share of the "Others" category on certain sites, discussions were held with EADS Astrium in order to try and identify the types of usage falling within that category. In fact 80% of this category corresponds to a P2P activity. The chart integrates the transfer of the 80% of "Others" to the P2P category.
In order to identify the types of user behaviour, we have compared average uses (as a percentage) by site. The table 2 highlights the average (as a percentage and over the duration of the experiment for each site) of the sites by category of usage and the overall average per type of usage (in yellow).

Site	Web	File transfert	Mail	On line Message	On line Game	P2P	Streaming	Voip	20% Others
Alpartir	34	5	0	0	0	48	1	0	11
Alconchel	27	20	0	0	0	41	0	2	10
Galloconta	48	0	1	0	2	38	1	0	9
Montaignac	10	9	2	0	0	64	1	0	13
Francescas	66	20	1	0	0	5	5	0	1
Cabolafuente	54	9	0	0	0	34	1	0	2
Luesia	63	1	2	1	0	14	17	0	3
Torrehermosa	59	1	0	0	0	30	3	0	7
Longas	49	2	0	0	0	42	2	0	5
Paracuellos	55	2	0	1	0	37	4	0	2
Average	49	5	0	0	0	35	4	0	6

Table 2. Average by category of usage by sites

The main uses are the Web and P2P. Together they represent approximately 80% (aggregated) of the traffic.
The percentages above the general average for the type of usage are marked in black. This highlights three groups of user behaviour:

- those with above average use of P2P: Gallocanta, Alpartir and Alconchel and Montaignac;
- those with above average use of the Web: Cabolafuente, Luesia and Torrehermosa and Francescas;
- those with above average use of both: Paracuellos and Longas

4.3 Conclusions on this phase

This analysis raises numerous questions. The data collected are expressed in volume (MB). It is therefore the traffic which is studied. Types of usage requiring little traffic are accordingly under-represented in the charts. In addition, the analysis is limited by the "time" dimension. The 13 sites were deployed only recently (less than six months earlier) and our results must be seen in the context of the limited duration of the experiment.

Accordingly, we added a qualitative phase to our research, based on interviews, in order to ascertain more accurately the actual types of usage.

In order to implement this new phase in our research we selected two sites: Lesponne/La Mongie and Tibiran-Jaunac (Hautes-Pyrénées).

5. The qualitative part of our analysis

5.1 The Lesponne and La Mongie experiment

The experiment took place from July 2005 in Lesponne and from October 2005 in La Mongie. In La Mongie, ADSL arrived in October 2005, following the signature of the innovative department's charter of France Télécom by the Hautes-Pyrénées department. The commune of La Mongie, included in France Télécom's deployment schedule, was connected more rapidly than initially planned. The operator accelerated the procedure.

Fig. 3. Localization of Lesponne/La Mongie source Google map

The valley of Lesponne (located in the Hautes-Pyrénées department) is a non-enabled area for mobile telephones and high-speed access, and for certain people, not having high-speed Internet access is perceived as an intolerable disadvantage.

The people taking part in the experiment in Lesponne are connected to two networks:

- the first, in the centre of the village connects directly the town hall and the school via satellite, business people (self-employed people, teleworkers, and an inn-restaurant) and three private individuals, thanks to a Wi-Fi relay installed on the church steeple;
- the second network is dedicated to the hotel resort of Ramonjuan which provides its guests with Wi-Fi access via a satellite connection.

5.1.1 Proposed uses

When the experiment was put in place in La Mongie, a series of high-speed services dedicated to mountain-related businesses and activities were planned: telemedicine, ski-related traumatology via videophone, etc. The administrative and medical centres intended to use the connection to meet their own needs and then to offer, on the one hand, access to tourists and, on the other hand, for telediagnoses with Rangueil and Purpan in Toulouse. However, the ADSL connection was installed at the end of 2005 and the bankruptcy of Aramiska, which occurred one and a half months after the start of the experiment, delayed initiatives and reflection on possible uses. This interruption in service (lasting 15 days in February 2006) did not make matters easier, but one of the reasons was the lack of involvement of the local authorities in these service projects.

5.1.2 Types of usage and satisfaction with the experimental system

The experimental system does not seem to behave in exactly the same way as ADSL technology, which the majority of people taking part in the experiment have already used, at a friend's place or at work, etc.[6].

On an everyday basis, the basic types of usage (searches, mail, browsing, etc.) have not given rise to any special problems. The types of usage of certain people taking part in the experiment have even developed, searches, online games, online shopping, managing Internet sites, etc. but they have all kept their old subscription and have used it several times (at the time of system failures, etc.). The Internet has replaced certain physical journeys to Tarbes (approximately 60 km away).

However, certain types of usage, which require more bandwidth, such as voice over IP (Skype), the use of instantaneous messaging software (MSN type voice and image) have apparently caused certain problems[7]. It appears that "synchronisation gaps" make videoconferences impossible. The signal is seen as too weak and the speed insufficient. However, the hotel resort of Ramonjuan is planning to build a conference room (with a movable amphitheatre) equipped with a videoconference installation. This installation would enable videoconferences to be organised, as well as live training sessions, on the Internet, during seminars. For the Ramonjuan hotel resort, the Internet plays an important role in developing its activities. Most of its estimates are sent by the Internet, some 75% of reservations and communications with guests are made by e-mail, and a number of seminars have been organised to take advantage of the Wi-Fi connection. Guests consider the quality of the connection to be good. Moreover, guests have been pleasantly surprised to have such a connection in the valley. Offering Wi-Fi access to guests is seen as a competitive advantage compared with other hotel resorts.

---

[6] It would be interesting to have data on the way in which ADSL behaves in terms of deploying technology.

[7] The satellite technical system was not designed to support this type of application and users were informed of this during the advanced Twister training they received at the start of the experiment. However, perhaps we should see this as reflecting a lack of information given to the users regarding the limits of the system used in the experiment.

The people taking part in the experiment consider that they are very fortunate to have high-speed Internet access. They are very satisfied with the service and the fact that it is free is considered as adequate compensation for the frequent technical incidents.
Interest in the experiment has increased over time. Several people have asked to join the project.

M. C.: co-ordinator in Lesponne.
Co-ordinator and technical intermediary between EADS Astrium and the people taking part in the experiment in Lesponne, M. C. also teaches IT at the school, on the three stations connected to the Twister system. Trained as a graphic designer and formerly an emergency physician, he has important professional needs and is passionately interested in high-speed access.
His main professional activity consists in creating and updating Internet sites. High-speed access enables him to work from home as a teleworker. He has noted several differences in the use of the system in comparison to ADSL technology. With ADSL, pages are displayed progressively (from top to bottom). A blank page is displayed, and then the full page is displayed all at once. In addition, transfers take longer with the satellite system. Therefore, when he wants to modify the Internet pages of a site, the modifications do not appear, because of the cache.
Skype does not work. Sound and video do not work. The MSN connection is often lost. Words are not synchronised, as a result of the problem of data sent and received via the cache. On the other hand, there are no problems for basic types of usage[8].
Although he has to deal with all the problems of the people taking part in the experiment (which means that he has to be reactive and a lot of his time is taken up resolving problems), he is pleased to have high-speed access.
Interview conducted on 6 April 2006

"Chez G." inn, Lesponne: the Internet "s a work tool".
The owners of the inn moved to Lesponne in 2003. They have two Internet sites devoted to their professional activities: one to the inn itself, with online reservations, which they set up when they arrived in 2003. The calculate that 80% of the inn's reservations are made via the Internet site and consider that the site is a very good means of communication and helps to enhance the inn's visibility beyond the local area. Numerous Anglo-Saxons visit the region and find an online service easier to understand and use. The second Internet site, devoted to their mountain guide activity, has existed since 2001. It gives details of prices, availability and other information, updated on daily basis, useful for tourists.
Since the beginning of the experiment, their types of usage have developed and increased: a large number of searches (for their children's schoolwork, news, etc.), online games, etc. They now spend more time connected to the Internet, but do not consider that there has been any substitution.
Interview conducted on 6 April 2006

[8] The satellite system was designed for basic Internet services, not to support Skype or MSN.

5.1.3 Problems perceived by users

The service has been subject to numerous incidents and interruptions. "Even if we do not pay for it, it still needs to work". "It must work properly, otherwise we need ADSL."

For most users, the system put in place is not perceived as an experiment or a test, but as a permanent way of having high-speed access (because it is often presented as such). Therefore, the problems encountered can provoke strong reactions, especially when the failures are too frequent, despite the fact that access is free.

Four types of problems/drawbacks have been encountered: technical problems, a lack of information and a failure to take on board the views of the people taking part in the experiment, the conflict between bandwidth and data security.

Technical problems have a variety of causes. It seems that they occur mainly during the weekend, from Saturday, because of the types of usage developed (MSN and downloading files) by the children of those taking part in the experiment.

The satellite dish is easily impaired by snow (even with little snow) and simple snowflakes or drops of water can interrupt the Internet connection. A solution is possible, but was not "taken into account" within the framework of the experiment. "This would involve investing in a protective system, the cost of which would vary, that is to say either a resistant cone or a ventilation system". Therefore a tree branch or a broom is used to clear the snow from the satellite dish, which is fixed at a height that makes it accessible for this purpose.

This problem has led to other problems, because "everyone has tinkered with the connection", giving rise to numerous individual "parameterization" problems.

Finally, numerous elements that cause failures and uncertainties need to be managed by the site co-ordinator: installation, personal connections, equipment, human handling, etc., making the task delicate and difficult. That is his role as defined by the Twister system. In this regard, a lack of information, the co-ordinator's level of expertise and availability can create problems.

Users are critical of what they perceive as a lack of communication and information provided within the framework of the experiment:

- technical failures and the means to be put in place to resolve them;
- the management of exceptional events. The bankruptcy of Aramiska raised concerns and questions, to which no answers were given[9];
- the installation of the equipment and material by technicians. The system was not co-constructed with the users, resulting in certain shortcomings in the installation. For example, apparently, connection sockets were installed "anywhere" (sensitive places, with a lot of comings and goings, for example, next to a door). "And these are very fragile, sensitive";
- At the Ramonjuan hotel resort, Wi-Fi is available free of charge to guests, but the system chosen, within the framework of the experiment, is based on the use of vouchers. A connection to the network requires a code composed of a ten letters and digits (given by the hotel), a login and a password. This is valid for one hour or one day depending on the voucher chosen. However, when the computer switches to standby mode, users need to recommence the identification system. This can be

[9] It is to be noted that the bankruptcy of Aramiska resulting in the service being interrupted for "only" 15 days.

a constraint given that the connection is free and people attending seminars stay at the hotel for several days[10].

There have been bandwidth conflicts between the people taking part in the experiment in Lesponne. The downloading of MP3 (at the beginning) was one of the causes of problems of connections and bandwidth sharing. To resolve this excessive use, the co-ordinator discussed the issue with users to raise their awareness of the problem. He "spread the word" and the problem was solved.

Finally, the security of data on the Wi-Fi network was deemed inadequate.

5.1.4 Resolving the problems

Three types of behaviour[11] were identified to resolve the failures and losses of connection, according to their "attachment" to the connection:

- the individual who has Wi-Fi access but whose use remains occasional or optional: this type of user lets things take their course and waits for the problem to solve itself. During this time (s)he uses the public switched telephone network (PSTN);
- the individual who has Wi-Fi access: this user unplugs and reconnects his modem. If the failure persists, he calls the co-ordinator, either to come and note the failure and try and solve it, or to call EADS Astrium (the co-ordinator is the direct contact person);
- the co-ordinator or the person in whose home the satellite dish in installed: he removes snow in the event of problems, irrespective of the time of day, and often helps the other people taking part in the experiment when they have problems.

During the period of the experiment, irrespective of the initial level of awareness of the users, most of them acquired rare skills, inherent in this technology, to resolve problems involving a loss of connection[12].

5.1.5 The future prospects

Although price information (on the proposed cost at the end of the experiment) was communicated in October 2006 when the sites in question were switched to a Eutelsat solution after the second bankruptcy, this time that of the operator Ouranos, users complained about a certain vagueness. The community of communes could intervene and finance part of the installations. High-speed Internet access has become indispensable for some of the people taking part in the experiment, who have even declared that they will move from the area if high-speed access does not become permanent.

In the Ramonjuan hotel resort, the end of the experiment has not yet been really anticipated (but several diversification attempts at the hotel resort have failed), the question of the cost has not been studied, but "it will not end there"[13]. Moreover, a consultant was

---

[10] If all this information feedback is interesting, it is to be noted that it was not obtained during the project, but once the project was terminated. This sociological behaviour may seem specific to the context of the experiment.

[11] This behavioural typology suggests that the user profile is a variable in its own right of the model, in the same way as the system, the application and the territory.

[12] It must be remembered, however, that TWISTER provided specific training for the local administrators as well as technical information for users.

[13] The words of the hotel's owner.

commissioned by the CIDAP to carry out a market research study for the hotel resort. This study, which was carried out by a consultant from Toulouse and financed in part by the regional council, concluded in particular that there was no regional market for industrial seminars that would justify the deployment of a videoconference system.
In La Mongie, all the people taking part in the experiment switched to ADSL as soon as that technology was deployed in the resort, well before the end of the experiment. ADSL is a system which seems to "kill" any alternative solution, irrespective of its quality.
The list of people taking part in the experiment is closed, because of the speed proposed and the need for sharing as well as the equipment to be financed, but several structures and individuals would like to participate in the experiment: a restaurant owner, as well as two people who have recently moved to the area: an English teacher who wants to organise online language courses and a person who wants to organise martial arts videoconferences. This situation creates a certain amount of jealousy among those who are "excluded" from the experiment.
Despite the failures experienced by the people taking part in the experiment, it appears that the Twister solution is unanimously perceived very positively and interests a growing number of people who have recently settled in the valley. The absence of ADSL and the fact that the experimental system is free encourages people who have recently arrived in the area (English people, neo-rural residents, etc.) to try and participate in the project. However, there is a problem with the quality of service of the system for users having large bandwidth needs.

5.2 The experiment in Tibiran-Jaunac

Tibiran-Jaunac-Las Aribes, a commune in the Hautes-Pyrénées, has 249 inhabitants, including a retirement home with 62 residents. Its territory is very vast: the subscriber node is 8 km from the nearest telephone subscriber and 14 km from the furthest subscriber. Tibiran is a perennial non-covered area, which, according to the mayor, does not interest the operators and whose network is not maintained by France Télécom.
The majority of the commune's inhabitants are elderly and are from the farming sector. Active inhabitants work mainly in Tarbes and Toulouse. The development plans include a communal estate.

Fig. 4. Localization of Tiriban-Jaunac Source: Google map

5.2.1 The genesis of the project

Local demand, composed of recently retired people and young people of school age, is strong. At the end of 2004 the mayor contacted France Télécom to connect the commune to ADSL. The operator considered that the commune was not eligible, but indicated that it could consider making a commercial offer if the commune financed the installation, at an estimated cost of €300,000.

There is still a strong demand within the commune and the mayor would like to find another cheaper solution. The project was then included in the Twister project at the end of 2005.

The mayor's personality and his ICT knowledge play a fundamental role in the high-speed project for the commune. He is familiar with the technology involved since he is seconded from the Interior Ministry to the ICT department of the Prefecture of the Hautes Pyrénées. The mayor decided, as soon as Tibiran was accepted as a Twister site, to put in place a veritable infrastructure policy for his territory by investing in a radio network to connect all the inhabitants.

5.2.2 The needs and projects

High-speed access is seen as a possible way of delivering more neighbourhood services to the commune's inhabitants:

- at the current time the town hall is open only on Saturdays: therefore the town hall intends to propose uses linked to the dematerialisation of procedures. It also wants to put in place an Internet access point, with three stations, linked to the town hall (training, etc.);
- the retirement home that provides medical care has considerable needs. These include, in particular, being able to connect to the local management system in Tarbes via a high-speed link; personal services are also planned (remote transmission, telediagnosis, etc.);
- since June 2006, it has been necessary to provide notaries public or purchasers of property with information regarding the commune and/or the risks related to the plot of land. The idea is to provide direct online access to such information;
- finally, the farm/inn present in the commune wants to develop Internet tools, such as an Internet site with online reservations.

5.2.3 The use made of the experimental system and the problems encountered

The system was deployed in Tibiran-Jaunac at the end of 2005. It connects the 12 people taking part in the experiment. The participants were selected by the mayor, based on the first applications, while trying to achieve a mix of ages and types of usage. Some participants were used to high-speed access, while others were primarily motivated by the geographical distance separating them from their families.

The installation of the Wi-Fi system required some trees, some of which were a hundred year old, to be pruned or chopped down. The owners had to be "pushed" to accept this sacrifice. Health concerns regarding the effects of Hertzien waves were also an obstacle.

The town hall considered that the bandwidth of 1Mb allocated by EADS Astrium[14] to all the people taking part in the experiment was less than the speed announced initially. No *peer-to-peer* problems have been encountered. An intelligent bandwidth manager was installed to this end (the last usage allocated was *peer-to-peer*).
The system has been affected by numerous losses of connection, due in particular to the failure of Aramiska, whose role was taken over by Ouranos Networks and other service interruptions to the satellite network. In the case of persistent failures, EADS Astrium's *hotline* is contacted. This is the only means used. The Wi-Fi network functions correctly.

Mrs B.: a participant aged 70
Mrs B. is a keen user of IT and the Internet. She has had a computer since 2001 and uses it intensively. She connects to the Internet for several hours every day. She uses it to make numerous searches and to pursue her leisure activities: information on embroidery and consulting forums on the subject, looking up train timetables, use of Google Earth to see how places she has visited have changed, genealogy searches, sending photos and looking more closely into information read in "Micro hebdo", etc. She also uses Skype every day (with headphones and Webcam) to call her daughter in Toulon. For her, high-speed access is indispensable.
She does not shop online (fears regarding bank card security) but looks up information, makes her choice and sends a shopping list to her daughter who then buys the items in question for her on eBay or other online shopping sites.
Thanks to the Internet and its search facilities, she avoids certain physical journeys and saves on the cost of telephone calls and postage stamps.
Finally, the only problem which she encounters is the English vocabulary specific to IT. Her dictionary is incomplete.
Interview conducted on 7 December 2006

Mr C.: a regular Internet user
Mr C. works at the University of Toulouse and commutes every day to and from Tibiran. He has used high-speed access for his professional activities since 1999. The Internet enables him to carry out a large number of searches on his passions in various fields: astronomy (in his view the photos on the Internet are better than direct viewing via an astronomical telescope), archaeology, the daily news, to supplement the radio and written press, local weather forecasts, etc.
The Internet is a means of putting daily information into perspective and making readers more responsible. "With a stone, you can crack a nut, or the neighbour's head."
Alongside the searches, which he also carries out at work, he also sends e-mails and photos to his friends. He neither uses Internet telephony services nor online shopping sites. The Internet enables him to avoid having to do certain tasks physically (buying train tickets) but, on the other hand, has had the opposite effect (discovering a museum online then physically visiting the museum).

[14] At the end of 2005, Aramiska (then Ouranos) provided the Tibiran site with a free access service of 2Mbps/512kbps. From October 2006, the speed was reduced to 1Mbps/256kbps.

He has sometimes carried out searches to help neighbours who did not want to take part in the experiment, but who now want to be connected.
Interview conducted on 7 December 2006

5.2.4 Assessment of the results of the experiment

Certain positive elements have emerged. At the end of the experiment, Wi-Fi technology is seen as both reliable and perennial. Local demand has increased strongly. Over and above the hard core of 12 people taking part in the experiment, thanks to word of mouth 300 people are currently awaiting a high-speed connection. For the town hall, which has a project for a communal estate, high-speed access is important to attract and keep people in the commune. The neighbouring communes also want to be connected to Tibiran-Jaunac's high-speed network.

There are however certain negative points:

- frequent losses of connection;
- the latency time is seen as long: this is a user perception of the system's performance. The relation with the latency time caused by the satellite may seem hazardous without any further information;
- the current project (the Twister experiment with its time and financial limits) does not allow more people to be connected[15];
- the Internet satellite service available on the market does not support VoIP and there is strong demand for the voice-data coupling, "at the current time the network is too poor";
- Wi-Fi nuisances (waves, old trees, leaves, etc.) have been important;
- the continuity of the system after the experiment was not envisaged in the project. It depends to a large extent on the mayor and his investment in Wi-Fi. Twister has contributed the satellite part.

5.2.5 The future prospects

Since December 2006, the commune has been fully Wi-Fi covered, thanks to 6 relays and one satellite access point. The whole of this infrastructure has been financed by the commune (€50,000) and subsidised for the time being solely by the State. The Midi-Pyrénées region is planning a call for projects to which the commune must respond. To that end, it must establish that no tenders have been submitted in response to its call for tenders (in accordance with the LEN), but the context is experimental, which changes things somewhat. The departmental council has not yet been mobilised on the question but lobbying has started.

For the experiment to be transformed into a commercial offering at an affordable cost for inhabitants, an economic model must be found. There are two thresholds fixed by the satellite offering and the ADSL offering in cities (not more than €3 for 1Mb per subscriber). There is also a psychological appropriation threshold by the user (irrespective of the user's location and the technology used) of the Internet access service which is that of ADSL, in urban areas. Indication: "The 2Mbs satellite connection is €700 a month. For 30 users, the

[15] The satellite technical system could connect more users, in exchange for a financial contribution.

cost is acceptable, but it will not be feasible to amortize the equipment and it will be difficult to renew the technical equipment every four years."
The town hall is holding discussion with TDF for the installation of a 2Mbs SDSL link to replace the satellite access point (approximately €400 a month). The aim is to conclude a public service delegation contract on a leasing basis with an operator to operate the network. But the changeover is both difficult (the software and equipment used for the network are proprietary systems and it would therefore be difficult for an operator to use them) and urgent (the project ends at the beginning of February 2007).
"For the time being the transition from experimentation to a commercial operation would be counter-productive. There would be some dissatisfied people."

5.3 Conclusion

The provisional evaluation of the Twister projects is positive. The solutions meet the high-speed needs identified beforehand. The deployment of the experiment has stirred interest in the valley, with increasing demand (successful school courses, reflection on the creation of a public Internet space, etc.). As a result other people have expressed an interest in benefiting from this offering which is attractive because, on the one hand, the equipment and material are fully financed and installed and, on the other hand, the connection is offered free of charge throughout the experiment. Subsequently, the community of communes could finance part of the installations to help maintain the system on the territory on a permanent basis, confirming the existence on this territory of a true digital development project, through the deployment of a technical system.
Nevertheless, the deployment of ADSL is developing as well as its technology (Re ADSL, etc.), the progressive covering of the territory by standard technology is a threat to the experiment. The technical difficulties encountered by the people taking part in the experiment and Aramiska's bankruptcy have hindered the development of this technology.
Finally, a problem of "governance" (lack of information, the fact that the technical system and the types of usage and services proposed were not co-constructed, etc.) was identified. The people taking part in the experiment want the circulation of information to be improved.

6. Conclusion

The research problem revolves around one question: that of analysing the *process of adoption and the prospects of wider deployment of technical systems*. The question of the wider deployment of the technical systems concerns not only the analysis of learning mechanisms and the appropriation of technology, but also that of the impact/effects on the territories used in the experiment and of the interplay of the stakeholders.
The analysis of the couplings referred to below is based on the means chosen to achieve the objective. Four "variables" interact here (technical systems; applications; territories and users) and can be considered from the point of view of various couplings:
the "applications/territories" coupling, in order to identify, for example, the variables explaining the adoption and appropriation of an application by a territory and the conditions necessary to ensure its effective transposition to other territories;

the "applications/technical systems" coupling, in order to identify, for example, the attributes of the alternative technologies used, which might be an obstacle to the use of an application: problems of latency, sharing bandwidth in the case of use by communities, etc.);
the "systems and territories" coupling: where the application plays the role of an intermediary variable in analysing the success factors of a technology, etc.;
the "technical systems and users" coupling: in order to identify, for example, the process of the adoption of a system by users.
Our analysis has enabled us to confirm the relevance of certain couplings of variables.
For the "applications-territories" coupling, certain territorial stakeholders want to develop more particularly structuring applications, such as, for example, e-administration services (for citizens), telemedicine (for retirement homers, ski resorts, etc.). However these projects are still on the drawing-board and have not been implemented. It seems that the first phase is to use the system and assess its reliability, credibility, etc., which is a key stage in the process of taking things further and actually implementing projects. However, satellite connections do not support certain uses for which there is now strong demand (basic uses are no longer enough. The interaction of the "applications-systems" coupling is in that case weak. However, interest in high-speed Internet access is not called into question. The "systems-users" coupling (through the appropriation process) and the "systems-territories" coupling (through the action of the local authorities) are very strong, confirming the major attraction of high-speed access.
Our assessment of the results of the experiment reveals certain negative elements.
1) problems regarding the quality of service in relation to types of usages, which can consume bandwidth or a demand that tends to replicate urban demand (success of VoIP);
2) a question of territorial governance: how to manage the sharing of bandwidth? Self-management or a 'technical' solution (software but who decides what and according to which criteria?);
3) the choice of territories used in the experiment: depending on the strategies of operators and regulatory and technological developments, the number of so-called perennial non-covered is shrinking.
There seem to be two types of territories: the transitional non-covered territories where satellite technology coupled with Wi-Fi seems to be a good temporary solution; in such cases the operators are preparing the market for the arrival of France Télécom since the business model is not competitive in the case of the French market. This raises several questions: how to transform these transitional solutions into a financially feasible market for operators? How long can this transitional period last? What prices should be charged? Is it necessary for the public authorities to contribute funding? For the perennial non-covered territories, satellite technology can be a solution if it can solve the problems of quality of service (see above) and provide a more efficient response to user requirements (*peer-to-peer*, VoIP);
4) finally, the local stakeholders participating in the experiment have quickly identified a decisive element, which is a stumbling block to territorial development: support services. The communes involved are small territories, which have limited financial resources and have difficulty not only finding the necessary technical and human competences locally, but also recruiting a "support co-ordinator" for the development of ICT.
However, numerous positive elements have undoubtedly emerged from the experiment.
1) The deployment of technology on territories, which satisfies the objectives of the public authorities to bring rural territories into the information society.

In France, the prevailing view favours the digital development of the territory and the deployment of high-speed access. This is confirmed by the numerous policies implemented and aids provided at the level of the European Commission (Plan i2010), the State (Plan RE/SO 2007) and its calls for projects (DATAR), etc. This prescriptive approach helps to stimulate (while accentuating the feeling of exclusion in the territories concerned) the high-speed offering by the operators irrespective of their type. The LEN (2004) supports this movement;

2) the development of a market on these territories; are we not seeing the ADSL operators arrive after or during the experiment and capture users?

These high-speed projects can find themselves competing with and "abandoned" by the sudden arrival of standard technology on the territory. Indeed, numerous experiments were cancelled, moved or under-used, owing to the deployment of ADSL in the territories concerned;

3) an ongoing improvement in the technical offering by the operators (Wi-Fi or satellite) in interaction with the complex forms of demand (peer-to-peer or VoIP). Satellite technical solution are evolving (undoubtedly less quickly than terrestrial solutions), but the commercialisation of offerings using these technologies is to a large extent held back by the difficulty of creating a market that is significantly large market (all the non-covered areas in France) and sustainable (competition with ADSL)

Initially, the project did not provide for bandwidth management, but in a Wi-Fi network, the bandwidth is shared and heavy consumption by one or more network can considerably undermine the quality of the service offered. At the request of the co-ordinators of certain sites, an intelligent bandwidth manager was installed locally. This is based on the prioritisation of the protocols used: the last protocol allocated being peer-to-peer;

4) the wider territorial deployment of the technology beyond the first people taking part in the experiment (waiting lists) and its extension to neighbouring communes.

The experiment has produced in most cases, a surge in local interest in the project by non-participating citizens and residents of neighbouring communes. This interest is also boosted by word of mouth, especially as regards the advantageous conditions of the experiment (free, equipment, etc.);

5) improved anticipation by local authorities of the needs of their citizens not only as regards infrastructure needs (reflection on costs, project management, etc.), but also as regards new innovative services (combined with a real territorial strategy) and, finally, in the area of types of usage (taking account of training needs).

In general, these experiments have proved a way of meeting needs expressed and have also helped to mobilise and energise the stakeholders involved locally (public authorities, citizens wanting high-speed access, etc.). As a result, local authorities are considering providing financial support so as not to interrupt the existing connections to increasingly insistent and demanding citizens, at the end of the end of experiments. When one measures the financial effort made and envisaged by small local authorities, even when helped by public co-financing, it is clear that these experiments have satisfied to a large extent the expectations of the territorial stakeholders.

7. References

Bijker, W. E. (1995). Of Bicycles, Bakelites, and Bulbs Toward a Theory of Sociotechnical Change, MIT Press, Cambridge

Bolton P.; Harris C. (1999). Strategic Experimentation. *Econometrica*, vol.67, n°2, pp. 349-374

Grossmann, S.J.; Kihlstrom, R.E.; Mirman, L.J. (1977). A Bayesian Approach to the Production of Information and Learning by Doing, *Review of Economic Studies*, vol. 44, pp. 533-547

Leach, J.C.; Madhavan, A.N. (1993). Price Experimentation and Security Market Structure, *The Review of Financial Studies*, vol. 6, n°2, pp. 375-404

Orlikowski, W. (1992). The duality of technology: rethinking the concept of technology in organizations, *Organization Science*, vol 3, pp. 398-427

Swanson, E.; Ramiller, N.C. (1997). The Organizing Vision in Information System Innovation, *Organization Science*, vol 8, n°5, pp. 458-474

Yin, R.K. (2003). Case study research: Design and methods, 3rd edition, Applied social research methods series, Sage publications, vol. 5, London

6

Design and Implementation of Satellite-Based Networks and Services for Ubiquitous Access to Healthcare

Georgi Graschew, Theo A. Roelofs, Stefan Rakowsky and Peter M. Schlag
Surgical Research Unit OP 2000, Experimental and Clinical Research Centre ECRC, Max-Delbrück-Centre for Molecular Medicine and Charité – University Medicine Berlin, Charité Campus Berlin-Buch, Lindenberger Weg 80, D-13125 Berlin Germany

1. Introduction and Background

Telemedicine describes the use of Information and Communication Technologies (ICT) for the delivery of medical services. It aims at equal access to medical expertise irrespective of the geographical location of the person in need. New developments in ICT have enabled the transmission of medical images in sufficiently high quality that allows for a reliable diagnosis to be determined by the expert at the receiving site (Pande et al., 2003; Lacroix et al., 2002). Through Telemedicine patients can get access to medical expertise that may not be available at the patients' site. Networks for Telemedicine enable the integration of distributed medical competence and contribute to the improvement of the quality of medical care, to the cost-effective use of medical resources and to quick and reliable decisions.

For optimal performance of telemedical applications, the networks and communication tools used must be optimised for medical applications, both with respect to the Quality-of-Service (QoS, a set of parameters characterising the performance of the communication channel per se, such as transmission bandwidth, delay, jitter, data loss, etc.) as well as to the Class-of-Service (CoS; a set of terms specifying the medical services offered in the network, like Telesurgery, Telepathology, Telesonography, Tele-Teaching, -Training & -Education, etc.).

The use of specifically designed networks for telemedicine (distributed medical intelligence) contributes to the continuous improvement of patient care. Experience over the last decade has shown that the goals of Telemedicine are not automatically reached by the introduction and use of singular new technologies per se, but rather require the implementation of integral services.

At the same time, however, these innovative developments in ICT over the last decade bear the risk of creating and amplifying a digital divide in the world, creating a disparity in the quality of life, e.g. between the northern and the southern Euro-Mediterranean area (Graschew et al., 2003a; Dario et al., 2005; Graschew et al., 2004a). In recent years different projects have demonstrated how the digital divide is only one part of a more complex

problem: the need for integration. (Wootton et al., 2005; Rheuban & Sullivan, 2005; Graschew et al., 2003b; Graschew et al., 2002).

2. Methodology

During the last years OP 2000 (Operating Room of the future) has designed, developed and validated various modules for interactive telemedicine services (Schlag et al. 1999; Graschew et al. 2000). One of the key elements is the interactive telecommunication module WoTeSa / WinVicos: WoTeSa, a dedicated Workstation for Telemedical applications via Satellite that uses the communication software WinVicos (Wavelet-based interactive Video communication system).
WoTeSa is a PC with sufficient processing capacity (≥ 3 GHz Pentium IV, ≥ 512 MB RAM), one or more Osprey video capture boards (Osprey 100 or Osprey 500), a camera with composite and s-video outputs as live source (e. g. Canon VC-C4); a second camera as document camera for transmission of non-digital images; standard headset or microphone with small loudspeakers. The different video inputs of the Osprey video capture card can be used for direct connection to various medical video sources. WoTeSa thus serves quasi as a medical video hub. It is noteworthy that WoTeSa is a dedicated workstation and yet it can be realised with off-the-shelf components, thus making it readily available and at the same time very flexible and adaptive.
WinVicos is an all-software high-quality interactive video communication system, supplying real-time video, still-image and audio-transmission. WinVicos is especially designed for telemedical applications (e.g. telesurgery, teleradiology, telepathology) using a hybrid speed-optimised wavelet-codec that is based on the concepts of Partition, Aggregation and Conditional Coding (PACC; Patent DE 197 34 542 A1 from "Deutsche Telekom", Darmstadt, Germany). In contrast to most mainstream video coding systems that are mostly optimised for cinema and home entertainment, the PACC codec does not employ motion estimation but maximises the frame resolution to allow for maximal details to be visualised. WinVicos communicates IP-based and allows for online scaling of the transmission parameters (bit rate, frame rate and frame size from 128x96 up to 640x480 pixel). It supports both point-to-point and multipoint communication scenarios. Besides high quality live video transmission using moderate bandwidths (0.5–1 Mbit/s) it also allows for still-image transmission. WinVicos is very easy to use: there is a main user interface that is sufficient for the standard actions of the user. This includes calling the video conference partner via a telephone book, adjustment of both local and remote transmission parameters, as well as speaker and microphone volume control (see Fig. 1).
In both the video windows and the still-image windows WinVicos supports the use of common cursors shared by the conference partners. As WinVicos is an all-software system not only the implementation of other video codecs can be readily realised, but also a continuous performance improvement is supported to keep up with recent developments in the field. The most recent version of WinVicos also supports the transmission of video streams in full high definition (HD) resolution. Further readings on WoTeSa and WinVicos can be found in the literature (Graschew et al., 2001).
Other telemedicine systems are used e.g. for tele-ultrasound in rural areas where telementoring by live videoconferencing allowed to guide the ultrasound technician to record additional images of the patient (O'Neill et al., 2000) for clinical assessment of

pediatric burns which showed a good agreement between the face-to-face consultation and seeing the patient via videoconference (Smith et al., 2004) and for home telecare services likely to improve quality of health services (Guillen et al., 2002). Other systems are described in Sable 2002; Latifi et al., 2004, Eadie et al., 2003.

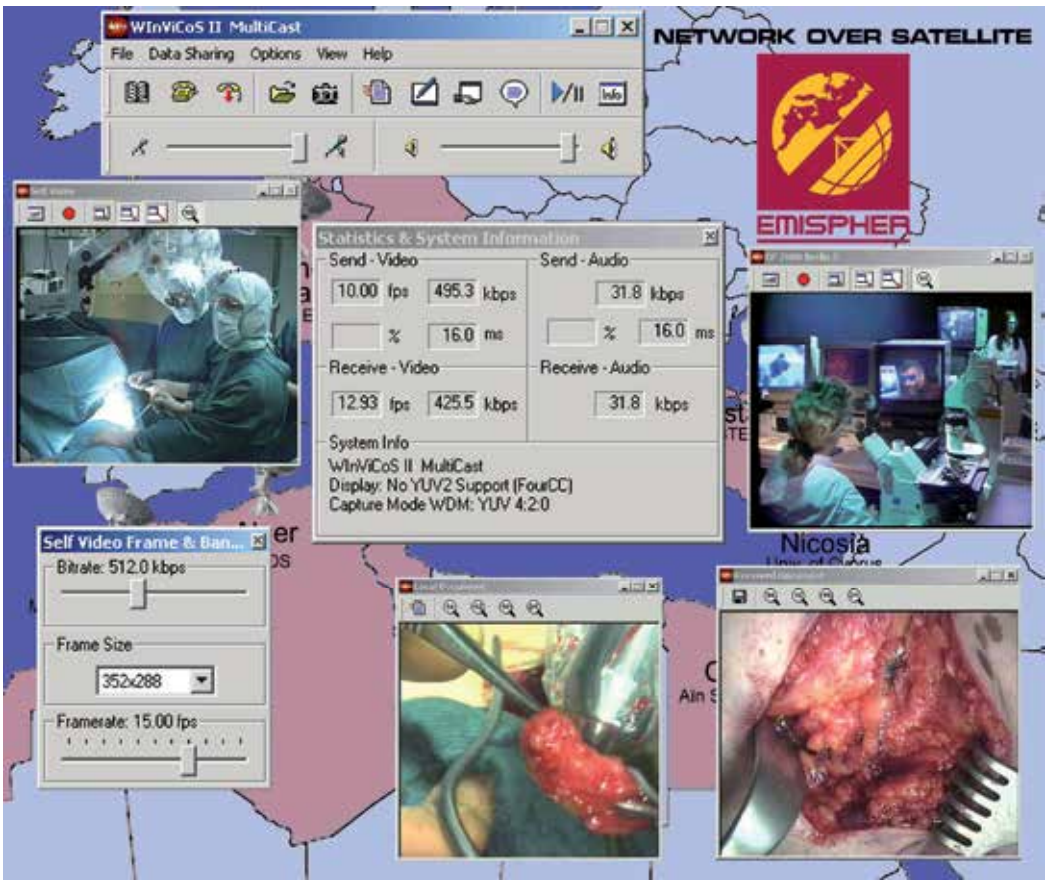


Fig. 1. WinVicos main user interface (top): two live video windows (top) and two still-image windows (bottom), online flexible adjustment of transmission parameters (bottom left), transmission performance can be monitored online (centre).

3. Satellite-based Networks and Services

OP 2000 has been a prominent partner in various telemedical networks using satellite-based technology for the deployment of services for interactive video communication and medical telepresence (Graschew et al., 2009). In the following, four projects will be described in which WoTeSa / WinVicos have been used as interactive tool for real-time telemedical application services and e-learning scenarios.

3.1 Generic Advance Low-cost trans-European Network Over Satellite - GALENOS

In cooperation with: Eutelsat (F), Nortel Dasa, (D), Alcatel Space Industries (F), Telespazio (I), MEDSAT (F), NCSR Demokritos (Gr); financially supported by the European Union (EU).

In the GALENOS project a European competence network via satellite, dedicated to telemedical applications, has been realised (Graschew et al., 2001). Owing to the participation of industrial partners in integrating the communication network and getting satellite transmission capacity available, several telemedical services (e.g. offline access to archived data, live consultation of experts, tele-teaching, etc.) have become available in a unified and low-cost technology. This possibility to get support from external experts, the improvement of the precision of the surgical treatment, and last but not least, online documentation and hence improved analysis of the available data on a patient, contributes to a continuous improvement in treatment and care of patients.

The trans-European communication network over satellite provides various collaborative services and applications, avoiding the use of specific technologies, thus promoting a wide availability. Furthermore, the network exhibits inter-operability with previously existing telecommunication networks. The network offers up to 2 Mbit/s interfaces with satellite link and covers a total of 14 clinics in Bulgaria, France, Germany, Greece, Italy and Tunisia. Access to the local area network (LAN) at each site is also provided for the exchange of patient data, medical images, videoconferencing, etc. Substantial reactivity and technical flexibility allow treatment protocols, adapted to the patient's pathology, in the shortest time.

The GALENOS network enables physicians not only to receive virtual "classical" hands-on medical training, but also trains them in using the state-of-the-art communication-, video- and computer-technologies needed for collaborative work in a network of distributed medical intelligence (interactive Teleteaching). Such distributed medical intelligence in the competence network enables e.g. local physicians or local hospitals that are confronted with unexpected test results to get online advice from the nearest academic hospital (intraoperative radiological Teleconsultation). It also becomes possible to perform intraoperative Telepathology where, after a biopsy has been taken and a slice has been prepared, this sample is put under a remotely-controllable, camera-equipped microscope and the diagnosis is then formulated by a remote expert. Or intraoperative Teleradiology, where anytime during the operation, the surgeon can access all the preoperative radiological data (like X-ray, CT, MRT, etc.) and discuss it with the radiologist in real-time.

3.2 Disaster Emergency Logistic Telemedicine Advanced Satellites System - DELTASS

The DELTASS Consortium consists of: Centre National D'Etudes Spatiales (F), MEDES (Institut de Médecine et de Physiologie Spatiales (F), SPACEBEL (B), Alcatel Space Industries (F), EADS – DORNIER (European Aeronautic Defence and Space Company (D), EADS – MS&I (European Aeronautic Defence and Space Company – Matra Systémes & Information (F), and Surgical Research Unit OP 2000 at Charité (D).

The DELTASS project was funded by the European Space Agency (ESA) as part of the ARTES-5 programme (Advanced Research in Telecommunications Systems).

Further general information on the DELTASS project can also be found at:
http://telecom.esa.int/telecom/www/object/index.cfm?fobjectid=750.

In the DELTASS-project (Graschew et al., 2008) a disaster scenario was analysed and an appropriate telecommunication system for effective rescue measures for the victims was set up and evaluated. Satellite-based systems are well suited for these circumstances, where generally ground infrastructures are partly or even totally destroyed. In such situations, even on a large geographic area or isolated area, space-based services can be easily and quickly deployed. DELTASS demonstrates operational performance of various services, covering the different aspects/phases of disaster emergency medicine. According to these phases, the DELTASS system is made up of the various corresponding subsystems (Fig. 2):

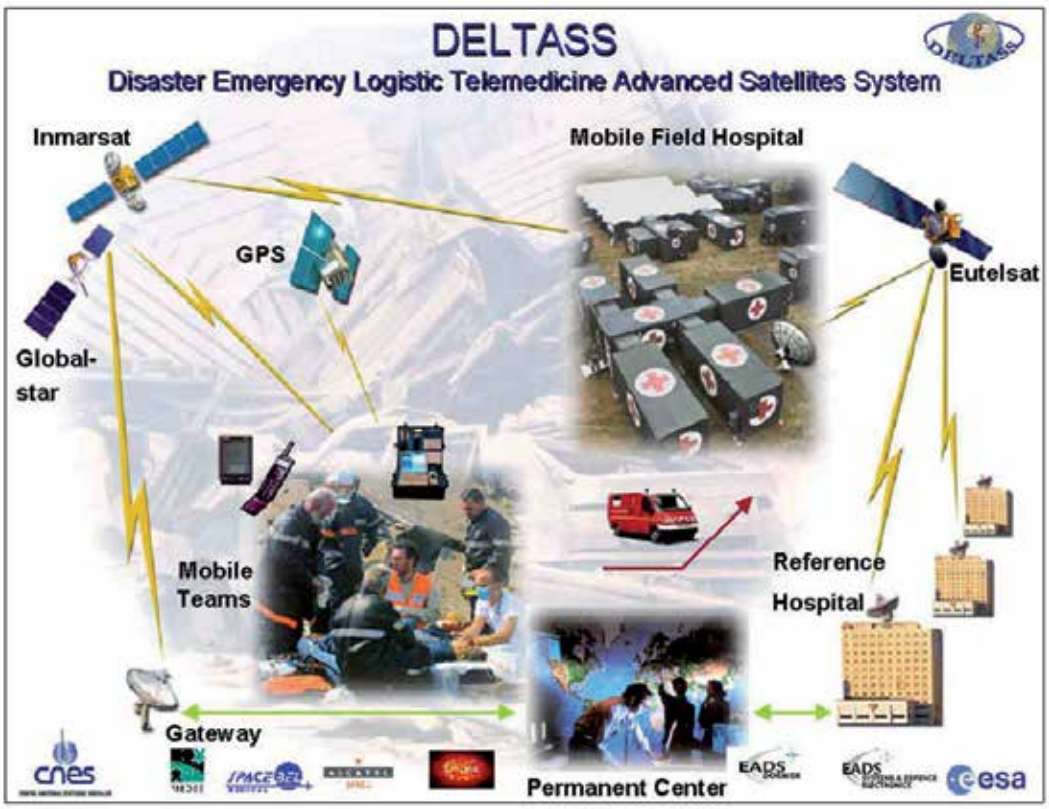


Fig. 2. DELTASS System Architecture: Mobile Teams, Permanent Centre (PC), Mobile Field Hospital (MFH) and Reference Hospital (RH) are interconnected via several satellite systems with different bandwidths. Additionally, terrestrial communication channels support the data exchange between the PC and the RH.

Mobile Teams

Mobile teams are deployed on the disaster site for search, identification, triage and evacuation of victims. They communicate with the coordination- and medical- teams located in Permanent Centre or Mobile Field Hospital via low-rate (Globalstar, 9.6 kbps) and medium-rate (Inmarsat, 64 kbps) satellite telecommunication systems. Their positions are tracked via the established Global Positioning System (GPS satellite system).

Permanent Centre
The Permanent Centre is located outside the disaster area. The Permanent Centre constitutes a new element in the architecture of support systems for disaster emergencies and is unique to the DELTASS system. In conventional set-ups the mobile teams at the disaster site are coordinated and supported by the staff of a Mobile Field Hospital deployed at or close to the disaster site. However, complete deployment of such a Mobile Field Hospital takes at least ~6 hours, usually ~12 hours, and consequently the activities of mobile teams in these first, highly critical hours, are ill-coordinated and far from optimal. To improve this bottleneck, DELTASS has a designated Permanent Centre that is in control of coordination and medical support to the mobile teams from time zero on. The Permanent Centre is equipped with terrestrial gateways to the Globalstar and Inmarsat satellite systems through which it receives all data from the mobile teams. It coordinates all actions of the mobile teams and manages all medical and logistic data, thus assuring efficient operation during the first critical phase. All data received at the Permanent Centre are processed, appropriate Reference Hospitals (RH; see below) are identified and the logistic and medical data are transferred to these RH via terrestrial telecommunication links.

Mobile Field Hospital (MFH)
A Mobile Field Hospital (MFH), which will be deployed at or close to the disaster site, provides all activities related to the co-ordination of the mobile teams on the disaster site, the victims' medical triage, reception, first aid treatment, conditioning for transportation, further medical expertise for some patients by teleconsultations between MFH and Reference Hospital(s).

Reference Hospital (RH)
The Reference Hospital(s) (RH), located outside of the disaster area, acts as an expert centre by providing telemedical services to the MFH using the high-bandwidth satellite link (VSAT, 2 Mbps). These services consist of off-line and on-line telediagnosis, access to external medical databases, as well as real-time interactive telemedical services such as live teleconsultations, live telesonography, intraoperative virtual reality simulation and interactive telemicrobiology (see Figs. 3 and 4).
Statistics show that in cases of disaster emergency medicine, approx. 40% more amputations are performed, as compared to normal situation. One of the aims of providing live second opinion by remote experts is to reduce this number of unneeded amputations, manipulations and subsequent complications substantially, by expert support during triage, diagnosis and medical treatment.
These interactive telemedical services between the MFH and the RH are realised using a dedicated WoTeSa (Workstation for Telemedical Applications via Satellite) with the communication software WinVicos (Wavelet-based interactive Video communication system). WoTeSa/WinVicos combines the user-friendliness and flexibility of IP-based communication protocols with the security and sufficiently-high quality of the live video transmission at a satellite bandwidth of only up to 2 Mbps. Medical experts at the RH support the medical treatments in the MFH and enable a quick and reliable decision concerning treatment and/or evacuation of the patient/victim. In this way, the quality of the provided medical service during and after disaster emergencies is strongly improved.

The performance of the DELTASS system has been shown during various full-size live demonstrations (http://telecom.esa.int/telecom/www/object/index.cfm?fobjectid=6324).

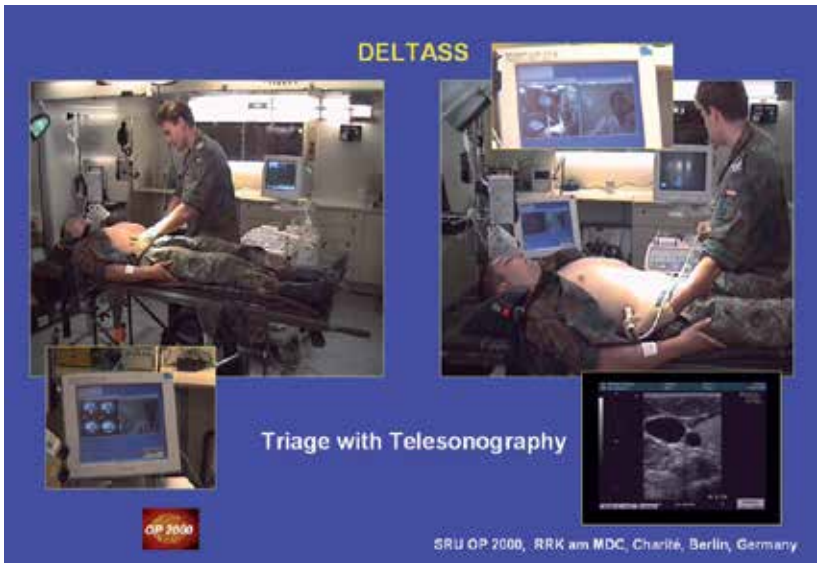


Fig. 3. Live Telesonography: The ultrasound video data stream (insert right bottom) is transmitted live via WoTeSa / WinVicos to the experts of the RH (insert right top). Additionally the signal of a room camera (e.g. for telementoring; insert right top) as well as further medical data (e.g. X-ray, CT, etc.; insert left bottom) can be transmitted live to the RH.

Fig. 4. Live Telesurgery of Inguinal Hernia: During the surgery the video data stream of a camera integrated in the centre of the surgical light can be transmitted live to the experts in the RH using WoTeSa / WinVicos (insert left bottom). In this way these experts receive a live video of the situs of surgery (inserts right bottom) and can advise the colleague in the MFH.

3.3 Medical Assistance for Ships - MEDASHIP

Cooperating partners: D'Appolonia (I), AVIENDA (UK), Eutelsat (F), NCSR Demokritos (GR); co-funded by the European Union (EU) under the eTEN programme.

The main objective of the service developed by the MEDASHIP project is to supply integrated solutions for medical consultations on-board of ships (Graschew et al., 2004b). The satellite-based telemedicine services address both passenger ships and merchant vessels and are intended to provide passengers and crew members with an effective medical assistance in cases of emergency and in all those cases where the on board medical staff requires second opinion. During the validation phase the service was tested on board of three ships with the possibility to have it connected to three land medical centres. (Fig. 5).

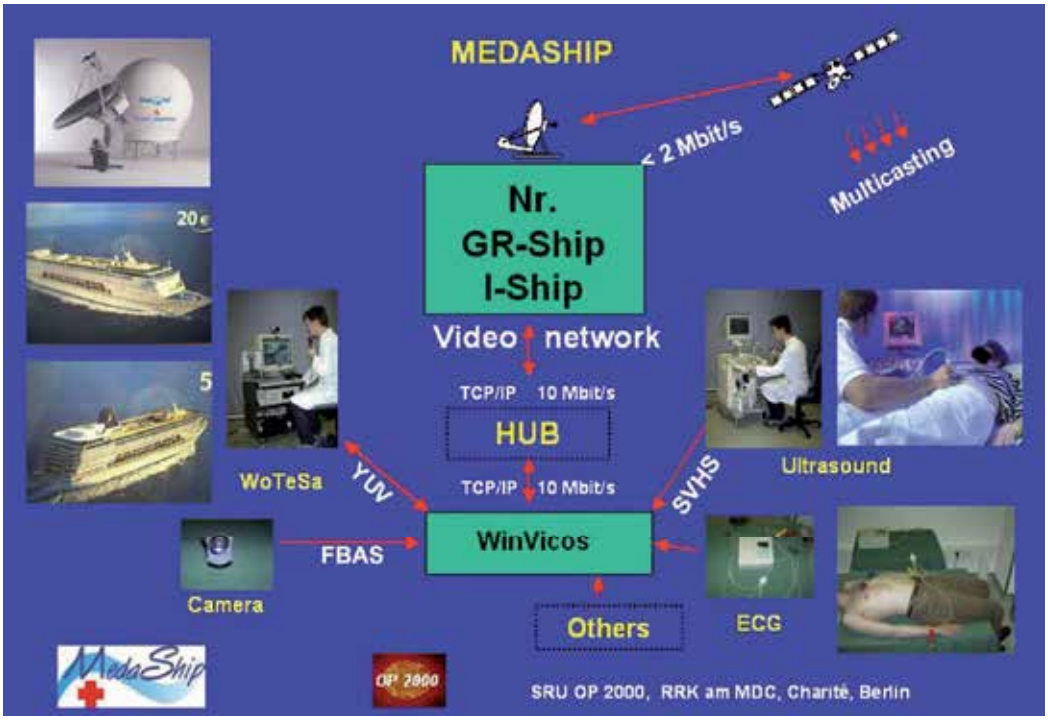


Fig. 5. Video network on-board of the ships: Video camera, ultrasound and electrocardiography equipment connected with the satellite modem via WoTeSa/WinVicos.

In addition to the standard medical equipment aboard the ships, two video cameras, an electrocardiograph (ECG) and an ultrasound (US) equipment are used. With this equipment the following telemedical services have been realised using satellite transmission at a bandwidth of 512 kbps up to 1 Mbps offering the required high quality of images and video transmission:

Teleconsultation

The live camera on-board of the ship can be used to transmit the image of the doctor who is leading the examination on-board of the ship or the image of the patient when being questioned by the land-based expert. It can also be used to show the land-based expert an

injured part of the patient's body which he needs to see for his consultation. Thus a very realistic and effective live communication is possible.

Electrocardiography
The ECG system is connected to WoTeSa on board the ship and can be controlled by the physician from this workstation. Via application sharing software also the expert can control the ECG system from the land-based workstation. The main menu that includes all the functions of the ECG as well as the patient's ECG is transmitted to the expert. Thus the expert and the physician on board can jointly acquire and analyse the ECG report.

Telesonography
The S-video output of the US equipment is directly connected to the Osprey video capture board. Satellite transmission tests have shown that not only still images can be transferred but also live ultrasound investigations can be transmitted at 500-700 kbps (see Fig. 6).
With a document camera analogous patient data can be captured and digitised by WinVicos as a document. For example X-ray or CT-images can be captured from an illumination board and displayed locally and transmitted using this document camera function.

Data Security
Respect for patient confidentiality is a clear requirement in the MEDASHIP system and the level of privacy protection has been addressed throughout the MEDASHIP project. In MEDASHIP, from the telemedicine platform viewpoint, the following requirements should be taken into account to prevent abuse and to safeguard the privacy and confidentiality of managed clinical and personal information of the patients.

- *Authentication:* Monitoring and verifying all the accesses to the information. A control over the users should be carried out at each access to the system, by utilising the users' credentials (username and password) in order to verify that the user is who he/she claims to be.
- *Encryption:* Scrambling a sender's transmission according to an algorithm that the recipient then uses to unscramble and decipher the transmission. All the process is not visible to the user, the client software knows all it needs.
- *Access control:* Authorising access to specific and clearly identified resources to certain users based on their company responsibilities and the security classification of the resources.
- *Integrity:* Developing or utilising applications and data management software that is secure from unauthorised modification of their code.
- *Confidentiality:* Developing or utilising applications and data management software that is secure from disclosure to unauthorised persons or programs.
- *Auditing and accounting system:* Monitoring the system and maintaining records of system and user activity.
- *Security policy:* Establishing clear security policies with customers and end-users.

Unlike open network satellite equipment technologies Linkway uses a proprietary acquisition and synchronisation technique making signal interception and decoding virtually impossible. Data Security will also be assured by the coding algorithm of the WinVicos software. The transmitted data can only be decoded by this software. As one

member owns the software licence this provides an initial level of data security through the licence distribution.

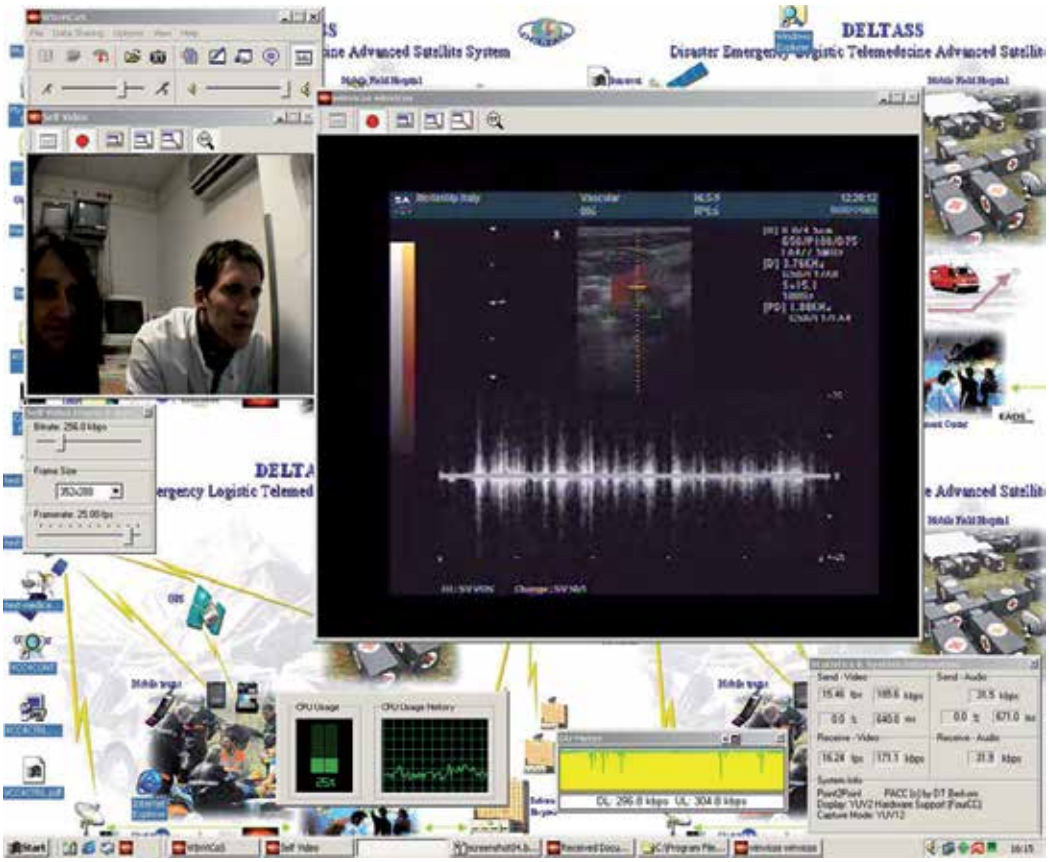


Fig. 6. Telesonography: the live signals of the ultrasound equipment on-board of the ship are transmitted to the reference hospital. A physician at Charité is consulting the ultrasound examination of a patient on board of the cruise ship.

Reduction of cost

The costs for emergency interventions for removing a passenger from the ship and hospitalisation abroad are not to be undervalued. The removal of a passenger in the Carribean can cost up to $ 11.000 and the cost for hospitalisation can range from 500-1000 € per day. Consequently, market trends force passenger shipping lines to offer services that help to improve the response to on-board clinical emergencies, improve the customer satisfaction and the companies' image.

3.4 Euro-Mediterranean Internet-Satellite Platform for Health, medical Education and Research - EMISPHER

In cooperation with: FMPC - Faculty of Medicine and Pharmacy of Casablanca, Morocco; ANDS - Agence National de Documentation de la Santé (Ministère de la Santé), Algiers, Algeria; Tunis - Faculty of Medicine of Tunis, Tunisia; ASU - Aïn Shams University, Cairo,

Egypt; NIFRT - Nasser Institute for Research and Treatment (Ministry of Health and Population, MOHP), Cairo, Egypt; UCY – University of Cyprus, Nicosia, Cyprus; ISTEM - Continuing Medical Education and Research Centre, University of Istanbul, Turkey; NCSR Demokritos, Athens, Greece; IsMeTT - Istituto Mediterraneo per i Trapianti e Terapie ad Alta Specializzazione, Palermo, Italy; CICE - Centre International de Chirurgie Endoscopique, Clermont-Ferrand, France; and Charité University Medicine Berlin, Germany; co-funded by the European Union (EU) under the EUMEDIS / MEDA programme.

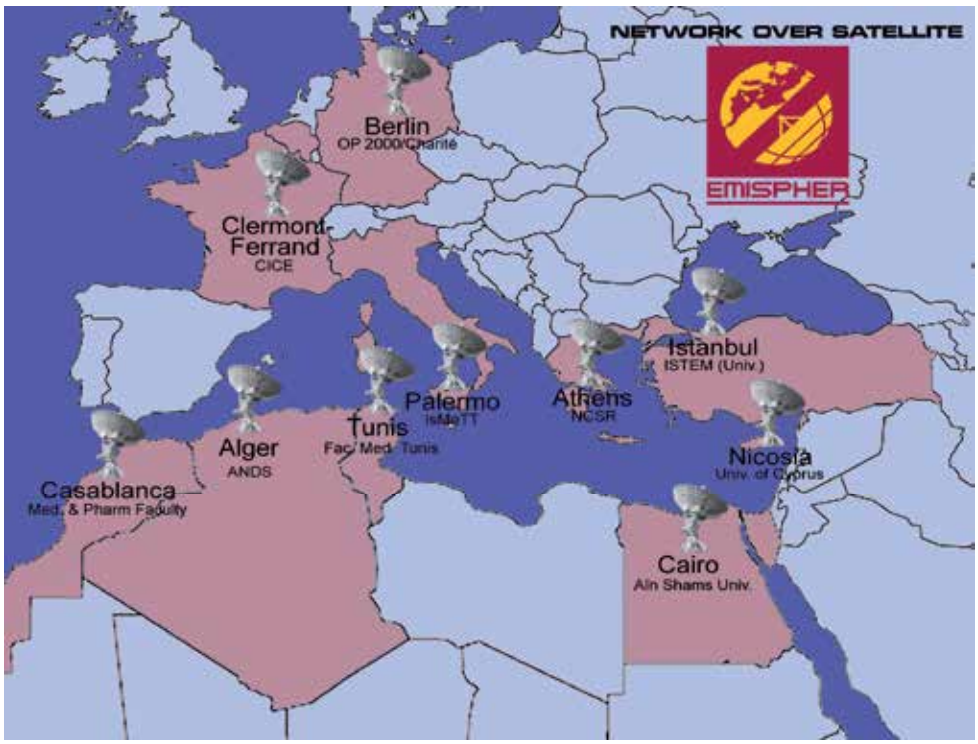


Fig. 7. Medical Centres in the EMISPHER Network

EMISPHER is dedicated to establish an equal access for most of the countries of the Euro-Mediterranean area to real-time and on-line services for healthcare in the required quality of service (see www.emispher.org). In the project an integrated Internet-Satellite platform has been set up on which three main areas of work have been realised: Virtual Medical University, Real-Time Telemedicine, and Medical Assistance (Graschew et al., 2005). The platform includes a bi-directional satellite network (up to 2 Mbps) between 10 Centres of Excellence in the Euro-Mediterranean region (Morocco, Algeria, Tunisia, Egypt, Cyprus, Turkey, Greece, Italy, France and Germany; see Fig. 7). For dissemination of the achieved results and for maximising its impact, EMISPHER has organised international conferences at each of the Mediterranean partner sites.

The EMISPHER Virtual Medical University

The formation and operation of the EMISPHER Virtual Medical University (EVMU) for e-learning (teleteaching) is one of the main efforts in the project. The EVMU uses real-time broadcast of lectures, live surgical operations and pre-recorded video sequences etc., as well as web-based e-learning applications. The target population of the EVMU is comprised of medical students (both undergraduate and postgraduate) hospital staff, general practitioners and specialists, health officers and citizens.

Each of the leading medical centres provides didactical material and modules for synchronous and asynchronous e-learning in their medical specialties. The central gateway to EVMU is the project's website: www.emispher.org.

Some of the multimedia teaching material needs to be presented in real-time. Live transmission of surgical operations from operating theatres, lectures, etc. from one site to one or several sites simultaneously (point-to-point or multipoint) are possible in the network between the 10 partners (Fig. 8).

Fig. 8. Interactive multipoint e-learning session with transmission of live ultrasound video data from Charité (Berlin) to Ain Shams University (Cairo), Agence Nationale de Sante (Algiers), Faculté de Medecine et de Pharmacie (Casablanca) and University of Cyprus (Nicosia).

Real-Time Telemedicine

EMISPHER has set up a satellite-based network using the combined WoTeSa and WinVicos modules for real-time telemedicine. In the field of real-time telemedicine the following

categories of applications are offered: second opinion, teleteaching & teletraining (demonstration and spread of new techniques), telementoring (enhancement of staff qualification), and undergraduate teaching courses and optimisation of the learning curve. The leading medical centres in the project provide expertise in the following medical fields: open and minimally-invasive surgery, multi-organ transplantation, endoscopy, pathology, radiology, interventional imaging, neurology, infectious diseases, oncology, gynaecology and obstetrics, reproductive medicine, etc. These real-time telemedical applications contribute to improved quality of patient care and to accelerated qualification of medical doctors in their respective specialty. The main target audience are specialist doctors (Fig. 9).

Fig. 9. Interactive multipoint radiological teleconsultation in EMISPHER. Experts from different centres interactively discuss radiology results from a particular patient case.

Medical Assistance

The third field of service operated in EMISPHER is medical assistance. As tourism constitutes a substantial economical factor in the Mediterranean region and because of the increasing mobility of the population, continuity of care through improved medical assistance is of major importance for improved healthcare in the Euro-Mediterranean region. Introduction of standardised procedures, integration of the platform with the various local communication systems and training of the medical and non-medical staff

involved in the medical assistance chain allow for shared management of files related to medical assistance (medical images, diagnosis, workflow, financial management, etc.) and thus for improved care for travellers and expatriates.

User Evaluation
Two types of evaluation were performed for the EMISPHER satellite-internet platform and the services delivered from the platform. Firstly, the evaluation of all technical aspects of the platform and connecting services (WinVicos and MEDSKY) and secondly the evaluation of the operational elements of the system meeting the medical user requirements.
Overall point-to-point connections were successful with normal delays (~650 ms). Realised maximum bandwidths for symmetric point-to-point connections are 512-640 kbps in each direction. Up to these bandwidths the video and audio streaming is smooth. When going beyond these settings frozen images and audio problems start occurring. For asymmetric video bandwidth allocation (for example for e-learning applications) the maximum bandwidth settings for good operation are 768 kbps in one direction and 192 kbps in the other direction. For multipoint connections between three sites the maximum bandwidth settings for good operation are 384, 384 and 512 kbps. If the results presented above are compared with the nominal bandwidths required for the various services, it is clear that the EMISPHER platform provides sufficient bandwidth for most of the telemedicine applications with smooth video and audio streams in sufficient resolution.
During the first months of full operation of the network continuous evaluations have been performed and user-feedback was collected. This resulted in the upgrade of the WinVicos telemedical communication software as well as various upgrades of the MEDSKY software platform. The valuable user feedbacks were taken into account and have lead to even improved functionalities of the system and thus improved quality of medical work.
This demonstrates that the satellite platform with WoTeSa/WinVicos telemedical stations is very suited for the envisioned variety of medical services and is easy to learn for a wide range of medical professionals from the various countries.

4. Barriers and their possible solutions

Although the market is promising and the technology ready (or nearly) to be used, the take-up and commercialisation of the telemedicine services is still uncertain due to a number of barriers:
External

- Decision-making in health is fragmented with respect to procurement policies, which hinders progression towards real integration.
- Establishing and building confidence with physicians takes time.
- It still seems that it will be a further 1-2 years before full deployment and therefore a critical mass can be built.
- How the services are packaged and attitudinal changes from within the health sector are also necessary.
- There are numerous organisational issues.

Economic

- It is difficult to predict the breakeven level of services and therefore to determine the minimum level of investment required for breakeven of the services.

- There is a lack of investment in health and therefore significant investment is still required.
- Considerable investment in improved management, training and education of personnel, re-designing of care and logistic processes is necessary.

Legal

- Ethical issues
- Regulatory framework issues
- Reimbursement issues
- Lack of standardisation

Technical

- Software development is still evolving.
- IP levels of connectivity are also required as a minimum.
- All the services need to be fully integrated into one platform.

In various pilot projects, the technology has been put in place, the necessary applications have been developed and it had been proved that it can be used successfully and meet the needs of its end users. Yet numerous trials and demonstrations carried out during the projects have also highlighted a certain number of issues that could potentially hinder the commercialisation process of e-services in the medical sector.
In fact, various market analysis reports prepared for the European Commission also insist on the fact that the telemedicine market growth will be dependent upon a number of vital conditions and enabling factors already noted above. It is generally accepted that four key action lines should be initiated in parallel: consolidation on the supply side, technical integration, investment on the demand side and "accompanying measures" that could be the enabling factors required to allow the health telematics market to achieve substantial and exponential growth.

Suggested solutions to overcome barriers
The solutions to such 'barriers' to commercialisation do not lie within the scope of the technology and appear to be generic to those 'up and running' tele-health activities worldwide. It is only in North America that new legislation has been introduced to respond to the particular needs of this practice and this legislation are not yet being applied on a federal level but rather on a state by state basis. However, the organisational and cultural aspects that must accompany any new form of practice need the input of other actors both on the governmental, legal and political scale. Tele-health is not a simple extension of current health systems and cannot be perceived as such.
Solutions to the "barriers" can be consolidated under the following three main categories:

- Awareness of telemedicine and tele-health as integral part of medical practice;
- The need for common standards and policies;
- The need for specific legislation.

5. Conclusions and Perspectives: Virtual Hospitals

Satellite-based networks and services are crucial to support ubiquitous access to healthcare. Appropriate enabling technologies must be deployed and interconnected via networks with

an appropriate design. In this chapter we have presented WoTeSa/WinVicos as a flexible high-end module for real-time interactive telemedical services. Besides video communication in medically expedient quality, the provision of interactivity for the remote control of medical equipment is indispensable. Both video communication and interactivity require a (nearly) real-time mode of bi-directional interactions. Various examples have been given of particular networks and services that have been deployed, each to support medical telepresence in specific functional scenarios (GALENOS, DELTASS MEDASHIP and EMISPHER).

However, despite substantial improvements that have been realised, these developments bear the risk of creating and amplifying digital divides in the world. To avoid and counteract this risk and to fulfill the promise of Telemedicine, namely ubiquitous access to high-level healthcare for everyone, anytime, anywhere (so-called ubiquitous Healthcare or u-Health) a real integration of both the various platforms (providing the "Quality-of-Service", QoS) and the various services (providing the "Class-of-Service", CoS) is required (Graschew et al., 2002; Graschew et al., 2003b; Wootton et al., 2005; Rheuban & Sullivan 2005; Graschew et al., 2006a). A virtual combination of applications serves as the basic concept for the virtualisation of hospitals. Virtualisation of hospitals supports the creation of ubiquitous organisations for healthcare, which amplify the attributes of physical organisations by extending its power and reach. Instead of people having to come to the physical hospital for information and services the virtual hospital comes to them whenever they need it. The creation of Virtual Hospitals (VH) can bring us closer to the ultimate target of u-Health (Graschew et al., 2006b).

The methodologies of VH should be medical-needs-driven, rather than technology-driven. Moreover, they should also supply new management tools for virtual medical communities (e.g. to support trust-building in virtual communities). VH provide a modular architecture for integration of different telemedical solutions in one platform (see Fig. 10).

Due to the distributed character of VH, data, computing resources as well as the need for these are distributed over many sites in the Virtual Hospital. Therefore, Grid infrastructures and services become useful for successful deployment of services like acquisition and processing of medical images (3D patient models), data storage, archiving and retrieval, as well as data mining, especially for evidence-based medicine (Graschew et al., 2006c).

The possibility to get support from external experts, the improvement of the precision of the medical treatment by means of a real medical telepresence, as well as online documentation and hence improved analysis of the available data of a patient, all contribute to an improvement in treatment and care of patients in all circumstances, thus supporting our progress from e-Health and Telemedicine towards real u-Health.

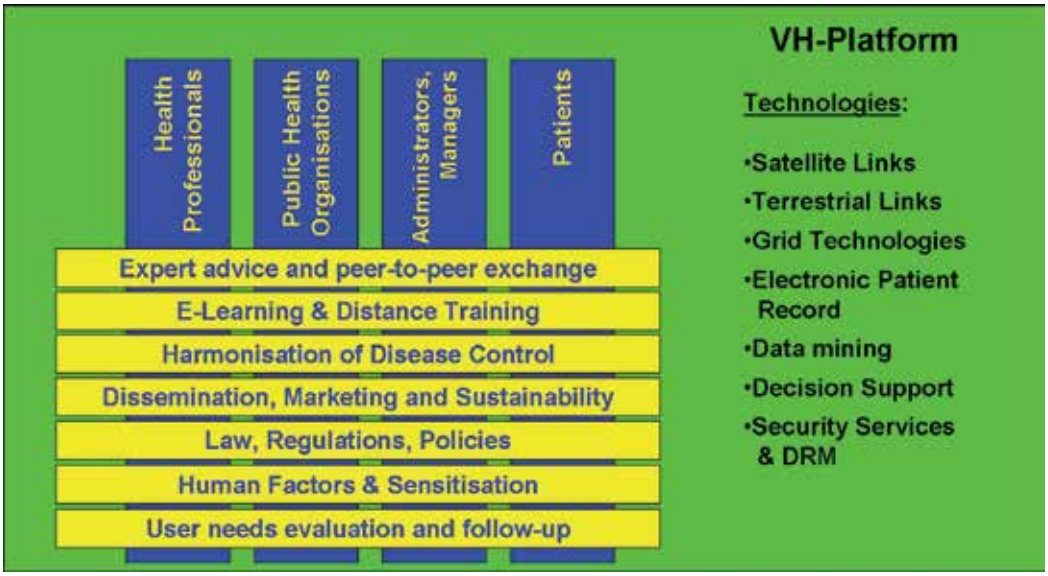


Fig. 10. Concept for the functional organisation of Virtual Hospitals (VH): The technologies of VH (providing the "Quality-of-Service", QoS) like satellite-terrestrial links, Grid technologies, etc. will be implemented as a transparent layer, so that the various user groups can access a variety of services (providing the "Class-of-Service", CoS) such as expert advice, e-learning, etc. on top of it, not bothering with the technological details and constraints.

6. References

Dario, C. et al. (2005). Opportunities and Challenges of eHealth and Telemedicine via Satellite. *Eur J. Med. Res., Vol. 10, Suppl I, Proceedings of ESRIN-Symposium, July 5, 2004, Frascati, Italy, (2005),* pp. 1-52.

Eadie, L.H. et al., (2003). Telemedicine in surgery. *Br. J. Surg.,* Vol. 90, pp. 647-58.

Graschew, G. et al. (2000). Interactive telemedicine in the operating theatre of the future. *J Telemedicine and Telecare* Vol. 6, Suppl. 2, pp. 20-24.

Graschew, G. et al. (2001). GALENOS as interactive telemedical network via satellite, In: *Optical Network Design and Management*, Proc. of SPIE, Vol 4584, pp. 202-205.

Graschew, G. et al. (2002). Broadband Networks for Interactive Telemedical Applications, *APOC 2002, Applications of Broadband Optical and Wireless Networks, Shanghai 16.-17.10.2002, Proceedings of SPIE,* Vol. 4912, pp. 1-6.

Graschew, G. et al. (2003a). Telemedicine as a Bridge to Avoid the Digital Divide World, *8. Fortbildungsveranstaltung und Arbeitstagung Telemed 2003, Berlin, 7.-8. November 2003, Tagungsband*, pp. 122-127.

Graschew, G. et al. (2003b). Telepresence over Satellite, *Proceedings of the 17th International Congress Computer Assisted Radiology and Surgery, London, 25.-28.6.2003,* International Congress Series, Vol. 1256, ed. H.U. Lemke et al., pp. 273-278.

Graschew, G. et al. (2004a). Interactive Telemedicine as a Tool to Avoid a Digital Divide of the World, In: *Medical Care and Compunetics 1*, L. Bos (Ed.), pp. 150-156, IOS Press, Amsterdam.

Graschew, G. et al., (2004b). MEDASHIP – Medizinische Assistenz an Bord von Schiffen, In: *Telemedizinführer Deutschland, ed. 2004,* A. Jäckel (Ed.), Deutsches Medizin Forum, Ober-Mörlen, Germany, pp. 45-50.

Graschew, G. et al., (2005). Überbrückung der digitalen Teilung in der Euro-Mediterranen Gesundheitsversorgung – das EMISPHER-Projekt, In: *Telemedizinführer Deutschland, ed. 2005,* A. Jäckel (Ed.), Ober-Mörlen, Germany, pp. 231-236.

Graschew, G. et al., (2006a). VEMH – Virtual Euro-Mediterranean Hospital für Evidenz-basierte Medizin in der Euro-Mediterranen Region, In: *Telemedizinführer Deutschland, Ausgabe 2006,* A. Jäckel (Ed.), Medizin Forum AG, Bad Nauheim, Germany, pp. 233-236.

Graschew, G. et al., (2006b). New Trends in the Virtualization of Hospitals – Tools for Global e-Health, In: *Medical and Care Compunetics 3*, L. Bos et al. (Eds.) Proceedings of ICMCC 2006, The Hague, 7-9 June 2006, IOS Press, Amsterdam, pp.168-175.

Graschew, G. et al., (2006c). Virtual Hospital and Digital Medicine – Why is the GRID needed?, In: *Challenges and Opportunities of HealthGrids,* V. Hernandez et al. (Eds.) Proceedings of HealthGrid 2006, Valencia, 7-9 June 2006, IOS Press, Amsterdam, pp.295-304.

Graschew, G. et al., (2008). DELTASS – Disaster Emergency Logistic Telemedicine Advanced Satellites System - Telemedical Services for Disaster Emergencies. *International Journal of Risk Assessment and Management* Vol. 9, pp. 351-366.

Graschew, G. et al., (2009). *New developments in network design for telemedicine.* Hospital IT Europe, Vol. 2 No. 2, pp. 15-18.

Guillen, S. et al., (2002). User satisfaction with home telecare based on broadband communication. *J. Telemed. Telecare,* Vol. 8, pp. 81-90.

Lacroix, L. et al., (2002). International concerted action on collaboration in telemedicine: recommendations of the G-8 Global Healthcare Applications Subproject-4. *Telemed. J. E-Health*, Vol. 8, pp. 149-157.

Latifi, R. et al., (2004). Telepresence and telemedicine in trauma and emergency care management. *Stud. Health Technol. Inform.*, Vol. 104, pp. 193-199.

O'Neill, S.K. et al., (2000). The design and implementation of an off-the-shelf, standards-based tele-ultrasound system. *J. Telemed. Telecare,* Vol. 6, suppl 2, pp. 52-53.

Pande, R.U. et al., (2003). The telecommunication revolution in the medical field: present applications and future perspective. *Curr. Surg.,* Vol. 60, pp. 636-640.

Rheuban, K.S. & Sullivan, E. (2005). The University of Virginia Telemedicine Program: traversing barriers beyond geography. *J. Long-Term Eff. Med. Implants,* Vol. 15, pp. 49-56.

Sable, C. (2002). Digital echocardiography and telemedicine applications in pediatric cardiology. *Pediatr-Cardiol.* Vol. 23, pp. 358-369.

Schlag, P.M. et al., (1999). Telemedicine – The New Must for Surgery. *Archives of Surgery* Vol. 134, pp. 1216-1221.

Smith, A.C. et al., (2004). Diagnostic accuracy of and patient satisfaction with telemedicine for the follow-up of paediatric burns patients. *J. Telemed. Telecare,* Vol. 10, pp. 193-198.

Wootton, R. et al., (2005). E-health and the Universitas 21 organization: 2. Telemedicine and underserved populations. *J. Telemed. Telecare*, Vol. 11, pp. 221-224.

7

Characterisation and Channel Modelling for Satellite Communication Systems

Asad Mehmood and Abbas Mohammed
Blekinge Institute of Technology
Sweden

1. Introduction

The high quality of service, low cost and high spectral efficiency are of particular interest for wireless communication systems. Fundamental to these features has been much enhanced understanding of radio propagation channels for wireless communication systems. In order to provide global coverage of broadband multimedia and internet-based services with a high signal quality to diverse users, seamless integration of terrestrial and satellites networks are expected to play a vital role in the upcoming era of mobile communications. The diverse nature of propagation environments has great impact on the design, real-time operation and performance assessment of highly reconfigurable hybrid (satellite-terrestrial) radio systems providing voice, text and multimedia services operating at radio frequencies ranging from 100 MHz to 100 GHz and optical frequencies. Therefore, a perfect knowledge and modelling of the propagation channel is necessary for the performance assessment of these systems. The frame work for most of the recent developments in satellite communications includes satellite land mobile and fixed communications, satellite navigation and earth observation systems and the sate-of-art propagation models and evaluation tools for these systems.

The organization of the chapter is as follows: Section 2 describes the multipath propagation impairments in land mobile satellite (LMS) communications. In Section 3, the probability distributions that characterize different impairments on radio waves are discussed. Section 4 provides an overview of statistical channel models including single-state, multi-state and frequency selective channel models for LMS communications. The chapter ends with concluding remarks.

2. Propagation Impairments Effecting Satellite Communication Links

The use of satellite communication systems for modern broadband wireless services involves propagation environments for radio signals different from that in conventional terrestrial radio systems. The radio waves propagating between a satellite and an earth station experience different kinds of propagation impairments: the effects of the ionosphere, the troposphere and the local fading effects as shown in Fig. 1. The combined effect of these

impairments on a satellite-earth link can cause random fluctuations in amplitude, phase, angles of arrivals, de-polarization of electromagnetic waves and shadowing which result in degradation of the signal quality and increase in the error rates of the communication links.

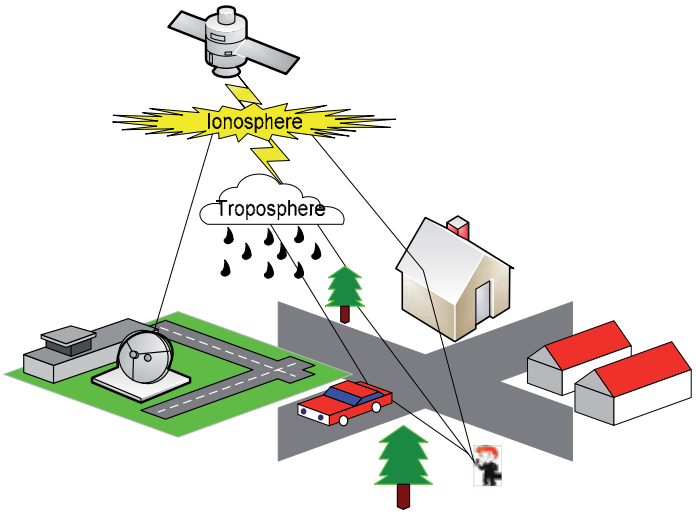


Fig. 1. The land-Mobile-Satellite Communication System

2.1 Ionospheric Effects

The effects of the ionosphere (an ionized section of the space extending from a height of 30 km to 1000 km) have adverse impact on the performance of earth-satellite radio propagation links. These effects cause various impairments phenomena such as scintillation, polarization rotation, refraction, group delays and dispersion etc, on the radio signals. The scintillation and polarization rotation effects are of foremost concern for satellite communications.

Ionospheric scintillations are variations in the amplitude level, phase and angle of arrival of the received radio waves. They are caused by the small irregularities in the refractive index of the atmosphere owing to rapid variations in the local electron density. The main effect of scintillation is fading that strongly depends on the irregularities or inhomogeneities of the ionosphere (Ratcliffe, 1973; Blaunstein, 1995; Saunders & Zavala, 2007). Scintillation effects are significant in two zones: at high altitudes (E and F layers of ionosphere) and the other is $\pm 20^{o}$ around the earth's magnetic equator. The effects of scintillation decrease with increase in operating frequency. It has been observed in various studies that at the operating frequency of 4 GHz ionospheric scintillations can result in fades of several dBs and duration between 1 to 10 seconds. The details about ionospheric scintillation can be found in International Telecommunication Union Recommendations (ITU-R, 2009a).

The orthogonal polarization (linear or circular) is used in satellite communication systems to increase the spectral efficiency without increasing the bandwidth requirements. This technique, however, has limitations due to depolarization of electromagnetic waves propagating through the atmosphere. When linearly polarized waves pass through the

ionosphere, the free electrons present in the ionosphere due to ionization interact with these waves under the influence of the earth's magnetic field in a similar way as the magnetic field of a motor acts on a current carrying conductor. This results in rotation of the plane of polarization of electromagnetic waves, recognized as Faraday rotation. The magnitude of Faraday rotation is proportional to the length of the path through the ionosphere, the geomagnetic field strength and the electron density, and inversely proportional to the square of the operating frequency. The polarization rotation is significant for small percentages of time at frequencies 1 GHz or less. The effect of Faraday rotation is much lower at higher frequencies even in the regions of strong ionospheric impairments and low elevation angles, e.g., at frequency of 10 GHz, Faraday rotation remains below 1^o and can be ignored (ITU, 2002). Cross-polarization can also be caused by the antenna systems at each side of the link. The effects of depolarization are investigated by two methods: cross-polarization discrimination (XPD) and polarization isolation. The details can be found in (Roddy, 2006; Saunders & Zavala, 2007).

2.2 Tropospheric Effects

The troposphere is the non-ionized lower portion of the earth's atmosphere covering altitudes from the ground surface up to a height of about 15 km of the atmosphere. The impairments of this region on radio propagations include hydrometeors, e.g., clouds, rain, snow, fog as well as moisture in atmosphere, gradient of temperature and sporadic structures of wind streams both in horizontal and vertical directions. The effects imparted by these impairments on radio signals are rain attenuation, depolarization, scintillation, refraction, absorption, etc. The radio waves are degraded by these effects to varying degrees as a function of geographic location, frequency and elevation angle with specific characteristics. The tropospheric effects in LMS communication links become significant when the operating frequency is greater than 1 GHz.

One of the major causes of attenuation for LMS communication links operating at frequency bands greater than 10 GHz (e.g., Ku-Band) is rain on the transmission paths in tropospheric region. The rain attenuation in the received signal amplitude is due to absorption and scattering of the radio waves energy by raindrops. The attenuation is measured as a function of rainfall rate and increases with increase in the operating frequency, rainfall rate and low elevation angles (Ippolito, 2008). The rainfall rate is the rate at which rain would accrue in a rain gauge placed in a specific region on the ground (e.g., at base station). The procedure to calculate attenuation statistics due to rainfall along a satellite-earth link for frequencies up to 30 GHz consists of estimating the attenuation that exceeds 0.001% of the time from the rainfall rate that exceeds at the same percentage of time and has been detailed in ITU-R recommendations (ITU-R, 2007).

The LMS channel utilization can be augmented without increasing the transmission bandwidth by the use of orthogonally polarized transmissions (linear or circular). The polarization of radio waves can be altered by raindrops or ice particles in the transmission path in such a way that power is transferred from the desired component to the undesired component, resulting in interference between two orthogonally polarized channels. The shape of small raindrops is spherical due to surface tension forces, but large raindrops adopt shape of spheroids (having flat base) produced by aerodynamic forces acting in upward

direction on the raindrops. When a linearly polarized wave passes through raindrops of non-spherical structure, the vertical component of radio wave parallel to minor axis of raindrops experiences less attenuation than that the horizontal component. As a result, there will be a difference in the amount of attenuation and phase shift experienced by each of the wave components. These differences cause depolarization of radio waves in the LMS links and are illustrated as differential attenuation and differential phase shift. Rain and ice depolarization have significant impacts on satellite-earth radio links for frequency bands above 12 GHz, especially for systems employing independent dual orthogonally polarized channels in the same frequency band in order to increase the capacity. The method of predicting the long-term depolarization statistics has been described in ITU-R recommendations (ITU-R, 2007).

A radio wave propagating through satellite-earth communication link will experience reduction in the received signal's amplitude level due to attenuation by different gases (oxygen, nitrogen, hydrogen, etc.) present in the atmosphere. The amount of fading due to gases is characterized mainly by altitude above sea level, frequency, temperature, pressure and water vapour concentration. The principal cause of signal attenuation due to atmospheric gases is molecular absorption. The absorption of radio waves occurs due to conversion of radio wave energy to thermal energy at some specific resonant frequency of the particles (quantum-level change in the rotational energy of the gas molecules). Among different gases only water vapours and oxygen have resonant frequencies in the band of interest up to 100 GHz. The attenuation due to atmospheric gases is normally neglected at frequency bands below 10 GHz. A procedure to find out the effects of gaseous attenuation on LMS links has been discussed in ITU-R recommendations (ITU-R, 2009b).

Scintillations (rapid variations in the received signal level, phase and angle-of-arrival) occur due to inhomogeneities in the refractive index of atmosphere and influence low margin satellite systems. The tropospheric scintillations can be severe at low elevation angles and frequency bands above 10 GHz. Multipath effects can be observed for small percentages of time at very low elevation angles ($\leq 4^o$) due to large scale scintillation effects resulting in signal attenuation greater than 10 dB.

2.3 Local Effects

In addition to the ionospheric and the tropospheric attenuation effects, radio waves suffer from energy loss due to complex and varying propagation environments on the terrain. An earth station is surrounded by different obstacles (buildings, trees, vegetation etc) of varying heights, dimensions and of different densities. These obstructions cause different multipath propagation phenomena: diffraction due to bending of the signal around edges of buildings, dispersion or scattering by the interaction with objects of uneven shapes or surfaces, specular reflection of the waves from objects with dimensions greater than the wavelength of the radio waves, absorption through foliage etc. In addition, the movement of mobile station on earth over short distances on the order of few wavelengths or over short time durations on the order of few seconds results in rapid changes in the signal strength due to changes in phases (Doppler Effect). All these effects result in loss of the signal energy and degrade the performance and reliability of LMS communications links. A detailed

discussion about local effects on LMS communication links can be found in (Goldhirsh & Vogel, 1998; Blaunstein & Christodoulou, 2007).

3. Probability Distribution Functions for Different Types of Fading

The performance of satellite-earth communication links depends on the operating frequency, geographical location, climate, elevation angle to the satellite etc. The link reliability of a satellite-based communication system decreases with the increase in operating frequency and at low elevation angles. In addition, the random and unpredictable nature of propagation environments increases complexity and uncertainty in the characterization of transmission impairments on the LMS communication links. Therefore, it is suitable to describe these phenomena in stochastic manner in order to assess the performance of LMS communication systems over fading channels. Various precise and elegant statistical distributions exist in the literature that can be used to characterize fading effects in different propagation environments (Simon & Alouini, 2000; Corraza, 2007). In general signal fading is decomposed as large scale path loss, a medium slowly varying component following lognormal distribution and small scale fading in terms of Rayleigh or Rice distributions depending on the existence of the LOS path between the transmitter and the receiver. In this section, we give a brief overview of standard statistical distributions used to model different fading effects on the LMS communication links.

3.1 Rayleigh Distribution

In case of heavily built-up areas (Urban Environments) the transmitted signal arrives at the receiver through different multipath propagation mechanisms (section 2.3). The resultant signal at the receiver is taken as the summation of diffuse multipath components characterized by time-varying attenuations, different delays and phase shifts. When the number of paths increase the sum approaches to complex Gaussian random variable having independent real and imaginary parts with zero mean and equal variance. The amplitude of the composite signal follows Rayleigh distribution and the phases of individual components are uniformly distributed in the interval 0 to 2π . The received signal (real part) can be written as:

$$R_{Ray} = \sum_{i=1}^{n} a_i(t)\cos\left(\omega_c t + \theta_i(t)\right) \qquad i = 0,1,2,\ldots,n \tag{1}$$

where $a_i(t)$ is the amplitude, $\theta_i(t)$ is the phase of the i^{th} multipath component and ω_c represents the angular frequency of the carrier. The corresponding probability density function (pdf) of the received signal's envelope is expressed in the following mathematical form:

$$P_{Ray}(r) = \frac{r}{\sigma^2}\exp(-r^2/2\sigma^2) \qquad r \geq 0 \tag{2}$$

where σ denotes the standard deviation and 'r' represents envelop of the received signal.

3.2 Rician Distribution

In propagation scenarios when a LOS component is present between the transmitter and the receiver, the signal arriving at the receiver is expressed as the sum of one dominant vector and large number of independently fading uncorrelated multipath components with amplitudes of the order of magnitude and phases uniformly distributed in the interval $(0,2\pi)$. The received signal is characterized by Rice distribution and is given as follows:

$$R_{Rice} = C + \sum_{i=1}^{n} a_i(t)\cos\left(\omega_c t + \theta_i(t)\right) \qquad i = 0,1,2,...,n \tag{3}$$

where constant 'C' represents the magnitude of the LOS signal between the transmitter and the receiver. Other parameters are the same as described for Rayleigh distribution. The pdf of the envelope of the received signal is illustrated in the following mathematical form:

$$P_{Rice}(r) = \frac{r}{\sigma^2}\exp\left[\frac{-\left(r^2+C^2\right)}{2\sigma^2}\right] I_0\left(rC/\sigma^2\right) \tag{4}$$

where I_0 represents the modified Bessel function of first kind and zero order, and $C^2/2$ is the mean power of the LOS component. If there is no direct path between the transmitter and the receiver (i.e., C = 0) the above equation reduces to Rayleigh distribution. The ratio of the average specular power (direct path) to the average fading power (multipath) over specular paths is known as Rician factor ($a^2/2\sigma^2$) and is expressed in dBs.

3.3 Log-Normal Distribution

In addition to signal power loss due to hindrance of objects of large dimensions (buildings, hills, etc), vegetation and foliage is another significant factor that cause scattering and absorption of radio waves by trees with irregular pattern of branches and leaves with different densities. As a result the power of the received signal varies about the mean power predicted by the path loss. This type of variation in the received signal power is called shadowing and is usually formulated as log-normally distributed over the ensemble of typical locals. Shadowing creates holes in coverage areas and results in poor coverage and poor carrier-to-interference ratio (CIR) in different places. The pdf of the received signal's envelope affected by shadowing follows lognormal distribution that can be written in the following mathematical form:

$$P_{\log normal}(r) = \frac{1}{\sqrt{2\pi}\sigma r}\exp\left[-\frac{1}{2}\left(\frac{(\ln r - \mu)^2}{2\sigma^2}\right)\right] \qquad r \geq 0 \tag{5}$$

where μ and σ are mean and standard deviation of the shadowed component of the received signal, respectively.

3.4 Nakagami Distribution

As discussed in (Saunders & Zavala, 2007), the random fluctuations in the radio signal propagating through the LMS communication channels can be categorized into two types of fading: multipath fading and shadowing. The composite shadow fading (line-of-sight and multiplicative shadowing) can be modelled by lognormal distribution. The application of lognormal distribution to characterize shadowing effects results in complicated expressions for the first and second order statistics and also the performance evaluation of communication systems such as interference analysis and bit error rates become difficult. An alternative to lognormal distribution is Nakagami distribution which can produce simple statistical models with the same performance. The pdf of the received signal envelope following Nakagami distribution is given by,

$$P_r(r) = \frac{2}{\Gamma(m)}\left(\frac{m}{2\sigma^2}\right)^m r^{2m-1}\exp\left(\frac{-mr^2}{2\sigma^2}\right) \qquad r \geq 0 \tag{6}$$

where $\Gamma(.)$ represents the Gamma function, $2\sigma^2 = E(r^2)$ is the average power of the LOS component and $m \geq \frac{1}{2}$ is the Nakagami-m parameter which varies between $\frac{1}{2}$ to ∞. When m increases the number of Gaussian random variables contributions increases and the probability of deep fades in the corresponding pdf function decreases. Non-zero finite small and large values of m correspond to urban and open areas, respectively. The intermediate values of m correspond to rural and suburban areas.

3.5 Suzuki Distribution

The Suzuki process is characterized as the product of Rayleigh distributed process and lognormal process (Pätzold et al., 1998). Consider a Rayleigh distributed random variable α with pdf $P_\alpha(r)$ and another random variable β following lognormal distribution with pdf $P_\beta(r)$. Let ξ be a random variable defined by the product of these two independent variables ($\xi = \alpha.\beta$). The pdf $P_\xi(r)$ of ξ can be written as follows:

$$P_\xi(r) = \frac{r}{\sqrt{2\pi}\sigma_0^2\sigma_\mu}\int_0^\infty \frac{1}{r^3}.\exp\left(\frac{-r^2}{2z^2\sigma_0^2}\right).\exp\left(-\frac{1}{2}\left(\frac{\ln r - m_\mu}{\sigma_\mu}\right)^2\right) \qquad r \geq 0 \tag{7}$$

This type of distribution can be used to represent propagation scenarios when LOS component is absent in the received signal.

4. Statistical Channel Models for Land-Mobile-Satellite Communications

The influence of radio channel is a critical issue for the design, real-time operation and performance assessment of LMS communication systems providing voice, text and multimedia services operating at radio frequencies ranging from 100 MHz to 100 GHz and optical frequencies. Thus, a good and accurate understanding of radio propagation channel

is of paramount significance in the design and implementation of satellite-based communication systems.

The radio propagation channels can be developed using different approaches, e.g., physical or deterministic techniques based on measured impulse responses and ray-tracing algorithms which are complex and time consuming and statistical approach in which input data and computational efforts are simple. The modelling of propagation effects on the LMS communication links becomes highly complex and unpredictable owing to diverse nature of radio propagation paths. Consequently statistical methods and analysis are generally the most favourable approaches for the characterization of transmission impairments and modelling of the LMS communication links.

The available statistical models for narrowband LMS channels can be characterized into two categories: single state and multi-state models (Abdi et al., 2003). The single state models are described by single statistical distributions and are valid for fixed satellite scenarios where the channel statistics remain constant over the areas of interest. The multi-state or mixture models are used to demonstrate non-stationary conditions where channel statistics vary significantly over large areas for particular time intervals in nonuniform environments. In this section, channel models developed for satellites based on statistical methods are discussed.

4.1 Single-State Models

Loo Model: The Loo model is one of the most primitive statistical LMS channel model with applications for rural environments specifically with shadowing due to roadside trees. In this model the shadowing attenuation affecting the LOS signal due to foliage is characterized by log-normal pdf and the diffuse multipath components are described by Rayleigh pdf. The model illustrates the statistics of the channel in terms of probability density and cumulative distribution functions under the assumption that foliage not only attenuates but also scatters radio waves as well. The resulting complex signal envelope is the sum of correlated lognormal and Rayleigh processes. The pdf of the received signal envelope is given by (Loo, 1985; Loo & Butterworth, 1998).

$$P(r) = \begin{cases} \frac{1}{r\sqrt{2\pi d_0}} \exp\left[-\frac{(\ln r - \mu)^2}{2 d_0} \right] & \text{for } r >> \sqrt{b_0} \\ \frac{r}{b_0} \exp\left(-\frac{r^2}{2 b_0} \right) & \text{for } r << \sqrt{b_0} \end{cases} \tag{8}$$

where μ and $\sqrt{d_0}$ are the mean and standard deviation, respectively. The parameter b_0 denotes the average scattered power due to multipath effects. Note that if attenuation due to shadowing (lognormal distribution) is kept constant then the pdf in (8) simply yields in Rician distribution. This model has been verified experimentally by conducting measurements in rural areas with elevation angles up to 30° (Loo et al., 1998).

Corraza-Vatalaro Model: In this model, a combination of Rice and lognormal distribution is used to model effects of shadowing on both the LOS and diffuse components (Corazza & Vatalaro, 1994) The model is suitable for non-geostationary satellite channels such as medium-earth orbit (MEO) and low-earth orbit (LEO) channels and can be applied to different environments (e.g., urban, suburban, rural) by simply adjusting the model parameters. The pdf of the received signal envelop can be written as:

$$P_{(r)}(r) = \int_0^\infty p(r/S)\, p_S(s) dS \tag{9}$$

where $p(r/S)$ denotes conditional pdf following Rice distribution conditioned on shadowing S (Corazza et al., 1994)

$$P(r/S) = 2(K+1)\frac{r}{S^2}\exp\left[-(K+1)\frac{r^2}{S^2} - K\right] I_0\left(2\frac{r}{S}\sqrt{K(K+1)}\right) \quad r \geq 0 \tag{10}$$

where K is Rician factor (section 3.2) and I_0 is zero order modified Bessel function of first kind. The pdf of lognormal of shadowing S, is given by:

$$P_S(S) = \frac{1}{\sqrt{2\pi}h\sigma S}\exp\left\{-\frac{1}{2}\left(\frac{\ln S - \mu}{h\sigma}\right)^2\right\} \qquad S \geq 0 \tag{11}$$

where $h = \ln(10)/20$, μ and $(h\sigma)^2$ are mean and variance of the associated normal variance, respectively. The received signal envelop can be interpreted as the product of two independent processes (lognormal and Rice) with cumulative distribution function in the following form (Corraza & Vatalaro, 1994):

$$P_r(r_0) = P(r < r_0) = \int_0^{r_0}\int_0^\infty \frac{P_S(S)}{S} P_r\left(\frac{r}{S}\right) dr dS = 1 - E_S\left\{Q\left(\sqrt{2K}, \frac{r_0}{S}\sqrt{2(K+1)}\right)\right\} \tag{12}$$

where $E(.)$ denotes the average with respect to S and Q is Marcum $Q-$ function.

The model is appropriate for different propagation conditions and has been verified using experimental data with wide range of elevation angles as compared to Loo's model.

Extended-Suzuki Model: A statistical channel model for terrestrial communications characterized by Rayleigh and lognormal process is known as Suzuki model (Suzuki, 1977). This model is suitable for modelling random variations of the signal in different types of urban environments. An extension to this model, for frequency non-selective satellite communication channels, is presented in (Pätzold et al., 1998) by considering that for most of the time a LOS component is present in the received signal. The extended Suzuki process is the product of Rice and lognormal probability distribution functions where inphase and quadrature components of Rice process are allowed to be mutually correlated and the LOS

component is frequency shifted due to Doppler shift. The pdf of the extended Suzuki process can be written as (Pätzold et al., 1998):

$$P_{\xi}(r) = \int_{-\infty}^{\infty} \frac{1}{|y|} P_{\alpha\beta}(\tfrac{r}{y}, y)dy \tag{13}$$

where $P_{\alpha\beta}(x, y)$ denotes the joint pdf of the independent Rician and lognormal processes $\alpha(t)$ and $\beta(t)$, and $x = r/y$ where y is variable of integration. The pdfs of Rice and lognormal processes can be used in (13) to obtain the following pdf:

$$P_{\xi}(r) = \frac{r}{\sqrt{2\pi}\psi_0\sigma} \int_0^{\infty} \frac{1}{y^3} \exp\left\{ -\left[\left((\tfrac{r}{y})^2 + p^2 \right) \Big/ 2\psi_0 \right] \right\} . I_0(rp/y\psi_0) . \exp\left[-\left((\ln y - m)^2 / 2\sigma 2 \right) \right] \quad r \geq 0 \tag{14}$$

where ψ_0 is the mean value of random variable x, m and μ are the mean and standard deviation of random variable y and p denotes LOS component.

The model was verified experimentally with operating frequency of 870 MHz at an elevation angle 15^o in rural area with 35% trees coverage. Two scenarios were selected: a lightly shadowed scenario and a heavily shadowed scenario with dense trees coverage. The cumulative distribution functions of the measurement data were in good agreement with those obtained from analytical extended Suzuki model.

Xie-Fang Model: This model (Xie & Fang, 2000), based on propagation scattering theory, deals with the statistical modelling of propagation characteristics in LEO and MEO satellites communication systems. In these satellites communication systems a mobile user or a satellite can move during communication sessions and as a result the received signals may fluctuate from time to time. The quality-of-service (QoS) degrades owing to random fluctuations in the received signal level caused by different propagation impairments in the LMS communication links (section 2). In order to efficiently design a satellite communication system, these propagation effects need to be explored. This channel model deals with the statistical characterization of such propagation channels.

In satellite communications operating at low elevation angles, the use of small antennas as well as movement of the receiver or the transmitter introduces the probability of path blockage and multipath scattering components which result in random fluctuations in the received signal causing various fading phenomena. In this model fading is characterized as two independent random processes: short-term (small scale) fading and long-term fading. The long term fading is modelled by lognormal distribution and the small scale fading is characterized by a more general form of Rician distribution. It is assumed, based on scattering theory of electromagnetic waves, that the amplitudes and phases of the scattering components which cause small scale fading due to superposition are correlated. The total electric field is the sum of multipath signals arriving at the receiver (Beckman et al., 1987):

$$E_{tot} = E\exp(j\psi) = \sum_{i=1}^{n} A_i \exp(j\phi_i) \tag{15}$$

where n denotes the number of paths, A_i and ϕ_i represent the amplitude and phase of the i^{th} path component, respectively. The pdf of the received signal envelope can be obtained as follows (Xie & Fang, 2000):

$$P_r(r) = \frac{r}{\sqrt{S_1 S_2}} \exp\left[-\frac{S_1 r^2 + S_2 \alpha^2 + S_1 \beta^2}{2S_1 S_2}\right] \times \frac{1}{2\pi} \int_0^{2\pi} \exp\left[\frac{2S_2 \alpha r \cos\theta + 2S_1 \beta r \sin\theta + (S_1 - S_2) r^2 \cos^2\theta}{2S_1 S_2}\right] d\theta \tag{16}$$

and the pdf of the received signal power envelope is given by:

$$P_p(W) = \frac{1}{2\sqrt{S_1 S_2}} \exp\left[-\frac{S_2 W + S_1 \alpha^2 + S_2 \beta^2}{2S_1 S_2}\right] \times \frac{1}{2\pi} \int_0^{2\pi} \exp\left[\frac{2S_2 \alpha \sqrt{W} \cos\theta + 2S_1 \beta \sqrt{W} \sin\theta + (S_1 - S_2) W \cos^2\theta}{2S_1 S_2}\right] d\theta \tag{17}$$

where the parameters S_1, S_2, α, and β denote the variances and means of the Gaussian distributed real and imaginary parts of the received signal envelope 'r', respectively, and 'W' represents the power of the received signal.

This statistical LMS channel model concludes that the received signal from a satellite can be expressed as the product of two independent random processes. The channel model is more general in the sense that it can provide a good fit to experimental data and better characterization of the propagation environments as compared to previously developed statistical channel models.

Abdi Model: This channel model (Abdi et al., 2003) is convenient for performance predictions of narrowband and wideband satellite communication systems. In this model the amplitude of the shadowed LOS signal is characterized by Nakagami distribution (section 3.4) and the multipath component of the total signal envelop is characterized by Rayleigh distribution. The advantage of this model is that it results in mathematically precise closed form expressions of the channel first order statistics such as signal envelop pdf, moment generating functions of the instantaneous power and the second order channel statistics such as average fade durations and level crossing rates (Abdi et al., 2003). According to this model the low pass equivalent of the shadowed Rician signal's complex envelope can as:

$$R(t) = A(t)\exp\left[j\phi(t)\right] + Z(t)\exp\left[j\varphi(t)\right] \tag{18}$$

where $A(t)$ and $Z(t)$ are independent stationary random processes representing the amplitudes of the scattered and LOS components, respectively. The independent stationary random process, $\phi(t)$, uniformly distributed over $(0, 2\pi)$ denotes the phase of scattered components and $\varphi(t)$ is the deterministic phase of LOS component. The pdf of the received signal envelop for the first order statistics of the model can be written as (Abdi et al., 2003):

$$P_r(r) = \left(\frac{2b_0 m}{2b_0 m + \Omega}\right)^m \cdot \frac{r}{b_0} \exp\left(-r^2 \big/ 2b_0\right) \bullet {}_1F_1\left(m,\ 1,\ \frac{\Omega r^2}{2b_0(2b_0 m + \Omega)}\right) \quad r \geq 0 \quad (19)$$

where $2b_0$ is the average power of the multipath component, Ω is the average power of the LOS component and ${}_1F_1(.)$ is the confluent hypergeometric function.

The channel model's first order and second order statistics compared with different available data sets, demonstrate the appropriateness of the model in characterizing various channel conditions over satellite communication links. This model illustrates similar agreements with the experimental data as the Loo's model and is suitable for the numerical and analytical performance predictions of narrowband and wideband LMS communication systems with different types of encoded/decoded modulations.

4.2 Multi-state Models

In the case of nonstationary conditions when terminals (either satellite or mobile terminal) move in a large area of a nonuniform environment, the received signal statistics may change significantly over the observation interval. Therefore, propagation characteristics of such environments are appropriately characterized by the so-called multi-state models.

Markov models are very popular because they are computationally efficient, analytically tractable with well established theory and have been successfully applied to characterize fading channels, to evaluate capacity of fading channels and in the design of optimum error correcting coding techniques (Tranter et al., 2003). Markov models are characterized in terms of state probability and state probability transition matrices. In multi-state channel models, each state is characterized by an underlying Markov process in terms of one of the single state models discussed in the previous section.

Lutz Model: Lutz's model (Lutz et al., 1991) is two-state (good state and bad state) statistical model based on data obtained from measurement campaigns in different parts of Europe at elevation angles between 13° to 43° and is appropriate for the characterization of radio wave propagation in urban and suburban areas. The good state represents LOS condition in which the received signal follows Rician distribution with Rice factor *K* which depends on the operating frequency and the satellite elevation angle. The bad state models the signal amplitude to be Rayleigh distributed with mean power $S_0 = \sigma^2$ which fluctuates with time. Another important parameter of this model is time share of shadowing '*A*'. Therefore, pdf of the received signal power can be written as follows (Lutz et al., 1991):

$$p(S) = (1-A).p_{Rice}(S) + A\int_0^{\infty} p_{Ray}(S/S_0)\,p_{LN}(S_0)\,dS_0 \tag{20}$$

The values of the parameters A, K, means, variances and the associated probabilities have been derived from measured data for different satellite elevations, antennas and environments using curve fitting procedures. The details can be found in (Lutz et al., 1991). Transitions between two states are described by first order Markov chain where transition from one state to the next depends only on the current state. For two-state Lutz' model, the probabilities P_{ij} $(i, j = g, b)$ represent transitions from sate i to state j according to good or bad state as shown in Fig. 2.

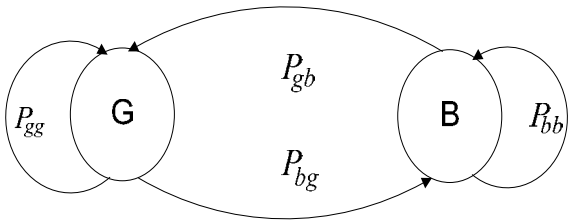


Fig. 2. Lutz's Two-state LMS channel model.

The transition probabilities can be determined in terms of the average distances D_g and D_b in meters over which the system remains in the good and bad states, respectively.

$$P_{gb} = vR/D_g \qquad P_{bg} = vR/D_b \tag{21}$$

where v is the mobile speed in meters per second, R is the transmission data rate in bits per second. As the sum of probabilities in any state is equal to unity, thus $P_{gg} = 1 - P_{gb}$ and $P_{bb} = 1 - P_{bg}$. The time share of shadowing can be obtained as:

$$A = \frac{D_b}{D_b + D_g} \tag{22}$$

The parameter A in this model is independent of data rate and mobile speed. For different channel models, the time share of shadowing is obtained according to available propagation conditions and parameters. For example in (Saunders & Evans, 1996) time share of shadowing is calculated by considering buildings height distributions and street width etc.

Three-State Model: This statistical channel model (Karasawa et al., 1997), based on three states, namely clear or LOS state, the shadowing state and the blocked state, provides the analysis of availability improvement in non-geostationary LMS communication systems. The clear state is characterized by Rice distribution, the shadowing state is described by

Loo's pdf and the blocked state is illustrated by Rayleigh fading as shown in Fig. 3(a), where a_1 denotes the LOS component, a_2 represents shadowing effects caused by trees and a_3 represents blockage (perfect shadowing). Similarly, multipath contributions in the form of coherently reflected waves from the ground are denoted by b_1 and incoherently scattered components from the land obstructions are represented by b_2. The pdf of the received signal envelop is weighted linear combination of these distributions:

$$P_R(r) = MP_{Rice}(r) + LP_{Loo}(r) + NP_{Rayleigh}(r) \tag{23}$$

where *M, L,* and *N* are the time share of shadowing of Rice, Loo and Rayleigh distributions, respectively. The distribution parameters for the model were found by means of the data obtained from measurements using "INMARSAT" satellite and other available data sets. The model was validated by comparing the theoretical cumulative distributions with those obtained from measurement data. The state transitions characteristics of the model were obtained using Markov model as shown in Fig. 3(b). The state occurrence probability functions P_A, P_B and P_c (where $P_A + P_B + P_C = 1$) can be computed as follows (Karasawa et al., 1997):

$$P_A = (90 - \theta)^2 / a \tag{24}$$

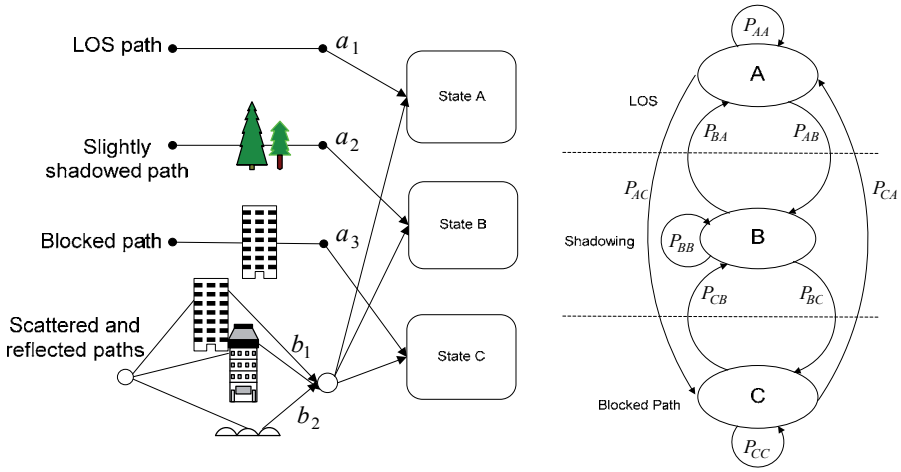
where θ is the elevation angle of satellite ($10^\circ \le \theta \le 90^\circ$) and 'a' is a constant with values:

$$a = \begin{cases} 7.0 \times 10^3 & \text{for urban areas} \\ 1.66 \times 10^4 & \text{for suburban areas} \end{cases}$$

$$P_B = \begin{cases} P_C / 4 & \text{for urban areas} \\ 4P_C & \text{for suburban areas} \end{cases} \tag{25}$$

In order to characterize the state duration statistics such as the average distances or time spans during which a particular state tends to persist, a model capable of providing time-variant features is essential. A Markov process suitable for this purpose is expressed as three-state model as shown in Fig. 3(b) (Karasawa et al., 1997). In this model short-term fluctuations in the received signal are represented by specific pdfs within the states and long-term fading is described by the transitions between the states. This model is also suitable for the performance assessment of satellite diversity.

A significant aspect of the LMS systems is that a single satellite is not adequate for achieving the desired coverage reliability with a high signal quality. Thus, it is desirable that different satellite constellations should be employed which can improve the system availability and signal quality by means of satellite diversity. If a link with one of the satellites is interrupted by shadowing, an alternative satellite should be available to help reduce the outage probability. This channel model also provides analysis for the improvement of the signal quality and service availability by means of satellite diversity where at least two satellites in LEO/MEO orbit, illuminate the coverage area simultaneously in urban and suburban environments.



(a) (b)

Fig. 3. Three-sate LMS channel model (a) Propagation impairments (b) Markov model.

Five-State Model: This channel model is based on Markov modelling approach in which two-state and three-state models are extended to five-state model under different time share of shadowing (Ming et al., 2008). The model is basically a composition of Gilbert-Elliot channel model and the three-state Markov channel model in which shadowing effects are split into three states: 'good' state represents low shadowing, 'not good not bad' state characterizes moderate shadowing and 'bad state' describes heavy or complete shadowing as shown in Fig. 4 (Ming et al., 2008). The 'good' state has two sub-states: clear LOS without shadowing and LOS state with low shadowing. Similarly, the 'bad' state has two sub-states: heavily shadowed areas or completely shadowed or blocked areas. A state transition can occur when the receiver is in low or high shadowing areas for a period of time. The transitions can take place from low and high shadowing conditions to moderate shadowing conditions but cannot occur directly between low and high shadowing environments.

For different shadowing effects, the statistical signal level characteristics in terms of the pdf are described as: low shadowing follows Rice distribution, moderate shadowing is represented by Loo's pdf and high shadowing conditions are described by Rayleigh-lognormal distribution. The pdf of the received signal power is a weighted linear combination of these distributions:

$$P(s) = X_1 P_{Rice1}(s) + X_2 P_{Rice2}(s) + X_3 P_{Loo}(s) + X_4 P_{Ray\_L1}(s) + X_5 P_{Ray\_L2}(s) \qquad (26)$$

where X_i $(i = 1,...,5)$ are time share of shadowing of the states S_i $(i = 1,...,5)$, respectively.

The state probability and state transition probability matrices are determined using the time series of the measured data. The channel model has been validated using available measured data sets and different statistical parameters are obtained using curve fitting procedures. The channel statistics like the cumulative distribution function, the level crossing rate, the average fade duration, and the bit error rate are computed which show a

good agreement with the statistics of the data obtained from measurements. The channel model is appropriate for urban and suburban areas.

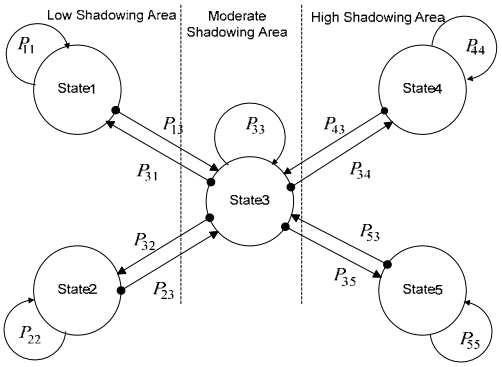


Fig. 4. Five-state Markov channel model for LMS communications.

Modelling Frequency Selective LMS Channel: The LMS propagation channel effects depend on the propagation impairments (section 2), geographical location, elevation angles and operating frequency band. Extensive measurements are needed for the characterization of LMS fading caused by different propagation impairments. When components of a signal travelling through different paths arrive at the receiver with delays significantly larger as compared to the bit or symbol duration, the signal will undergo significant amount of distortion across the information bandwidth, it results in frequency selective fading or wideband fading (e.g., in the case of broadband services or spread spectrum). The impulse response of a wideband channel model (also known as tapped-delay line model) under wide sense stationary uncorrelated scattering (WSSUS) assumption can be written as:

$$h(t,\tau) = \sum_{i=1}^{N} a_i(t)\delta\left[t-\tau_i(t)\right]\exp\left[j(2\pi f_{d,i}(t)+\theta_i(t))\right] \tag{27}$$

where $a_i(t)$, $\tau_i(t)$, $f_{d,i}$ and $\theta_i(t)$ are the amplitude, delay, Doppler shift and phase of the i^{th} component of the received signal, respectively, and $\delta(t)$ denotes the Dirac delta function.

A tapped-delay line model that describes the wideband characteristics of LMS communication link has been given in (Jahn, 2001). The parameters for this model are extracted using extensive measurement data at L-band for different applications, scenarios and environments. In order to adopt the channel for LMS communications, the channel impulse response is divided into three components: the direct path, near echoes and far echoes as shown in Fig. 5 (Jahn, 2001). The delays τ_i $(i=1,2,...,N)$ of the taps are taken with respect to the delay of the direct path. The power of all taps is normalized to the power of the direct path. The amplitude distributions of the echoes follow Rice or Rayleigh distribution (section 3) depending on the presence of LOS or non-LOS situations,

respectively. The number N^n of near echoes in the locality of the receiver follows Poisson distribution with parameter λ $(i.e., f_{Poisson}(N) = (\lambda^N/N!)e^{-\lambda})$ and the corresponding delays τ_i $(i = 1,2,...,N)$ characterizing near echoes follow exponential distribution with parameter b $\{i.e., f_{\exp}(\tau_i^n) = (e^{-\tau_i^n/b}/b)\}$. The power of the taps decay exponentially. The far echoes $N^f = N - N^n - 1,$ which are few in numbers are characterized by Poisson distribution. The amplitude distributions of the far echoes are described by Rayleigh distribution. The description of different regions of the wideband LMS channel impulse response can be found in (Jahn, 2001). Another physical-statistical channel model that deals with the frequency selectivity of LMS channels is found in (Parks et al., 1996). This model consists of two cascaded processes. The first one deals with propagation effects from satellite to earth and the second process illustrates the terrestrial propagation impairments.

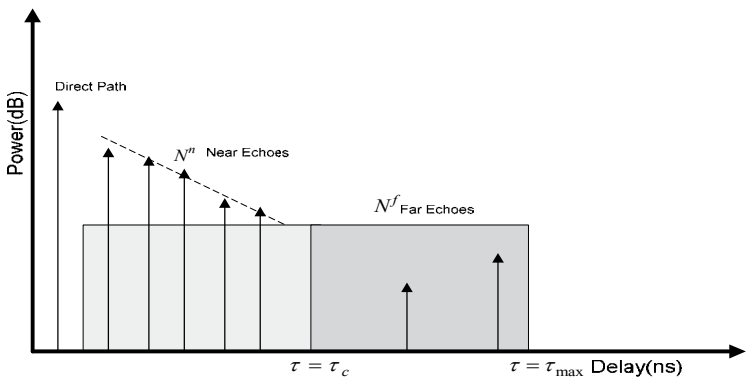


Fig. 5. Wideband LMS channel impulse response with different regions.

5. Conclusions

This chapter provides an overview of propagation impairments on LMS communication links, probability distributions describing these fading effects and channel models developed using these probability distributions. Proper knowledge of propagation impairments and channel models is necessary for the design and performance assessment of advanced transceiver techniques employed to establish reliable communication links in LMS communication systems. The main focus lies on highlighting which are the effects and the relevant propagation models need to be considered for LMS communication links in order to accurately estimate the propagation impairments. The performance of LMS communication systems depend on different factors including operating frequency, elevation angles, geographic location, climate etc. Different approaches can be used to find the effects of these factors on LMS communication links such as physical-statistical channel models which are more accurate but require long simulation times and are complex. On the other hand statistical methods are simple and require less computational efforts. In addition, due to diverse nature of propagation environments, it is appropriate to use stochastic approaches for the performance assessment of LMS communication links.

6. References

Abdi, A., Lau, C. W., Alouini, M., & Kaveh, M. (2003). A New Simple Model for Land Mobile Satellite Channels: First- and Second-Order Statistics. *IEEE Trans. Wireless Comm.*, 2(3), 519-528.

Blaunstein, N., & Christodoulou, C. G. (2007). Radio Propagation and Adaptive Antennas for Wireless Communication Links. John Wiley & Sons, Inc., Hoboken, New Jersey.

Corraza, G. E., & Vatalaro, F. (1994). A Statistical Channel Model for Land Mobile Satellite Channels and Its Application to Nongeostationary Orbit Systems. *IEEE Trans. Vehicular Technology,* 43(3), 738-742.

Corazza, G. E. (2007). Digital Satellite Communications. Springer Science plus Business Media, LLC, New York.

Goldhirsh, J., & Vogel, W. J. (1998). Handbook of Propagation Effects for Vehicular and Personal Mobile Satellite Systems, Over of Experimental and Modelling Results.

Ippolito, J. L., Jr. (2008). Satellite Communications Systems Engineering, Atmospheric Effects, Satellite Link Design and System Performance. John Wiley & Sons Ltd.

ITU. (2002). Handbook on Satellite Communications, Wiley-Interscience, 3rd Edition.

ITU-R. (2007). Ionospheric Propagation data and Prediction Methods Required for the Design of Satellite Services and Syatems. ITU-R P. 618-9.

ITU-R. (2009a). Ionospheric Propagation data and Prediction Methods Required for the Design of Satellite Services and Syatems. ITU-R P. 531-10.

ITU-R. (2009b). *Attenuation by Atmospheric Gases*. ITU-R P. 676-8.

Jahn, A. (2001). Propagation Considerations and Fading Countermeasures for Mobile Multimedia Services. *Int. Journal of Satellite Communications*, 19(3), 223-250.

Karasawa, Y., Kimura, K. & Minamisono, K. (1997). Analysis of Availability Improvement in LMSS by Means of Satellite DiversityBased on Three-State Propagation Channel Model. *IEEE Trans. Vehicular Technology,* 46(4), 1047-1056.

Loo, C. (1985). A Statistical Model for a Land Mobile Satellite Links. *IEEE Trans. Vehicular Technology*, Vol. 34, no. 3, pp. 122-127.

Loo, C., & Butterworth, J. S. (1998). Lan Mobile Satellite Measurements and Modelling. *IEEE Proc.*, 86(7), 1442-14462.

Lutz, E., Cygan, D., Dippold, M., Donalsky, F., & Papke, W. (1991). The Land Mobile Satellite Communication Channel- Rceording, Statistics and Channel Model. *IEEE Transactions on Vechicular Technology,* 40(2), 375-386.

Ming, H., Dongya, Y., Yanni, C., Jie, X., Dong, Y., Jie, C. & Anxian, L. (2008). A New Five-State Markov Model for Land Mobile Satellite Channels. *Int. Symposium, Antennas, Propagation and EM Theory,* 1512-1515.

Parks, M. A. N., Saunders, S. R., Evans, B. G. (1996). A wideband channel model applicable to Mobile Satellite Systems at L-band and S-band. *IEE Colloquim on Propagation Aspects of Future Mobile Systems*, 12, 1-6.

Pätzold, M., Killat, U., & Laue, F. (1998). An Extended Suzuki Model for Land Mobile Satellite Channels and Its Statistical Properties. *IEEE Trans. Vehicular Technology,* 47(2), 617-630.

Ratcliffe, J. A. (1973). Introduction in Physics of Ionosphere and Magnetosphere. Academic Press, New York.Blaunstein, N. (1995). Diffusion spreading of middle-latitude ionospheric plasma irregularities. Annales Geophasice, 13, 617-626.

Roddy, D. (2006). Satellite Communications, The McGraw Hill Companies, Inc, Fourth Edition.

Saunders, S. R., & Evans, B. G. (1996). Physical Model for Shadowing Probability for Land Mobile Satellite Propagation. *IEE Electronic Letters,* 32(17), 1248-1249.

Saunders, S. R., & Zavala, A. A. (2007). Antennas and Propagation for Wireless Communication Systems. J. Wiley & Sons, New York.

Simon, M., & Alouini, M. (2000). Digital Communication over Fading Channels: A Unified Approach to Performance Analysis. John Wileys & Sons, Inc, ISBN 0-471-31779-9.

Suzuki, H. (1977). A Statistical Model for Urban Radio Propagation. *IEEE Trans. Comm.,* 25(7), 673-680.

Tranter, W., Shanmugan, K., Rappaport, T., and Kosbar, K. (2004). Principles of Communication Systems Simulation with Wireless Applications. Pearson Education, Inc.

Xie, Y., & Fang, Y. (2000). A General Statistical Channel Model for Mobile Satelllite Systems. *IEEE Trans. Vehicular Technology,* 49(3), 744-752.

8

Combining satellite and geospatial technologies for exploring rainstorm hazard over Mediterranean Central Area[1]

Nazzareno Diodato
MetEROBS – Met European Research Observatory, GEWEX-CEOP Network, World Climate Research Programme, via Monte Pino snc, 82100 Benevento Italy
e-mail: scodalabdiodato@gmail.com

1. Introduction

> Modelling is not an alternative to observation but,
> under certain circumstances, can be a powerful tool
> in understanding observations and in developing and testing theory.
>
> Mulligan M., and Wainwright J., 2004. Modelling and Model Building.
> In: *Environmental Modelling*, Wiley, p. 2

Multiple Damaging Hydrological Events (MDHE, Petrucci & Polemio, 2003) are rapidly developing into deluges, flashfloods, floods, mudflows, accelerated erosion, and landslides (Kar & Hodgson, 2008; Younis et al., 2008), with tragic consequences on the viable habitat for humankind and ecosystems, and agriculture (Clarke & Rendell, 2005). In this context, MDHE could have more impact than the frequently cited hazard of global warming due to intensification of the hydrological cycle and the concentration of rainfall in sporadic- but more intense events (Allen & Ingram, 2002).

There is, in fact, evidence available from different parts of the world of a rising trend of natural disasters since 1993 (Sivakumar, 2005), included Medietarrean basin (Diodato & Bellocchi, 2010). For Southern Italy, in particular, the catstrophic events of Sarno in 1998 (Mazzarella & Diodato, 2002), with the more recent devastating deluges in Naples in 2001, 2003, 2004, 2006, and in southeastern of Sicily in 2009, were caused by extremes rain of 100-400 mm fallen in few hours over little areas. Therefore, global vision in remote sensing coverage and surveillance loop are important, since we do not know where an event might take place (Bacon et al., 2008). However, estimating rainfall from satellite imagery is rather complex (Ymeti, 2007), and due to limited success of deterministic rainstorm impact modelling techniques (Heneker et al., 2001).

---

[1] This chapter is a revision of the paper appeared on **The Open Environmental Engineering Journal, 2009, 2, 97-103**.

Also, while the literature on general model theory is vast, the aims of modellers usually consist of improving our understanding of a phenomenon and its process, and ultimately predicting the response of the landscape (Kelly et al., 2004; Diodato, 2005). In this context, data assimilation models, that combine ground measurements with remote sensing of rain-data, need to accommodate many specific aspects of the observations and models (Pan et al., 2008).

While surface data will always remain important cornerstones of reference for monitoring and modelling geospatial data, ground data suffers especially due to mutability of their patterns, even as the modeller is compelled to adapt frequently to maintain sufficient condition of temporal and spatial homogeneity, with time-series that are difficult to update.

The advent of Geographical Information Science (GISsci) can confer an innovative role on hazard modelling development, satellite data assimilation, model outputs uncertainty assessment, spatial data scaling, and mapping visualization. Although satellite data are regarded as indirect information and not as reliable as surface data, they can be of great help when used for scaling and assisting the modelling of a dynamic system (Su et al., 2008). However, the problem is that we have a significant increase in uncertainty when the measurements and forecasts move from the global to local scale, especially in their landscape response to change, such as downpours, heavy runoffs and flash-floods, deluges, sediment transport, and urban stormwater (after Beven, 2008). An interesting study for assessing rainfall impact was recently done by (Shoji & Kitaura, 2006) that analyzed precipitation with the parametric geostatistical approach in order to obtain information for predicting natural hazards caused by heavy rains.

In this paper, a different geostatistical criterion was applied – specifically a non-parametric approach – by transforming ground and satellite information into a continuous probabilistic response consistent with *soft* descriptions of hazards which is referred to in this study to mitigate the uncertainties in downscaling and geocomputational tracking (e.g., spatio-temporal non-homogeneity in the primary variable pattern, accuracy of the supplementary variables, errors involving sampling and hazard modelling). Processes operating to these multiple spatial and temporal scales, however, challenge the predictive capability of environmental models and integration or scaling of data from different sources (Allen et al., 2004). Non-parametric geostatistical multivariate analysis, via co-indicator coding criteria, is able to combine rainstorm indicators (which are recorded at sparse raingauge station-points) and supplementary satellite rain data (which are recorded across regular patterns). So that, the novelty of our approach lies in how methods and different tools might incorporate uncertainty associated with satellite data into a model of rainstorm hazard accounting, and to illustrate how model performs at sub-regional scale. In this way, the expansion of a Rainstorm Hazard Index (*RHI*) data from point to spatial information can be assessed with the Indicator CoKriging (ICK) technique, using Tropical Rainfall Mission Monitoring (TRMM–NASA) satellite rain data as covariate. Thus, spatial information is visualized with examples of probability estimations for different precipitation durations – ranging from 3 to 48 hours – and the quantification of hydrological hazard fields is done using probability maps of damaging rainstorms prone-areas.

2. Reference Data Sets and Methodology

2.1 Study area and problem setting

Heavy rainfall between 1951 and 2007 show Northern Mediterranean more affected than Southern one (Fig. 1a). Worldwide temporal pattern is also shown with a trend of hydrological disasters strongly increasing (Fig. 1b).
The rainstorms most perceived by the public are the large-scale damaging events; however, there is evidence that the most deadly floods are those with short lead times – flash floods – which in Mediterranean Europe have mostly a spatially limited character and can occur far away from major rivers (Lalsat et al., 2003).

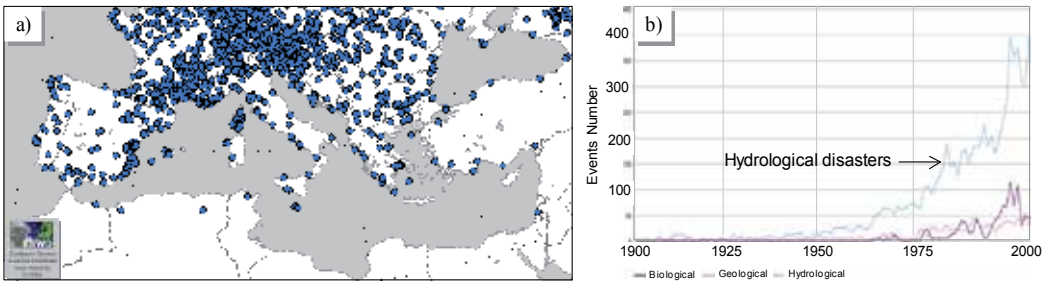


Fig. 1. (a): Occurrence of the heavy rain and hail during 1951–2007 period across Mediterranean lands (http://essl.org/cgi-bin/eswd/eswd.cgi); (**b**): Global natural disasters trends upon 1900-2005 period from EM-DAT (OFDA/CRED International Disaster Database, http://www.emdat.be).

In this respect, a test-area extending approximately 60000 km^2, was selected from Mediterranean central area (Fig. 2a corner). SCIA-APAT Database (www.apat.it/) was utilized for collecting rainfall ground data. However, ground data are not always updated and not all the networks uniformly coincide at all times with this database. Then satellite rain-data were also derived from the TRMM-NASA platform, algorithm 3B42 multi-satellite precipitation estimates (Huffman et al., 2007), that uses an optimal combination (HQ) of 2B-31, 2A-12, SSMI, AMSR, and AMSU precipitation estimates, with a resolution of 0.25x0.25 degree (about 25x25 km) grid boxes (http://disc.sci.gsfc.nasa.gov/).
In this way, a reference classification was constructed from RHI, driven by rainstorm events on 14 November 2004, 24 January 2003, and 4-5 May 1998. Data assimilation pattern in the region under study were obtained from 64 raingauges (Fig. 2a), and 143 supplementary satellite rain grid-data (Fig. 2b).

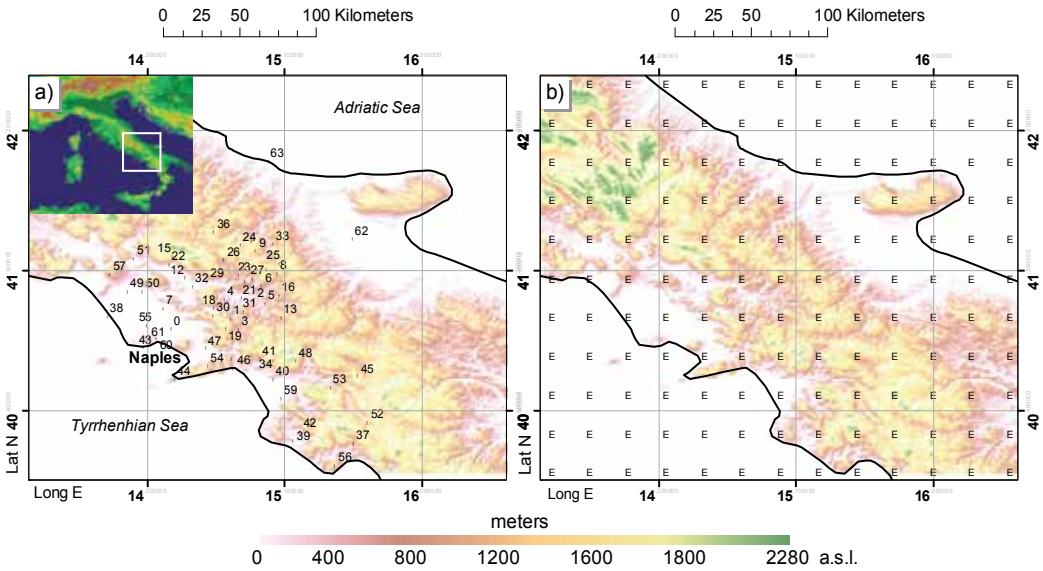


Fig. 2. (**a**): Geographical setting and data assimilation patterns from in-situ-raingauges with coded-station-points, and (**b**): TRMM-RS satellite rain data pixel centroid grid of 25 x 25 km, superimposed on elevation data of hillshade land derived from DEM (SRTM)-90 meters (http://srtm.csi.cgiar.org/).

2.2 Rainstorm hazard problem-solving logic process

Expert systems can be designed to model processes when carried out using the IF-THEN logic statement to impose an event contingent upon the condition (Moody & Katz, 2004).
Problem–solving logic process frameworks include first an invariant spatial model recognizing critical-thresholds from the response ratios between the two following components of the landscape:

- pulsing force that disturbs the system, including current rainstorm depth, and;
- resistance force, including storm variability that occurred in the system's climate history.

As a more concrete application, we can incorporate, for each rainy step of duration h at sampled location $\mathbf{s}_\alpha$, two processes into the rainstorm logic statement linking the the *RHI* to the following power equation (after Diodato, 2006; Diodato & Petrucci, 2009):

$$RHI_h\left(\mathrm{s}_\alpha\right) = \max : \left\{ \left[\frac{1 + RSD_h}{f_{(Rclim)}} \right]^2_{h(\mathrm{s}_\alpha)} \quad \forall \quad h = 1...48 \quad \text{hours} \right. \tag{1}$$

where RSD_h is the Rain-Storm Depth (mm), that represents the **pulsing force** that disturbs the system during an event of duration h, and:

$$f_{(Rclim)} = \mathrm{Med}\left(RSD_h\right) - \left(8 - \sqrt{h}\right) \cdot S_{wet} \tag{2}$$

is a function that represents the **system resistance state**, that is the intrinsic ability of the system to resist change because of its history (recent and past). Med(RSD_h) –the threshold value – is the median of the annual maximum rainfall (mm) of duration h, and the term (8–$\sqrt{h}$)S_{wet}, is a function adjusting the threshold value with the current variation of the soil humidity. As proxies of the soil humidity, three coefficients was introduced as S_{wet} equal to 0.5, 0, and 2 according to dry, humid, and very humid soil conditions before the event, respectively; these coefficients can be derived, in turn from remote sensing; the duration of rainstorm (h) under square root is to explain a major accommodation of the system for rainfall spanning over a longer period. Whereas, for each sampled location with a Rainstorm Hazard Index (RHI)→0, no-rainstorm hazard occurs, and with RHI value barely over 1, the probability of occurrence of a rainstorm hazard commences at 0.50.

2.3 Matching coding approach for decision–making under uncertainty

While the above RHI–model is utilized to arrive at conclusions at the puntual-scale, the use of geostatistics method may help to overcome the inherent difficulties in spatial scaling, when the above RHI discrete data must accommodate a continuous spatial solution and data collection across sampled- and unsampled locations. Thus, the RHI–results are converted to a binary vector and matched to satellite rain-data under a GIS flow and supported by indicator cokriging technique (Fig. 3).
Consider the following information obtained over the study area:

- values of the random primary variable Z (RHI), at m locations $\mathbf{s}_\alpha$, $z(\mathbf{s}_\alpha)$, α = 1,2 … n_1; and
- $y(\mathbf{s})$ TRMM satellite rain-data at supplementary grid locations $\mathbf{s}$ within the area.

Indicator approach of the primary variable requires that all data be coded as local prior probability values. Precise measurements of z_k at hard data locations $\mathbf{s}_\alpha$ are then coded into a set of K binary (hard) indicator data defined as:

with
$$i\left(\mathrm{s}_\alpha; z_k\right) = \begin{cases} 1 \rightarrow z\left(\mathrm{s}_\alpha\right) > z_k \\ 0 \rightarrow \text{otherwise} \end{cases} \tag{3}$$

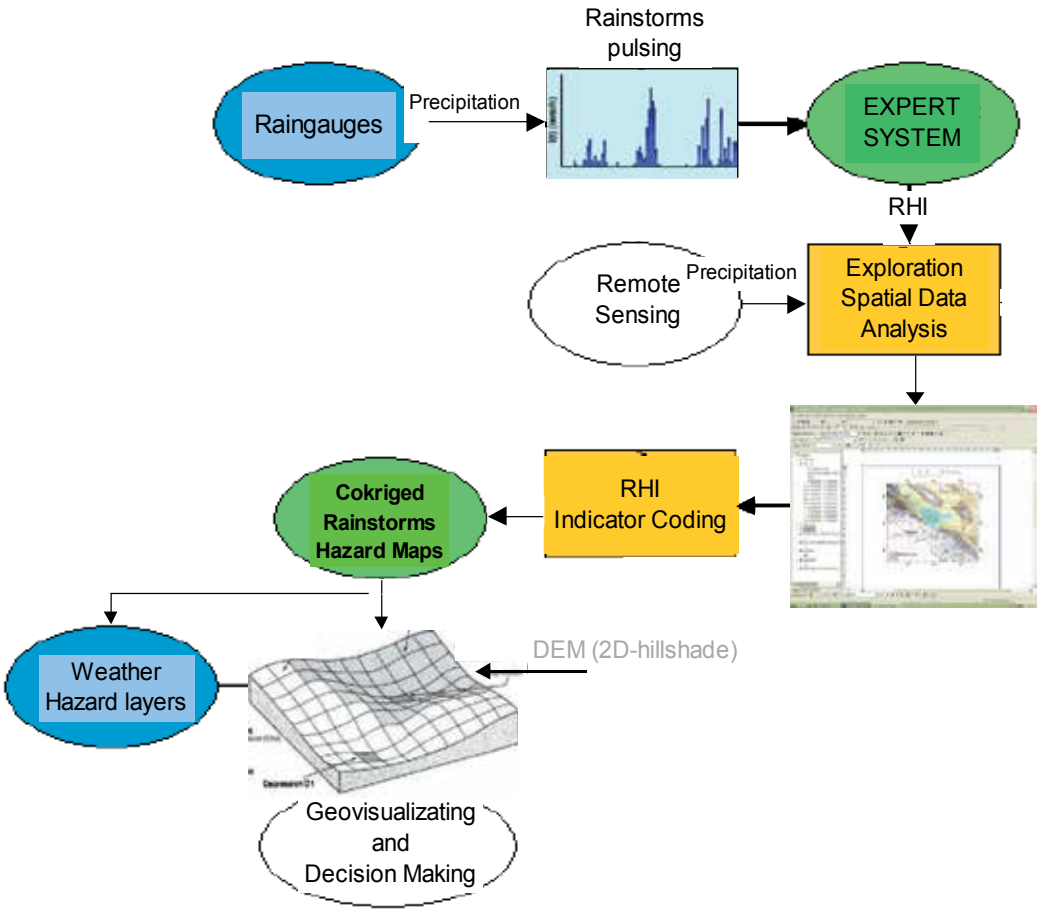


Fig. 3. Flow chart of process for estimation of rainstorm hazard mapping via GIS rules.

The z–values are *hard* in both senses: (1) they are directly derived from measurements of ground rainfall-data, and (2) are successively transformed into binary vector data. These measurements are often supplemented by a relatively large amount of indirect data, such as those conditioned on remotely sensed spectral response $y(\mathbf{s})$.
Each of these data provides only indirect information about the value of the variable Z. Using both *ground* and *satellite* information such as matching data, the approach is aimed at assessing the probability that the value of z at any unsampled site $\mathbf{s}$ is greater than a given critical z_k value. In this way, Indicator CoKriging (ICK) is able to take into account both the information to be processed together, and then used in the ordinary cokriging equations (Goovaerts, 1997; Johnston et al., 2001). To account for both categorial (*RHI*) and continuous (satellite data), we used standardized variables to produce composite indices compatible to indicator cokriging (Johnston et al., 2001; Hengle et al., 2004). So that, both covariance and cross-covariance functions were applied on the above standardized primary and auxiliary variables for incorporating exhaustively sampled satellite data using the indicator datum that is collocated with the location being estimated. Availability of coregionalization between indicator *ground* and *satellite* at critical values of *RHI* for each location $\mathbf{s}_o$ within the study area

allows a grid layer of: the hazard $\alpha(\mathbf{s}_\alpha)$ of declaring a location vulnerable to damage by rainstorms on the basis of the estimate $\left[I(s;z_k)\right]_{\text{oIOK}}^{*}$ when actually $Z(\mathbf{s}) > z_k = p_c$ (critical value = 1).

3. Results and Discussions

3.1 Mapping the rainstorms hazard prone-areas

Figure 4 (a,b,c) shows high-probability cokriged (p>50%) maps of areas prone to rainstorm hazards (dark grey zones), and superimposed by areas where multiple damaging hydrological events (MDHE) were observed. It was found that areas with high probability of predicted hazard matched the area actually subject to injurious phenomena, such as severe erosion, landslides, floods, and mudflows.

The severity of the damage suffered in these areas was not uniform for each rainstorm level, i.e., the damage observed depended not only on the amount of rainfall but also on the sensitivity of each specific landscape and on soil humidity (others topographical conditions were not considered in this work).

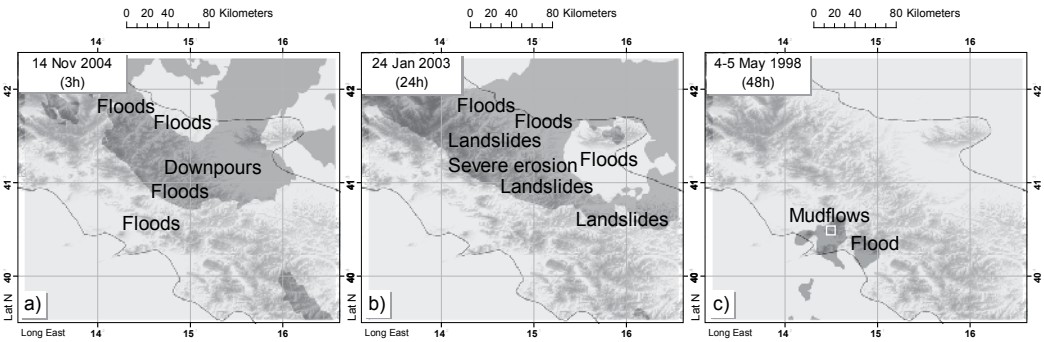


Fig. 4. High-probability (p>50%) cokriged maps of areas prone to rainstorm hazards (dark grey zone) for rainstorms of duration 3, 24, and 48 hours (a, b, and c, respectively). Note: the damaging hydrological events superimposed were almost all matched by the cokriged model.

The most extreme hydrogeomorphological processes occur over orographically complex terrain where vegetation is sparse (especially lands that are under autumn tilling, or after the rainy season), and where drainage systems may be obstructed by sediment erosion to contain large volumes of runoff. Prolonged rain usually occurs during the winter season only, when rainstorm prone-areas assume winding configurations, as for instance in Fig. 4b. On the contrary, MDHE are more spatially limited in the warm season (May-September), but more intensive, such as those which occurred on 14 November 2004 (Fig. 4a). Although the event of May 4-5, 1998 was expected to be of lower intensity, because of its long duration, the impact was catastrophic at the Sarno location (Campania region), where the several mudflows destroyed over one hundred people (Fig. 4c). This occurred because the meteorological perturbation originated from convective-clouds in larger systems, which are today more dominant in rain-producing mechanisms of high-impact over small areas (Dünkeloh & Jacobeit, 2003). This complexity is reflected in stormy pattern concentrate to the end of summer, as it is possible observe by positive anomalies of rain amount in

September occur during the recent decade (1999-2008), compared to the climatological period (1950-2000) (Fig. 5a).
The graph of Figure 5b also suggests that Southern Italy is subjected to a increasing precipitation, in term of intensity and of course of hazard, within September months spanning from 1948 to 2009. After 1996 in fact, rain rates were unusually very high, showing an abrupt change during the last decade. This can be an important advice when RHI modelling is used in an operative phase, where the climate information and human experiences becoming essential for successful completion an alert forecasts.

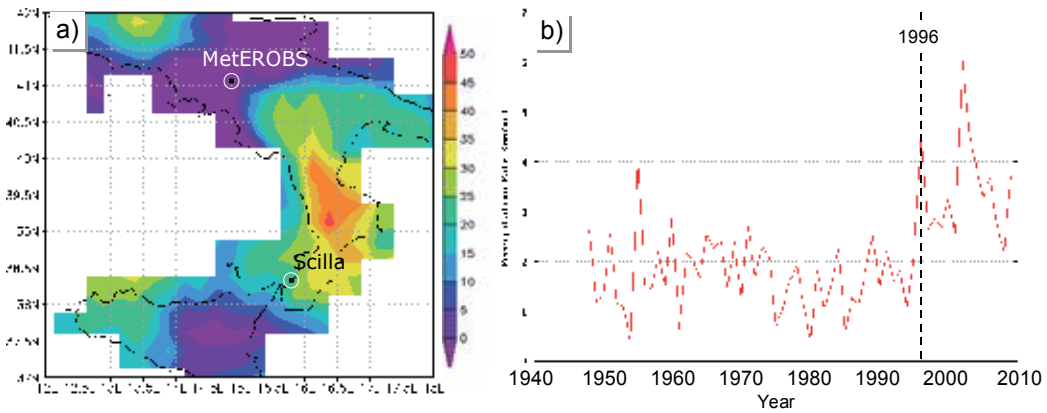


Fig. 5. (**a**): Spatial pattern of the September rain anomalies (mm) during recent decade (2000-2008) compared to climatological rain (1950-2000) upon Southern Italy (arranged from TRMM remote sensing by NASA Earth Science Data); (**b**): September rain rate evolution from 1948 to 2009 for the same region (from NOAA, ESRL Physical Sciences Division http://www.esrl.noaa.gov/ Acker & Leptoukh, 2007).

