CAD and Circuit Techniques for Low Power, Variation-Aware SRAM Design

---

A Thesis

Presented to

the faculty of the School of Engineering and Applied Science

University of Virginia

In Partial Fulfillment

of the requirements for the Degree

Master of Science (Electrical Engineering)

by

Peter Beshay

May 2014

APPROVAL SHEET

The thesis is submitted in partial fulfillment of the

requirements for the degree of

Master of Science (Electrical Engineering)

Peter Beshay

This thesis has been read and approved by the examining Committee:

Benton H. Calhoun (Thesis advisor)

Mircea R. Stan

John Lach (Chair)

Accepted for the School of Engineering and Applied Science:

James H. Aylor (Dean, School of Engineering and Applied Science)

May 2014

Acknowledgements

I would like to thank my advisor, Prof. Benton H. Calhoun, for giving me the opportunity to work with him and for his encouragement and support throughout my academic work at University of Virginia. I would also like to thank my committee members, Prof. John Lach and Prof Mircea Stan for useful discussions. I want to thank my colleagues and collaborators Pete Stevenson, Dinesh Patel from Stanford University, Arijit Banerjee, James Boley, Jonathan Bolus, Saad Arrabi and Joseph Ryan from the University of Virginia for interesting discussions and a great college experience. I also want to thank my family for their continuous support and encouragement.

Table of Contents

1. INTRODUCTION

1.1 Low Energy Applications

Nowadays, various biomedical and mobile applications require ultra-low-power hardware implementation. Examples on mobile multimedia devices are portable handsets and tablets. Those portable devices are not expected to be connected constantly to a power supply source and therefore their implementation should have higher energy efficiency than other devices that can be connected to a constant source of power supply like desktops for instance. Those portable devices accordingly need the time between the battery charges to be extended in the order of days. With the advancement of technology, the capabilities of these devices are being extended and improved, for instance, the delivery of high definition video. This leads to increased energy consumption and hence careful low power hardware design is essential to enable the device to meet the low energy constraints and performance requirements.

Another example on energy constraints applications are the biomedical devices (implantable devices, pacemakers, implants, and stimulators) as shown in Table 1. In those applications the life time of the power supply (battery) sets the time between surgical replacements and therefore sets a limit on the energy consumption of the system [1]. As demonstrated in Table 1, although the energy constraints are tight, the performance constraints are relaxed in those applications (< 1MHz). Hence the system can be optimized to achieve ultra-low power implementation on the cost of performance.

Static Random Access Memory (SRAM) is essentially needed and commonly used in most of these applications to either store the data recorded by the device and/or store the instructions to be executed by the device. SRAMs contribute significantly to the total power consumption of these applications [1]. Figure 1.1 shows the energy consumed by SRAMs in varies applications (Intel core 2 processor, ARM 1176JZ

processor and MSP 430) [1]. The figure indicates that in energy constraints applications like the MSP case, the energy consumed by the SRAM memory ~70% is more significant than the energy consumed in the performance constrained ones (~20% Intel Core 2), and thus the reduction of the SRAM power consumption is crucial for these applications. This thesis is focused on the implementation of low energy SRAM memories that are used inside energy constraints systems to enable low power hardware implementation of those systems and at the same time meets their performance specifications.

Table 1. Performance specification of energy constrained biomedical applications.

Application	**Performance Specification**		
	Power	**Performance**	**Source of supply**
Pacemakers & defibrillators	<10uW	1kHz DSP	10-year life time battery
Hearing aid & Cochlear implants	100-2000 uW	32kHz-1MHz DSP	1-week life time battery

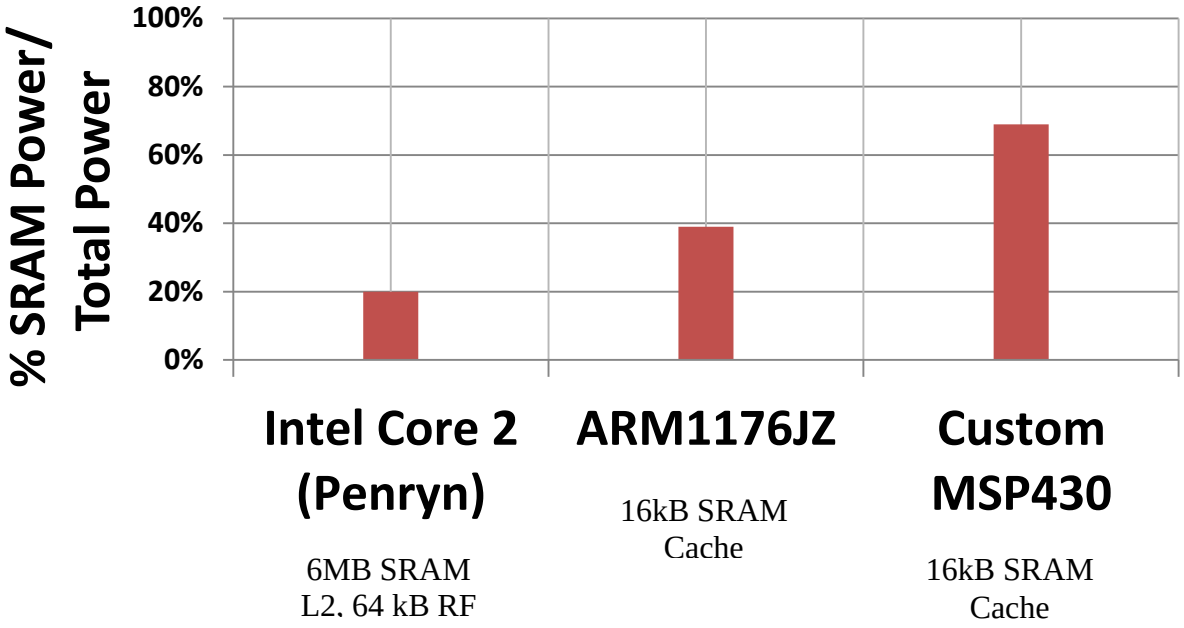


Figure 1.1 SRAM power consumption contribution in various applications.

1.2 Low Energy SRAM Design and Requirements

Figure 1.2 shows the trade-off in the design of SRAMs. In some applications like desktops and servers, high performance operation of SRAMs is required, in other applications like portable devices low density and power are the main objective on the cost of reduced performance. In this work, we focus on the design of low power SRAMs. The SRAM bit-cells occupy a large portion of the total chip area ~50-75% [1] and accordingly are usually designed for near minimum sizes. Therefore, they are susceptible to process variation and are not able to reap the reduction in power consumption benefits of technology scaling as other blocks. We tackle this major problem by proposing circuits that combat the effect of process variations to reduce the dynamic energy consumption. In chapter 2 and 3 the circuits are presented to reduce the read and write energy respectively through combating the effect of process variations during the read and write operations and in chapter 4 a CAD method is presented to locate the optimal structure of the SRAM macro (architecture and circuit design parameters) to reduce the array energy consumption.

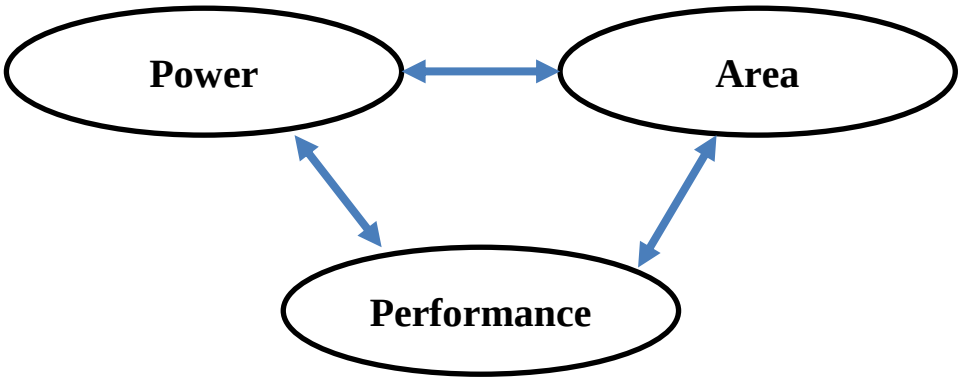


Figure 1.2 SRAM design trade-off.

1.3 Contibution of the thesis

The contributions are listed below

- Circuits and methods that minimizes the energy of the SRAM Read operation through improving the intrinsic offset of the sense amplifier circuit (Chapter 2)
- A digital sensor that modulates the Wordline (WL) pulse duration of the SRAM memory to minimize the energy of both the Read and Write operations while maintaining robust operation (Chapter 3)
- A scheme that co-optimize the architecture of the SRAM memory and the underlying circuits to generate optimal designs (Chapter 4)

1.4 Outline of the thesis

Chapter 2 – SRAM Sense Amplifier Offset Cancellation: This chapter presents methods to improve the offset of the sense amplifier circuit to minimize the energy of the SRAM Read operation. In the first section, we present a technique that uses typically detrimental aging effect to improve the SRAM sense amplifier offset. The method can be integrated to the post fabrication process and used in addition to other compensation techniques as an initial step to reduce the offset. In the second section we presented an online offset compensation circuit that uses auto-zeroing technique that is specifically useful in sub-threshold region and we include a comparison to the state of the art compensation methods.

Chapter 3 – A Digital Dynamic Write Margin Sensor for Low Power Read/Write Operation in 28nm SRAM: This chapter presents a WL adaptation scheme that minimizes the energy of the Write and Read operations of the SRAM while maintaining roust operations. The scheme utilizes a digital sensor, Built-In-Self-Test Circuit and WL Quantizer circuit to adapt the WL pulse duration with Process, Voltage and Temperature variations.

Chapter 4 – A Hybrid Scheme for Architecture/Circuit Co-design for Complete SRAM Macros: This chapter presents a hybrid optimization methodology for SRAM circuit-architecture co-design that combines convex optimization (CO) to explore the power-performance trade-off in lower level blocks with macro level optimization. We apply this methodology to explore the power-performance tradeoff in 4kB, 16kB, and 64kB SRAMs in a 90nm technology. We showed that the hybrid optimization approach generates better designs than a top level approach that omits circuit level optimizations.

Conclusion: A summary of the contributions and future directions

Publications: A list of publications and patents that are published on this work.

Bibliography: A list of references used in this thesis.

1.5 Collaboration

Section 2 in Chapter 2 (using Auto-zeroing circuit for offset compensation) is a collaborative work between the author of this thesis and Joseph Ryan, a PhD graduate from the University of Virginia. Joseph was responsible for the complete design of the circuits and the layouts. The author of this thesis performed the chip measurements, generated the simulation data and wrote publications on this work. In addition, the author modified the circuit to minimize the dynamic power consumption. Other chapters and sections are the main work of the author.

2. SRAM SENSE AMPLIFIER OFFSET CANCELLATION

2.1 Motivation and Background

Device variability in modern processes has become a major concern in SRAM design leading to degradation of both performance and yield. While low swing bit-lines reduce the power consumption and improves performance, offset in the sense amplifiers due to variability hinders the scalability of this technique [2]. The effect aggravated in the sub-threshold region. To illustrate the variation of the SAs offset across technologies, 1000 Monte-Carlo simulations were performed to evaluate the offset voltage of a latch-based SA using 45, 65, 90 and 130 nm commercial technology models. Figure 2.1 illustrates the 3σ value of the SA offset voltage across technologies. The results indicate that the largest offset voltage belongs to 45 nm technology, followed by 32 nm, 65 nm, 90 nm and 130 nm respectively. Although the offset behavior is not monotonically increasing with technology scaling, the plot indicates a trend of increased offset in emerging technologies, *i.e.*, 32, 45 nm. Since offset voltage is increasing with newer technologies, compensation becomes more essential as technology scales. In this chapter we present methods to improve the SRAM sense amplifier offset. A big portion of this chapter is from (1)(2)(3) in the publication section. Several attempts have been made before to tackle the problem of the offset voltage in the sense amp including redundancy [3], transistor upsizing [4], and digital compensation [5][6]. In section 1 we present a post fabrication technique using typically detrimental aging effect to improve the SRAM sense amplifier offset. Unlike the aforementioned approaches, the method is not considered during design time. It can be integrated to the post fabrication process and used in addition to other compensation techniques mentioned above as an initial step to reduce the offset. In section 2 we presented an online offset compensation technique that uses auto-zeroing technique. The circuit enables flexible tuning of the offset voltage. The main advantages of the approach are the near-zero offset after cancellation, offset tuning, and the automatic temperature, voltage, and aging tracking achievable using a repeated offset calibration phase, which makes

the design useful in the sub-threshold region due to the high offset voltage sensitivity to supply voltage and temperature variations. We include a comparison to the state of the art online compensation methods.

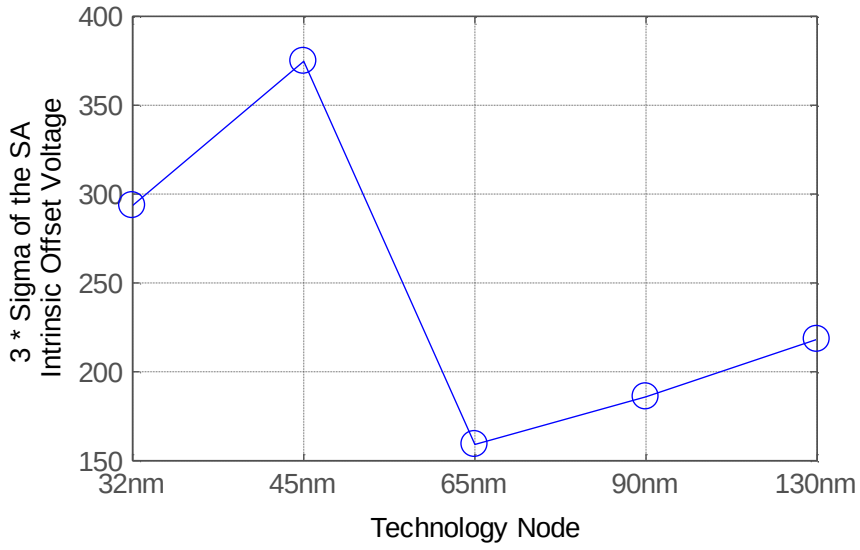


Figure 2.1 3σ of the intrinsic voltage of the SA across technology nodes.

2.2 Using BTI Stress

2.2.1 Shceme

Bias temperature instability (BTI) is one of the major aging issues that weaken transistors over time. Many studies have examined the impact of BTI on offset voltage and proposed techniques to reduce BTI effects [5][7]. In lieu of only considering BTI as a variation source, it can however be used as a mean of variation compensation. More interestingly, due to its dependence on the stress condition, we can exploit BTI to combat random variation (i.e. mismatch) by applying different stress conditions for different transistors. Our approach is adapted from [8]. For a sense amp with mismatch, when we only stress transistors with higher strength, BTI induced change can offset mismatch and thus contribute to balancing the sense amp.

The scheme is applied on a latch based sense amp shown in Figure 2.2. The main contribution of the sense amplifier offset comes from the NMOS transistors as they are responsible to drive the output nodes to the trigger point while both the PMOS are still off, so stressing the PMOS transistors will have negligible impact on the offset voltage compared to stressing the NMOS. Variations in the NMOS access transistors M2, M3 are more dominant in determining the offset than M4 and M5. Consider a case where the threshold voltage of M2 is lower than that of M3 as shown in Figure 2.3. If a zero-differential input is applied to the sense amp, the output node “OUT” will be driven to 0. M1, M2, M3, M5 and M6 will be in strong inversion indicated by the red circles and will be affected by the BTI stress during compensation. M4 and M7 will be off.

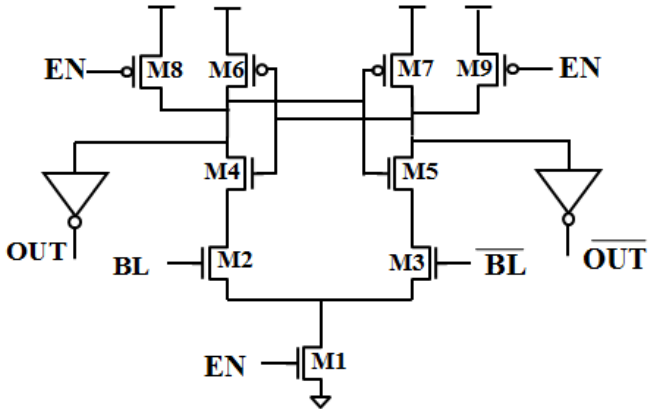


Figure 2.2 Latch-based sense amp

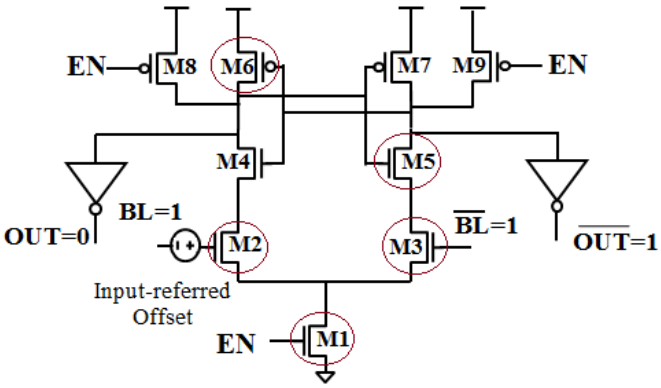


Figure 2.3 Sense amp latches according to the input-referred offset

To avoid stressing both the access transistors, the output is feedback to the input as shown in Figure 2.4 to turn off M2 during stress. The feedback saturates the sense amp in the offset direction leaving only M1, M3 and M6 in strong inversion to be affected by the stress as indicated by the red circles. The scheme has negligible power budget. It can be easily integrated into burn-in test and use burn-in stress to offset mismatch, it can also be periodically used through the life time of the chip to offset variation from real-time aging and could be easily integrated with other compensation techniques as an initial step to reduce the offset variations.

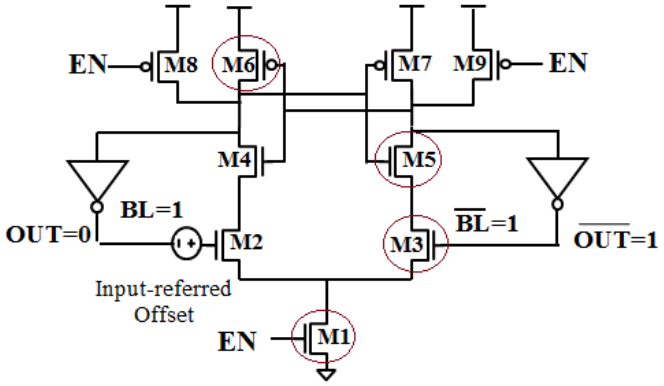


Figure 2.4 Feeding output to input during stress to avoid stressing M2

Overstress Avoidance

Applying the stress only one time might cause an extra increase in the threshold voltage that might move sense amplifier offset to the other polarity rather than canceling it. The alternative solution is to repeatedly power-up and stress the array. According to the new offset value, transistors responsible for deriving the output node will be in strong inversion which will then be affected by the stress. Stress is repeatedly applied after latching sense amp according to the new offset value to completely cancel the offset by allowing small recovery every power-up.

Figure 2.5 shows the effect of applying 80mv stress once vs. applying it on four increments of 20mv with latching the sense amp after each stress step. The sense amp initial offset was 50mv. Applying stress once overstressed the sense amp in the opposite direction of the offset causing the offset to shift to the other polarity. Repeated stress allows fine tuning of the offset and it can compensate for very small offset values.

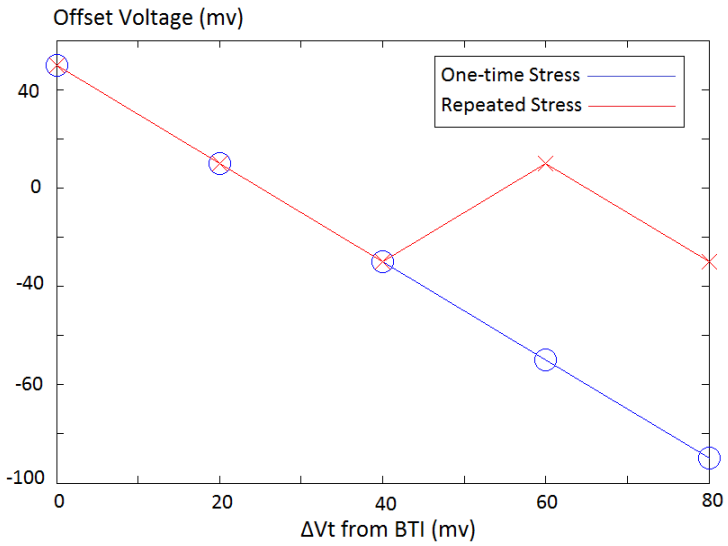


Figure 2.5 Effect of One-time and repeated BTI stress on the offset voltage Simulation results

Monte-Carol simulations are done to account for variations in the threshold voltages of M2, M3, M4 and M5. Figure 2.6 and Figure 2.7 shows the mean and standard deviation of the offset voltage after each stress step respectively.

