A 64 Mb SRAM in 32 nm High-k Metal-Gate SOI Technology With 0.7 V Operation Enabled by Stability, Write-Ability and Read-Ability Enhancements

Harold Pilo, *Member, IEEE*, Igor Arsovsi, Kevin Batson, Geordie Braceras, John Gabric, Robert Houle, Steve Lamphier, Carl Radens, and Adnan Seferagic

***Abstract*—A 64 Mb SRAM macro has been fabricated in a 32 nm high-k metal-gate SOI technology. The SRAM features a 0.154 μm^2 bit-cell, the smallest to date for a 32 nm SOI product. A 0.7 V VDD_{MIN} operation is enabled by three assist features. Stability is improved by a bit-line regulation scheme which reduces charge injection into the bit-cell. Enhancements to the write path include an increase of 40% of bit-line boost voltage. Finally, a bit-cell-tracking delay circuit improves both performance and yield across the process space.**

***Index Terms*—CMOS memory integrated circuits, high-k metal-gate, read assist, stability assist, static random-access memory (SRAM), 32 nm, Vddmin, write assist.**

I. Introduction

As on-chip memory demand increases, 6T SRAM area and minimum operating voltage scaling becomes major challenge in achieving high-density, low-power designs. In today's System on Chips memory plays a major role in overall system competitiveness. Fig. 1 shows a floor-plan of a typical 32 nm ASIC design where memory covers 56% of the total active area, it consumes more that 40% of the total power, and since SRAM limits the minimum operating voltage it also has a secondary impact on overall power by elevating the required power-supply voltage. Embedded-DRAM has been used to improve both area and power, however for fast random access applications the 6T SRAM based memory still remains the memory of choice. As evident in Fig. 1 improvements in both SRAM area and minimum operating voltage would have a significant effect on design competitiveness. This paper will discus the challenges associated with nano-scale SRAMs, and present three circuits that are used to maintain Moore's law scaling into 32 nm technology node.

Manuscript received April 25, 2011; revised June 18, 2011; accepted June 23, 2011. Date of publication October 03, 2011; date of current version December 23, 2011. This paper was approved by Guest Editor Ken Takeuchi.

H. Pilo, I. Arsovsi, K. Batson, G. Braceras, J. Gabric, R. Houle, S. Lamphier, and A. Seferagic are with the IBM Systems and Technology Group, Essex Junction, VT 05489 USA (e-mail: hpilo@us.ibm.com).

C. Radens is with the IBM Systems and Technology Group, Hopewell Junction, NY.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSSC.2011.2164730

The paper is organized as follows. Section II provides an overview of 6T SRAM scaling challenges. Sections III–V, cover the design of the three SRAM assist circuits: Stability Assist, Write Assist, and Read Assist and present the SRAM fail reduction for each. Section VI then describes the overall test chip implementation and presents the large improvement in SRAM minimum operating voltage when all of these circuits work in unison. Section VII concludes with a summary of this work.

II. Scaling Challenges

As 6T SRAM enters nanometer geometries, fewer than 100 dopant atoms define SRAM device VTs, producing large random device variation that increases with each technology node. This random device variation coupled with conflicting metrics for SRAM stability and write-ability have effectively halted conventional 6T SRAM bit-cell scaling. Fig. 2 highlights the 6T SRAM challenges through actual silicon data on a conventional SRAM design. The large ellipse shows the full manufacturing process; the slow process corner is shown in the bottom left, the fast process corner shown in the top right, and NFET/PFET skew corners are shown along the minor axis of the ellipse. The white square markers throughout the ellipse represent chips that are either perfect or have few fails that are fully fixable with redundancy. The black square markers found in the slow NFET fast PFET corner represent chips that have become unfixable because of large number of write fails, while the trapezoidal markers found in the fast NFET slow PFET corner represent chips that have become unfixable because of large number of stability fails, showing the conflicting nature of the SRAM write-ability and stability metrics. In addition to write-ability and stability, SRAMs also suffer from yield loss due to readability, shown with triangular markers, and leakage screens (pentagons) which reduce the yield at the fast NFET fast PFET. The 6T SRAM bit-cell shown in Fig. 3 contains some of the most studied six transistors in the semiconductor industry. Each of these transistors impacts various metrics of the SRAM design—whether it is functionality, area, power or performance. However, the SRAM pass-gate has by far the largest impact on SRAM bit-cell stability and write-ability. The graph on the right shows the stability and writability design sigma as a function of pass gate strength. Reducing Pass-Gate strength

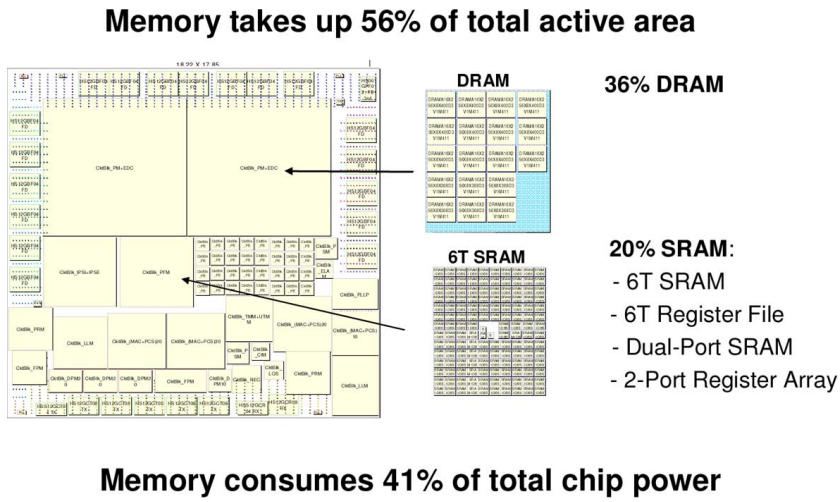


Fig. 1. 6T SRAM motivation.

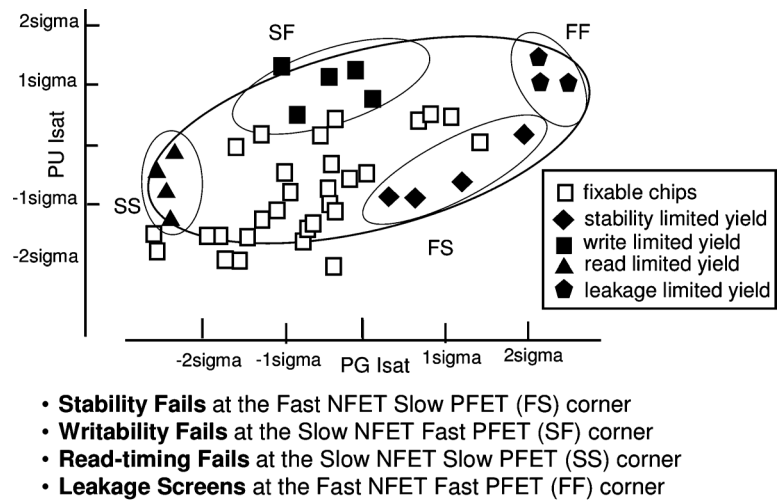


Fig. 2. 6T SRAM scaling challenge.

