Circuit Design for Embedded Memory in Low-Power Integrated Circuits

by

Masood Qazi

Submitted to the Department of Electrical Engineering and Computer Science

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2012

© 2012 Massachusetts Institute of Technology. All rights reserved.

Author: _____

Department of Electrical Engineering and Computer Science

May 23, 2012

Certified by: _____

Anantha P. Chandrakasan

Joseph F. and Nancy P. Keithley Professor of Electrical Engineering

Thesis Supervisor

Accepted by: _____

Leslie Kolodziejski

Chair, Department Committee on Graduate Theses

Circuit Design for Embedded Memory in Low-Power Integrated Circuits

by
Masood Qazi

Submitted to the Department of Electrical Engineering and Computer Science
on May 23, 2012, in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

Abstract

This thesis explores the challenges for integrating embedded static random access memory (SRAM) and non-volatile memory—based on ferroelectric capacitor technology—into low-power integrated circuits.

First considered is the impact of process variation in deep-submicron technologies on SRAM, which must exhibit higher density and performance at increased levels of integration with every new semiconductor generation. Techniques to speed up the statistical analysis of physical memory designs by a factor of 100 to 10,000 relative to the conventional Monte Carlo Method are developed. The proposed methods build upon the Importance Sampling simulation algorithm and efficiently explore the sample space of transistor parameter fluctuation. Process variation in SRAM at low-voltage is further investigated experimentally with a 512kb 8T SRAM test chip in 45nm SOI CMOS technology. For active operation, an AC coupled sense amplifier and regenerative global bitline scheme are designed to operate at the limit of on current and off current separation on a single-ended SRAM bitline. The SRAM operates from 1.2 V down to 0.57 V with access times from 400ps to 3.4ns. For standby power, a data retention voltage sensor predicts the mismatch-limited minimum supply voltage without corrupting the contents of the memory.

The leakage power of SRAM forces the chip designer to seek non-volatile memory in applications such as portable electronics that retain significant quantities of data over long durations. In this scenario, the energy cost of accessing data must be minimized. This thesis presents a ferroelectric random access memory (FRAM) prototype that addresses the challenges of sensing diminishingly small charge under conditions favorable to low access energy with a time-to-digital sensing scheme. The 1 Mb 1T1C FRAM fabricated in 130 nm CMOS operates from 1.5 V to 1.0 V with corresponding access energy from 19.2 pJ to 9.8 pJ per bit. Finally, the computational state of sequential elements interspersed in CMOS logic, also restricts the ability to power gate. To enable simple and fast turn-on, ferroelectric capacitors are integrated into the design of a standard cell register, whose non-volatile operation is made compatible with the digital design flow. A test-case circuit containing ferroelectric registers exhibits non-volatile operation and consumes less than 1.3 pJ per bit of state information and less than 10 clock cycles to save or restore with no minimum standby power requirement in-between active periods.

Thesis Supervisor: Anantha P. Chandrakasan
Joseph F. and Nancy P. Keithley Professor of Electrical Engineering

Acknowledgments

I would like to acknowledge the following people and entities that have been involved in my graduate experience:

- Advisor: Anantha Chandrakasan
- Committee: Devavrat Shah and Ajith Amerasekera
- Additional collaborators: Lara Dolecek, Leland Chang, Kevin Stawiasz, Michael Clinton, Steve Bartling, Mehul Tikekar
- Sponsorship: FCRP/C2S2
- Chip fabrication access: IBM, Texas Instruments, and Dennis Buss
- Colleagues from Taiwan Semiconductor Manufacturing Corporation: Derrick Lin, Rinus Tek Po Lee, Brian Yang, Ching Tien Peng, JT Tzeng, HT Chen, David Scott, Bing J Sheu, Ali Keshavarzi, Sreedhar Natarajan
- Colleagues from Texas Instruments (too many to name)
- Administration: Margaret Flaherty
- MIT Lab experiences: Ananthagroup members, past and present (too many to name)
- My Mother, Father, and Sister
- Anicham

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The design of low-power integrated circuits necessitates the re-design of how computational state is managed—in the form of high density arrays of memory cells and in the form of registers to facilitate sequential processing. Because memory comprises a significant portion of chip area in most integrated circuits, it also plays a significant role in the power consumption, speed of operation, and manufacturing yield. This thesis concentrates on the circuit design challenges to integrating embedded memory into low-power integrated circuits.

1.1 Areas of Challenge in Embedded Memory and Low-Power

The first memory that is considered is SRAM. Embedded SRAM is a vital component of digital integrated circuits and often constitutes a dominant portion of chip area. For example, over half the area of a state-of-the art microprocessor shown in Fig. 1.1 is comprised of embedded SRAM and memory interface circuits. Achieving such a level of integration is not straightforward, but integrating more memory on chip provides an effective means to use silicon because of memory's lower power density, layout regularity, and performance and power benefits from reduced off-chip bandwidth. Consequently, the amount of embedded SRAM has grown from 2 KB [6] to 12 MB [1] in 25 years.

Shown in Fig. 1.2(a) is a plot of reported cell areas in fully functional SRAM macros versus the technology node for the past few years. The cell area has scaled with the scaling of the critical feature size. Fig. 1.2(b) plots an unconventional metric—the number of SRAM bits per mm^2 of silicon in high performance microprocessor chips—that reveals reduced SRAM cell area does not readily translate into increased SRAM utilization. The amount of SRAM is dominated by the last-level cache in these microprocessors, though a smaller quantity of SRAM is used for local high-speed cache. This discrepancy in trends is due to a number of limitations of SRAM, all related to local variation: SRAM often needs a separate, elevated power supply; excessive SRAM timing variation degrades performance; and, uncertain aging effects show up first in the highly integrated and highly sensitive SRAM cells.

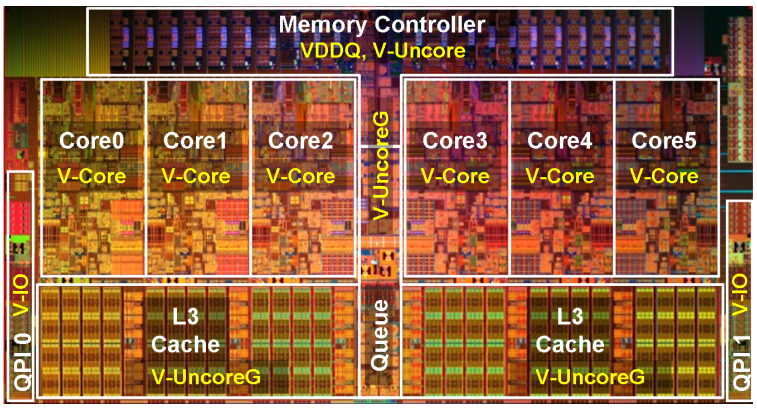
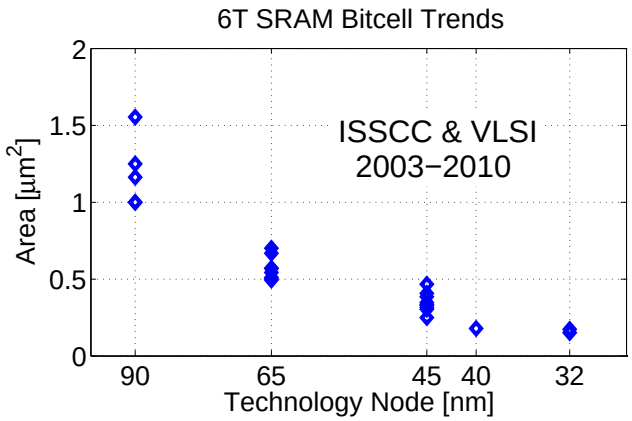
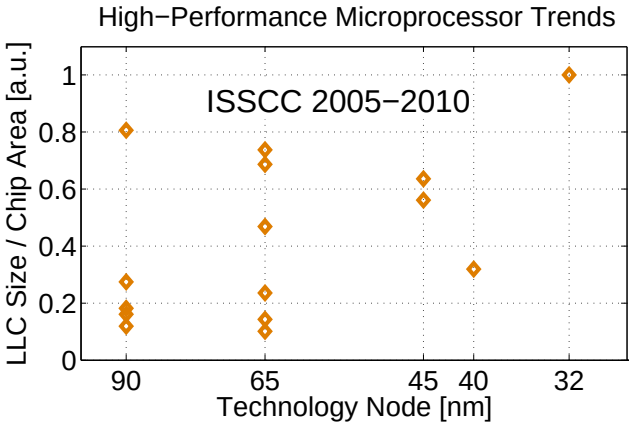


Figure 1.1: State of the art microprocessor chip whose area is half SRAM [1]

Though SRAM provides the most convenient, fast, and potentially lowest access energy embedded memory option, many systems require non-volatile memory for its essential feature of consuming zero standby power while retaining data. A cardiac loop recorder exemplifies one such system and is shown in Fig. 1.3. It is an implantable medical device that processes critical patient data. It monitors the ECG waveform of a patient and can store a buffered amount of the patient data after a syncopal episode (fainting) so that the physician can examine the ECG data to look for arrhythmia, a very serious condition. Such a system is limited in processing capability, lasts only 14 months, and can store only 40 minutes of data because it must operate from a battery that cannot be replaced without a surgery [7]. By applying some of the solutions in this thesis for low voltage SRAM cache and low energy



(a)



(b)

Figure 1.2: (a) SRAM bitcell area scaling and (b) resulting utilization of SRAM on recent high performance microprocessor designs.

FeRAM storage, greater functionality can be delivered in the patient care. For example, more sophisticated processing (as in [8]) can be implemented locally to automatically detect abnormalities without requiring the patient to explicitly initiate the recording of data.