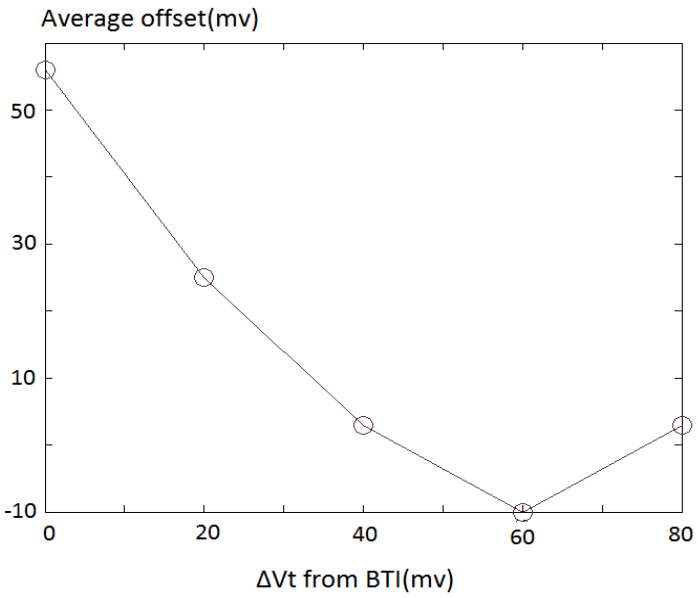


Figure 2.6 Average value of the offset voltage after each stress step

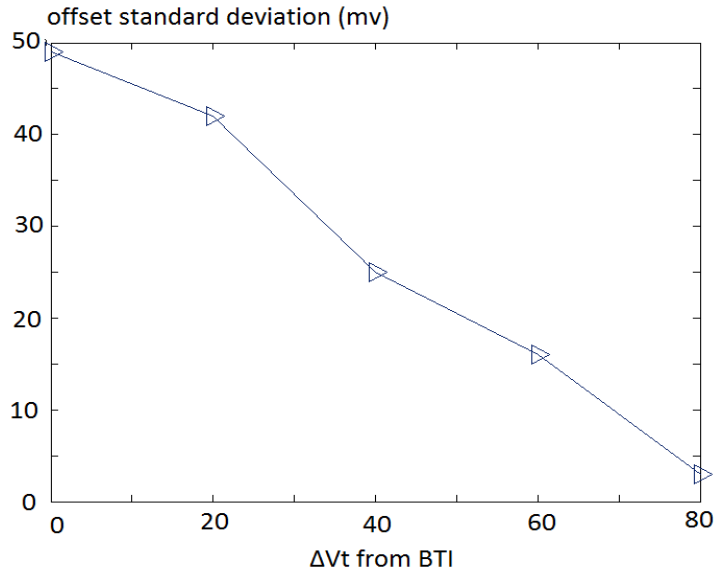


Figure 2.7 Standard deviation of the offset voltage after each stress step

Its noticeable that both the average and the standard deviation doesn't change linearly with the induced threshold, the reason for that is while the BTI stress decrease the offset for some sense amps it increases it for those with offset voltage close to zero. This effect can be minimized by decreasing the period of stress to avoid overstressing the sense amp whose offset is close to zero.

Increasing the threshold voltage decreases the leakage power but has the downside of degrading the sense amp speed. Figure 2.8 shows the average delay after each stress step. Figure 2.9 shows the average of the total leakage power after each stress step.

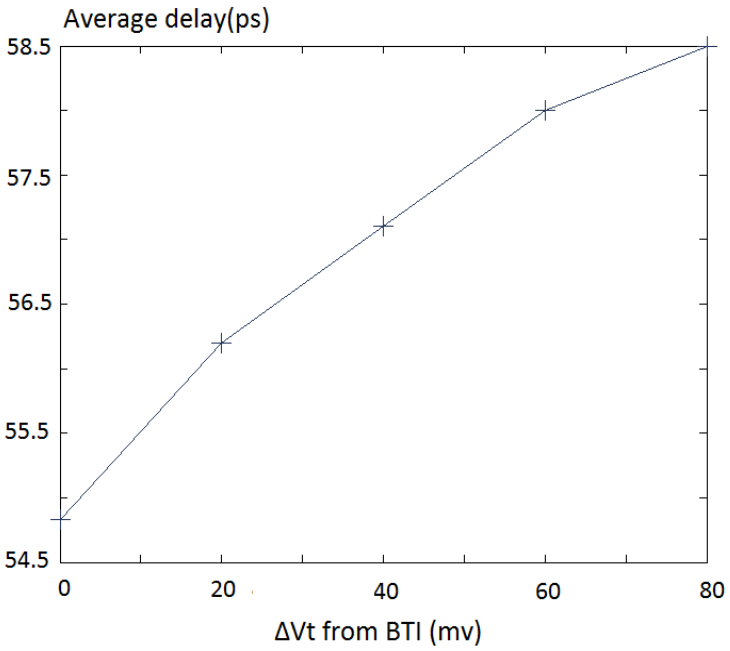


Figure 2.8 Average delay after each stress step

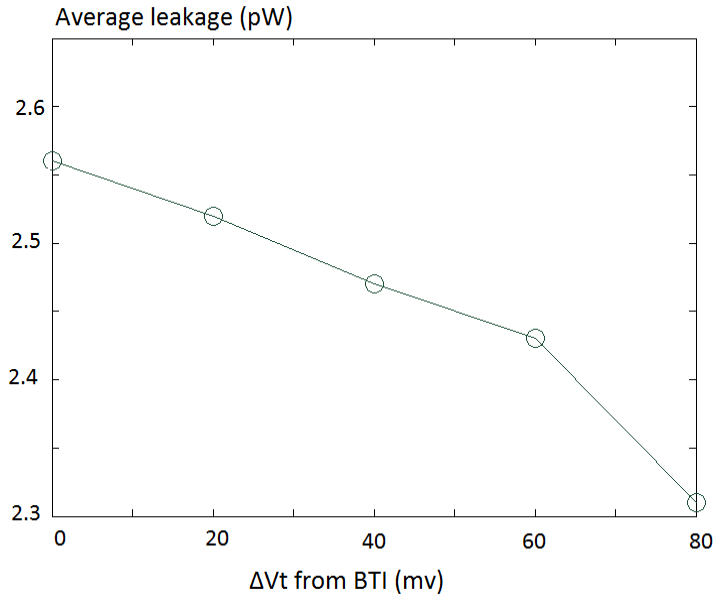
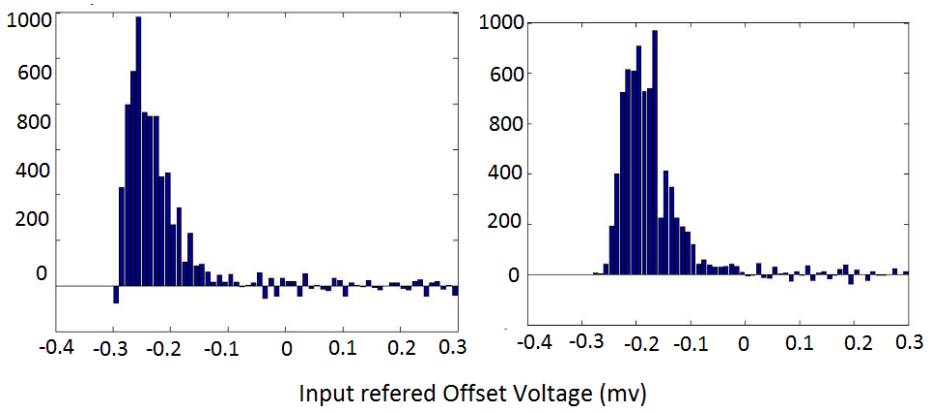


Figure 2.9 Average leakage after each stress step

2.2.2 120nm Test Chip Measurements

A chip is fabricated in 120nm technology constitutes 480 sense amplifiers, used for testing the scheme. To verify the scheme, the chip was stressed at 1.7× of the nominal VDD and 45°C for 12 hours, measurement of the offset voltage were recorded every 4 hours. Figure 2.10 shows the initial and final distribution of the offset voltage. Figure 2.11 shows the average offset voltage after each stress step.



(a) (b)

Figure 2.10 a. Initial distribution of the offset voltage b. Final distribution of the offset voltage

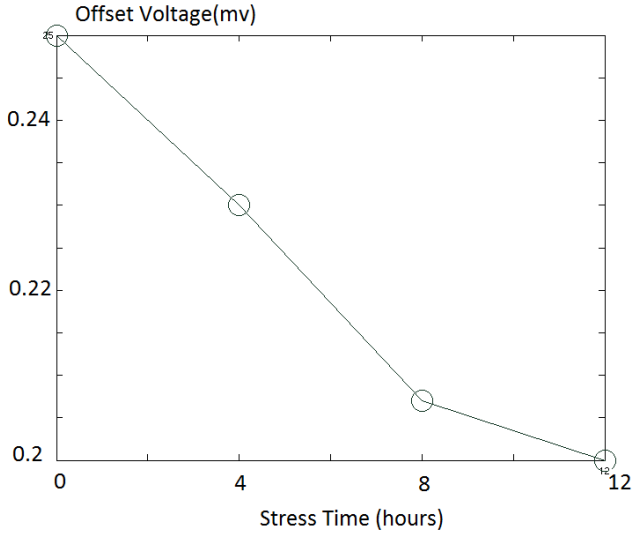


Figure 2.11 Average offset voltage after each stress step

2.2.3 Conclusion

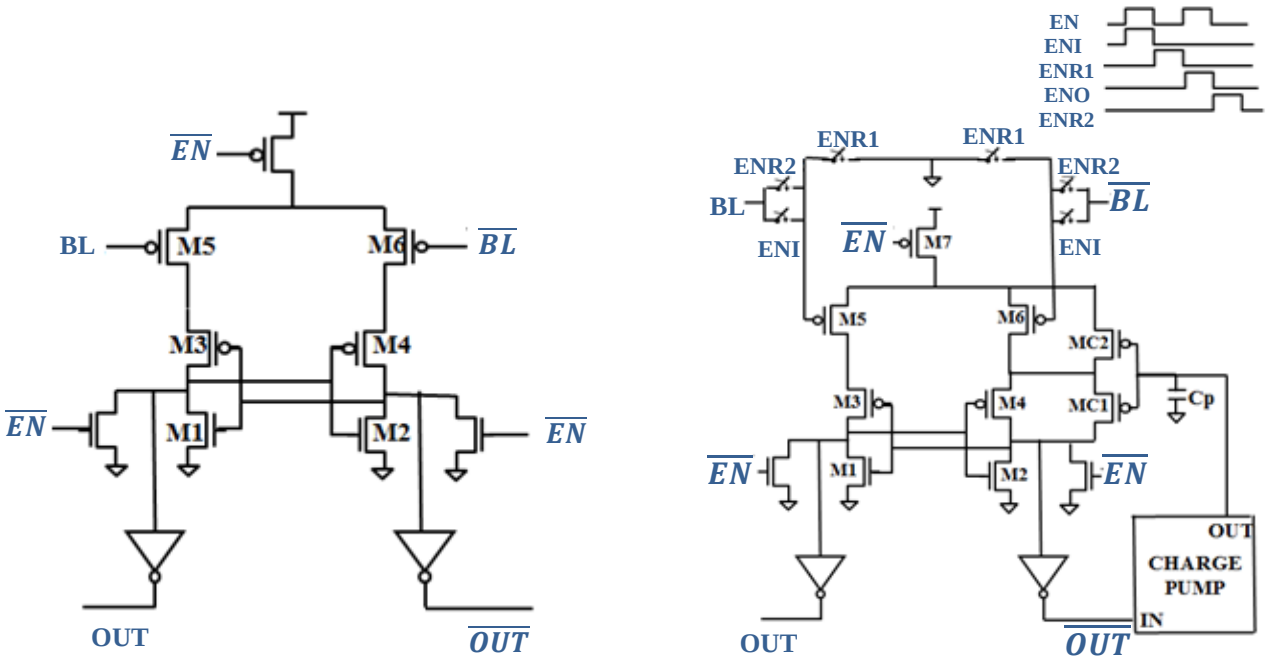
We presented a scheme that uses the typically detrimental aging effect to improve the SRAM sense amplifier offset. Offset compensation of 50mv was achieved using stress of 1.7× of the nominal VDD and 45°C for 12 hours. It can be periodically used through the life time of the chip to offset variation from real-time aging and could be easily integrated with other compensation techniques as an initial step to reduce the offset variations.

2.3 Using Auto- Zeroing Circuitry

2.3.1 Scheme

Our approach to eliminating offset is a digital auto-zeroing (DAZ) scheme inspired by analog amplifier offset correction. The main advantages of the approach are the near-zero offset after cancellation, offset tuning, and the automatic temperature, voltage, and aging tracking achievable using a repeated offset calibration phase, which makes the design useful in the sub-threshold region due to the high offset voltage sensitivity to supply voltage and temperature variations. In [9], a similar scheme is used, the dynamic compensation. A group of transistors are selectively coupled to high and low voltage levels via multi-phase timing. This results in a voltage level on nodes of interest that is a function of transistor mismatch. The voltage levels act to compensate for the transistor mismatch. This scheme is similar to the auto-zeroing scheme presented in this work. However the presented scheme uses a compensation capacitor, charge pump and feed-back circuit. Hence, the calibration phase is not necessarily needed prior to every sensing cycle. This improves the SA power consumption as will be illustrated.

Our auto-zeroing scheme uses a split-phase clock and charge pump feedback circuit. Figure 2.12a shows a conventional latch-based sense amp with PMOS inputs (e.g., to support near-V_{SS} sensing on a low swing bus). Figure 2.12b shows the auto-zeroing circuit attached to the sense amp.



(a) **(b)**

Figure 2.12 (a) Latch-based sense amp for near-V_{SS} inputs; (b) Auto-zeroing circuit attached to the sense amp.

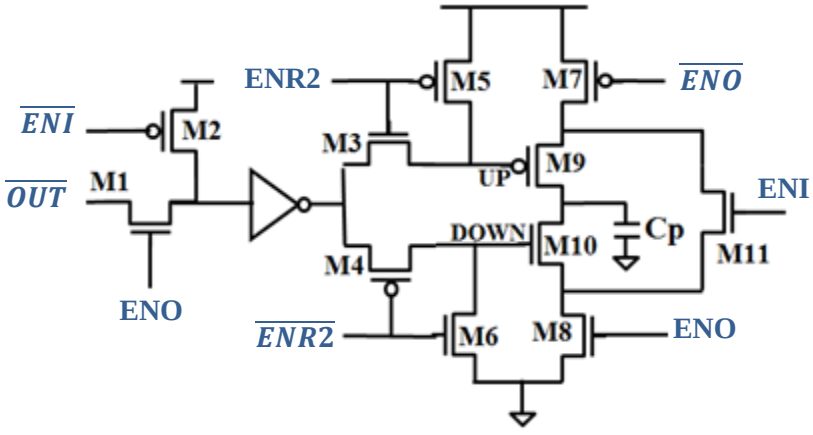


Figure 2.13 Charge pump circuit for adjusting the voltage on Cp.

The same scheme can apply to a SA with NMOS inputs in an SRAM. The charge pump circuit is shown in Figure 2.13. ENI and ENO are the input voltage differential and offset tuning phases respectively. ENR1 and ENR2 are reset phases. During ENR1, a zero differential input is applied to the sense amp. The ENO phase then occurs, and the SA resolves based on its intrinsic offset. The sense amp output is fed to the charge pump circuit that charges the capacitor, Cp, up or down. During ENR2, the differential input is applied to the sense amp. ENI then occurs, and the SA resolves based on the differential input. Note that phases ENR1 and ENO can be omitted or included based on how often re-calibration is needed. Transistors MC1 and MC2 control the drive strength of the right side of the sense amp to compensate for the offset. The charge pump controls the drive current in both transistors to equalize the strength of the SA right and left sides to reduce the offset. The offset is compensated with minimal capacitive loading at the output and is independent of input DC bias (V_{INDC}). A supply voltage and clock frequency of 0.5 V and 1MHz are used in the simulations. The output voltage of the sense amp and the voltage on Cp are illustrated in Figure 2.14 for an input differential of −10 mV. The initial voltage on Cp is zero. This causes an intrinsic positive offset voltage that set the SA output voltage to 1. Simulations indicate that the voltage on Cp required for a zero offset is 142 mV. For a 10 mV offset, the voltage on Cp can vary within ±12 mV.

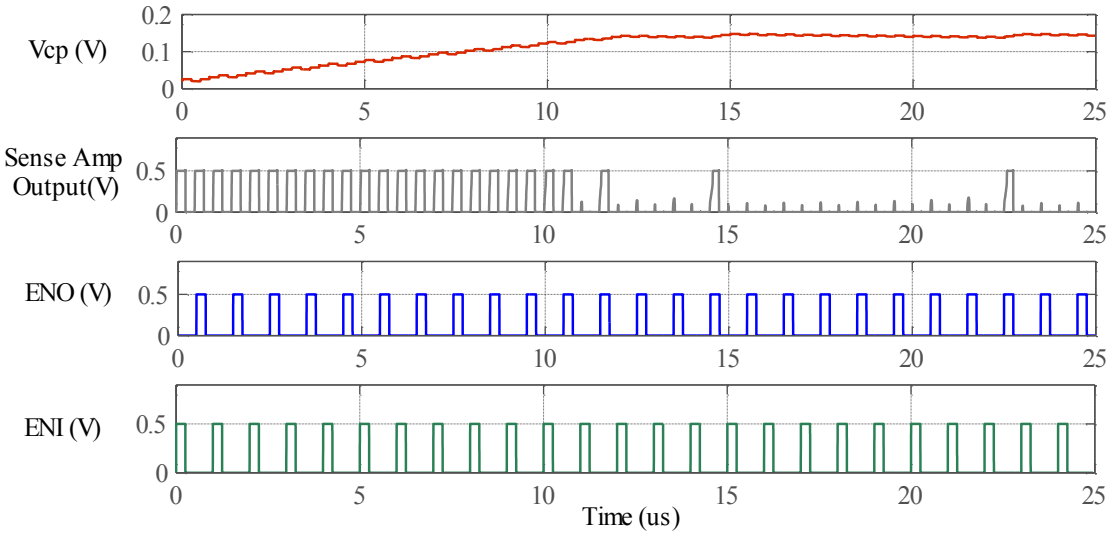


Figure 2.14 Simulated output voltage of the sense amp and Cp voltage for a −10 mV differential input voltage at 0.5 V and 1MHz in 45 nm CMOS.

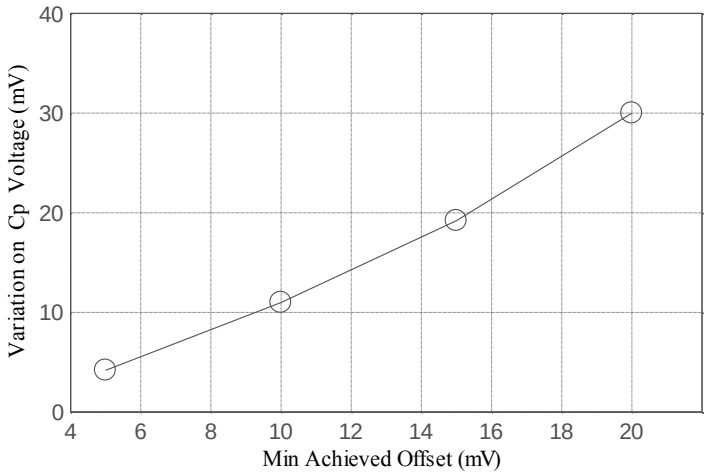


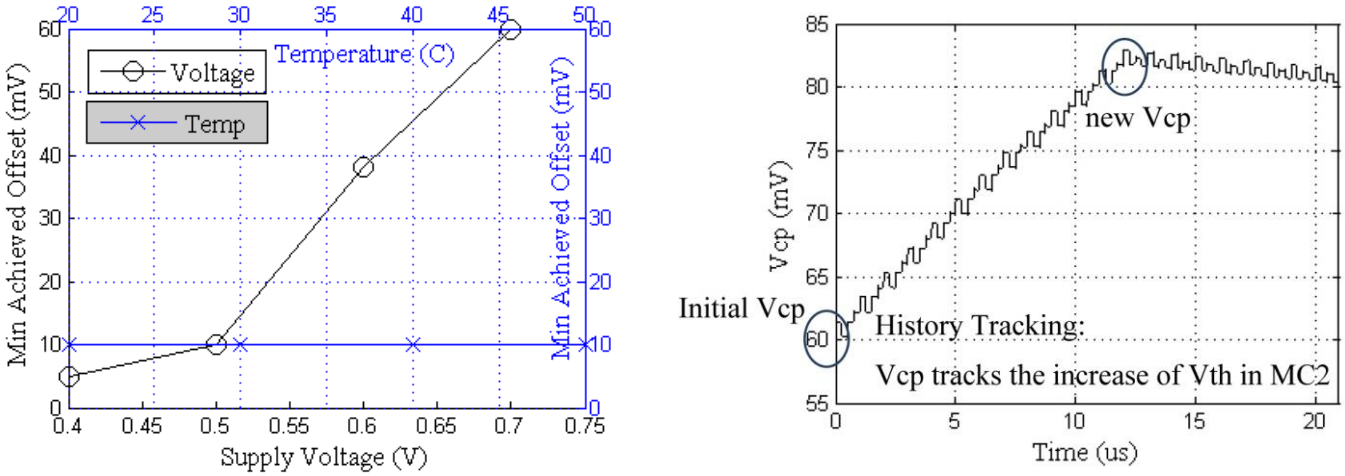
Figure 2.15 Variation on Cp voltage *vs.* Minimum Achieved Offset.