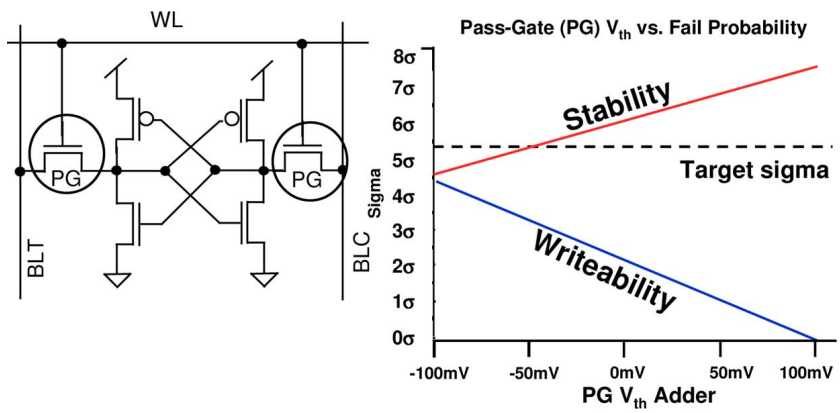


Fig. 3. Stability versus writeability trade-off.

improves Stability at the cost of degrading Write-ability. The opposite is true for the increasing the Pass-Gate strength. The trade-off is roughly 1:1.

Ideally the Pass-Gate strength is modulated based on the operation being performed. The function of the SRAM assist circuits is to modulate the strength of the Pass-Gate. Stability Assist reduces the Pass-Gate strength during a read or a half select operation to improve SRAM bit-cell stability. On the other hand, write assist increases the Pass-Gate strength on the selected bit-cells during a write to improve SRAM bit-cell write-ability. The

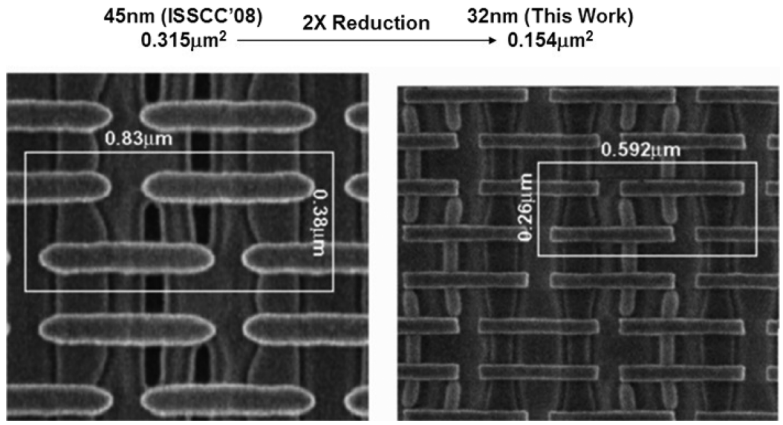


Fig. 4. 45 nm to 32 nm technology scaling of 6T SRAM bit-cell.

Read assist circuits improve performance by the close tracking of SRAM bit-cell device characteristics to determine the optimal delay between Word-line activation and Sense-Amplifier set time.

These three SRAM assist circuits were implemented on a 64 Mb SRAM test-chip designed in 32 nm High-K Metal Gate SOI process. Fig. 4 shows the 0.154 μm^2 bit-cell. A $2\times$ reduction of macro size from the previous 45 nm design [2] is enabled by an equal $2\times$ reduction in bit-cell area. No corner rounding of bit-cell gates allows tighter overlay of gate electrode and active area. The introduction of High-k Metal-Gate provides a significant reduction in the equivalent oxide thickness, thereby reducing the Vt mismatch. This reduction allows aggressive scaling of device dimensions needed to achieve the small area footprint. Power scaling continues to drive the demand for lower $\mathrm{VDD_{MIN}}$; this is enabled by innovation in the periphery assist features and is described in the Section III.

III. Stability Assist

Stability assist by reducing the pass-gate strength is an effective method to decrease SRAM bit-cell stability failure rate. However, several of these methods [4], [5] interfere with other operations and require external modulation to preserve the balance between stability and write-ability as shown in the chart of Fig. 3. An alternative method to achieve stability improvement without degrading the write margin, is accomplished by pre-charging bit-lines to a reduced level, VBLH [3]. As shown in Fig. 5, the reduced bit-line voltage results in a lower bump level, VBUMP, on the bit-cell's internal "low" node. This improves the noise margin of the cross-coupled inverter's trip-point. The reduction of VBUMP is accomplished by the reduction of the pass-gate drain-to-source voltage, Vds, effectively improving the SRAM bit-cell beta ratio. The waveforms in Fig. 5 show bit-line levels during a read operation with bit-lines pre-charged to both VDD and VBLH voltages. The VBLH voltage supply is regulated from the VDD supply; the regulation system is shown in Fig. 6. The system consists of a Voltage Reference circuit (transistors T0, T1 and T2), and a Push-Pull Regulator with a distributed PFET output device, Treg. The Voltage Reference circuit is PVT compensated by body-contacted device T0 to maintain a relatively constant reference voltage, Ref. Diode

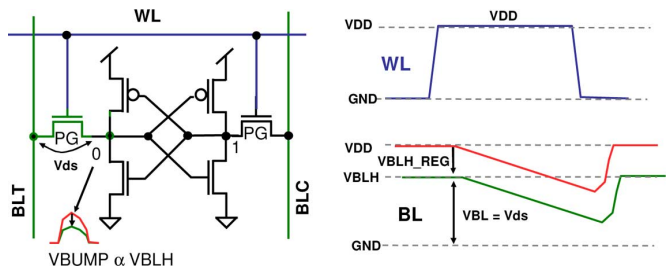


Fig. 5. Stability assist concept.

connected PFET, T1, modulates the body of T0 to compensate for changes in Vt as a result of process, voltage or temperature variation. For example, an increase in temperature lowers the Vt of transistor T0 which in turns raises the voltage of Bbias to counter-act the effect of the temperature increase by raising the Vt of T0. Likewise a decrease in temperature is compensated by a decrease in the body bias of T0 causing its Vds to be kept near constant. Fig. 7 shows the regulator reference voltage (Ref) variation from the VDD terminal $(\mathrm{Vdiff} = \mathrm{VDD\text{-}Ref})$ measured as a function of VDD across process and temperature. Each line represents a given process-temperature ranging from -40 C to 125 C and 3-sigma slow to 3-sigma fast process window. The top graph in Fig. 7 shows the original reference circuit design with no threshold compensation, while the bottom graph shows the proposed reference circuit as shown in Fig. 6. In the proposed design the maximum variation is decreased by more than $3\times$ to 50 mV.