In this respect, the our results show that sub-regional rainstorm hazard modelling can provide probability maps for damaging events in Italy with a spatial variability resolution of approximately 20 km. Spatially finer estimates (e.g., at local-scale: < 10 km) can be ensured only with the availability of more accurate and detailed supplementary satellite-rain data, although, as noted by Anagnostou (2004), all satellite sensors are affected by errors originating from the non-unique, non-linear relationship of rainfall characteristics to observations and by sampling frequency and sensor resolution issues.

4. Conclusion

The model presented here provided the minimum but valuable set of data from which a rough tool for estimating early impacts soon after rainstorms can be derived. Damaging rainstorms collected for this retrospective experiment are documented in the category of localized events. Impact of the damage was determined by an optimum scaling critical value for predicting hazard prone-areas of three rainstorm types, although the *RHI*–model is capable of performing with data of storms of different intensities. These first results show that sub-regional rainstorm hazard modeling can provide probability maps for damaging events in Italy with a spatial variability resolution of approximately 20 km. Spatially finer estimates can be ensured only with the availability of more accurate- and detailed satellite-

rain data, or during forecast stages, if real-time monitoring is implemented on an operational basis, where supplementary satellite information is then replaced by Quantitative Precipitation Forecasting.

5. References

Acker J.G. & Leptoukh, G. (2007). Online analysis enhances use of NASA earth science data. *EOS Trans.* AGU 88, 14–17.

Allen, M.R. & Ingram, W.J. (2002). Constraints on future changes in climate and the hydrologic cycle. *Nature* 419, 224-231.

Allen, T.R.; Walsh, S.J.; Cairns, D.M., Messina, J.P., Butler, D. & Malanson, G.P. (2004). Geostatistics and spatial analysis: Characterizing form and pattern at the Alpine treeline. In: Geographical Information Science and Mountain Geomorphology, Bishop, M.P. & J. F. Shroder Jr, (eds.). Springer-Praxis Publishing: Chichester, 190-218.

Anagnostou, E.N. (2004). Overview of overland satellite rainfall estimation for hydro-meteorological applications. *Surv. Geophys.*, 25, 511-537.

Bacon, D.P.; Ahmad, N.N.; Dunn, T.J.; Montheit, M.C. & Sarma, A. (2008). An operational multiscale system for hazards prediction, mapping, and response. *Nat. Hazards*n 44, 317-327.

Beven, K. (2008). Models, Management and Uncertainty: The future of Hydrological Science. In: *Rivers Basins-From Hydrological Science to Water Management*, I. Tchiguirinskaia, S. Demuth, and P. Hubert (eds.), IAHS Publication No. 323 ISBN 978-1-901502-69-5, 154.

Clarke, M.L. & Rendell, H,M. (2005). Climate, extreme events and land degradation. In: *Extreme Weather Events and Public Health Responses*, W. Kirch, B. Menne, and R. Bertollini (eds.), Springer, 136-152.

Diodato, N. (2005). Geostatistical uncertainty modelling for the environmental hazard assessment during single erosive rainstorm events. Environ. Monit. Assess. 105, 25-42, 2005.

Diodato, N. (2006). Spatial uncertainty modeling of climate processes for extreme hydrogeomorphological events hazard monitoring. ASCE's *J. Environ. Eng*. 132, 1530-1538.

Diodato, N.; Petrucci, O. & Ceccarelli, M. (2009). Rainstorm hazard problem-solving spatial-time scale invariant process model designing. *Geophys. Res. Abstracts*, vol. 11, EGU2009-4107, 2009.

Diodato, N. & Bellocchi G. (2010). Storminess and Environmental Changes in the Mediterranean Central Area. Earth Interaction (in press).

Dünkeloh, A. & Jacobeit, J. (2003). Circulation dynamics of Mediterranean precipitation variability 1948-1998. *Int. J. Clim.* 23, 1843-1866.

Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation*, Oxford: Oxford University Press.

Hengle, T.; Heuvelink, G.B.M. & Stein, A. (2004). A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma* 120, 75–93.

Heneker, T.M.; Lambert, M.F. & Kuczera, G. (2001).A point rainfall model for risk-based design. *J. Hydrol.*, 247, 54-71.

Huffman, G.J.; Adler, R.F.; Bolvin, D.T.; Gu, G.; Nelkin, E.J.; Bowman, K.P.; Hong, Y.; Stocker, E.F. & Wolff, D.B. (2007). The TRMM Multi-satellite Precipitation Analysis: Quasi-Global, Multi-Year, Combined-Sensor Precipitation Estimates at Fine Scale", *J. Hydrometeorology* 8, 38-55.

Johnston, K.; Ver Hoef, J.M.; Krivoruchko, K. & Lucas, J. (2001). *Using ArcGis Geostatistical Analyst*, New York: ESRI.

Kar, B. & Hodgson, M.E. (2008).A GIS-Based Model to Determine Site Suitability of Emergency Evacuation Shelters. *Trans. in GIS* 12, 227–248.

Kelly, E.J.; Drake, N.A. & Barr, S.L. (2004). Spatial modelling of the terrestrial environment: the coupling of remote sensing with spatial models. In: *Spatial modelling of the terrestrial environment*, E.J. Kelly, N. A. Drake, & S.L. Barr (eds.), Wiley & Sons Ltd: Chichester, 1-6.

Llasat, M.C.; Rigo, T. & Barriendos, M. (2003). The 'Montserrat-2000' flash flood event: A comparison with the floods that have occurred in the northeast Iberian Peninsula since the 14th century. *Int. J. Climatol.* 23, 453–469.

Mazzarella, M. & Diodato, N. (2002). The alluvial events in the last two Century at Sarno, southern Italy: their classification and power-low time occurrence. *Theor. App. Climatol.* 72, 75-84.

Moody, A. & Katz, D.B., (2004). Artificial intelligence in the study of mountain landscapes", in *Geographical Information Science and Mountain Geomorphology*, M. P. Bishop, M.P., and J. F. Shroder Jr, Ed. Springer-Praxis Publishing: Chichester, 220-251, 2004.

Pan, M.; Wood, E.F.; Wojcik, R. & McCabe, M.F. (2008). Estimation of regional terrestrial water cycle using multi-sensor remote sensing observations and data assimilation. *Remote Sens. Environ*. 112, 1282-1294.

Petrucci, O. & Polemio, M. (2003). The use of historical data for the characterisation of multiple damaging hydrogeological events. *Nat. Hazards Earth Syst. Sci.*, 3, 17–30.

Shoji, T. & Kitaura, H. (2006). Statistical and geostatistical analysis of rainfall in central Japan. *Comput. Geosci*. 32, 1007-1024.

Sivakumar, M.V.K. (2005). Impact of natural disasters in agriculture, rangeland and forestry: an overvie. In: *Natural Disasters and Extreme Events in Agriculture*, M. V. K. Sivakumar, R. P. Motha, & H. P. Das (eds.), Springer-Verlag: Berlin, 1-22.

Su, H.; Wood, E.F.; Wang, H. & Pinker, R.T. (2008). *Spatial and Temporal Scaling Behavior of Surface Shortwave Downward Radiation Based on MODIS and In Situ Measurements. Geosci. Rem. Sens. Lett. IEEE* 5, 542-546.

Ymeti, I. (2007). Rainfall estimation by Remote Sensing for conceptual rainfall-runoff modelling in the Upper Blue Nile basin. M.S. Thesis, International Institute for Geo-information Science and Earth Observation, The Netherlands.

Younis, J.; Anquetin, S. & Thielen, J. (2008). The benefit of high-resolution operational weather forecasts for flash flood warning. *Hydrol. Earth Syst. Sci.* 12, 1039–1051.

9

Design and Simulation of a DVB-S2-like Adaptive Air interface Designed for Low Bit Rate Emergency Communications Satellite Link in Ku/Ka/Q/V Bands

Ponia Pech[1], Marie Robert[2], Alban Duverdier[2] and Michel Bousquet[3]
[1]TeSA, [2]CNES, [3]ISAE/SUPAERO
France

1. Introduction

Access to information is of paramount importance in emergency situations when a disaster or a natural catastrophe (earthquakes, landslides, tsunamis, tidal wave, volcanic eruptions, floods, cyclones, other tornados or forest fires) occurs in a given place. The rescuers must have access to telecommunications means in order to be able to transmit critical information (such as the number of victims, of persons needing medical assistance or to be transported towards care centers, hospital locations, the nature of injuries, etc.) to the center that is in charge of coordinating the heavy logistical rescue operations. However, when an extensive area is affected (in developing countries in particular, but this holds true in first world countries also), it is most often impossible to use some telecommunications equipments, because of a destruction of the base stations or an unavailability of electric power. Thus, it is an absolute priority, in order to manage a situation of post-disaster, to restore the communications infrastructure. Three distinct and consecutive phases can be distinguished (Pech et al., 2008):

- **Phase 1**: This is the phase that immediately follows the disaster, where no means of communications is available, and where minimal emergency communications must be established as quickly as possible. This is in this phase that the very rapid deployment (less than some hours) of a satellite system seems most suited. The ground equipments that can be envisaged in this type of system are transportable or portable terminals, not necessarily mobile, that can be operational everywhere between 5 and 30 minutes without the assistance of a technical expert staff. Transmissions must be secured and have a strong availability.
- **Phase 2**: In a second stage, the restoration of a more elaborated communications network involving a transportable user terminal with a higher capacity also ensuring wireless connectivity (GSM, Wi-Fi, WiMax, etc.) is necessary to exchange more complex information. This local infrastructure can be linked to the national telecommunications

public network through a DVB-RCS-like professional satellite terminal, or an Inmarsat's BGAN (Global Broadband Area Network) device (Inmarsat).

- **Phase 3**: After several weeks, even several months, once the first necessities are met and the situation is being reestablished little by little, a restoration of the nominal communication infrastructure can be envisaged.

Emergency satellite systems intended to address the needs of the rescue teams both in phase 1 and phase 2 have been the subject of numerous recent studies and developments, such as for instance the FP6 WISECOM project (cf. Fig. 1) (WISECOM, 2007), diverse standardization works TISPAN (ETSI. EG 202 339, 2004), EMTEL (ETSI. TS 102 181, TS 102 182 and TR 102 410), and ITU-R WP-4B work on wideband spreading signals (IUT-R. IUT-R WP 4B, 2006), and research studies on UWB (*Ultra Wide Band*) satellite transmission carried out with the French Space Agency (CNES - *Centre Nationale d'Etudes Spatiales*) and Thales Alenia Space France (Leconte et al., 2006; Dervin, 2007). Another related research field which has drawn much attention these last years from academia and industries alike is telemedicine, for instance with the OURSES (French acronym meaning *Offer of Services Rural Use using Satellite*) project (Girault et al., 2008; Mailhes et al., 2008).

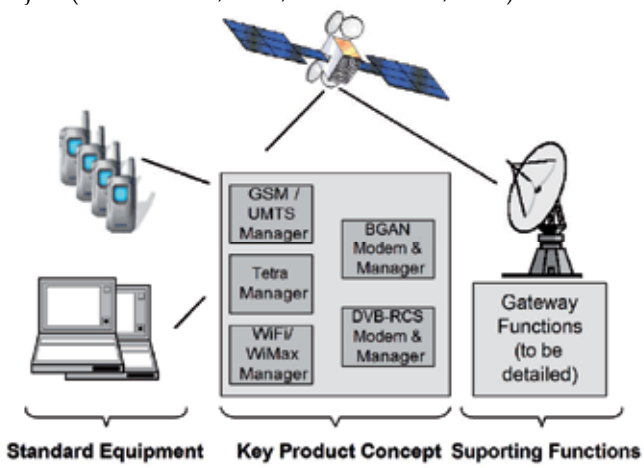


Fig. 1. The WISECOM architecture

The present study inserts itself only within the framework of the aforementioned phase 1. The purpose here is to propose and study a solution allowing to establish very quickly a minimal low bit rate satellite link (in the order of several tens of kbps) in an emergency situation, while using the available resources of the satellite, and a single-user ground terminal with a low transmission power, a small diameter antenna, and a dedicated Tx/Rx system, characterized by an electric consumption which shall be as low as possible (battery or solar panels). The bi-directional transmission link is modeled within the Juzzle [4] open source environment software with an emphasis on the return channel, deploying an adaptive strategy based on the DVB-S2 adaptive modulation and coding (ACM) scheme. This paper expounds the link budget dimensioning and a customized, enhanced DVB-S2 air interface proposed to support minimal emergency communications in a severely impaired channel environment in high frequencies; it also outlines the combined Excel/Matlab/Juzzle high-level transmission link software simulation platform that is being developed in order

to assess the performance of the proposed transmission scheme. Focus is placed on the underlying theoretical models involved in the simulator which follows a cross-layer approach, integrating propagation, DVB-S2-based physical layer, and traffic components.

2. System architecture, scenario, and traffic characterization

2.1 System architecture and scenario

The proposed system architecture and scenario are as follows (Pech et al., 2008): the emergency mission signals are superimposed with those of the primary system characterized by a star topology, a classic multibeam, multicarrier, broadband bent-pipe satellite operating either in Ku/Ka or Q/V-bands, and with a 120-MHz transponder (cf. Fig. 2). Thus a dedicated channel for the emergency mission is not required. The gateway has at its disposal a bandwidth of 480 MHz, while the user links share a total bandwidth of 240 MHz, in four 120-MHz sub-bands and two polarizations, with a reuse frequency factor of 1 over 4. The frequencies used for the return link are enlisted in Table 1. Furthermore, all-IP (Internet Protocol) architecture is assumed.

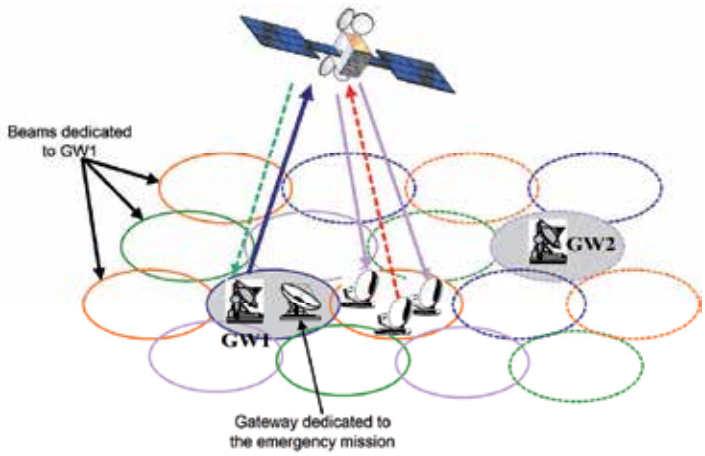


Fig. 2. System architecture

The main characteristics of the satellite are summarized in Table 2. For the emergency mission, a set of 4 different types of user terminals is considered: two mobile user terminals of very low transmission power (between 0.5 and 2 W), the first one (UTA) having a patch antenna, and the second one (UTB) an omnidirectional antenna; and two deployable or transportable user terminals: UTC is a rapidly deployable "mini-gateway" mounted on a van, transmitting at up to 50 W, while UTD can be transported by a human user and has a transmission power of up to 5 W.

2.2 Traffic characterization

In phase 1 of a post-disaster emergency situation, the services which are most required to be provided as a minimum requirement are voice services, and transfer of small text or video files (SMS - Short Message Service / MMS - Multimedia Messaging System). Special emphasis will be laid upon Voice Over IP (VoIP), and more generally IP applications, as the

ever increasing widespread use of IP technologies will greatly contribute to enhancing mobility and nomadicity between private and corporate access. The data to be transmitted are characterized mainly in terms of bit rate, error rates (Bit Error Rate – BER, or Packet Error Rate - PER), delay, jitter, average and maximum packet sizes. The salient Quality of Service (QoS) features of the two types of applications which are envisaged are presented hereafter:

- Real-time, delay-sensitive but not or very little loss-sensitive applications: especially VoIP. A strictly minimum bit rate of 5.3 kbps (assuming ITU H.323 G.723.1 ACELP codec) is required (Nguyen et al., 2001). It has been shown however that optimal bandwidth occupation for VoIP over satellite is around 12 kbps. VoIP bit rate also varies depending on the codec, on whether RTP (Real-Time Protocol) is compressed, and on the overhead introduced by the header of the protocol suite (Ethernet, IP, UDP, RTP). For instance, with RTP, a total of 54 bytes are transmitted, out of which only 20 belong to the payload. The bit rate can thus be considered to range between 5.3 and 13 kbps. A minimum MOS (Mean Opinion Score) requirement of 3.5 as suggested in WISECOM (WISECOM, 2007) ensures a good voice quality. In this respect, ITU-T G.114 Recommendation (IUT-T, 2000) specifies a maximum latency value of 150 ms for one-way VoIP communications. As for the jitter, the mean packet inter-arrival time at the receive side must be roughly close to that at the transmit side, with a small standard deviation. Lastly, in terms of packet corruption and loss, some experiments have shown that the satellite link is quite robust to packet corruption in clear sky or moderately degraded channel environment with a BER of up to a 10^{-5} (that is, Frame Erasure Rate or FER of 2%) (Nguyen et al., 2001). The performance of VoIP over satellite as obtained by Nguyen et al. (Nguyen et al., 2001) from the COMSAT laboratories will be used as a reference test bench.

- Data-like loss-sensitive, but not or very little delay-sensitive applications, for instance SMS/MMS, email applications, file exchange and Internet browsing: the transmitted mean bit rate shall be at least 32 kbps for Web browsing and file exchange, and 200 kbps for email applications (Inmarsat). BER values of up to 10^{-6} can be supported (WISECOM, 2007). Moreover, the time interval between the sending of an SMS and its reception by the receiver must be between 6 and 8 s in average, given that actually 98% of sent SMSs are successfully delivered by a mobile user to a fixed network within a 5-s time period, according to some telecom operators [ETSI, February 2006; ETSI, October 2004]. Since the integrity of SMS messages is 100%, it is obvious that SMSs are well fitted to emergency communications.

3. Air interface, channel modelling and link budget

3.1 Enhanced DVB-S2 air interface

It is proposed to adopt the ETSI DVB-S2 ACM MODCODs (ETSI, EN 302 307, January 2004) for the return channel as it provides excellent performance close to Shannon's theoretical limit due to an advanced Forward Error Correction (FEC) scheme (concatenated BCH and LDPC codes), and allows an attractive waveform flexibility in presence of channel fading, with its inherent ACM capability. Incidentally, this adoption of DVB-S2 ACM schemes for the return channel was recently standardized within the DVB-RCS+M working group (ETSI, EN 301 790 V1.5.1, May 2009), but in the latter standard, very short (4 kbps) DVB-S2

PLFRAMEs were envisaged instead of normal and short lengths (64,800 bits for the normal frame, and 16,200 bits for the short frame). By contrast, here, standard-length DVB-S2 PLFRAMEs are assumed. In quasi-error free (QEF) environment (PER of 10^{-7}), in an Additive White Gaussian Noise (AWGN) channel, DVB-S2 operates at ideal E_s/N_0 ranging from 16 dB down to -2.35 dB. The performance of the receiver in terms of signal acquisition/ reacquisition time, decoding thresholds, etc., all are are well known (ETSI, TR 102 376, February 2002). The present novelty is the use of DVB-S2 along with spread spectrum and other adaptive mechanisms for return channel interactive services while DVB-S2 was specifically designed for the forward link and broadcasting services. This thus raises some performance challenges to be coped with.

Since the considered applications are all assumed here to be IP-based ones, the DVB-S2 waveform is coupled with an efficient encapsulation mechanism, namely the GSE (*Generic Stream Encapsulation*) protocol at the Segmentation and Reassembly (SAR) layer, aimed at segmenting network layer IP datagrams (PDUs or *Payload Data Units*) into link layer DVB-S2 basic data units called BBFRAMEs (*base band frames*). The GSE mechanism was designed with the purpose of fully taking advantage of the innovative features of the DVB-S2, primarily in terms of reliability, flexibility, and enhanced capacity. MPE (*Multi-Protocol Encapsulation*) and ULE (*Unidirectional Lightweight Encapsulation*) have been the standard encapsulation techniques that were classically used in DVB-typed satellite systems, and as such they received abundant attention in the literature. Nonetheless, GSE constitutes a much more efficient encapsulation scheme fitted to the DVB-S2 standard in that it allows to fully exploit the adaptive ACM capability, implementing QoS scheduling decisions and flexible placement and enhanced fragmentation of PDUs in the flow (Cantillo, 2008).

In particular, DVB-S2 GS (*Generic Stream*) data flows may be packetized or continuous streams. The first ones are suited to carrying PDUs of constant size, whereas the latter category was designed to seamlessly adapt to input stream of any format, including continuous bit streams and variable-sized PDUs such as IP datagrams. GSE can avoid using MPEG2 packets as with MPE and ULE, which would be sub-optimal in the framework of DVB-S2. In effect, the GS flow is more suited to interactive services as it overcomes the inadequate MPEG2-TS constraints of constant bit rate and end-to-end delay. In addition, due to the large sizes of a BBFRAME payload (up to 40 times as long as an MPEG2 packet), datagram fragmentation occurs less often. Measures in the Internet network backbone show that the mean size of an IP datagram is about 500 bytes, which roughly amounts to 7000 \ 500 ≈ 14 IP datagrams carried in the longest available BBFRAME, against 2 or 3 fragmentations on the MPEG2-TS layer, and up to 10 in the case of ATM (Cantillo, 2008). GS streams are tailored into 21 possible BBFRAME frames, thus offering a variety of efficiency versus error protection compromises, and predefined sizes ranging from 384 to 1,779 bytes (short BBFRAMEs), or 2,001 to 7,274 bytes (for long BBFRAMEs). Consequently, all these characteristics of GSE result in IP datagrams being delivered more rapidly, efficiently and optimally in a cross-layer perspective, with reduced overhead and complexity (Cantillo, 2008).

A last technique deployed for the purpose of enhancing the DVB-S2 transmission performance is Spread Spectrum (SS) in its Direct Sequence (DS) variant. Besides its resistance to interference, jamming and multipath impairments, a quite powerful property of SS exploited in the framework of this study is its processing gain G_P, defined as the ratio of the spread bandwidth over the original bandwidth in dB. This processing gain can thus

be added to the signal side of the SNR calculation (Ayala et al., 2004). This fruitful property is due to the power–bandwidth trade-off that exists in any radio communication system: using a spread spectrum signal enables the system to operate at negative signal to noise ratios, thus allowing to deploy smaller terminals with reduced transmission power with respect to the non-spread case. This consequently means improved battery lifetime in the case of portable terminals, as well as an easier and quicker deployment of the terminals. Therefore the adequacy of SS is straightforward for emergency communications, in heavy rain environment (Yoon et al., 2008).
It must be pointed out that spread spectrum should not be confused with the standard DVB-S2 scrambling process. Spreading must be applied to each symbol of the PLFRAME including the PLHEADER and the pilot symbols, and is followed by scrambling, which applies a scrambling code to the spread signal (ETSI, EN 302 307, April 2009). Although there is some similarity between the two processes since both multiply an original signal with a pseudo-random noise (PRN), SS enlarges the bandwidth of the signal whereas scrambling does not since it only randomizes the ($I+jQ$) samples of the PLFRAME for energy dispersal (ETSI, EN 301 790 V1.5.1, May 2009). The Direct Sequence Spread Spectrum (DS-SS) block can only be inserted before the base-band filter and the modulator. In this technique, the PRN is directly applied to the data entering the carrier modulator. The modulator therefore sees a much higher bit rate, which corresponds to the chip rate of the PRN sequence. The purpose of modulating an RF carrier with such a code sequence is to produce a direct-sequence-modulated spread spectrum with $((sin\ x)/x)^2$ frequency spectrum, centered at the carrier frequency. There is no changing the point in the system where the DS-SS must be placed otherwise it would be a quite different SS technique. As a result, that means that a standard DVB-S2 transmitter cannot be used as a black box, but that the transmission chain must undergo some design adaptations, so as to conform to the provision envisaged in clause 5.1 of the mobile version of the DVB-RCS standard (ETSI, EN 301 790 V1.5.1, May 2009).

3.2 Channel modelling and link budgets

On the basis of such an air interface, a thorough point-to-point link budget analysis has been carried out for the four selected frequency bands Ku, Ka, Q and V in order to determine, first how much admissible capacity on the return link is provided for each MODCOD for the different user terminals, UTA, UTB, UTC and UTD as referred to in section *§2.1*; and second, which DVB-S2 MODCOD is required when a specific attenuation threshold is crossed (or equivalently to meet a given system availability). Link budget calculations and analyses have been performed both with and without SS, and have carefully focused on interference calculations, since the ratio *C/I* is of paramount importance in the link budget. However, link budget computation over the whole coverage is not needed, since it is assumed that the geographical area affected by the disaster is quite limited in extent (less than 100 km² which is the size of a large city such as Paris in France) which allows to consider that the link budget parameters are roughly constant over the area.
For this purpose, static (that is to say, temporal and spatial variability is not taken into account) statistical ITU-R channel models are needed to calculate the total atmospheric attenuation to be introduced into the link budgets.
Actually, the effect of each attenuation component may strongly depend on the link frequency. From Ku band to higher frequency bands, propagation through the satellite

channel is mostly impaired by rain, gas (oxygen), clouds, water vapor, amplitude scintillation, depolarization, and a degradation of the receive terminal figure of merit (*G/T*) due to an increase of its noise temperature (Castanet et al., 2003).
For example, oxygen attenuation is neglectable at Ku/Ka bands, but is permanently present in the atmosphere at higher frequencies, and exhibits close lines of strong absorption between 50 and 70 GHz, making satellite communications impossible at 60 GHz. Rain attenuation impairs the propagation channel most, with fade depth of up to 20 dB already at Ka band, for 0.01 % of the time (cf. Fig. 3).

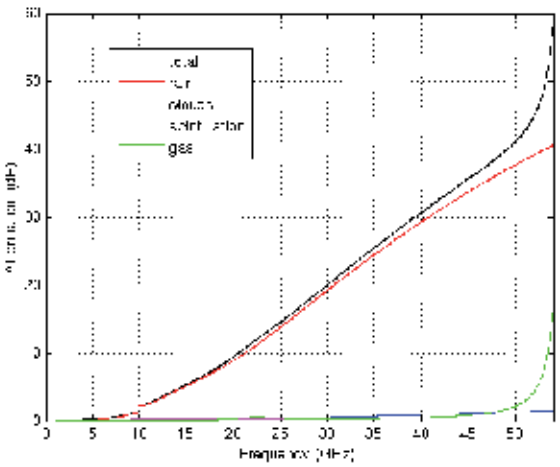


Fig. 3. Total attenuation vs the frequency at Louvain la Neuve, Belgium (Pech, 2003)

The acceptable maximum bit rates for each user terminal were determined according to the channel propagation state in Ku, Ka, Q and V bands, and the calculations also took into account the SS spreading factor *L* (*L*=1 corresponds to the case where SS is unused), as well as the size of the DVB-S2 PLFRAME (64,800 bits for normal frame, or 16,200 bits for short frame). A sample of obtained results is given in Table 1 for an uplink attenuation corresponding to 0.01 % of the time in the case of normal PLFRAME, and assuming that all *M* = 200 user terminals are of the same type. When there is only one user terminal transmitting, no spread spectrum interference is present. As a matter of fact, the link budget computations have to be refined by taking into account the effect of the mutual interferences due to spread spectrum. Such analysis was presented for instance by Viterbi in (Viterbi, 1985). It leads to considering the following signal energy per chip (of duration T_s and bandwidth $W_S = 1/T_S$) over noise ratio as the actual operating point of the system:

$$\frac{E_s}{N_0'} = \frac{E_b/N_0}{1+(M-1)(R_b/W_s)(E_b/N_0)} \tag{1}$$

where there are *M* identical user terminals transmitting to a same gateway at the same power and with spread spectrum; N_0 is the background noise spectral density received by each user; E_b is the received energy per information bit of duration T_b and rate $R_b = 1/T_b$ bps. N_0 is the sum of thermal noise N_0W_s and of the interference contribution of the other *M*-1 users. In other words, the total noise spectral density is related to N_0 through the equation:

$$N_0^{'} = N_0 + (M-1)E_s \quad (2)$$

Frequency band	Uplink atten. (dB)	Maximum bit rate (kbps) (L, MODCOD)			
		User terminal			
		UTA	*UTB*	*UTC*	*UTD*
Ku (14 GHz)	6.5	**889.7 (1,9)**	**889.7 (1,9)**	889.7 (1,9)	889.7 (1,9)
Ka (30 GHz)	24.7	236 (32,5)	**102 (32,6)**	3977 (8,28)	3977 (8,28)
Q (45 GHz)	43.5	0	0	3977 (16,28)	**591 (8,5)**
V (54 GHz)	71.5	0	0	1770 (8,13)	**29 (32,1)**

Table 1. Maximum allowed bit rates for an uplink attenuation corresponding to 0.01% of the time, with their required spread factors *L* and MODCODs. The 28 MODCODs of the DVB-S2 standard are involved, and referred to by their indices from 1 (QPSK 1/4) to 28 (32 APSK 9/10). By required MOCODs, it is meant the least robust, but robust enough MODCODs required by the terminals to establish a communication link.

From the results presented in Table 1, the following conclusions can be drawn:

- In Ku band, the portable/mobile terminals UTA and UTB yield quite sufficient bit rate performance, for a moderate channel attenuation level of 6.5 dB.
- In Ka band, these terminals can still transmit very reasonably: maximum bit rates of 236 kbps and more than 100 kbps are yielded by UTA and UTB terminals respectively.
- In Q band, the portable user terminals UTA and UTB can still transmit at low bit rates for an uplink attenuation of up to 30 dB, but require the activation of DS-SS. For A_{UL} greater than 32 dB, increasing L is not sufficient enough to establish the communications link.
- In Q and V bands, the link cannot be established for these low transmit power terminals even though SS is used when the uplink attenuation exceeds 32 dB. Therefore the use of a more powerful terminal, at least UTD (man-transportable terminal, with a transmit power of 5W) is necessary. For the latter, a maximum bit rate of almost 600 kbps is attainable for an uplink attenuation of 43.5 dB (Q band).
  - In V band, the link cannot be established anymore for the low to moderate transmit power terminals UTA, UTB and UTD even though SS is used. Nonetheless, minimum communications with VoIP could be established with an uplink attenuation of up to 60 dB, but at the cost of having a very high spread factor *L*, in the order of 5,000, which is not permitted by the bandwidth constraint. UTD is still usable with a maximum bit rate yielding 29 kbps. As the transmit power must be kept as low as possible, it is straightforward that the mini-gateway UTC does not need to be deployed to reach the goal of establishing a minimum emergency link.

4. Integrated simulation platform

4.1 General description

A software model of the enhanced DVB-S2 transmission link was developed in mixed standard C and Java languages within an open source software environment called Juzzle (SILICOM). The whole simulation platform is an integrated Excel/Matlab/Juzzle software package which also implements the C DISLIN (Michels, 2008) data plotting library. GSL GNU scientific library (Galassi, 2008) could also be incorporated in a future revision of the simulator. The former library is a set of C subroutines and functions that are used to display data graphically as post-processing functions within Juzzle, whereas the latter is a collection of routines written from scratch in C for numerical computing. These routines present a modern Applications Programming Interface (API) for C programmers, allowing wrappers to be written for very high level languages.

The philosophy of the integrated simulation platform tool is as follows:

Step 1: The user has the possibility of tuning the values of all the parameters of the point-to-point link budget in the Excel sheet, including the detailed characteristics of the terminals (location, EIRP, transmit power, etc.), the operating frequency band, and the DVB-S2 physical layer parameters (modem roll-off, type of PLFRAME, etc.). Once the system parameters have been validated, an Excel macro enables to create all the system and transmission parameters input files that feed both an off-line propagation pre-processing Matlab routine and the Juzzle core component.

Step 2: The offline pre-processing propagation module written in C must be run in order to generate the needed attenuation time series for the return link in Ka and EHF bands using the enhanced stochastic Maseng-Bakken model developed by Lacoste and Carrié (Lacoste, 2005; Lacoste et al., 2005; Castanet et al., 2008; Carrié et al.; Lemorton et al., 2007). The module uses the Juzzle C data-driven engine. An example of such an attenuation time-series is shown in Fig. 4 (taken from Pech et al., 2009). The synthesizer relies on the theoretical principles of the enhanced Maseng-Bakken model which was used by Lacoste (Lacoste et al., 2005; Castanet et al., 2008), and recently improved by Carrié (Carrié et al.). Actually, this new stochastic model enables to either stochastically interpolate initial samples to synthesize fast fluctuations of rain attenuation or generate "on-demand" rain attenuation events starting from the three parameters: duration of the event, maximum attenuation and position of the peak attenuation.
The model improvement lies in the characterization of the conditional probability $p(A(t)\,|\,A(t\text{-}Dt_1),\ A(t\text{+}Dt_2))$ which enables very fast interpolation or stochastic generation of rain events "on-demand". Three of the input parameters (m, s, A_{offset}) can be assessed for all link configurations using ITU-R recommended models. For the last parameter, b, a rough estimate equal to 10^{-4} s^{-1} can be used independently of the sampling rate and whatever the link considered in North-western Europe for elevation angles between 25° and 38° and frequencies between 12 GHz and 50 GHz (Carrié et al.).
The "event-on-demand" time series synthesizer derived from this first model offers the possibility to configure the maximum level and the duration of the rain attenuation events to be generated. This model exhibits the physical soundness of the real rain impairments

phenomenon and enables generation of rain attenuation events databases representative of experimental databases (Carrié et al.).

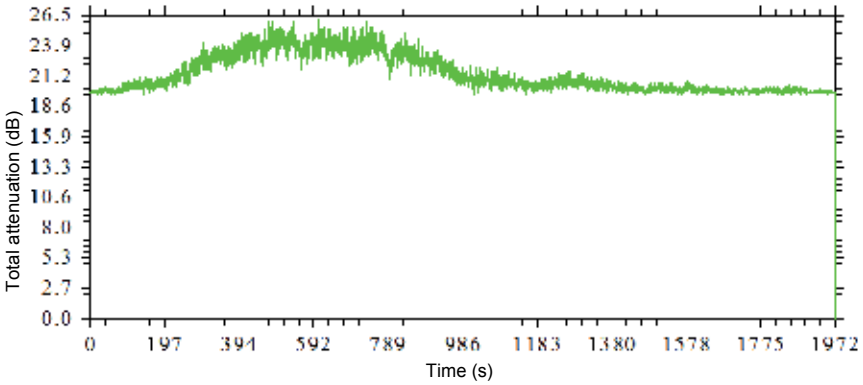


Fig. 4. Total attenuation time-series (in dB) generated by the offline module within the Juzzle simulator at 54 GHz

Step 3: The Juzzle on-line core component contains the complete high level system link model and enables to post-process and display a number of performance results.

Step 4: Last but not least, a Matlab routine is provided with the aforementioned simulator components, allowing to post-process the results file containing the collected performance parameters. It could be envisaged to replace this Matlab component in the future with a GSL-based module embedded in the Juzzle environment, as mentioned previously.

4.2 The Excel sheet link budget calculator

The Excel sheet link budget calculator was developed in order to help dimension the satellite emergency system to be simulated. It is a powerful tool allowing to:

- configure the whole system (satellite, air interface, characteristics of Earth terminals both from the emergency system and the primary system, etc.);
- compute the return link budgets both for the emergency system and the primary system.

The Excel sheet is composed of 12 window tabs. The tool was developed with a library of Visual Basic functions grouped into 10 distinct categories. It also calls an extern dynamic link library (DLL) named "propa.dll" (Lacoste, 2006; Lacoste, February 2006), which was developed by the CNES and comprises routines that enable to calculate the various propagation attenuation and scintillation components on an Earth-space link according to standardized ITU statistical prediction methods:

- Rec. ITU-R P.676 (ITU-R, 2005) due to dry air and water vapour;
- Rec. ITU-R P.840 (ITU-R, 1999): attenuation due to clouds;
- Rec. ITU-R P.618 (ITU-R, 2007): attenuation due to rain and melting layer, scintillation fade, total impairment.

4.3 The Juzzle simulation platform

The Juzzle simulation platform is composed of two separate and standalone Juzzle components (a propagation generator component and a processing core component), and an extern post-processing Matlab routine.

The architecture of the Juzzle link model is shown in Fig. 5 hereafter with all its components:

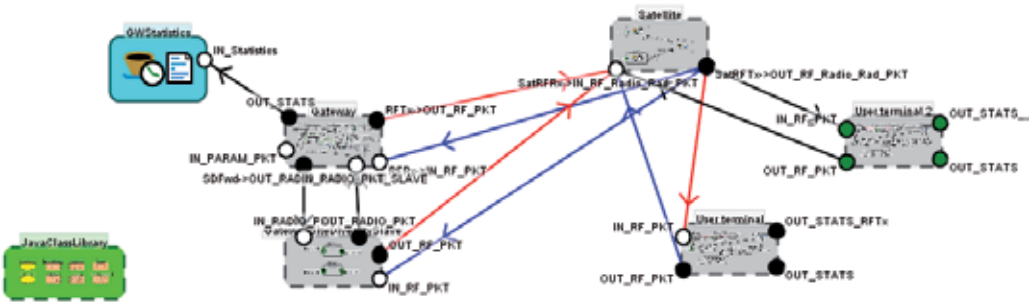


Fig. 5. Juzzle link model

The core processing module was developed in Java due to the portability, reusability, and object-oriented feature of the programming language. This module uses the Java event engine allowing to model the various delays within the system more easily.
The module models the emergency satellite system at high level as shown in Fig. 5, with the following nodes: two user terminals ("User terminal 1" and "User terminal 2"), the gateway, and the satellite, especially in terms of radio packet transmission, and its related link budget computation functions. As a result, the simulator is a powerful and complex integrated simulation platform handling more than 700 parameters, and being composed of 44 Java classes.
The component also possesses a statistics node whose tasks is to write down in an output ASCII text file a number of probe parameters allowing to compute performance statistics related to the link quality (BER, system margin, $C/(N_0+I_0)$, MODCOD, DS-SS spread factor, bit rate, etc.). As of today, exactly 30 parameters are monitored, but their number can be increased at will, with the restriction that the larger the number of parameters, the larger the results file. For instance, a simulation over a simulated time of 59,900 seconds produces a 1.428-GB output file.

The software simulation platform is described in more details in the next sections.

4.3.1 Traffic modules

Four types of traffic are modelled: VoIP, Web browsing, and SMS for the return link, and Internet aggregate traffic as an example of continuous forward link traffic. The corresponding underlying theoretical models employed are given in the following:

a) VoIP traffic

This source is modelled by a classic ON/ OFF state machine (Pech, 2003). The model is represented by an ON/OFF source which operates as an alternating process of talk and

silence periods which are distributed according to negative exponential laws with means a^{-1} and b^{-1} respectively.

b) Web browsing traffic model

Basically, the Web traffic model relies on an ON/OFF two-state model. The specificity of the model lies in its representing the traffic from a user behavioural point of view, by modeling a user's session, that is, a user Web request instead of focusing on the notion of a Web page. The overall traffic stream of a user is composed of a superimposition of the packet arrival processes of all TCP connections within a user's session. Moreover, it is a hierarchical model in which session, connection and IP levels are included (Choi & Limb, 1999).

c) SMS traffic

SMS traffic is simply modeled by a generator of SMS messages whose mode of operation is as follows (ETSI, 3GPP TS 23.040, October 2004; ETSI, 3GPP TS 23.038, September 2004): a random number N ranging from 1 to 10 (produced by a uniform random number generator) of concatenated 140-byte (amounting to 160 alphanumeric characters) SMS messages is generated every period T_{SMS} where T_{SMS} is distributed according to a negative exponential distribution. The number N is generated each time the SMS traffic source is invoked, that is, when the basic service is switched into. The scheme is repeated (ARQ: Automatic Repeat reQuest) every two minutes provided the basic service (i.e. a bit rate of less than 4 kbps) is still selected.

d) IP aggregate traffic source

The traffic source is modeled by a Fractional Gaussian Noise (FGN) using the Fast Fourier Transform (FFT). Paxson's FFT method (Paxson, 1997) consists in synthesizing a sample path having a same power spectral density (PSD) as a Fractional Gaussian Noise (FGN) process. This sample path can then be used in simulations as traces of real self-similar traffic. The algorithm is basically based on a fast approximation of the PSD of an FGN process using the FFT. The algorithm relies on an implementation of Paxon's self-similar generator written in ANSI C, and provided with in the form of a routine fft_fgn.c developed by Christian Schuler.

4.3.2 Adaptive strategy module

The adaptive strategy mechanism is at the heart of the simulator, and relies on a combination of several techniques aimed at ensuring high availability of the transmission links in spite of severe channel impairments in the selected frequency bands (*e.g.* about 20 dB in Ka band and more than 80 dB in Q band 0.01 % of the time). The techniques employed are: ACM, spread spectrum with varying DS-SS spread factor (L), gateway site diversity (SD) (to improve the downlink budget for the return channel when it is impaired by rain), and ARQ-like time diversity (TD). ARQ is only applied to SMSs, the user terminal automatically attempting at retransmitting the same SMS message at different times, when the channel conditions improve. Currently, retransmission is done on a pure deterministic basis, that is, periodically after a fixed time interval. Nevertheless, in the future, more

elaborate strategies could be devised exploiting fade duration, fade slope, and inter-fade statistics so as to implement an efficient method able to predict the channel attenuation in the medium term (several minutes).

It is moreover assumed that all of the user terminals of the emergency mission which are deployed over the geographical area affected by the disaster are of the same type and undergo the same ACM mode. This makes sense since the area is assumed to be quite limited in extent, more precisely smaller than 100 km², which allows to consider that the link budget and transmission parameters are roughly constant over the whole area. Nevertheless, it must be kept in mind that this uniform ACM scheme over the whole area is also a consequence of a simplistic assumption concerning the channel propagation modelling, in which no spatial variability is taken into account. Only a temporal variability is modeled using Carrié's enhancement of Lacoste's CNES/ONERA rain fading time series stochastic generator based on Maseng-Bakken model (Lacoste, 2005; Lacoste et al., 2005; Castanet et al., 2008; Carrié et al.; Lemorton et al., 2007).

The tabulated results of the link budget analysis previously mentioned enabled to carefully construct appropriate and efficient channel-aware link adaptive strategies, specific to each frequency band, the purpose of which is to ensure very high availability of the link, and enough capacity to authorize at least SMS communications even during a severe rain event (extremely strong storms), and VoIP as a minimum guaranteed service the rest of the time.
The adaptive strategy combines ACM with spread-factor-varying SS, the spread factor L being adapted (increased) before ACM activation depending on the attenuation level with respect to predefined thresholds, and with three different service classes being defined:

1. *Premium service*: VoIP, and Web browsing/SMS with a bit rate of about 180 kbps;
2. *Gold service*: VoIP/SMS with a bit rate of 4 to 16 kbps;
3. *Basic service*: SMS only with a bit rate of less than 4 kbps.

For each frequency band, these services are switched from each other through the DS-SS spread factor L (from 1 to up to several thousands). Within each service (or primary bit rate mode), the channel fluctuations are dynamically compensated for by full DVB-S2 ACM MODCOD switching.

Whenever the return link downlink channel is impaired by rain above a pre-determined attenuation threshold, and when the above compensation techniques still do not allow to close the link budget, SD is deployed on the feeder link. In other words, the downlink signal from the satellite is rerouted towards a slave gateway station which is located in a place not affected by rain. This brings an additional site diversity gain $G_{SD,DL}$ into the link budget, that can be calculated with the ITU-R P.618-9 Recommendation (ITU-R, P.618-9, January 2007).

It should now be noted that an optimal adaptive mechanism would obviously need to exploit, not only the static statistical characteristics of the channel fade components, but also their dynamic properties – in other words, an informed knowledge of fade / inter-fade duration, and fade slope ITU-R statistics and models (cf. Fig. 6). Furthermore, two additional considerations must be pondered when devising a relevant adaptive strategy to meet the system requirements for the targeted emergency mission: *(a)* the system total delay requires some short-term fade prediction to be made within a time horizon of several

seconds; *(b)* and a higher layer MODCOD switching criterion other than merely the E_s/N_0 or the BER, such as the PER could be used so as to yield better performance at the application level. The former consideration is due to a necessity of correctly estimating at time t the channel attenuation $A(t+\Delta t)$ at time $t+\Delta t$, knowing that otherwise using $A(t)$ to decide at time t which MODCOD is to be used would be mostly inadequate since quite rapid variations may occur in the channel attenuation time evolution. To properly model a rain attenuation predictor, an accurate link delay budget must be established, including in particular the 500-ms satellite Round-Trip Delay (RTD), and the channel estimator calculation time (Aroumont et al., 2006) (using the DVB-S2 pilot symbols which are spread over different frames). The second consideration relies on some error modelling that enables to derive the PER from a physical layer-oriented QoS parameter like the BER (Pech et al., 2002).

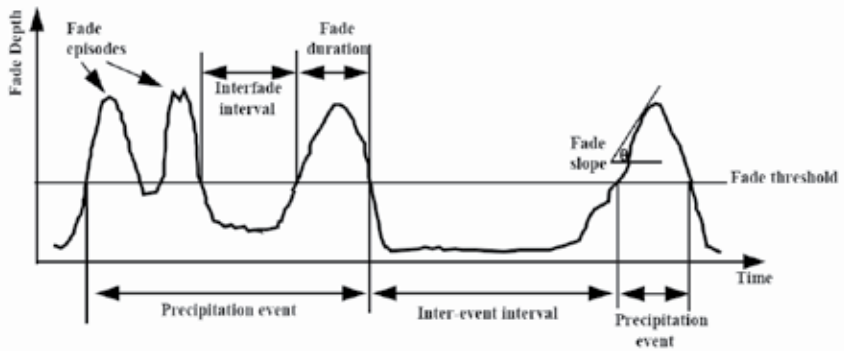


Fig. 6. Secondary statistics of fade

Currently, to obtain the attenuation to be applied to the channel, the simulator performs an attenuation prediction by linear interpolation or extrapolation, and takes into account the satellite geostationary Round Trip Propagation Delay (RTPD).

4.4 Extern Matlab post-processing module

This module is a post-processing module in charge of computing the transmission link performance mainly in terms of link availability, bit rate, error rates (BER or PER), delay, jitter, average and maximum packet sizes in order to characterize the application-level QoS, but also in terms of IP throughput, IP traffic characteristics (Tou et al., 2008) (inter-packet delay, IP delay variation, one way delay with maximum and minimum values, packet losses), and IP over frame efficiency over the DVB-S2 (Girault et al., 2008) (in connection with the encapsulation overhead and the MODCODs selected, and more generally the implemented adaptive strategy), as well as in terms of attenuation prediction errors. Some performance figures related to VoIP over DVB-S2 links have been made available in recent studies (Jegham et al., 2008), and could serve as a reference basis for the present project, in particular regarding a comparison between GSE and MPE/ULE encapsulation schemes. The module also allows to investigate the impacts of rain fading on the service performance. It provides a set of post-processing graphical plotting functions that will enable to optimize the performance of the adaptive strategy, tune its parameters (*e.g.* hysteresis and detection margins), and assess the ability of the proposed air interface to ensure a robust low bit rate satellite communication link for emergency mission.

5. Some results

The acceptable maximum bit rates for each user terminal were determined according to the channel propagation state in Ku, Ka, Q and V bands, and the calculations also took into account the SS spreading factor *L* (*L*=1 corresponds to the case where SS is unused), as well as the size of the DVB-S2 coded frame FECFRAME (64,800 bits for normal frame, or 16,200 bits for short frame). A sample of obtained results regarding to the admissible maximum bit rates and their associated MODCODs and spread factors was given in Table 1 which is copied into Table 2 hereafter for the sake of commodity:

Frequency band	**Uplink atten. (dB)**	**Maximum bit rate (kbps) (L, MODCOD)**			
		User terminal			
		UTA	***UTB***	***UTC***	***UTD***
Ku (14 GHz)	6.5	**889.7 (1,9)**	**889.7 (1,9)**	889.7 (1,9)	889.7 (1,9)
Ka (30 GHz)	24.7	236 (32,5)	**102 (32,6)**	3977 (8,28)	3977 (8,28)
Q (45 GHz)	43.5	0	0	3977 (16,28)	**591 (8,5)**
V (54 GHz)	71.5	0	0	1770 (8,13)	**29 (32,1)**

Table 2. Maximum bit rates allowable for an uplink attenuation corresponding to 0.01% of the time, with their required spread factors *L* and MODCODs

The results obtained without SS show in particular that the portable/mobile terminals UTA and UTB yield quite sufficient bit rate performance in Ku band, and for a moderate channel attenuation level of 6.5 dB. These terminals can still transmit very reasonably in Ka band. But in Q and V bands, the link cannot be established for these low transmit power terminals even though SS is used. Therefore, the use of at least UTD is necessary. For the latter, a maximum bit rate of almost 600 kbps is attainable for an uplink attenuation of 43.5 dB (Q band). In V band, UTD still is usable with a maximum bit rate yielding 29 kbps. The behaviour of the adaptive strategy for UTD in function of the total uplink attenuation is shown in Fig. 7. As the transmit power must be kept as low as possible, it is straightforward that the mini-gateway UTC does not need to be deployed in order to establish a minimal emergency link.

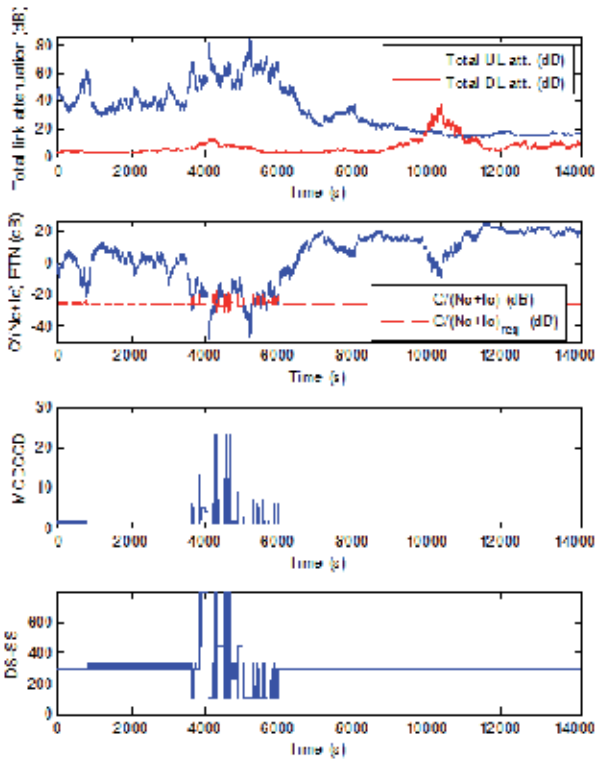


Fig. 7. Adaptive strategy behaviour for UTD in V band vs the total uplink attenuation

6. Conclusion

This chapter has presented the architecture of a multibeam, bent-pipe satellite system, used to rapidly establish a low bit rate link for emergency communications in Ku/Ka and Q/V bands. The characteristics of the proposed system have been described. An enhanced DVB-S2-like air interface has been proposed, involving DS-SS technique and other adaptive mechanisms such power control, and site diversity. Link budget analyses have shown that even though SS may be deactivated, very low transmit power user terminals UTA and UTB yield reasonable performance for the envisaged purpose up to the Q band, but for rather moderate channel attenuation values. This highlights the relevance and efficiency of a solution combining ACM, SS, bit reduction, and SD techniques in a new adaptive strategy in order to improve transmission performance. Such an adaptive strategy has been successfully implemented within the more general framework of an integrated Excel/Juzzle/Matlab software simulation platform designed to model the system in a high level cross-layer approach mixing propagation, physical layer, and higher layers components. Some selective results of the simulations carried out using this DVB-S2-like satellite link software simulator have been partially presented highlighting its ability to perform relevant analyses of the performance of the system both at the physical layer and the network level.

7. Acknowledgments

The work was carried out within the framework of an R&T project funded by the French Space Agency (CNES, Centre National d'Etudes Spatiales), Toulouse, France. The project developed in three phases: I. *Traffic and service characterization*, II. *Air interface and link budgets*, and III. *Computer link modelling and simulations*.

8. References

Ponia Pech, Puming Huang, Michel Bousquet. Low Bit Rate Satellite Link for Emergency Communications, *International Workshop on Satellite and Space Communications 2008* (IWSSC'2008), Toulouse, France, 1-3 October 2008.

Inmarsat. Inmarsat BGAN description: http://www.inmarsat.com/Services/Land/BGAN/default.aspx.

WISECOM. Deliverable 1.1-1; Survey of use cases, contract number 034673, 9 August 2007, available at: http://www.wisecom-fp6.eu/deliverables/D1.1-1_Survey_of_Use_Cases.pdf.

SILICOM. Official Juzzle Web site: http://www.juzzle.org/.

Thuan Nguyen, Ferit Yegenoglu, Agatino Sciuto, and Ravi Subbarayan. Voice over IP Service and performance in satellite, *IEEE Communications Magazine* **March 2001; 39**, pp. 164 - 171.

ITU-T G.114 Recommendation: Transmission systems and media, general recommendations on the transmission quality for the entire Internet telephone connection: One-way transmission time, number 114 in G. ITU-T, 2000.

ETSI TR 102 444 V1.1.1: Analysis of the short message service (SMS) and cell broadcast service (CBS) for emergency messaging applications; emergency messaging; SMS and CBS, February 2006.

ETSI. Digital cellular telecommunications system (Phase 2+); Universal mobile telecommunications system (UMTS); Technical realization of short message service (SMS) (3GPP TS 23.040 version 5.8.1 Release 5), October 2004.

ETSI. EN 302 307 V1.1.1: Digital video broadcasting (DVB); Second generation framing structure, channel coding and modulation systems for broadcasting, interactive services, news gathering and other broadband satellite applications, January 2004.

ETSI. ETSI TR 102 376 V1.1.1: Digital video broadcasting (DVB); User guidelines for the second generation system for broadcasting, interactive services, news gathering and other broadband satellite applications (DVB-S2), February 2002.

Juan Cantillo. Codage multi-couches pour systèmes de communication par satellites, Ph.D thesis report, Télécom Paris, Toulouse, France, 19 May 2008.

Mario Reyes Ayala, Edgar Alejandro Andrade Gonzáles, and José de Jesús Roa Franco. Performance in QPSK and BPSK synchronous sequence DS-CDMA satellite systems, *1st International Conference on Electrical and Electronics Engineering (ICEEE)*, Iaghouat, Algeria, 24-27 June 2004.

Ju-Hyun Yoon, Jae-Kwon Lee, Dae-Ig Chang, and Deock-Gil Oh. Spread spectrum technique for digital broadcasting system in rain environment, *10th International Conference on Advanced Communication Technology (ICACT),* Phoenix Park, South Korea, 17-20 February 2008.

Laurent Castanet, Laurent Féral, and Frédéric Lacoste. Adaptive coding-modulation techniques for Ka/Q-band systems; Review of available Ka/Q-band propagation measurements and models, ESA contract confidential technical report, ESA/ESTEC/Contract n°16533/02/NL/EC. Ref.: IFMT-ACM_TR1, issue 2, 3 March 2003.

Ponia Pech. Incidence de la prise en compte des effets de techniques de mécanismes de lutte contre les affaiblissements (FMT) en bande Ka sur la gestion des ressources dans un système d'accès multimédia par satellite géostationnaire, Ph.D thesis report, Supaero, Toulouse, 19 December 2003.

A. J. Viterbi. When Not to Spread Spectrum – a Sequel, *IEEE Communications Magazine* **April 1985; 23(4)**.

Helmut Michels, DISLIN 9.4, A Data Plotting Library, 15 October 2008. Downloadable at the following address: http://www.mps.mpg.de/dislin/document.html.

Mark Galassi, Jim Davies, James Theiler, Brian Gough, Gerard Jungman, Michael Booth, and Fabrice Rossi. GNU Scientic Library, Reference Manual, Edition 1.11, for GSL Version 1.11, 5 February 2008. Downloadable at the following address: http://www.gnu.org/software/gsl/.

Frédéric Lacoste. Modelling of the dynamics of the Earth-space propagation channel at Ka and EHF bands, Ph.D thesis report, SUPAERO, Toulouse, France, September 2005.

Hyoung-Kee Choi, John O. Limb. A Behavioral Model of Web Traffic, *Proceedings of the 7th International Conference on Network Protocols (ICNP'99)*, Toronto, Ontario, Canada, October/November 1999.

Alexander Klemm, Christoph Lindemann, and Marco Lohmann. Traffic Modeling and Characterization for UMTS Networks. *Proceedings of the Globecom, Internet Performance Symposium*, San Antonio TX, United States, November 2001.

Vern Paxson. Fast, Approximate Synthesis of Fractional Gaussian Noise for Generating Self-Similar Network Traffic, *Computer Communication Review* **October 1997; 27(5)**, pp.5-18.

W. Willinger, M. S. Taqqu, R. Sherman, D. V. Wilson, Self-Similarity Through High-Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level, *IEEE/ACM Transactions on Networking* **1997; 5**, pp. 71-86.

Frédéric Lacoste, Michel Bousquet, Laurent Castanet, Frédéric Cornet, Joël Lemorton. Improvement of the ONERA-CNES rain attenuation time series synthesiser and validation of the dynamic characteristics of the generated fade events, *Space Communication Journal* **2005; 20(1-2)**.

Laurent Castanet (editor) et al. Influence of the variability of the propagation channel on mobile, fixed multimedia and optical satellite communications, *SatNEx JA-2310 book*, ISBN 978-3-8322-6904-3, Shaker, 2008.

Guillaume Carrié, Frédéric Lacoste, Laurent Castanet. New "on-demand" channel model to synthesise rain attenuation time series at Ku-, Ka- and Q/V-bands, submitted to *International Journal of Satellite Communications and Networking*, Special issue on Channel Modelling and Propagation Impairment Simulation.

Anbazhagan Aroumont, Ana Bolea Alamanac, Laurent Castanet, Michel Bousquet, Stefano Cioni, and Giovanni E. Corazza. Performance of channel quality estimation algorithms for fade mitigation techniques with Ka/Q/V band satellite systems, *9th International Workshop on Signal Processing for Space Communication*, Noordwijk, Netherlands, September 2006.

Ponia Pech, Laurent Castanet, and Michel Bousquet. A prediction model to convert propagation distributions in statistics of quality of service performance parameters, *1st COST 280 International Workshop on Propagation Impairment Mitigation for Millimetre Wave Radio Systems*, Malvern, United Kingdom, 1-3 July 2002.

Ihsane Tou, Mathieu Gineste, Thierry Gayraud, Pascal Berthou. Quality of Service Evaluation in Satellite Systems, *International Workshop on Satellite and Space Communications 2008* (IWSSC'2008), Toulouse, France, 1-3 October 2008.

Nicolas Girault, Nizar Jegham, Nicolas Lerouge, Cédric Le Guern, André-Luc Beylot. OURSES: Efficiency of IP Encapsulation over DVB-S2 Links, *International Workshop on Satellite and Space Communications 2008* (IWSSC'2008), Toulouse, France, 1-3 October 2008.

Nizar Jegham, Nicolas Girault, Cédric Le Guern, Gilles Roussel, Stéphane Lohier, André-Luc Beylot. VoIP over a DVB-S2 ACM link, *International Workshop on Satellite and Space Communications 2008* (IWSSC'2008), Toulouse, France, 1-3 October 2008.

Joël Lemorton, Laurent Castanet, Frédéric Lacoste, Carlo Riva, Emilio Matricciani, Uwe-Carsten Fiebig, Max Van de Kamp, and Antonio Martellucci, Development and validation of time-series synthesizers of rain attenuation for Ka-band and Q/V-band satellite communication systems, *International Journal of Satellite Communications and Networking*, 2007; 25:575–601.

International Telecommunication Union/ITU Radiocommunication Sector, Recommendation ITU-R P.618-9 - Propagation data and prediction methods required for the design of Earth-space telecommunication systems, 1 January 2007.

International Telecommunication Union/ITU Radiocommunication Sector, Recommendation ITU-R P.676-6 - Attenuation by atmospheric gases (Question ITU-R 201/3), 2005.

International Telecommunication Union/ITU Radiocommunication Sector, Recommendation ITU-R P.840-3 -Attenuation due to clouds and fog (Question ITU-R 201/3), 1999.

Ponia Pech, Puming Huang, Michel Bousquet, Marie Robert, and Alban Duverdier. Simulation of an Adaptive Strategy Designed for Low Bit Rate Emergency Satellite Communications Links in Ku/Ka/Q/V Bands. *International Workshop on Satellite and Space Communications 2009* (IWSSC 2009), 10th-11th September 2009, Siena-Tuscany, Italy.

ETSI, « Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); Alphabets and language-specific information (3GPP TS 23.038 v6.1.0)", September 2004.

CNES freeware PROPAGATION Dynamic Link Library (DLL). http://logiciels.cnes.fr/PROPA/en/logiciel.htm.

Frédéric Lacoste. IUT-R propagation models software library. CNES, 13 February 2006. http://logiciels.cnes.fr/PROPA/en/usermanual.pdf.

ETSI EN 301 790 V1.5.1. Digital Video Broadcasting (DVB); Interaction channel for satellite distribution systems. May 2009.

Netdish VoIP over satellite. http://www.netdish.com/index.php?option=com_content&task=view&lang=en&id=37&Itemid=55.

ETSI. EG 202 339 v1.1.1: Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN); Definition of requirements on the functional architecture for supporting Emergency and Priority user services. September 2004.

ETSI. TS 102 181 V1.1.1: Emergency Communications (EMTEL); Requirements for communication between authorities/organizations during emergencies, December 2005.

ETSI. TS 102 182 V1.2.1: Emergency Communications (EMTEL); Requirements for communications from authorities/organizations to individuals, groups or the general public during emergencies. December 2006.

ETSI. TR 102 410 V1.1.1: Emergency Communications (EMTEL); Basis of requirements for communications between individuals and between individuals and authorities whilst emergencies are in progress. August 2007.

Katia Leconte, Martial Coulon et Marie-laure Boucheret. Analyse d'applicabilité de standards de télécom terrestres aux systèmes de télécommunication par satellite - Signal très large bande pour scénario d'urgence. Rapport d'étude du CNES, version 02, 1er septembre 2006.

Matthieu Dervin. Définition des missions et des scénarii de référence. Rapport d'étude du CNES, 1er septembre 2007.

IUT-R. IUT-R WP 4B: Draft new recommendation ITU-R: Characteristics of fixed-satellite service systems using wideband spreading signals, ref 4/92 (Rev.1)-E. 3 October 2006.

C. Mailhes, B. Comet, H. de Bernard, E. Campo, A. Prieto, and S. Bonhomme. Telemedicine Applications in OURSES Project, *International Workshop on Satellite and Space Communications 2008* (IWSSC'2008), Toulouse, France, 1-3 October 2008.

10

Mapping and Estimation of Chemical Concentrations in Surface Soils Using LANDSAT TM Satellite Imagery

B.B. Maruthi Sridhar and Robert K. Vincent
Department of Geology, Bowling Green State University, Bowling Green, OH 43403 United States

1. Introduction

Application of fertilizers, manures and other chemicals to agricultural fields is a continuous practice that is required for the optimum crop production. The rate of fertilizer application to soils is often determined by farmers based on the results of periodic soil sampling and analysis that were conducted prior to the cropping season. However conventional methods of soil sampling and analysis for soil variability in chemical characteristics are too time-consuming and expensive for multi-seasonal monitoring over large-scale areas. Hence, remote sensing has been used as an alternative method for determining and mapping the physical and chemical characteristics of the soil.

High resolution aerial imagery were used to map the organic carbon (Chen et al., 2000), clay content (Sullivan et al., 2005), organic matter and Bray-1 phosphorus concentration (Varvel et al., 1999), total soil phosphorus, copper and sulfur (Sridhar et al., 2009a,b) in bare soils. Dematte et al. (2003) reported that chemical variations in soil resulting from fertilizer applications can be detected, based on the intensity of reflectance. Several studies showed the use of spectral reflectance to determine the soil color (Post et al., 2000), texture and particle size distribution (Chang et al., 2001), soil moisture (Lobell & Asner, 2002), iron oxides (Ji et al., 2002), carbonates (Ben-Dor & Banin, 1990), clay (Ben-Dor & Banin, 1995), organic carbon (Dalal & Henry, 1986; Morra et al., 1991; Reeves et al., 2002) organic matter (Henderson et al, 1992) and soil phosphorus (Bogreckci & Lee, 2005, 2007; Sridhar et al., 2009a).

In this chapter, we are focuing on the use of LANDSAT TM data to map the changes in elemental concentrations of the soils that were applied with sewage sludge.

1.1. Land application of sewage sludge

Application of treated sewage sludges (biosolids) to agricultural land has become a prominent and acceptable method of waste disposal in recent years. Largely because of the 1991 ban on ocean dumping of sewage sludge and the higher costs of sludge incineration and land-filling, the application of sewage sludge to agricultural fields has gained momentum. Sewage sludge is the solid, semi-solid or liquid residue generated during the

treatment of domestic sewage in waste water treatment plants (USEPA, 2002). Sewage sludge that is land applied has to meet the treatment criteria and the standard pollutant and pathogen limits set by USEPA under Part 503 rule (USEPA, 2002).

Biosolids are known to improve soil physical characteristics (Epstein et al., 1975; Wei et al., 1985), increase the organic matter and cation exchange capacity and supply the nutrients required for crop growth (Sommers, 1977; Singh & Agrawal, 2008). However, the potential for excess application of biosolids, resulting in a build up of nitrogen, phosphorus (Mantovi et al., 2005), zinc, copper, lead (Mantovi et al., 2005; Udom etal., 2004; Nyamangara & Mzezewa, 1999) and cadmium (Bergkvist et al., 2003) in the surface soils of agricultural fields continues to be an area of concern. Accumulation of phosphorus at high concentrations is a major environmental concern, as it affects the water quality of lakes and rivers in the event of runoff (Shober & Sims, 2003).

Hence, there is an increasing need to continuously monitor the extent of soil contamination in biosolid-applied fields. Even though conventional methods of soil sampling and testing are being used for this purpose, they are often expensive, time-consuming and unsuitable for mapping soil contamination over large areas. The addition of soil contaminants as a result of biosolid application tend to be concentrated in surface soil samples (Mantovi et al., 2005; Bergkvist et al., 2003; Udom etal., 2004; Nyamangara & Mzezewa, 1999).

2. Methodology

2.1. Soil sampling and chemical analysis

Two adjacent agricultural fields, F34 and F11, that received a cummulative amount of 34 ton $acre^{-1}$ (76 Mg hac^{-1}) and 11 ton $acre^{-1}$ (25 Mg hac^{-1}) of Class B biosolids on a dry weight basis during the period of 1985-2002 were selected for this study (Fig 1). Soil samples were collected at 0, 30, and 50 cm depths from each of the 70 sampling locations across the two fields. These fields were selected because they are representative of large areas of northwest Ohio where land application of biosolids has become an important agricultural practice. The soil samples were collected on May 19 of 2005, one day prior to LANDSAT over pass, and the sampling locations were marked using a Trimble GeoExplorer (Trimble Navigation Limited, CA, USA) global positioning system (GPS) receiver. The collected soil samples were dried and passed through a 2 mm sieve. The moisture content of the surface soil samples was measured using the gravimetric method. The source of sewage sludge for the agricultural fields in the study area was the Oregon Waste Water Treatment Plant (OWWTP). The basic composition of the sewage sludge of OWWTP is typical of the bio-solids produced in Ohio, which is regulated within the limits set by the U.S. Environmental Protection Agency (USEPA) under the part 503 rule (USEPA 2002).

Soil samples (approximately 0.5 g) were digested with concentrated HNO_3, according to USEPA method SW846-3051A (USEPA, 1998) using a Mars Xpress microwave digestion unit (CEM, Matthews, NC, USA). The digested solution was filtered and then analyzed for As, B, Be, Ca, Cd, Cr, Cu, Fe, K, Mg, Mn, Mo, Na, Ni, P, Pb, S, Se, Si and Zn concentrations using inductively coupled plasma-optical emission spectrometry (ICP-OES) (IRIS Intrepid II, Thermo Scientific, Waltham, MA, USA). Quantification was achieved using matrix matched high and low concentration standards for each element. Internal quality controls and blanks were run every ten samples in order to quantify cross-contamination and recovery efficiencies.

Analysis of variance (ANOVA) was used to compare the accumulation of each element in the F34 and F11 fields using SAS version 9.1 statistical software (SAS Institute Inc., Cary, NC, USA). An alpha level of 0.05 was used to determine the significance.

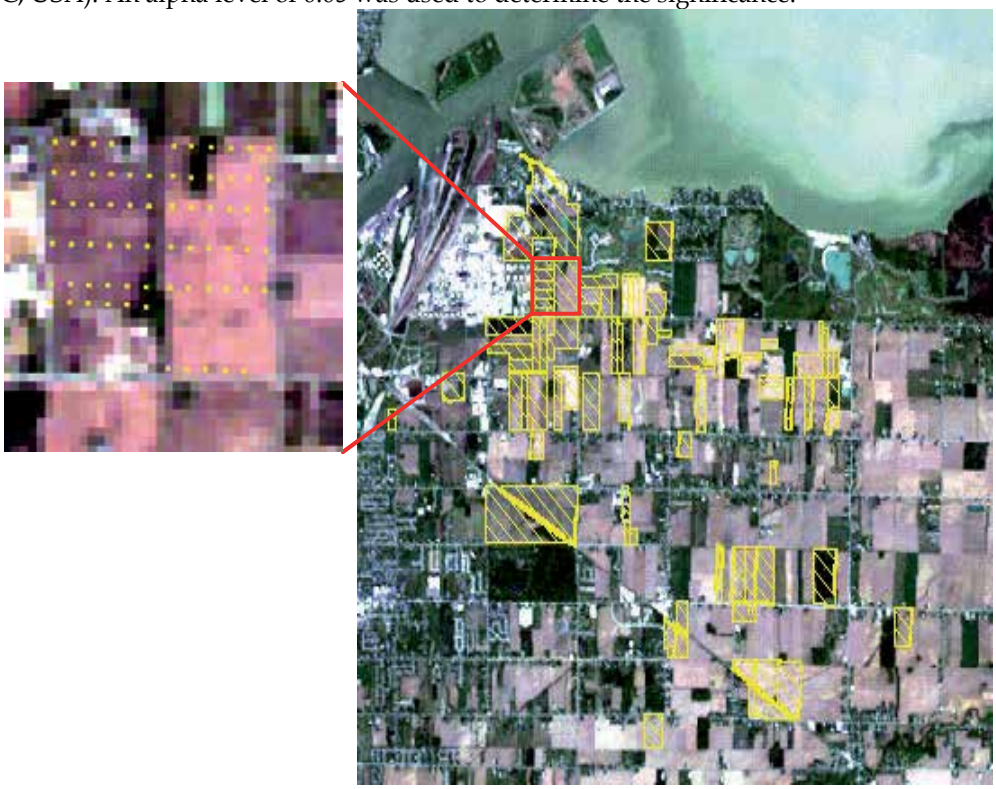


Fig. 1. The LANDSAT 5 TM natural color image (TM bands 1, 2, and 3 displayed as BGR, respectively) obtained on May 20, 2005 showing the eastern part of Lucas County in northwest Ohio; this area drains into Lake Erie, which is towards the northern side (top) of the image. The fields permitted for Class B biosolid application in the area are outlined in the image. The fields marked with red borders are the experimental fields used in this study. Soil sampling locations of the study area were shown as yellow dots in the insert image. (Reprinted from Science of Total Environment 47, Sridhar et al., 2009a, with permission from Elsevier)

2.2. LANDSAT data acquisition and analysis

The LANDSAT image frames of May 20 and June 5, 2005, covering the study area were downloaded soon after soil sampling. They were then processed with the ER Mapper image processing software, a commercial product of Earth Resources Mapping, Inc. The study area was located within the LANDSAT overpass region of Path 20, Row 31. The natural color image of the study area, overlaid with outlines of the fields permitted for Class B biosolid applications, is shown in Figure 1. The locations of all the 70 soil sampling points collected one day prior to LANDSAT 5 overpass were also shown separately in Figure 1 on the natural color image of the study area. The study site was dry, without any vegetation, implying that image spectral reflectances represent the spectral reflectance of soil. The

procedure for developing the GIS database of the Class B biosolid permitted fields in Wood and Lucas counties of northwest Ohio was reported in detail by McNulty (2005).
Based on the locations of the 70 soil samples, the dark object subtracted (DOS) pixel values corresponding to the LANDSAT TM bands 1-5 and 7 were derived from the original May 20, 2005 image. The spectral range of these LANDSAT TM bands are as follows: Band 1: 450-520 nm; Band 2: 520-600 nm; Band 3: 630-690 nm; Band 4: 760-900 nm; Band 5: 1550-1750 nm; and Band 7: 2080-2350 nm. The dark object of each spectral band is defined as one value less than the minimum digital number found in all the pixels of the image (Vincent et al., 2004). The detailed procedure for DOS and its effects on removal of atmospheric haze was given in Vincent (1997) and Vincent et al. (2004). From the DOS-corrected digital number (DN) values of the six LANDSAT single bands, 15 non-reciprocal spectral ratios were calculated. These spectral ratios are: $R_{2,1}$; $R_{3,1}$; $R_{3,2}$; $R_{4,1}$; $R_{4,2}$; $R_{4,3}$; $R_{5,1}$; $R_{5,2}$; $R_{5,3}$; $R_{5,4}$; $R_{7,1}$; $R_{7,2}$; $R_{7,3}$; $R_{7,4}$; $R_{7,5}$ where R represents the ratio and the numbers represent the LANDSAT TM band numbers (Vincent, 1997). The spectral ratios were calculated using the MINITAB statistical software (MINITAB Inc., State College, PA, USA).

2.3. LANDSAT TM best spectral ratio model development and validation

The relationship between the chemical concentrations of the surface soil samples and the DOS DN values corresponding to the six single bands and the 15 non-reciprocal spectral ratios were developed by regression analysis. Using the MINITAB regression analysis component the best subsets regression was employed, and only the top two models with highest R^2 adjusted values were chosen to report for each number of variables. The best subsets procedure was used for sequentially entering independent variables one at a time to improve the regression equation's predictive ability. The reported models from the best subset regression output were tested for autocorrelation with a Durbin-Watson (DW) statistical test (Durbin & Watson, 1951). This tests for autocorrelation in the input parameters. Finally, the model which had the highest R^2 adjusted and that also passed the DW test was selected as the best model for given inputs. This procedure was reported in detail elsewhere by Vincent (2000) and Vincent et al. (2004). The identified best model was then applied to the same May 20, 2005, LANDSAT image, which was used in developing the model to map the elemental concentration of the surface soils. The model was also applied and validated using the June 5, 2005, LANDSAT image, which was obtained 17 days after the soil sampling. In the LANDSAT images that were applied with the best model, masks were created to limit the display to only bare soil fields.

3. Results and discussion

3.1. Soil chemical concentration

The chemical concentration of the soils at 0, 30 and 50 cm depths and the moisture content of the surface soils in both the F34 and F11 treated fields are shown in Table 1. Among all the chemicals that were analyzed, the accumulation of Ba, Cd, Cu, S and P were significantly (p<0.05) higher in the surface soils of F34, compared to F11 (Table 1). There was no significant difference in the chemical concentrations at 30 and 50 cm depths among the F34 and F11 soils. Also, the moisture content of the surface soils in both the fields was similar (Table 1). The soils were of the prevalent latty silty clay type with surface soils having 40-55% of clay and 3-5% of organic matter (Soil Survey Staff, 2007).

Treatment	Soil Depth (cm)	Ba (mg/kg)	Cd (mg/kg)	Cu (mg/kg)	S (mg/kg)	P (mg/kg)	Moisture (%)
F34	0	161 (± 28.5)	5.7 (± 0.5)	55 (± 5)	405 (± 57)	2550 (± 625)	9.7 (± 7.3)
	30	103 (± 16.7)	5.7 (± 0.46)	37 (± 5.1)	197 (± 55)	796 (± 439)	
	50	97 (± 14.4)	5.6 (± 1.1)	35 (± 6.1)	154 (± 31)	557 (± 103)	
F11	0	98 (± 19.1)	3.6 (± 1.9)	37 (± 6.6)	265 (± 68)	988 (± 303)	5.5 (± 2.5)
	30	100 (± 13.3)	3.5 (± 0.9)	31 (± 5.1)	165 (± 49)	588 (± 177)	
	50	99 (± 13.5)	3.7 (± 0.9)	31 (± 4.9)	132 (± 51)	558 (± 149)	

Table 1. Chemical concentration of soils applied with 34 ton acre^{-1} (F34) and 11 ton acre^{-1} (F11) of class B biosolids. The given values are means ± standard deviation of 35 replicates. (Reprinted from Science of Total Environment 47, Sridhar et al., 2009a, with permission from Elsevier)

3.2. LANDSAT spectral ratio model

Regression equations were established to determine the chemical concentrations of Ba, Cd, Cu, S and P, which are significantly (p<0.05) higher in the surface soils of F34 compared to F11, using the DOS-corrected six TM bands and the 15 non-reciprocal spectral ratios. The best spectral ratio input models that pass the DW test of significance along with their R^2 adjusted and standard error values are given in Table 2.

Chemical	Best spectral ratio model	R^2 Adjusted (%)	SE (mg/kg)
Phosphorus	4156 -1690 (R51) + 2257 (R73)	67.9	531.2
Copper	75 -17.9 (R51) + 21.9 (R73)	59	6.9
Sulfur	507 - 14.7 (R51) + 214 (R73)	49.3	66.8

Table 2. Best spectral ratio input models for phosphorous, copper, and sulfur that pass the Durbin Watson test along with the values of R^2 adjusted and SE (standard error). Note: The models developed for Cd and Ba did not pass the Durbin Watson test at 5% level of significance. (Reprinted from Science of Total Environment 47, Sridhar et al., 2009a, with permission from Elsevier).

None of the single band models passed the DW test. Phosphorus had the highest R^2 adjusted value (67.9%) followed by Cu among the chemical attributes that passed the DW test (Table 2) and are considered for mapping with LANDSAT TM data. Hence, only the P and Cu results were shown in this paper. The P values obtained from chemical analysis of the 70 surface soil sampling locations versus the predicted values of P for the same locations

obtained by applying the P spectral ratio model P(mg/kg) = 4156 -1690 (R51) + 2257 (R73) to the May 20, 2005, LANDSAT TM frame is given in Figure 3.

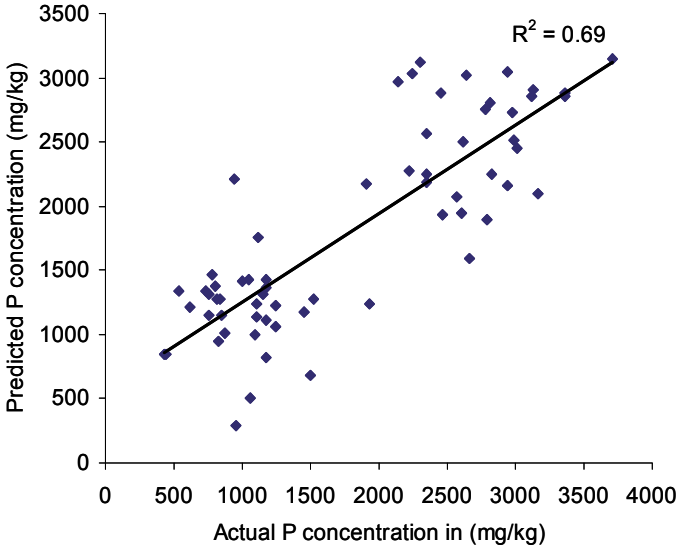


Fig. 3. Actual versus predicted P concentration (in mg/kg) of surface soil samples using the dark object subtracted best P spectral ratio model being applied to the LANDSAT 5 TM frame of May 20, 2005, which was also used for developing the model. (Reprinted from Science of Total Environment 47, Sridhar et al., 2009a, with permission from Elsevier)

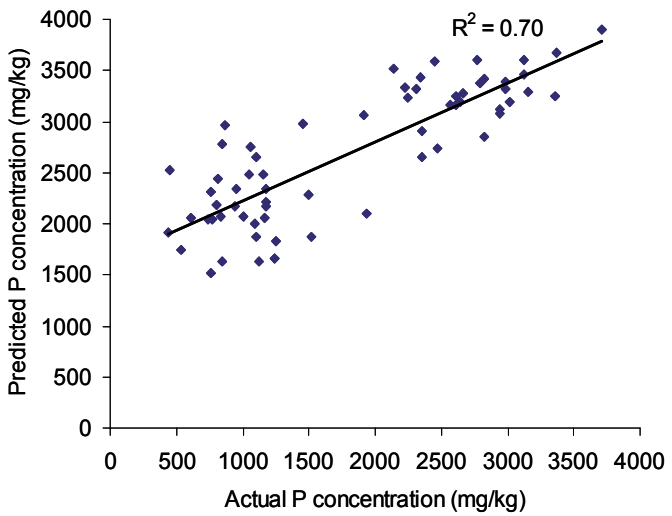


Fig. 4. Actual versus predicted P concentration (in mg/kg) of surface soil samples using the dark object subtracted best P spectral ratio model being applied to the LANDSAT 5 TM frame of June 5, 2005. (Reprinted from Science of Total Environment 47, Sridhar et al., 2009a, with permission from Elsevier).

The P spectral ratio model was also applied to the June 5, 2005 LANDSAT image frame and the predicted P values were plotted against the P values obtained by the soil analysis (Fig. 4). The model performed well in predicting the P concentrations of surface soil when applied to either of the LANDSAT TM images.

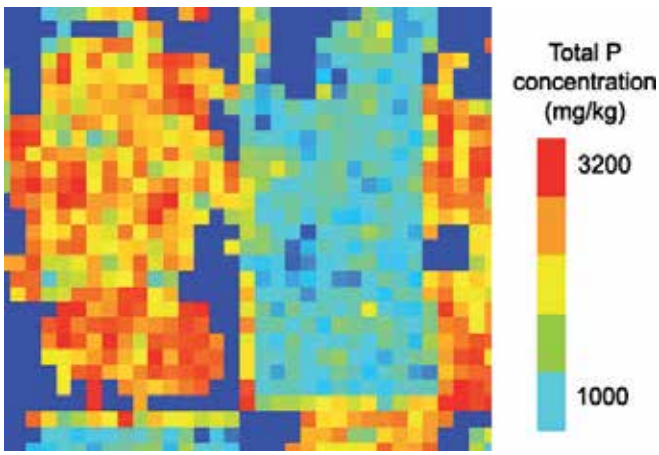


Fig. 5. Image showing the total P concentration (mg/kg) in surface soil samples of fields F34 (left side of the image) and F11 (right side of the image), displayed as red (high P content) to Turquoise (low P content), obtained by applying the best P spectral ratio model to the LANDSAT 5 TM frame of May 20, 2005 which was also used for developing the model. (Reprinted from Science of Total Environment 47, Sridhar et al., 2009a, with permission from Elsevier)

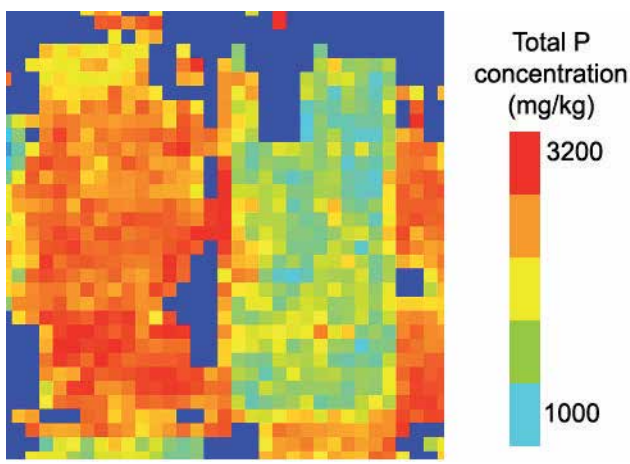


Fig. 6. Image showing the total P concentration (mg/kg) in surface soil samples of fields F34 (left side of the image) and F11 (right side of the image), displayed as red (high P content) to Turquoise (low P content), obtained by applying the P spectral ratio model to the LANDSAT 5 TM frame of June 5, 2005. (Reprinted from Science of Total Environment 47, Sridhar et al., 2009a, with permission from Elsevier)

The application of the best P spectral ratio model to the LANDSAT 5 TM frame of May 20, 2005, which was also used in developing the model, is shown in figure 5. The redder color in this image corresponds to higher amounts of P in surface soil. Figure 6 shows the image of the same spectral ratio model that was developed using the LANDSAT 5 frame of May 20, 2005, being applied to the LANDSAT 5 frame of June 5, 2005. Note that the P concentration in the F34 field is significantly higher than the F11 field in both the images (Fig 5 and 6). The application of the best P spectral ratio model to the May 20, 2005, LANDSAT TM image, showing the part of the watershed that drains into Lake Erie, is given in Figure 7. The fields outlined in this figure are permitted for Class B biosolid application.

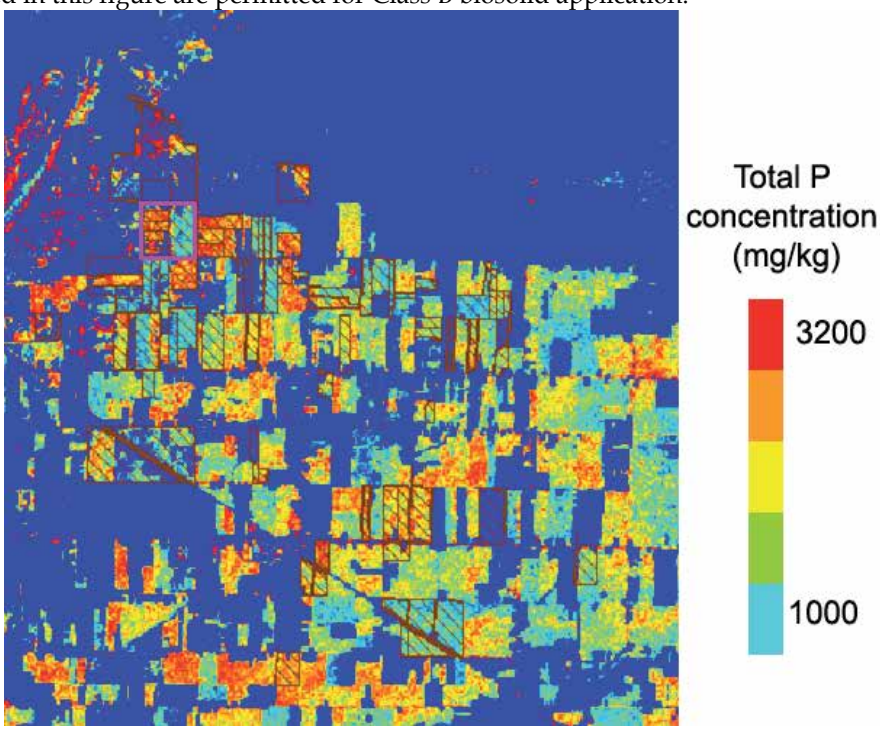


Fig. 7. The total P concentration (mg/kg) in surface soil samples of the bare soil fields in the eastern part of the Lucas County of northwest Ohio, which is a part of the drainage basin of Lake Erie, which is located at the northern side (top) of the image. The image is obtained by applying the P spectral ratio model to the LANDSAT 5 TM frame of May 20, 2005 which was used for developing the model. The fields permitted for Class B biosolid application in the vicinity are outlined in the image. The fields marked with a pink border are the experimental fields used in this study. Not all of the permitted fields in this image would have recently received sewage sludge, and some non-permitted fields could have recently received other types of fertilizers. (Reprinted from Science of Total Environment 47, Sridhar et al., 2009a, with permission from Elsevier)

The Cu values obtained from chemical analysis of the 70 surface soil sampling locations versus the predicted values of Cu for the same locations obtained by applying the Cu spectral ratio model Cu(mg/kg) = 75 -17.9 (R51) + 21.9 (R73) to the May 20, 2005, LANDSAT TM frame is given in Figure 8. The Cu spectral ratio model was also applied to the June 5,

2005 LANDSAT image frame and the predicted Cu values were plotted against the Cu values obtained by the soil analysis (Fig. 9). The model performed well in predicting the Cu concentrations of surface soil when applied to either of the LANDSAT TM images.

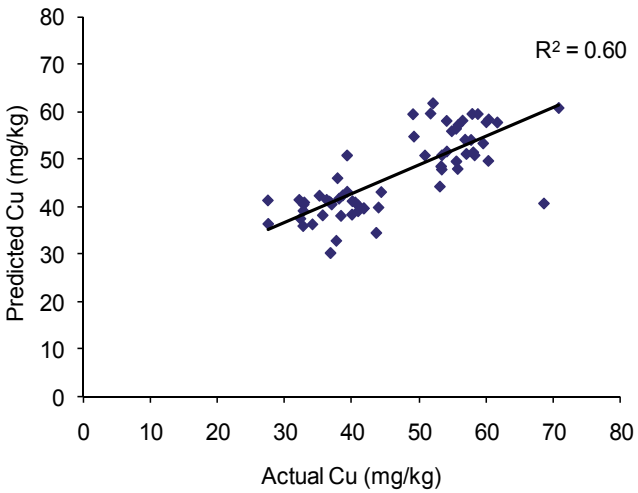


Fig. 8. Actual versus predicted Cu concentration (in mg/kg) of surface soil samples using the dark object subtracted best Cu spectral ratio model being applied to the LANDSAT 5 TM frame of May 20, 2005, which was also used for developing the model.

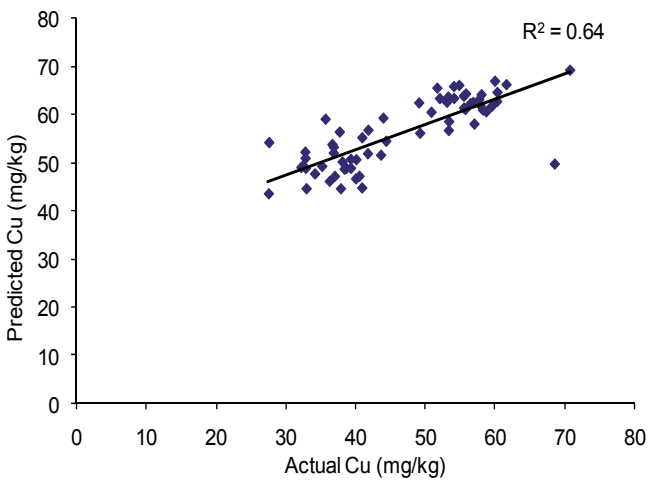


Fig. 9. Actual versus predicted Cu concentration (in mg/kg) of surface soil samples using the dark object subtracted best Cu spectral ratio model being applied to the LANDSAT 5 TM frame of June 5, 2005.

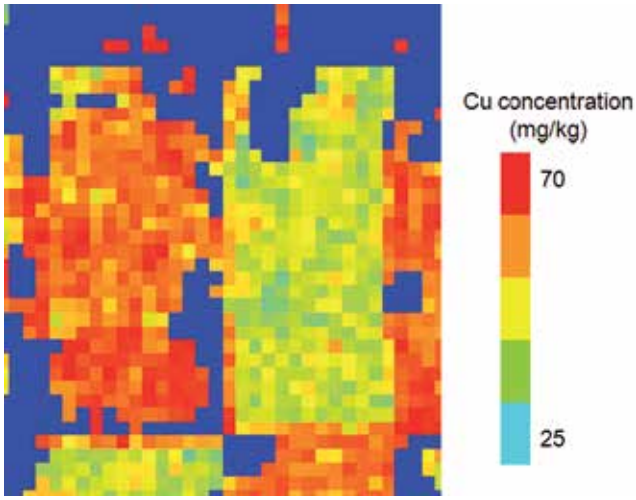


Fig. 10. Image showing the total Cu concentration (mg/kg) in surface soil samples of fields F34 (left side of the image) and F11 (right side of the image), displayed as red (high Cu content) to Turquoise (low Cu content), obtained by applying the best Cu spectral ratio model to the LANDSAT 5 TM frame of May 20, 2005 which was also used for developing the model.

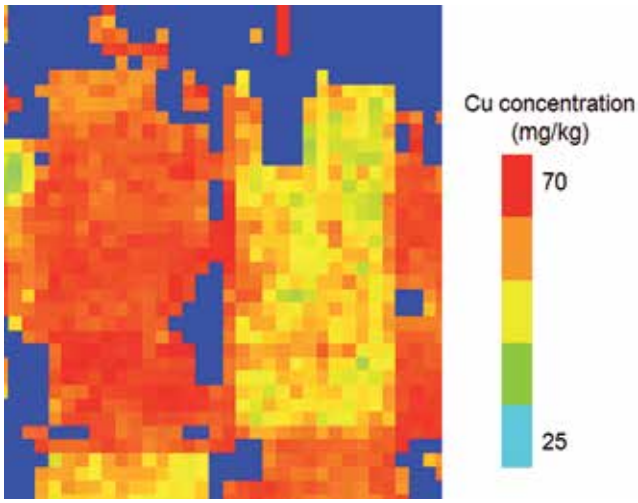


Fig. 11. Image showing the total Cu concentration (mg/kg) in surface soil samples of fields F34 (left side of the image) and F11 (right side of the image), displayed as red (high Cu content) to Turquoise (low Cu content), obtained by applying the P spectral ratio model to the LANDSAT 5 TM frame of June 5, 2005.

The application of the best Cu spectral ratio model to the LANDSAT 5 TM frame of May 20, 2005, which was also used in developing the model, is shown in figure 10. The redder color in this image corresponds to higher amounts of Cu in surface soil. Figure 11 shows the image of the same spectral ratio model that was developed using the LANDSAT 5 frame of May 20, 2005, being applied to the LANDSAT 5 frame of June 5, 2005. Note that the Cu

concentration in the F34 field is significantly higher than the F11 field in both the images (Fig 10 and 11). The application of the best Cu spectral ratio model to the May 20, 2005, LANDSAT TM image, showing the part of the watershed that drains into Lake Erie, is given in Figure 12.

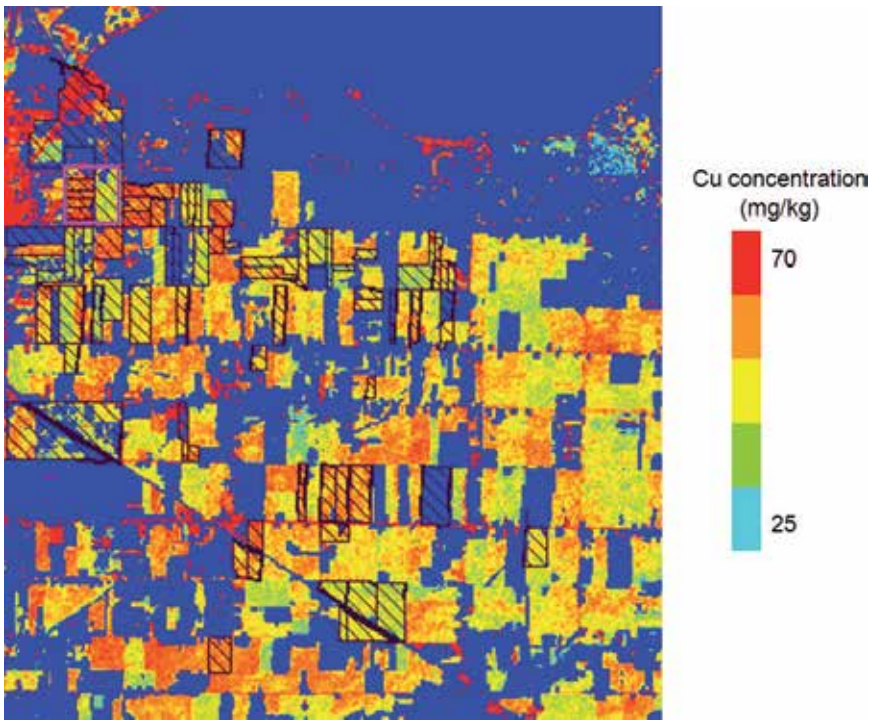


Fig. 12. The total Cu concentration (mg/kg) in surface soil samples of the bare soil fields in the eastern part of the Lucas County of northwest Ohio, which is a part of the drainage basin of Lake Erie, which is located at the northern side (top) of the image. The image is obtained by applying the Cu spectral ratio model to the LANDSAT 5 TM frame of May 20, 2005 which was used for developing the model. The fields permitted for Class B biosolid application in the vicinity are outlined in the image. The fields marked with a pink border are the experimental fields used in this study.

The analytical results showed that the accumulation of P and Cu in surface soil samples of F34 was about 2.6 and 1.5 times higher than for the F11 soils. This confirms the report of Chang et al. (1983) that five continuous years of biosolids application in two California soils at 0, 22.5, 45 and 90 ton per hectare increased the total P concentration of surface soil (0-15 cm) from 515-540 mg/kg to 1092-1312, 1657-2163 and 2617-3470 mg/kg, respectively. Similarly Maguire et al. (2000) reported that the concentration of total soil P in surface soils (0-20 cm) of biosolid amended soils was 738 mg/kg, or nearly double the values in unamended soils, where the total soil P was 403 mg/kg. High concentrations of Cd and Cu in the surface soils of F34 compared to F11 agree with the reports of Nyamangara and Mzezewa (1999), that the long-term application of biosolid increases the accumulation of Cd and Cu in the surface soils

4. Discussion

LANDSAT TM data can be used to estimate and map some chemical characteristics of soils, such as total phosphate content, as shown in this study. Our results enabled us to conclude that remotely sensed imagery of bare soil fields can be used to quantify and map the spatial variation of total P and Cu concentration in surface soils. The technology is simple enough to be applied to the entire watershed. The P and Cu spectral ratio models were more robust and reliable than the single band input models and can be applied to bare soil fields with low (<13%) soil moisture.

Nanni and Dematte (2006) have successfully employed LANDSAT TM data to estimate the sand, silt, clay, organic matter, cation exchange capacity (CEC) and sum of cations in Brazilian soils. They derived spectral reflectance values from the corrected LANDSAT image to develop multiple regression equations in order to estimate the different physical and chemical characteristics of the soils; however, no soil maps were presented in that study (Nanni & Dematte, 2006). Our study is significant because it represents the first successful effort in using LANDSAT TM data to estimate and map P concentration in surface soils. We also successfully validated the P spectral ratio model by applying to another LANDSAT image obtained on June 5, 2005.

Aerial imagery was used to map the organic carbon (Chen et al., 2000), clay content (Sullivan et al., 2005), organic matter and Bray-1 phosphorus concentration (Varvel et al., 1999) and LANDSAT TM imagery was used to estimate the physical and chemical properties (Nanni & Dematte, 2006) of surface soils in the previous studies. However, the algorithms developed in these studies were based on the reflected image intensity values of the soils, which required correction for atmospheric haze with atmospheric models before applying the algorithms to another date. The P spectral ratio model developed in our study is based on the DOS-corrrected spectral ratios and is more robust that any model that could be derived from a combination of single spectral bands. Vincent et al. (2004) showed that the DOS spectral ratio models were more robust than single band models and can be applied with reasonable accuracy to different times of data collection, though their subject was cyanobacteria blooms in lakes or streams, and the present study is about P concentrations in bare soils on dry land.

By applying our P and Cu spectral ratio model, we can identify and map the P and Cu concentration in surface soils as a result of biosolid application. Because P accumulation in soils can also result from the application of biosolids, animal manures, and man-made fertilizers, this research has significant implications in identifying the fields with high concentrations of surface soil P, thus helping implementation of P-based management practices on agricultural fields, with an aim toward reduction of P runoff into nearby surface water bodies.

Shober and Sims (2003) reported that twenty-four of the states and territories in the United States now have regulations to restrict the land application of biosolids, based on phosphorus concentration in soil. Thirteen of these 24 states have established actual numerical limits for soil test phosphorus (STP), with an aim to cease the application of biosolids once these limits are reached. As the total soil P and STP are linearly related to each other (Allen & Mallarino, 2006), our P spectral ratio model can be used to monitor P levels in surface soils.

The remainder of this chapter is devoted for the presentation of a sample application which demonstrates how quantitative remote sensing of soil P and Cu concentrations can be used

to meet the variety of information needs. This particular case study shows the use of our remote sensing models for monitoring a recent environmental disaster.

4.1 Sample application

a. Coal fly ash spill

A vast amount of fly ash, a by-product of coal incineration, was spilled over a wide area on December 22, 2008, at approximately 0100 hours EST, when an earthen wall of a fly ash disposal pond broke at the Tennessee Valley Authority (TVA) Kingston Fossil Plant, located at Harriman, Roane County, Tennessee (TVA, 2009; Fig 13). This is the largest environmental disaster of its kind involving a coal fly ash spill in the US history (New York Times, 2008). Approximately 5.4 million cubic yards of fly ash were spilled over an area of 300 acres outside the ash storage ponds. Fly ash was also spilled into the Emory River and the tributaries that flow into the Emory River, which serves as a source of drinking water (TVA, 2009; New York Times, 2008; Fig 13).

Fig 13. The aerial image of the Tennessee fly ash spill area acquired soon after the spill. (Image courtesy of Tennessee Valley Authority and can be viewed at http://www.tva.gov/kingston/photo_gallery/pages/123002.htm)

Accumulation of large amounts of fly ash as a result of coal combustion is becoming a major environmental concern in the United States. With the Nation's increasing demand for energy, the increase in coal combustion has required the disposal of large quantities of coal combustion residues. Approximately 131 million tons of coal combustion residues were generated within the United States in 2007, of which 36% was disposed in landfills, 21% in surface impoundments, and the remainder was reused for beneficial purposes. There are approximately 300 landfills and 300 surface impoundment facilities used by 440 coal fired plants across the United States (Fig 13).

The physical and chemical properties of the fly ash generated at a given coal-fired plant depends on the nature of the coal being burned, type of the combustion method, and the storage and handling methods involved. In general, fly ash is substantially rich in many elements (including heavy metals) and is usually stored in either landfills or artificial lagoons on open land. The general chemical composition of fly ash is given in Table 3 (Page *et al*., 1979).

Element	Fly ash concentration
Major elements (%)	
Al	0.1-17.3
Ca	0.11-22.2
Fe	1-29
Mg	0.04-7.6
Na	0.01-2.03
K	0.15-3.5
S	0.002-0.004
P	0.04-0.8
Trace elements (ppm)	
As	2.3-6300
B	10-618
Cd	0.7-130
Co	7-520
Cr	10-1000
Cu	14-2800
Hg	0.02-1
Mn	58-3000
Mo	7-160
Ni	6.3-4300
Pb	3.1-5000
Se	0.2-134
Zn	10-3500

Table 3. General chemical characteristics of Fly ash. (After Page et al., 1979; Reprinted from Photogrammetric Engineering and Remote Sensing 75, Sridhar et al., 2009b)

Mapping the spatial distribution of chemical concentrations as a result of fly ash spill is important for determining its effect on human and environmental health and for conducting remediation and recovery efforts. Previous studies have shown that fly ash pollution can be monitored through several methods: conventional soil sampling and analysis, measurement

of Sr-87 to Sr-86 ratios in plant samples growing in the vicinity, and measurement of ferromagnetic mineral concentrations present within the fly ash by magnetic mapping. However, all of these methods are based on point measurements at the ground-level, requiring intensive sampling of large-scale contaminated areas, which is expensive, time-consuming and sometimes not even possible as the areas may be sufficiently hazardous (e.g., poor slope stability) and inaccessible to scientists and engineers.

b. LANDSAT TM Images

Remote sensing has been used as an alternative method for determining and mapping the physical and chemical characteristics of exposed soils. In this study, we have used LANDSAT Thematic Mapper (TM) sensor data to identify and map the TVA's fly ash spill in Tennessee. The Thematic Mapper and Enhanced Thematic Mapper plus (ETM+) sensors were aboard the LANDSAT 5 and 7 respectively, each of which has a 16-day repeat cycle with a pixel size of 30 m X 30 m for bands 1-5 and 7. Because the two satellites have an 8-day offset from one another, it is possible to obtain coverage of a given ground area every 8 days, though the TM sensor aboard LANDSAT 7 (called ETM+) has approximately a quarter of its pixels missing, due to the loss of its scan line converter since May 31, 2003. In order to monitor the before and after fly ash spill events of the Kingston fly ash plant, we downloaded LANDSAT 5 TM data acquired on November 20, 2008 (32 days before fly ash spill); December 22, 2008 (about 9 hours after fly ash spill) and February 1, 2009 (40 days after fly ash spill) for this study. The LANDSAT TM images were then processed with the ER Mapper image processing software, a commercial product of Earth Resources Mapping, Inc., as mentioned earlier for the sewage sludge study.

For the purpose of measuring the P and Cu concentrations in the exposed fly ash and soil in the spill vicinity, the spectral ratio models given in Table 2 were applied to each of the three LANDSAT TM images covering the fly ash spill study area.

c. Results of Image Analysis

Figure 14 shows the natural color images, total P and Cu concentration images of the fly ash-spill vicinity for each of the following dates of LANDSAT TM overpass: November 20, 2008, December 22, 2008, and February 1, 2009, representing the periods of 32 days before the fly ash spill, 9 hours after, and 40 days after the fly ash spill, respectively. The fly ash is seen as a grey-colored area in the December 22, 2008 natural color imagery. Image interpretation of a standard natural color image (TM bands 1, 2, and 3 displayed as blue, green, and red, respectively) yields limited success for mapping the fly ash deposits because the contrast between fly ash and background soils is small in visible wavelength regions. The LANDSAT TM derived P concentrations range from 1,500 to 4,500 ppm and the Cu concentrations from 25 to 75 ppm in the November, 2008 and February, 2009 images. We have assigned turquoise, yellow, orange and red colors to increasing P and Cu concentrations as indicated by the respective color coding bars in Figure 1. The intense red color in the December 22, 2008 image clearly stands out due to larger areas of high P and Cu concentrations exposures at the surface, compared to the November 20, 2008 image acquired before the fly ash spill. The 25 to 75 ppm range of Cu concentrations derived from the December 22, 2008 image correlates well with the 29.9 to 69.4 ppm range of Cu values obtained through laboratory analysis from eight (8) ash sampling locations collected from

December 23, 2008 through January 5, 2009 in the fly ash spill vicinity (Tetra Tech, 2009). The P concentrations were not reported through soil analysis (Tetra Tech, 2009).

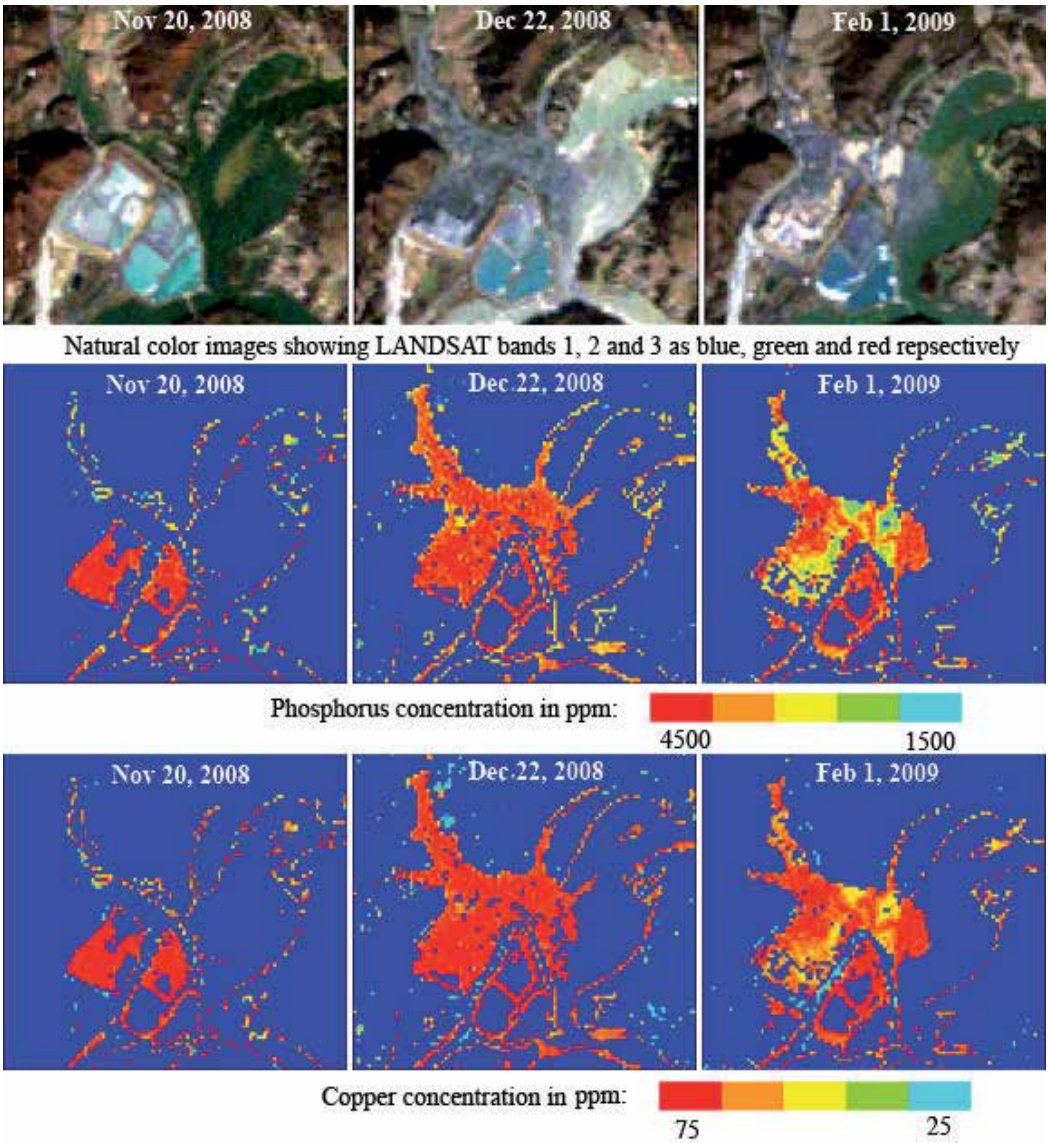


Fig. 14. The LANDSAT TM images of the Tennessee fly ash spill area acquired on November 20, 2008 (32 days before fly ash spill); December 22, 2008 (9 hours after fly ash spill) and February 1, 2009 (40 days after fly ash spill), respectively. The images in the rows 1-3 represent the natural color image, surface P concentration image, and surface Cu concentration image, respectively. (Reprinted from Photogrammetric Engineering and Remote Sensing 75, Sridhar et al., 2009b, with permission from the American Society for Photogrammetry & Remote Sensing)

In November, before the fly ash spill, the high concentrations of P and Cu were shown to be confined to the fly ash ponds. In December, a large extent of increased P and Cu concentrations is clearly evident towards the northern and north eastern side of the ash holding ponds as shown in Figure 14. These high P and Cu concentration features agree with the fly ash spill towards the northern and north eastern sides of the ash holding ponds (TVA, 2009). In February, the P and Cu concentrations mapped with the same LANDSAT TM algorithms were lower than those recorded in December. As an action item for dust suppression, TVA spread 85 tons of winter rye grass, 650 tons of straw, and other erosion-control mulch on the surface of the fly ash spill areas from January 3 through January 15, 2009 (TVA, 2009). For areas of where grass and straw were spread over, the February 1, 2009 image shows lower P and Cu concentrations, compared to the December 22, 2008 image. This series of images shows that LANDSAT TM data can be used to quantitatively monitor the remediation and recovery efforts of P and Cu contaminated sites.

5. Conclusions

In summary, the application of P and Cu mapping algorithms from LANDSAT TM images have allowed us to map and estimate the increase in surface concentrations of these two elements in surface soil samples. These results show the effective use of multispectral satellite image analysis in determining surface soil chemical concentrations. Compared to point measurements made by ground-based soil sampling and chemical analysis, the satellite-based measurements have a great advantage in mapping the spatial distribution and concentration of elements and chemical compositions over time. As LANDSAT TM has 30-m spatial resolution, there is a measurement by the satellite for every 1/5th of an acre (the area covered by one pixel) during each overpass, which would be cost prohibitive for point measurements that require manual soil sample collection and laboratory analysis of each sample. Satellite monitoring of surface soil elemental concentrations for environmental purposes can surpass point measurements on the ground and measured in the laboratory, at least for P and Cu. We have looked at several other elements for similar algorithms but found none as good as these two. The limitations of our spectral ratio models is that they were developed using bare soil fields that had a low soil moisture (<13%) and the soils tested were of the prevalent soil type (Latty Silty Clay) in this region. Also the P and Cu concentration that went into these models were in the ranges of 700-3000's and 30-60's respectively. Thus we have no evidence that these models will be accurate for soils with P and Cu concentrations outside these limits and with different soil types of higher soil moisture contents. Further research is being done to test the model to determine the P and Cu concentration in surface soils of other soil types in this region.

6. References

Allen, B.L. & Mallarino, A.P. (2006) Relationships between extractable soil phosphorus and phosphorus saturation after long-term fertilizer or manure application. *Soil Sci Soc Am J*, Vol.70, pp.454-463.

Ben-Dor, E. & Banin, A. (1990) Near infrared reflectance analysis of carbonate concentration in soils. *Appl. Spectrosc*, Vol.44, pp.1064-1069.

Ben-Dor, E. & Banin, A. (1995) Near infrared analysis (NIRA) as a method to simultaneously evaluate spectral featureless constituents in soils. *Soil Sci*, Vol.4, pp.259-270.

Bergkvist, P.; Jarvis, N.; Berggren, D. & Carlgren, K. (2003) Long-term effects of sewage sludge applications on soil properties, cadmium availability and distribution in arable soil. *Agric. Ecosyst. Environ*, Vol.97, pp.167-179.

Bogrekci, I. & Lee, W.S. (2005) Spectral soil signatures and sensing phosphorus. *Biosyst. Eng*, Vol.92, pp.527-533.

Bogrekci, I. & Lee, W.S. (2007) Comparison of ultraviolet, visible and near infrared sensing for soil phosphorus. *Biosyst. Eng*, Vol.96, pp.293-299.

Chang, A.C.; Page, A.L.; Sutherland, F.H. & Grgurevic, E. (1983) Fractionation of phosphorus in sludge affected soils. *J. Environ. Qual*, Vol.12, pp.286-290.

Chang, C.W.; Laird, D.A.; Mausbach, M.J. & Hurburgh, Jr C.R. (2001) Near-infrared reflectance spectroscopy – principal component regression analysis of soil properties. *Soil Sci. Soc. Am. J*, Vol.65, pp.480-490.

Chen, F.; Kissel, D.E.; West, L.T. & Adkins, W. (2000) Field-scale mapping of surface soil organic carbon using remotely sensed imagery. *Soil Sci. Soc. Am. J*, Vol.64, pp.746-753.

Dalal, R.C. & Henry, R.J. (1986) Simultaneous determination of moisture, organic carbon, and total nitrogen by near infrared reflectance spectrophotometry. *Soil Sci. Soc. Am. J*, Vol.50, pp.120-123.

Dematte, J.A.M.; Pereira, H.S.; Nanni, M.R.; Cooper, M. & Fiora, P.R. (2003) Soil chemical alterations promoted by fertilizer application assessed by spectral reflectance. *Soil Sci*, Vol.168, pp.730-747.

Durbin, J. & Watson, G.S. (1951) Testing for serial correlation in least squares regression: II. *Biometrica,* Vol.38, pp.159-178.

Epstein, E.; Taylor, J.M. & Chaney, R.L. (1975) Effects of sewage sludge on some soil physical properties. *J. Environ. Qual*, Vol.4, pp.139-142.

Henderson, T.L.; Baumgardner, M.F.; Franzmeieir, D.P.; Stott, D.E & Coster, D.C. (1992) High dimensional reflectance analysis of soil organic matter. *Soil Sci. Soc. Am. J*, Vol.56, pp.865-872.

Ji, J.F.; Balsam, W.L.; Chen, J. & Liu, L.W. (2002) Rapid and quantitative measurement of hematite and goethite in the Chinese Loess-Paleosol sequence by diffuse reflectance spectroscopy. *Clay Miner*, Vol.50, pp.208-216.

Lobell, D.B. & Asner, G.P. (2002) Moisture effects on soil reflectance. *Soil Sci. Soc. Am. J*, Vol.66, pp.722-727.

Maguire, R.O.; Sims, J.T. & Coale, F.J. (2000) Phosphorus fractionation in biosolids–amended soils: Relationship to soluble and desorbable phosphorus. *Soil Sci. Soc. Am. J*, Vol.64, pp.2018-2024.

Mantovi P, Baldoni G, Toderi G. Reuse of liquid, dewatered, and composted sewage sludge on agricultural land: effects of long-term application on soil and crop. Water Res. 2005; Vol.39, pp.289-296.

McNulty, W.S. (2005) The creation of a GIS database and the determination of sludge's spectral signature in an agricultural setting. M.S. Thesis. Department of Geology, Bowling Green State University, Bowling Green, OH, USA.

MINITAB Statistical Software Version 15. State College, PA: MINITAB Inc.; 2007-2008.

Morra, M.J.; Hall, M.H. & Freeborn, L.L. (1991) Carbon and nitrogen analysis of soil fractions using near infrared reflectance spectroscopy. *Soil Sci. Soc. Am. J*, Vol.55, pp.288-291.

Nanni, M.R. & Dematte, J.A.M. (2006) Spectral reflectance methodology in comparison to traditional soil analysis. *Soil Sci. Soc. Am. J*, Vol.70, pp.393-407.

New York Times (2008) Tennessee ash flood larger than initial estimate, Published on December 26, 2008, URL: *http://www.nytimes.com/2008/12/27/us/27sludge. html?_r=2&ref=us* (last date accessed: 1 July 2009).

Nyamangara, J. & Mzezewa, J. (1999) The effect of long-term sewage sludge application on Zn, Cu, Ni and Pb levels in a clay loam soil under pasture grass in Zimbabwe. *Agric. Ecosyst. Environ*, Vol.73, pp.199-204.

Page, A.L.; Elseewi, A.A. & Straughan, I.R. (1979) Physical and chemical properties of fly ash from coal-fired power plants with special reference to environmental impacts. *Residue Reviews*, Vol.71, pp.83-120.

Post, D.F.; Fimbres, A.; Matthias, A.D.; Sano, E.E.; Accioly, L.; Batchily, A.K. & Ferreira, L.G. (2000) Predicting soil albedo from soil color and spectral reflectance data. *Soil Sci. Soc. Am. J*, Vol.64, pp.1027-1034.

Reeves, III J.B.; McCarty, G.W.; Mimmo, T. (2002) The potential of diffuse reflectance spectroscopy for the determination of carbon inventories in soil. *Environ. Pollut*. Vol.116, pp.S264-S277.

SAS Institute. SAS Software Version 9.1. Cary, NC: SAS Institute, Inc.; 2002-2003.

Shober, A.L. & Sims, J.T. (2003) Phosphorus restrictions for land application of biosolids: current status and future trends. *J. Environ. Qual*, Vol.32, pp.1955-1964.

Singh, R.P. & Agrawal, M. (2008) Potentail benefits and risks of land application of sewage sludge. *Waste Management,* Vol.28, pp.347-358.

Soil Survey Staff, Natural Resources Conservation Service, United States Department of Agriculture. Web Soil Survey. http://websoilsurvey.nrcs.usda.gov. 2007.

Sommers, L.E. (1977) Chemical composition of sewage sludges and analysis of their potential use as fertilizers. *J. Environ. Qual*, Vol.6, pp.225-232.

Sridhar, B.B.M. & Vincent, R.K. (2009) Mapping and estimation of phosphorus and copper concentrations in fly ash spill area using LANDSAT TM Images. *Photogrammetric Engineering and Remote Sensing*, Vol.75, Nb.9, pp.1030-1033.

Sridhar, B.B.M.; Vincent, R.K.; Witter, J.D. & Spongberg, A.J. (2009) Mapping the total phosphorus concentration of biosolid amended surface soils using LANDSAT TM data. *Science of Total Environment*, Vol.47, pp.2894-2899.

Sullivan, D.G.; Shaw, J.N. & Rickman, D. (2005) IKONOS imagery to estimate surface soil property variability in two Alabama physiographies. *Soil Sci. Soc. Am. J*, Vol.69, pp.1789-1798.

Tennessee Valley Authority (2009) Corrective Action Plan for the TVA Kingston Fossil Plant Ash Release, URL: *http://www.tva.gov/kingston/cap/ TVA_ Corrective _ Action_ Plan_Draft _D5.pdf* (last date accessed: 1 July 2009).

Tetra Tech EM Inc. (2009) Final CERCLA emergency response report, Kingston fossil plant Fly ash response Harriman, Roane County, Tennessee Tetra Tech Inc. Soil and ash sampling results Kingston fossil fly ash response Harriman, Roane County, Tennessee, URL: *http://www.epaosc.org/sites/4642/files/ erfinal reporttvakingston.pdf* (last date accessed: 1 July 2009).

U. S. Environmental Protection Agency. Standards for the use or disposal of sewage sludge. Office of Water, Washington D. C, 2002.

U. S. Environmental Protection Agency. Test methods for evaluating solid waste. Office of Solid Waste and Emergency Response, Washington D. C, 1998.

Udom, B.E.; Mbagwu, J.S.C.; Adesodun, J.K. & Agbim, N.N. (2004) Distributions of zinc, copper, cadmium and lead in a tropical ultisol after long-term disposal of sewage sludge. *Environ. Int*. Vol.30, pp.467-470.

Varvel, G.E.; Schlemmer, M.R. & Schepers, J.S. (1999) Relationship between spectral data from an aerial image and soil organic matter and phosphorus levels. *Precision Agric*, Vol.1, pp.291-300.

Vincent, R.K. (1997) Fundametals of geological and environmental remote sensing. Prentice Hall, Upper Saddle River, NJ.

Vincent, R.K. (2000) Forecasts of monthly averaged daily temperature highs in Bowling Green, Ohio from monthly sea surface temperature anomalies in Eastern Pacific ocean during the previous year. *Photogramm. Eng. Remote Sens*, Vol.66, pp.1001-1009.

Vincent, R.K.; Qin, X.; McKay, R.M.L.; Miner, J.; Czajkowski, K.; Savino, J. & Bridgeman, T. (2004) Phycocyanin detection from LANDSAT TM data for mapping cyanobacterial blooms in Lake Erie. *Remote Sens. of Environ*, Vol.89, pp.381-392.

Wei, Q.F.; Lowery, B. & Peterson, A.E. (1985) Effect of sludge application on physical properties of a silty clay loam soil. *J. Environ. Qual*, Vol.14, pp.178-180.

OLFISH - A complete, paperless solution for the collection, management and dissemination of marine data

Dr. Amos Barkai, Fatima Felaar, Karl Geggus,
Zahrah Dantie and Arno Hayes
Olrac (Ocean Land Resource Assessment Consultants)
South Africa

1. Introduction

Fisheries management is continually frustrated by the lack, or poor quality, of critical data on fishing operations (catches, duration, gear, locations and relevant environmental conditions). While quantitative methods for managing fisheries have developed with considerable complexity, the quality of the available data remains an obstacle for meaningful advances in fisheries management. There are a number of aspects to the problem, not all of which are technical. A culture of protecting catch data and disinformation is common amongst fishers, fishing companies and even formal state-run offices, and significant education is needed in order to change this culture.

Another problem is the poor quality of historic data in many fisheries around the world. Much energy is wasted and important opportunities lost because of the uncertainty surrounding crucial historic data. For example, there are typically many factors related to catch-per-unit-effort data, a key index of trends in resource abundance, which are not recorded, and hence cannot be incorporated into statistical analyses. Frequently, these missing data are crucial to management decisions. For scientists, unreliable data leads to a poor basis for stock assessment models and management programs. For industry, the lack of sound data significantly reduces its fishing efficiency, since past performance cannot be studied properly. As a result, poor management decisions based on unreliable analyses are made, often with substantial cost and risk to fish resources and the fishing industry.

Although there is presently greater awareness amongst scientists and fisheries managers about the importance of collecting fishing data, there is still confusion about exactly which data are needed, and how to collect and store them. It is common for skippers to record scientific data on one form, for shore managers to use another for commercial purposes, and for skippers to keep separate fishing logbooks. These data are then transferred to different computer systems, often complex spreadsheets, or, sometimes, are left in paper format in large, inaccessible books and files. There is a degradation in the quality of data because of the multi-stage process of transcription from handwritten logbook sheets to paper forms and then to computer databases.

So even when good will is present, technically, the absence of a flexible and comprehensive system for capturing essential data during fishing operations is a major obstacle. A large amount of logistical and environmental data is lost simply because of the difficulty of recording this information easily in real time. This is despite the advent of a complex array of sensory equipment available in the bridge of modern fishing vessels. As a result, environmental patterns become part of a skipper's experience and seldom, if ever, become formally available to scientists or managers of fishing operations.
The most logical first point of data entry, through the fishing vessel skipper, should occur in digital format directly into a computer. One of the difficulties with fisheries data is the complexity of the logical linkages between the different types of data. Any reasonable approach to the problem requires the use of modern relational databases which are able to address the multidimensional complexity of the problem.
In order to address many of the problems described above, Olrac (www.olrac.com) a South African company, has developed a data collection and management system it has named Olfish (www.olfish.com) for the specific use of operators and managers in the marine environment with a special focus on the commercial fishing industry.

2. Electronic Logbooking

2.1 Benefits

An obvious approach to the "data crises" is to bring modern data and information technology (Elog) to the marine environment in general, and to the commercial and recreational fishing industry in particular. Providing fishers with accurate yet easy to use data logging tools could potentially transform the entire fishing fleet and the fishers community into the largest surveyors group of the marine environment in the world. The calibre of data produced through electronic logbooks has the potential to benefit all sectors of the fishing industry, from the fishers themselves to seafood consumers, resource managers, scientists and government enforcement agencies in between. In addition, the international shift towards a greater emphasis on output control measures, such as annual catch limits (ACL's) and total allowable catches (TAC's), requires the implementation of sophisticated catch monitoring tools in order to allow for a near-real-time auditing of catch versus TAC. However, it is important to note that the benefits of electronic data logging go beyond merely adhering to regulations. It is crucial for the industry to realise that it will ultimately be the greatest beneficiary of accumulated good quality data. A few obvious benefits derived from the collection of a large amount of accurate data in a near-real-time environment are:

2.1.1 Better Stock Assessments

The accuracy and timely delivery of electronically recorded data will allow for more exact indications of catch in a current year. In the past, due to the delays of paper-based reporting, incomplete data from preceding years has been used to estimate the TAC of the following year. The uncertainty associated with such calculations has resulted in conservative stock assessments which lead to overly restrictive TAC's. This means that there is often an over-discard of fish which would otherwise be commercially viable. Electronic data logging would allow for up-to-date and accurate data to be used for TAC estimation, thus eliminating much uncertainty and adding weight and justification to the TAC's allocated.

2.1.2 Better Targeting and Gear Utilization

The security and verification features of electronic logbooks, (see Security and Data Integrity – 5.4.7 below), as well as multimedia photographic and video utilities (see Multinote Taker and Notebook – 6.2 below), can potentially replace the role of an observer onboard a vessel. This can then be adapted into an incentive scheme for improved gear and fishing-ground selectivity, thus reducing unintended bycatch. Capturing target species may also lead to a decrease in days at sea, which is often beneficial for the skipper.

2.1.3 Faster Transmission

Faster transmission of information from sea to shore allows for "on-time reaction", i.e. decisions made on a regulatory, managerial, commercial or environmental basis are relevant to what is actually happening at sea. Back-logged, non-electronic reporting means that any event at sea is only registered on shore sometimes up to several weeks after it has occurred. Responding to month-old information, particularly in an ever-dynamic ocean environment, is practically pointless. Faster transmission will have a substantially positive effect on, for example, quota management, conservation and even commercial decision-making.

2.1.4 Catch Prediction and Management

Built-in analytical tools available within electronic logbook software (see Olfish Explorer - 6.4 below) are able to harness historical information stored in their electronic databases to help fishers calculate and predict fish migration, fishing hotspots etc. This greatly increases efficiency in a number of fields, such as targeting areas and the selection of fishing grounds and techniques. Similarly, fishers will be able to avoid "dry" areas, maximizing their time at sea and ultimately reducing discarding rates.

2.1.5 Traceability

Traceability is the ability to locate the source and "journey" of a fish from ocean to supermarket shelf. Legal organizations, such as the Marine Stewardship Council, prohibit fish without certification logos from entering the market. Such logos are obtained through traceability, i.e. proving that the fish in question had been caught in a certified area under certified conditions. Electronic data logging makes traceability a simple and speedy process. Information from the vessel at sea can be efficiently transmitted to market authorities who can then clear the catch for sale. Furthermore, electronic data logging allows for a highly detailed recording of catch information. Thus, catch freshness can easily be proven, increasing its market value. Buyers then benefits from being able to accurately estimate the shelf-life of the product they have bought. None of this would be possible without verifiable and immediate traceability.

2.2 The Olfish System: A Short Overview

Olfish is a third-generation, data logging and data management, software tool which was initially developed for the commercial fishing industry, but now provides a complete solution for the collection, management and reporting of other vessel-based activities, such as commercial and recreational fishing trips, oceanographic surveys, marine inspections, cargo and service trips, surveillance missions, etc.

The present version of Olfish includes three basic components in order to cater for the entire data flow, from at-sea collection to the generation and dissemination of reports.
The onboard, data collection component named Olfish Dynamic Data Logger (Olfish-DDL) is a standalone data collection tool installed onboard the vessel's PC.
Olfish-DDL also has a shore component which is identical to the vessel version but allows data from many vessels to be stored and viewed on one user-interface. This component is available in two versions:

a. A Single Fleet unit that aggregates operational data from vessels of a single company or organisation.
b. A Meta-Shore unit, which can aggregate operational data received from many shore units. The Meta-Shore unit can be used by a government agency, fishing association or even a union of states to manage data from a number of countries/states.

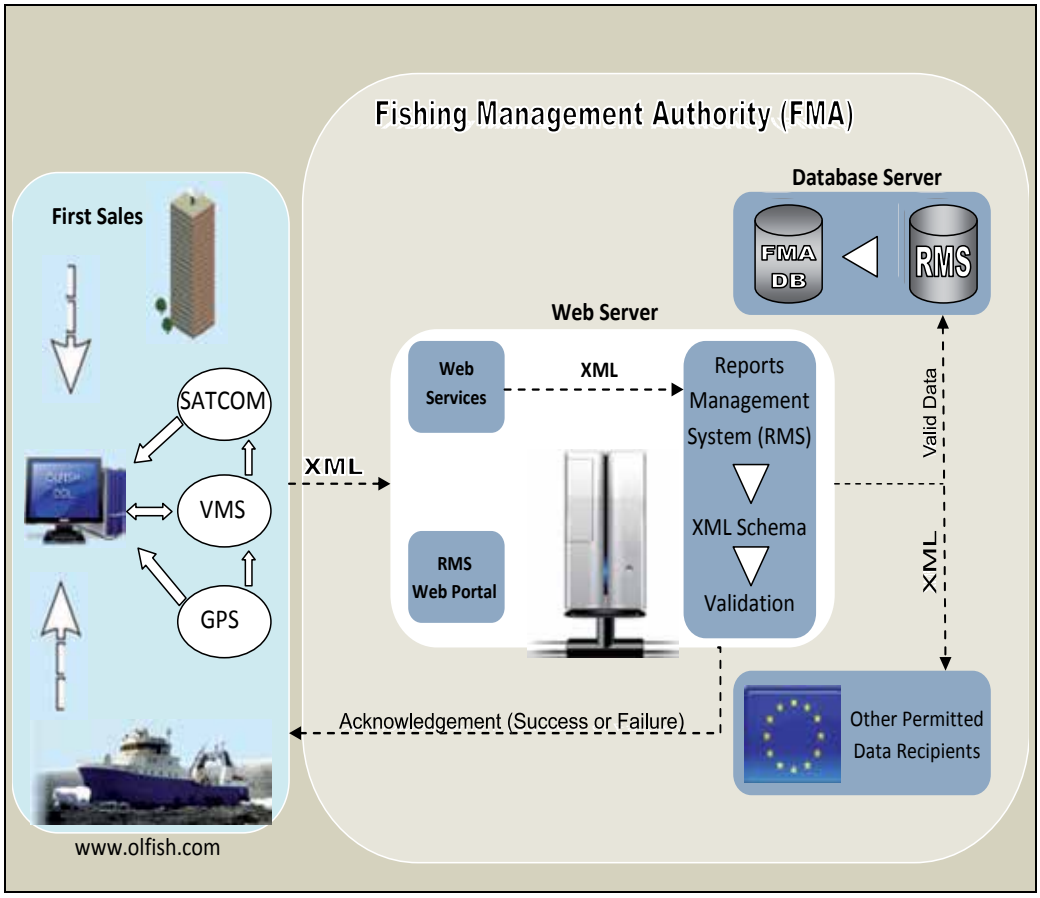


Fig. 1. Overall structure of the Olfish data collection and management system

The third component of Olfish is a web application named Olfish- Report Management System (Olfish-RMS™) and its main function is to receive, store and disseminate reports

coming from Olfish-DDL (or, if necessary, other, third party, data logging systems). Olfish-RMS also allows for the direct entry of data via an internet interface for cases where the use of an onboard data-logger is not practical (cost or unsuitable working environment). With Olfish-RMS the entire fleet of vessels can be managed. It includes a vessel registry, a full quota management system and an elaborate administrative component which allows Olfish-RMS to be customised to satisfy many needs.

2.3 Olfish Dynamic Data Logger

2.3.1 Basic Functionality

Olfish-DDL is a touch-screen-ready utility that captures data in real-time and/or after the fishing activity has taken place. Olfish-DDL can read GPS input via an additional GPS logging utility and it incorporates GIS capabilities for easy viewing of vessel movements and other operational fishing data. With Olfish-DDL, the user can collect any type of data in any form. These include images, video clips, numerical and alphanumeric fields, free text comments, date, time, location, etc. Olfish allows data to be inserted from guiding images ("infograph") to guide it through complex data entry needs. Each mode of data entry has its own unique data entry interface, specifically designed for the type of data recorded. Olfish-DDL is highly customisable and can be easily modified to address vastly different data recording and reporting needs.

Fig. 2. Olfish vessel unit on a tablet PC

2.3.2 Overall Structure

Olfish-DDL consists of the following:

- Configuration files defining levels, fields, parameters
- Database for working data
- Database for archived data
- User interface elements: Data Entry, Data Browser, Mapper, Data Centre, Mini Reporter, Explorer
- Input/output modules for the following types of data:
  a. Reports to specific agencies and third-parties
  b. Import / export of operational data
  c. Backup of the complete system
  d. Error / exception handling reports to Olfish Support

There are two main levels of configuration in Olfish-DDL:

User interface: This is a developer-level configuration which governs the way the command bar menu (Dynamic Commands Bar – DCB) functions, based on client specific needs.

Field and lookup values: As a business model, Olrac ships Olfish-DDL with as many predefined fields as possible. However, within Olfish-DDL, the user can:

- modify field parameters, such as: display names, maximum and minimum values, set mandatory and carry over fields, capture on start/on end, make visible etc. Olfish-DDL ensures that changes which could affect underlying data capture logic are not allowed.
- hide and show lookup table records.
- add, edit and delete lookup table records.
- add fields – these fields have as much functionality and legitimacy as any original predefined fields.

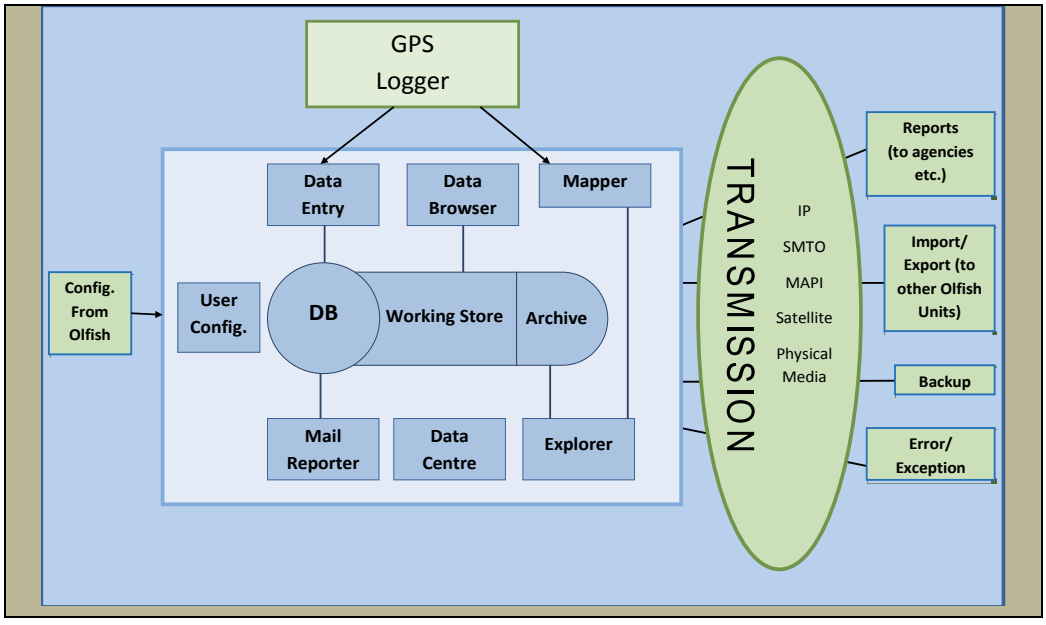


Fig. 3. Olfish-DDL basic structure

Olfish-DDL makes extensive use of dropdown lists whenever possible. The use of dropdown lists to enter data helps to maintain data integrity, thus minimising typos and saving time. However, new fields and values can be added by users if necessary.
While Olfish-DDL is normally shipped with the client's basic user configuration, it can be easily configured to fit the "taste" and needs of different users of the same basic configuration. With Olfish-DDL, the user can decide which fields should be visible, which are compulsory and which are remembered from previous entries. The user can also decide which data fields should be visible in which phase of the vessel operation. Examples are 'trip start', 'trip end' and, within a trip activity, 'start' and 'end'. Olfish-DDL also allows users to set up upper and lower limits for any numerical field in order to reduce the chance of typos.

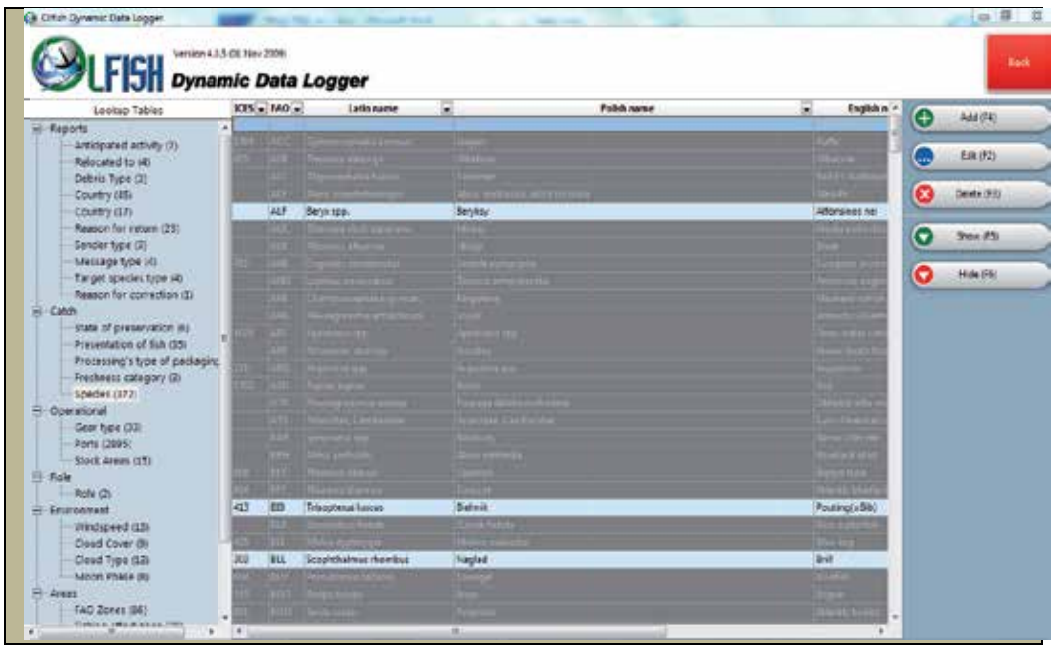


Fig. 4. Olfish-DDL lookup table customisation form

Another feature of Olfish-DDL is a Dynamic Commands Bar (DCB) which can be configured to "intelligently" guide the user during its data logging activities. The DCB can be configured to reflect data logging actions of vastly different activities. For example, the same basic underlying user interface can have different DCBs and can be used to collect data for totally different forms of fishing (trawl, longline, purse seine, traps, etc.) or other vessel activities (sea-farms maintenance, cargo delivery, coastal guard patrols, oceanographic surveys, etc.).

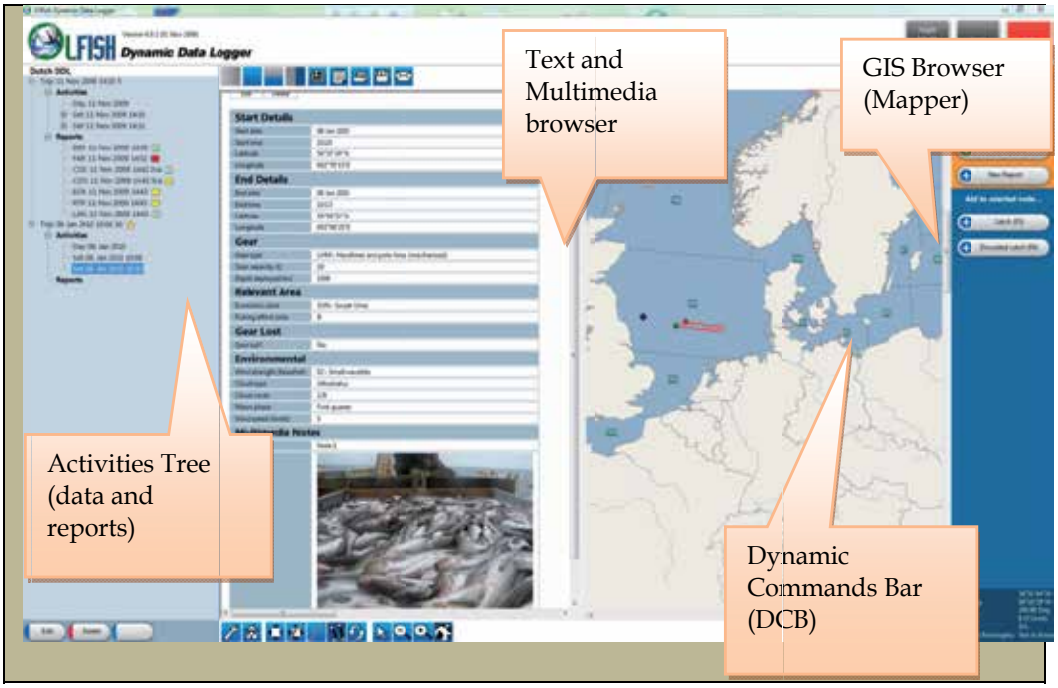


Fig. 5. Olfish-DDL main dashboard screen

Olfish-DDL has been designed to be a highly customisable data logging tool. However, under certain circumstances, it is undesirable to allow the user to change the basic configuration and customisation pre-setup. With Olfish-DDL, it is possible to prevent unwanted configuration changes in cases where data definition is strictly controlled by a higher level management body (examples are: company head offices, management agencies, scientific program managers, etc.). In such cases, it is possible to "hard" configure Olfish-DDL and constrain the user's ability to hide or ignore certain fields. This is mainly done in order to "force" uniformity and full data logging execution when Olfish-DDL is used for regulation-controlled data logging activities.

Data collected by Olfish-DDL can be used to generate any type of report in any format (XML, HTML, CSV, PDF, etc.). These reports can be saved and transferred to other databases (such as Olfish-DDL shore version, Olfish-RMS or other third party databases) either directly, using portable storage devices, or in real-time using the onboard VMS (Vessel Monitoring System) or other onboard satellite communication systems.

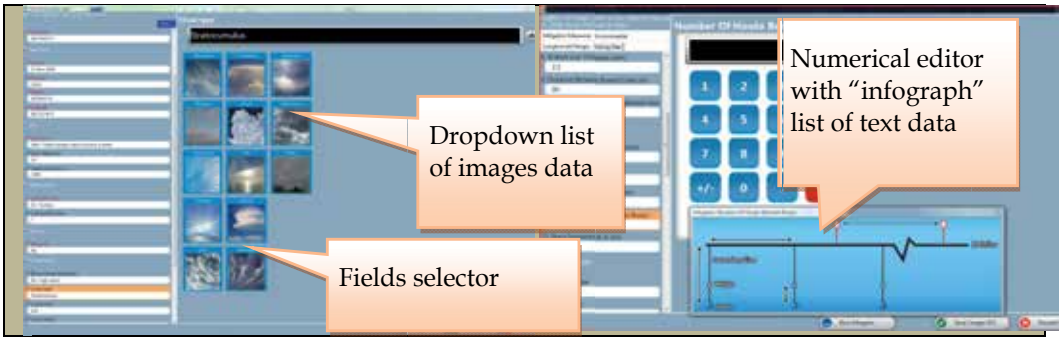


Fig. 6. Examples for Olfish-DDL Data Editor

2.3.3 GPS-Logger

Olfish-DDL can plot vessel movements and trips and set tracks, as well as automatically fill in date, time and location and other GPS related fields, if it has access to a GPS unit. The GPS unit can either be a VMS transponder or a standard GPS outputting NMEA strings on a serial or USB connection.

The Olfish-DDL application does not, in fact, talk directly to the GPS unit. Rather, a small, "light-weight", stand-alone application runs continually on the computer hosting Olfish-DDL. This application is the GPS-Logger, developed by Olrac. Olfish-DDL communicates with this GPS-Logger via a simple application programming interface, allowing all the low-level interfacing with the various GPS units to be handled exclusively by the GPS-Logger. This means that only the GPS-Logger application needs to be updated as Olfish develops support for new GPS or VMS units.

Another advantage to having the GPS-Logger run continuously is that it still logs GPS information even if Olfish-DDL is not running. Olfish-DDL can then use this logged information to provide marker values by means of a small "time machine" utility, when marker data are needed during non-real-time data recording activities. The data stored in the GPS-Logger can also be used to plot vessel tracks even if the user has not actively recorded GPS points.

The GPS-Logger can actually read any serial port information and can be extended to extract data from any set of NMEA 0183 sentences. This allows the GPS-Logger to record not only GPS information, but also any information outputted by devices conforming to the NMEA standard. These devices could include many analogue sensors such as echo sounders and anemometers amongst others.

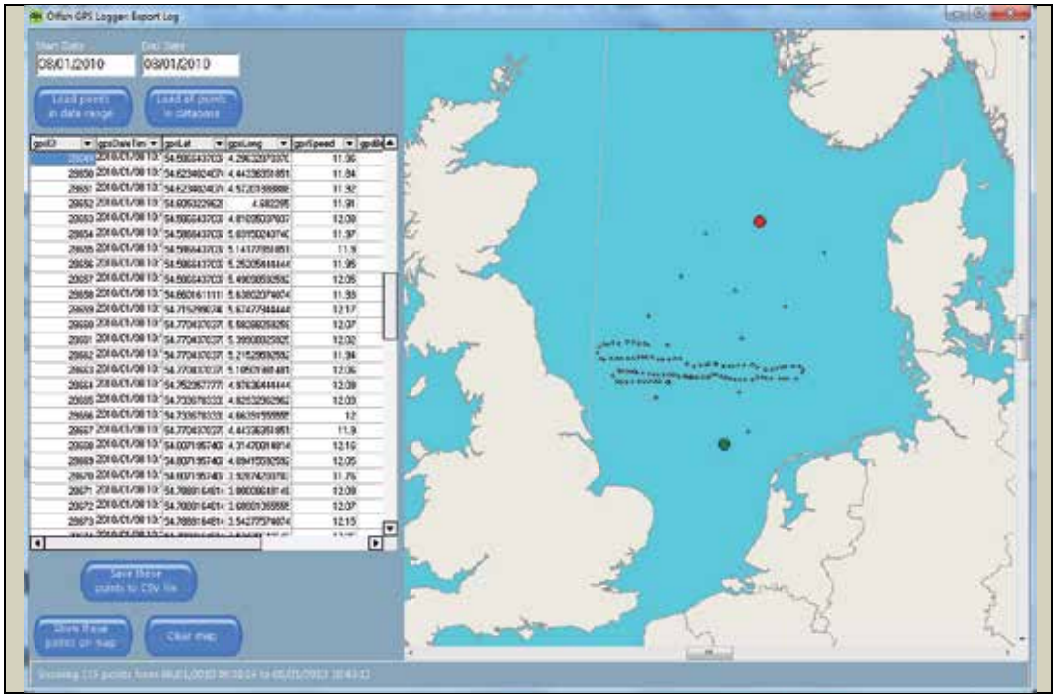


Fig. 7. GPS-Logger

3. Data Management Approach

3.1 Real-Time Activities and Events

Olfish-DDL data classes can be grouped into the following three kinds:

Activities – the data class encapsulates an activity, that is, something that has a start time and location as well as an end time and location. The combination of time and location properties is called a Marker. Examples include Trips, Shots, Sets, Hauls, etc.

