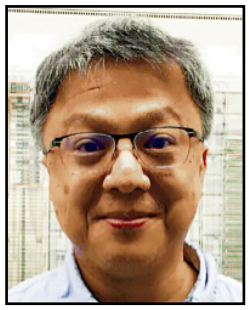
Session 12 Overview: *SRAM*

MEMORY SUBCOMMITTEE

Session Chair: *Fatih Hamzaoglu, Intel, Hillsboro, OR*



Session Co-Chair: *Chun Shiah, Etron, Hsinchu, Taiwan*

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

The growing demand for battery-powered mobile devices is the major driver to keep pushing power and area scaling for SoCs. This year, the SRAM session is headlined by the most advanced 7nm SRAM designs from both TSMC and Samsung. A novel dual-rail low-power SRAM design in 10nm from TSMC, and a two-phase-precharge ML sensing TCAM design from Globalfoundries in 14nm, are demonstrated.

978-1-5090-3758-2/17/$31.00

10:15 AM

12.1 A 7nm 256Mb SRAM in High-K Metal-Gate FinFET Technology with Write-Assist Circuitry for Low-V_{MIN} Applications

J. Chang, TSMC Design Technology, Hsinchu, Taiwan

In Paper 12.1, TSMC presents a 7nm 256Mb SRAM in high-K metal gate FinFET technology with write-assist circuitry for low V_{min} Applications. It reports the smallest SRAM bit cell published to date at 0.027μm^2.

10:45 AM

12.2 A 7nm FinFET SRAM Macro Using EUV Lithography for Peripheral Repair Analysis

T. Song, Samsung Electronics, Hwasung, Korea

In Paper 12.2, Samsung presents a 7nm FinFET SRAM macro using EUV lithography for peripheral repair analysis. A 512Kb SRAM macro is implemented with detour logic to analyze the failure phenomena of bit cell array and peripheral. The proposed peripheral repair expects to improve V_{min} by 39.9 mV based on the failure analysis.

11:15 AM

12.3 A Low-Power and High-Performance 10nm SRAM Architecture for Mobile Applications

M. Clinton, TSMC Design Technology, Austin, TX

In Paper 12.3, TSMC presents a low-power and high-performance 10nm SRAM architecture for mobile applications, with innovative dual-rail SRAM architecture. It achieves superior power savings and performance scaling when compared to the previous 16nm technology node.

11:45 AM

12.4 1.4Gsearch/s 2Mb/mm^2 TCAM Using Two-Phase-Precharge ML Sensing and Power-Grid Pre-Conditioning to Reduce Ldi/dt Power-Supply Noise by 50%

I. Arsovski, Globalfoundries, Essex Junction, VT

This paper describe two Ldi/dt management techniques implemented in a 14nm FinFET 2Kx640b TCAM running at 1.4Gsearches/sec while achieving a density of 2Mb/mm^2. To reduce within-cycle noise, a Two-Phase Match-Line (ML) Pre-charge cuts the current on easy-to-detect multi-bit mismatched MLs early to save 60% of the ML power and reduce within cycle noise by 52%. To reduce multi-cycle noise, targeted dummy search operations are inserted during low-current demand periods to flatten out current demand and reduce Ldi/dt noise by another 50%.

12.1 A 7nm 256Mb SRAM in High-K Metal-Gate FinFET Technology with Write-Assist Circuitry for Low-V_{MIN} Applications

Jonathan Chang[1], Yen-Huei Chen[1], Wei-Min Chan[1], Sahil Preet Singh[1], Hank Cheng[1], Hidehiro Fujiwara[1], Jih-Yu Lin[1], Kao-Cheng Lin[1], John Hung[1], Robin Lee[1], Hung-Jen Liao[1], Jhon-Jhy Liaw[2], Quincy Li[2], Chih-Yung Lin[2], Mu-Chi Chiang[2], Shien-Yang Wu[2]

[1]TSMC Design Technology, Hsinchu, Taiwan
[2]TSMC, Hsinchu, Taiwan

The growing demand for battery powered mobile devices is a major driver for reducing power and continued area scaling in SOC chips. Continued scaling of the transistor and metal interconnection geometry is accompanied by increasing random V_t variation and increased wire routing resistance and capacitance variation in advanced technologies. Such variation degrades SRAM performance and its minimum operating voltage, which then seriously impact the battery life of mobile devices. FinFET technology provides a superior short-channel effect and less random dopant fluctuation. However, the quantized channel width and length force constrains on transistor sizing of high density SRAM bitcells. Figure 12.1.1(a) shows the layout of a high density 6T SRAM bit cell with a 0.027μm^2 area in a leading edge 7nm FinFET technology. In order to achieve minimum area, all transistors (PU, PG, PD) in this bitcell have to be sized as single fin. Figure 12.1.1(b) shows a contention between the pull-up (PU) and the pass-gate (PG) transistors during a write operation. A stronger PU transistor results in better read stability, but the write margin is significantly degraded and results in elevation of minimum operation voltage for write operation. The negative bit-line (NBL) technique was proposed to improve write V_{MIN} in previous work [1-6]. In addition to transistor scaling, the geometric scaling of metal and via routing increases the back-end wire RC load, which also significantly degrades SRAM operation speed. In this work, we use a flying BL (FBL) and double WL (DWL) design to mitigate the RC wire load impact in order to improve SRAM array access performance.

Figure 12.1.2(a) shows a schematic of an SRAM design equipped with the NBL write assist scheme. In order to track NBL signal timing to different SRAM array configurations, a replica BL is used. In the write operation, a write enable signal triggers the replica write driver to pull the replica BL (RBL) low to generate a negative BL enable signal (ENB_NBL). The ENB_NBL signal will propagate and become the coupling signal (NBL_FIRE). Then, the falling edge of NBL_FIRE signal is going to couple to a capacitor (C1) to generate a negative coupling signal (NVSS). Next, the instant negative bias will be transferred into the selected bitcell through the write driver (WD1) and the write multiplexer (N1). Figure 12.1.2(b) shows the required negative BL bias (blue line) and simulated coupling NBL voltage levels (red line) of NBL write assist scheme. Since aggressive negative bias is needed to achieve the write V_{MIN} target, the coupled negative-bias level has to be more negative to provide the required BL write voltage. Due to the signal coupling technique, the negative-bias voltage level is proportional to the voltage level of the coupling signal. The cross point of the red line and blue line represents the write V_{MIN} with NBL write assist. Compared to the intrinsic write V_{MIN} without write assisted scheme, the write V_{MIN} with NBL assist can be improved by 150mV.

Figure 12.1.3 shows the flying BL and double WL schemes. The flying BL uses (M_{X+2}) metal layer and double WL is implemented by (M_{X+3}) metal layer. In order to support the long BL load without sacrificing the SRAM performance. A single bank array with 256cells/BL is separated into two segments by a strap row which is placed in the middle of the SRAM array. The BL of top segment is connected to the upper metal (M_{X+2}) at the middle strap row and flies over the bottom segment to connect to the multiplexer, which is placed on the bottom of the SRAM arrays. Because each memory segment contains only 128cells/BL, the effective BL load can be reduced significantly through the flying BL scheme. In order to reduce area overhead, a two-to-one multiplexer is implemented in the local read/write block to separate BL connection, as the top and bottom segments share the sense-amplifier and the rest of read-out circuits. In addition, a double WL scheme is implemented by using the upper metal layer (M_{X+3}) that is in parallel with local WL tracks (M_{X+1}) to reduce the WL metal resistance for long WL SRAM configurations.

Figure 12.1.4(a) illustrates the effective BL capacitance improvement by the flying BL scheme. Since the flying BL scheme removes the front-end load of the bottom segment, the effective BL capacitance of the top segment can be improved by 42%. Figure 12.1.4(b) shows the WL resistance improvement by the double WL scheme. With the double WL scheme, the WL resistance for the furthest bitcell can be reduced by 22%. Figure 12.1.4(c) shows the simulation waveforms of the WL pulse, and the BL discharge behavior for comparison. The blue line shows the WL/BL waveforms without flying BL and double WL schemes, and the red line represents WL/BL waveforms with the flying BL and double WL schemes. Due to the double WL scheme, the slew rate of the WL pulse can be significantly improved. Combined, the flying BL and double WL schemes improve the SRAM array read access time (defined as WL rising to BL discharging to VDD-100mV) by 40%.