The Push-Pull Regulator's differential amplifiers compare the reference voltage input, Ref, with the regulator output voltage, VBLH. The diff-amps respond to changes in the VBLH by driving outputs Gbias and Gnbias to alternately turn on either PFET Treg or NFET TL until VBLH equals Ref. The PFET regulator device, Treg, is physically distributed in parallel devices and located in each sense amp, near the bit-line and sense amp pre-charge devices that it sources. The physically distributed Treg facilitates the array grow-ability as it is compiled along the word-line dimension by automatically adding or subtracting regulator devices as the number of sense-amplifiers change. Bit-cell leakage causes the VBLH supply to drift towards VDD which decreases the stability

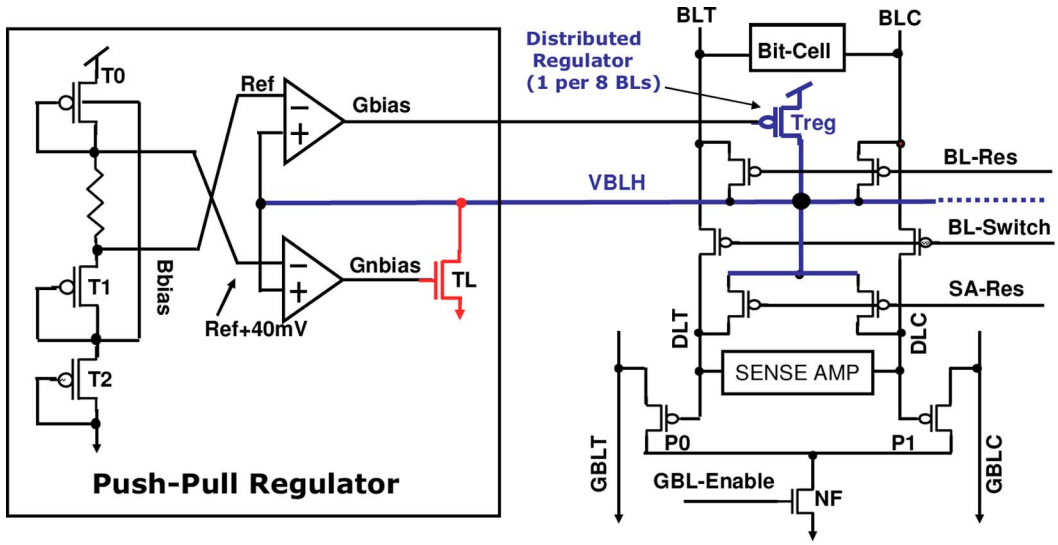


Fig. 6. Stability assist implementation.

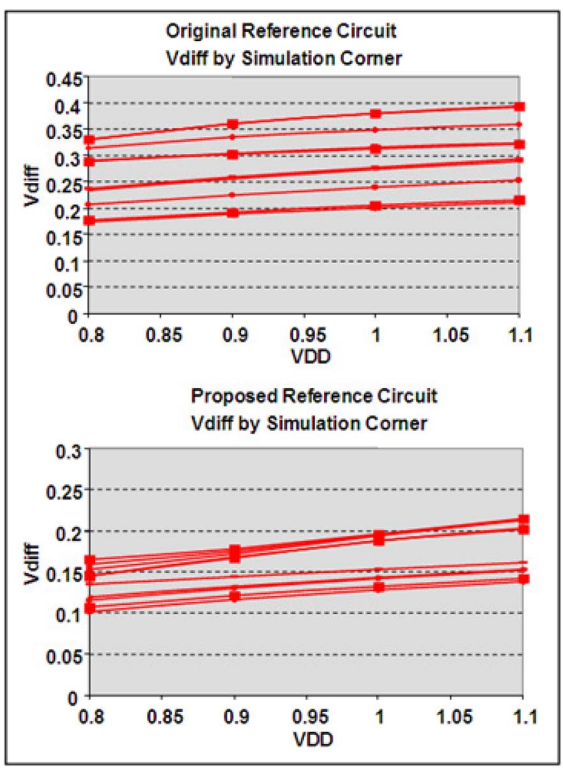


Fig. 7. Push-Pull regulator reference design comparison.

advantage. To prevent VBLH from drifting, a single pull-down device, TL is driven by the push-pull regulator signal Gnbias. Overlap current between Treg and TL is minimized with a 40 mV dead zone built into the regulator. The voltage translation from the VBLH to the VDD domain occurs in the transition from the SA data lines DLT/DLC to the global read data lines GBLT/GBLC. The reduction in bit-line pre-charge voltage also reduces the Drain-to-Source voltage across the bit-cell pass-gate during the read operation. This has an effect of reducing the bit-cell read current by 7% and a similar slow down effect to the sense-amplifier sensing time as the common mode of the sense-amplifier is reduced. Conversely, the global read data lines selection is slightly faster from the reduced input swing. The overall macro access time delay overhead from this technique is 2%. Overall power consumption is reduced with stability assist. The push-pull regulator and voltage reference consume less than 1 mA DC current at the fast process corner. The AC power consumption of the regulator is very small as the load of the regulator device has a slow-frequency, limited signal swing. The additional power consumption of the regulator system is small compared to the savings from 1) reduced pass-gate leakage from the lower pre-charge voltage and 2) decreased dynamic power from the reduced swing of the bit-line during the pre-charge. A 1.5% area overhead for the stability assist includes the regulator and reference circuits, which are located in the control section of the macro (one for each 512 Kb macro). To improve the stability and frequency response of the regulator system, decoupling capacitance is added to the VBLH network. The response time of the regulator has to be kept constant as the memory configuration changes in both number of banks and I/O width directions. The distributed topology of the regulator facilitates the different memory configurations in the I/O dimension as both the number of bit-lines to be regulated (load) and the number of regulator devices are added or subtracted in equal increments. As the number of banks is decreased, the load to the regulator is kept the same as the number of active bit-lines is unchanged. Therefore, the VBLH decoupling capacitance has to be maintained for all bank configurations. This results in a larger area overhead for smaller bank configurations.

Fig. 8 shows the improvement in failure rate as a function of VBLH level, plotted as a fraction of VDD supply at the process corner with the worst SRAM stability. The regulator is designed to operate with a VBLH range of 68–78% of VDD. The VBLH range takes into account transients effects of the regulator when operating from a “cold” start where the memory is idle for long time periods before bit-line pre-charge occurs. A 0.6-sigma improvement in stability failure rate is simulated when VBLH is operating in the transient operating range as circled in Fig. 8.

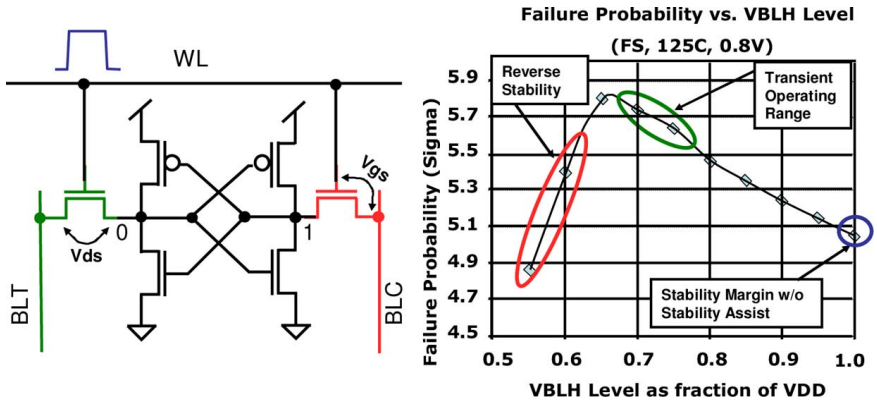


Fig. 8. Bit-Line voltage versus stability sigma.

