A Variation-Tolerant Sub-200 mV 6-T Subthreshold SRAM

Bo Zhai, Scott Hanson, *Student Member, IEEE*, David Blaauw, *Member, IEEE*, and Dennis Sylvester, *Senior Member, IEEE*

***Abstract*—In this paper, we present a deep subthreshold 6-T SRAM, which was fabricated in an industrial 0.13 μm CMOS technology. We first use detailed simulations to explore the challenges of ultra-low-voltage memory design with a specific emphasis on the implications of variability. We then propose a single-ended 6-T SRAM design with a gated-feedback write-assist that remains robust deep in the subthreshold regime. Measurements of a test chip show that the proposed memory architecture functions from 1.2 V down to 193 mV and provides a 36% improvement in energy consumption over the previously proposed multiplexer-based subthreshold SRAM designs while using only half the area. Adjustable footers and headers are introduced, as well as body bias techniques to extend voltage scaling limits.**

***Index Terms*—Low voltage, subthreshold, variation-tolerant SRAM.**

I. Introduction

SUBTHRESHOLD operation holds promise for ultra-low-energy operation in emerging applications such as environmental and biomedical sensing and supply chain management. In addition to sensor-based applications, subthreshold operation is attractive for mid- to high-performance applications where power has become a limiting constraint. Highly parallel near-threshold or subthreshold systems can eliminate the performance penalty associated with low-voltage operation while also leveraging the energy benefits [1]. The design of robust, high-density subthreshold SRAM will play a pivotal role in determining the viability of these systems.

Subthreshold operation has been strongly motivated by previous work. It was shown in [2], [3] that there exists an energy optimal supply voltage for CMOS digital circuits which typically resides in the subthreshold regime. Fig. 1 shows the simulated energy consumption of an inverter chain in a 0.13 μm technology with a threshold voltage ($V_{\rm th}$) of $\sim$0.4 V, where energy is defined as the power delay product. In the superthreshold regime [Fig. 1(a)], the active energy consumption, $E_{\rm act}$, scales down quadratically with supply voltage. In this region, active energy (left axis) dominates leakage energy (right axis). Due to the dominance of $E_{\rm act}$, it is always beneficial to scale down supply voltage in this regime. However, this is no longer the case when $V_{\rm dd}$ drops below $V_{\rm th}$, as shown in Fig. 1(b). The quantity

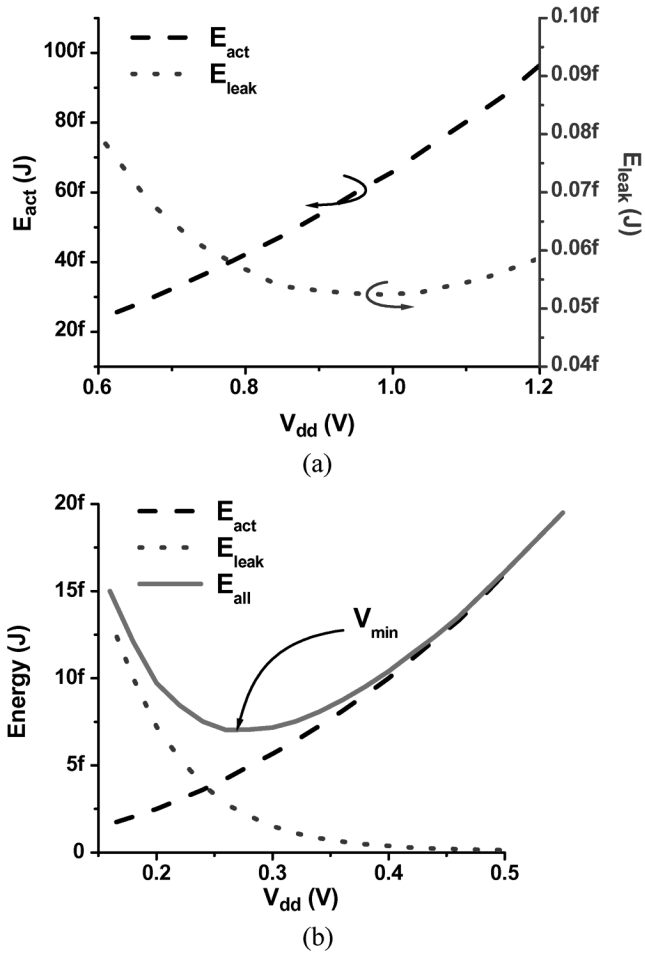


Fig. 1. Illustration of energy optimal voltage for an inverter chain.

$E_{\rm act}$ continues to scale quadratically, but $E_{\rm leak}$ rises quickly due to exponentially increasing delay, creating a minimum energy voltage, $V_{\rm min}$. From an energy perspective, it is not advantageous to operate a digital circuit below $V_{\rm min}$. The location of $V_{\rm min}$ was shown in [2], [3] to be dependent on the balance between $E_{\rm act}$ and $E_{\rm leak}$, which is a strong function of switching activity in the circuit under test. For typical circuit topologies and switching activity rates, the balance between $E_{\rm act}$ and $E_{\rm leak}$ usually occurs in the subthreshold regime, making subthreshold operation optimal [2], [3], [6], [8]. Digital logic has been shown to operate correctly at <200 mV in previous work [4], [6], [8], suggesting that operation at $V_{\rm min}$ is feasible.

However, the design of robust and dense memories becomes challenging at low voltages. Mismatch induced by process variability, in particular, is problematic for low-voltage SRAM

Manuscript received August 16, 2007; revised April 25, 2008. Current version published October 8, 2008.

The authors are with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: bzhai@umich.edu; hansons@umich.edu; blaauw@umich.edu; dmcs@umich.edu).

Digital Object Identifier 10.1109/JSSC.2008.2001903

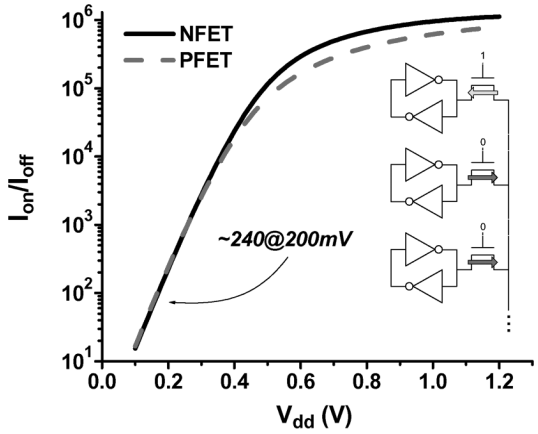


Fig. 2. On/off current ratio as a function of supply voltage. The inset shows contention between read current and bitline leakage during a read operation.

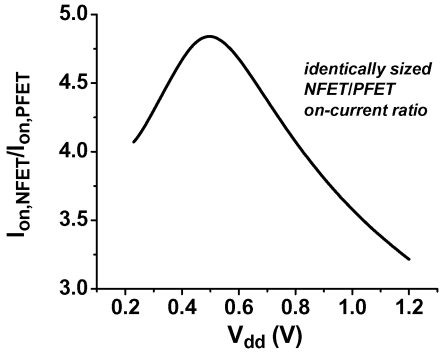


Fig. 3. Identically sized NFET and PFET on-current ratio as a function of supply voltage.

design and creates serious questions about the viability of large-scale subthreshold systems with dense SRAM. In this work we explore the challenges of subthreshold memory design and propose a dense, low energy 6-T cell that effectively combats variability. Our proposed cell, which uses a single-ended cell with gated-feedback write-assist as well as transistor upsizing to fight variability, is capable of operating below 200 mV. We begin in Section II with an in-depth discussion of the challenges facing low-voltage SRAM designers as well as an overview of previously proposed solutions. In Section III, we discuss our proposed SRAM design, which has been implemented in a 0.13 μm technology. We present extensive measurements of the proposed SRAM in Section IV.

