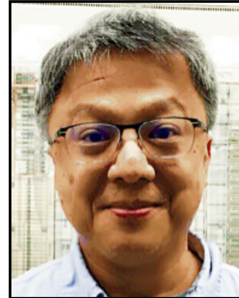
Session 11 Overview: *SRAM*

MEMORY SUBCOMMITTEE

Session Chair:
Jonathan Chang
TSMC, Hsinchu, Taiwan



Associate Chair:
Chun Shiah
Etron, Hsinchu, Taiwan

Subcommittee Chair: ***Leland Chang,*** *IBM, Yorktown Heights, NY*

SRAM continues to be the critical technology enabler for a wide range of applications from low-power to high-performance computing. This session showcases the leading-edge SRAM developments from the semiconductor industry. Intel presents the smallest SRAM bitcell for 10nm technology, with design assist techniques to enable low V_{MIN} operation. Samsung presents the smallest bitcell for 7nm technology and shows a double-write driver technique to further improve V_{MIN}. TSMC demonstrates a 7nm 5GHz L1 cache for high-performance computing.

 978-1-5090-4940-0/18/$31.00 ©2018 IEEE

8:30 AM

11.1 A 23.6Mb/mm² SRAM in 10nm FinFET Technology with Pulsed PMOS TVC and Stepped-WL for Low-Voltage Applications

Z. Guo, Intel, Hillsboro, OR

In Paper 11.1, Intel presents a 23.6Mb/mm² SRAM in 10nm FinFET with the smallest 10nm SRAM bitcell. It adopts column-based transient voltage collapse and a stepped wordline to lower the minimum operation voltage (V_{MIN}).

9:00 AM

11.2 A 7nm FinFET SRAM Using EUV Lithography with Dual Write-Driver-Assist Circuitry for Low-Voltage Applications

T. Song, Samsung Electronics, Hwaseong, Korea

In Paper 11.2, Samsung Electronics presents a 7nm FinFET SRAM using EUV lithography. It adopts a 0.026μm² bitcell and V_{MIN} is improved with a proposed dual write-driver (DWD) scheme in combination with a negative bitline scheme.

9:30 AM

11.3 A 5GHz 7nm L1 Cache Memory Compiler for High-Speed Computing and Mobile Applications

M. Clinton, TSMC, Austin, TX

In Paper 11.3, TSMC presents a 7nm L1 cache memory compiler, which operates at a 5GHz clock frequency. It implements a self-timing scheme with small-signal sensing and a folded architecture to increase the performance.

11.1 A 23.6Mb/mm² SRAM in 10nm FinFET Technology with Pulsed PMOS TVC and Stepped-WL for Low-Voltage Applications

Zheng Guo, Daeyeon Kim, Satyanand Nalam, Jami Wiedemer, Xiaofei Wang, Eric Karl

Intel, Hillsboro, OR

The emergence of cloud computing and big data analytics, accompanied by a sustained growth of battery-powered mobile devices, continues to drive the importance of energy and area efficient CPU and SoC designs. Low-voltage operation remains one of the primary approaches for active power reduction, but SRAM V_{MIN} can limit the minimum operating voltage. Device size quantization continues to be a challenge for compact 6T SRAM design in FinFET technologies, where careful co-optimization of the technology and assist circuit design is required for high-density low-voltage array implementations. This paper presents two SRAM array designs in a 10nm low-power CMOS technology featuring 3rd generation FinFET transistors: a high-density 23.6Mb/mm² array and a low-voltage 20.4Mb/mm² array.

Figure 11.1.1 shows the layout diagrams of a 0.0312μm² high-density 6T SRAM cell (HDC) and a 0.0367μm² low-voltage 6T SRAM cell (LVC) in a 10nm FinFET technology. The HDC utilizes minimum sized devices, with a fin ratio of 1:1:1 (PU:PG:PD), to minimize cell area, while the LVC features a larger PD device (1:1:2) for improved read stability at low voltage. Self-aligned quad patterning (SAQP) is introduced on critical layers to achieve fin pitches down to 34nm and metal pitches down to 36nm with 193nm immersion lithography [1], enabling a 0.62x area scaling of the 6T SRAM cell relative to a 14nm technology [2]. To further maximize density scaling of the 10nm technology, several key architectural features have been added to achieve further array area scaling of the 128kb HDC and LVC macros: achieving a 0.58x and 0.57x reduction relative to 14nm equivalents. Figure 11.1.1 highlights the cell area (μm²) and array area (mm²/Mb) of recently reported 6T SRAM designs from 14nm, 10nm and 7nm technologies [2-4].

Figure 11.1.2 details two architectural features of the 10nm technology for improved density [1]. The first eliminates the need for isolation dummy gates by introducing a minimum isolation step at the source/drain boundary to isolate neighboring transistors by the width of a single gate. The second enables the placement of gate contacts over active transistors, thus eliminating the need for gate extension over isolation to land contacts. The tables in Fig. 11.1.2 summarize the area scaling of critical SRAM periphery circuits, and the array efficiency and density of 128kb SRAM macros in 14nm and 10nm technologies. The combination of single-gate isolation, enabling contacts over active gates, along with the improved pitch scaling of critical interconnect layers has enabled aggressive area scaling of critical SRAM peripheral logic from 14nm to 10nm with minimum fin depopulation. As a result, a 77.1% array efficiency and a 23.6Mb/mm² density are achieved for a 128kb HDC macro: a 5.4% area efficiency improvement over a comparable 14nm design. A 78.4% array efficiency and a 20.4Mb/mm² density are achieved for a 128kb LVC macro: a 6.8% area efficiency improvement over a comparable 14nm design.