Figure 12.1.5 shows the floor plan and area of a 128kb SRAM macro using a 0.027μm^2 SRAM bitcell. The SRAM macro configuration is 4096×32b with 258 bits/BL and 272 bits/WL, including row/column redundant cells. The WL decoder/driver (WLDRV) and main control block (MCTRL) are placed in the middle of the SRAM macro. The two-to-one multiplexer for flying BL scheme is placed on the boundary of SRAM bitcell array and read/write block. The NBL scheme is placed at the bottom of the read/write circuits. The area overhead of the flying BL and NBL schemes is about 3% and 2%, respectively.

Figure 12.1.6 shows the Si cumulative plot of 256Mb SRAM V_{MIN} with and without write assist at 25°C. The blue line represents the SRAM V_{MIN} when write assist schemes are disabled. Without the write assist, the SRAM V_{MIN} is wide spread across to the higher voltage range due to write failure at the lower operation voltage. The red line represents the SRAM V_{MIN} when the NBL write assist scheme is used, it can successfully improve the SRAM V_{MIN} over 150mV at the 95% percentile for the 0.027μm^2 SRAM bitcell in a 256Mb test chip.

Figure 12.1.7 shows the die photo of the 256Mb SRAM test chip, which is equipped with an electrically programmable fuse for post-silicon tuning for redundancy and write assisted options. The die area of the test chip is 42.6mm^2 with 2048 (4096×32) SRAM macros.

Acknowledgements:
The authors would like to thank R.S. Chen, Hanson Hsu, and L. J. Tyan for layout and chip implementation; the RD teams for wafer manufacturing; the test department for chip measurements on this work.

References:
[1] Y. Fujimura, et at.,"A Configurable SRAM with Constant-Negative-Level Write Buffer for Low-Voltage Operation with 0.149um^2 Cell in 32nm High-K Metal-Gate CMOS", *ISSCC*, pp 348-349, Feb. 2010.
[2] Y. Wang, et at., "Dynamic Behavior of SRAM Data Retention and a Novel Transient Voltage Collapse technique for 0.6V 32nm LP SRAM", *IEDM*, pp. 32.1.1-32.1.4, Dec. 2011.
[3] H. Pilo, et at., "A 64Mb SRAM in 32nm High-k Metal Gate SOI Technology with 0.7V Operation Enabled by Stability, Write-Ability and Read-Ability Enhancements", *ISSCC*, pp. 254-256, Feb. 2011.
[4] E. Karl, et at., "A 4.6GHz 162Mb SRAM Design in 22nm Tri-Gate CMOS Technology with Integrated Active Vmin Enhanced Assist Circuitry", *ISSCC*, pp. 230-231, Feb. 2012.
[5] J. Chang, et at., "A 20nm 112Mb SRAM in High-K Metal-Gate with Assist Circuitry for Low-Leakage and Low-Vmin Applications", *ISSCC*, pp. 316-317, Feb. 2013.
[6] Y. H. Chen, et at., "A 16nm 128Mb SRAM in High-κ Metal-Gate FinFET Technology with Write-Assist Circuitry for Low-VMIN Applications", *ISSCC*, pp. 238-239, Feb. 2014.

 978-1-5090-3758-2/17/$31.00

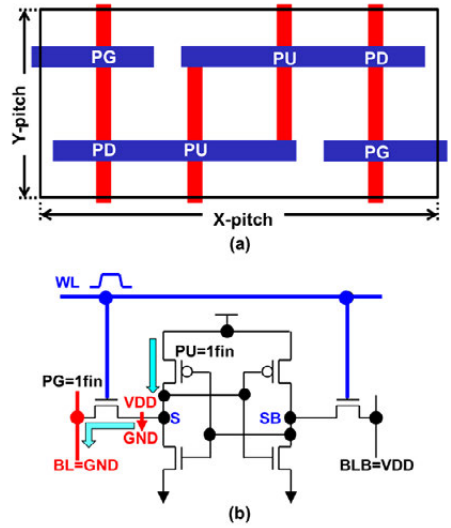


Figure 12.1.1: (a) Layout of the 0.027μm² SRAM bitcell. (b) Write contention between PU and PG of an 6T SRAM bitcell.

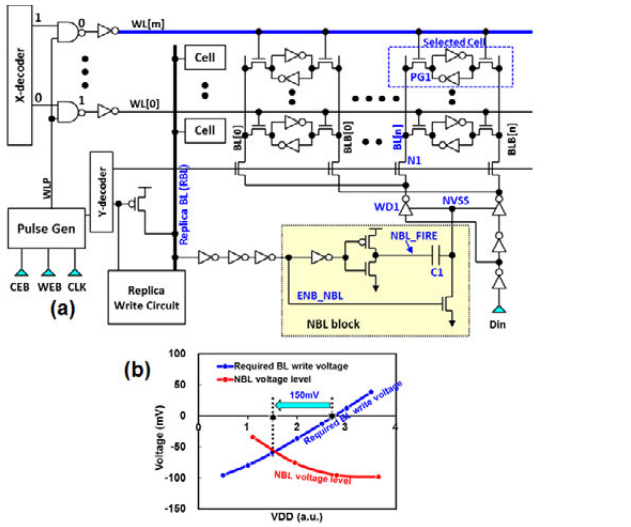


Figure 12.1.2: (a) SRAM design equipped with NBL write assist scheme. (b) Negative bitline voltage versus required write bitline voltage.

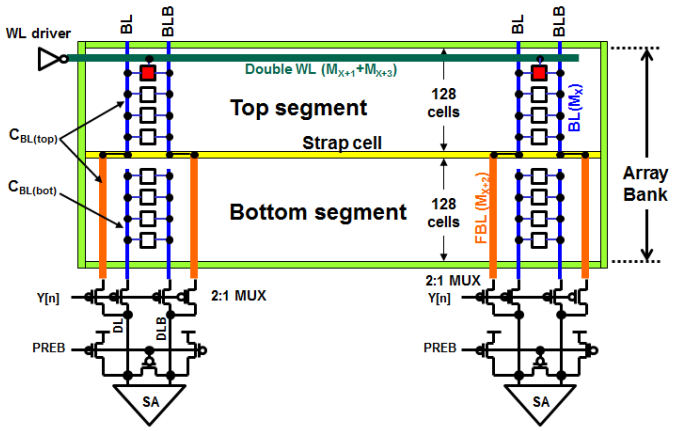


Figure 12.1.3: Flying BL and double WL schemes in an SRAM array.

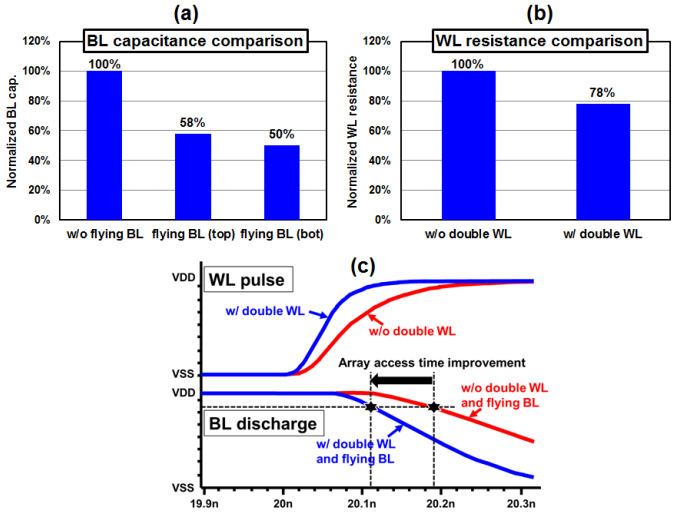


Figure 12.1.4: (a) BL capacitance comparison, (b) WL resistance comparison, (c) simulation waveforms of array read access time.

Technology	7nm HK-MG FinFET
Metal scheme	1P7M
Supply voltage	Core: 0.75V IO: 1.8V
Bit cell size	0.027μm²
SRAM macro configuration	4096x32 MUX=16 258 bits/BL, 272 bits/WL
SRAM capacity	256Mb
Test Features	Row/Column Redundancy Programmable E-fuse
Chip size	5903μm x 7223μm = 42mm²

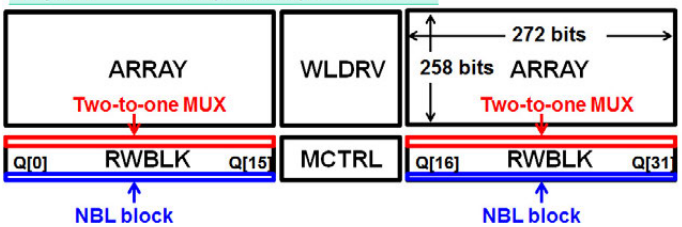


Figure 12.1.5: SRAM macro floor-plan with NBL scheme.

