18.4 A 65nm 8T Sub-V_t SRAM Employing Sense-Amplifier Redundancy

Naveen Verma, Anantha P. Chandrakasan

Massachusetts Institute of Technology, Cambridge, MA

The subthreshold regime is a critical biasing space as it enables minimum energy operation for logic circuits [1]. However, practical systems rely heavily on SRAMs, which conventionally limit the minimum V_{DD} to above V_t. SRAMs often dominate the total die area and power, and minimizing their energy requires scaling V_{DD} as low as possible. In this work, a 256kb SRAM in 65nm CMOS is presented that operates in sub-V_t (at 350mV) despite the exponential effect V_t variations have on device strength.

The 6T bit-cell in Fig. 18.4.1 provides a good balance between stability, performance, and density. However, in the presence of variation, it fails to operate in sub-V_t. Figure 18.4.1 shows a Monte Carlo simulation of the SNM [2] for both read and hold cases of a 65nm cell. At 350mV, hold stability is preserved, but read failures are prominent. Write SNM violations (not shown) appear in a similar manner. Functional errors are also caused by severely degraded I_{READ}. Figure 18.4.1 considers the case of 256 cells per column. In sub-V_t, the values stored in the unaccessed cells can result in an aggregate leakage current on the shared bitlines that is greater than the 3σ and 4σ read currents, implying that the data in the accessed cell is indistinguishable from bitline leakage.

To overcome these challenges, the 8T bit-cell shown in Fig. 18.4.2 is developed. Buffered read eliminates the read SNM limitation; peripheral footer circuitry eliminates bitline leakage; peripheral write drivers and storage-cell supply drivers interact to reduce the cell supply voltage during write operations; and sense-amp redundancy provides a favorable trade-off between offset and area. Previous implementations of sub-V_t memories deal with stability, read-current, and bitline leakage by adding devices within the cell or employing hierarchy to limit fan-in/out. For instance, a 10T cell operates at 400mV [3], and a register-file uses multiplexed read to operate at 310mV [4]. In this design, peripheral circuit assists are used to maximize density and reduce the leakage paths to those of a 6T SRAM.

In Fig. 18.4.2, the read buffer is composed of M7-M8. Instead of statically connecting its foot to ground, however, a foot-driver is used in the periphery. As shown in Fig. 18.4.3, the buffer-foots of all cells of the same word are shorted, and their foot-driver is shared. During a read, only the foot of the accessed word is driven low; all others remain at V_{DD}. Accordingly, after RDBL is precharged, the read-buffers of the unaccessed cells have no voltage drop across them, and their access devices have a negative V_{GS}. Consequently, they impose no sub-V_t leakage, and dynamically held data values of “1” on RDBL can be sensed successfully.

The foot-driver is required to sink the read current from all of the accessed cells. Use of a large NMOS to accomplish this is impractical since it would impose a significant area and leakage-power overhead. Instead, the sub-V_t charge-pump circuit shown in Fig. 18.4.3 is used. The voltage boost provided by typical charge-pump implementations suffers from V_t drops, and would be inadequate for this application. Instead, the circuit of Fig. 18.4.3 uses a PMOS (M1) to precharge C_{BOOST}. The charge-pump generates a swing of nearly $2V_{DD}$ at the input of the foot-driver, enhancing its current by over two orders of magnitude while reducing V_t variation dependencies on its devices. This allows the devices of the foot-driver to be near minimum sized so that their leakage-power is insignificant. Further, since the charge-pump drives minimal load, its devices and boost capacitor can be small, consuming negligible power and area.

Write operations fail when the cell pass devices cannot overpower the internal cell feedback. In this design, write (Fig. 18.4.4) is performed by boosting WL by 50mV and, more importantly, reducing VV_{DD} through a supply driver. Simultaneously, new data is written primarily by pulling the desired storage node low through the NMOS pass device. Although, the opposite storage node is only weakly pulled high, its load PMOS provides a current path to VV_{DD}. Accordingly, all cells in the accessed word contribute to driving VV_{DD} high through one of their NMOS pass devices. Relatively large devices are used in the supply driver, and the net variation in the pass devices and write drivers tends to average; hence, sizing accurately allows VV_{DD} to be set to a low intermediate voltage.