II. Low-Voltage SRAM Challenges

The design of robust low-voltage SRAM is extremely challenging and has been the subject of intense study recently [7], [11]–[15]. In this section, we introduce the key challenges in low-voltage memory design. Our proposed solution to these challenges will be described in the next section.

A. On-Current to Off-Current Ratio

The dramatic reduction in the on-current to off-current ratio $(I_{\rm on}/I_{\rm off})$ observed in the near- threshold and subthreshold regions is one of the most fundamental challenges facing low-voltage memory designers. Fig. 2 shows $I_{\rm on}/I_{\rm off}$ for NFET and PFET transistors in a 0.13 μm technology. The ratio $I_{\rm on}/I_{\rm off}$, which goes as low as $\sim$240 for the NFET at 200 mV, determines the theoretical upper bound of the number of cells sharing one bitline. When this ratio is small, it becomes difficult to distinguish between the read current of the accessed cell and the cumulative leakage current of unaccessed cells (inset in Fig. 2). Note that the data in Fig. 2 represent nominal conditions. Under process variation, $I_{\rm on}/I_{\rm off}$ for an NFET at 200 mV can be as low as $\sim$190 at the 99.5% confidence point, forcing the use of very small bitlines.

B. Sizing Constraints

The second challenge facing low-voltage memory designers is a change in gate sizing requirements. Recall that subthreshold current is exponentially dependent on $V_{\rm th}$, so any skew between the nominal threshold voltages of PFET and NFET devices (which is highly technology dependent) can lead to dramatic shifts in the β ratio at low voltages. The read stability and write stability of a conventional 6-T SRAM cell are heavily dependent on the relative strengths of the pull-up, pull-down and pass transistor devices. The skewed β ratios observed at low voltage may therefore lead to an unstable memory cell in addition to unfavorable timing. Fig. 3 shows the simulated β ratio as a function of supply voltage. In this technology, the β ratio rises considerably when $V_{\rm dd}$ drops from 1.2 V to 0.2 V. For robust operation at low voltage, it is important to account for changing device sizing requirements. It has also been noted in recent work that transistors exhibit strong reverse short-channel effects (RSCE) in the subthreshold regime [6], [14] due to reduced drain-induced barrier lowering (DIBL). Larger than minimum channel length access transistors [14] therefore may lead to stronger (rather than weaker, as in typical superthreshold operation) transistors, greatly changing the memory cell ratio. However, as we will see in the next section, sizing transistors to meet nominal β ratio requirements can be ineffective due to the impact of $V_{\rm th}$ variations.

C. Variability

The final, and arguably most important, challenge in low-voltage SRAM design is the heightened sensitivity to process, voltage, and temperature variations. Due to the exponential dependence of drive current on $V_{\rm th}$, $V_{\rm dd}$, and temperature, even small variations lead to large fluctuations in transistor drive current. Mismatch between the cross-coupled inverters is particularly concerning since it can lead to widespread functional failure.

In general, variability can be grouped into two classes: global and random. Global variability is shared among all devices and is therefore not a significant threat to functionality. It is instead a threat to parametric yield in terms of energy and delay. Global variation typically comes in the form of chip-wide variations in $V_{\rm th}$ or global fluctuations in temperature. Random variability does pose a significant threat for SRAM designers, however, since it introduces mismatch between the cross-coupled inverters. There are two dominant sources of random variability in a typical technology: gate length $(L_{\rm eff})$ and $V_{\rm th}$. Gate length variations arise from irregularities during the lithography process [16] and induce $V_{\rm th}$ variation in superthreshold devices

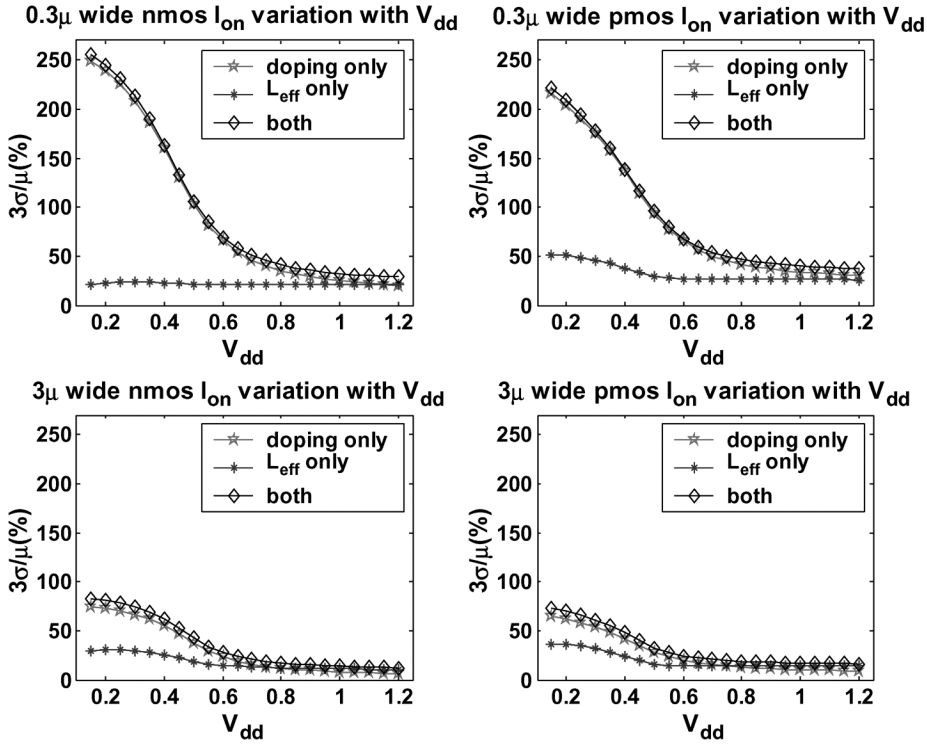


Fig. 4. $3\sigma/\mu$ of $I_{\rm on}$ due to different variation sources over a wide range of $V_{\rm dd}$, showing the dominance of RDF in subthreshold operation.

due to short channel effects. Variations in $V_{\rm th}$ are also caused by random variations in both number of dopants and the positions of those dopants in the channel. These variations are typically called random dopant fluctuations (RDF) and are known to exhibit an inverse dependence on the square root of gate area [10]. Fig. 4 shows simulated on-current variation due to random doping and $L_{\rm eff}$ variation for two different NFET and PFET sizes in a 0.13 μm technology. At high voltage, the relative importance of gate length and doping variations are comparable. As the voltage reduces, the relative importance of doping variations grows. Note that channel length variation induces $V_{\rm th}$ variations due to DIBL. Since DIBL becomes less pronounced at low voltages, the magnitude of $V_{\rm th}$ variation arising from channel length uncertainty rapidly falls off as $V_{\rm dd}$ reduces. However, since $I_{\rm on}$ at low voltages becomes more sensitive to $V_{\rm th}$ fluctuations (exponentially dependent in the subthreshold region), the net result is that $I_{\rm on}$ variation due to DIBL remains roughly constant. On the other hand, the uncertainty in $V_{\rm th}$ due to RDF is independent of $V_{\rm dd}$ and solely a function of channel area [10]. Therefore, $I_{\rm on}$ variation resulting from RDF becomes the dominating component as $V_{\rm dd}$ nears $V_{\rm th}$ as shown in Fig. 4. Note that the total current variation reduces dramatically when the gate width is increased from 0.3 μm to 3 μm.