This imposes a minimum and maximum limit on Cp voltage to 130 mV and 154 mV in order to maintain an offset less than 10 mV. The deviation of Cp voltage from to the value corresponding to zero offset (142 mV in this case) is plotted in Figure 2.15 for desired final offset voltages of 5 mV, 10 mV, 15 mV, and 20 mV. Low offset voltages are usually realized using a higher value of Cp. In Figure 2.14, the offset compensation is completed when the voltage on Cp settles to its final value within the 130 mV to 154 mV range. The sense amp then resolves its output correctly to 0 during the input phase. A zero differential voltage is applied to the SA input during the offset phase. This sets the SA output to "1" when Cp voltage drops below 142mV and "0" otherwise. In this design, rate at which Cp charges up is higher than its charge down rate. This helps to minimize the power consumption as will be discussed in Section 5.

Voltage, Temperature, and Aging Tracking

To demonstrate temperature, voltage, and aging tracking, the offset voltage that remains after compensation is calculated for various voltages and temperatures as shown in Figure 2.16a. Simulations in a commercial 45 nm process show that the circuit maintains a constant offset across temperature. The accuracy of voltage tracking depends on the supply voltage. Higher supply voltage causes more charge to be pumped to Cp

during each offset calibration cycle, and this larger change in charge leads to a coarser resolution, as Figure 2.16a illustrates.



(a) (b)

Figure 2.16 (a) Voltage and temperature tracking; (b) Aging tracking

The auto-zeroing scheme also has the ability to compensate for any changes in device characteristics after circuit deployment. One common cause for such changes is effective threshold voltage shifting due to Bias Temperature Instability (BTI), hot carrier injection, or other aging effects. To demonstrate how this circuit can compensate for such changes, Figure 2.16b shows the capacitor voltage after an abrupt increase in the threshold voltage of MC2, to emulate an aging effect. The charge pump boosts the voltage on Cp to decrease the drive strength of MC2 in response and rapidly restores the compensated offset voltage.

Offset Tuning

We define settling time as the difference between the time when the zero differential-input is first applied and the time when the voltage of the output capacitor settles as shown in Figure 2.17. Changing the size of the output capacitor (Cp) affects the amount of charge added during the offset compensation phase (ENO) and so controls both the offset and the settling time. Figure 2.18 demonstrates the trade-off between accuracy (min achieved offset) and settling time using different values of output capacitors.

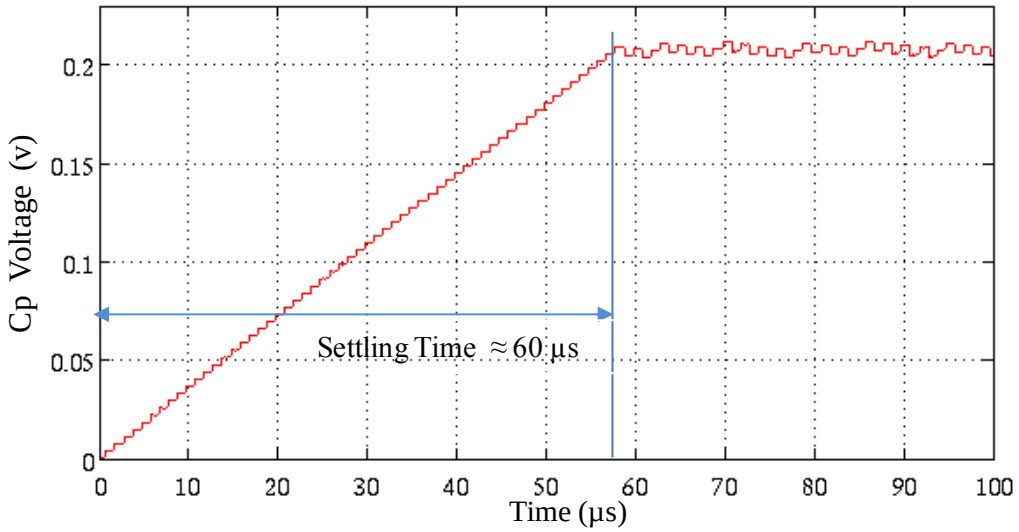


Figure 2.17. Settling time simulation

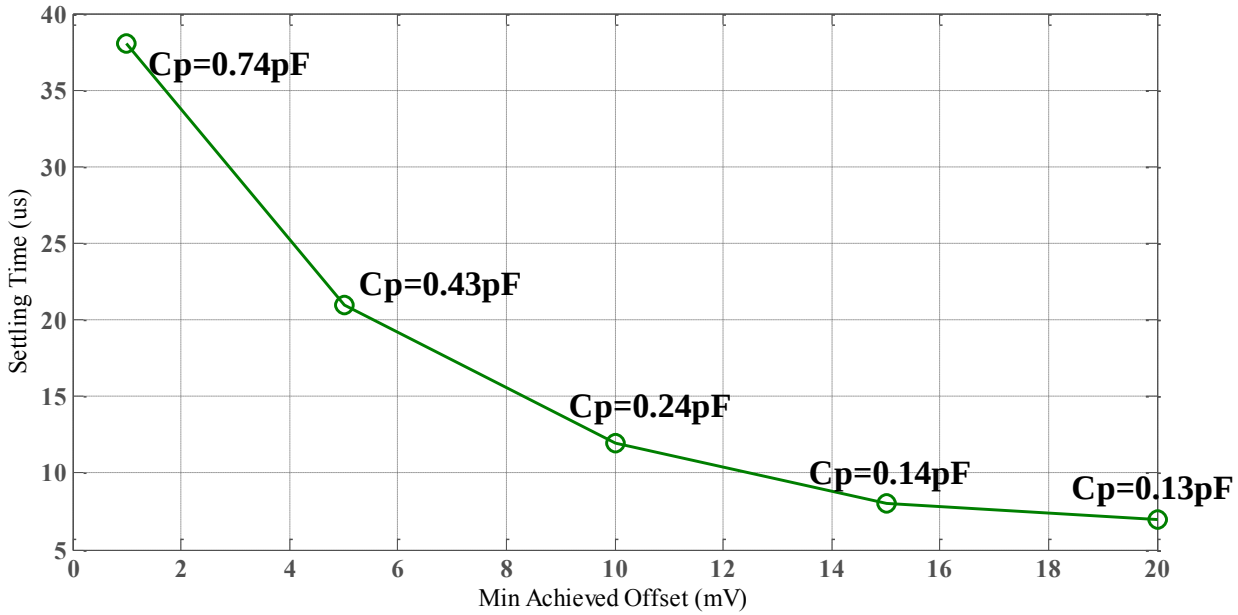


Figure 2.18 Min achieved offset *vs.* settling time for different values of output capacitor

Power Consumption

The main contribution to the power consumed by the DAZ SA comes from the continuous calibration. Decreasing the number of cycles of calibration phase (ENI and ENO) relative to the input phase decreases the switching power of the feedback circuit and the power consumed in charging and discharging (Cp) but is limited by the leakage at the output capacitor (Cp). The overhead area of the scheme includes the area of the

timing circuit, the charge pump circuit, and the output capacitance (Cp). For an offset voltage of 1mV, a 0.74pF output capacitance is needed. In this case, the area of Cp can dominate the total area overhead. In Figure 2.19, the offset calibration phase occurs once every 15 clock cycles. The maximum calibration period or the minimum number of offset calibration cycles needed is limited by the leakage on Cp. Simulation results indicated a maximum calibration period of 200 µs. This high period makes the difference in power consumption between the DAZ SA and the Latch SA insignificant. The total power consumption of the DAZ SA and the Latch SA is 2.02 nW and 2 nW respectively. The minimum number of offset calibrations is independent of the required offset or the value of Cp, but it depends on the charge pump current. In this design the charging current is 0.5 µA. High current allows fewer number of calibration cycles. Increasing the charge pump current however increases the dynamic power consumption.

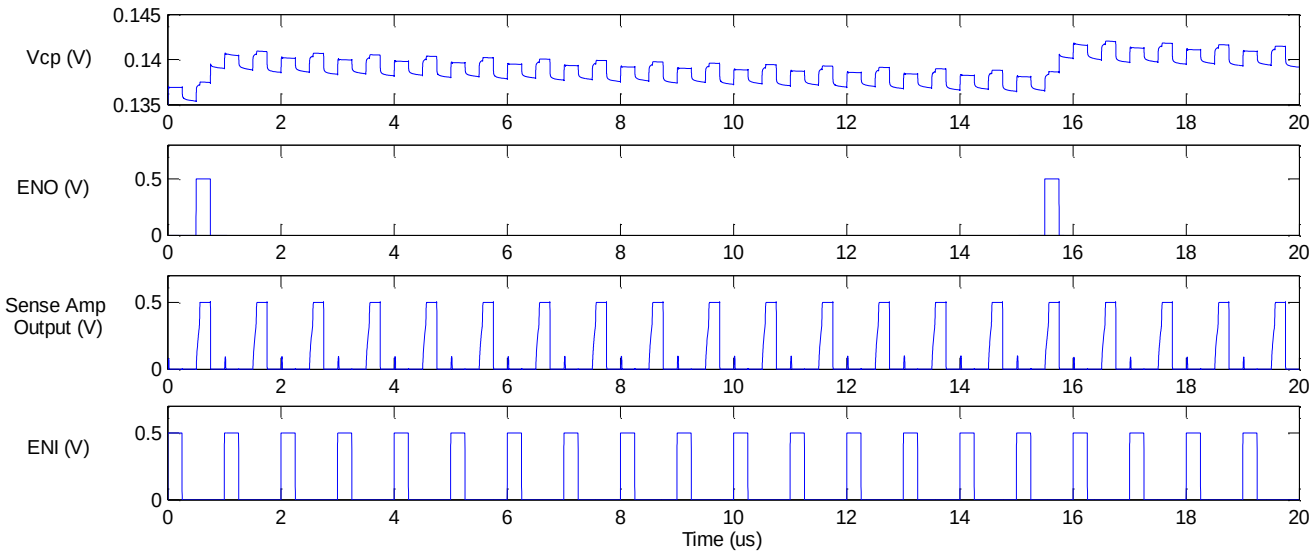


Figure 2.19 Offset compensation clock period for a 10 mV offset voltage.

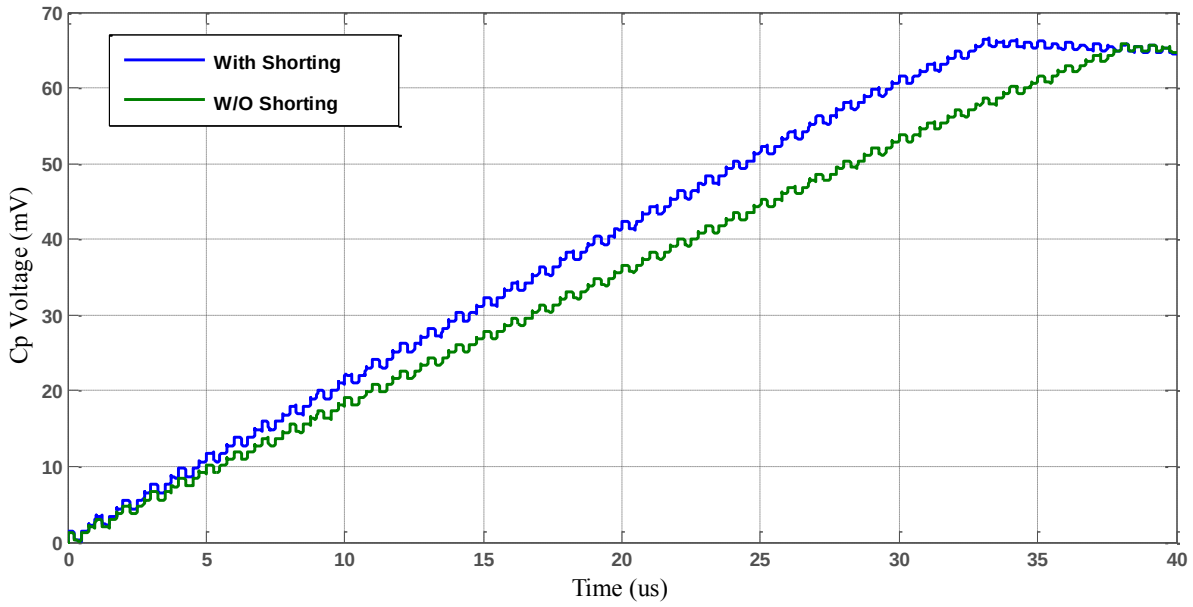


Figure 2.20. Voltage on Cp with and without shorting the virtual supply nodes of the charge pump.

Shorting the output virtual nodes of the charge pump through M11 can decrease the leakage by reducing V_{DS} of the switches and improve settling time as shown in Figure 2.20. The switching power can also be decreased by strengthening M9 in the charge pump circuit relative to M10 to avoid the continuous toggling of the sense amp output during offset compensation phase (ENO) after settling as shown in Figure 2.21. Strengthening M9 has the downside effect of increasing the settling time when Cp is moving to lower voltages; the time Cp takes to discharge will increase. However, the compensation usually starts with zero-initial voltage on the capacitor Cp that makes the settling time mostly dependent on the charging rate.

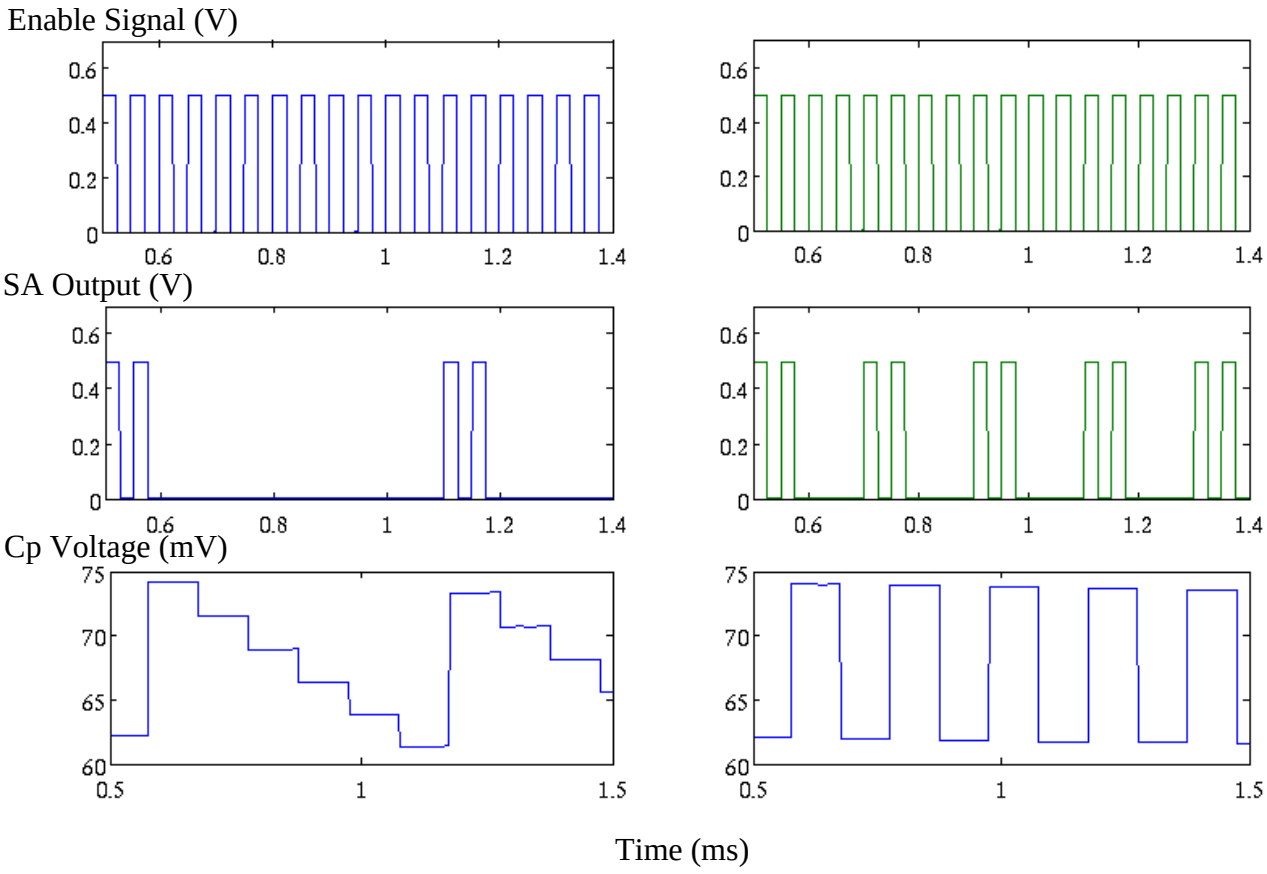


Figure 2.21 Strengthening the pull up transistor of the charge pump (left column of sims) reduces the rate at which the SA output switches high relative to equal strength devices (right column). This reduces power consumption

Offset Sensitivity

The sensitivity of the offset compensation depends on the split phases, charge pump circuit, and the output capacitance. The accuracy of the split phases has the dominant influence on the resolution. A small overlap between ENO and ENR2 phases can dramatically degrade the accuracy by connecting M1 and M2 to the supply rails during charging. That leads to a significant increase in the charge

pump rate degrading the accuracy as shown in Figure 2.22b, where the min achieved offset is plotted against the error in split phase timing, measured as the percentage of time overlap between ENO and ENR2. The scheme is also sensitive to variations in the M9 and M10 transistors in the charge pump circuit. They are responsible for charging/discharging Cp, and so the one with more drive strength determines the final offset value. Figure 2.22a shows the sensitivity of the offset voltage to the output capacitance Cp.

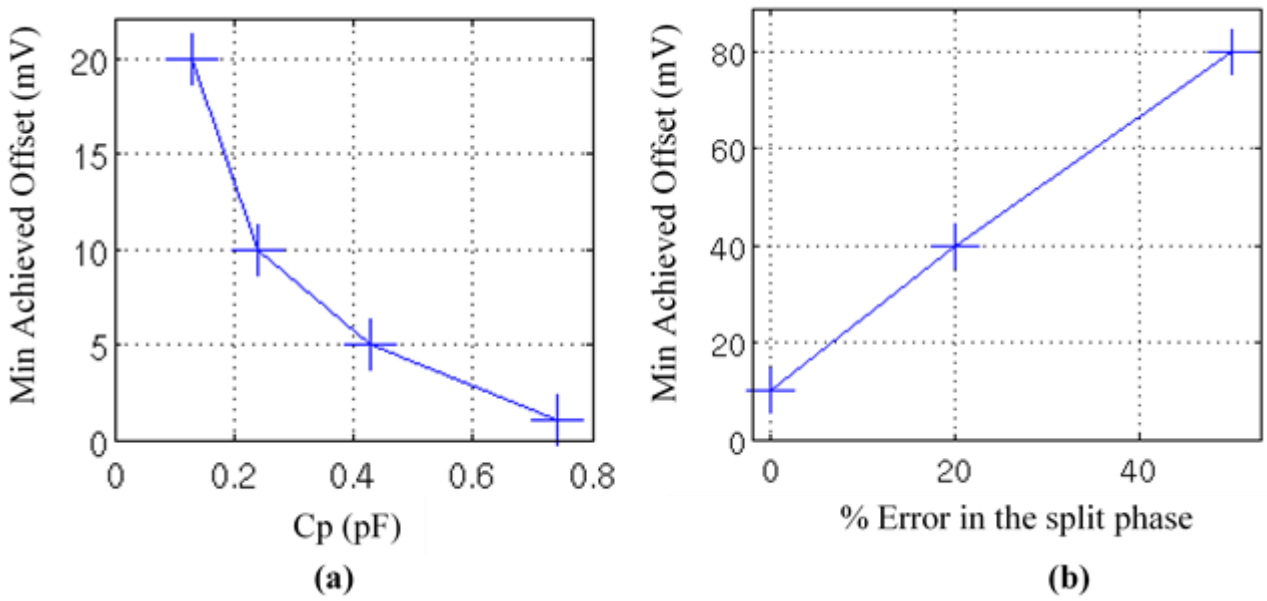


Figure 2.22 (a) Offset voltage sensitivity to output capacitance; (b) sensitivity to split phases

The offset voltage is also sensitive to the frequency of the split phase. The increase in the split phase frequency increases the enable signal switching and degrades the compensated offset voltage.

16 kB SRAM Design

In this section, we investigate the effect of utilizing the DAZ SA in a 16kB SRAM memory. The power consumption of the DAZ SA is higher than the uncompensated one due to the clock generator, charge pump and the buffer stages needed for the non-overlapping clock. The sense amplifier delay is also higher due to the high capacitive loading. Reducing the sense amp offset reduces the necessary bit-line swing, which decreases both the precharge and bitcell energy and delay during the read operation. The reduction in

the read energy and delay depends on the number of banks, rows and words per row of the memory. The energy and delay of the 16 kB SRAM is calculated using a 20 mV DAZ SA for all possible configurations and plotted in Figure 2.23. Each point is annotated with (B, R, W) where B is the number of banks, R is the number of rows, and W is the number of words per row. The results indicate a significant improvement in both the energy and delay for cases with large numbers of rows and small improvement or degradation for cases with small number of rows. The design point of 1 bank, 512 rows and 2 words per row has the biggest improvement of 10% in energy and 24% in delay. The design point of 4 banks, 32 rows and 8 words per row has the biggest degradation of 6% in energy and 5% in delay. The DAZ SA created 3 new optimal design points (1, 256, 4), (1, 128, 8) and (8, 128, 1). The improvement in energy and delay for (1, 256, 4) is 12% and 13% respectively. For (1, 128, 8) it improves the energy by 13% and degrades the delay by 5%. The energy of (1, 128, 8) with DAZ SA is the lowest. This could not be achieved using a Latch SA. The system level parameters of the SRAM that satisfy the energy/delay requirements changed (*i.e.*, design point (16, 64, 1) is the minimum energy/delay point with uncompensated SA. Using DAZ SA, the min delay/energy design changed to (1, 128, 8).

