LA-UR-08-6246

Roadrunner: What makes it tick?

Los Alamos Computer Science Symposium
October 14, 2008

Ken Koch

Roadrunner Technical Manager,
Computer, Computational, and Statistical Sciences Division,
Los Alamos National Laboratory

Work presented was performed by a large team of Roadrunner project staff!

The messages this talk will convey are:

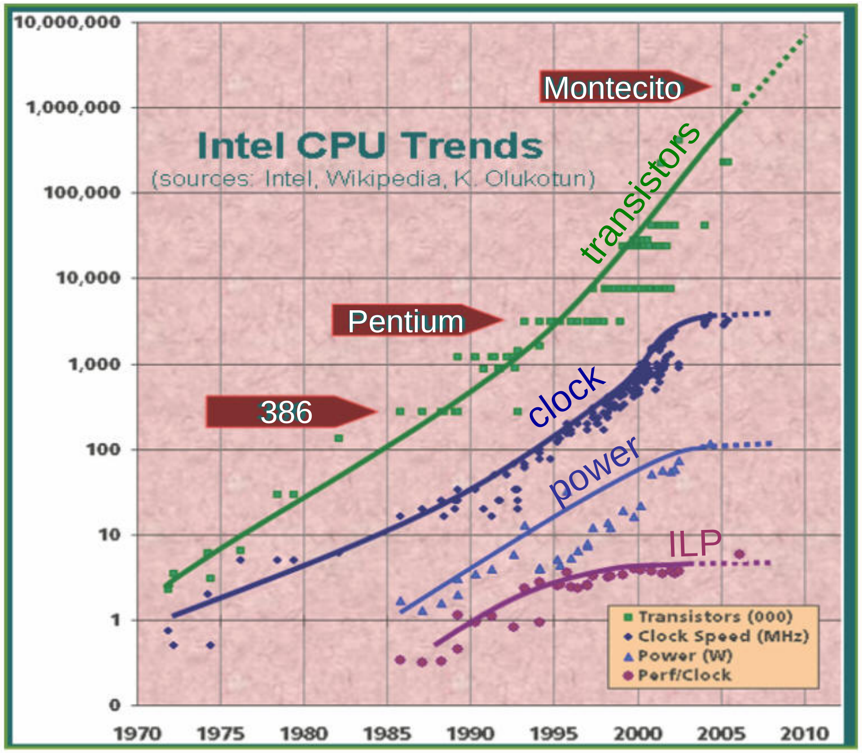
- Why Roadrunner? Why Cell?
  - *A bold but important step toward the future*
- What does Roadrunner look like?
  - *Cluster-of-clusters with node-attached Cells*
- Concepts for Programming Roadrunner
  - *MPI, Opteron+Cell, “local-store” memory & DMA transfers*
- Status and plans for Roadrunner
  - *Unclassified Science opportunities*

The Cell Processor

a harbinger of the future

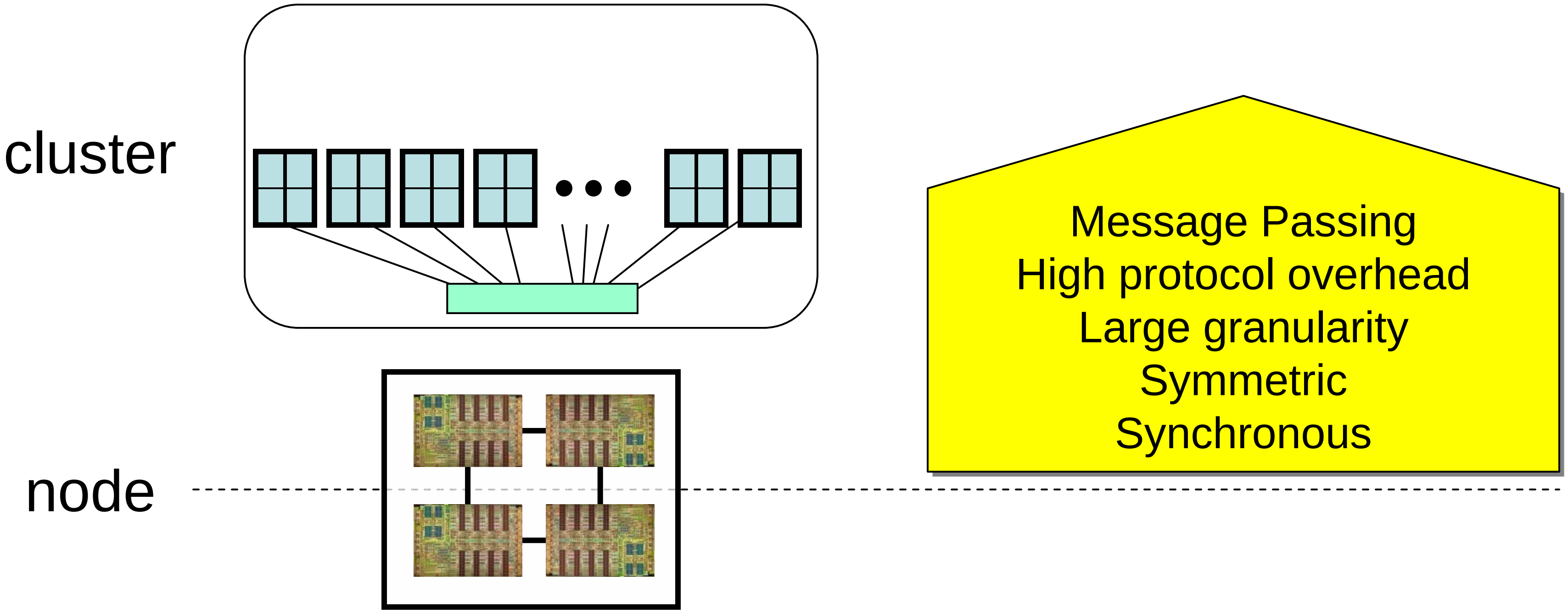
Microprocessor trends are changing

- Moore’s law still holds, but is now being realized differently
  - *Frequency, power, & instruction-level-parallelism (ILP) have all plateaued*
  - *Multi-core is here today and many-core (≥ 32) looks to be the future*
  - *Memory bandwidth and capacity per core are headed downward (caused by increased core counts)*
  - *Key findings of Jan. 2007 IDC Study: “Next Phase in HPC”*
    - ***new ways of dealing with parallelism will be required***
    - *must focus more heavily on bandwidth (flow of data) and less on processor*



From Burton Smith, LASCI-06 keynote, with permission

We are programming thousands of processors with MPI

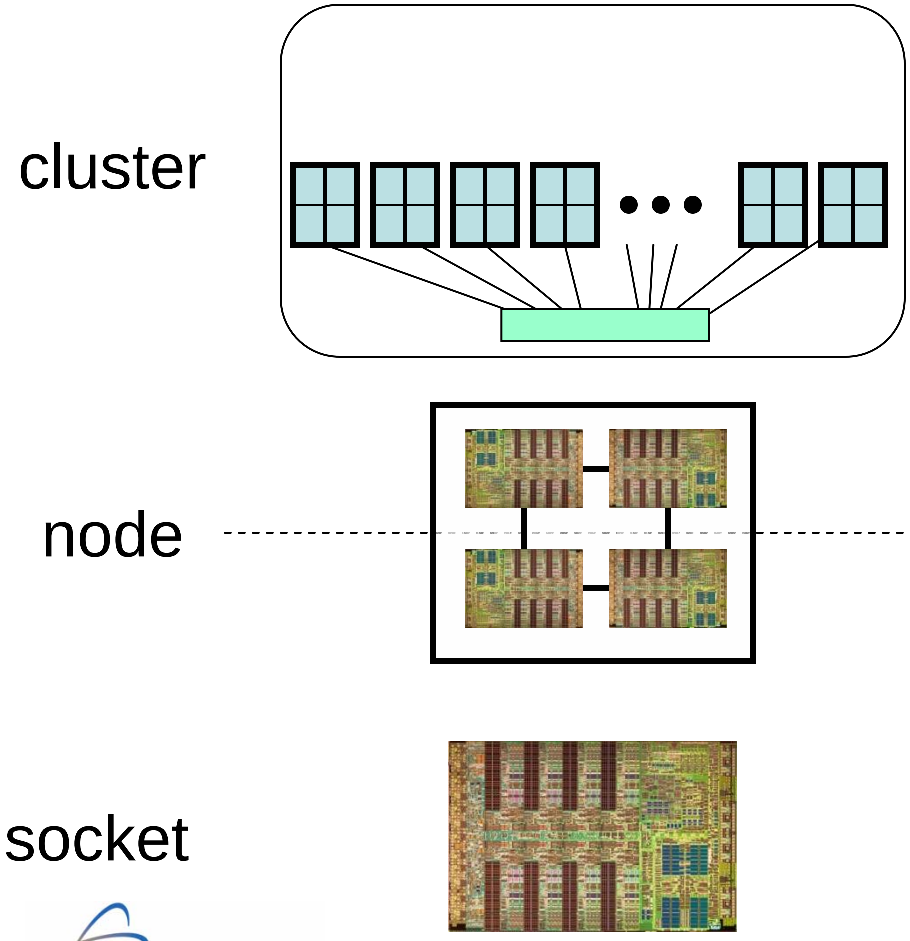


cluster

node

Message Passing
High protocol overhead
Large granularity
Symmetric
Synchronous

Future supercomputers will require new programming models



Message Passing
High protocol overhead
Large granularity
Symmetric
Synchronous

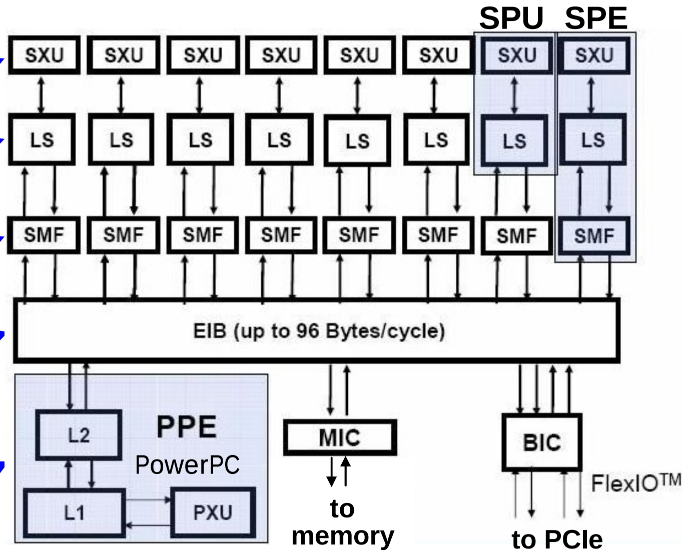
Not Message Passing

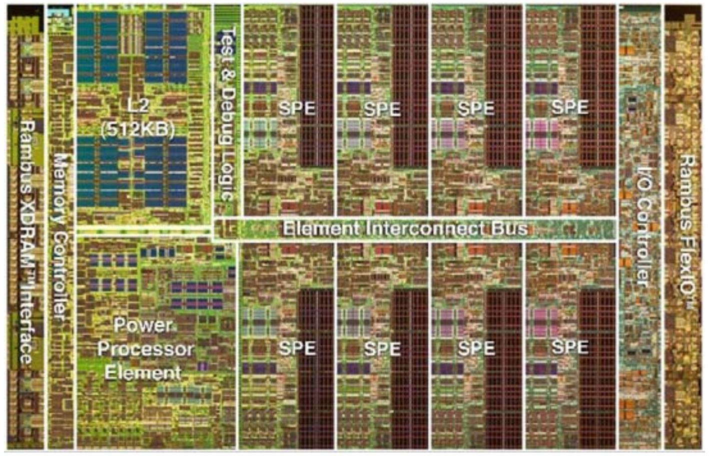
Parallelism and heterogeneity require new approaches:

Threads, OpenMP, Accelerators …

The Cell processor is an (8+1)-way heterogeneous parallel processor

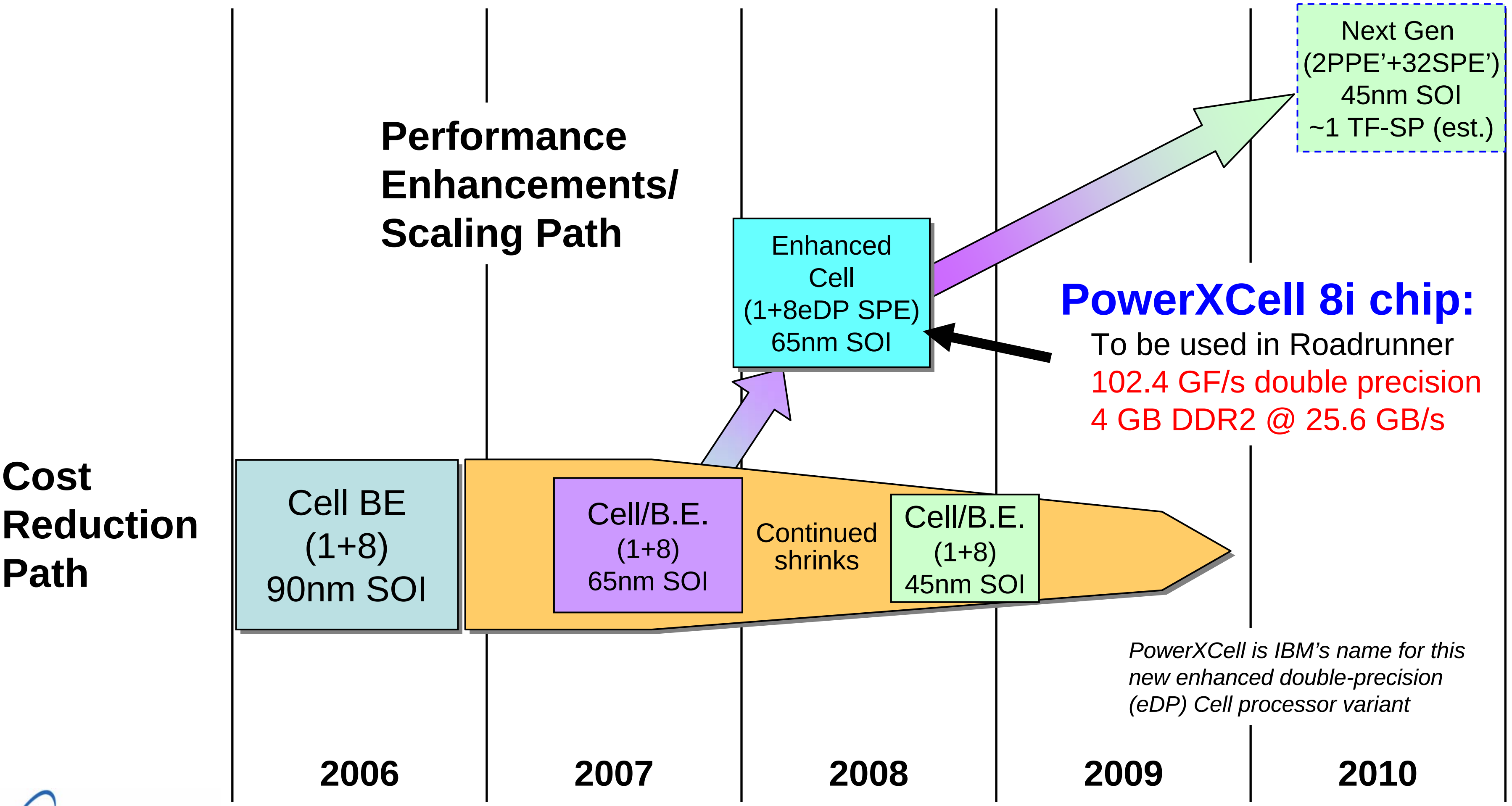
- Cell Broadband Engine (CBE*) developed by Sony-Toshiba-IBM
  - used in Sony PlayStation 3
- 8 Synergistic Processing Elements (SPEs)
  - 128-bit vector engines
  - 256 kB local memory (LS = Local Store)
  - Direct Memory Access (DMA) engine (25.6 GB/s each)
  - Chip interconnect (EIB)
  - Run SPE-code as POSIX threads (SPMD, MPMD, streaming)
- PowerPC PPE runs Linux OS
- Current Cell performance:
  - 204.8 GF/s SP & 13.65 GF/s DP
  - 512 MB @ 25.6 GB/s XDR memory
  - **Insufficient for a Petaflop/s machine**





* trademark of Sony Computer Entertainment, Inc.

IBM is creating new Cell processors



PowerXCell 8i chip:

To be used in Roadrunner

102.4 GF/s double precision

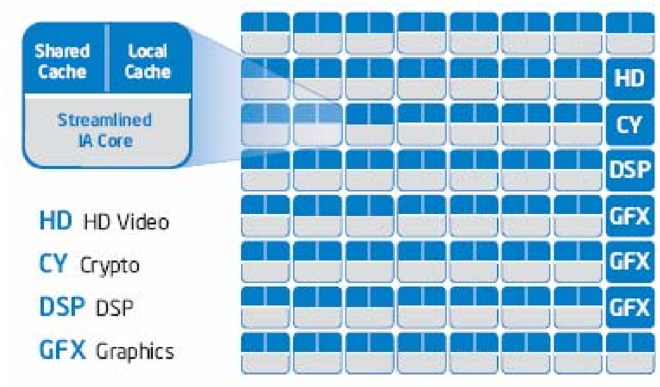
4 GB DDR2 @ 25.6 GB/s

PowerXCell is IBM's name for this new enhanced double-precision (eDP) Cell processor variant

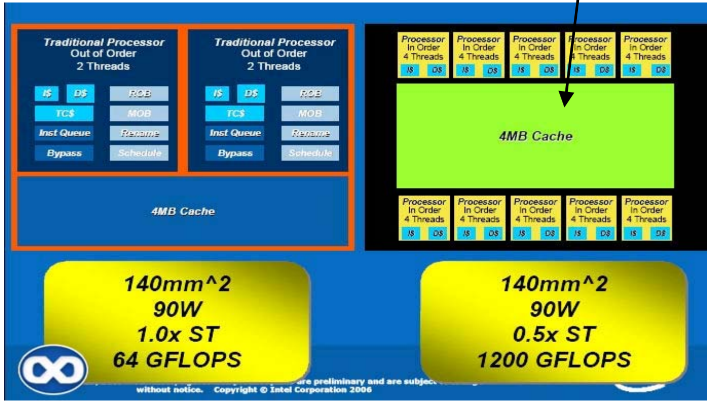
All future dates and specifications are estimations only; Subject to change without notice. Dashed outlines indicate concept designs.

Industry presentations show changing trends in processors

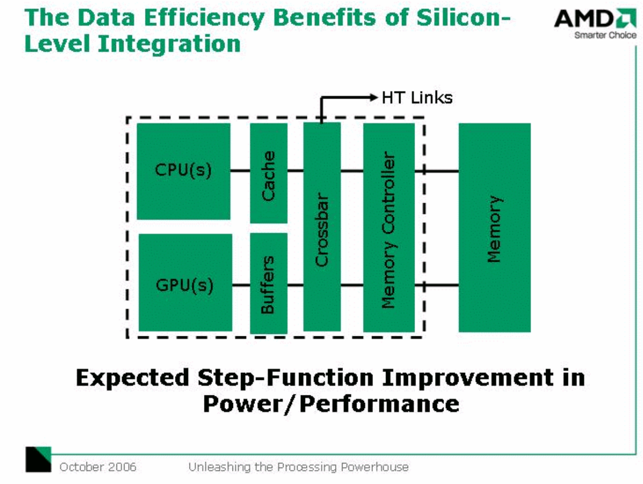
Intel's Microprocessor Research Lab



Intel's Visual Computing Group - Larabee


Traditional Processor Out of Order 2 Threads
I$
D$
ROB
TC$
MOB
Inst Queue
Rename
Bypass
Schedule
Traditional Processor Out of Order 2 Threads
4MB Cache
Processor In Order 4 Threads
I$
D$
4MB Cache
140mm^2
90W
1.0x ST
64 GFLOPS
140mm^2
90W
0.5x ST
1200 GFLOPS
are preliminary and are subject
without notice. Copyright © Intel Corporation 2006

AMD Fusion

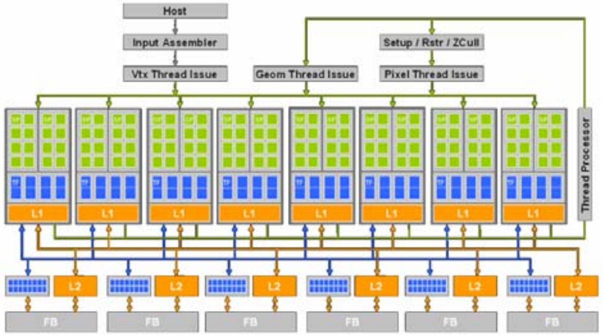


The Data Efficiency Benefits of Silicon-Level Integration

Expected Step-Function Improvement in Power/Performance

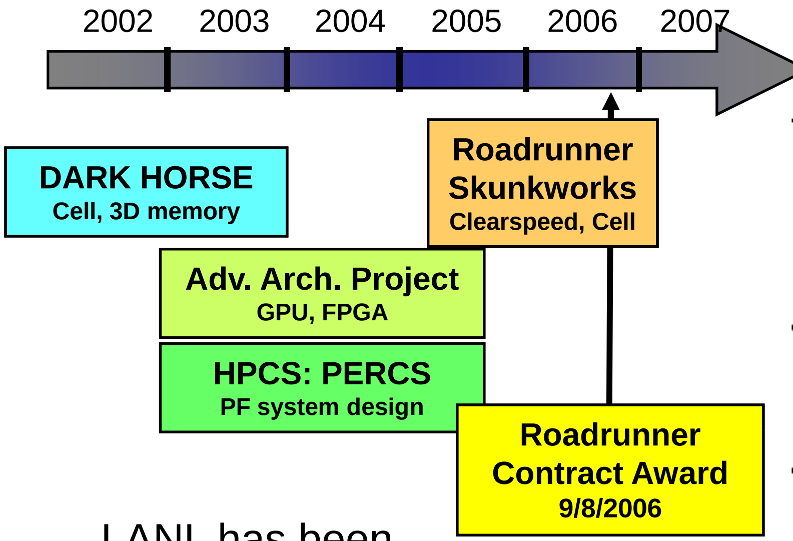
October 2006 Unleashing the Processing Powerhouse

nVidia G80 - 2006



Taken from publicly available information

Roadrunner is on a different path to a petascale



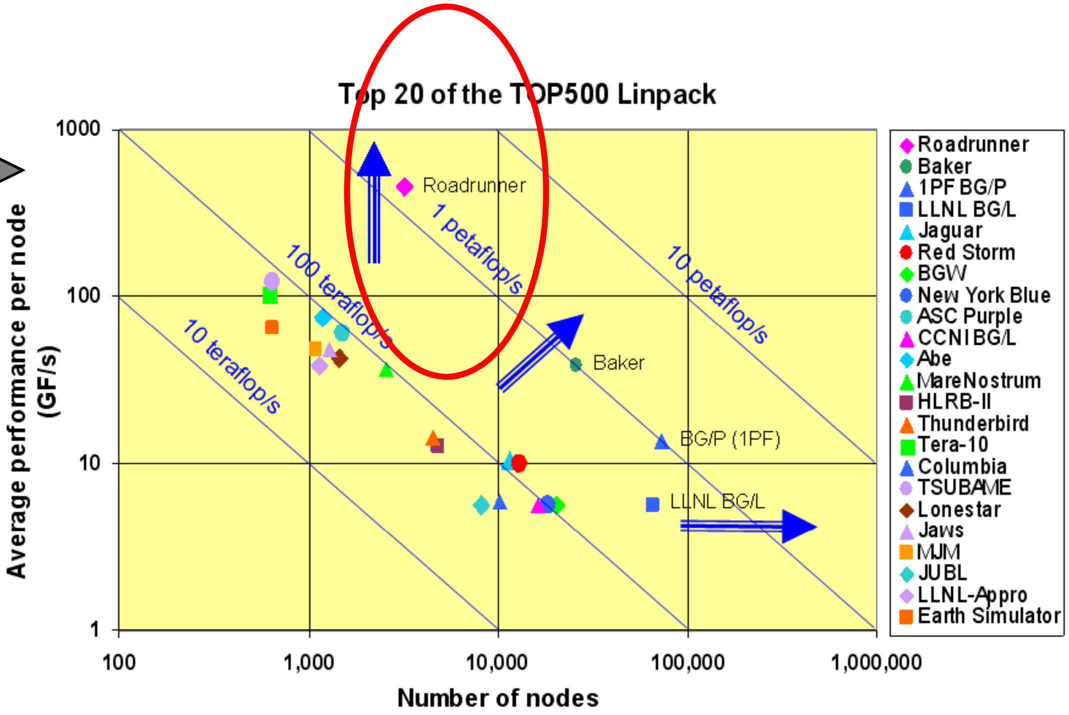
DARK HORSE
Cell, 3D memory

Roadrunner Skunkworks
Clearspeed, Cell

Adv. Arch. Project
GPU, FPGA

HPCS: PERCS
PF system design

Roadrunner Contract Award
9/8/2006



LANL has been looking at hybrid & petascale computing for some time

Cell is fast
Cell is energy efficient
Cell is commodity
Cell brings heterogeneity
Cell brings fine-scale paralleism