The write mechanism, which is essential for sub-V_t operation, requires each word to have a separate VV_{DD}. As shown in Fig. 18.4.5, this implies that columns of different blocks cannot be interleaved in layout, and adjacent columns can no longer share a multiplexed sense-amp. Hence, the number of sense-amps required increases, and each must fit in a column pitch. Nominally, the approach of large-signal read, which is advantageous in high-density, scaled SRAMs [5], is used; nonetheless, the BL voltage levels are degraded, due to gate-leakage and other noise mechanisms, and sense-amp offsets still limit yield. To remedy this, sense-amp redundancy is employed. Erroneous reads occur when the net offset of each sensing network is greater than the input voltage swing. Increasing device sizes reduces local variation, accordingly reducing sense-amp offset. Redundancy, however, allows exclusive selection of the sense-amp that minimizes the achievable offset. Hence, errors now depend on the joint probability that all sense-amps have an offset greater than the input voltage swing. As shown in Fig. 18.4.5, the error probability for a half-sized sense-amp is greater than that for a unit-sized sense-amp; however, Monte Carlo simulation shows that the joint error probability for two half-sized sense-amps is lower than that for a unit-sized sense-amp. Specifically, a factor of five improvement is observed at the input swings of interest (i.e., 50mV). This only applies where the errors due to offset are uncorrelated, so, a pseudo-differential sense-amp structure is employed to cancel the effects of global variation.

Increased redundancy yields further improvement, but the overhead of selecting between redundant sense-amps and storing that selection state also increases. In this design, two sense-amps are used, requiring the minimal support circuitry of two flip-flops and a few logic gates. On start-up a selection routine determines which sense-amp can correctly read both logic “0” and “1”, and enables only the corresponding structure.

The SRAM is fabricated in a 65nm CMOS process (Fig. 18.4.7). The 256kb array is arranged into 8, 256 row × 128 column blocks. Full read and write functionality is achieved with a V_{DD} of 350mV (and 50mV boosting of WL drivers). At this voltage, the SRAM operates at 25kHz and consumes 2.83μW during read and 3.96μW during write. As shown in Fig. 18.4.6, data is held to 300mV where the leakage power is 1.92μW. At 325mV fewer than 0.05% read/write errors are observed.

Acknowledgements:
This work is funded by DARPA, and IC fabrication is provided by Texas Instruments. N. Verma is supported by an Intel Ph.D. Fellowship. The authors thank B. Calhoun and M. Bhardwaj for their support and feedback.

References:
[1] A. Wang, A. Chandrakasan, and S. Kosonocky, “Optimal Supply and Threshold Scaling for Sub-Threshold CMOS Circuits,” *Proc. IEEE Comp. Society Annual Int. Symp. VLSI*, pp. 5-9, Apr., 2002.
[2] E. Seevinck, F. List, and J. Lohstroh, “Static Noise Margin Analysis of MOS SRAM Cells,” *IEEE Journal of Solid-State Circuits*, SC-22, no. 5, pp. 748-754, Oct. 1987.
[3] B. Calhoun and A. Chandrakasan, “A 256kb Sub-Threshold SRAM in 65nm CMOS,” *ISSCC Dig. Tech. Papers*, pp. 628-629, Feb., 2006.
[4] J. Chen, L. Clark, and T.-H. Chen, “An Ultra-Low-Power Memory With a Subthreshold Power Supply Voltage,” *IEEE J. Solid-State Circuits*, vol. 41, no. 10, pp. 2344-2353, Oct., 2006.
[5] K. Zhang, K Hose, V. De, et al., “The Scaling of Data Sensing Schemes for High-Speed Cache Design in Sub-0.18μm Technologies,” *Symp. VLSI Circuits*, pp. 226-227, Jun., 2000.