Wordline underdrive (WLUD) is used to improve the low-voltage read and half-select stability of an SRAM cell: trading off performance for V_{MIN} [2]. To minimize the impact of interconnect resistance on WL voltage uniformity between different rows of the decoder, WLUD PMOS devices are implemented locally in the WL driver using a matched layout and routing across neighboring rows. To improve the low-voltage write margin, a column-based transient voltage collapse (TVC) scheme is employed to weaken the PU transistor during a write [2]. In this work, a PMOS device (PWR) is used to discharge the memory cell supply (V_{CS}). Compared to an NMOS device, a PMOS device improves V_{CS} control, but at the cost of discharge speed. V_{CS} can be regulated by PMOS bias devices (PB[1:0]), as is illustrated in Fig. 11.1.3. To minimize write energy overhead, a pulsed V_{CS} collapse can be applied with no bias current [2]. To avoid half-select instability along the column, due to a low V_{CS}, careful control of the TVC pulse must be implemented across a range of array configurations when using pulsed TVC. Since the PMOS transistor drive strength degrades super-linearly with a falling V_{CS} in this configuration, wider TVC pulses can be applied without requiring a bias current to avoid half-select instability. V_{CS} sensitivity to the TVC pulse-width and/or array configuration is also reduced, and can be further adjusted by tuning the PMOS transistor V_t. If PMOS bias is required, V_{CS} can be determined by the voltage division across PWR and PB[1:0]. The resulting voltage level correlates well to the write margin, compared to an NMOS TVC that can produce a higher V_{CS} under process skew with a lower NMOS:PMOS drive ratio, where the SRAM write margin is degraded.

While WLUD is effective for enhancing read stability, it degrades the write margin. To independently improve the 6T cell's read and write V_{MIN}, a stepped-wordline (S-WL) scheme [5,6] is implemented to complement the pulsed PMOS TVC write assist. Figure 11.1.3 details the design of the S-WL and PMOS TVC in a 128kb SRAM macro. WLUD pulses (WLUDPULSE[2:0]) are generated from static WLUD bias controls (WLBIAS[2:0]) and the read/write clock. WL suppression is first enabled to create a sufficient BL separation that reinforces cell stability, before WLUDPULSE[2:0] are adjusted to restore the WL to a higher voltage level. Since the required BL differential needed to improve read stability is higher than the voltage sensing margin, the read performance is not impacted by S-WL operation. To maximize the effectiveness of the TVC write assist, the TVC pulse is delayed to align it with WL restoration. To minimize interconnect delay along the WLCLK# and WLUDPULSE paths, local buffers are implemented to drive across the 32b sections of the 256b decoder. This reduces the distributed WLCLK# and WLUDPULSE gate loading by 8×, while maintaining the same logic depth for WL generation. Control logic for S-WL is implemented in the timer/control region with a negligible area overhead.

Figure 11.1.4 shows the simulation waveforms during a write cycle using static WLUD, no WLUD and S-WL. With static WLUD, WL is suppressed for the duration of the WL pulse to maintain cell stability while degrading the write margin. Turning off WLUD improves write margin, but compromises read stability. When S-WL is enabled, the WL is suppressed during the first phase to maintain cell stability. After a sufficient BL separation is achieved, WLBIAS#[2:0] are adjusted to raise the WL voltage, aligned to the TVC pulse, to improve the write margin.

Figure 11.1.5 shows the measured voltage-frequency shmoos of the HDC SRAM using pulsed PMOS TVC write assist, complemented by static WLUD, no WLUD and S-WL. The WL voltage level during the first phase of S-WL matches the static WLUD level. S-WL enables an 80mV and a 100mV improvement to V_{MIN}, compared to a static WLUD and no WLUD. Compared to S-WL operation, array performance is improved with no WLUD due to the higher WL voltage over the entire WL pulse, but V_{MIN} increases due to degraded read stability. In contrast, array performance for static WLUD is limited by the suppressed WL write operation. Figure 11.1.6 summarizes the write and read V_{MIN} measurements for HDC and LVC using a static WLUD and S-WL, with pulsed PMOS TVC. Similar to the voltage-frequency measurements, the WL voltage level during the first phase of S-WL matches the static WLUD level, as determined by the memory cell read stability requirement. S-WL operation enables a 150mV write V_{MIN} improvement for HDC at the 90th percentile and a 60mV write V_{MIN} improvement for LVC at the 90th percentile, without degrading the read V_{MIN}. Decreased write V_{MIN} improvement is observed for LVC due to the reduced WLUD applied. A die micrograph of the 10nm test vehicle with 72Mb of LVC SRAM and 54Mb of HDC SRAM is shown in Fig. 11.1.7.

References:

[1] C. Auth, et al., "A 10nm High Performance and Low-Power CMOS Technology Featuring 3rd Generation FinFET Transistors, Self-Aligned Quad Patterning, Contact over Active Gate and Cobalt Local Interconnects", *IEDM*, 2017.
[2] E. Karl, et al., "A 0.6V, 1.5GHz 84Mb SRAM Design in 14nm FinFET CMOS Technology", *ISSCC,* pp. 310-311, 2015.
[3] K.-I. Seo, et al., "A 10nm Platform Technology for Low Power and High Performance Application Featuring FINFET Devices with Multi Workfunction Gate Stack on Bulk and SOI", *Symp. VLSI Tech.*, pp. 12-13, 2014.
[4] J. Chang, et al., "A 7nm 256Mb SRAM in High-K Metal-Gate FinFET Technology with Write-Assist Circuitry for Low-V_{MIN} Applications", *ISSCC*, pp. 206-207, 2017.
[5] K. Takeda, et al., "Multi-step Word-line Control Technology in Hierarchical Cell Architecture for Scaled-down High-density SRAMs", IEEE Symp. *VLSI Circuits,* pp. 101-102, 2010.
[6] J. Chang, et al., "A 20nm 112Mb SRAM in High-κ Metal-Gate with Assist Circuitry for Low-Leakage and Low-V_{MIN} Applications", *ISSCC*, pp. 316-317, 2013.

 978-1-5090-4940-0/18/$31.00 ©2018 IEEE

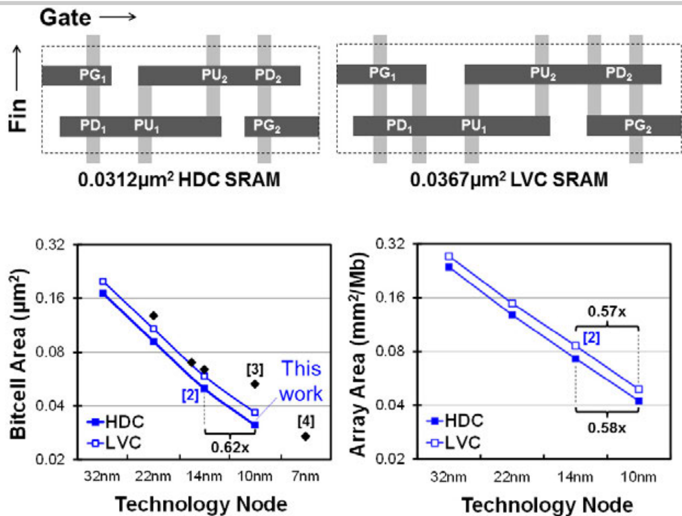
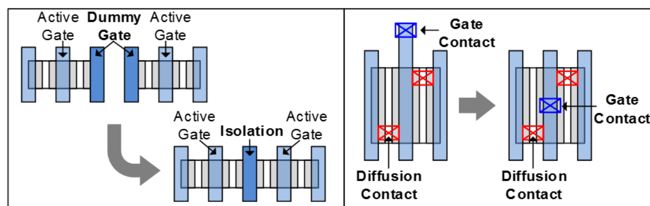


Figure 11.1.1: (top) 10nm HDC and LVC 6T SRAM cells. (bottom) Bitcell and array area scaling trends.