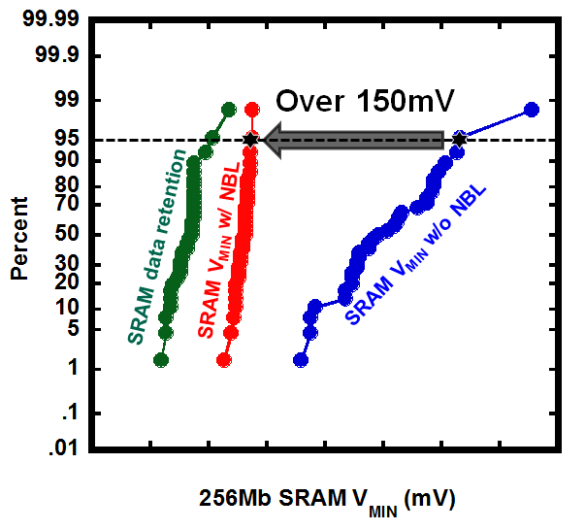


Figure 12.1.6: Si cumulative plot of 256Mb SRAM write V_{MIN} with and without out NBL write assist.

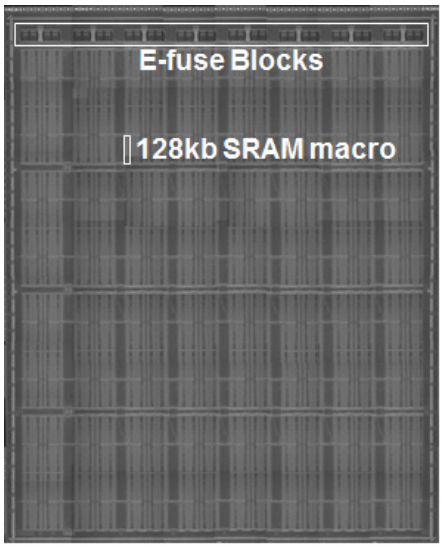


Figure 12.1.7: Die photo of the 256Mb SRAM test chip.

978-1-5090-3758-2/17/$31.00

12.2 A 7nm FinFET SRAM Macro Using EUV Lithography for Peripheral Repair Analysis

Taejoong Song, Hoonki Kim, Woojin Rim, Yongho Kim, Sunghyun Park, Changnam Park, Minsun Hong, Giyong Yang, Jeongho Do, Jinyoung Lim, Seungyoung Lee, Ingyum Kim, Sanghoon Baek, Jonghoon Jung, Daewon Ha, Hyungsoon Jang, Taejung Lee, Chul-Hong Park, Bongjae Kwon, Hyuntaek Jung, Sungwee Cho, Yongjae Choo, JaeSeung Choi

Samsung Electronics, Hwasung, Korea

Conventional patterning techniques, such as self-aligned double patterning (SADP) and litho-etch-litho-etch (LELE), have paved the way for the extreme ultraviolet (EUV) technology that aims to reduce the photomask steps [1,2]. EUV adds the extreme scaling to the high-performance of FinFET technology, thus opening up new opportunities for system-on-chip designers: delivering power, performance, and area (PPA) competitiveness. In terms of area, peripheral logic has scaled down aggressively in comparison to the bitcell given the intense design-rule shrinkage. Figure 12.2.1 shows the bitcell scaling trend and the peripheral logic unit area across different process nodes. Compared to the 10nm process node, the peripheral logic unit area is closer to the bitcell area in a 7nm process node aided by EUV, which allows bi-directional metal lines for scaling. Complex patterns and intensive scaling induce defective elements in the SRAM peripheral logic. Therefore, the probability of yield-loss due to defects is high, which necessitates the need for a repair scheme for the peripheral logic in addition to the SRAM bitcell. Despite the varied literature on bitcell repair, such as the built-in self-repair that analyzes the faulty bitcells to allocate the repair efficiently for a higher repairable rate [3], literature that discusses peripheral logic repair is sparse. Early literature [4] discusses the usage of a sense-amplifier, designed with redundancy, to address the sense-amplifier offset. Nevertheless, it is not related to the peripheral logic repair for yield improvement. This paper exclusively addresses the peripheral logic repair issue to achieve a higher repairable rate. A separate analysis of SRAM macro defect failures, in the bitcell and peripheral logic, provides a deeper understanding so as to increase the maximum repairable rate under random defect conditions.

Figure 12.2.2 shows a conventional column and peripheral repair approach; in which a bitcell defect (*B*) or a peripheral-defect (*P)* is repaired using a bitcell and peripheral tie. As such, it is possible to create a wasted resource (*W*) that could be, instead, used for additional repairs. Otherwise, an additional defect (*AU*) becomes irreparable due to the lack of redundancy in the conventional column repair; specifically, once all redundancies (*R*) are used. We propose finer grain redundancy control to provide a higher coverage of additional defects (AR). Figure 12.2.3 illustrates the repair decision diagram for the conventional and proposed methods**.** The conventional method begins by checking for a bitcell or peripheral defect, and uses available array redundancy to replace the whole of the defective column: bitcell and peripheral logic regardless of where the actual defect is located. In contrast, in this work a defect in the peripheral logic is replaced using the available peripheral redundancy, and the bitcell is addressed for additional repair independently. Handling the bitcell array and the peripheral separately helps increase the repairable rate using the peripheral repair method. Figure 12.2.3 compares the maximum repairable rate versus the available redundancy. The peripheral repair method can achieve up to a 2× higher maximum repairable rate compared to the conventional method.

An SRAM macro is designed to validate the failure phenomena for the effectiveness of the peripheral repair. Fig. 12.2.4 illustrates the SRAM test-chip with the peripheral repair analysis circuit. The efficiency of the peripheral repair scheme is assessed by analyzing the failure map for the test-chip. Using the divide and conquer rule, detour logic is implemented between the array and the peripheral logic to validate the possible failure spots. Figure 12.2.5 illustrates the SRAM macro functional blocks: sense amplifier, write-driver, and the detour logic. An SRAM write-assist scheme is implemented using WL overdrive (WLOD) and negative BL (NBL). WL underdrive (WLUD) is also implemented for read-assist.

The detour logic, Fig. 12.2.4, bypasses the internal signals to detect the defective failure spot. DETOUR-I checks the functionality of the SRAM data in and out (DQ) block. Similarly, DETOUR-II and DETOUR-III check the functionality of the peripheral logic and the bitcell. When the detour mode is enabled, the detour logic (Fig. 12.2.5c) disconnects the IN_{orig_B}-OUT_{orig_B} path and then configures IN_{orig_A}-OUT_{orig_B} path as shown in Fig. 12.2.5(c), which helps to validate the failure between the bitcell array and peripheral logic.

A 512kb SRAM macro is designed in a 7nm FinFET technology using EUV lithography. The 6T high-density (HD) SRAM bitcell is designed with a PU:PG:PD=1:1:1 fin number. As shown in Fig. 12.2.1, the 7nm FinFET 6T-HD SRAM bitcell exhibits the best scaling (smallest bitcell) published thus far.

Figure 12.2.6 shows the measured results of the macro using the detour test. Various failure maps are obtained by applying the detour test-modes. This helps to analyze the defective and the repairable failures. Silicon test results have shown that DETOUR-I skips the bitcell array and the peripheral logic without any failures. DETOUR-II skips the bitcell array by highlighting the entire column failure in a certain column array, and the DETOUR-III highlights a mix of bitcell and peripheral logic failures. According to silicon results, we have proved that there is a higher probablilty to increase SRAM yield with peripheral logic repair in 7nm technology. By extension, this probability increases as the area of SRAM peripheral logic shrinks, and thus probability of failure increases, with EUV technology.

Since 16Kb of redundancy was designed for the 512kb of bitcells, there is a 3% maximum repairable rate using the conventional repair scheme, and a 6% repairable rate using the proposed peripheral repair scheme, as shown in Fig. 12.2.6. This increase in the repairable rate, when applied to bitcell failures under a reduced supply-voltage, improves V_{MIN} by 39.9mV. The peripheral repair scheme also requires an additional multiplexer between the bitcell array and the peripheral logic to bypass the column-to-peripheral signals. However, the 1% area overhead for the peripheral repair method is negligible when compared to the additional number of repair columns needed to achieve a similar repairable rate. A 3% latency overhead is also observed, which is attributed to the switch logic between the bitcell array and the sense amplifier. Although the SRAM peripheral repair scheme has an additional latency overhead, the increasing defective yield-loss for a 7nm technology aided by EUV necessitates the need for an SRAM repair scheme. The 7nm FinFET 6T-HD SRAM is also evaluated for V_{MIN} improvement using assist. Experimental results show that the write-assist of NBL and read-assist of WLUD improves V_{MIN} by 150mV. Fig. 12.2.7 shows a die micrograph of a 7nm FinFET 512Kb SRAM test-chip with standard cell and I/O, which are designed with EUV lithography.