Events – the data class encapsulates an event, that is, something that is of short (negligible) duration or has duration of no interest, and recorded as a single temporal point. Examples include placing a trap and retrieving catch from a long-line hook.

Non-spatial, non-temporal data – the data class encapsulates data without time or space markers. Examples include catch records and gear details.

The vessel unit of Olfish- DDL allows for the "real-time" data entry of activities and events, guiding the user in data entry to ensure a linear sequence of events (as they happen). For example, only once a trip has been started can a fishing operation be started, and only once a fishing operation has been started can catch events and catch be recorded. With Olfish-DDL it is possible, however, to switch off the "real time" mode and to enter data at "post event" mode when it might be more convenient for the user to do so. In non-real time mode the entire operation data can be captured on one form in any order desired by the user. In this mode it is assumed that the user will read his/her data from other data logging platform(s) such as a paper logsheet, spreadsheet, or even form memory. However, in order to ease data entry in non-real time mode, the user can access the entire GPS-Logger database and, for example, enter vessel location at any particular time by simply pointing at the

relevant date/time point in the GPS-Logger database. This utility in Olfish-DDL is called the "Time Machine".

3.2 Data Storage

The initial Olfish configuration defines this hierarchical data structure of classes and fields, as well as operating parameters for each level and field. Olfish-DDL uses a hybrid approach to data storage. It uses XML files to store "active" trip data during vessel operations and relational databases to store historic data ("MS Access" on the onboard PC and MySQL on the shore unit PC). Active trip data can be archived into the storage database once the trip has ended and un-archived back, anytime, if any editing is needed (Olfish-DDL does not allow archived data to be edited directly and it is possible to block data from being un-archived if there is such a request). Regardless of which database is used, the database structure is always kept in synchronization with the configuration of the user-specific version of Olfish-DDL. This is necessary, as configuration may change (either by updates from the development team or configuration and customisation changes by the user). On start-up, Olfish-DDL compares the data store structures with the configuration and automatically reconfigures the structure, if necessary and without data loss, to bring it into alignment with the configuration.

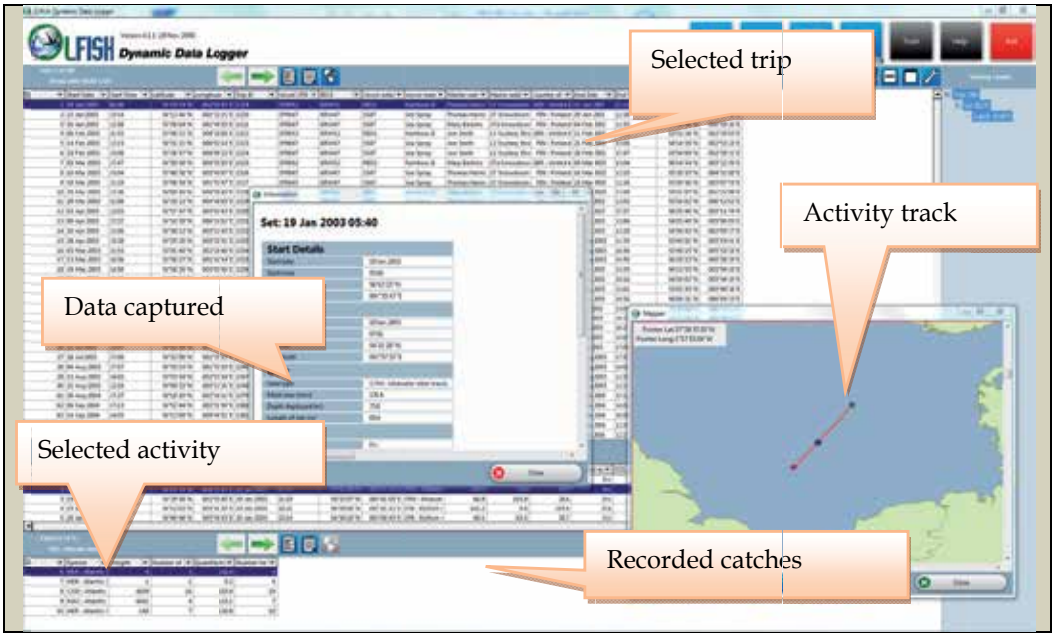


Fig. 8. Olfish-DDL main database interface (Data Centre)

All data captured by Olfish-DDL are organised into classes (or levels or tables). Each class has fields and a set of children classes. For example, one Olfish-DDL might let the user capture information for levels called Trips (a term often used to define the period between vessel departure and return) Sets (a defined activity within a trip), and Catches. A Trip class can have many Sets and a Set can have many Catches. Each class can contain any number of

fields. Typical fields for a Trip could be, for example, Departure Date, Departure Time, Departure Port and Skipper Name. Possible Set fields could include Start Time, Start Latitude, Start Longitude, Gear Used, etc. Possible Catch fields include Species Caught, Weight, Products, etc.

4. System Configuration

4.1 The Need for Flexibility

An electronic logbook is not and cannot be a simple "off-the-shelf" product. Apart from multiple discrepancies within the fishing industry (differing regulations for different countries, different vessel-types, fishing methods, gear, targeted species etc.), the product must also cater for the data logging needs of conservation groups, observers and recreational fishers, to name a few. A rigid software configuration would be unable to adapt itself to the unique data recording requirements of each party. It is therefore necessary that any Elog tool be highly customizable, either at a developmental or user level. In this way, each company or organization will have a logbook configured specifically for its recording needs, with only the fields, input data and report types relevant to its operation appearing on the interface. Such configuration prevents clutter and confusion, and ensures maximum accuracy regarding data collection and transmission.

4.2 Editing Field Properties and Adding New Fields

The user can define new fields either on the vessel or on the shore units. These field definitions are then added to the configuration store, in the same manner as the Olfish-defined fields. These user fields are therefore then indistinguishable from fields defined by Olfish. Olfish-DDL automatically updates the database structure in accordance with these new field definitions.

Fig. 9. Olfish-DDL user configuration form

When data is exported, new field definitions are included in the export file. Then, during import on the target system, Olfish-DDL adds the new fields to the configuration store and updates the database, providing a suitably configured data store for all the data in the imported file.

5. Reporting

Faster transmission of accurate data from vessel to land-based authorities can allow for near–real-time quota management. Current paper-based data-logging systems cause serious delays regarding quota calculations. In order for the quota allocation to ensue, data needs to be submitted, cross-checked and processed, a procedure which, using current data-logging techniques, may take up to a few months. This can substantially influence the last few months of a fishery, where delays in data reception can lead to some quota not being taken or, in the eyes of many, even worse: to be over- taken. Without updated knowledge as to how much quota has been thus far consumed, quota calculations have to be over-conservative and fishers often find themselves ultimately not reaching their entire quota. An electronic dat-logging system, which allows for near real time data entry and submission, and which can communicate via the web with all necessary authorities, reduces transmission time to days (see Web Based reporting, 7 below). Fishers will be able to use up their entire quota confidently, as it will be managed by near-real-time, good quality data. Similarly, management decisions will then also be made in near-real-time, decreasing the risk of over-fishing, as quota deficiency can also be identified and prevented in time.

5.1 An Ecosystem-Based Approach to Fisheries Management

A global move towards a more conservation-directed fisheries management system is underway. In the past, the only preventative measure which vessels were bound to take, regarding ecosystem conservation, involved using only approved gear types. There was very little surveillance and research done on the effect of the non-target-species by-catch on the ecosystem. This is changing. A paper from the Joint Nature Conservation Committee (Pope, J. G. & Symes, D. 2002) of the UK lists a few of the new regulations which are to be implemented in order to achieve this change. Amongst them:

- Scientists are required to provide ecosystem impact assessments alongside the advice on fisheries management options;
- Scientists are required to establish limit reference points for all target and non-target species within the ecosystem and managers should determine the best fishing practice to guarantee minimum levels of disturbance to the ecosystem;
- Fisheries managers are required to take all necessary actions to ensure that damage to the ecosystem (viz. excessive non-target species mortality and degradation of marine habitats) is not caused by intensities or forms of fishing activity beyond those required for rational and responsible exploitation of target species within commercial fisheries. In effect, this will require most - if not all - fishing effort to be reduced to levels commensurate with limit reference points.

Achieving such goals is dependent on the ability to collect data other than retained catches. These should include non-commercial and commercial by-catch, sea-bird and marine mammal interactions, impact on benthic species, detailed information on gear used and

many others. For such a quantity and variety of data, paper logbooks are hopelessly inadequate.

5.2 Real-Time Management

Another advantage of electronic logbook near-real-time reporting is the ability to report trial fishing results (common in mixed species pelagic fisheries) almost immediately. The outcome of such trial fishing can determine if permission to fish is giving or not. Such permission, if not granted immediately (when supported by catch results), can compromise the ability of fishers to fish the discovered shoal of fish since pelagic shoal are often dispersed or move away in a very short time. Presently, it common that the time-lapse between the provision of catch results and the warranty of permission is too long and fish shoals disappear before the vessel is given an opportunity to fish them. Through near-real-time reporting, a sample-catch report can be sent, received, analyzed and responded to in time for the resource to be properly exploited.

5.3 Security

Finally, electronic reporting, able to harness the latest in encryption technology, can heighten report confidentiality and security to the maximum. Near real time reporting means that little time lapses between event and report delivery, thus minimizing any opportunity for tampering with the report by third-party "messengers". Furthermore, not only can reports themselves be encrypted, but access to the reports can be restricted through the use of unique codes and passwords.

5.4 Reporting Through Olfish

5.4.1 Data Transmission Process

Data collected by Olfish-DDL can be transmitted to the shore in two basic forms:

- Raw, original data
- Summarized reports

In principle, Olfish-DDL transmits original data only to its shore mirror units while other data recipients receive data in a summarised form as a report.

With Olfish-DDL onboard, groups of vessels can "export" vessel data to different local companies/agencies (each running an Olfish shore unit) and the local companies can report in turn to a global company/agency running an Olfish meta-shore unit. Operational data can only flow from the vessel units to the shore units. Configuration data (lookup table records, field display names, new field definitions and field properties such as max, min values and mandatory, carryover status) can flow in either direction (from vessel to shore or from shore to vessel). This is illustrated in Figure 10 below.

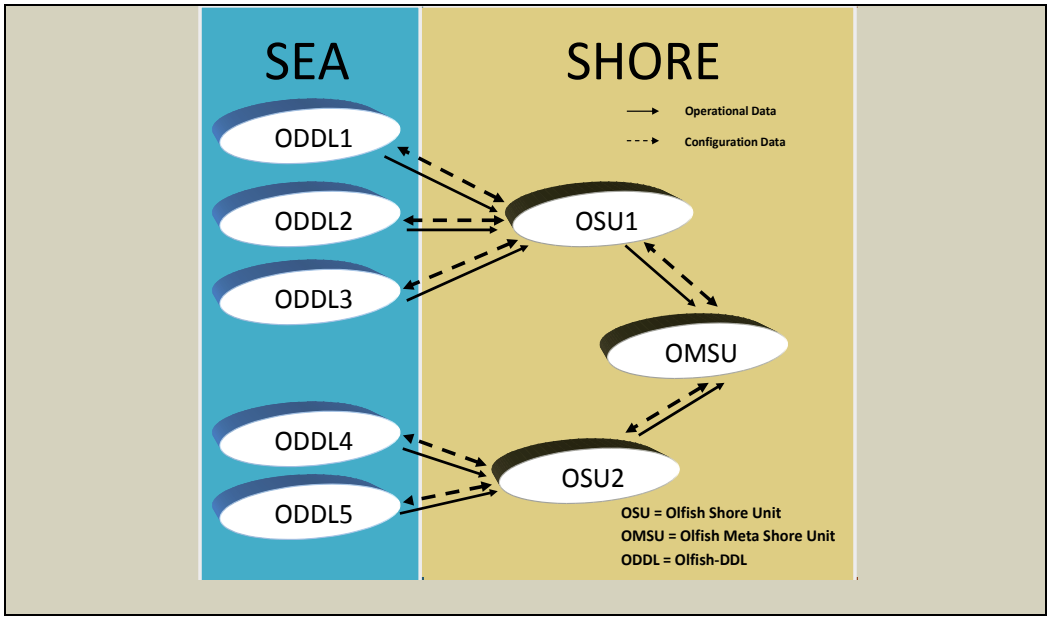


Fig. 10. Data flow between Olfish-DDL vessel units to Olfish-DDL shore units

5.4.2 Reports Transmission Process

Olfish-DDL can also generate specific reports as required by the regulatory bodies and compliance agencies. These specific reports use only a subset of the operational data recorded by the onboard unit. The recipient's of these reports can be either or both Olfish-RMS and other, third party, databases (European ERS for example).

The flow of reports amongst all these entities is configurable and can be customized to match different requirements. Implementations can include one or several of the following:

- Vessel units exporting all operational data to the shore unit. The shore unit then extracts and generates relevant reports for Olfish-RMS or other agencies.
- Vessel unit generates reports and sends them to Olfish-RMS. These reports can be accessed via the internet by permitted users directly or can be passed to other agencies' and/or countries' Reports Management Systems.
- Vessel unit generates reports and sends directly to other agencies' databases.
- Olfish-DDL can compile reports required by commercial clients for their own internal use. However, reports can also be used by any other authorised data recipients such as: compliance agencies, scientific and conservation groups and any other permitted third party commercial or non-commercial data recipients. Eligible third party data recipients can receive data selectively based on their needs and permission level.

Olfish DDL can provide output in a generic (native) XML or CSV format, or can comply with regulations and XSD schemas. Examples of management agencies whose regulations and schemas have been implemented by the Olfish-DDL are: EU regulations (EU 1077/2008), AFMA (Australian Fisheries Management Authorities) regulations etc.

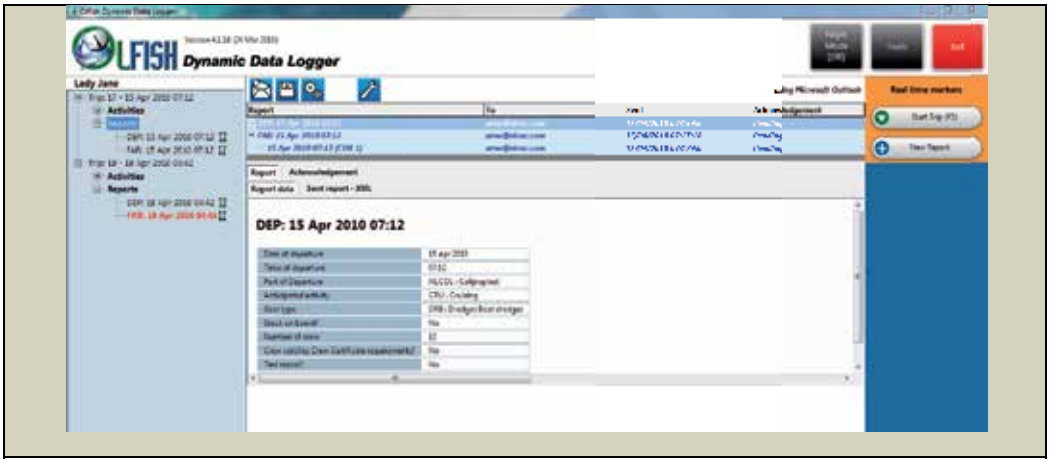


Fig. 11. Report Manager

In addition Olfish- DLL can generate output reports as PDF, HTML, BMP files or even as formal hardcopy logheets as may be required by the compliance authorities.
Reports can be generated automatically by predefined triggered events such as the crossing of certain geographical lines and date/time signals or manually by the user as required. Once generated, reports can be saved to a report queue, viewed, edited, sent and resent. The user can view the reports in legible HTML format or as XML or CSV before sending them out.
Reports can be compressed, authenticated and encrypted to fit any set of transmission requirements. For example, Olfish-DDL can implement secure, end-to-end transmission protocol that allows users to transmit secure XML reports, using X.509 digital certificates according the W3C XML security standards.
In addition to XML or CSV reports, Olfish-DDL can generate reports in the form of SQL scripts (with embedded data) that load the data into the third party database directly, without the need for a front end data loader tool.
Reports can be sent as email attachments, embedded text in the email message or as saved files on any mobile storage container. Olfish-DDL can also send data via a secured, dedicated FTP port if this is the preferred or required transfer mechanism.
Olfish-DDL includes an administrative module for the creation and mailing of predefined reports. These include agency compliance reports as defined by the EU regulations (EU 1077/2008), the Australian Fisheries Management Agency (AFMA) regulations and other management agencies (Poland, Holland, New Zealand, Canada and the USA amongst others).
Sending a report from Olfish-DDL to a third party on the shore can involve some or all of the following steps (Figure 12 below):

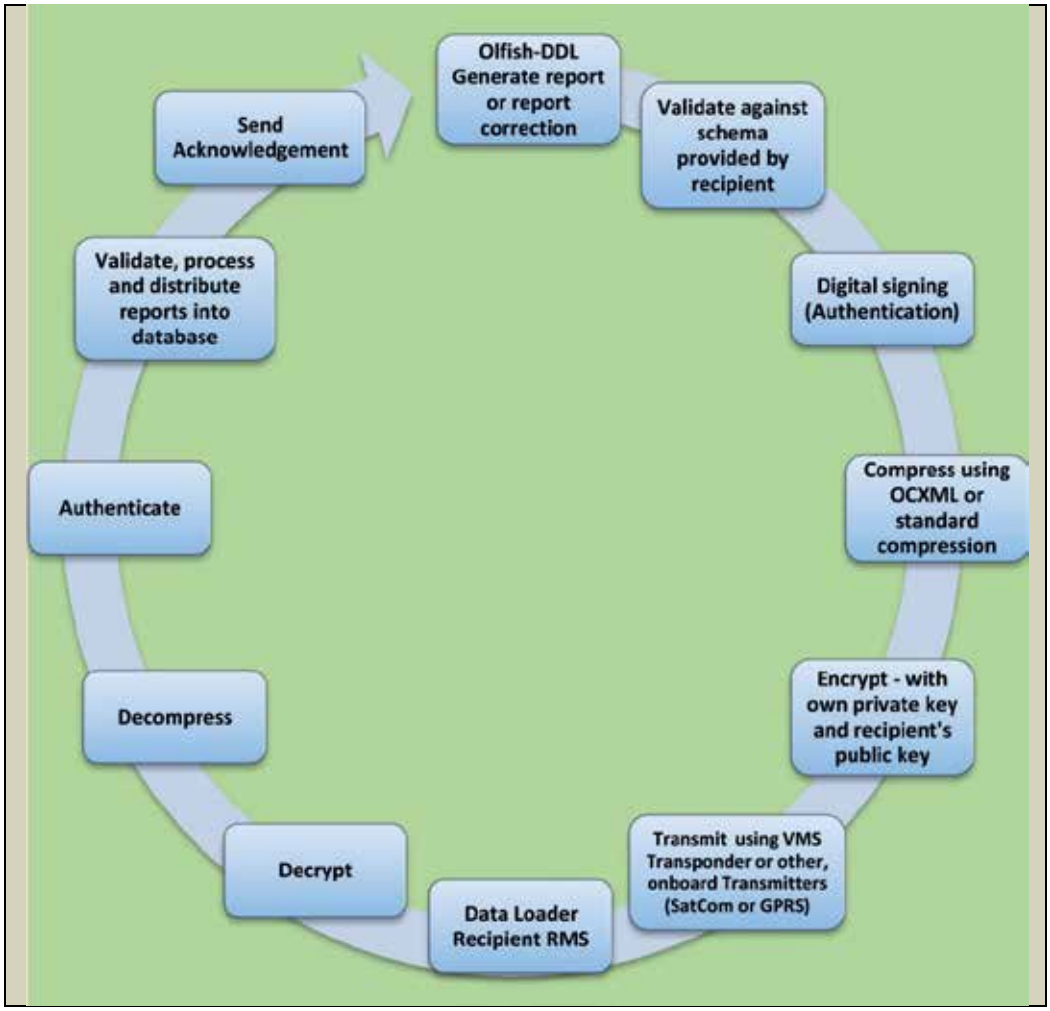


Fig. 12. Transmission process from Olfish-DDL to Shore-RMS

Olfish-DDL can, potentially, generate reports in any format which the client requires. Currently, formats catered for are XML and CSV. Olfish-DDL uses XSLT 2.0 to generate reports in XML format. In Poland, where the shore Reporting Management System was also developed by Olrac Olfish-DDL, EU and Polish specific reports are generated as XML files, which are then compressed as OCXML before being sent to the Polish RMS. On the shore these files are decompressed using ODCXML before validation and data integration take place. Once the reports are validated and processed successfully, a success message is sent back to the sender Olfish-DDL. If the report is not processed successfully, an error message is sent back to the sender, along with a reason for the process failure.

5.4.3 Correction Reports

Olfish-DDL makes a very clear distinction between data and reports. Data is information such as catch, gear used, fishing time, etc. which is captured in a raw (original) form by

Olfish-DDL. Reports are extracts of specific data as required by different schemas. As such, reports are really short documents (XML, CSV, HTML, etc.) compiled from the original data. If data is changed after the report has been sent, it is a requirement in some cases to send a new, corrected report. Any change of data triggers the Olfish-DDL to automatically update reports which make use of this data. Thus, any change of data means a change in the corresponding report.
Olfish-DDL marks sent reports, which have been corrected in this way, visually on the interface. As such, it is very clear to the user that a report has been corrected and must be resent. When the report gets resent, Olfish-DDL adds additional elements to the XML report, indicating that the message is a correction, the reason for the correction, as well as which report is being corrected. Upon receipt of a correction message, the original message is discarded by the recipient authorities, and the new message is processed. Each management authority has a different set of regulations regarding the handling of corrections and updated reports. AFMA, for example, allows only ten corrections per fishing trip or non-fishing period.

5.4.4 XML File Transmission

Many reports sent from Olfish-DDL to Olfish-RMS or other third party RMS, such as enforcement/regulatory agencies, are in XML (Extensible Markup Language) format. XML is not an ideally compact data format, and transmission between vessels and shore using onboard satellite communication systems is often costly. It is therefore useful to compress the XML before sending it. Olrac has developed a method for compressing XML (OCXML) that achieves higher data compression than generally available compression utilities.

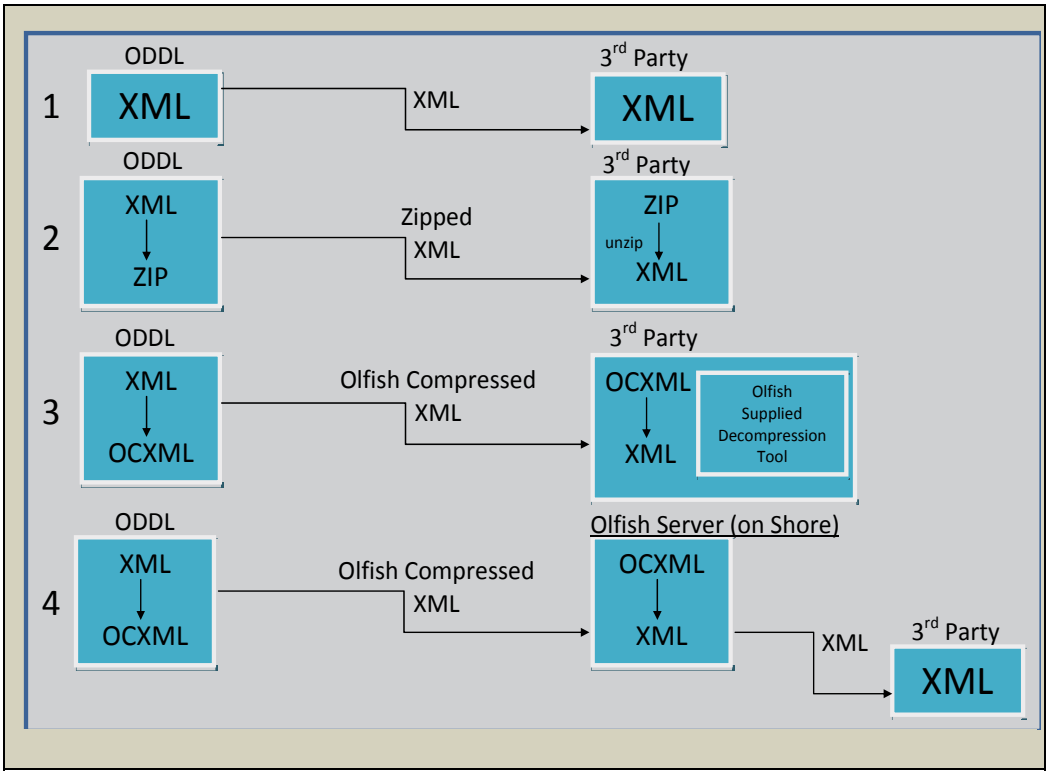


Fig. 13. Transmission of XML files from Olfish- DDL (vessel unit) to third-party (Olfish-DDL shore unit or RMS)

Figure 13 above shows four different ways in which the XML files could be sent:

- uncompressed
- compressed using a mutually agreed-upon compression standard (e.g. ZIP)
- compressed as OCXML, with a paired ODCXML decompression utility on the recipient RMS (either Olfish-RMS or third party RMS).
- compressed as OCXML, sending to a shore-based intermediate Olfish server, which decompresses the OCXML and forwards the standard uncompressed XML file on to the intended third party RMS.

5.4.5 Secure Transmission

A number of transmission methods have been developed by Olrac in order to address different client requirements. In the Netherlands, various reports, as determined by EU regulations 1077/2008 and Dutch regulations, are sent to authorities as XML documents. These XML documents conform to the XSD (XML Schema Definition) offered by Holland's VISHUB agency. The VISHUB agency provides an encryption add-on, by means of a user specific USB key. The type of encryption implemented in Holland is an asymmetric private/public key encryption.

After a report is generated, Olfish-DDL validates the report against the VISHUB's schema. If the report validates successfully, Olfish-DDL invokes the Dutch provided encryption add-

on to compress, encrypt and sign reports generated aboard the vessel. The encryption add-on performs these functions in the following way:

- Compresses the report using g-zip.
- Generates random AES key.
- Encrypts the report with AES key.
- Encrypts AES key with recipient's public key.
- Signs the message with sender's private key – digital signature.
- Concatenates and encodes the contents of the report along with signature and encrypted AES key.

  Upon receipt of the report from the vessel, the VISHUB follows the following process:
- Separates message into report, signature content etc.
- Decrypt encrypted AES key with recipient's private key.
- Decrypt encrypted report with AES key.
- Decompress message.
- Verifies signature using the sender's public key.

The end result of this process is an unencrypted report, exactly as it was sent from the vessel. After the report has been verified, the VISHUB validates the report against its schema. If the report validates successfully, a success message is sent back to the sender. If the report does not validate, an error message is sent back to the sender, along with a reason for the validation failure. If the report validates successfully, the VISHUB forwards a portion of the message to the AID agency.

In Australia, where Olfish-DDL needs to conform with the Australian Fisheries Management Agency (AFMA) regulations, Trip and non-fishing period reports are sent as XML documents that conform to AFMA's published XML Schema. The reports are sent to AFMA's native RMS (eLogs) with a pre-allocated username and password attached to them.

Before a report gets transmitted, Olfish-DDL validates the report against AFMA's XSD. If the report validates successfully, Olfish-DDL encrypts the sender's username and password using AFMA's public key. The encrypted username and password is added to the report. The report is compressed by Olfish-DDL using a standard Zip utility before transmission takes place.

Upon receipt of the report by eLogs, a locally developed data loader utility decrypts the username and password using its private key, thereby verifying that the report comes from a valid sender. Once authenticated, the report is validated by AFMA against its own XSD. If the data validates, AFMA imports the report data into its database.

5.4.6 Communication Techniques

Olfish-DDL allows data and reports to be sent out using a variety of transmission methods, depending on their availability on the vessel.

MAPI: Olfish-DDL can send email using any standard or maritime email software that supports the MAPI standard (the "simple" MAPI subset is sufficient). MAPI is the Messaging Application Programming Interface, a messaging architecture for Microsoft Windows. The email software must be:

installed on the same computer as Olfish-DDL

correctly configured to use some transmission mechanism (such as TCP/IP or Inmarsat-C) and

set up to be the default email client.
Examples of e-mail clients that can be used to send e-mail using MAPI:
Stratos AmosConnect
Vizada SkyFile
Outlook
Any other e-mail client that offers simple MAPI support
Direct interface to satellite transceiver on vessel: Olfish-DDL can send email directly using several different types of satellite transceivers. Olfish-DDL achieves this by using the Olfish GPS-Logger utility, which communicates directly via a serial port connection to the transceiver and lets the transceiver handle the transmission using its own proprietary technology to send emails and attached reports. Examples are:
Thrane & Thrane Inmarsat-C transceivers (Inmarsat)
BlueTraker (GPRS/Iridium hybrid)
Integrating with maritime e-mail solutions: Olfish-DDL can send mail using these programs in the following way: The report is generated by Olfish-DDL and then copied automatically to a special "outbox" folder. In some cases, the user then browses to this folder from within one of these third-party email applications, finds the report that has just been generated, and sends it. In other cases, the third-party software automatically "polls" the outbox folder, and sends the reports at set intervals. Sometimes reports cannot be sent as binary attachments by the third party software. In these cases, Olfish-DDL uuencodes the binary reports into text before they are placed in the outbox folder. The recipient e-mail client then uudecodes the text back into its binary report form.
Examples of third-party software with which Olfish-DDL interacts are:
Boatracs
SkyMate
Faria WatchDog
METOCEAN MetMailer
Thrane & Thrane EasyMail
Direct SMTP and POP: Olfish-DDL can send and receive email via its own internal emailing utility (Olfish-Mailer), using internet SMTP (Simple Mail Transfer Protocol) and POP (Post Office Protocol), if the vessel has access to the Internet (TCP/IP), either broadband or dialup. Olfish-DDL then sends email without any need for third party email software.
Other direct internet protocols: If the vessel has access to the internet (TCP/IP), then Olfish-DDL can be configured to make use of several different internet transfer protocols. These include SMTP and POP as described above (specifically for email) but also FTP (File Transfer Protocol) and HTTP/HTTPS (Hypertext Transfer Protocol/Secure Sockets layer), for text and binary files.

5.4.7 Security and Data Integrity

Olfish-DDL includes numerous security and validation features to ensure data security and integrity. Examples are:
User logins – the software requires a unique user ID and password to access each software installation. Each software installation can have a number of authorized users, for example an administrator, owner/operator and mate.

Edit/Delete control - although with the standard Olfish-DDL the user has the ability to delete and/or edit any recorded data, these features can be shut off or amended to meet the data standards of the customer.
Criteria monitoring – many fields in Olfish are valid or permitted only when certain criteria have occurred. With Olfish-DDL, only valid fields and values are visible depending on predefined combinations of criteria.
Data validation: Olfish-DDL allows the user to set up a minimum and maximum for any numerical data to prevent the entry of out-of-range typos.
GPS Puller – a feature located on a remote computer, which allows authorities to call for and receive location data directly from the Olfish GPS logger at any point during a vessel's trip.

6. Add-On Modules and Other Utilities

6.1 Change Control

In addition Olfish-DDL allows permitted users to track back any change that was made to data previously saved. This utility is password protected and can be activated only as part of an overall agreement between management authorities and the user. When in auditing mode, the user can see any sequences of changes made to previously saved data. These include the date and time of a change which was made, the username and the original value/s.

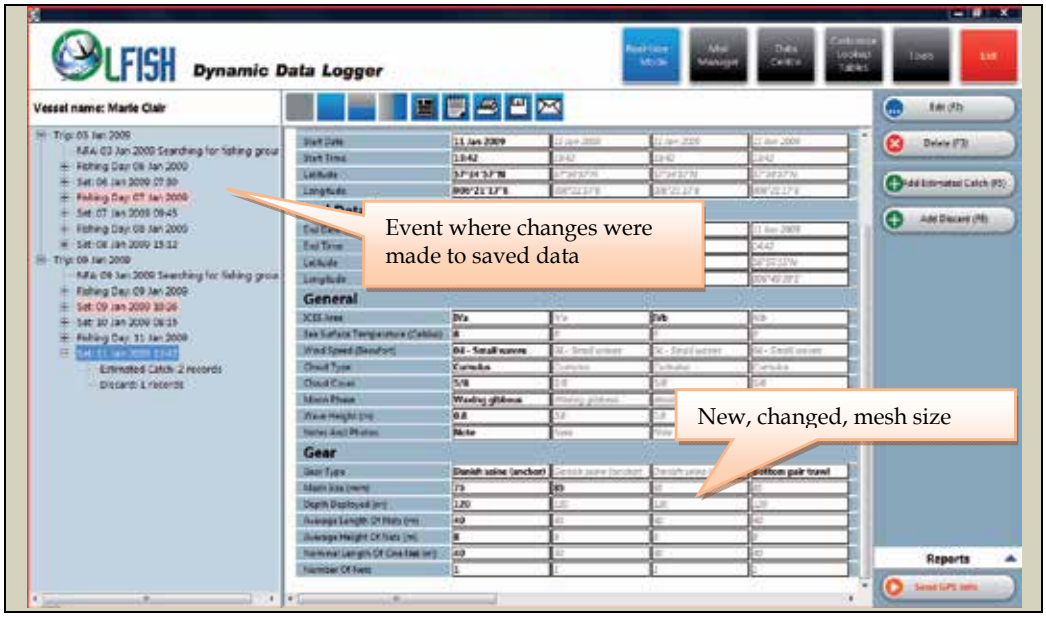


Fig. 14. Change-Control utility: pink markers on the activity tree show places where changes took place, bold figures in 2nd, 3rd and 4th columns were changed.

6.2 Multinote Taker and Notebook

Olfish DDL includes a utility named Multinote taker which allows users to enter multimedia data in the form of free text, pictures and video clips. These notes are part of the

Olfish-DDL database and the user can browse them anytime by pointing at the activity where the Multinote was taken or by using a specially built Notebook browser which allows users to browse through all their notes, pictures and videos in one location while keeping the note traced to its original location on the DLL activity tree.

Fig. 15. Multinote Taker

Fig. 16. Multimedia Notebook browser

6.3 Mini Reporter

The mini-report query utility (Mini-Reporter) is a relatively simple query tool which allows users to extract basic statistics of recorded data during an active trip before they are

archived in simple tabulated form. The mini-reporter allows users to group data by different grouping variables.

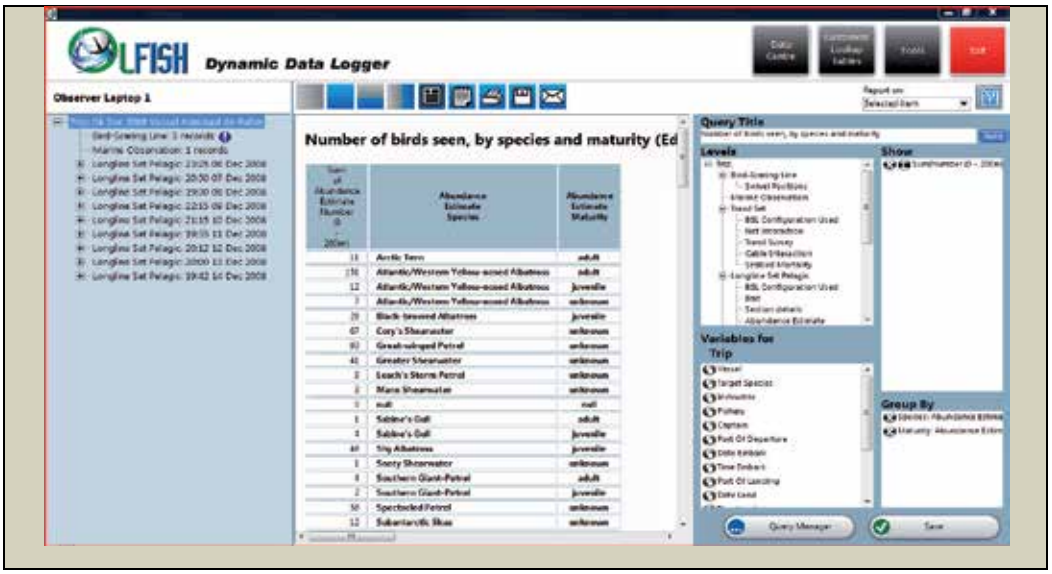


Fig. 17. Mini-Reporter: The HTML browser in the centre is changed to reflect the query output as was set up on the right.

6.4 Olfish Explorer

Olfish-Explorer is an add-on, optional, visual data analysis module specifically designed to work with the Olfish-DDL main database. It allows the user to analyse subsets of data captured by Olfish-DDL. The DDL lets the user capture data grouped into classes. Each class has fields. For example, one DDL might let the user capture information on Trips, Sets, and Catches. A trip has, for example, Departure Date, Departure Port and Skipper Name fields. Possible Set fields include Start Time, Start Latitude, Start Longitude, Gear Used, etc. Possible Catch record fields include Species, Weight, etc. From a selected set the user can create Subsets of data for a particular analysis or presentation. For example, graphs can be drawn showing CPUE (Catch per Unit Effort) as a function of time, moon phase, current strength etc. Also, spatial CPUE density distributions can be plotted on a map, which can then be filtered for different target species or to reflect differing environmental conditions. Subsets can be swapped in order to explore different scenarios and data relationships. For example, CPUE at new moon time compared to CPUE at full moon. The subset definitions, i.e. the list of classes and fields selected for the subset can be saved for re-use.

With Olfish-Explorer it is possible to add calculated fields and to use them just as pre-entered data. For example, the calculations of CPUE by dividing catch weight by fishing duration. The Explorer allows the user to create and set up any number of calculated fields and to edit and delete them.

The table created by the Olfish-Explorer query engine can be shown on a graph (such as a bar graph or a pie chart) or shown in a density distribution map (if the data contains longitude/latitude information for each row in the table).

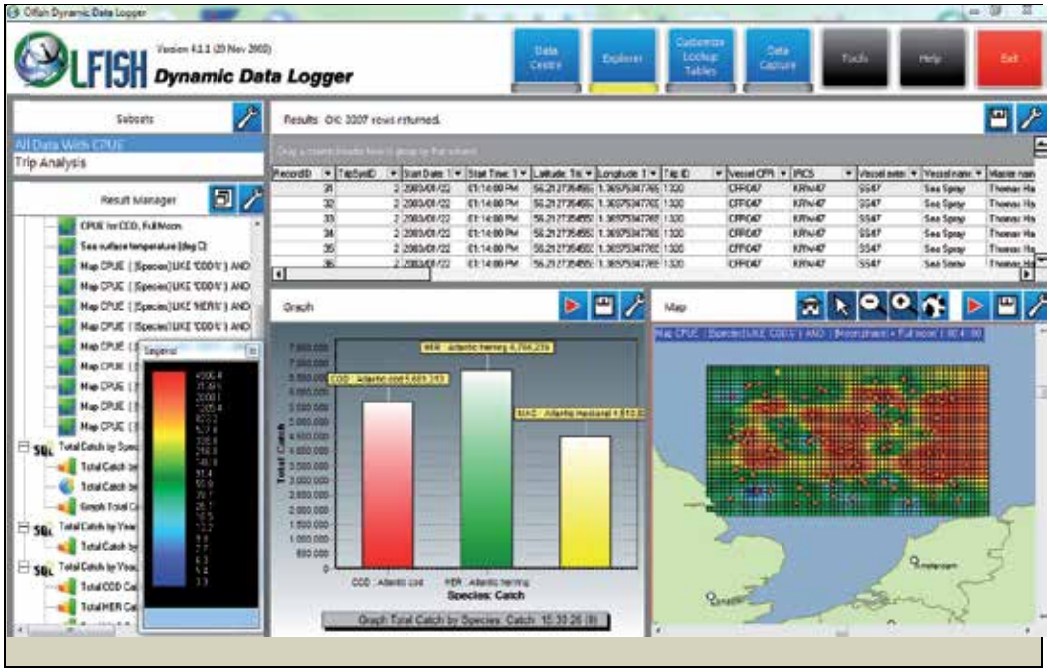


Fig. 18. Olfish-Explorer view mode

6.5 Fishing Consultant

The Fishing Consultant can predict the future catch rates of a particular species for specified fishing areas and environmental conditions. The consultant uses catch rate and any other information that is available about the conditions in which each recorded catch occurred to produce a density distribution, indicating which areas are most likely to have the densest population of fish.

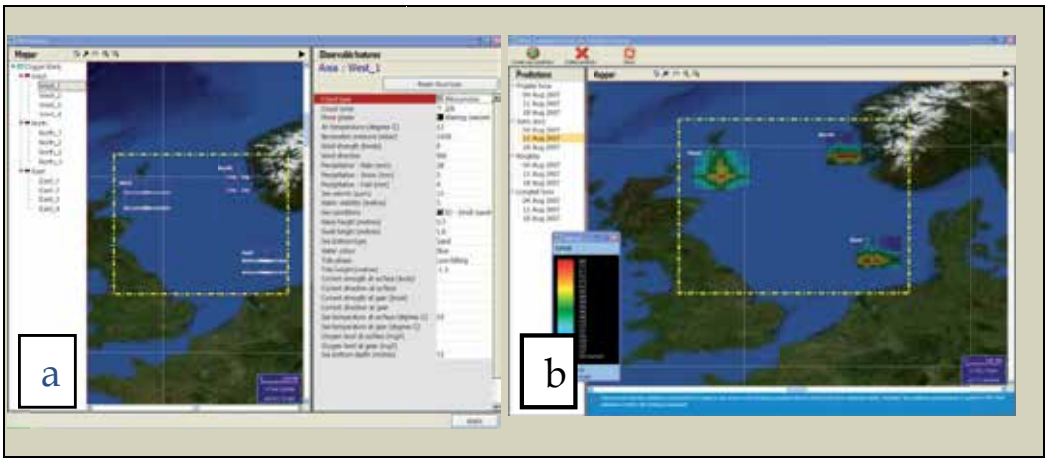


Fig. 19. Fishing Consultant. Setting up prediction parameters (a). Projected good fishing sites (b)

A mathematical prediction model is trained on the historical data of catch and all recorded environmental factors. The trained model can then predict future catch rates under conditions specified by the user. The predictive accuracy of the model depends on the quality and quantity of data available in the database. The predicted catch rates can be visualised on a GIS component that provides a colour-coded predicted catch rate grid of each fishing area specified by the user.

6.6 Fleet Activities Optimizer

OLFISH Fleet Activities Optimiser is a utility which has been designed to aid fishing managers in scheduling ship time optimally in order to meet certain objectives. The program can manage a fleet of numerous vessels, fishing in multiple areas, for several species of fish, one of which will usually be identified as the primary target species. The scheduling mechanism can be set to achieve any of the following tasks:

- Land a specified tonnage of the primary species at the minimum possible cost.
- Land a specified tonnage of the primary species in the shortest possible time
- Land the maximum possible catch of the primary species in a specified time period.
- Maximise the value of the catch (all species) in a specified time period.
- Land the primary species at minimum cost, in order to maintain stock for processing at a specified daily rate without exceeding specified stock limits.

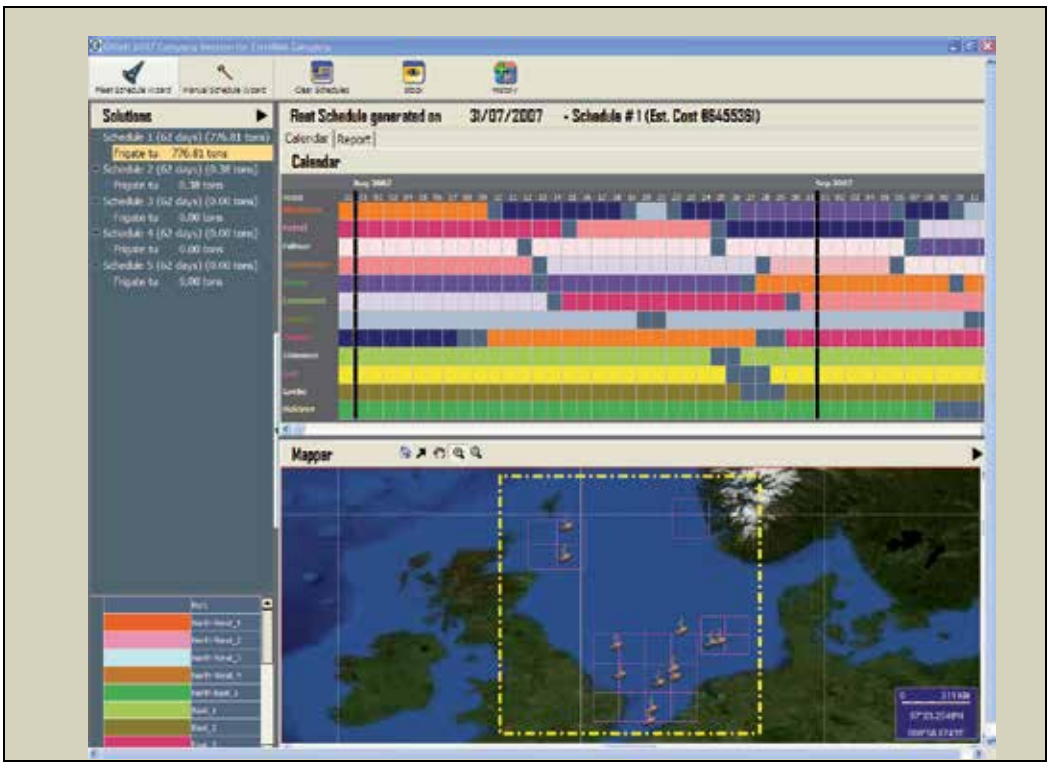


Fig. 20. Fishing fleet activity optimiser – vessel locations and proposed schedule

In determining a solution, the program uses input information with regard to vessel specifications, the distance from port to each area, the expected catch rate in each area for each species (which may be generated using the "OLFISH Fishing Consultant") and the specified constraints. The output of the program is a visual representation of a number of possible fleet schedules, each of which is close to the optimal solution, but which may differ with respect to bycatch landings etc., allowing the user to choose the most appropriate solution for his/her needs.

6.7 Form Maker

It is also envisaged that there will be a transition period, where a vessel may be using Olfish-DDL to log vessel activities and catches, but those vessel activities and catches will still need to be reported to agencies and government authorities using paper-based industry standard log sheets. In order to address such needs, Olfish-DDL allows any paper-based log-sheet to be scanned, incorporated into, and completed within Olfish-DDL, and then to be printed out and delivered to the relevant authority. The paper logsheet can be populated automatically by the form maker utility if data have been entered using Olfish-DDL.
The form maker utility is a developer utility and cannot be used by the user. It does, however, make the creation of hard-copy forms within Olfish-DDL an easy and fast process. It is our hope, nevertheless, that paper forms will soon be a thing of the past since the secure and efficient way of recording and transferring data offered by Olfish-DDL eliminates the need for paper forms of any kind for this process.

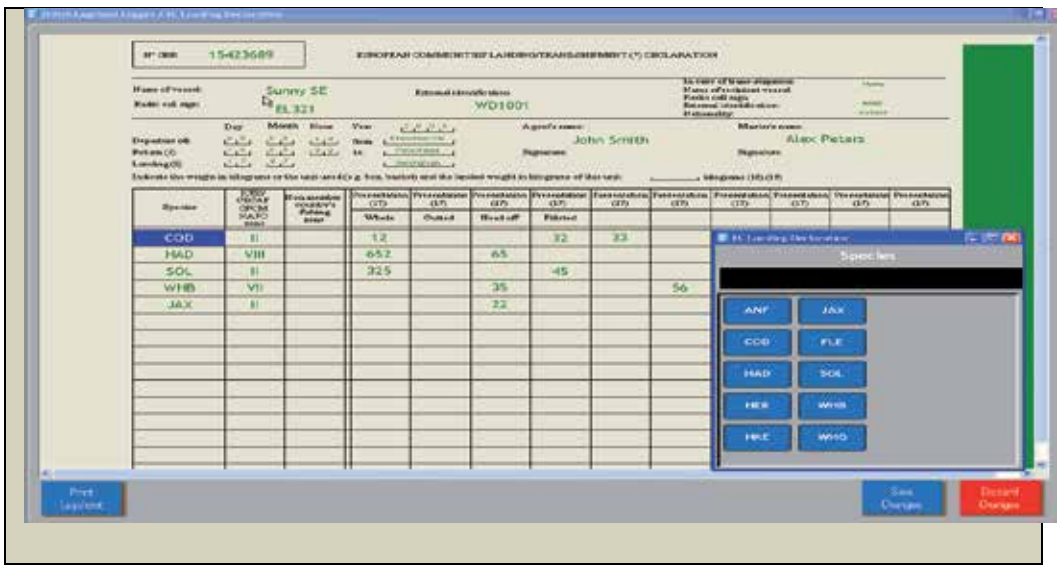


Fig. 21. Form Maker

6.8 Scales Data Logging Utility

The Scales data logging utility facilitates the incorporation of data obtained from mechanical scales (regarding the species, weight, grade and the processing state of catch). The data received from the scales is sent to a computer by either a TCP/IP network connection or a

standard RS232 communications port. Olfish-DDL currently caters for four different industry standard scales. These scale types are NESCO, Pols, Scanvaegt and Marel scales. A scales code component allows a user to link data such as species, weight, grade and the processing state used in the software to the scales codes that are used by the scales. An individual scales code might indicate a particular species of a particular grade and processing state that is being weighed. The scale sends only the scales code and the weight of the fish currently being weighed to the software. The Scales Interface component uses the scales code to retrieve the species, grade and processing state and then ties these values to the received weight.

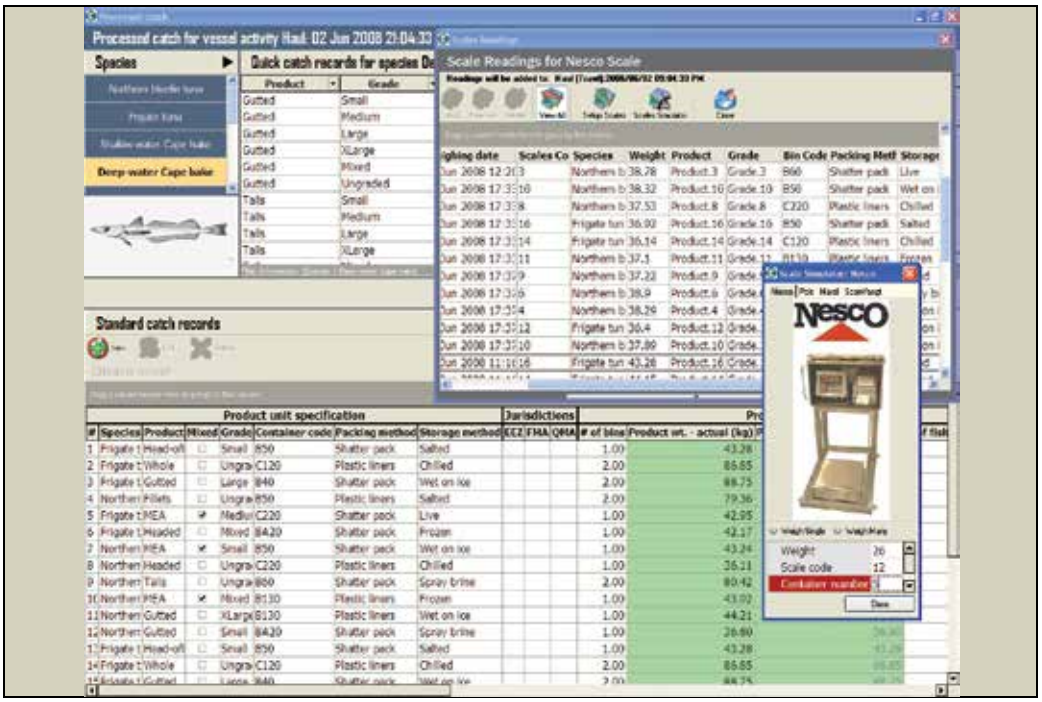


Fig. 22. Scale interface

6.9 Fishing Inspector

The Olfish-DDL inspector utility is a small program which can reside on the compliance inspector's USB key. This utility allows permitted inspectors to extract certain (predefined) information from the user's Olfish-DDL without the need to give them access to the full software.

Fig. 23. Olfish-DDL Inspector utility

7. Olfish-RMS: A Web-based Reports Management System

7.1 Functional Overview

In those installations for which a client does not have or does not wish to use a pre-existing database, Olfish has developed a shore management reporting system it has named Olfish-RMS. This system is an in-house implementation of the Electronic Reporting System (ERS) as requested by the European Union. Olfish manages reports generated by vessels and allows users to access their sent reports via the internet. The system allows the administrator to see, in real time, reports coming from different vessels into the RMS. Once reports are saved and processed, they can be viewed and data can be summarized (such as total catch by vessel and by species). The RMS also incorporates Google Maps to graphically visualize vessels and catch locations.

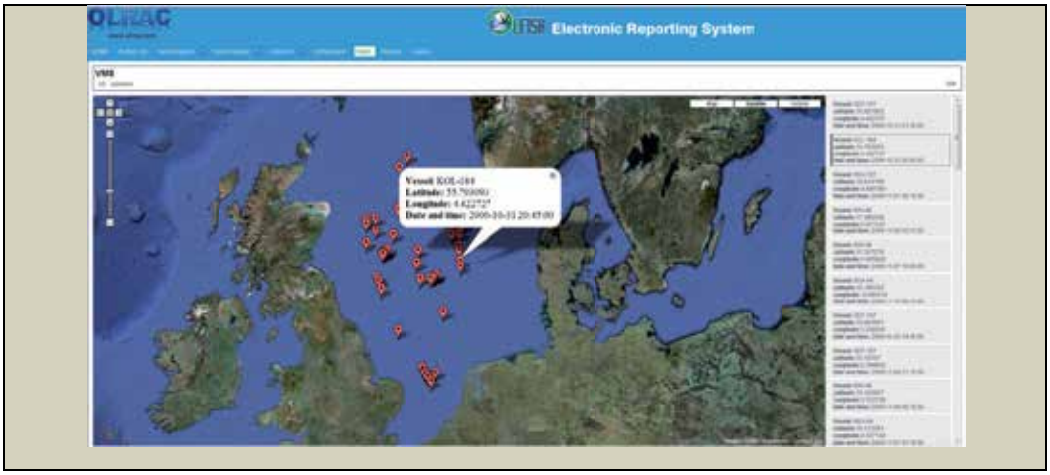


Fig. 24. Olfish-RMS WebViewer: Vessels location using Google Maps

Olfish-RMS is a complete system for the management and dissemination of reports (catch, landing, vessel movement, sales and transhipment) received from fishing vessels or entered directly when relevant. The Olfish-RMS server was constructed according to the EU regulation (EC) 1077/2008 as was adopted by the Commission in 2008. However, the Olfish-RMS server has been designed to allow its scope to broaden to include other national data and reporting needs. Olfish-RMS is deployed on an MS Windows Apache server running PHP version 5 and MySQL Version 5.

The Olfish-RMS consists of 2 components: A web-based interface (WebViewer) and a communication module (MailPuller).

The web-based interface is controlled by user access control. The levels of access are:

- Administrator – Has complete control to add/edit/delete data on the system as well as viewing of aggregate data.
- Inspector – Has access to add/edit/delete data related to EU reports and the viewing of lookup fields as well as the viewing of aggregate data.
- Vessel user – Allows the vessel owner/master to view all reports submitted by the vessel.
- Member state – Allows other member states to view landings when this occurs in their waters, as well as prior notification reports to their ports.

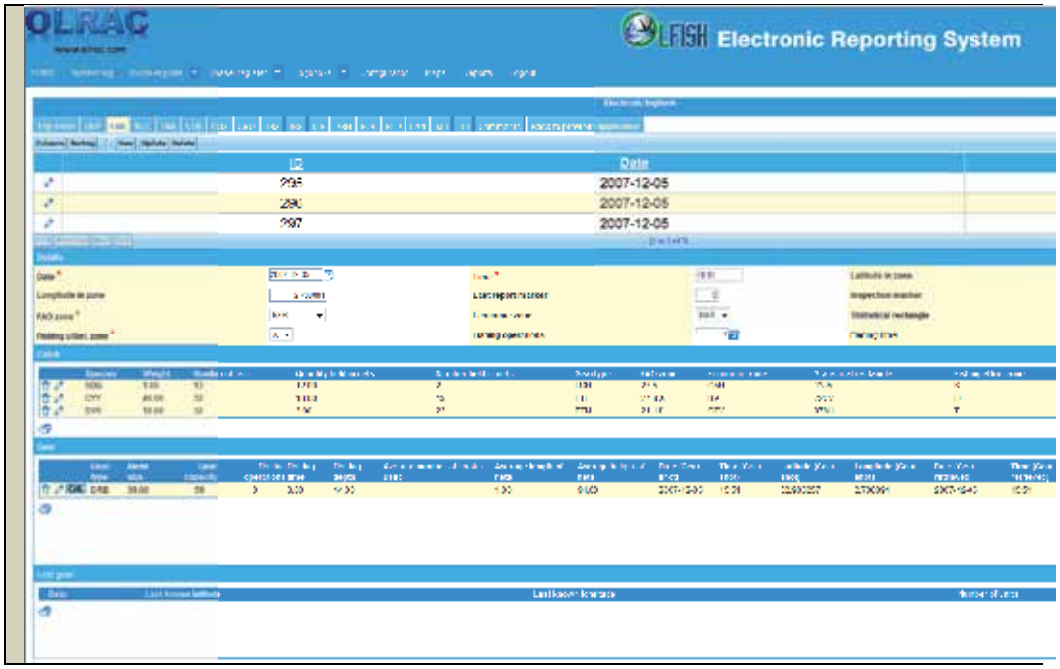


Fig. 25. Olfish-RMS WebViewer. Example listing - Vessel Registers

The communication module (MailPuller) is configurable to connect to any SMTP server account dedicated to the receipt of electronic reports from Vessels and First Sale units as well as other member states. The MailPuller also scans all incoming messages, and where relevant (according to regulation EC 1077/2008) forwards landing and prior notification reports to other member states. The configuration of which member states to send to is done from within the Olfish-RMS web interface. Once incoming reports have been validated, they are imported into the Olfish-RMS MySQL database. If validation fails, a failure report is sent back to the sender. If the report is imported, a success acknowledgement is sent back to the sender.

If required, inspectors can manually add report data. This may be necessary due to the failure of onboard systems or for vessels that have not yet adopted an onboard reporting system. Inspectors are also able to enter comments on any submitted report and reports are marked as pending until they are accepted by an inspector.

Olfish-RMS allows the user to view reports on an individual level (in their original format) as well as aggregated reports based on imported data. The system also allows for the import of corrected data as per EU regulation (EC) 1077/2008. Vessel location and summary data can also be viewed spatially, using Google Maps. Olfish-RMS facilitates data cross-checking, system status, data validation and the conditional dissemination of data and reports to other member states as dictated by EU regulations. Olfish-RMS is also preconfigured with all lookup values as specified by (EC) 1077/2008.

Olfish-RMS allows the user to swap between languages. The interface, as well as all lookup values, change based on the language selection. Olfish-RMS is able to export all lookup data and reports to Excel and Word formats.

Fig. 26. Olfish-RMS WebViewer. Example of data entry screen for manual entry of report data

7.2 Technical Overview

7.2.1 Definitions

Olfish-RMS – The web-based Olfish Report Management System for the processing of electronic reports
Mail Processor – Windows-based application that connects to a mail account and downloads, processes and acknowledges electronic reports
MS – member states / other countries
Flag state – vessel country of registration
Windows service – always on Windows application running in the background
Web service – A standard method of integrating web-based applications over the Internet. All communication is done using XML
SOAP – a protocol for exchanging structured data in the implementation of web-services. It essentially provides the envelope for sending web-service messages.
Apache – a deployment platform on the server on which Olfish-RMS is installed which returns a web page to a web browser when the web browser requests it.
PHP - a widely-used general-purpose scripting language that is especially well-suited for Web development and can be embedded into HTML
MySQL – open source relational database
Ajax - a group of interrelated web development techniques used on the client-side to create interactive web applications
Delphi – the programming language used to write the data logging, data pulling and data pushing applications

7.2.2 The Basic Functional Infrastructure

Olfish-RMS is developed to run on a Microsoft Windows platform with an Apache distribution containing MySQL and PHP. The exact version requirements are:

Apache 2.2
MySQL 5.1
PHP 5.3.1

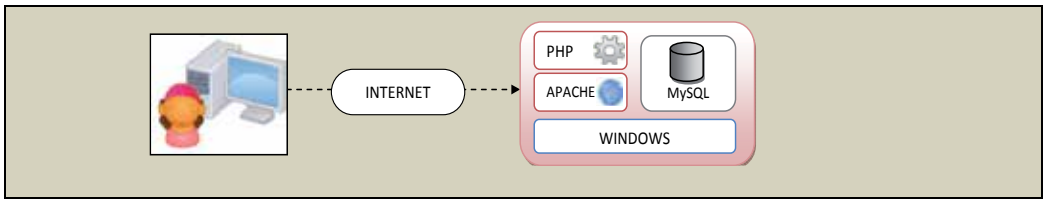


Fig. 27. Olfish-RMS infrastructure

7.3 System Components

Olfish-RMS consists of four components:

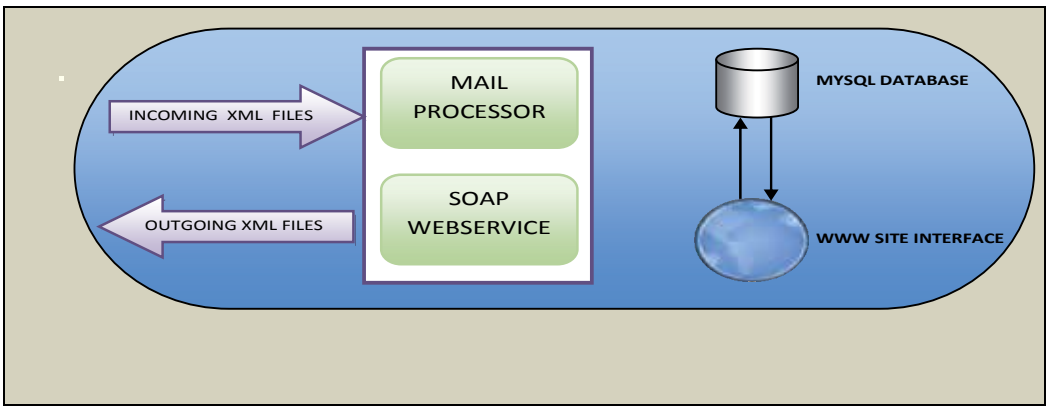


Fig. 28. Olfish-RMS components

7.3.1 MySQL Database

Olfish-RMS uses a well-designed relational database that makes extensive use of indexes and stored procedures for the fast execution of reporting requirements on growing databases. The database implements the following modules:
Security – controls access to the interface
System logs – logs all automatic and manual processes on the database. This includes the receipt and sending of electronic reports
Electronic reports – stores received electronic reports in a relational way with links between report data. E.g. a departure report and a return to port report are linked to the same trip
"Paper logbook" reports – stores manually inputted data from the paper logsheets until these are to be phased out completely.
Monthly reports – stores manually inputted data from the monthly paper logsheets in use by smaller vessels
Quota management – stores yearly TAC allocations as well as the distribution of these allocations to vessels and vessel groups

Vessel registry – stores entire fleet details, movement into and out of the fleet as well as all changes to the vessels (physical or management related)
Lookups – stores all lookups in use throughout the system. All regulated lookups come pre-populated and are not modifiable

7.3.2 Web Site/ Interface

Olfish-RMS has a PHP written front-end utilising AJAX technology for ease of use and speed. All forms and grids are implemented uniformly for quick understanding of the user interface. All grids throughout the system have a header section with one or more of the following functionalities:
Search – the user can specify filtering for each field in the grid.
Columns – the user can select which columns to display in the grid
Sorting – the user can set the column sorting (ascending or descending)
XLS – export entire grid contents to Excel format
Print – prepare the grid for printing
Likewise, a footer section with one or more of the following functionalities:
First, Previous, Next, Last – Move page by page through the dataset in the grid
Go to – jump to a particular page
View – change the number of records displayed in a grid on each page.
The website provides an interface for the eight database modules listed above as well as five additional interfaces:
Reporting – Interrogates tables for user defined reports in Excel and third party data formats
Cross-checks – compares reported catch weights across different reports
Maps – visually displays vessel locations across different reports using Google Maps.
Tables – Tabulated presentation of data such as catch distribution, quota caught, etc.
Graphs - graphic presentation of similar data (above).

7.3.3 Mail Processor

Olfish-RMS utilizes a Delphi written application that registers three Windows services performing the following:
Monitors a POP3 account for incoming electronic messages from vessels. In order to configure the Mail Processor POP account, the user needs three pieces of information:
The name of the user ISP's mail server that holds the user email. Typically it's something like "mail.example.com".
The name of the account the user was assigned by the user ISP. This may or may not be the user email name, or something like it, or something completely unrelated.
The password to the user account.
Processes electronic messages in the following steps:
Decodes if required
Decompresses
Validates against schema
Saves to relational database
Prepares acknowledgements
Connects to SMTP server for the sending of electronic messages (acknowledgements) to vessels.

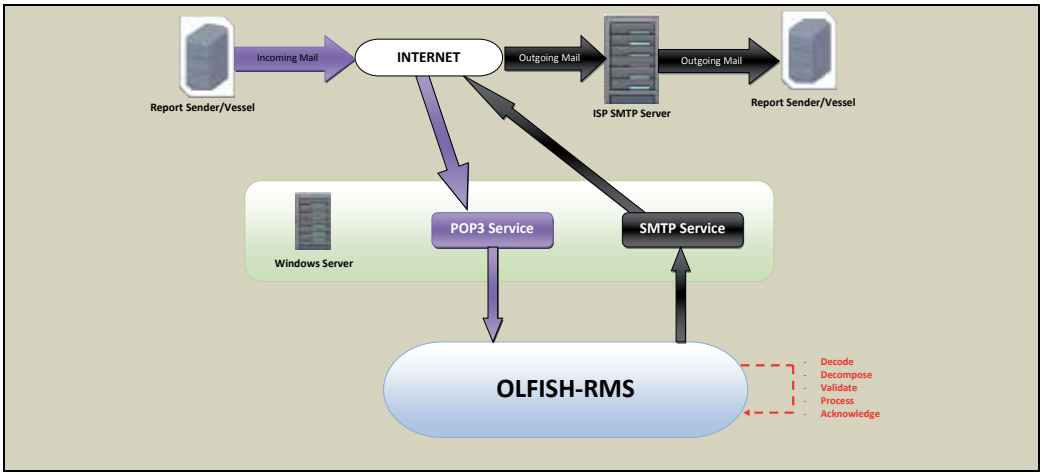


Fig. 29. Processor processes

7.3.4 SOAP Web-Service

Olfish-RMS registers a SOAP web service for interaction with other member states. This allows for the sending and request of reports between MS. The unique procedure of the webservice is called setERS - this is the only entry point of the webservice. MS only need to communicate the IP address and port of the server the webservice is running on. The procedure is called using the SOAP protocol and the parameter of the setERS procedure is an XML message containing an OPS element.

OPS elements have the following structure:

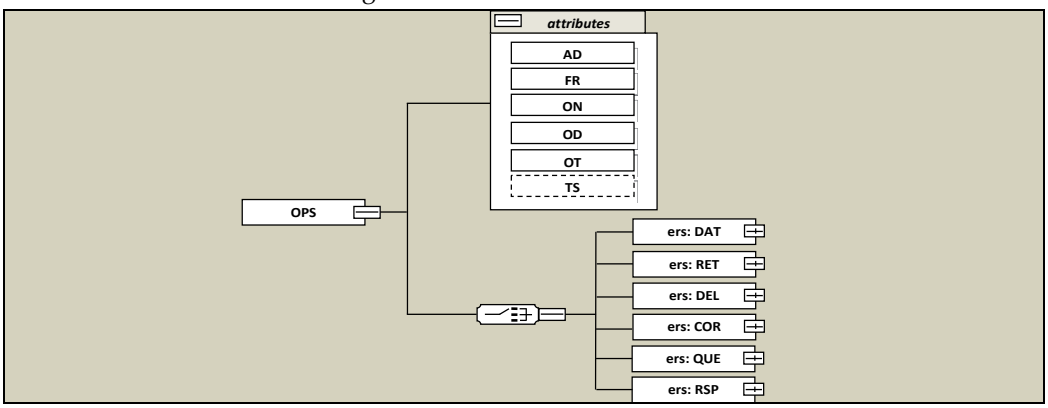


Fig. 30. Operations XSD

The attributes are:

AD: 2 letter country code of the recipient MS
FR: 2 letter country code of the sender MS
ON: Operation Number (AAA 99999999 999999)
OD: Operation date (Date the operation was initiated)
OT: Operation time (Time the operation was initiated)
TS: Test flag (if TS is present and free text is filled in)

Sub-elements (Operations) are:

DAT: Pushing of data to another MS. This happens under the following circumstances:

- When a vessel lands its catch in an MS other than the flag MS
- When a vessel intends to enter a port in an MS other than the flag MS
- When the first marketing takes place in an MS other than the flag MS
- When the first marketing does not take place in the MS where the fish was landed

RET: Acknowledgement of a previous operation
DEL: Deletion of previously sent data
COR: Correction to previously sent data
QUE: Query to pull data from another MS
RSP: Response to a pull query (QUE)

7.3.5 Web-Service Security

Data exchanges between MS are secured using SSL certificates. The web service uses a double hand-shake mechanism when establishing the SSL connection. The coastal state certificate is checked by the flag state and the coastal state checks the flag state certificate. No maximum downtime is set, best efforts are deployed to maintain the web-service availability 24/7.

8. References

022615 CEDER, 2008. Final Activity Report [online] available: https://ceder.jrc.ec.europa.eu/c/document_library/get_file?folderId=68984&name=DLFE-11544.pdf

SSP8-CT-2003-502153 SHEEL. 2004. Electronic logbook information to be exchanged. Report 1.2.Specific Targeted Research Project (STREP)

Barkai, A. & Bergh, M. Use and Abuse of data in fishery management. Deep Sea 2003: Conference on the Governance and Management of Deep-sea Fisheries. 27- 29 November 2003. Dunedin. Theme 4. Technology requirements.

FAO Fisheries Report. No. 761. Rome, FAO. 2005. 16p Data Formats and Procedures for Monitoring, Control and Surveillance. Report of the Expert Consultation on Bergen, Norway, 25-27 October 2004.

Henninger, H. 2009 Environmental Defense Fund Electronic Logbook Pilot (Phase 1) Final Report

Pope, J. G. & Symes, D. 2002. An ecosystem based approach to fisheries management, English Nature, ISSN, jncc.gov.uk [online] available: http://www.jncc.gov.uk/pdf/Achieving_2.pdf

Richard Banks Ltd. 2004. Evaluation of the NAFO OBSERVER SCHEME. Final Report. Fish 2002/03

12

Vegetation Mapping of the Mond Protected Area of Bushehr Province (SW Iran)

Ahmadreza Mehrabian*[1], Alireza Naqinezhad[2], Abdolrassoul Salman Mahiny[3], Hossein Mostafavi[1], Homan Liaghati[4] and Mohsen Kouchekzadeh[5]

[1]*Department of Botany, Faculty of Biological Sciences, Shahid Beheshti University, Tehran, Iran*

[2]*Department of Biology, Faculty of Sciences, University of Mazandaran, Babolsar, Iran*

[3]*College of the Environment, Gorgan University of Agriculture and Natural Resources Sciences, Iran*

[4]*Department of environmental and resources economic, Environmental Sciences Research Institute, Shahid Beheshti University, Tehran, Iran*

[5]*Department of planning and design of the environment, Environmental Sciences Research Institute, Shahid Beheshti University, Tehran, Iran*

1. Introduction

Ecosystems dominated by natural and semi-natural vegetation, occupy large portions of the Earth's surface and provide important services that should be preserved (Balvanera et al., 2001). Vegetation mapping is one of the most important phases of vegetation conservation. Satellite data such as those produced by Landsat and Spot have became ever more available to the public (Mahiny 2004) and advances in the automatic classification of satellite data make this technique an important tool for vegetation mapping nowadays (Jenes, 1996). The main goal of traditional vegetation mapping has been the identification of plant communities which are defined as the repetitive combination of species, structural types, growth forms and other terrain attributes (e.g. McGraw and Tueller, 1983; Wallens et al., 2000; Calarck et al., 2001; Zak and Cabido, 2002; Tobler et al., 2003). The mixing of traditional and advance methods can be used for comprehensive studies in the vegetation mapping. Due to the vulnerability of arid regions, comprehensive vegetation studies are necessary in these areas. The arid regions of the world occupy 26-35 percent of the earth's land surface, much of this wide region lying latitudes of between 15 and 30 degrees Northern and reflects various types based on the climatic conditions (Archibold, 1995). There is a shortage of knowledge on the vegetation of the Middle East, but investigations have been carried out on the ecology of individual plants and their associations (Zohary, 1973). The coastlines in the Middle East can support a diverse range of flowering plants; some are tolerant to highly saline soil and inundation to various degrees, while others inhabit low salinity soil. Regional climatic, topographic and geographic conditions are

assumed to be the main causes of vegetation forming in the desert and semi desert areas of Iran (Zohary, 1966-1986). In the hot southern parts of Iran with relatively high temperatures in both winter and summer and scant rainfall, a climatic regime governs which is similar to that of tropical northeast Africa, and the hot Sindian desert dominates, with occasionally more server temperature maxima and minima (Rechinger, 1963-1999; Zohary, 1966-1986 and Assadi, 1984). Iran is the classic country of great salines and Kavirs; Saline and alkaline soils are expanding in arid and semi-arid regions and cover 12.5% of the total land area of the country. These include Solenchak and Solontez soils, salt marsh soils, desert soils, Sierozem mixed with Solenchalk soils and saline alluvial soils (Dewan & Famouri, 1964). The elevation of the regions varies between -28 m on the shores of the Caspian Sea to about 1650 m in Kavire-Meyghan, Markazi Province (Akhani & Ghorbanli, 1993). Halophytic communities in Iran have been studied by many researchers. Zarinkafshe (1977) studied salty regions of Iran for flora, while Kunkel (1977) addressed the plants in the Hormoz, Qeshm and neighbouring islands. Moreover, some investigations have been carried out on the plants and vegetation of the Qeshm and Kish Island, the Persian Gulf region was also studied by Termeh & Moussavi (1982), Hamzehée (2001) and Attar et al. (2004).

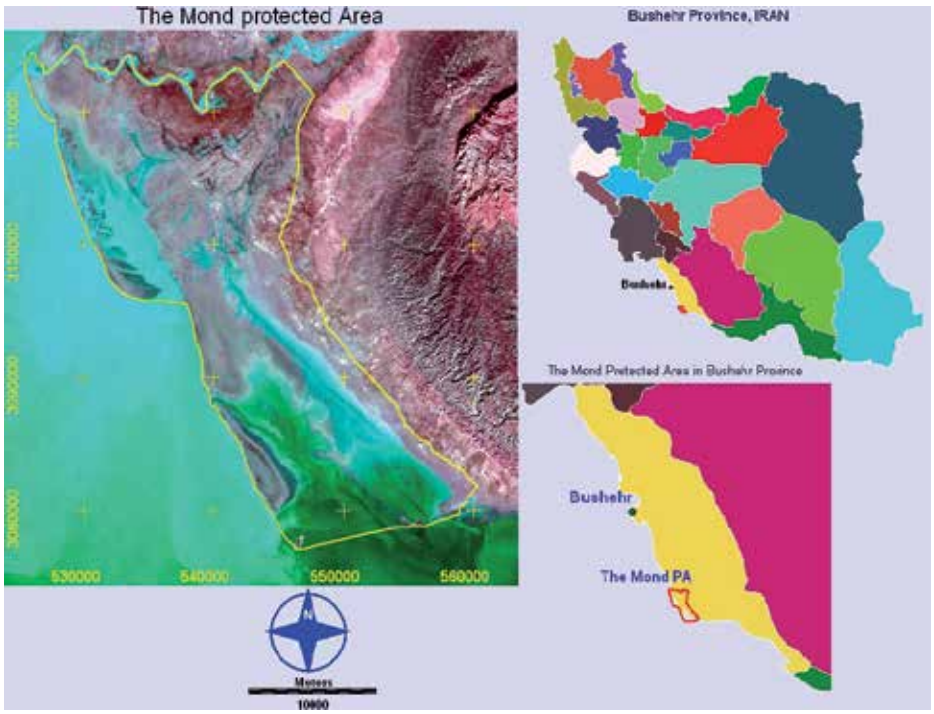


Fig. 1. Location of the Mond Protected Area in the coastal zone of Persian Gulf, Southern Iran.

The distribution of halophytic communities has been depicted cartographically by Mobayen & Tregubov (1970), Mobayen (1976), Freitag (1977), Carle and Frey (1977), Frey (1982), and Kramer (1984). Further physiognomic and ecological-geographic data on such plant communities have been provided by Kunkel (1977), Ghorbanli and Lambinon (1978), Breckle (1983), Assadi (1984), Frey and Probst (1986), and Akhani and Ghorbanli (1993).

Many researchers have been carried out on salt desert vegetation e.g., Zohary (1963, 1973), Termeh and Moussavi (1976), Leonard (1981-1988), Asri et al. (1995) and Asri and Ghorbanli (1997), Mehrabian et al., 2008). Due to the ecological and conservational values of Mond Protected area (Bushehr Province, Iran), this area was selected for vegetation mapping based on an integrative description of community structure and floristic attributes. The most important goals of this paper are 1) to provide a case of vegetation type mapping in the arid study area using field work, GIS and RS techniques, and 2) to compare these results with those other arid regions of the world.

2. Important

Mond Protected Area covers 53227 hectares and is located to the southwest of Bushehr between Northern latitude 27˚15' to 28˚ 45' and Eastern longitude 51˚15' to 51˚35' (Fig. 1). The average yearly temperature is 14 0C and annual precipitation is 155 mm. The study area is very flat, with its highest altitude at only 12 m. There are three physiographical units in the area including alluvial and colluvial fans, river alluvial plains and lowlands. The soils consist of alluvial, regosols, saline alkaline soils, solonchak and solontez. Administratively, tree islands called Omolgorm, Tahmadoon and Nakhiloo have been included in the area. Soils of the islands belong to the saline-alkaline type with a sandy texture (Fig. 2). The Mond area can be phytogeographically classified within the Sahara-Sindian region (Leonard, 1981-1988). However, it can also be classified in the Sudanian region (sensu Zohary, 1973).
Vegetation sampling was carried out during 2005 to 2007 when the soils and vegetation map units were studied. We used all four bands of the Spot5 Satellite XS imagery acquired on 26 January 2005 to investigate the vegetation attributes. Image projection was WGS 84, and the zone number was 39n. Unsupervised classification was conducted and sampling units were chosen for the field work. Owing to the sparse vegetation of the area and based on a visual examination of the image, we found that a combined visual, unsupervised and supervised method should be used for vegetation mapping of the area. For visual assessment, we generated a pseudo-color composite image using bands 2, 3 and 4 of the Spot5 imagery. We also used bands 2 and 4 to produce a preliminary NDVI layer (Normalized Difference Vegetation Index) showing crude vegetation density for the area. This was used along with unsupervised map and other ancillary data to sample vegetation in the field. Vegetation sampling was conducted following Braun-Blanquet cover scale (Braun-Balnquet, 1964). We used 156 geographically positioned sampling points to assess vegetation. The size of samples varied between 4 m2 to 32 m2 based on the minimal area taken at each point. The field work and satellite images were mutually complementary. Dominant and companion species and their coverage were recorded in samples. Vegetation types for each area were recognized according to the occurrence of specific perennial species accompanied by some companion species. These dominant species were used for naming each vegetation type. Species emerging in each season were added to the plant list of each vegetation type during the investigations. Geo-positioning of sampling points made using with GPS The visual boundary of the map units was digitized and stored on GIS for future analysis. Using data gathered on the field, unsupervised classification of the Spot5 XS bands through is cluster module of the Erdas Imagine 8.4 software (Leica Geosystems Geospatial Imaging) and visual examination of the pseudo-color composite of the area, we distinguished different vegetation types (as map units on the GIS map) delineated them on the image and produced

a final vegetation map. Vegetation map units are defined as areas where vegetation is relatively homogenous (Samira et al., 2001). A map unit is defined and named according to the taxonomic classification of the dominant community. Each map unit for the area comprised a vegetation type with the exception of those areas empty of vegetation. Water bodies and bare lands. Information about soils in the study area is based on previous soil studies in different parts of the area. Based on these studies, four major types of soil were recognized which can be subdivided to 13 detailed soil units (Fig. 2). Moreover, a classification of the habitats in the study area was provided according to fieldworks and complementary GIS methods which helped in vegetation type mapping.

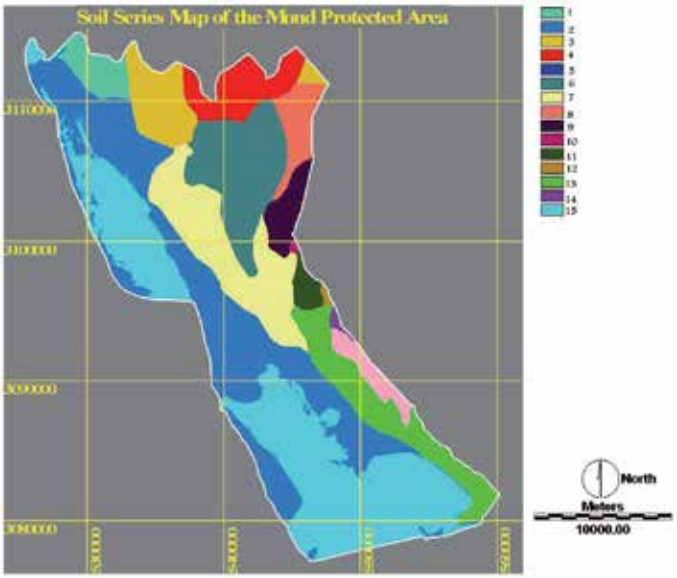


Fig. 2. Soil map of the Mond Protected Area. (1,3,4,13=Alluvial 2, 7, 8,10,11, 12 14=Solenchak 9=Regosoil, 15=Water. Different number represent gradient in each one.

Based on field observation and supported by satellite maps, three major habitat zones in the study area i.e. coastal zone, riverine zone and inland zone were recognized. These habitat zones are covered with three broad plant formations in the area. These are shrublands (northern parts), bushy grasslands (inland parts) and mangrove forests (southeastern parts of the coastal zone). In each formation, different vegetation types were recognized on the basis of field vegetation sampling guided by an unsupervised classification of the Spot XS data. Twelve vegetation types were recognized in the field that showed a good compatibility with the satellite image (map units). Moreover, large parts of the study area near the sea coast were bare lands or filled by sea water. These parts together with cultivated areas were defined as separate map units on the final map manipulated by GIS (Fig. 3). The vegetation types were variable in size and flora composition. Some vegetation types covered more than 20% of the area while others had coverage of less than 5% (Table 1). Some vegetation types, e.g. Halocnemum strobilaceum, Suaeda aegyptiaca, Lycium edgworthii are widely distributed, but Ephedra foliata (Nakhiloo island), Salsola drummundi (Eastern area), Atriplex leucoclada (Nakhiloo island), Salicornia europaea-Suaeda heterophylla (northwestern area) and Avicennia marina (south eastern area) are restricted to small

habitats (Fig. 3). There are three vegetation types (Arthrocnemum macrostachyum, Ephedra foliata and Cyperus conglomerates-Halopyrum mucronatum) on Omolgorm Island, two vegetation types (Arthrocnemum macrastachyum and Cyperus conglomerates-Halopyrum mucronatum) on Tahmadoon Island and three vegetation types (Arthrocnemum macrostachyum, Cyperus conglomerates-Halopyrum mucronatum and Atriplex leucoclada on Nakhiloo Island (Fig. 3). The density of vegetation was presented as a map using bands 2 and 4 of the Spot XS data in NDVI calculation. The density indicated an increase in vegetation southward to northward and westward to eastward (Fig. 4).

A-Shrubland formation (along the Mond River)

1-Tamarix leptopetala-Phragmites australis vegetation type (no. 16 in Fig. 3)

Tamarix leptopetala Bge. and Phragmites australis (Cav.) Trin. ex Steud are two dominant species of this vegetation type. Phragmites australis is a hygrophilous plant in rivers and saline marshes (Asri and Ghorbanli 1997). Tamarix leptopetala is one of the most characteristic genera in the Middle East (Zohary 1973). It comprises of about 35 species in the Middle East, many occurring in saline habitats, saline river beds and desert wadies, saline and sandy soils, estuaries of central depressions and vast areas of inland salines with a relatively high water table (Zohary 1973).

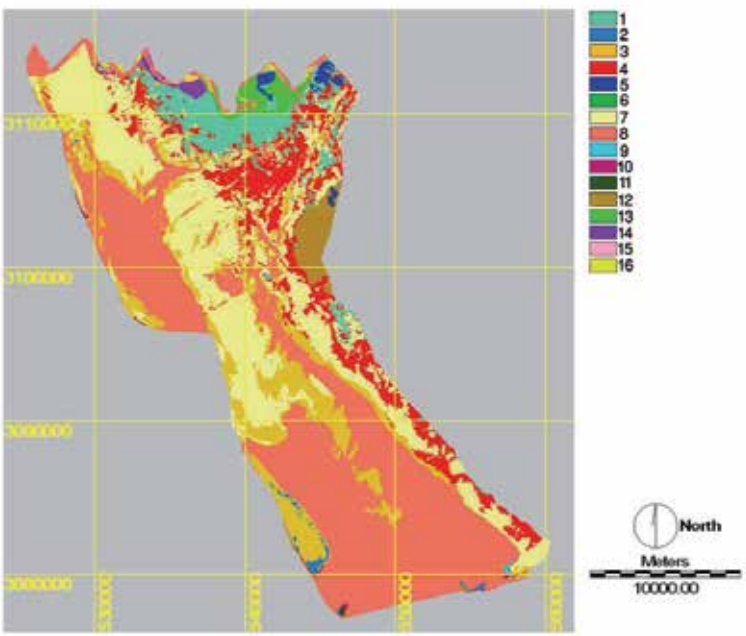


Fig. 3. Vegetation types and map units in the Mond Protected Area (Zohary, 1973): 1- Suaeda aegyptiaca, 2- Arthrocnemum macrastachyum, 3- Bare lands, 4- Halocnemum strobilaceum ((high density), 5- Farmlands, 6- Halocnemum strobilaceum (low density), 7- Water, 8- Avicennia marina, 9- Cyperus conglomerates-Halopyrum mucronatum, 10- Atriplex leucoclada, 11- Salsola drummondii, 12- Lycium edgworthii, 13- Suaeda fruticosa, 14- Salicornia europaeae- suaeda heterophylla, 15- Tamarix leptopetala

The Tamarix leptopetala- Phragmites australis vegetation type is situated in the banks of the Mond River. The first zone of this riverine vegetation belt comprises Phragmites australis and towards the inland Tamarix leptopetala replaces it and dominates over a wide area and is also dominant in many small dried rivulets and stream beds inside the area and this vegetation type is the most important vegetation type of the Iranian salt lands. This vegetation type shows a coverage of 50-75 % over the area. The most important companion species are Alhagi persarum Boiss. & Buhse., Artemisia scoparia Waldst., Cressa cretica L., Cyperus rotundus L. Spergula fallax (Lwe.) E. H. L. Krause and Suaeda aegyptiaca (Hasselq.) Zoh. This vegetation type is situated in alluvial soils.

2-*Lycium edgworthii* vegetation type (no. 12 in Fig. 3)

Lycium edgworthii is distributed over certain parts of Iran. The vegetation type dominated by latter species is found on the margins of wet salty inland soils and also the external zone of the Mond River after the Tamarix-Phragmites vegetation type. Lycium edgworthii has a high density in some parts of the river margin. The coverage of this vegetation type varies between 60-70%. The most important companion species of this vegetation type are Aloina aloides (Schultz) Kindb., Anagallis arvensis L., Bromus rubens L., Calendula persica C. A. Mey., Centaurium pulchellum (Swartz.) Druce., Cuscuta chinensis Lam., Lophochlora phleoides (Vill.) Reichenb., Phlaris minor Retz. This vegetation type occupies alluvial soils in the study area.

3-*Suaeda fruticosa* vegetation type (no. 13 in Fig. 3)

Suaeda fruticosa (L.) Forssk. is a dark green bushy plant which distributed across saline lands of the Sahara-Sindian region that in places penetrate into the Irano-Turanian region. This species is geographically distinct in central and southern saline (Zohary, 1973). The species is dominant in the vegetation type distributed over northern parts regions of the study area as well as Omolgorm, Tahmadoon and Nakhiloo Islands with alluvial soils. The coverage of this vegetation type varies between 75 and 100 %. The most important companion species of this type are Aeluropus lagopoides (L.) Trin. ex Thwaites, Cyperus rotundus L. Ephedra foliata Boiss and Kotschy., Lycium edgeworthi and Salsola drummondii Ulbrich.. Although this vegetation type shows some mixed situations with Lycium edgworthii vegetation type in some parts of the study area, there are many pure spots of this vegetation type dominated by Suaeda fruticosa in the area.

B-bushy and grassland formations (vast inland area)

4,7- *Halocnemum strobilaceum* vegetation type (no. 4 & 7 in Fig. 3)

Halocnemum strobilaceum is a dwarf shrub or richly branched perennial herb turning dark green as an adult. This species is an penetrative element to coast lines and inland marshes. It occupies broad belts on the fringe of salt lakes and Kavirs with relatively higher water table (Akhani & Ghorbanli, 1993). In Iranian inlands, it forms dense and almost pure stands for hundreds of miles around the smaller and large salt pans and also in «lost rivers» (Zohary, 1973). It also covers broad zones in the South and South west of Iran, extending far inwards from the seashores of the Persian Gulf, the Gulf of Oman and the Arabian Sea (Zohary, 1973). On the peripheries of most of the inland salines, it forms a pioneer halophytic

community or the second phase after the Salicornia europaea vegetation (Zohary, 1973). The vegetation type dominated and characterized by Halocnemum strobilaceum is the largest vegetation type and distributed in almost all of inland parts of the study area with alluvial soils. The coverage of this type is 5-75%. Due to the intensively salty conditions of the habitats of this vegetation type, companion species are very poorly represented. The most important companion species are Aelurupus lagopoides (L.) Trin. Ex Thwaites., Asphodelus tenuifolius Cav., Gynandriris sisyrinchium (L.) Parl., Plantago amplexicaulis Cav., Plantago coronopus L., Plantago psyliium L., Plantago stocksii Boiss. & Decne., Psylliostachys spicata (Willd.), Sonchus tenerrimus L. and Suaeda heterophylla (Kar. et Kir.) Bge.

6-*Salsola drummundi* vegetation type (no. 12 in Fig. 3)

The genus Salsola comprises about 30 species in the Middle East. Except for a few annual and uncommon species, they are mostly dominant species in various plant communities (Zohary, 1973). Most Salsola species are xero-halophytes (Zohary, 1973). Salsola drummundi is the dominant species for the vegetation type distributed over the eastern parts of the study area with the Solonchak soils. The coverage of this vegetation type is 50-75 %. The most important companion species are Atriplex leucoclada (Boiss.) Aellen., Limonium iranicum (Bornm.) Lincz., Plantago psyliium L., Salsola cyclophylla Barker. and Suaeda aegyptiaca (Hasselq.) Zoh.

7-*Arthrocnemum macrastachyum* vegetation type (no. 2 in Fig. 3)

Arthrocnemum macrastachyum as a leafless, bushy succulent species with rather deep roots that is very common in the west over part the Middle East (Zohary, 1973). This species with the main distribution in Mediterranean region, occupies large stretches of littoral marshes (Akhani & Ghorbanli, 1993). It penetrates however deeply into desert areas such as the Dead Sea area, inner Anatolia, the Syrian Desert and Iraq. In the coastal marshes of the East Mediterranean, this species forms large pure stands along the salt-water bodies. Arthrocnemum macrostachyum is distributed in northwestern, southern (Omolgorm Island) and southwestern (Tahmadoon and Nakhiloo Islands) areas. The Arthrocnemum macrostachyum vegetation type is unique to high salty and wet soils on the margins of salt lakes, banks and estuaries of high saline rivers and streams and of littoral marshes of the Persian Gulf. In other localities it is less exclusive but still very abundant (Zohary, 1973). The coverage of this vegetation type is 75-100 % and it occurs on alluvial soils. Companion species are Atriplex leucoclada (Boiss.) Aellen, Cistanche tubulosa (Schrenk.) R. Wight., Halocnemum strobilaceum M. B., Limonium Iranicum (Bornm.) Lincz., Salicornia europaea L. and Suaeda heterophylla (Kar. et Kir.) Bge.

8-Salicornia europaea-Suaeda heterophylla vegetation type (no. 14 in Fig. 3)

Salicornia europaea and Suaeda heterophylla are two dominant annual species in this vegetation type distributed over the north west of the area. This vegetation type constitutes the first vegetation zone in salty habitats near maritime and estuary areas with Solenchak soils. The coverage of this type is 75-100%. Companion species of this vegetation type are Arthrocnemum macrastachyum and Halocnemum strobilaceum. This vegetation type was previously considered to be one of the obligatory hygro-halophtic communities in the classification presented by Akhani and Ghorbanli (1993).

9-Cyperus conglomeratus-Halopyrum mucronatum vegetation type (no. 10 in Fig. 3)
Cyperus conglomeratus and Halopyrum mucronatum are dominant species of this vegetation type which is located on the coastal shores to the south east of the study area with coverage of 25-50 %. The most important companion species are Atriplex leucoclada (Boiss.) Aellen., Cistanche tubulosa (Schrenk.) R. Wight., Heliotropium bacciferum Forssk., Salsola jordanicola Eig., Senecio vulgaris L. and Helianthemum lippi (L.) Pers. This vegetation type is found in alluvial or sandy soils.

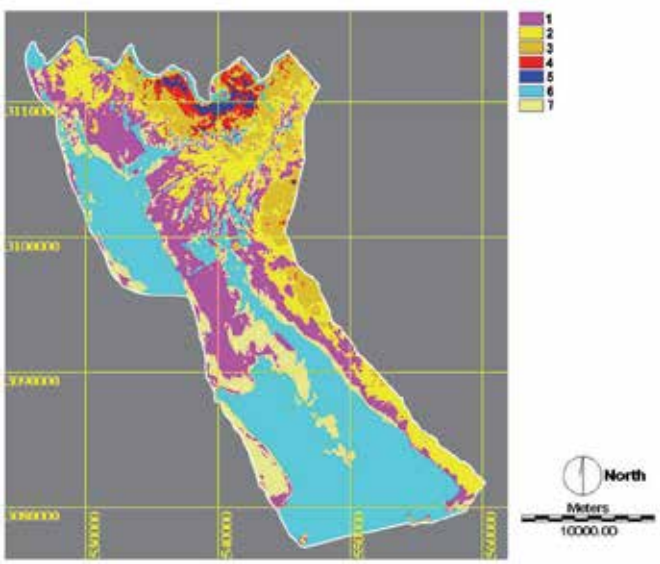


Fig. 4. Vegetation density map of the Mond Protected Area (1=very scanty, 2=scanty, 3=semi-scanty, 4=low dense, 5=dense, 6=water, 7=bare land)

10- Atriplex leucoclada vegetation type (no. 11 in Fig. 3)
The genus Atriplex is represented by the number of dominant species in various vegetation types of which some species (e.g. A. halimus and A. leucoclada) are distributed in salty habitats (Zohary, 1973). The Atriplex leucoclada vegetation type is distributed on Nakhiloo Island (south west part). The coverage of this vegetation type is 50-75 %. The most important companion species are Aizoon canarensis L., Cornulaca aucheri, Heliotropium bacciferum Forssk., Spergularia marina (L.) Griseb., Polycarpon tetraphyllum (L.) L. and Senecio vulgaris L. This vegetation type occupies sandy soils.

11-Ephedra foliata vegetation type (no. 6 in Fig. 3)
Ephedra is an excellent example of a world-wide genus represented in the Middle East by about ten species (Zohary, 1973). The Ephedra foliata is a Sudanian element in the Middle East and confined to the Gulf of Eilat, southern Arabia and southern Iran (Zohary, 1973). Ephedra foliata vegetation type is distributed only on Omolgorm Island, and over very small patches in northern parts of the area. The coverage of this type is 25-50 %. The companion species are Cyperus conglomerates Rottb., Senecio vulgaris L. and Suaeda fruticosa (L.) Forssk. This vegetation type occupies sandy soils in the study area.

C-Mangrove forest formations (coastal vegetation)

12-*Avicennia marina* vegetation type (mangrove forest) (no. 9 in Fig. 3)

Mangrove forests are sensitive habitats due to their ecotonic (transitional) condition, i.e. they are affected by both marine and land ecosystems. These forests are confined to shores of the Persian Gulf (including its islands) and Oman Sea in the Middle East. The Avicennia marina vegetation type in Mond protected area is the furthest western range of Mangrove distribution in the world. Avicennia marina is the typical component of this vegetation, a species belonging to eastern mangroves that are distributed broadly in throughout world (Zohary, 1973). This vegetation type is somewhat mixed with other vegetation types, e.g. Arthrocnemum macrastachyum and Halocnemum strobilaceum vegetation types. The coverage of this type is 50-100%. It occupies alluvial and Solenchak soils in the area. Other companion species of this vegetation type are Arthrocnemum macrastachyum, Cistanche tubulosa (Schrenk) R. Wight and Salicornia europaea.

D-vegetation affected by human activities (anthropogenic effects)

13-Suaeda aegyptiaca vegetation type in abandoned farmlands (no. 1 in Fig. 3)

There are about 22 species of this genus in the Middle East, of which only a few are important for their broad distribution. They can be divided into hydro- and xero-halophytes and occur in both littoral and inland saline habitats. Suaeda aegyptiaca is a ruderal species and therefore the vegetation type dominated by this species establishes in abandoned farmlands. This vegetation type with coverage of 50-75 % occurs mainly in northern parts of the area with alluvial soils. The companion species includes Artemisia scoparia Waldst. & Kit., Atriplex leucoclada, Calendula persica C. A. Mey., Chrozophora hierosolymitana Spreng., Senecio vulgaris L., Sonchus oleraceus L., Sporobolus arabicus Boiss. and Stellaria media (L.) Cyr.

14-Farmlands (anthropogenic vegetation) (no. 5 in Fig. 3)

One of the important categories in the vegetation map prepared is farmlands representing areas strongly affected by man and livestock. This map category is largely the result of human agricultural activity, and is comprised of three large units, namely cultivated, segetal and ruderal (Zohary, 1973). This map unit is distributed over the margins of the northeastern and eastern parts with alluvial and Solenchak soils.

E-Map units without vegetation

15-Bare lands (no. 3 in Fig. 3)

Bare lands constitute one of the largest units in the vegetation map of the study area located along the coastal line. Due to high salinity, vegetation density was as low as nearly 0. This vegetation type is established in Solenchak soils.

16-Water (no. 8 in Fig. 3)

The eastern and southern parts of the study area are fully occupied by water connected to the Sea (Fig. 3).

3. Conclusion

The current study is a new approach to vegetation mapping in Iran using remote sensing (RS) and the geographic information system (GIS). One of the most important problems in remote sensing of desert vegetation is that the reflectance from soil and rocks is often much greater than that of sparse vegetation and this makes it difficult to separate out the vegetation signal (Gates et al. 1965); and there is spectral variability within shrubs of the same species (Duncant et al., 1993). These properties hamper accurate classification of vegetation in these areas. This study proved the usefulness of the Spot XS imagery for vegetation mapping but also it showed that in arid regions, mapping can only be completed satisfactorily if it is accompanied by extensive field sampling, visual image interpretation and hybrid classification methods. Thus, hybrid approaches that include field work, GIS and RS are required in such circumstances.

In a study of the halophytic vegetation of the Middle East, Zohary (1973) discussed that the most halophytic communities of Iran belong to the phytosociological classes Halocnemum strobilacei irano-anatolica. Breckle (1983) classified the halophytic vegetation of Iran and Afghanistan as follows: (1) saline flats (very sparse vegetation, soil with very high salt content); (2) euhalophytic vegetation (halo-hammada on gravel-sandy but probably clay soil) and (3) mesohalophytic vegetation (with less salt in the soil profile). Frey and Probst (1986) provided a geographical classification for total halophyte vegetation including (1) salty pans of the central Iranian undrained basin and their peripheries; (2) shore zones of salt lakes; (3) areas the Persian Gulf and (4) southern Caspian coastal zone. Akhani and Ghorbanli (1993) with a geographical-ecological approach classified halophytic communities of Iran as (1) Halocnemum strobilaceum communities (on muddy salt flats); (2) obligatory hydro-halophytes communities; (3) Tamarix communities; (4) Hydrophilous euryhlophytic communities; (5) Mangrove communities; (6) Hydrohalophytic plant communities; (7) Halophytic shrub communities on salty and dry soils; (8) Herbaceous perennial and hemicryptophyte halophytic communities; (9) xeromorphytic communities with salt-tolerant xerophytes and (10) annual halophytic communities.

Mond Protected Area is one of the best indicators of halophytic vegetation in the arid lands of Iran. Salty river, salty pans, coastal and muddy salt flats induce habitat diversity supporting the establishment of different halophytic vegetation types (see Figs. 2, 3, 4). This study identified 15 map units of which 12 are concerned with vegetation types. These vegetation types reveal plant communities adapted to different habitats and environmental gradients in the area.

Most vegetation types identified in the current study were previousy recognized in the other arid or salty areas of Iran (Akhani and Ghorbanli, 1993; Asri & Ghorbanli, 1997; Ghahreman et al., 2000; Alaei et al. 2001). The occurrence of Halocnemum strobilaceum, Avicennia marina, Salicornia europaea-Suaeda heterophylla, Suaeda fruticosa, Tamarix leptopetala and Arthrocnemum macrostachyum vegetation types is identical with the results of Akhani and Ghorbanli (1993). The study area can be considered as the first and third vegetation zones defined in the classification offered by Breckle (1983). This research revealed that the establishment of the vegetation types is largely regulated by edaphic factors (texture, chemical composition and humidity). Some parts of the area are impacted by agricultural activity, grazing and the destruction of vegetation. These activities have led to major changes in habitats and ecosystems and have threatened wildlife. These ecosystems represent landscapes that can be restored and managed for uses such as ecotourism and recreation.

Vegetation types	Surface (ha)	Percentage (%)
Halocnemum stroboilaceum	19320.1	36.30
Suaeda aegyptica	3044.52	5.72
Salsola drummondii	1080.88	2.03
Lycium edgworthii	708.72	1.33
Suaeda fruticosa	263.96	0.50
Arthrocnemum macrastachyum	209.52	0.39
Tamarix leptopetala	189.04	0.36
Cyperus conglomerates-Halopyrum mucronatum	87.04	0.16
Atriplex leucoclada	18.92	0.04
Avicennia marina	14.56	0.03
Ephedra foliata	8.92	0.02
Salicorniaeuropaea-suaeda teterophylla	9	0.02
Farmalnds	364.52	0.68
Water	21052.2	39.55
Bare lands	6855.12	12.88
Total	53227	100.00

Table 1. Areal coverage of vegetation types in the Mond Protected Area

4. References

Akhani. H., Ghorbanli M. (1993). A contribution to the halophytic vegetation and flora of Iran, in H. Leith and A. Al- Masoom (eds.). Towards the Rational use of High Salinity Tolerant Plants, vol. 1, p. 35-44, Kluwer Academic publishers, Netherlands.

Archibold, O. W. (1995). Ecology of World vegetation. London. Chapman & Hall. 1-510.

Asri Y., Ghorbanli M. (1997). The halophilous vegetation of the Orumieh lake salt marshes, NW. Iran. Plant Ecology 132, 155-170.

Asri Y., Hamzehee B., Ghorbanli M. (1995). Etude Phytosociologique de la vegetation halophile de l'est du lac Orumieh (Nord Ouest de l'Iran). Doc. Phytosociology, 15: 299-308.

Assadi M. (1984). Studies on the autumn plants of Kavir, Iran. Iran. j. Bot. 2, 125-148.

Attar F., Hamzehée B, Ghahreman A (2004). A Contribution to flora of Qeshm Hsland, Iran. Iran. J. Bot. 10, 199-218.

Balvenera, P., Daily, G., Ehrlich, P., Ricketts, T., Bailey, S., Kark, S., Kermen, C. & Pereira, H (2001). Conserving biodiversity and ecosystem services. Sciences, 291, 2047.

Braun-Blanquet J. (1964). Pflanzensoziologie: Grundzüge der Vegetationskunde. 3. neu bearb. Aufl. – Springer-Verlag. Wien.

Breckle S.W. (1983). Temperate deserts and semi-deserts of Afghanistan and Iran. In: N. E. West (ed). Ecosystems of the World. 5 : Temperate Deserts and Semi-Deserts. Elsevier. Amsterdam. 271-319.

Calarck P.E., Seyfried M.S., Harris B. (2001). Intermountain plant community classification using Landsat TM and spot HRV data. Range Management 54, 152-160.

Carle R., Frey W. (1977). Die vegetation des Maharlu-Beckenis bei Shiraz (Iran) Unter besonderer berucksichitigung der vegetation im bereich der suss-und salzawasserquellen am seeufer. Beih. TAVO, Reihe A (Naturwiss) 2, Dr. Ludwig Reuchert Verlag, Wiesbaden.

Dewan M., Famouri J. (1964). The Soils of Iran. FAO, Rome.

Duncant J., Stow D., Franklin J., Hope A. (1993). Assessing the relationship between spectral vegetation indices and shrub cover in the Jourdan Basin, New Mexico. International journal of Remote Sensing, 14, 3395-3416.

Frey W., Probst W. (1986). A synopsis of the vegetation of Iran. In : H Kürschner (ed). Contribution of the vegetation of southwest Asia. Beih. TAVO. Naturwiss, 24, 9-24.

Frey W. (1982). Maharlu-Becken bei Shiraz (Iran). Mittlerer Teil. Vegetation, 1: 100,000, Karte AVI 10.2 TAVO, Dr. Ludwig Reichert Verlag, Wiesbaden.

Freitag H. (1977). Turan Biospher Reserve, prelimnary vegetation map, pp. 86-89. im : Spooner, B. (ed.). Case Study on Desertification Iran : Turan. Department of Environment, Tehran.

Gates D.M., Keegan H.J., Schleter J.C., Weidner V.R. (1965). Spectral properties of plants. Applied Optics, 4, 11-20.

Ghahreman A., Hamzehée B., Attar F. (2000). Natural Vegetation Map of the Qeshm island, Publication of Qeshm Free Area.

Ghorbanli M., Lambinon J. (1978). Permier. Apereu de la zonation de la vegetation halo-gypsophile du lac Ghom. Lenjunia Rev. Bot. 92 :1-20.

Hamzehée B. (2001). Application of Anaphyto Software in Phytosociological data analysis (A Case Study: Eroded Terraces of the Queshm Island). – Research Institute of Forests and Rangelands. Tech. Publ. 237. Tehran. 379 pp. (In Persian).

Jensen, J. R. (1996). Introductory Digital Image Processing. A Remote Sensing Perspective (2nd ed). New Jersey : Prentice Hall.

Kramer W. (1984). Mittleter Zagros (Iran) Vegetation, 1:600,000, Karte AVI 6 TAVO, Dr. Ludwig Riechert Verlag, Wiesbaden.

Kunkel G. (1977). The vegetation of Hormoz, Qeshm and neighboring islands (Persian Gulf area). Flora et vegetatio Mundi 6. 186p.

Leica Geosystems Geospatial Imaging (2002). Erdas Imagine 8.4 software, USA.

Leonard J. (1981-1988). Contribution a l'étude de la flore et de la végétation des déserts d Iran. Fasc 1-9. Meise.

McGraw J.F., Tueller P.T. (1983). Landsat computer-aided analysis teqniques for range vegetation mapping. Journal of Range Managements, 36: 627-631.

Mehrabin A. R., Naqinezhad A., Mostafavi H., Kiabi B., Abdoli A. Flora and habitats of Mond Protected area. Mohit Shenasi., Vol 46(1) : 1-18

Mobayen S. (1976). Structure geobotanique du Lut. Acta Ecol. Iranica 1, 73-86.

Mobayen S., Tregubov V. (1970). Carte de la végétation naturelle de l,Iran, 1; 250,000. University Tehran, UNDP/FAO no.Ira 7.

Rechinger K.H. (ed.) (1963-1999). Flora Iranica. Vien, Graz : Akademische Druck.

Salman Mahiny A. (2004). A Modelling Approach to Cumulative Effects Assessment for Rehabilitation of Remnant Vegetation. Ph.D. Thesis, ANU, Australia.

Omar. Samir. A. S., Misak. R., King. P., Shahid. Shabbir. A., Abo-Rizk. Hanna, Grealish. G. and Roy. W. (2001) Mapping the vegetation of Kuwait through reconnaissance soil survey. Journal of Arid Environment, 48: 341-355.

Termeh F., Moussavi M. (1976). Contribution a l'étude de la végétation automnale du Dashte-Lut. Institut de Recherches Entomologiques et Phytopathologiques d Evine, Department of Botany. Tehran. No. 7.

Termeh F., Moussavi M. (1982). Plants of Kish Island, Iran. Wildenowia, 12, 253-286.

Tobler M.W., Cochard R., Edward P.J. (2003). The impact of cattle ranching on large-scale vegetation patterns in coastal Savanna in Tanzannia. journal of Applied ecology, 30, 430-444.

Wallens J., Milington A.C., Hichkin W., Arquepino R., Jones S. (2000). Mapping humid tropical vegetation in eastern Bolivia in R. Alexander and A. C. Millington (Eds) Vegetation Mapping. Chichester : John Willey and sons.

Zak M.R., Cabido M. (2002). Spatial pattern of the Chaco vegetation of central Argentina. Integration of remote sensing and phytosociology. Applied Vegetation Sciences, 5, 213 -226.

Zarrin-kafche M. (1977). Etude de la relation existent entre la compositon chemique des plants et celle des Sols du Sud de karadj (pressé de Tehran). Acta Ecology. Iran. 1, 60-69.

Zohary M. (1963). On the Geobotanical Structure of Iran. Bull. Res. Counc. Isr., Sect. d, Bot., Suppl. 1. 113p.

Zohary M. (1966-1986). Flora Palestine, Vol : 1-4. The Israel Academy of Sciences and Humanities.

Zohary M. (1973). Geobotanical Foundations of the Middle East. 2 vols. Gustav Fischer Verlag. Amesterdam.

13

Earth to space link

Mandeep Jit Singh, Mardina Abdullah, Baharudin Yatim,
Mahamod Ismail and Wayan Suparta
Malaysia

1. Introduction

The measurements were taken continuously from 1st January 2006 to 31st December 2008. The measured rainfall data was analyzed and presented in this chapter. Three years averaged measured rain rate and rain attenuation were compared with the cumulative probability distributions existing prediction models. Some rainfall data from selected tropical climates sites were used to make a comparison of rain rate and rain attenuation prediction.

2. The Variation of Rainfall Amount

Malaysia experiences heavy rain throughout the year and the rainfall distribution is patterned by monsoon activities. The Northeast monsoon is from October to March and the Southwest monsoon is from May to September. The Northeast Monsoon is the major rainy season in the country. The Northeast Monsoon blows from South China Sea during the months of October to March, this affects high brought rainfall to the East Coast of Peninsular Malaysia and a few other locations at Peninsular Malaysia. It was recorded by Malaysian Meteorological Department that the Northeast Monsoon has give the highest rainfall in all location in Malaysia because the wind of this monsoon blows from South China Sea without obstruction by island, peninsular and mountain. Monsoon weather systems that develop in the conjunction with cold air outbreaks from Siberia produce heavy rains, which often cause severe floods along the east coast of Malaysia. The Southwest monsoon is comparatively drier throughout the country except for the state of Sabah. During this season, most states experience monthly minimum rainfall. This is attributed to relatively stable atmospheric conditions in the equatorial region. In particular, the dry condition in Peninsular Malaysia is accentuated by the rain shadow effect of the Sumatran mountain range. During the inter-monsoon periods, winds are light and variable. Morning skies are often clear and this favours thunderstorm development in the afternoon. In the states of west coast of Peninsular Malaysia, thunderstorms contribute to a mean monthly rainfall maximum in each of the two transition period.
Figure 1 shows the monthly variation of rainfall amount at USM. A relationship between the monsoon activities and monthly rainfall amount can be studied. Based on the measurement, the total rainfall for the year 2006 was 1835 mm, for the year 2007 was 2065 mm and for the

year 2008 was 2111 mm. The total average annual rainfall at USM for three years was 2004 mm. The total average rain accumulation for the Northeast Monsoon (October – March) and Southwest Monsoon (May – September) was 1192 mm (59% of the annual averaged) and 640 mm (32% of the annual averaged), respectively. The maximum rainfall amount was in November for the year 2006 that is 274 mm, in September for the year 2007 that is 287 mm and in December for the year 2008 that is 280 mm. The minimum rainfall amount was in July for year 2006 that is 66mm, in August that is 58 mm and in May that is 5 mm. From this figure, it indicates that the rainfall amount for Northeast Monsoon is higher than Southwest Monsoon.

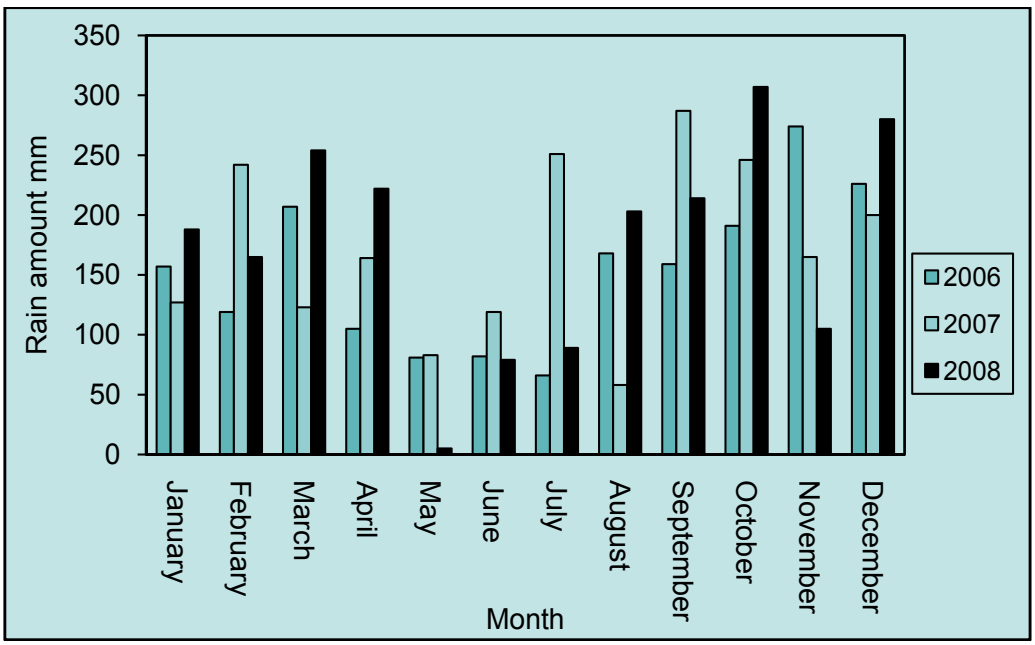


Fig. 1. Variation of monthly rainfall amount at USM for the year 2006, 2007 and 2008

3. Testing of Prediction Models

The achievement of high availability targets in advanced satellite link design requires a deep knowledge of the radio channel behavior. The effects due to different atmospheric causes, such as precipitation, clouds, atmospheric gases and tropospheric scintillation can be measured quite accurately by means of satellite beacon signals. Most prediction models of probability distribution function (pdf) of exceeding rain attenuation in slant path, based on rain rate pdf, when tested against concurrent satellite beacon measurements show large errors (Matricciani, *el at*., 2008).

Two types of prediction tests used for this analysis:

1. The percentage error
2. The Real Mean Square (RMS) error

For certain percentage of time (from 0.001 to 1 percent of the year), for which data are available, percentage relative error, E_{rel} (percent) between the predicted value and the measured value are calculated

$$E_{rel} = \frac{A_{predicted} - A_{measured}}{A_{measured}} \times 100\% \quad (1)$$

The mean error, μ_e and standard deviation, σ_e are used to calculate the Root Mean Square, D_e (RMS). The parameter is defined as follows

$$D_e = [(\mu_e)^2 + (\sigma_e)^2]^{1/2} \quad (2)$$

According to evaluation procedures adopted by the CCIR the preferred prediction method is the one producing the smallest RMS values (CCIR, 1983)

4. Statistical analysis rainfall

4.1 The correlation of Rain Rate and Rain Attenuation

Fig. 2 shows a typical variation of rain rate and rain attenuation measured on a rainy day (2nd March 2008). The rapid small-scale fluctuations of the signal level are because of ionospheric and tropospheric scintillations and slower variation correspond to rain events - fade. At certain times, the decrease in the relative signal strength (attenuation) and the increase in the rain rate are coincidental, however at other times, the decrease in the signal strength was leading the increased rain rate. It is probably the result of raindrops held aloft by updrafts in the cloud, and then their time of arrival at the ground would be considerably delayed. This delay time is variable and is as long as 4–8min (Ramachandran, *et. al.*, 2004, Mandeep, 2008). In the data presented in Fig. 2, delay (corrections) time of 8 min was accounted for when investigating the correlation between rain rate and rain attenuation.

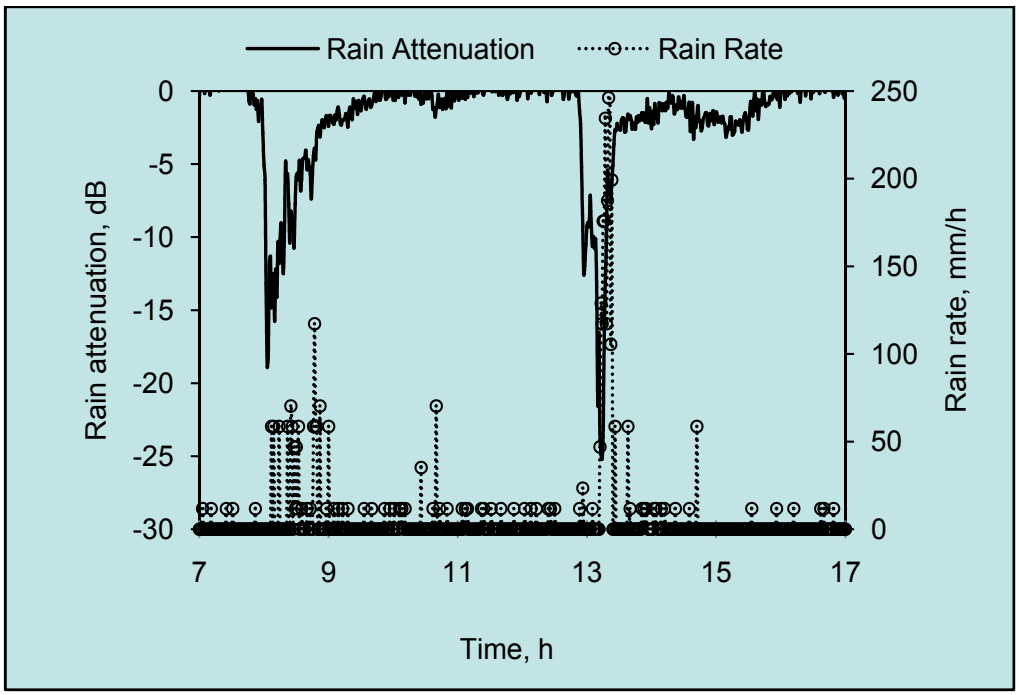


Fig. 2. Time series record of attenuation and rain rate

To determine the nature and strength of the relationship between rain rate and rain attenuation, correlation analysis has been done to indicate the strength of the relationship between rain rate and rain attenuation is high. Fig 3 shows the relation between rain rate and rain attenuation. Form that figure, the dots are obtained by straightforwardly plotting rain attenuation versus rain rate and the solid line is the best fit curve obtained by fitting rain attenuation versus rain rate. The R- square is 91.39%. It indicates that the correlation between rain rate and rain attenuation is high.
The attenuation for the site was a logarithmic function of rain rate and can be simplified as:

$$A\ (dB) = 0.054R^{1.2315}$$

where A denotes as attenuation and R denotes as the site rain rate, mm/h. This empirical equation would be helpful to system designer to determine the link margin at the site. Calculating by using this empirical equation, a threshold of 19.6 dB is considered as the economical limits.

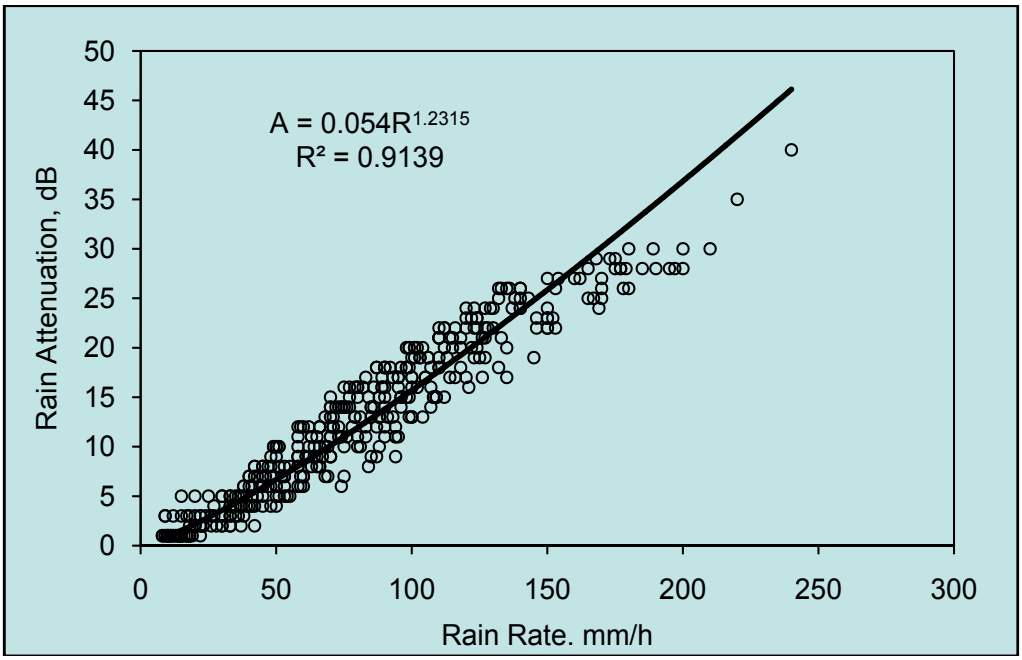


Fig. 3. The correlation between rain rate and rain attenuation

4.2 Statistical Confidence Level and Interval of the Measured Data

In statistics, a confidence interval (CI) is an interval estimate of a population parameter. Instead of estimating the parameter by a single value, an interval likely to include the parameter is given. Thus, confidence intervals are used to indicate the reliability of an estimate. How likely the interval is to contain the parameter is determined by the confidence level or confidence coefficient. The end points of the confidence interval are referred to as confidence limits. Increasing the desired confidence level will widen the confidence interval. For analysis purpose, a 95% confidence level was used because it is recommended by many statistical analysts. By using Minitab software, the confidence limits for 95%

confidence level were calculated. Fig 4, 5 and 6 show the confidence level of cumulative distribution of measured rain rate data in the year 2006, 2007 and 2008, respectively. In any statistical investigation, the variations of values in a measured data are due to random and systematic errors. The measured rain rate data for the year 2006, 2007 and 2008 lie in between the 95% confidence limits. The R-square is 99.7%, 99.2%, 97.8%, respectively. Therefore, these rain rate data are suitable to be used for comparison with models.

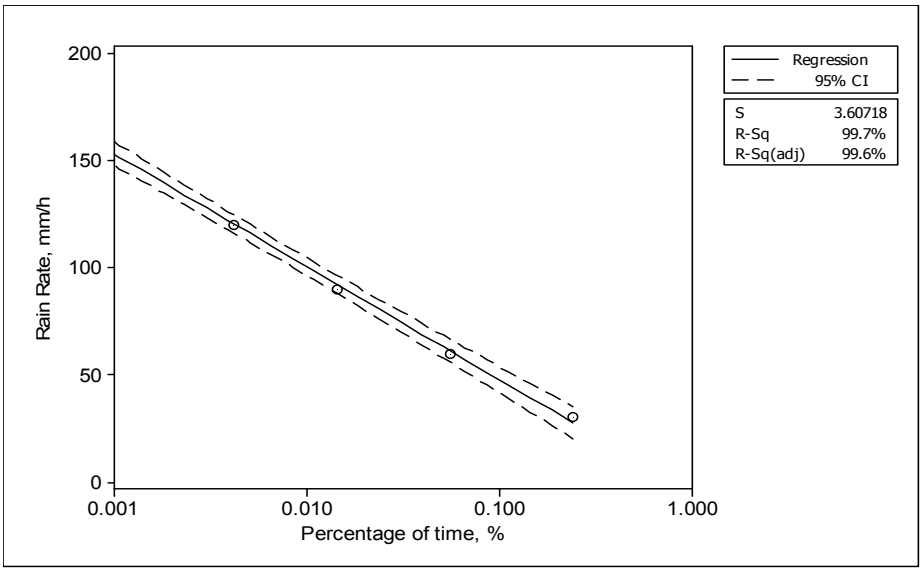


Fig. 4. The confidence level of cumulative distribution of measured rain rate data in year 2006

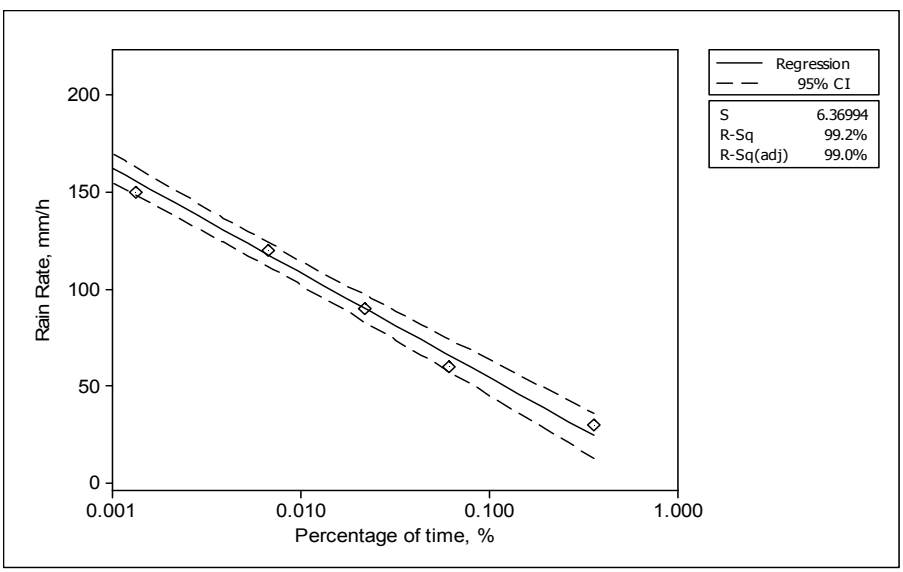


Fig. 5. The confidence level of cumulative distribution of measured rain rate data in year 2007

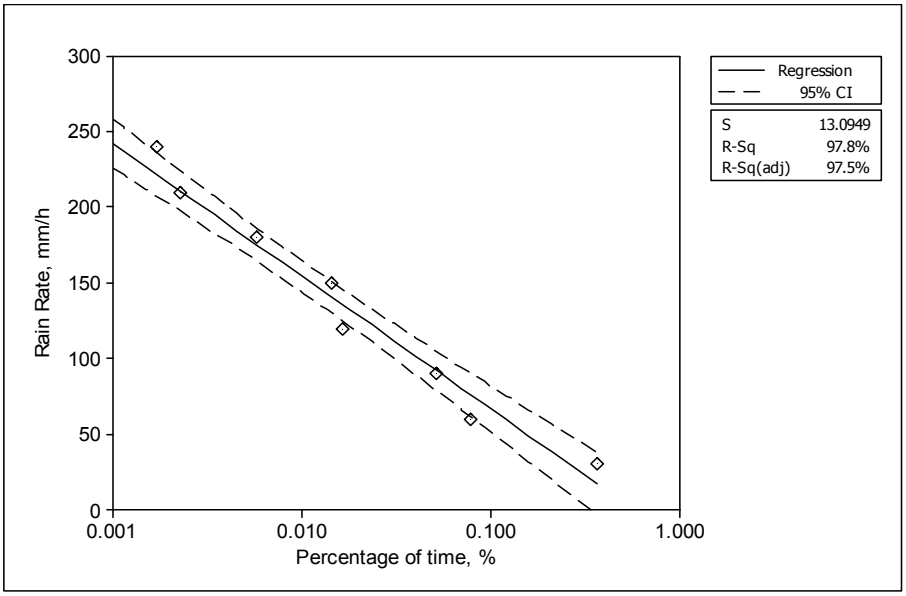


Fig. 6. The confidence level of cumulative distribution of measured rain rate data in year 2008

To calculate the confidence limit, having drawn a random sample and calculated the mean and standard deviation, the lower confidence limit (LL) is found by

$$LL = \bar{X} - z\left(\frac{\alpha}{2}\right) s_{\bar{x}} \tag{3}$$

In a similar way, the upper confidence limit (UL) is found by

$$UL = \bar{X} + z\left(\frac{\alpha}{2}\right) s_{\bar{x}} \tag{4}$$

where a is the significance level. At 95% confidence level, $z\left(\frac{\alpha}{2}\right)$ is equal to 1.96 and $s_{\bar{x}} = \frac{s}{\sqrt{N}}$, where s is the standard deviation and N is the sample size.
By using equation (3) and (4), the confidence limits were calculated for 95% confidence level. The confidence limits are calculated for the rain attenuation data in year 2006, 2007 and 2008 data are shown in Fig 7, 8 and 9. The measured rain attenuation data lies in between the 95% confidence limits for the entire measurement time. Hence, these rain attenuation data are suitable to be used for comparison with models.

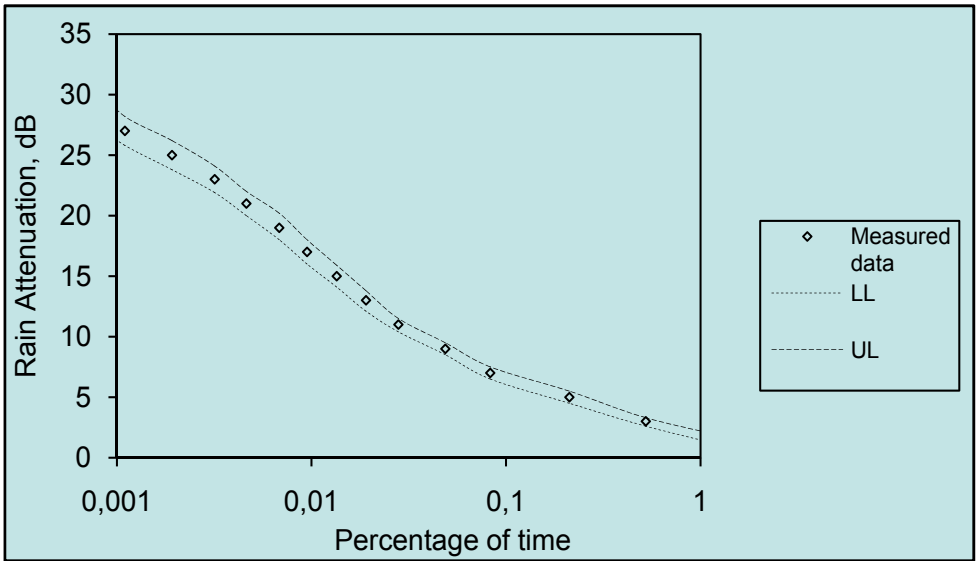


Fig. 7. The confidence level of cumulative distribution of measured rain attenuation data in year 2006

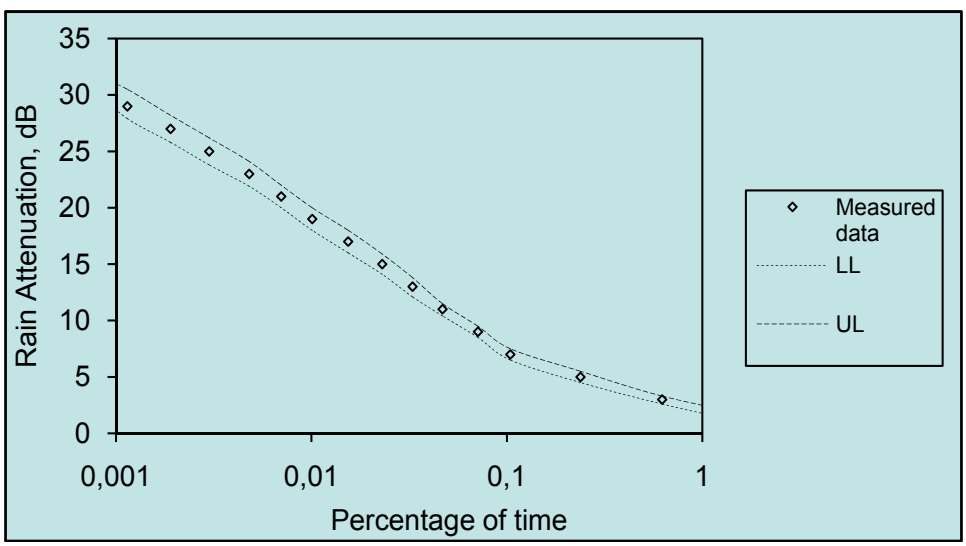


Fig. 8. The confidence level of cumulative distribution of measured rain attenuation data in year 2007

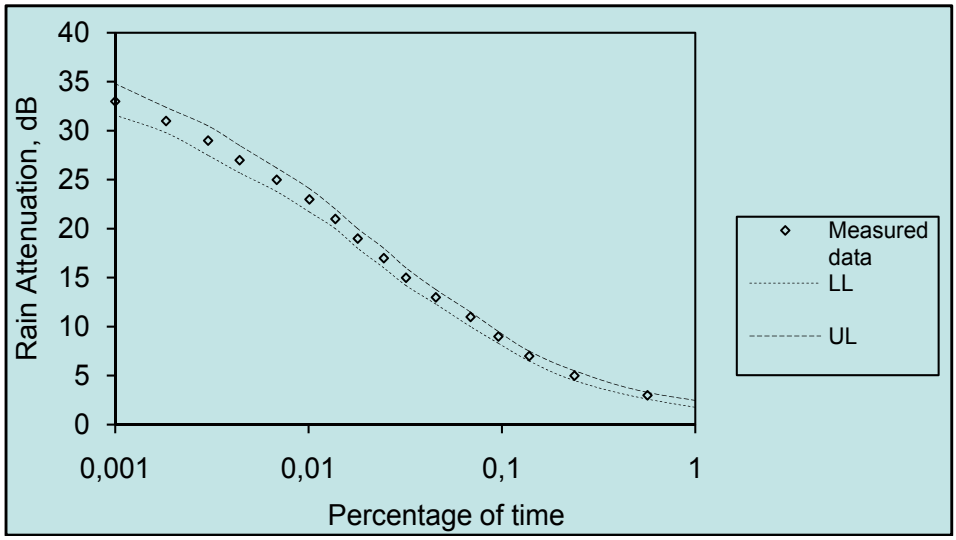


Fig. 9. The confidence level of cumulative distribution of measured rain attenuation data in year 2008

5. Rain Rate, Rain Attenuation Analysis

Communication systems operating at frequencies beyond 10 GHz in tropical and equatorial climates are subjected to many fade occurrences due to heavy rain. The information on rain rate and attenuation statistic is useful in link budget design. The considerable average worst month from year to year and within individual years is important in the planning of satellite earth-link design.

Rain rate is the volume of liquid water that falls through a unit area per unit time period. Rain attenuation is the depletion of electromagnetic energy during propagation through rain, caused by raindrop scattering and absorption. Worst month distribution of rain rate and attenuation is defined as the highest rain rate and attenuation, respectively within a set of individual monthly distributions of rain rate and attenuation for a year.

Best fit lines of the annual cumulative distribution of measured rain rate and rain attenuation for these three years have been used to make the comparison.

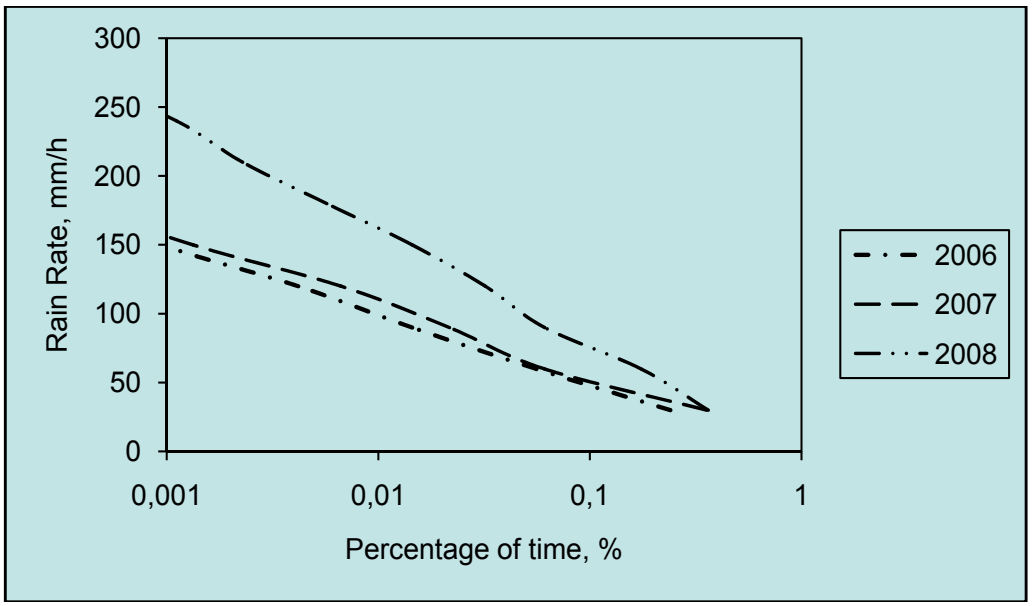


Fig. 10. Cumulative distribution of measured rain rate

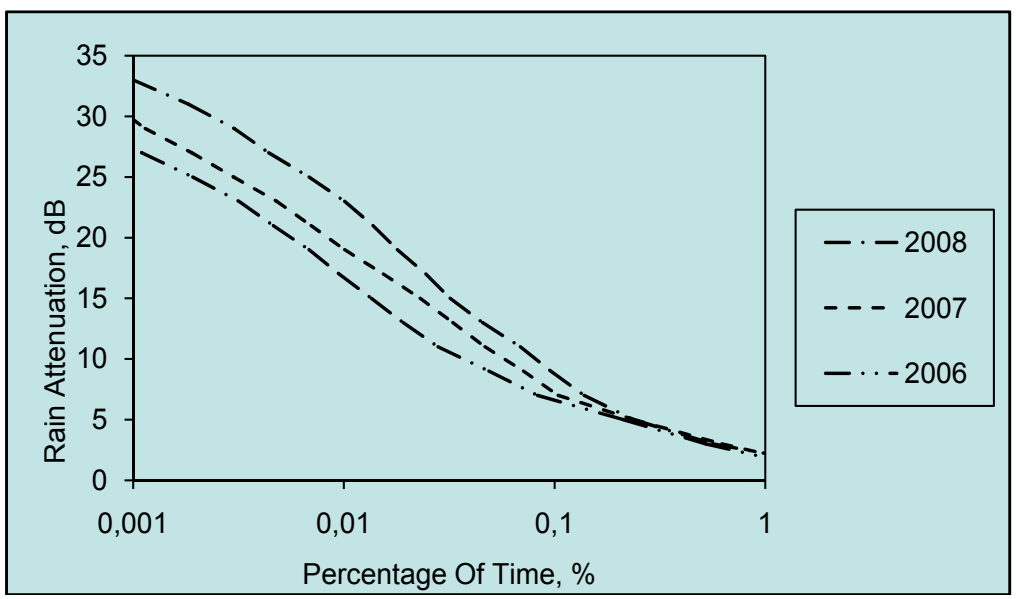


Fig. 11. Cumulative distribution of rain attenuation

Fig 10 and Fig 11 show the annual cumulative distribution of rain rate and rain attenuation for three years (2006-2008), respectively. Table 1 shows the record of rain rate and rain attenuation at point 0.01% of time and maximum point for the year 2006, 2007, 2008.

Year	Rain Rate. mm/h		Rain Attenuation, dB	
	At point 0.01%	Maximum	At point 0.01%	Maximum
2006	100	180	17	29
2007	110	210	18	31
2008	160	270	28	31

Table 1. The record of rain rate and rain attenuation at point 0.01% of time and maximum point for the year 2006, 2007, 2008

6. Worst Month Statistics

Worst month distribution of rain attenuation is defined as the highest rain attenuation within a set of individual monthly distributions of rain attenuation for one year. Rain attenuation data for year 2006, 2007 and 2008 at USM was analyzed.

In planning of the design of reliable communication systems, the required statistics of propagation effects is relevant to the worst month reference. On the other hand, the reference statistics for several propagation prediction methods are the long term average annual distribution. Therefore, the conversion of the yearly statistic to the worst month is important. The annual and worst-month rain attenuation exceedance curve is shown in Fig.12, Fig 14 and Fig 16 for year 2006, 2007 and 2008, respectively. The worst month for 2006, 2007 and 2008 were November, October and March, respectively. The Q factor as a function of annual percentage of exceedance for Ku-band rain attenuation is shown in Fig.13, Fig 15 and Fig 17 for year 2006, 2007 and 2008, respectively.

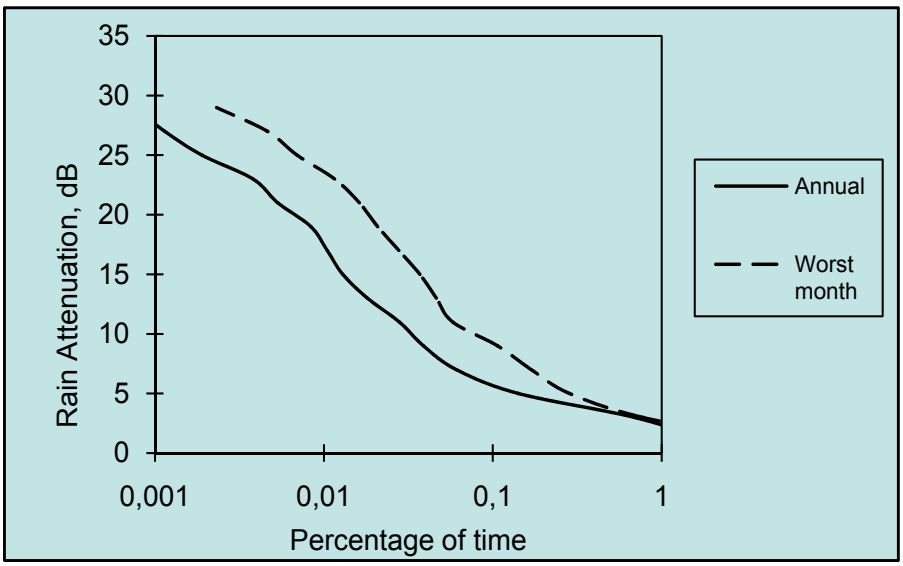


Fig. 12. The annual and worst-month rain attenuation exceedance curve at Ku-band for year 2006

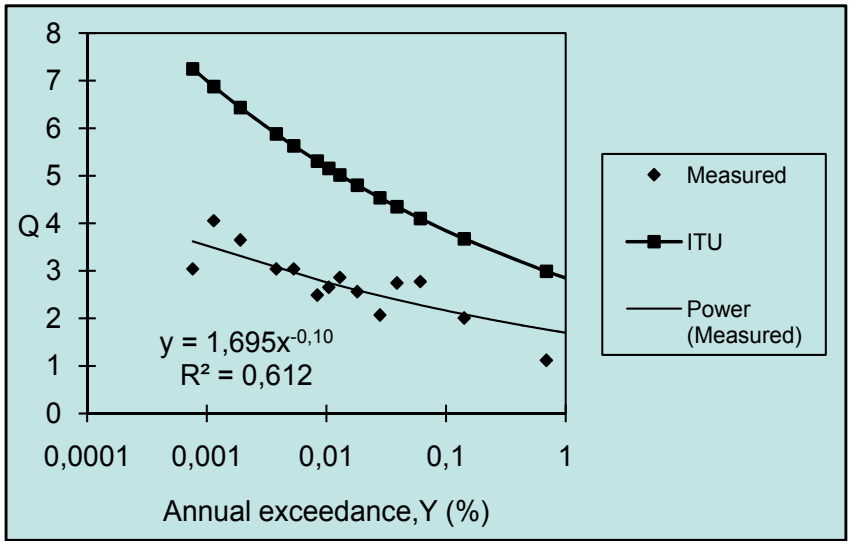


Fig. 13. The Q factor as a function of annual percentage of exceedance for rain attenuation at Ku-band for year 2006

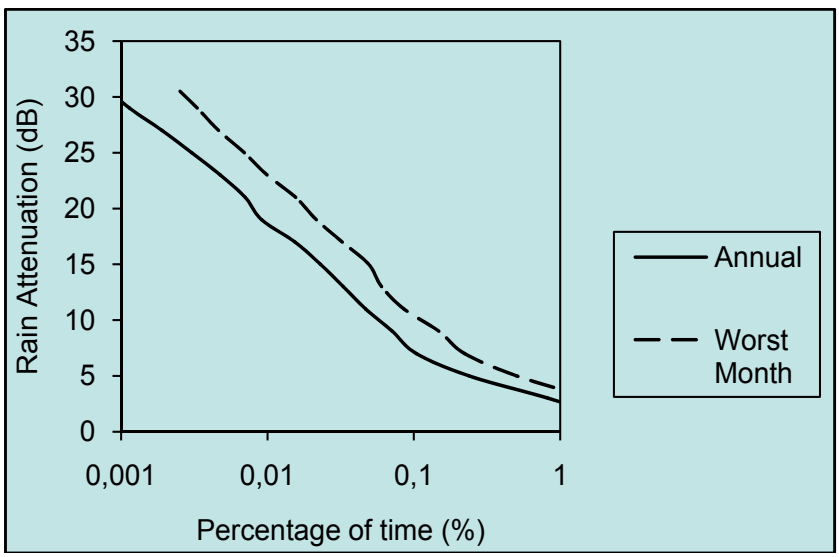


Fig. 14. The annual and worst-month rain attenuation exceedance curve at Ku-band for year 2007

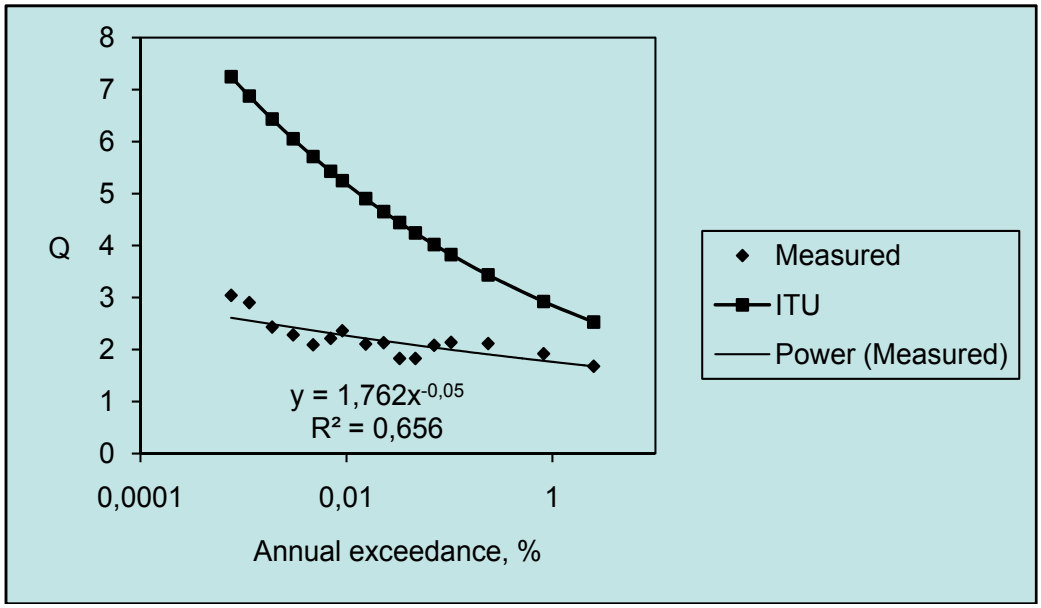


Fig. 15. The Q factor as a function of annual percentage of exceedance for rain attenuation at Ku-band for year 2007

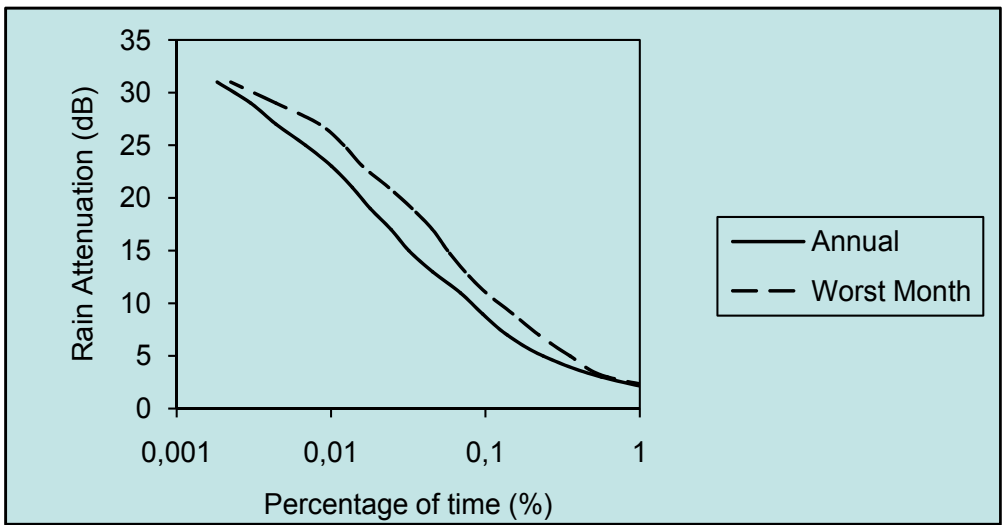


Fig. 16. The annual and worst-month rain attenuation exceedance curve at Ku-band for year 2008

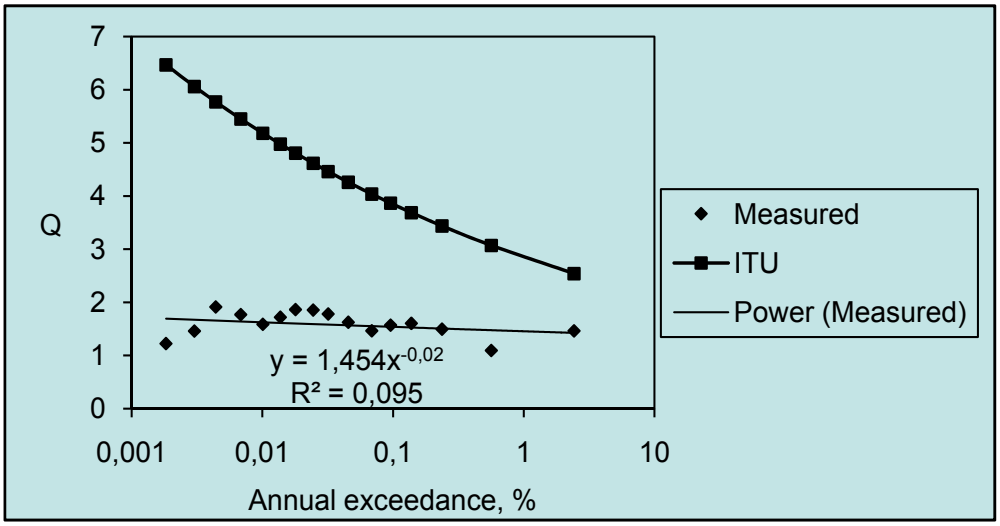


Fig. 17. The Q factor as a function of annual percentage of exceedance for rain attenuation at Ku-band for year 2008

The Q factor for rain attenuation was found to follow the power law of the form $Q=AY^{-\beta}$. Table 2 shows the parameter A and β for year 2006, 2007 and 2008.