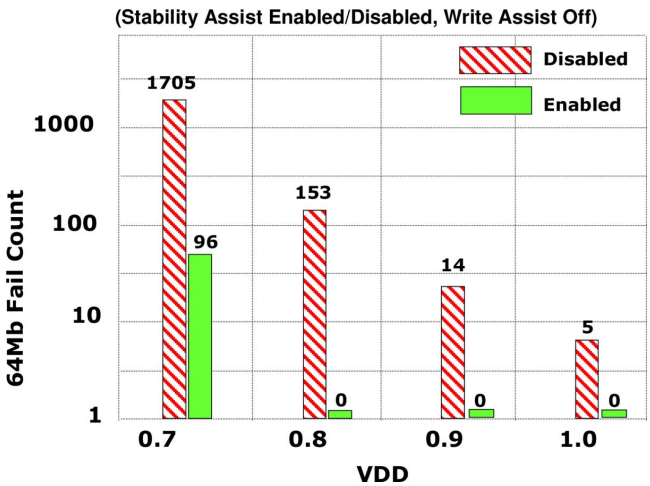


Fig. 9. Stability assist hardware results at FS corner.

The rapid increase in failure rate that occurs as VBLH is lowered beyond this operating range is caused by reverse stability fails and also shown in the schematic of Fig. 8. A reverse stability fail occurs when the pass-gate connected to the "1" side of the bit-cell begins to conduct and discharges the "1" node from the full bit-cell potential (Vcs). Fig. 9 shows hardware results of the stability assist feature at the Fast-NFET, Slow-PFET (FS) process corner which is the most susceptible for stability failures. The cross-hatched bars indicate single-cell failure rates for 64 Mb SRAM die operating without stability assist. A large improvement in failure rate is observed when stability assist is enabled (solid bars). Failure rate can be further reduced when write-assist is enabled, as described in the Section IV.

IV. Write Assist

Negative bit-line boosting is an effective technique to improve write margin [4]. Fig. 10 shows the bit-cell and corresponding waveforms for negative bit-line boosting technique. The bit-line boost increases the V_{GS} of the Pass-Gate, facilitating the discharge of the internal bit-cell node. Limitation in the boost voltage in previous work [4] is caused by partial capacitor discharge which reduces the charge transfer into the bit-lines. Another limitation is the loss of boost signal from leakage

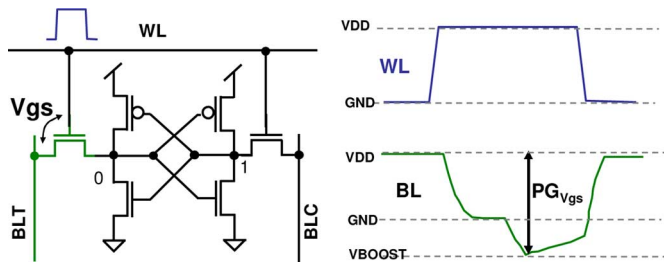


Fig. 10. Write assist concept.

from the unselected transistors in the write path that begin to conduct when source voltages are boosted below GND. Conventional negative bit-line boosting schemes are also susceptible to over-voltages of circuitry in the write path at the higher VDD voltage range as the boost voltage can reach its maximum value. This may cause gate-oxide reliability wear-down on critical write driver transistors, accelerating the end-of-life. Fig. 11 shows a schematic of the write driver with boost control that features a 40% improvement in boost voltage compared to the previous design [2]. Boosted node Nboost connects to eight physical bit-line pairs, segmented into upper (Ntu/Ncu) and lower half (Ntl/Ncl) partitions. Only the upper-half partition is shown in detail for clarity. Nboost is pre-charged to GND by Nd at the end of the write cycle. Boost capacitor, Cboost is also charged during this time by the transition of WS1n to VDD also during pre-charge. Cboost is built using a thin-oxide depletion-MOS capacitor to maximize capacitance per unit area. Also shown in Fig. 11 is the boost control circuitry which includes the high-VDD boost attenuation control. To write a "1" into the bit-cell, BLT is discharged to GND through bit-switch device Nt0 and segment device Ntu. Shortly after BLT reaches GND, the gate of Nd is shut off by the falling edge of signal WS0n and WS1n transitions to GND to boost BLT below GND. Ntu is selected by the combination of true write data, WDTn and upper write select, WSELn. Boost voltage is increased as the gates of the three unselected segment devices (Gcu/Gtl/Gcl) are also boosted below GND. A 0 V V_{GS} across the unselected segment devices guarantees full isolation and no loss of charge. As a result, a boost with minimal charge loss is delivered to the selected bit-line for writing. Most of the improvement in boost voltage is attributed to this scheme alone.

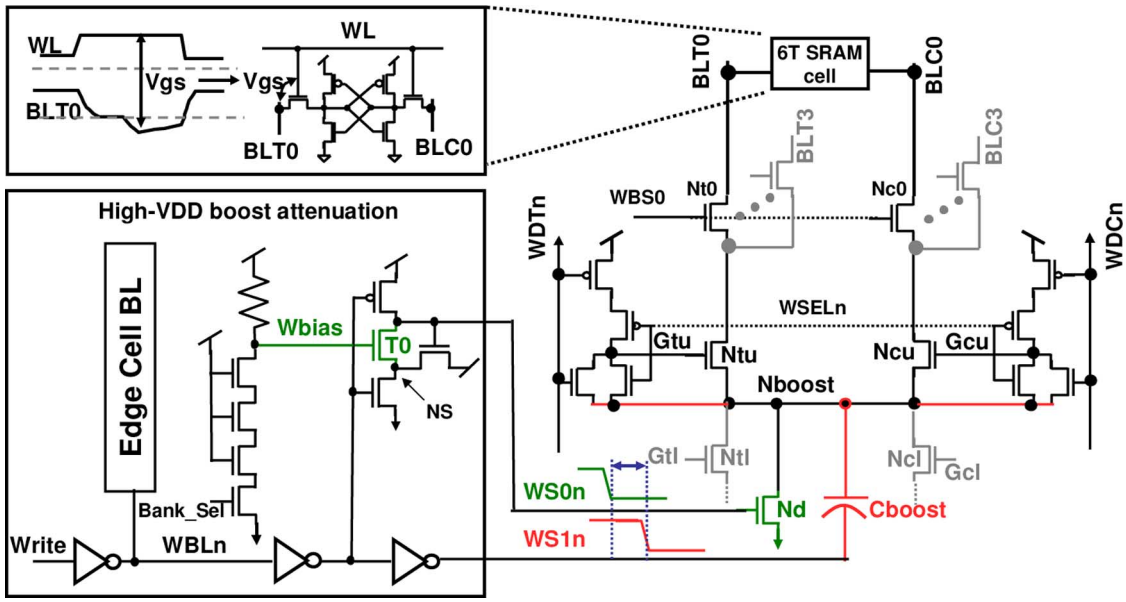


Fig. 11. Write assist implementation.