References:

[1] A. Veloso, et al., "Demonstration of scaled 0.099μm² FinFET 6T-SRAM cell using full-field EUV lithography for (Sub-)22nm node single-patterning technology", *IEDM*, pp. 12.4.1-12.4.4, Dec. 2009.

[2] N. Horiguchi, et al., "High yield sub-0.1μm² 6T-SRAM cells, featuring high-k/metal-gate finfet devices, double gate patterning, a novel fin etch strategy, full-field EUV lithography and optimized junction design & layout", *Symp. VLSI Tech.*, pp. 23-24, June 2010.

[3] J. F. Li, et al., "A built-in self-repair design for RAMs with 2-D redundancy," *IEEE TVLSI*, vol. 13, no. 6, pp. 742-745, June 2005.

[4] N. Verma, et al., "A 65nm 8T Sub-Vt SRAM Employing Sense-Amplifier Redundancy", *ISSCC*, pp. 328-329, Feb. 2007.

[5] E. Karl, et al., "A 4.6GHz 162Mb SRAM design in 22nm tri-gate CMOS technology with integrated active VMIN-enhancing assist circuitry", *ISSCC*, pp. 230-231, Feb. 2012.

[6] E. Karl, et al., "A 0.6V 1.5GHz 84Mb SRAM design in 14nm FinFET CMOS technology", *ISSCC*, pp. 309-310, Feb. 2015.

[7] T. Song, et al., "A 10nm FinFET 128Mb SRAM with assist adjustment system for power, performance, and area optimization", *ISSCC*, pp. 306-307, Feb. 2016

[8] 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.*, June 2016.

 978-1-5090-3758-2/17/$31.00

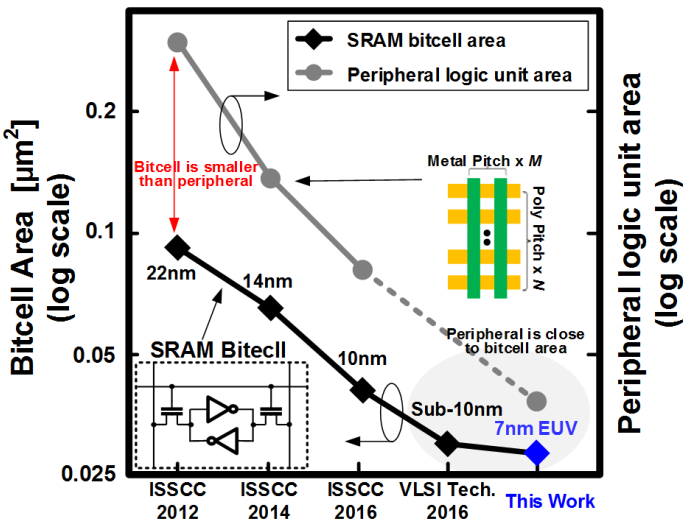


Figure 12.2.1: SRAM bitcell and peripheral logic unit area for different technology nodes.

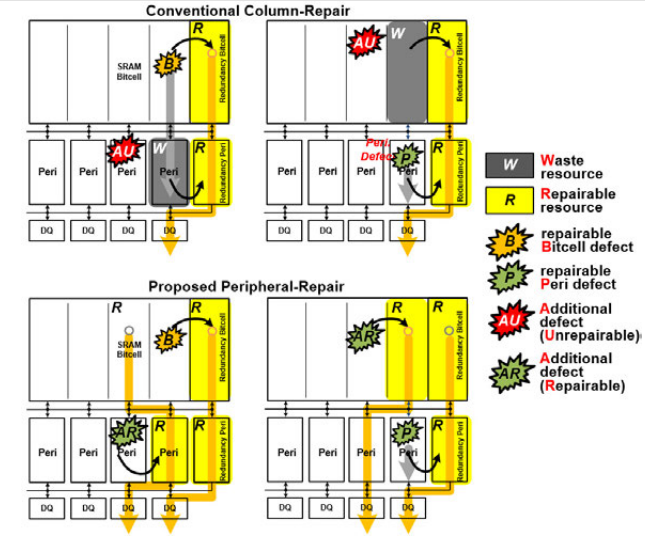


Figure 12.2.2: Conceptual behavior of the conventional column repair and the proposed peripheral repair.

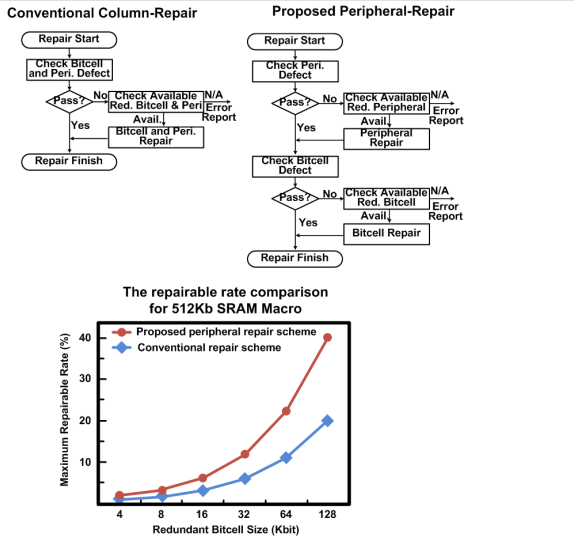


Figure 12.2.3: (a) Comparison of repair flow between the two schemes, and (b) a comparison of the maximum repairable rate.

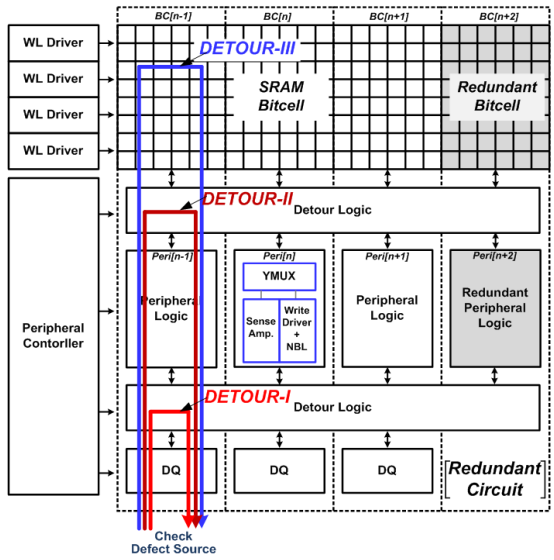


Figure 12.2.4: The SRAM test-chip for peripheral repair analysis.

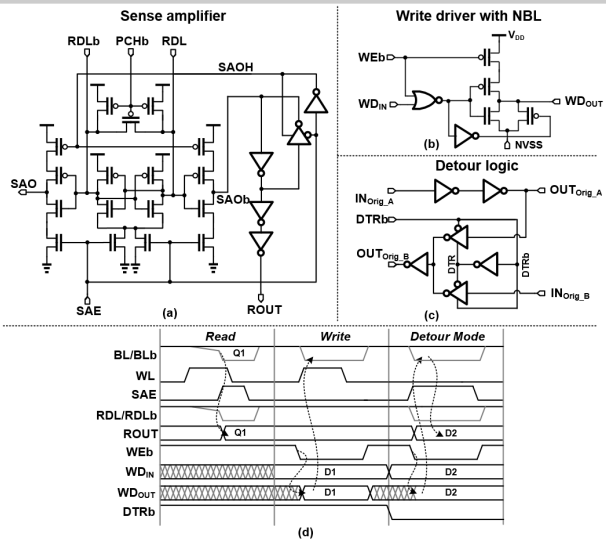


Figure 12.2.5: Circuits used: (a) sense amplifier, (b) write driver with NBL, and (c) detour logic. (d) timing diagram of the SRAM-chip.