Year	A (Proposed by ITU-R = 2.82)	B (Proposed by ITU-R = 0.15)
2006	1.6953	0.106
2007	1.7624	0.055
2008	1.4547	0.024

Table 2. The parameter A and β for year 2006, 2007 and 2008

For global rain rate applications, the ITU P.841-4 has recommended values of **A** = 2.82 and β = 0.15 for tropical, subtropical and temperate climate regions with frequent rain. The percentage of time for worst month distribution of rain attenuation is significantly higher than annual distribution of rain attenuation. There is a large difference between the A and β values obtained for tropical area with the ITU proposed. It shows that the ITU values are not suitable for use in worst-month analysis for tropical area. This indicates that the Q factors are climatic dependant. The worst month curve was strongly influenced by the Northeast Monsoon during which the highest levels of attenuation occurred. Worst month attenuation statistics are very important for the study of the performance of a communication system during periods of up to 31 days.

Ku-band TV services are affected by outages for time-critical transmission such as real-time news and sports broadcasting. Service providers need to consider the use of appropriate forward error correction codes, the choice of modulations, and the range of uplink power controls to use during severe rain fade periods in the overall design of their communications networks (Pan, *et al*, 2008). These techniques can be used to provide a significant improvement in both performance and availability for worst month attenuation.

7. Specific Attenuation Analysis

The measurement ran continuously from 1st January 2006 to 31st December 2008. Three years of cumulative distributions are obtained. The measurement ran continuously from 1st January 2006 to 31st December 2008. Three years of cumulative distributions are obtained. The relation between rain attenuation and rain rate in year 2006, 2007 and 2008 were shown in Fig. 18, 19 and 20. In these figures, the symbols × are obtained by directly plotting the rain attenuation against rain rate and the solid line is the best-fit curve obtained by fitting the rain attenuation against rain rate. The effective path length, L_{eff}, is a function of rain rate and has a direct correlation with the measured rain rate. The effective path length has been found by using equation 2.5 and 2.6 and shown in Fig. 21. The effective path length for the site is a power-fitting function of rain rate and can be simplified as

$$L_{eff}\,(\text{km}) = 13.367\, R^{-0.21} \tag{5}$$

The L_{eff} is obtained from the equation 5 is used to divide the total attenuation to obtain specific attenuation, γ at different rain rate. Fig. 22, 23 and 24 show the relation between specific attenuation, γ and rain rate in the year 2006, 2007 and 2008.

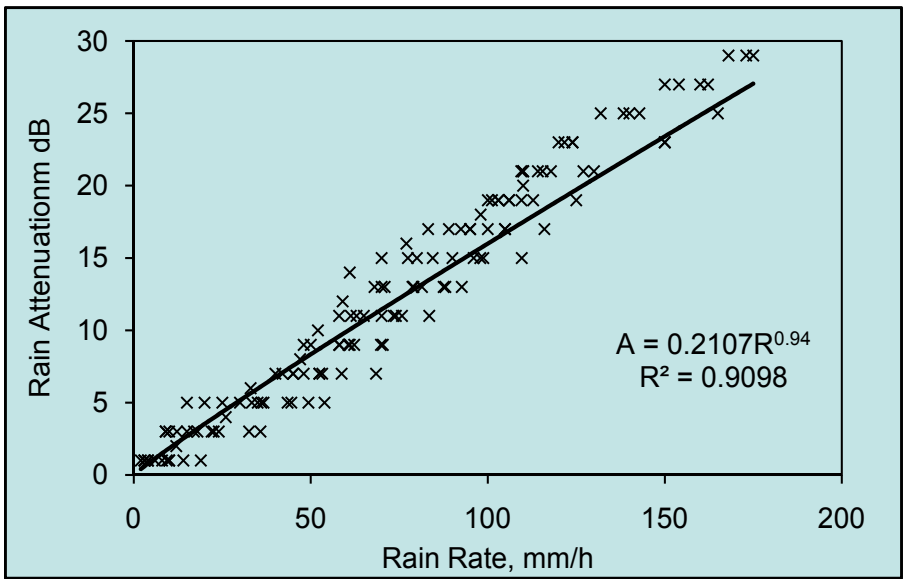


Fig. 18. The correlation between rain attenuation and rain rate in year 2006

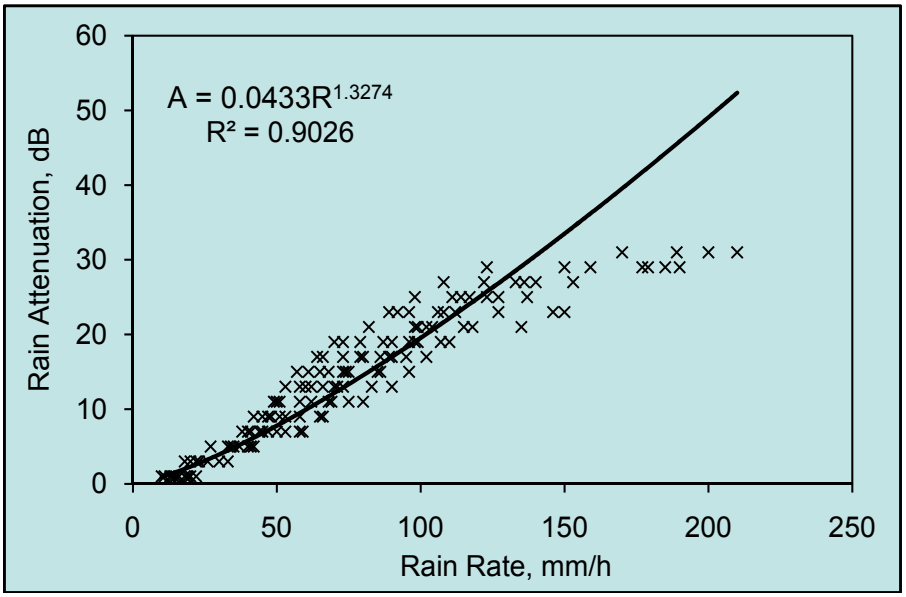


Fig. 19. The correlation between rain attenuation and rain rate in year 2007

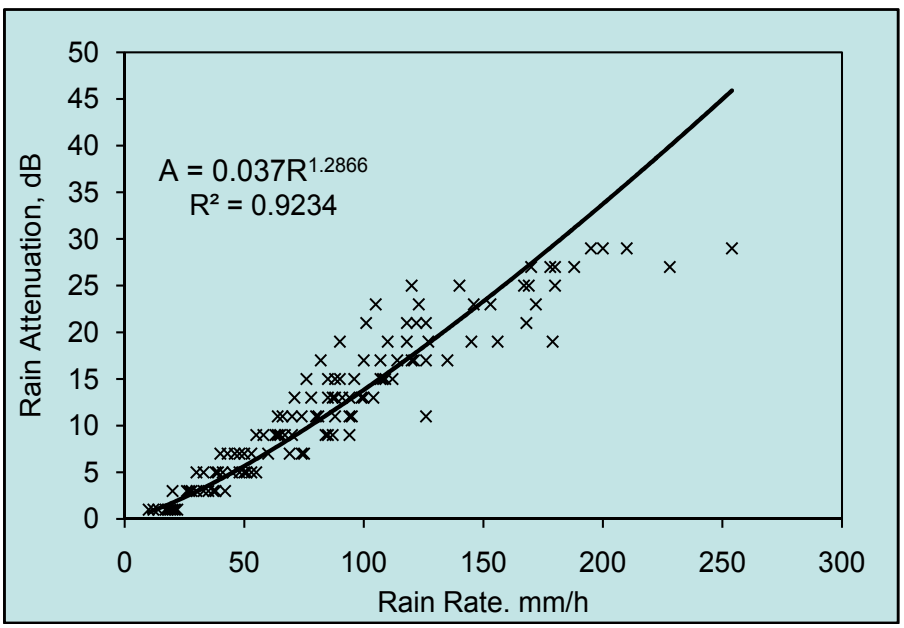


Fig. 20. The correlation between rain attenuation and rain rate in year 2008

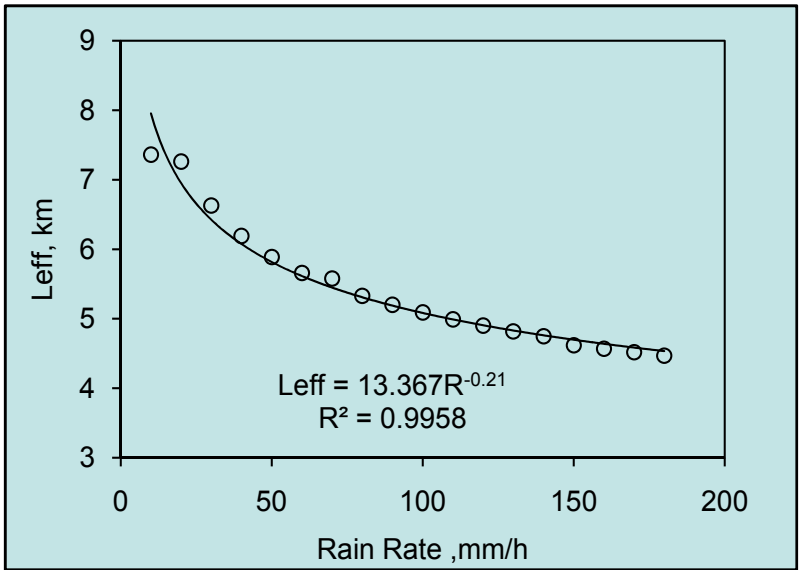


Fig. 21. The correlation between the effective path length and rain rate.

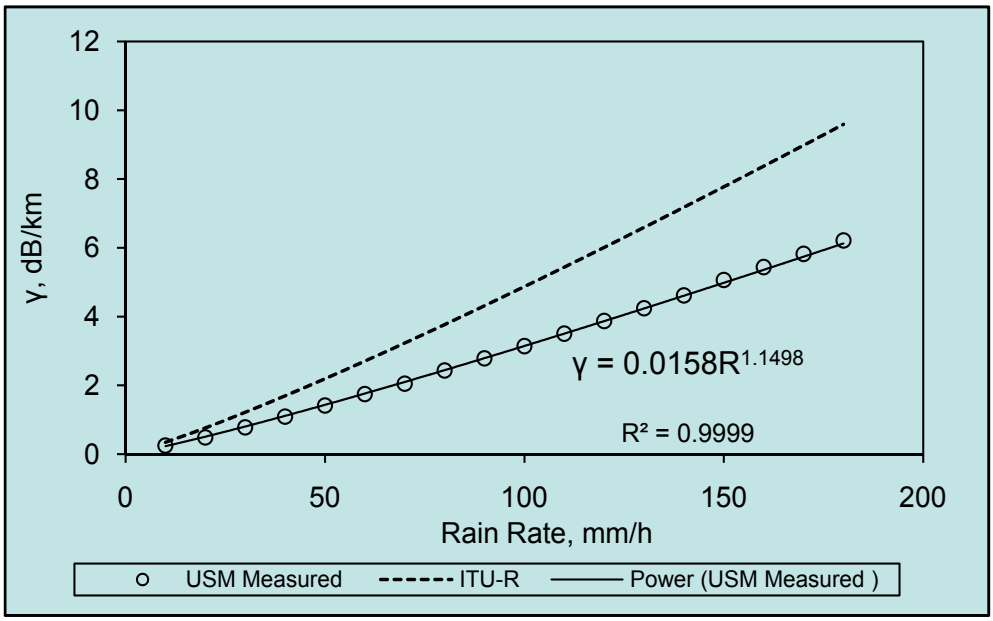


Fig. 22. Relationship between specific attenuation and rain rate compared with ITU-R in year 2006

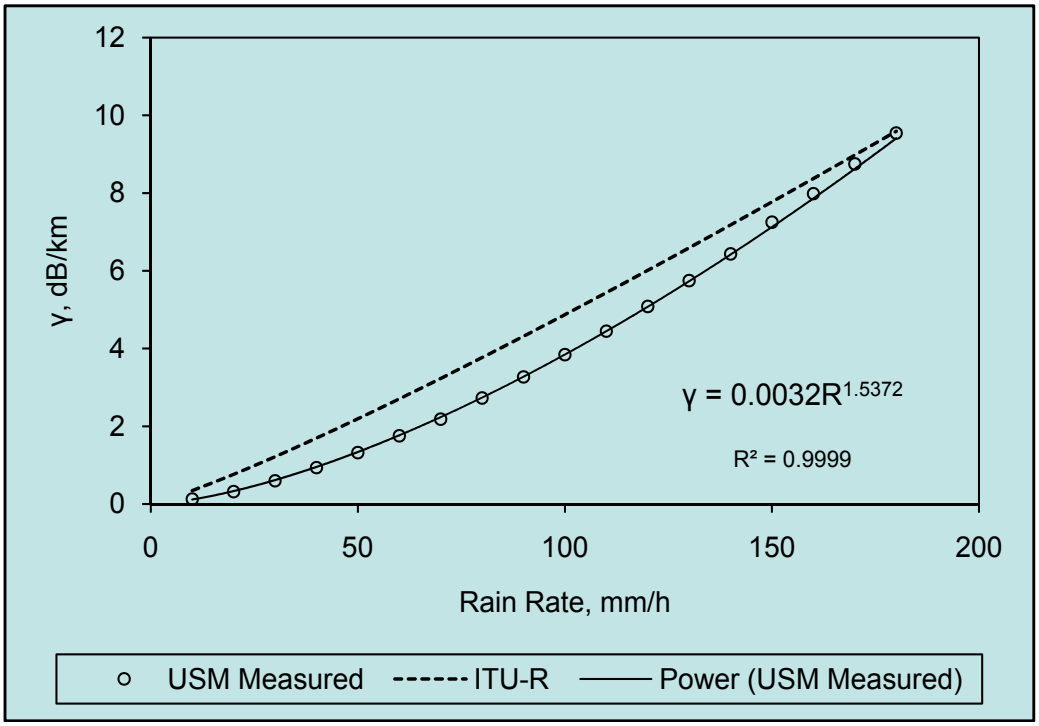


Fig. 23. Relationship between specific attenuation and rain rate compared with ITU-R in year 2007

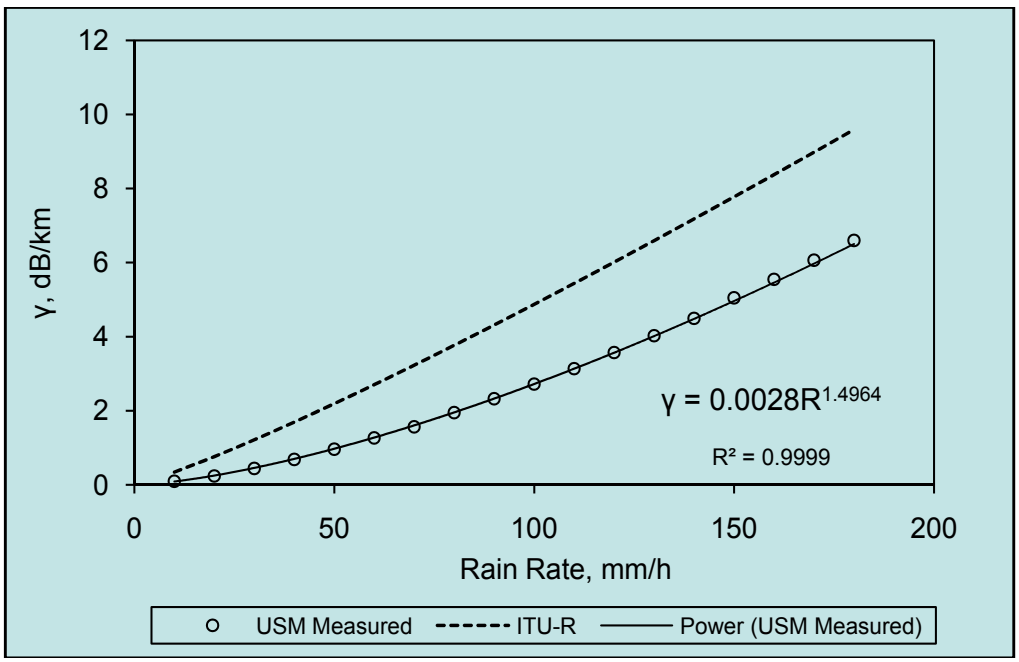


Fig. 24. Relationship between specific attenuation and rain rate compared with ITU-R in year 2008

Using equation 2.2 and 2.3, k and a that obtained are 0.0242 and 1.152 respectively. Table 3 shows the regression coefficients for k and α by using empirical procedure.

Year	k	α
2006	0.0158	1.1498
2007	0.0032	1.5372
2008	0.0028	1.4964

Table 3. Regression coefficients for k and α by using empirical procedure

Based on rain rate and rain attenuation measurements, the ITU-R has overestimated the specific rain attenuation due to tropical rainfall at least in the 3 years term view. The coefficients of k and α are found that can significantly vary and be considerably different from the ITU-R proposed for regression coefficients and it implies that the raindrop size distribution (DSD) in Malaysia's tropical region is quite different from that adopted by ITU-R, at least in our experiment period. There are many factors influencing the specific attenuation. This is considered due to the verity of the drop size from temperate regions to the tropical region. The availability and accuracy of measured data is the factor to influence the empirical value. Therefore, ITU-R recommendation for regression coefficients of rain specific attenuation is not suitable use in predicting rain attenuation for Malaysia.

8. Analysis of One-Minute Rain Rate Measured Data with Existing Models

The comparison of the measured one minute rain rate values with existing rain rate models is shown in this section. There are 5 tropical climates sites (e.g. USM, Bangkok, Bandung, Manila, Fiji) 2 years average (from the years 2002 to 2003) measured one-minute rain rate that used in comparison. The 2 years (from 1st January 2007 to 31st December 2008) average USM measured rain rate has been used in the comparison. The existing models that applied in the prediction one minute rain rate are Moupfouma model, ITU-R model, KIT simplified model, Rice & Holmberg model and Dutton & Dougherty model. The prediction rain rate depends on the annual rainfall values. The annual rainfall values for these tropical climates sites are shown below:

Sites	Annual rainfall (mm)
USM	2088.0
Bangkok	1565.0
Bandung	1956.0
Manila	2300.0
Fiji	3087.5

Table 4. The average annual rainfall

The comparison of one minute rain rate prediction models with measured data for the 6 tropical climates sites are shown in Fig 25, 26, 27, 28 and 29.

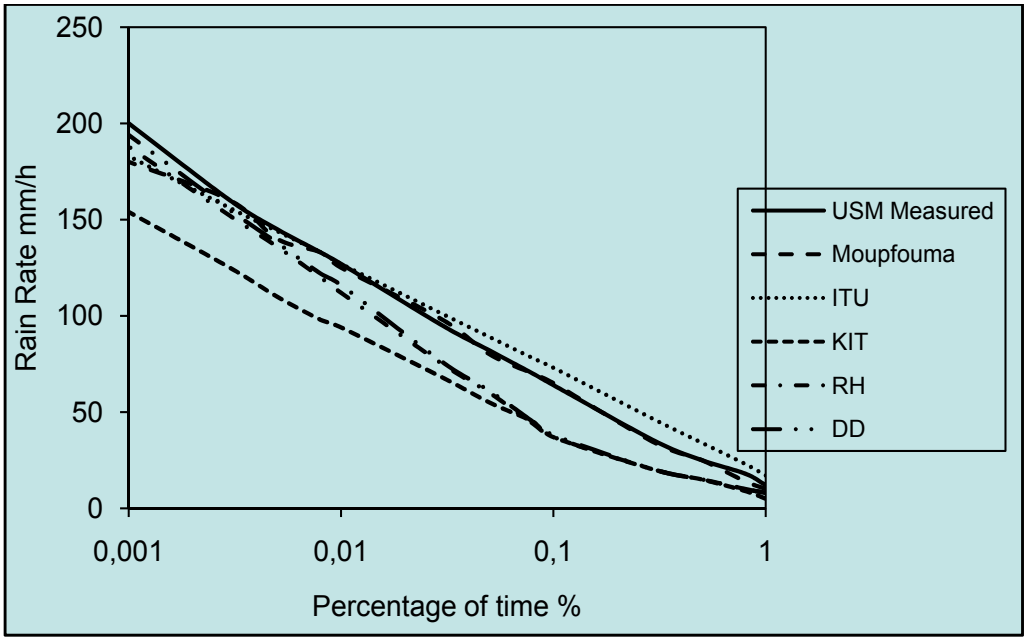


Fig. 25. Comparison of one minute rain rate prediction models with measured data for USM site.

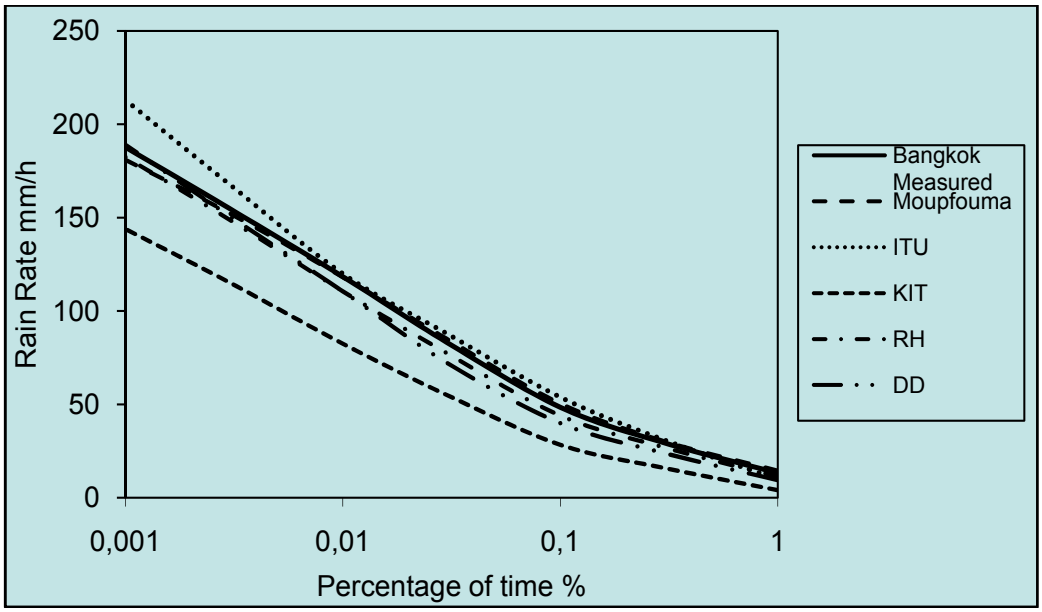


Fig. 26. Comparison of one minute rain rate prediction models with measured data for Bangkok site.

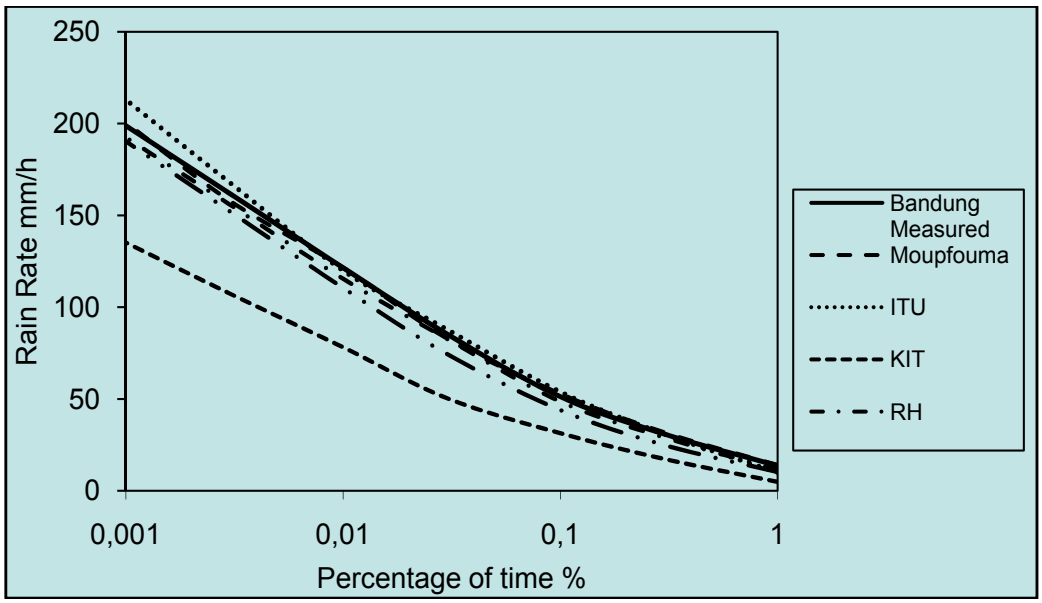


Fig. 27. Comparison of one minute rain rate prediction models with measured data for Bandung site.

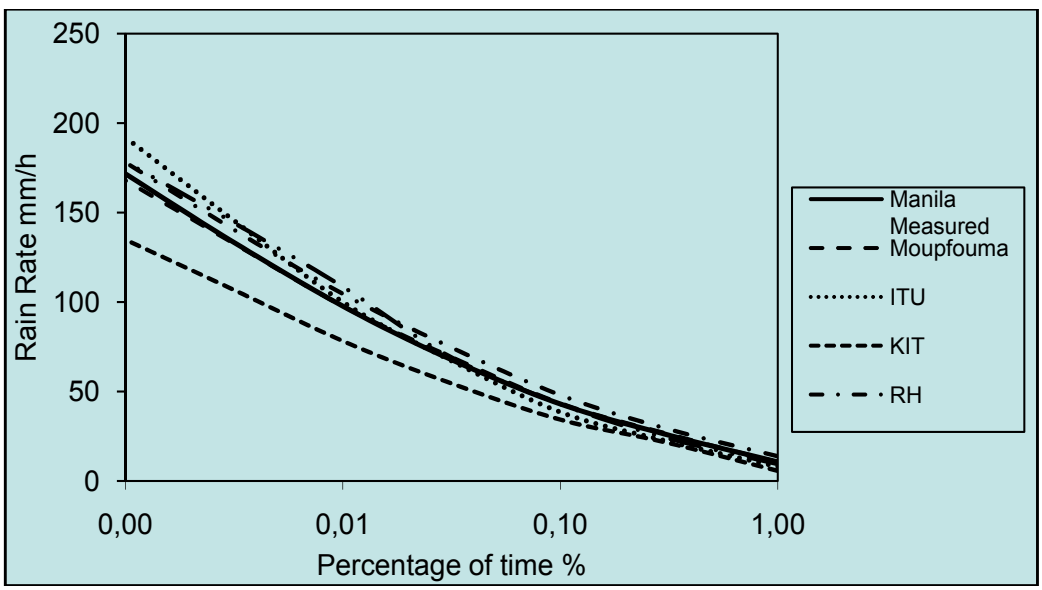


Fig. 28. Comparison of one minute rain rate prediction models with measured data for Manila site.

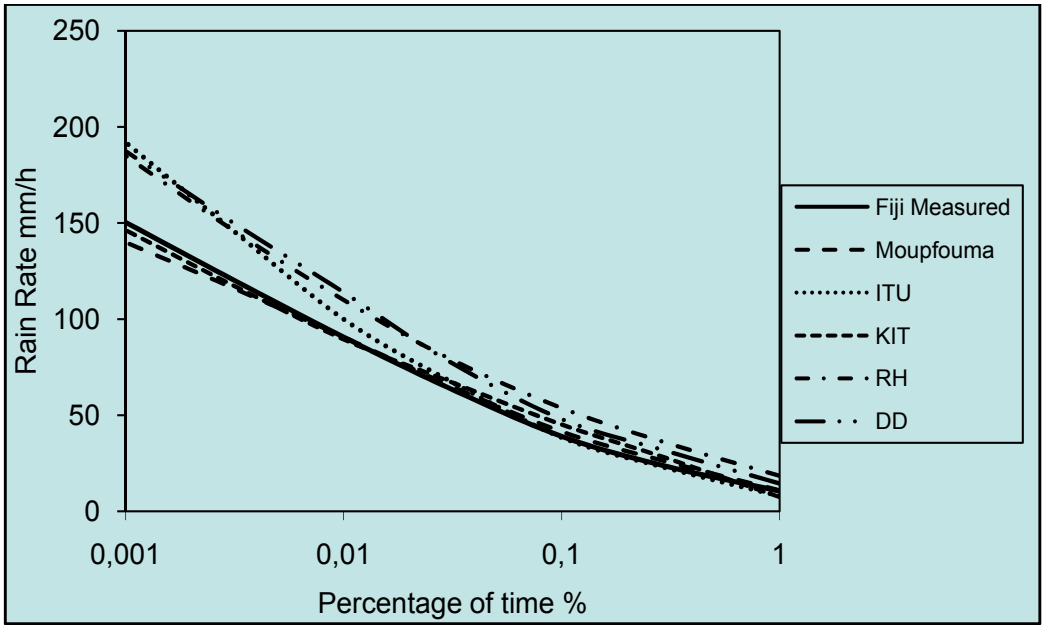


Fig. 29. Comparison of one minute rain rate prediction models with measured data for Fiji site.

The Moupfouma model overestimates the one minute rain rate from 0.01% to 1% of time and underestimates the rain rate from 0.001% to 0.01% at USM sites. The model gave a RMS value of 8.64% for USM. This is because the model has a probability law behavior that underlines the complexity of the rain rate distribution according to the climate of the zone of interest. For Bangkok, Bandung, Manila and Fiji sites, the model follows closely the measured rain rate values throughout the entire percentage of time that the rain rate is exceeded. The model gave a low RMS value for those tropical sites. The model's RMS values were 53% (Bangkok), 2.33% (Bandung), 1.69% (Manila) and 6.22% (Fiji). The coefficients of λ and Y values from the slope of rain rate curve in equation 2.18 depend strongly on the measured rain rate data. For tropical and sub-tropical localities, $\lambda = 1.066$ and $Y = 0.214$ are used in calculation of rain rate cumulative distribution slopes.

The ITU-R model overestimates the one minute rain rate from 0.01% to 1% of time and underestimates the rain rate from 0.001% to 0.01% at USM sites. The model gave a RMS value of 20.72% for USM. For Bangkok, Bandung, Manila and Fiji sites, the model follows closely the measured rain rate values up to 0.01% of time that rain rate is exceeded before the model overestimates the measured values. The model gave a RMS value of 13.18% for Bangkok site, 11.75% for Bandung, 13.65% for Manila and 17.90% for Fiji. The ITU-R has the climate zones used in the equatorial region are subdivided further that includes region with the similar rain rate characteristics and a large number of measured rain rate database that are available for equatorial region.

For USM, Bangkok, Bandung and Manila, the Kitami Institute of Technology (KIT simplified) model underestimates the measured rain rate throughout the entire percentage of time that the rain rate is exceeded. The model's RMS value was 36.56% (USM), 41.59% (Bangkok),

42.72% (Bandung) and 25.57% (Manila). The model gave a high RMS value for these sites because the annual rainfall amount at these sites were not more than 2300 mm. The KIT model prediction at Fiji, gave a low RMS value of 15.62%. The model follows closely the measured rain rate values at the entire percentage of time that the rain rate is exceeded. This is because the annual rainfall at Fiji was above 3000 mm. The KIT model states that the accuracy of the model depends largely on the annual rainfall values, where the higher the annual rainfall values better the prediction gets.

The RH model underestimates the measured rain rate at USM, Bangkok and Bandung throughout the entire percentage of time that the rain rate is exceeded. The model gave a RMS value of 29.65% at USM, 8.59% at Bangkok and 7.58% at Bandung. The RH model overestimates the measured rain rate at Manila and Fiji throughout the entire percentage of time that the rain rate is exceeded. The model gave a RMS value of 149% at Manila and 42.16% at Fiji. The RH considered the convective rain activity and stratiform rain activity was neglected. The thunderstorm ratio, β was based on thunderstorm rain but on the convective rain activity days to total rain days. The model gave a high RMS value at Fiji site because the β value given by RH is 0.3, however the β value calculated to be 0.75.

The Dutton and Dougherty (DD) model underestimates the measured rain rate at USM, Bangkok and Bandung and overestimates the measured rain rate at Manila and Fiji throughout the entire percentage of time that the rain rate is exceeded. The model gave a RMS value of 29.04% at USM, 16.03% at Bangkok, 186% at Bandung, 7.73% at Manila and 28.10% at Fiji. The M (average annual total rainfall depth, ,mm) values used to calculate the coefficient constant in Europe were below 1200mm per year, but the annual rainfall, M is above 1800mm per year in tropical climate.

A summary of the info is as show in Table 5 and a conclusion of the best and worst model is given. The comparison of rain rate was done between measured data and five pre-existing mathematical models. For the tropical region, it was found that Moupfouma model revealed a close fit to the measured data for low, medium and high rain rates. The Moupfouma model is judged suitable for use in predicting rates in tropical climates. The KIT simplified model exhibited poor performance in comparison.

Site	Annual rainfall, mm	RMS value, %				Conclusion		
		Moupfouma	ITU-R	KIT	RH	DD	Best Model	Worst Model
USM	2088.00	8.64	20.72	36.56	29.65	29.04	Moupfouma	KIT
Bangkok	1565.00	35	13.18	41.59	8.59	16.03	Moupfouma	KIT
Bandung	1956.00	2.33	11.75	42.72	7.58	186	Moupfouma	KIT
Manila	2300.00	1.69	13.65	25.59	149	7.73	Moupfouma	KIT
Fiji	3087.50	6.22	17.90	15.62	42.16	28.10	Moupfouma	RH

Table 5. The summary of the comparison of rain rate prediction model

9. Analysis of Rain Attenuation Measured Data with Existing Models

The rain attenuation prediction models exposed in literature calculate the attenuation related to a given rain rate or else to a given percentage of time. For terrestrial as well as satellite microwave links, one of the fundamental needs for the link designer is to have at his disposal an effective model that predicts attenuation caused by rain on the propagation path with a good accuracy. Most of the models available are empirical or semi empirical and their accuracy are based on the accuracy of the measured rain rate cumulative distribution (Moupfouma, 2009).

The comparison of the measured rain attenuation values with existing rain attenuation models is shown in this section. There are 5 tropical climates sites that are USM, Bangkok, Bandung, Manila, Fiji measured rain attenuation that is used for the comparison. The 2 years (from 1st January 2007 until 31st December 2008) average USM measured rain attenuation has been used in the comparison. The existing models that applied in the prediction rain attenuation are ITU-R model, Ong model, Ramachandran and Kumar model, CETUC model, Leitao and Watson model, Garcia-Lopez model, SAM model and Assis-Einloft model. The comparison measured rain attenuation with existing predicted models at these 8 tropical climates sites are shown in Fig 30, 31, 32, 33 and 34.

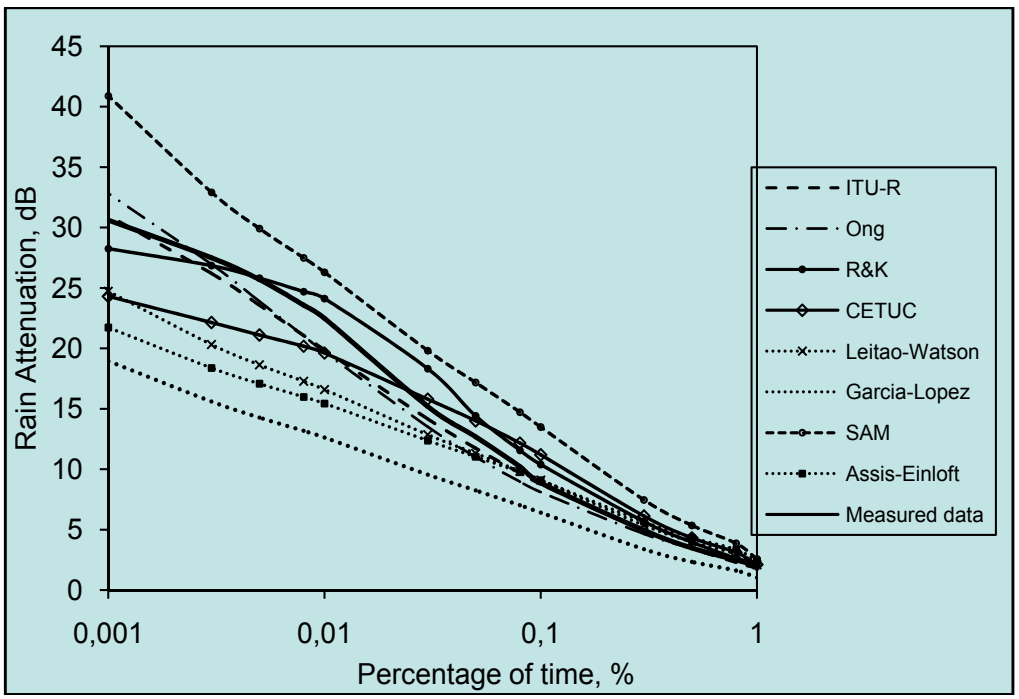


Fig. 30. The comparison measured rain attenuation with existing predicted models at USM

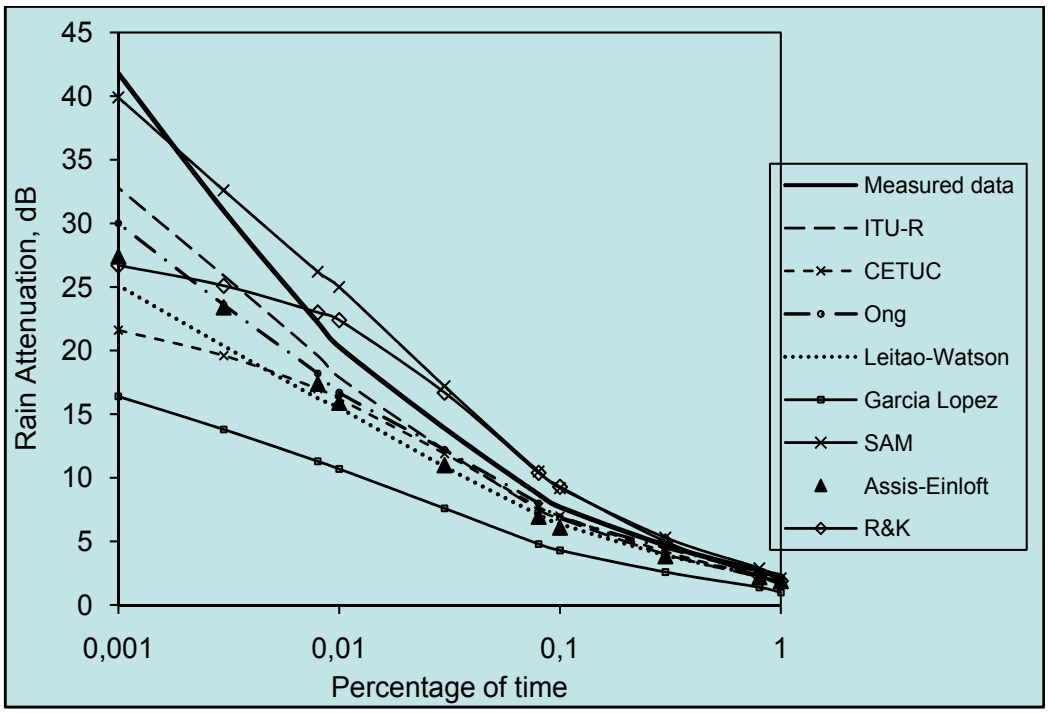


Fig. 31. The comparison measured rain attenuation with existing predicted models at Bangkok

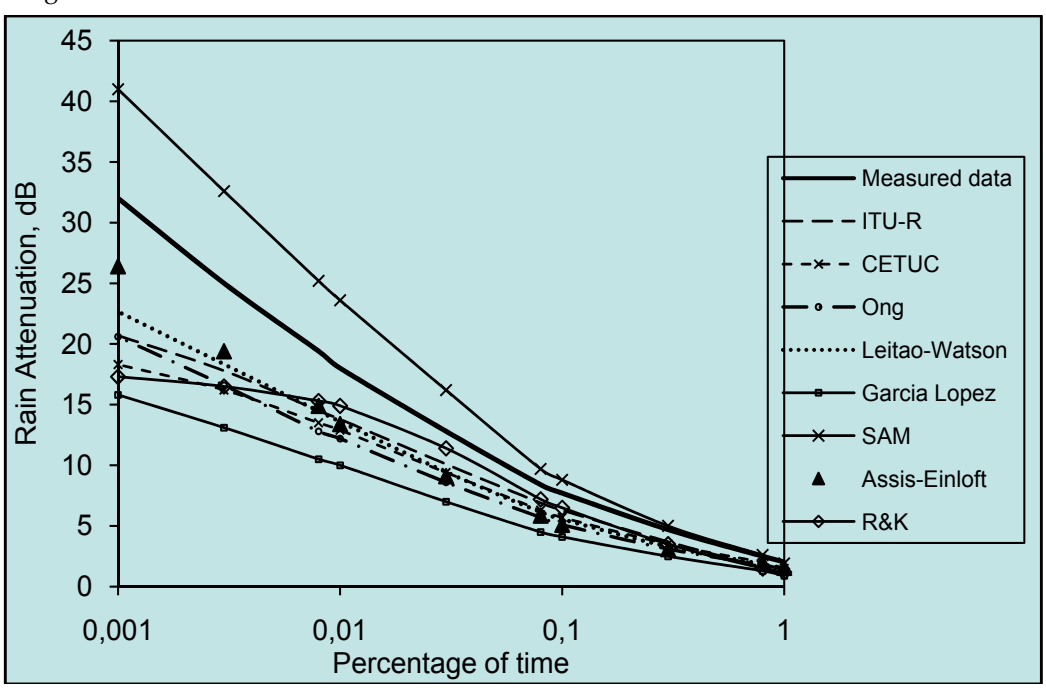


Fig. 32. The comparison measured rain attenuation with existing predicted models at Bandung

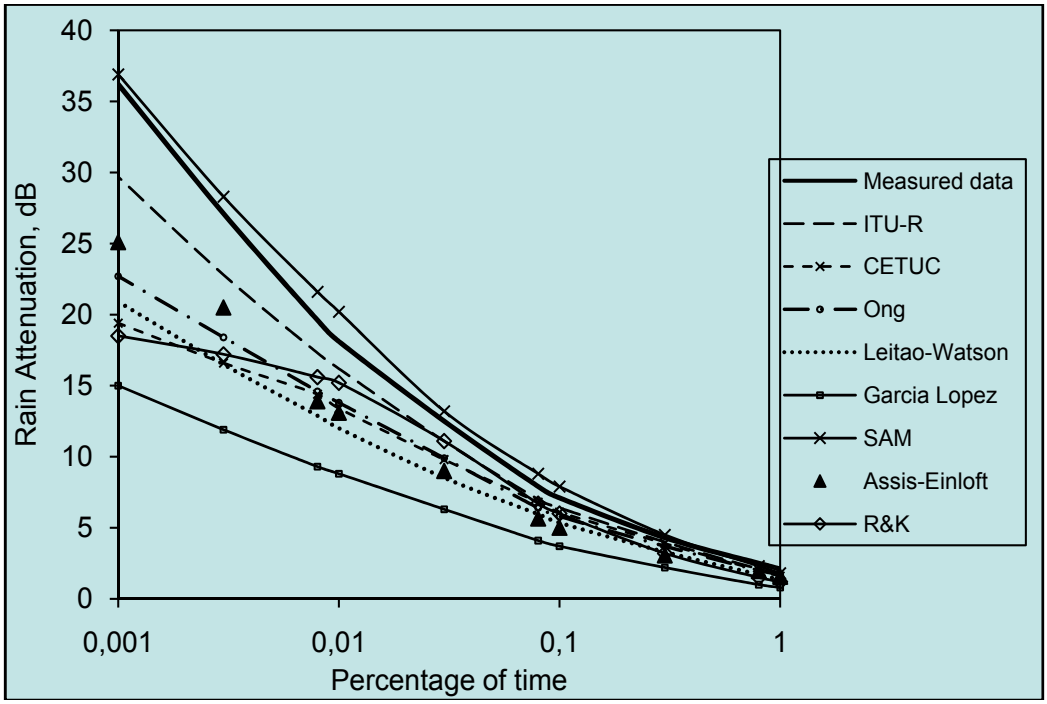


Fig. 33. The comparison measured rain attenuation with existing predicted models at Manila

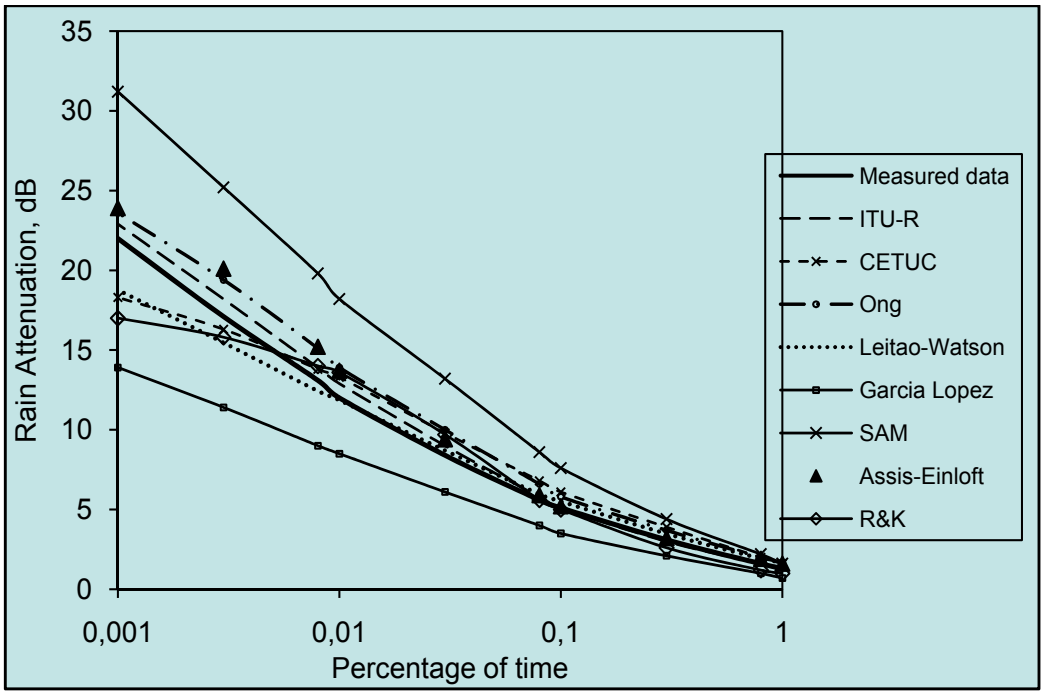


Fig. 34. The comparison measured rain attenuation with existing predicted models at Fiji

The ITU-R model underestimates all of these 5 tropical climates, except Fiji throughout the entire percentage of time. The model follows closely with the USM measured data from 0.05% - 1% of time. ITU-R underestimates the rain attenuation at the lower percentage of time because of the roll over effect, where as the rain rate increases, the attenuation reduce. This is because of the lack of high rain rate data from tropical climates. The rain column height is constant and maximum (10 km) when the rain reaches its saturation point, but the rain-cell diameter continues to decrease with increasing rain rate. Hence, the proportional increase of rain volume, which is a combination of rain-cell diameter, rain column height and rain rate would cause saturation (Ramachandran and Kumar, 2004). The vertical path reduction coefficient was used to minimize the prediction error. At Bangkok, Manila and Fiji, the ITU-R models gave a lower RMS value. At Fiji, the ITU-R model follows closely the measured rain attenuation throughout the entire percentage of time. The model gave a low RMS value because the rain rate of 90.7 mm/h was used for calculating the rain attenuation at 0.01% of time. This model was developed based on low rain rate of 85 mm/h at 0.01% of time from temperate climates. At Bangkok and Manila, the model gave a high RMS value because of the high rain rate vales at 0.01% of time have been used in calculating the rain attenuation. At Bandung, the model gave a high RMS values. At Bandung, the high elevation angle of above 60^0 was applied in experimental. The station height above sea level that used was 700m, whereas this model was developed by station heights above mean sea level from 20m to 400m.
The Ong model at USM underestimates rain attenuation at the entire measurement time. The model gave a percentage error of ±14% with a range RMS value of 9.62% at USM. This model was revised from ITU-R model. The model has a roll over effect at lower percentage of time, because it was developed for 4/6 GHz. At Fiji, the Ong model follows closely the measured rain attenuation for the entire measurement time. The station height above sea level at both of these sites is below 60m. The station height above sea level that was used to develop this model was below 60m. At Bangkok, the model agrees reasonably well with the measured rain attenuation down to an outage time of 0.03% and deviates considerably from the measured values from 0.03% to 0.001%. At lower percentage of time above 0.01%, the model relative error increases because of the model was developed for 4/6 GHz. When the higher operational frequency gets, the higher rain attenuation will be at lower percentage of time. The Ong model at Bandung and Manila give poor performance for the entire measurement time. The model gave a high RMS value at these sites. It is because the station heights above sea level are above 80m and the elevation angle of the measurement site was above 55°.
At USM, Bangkok and Fiji, Ramachandran and Kumar model (R&K) follows closely with the USM rain attenuation measured from 0.03% to 1% of time. However, the model underestimates the rain attenuation from 0.03% to 0.001% of time. For this model takes into account the effect of the breakpoint to predict the attenuation exceedance in the tropics. In the tropics when the rain rate increase and approach the breakpoint the rain structure gradually changes from stratiform to convective. If the breakpoint is reached at a lower rain rate, then the rain tends to saturate fast (Ramachandran and Kumar, 2004). Because of this reason, the model has a roll over effect at lower percentage of time. The model gave a RMS value of 19.91% at Bangkok and 16.41% at Fiji. For the $0.003 \leq p \leq 1$, the rain attenuation increase gradually with increasing rain rate. Beyond 0.003% of time, the rain attenuation tended to saturation finally leading to total outage. At Bandung and Manila, the model is rejected for prediction for the entire measurement time. The model gave high RMS values for these sites. This is because of the rain rate (R_{AB}) at the breakpoint is above 70mm/h at these measurement sites. The rain rate of 58mm/h at the breakpoint was used to determine the model coefficient.

The CETUC model is simple to apply and uses the full rain rate distribution to predict the attenuation distribution, avoiding extrapolations functions dependent on the percentage of time. The model keeps the concept of an equivalent rain cell. The attenuation dependence on frequency is completely described by the parameters k and α (ITU-R recommendation parameters that used in calculating specific attenuation). At Fiji, the CETUC model agrees reasonably well with the measured values from 0.008% -1% of time and deviates considerably from the measured values from 0.001% to 0.008% of outage time. It gave a percentage error of ±25% with a RMS value of 18.96%. The model gave a low RMS value because the elevation angle was 45.5^0 and the station height was 13m. The model was developed based on an average elevation angle 42^0 and the altitudes below 50m. At USM, the model gave a high RMS value of 19.34% at USM, because the parameters k and α that recommended are not suitable used in USM. The highest rain rate and rain attenuation values were 200 mm/h and 30 dB, respectively used to apply the model. At Bandung, Manila and Bangkok, the CETUC model deviated considerably the measured rain attenuation for the entire measurement time. At Bandung and Bangkok sites, the model gave a high percentage error with a RMS value because the rain height calculated by CETUC was 3.18 km. The height above sea level for Bandung station was 700m. However, the rain height used to develop the model was based on limited number of stations with height above sea level below 50m. At Manila, the model gave a high percentage error of ±38.7% with a RMS value of 26.6%. This is because the effective length of the rain cell was developed by the rain rate values from 12 mm/h at 1% to 150 mm/h at 0.001% of time and the rain height calculated by CETUC was 3.18 km.

At Fiji, the Leitao-Watson model appears to work well down to the entire measurement time. The model gave a lower RMS values. The model parameters *s, t, u, v* and *w* are suitable to be used at Fiji site. Besides Fiji, the Leitao-Watson model underestimates the rain attenuation for the entire measurement time at the other sites. The model gave high RMS value of 24% and above. The model developed according to radar observation of rainfall structure, proposed the same set of equations with different parameters for widespread and convective rain (discrimated by a rain rate threshold of 20 mm/h) (Capsoni, *et. al.*,2009). The model was developed by using Europe rainfall data. It will make the model cannot perform well and give a high RMS value in predicting rain attenuation in tropical countries.

The Garcia-Lopez gave a high RMS value of above 30% and above for all these measurement sites. The model is underestimates the rain attenuation values for the entire outage time of an average year at all these measurement sites. This is because the rain height of 4km was given by Garcia-Lopez. The rain heights in tropical countries are above 5km, which are given by the ITU-R map of rain height above mean sea level. The coefficient constant of the model was obtained based on low rain rate of 60 mm/h at 0.01% of time. The range of rain rate at 0.01% of time for tropical climates is from 100mm/h to 130mm/h of time depending on the geographical area. The station height used to determine coefficient constant was averaged to 200m above mean sea level.

The SAM model at USM and Fiji site overestimates and shows poor performance in predicting rain attenuation. The model gave a high RMS value. This is because the model of the parameter controlling the rate of decay of the horizontal profile (γ). The model would give a lower RMS value below 10% if γ parameter was optimized against the set of data obtained from the measurement sites (Stutzman & Yon, 1986). At Bangkok and Bandung, the model follows closely the measured rain attenuation from 0.08%-1% of time and overestimates the rain attenuation from 0.08% to 0.001%. At low percentage of times, the

large errors are due to the fact that the predicted rain rates are less accurate in regions of high occurrence levels. The median values of the observations were estimated for each probability level in order to develop the model because long data sets were not available and the pooling of data from a number of locations was necessary to reduce the estimation error. The model appears to work well for the entire measurement time at Manila. The model gave a low RMS value of 9.8% because the parameter γ value given by SAM was optimized against the measured data sets.

At Fiji, Assis-Einloft model agrees reasonably well with the measured values for the entire measurement time. This is because the development of the reduction factor for this model was based on the measurement done at temperate and tropical climates, whereby at tropical climates 80 data sets at antenna elevation angles from 40^0 to 50^0 were used in order to reduce the prediction error at high rainfall intensity regions. Assis-Einloft model is not show the good agreement for the entire measurement time at Bangkok, Bandung and Manila. This is because the antenna elevation angles that used at these measurement sites were above 50^0. The path length that was considered by Assis was from 6km to 20.7km, but the path lengths at these measurement sites were below 6km. At USM, the model gave a high RMS value. A uniform rain rate was assumed for developing the model by introducing the concept of path length reduction factor. The path length at USM sites was below 6km.

The summary of the best model and worst model at the comparison sites was done and shown in Table 6. In this section, the comparison of rain attenuation was done for the measured data. For the tropical climate, it was found that no model revealed a close fit to the measured data for low, medium and high rain rates. The models do not predict rain attenuation at the lower percentage of time. The noticeable difference between the measured and the predicted variation is the existence of the breakpoint. The exceedance curves show that as the rain rate increases, the trend of the slope of the curve gradually decreases from large negative value, and then the trend that changes is referred to as the breakpoint in the exceedance curve (Ramachandran, *et. al.*, 2004, Mandeep, *et. al.*, 2008). The breakpoint exceedance curve usually occurs at high rain rate. When the rain structure is stratiform, the rainfall is widespread with low rain rates.

Site	RMS value, %								Conclusion	
	ITU-R	Ong	R&K	CETUC	Leitao	Garcia	SAM	Assis	Best Model	Worst Model
USM	7.11	9.62	11.50	19.34	18.75	38.58	37.96	25.80	ITU-R	Garcia
Bangkok	17.36	15.82	19.91	221	251	49.98	16.45	22.35	Ong	Garcia
Bandung	28.78	32.79	29.43	29.29	27.35	47.86	22.36	26.70	SAM	Garcia
Manila	13.02	202	30.45	26.60	32.73	53.75	9.57	27.63	SAM	Garcia
Fiji	5.57	18.91	16.41	18.96	137	33.85	46.49	100	ITU-R	SAM

Table 6. The summary of the comparison of rain attenuation prediction models

The ITU-R model is judged suitable for use in predicting rain attenuation in these measurement tropical climates sites. The Garcia-Lopez model exhibited poor performance in comparison. The results are particularly important for the tropical and equatorial region because not much of research that has been done in these regions.

Acknowledgment
The authors are grateful thanks to Electrical and Electronic School of USM Engineering Campus for technical support and like to acknowledge Ministry of Science, Technology and Innovation (MOSTI) 01-01-02-SF0670 for financial support. The authors would like to express sincere thanks to the reviewers for their comments.

14

Guidelines for Satellite Tracking

Dusan Vuckovic, Petar Rajkovic and Dragan Jankovic
Faculty of Electronic Engineering, University of Nis
Serbia

1. Looking at the sky

Whether you are willing to confess or not, the most useful instrument for observing the sky, by day or by night, is the naked eye. For the ones with less experience, it is important to become familiar with the general aspects of the sky. Wide field of view that human eye is naturally equipped with is ideal for this purpose. Binoculars and telescope are also indispensable tools, but they may be introduced, once the familiarity with the sky is established.

Since celestial bodies have a tendency of moving, it is very hard to predict their future position. It is the big mathematical engine that will help us pinpoint the object in the sky, so we would be able to look there, or point our telescope. Many objects in the sky look very similar. In NASA's[1] 1997 report it's being said that there are 25,000 man-made objects in space. 8681 are currently orbiting Earth, and 16,000 objects are in a state of decay. Course, situation is quite different today, considering that's been 13 years since that report was published.

Large portion of 16,000 objects are orbital debris: parts, such as nosecone shrouds, lens, hatch covers, rocket bodies, payloads that have disintegrated or exploded, and even objects that "escape" from manned spacecraft during operations. Most of these objects are small, but quite a few are large enough to be seen with unaided eye.

Additional difficulties come with the fact that many military or government satellites are painted in black, thus very hard to spot even with very sophisticated optical instruments.

Three things become immediately obvious when you take a look at the night sky. Not all the stars look the same, also stars twinkle, and finally, it's easy to arrange stars into recognizable patterns. Some stars are bright, some are less bright and some almost too dim to see with naked eye. Astronomers have had to devise some means of expression the variation precisely.

Ancient Greek astronomer Hipparchos was the first to produce the star catalogue, introducing the new value of importance, or magnitude, the star had. He thought that important stars should be of higher importance and termed them magnitude 1. Those somewhat dimmer he considered of second importance (magnitude 2), and so on. Altogether, he created six classes of star's magnitudes. At that time, he was not aware of the fact that his six divisions were based on the way the human eye recognizes a brightness

[1] NASA – US National Aeronautics and Space Administration

difference, where one object seems half as bright as another does. It was later discovered that Hipparchos' six magnitudes give a difference between magnitude 1 and 6 of 100 times. Although this system has been refined, most scientists use six degrees magnitude system to explain the stars, or other celestial body's brightness in the sky.
Predictions are necessary if you want to observe satellites or other objects in the sky. You have to know where to look and when to look. Time is very important, so a stopwatch with an accuracy of at least 0.5 seconds is essential. Set the watch by radio time signals, or by telephone system's speaking clock. More innovative method would include mobile phone with stopwatch that synchronizes the clock through network operator time, or over the network, with NTP[2] protocol.
Next step in simplified satellite tracking would include gauging the positions of a satellite by imaging a line drawn in the middle of two fairly close stars between which the satellite passes, or by imaging a vertical line drawn down from a particular star. The precise moment at which the satellite crosses this line is the most important one. With the help of a Star Atlas, satellite track can be indicated. Identification of a position is not too difficult, but the problem is accurate satellite's transit time. Reliable results require constant repetition and practice.
Artificial satellites may be divided in three categories, depending on their brightness: Bright, visible to the naked eye (magnitude < 3), dimmer (3 < magnitude < 9) and extremely faint ones. Satellites dimmer than magnitude 6 require at least binoculars to be able to spot them.
More sophisticated satellite tracking method would involve specialized computer software, and mathematical apparatus. Today, time synchronization, which is the requirement for better preciseness is simplified in a great manner with the introduction of NTP boxes, primarily used in network performance monitoring. With a set of relatively simple mathematical procedures and the selection of adequate, so called, "propagation model", we'll be able to spot, follow or wait for a satellite to pass above us in the evening sky.
Necessary steps needed to achieve this goal will be explained in detail in the following chapters.

2. Propagation Models

To be able to distinct satellite orbits we need to describe them. There are seven, and sometimes, eight numbers to tell us the orbit of each satellite. Those are orbital elements, or sometimes "Keplerian" [3] elements. The ellipse oriented about the Earth with the satellite on a specific position in time is described with these numbers. Keplearian model introduced orbits with constant shape and orientation, with the Earth at one focus of the ellipse, not the centre, unless the orbit itself is circular.
To increase the exactness of the tracking model, another value is introduced. This value represents the corrections of Keplerian model known as perturbations. Corrections are introduced due to lumpiness of the Earth's gravitational field, and the "drag" on the satellite due to atmosphere. That way, Drag becomes an optional eight orbital element.
Basic orbital elements are:

---

[2] Network Time Protocol, protocol for synchronizing the clocks of computer systems over packet-switched, variable-latency data networks.
[3] Johann Kepler [1571-1630] was a German mathematician, astronomer and astrologer, and key figure in the 17th century scientific revolution.

1. **Epoch**, or "Epoch Time" (T0) is a number that specifies the time at which the snapshot of other orbital elements was taken.

2. **Orbital inclination** (I0) indicates the angle between the equator and the orbit when looking from the centre of the Earth. It ranges from 0 to 180 degrees. (Fig. 1)

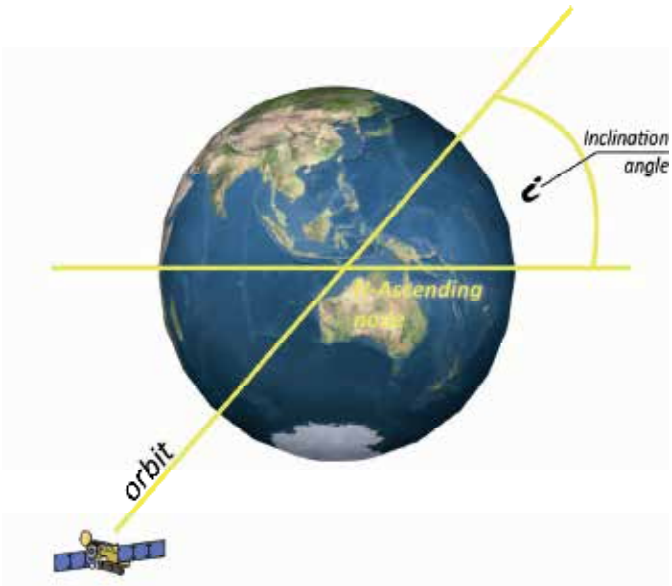


Fig. 1. Inclination angle

3. **Right Ascension of Ascending Node** (RAAN or O0) is the second number that orients the orbital plane in space. The first one is Inclination. Ascending node is the place where the satellite crosses the equator while going from the Southern Hemisphere to the Northern Hemisphere. Due to Earth's rotation, fixed object in space is necessary to measure Ascension. (Fig. 2.)
Solution is to use Aries, which has the same location as vernal equinox, so the angle between the Aries and Ascending node is called Right Ascension of Ascending Node.

4. **The Eccentricity** defines the flatness of the orbit. If the orbit is a perfect circle, then eccentricity is 0, and 1 when very flat.

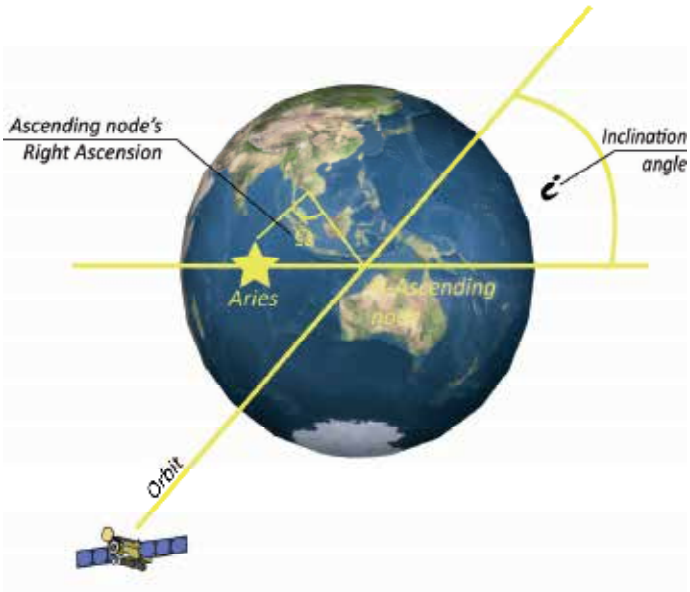


Fig 2. Ascending node's Right Ascension

5. Due to elliptical shape of the orbit, satellite will sometimes be closer and sometimes further from the Earth. The point where the satellite is closest to the Earth is called Perigee, and the point where it's furthest from the Earth is called Apogee. (Fig 3.)

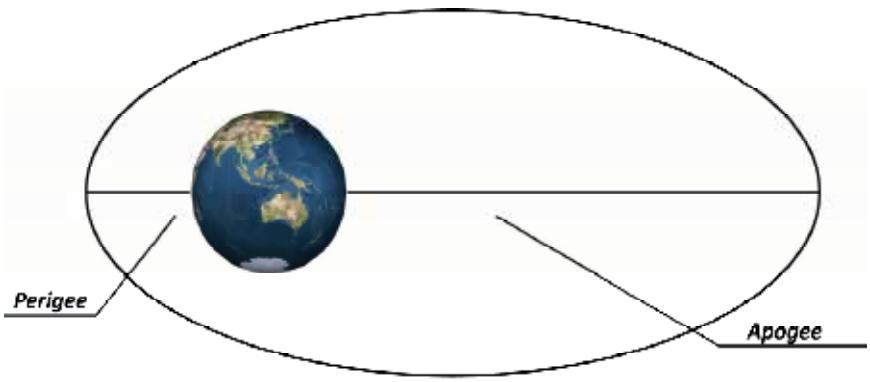


Fig. 3. Perigee and Apogee

The angle formed between the perigee and the ascending node is called **the Argument of Perigee**. The argument of perigee would have the value 0 if the perigee would occur at the ascending node. (Fig. 4.)

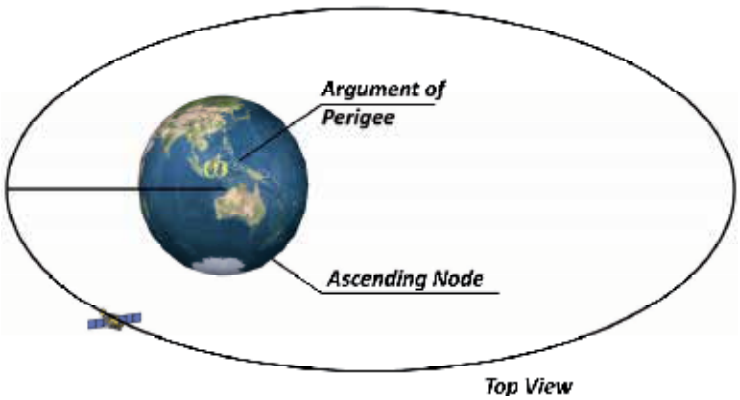


Fig. 4. Argument of Perigee

6. **The Mean Motion** is the value that illustrates how fast the satellite is going. According to Kepler's Law:

$$v^2=GM/r$$

v = velocity of the satellite (m/s)
M = Earth's mass ($5.98*10^{24}$)
G = gravitational constant ($6.672*10^{-11}$ Nm^2/kg^2)
r = distance between the satellite and centre of the Earth (m)

The closer the satellite gets to the Earth, more speed it gets. This value will help us obtain the satellite's altitude.

7. **The Mean Anomaly** ranges from 0 to 360 degrees, it's referenced to the perigee and represents satellite's position on orbital path. In the perigee itself, mean anomaly would be 0.

8. **The Drag** is final, optional value that encapsulates atmospheric drag as well as gravitational pull from stellar bodies such as Sun or the moon. Knowledge of the atmospheric density at the satellite's position is required to calculate the drag force on the satellite. It may also be required in terms of establishing a reliable linkto the satellite. Nevertheless, this may require more information than density alone. (King-Hele, 1983.)
Just for density calculations two models should be mentioned. The Harris-Priester model and the Jacchia model (Bindebrink et al., 1989.)

If only gravitational force is assumed by Newton's law is acting on the satellite the first five parameters are constant and the orbit is an ideal Keplerian orbit (Bunnell, P. 1981.)

Considering the tracking of Earth orbiting objects, institution that has the leading role in providing data related to all satellites in the orbit is NORAD[4]. NORAD periodically releases

[4] NORAD – North American Aerospace Defense Command

element sets that, in conjunction with certain models, can be used to track satellites and all other objects. These element sets are periodically refined so as to maintain a reasonable prediction capability.
It is very important to emphasize that not just any prediction model will do the job. The NORAD element sets are "mean" values obtained by removing periodic variations in a particular way. In order to obtain good predictions, periodic variations must be reconstructed. Reconstruction is done by prediction model in the exactly the same way they were removed by NORAD. To gain maximum prediction accuracy NORAD element sets must be used with one of the following models:

2.1 SGP/SDP propagation models

SGP propagation model, developed by Hilton and Kuhlman in 1966 is used for near-Earth satellites. This model simplifies drag effect on mean motion by taking it as linear in time.
SGP4 model was developed by Ken Cranford in 1970 and is also used for near-Earth satellites. Afterwards SGP4 model was extended with deep-space equations by Hujsak in 1979 so it incorporated gravitational effects of the moon and Sun, sectoral and tesseral Earth harmonics important for calculation of half-day and one-day period orbits. This model was named SDP4 model as for deep-space model.
Hoots created SGP8 model in 1980 for near-Earth satellites as a simplification of his own theory that uses same gravitational and atmospheric models as Lane and Cranford but with differential equations integrated in a different manner. As a follow up SDP8 model was formed for deep-space satellites.
As previously mentioned the first orbit model was developed in 1966 and used to provide orbital elements in the form of two-line elements. They contain the Keplerian elements (Montenbruck et al., 2000.) inclination i, eccentricity e, right ascension Ω, argument of perigee ω and Mean anomaly M_0 at epoch time t_0. Instead of the semi-major axis a, anomalistic mean motion n_0 is used. Modified model SGP4 is used since 1970 for the production of two-line element sets for low-Earth satellites by NORAD. It accounts Earth gravitational field through zonal terms J_2, J_3, J_4 and atmospheric drag (assuming a non-rotating spherical atmosphere) through a power density function.
The first step in SGP4 calculation is constants calculation:

$$a_1 = \left(\frac{k_e}{n_o} \right)^{\frac{2}{3}} \tag{1}$$

whereas $k_e = \sqrt{GM}$, and G is Newton's universal gravitational constant.
M = the mass of the Earth
n_o = the SGP type "mean" motion at epoch

$$\delta_1 = \frac{3}{2} \frac{k_2}{{a_1}^2} \frac{\left(3\cos^2 i_o - 1\right)}{\left(1 - {e_o}^2\right)^{\frac{3}{2}}} \tag{2}$$

e_o is the "mean" eccentricity at epoch

i_o is the "mean" inclination at epoch

$$k_2 = \frac{1}{2} J_2 {a_E}^2 \tag{3}$$

J_2 is , as previously mentioned, the second gravitational zonal harmonic of the Earth, and a_E is the equatorial radius of the Earth.

$$a_o = a_1 \left(1 - \frac{1}{3}\delta_1 - {\delta_1}^2 - \frac{134}{81}{\delta_1}^3 \right) \tag{4}$$

$$\delta_o = \frac{3}{2}\frac{k_2}{{a_o}^2}\frac{\left(3\cos^2 i_o - 1\right)}{\left(1 - {e_o}^2\right)^{\frac{3}{2}}} \tag{5}$$

The original mean motion n_o'' and semimajor axis a_o'' are first recovered from the input elements by the equations

$$n_o'' = \frac{n_o}{1 + \delta_o} \tag{6}$$

$$a_o'' = \frac{a_o}{1 - \delta_o} \tag{7}$$

If the perigee stands between 98 kilometers and 156 kilometers, the value of the constant s used in SGP4 model will change to

$$s^* = a_o''\left(1 - e_o\right) - s + a_E \tag{8}$$

s is parameter for the SGP4/SGP8 density function

For perigee below 98 kilometers, s becomes

$$s^* = 20/\mathrm{XKMPER} + a_E \tag{9}$$

XKMPER = 6378.135kilometers/Earth radii.

If the value of s is changed, then the value of $\left(q_o - s\right)^4$ must be replaced by

$$\left(q_o - s^*\right)^4 = \left[\left[\left(q_o - s\right)^4\right]^{\frac{1}{4}} + s - s^*\right]^4 \tag{10}$$

Whereas q_o is parameter for the SGP4/SGP8 density function.

The next step is calculation of the constants with the use of appropriate values of s and $\left(q_o - s\right)^4$.

$$\theta = \cos i_o \tag{11}$$

$$\xi = \frac{1}{a_o'' - s} \tag{12}$$

$$\beta_o = \left(1 - {e_o}^2\right)^{\frac{1}{2}} \tag{13}$$

$$\eta = a_o'' e_o \xi \tag{14}$$

$$C_2 = \left(q_o - s\right)^4 \xi^4 n_o'' \left(1-\eta^2\right)^{-\frac{7}{2}} \left[a_o'' \left(1 + \frac{3}{2}\eta^2 + 4e_0\eta + e_o\eta^3\right) + \frac{3}{2}\frac{k_2\xi}{\left(1-\eta^2\right)}\left(-\frac{1}{2} + \frac{3}{2}\theta^2\right)\left(8 + 24\eta^2 + 3\eta^4\right)\right] \tag{15}$$

$$C_1 = B^* C_2 \tag{16}$$

B^* is the SGP4 type drag coefficient

$$C_3 = \frac{\left(q_o - s\right)^4 \xi^5 A_{3,0} n_o'' a_E \sin i_o}{k_2 e_o} \tag{17}$$

$$A_{3,0} = -J_3 {a_E}^3, \tag{18}$$

Whereas J_3 is the third gravitational zonal harmonic of the Earth.

$$C_4 = 2n_o''\left(q_o - s\right)^4 \xi^4 a_o'' {\beta_o}^2 \left(1-\eta^2\right)^{-\frac{7}{2}} \left(\left[2\eta\left(1 + e_o\eta\right) + \frac{1}{2}e_o + \frac{1}{2}\eta^3 \right] - \frac{2k_2\xi}{a_o''\left(1-\eta^2\right)} \times \right.$$
$$\left. \left[3\left(1 - 3\theta^2\right)\left(1 + \frac{3}{2}\eta^2 - 2e_o\eta - \frac{1}{2}e_o\eta^3\right) + \frac{3}{4}\left(1-\theta^2\right)\left(2\eta^2 - e_o\eta - e_o\eta^3\right)\cos 2\omega_o \right] \right) \tag{19}$$

ω_o is the "mean" argument of perigee at epoch.

$$C_5 = 2\left(q_o - s\right)^4 \xi^4 a_o'' {\beta_o}^2 \left(1-\eta^2\right)^{-\frac{7}{2}} \left[1 + \frac{11}{4}\eta\left(\eta + e_o\right) + e_o\eta^3 \right] \tag{20}$$

$$D_2 = 4a_o''\xi {C_1}^2 \tag{21}$$

$$D_3 = \frac{4}{3}a_o''\xi^2\left(17a_o'' + s\right){C_1}^3 \tag{22}$$

$$D_4 = \frac{2}{3}a_o''\xi^3\left(221a_o'' + 31s\right){C_1}^4 \tag{23}$$

The secular effects of atmospheric drag and gravitation are included through the equations

$$M_{DF} = M_o + \left[1 + \frac{3k_2\left(-1 + 3\theta^2\right)}{2a_o''^2 {\beta_o}^3} + \frac{3{k_2}^2\left(13 - 78\theta^2 + 137\theta^4\right)}{16a_o''^4 {\beta_o}^7} \right] n_o''\left(t - t_o\right) \tag{24}$$

Where M_o is the "mean" mean anomaly at epoch and $\left(t - t_o\right)$ is time since epoch

$$\omega_{DF} = \omega_o + \left[-\frac{3k_2\left(1-5\theta^2\right)}{2a_o''^2\beta_o^4} + \frac{3k_2^2\left(7-114\theta^2+395\theta^4\right)}{16a_o''^4\beta_o^8} + \frac{5k_4\left(3-36\theta^2+49\theta^4\right)}{4a_o''^4\beta_o^8} \right] n_o''\left(t-t_o\right) \quad (25)$$

$$k_4 = -\frac{3}{8} J_4 a_E^4 \quad (26)$$

According to previous cases, J_4 is the fourth gravitational zonal harmonic of the Earth

$$\Omega_{DF} = \Omega_o + \left[-\frac{3k_2\theta}{a_o''^2\beta_o^4} + \frac{3k_2^2\left(4\theta-19\theta^3\right)}{2a_o''^4\beta_o^8} + \frac{5k_4\theta\left(3-7\theta^2\right)}{2a_o''^4\beta_o^8} \right] n_o''\left(t-t_o\right) \quad (27)$$

And Ω_o is the "mean" longitude of ascending node at epoch.

$$\delta\omega = B^* C_3\left(\cos\omega_o\right)\left(t-t_o\right) \quad (28)$$

$$\delta M = -\frac{2}{3}\left(q_o - s\right)^4 B^* \xi^4 \frac{a_E}{e_o\eta}\left[\left(1+\eta\cos M_{DF}\right)^3 - \left(1+\eta\cos M_o\right)^3\right] \quad (29)$$

$$M_p = M_{DF} + \delta\omega + \delta M \quad (30)$$

$$\omega = \omega_{DF} - \delta\omega - \delta M \quad (31)$$

$$\Omega = \Omega_{DF} - \frac{21}{2}\frac{n_o'' k_2\theta}{a_o''^2\beta_o^2} C_1\left(t-t_o\right)^2 \quad (32)$$

$$e = e_o - B^* C_4\left(t-t_o\right) - B^* C_5\left(\sin M_p - \sin M_o\right) \quad (33)$$

$$a = a_o''\left[1 - C_1\left(t-t_o\right) - D_2\left(t-t_o\right)^2 - D_3\left(t-t_o\right)^3 - D_4\left(t-t_o\right)^4\right]^2 \quad (34)$$

$$IL = M_p + \omega + \Omega + n_o''\left[\frac{3}{2}C_1\left(t-t_o\right)^2 + \left(D_2 + 2C_1^2\right)\left(t-t_o\right)^3 + \frac{1}{4}\left(3D_3 + 12C_1D_2 + 10C_1^3\right)\left(t-t_o\right)^4 + \frac{1}{5}\left(3D_4 + 12C_1D_3 + 6D_2^2 + 30C_1^2D_2 + 15C_1^4\right)\left(t-t_o\right)^5\right] \quad (35)$$

$$\beta = \sqrt{\left(1-e^2\right)} \quad (36)$$

$$n = k_e / a^{3/2} \quad (37)$$

where $\left(t-t_o\right)$ is time since epoch.

In cases where perigee height is less than 220 kilometers, the equations for a and IL should be truncated after the C_1 term, and all terms involving C_5, $\delta\omega$, and δM should be dropped.
Add the long-period periodic terms

$$a_{xN} = e\cos\omega \tag{38}$$

$$IL_L = \frac{A_{3,0}\sin i_o}{8k_2 a\beta^2}\left(e\cos\omega\right)\left(\frac{3+5\theta}{1+\theta}\right) \tag{39}$$

$$a_{yNL} = \frac{A_{3,0}\sin i_o}{4k_2 a\beta^2} \tag{40}$$

$$IL_T = IL + IL_L \tag{41}$$

$$a_{yN} = e\sin\omega + a_{yNL} \tag{42}$$

Next step involves solving the Kepler's equation for $\left(E+\omega\right)$by defining

$$U = IL_T - \Omega \tag{43}$$

$$\left(E+\omega\right)_{i+1} = \left(E+\omega\right)_i + \Delta\left(E+\omega\right)_i \tag{44}$$

$$\Delta\left(E+\omega\right)_i = \frac{U - a_{yN}\cos\left(E+\omega\right)_i + a_{xN}\sin\left(E+\omega\right)_i - \left(E+\omega\right)_i}{-a_{yN}\sin\left(E+\omega\right)_i - a_{xN}\cos\left(E+\omega\right)_i + 1} \tag{45}$$

$$\left(E+\omega\right)_1 = U \tag{46}$$

The following equations are used to calculate preliminary quantities needed for short-period periodics.

$$e\cos E = a_{xN}\cos\left(E+\omega\right) + a_{yN}\sin\left(E+\omega\right) \tag{47}$$

$$e\sin E = a_{xN}\sin\left(E+\omega\right) - a_{yN}\cos\left(E+\omega\right) \tag{48}$$

$$e_L = \left({a_{xN}}^2 + {a_{yN}}^2\right)^{1/2} \tag{49}$$

$$p_L = a\left(1 - {e_L}^2\right) \tag{50}$$

$$r = a\left(1 - e\cos E\right) \tag{51}$$

$$\dot{r} = k_e\frac{\sqrt{a}}{r}e\sin E \tag{52}$$

$$r\dot{f} = k_e \frac{\sqrt{p_L}}{r} \tag{53}$$

$$\cos u = \frac{a}{r}\left[\cos\left(E+\omega\right) - a_{xN} + \frac{a_{yN}\left(e\sin E\right)}{1+\sqrt{1-{e_L}^2}}\right] \tag{54}$$

$$\sin u = \frac{a}{r}\left[\sin\left(E+\omega\right) - a_{yN} - \frac{a_{xN}\left(e\sin E\right)}{1+\sqrt{1-{e_L}^2}}\right] \tag{55}$$

$$u = \tan^{-1}\left(\frac{\sin u}{\cos u}\right) \tag{56}$$

$$\Delta r = \frac{k_2}{2p_L}\left(1-\theta^2\right)\cos 2u \tag{57}$$

$$\Delta u = -\frac{k_2}{4{p_L}^2}\left(7\theta^2-1\right)\sin 2u \tag{58}$$

$$\Delta\Omega = \frac{3k_2\theta}{2{p_L}^2}\sin 2u \tag{59}$$

$$\Delta i = \frac{3k_2\theta}{2{p_L}^2}\sin i_o \cos 2u \tag{60}$$

$$\Delta\dot{r} = -\frac{k_2 n}{p_L}\left(1-\theta^2\right)\sin 2u \tag{61}$$

$$\Delta r\dot{f} = \frac{k_2 n}{p_L}\left[\left(1-\theta^2\right)\cos 2u - \frac{3}{2}\left(1-3\theta^2\right)\right] \tag{62}$$

The osculating quantities are calculated with the addition of short-period periodics

$$r_k = r\left[1-\frac{3}{2}k_2\frac{\sqrt{1-{e_L}^2}}{{p_L}^2}\left(3\theta^2-1\right)\right] + \Delta r \tag{63}$$

$$u_k = u + \Delta u \tag{64}$$

$$\Omega_k = \Omega + \Delta\Omega \tag{65}$$

$$i_k = i_o + \Delta i \tag{66}$$

$$\dot{r}_k = \dot{r} + \Delta\dot{r} \tag{67}$$

$$\dot{rf}_k = \dot{rf} + \Delta\dot{rf} \tag{68}$$

Unit orientation vectors are calculated by

$$\mathbf{U} = \mathbf{M}\sin u_k + \mathbf{N}\cos u_k \tag{69}$$

$$\mathbf{V} = \mathbf{M}\cos u_k - \mathbf{N}\sin u_k \tag{70}$$

where

$$\mathbf{M} = \begin{Bmatrix} M_x = -\sin\Omega_k \cos i_k \\ M_y = \cos\Omega_k \cos i_k \\ M_z = \sin i_k \end{Bmatrix} \tag{71}$$

$$\mathbf{N} = \begin{Bmatrix} N_x = \cos\Omega_k \\ N_y = \sin\Omega_k \\ N_z = 0 \end{Bmatrix} \tag{72}$$

Then, position and velocity are given by

$$\mathbf{r} = r_k \mathbf{U} \tag{73}$$

and

$$\dot{\mathbf{r}} = \dot{r}_k \mathbf{U} + \left(\dot{rf}\right)_k \mathbf{V} \tag{74}$$

2.2 Propagation models modifications

SGP propagation model was modified in time. Several minor points in the original SGP4 paper emerged where performance of SGP4 could be improved. To maximise the usefulness of all of these features, one should ideally use Two Line Elements formed with differential correction, using an identical model as well (Vallado, D. et al. 2006). Next chapter will shed some light on what Two Line Elements are.

3. Two Line Elements

Orbit tracking programs require information about the shape and orientation of satellite orbits. That information was available from different websites. One of most common quoted sources is CelesTrak.com website, maintained by the group of satellite tracking enthusiasts. As a result of legislation passed by the US Congress and signed into law on 2003 November 24 (Public Law 108-136, Section 913), which was updated in 2006 (US National Archives, 2006.), Air Force Space Command (AFSPC) has embarked on a three-year pilot program to provide space surveillance data—including NORAD two-line element sets (TLEs)—to non-US government entities (NUGE). Since US Public Law prohibits the redistribution of the data obtained from this new NUGE service "without the express approval of the Secretary of Defence"a lot of other sources were immediately shut down.

CelesTrak has received continuing authority to redistribute Space Track data from US government and that way become one of the most useful information sources for the community.
TLE's are redistributed in a form shown in Fig. 5. All relevant parameters are color-coded and explained in Table 1.

```
ISS
1 25544U98067A10102.85853206  .00025654 00000-017456-309629
2 25544  51.6472205.9374 0004892 166.2878 293.9622 15.74716373653188
```

Fig. 5. Two Line elements set

Element	Description
ISS	Satellite name
1 25544U 2 25544	Satellite number
98067A	International designator
51.6472	Inclination
10102.85853206	Epoch Year & Day Fraction
205.9374 0004892	Right Ascension of the Ascending Node
.00025654	First Derivative of Mean Motion
00000-0	Second Derivative of Mean Motion
166.2878	Eccentricity
293.9622	Argument of Perigee
15.747163736	Mean Anomaly
53188	Revolution Number at Epoch & Checksum
17456-3	Drag Term
0	Ephemeris Type
9629	Element number & Checksum

Table 1. Two Line Elements Explained

There are several things to consider. The accuracy of the original TLEs is not known. Some TLE data propagates into future quite well, while, the next set of elements can depart dramatically after only a day or less. Methods to overcome this problem are explained in (Vallado, D. et al. 2006).

4. Integrating Mathematical Models

Our intention was to integrate all orbital propagation models into one C# program. Integral version of the program can be downloaded from
http://medlab.elfak.ni.ac.rs/spacetrack/sgpsdp.rar. It is important to highlight main program methods used for satellite position calculations:

```
publicvoid SGP(int IFLAG, double TSINCE)
publicvoid SGP4(int IFLAG, double TSINCE)
publicvoid SDP4(int IFLAG, double TSINCE)
publicvoid SGP8(int IFLAG, double TSINCE)
publicvoid SDP8(int IFLAG, double TSINCE)
```

Previous FORTRAN IV code produced by T.S. Kelso in 1988 according to (Hoots, F. R et al. 1980) was not optimized and hard to execute on modern parallel (multi-core) architectures.
The FORTRAN implementation of the SGP4 and SDP4 model in respective methods is rudimentary for the propagation process. It was necessary to produce functions which would help us calculate position $\vec{r}$ and velocity $\vec{r}$ of a satellite at any given time by using the TLE data. Models specified in (Hoots, F. R. and al., 1980) from the original FORTRAN IV code are ported to C# in respect to the corrections made during the years, especially in the SDP4 subroutine DEEP. C# code contains the same variable names and structures as in the original FORTRAN routines to ensure compatibility and expandability.
Additional encapsulation was done with the creation of ActiveX component ready to be integrated in any .NET project.

5. NAVSTAR Satellite Tracking Software

NAVSTAR satellite tracking software presented in this paper is also based on the mathematical SGP4/SDP4 model. Program uses two line elements set as an input to calculate and visualize satellite's position in Space. It can be used to navigate telescopes to space objects passing over certain point on Earth. The complete mathematical model is encapsulated in ActiveX control, so it acts like a black box. The data is provided from TLEs and on the other end viewport coordinates are calculated.
NAVSTAR has three basic functions:

- Graphical display of satellite positions in real-time, simulation, and manual modes;
- Tabular display of satellite information in the same modes;
- Generation of tables (ephemerides) of past or future satellite information for planning or analysis of satellite orbits.

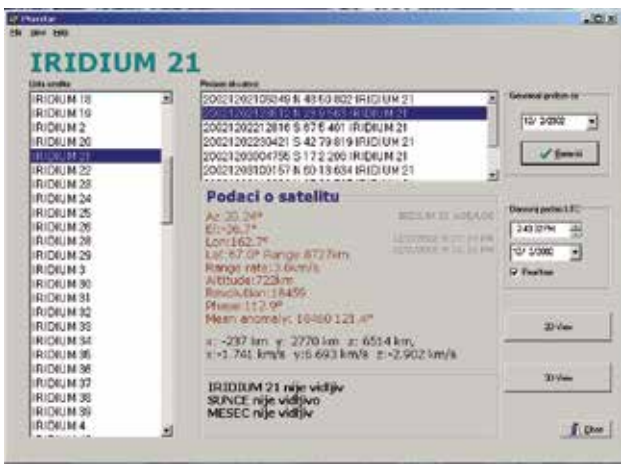


Fig. 6. Satellite selection dialog and Table Window

The principal feature of NAVSTAR is a series of Map Windows, which display the current position of satellites and observers on a simple world map, together with information such as bearing (azimuth), distance, and elevation above the observer's horizon. The maps may

be updated in real time, simulated time, or manually set to show the situation at any given moment of time, past or future.
An additional Table Window displays much more-detailed information about one or more satellites in a tabular form. The tabulated items can be selected and rearranged to fit the screen. This information can be updated in real-time, manual, or simulation modes as illustrated in Fig. 6.

Also, satellite 2D footprint (Fig.7) tracking is available, as well as a 3D view (Fig.8). Tracking algorithms SGP4 and SDP4 give considerable accuracy and opportunity of efficient computation of viewing opportunities. It's also possible in 3D view to make a prediction on satellites position in the future, or to see its position in the past. All is based on the information gathered from TLE's.
The preciseness of visualization depends on accuracy and age of gathered TLE data.

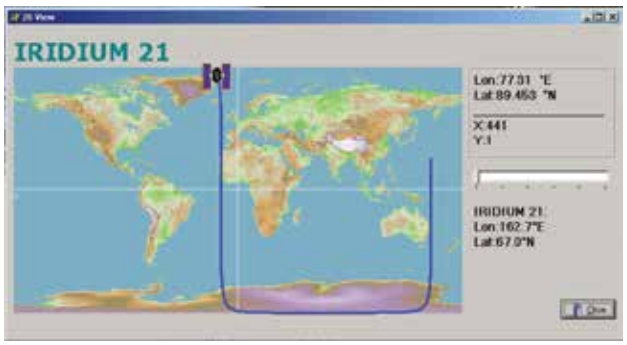


Fig. 7. 2D View

Regarding the 3D View (Fig.8), options for variable view angle, zoom and time increment are implemented. This gives a user the opportunity to view satellite from all angles and possibility to see its path (orbit), area on the Earth covered by its signal (in a form of beam) and real-time movement, as well as possible faster movement caused by a time speed up.

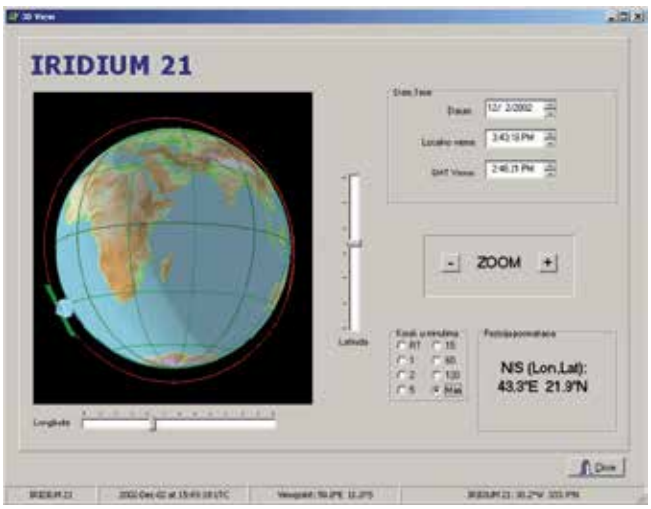


Fig. 8. 3D Satellite tracking View

The part of Earth not covered with the Sun light is dimmed on the globe, so the user can predict when it will become possible to see the satellite by a naked eye.
This software can be used to visualise the orbit trajectory of a satellite under different points of view. It gives the user the freedom of being able to study the satellite's ground conjunctions by tracking the satellite over the ground, or, with another approach, by calculating the elevation and azimuth angle of the satellite from a static ground station position.
Another very useful functionality, previously mentioned is satellite's footprint visualisation. These functionalities can be used, from an engineering point of view, to adjust ground station dishes in order to establish reliable links to the satellite by calculating it's precise position with the help of TLE data sets.
3D view can be used to simulate satellite's orbit around the Earth. Time lapse function enables user to see the orbit in the future, and exact satellite's position so we would be able to see which orbit parameters the satellite has to have in order to fulfil its task.

6. Conclusion

It is assumed that "space age" started with the first artificial satellite in the orbit around the planet. Nowadays, satellites are used for various different purposes: telecommunications, broadcasting, observation, imaging and even espionage. What they all have in common is the fact that they all must obey the rules of celestial mechanics. To be able to visualise the motion, software presented in this chapter was created.
For the satellite dynamics, the SGP4 and SDP4 models by NORAD were implemented. Both SGP4 and SDP4 are based on fundamental laws stated by Newton and Kepler. One of the biggest advantages of SGP4/SDP4 models is that they've been recognized and verified by NORAD thus providing a precise and manageable mathematical framework for the orbital calculations. But bear in mind that those are not perfect models. They work with mean values. NORAD has removed periodic variations in a particular way, and the models, in their present form do not contain numerical integration methods.
Future investigations and updates will improve propagation making it more precise. This process will certainly increase the complexity, so the balance between complexity and preciseness must be kept.

7. References

Binderink, A. L.; Radomski, M.S.; Samii M. V. (1989) Atmospheric drag model calibrations for spacecraft lifetime prediction; *In NASA, Goddard Space Flight Center, Flight Mechanics/ Estimation Theory Symposium*; 445-458

Bunnell, P. (1981) Tracking Satellites in Elliptical Orbits; *Ham Radio Magazine*; 46-50.

Hoots, F. R.; Roehrich R. L. (1980) Models for propagation of NORAD element sets; *Project Spacecraft Report No. 3*; Aerospace Defence Command, United States Air Force, 3-6

King-Hele, D.G (1983), Observing Earth Satellites, Macmillan

Montenbruck, O.; Gill, E.. (2000) Real-Time estimation of SGP4 Orbital Elements from GPS Navigation Data; *International Symposium SpaceFlight Dynamics, MS00/28*; 2-3

US National Archives and Records Administration.(2006) US Public Law, *Section 109–364*; Oct. 17, 2006, Stat. 2355.

Vallado A. D.; Crawford P. (2006) SGP4 Orbit Determination; *American Institute of Aeronautics and Astronautics publication*; 19-21.

15

Interference in Cellular Satellite Systems

Ozlem Kilic[1] and Amir I. Zaghloul[2,3]
[1]*The Catholic University of America, Washington, DC, U.S.A.*
[2]*Virginia Polytechnic Institute and State University, VA, U.S.A*
[3]*US Army Research Laboratory, MD, U.S.A.*

1. Introduction

In cellular satellite communications systems, a given coverage area is typically filled with a network of contiguous spot beams, which carry concentrated radiation along preferred directions. The coverage regions for such applications are typically large areas, such as continents and many beams need to be generated.

Due to bandwidth limitations in cellular communications, the same bandwidth is allocated to beams which are isolated spatially. This is known as frequency reuse, and the beams operating at the same frequency are referred as co-channel beams. While this approach allows a large coverage area with limited bandwidth, the co-channel beams have the potential to interfere with each other. This is known as co-channel interference and its nature and how it could be reduced is the focus of this chapter.

The interference in multiple beam satellite communications systems will be investigated under two different approaches. First approach, which is the conventional way of defining beam coverage on earth, is discussed in Section 2. This will be referred to as spot beam coverage as explained in further detail later. The interference will be investigated for two cases; first is the uplink where interference at the satellite antenna is the main concern, and the second is the downlink where interference at the user terminal is calculated. Section 3 discusses a new method of defining beam coverage on earth, referred to as sub-beam coverage. The motivation is to keep the coverage on earth identical but reduce the satellite antenna size as much as 50% (Kilic & Zaghloul, 2009). The advantages and overall performance of the sub-beam approach in terms of interference is the subject of Section 3.[1]

2. Interference in Cellular Satellite Systems

In multibeam satellite systems, the coverage area is divided into a number of beams often referred to as spot beams, which are much smaller in size and cover the area contiguously.

---

[1] Copyright 2009 American Geophysical Union, This chapter has material substantially reproduced, with permission, from Radio Science, Volume 44, No. 1, January 2009, „Antenna Aperture Size Reduction Using Subbeam Concept in Multiple Spot Beam Cellular Satellite Systems,, O. Kilic and A.I. Zaghloul.

Since satellite systems are bandwidth limited, the sub-division of beams into smaller portions allow for frequency reuse to increase capacity. The available bandwidth is shared among these beams as depicted in Figure 1 for reuse factor of n.

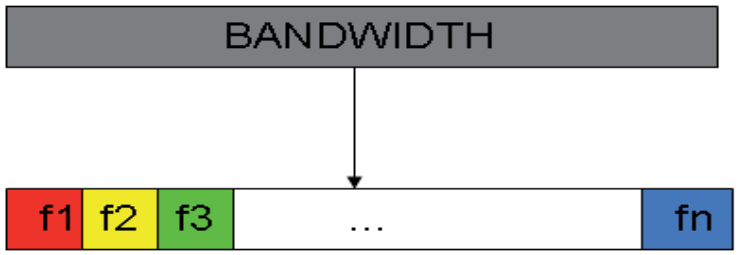


Fig. 1. Frequency reuse in multi-beam satellite communications

The size of the antenna that generates these beams on earth is related directly to the peak gain at the center of the spot beams and the smallest spot beam size. The spot beams are typically defined by the contours at 3 or 4 dB down from the peak power at the center of the beam.

2.1 Coverage on Earth

Achieving a contiguous coverage is important so that there are no regions without service in the coverage area. Since the beams are defined by the projection of antenna patterns on earth at a certain contour value, they tend to be close to circular shape. These circles on earth need to be structured so that they overlap with each other to avoid any gaps in the coverage area. In order to have a systematic approach, these can be represented by various geometric lattices that tessellate. A few such possibilities are shown in Figure 2. It is often the hexagon that is used in the system design as it closely represents a circle, i.e. for the same distance from the center to the edge, the hexagon area is closest to that of the circle that circumscribes it. Therefore the hexagon represents the beam which circumscribes it as shown in Figure 3. This assures that there are no gaps between beams. Then the system is designed based on this artificial hexagonal geometry on earth as depicted in Figure 4.

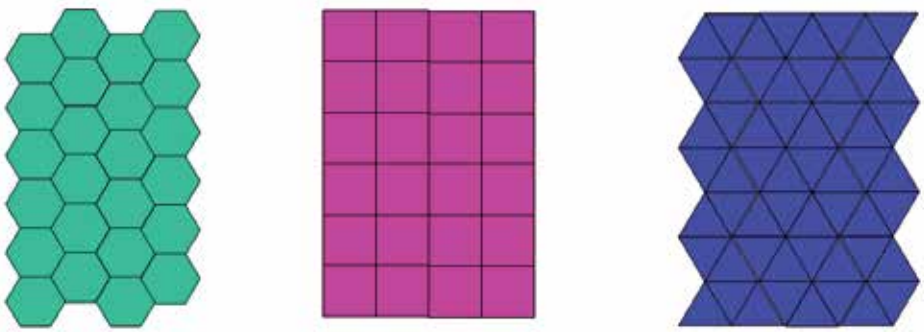


Fig. 2. Contiguous coverage on earth using tessalating shapes – hexagonal, square and triangular lattices

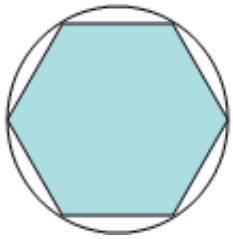


Fig. 3. Hexagonal representation of a circular beam

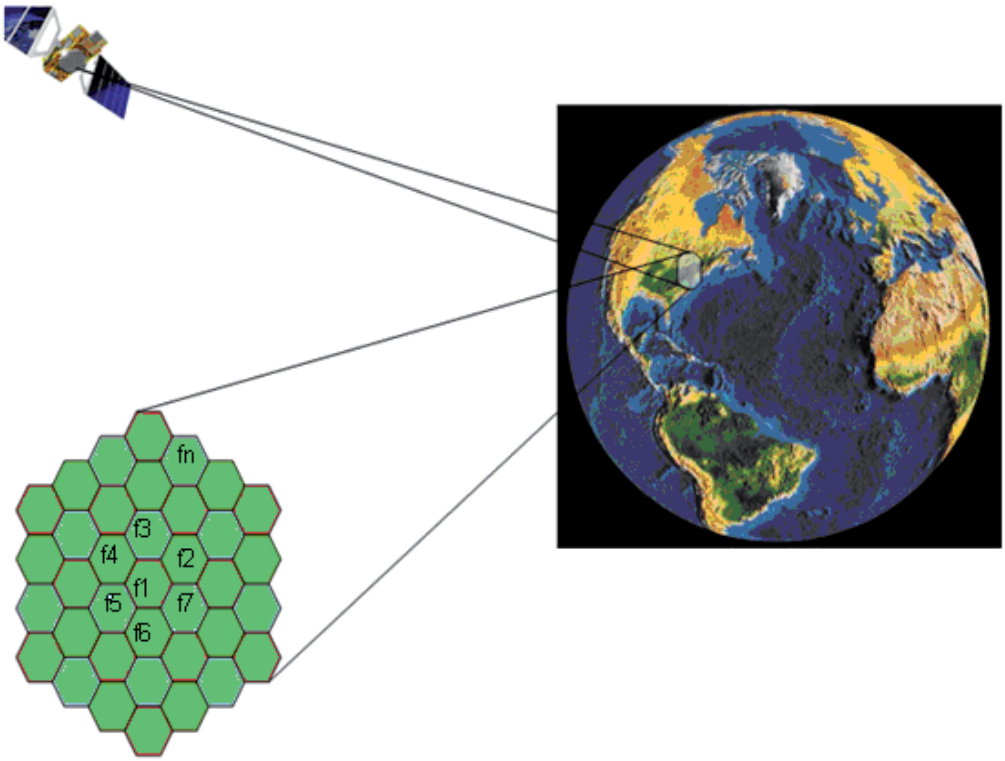


Fig. 4. Hexagonal Coverage on Earth

2.2 Frequency Reuse

Since the satellite systems typically serve large ares such as countries or continents, a large number of beams need to share the available bandwidth. Therefore, the available bandwidth within a beam becomes a very limited resource, as is implied in Figure 1 earlier. To circumvent this, a frequency reuse scheme is often utilized. This is based on reusing the same frequencies in spatially isolated beams. Therefore, the available bandwidth is divided into a smaller number of beams than the total number of beams in the coverage area. The set of contiguous beams that share the total available bandwidth is known as a cluster. The clusters are then repeated in the coverage area relying on the fact that the beams operating at the same bandwidth will be separated from each other sufficiently so that they do not interfere with each other.

There are only a discrete set of possible cluster sizes, N, to accommodate a contiguous coverage for a hexagonal geometry [Mehrotra, 1994]. The possible number of beams in a cluster which would form a tessalating shape is given by:

$$N = i^2 + j^2 + i \times j \tag{1}$$

where N, is the number of beams in the cluster; i.e. the number of beams that share the total bandwidth, and i, j are non-negative integer numbers. This results in possible cluster sizes of 3, 4, 7, 9, etc. An example of how these clusters are formed is shown in Figure 5.

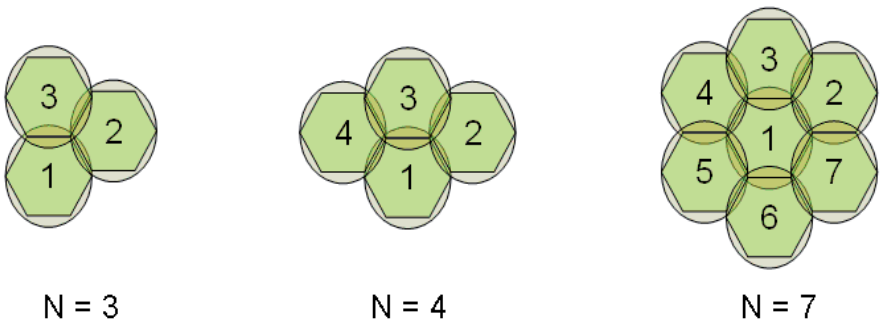


Fig. 5. Different cluster options for hexagonal lattice

2.3 Co-channel Beams and Tiers

The clusters as depicted in Figure 5 for reuse factors of 3, 4 and 7, are repeated to fill the required coverage area on earth. An example of how a cluster size of three (i.e. i=1, j=1) would be used to fill a given area is shown in Figure 6. A cluster size of three implies that the total available bandwidth is shared between three beams. The numbers 1, 2, 3 are used to identify the beams using the corresponding bandwidth segments. Therefore, beams with the same number imply that they use the same frequency.

Beams operating at the same frequency are known as co-channel beams. In Figure 6, the co-channel beams are shown in the same color and labeled with the same bandwidth segment number. It can be observed that the location of co-channel beams follow a pattern. They can be grouped by their distance with respect to a reference beam. The set of co-channel beams which have same distance from a reference beam are said to fall on a tier. Therefore tiers define a set of beams equidistant from a reference point. In a hexagonal geometry, each tier consists of six or twelve beams. Figure 7 shows the first four tiers for the frequency reuse of three.

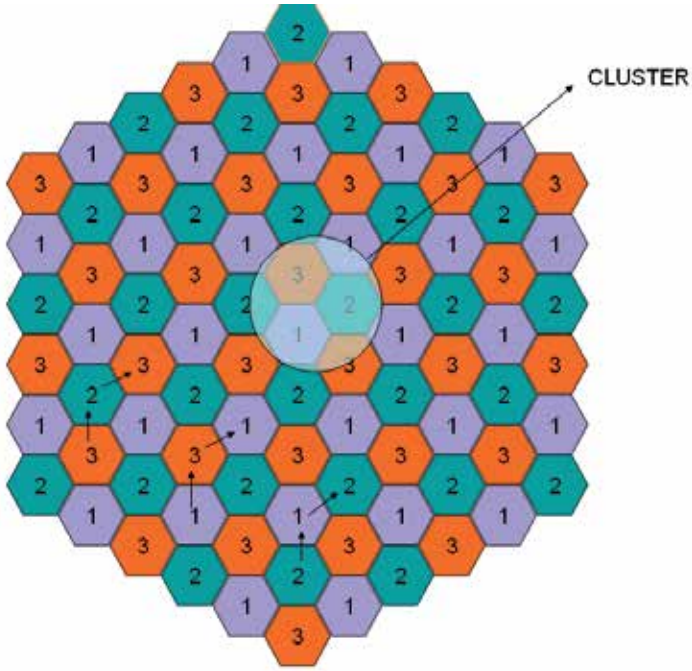


Fig. 6. Cluster coverage, N=3

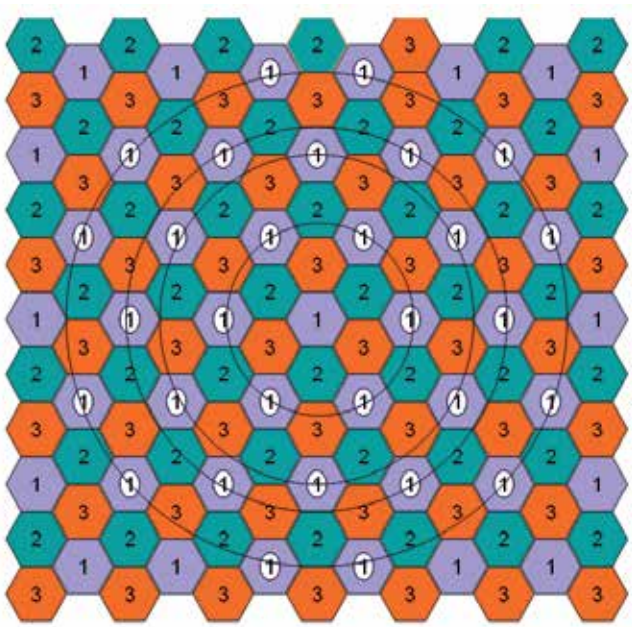


Fig. 7. Tiers in frequency reuse

The beams that lie on a tier are equidistant from the beam at the center of the tier, and for an azimuthally symmetric power distribution, beams on a tier would contribute the same amount

of interference. As the tier's number increases, the diameter hence the distance from the center beam increases, reducing the contribution from the beams in that tier compared to the beams on closer tiers assuming the radiation decreases away steadily from the antenna peak.

As can be observed from Figures 6 and 7, higher frequency reuse numbers result in tiers with larger diameters, thereby increasing the distance between co-channel beams and reducing the total number of beams operating at the same frequency. However, this is done at the expense of reduced bandwidth within a beam, a trade off which needs to be decided by system engineers based on the requirements of a particular system.

2.4 Antenna Pattern and Spot Beam Generation

Due to their ability to generate multiple beams simultaneously, phased arrays are a natural choice for multi-beam satellite antennas. Each beam in the coverage is generated by electronically scanning the beam. A key parameter in satellite antenna design is the directivity of antenna, which defines how well the antenna focuses the power in the desired direction.

The radiation pattern of an array antenna depends on the array factor. For a MxN planar array, the array factor is given by (Balanis, 2005)

$$AF = \sum_{m=1}^{M}\sum_{n=1}^{N} I_{mn} e^{j(m-1)(kd_x \sin\theta\cos\phi+\beta_m)} e^{j(n-1)(kd_y \sin\theta\sin\phi+\beta_n)} \tag{2}$$

where I_{mn} are the voltages that feed the mn^{th} element in the array. The array factor is related to the antenna directivity which measures how well the input power is focused along a given direction and can be computed as follows (Stutzman, 1998)

$$D_0 = \frac{4\pi[\mathrm{AF}(\theta_0,\phi_0)][\mathrm{AF}(\theta_0,\phi_0)^*|_{\max}}{\int_0^{2\pi}\int_0^{\pi}[\mathrm{AF}(\theta_0,\phi_0)][\mathrm{AF}(\theta_0,\phi_0)]^* \sin\theta\cdot d\theta\cdot d\phi} \tag{3}$$

In the case of a large array with uniform excitation, this equation can be approximated by

$$D_0 = \frac{\pi^2}{\Omega_A} \tag{4}$$

where the denominator is the subtended beam angle defined by

$$\Omega_A = \frac{\theta_{x0}\theta_{y0}\sec\theta_0}{\left[\sin^2\varphi_0 + \frac{\theta_{y0}^2}{\theta_{x0}^2}\cos^2\varphi_0\right]^{\frac{1}{2}}\left[\sin^2\varphi_0 + \frac{\theta_{x0}^2}{\theta_{y0}^2}\cos^2\varphi_0\right]^{\frac{1}{2}}} \tag{5}$$

In equation (5) the angles φ_0 and θ_0 represent the direction at which the directivity is computed, and θ_{x0} and θ_{y0} denote the half power beamwidths of the array along the directions x and y, the plane on which the array resides. In this chapter we will assume a 40x40 square planar array with isotropic elements separated by half wavelength as shown in Figure 8.

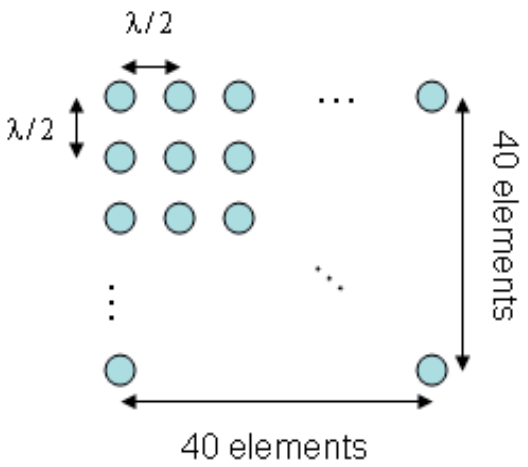


Fig. 8. A uniformly spaced 40x40 element planar array

The plot of the directivity in two different planes for this particular array is shown in Figures 9-10. The 40x40 element antenna is sized to have an edge gain of 36 dB defined at its -4dB contour relative to the peak. Thus, the peak gain of the antenna is 40 dB, with a beamwidth of 1.44 degrees at the - 4 dB down contour from the peak gain. The spot beam in the coverage area is defined by this contour. The contour plot for the same antenna is shown in Figure 11. The beam is generated at the center as denoted by the white circle.

As seen in the contour plot, the energy is radiated to the whole coverage area even though most of the power is focused in the spot beam. This "leak" into other regions is not a problem for beams operating at different frequencies as filtering will eliminate this energy. But for co-channel beams; i.e. beams operating at the same frequency, this leaking energy creates the interference which is the focus of this chapter.

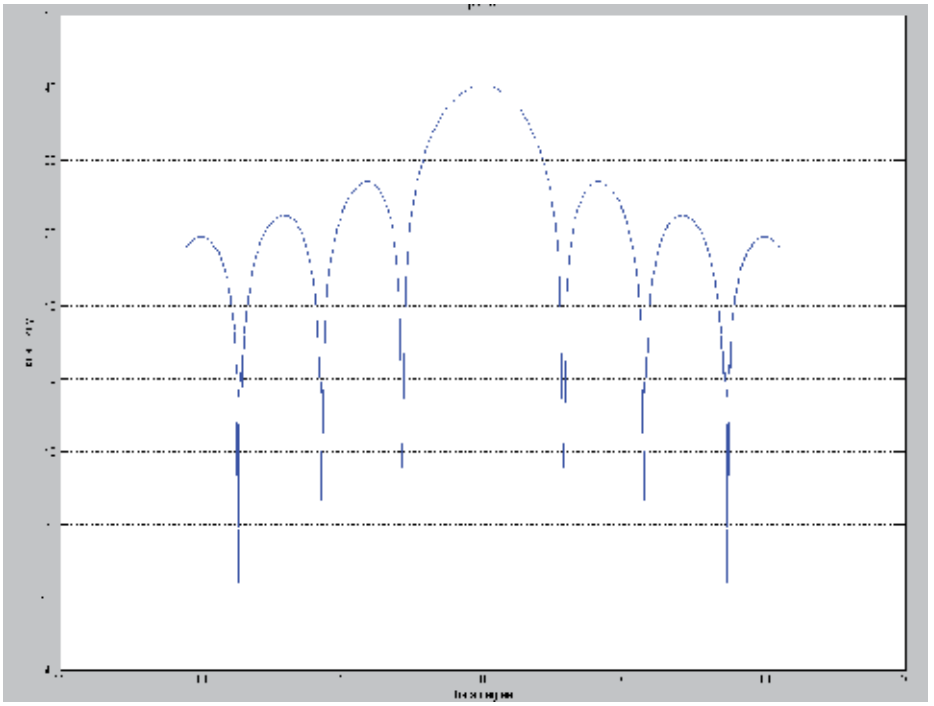


Fig. 9. Array factor for the 40x40 element array, φ = 0 plane

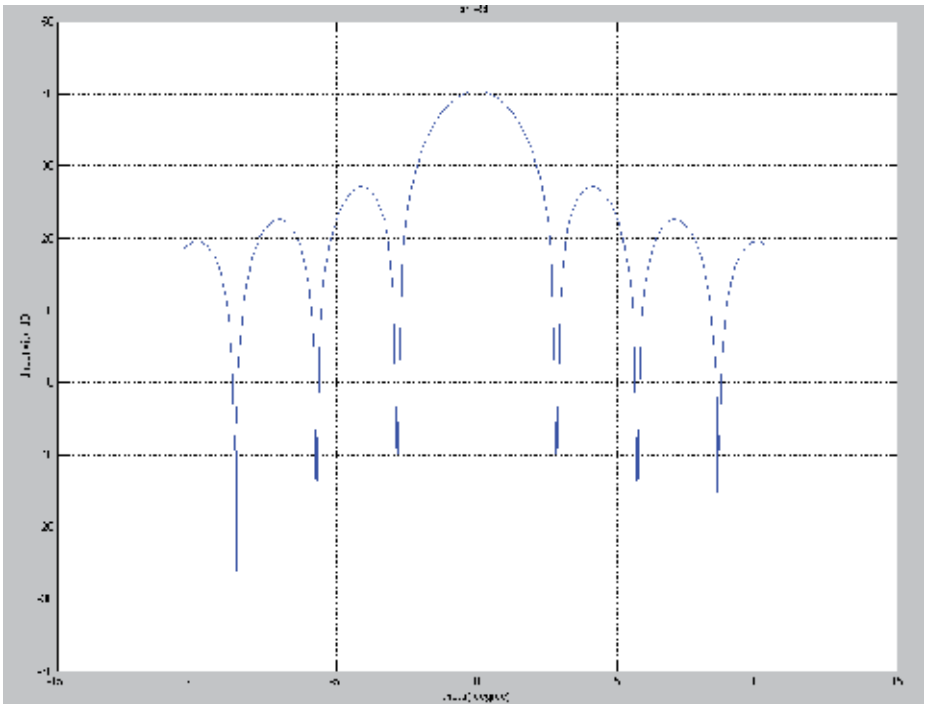


Fig. 10. Array factor for the 40x40 element array, φ = 90 plane

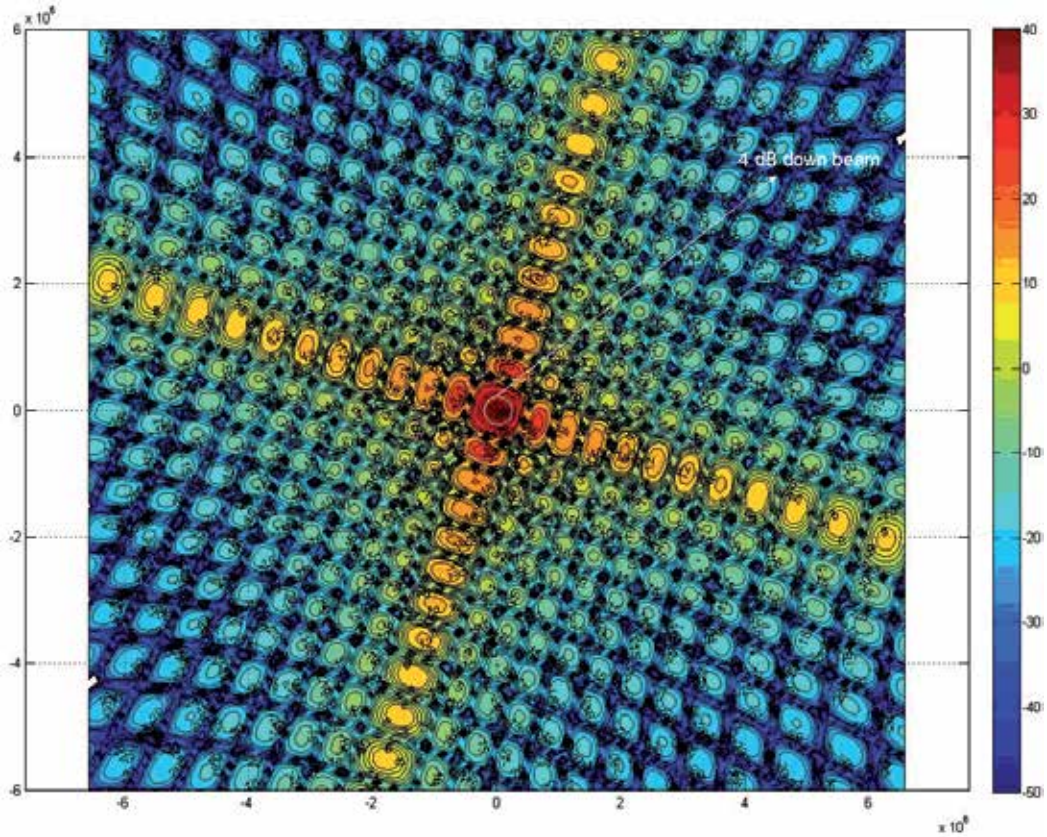


Fig. 11. Contour plot of directivity for 40x40 element array

2.5 Co-channel Interference

The energy leak from a beam into other beams operating at the same frequency causes co-channel interference. Figure 12 shows a simplistic diagram of the co-channel interferers in both the uplink and the downlink. The downlink interference is primarily a function of the reuse number and of the aggregate power due to the power in the side lobes of the interfering co-channel spot beams that is received in an earth station receiver. On the other hand, the uplink co-channel carrier-to-interference ratio is dependent upon the reuse number and the number of co-channel users transmitting simultaneously and received at the side lobes of the interfered beam. This section will focus solely on the downlink interference.

The farther the co-channel beams are separated from each other, the lower is the expected interference levels as the antenna pattern gradually tapers off with increasing angles away from the peak point. However, this gradual taper in the antenna pattern is oscillatory due to the side lobes. Therefore, the co-channel interference in a multiple beam system depends on the antenna pattern as well as the cluster configuration.

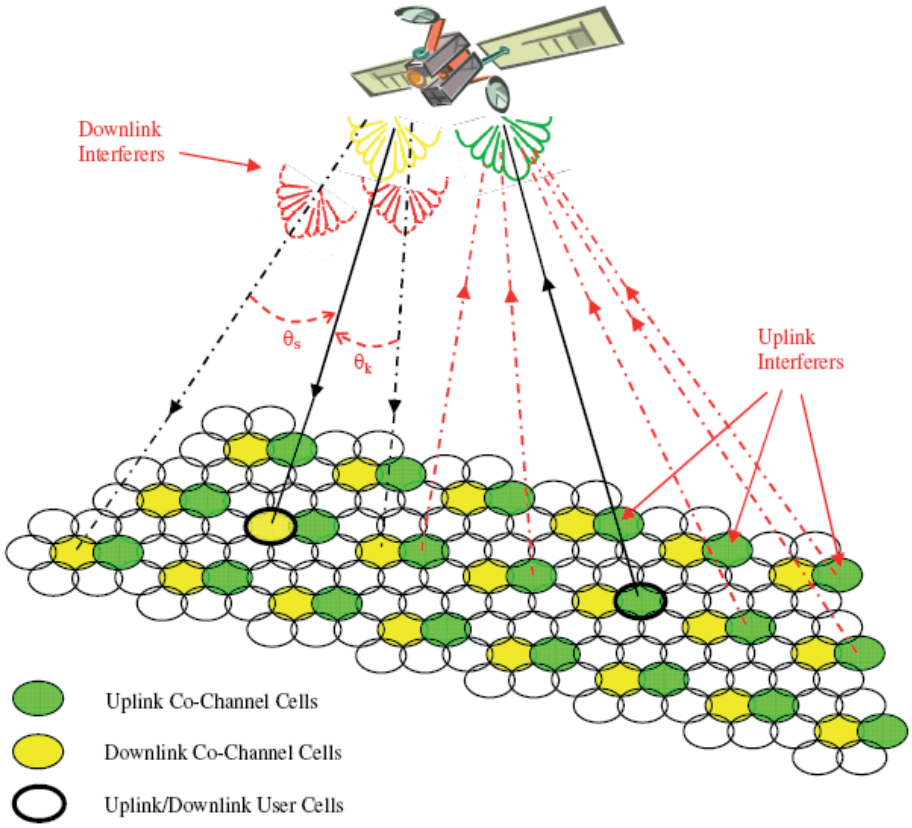


Fig. 12. Co-channel interference concept in multibeam satellite communication systems

A parameter for the system performance is the carrier to interference ratio, which defines how much better the intended signal is compared to the interfering signals:

$$\left.\frac{Carier}{Interference}\right|_{\text{at beam center}} = \left.\frac{C}{I}\right|_{\text{at beam center}} = \frac{D_{Max}}{D_{Interference}} \tag{6}$$

where the interference term is the sum of powers received from all co-channel beams. We can express equation (6) equivalently in dB as:

$$10\log(\frac{C}{I}) = 10\log(D_{\max} - \log D_{Interference}) \tag{7}$$

For an azimuthally symmetric antenna pattern, the co-channel beams lying on the same tier would result in equal interference levels as they are equidistant from the reference point. Therefore, categorizing all co-channel beams by their tier number helps simplify the interference calculations. Although the square array does not produce an azimuthally

symmetric pattern, the power levels tend to decrease as the tier number increases, i.e. as the co-channel beams are farther from the reference point. This concept is shown in Figures 12 and 13 for frequency reuse factors of 3 and 7. We observe that for the larger size cluster (i.e. N=7), the diameter of the tiers are larger for the same tier number. This implies that, larger cluster sizes will likely have less co-channel interference.

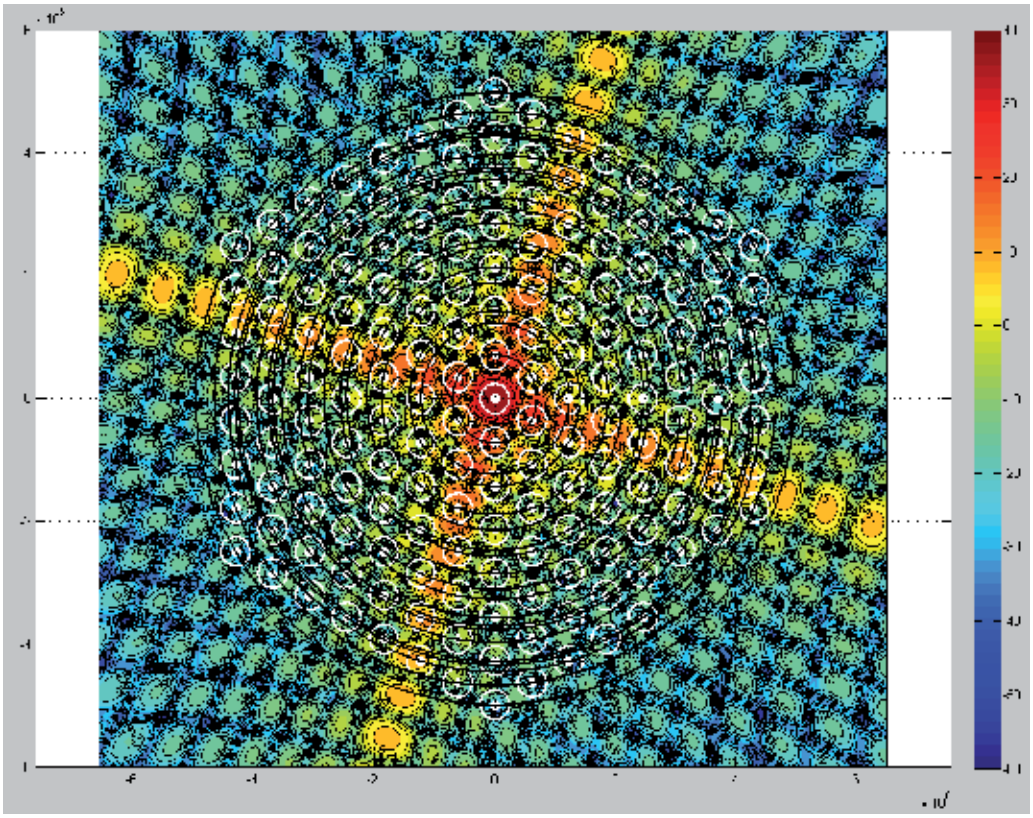


Fig. 13. Tiers for reuse factor = 3 overlaid on the contour plot of the 40x40 array

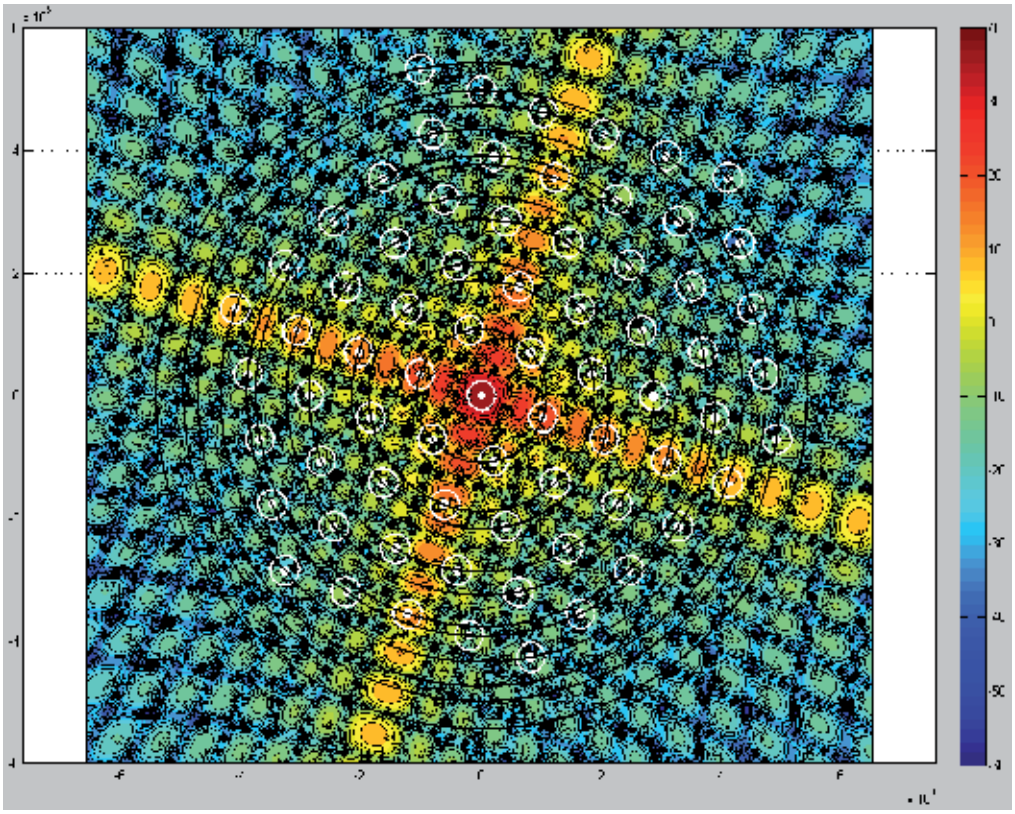


Fig. 14. Tiers for reuse factor = 7 overlaid on the contour plot of the 40x40 array

2.6 Example – LEO System Downlink Interference

In this section we calculate the total downlink interference in using a Low Earth Orbiting (LEO) satellite system. The relevant parameters of the system are as follows:

Altitude = 9400 km
Peak Gain = 40.0 dB
Spot Beam edge @ 4dB (36.0 dB)
Frequency = 1.9 GHz
Minimum elevation angle = 22^{o}
Number of beams on earth = 500

The spot beams generated by the 40x40 element phased array antenna for frequency reuse of three are shown in Figure 15. The radii of all spot beams are assumed equal to that of the center spot beam, and are computed using the beam width at -4dB relative to the peak gain and the altitude as follows:

$$Radius = H(altitude) * \tan(\theta_{4dB\_down}) \tag{8}$$

For the 40x40 element array antenna, the beam width is 0.0252 radians (1.44 degrees), resulting in a radius of 236km. The beam coverage on earth for this system for reuse factor of 3 is shown in Figure 15. The red lines depict the different tiers for this reuse factor, and same numbers denote the same bandwidth of operation, hence co-channel beams.

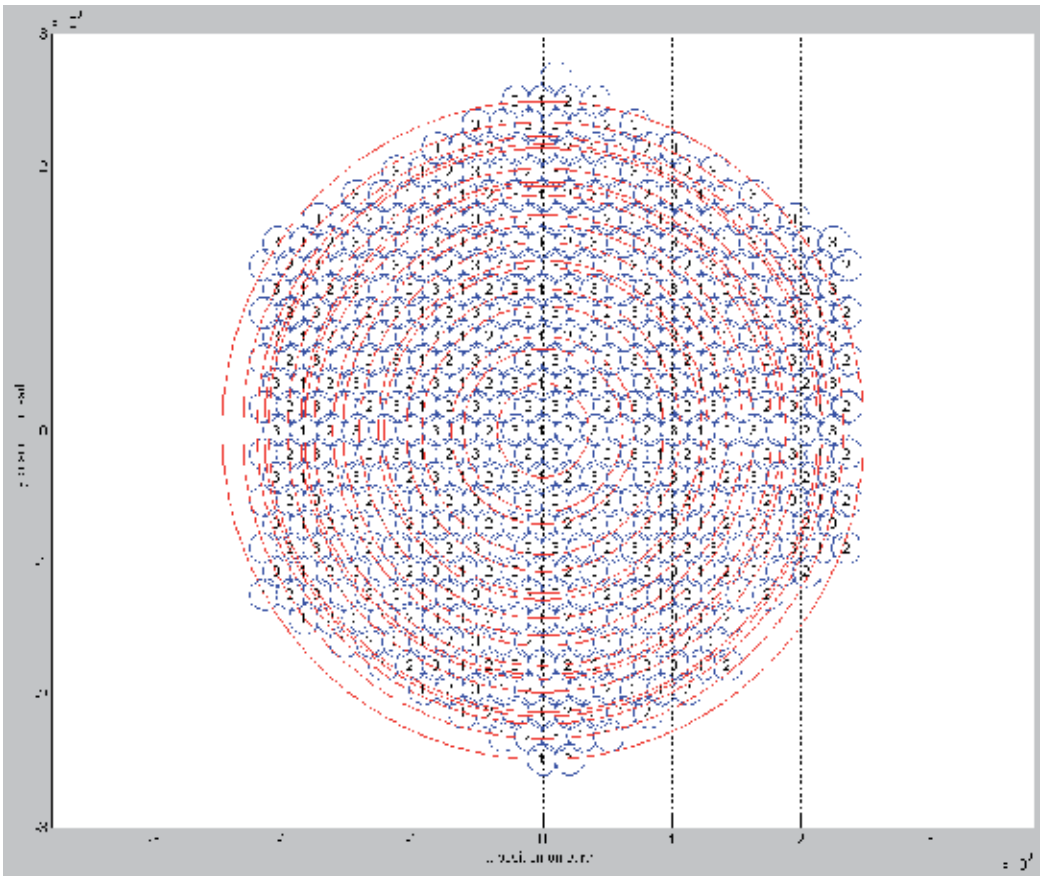


Fig. 15. LEO system, reuse of 3

The total interference from each tier for the center beam is shown for different reuse factors in Figure 16. We observe that there are fewer tiers for the same coverage when a higher frequency reuse factor is used. It is also observed that the interference levels are higher (hence the carrier to interference ratio is lower) from the tiers closest to the beam of interest. Since the square array is not azimuthally symmetric, we observe oscillations in the interference levels between tiers due to the side lobe level variations. We also observe that higher frequency reuse factor results in lower interference as there are fewer co-channel beams in the coverage area and they are spread farther apart.

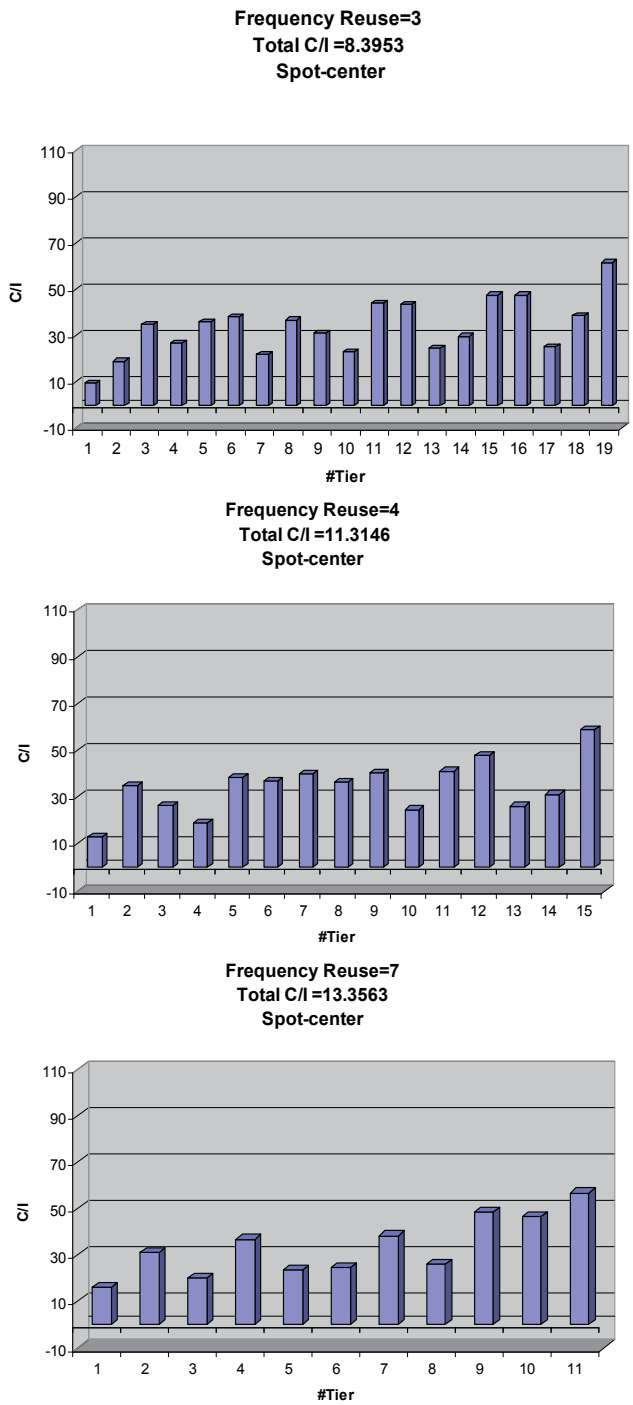


Fig. 16. Interference versus tier number (a) reuse = 3 (b) reuse = 4 (c) reuse = 7

3. Sub-Beam Concept

As we have seen so far, the coverage area in multibeam satellite systems is divided into a number of beams often referred to as spot beams, which are much smaller in size. As discussed before, since satellite systems are bandwidth limited, the sub-division of beams into smaller portions allow for frequency reuse to increase capacity. The size of antenna that generates these beams on earth is related directly to the peak gain at the center of the spot beam and the smallest spot beam size. The spot beams are typically defined by the contours at 3 or 4 dB down from the peak gain at the center of the beam. While it is the edge gain requirement which needs to satisfy the communication link, the satellite antennas have to be sized for the maximum gain to achieve the 3-4 dB drops around the perimeter of the spot beam. Therefore it can be argued that the satellite antenna is typically oversized as lower gain levels within the spot beam would satisfy the link requirements. Having a larger sized antenna leads to higher cost. To achieve 3 dB higher gain levels, one needs to double the aperture size, which directly translates to cost.

Recently a new method has been presented for reducing the cost of satellite antennas by sizing the antennas for lower power levels than the conventional approach, while maintaining the edge gain requirements, (Kilic and Zaghloul, 2009). This approach relies on the partitioning of the spot beam into a number of sub-beams. The basic idea behind the sub-beam partitioning approach is to replace the spot beams with a number of smaller sub-beams with lower peak gain and same edge gain values. This translates to an increase in the total number of beams to be generated by the antenna. However, since the sub-beams have lower peak gain, the antenna size can be reduced significantly. The increase in the number of beams results in a more complex beamforming structure, but this can be handled digitally without significant cost effects. Therefore the sub-beam partitioning results in significant reduction in the cost of the antenna compared to the spot beam approach.

Although the antenna cost is reduced in sub-beam approach a key issue which remains to be addressed is the interference effects in the new beam configuration. It is necessary to satisfy the carrier to interference ratio throughout the coverage area for this approach to be of practical value.

The objective of this section is to investigate interference performance of the sub-beam approach in comparison to the spot beam configuration. First the definition of the spot beam concept is presented and the sources of interference in a multiple beam communication system are identified. Then the sub-beam concept is introduced. Finally, the performance of the conventional spot beam approach with the sub-beam partitioning technique is compared in terms of co-channel beam interference.

3.1 Partitioning of spot beams into sub-beams

The sub-beam design reduces the peak gain requirements at the beam center while maintaining the edge gain requirements. In order to do this, each spot beam in the coverage area is divided into a number of smaller beams defined by contour levels less than the typical 3 or 4 dB, as depicted in Figure 17, where four sub-beams are used to replace a spot beam. For an edge gain requirement of 36 dB, it is observed that the sub-beam approach

provides a more uniform gain distribution within a beam's coverage area (36-37 dB) compared to the spot beam approach (36-40 dB). A group of these smaller sub-beams, i.e. "sub-beam clusters," represent each spot beam.

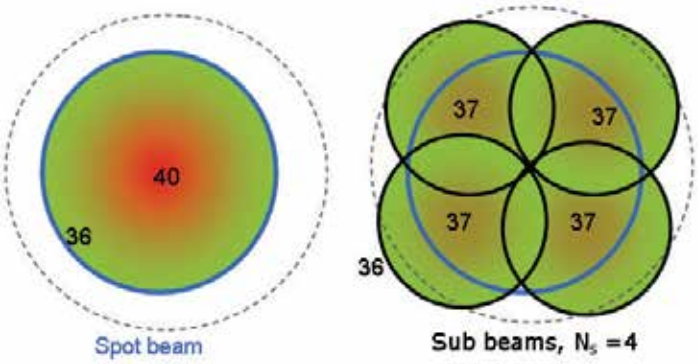


Fig. 17. Sub-beam representation of a spot beam

To satisfy the contiguity requirements of the coverage, the sub-beams intersect at the boundaries of the spot beam. Figure 18 demonstrates the sub-beam concept, where the hexagons correspond to the spot beams in the coverage area. The numbers at the center of each hexagon denote the frequency band assigned in each beam. A frequency reuse index (or spot beam cluster size) of seven is depicted in the figure. The large circles encompassing sets of seven hexagons represent the spot-beam clusters based on this reuse factor. The sub-beams are shown for the center spot beam of each cluster. A sub-beam cluster size of four is assumed in the demonstration.

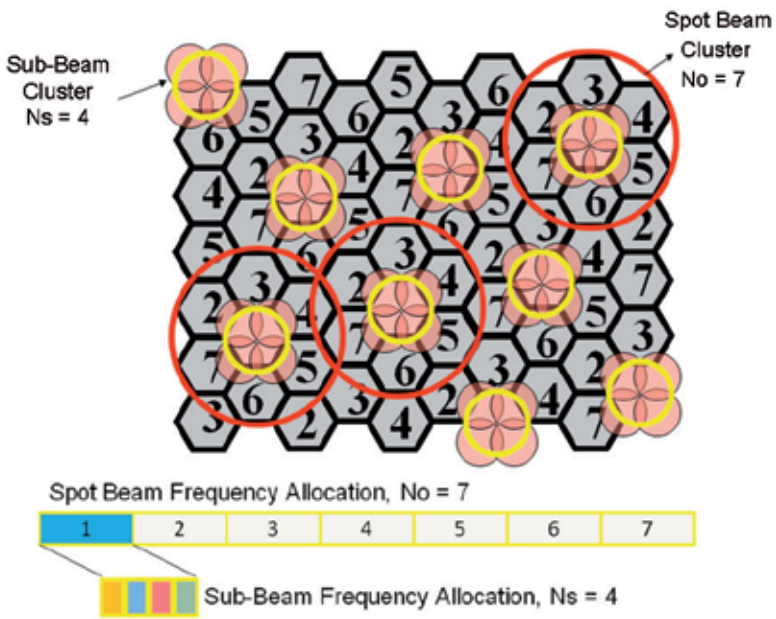


Fig. 18. Sub-beam partitioning for spot beams, reuse =7

As a result of this approach, the antenna is required to generate four times more beams than the conventional spot beam design. However, the sub-beams do not need to satisfy the peak gain requirements of the spot beam. Since only the edge gain needs to be satisfied, the sub-beam peak gain is lower and the overall antenna size is reduced as explained in the following sections.

3.2 Spectrum Subdivision with the Sub-beams

The frequency reuse configuration is conserved with the sub-beam approach. The allocated bandwidth per spot beam is shared among the sub-beams in a sub-beam cluster. As Figure 18 suggests, each sub-beam in a sub-beam cluster uses the bandwidth allocated to the spot beam they represent. While each sub-beam is designated a smaller portions of the bandwidth, the bandwidth allocation for the entire spot beam remains the same. Hence, the capacity per spot beam is conserved. Furthermore, this approach makes it possible to have a non-uniform distribution of the available bandwidth in a spot beam to address local high traffic areas. Depending on the sub-beam cluster size, some areas are served by more than one sub-beam from different spot beams, adding to the flexibility of traffic assignment.

3.3 Gain Reduction Concept of the Sub-beam Approach

The amount of reduction in the antenna size depends on the edge gain relative to the peak, and beam width requirements of the coverage. The sub-beam cluster generates an equivalent coverage to the spot beam by using a higher number of smaller beams. This enables using smaller antennas. The number of beams in a sub-beam cluster needs to be chosen such that a contiguous coverage is achieved. This results in values such as 3, 4, 7 …etc.

The relationship between the relative edge gain of the spot beam and the relative edge gain of the sub-beams can be determined as a function of the number of sub-beams that make up a sub-beam cluster. In the derivations that follow, the subscripts ‘o’, and ‘s’ denote the spot beam and sub-beams, respectively. The sub-beam edge contours are defined by x_s dB down relative to the peak gain of G_s dB. The spot beam contours are assumed to be x_0 dB lower than the peak gain, which is denoted by G_0 dB. Since the edge gain requirements are satisfied for both cases, the gain relationship can be written as follows:

$$G_o - x_o = G_s - x_s \tag{9}$$

This equation can be rewritten in terms of gain reduction as:

$$\Delta G \equiv G_o - G_s = x_o - x_s \tag{10}$$

It is observed that the reduction in gain is directly proportional to the difference of the beam contour levels of the spot beam and the sub-beams. While the typical contour level to define a spot beam is 3-4 dB, the sub-beam configuration is flexible, and the choice is based on increasing the gain reduction. As equation (10) suggests, defining the sub-beams at a low level would increase the reduction in gain. However, the sub-beam cluster size, Ns, constrains the beam width of the sub-beams, and the value of x_s cannot be assigned

arbitrarily. The choice will depend on the edge gain of the spot beam as well as the partition size, N_s.