Lastly, low-power memory solutions must enable an overall low-power system. The central feature of a low-power system is the ability to turn on and off rapidly. This behavior may apply to the whole system or specific portions of the system that are idle. A driving application is a wireless sensor network in which each node derives power from ambient sources. This type of system would be useful for monitoring the environment of office buildings as in Fig. 1.4. Building operating costs costs constitute 70% of US and 40% of worldwide energy

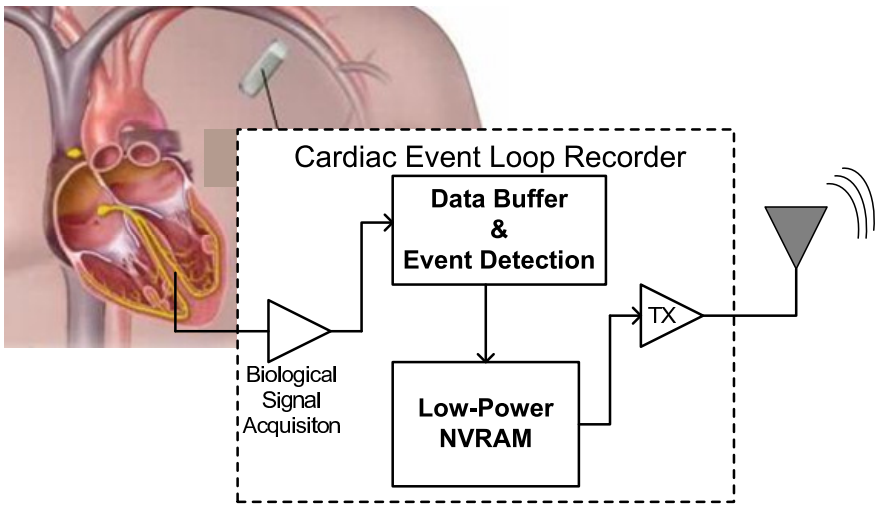


Figure 1.3: Example Application: Cardiac Loop Recorder Block Diagram

expenditure [9]. A large fraction of this energy consumption is wasted in lighting and climate control of unoccupied areas. In order to make buildings more efficient, one can consider upgrading the current infrastructure or simply building brand new efficient buildings. Because of the cost barrier, great impact in the reduction of global energy use can be achieved by simply monitoring building environments more precisely in terms of spatial and temporal resolution, and then using the information to dynamically actuate power expenditure for only the locations that require climate control and lighting.

In this scenario of monitoring the internal environment of a building, all the sensors cannot be wired together, nor can the batteries of so many devices be regularly replaced. Autonomous operation from harvested energy is a critical requirement in order to free the size and computation constraints imposed by the battery. Almost poetically, the key to reducing energy consumption on a global scale requires ultra-low-energy electronics that can operate with nanojoules.

Figure 1.4: Smart Building Application Overview

1.2 Contributions of this Thesis

This thesis covers the four topics of SRAM statistical analysis, low-voltage SRAM design, low-energy non-volatile memory design, and non-volatile computation. The first three topics have substantiation by the measurement of silicon prototypes and the last topic is validated with simulation analysis.

1.2.1 Statistical SRAM Simulation Methodologies

This area of work presents a technique to evaluate the timing variation of SRAM, which becomes especially challenging for low-voltage low-power designs [10]. Specifically, a method called loop flattening that reduces the evaluation of the timing statistics in the complex, highly structured circuit to that of a single chain of component circuits is justified. To then very quickly evaluate the timing delay of a single chain, a statistical method based on Importance Sampling augmented with targeted, high-dimensional, spherical sampling can be employed. The overall methodology has shown 650X or greater speed-up over the nominal Monte Carlo approach with 10.5 % accuracy in probability. Examples based on both the

large-signal and small-signal SRAM read path are discussed and a detailed comparison with state of the art accelerated statistical simulation techniques is given.

Contributions: A justification based on numerical experiment and mathematical proof is provided for the loop flattening technique. In addition, a detailed exposition of the Spherical Importance Sampling algorithm is provided with a set of parameter values that work for a wide range of probabilities and multiple simulation types.

1.2.2 Voltage scaling in SRAM

This work presents an 8T SRAM fabricated in 45 nm SOI CMOS that exhibits voltage scalable operation from 1.2V down to 0.57 V with access times from 400 ps to 3.4 ns [11]. Timing variation and the challenge of low voltage operation are addressed with an AC-coupled sense amplifier. An area efficient data path is achieved with a regenerative global bitline scheme. Finally, a data retention voltage sensor has been developed to predict the mismatch-limited minimum standby voltage without corrupting the contents of the memory.

Contributions: The schematic and physical implementation details of the AC coupled sense amplifier and regenerative global bitline scheme are provided. For the data retention voltage sensor, a layout structure and an algorithm is developed to apply accelerated statistical analysis on chip.

1.2.3 Low-Energy FRAM test chip

In the effort to achieve low access energy non-volatile memory, challenges are encountered in sensing data at a low power supply voltage. This area of work presents the design of a ferroelectric random access memory (FRAM) as a promising candidate for this need [12]. The challenge of sensing diminishingly small charge and developing circuits compatible with the scaling of FRAM technology to low voltage and more advanced CMOS nodes are addressed with a time-to-digital sensing scheme. In this work, the 1T1C bitcell signal is analyzed, the circuits for a TDC-based sensing network are presented, and the implementation and operation details of a 1 Mb chip are described. The 1 Mb 1T1C FRAM fabricated in 130 nm

CMOS operates from 1.5 V to 1.0 V with corresponding access energy from 19.2 pJ to 9.8 pJ per bit. This approach is generalized to a variety of non-volatile memory technologies.

Contributions: The time-to-digital-sensing scheme is illustrated with schematic, layout, and operation details. The architecture of a high-density FRAM with platelines running parallel to bitlines is highlighted in the construction of a full 1 Mb prototype.

1.2.4 Non-volatile Processing in a Digital Integrated Circuit with Non-volatile Registers

The computational state of sequential elements interspersed in CMOS logic, also restricts the ability to power gate. To enable simple and fast turn-on, ferroelectric capacitors are integrated into the design of a standard cell register, whose non-volatile operation is made compatible with the digital design flow. A test-case circuit exhibits non-volatile operation with the ferroelectric register, requiring less than 1.3 pJ per bit of state information and less than 10 clock cycles to save or restore while consuming zero standby power in-between active periods. The area overhead of the current implementation is estimated to be 49 %. Also accompanying the non-volatile implementation of a digital integrated circuit, is a power management unit and energy harvester interface that ensures sufficient energy to manage computational state before toggling the power of the digital circuit.

Contributions: A circuit topology for a ferroelectric non-volatile register that exhibits self-timed operation is developed. Steps to integrate the non-volatile register into the digital design flow—a power management unit interfacing with an energy harvester, modifications to the synthesis and physical design flow—are described so that a system level demonstration of a battery-less computing node can be constructed.

Chapter 2

Techniques to Efficiently Evaluate SRAM Failure

As its overarching goal, the work in this chapter seeks to increase SRAM utilization by propagating the physical trend of shrinking cell area into the overall system-on-chip improvement. This goal can be achieved if designers have a way to quickly assess the impact of circuit solutions on the operating constraints (e.g., minimum V_{DD}, frequency) to ultimately preserve the overall chip yield.

Specifically, this work focuses on read access yield. It has been observed in measurements that AC fails, manifested as too slow of an access time from one or more addresses, are encountered before DC failures, manifested as the corruption of data at one or more addresses [13]. Therefore, DC stability (write and read margin) is necessary but not sufficient for yielding a memory chip. A significant degree of additional margin must be inserted to meet performance requirements. To meet the target performance, a number of questions must be answered: How much margin? What types of sensing schemes should be employed? Should the bitcell be up-sized?

Further exacerbating the design is the fact that, in general, the exact distributions of the relevant performance metrics are not known. As a consequence, any statistical simulation method unavoidably resorts to numerical solvers such as SPICE. Classical approaches like the

Monte Carlo method require too many iterations of such SPICE evaluations because of the circuit complexity and extremely low tolerable failure probabilities of individual components (10^{-8} and below). Thus, the primary challenges to any statistical simulation method are: (a) dealing with the structural complexity of the timing delay evaluation problem, and (b) estimating timing delay statistics to a very high accuracy.

A lot of exciting recent work has made important progress towards the eventual goal of designing generically applicable, efficient simulation methodologies for circuit performance evaluation. However, this line of work uniformly falls short in addressing the above stated challenges regarding the evaluation of timing delays of integrated SRAMs.

To begin with, in [14, 15, 5, 2, 16], the authors developed efficient sampling based approaches that provide significant speedup over the Monte Carlo. However these works do not deal with the interconnection complexity, i.e., do not address the challenge (a) stated above. And hence, as is, these approaches suffer from the curse of dimensionality that results from a very large number of transistors in relevant circuit applications.

As noted earlier, the main bottleneck for efficient simulation is the utilization of a circuit solver like SPICE, as this approach is necessitated by the lack of an exact analytic description of the variation of performance metrics. In [17], by modeling the bitline signal and the sense amplifier offset (and the timer circuit) with Gaussian distributions, the authors proposed a linearized model for the read path. As this model can be simulated in MATLAB, the SRAM structure can be emulated and the evaluation time can be significantly improved. This approach, though interesting, is limited since an exact analytic description is unlikely to be known in general, or even considered as a truthful approximation of reality. It is then only reasonable to look for a generic simulation method that can work with *any* form of distributions implied by the circuit solver such as SPICE.