Periphery Logic Scaling vs. 14nm:

Circuit	Fin Count Scaling	Area Scaling
RD/WR MUX	0.91x	0.37x
Sense Amplifier	0.95x	0.42x
Row Decoder	0.88x	0.51x
Timer/Control	0.79x	0.38x

Subarray Summary:

Technology	Bitcell	Configuration	Array Efficiency	Array Density
14nm	HDC	516R x 272C	71.7%	13.7 Mb/mm²
	LVC	516R x 272C	71.6%	11.6 Mb/mm² [2]
10nm (This Work)	HDC	516R x 272C	77.1%	23.6 Mb/mm²
	LVC	516R x 272C	78.4%	20.4 Mb/mm²

Figure 11.1.2: (top) 10nm technology features. (middle and bottom) Impact on array efficiency and array density.

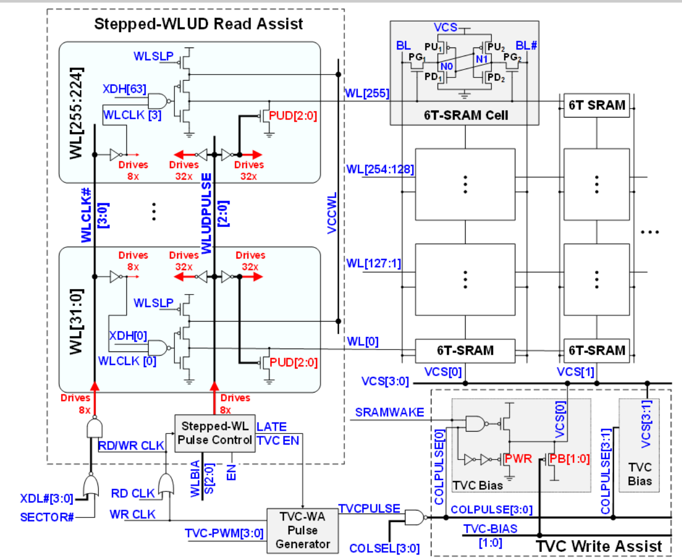


Figure Figure 11.1.3: PMOS TVC circuit, and WLUD circuit with S-WL feature.

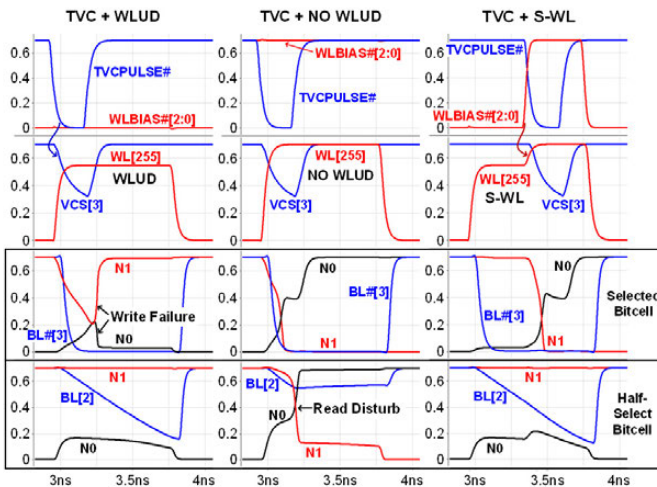


Figure 11.1.4: Simulation waveforms with (left) PMOS TVC and static WLUD, (middle) TVC and no WLUD, and (right) TVC and S-WL.

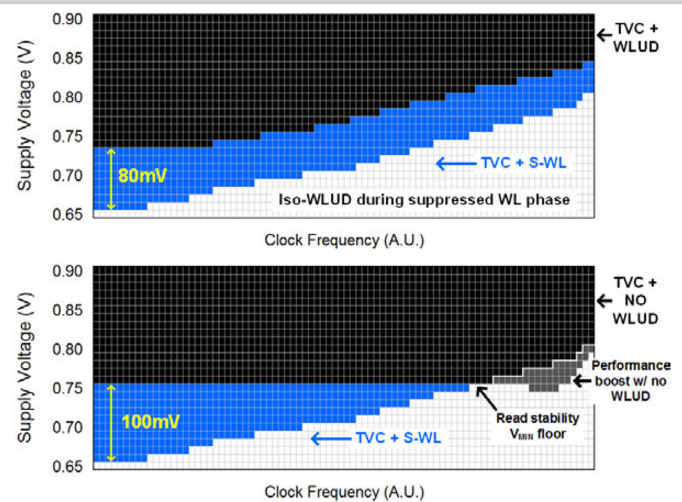


Figure 11.1.5: Measured voltage-frequency shmoo for HDC with (top) PMOS TVC and static WLUD compared to TVC and S-WL, and (bottom) TVC and no WLUD compared to TVC and S-WL.

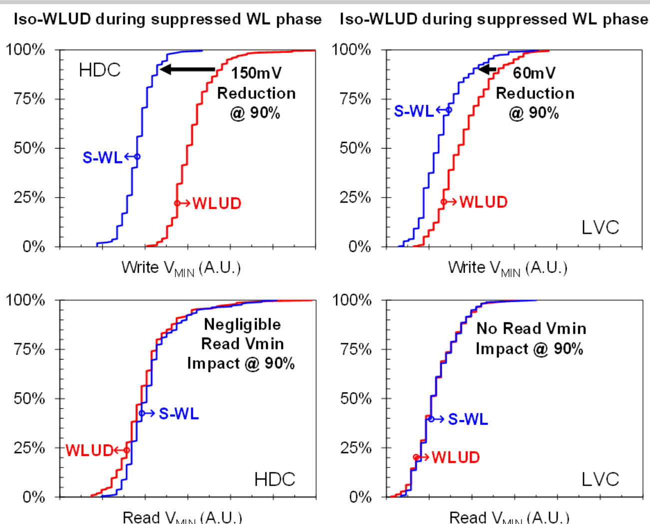


Figure 11.1.6: Measured write and read V_{MIN} distribution for HDC and LVC with PMOS TVC plus static WLUD and S-WL.