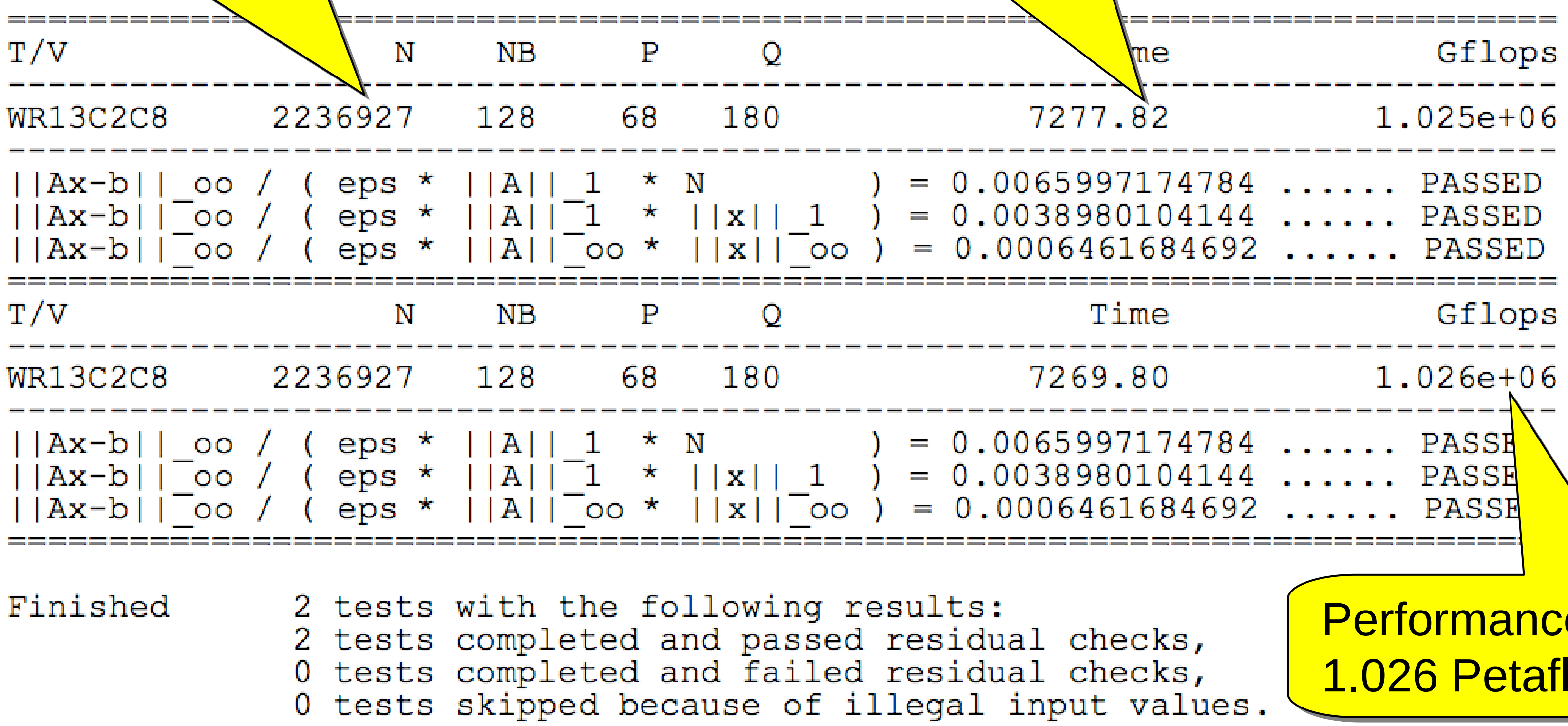
A Roadrunner is born

IBM built hybrid nodes in Rochester, MN and assembled the system in Poughkeepsie, NY

Roadrunner broke the 1 Petaflop/s mark on May 26th, 2008

Matrix: ~5 trillion entries

Calculation: ~2 hours



```
================================================================================
T/V                N     NB      P      Q                 me          Gflops
--------------------------------------------------------------------------------
WR13C2C8      2236927   128     68    180            7277.82       1.025e+06
--------------------------------------------------------------------------------
||Ax-b||_oo / ( eps * ||A||_1  * N        ) = 0.0065997174784 ...... PASSED
||Ax-b||_oo / ( eps * ||A||_1  * ||x||_1  ) = 0.0038980104144 ...... PASSED
||Ax-b||_oo / ( eps * ||A||_oo * ||x||_oo ) = 0.0006461684692 ...... PASSED
================================================================================
T/V                N     NB      P      Q               Time          Gflops
--------------------------------------------------------------------------------
WR13C2C8      2236927   128     68    180            7269.80       1.026e+06
--------------------------------------------------------------------------------
||Ax-b||_oo / ( eps * ||A||_1  * N        ) = 0.0065997174784 ...... PASSE
||Ax-b||_oo / ( eps * ||A||_1  * ||x||_1  ) = 0.0038980104144 ...... PASSE
||Ax-b||_oo / ( eps * ||A||_oo * ||x||_oo ) = 0.0006461684692 ...... PASSE
================================================================================

Finished      2 tests with the following results:
              2 tests completed and passed residual checks,
              0 tests completed and failed residual checks,
              0 tests skipped because of illegal input values.
```

Performance:
1.026 Petaflop/s

Only 3 days after the full machine was finally assembled!

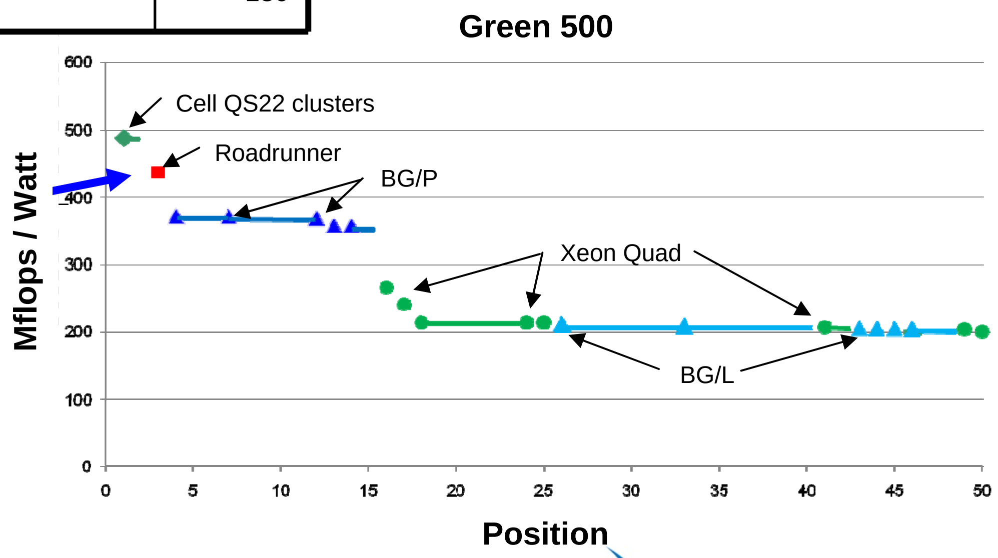
Roadrunner is a TOP performer!

#	SITE	SYSTEM	TF/sec
1	**DOE/NNSA/LANL** **United States**	**Roadrunner, QS22/LS21** **IBM**	**1026**
2	DOE/NNSA/LLNL United States	Blue Gene/L IBM	478
3	Argonne National Laboratory United States	Blue Gene/P IBM	450
4	Texas Adv. Comp. Center United States	SunBlade Opteron IB Cluster Sun	326
5	DOE/ORNL United States	Jaguar, XT4-QuadCore Cray	205
6	Forschungszentrum Juelich Germany	Blue Gene/P IBM	180