To minimize the gate-dielectric stress of the write path as the bit-line is boosted below GND, a boost control scheme reduces the boost voltage at higher VDD where full assist is not required. The amount of boost is determined by the separation of WS0n and WS1n that control Nd and Cboost, respectively. An array edge-cell bit-line is used as a write path load to accurately time the initiation of the boost after bit-line is discharged to GND. This is a key requirement; applying the bit-line boost too early decreases the effectiveness of the write operation as the charge transfer of Cboost would occur before the bit-line is pre-discharged to GND. Similarly, applying the boost too late will produce the largest boost voltage at the bit-line, but will also reduce the write window as the WL is unselected shortly after the boost is applied. A write bias generator, Wbias is activated during a write cycle to the decoded bank. Wbias sets the V_{GS} of T0 (Wbias—NS) to modulate the delay of WS0n with respect to WS1n. At high VDD, Wbias is lowered by the increased strength of the four-NFET stack and WS0n is delayed compared to WS1n. The transition of WS1n before WS0n depletes the charge on Cboost by on-device Nd and the boost is attenuated. At lower VDD, WS0n switches low prior to WS1n to prevent the charge of Cboost to drain across Nd and the boost is maximized. The largest boost occurs at low VDD where write margin is needed the most. Fig. 12 compares waveforms at 0.7 V (-198 mV boost) and 1.0 V (-66 mV boost). The timing relationship between WS0n/WS1n for the two VDD cases is shown. The Wbias and NS waveforms show that the delay of WS0n in relation to WS1n is caused by the reduced T0 overdrive as indicated in the figure. The maximum bit-line boost as a function of VDD is also plotted in Fig. 12. The dotted lines represent the bit-line boost level without attenuation. Without boost attenuation, the amount of bit-line boost voltage increases incrementally as VDD increases. Voltage stress is reduced by 200 mV at 1.2 V/-40C. Both temperature and process variation can affect the timing relationship between WS0n and WS1n. The primary driver in the delay difference between WS0n and WS1n is the V_{GS} bias of transistor T0 in Fig. 11. To minimize the variation of V_{GS} for T0 as temperature and process varies, Wbias is generated using a precision resistor that has a small variation to process and temperature. More importantly is to ensure that lower voltage boost operation is not compromised at the worse process corner for write-ability (slow-NFET, fast-PFET) as temperature varies. The separation between WS0n and WS1n is designed with enough delay at lower voltages to not attenuate the boost. The write assist scheme carries an additional 8% AC power consumption overhead for the wide-IO macro configuration and 128-bit-cell macro granularity.

Fig. 13 shows hardware results for the Slow-Pass-Gate, Fast-Pull-Up (SF) process window which typically sets the minimum operating voltage for a write cycle. The vertical axis shows the fail count for 64 Mb SRAM die. The cross-hatched bars indicate fail count when write assist is disabled. This is compared with write assist enabled (solid bars). A reduction in fail count of nearly five orders of magnitude is measured in this extreme process corner at 0.7 V.

V. Read Assist—Signal Margin Tracking

To achieve high yielding SRAM memory over a wide process, voltage and temperature window, it is necessary to employ separate Vt implants for the SRAM devices (i.e., Pass-Gate, Pull-Down and Pull-Up devices) that are decoupled from the logic transistors. This, however, may cause conventional timing circuits composed only of logic-type transistors to track poorly with critical SRAM timings. This is especially true in the establishment of the sense amp set time. This is the key component of the SRAM access time and is critical to the SRAM cycle time in that it determines the minimum word line pulse width during a read operation. Ideally, the timing circuit that establishes the set time should be only a function of the Iread current through the Pass-Gate and Pull-Down devices connected to the low storage node of the selected bit-cell [6].

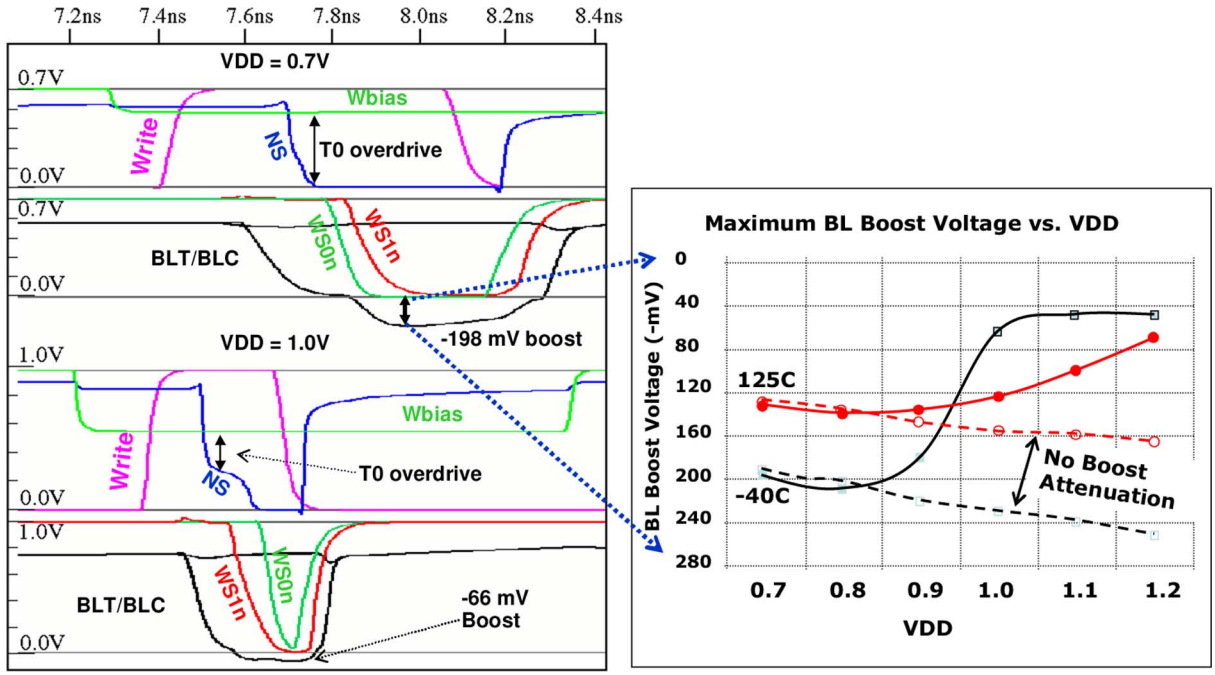


Fig. 12. Write assist simulation.