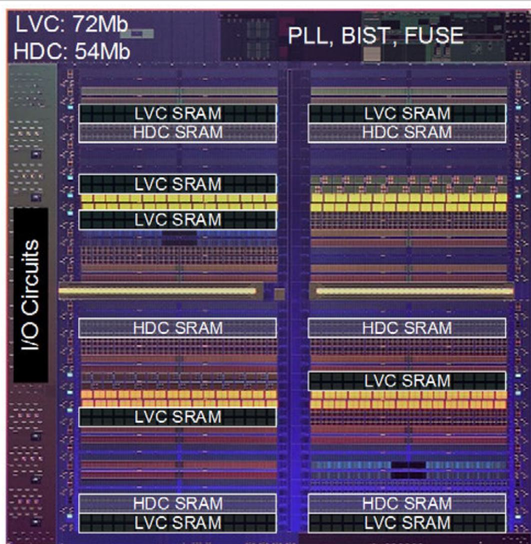


Figure 11.1.7: Micrograph of 10nm 72Mb LVC and 54Mb HDC test chip.

978-1-5090-4940-0/18/$31.00 ©2018 IEEE

11.2 A 7nm FinFET SRAM Using EUV Lithography with Dual Write-Driver-Assist Circuitry for Low-Voltage Applications

Taejoong Song, Jonghoon Jung, Woojin Rim, Hoonki Kim, Yongho Kim, Changnam Park, Jeongho Do, Sunghyun Park, Sungwee Cho, Hyuntaek Jung, Bongjae Kwon, Hyun-Su Choi, JaeSeung Choi, Jong Shik Yoon

Samsung Electronics, Hwaseong, Korea

SRAM plays an integral role in the power, performance, and area of a mobile system-on-a-chip. To achieve low power and high density, extreme ultraviolet (EUV) technology is adopted for the 7nm FinFET technology [3-4]. Conventional ArF immersion with a single exposure for an extreme high-resolution patterning shows the limitation of lithographic patterning. Therefore, multi-patterning lithographic technique is applied to support a high-resolution lithography. However, this also includes process variations due to using multi-pattering masks. Alternatively, EUV offers competitive scaling with a single-mask with the benefit of smaller wavelength, which provides smaller process variation with less additional pattering. Figure 11.2.1 shows a 7nm EUV FinFET 6T high-density (HD) SRAM bitcell with an area of 0.026μm^2. The pull-up, pass-gate, and pull-down ratios are 1:1:1 for high-density and low-power applications. Another benefit of EUV technology also features a bi-directional metal layer with a scaled pitch that provides an extra degree of freedom for signal and power routing. Figure 11.2.2 highlights EUV benefits in accordance with bi-directional metals. A uni-directional metal layer requires different metal layer to connect two nets, and have no choice but to support the limited via between two perpendicular metal lines with the limited metal width. A wider metal allows placement of more vias between the metal lines, but it does not demonstrate optimum Power, Performance, and Area (PPA) with redundant parasitic capacitance. However, EUV provides bi-directional metal lines, where the different layers of metal are coherent in the same direction. Therefore, more vias can be placed to reduce the IR-drop with smaller parasitic capacitance and resistance. Figure 11.2.2 illustrates the delay impact versus stacked-via distance in a standard cell array. It shows that the timing penalty is directly proportional to the stacked via distance in a uni-directional metal routing.

SRAM assist is a common technique for achieving low power in recent technologies [2-5]. Since the 6T-HD bitcell does not cover low-voltage ranges in write and read operation, especially in the FinFET technology, SRAM assist techniques are selectively applied to write and/or read operations. Figure 11.2.3 illustrates conventional SRAM assist schemes that control the WL, BL, and bit-cell voltage (V_{DDC}) independently or together to affect bitcell characteristics favorably. WL is controlled to help bolster the Access-Disturbance Margin (ADM) by trading off against the Write-Margin (WRM) temporarily. V_{DDC} is lowered to skew the WRM within the safe range of bitcell retention. Meanwhile, a negative BL (NBL) scheme is used as a write-assist technique to improve WRM without affecting ADM. However, the NBL technique is limited in application due to BL resistance; Fig. 11.2.3 illustrates the WRM degradation as the number of rows per BL (RPB) increases. The NBL effect diminishes for a large RPB, and is worse for the bitcell farthest from the write-driver. Otherwise, WL is used as an assist-knob to avoid the resistance impact, since WL is connected to the gate, not source of pass transistor. The WL voltage is turned-on slowly low-to-high for both ADM and WRM [5] in connection with the timing penalty for a safe ADM. Meanwhile, the BL line is designed with wider width of metal through the part of bitcell array to mitigate the BL resistance [2]. Therefore, BL resistance decreases up to 50% of the original BL by widening the infinitive width of half-BL at most, which degrades the performance with large capacitance of BL instead.

Conventionally, the write driver is located at the bottom of SRAM macro to drive the whole bitcell array. Therefore, the top bitcell, which is located farthest from the write driver, suffers from the worst WRM due to the largest BL resistance among the bitcell array under the same condition of bitcell variation itself. To minimize BL resistance effectively, the Dual Write-Driver (DWD) is proposed as a write-assist as shown in Fig. 11.2.4. The DWD uses two write drivers on the top and bottom, which act coherently in a short time. Since the two write drivers are designed by half-size of the conventional single write driver, the DWD has a similar area to the conventional one. Moreover, the farthermost bitcell from the write driver is located in the middle of the bitcell array, neither in the top nor bottom. The effective resistance of DWD is calculated using simple methods: (1) Since BL length from the write driver to the farthermost bitcell is cut by half, each BL resistance is reduced by 2x. (2) Also, the two write-drivers drive the middle bitcell in parallel at the very same time, thus reducing the BL resistance by 2x again. (3) Therefore, the effective BL resistance sums up to be 0.25x of the conventional single write-driver for the farthermost bitcell from the write-driver finally. The top write-driver features a Global Write BL (GWBL) that is designed to be enabled with the bottom write-driver in a short time. There are other approaches to decrease the BL resistance in the conventional SRAM design: (1) BL is designed using a 4x width that proportionately decreases the ADM. Therefore, there is a limitation to increase BL width using the optimum bitcell margin. Moreover, there is a PPA trade-off such as performance degradation due to a large BL width. (2) Alternatively, a multi-bank architecture is also adopted to provide a smaller BL resistance in each chunk of BL. However, a multi-bank SRAM macro requires white-space that tends to increase area at the boundary between the bitcell array and the peripheral, even more in a recent cutting-edge technology [3]. However, the DWD is effective to reduce the BL resistance by maintaining the BL capacitance with additional write-driver path. It mitigates the potential technology challenges, which make design overhead with conventional approaches. The DWD can handle 4x larger RPB effectively without trading off with the bitcell stability and scaling, which is not easily accomplished in the conventional SRAM design.