In [18] the authors extended upon the conventional worst-case approach to SRAM design in which the largest offset sense amplifier is required to support the weakest memory cell. Their proposed method requires that an underlying Gaussian distribution models the bitcell current—particularly in the extreme tails of the distribution—in order to enable the

numerical convolution of a Gumbel distribution of the worst-case read current with the sense amplifier offset. Nevertheless, their work does not put forth an unambiguous approach as it requires a designer to heuristically allocate the total margin between memory cells and sense amplifiers. A further extension of this type of approach to other relevant circuit blocks based on sensitivity analysis was subsequently introduced in [19]. Because these solutions do not run a full-scale SPICE simulation to directly evaluate the extreme delay statistics, they compromise on the challenge (b) stated previously. A summary of the previous work is shown in Table 2.1.

Table 2.1: Summary of previous work for SRAM statistical analysis

Publication	Contribution
[14]	Application of Importance Sampling to SRAM bitcell simulation, uniform exploration of parameter space, Importance Sampling simulation with mixture of uniform and Gaussian distributions
[15]	Application of Extreme Value Theory to SRAM circuit simulation, parameter sampling filter (blockade) to more frequently evaluate rare conditions
[5]	Implementation of Importance Sampling with a mean-shifted distribution, uniform exploration of parameter space and norm-minimization to identify optimum mean-shift
[2, 16]	Adaptive Importance Simulation algorithms that can adjust the sampling distribution with every new simulation trial
[17]	Simplified modeling of the SRAM read path accompanied by emulation of the memory structure in a MATLAB framework
[18]	An analytical model of worst-case *bitline* current based on bitcell distribution and a methodology based on allocating margin between bitline current and sense amplifier offset
[19]	Sensitivity characterization of different types of circuits in an SRAM (denoted as islands) and a methodology to use the sensitivities to evaluate the interaction among circuit islands and determine yield
This work	A justification of loop flattening to simplify the structure of multiplexed paths to a critical path chain, a two-step Spherical Importance Sampling method to accelerate the spice-based Monte Carlo simulation of SRAM timing

In this chapter, the two challenges for the timing delay analysis of SRAM will be overcome by means of two proposed methods of *Spherical Importance Sampling* and *Loop Flattening*,

respectively.[1] First, the Spherical Importance Sampling technique is developed to evaluate the extremes of a distribution dependent on 12 parameters, wherein a standard Monte Carlo is clearly not useful. The importance sampling-based approach in a recent work [5] employs 'uniform sampling,' and as such does not scale to higher dimensionality (12 dimensions in the case of interest). To cope with the dimensionality, a spherical sampling based approach is used and samples the parameter space in an adaptive manner.

In order to apply Spherical Importance Sampling to the SRAM read path, the Loop Flattening approximation is first justified. It is shown that, surprisingly, the naïve adaptation of the conventional critical path methodology—in which a Monte Carlo simulation of a chain of component circuits (such as row driver, memory cell, and sense amplifier) disregards the relative rate of replication—produces an estimate that is *always* conservative and highly accurate. Indeed, the proposed technique directly contradicts the approaches in [17, 18, 19, 20]. If the loop flattening -based approach indicates that the delay exceeds 130 ps with probability 10^{-5}, then the actual delay will exceed 130 ps with probability less than or equal to 10^{-5}. More importantly, unlike conventional worst-case methodologies, this *conservative approach is increasingly accurate at lower failure levels.*

The elements of this framework will now be presented in detail. Section 2.1, begins with a discussion of Importance Sampling as a fast simulation technique, which shows how an efficient spherical search can greatly simplify the identification of the biasing distribution in a high-dimensional parameter space. The resultant sampling algorithms synthesizes these ideas. Next, Section 2.2 discusses the surprisingly accurate Loop Flattening approximation. Under a representative structure of multiplexed paths and for low probability events, Loop Flattening offers a substantial simplification of the SRAM critical path evaluation. Section 2.3 provides experimental results for the small-signal read path. Section 2.4 delivers the conclusions and indicates future directions.

[1]The work in this chapter is joint work with Mehul Tikekar, Lara Dolecek, Devavrat Shah, and Anantha Chandrakasan. It has been reported in [5] and [10].

2.1 Spherical Importance Sampling

In this section, the Monte Carlo method is introduced and then the Spherical Importance Sampling method is developed with the purpose of significantly reducing the number of SPICE simulations required. Suppose one is given a circuit that fails with probability p, and wishes to identify the value of this failure with a Monte Carlo simulation that produces an estimate $\hat{p}_{\mathrm{MC}}$. Based on a Chernoff bound [21], the number of required Monte Carlo simulation runs is given by:

$$N_{\mathrm{MC}} > \frac{2\ln\left(\frac{1}{\delta}\right)}{p\epsilon^2} \tag{2.1}$$

where $\delta = \Pr\left(\hat{p}_{\mathrm{MC}} < (1-\epsilon)p\right)$. In plain English, Eq. (2.1) says that with probability $1-\delta$, the Monte Carlo estimate $\hat{p}_{\mathrm{MC}}$ will not underestimate the true failure p by more than $\epsilon\times 100\%$. For typical values of δ (0.01 to 0.1) and ϵ (0.05 to 0.3), this expression indicates $N_{\mathrm{MC}} > 100/p$ to $1000/p$. [2]

To accurately estimate low failure probabilities—10^{-6} to 10^{-10} for embedded SRAM—the standard Monte Carlo approach would require too many SPICE simulations as seen from Eq. (2.1). Fortunately, Importance Sampling provides a way to speed up such a simulation [23]. Indeed, because of the ease of implementation (one needs only to provide an indicator of pass/fail from the SPICE simulation), several examples of Importance Sampling applied to circuit design have emerged [14, 5, 24, 25, 16, 2].

This work focuses on identifying the most likely region of failure in the parameter space, relying on the well-known notion of a worst-case point [26, 27, 28, 29]. The Spherical Importance Sampling approach defines a dominant point of failure as the point that minimizes the quadratic norm (in a parameter space normalized by the standard deviation along each coordinate). This point is then used as a mean shift in an IS simulation to quickly evaluate the circuit failure probability [30]. The general picture of this strategy is illustrated by the

[2] From the perspective of a designer who does not know the true failure p, the estimator $\hat{p}$ is a random variable whose distribution and related parameters, such as mean and variance, are unknown. In such cases, sample mean and sample variance are used to determine the confidence level and confidence interval under a Gaussian approximation. For nominal Monte Carlo simulations, confidence intervals can be obtained using other techniques that do not require estimating the parameters of the underlying distributions [22].

cartoon in Fig. 2.1. Furthermore, the method presented herein presents an effective, general way to find the optimal sampling distribution without expending too many simulations relative to the nominal Monte Carlo approach.

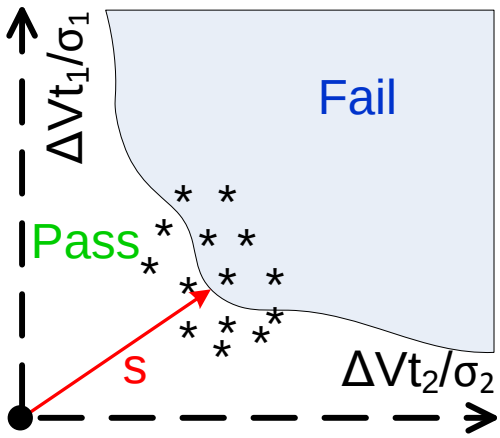


Figure 2.1: A cartoon of statistical analysis techniques focused on identifying the dominant failure mechanism by minimizing the distance between the nominal condition (origin) and the failure boundary.

First, the basics of the Importance Sampling are revisited in the context of this setup in which a SPICE simulation is performed for every realization of a set of 12 random circuit parameters, say $A_\ell, 1 \leq \ell \leq 12$, that correspond to the random threshold voltage variation of transistors in the circuit to be studied (which is presented in full detail in Section 2.3). To capture local mismatch process variation in SRAM,[3] these variables are typically modeled as independent Gaussians with means $\mu_\ell, 1 \leq \ell \leq 12$, and variances $\sigma_\ell^2, 1 \leq \ell \leq 12$. Since the relationship between A_ℓ's and circuit failure is described via an implicit function in SPICE, a simulation-based approach is the primary method for estimating failure. Failure is evaluated from a transient SPICE simulation, checking for a polarity of sense amplifier differential output that is consistent with the bitcell data.

In a nutshell, one draws values for $A_\ell, 1 \leq \ell \leq 12$ as per normal distributions with means $\mu_\ell + s_\ell, 1 \leq \ell \leq 12$, and variances, $\sigma_\ell^2, 1 \leq \ell \leq 12$, where $s_\ell, 1 \leq \ell \leq 12$ are *cleverly* chosen mean shifts so that a circuit failure becomes likely under the new shifts. Under these

[3]Global variation from one chip to another is modeled as a common shift in the mean of threshold voltage distributions. This type of variation is also a concern for chip yield (but does not dominate simulation cost) and must be separately treated after the impact of local variation is evaluated.

shifted variables, one estimates the probability of circuit failure highly accurately, while using only a few samples, say N. The explicit transformation between the original and the new sampling domains reverts this estimate back to the original domain at no additional cost. For $1 \leq \ell \leq 12$, let the values of A_ℓ for these N samples be $a_\ell(n), 1 \leq n \leq N$. Let $\chi(n) = 1$ if the circuit behaves incorrectly, and 0 otherwise (for example $\chi(n) = 1$ if the read-out data is incorrect for a specific set of values $\{a_\ell(n), 1 \leq \ell \leq 12\}$). Then, the Importance Sampling estimator of the probability p of circuit failure based on the shifts $\mathbf{s} = (s_\ell)_{\ell=1}^{12}$ is given by

$$\hat{p}(\mathbf{s}) = \frac{1}{N} \sum_{n=1}^{N} \chi(n)\ w(n, \mathbf{s})\ , \tag{2.2}$$

where

$$w(n, \mathbf{s}) = \exp\left[- \sum_{\ell=1}^{12} \left(\frac{s_\ell(2a_\ell(n) - 2\mu_\ell - s_\ell)}{\sigma_\ell^2} \right) \right]. \tag{2.3}$$

The non-triviality lies in finding an appropriate mean shift vector $\mathbf{s}$ so that the estimate $\hat{p}(\mathbf{s})$ converges quickly—that is, in very few samples N. The overall cost of estimating p is then the sum of (i) the number of samples required to discover a good shift $\mathbf{s}$, and (ii) the number of samples required to obtain a convergent estimate $\hat{p}(\mathbf{s})$ based on the shift $\mathbf{s}$.