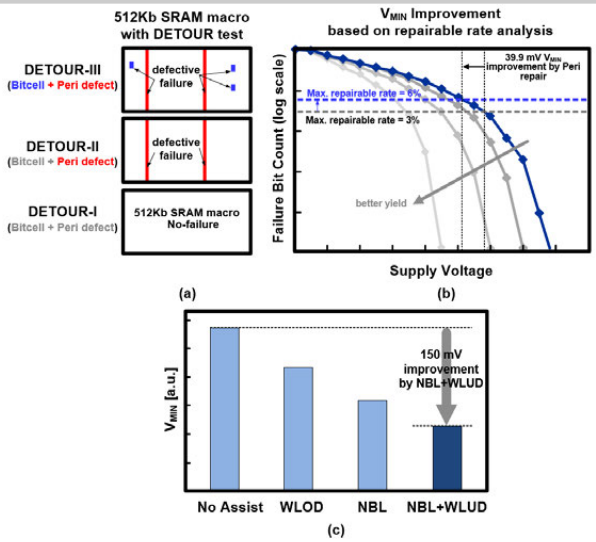


Figure 12.2.6: Silicon results: (a) DETOUR test, (b) V_{MIN} improvement with peripheral repair, and (c) V_{MIN} improvement with SRAM assists.

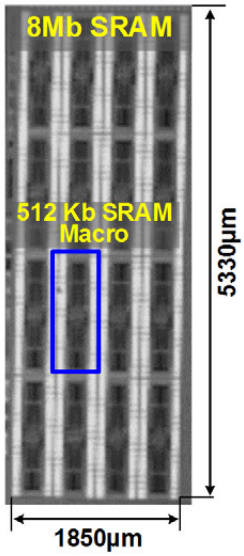


Figure 12.2.7: 7nm FinFET SRAM test-chip micrograph.

978-1-5090-3758-2/17/$31.00

12.3 A Low-Power and High-Performance 10nm SRAM Architecture for Mobile Applications

Michael Clinton[1], Hank Cheng[2], HJ Liao[2], Robin Lee[2], Ching-Wei Wu[2], Johnny Yang[2], Hau-Tai Hsieh[2], Frank Wu[2], Jung-Ping Yang[2], Atul Katoch[3], Arun Achyuthan[3], Donald Mikan[1], Bryan Sheffield[1], Jonathan Chang[2]

[1]TSMC Design Technology, Austin, TX
[2]TSMC Design Technology, Hsinchu, Taiwan
[3]TSMC Design Technology, Ottawa, Canada

Mobile applications, such as smartphones streaming HD videos or virtual-reality headsets rendering 3D landscapes, need SRAM memories that can be put in a low-power state to extend battery life, but can also offer high performance operation when required [1]. This paper will merge a 10nm technology with a dual-rail SRAM architecture to achieve superior power savings and performance scaling in comparison to the previous 16nm technology node [2]. Due to its simple design and area efficient layout, the 6T SRAM bitcell continues to be the primary memory technology used in almost all SoC and processor designs in high volume manufacturing today. The 10nm technology uses low-leakage, high-performance, second-generation FinFET transistors; it also offers a 6T cell (0.042μm^2), for area and power savings, that does not require read or write assist circuits to achieve low voltage (V_{min}) operation. This bitcell uses a fin ratio of 1:2:2 (PU:PG:PD), as illustrated in Fig. 12.3.1.

The 6T bitcell uses a common read/write port, which introduces contradictory transistor strength requirements between the read and write operation. The relative strength of the six transistors, which allow the cross-couple latch to be easily written through the NMOS pass gate, necessarily means that this same bit will be less stable during a read operation. Conversely, transistor strengths that result in a cell immune to read disturb will, by the nature of the circuit, be more difficult to write. As a result, the minimum V_{min} of an SoC design using embedded SRAM is limited by the 6T cell, due to the conflicting read and write requirements, and voltage margin loss due to random device parameter variation, such as V_t. Various read and write assist schemes have been proposed and shown as possible solutions for this V_{min} limitation [3-5].

A dual-rail architecture can be used to solve SRAM V_{min} limitations. Two variations of the dual-rail architecture have been previously described in the literature [6]. Both use independent power supplies: V_{DDM} to power the SRAM array, and V_{DD} to power the rest of the SoC logic. As such, V_{DD} is de-coupled from the V_{min} limitations of the 6T cell. The majority of an SoC's dynamic power is due to the logic that exists outside of the memory macro, hence decoupling it's V_{min} limitation from that of the memory allows one to optimize it for the lowest energy design. In this scheme V_{DDM} is only used to power the 6T cells and the WL driver. The advantage of this scheme is low energy, since everything except the bitcell and WL driver, is powered from V_{DD}. Another popular dual-rail architecture places level-shifters at the interface between the SRAM macro and the SoC, so that the 6T cells and all periphery circuits are powered from V_{DDM}. Since this dual-rail architecture powers all periphery circuits with the higher V_{DDM} power supply, it will lead to the best SRAM performance, but also the highest SRAM macro energy consumption.

A dual-rail architecture that has high energy efficiency and maintains high performance, even as the V_{DD} supply voltage is lowered, is necessary. This paper presents a hybrid dual-rail (HDR) architecture, which is based on two observations. (1) The highest component of dynamic power in the SRAM macro is due to the charging and discharging the highly capacitive BLs. (2) Activating the WL high as fast as possible is critical to achieving high performance. An HDR SRAM uses level-shifters at the memory macro interface and all periphery circuits operate off the higher V_{DDM} voltage for high performance, except the BL which are precharged to the lower V_{DD} voltage for improved energy efficiency.

The actual HDR implementation is slightly more complicated as shown in the block diagram in Fig. 12.3.3. All inputs are generated in the V_{DD}-domain of the SoC, and these power domain level-crossing signals require level shifters, which introduces additional gate delays. To avoid this gate delay, the CLK input is fed directly into a dynamic clock generator, which triggers the internal SRAM clock in the V_{DDM} domain with just one gate delay. Since all circuits from the CLK input to the WL use V_{DDM}, no level shifting is required and the WL is driven high as soon as possible. Signals driven by V_{DD}-powered gates slow down dramatically at very low V_{DD} levels. Hence, in order to retain the performance gain given by the HDR architecture, it is necessary for the tracking WL, BL and most of the circuits used to generate the sense-amplifier enable signal (in the control block) to be kept in the V_{DDM} domain.

The IO block contains both V_{DD} and V_{DDM} domain circuits, as can be seen in Fig. 12.3.3. The BLs are precharged to V_{DD} and the output (Q) must be driven with V_{DD}, so the read path is kept completely within the V_{DD}-domain. This requires all of the read control signals which are generated in the V_{DDM}-domain in the control block, to be level shifted back to the V_{DD}-domain. The global read path signals are locally buffered in each IO, which is where the level-shifting is accomplished. The cross-domain signals (global read signals) are low during standby and switched high only during a read, thereby avoiding additional standby leakage in the level shift inverter if the V_{DD} level is ever greater than the V_{DDM} level. Write circuits, except for the input level shifters and the write driver PMOS, are kept completely in the V_{DDM} domain. An important benefit of this scheme is the NMOS in the write driver is driven with a V_{DDM}-domain signal, insuring sufficient overdrive to guarantee a timely write of the BL.

When designs use multiple power domains, having robust power-up and power-down operation, which avoids excessive current, is critical to system quality and reliability. Since the SoC uses V_{DD}, this power supply is ramped up before V_{DDM} at power-up and ramps down after V_{DDM} during power-down. For an HDR SRAM, with the majority of the peripheral circuits powered from the V_{DDM} supply, a lot of internal control signals will be undefined until V_{DDM} is stabilized. To ensure a safe power-up and power down, we use a power detect circuit (Fig. 12.3.4) which will keep the internal SRAM macro header switches off until V_{DDM} is sufficiently powered-up relative to V_{DD}. In a similar manner, the circuit will turn the header switches off, once V_{DDM} has powered down sufficiently, relative to V_{DD}. The HDR SRAM is designed to operate over a wide range of V_{DD} and V_{DDM} voltages, so it is important that the power detect circuit does not erroneously trigger a power down condition and turn off the header switches during normal operation.

This work demonstrates that as the V_{DD} supply levels are reduced, for SoC power savings, the hybrid dual-rail architecture with extensive use of V_{DDM} domain circuits and signals, will achieve better performance scaling compared to a conventional dual-rail SRAM. Figure 12.3.5 shows that compared to the interface dual-rail architecture [6], the HDR architecture with BL precharge to V_{DD}, can deliver ~25% active power reduction as V_{DD} is lowered 300mV below V_{DDM}. A power detect circuit is used to guarantee robust power-up and power-down, and the silicon shmoo shown in Fig. 12.3.6 demonstrates that the HDR SRAM macro designed in a 10nm FinFET technology can operate over a very wide voltage window. A die micrograph of a 2kb×72 hybrid dual-rail SRAM macro is shown in Fig. 12.3.7.