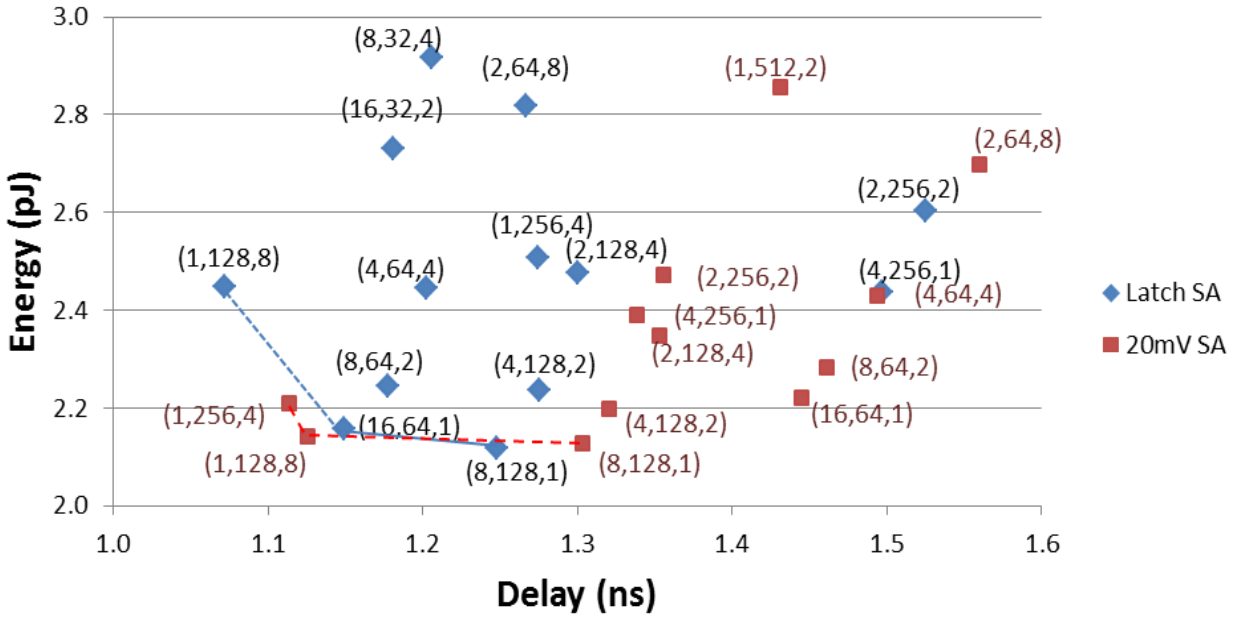


Figure 2.23 Design space of 16kB SRAM Memory with uncompensated and 20 mV digital auto-zeroing (DAZ) SA.

Comparison to other Offset Compensation Schemes

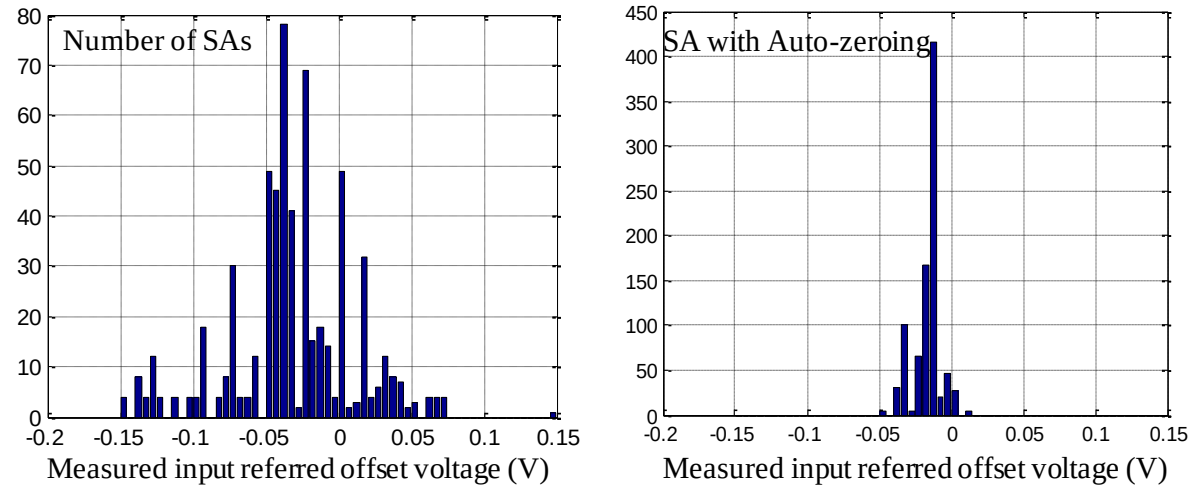
The main advantages of the scheme are the continuous calibration that makes it specifically useful for sub-threshold operation and the flexibility to tune the offset voltage. The latter provides different design options that can be utilized in the SRAM design process. Approaches like redundancy [3], transistor upsizing [4], and digitally controlled compensation [5][6] do not support continuous calibration and hence would not be tolerant to voltage and temperature variation. The approach in [9] provides continuous calibration. The power consumed in this approach is only for compensation clock phase generation. There is no charge pump or additional circuitry. However this approach requires the calibration phase to essentially occur before every sensing cycle. As explained in Section 5, the DAZ SA can perform compensation every N number of cycles. High charge pump current can be used to increase N at a cost of higher dynamic and leakage power of the charge pump circuit. The power consumption of [9] is compared to that of DAZ SA with the offset calibration phase occurring every cycle (cycle period = 1 µs) and every 200µs with a controllable offset phase. The controllable offset phase logic is employed to force calibration every cycle at the beginning. The logic then enables calibration every N cycles once the voltage on Cp settled to its final value. The results are shown in Table 2. The settling time of both schemes is compared. The DAZ SA with controllable offset phase consumes the lowest power with a 12 µs settling time.

Table 2. Power consumption of dynamic offset compensation and auto-zeroing circuit.

Offset Compensation Scheme	Power Consumption	Settling time
DAZ SA	6 nW	12 µs
Dynamic Compensation [6]	4 nW	0.5 µs
DAZ SA with controllable offset phase	2.5 nW	12 S

2.3.2 45nm Test Chip Measurements

A test chip fabricated in 45 nm technology is used to verify the scheme. The chip contains one regular SA array for benchmarking and another array that uses SAs with the auto-zeroing circuitry, with Cp equal to 32fF. The layout of the output capacitor consumes an area of 2.97 µm × 3.9 µm. The sense amplifier and the charge pump layout consumes an area of 4.39 µm × 5.29 µm. The supply voltage is set to 0.6V during measurements to mitigate the effect of noise on the measured results. The control signals are supplied to the auto-zeroing circuit at 1 MHz. Figure 2.24 shows the measured offset distribution of both banks. The positive terminal of the SAs is connected to 0.45 V. The negative terminal of the SAs is swept from 0.3 V to 0.6 V in increments of 5 mV. The SAs are enabled during each increment, and measurements of the SAs outputs are recorded. This information is then used to construct the SAs offset distribution in Figure 2.24. The measured mean (µ) and standard deviation (σ) of the uncompensated SA banks is −31 mV and 45 mV respectively. The auto-zeroing circuitry reduced the value of µ to −13mV and lowered σ to 9.3 mV. This indicates an 80% improvement in σ. The scheme limits the absolute value of the maximum offset to 50 mV.



(a) (b)

Figure 2.24. Measured offset voltage distribution (a) regular SA; (b) SA with auto-zeroing circuitry

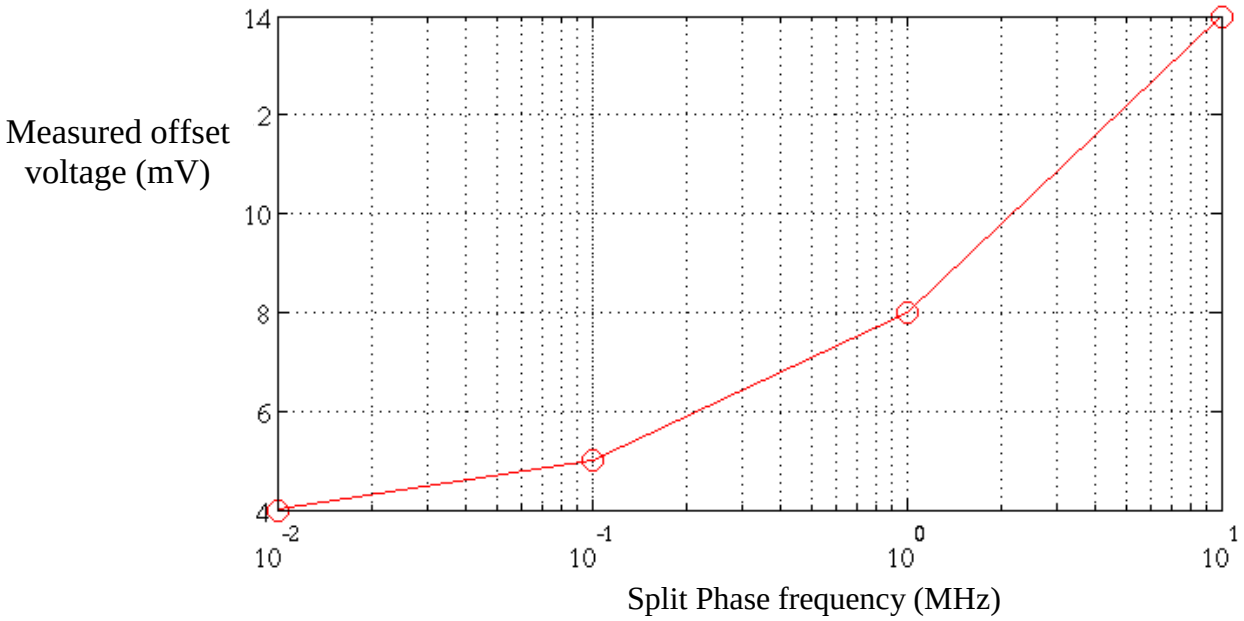


Figure 2.25. Measured offset voltage *vs.* split phase frequency

To verify the offset sensitivity to split phases, the offset of a sample DAZ SA is measured for different split phase frequencies. Figure 2.25 shows the offset voltage values for different split phase frequency.

2.3.3 Conclusion

We presented a circuit that is capable of improving sense-amp offset to near zero, which is valuable for sub-threshold operation due to the heightened effect of mismatch. Simulations of the design (0.5 V, 1 MHz) show a compensated offset voltage of 1mV, settling time of 37 μs, and total power consumption of 12 nW. Measurements from a test chip fabricated in 45 nm technology showed the circuit's ability to improve σ of the offset voltage by 80% and limited the absolute maximum value of the offset voltage to 50 mV using a 1 MHz split phase frequency and 32fF output capacitance. Using the circuit in a 16 kB SRAM showed a reduction in the total energy and delay of 10% and 15% respectively. The trade-off between the sense amp compensated offset and power consumption is demonstrated. This makes the circuit able to provide the offset/power values that can generate the optimal SRAM design.

3. A DIGITAL DYNAMIC WRITE MARGIN SENSOR FOR LOW POWER READ/WRITE OPERATIONS IN 28NM SRAM

3.1 Motivation and Background

The conventional guard band design approach increases the SRAM Wordline (WL) pulse duration to operate successfully in all the process, voltage and temperature (PVT) corners. This can significantly increase the dynamic energy. This work presents a digital circuit that is able to track and control the WL pulse duration of the SRAM memory across PVT variations, to minimize the dynamic energy while maintaining robust operations. A big portion of this chapter is from (4) in the publication section. The circuit is applied on a 78kbit SRAM. The results are compared to the worst case margin approach and show a maximum write energy savings of 45% and 49% relative to margining voltage/temperature (VT) and process variations, respectively.

The static noise margin has been traditionally used for SRAM cell stability characterization. However, recently it has been shown that the dynamic noise margin that utilizes the critical WL pulse width or TCRIT to estimate SRAM cell stability is more precise [10][11] because static read margin overestimates failures and static write margin underestimates failures. TCRIT is defined as the minimum WL pulse duration required to successfully write to the memory. As technology scales and the supply voltage is lowered, variability increases. This causes an SRAM cell to have a wider distribution of Wordline (WL) pulse duration required to perform successful read and write operations. The conventional worst case margin design methodology increases the WL voltage or pulse duration to operate successfully in all PVT corners. This can lead to operating the memory at much dynamic energy than what the memory requires for successful read/write operations [12]. This becomes more problematic with technology scaling due to the heightened variability effect [13] which can be a big problem in high performance System on Chips (SOCs) and multi-core processors where the SRAM memory contributes significantly to the total power

consumption. Increasing the WL pulse duration increases the Read and Write energies due to the increased Pre-charge, WL driver, and Bit-line (BL) driver energies. Figure 3.1(a) shows that longer WL pulses results in larger differential voltage between BL and BLB which increases the Pre-charge energy during the Read operation. In Figure 3.1(a), increasing the WL pulse from W1 to W2 results in an energy increase of $C_{BL}V_{DD*}(\Delta V)$ where C_{BL} is the capacitance of the BL, V_{DD} is the supply voltage and ΔV is the difference between the developed differential voltages on BL/BLB as shown in Figure 3.1(a). Similarly Figure 3.1(b) shows that longer WL pulses results in an increased developed voltage on the WL, and hence an increased WL driver energy during both the Read and the Write operations. In Figure 3.1(b), increasing the WL pulse from W1 to W2 results in an energy increase of $C_{WL}V_{DD*}(\Delta V)$ where C_{WL} is the capacitance of the WL. Figure 3.2 shows the distribution of the WL pulse duration required to successfully perform robust read/write operations to a 78kbit SRAM array across global and local process in 28nm technology node. The array has 256 rows and 312 columns. The read and write energy that corresponds to the WL pulse duration range is shown in Figure 3.3.

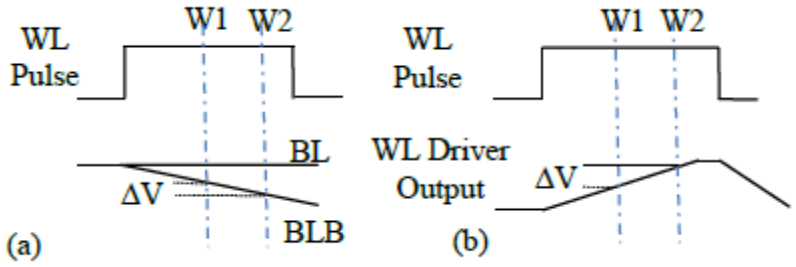


Figure 3.1. (a) Longer WL pulse increases the Pre-charge energy (b) Longer WL pulse increases the WL driver energy.

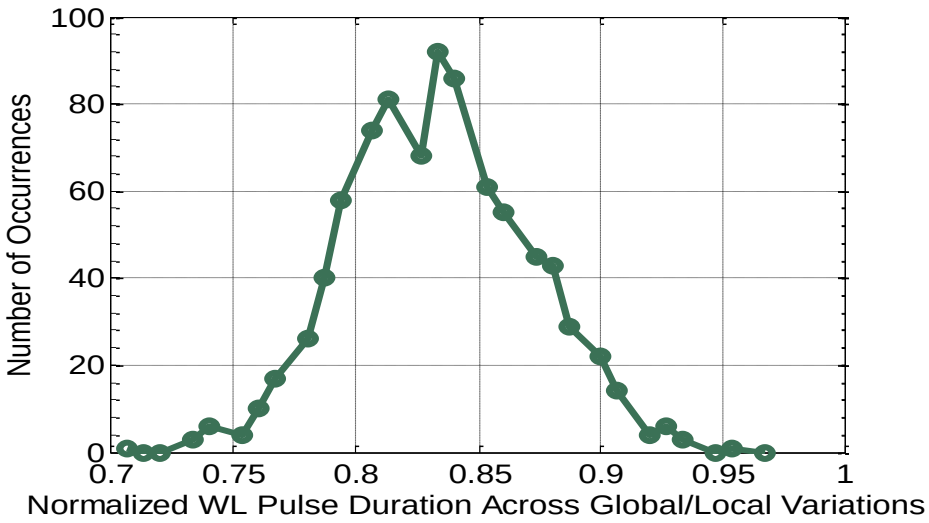


Figure 3.2 Distribution of the minimum WL Pulse in a 78kbit SRAM

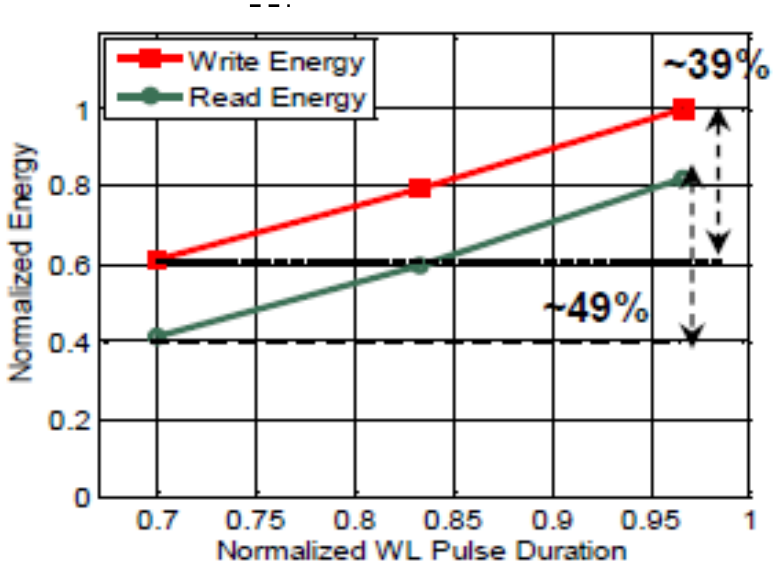


Figure 3.3 Total Energy versus the WL pulse

The figure indicates energy savings of up to 39% and 49% can be achieved in the write and read energies, respectively by adjusting the WL pulse duration with the global and local process variations. Usually worst case margin design approach guards for ±10% variation in the supply voltage and a 0-100 C temperature operation range. Figure 3.4 shows the WL pulse required to perform robust operations across the aforementioned voltage and temperature range at the typical process corner for the 78kbit SRAM array and

the corresponding dynamic energy. As shown, energy savings of up to 34% and 45% can be achieved in the write and read energies, respectively by adjusting the WL pulse with the supply voltage and temperature variations at the typical process corner.

Hence, significant energy savings can be achieved by adjusting the WL pulse duration across PVT variations. Utilizing adaptive circuits that control the memory to operate at the minimum dynamic energy without incurring high power and area overhead is challenging [13][14]. Prior works [14]-[19] have addressed this issue. In [17]-[19], replica rows/columns were used to track the memory margins across PVT. Replica rows/columns use extra row or column of the memory to track the margins, which have high area overhead. The energy overhead is also high since the replica rows/columns are used anytime the memory is accessed. In addition they can precisely track the global variations' effect on the margins not the local variations; hence the memory still operates at higher dynamic energy than the minimum achievable energy. In [14][15], analog sensors are used to control the WL voltage across PVT. Analog sensors are less scalable than their digital counterparts, which make them less attractive for scaled technologies nodes (32nm and beyond). In [16], a Built In Self-Test (BIST) was used to tune the WL pulse duration for minimum dynamic energy after powering up the memory. However, the proposed scheme failed to track the changes in that WL pulse across voltage and temperature (VT) variations. The WL pulse duration was overdesigned to account for VT variations, which can significantly increase the dynamic energy as shown earlier in this work. Hence, with technology scaling, the SRAM will no longer be able to fully realize the power/performance benefits of scaling unless adaptive circuits that can operate in scaled technologies without much energy and area overhead are utilized to improve the design margin while satisfying the power and performance requirements. In this chapter, we present a digital sensor that is calibrated once at the beginning of the memory operation and then used throughout the operation of the memory to adaptively tune the WL pulse duration across PVT variations to minimize the dynamic energy

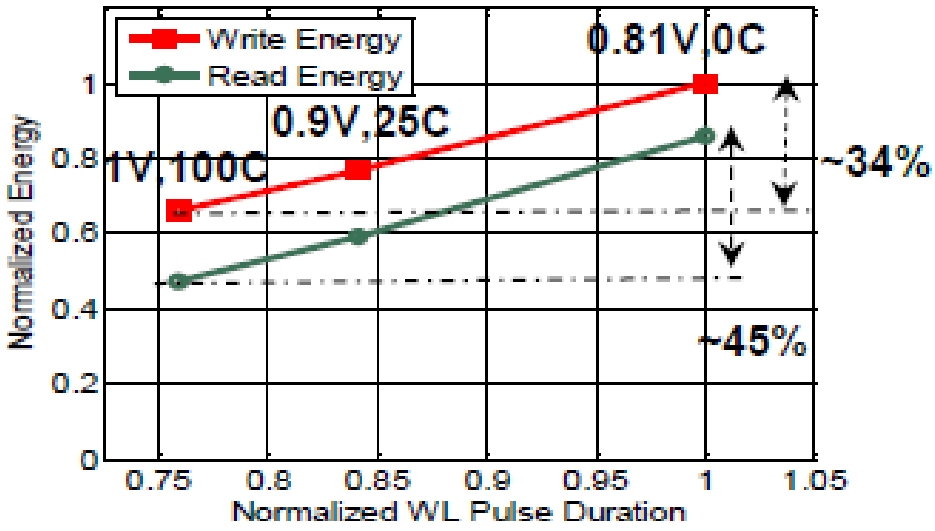


Figure 3.4. Total energy versus voltage and temperature at the typical process corner

while maintaining robust operations. The circuit is digital and hence can be adapted in scaled technologies. The contribution of this work is the sensor and a low power WL control scheme that utilizes the sensor. The rest of this paper is organized as follows. Section 3.2 and Section 3.3 discuss the sensor circuit and the calibration step respectively. Section 3.4 describes the sensor precision and overhead. Section 3.5 demonstrates the utilization of the sensor in a low power WL pulse control scheme.

