19.5 A 512kb 8T SRAM Macro Operating Down to 0.57V with An AC-Coupled Sense Amplifier and Embedded Data-Retention-Voltage Sensor in 45nm SOI CMOS

Masood Qazi[1], Kevin Stawiasz[2], Leland Chang[2], Anantha Chandrakasan[1]

[1]Massachusetts Institute of Technology, Cambridge, MA
[2]IBM T. J. Watson, Yorktown Heights, NY

There is a need for large embedded memory that operates over a wide range of supply voltage compatible with the limits of static CMOS logic that also minimizes standby power [1,2]. A 512kb 8T SRAM macro in 45nm SOI CMOS is developed as a building block for this application. It addresses the design challenges related to area efficiency and process variation with three contributions: 1) an AC coupled sense amplifier (ACSA) that operates at a power supply ultimately limited by the worst-case bit line on/off current ratio; 2) area efficient, regenerative driving of long data lines to permit bidirectional signaling on a single metal 4 wiring track on the memory cell column pitch; 3) a data retention voltage (DRV) sensor to determine the mismatch-limited minimum standby supply voltage without corrupting the contents of the memory.

The ACSA is shown in Fig. 19.5.1. During precharge (ϕ=0), device M5 diode-connects the sensing PMOS M3, charging the input capacitor to the V_T of M3. Loading from devices M6-M8 and stacking from M4 establishes a steep transfer characteristic near the tipping current level at the dynamic output Z. A thick oxide MOS capacitor achieves a coupling ratio close to 1 while keeping the area of the entire ACSA under 9μm^2. During sensing (ϕ=1, RWL=1), node X is initialized 50mV above the tipping point of the ACSA through C_{gd} coupling and charge injection from M5. When reading a 1, devices M2 and M1 discharge the local bitline, and after 100mV of signal development the output current quickly charges node Z to VDD. Waveform plots B and C in Fig. 19.5.3 show amplification when reading 1 and stability when reading 0.

The conventional domino read path [3], which replaces the ACSA by an optimally sized PMOS device of the same V_T, is 1.8x slower (Fig. 19.5.1). Monte Carlo simulation of the 5σ timing window reveals that additional variation of the dynamic PMOS also restricts the minimum operating VDD to 0.9V; whereas, the ACSA works down to 0.53V (FF corner, 85C). Small-signal, single-ended sensing with offset compensation has also been employed in [5] and [6]. By AC coupling, this work avoids the unconditional full V_T swing on the bitlines and the requirement for interlock between wordline access and precharge found in [6]. By exploiting the sharp cutoff characteristics of a PMOS stack M3-M4 to separate 1 from 0, this work avoids the 2 capacitors, 13 transistors, and power penalty related to regenerative feedback in [5]. Hence, 128 parallel ACSAs can be laid out on the bitline pitch, satisfying the non-interleaved column constraint of low-voltage 8T SRAM to avoid half-selected write operation.

After local sensing, data is forwarded across eight 64kb banks in the read cycle via the regenerative global bitline scheme (RGBS) shown in Fig. 19.5.2. Two ACSAs with shorted outputs are buffered to the global pull-up PMOS Mgp. When accessing bank0, device Mgp needs only to charge up the GBL input to the neighbor bank1 above the V_T of NMOS Mxu/Mxl to trigger the next ACSA, turning on device Mgp in bank1 and so on until the data traverses the 1.7mm global bitline (plot D Fig. 19.5.3). The simulated delay of the RGBS is 40% faster than the conventional pull down NMOS. The RGBS scales linearly with distance, preserving sharp transitions and performance comparable to a buffering daisy chain [7]; yet the RGBS avoids cross-over currents and enables bidirectional signaling—preventing routing congestion of the 128b word—through a single metal 4 track on the memory cell pitch.

The access times in Fig. 19.5.3 equal the minimum difference between the falling edge of CLK and the rising edge of ACLK (Fig. 19.5.2). Each measurement corresponds to 100% pass of all 512kb under alternating checkerboard and blanket data patterns. Two measurements below 0.65V require partial turn-on of a bleeder PMOS device on the bitlines to compensate for fast process corner leakage. Performance scales from 400ps at 1.2V to 3.4ns at 0.57V. Measured leakage power scales down 9.4x over this range. The read path benefits from CMOS scaling by avoiding differential sense amplifiers, which are sensitive to mismatch, and variable timing control signals [4,5].