Figure 11.2.5 illustrates the 7nm EUV FinFET 256Mb SRAM array's V_{MIN} operation with NBL and/or DWD schemes. In order to exclude the impacts of ADM among V_{MIN} distribution, WLUD is applied using a 10% lower V_{DD} as a read-assist. Silicon shows that DWD itself improves V_{MIN} by 120mV, and NBL by 200mV, compared to no-assist. Then, when both DWD and NBL are applied, V_{MIN} improves by 300mV. Moreover, when V_{MIN} is measured at different positions in the bitcell array, DWD shows a smaller V_{MIN} variation. Conventionally, V_{MIN} is worse at the farthermost position from the write-driver as explained in the previous section. As shown in Fig. 11.2.5, the top-most bitcell (256th row among 256 RPB) shows the worst WRM, and a lower bitcell (64th row of 256 RPB) has better WRM in either no-assist or NBL. However, DWD shows smaller V_{MIN} variation over the bitcell array, which provides better controllability of process margin for mass production. Silicon shows that DWD reduces V_{MIN} variation by up to 8x compared to without DWD.

The SRAM macro area overhead is assessed for the different write-assist schemes in Fig. 11.2.5. NBL requires about 5% area overhead, due to the charge pump and additional buffer. However, DWD shows no more than a 0.5% area overhead due to the additional driver. Otherwise, a multi-bank architecture can be applied to decrease the resistance per BL. For example, a 4-bank architecture is adopted to implement 64 RPB with a similar V_{MIN} as shown in the silicon result. However, a 4-bank architecture requires four times white-space at the bitcell array boundary, compared to a 1-bank architecture. The SRAM macro area also increases by up to 30% for a 4-bank architecture with 64 RPB, versus a 1-bank architecture with 256 RPB.

Figure 11.2.6 shows the V_{MIN} distribution of the 7nm EUV FinFET 6T-HD SRAM with write-assist. The 64 RPB V_{MIN} distribution shows that DWD is expected to improve V_{MIN} additionally over the operating voltage-range. Figure 11.2.7 shows the die-photo of the 7nm EUV FinFET SRAM test-chips. Chip-A is designed using a 256Mb SRAM macro that explores NBL and DWD write-assist schemes. Chip-B is configured using 512Kb SRAM macros using the 0.026μm^2 6T-HD bitcell, which shows a V_{MIN} distribution with NBL assist and DWD impact.

References:

[1] S. Y. Wu, et al., "Demonstration of a sub-0.03 um2 high density 6-T SRAM with scaled bulk FinFETs for mobile SOC applications beyond 10nm node," *IEEE Symp. VLSI Tech.*, 2016.
[2] J. Chang, et al., "A 7nm 256Mb SRAM in high-k metal-gate FinFET technology with write-assist circuitry for low-VMIN applications," *ISSCC*, pp. 206-207, 2017.
[3] T. Song, et al., "A 7nm FinFET SRAM macro using EUV lithography for peripheral repair analysis," *ISSCC*, pp. 208-209, 2017.
[4] D. Ha, et al., "Highly manufacturable 7nm FinFET technology featuring EUV lithography for low power and high performance applications," *IEEE Symp. VLSI Tech.*, 2017.
[5] T. Song, et al., "A 10nm FinFET 128Mb SRAM with assist adjustment system for power, performance, and area optimization," *ISSCC,* pp. 306-307, 2016 .

 978-1-5090-4940-0/18/$31.00 ©2018 IEEE

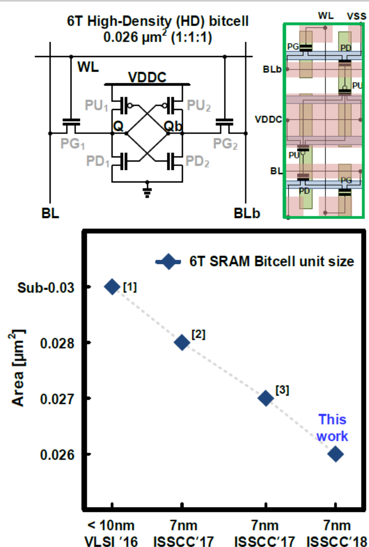


Figure 11.2.1: 7nm EUV FinFET 6T HD 0.026 μm² SRAM bitcell.

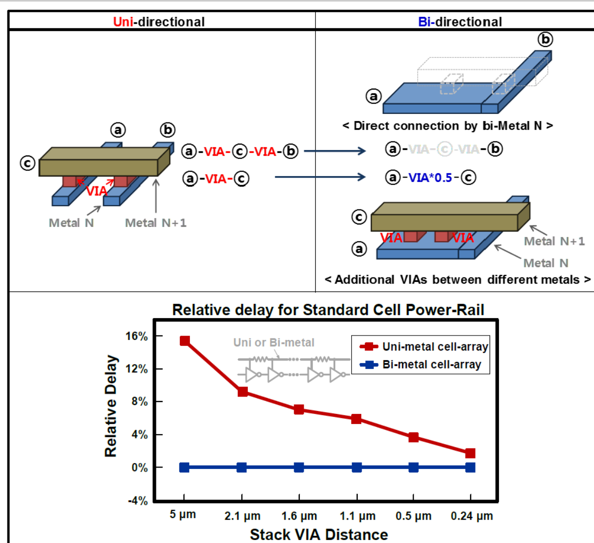


Figure 11.2.2: EUV design flexibility with smaller IR-drop impact.