3.2 Sensor's Circuit

In scaled technologies it has been shown that write failures are higher than the read failure [20][21]. In other words, TCRIT is wider than the read WL pulse duration. Therefore we can use TCRIT to successfully perform both the read and write operations. This work utilizes a sensor that generates TCRIT for different PVT conditions. In other words, it tracks TCRIT across PVT corners. Figure 3.5 shows a block diagram of a WL control scheme that utilizes the sensor to track and control the WL pulse duration across PVT conditions. In this work, we will refer to the bitcell with the longest TCRIT as the worst case bitcell. The sensor is calibrated to generate TCRIT of the worst case bit-cell. When a coarse change in the temperature

or voltage occurs, a trigger signal is generated. The sensor then generates the new TCRIT and sends it to digital logic that digitizes and stores TCRIT and uses it to generate the WL pulse for the memory during normal operation. It is worth mentioning that the worst case bitcell doesn't change with VT variations i.e. when the temperature or voltage changes, TCRIT of the worst case bitcell changes but it remains the longest TCRIT among all bitcells of the memory.

The sensor circuit is shown in Figure 3.6. The waveforms of the circuit are shown in Figure 3.7. The sensor consists of a 6T bitcell and circuits that measure TCRIT of this bitcell. The bitcell is initially calibrated to have the same TCRIT as the worst case bitcell of the memory as will be shown in section 3 that discusses the calibration procedure. The calibration ensures that TCRIT of the sensor's bitcell follows that of the worst case bitcell with voltage and temperature variations. The sensor measures TCRIT of its bitcell as follows the bitcell initially stores 0. Once the sensor is enabled, a Trigger signal is generated that asserts the sensor's Wordline to write 1 to the bitcell. Q and QB starts charging and discharging, respectively. Once Q crosses QB, a comparator asserts a Reset signal. The Reset signal turns off the Wordline indicating the end of the write operation and asserts PCH and PCHB to 0 back again to the bitcell to be ready for the next operation. The width of the WL pulse is the measured TCRIT.

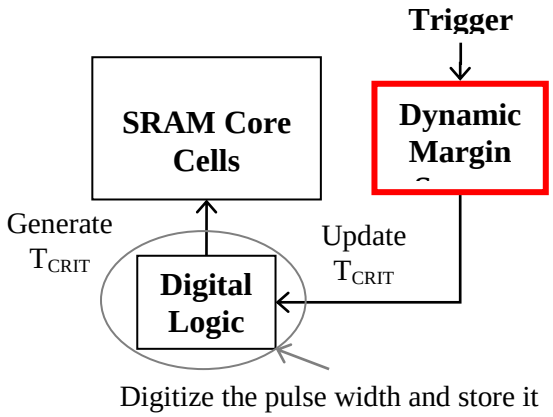


Figure 3.5 Dynamic Margin Sensor in a WL control scheme

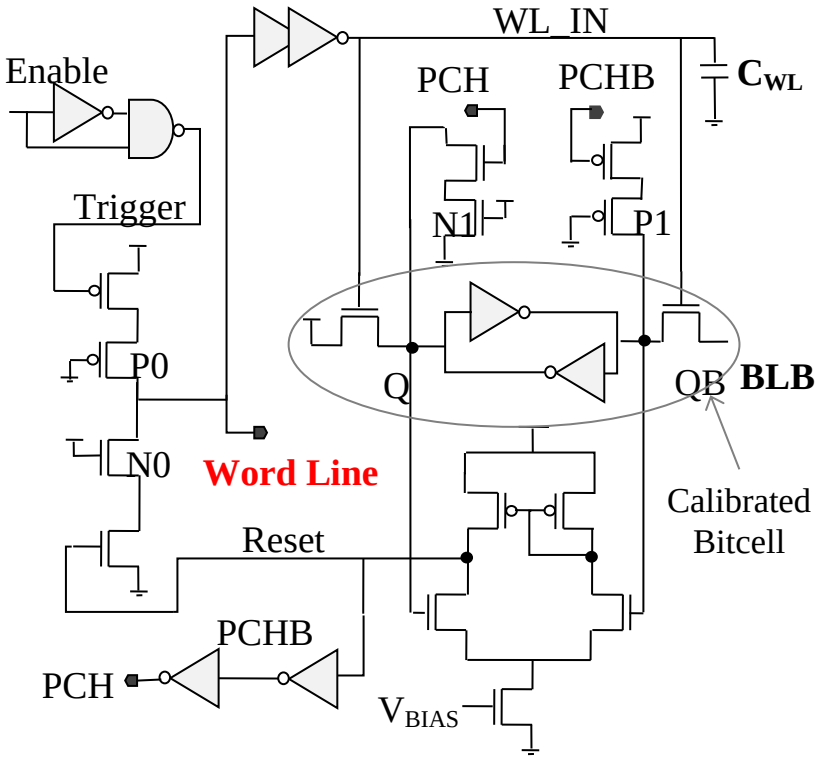


Figure 3.6 Dynamic Write Margin Sensor.

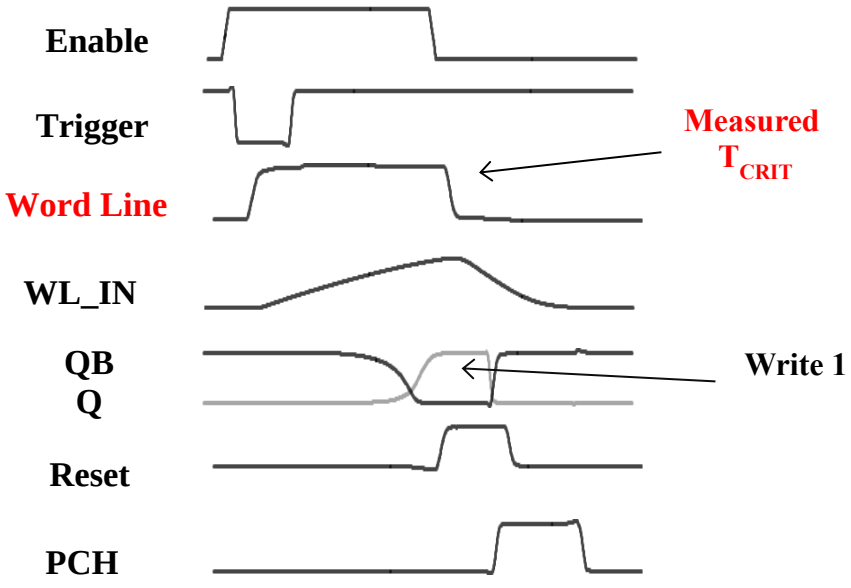


Figure 3.7 Write Margin Sensor Waveforms.

Transistors P0 and N0 are used to decouple the switching effect from the Wordline signal. Similarly, transistors P1 and N1 are used to decouple the switching effect from the Q and QB nodes. The pulse generator circuit that asserts the Trigger signal is designed to ensure long output pulse enough to turn on the Wordline across process corners. The capacitor CWL is attached to the bit-cell and its value is equivalent to the WL capacitance of one row in the SRAM memory (CWL = number of bitcells/row * 2Cgate), where Cgate is the gate capacitance of the access transistors. CWL is utilized to have similar initial dynamic margin of the sensor as that of the SRAM bitcell. VBLB is the voltage set by the digital controlled voltage divider during the calibration phase.

3.3 Sensor's Calibration

The calibration of the sensor's bitcell to the worst case bitcell of the memory is done through digitally setting the value of BLB of the sensor's bitcell. A Built-In Self-Test (BIST) is needed at first to measure TCRIT of the worst case bitcell. BIST are commonly used in modern memories [7] and can be easily integrated to this work. Control logic is then used to digitally tune BLB to adjust TCRIT of the sensor's bitcell to be the same as that of the worst case bitcell. BLB is tuned using the voltage divider circuit shown in Figure 3.8. Using the voltage divider for tuning BLB is critical to keep the bitcell tuned to the worst case with the supply variations. If a reference bias is used instead to set BLB, for instance, an increase in the supply voltage will increase the sensor's bitcell dynamic margin more than that of the worst case bitcell, and hence will not be kept tuned to the worst case bitcell of the memory. Another challenging point is that local process variations can set TCRIT of the sensor's bitcell to be higher or lower than that of that of the worst case memory bitcell. A negative voltage might be needed to bias BLB in this case. Most SOC don't have a negative supply rail. To avoid this, we upsize the right access transistor of the sensor bitcell to 3X the access transistor of the SRAM bitcell. This ensures that the sensor's bitcell will initially have smaller TCRIT than

that of the worst case SRAM bitcell and hence only a positive bias of BLB is required. Figure 3.9a shows the initial distribution of TCRIT of the sensor's bitcell and the core SRAM bitcell across global and local process variations. Figure 3.9b shows the distribution after calibration using the voltage divider circuit in Figure 3.8. The choice of the value of R in the voltage divider controls the area/energy overhead tradeoff. Large value of R decreases the energy overhead on the cost of higher area overhead. The circuit energy and area overhead will be included in the control scheme overhead discussion in Section 5. Figure 3.9 shows TCRIT of the sensor's bitcell versus the BLB/VDD ratio set by the voltage divider. As shown the maximum error in TCRIT due to calibration is ~3.5% TCRIT.

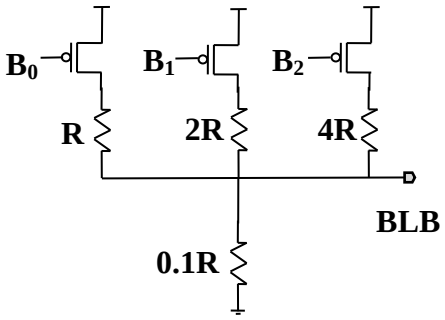
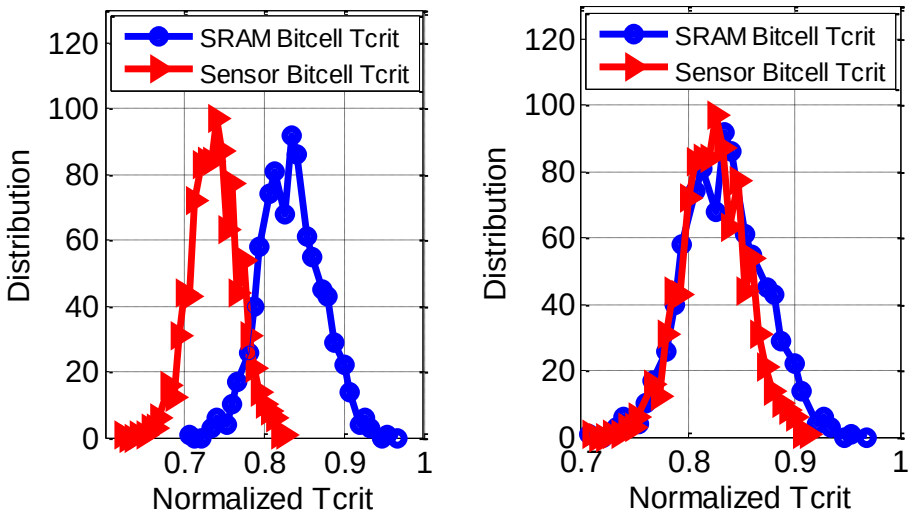


Figure 3.8 Digitally Controlled Voltage Divider for calibrating the sensor's bitcell.



(a) (b)

Figure 3.9 (a). Initial Distribution of T_{CRIT} of the sensor bitcell and the SRAM bitcells across global and local process variations. (b) T_{CRIT} distribution after calibration.

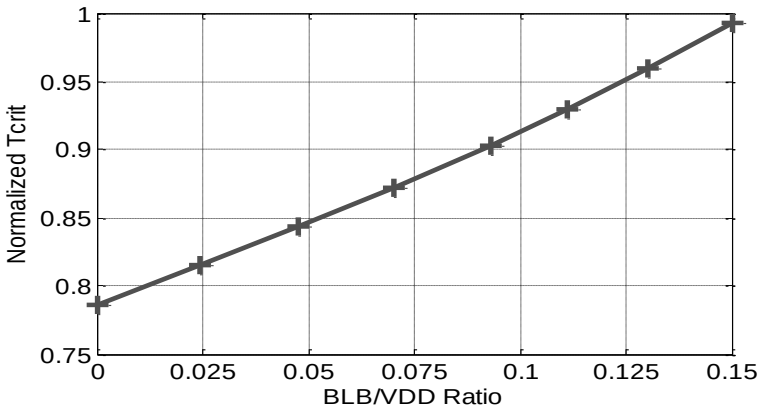


Figure 3.10. Tcrit changes with BLB/VDD ratio.

3.4 Precision and Overhead

We studied the precision of the sensor in controlling TCRIT .The sensor shown in Figure 3.6 is calibrated to the worst case bitcell of the 78kbit memory at the typical operating point (0.9V, 25C). Figure 3.11 shows TCRIT generated by the sensor and TCRIT of the worst case bitcell at 0.81V,0C and 1V,100C. As shown, the sensor precisely adjusts the WL pulse duration needed to write to the worst case bitcell, with a slight deviation of 2% at 0.81V, 0C. We studied the process variation effects on the sensor. The analysis indicated that the sensor can be tuned to the worst case bitcell in the memory independent of process variations as shown in Figure 3.12 that shows the correlation of TCRIT measured by the sensor to the worst case bitcell TCRIT at 0.8V for 1000 Monte Carlo (MC) points after the calibration step is first done at the typical operating voltage 0.9V. Only few points of the MC results are shown for clarity. The energy overhead of the margin sensor is 0.5X the energy of a single write operation to the 78kbit memory. However the sensor only operates when a coarse change in the

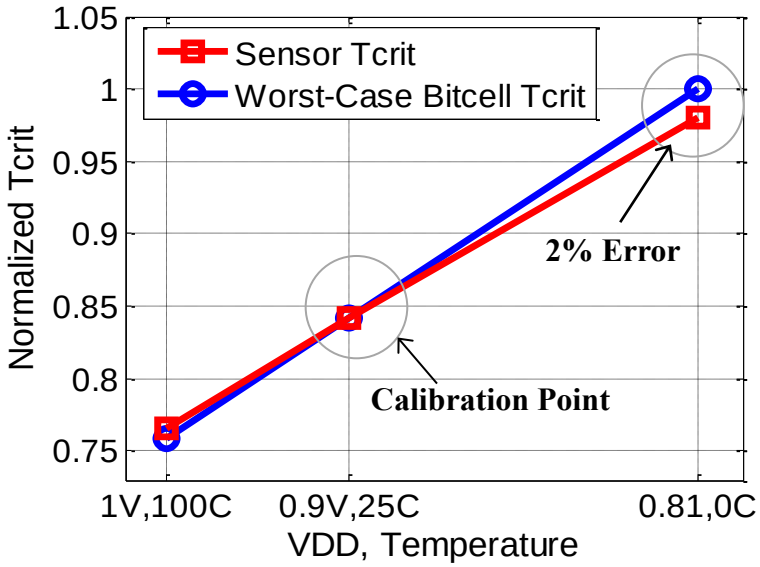
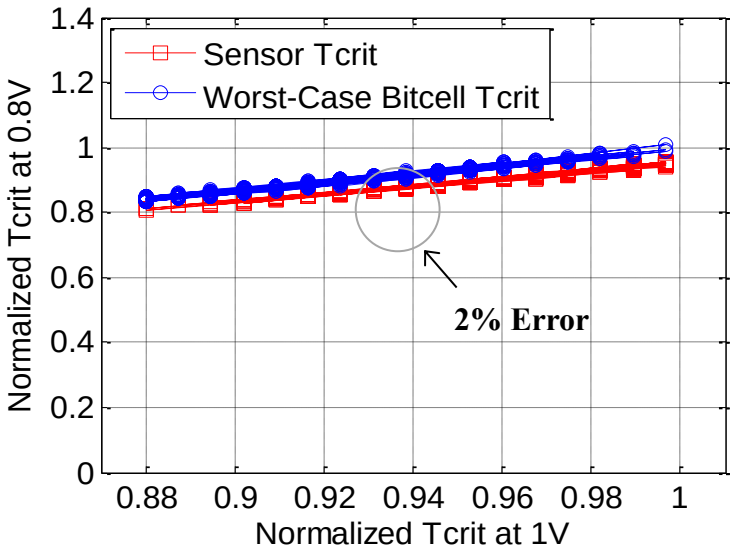


Figure 3.11. T_{CRIT} measured by the sensor follows T_{CRIT} of the worst case bitcell in the memory.



1000 Monte Carlo (MC) to verify the precision of the sensor across process variations

Figure 3.12. Correlation of T_{CRIT} measured by the sensor to T_{CRIT} of the worst case bitcell across global/local variations for 1000 Monte Carlo (MC) points

voltage/temperature is sensed. In modern processors, where the frequency of operations is in the GHz range, the rate of coarse change of temperature or voltage is much smaller than the cycle time and hence the effective energy overhead of the sensor is amortized. For instance, if the rate of voltage/temperature coarse change is assumed to be 10X of the cycle time which is an extreme upper limit, the effective energy

overhead of the sensor will be 5% the energy of the memory write operation. The area overhead is 7X the area of single bit-cell or ~0.0087% the area of the core SRAM array. The area overhead can be minimized if the Wordline driver and the Wordline capacitor are shared between the sensor and the memory array.

3.5 Wordline Control Scheme

Figure 3.13 shows one realization of the sensor in a Wordline (WL) pulse control scheme. The calibration flow is shown in dotted lines, while the operation flow is shown in solid lines. Figure 3.14 shows a flow chart of the scheme. The control scheme uses a BIST circuit to measure the minimum WL pulse required to successfully write to the worst case bitcell in the memory. The measured pulse is then digitized and stored in the WL register. Afterwards, calibration logic uses the WL register to calibrate the sensor to enable the sensor to track TCRIT of the worst case bitcell across VT variations. This calibration step is done at powering up the memory and repeated to compensate for the aging effect. Once the temperature and/or voltage changes, the margin sensor generates the new WL pulse and sends it to a WL quantizer circuit that digitizes the pulse and stores it in the WL digital register. The WL register is then used throughout the operation of the memory to generate the read/write WL pulses. External voltage and temperature sensors are used to sense a coarse change in voltage and temperature (VT) and trigger the margin sensor to generate TCRIT for this instantaneous VT condition. VT sensors are not discussed in this work but they are extensively used in advanced processors [22]-[24] and can be easily integrated to this work. It is worth mentioning that the precision of adjusting the WL pulse depends solely on the margin sensor accuracy, not the VT sensors which usually have an analog implementation and might suffer from process variations. The VT sensor role is to only trigger the margin sensor. The WL quantization circuit is shown in Figure 3.15. The circuit consists of 16 delay stages, a thermometer-to-binary encoder (TBE), and 4 output registers. In [25] [26], the conventional delay line approach for time to digital conversion was used, which is similar to this work. In [25], the circuit was used to characterize SRAM dynamic margins.