For idle banks, lowering the supply to the DRV—measured to lie between 0.375V and 0.4V—reduces leakage power by 29x going from 1.2V to 0.4V. Generally the minimum supply voltage is determined by the one memory cell out of 524,288 with the largest DRV, resulting from local mismatch variation and its functional relation to the static noise margin (SNM) of the memory cells. This relation changes with process corner, temperature, and end-of-life degradation. Prior work [8,9] recognizes the importance of tracking the DRV but relies on apriori characterization of the main memory. The primary challenge remains: how to determine the DRV without corrupting the contents of the main memory?

In this work, a separate column of 256 cells is employed in conjunction with statistical techniques as a DRV sensor to predict the 5σ failure rate of the main memory (Fig. 19.5.4). The test chip contains 4096 DRV sensor cells but only one segment of 256 is measured to demonstrate the area efficiency of the technique. The layout of DRV sensor cells is identical to the main memory cells except for one modification: the metal 2 wire supplying VDDAR is split into two separate VDDL and VDDR supply lines. Along with VSSL and VSSR, the four independent supplies of the DRV sensor cells are used to program and skew without accessing the main memory.

Given a 1σ measurement of V_T variation in memory cell devices, the DRV sensor is skewed in multiple directions to reconstruct the SNM as a linear function of V_T. The resulting coefficients are combined with σ_{Vt} to estimate failure. The detailed algorithm in Fig. 19.5.5 searches the magnitude of three linearly independent voltage skews that collapse the nominal SNM, as observed by 50% failure in the 256 memory cells. The application of voltage skews emulates an effective shift in threshold voltage—through a transformation represented by T in the matrix equation of Fig. 19.5.4—since the drain current depends on ($V_{GS} - V_T$) and the voltage skews directly add to or detract from V_{GS}. By simplifying coefficients (assuming M5 and M6 do not influence retention and M3 and M2 have similar influence on the SNM), a system of three equations and three unknowns can be solved. The measurement results in Fig. 19.5.5 show that the predicted retention failure matches the actual retention failure in the main memory. A conservative estimate is also produced by replacing the failure criterion of 50% with 25% in the prediction algorithm, introducing a measured margin of 40mV. Thus, for real-time embedded operation, the DRV sensor can permit aggressive reduction of retention standby voltage with negligible risk of losing data. This technique is relevant to state-of-the-art embedded SRAM that requires the retention voltage as an input to standby power regulation circuits [10].

Figure 19.5.6 lists key features of the fabricated 8T SRAM, a viable candidate for voltage scalable operation down to 0.57V. The die photo in Fig. 19.5.7 illustrates the 2.0mm x 0.35mm chip floor plan containing 8 banks of 64kb along with on-chip testability circuits.

Acknowledgements:
This work was funded in part by the FCRP Focus Center for Circuit & System Solutions (C2S2), under contract 2003-CT-888. The authors thank Robert Montoye for technical discussion and Keith Jenkins for test support.

References:

[1] D. J. Frank, W. Haensch, G. Shahidi, O. Dokumaci, "Optimizing CMOS Technology for Maximum Performance," *IBM J. Res. Dev.,* vol. 50, pp. 419, Jul., 2006.
[2] R. Krishnamurthy, et al., "High-performance, Low-power, and Leakage-tolerance Challenges for Sub-70nm Microprocessor Circuits," *ESSCIRC Dig. Tech. Papers,* pp. 315-322, Sept., 2002.
[3] L. Chang, et al., "A 5.3GHz 8T-SRAM with Operation Down to 0.41V in 65nm CMOS," *Dig. Symp. VLSI Circuits,* pp. 252-253, Jun., 2007.
[4] K. Zhang, et al., "The Scaling of Data Sensing Schemes for High Speed Cache Design in Sub-0.18μm Technologies," *Dig. Symp. VLSI Circuits,* pp. 226-227, Jun., 2000.
[5] N. Verma and A. P. Chandrakasan, "A High-Density 45nm SRAM Using Small-Signal Non-Strobed Regenerative Sensing," *ISSCC Dig. Tech. Papers,* pp. 380-381, Feb., 2008.
[6] I. Arsovski and R. Wistort, "Self-referenced Sense Amplifier for Across-chip-variation Immune Sensing in High-performance Content-Addressable Memories," *CICC Dig. Tech. Papers,* pp. 453-456, Sept., 2006.
[7] K. Zhang, et al., "A Fully Synchronized, Pipelined, and Re-Configurable 50Mb SRAM on 90nm CMOS Technology for Logic Application," *Dig. Symp. VLSI Circuits,* pp. 253-254, Jun., 2003.
[8] J. Wang and B. Calhoun, "Canary Replica Feedback for Near-DRV Standby V_{DD} Scaling in a 90nm SRAM," *CICC Dig. Tech. Papers,* pp. 29-32, Sept., 2007.
[9] H. Qin, Y. Cao, D. Markovic, A. Vladimirescu and J. Rabaey, "SRAM Leakage Suppression by Minimizing Standby Supply Voltage," *ISQED Dig. Tech. Papers,* pp. 55-60, Mar., 2004.
[10] Y. Wang, et al., "A 4.0 GHz 291Mb Voltage-Scalable SRAM Design in 32nm High-k Metal-Gate CMOS with Integrated Power Management," *ISSCC Dig. Tech. Papers,* pp. 456-457, Feb., 2008.