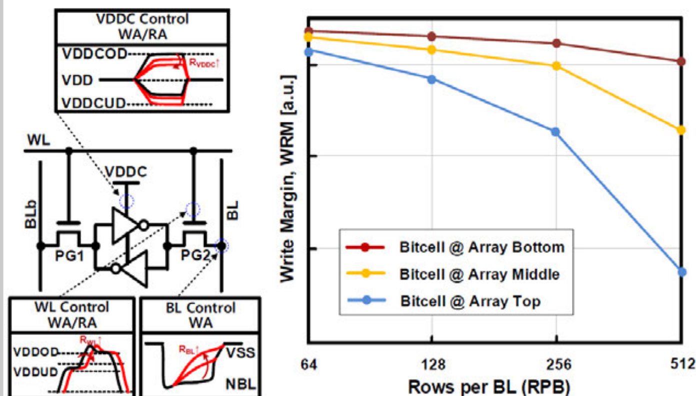
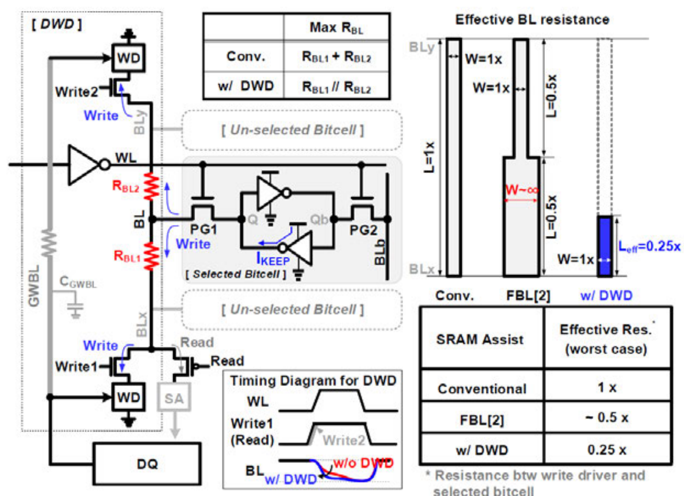


Figure 11.2.3: Conventional SRAM assist, and WRM versus rows per BL.



	Max R_{BL}
Conv.	$R_{BL1} + R_{BL2}$
w/ DWD	$R_{BL1} // R_{BL2}$

SRAM Assist	Effective Res. (worst case)
Conventional	1 x
FBL[2]	~ 0.5 x
w/ DWD	0.25 x

Figure 11.2.4: The proposed Dual Write Driver (DWD) SRAM write-assist and effective BL resistance.

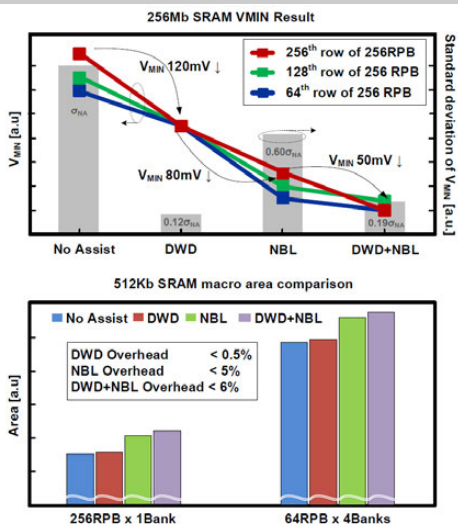


Figure 11.2.5: 256Mb SRAM silicon result with DWD or/and NBL, and SRAM macro area comparison for write-assist schemes and bank-architectures.

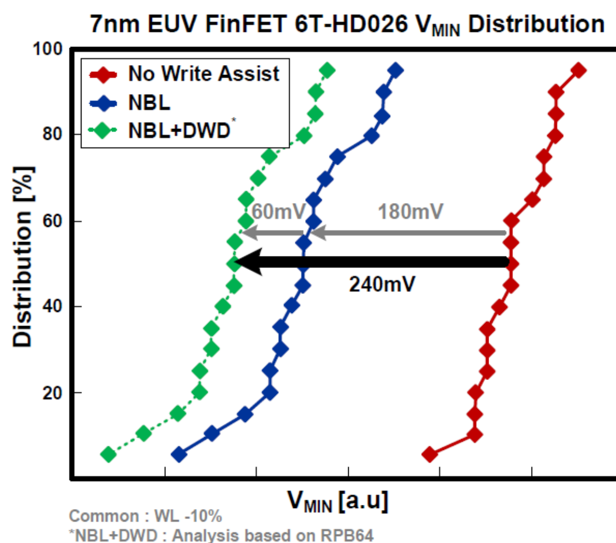


Figure 11.2.6: V_{MIN} distribution of 6T-HD SRAM bitcell.

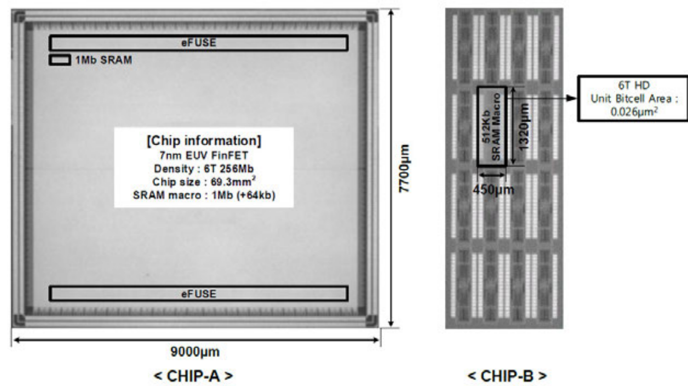


Figure 11.2.7: 7nm EUV FinFET 6T SRAM test-chips.