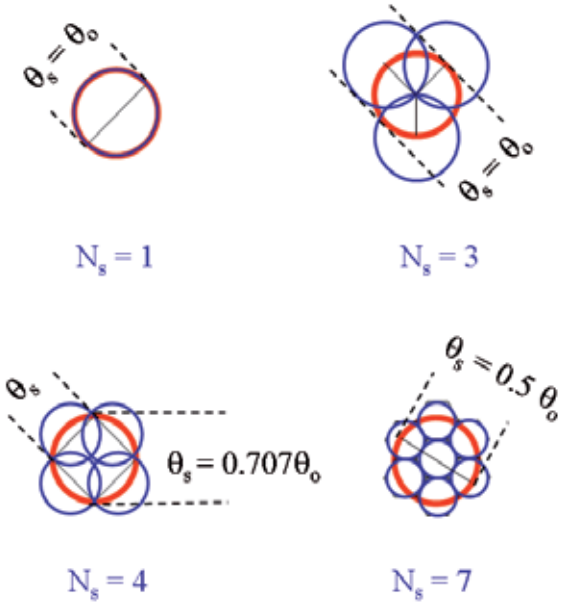


Fig. 19. Possible Sub-beam Partitions for Contiguous Coverage

3.4 Defining the Sub-beams to Achieve Desired Gain Reduction

A contiguous coverage of the sub-beam clusters can be obtained by satisfying Equation 1, in which case the number of sub-beams that define a spot beam are restricted to values such as 1, 3, 4, 7, …, etc. Examples of the geometry for different cluster sizes are shown in Figure 19, where the circle at the center corresponds to the spot beam and the surrounding circles denote the sub-beam contours. The beam width ratio, θ_s/θ_o depends on the sub-beam cluster geometry, and is equal to 1, 1, 0.707 and 0.5 for sub-beam cluster sizes of 1, 3, 4 and 7, respectively. The higher sub-beam clusters approximate the spot beam better and reduce the overlap of sub-beams from different clusters. Higher cluster sizes also reduce the sub-beam peak gain that is needed to satisfy the spot beam edge gain requirement.

An empirical relationship has been obtained to relate the beam width of a phased array at an arbitrary contour level, x to its HPBW by using the definition of the array factor as:

$$\frac{\theta_X}{\theta_3} = 0.59x^{0.4806} \tag{12}$$

where x is in dBs, (Zaghloul , et. al. 2000). This empirical relationship assumes isotropic radiating elements, which combine in space to form the directive pattern, and has been verified for large array sizes for broadside radiation. The expression is determined by fitting a curve to different size square arrays that range from 20x20 to 50x50 elements with half wavelength spacing between them. The results of the simulations and the agreement with the empirical relation above are shown in Figure 20.

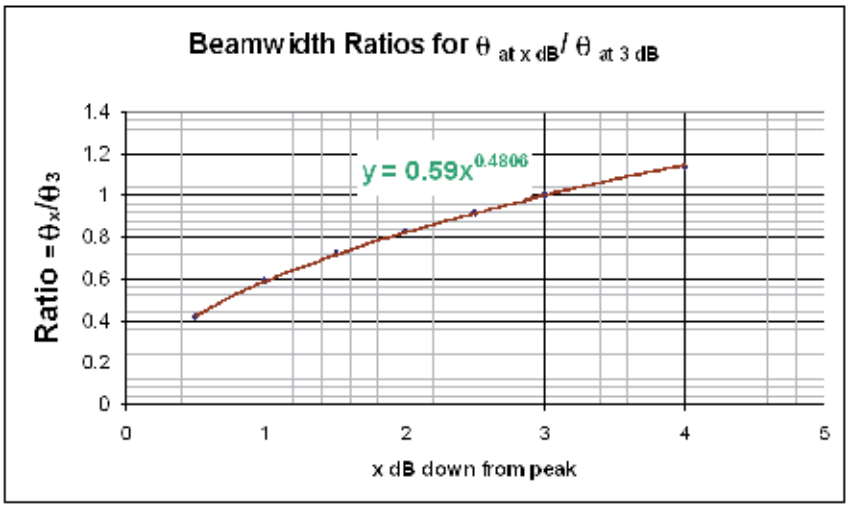


Fig. 20. Beam width at arbitrary contour levels for large arrays

Using equation (12) in (2), the contour levels of the spot and sub-beams can be related to their beam widths, and as follows:

$$9.612\log\left(\frac{x_s}{x_o}\right)+x_o-x_s=20\log\left(\frac{\theta_s}{\theta_o}\right) \tag{13}$$

Equation 13 can be rewritten in terms of gain reduction, using (10) as

$$9.612\log\left(\frac{x_s}{x_o}\right)+\Delta G=20\log\left(\frac{\theta_s}{\theta_o}\right) \tag{14}$$

where $\Delta G = G_o - G_s$ is the gain reduction achieved by using the sub-beam concept and the beam width ratio on the right is a constant determined by the sub-beam cluster configuration. Thus, for a given sub-beam configuration, the peak gain reduction and the ratio of the relative edge gains are uniquely related. In other words, if two of the parameters ΔG, x_c and x_s are given, the third is uniquely determined for a specific sub-beam cluster configuration.

The relation between reduction in gain and beam contour levels is demonstrated in Figure 21. Two cases are plotted for comparison. The blue line corresponds to the sub-beam cluster size of 4, while the red line represents the sub-beam cluster size of 7. The dashed lines show the contour level for the sub-beams as a function of gain reduction, and the solid line describes the relation between spot-beam contour levels and the gain reduction. It is observed that, for a desired amount of gain reduction, the beam contour levels for both sub-beam and spot beam are uniquely defined for each sub-beam cluster size.

Since the sub-beam approach is offered as a solution to reduce the antenna aperture size, it is worth investigating the improvement by assuming that the spot beam contour levels and the peak gain are already defined for the system to be replaced. For instance, if a spot beam width at 4 dB contour levels is being replaced with sub-beams of lower contour levels, the

graph in Figure 21 suggests that a gain reduction of 3dB can be achieved for Ns = 4 by defining the sub-beam contours at 1 dB down from the sub-beam peak gain. Similarly, for Ns = 7, a gain reduction of 3.6 dB is achieved by defining the sub-beams at 0.4 dB down from their peak gain.

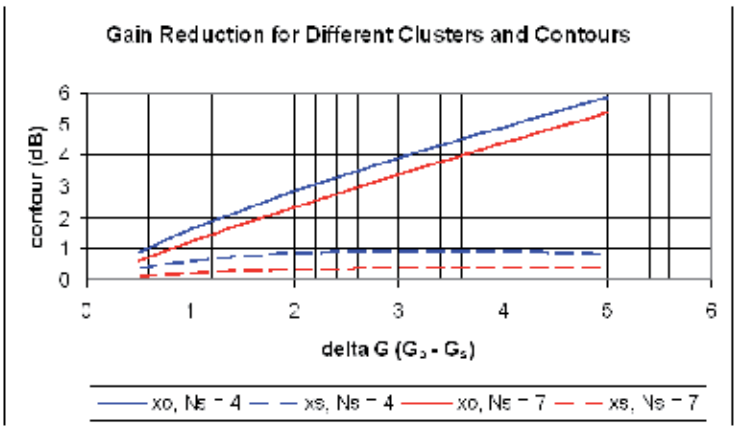


Fig. 21. Contour levels of spot beams and sub-beams as a function of gain reduction.

This example implies that for a given spot beam configuration, the larger sub-beam clusters can offer higher gain reduction, which is associated with a corresponding reduction in the antenna aperture size and cost. However, since the higher cluster sizes define beams at lower contour levels, a legitimate concern is to make sure the energy in the main beam of these antennas do not adversely affect the co-channel interference levels.

The reduction in peak gain results in a reduction in the aperture size and correspondingly a reduction in the number of elements in the array. This is illustrated in Figure 22 where a 28 x 28 or 784-element array for a conventional spot beam (corresponds to Ns =1) is replaced with arrays of fewer elements as the number of sub-beams increase in the cluster. It is observed that the limit case is approached for sub-beam cluster sizes of 7, where less than 300 elements are required.

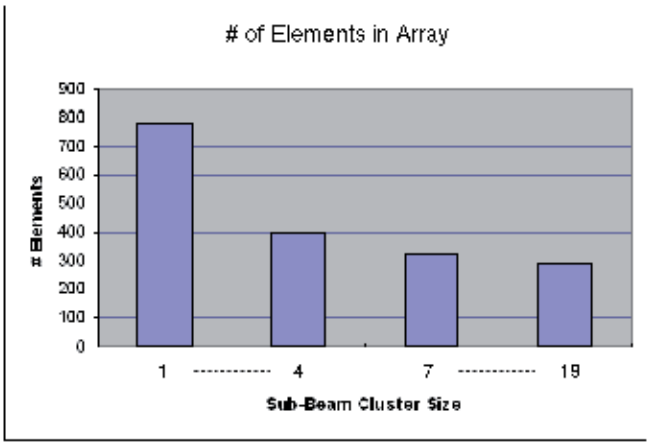


Fig. 22. Array size reduction with sub-beam partitioning

The reduction in the receive and transmit antenna aperture sizes result in fewer components. However, the generation of sub-beam clusters requires further de-multiplexing and increase in the size of the digital traffic router. Corresponding multiplexing on the transmit side is also required. However, the added complexity occurs in the software-driven processor on board the satellite, with minimal addition to the hardware.

3.5 Interference in Sub-beam Partitioned Coverage

The same LEO system as in section 2.5 is considered for the interference analysis of the sub-beam partitioned coverage. A comparison of beam coverage on earth for spot beam and sub-beam configuration with partition size of 3 is shown in Figure 23. The beams are overlaid on the contour plot of the antenna pattern that generates the center beam, the beam of interest for interference calculations.

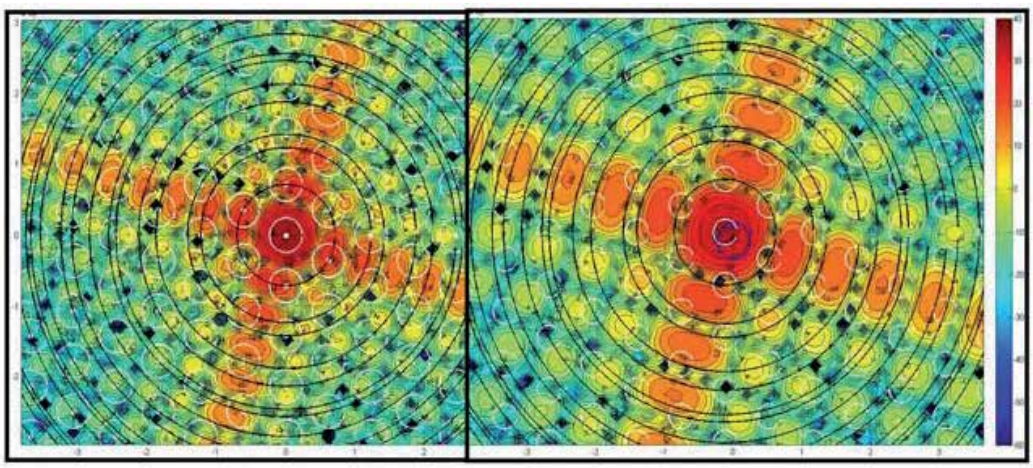


Fig. 23. Coverage on Earth – Spot beam pattern (left) versus Sub-beam pattern (right)

Sub-beam partition sizes of four and seven will be considered for the interference analysis. Figure 24 shows a zoomed in depiction of which one of the sub-beams that replace a spot beam are used for interference calculations. The spot beam to be replaced is the blue circle, while the red circles are the sub-beams for a partition size of four (left) and seven (right). The green circle is the sub-beam that will be selected in the interference analysis.

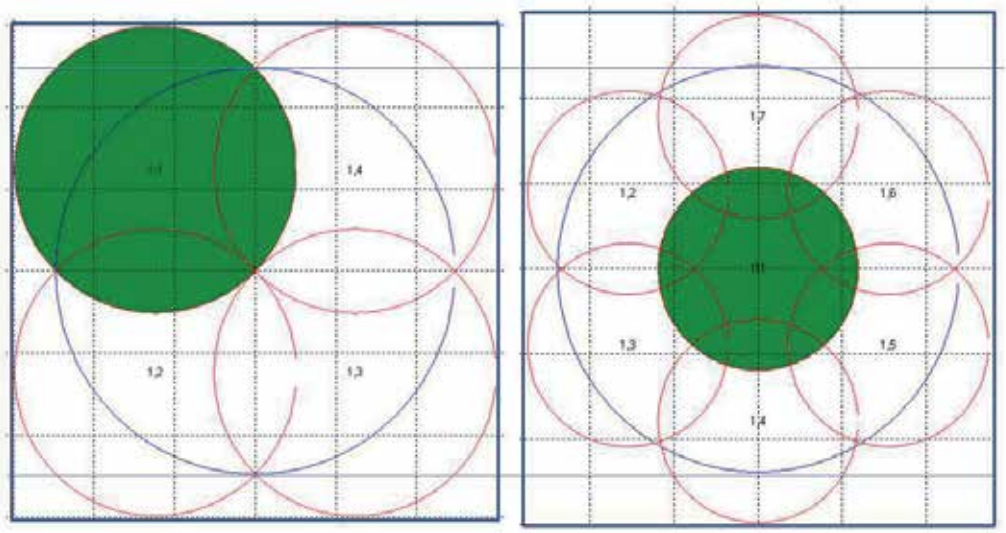


Fig. 24. Sub-beam selection for interference calcualtions, Ns = 4

The carrier to interference ratio and total interference are shown for frequency reuse of 3, 4, 7, 9 and 13 in Figure 25 for sub-beam partition sizes of four and seven. It is observed that the C/I at the center of the beam (where carrier power is maximum) are better for the spot beam configuration for large cluster sizes (i.e. N > 9). However in certain cases such as No = 3 and No = 9, the performance of the sub-beam approach is better or comparable to the spot beam configuration. It should be noted that this comparison presents a best case scenario for the spot beams as the carrier power is at its peak value. The total interference shown in Figure 25(b) indicates that sub-beams may have lower interference values relative to the spot beam cases of the same frequency re-use number. A large cluster size separates co-channel cells, and the highly directive spot beam antenna tends to have lower sidelobe levels at these separations.

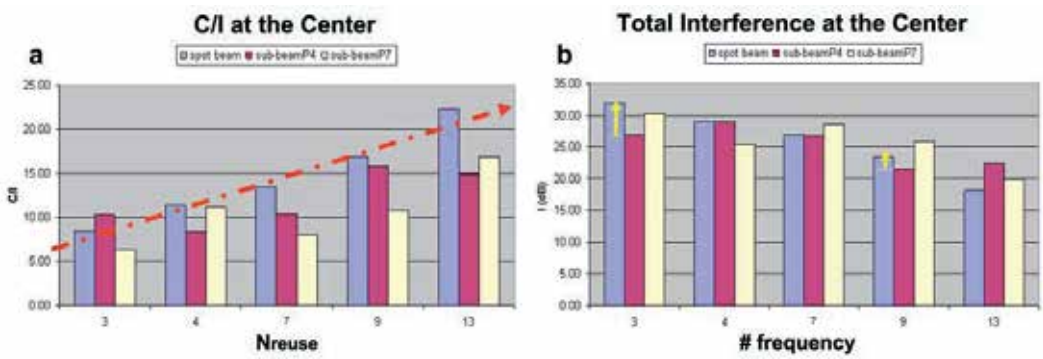


Fig. 25. Interference at the center by tier for sub-bem partition=4, 7 and frequency reuse 3, 4, 7, 9, 13. (a) C/I (b) Total interference

To better understand the sub-beam performance, a better analysis is calculating C/I as one moves within the spot beam to be replaced by the sub-beams as depicted in Figure 26. The

spot beam shown at the center with the blue circle is replaced by 4 sub-beams. The C/I is calculated at discrete distances from the center (d) and at angular positions (α). The chart on the right shows the C/I for the spot beams (blue lines) and sub-beams (red lines) for different d and α values. The x-axis shows the ratio of the distance from the center to the spot beam radius, R.

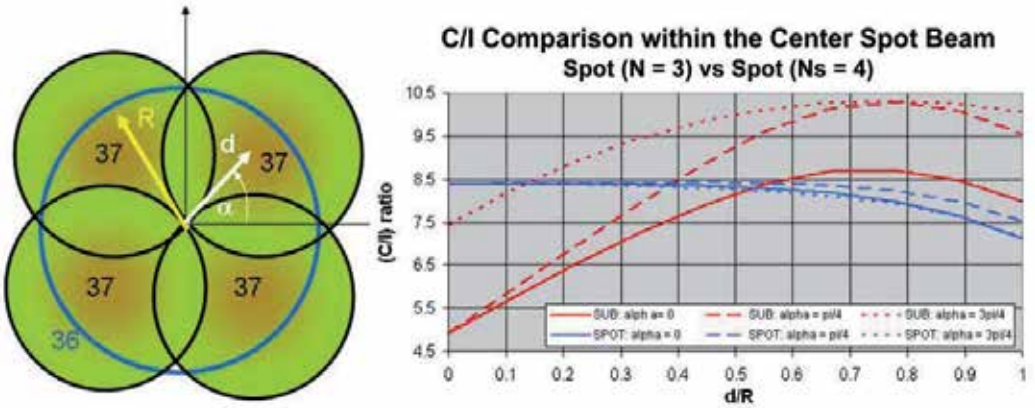


Fig. 26. C/I performance comparison within the spot beam

It is observed that partitioning the spot beam into sub-beams performs superior towards the spot beam edge because the carrier power in the sub-beam gets stronger. This is typically where the spot beam performs worst and a significant advantage for the sub-beam approach results. The main region where the spot beam seems to perform better is at the center of the beam as expected due to the high carrier power levels in the spot beam. In the regions near the beam center, it is observed that the sub-beam approach performs worst where the spot beams have the best performance. As expected, this is due to the high carrier power the spot beam has at the center. Typically the C/I value at the beam center has a higher margin as the systems are designed for the edge gain and the degraded performance of the sub-beam at the center may not pose a problem. To mitigate the relatively poor performance at the center, it would be possible to optimize the sub-beam antennas to perform better at the beam centers. If necessary a further step could be to modify the frequency allocation. However, this could result in lower system capacity, and should be attempted only if the antenna size and cost is the primary concern. Overall, the findings suggest that the sub-beam approach holds promise as its C/I is mostly not degraded compared to the spot beam approach, while its peak gain is lower and thus requires smaller antenna aperture.

4. Conclusion

Co-channel beam interference in multi-beam satellite communications systems was investigated particularly for the downlink. Concept of frequency reuse was explained and the role of satellite antenna size and pattern was examined. Conventional spot beam coverage and its impact on determining the antenna size on board was discussed.

A method was presented to reduce the array antenna aperture on board the satellite for multiple-spot-beam cellular coverage. The overlapped spot beams have edge gain that is determined based on link budget, coverage requirement, frequency band allocations, and corresponding channel capacity allocations. The edge gain relative to the peak gain of the spot beam determines the size of the required antenna aperture. Dividing the spot beam into a number of sub-beams that overlap at the same gain level as the spot beam but with lower gain taper over the sub-beam results in lower peak gain and consequently smaller aperture. The frequency band allocated to the spot beam is divided among the constituent sub-beams and can be assigned with enough flexibility to meet projected traffic demands.

The reduction in the antenna aperture with this approach translates into significant reductions in number of array elements, RF components, and A/D and D/A converters. These savings are obtained at the expense of more complex digital beam forming and traffic routing algorithms. Analysis has shown that in spite of the smaller aperture and the broader beams of the sub-beams, the co-channel interference between sub-beams using the same frequency segment is not adversely affected.

5. References

Balanis C.A., (2005) Antenna Theory: Analysis and Design, John Wiley and Sons, New York, 3rd Edition, ISBN 0471592684.

Kilic, O., and A. I. Zaghloul (2009), Antenna aperture size reduction using subbeam concept in multiple spot beam cellular satellite systems, *Radio Sci.*, 44, RS3001, doi:10.1029/2008RS004052

Mehrotra, (1994) A. *Cellular Radio Performance Engineering*, Artech House, Boston, ISBN 0890067481.

Stutzman, W.L. and Thiele, G.A., (1998) *Antenna Theory and Design*, John Wiley and Sons, New York, 2nd Edition, ISBN 0471025909.

Zaghloul, A.I., O. Kilic and A.E. Williams(2000), "Aperture Size Reduction Using Sub-Beam Design Concepts for Multiple Spot Beam Satellites," *IEEE International Symposium on Antennas and Propagation*, Salt Lake City, Utah, July 2000, pp. 534-537.

16

Beyond life-cycle utilization of geostationary communication satellites in end-of-life

Shi Hu-Li, Han Yan-Ben, Ma Li-Hua, Pei Jun, Yin Zhi-Qiang and Ji Hai-Fu
National Astronomical Observatories, Chinese Academy of Sciences, 20A Datun Road, Chaoyang District, Beijing 100012 China

1. Introduction

In 2002, we proposed an idea that carrier signals with ranging codes and navigation data from ground navigation master station are transmitted in transponders of geostationary (GEO) communication satellites and then ground users receive those signals and realize passive navigation. In 2003 we began to lease transponders of GEO communication satellites to carry out experimental verification, and a success was achieved in 2005[11, 12].

The most challenging problem in the verification is that all rental GEO communication satellites are located in GEO orbit and therefore three dimensional positioning is impossible. To improve geometric configuration of satellites constellation and positioning precision, we hope that several inclined geosynchronous orbit (IGSO) satellites can be adopted except for GEO satellites to constitute the constellation. Considering the long designing period of IGSO satellite, a concept that we try to utilize IGSO satellites was proposed. Unfortunately, those efforts are in vain, and therefore we proposed the concept turning GEO communication satellites in end-of-life to IGSO by regulation.

Communication satellites running in GEO orbit with altitude of 35786 km above equator are maintained in fixed track with up to about ±0.1 degree precision and some even with better than about ±0.05 degree precision. However, because of the perturbation from other celestial bodies, the position of GEO satellites are changing relative to the Earth, and the drift of satellite's fixed position relative to the equator are constantly increasing and even exceeds more than the ITU fixed-point drift value. In order to control the satellite's drift, we must adjust the satellite's orbital position and attitude, which will inevitably consume fuel, when fuel in the satellite is consuming step by step, the satellite's service life will come to end. We call such a geostationary communications satellite "late-life" or "end-life" GEO satellite. In fact, the life of satellite's solar battery and signal transponders has a certain redundancy and it can still work for many years, just because the remaining fuel in satellite has almost run out, this satellite with heavily investment in the process of development and launch has to been abandoned. To ensure custom's business is not interrupted, satellites company must launch a new satellite to replace the GEO communication satellite in end-of-life. According to the IADC's suggestion on retiring GEO communication satellites, the satellite solar power

and related equipment have to been closed and satellite's must been pulled up about 200 ~ 300 km to the grave orbit and then those retired satellites become space junk[14, 15].
According to IADC's monitoring, there have been approximately more than a thousand satellites in the vicinity of GEO orbit within about 60 km, but only about 300 satellites work normally[16]. Increasing space junks seriously affect the space environment, especially the valuable space environment near the equatorial orbit. How to extend the life of the satellite in orbit and thereby turning waste into wealth has became one of important topics for space technology researchers.
To this end, according to concept of modern design and advanced manufacturing technology, we explored the issue of reusing satellite in end-of-life[1, 9, 13]. Specifically, with use of remaining fuel resources in retired satellites in end-of-life, gradually turning GEO satellite in end-of-life to slightly inclined geostationary orbit (SIGSO) satellite by control of satellite attitude. Utilizing those SIGSO satellites to constitute a new navigation and positioning constellation is for the purpose of improving the navigation and positioning accuracy of the transmitting satellite navigation system. In addition, the abundant transponders resources in those satellite can also been used to carry out satellite communication business or open up new satellite communication services, that is to say, it is possible to develop a second life-cycle of the satellite in end-of-life.

2. Orbital dynamic analysis of GEO communication satellite orbit

In order to explore the issue of reusing the GEO satellite in end-of life, we must thoroughly understand and analyze the dynamic theory of GEO satellite[34].

2.1 Description of GEO orbit

GEO is a special orbit, on which the satellites rotating around the Earth has the same rotation cycle with Earth's rotation period and its nadir position in the Earth's surface should be fixed. Generally speaking, GEO has the following three characteristics:

(1) the shape of orbit is circular, eccentricity e = 0;
(2) orbit is located in the Earth equatorial plane, the inclination i = 0;
(3) orbit's spin cycle is same with the earth rotation cycle, $T = 2\pi / \omega_e$ = 23 hours 56 minutes 4.1 seconds; the orbit altitude is 35786 km, or semi-major axis a = 42164km, where ω_e is for the Earth's rotation angular velocity.

However, due to the Earth's non-spherical gravity, moon and sun gravitational perturbation and solar light pressure, satellite orbit elements (including a, e, i) are constantly changing and deviating from the fixed location of the satellite.

2.2 The relationship between deviation of satellite orbital elements and drift

Semi-major axis deviation causes long-term east-west direction drift, At t time, sub-satellite point longitude λ is:

$$\lambda = \arctan(\cos i \tan u) + \Omega - \omega_e (t - t_0) - s_{g0} \tag{2.1}$$

Where, s_{g0} is GMT at t_0 time; u is satellite longitude argument, $u = \omega + f$, namely the angle between satellite and ascending node; is perigee argument, namely the angle between orbit ascending node and perigee; ω is true anomaly, that is, angle between satellite and perigee; Ω is RAAN, namely angle between ascending node and the equinox.
When $i = 0$, $e = 0$, we have $f = M$, at t time, sub-satellite point longitude is:

$$\lambda = \omega + M + \Omega - \omega_e(t - t_0) - s_{g0} = \omega + \mathrm{M}_0 + n_e(t - t_0) + \Omega - \omega_e(t - t_0) - s_{g0} \quad (2.2)$$

In initial time t_0, satellite longitude λ_0 is

$$\lambda_0 = \omega + \mathrm{M}_0 + \Omega - s_{g0} \quad (2.3)$$

And then can been simplified as

$$\lambda = n_e(t - t_0) - \omega_e(t - t_0) + \lambda_0 \quad (2.4)$$

When semi-major axis a has no error, we have

$$n = \omega_e \quad (2.5)$$

$$\lambda \equiv \lambda_0 \quad (2.6)$$

And when semi-major axis has error, there is

$$\Delta n = -\frac{3}{2} n_e \frac{\Delta a}{a} \quad (2.7)$$

Therefore, when a has deviation Δa , satellite drift in east-west direction is

$$\dot{\lambda} = -0.012806 \cdot \Delta a \ (^\circ/\mathrm{d}ay) \quad (2.8)$$

When $\Delta a > 0$, $\dot{\lambda} < 0$, satellite drifts to west; when $\Delta a < 0$, $\dot{\lambda} > 0$ satellite drifts to east.

2.3 Oscillation in east-west direction caused by eccentricity deviation

The instantaneous sub-satellite point longitude can be denoted as:

$$\lambda = \lambda_0 + 2e \cdot \sin M \quad (2.9)$$

where λ_0 is central point longitude, M is mean anomaly; when $e \neq 0$, satellite swings periodically around the center of λ_0, the largest Oscillation argument is $2e$(rad), Oscillation cycle is equal to satellite orbit period.

2.4 Oscillation in north-south and east-west direction caused by Orbit inclination deviation

When $\Delta a = 0, e = 0$, if satellite orbital plane deviates from the earth equator, namely there is deviation Δi.

Satellite geocentric longitude is

$$\sin\phi = \sin\Delta i \sin u \tag{2.10}$$

While is Δi *is* small, then

$$\phi = \Delta i \sin u \tag{2.11}$$

The former equation indicates that when $\Delta i \neq 0$, satellite swings in latitude direction with the period of one day, and the argument in south-north direction is i. u is latitudinal argument, with the range from 0° to 360° in one day. When inclination is not zero, satellite will has oscillation both in south-north direction and east-west direction, and now sub-satellite point longitude is

$$\lambda = \arctan(\tan u) - \frac{i^2}{4}\sin 2u + \Omega - \omega_e(t - t_0) - s_{g0} = \lambda_0 - \frac{i^2}{4}\sin 2u \tag{2.12}$$

The former formulation denotes: when the inclination is not zero, satellite has oscillation with a half-day period in longitude direction, and argument of oscillation is $i^2/4$, the track of sub-satellite point has "8" shape which is shown in figure 2.1.

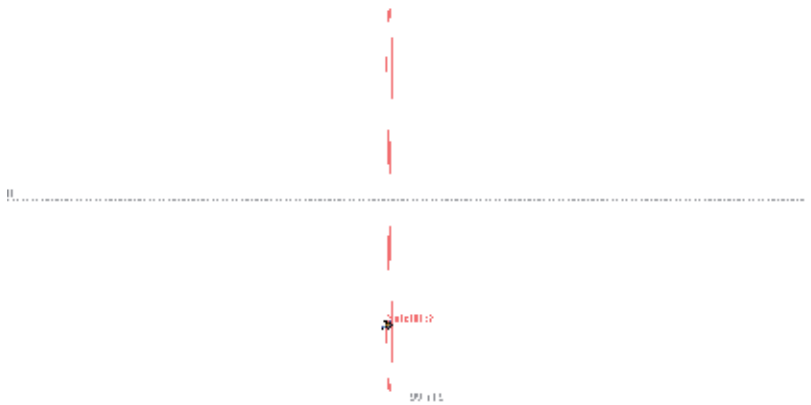


Fig. 2.1 Diagram of sub-satellite point "8" shape due to the change of GEO satellite orbital inclination

Due to various perturbations, GEO satellite position relative to Earth is not absolutely static, in the actual project implementation; it needs to continually and periodically modify its orbit[34, 35], restricting its motion in a small area around its specified location. Figure 2.2 shows that of Canada Anik-C3 satellite still has typical latitude and longitude changes after position maintenance[7].

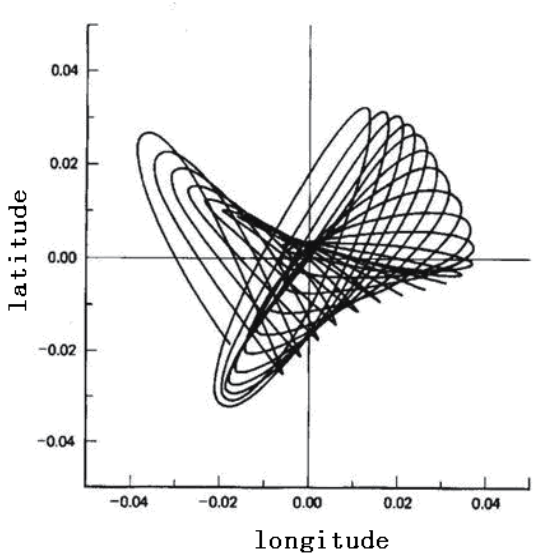


Fig. 2.2 Typical track of GEO satellite's sub-satellite point

2.5 GEO satellite orbit perturbations

GEO satellites are subject to various perturbations, including: additional gravitational force caused by non-spherical shape and the uneven quality of the earth, the sun and the moon's gravity and solar radiation pressure. Under the perturbations, satellite orbit no longer follows the two-body orbital motion law, the period, eccentricity, RAAN and inclination are constantly changing, and thereby causing satellite orbit drift. The details of these perturbations are as follows:

1) Earth's non-spherical gravitational perturbation
   Earth's non-spherical gravity will mainly cause changes in sub-satellite point longitude; especially Earth's equatorial ellipticity has a long-term accelerating effect on sub-satellite point longitude. The difference between semi-major axis and semi-minor axis is about 68 meters, the short axis pointed at west longitude 75 degrees and west longitude 105 degrees, the long axis point and short point axis has a 90 degrees in longitude direction. As in the geostationary orbit, semi-minor axis point is stable and the semi-major axis point is not, so the equatorial ellipticity will bring about perturbation which will cause the satellite drift back and forth near the center of semi-minor axis point in east-west direction.

Considering the first 4-order harmonic field perturbation primarily, the average longitude perturbing acceleration can be described as follows[39]:

$$\ddot{\bar{\lambda}} = 0.00168\sin 2(\lambda - \lambda_{22}) - 7.79\times10^{-5}\sin(\lambda - \lambda_{31}) + 2.34\times10^{-4}\sin 3(\lambda - \lambda_{33}) - 8.95\times10^{-6}\sin 2(\lambda - \lambda_{42}) + 1.14\times10^{-5}\sin 4(\lambda - \lambda_{44})°/d^2 \quad (2.13)$$

Where λ_{lm} is defined as phase longitude[40], data of λ_{lm} is presented in Table 2.1

Table 2.1 Data table[51] for λ_{lm}

λ_{22}	λ_{31}	λ_{33}	λ_{42}	λ_{44}
-14.91°	7.0°	21.1°	31.2°	30.65°

In the east longitude 0° ~ 180°, the average longitude perturbing acceleration caused by the Earth's gravitational perturbation is shown in Figure 2.3. As can be seen from the diagram, average longitude drift acceleration reaches the maximum value of about -0.00197° / day^2 near east longitude 118° and equals to zero in the vicinity of east longitude 75° and 165°.

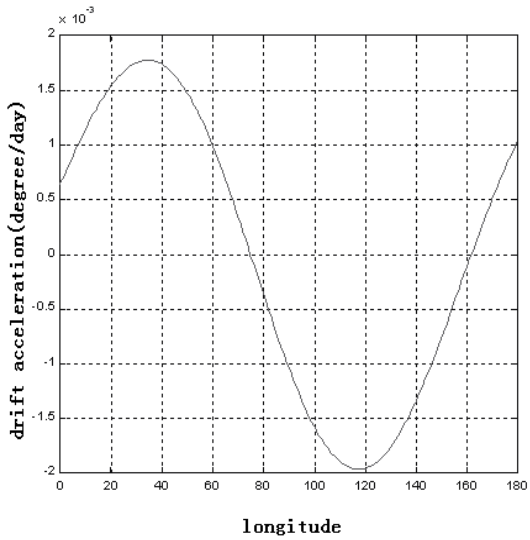


Fig. 2.3 Mean longitude perturbing acceleration caused by the Earth's gravitational perturbation

2) The sun light pressure perturbation
Sun light pressure perturbation mainly affects eccentricity of geostationary satellites orbit. Eccentricity vector components is defined as:

$$e_x = e\cos(\Omega+\omega) \quad (2.14)$$

$$e_y = e\sin(\Omega+\omega) \quad (2.15)$$

where Ω is RAAN, ω is perigee argument.

The trajectory of eccentricity vector Endpoint forms an ellipse with one-year cycle, its semi-major axis and semi-minor axis are as follows:

$$a_e = 0.0114 \frac{A}{m} \tag{2.16}$$

$$b_e = 0.0105 \frac{A}{m} \tag{2.17}$$

Where A is surface area of satellite, m is satellite quality. The eccentricity vector caused by visible sun light pressure is mainly depends on the ratio of surface area to mass. Figure 2.4 shows the changes of eccentricity vector curve caused by various factors in one year. The value of eccentricity equals the distance between the point in the figure and the origin, the size of large circular curve mainly caused by the pressure of sunlight, large wavy lines caused by the lunar gravity, and small zigzag line is derived from several factors such as the Earth's gravity, solar gravity, and lunar gravity.

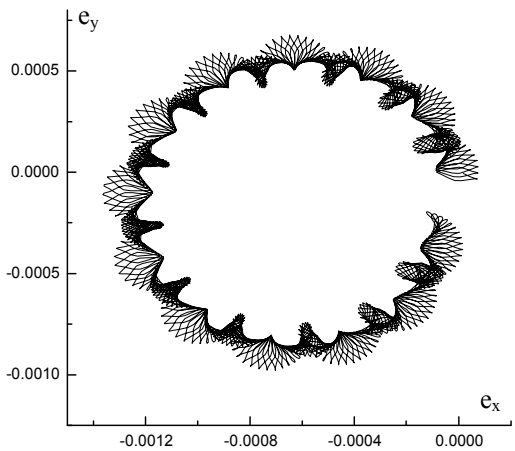


Fig. 2.4 Various perturbations' affects on eccentricity in one year

3) Sun and Moon gravitational perturbation
Sun and Moon gravitational perturbation is most important factor causing the changes of GEO orbit inclination. Orbital inclination vector (or orbital polar) is defined as the unit vector for the normal direction of orbital plane. Under the combined effect of gravity field J_2 items of the sun, the lunar and the earth (the earth with harmonic coefficients), the normal direction has gradually deviated from the Arctic. Orbital normal direction generally rotates around a certain direction which is in the Arctic pole plane and ecliptic pole plane, between the ecliptic pole and the Arctic pole, and has an angle of about 7.5° from Arctic pole. With a 52 years cycle, orbital inclination reached its maximum value of about 15° in about 26 years and back to zero 52 years later which means normal direction meets Arctic Pole again.

As the lunar orbit has an angle between 23.5° ± 5.14° relative to equatorial plane, the drift of satellite's inclination is changing every year, the specific value of which depends on the 18.6-year -cycle average value of lunar orbit inclination. Drift rate ranges from 0.7°/ year to 0.95° / years and its direction varies in different years as the lunar orbit pole varies. The most maximum value of drift occurred in 2006, the lunar orbit inclination reached maximum value at the same time.

Figure 2.5 is two-dimensional inclination vector diagram for the satellite, the moon and the sun. As can be seen from the figure, the angle between terrestrial pole and the ecliptic pole is 23.5°, the angle between the lunar orbital pole and ecliptic pole was 5.14°. Lunar orbital plane is not fixed, its orbital pole regress around ecliptic pole with a period of 18.6 years. If we don't maintain satellite position in north-south direction, the satellite's orbit rotates around the pole which has a 7.5° angle with terrestrial pole and a period of about 52 years. If the initial orbit inclination satellite is 0°, the orbital inclination will reach the maximum value of about 15° and then gradually decreasing with year.

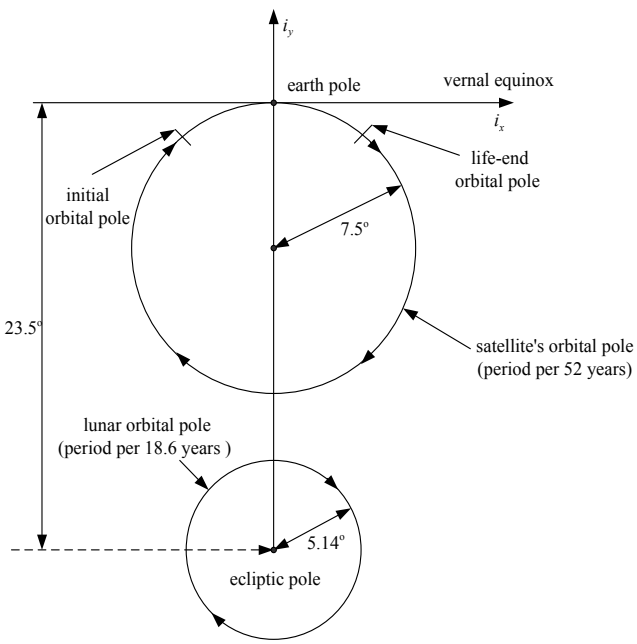


Fig. 2.5 Two dimensional diagram of GEO satellite orbit, moon's path and ecliptic

The effects on inclination vector caused by solar gravity, lunar gravity and the Earth's non-spherical gravitational perturbations vector can been defined as

$$i_x = \sin i \cos\Omega \tag{2.18}$$

$$i_y = \sin i \sin\Omega \tag{2.19}$$

Lagrangian perturbation equation defined with i is as follows[40]:

$$\frac{di_x}{dt} = -n_e(5.21 i_y - 0.099 \sin \Omega_m) \times 10^{-5} \ rad/s$$
$$\frac{di_y}{dt} = n_e(-4.93 \ i_x + 0.074 \cos \Omega_m + 0.648) \times 10^{-5} \ rad/s \tag{2.20}$$

Where n_e is the average angular velocity of the earth's rotation, namely:

$$n_e = \sqrt{\frac{G m_e}{r_e^3}} \tag{2.21}$$

Here Ω_m is ecliptic longitude of ascending node of lunar orbit. Based on AGI's STK, the following Table 2.2 and Table 2.3 gives the changes of i and Ω for the perturbed geostationary orbit in the situation of different RAAN, with the initial inclination 0.1° and 0.5° .

Table 2.2 Changes of orbit in varied RAAN (*Ω*) with initial inclination 0.1°

Ω		3 monthes	6 monthes	1 years	2 years	3 years	4 years
0°	*i*	0.266	0.58	1.04	1.89	2.82	3.79
	Ω	80.45	82.35	84.6	82.86	79.4	75.1
45 °	*i*	0.334	0.65	1.02	1.97	2.91	3.86
	Ω	87.07	85.36	86.3	83.54	79.6	75.26
90 °	*i*	0.368	0.678	1.05	2.01	2.96	3.90
	Ω	98.242	91.389	90.05	85.5	80.86	76.2
135 °	*i*	0.359	0.66	1.03	1.99	2.93	3.89
	Ω	110.15	97.73	94.12	87.60	82.35	77.36
180 °	*i*	0.31	0.60	0.97	1.93	2.89	3.83
	Ω	120.14	101.82	96.58	88.88	83.12	78.03
225 °	*i*	0.24	0.52	0.90	1.86	2.82	3.75
	Ω	123.24	100.88	95.29	88.46	82.92	77.86
270 °	*i*	0.18	0.48	0.85	1.81	2.39	3.71
	Ω	109.77	93.61	91.72	86.44	83.56	76.93
315 °	*i*	0.19	0.50	0.88	1.83	2.78	3.72
	Ω	86.25	85.15	86.76	84.13	80.08	75.77

Table 2.3 Changes of orbit in varied RAAN (Ω) with initial inclination 0.5°

Ω		3 months	6 months	1 years	2 years	3 years	4 years
0°	*i*	0.510	0.656	1.020	1.900	2.816	3.736
	Ω	29.48	43.09	61.70	70.81	71.149	69.070
45°	*i*	0.682	0.882	1.321	2.249	3.176	4.101
	Ω	63.19	66.41	73.314	75.52	73.618	70.536
90°	*i*	0.766	0.975	1.449	2.408	3.363	4.294
	Ω	93.168	89.463	88.211	83.986	79.487	74.990
135°	*i*	0.745	0.918	1.380	2.333	3.286	4.231
	Ω	122.480	112.562	103.53	93.111	86.022	80.131
180°	*i*	0.623	0.723	1.129	2.048	2.987	3.938
	Ω	153.281	135.878	116.831	99.877	90.641	83.681
225°	*i*	0.427	0.419	0.758	1.665	2.610	3.560
	Ω	190.385	160.25	122.407	100.05	90.39	83.450
270°	*i*	0.245	0.055	0.460	1.420	2.369	3.317
	Ω	253.177	199.55	99.09	89.273	83.618	78.605
315°	*i*	0.304	0.328	0.635	1.534	2.465	3.397
	Ω	341.197	18.153	62.185	74.94	74.757	72.170

As can be seen from the above results, for the geostationary orbit whose initial inclination is in the vicinity of 0°, the general trend is that its orbital inclination is to increase. However, in the different case of RAAN, there are still some slight differences in the changes of orbital inclination. Specifically, for Ω = 90°, orbital inclination will directly increase, and for Ω = 270°, the inclination will decrease first and then increase. Ω is to change in the direction of 90 ° with a 52-year period. to the specific, Ω will be 90° from the first beginning 0 ° after 26 years and 270° after 52 years (in 52 years, the changing law of Ω is 90° -> 0°, 360° -> 270°). This variation does not change with sub-satellite point longitude.

2.6 Orbit control scheme for GEO satellite

To maintain GEO satellites in the fixed position with the required accuracy, it is necessary to regulate the position of satellite's orbit periodically based on the characteristics of the orbit's change. In order to keep GEO satellites operating in the designated area, it is required to increase the velocity by jet to overcome various perturbation imposed on satellite. For GEO satellites, the velocity increment in north-south direction resulting from north-south position maintenance is to overcome satellite inclination's changes due to the sun and moon gravity, and while the velocity increment in north-south direction resulting from east-west position maintenance is to overcome the changes of eccentricity caused by solar light pressure and satellite's drift in east-west direction caused by harmonic perturbation of the Earth's gravitational field. East-west position maintenance for the GEO satellite is to change inclination by control of orbital normal direction in the vicinity of RAAN. Taking a longer time interval for East-west position maintenance into account, RAAN should be adjusted to

270 ° at the same time of controlling orbital inclination. Velocity increment for north-south position maintenance has a relative large value of about 42 m/s ~51m/s every year.
East-west position maintenance for the GEO satellite is mainly to overcome the long-term accelerated drift of average longitude caused by the tangential perturbation resulted from the Earth's non-spherical gravity. Through the control method with limit cycle[40], satellite can be restricted to do the long term limit cycle oscillatory movement near the fixed-point. Control mode of drift limit cycle will be introduced in the following via typical cycle of east-west amendments (as shown in Figure 2.6).

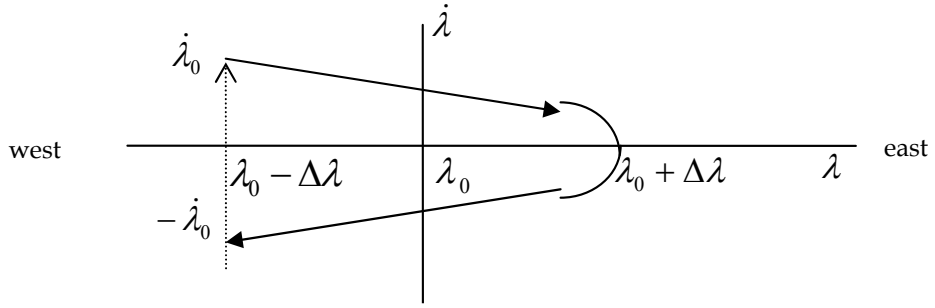


Fig. 2.6 Limit cycle of east-west position maintenance

Supposed λ_0 is sub-satellite point, $\Delta\lambda$ is the allowed longitude deviation, $\ddot{\lambda}$ is for the longitude drift acceleration which is a function of longitude (in China's space, $\ddot{\lambda} < 0$). If Satellite's longitude is λ_0 at time t_0, Satellite's longitude will be as followings at time t:

$$\lambda = \lambda_0 + \dot{\lambda}_0 t + \frac{1}{2}\ddot{\lambda} t^2 \tag{2.22}$$

where $\dot{\lambda}_0$ is initial drift velocity. Due to $\ddot{\lambda}$, satellite will has drift velocity. Specifically, satellite will have westward drift when $\dot{\lambda}_0 < 0$ and eastward drift when $\dot{\lambda}_0 > 0$. as in the vicinity of china area $\ddot{\lambda} < 0$, control is usually implemented in the western boundary of the limit cycle to give the satellites a long running time within the fixed-point accuracy (such as ± 0.1° the east-west direction), which will equip the satellite an eastward drift velocity . In the role of $\ddot{\lambda}$, the velocity of satellite will gradually slow down during the eastward drift until satellite drifts to the east boundary of limit cycle. At that time, the satellite will have westward drift velocity and then drift to west continuously by the effect of $\ddot{\lambda}$. When the satellite reaches the western boundary of limit cycle $\lambda_0 - \Delta\lambda$, it can be equipped with drift velocity $\dot{\lambda}_0$ by the control of executive organization and thereby drifts to the east again. A drift cycle has been formed after several repetition of former process.

Velocity increment for east-west position maintenance each year is generally less than 2 m/s, especially in the equilibrium point $\ddot{\lambda}$ will be close to zero and here the required velocity increment is also approaching zero.

2.7 issue of GEO orbital phase

For geostationary orbit with a certain inclination, the satellite phase we actually call refers to Greenwich sidereal time when satellite passes the RAAN. As can been seen from $\lambda = \omega + \mathrm{M} + \Omega - s_g$, if $\omega + \mathrm{M} = 0$, while satellite passing the RAAN, there will be, $\varphi = s_{g(\mathrm{M}+\omega=0)} = \Omega - \lambda$. Therefore satellite phase can been denoted as $\Omega - \lambda$, where λ is sub-satellite point longitude and Ω is RAAN. When there is no north-south position maintenance for GEO satellites, they generally have the same law for the changes of orbital inclination and RAAN. Therefore their phases more depend on the differences of fixed position. For the two satellites with 10^o difference in fixed position, their phase difference will be about 10^o. The change of two satellites' phase-difference can be seen from $\phi = \Omega - \lambda$ and realized by altering their Ω or λ.

3. Technical scheme on altering control mode of the satellite's attitude

In precious section, we have explored the characteristics of IGSO satellite orbit and its running law. In this section, technical scheme on altering control mode of satellite's attitude in end-of-life will be introduced.

3.1 Technical schemes on regulating GEO satellite in end-of-life into IGSO satellite

Known from the previous section, due to various perturbations such as gravity of sun and moon, GEO satellite's inclination will change. Satellite will gradually deviate from the original position and drift in both east-west and north-south direction which is more apparent. There is 0.269° increase in IGSO satellite's inclination each year caused by solar gravity and 0.478°~0.674° increase caused by lunar gravity. The inclination drifts 0.75°~0.95° in north-south direction each year under joint effects of sun and moon. The inclination changes periodically with period of 18.6 years and its largest range ability is about 14.67°. After implementing the new control model, GEO satellite will become a new kind of satellite, namely SIGSO whose inclination is between 1° and 10°.

So if there is just east-west position maintenance and no north-south position maintenance which means satellite can drift in latitudinal direction freely, the retired GEO satellite can become half-controlled SIGSO satellite. As the fuel that north-south position maintenance consume is 9 times as much as the fuel that east-west position maintenance consume, the using time of remained fuel in retired satellite can be prolonged about 10 times with the new orbit regulation model. As mentioned earlier, the satellite's electronic equipment such as solar cells, signal transponders has a relative longer service life, SIGSO satellites can service for many years in transmitting navigation and positioning satellite system. Meanwhile, due to retire GEO communications satellites are no longer bear the original communication and broadcasting tasks, they can be regulated to other required and available track in a larger range in accordance with the requirements about formation of navigation constellation. The

satellite management departments regulating the obits in this way has some technical difficulties, but much easier than developing and launching new satellites, and the satellite transponder lease costs is much less than the costs of developing and launching dedicated navigation satellites. Therefore the formations of navigation constellation composed of the SIGSO satellites which are regulated from GEO satellites make retired GEO communications satellites playing a role in improving the accuracy of navigation. The usage of the retired GEO satellites can not only shorten the time and saves a lot of money to organizing a navigation system, but also delays the time of satellites into space junk which will pollute the space environment. CAPS verification system using SIGSO satellites[1, 9, 10, 13] which is deviated from regulating retired satellites has proven effective.

Of course, utilizing this kind of satellites need regulate the SIGSO satellites to required orbits and meet the provisions of the ITU about anti-interference. It also has extremely high demand of orbits determination since the orbits of satellites are the resources for navigation. CAPS project team using two-way time transfer technology and code-related dispreading demodulation makes satellite ranging accuracy better than 1 cm and orbit observation residuals less than 9 cm[3, 17] which can meet the requirement of transmitting satellite navigation and positioning system.

3.2 Antenna's direction and its Isotropic Radiated Power

After retired satellites regulated to the new orbit, the ground beacons should be re-established or technology of geocentric point mode should be adopted. To meet the navigation's requirement on the effective coverage area and power distribution, effective coverage area of the satellite's antenna and EIRP allocation should be altered by adjusting antenna's direction.

3.3 The method of changing sub-satellite point longitude λ

For more than two satellites, their intervals in longitude direction must be drawn away. Method of the operation is not complicated, as long as one of the satellites has been drifted away, the purpose of changing the fixed-point longitude could be achieved, since changes in phase is consistent with changes in longitude. About 5~6m/s velocity increment is enough for drifting satellite and drift speed is 1°/ day. giving the scope of China's territory, drift of the longitude range can be limited to 50^{o} E ~ 170^{o} E. when satellite is in the new position, it is also necessary to consider electromagnetic compatibility of other satellites nearby.

3.4 The method of changing RAAN

As described above, each satellite's orbital inclination i has the same change with its RAAN when there is no north-south position maintenance. If there is some time difference in suspending the two satellite's north-south position maintenance, a certain Right Ascension difference will be retained during process of change. As is shown in Table 2.2, when the first satellite has been no north-south position maintenance for four years and the second satellite 2 years, there will be 10° difference in Right Ascension of the two satellites.

Changing the satellite's RAAN usually consumes a large amount of fuel. Taking two satellite with inclination 3° and RAAN 75° for example, velocity increment of 55m/s is required to bring about 20° difference in RAAN and velocity increment of 55m/s for 60° difference in RAAN, which is unrealistic for the huge cost. The initial inclination is about 0°

in the beginning of suspending north-south position maintenance, and therefore RAAN is meaningless and thereby setting Right Ascension difference is relatively easy. To set the Right Ascension of two satellites arbitrarily, the only thing need to do is that one of satellite has been established a non-zero angle at a certain time. For example, when satellite passes Right Ascension 90°, RAAN of satellite can be changed to 90° with north direction position maintenance. But the Right Ascension we have set will decrease generally during the following perturbed evolution, the extent of decrease depend on the inclination when setting the RAAN.

For example, RAAN of the two satellites were set to 0° and 180° respectively and inclination is controlled to 0.1°, velocity increment is about 5.5m/s and thereby consuming less fuel. Table 2.2 shows the subsequent changes, as can be seen from Table 2.2, the RAAN of the satellite with the initial RAAN 0° will change to 84.6° in a year and 79.4° in three years, and the RAAN of the satellite with the initial RAAN 180° will change to 96.6° in a year and 78° in three years. The right ascension difference will decrease to 12° in one year from the beginning 180° and to 1.4° three years later. If the inclination of the two satellites is set to 0.5°, while the RAAN is still set to 0° and 180° respectively, the reducing trend of right ascension difference will slow down.

The subsequent evolution of the situation is shown in Table 2.3, the RAAN of the satellite with the initial RAAN 0° will change to 61.7° in a year and 71.7° in three years, and the RAAN of the satellite with the initial RAAN 180° will change to 116.8° in a year and 90.6° in three years. The right ascension difference will decrease to 55° in one year from the beginning 180° and to 15.5° three years later. However, if the inclination has been set too larger, the consumption of propellant will have a substantial increase which is equal to the amount of half-year's consuming fuel and velocity increment will have to be about 27m/s. If we further increase the initial inclination, the reduction of the two satellite's RAAN difference will further slow down but the consumption of propellant will have a substantial increase, which can be seen from the table 2.3, 2.5, 2.5.

3.5 Conclusions

Based on the above analysis, we can obtain the following conclusion:

1. GEO satellite's orbital inclination and RAAN have a 52-year cycle of change due to solar and lunar perturbations, right ascensions reducing from about 90° to 0°, then to 270°, and at last to 270° directly. Therefore, GEO satellites without north-south position maintenance have same changes in their orbital inclination and RAAN.
2. There are two ways to change the time when the two retired satellites pass the equator: one is changing sub-satellite point longitude difference $\Delta\lambda$ between the two satellites, the other is to change the satellite's RAAN difference $\Delta\Omega$.
3. Altering the time when the two retired satellites pass the equator by changing the sub-satellite point longitude can improve navigation performance via the improvements of satellites constellation;
4. It is easier to change the sub-satellite point longitude with control volume 5 ~ 6m/s; consider the scope of China's territory, drift in the longitude range can be limited to 50^{o} E ~ 170^{o} E. However, we should note that the re-coordination of the orbit is required;
5. Altering the orbital phase by the change of the RAAN can improve navigation performance actually via increasing geocentric angles of two satellites, but in case of small

inclination, geometrical distribution can't be significantly improved because the increasing amount of the geocentric angle is not more than the sum of two satellites' inclination;
6. The method that we alter the orbital phase by the change of the RAAN can be optimized in different initial conditions. But even so, the control volume has the same amount of the normal north-south position maintenance and the initial set of orbital phase differences will gradually reduce and have the same value eventually.

4. Contribution of SIGSO satellite for transponder satellite navigation constellation

In this section, through simulation analysis, we will discuss the SIGSO communications satellites in the transponder satellite navigation system how to affect the navigation and positioning constellation structure and improve the navigation and positioning solutions.

4.1 Requirements of positioning on the constellation structure

Generally, position accuracy in satellites navigation system can be represented as:

$$\sigma_x = \mathrm{GDOP}\ \sigma_\rho \tag{4.1}$$

Where $\sigma_x = \sigma_{(x,y,z,t)}$ is positioning and time service error, σ_ρ is pseudo-range measurement error. GDOP (Geometric Dilution of Precision) is an important index to describe the quality of the space section of a positioning system, and its value determines position accuracy[42, 49, 53, 56].

An excellent positioning system requests PDOP value to be small in effective covering region in any time. The GPS system is formed by more than thirty satellites and corresponding PDOP value in covering region is smaller than 4 or less [26].

When GEO communications satellites are used in constellation of regional navigation system, the system needs three communications satellites with the large orbital inclination (IGSO) too. Thus the coverage of the system will be of about one-third of the earth. Three such regional systems can cover the whole world.

In 2003, the problem existing in CAPS test system is that we have not appropriate IGSO satellites to apply in the system. Therefore, we propose to use GEO satellites which are going to be retired for solving the problem. That is, using new mode to control thus GEO satellites to become SIGSO satellites for navigation and positioning. However, the improvement of constellation structure when SIGSO satellites are used should be studied in depth. In this section, by calculating the distribution and variation of constellation PDOP within a certain coverage we mainly analyze quality of the constellation.

4.2 Improvement to PDOP value in satellites navigation system by SIGSO satellites

Based on the theoretical analysis of the constellation and PDOP changing with time, we developed a set of application software of calculation and analysis on PDOP change. For the navigation and positioning constellation includes some satellites with different orbital parameters, the software can calculate the PDOP change in any place on the ground at any time, calculate the PDOP change in the designated area within a certain period, and can

produce various charts of PDOP by the time and distribution of regional areas for simulation and analysis constellation quality.

(1) Analysis of the impact of SIGSO satellites for navigation constellation PDOP

We will analyze the function of SIGSO satellite through studies of distribution and variation of PDOP values for various constellation structures. In simulation, the satellite number of constellation is designed to be six, including the GEO satellites and SIGSO satellites. The number of SIGSO satellite changes from one to six, and number of GEO satellite changes from five to zero correspondingly. The longitude of the satellites is 57°, 87.5°, 103°, 122°, 142° and 163° respectively, and the least of the satellite's elevation for users is 5^o. The orbit inclination of SIGSO satellite is 3^o and 5^o. For each $2^o \times 2^o$ grid point to calculate the 10-minute intervals PDOP values, and plotted PDOP distribution. If the PDOP values of a grid point is always less than 1000 in the required effective period (such as 95%, 98% or 100% of the day), the points will be called effective grid points and they will compose the effective coverage area. Effective area of the PDOP calculated respectively the minimum, maximum and average. The figures and data will be used in analysis of the function of SIGSO satellites.

(a) Constellation including SIGSO and GEO satellites can achieve three-dimensional navigation.

If the positioning system only uses GEO communication satellites without other assistant methods, it cannot do the 3-D positioning since these satellites are all in the earth's equator plane. The calculations show that the DOP value of this constellation is too large for navigation and positioning. For example, when using only the six GEO satellite, the PDOP value of Beijing area approximately is one million or more. Supposing an retired GEO satellite (such as at 110.5° E) has been controlled to be a SIGSO satellite with i = 3°, the effective time is 95% of one day, the calculated PDOP distribution is shown in Figure 4.1. In the scope of about 60° E—160° E and about 65° S—65° N, PDOP minimum is 57.1, and average is 76.2, PDOP <60 in the center of the covered area and PDOP <70 in a considerable area. Although the effective time has not yet reached whole day, and the value of PDOP is also larger, the significant change is the 3-D positioning equation could be solved and fundamentally change the situation that the equation cannot be solved when GEO satellites are used only.

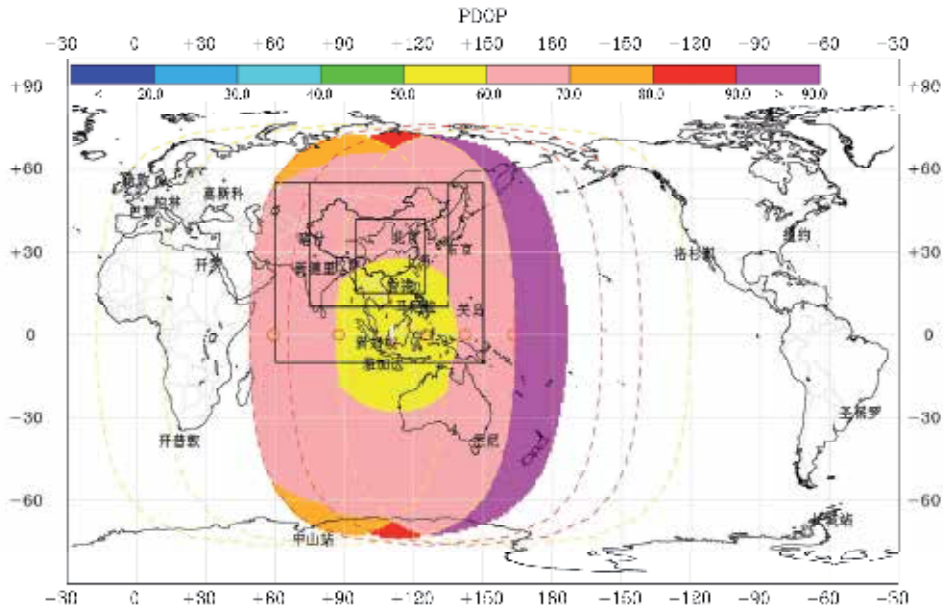


Fig. 4.1 Distribution of PDOP equal-value-area of the constellation of one IGSO (i=35°,e=0) and five GEO satellites.

Notes: Different color stands for the distribution grade levels of the PDOP daily mean and the same color area is called the equal-value-area. The circumscription value of the equal-value-area is marked on the top of the figure, such as PDOP<20 in the blue area and on the analogy of this. The red circles marked with 'G' along the equator are the ground projecting of GEO satellite, the corresponding coverage boundary are shown by dotted lines. The '8' shaped ground track of IGSO satellite is plotted with white dash-dot line, on which the red star represents the position of IGSO satellite, while those white points are the positions at every hour. The three key areas marked by the black line, they are Chinese hinterland, most of the territory and surrounding areas, and it is the same below.

(b) Increase in the number of SIGSO satellite PDOP of constellation can be improved.
Improvement of the PDOP value of constellation is relatively obvious along with the increasing in the number of SIGSO satellites. Based on above case, if retired GEO satellite at 142° E is also controlled to be a SIGSO satellite with i = 3°, that is the combination of 4 GEO satellites and 2 SIGSO satellites is formed, the PDOP is improved obviously. In the cases of effective time is 98% of one day, the minimum reduces from 57.1 to 36.0 and the average decrease to 47.3 from 76.2 in the entire covered region, and the largish region will reach PDOP <50. Assuming an retired GEO satellite at 87.5° E is controlled to be a SIGSO satellite with i=3° and the combination of 3 GEO and 3 SIGSO satellites will be formed and ask the ratio of effective time to reach to 100% of one day, the PDOP distribution is shown in Figure 4.2. PDOP minimum further reduces to 15, average of the entire coverage region reduces to 28.6, the largish region reaches to PDOP <17.
Further assuming the retired GEO satellite at 125° E is controlled to be a SIGSO satellite with i=3° and the combination of 2 GEO and 4 SIGSO satellites will formed, the PDOP is further improved (see Figure 4.3). The minimum will reduce to 12.8, average of the entire coverage

region will reduce to 16.6, the largish region will reach to PDOP <14. If the number of controlled SIGSO satellite increases again, for example, all 6 GEO satellites are controlled to SIGSO ones with i = 3°, PDOP minimum will further reduce to 11.9. For comparison, simulated PDOP values of some constellations are listed in Table 4.1 (i stands the orbital inclination, T is the percentage of effective coverage time in a whole day, min, max and mean show minimum, maximum and regional averages of daily average PDOP respectively, H means using atmospheric pressure altimeters).

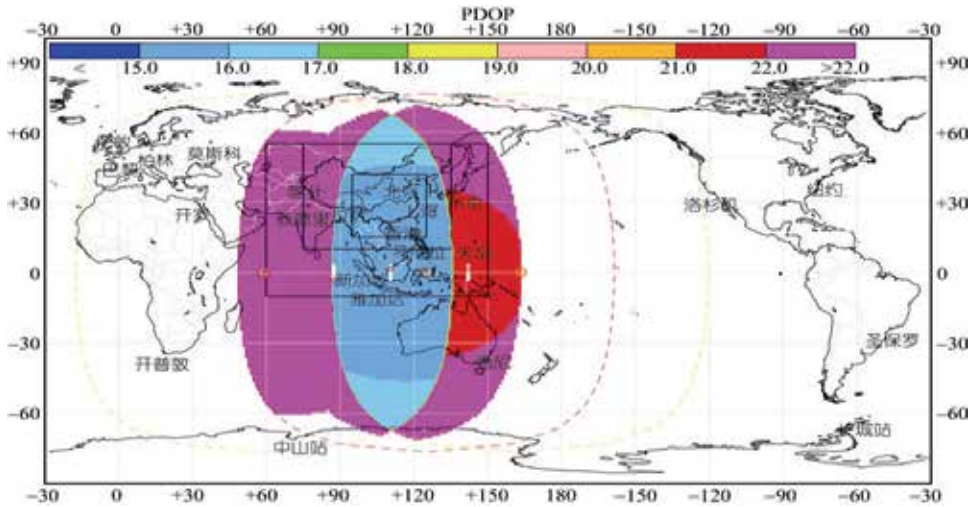


Fig. 4.2 Distribution of PDOP equal-value-area of the constellation of three GEO and three SIGSO satellites with i=3°.

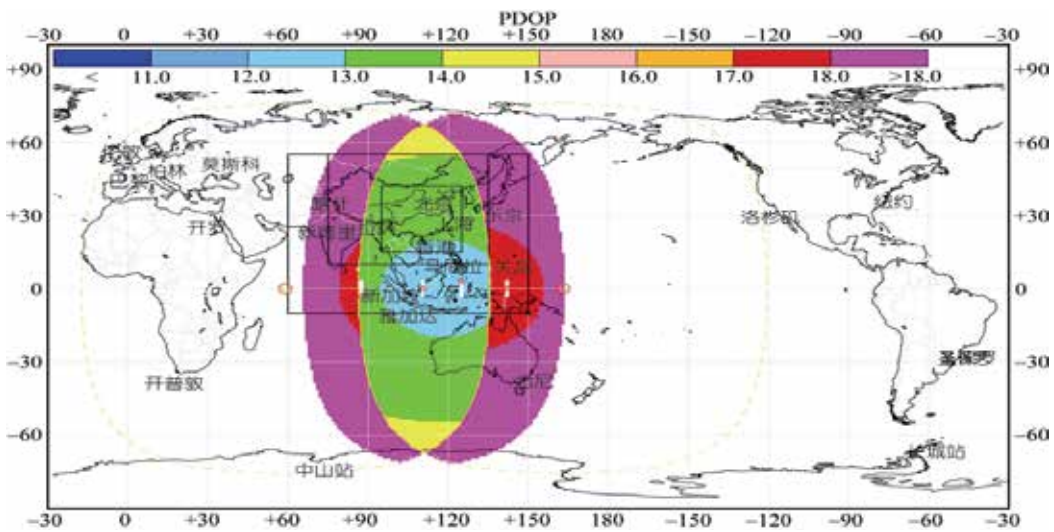


Fig. 4.3 Distribution of PDOP equal-value-area of the constellation of two GEO and four SIGSO satellites with i=3°.

It should be noted that PDOP maximum became larger for the two groups of using atmospheric pressure altimeters in the table 4.1. It is the reason that the effective coverage area increases and the PDOP is relatively large in the marginal parts of the effective coverage area. As the parts are smaller, the impact on the average is also small, and it will not adversely affect the largish region we care about.

Table 4.1 PDOP values of various constellations
(The minimum elevation of satellite is 5°)

Table 4.1 "PDOP" values of various constellations (min elevation 5 degree)

constellation	i	7%	Min	Max	mean	备注
5GEO+18IGSO	1°	95	139.0	170.4	151.8	
	3°	95	57.1	149.1	76.2	
4GEO+2SIGSO	3°	98	36.0	75.1	47.3	
3GEO+3SIGSO	3°	100	15.0	64.0	28.6	
	3° H	100	2.9	108.3	12.2	
2GEO+4SIGSO	3°	100	12.8	33.4	16.6	
	5°	100	7.9	23.5	10.5	
	7°	100	5.9	18.9	8.0	
1GEO+5SIGSO	5°	100	12.1	61.5	16.9	satellite phase 60°
	5°	100	9.1	35.8	13.3	satellite phase 75°
0GEO+6SIGSO	3°	100	24.0	77.2	36.6	
	5°	100	14.7	55.9	22.3	satellite phase 60°
	5°—H	100	3.1	106.4	12.8	satellite phase 120°
	7°	100	10.8	40.4	16.4	
	5°	100	6.6	51.7	11.8	

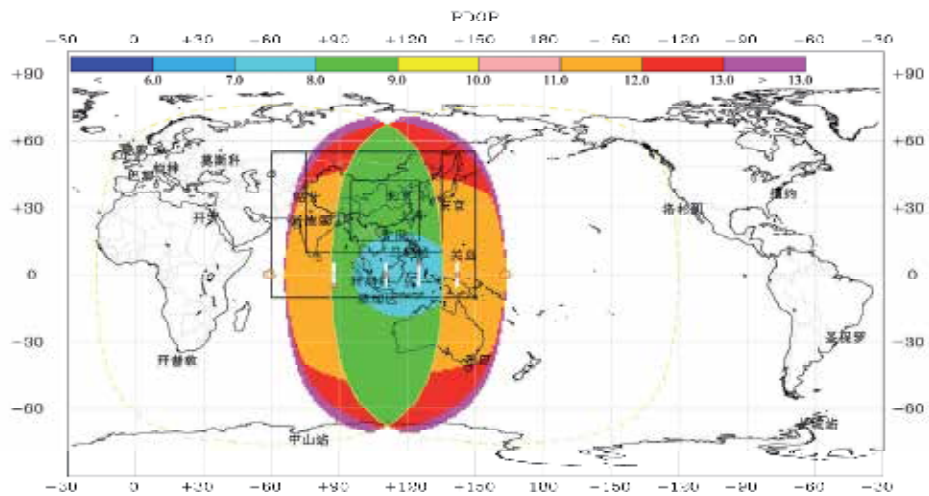


Fig. 4.4 Distribution of PDOP equal-value-area of the constellation of two GEO and four SIGSO satellites with *i*=5°.

(c) The relatively greater inclination of the SIGSO satellite is, the greater the degree of improvement of PDOP value is.

Simulation shows that even if only one SIGSO satellite exists and the orbital inclination *i* increases to 3°, the PDOP minimum will reduce to 57.1. Here we use the combination of two GEO and four SIGSO satellites to further display the impact of the orbital inclination on PDOP. Four SIGSO satellites locate at 87.5° E, 110.5° E, 125° E and 142° E and *i* changes from 3° to 5° and 7° for the simulation. The result of *i*=3° has been showed in Figure 4.3. The results of *i*=5° and 7°, respectively, are showed in Figure 4.4 and 4.5. The comparison of results in Table 4.1 and Figures 4.3, 4.4 and 4.5 show that PDOP value is obviously improved along with the increasing of inclination of SIGSO. The minimum PDOP reduces from 12.8 to 7.9 then to 5.9 and the average decreases from 16.6 to 10.5 and then to 8.0. At the same time, when *i* = 5°, 4/5 of Chinese domain achieves PDOP <9. The change curves of PDOP per minute in 24 hours of some representative locations are in Figure 4.6. The results show that the variation of PDOP per minute is relatively stable and the value is relatively small except for the PDOP is relatively large in north-west marginal region of China which cannot observe 163° E satellite. Using SIGSO satellites are from retired GEO satellites to form constellation and then get such improvements of PDOP that may be not imagine before. In Figure 4.5, we can see that 4/5 of Chinese domain achieves PDOP<7 when *i* = 7°. However it needs a longer time that SIGSO is controlled to be the orbital inclination of 7°, and there should be more of remaining fuel, as well as solar battery and the transponder has a long working life.

In the simulation, if three SIGSO satellites are used and their orbit inclination is about 5°, the PDOP average will be about 15.7, PDOP <12 in a majority of the satellite coverage area and a larger area of PDOP <10 appears in the middle (see Figure 4.6). When using four IGSO satellites, PDOP reduced to an average of 11.8, and the area of PDOP <8 appears in the middle (see Figure 4.7). Compared with the combination of the number of IGSO satellites of *i* = 3°, the improvement of PDOP is obvious when the SIGSO satellites are used, of course, the navigation and positioning accuracy will be improved significantly (as shown in Figure 4.4).

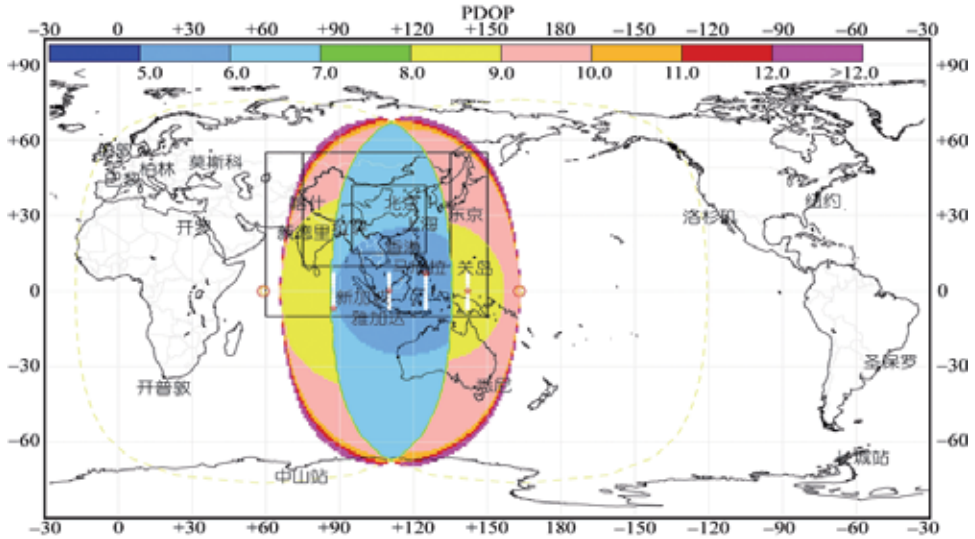


Fig. 4.5 Distribution of PDOP equal-value-area of the constellation of two GEO and four SIGSO satellites with i=7°.

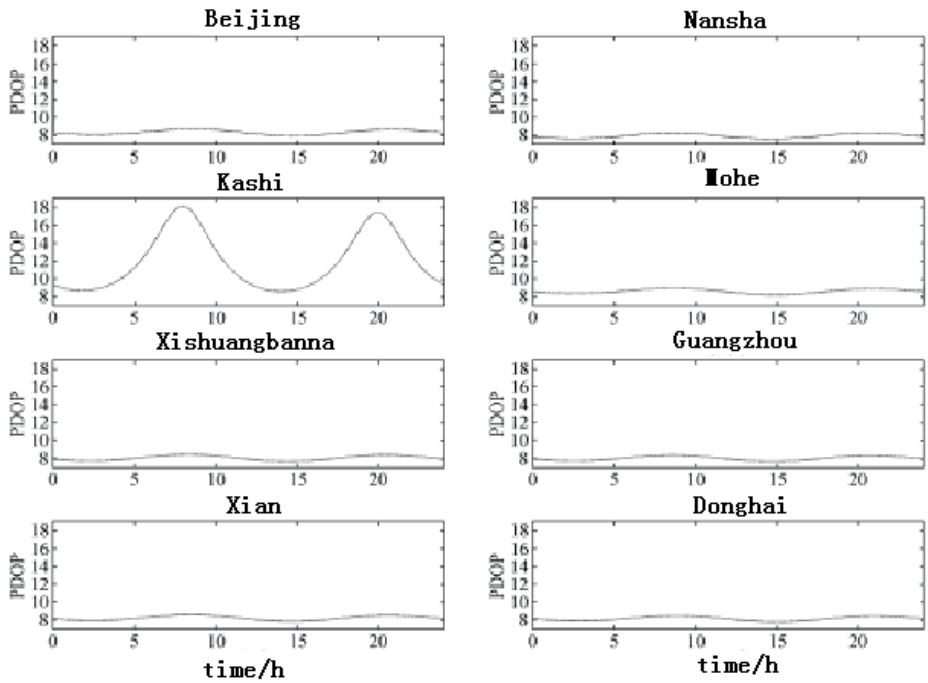


Fig. 4.6 the PDOP value per minute of a constellation with two GEO and four SIGSO (i=5°) satellites in different sites in Chinese.

In order to make better use of contribution of SIGSO satellites for improving the distribution of satellite constellation, it is needful to take into account the phase difference of the SIGSO satellites. These GEO satellites have the same ascending node when they are launched, there are two ways to change their phase. We can adjust the phase directly, of course this way will consume more fuel and it is not advisable since SIGSO satellites have a spot of fuel. Another way is to move SIGSO satellites in the longitude direction in order to increase the interval between satellites and then their phase difference will be increased and the satellite coverage will be increased.

In summary, joining of SIGSO satellites changed the situation that the three-dimensional positioning cannot be made if only use GEO satellites, the three-dimensional positioning can be actualized. More SIGSO satellites and greater orbital inclination, the improvement of PDOP is more obvious. To increase the phase difference between SIGSO satellites, can be used pull SIGSO satellite in longitude direction interval.

(2) SIGSO satellites for improvement of PDOP value of CAPS validating system

The coverage of CAPS focuses on Chinese land and coastal areas, considering there should be a certain redundancy for the coverage, the simulation will be extended to Chinese surrounding areas (60° E—150° E, 0° —65° N). The GEO satellites involved in CAPS are at 87.5° E, 110.5° E, 125° E and 142° E, their span in west-east direction is relatively small. We suppose to launch new GEO communication satellites at 59° E and 163° E, or retired GEO satellites are controlled to become SIGSO ones and moved there (here we mainly consider Chinese existing GEO communication satellites and the possible new ones, as well as SIGSO

satellites from the retired GEO satellites). For the above-mentioned 6 satellites, they gradually enter the end of the life and become SIGSO satellites in sequence, their orbital inclination will reach to 1°, 3° and 5° etc., and minimum altitude angle of satellite is 5°. Under these assumptions, the simulation will provide PDOP value every 10 minutes, at each 1° × 1° grid (latitude × longitude) in the effective coverage area, as well as the daily average and minimum value of PDOP. For regional navigation and positioning system, we are concerned about the value of PDOP in the region and its distribution, so we draw the value of PDOP distribution in different levels. At the same time, in order to fully understand the details of the PDOP changes and further comparative analysis, we calculate PDOP changes per minute in one day and draw PDOP curve for some selected representative sites within the coverage area. Through the analysis of the change and distribution of PDOP, we can get the following results:

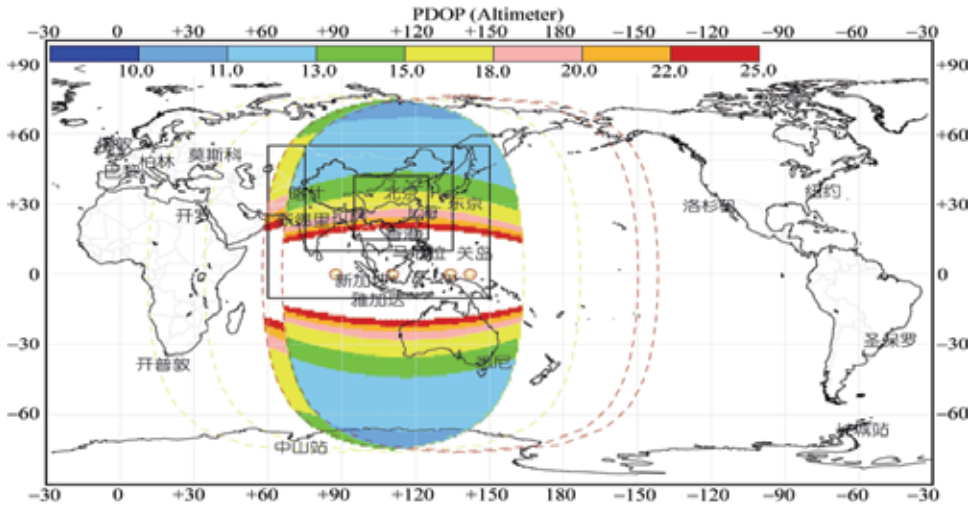


Fig. 4.7 Distribution of PDOP equal-value-area of the constellation of CAPS validating system formed by four GEO satellites and atmospheric pressure altimeters are used.

To verify the function of retired satellites used in the CAPS validating system, based on the situation of using atmospheric pressure altimeters we simulate the constellation is formed by above-mentioned 4 GEO communication satellites, as well as the constellation when GEO satellites at 134° E and 138° E retired and are controlled to be SIGSO satellites with i=3° respectively. The results are shown in Figures 4.7 and Figures 4.8. In absence of retired satellites, the average, minimum and maximum of PDOP value is 43.0, 10.9 and 970.8 (in Figure 4.7, the part for PDOP>25 is set to be white, and the same below). When two satellites retired and became SIGSO ones with i=3°, these values of PDOP are improved as 19.4, 10.9 and 131.8, respectively (see Figure 4.8). It is obvious that retired satellite is certain valuable for improving PDOP value of CAPS validating system. With their orbital inclination increases, the improvement on PDOP will be relatively more pronounced. When the orbit inclination of the two satellites is about 5°, these PDOP values are 14.7, 10.8 and 124.7 (see Figure 4.9). We could see that, although the minimum of PDOP unchanged with the increase of the orbital inclination, the change of the average is obvious, the region with smaller PDOP expanded. For example, the south boundary of the region of 11 <PDOP ≤13 is at the

north of Beijing at first, and then it extends to the southern of Beijing when i = 3°, then the boundary extends to Hainan island, and the region of PDOP ≤ 11 exists in the eastern part of China when i = 5°. Therefore, using relative more SIGSO satellites are from retired GEO satellites, and the orbital inclination is greater relatively, the improvement for the PDOP of CAPS validating system is relatively more pronounced.

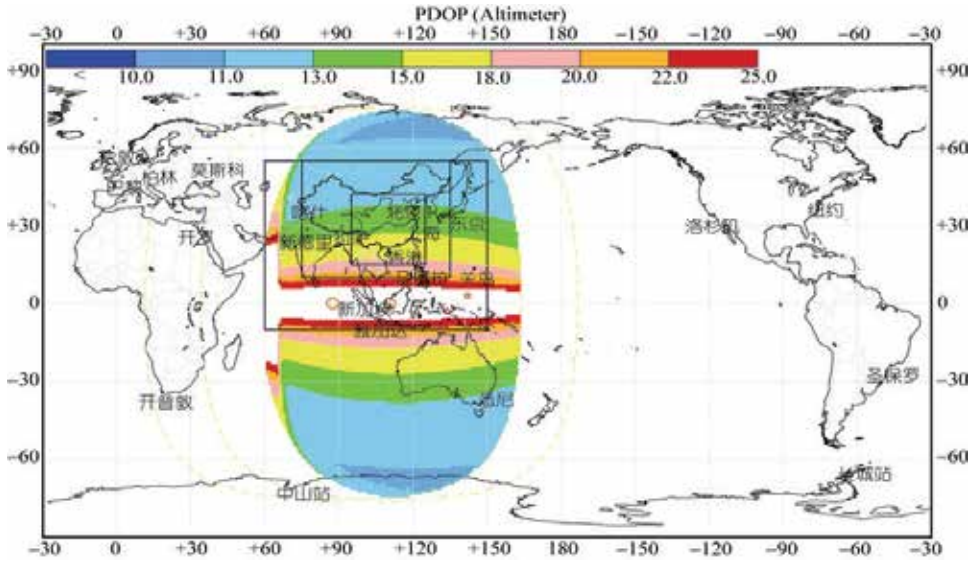


Fig. 4.8 Distribution of PDOP equal-value-area of the constellation of CAPS validating system formed by two GEO satellites and two SIGSO satellites with i=3° and atmospheric pressure altimeters are used.

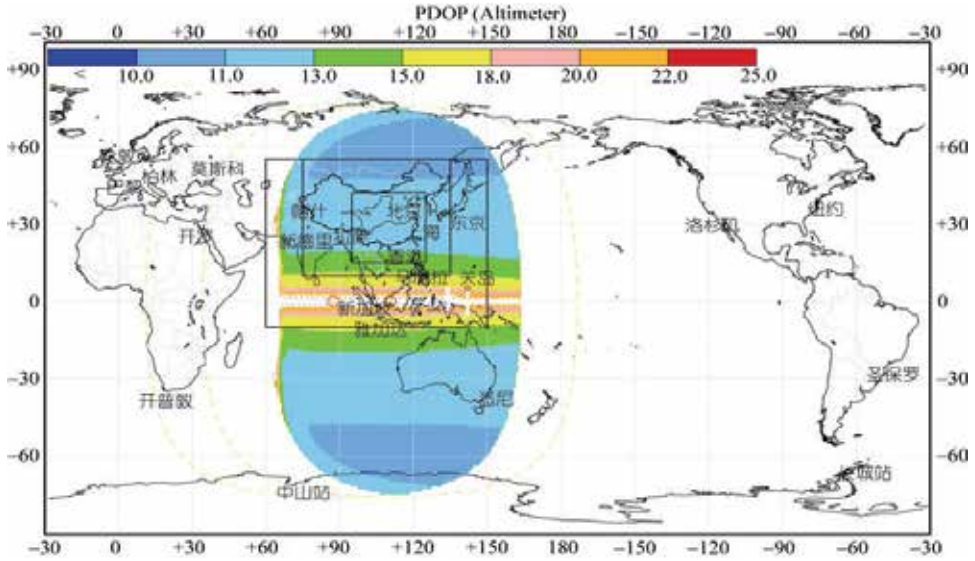


Figure 4.9 Distribution of PDOP equal-value-area of the constellation of CAPS validating system formed by two GEO satellites and two SIGSO satellites with i=5° and atmospheric pressure altimeters are used.

The results show that the retired GEO communication satellite which is controlled to be SIGSO satellite can be used to the transmitting navigation and positioning system, and 3-D navigation and positioning can be actualized. Of course, a certain number of IGSO satellites are necessary to form an ideal constellation for navigation and positioning even if the GEO and SIGSO communication satellites are together used in the constellation or all members of the constellation are SIGSO satellites. The analysis shows that controlled SIGSO satellite from retired GEO satellite can improve PDOP value of the constellation, and its function for improvement of PDOP is better than one of GEO satellite. More SIGSO satellites and greater orbital inclination, the improvement of PDOP is more obvious.

Studies showed that the CAPS needs SIGSO satellites not only lie within the range of 80° E—140° E, but also lie within the ranges of 50° E—80° E and 150° E—180° E, thus navigation accuracy and effective coverage range of the CAPS will be improved significantly. However, Chinese GEO communications satellites launched distributed in 80° E—140° E generally. Through studies, the research team adopted the way to change the position of retired GEO communication satellites and move the satellites into the ranges of 50° E—80° E and 150° E—180° E respectively, in order to form a relatively better navigation constellation. If retired GEO communication satellites do not be used, thus we need launch two GEO communications satellites to the two areas, this must take longer time and spend more fund.

4.3 The capability of anti-jamming

As a lot of C-band transponders in SIGSO satellites with 500/700MHz bandwidth to generate three-frequency, multi-frequency and frequency navigation signals for enhancement of corresponding ground receiving terminals can be used for user's navigation and positioning, it can overcome the limitations of using fixed navigation signal in current navigation system. The implementation procedures are to choice different and contiguous navigation frequencies which can be covered by a ground receiving antenna to convert them down to expected navigation base band signal, and then measure and compensate their converting time delay by software to achieve finally the accuracy navigation and positioning[52, 2]. As the fixed navigation frequency signals with a low power and easily interfered are used in common satellite navigation systems, frequency-hopping technology[1, 6, 7] may be applicable to anti-jamming in satellite navigation, therefore, an innovative satellite navigation system can be built up with the functions of multi-frequency and hopping for navigation, they will significantly improve anti-jamming capability and increase more functions in navigation system.

5. A low data rate satellite communications technique

A low data rate satellite communication system is being developed in the navigation and positioning system based GEO communications satellites as SIGSO satellites around end of life span. Such a communication system has its especially technical issues and demands, of course, and its technical difficulties.

5.1 Why developing this type of satellite communication techniques

By means of changing the attitude control mode, GEO communication satellite around the end of life span will drift into a small angle inclined orbit (SIGSO) satellite. Except 2-3 band

transponders for navigation purpose, if remaining 21 to 22 transponders resource can be used well? They are very cheap space resources. In general, human beings demand more increasing communications, so satellite communications techniques tend to improve bandwidth efficiency and data transmission rate with higher cost-effective .Would SIGSO satellite meet these needs?

As SIGSO satellites have "8" shape shift daily movement and continue increasing drift, the communication terminal antenna must point to the satellite. It will be required to be able to control the antenna beam or antenna beam directive should be able to cover a very wide range, in particular, the composition and structure of the terminal installed on the mobile carrier will become more complex.

There are various demands of application, for example, vehicle travel monitoring carrying dangerous goods; the transmission of measured data in meteorological and hydrological monitoring stations in remote mountains, ocean and desert etc.; emergency information reporting in real-time in case of disaster; corrosion monitoring for petroleum pipeline and so on. Particularly, after GPS is being used widely, people needs increasingly location and time information transmission service which needs an innovative medium with capacities of long-distance transmission in real-time, low information transmission under a strict error code rate and stronger anti-interference. There are potentially numerous users who hopefully need cheaper device and lower using cost. Thus, a new classified satellite communications with an extra low-information rate are being developed widely.

5.2 Configuration of navigation and communications system

Navigation and communication system is composed of SIGSO communications satellites, master communication ground station, user terminal, data distribution and network management system. User segment is user receivers that are classified into navigation/ communication or combined receivers. User receivers in functions may include Positioning, Timing, Velocity and Communication; carrier feature for user terminals could be classified as static, dynamic and highly dynamic motion. The sorts of device include portable devices, vehicle terminal, mobile terminals and static terminals.

CAPS project now has two SIGSO satellite communications with capability of communications for users, and provide positioning and navigation services with higher positioning accuracy—similar to GPS positioning accuracy, and higher timing precision in all weather and day in service area (longitude 60° ~ 160°, latitude 20° ~ latitude 60°) user terminals are featured by short message and low data rate voice communication and compressed image transmission. CAPS communications hub station locates in Huairou region Beijing with 16 meters aperture antenna operated by automatic tracking. Test system shown in Figure 5.1.

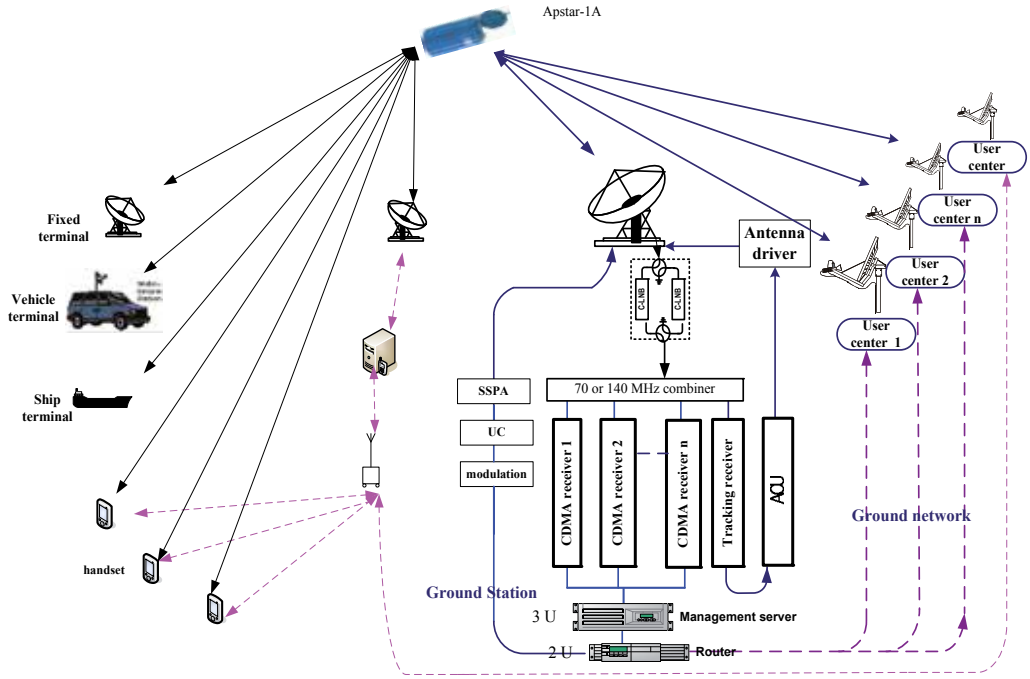


Fig. 5.1 Test principle of CAPS-based integrated communications and navigation in CAPS

5.3 Optimal Design of Link

A portable mobile devices with lighter weight, lower power and lower cost seems to be incompatible in a low data rate satellite communication system, development difficulties focus on user terminal with low power, miniaturized and cost-effective, especially, the design of miniaturized antenna in portable user terminal without tracking systems have to be solved, so, the design of the transmission power link must be optimized.

Power link margin is usual expressed by decibel difference between system Eb/N0 and demodulation threshold (Eb/N0). Eb/N0 obtained by the following formula

$$\left[E_b/N_0\right]=\left[C/T\right]_t-\left[\mathrm{R_b}\right]-\left[k\right], \tag{5.1}$$

where $\left[C/T\right]_t$ (C / T) t is total link carrier power and noise power ratio, shown in (5.2) as follows:

$$\left(\frac{C}{T}\right)_t=\left[\left(\frac{C}{T}\right)_u^{-1}+\left(\frac{C}{T}\right)_d^{-1}+\left(\frac{C}{T}\right)_{co}^{-1}+\left(\frac{C}{T}\right)_{im}^{-1}+\left(\frac{C}{T}\right)_{adj}^{-1}\right]^{-1}, \tag{5.2}$$

where, C/T_u, C/T_{co}, C/T_{im}, C/T_{adj} uplink CNR, co-channel interference CNR, inter-modulation interference CNR and adjacent satellite interference CNR, respectively, their expression are shown as follows:

$$[C/T]_{\mathrm{u}} = [\mathrm{EIRP}]_{\mathrm{e}} - [L_{\mathrm{u}}] - [L'] + [G/T]_{\mathrm{s}}, \tag{5.3}$$

$$[C/T]_{\mathrm{co}} = 23 + [k] + [B], \tag{5.4}$$

$$[C/T]_{\mathrm{im}} = 20 + [k] + [B], \tag{5.5}$$

$$[C/T]_{\mathrm{adj}} = [G_{\mathrm{r}}] - 29 + 25 \times \log(2) + [k] + [B], \tag{5.6}$$

where, L_{u} is free space path loss in uplink transmission, L' are other losses, including the uplink rain attenuation, antenna pointing error in master control navigation station; $[G/T]_{\mathrm{s}}$ are the gain and noise temperature ratio of satellite receiving antenna.

The three parameters C/T_{co}, C/T_{im}, C/T_{adj} of total Carrier to Noise ratio $[C/T]_{\mathrm{t}}$ in satellite communications system relate with carrier noise bandwidth and temperature parameters, and C/T_{co} relate with the number of CDMA users working in same frequency and cross-correlation of their PN codes. If 20 groups GOLD code are selected as PN codes with 2047 code length, the above three values are generally 130 ~ 150 dBW / K in inbound communication link, the C/T_{u} at the margin of APstar-1 coverage is about 190 dBW / K, therefore, in equation (5.2), C/T_{u} and the C/T_{d} are dominant. $(G/T)_{\mathrm{s}}$ is about 7 dB / K at the margin of APstar coverage, if R_{b} takes 50 bps, demodulation threshold $(E_{\mathrm{b}}/N_0)_{\mathrm{th}}$ is 10.6dB, it will be known by Eq. (5.3), $(C/T)_{\mathrm{u}}$ value is 195.3 dBW / K, the uplink C/N is very low, and the demodulation threshold of signal to satellite is close to system threshold. Similarly, for the outbound link, as receiving antenna gain is lower, downlink or C/N is also limited power system.

In order to achieve miniaturization of user terminal, the terminal's power amplifier configuration should not be too high. Data rate is less to 50-300bps for short message transmission with antenna gain of 5-6dBi. CNR uplink signal to satellite is close to demodulation threshold with very weak signal, so a large size antenna need to be configured to center ground station in this system so that downlink CNR is not dominant to have weak signal receiving and demodulation. The link power budget in the existing APstar communications satellite are shown as follows.

The link power budget results in inbound (terminal transmitting and receiving in central ground station) and outbound (transmitting in central ground station and terminal receiving), respectively are shown in Table 5.1 and Table 5.2.

Table 5.1 Inbound link power budget in C-band mobile communication

APstar-I 2A-50bps--single-carrier transmitting in margin region and received in Beijing	
	Edge coverage - Beijing
A. Satellite Parameters	**Parameters**
1. Satellite Name	APstar-I
2. Orbital position (deg.E)	142
3. Transponder No.	1A

4. Transponder gain attenuation(dB)	2
5. Satellite saturated effective isotropic radiated power (dBw)	36.00
6. Satellite Saturation Flux Density (dBw/m^2)	-90.82
7. Satellite receiving antenna gain and noise temperature ratio (dB/k)	-2.00
8. Transponder Output backoff (dB)	4.50
9. Transponder input backoff (dB)	10.50
10. Uplink Frequency (MHz)	5965.00
11. Downlink Frequency (MHz)	3740.00
12. Transponder Bandwidth (MHz)	72.00
B. Ground station parameters	
1. Transmission parameters□	
a. Antenna diameter (m)	Monopole Antenna
b. Ground station location	Coverage
c. Feed losses (dB)	1.00
d. Power amplifier to feed insertion loss (dB)	0.50
e. Type of power amplifier (klystron, traveling wave tube, SSPA)	SSPA
f. Power amplifier output (dBw)（10W）	10.00
g. Power amplifier maximum output power (dBw)	10.79
h. Amplifier output margin (dBW)	0.79
i. Antenna efficiency (%)	45.00
l. Antenna gain (dB)	5.70
k. Antenna tracking capability	no
l. Type of antenna	mobile
2. Transmission parameters	
a. Antenna diameter (m)	16.00
b. Ground station position	Beijing
c. Longitude (deg.E)	116.46
d. Latitude (deg.N)	39.92
e. To the satellite distance (km)	38031.35
f. Antenna elevation angle (deg)	36.85
g. Antenna Azimuth (deg)	143.33
h. Antenna efficiency (%)	60
i. Antenna gain (dBi)	53.70

j. Receiving system noise temperature (dBk)	19.54
k.Antenna gain and noise temperature ratio (dB/k)	34.16
l. Antenna tracking capability	Automatic
m. Antenna type	Fixed
C. Space loss	
1. Uplink Rain Attenuation (dB)	0.5
2. Downlink Rain Attenuation (dB)	0.3
3. Uplink free space attenuation (dB)	200.50
4. Downlink free-space attenuation (dB)	195.51
5. Uplink polarization loss	3.00
D. Carrier Parameters	
1. Carrier Type	DSSS
2. Multiplexed mode	CDMA
3. Modulation	BPSK
4. Data Rate (bps)	50
5. Chip rate (kbps)	2047.00
6. Spreading frequency gain	46.12
7. Carrier Noise Bandwidth (khz)	2800.00
8. Carrier allocated bandwidth (khz)	3000.00
9. Threshold E_b/N_0 (dB)	10.00
E. Link Budget	
1. Uplink C/T	
a. Ground station effective isotropic radiated power (dBw)	11.20
b. Uplink free space loss (dB)	200.50
c. Loss of antenna pointing (dB)	2.00
d. Antenna gain per square meter (dB/m^2)	36.96
e. Satellite carrier flux density (dBw/m^2)	-141.55
f. Transponder saturation flux density (dBw/m^2)	-90.82
g. Carrier input backoff (dB)	50.73
h. Satellite antenna gain and noise temperature ratio (dB/k)	-2.00
i. Uplink C/T (dBw/k)	-193.80
2. Downlink C/T	

a. Satellite effective isotropic radiated power saturation (dBw)	36.00
b. Carrier Output backoff (dB)	44.73
c. Carrier Downlink effective isotropic radiated power (dBw)	-8.73
c. Downlink carrier per channel effective isotropic radiated power (dBw)	-19.52
d. Downlink free-space loss (dB)	195.51
e. Receiving antenna pointing error (dB)	0.30
f. Antenna gain and noise temperature ratio (dB/k)	34.16
g. Downlink C/T (dBw)	-181.47
3. Co-channel interference C/T (dBw/k)	-162.31
4. Inter-modulation interference C/T (dBw/k)	-144.13
5. Satellite interference C/T (dBw/k)	-131.90
6. Total C/T and C/N	
a. Total C/T (dBw/k)	-194.05
b. Boltzmann constant (dBw/k-Hz)	-228.6
c. Receiver Noise Bandwidth (dB-Hz)	64.47
d. Total C/N (dB)	-29.92
e. Threshold E_{bi}/N_0 (dB)	10
f. Threshold C/N (dB)	-37.48
g. Link margin (dB)	7.56

Table 5.2 Outbound link power budget in C-band mobile communication

APstar-I 2A-50bps—Beijing station transmitting and terminal received in margin region	
	Beijing-Edge coverage
A. Satellite Parameters	**Parameters**
1. Satellite Name	Apstar-1
2. Orbital position (deg.E)	142
3. Transponder No.	5A
4. Transponder gain attenuation(dB)	8
5. Satellite saturated effective isotropic radiated power (dBw)	36.00
6. Satellite Saturation Flux Density (dBw/m^2)	-86.73

7. Satellite receiving antenna gain and noise temperature ratio (dB/k)	-1.12
8. Transponder Output backoff (dB)	4.50
9. Transponder input backoff (dB)	10.50
10. Uplink Frequency (MHz)	6105.00
11. Downlink Frequency (MHz)	3880.00
12. Transponder Bandwidth (MHz)	36.00
B. Ground station parameters	
1. Transmission parameters	
a. Antenna diameter (m)	16.00
b. Ground station location	Beijing
c. Longitude (deg.E)	116.46
d. Latitude (deg.N)	39.92
e. Distance to satellite (km)	38031.4
f. Elevation angle (deg)	36.85
g. Azimuth angle (deg)	143.33
h. Feed losses (dB)	0.30
i. Power amplifier to feed insertion loss (dB)	3.00
j. Type of power amplifier (klystron, traveling wave tube,SSPA)	SSPA
k. Power amplifier output (dBw)	2.21
l. Power amplifier maximum output power (dBw)	20.00
m. Amplifier output margin (dBW)	17.79
n. Antenna efficiency (%)	60.00
o. Antenna gain (dB)	57.98
p. Antenna tracking capability	Automatic
q. Type of antenna	Fixed
2. Transmission parameters	
a. Type of antenna	Double spiral antenna
b. Station position	Edge coverage
c. Antenna gain (dBi)	5.00
d. Receiving system noise temperature (dBk)	24.13
e. Antenna gain and noise temperature ratio (dB/k)	-19.13
f. Antenna tracking capability	None
g. Antenna type	Mobile

C. Space loss	
1. Uplink Rain Attenuation (dB)	0.5
2. Downlink Rain Attenuation (dB)	0.3
3. Uplink free space attenuation (dB)	199.77
4. Downlink free-space attenuation (dB)	196.50
5. Uplink polarization loss	0.00
6. Downlink polarization loss	3.00
D. Carrier Parameters	
1. Carrier Type	DSSS
2. Multiplexed mode	Single carrier
3. Modulation	BPSK
4. Data Rate (bps)	50
5. Chip rate (kbps)	2047.00
6. Spreading frequency gain	49.13
7. Number of path	1
8. Carrier Noise Bandwidth (khz)	2800.00
9. Carrier allocated bandwidth (khz)	3000.00
10. Threshold E_b/N_0 (dB)	10.00
E. Link Budget	
1. Uplink C/T	
a. Ground station effective isotropic radiated power (dBw)	56.89
b. Uplink free space loss (dB)	199.77
c. Loss of antenna pointing (dB)	0.30
d. Antenna gain per square meter (dB/m^2)	37.16
e. Satellite carrier flux density (dBw/m^2)	-105.71
f. Transponder saturation flux density (dBw/m^2)	-86.73
g. Carrier input backoff (dB)	18.98
h. Satellite antenna gain and noise temperature ratio (dB/k)	-1.12
i. Uplink C/T (dBw/k)	-144.80
2. Downlink C/T	
a. Satellite effective isotropic radiated power saturation (dBw)	36.00
b. Carrier Output backoff (dB)	12.98
c. Carrier Downlink effective isotropic radiated power (dBw)	23.02

c. Downlink carrier per channel effective isotropic radiated power (dBw)	23.02
d. Downlink free-space loss (dB)	196.50
e. Receiving antenna pointing error (dB)	2.00
f. Antenna gain and noise temperature ratio (dB/k)	-19.13
g. Downlink C/T (dBw)	-197.92
3. Co-channel interference C/T (dBw/k)	-141.13
4. Inter-modulation interference C/T (dBw/k)	-144.13
5. Satellite interference C/T (dBw/k)	-180.60
6. Total C/T and C/N	
a. Total C/T (dBw/k)	-198.00
b. Boltzmann constant (dBw/k-Hz)	-228.6
c. Receiver Noise Bandwidth (dB-Hz)	64.47
d. Total C/N (dB)	-33.87
e. Threshold E_{bi}/N_0 (dB)	10
f. Threshold C/N (dB)	-40.49
g. Link margin (dB)	6.62

From Table 5.1 and Table 5.2, inbound and outbound margins are about 7dB, and relatively abundant, and can ensure reliable information transmission through satellite.

5.4 The key technologies

The following technologies are adopted to solve limited power and meet special design requirements for small terminal device, as well as SIGSO satellites operational requirements.

1) Broadband spread spectrum technology

As bandwidth resources in the SIGSO satellite are abundant, so it can be used to wideband spread-spectrum communication. The trade off using frequency bands for power solves the limited power and enhancement of C/N. the bandwidths of spread spectrum are optimized in terms of meeting the relevant ITU provisions so that we can solve adjacent satellite interference in using wide beam antenna, reducing the power flux density, and improving the system of anti-jamming capability.

2) Large size receiving antenna in ground station

The small size and wide beam antenna with lower gain and power result in power and CNR limitation in satellite transmission links, it can be solved by adjusting attenuation in the transponder, or adopting FEC technique, besides that, a practical solution is to use large size receiving antenna on ground station. Thus, CNR in the downlink cannot effect on the SNR of the whole link which can be met well.

3) Shaped wide beam antenna

As SIGSO satellite drift all the time, and its daily motion is "8" shape (Figure 5.2). Such as the APstar-1 satellite drifts over ± 4°/day in north-southern direction, the azimuth and elevation of satellite observing from antenna change larger per day (Figure 5.3).

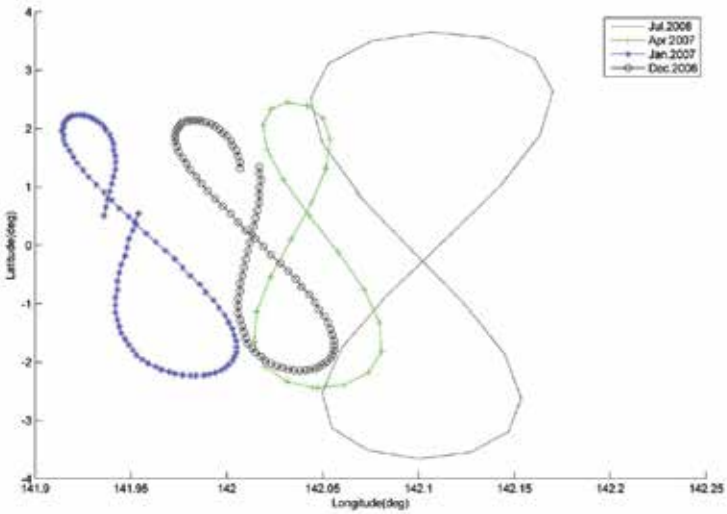


Fig. 5.2 APstar-1 track of subsatellite point

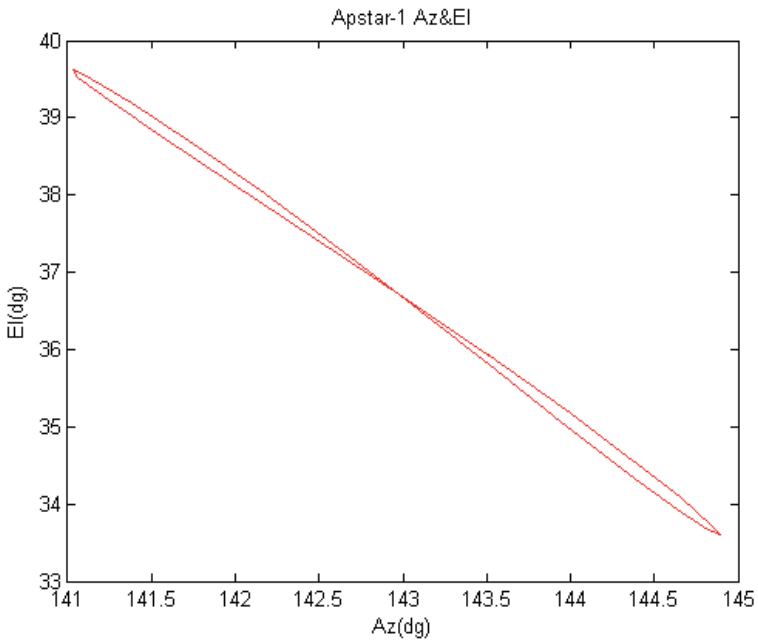


Fig. 5.3 The azimuth and elevation in Beijing earth station

In the mobile satellite communication systems, antenna is a crucial component. Antenna directive, tracking accuracy and its cost restrict the overall system performance and price in certain extent, restricted definitely the availability and promotion of the system. Therefore, we developed a monopole shaped beam antenna. Antenna pattern is shown in Figure 5.4, where it is omnidirectional and wide-beam with about 60° elevation angle, the antenna maximum direction can be designed to required direction. It totally does not need to change the antenna attitude to meet the demand for satellites drift. And even devices would be installed in the mobile vehicle, they can meet the moving dynamic requirements, so it is very simple and cheap.

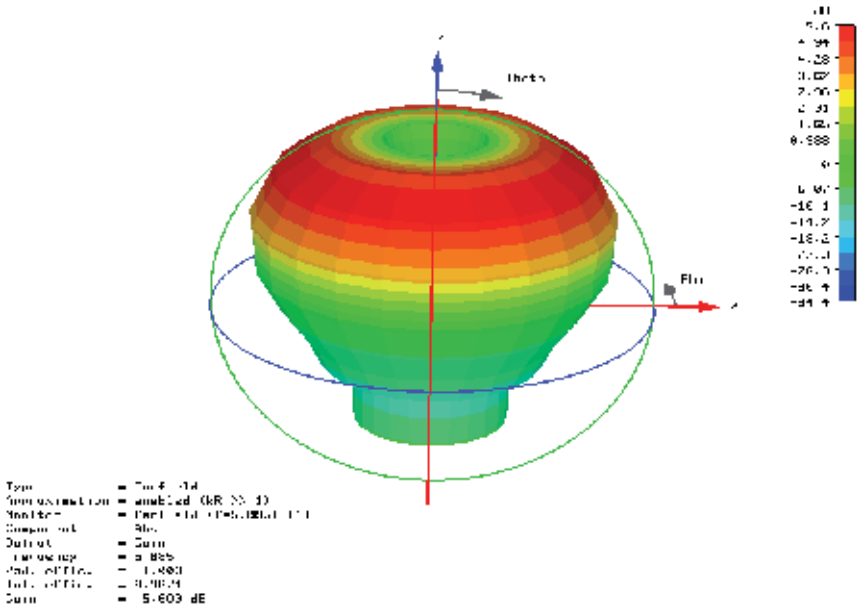


Fig. 5.4 Monopole Antenna 3-D pattern

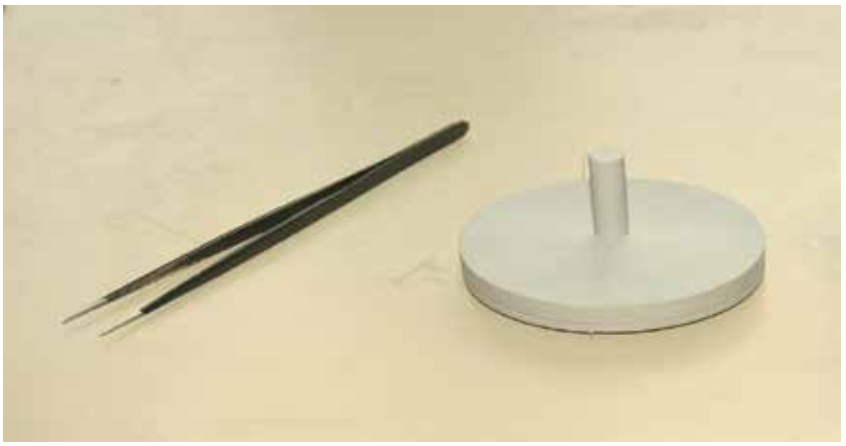


Fig. 5.5 Monopole Antenna

Terminal antenna is vertical polarization; polarization loss is relatively small between antenna and satellite. Antenna shape is shown in Figure 5.5. The disc of its bottom is reflection plate and its diameter is 6cm.

Other kind receiving antenna is double spiral antenna; its 3-D antenna pattern is shown in Figure 5.6.

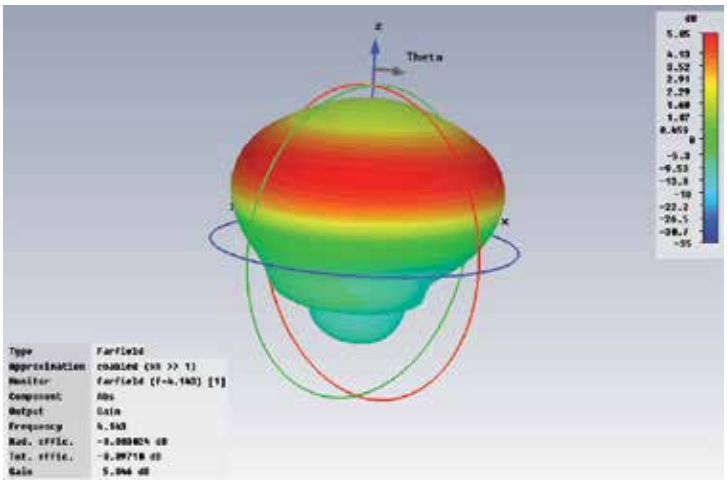


Fig. 5.6 Three-dimensional pattern of double spiral antenna

Double spiral antenna gain is relatively low around 5dBi with whip-like in (as shown in Figure 5.7).

Fig. 5.7 Double spiral antenna

In order to prevent carrier signal interfere from earth stations with neighboring satellites, the International Radio Consultative Committee and the International Telecommunication Union (ITU) on the uplink station carrier axis radiation power density is strictly required; in order to prevent interference to ground microwave communications, ITU regulated

international standards to allow power flux density on the ground transmitted from space communications systems. In addition, the coordination is required about above two issues by satellite operators. From Figure 5.4 and Figure 5.6 it can be seen that such antenna beam is wide, and even adjacent satellite is radiated by antenna main beam, so, the measures must be taken to avoid interference with adjacent satellite.

It is very important to three aspects mentioned above for low data rate satellite communication system design which is typically different from normal satellite communication system, of course, in addition, technologies generally used in satellite communications need to be optimized, such as the use of data compression, optimized multiple access and modulation and so on.

5.5. The system characteristics and advantages

1) Information exchange

As CAPS has successfully developed satellite communications technology with a low data rate, it can exchange and transmit spatial and time information. This system can provide the spatial location and time information.

2) Stronger anti-interference

As the system uses wideband PN spread spectrum system, signal will be buried in the noise below, so the signal will not be easy to be intercepted and deciphered, but also it is difficult to interfere with its transmission and has anti-interference ability.

3) Cost-effectiveness

Constellation adopts GEO communications satellite around late life span; it is prolonged 10 times life expectancy and satellite by changing attitude control with very cheap transponder. System is designed to use small devices and small antenna; equipment development costs are relatively small. Ground station can continue to use the original communication equipments and operate with a little modified equipment. So, the whole system invests less, system cost-effective and easy to build up.

5.6 Typical application

1) Based on the position and time service

The development of a low data rate of satellite communications technology in satellite navigation system based on communication satellite enables to communicate each other to position message and create more applications in combining of satellite navigation and position and timing service, it solve the limitations of individual navigation service in GPS, and to realize the group and center system navigation, and center. Figure 5.8 is shown terminal vehicle positioning information in the electronic map picture.

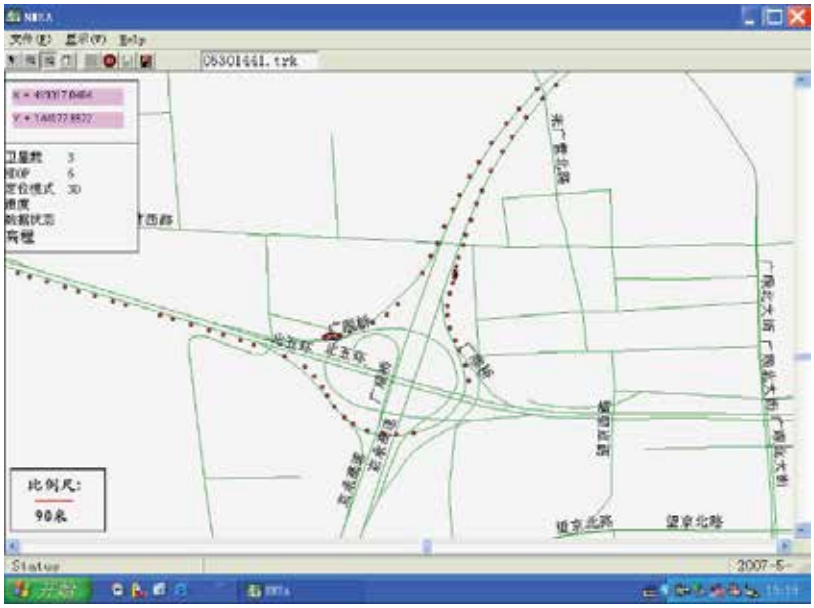


Fig. 5.8 Display on electronic map of vehicle terminal location

2) Applicative for scattered users, rare routing and small traffic communication network, data transmission in automated and unattended station.

3) Applicative for anti-terrorism, disaster relief, emergency communications, When the earthquake disaster and other emergencies happen, it is very important the timely to report the disaster. Disaster relief center also need to control the situation in real-time, directly command and deploy disaster relief teams. People need emergency communication means as facing terrorist activities, also need reliable means of communication, which requires. The project develops a low information rate of satellite communications to meet the emergency communication needs.

4) A new satellite monitoring function

Low information rate satellite communication system can send timely data to control center, for example, operating parameters of large equipments like scraper, oil pipeline corrosion data, measurements data in marine, hydrological and meteorological investigation, logistics management information of ship and vehicle, displacement transfer parameters of reservoirs, rivers, lakes, and dams and so on, contrarily, control center can also transmit instructions to the monitoring terminal in time, to change work parameters and set working status for monitoring.

Totally, the low information rate satellite communications experiments for integration of navigation and communication and multi-satellite communication applications, and to be a practical technology orient to market applications.

6. The application value of GEO communications satellites with extra-live span

We proposed the technological methods to reuse retired communication satellite[1], it has obtained significant results through the conceptive research, simulation and project validation, as well as practical application.
CAPS system now use Apstar-1 and Apstar-1A satellites, where orbital position of Apstar-1 satellite was adjusted from the 138° E to the current 142° E orbital position, Apstar-1A satellite from the 134° E to current 130° E. Apstar-1 and Apstar-1A satellite began to be adjusted to SIGSO satellite in August 2004 and in May 2005 respectively. They still keep orbital precision to the original ± 0.05° [19, 20] along latitudinal direction without control in longitudinal direction. Apstar-1 satellite's inclination had reached 4.1°, Apstar-1A satellite's inclination was 3.5° till March 2009. The following detailed description of the project application value:

1) The protection of spatial environmental resources
Research and realization of retired GEO communications satellite remaining payload fuel without abandonment can be used as constellation of satellite navigation and positioning system. they are only be maintained and adjusted along longitude in orbit spaces, the working mode can save 90% fuel keeping the new control method for CAPS and significantly extend the life of satellites, a retired communications satellite achieves to reuse, creating a development and utilization of extra life-cycle of satellite.
A new model of satellite orbit control has been study and proposes new methods and technologies of reuse retired GEO communications satellites. For example, AP-star 1 satellite reuse, the remained using period of fuel in normal control mode was 3 to 6 months while it was replaced by other new satellite in 2004. At this time, the satellite's solar cells and electronic equipment is still intact with 24 transponders still working, even six spare transponders not yet being used. In the new orbit control mode, the retired satellite became the SIGSO satellite, and has worked more than five years, the GEO communications satellite resources are ultimately used and delay the period to become space junk and pollute the space environment, it plays a very important role to protect the space environment resulting in significant social and economic benefits. In contrast, in previous years, without this new technique, the AP-star 1 satellite in end of life span would be replaced by new satellite and retire to be abandoned as space junk.

2) The constellation of satellite navigation and positioning system, improvement of navigation and positioning accuracy
In the new satellite control mode, the retired GEO communications satellite were regulated as SIGSO satellite to improve the geometry of the navigation constellation, reducing constellation DOP, and can further improve the navigation and positioning accuracy, therefore, it is obvious augmentation due to SIGSO satellite.
Chinese satellite navigation and positioning system (CAPS) purchased retired APstar-1 GEO communications satellite at 138° E orbit, and maneuvers to the 142° E orbital position to SIGSO satellite, it realized GEO communications satellites to reuse. A preliminary application system was configured by adding other two transponders in two GEO satellites, and barometer supplement to realize 3-D positioning. In 2005, the system has reached the GPS navigation and positioning accuracy in coarse code level, positioning accuracy of 10 m,

velocity accuracy is 0.2 m / s, timing accuracy of 12 ns [2, 12]. Communication data rate 50bps ~ 1.2kbps; spread spectrum bandwidth of ± 1.5M; transmitting power 3W ~ 20W; BER 1 × 10-5 ~ 1 × 10-6; modulation is BPSK. Reusing retired satellites can extend the life span of the satellite more than 8 to 10 times, and the band-costing decreasing by about 80%. It will save investment and costing less than one-tenth of the new satellite launching to buy this satellite, or lease their transponders, this can generate significant economic benefits in terms of deploying navigation system and investment. The innovative satellite navigation and positioning system is featured lower cost, shorter deployment period and better performance.

Retired satellites often have full-band (C band) transponders, it can be used to navigation frequency change to become a multi-band satellite navigation system and find new way to improve anti-interference capability, lay the foundation for reuse retired GEO satellites. For example, currently, it is a rather large PDOP <10 region. SIGSO satellite orbital inclination increases along with regulation time (for example, when at about 7^{o}), PDOP value is further reduced,

3) Developing a satellite communication system with a low data rate transmission and small satellite user terminals

GPS class satellite navigation system can provide user position, but its positioning message cannot transfer and communicate each other within the system. The satellite navigation system based communications transponders have a strong communication transmission capacity, easily achieved communicating positioning message, and enhance information value. Conventionally, we need to lease other GEO communication satellite transponders, but do need to increase the cost of transponder rental. Fortunately, we bought a retired GEO communications satellite with 24 transponders where two transponders use for navigation, in addition, using three transponders for tri-frequency measurement, so there are 21 transponders can be used for communication. Reuse of retired communications satellite is to turn waste into treasure, and make lower channel cost in this satellite communication system. Mobile Micro-Satellite Terminal(MMSAT) with wide-beam antenna are applicable to mobile communications, especially time, position and state message service in navigation system, the development of the spread spectrum communication system with low data rate and small antenna and user terminal can be used for short messaging, voice communication and transmission of static images in various fields such as traffic safety and vehicle management, calling, emergency rescue, search and rescue, emergency rescue and disaster relief, mobile TV and video broadcast. Satellite MMSAT terminal series can be formed to provide a short message and data communication services, which promote the application of satellite communications. Therefore, it is necessary to use more retied GEO communications satellite which provides a plentiful space transponder resource to give the full enhancement of combined navigation and communication.

NAOC in china has developed a class low data rate satellite communications technology with short text and data messages more than four years in vehicle monitoring and oceanic observation, recently, a class of low data rate voice communication system is developing.

7. Reference

1. Inventor: Shi H L, Ai G X, Chen J B, Han Y B, Shang J N, Geng J P, Ma L H, applicant: National Astronomical observatories, CAS, patent name: methods of changing retired satellites into SIGSO, patent number: 200610055909.2, date of application: October 5, 2006, date of approval: June 3, 2009
2. Inventor: Ai G X, Shi H L, Wu H T, Yan Y H, Bian Y J, Hu Y H, Li Z G, Guo J, Cai X D, applicant: National Astronomical observatories, CAS, National Time Service Center, CAS, patent name: the transmitting satellite navigation and positioning system, patent number: 20041004760264.1, date of application: June 3, 2004, date of approval: June 5, 2009
3. Inventor: Li Z G,Ai G X,Shi H L, Yang X H, applicant: National Astronomical observatories, CAS, National Time Service Center, CAS, patent name: a new method for determination of satellite orbits by transfer, patent number: 200310102197.1, date of application: December 30, 2003, date of approval: July 5,2006
4. Inventor: Shang J N, Shi H L, Lu X C, Zhou X applicant: National Astronomical observatories, CAS, National Time Service Center, CAS, patent name: integrated system of navigation and communication based on PT and status messages, patent number: 200610012056.4, date of application: May 31, 2006
5. Inventor: Shi H L, Yang P C, Liang K, Pei J, Cui J X, Ji Y F, Li S M, Ning C L applicant: National Astronomical observatories, CAS, patent name: parasitic anti-interference method and system for transmitting satellite navigation signal, patent number: 200710081260.6, date of application: May 10, 2007
6. Inventor: Sun X Y, Ai G X, Shi H L, Ji Y F, Wang X N, Pang F, Ning C L, applicant: National Astronomical observatories, CAS, patent name: method of realizing satellite navigation and positioning with different navigation frequency, patent number: 200810056250.1, date of application: January 16,2008
7. Inventor: Shi H L, Cui J X, Han Y B, Pei J, Du X H, Hu Z Q, Ji Y F, Wang M, Zhu H L, applicant: National Astronomical observatories, CAS, patent name: signal structure for integrated system of navigation and communication, patent number: 200910084033.8, date of application: May 13, 2009
8. SHI HuLi, AI GuoXiang, HAN YanBen, MA LiHua, CHEN JiBin, GENG JianPing. Multi-life cycles utilization of retired satellites. Science in China Series G: Physics, Mechanics & Astronomy, 2009, 52(3):323-327.
9. HAN YanBen, MA LiHua, QIAO QiYuan, YIN ZhiQiang,SHI HuLi, AI GuoXiang, Function of retired GEO communication satellites on improving PDOP value of CAPS, Science in China Series G: Physics, Mechanics &Astronomy, 2009, 52(3):423-433.
10. Ai G X, Shi H L, Wu H T, et al. Positioning system based satellite communication and Chinese Area Positioning System (CAPS). Chin J Astron Astrophys, 2008, 8(6): 611—635.
11. AI GuoXiang, SHI HuLi1, WU HaiTao2, LI ZhiGang2 & GUO Ji2. The principle of the positioning system based on communication satellites, Science in China Series G: Physics, Mechanics & Astronomy, 2009, 52(3):472-488.
12. Shi H L, Han Y B, Ma L H, exploration and practice on multi-life cycles utilization of GEO satellites, Satellite and Network, 2007, (8): 34-37

13. HAN YanBen, MA LiHua, QIAO QiYuan, YIN ZhiQiang, AI GuoXiang, Selection of satellite constellation framework of CAPS, Science in China Series G: Physics, Mechanics & Astronomy, 2009, 52(3):458-471.
14. CUI JunXia, SHI HuLi & CHEN JiBin. Transmission Link of CAPS Navigation and Communications system. Science in China Series G: Physics, Mechanics & Astronomy, 2009, 52(3):402-411.
15. LI Zhigang, YANG Xuhai, SHI Huli, AI Guoxiang, QIAO Rongchuan & FENG Chugang. New Method for Determination of Satellite orbit by Transfer. Science in China Series G: Physics, Mechanics & Astronomy, 2009, 52(1), in press
16. Huang Yong, Hu XiaoGong, Huang Cheng, Yang QiangWen & Jiao WenHai. Precise orbit determination of a maneu-vered GEO satellite using CAPS ranging data. Science in China Series G: Physics, Mechanics & Astronomy, 2009, 52(1), in press
17. SHI Huli & PEI Jun. The solutions of navigation observation equations for CAPS. Science in China Series G: Physics, Mechanics & Astronomy, 2009, 52(3):434-444.
18. YANG Xuhai, LI Zhigang, FENG Chugang, GUO Ji, SHI Huli, AI Guoxiang, WU Fenglie & QIAO Rongchuan. Rapid method of orbit forecasting after maneuver for geostationary satellite. Science in China Series G: Physics, Mechanics & Astronomy, 2009, 52(1), in press
19. Shi H L, Cui J X, Du X H, one kind of satellite communication technique with low information rate, China Science and Technology Achievements
20. Shi H L, Sun X Y, Li Z G, principles on the transmitting satellite navigation and positioning system, Beijing: Science Press, 2009
21. Ma L H, Han Y B, Qiao Q Y, Some questions about Satellite static positioning, Global Positioning System, 2005, 4: 9—11.
22. Ma L H, Han Y B, Qiao Q Y, calculating DOP of artificial satellites with MATLAB, Application Research of Computers (supplement), 2007, 24: 1716.
23. Deng C M, Zhuo C N, Wu T Y, Constellation of a Regional Satellite Navigation System, JOURNAL OF UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA, 2006, 35(5): 725—728.
24. International Telecommunication Union. Recommendation ITU-R S.484-3. Station-keeping in longitude of geostationary satellites in the fixed-satellite service Geneva: International Telecommunication Union Electronic Publishing Service.
25. International Telecommunication Union, Recommendation ITU-R S.743-1, ITU-R Recommendations on CD-ROM, Electronic Publishing Service, 2004.
26. Kaplan E D, Hegarty C J. Understanding GPS: Principle and Application. 2nd ed. Boston: Artech House, 2006.
27. Hanson J M, Evans M J, Turner R E. Designing good partial coverage satellite constellations. J Astr Sci, 1992, 20(2): 215—239.
28. IADC/WG4/WGD01, Draft IADC space debris mitigation guideline, IADC/WG4/ WGD01, 2001.
29. Jing Y X, Zhang Y T, Analyzation of GEO satellite phase, CAPS Technical report. 2009, 12
30. Li Y H, principle and implementing strategy for position maintenance of GEO satellites, Journal of Spacecraft TT&C Technology, 2003, 22(4): 53-61.
31. Liu L, Hu S J, Wang X, an introduction of astrodynamics, Nanjing University Press, 2006
32. Yang J X, orbits dynamics and control of spacecraft, China Astronautic Publishing House, 1995.

33. Zhang R W, orbits dynamics and control of satellite, BEIHANG University
34. Hu S J, Chen L, Liu L, The Structure Evolution of Satellite Constellation, Acta Astronomica Sinica, 2003, 44(1): 46—54.
35. Zheng J, Chen H. Study on GEO Satellite's Re-Orbit Control Strategy in End of Life, AEROSPACE SHANGHAI, 2007, 3: 38—41
36. Wu B Y, Consideration on Inclined GEO Satellite Communication System, TELECOMMUNICATIONS SCIENCE, 2002, 11: 7—11.
37. Xu, Q F, Satellite Constellation of Local Navigation System, ENGINEERING OF SURVEYING AND MAPPING, 2001, 10(1): 1—5.
38. Liu H J, Zhang N T, CONSTELLATION DESIGN FOR REGIONAL SATELLITE POSITIONING SYSTEM, JOURNAL OF ELECTRONICS & INFORMATION TECHNOLOGY, 2001

17

Planar Antennas For Satellite Communications

Jorge Sosa-Pedroza, Fabiola Martínez-Zúniga and Mauro Enciso-Aguilar
Instituto Politécnico Nacional. Escuela Superior de Ingeniería Mecánica y Eléctrica
México

1. Introduction

Modern Satellite Communications requires the development of very small size, low-cost, low-profile, high gain and high directivity antennas. Antennas form the link between transmitting and receiving equipment and the space propagation path. Even the general principles may apply to satellite communications antennas, imposed limitations of gain, field pattern and mainly the physical environment lead to design requirements which should be taken in account. Most satellite antennas are designed to give coverage over a specified restricted area defining restrictions over antenna gain; on the other hand payload restrictions of weight and size lead to small size and weight antennas. Especially for these characteristics, planar antennas could have wide application on communications satellites. Since the late 1970s, the international antenna community has devoted much effort to the theoretical and experimental research on microstrip and printed antennas, which offer the advantages of low profile, compatibility with integrated circuit technology and conformability to a shaped surface. The results of this research have contributed to the success of these antennas not only in military applications such as aircraft, missiles, and rockets but also in commercial areas such as mobile satellite communications, the Direct Broadcast Satellite (DBS) system, global position system (GPS) and remote sensing but their main applications are on ground transceivers. Last two decades have been especially worthwhile on planar antennas development, mainly for mobile communications, but its applications should be extended to satellite systems. Particularly for the first Mexican satellite system (Morelos I and II), primary array receiving Ku band , was a planar array to produce an elliptical footprint over Mexican territory. For the second generation of Mexican satellites, was introduced the L band circular polarization array, formed by 16 parabolic reflectors feed by crossed short dipoles producing the needed circular polarization and also an elliptical footprint. It is clear that if there were used a planar array, as those that technology has developed in the last two decades, the cost and the weight would be reduced to a fourth part. This chapter of the Satellite Communications book, is devoted to planar antennas, not only for that already in use but proposing other kind that could be applied for satellite communications. It starts with a brief description of planar antennas, their characteristics, and their applications as antenna arrays. The next part will describe actual planar antennas used in satellite communications systems and will finish with a proposal of new developments of planar antennas that could be used in the near future.

1.1 A Brief History

Historically microstrip antennas and consequently planar antennas, have associated with low cost and low profile, but this simplistic description is inadequate as state in (James et al.,

1989), considering that the feasibility of a low profile printed radiator has inspired the system creator with abundance of examples, as the printed paper antennas, or adaptive conformal antenna arrays and many other examples. The success on cellular telephony is based in much in the planar antenna development, as the services offered by cellular company's increase every day, planar ultra wideband antennas are an actual challenge for those working in the area. Research and engineering publications are also devoting more space to development in planar antennas; one can open any proceedings on antennas and will find at least one or two articles related with this subject. Satellite communications are not outside of this boom, especially when, as in no other industry, low weight, low profile and low cost are a daily challenge for antenna design. The invention of the microstrip concept has been attributed to many sources and the earliest include (Greig et al., 1952), and Deschamps (1953), the idea was to create a new way for connecting electronic circuits even knowing the high unwanted radiation, leading to reduce both, substrate and conducting strip. Whether the advent of the transistor influenced the rapid development of printed circuits, the main interest was the development of low cost microwave circuits (James et al., 1981). There were few proposals to use the technique for antennas in those early years, until the 1970's when the new generation of missiles create the need of low profile antennas, after that the development of microstrip and planar antennas started to fill up research publications. A key point on the design of microstrip and planar antennas, was the development of substrates, not only to define characteristics of permittivity and low loss tangent but also to make materials capable to work on extreme ambient conditions; the 80's was the decade when bigger steps were done in substrate design, however as the frequency demand increases in communication systems, the development of new materials has lead to manufacturers in tightened their specifications, and reducing substrate costs. Table 1 shows examples of materials actually used in microstrip and planar antennas.

Type	Composition	Thickness	ϵ_r	$\tan\delta$
RO3010	PTFE Ceramic	0.005"-0.050"	10.2±0.30	0.0023
RO3206	PTFE Woven Glass	0.005"-0.050"	6.15±0.15	0.0027
RO4350B	Hydrocarbon Ceramic	0.004"-0.06"	3.48±0.05	0.0037
RT/Duroid 5880	PTFE Glass Fiber	0.005"-0.125"	2.20±0.02	0.0009
RT/Duroid 6006	PTFE Ceramic	0.005"-0.100"	6.15±0.15	0.0027
TMM 3	Hydrocarbon Ceramic	0.15"-0.125"	3.27±0.032	0.0020
TMM 10	Hydrocarbon Ceramic	0.015"-0.100"	9.20±0.23	0.0022
FR-4	Woven Glass/Epoxy Resin	0.059"-0.118"	4.8 and 2.2	0.017

Table 1. Typical substrate materials for planar antennas

Planar antennas have influenced other fields in electromagnetics, those related with materials and the development of transmission lines; actually engineers include the feeders as a part of the design of patch antennas to get a better control over the input impedance characteristics, this is even more important for large arrays architectures where the feeders and antennas are regarded as a complete entity, emphasizing the importance of choice of array topology and the fact that feeders cannot necessarily be freely attached to printed elements. On the other hand, analytical development of antennas is no longer applied; the actual printed substrate technology uses mathematical models as the Finite Difference on Time Domain (FDTD) (Yee, 1966) or the Method of Moments (Harrington, 1992) and others, managed by computer simulators, reducing the highly mathematical analysis to a drawing in a CAD system, which determines current distribution, and after that, radiation patterns, gain, feeder impedance and

other parameters of antennas and antenna arrays. Computational Electromagnetics is a field in constant development, seeking for better computer characteristics and computer architecture, as the parallel computer systems, as well for better programming efficiency.

2. Microstrip Antennas

Microstrip antennas are very attractive to be used as radiators in satellite communications systems because of its several interesting features like low profile, light weight, easy fabrication, robust nature, conformability to mounting hosts, compatibility with microwave monolithic integrated circuits (MMICs) and optoelectronic integrated circuits (OEICs) technologies, they are very versatile in terms of resonant frequency, polarization, radiation pattern and impedance Microstrip antennas were first introduced in the mid of 70s (Howell, 1975; Munson, 1974) since then a really intense research activity on microstrip antenna has been taken place around the world for both academic and industrial entities. The basic configuration of a microstrip patch antenna is shown in Fig. 1.

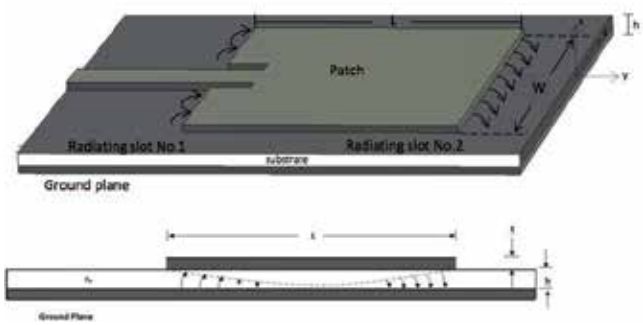


Fig. 1. Microstrip patch antenna configuration

It consist of a dielectric substrate sandwiched structure. The upper surface supports a printed conducting radiator strip or *patch* (even if a microstrip antenna is essentially different to a *patch* antenna, we will use both terms indifferently in this subsection) which is well contoured while the entire lower surface of the substrate is backed by the ground plane. The microstrip patch is designed in a such way that its maximum radiation pattern is normal to the surface of the patch (broadside radiator). This can be achieved by choosing the mode of excitation beneath the patch (Balanis, 2005). The patch radiates from fringing fields around its edges. Impedance match occurs when a patch resonates as a resonant cavity. When matched, the antenna achieves peak efficiency. For microwave applications, thinner substrates with higher dielectric constants are desirable because they require tightly bound field to decrease parasitics effects like undesired radiation and coupling. This also leads to smaller element sizes.
The shape of the patch can take several geometrical forms like square, rectangular, circular, triangular, elliptical, linear (thin strip dipole) or any other configuration. The most common patch shapes are square, rectangular, and circular due to its simplicity of design, manufacture and analysis. They are also attractive because of their radiation characteristics, especially for low cross-polarization radiation. Linear and circular polarization can be obtained by either one element or an array of microstrip antennas. Rectangular patches tend to have the largest impedance bandwidth, simply because they are larger than the other shapes. Circular and elliptical patches are slightly smaller than their rectangular counterpart and as a result have

slightly lower gain and bandwidth. One of the primary reasons the circular geometry was intensively investigated in the past was because of its symmetry allowed full-wave analysis tools utilizing a spectral domain technique. Circular patch antennas provide circularly polarized conical patterns for effective data transmission from a satellite to the earth at UHF and S-band frequencies. In the following sections we shall briefly introduce the theoretical framework of only two types of patch: rectangular and circular. Full wave methods like Method of Moments and FDTD along computational tools are discussed later in this chapter.

3. Main Figures of Merit of Microstrip Antennas

3.1 Resonant Frequency

In general, microstrip antennas are half-wavelength structures and are operated at the fundamental resonant mode TM_{01} or TM_{10}, with a resonant frequency given by (James et al., 1989):

$$f_r = \frac{c}{2L\sqrt{\epsilon_r}} \tag{1}$$

The above expression is only valid for a rectangular microstrip antenna with a thin microwave substrate, where c is the speed of light, L is the patch length of the rectangular microstrip antenna, and ϵ_r is the relative permittivity of the grounded microwave substrate. From 1 it is clear that the microstrip antenna has a resonant length inversely proportional to $\sqrt{\epsilon_r}$, hence the use of a microwave substrate with a larger permittivity thus can result in a smaller physical antenna length at a fixed operating frequency.

If the open-end extension due to fringing effect is considered then the resonant frequency for a rectangular microstrip antenna becomes (Hirazawa et al., 1992; James et al., 1989):

$$f_r \cong \frac{c}{2L_{eff}\sqrt{\epsilon}_{eff}} \tag{2}$$

where:

$$L_{eff} = L\left\{1 + 0.824\frac{h}{L}\frac{(\epsilon_{eff}+0.3)(L/h+0.262)}{(\epsilon_{eff}-0.258)(L/h+0.813)}\right\} \tag{3}$$

$$\epsilon_{eff} = \frac{\epsilon_r+1}{2} + \frac{\epsilon_r-1}{2}\left(1+10\frac{L}{h}\right)^{-\frac{1}{2}} \tag{4}$$

with h being the substrate thickness. If the microstrip antenna is circular then its resonant frequency is approximated by:

$$f_r \cong \frac{\chi_{11}c}{2a_{eff}\sqrt{\epsilon}_r} \tag{5}$$

where

$$a_{eff} = L\left\{1 + \frac{2h}{\pi a\epsilon_r}\ln\frac{\pi a}{2h} + 1.7726\right\}^{\frac{1}{2}} \tag{6}$$

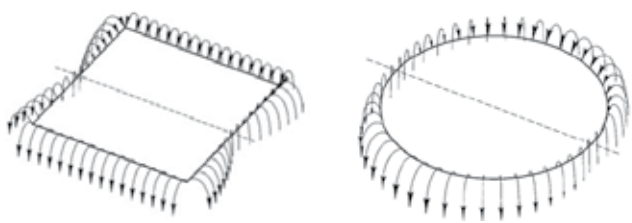
and

$$\chi_{11} = 1.841 \tag{7}$$

a being the radius of the patch.

3.2 Radiation Pattern

It is quite hard to talk on radiation pattern characteristics without associate it with the internal construction consideration. Here we only briefly discuss the associated internal structures that affect the pattern. Rectangular and circular are the most common shapes for patch and they radiate similar patterns. When the cavity is loaded in order to shrink its size, it radiates wider beamwidth patterns that decreases directivity. Structures that couple to coplanar patches to increase the impedance bandwidth will radiate narrower beams, but the basic patch has a wide beamwidth. If multiple coplanar patches are coupled, one can expect the pattern to narrow or vary its shape as the mixture of modes on the various patches changes over the frequency range of operation(Milligan, 2005). Radiation pattern of microstrip antennas have been investigated by several techniques including the equivalent magnetic current method (Derneryd, 1976; James et al., 1989; Long, 1978), full-wave analysis (Chang, 1989; Itoh et al., 1981), and method of moments (Agrawal et al., 1977; Chang, 1989). The former has been often used because its procedural simplicity and extensive applicability. This method considers magnetic current sources $\mathbf{K}_{m1}$ and $\mathbf{K}_{m2}$ laying at the edge or within the vicinity of the edge of the patch that can be approximated by a magnetic current loop antenna. The radiation pattern can be determined by using a Hertz or potential vector. Fig. 2 depicts the fringing electric fields around the edges of square and circular patch antennas excited in the lowest-order cavity modes. The arrow sizes indicate the magnitude of the fields. As seen the square patch exhibit nearly uniform fields along two edges called the width, and a sinusoidal evolution along the other two edges, called the resonant length. The fields vanish along a virtual electrically short-circuited plane halfway across the patches. On either side of the short-circuit plane, the fields are directed in opposite directions.
The circular patch fringing fields distribution varies as $\cos\phi$, where the angle ϕ along the edge is measured from the peak electric field. Magnetic currents found from the fringing electric fields can replace the electric currents located on the patch and the surrounding ground plane for pattern analysis.



(a) Square (b) Circular

Fig. 2. Fringing electric fields around microstrip patches: (a) square; (b) circular.(After (Milligan, 2005))

The source of equivalent magnetic current corresponding to the E_z field component (along the vertical direction) of a dominant mode is expressed taking the image component as $\mathbf{K}_m = 2\mathbf{E}_z \times \mathbf{n}$, $\mathbf{n}$ being a unity normal vector. The pattern is determined by the line integration of $\mathbf{K}_m$ at the rim of the patch over the adjacent source region (James et al., 1989). For a circular patch with a TM_{110} dominant mode, the expressions to approximate the pattern are: (Hirazawa et al., 1992; James et al., 1989; Long, 1978):

$$E_\theta = -jK_d a\cos(\phi)\sin(m)\left\{\frac{J_1'(n)}{\cos\theta}\right\} \tag{8}$$

$$E_\phi = \frac{jK_d}{k_0}\sin(\phi)\cos(m)\left\{\frac{J_1(n)}{\cos\theta}\right\} \tag{9}$$

with

$$m = k_0 h\cos\theta;\quad n = k_0 a\sin\theta;\quad K_d = E_0\frac{e^{-jk_0R}}{R}J_1(ka) \tag{10}$$

For a rectangular patch with a dominant TM_{100} mode the expression for the radiation pattern are given by:

$$E_\theta = -jK_r f(\theta,\phi)\cos\phi\left\{\frac{\epsilon_r - \sin^2\theta}{\epsilon_r - (\sin\theta\cos\phi)^2}\right\} \tag{11}$$

$$E_\phi = jK_r f(\theta,\phi)\cos\theta\sin\phi\left\{\frac{\epsilon_r}{\epsilon_r - (\sin\theta\cos\phi)^2}\right\} \tag{12}$$

with:

$$f(\theta,\phi) = \left\{\frac{\sin(m)}{m}\right\}\cos n;\ m = \frac{k_0 L}{2}\sin\theta\sin\phi;\ n = \frac{k_0 L}{2}\sin\theta\cos\phi;$$
$$K_r = \frac{V_0 k_0 L}{\pi}\frac{e^{-jk_0R}}{R};\ V_0 = E_0 h \tag{13}$$

where k_0 is the wave number in free space, k is the wave number in the substrate, $J(n), J'(n)$ first kind of Bessel function and its derivative, respectively. E_0 is an arbitrary constant, h is the substrate thickness, a is the radio of the circular patch , L is the patch length of the rectangular shape.

3.3 Quality Factor, Radiation Efficiency, Bandwidth and Gain

The quality factor, bandwidth, and efficiency are antenna figures-of-merit, which are interrelated, and there is no complete freedom to independently optimize each one.The quality factor Q of a microstrip antenna becomes one of the most important design parameters. Quality Factor is a measure of the radiation loss in the microstrip antenna. For a given substrate characteristics like ϵ_r, $\tan\delta$, σ and a radiation pattern, the unloaded quality factor Q_0 and the radiation efficiency is (James et al., 1989; Long, 1978):

$$\frac{1}{Q_0} = \frac{1}{Q_r} + \frac{1}{Q_c} + \frac{1}{Q_d} \tag{14}$$

where

$$\eta = \frac{Q_0}{Q_r};\quad Q_d = \frac{1}{tan\delta};\quad Q_c = \frac{h}{\delta_s};\quad \delta_s = \frac{1}{\sqrt{\pi f\mu_0\sigma}};\quad Q_r = \omega\frac{W_T}{P_{rad}};\quad W_T = \frac{1}{2}\epsilon\int_v |E|^2 dv \tag{15}$$

and

$$P_{rad} = \frac{1}{2}\Re\left\{\int_{sh}((E\times H)^*\cdot nds)\right\} = \frac{1}{2Z_0}\int_0^{2\pi}\int_0^{\frac{\pi}{2}}(|E_\theta|^2 + |E_\phi|^2)R^2\sin\theta d\theta d\phi \tag{16}$$

Q_r being the quality factor Q of radiation loss, δ_s skin depth, Q_c is Q of conductor loss, Q_d Q of dielectric loss, P_{rad} radiated power, P_{in} input power, Z_0 free space intrinsic impedance, $\Re$ real part, Sh sphere. Those expression agree well with measured values of Q_0 and η for common design ranges as long as the substrate is not thick enough to excite the surface wave and higher order modes (Hirazawa et al., 1992). In the general case the bandwidth of the antenna is expressed as:

$$BW = \frac{(VSWR - 1)}{Q_0\sqrt{VSWR}} \tag{17}$$

Then the BW of a practical antenna is determined for a given value of Q_0 and $VSWR$.