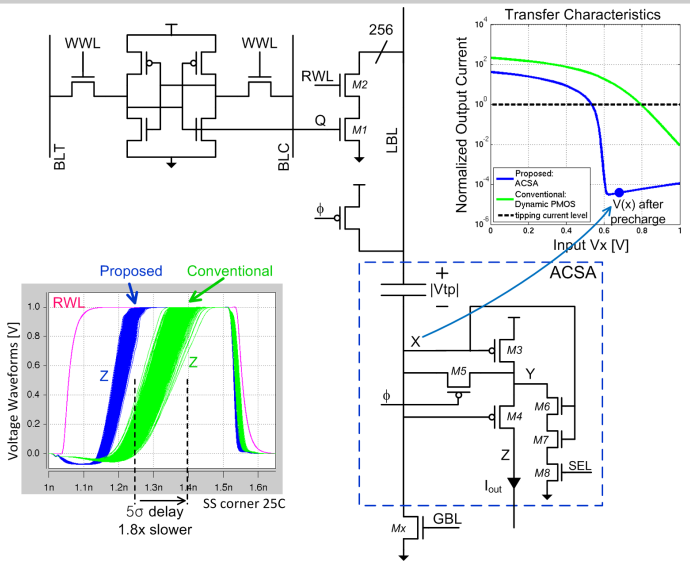

Figure 19.5.1: The AC coupled sense amplifier supports a bitline of 256 cells.

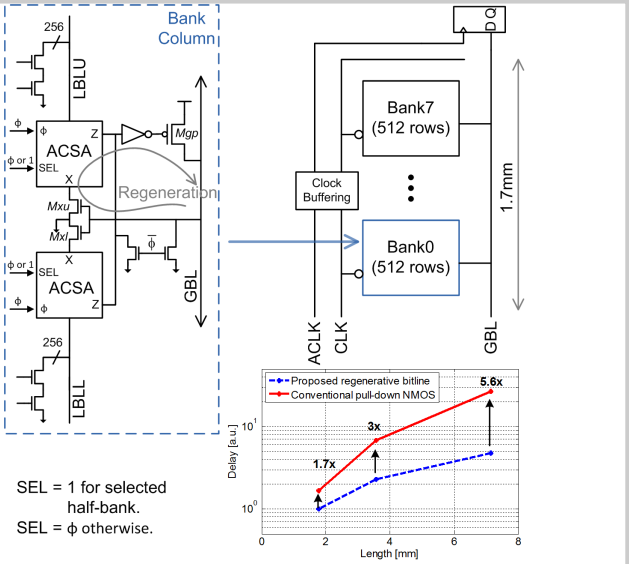

Figure 19.5.2: The regenerative global bitline scheme reuses the ACSA.

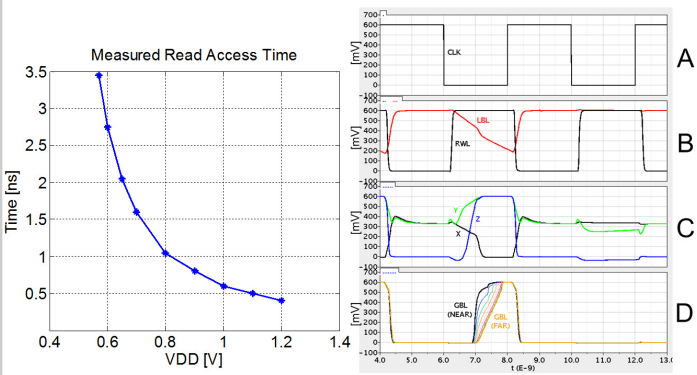

Figure 19.5.3: Read access time measurement and simulated operational waveforms at 0.6V.