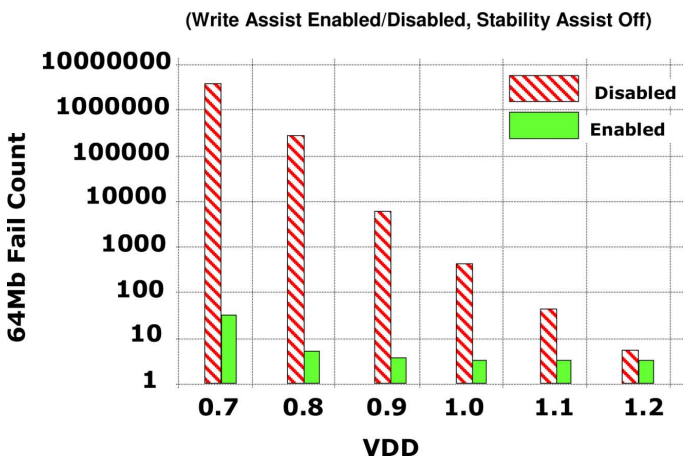


Fig. 13. Write assist hardware results at SF corner.

This, of course, would vary with every bit-cell due to the VT dopant variations, so averaging is required.

To improve the tracking process, the timing circuit shown in Fig. 14 is used for this design. This circuit consists of two separate paths that are each pre-charged high when the macro clock is low. One path (Nbitcell) contains multiple SRAM bit-cells connected in parallel and the other path (Nlogic) uses only conventional logic gates. When the macro clock rises, both paths discharge separate timing capacitors and the slower of the two paths determines the tripping of the NOR circuit to activate the sense amp and to terminate the word line pulse.

The Nbitcell path consists of a small SRAM block (8 rows by 2 columns) surrounded by a full complement of edge bit-cells. Each bit-cell uses exactly the same layout as that in the full SRAM macro, but the Metal-2 True bit-line and the Metal-2 wire that normally connects each bit-cell's "true-side" Pull-Down device to GND are connected to the bit-cell supply terminal (Vcs) instead. In this way, all the true nodes in the timing array are forced to be high. The complement bit-line for the 16 bit-cells is connected to a fixed capacitor and to an edge-cell bit-line from the nearest sub-block in the macro SRAM. The edge-cell bit-line is a replica bit-line structure that has equal capacitance to a normal bit-line. This allows the capacitive load to vary when the array is configured with a different number of bit-cells per bit-line. The fixed capacitor is built using a thin-oxide depletion-MOS capacitor similar to that used for the write-assist. The word lines for the timing bit-cells are activated when the macro clock goes high (gating the macro clock signal to different number of rows in the timing array and multiplexing different capacitive loads in both paths of the timing circuit were used to vary the set time after the design is manufactured). Sixteen bit-cells were used in parallel to reduce the Iread variation relative to the Iread mean value (i.e., σ/μ) to one fourth that of a single bit-cell. Larger timing arrays could be used to reduce this further, but they would require larger fixed capacitors and use more silicon area. Pass-Gate devices typically have a much higher Vt than the NFET logic devices (for bit-cell stability reasons); therefore, the Nbitcell path is much slower at low voltages than the Nlogic path, but at higher voltages, the Nlogic delay is slower than the Nbitcell delay. Consequently, the Nlogic path's device sizes and capacitor are chosen to just achieve reliable word line pulse widths (and up levels) at the fastest process corners and highest operating voltages. Given the typical mismatches of the RC time constant between Word-line signals and sense-amplifier set signals, it is required to have a minimum base delay beyond what the bit-cell tracking can produce. This guarantees that at high-voltage, high temperature operation, where the RC time constant is the slowest, but device performance is fastest, the sense-amp signal development budget is not compromised. In an ideal design where signal slew rates are perfectly matched

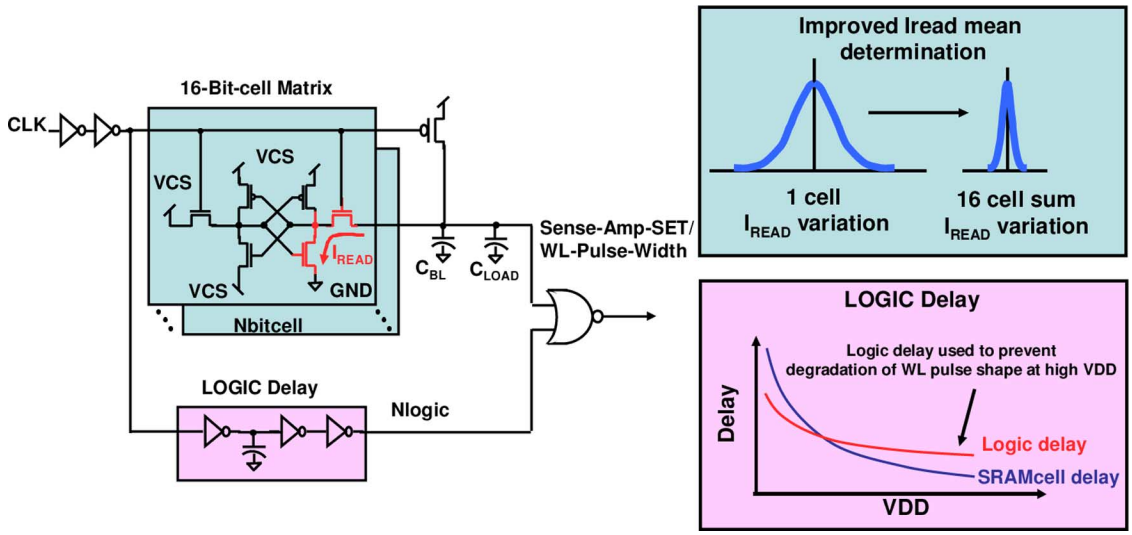


Fig. 14. Read assist (bit-cell tracking) concept.

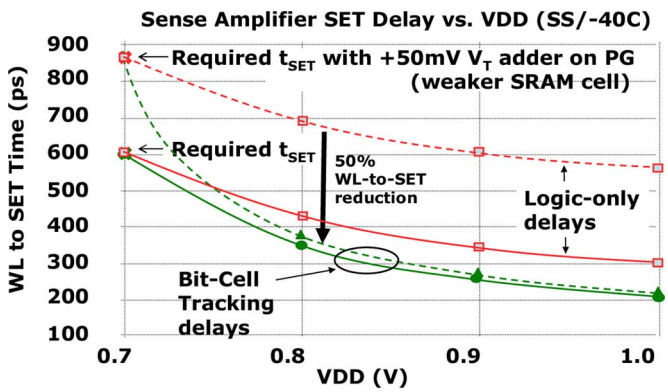


Fig. 15. Read assist simulation.

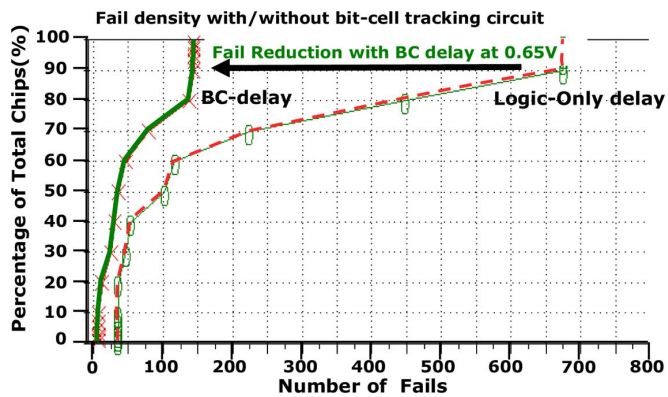


Fig. 16. Read assist hardware results.

between Word-line and sense-amplifier set signals, this logic delay component may not be required.