3.4 Directivity Gain

Directivity gain can be obtained by integrating the radiation pattern over an arbitrary surface containing the antenna. A useful expression to compute the directivity gain is:

$$G_d(\phi,\theta) = \frac{|E(\theta_0,\phi_0)|^2}{\frac{1}{4\pi}\int_0^{2\pi}\int_0^{\pi}|E(\theta,\phi)|^2\sin\theta d\theta d\phi} \tag{18}$$

where

$$|E(\theta,\phi)|^2 = |E_\theta(\theta,\phi)|^2 + |E_\phi(\theta,\phi)|^2 \tag{19}$$

and E_θ and E_ϕ are given by (8,), (9), (11,) and (12,)

3.5 Impedance Characteristics

Impedance characteristics of a microstrip antennas has been treated mainly from two points of view: self impedance and auto-impedance. Here the basic procedures for computing the impedance characteristics are briefly described.

3.5.1 Input Impedance

The Input impedance or self-impedance of a microstrip antennas has been studied by several models like Transmission Line Model (Derneryd, 1976; James et al., 1989) cavity model (Richards et al., 1979; 1981) and full wave models that includes several numerical techniques like method of moments (Gupta et al., 1988; Pozar, 1982) finite element (Volakis et al., 1998) and finite difference time domain method (FDTD) (Reineix et al., 1989; Sheen et al., 1990) among others. Transmission Line Model has been widely used because of its simplicity to describe the impedance of a general microstrip and provides a good physical insight, however it may lack in accuracy. In contrast, the cavity model is more accurate but at the same time, more complex although it also provides a good physical insight. Full wave models are in general very accurate and versatile and can manage a wide number of elements, complex patch shapes and coupling problems, but they are more complex models with less physical insight. Here we shall discuss the transmission line model only. In general, input impedance is complex and includes both a resonant an nonresonant components which is commonly reactive. Both real and imaginary parts of impedance vary with frequency. Ideally the reactance and the resistance exhibit symmetry about resonant frequency. At resonance, the reactance is equal to the average of the addition of its maximum and minimum values.

3.5.2 Transmission Line Model

The simplest analytical description of a rectangular microstrip patch utilizes transmission-line theory and models the patch as two parallel radiating slots (Munson, 1974). Each radiating edge of length L (Fig. 3) is modeled as a narrow slot radiating into a half-space,with a slot admittance given by (Harrington, 1961)

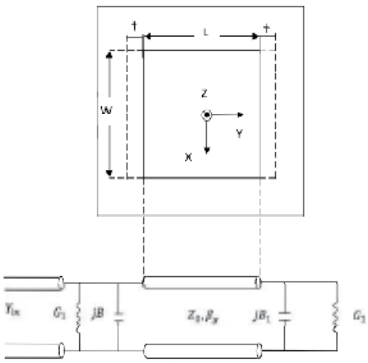


Fig. 3. Top view of a rectangular patch antenna and its associated transmission line equivalent circuit

$$G_1 + jB_1 \cong \frac{\pi q}{\lambda_0 \eta_0}[1 + j(1 - 0.636 \ln k_0 w)] \tag{20}$$

where λ_0 is the free-space wavelength, η_0 the intrinsic impedance of the free-space, $k_0 = \frac{2\pi}{\lambda_0}$ and w is the slot width, approximately equal to the substrate thickness h. Assuming that the field remains constant along the direction parallel to the radiating edge, the characteristic impedance is given by:

$$Z_0 = \frac{1}{Y_0} = \frac{t\eta_0}{L\sqrt{\epsilon_r}} \tag{21}$$

and the propagation constant is approximated by :

$$\beta_p \cong k_f \sqrt{\epsilon}_{eff} \tag{22}$$

where k_f is the propagation constant in free space. ϵ_{eff} is the effective dielectric constant given by:

$$\epsilon_{eff} = \frac{\epsilon_r + 1}{2} + \frac{\epsilon_r - 1}{2}\left(1 + 12\frac{h}{L}\right)^{-\frac{1}{2}} \tag{23}$$

The conductive component G_1 and the reactive component B_1 in (20) are related to the fringing effect and the radiation loss, and are approximated respectively by:

$$B_1 = \frac{k_f \Delta l}{Z_0}\sqrt{\epsilon_{eff}} \tag{24}$$

$$G_1 \approx \begin{cases} \frac{L^2}{90\lambda_0^2}, & L < 0.35\lambda_0 \\ \frac{L}{120\lambda_0} - \frac{1}{60\pi^2}, & 0.35\lambda_0 \leq L \leq 2\lambda_0 \\ \frac{L}{120\lambda_0}, & 2\lambda_0 < L \end{cases} \tag{25}$$

where Δl denotes the line extension due to fringing effect. This value can be expressed by:

$$\Delta l \approx 0.412t\frac{(\epsilon_{eff}+0.3)}{(L/h+0.264)} \tag{26}$$

The input admittance of the patch antenna on Fig. 3 can be treated as two slot antennas interconnected by a transmission line having characteristic admittance and propagation constant approximated by (20) and (22):

$$Y_{in} = G_1 + jB_1 + Y_0\frac{(G_1+jB_1)+jY_0\tan(\beta_p L)}{Y_0+j(G_1+jB_1)\tan(\beta_p L)} \tag{27}$$

When the average stored magnetic and electric energies are equal then resonance occurs (Pozar, 2005), under these conditions the imaginary part of the admittance must vanish.

$$\Im\{Y_{in}\} = 0 \tag{28}$$

leading to:

$$\tan(\beta_p L) = \frac{2Y_0 B_1}{G_1^2 + B^2 - Y_0} \tag{29}$$

This condition is used to determine the resonant frequency for a given patch length L or inversely to establish the resonant leght L for a given resonant frequency. The input admittance at resonance is:

$$Y_{in} = 2G_1 \tag{30}$$

3.6 Arrays and feed networks

Even that most of the planar antennas are low and medium gain, wide pattern antennas, its versatility and mainly its low profile make them very popular in arrays. Its use to synthesize a required pattern is a nowadays application in many communication systems; most of the microstrip arrays are designed for fixed-beam broadside applications, increasing directivity and performing other functions, as for scanning purposes; those functions are difficult to obtain with a single radiator or even other classes of antennas; often the feed network is located coplanar with the array elements. The main versatility of planar antenna arrays is that the feed network usually is part of the printed environment; the elements can be feed by a single line in a series-feed network, or by multiple lines as a corporate-feed network (Balanis, 2005), including both parallel feed and hybrid series/parallel feed. Series-feed arrays are usually built using photolithography for both the radiating elements and the feed network, reducing its application to fixed beams, linear or planar arrays with single or dual polarization. In this configuration multiple elements are arranged linearly and feed serially by a single transmission line, multiple linear arrays can then be fed either serially or parallel to form a two dimensional planar array (Lee, 1997). Series-feed line can be in-line feed when all radiators are arranged in the same line or out-of-line feed when elements are parallel each other. The series-feed array can be classified as resonant and traveling-wave. For the first kind, impedances at the transmission line junctions and the patch elements are not matched and elements are spaced multiple integers of one wavelength apart creating a broadside beam, with a very narrow resonant bandwidth, around 1% . For the traveling wave array type, transmission line and elements are matched, with spacing of one wavelength for broadside or less for off-broadside radiation. The traveling current wave over the feeding line moves almost without reflections reducing almost to zero the energy at the end of array which can be either

absorbed by a matching network or reflected back to be reradiated in phase. Corporate feed arrays, also called parallel feed arrays, are more versatile and the designer has more control of the phase and amplitude of each feeding element, making them ideal for scanning phased arrays, multibeam arrays, or shaped-beam arrays.

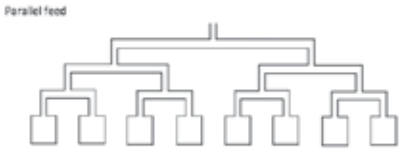


Fig. 4. Parallel feed configuration for microstrip arrays. After (Lee, 1997)

Fig. 4 shows how the patches are feeding by power division lines. Transmission lines divides into two branches and each branch divides again until it reaches the patch elements. For a broadside array all the divided lines are of the same length. A disadvantage of parallel feeding is that the insertion loss is higher than that of a series feeding, however is less affected by phase changes due frequency changes, because relative phases between all elements will remain the same. The corporate feed array can achieve a bandwidth of 15% or more, depending on the design.

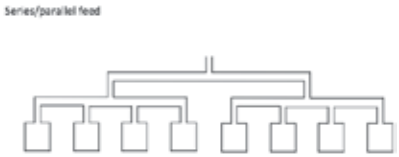


Fig. 5. Series-Parallel feed configuration for microstrip arrays. After (Lee, 1997)

The hybrid series/parallel array is shown in Fig. 5 where a combination of series and parallel feed lines are used, feeding elements in this way is possible to have a wider bandwidth than the series feed array but with a higher insertion loss due the parallel feeding, but the technique allows to make design trade-offs between bandwidth and insertion loss.
A microstrip can be designed in a single layer or multilayer configuration, decision is related with complexity and cost, sidelobes, cross polarization, bandwidth and other factors. Using a single layer reduces manufacturing costs but other characteristics are degraded. When low sidelobe or cross polarization is needed the double layer design seems to be the better choice. Fig. 6 shows a multilayer of a dual polarization antenna, where the antenna feeding is obtained with crossed slots on the ground plane and the feeding of the two polarizations is obtained using two orthogonal microstrip lines. A disadvantage of planar arrays is the influence between elements and feed lines, which affects the performance of the others, then in the design is very important to take into account effects as mutual coupling and internal reflections; coupling between elements generates surface waves within the substrate which can be eliminated using cavities in conjunction with microstrip feeding elements, but these effects are difficult to analyze for common analytical methods, therefore for accurate results should be used full wave solutions as those presented in this chapter, applied in most of the actual computational tools of present days.

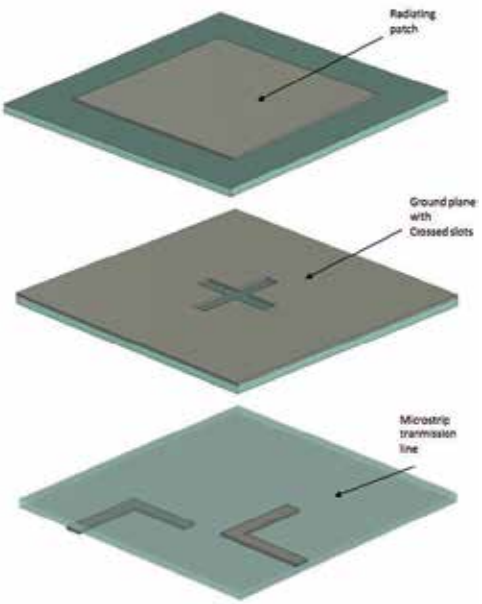


Fig. 6. Multilayer dual-polarized microstrip patch element. After (Lee, 1997)

4. Computational Tools

At the beginning of 80's, when the boom of planar antennas started the only available computational tools where the CAD packages, but their applications were constrained to low frequency electronic systems (Harrington, 1992) and the microwave programs were limited and expensive. The situation has changed radically since the arrival of personal computers and mainly the permanent improvement of their characteristics, making them a powerful tool for analysis of electromagnetic problems; on the other hand, the use of mathematical models had become in the development of specialized software tools widely used nowadays in the field of planar antenna analysis and design. Although the appearance of many commercial tools is a common situation every day, many researchers use their own programs; authors have been working since some years ago in development of their own software, applying it to antenna and microstrip devices (Barrera-Figueroa et al., 2007; 2009; Sosa-Pedroza et al., 2008; 2009) considering both, saving economical resources and "learning doing" mainly for academic reasons. Most of the actual work on antenna computational methods is based on solution of Maxwell Equations in integral or differential form, Method of Moments is an example and maybe the most applied for integral Methods and Finite Elements (FE) and Finite Difference on Time Domain (FDTD) for differential methods.

4.1 Method of Moments (MoM)

Integral Methods solve Maxwell's Equations in its integral form, describing the electromagnetic problem. As the current density on the conductor is related with the Electric Field, some equations have been derived from Maxwell equations, two examples of those are Pocklington equation and Hallén Equation, both can be studied in the literature (Balanis, 2005; Kraus et al., 2002). For an arbitrary shaped wire (Sosa-Pedroza et al., 2007) as the one shown in Fig. 7, is possible to deduce Pocklington Equation given by:

$$E_s^I = -\frac{j}{\omega\epsilon}\int_s' I_s(s')\left[k^2\mathbf{s}\cdot\mathbf{s} + \frac{\partial^2}{\partial s\partial s'}\right]\frac{e^{-jkr-r'}}{4\pi r - r'} \tag{31}$$

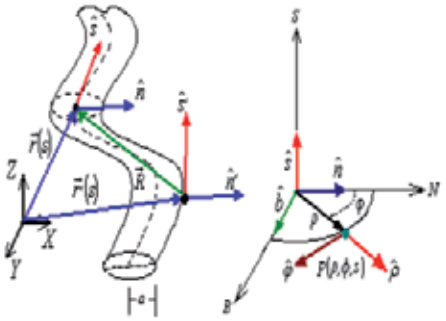


Fig. 7. Arbitrary shaped wire

E_s^I is the tangential incident electric field. As it seen the unknown distribution current is inside the integral and solution is not possible. The MoM formulation, introduced by Harrington in the 60's (Harrington, 1961), is used to get a numerical solution of (31) expressing the unknown function as a linear combination of n linearly independent functions, called basis functions:

$$I_s(s') = \sum_{n=1}^{N} c_n i_n(s') \tag{32}$$

where c_n are the unknown coefficients to be determined. Substituting (32) into (31) results the equation with N unknowns:

$$E_s^I = -\frac{j}{\omega\epsilon}\sum_{n=1}^{N} c_n \int_s w_m \int_{s'} i_n(s')\left[k^2\mathbf{s}\cdot\mathbf{s} + \frac{\partial^2}{\partial s\partial s'}\right]\frac{e^{-jkr-r'}}{4\pi r - r'} \tag{33}$$

with $m = 1, 2, ..., N$.
for a consistent equation system, is necessary to find N linearly independent equations, then is used the inner product of (33) with other set of N chosen linearly independent functions $w_m(s)$ named weight function, then

$$\int_s E_s^I ds = -\frac{j}{\omega\epsilon}\sum_{n=1}^{N} c_n \int_s w_m \int_{s'} i_n(s')\left[k^2\mathbf{s}\cdot\mathbf{s} + \frac{\partial^2}{\partial s\partial s'}\right]\frac{e^{-jkr-r'}}{4\pi r - r'} \tag{34}$$

Which can be reduced in a matrix form as:

$$\begin{pmatrix} Z_{11} & Z_{12} & \dots & Z_{1N} \\ Z_{21} & Z_{22} & \dots & Z_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{N1} & Z_{N2} & \dots & Z_{NN} \end{pmatrix}\begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_N \end{pmatrix} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_N \end{pmatrix} \tag{35}$$

where the elements Z_{mn} are obtained from:

$$Z_{mn} = -\frac{j}{\omega\epsilon}\int_s w_m \int_{s'} i_n(s')\left[k^2\mathbf{s}\cdot\mathbf{s} + \frac{\partial^2}{\partial s\partial s'}\right]\frac{e^{-jkr-r'}}{4\pi r - r'} \tag{36}$$

and the elements v_m are:

$$v_m = \int_s w_m E_s' ds \tag{37}$$

c_n represent the system's unknowns. Matrices of (35) are known as impedance matrix (Z_{mn}), current matrix (c_n), and voltage matrix (v_m). The solution for (35) is:

$$(c_n) = [Z_{mn}]^{-1}(v_m) \tag{38}$$

where the inverse matrix$(Z_{mn})^{-1}$ is obtained by a numerical technique. It is important to mention that both, basis and weight functions are arbitrary functions, selected considering computational resources and time and accuracy of solution. As is seen the Kernel of the integral includes Green's function, representing the electromagnetic influence on the entire surrounding space. Solution is valid at every point of an infinite space, including the far field radiation phenomena that are vital for antenna analysis.
Integral equations are established as a multilayered Green's function, such that the background can consist of an arbitrary number of horizontal, infinitely stretched layers, containing dielectric substrates and conducting ground planes, always present on planar antennas. The main components on the antenna are replaced with equivalent surface/volume currents, appearing as the primary unknowns in the resulting integral equations, solved by MoM. From (35) we can see that the method creates a dense matrix equation which solution gives the current distribution on the environment, after that, all other antenna parameters are easily obtained.

4.2 Finite Difference on Time Domain (FDTD)

Differential methods solve Maxwell Equations in their differential form; most of the computational solvers use the Finite Element Method (FEM) and the Finite Difference on Time Domain (FDTD).
The FDTD method, introduced by (Yee, 1966), transforms the Differential Maxwell Equations in Finite Difference Equations, used to generate an algorithm and a program code to solve them, over a specific propagation region; it uses the Yee's cell algorithm in a central difference scheme, considering field variations in time and space as the original Maxwell Equations, the characteristics of media are also defined in the method by means of ϵ, μ and σ characteristics, the algorithm is processed by a computer to analyze behavior of EM field moving over an environment of any kind. The interacting electric and magnetic fields are given by:

$$\frac{\partial H_x}{\partial t} = \frac{1}{\mu}\left[\frac{\partial E_y}{\partial z} - \frac{\partial E_z}{\partial y} - (M_{source_x} + \sigma^* H_x)\right] \tag{39}$$

$$\frac{\partial H_y}{\partial t} = \frac{1}{\mu}\left[\frac{\partial E_z}{\partial x} - \frac{\partial E_x}{\partial z} - (M_{source_y} + \sigma^* H_y)\right] \tag{40}$$

$$\frac{\partial H_z}{\partial t} = \frac{1}{\mu}\left[\frac{\partial E_x}{\partial y} - \frac{\partial E_y}{\partial x} - (M_{source_z} + \sigma^* H_z)\right] \tag{41}$$

$$\frac{\partial E_x}{\partial t} = \frac{1}{\epsilon}\left[\frac{\partial H_z}{\partial y} - \frac{\partial H_y}{\partial z} - (J_{source_x} + \sigma E_x)\right] \tag{42}$$

$$\frac{\partial E_y}{\partial t} = \frac{1}{\epsilon}\left[\frac{\partial H_x}{\partial z} - \frac{\partial H_z}{\partial x} - (J_{source_y} + \sigma E_y)\right] \tag{43}$$

$$\frac{\partial E_z}{\partial t} = \frac{1}{\epsilon}\left[\frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} - (J_{source_z} + \sigma E_z)\right] \tag{44}$$

These equations are transformed in a discrete form using the Yee algorithm, which can be solved by computational methods, as an example is presented only one of them:

$$\frac{E_x\,|^{n+1/2}_{i,j+1/2,k+1/2} - E_x\,|^{n-1/2}_{i,j+1/2,k+1/2}}{\Delta t} = \frac{1}{\epsilon_{i,j+1/2,k-1/2}} \cdot$$
$$\left\{\frac{H_z\,|^{n}_{i,j+1,k+1/2} - H_z\,|^{n}_{i,j,k+1/2}}{\Delta y} - \frac{H_y\,|^{n}_{i,j+1/2,k+1} - H_y\,|^{n}_{i,j+1/2,k}}{\Delta z}\right.$$
$$\left. - J_{source_x}\,|_{i,j+1/2,k+1/2} - \sigma_{i,j+1/2,k+1/2} E_x\,|_{i,j-1/2,k+1/2}\right\} \tag{45}$$

Yee Algorithm discretizes both time and space, represented by parameters n, i, j, k with intervals of Δt and Δ respectively. As seen in equation (45) the media characteristics are specially considered as ϵ, μ and σ which position is defined using the (i, j, k) subindex, then is possible to analyze the effects of any material at any position on the computational space. A computational code of any program language permits to know the EM behavior over the entire computational space. The FDTD and also the FE differential-equation methods are particularly suitable for modeling full three-dimensional volumes that have complex geometrical details. They are extremely efficient for smaller close-region problems involving inhomogeneous media (James et al., 1989).

4.3 Computational tools comparison

An excellent summary and comparison of actual available commercial software used on planar antenna analysis and design is presented in (Vasylchenko, 2009), they analyze 5 commercial tools and one "in house", comparing all of them in the analysis of planar antennas looking to guarantee the optimal use of each of the software packages, to study in detail any discrepancies between the solvers, and to assess the remaining simulation challenges. Even their work is not the first one on the theme, mentioning references strengthening their vision that, an extensive benchmark study over a large variety of solvers and for several structures has not yet been documented.

As the operation of EM solvers is based on the numerical solution of Maxwell's equations in differential or integral form, one or other influences the efficiency and accuracy and users may get the wrong impression that a given solver is automatically suited to solve any kind of problem with arbitrary precision. Comparison in the Vasylchenko work verifies the plausibility of such expectations by presenting an extensive benchmark study that focuses on the capabilities and limitations of the applied EM modeling theories that usually remain hidden from the antenna designer. The integral solvers they analyze are the one they designed in K. U. Leuven's: MAGMAS 3D, the others are IE3D from Zeland Software, FEKO from EM Software & Systems, and ADS Momentum from Agilent. On the other hand they analyze the two leading differential EM tools, HFSS from Ansoft for the finite-element method, and CST Microwave

Studio for the FDTD method. After a careful analysis, comparing results with measurement of 4 common planar antennas, their conclusion is as follows:
Classical patch antennas could be predicted by every simulation program with a deviation not beyond 1.5 %. The simulation based on MoM was inherently faster and are more attractive in price. On the other hand the FEM and FDTD are inherently able to analyze much more general structures, but require the inversion of much larger, but sparse, matrices, requiring higher memory resources. Although the calculation times were not that different at the time of experiment, they presented a reference in which it seems that dedicated inversion techniques for MoM solvers are nowadays fully in development, opening the possibility that better times can be obtained for differential equations solvers.
Proper mesh generation and a correct feeding model are two crucial issues predetermining the successful simulation in the software packages reviewed. In general, a very neat adaptive mesh refinement, implemented in Ansoft's HFSS and as an option in CST's MWS, allows better handling of a design with difficult electromagnetic coupling between its different parts. Such characteristics pertain to applications in mobile gadgets, such as the GSM antennas. Having no mesh refinement option, MoM-based programs require more careful consideration of the initial meshing. MoM solvers can provide an improvement in simulation results and time using so called edge-meshing features, while avoiding excessive meshing on the bulk of the metal structure. However the study concludes that the meshing schemes in all solvers are adequate.
Some designs, such as the GSM and UWB antennas, require finite substrate effects to be taken into account, such as diffraction from substrate edges. MoM based solvers show better convergence when a dielectric substrate is infinite, but the trend toward miniaturizing antennas diminishes the advantage of using these solvers, then they conclude that at present, differential equations programs are better suited for modeling small antennas. On the other hand(Vasylchenko, 2009) suggest that the feeding models, as implemented today in the widespread commercial 35 solvers, are probably unsatisfactory in the case of small structures with complicated electromagnetic-coupling behavior, but HFSS and CST MWS solvers are better suited to handle the problem.
As a final guideline, authors recommend the use of two different solvers, based on different theoretical methods (integral and differential), to characterize a specific device if both results are in good agreement, it is reasonable to expect that the results can be trusted, if the two results are in disagreement, a deeper investigation of the structure and its modeling is absolutely necessary.

5. Planar antennas on space applications

When a designer decide to use planar or microstrip antennas on a space applications should take in account three factors among those related with the inherent design of the radiator (Lee, 1997); those factors are critical and need to be considered. One is that the antenna must be able to support the high vibration produced during the launch from the Earth; acceleration can be as high as 10 Gs or more, under this conditions soldering junctions and laminating of multilayer antennas tend to breakdown, then they should be made strong enough to survive the vibration, a solution could be the use of noncontacting feeds as proximity, capacitive or aperture coupling. The second factor is related with the extreme temperature difference which can be as high as 100°C to -70°C, whether the antenna "sees" the sun or not, behind a shaded area. Under this condition, the laminating adhesive material must survive physically and electrically into this environment. Third factor is the space vacuum, as is known at low

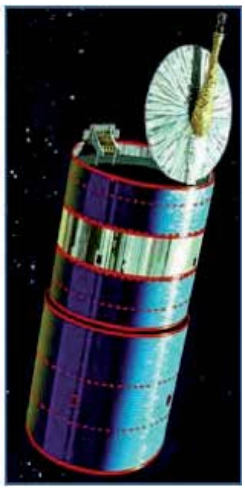
pressures, electrons are almost free to leave an electrode and move across to the opposite electrode, a phenomenon known as multipacting. For a microstrip antenna, the two electrodes are the patch and the ground plane, when the phenomenon is present reduces the capacity of power handling of the antenna then it should be designed with the proper thickness. These three factors limit the use of planar and especially microstrip antennas, nevertheless there are many examples of spacecrafts which can be mentioned: Earth Limb Measurements Satellite, Shuttle Imaging Radar, Geostar system and especially the Mars Pathfinder using a small X band microstrip antenna providing circular polarization with a peak gain of 25 dB. Antenna was constructed with a parallel feed power divider and electromagnetically coupled dipoles. The divider and the dipoles were printed on multilayer honeycomb substrates which have open vented cells for space applications.

5.1 Morelos: First Mexican Satellite System

Historically the first satellites using planar antennas could be the Mexican Morelos System, constructed by Hughes Aircraft Company (Satmex, 2010);. They were launched on the space Shuttle in June 17 and November 27, 1985 and they were the first in use the HS-376 platform as a hybrid satellite operating in two frequency bands (C and Ku) simultaneously. The four Ku-band channels used the planar arrays for reception only having a bandwidth of 108 MHz with a minimum effective isotropic radiated power (EIRP) of 44 dBW throughout Mexico. Transmit and receive beams in the C-band and the transmit beams in the Ku-band were created by a 1.8 m wide shared aperture grid antenna with two polarization-selective surfaces. The front surface was sensitive to horizontally polarized beams and the rear was sensitive to vertically polarized beams. Separate microwave feed networks are used for the two polarizations. Fig. 8(a) shows the spacecraft with the planar array and Fig. 8(b)the antenna and the reflector in the construction bay. Morelos Satellites were a very successful communications system; Morelos 1 exceeded his life from 9 years to 10, when it was substituted in 1996 for the first satellite of 2^{nd} generation of Mexican satellites, but Morelos 2 was in operation until to 2002, almost doubling its life designed time.

5.2 The IRIDIUM Main Mission Antenna Concept

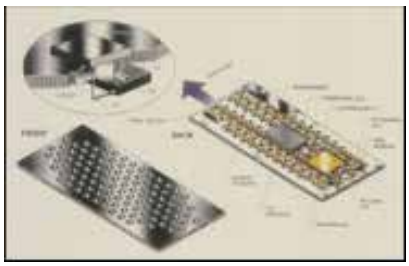
A commercial satellite system using planar antennas is the MOTOROLA's IRIDIUM (Schuss et al., 1990) shown in Fig. 8(c) used for personal satellite communications with a constellation of 66 satellites placed in low earth orbit, positioned in six polar orbital planes with 11 satellites plus one spare per plane. The main mission antenna (MMA), consists of three fully active phased-array panels providing the band link from the satellite to the ground user. Each phased-array panel produces 16 fixed simultaneous beams for a total of 48 beams per satellite linked to hand-held phones having low-gain antennas. The MMA radiates multiple carriers into multiple beams with high efficiency and linearity as well as being lightweight and able to function in the thermal and radiation environment of space. MMA was optimized for the highest link margin accordingly with its size and the budgeted RF power per carrier. The architecture of the MMA phased-array panel is shown in Fig. 8(d); each array consists of over 100 lightweight patch radiators, each of which is driven by a Transmitter/Receiver (T/R) module, which are in turn collectively excited by an optimized beamformer network. The beamformer network forms the 16 optimized shaped beams for both transmit and receive operation with the T/R modules maintaining a high G/T in receive operation and efficient EIRP generation for transmit operation. The satellite can receive or transmit through each beamport, providing the RF access to a particular fixed beam. In general, several or all beams



(a) The *Morelos* satellite

(b) The *Morelos* at the construction bay

(c) IRIDIUM space vehicle (©(1999) IEEE)

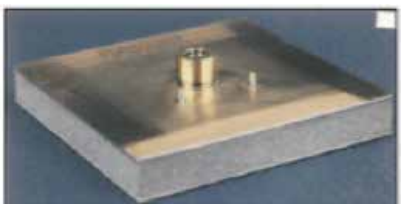


(d) MMA panel construction (©(1999) IEEE)

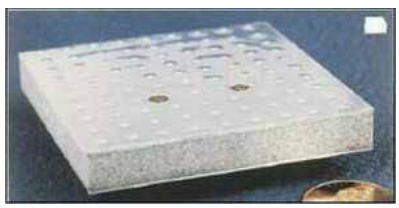
Fig. 8. The use of planar antennas in commercial satellites and space vehicles

can be utilized at once in either transmit or receive operation with the only limitation being the MMA capacity constraints on transmit.

5.2.1 Patch Radiator



(a) Bottom view of patch radiator (©(1999) IEEE)



(b) Top view of patch radiator (©(1999) IEEE)

Fig. 9. Patch radiator developed for the MMA

Fig. 9(a) and Fig. 9(b), show the patch radiator developed for the MMA, which was manufactured as a separate component and bonded onto the MMA panel during array assembly; its radiator is built as one assembly and contains the matching and polarizing networks; a single 50 Ω input connector is provided on the underside of the patch for connection to the T/R module. The radiator cavity is loaded with an artificial dielectric substrate whose weight is approximately one tenth that of teflon, but which has a dielectric constant of approximately two. This dielectric constraint is needed to obtain the desired scan and polarization performance of the array. The artificial dielectric also permit efficient heat radiation out the front face of the array during peak traffic loads.

5.3 Antennas for Modern Small Satellites

Many examples of planar antennas application are discussed in literature, but its major application could be the modern small satellites (MSS) which are revolutionizing the space industry (Gao et al., 2009). They can drastically reduce the mission cost, and can make access to space more affordable.

These modern small satellites are useful for various applications, including telecommunications, space science, Earth observation, mitigation and management of disasters (floods, fire, earthquake, etc.), in-orbit technology verification, military applications, education, and training. Typical antenna coverages ranges from low-gain hemispherical, to medium-gain antennas. The basic radiator designs used are normally helices, monopoles, patches, and patch-excited cups (PEC), depending on frequency and range, coverage requirements, and application. As antenna examples of small satellites are mentioned various monopole antennas, printed inverted-F-shaped antennas (PIFAs), microstrip-patch antennas, helices, and patch-excited cup antennas, developed for telemetry, tracking, and command in the UHF, VHF, S, C, and X bands. These antennas are simple, cheap, easy to fabricate, and have wide radiation-pattern coverage; the satellite thus does not need accurate control of attitude.

Universities have played an important role in satellites development, since the beginning of space era; professors were interested in the new research area, either as academic developers or as a part of contracts with satellite industry, but small satellites seems to be a very appropriate area to be working in by universities, due the few economical resources needed. As an example we can mention universities in Mexico, creating clusters to design small satellites; institutions as CICESE (Centro de Investigación Científica y de Educación Superior de Ensenada) in north of Mexico developing transponders and the Instituto Politécnico Nacional working with satellite structures and integration into a clean room, design of monopoles and planar antennas for satellite applications and also exploring the capabilities of new active devices as candidates for LNA amplifiers (Enciso et al., 2005). An especial mention should be make to the Universidad Nacional Autónoma de México (UNAM) which has been working towards the design of a femto satellite.

Other illustrative example is the University of Surrey, which has been developing modern small satellite technology since starting its UoSAT program in 1978. UoSAT-l, developed by Surrey, was launched in 1981. This was followed by UoSAT-2 in 1984. UoSAT-l continued to operate for eight years, while UoSAT-2 was still operational after 18 years in orbit. During the past 30 years, the University of Surrey's spinoff company, Surrey Satellite Technology Ltd. (SSTL), together with Surrey Space Centre (SSC), have successfully designed, developed and launched 32 modern small satellites for various countries around the world. (Gao et al., 2009) have a complete description of various small satellites, which are described in the next lines and figures. Fig. 10 shows a photograph of the S-band microstrip-patch antenna used at SSTL;

it employs a circular microstrip patch, fed by a 50Ω probe feed at the bottom. It can operate within a tunable frequency range of 2.0-2.5 GHz. Left-hand or right-hand circular polarization can be achieved by using a single feed combined with patch perturbation, or a 90°microstrip hybrid combined with a circular patch. It achieves a maximum gain of about 6.5 dBi, has a size of 82 x 82 x 20 mm, and a mass of less than 80 g. It can operate within -20°C to +50°C, is radiation tolerant to 50 kRad, and qualified to 50 Gs rms random vibration on three axes.

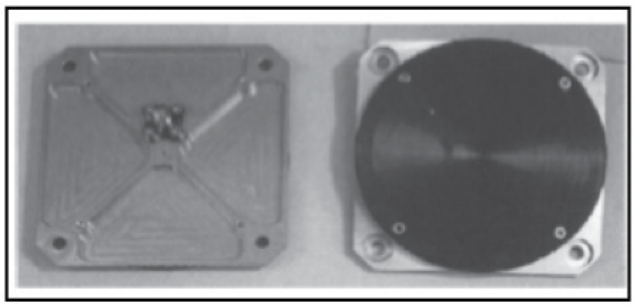


Fig. 10. An S-band patch antenna SSTL. (©(2009) IEEE)

To respond the need for single-frequency low-profile and low-weight hemispherical or near-hemispherical antennas, working at S, C, or X band, patch-excited cup antennas were developed at RUAG Aerospace Sweden. They consist of a short cylindrical cup, with a circular cross section and an exciter. The cup is excited using a stacked circular dual-patch element, or a single patch. The lower patch or the single patch is fed at one point, and the patch has two opposite perturbations for generating circular polarization. The antennas have special features to minimize their coupling to the surrounding spacecraft environment, as this is a common problem for low-gain antennas of this type, and it has an effect on the installed performance. The antenna's diameter is 60 mm for the C band antenna, and 40 mm for the X-band antenna. The mass is less than 90 g for the C-band antenna, and less than 20 g for the X-band antenna. They are both almost all metal antennas (which is a preferred property), with dielectric material only in the interface connector.
Fig. 11 shows the X-band patch-excited cup antennas that can be used for the telemetry, tracking, and command function. Fig. 12(a) shows the S-band patch-excited cup antenna, developed at Saab Space. It consists of three patches, mounted within a thin aluminum cup with a rim height of about a quarter wavelength. Two lower patches form a resonant cavity, allowing broadband or double tuning. The top patch acts as a reflector that affects the illumination of the aperture, and is used to improve the aperture efficiency. To achieve circular polarization, the lower patch is fed in phase quadrature at four points by a stripline network. It achieves a maximum gain of about 12 dBi. A patch-excited cup antenna development performed at Saab Space is the update of the antenna in Figure 6, to be used for other missions; it has a radiator tower that is modified compared to the original design. It is now an all-metal design, and has a new feed network configuration: an isolated four-point feed design, antenna is shown in Fig. 12(b).
Surrey also pioneered the use of GPS and global navigation satellite systems (GNSS) in space. A GPS receiver can provide accurate position, velocity, and time for LEO satellites. For this application, the antenna needs to be compact, low profile, able to operate at GPS frequencies in the L1 (1.575 GHz) and L2 (1.227 GHz) bands with stable performance, and produce low backward radiation towards the small satellite body.

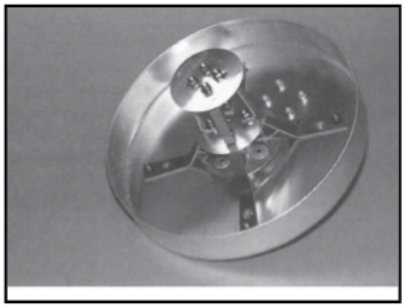
Fig. 11. An X-band patch-exited cup antenna (©(2009) IEEE).

A medium-gain antenna, shown in Fig. 13(a), was launched on the UK-DMC satellite of SSTL for the purpose of collecting reflected GPS signals in orbit. This satellite has begun to collect reflected signals under a variety of sea conditions, and over land and ice. The antenna is a three-element, circularly polarized microstrip-patch array with a gain of 12 dBi. Antenna-design challenges remain in terms of further reducing antenna size, improving the antenna's efficiency, multi-band (L1/L2/L5 band) operation, constant phase center, multipath mitigation, etc.

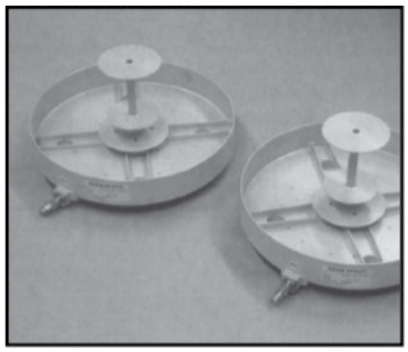
Fig. 13(b) shows the patch-excited cup antenna developed at RUAG Aerospace Sweden. It consists of two patches placed in a circular cup. To obtain a stable antenna covering two GPS frequency bands (Ll, L2), the bottom patch was capacitively fed by four probes and an isolated feed network. The antenna achieved a coverage out to 80^0 in zenith angle, and low backward radiation. The antenna's diameter is 160 mm, and the mass is 345 g. This antenna shows how shorted-annular-patch can achieve high-accuracy GPS/GNSS performance without compromising the physical constrains.

6. Some proposals for future applications

Spacecraft development and research never ends and antenna improvements are not the exception, even thinking that some of them were designed for other applications, always is possible to extrapolate to space applications, but antenna research and design for satellites and spacecrafts is an area of permanent expansion. Starting with airborne applications, where planar antennas have a permanent development, to meet the low profile and conformal challenges, is possible to extrapolate them to satellite systems. For airplanes as for satellite and spacecrafts, an array antenna should have good isolation, high efficiency, and ease of integration, also a simple feeding-line network with lower loss and high isolation is generally desired. Microstrip series-fed arrays have been shown to have a structure that enhances the antenna's efficiency. This is because the array feeding-line length is significantly reduced, compared to

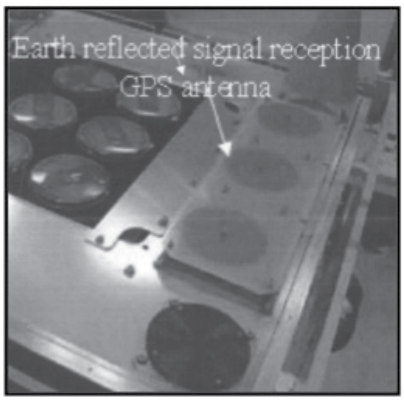


(a) Cup antenna at RUAG (©(2009) IEEE)

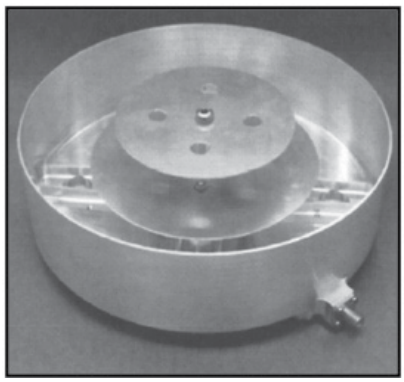


(b) Medium-dowlink antennas (©(2009) IEEE)

Fig. 12. S-band patch-excited cup antenna.



(a) For the UK DMC satellite at SSTL (©(2009) IEEE)



(b) Antennas at RUAG (©(2009) IEEE).

Fig. 13. GPS antennas.

the conventional corporate feeding-line network. A planar structure with a thin and flexible substrate is a good choice, because it will not disturb the appearance of the aircraft, and can be easily integrated with electronic devices for signal processing.

6.1 The Shih planar antenna

An example of a planar antenna first designed for aircrafts is the dual-frequency dual-polarized array antenna presented by (Shih et al., 2009). It consists of a multilayer structure of two antennas separated on different layers, adopted for dual-band operation, working in the S band and X band frequencies. To reduce the array's volume and weight, a series-fed network is used. An ultra-thin substrate is chosen in order to make the array conformal, and the array can be easily placed on an aircraft's fuselage, or inside the aircraft.

6.1.1 S-band Array Design

The multilayer array structure for dual-band operation is shown in Fig. 14. The S-band antenna elements sit on the top layer, and the X-band antennas are on the bottom layer. A foam layer (h_2) serves as the spacer, and is sandwiched between the two substrate layers. One of the important design considerations for this multilayer dual-band array is that the S-band antenna element should be nearly transparent to the X-band antenna elements. Otherwise, the S-band element may degrade the performance of the X-band antenna. Two RTlDuroid 5880 substrates ($\epsilon_1 = \epsilon_3 = 2.2$) and a foam layer ($\epsilon_2 = 1.06$) form the multilayer structure. The thicknesses of the substrates (h_1 and h_2) are both only 0.13 mm. These ultra-thin and flexible substrates make it possible for the array to be easily attached onto the aircraft's fuselage, or installed inside the aircraft. The foam layer has a thickness of h_2 =1.6 mm.

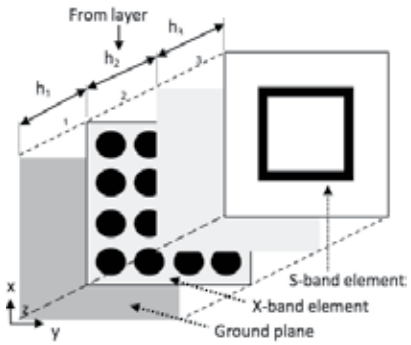


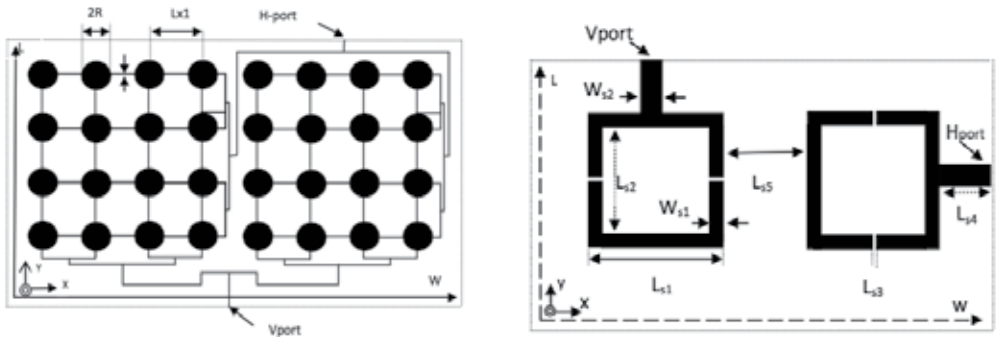
Fig. 14. The multilayer structure of dual-band dual polarized array antenna (©(2009) IEEE).

6.1.2 X-Band Antenna and Subarray

The X-band array uses the circular patch as its unit antenna element. The circular patches are fed with microstrip lines at the circumferential edge, as shown in Fig. 15(a) for a single circular patch, two microstrip feeding lines are used to feed the circular patch to generate two orthogonally radiating TM_{11} modes for dual polarized operation. Two feed points are located at the edge of the patch, 90ą away from each other, so that the coupling between these two ports can be minimized. The port isolation also depends on the quality factor of the patch. Increasing the substrate's thickness decreases the isolation, therefore using thin substrates could improve the quality of isolation.

Fig. 15(a) shows a 4 x 8 dual-polarized X-band array. The V port and the H port are the input ports for the two orthogonal polarizations (vertical and horizontal). The array is composed of two 4 x 4 subarrays. The corporate-fed power-divider lines split the input power at each port to the subarrays. Within each subarray, the circular patches are configured into four 4 x 1 series-fed resonant type arrays, which make the total array compact and have less microstrip line losses than would a purely corporate-fed type of array. An open circuit is placed after the last patch of each 4 x 1 array. The spacing between adjacent circular-patch centers is about one guided wavelength (λ_g =21.5 mm at 10 GHz). This is equivalent to a 360^o phase shift between patches, such that the main beam points to the broadside. The power coupled to each patch can also be controlled by adjusting the size of the individual patch to achieve a tapered amplitude distribution for a lower-sidelobe design.

As shown in Figure Fig. 15(b), the S-band antenna elements are printed on the top substrate, and are separated from the X-band elements by the foam layer. To reduce the blocking of



(a) X-band antenna (©(2009) IEEE) (b) S-band antenna (©(2009) IEEE)

Fig. 15. Microstrip Antenna Arrays

the radiation from the X-band elements at the bottom layer, the shape of the S band elements has to be carefully selected. A ring configuration was a good candidate, since it uses less metallization than an equivalent patch element. Here, a square-ring microstrip antenna is used as the unit element of the S-band array. Because antenna elements at both frequency bands share the same aperture, it is also preferred that the number of elements on the top layer be as small as possible, to minimize the blocking effects.

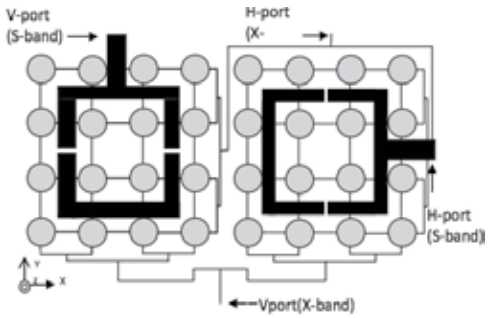


Fig. 16. Geometry of dual antenna (©(2009) IEEE)

The stacked X-band and S-band array antennas are shown in Fig. 16. As can be seen in the figure, the four sides of the square-ring element are laid out in such a way that they only cover part of the feeding lines on the bottom layer, but none of the radiating elements. Unlike an ordinary microstrip-ring antenna that has a mean circumference equal to a guided wavelength, the antenna proposed here has a mean circumference of about $2\lambda_g$ (λ_g = 82.44mm at 3 GHz). Although the size of the proposed unit element is larger than an ordinary ring antenna, its gain is about twice as high, because of its larger radiation-aperture area. The ring is loaded by two gaps at two of its parallel sides, these make possible to achieve a 50 Ω input match at the edge of the third side without using a small value of L_{s2}/L_{s1}. For an edge fed microstrip ring, if a second feed line is added to the orthogonal edge, the coupling between the two feeding ports will be high. The V-port and H-port feeds are therefore placed at two individual elements, so that the coupling between the two ports can be significantly reduced. Using separate elements seems to increase the number of antenna elements within a given aperture.

However, this harmful effect could be minimized by reducing the number of elements with the use of larger-sized microstrip rings.

6.2 The Cross Antenna

The cross antenna is another possibility of use in spatial applications, it is a traveling wave antenna with circular polarization formed by conductors over a ground plane, proposed by (Roederer, 1990). Antenna can be constructed as a wire or printed antenna. Roederer's paper do not describe completely the antenna but it was reanalyzed by authors (Sosa-Pedroza et al., 2006).

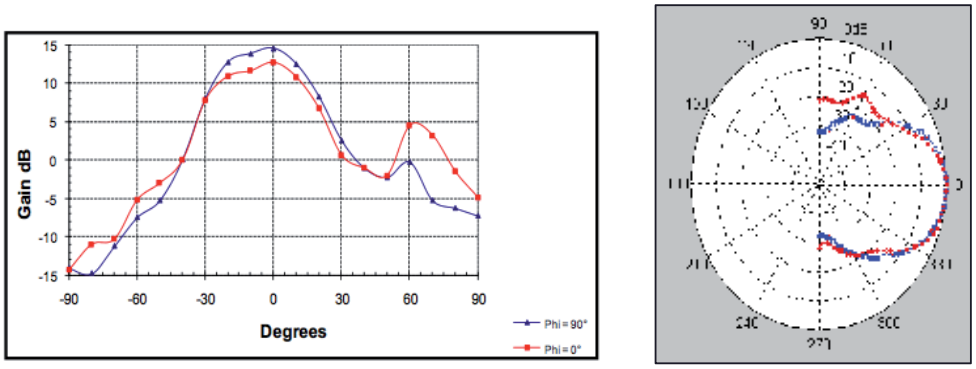
The cross antenna is a printed structure of medium gain and circular polarization, consisting of a conductor or microstrip over a ground plane following the contour of a cross with four or more arms and a diameter of about 1.5 wavelengths. The antenna is feeding on one end by a coaxial line and finished on the other end by a load impedance, considering behavior of travelling wave. Even the antenna was primarily designed for applications in L Band (1500 MHz) mobile communications, the design and experimental characterization was made at 10 GHz and for an eight arms antenna besides original four arms antenna, showing the possibility of extrapolation for other applications as satellite communications. For the cross antenna, feed connector and load position define the right or left circular polarization; it can be used as a unique radiator or as a part of an array, a proposal is that could be used as primary antenna for parabolic reflector with wide focal length and diameter relationship. The main advantage of the cross antenna is its gain (12-15 dBi) compared with its size and weight, ideal for space communications.

Fig. 17. The cross wire antenna

Arm length	λ_{eff}
Arm width	$0.25\lambda_{eff}$
Cross diameter	$2.5\lambda_{eff}$
Wire diameter	$0.01\lambda_{eff}$

Table 2. Geometric characteristics of cross antenna

The power at the end of the antenna is controlled by the load impedance and is limited to a small percentage, changing the height of the conductor over the ground plane (typically $\lambda_{eff}/20$ to $\lambda_{eff}/4$) which also affects the axial rate. The bandwidth of the cross antenna is around 5% depending on the number of arms. Fig. 17 shows photograph of a 8 arm radiator,



(a) Gain (b) Radiation Pattern

Fig. 18. Electrical characteristics of the Cross antenna

which was constructed both, as a microstrip antenna using a 3.6 mm thick RTDuroid with 2.3 of ϵ_{eff}=2.3 and as a wire antenna using copper wire, supported over the ground plane by small Teflon fragments giving flexibility to move up the structure to analyze the effect of height over the ground plane. Table 2 shows dimensions of the antenna. On the other hand Fig. 18(a) and Fig. 18(b) show the gain and the radiation pattern respectively, for one of the antennas.

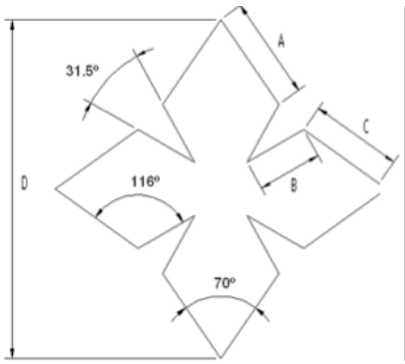
6.3 Rhombic cross antenna

A variation over cross antenna is a four arm rhombic cross antenna (Lucas et al., 2008), it is also a medium gain and circular polarization structure made of a conductor or strip line over a ground plane, following a rhombic contour of four branches. One end is connected to the feed line and the other is grounded by a load impedance. Antenna was analyzed using Method of Moments and constructed for experimental analysis using both, a 12 AWG wire over a ground plane and printed as a microstrip structure working in 4.2 GHz. The rhombic antenna shows a better performance compared with the four arms Roederer's antenna, with almost 15 dB gain and 1.4 dB for axial ratio. The antenna can be used in mobile communication or as primary radiator of parabolic reflectors, when circular polarization is needed. The construction repeatability is very easy as well the facility to obtain 15 dB gain in a very small antenna.

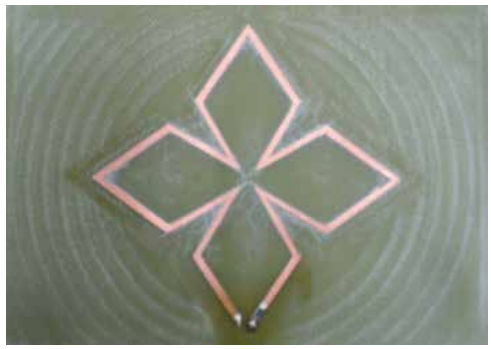
A	$0.430\lambda_{eff}$
B	$0.276\lambda_{eff}$
C	$0.3911\lambda_{eff}$
D	$1.4112\lambda_{eff}$

Table 3. Dimensions of rhombic antenna

The rhombic cross antenna geometry is shown in Fig. 19(a), and antenna dimensions as function of effective wavelength, are given in Table 3.
There were constructed several antennas, both wire (air dielectric) and strip line (fiber glass dielectric), the last one is shown in Fig. 19(b); wire antenna uses Teflon supports over the

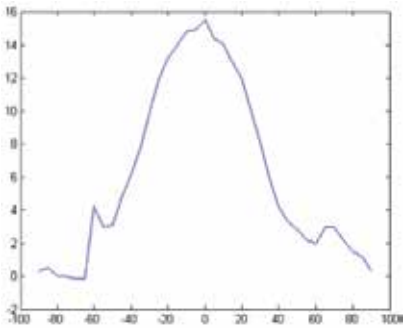


(a) Scheme of rhombic antenna

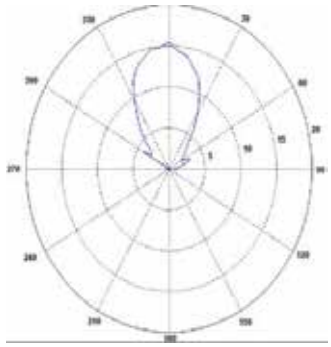


(b) Microstrip antenna

Fig. 19. Physical characteristics of Rhombic Antenna



(a) Rhombic antenna gain



(b) Rhombic antenna field pattern

Fig. 20. Physical characteristics of Rhombic Antenna

ground plane, giving flexibility to change the height over it. Results for gain and field pattern are shown in Fig. 20(a) and Fig. 20(b) respectively, for a 50 Ω load impedance; the feed impedance is $Z = 38.6 - j56.8\Omega$ for 2.4 GHz:
Even the proposed antennas have not been used yet for spatial applications, their profiles can match for it, in frequencies ranging from L band, S band, commercial C band or X band, either as single structures or as arrays.

7. References

Agrawal, P. K., Bailey M. C. (1977) *An Analysis Technique for Microstrip Antennas*. IEEE Trans. on Antennas and Propagation AP-25, pp.756-759

Balanis, C.(2005).*Antenna Theory Analysis and Design*, Wiley-Interscience ISBN 0-471-66782-X

Barrera-Figueroa, V.; Sosa-Pedroza, J.; Lopez-Bonilla, J. (2007). *Numerical approach to King's analytical study for circular loop antenna*, Journal of Discrete Mathematical Sciences & Cryptography, Vol. 10, No. 1 February 2007, pp 82-92.

Barrera-Figueroa, V.; Sosa-Pedroza, J.; Lopez-Bomilla, J. (2009) *Pocklington Equation via circuit theory* Apeiron, on line Journal, Vol 16, No. 1.

Chang, K.(1989). *Handbook of Microwave and Optical Components*. Vol. 1, John Wiley & Sons, New York, USA

Derneryd, A.,G.(1975). *Linear Polarized Microstrip Antennas*. IEEE Trans. on Antennas and Propagation, AP-24, pp. 845-850.

Deschamps, G. A. (1953) *M*icrostrip microwave antennas. *3rd USAF Symposium on Antennas*.

Enciso-Agilar, M.; Crozat, P.; Hackbarth, T; Herzog, H,; Aniel F.(2005). *Microwave Noise Performance and Modeling of SiGe-Based HFETs*. IEEE Trans. on Electron. Dev., Vol. 52, No. 11, November 2005

Gao S.; Clark K.; Unwin M.; Zackrisson J.; Shiroma W.A.; Akagi J.M.; Maynard K.; Garner P.; Boccia L.; Amendola G.; Massa G.; C. Underwood; Brenchley M.; Pointer M.; Sweeting M.N. (2009). *Antennas for Modern Small Satellites.* IEEE Antennas and Propagation Magazine (August 2009), Vol. 51 No.4, pp. 40-56. IEEE. ISSN 1045 9243/2009.

Greig D., D.; Engleman H., F. (1952) Microstrip a new transmission technique for the kilomegacycle range. *Proceedings of IRE*, No. 40.

Gupta,K.C.; Benalla, A. eds. *Microstrip Antenna Design*. Artech House, Norwood MA, USA

Harrington, R.F.,(1961). *Time-Harmonic Electromagnetic Waves*, McGraw-Hil, New York, USA

Harrington,F.,R., *Matrix Methods for Field Problems*, Proc. IEEE, Vol. 55, No. 2, Feb. 1967.

Roger F. Harrington, (1992). *Field Computation by Method of Moments*, IEEE Press Series on Electromagnetic Waves, ISBN 0 7803 1014 4, Piscataway, NJ, USA.

Hirasawa, K.; Haneishi, M. (1994). *Analysis, Design and Measurement of Small and Low Profile Antennas*, Artech House, ISBN 0-89006-486-5,USA

Howell, J., Q.(1975) *Microstrip Antennas*. IEEE IEEE Trans. on Antennas and Propagation, vol. AP22 , pp. 74-78.

Itoh, T.; Menzel,W. (1989) *A Full Wave Analysis Method for Open Microstrip Structures*. IEEE Trans. on Antennas and Propagation AP-29 , 63-68.

James, J.R., Hall, P. S. (1989). *Handbook of Microstrip Antennas*, Vol 1. Peregrinus Ltd, IEE Electromagnetic Waves Series, ISBN 0 86341 150 9 Peter, London United Kingdom.

James, J. R.; Hall, P., S.; and Wood, C. (1981). *Microstrip antenna theory and design.* IEE Peter Peregrinus.

Kraus, J.,D.; Marhefka, R.,J. (2002) *Antennas 3rd. Edition*, ISBN 0 07 232103 2, New York NY USA.

Lee H.,F.; Chen, W. (1997). *Advances in Microstrip and Printed Antennas*. John Wiley & Sons, New York, USA

Long, S., A.; Shen, L.,C.; & Morel,P.,B. (1978). *A Theory of the Circular Disk Printed Circuit Antenna*. Proc IEE, Pt. H, Vol. 125, October 1978, pp. 925-928

Lucas-Bravo A., Sosa-Pedroza J., Barrera-Figueroa V. (2008). Experimental and numerical results of a rhombic cross antenna. *5o Congreso Internacional de Ingeniería Electromecánica y de Sistemas*, pp. 1178-1182. Mexico D.F., November 2008.

Thomas, A., Milligan.(2005). *Modern Antenna Design*. John Wiley & Sons. New Yersey, USA

Munson, R., E. (1974). "Conformal Microstrip Antennas and Microstrip Phased Arrays" *IEEE Trans. Antennas and Propagation*, Vol. AP-22, January 1974, pp 74-78, ISSN 0018-926X

Pozar,D.,M. (1982) *Input impedance and mutual coupling of rectangular microstrip antennas*. IEEE Trans Antennas Propagatation. AP-30, 1191-1196.

Pozar,D.M. (2005). *Microwave Engineering*, John Wiley & Sons, USA

Reineix, A.; Jecko, B. (1989) *Analysis of Microstrip Patch Antennas Using Finite-Diference Time Domain Method*. IEEE Trans. on Antennas and Propagation AP-37 pp. 1361-1369.

Richards,W.,F.; Lo,Y., T.; Harrison, D.,D. (1979) *An improved theory for microstrip antennas* .Electron. Letters, Vol 15 , 42-44.

Richards,W.,F.; Lo,Y., T.; Harrison, D.,D. (1981) *An improved theory for microstrip antennas and applications*. IEEE Trans. on Antennas and Propagation AP-29 , 38-46.

Roederer A. (1990). *The Cross Antenna: a New Low Profile Circularly Polarized Radiator*. IEEE Transactions on Antennas and Propagation, Vol. 38, no. 5, May 1990.

Schuss J. J., Upton J., Myers B., Sikina T., Rohwer A., Makridakas P., Francois R., Wardle L., and Smith R. (1999). *The IRIDIUM Main Mission Antenna Concept*. IEEE Transactions on Antennas and Propagation, Vol. 47, No. 3, March 1999, pp. 416-424, IEEE, ISSN 0018 926X/99

Sheen, D.,M.; Ali, S.,M.;Abouzahra M.,D.; Kong J.,A. (1990). *Application of Three-Dimensional Finite-Difference Time-Domain Method to the Analysis of Planar Microstrip Circuits*. IEEE Trans. on Microwave Theo. and Tech. vol. 38, pp. 849-857

Shih-Hsun H., Yu-Jiun R., and Kai C. (2009). *A dual-Polarized Planar-Array Antenna for S-Band and X-Band Airbone Applications.* IEEE Antennas and Propagation Magazine, Vol. 51, No.4, August 2009, pp.70-78

Sosa-Pedroza J., Lucas-Bravo A., Lopez-Bonilla J. (2006). Numerical and Experimental Analysis for a Cross Antenna. *International Caribbean Congress on Devices, Circuits and Systems* pp 207-211, ISBN: 1 4244 0042 2, Playa del Carmen, Quintana Roo, Mexico, April 2006.

Sosa-Pedroza, J.; Enciso-Aguilar, M.; Benavides-Cruz, M. (2008). A 9 slots antenna designed by Chebyshev technique and modeled by Finite Difference Time Domain, *Proc. of 7th International Caribean Conference on Devices Circuits and Systems*, Cancun Quintana Roo, Mexico, April 2008, IEEE.

Sosa-Pedroza, J., Barrera-Figueroa,V., Lopez-Bonilla, J. Equidistant and non-equidistant sampling for method of moments applied to Pocklington Equation. *The 18th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, Cancun Mexico, 2007.

Sosa-Pedroza J.; Enciso-Aguilar, M.; Benavides-Cruz, M.; Nieto-Rodríguez, M.; Galaz-Larios, M. A parametric analysis of perfect matched layer model of Finite Difference Time Domain Method, *Proc. of PIERS* 2009, Moscow, Russia.

Vasylchenko, A., Schols, Y., De Raedt , W., and Vandenbosch, G. A. E. *Quality Assessment of Computational Techniques and Software Tools for Planar-Antenna Analysis*, IEEE Antennas and Propagation Magazine, Vol. 51, No.1, February 2009

Volakis, J.,L.; Chatterjee, A.; and Kempel, L.,C. (1998). *Finite element method for electromagnetics* IEEE Press, Chap. 7, 1998.

www.satmex.com. (2010)

Yee, K. S (1966). *Numerical solution of initial boundary value problem involving Maxwell's equations in isotropic media.* IEEE Trans. on Antennas and Propagation, vol. 14, pp. 302-307.

18

Power and Spectral Efficient Multiuser Broadband Wireless Communication System

Santi P. Maity
Bengal Engineering & Science University, Shibpur
India

1. Introduction

Communication Satellite plays significant role in long distance broadband signal transmission in recent times. Development of an efficient high data rate communication system for multiusers becomes important and challenging. To meet ever-increasing demand for broadband wireless communications, a key issue that should be coped with is the scarcity of power and spectral bandwidth. In the conventional communication scenario, supporting ubiquitous users with high quality data communication services requires huge bandwidth. However, since the spectrum of nowadays is a very costly resource, further bandwidth expansion is impracticable. Thus, the use of efficient multiple access is critical. Moreover, for a given limited bandwidth and limited number of antennas, the only option that can increase data rate is to increase transmit power as suggested by Shannon channel capacity theorem. Therefore, development of power and bandwidth efficient coding, modulation, and multiple-access techniques is essential for the future wireless communication systems implemented through Satellite link.

Considering the above issues, this chapter discusses a new communication system for Satellite system that can achieve high power and spectral efficiency in broadband wireless communication. To achieve the goal, development of a new and complete communication system with high user capacity and variable data rate become essential. The system includes multicarrier code division multiple access (MC-CDMA) with peak-to-average power ratio (PAPR) reduction using channel coding, optimization in MC-CDMA, estimation of wireless channel condition and MC-DMA with multi-user detection (MUD). The chapter proposal focuses on different aspects for different part of a communication system, namely PAPR reduction in transmitter, channel estimation for design of adaptive and optimized system, multiuser detection at the receiver for increase in user capacity. The different issues are described under four broad subheadings. Prior to that a brief literature review on the above issues are also presented.

The organization of the chapter is as follows: Section 2 presents literature review, while proposed system model is described in Section 3. Design of power and spectral efficient system is described in Section 4. Performance evaluation of the system is presented in Section 5, while conclusions and scope of future works are highlighted in Section 6.

2. Literature Review

In this section, we present a brief literature review related to PAPR reduction, multiuser detection in CDMA, channel estimation and optimization in CDMA/MC-CDMA system. The objective of this review is to discuss the merits and the limitations of the existing works, scope of the present work and finally to compare the performance of this work with respect to the related existing works.

A growing number of techniques have been proposed in the recent literature to alleviate power efficiency problem in multicarrier system through the development of PAPR reduction methods. The approaches include simplest filtering and clipping (Ochiai & Ima; 2000), partial transmit sequence (PTS) (Lim et al; 2006), selective mapping (Yoo; 2006) to the sophisticated trellis shaping approach (Ochia; 2004). Ochiai et al (Ochia; 2004) propose a new trellis shaping design based on recursive minimization of the autocorrelation sidelobes for reducing PAPR of the bandlimited orthogonal frequency division multiplexing (OFDM) signals. The exponentially increasing complexity of the PTS scheme prevents its use for MC-CDMA. Kang et al. (Kang et al; 2005) propose an efficient PTS scheme through phase factor optimization to reduce the complexity. Natarajan et al (Natarajan; 2004) address signal compactness issues in MC-CDMA employing CI codes using the measure of crest factors.

Literature on multiuser detection in CDMA is quite rich. The optimum multiuser detector proposed in (Vedu; 1986) achieves significant performance improvement relative to single user receiver but the computational complexity increases exponentially with the number of users. This has motivated the use of low complexity linear (Lupas & Vedu; 1989) and decision driven suboptimal multiuser detection techniques. To overcome the disadvantage intrinsic to the total parallel interference cancelation (PIC), some modified versions like partial parallel interference canceling (Divsalar et al; 1989) and linear parallel interference cancelation technique (Kim & Lee; 2001) have been proposed. Parallel interference weighted canceler has been proposed in (Xiao & Liang; 1999) to mitigate the degrading effects of unreliable interference estimation. Two variants of PIC known as threshold PIC and block PIC have been proposed in (Thippavajjula & Natarajan; 2004) for synchronous CI/MCCDMA uplink. Block PIC is shown to provide performance gain relative to conventional PIC. Maity et al (Maity at al; 2008a) further improves bit error rate (BER) performance of diversity assisted block PIC using genetic algorithms.

A growing number of techniques have been proposed in the recent literature to estimate wireless channel parameters. Sgraja et al (Sgraja & Linder; 2003) propose an algorithm to estimate the whole channel matrix using Wiener filtering. Such an estimator requires a Wiener filter, a multiplication operation for each element of the channel matrix and matrix inversion, which increases the complexity of the receiver for large number of sub-carriers. In some systems, a known training sequence is sent by the transmitter and a training algorithm is performed by the receiver on the observed channel output and the known input to estimate the channel (Chow et al; 1991), (Ziegler & Cioffi; 1992) The deterministic least squares (DLS) channel identification algorithm in (Ziegler & Cioffi; 1992) is such a simple but widely used training approach. However, it is not suited for time varying system. In (Wang & Ray; 1999), authors propose an adaptive channel estimation algorithm using the cyclic prefix. This algorithm can adaptively track the channel variation without additional training sequences.

In (Choi et al; 2001), time-domain channel estimation followed by symbol detection based on zero forcing (ZF) and minimum mean square error (MMSE) criterion have been proposed. Pilot based channel estimation is also widely used in many algorithms. The most common pilot based channel estimation scheme is the least square (LS) method of channel estimation

(Coleri et al; 2002), (Schramm et al; 1998). Blind adaptive channel estimation, based on least mean square (LMS) for OFDM communication, is proposed in (Doukopoulos & Moustakides; 2004). Kalman filter is also used for channel estimation in (Gupta & Mehra; 2007). Finite affine projection (FAP) algorithm is used to estimate the channel response for OFDM system in (Gao et al; 2007). Gao et al (Gao et al; 2007) propose an efficient channel estimation for multiple input multiple output (MIMO) single-carrier block transmission with dual cyclic timeslot structure. An optimal channel estimation is then investigated in the minimal mean square error (MMSE) sense on the time slot basis.
This channel state information (CSI) is exploited to design optimized system and several such algorithms for CDMA and MC-CDMA systems are reported in literature. The optimized systems include adaptive algorithm to minimize the total transmitted power (Lok & Wong; 2000), signal-to-interference ratio (SIR) balanced optimum power control in CDMA (Wu; 1999), joint optimization of spreading codes and a utility based power control (Reynolds & Wang; 2003), joint rate and power control (Kim; 1999), joint transmitter-receiver optimization using multiuser detection and resource allocation for energy efficiency in wireless CDMA networks (Buzzi & Poor; 2008), using transmitter power control, receiver array processing and multiuser detection (Seo & Yang,2006) etc. The objective of the optimized methods, in general, is to minimize transmit power and maximize per user's data transmission rate.

3. Proposed System

MC-CDMA system was first proposed in (Yee; 1993) and is a combination of CDMA and OFDM with the spreading codes applied in frequency domain. We use carrier interferometry (CI) code as spreading code. Interferometry, a classical method in experimental physics, refers to the study of interference patterns resulting from the superpositioning of waves. CI codes of length N supports N users orthogonally and then, as system demand increases, codes can be selected to accommodate upto addition N-1 users pseudo-orthogonaly. Additionally, there is no restriction on the length N of the CI code (i.e. $N \in I$), making it more robust to the diverse requirements of wireless environments. Since the present MC-CDMA system employs CI code as spreading code, this multiple access scheme is called as CI/MC-CDMA system. In other words, the CI/MC-CDMA is an MC-CDMA scheme employing complex carrier interferometry spreading codes. Assuming that there are K users and N subcarriers in the MC-CDMA system, the CI code for the k-th user $(1 \leq k \leq K)$ is given by (Natarajan et al; 2001):

$[1, e^{j\Delta\theta_k}, ...e^{j(N-1)\Delta\theta_k}]$ where $\Delta\theta_k = 2\pi.k/N$.

In this chapter, we have discussed the operation and performance of a newly developed CI/MC-CDMA model which is a variation that is discussed in (Natarajan et al; 2001). The system proposed here is capable of supporting high capacity with reduction in PAPR as well as low BER values at the receiver. Therefore, we present a brief review of the model in (Natarajan et al; 2001) followed by our new model.

3.1 Transmitter Model

In CI/MC-CDMA transmitter (Natarajan et al; 2001), the incoming data $a_k[n]$ for the k-th user, is transmitted over N narrow-band sub-carriers each multiplied with an element of the k-th user spreading code. Binary phase shift keying (BPSK) modulation is assumed, i.e. $a_k[n] = \pm 1$, where $a_k[n]$ represents n-th bit of k-th user. The transmitted signal corresponding to n-th bit of the incoming data is given by

$$S(t) = \sum_{k=0}^{N-1} \sum_{i=0}^{N-1} a_k[n] exp^{j(2\pi.f_i t + i\Delta\theta_k)} p(t) \quad (1)$$

where $f_i = f_c + i.\Delta f$ is the i-th subcarrier, and p(t) is a rectangular pulse of duration T_b. As with the traditional MC-CDMA, the Δf's are selected such that the carrier frequencies are orthogonal to each other, typically $\Delta f = 1/T_b$, where T_b is the bit duration.
The transmitter model developed in (Natarajan et al; 2001) allows high data rate transmission for (2N-1) users employing 'N' number of sub-carriers. However, in many practical situations, users may have transmission of variable data rate. It is logical to consider variable data rate transmission in multimedia signals environment as diverse classes of traffic signals have different requirement of quality of services (QoS)(Maity et al; 2008b). We like to modify the above transmitter model by allocating all sub-carriers to the users of high data rate transmission and odd and even sub-carriers are shared alternately among the users of low data rate. In other words, we can split the sub-carriers in even and odd parts as well as the 'N' length PO-CI (pseudo-orthogonal) codes in N/2 odd and N/2 even parts. The mathematical form for the transmitted signal S_1(t) becomes

$$S_1(t) = [\sum_{k=0}^{N-1} \sum_{i=0}^{N-1} a_k[n] exp^{j(2\pi.f_i t + 2\pi.i.k/N)} + \sum_{k=N}^{3N/2-1} \sum_{\forall i=odd}^{N-1} a_k[n] exp^{j(2\pi.f_i t + 2\pi.i.k/N + i\Delta\phi_1)}$$
$$+ \sum_{k=3N/2}^{2N-1} \sum_{\forall i=odd}^{N-1} a_k[n] exp^{j(2\pi.f_i t + 2\pi.(i+1).k/N + i\Delta\phi_2)} + \sum_{k=2N}^{5N/2-1} \sum_{\forall i=even}^{N-1} a_k[n] exp^{j(2\pi.f_i t + 2\pi.i.k/N + i\Delta\phi_3)}$$
$$+ \sum_{k=5N/2}^{3N-1} \sum_{\forall i=even}^{N-1} a_k[n] exp^{j(2\pi.f_i t + 2\pi.(i+1).k/N + i\Delta\phi_4)}] p(t - n.T_b) \quad (2)$$

where $\Delta\phi_1 = +\pi/2$,$\Delta\phi_2 = -\pi/2$,$\Delta\phi_3 = -\pi/N - \pi$ and $\Delta\phi_4 = -\pi/N$ indicate respective phase shift angles with respect to orthogonal CI codes. The phase shift are set in order to make the same even or odd subcarriers used by the different users through the respective spreading codes in out-of-phase. This out-of-phase condition reduces cross-correlation values among the spreading code patterns that not only leads to the reduction in PAPR values but also lower BER values at the receiver. This fact is analyzed mathematically as well as is supported by simulation results.

3.2 Channel Model

We also introduce here the channel model. The particular channel model needs to be considered since we will analyze later BER performance of the proposed system and its relative improvement/ degradation due to reduction in PAPR. An uplink model has been considered where all the users' transmissions are assumed to be synchronized for simplification of analysis, although, this condition may seem difficult to be valid practically. It is also assumed that every user experiences an independent propagation environment that is modeled as a slowly varying multipath channel. Multipath propagation in time translates into frequency selectivity in the frequency domain (Proakis, 1995). Frequency selectivity refers to the selectivity over the entire bandwidth of transmission and not over each subcarrier transmission. This is because $\frac{1}{T_b} \ll (\Delta f_c) \ll BW$, where Δf_c is the coherence bandwidth and BW is the total bandwidth.

3.3 Receiver Design for the proposed system

In this subsection, we describe the multiuser receiver operation when all the users transmit data through all sub-carriers as developed $S(t)$ in Eq. (1). The received signal $r(t)$ for this $S(t)$ corresponds to

$$r(t) = \sum_{k=1}^{K} \sum_{i=0}^{N-1} \alpha_{ik} a_k[n] cos(2\pi.f_i t + i.\Delta\theta_k + \beta_{ik}).p(t) + \eta(t) \tag{3}$$

where α_{ik} is the Rayleigh fading gain and β_{ik} is uniformly distributed phase offset of the k-th user in the i-th carrier and the symbol $\eta(t)$ represents additive white Gaussian noise (AWGN). The received signal is projected onto N-orthogonal carriers and is despread using j-th users CI code resulting in $r^j = (r_0^j, r_1^j, ...r_{N-1}^j)$, where r_i^j corresponds to

$$r_i^j = \alpha_{ij}.a_j[n] + \sum_{k=1,k\neq j}^{K} \alpha_{ik} a_k[n] cos(i(\Delta\theta_k - \Delta\theta_j) + \beta_{ik} - \hat{\beta}_{ij}) + \eta_i \tag{4}$$

where η_i is a Gaussian random variable with mean 0 and variance $N_0/2$. Exact phase and frequency synchronization for the desired user is assumed i.e.,$\hat{\beta}_{ij} = \beta_{ij}$. Now, a suitable combining strategy is used to create a decision variable D_j, which then enters a decision device that outputs $\hat{a}_j$. Minimum mean square error combining (MMSEC) is employed as it is shown to provide the best performance in a frequency selective fading channel (Cal & Akansu; 2000). The decision variable D_j for the n-th bit is (Natarajan et al; 2001)

$$D_n^j = \sum_{i=1}^{N} r_i^j w_{ij} \tag{5}$$

where

$$w_{ij} = \frac{\alpha}{(var(a_k)A_{ij} + N_0/2)} \tag{6}$$

where $A_{ij} = \sum_{k=1}^{K} \alpha_{ik}^2 cos(i\delta\theta_k - i\delta\theta_j + \beta_{ik} - \beta_{ij})^2$ and $var(a_k) = 1$. Thus, the outputs as the single user detector for all the users generate a decision vector $D = [D^1, D^2, ...D^K]$ which is used to obtain the initial estimates of the data $\hat{a} = (\hat{a}_1, \hat{a}_2, ...\hat{a}_k)$. These initial estimates are then used to evaluate multiple access interference (MAI) experienced by each user in the interference cancelation technique.

4. Design of power and spectral efficient system

This section describes multi-carrier code division multiple access (MC-CDMA) with PAPR (peak-to-average power ratio) reduction using channel coding, optimized system design for number of subcarriers and users, estimation of wireless channel condition, and finally MC-CDMA with multi-user detection (MUD). The overall discussion focuses on different issues for different section of a communication system, namely PAPR reduction and optimized system design in transmitter, estimation of parameters for the channel, multiuser detection at the receiver for increase in user capacity. The different issues are described under four broad subheadings as follows.

4.1 MC-CDMA with PAPR reduction using channel coding

In this section, we will first define PAPR mathematically and then see whether this PAPR is related to the properties of the spreading codes in multiuser systems. The rationale behind such relation lies as each subcarrier is multiplied by chip of individual user's spreading code which is generated from phase shift for CI codes. It would not be irrelevant to mention here that high PAPR occurs due to the superpositioning of several in-phase or near in-phase subcarriers. So cross-correlation and auto-correlation values of CI codes expect to have an influence on PAPR values of the resultant multiuser signals. In MC-CDMA system, the corresponding PAPR definition per discrete-time symbol is given by

$$PAPR = \frac{\max_{0\le i\le N-1} |S(i)^2|}{E[S(i)^2]} \tag{7}$$

where $E[|S(i)|^2]$ and $\max_{0\le i\le N-1} |S(t)|^2$ denote the the average power and the peak power, respectively in one symbol interval. The power of i-th time-domain sample denoted by $P(i)$, $1 \le n \le N$, from equation (2) for complex signal is

$$\begin{aligned} P(i) &= (\sum_{k=1}^{K}\sum_{i=0}^{N-1} a_k[n]c_k^i exp(j2\pi f_i t))(\sum_{k=1}^{K}\sum_{i1=0}^{N-1} a_k[n]c_k^i exp(j2\pi f_i t))^* \\ &= \sum_{k=1}^{K} |a_k[n]|^2 + \sum_{k=1}^{K}\sum_{k=1,k\neq k1} a_k[n]a_{k1}[n] \sum_{i2=-(N-1)}^{N-1} Z^{(k,k1)}(i2)exp(i2\pi f_{i2}t) \\ &\quad + \sum_{k=1}^{K} a_k[n]^2 \sum_{i2=-(N-1),i2\neq 0}^{N-1} Z^{(k,k)}(i2)exp(j2\pi f_{i2}t) \end{aligned} \tag{8}$$

The symbol $a_k[n]$ indicates n-th symbol of k-th user. Here $Z^{k,k1}(i2)$ is the aperiodic crosscorrelation function of spreading codes between user k and k1,

$$Z^{(k,k1)}(i2) = \sum_{m=1}^{N-i2} c_k^m (c_{k1}^{m+i2})*$$

and $Z^{k,k}(i2)$ is the aperiodic autocorrelation function of the k-th user. It can be found that the signal power of MC-CDMA system is partially determined by the correlation property of the selected set of spreading codes. We now briefly present proposed PAPR reduction method using code optimization, phase optimization and trellis coding.

A. Code Optimization

As defined earlier, the CI code for the k^{th} user is given by the spreading sequence,

$$\begin{aligned} \{\beta_k^0, \beta_k^1, \beta_k^2, ...\beta_k^{N-1}\} &= \{e^{j\Delta\theta_k^0}, e^{j\Delta\theta_k^1} ... e^{j\Delta\theta_k^{N-1}}\} \\ &= \{e^{j(2\pi/N)0.k}, e^{j(2\pi/N)1.k} ... e^{j(2\pi/N)(N-1).k}\} \end{aligned} \tag{9}$$

Here, $k\epsilon[0, 1, 2, ...K-1]$. However, low value of PAPR is achieved if we give cycle shift of the code that leads to new code. We can write the new code as

$$\begin{aligned} \text{New code } &= \{\beta_{k+shift}^0, \beta_{k+shift}^1, ...\beta_{k+shift}^{N-1}\} \\ &= \{e^{j\Delta\theta_{k+shift}^0}, e^{j\Delta\theta_{k+shift}^1} ... e^{j\Delta\theta_{k+shift}^{N-1}}\} \\ &= \{e^{j(2\pi/N)0.(k+shift)}, e^{j(2\pi/N)1.(k+shift)} ... \\ &\qquad e^{j(2\pi/N)(N-1).(k+shift)}\} \end{aligned} \tag{10}$$

Here 'shift' indicates the number of cycle shift which is fixed for all users for a given period of time, $[0 \leq' shift' \leq K]$. It is seen from the simulation results that for a minimum value of PAPR, the 'shift' is either [1,2,3] or [(K-1) to (K-3)]. It is also found that performance achieved by doing (K-1) to (K-3) shifting operations is same as is obtained by performing 1 to 3 times reverse shifting. This means that as far as the code optimization is concerned, code sequence for the users is shifted in forward or reverse cyclic shifting by an amount of 1 to 3 shifts.

B. Phase Optimization

To achieve lower PAPR value, we can use the iterative method where we shift the phase of each set of users with additional rotation in between $\Delta\phi_{(max)}$ and $\Delta\phi_{(min)}$ so that the condition $\Delta\phi_{(max)} > \Delta\phi > \Delta\phi_{(min)}$ is satisfied. It is also to be mentioned here that $\Delta\phi_{(max)}$ is allowed to take value of $\Delta\phi + \pi/32N$, while $\Delta\phi_{(min)}$=$\Delta\phi - \pi/32N$. Fig. 1 shows the proposed code and phase optimized variable data rate CI/MC-CDMA system with reduction in PAPR value.

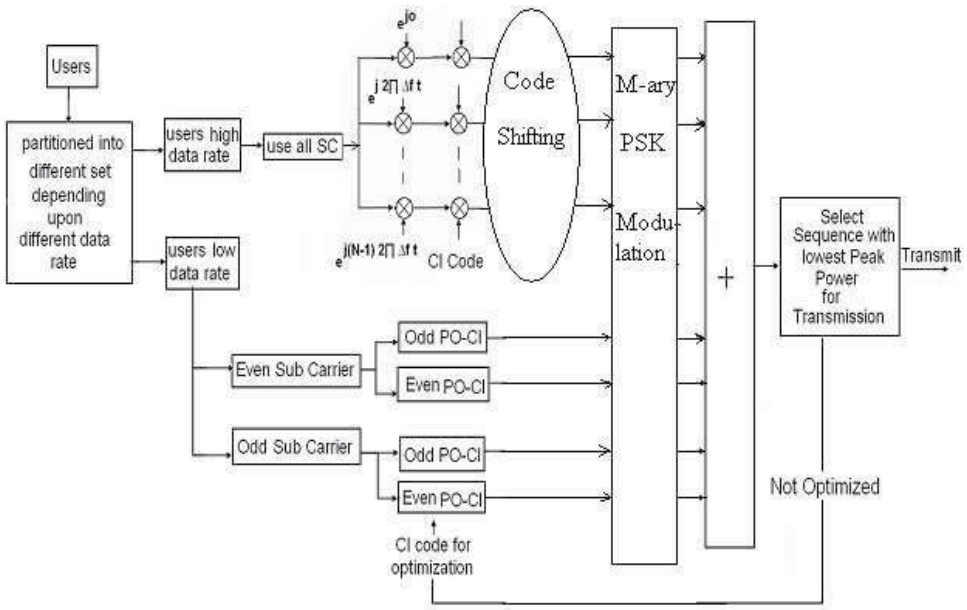


Fig. 1. Block diagram representation of proposed M-ary CI/MC-CDMA PAPR reduction scheme

C. Trellis coding

We also incorporate the effect of trellis coding for further improvement in PAPR reduction. Trellis coding is a kind of convolution coding used here in combination with code and phase optimization so that symbols for the different users are placed in the constellation that may minimize the occurrence of high peak generation in the same odd and even sub-carriers. Moreover, the usage of trellis coding offers the benefit of reduction in bit error rate (BER) performance at the receiver. Fig. 2 shows the simplified block diagram of Fig. 1 with trellis coding.

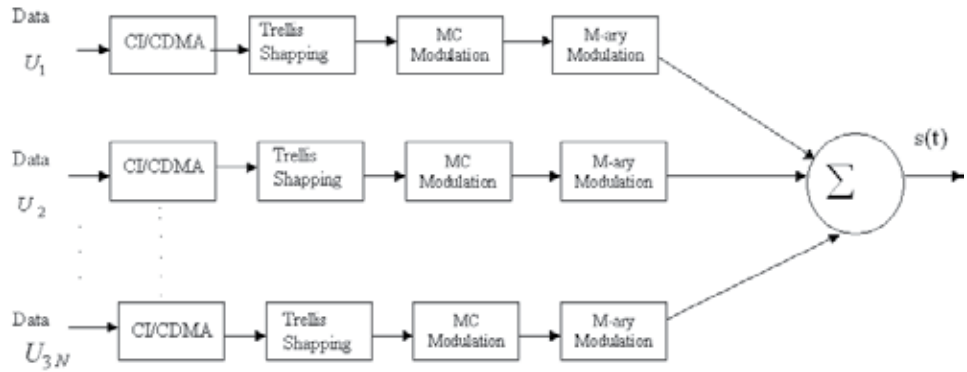


Fig. 2. Simplified block diagram representation for the proposed M-ary CI/MC-CDMA scheme with trellis coding

4.2 Proposed GA based channel estimation

This subsection first defines fitness function 'F' and GA based minimization is presented later.

A. Formation of Fitness function

We assume that actual fading gain at n-th subcarrier is α_n and its estimated value is denoted by $\hat{\alpha}_n$. The error in estimated value is represented by e_n. We can think of probability density function (pdf) of the estimation error as central Chi-square distribution (Ahamad & Assaad; 2009) and can be written as follows:

$$f(|e_n^2|) = \frac{1}{\sigma_n^2} e^{\frac{-|e_n|^2}{\sigma_n^2}} \tag{11}$$

where σ_n^2 is the variance of estimation error. So, in terms of estimated channel gains and its corresponding error terms, the expression of (4) can be written as follows:

$$\begin{aligned} r_i^j &= \hat{\alpha}_i + e_i + \sum_{k=1,k\neq j}^{K} \hat{\alpha}_i \rho_{kj} + \sum_{k=1,k\neq j}^{K} e_i \rho_{kj} + \eta_j \\ &= \hat{\alpha}_i + e_n + \sigma_I^2 + \sigma_{Ie}^2 + \sigma_N^2 \end{aligned} \tag{12}$$

The first, second, third, forth and fifth term of (12) can be designated for the j-user as signal term in $i-$ th subcarrier, estimated error in signal term, variance of interfere i.e. interference power due to all other users at i-th subcarrier, interference power for estimation error, and noise power at i-th subcarrier, respectively. For large number of users, the third and the forth term correspond to a random variable with normal distribution (according to central limit theorem).

We define signal-to-interference noise power ratio (SINR) corresponding to j-th user's bit at i-th subcarrier as follows:

$$(SINR)_i^j = \frac{(|\hat{\alpha}_i| + |e_i|)^2}{\sigma_I^2 + \sigma_{Ie}^2 + \sigma_{N_i}^2} \tag{13}$$

We assume SINR for all subcarriers are independent, so total SINR for j-th user's bit is

$$(SINR)^j = \sum_{i=1}^{N} (SINR)_i^j \tag{14}$$

The channel capacity corresponding to j-th user's is

$$C^j = log(1 + SINR^j)\ bits/sample \tag{15}$$

Now total channel capacity can be calculated as $C = \sum_{\forall j} C^j$.
We now define 'F' as weighted average of channel capacity and detection probability p_e i.e. $F = f(C, p_e)$. It is preferable to minimize 'F', while target is to maximize C and minimize p_e. It is logical to express C as $C_{norm} = \frac{C_{\hat{\alpha},e}}{C_{\alpha=1}}$, that indicates normalization of the channel capacity. The symbol $C_{\alpha=1}$ corresponds to non-fading situation and obviously channel capacity with reliable decoding will be high. It is quite true that the value of $\frac{C_{\hat{\alpha},e}}{C_{\alpha=1}}$ is less than 1 but our target is to achieve this value close to 1. The p_e is calculated as follows:

$$p_e = \frac{1}{K}\sum_{k=1}^{K}(b_k - \hat{b}_k) \tag{16}$$

where, K is total number of users, b_k and $\hat{b}_k$ are the transmitted and the received k-th user bit, respectively.
The objective function 'F' can be defined as

$$F = w_1(1 - C_{norm}) + w_2 p_e = w_1(1 - \frac{C_{\hat{h},e}}{C_{h=1}}) + w_2 p_e \tag{17}$$

where w_1 and w_2 are the weight factors of channel capacity and detection reliability, respectively. Each weight factor represents how important each index is during the searching process of GA. For example, if both indices are equally important, each one should be 0.5 so that the relationship $w_1 + w_2 = 1.0$ must hold.
B. Optimization of Fitness function using GA
The experimental conditions of GA for the present problem are depicted as follows:
(i) size of population is 10,(ii) number of generation/iterations 100, (iii) probability of crossover per generation is 0.80, and (iv) probability of mutation per bit is 0.07.
Initialization of twenty sets of random values for $\hat{\alpha}_1$,$\hat{\alpha}_2$,.....$\hat{\alpha}_n$, e_1,e_2,.....e_n are done. The values of $\hat{\alpha}_i$ are taken from Rayleigh distribution, while the values of e's are taken from (11). Then the value of C and p_e are calculated for each set. Using the procedure outlined in previous subsection, the value of fitness function 'F' is calculated for each of $\hat{\alpha}_1$,$\hat{\alpha}_2$,.....$\hat{\alpha}_n$, e_1,e_2,.....e_n using (17). An upper bound of F value (F_U) is determined based on the calculated 'F' values. The value of (F_U) acts as a threshold and is adjustable. This is required so that the needful number of sets of $\hat{\alpha}_1$,$\hat{\alpha}_2$,.....$\hat{\alpha}_n$, e_1,e_2,.....e_n values for which 'F' values lie below the (F_U) are duplicated. The remaining sets having 'F' values higher than (F_U) are ignored from the population. This process is done from the concept of selection of GA based algorithm. A binary string is generated through decimal-to-binary conversion for each selected set of $\hat{\alpha}_1$,$\hat{\alpha}_2$,.....$\hat{\alpha}_n$, e_1,e_2,.....e_n value and thus a set of strings are calculated for all selected combinations. Now, crossover and mutation operations are done with above probabilities.Operations as described, when applied to the selected sets, generate a new set of $\hat{\alpha}_1$,$\hat{\alpha}_2$,.....$\hat{\alpha}_n$, e_1,e_2,.....e_n value. This set is considered as population for next iteration/generation of the proposed GA based optimization problem. The operations are repeated for desired number of iterations or till a predefined acceptable values for C and p_e are achieved.

4.3 Subcarrier PIC scheme

Multicarrier (MC) signal can be modeled as a vector in N-dimensions where N-mutually orthogonal subcarriers form the basis signals. Accordingly, MC-CDMA signals with high and low data rates can be modeled as vectors of different dimensions in subcarrier signal space. In MUD, CDMA signal is treated as a vector and is projected onto individual orthogonal spreading function to estimate the interference of other users for the respective user. In PIC, for a given user, CDMA composite signal is projected simultaneously to all spreading codes except the desired one and the process demands high computation cost. The multiple access interference (MAI) estimation and its subsequent cancelation in MC-CDMA with variable data rates can be implemented in the respective subcarrier component rather than projecting it onto spreading functions followed by calculation of vector magnitude (Eq. 5 and Eq. 6), as is done in case of conventional PIC method. This very simple concept is used to develop a low computation cost PIC scheme when MAI is subtracted at the sub-carriers level. Fig. 3 shows the block diagram representation of the proposed PIC method. The users are divided in two different blocks, block 1(B-1) with high data rate transmission and block-2 (B-2) with low data rate transmission.

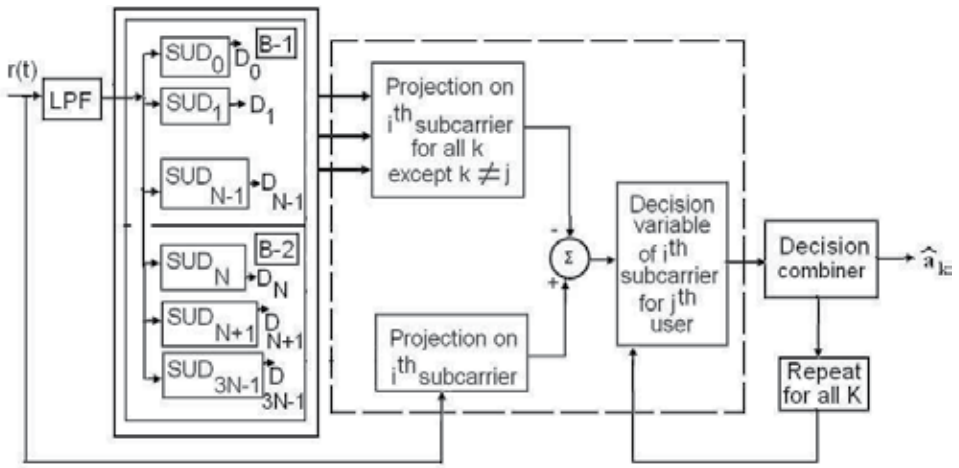


Fig. 3. Block diagram representation of proposed PIC scheme

The received multiuser signal $r(t)$ is first passed through a lowpass filter followed by single user detection through MMSEC. The decision vector $D = [D^1, D^2, ...D^K]$ then generates the initial estimates of the data $\hat{a} = (\hat{a}_1, \hat{a}_2, ...\hat{a}_k)$ which are used to determine the polarity of the interference. To estimate the interference experienced by j-th user at i-th subcarrier, the received signal r(t) and the signal of all other users, except the j-th user, are projected onto i-th subcarrier. The interference of other users are subtracted collectively from the former.
We now analyze mathematically the interference free estimation for the n-th bit of j-th user for conventional PIC, two block PIC and the proposed subcarrier PIC. The expression of the same for the conventional PIC system is

$$\hat{a}_n^j = r_n - \sum_{k=1,k\neq j}^{K} \hat{a}_n^k \tag{18}$$

The similar expression for an arbitrary j-th user belonging to strong group and when belong to weak group, respectively (of a two block PIC system) can be written as follows:

$$\hat{a}_n^j = r_n - \sum_{\forall k_1 \in S_1, k_1 \neq j} \hat{a}_n^{k_1} \tag{19}$$

$$\hat{a}_n^j = r_n - \sum_{\forall k_1 \in S_1} \hat{a}_n^{k_1} - \sum_{\forall k_2 \in S_2, k_2 \neq j} \hat{a}_n^{k_2} \tag{20}$$

where the symbols of $\hat{a}_n^j, r_n$, S_1 and S_2 represent n-th estimated bit of j-th user, resultant received signal for n-th bit, set of strong and weak user groups, respectively.
The interference free estimation of the i-th subcarrier for the j-th user of block B_1,

$$r_i^j = r_i - \sum_{\forall k_1 \in B_1, k_1 \neq j} r_i^{k_1} \tag{21}$$

while the expression for the same belonging to block B_2,

$$r_i^j = r_i - \sum_{\forall k_1 \in B_1} r_i^{k_1} - \sum_{\forall k_2 \in S_2, k_2 \neq j} r_i^{k_2} \tag{22}$$

where r_i^j and r_i indicate the i-th subcarrier component of the j-th user and the i-th subcarrier component of resultant received signal, respectively.
Now, the decision variable D_n^j for the n-th bit of an arbitrary j-th user of block B_1 can be written accordingly to Eq. (5). The equation is rewritten in Eq. (22) for convenience of further analysis. The similar expression for the j-th user of the block B_2 is written in Eq.(23)

$$D_n^j = \sum_{i=1}^{N} r_i^j w_{ij} \tag{23}$$

$$D_n^j = \sum_{\forall i = odd\ or\ even}^{N} r_i^j w_{ij} \tag{24}$$

The values of r_i^j in Eq. (22) and Eq. (23) can be put from Eq. (20) and Eq. (21), respectively.
Mathematical analysis of Eq. (20)-Eq.(23) show that stable decision variable D_n^j can be achieved for subcarrier PIC method compared to conventional PIC and block PIC. This is because interference due to other users are subtracted before forming D_n^j in the case of former unlike the latter methods where D_n^j is formed first and then interference is subtracted. This stable decision variable D_n^j gives rise to better estimation of $\hat{a}_j$ leading to an improvement in BER through interference cancelation. Moreover, the computation cost and time requirement is less compared to the conventional PIC and other block PIC (Thippavajjula; 2004). This can be well understood considering the fact that the users of block-1, transmit data at 'q' times higher rate (say) compared to the users of block-2. So, for 'q' consecutive bits detection of high data rate users, estimated interferences for the single bit due to low data rate users can be made use. In other words, interferences estimated for low users' single bit can be applied to improve subsequent BER of high user data. It is to be mentioned here that the projection of the received signal on the mutually orthogonal subcarriers is essentially the same for both

the conventional PIC and the proposed PIC. However, the interference in the former is subtracted after combining them as decision vector while in the later, interference is subtracted in subcarrier level itself. The computation complexity, for CI codes of length N, is $O(2N-1)$ for conventional PIC, 2 $O(N)$ for block PIC (Thippavajjula; 2004) and $O(N)$ for the proposed subcarrier PIC.

4.4 Optimized system design

In the present study, optimum values for N, K, and SNR are calculated based on the minimization of a fitness function 'F' that depends on PAPR, BER and average data rate (ADR). We first define 'F' and then GA based minimization technique is presented. Let us define the following symbols:

P_{diff}: Difference in threshold PAPR and actual PAPR

B_{diff}:Difference in threshold BER and actual BER

ADR:Average data rate which is normalized with respect to the collective data of all users transmitting at high data rates.

The predefined values of PAPR and BER indicate threshold of the permissible values for the respective parameters. It is to be noted that we have considered P_{diff} and B_{diff}, instead of considering their actual values, while forming the objective function and its subsequent minimization process is performed. This is done so as these two measures are related in a conflicting manner with their associated parameters and the simultaneous minimization of both of these is not optimally achievable with respect to the overall optimal system design is concerned. So the objective is to reduce P_{diff} and B_{diff} whether their actual values are below or above the threshold value. The fitness function 'F' is defined as follows:

$$F = w_1 P_{diff} + w_2 B_{diff} + w_3(1 - ADR) \tag{25}$$

where w_1,w_2 and w_3 are defined as follows:

$$w_1 = \frac{B_{diff} \times (1 - ADR)}{(P_{diff} + B_{diff})(1 - ADR) + B_{diff} P_{diff}} \tag{26}$$

$$w_2 = \frac{P_{diff} \times (1 - ADR)}{(P_{diff} + B_{diff})(1 - ADR) + B_{diff} P_{diff}} \tag{27}$$

$$w_3 = \frac{P_{diff} \times B_{diff}}{(P_{diff} + B_{diff})(1 - ADR) + B_{diff} P_{diff}} \tag{28}$$

so that $0 < w_1$,w_2,$w_3 < 1$ and w_1+w_2+$w_3 = 1$ relation must hold.

The values of w_1,w_2 and w_3 when put in Eq. (24), the expression of 'F' becomes

$$F = \frac{3P_{diff} \times B_{diff} \times (1 - ADR)}{(P_{diff} + B_{diff})(1 - ADR) + B_{diff} P_{diff}} \tag{29}$$

We now describe very briefly GA based optimization process. Initialization of twenty sets of random $\{N, K, SNR\}$ values within the predefined range are done. The values for P_{diff}, B_{diff} and ADR as well as the value of fitness function 'F' for each set of {N,K,SNR} are then calculated. A predefined threshold (F_U) value of 'F' is assigned to identify the fit parameter sets. The particular sets of {N,K,SNR} for which 'F' values lie below the F_U are duplicated and the other sets are ignored from the population following the spirit of GA. Now, crossover

and mutation operations are done on the strings. The experimental conditions of GA for the present problem are depicted as, (i) size of population is 20, (ii) probability of crossover per generation is 0.80, and probability of mutation per bit is 0.09, user capacity i.e. number of users vary from 15-75, number of subcarriers range from 10-25, value of SNR range from 7-14 dB. The operations are done for the desired number of iterations or till a predefined acceptable values for PAPR and BER are achieved for optimal values of N, K and SNR.

5. Performance Evaluation

In this section, we present performance analysis of PAPR reduction, channel estimation, subcarrier PIC for improved receiver operation and finally optimization system for power and spectral efficient system design.

5.1 Performance for PAPR reduction

This subsection presents the performance of the proposed PAPR reduction technique in synchronous M-ary CI/MC-CDMA system. As figures of merit, we evaluate the complementary cumulative distribution function (CCDF) of PAPR under various phase modulation. We also show through simulation results that the proposed CI/MC-CDMA method offers lower PAPR and BER values compared to conventional CI/MC-CDMA system (Natarajan et al; 2001) and other recent work [12]. We first show the cross-correlation values for the different user's spreading codes under the phases as described in equation (7). The cross correlation (CC) between user k's signature waveform $c_k(t)$[created using the spreading sequence (1, $e^{j\Delta}\theta_k$,... $e^{j(N-1)\Delta}\theta_k$) and user j's signature waveform [created via the spreading sequence (1, $e^{j\Delta}\theta_j$,... $e^{j(N-1)\Delta}\theta_j$) can be shown as follows

$$R_{k,j}(\tau) = \frac{1}{2\Delta f} \sum_{i=0}^{i=N-1} cos(i(2\pi.\Delta f\tau)) \tag{30}$$

$$R_{k,j}(\tau) = \frac{1}{2\Delta f} \frac{sin(1/2N.2\pi.\Delta f.\tau)}{sin(1/2.2\pi.\Delta f.\tau)} cos(\frac{(N-1)}{2} 2\pi\Delta f\tau) \tag{31}$$

where $\tau = \Delta t_k - \Delta t_j = (\Delta\theta_k - \Delta\theta_j)/2\pi.\Delta f$. Table 1 shows the cross-correlation values for some arbitrary pair of spreading codes under various phase shifts. Numerical values show that $R_{j,k}$ values are reduced significantly for the proposed phase shifts of the code patterns.

We now show cumulative complementary distribution function (CCDF) of PAPR values for different M-ary modulations with separate code and phase optimization of the spreading codes in Fig. 4 and as well as with the combined effect of code and phase optimization in Fig. 5. It is also to be mentioned here that the proposed PAPR reduction algorithm is very fast. It takes approximately 2.25 second for 10 time iteration in MATLAB 7 platform running on a Pentium IV 2.79 GHz, 512 MB RAM PC system. The graphical representation shows that best PAPR reduction performance is achieved for the proposed algorithm with QPSK system, while the worst PAPR reduction performance is shown in BPSK performance. Similarly, PAPR reduction performance of 16-ary PSK is comparatively better than 8-ary PSK system. The results obtained are quite consistent with the cross-correlation values shown in the Table 1 for any arbitrary pair of code patterns and different M-ary modulation. While phase quadrature in QPSK system is not much influenced by the polarity of the symbols, but in case of BPSK modulation the polarity of symbols makes superimposition of subcarriers for in-phase that leads to an increase in PAPR value. It is interesting to see that for BPSK, QPSK and 8-ary PSK

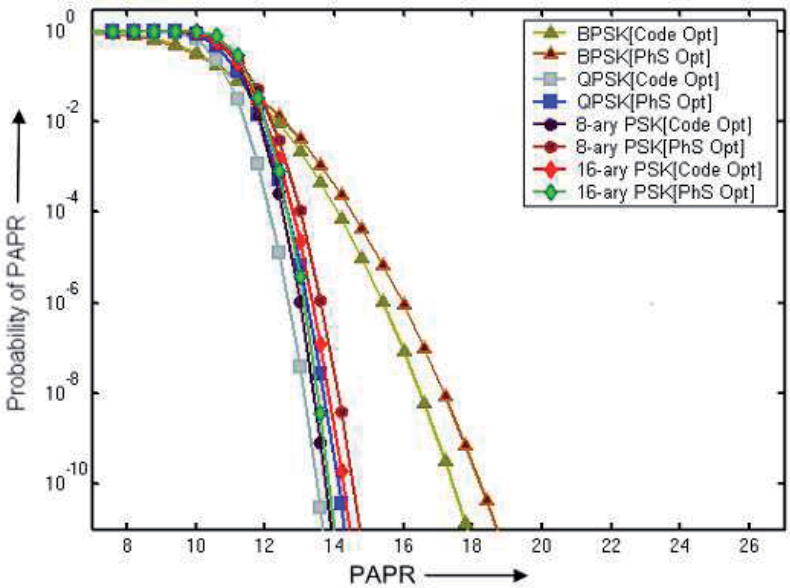


Fig. 4. CCDF of PAPR for N=8 and K=24 with code and phase optimization separately

code optimizations are more effective for reduction in PAPR values. On the other hand, for 16-ary PSK, phase optimization is more effective. Significant improvement in PAPR reduction is possible to achieve by exploiting the combined effect of code and phase optimization. Fig. 5 shows that PAPR values are not only reduced but also the probability for the PAPR values over a particular threshold value also become low. Fig. 6 shows improvement in PAPR using trellis coding.

We also test PAPR reduction performance of the proposed system in real time system of Agilent made N5182A MXG RF vector signal generator (VSG). The signal characteristics is 100 kHz to 3 GHz, +23 dBm output power at 1GHz = -121 dBc/Hz (typ), phase noise at 1 GHz and 20 kHz offset =1.2 ms switching speed in SCPI mode; $\leq$ 900 micro-second simultaneous frequency, amplitude, and waveform switching in list mode. Fig. 7 shows CCDF for PAPR reduction when the proposed algorithm is generated in VSG before code and phase optimization while the same is shown in Fig. 8 after code and phase optimization. The results show that due to optimization peak power is not only reduced from 6.37 dB to 4.06 dB, but also probability for attaining 0dB value also increases from 37% to 46 %. Similarly, 10% probability is found for PAPR value with 3.52dB, while after optimization this value reduces to 3.06dB. So theoretical performance improvement of PAPR reduction is also verified in hardware system indicating the importance of implementation of the proposed algorithm in practical system.

PAPR reduction for the proposed algorithm is also verified in VSG for 8-ary PSK modulation. Results are shown in Fig. 9 and Fig. 10. The results show that due to optimization not only peak power is reduced from 7.27 dB to 6.13 dB, but also probability for attaining 0dB value increases from 29% to 35 %. Similarly, 10% probability is found for PAPR value with 3.35dB, while after optimization this value reduces to 3.00dB. Verification results once again support the theoretical results that PAPR reduction performance for the proposed algorithm with 8-ary and 16-ary modulation system lie between performance for BPSK and QPSK modulation.

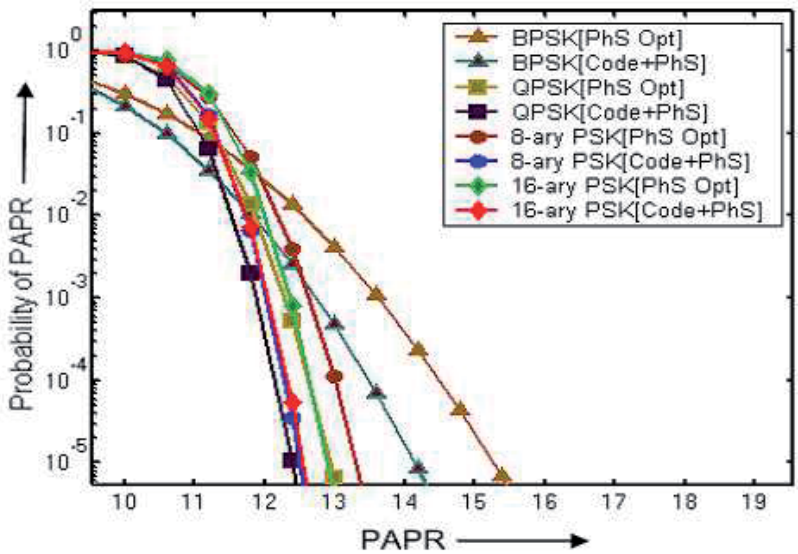


Fig. 5. CCDF of PAPR for N=8 and K=24 with combined code and phase optimization

5.2 Performance evaluation for channel estimation

This subsection represents performance analysis of the proposed channel estimation method. We have evaluated BER for different number of users at SNR=7dB and the number of subcarriers N=10. Fig. 11 shows that BER values corresponding to the estimated channel parameters become more closer to the values corresponding to the actual values of the channel parameters. Simulation results show that for N=10, difference in BER value corresponding to the actual and the estimated channel parameters become of the order of 0.085 for the number of users 20, and is of the order of 0.055 for the number of users 30. We also test the effect of the number of iterations/generations on the estimated channel parameters in terms of detection reliability and channel capacity. Fig. 12 shows that BER values increase with the increase of number of users but decreases with the number of generations. At the same time normalized (maximum values is 1) data transmission capacity is increased significantly with the number of generations which is shown in Fig. 13.

We also compare performance of the proposed method with the method in (Doukopoulos & Moustakides; 2004). For compatibility, orthogonal frequency division multiplexing (OFDM) system of (Doukopoulos & Moustakides; 2004) is designed as MC-CDMA system for the same number of subcarriers N=10 and simulation is performed at SNR=7 dB. Simulation results shown in Fig. 14 reflects the fact that the proposed channel estimation offers better BER performance due to the optimal selection of GA. The computation complexity for the proposed method is $O(n)$, while the same for LMS method in (Doukopoulos & Moustakides; 2004) is $O(n^2)$.

5.3 Performance evaluation of subcarrier PIC

Fig. 15 shows BER performance of the proposed scheme at 14 dB SNR and with N=16 subcarriers for 3N users system where N number of users transmit high data rate and 2N number of users transmit low data rate. The graphical result shows that when capacity is low (less

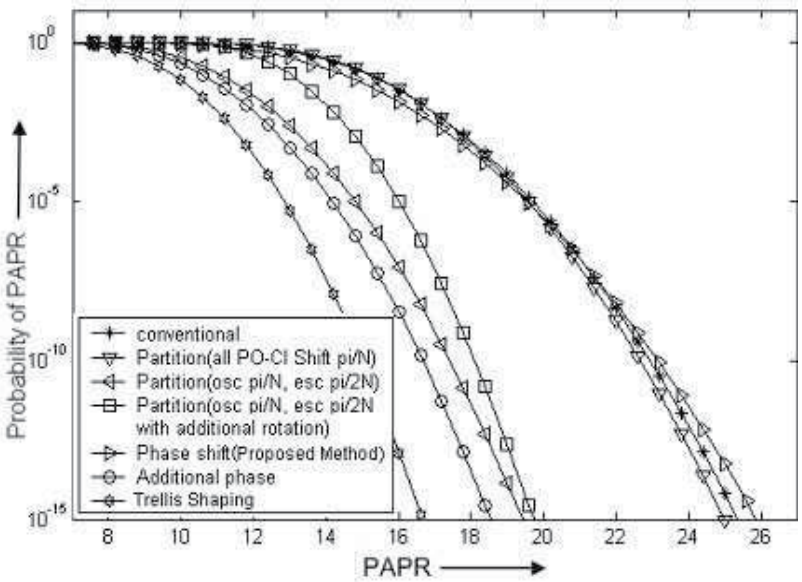


Fig. 6. CCDF of PAPR for N=8 and K=24 using trellis coding

than 20 users), MMSEC receiver of conventional CI/MC-CDMA system (shown as Con. MMSEC in the figure) performs significantly better compared to MMSEC receiver of the proposed system (shown as MMSEC in the figure). This is due to the fact that all users in the former transmit data using all sub-carriers while some of the users in the latter transmits data using either odd or even sub-carriers. However, significant improvement in BER performance is achieved in the latter case compared to the former when the number of users are gradually increasing over 20. Further significant improvement in BER performance can be achieved after different stages of the proposed subcarrier PIC scheme. Simulation results show that the proposed subcarrier PIC scheme after third stage iteration, can support the number of users three times the number of sub-carriers with BER of the order of 0.0428 (0.11071 for (Natarajan et al; 2001)), and it can support users upto four times the number of sub-carriers with BER of the order of 0.1350 (0.2193 for (Natarajan et al; 2001)).

We also test BER performance of the proposed scheme with the increase of number of users transmitting at high data rate. Fig. 16 shows BER performance of the proposed scheme at SNR 14 dB with N=16 sub-carriers for 2.5N users system. Here 1.5 N number of users transmit at high data rate and N number of users transmit low data rate. Simulation results show that the system supports the number of users two-and-half and three times the number of sub-carriers with BER values of 0.0591 and 0.104, respectively after three stage iterations. The relative degradation in BER performance for 2.5 N system, over 3N system, based on the proposed subcarrier PIC is due to the increase in overall data transmission rate for the former compared to the latter. The over all data transmission rate is 0.7742 times for 3N user and 0.9032 times for 2.5 N system with respect to the conventional CI/MC-CDMA system (Natarajan et al; 2001). The numerical values specified here for data transmission rate is obtained when transmission rate between high and low data rate user differs by a factor of 4.

We also compare BER performance of the proposed subcarrier PIC scheme (SCPIC) and block PIC (BPIC) scheme (Thippavajjula; 2004). The results are reported with the performance of

Arbitrary code pairs	Diff. PSK Sys.	Two Ort. cod. (j,k)	Orth.Code (j) $\pm\pi/M$ Shift. codes	$+\pi/M$ & $-\pi/M$.shift. codes
1st pair	BPSK	0.3953	0.3750	0.3536
	QPSK	0.2812	0.2789	0.2542
	8- PSK	0.3162	0.3149	0.3020
	16- PSK	0.3617	0.3411	0.3211
2nd pair	BPSK	0.3953	0.3750	0.3536
	QPSK	0.2834	0.2791	0.2590
	8- PSK	0.3179	0.3151	0.3001
	16- PSK	0.3619	0.3459	0.3291
3rd pair	BPSK	0.3953	0.3750	0.3536
	QPSK	0.2834	0.2791	0.2590
	8- PSK	0.3179	0.3151	0.3001
	16- PSK	0.3619	0.3459	0.3291
4rth pair	BPSK	0.3962	0.3690	0.3542
	QPSK	0.2842	0.2709	0.2581
	8- PSK	0.3287	0.3124	0.3054
	16- PSK	0.3754	0.3589	0.3205

Table 1. Cross correlation values for arbitrary code pair

Sl. no	N	K	SNR	PAPR	BER	ADR
1	22	64	10	10.2649	0.0482	0.7444
2	24	70	10	11.4017	0.0450	0.6700
3	21	60	14	10.0973	0.0125	0.8100
4	23	68	14	11.4012	0.0650	0.5000
5	18	52	11	10.4698	0.0464	0.5577
6	21	65	12	10.8132	0.0414	0.5407
7	23	67	12	11.7433	0.0701	0.5135
8	19	57	14	10.066	0.0422	0.5716

Table 2. GA based optimization for the proposed method

(Natarajan et al; 2001) through interference cancelation (IC). Fig. 17 shows that BER performance of the proposed subcarrier PIC scheme is significantly better compared to that of block PIC scheme and needless to mention its superior BER performance compared to MMSE scheme of (Natarajan et al; 2001). The performance improvement for the proposed subcarrier PIC is due to the twofold advantages in interference cancelation. Since the high data rate transmission uses all sub-carriers, the data can be decoded with greater reliability and interference due to these users can be estimated with greater accuracy. This interference when subtracted from the resultant received signal improves detection performance of the low data rate users. On the other hand, low data rate transmission uses alternate sub-carriers, so subcarriers of high data rate users experience less interference that leads to an improvement in BER performance of the latter. This cumulative effect on BER performance in multistage interference cancelation significantly improves overall BER performance of the proposed system unlike to that of the block PIC scheme in (Thippavajjula; 2004). In block PIC, in any stage

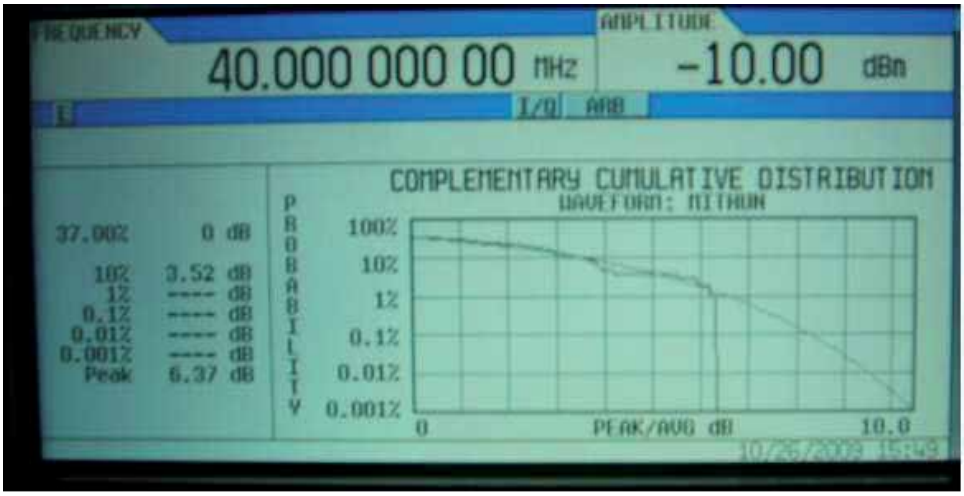


Fig. 7. CCDF of PAPR in QPSK for N=24 before code and phase optimization

Sl. no	N	K	SNR	PAPR	BER	ADR
1	25	54	14	18.6040	0.0770	1.000
2	23	48	14	16.0504	0.0691	1.0000
3	21	44	13	14.1088	0.0805	1.0000
4	20	35	10	13.1250	0.0691	1.0000
5	22	43	12	15.3944	0.0714	1.0000
6	18	33	11	13.1597	0.0670	1.0000
7	24	46	10	16.9413	0.0683	1.0000

Table 3. GA based optimization in

of interference cancelation, weak users data noway benefits BER performance of strong users data unlike the proposed subcarrier PIC scheme.
Fig. 18 shows graphical representation for BER performance with the number of users for subcarrier parallel interference cancelation (PIC) (Maity & Mukherjee; 2009), code and subcarrier PIC and trellis coded system for 'N'=16 and SNR=14 dB. It is found that trellis coded system provides significant improvement in BER performance even at less number of interference cancelation stage compared to the same for higher stage interference cancelation of subcarrier PIC and combined code & phase PIC.

5.4 Performance evaluation of optimized system

Table 2 and Table 3 show the performance of the optimization for the proposed system and CI/MC-CDMA system in (Natarajan et al; 2001), respectively. Simulation results clearly specify the importance of the optimization problem. The values of PAPR, BER and ADR for both the optimized systems are quite consistent for the particular combinations of K, N and SNR values i.e. large N values offer lower BER and increased data transmission rate, while large K values yield increased BER. The values of SNR have both way effect on BER performance in multiuser communication system. As a matter of fact, a set of N, K, SNR values are (at least) near optimal for the set of PAPR, BER and ADR values with respect to the status of the wireless channel condition. For example, if we see the results depicted in 4th row (Sl. no 3)

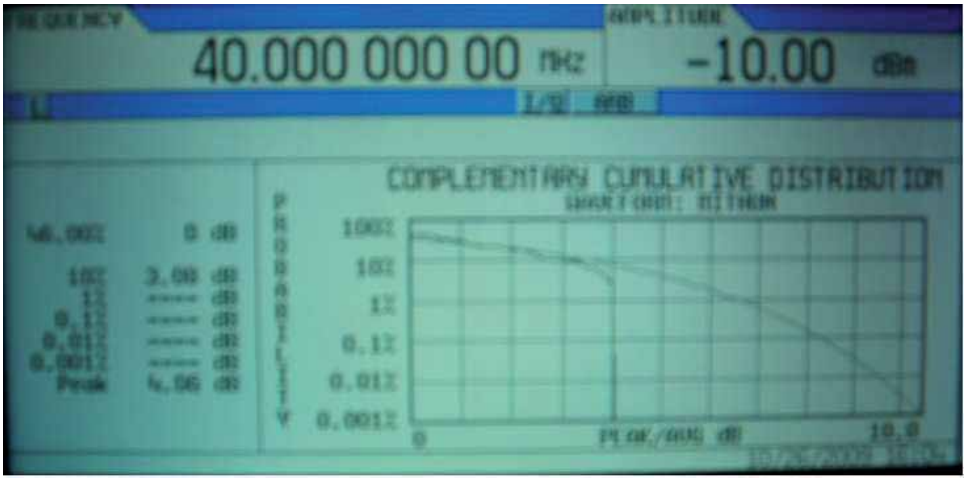


Fig. 8. CCDF of PAPR in QPSK for N=24 after code and phase optimization

of Table 2 and Table 3, for N=21, and for similar SNR values (14 dB for the proposed system and 13 dB for (Natarajan et al; 2001)), PAPR values of proposed system is lower compared to (Natarajan et al; 2001) due to improved PAPR reduction performance for the proposed system. At the same time a significant improvement in BER is achieved for our method due to novelty of the proposed subcarrier scheme, even at nearly 1.5 times increase in user capacity. Similar explanation is applicable for other set of results in Table 2 and Table 3.

6. Conclusion

This chapter discusses a new model of high capacity CI/MC-CDMA system with variable data rates along with simple, fast and efficient PAPR reduction at transmitter and subcarrier PIC scheme at receiver. PAPR reduction is achieved through phase shift of pseudo-orthogonal codes with respect to the orthogonal spreading codes assigned for low and high data rate transmission, respectively. The algorithm has been extended for M-ary PSK system. Significant reduction in PAPR is achieved with combined code and phase optimization in conjunction with trellis coding. Simulation results show that code optimization is more effective for PAPR reduction in BPSK, Q-PSK and 8-ary PSK while phase optimization is effective for the same in case of 16-ary PSK. In the receiver, a simple, fast and efficient subcarrier PIC scheme is proposed. BER performance of the proposed method not only shows improved result compared to the conventional PIC and block PIC system but also requires low computation complexity. The scope of usage of genetic algorithms for the estimation of channel parameters for the proposed MC-CDMA system is then explored. The results reported here show that with the increase of number of users, BER values corresponding to the estimated parameters closely follow to that of BER values obtained for actual parameters values. Simulation results also show that with the increase of number of generations both BER values decrease and channel capacity increases. Finally, GA based optimized system is designed to achieve acceptable values of PAPR, BER and ADR for optimal set of the number of users, the number of subcarriers and SNR values based on the status of the wireless channel.The prposed system may be used as a potential multiple access with broadband data transmission for both uplink and downlink satellite system in conjunction with mobile communication.

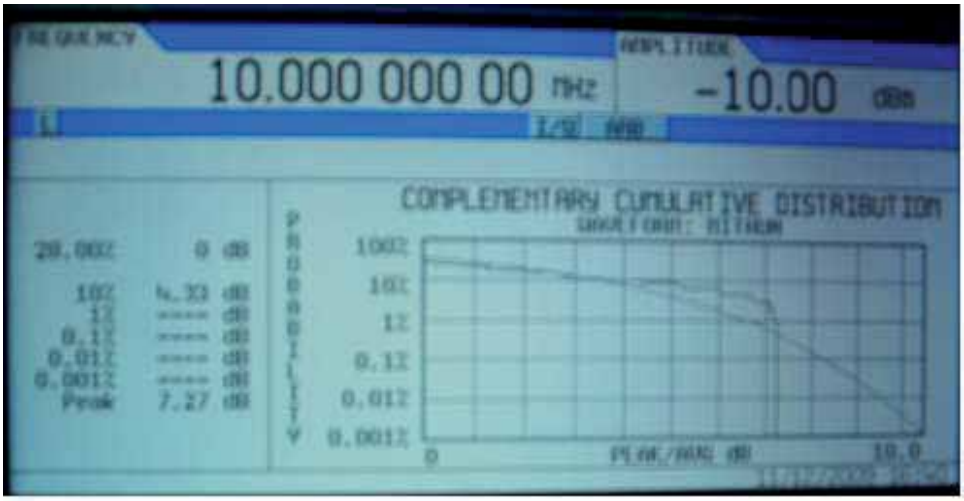


Fig. 9. CCDF of PAPR in 8-ary PSK for N=24 before code and phase optimization

Fig. 10. CCDF of PAPR in 8-ary PSK for N=24 after code and phase optimization

Acknowledgment

The author acknowledge financial support for the project on “Development of high power and spectral efficiency multiuser system for broadband wireless communication" funded by Ministry of Communication and Information Technology, Govt. of India vide administrative approval no. 13(2)/2008-CC & BT dated 31.03.2008.

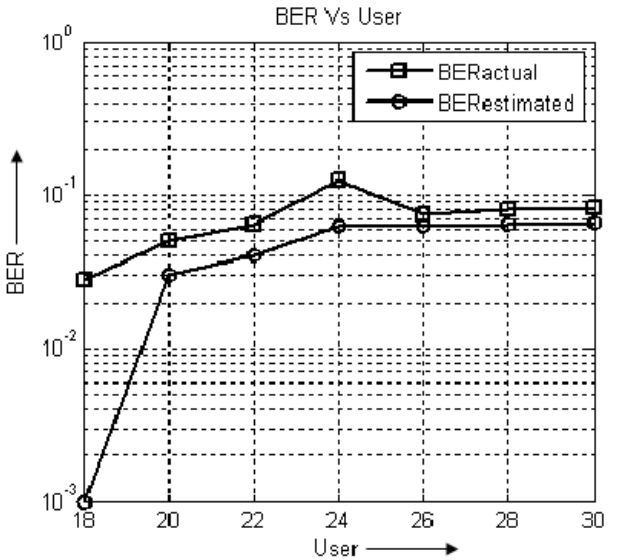


Fig. 11. BER comparison for estimated and actual channel parameters

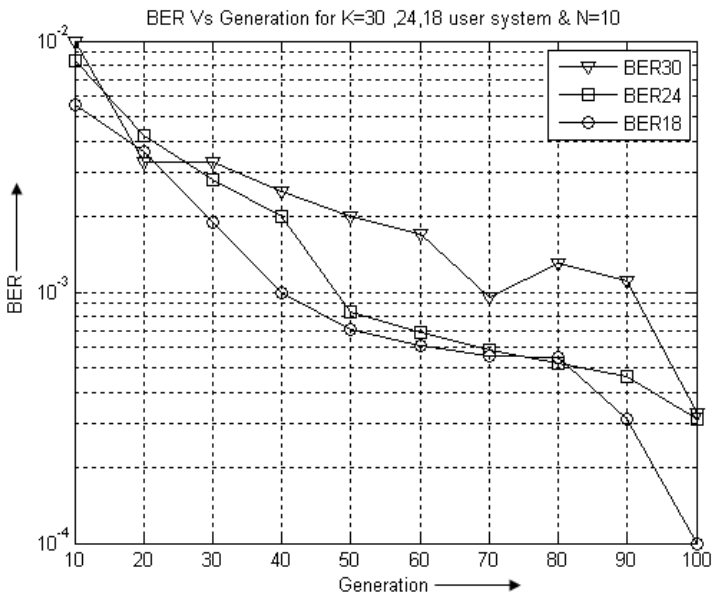


Fig. 12. BER performance with the number of generations

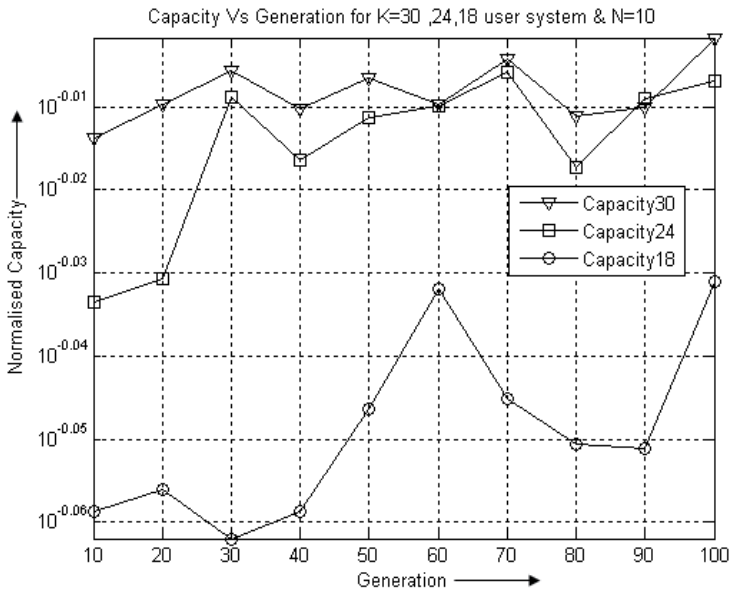


Fig. 13. Channel capacity with number of generations

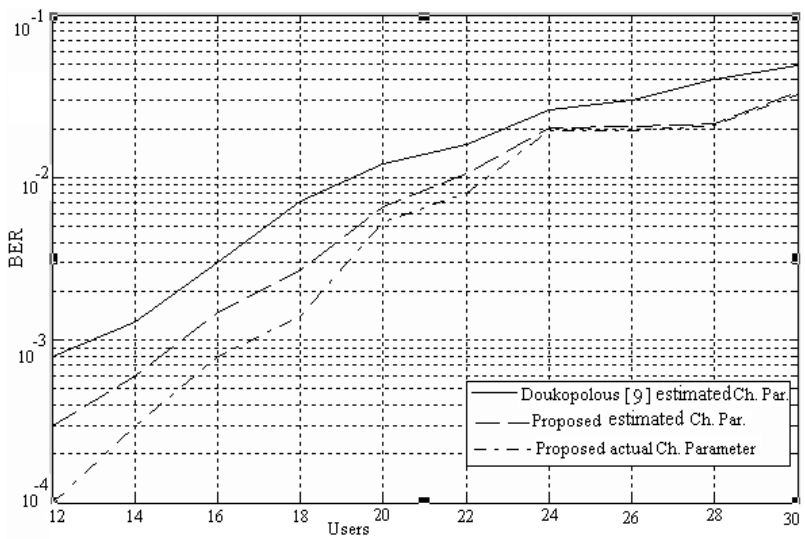


Fig. 14. Comparison of BER performance through channel estimation using N=10 and SNR=7dB

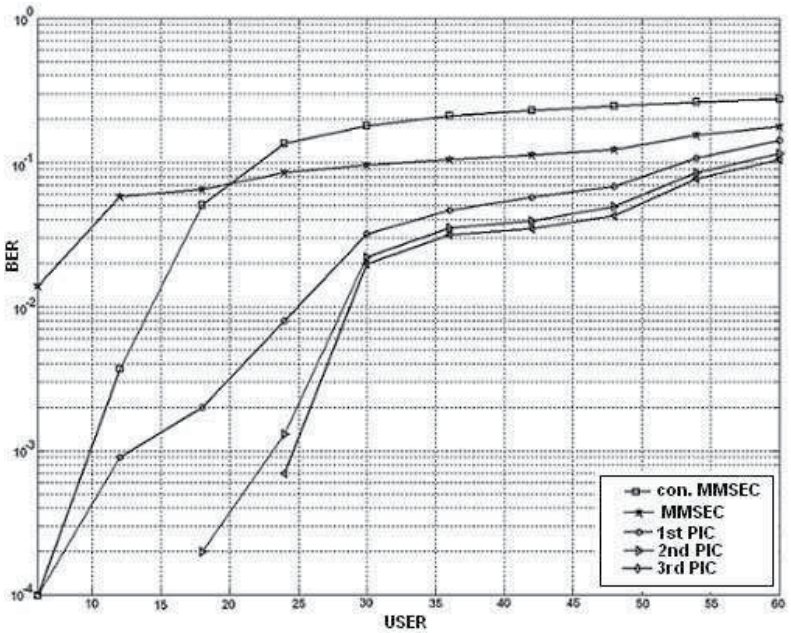


Fig. 15. Performance of subcarrier PIC scheme for 3N user system

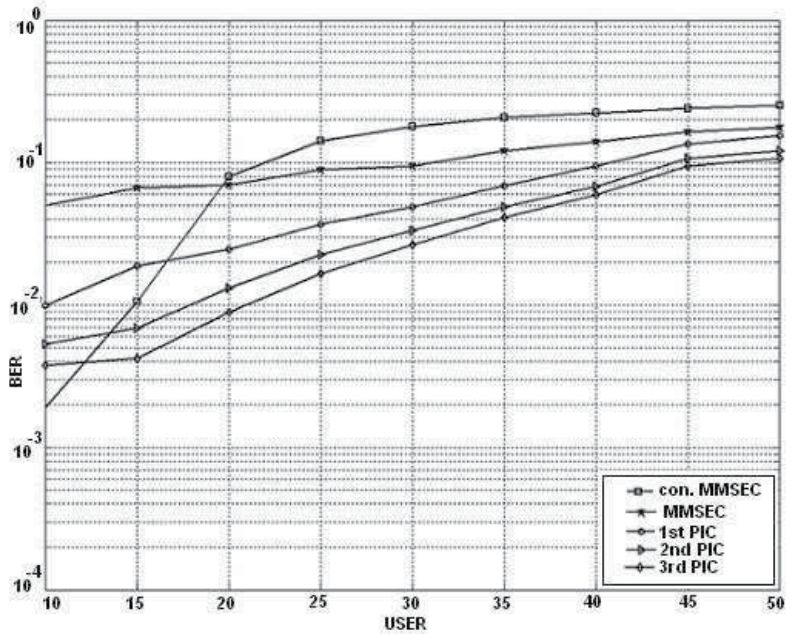


Fig. 16. Performance of subcarrier PIC scheme for 2.5N users system

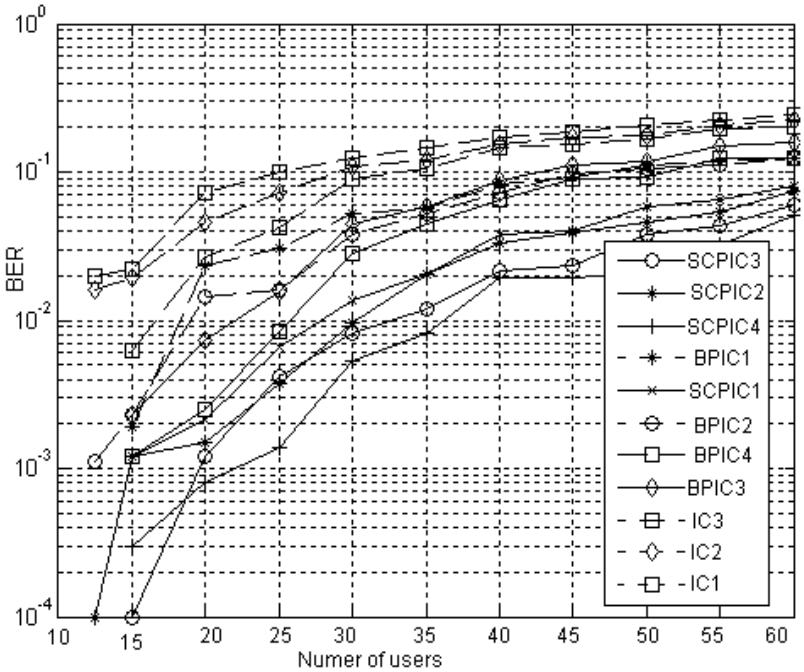


Fig. 17. Performance comparison of subcarrier PIC & block PIC schemes for 3N users system

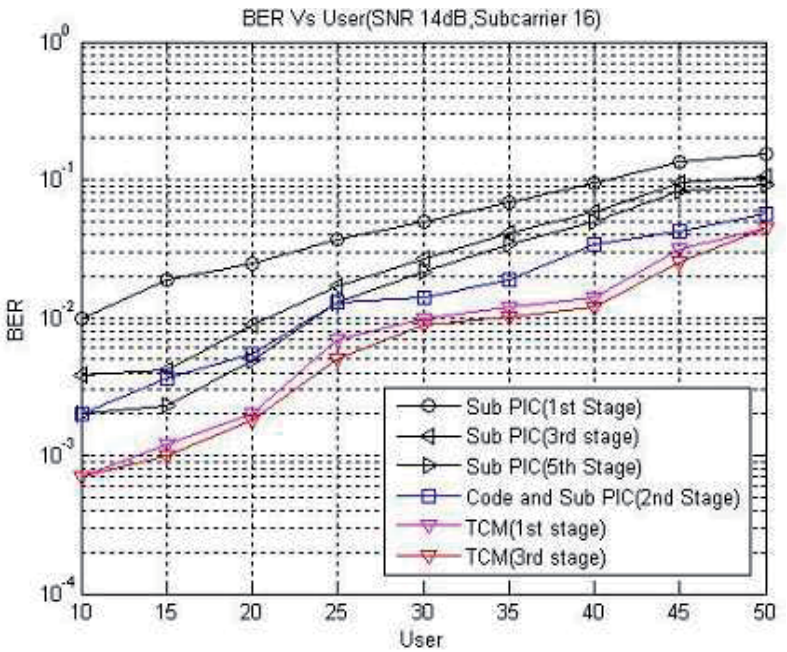


Fig. 18. BER performance for subcarrier PIC, Code and subcarrier PIC and trellis coded system for N=8 and K=24

7. References

Ochiai H. and Ima H.(2000). Performance of the deliberate clipping with adaptive symbol selection for strictly bandlimited OFDM systems. *IEEE Journal on Selected Areas in Communications*, Vol. 18, No. 11, (2000) (2270-2277)

Lim D. W., Heo S. J., No J. S., and Chung, H. A New PTS OFDM Scheme with Low Complexity for PAPR Reduction. *IEEE Tran. Broadcasting* Vol. 52, No. 1, (2006)(77-82).

Yoo, S., Yoon, S., Kim S. Y. and Song, I. A novel PAPR reduction scheme for OFDM systems: Selective Mapping of Partial Tones (SMOPT). *IEEE Trans. on Consumer Electronics*, Vol. 52, No. 1, (2006) (40-43)

Ochia, H. A novel trellis-shaping design with both peak and average power reduction for OFDM systems. *IEEE Trns. on Communication*, Vol. 52, No. 11, (2004)(1916-1926).

Kang, K., Kim, S., Ahn, D. and Lee, H.J. Efficient PAPR reduction scheme for satellite MC-CDMA systems, *IEE Proc. on Communication*, Vol. 152, No. 5, (2005)(697-703).

Vedu, S. Minimum probability of error for asynchronous gaussian multiple access channels. *IEEE Transactions on Inform. Theory*, Vol. 32, (1986)(85-96)

Lupas, R. and Verdu, S. Linear multiuser detectors for synchronous code division multiple access channels. *IEEE Transactions on Inform. Theory*, Vol. 35,(1989)(123-136)

Divsalar, D., Simon, M. K. and Raphaeli, D. Improved Parallel Interference Cancelation for CDMA. *IEEE Trans. Communication*, Vol. 46, No. 2 (Feb 1998)(258-268)

Kim, S. and Lee, J. H. Performance of iterative multiuser detection with a partial PIC detector and serial concatenated codes, *IEEE VTS 54th Vehicular Technology Conference*, Vol. 1, pp.487-491, 2001.

Xiao, L. and Liang, Q. The study of parallel interference weighted canceler multiuser detection, *IEEE VTS 50th Vehicular Technology Conference*, Vol. 5,pp.3009-3013, 1999.

Thippavajjula, V. and Natarajan, B. Parallel interference cancelation techniques for synchronous carrier interferometry/MC-CDMA uplink, *IEEE Vehicular Technology conference*,pp.399-403, 2004.

Maity, S. P., Hati, S. and Maity, S. Diversity Assisted Block PIC for Synchronous CI/MC-CDMA Uplink Systems Using Genetic Algorithm, *Proceedings of the third IEEE International Conf. on Industrial and Information System*, Indian Institute of Technology, Kharagpur, India, (December 2008).

Sgraja, C. and Linder, J. Estimation of Rapid Time- Variant Channels for OFDM using Wiener Filtering, Proc. IEEE Int. Conf.on Comm., Vol. 4, pp. 2390-95, 2003.

Chow, J. S., Tu, J. C. and Cioffi, J. M. A discrete multitone transceiver system for HDSL application. *IEEE J. Select.Areas Communication*, Vol. 9,(Aug. 1991)(895Ű-908).

Ziegler, R. A. and Cioffi, J. M. Estimation of time-varying digital radio channel. *IEEE Trans. Veh. Tech.*, Vol. 41, (1992)(134Ű-151).

Wang, X. and Ray Liu, K. J. Adaptive channel estimation using cyclic prefix in multicarrier modulation system. *IEEE Commun. Lett.*, Vol. 3, No. 10, (1999)(291-293).

Choi, Y. S., Voltz, P. J. and Cassara, F. A. On channel estimation and detection for multicarrier signals in fast and selective Rayleigh fading channels. *IEEE Trans. on Communication*, Vol. 49, No. 8,(2001)(1375-1387).

S. Coleri, M. Ergen and A. Puri, A study of channel estimation in OFDM systems, IEEE Globecom, 2002.

P. Schramm and R. Mullar, Pilot symbol assisted on Rayleigh fading channels with diversity: Performance analysis and parameter optimization. *IEEE Trans. on Communication*, Vol. 46, No. 12, (1998)(1560-1563).

Doukopoulos X. G. and Moustakides, G. V. Blind adaptive channel estimation in OFDM systems, *Proc. Of IEEE ICC*, Vol. 4, (2004)(20-24).

Gupta, P. and Mehra, D. K. Kalman filter based equalization for ICI suppression in High mobility OFDM systems, *Proc. of 13th National Conf. on Commun.*, (NCC-07), IIT Kanpur, pp.21-25, 2007.

Ramesh, C., Jawakar P. K., and Vaidehi, V. Pilot based adaptive channel estimation for OFDM system using GS FAP algorithm, *Proc. of 12th National Conf. on Commun. (NCC-2006)*, IIT Delhi, pp. 94-98, 2006.

Gao, X., Jiang, B.,You, X., Pan, Z., Xue, Y. and Schulz, E. Efficient Channel Estimation for MIMO Single-Carrier Block Transmission With Dual Cyclic Timeslot Structure. *IEEE Trans. on Communications*, Vol. 55, no. 11, (November 2007), (2210-2223).

Lok, T. M. and Wong, T. F. Transmitter and Receiver Optimization in Multicarrier CDMA Systems. *IEEE Transaction on Communication*, (2000)(1197-1207).

Wu, Q. Performance of optimum transmitter power control in CDMA cellular mobile systems. *IEEE Transaction on Vehicular Tech.*, Vol. 48, (1999).

Reynolds, D. and Wang, X. Adaptive transmitter optimization for blind and group-blind multiuser detection. *IEEE Trans. on Signal Proc.*, Vol. 51, (2003)(825-38).

Kim, D. Rate-regulated power control for supporting flexible transmission in future CDMA mobile networks. *IEEE Journal on Selected Areas Commmunications*, Vol. 17, (1999)(968-977).

Buzzi, S. and Poor, H. V. Joint Transmitter and Receiver Optimization for Energy-Efficient CDMA Communications. *IEEE Journal Selected Areas Communication -Special issue on multiuser detection for adv. commun. and networks*, Vol. 26,(Apr. 2008)(pp. 459-472).

Seo, K. and Yang, L. Joint transceiver optimization in MC-CDMA systems exploiting multipath and spectral density, *IEEE GLOBECOM Proceedings*, pp. 1-5, 2006.

Yee, N., Linnartz, J. P. and Fettweis, G. BER of multicarrier CDMA in indoor wireless networks. *IEEE Trans. on Commun.*, Vol. 1, (1993)(2817-2821).

Natarajan, B., Nassar, C.R., Shattil, S., Michelini, M. and Wu, Z. High performance mc-cdma via carrier interferometry codes. *IEEE Transactions on Vehicular Technology*, Vol. 50, (2001)(1344-1353).

Maity, S. P., Hati, S., Maity, S. and Mandal, M. K. Transmitter Optimization in Diversity Assisted Synchronous CI/MC-CDMA Uplink Systems Using Genetic Algorithm, *24th IEEE Queen's Biennial Symposium on Communications (QBSC-2008)*, Canada, pp. 62-67, 2008.

Proakis, J.G. *Digital communications*, 3rd Ed,McGraw-Hill, 1995.

Cal, X. and Akansu, A. N. Multicarrier CDMA systems with transmit diversity. *IEEE Transactions on Vehicular Technology*, Vol. 2, (2000)(658-661)

Ahamad, A. and Ali Assaad, M. Margin adaptive resources allocation in downlink OFDMA system with outdated channel state information, *PIMRC 2009*, Japan.

Doukopoulos, X. G. and Moustakides, G. V. Blind adaptive channel estimation in OFDM systems, *Proc. Of IEEE ICC*, Vol. 4, pp. 20-24, 2004.

Maity S. P. and Mukherjee M., Subcarrier PIC scheme for high capacity CI/MC-CDMA System with Variable Data Rates, IEEE Mobeile WiMAX'09, July 9-10, Napa Valley, California, pp. 135-140, 2009.

19

Quantum Based Information Transfer in Satellite Communication

Laszlo Bacsardi and Sandor Imre
Department of Telecommunications, Budapest University of Technology and Economics
Hungary

1. Introduction

The first electronic computer, the ENIAC (Electronic Numerical Integrator And Computer) was developed in 1943 at the University of Pennsylvania to calculate artillery firing tables. It contained around 17500 vacuum tubes and it weighed about 27 tonnes. Since that we construct smaller and smaller computers from year to year, whose performance is becoming higher and higher. Gordon Moore, co-founder of the Intel Corporation examined the number of transistors that can be placed inexpensively on an integrated circuit in 1965. He found that this number had doubled every second year. In his original paper he examined the time interval between 1958 and 1965. However, the trend – called Moore-law - has continued more than half a century and is not expected to stop in the next five years (Moore, 1965). The law is represented on Fig. 1.

Capabilities of many electronic devices are linked to the Moore-law, for example processor speed, memory capacity etc. We can observe a continuing size decreasing in the field of integrated circuits as well. The growth in the performance of the processor is due to putting more and more transistors on the microchip of same size. This requires smaller and smaller transistors, which can be achieved if we are able to draw thinner and thinner lines onto the surface of a semiconductor disk. The big question is how long this trend can continue? We will reach the limit of our technology and won't be able to place more transistors on an integrated circuit. Researches offer different solutions for this problem like using parallel computers, DNS-technology or informatics based on quantum mechanics. Why quantum mechanics? If we want to place more transistors on an integrated circuit of a given size, the size of transistors have to be decreased. At a point we will cross the line to the world of the atoms. In that world the classical Ebers-Moll equals are not valid anymore, and quantum mechanical equals have to be used instead. Informatics based on quantum mechanical models is called quantum informatics.

In the last years, quantum theory has appeared in satellite communications offering answers for some of nowadays' technical questions. Although quantum computers are going to be the tools of the far future, there exist already algorithms to solve problems which are very difficult to be solved by traditional computers (Imre & Ferenc, 2005).

The quantum informatics can play a key role in the field of cryptography. In present classical cryptographic methods, the key exchange is generally based on public key

methods. The security of modern cryptographic methods like asymmetric cryptography, relies heavily on the problem of factoring integers. In the future, if quantum computers become reality, any information exchange using current classical cryptographic schemes will be immediately insecure. Current classical cryptographic methods are not able to guarantee long-term security. Other cryptographic methods, with absolute security must be applied in the future (Gyongyosi & Imre, 2009). The quantum cryptography gives better solutions for communication problems than the classical cryptographic methods.

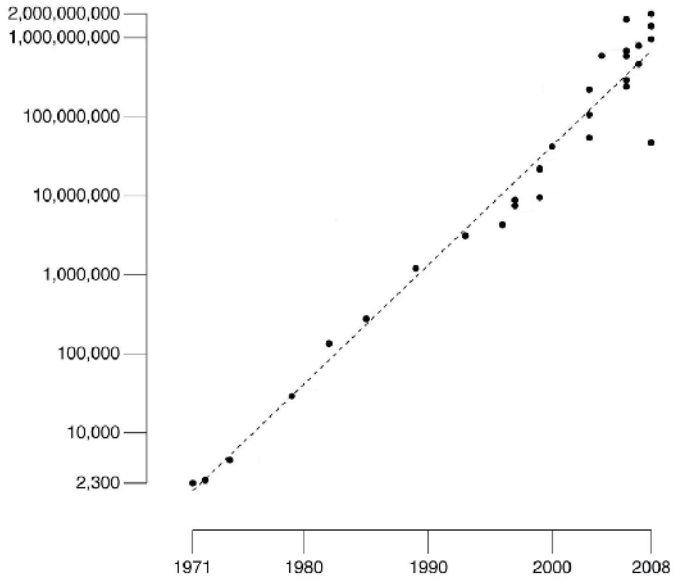


Fig. 1. One representation of the Moore-law. Horizontally the years, vertically the number of transistors in a CPU are represented. The points are for different CPU's between 1971 and 2008. The dashed line represents the Moore-law.

One of the interesting communication problems is how we can distribute a secret key for a secure communication between different parties. This is the so-called key distribution. The free-space Quantum Key Distribution (QKD) has a 16-year-old history. The first quantum cryptography protocol, the BB84 was introduced in 1984 and offered a solution for secure key distribution based on quantum theory principles like No Cloning Theorem.
The free-space quantum communications can be extended to ground-to-satellite or satellite-satellite quantum communications, which could be an ideal application for global quantum cryptography (Bacsardi, 2005).
One of the primary requirements of long-distance and free-space quantum communications is the capability of the effective transmission of quantum states in non-ideal, noisy environments. The free-space and satellite quantum channels are possible ways to increase significantly the distance limit of current quantum communication systems. To exploit the advantages of free-space quantum channels, it will be necessary to use space and satellite technology. The free space optical technology has been combined successfully with entangled pairs and satellite communications.
One of the main advantages of the usage of space for future quantum communication is the loss-free and distortion-free optical communication. In space, communication between

satellites can exploit the advantages of vacuum, where the noise of the channel can be negligible. Entanglement can be used in satellite communication to enhance the security level of key agreement process, and to realize a more secure communication compared to faint pulse quantum-key distribution technology.
This chapter is organized as follows. At first we introduce basics of quantum computing (Section 2) and quantum communication (Section 3). In Section 4, we discuss the possible connections between quantum and satellite communication including different approaches for quantum based information transfer in satellite communication, which can help to establish a secure communication link. Section 5 introduces our solutions with zero redundancy error correction which can help to establish an efficient communication link.