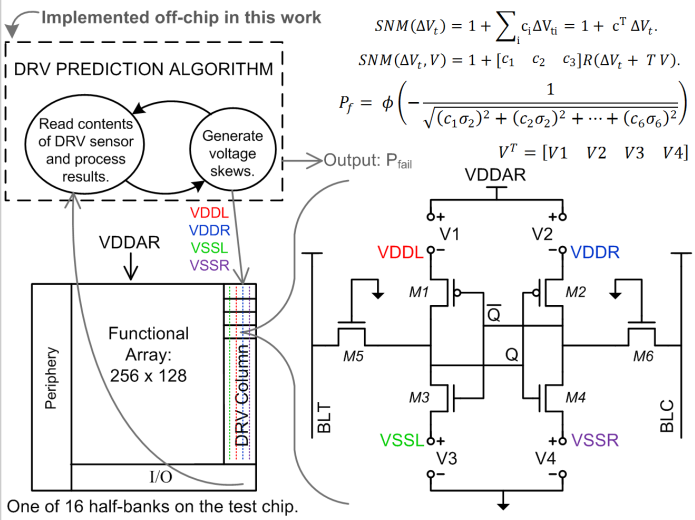

Figure 19.5.4: The DRV sensor with voltage skews illustrated.

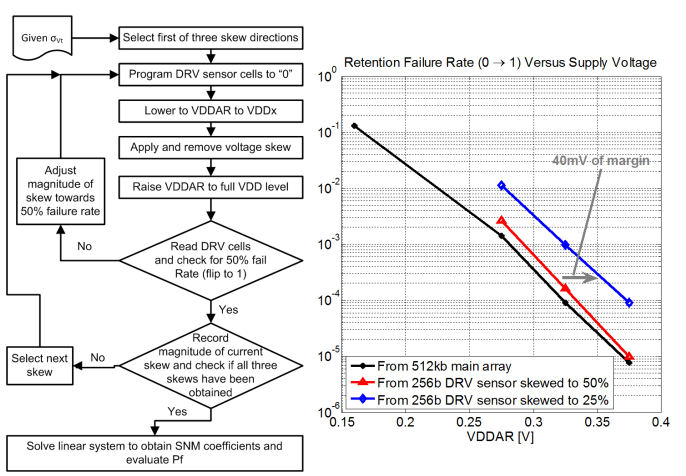

Figure 19.5.5: DRV Sensor algorithm and measurement results from test chip.

Organization	4096 words x 128b (in 8 banks of 64kb)
Technology	45nm High-Performance SOI
Cell area	0.578um^2
Sense Circuit Area Supporting 512b	20.9um^2
Access time at 1.2V	400ps
Access time at 0.57	3.4ns
Active Power at 1.2V (extrapolated from 100MHz to 1.25GHz)	169mW
Leakage Power at 1.2V	338mW

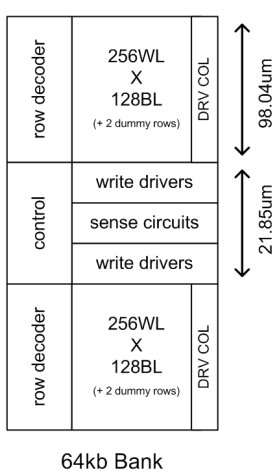

Figure 19.5.6: Summary of chip characteristics and 64kb bank floor plan.

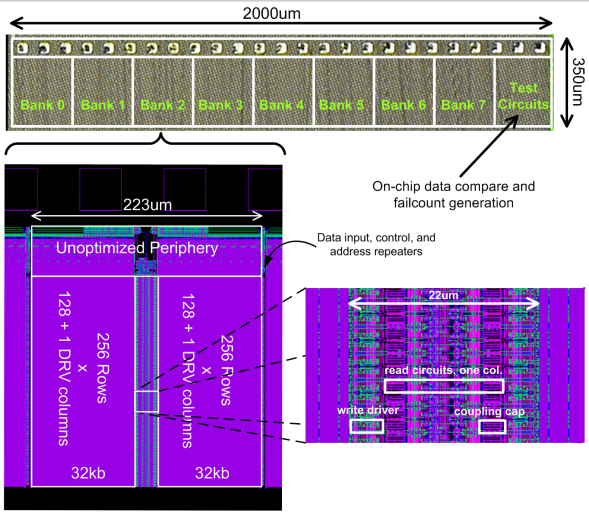

Figure 19.5.7: Die photo of the 512kb SRAM test chip in 45nm SOI CMOS along with a close-up snapshot of the bank layout. Global bit lines run 1.7mm long from left to right across 8 banks of 64kb each.

978-1-4244-6034-2/10/$26.00