For a uniform characterization of such an estimator, the relative sample variance is employed as the figure of merit (FOM) that describes a predetermined level of confidence and accuracy [5]. The FOM $\rho(\hat{p})$ is

$$\rho(\hat{p}) = \frac{\sqrt{\mathrm{VAR}(\hat{p})}}{\hat{p}}\ , \tag{2.4}$$

where $\mathrm{VAR}(\hat{p})$ is the sample variance:

$$\mathrm{VAR}(\hat{p}) = \frac{1}{N(N-1)} \sum_{i=1}^{N} \chi(n) w(n, \mathbf{s})^2 - \frac{1}{N-1} \hat{p}^2\ . \tag{2.5}$$

The above equation also holds for a nominal Monte Carlo simulation under which $\mathbf{s} = 0$.

To minimize this overall sampling cost an approach, called Spherical Importance Sampling is devised. See Fig. 2.2 for the projection of the proposed algorithm onto a 2D parameter

space. The dashed line is the boundary between passing and failing circuits. We wish to find the point on the boundary closest to the origin in the parameter space, normalized by the standard deviation of each coordinate. The key idea is that Step 1 quickly gives a coarse sense of direction of the appropriate mean-shift vector, and that Step 2 fine-tunes the vector selection across the local neighborhood in the parameter space.

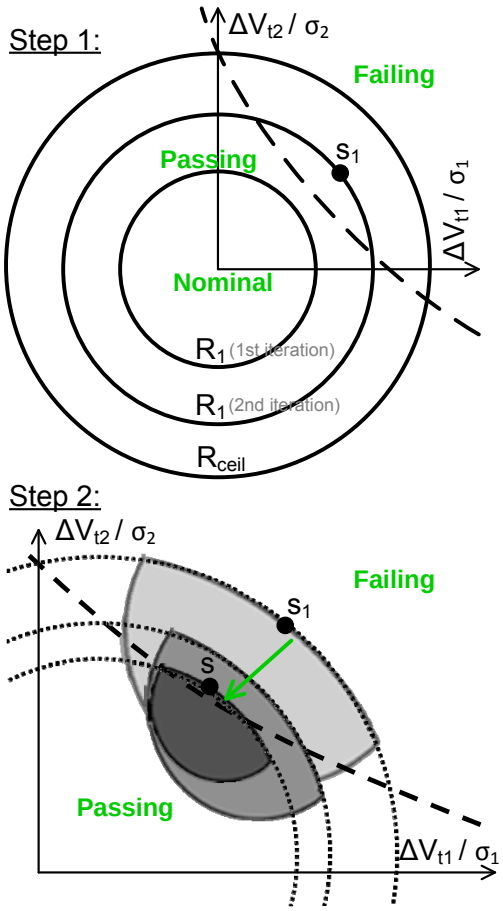


Figure 2.2: Illustration of Spherical Importance Sampling Algorithm. In Step 1, samples on spherical shells quickly find a good starting direction enabling the local exploration in Step 2 to gravitate to a candidate mean shift that exhibits the minimum quadratic norm. Acknowledgment: M. Tikekar for partial figure drafting.

Now, the main steps of the overall procedure are outlined, assuming without loss of generality that all random variables are $\mathcal{N}(0,1)$.[4] Namely, the random variables are scaled prior

[4]The proposed Spherical Importance Sampling method can be extended to the case of non-Gaussian input parameters by replacing the minimization of quadratic norm with the maximization of the joint pdf.

to running the SPICE simulation but after sampling from standard normal distributions.

Step 1. Perform spherical search.

Given the initial tolerable failure floor p_{floor} from a user (e.g., 10^{-12}), initialize the algorithm parameters: $R_{\text{high}} = 2\phi^{-1}(1 - p_{\text{floor}})$, $R_{\text{low}} = 0$, and $N_{\text{iter}} = 0$. Here, $\phi(\cdot)$ is the standard normal CDF. For the allocated complexity parameter N_1, set L and U as the lower and upper bounds on the target number of fails. A good design choice is $U = 0.5N_1$ and $L = 1$.

(a) While $M_f \notin [L, U]$:

- Set $R_1 := \frac{1}{2}(R_{\text{low}} + R_{\text{high}})$
- Sample the radius-R_1 spherical shell N_1 times, and record the total number of failures, M_f.
- If $M_f > U$, set $R_{\text{high}} := R_1$
- If $M_f < L$, set $R_{\text{low}} := R_1$
- $N_{\text{iter}} := N_{\text{iter}} + 1$.

(b) Then:

- Record the final number of iterations as $B_1 = N_{\text{iter}}$.
- Average over the direction vectors associated with all the failing vectors in the last iteration in step (a).
- Initialize $\mathbf{s}_1$ with this average as the starting mean-shift for Step 2.
- Record the quadratic norm of this shift as the current minimum norm, $minNorm = \text{norm}(\mathbf{s}_1)$.

Fig. 2.2 illustrates the iteration in Step 1 over two values of R_1. The second value of R_1 corresponds to a sampling sphere that intersects with the failure region. The failing points on this second sphere are averaged to produce $\mathbf{s}_1$.

Step 2. Perform local exploration.

Let N_2 be the allocated complexity of this step. Set $runCount = 0$, initialize $R_2 = R_1/2$, and $\alpha = (0.05/R_2)^{\frac{1}{N2}}$. The latter parameter governs the gradual increase in the resolution of the local exploration.

While $runCount \leq N_2$:

- Sample uniformly from a spherical distribution of radius R_2, around the point $\mathbf{s}_1$. Call this point $\mathbf{s}_x$.
- If $\text{norm}(\mathbf{s}_x) < minNorm$,
  - Set $runCount := runCount + 1$.
  - Run a SPICE simulation with $\mathbf{s}_x$ as its input. If the simulation results in a failure, record the displacement $d = \text{norm}(\mathbf{s}_1 - \mathbf{s}_x)$ and then update the mean shift vector : $\mathbf{s}_1 = \mathbf{s}_x$. Otherwise $d = 0$.
  - Geometrically shrink R_2, while factoring in the displacement:

$$R_2 := \alpha R_2 + (1 - \alpha)d \ .$$

Fig. 2.2 illustrates Step 2, wherein the local exploration gravitates towards a point with a lower quadratic norm while the resolution simultaneously increases.

Step 3. Run Importance Sampling.

This step is done as per (2.2) with $\mathbf{s} = \mathbf{s}_1$. The value of N_3 is assigned to the number of steps it takes the estimator to reach the FOM value of 0.1.

The overall cost of the proposed approach is then $N_{total} = N_1 \times B_1 + N_2 + N_3$. In previous work [14, 5], the exploration stages employed random, uniform sampling in six dimensions to identify a suitable shifted distribution for Importance Sampling. This uniform sampling approach breaks down in higher dimensions such as 12, which motivates the two-step spherical search. In Section 2.3 we shall see how the proposed procedure can be effectively applied to SRAM circuit design.

2.2 Loop Flattening for Timing variation

The problem of SRAM read access yield has additional challenges beyond the mere fact that very low failure probabilities need to be evaluated. In this section, the problem of statistically analyzing the SRAM read path, which contains circuit blocks repeated at different rates, is

described. Then, the Loop Flattening approximation is introduced and justified to enable the application of accelerated statistical simulation techniques.

In the representative block diagram of an SRAM array of Fig. 2.3(a), there are highly repeated structures: memory cells, sense amplifiers, row decoders and drivers, and timing circuits. There are several distinct, cascaded circuit stages, some of which may be correlated. The circuit is also big. A straightforward way to simulate this circuit is to take a complete schematic and address each location in simulation while noting the behavior for each address location. This method would cost too much computational resources, so a circuit designer turns to a simulation of a critical path by taking a representative member of each group of circuits and adding appropriate parasitic loading in parallel.

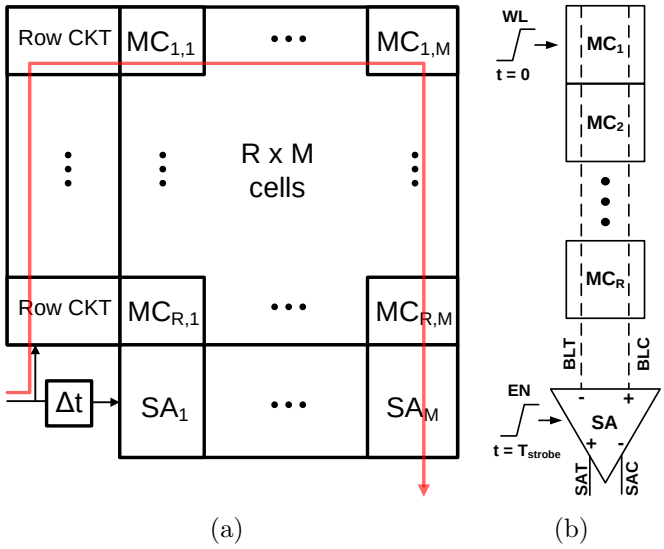


Figure 2.3: (a) a simplified schematic of the small signal read path and (b) a representative SRAM array.