2. What is Quantum Computing?

2.1 Short Introduction to Quantum Informatics

From the viewpoint of quantum informatics the traditionally used communication methods are called classical methods. Communication algorithms based on classical methods are called classical algorithms. Quantum research started more than 25 years ago, and a lot of interesting results has been published since that. Although Deutsch has published the theoretical plan of a quantum computer, until now it hasn't been possible to build a real working quantum computer. Researches have had a lot of success in this area, and a lot of interesting physical implementation has been demonstrated. However, quantum informatics could not play a key role because quantum based algorithms are impossible to use without a working computer. These algorithms are very different from classical ones. Their properties have advantages in factoring, encrypting messages or creating unbreakable cryptography methods. Such solutions can be bought for commercial use from different quantum companies like id Quantique, MagiQ Technologies, Quintessence Labs.

The mathematical background of Quantum Informatics can be described by four postulates. In the first postulate the state space is defined. The second axiom describes the evolution of a closed system. The third postulate deals with measurements to create connection between quantum and classical world. In the fourth one composite systems are specified (Nielsen & Chuang, 2000).

1st postulate. The actual state of any closed physical system can be described by means of a so-called state vector **v** having complex coefficients and unit length in a Hilbert space V, i.e. a complex linear vector space equipped with an inner product.

2nd postulate. The evolution of any closed physical system in time can be characterized by means of unitary transforms depending only on the starting and finishing time of evolution.

3rd postulate. Let X be the set of possible results of the measurement. A quantum measurement can be described by means of a set of corresponding measurement operators .

$$M = \{\mathbf{M}_x\}, x \in X, \mathbf{M}_x \in \mathrm{H} \tag{1}$$

The operators should satisfy the completeness relation:

$$\sum_x \mathbf{M}_x^{\dagger} \mathbf{M}_x = \mathbf{I} \tag{2}$$

The probability of measuring x if the system is in state $|\varphi\rangle$ can be calculated as

$$p_x = \langle\varphi|\mathbf{M}_x^T\mathbf{M}_x|\varphi\rangle \tag{3}$$

The state of system after measurement is the following:

$$|\varphi\rangle = \frac{\mathbf{M}_x|\varphi\rangle}{\sqrt{p_x}} \tag{4}$$

4th postulate. The state space of a composite physical system W can be determined using the tensor product of the individual system V and Y:

$$W = V \otimes Y \tag{5}$$

2.2 Quantum bits

In classical information theory, the smallest unit is the bit. In digital computers, the voltage between the plates of a capacitor represents a bit of information: a charged capacitor denotes bit value 1 and an uncharged capacitor bit value 0. The smallest unit of the quantum informatics is the quantum bit (or qbit). One bit of information can be encoded using two different polarisations of light or two different electronic states of an atom. However, if we choose an atom as a physical bit, then apart from the two distinct electronic states the atom can be also prepared in a coherent superposition of the two states according to the rules of quantum mechanics. Therefore the atom is both in state 0 and state 1. Quantum computers use quantum states which can be in a superposition of many different numbers at the same time. In long distance communication photons are used as carriers of quantum bits. The channel can be a wired optical cable or the free-space. The problem is caused by No Cloning Theorem (NTC). According to NCT, copies can not be made of a non classical state, which means it is impossible to copy an electron spin based quantum bit to a photon based quantum bit without destroying the original quantum bit (Wootters & Zurek, 1982).

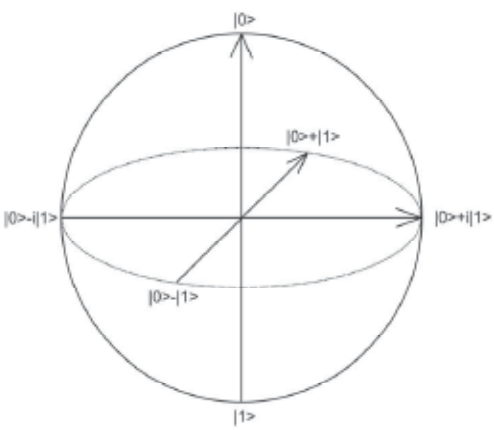


Fig. 2. Bloch sphere – a special visual representation of a quantum bit

A simple quantum system is a half-state of the two-level spin. Its basic states, spin-down $|\downarrow\rangle$ and spin-up $|\uparrow\rangle$, may be relabelled to represent binary zero and one, i.e. $|0\rangle$ and $|1\rangle$, respectively. The state of a single such particle is described by the wave function $|\psi\rangle = \lambda |0\rangle + \beta |1\rangle$. The squares of the complex coefficients – $|\lambda|^2$ and $|\beta|^2$ – represent the probabilities for finding the particle in the corresponding states. The representation of a two dimensional quantum bit can be seen in Fig.2.
For example, $|\psi\rangle = 0.6 |0\rangle + 0.8 |1\rangle$ means that we get 0 as result after the measurement with probability of 0.6, and we get 1 as result after the measurement with probability of 0.8.
Generalizing this to a set of k spin-1/2 particles we find that there are now 2^k basis states which equals to $2k$ possible bit-strings of length k (Nielsen & Chuang, 2000).

2.3 Quantum algorithms

ased on the postulates of quantum informatics quantum gates can be created, which perform a typical operation and/or transformation like identity, rotation, controlled NOT etc. A quantum gate can be described with its result or with its transformation matrix.
Some important gates are the following

$$I \equiv |0\rangle\langle 0| + |1\rangle\langle 1| \tag{6}$$

$$X \equiv |0\rangle\langle 1| + |1\rangle\langle 0| \tag{7}$$

$$\mathbf{Y} = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix} \tag{8}$$

$$\mathbf{Z} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \tag{9}$$

$$\mathbf{H} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \tag{10}$$

where I is the identity transformation, X is the bit flip, Z is the phase flip, Y exchanges the probability amplitudes multiplied by j, and H is the Hadamard transformation.

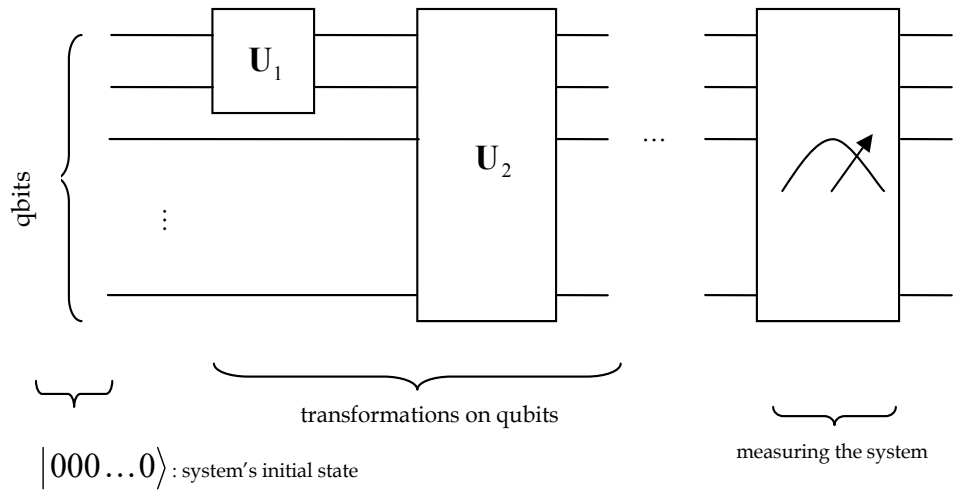


Fig. 3. General model of a quantum circuit

From the quantum gates, quantum circuits can be built. The quantum circuit is equivalent to the quantum algorithm, therefore quantum algorithms can be described by quantum circuits. The general model of a quantum circuit is illustrated in Fig.3. A quantum computer manipulates qbits by executing a series of quantum gates, each being unitary transformation acting on a single qbit or pair of qbits. At the end of the process, a classical result like 0 or 1 is determined by a measurement.
Let us create a quantum circuit like in Fig. 4. H is for Hadamard transformation, and the other gate is the Controlled-NOT gate (Imre & Ferenc, 2005). The final state of the system is

$$|\psi\rangle = \mathbf{CNOT} \cdot (\mathbf{H} \otimes \mathbf{I})(|0\rangle \otimes |0\rangle) = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle) \tag{11}$$

If we try to decompose the $|\psi\rangle = |00\rangle + |11\rangle$, our effort provides an interesting result because individual one-qbit states do not exist . States whose decomposition comprise one-qbit states are called product states while qbit and qregisters bounded together by a special phenomenon are referred to as entangled states. There are four distinguished entangled pairs called EPR pairs (named after Einstein, Podolsky and Rosen) or the Bell states.

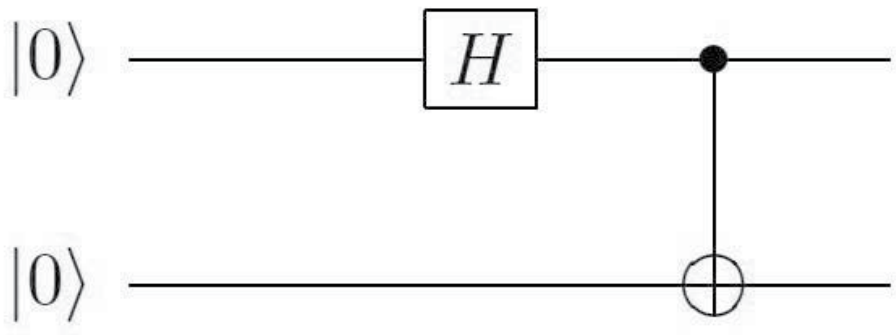


Fig. 4. EPR pairs generator quantum circuit

There are many interesting quantum algorithms, we would like to highlight some of them (Imre & Ferenc, 2005). In the quantum teleportation we use entangled pairs to transport information from point A to point B. The search in an unsorted database is faster with the Grover-algorithm than with the classical methods. We can break and decrypt the keys based on RSA algorithm with the help of Shor-algorithm. With the so-called superdense coding algorithm, two classical bits can be sent over a channel with only one quantum bit, using EPR pairs. This algorithm is illustrated in Fig.5.
From the viewpoint of the communication system, a secure key distribution is essential. In present classical cryptographic methods, the key exchange is generally based on public key methods. The security of modern cryptographic methods like asymmetric cryptography, relies heavily on the problem of factoring integers. In the future, if quantum computers become reality, any information exchange using current classical cryptographic schemes will be immediately insecure. Current classical cryptographic methods are not able to guarantee long-term security.

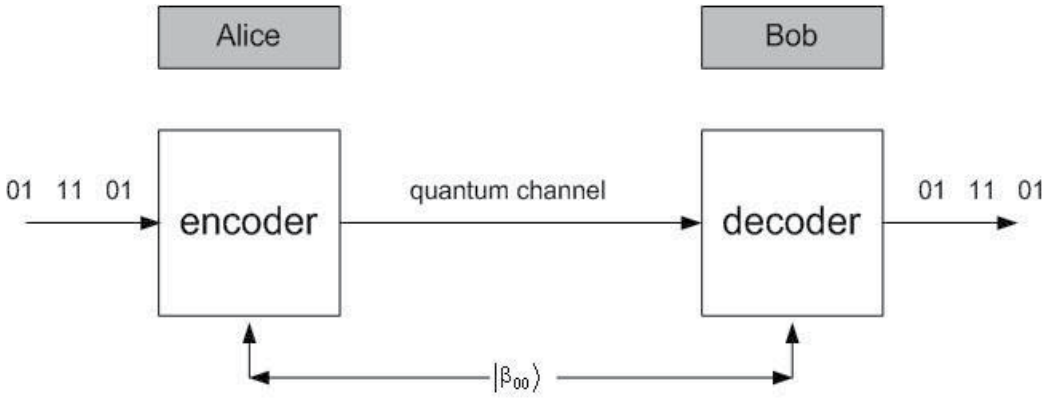


Fig. 5. In the superdense coding scenario we can send two classical bits over a channel with only one quantum bit.

3. Quantum Computing and Key Distribution

Quantum cryptography provides new ways to transmit information securely, using the fundamental principles of quantum-mechanics. As classical cryptography uses and manipulates classical bits, quantum cryptography does the same with qbits to realize provable absolute secure communication (Gyongyosi & Imre, 2009). In quantum cryptographic schemes, the secret information is not encoded directly into the quantum states, the qbits are used only to generate a secret cryptographic key, shared between two legal parties, called Alice and Bob. The main idea behind the quantum cryptographic protocols was the absolute secure key distribution. This is why we rather call these cryptographic methods as Quantum Key Distribution (QKD) systems.

These systems can provide unbreakable quantum communication. Therefore the QKD schemes may have a big relevance in future's space communication. There exist several fields, where private information must be sent through a free-space quantum communication channel. The QKD schemes solve the problem of key exchange, however some security steps are integrated after the process of quantum-based key exchange to provide provable safeness. Using quantum cryptography, the safety of future's space communication against various forms of possible attacks can be guaranteed.

The first QKD protocol was the BB84 protocol. The inventors of the BB84 protocol were C. H. Bennett and G. Brassard, and they published it in 1984 (Bennett & Brassard, 1984). C. H. Bennett in 1992 published the B92 protocol, the simplified version of the four state BB84 protocol (Bennet et al., 1992). The B92 protocol based on the similar principles of quantum mechanics like the BB84, however it uses only two polarization states.

The QKD schemes can solve the problem of key exchange and help to support sending private information through a free-space quantum channel. Using quantum cryptography, the safety of future's space communication against various forms of possible attacks can be guaranteed.

The QKD protocols are symmetric cryptographic systems. In these protocols, the action of encryption transforms the original message into an encoded message. The reverse operation of encryption is called decryption, which decodes the encoded message into the original

message. The decryption operation is based on a secret key, which is a symmetric key, hence the sender and the receiver have the same one to encode and decode the given message.
In the BB84 protocol Alice draws two random series of bits. One of these are the bits for the key, the second one is for the basis of measurement. If she has 0, she sends the $|0\rangle$ state, in other case she sends $|1\rangle$ state. If the base of the measurement is 1, than she performs a Hadamard transformation on the qubit. She sends the qubit to Bob. Bob draws a random series of bits. If he has 1 bit, he performs a Hadamard transformation on the received qbit. If he has a 0 bit, he performs an Identity transformation. Bob measures the quantum bit. Alice and Bob reconcile the base of their measurement on a public channel. They delete the quantum bits on which they used different measurement base. If the channel is noiseless and there is no eavesdropper in the channel, then they have the same series of qbits. To get ensured, they conciliate some selected bits.
If Eve wants to eavesdrop the channel, then she must measure Alice's qbits with a random selected measurement base, and send new qbits. However, if Alice and Bob conciliate their bits, they will notice the errors occurred by Eve's random selection. If they conciliate *N* bits, then they can find Eve's present Eve (?) with $1-\left(\frac{3}{4}\right)^N$ probability.

4. Quantum Satellite Communications

The quantum cryptography means distributing secret keys over a public communication channel, and not encrypting and decrypting of qbits like in the classical cryptography. The main task of the QKD protocols is to achieve an absolute secure key exchange process on the quantum channel.
Currently we use quantum algorithms in space communication mostly to secure key exchange. The values of quantum bits are encoded by photon polarizations. For the communication multilaser sources and optical receiver units are used. The current satellite and free-space schemes use weak laser pulses, instead of single photon communication. The polarization states of photons are related to the logical values of quantum states, and the particles of special photon-pairs can form entanglement (Lo & Chau, 1999).
The free-space QKD was first introduced over an optical path of about 30 cm in 1991. Since that several demonstrations (like indoor optical paths of 205 m and outdoor optical paths of 75 m) increased the utility of QKD by extending it to line-of-site laser communications systems during the 90's. In 1998, a research group at Los Alamos National Laboratory, New Mexico, USA developed a free-space QKD over outdoor optical paths of up to 950 m under night time conditions (Buttler et al., 1998). Four years later, in 2002 the same laboratory demonstrated that free-space QKD is possible by daylight as well as at night. The optical path had a length of appr. 10 km (Hughes et al., 2002). In 2006, an European research group reached the distance of 144 km (Schmitt-Manderbach et al., 2007). Space quantum communication systems have been implemented over a free-space distance of 23.4 km in Germany. The proposed satellite based quantum key distribution scheme used optical devices, improved spatial filtering and narrowband passfilter to control the transmitter lasers. In this implementation, the error rate was below 6%, and the polarization preparation devices and analysis modules were stable.

Establishing a successful free-space communication is not simple. The errors would arise from background photons collected at the satellite. The background rate depends on full or new moon: the error rate will be dominated by background photons during full moon periods, and by detector noise during new moon. During the daytime orbits the background radiance would be much larger

According to our models the quantum computing algorithms can be used to affirm our free-space communication in four different ways, which are the open-air communications, earth-satellite communications, satellite broadcast and inter-satellite communications (Bacsardi, 2007).

1. *Open-air communications:* a 'horizontal' communication channel is used. The communication happens below 100km height.
2. *Earth-satellite communications:* the communications take place through greater heights than in the Open-air communication, usually between 300 and 800 km altitude. Signal encoding and decoding can be used to produce quantum error correction that allows operation in noisy environment.
3. *Satellite broadcast:* the broadcast satellites are in orbit at 36,000 km using 27 MHz frequency for signalling. The quantum algorithms can play an important role to improve the effective bandwidth, thus the brand is better utilized as in traditional cases.
4. *Inter-satellite communication:* Equals the satellite-to-satellite communications. In this case the channel is the free-space. Any kind of coding and encoding can be used, to increase stability.

The future's free-space quantum communication should be able to realize secret key distribution over long distances using low, median and geostationary satellite system. The QKD can be extended to achieve global quantum data protection. The future free-space satellite QKD schemes can be applied in two schemes [24, 25]:

1. Global quantum key distribution based on satellite-to-ground faint pulse quantum communication,
2. Simultaneous key generation between ground stations using EPR-based QKD scheme.

In Fig. 6 we illustrated the schematic of a global key exchange QKD system in free-space. The global key free-space QKD scheme uses symmetric key encoding at the space satellite modules, and the decoding process is implemented at ground level. According to the security requirements of free-space communication, the global key free-space method requires the generation of new keys regularly. In this scheme, QKD is applied during the key upload to satellites, which (?) key will be used by the satellite module to data scrambling. The satellites scramble the transmitted data to the licensed users only, and in this process, a QKD based key can be applied efficiently to ensure the advantages of quantum mechanics.

The global key exchange based QKD scheme uses faint pulse quantum cryptography. The bitrate of the key agreement process is about 1000 bit/sec. The tested implementations have worked between 600 km and 2000 km distances. The global key exchanging satellite QKD systems can be implemented with more than one ground receiving stations for a single satellite module. The exchanged keys can be used by ground stations on long distances (Koashi et al., 2008).

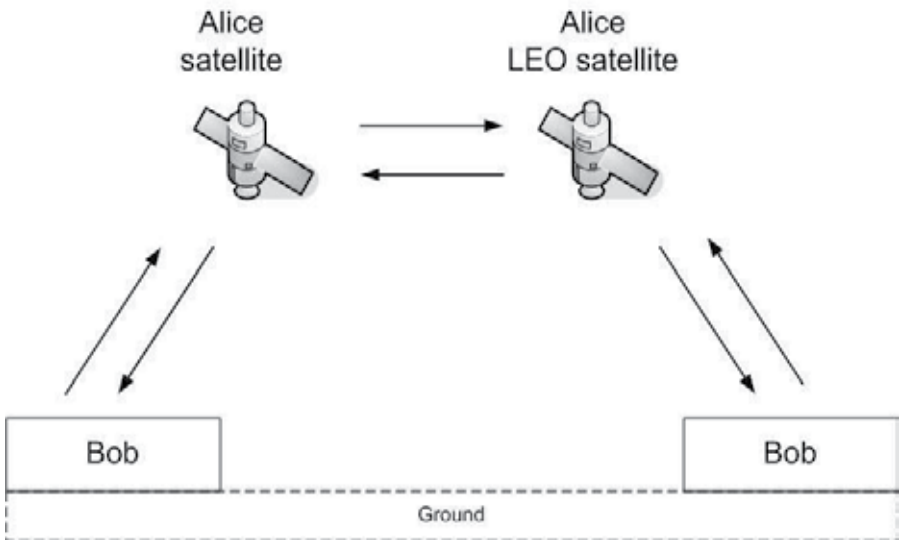


Fig. 6. Global key exchange based free-space QKD

We illustrated a satellite-to-ground free-space QKD scheme in. Fig 7. This method uses entangled photons. The bitrate of the key agreement process in the current implementations is greater than 100 bit/s.

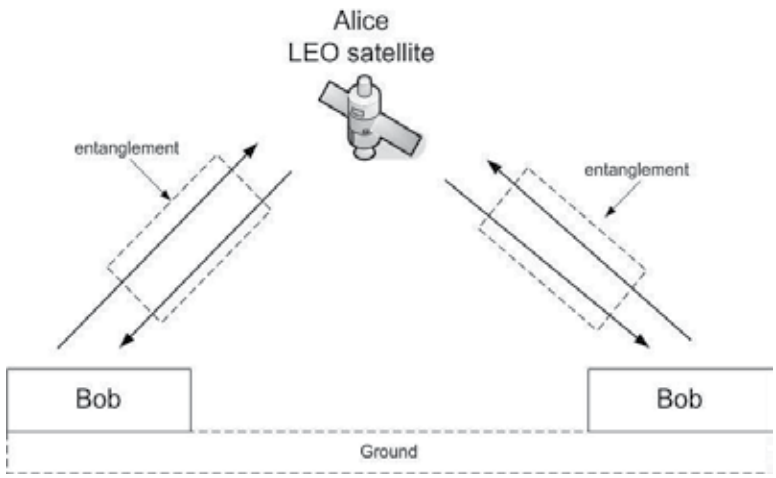


Fig. 7. Satellite-to-ground free-space QKD

The secret key agreement process requires two optical paths from the space module to the ground communication module, which increments the error rate in the current solutions. Another drawback of polarization based satellite-to-ground QKDs, that Alice and Bob rotate to each other according to the movement of the Earth, thus it is necessary to continuously compensate this movement in polarization detection (Schmitt-Manderbach et al., 2007). This kind of cryptography uses EPR-based QKD scheme. There are still many challenges in practical implementation to use entanglement. However, the free-space EPR technology can

be implemented easily in current free-space QKD realizations (Ho et al., 2008; Ling et al., 2008; Tsurumaru & Tamaki, 2008).

5. Redundancy-free Quantum Channels

5.1. Importance of Redundancy-free Quantum Channels

In the classical communication we need a channel coding to handle the errors appearing in a communication channel. The error correction can be performed only with the help of redundancy in the classical system. The error correction capabilities are required for any large scale computation and communication. In these classical systems the simplest form to give redundancy to the communication is to encode the bits more than once. However, the quantum communication can be performed by more complex strategies.

In quantum computing the classical error coding methods could not be used because of the following three reasons (Nielsen & Chuang, 2000):

1. The errors are continuous. The errors can result either amplitude or phase decoherence. Moreover both errors have complex coefficients. This means that their co-domains are continuous.
2. Through the No Cloning Theorem a simple copy-based redundancy is not possible.
3. Problems occur at the measurement of the transmitted states. For the error correction the type of error has to be known. But if the quantum bits are measured for determination of the failure, then the original bits are lost.

Many quantum error correction methods have been introduced to overcome these problems and to try to handle the limitations of quantum theory principles. Since the quantum states cannot be cloned perfectly, or cannot be measured nondestructively, a simple copy-based solution is not possible. In these proposals some redundancy is required for successful error correction. However, best would to implement a redundancy-free proposal.

The redundancy-free solutions could be very useful in the long-distance free-space communication, because there would be no need to use redundant error correction codes as nowadays. With the redundancy-free techniques the effective capacity of the satellite link could also be increased

In our solution we would like to provide error correction by sending certain amount of qbits over a noisy quantum channel. The qbits are independent. Each contains information that needs to be processed. We show two different redundancy-free solutions. In the first one the noise of the quantum channel is modeled by a rotation angle. In the second one we consider the redundancy-free implementation of a unitary error correcting operator $\mathcal{R}_\theta$.

5.2. Redundancy-free Quantum Channel – Solution 1.

Our initial assumption is that the channel rotates the qbit with a ω degree which is considered to be constant. We wish to create a system where error correction is possible. To achieve this we mix the qbits and send them over the channel, as shown in Fig. 8.

We use n long qbits for the communication so $2^n = N$, where n is the length of the qbits and N is the size of the space. If we code the classical states into the eigenvectors of the U matrix, we can construct a description which leads to a redundancy-free solution With the appropriate selection of the matrix A, we can restore one quantum bit sent over the channel without any other (redundant) information.

Because U is unitary, it can be written in the following form

$$U = \sum_i \lambda_i \left| u_i \right\rangle, \P \tag{12}$$

where λ_i, $\left| u_i \right\rangle$ are the eigenvalues and the eigenvectors of matrix U and

$$\lambda_n = e^{j\alpha_n}. \P \tag{13}$$

Because U is unitary, it acts on each qbits and changes it as

$$\left| \psi \right\rangle = U \left| \psi_k \right\rangle. \P \tag{14}$$

Using the eigenvalues we get the following matrix for U

$$U_k = \begin{bmatrix} e^{j\alpha_{k1}} & 0 \\ 0 & e^{j\alpha_{k2}} \end{bmatrix}. \P \tag{15}$$

For the eigenvalues, we have two cases. In the first case we can describe U as

$$U = \begin{bmatrix} e^{j+\alpha} & 0 \\ 0 & e^{j+\alpha} \end{bmatrix} \otimes \begin{bmatrix} e^{j+\alpha} & 0 \\ 0 & e^{j+\alpha} \end{bmatrix} = \begin{bmatrix} e^{j2\alpha} & 0 \\ 0 & e^{j2\alpha} \end{bmatrix}, \P \tag{16}$$

In the second case we can describe U as

$$U = \begin{bmatrix} e^{j+\alpha} & 0 \\ 0 & e^{j+\alpha} \end{bmatrix} \otimes \begin{bmatrix} e^{j-\alpha} & 0 \\ 0 & e^{j-\alpha} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I. \P \tag{17}$$

This description leads to a redundancy-free solution because the classical states are coded into the eigenvectors of the U matrix. The eigenvalues can be written in the form shown in (13) in case of a unitary transformation. The whole mathematical background of the algorithm is described in (Bacsardi et al., 2009).

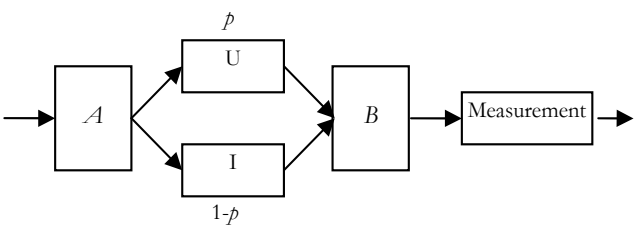


Fig. 8. Our channel model for redundancy free-error correction.

5.3. Redundancy-free Quantum Channel – Solution 2.

In this solution we consider the redundancy-free implementation of a unitary error correcting operator $\mathcal{R}_\theta$. The solution achieves the redundancy-free quantum communication using local unitary operations and unitary matrices. The error correcting operator is illustrated on Fig. 9.

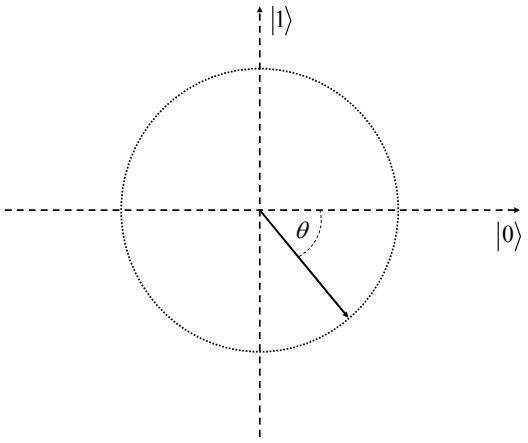


Fig. 9. Redundancy-free error correction, circle

The error of the satellite quantum channel can be modeled by a unitary rotation $\mathcal{R}_\theta^\dagger$, thus the error of the satellite quantum channel can be expressed as an angle: $\theta_i \in [0, 2\pi)$. At the beginning of the communication, Alice sends her quantum state $|\psi_A\rangle$ on the quantum channel, which transforms it to $|d\rangle = \mathcal{R}_\theta^\dagger(\psi_A)$ with given probability *p*. The error of the quantum channel is denoted by $\mathcal{R}_\theta^\dagger$. In the error-correcting process while reading the sent quantum state, Bob doesn't know the properties of the noise on the quantum channel. In the redundancy-free coding mechanism, Alice's initial state is $|\psi_A\rangle$, the correction transformation denoted by $\mathcal{R}_\theta$. Bob uses a CNOT to correct the error of the quantum channel. In order to read the sent quantum bits correctly, Bob must rotate the *i*-th data quantum bit by the angle θ_i in the opposite direction of that rotated by the error of the quantum channel. The method is illustrated on Fig. 10.

Bob has a chance not greater than $\varepsilon = \sin^2(\theta_i)$ to correct the sent states, because he doesn't know the original rotation angle θ_i of the quantum channel's error on the *i*-th sent qbit. The rotation operation $\mathcal{R}_\theta$ of the error correcting mechanism can be given by the angle $|\theta\rangle$, where

$$|\theta\rangle = \frac{1}{\sqrt{2}}\left(e^{i\frac{\theta}{2}}|0\rangle + e^{-i\frac{\theta}{2}}|1\rangle\right) \tag{18}$$

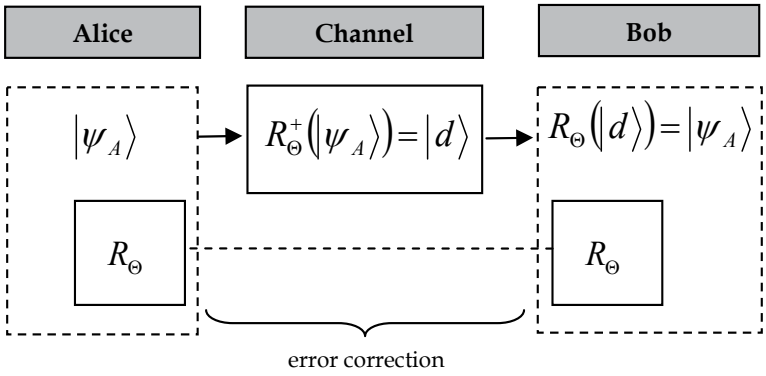


Fig. 10. Redundancy-free error correction, solution 2.

The error-correcting method consists of a control qbit, which corresponds to the modified qbit $|d\rangle$, and a target qbit, which is equal to the error-correction angle state $|\theta\rangle$. To correct state $|d\rangle$ to ket(ψ_A), Bob uses a simple CNOT transformation, thus our state is transformed to

$$|d\rangle\otimes|\theta\rangle\rightarrow\frac{1}{\sqrt{2}}\left(\mathcal{R}_\theta|d\rangle\otimes|0\rangle+\mathcal{R}_\theta^\dagger|d\rangle\otimes|1\rangle\right), \tag{19}$$

and therefore a projective measurement in the $\{|0\rangle,|1\rangle\}$ base of the correction-state $|\theta\rangle$ will make the modified qbit $|d\rangle$ collapse either into the desired state $\mathcal{R}_\theta|d\rangle$ or into the wrong state $\mathcal{R}_\theta^\dagger|d\rangle$. Bob cannot determine the received state exactly, since he does not know the angle of the error θ_i. In this phase, Bob can not be sure whether the *i*-th quantum state $|d_i\rangle$ is identical to the original sent state $|\psi_i\rangle$ or not. The whole algorithm is described in (Bacsardi et al., 2009).

6. In the near future

Although the free-space QKD seems to be a tool for the future, the QKD implementations include some challenges, like the extremely high precision of optical devices, the nano-second time interval synchronization of key exchange process, and the generation of entangled photons for long distances (Ho et al., 2009; Ling et al., 2008). The future free-space QKD applications can include direct communication between satellites, where the secret quantum key agreement process will be achieved between satellites in the space. The current low, medium and geostationary earth orbit satellites between the range of 1000 km and 35,000 km can be used to achieve secure future space quantum communication (Schmitt-Manderbach et al., 2007). The fiber based results are promising as well. The Japanese Researches at National Institute of Information and Communications Technology

(NICT) reported a QKD field test through a 97 km installed fiber. Their experiments established the clock transmission technique for long distance QKD without degrading the quantum signal, using an optical amplifier for the clock signal (Tanaka et al., 2008). The first experimental implementation of a scheme for single-photon exchange between an Earth-based satellite was reported by Italian researches. They have built such an experiment which has a single photon source on a satellite, exploiting the telescope at the Matera Laser Ranging Observatory to detect the transmitted photons. Returning signals were detected from a low-earth orbit geodetic satellite (Ajisai), whose orbit has a perigee height of 1485 km.

Space-QUEST (Quantum Entanglement in Space Experiments) is a promising project. This project is led by a European research consortium and supported by the European Life and Physical Sciences in Space Program of ESA. The aim is to make a quantum communication space-to-Earth experiment from the International Space Station According to their mission scope, bringing quantum entanglement for the first time in the space environment will open a new range of fundamental physical experiments (Ursin et al., 2009).

7. Conclusions

In the next years we have to handle the problems originated from the decreasing size of microchips. As everyday life is getting faster it requires researchers and engineers to respond to the growing needs. In telecommunication applications one of the most crucial issue to solve is how to transmit as many bytes at a given frequency in a second as possible. One way to determine the system's efficiency is to measure the number of bytes sent per second.

In this chapter we presented the base of quantum mechanic based communication. We showed that there exist already algorithms to solve problems which are very difficult to be solved by traditional computers. The quantum theory has appeared in satellite communications offering answers for some of nowadays' technical questions. We introduced two solutions which can be useful in redundancy-free communications. One of the primary requirements of long-distance and free-space quantum communication is the capability of the effective transmission of quantum states in non-ideal, noisy environments. The free-space and satellite quantum channel could be the way to increase significantly the distance limit of current quantum communication systems. The current earthbound free-space quantum channels have the advantage in that they can be combined with satellite quantum communication. In the future, we will be able to overcome the current distance limits in quantum communication by transmitting EPR-states from space to Earth.

8. References

Bacsardi, L. (2005). Using Quantum Computing Algorithms in Future Satellite Communication. *Acta Astronautica*, Vol 57., No. (2005) 224-229, ISSN 0094-5765

Bacsardi, L. (2007). Satellite communication over quantum channel. *Acta Astronautica*. Vol 61. No. 1-6. (June-August 2007), 151-159, ISSN 0094-5765

Bacsardi, L.; Gyongyosi, L. & Imre, S. (2009). Solutions for Redundancy-free Error Correction in Quantum Channel, *Proceeding of the QuantumComm 2009*, Vico Equenso, Italy, October 2009, ICST, Gent

Bennett, C. H. & Brassard, G. (1984). Quantum cryptography: Public key distribution and coin tossing. *Proceedings of the IEEE International Conference on Computers, Systems and Signal Processing*, Bangalore, India, December 1984, IEEE, New York

Bennett, C. H.; Brassard, G. & Ekert, A.K. (1992). Quantum cryptography. *Scientific American*. Vol. 267(4) Issue. 50 (October 1992), ISSN 0036-8733

Buttler, W. T.; Hughes, R. J.; Kwiat, P. G. ; Lamoreaux, S. K.; Luther, G. G.; Morgan, G. L.; Nordholt, J. E.;Peterson, C. G. & Simmons, C. M. (1998) Practical free-space quantum key distribution over 1 km. *Phys. Rev. Lett.* Vol 81., October 1998, ISSN 1079-711

Gyongyosi, L & Imre, S. Fidelity Analysis of Quantum Cloning Based Attacks in Quantum Cryptography, *Proceedings of the 10th International Conference on Telecommunications*, Zagreb, Croatia, 2009

Ho, C; Lamas-Linares, A & Kurtsiefer, C (2009). Clock synchronization by remote detection of correlated photon pairs. *New Journal of Physics* Vol. 11, April 2009, ISSN 1367-2630 (2009)

Hughes, R. J.; Nordholt, J. E; Derkacs, D. & Peterson, C. G (2002). Practical free-space quantum key distribution over 10 km in daylight and at night. *New Journal of Physics*. Vol. 4., Issue 1, Jul 2002, ISSN 1367-2630

Imre, S. & Ferenc, B. (2005) *Quantum Computing and Communications: An Engineering Approach.* Wiley, ISBN 9780470869024

Ling, A. ; Peloso, M.P.; Marcikic, I.; Scarani, V. ; Lamas-Linares, A. & Kurtsiefer:, C. (2008). Experimental quantum key distribution based on a Bell test. *Physical Review A*. Vol.78 , Issue 2, 2008, ISSN 1050-2947 (2008)

Lo, H.-K. & Chau, H.F. (1999). Unconditional security of quantum key distribution over arbitrarily long distances. *Science* Vol. 283 No. 2050 (March 1999), ISSN 0036-8075

Koashi, M. ; Adachi, Y. ; Yamamoto, T. & Imoto, N. (2008). Security of Entanglement-Based Quantum Key Distribution with Practical Detectors. http://arxiv.org/abs/0804.0891, 2008.

Nielsen, M.A. & Chuang, I.L. (2000). Quantum Computation and Quantum Information. Cambridge University Press, ISBN 9780521635035, Cambridge

Schmitt-Manderbach, T., Weier, H; Fürst, M; Ursin,R; Tiefenbacher, F; Scheidl, T; Perdigues, J.; Sodnik, Z.; Kurtsiefer, C.; Rarity, J. G., Zeilinger, A.; & Weinfurter, H (2007). Experimental Demonstration of Free-Space Decoy-State Quantum Key Distribution over 144 km, *Physical Review Letters,* Vol. 98, Issue 1, 2007, ISSN 1079-7114

Tanaka, A; Fujiwara, M.; Nam, S. W.; Nambu, Y.; Takahashi, S. ; Maeda, W. ; Yoshino, K.; Miki, S.; Baek, B.; Wang, Z.; Tajima, A.; Sasaki, M. & Tomita, A. (2008) Ultra fast quantum key distribution over a 97 km installed telecom fiber with wavelength division multiplexing clock synchronization. *Optics Express* Vol. 16, Issue 15, July 2008, ISSN 1094-4087

Tsurumaru, T. & Tamaki, K. (2008). Security Proof for QKD Systems with Threshold Detectors. http://arxiv.org/abs/0803.4226, 2008.

Ursin, R. et al. (2009), Space-QUEST. Experiments with quantum entanglement in space. *Europhysics News* Vol. 40, Issue 3, 2009, ISSN 0531-7479

Wootters, W. K. & Zurek, W. H. (1982) A single quantum cannot be cloned, *Nature* Vol. 299 No. 802–803, 1982, ISSN 0028-0836

20

Satellite coverage optimization problems with shaped reflector antennas

Adriano C. Lisboa and Douglas A. G. Vieira
ENACOM - Handcrafted Technologies
Brazil

Rodney R. Saldanha
Universidade Federal de Minas Gerais
Brazil

This chapter is devoted to optimization formulations of satellite coverage problems and their solution. They are specially useful in satellite broadcasting applications, where the information goes from one to many. Due to distance from target and power supply constraints, highly directive efficient reflector antennas are the best choice. There are many possible optimization formulations for this problem, depending fundamentally on specifications and degrees of freedom in the system. Some specific coverage problems considering broadband, radiation pattern constraints and interference are presented, where the degrees of freedom are purely geometric.

1. Satellite coverage problems

A satellite coverage problem queries for antennas in geostationary orbit of a planet, as shown in the example in Fig. 1 for the Brazilian territory, whose radiation pattern over a target is as close as possible to a specification.

Fig. 1. Graphical illustration of a satellite coverage problem for the Brazilian territory.

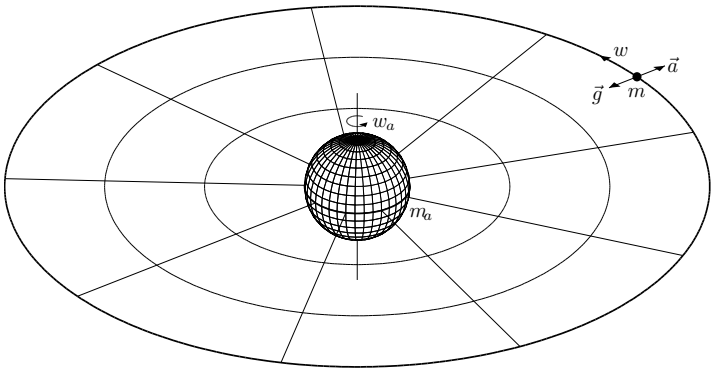


Fig. 2. Geostationary orbit (true scale) around a planet of mass m_a and angular speed w_a.

1.1 Satellite's orbit

Geosynchronous orbit is a requirement because it is considered a fixed target on the planet surface. Geostationary orbit (a special case of geosynchronous orbit) is a further requirement to reduce orbital station-keeping of the satellite. Geostationary means that the satellite centripetal force $m\vec{a}$ and the planet gravitational force $m\vec{g}$ are equal in modulus and opposite in direction (see Fig. 2), i.e. $m\vec{a} = -m\vec{g}$, where m is the satellite mass, $\vec{a}$ is the satellite centripetal acceleration and $\vec{g}$ is the planet gravitational acceleration. The direction constraint implies that the satellite is above the planet's equator. Using the notation $\vec{a} = a\hat{a}$ to denote the modulus $a = |\vec{a}|$ and unit direction $\hat{a} = \vec{a}/|\vec{a}|$ of a vector $\vec{a}$, the modulus constraint implies $ma = mg$, where

$$a = w^2 r \tag{1}$$

$$g = \frac{Gm_a}{r^2} \tag{2}$$

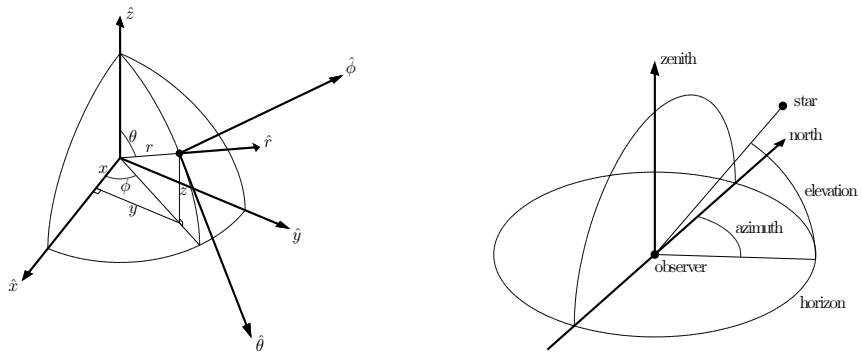
where w is the satellite angular speed, G is the gravitational constant ($G \approx 6.674 \times 10^{-11}\mathrm{m}^3/\mathrm{kg}/\mathrm{s}^2$), m_a is the planet mass ($m_a \approx 5.974 \times 10^{24}$kg for Earth), and r is the distance from the planet center to the satellite, so that the geostationary orbit radius is given by

$$r_s = \sqrt[3]{\frac{Gm_a}{w_a^2}} \tag{3}$$

where w_a is the angular speed of the planet's rotation around its own axis ($w_a \approx 7.292 \times 10^{-5}$rad/s for Earth). The geostationary orbit radius depends only on planet properties. For Earth it is about 42.2×10^6m.

1.2 Coordinate systems

A very important, yet simple, point in the mathematical model of the satellite is the definition of coordinate systems. In the context of satellite coverage problems, they can be defined by taking the planet, the antenna reflector or the antenna feed as reference. For now, consider planet coordinate system as any coordinate system that takes the planet center as its origin and the planet rotation axis as its z-axis.



(a) (b)

Fig. 3. Spherical coordinate system (a) and hoizontal coordinate system (b).

Denote $\vec{p} = x_p\hat{x} + y_p\hat{y} + z_p\hat{z}$ a point given in Cartesian coordinate system, and $\vec{p} = r_p\hat{r} + \theta_p\hat{\theta} + \phi_p\hat{\phi}$ a point given in spherical coordinate system (see Fig. 3(a)), so that

$$\begin{aligned} x_p &= r_p \sin\theta_p \cos\phi_p \\ y_p &= r_p \sin\theta_p \sin\phi_p \\ z_p &= r_p \cos\theta_p \end{aligned} \tag{4}$$

Denote $\vec{a} = a_x\hat{x} + a_y\hat{y} + a_z\hat{z}$ a vector given in Cartesian coordinate system, and $\vec{a} = a_r\hat{r} + a_\theta\hat{\theta} + a_\phi\hat{\phi}$ a vector given in spherical coordinate system, so that

$$\begin{bmatrix} a_x \\ a_y \\ a_z \end{bmatrix} = \begin{bmatrix} \sin\theta\cos\phi & \cos\theta\cos\phi & -\sin\phi \\ \sin\theta\sin\phi & \cos\theta\sin\phi & \cos\phi \\ \cos\theta & -\sin\theta & 0 \end{bmatrix} \begin{bmatrix} a_r \\ a_\theta \\ a_\phi \end{bmatrix} \tag{5}$$

and, since the transformation matrix U is unitary (i.e. $U^{-1} = U^T$),

$$\begin{bmatrix} a_r \\ a_\theta \\ a_\phi \end{bmatrix} = \begin{bmatrix} \sin\theta\cos\phi & \sin\theta\sin\phi & \cos\theta \\ \cos\theta\cos\phi & \cos\theta\sin\phi & -\sin\theta \\ -\sin\phi & \cos\phi & 0 \end{bmatrix} \begin{bmatrix} a_x \\ a_y \\ a_z \end{bmatrix} \tag{6}$$

where it is important to note that the vector components depends on the angular components θ and ϕ of the point where the vector itself is defined. Coherently to this, a vector field (e.g. electric field $\vec{E}$ and current density $\vec{J}$) is a function of the position where each vector is defined (e.g. $\vec{E}(\vec{r})$ and $\vec{J}(\vec{r}')$, where $\vec{r}$ is an observer point and $\vec{r}'$ is a source point). Also because the transformation matrix is unitary, the modulus of a vector does not depend on which coordinate system it is evaluated: just take the square root of the sum of its squared components.

The horizontal coordinate system, as shown in Fig. 3(b), is slightly different. It is a kind of celestial coordinate system, where positions are mapped on a sphere centered in the observer, originally taking the north pole of Earth as origin and the zenith as reference. It is defined by two components: azimuth (horizontal angular distance to the north) and elevation (vertical angular distance to the ground). It is very convenient to map deviations around a direction and it will be used to map radiation patterns where the observer is the satellite.

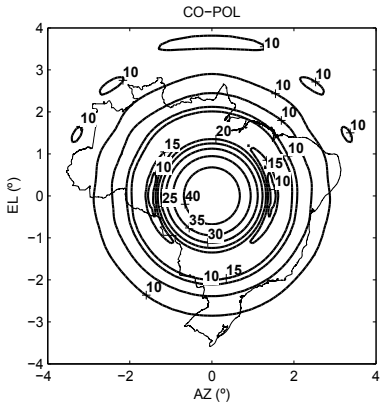


Fig. 4. Radiation pattern of a parabolic antenna illuminating the Brazilian territory. It is shown the level sets of the co-polarization gain (dBi) relative to an isotropic radiation in each direction given in reflector horizontal coordinate system.

1.3 Degrees of freedom in satellite's position and direction

Considering a geostationary orbit of radius r_s relative to the planet center, the remaining degree of freedom in the satellite position given in planet spherical coordinate system

$$\vec{p} = r_s\hat{r} + \frac{\pi}{2}\hat{\theta} + \phi_p\hat{\phi} \tag{7}$$

is its angular position ϕ_p (it defines a location on the geostationary circle). The satellite is directed to a point on the planet surface

$$\vec{t} = r_a\hat{r} + \theta_t\hat{\theta} + \phi_t\hat{\phi} \tag{8}$$

where r_a is the planet radius ($r_a \approx 6.4 \times 10^6$m for Earth). The center of mass $\vec{c}$ of the target is a good reference, as shown in Fig. 1, so that $\theta_t = \theta_c$ and $\phi_t = \phi_p = \phi_c$. However, the degrees of freedom ϕ_p, θ_t and ϕ_t may be let as optimization variables for the satellite coverage problem.

1.4 Radiation pattern specification

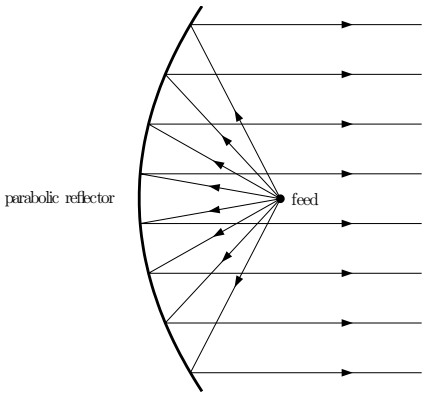
A typical radiation pattern of a classical reflector antenna is shown in Fig. 4, where the target is the Brazilian territory. This highly directive pattern is desired for one-to-one communications (i.e. when target is a point), but can be improved when one-to-many communications (i.e. when target is an area) are considered. The most fair radiation pattern for broadcasting is the one that uniformly illuminates the target area, and only that. However different types of coverage can also be specified by gain lower or upper bounds, energy confinement inside target, radiation polarization or broadband operation.

2. Electromagnetic model

2.1 Antenna type selection

Since energy must be collimated within a few degrees around its main radiation direction, reflector antennas are very suitable and the natural choice.

A relector antenna is composed by feed and reflector, as shown in Fig. 5. The main function of the feed is to supply energy. The main function of the reflector is to collimate energy.



(a)

Fig. 5. Parabolic reflector antenna (a).

2.1.1 Parabolic reflector antenna

A point $\vec{b} = (z, \rho)$ belongs to a parabola, as shown in Fig. 6, when it is at the same distance from the focus $\vec{a} = (F, 0)$ and its projection $\vec{c} = (-F, \rho)$ onto the directrix $z = -F$, which leads to the implicit equation

$$z = \frac{\rho^2}{4F} = \frac{x^2 + y^2}{4F} \tag{9}$$

where F is called the focal distance.
The unit normal vector to the parabola can be written as

$$\hat{n} = \frac{2F\hat{z} - \rho\hat{\rho}}{\sqrt{4F^2 + \rho^2}} = \frac{2F\hat{z} - x\hat{x} - y\hat{y}}{\sqrt{4F^2 + x^2 + y^2}} \tag{10}$$

The angle of the normal vector $\hat{n}$ to $\hat{z}$ is

$$\theta = \arctan\left(\frac{|\rho|}{2F}\right) = \arctan\left(\frac{\sqrt{x^2 + y^2}}{2F}\right) \tag{11}$$

The direction of a ray from the focus to a point on the parabola can be written as

$$\hat{d} = \frac{(z - F)\hat{z} + \rho\hat{\rho}}{z + F} \tag{12}$$

so that

$$\begin{aligned} -\hat{d} \cdot \hat{n} &= \frac{-2zF + 2F^2 + \rho^2}{(z + F)\sqrt{4F^2 + \rho^2}} = \frac{-2zF + 2F^2 + 4zF}{(z + F)\sqrt{4F^2 + \rho^2}} \\ &= \frac{2F}{\sqrt{4F^2 + \rho^2}} \\ &= \hat{n} \cdot \hat{z} \\ &= \cos\theta \end{aligned} \tag{13}$$

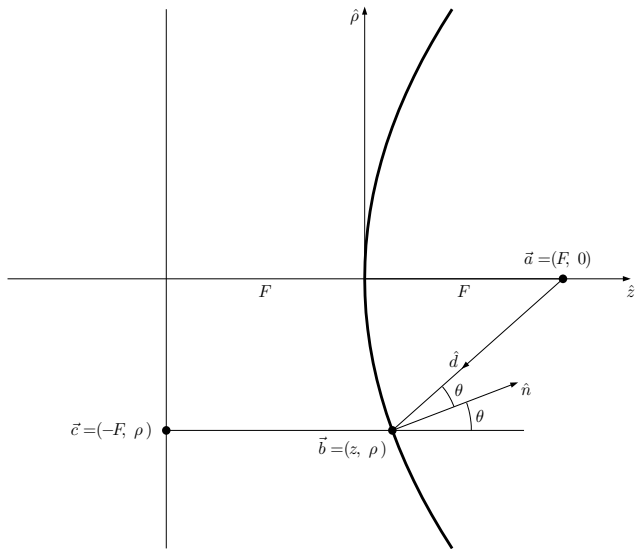


Fig. 6. Parabola with focus $\vec{a} = (F, 0)$ and directrix $z = -F$, defined by points $\vec{b} = (z, \rho)$.

Hence, the angle between $-\hat{d}$ and $\hat{n}$ is the same as the angle between $\hat{n}$ and $\hat{z}$. If the reflection is perfect (e.g. for electromagnetic waves reflector must be a perfect electric conductor), then θ is the angle of reflection of the ray, which equals the angle of incidence, so that all reflected rays go towards $\hat{z}$.
A last notable feature of parabolic reflectors is that the distance traveled by two reflected rays originated from the focus up to a given $z = z_a$ plane are exactly the same and valued to

$$\begin{aligned} \left|\vec{b} - \vec{a}\right| + z_a - z &= \left|\vec{b} - \vec{c}\right| + z_a - z \\ &= F + z + z_a - z \\ &= F + z_a \end{aligned} \tag{14}$$

Therefore, if all rays from the feed are in phase, then they will still be in phase after reflecting and reaching $z = z_a$. This is highly desired, considering that an aperture (i.e. a bounded region at a given $z = z_a$) is most directive when the field distribution on it is uniform in phase and amplitude. Unfortunately, the same does not hold true for the amplitude, since the dispersive behavior from the feed up to the reflector attenuates non-uniformly each ray according to the respective distance $F + z$.
Fig. 7 shows a parabolic reflector antenna with circular aperture of radius R. The reflector center was displaced H towards $\hat{y}$ together with a feed tilt of τ in the $x = 0$ plane, so that the feed do not block reflected waves.

2.2 Geometric parameterization of the antenna

In optimization, a parameterization that considers the expected configurations to meet specified requirements reduce the search space and, hence, the computational effort to solve the optimization problem.
For a single reflector antenna, the expected shape is nearly parabolic, convex, continuous and smooth. A suitable parameterization would be a set of basis functions capable to represent

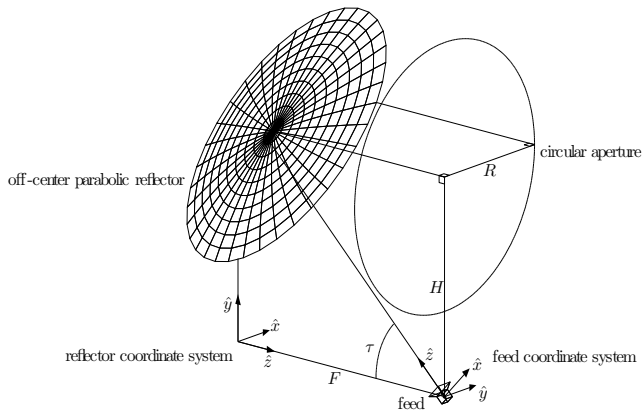


Fig. 7. Off-center parabolic reflector antenna.

any convex function, just like Fourier series are for periodic functions. Unfortunately, this is still an open problem. A parameterization that takes into consideration the expected shapes for a single reflector antenna is based on four triangular Bézier patches (Lisboa et al., 2006) to represent the reflector shape. Bézier patches are very common in computer aided geometric design. The shape of the object they represent is controlled by the position of vertices of a grid, called control points.
The feed position and direction can represent up to 5 geometric degrees of freedom for the antenna.

2.3 Electromagnetic analysis

In optimization, the antenna must be simulated (analysed) so that specified measures can be evaluated. These measures are evaluated at the so called observer points, denoted by $\vec{r}$. In spherical coordinates, observer points are written as $\vec{r} = r\hat{r} + \theta\hat{\theta} + \phi\hat{\phi}$. Analogously, source points are denoted by $\vec{r}'$.
Some key points concerning the electromagnetic model of the reflector antenna are depicted next.

2.3.1 Harmonic fields

A very useful simplification in field propagation formulation is to consider harmonic components of it. This assumption allows to express the instantaneous field as the product of two terms: one depending only on space (i.e. amplitude, phase and direction at some spacial position) and another one depending only on time. The instantaneous field relative to a single harmonic component can be written as

$$\begin{aligned}\vec{E}(\vec{r},t) &= \mathrm{Re}\left(\vec{E}(\vec{r})\,\mathrm{e}^{jwt}\right)\\ &= \mathrm{Re}(\vec{E}(\vec{r}))\cos(wt) - \mathrm{Im}(\vec{E}(\vec{r}))\sin(wt)\\ &= \frac{1}{2}\left(\vec{E}(\vec{r})\,\mathrm{e}^{jwt} + \vec{E}^{*}(\vec{r})\,\mathrm{e}^{-jwt}\right)\end{aligned} \tag{15}$$

where t is the time, $\vec{E}(\vec{r})$ is an harmonic field at angular frequency $w = 2\pi\nu$, ν is the frequency, and j is the imaginary unit. Since any time independent phase angle or attenuation could be

put into $\vec{E}(\vec{r})$ (i.e. $A\,e^{jwt+\alpha+j\beta} = A\,e^{\alpha}\,e^{j\beta}\,e^{jwt} = A'\,e^{jwt}$), they were set to zero. Analyzing field distributions at a single frequency (harmonic) is enough in many applications and, even if it is not, it may be possible to formulate a field distribution as a sum of harmonic components.

2.3.2 Feed model

The feed of a reflector antenna must radiate the energy that will be collimated by the reflector. It must be directive in order to waste less energy by missing the reflector.
One of the most simple mathematical feed model is the raised cosine, which can be written in feed coordinate system as

$$\vec{E}(\vec{r}) = \begin{cases} \frac{e^{-jkr}}{r}\cos^q\theta\left(\cos\phi\hat{\theta} - \sin\phi\hat{\phi}\right), & \theta \le 90^\circ \\ 0, & \theta > 90^\circ \end{cases} \tag{16}$$

where q is an exponent that controls the feed directivity, $k = w\sqrt{\mu\epsilon}$ is the wavenumber, μ and ϵ are the magnetic permeability and electric permittivity of free space, and j is the imaginary unit. Its simplicity allows some analytical analysis and its radiation pattern approximate real world feeds.
The raised cosine formulation is a typical far field equation: no field component towards $\hat{r}$, phase factor e^{-jkr}, and attenuation factor $1/r$. The remaining terms are related to field strength and polarization in each direction defined by ϕ and θ. Given the electric field $\vec{E}(\vec{r})$ in free space at $r >> 1/k$, the magnetic field $\vec{H}(\vec{r})$ is given by

$$\vec{H}(\vec{r}) = \frac{1}{\eta}\hat{r} \times \vec{E}(\vec{r}) \tag{17}$$

where η is the intrinsic impedance of free space.

2.3.3 Radiated power

The electromagnetic power density is given by the Poynting vector

$$\begin{aligned} \vec{W}(\vec{r},t) &= \vec{E}(\vec{r},t) \times \vec{H}(\vec{r},t) = \frac{1}{2}\left(\vec{E}(\vec{r})\,e^{jwt} + \vec{E}^*(\vec{r})\,e^{-jwt}\right) \times \frac{1}{2}\left(\vec{H}(\vec{r})\,e^{jwt} + \vec{H}^*(\vec{r})\,e^{-jwt}\right) \\ &= \frac{1}{2}\,\mathrm{Re}\left(\vec{E}(\vec{r}) \times \vec{H}^*(\vec{r})\right) + \frac{1}{2}\,\mathrm{Re}\left(\vec{E}(\vec{r}) \times \vec{H}(\vec{r})e^{j2wt}\right) \end{aligned} \tag{18}$$

where the first term independs on time t (mean power density) and the second term oscillates at $2w$.
The radiated power in free space of a source can be easier found by integrating the power density over a sphere in the far field region, where no field component towards $\hat{r}$ is a good approximation. It can be written as

$$P_{rad} = \frac{1}{2}\oint_s \mathrm{Re}\left(\vec{E}(\vec{r}) \times \vec{H}^*(\vec{r})\right) \cdot d\vec{s} = \frac{1}{2\eta}\int_0^{\pi}\int_0^{2\pi}\left|\vec{E}(\vec{r})\right|^2 r^2 \sin\theta d\phi d\theta \tag{19}$$

where s is an sphere of large radius and η is the intrinsic impedance of free space.
The radiated power of a raised cosine feed (16) is

$$\begin{aligned} P_{rad} &= \frac{1}{2\eta}\int_0^{\pi/2}\int_0^{2\pi}\frac{\cos^{2q}\theta}{r^2}r^2\sin\theta d\phi d\theta = \frac{\pi}{\eta}\int_0^{\pi/2}\cos^{2q}\theta\sin\theta d\theta \\ &= \frac{\pi}{\eta(2q+1)} \end{aligned} \tag{20}$$

where q is the raised cosine exponent.

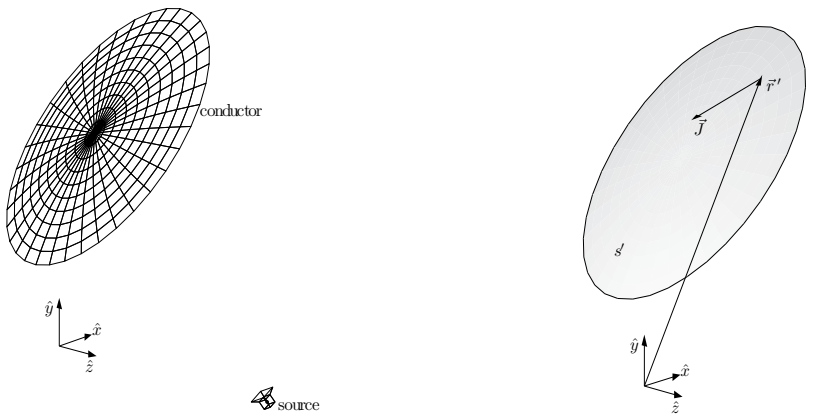


Fig. 8. Equivalence theorem: any field sources and scatters (a) can be replaced by equivalent current distributions (b).

2.3.4 Equivalent current sources

A common way to analyze reflector antennas is to replace the feed and reflectors (scatterers) by equivalent current sources, which is supported by the equivalence theorem (see Fig. 8).
For a single, smooth and perfect electric conducting reflector, the physical optics approximation is accurate and can be written as

$$\vec{J}(\vec{r'}) \approx 2\hat{n}(\vec{r'}) \times \vec{H}_i(\vec{r'}) \tag{21}$$

where $\hat{n}$ is the unit normal vector at the source point $\vec{r}'$, and $\vec{H}_i$ is the incident magnetic field at the source point $\vec{r}'$. Multiple reflectors may interact with each other, so that more complex procedures (e.g. method of moments) may have to be considered in order to determine accurate equivalent current sources.

2.3.5 Far field

Given a current density source $\vec{J}(\vec{r}')$ over a surface s' in free space as shown in Fig. 9, the electric field can be accurately approximated for $r >> 1/k$ by the integral

$$\vec{E}(\vec{r}) = -j\frac{k\eta}{4\pi}\frac{\mathrm{e}^{-jkr}}{r}\int_{s'}\left[\vec{J}(\vec{r}') - \left(\vec{J}(\vec{r}')\cdot\hat{r}\right)\hat{r}\right]\mathrm{e}^{jk\vec{r}'\cdot\hat{r}}\,ds' \tag{22}$$

where $k = w\sqrt{\mu\epsilon}$ is the wavenumber, μ and ϵ are the magnetic permeability and electric permittivity of free space, η is the intrinsic impedance of free space and j is the imaginary unit.
The far field integral can be evaluated by standard numerical methods (e.g. trapezoidal rule or Gaussian quadrature).

2.3.6 Definition of polarization

Polarization of a radiation is related to the orientation of the electric field. The third definition of Ludwig (1973) is very commonly used. It can be written as

$$E_{co}(\vec{r}) = \cos\phi E_\theta - \sin\phi E_\phi \tag{23}$$

$$E_{cx}(\vec{r}) = \sin\phi E_\theta + \cos\phi E_\phi \tag{24}$$

Fig. 9. Farfield $\vec{E}$ at observer point $\vec{r}$ relative to a current distribution $\vec{J}$ at source points $\vec{r}'$ over s' .

Note that the definition of co and cross polarization must be coherent to the feed radiation. For the raised cosine model (16), the field components in each polarization are

$$E_{co}(\vec{r}) = \frac{\cos^q \theta \, \mathrm{e}^{-jkr}}{r} \tag{25}$$

$$E_{cx}(\vec{r}) = 0 \tag{26}$$

for $\theta \leq 90^\circ$. Unfortunately, reflectors may insert cross polarized components into the collimated field, even when the source does not radiate them.

2.3.7 Gain relative to an isotropic radiation

An isotropic radiation is the one that has the same intensity regardless of the direction. Hence, if an isotropic antenna radiates a power P_{rad}, the power density at a distance r away from it is

$$\vec{W}(\vec{r}) = \frac{P_{rad}}{4\pi r^2}\hat{r} \tag{27}$$

considering that P_{rad} is uniformly distributed over the area of an sphere of radius r.
The power density relative to an electric field $\vec{E}$ in free space at $r >> 1/k$ away from the source is given by

$$\begin{aligned} \vec{W}(\vec{r}) &= \frac{1}{2}\,\mathrm{Re}\left(\vec{E}(\vec{r}) \times \vec{H}^*(\vec{r})\right) = \frac{1}{2\eta}\,\mathrm{Re}\left(\vec{E}(\vec{r}) \times \hat{r} \times \vec{E}^*(\vec{r})\right) \\ &= \frac{\left|\vec{E}(\vec{r})\right|^2}{2\eta}\hat{r} \end{aligned} \tag{28}$$

The gain relative to an isotropic radiation can be written as

$$G(\vec{r}) = \frac{W(\vec{r})}{\bar{W}(\vec{r})} = \frac{2\pi r^2 \left|\vec{E}(\vec{r})\right|^2}{\eta P_{rad}} \tag{29}$$

where P_{rad} is the feed radiated power and η is the intrinsic impedance of free space. When the gain is given in decibels (i.e. $10\log_{10} G$), the unit is called dBi to indicate that it is relative to an isotropic radiation.

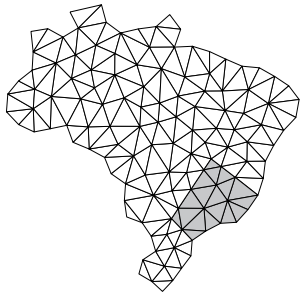


Fig. 10. Mesh over Brazilian territory used to define sample observer points.

3. Optimization problem formulations and solutions

3.1 Sampling of observer points

To formulate optimization problems, sample observer points $\vec{r}_i \in \mathbb{P}$, $i = 1, ..., n_r$, are uniformly spread all over the target. Fig. 10 shows a target instance, the Brazilian territory, where a triangular mesh is used to support the sampling. The antenna gain is evaluated in the vertices $\mathbb{V} \subset \mathbb{P}$ for integrative measures (e.g. energy confinement) and in the centroids of triangles $\mathbb{B} \subset \mathbb{P}$ for average measures (e.g. coverage). Subsets of samples can be differentiated to allow distinct specifications.

3.2 Useful measures

3.2.1 For coverage

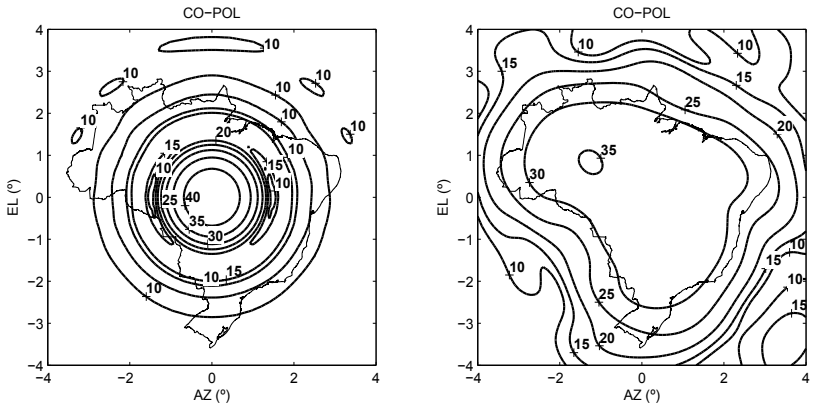
A good measure of coverage is the average gain observed at the sample points. Considering the gain given in dB and that a maximum coverage is desired, an implicit uniformity measure is imbued in the average value.

3.2.2 For energy confinement

Ideally, coverage also means energy confinement since illuminating a target implicitly means only that target. However, due to physical constraints, transition to outside regions is not discontinuous, so it is meaningful to formulate apart an energy confinement measure. The energy ratio inside the target can be numerically computed using the supporting mesh for sample observer points.

3.3 Instance formulations of satellite coverage problems

The following instances of satellite coverage problems consider the Brazilian territory as target of a single feed single reflector antenna. The reference reflector is an off-center parabolic with circular aperture of radius $R = 0.762$m, focal distance $F = 1.506$m, and offset $H = 1.245$m along $\hat{y}$ (Duan & Rahmat-Samii, 1995). The feed is modeled by a raised cosine radiation with exponent $q = 14.28$ at frequency $\nu = 11.95$GHz, whose reference position is the reflector focus and reference direction is $\tau = 42.77^\circ$ above the $-z$-axis on the $x = 0$ plane (Duan & Rahmat-Samii, 1995). The satellite (reflector) is directed to the center of mass of the target. The reflector shape is parameterized by four triangular Bézier patches and the feed is parameterized by its position and direction on the plane $x = 0$ of the reflector coordinate system (Lisboa et al., 2006). The number of design variables is $n = 38$ and they compose the respective vector of design variables $x \in \mathbb{R}^n$. The sampling is supported by a mesh with 125 vertices $\mathbb{V}$ and 190 triangle centroids $\mathbb{B}$, where 18 triangle centroids are labeled as southeast region $\mathbb{B}'$.



(a) (b)

Fig. 11. Radiation patterns of the reference parabolic antenna (a) and of the optimal antenna for maximum uniform coverage (b).

Some instances of satellite coverage problems are shown next. Each one is intended to tackle a specific point on design of reflector antennas. Hence, the formulation techniques can be combined to produce more complete optimization problem formulations.

3.3.1 Maximum uniform coverage

The most simple satellite coverage problem aims to uniformly cover the target with the maximum energy possible. It can be formulated as the monoobjective optimization problem (Duan & Rahmat-Samii, 1995)

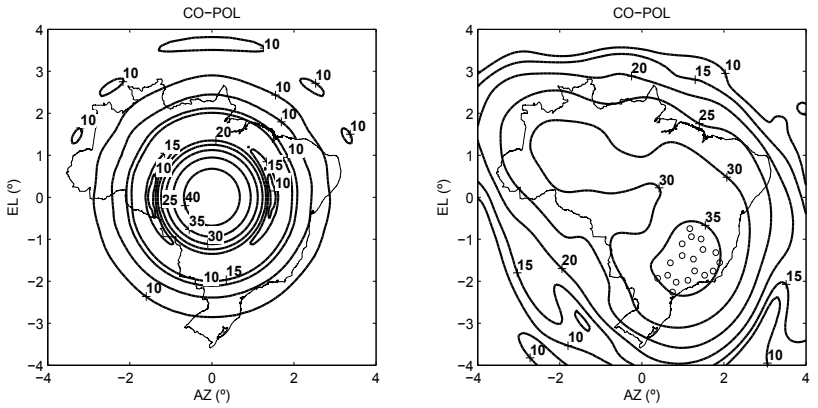
$$\text{minimize} \quad f(x) = -\frac{1}{n_b}\sum_{i=1}^{n_b} 10\log_{10} G(\vec{r}_i, x) \tag{30}$$

$$\text{subject to} \quad x_{\min} \leq x \leq x_{\max} \tag{31}$$

where $x \in \mathbb{R}^n$ is the vector of design variables, G is the antenna gain at an observer point $\vec{r}$, and $\vec{r}_i \in \mathbb{B}$ denotes an observer point derived from triangle centroids.

The optimization algorithm iteratively queries the oracle of the optimization problem (antenna simulator in this case) with a vector of design variables x_k at iteration k, which in turn must answer with the respective values of the optimization functions at x_k (only $f(x_k)$ in this case). Since the oracle query is expensive, the fewer iterations required for convergence, the better. Considering that the optimal antenna must still collimate energy since the target is far away, a highly directive classical reflector antenna is a good starting point.

Fig. 12 shows the optimal radiation pattern for maximum uniform coverage of the Brazilian territory (Lisboa et al., 2006). It was achieved by a search direction optimization algorithm with theoretical guarantees on improving the objective functions every iteration (Vieira, Takahashi & Saldanha, 2010), starting from a classical off-center parabolic reflector antenna.



(a) (b)

Fig. 12. Radiation patterns of the reference parabolic antenna (a) and of the optimal antenna for maximum uniform coverage constrained to lower bound of 35dBi in the southeast region (b).

3.3.1.1 Minimum gain constraints

To specify lower bounds of gain at subregions of the target, a subset of sample points can be used. It can be formulated as the monoobjective optimization problem (Vieira, Lisboa & Saldanha, 2010)

$$\text{minimize} \quad f(x) = -\frac{1}{n_b}\sum_{i=1}^{n_b} 10\log_{10} G(\vec{r}_i, x) \tag{32}$$

$$\text{subject to} \quad g_j(x) = G_{\min,i} - G(\vec{r}_i, x) \leq 0, \ \ \forall \vec{r}_i \in \mathbb{B}' \subset \mathbb{B} \tag{33}$$

$$x_{\min} \leq x \leq x_{\max} \tag{34}$$

where $x \in \mathbb{R}^n$ is the vector of design variables, G is the antenna gain at an observer point $\vec{r}$, $\vec{r}_i \in \mathbb{B}$ denotes an observer point derived from triangle centroids, and $G_{\min,i}$ is the desired minimum gain at $\vec{r}_i \in \mathbb{B}'$. Each sample observer point inside the subregion $\mathbb{B}'$ defines one constraint.

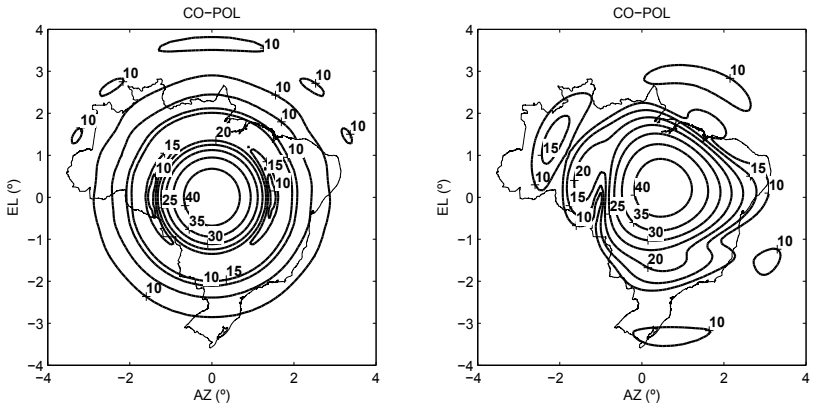
Fig. 12 shows the optimal radiation pattern for maximum coverage with a minimum gain 35dBi in the southeast region (Vieira, Lisboa & Saldanha, 2010), leading to a problem with 18 constraint functions. It was achieved by an ellipsoid method with an improved constraint functions treatment (Shor, 1977; Vieira, Lisboa & Saldanha, 2010; Yudin & Nemirovsky, 1977), starting from a classical off-center parabolic reflector antenna.

3.3.2 Maximum energy confinement

The search for maximum energy confinement can be formulated as the monoobjective optimization problem (Lisboa et al., 2008)

$$\text{minimize} \quad f(x) = -\sum_{i=1}^{n_b} \frac{A_i}{4\pi r_a^2}\frac{1}{3}\sum_{j=1}^{3} G(\vec{r}_{ij}, x) \tag{35}$$

$$\text{subject to} \quad x_{\min} \leq x \leq x_{\max} \tag{36}$$



(a) (b)

Fig. 13. Radiation patterns of the reference parabolic antenna (a) and of the optimal antenna for maximum energy confinement inside target area (b).

where $x \in \mathbb{R}^n$ is the vector of design variables, G is the antenna gain at an observer point $\vec{r}$, A is the area of triangles, and $\vec{r}_{ij} \in \mathbb{V}$ is used to denote the j-th vertex of the i-th triangle.
Fig. 13 shows the optimal radiation pattern for maximum energy confinement inside target (Lisboa et al., 2008). It was achieved by a search direction optimization algorithm with theoretical guarantees on improving the objective functions every iteration (Vieira, Takahashi & Saldanha, 2010), starting from a classical off-center parabolic reflector antenna.

3.3.3 Maximum coverage and energy confinement

Optimal tradeoffs between maximum coverage and energy confinement can be formulated as the solution of the multiobjective optimization problem (Lisboa et al., 2008)

$$\text{minimize} \quad f(x) = \begin{bmatrix} -\dfrac{1}{n_b}\displaystyle\sum_{i=1}^{n_b} 10\log_{10} G(\vec{r}_i, x) \\ -\displaystyle\sum_{i=1}^{n_b} \dfrac{A_i}{4\pi r_a^2}\dfrac{1}{3}\displaystyle\sum_{j=1}^{3} G(\vec{r}_{ij}, x) \end{bmatrix} \tag{37}$$

$$\text{subject to} \quad x_{\min} \leq x \leq x_{\max} \tag{38}$$

where $x \in \mathbb{R}^n$ is the vector of design variables, G is the antenna gain at an observer point $\vec{r}$, A is the area of triangles, $\vec{r}_i \in \mathbb{B}$ denotes an observer point derived from triangle centroids, and $\vec{r}_{ij} \in \mathbb{V}$ is used to denote the j-th vertex of the i-th triangle.
Fig. 15 shows the optimal Pareto front for coverage and energy confinement for the Brazilian territory, whose respective radiation patterns of marked points are shown, from maximum coverage to maximum confinement, in Fig. 14(a), 14(b), 14(c) and 14(d). It was achieved by launching several times a search direction optimization algorithm with theoretical guarantees on improving all the objective functions every iteration (Vieira, Takahashi & Saldanha, 2010), starting from a classical off-center parabolic reflector antenna.

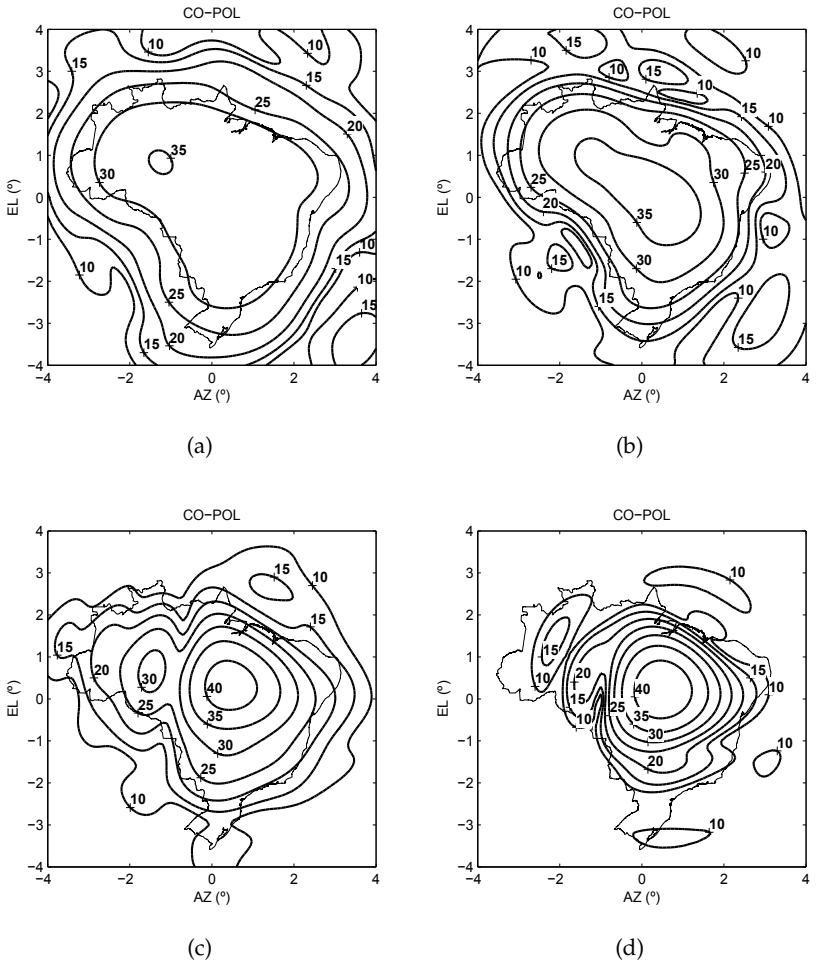


Fig. 14. Radiation patterns for maximum coverage (a), intermediate tradeoff for coverage and energy confinement (b-c), and for maximum energy confinement (d).

3.3.4 Broadband operation

To cope with broadband specifications, sample frequencies can be chosen inside the range so that the set of objective functions is considered for each sample frequency (Lisboa et al., 2006; 2008). For example, 5 sample frequencies for coverage and energy confinement lead to an optimization problem with 10 objective functions.

Another possible formulation is to take the worst case of each objective function inside the frequency range, which can be even more meaningful in practice. However, the convergence rate will be slower, considering that optimization algorithms that can directly cope with multiple objectives without any scalarization procedure (Vieira, Lisboa & Saldanha, 2010; Vieira, Takahashi & Saldanha, 2010) are faster when the number of objective functions increases. A final remark in this sense, is that the number of sample frequencies can be increased until it surpasses the number of design variables. From this point on, the Pareto set is likely to have

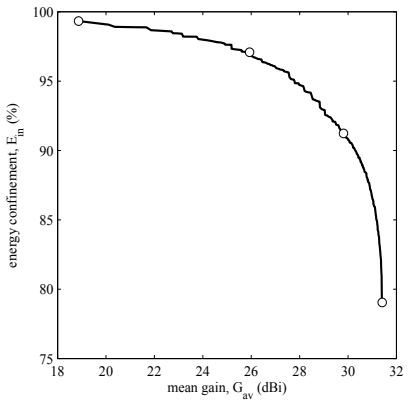


Fig. 15. Pareto front for coverage and energy confinement.

the same dimension of the search space, making numerically much easier to assert if a point is a local optimum.

Acknowledgements
This work was supported by FAPEMIG, CAPES and CNPq, Brazil.

4. References

Duan, D.-W. & Rahmat-Samii, Y. (1995). A generalized diffraction synthesis technique for high performance reflector antennas, *IEEE Transactions on Antennas and Propagation* **43**(1): 27–39.

Lisboa, A. C., Vieira, D. A. G., Vasconcelos, J. A., Saldanha, R. R. & Takahashi, R. H. C. (2006). Multi-objective shape optimization of broad-band reflector antennas using the cone of efficient directions algorithm, *IEEE Transactions on Magnetics* **42**: 1223–1226.

Lisboa, A. C., Vieira, D. A. G., Vasconcelos, J. A., Saldanha, R. R. & Takahashi, R. H. C. (2008). Decreasing interference in satellite broadband communication systems using modeled reflector antennas, *IEEE Transactions on Magnetics* **44**: 958–961.

Ludwig, A. C. (1973). The definition of cross polarization, *IEEE Transactions on Antennas and Propagation* **21**(1): 116–119.

Shor, N. Z. (1977). Cut-off method with space extension in convex programming problems, *Cybernetics* **12**: 94–96.

Vieira, D. A. G., Lisboa, A. C. & Saldanha, R. R. (2010). An enhanced ellipsoid method for electromagnetic devices optimisation and design, *IEEE Transactions on Magnetics* **(accepted for publication)**.

Vieira, D. A. G., Takahashi, R. H. C. & Saldanha, R. R. (2010). Multicriteria optimization with a multiobjective golden section line search, *Mathematical Programming* **1-31**.

Yudin, D. B. & Nemirovsky, A. S. (1977). Informational complexity and effective methods for solving convex extremum problems, *Matekon* **13**: 24–45.

Satellite Laser Communication With Widely Dispersed Ground Stations

Paul Christopher
Leesburg, VA 20175
pfchristop@aol.com

Abstract

We use a cloud attenuation model derived from key Italsat results to find laser cloud attenuation for satellite-ground links. Gaseous attenuation is included for 2- 10 micron attenuation with the aid of a MODTRAN model. New silicon laser developments at 3 microns allow special 10 and 3 micron comparisons. Ten microns are emphasized. High worldwide laser attenuation and relatively low link availability are indicated for latitudes south of Washington, DC.

Link availability is raised with suitable Northern Latitude satellites, such as Earth Observation Satellites. Soviet cloud correlation results indicate that link availability would be raised to acceptable levels with ground sites separated by 100-390 km.

1. Introduction

The Jet Propulsion Laboratory and Kim (1,2) have recognized that optical communication near one micron wavelength can offer significant broadband advantages for satellite to ground communication links. They recognized that optical advantages over conventional satellite frequencies at 12-14 GHz included small size, low cost convenient signal processing, and freedom from government frequency regulations. Laser communication links also hold the promise of high data rate. Satellite-ground links would, however, be complicated by propagation loss and sporadic availability.

Chu and Hogg (3) allowed us to note (4) that 10 micron lasers would have less atmospheric attenuation than 1 micron lasers. The 10 micron attenuation (function of longitude, latitude, and link availability) is available as a Mathematica function (4). The results can be shown as zenith attenuation maps, and we emphasize modest 80% availability here for reasonable link attenuation. We include attenuation, size, and cost features here for a fuller comparison of lasers for satellite-ground links. We combine the Chu and Hogg cloud attenuation results with the discrete atmospheric windows shown by MODTRAN (5). These windows include favorable observations near 2, 3, and 10 microns.

Laser links might be especially useful for sun synchronous Earth Observation Satellites. Landsat 4 and 5 missions were precursors to the valuable missions which may occur for EOS, but we reconsider the 701 km Landsat altitude. Large clouds of space debris were released by the Chinese antisatellite test of January 2007. The average altitude of the debris

is near 800km, and we suggest new sun synchronous orbits near 1000 km. We show a key equation for required sun synchronous inclination (Appendix A) as a function of semi major axis, and we show the altered inclinations for the new altitudes. We show figures of the new earth observation orbits.

The Chu and Hogg cloud attenuation implied notable advantages for 10 micron lasers relative to the shorter wavelengths. In addition, we use a MODTRAN model to show atmospheric transmittance regions between 1 to 10.6 microns. The combination of Chu and Hogg cloud attenuation with MODTRAN gaseous attenuation is used as a basis for comparisons of laser links. The combination often favors 3 and 10 micron links. The attenuation results are also found to be a strong function of Latitude and Longitude.

Worldwide attenuation plots are emphasized for the 10 and 3 micron regions. The shorter wavelengths are indicated to become competitive with 10 micron attenuation at Latitudes greater than 50 degrees north. Communication links at Northern latitudes are especially interesting for EOS because downlinks are concentrated near the poles. The polar EOS tends to have successive repeat passes over Northern ground stations.

The shorter wavelengths also have some intrinsic advantages in size and system cost. Satellite system cost studies as Teichmann (6,7,8) tend to place high emphasis on system antenna gain. More than half of the system cost for optimum system studies was implied to reside in the antenna subsystem to achieve high gain. The shorter wavelength lasers would overcome part of their attenuation disadvantages by achieving high gain at small apertures. We discuss the system cost relations of lasers in the 2 to 10 micron region, and find that 3 micron lasers may be competitive with 10u EOS downlinks at high latitudes.

2. Recent Developments

Recent Silicon Laser Developments

We applied the Chu and Hogg insights to the cloud water content which can be derived from key millimeter wave results of Barbaliscia et al (9,10) by analytical methods (11,12). Ten micron communication links appeared attractive at mid and high latitudes. Also, Frank Hanson (13) had fascinating comparisons of 10 microns and 3 micron links.

Since then (14), important new silicon laser developments have been proceeding rapidly. Bahram Jalali observed that silicon allowed photon-phonon exchange (the Raman Effect) to proceed better than in other materials. He has noted that silicon lasers are promising for 2.3 to 7 micron radiation. Fig. 2-1 is a schematic of the Jalali amplitude as the radiation traverses a silicon light pipe. The amplitude builds rapidly along the length of the silicon.

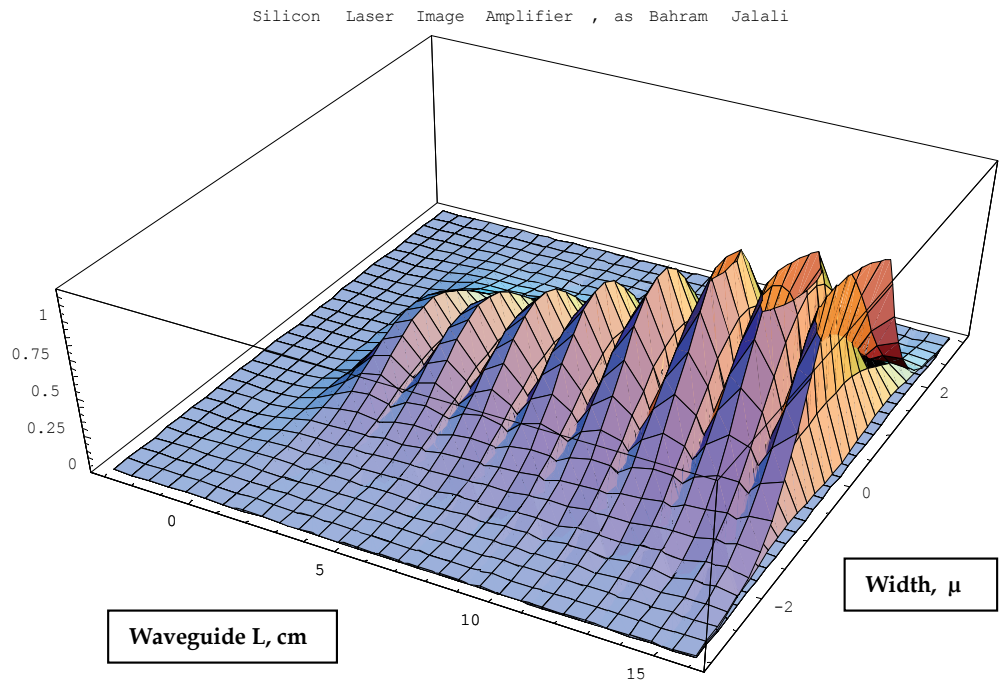


Fig. 2-1. Amplitude v. Length for Silicon Laser, after Jalali May10—Jalali2.nb.

The effectiveness and reliability of the silicon lasers in the estimated 2.3- 7 micron range is a strong inducement for us to include this range of wavelengths for Earth Observation Satellite laser link studies.

Intense Satellite Debris Field Developments

A Chinese anti-satellite test in January 2007 left an enlarged debris field between centered near 800 km altitude. It stretches from 700 km to over 900 km altitude. In addition, a Russian satellite hit a valuable Iridium satellite in 2009 at lower altitude. Therefore, we alter a useful 701 km Landsat orbit to nearly 1000 km altitude. The inclination must also be altered to keep the useful sun-synchronous features. Fig. 2-2 indicates the mean debris field, and Fig. 2-3 the higher orbit required to clear the debris. Appendix A discusses the change in inclination needed for the higher orbit.

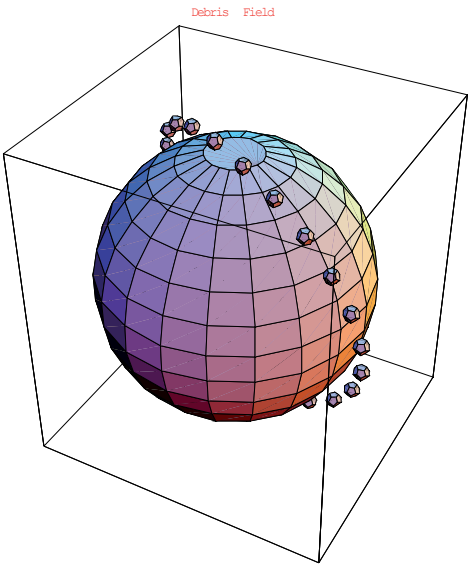


Fig. 2-2 ASAT Debris Field, Mean Altitude 800 km

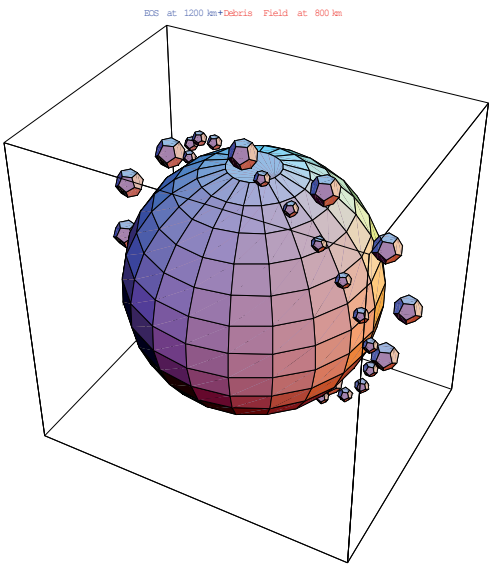


Fig. 2-3 EOS Higher (1000km) than Debris Field

Mathematica Apr27—EOScx.nb.

Valuable EOS views of Earth include the endangered Arctic region (Fig. 2-4). The coastlines are shown in relief so that visibility regions can be found quickly.

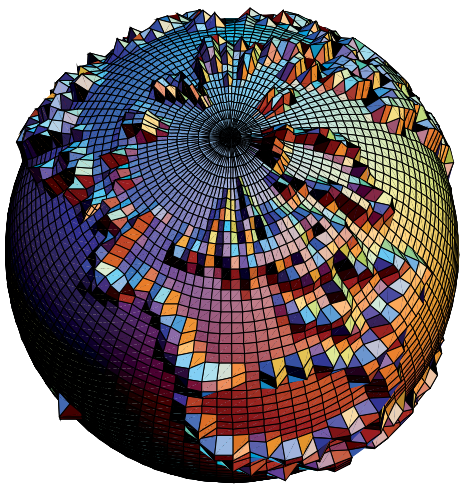


Fig. 2-4 EOS View of Canada, Alaska, Arctic Ocean Jan8ConstEOS--.nb

3. Laser Propagation Comparisons

Four decades ago, Chu and Hogg attempted to compare optical communication with infrared communication at 10.6 microns for terrestrial communication. They found that 10 micron signals propagated much better through fog than did the optical communication.

Fig. 3-1 indicates the kind of signal loss they anticipated over a wide range of frequencies. The upper curve indicates loss through 1 km of fog with 0.1 gm/m3 liquid water density, vs Frequency on the abscissa.
Note that 10 microns corresponds to 30 THz and the 1 micron region corresponds to 300THz. The fog comparison for this figure indicates about 60 dB and 200 dB loss, respectively, for the two frequencies.

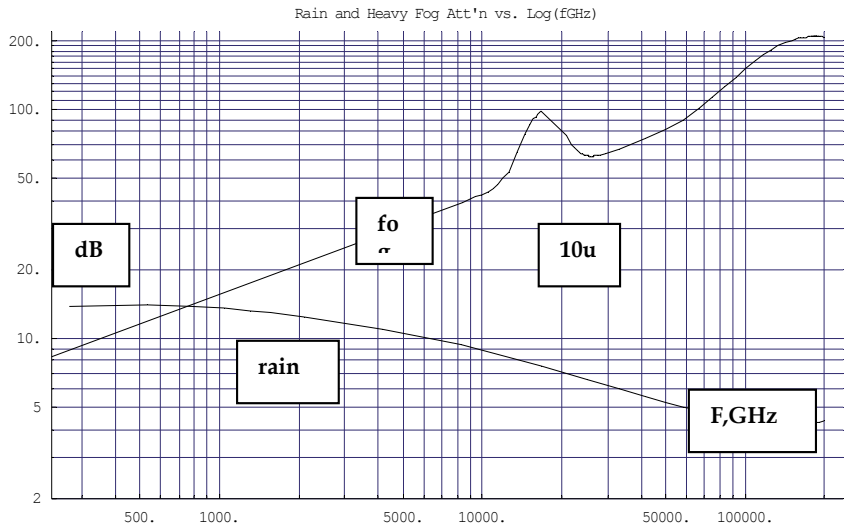


Fig. 3-1 Chu and Hogg's Fog (top) and Rain Attenuation (dB) v. Frequency
Note, 10u at 30,000 GHz and 1u at 300,000 GHz

Cloud attenuation can be estimated as a function of worldwide location and desired availability. An 80% availability was used for 10 micron propagation in order to keep propagation loss to reasonable levels. A form of Quad-Diversity was chosen for the ground site to get desired overall link availability. Four sites on the order of 15 km apart were used to approach 98-99% availability. The diversity section below will indicate even higher link availability at hundreds of kilometer separation.
We also add gaseous attenuation here for completeness. A smoothed MODTRAN result is seen as Fig. 3-2. It includes gaseous attenuation for a moderate atmosphere.

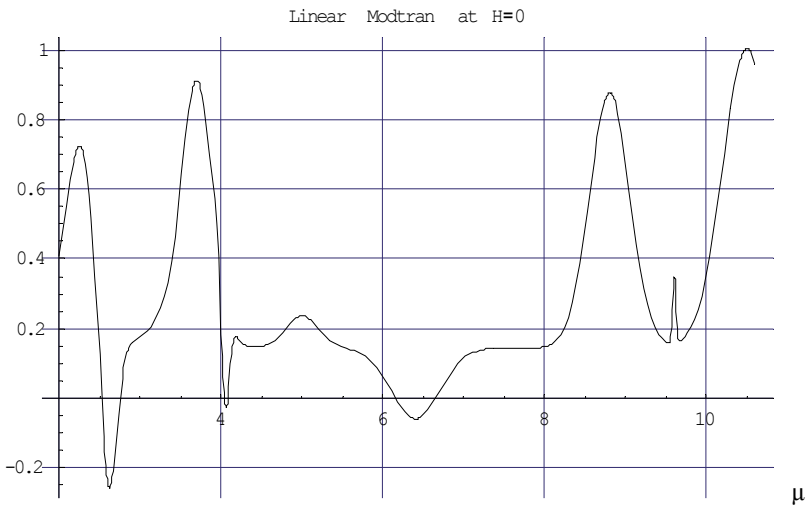


Fig. 3-2 Smoothed MODTRAN Transmission v. Wavelength, Microns

The MODTRAN model can be combined with the cloud zenith attenuation model indicate low attenuation at 10u and 3 u.

Variations in attenuation for the 10 micron and 3 micron maps may be seen as Figs. 3-3, -4.

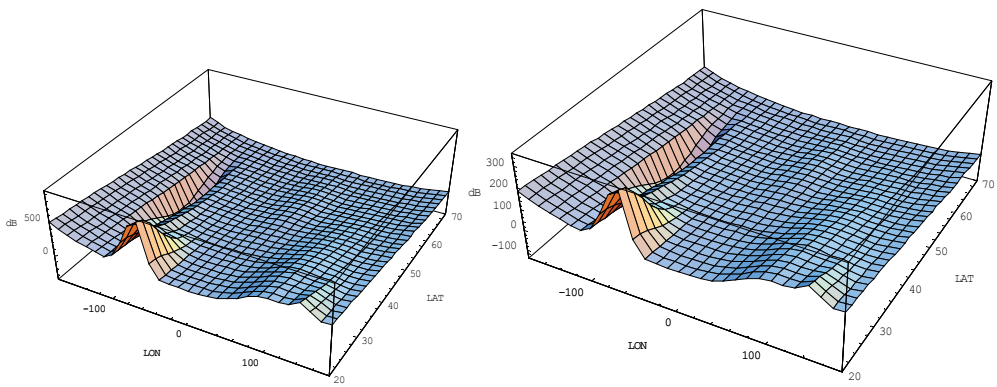


Fig. 3-3 10 Micron Attenuation Map 80% Availability

Fig. 3-4 3 Micron Attenuation Map

The ten micron cloud attenuation varies, as clouds form and dissipate. The attenuation may be described by an exponential probability density function, with high attenuation at the 90^{th} percentile level. For convenient probability calculations, this may be referred to as the 10 percent exceedance level. Fig. 3-5 shows the intense attenuation that a ground observer would experience by looking straight up (zenith), looking through the clouds. An observer near Miami may experience hundreds of dB attenuation.

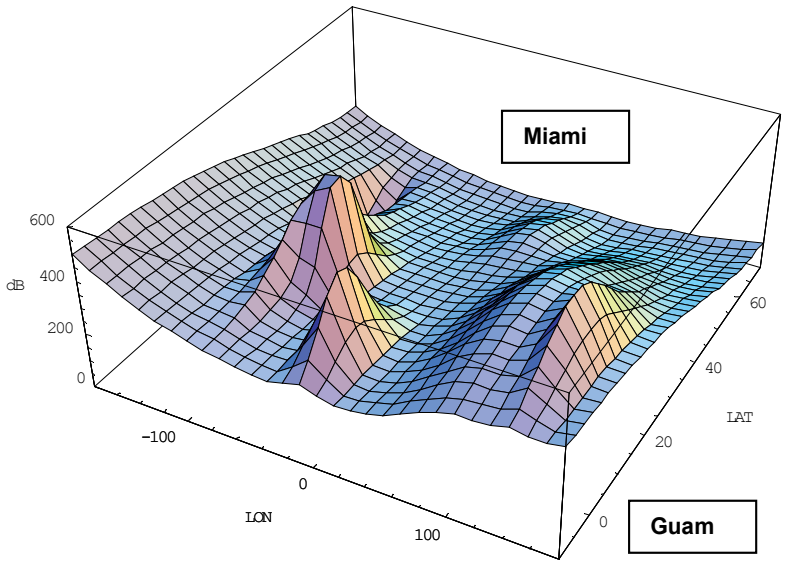


Fig. 3-5 10u Zenith Attenuation in Northern Hemisphere 90% Availability [PR=.1] Mathematica Jan22010--.nb.

Many successful systems would require attenuation less than 10 or 15 dB. Lower link availability, as Fig. 3-6, would be required. Fig 3-6 indicates zenith attenuation at 80% availability, and many important ground sites in northern New England (note Boston 42 deg North) may enjoy reasonable attenuation as 10 dB.

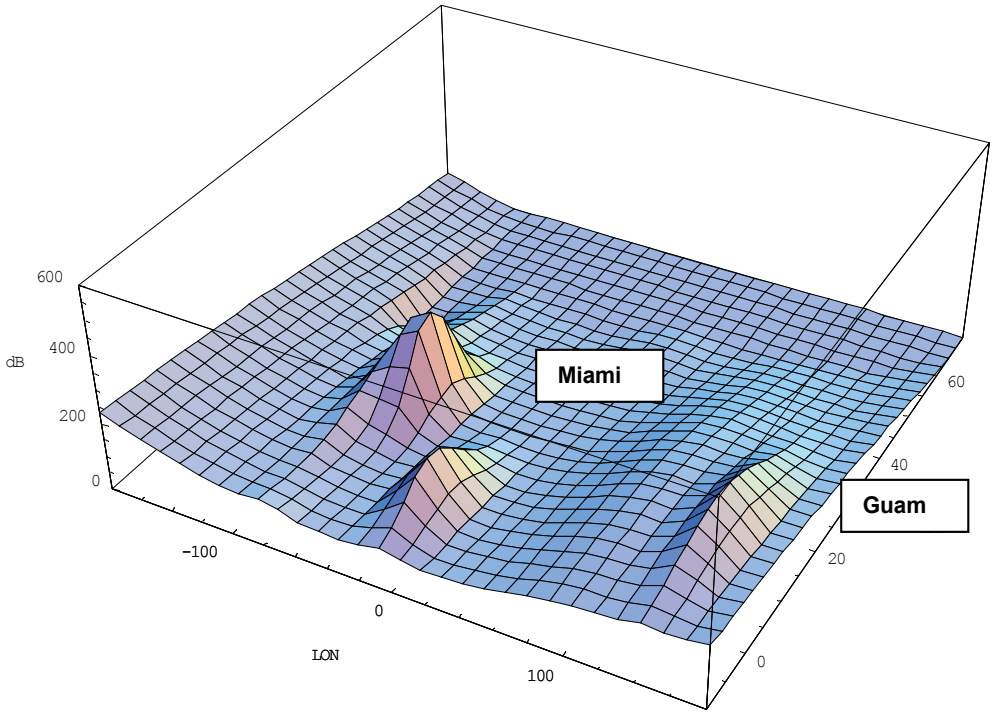


Fig. 3-6 10u Zenith Attenuation in Northern Hemisphere
80% Availability[PR=.2] Mathematica Jan22010--.nb .

Zenith attenuation at 10 microns is notably less than attenuation at 3u (Fig 3-7).

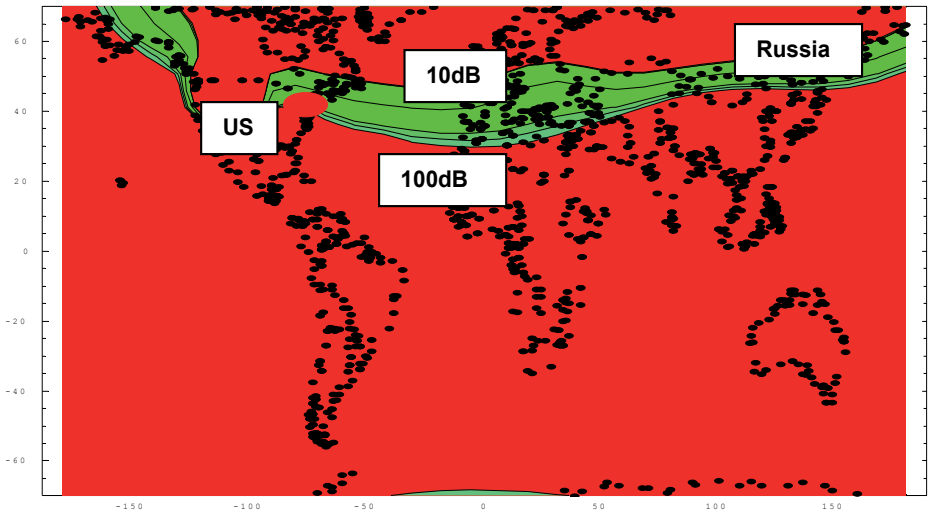


Fig. 3-7 Worldwide 3u Attenuation Contours, 80% Availability [PR=.2] Jan12010—nb.

The modest 10 micron attenuation at 80% availability allows attractive communication sites to appear, as Bangor Maine near 10 dB loss and 50dB near Washington DC, as in Fig 3-8. Difficult or untenable attenuation as 80-100 dB is indicated near Charleston, SC.

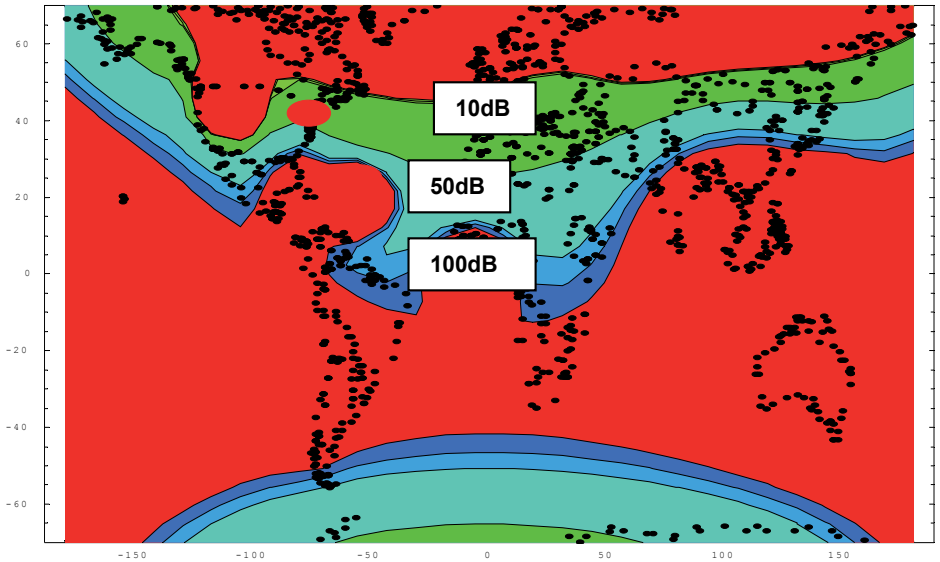


Fig. 3-8 Worldwide 10u Attenuation Contours, 80% Availability [PR=.2] Jan12010—nb.

The asymmetry in the attenuation contours between the northern and southern hemispheres is apparent in Fig. 3-8. In addition, the attenuation of the southern hemisphere has not been as thoroughly studied as the northern hemisphere.

The 80% attenuation can offer interesting ground sites, but if we let availability standards slip even further to 70%, the area of ground sites expands sharply, as in Figs. 3-9 and -10.
The areas east of Cuba and off West Africa are the only intense propagation loss regions remaining at both 3u and 10u in Figs. 3-9 and -10, respectively.

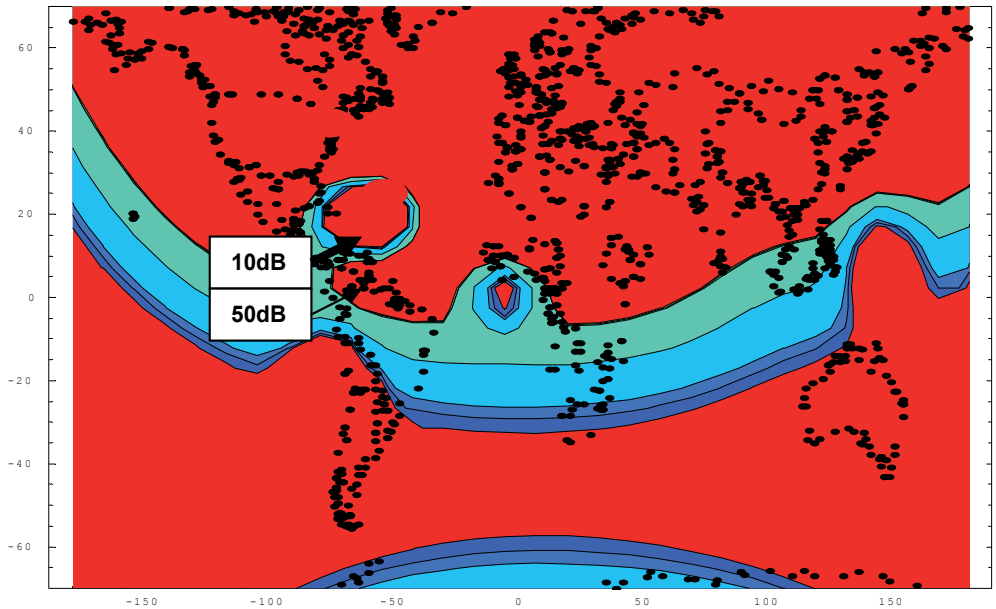


Fig. 3-9 Worldwide 3u Attenuation Contours, 70% Availability [PR=.3]

The attractive coverage for 3u at 70% availability extends even further into the Southern Hemisphere for 10u, as in Fig 3-10. Appendix B lists the 10u attenuation equation.

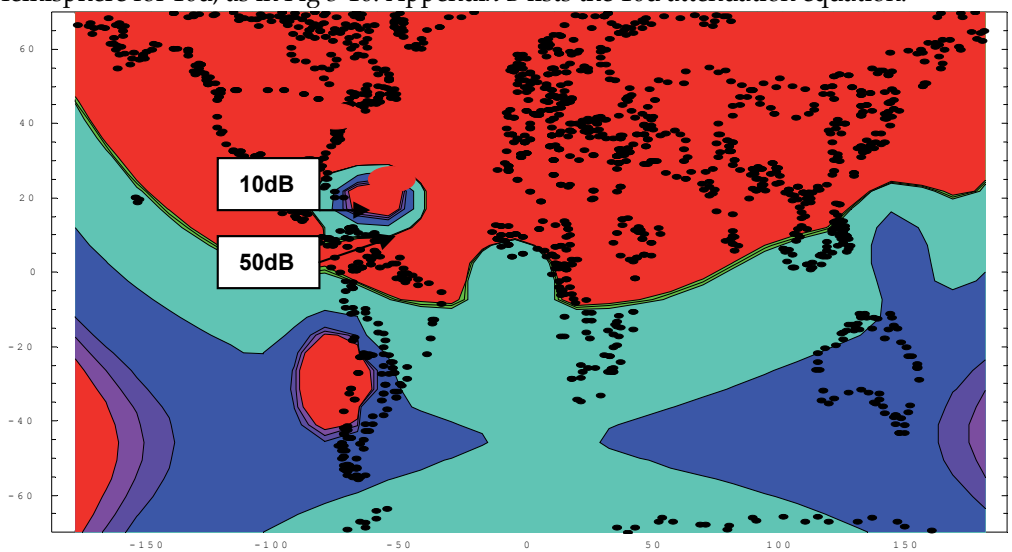


Fig. 3-10 Worldwide 10u Attenuation Contours, 70% Availability [PR=.3] Jan22010--.nb.

Of course, the limited single link availability we have discussed here is not acceptable for most important communication applications. Key Japanese tests (15) of 1u laser links addressed the question of extending the observed 20% single link availability. They placed laser receiver sites throughout the country, and were delightfully surprised that availability increased to greater than 90%. The Soviets may have anticipated this sharp rise in

availability with cloud correlation studies in the early '70s. We will address the Soviet cloud correlation function in Section 4, to imply much better availability when 100-390 km site separation is available.

The ratio of 3micron/ 10micron attenuation can also be found. The ratio appears useful for estimating where 3 micron lasers could be seriously considered as an alternative to 10 microns. Fig. 3-11 uses the ratio, and draws contours at locations where 3 micron lasers are estimated to be competitive. Regions north of the contours of Fig. 3-6 may be considered as reasonably competitive 3 micron regions. Note that regions north of Labrador and Denmark appear to be a demarcation for 3 micron laser use.

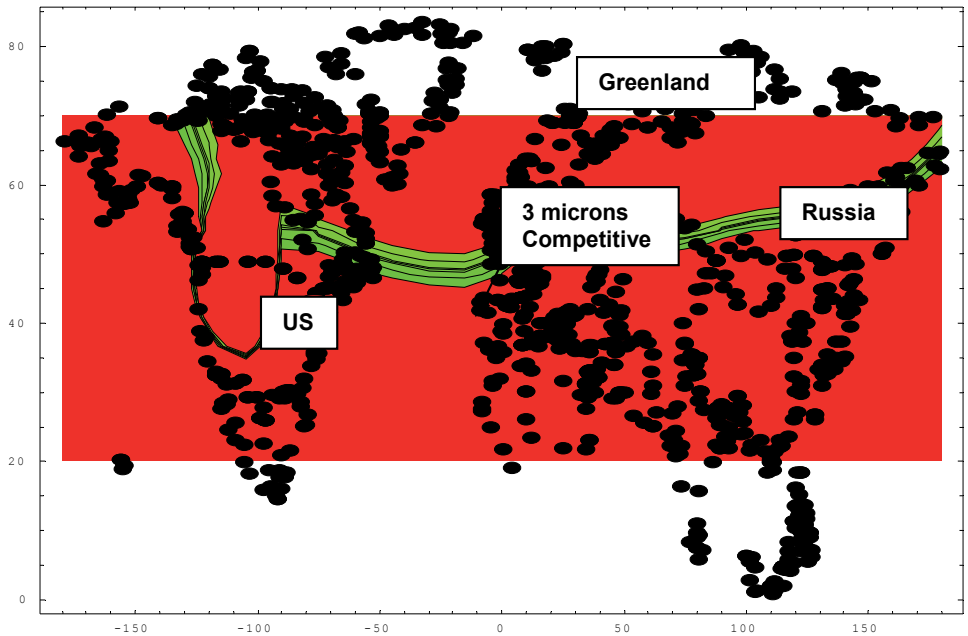


Fig. 3-11 Contours of 10u/3u Attenuation; Transition Contours to 3u at Labrador-Denmark Mathematica May7—CountryBig.nb.

Cost Advantages for 3 Microns

Key satellite system optimization studies (6,7,8) have placed a heavy emphasis on system antenna gain. Further, high gain at small size and cost is emphasized. Lasers can provide this high required antenna gain, but with disadvantages of high propagation loss.

A cost term as

$$\text{Cost/Bit} = \left(\frac{\text{loss}}{f^2} \right)^{1/3}$$

was found (5) to include both the loss disadvantages and the gain advantages of a satellite system. This is also closely related to a method of minimizing (Loss,dB – 20 LogF) for millimeter wave system. This cost per information bit method can be applied to Fig. 3-6 to find the transition line moves slightly south for the system cost comparisons, as Fig. 3-12.

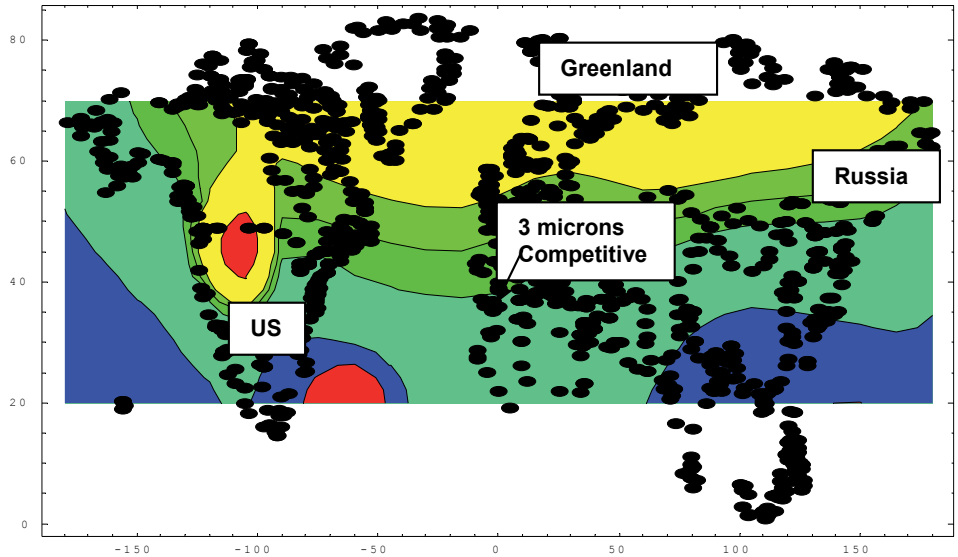


Fig. 3-12 Estimated Cost Contours for 3 Micron/ 10 Micron Comparisons

A transition region to 3 microns is now indicated to move south from Labrador to Northern Maine.

Intense 10u Attenuation at 90% Availability

The modest 80% availability attenuation of Figs 3-8 would require ground site diversity for useful communication. Alternatively, one might consider 90% availability for single sites (Fig. 3-13).

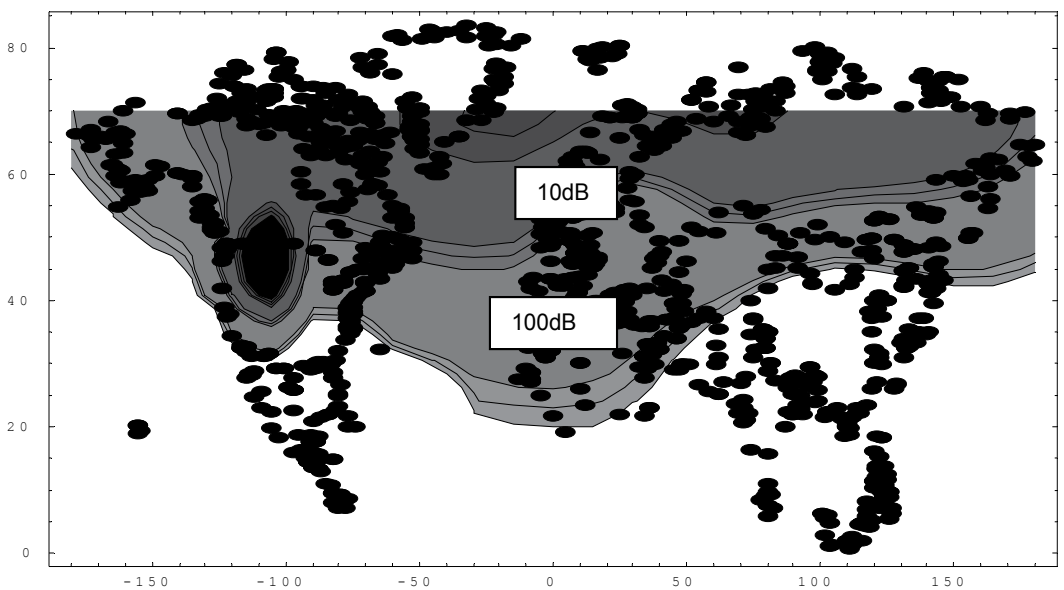


Fig. 3-13 Intense 10 Micron Attenuation Contours for 90% Availability May17-PR1.nb

Severe 100dB zenith attenuation contours are indicated at Labrador and Denmark if 90% availability is required for single links.

4. Diversity Analysis

Boldyrev and Tulupov (16) developed cloud correlation functions to describe the cloudy/ clear relations between two separated ground stations. The correlation functions were compiled from many satellite observations, with large scale characteristics for hurricanes and much smaller sizes for more typical observations. Fig. 4-1 shows correlation function results for a number of different seasons.

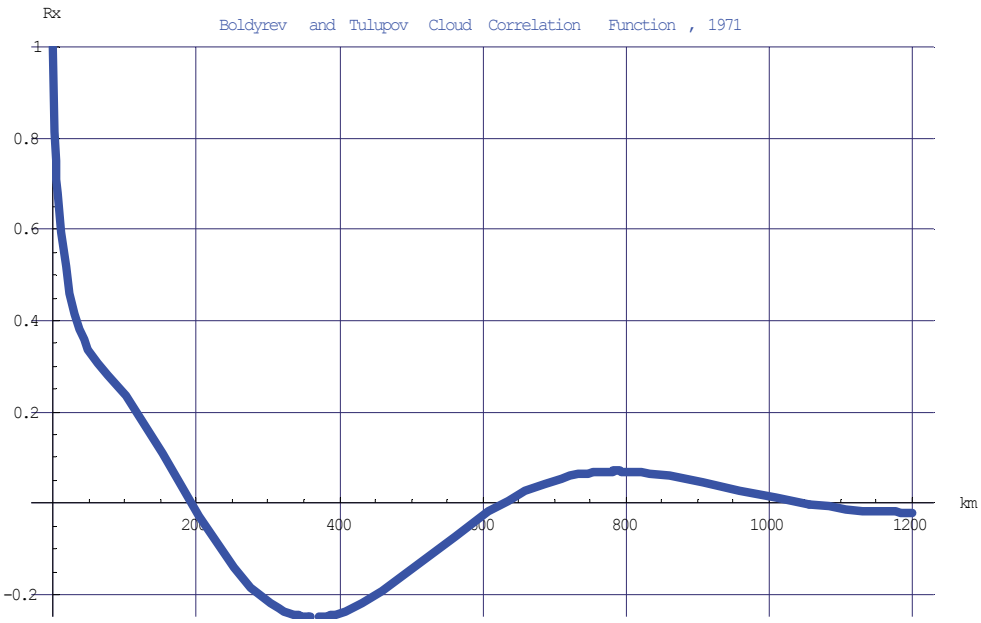


Fig. 4-1 Boldyrev and Tulupov Cloud Correlation Function

Boldyrev and Tulupov recognized that a correlation for all clouds could be represented conveniently by

$$\text{Cloud correlation}=0.2e^{-x}+e^{-0.036x}-e^{-0.15x}+0.8e^{-0.003x}\text{Cos}[0.0075x] \quad (4\text{-}1)$$

where x= separation, km

Their insight into the importance of the correlation function will next be seen with a description of Pierce's bivariate exponential probability density function.

Boldyrev's function can be used as a direct input into a correlated bivariate exponential pdf, and then to develop quantitative results for diversity advantages. We use the form of the correlated exponential pdf as:

$$p(A1,A2) = \frac{e^{-\frac{A1+A2}{B(1-r)}}\, p\, I_0\left(2\sqrt{\frac{A1\,A2\,r}{B^2(1-r)^2}}\right)}{B^2(1-r)} \tag{4-2}$$

where A1 and A2 are the rain attenuation (dB) at the 2 sites
r= correlation between sites, found from Boldyrev
B=standard deviation of the univariate pdf, and
p= conditional probability that rain attenuation is observed; as 0.80

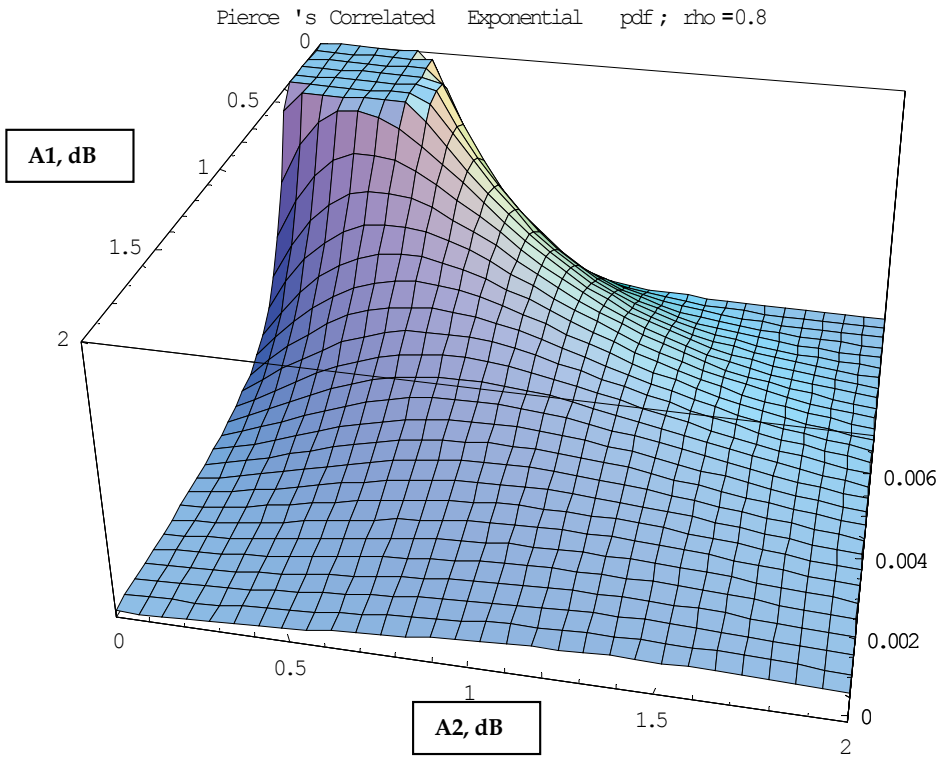


Fig. 4-2. Pierce's Correlated Exponential pdf (Eq. 4-2)

The double integral on the density function should be evaluated to find the joint probability of exceeding arbitrary attenuation levels. The probability of both sites having attenuation greater than AR (dB) may be functionally shown as Eq. 4-3.

P[A1>AR, A2>AR]=

$$\frac{p\int_{AR}^{\infty}\int_{AR}^{\infty} e^{-\frac{A1+A2}{B(1-r)}} I_0\left(2\sqrt{\frac{A1\,A2\,r}{B^2(1-r)^2}}\right) dA2\,dA1}{B^2(1-r)} \tag{4-3}$$

The low values of the Boldyrev correlation function (r) will be key to finding low probability of attenuation AR. System operators will recognize AR (dB) as the rain attenuation available with switched diversity, when they can choose the site with the least rain attenuation.
The integral has not yielded to attempts to integrate it exactly, and it turns out to be a very lengthy numerical integration. An upper limit is chosen as (10 B) rather than infinity. The exceedance probability (Eq. 2-3) is abbreviated as (pr) below.
The result of the double integration can be found as Figures 4-3 and 4-4. Fig. 4-3 shows exceedance probability (PR=Log10(pr)) v. normalized (AR/B) and correlation coefficient r. Exceedance probability contours may also be shown as Fig. 4-4 for 0<r<1.

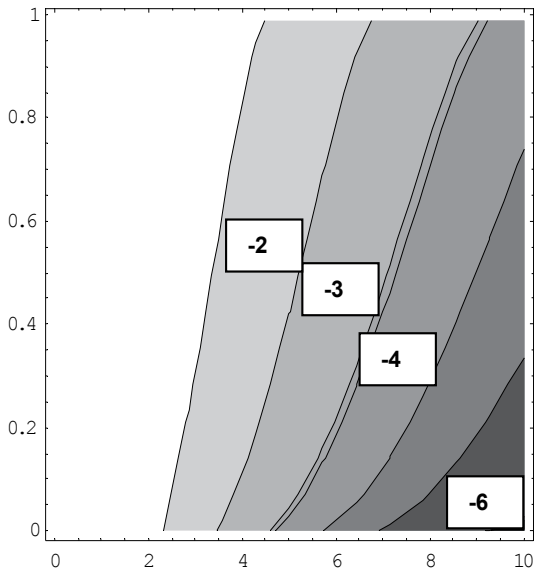


Fig. 4-3. PR v. (AR/B), r

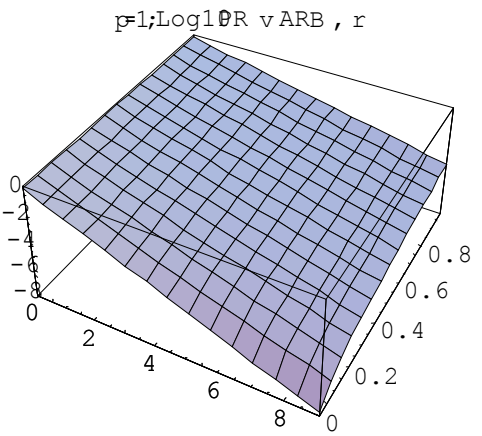


Fig. 4-4 PR Contours v. (AR/B), r

The bivariate function of Fig. 4-3 is defined for all possible weather events, even clear weather at both ground sites. Instead, the exponential density function should be defined for rain events in the 1% to 0.1% range: This is where the exponential pdf has the most relevance. We note again that PR= -0.1 corresponds to 90% availability.

The Boldyrev function offers even more promise for ground diversity, however.

The integral can be evaluated numerically even for negative correlation, as occurs at 390 km for the Boldyrev function. Sharp new benefits can be found in the results of Fig 4-5, with the negative correlation in the 200km-390 km interval as specified by Boldyrev.

The joint exceedance probability is seen to take a quantum step down in the 200-390 km region. This means that if one ground site has bad weather, another ground site 200-390 km away probably has good weather. Indeed, one can test this by experience: if you get a blizzard, your cousin 300 km away is likely to have brilliant blue skies. Two widely separated ground sites are not just better than independent sites, but are much better. Fig. 4-5 extends 4-4 to negative correlation, and the sharp change in PR can be seen at correlation r= -0.2.

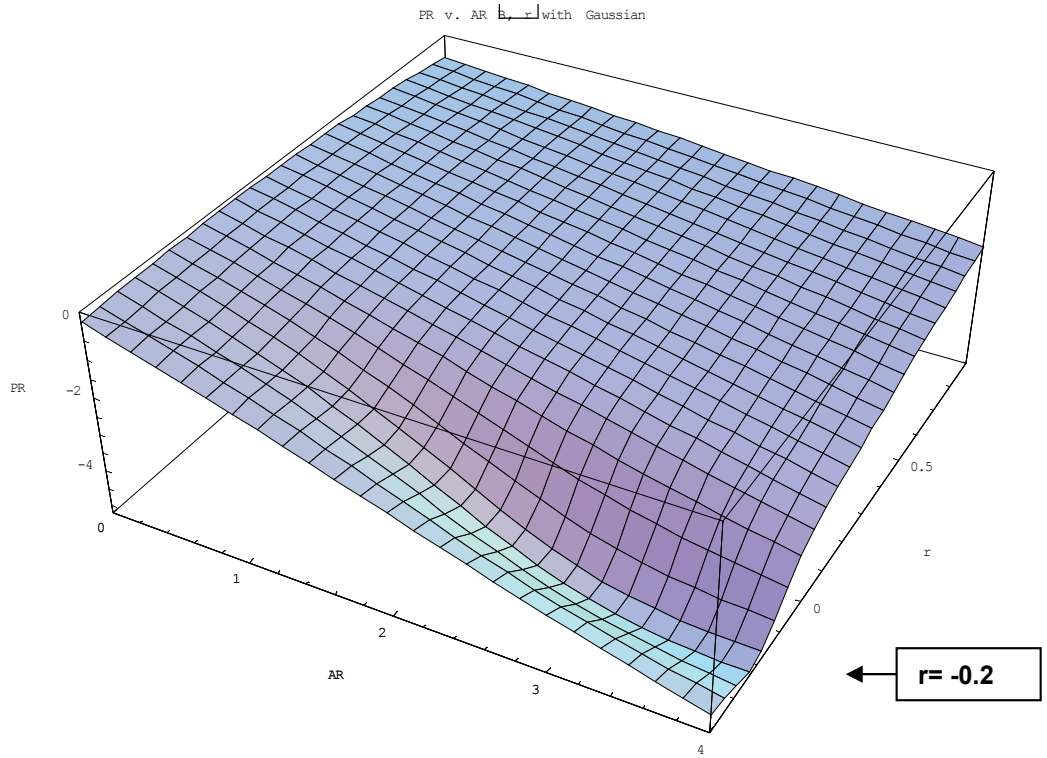


Fig. 4-5 PR v. Single Link Attenuation AR, Cloud Correlation r Mathematica Jan18PROBlistb--.nb.

A short approximation for the exceedance probability PR can be found as:
Log10[Exceedance]=

$$0{,}131989 - 0{,}861214\,AR - 0{,}0113578\,AR^2 - 1{,}69174\,e^{\frac{1}{2}(3-AR)^2-50.(-0.2-r)^2} + 0{,}0383318\,r + 0{,}603411\,AR\,r - 0{,}651209\,r^2 \quad (4\text{-}4)$$

Where AR is single link attenuation(dB), with standard deviation B=1
r= cloud correlation coefficient

The probability contours may be seen more clearly as Fig. 4-6.

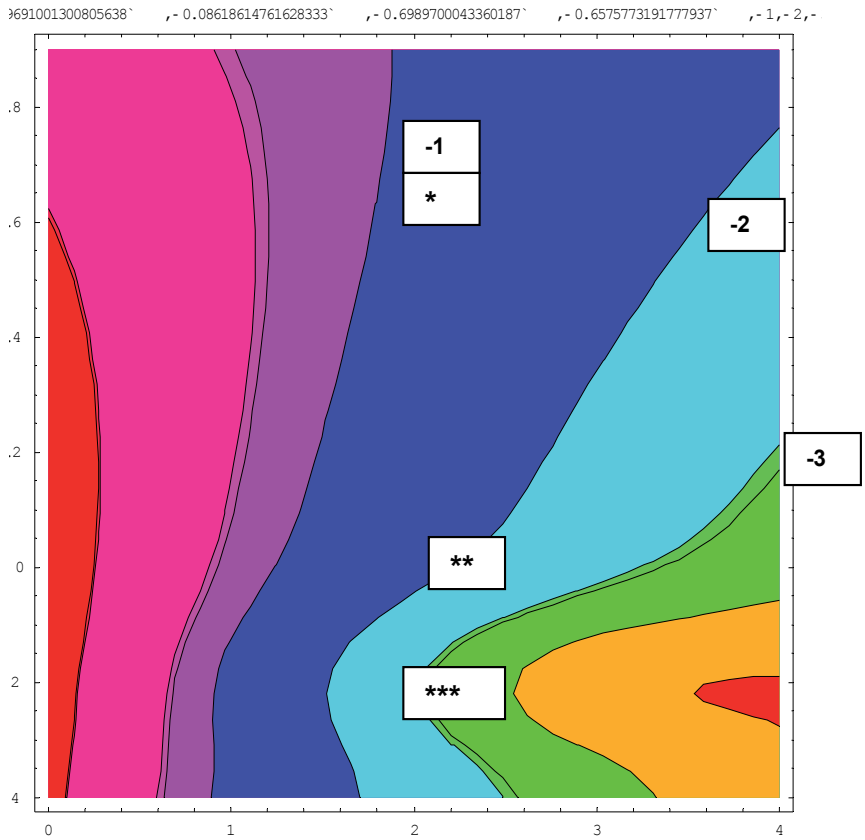


Fig. 4-6 Probability Contours v. AR/B, Correlation Coefficient r
Mathematica Jan18PROBlistb--.nb.

The PR contours can be checked immediately by comparing the PR=-1 and PR=-2 curves.
At constant attenuation AR, point (*) at correlation=1 corresponds to point (**) at correlation=0.
This is necessary because 0.1 exceedance must correspond to 0.01 at zero correlation. The benefits of negative correlation, as seen by Boldyrev, are seen by further dropping to point 3 (***) at correlation= -0.2. Point 3 shows much better probability, approaching PR=-3 or 0.001 exceedance probability, or 99.9% availability.

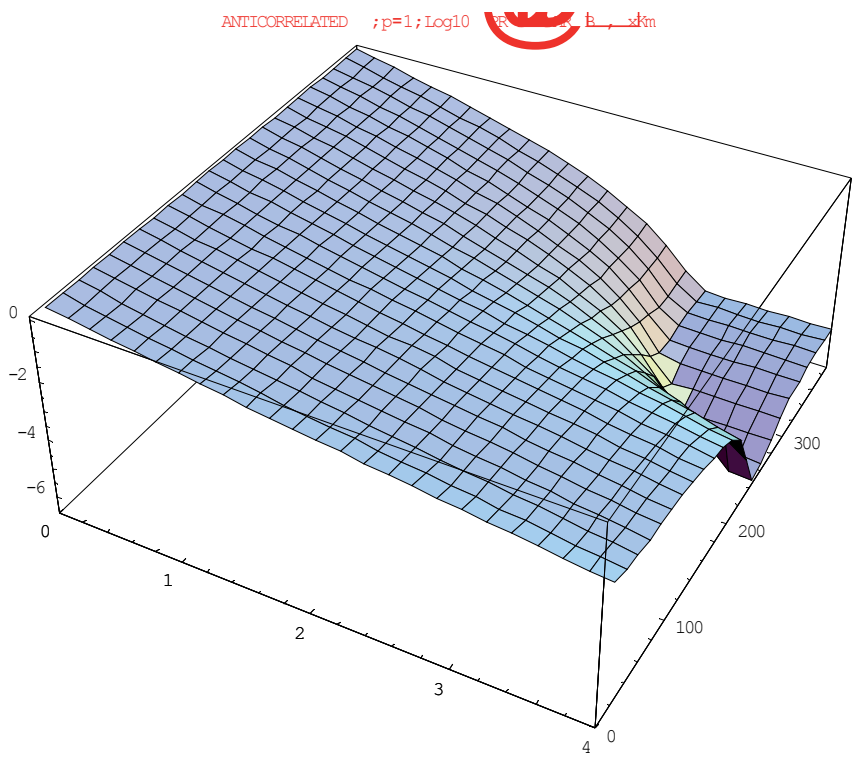


Fig. 4-7 PR v. Attenuation, Distance (km) by Boldyrev and Pierce Mathematica Jan122010—Boldeyrev.nb

Fig. 4-7 imposes the Boldyrev correlation function as a function of distance. The cloud correlation function is 1 at x=0 and drops to 0 at x=200 km. It then drops to negative values for 200<x<390 km, and exceedance probability PR drops abruptly to give clear advantages over independent sites. With independent sites (correlation=0) a single site outage probability of 20% would drop to 4%, or 96% availability. Site separation as 300- 390 km would imply even better availability, from further analysis of Fig. 4-6.

The 80% exceedance (20% availability) of the Japanese experiments would be implied to improve to nearly 85% availability, with an approximate solution of Fig 4-7. The approximate solution is seen as the contours of Fig 4-8. The Japanese experimenters were delighted with results near 90% availability for widely separated ground stations in Japan.

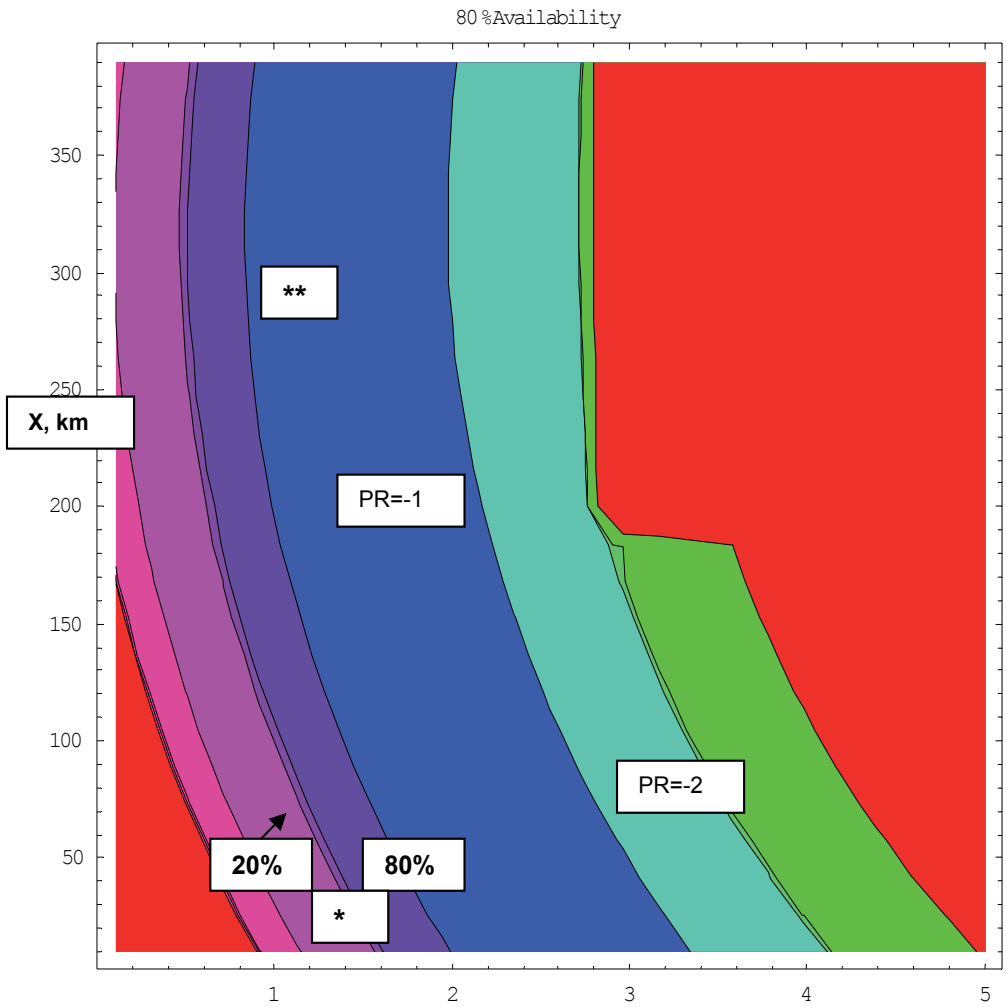


Fig. 4-8 Approximate Availability Contours v. Attenuation, Distance
Compare *20% Availability at 0km with **85% at 300 km

6. Summary

We applied Chu and Hogg's key cloud attenuation results to Italsat observations (9) and cloud water analysis (11) to estimate worldwide laser attenuation for satellite- ground communication links. Section 3 gave worldwide zenith attenuation estimates for 80% availability, 90% availability, and 70% availability. The equation for cloud attenuation at 70% availability is given in Appendix B. We` emphasized 10 micron wavelengths and 80% availability for modest attenuation. Reasonable availability was suggested by using Boldyrev and Tulupov's negative cloud correlation at large distances, as 200 – 390 km. Section 4 outlined the analysis and anticipated results for dual switched ground sites. Fig. 4-8 anticipated the encouraging current Japanese (15) results: 20% single site availability growing to over 85% at 300 km site separation. Eighty percent availability and zero

correlation at 200 km were indicated to give 96% availability for the dual sites, and 390 km was indicated as notably better. Even better performance should be expected for quad diversity.

We addressed possible 10 micron laser reliability issues by considering new silicon lasers in the 3 -10 micron range. The cloud attenuation model (1,2) and a smoothed MODTRAN atmospheric transmission model indicated the 3 micron space-ground attenuation as too high except in regions north of Labrador-Denmark. The 3 micron lasers might be well suited to small Earth Observation Satellites with widely dispersed ground sites at far northerly latitudes.

Laser space-ground links would be similarly attractive for Molniya satellites and temperate and Arctic ground locations. Indeed, Molniya satellites may avoid a difficult acquisition concern intrinsic to the fast moving EOS. The Molniya satellites would have an additional concern: attenuation would be slightly higher than the zenith attenuation shown here. This has been addressed at the Ka Conference in Naples (4).

Acknowledgements

Daniel Brandel of ITT noted concerns for 10 micron CO2 laser reliability. David Brown of Wolfram Research (17) gave key advice on Mathematica worldwide mapping functions. Analytical Graphics (AGI) gave excellent descriptions of the Chinese ASAT debris.

7. Reference

K.E. Wilson et al, “Overview of the Ground-to-Orbit Lasercom Demonstration,” Free-Space Laser Comm. Tech IX, SPIE, Jan. 1997. The GOLD experiment was intended for ground to GEO communication with the Japanese ETS-VI satellite. Instead, the ETS-VI was stuck in the transfer orbit and only partial data was collected. The 1.024 MB/s links showed 10E-4 BER on the uplink and 10E-5 BER on the downlink.

I. Kim et al., “(Very) Preliminary Results of the STRV-2 Satellite-to-ground Lasercom Experiment,” Free-Space Laser Comm. Tech. XII, SPIE, vol. 3932, Jan. 2000.

T.S.Chu and D.C.Hogg, “Effects of Precipitation at 0.63, 3.5, and 10.6 Microns”, Bell System Technical Journal, May-June 1968

Paul Christopher, “Considerations for Lunar Relays with Millimeter Wave and IR Communication,” Ka and Broadband Conference, Naples, Sept. 2006.

Report of the APS Study Group of boost- phase intercept systems for national missile defense. American Physical Society, NY, July 15, 2003.

T. Teichmann, “Some Elementary Considerations of Satellite Earth Communication Systems,” IRE Trans. Space Electronics and Telemetry, VOL SET-7, June 1960.

P. Christopher, “Optimum Space-Ground Tradeoffs for Millimeter Wave Satellite Communication,” Proc. Int’l Mobile Satellite Conference, Pasadena CA, June 1997.

Richard Orr, Leonard Schuchman, Proc. Ka and Broadband Conference, Naples, Italy, Sept. 2006.

F. Barbaliscia, M. Boumis, A. Martellucci, “World Wide Maps of Non Rainy Attenuation for Low-Margin Satcom Systems Operating in SHF/EHF Bands,” Ka Band Conference, Sept. 1998.

F. Barbaliscia, M. Boumis, A. Martellucci, "Characterization of Atmospheric Attenuation in the Absence of Rain in Europe in SHF/EHF Bands for VSAT Satcom Systems Applications," Ka Band Conference, Sorrento, Italy, Sept. 1997.

Paul Christopher, "World Wide Millimeter Wave Attenuation Functions from Barbaliscia's 49/22 GHz Observations," Ka Band Conference, Taormina Sicily, Oct. 1999.

Paul Christopher, "Millimeter Waves for Broadband Satellite Communication, 75- 98 GHz;" Proc. Ka and Broadband Conference, Isle of Ischia, Oct. 2003

Frank Hanson, "Coherent laser radar performance in littoral environments- a statistical analysis based on weather observations," Optical Engineering, Vol. 39 No. 11, Nov. 2000, 3044-3052.

Bahram Jalali, "Making Silicon Lase," Scientific American, pages 58-65, February 2007.

Yoshihisa Takayama, personal communication on Japanese Satellite Laser Experiments, at Ka and Broadband Conference, Cagliari, Sicily, Sept. 2009.

Boldyrev and Tulupov, Cloud Correlation Function, COSPAR Space Research XI, Leningrad USSR 20-29 May 1970, Vol. 1 Akademie-Verlag Berlin, 1971.

Michael Trott, The Mathematica Guidebook for Symbolics, Springer, New York, NY, 2006.

Appendix A Required Inclination for Higher EOS Orbit

The typical Landsat 4 orbit, at 701 km altitude, and 98.4 degrees inclination, must be altered for higher altitudes. The method of Osculating Elements shows the rate of change of Right Ascension as

$$\frac{3\,J2\,\sqrt{\frac{mu}{a}}\,r^2\,(5\cos^2(i)-1)}{4\,a^3}$$

with i as the inclination, a as the semimajor axis, and mu the Earth gravitational constant.