The device-level implications of RDF-induced $V_{\rm th}$ variations are clear, but it is important to relate these problems to changes at the circuit level. Mismatch between the relative strengths of feed-forward and feed-back inverters can lead to dramatic reductions in noise margins. To investigate the implications of mismatch, we run Monte Carlo simulations with 1000 trials on a minimum size SRAM cell. Both global and random variations in $L_{\rm gate}$ and $V_{\rm th}$ are modeled using normal distributions. Fig. 5 shows the variation in low hold noise margins for the simulated SRAM cell at 0.3 V and 1.2 V. The mean noise margin reduces

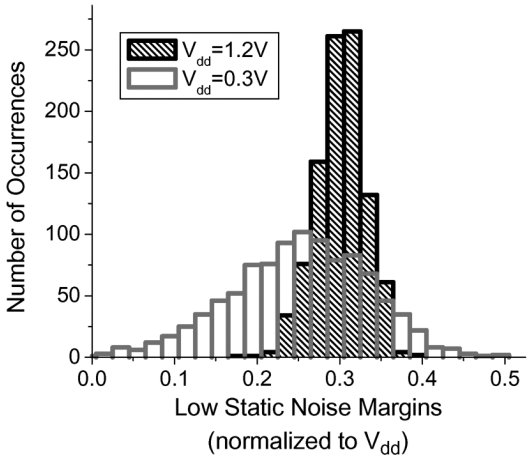


Fig. 5. Variation in hold noise margins in superthreshold (1.2 V) and subthreshold (0.3 V) regions.

from 30% of $V_{\rm dd}$ at 1.2 V to 25% of $V_{\rm dd}$ at 0.3 V. This reduction in noise margins is not as alarming as the $\sim$3.4X increase in variability as measured by $\sigma_{\rm SNM}/\mu_{\rm SNM}$.

Maintaining a balance between read and write requirements is also very difficult at low $V_{\rm dd}$ under random $V_{\rm th}$ variability [17]. In a typical 6-T SRAM cell, this balance is typically achieved by sizing the pull-down, pull-up and pass transistors to achieve desired relative strength ratios while also meeting density requirements. Given the exponential sensitivity of subthreshold current to $V_{\rm th}$, it is impractical to rely on sizing ratios with linear

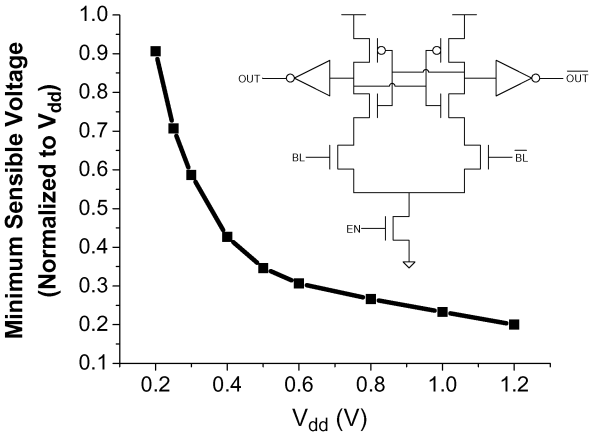


Fig. 6. Minimum sensible voltage for a sense amplifier.

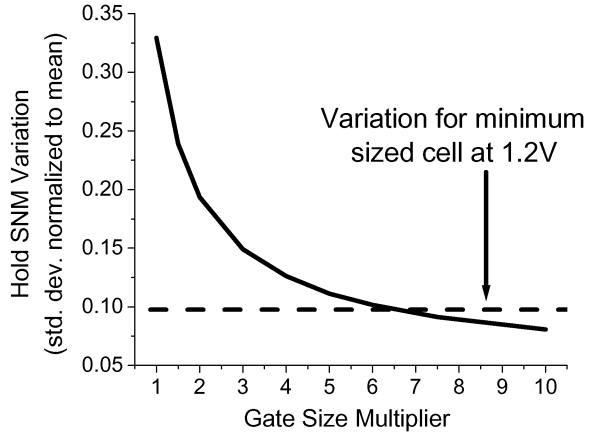


Fig. 7. 3σ hold noise margins for different cell areas at 0.3 V. All transistor sizes in the cell are multiplied by an identical scaling factor.

sensitivities to achieve balanced read and write characteristics. Rather, alternative cell topologies must be used to decouple read and write requirements. Several such topologies, along with our own proposal, will be discussed in subsequent sections.

Until now, the discussion in this paper has focused entirely on bitcell design. Traditional differential 6-T SRAM also requires a sense amplifier (SA) for readout. SA design is complicated at low voltages, especially when considering variability. Due to the differential nature of SA operation, the SA is particularly susceptible to mismatch introduced by random RDF-induced $V_{\rm th}$ variations. To examine this further, we run Monte Carlo SPICE simulations with 1000 trials on the sense amplifier shown on the inset in Fig. 6. The minimum DC voltage required at the bitline (BL) inputs to ensure proper output switching across 99% of the trials is plotted as a function of $V_{\rm dd}$ in Fig. 6. The minimum sensible voltage increases from 20% of $V_{\rm dd}$ at 1.2 V to 59% of $V_{\rm dd}$ at 0.3 V. This increase in relative sensible voltage translates to a dramatic increase in read delay with reduced $V_{\rm dd}$ (relative to nominal inverter delay).

Given the gate area dependence of RDF, the simplest solution to the variability problem is to use larger gate sizes. We again conduct Monte Carlo simulations using the methodology originally highlighted in Fig. 5. Fig. 7 shows the simulated variation in hold SNM (as measured by σ/μ) for different 6-T SRAM cell sizes at 0.3 V. For every point along the horizontal axis, the size of each transistor in the SRAM cell is multiplied by the same factor. At 0.3 V, the transistor sizes must be increased by $\sim$6.5X in order to achieve noise margin variation equal to that observed at 1.2 V (relative to $V_{\rm dd}$). Larger gate sizes may be similarly used to improve SA reliability. The use of increased gate sizes leads to obvious array density problems, so it is important for subthreshold SRAM designers to be aware of the trade-offs between robustness and cell area. Furthermore, designers must simulate extensively at the intended operating voltage since these trade-offs will change with $V_{\rm dd}$.

In addition to simulation-driven transistor sizing, variation-tolerant subthreshold SRAM design will require circuit innovations. We explore several previously proposed circuit solutions in Section II-D before presenting our proposed solution in Section III.

D. Previous Solutions

To this point we have focused our discussion on the traditional 6-T SRAM cell. A number of alternative cell architectures have been proposed recently to help cope with the problems discussed in this section. One of the first attempts at subthreshold memory design used latch-based memory cells with multiplexer (mux)-based decoders [4]. This design remains functional below 200 mV but has unacceptable density and performance characteristics for commercial applications. Furthermore, the prohibitive area implies substantial switching capacitance and leakage current, thus minimizing the energy savings. An 8-T cell with a 2-T read buffer was shown to be functional in the near- subthreshold regime in [11], [12]. The extra transistors isolate memory cells from the read bitline to improve read stability and decouple the read and write requirements. The authors of [7] proposed a 10-T cell with a 4-transistor read buffer that remained functional with the supply voltage as low as 380 mV. More recent work has also expanded upon both of these designs with interesting techniques for improving robustness [13], [14]. However, all of the past work has relied on the addition of transistors to the traditional 6-T cell. In the next two sections we discuss the design and test of a subthreshold SRAM cell that uses only 6 transistors.