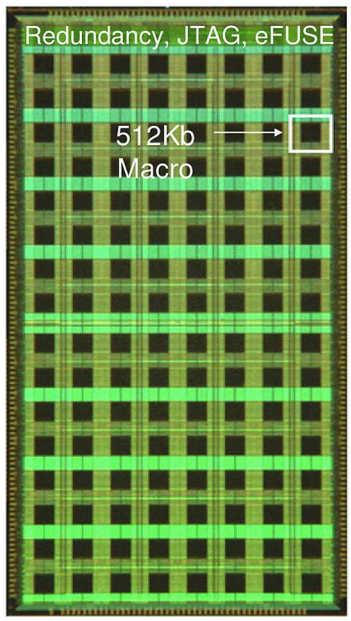
Fig. 15 shows the advantages of this scheme when low voltage functionality is required along with performance requirements at a higher voltage. The performance requirement is at 0.8 V, but the circuit must also operate at 0.7 V (at significantly lower performance and lower power). The bottom solid line is the set time produced by this circuit at the Slow-NFET, Slow-PFET (SS) corner, at -40 C, for various Vdd settings. The circuit in Fig. 14 generates a 600 ps set time at 0.7 V and a 360 ps set time at 0.8 V. A logic-only timing implementation (shown by the bottom dashed line) would produce a set time of 420 ps at 0.8 V in order to achieve the required functional 600 ps set time at 0.7 V (i.e., a 16% increase in set time at the performance corner). If the SRAM Pass-Gate and Pull-Down devices are systematically slower than the NFET logic devices, the situation is much worse. Suppose the Pass-Gate is 50 mV weaker. The circuit produces a set time (shown by the top solid line) of 875 ps at 0.7 V and 390 ps at 0.8 V. The logic-only timing circuit (top dashed line) requires a 690 ps set time at 0.8 V in order to obtain a 875 ps delay at 0.7 V which represents a 77% increase in set time.

Fig. 16 shows hardware results for the read assist scheme. The vertical axis represents the percentage of 64 Mb die that meet the criteria in the horizontal axis: fail density. The dashed line shows the number of fails for a logic-only delay and the solid line shows the number of fails for the proposed bit-cell-tracking delay scheme. A large fail count reduction is measured for all die at 0.65 V.

VI. Test Chip Implementation and Hardware Results

Fig. 17 shows the micrograph of the test chip and its design features. The 64 Mb SRAM is built from 128 512 Kb macros each 331 μ × 339 μ in size. The Sub-Array configuration consists of 128 Word-lines × 256 Bit-lines. The operating voltage range of the core and the SRAM is 0.7 V –1.0 V with a performance target of 1.4 GHz at a Slow process corner and 0.8 V, 0 °C. This design is used as the principal building block for high-performance ASIC SoC. The test chip design includes full redundancy capability as well as design-for-test diagnostics, including at speed cycle-time, access time and setup/hold testing.

Fig. 18 shows hardware results at 85 C. The fail-count is compared with assist features disabled (left) and enabled (right). The VCS supply (array and WL-driver) is plotted against the VDD supply (periphery). SRAM operation is shown down to 0.7 V. An overall improvement of 400 mV of VCS is observed when these features are enabled compared to the default state. The



Technology	32nm PD SOI with High-k Metal Gate
Cell Size	0.154μm²
512Kb Macro Size	331μ x 339μ
Sub-Array Configuration	128 Word-Line x 256 Bit-Line
Operating Voltage	Core: 0.7V – 1.0V (0.9V typ.) SRAM: 0.7V – 1.0V (0.9V typ.)
Performance Target	1.4GHz (Slow process, 0.8V, 0C)

Fig. 17. Micrograph of 64 Mb SRAM test chip and features.

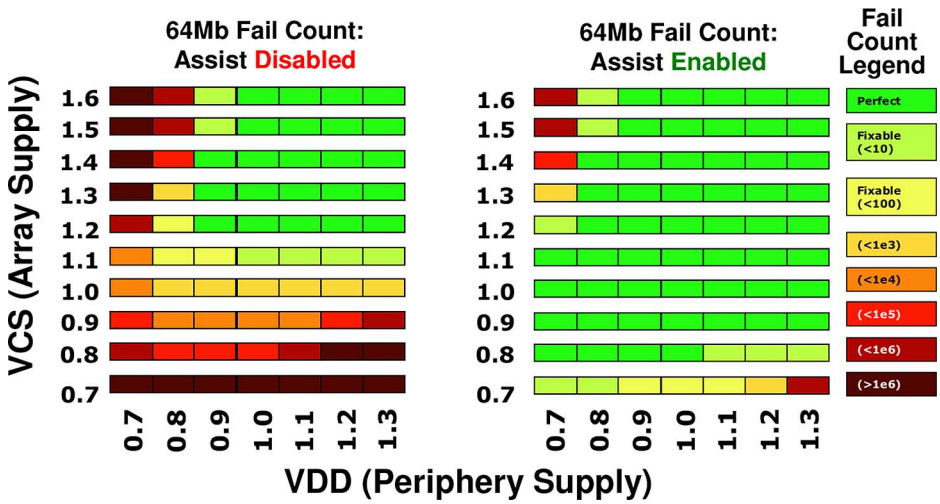


Fig. 18. Hardware results: 64 Mb SRAM 85°C VCS/VDD voltage shmoo.

area overhead for the stability and write assist is 1.5% and 1.2%, respectively. An overall array efficiency of 71.6% is achieved with a 128-Word-line sub-array configuration.

VII. Conclusion

The introduction of High-k Metal-Gate in 32 nm SOI has enabled a significant reduction in the equivalent oxide thickness, facilitating the aggressive scaling of device dimensions and overall area footprints of SRAM memories. However, technology innovation alone is not sufficient to maintain the aggressive area and power scaling demands of future System on Chip designs. Recent assist methods such as stability assist have been effective for decreasing the SRAMs failure rate. However, several of these methods interfere with other operations and require external modulation to preserve the balance between bit-cell stability and write-ability. This work presents three innovative assist features that enhance the stability, write-ability, and read-ability of SRAM memories to allow a reliable 0.7 V $\mathrm{VDD_{MIN}}$ operation and enable a $2\times$ size reduction from the previous 45 nm design. Stability was improved by a bit-line regulation system that does not degrade the write margin. Low voltage write-ability is achieved with a bit-line boost scheme and boost control that features a 40% improvement in bit-line boost voltage compared to the previous designs. Finally, a bit-cell-tracking delay circuit is implemented to improve both performance and yield across the process space. The assist features are implemented on a 64 Mb SRAM test-chip fabricated in a 32 nm High-k Metal-Gate SOI technology using a 0.154 $\mu\mathrm{m}^2$ bit-cell.