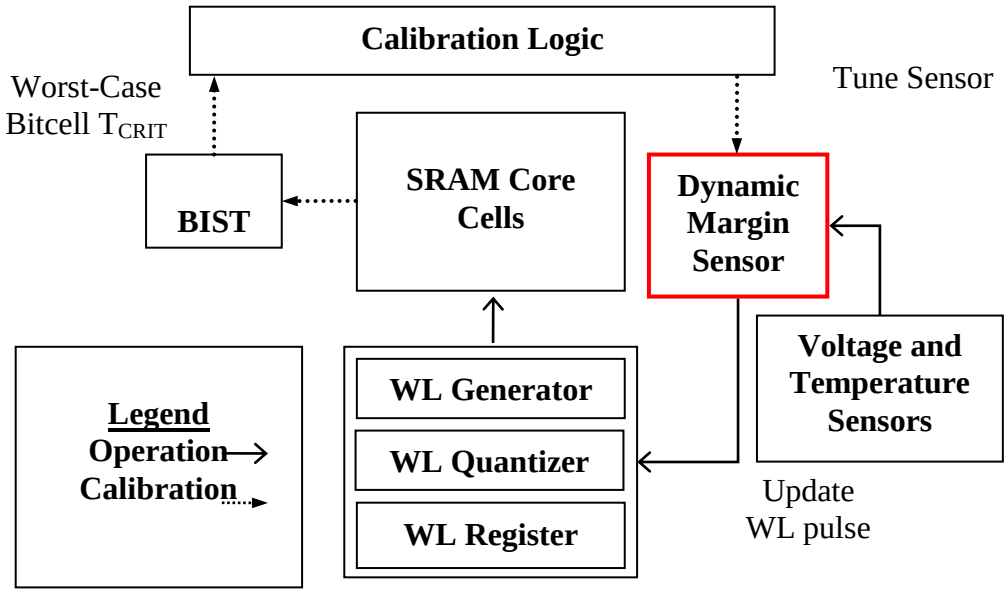


Figure 3.13. Wordline control scheme

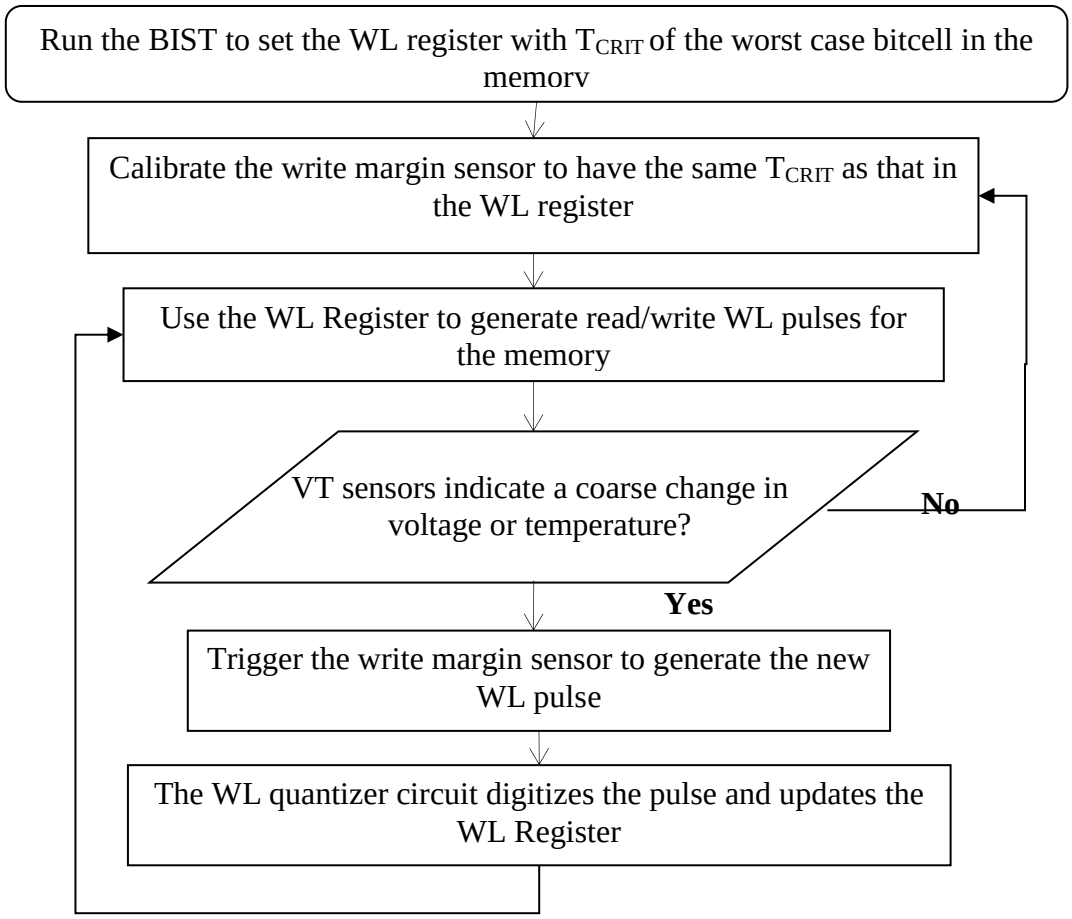


Figure 3.14. Margin control scheme

The delay of all the delay line stages was set to the minimum sized delay in both approaches. Unlike the aforementioned approaches, we statistically designed the first stage to have a delay equal to the minimum WL pulse duration across all PVT corners and the rest of the delay line stages to have the minimum sized delay. This minimizes the quantization energy by ~70% through only quantizing the range of the WL pulse instead of the WL pulse itself. VA is the bias voltage of the first stage. It is set during the design time statistically to tune the first stage delay to the memory minimum WL pulse duration. In addition, we sensed the delay nodes using pass transistors P0, P1, P2 as shown in Figure 3.15. This decreases the capacitive loading on the delay line stages, and minimizes the quantization energy by ~40% compared to the conventional TDC approach [25][26] that loads the delay line nodes with latch circuits. The error resulting from quantizing the WL pulse is ~6% of the minimum WL pulse and can be minimized by increasing the number of stages of the delay line. The WL generator circuit is shown in Figure 3.16. A binary to thermometer encoder converts the 4 binary registers to a thermometer code T0, T1, till T15. The delay line nodes are initially set to zero. When the clock signal (clk) changes from 0 to 1, a pulsed trigger signal is generated. The trigger signal is applied to transistor P0 to initially set the WL voltage. A pulse then propagates in the delay line and set the delay line nodes to '1'. When the WL pulse timing is met, the pull down transistors N0, N1, N15 reset the WL pulse. Figure 3.17 shows the waveforms of the WL generator circuit equivalent to a WL code of "0101". The waveform of only the first 8 delay line nodes were shown for clarity. The energy overhead of the WL quantizer and generator circuits is 2% and 1.2% the energy of a single write operation. The effective energy overhead of the quantizer is less than this value since it only operates when a coarse change in the voltage and/or temperature occurred. The area overhead of the WL quantizer and generator circuits is 0.02% and 0.01% of the SRAM core array respectively. Figure 3.18 shows the flowchart of the BIST margin characterization scheme. The circuit first sets the WL pulse width to the min WL pulse of the design (TWL_MIN). This is the smallest critical WL pulse duration of the design across all PVT corners. This is also the delay of the first stage of the WL generator. The BIST then scans

each row in the memory and performs a write followed by read operation to check if a successful write occurs. If the write operation didn't finish successfully the circuit increment the WL pulse duration by 1 delay unit, store the new WL pulse in the WL register and scans the next row. The min WL pulse is then stored in the WL register after scanning all the rows of the memory and used to calibrate the sensor. The total energy overhead of the control scheme is ~6.5% the energy of a single write operation to the memory including the energy of the margin sensor, WL quantization and generation circuits, calibration registers and the voltage divider circuit. The overall area overhead of the control scheme is ~0.12% the area of the SRAM core array including the aforementioned circuits. The overhead of the BIST and the VT sensors are not included since they exist in modern processors [22]-[24] for various purposes, their output are available and can be easily integrated to this work.

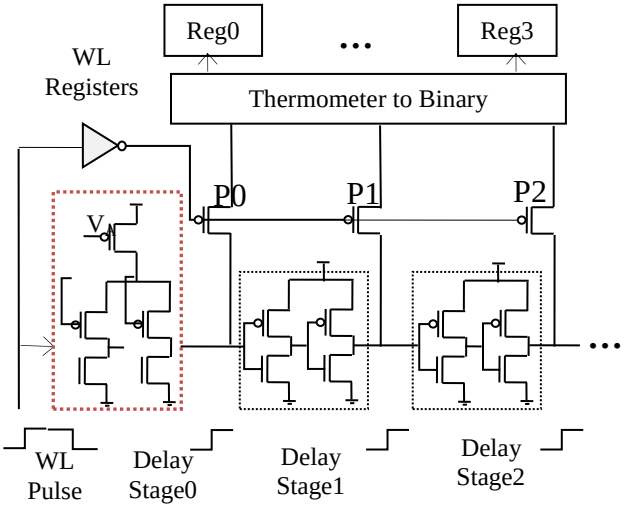


Figure 3.15. Wordline quantizer

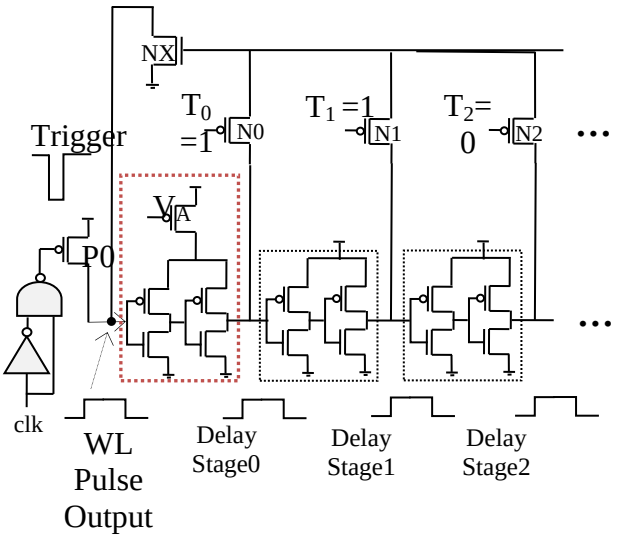


Figure 3.16. Wordline generator

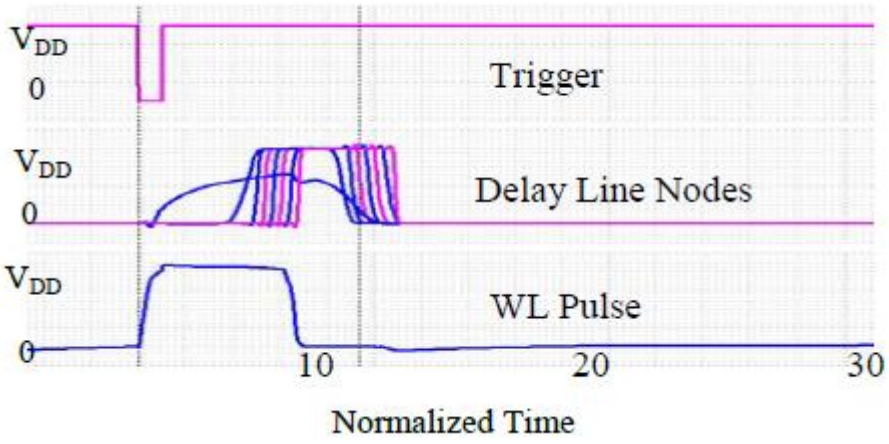


Figure 3.17. Wordline generator waveforms

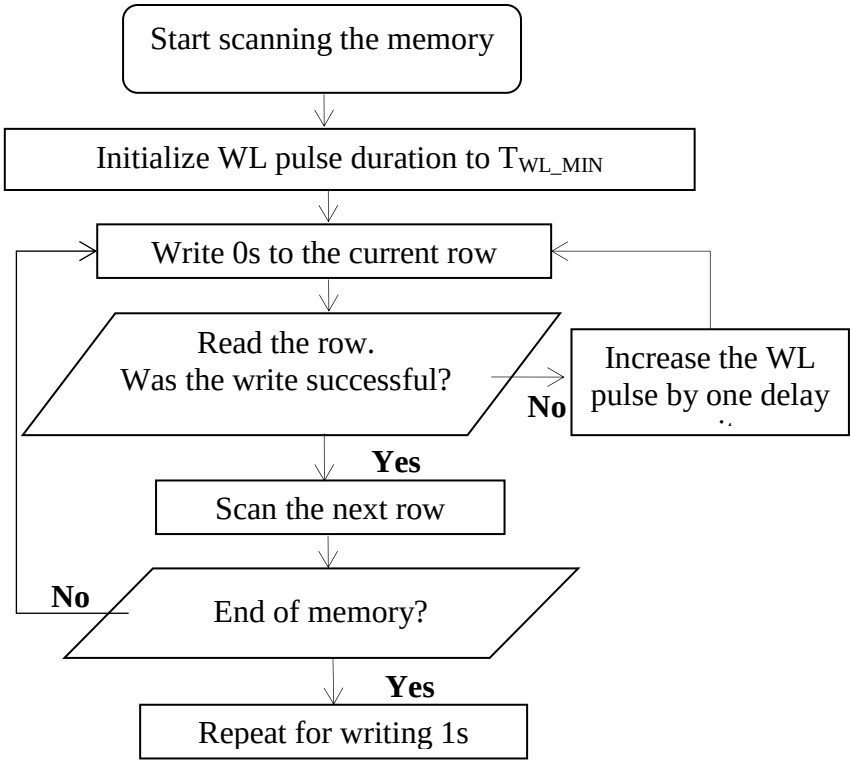


Figure 3.18. BIST characterization scheme

3.6 Conclusion

In this chapter, we presented a dynamic margin sensor that is capable of tracking the critical WL pulse duration (TCRIT) of a memory across PVT variations. The sensor is digital and hence suitable for scaled technologies. The sensor was used in a time control scheme that utilizes WL quantization and generation circuits. The scheme was applied on a 78kbit SRAM memory with 256 rows and 312 columns in 28nm commercial technology node. The results are compared to the worst case margin approach and show a maximum write energy savings of 45% and 49% relative to margining voltage/temperature (VT) and process variations, respectively. The total energy overhead of the control scheme is ~6.5% the energy of a single write operation to the memory. The total area overhead of the scheme is ~0.12% the area of the array.

4. A HYBRID OPTIMIZATION SCHEME FOR CIRCUIT/ARCHITECTURE CO-DESIGN OF COMPLETE SRAM MACROS

4.1 Motivation and Background

Transistor scaling and the accompanying advent of bit-cell topologies, periphery logic, circuit techniques, assist methods, and architectures have expanded the SRAM design space and complicated SRAM design. Various optimization techniques have been proposed, but they tend to address only individual components of a memory. In this chapter, we present a hybrid optimization methodology for SRAM circuit-architecture co-design that combines convex optimization (CO) to explore the power-performance trade-off in lower level blocks with macro level optimization. A big portion of this chapter is from (5)(6) in the publication section. We combine the optimal circuit design points from CO at the architecture level to explore the power-performance trade-off for the complete SRAM macro with a simulated annealing (SA) algorithm. We apply this methodology to explore the power-performance tradeoff in 4kB, 16kB, and 1MB SRAMs in a 90nm technology. Circuit level options including offset compensated sense amplifiers (SAs), write assist techniques, and varying periphery circuits sizing were included in the analysis, which shows that the hybrid optimization approach generates better designs than a top level approach that omits circuit level optimizations. SRAM in the future will host billions of transistors to implement heterogeneous collection of functional blocks. The growth in available transistors has outpaced the CAD tools ability to design them. In addition, technology scaling leads to smaller transistors, more variation, and reduced operating margins. This degrades the bitcell performance and necessitates a variety of compensation schemes that are both architectural and circuits based. For instance, circuit techniques such as read/write periphery assist features [27]-[31], single ended 8T cells [32]-[34] and short bitlines [35] have been proposed to ensure robust operations. All of these changes complicate the process of designing SRAM macros and lead to a substantial increase in the size and complexity of the design space.

Several approaches proposed optimization methodologies before to tackle the increasing challenge of designing SRAMs [36]-[39], but they tend to address only individual components of the memory. In [36] a multi-objective approach was proposed to optimize the SRAM bit-cell and the periphery circuits. However, each component were optimized separately. In [37] analytical models for each of the SRAM components were used to study the effect of different optimization knobs on the total energy and delay. In [38][39] geometric programming was used to optimize the power and performance of the SRAM bit cell, and the row decoder respectively. Each of the aforementioned approaches mainly targets the design of one or more component in the SRAM. Whilst all of the methods provide useful results, a modification in any of the memory component can impact the optimal design at both the architectural level and circuit level. Further, designs targeting different constraints (e.g., high speed vs. low power) may optimally use dramatically different circuit choices, and those circuit choices affect the architectural decisions. In [40][41] the authors presented a tool that uses exhaustive search for SRAM macro optimization. However, exhaustive search is not efficient in rapidly exploring and optimizing SRAM design space.

Thus, to fully explore the design space, there is a need for a methodology and tool to consider the effect of each memory components on the overall system design and co-optimize all the components for different architectures of the SRAM macro. This also helps to perform an efficient full re-design and optimization each time a new circuit is added, support assist features and makes the design easily portable across technology nodes. The optimization methodology is incorporated in the framework proposed in [42]. In [42], the authors proposed a tool, ViPro, that enhances the ability of a knowledgeable human designer to generate base case memories in a new technology, estimate macro level metrics before sub-components are completed or designed, compute the effects of a low level circuit change on the overall SRAM macro, and rapidly evaluate varying prototypes of arrays using new circuits or new cell technologies. The tool improves the memory design process across a diverse market space. The authors used brute force to perform the

macro optimization. The option to perform transistor sizing optimization was not included in the tool flow. In this work, we incorporated our hybrid optimization scheme to the tool and show the capabilities of the tool to efficiently generate optimized macros design with better metrics (energy/delay) than a top level approach that omits circuit level optimizations. In [43], an approach was presented to perform circuit architecture co-design of superscalar processors. Each underlying component of a superscalar processor was characterized to generate Energy/Delay (ED) Pareto-optimal curves. This information was then used to perform the architecture design. Characterizing the underlying circuit components for energy and delay tradeoff, and passing this information to perform circuit-architecture optimization has been proved to be efficient [43]. In addition, in SRAM design, the existence of multiple repeated instances in the SRAM macro (bit-cells, periphery logic) makes the concept of utilizing the optimization results of each instance attractive. We similarly characterize each component in the SRAM macro to generate its ED Pareto-optimal designs, and combine this information to generate the macro optimal ED design points. Furthermore, we use the simulated annealing algorithm to efficiently locate an optimal macro design under energy/delay constraints if the full macro Pareto-optimal curve is not needed.

4.2 Optimization Scheme

In this section, we describe the optimization scheme that we added to the tool, to efficiently perform the following functions

+ Efficiently locate an optimal virtual prototype SRAM macro that meets input defined energy and delay constraints.

+ Generate the Pareto-optimal energy and delay designs of an SRAM macro.

The optimization scheme can perform these functions given various input defined circuit design options that defines the design space. The design options include various periphery circuit topologies (i.e. compensated sense amplifier, Latch sense amplifier), assist features (Wordline boosting, negative bitline) and the option to include transistor sizing optimization for each circuit.

The problem of SRAM optimization can be decomposed into system level and circuit level optimization. We characterize each component in the SRAM macro to generate its ED Pareto-optimal curve using convex optimization, and then utilize the simulated annealing algorithm to efficiently locate an optimal macro design under energy/delay constraints. The following sub-sections describe the scheme in details.

Circuit Optimization

There exist a number of tools that can optimize a circuit along one or more of these circuit parameters to produce power-performance tradeoff curves. We utilized the tool, SCOT [44]-[47] in this work, to perform convex optimization to find the optimal circuit sizing and generate the Pareto-optimal design points of each circuit. By trying a few discrete circuit topologies and circuit styles, one can construct the overall tradeoff space for a circuit. SCOT uses convex optimization algorithms which have been proven efficient in solving transistor sizing optimization problems [45]. Transistor sizing optimization problem can be mapped to a geometric program (GP) [45]; a GP consists of a set of posynomial constraints, where each posynomial is

the sum of any number of positive monomial terms. GPs can be mapped to a convex space through a logarithmic transformation and convex optimizers can then quickly search the space to reliably find the global optimum. SCOT models the energy and delay of the underlying circuits of SRAM macros components a GP, thus enabling efficient optimization. For more information about SCOT the reader can refer to [44][45].

Architecture/Circuit Co-Design

First, circuit characterization is applied to the underlying macro circuits. The characterization step is repeated for different SRAM circuits' topologies. After the characterization step is performed, the ED Pareto-optimal design points of each underlying component in the SRAM macro will be available. We then perform the design exploration as follows. For each structure of the macro (number of banks, number of rows, number of words per row) we construct an ED Pareto-curve for the macro. The process is repeated for all structures of the memory to generate the circuit/architecture optimal Pareto-curve of the macro at the end. *Figure 4.1* illustrates the optimization scheme. *Figure 4.2* shows the flowchart of the scheme.

If the optimal Pareto-curve of the macro is not required, and only a certain optimal macro design is required that satisfies input defined energy/delay constraints, to save time, we locate the optimal design as follows. The scheme uses the SA algorithm to perform circuit architecture co-optimization. The algorithm is illustrated in Algorithm1 section. The circuit characterization is first applied to the underlying macro circuits and repeated for different circuits' topologies. The information needed by the simulated annealing algorithm about the underlying macro circuits is only the ED Pareto-optimal design points calculated by SCOT. The algorithm constitutes a certain SRAM design by randomly selecting from each component one of the ED Pareto-optimal design points with random values of the architectural-level design knobs (number of banks, number of rows, number of words per row). The algorithm then evaluates the objective function as a weighted function of the total energy and delay of the design. According to the objective function value,

the algorithm will decide to either accept the current design or not. Then, the algorithm randomly constitutes another design point, by modifying a single parameter in the design. The parameter could be one of the following, the number of banks, number of rows, number of words per row, a different design point in the ED Pareto-curve of any of the underlying circuits (decoder, pre-charge devices, sense amplifier), or choosing a different topology of circuits (i.e. using offset compensated sense amplifier instead of Latch sense amplifier). The algorithm will evaluate the objective function again and decide whether or not to accept the design as described above.