A statistical analysis of a memory critical path requires additional insight into the architecture. For now, consider a single column of 256 memory cells as in Fig. 2.3(b) with $R = 256$. When the wordline goes high at time $t = 0$, the memory cell develops a differential signal (voltage difference between BLT and BLC), and when the enable signal goes high at time $t = T$, the sense amplifier resolves that differential signal to a logic-level output (voltage difference between SAT and SAC). One can represent the bitcell signal of cell i as

$TX_i = T(\sigma_X \tilde{X}_i + \mu_X)$ and the sense amplifier offset as $Y = \sigma_Y \tilde{Y}$ ($\tilde{Y}$ and $\tilde{X}_i$ are $\mathcal{N}(0,1)$). The failure of this read operation is determined by the interaction of two random variables sampled at different rates. The probability P_f that this single column memory fails for a given strobe timing T is the probability that the sense amplifier offset overpowers the bitcell signal for one or more paths in the column:

$$P_f := \Pr\left(\bigcup_{i=1}^{R} \{Y - TX_i > 0\}\right) \tag{2.6}$$

$$\leq \quad R \cdot \Pr\left(Y - TX_1 > 0\right) \quad =: P_u, \tag{2.7}$$

where P_u is the conservative union bound estimate of P_f.

Because of the different rates of repetition, a proper Monte Carlo simulation on a critical path with one cell and one sense amplifier must sample variables in a nested for loop: for each sense amplifier, simulate over 256 cells and check for one or more failing paths, then sample a new sense amplifier and repeat over 256 new cell realizations and so on, as suggested in [17]. If one wishes to apply an accelerated statistical simulation to evaluate the failure of this circuit, the "for loop" sets an unfavorable lower bound on the number of simulations needed just to emulate the architecture. In order to bypass this requirement, there are two conflicting intuitions that one can apply to this circuit: (1) the sense amplifier variation is less severe because of the lower replication rate so the bitcell variation should somehow be amplified in a flattened critical path, or (2) the sense amplifier variation is just as critical as the memory cell variation because every path involves both a memory cell and a sense amplifier.

It is observed that the latter interpretation is more appropriate, thus enabling a simple union-bound estimate for P_f. Hence, only the path failure probability needs to be simulated and the result is multiplied by R. The estimate is guaranteed to be conservative and in practice agrees very well at low levels of failure probability.

Shown in Fig. 2.4, is a numerical evaluation of two interacting normal random variables. As in [31] the bitline signal development and sense amplifier offset are parametrized by

Gaussian random variables—$\{\mu_X = 1mV/ps, \sigma_X = 0.10 \times \mu_X, R = 256, \sigma_Y = 25mV\}$ for the expression in Eq. 2.6. The solid (red) curve gives the failure of a single memory column as determined from a proper Monte Carlo simulation with nested sampling. For 50% pass probability, a $25ps$ strobe timing is needed. For 99% yield, a $75ps$ strobe timing is needed. Such a wide transition in terms of decreasing failure highlights the challenging statistical nature of this problem. The small-dashed (black) curve represents a conventional worst-case methodology. It takes the worst-case bitcell and the worst-case sense amplifier and associates them with a common threshold of signal. Passing is defined by the joint condition of the bitcell having more signal than the threshold and the sense amplifier offset not exceeding the threshold. Because the weakest cell is rarely read by the largest offset sense amplifier, this method is overly conservative.

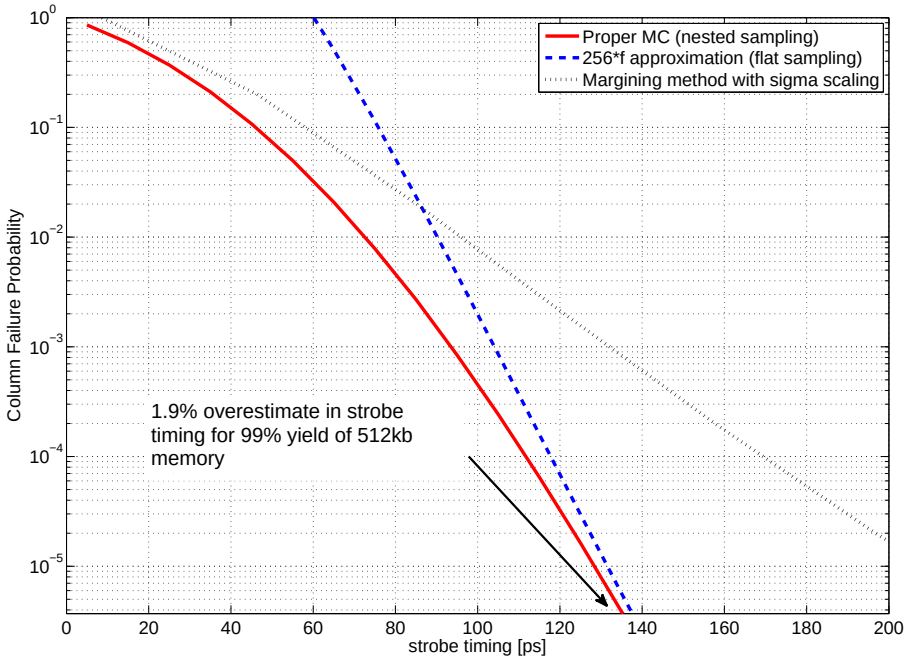


Figure 2.4: A Comparison of three failure estimates for the small signal read path: nominal Monte Carlo (nested sampling), worst-case design, and loop flattening.

The long-dashed (blue) curve represents the loop flattening estimate: failure is obtained from one cell's normal distribution convolved with one sense amplifier's offset, then multiplied

by 256. For a modestly sized 512 kb memory (2048 memory columns) at a yield of 99%, the loop flattening estimate is only 1.9% pessimistic in strobe timing. This estimate gets only tighter at even lower levels of failure. This convergence at low failure probabilities suggests that, for a fixed memory size, the dominant failure mechanism in terms of bitcell signal degradation and sense amplifier offset is independent of the number of cells per bitline. Even in MATLAB, the exhaustive Monte Carlo evaluation took over six hours; whereas, the loop flattened curve was obtained instantaneously as it arises from the sum of two Gaussians. In the experimental results in Section 2.3, the proposed approach exhibits 650X or greater speedup over a Monte Carlo approach, while still using the more generally applicable SPICE solver.

The schematic in Fig. 2.7 is the schematic tree of the large signal read path. For the case of cascaded random delays, one can also see the applicability of the loop flattening estimate. This circuit is simulated in a production quality 45 nm CMOS technology, where each shaded transistor (or gate input) exhibits local mismatch modeled as a fluctuation of its threshold voltage. Fig. 2.6 shows the Monte Carlo SPICE simulation result of this circuit for eight cells per local bitline ($N_{LBL} = 8$) and 16 local evaluation networks ($N_{SEG} = 16$). In this picture, there is a delay Z_i ($1 \leq i \leq 256$) associated with each of the 256 cells on the column and the probability of failure associated with a target delay t is

$$P_f = \Pr\left(\bigcup_{i=1}^{R} \{Z_i \geq t\}\right) \tag{2.8}$$

with $R = 256$. The solid curve gives the proper Monte Carlo simulation result by sampling random variables in proportion to the rate of repetition of their associated transistors. The dashed curve gives the loop flattening estimate in which a simple chain of representative transistors is simulated with all random variables sampled at the same rate. Even for this example, in which the delays are not perfectly normal and delay stages are correlated, the loop flattening technique produces a tight estimate.

The single, solid black dot gives a specific example for how an Importance Sampling (IS) simulation [5] (discussed in detail in Section 2.1) with an appropriately chosen mean shift

can evaluate the loop flattening (dashed curve) with significant speedup, consuming approximately 1.1 thousand SPICE simulations in this example. The loop flattening approximation suggests that this IS estimate in turn will match the result produced by a proper Monte Carlo simulation with nested sampling, which requires 1.7 million SPICE simulations to observe the 1% failure probability of this single column memory.

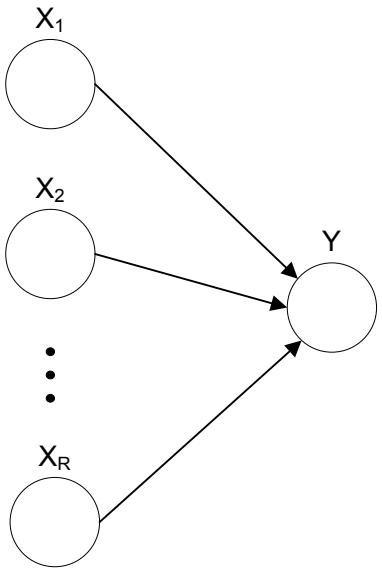


Figure 2.5: A simple tree for analyzing the loop flattening approximation.