III. Proposed Design

In order to address the challenges covered in the last section, we propose a new 6-T memory cell design [15] targeting low-voltage operation. The schematic and layout of the memory cell are shown in Fig. 8. Instead of using the traditional differential structure, we employ a single-ended cell with a full transmission gate at one side. The penalty of having one additional wordline (*WL_*) is offset by the elimination of the second bitline. One clear advantage of this design is that the bitline can be driven from rail to rail, eliminating the need for a sense amplifier (which can lead to density and variability problems in differential designs). Furthermore, noise during a read operation is isolated to the single bitline, making single-ended design inherently more robust to read upsets than conventional differential design. To recover lost write margins, we gate the supply voltage on the feedback inverter during the write operation. The use of the single-ended cell in combination with the gated-feedback

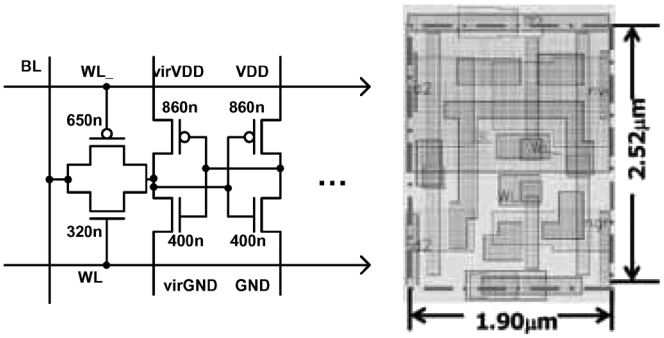


Fig. 8. Proposed SRAM cell design.

write-assist effectively decouples the read and write operations, enabling balanced read and write margins in the face of high variability at low voltage.

The use of increased gate sizes to fight RDF was also a critical piece of our design strategy. Using Monte Carlo SPICE simulations, transistor widths were set to the values shown in Fig. 8 to meet robustness requirements with all transistor lengths set to 0.12 μm. The minimum device width was set to 0.32 μm to limit the amount of RDF-induced $V_{\rm th}$ mismatch, which was shown in Section II-C to be a strong function of gate area. The PFET device is sized relative to the NFET to accommodate the changing β ratio observed in the subthreshold regime (Fig. 3). Unlike a traditional 6-T cell in which the PFET acts as a resistive load (and thus can be small), a single-ended design relies on the PFET to pull up the bitline. We find that the use of identically-sized feedback and forward inverters effectively balances read and write capabilities. The area of the bitcell is 4.788 μm^2 as shown in Fig. 8 and is $\sim$2X larger than that of a typical traditional 6-T cell as given by the ITRS (2.366 μm^2 in 130 nm [9]). It is important to note that cell device sizes can be reduced significantly if a less stringent supply voltage floor is required and if memory design rules are available (logic rules were used).

Fig. 9 shows the architecture of the proposed 6-T-based SRAM. There are 16 bitcells connected to one bitline. Monte Carlo simulations indicate that, at 16 bitcells per bitline, the read current of a single bitcell is always greater than the cumulative bitline leakage due to the unaccessed devices on the same bitline. Additionally, for performance reasons it is important to have a short bitline (with small capacitance) in a single-ended design since the sensing element is a simple inverter that requires a bitline swing of nearly $V_{\rm dd}$. In contrast to conventional sense-amp based designs, the area penalty of a short bitline is minimal since each added bitline requires only two CMOS inverters in the readout path.

The readout path, shown in Fig. 10, consists of a 16-to-1 column mux and a pulsed latch. A near-minimum sized inverter is used as the sensing element to reduce bitline capacitance and minimize the likelihood of a read upset. The second inverter in the read-out path is larger for robustness reasons and, based on simulation, is able to drive a tri-state line with up to 64 units. In the implemented design, the second level mux is restricted to 16 inputs, since 2 kb is sufficient for the targeted sensor applications. Signal *latch_en* is pulsed at the end of one clock cycle to latch the output.

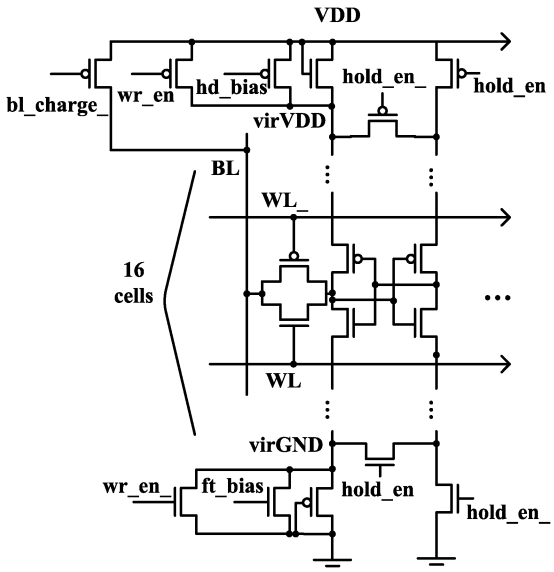


Fig. 9. Cell array structure.

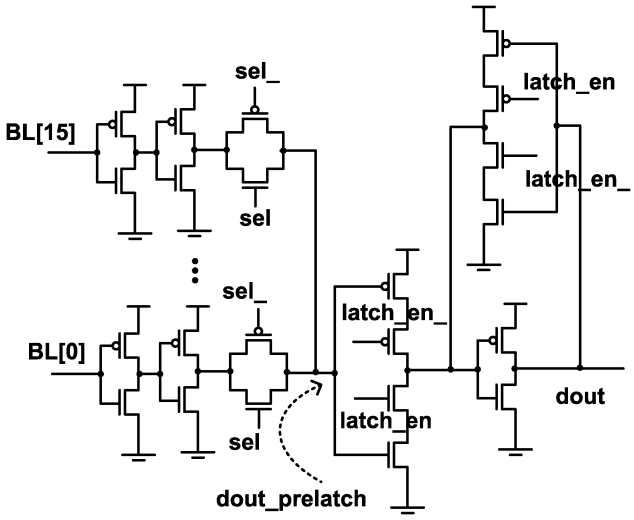


Fig. 10. Read-out with 16-to-1 mux sized to ensure reliability.

To recover the write stability sacrificed in the single-ended design, adjustable strength header and footer devices are used. A write mode abstraction of the cell array is shown in Fig. 11. The goal is to break the feedback loop between the two cross-coupled inverters by weakening the feedback inverter. When a write occurs, the *wr_en* signal is asserted and the strong PFET transistor at the top and the NFET transistor at the bottom are both turned off with only the weak headers and footers enabled. This results in a temporary supply voltage droop, which allows the stored state to be easily overwritten. In this design, we adopt a NFET/PFET as the weak device at header/footer since we find that even a minimum sized PFET/NFET header/footer is not sufficiently resistive. To minimize area, only one header/footer supply throttling circuit is used per bitline with all 16 cells sharing the same virtual $V_{\rm dd}$ and virtual ground (Fig. 11). Despite the supply droop, the state of the non-accessed cells is retained.

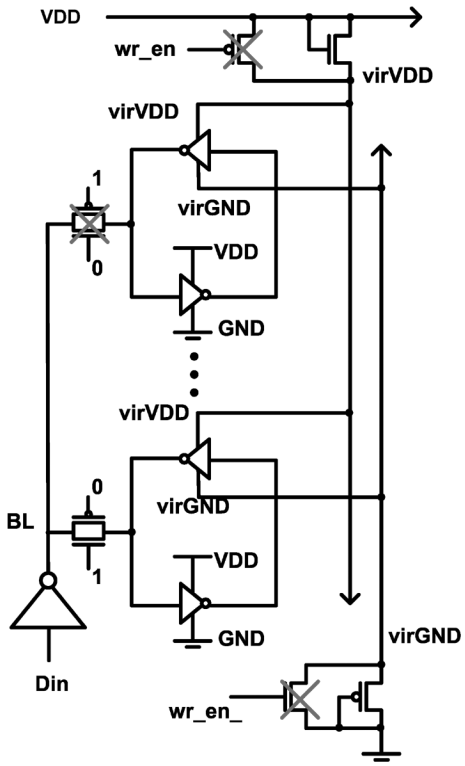


Fig. 11. Write mode abstraction.