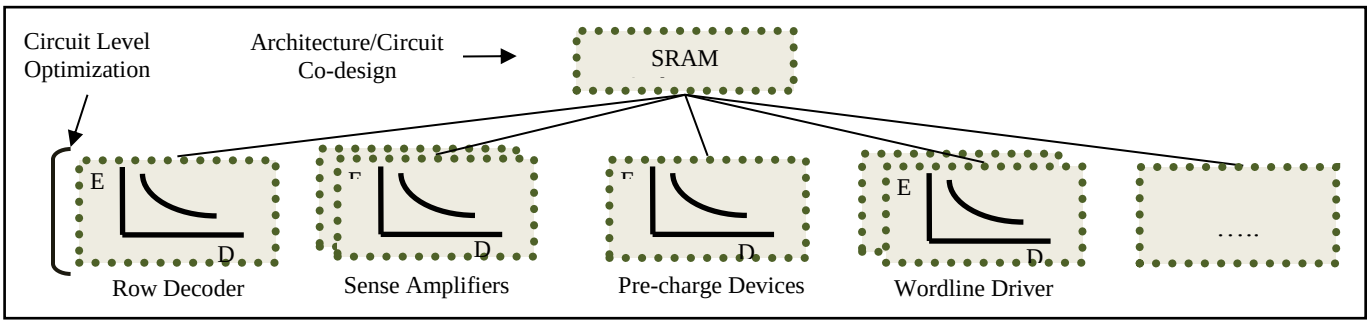


Figure 4.1. Demonstration of the optimization scheme

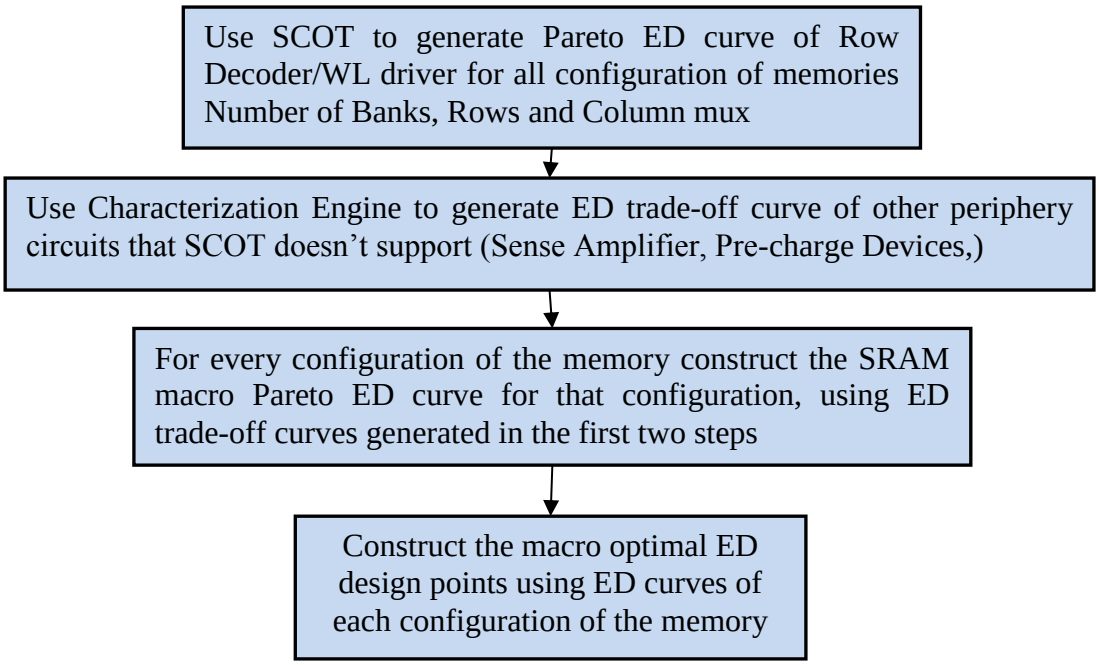


Figure 4.2. Flow chart of the optimization scheme

The algorithm will keep searching the design space using the methodology discussed till it meets the end criterion, or satisfied the design criterion (energy/delay constraints), whatever comes first.

Algorithm1 - Circuit Architecture Co-optimization

1: Calculate the pareto-optimal curve of each SRAM component.

2: Calculate initial Temperature through N Random Moves

3: WHILE Stopping Criterion IS NOT true DO

4: Randomly construct SRAM Component Knobs */*Rows, Words per Row, Banks, periphery circuit topologies, transistor sizing, assist methods)*/*

5: Evaluate Objective Functions (Energy, Delay)

6: IF Objective Function improved? THEN

7: Accept Design

8: ELSE

9: IF Prob(delta) > Random Value THEN

10: GO TO 6

12: ELSE

12: Reject Design

13: END IF

14: END WHILE

15: Return the optimal design

4.3 Optimization Framework

We modified the tool structure in [42] to incorporate the optimization scheme presented in section II. The new structure is shown in *Figure 4.3*. The implementation is done in C++ to easily interface with the optimization algorithm. The inputs to the framework are the optimization constraints of the objective functions, the design knobs (different SRAM components, assist methods, gate sizing, etc.) and the input specification (memory size, technology node, etc.). A circuit designer may consider potential circuit topologies (e.g. different styles of decoders, pre-decode, offset compensated sense amplifiers, word-line driver, different muxing schemes, etc.). The hierarchical memory model stores a virtual prototype of the SRAM macro. The characterization engine (CE) contains technology agnostic templates for characterizing the energy and delay of each underlying SRAM circuit of the macro as well as Monte Carlo (MC) templates for yield characterization. Yield characterization and optimization is added as a new feature to the tool but is not discussed in this work. The reader can refer to [42] for detailed information about the hierarchical model and the characterization engine. We included SCOT to the CE to generate the Pareto-optimal curve for the row decoder, the Wordline drivers and the BL drivers. We utilize the CE to investigate in the energy-delay trade-off of the other non-Complementary logic periphery circuits that SCOT doesn't support (Sense Amplifiers, Pre-charge devices). The optimization engine constitutes the adaptive simulated annealing algorithm discussed in the previous section.

Hierarchical Memory Model

The hierarchical memory model represents an SRAM macro design. The top level class is called SRAM, and it contains a handle to the input specification (supply voltage, memory capacity, word size, number of banks, and flags for using assist features, or various circuit topologies like offset compensated sense amplifiers.) These input specifications are inherited by each component class, which additionally contains

its own local parameters. For instance, the Wordline driver class has parameters such as WLBoost, to indicate where to simulate the Wordline driver circuit or the Wordline driver with a charge pump to post its output voltage. Also, the sense amp class has parameters such as offset, topology type, etc. which are specific to that component. Once the virtual prototype is defined, components with fully defined netlists are characterized by the CE, which returns the energy and delay of the component. Once all of the components have been defined, the model calculates the energy and delay of the SRAM macro and either reports back to the optimization engine or outputs the results to the user. The main advantage of the hierarchical model is that it is scalable and able to capture the interactions between the various components.

Characterization Engine (CE)

The CE contains a library of simulation templates which can be used for characterizing a technology (e.g. I-V curves, leakage current, inverter delay), characterizing the energy and delay of an SRAM component, The configuration and technology specific parameters is then passed to the CE by the memory model, before the simulation can be executed. After simulation, the data is sent back to the memory model and stored in the virtual prototype.

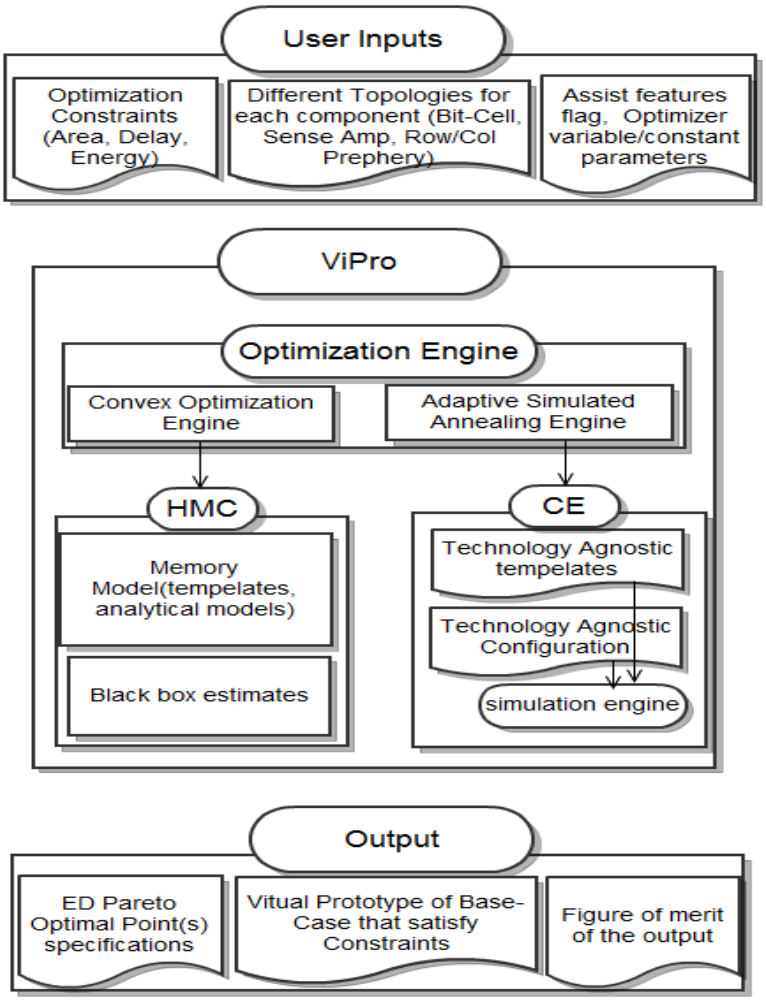


Figure 4.3. Optimization Framework.

The CE is modified from [42] by adding SCOT and other characterization test to generate the Pareto-curve of the underlying circuits of the SRAM macro.

Optimization Engine

The optimization engine contains the simulated annealing optimization algorithm. The energy, yield, and delay constraints are passed to the engine from the top level program. The optimization engine creates a hierarchical memory model and set the input specification defined by each iteration in the model. The model passes the energy/delay values at the end of the simulation to the engine to decide whether to accept the move or not, then the engine adjust the input specifications and set it in the hierarchical model till a design

that meets the specification is located or the end criterion achieved as discussed. The next section will discuss the results of running the optimization scheme on various sizes of SRAM macros in 90nm technology node and analyze the energy-performance trade-off of the results.

4.4 Simulation Results

In this section we will show the capabilities of scheme to perform architecture optimization for different sizes of SRAM macros and show that the hybrid optimization approach generates better designs than a top level approach that omits circuit level optimizations. Furthermore, we utilized ViPro to investigate in SRAM macros that use write assist circuits.

SRAM Macros Optimization

To demonstrate the benefits of the hybrid optimization we used ViPro to generate the Pareto-optimal ED design points of different sizes of SRAM macros for a fixed sizing of the periphery circuits (row decoder, WL driver, Pre-charge devices and Sense Amplifier). We then enable the hybrid optimization engine inside ViPro to re-generate the Pareto-optimal ED design with optimized periphery circuits and compared it to the fix periphery circuits' case. First to illustrate the scheme and its benefits, we demonstrate in Figure 4.4 the effect of exploring the ED trade-off of the row decoder/WL driver on the design space of a 16kB SRAM macro in 90nm technology node. As shown, exploring different design choices of the row decoder/WL drivers creates new optimal ED design points, and minimizes the min energy point by ~6%. Similarly, Figure 4.5 shows the effect of exploring the ED trade-off of the sense amplifier on the 16kB SRAM macro, which shows also those exploring different design choices of the row decoder/WL drivers creates new optimal ED design points at the macro level. The results indicate an improvement of the min delay of 15% was achieved after optimization. We combine the ED trade-off of both the row decoder/WL driver and the

sense amplifier to generate the optimized ED curve of the 16kB SRAM macro in Figure 4.6. The results indicate that utilization of the optimization scheme resulted in better design points that cannot be achieved with the top down design approach, and that the optimal design point cannot be achieved without exploring the circuit block energy-performance trade-off that confirms with [43]. The results also indicate an improvement of ~30% in the Energy for a given delay constraints when the optimization scheme is utilized.

Similarly, Figure 4.7 shows the Pareto-optimal curve of a 64kB SRAM macro with and without using ViPro for hybrid optimization, and Figure 4.8 shows the Pareto-curves of a 4kB SRAM macro. In Figure 4.7, the fixed sizing of the row decoder/WL driver and the sense amplifier was designed close to the min delay point. Therefore, using the hybrid optimization scheme resulted only in slight improvement in the min delay and significant improvement in the min energy points. The results indicate an improvement of ~30% in the Energy for a given delay constraints when the optimization scheme is utilized. In Figure 4.8 demonstrates the results of optimization when the periphery circuits were initially sized towards min energy. ViPro achieved significant improvement in the min energy points ~13% and a slight improvement in the min delay points ~4%. ViPro was used to locate design point A in Figure 4.6 (minimize delay, under energy constraints < 4pJ). The average speed up is ~10X compared to exploring all the macro configurations (banks, rows, col) to generate the optimal ED points. ViPro was also used to locate design point B in Figure 4.7 (minimize energy, under delay constraints < 1.1ns). The average speed up is ~12X compared to exploring all the macro configurations (banks, rows, col) to generate the optimal ED points.

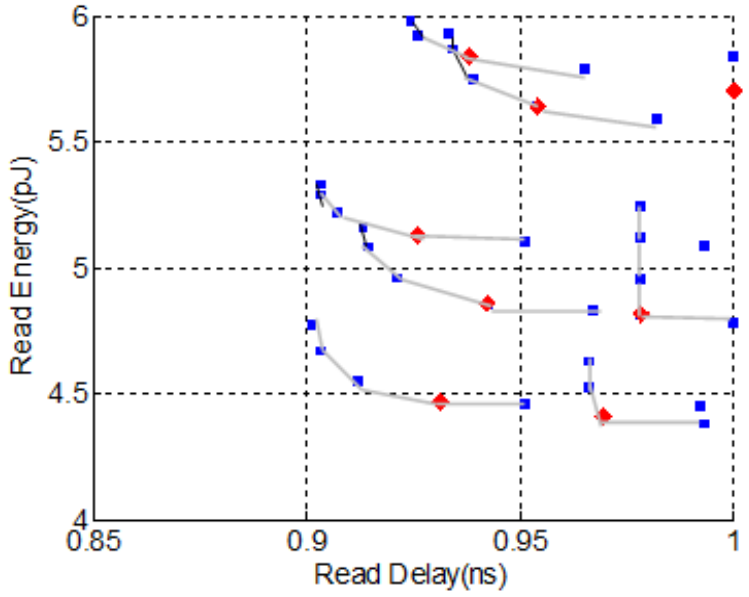


Figure 4.4. The effect of exploring the ED trade-off of the row decoder/WL driver on the design space of a 16kB SRAM macro in 90nm technology node

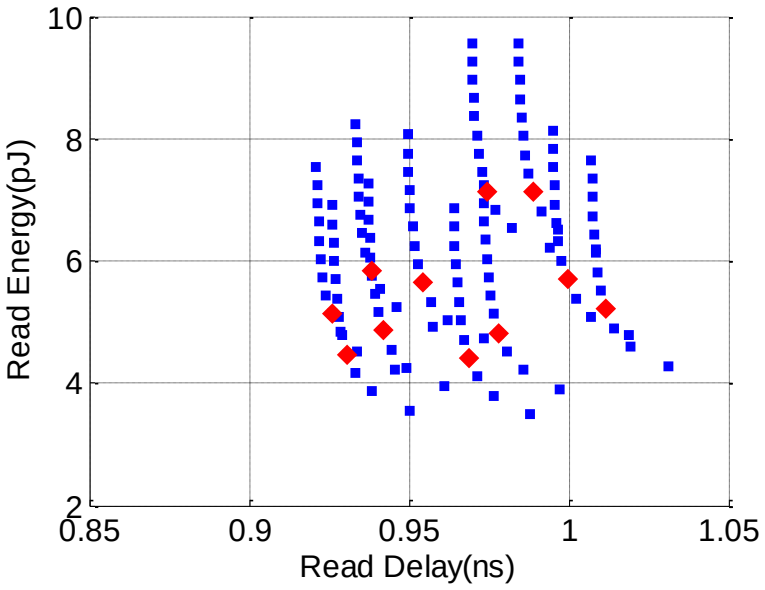


Figure 4.5. The effect of exploring the ED trade-off of the sense amplifier on the design space of a 16kB SRAM macro in 90nm technology node

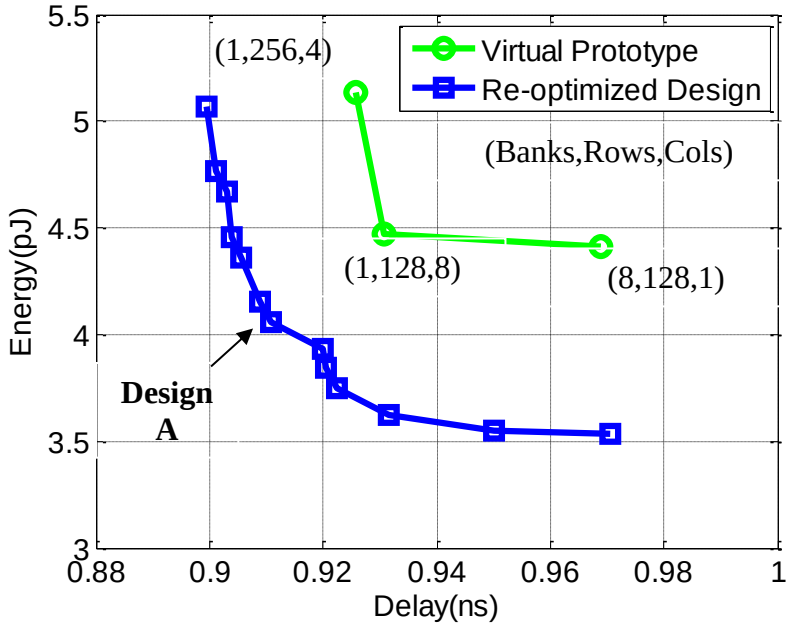


Figure 4.6. The Read Energy and Delay Pareto curve for a 16kB SRAM after optimizing the transistor sizing of the row decoder, WL driver and the sense amplifier.

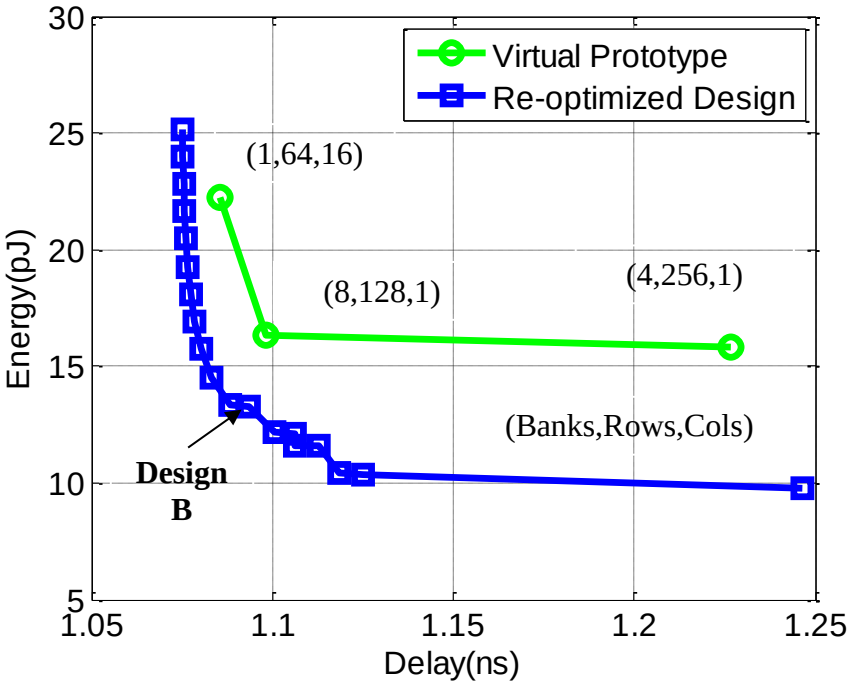


Figure 4.7. The Read Energy and Delay Pareto curve for a 64kB SRAM after optimizing the transistor sizing of the row decoder, WL driver and the sense amplifier.

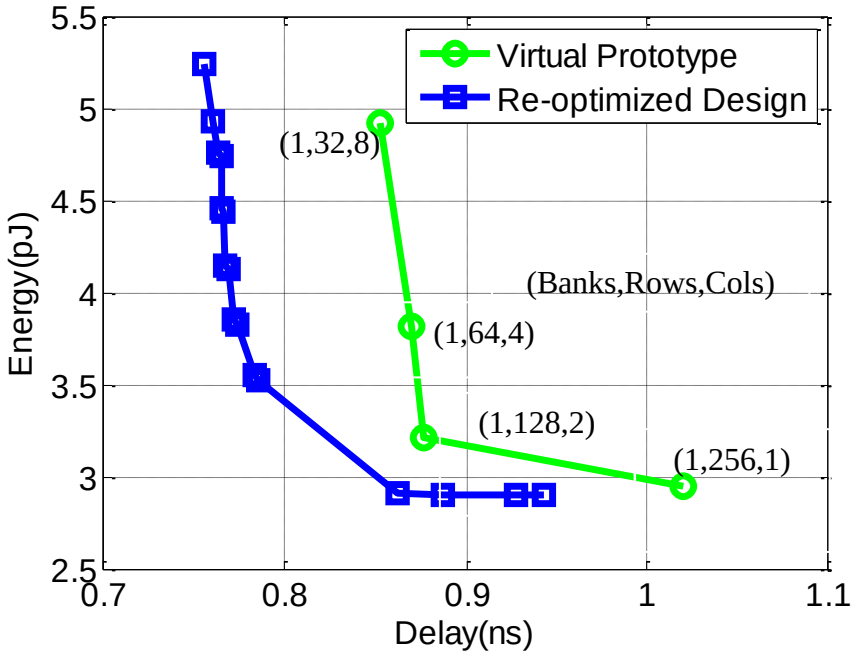


Figure 4.8. The Read Energy and Delay Pareto curve for a 4kB SRAM after optimizing the transistor sizing of the row decoder and WL driver and the sense amplifier.