#1 on the TOP500

From June 2008 Top 500 List

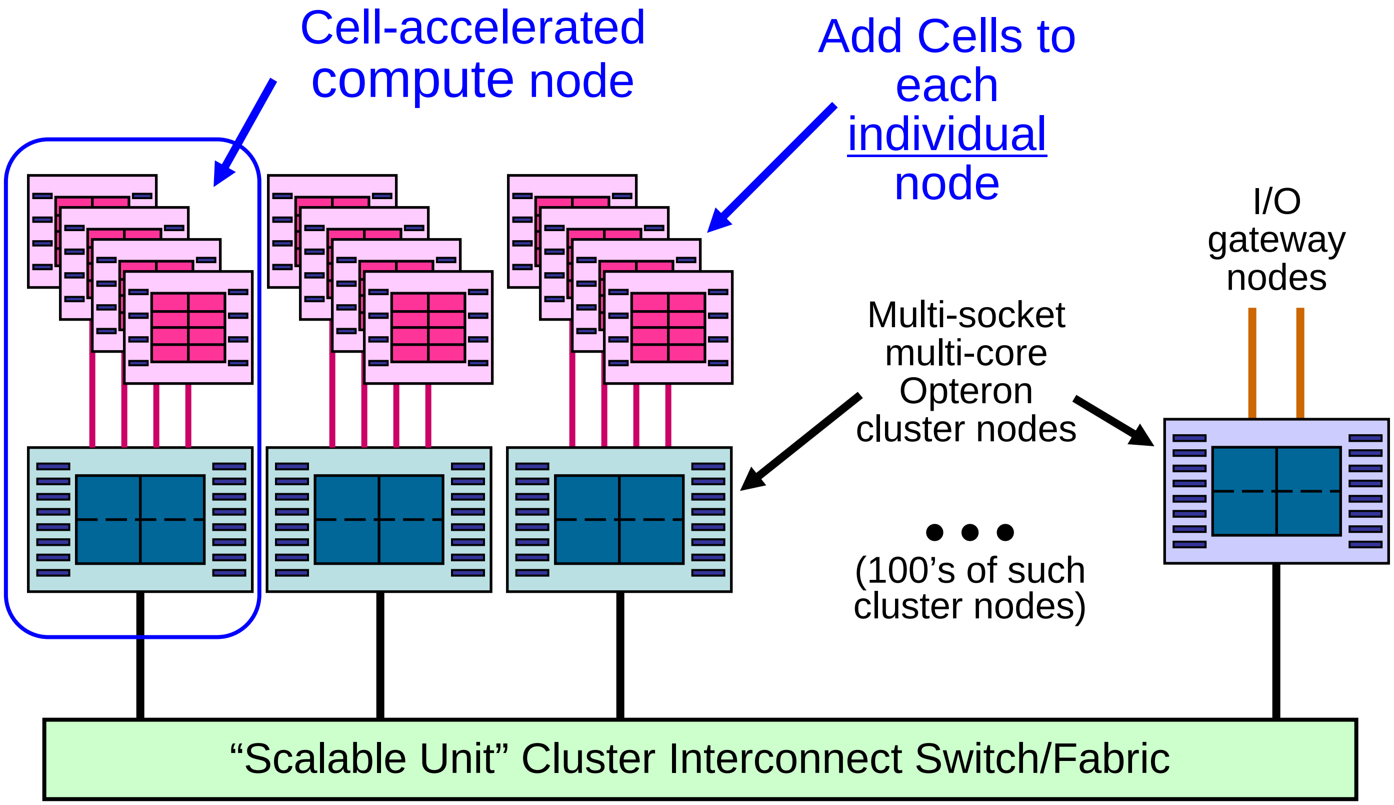
#3 on the Green500



Green 500

Roadrunner System Configuration

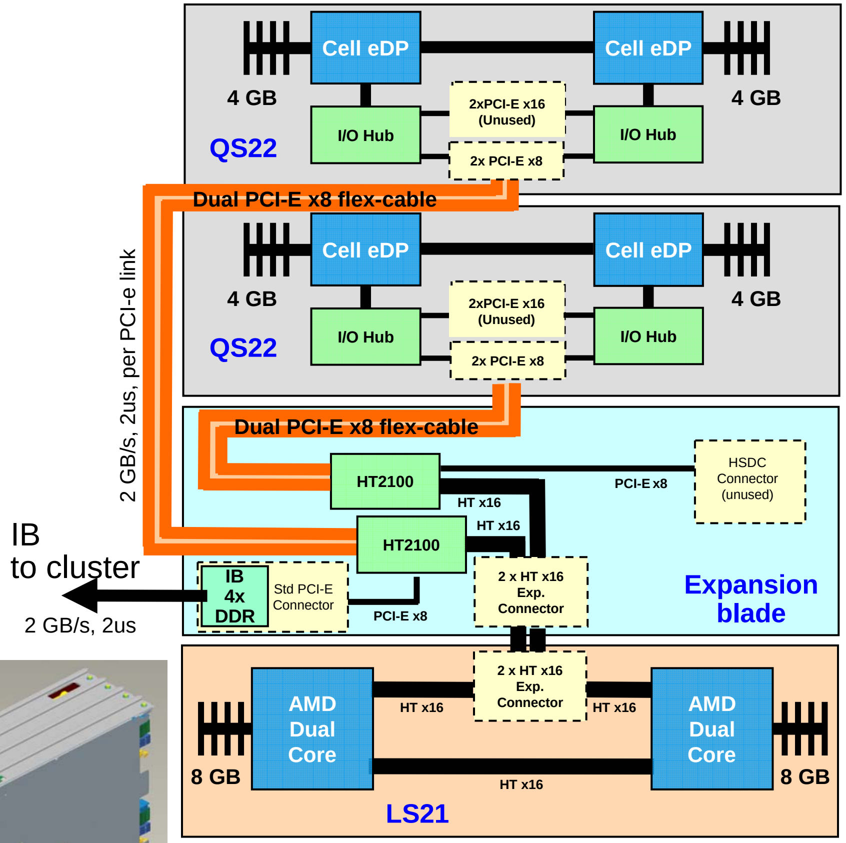
Roadrunner Phase 3 is Cell-accelerated, not a cluster of Cells

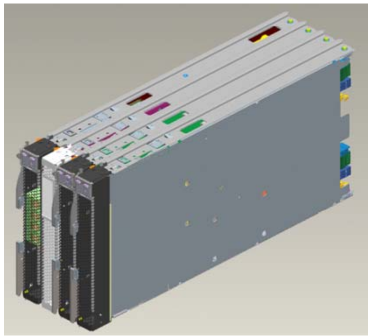

Cell-accelerated compute node
Add Cells to each individual node
I/O gateway nodes
Multi-socket multi-core Opteron cluster nodes
• • •
(100's of such cluster nodes)
"Scalable Unit" Cluster Interconnect Switch/Fabric

Node-attached Cells is what makes Roadrunner different!

A Roadrunner TriBlade node integrates Cell and Opteron blades

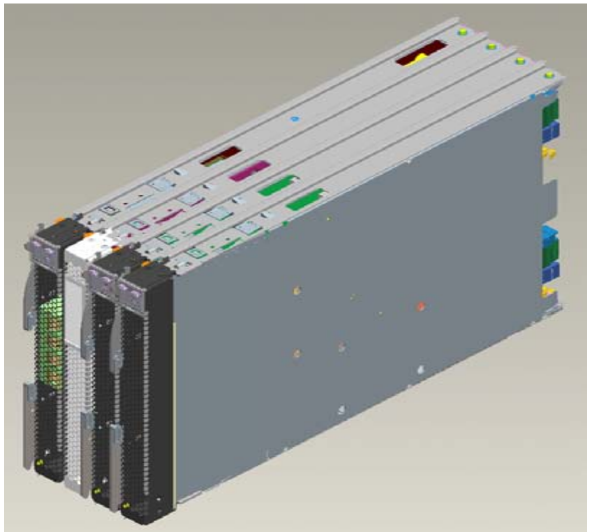
- QS22 is an IBM Cell blade containing two new enhanced double-precision (eDP/PowerXCell™) Cell chips
- Expansion blade connects two QS22 via four PCI-e x8 links to LS21 & provides the node’s ConnectX IB 4X DDR cluster attachment
- LS21 is an IBM dual-socket Opteron blade
- 4-wide IBM BladeCenter packaging
- Roadrunner Triblades are completely diskless and run from RAM disks with NFS & Panasas only to the LS21
- Node design points:
  - *One Cell chip per Opteron core*
  - *~400 GF/s double-precision & ~800 GF/s single-precision*
  - *16 GB Opteron memory PLUS 16 GB Cell memory*
  - *1 PCI-E x8 to each Cell*

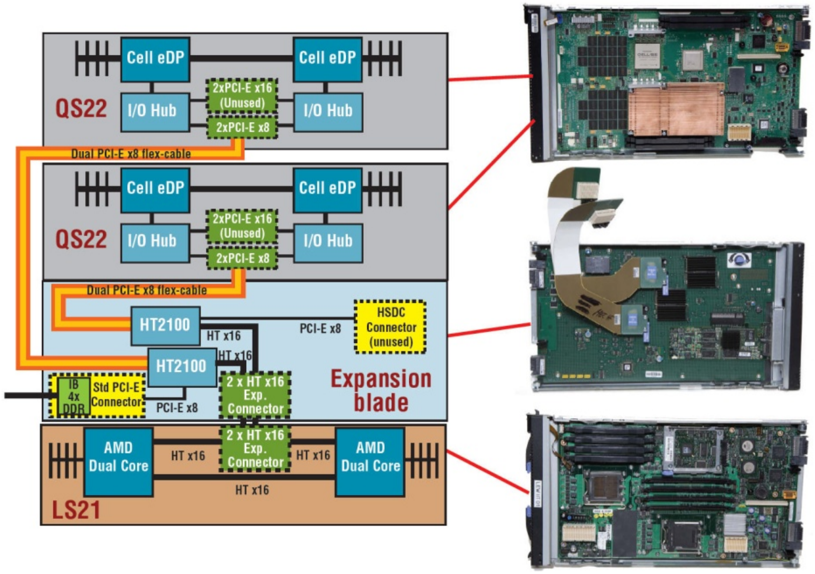




Design point:
One Cell per Opteron core

A Roadrunner TriBlade node integrates Cell and Opteron blades



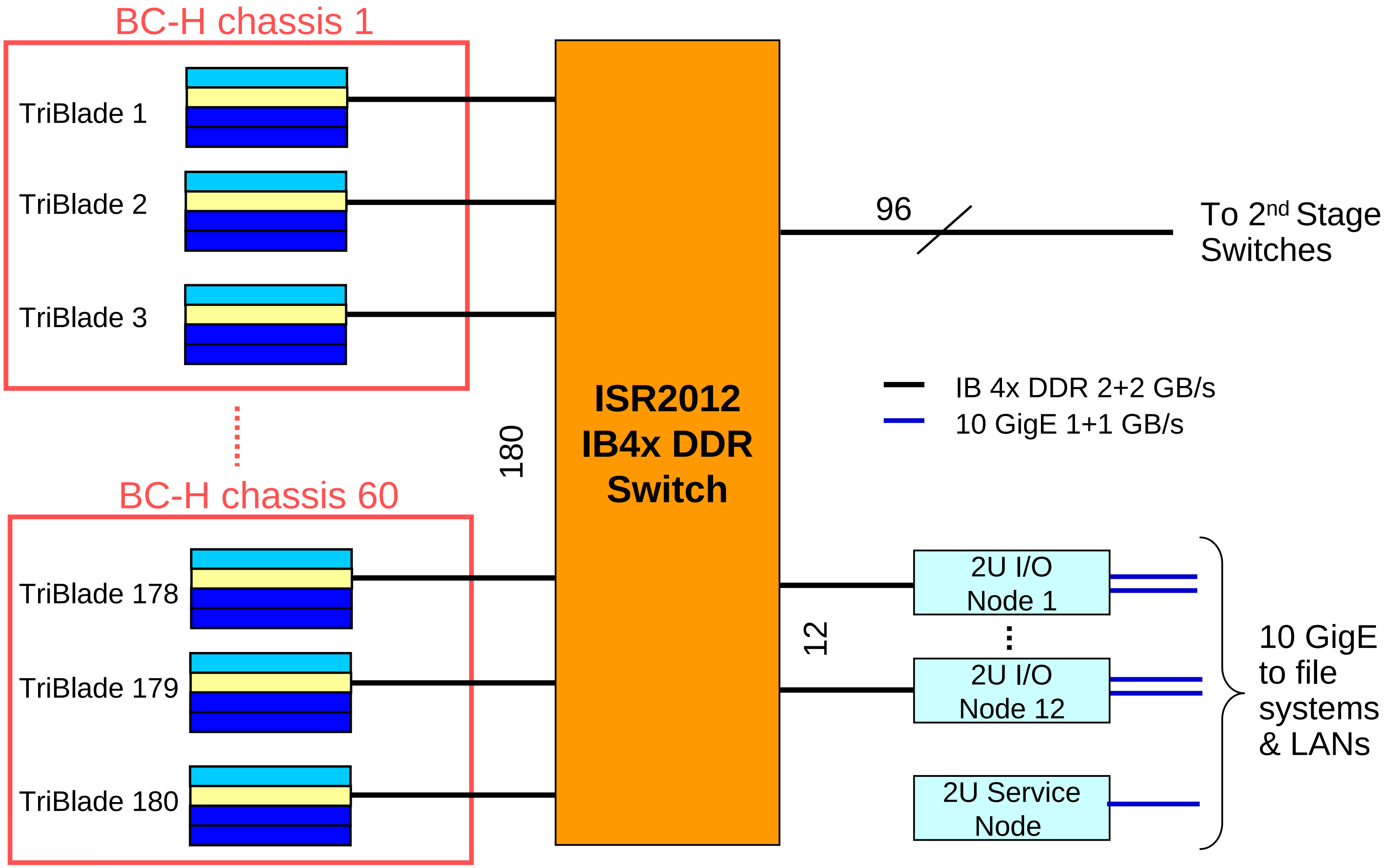


Two QS22's with 2 Cells each

Expansion blade

LS21 with two dual-core Opterons

A Connected Unit (CU) forms a building block



A Connected Unit (CU) is a powerful cluster

Connected Unit Specifications:

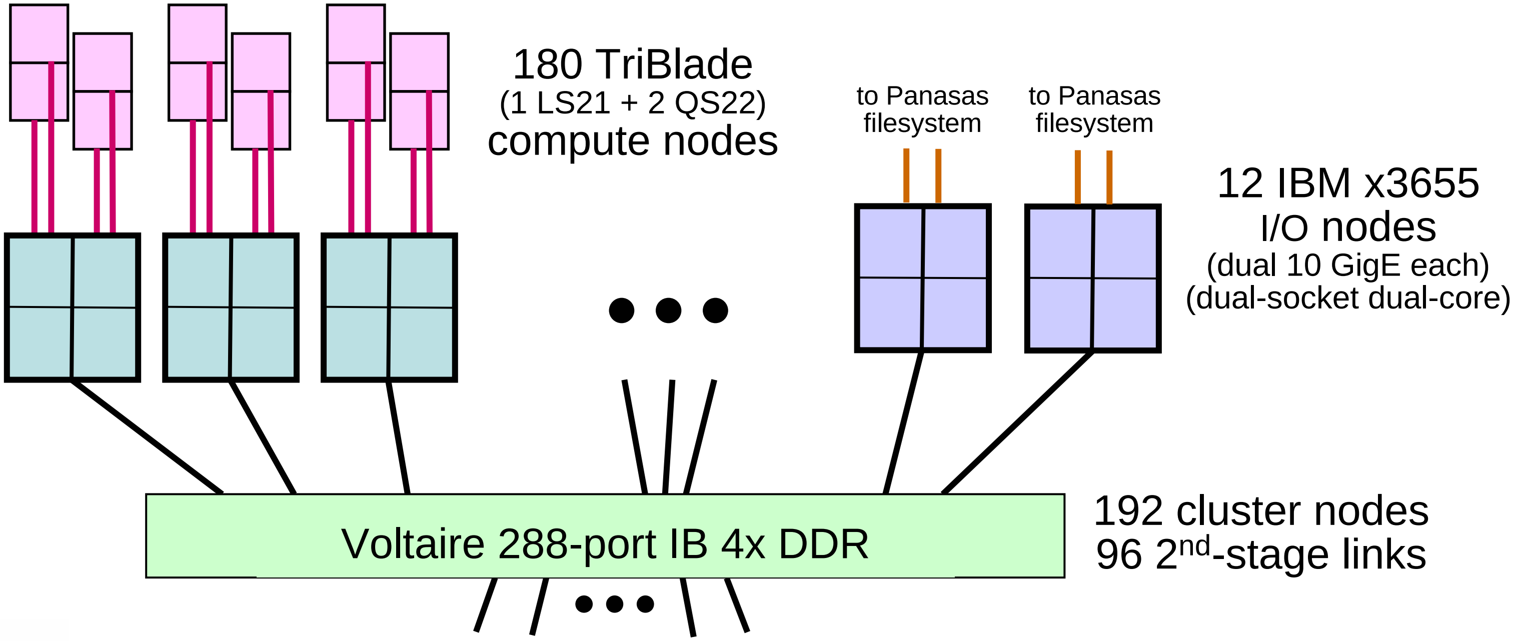
360 1.8 GHz dual-core Opterons
- 2.59 TF DP peak Opteron
- 2.88 TB Opteron memory

24 2.6 GHz dual-core Opterons
- in I/O nodes

720 PowerXCell chips
- **73.7 TF DP peak Cell**
- 2.88 TB Cell memory
- 18.4 TB/s Cell memory BW

192 IB 4X DDR cluster links
- 768 GB/s aggregate BW (bi-dir)
- 384 GB/s bi-section BW (bi-dir)

24 10 GigE I/O links on 12 I/O nodes
- 24 GB/s aggregate I/O BW (uni-dir) (IB limited)



180 TriBlade
(1 LS21 + 2 QS22)
compute nodes

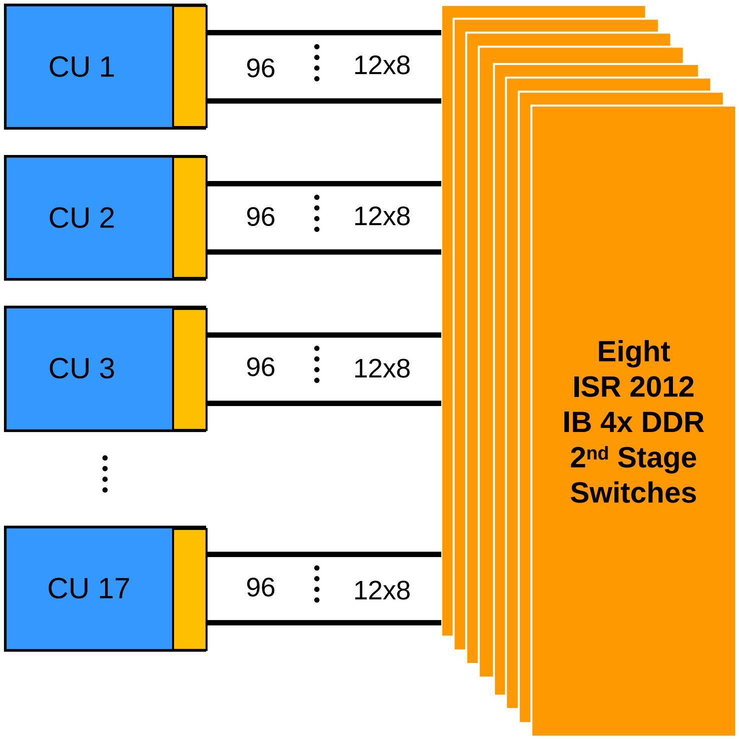
to Panasas filesystem

to Panasas filesystem

12 IBM x3655
I/O nodes
(dual 10 GigE each)
(dual-socket dual-core)

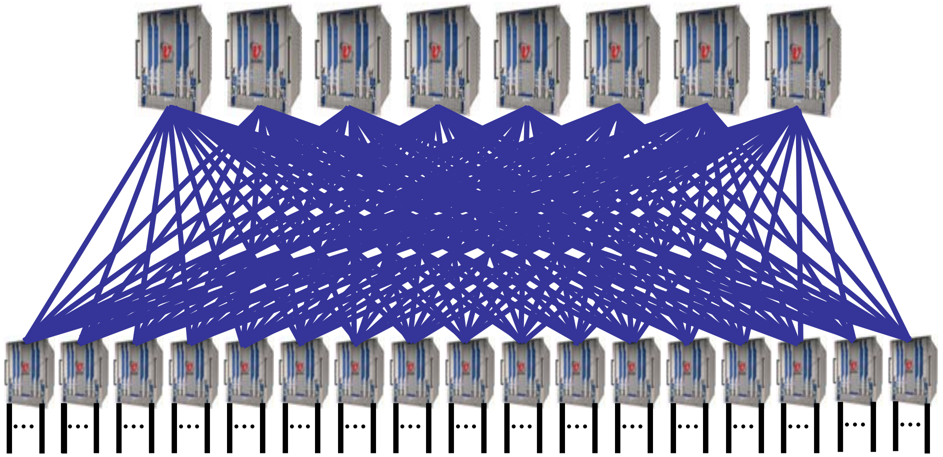
192 cluster nodes
96 2nd-stage links

Now build a cluster-of-clusters…



Eight
ISR 2012
IB 4x DDR
2nd Stage
Switches

2nd–stage switches form
a half-bandwidth fat-tree



17 CUs with CU switches, 3264 IB nodes

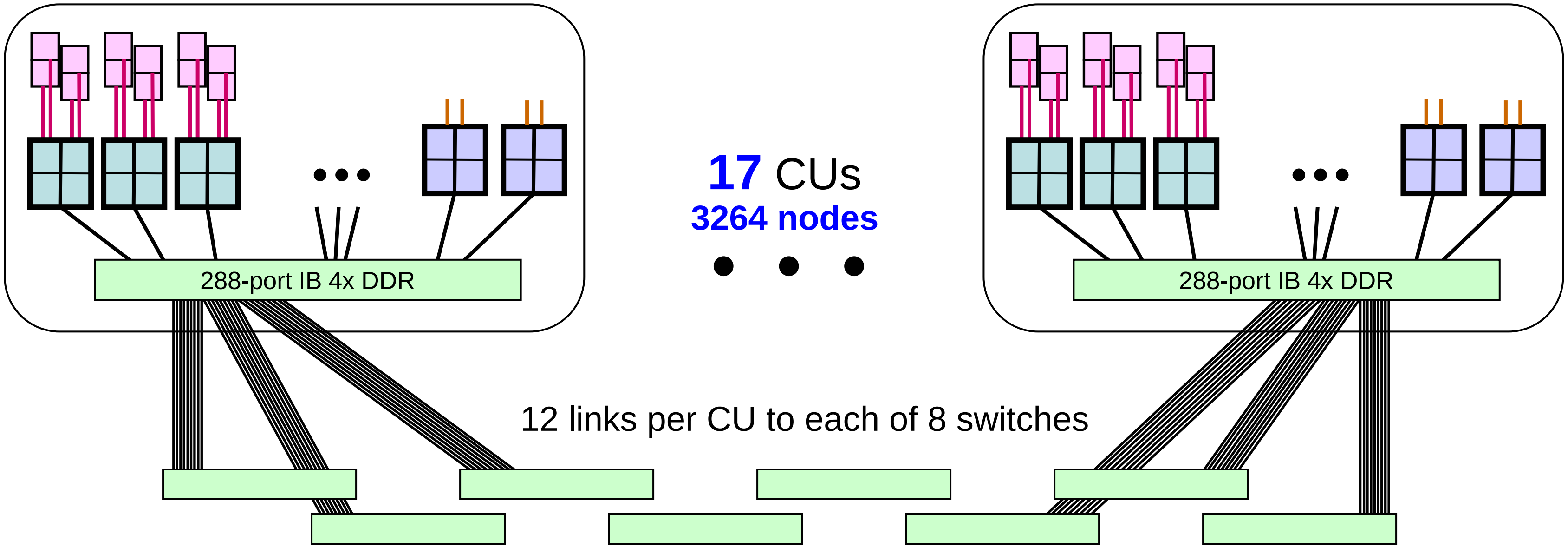
Extra 2nd–stage switch
ports allow expansion
up to 24 CUs

Roadrunner is a hybrid petascale system of modest size delivered in 2008

Connected Unit cluster
180 compute nodes w/ Cells
12 x3655 I/O nodes

12,240 PowerXCell 8i chips ⇒ 1.33 PF, 49 TB
6,120 dual-core Opterons ⇒ 44 TF, 49 TB

*** I/O nodes not counted**



288-port IB 4x DDR

17 CUs
3264 nodes

288-port IB 4x DDR

12 links per CU to each of 8 switches

Eight 2nd-stage 288-port IB 4X DDR switches

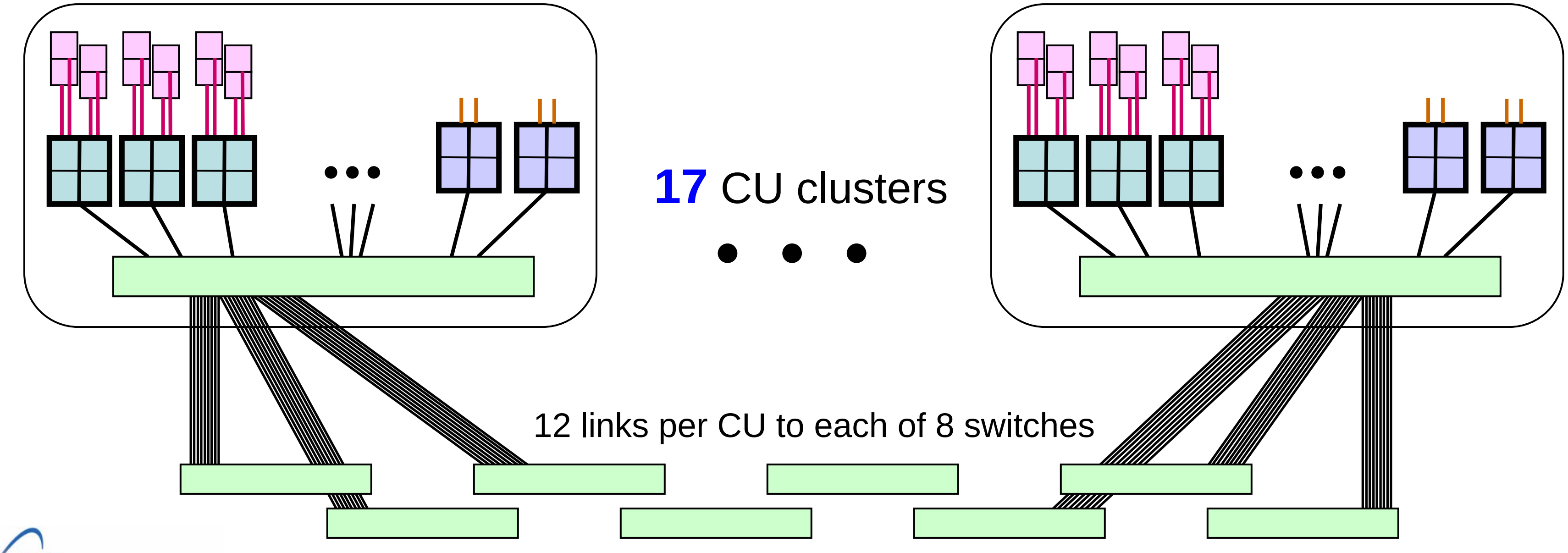
Roadrunner is a petascale system in 2008