For the case of independent, random additive delays, it can be shown analytically that the loop flattening estimate converges to the true failure. Consider the simple tree in Fig. 2.7 where this time the random variables X_i and Y represent delays, and the overall delay of any path is $Z_i = X_i + Y$, associated with the node whose delay is X_i. Then, given a time, t, the failure probability is defined as the probability that one or more paths in the tree exceeds t as in Eq. (2.8). The proposed loop flattening estimate treats these paths as independent at low probabilities and approximates the failure probability with the upper bound:

$$P_u := R \cdot \Pr\left(Z_1 \geq t\right) \ . \tag{2.9}$$

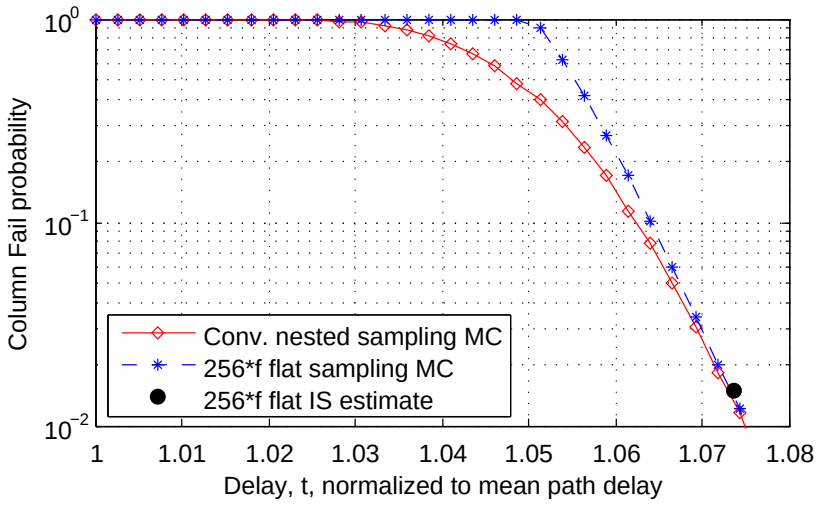


Figure 2.6: Demonstration with SPICE simulation that the loop flattening estimate works in practice for the complicated structure of the large signal read path.

For X_i and Y independent normal random variables, the result in Section 2.5 shows that:

$$\lim_{t\to\infty} \frac{P_u - P_f}{P_f} = 0 \ . \tag{2.10}$$

A similar argument can be developed for the sense amplifier strobe timing case.

For finite t, the ratio in Eq. (2.10) represents the overestimate of the column failure probability as a fraction of the true failure probability ($P_u = P_f(1+\epsilon)$ with $\epsilon \to 0$). For a memory with M independent columns, the overall memory failure estimate from the loop flattening approximation is:

$$1 - (1 - P_f(1+\epsilon))^M \approx 1 - (1 - MP_f - M\epsilon P_f) = MP_f(1+\epsilon).$$

Therefore, the failure of the overall memory is also overestimated by only $\epsilon \times 100\%$. For a variety of cases including numerical examples and a formal proof, the loop flattening estimate has been justified. This approximation eliminates the constraint of nested sampling which would set an unfavorable lower bound on the required number of Monte Carlo simulations. Accelerated statistical methods such as those described in Section 2.1 can be directly applied to quickly determine the value of the loop flattening estimate which produces an accurate,

conservative value for the memory failure probability.

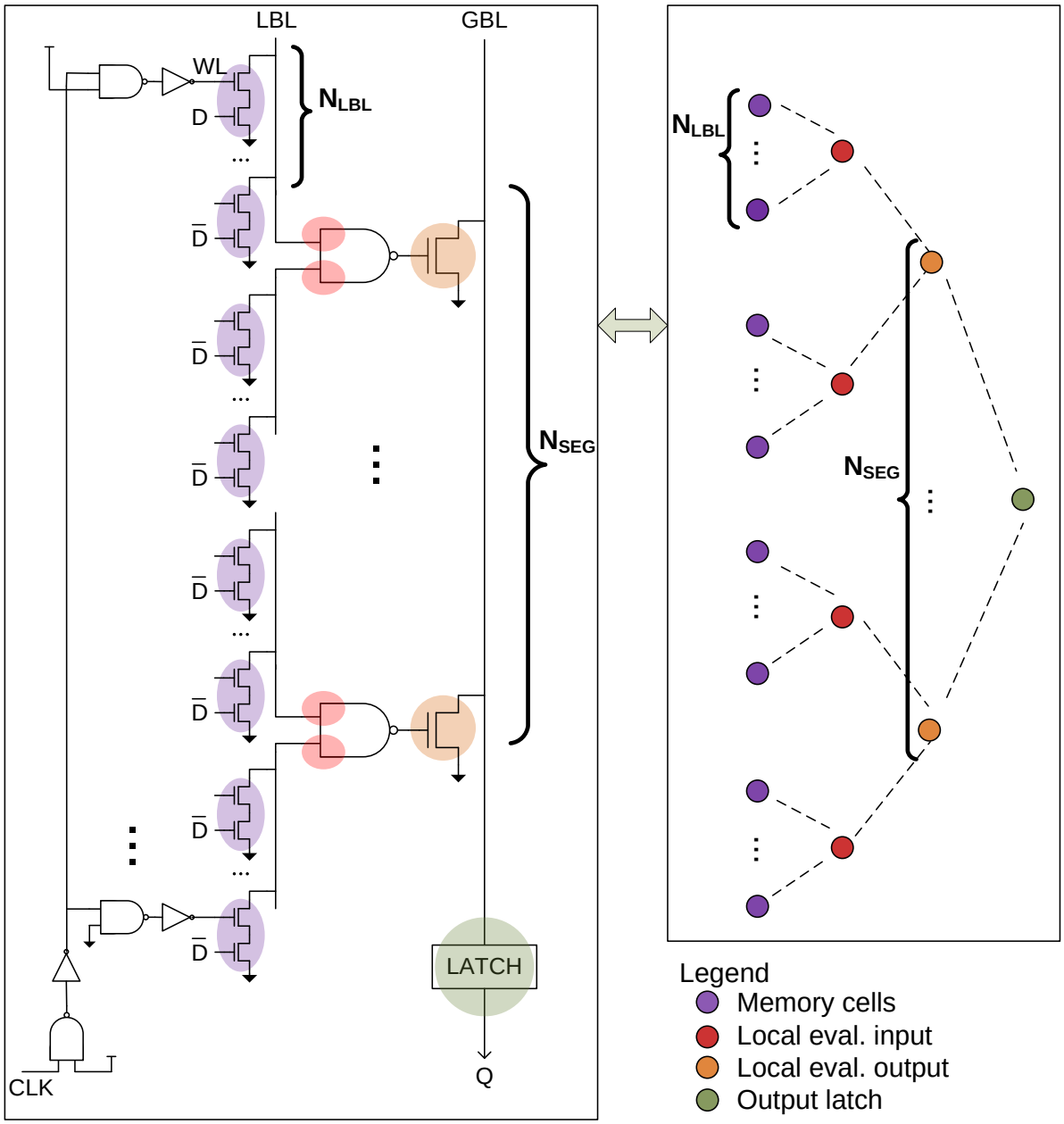


Figure 2.7: Schematic and tree structure of the large signal read path.

2.3 A Detailed Example of the Small-Signal Read Path

In this section the Spherical Importance Sampling method is applied to analyze the small-signal SRAM read path and use the loop flattening estimate to report the read access failure probability. The circuit in Fig. 2.8 is simulated in a generic 45nm technology [32] with High-k Metal Gate 45nm PTM models [33] at the nominal process corner. Based on the measurement of transistor mismatch in [34] the local random variation of transistor threshold voltage is modeled as:

$$\sigma_{V_t} = \frac{1.8mV}{\sqrt{W_{\text{eff}} L_{\text{eff}}}} \ .$$

Furthermore, the memory cell was designed and laid out to exhibit a read instability probability of less than 10^{-9}. The layout of this 6T memory cell in the freePDK 45nm technology is shown in Fig. 2.9. The parasitic extraction of the layout gives the bitline capacitance of $0.24fF/\mu m$ and it also gives the physical cell dimensions of $1.54\mu m \times 0.38\mu m$. The cell device sizings (W/L) for the pull-up, pull-down, and access transistors are 85 nm/50 nm, 100 nm/50 nm, and 75 nm/50 nm respectively to satisfy the bitcell stability requirement. The sense amplifier and column multiplexer devices were sized to occupy the area of roughly eight memory cells.

The waveforms of a nominal Monte Carlo simulation with an aggressive timing are shown in Fig. 2.10. About $40mV$ of highly variable bitline signal (BLT - BLC) interacts with the sense amplifier offset to produce a logic level on the output of the sense amplifier (SAT - SAC). A larger strobe time setting, t_{STR}, allows more bitline signal to develop and therefore reduces the probability of incorrect output. The complex nature of the waveforms shows how no suitable performance metric can be defined to facilitate analytical modeling or even to properly formulate numerical derivatives in the space of variation parameters. Therefore, techniques based on sensitivity analysis [27], hypersphere or directional cosine searches [29], or approximations for simplified Monte Carlo simulation [17] can significantly compromise accuracy.

On the other hand, the processing of the read path simulation results leads unambiguously

to a binary indicator of pass/fail, and, in general, the formulation of a binary indicator is feasible for arbitrary circuits. In order to evaluate the probability of incorrect read-out versus strobe time setting, the proposed Spherical Importance Sampling algorithm is applied directly to this complicated simulation as it requires nothing more than a binary simulation output—1 if the sense amplifier output is incorrect and 0 otherwise.

Shown in Fig. 2.11 is the evolution of the failure probability estimate versus simulation run index for both the Importance Sampling stage (step 3) of the proposed algorithm and nominal Monte Carlo. The strobe time setting of $40ps$ results in a path failure probability of $1.01 \cdot 10^{-4}$ which, by the loop-flattening technique described in Section 2.2, results in a column failure of $5.2 \cdot 10^{-2} = 128 \cdot 4 \cdot 1.01 \cdot 10^{-4}$. The raw speed-up of step 3 is 1800X at a common level of confidence ($\rho = 0.1$), and taking into consideration the cost of step 1 and step 2, the total speedup is 650X. This reflects a 4.3X speedup over the work in [5] at the same level of failure probability, *and with twice the dimensionality—12 instead of 6.* The specific algorithm parameters used for all runs were: $p_{floor} = 10^{-12}$, $N_1 = 500$, $N_2 = 500$.