978-1-5090-4940-0/18/$31.00 ©2018 IEEE

11.3 A 5GHz 7nm L1 Cache Memory Compiler for High-Speed Computing and Mobile Applications

Michael Clinton[1], Rajinder Singh[1], Marty Tsai[1], Shayan Zhang[1], Bryan Sheffield[1], Jonathan Chang[2]

[1]TSMC, Austin, TX
[2]TSMC, Hsinchu, Taiwan

In high performance computing (HPC) applications, the speed of the L1 cache will typically determine the maximum frequency (f_{MAX}) of the processor core. Companies that mass produce high-performance microprocessors commonly have the L1 cache consist of fully-custom macros: to ensure that the performance of the L1 cache does not limit the f_{MAX} or throughput of the processor. In addition, it is also common for the custom L1 cache designs to use a two-port 8T or a large 6T bitcell, along with domino read logic and very short BL [2,3]. These designs tradeoff density and area for high performance. This paper presents a different approach, one which can satisfy a range of different applications; a memory compiler that can generate more than 10,000 different high-speed L1 cache macro configurations is proposed. The 7nm L1-cache compiler described in this paper uses a high-current (HC) 6T bitcell, which is more area efficient than an 8T bitcell. The HC bitcell, along with small-signal sensing, allows for long BL (256b), leading to further area efficiency improvements. Since these L1 macros are just as likely to be used in mobile applications as they are to be used in HPC applications, they were implemented using the array dual-rail (ADR) architecture [4]. The ADR architecture (Fig. 11.3.1) allows the periphery circuits of the L1 macro to operate at the same voltage as the processor core: a lower V_{DD} results in dynamic power savings. ADR performance is also improved, over an interface dual-rail, when the SRAM and logic supplies are equivalent, as ADR design does not suffer from a level-shifter delays on the inputs or outputs.

Quickly activating the WL is critical for a high-speed L1 cache. The L1 macro is built using a standard SRAM butterfly architecture and places the row-decoder and WL drivers in the center of the macro, which reduces the WL RC delay by 4×. Due to the increased wiring and via resistance in advanced nodes (i.e. 7nm) careful layout construction is required to guarantee that the upper and lower WL's are activated at exactly the same time. Within our power, performance and area constraints, we found that a four-WL clock-drive scheme resulted in the best address setup, access time and wiring/circuit area optimization (Fig. 11.3.2). Using wider than minimum WL clock (wl_clk<3:0>) wires, and reducing the gate load by a factor of four helped speed up WL activation. In addition, by controlling the WL pulse width independently for read and write cycles, we are able to shorten the WL pulse during a write and reduce the dynamic power associated with dummy reads.

In an ADR design, the WL driver must use the bitcell voltage (V_{DDM}) for proper bitcell operations. The L1 cache performs voltage level translation from the periphery's supply to V_{DDM} using the NAND gate in the WL decode path. The self-timer scheme used in the L1 cache and described in more detail in the next paragraph, depends on all of the various delays in the normal and self-time path matching. In this design we copy the entire 4-WL decode/driver block and use it to activate the single tracking WL. This allows us to replicate the layout context and layout dependent effects (LDE) for this critical portion of the access path.

The rising edge of CLK generates the internal clock (iclkz) and starts an access to the L1 cache. The self-timing scheme controls the setting time of the sense amplifier and the timing of the restore sequence. The internal clock has a very high fan-out, but we are able to generate iclkz and drive it with only one gate delay by using a dynamic clock generator circuit (Fig. 11.3.3).

The self-timing scheme consists of tracking bitcells, which are base-layer identical to normal bitcells, and therefore can track the normal bitcell read current (I_{CELL}) closely. The tracking BL has the same wire and diffusion loading as a normal BL, thus tracking the rate of voltage change very closely and proportionately to the rate of differential development on a normal BL. This scheme uses a tracking WL which is tuned to match the rise time of a normal WL across the full range of columns of the L1 compiler. The differential on the BL's at sense time is flat as a function of columns, which allows us to drive the global IO signals with fast edges. The restore operation start is timed from the sense enable trigger signal, which helps to minimize cycle time.

The HC bitcell can meet the performance targets with a 256b long BL, but there is a significant performance improvement when the BL length is cut in half. This is exactly what is done with what we refer to as the folded option. For this option, we fold the L1 macro over its right edge and reduce the BL length in half (Fig. 11.3.4). The capacity of the macro remains the same, but the BL length is halved leading to a 15-20% reduction in access and cycle time. In our current implementation of the folded macro, the area penalty is approximately 15%. The folding option can offer a sufficient performance boost, for example by pushing the minimum cycle time of the largest macro (72kb) to over 5GHz.

We recognize that the minimum differential, even with a 6σ weak bitcell, increases as the SRAM bitcell voltage is increased. We take advantage of this fact by offering a turbo mode at higher voltages, where the sense enable timing is advanced. Putting the largest L1 macros into turbo mode at high voltage, can result in an additional 5% performance boost.

Compared to a 16nm L1 cache [5] that uses the same architecture, the presented 7nm cache is over 60% smaller (Fig. 11.3.5). The high-speed 7nm L1-cache complier described in this paper has been verified in silicon. Cycle time measurements made at room temperature and -40°C are presented for a 512×36 and a 1024×72 macro. The measured results were performed on a slow-corner lot.. The -40°C measured results show that the 18kb macro is able to run at 5.36GHz at 1.115V, while the largest 72kb macro is able to achieve 4.4GHz operation at this voltage (Fig. 11.3.6).

The authors would like to thank Van Sisourath for physical design support, and Rao Kodali for logic verification.

References:

[1] J. Chang, et al., "A 7nm 256Mb SRAM in High-K Metal-Gate FinFET Technology with Write-Assist Circuitry fo Low-V_{MIN} Applications," *ISSCC*, pp. 206-207, 2017.
[2] J. Davis, et al., "7GHz L1 Cache SRAMs for the 32nm zEnterprise EC12 Processor," *ISSCC*, pp. 324-325, 2013.
[3] J. Kulkarni, et al., "Dual-Vcc 8T-bitcell SRAM Array in 22nm Tri-Gate CMOS for Energy-Efficient Operation across Wide Dynamic Voltage Range," *IEEE Symp. VLSI Tech.*, pp. 126-127, 2013.
[4] M. Clinton, et al., "A Low-Power and High-Performance 10nm Architecture for Mobile Applications," *ISSCC*, pp. 210-211, 2017.
[5] J. Chang, et al., "Embedded Memories for Mobile, IoT, Automotive and High Performance Computing," *IEEE Symp. VLSI Tech.*, pp. 26-27, 2017.