Full Roadrunner Specifications:

- 6,120 dual-core Opterons
  - 44.1 TF DP peak Opteron
  - 49 TB Opteron memory
- 408 dual-core Opterons
  - in I/O nodes

- 12,240 PowerXCell 8i chips
  - 1.33 PF DP peak Cell
  - 2.59 PF SP peak Cell
  - 49 TB Cell memory
  - 313 TB/s Cell memory BW

- 3,264 nodes on 2-stage IB 4X DDR
  - 13.1 TB/s aggregate BW (bi-dir) (1st stage)
    - 6.5 TB/s aggregate BW (bi-dir) (2nd stage)
    - 3.3 TB/s bi-section BW (bi-dir) (2nd stage)
- 408 10 GigE I/O links on 204 I/O nodes
  - 408 GB/s aggregate I/O BW (uni-dir) (IB limited)



17 CU clusters

• • •

12 links per CU to each of 8 switches

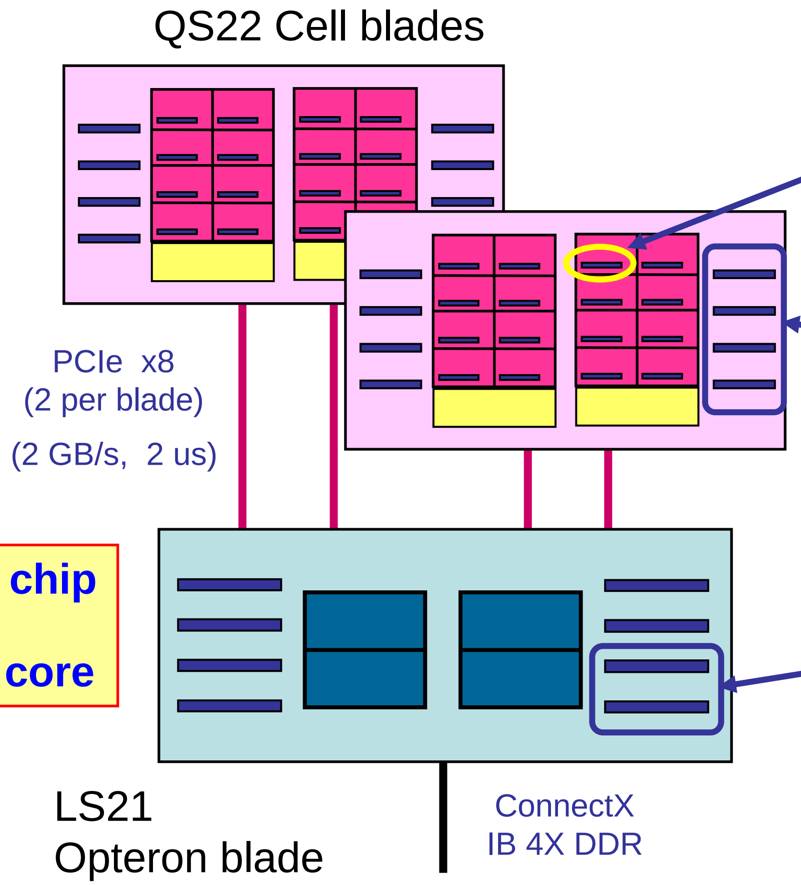
Eight 2nd-stage IB 4X DDR switches

Roadrunner at a glance

- **Cluster of 17 Connected Units (CU)**
  - *12,240 IBM PowerXCell 8i chips*
  - *1.33 Petaflop/s DP peak (Cell)*
  - *1.026 PF sustained Linpack (DP)*
  - *6120 (+408) AMD dual-core Opterons*
  - *44.1 (+4.4) Teraflop/s peak (Opteron)*
- **InfiniBand 4x DDR fabric**
  - *3264 nodes, 2-stage fat-tree; all-optical cables*
  - *Full bi-section BW within each CU*
    - 384 GB/s (bi-directional)
  - *Half bi-section BW among CUs*
    - 3.26 TB/s (bi-directional)
- **~100 TB aggregate memory**
  - *49 TB Opteron (compute nodes)*
  - *49 TB Cell*
- **204 GB/s sustained File System I/O:**
  - *204x2 10G Ethernets to Panasas*
- **Fedora Linux**
  - *On LS21 and QS22 blades*
- **SDK for Multicore Acceleration**
  - *Cell compilers, libraries, tools*
- **xCAT Cluster Management**
  - *System-wide GigEnet network*
- **2.35 MW Power:**
  - *0.437 GF/Watt*
- **Area:**
  - *280 racks*
  - *5200 ft^2*

Programming Concepts

Roadrunner nodes have a memory hierarchy



One Cell chip per Opteron core

256 KB of
“working” memory
(per SPE)

25.6 GB/s
off-SPE BW

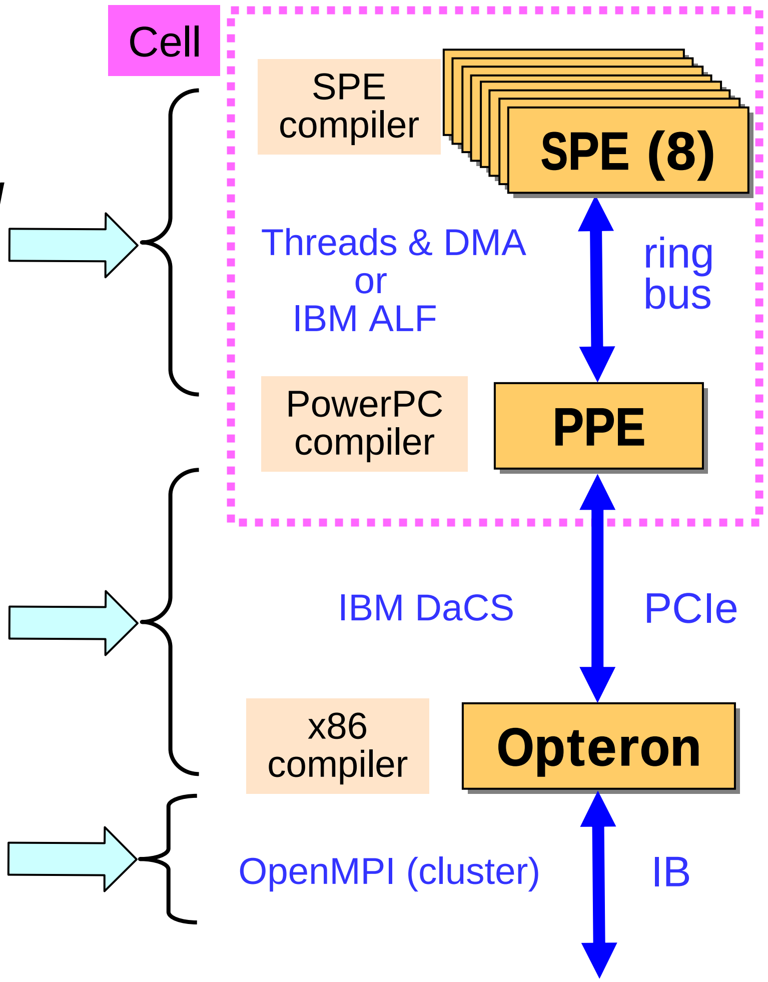
~200 GB/s per
Cell on EIB bus

4 GB of shared memory (per Cell) 25.6 GB/s/chip (w/ 800 DDR2)	8 GB of NUMA shared memory (per blade)	16 GB of distributed memory (per node)
↕	↕	↕
4 GB of memory (per core)	8 GB of shared memory (per socket) 5.4 GB/s/core	16 GB of NUMA shared memory (per node)

“equal memory size” concept

Three types of processors work together

- Parallel computing on Cell
  - *data partitioning & work queue pipelining*
  - *process management & synchronization*
- Remote communication to/from Cell
  - *data communication & synchronization*
  - *process management & synchronization*
  - *computationally-intense offload*
- MPI remains as the foundation

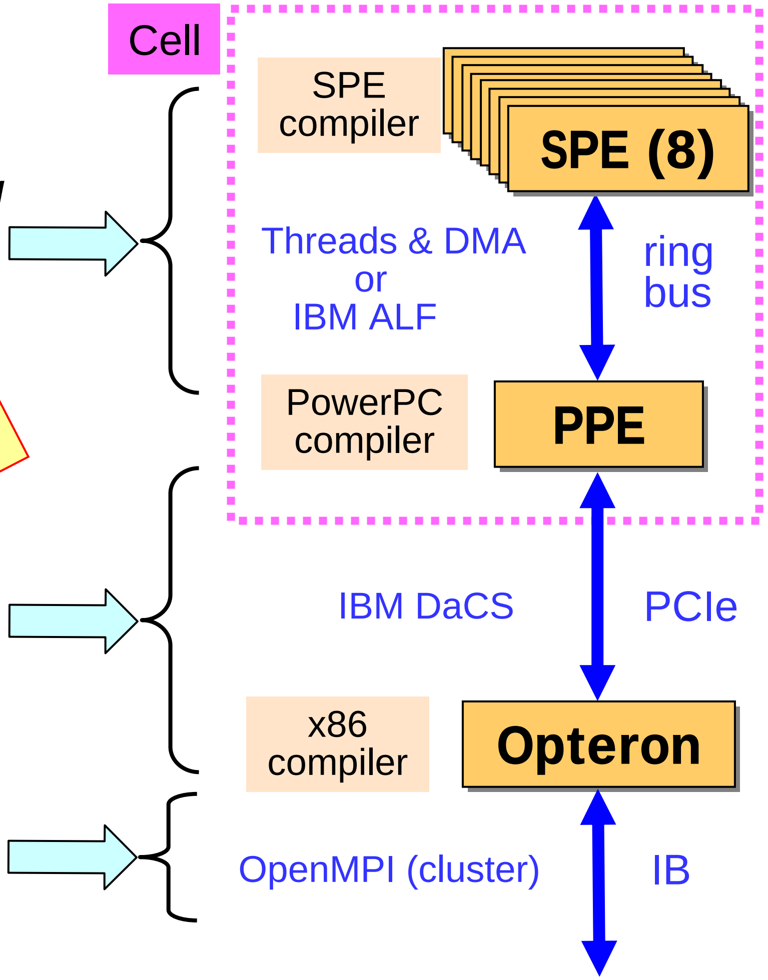


Three types of processors work together

- Parallel computing on Cell
  - *data partitioning & work qu ng*
  - *process management ation*
- Re ication t
  - *munication ation*
  - *ess mana chronization*
  - *computat e offload*
- MPI remains as the foundation

Parallel-in-parallel

This can be done one algorithm at a time in a multi-physics code!