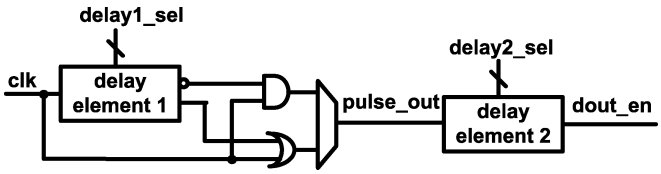


Fig. 12. Pulse generator with programmable width and distribution delay to improve robustness to variability.

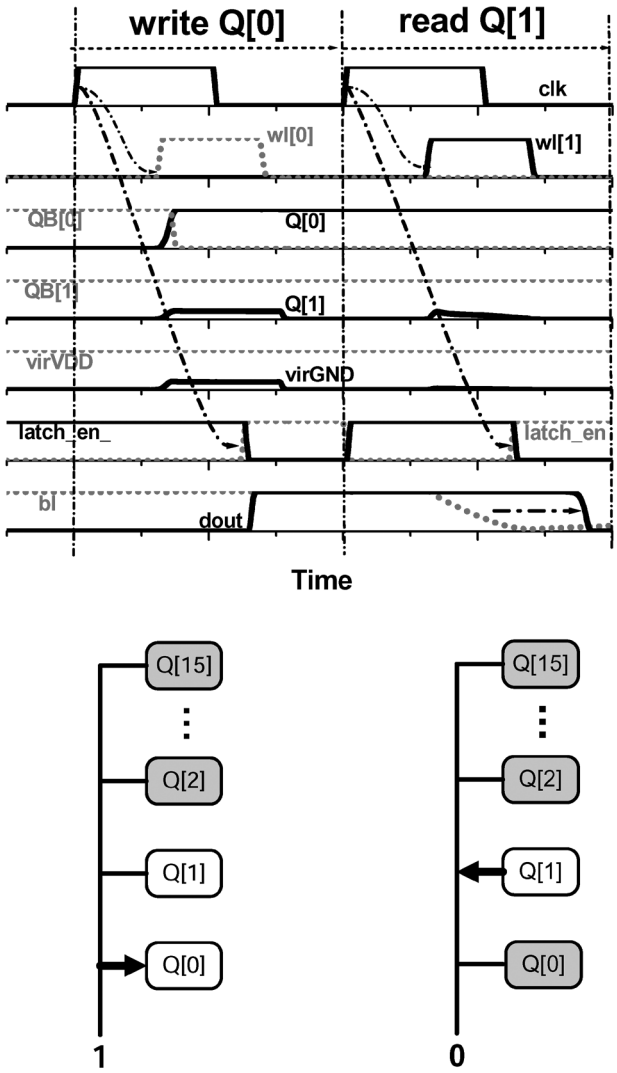


Fig. 13. SPICE simulation of proposed 6-T SRAM showing consecutive read and write operation and supply voltage suppression through tunable headers and footers.

The timing generation is shown in Fig. 12 and is programmable to allow for improved timing robustness and performance. To address the increased variability in the subthreshold regime, we implemented both NAND type and NOR type pulse generation to achieve a tunable pulse window that extends beyond the half cycle point. Fine pulse width and delay correction between the critical signals such as *wl* and *latch_en* can be tuned, but in practice, measurements indicate that one setting is sufficient for all the dies to function properly.

Fig. 13 illustrates the simulated waveforms for two consecutive read and write operations using SPICE simulation. Memory cells 2 through 15 (denoted *Q[2]–Q[15]*) are initialized with a logic value of 1 while *Q[0]* and *Q[1]* are initialized to 0. During the first clock cycle, a write operation to *Q[0]* is performed. The wordline pulses are derived from the rising edge of the input *clk*, and *Q[0]* data is overwritten shortly after wordline 0 is asserted. As expected, we see a *virGND* disturbance during this cycle while *virVDD* is unaffected because the bitline driver fights the footer transistors only when writing a "1". The same voltage droop is also seen by the unaccessed cells as shown by the *Q[1]* and *QB[1]* waveforms. On the following cycle, a read from *Q[1]* represents the worst case condition for data dependency since all the remaining cells are holding opposite value and leaking towards *Q[1]*. Soon after, the wordline for *Q[1]* (*wl[1]*) is turned on and the bitline (*bl*) is pulled down and latched at the end of the clock cycle. Monte Carlo simulations of this worst-case write/read cycle were used to drive design-time sizing optimization and yield estimation.

Since the bitline is directly connected to the read-out inverter, static current can be high when the bitline is floating. Therefore, *bl_charge*, as well as *hold_en*, is asserted during standby mode. By enabling *hold_en*, *virGND* and *virVDD* are shared by the forward driving inverters, creating a stronger stack effect and reducing leakage by current starving.

Fig. 14 shows the top level view of the SRAM micro-architecture. The 2 kb SRAM is organized into 256 words with 8 bits per word. An 8-bit address is divided among a 4-bit wordline decoder and 4-bit column decoder.

IV. Measurements

A test chip with a 2 kb SRAM has been designed and fabricated in a commercial 0.13 μm CMOS bulk technology with

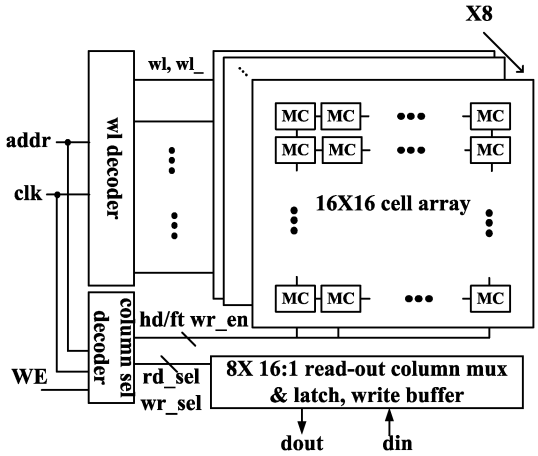


Fig. 14. Microarchitecture of the 2 kb SRAM.

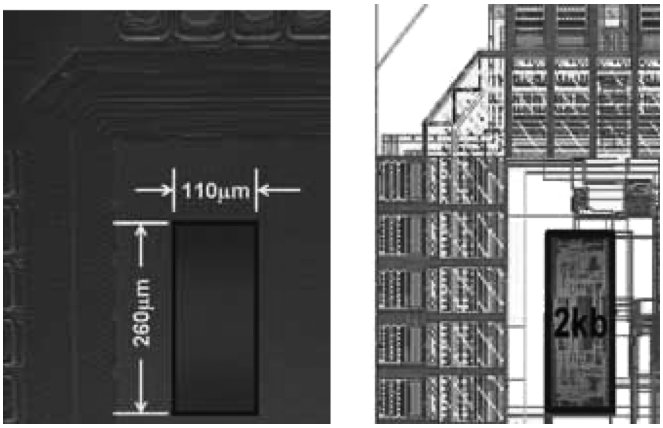


Fig. 15. Die photograph and layout.

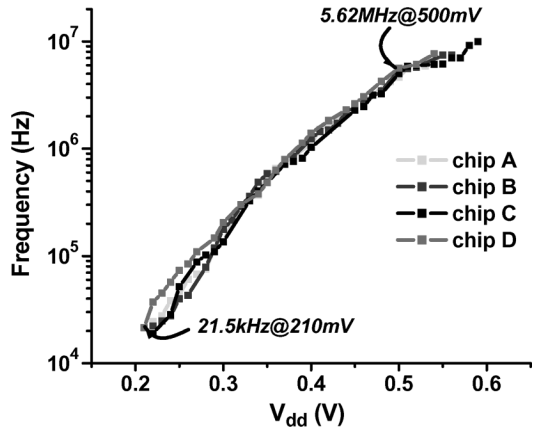


Fig. 16. Frequency with $V_{\rm dd}$ scaling for four dies.