Shown in Table 2.2 is a summary of simulation results for four strobe time settings. One million Monte Carlo runs takes seven days on one Linux workstation utilizing 1 Intel 3.2GHz CPU with 16GB of RAM; therefore, verification with a SPICE Monte Carlo benchmark for failure probabilities of 10^{-5} or lower is infeasible as it requires ten million or more runs. For strobe timings of $50ps$, $65ps$, $80ps$, the speed-up is compared against a conservative projection of required Monte Carlo trials with the expression $100/p_{\text{path}}$. As additional verification, the Spherical Importance Sampling algorithm was verified at these failure levels with a linear function in MATLAB, and the step 3 Importance Sampling run was extended 3X beyond the required number of runs to verify stable convergence of the failure estimate. Table 2.2 shows how the simulation cost gradually increases with exponentially lower levels of failure probability, achieving a speed-up of over twenty million at $p_{path} = 1.91 \cdot 10^{-9}$.

The increasing cost comes from needing more runs in step 1 to find a suitable sampling shell radius, and from needing more runs in step 3 to achieve the desired level level of confidence. Generally, no more than two radius trials were required to find a useful direction

Table 2.2: Summary of S-IS simulation results on SPICE simulation of the read column in Fig. 2.8. The path failure probability is multiplied by $512 = 128 \cdot 4$ to produce the overall column failure. Speed-up over nominal Monte Carlo is 650X or higher. Values marked with †are conservative projections on the number of Monte Carlo using $100/p_{\text{path}}$.

t_{STR}	p_{path}	cost	speed-up	p_{col}
40ps	$1.01 \cdot 10^{-4}$	1534	$6.50 \cdot 10^{2}$	$5.2 \cdot 10^{-2}$
50ps	$9.08 \cdot 10^{-6}$	1660	$6.63 \cdot 10^{3}$†	$4.6 \cdot 10^{-3}$
65ps	$1.33 \cdot 10^{-7}$	2214	$3.40 \cdot 10^{5}$†	$6.8 \cdot 10^{-5}$
80ps	$1.91 \cdot 10^{-9}$	2423	$2.16 \cdot 10^{7}$†	$9.8 \cdot 10^{-7}$

for the initialization of the local exploration in step 2. As an example, Table 2.3 shows the evolution of the mean shift after completing spherical sampling ($\mathbf{s}_1$) and then after completing local exploration ($\mathbf{s}$), when evaluating the strobe timing of $80ps$ (failure probability around $2 \cdot 10^{-9}$.

Table 2.3: Evolution of mean shift vector after spherical sampling (step 1) and local exploration (step 2) for strobe timing of $80ps$.

	Sense Amplifier				Col. Mux		Memory Cell						
shift	ΔV_{t1}	ΔV_{t2}	ΔV_{t3}	ΔV_{t4}	ΔV_{t5}	ΔV_{t6}	ΔV_{t7}	ΔV_{t8}	ΔV_{t9}	ΔV_{t10}	ΔV_{t11}	ΔV_{t12}	L_2 norm
$\mathbf{s}_1$	-0.89	-1.45	-3.61	5.59	0.03	-0.93	-3.29	0.19	-3.58	6.27	-0.53	-0.28	10.55
$\mathbf{s}$	-0.65	0.35	-3.60	3.27	0.20	0.00	-0.02	0.08	-0.13	0.69	0.18	3.24	5.90

The first row shows how the spherical sampling gets a coarse direction correct along a few key dimensions (1, 3, 4, 8) after 1000 simulations, and then the local exploration greatly improves the accuracy of the shift both in terms of magnitude and direction after an additional 500 simulations. The resulting shift $\mathbf{s}$ in the second row matches circuit insight: 1) the column multiplexer devices have little influence; 2) devices associated with the non-discharging side (BLC) in Fig. 2.10 have small magnitude; 3) the read stack in the memory cell (10,12) is significantly weakened; 4) the mismatch between NMOS devices (3,4) in the sense amplifier is most critical. Going from $\mathbf{s}_1$ to $\mathbf{s}$, the mean shift for the Importance Sampling run, took only 500 trials with an initial exploration radius (R_2) of 5.25 tapering down to 0.06. As discussed in Section 2.1, the local exploration radius can at most shrink down to a magnitude of 0.05. The actual value of the final radius being close to 0.05,

indicates that most of the local exploration runs resulted in a small displacement. A larger final radius would have suggested the need for a larger complexity parameter N_2. Finally, having invested 1,500 samples to identify a good mean shift, the final Importance Sampling simulation took an additional 923 simulations to produce a high confidence estimate with $\rho = 0.1$ even though the failure probability was $1.91 \cdot 10^{-9}$.

A closer look at the simulation cost breakdown is presented in Table 2.4. Across a range of probabilities from $9 \cdot 10^{-6}$ to $2 \cdot 10^{-9}$, the exploration cost of Spherical Importance Sampling varied from 1000 to 1,500 simulation runs. The subsequent Importance Sampling stage took 660 to 923 runs. Compared to the previous ISNM work [5], at half the dimensionality (6 instead of 12) and higher probability levels, the simulation cost was higher. This work's improvement comes from the two-step spherical exploration finding a better shift for the Importance Sampling run. It is also worth highlighting that the local exploration in step 2 is computationally less costly than the directional search in step 1. The two-stage Spherical search effectively handles the dimensionality of 12, considering that over 4,000 simulations are needed just to check all the corners of a 12D cube. With Spherical Importance Sampling much fewer directions are examined while still identifying a suitable mean shift.

Table 2.4: Comparison of simulation cost between this work and [5]

	This Work 12 Dimensions 2-step Spherical Samples			[Dolecek 2008 ICCAD] 6 Dimensions Uniform Exploration		
P	9.08×10^{-6}	1.33×10^{-7}	1.91×10^{-9}	4.9×10^{-3}	4.4×10^{-4}	3.0×10^{-6}
Step 1	500	1000	1000	-	-	-
Step 2	500	500	500	-	-	-
Total Exploration	1000	1500	1500	1000	1000	2000
IS run	660	714	923	1000	2000	2000
Total	1660	2214	2423	2000	3000	4000

A general comparison across a variety of simulation methods [5, 14, 35, 16, 2] is presented in Fig. 2.12. The y-axis gives failure probability levels and the horizontal axis gives the number of total circuit simulations (e.g., SPICE runs) to evaluate the failure probability.

Also indicated is the dimensionality of the problem (number of random variables). All methods require less than 10,000 circuit evaluations to identify failure probabilities from 10^{-3} to 10^{-9}, and the relation between failure probability and number of simulation trials is much steeper than the Monte Carlo behavior of $\approx 100/p$. The Spherical Importance Sampling method compares favorably with other works. Not plotted on the graph is the result in [2] which evaluates a failure probability of $1.8 \cdot 10^{-7}$ in 300,000 circuit evaluations for a 24-dimensional parameter space. Indeed, Monte Carlo simulation is much less sensitive to dimensionality than the accelerated statistical evaluation techniques in Fig. 2.12 which all rely on some type of classification of the parameter space. Developing an accelerated simulation method for a higher dimensional parameter space (e.g., 50) will broaden the applicability of quick statistical simulation techniques.

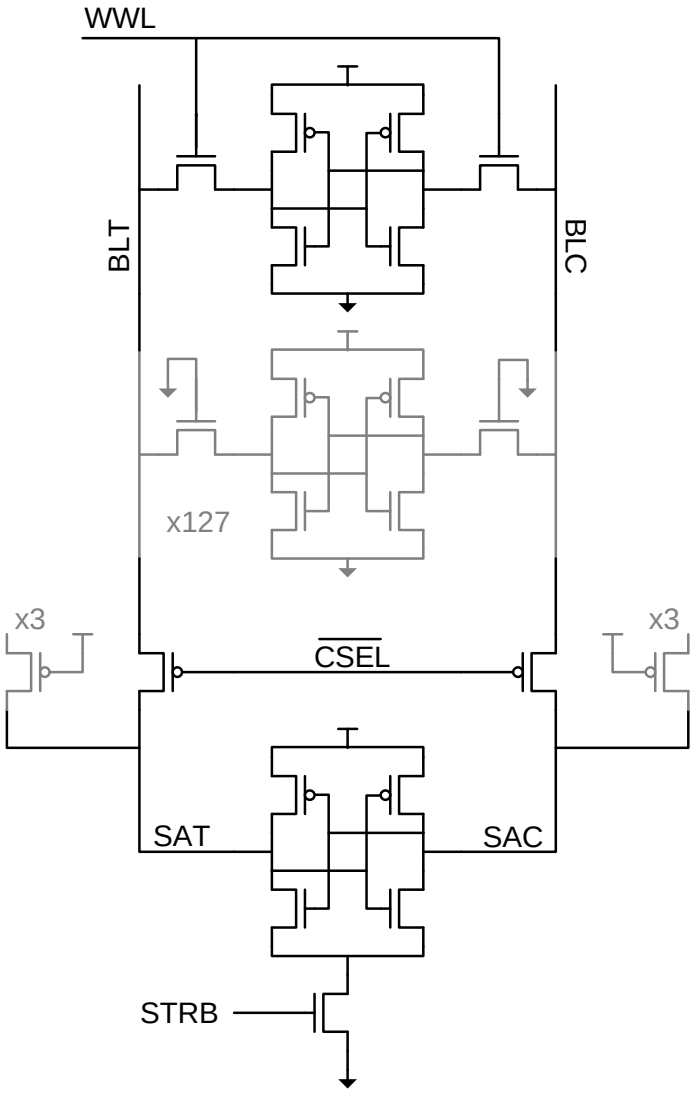


Figure 2.8: Transistor level schematic of a representative small-signal (sense amplifier -based) memory column with 128 cells per bitline and additional multiplexing of 4 columns per sense amplifier. Precharge devices are not shown for clarity.