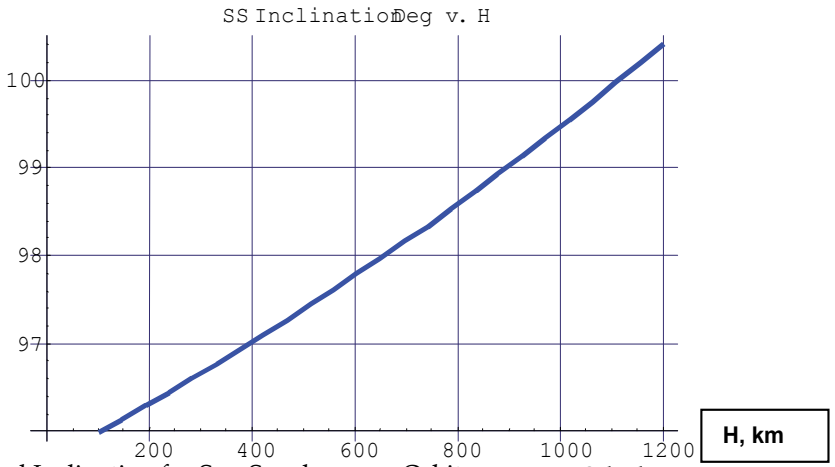


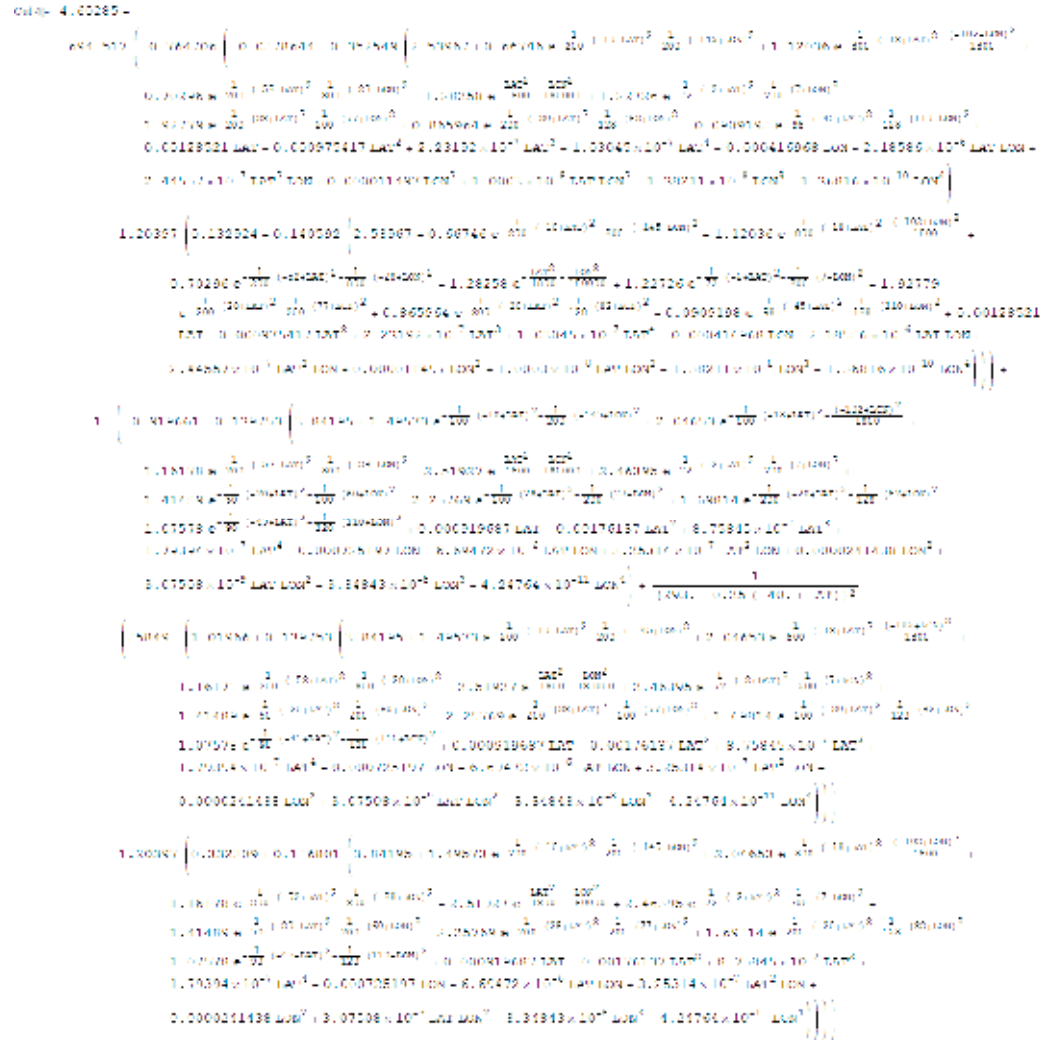
Fig. A-1 Required Inclination for Sun Synchronous Orbits May17 – LS4b.nb .

The shift from the Landsat4 altitude to 1000 km is seen to require almost 99.5 deg inclination.

Appendix B 10 Micron Zenith Attenuation, dB PR=.3

This Mathematica program is available from the author. It gives zenith attenuation maps at 70% availability.

Zenith Attenuation [70%]=



dB

(B-1)

Where LON= Longitude, deg East

LAT= N. Latitude

Mathematica [Jan2—Point3.nb]

22

Satellite Motion

Miljenko Solarić
University of Zagreb - Faculty of Geodesy
Croatia

1. Introduction

What is satellite?
The word satellite is coming from the Latin language (Latin *satelles* – escort, companion). Satellites are objects that rotate around the planets under the influence of the gravitational force. For example, the Moon is natural satellite of the Earth.
What is the artificial Earth satellite?
The artificial Earth satellites are artificial objects which are launched into orbit around the Earth by a rocket vehicle. This kind of satellites was named, the Human-Made Earth Satellites.
How do rockets function?
Aeroplanes work on the principle of buoyancy difference on their wings. This is the reason why aeroplanes can fly only in the air but not in the vacuum. Thus, an airplane cannot be used for launching satellites in their orbit around the Earth (Fig. 1).

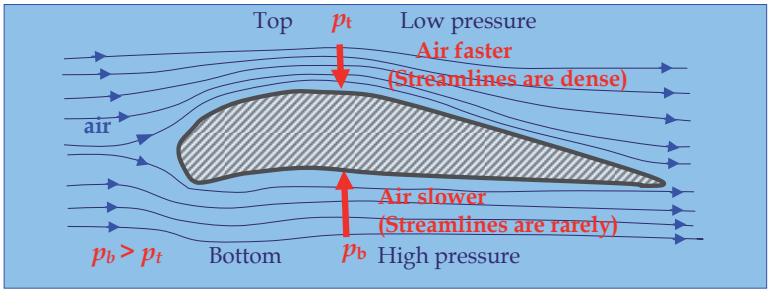


Fig. 1. When the aeroplanes have velocity in the air on the bottom of his wings they have higher pressure then on the top of wings. This difference of pressure is giving the force of buoyancy and the aeroplanes can fly.

It is also not possible to launch a human-made satellite into the orbit around the Earth with a cannon or a gun because a cannon-ball has the velocity of about 0.5 km/s. It means that this velocity of cannon-ball is about 15 times smaller than the first cosmic velocity (7.9 km/s). So for it has been possible to launch a satellite into an orbit around the Earth where there is vacuum is possible only with using rockets.
The word *rocket* comes from the Italian *Rocchetta* (i.e. *little fuse*), a name of a small firecracker. It is commonly accepted that the first recorded use of a rocket in battle was by the Chinese

in 1232 against the Mongol hordes at Kai Feng Fu. The Mongols were the first to have applied rocket technology in Europe as they conquered some parts of China and of Russia, Eastern and Central Europe.

Konstantin Tsiolkovsky (1857-1935) (Fig. 2) from the Imperial Russia and after from the Soviet Union published the first serious scientific work on space travels titled ''*The exploration of Cosmic Space by Means of Reaction Devices*'' in 1903. He is considered by many to be the *father of theoretical astronautics.* He also advocated the use of liquid hydrogen and oxygen for propellant, calculating their maximum exhaust velocity. His work inspired further research, experimentation and the formation of Society for Studies of Interplanetary Travel in 1924. Also in 1924, Tsiolkovsky wrote about multi-stage rockets, in ''*Cosmic Rocket Trains*''.

Fig. 2. Konstantin Tsiolkovsky

Fig. 3. Robert Goddard

Fig. 4. Sergei Korolev

Fig. 5. Wernher von Braun

In the USA, Robert Goddard (Fig. 3) began a serious analysis of rockets in 1912. It can thus be concluded that conventional solid-fuel rockets needed to be improved in three ways. One of them is that rockets could be arranged in stages. He also independently developed the mathematics of rocket flight. For his ideas, careful research, and great vision, Goddard was called the *father of modern astronautics.*

After the World War II in the USA Wernher von Braun (Fig. 5) and Sergei Korolev (Fig. 4) in the Soviet Union were the leaders in the advancing rockets technology.

The operational principle a rocket can be explained by means of a balloon (Fig. 6).

In a balloon the pressure of gas is practically equal on all sides (Fig. 6 a). When this balloon has an aperture then a particle of gas on this aperture will under pressure of gas be thrown out with velocity v_g (Fig. 6 b). The pressure in the balloon in opposite direction of the aperture will produce the pressure on the balloon and will give it the velocity v_B. The pressures in the other directions will be mutually cancelled in opposite direction. The rocket operates on this principle.

A rocket travelling in vacuum is accelerated by the high-velocity expulsion of a small part of its mass (gas). The Fig. 7. represents a rocket with the situations before and after the explosion. This is closed material system and for this system linear momentum needs to be conserved. So we can say that the momentum in the beginning position ($M \cdot \mathbf{v}$) is equal to momentum of this system after the explosion when a particle with the mass (dm) is be thrown out with velocity ($\mathbf{u}_{rel}$) in the opposite direction of this rocket velocity.

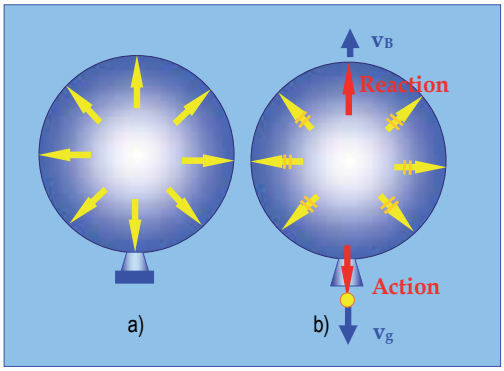


Fig. 6. A balloon closed under gas pressure and after opening an aperture.

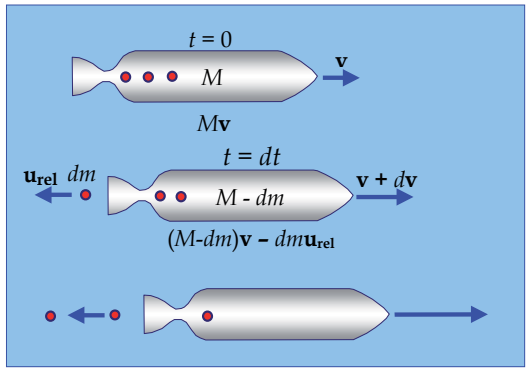


Fig. 7. The principle of a rocket operation.

It means that the momentum for a particle is negative ($-dm \cdot \mathbf{u}_{rel}$) and we can write the equation (Carton, 1965) and (Danby, 1989):

$$M \cdot \mathbf{v} = (M - dm) \cdot (\mathbf{v} + d\mathbf{v}) - dm \cdot \mathbf{u}_{rel} . \tag{1}$$

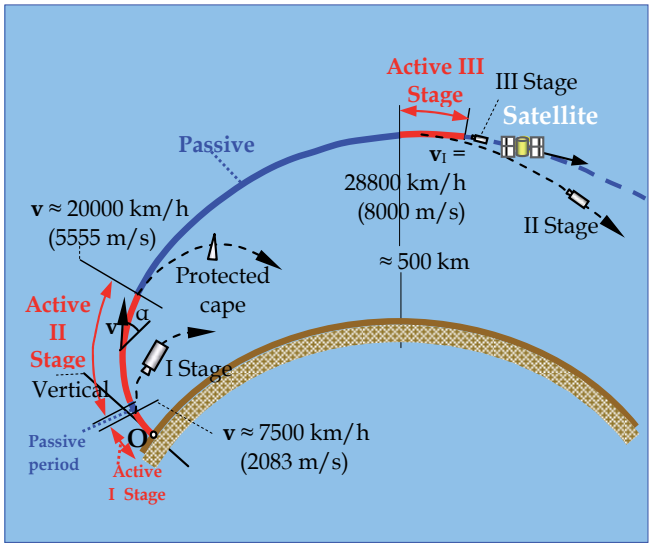


Fig. 8. Imagery of launching a satellite by rocket with the three stages.

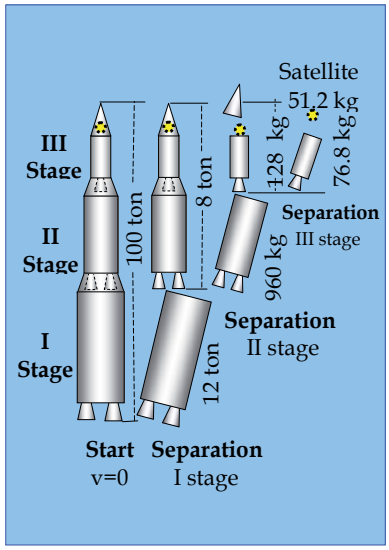


Fig. 9. For example during the start a three stages rocket may have the mass 100 t but mass of a space vehicle will be only 51.2 kg.

It follows from this equation that the velocity of rocket increased for the elementary magnitude

$$d\mathbf{v} = \frac{dm(\mathbf{v} + \mathbf{u}_{\text{rel}})}{M - dm} . \tag{2}$$

From this equation it is possible to see that the increase of the rocket velocity is larger if the velocity $\mathbf{u}_{\text{rel}}$ of the particles of gas is maximally the greater when the mass of particle dm has some magnitude. This is the reason why the constructors of rockets like to make rockets with very high (maximal) velocity of the particles (gas) of the rocket.
Usual rockets have vertical start. Longer delaying of rockets in the Earth gravitation field causes the loss of velocity but also to big thrust during the start is not suitable. So rockets are usually made in same stages (Fig. 8 and 9). During the starts of rockets the consummation of fuel is very large so that a satellite or a space vehicle enters into the orbit with a mass practically next to nothing (see for example Table 1).

Table 1. Data on initial mass of a rocket in a start, spend fuel, thrown parts for a launching space vehicle of the mass 51.2 kg

Stage	1	2	3
Initial mass	100 t	8 t	640 kg
Spend fuel	80 t	6.4 t	512 kg
Thrown part	-	12 t	960 kg
Definitely mass	20 t	1.6 t	128 kg
Velocity	3860 (m/s)	7720 (m/s)	11580 (m/s)

2. Planet Motion

Until the 17th century people were thinking that the Sun and planets are rotated around the Earth by circles. Such opinions were practically usual until Johann Kepler.

2.1 Kepler's Laws of Planetary Motion

Johann Kepler (1546 -1601) discovered the laws of planetary motion empirically from Tycho Brahe's (1546 -1601) astronomical observations of the planet Mars. The first and the second laws he published in *Astronomia Nova (New Astronomy*) in 1609, and the third law in *Harmonices mundi libri V (Harmony of the World*) in 1619.

a) Kepler's First Law of Planetary Motion
This law can be expressed as follows:

The path of each planet describes an ellipse with the Sun located at one of its foci.
(The Law of Ellipse)

This first Kepler's Law (Fig. 10 and 11) is sometimes referred to as *the law of ellipse* because planets are orbiting around the Sun in a path described as an ellipse. An ellipse is a special

curve in which the sum of the distances from every point on the curve to two other points (foci F_1 and F_2) is a constant.

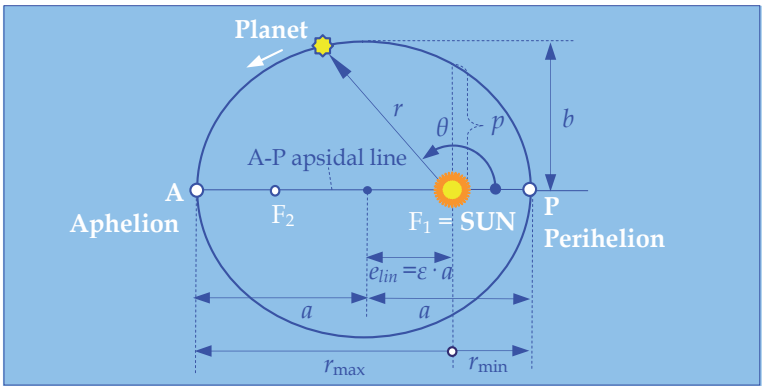


Fig. 10. The Orbit of a planet is an ellipse and its elements.

Before Kepler the Greek astronomer Ptolemy and many others after him were thinking that the Sun and planets travel in circles around the Earth. The ellipse can be mathematically expressed in the polar coordinate system by this equation:

$$r = \frac{p}{1 + \varepsilon \cos\theta}, \tag{3}$$

where (r, θ) are heliocentric polar coordinates for the orbit of planet (r – the distance between the Sun and a planet, θ – angle from the perihelion to the planet as seen from the Sun, respectively known as *the true anomaly*), p is the semi-latus rectum, and ε is the numerical eccentricity.
At $\theta = 0^\circ$ the minimum distance is equal to

$$r_{\min} = \frac{p}{1 + \varepsilon}. \tag{4}$$

At $\theta = 90^\circ$ the distance is equal p.
At $\theta = 180^\circ$ the maximum distance is

$$r_{\max} = \frac{p}{1 - \varepsilon}. \tag{5}$$

The semi-major axis is the arithmetic mean between $r_{\min}$ and $r_{\max}$:

$$a = \frac{r_{\max} + r_{\min}}{2} = \frac{p}{1 - \varepsilon^2}. \tag{6}$$

The semi-minor axis is the geometric mean between $r_{\min}$ and $r_{\max}$:

$$b = \sqrt{r_{\min} \cdot r_{\max}} = \frac{p}{\sqrt{1-\varepsilon^2}} = a\sqrt{1-\varepsilon^2} \ . \tag{7}$$

The semi-latus rectum p is equal to

$$p = \frac{b^2}{a} \ . \tag{8}$$

The area A of an ellipse is

$$A = \pi ab \ . \tag{9}$$

In the special case when $\varepsilon = 0$ then an ellipse turns into a circle where $r = p = r_{\min} = r_{\max} = a = b$ and $A = \pi r^2$.

Using ellipse-related equations Kepler's procedure for calculating heliocentric polar coordinates r, θ, for planetary position as a function of the time t from Perihelion, and the orbital period P, follows four steps:

1. Compute the *mean anomaly* M_a from the equation $M_a = \frac{2\pi t}{P}$. (10)
2. Compute the *eccentric anomaly* E by numerically solving Kepler's equation:

$$M_a = E - \varepsilon \sin E \ . \tag{11}$$

3. Compute the *true anomaly* θ by the equation $\tan\frac{\theta}{2} = \sqrt{\frac{1+\varepsilon}{1-\varepsilon}} \tan\frac{E}{2}$. (12)
4. Compute the *heliocentric distance* r from the equation $r = \frac{p}{1+\varepsilon\cos\theta}$. (13)

For the circle $\varepsilon = 0$ we have simple dependence $\theta = E = M_a$.

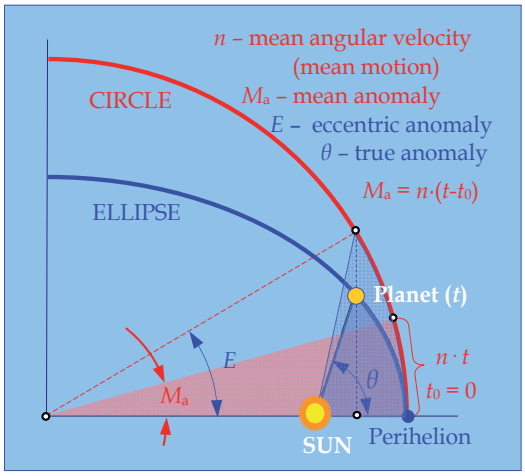


Fig. 11. Elements of parameters of a satellite orbit.

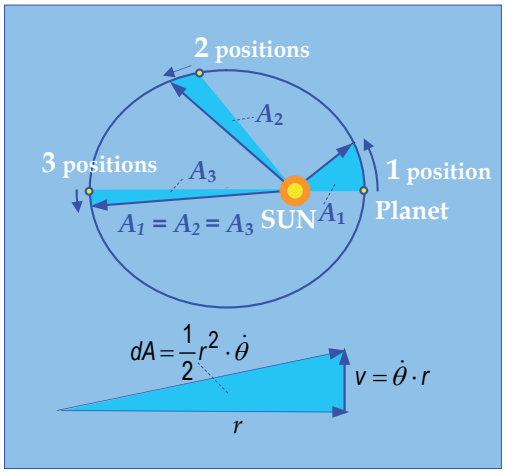


Fig. 12. The radius vector drawn from the Sun to a planet covers equal areas in equal times.

b) The Second Kepler's Law of Planetary Motion
This law can be expressed as follows:

The radius vector drawn from the Sun to a planet covers equal areas in equal times. (The Law of equal areas)

Mathematically this law can be expressed with the equation:

$$\frac{d}{dt}\left(\frac{1}{2}r^2\dot{\theta}\right)=0\,, \tag{14}$$

where $\dot{\theta}$ is angular velocity of true anomaly, $\frac{1}{2}r^2\dot{\theta}$ is the ''*areal velocity*'' that the radius vector *r* drawn from the Sun to the planet sweeps in one second (Fig. 12).

From this law it follows that the speed at which any planet moves through space is continuously by changing. A planet moves most quickly when it's closer to the Sun and more slowly when it is further from the Sun.

c) The Third Kepler's Law of Planetary Motion
This law can be expressed as follows:

The squares of the periodic times of the planets are proportional to the cubes of the semi-major axes of their orbits. (The harmonic law)

This law is giving the relationship between the distance of planets from the Sun and their orbital periods. Mathematically and symbolically it's possible to express as follows:

$$P^2 \propto a^3\,,$$

where *P* is the orbital period of circulate planet around the Sun and *a* is the semi-major axis of this orbit. Because this proportionality is the same for any planet which rotates around the Sun it's possible to write the next equation:

$$\frac{P^2_{planet1}}{a^3_{planet1}}=\frac{P^2_{planet2}}{a^3_{planet2}}\,,\quad \text{namely}\quad \frac{P^2_{planet1}}{P^2_{planet2}}=\frac{a^3_{planet1}}{a^3_{planet2}}\,. \tag{15}$$

3. The Physical Laws of Motions

Sir Isaac Newton's formulated three fundamental laws of the classical mechanics and the law of gravitation in his great work *Philosophieæ Naturalis (Principia Mathematica)* published on July 5, 1687. Before Isaac Newton the great contribution to the advance of mechanic was given by Galileo, Kepler and Huygens.

3.1 The First Law of Motion – Law of Inertia

This law can be expressed as follows:

Everybody persists in its state of being at rest or moving uniformly straight forward, except insofar as it is compelled to change its state by force impressed.

Newton's first law of motion is also called *the law of inertia*. It states that if the vector sum of all forces acting on an object is zero, then the acceleration of the object is zero and its velocity is constant. Consequently:

- An object that is at rest will stay at rest until a balanced force acts upon it.
- An object that is in motion will not change its velocity until a balanced force acts upon it.

If the resultant force acting on a particle is zero, the particle will remain at rest (if originally at rest) or will move with constant speed in a straight line (if originally in motion).

3.2 The Second Law of Motion – Law of Force

This law can be expressed as follows:

Force equals mass times acceleration.
If the resultant force acting on a particle is not zero, the particle will have acceleration proportional to the magnitude of the resultant and in the direction of this resultant force.

This law may be expressed by the equation:

$$\mathbf{F} = m\mathbf{a}\,, \tag{16}$$

where **F** is the vector of force, m the mass of particle and **a** is the vector of acceleration (Fig. 13. a).

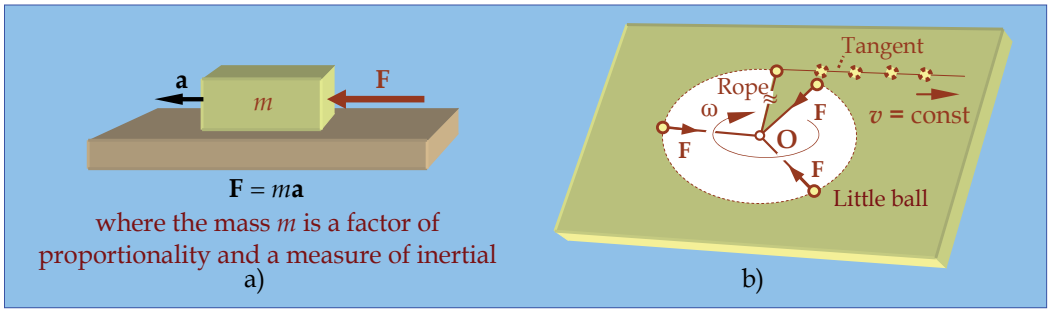


Fig. 13. a) An acceleration **a** of a free body on a horizontal plane under influence of a force **F**, b) In the rotation when the rope is broken, a little ball shall start moving with constant velocity along the line of tangent in the horizontal plane.

Really, this is differential equation which represents a basic equation of motion or basic equation of dynamic.
Alternatively this law can be expressed by the equation:

$$\mathbf{F} = \frac{d}{dt}(m\mathbf{v})\,, \tag{17}$$

where the product $m\mathbf{v}$ is the momentum: m - the mass of particle and $\mathbf{v}$ - the velocity. So, we can say:

The force is equal to the time derivative of the body's momentum.

3.3 The Third Law – Law of action and reaction

This law can be expressed in the following way:

To every action there is an equal by magnitude and opposite reaction (Fig. 14. a).

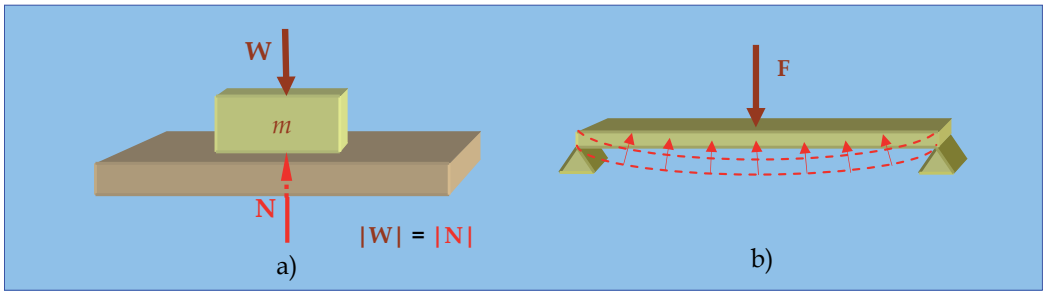


Fig. 14. a) Under the influence of the weight **W** of a body, a normal reaction of its support occurs, b) A beam under loading by a force **F** will be deformed as reaction to an active force **F**.

This law can be also expressed:

The forces of action and reaction between bodies in contact have the same magnitude, same line of action and opposite sense.

3.4 The Law of Gravitation – The Law of Universal Gravitation

Isaac Newton stated that two particles at the distance r from each other and, respectively, of mass M and m, attract each other with equal and opposite forces $\mathbf{F_1}$ and $\mathbf{F_2}$ directed along the line joining the particles (Fig. 15). The common magnitude F of these two forces is:

$$F = G\frac{Mm}{r^2}, \tag{18}$$

where G is the *universal constant of gravitation* $G \approx 6.67428 \times 10^{-11}$ $m^3kg^{-1}s^{-2}$),or approximately $3.44\times10^{-8}ft^4/lb\text{-}sec^4$ in British gravitational system of units (Beer & Johnston 1962).

The force of attraction exerted by the Earth on body of the mass m located on or near its surface is defined as the weight (W) of the body (Fig. 16)

$$W = mg, \tag{19}$$

where g is the acceleration of gravity, being also the acceleration of force of weight.

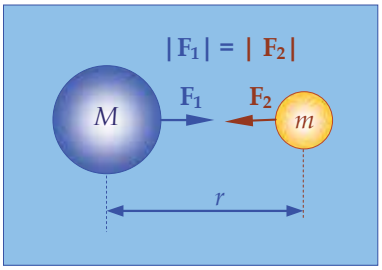


Fig. 15. Newton's Law of the universal gravitation.

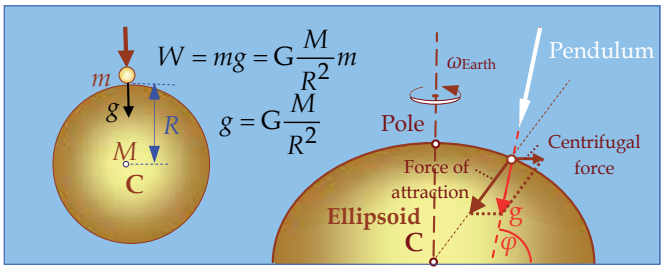


Fig. 16. The weight of a body on the surface of the Earth and the influence of centrifugal forces.

Because this force is really the force of universal gravitation it's possible to say

$$W = mg = G\frac{M}{R^2}m\,. \tag{20}$$

From this equation it follows that the acceleration of gravity is

$$g = \frac{GM}{R^2}\,. \tag{21}$$

The Earth is not truly spherical so the distance R from the centre of the Earth depends on the point selected on its surface. This will be the reason why the weight of the same body will also not be the same weight on different geographical latitude and altitude of the considered point. For more accurate definition of the weight of a body it's necessary to include a component representing the centrifugal force due to the rotation of the Earth. So, the values of g for a body in rest at the sea level vary from 9.780 m/s² (32.09 ft/s²) at the equator to 9.832 m/s² (32.26 ft/s²) at the poles.

3.5 D'Alembert's Principle

Jean le Rond d' Alembert (1717-1783) postulated the principle called by his name from the basic equation of dynamic:

$$\mathbf{F} = m\cdot\mathbf{a}\,. \tag{22}$$

This equation can be written in this form:

$$\mathbf{F} + (-m\cdot\mathbf{a}) = 0\,. \tag{23}$$

From this equation the magnitude $(-m\cdot\mathbf{a})$ is called *inertial force*. So, this equation represents an equation of fictive equilibrium where **F** represents resultant of all active and reactive forces, and inertial force which has the magnitude $m{\cdot}a$, but in opposite direction of the acceleration **a**. This equation is named *equation of dynamic equilibrium.*

For example, when a body is rotating in a circle with constant velocity v restrained by a rope length R then centrifugal force appears (Fig. 17).

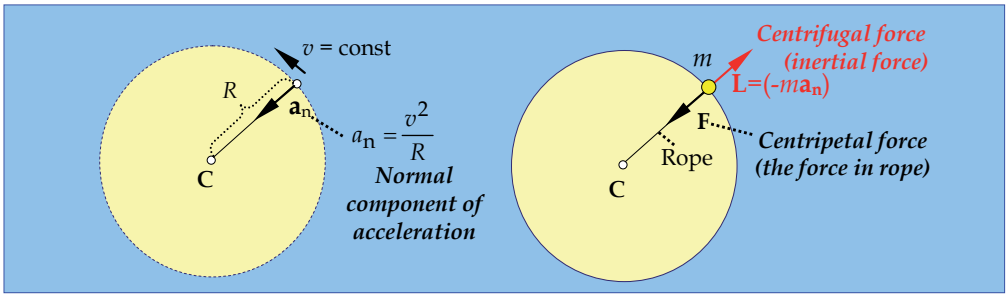


Fig. 17. Centrifugal force at rotation with constant magnitude of velocity v along a circle, and then centrifugal force **L** appears.

3.6 Potential due to a Spherical Shell

A basic result proved by Isaac Newton is that spherical shell which is homogeneous (with constant density) attracts an exterior point with mass m=1 as if all of the mass M of the spherical shell concentrated at its centre C (Fig. 18). This is the same, as if we have homogenous concentric layers but with different densities and whole masses M then an exterior mass point m attracts as if all of the mass M of the spheres was concentrated at its centre (Fig. 19).

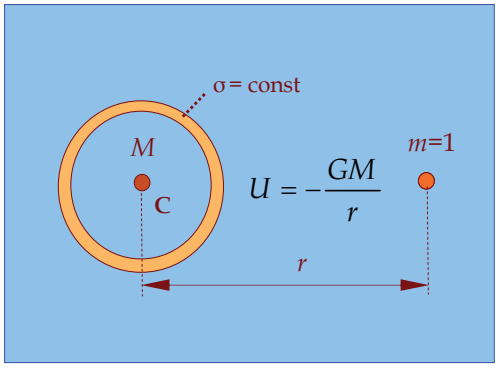


Fig. 18. The potential due to the solid spherical shell.

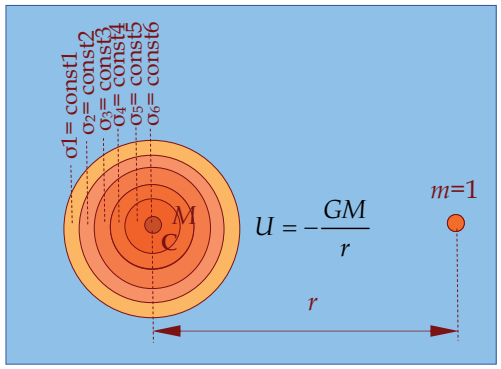


Fig. 19. The potential due to the concentric solid homogeneous spherical shells.

This fundamental result allows us to consider that the attraction between the Earth and the Sun, for example, to be equivalent to that between two mass points.
So we can say:

The solid sphere of constant density attracts an exterior unit mass though all of its masses were concentrated at the centre.

The potential, therefore, due to a spherical body homogeneous in concentric layers, for a point outside the sphere is

$$U = -\frac{GM}{r}, \tag{24}$$

where r is the distance from the point with mass m to the centre C of the mass of the homogeneous sphere or to the centre C of the concentric homogeneous spheres.

4. Determination of Orbits

Jacques Philippe Marie Binet (1786-1856) derived the differential equation in the polar coordinate system of the motion free material particle under action of the central force when areal velocity by the second Kepler's law is constant. This Binet's differential equation can be put down in writing

$$-mC^2u^2\left[\frac{d^2u}{d\theta^2}+u\right] = F_{\text{rad}}\ , \tag{25}$$

where is $u = \frac{1}{r}$, C - double areal velocity ($C = \dot{\theta}r^2$), F_{rad} gravitation force of the central body with mass M on the free particle with mass m ($F_{\text{rad}} = -GMmu^2$), where minus sign indicates that this is an attracting force, and plus sign stands for the repulsive force.
This differential equation is equation of free particle motion in a plane displayed in the polar coordinate system. Thus, inhomogeneous differential equation is obtained

$$\underbrace{\left[\frac{d^2u}{d\theta^2}+u\right]}_{\text{Homogeneous Part}} = \underbrace{\frac{GM}{C^2}}_{\text{Inhomogeneous Part}}\ . \tag{26}$$

The solution for the homogeneous part of this equation is

$$u_1 = B\cos(\theta-\theta_0)\ , \tag{27}$$

where B and θ_0 are the constants of integration. Choosing the polar axis so that $\theta_0 = 0$ we can write

$$u_1 = B\cos(\theta) \tag{28}$$

and for the inhomogeneous part of the equation

$$u_2 = \frac{GM}{C^2} = \frac{1}{p}\ . \tag{29}$$

The solution of this inhomogeneous differential equation (26) is

$$u = u_2 + u_1 = \frac{1}{p}\left[1 + \frac{C^2}{GM} B\cos(\theta)\right]. \tag{30}$$

The equation for the ellipse and for the other conic section in the polar coordinate system can be written in the form

$$\frac{1}{r} = \frac{1}{p}\left[1 + \varepsilon\cos\theta\right]. \tag{31}$$

After comparing the equations (30) and (31) it is possible to see that the equation (30) truly represents the equation of a conic section. The product of the constants B and $C^2/(GM)$ defines the *eccentricity* ε of the conic section. So it can be expressed by the equation (Beer & Johnston, 1962)

$$\varepsilon = \frac{B}{\frac{GM}{C^2}} = \frac{BC^2}{GM}. \tag{32}$$

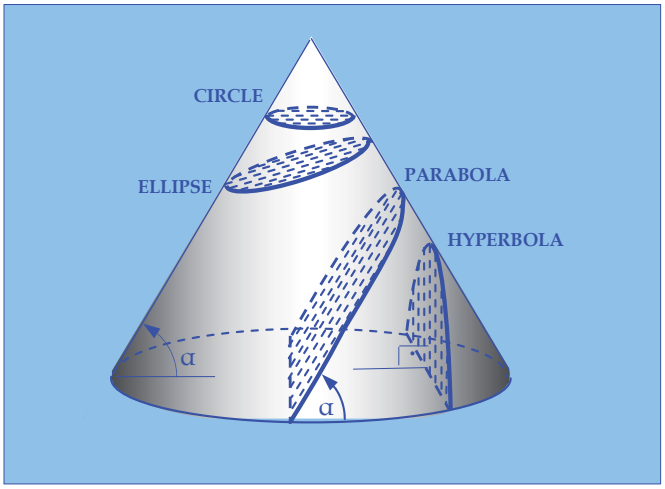


Fig. 20. The conic sections: circle, ellipse, parabola and hyperbola.

Four cases may be distinguished for different eccentricities (Fig. 20).

1) The conic section is a *circle* when is $\varepsilon = 0$.
2) The conic section is an *ellipse* when $0 < \varepsilon < 1$.
3) The conic section is a *parabola* when $\varepsilon = 1$.
4) The conic section is a *hyperbola* when $\varepsilon > 1$.

Of cause for the planets and for the satellites orbits can be only circulars or ellipses.

5. The Two-Body Problem

It is possible to investigate the motion of two bodies that are only under their mutual attraction. It can also be assumed that the bodies are symmetrical and homogeneous and that they can be considered to be point masses. So we can do analysis of the motion of planets and the Sun.

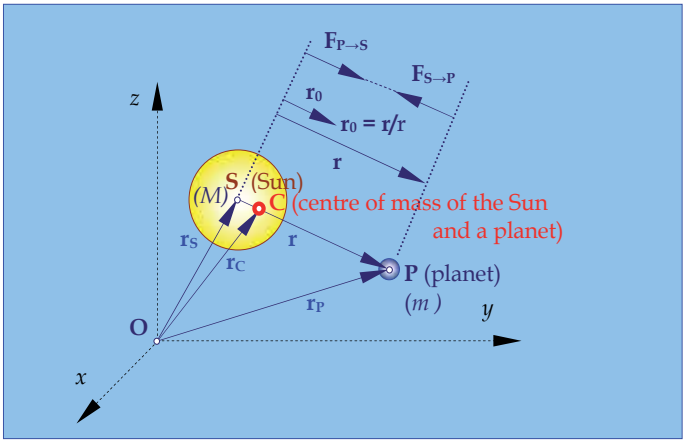


Fig. 21. Motion of the Sun and a planet in two-body problem.

The differential equation of the Sun motion (Fig. 21) is

$$M\frac{d^2\mathbf{r}_S}{dt^2} = +G\frac{Mm}{r^2}\frac{\mathbf{r}}{r}. \tag{33}$$

The sign + is because the force $\mathbf{F}_{P\to S}$ has the same orientation as the vector $\mathbf{r}_0$.
The differential equation of the planet motion is

$$m\frac{d^2\mathbf{r}_P}{dt^2} = -G\frac{Mm}{r^2}\frac{\mathbf{r}}{r}. \tag{34}$$

The sign – is because the force $\mathbf{F}_{S\to P}$ has the same orientation as the vector $\mathbf{r}_0$.
After summing up the equations (33) and (34) we can write

$$M\frac{d^2\mathbf{r}_S}{dt^2} + m\frac{d^2\mathbf{r}_P}{dt^2} = 0 \quad \text{or} \quad \frac{d^2}{dt^2}\left(M\mathbf{r}_S + m\mathbf{r}_P\right) = 0\,. \tag{35}$$

From the static it is known that the sum of the moment forces is equal to the moment of resultant. So we can say that the sum of the moment masses is equal to the moment of resultant mass. Now it is possible to write

$$M\mathbf{r}_S + m\mathbf{r}_P = (M+m)\mathbf{r}_C\,. \tag{36}$$

After the first and the second derivation we have

$$\frac{d}{dt}\left(M\mathbf{r}_S + m\mathbf{r}_P\right) = \frac{d}{dt}\left[(M+m)\mathbf{r}_C\right] \qquad \text{and}$$

$$\frac{d^2}{dt^2}\left(M\mathbf{r}_S + m\mathbf{r}_P\right) = \frac{d^2}{dt^2}\left[(M+m)\mathbf{r}_C\right] = \frac{d^2\mathbf{r}_C}{dt^2}(M+m). \tag{37}$$

From the equations (35) and (37) it follows

$$\frac{d^2\mathbf{r}_C}{dt^2}(M+m) = 0 \quad \text{or} \quad \frac{d^2\mathbf{r}_C}{dt^2} = 0\,. \tag{38}$$

Therefore the centre of this material system has no acceleration, namely this material system is in the inertial motion with the possibility to move with constant velocity $\mathbf{v}_C$, or remain at rest.

The equation (33) may be multiplied by m, and equation (34) by M and after subtracting the equation (33) from (34) we can write

$$Mm\left[\frac{d^2\mathbf{r}_P}{dt^2} - \frac{d^2\mathbf{r}_S}{dt^2}\right] = -G\frac{mM}{r^2}\frac{\mathbf{r}}{r}\left(M+m\right), \tag{39}$$

$$Mm\frac{d^2}{dt^2}\left[\mathbf{r}_P - \mathbf{r}_S\right] = -G\frac{mM}{r^2}\frac{\mathbf{r}}{r}\left(M+m\right) \tag{40}$$

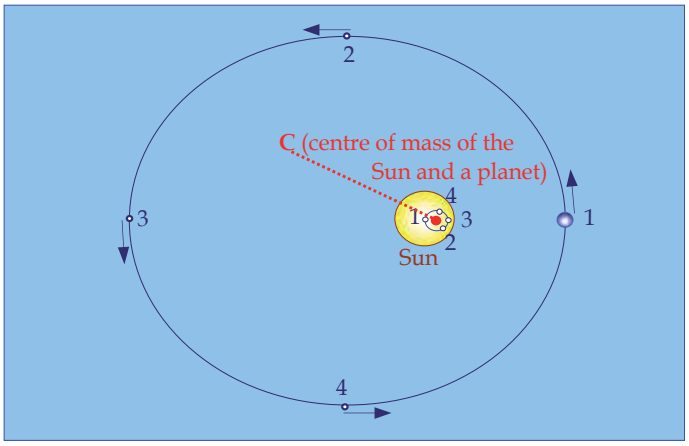


Fig. 22. The Sun is moving also around the centre C of mass of the Sun and a planet by a small ellipse and a planet is moving about the same centre C by the bigger ellipse, not around the geometrical centre of Sun.

After dividing the equation (40) by M and taking from Fig. 21 that $\mathbf{r}=\mathbf{r}_P-\mathbf{r}_S$ we can write

$$m\frac{d^2\mathbf{r}}{dt^2} = -G\frac{m(M+m)}{r^2}\frac{\mathbf{r}}{r}\,. \tag{41}$$

This is differential equation (41) of the planet motion when taken into account and the planet acting on the Sun. It is easy to prove that this planet is really rotating around the centre of mass (C) of the Sun and the planet. Also the Sun's geometrical center is rotating by the small ellipse around the centre of mass (C) (Fig. 22).
Hence:

The planet is rotating around the centre of masses (C) of the Sun and the planet by the bigger ellipse. The geometrical centre of the Sun also rotates around the centre of masses (C) by a small ellipse.

6. Satellite Motion

The problem, of two bodies is solved exactly in the celestial mechanics, but only in the special case if both bodies are having small dimensions, i.e. if the Sun and a planet can be thought of as particles. In this special case the motions of particles around the body with finite dimensions is also included, if this body with finite dimensions has the central spherical field of forces. (For example, as a homogeneous ball (Fig. 18) or concentric solid homogeneous spherical shells with different densities (Fig. 19)). Just because our Earth is not a ball and with homogeneous masses some discrepancies appear at satellite motions around the Earth from the exact solutions of two bodies when we imagine whole mass of the Earth as concentrated in it the centre of mass. For the solution of the problem of the motion of bodies (two particles) exactly valuable are three Kepler's laws from which fallow that the satellite would be moving constantly in the same plane by the ellipse with constant areal velocity.

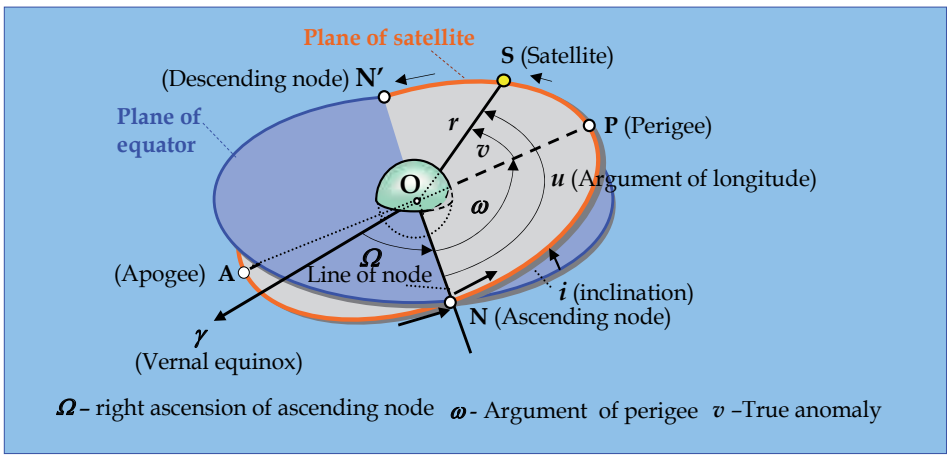


Fig. 23. Keplerian orbital parameters.

The positions of satellites are determined with six Keplerian orbital parameters: Ω, i, ω, a, e and v or t (Fig. 23):

- *The orientation of orbits* in space is determined by:

Ω - *the right ascension of ascending node* (the angle measured in the equator plane between the directions to the vernal equinox and ascending node N where the satellite crosses

equatorial plane from the south to the north celestial sphere), i – *the inclination of orbit*, (the angle between the equatorial plane and orbital plane) and ω – *the argument of perigee* (the angle between the ascending node and the direction to perigee (as the nearest point of satellite)).

- *The dimensions of orbit* are determined by: a - *the semi-major axis* and ε – *the numerical eccentricity* of an ellipse.
- *The position of satellite on its orbits* is determined by: v - *the true anomaly* (as the angle between the directions to perigee and instantaneous position of satellite) or by t - the difference of time in instantaneous position and the time in perigee.

All Kepler's laws and Newton's laws for a planet motion are valued also for the Earth's satellites motion but at satellites there are some more perturbations.

6.1 Required Velocity for a Satellite

A body will be a satellite in a circular orbit around the Earth if it has velocity in the horizontal line so that centrifugal force is equal to centripetal force which is produced by the Earth's gravitation attraction (Fig. 24). So it can be expressed with the equation:

$$m\frac{v_{\mathrm{I}}^2}{R+H_{\mathrm{S}}}=\mathrm{G}\frac{mM}{(R+H_S)^2}, \tag{42}$$

where: m - mass of a satellite, M – the Earth's mass, R - radius of the Earth, H_s - altitude of a satellite above the surface of the Earth, G constant of universal gravitation and v_{I} velocity of a body which will become the satellite.

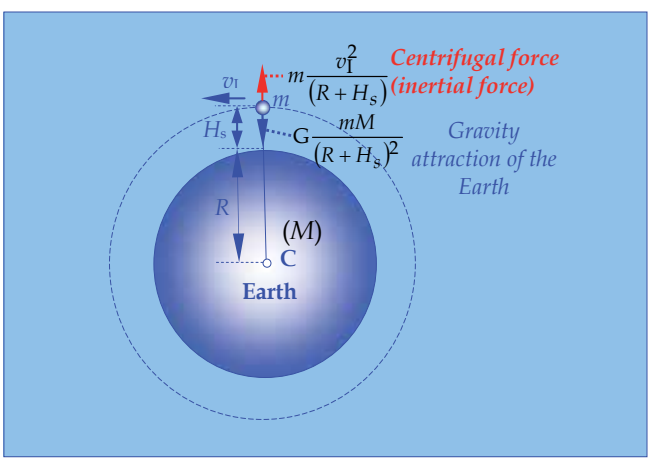


Fig. 24. On the satellite in orbit act the Earth's gravity attraction and the centrifugal force.

From this equation (42) next the equation follows

$$v_{\mathrm{I}}=\sqrt{\frac{\mathrm{G}M}{R+H_{\mathrm{S}}}}\,. \tag{43}$$

The equation (21) can be expressed as

$$GM = gR^2 \tag{44}$$

and after including this equation (44) into (43) we have

$$v_{\mathrm{I}} = \sqrt{\frac{gR^2}{R + H_{\mathrm{S}}}}\ . \tag{45}$$

This velocity v_{I} is sometimes called the *first cosmic velocity,* see Table 2.

Table 2. The first cosmic velocity depending on the heights over the Earth.

The heights over the Earth (km)	200	300	400	500	1000	3000	5000
Velocity v_{I} (m/s)	7791	7732	7675	7619	7356	6525	5924

6.2 Period of Circulation a Satellite around the Earth by a Circle

Since a satellite makes a circle $2\pi(R+H_{\mathrm{S}})$ long around the Earth in a time P (period) and because its velocity is constant then the equation (45) can be expressed as

$$2\pi(R+H_{\mathrm{s}}) = \sqrt{\frac{gR^2}{R+H_{\mathrm{s}}}}\ P. \tag{46}$$

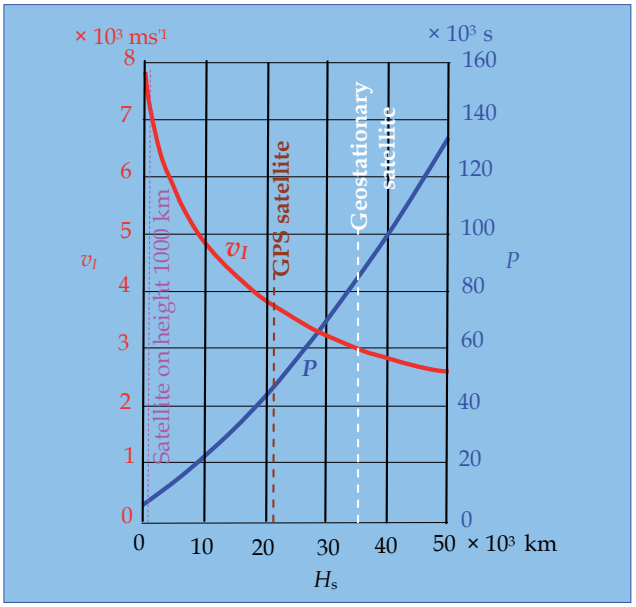


Fig. 25. Dependence the velocity v_{I} on the circular orbit and the period P of satellite height over the Earth's surface.

From this equation it follows

$$P = 2\pi\sqrt{\frac{(R+H_S)^3}{gR^2}}\,. \tag{47}$$

With the help of the equations (45) and (47) graphs on the Fig. 25 are constructed.

6.3 Escape Velocity

The escape velocity is minimum velocity of a body that a spacecraft must have in order to escape from the Earth's gravitation field or of other celestial bodies.

In order to launch a spacecraft into the space out of the Earth's gravitation field, it's necessary to do mechanical work for lifting a spacecraft from r=R to r=∞ (Fig. 26).

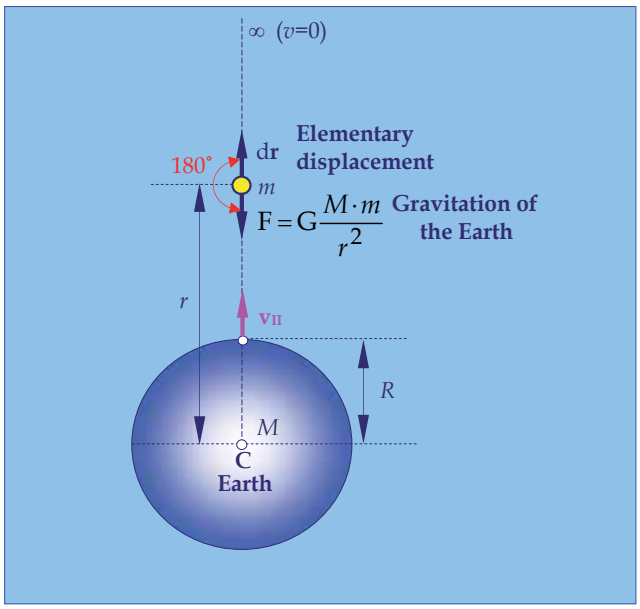


Fig. 26. Launching a spacecraft in the space out of Earth gravitation field (the second cosmic velocity).

The elementary work of the gravitation force **F** during the displacement $d\mathbf{r}$ is defined as the scalar quantity

$$dW = F \cdot dr \cos 180^\circ = -F \cdot dr\,. \tag{47}$$

For the work needed by the a spacecraft from R to the infinite ∞ we need to do the integration

$$W = -GMm\int_R^\infty \frac{dr}{r^2} = -GMm\frac{1}{R}\,. \tag{48}$$

The kinetic energy of a particle at the end minus the kinetic energy at the start is equal to the work, so we have

$$0-\frac{mv_{\mathrm{II}}^2}{2}=-GMm\frac{1}{R},\quad \text{namely}\quad v_{\mathrm{II}}=\sqrt{2GM\frac{1}{R}}\,. \tag{49}$$

Since is $GM = gR^2$, according to the equation (44), the equation (49) can be written as

$$v_{\mathrm{II}}=\sqrt{2gR}\,. \tag{50}$$

When a body has the velocity v_{II} then it will abandon gravitation field of the bigger body. This velocity v_{II} is called by the some people the *second cosmic velocity*.
From the equation (50) for the Earth it follows that the second cosmic velocity is 11.2 km/s (6.96 mi/second) and in this case spacecrafts will abandon gravitation field of the Earth.

6.4 The Potential of a Body with Irregular Disposition of Mass

The elementary potential of an attraction force of a parcel P_1 with mass dM in the point P (Fig. 27) is equal

$$dU=G\frac{dM}{\Delta}\,. \tag{51}$$

where: G – the universal gravitational constant and Δ the distance point P of P_1.

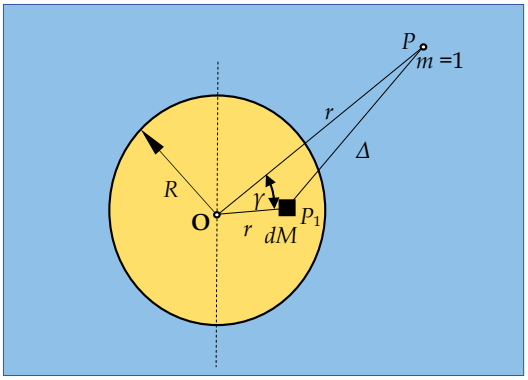


Fig. 27. The elementary potential of an attraction force of a body with arbitrary disposition of mass.

For a body with finite dimensions and with arbitrary disposition of mass it can be written in this form

$$U=G\int_M\frac{dM}{\Delta}\,. \tag{52}$$

This equation can be expressed by spherical functions in the form

$$U=G\frac{M}{r}\left[1+\sum_{n=2}^{\infty}\sum_{m=0}^{n}\left(\frac{R}{r}\right)^n P_{nm}\left(\sin\phi\right)\left(c_{nm}\cos m\lambda+s_{nm}\sin m\lambda\right)\right], \tag{53}$$

where the potential is evaluated at a point whose geocentric spherical coordinates are: r – the geocentric distance, φ – the geocentric latitude, λ – the longitude measured eastwards and R - the Earth's equatorial radius, c_{nm} and s_{nm} – the coefficients of the Earth's harmonics. The origin of the coordinate system is in the centre of Earth's mass, the geocentric longitude λ is calculated starting from Greenwich meridian and can have the values $0 \le \lambda \le 2\pi$, the geocentric latitude φ can have the values $(+\pi/2) \ge \varphi \ge (-\pi/2)$. In the equation (53) the expression P_{nm} (sin φ), is spherical function of sin φ which is determined by this expression

$$P_{nm}\left(\sin\phi\right) = \left(1-\sin^2\phi\right)^{m/2} \frac{d^m P_{n0}\left(\sin\phi\right)}{d\left(\sin\phi\right)^m}, \tag{54}$$

where is

$$P_{n0}\left(\sin\phi\right) = \frac{1}{2^n n!} \frac{d^n\left(\sin^2\phi-1\right)^n}{d\left(\sin\phi\right)^n}. \tag{55}$$

The functions:

- $P_{n0}(\sin\varphi)$ are the *zonal spherical function* or shorter *zonal harmonics* (Fig. 28. a),
- $P_{nn}(\sin\varphi)\cos n\lambda$ and $P_{nn}(\sin\varphi)\sin n\lambda$ are *sectorial spherical harmonics* or sectorial harmonics (Fig. 28. b) and
- $P_{nm}(\sin\varphi)\cos m\lambda$ and $P_{nm}(\sin\varphi)\sin m\lambda$ are *tesseral spherical harmonics* or tesseral harmonics (Fig. 28. c).

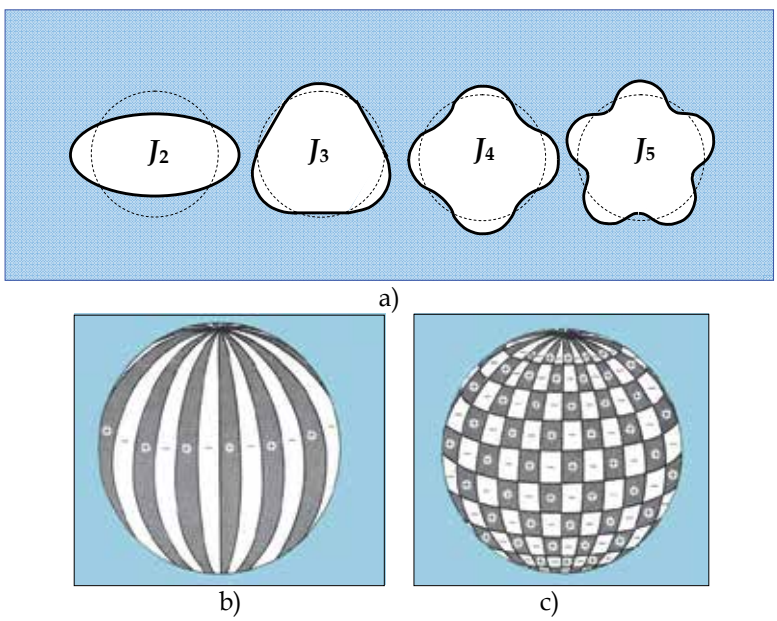


Fig. 28. Images of influences a) zonal hamonics, b) sectorial harmonics and c) tesseral harmonics.

If we suppose that the disposition of the mass in interior of the Earth is symmetrical with respect to the rotation axis than we obtain a somewhat simpler equation:

$$U = G\frac{M}{r}\left[1+\sum_{n=2}^{\infty} J_n\left(\frac{R}{r}\right)^n P_{n0}\left(\sin\phi\right)\right], \tag{56}$$

where J_n=-c_{n0}. Hence, than only zonal harmonics remain. The asymmetry of the north and of the south hemisphere is characterised by the coefficients of the odd zonal harmonics. If we suppose that the Earth is symmetrical in respect of the equator than will coefficients of the odd zonal harmonics disappear. So, it can be expressed with the equation

$$U = G\frac{M}{r}\left[1-J_2\left(\frac{R}{r}\right)^2 P_{20}\left(\sin\phi\right)-J_4\left(\frac{R}{r}\right)^4 P_{40}\left(\sin\phi\right)-....\right]. \tag{57}$$

6.5 Influence of the Earth's Gravitation on Movement of Artificial Satellites

Because the Earth is not a ball, i. e. not homogeneous ball, and also because it does not consist of concentric homogeneous spherical shells, the deviations of satellites orbits from an ellipse which corresponds to Kepler's laws will appear. Consequently to the second Newton's law can be written following the differential equations of satellites motion in rectangular coordinate system:

$$\ddot{x} = \frac{\partial U}{\partial x}, \qquad \ddot{y} = \frac{\partial U}{\partial y}, \qquad \ddot{z} = \frac{\partial U}{\partial z}. \tag{58}$$

Only the gravitation force of the Earth is taken into account but the other influences are neglected (the resistances of the air, the attraction of the Sun and the Moon, the radiation pressure of the Sun and so on). In this case the force of disturbance is

$$R^* = U - \frac{GM}{r} \tag{59}$$

and the equations (58) can be written in the form

$$\frac{d^2x}{dt^2}+\frac{\mu x}{r^3}=\frac{\partial R^*}{\partial x}, \qquad \frac{d^2y}{dt^2}+\frac{\mu y}{r^3}=\frac{\partial R^*}{\partial y}, \qquad \frac{d^2z}{dt^2}+\frac{\mu z}{r^3}=\frac{\partial R^*}{\partial z}, \tag{60}$$

where μ=G(1+M). The solution of these systems of differential equations and replacing the coordinates by the elliptical elements yields the parameters of orbit depending on time and the position of satellite.

Very complicated equations are obtained that can be found in the professional experts literature (Kozai 1959 and 1966), (McCuskey 1963). They are made of the starting or the mean magnitude of orbital parameters, their secular perturbations (which are linearly progressive in the time), the long-periodic and the short-periodic perturbations (Fig. 29) (Escobal, (1965).

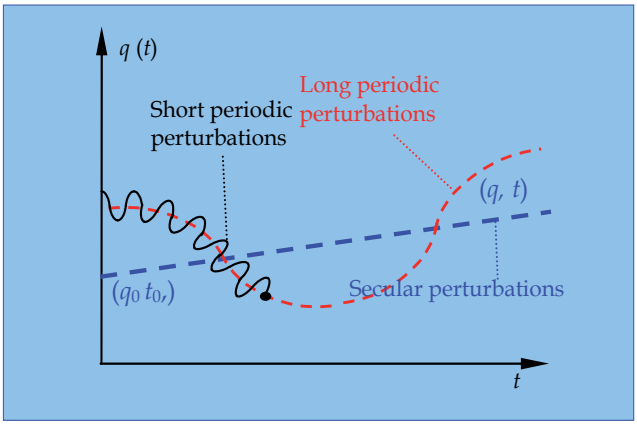


Fig. 29. The typical changes of orbital elements ω, Ω and M_0 by time.

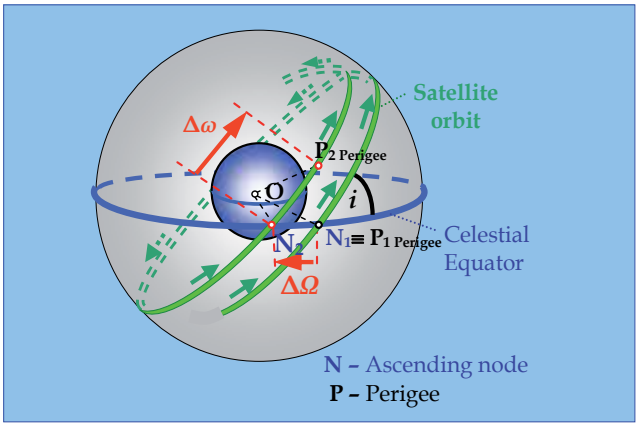


Fig. 30. Perturbations of the ascending node and perigee.

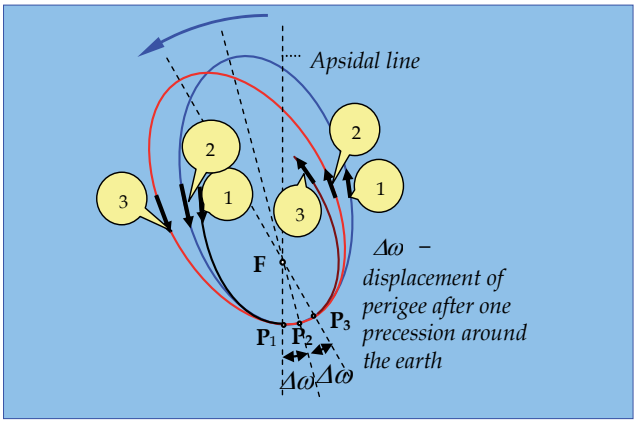


Fig. 31. Perturbations of the perigee P.

The parameter a has only short periodic perturbations, the parameters ε and i have long and short periodic perturbations and ω, Ω and M_a have secular, long and short periodic perturbations (Hofmann-Wellenhof et al. 1994) and (Kaula 1966).
From this solutions it follow that the ascending node will change position from the point N_1 in the point N_2 after one satellite precession around the Earth and so the longitude of node will dislocate for the angle $\Delta\Omega$ (Fig. 30). This phenomenon is named the precession. Also the point of perigee $N_1 \equiv P_1$ will change position in the point P_2 (Fig. 30 and Fig. 31) and thus, it will arise the argument of perigee for the angle $\Delta\omega$. Also the apsidal line will rotate.

The secular perturbation of node longitude and the argument of perigee can be determined from approximate equations

$$\dot{\Omega} = -\frac{3}{2} n a_E^2 \frac{\cos i}{a^2 \left(1-\varepsilon^2\right)^2} J_2 + \ldots\ldots , \tag{61}$$

$$\dot{\omega} = \frac{3}{4} n a_E^2 \frac{5\cos^2 i - 1}{a^2 \left(1-\varepsilon^2\right)^2} J_2 + \ldots\ldots , \tag{62}$$

where are: $\dot{\Omega}$ and $\dot{\omega}$ are the temporal changes of angles in the degrees per day, n – the mean angular satellite velocity ($n = 2\pi / P = \sqrt{GM / a^3}$, G – the universal constant of Gravitation, M – the Earth's mass), a_E – the semi-major axis of the Earth, a – the semi-major axis of satellite orbit, ε – the numerical eccentricity of elliptic orbit, i – the inclination of orbital plane and J_2 – the zonal coefficient of Earth's harmonics. This coefficient is about 500 and more time bigger than other coefficients and characterizes the flatness on poles and the bulginess on equator, which is caused by the Earth's rotation.
From the equation (61) it is possible to make the decision that $\dot{\Omega}$ =0°/day for i = 90˚ when satellites are in polar orbits. Then the ascending node will not move. From the eqution (61) it also thus follows that $\dot{\Omega}$ has maximal value for i = 0˚ when the satellite rotates in the equatorial plane.
From the equation (62) it is possible to see that $\dot{\omega}$ =0°/day for $\cos^2 i$ = 0.2 and thus, it is obtained that the inclination of orbit is i = 63.4˚ or i = 166.6˚. The maximal value of $\dot{\omega}$ is when is i = 90˚ that is when the satellite rotates in the polar orbit.
In the equations (61) and (62) there is ratio a_E^2/a^2 which is in every case less than 1. The satellites with bigger distance from the Earth's centre have the lower value of ratio a_E^2/a^2 and then the dislocations of nodes and apsidal lines will be much smaller when the satellites are farther from the Earth.
The zonal coefficient J_2 has the largest influence on satellites motion. The other coefficients have considerably smaller influence but in spite of this fact the artificial satellites made possible to determine the true shape of the Earth that is the geoid (Fig. 32). For this purpose usually the low satellites were used (altitudes about 1000 km), but before calculation we are obliged to remove the other influences which is a very complicated mode of proceedings. In the recent time the satellite GOCE was launched for the determination Earth's gravitation field with the very high accuracy.

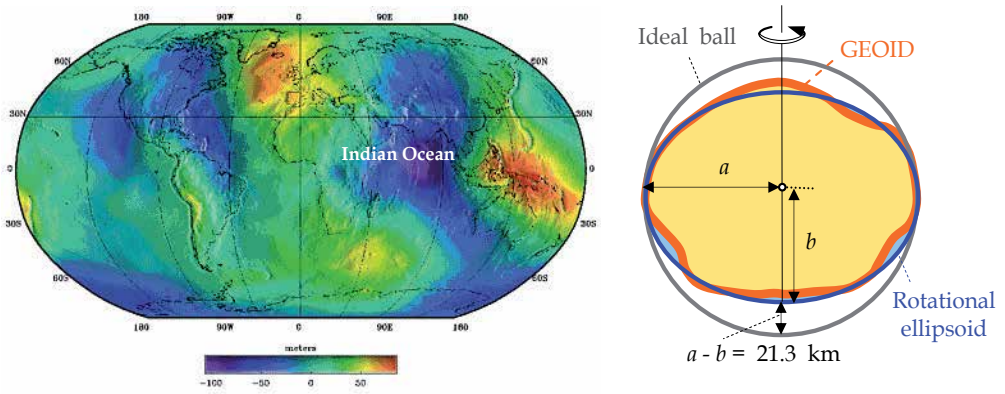


Fig. 32 The geoid is the true shape of the Earth – the maximal elevation over the reference ellipsoid is 85.4 m and the largest hollow 107.0 m in Indian Ocean (Geoid 050 NASA).

6.6 Influence of the Atmosphere on the Satellite Motion

The influence of atmosphere on the satellite motions is bigger if the distance of satellite from the Earth's surface is smaller. This effect is the biggest immediately after the effects of Earth's flatness for the low Earth satellites. The force of satellite retardation (F_0) can be determined from the equation

$$F_0 = -0.5 C_D \rho v_{rel}^2 \frac{A}{m}, \tag{63}$$

where: C_D – the aerodynamic coefficient, ρ – the density of atmosphere, v_{rel} – the relative velocity of a satellite in respect to the atmosphere, A – the area of satellite cross section and m – the mass of a satellite.

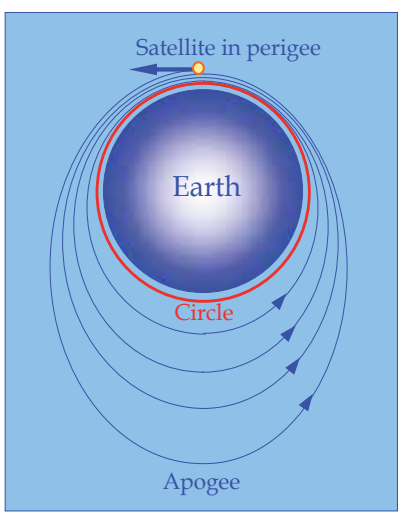


Fig. 33. Image of influence the resistance

The density of atmosphere is changeable and depends of altitudes but bigger changes caused by the activity of the Sun (phenomenon of the Sun spot). Because the resistance of the air depends on its density which is changeable and depends on more factors, this influence can be taken in account with more limited accuracy. The eccentricity of elliptic orbits in low orbits is progressively smaller (Fig. 33).
From the parameters of satellites motions in the low Earth orbit the effect of air resistance can be removed only with limited accuracy.

6.7 The Influence of Gravitation of Moon and Sun

The gravitation influence of the Moon and the Sun can be eliminated by equations which are obtained solving the differential equations of satellite motion in which the functions of the Moon and the Sun attractions forces are inserted.
Secular motions due to the lunar and solar attractions are denoted by $\Delta\omega_{\mathrm{LS}}$ and $\Delta\Omega_{\mathrm{LS}}$ are expressed as,

$$\Delta\omega_{\mathrm{LS}} = \frac{1}{2}\psi\left(-1+5\theta_i^2+\varepsilon^2\right) \quad \text{and} \quad \Delta\Omega_{\mathrm{LS}} = \Psi\left((1+\frac{1}{2}\varepsilon^2\right), \tag{64}$$

where: ε is the numerical eccentricity of the satellite orbit at the beginning for $t = 0$,
$\theta_i = \cos(i)$ for inclination of the satellite orbit i at beginning for $t = 0$,

$$\Psi = \frac{3}{4}\frac{1}{n^2\sqrt{1-\varepsilon^2}}\left[n_{\Theta}^2\left(1-\frac{3}{2}\sin^2\zeta\right)+n_{\Diamond}^2 m_{\Diamond}\left\{1-\frac{3}{4}\sin^2 J\left(1+\cos^2\zeta\right)-\frac{3}{2}\sin^2\zeta\cos^2 J\right\}\right]. \tag{65}$$

In this expression n_{Θ} and $n_{\Diamond}$ are mean motions of the Sun and the Moon, respectively, $m_{\Diamond}$ is the mass of the Moon expressed in units of the mass of the Earth, ζ is then obliquity, that is the inclination between the ecliptic and the equator, and J is the inclination of the lunar orbit with reference to the ecliptic and is about 5˚. When n is given in degrees per day, Ψ, the dimensionless quantity, takes the following value (Kozai 1959 and Kozai 1966):

$$\Psi = \frac{1.762}{n^2\sqrt{1-\varepsilon^2}}\,. \tag{66}$$

6.8 Tidal Effects

The Sun and the Moon deform the Earth by their tidal forces not only in the ocean but also in the body (the land). In this manner the Earth's shape is temporally changeable and the tidal effects are apparent. Merson, R. H.; Kaula, W. M.; Newton, R. R. and Kozai, Y. are the first who examined the tidal effects on the motion of polar Earth's satellites and they found out that there are small changes in inclinations of the polar satellites orbits but this results were not consistent. Today the instrument for determining the positions of satellites are more precise and this effect needs to be taken into account for better determination parameters of satellite's orbit (Hofmann-Wellenhof et al. 1994).

6.9 Solar Radiation Pressure (The Pressure of Light)

In the first time when the first satellites were launched in the Earth's orbit solar radiation (the pressure of light) was neglected. The motions examinations of the satellite "Vangard 1" indicated that there are some discrepancies. These perturbations were possible to explain by the pressure of light. The pressure of light on the Earth's surface is very tiny and really practically zero, but in the orbit this effect can change position of satellite especially if the mass of satellite is small and its effective cross section is very big. So this effect could be studied better at the satellite EHO 1 which really was a weightless balloon with small mass and big cross section. The module of acceleration ''a_a'' of a satellite under influence of the solar radiation pressure can be determined by the equation

$$a_a = P_r \frac{A}{m}. \tag{67}$$

where: P_r - the force of light pressure on the unit of area, A – the area of satellite cross section and m – the mass of satellite.

It should be noted here that the satellites enter into shadow of the Earth and so the problems with the pressure of light are more complex, although small.
The solar radiation pressure which is reflected back from the Earth's surface causes an effect called albedo. The pressure of reflected light also affects on the satellites but this effect can be for the most number of cases practically neglected. From the perturbations of Earth's satellite motion it was determined in the paper (Sehnal 1978) which parts of the Earth's surface give bigger or smaller albedo.
The solar radiation pressure which is reflected to the Earth's satellite from the Moon also affects on the Earth's satellite but this effect is smaller than this of Earth's albedo.

6.10 The other influences

▪ The low satellites have very big velocities something more of 7 km/s but this velocities are relatively small comparing to the velocity of light (300,000 km/s). Meanwhile, this *relativistic effect* needs also to be taken in the accurate calculation of parameters of the satellite orbits. This effect can cause the perturbing acceleration results in an order of $3 \cdot 10^{-10}$ ms^{-2}.
▪ The influences of planets gravitations can be neglected because the perturbation of planets is really too small.
▪ On the satellites there are different electronic equipments which have proper magnetic and electric fields. Since the satellites are in the motion in the Earth's the magnetic field the perturbation forces appear which are provoked by the acting of this field. Such effects are very small and poorly examined.

6.11 Sorts of Satellite Orbits

Satellites are launched into orbits, which is to say that they are shot up into the sky on rockets to get them up above the atmosphere where there is no friction. The idea is to get them flying so fast, that when they fall back to the Earth, they fall towards the Earth at the same rate the Earth's surface falls away from them. When an object's path around the Earth ''trajectory'' matches the Earth curvature, the object is said to be ''in orbit''.

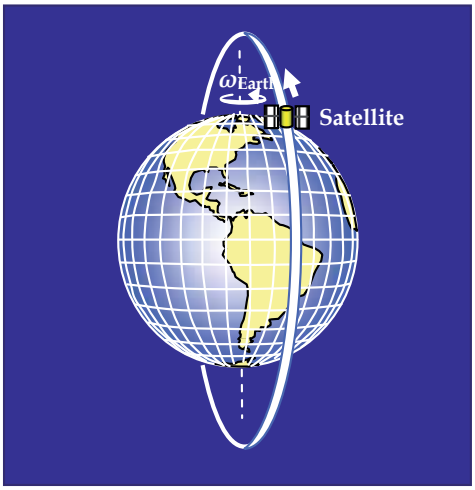


Fig. 34 Polar orbit of a satellite.

The orbits are defined by 3 factors:
▪ The first is the ***shape of the orbit,*** which can be
- *circular* (with equal distance from the Earth's centre) or
- *elliptic* (an oval shaped orbit with the *perigee* as the closet point to the Earth and the *apogee* as the farthest point from the Earth).

▪ The second is the ***altitude of the orbit.*** The altitude is constant for a circular orbit but changes constantly for an elliptical orbit.
▪ The third factor is the ***angle of the inclination,*** the angle which plane of satellite orbit makes with the plane of equator. An orbit when a satellite is passing over the Earth's poles or close to it has a large angle of inclination (i = 90˚ or near 90˚). These satellites are called the *polar satellites* (Fig. 34). An orbit that makes the satellite stay close to the equator has a small angle of inclination (i = 0˚ or near 0˚). These satellites are called the *equatorial satellites* (Fig. 35).

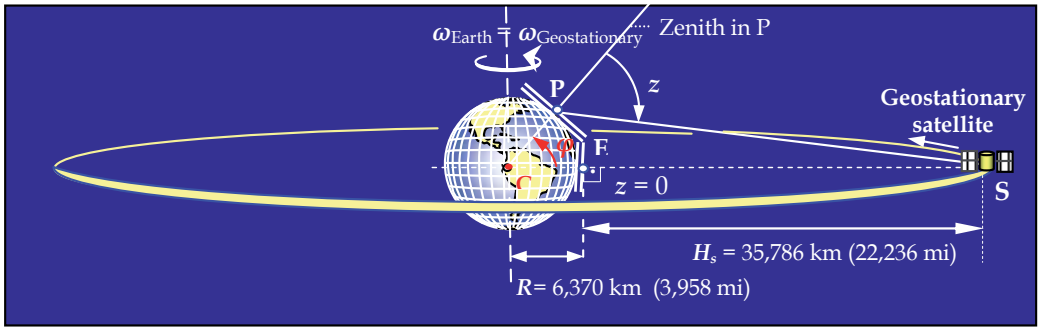


Fig. 35. Imagery of a geostationary satellite in the equatorial orbit.

□ ***Low Earth Orbits*** **(LEO)**
Satellites in low Earth orbit (LEO) orbit around the Earth at altitudes of less than 2000 km (1240 miles). Communication satellites in LEO can give clearer surveillance images and require much less power to transmit their data to the Earth.

For example by the *Iridium satellite network* voice and data messages can be routed anywhere in the world. This Iridium satellite constellation was originally planned to have 77 active satellites and was redesigned to on the 66 active satellites of day. These satellites are in six polar low Earth's orbital plane at a height of approximately 781 km (485 miles) and have inclinations i = 86,4°. Each satellite completely orbits the Earth in 100 minutes travelling. From the horizon to the horizon it takes 10 minutes. As satellites move out of view from the satellite phone user the call is handed over the next satellite coming into view (URL 4) and (URL 5).
The ***apparent trajectory*** for LEO satellites in their circular orbits or elliptic orbits for a motionless spectator on the Earth looks like as a straight line or a mild curved line.

□ ***Medium Earth Orbits*** **(MEO)**
Satellites in medium Earth orbit (LEO) circling around the Earth are at altitudes of about 2000 km (1240 miles) to just below geosynchronous orbit at 35,786 km (22,240 miles). These kinds of orbits are known as *intermediate circular orbits*.

□ ***Geosynchronous satellites***
The geosynchronous satellites can be: the geostationary satellites and the 24-hours Earth synchronous satellites.

- ***Geostationary satellites*** **(GEO)**
A satellite in geostationary orbit circles the Earth in exactly 1 day and is placed above the equator. The geostationary satellites have angles of inclination of orbits equal 0° and they are at the elevation of approximately 35,786 km (22,236 miles) above the ground of Earth (Fig. 35). These satellites have the same directions of rotation as the Earth and so follow the Earth's rotation. For one full orbit they need 23 hours, 56 minutes, 4.091 seconds (one sidereal day). As a result the satellite seems to the man who is standing on the Earth that the satellite (GEO) is all the time in the same position without motion.

The geostationary satellites are usually launched as the first in an elliptical orbit (Fig. 36) and after when the satellite is in the apogee of this orbit it obtains velocity necessary for maintaining the height over the Earth.

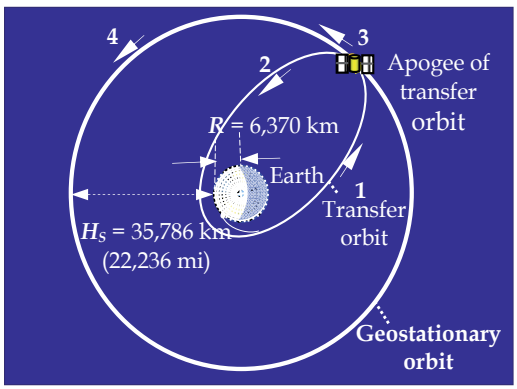


Fig. 36. Launching a satellite in the geostationary orbit with the help of an elliptic transfer orbit (Bazjanac 1977).

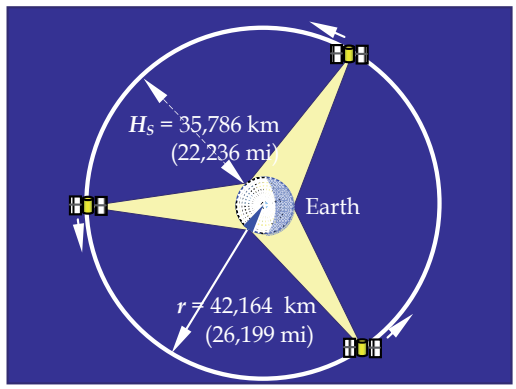


Fig. 37. All the Earth's surface is covered with only three geostationary satellites.

These satellites are usually used for communications, satellite TV, meteorology and remote sensing. With only three geostationary satellites it is practically possible to cover all the Earth's surface (Fig. 37) from the equator to the geographical latitudes +81.3° and – 81.3°.
A geostationary satellite can be immovable in the zenith only in a point on the equator, but in the point on the Earth with higher latitudes it is not possible to see immovable satellites in the zenith. This means that a satellite cannot be geostationary in the zenith at any point on the Earth with higher geographic latitudes. The planes of satellites need to pass through the centre of Earth's mass and this is the reason why the satellites cannot orbit around the Earth in the plane of parallel and to be all the time immovable in the zenith of this point on the parallel.

- ***Geosynchronous 24 Hour Earth Satellites***

These kinds of satellites have the same orbital period as the geostationary satellites but their orbital plane is not placed in the equator plane, namely they have an inclination greater than 0°. This orbit of 24-hour Earth satellite in reality may be a circular or an elliptic.

a) Apparent Motion of 24-Hour Earth Satellites in Circular Orbits

The 24-hour satellites in the circular orbits and under certain inclination angles i in relation to equatorial plane can be described as a *number eight* on the sky for the "standing spectator" on the Earth. This apparent motion of 24-hour satellites is caused by combining rotation of the Earth around its axis and rotation of the satellite around the Earth around the axis inclined under the angle i in relation to the equatorial plane.
This occurrence can be best explained by Fig. 38. In the initial moment $t = 0$ let the 24-hour satellite be in the equatorial plane. After the time t, the Earth will turn around its rotation axis for the angle β_{Earth}. At the same time t 24-hour satellites will rotate around their orbit for the equal angle β_{Sa}. Namely, angle speed of the Earth ω_{Earth} and angle speed of 24-hour satellite ω_{Sa} are equal, so the banking angles of longitude plane and of 24-hour satellite in its plane will be equal, that is $\beta_{\text{Earth}} = \beta_{\text{Sa}}$.
However, these inclination angles β_{Earth} and β_{Sa} do not lie in the same plane, but their planes close between themselves an inclination angle i of the 24-hour satellite orbit. This is why the projection of banking angle of the satellite in the plane of the satellite's orbit β_{Sa} will be projected in the equatorial plane in the S′ spot, e.g. it will be shorter from β_{Earth}. This means that a spectator on the Earth will at first see 24-hour satellite lapses on the west side. It will remain this way until the plane of the Earth's longitude does not turn for 90°, when the 24-hour satellite appears in that displaced longitude's plane.
In order to mathematically determine this apparent motion of 24-hour satellites, one can observe the spherical triangle E_1SS' and calculate the size of the projection of the spherical side angle β_{Sa} in the equatorial plane, which has been marked with letter S′ on the Fig. 38. In this way we get:

$$s = \arctan\left[\tan(\beta_{\text{Sa}}) \cdot \cos(i)\right]. \tag{68}$$

So the lagging angle of the 24-hour satellite in the equatorial plane is:

$$\Delta\beta = \beta_{\text{Earth}} - s. \tag{69}$$

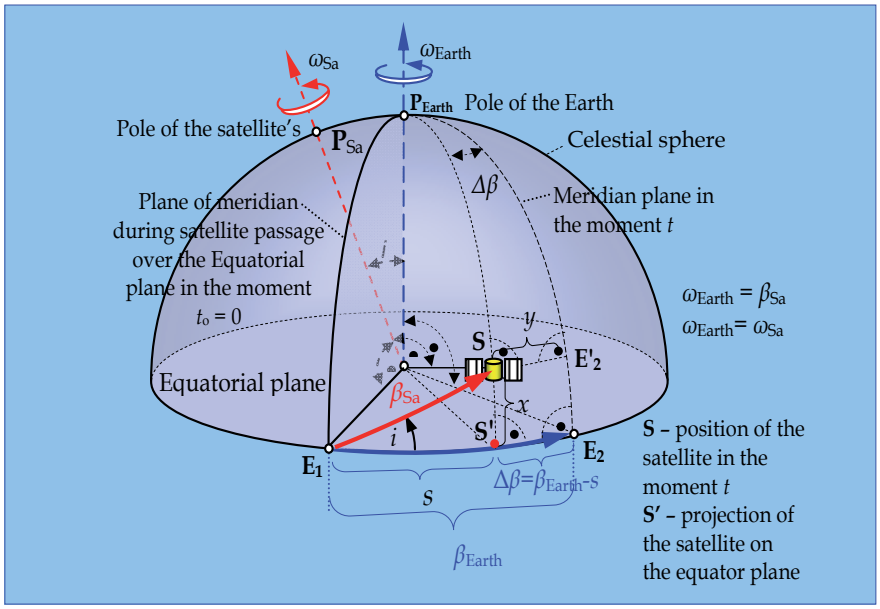


Fig. 38. Graphic illustration of why the spectator on the Earth sees relatively (apparent) motion of 24-hour satellites in circular orbit in relation to ''still'' Earth.

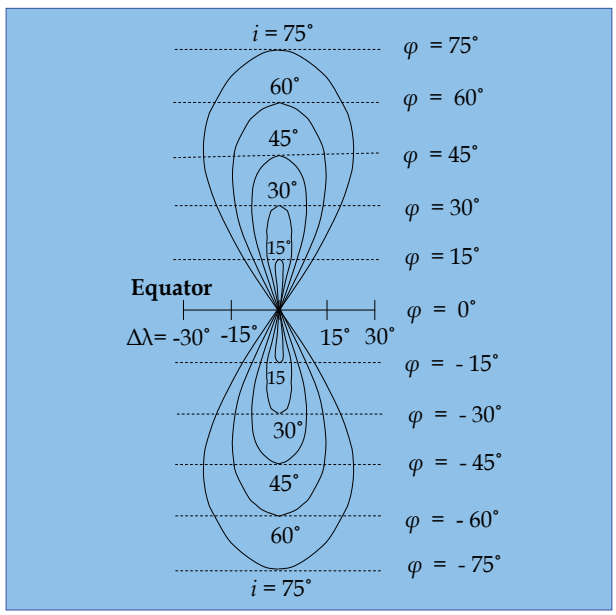


Fig. 39. Graphical illustration of apparent orbits of relative motion of 24-hour satellites on their circular orbit for different inclinations of their orbits, as a spectator on the Earth sees them.

Now it can be calculated, from the very same spherical triangle E_1SS' the value of angle distance x from the equatorial plane and the plane the satellite passes through. In this way, we get:

$$x = \arctan\left[\sin(s) \cdot \tan(i)\right]. \tag{70}$$

The coordinate y expressed in degrees can be calculated from the spherical triangle $P_{Earth}SE'_2$. We get:

$$y = \arctan\left[\cos(x) \cdot \tan(\Delta\beta)\right]. \tag{71}$$

For different inclinations i of the satellite's orbit, with the help of equations (68), (69), (70) and (71), angle values of x and y were calculated for drawing of graphical display of orbits of apparent (relative) motion of 24-hour satellites, as it is seen on the Earth, which is shown in Fig. 39 (Solarić, 2007).

From the chart in the Fig. 39, it is obvious that 24-hour satellite in its apparent relative motion reaches latitudes that are equal to their orbital inclinations. Besides, these *symmetrical eights* become greater when the inclination i of the satellite's orbit are greater.

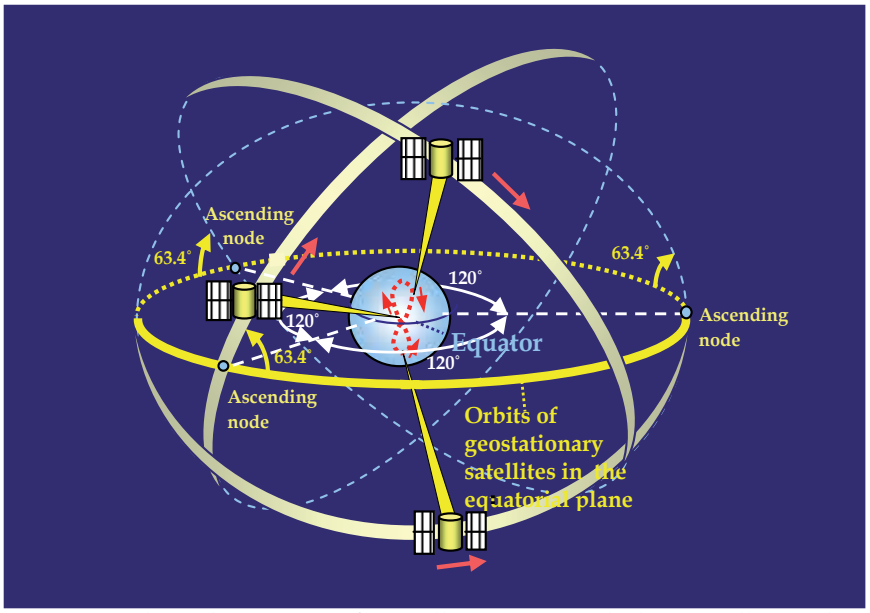


Fig. 40. Imaging in three dimensions of a satellite system with three satellite in three orbital planes.

b) Apparent Motion of 24-Hours Earth Satellites in Elliptic Orbits

The geosynchronous satellites on elliptical orbits with a high inclination (usually near 63.4°) and an orbital period of one sidereal day (23 hours, 56 minutes, 4.091 seconds) called *Tundra* (URL 7). A satellite placed in this orbit spends most of its time over a chosen area of the Earth.

From the second Kepler's law that satellite's velocity is lowest in the apogee, the greatest in the perigee, while the velocity of the satellite in any other point of elliptical orbit is between

these values, that is: $v_{perigee} > v > v_{apogee}$. So the satellites in the apogee point of an elliptic orbits spend the most time, and because they have usually the critical inclinations $i = 63.4°$ they have not the secular perturbation of the argument of perigee caused by the Earth's equatorial bulge.

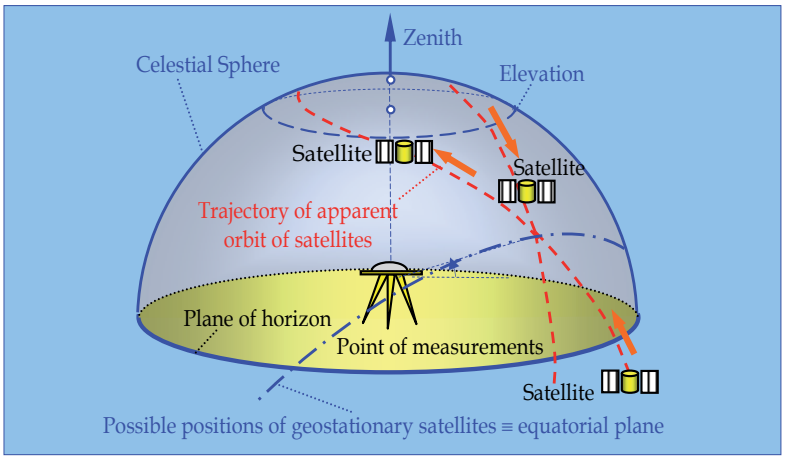


Fig. 41. View of celestial sphere on the point on the north hemisphere with apparent (relative) orbits a satellite system with the 24-hour satellites in three different orbits, as the standing spectator sees them. The possible positions of geostationary satellites have been represented as well.

The apparent ground track of a satellite in *a tundra orbit* is a closed ''*asymmetrical– eight*'' lemniscate's.

The Sirius Satellite Radio (the USA) uses the tundra orbits in three orbital planes with the inclination $i = 63.4°$ and the eccentricity $\varepsilon = 0.2684$ so that their ascending nodes in the equatorial plane are moved on the longitude for 120°. Also the difference of mean anomaly of each satellite is 120°. When one satellites moves out of position, another has passed perigee and is ready to take functions over the territory of the USA and the Canada. Every satellite of these satellites is every day 16 hours over the north hemisphere and 8 hours over the south hemisphere. (Simile as at Fig. 40.)

In order to numerically determine apparent trajectory of tundra orbits with elliptical orbit, as a "standing" spectator sees it on the Earth, one should:

- project position of a satellite from the elliptical orbit to sphere,
- out of these coordinates from the sphere, according to equations (68) – (71), determine coordinates x and y the point of tundra orbits, expressed in degrees,
- graphically demonstrate sequential positions of satellite's apparent orbit.

Japan decided to set up their satellite communication and positioning system called ''*Quasi Zenith Satellite system*'' (abbreviated QZSS) or in Japanese ''Jun-Ten-Cho''. As expected this satellite system will up to schedule operate in 2013 (URL 8).

It is being anticipated that 3 of these 24-hour QZSS satellites would be substituting every 8 hours on the near zenith above Japan. In this way, there will always be one satellite approximately in zenith, and another QZSS satellite would be on somewhat greater zenith distance. The third QZSS satellite would then be somewhere on the south, so it can be for

shorter time under the horizon in Tokyo. This could be done in the way that every QZSS satellite has its own orbital plane and that their ascending nodes in the equatorial plane are moved on the longitude for 120° (Fig. 42). The Japanese called this satellite system quasi-zenith because there will always be one QZSS satellite near the zenith.

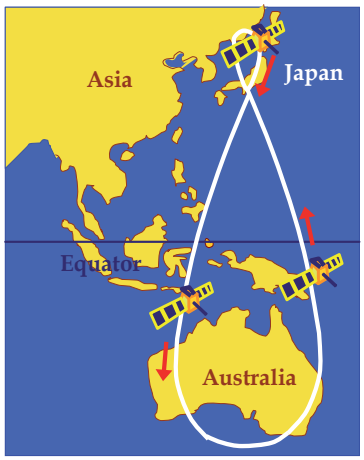


Fig. 42. Apparent orbits of three planned 24-hour QZSS-satellites with inclination i = 45° and eccentricity e = 0.099.

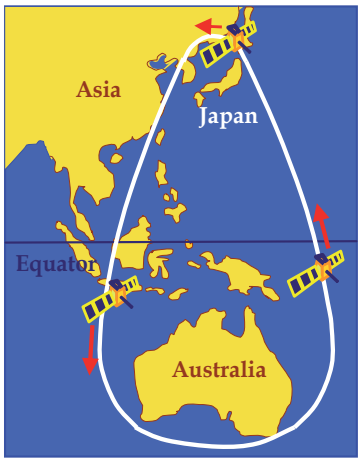


Fig. 43. Apparent orbits of three planned 24-hour QZSS-satellites with inclination i = 42.5° and eccentricity e = 0.21.

In the article (Petrovski et al. 2003), experiments for possible usage of 24-hour satellites' orbits with different eccentricities ε and for different inclinations of orbits' planes i, have been conducted.

Orbit of 24-hour satellite will have different form for different values of eccentricities of apparent motion:

a) For the inclination of orbital plane i = 45° and eccentricity ε = 0.099 apparent orbit of 24-hour satellite will have the shape of *asymmetrical eight* (Fig. 42), while they are symmetrical with the circular orbits of 24-hour satellites.
b) For the grater eccentricity of elliptical orbit, apparent orbit of 24-hour satellite motion will no longer have the shape of an asymmetrical eight, but it will assume the shape of *rain drops*. It can be seen in Fig. 43. for the eccentricity of elliptical orbits of 24-hour satellites ε = 0.21 and orbital inclination i = 45°. Apparent motion orbit of QZSS satellite will expand even more for somewhat greater eccentricity ε. This kind of orbit of a QZSS satellite would be appropriate for an area wider than Japan, so, for this case, possibilities with 3 and 4 QZSS satellites have been tested.
Clear spatial image could also be constructed for this case, when QZSS has three QZSS satellites in elliptical orbits, as it is demonstrated in Fig. 40, for circular orbits of QZSS satellite.

□ ***Semi - Synchronous Orbit***
The satellite in the semi-synchronous orbit has the period of the 0.5 sidereal day (about 11 hours, 58 minutes) of travelling about the Earth. A Molniya orbit is a type of highly elliptical orbit with an inclination of 63.4˚ and an orbital period of precisely one half of sidereal day. Soviet/Russian communications satellites use these kinds of orbits which have been using this type of orbit since the mid 1960s. (Molniya (Russian mean: ''lightning'')). A satellite in a Molniya orbit spends most of its time over the Earth as a result of ''apogee dwell'' (Fig. 44). Only three such satellites are needed to give 24 hour coverage.
Also the satellites from the global position system (GPS) have the semi-synchronous period of orbits around the Earth.

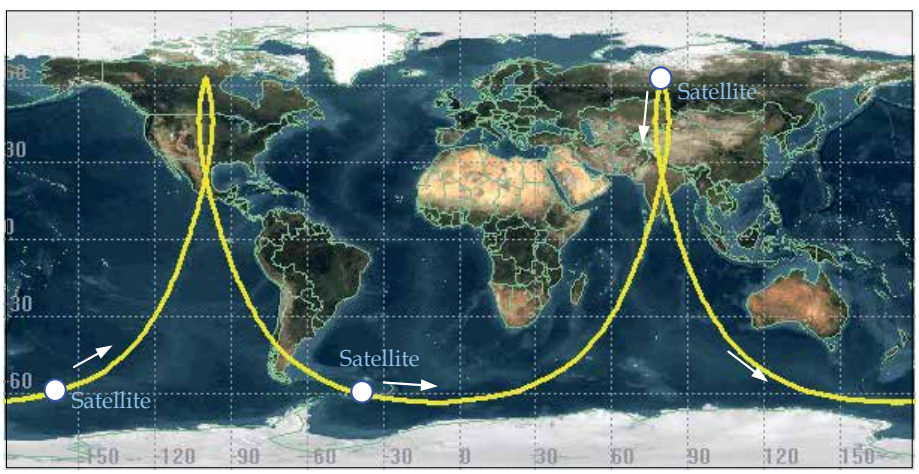


Fig. 44. Apparent trajectory of the satellites Molniya (URL 6).

□ ***Sun - Synchronous Orbit (SO)***
A Sun-synchronous orbit of the Earth's satellite is a geometric orbit which combines altitude and inclination in such a way that an object on that orbit ascends or descends over any given point of the Earth's surface at the same local mean solar time. In this way the Earth's surface illumination angle will be nearly in the same time.

Because the Earth is bulgy on the equator it follows from the equation (61) that the right ascension of ascending node has a constant negative velocity. In this way the plane of satellite orbit does not remain in the same position. So the ascending node is moving for approximately one degree per day which is equal to the angle of the Earth's shift during one day on the Earth orbit around the Sun.
Typical Sun-synchronous orbits are about 600-800 km in altitude, with periods in the 96-100 minute range, and inclinations of around 98˚. Variations on this type of orbit are possible; a satellite could have a highly eccentric sun-synchronous orbit.

7. References

Bazjanac, D. (1977): Seminar satellites – servants from space, Sveučilišna naklada Liber, Zagreb, Croatia

Beer, F. P. & Johnston, R. E. (1962): Mechanics for Engineers (Second Edition), McGraw-Hill Book Company, Inc., New York, San Francisco, Toronto, London

Carton, D. S. (1965): The propulsion and motion of rigid rocket – propellend launch vehicles, In: Dynamic of Rockets and Satellites, Groves, G. V. (Ed.), 1-30, North-Holland Publishing Company, Amsterdam

Danby, J. M. A. (1989): Fundamentals of Celestial Mechanics, Second Edition, Revised & Enlarged, Willmann-Bell, Inc. Richmond, Virginia, USA

Escobal, P. R. (1965): Methods of Orbit Determination, John Wiley & Sons, Inc., New York, London, Sydney

Hofmann-Wellenhof, B.; Lichtenegger, H. & Collins, J. (1994): GPS Theory and Practice, Spring-Verlag, Wien, New York.

Groves, G. V. (1965): Dynamic of Rockets and Satellites, North-Holland Publishing Company, Amsterdam

Kaula, W. M. (1966): Theory of Satellite Geodesy, Blaisdell Publishing Company, Waltham, Massachusetts, Toronto, London

Kozai, Y. (1959): Smithsonian Astrophysical Observatory Special Report No 22, 10.

Kozai, Y. (1966): The Earth Gravitational Potential Derived from Satellite Motion, Space Science Reviews 5, 818-879, D. Reidel Publishing Company, Dordrecht-Holland.

McCuskey, S. W. (1963): Introduction to Celestial Mechanics, Addison-Wesley Publishing Company, Inc. Reading, Massachusetts, Palo Alto, London

Petrovski, I.; Makoto, I.; Hideyuki, T.; Hitoshi, K.; Toshio, F.; Masayuki, S.; Toshiyuki, T. & Hiroaki, M. (2003): QZSS – Japan's New Integrated Communication and Position – Service for Mobile Users, GPS Word, June 2003, pp. 24-29.

Sehnal, L. (1978): Spherical Harmonic Analysis of the Earth's Albedo, Obserwacje Sztucznych Satelitów Ziemi, 18, 315-322

Solarić, M. (2007): Japan's Quasi-Zenith Communication and Position Satellite System, Cartography and Geoinformation, Vol 6, Special issue, June 2007, pp. 234-245.

URL 1: Robert Goddard
http://en.wikipedia.org/wiki/File:Dr._Robert_H._Goddard_-_GPN-2002-0001.31.jpg (12.4.2010)

URL 2: Sergei Korolov
http://www.centennialofflight.gov/essaySPACEFLIGHT/korolev/SP5G1.htm (12.4.2010)

URL 3: Wernher von Braun:
http://en.wikipedia.org/wiki/File:Wernher_von_Braun_crop.jpg (12.4.2010)
URL 4: How the Iridium Network Works
http://www.satphoneusa.com/iridium/network.html (20.1.2010)
URL 5: Iridium satellite constellation
http://en.wikipedia.org/wiki/Iridium_satellite_constelation (20.1.2010)
URL 6: Molniya orbit http://en.wikipedia.org/wiki/File:Molniya.jpg (23.1.2010)
URL 7: Tundra orbit http://en.wikipedia.org/wiki/Tundra_orbit (25.1.2010)
URL 8: Quasi-Zenith Satellite System
http://en.wikipedia.org/wiki/Quasi-Zenith_Satellite_System (18.4.2010)

23

System Aspects of Active Phased Arrays[1]

Amir I. Zaghloul[1,2], Ozlem Kilic[3] and Eric C. Kohls[4]
[1]*Virginia Polytechnic Institute and State University*
[2]*US Army Research Laboratory*
[3]*The Catholic University of America*
[4]*Lockheed Martin Corporation*
USA

1. Introduction

Demands in satellite communication systems for large number of spot beams with high beam-to-beam isolation and the need for flexibility in steering the beams impose requirements that are easier met with active phased arrays. The advances in MMIC and digital technologies played a major role in realizing these arrays for such systems. However, the added requirements and active array characteristics introduce a set of transmission impairments in the satellite link that did not necessarily exist in the conventional reflector-based satellite systems.

The design of the antennas for multiple beam communication systems depends on the beam definition, which in turn is a function of the system capacity and projected traffic patterns. Several systems require a large number of fixed narrow spot beams to cover the service area. A single reflector with a large number of feeds may provide such coverage. However, if the scanning loss is excessive due to the large number of beams scanned in one direction, multiple reflectors may be required. Alternatively, single or multiple phased arrays may be used. The phased array may have lower scan loss and a single phased array can handle a large number of beams. The phased array solution also offers higher reliability due to the use of the beam former versus the switching arrangement in a focal-region-fed reflector antenna. For a small number of beams at a time, a microwave beam former represents a simple and attractive solution, while if the number of beams is large a digital beam former is a more viable alternative. In many applications the choice between the microwave and the digital beam formers becomes a system issue.

Active phased arrays can also meet high flexibility requirements that have become the feature of several satellite communications systems. Among the flexibility requirements are the ability to form multiple beams, provide power sharing among beams through distributed amplification, and rapidly reconfigure and/or repoint the beams. The

---

requirements also include high reliability and the ability to readjust the feed or array coefficients to compensate for component failures. The conventional drawbacks of active antennas pertaining to mass, non-uniformity of performance of a large number of elements, and reliability have been largely overcome by the advances in monolithic microwave integrated circuits (MMICs), printed-circuit antennas and digital beam forming networks. The MMICs allow the beam-forming network, which may include signal-conditioning circuitry such as MMIC phase shifters and attenuators, to be incorporated in a lightweight transmission medium. The resulting antenna configurations have the potential for dramatic reductions in mass while providing increased efficiency, flexibility, and capacity. Similar features apply also to digital beam formers.

The satellite industry conducted evolutionary research and development programs to assess the feasibility of implementing MMIC-driven active phased-array antennas in high-capacity communications systems. A number of advanced concepts employing active phased arrays have been investigated. Among the early developments are Ku-band and X-band arrays that were sponsored by INTELSAT, COMSAT Corporation and the US Air Force (Zaghloul et al., 1994), (Zaghloul, 1996). Another array was also developed at COMSAT Laboratories at C-band to produce eight simultaneous and independently shaped and steerable beams. Extensive development was performed in parallel at Lockheed Martin Commercial Satellite Systems and accomplished high level of integration in active phased arrays (Jacomb & Lier, 2000). The design uses modular concepts to increase the design flexibility and afford efficient in-orbit operation using MMIC components and high-efficiency SSPAs. The characteristics of the communications systems that use phased arrays have been measured and analyzed to show the effects of the array's signal formation mechanisms on the transmission parameters. The following section reviews some of the reported development of these active phased arrays. The transmission impairment mechanisms associated with such systems are subsequently addressed. Other and similar characterizations of the transmission mechanisms, interference parameters and bit error analyses were reported (Maalouf et al., 1998), (Purdy & Lier, 2000), (Lier & Maalouf, 2003), (Maalouf & Lier, 2004). Of special interest is the analysis of intermodulation components that result from multi-carrier operation of active phased arrays (Sandrin, 1973), (Lier & Charrette, 2005). A reliability model to support the design of active phased arrays in satellite communications systems was also reported (Ruggieri, 1997).

2. Early Developments of Active Phased Arrays for Satellite Systems

2.1 Ku-Band Active Phased Array

The first example is a Ku-band array that consists of 24 active radiating elements, each equipped with an ortho-mode transducer (OMT) for dual-linear polarization (Zaghloul et al., 1994). Figure 1 shows a block diagram of the array components and Figure 2 shows the array assembly. The radiating elements are dual-linearly polarized square horns of 3-λ aperture each; a size chosen to reduce the number of elements in the overall aperture, reduce the mutual couplings between the elements, and eliminate the grating lobes within the ± 9° field of view of the geostationary satellite.

To form a number of simultaneous and independent beams, a beam-forming matrix (BFM) is used. Directly behind each element is a 2-W Solid-State Power Amplifier (SSPA). The SSPAs are located in separate housings that are mechanically cooled by a liquid cooling loop. The outputs of the amplifiers are probe coupled to the input of the OMTs. The inputs

to the amplifiers are connected to the respective outputs of the BFM, which is capable of forming four simultaneous beams. Each beam is independently steerable through the use of digital phase shifters located in the BFM. The logic circuitry necessary to drive and store the respective states of the phase shifters is also located within the BFM. The SSPA and BFM designs both make extensive use of MMICs. Phase-matched RF cables provide the signal path from the outputs of the BFM to the inputs of the amplifiers.

The BFM consists of four 1-to-24-way power dividers and 24 four-way power combiners. Each transmission path in the BFM consists of an MMIC 5-bit phase-shifter module, input and output distribution networks. A total of 96 MMIC phase-shifter modules are located in the BFM housing. A modular architecture was selected for implementation of the MMIC-populated BFM in the form of 4-input, 8-output shelves, corresponding to the four beams and eight elements, respectively. The three shelves are identical in assembly and performance. Each shelf shown in Figure 3 contains four 1-to-8-way dividers, eight 4-to-1-way combiners, thirty-two MMIC phase-shifter modules and associated control electronics. Four 1-to-3-way dividers distribute the signals for each beam to the three shelves. The input plane of the BFM shelf contains four eight-way Wilkinson dividers and MMIC phase shifter modules. The MMIC phase-shifters are assembled on separate carriers for individual testing. The output plane of the BFM shelf also contains the phase-shifter driver and level-shifter electronics that are mounted on multilayer alumina boards. The ground plane in the middle serves as an RF and DC ground and also provides isolation between input and output planes.

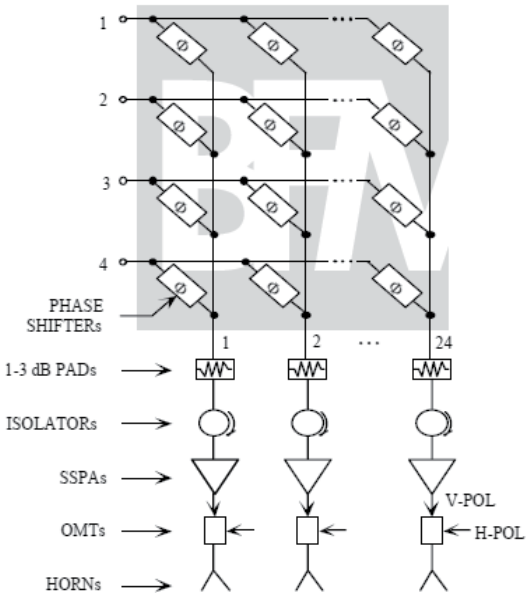


Fig. 1. Block diagram of multiple-beam active phased array

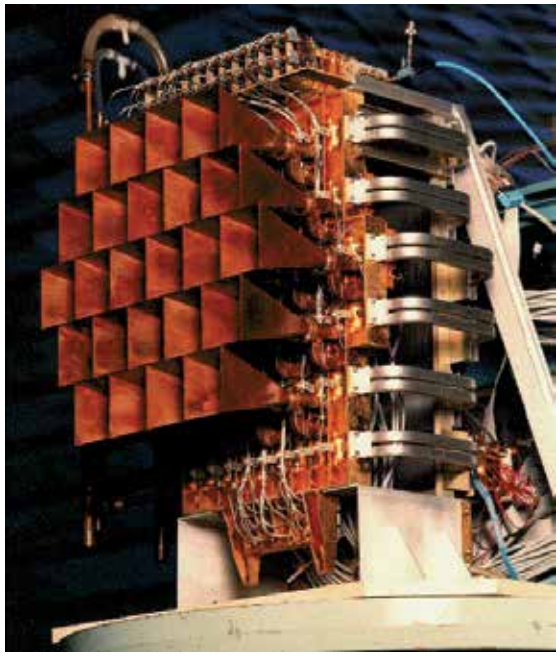


Fig. 2. Ku-band Active Phased Array

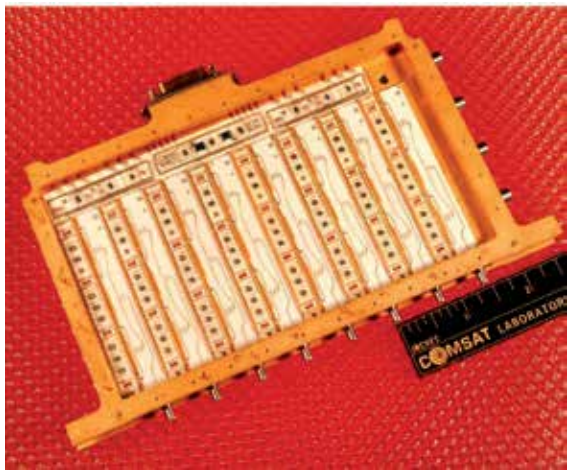


Fig. 3. 4x8 Shelf in Ku-band BFM

This design approach for the BFM provides several advantages including modularity, reproducible assembly and performance. Broadband Wilkinson dividers and combiners provide inherently good amplitude and phase balance between the paths. Furthermore, path-to-path insertion loss variations can be minimized by the appropriate selection of MMIC phase shifter modules so that low loss phase shifters are matched with high loss paths, and vice versa.

The power amplifier design combines MMIC and quasi-monolithic technology to achieve optimum DC-to-RF conversion and linearity, while simultaneously meeting output power

and size requirements. The amplifiers are designed to be integrated directly in back of the array to minimize total output losses.

2.2 X-Band High-Power Active Phased Array

A study performed at COMSAT Laboratories addressed a set of requirements for a future DSCS system (Zaghloul, 1996). A key antenna specification is to transmit four beams simultaneously. These beams are to be shaped and steered independently, and can range in size from the minimum of 2° to the full earth coverage of 17°. An EIRP of 50 dBW is required for the 2° beams, and the overall antenna mass and DC power consumption are to be minimized with goals of 200 lb. and 400 Watts, respectively.

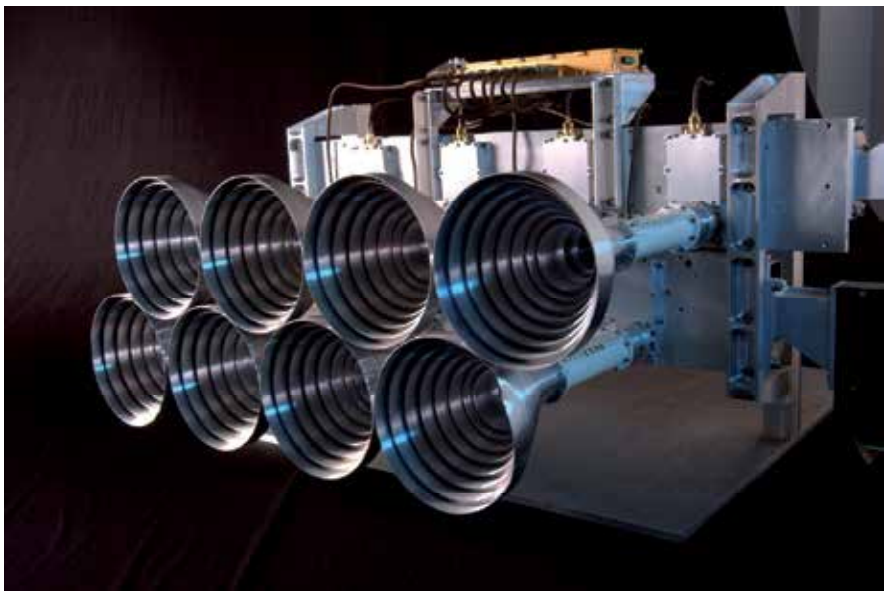


Fig. 4. Demonstration X-band array for DSCS system

A proof-of-concept array was developed as a subarray of such an antenna, and aims at demonstrating the most critical components required of a full-up active array antenna. This demonstration array, shown in Figure 4, utilized a total of eight radiating elements to form two independently steerable and reconfigurable beams. A key component developed for the program is the Beam Forming Matrix (BFM) whose function is to shape and steer the beams. The BFM implements modular MMIC technology in its phase shifters and attenuators which provide highly repeatable and uniform performance that is critical for this application. The lightweight radiating elements are composed of scalar ring horns and pin-polarizers that together provide high polarization purity required in a dual-polarization system should that option be exercised. Another key development is the highly efficient 2 Watt power amplifiers which are positioned within the transmit modules behind each radiating element.

2.3 C-Band Multiple-Beam Active Phased Array

The linearly polarized Ku-band array described above demonstrated the concept of active phased arrays for communications systems. It used large pyramidal horns as the radiating elements. One of the most desirable features, which have to be achieved, is lightweight. Because the size of the array can be large, especially at lower frequencies, the need for compact, lightweight radiating elements became paramount. The next array in this development series used printed-circuit patch radiators in order to achieve this goal. The array, shown in Figure 5, was designed at C-band (3.6-4.2 GHz) and featured printed-circuit radiators, a highly modular approach for the beam-forming matrix as well as for the active aperture, and an integrated thermal control system (Zaghloul et al., 1994). The array consists of 69 electromagnetically coupled patch (EMCP) elements, each producing two orthogonal circular polarizations. The elements are fed with 2-Watt SSPAs that are integrated in the array structure. The BFM provides independent control of eight simultaneous beams.

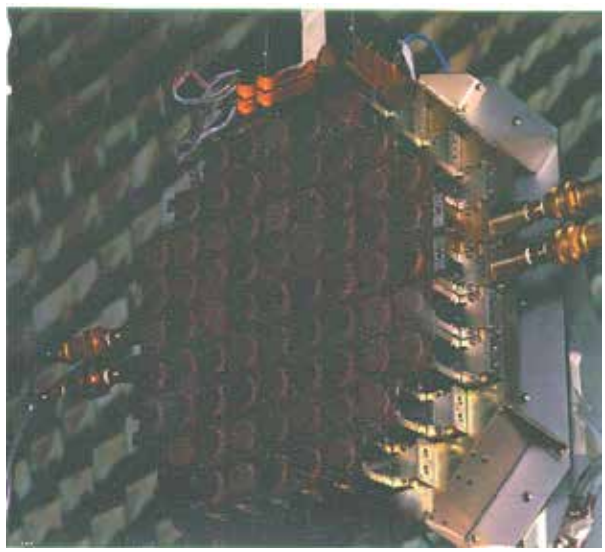


Fig. 5. C-Band active phased array

The array active elements demonstrate a high level of integration. The integrated element includes the EMCP radiating element, redundant SSPAs, MMIC gain blocks, redundancy switches, and monitoring circuit at the SSPA output. The SSPA consists of a high-efficiency power amplifier preceded by multistage MMIC preamplifiers and an MMIC linearizer to predistort the signal such that high linearity is achieved simultaneously with the required efficiency.

The active array is controlled by an 8X69 BFM, which consists of nine 8x8, shelves and accommodates up to eight beams connected to a maximum of 72 elements. For the 69-element array, three outputs are match terminated. The modular architecture of the Ku-band BFM described above has been further enhanced by implementing the C-band BFM shelf with eight 1x8 BFM modules. Each of these modules is realized by modular MMIC packages, each containing phase and amplitude control elements and associated digital control for each BFM cross point. This additional modularity further offers improved design flexibility for larger size BFMs, ease of repair and reduced manufacturing costs. The 8 x 8

shelf, shown in Figure 6, is composed of eight 1x8 BFM modules on the topside and eight 8-way power combiners on the bottom surface. Each 1x8 BFM module is an independently testable unit and includes the control circuits for eight MMIC packages and also provides the RF signal paths from the top to the bottom of the BFM shelf. Each 1x8 module consists of an eight-way power divider, eight MMIC packages and the digital control circuits. The most basic building block of the BFM is the MMIC package, shown in Figure 7, which contains an MMIC 5-bit phase-shifter, a 5-bit MMIC attenuator and three digital control ICs. This MMIC package can be individually tested and hermetically sealed to provide a rugged design.

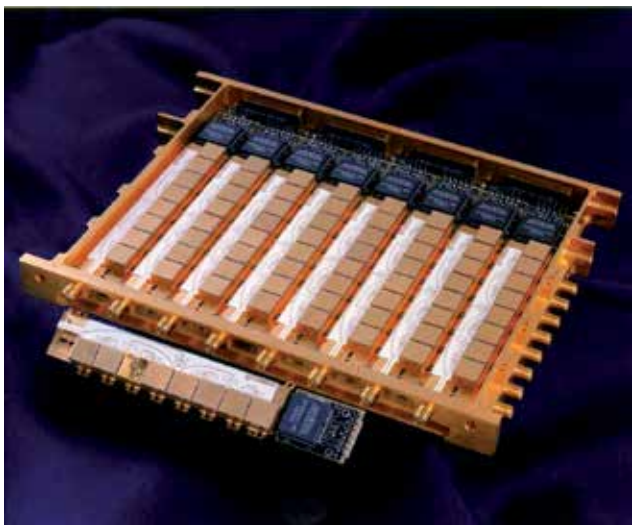


Fig. 6. 8 X 8 Shelf in C-band BFM

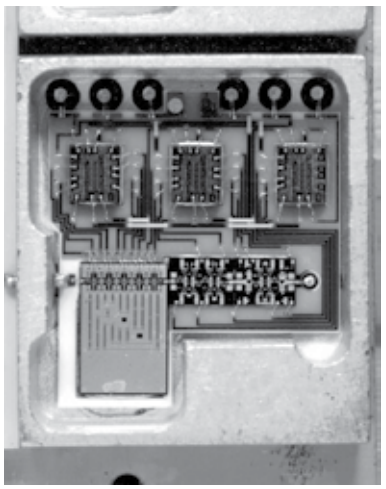


Fig. 7. Basic C-band MMIC module

3. Evaluation of Transmission Parameters

This section presents evaluations of three transmission parameters through analysis and/or measurements: radiation patterns of intermodulation products, bit error ratios and co-channel interference between beams.

3.1 Far-field radiation pattern measurements of intermodulation products

Intermodulation products (IMP) result from the nonlinearities in the power amplifiers that are used in the active phased arrays. The distributed amplifiers are shared between all channels that are transmitted through all elements in the array. Due to the periodicity of the array structure and the nature of the intermodulation product generation, the location of such undesired and interference-like radiation can be predicted. The intermodulation products generated by a pair of carriers whose frequencies are f_1 and f_2 form distinctive beams with peak positions that are the vector sums of the carriers' positions. The frequency of the intermodulation product is defined by:

$$f_{third} = 2*f_2 - f_1$$

The direction of the intermodulation product beam can be predicted by (Sandrin, 1973):

$$\bar{k}_{third} = 2*\bar{k}_2 - \bar{k}_1$$

where:

$$\bar{k} = 2\pi\frac{d}{\lambda}\left(\cos\varphi\ \sin\theta\ \hat{x} + \sin\varphi\ \sin\theta\ \hat{y}\right)$$

is the direction vector of the beam. The levels of the intermodulation products can be estimated knowing the nonlinear input/output characteristics of the amplifiers. The nonlinearities are represented in Bessel function expansion series that leads to power level evaluations (Fuenzalida et al., 1994). Typical amplifier characteristics are shown in Figure 8. The output power levels are plotted against the input back-off (IBO). The IBO is the back-off of the amplifier total input power level that corresponds to the output back-off (OBO), which is defined as the ratio of the peak single-carrier output power to the total multi-carrier output power at the operating point.

Multiple-beam intermodulation measurements for the 24-element Ku-band array described above, and shown in Figures 1 and 2, were performed in an anechoic chamber, where the antenna under test operated in transmit mode at relatively high power levels. Array control, measurement control, and data collection proceeded in essentially the same manner as during regular single carrier pattern measurements (Ekelman et al., 1994). The signal sources were composed of a synthesizer driving a 20-Watt Traveling Wave Tube Amplifier (TWTA) through a band-pass filter and 10-dB directional coupler. Three input signals were coupled off at the BFM inputs, and drive levels were continually monitored for the duration of the measurement. The drive level for all carriers was 33 dBm so as to provide the maximum drive into the SSPA's without incurring gain compression at the phase shifters in the BFM. Carrier frequencies and beam locations were chosen as shown in Figure 9 and Table 1, and designated as frequencies a, b and c.

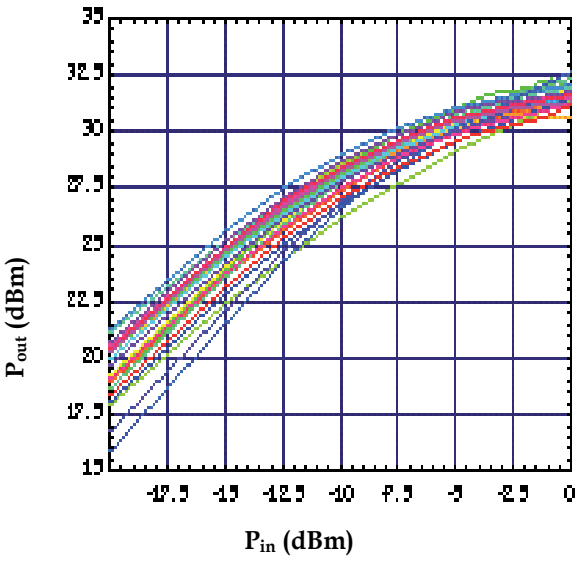


Fig. 8. Transfer characteristics of SSPA models

Signal Type	Frequency (GHz)	Generator Polynomial	Location (Θ^o, Φ^o)
IMP	11.67	2 • a - c	(0°, 4°)
IMP	11.72	a + b - c	(4°, 4°)
IMP	11.77	2 • b - c	(8°, 4°)
IMP	11.80	2 • a - b	(-4°, 0°)
Carrier	11.85	a	(0°, 0°)
Carrier	11.90	b	(4°, 0°)
IMP	11.95	2 • b - a	(8°, 0°)
IMP	11.98	c + a - b	(-4°, 4°)
Carrier	12.03	c	(0°, 4°)
IMP	12.08	b + c - a	(4°, 4°)
IMP	12.16	2 • c - b	(-4°, 8°)
IMP	12.21	2 • c - a	(0°, 8°)

Table 1 Carrier and Intermodulation Parameters

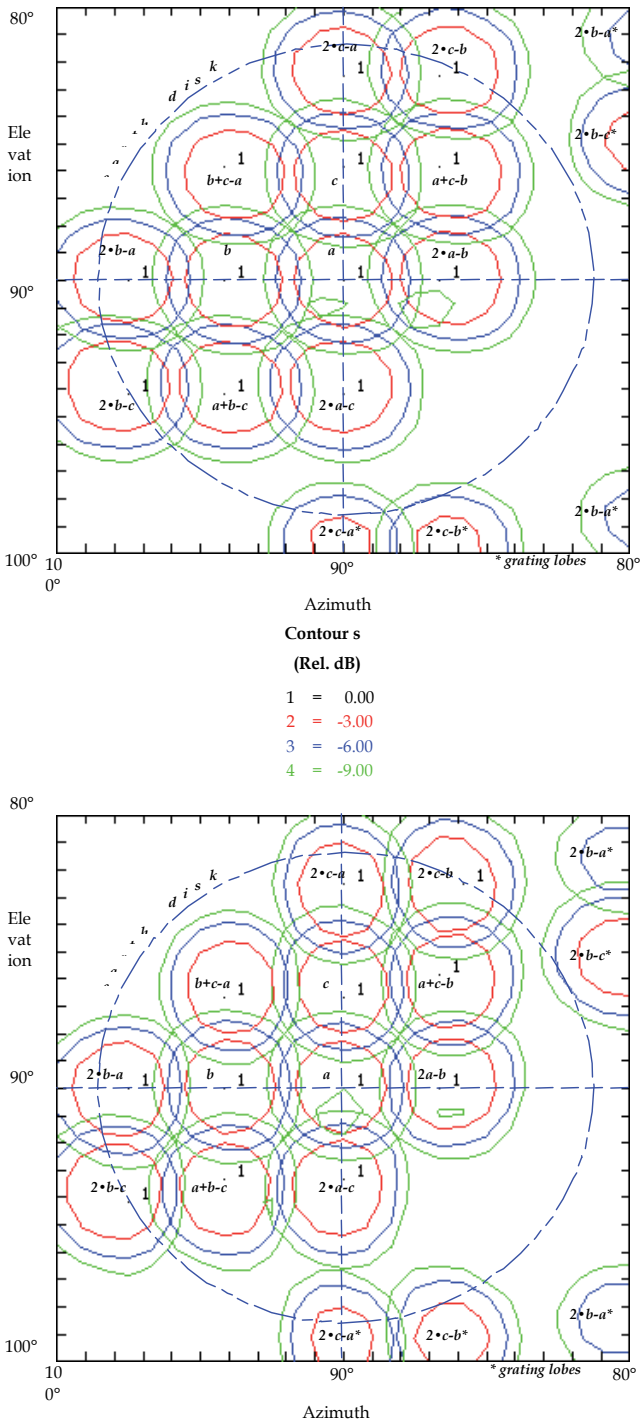


Fig. 9. Radiation far field intermodulation contours, Top: predicted, Bottom: measured

Predicted third order IMP locations are shown in Figure 9 (top). Pattern contours were measured at the 3 carrier frequencies and the 9 in-band 3rd-order IMP frequencies, and are shown in Figure 9 (bottom). The carrier and intermodulation component locations are listed in Table 1. Measured pattern contours demonstrate that the predicted patterns are accurate in regard to the beam shapes and locations.

3.2 Bit-Error Ratio Measurements

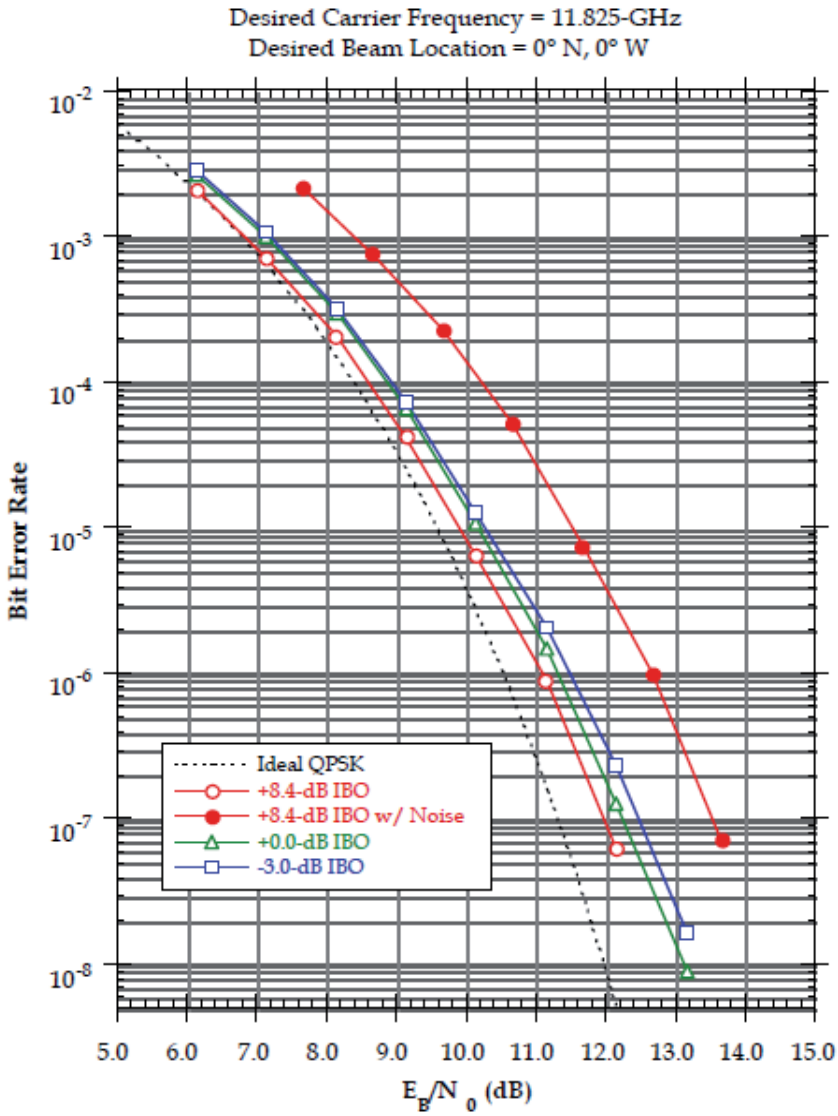


Fig. 10. BER for single carrier

The bit-error ratio (BER) measurements of the same Ku-band active phased array described above were performed under different operating conditions: single channel and multi-channel operations at different power levels in one beam with and without traffic loading in a different beam (Kohls et al., 1995). In order to simulate the traffic loading represented in multitude of populated signal channels, a single white noise source was connected to one of the BFM input ports. The noise-loading beam was scanned to a position isolated from the beam under test. When in operation, the drive level of the noise-loading channel was maintained at a level equivalent to a fully populated transponder bank. At such a drive level, it is possible to observe the effects of driving the SSPA's into saturation.

The purpose of the single channel measurements is to assess independently the impact of noise loading and carrier input drive level on BER performance. Figure 10 shows the results of the single channel measurements where four operating conditions are investigated: the desired carrier at -3.0, 0.0, and +8.4 input back-off (IBO) levels with no noise loading, and the desired carrier at +8.4-dB IBO in the existence of a noise loading at a level of +0.7-dB IBO. The dotted line indicates the theoretical BER limit for a QPSK signal. The changes in the BER characteristics as the channel drive level is increased indicate the degradation that is attributable to the non-linear distortion effects as the SSPA's are driven towards saturation. The difference between the curves with and without the noise loading indicates that the degradation is primarily attributable to carrier suppression as most of the available SSPA power is expended in the loading channel. As the traffic or the number of carriers represented by the noise loading increase, the relative level of the carrier under test decreases within the available SSPA power. This causes the carrier suppression and the corresponding degradation in the BER performance. The same degradation due to non-linear distortion is still present in the unloaded case, as in the +0.0-dB back-off case, but the impact is secondary by comparison.

The purpose of the multi-channel measurements is to assess the impact of adjacent channel interference on BER performance, while the adjacent-channel beam is either in the same location as, or scanned away from, the desired-channel beam. It is also the purpose of the multi-channel measurements to assess the impact of co-channel interference on BER performance, while the co-channel interferer beam is scanned away from the desired-channel beam. All multi-channel measurements were performed with and without noise loading. Figure 11 shows the results of one set of multiple channel measurements. In this case, the BER of the desired channel was measured with the desired channel set at a +8.4-dB IBO drive level. The BER was then re-measured after the addition of the interfering channel, which was also set at a +8.4-dB back-off drive level. Finally, the desired channel BER was measured with the addition of the noise-loading channel, which was set to +1.5-dB back off. Similar to the single channel case, the most degradation occurs due to the carrier suppression as a result of the noise loading.

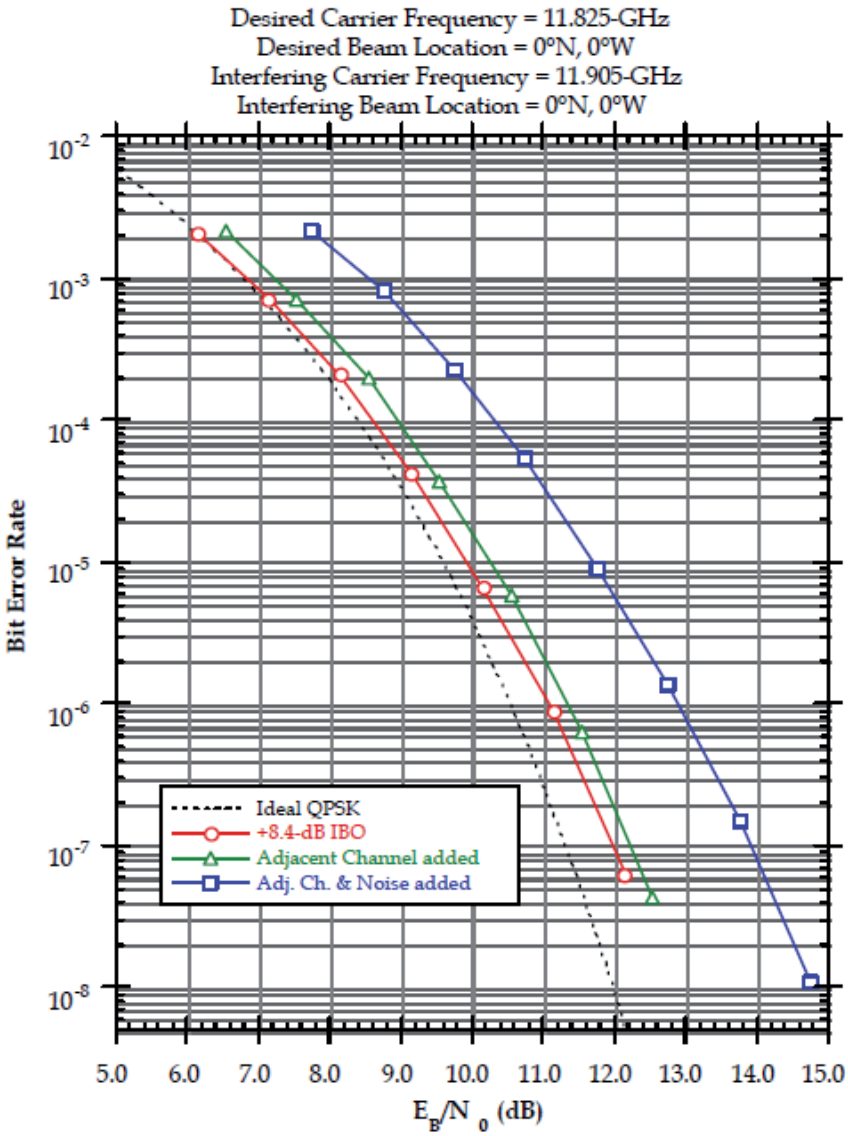


Fig. 11. BER for a carrier with on-axis adjacent channel interferer

3.3 Co-Channel Beam Interference

In cellular satellite communication systems employing multiple beam antennas, mutual interference among beams operating at the same frequency can be severe especially for systems that use a large number of beams. The co-channel interference depends on the coverage requirements, antenna performance for both user terminals and satellite payload, and user traffic within each cell. The carrier-to-interference ratio (CIR), which is the ratio of the power levels of the desired signal and the aggregate interference, is used to quantify this effect. Multiple parameters affect the CIR, some of which are readily predictable from the

system parameters (such as the total number of co-channel beams, and their separation), while the others need to be determined by statistical means (such as the antenna pointing error and user traffic).

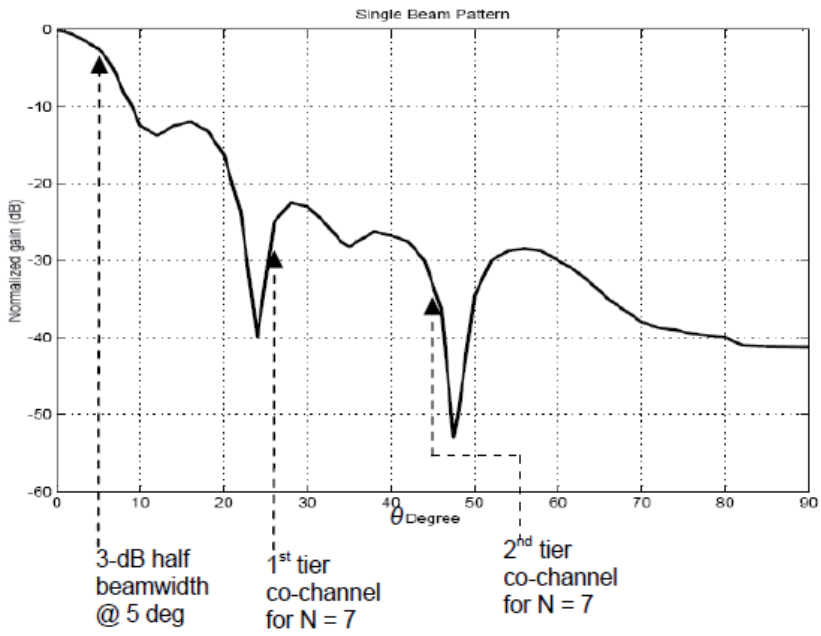


Fig. 12. Satellite antenna pattern used in interference calculations

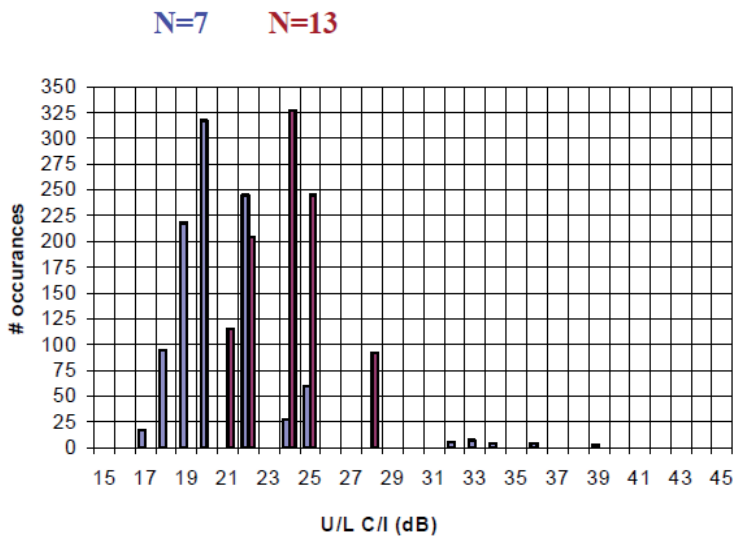


Fig. 13. Example of simulation for uplink interference

The downlink CIR is measured at the user terminal and is a function of the payload antenna characteristics and the location of the user terminal within the beam. The uplink CIR is measured at the payload and is a function of the user traffic within all co-channel beams at a

given time, as well as their locations within the beam (Kilic et al., 1999; Chhabra, et al., 2005; Chhabra et al., 2009). To demonstrate these interferences, array patterns for the satellite antenna are assumed based on large apertures consistent with the cell size. Designs similar to the ones described above can be used with the aperture distribution optimized to minimize the interference. In the results to follow, the pattern shown in Figure 12 was used to represent the relative gain function of all cellular satellite beams in a cellular low-altitude system. The pattern is based on an actual satellite antenna design provided by a satellite manufacturer. The figure shows the location of the co-channel interference beams for a hexagonal cellular system with a frequency re-use pattern of N = 7. The antenna pattern can be optimized further to position the pattern nulls at the center of the interfering cells.

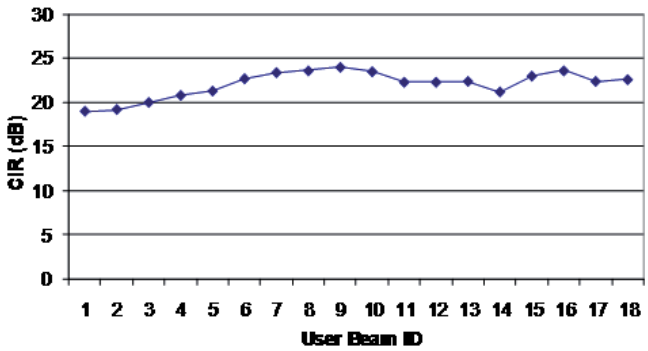


Fig. 14. Example of simulation for downlink interference

Figure 13 demonstrates a statistical histogram for the uplink CIR for two cases where the number of frequency segments in the frequency re-use pattern N is 7 and 13, respectively. The uplink CIR is calculated hundreds of times with uniform statistical distribution of the interfering users within the co-channel beams. The number of times the CIR falls within certain range (# of occurrences) is recorded for different ranges and the results constitute the statistical histogram. As the number of frequency segments increases, the number of interfering beams that use the same frequency segment decreases, leading to higher levels of CIR. Figure 14 shows the downlink CIR levels at different beams that are identified in an ascending order starting from the coverage center. All users in these examples are assumed to be located at the beam center. As Figure 13 suggests, the uplink CIR can vary as much as 12 dB for N = 7 and 6 dB for N = 13, depending on the user traffic within each beam. The downlink CIR, on the other hand, is more stable due the fact that all co-channel interferers are always on. However, depending on the location of the beam in the coverage area, the variation in downlink CIR can be as much as 5 dB as the simulation for the case shown in Figure 14 indicates.

4. Conclusion

Active phased arrays provide significant advantages in satellite communications systems that require the generation of a large number of beams. The advantages include the flexibility in generating the beams, sharing of power between beams and achieving a large

number of frequencies re-use. These advantages come at a cost at the system level performance that has to be studied and can be minimized. The system-level performance of active phased arrays, which employ distributed amplifiers across the array aperture can be accessed through measurements and computer simulation. The primary parameter in identifying the goodness of the system is the BER, which is a function of the net carrier to noise and interference ratio (CNIR).

In a multiple beam phased array system, intermodulation products and co-channel interference are two key contributors to the degradation of the CNIR. The far-field intermodulation radiation follows certain patterns, which can be predicted using software tools that are verifiable by measurements. The intermodulation radiation degrades the CNIR if it falls within certain frequency bands in the directions of the beams that use those bands. Judicial choices of the operating frequencies and associated beams can reduce or mitigate such degradation.

The two components of the co-channel interferences, downlink and uplink, are functions of both deterministic parameters of the system, such as the total number of beams, antenna performance and frequency re-use scheme, as well as stochastic parameters, such as the traffic pattern, location and equipment pointing errors of the users. While these effects can cause significant variations in the co-channel CIR, the system can be designed based on the worst-case predictions in order to offer higher levels of margin.

An associated phenomenon is the BER degradation due to carrier suppression as a result of traffic loading. This exceeds the degradation that results from driving the desired channel in the non-linear region of the amplifier or at saturation. It also exceeds the degradation that results from adjacent channel or co-channel interference. This type of BER degradation can be controlled by proper loading of the satellite channels.

5. References

Chhabra, S.; Zaghloul, A.I. & Kilic, O. (2005). Co-Channel Interference in Satellite-Based Cellular Communication Systems, *Proceedings of the URSI General Assembly*, New Delhi, India, October 2005.

Chhabra, S.; Kilic, O. & Zaghloul, A.I. (2009). *Co-Channel Interference in Satellite Links Utilizing Frequency Reuse,* VDM Publisher, ISBN 978-3-639-17714-5, Saarbrucken, Germany.

Ekelman, E.P.; Kohls, E.C.; Zaghloul, A.I. & Assal, F.T. (1994). Measured Performance of a Ku-Band Multibeam High-Power Phased-Array, *Proceedings of the IEEE International Symposium on Antennas and Propagation*, pp. 844-847, Seattle, Washington, June 1994.

Fuenzalida, J.C., Shimbo, O. & Cook, W.L. (1973). Time-Domain Analysis of Intermodulation Effects Caused by Nonlinear Amplifiers, *COMSAT Technical Review*, 1973.

Jacomb-Hood, A. & Lier, E. (1999). Multibeam Active Phased Arrays for Communications Satellites, *IEEE Microwave Magazine*, pp. 40-47, December 2000.

Kilic, O.; Zaghloul, A.I.; Thai, M. & Nguyen, L. (1999). Beam Interference Modeling in Cellular Satellite Communication Systems, *Proceedings of the URSI General Assembly*, Toronto, Canada, 1999.

Kohls, E.; Ekelman, E.; Zaghloul, A.I. & Assal, F. (1995). Intermodulation and BER Performance of a Ku-Band High Power Active Phased Array, *Proceedings of the IEEE International Symposium on Antennas and Propagation*, pp. 1404-1408, Newport Beach, California, June 1995.

Lier, E.; Purdy, D. & Maalouf, K. (2003). Study of Deployed and Modular Active Phased Array Multibeam Satellite Antennas, *IEEE Antennas and Propagation Magazine*, vol. AP-45, no. 5, pp. 34-45, October 2003.

Lier, E. & Cherrette, A. (2005). An Intermodulation Suppression Technique for Transmit Actrive Phased Array Satellite Antennas with Multiple Shaped Beams, *IEEE Transactions on Antennas and Propagation*, vol. AP-53, no. 5, pp. 1853-1858, May 2005.

Maalouf, K.; Gaus, Jr., R.; Sowelam, S. & Jacomb-Hood, A. (1998). Error Rate Estimation in a Multi-Channel Active Phased Array," *Proceedings of the IEEE International Conference on Communications*, pp. 402-406, June 1998.

Maalouf, K. & Lier, E. (2004). Theoretical and Experimental Study of Interference in Multibeam Active Phased Array Transmit Antenna for Satellite Communications, *IEEE Transactions on Antennas and Propagation*, vol. AP-52, no. 2, pp. 578-592, February 2004.

Purdy, D. & Lier, E. (2000). "EIRP and C/I Analysis of Multibeam Active Phased Array Transmit Antennas for Communication Satellites," *Proceedings of the IEEE International Conference on Phased Array Systems and Technology*, Dana Point, CA, pp. 203-206, May 2000.

Ruggieri, M. (1997). A (Design-Oriented) Reliability Model for Active Phased Arrays in Satellite Communication Systems, *IEEE Aerospace and Electronic Systems Magazine*, vol. 12, no. 10, pp. 23-29, October 1997.

Sandrin, W. (1973). Spatial Distribution of Intermodulation Products in Active Phased Array Antennas, *IEEE Transactions on Antennas and Propagation*, vol. AP-26, no. 6, pp. 864-868, November 1973.

Zaghloul, A.I.; Gupta, R.K.; Assal, F.T.; Martin, J.E. & Wernek, S.B. (1994). Active Phased-Array Development at C- and Ku-Band, *Proceedings of the 15th AIAA International Communications Satellite Systems Conference*, pp. 1683-1691, San Diego, California, February – March, 1994.

Zaghloul, A.I. (1996). X-Band Active Transmit Phased Array, *US Air Force Report*, RL-TR-96-164, September 1996.

Edited by Nazzareno Diodato

This study is motivated by the need to give the reader a broad view of the developments, key concepts, and technologies related to information society evolution, with a focus on the wireless communications and geoinformation technologies and their role in the environment. Giving perspective, it aims at assisting people active in the industry, the public sector, and Earth science fields as well, by providing a base for their continued work and thinking.

ISBN 978-953-51-5932-2

Photo by PhonlamaiPhoto / iStock

IntechOpen