References:

[1] H. T. Mair, et al., "A 20nm 2.5GHz Ultra-Low-Power Tri-Cluster CPU Subsystem with Adaptive Power Allocation for Optimal Mobile SoC Performance", *ISSCC*, pp. 76-77, Feb. 2016.

[2] Y. H. Chen, et al., "A 16nm 128Mb SRAM in high-κ metal-gate FinFET technology with write-assist circuitry for low-VMIN applications", *ISSCC*, pp. 238-239, Feb. 2014.

[3] T. Song, et al., "A 10nm FinFET 128Mb SRAM with Assist Adjustment System for Power, Performance and Area Optimization", *ISSCC*, pp. 306-307, Feb. 2016

[4] E. Karl, et al., "A 0.6V 1.5GHz 84Mb SRAM design in 14nm FinFET MOS technology", *ISSCC*, pp. 309-310, Feb. 2015.

[5] J. Chang, et al., "A 20nm 112Mb SRAM in High-k Metal-Gate with Assist Circuitry for Low-Leakage and Low Vmin Applications", *ISSCC*, pp. 316-317, Feb. 2013.

[6] Y. H. Chen, et al., "A 0.6V Dual-Rail Compiler SRAM Design on 45nm CMOS Technology With Adaptive SRAM Power for Lower VDD_min VLSIs", *IEEE JSSC*, vol. 44, no. 4, pp. 1209-1215, Apr. 2009.

 978-1-5090-3758-2/17/$31.00

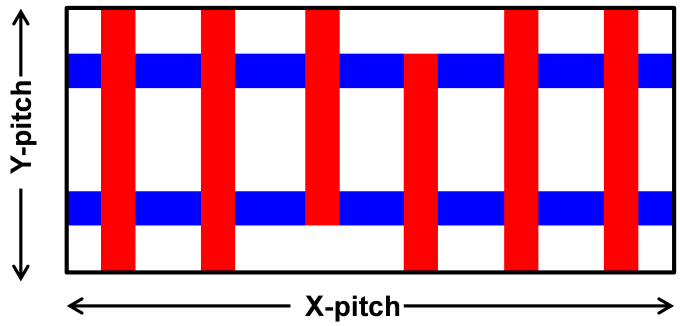


Figure 12.3.1: 10nm FinFET 0.042μm² bitcell layout.

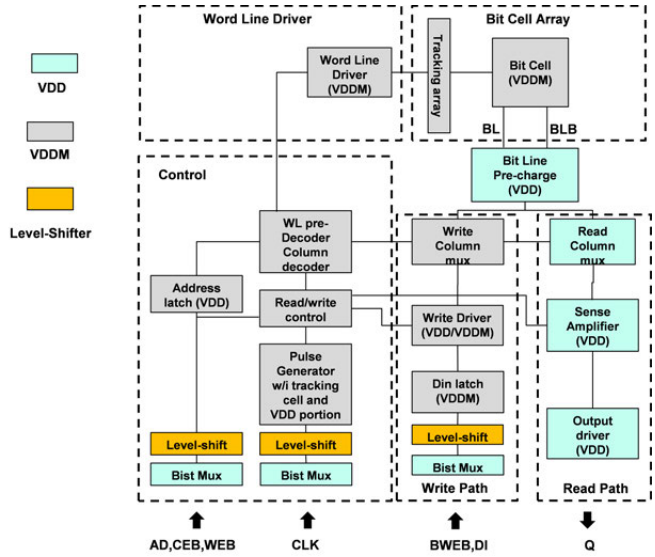


Figure 12.3.2: Hybrid dual rail (HDR) power domain block diagram.

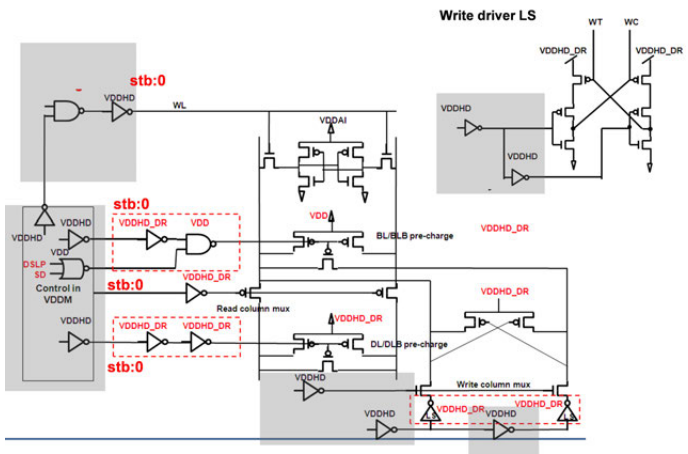


Figure 12.3.3: Hybrid dual rail circuit implementation.

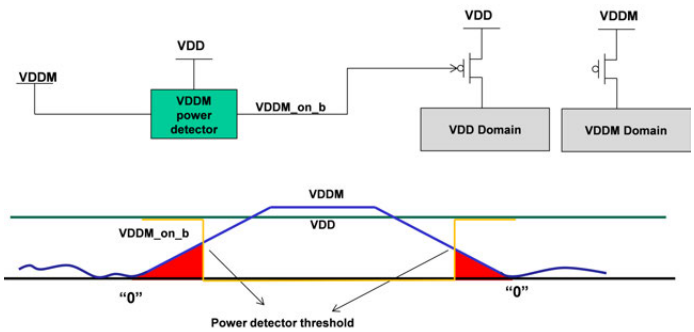


Figure 12.3.4: Power detector implementation.

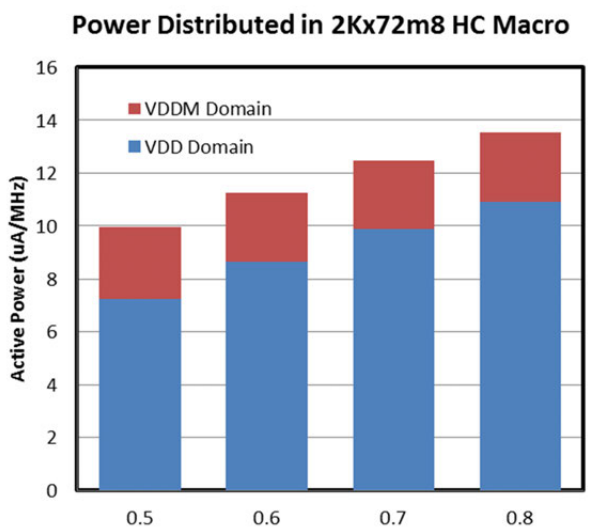


Figure 12.3.5: V_{DD} BL precharge power savings vs. precharging to V_{DDM}.

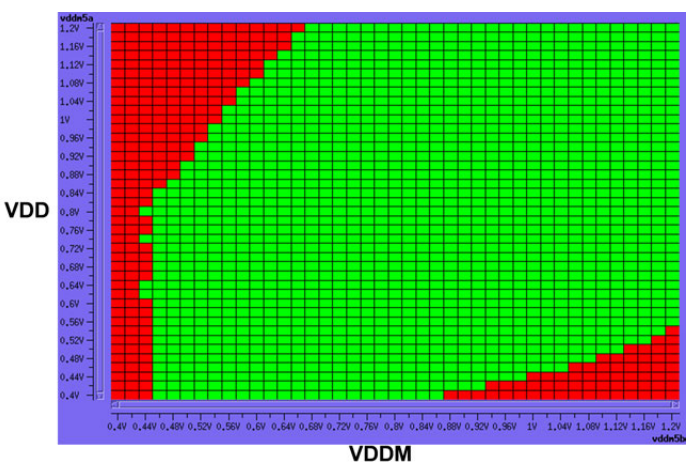


Figure 12.3.6: Silicon test data - voltage shmoo.