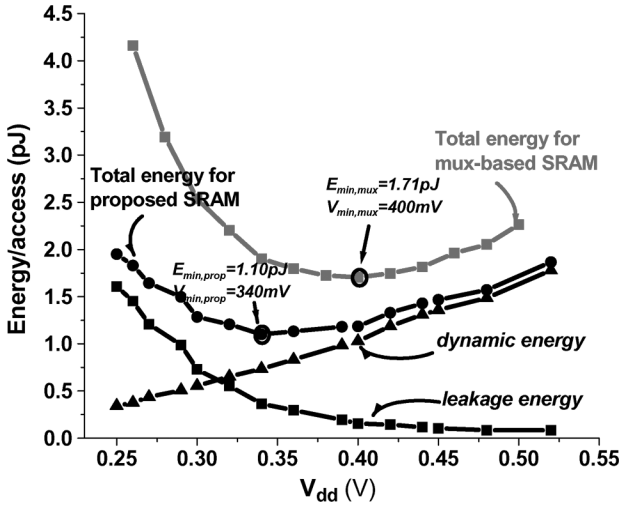


Fig. 17. Energy consumption comparison to mux-based design in the same technology.

$V_{\rm th}$ ~0.4 V to demonstrate our proposed SRAM architecture. The die photo and layout diagram are shown in Fig. 15. With no die-specific adjustments, the SRAM is fully functional at a supply voltage of 208 mV, while enabling on-chip tuning allows operation down to 193 mV. To our knowledge, this is the first 6-T SRAM capable of operating substantially below threshold voltage. A total of 24 dies have been measured and all are functional.

A. *Performance and Energy Analysis*

Fig. 16 shows frequency measurements for 4 typical dies, showing the expected exponential dependence of frequency with $V_{\rm dd}$ in the subthreshold regime. The array achieves a frequency of 5.6 MHz at 0.5 V, a reasonable speed for microcontrollers and 21.5 kHz at 210 mV, which matches a previously reported subthreshold processor [8].

The measured energy per access for the proposed SRAM is compared with that of a mux-based memory [4] fabricated in the same technology. The power is measured with identical random input traffic with an activity rate of 0.5 accesses/cycle in both cases. Fig. 17 shows the results for both SRAM. For equal supply voltages, the proposed SRAM consumes 31% less energy with 20% better performance. The energy optimal supply voltage $V_{\rm min}$ lies at 340 mV for the proposed SRAM and at 400 mV for the mux-based memory because there is more relative leakage in the mux-SRAM. Fig. 17 also shows the energy breakdown between active and leakage for the proposed design. Due to the exponential increase of circuit delay and the dominance of leakage current in this voltage regime, the energy per access increases at supply voltages below $V_{\rm min}$ as expected. At the respective $V_{\rm min}$ voltages, the proposed SRAM consumes 36% less energy than the mux-based memory. Furthermore, unlike the mux-based memory, $V_{\rm min}$ for the proposed SRAM matches more closely to that of a typical sensor processor core [8], which could allow both memory and core to operate at peak energy efficiency with a single supply voltage. In Fig. 18 the energy and frequency distributions of 14 tested dies are presented for three operating voltages. As expected, the variation reduces as voltage increases.

The area of the proposed SRAM is 28600 μm^2, nearly half that of the mux-based memory (54000 μm^2). Table I provides a comparison of the circuit properties. A 2 kb SRAM using a traditional 6-T cell from a commercial library is 30% smaller than our proposed SRAM.

TABLE I
COMPARISON OF THIS WORK WITH MUX-BASED DESIGN

	This work	Mus-based
Area	28,600μm^2	54,000μm^2
Speed	205kHz@300mV	162kHz@300mV
Minimum energy per access	1.10pJ	1.71pJ
V_{min}	340mV	400mV

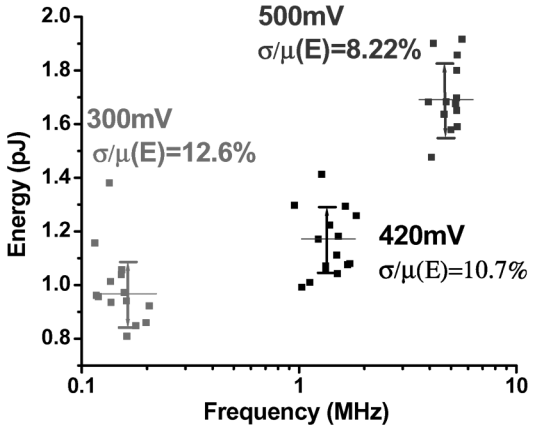


Fig. 18. Frequency–energy scatter plot at different supply voltages.

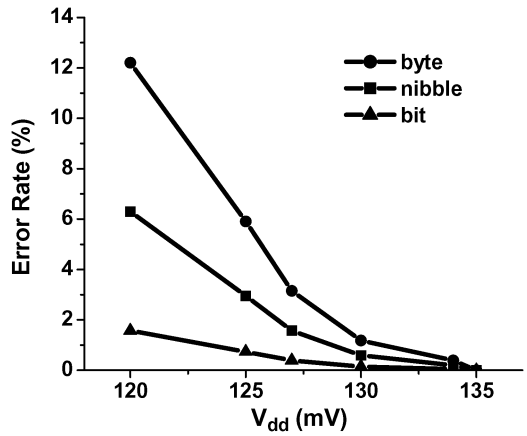
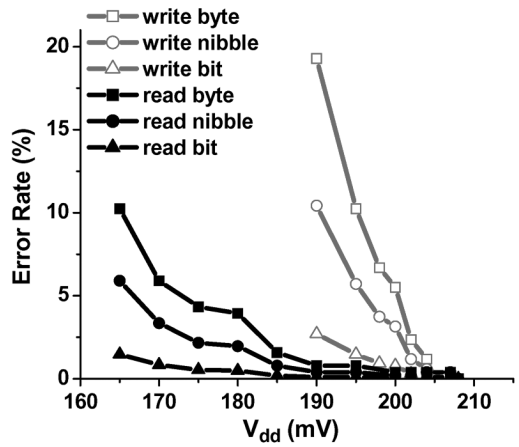
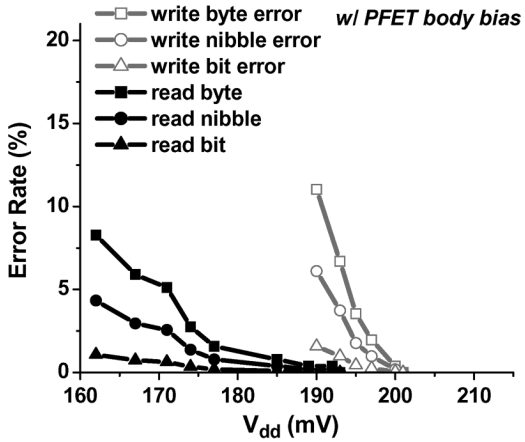


Fig. 19. Hold failure measurements in bit, nibble and bytes.



(a)



(b)

Fig. 20. (a) Read and write error rates versus supply voltage with zero body bias. (b) Read and write error rates with a small (<15 mV) PFET body bias.

B. Failure Rate Analysis

Fig. 19 shows the hold failure rate measurements for a typical die. We include nibble-level (4-bit) and byte-level (8-bit) error rates in addition to bit-level error rates to highlight the system-level impact of hold failures. Note that a nibble or byte is considered erroneous if it contains even a single incorrect bit. Consequently, error rates are much higher at the nibble and byte levels than at the bit level.