Cell

SPE compiler

SPE (8)

Threads & DMA or IBM ALF

ring bus

PowerPC compiler

PPE

IBM DaCS

PCIe

x86 compiler

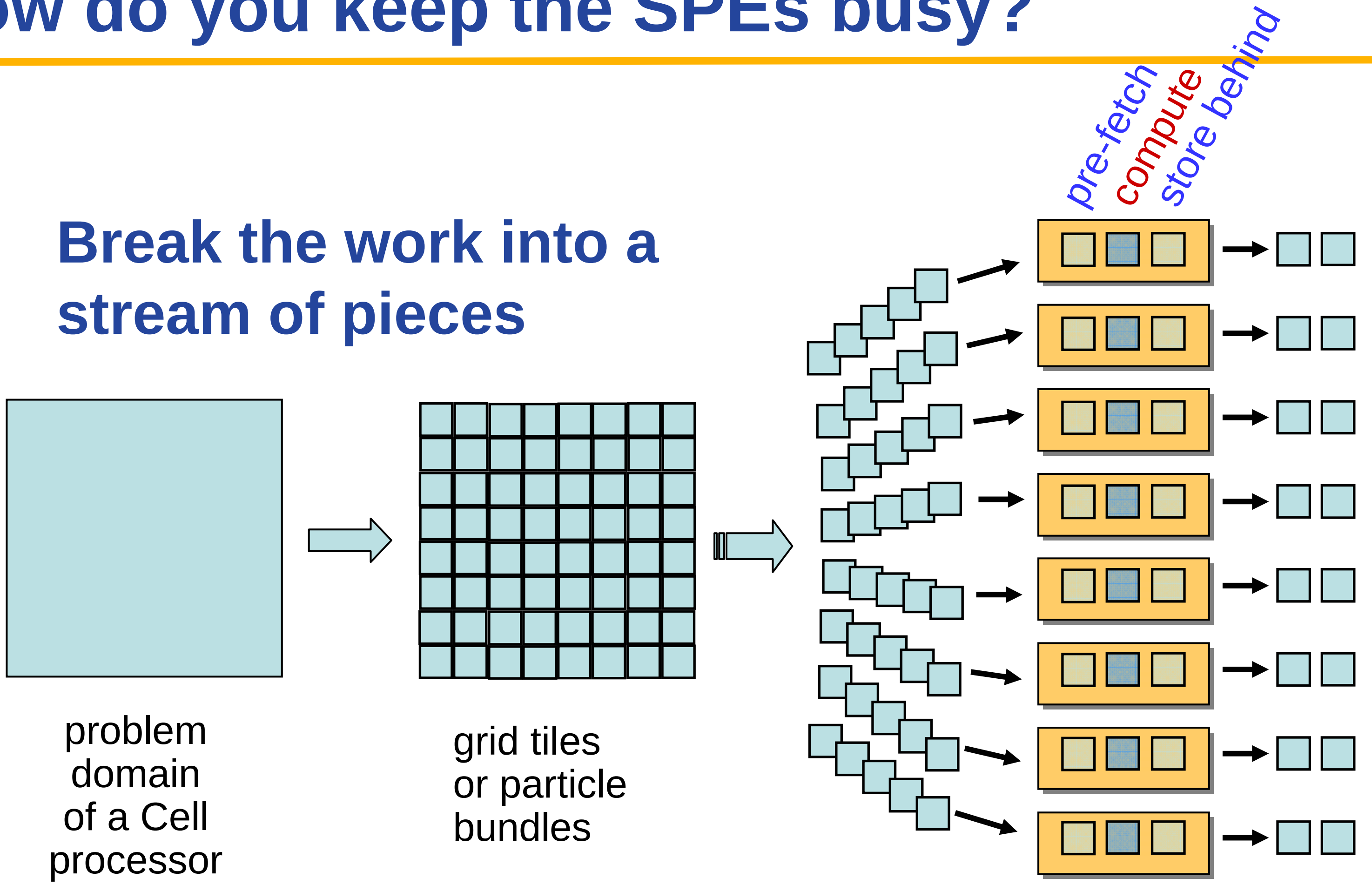
Opteron

OpenMPI (cluster)

IB

How do you keep the SPEs busy?

Break the work into a stream of pieces

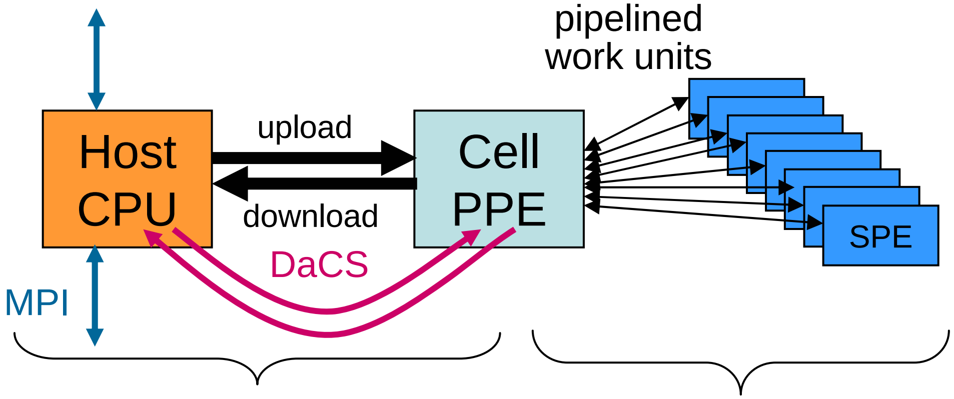


problem domain of a Cell processor

grid tiles or particle bundles

data chunks stream in & out of 8 SPEs using asynch DMAs and multi-buffering

Put it all together: MPI+DaCS+DMA+SIMD



“relay” of DaCS ⇔ MPI messages

Compute & memory DMA transfers are overlapped in HW!

MPI & DaCS can also be fully asynchronous

DMA Get (first prefetch)
Switch work buffers

DMA Get (prefetch)
DMA Wait (complet current)
Compute
DMA Put (store behind)
DMA Wait (previous put)
Switch work buffers

DMA Wait (put)

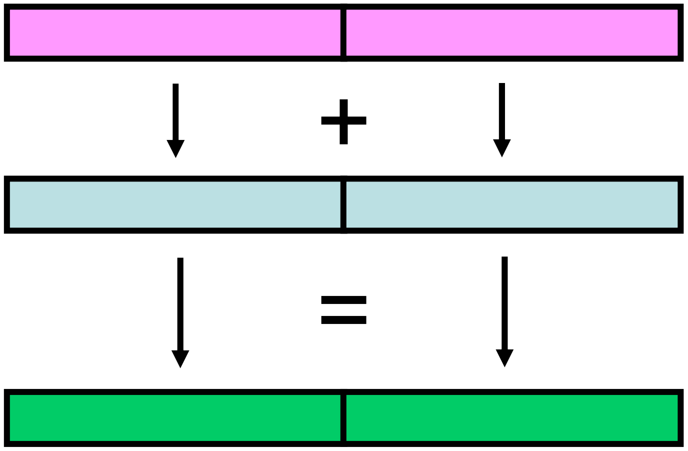
- DMAs are simply block memory transfers
  - *HW asynchronous (no SPE stalls)*
  - *DDR2 memory latency and BW performance*

DMA Get:
```
  mfc_get( LS_addr, Mem_addr, size, tag, 0, 0);
```

DMA Put:
```
  mfc_put( Mem_addr, LS_addr, size, tag, 0, 0);
```

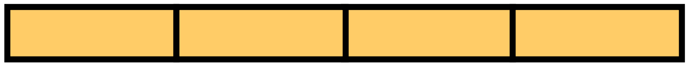
DMA Wait:
```
  mfc_write_tag_mask(1<<tag);
  mfc_read_tag_status_all();
```

Pick data structures & alignment to allow SIMD



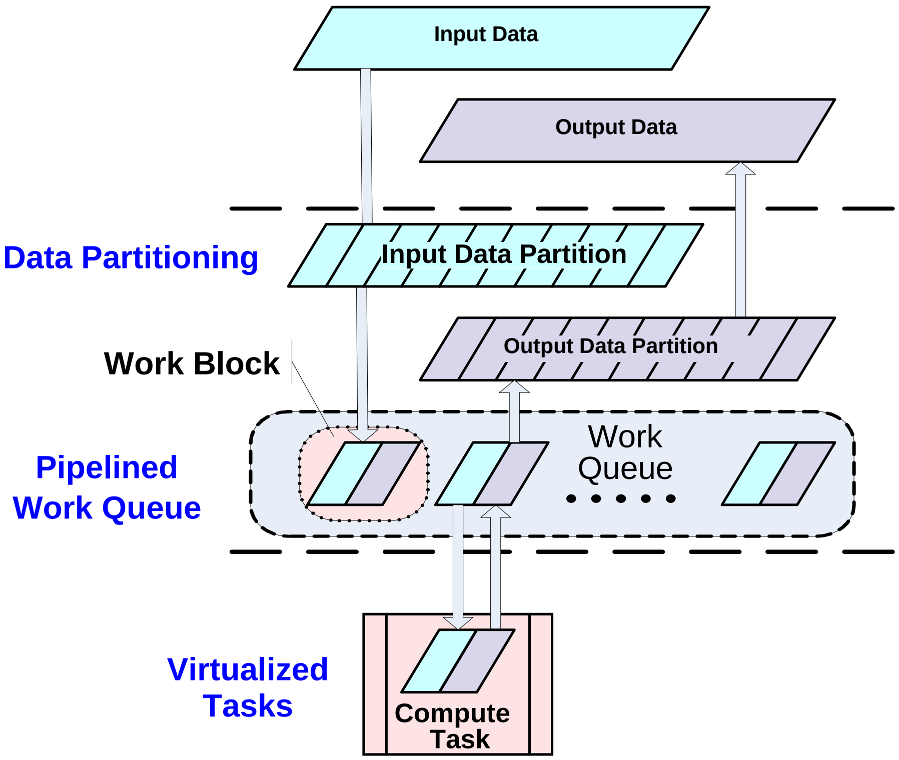
128 bits = 2 doubles
Work on aligned data
c[i] = a[i] + b[i]

Cross aligned operations are really bad!
c[i] = a[i] + a[i+1]