References

[1] B. Greene *et al.*, "High performance 32 nm SOI CMOS with high-k/metal gate and 0.149 $\mu\mathrm{m}^2$ SRAM and ultra low-k back end with eleven levels of copper," in *Proc. Symp. VLSI Technology*, 2009.

[2] H. Pilo *et al.*, "A 450 ps access-time SRAM macro in 45 nm SOI featuring a two-stage sensing-scheme and dynamic power management," in *IEEE ISSCC Dig.*, 2008, pp. 378–379.

[3] A. Bhavnagarwala *et al.*, "A sub-600 mV, fluctuation tolerant 65 nm CMOS SRAM array with dynamic cell biasing," in *Proc. Symp. VLSI Circuits*, Jun. 2007, pp. 78–79.
[4] Y. Fujimura *et al.*, "A configurable SRAM with constant-negative-level write buffer for low-voltage operation with 0.149 μm^2 cell in 32 nm high-k metal-gate CMOS," in *IEEE ISSCC Dig.*, 2010, pp. 348–349.
[5] P. Kolar *et al.*, "A 32 nm high-k metal gate SRAM with adaptive dynamic stability enhancement for low-voltage operation," in *IEEE ISSCC Dig.*, 2010, pp. 346–347.
[6] B. Amrutur *et al.*, "A replica technique for wordline and sense control in low-power SRAMs," *IEEE J. Solid-State Circuits*, pp. 1208–1219, Aug. 1998.

Harold Pilo (M'07) received the B.S.E.E. degree from the University of Florida, Gainesville, in 1989.

He was with Motorola in Austin, TX, from 1989 to 1993 where he designed specialty memory products. In 1993, he joined IBM in Burlington, VT, to develop IBM's OEM SRAM and eDRAM products for the IT industry. He is presently a Senior Technical Staff Member responsible for circuit IP development for ASIC SRAM.

Mr. Pilo has presented many papers and lectures at the ISSCC, ITC, IEDM and VLSI Circuits Symposium. In 2003 he was the recipient of the ISSCC Beatrice Winner Award for Editorial Excellence and in 2005 he received the Special-Topic Session Award at the ISSCC. He was member of the ISSCC Memory Subcommittee from 2008 to 2011 and has authored several papers for the IEEE JOURNAL OF SOLID-STATE CIRCUITS and other publications. He presently holds more than 60 U.S. patents.

Igor Arsovski received the B.S. and M.S. degrees in electrical and computer engineering from the University of Toronto, ON, Canada, in 2001 and 2003, respectively.

He joined IBM in Vermont in 2003 to work on development of Ternary Content Addressable Memory (TCAM) for the networking applications. He has been the main architect of the IBM's TCAM compiler responsible for the TCAM architecture and circuits for four TCAM generations (90 nm to 32 nm). He currently works on ASIC memory strategy and advanced custom memory solutions. He has filed 40 patents and published numerous papers at premier conferences.

Mr. Arsovski received the best paper award from IEEE CICC 2006, the University of Toronto Centennial Thesis Award, and IBM's Outstanding Technical Achievement Award. He is currently a Solid-State Memory reviewer for the IEEE JOURNAL OF SOLID-STATE CIRCUITS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS and the *VLSI Journal*.

Kevin Batson received the B.S. and M.S. degrees in electrical engineering from the Polytechnic Institute of NYU, New York City, in 1989 and 1993, respectively.

He joined IBM as an SRAM Circuit Designer in 1993. He has worked in the design of ASIC SRAM arrays and test-sites, stand-alone SRAM, Content Addressable Memory (CAM), and embedded SRAM Cache memory. He has filed 14 patents.

Geordie Braceras received the B.S.E.E. degree from Lehigh University, Bethlehem, PA, in 1983.

He has focused on SRAM design since joining IBM in 1983, completing numerous product designs and qualifications, from embedded cache design to stand-alone SRAM products. He presently has 46 issued patents, and has authored or co-authored numerous papers on SRAM design.

Mr. Braceras is presently an IBM Distinguished Engineer leading the ASIC embedded memory effort for the Silicon Solution Engineering organization.

John Gabric is a Senior Memory Development Manager at IBM, Essex Junction, VT. He joined IBM in 1970 after graduating from the University of Akron with a B.S.E.E. His contributions include DRAM product design and development from 64Kb to 512Mb and SRAM development for stand-alone and embedded applications in Bulk and SOI technologies. He has seven issued patents and has authored ten publications and eight technical papers. He is currently managing product development of MRAM, PCM, NVRAM, and SRAM in the Systems and Technology Group. His interests include 22nm ASIC compiler development and 14nm SRAM bit cell optimization and SRAM yield optimization vehicles.

Robert Houle received the B.A. degree in mathematics in 1974, the B.S. degree in electrical engineering in 1976, both from the University of Vermont, and the M.S. degree in electrical engineering in 1980 from the University of Connecticut.

He worked for the Naval Underwater Systems Center in New London, CT, for four years before joining IBM in Essex Junction, VT, in 1984. He has worked primarily on circuit design for microprocessors, with an emphasis on SRAM since 1996.

Steve Lamphier received the B.S. degree in electrical engineering from Clarkson University, Potsdam, NY, in 1992.

He joined IBM, Essex Junction, VT, in 1992, in the High Performance SRAM area. His work experience includes chip floor planning, wafer characterization, cell design, package laminate design, ESD network, chip logical verification, chip design rule verification, chip logical to physical verification, redundancy, I/O circuits, chip critical timing analysis, and microprocessor integration. He is currently an Advisory Engineer working on high performance SRAM ASIC macros.

Carl Radens received the B.A. degree in physics from Oberlin College, Oberlin, OH, in 1983, and the Ph.D. degree in electrical engineering from the University of Cincinnati, Cincinnati, OH, in 1990.

He is a Senior Technical Staff Member (STSM) at the IBM Semiconductor Research and Development Center (SRDC) working on the development of memory technology and SRAM. Since joining IBM in 1990 he has worked on memory cell design, FEOL and BEOL process integration, device, interconnect, SOI and bulk technology, SRAM, DRAM, and process development.

Dr. Radens has served on the IEEE IEDM Integrated Circuits and Manufacturing subcommittee, as an invited panelist at the IEEE VLSI Technology and Circuits Symposium, has coauthored more than 36 technical publications, and holds over 239 issued US patents.

Adnan Seferagic received the B.S.E.E. and M.S.E.E. degrees from the University of Vermont, Burlington, in 2006 and 2009, respectively.

In 2004, he interned with Huber & Suhner in Essex Junction, VT, where he was involved in RF characterization of wireless cable assembly products for the communication industry. In 2006 he joined IBM in Essex Junction, VT, where he worked as an RF characterization engineer responsible for a Bipolar and CMOS device measurements in support of IBM's device modeling group and device design kit efforts. Since 2008 he has been working on SRAM designs for ASIC technology development test chips.