Write Assist

In this section we utilized the hybrid optimization scheme to investigate the effect of utilizing write assist circuits in 16kB SRAM macro. As SRAMs continue to scale, peripheral assist methods will be needed to allow for continued voltage scaling [30]-[35]. Boosting the WL above V_{DD} strengthens the pass-gate transistor in order to improve write-ability. The downside is that it reduces read stability in half-selected cells. However by increasing the read current through the pass-gate, it also reduces the read delay. The implementation of WL boosting is typically done using a charge pump circuit. In our experiments, the boost capacitors were tuned to provide boosting voltages of 200 mV and 400 mV of boost to the WL voltage. The trade-off is that boosting the WL above V_{DD} minimizes the write delay on the cost of increased write energy. Using ViPro, we are able to directly calculate the overhead of using a boosted WL scheme. Due to its hierarchical design, the tool can easily select between a standard WL driver and the boosted WL circuit.

Once the boosted WL driver is selected, the tool automatically adjusts the WL voltage going to the bitcell to account for the boosted voltage. Figure 4.9 shows the results of using the hybrid optimization scheme to generate the optimal ED curve of a 16kB SRAM in the nominal case, 200 mV and 400 mV boosted WL voltage cases. The results indicate that, at the macro level, the WL boosting creates new optimal ED design points at the min delay designs. The WL boosting circuit does not signficantly affect the optimal configurations, however it does shift the total energy slightly higher, and the worst case delay slightly lower. The average energy overhead is approximately 7%. The word line boosting circuit reduces the delay of the optimal configuration by approximately 5%.

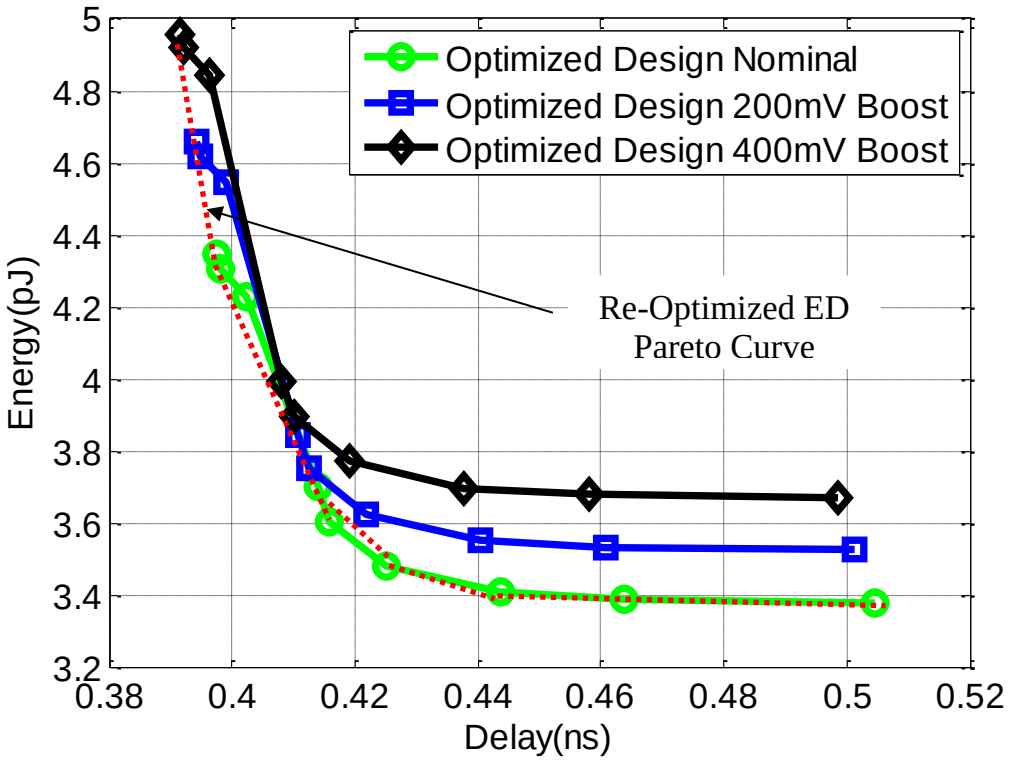


Figure 4.9 The Write Energy and Write Delay Pareto curve for a 16kB SRAM after optimizing the transistor sizing of the row decoder and WL driver and the sense amplifier.

4.5 Conclusion

In this chapter, we presented a hybrid optimization methodology for SRAM circuit-architecture co-design that combines convex optimization (CO) to explore the power-performance trade-off in lower level blocks with macro level optimization. We combined the optimal circuit design points from CO at the architecture level to explore the power-performance trade-off for the complete SRAM macro with a simulated annealing (SA) algorithm. We apply the scheme on to explore the power-performance tradeoff in 4kB, 16kB, and 64MB SRAMs in a 90nm technology. Improvement of up to ~30% in energy/delay was achieved after optimizing the transistor sizing of the row decoder, WL driver and the sense amplifier. The scheme provided a speedup of up to 12x compared to utilizing only convex optimization for architecture circuit co-design. Circuit level options including write assist techniques, and varying periphery circuits sizing were included in the analysis, which shows that the hybrid optimization approach generates better designs than a top level approach that omits circuit level optimizations.

5. THESIS CONCLUSION

Process variations are a main challenge on the design of SRAM in advanced technology nodes, and can limit the min supply voltage and the min energy consumption. In this work, we focused on circuits that combat these variations effect enabling ultra-low power operation. In addition, we proposed a CAD scheme to co-optimize the architecture and circuit structure of the SRAM to further achieve optimal low power operation. A summary of the contribution is highlighted below

5.1 Summary of Contributions

- A typically detrimental aging effect that improves the SRAM sense amplifier offset to 50mv is presented using stress of 1.7× of the nominal VDD and 45°C for 12 hours.
- An online offset compensation circuit is presented that improves σ of the sense amplifier offset voltage by 80% and limited the absolute maximum value of the offset voltage to 50 mV using a 1 MHz split phase frequency and 32fF output capacitance. The circuit is used to design a 16 kB SRAM and showed a reduction in the total energy and delay of 10% and 15% respectively.
- A digital WL control circuit is presented that tracks TCRIT of a 78kbit SRAM memory across PVT variations. Energy savings of 45% is achieved compared to the worst case guard band design approach with a total energy overhead of ~6.5% single write operation of the memory and a total area overhead of ~0.12% the area of the array.
- A CAD technique is presented that optimize the SRAM circuit-architecture design. The scheme located the min energy points of a 4kB, 16kB, and 64MB SRAMs in a 90nm technology.

5.2 Concluding Thoughts and Future Directions

The second chapter in this work proposed a circuit that limited the sense amplifier offset voltage to 50mV. This scheme reduced the energy of the SRAM by 10% compared to the uncompensated case. However, in most cases the area overhead imposed by the compensation circuitry lends the concept of compensation unattractive in high density SRAMs. A compensation circuit that achieves the same offset compensation with a minimal area overhead could be an area of research for high density SRAMs.

In this third chapter, we analyzed a WL control scheme that tracks the WL pulse of a 78kB SRAM in 28nm across PVT conditions. Analysis of the scheme for various sizes of SRAMs is needed to determine the energy savings relation to the size of the memory. Accordingly, the scheme can be analyzed and designed to control multiple L1-2 cache memories and shows the potential savings. Also, the scheme uses a capacitive load at the output of the sensor (CL) equivalent to the capacitive load of the Wordline of the array, to achieve similar TCRIT distribution to the array bitcells. This dramatically increases the sensor energy and area (energy overhead=0.5X single write operation energy, area overhead=7X bitcell area). If a small load is used instead, the overhead can be reduced.

The fourth chapter presented an optimization scheme to co-design the circuit and architecture of the SRAMs. Thorough benchmarking of the scheme versus other optimization scheme is important to evaluate the method, which was part of the work the author intended to do but couldn’t because of the time limit.

6. PUBLICATIONS

(1) **P. Beshay**, J. Bolus, T. Blalock, V. Chandra, and B. H. Calhoun "SRAM Sense Amplifier Offset Cancellation Using BTI Stress" Subthreshold Microelectronics Conference (SubVT), 2012

(2) **P. Beshay**, J. Ryan, and B. H. Calhoun "Sub-threshold Sense Amplifier Compensation Using Auto-zeroing Circuitry" Subthreshold Microelectronics Conference (SubVT), 2012

(3) **P. Beshay**, J. Ryan, and B. H. Calhoun. "A Digital Auto-Zeroing Circuit to Reduce Offset in Sub-Threshold Sense Amplifiers." *Journal of Low Power Electronics and Applications* 3.2 (2013): 159-173.

(4) **P. Beshay**, V. Chandra, R. Aitken and B. H. Calhoun "A Digital Dynamic Write Margin Sensor for Low Power Read/Write Operations in 28nm SRAM" ISLPED 2014 (Accepted)

(5) J. Boley, **P. Beshay** and B. H. Calhoun "Virtual Prototyper (ViPro): An SRAM Design Tool for Yield Constrained Optimization" ICCAD 2014 (Under Submission)

(6) **P. Beshay**, J. Boley and B. H. Calhoun "A Hybrid Optimization Scheme for Circuit / Architecture Co-design of Complete SRAM Macros" CICD 2014 (Under Submission)

7. BIBLIOGRAPHY

[1]. N. Verma et al. "Ultra-Low-Power SRAM Design In High Variability Advanced SRAM CMOS" MIT, PhD Thesis 2009

[2]. J. Wang et al. "Improving SRAM Vmin and Yield by Using Variation Aware BTI Stress" CICC 2010.

[3]. N. Verma "A 256 kb 65 nm 8T Sub-threshold SRAM Employing Sense-Amplifier Redundancy" ISSCC 2008.

[4]. S. Cosemans "3.6pJ/Access 480MHz, 128Kbit on-Chip SRAM with 850MHz Boost Mode" ESSCIRC 2008.

[5]. A.T. Krishnan et al. "SRAM cell static noise margin and Vmin sensitivity to transistor degradation" IEDM 2006.

[6]. M. Bhargava et al. "Low-Overhead, Digital Offset Compensated, SRAM Sense Amplifiers" CICC 2009. L. Pileggi "Mismatch Analysis and Statistical Design" CICC 2008.

[7]. J.C. Lin et al. "Time dependent vccmin degradation of SRAM fabricated with high-k gate dielectrics" IRPS 2007.

[8]. Ryan, J.F.; Calhoun, B.H. Minimizing Offset for Latching Voltage-Mode Sense Amplifiers for Sub-Threshold Operation. In Proceedings of the 9th International Symposium on Quality Electronic Design, San Jose, California, USA, 17–19 March 2008; pp. 127–132.

[9]. Sachdev, M.; Sharifkhani, M.; Shah, J.S.; Rennie, D. Sense-amplification with Offset Cancellation for Static Random Access Memories. U.S. Patent Application 12/757,033, 8 April 2010.

[10]. Sharifkhani, Mohammad, and Manoj Sachdev. "SRAM cell stability: A dynamic perspective." *Solid-State Circuits, IEEE Journal of* 44.2 (2009): 609-619.

[11]. Dong, Wei, Peng Li, and Garng M. Huang. "SRAM dynamic stability: theory, variability and analysis." *Computer-Aided Design, 2008. ICCAD 2008. IEEE/ACM International Conference on*. IEEE, 2008.

[12]. Aitken, Robert, and Sachin Idgunji. "Worst-case design and margin for embedded SRAM." *Design, Automation & Test in Europe Conference & Exhibition, 2007. DATE'07*. IEEE, 2007.

[13]. Chandra, Vikas, Cezary Pietrzyk, and Robert Aitken. "On the efficacy of write-assist techniques in low voltage nanoscale SRAMs." *Proceedings of the Conference on Design, Automation and Test in Europe*. European Design and Automation Association, 2010.

[14]. Nho, Hyunwoo, et al. "A 32nm High-k metal gate SRAM with adaptive dynamic stability enhancement for low-voltage operation." Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2010 IEEE International. IEEE, 2010.

[15]. Carlson, Andrew, et al. "Compensation of systematic variations through optimal biasing of SRAM wordlines." *Custom Integrated Circuits Conference, 2008. CICC 2008. IEEE*. IEEE, 2008.

[16]. Abu-Rahma, Mohamed H., Mohab Anis, and Sei Seung Yoon. "Reducing SRAM power using fine-grained wordline pulsewidth control." Very Large Scale Integration (VLSI) Systems, IEEE Transactions on 18.3 (2010): 356-364.

[17]. Amrutur, Bharadwaj S., and Mark A. Horowitz. "A replica technique for wordline and sense control in low-power SRAM's." Solid-State Circuits, IEEE Journal of33.8 (1998): 1208-1219.

[18]. Heald, Raymond, and Ping Wang. "Variability in sub-100nm SRAM designs."Proceedings of the 2004 IEEE/ACM International conference on Computer-aided design. IEEE Computer Society, 2004.

[19]. Yamaoka, Masanao, et al. "Low-power embedded SRAM modules with expanded margins for writing." Solid-State Circuits Conference, 2005. Digest of Technical Papers. ISSCC. 2005 IEEE International. IEEE, 2005.

[20]. Bhavnagarwala, Azeez, et al. "Fluctuation limits & scaling opportunities for CMOS SRAM cells." *Electron Devices Meeting, 2005. IEDM Technical Digest. IEEE Internation-al*. IEEE, 2005.

[21]. Boley, James, et al. "Leveraging sensitivity analysis for fast, accurate estimation of SRAM dynamic write V MIN." *Proceedings of the Conference on Design, Automation and Test in Europe*. EDA Consortium, 2013.

[22]. Souri, Kamran, Youngcheol Chae, and Kofi Makinwa. "A CMOS temperature sensor with a voltage-calibrated inaccuracy of±0.15° C (3σ) from− 55 to 125° C." *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2012 IEEE International*. IEEE, 2012.

[23]. Li, Y. William, et al. "A 1.05 V 1.6 mW 0.45 C 3σ-resolution ΔΣ-based temperature sensor with parasitic-resistance compensation in 32nm CMOS."*Solid-State Circuits Conference-Digest of Technical Papers, 2009. ISSCC 2009. IEEE International*. IEEE, 2009.

[24]. Chen, Shi-Wen, et al. "Fully on-chip temperature, process, and voltage sensors." *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*. IEEE, 2010.

[25]. Joshi, R., et al. "6.6+ GHz Low Vmin, read and half select disturb-free 1.2 Mb SRAM." VLSI Circuits, 2007 IEEE Symposium on. IEEE, 2007.

[26]. Helal, Belal M., et al. "A low jitter 1.6 GHz multiplying DLL utilizing a scrambling time-to-digital converter and digital correlation." VLSI Circuits, 2007 IEEE Symposium on. IEEE, 2007.

[27]. Chang, Jonathan, et al. "A 20nm 112Mb SRAM in High-κ metal-gate with assist circuitry for low-leakage and low-V MIN applications." *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2013 IEEE International*. IEEE, 2013.

[28]. Karl, Eric, et al. "A 4.6 GHz 162Mb SRAM design in 22nm tri-gate CMOS technology with integrated active V MIN-enhancing assist circuitry." *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2012 IEEE International*. IEEE, 2012.

[29]. Song, Taejoong, et al. "13.2 A 14nm FinFET 128Mb 6T SRAM with V MIN-enhancement techniques for low-power applications." *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2014 IEEE International*. IEEE, 2014.

[30]. Mann, Randy W., et al. "Impact of circuit assist methods on margin and performance in 6T SRAM." *Solid-State Electronics* 54.11 (2010): 1398-1407.

[31]. Chandra, Vikas, Cezary Pietrzyk, and Robert Aitken. "On the efficacy of write-assist techniques in low voltage nanoscale SRAMs." *Proceedings of the Conference on Design, Automation and Test in Europe*. European Design and Automation Association, 2010.

[32]. Kulkarni, Jaydeep P., et al. "A read-disturb-free, differential sensing 1R/1W port, 8T Bitcell array." *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on* 19.9 (2011): 1727-1730.

[33]. Kulkarni, Jaydeep, et al. "Capacitive-coupling wordline boosting with self-induced V CC collapse for write V MIN reduction in 22-nm 8T SRAM." *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2012 IEEE International*. IEEE, 2012.

[34]. Khayatzadeh, Mahmood, and Yong Lian. "Average-8T Differential-Sensing Subthreshold SRAM With Bit Interleaving and 1k Bits Per Bitline." 1-1.

[35]. Lee, Donghyuk, et al. "Tiered-latency DRAM: A low latency and low cost DRAM architecture." *High Performance Computer Architecture (HPCA2013), 2013 IEEE 19th International Symposium on*. IEEE, 2013.

[36]. "Device-Architecture Co-Optimization of STT-RAM Based Memory for Low Power Embedded Systems" 2011.

[37]. "Noise Margin, Critical Charge and Power-Delay Tradeoffs for SRAM Design" 2011.

[38]. "Statistical DOE–ILP based power–performance–process (P3) optimization of nano-CMOS SRAM" 2012.

[39]. "Energy-Delay Tradeoffs in Combinational Logic using Gate Sizing and Supply Voltage Optimization" 2002.

[40]. Muralimanohar N. et al, "Optimizing NUCA Organizations and Wiring Alternatives for Large Caches With CACTI 6.0" 2007

[41]. Sheng Li, "CACTI-P: Architecture-Level Modeling for SRAM-based Structures with Advanced Leakage Reduction Techniques" 2011

[42]. Nalam, Satyanand, et al. "Virtual prototyper (ViPro): an early design space exploration and optimization tool for SRAM designers." *Proceedings of the 47th Design Automation Conference*. ACM, 2010.

[43]. Azizi, Omid, et al. "Energy-performance tradeoffs in processor architecture and circuit design: a marginal cost analysis." *ACM SIGARCH Computer Architecture News*. Vol. 38. No. 3. ACM, 2010.

[44]. D. Patil, S. Kim, “Stanford Circuit Optimization Tool (SCOT) User Guide,” private correspondence with authors: ddpatil@stanford.edu, sjkim@stanford.edu.

[45]. Boyd, Stephen P., et al. "Digital circuit optimization via geometric programming." *Operations Research* 53.6 (2005): 899-932.

[46]. Huntzicker, Steven, et al. "Energy-delay tradeoffs in 32-bit static shifter designs." *Computer Design, 2008. ICCD 2008. IEEE International Conference on*. IEEE, 2008.

[47]. Patil, Dinesh, et al. "Robust energy-efficient adder topologies." *Computer Arithmetic, 2007. ARITH'07.a 18th IEEE Symposium on*. IEEE, 2007.

[48]. Beshay, Peter, Benton H. Calhoun, and J. F. Ryan. "Sub-threshold Sense Amplifier Compensation Using Auto-zeroing Circuitry." *variations* 1 (2012): 5.

[49]. Beshay, Peter, Joseph F. Ryan, and Benton H. Calhoun. "A Digital Auto-Zeroing Circuit to Reduce Offset in Sub-Threshold Sense Amplifiers." *Journal of Low Power Electronics and Applications* 3.2 (2013): 15

8. APPENDIX

8.1 Terminology

CMOS: Complementary MOS, circuts that contain both NMOS and PMOS devices.

MOSFET: Metal Oxide Semiconductor Field-Effect Transistor, a transistor that uses a metal oxide as an insulator between a polysilicon gate and a semiconductor. An electric field can be used to create an inversion layer or channel between the source and drain terminals of the transistor.

NMOS: A MOSFET that utilizes an n-type inversion layer for conducting current.

PMOS: A MOSFET that utilizes a p-type inversion layer for conducting current.

Transistor: Refers to a MOSFET in this thesis.

VDD: Reference for the high potential power supply.

VSS: Ground, reference for the low potential power supply (0 V).

VT: Threshold voltage, the voltage at which the channel in a transistor undergoes strong inversion and begins conducting.

Sense Amplifier (SA): An analog circuit that amplifies a differential voltage. It is used to speed up reading by sensing and amplify the differential between BL and BLB. It also helps in avoiding the energy overhead of fully discharging the bitlines which have large capacitances.

SRAM: Static Random Access Memory, which stores data statically using a cross-coupled inverter pair.

Bitcell (cell): Basic unit of an SRAM that stores one bit of data. Essentially composed of a cross-coupled inverter pair and zero or more access ports or transistors.

BL: Bitline, a wire that connects the bitcell, possibly through an access transistor, to the sense amplifiers and the Bitline drivers which supply the data during a write.

BLB: Bitline-bar, a complementary Bitline, present in a conventional 6T bitcell.

WL: Wordline, a wire that controls the gates of the access transistors of a bitcell.

MC Simulation: Monte Carlo simulation, the technique of simulating a circuit over a wide range of randomly chosen values for device parameters.

PVT Variations: The effect of the variations of the temperature, supply voltage or manufacturing process on the circuit electrical behavior.