The first retention failure is observed at 134 mV for one typical die. The standby power at this voltage is ∼26 nW for the entire array. The bit error rate remains below 2% at 120 mV. Since minimum retention voltage is a strong function of transistor sizing, these results show the efficiency of the employed cell sizing strategy.

Fig. 20(a) shows measured read and write failure rates across supply voltage. The first read and write errors occur at 208 mV and 205 mV, respectively. These measurements suggest that read and write failures are well-balanced, which was a goal of the SRAM sizing strategy. If we assume a 2% bit redundancy rate, the effective operational voltage could be extended to 195 mV. At 195 mV, further voltage scaling is limited by the rise of write errors. Despite well-matched first read and write error voltages, the write error rate curves are steeper than the read curves. We will look at solving this problem in the next section.

The error rates can by further improved by applying a small forward body bias (<15 mV) on the PFETs, as shown in Fig. 20(b). The application of a body bias to the PFET devices compensates for chip-wide mismatch between PFET and NFET devices and improves noise margins. It has been shown in previous work that body biasing is particularly effective in the subthreshold region due to the exponential sensitivity of current to $V_{\rm th}$ [6]. The first read and write errors occur at 193 mV and 200 mV, respectively, with write errors becoming significant

TABLE II
ERROR RATE IMPROVEMENT WITH DIFFERENT TECHNIQUES

	first read error	first write error	Functional voltage (assuming 2% bit redundancy)
no body bias	208mV	205mV	195mV
with PFET body bias	193mV	200mV	190mV
with body bias and header/footer bias	193mV	192mV	<170mV (>22% power savings)

TABLE III
OVERALL SPECIFICATION OF THE SRAM

Technology	0.13µm 8-metal CMOS
SRAM Size	2k bits
Area	28,600µm^2
Functional Supply Voltage	1.2V-193mV
Frequency	5.62MHz@500mV 21.5kHz@210mV
Energy/access	780fJ@300mV
Power	50nW@210mV

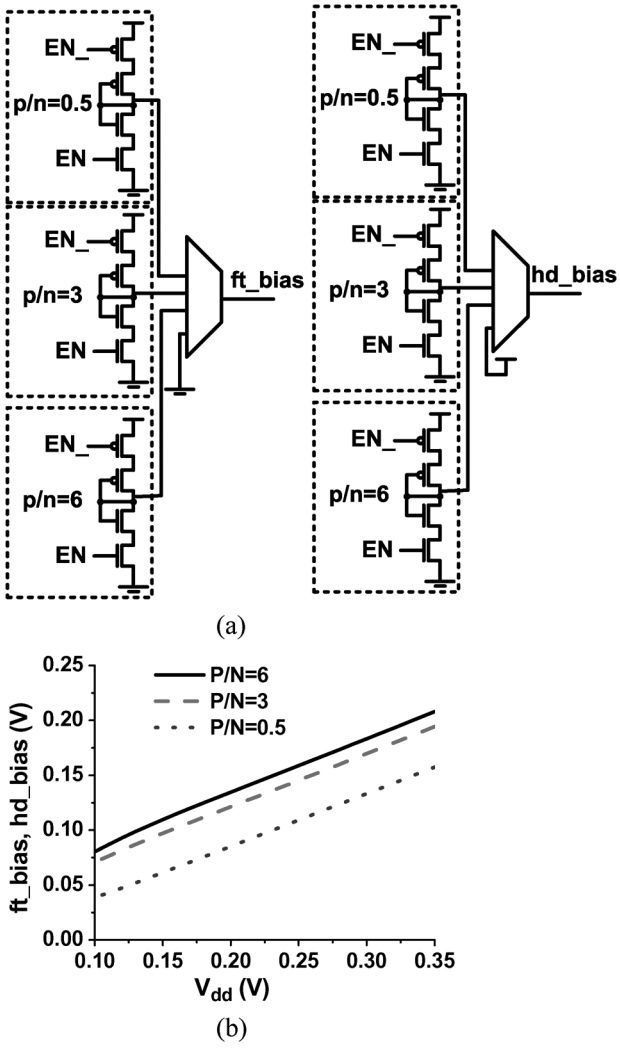


Fig. 21. (a) On-chip bias generator. (b) Simulated footer and header bias voltages (*ft_bias*, *hd_bias*) for different β ratios.

before any read errors are observed. With 2% bit redundancy, the new effective operational voltage is 190 mV.

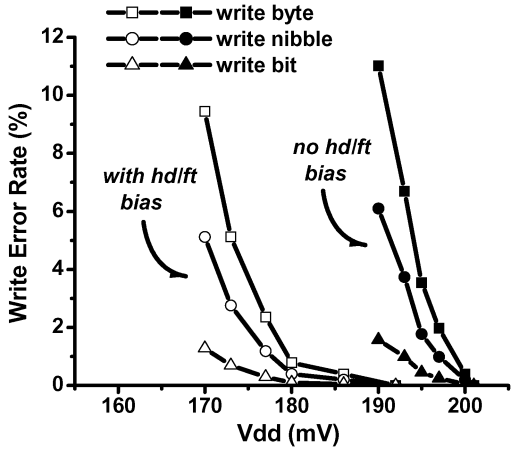


Fig. 22. Write error rates before and after applying header/footer bias (PFET body bias is applied in both cases).

C. Extending the Operational Voltage

Below 200 mV, SRAM operation is limited by write stability due to weak headers/footers. The weak NFET header device causes excessive voltage fluctuations during a write operation leading to destruction of data in unaccessed memory cells on the same bitline. We can solve this problem by employing a partially turned-on PFET/NFET in parallel with weak NFET/PFET as the header/footer. The controlling voltage (*hd_bias*/*ft_bias*) allows each die to be individually tuned. Using an on-chip bias generator, *ft_bias* and *hd_bias* can be programmed to one of 6 values between 0 and $V_{\rm dd}$. The schematic of the on-chip bias generator is shown in Fig. 21(a). Since it is only used at very low supply voltages, the static current of the bias generator is minimal compared to the leakage of the entire array. Fig. 21(b) shows the available bias voltages for different supply voltages based on SPICE simulation.

The effectiveness of the tunable header/footer circuits are shown in Fig. 22. The first observed write error is seen at 192 mV after applying header/footer bias, compared to 200 mV without the application of a bias. While enabling the tuned header/footer does not significantly reduce the point of first failure, it does improve the failure rate increase below this

point, making redundancy more advantageous. With 2% redundancy the effective minimum operating voltage is extended from 190 to 170 mV, which translates to ∼22% power savings. Note that although we are mainly focused on energy efficiency, certain applications require low power consumption. An example would be a system harvesting power from on-chip solar cells. Table II summarizes the error rate analysis for each of the studied techniques. The key specifications and measurements from this section have been summarized in Table III.

V. Conclusion

In this work we proposed the first deep subthreshold 6-T SRAM design. A 2 kb test SRAM has been designed and fabricated in a commercial 0.13 μm CMOS bulk technology. The measurements show significant improvement over a previous mux-based design. The proposed design is approximately half the size of the mux design and consumes 36% less energy per access. It is fully functional from 1.2 V to 193 mV, demonstrating that careful memory design will make subthreshold operation a viable design option.

References

[1] B. Zhai, R. Dreslinski, D. Blaauw, T. Mudge, and D. Sylvester, "Energy efficient near-threshold chip multiprocessing," in *Proc. IEEE Int. Symp. Low Power Electronics and Design (ISLPED)*, 2007.