 978-1-5090-4940-0/18/$31.00 ©2018 IEEE

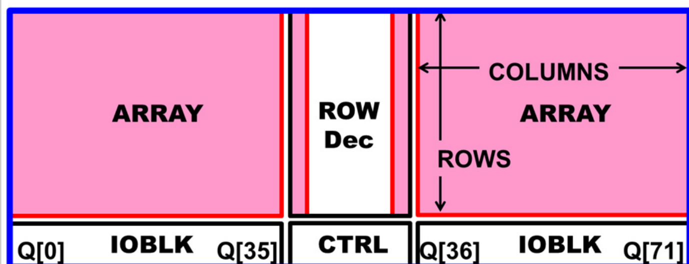


Figure 11.3.1: SRAM *butterfly floorplan* with array dual-rail power distribution.

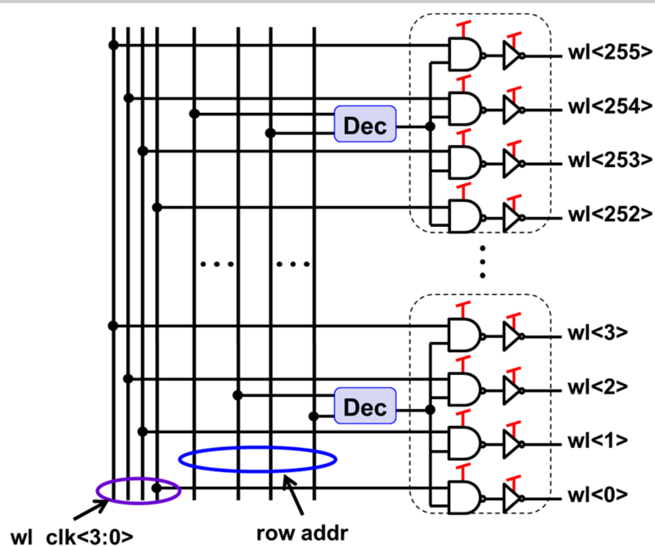


Figure 11.3.2: Row decoder, level shifter and WL driver.

11

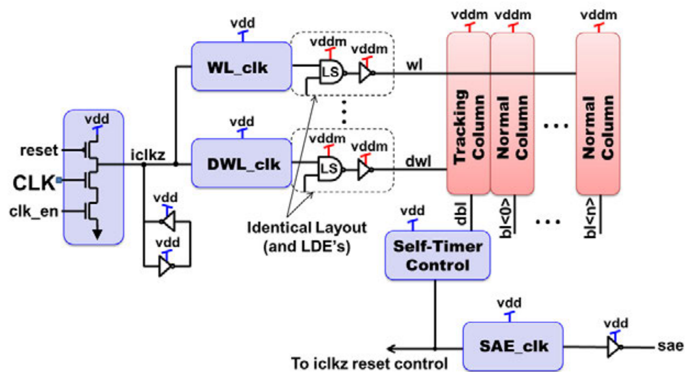


Figure 11.3.3: CLK generator, WL activation and self-timing scheme.

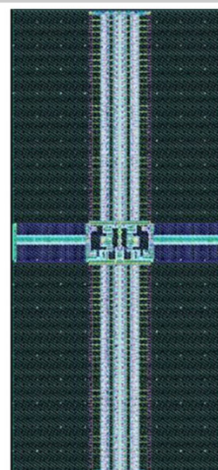


Figure 11.3.4: Layout for high-speed *folded* 7nm L1-cache macro.

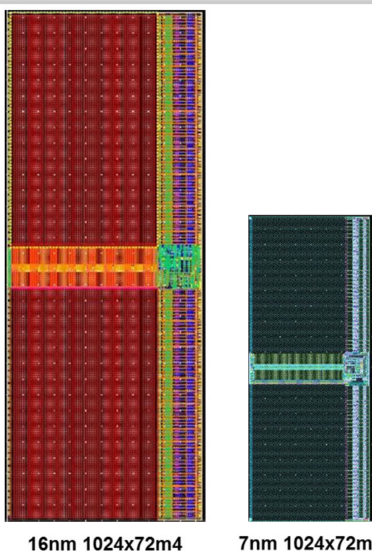


Figure 11.3.5: Layout view of 16nm and 7nm L1-cache macros.

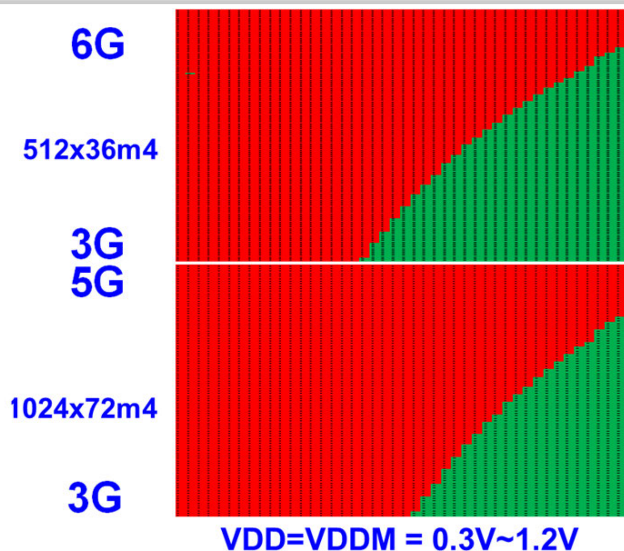


Figure 11.3.6: Silicon cycle time shmoo:for 512×36m4 and 1024×72m4.

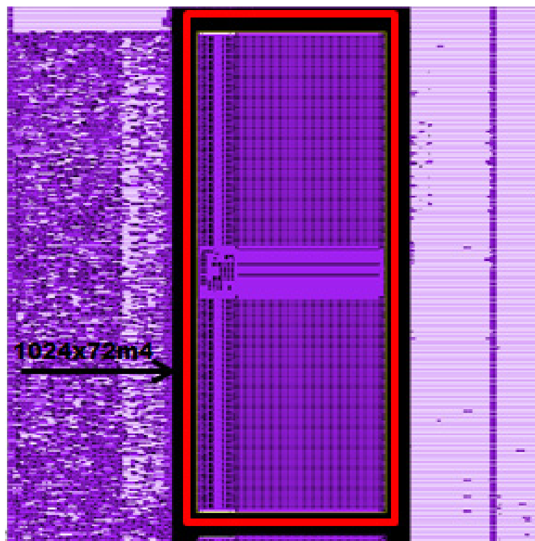


Figure 11.3.7: 1024×72 L1 cache macro die photograph.

978-1-5090-4940-0/18/$31.00 ©2018 IEEE