4 singles or integers work similarly at **twice** the performance

IBM-provided ALF is a simple work-queue approach for abstracting parallelism



Input Data

Output Data

Data Partitioning

Input Data Partition

Output Data Partition

Work Block

Pipelined Work Queue

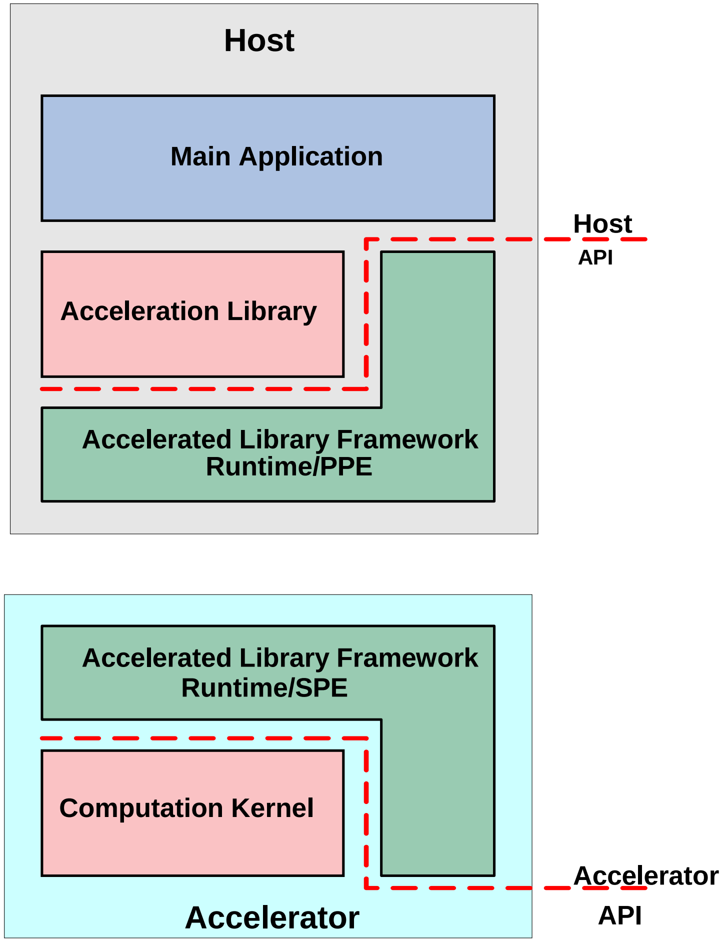
Work Queue

• • • • •

Virtualized Tasks

Compute Task

Accelerated Library Framework



Host

Main Application

Host API

Acceleration Library

Accelerated Library Framework Runtime/PPE

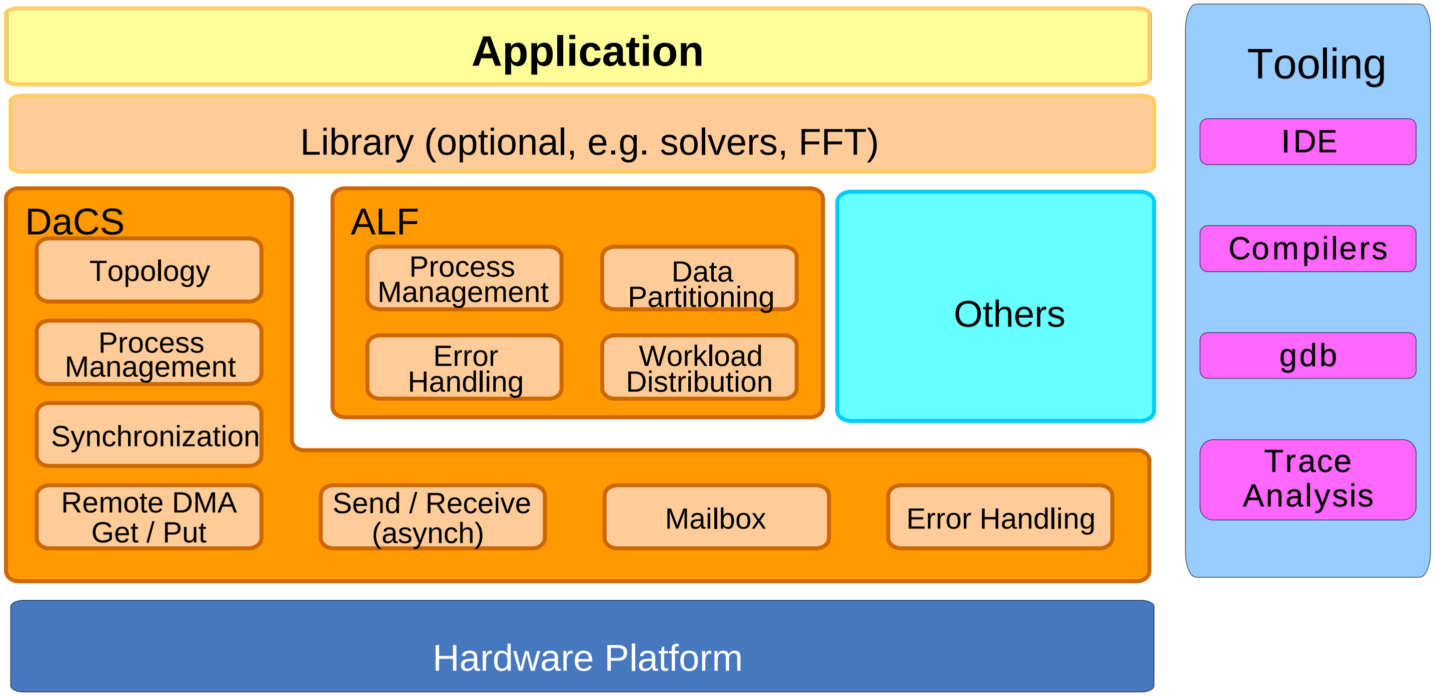
Accelerated Library Framework Runtime/SPE

Computation Kernel

Accelerator API

Accelerator

ALF & DaCS: Broader than Cell & Roadrunner



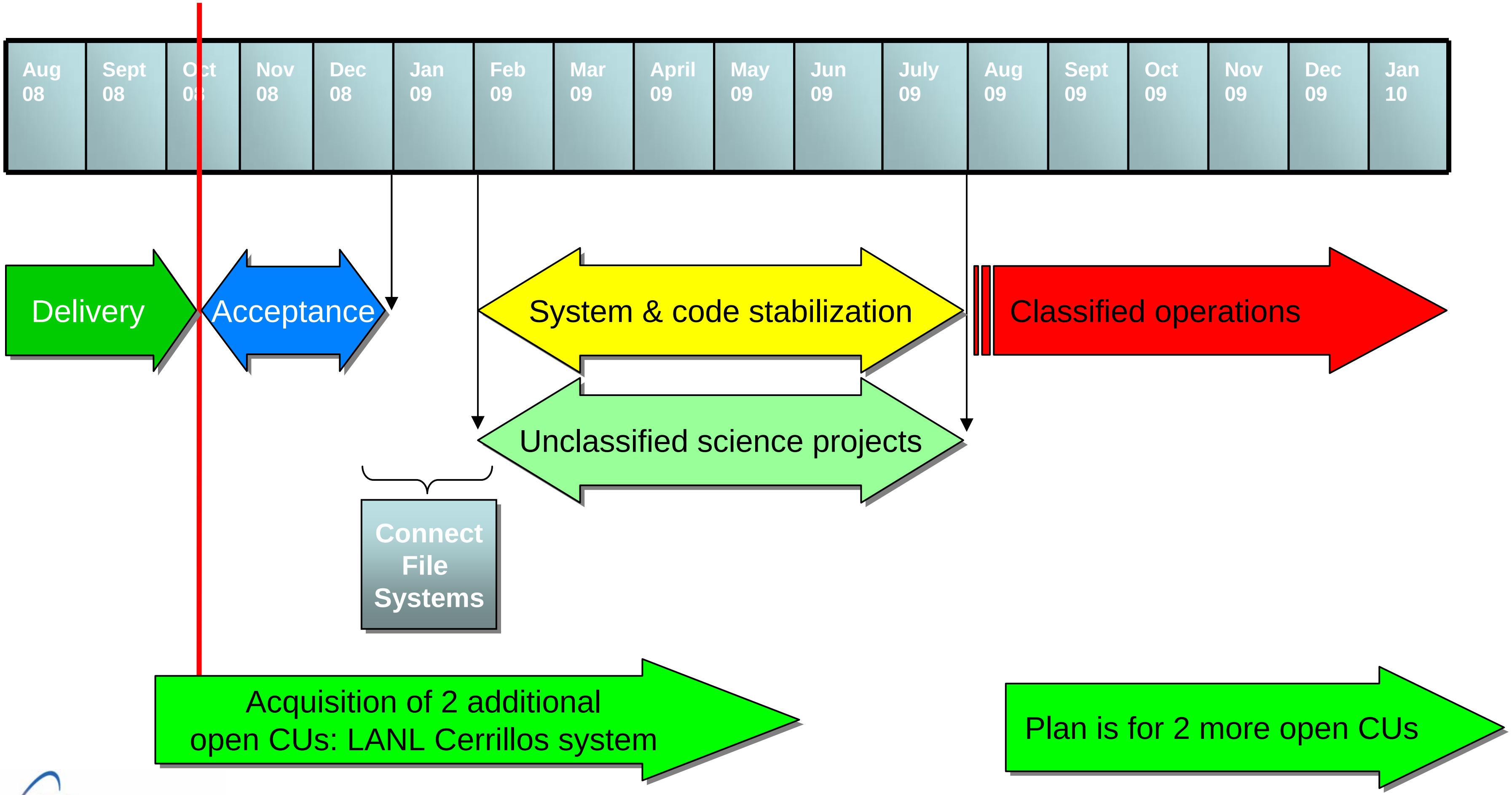
- Designed by IBM & LANL to be HW agnostic
  - Cell PPE+SPEs and also Opterons+direct-SPEs
  - multicore/GPU/Cell, interconnect, even possibly cluster-wide
  - desire technical community participation to extend range

Programming approach has now been demonstrated and is Tractable

- Two levels of parallelism:
  - *node-to-node: MPI & DaCS-MPI-DaCS relay*
  - *within-Cell: threads, pipelined DMAs, & SIMD*
- Large-grain computationally intense portions of code are split off for Cell acceleration within a node process
  - *Usually an entire tree of subroutines*
  - *This is equivalent to "function offload" of entire large algorithms*
- Threaded fine-grained parallelism introduced within the Cell itself
  - *Create many-way parallel pipelined work units for the 8 SPEs*
  - *Good for both multicore/manycore chips and heterogeneous chip trends with dwindling memory bandwidth*
- Communications during Cell computation are possible between Cells via DaCS-MPI-DaCS "relay" approach
- Considerable flexibility and opportunities exist

Roadrunner Status and Future Plans

LANL has two tracks for Open Science



Aug 08 | Sept 08 | Oct 08 | Nov 08 | Dec 08 | Jan 09 | Feb 09 | Mar 09 | April 09 | May 09 | Jun 09 | July 09 | Aug 09 | Sept 09 | Oct 09 | Nov 09 | Dec 09 | Jan 10

Delivery

Acceptance

System & code stabilization

Classified operations

Unclassified science projects

Connect File Systems

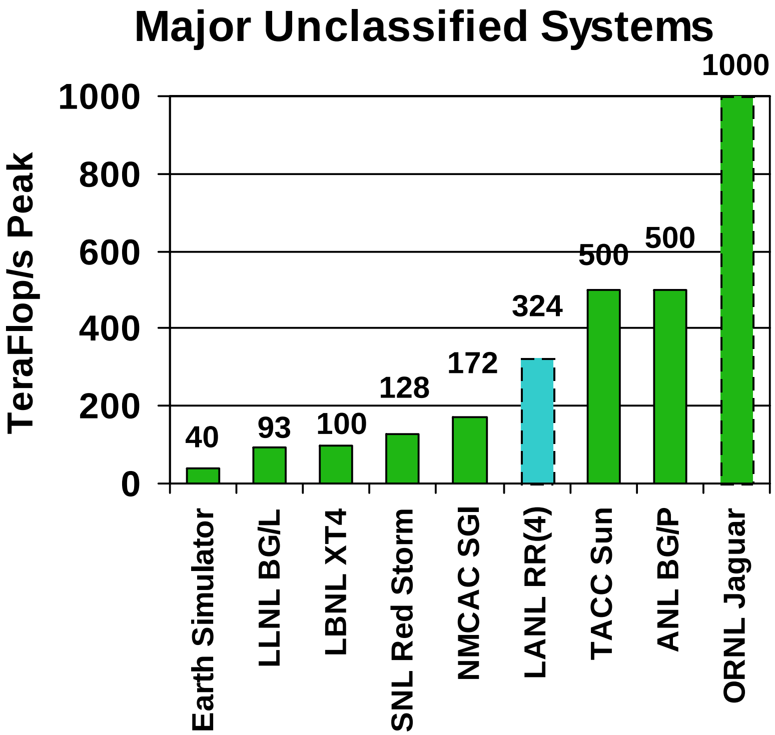
Acquisition of 2 additional open CUs: LANL Cerrillos system

Plan is for 2 more open CUs

LANL has two tracks for Open Science

- Call for open science proposals on full Roadrunner during stabilization
  - *Important side effects*
    - increase the cadre of expert Cell programmers
    - Increase the number of codes that can take advantage of Roadrunner architecture
- There were 29 proposals submitted
  - *Requests for 181 M Cells hours (5x available resources)*
  - *Requests for $9M in LDRD support (3x available resources)*
- Eight projects were selected

Additional LANL open RR resources are required to support open science.



There are very exciting opportunities among the 8 selected proposals for full Roadrunner time.

Kinetic Thermonuclear Burn Studies with VPIC on Roadrunner	VPIC
Multibillion-Atom Molecular Dynamics Simulations of Ejecta Production and Transport using Roadrunner	SPaSM
New frontiers in viral phylogenetics	ML
Three-Dimensional Dynamics of Magnetic Reconnection in Space and Laboratory Plasmas	VPIC
The Roadrunner Universe	MC^3
Implicit Monte Carlo Calculations of Supernova Light-Curves	IMC + Rage
Instabilities-Driven Reacting Compressible Turbulence	CFDNS
Cellulosomes in Action: Peta-Scale Atomistic Bioenergy Simulations	GROMACS
Parallel-replica dynamics study of tip-surface and tip-tip interactions in atomic force microscopy and the formation and mechanical properties of metallic nanowires	SPaSM + PAR-REP
Saturation of Backward Stimulated Scattering of Laser In The Collisional Regime	VPIC

🟩 Indicates new work 🟨 Indicates new + old

The LANL Roadrunner web site is

http://www.lanl.gov/roadrunner/

Roadrunner architecture
Early applications efforts
Upcoming Open Science efforts
Cell & hybrid programming
Computing trends
Related Internet links