[2] B. Zhai, D. Blaauw, D. Sylvester, and K. Flautner, "Theoretical and practical limits of dynamic voltage scaling," in *Proc. IEEE/ACM Design Automation Conf.*, 2004, pp. 868–873.

[3] B. Calhoun and A. Chandrakasan, "Characterizing and modeling minimum energy operation for subthreshold circuits," in *Proc. IEEE Int. Symp. Low Power Electronics and Design (ISLPED)*, 2004, pp. 90–95.

[4] A. Wang and A. Chandrakasan, "A 180 mV FFT processor using subthreshold circuit techniques," in *IEEE Int. Solid-State Circuits Conf. Dig.*, 2004, pp. 292–293.

[5] L. Nazhandali, B. Zhai, R. Helfand, M. Minuth, J. Olson, S. Pant, A. Reeves, T. Austin, and D. Blaauw, "Energy optimization of subthreshold-voltage sensor processors," in *Proc. ACM Int. Symp. Computer Architecture*, Jun. 2005, pp. 197–207.

[6] S. Hanson, B. Zhai, M. Seok, B. Cline, K. Zhou, M. Singhal, M. Minuth, J. Olson, L. Nazhandali, T. Austin, D. Sylvester, and D. Blaauw, "Performance and variability optimization strategies in a sub-200 mV, 3.5 pJ/inst, 11 nW subthreshold processor," in *Symp. VLSI Circuits Dig.*, 2007, pp. 152–153.

[7] B. Calhoun and A. Chandrakasan, "A 256 kb sub-threshold SRAM in 65 nm CMOS," in *IEEE Int. Solid-State Circuits Conf. Dig.*, Feb. 2006, pp. 628–629.

[8] B. Zhai, L. Nazhandali, J. Olson, A. Reeves, M. Minuth, R. Helfand, S. Pant, D. Blaauw, and T. Austin, "A 2.60 pJ/Inst subthreshold sensor processor for optimal energy efficiency," in *Symp. VLSI Circuits Dig.*, Jun. 2006, pp. 154–155.

[9] International Technology Roadmap for Semiconductors, ITRS. [Online]. Available: http://www.itrs.net

[10] M. J. M. Pelgrom, A. C. J. Duinmaijer, and A. P. G. Welbers, "Matching properties of MOS transistors," *IEEE J. Solid-State Circuits*, vol. 24, pp. 1433–1440, 1989.

[11] L. Chang, D. Fried, J. Hergenrother, J. Sleight, R. Dennard, R. Montoye, L. Sekaric, S. McNab, A. Topol, C. Adams, K. Guarini, and W. Haensch, "Stable SRAM cell design for the 32 nm node and beyond," in *IEEE Symp. VLSI Technology*, 2005, pp. 128–129.

[12] L. Chang, Y. Nakamura, R. Montoye, J. Sawada, A. Martin, K. Kinishita, F. Gebara, K. Agarwal, D. Acharyya, W. Haensch, K. Hosokawa, and D. Jamsek, "A 5.3 GHz 8T-SRAM with operation down to 0.41 V in 65 nm CMOS," in *Symp. VLSI Circuits Dig.*, 2007, pp. 252–253.

[13] N. Verma and A. Chandrakasan, "A 65 nm 8T sub-Vt SRAM employing sense-amplifier redundancy," in *IEEE Int. Solid-State Circuits Conf. Dig.*, 2007, pp. 328–329.

[14] T. Kim, J. Liu, J. Keane, and C. Kim, "A high-density subthreshold SRAM with data-independent bitline leakage and virtual ground replica scheme," in *IEEE Int. Solid-State Circuits Conf. Dig.*, 2007, pp. 330–331.

[15] B. Zhai, D. Blaauw, D. Sylvester, and S. Hanson, "A sub-200 mV 6T SRAM in 0.13 μm CMOS," in *IEEE Int. Solid-State Circuits Conf. Dig.*, 2007, pp. 332–333.

[16] K. Bernstein, D. Frank, A. Gattiker, W. Haensch, B. Ji, S. Nassif, E. Nowak, D. Pearson, and N. Rohrer, "High-performance CMOS variability in the 65-nm regime and beyond," *IBM J. Res. Devel.*, vol. 50, pp. 433–449, 2006.

[17] B. Calhoun and A. Chandrakasan, "Static noise margin variation for sub-threshold SRAM in 65 n-nm CMOS," *IEEE J. Solid-State Circuits*, vol. 41, pp. 1673–1679, 2006.

Bo Zhai received the B.S. degree in microelectronics from Peking University, Beijing, China, in 2002, and the M.S. and Ph.D. degrees in electrical engineering from the University of Michigan, Ann Arbor, in 2004 and 2007, respectively.

His research focuses on low power VLSI design. He is a currently a Senior Design Engineer at Advanced Micro Devices, Austin, TX.

Scott Hanson (S'05) received the Bachelor's and Master's degrees in electrical engineering from the University of Michigan, Ann Arbor, in 2004 and 2006, respectively. He is currently pursuing the Ph.D. degree in electrical engineering at the University of Michigan.

His research interests include low-voltage circuit design for ultra-low-energy applications, variation tolerant circuit design, and energy efficient high performance circuit design. He is the recipient of an SRC fellowship.

David Blaauw (M'94) received the B.S. degree in physics and computer science from Duke University, Raleigh, NC, in 1986, and the Ph.D. degree in computer science from the University of Illinois, Urbana, in 1991.

Until August 2001, he worked for Motorola, Inc. in Austin, TX, where he was the manager of the High Performance Design Technology group. Since August 2001, he has been on the faculty at the University of Michigan as an Associate Professor. His work has focused on VLSI design and CAD with particular emphasis on circuit design and optimization for high performance and low power applications.

Dr. Blaauw was the Technical Program Chair and General Chair for the International Symposium on Low Power Electronics and Design and was the Technical Program Co-Chair and member of the Executive Committee the ACM/IEEE Design Automation Conference. He is currently a member of the ISSCC Technical Program Committee.

Dennis Sylvester (S'96–M'97–SM'04) received the B.S. in electrical engineering (*summa cum laude*) from the University of Michigan, Ann Arbor, and the Ph.D. degree in electrical engineering from University of California, Berkeley, in 1999. His dissertation research was recognized with the David J. Sakrison Memorial Prize as the most outstanding research in the UC-Berkeley EECS department.

He is now an Associate Professor of Electrical Engineering and Computer Science at the University of Michigan, Ann Arbor. He previously held research staff positions in the Advanced Technology Group of Synopsys, Mountain View,

CA, Hewlett-Packard Laboratories, Palo Alto, CA, and a visiting professorship in Electrical and Computer Engineering at the National University of Singapore. He has published numerous articles along with one book and several book chapters in his field of research, which includes low-power circuit design and design automation techniques, design-for-manufacturability, and interconnect modeling. He also serves as a consultant and technical advisory board member for several electronic design automation and semiconductor firms in these areas.

Dr. Sylvester received an NSF CAREER award, the Beatrice Winner Award at ISSCC, an IBM Faculty Award, an SRC Inventor Recognition Award, and several best paper awards and nominations. He is the recipient of the ACM SIGDA Outstanding New Faculty Award and the University of Michigan Henry Russel Award for distinguished scholarship. He has served on the technical program committee of numerous design automation and circuit design conferences, the steering committee of the ACM/IEEE International Symposium on Physical Design, and was general chair for the 2005 ACM/IEEE Workshop on Timing Issues in the Synthesis and Specification of Digital Systems (TAU). He is currently an Associate Editor for IEEE TRANSACTIONS ON CAD and previously served as Associate Editor for IEEE TRANSACTIONS ON VLSI SYSTEMS. He is a member of the ACM, American Society of Engineering Education, and Eta Kappa Nu.