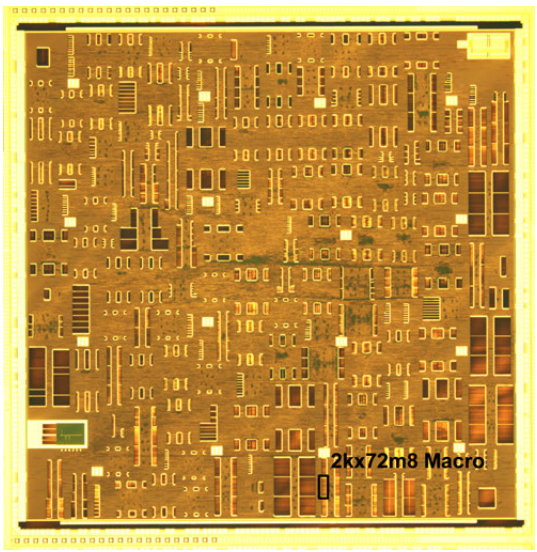


Figure 12.3.7: Die micrograph of a 2kb×72 hybrid dual-rail SRAM macro.

978-1-5090-3758-2/17/$31.00

12.4 1.4Gsearch/s 2Mb/mm² TCAM Using Two-Phase-Precharge ML Sensing and Power-Grid Pre-Conditioning to Reduce Ldi/dt Power-Supply Noise by 50%

Igor Arsovski[1], Michael Fragano[1], Robert M. Houle[1], Akhilesh Patil[1], Van Butler[1,2], Raymond Kim[3], Ramon Rodriguez[1], Tom Maffitt[4], Joseph J. Oler[1], John Goss[1], Christopher Parkinson[5,6], Michael A. Ziegerhofer[1], Steven Burns[1]

[1]Globalfoundries, Essex Junction, VT
[2]Green Mountain Semiconductor, Burlington, VT
[3]Globalfoundries, Endicott, NY
[4]IBM Research, Essex Junction, VT
[5]Globalfoundries, Raleigh, NC
[6]ASIC North, Raleigh, NC

Ternary Content Addressable Memory (TCAM) executes a fully parallel search of its entire memory contents and uses powerful wild-card pattern matching to return search results in a single clock cycle. This capability makes TCAM attractive for implementing fast hardware look-up tables in network routers, processor caches, and many pattern recognition applications. However, the push for higher performance and increased memory density coupled with parallel TCAM array activation during search operation creates large *Ldi/dt* power supply noise challenges that could result in timing fails in both TCAM and its surrounding logic.

In this paper we describe two *Ldi/dt* management techniques implemented in a 2K×640b TCAM core running at 1.4Gsearches/s while achieving a density of 2.01Mb/mm², 15% better performance and 10% better density than previous state-of-art TCAM [1]. To reduce within-cycle noise, a two-phase Match-Line (ML) pre-charge cuts the current on easy-to-detect multi-bit mismatched MLs early in the cycle and reduces ML power by 60%. To reduce multi-cycle noise, targeted dummy search operations are inserted during low-current demand periods to flatten out current demand and reduce *Ldi/dt* noise by 50%.

Figure 12.4.1 highlights the challenge with the data-dependent current-demand of the high-performance Self-Referenced ML Sense Amplifier (SRSA) [2]. When fully active the TCAM activates every Search-Line (SL) and every Match-Line (ML) in the array, creating a current demand >10× that of a similar capacity SRAM array. To reduce power conventional SRSA uses precharge-to-GND ML sensing, which eliminates the need for SL reset and reduces SL power by an average of 50%. However, the SRSA precharge phase also creates active shoot-through current through the mismatched bit-compare circuits, causing mismatched-bit-dependent ML current demand. Figure 12.4.1 illustrates the normalized current demand for the hardest to distinguish ML cases (full match (ML0) and 1b miss (ML1)) vs. that of the easiest to detect all-bit mismatched MLs (MLN, where N>>0). With most MLs having >8b mismatches, the current demand for an average SRSA ML is 2.4× that of a match, nearing the ML current of precharge-to-V_{DD} ML sensing schemes [3].

To maintain the performance and power benefits of the SRSA while also reducing the worst-case power consumption the TCAM described in this paper uses a two-phase-precharge SRSA (TPP-SRSA) shown in Fig. 12.4.2. The early pre-charge phase is used to first make a gross differentiation between likely matches (MLN, where N<8) vs. large mismatches (MLN, where N>8), while the second pre-charge phase spends additional current to further differentiate between ML0s vs. ML1s. To cut the current on easy-to-detect large-mismatches, this circuit starts the phase-one of the ML precharge with the EARLY_PRE signal. With the ML reset to GND, the pre-charge current flows though P1 and N1 and starts charging the MLs. MLs with a few bit mismatches charge up quickly thereby causing ML_{OUT} to fall shortly after EARLY_PRE is asserted. In contrast, MLs with many mismatches will charge up more slowly, which will prevent or delay the fall of ML_{OUT}. If ML_{OUT} is still high, after a replica-bias generated delay, when LATE_PRE starts the second phase of the pre-charge, the INV-AND-OR will stop the pre-charge current. Each SA makes its own decision whether to continue to supply current to the likely ML0s or cut the current and save power on (MLN, where N>>0). MLs whose ML_{OUT} signal is low continue to receive current, while MLs where the ML_{OUT} signal is high stop pre-charge, saving >60% of the ML current.

After the two-phase pre-charge phase completes the MLs are left floating allowing MLN (where N>0) to discharge and trigger a miss, while keeping ML0s precharged high resulting in a HIT. Similar to previous SRSA work [2], a LATCH signal (not shown for brevity) is used to separate between these two ML_{OUT} cases. Fig. 12.4.3 shows both the current demand as a function of number of mismatched ML bits and the associated power-supply collapse comparison between the conventional SRSA and the novel TPP-SRSA. The TPP-SRSA effectively detects eight-bit and larger mismatches (MLN where N>8) and shuts off the shoot-through current through P1 and N1 pre-charge stack. Since statistically most MLs have >>8b mismatches on a 160b wide ML this scheme saves an average of 60% of the ML current power and reduces fast transient *Ldi/dt* power-supply noise by 52%.

To further reduce *Ldi/dt* noise, this TCAM also employs multi-cycle *di/dt* reduction architecture that allows a gradual TCAM activation from low-power IDLE state, with no dynamic power consumption, to moderate-power HUM mode, where targeted dummy SEARCH operations are inserted in inactive TCAM banks during low-power periods (such as NOOPs, READs, and WRITEs) to minimize *di/dt*. The top portion of Fig. 12.4.4 shows the gradual HUM mode activation when chip-enable (CE) is activated, causing the TCAM to gradually transition from low-power IDLE mode (blocks in green) to moderate-power HUM mode (blocks in red). To minimize the impact on the neighboring TCAM logic the insertion of dummy SEARCH operations starts from the inner-most TCAM banks, and over multiple cycles, activates the outer TCAM banks where *Ldi/dt* voltage collapse can cause timing fails in the logic that is surrounding the TCAM.

The bottom portion of Fig. 12.4.4 also shows a power-supply integrity simulation, illustrating the benefit of the HUM mode once fully active. With the HUM mode disabled (red waveform) the NOOP-to-SEARCH TCAM transition starting at t=140ns shows a 20% collapse on the power-supply rail. With the HUM mode enabled at t=40ns (green waveform) the identical NOOP-to-SEARCH transition now executing on a pre-conditioned power supply only sees 10% collapse. By using this technique, the TCAM can not only improve power-supply noise, but also allow a lower operating voltage for a set performance, significantly reducing dynamic power. To reduce energy consumption associated with the HUM mode, which is actively consuming current to reduce *di/dt* events, the TCAM is also equipped with CE pin that allows a gentle multi-cycle transition from IDLE to HUM and HUM to IDLE modes.

Figure 12.4.5 shows the microphotograph of the largest compiler generated TCAM instance, 2048×640b implemented using a 16×4 matrix of 256×160b banks in a 14nm FinFET process spanning an area of 0.649mm² to achieve a memory density of 2.01Mb/mm². Figure 12.4.5 also shows a McLeod [4] loop hardware measurement of this instance showing peak cycle time of 1.4GHz at a V_{DD}/V_{CS} (logic supply to cell supply) of 0.80V/0.90V, 85°C while consuming 0.58W of power. Figure 12.4.6 shows how this work compares to previous state of art. By using TPP-SRSA this design achieves a 15% higher performance, and 10% higher density than previous state of art [1], while also significantly reducing the load-step on the power supply.

Acknowledgements:
The authors thank the worldwide GLOBALFOUNDRIES IP design community for their insightful discussions, and G. Braceras, M. Lang, R. McMahon, T. Corrigan, J. Chickanosky, K. O'Buckley, and R. Cook for encouragement and support.