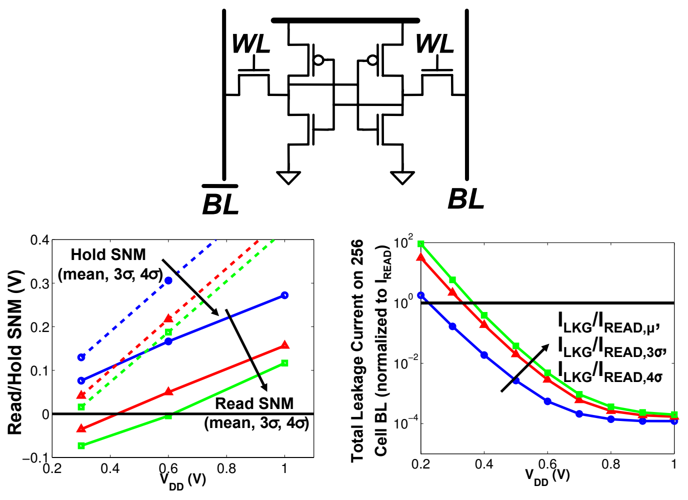

Figure 18.4.1: 6T cell SNM and bitline leakage (normalized to I_{READ}) demonstrating loss of functionality at low voltages.

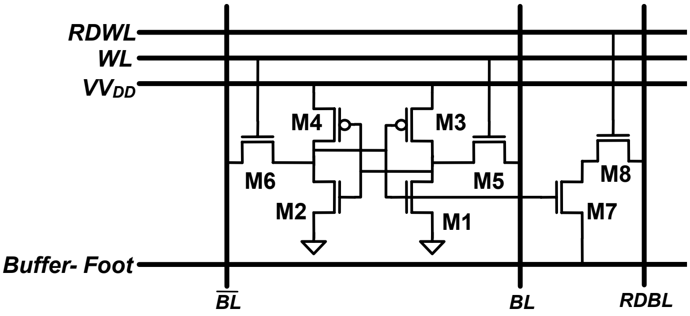

Figure 18.4.2: 8T cell enabling low-voltage read/write and sensing.

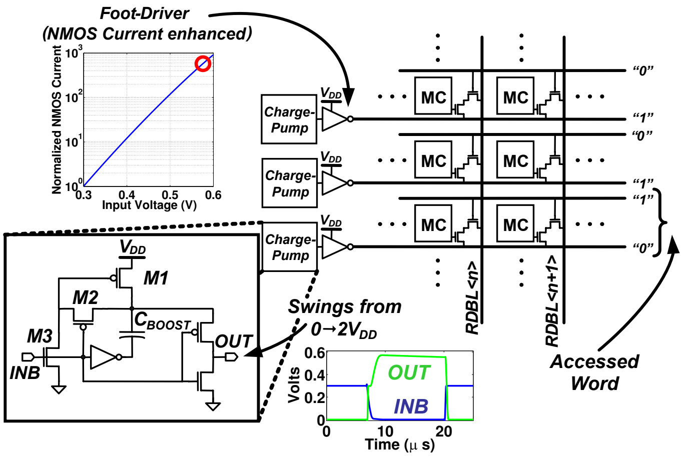

Figure 18.4.3: Circuitry to eliminate sub-V_t leakage from unaccessed read-buffers. Peripheral charge-pumps ensure buffer-foot drivers do not limit I_{READ}.

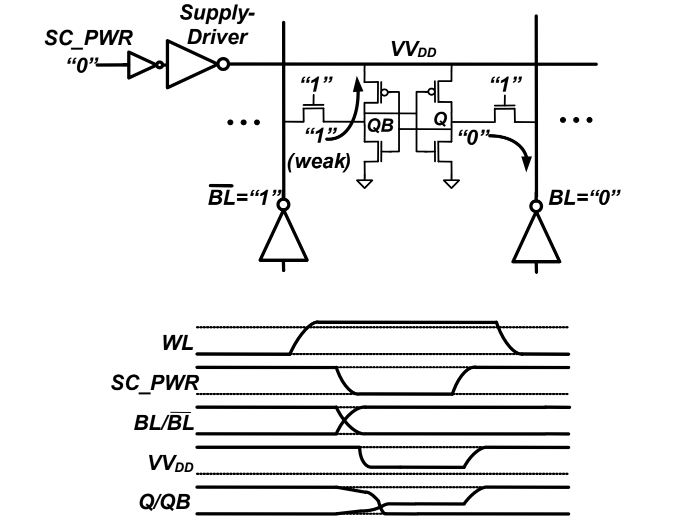

Figure 18.4.4: Cell write performed by weakening local feed-back. Cell supply settles to low intermediate voltage determined by supply driver and write drivers.

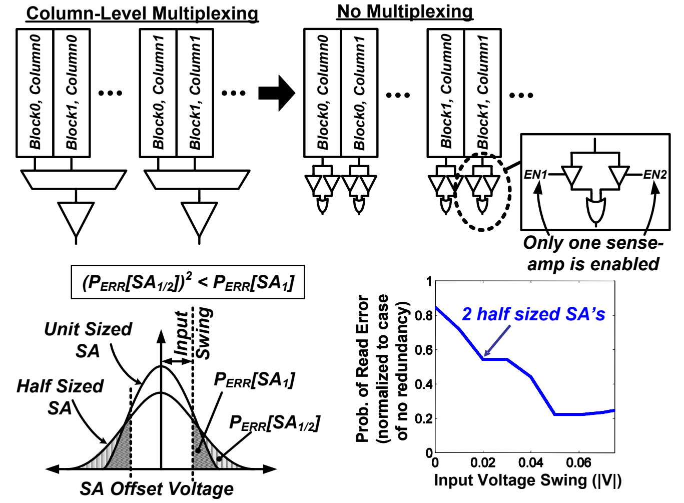

Figure 18.4.5: Without multiplexing, sense-amplifiers have stringent offset and area requirements. With redundancy, errors depend on joint probabilities, improving offset for a given area constraint.

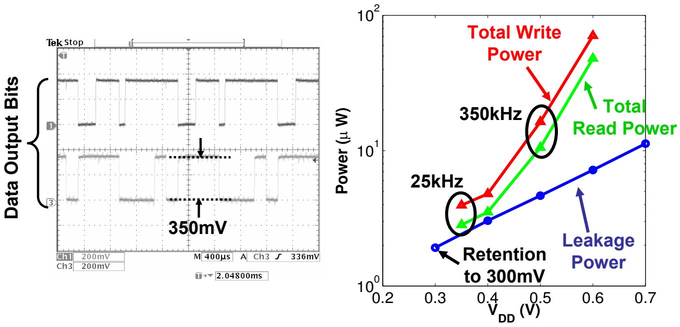

Figure 18.4.6: Scope output and measurements of 65nm test-chip. Array reads and writes at 350mV. Data is correctly retained at 300mV.

Continued on Page 606

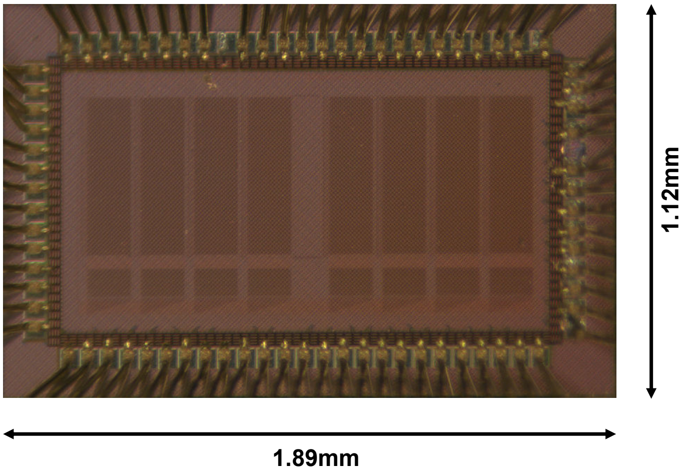

Figure 18.4.7: Die photograph of 256kb 8T SRAM in 65nm CMOS.