References:
[1] Y. Tsukamoto, et al., "1.8 Mbit/mm² Ternary-CAM macro with 484 ps Search Access Time in 16 nm Fin-FET Bulk CMOS Technology", *IEEE Symp. VLSI Circuits*, pp. 274-275, June 2015.
[2] I. Arsovski, et al., "A 32 nm 0.58-fJ/Bit/Search 1-GHz Ternary Content Addressable Memory Compiler Using Silicon-Aware Early-Predict Late-Correct Sensing With Embedded Deep-Trench Capacitor Noise Mitigation", *IEEE JSSC*, vol. 48, no. 4, pp. 932-939, Apr. 2013.
[3] K. Nii, et al., "A 28nm 400MHz 4-parallel 1.6Gsearch/s 80Mb ternary CAM", *ISSCC*, pp. 240-241, Feb. 2014.
[4] O. Wagner, et al., "A new method for improved delay characterization of VLSI logic", *ESSCIRC*, pp. 102-105, Sept. 1982.

 978-1-5090-3758-2/17/$31.00

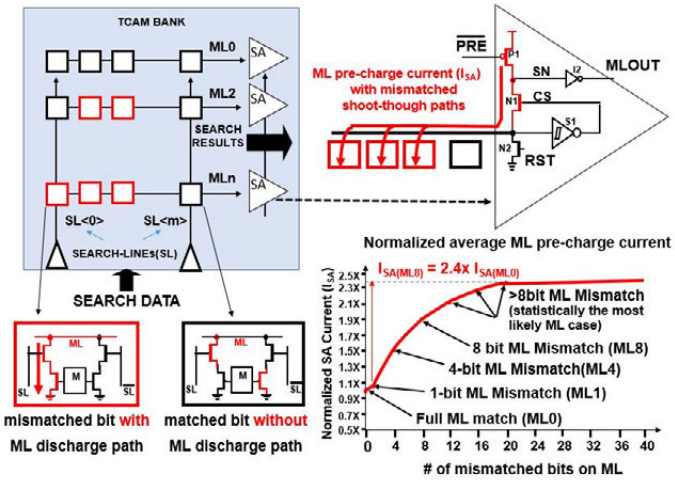


Figure 12.4.1: Self-referenced ML sense-amplifier (SRSA) uses progressively more current with the number of mismatched bits. Easiest to detect all-bit miss consumes the most power.

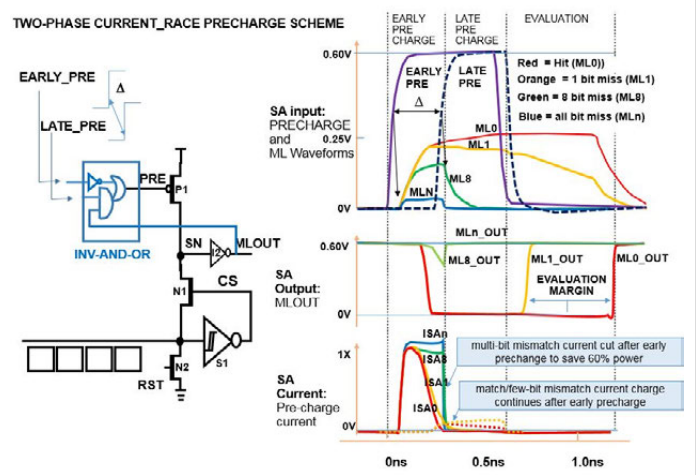


Figure 12.4.2: Two-phase ML precharge uses current during the early precharge phase to differentiate between likely matches and multi-bit mismatches, and then shuts off current to multi-bit mismatches during the late precharge phase to reduce power overall power by 60%.

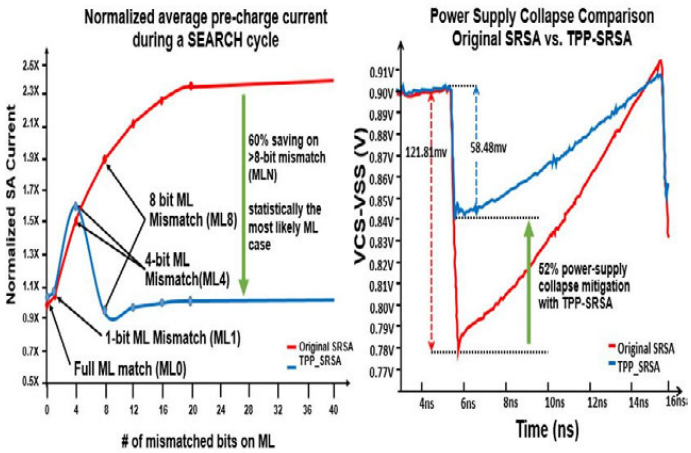


Figure 12.4.3: The two-phase-precharge SRSA cuts-off the precharge current early for MLs with multi-bit mismatches resulting in 60% current saving and 52% less voltage collapse in comparison to a conventional SRSA.

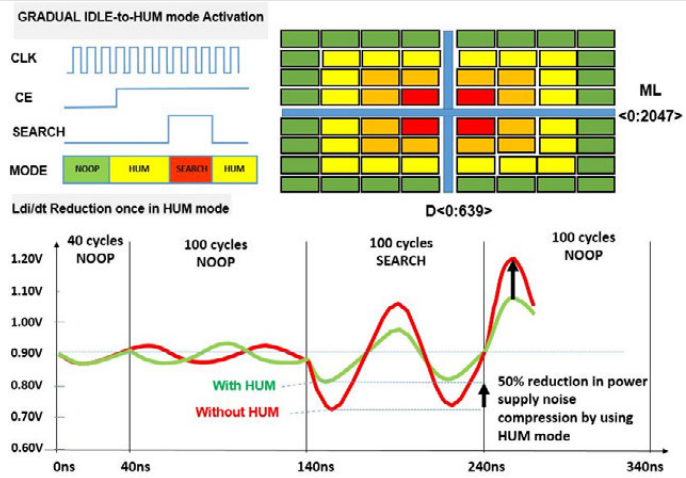


Figure 12.4.4: Gradual HUM mode bring-up minimizes startup *Ldi/dt* (red blocks enter HUM mode first followed by yellow and green). Once active, HUM mode inserts dummy search operations during NOPs to reduce TCAM *Ldi/dt* by more than 50%.

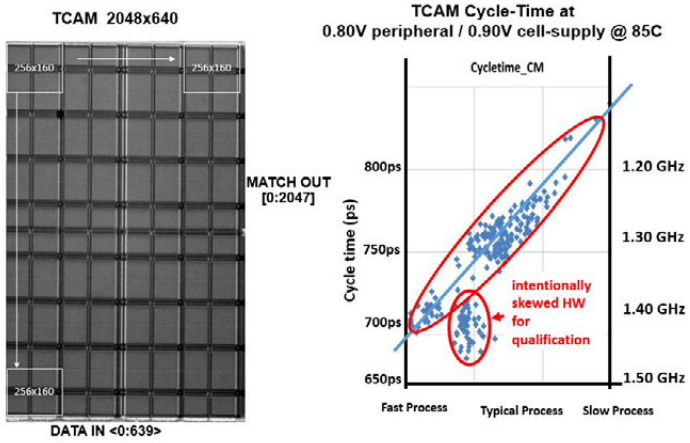


Figure 12.4.5: 2K×640b TCAM micrograph and McLeod [6] loop measurements showing a 1.4GHz maximum performance while consuming 580mW of power.

	JSSC 2013 [2]	ISSCC 2014 [3]	VLSI 2015 [1]	This Work
Tech	32nm HKMG SOI	28nm HKMG	16nm Fin-FET	14nm Fin-FET
Density	0.84 Mb/mm2	0.61 Mb/mm2	1.80 Mb/mm2	2.01 Mb/mm2
Max Search Rate	1000 Msrch/s	400 Msrch/s	1250 Msrch/s	1400 Msrch/s
VDD VCS	0.95V -	0.85V -	0.80V -	0.80V 0.90V

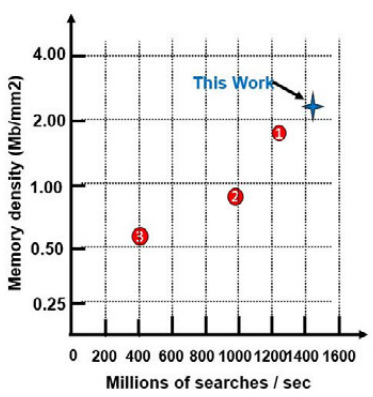


Figure 12.4.6: Comparison of this work to previously published work shows a 10% better density and a 15% higher performance to the previous state-of-art.