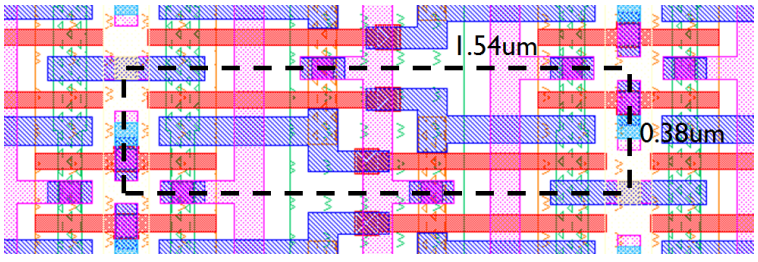


Figure 2.9: The layout of the 6T memory cell in the freePDK 45nm technology. Extraction from layout reveals 0.24 fF/μm of bitline capacitance. Courtesy of Mehul Tikekar.

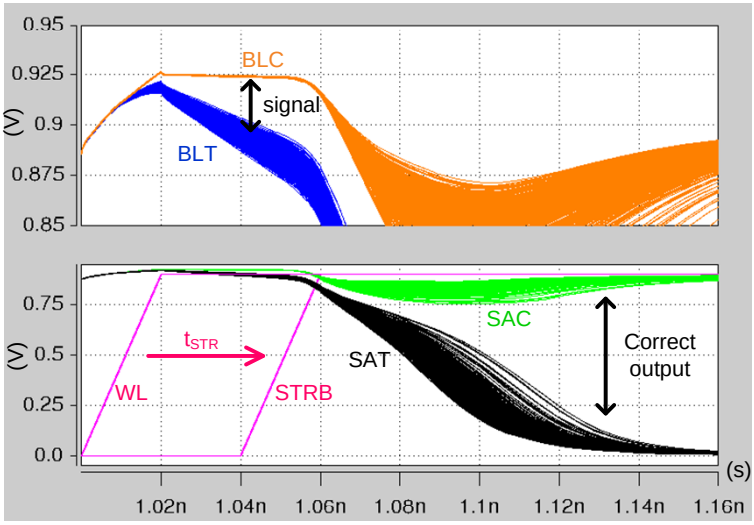


Figure 2.10: Operational waveforms of the small-signal read column illustrate the pronounced effect of device variation and the difficulty in defining a performance metric conducive to analytical modeling or even the formulation of numerical derivatives.

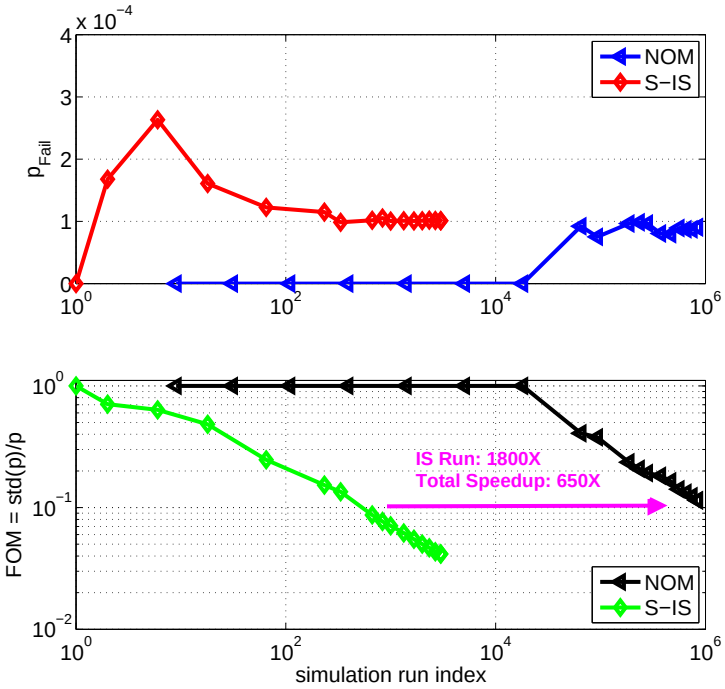


Figure 2.11: Evolution of failure probability estimate from Spherical Importance Sampling (S-IS) compared with nominal Monte Carlo for a strobe time setting of $40ps$.

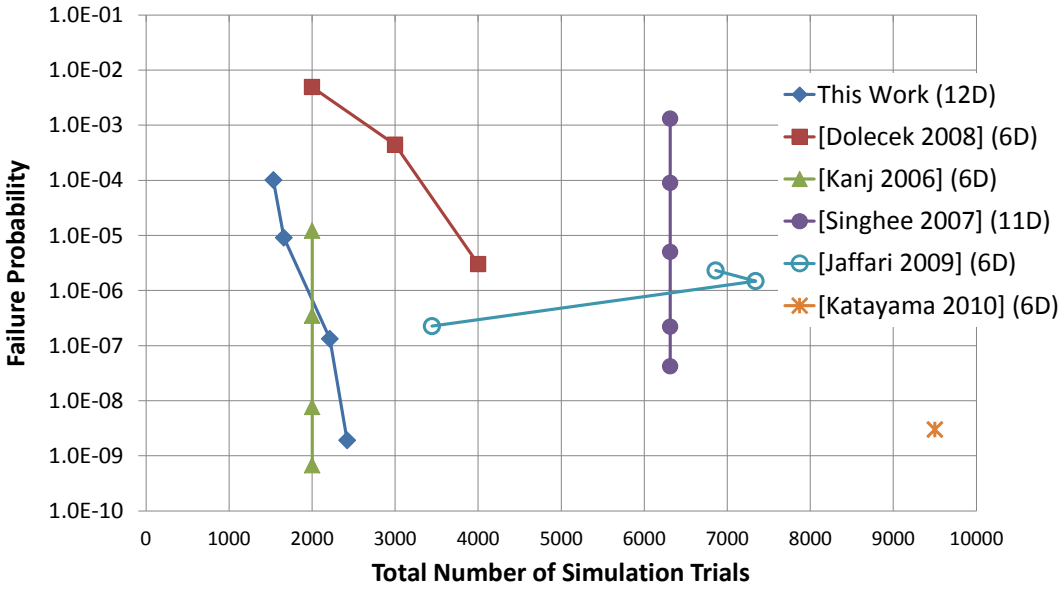


Figure 2.12: Simulation cost comparison. Not shown is a point for evaluating a failure probability of $1.8 \cdot 10^{-7}$ in 300,000 simulations in a 24 dimensional space by [2].

2.4 Concluding Remarks on SRAM Statistical Analysis

In this chapter two techniques were presented—*Spherical Importance Sampling* and *Loop Flattening*—and a method to synthesize them to reduce the statistical analysis of an SRAM block to an Importance Sampling simulation of a chain of component circuits. The key challenge of searching for the most likely failure mechanism in a high dimensionality (12 in this work) parameter space is addressed by a two-stage process in which a coarse direction is obtained first, followed by a local sampling of increasing resolution.

It has been demonstrated that when the failure probability is small, the statistical analysis of a system amounts to evaluating probabilities of simple constituent structures and summing them together. This observation greatly simplifies the analysis and offers a cost-effective alternative to the nominal nested approach. As a contrast with prior SRAM yield analysis, the consideration of intermediate metrics (bitline signal, sense amplifier offset) was replaced by a full-scale SPICE simulation requiring only the indication of pass or fail. As future work, this method can be extended to the complete, global row and column path of large embedded SRAM, as well as to other highly structured circuits such as adders, FIR filters, and FFT accelerators. Such highly symmetric, multiplexing structures will become more prevalent in the ascent of multi-core chip design.

2.5 Proof of the Loop Flattening Approximation Convergence for Independent Additive Delays

Here, Eq. (2.10) is shown for the case where X_i and Y are $\mathcal{N}(0,1)$ and independent additive delays ($Z_i = X_i + Y$).[5] Recall that failure is given by Eq. (2.8) and the conservative loop flattening estimate P_u is defined in Eq. (2.9).

[5]This proof assumes zero mean standard normal variables. The same arguments can be applied for the general case of non-zero mean normal random variables of different variance and will hold for generally well-behaved distributions that decay sufficiently fast.

By the union bound,

$$P_f := \Pr\left(\max_{1\le i\le R} Z_i \ge t\right) = \Pr\left(\bigcup_{i=1}^{R} Z_i \ge t\right)$$
$$\le \sum_{i=1}^{R} \Pr\left(Z_i \ge t\right) = R\cdot \Pr\left(Z_1 \ge t\right) =: P_u \ .$$

Then, by incorporating the pair-wise intersection probabilities, we can introduce an optimistic (lower bound) estimate P_x:

$$P_f = \Pr\left(\bigcup_{i=1}^{R} Z_i \ge t\right)$$
$$\ge \sum_{i=1}^{R} \Pr\left(Z_i \ge t\right) - \frac{1}{2}\sum_{\substack{(i,j)\\ i\ne j}}^{R} \Pr\left(Z_i \ge t \bigcap Z_j \ge t\right)$$
$$\ge P_u - \frac{R(R-1)}{2}\Pr\left(Z_1 \ge t \bigcap Z_2 \ge t\right) =: P_x \ .$$

This implies

$$\frac{P_u - P_f}{P_f} \le \frac{P_u - P_x}{P_x} = \frac{1}{\frac{2}{R-1}\frac{\Pr(Z_1\ge t)}{\Pr(Z_1\ge t \bigcap Z_2\ge t)} - 1} \ . \tag{2.11}$$

The following well-known bound on the tail probability of the standard Gaussian (let $W \sim \mathcal{N}(0,1)$, $w > 0$) is used:

$$\frac{1}{\sqrt{2\pi}w}\left(1 - \frac{1}{w^2}\right)e^{-\frac{w^2}{2}} < \Pr\left(W \ge w\right) \le \frac{1}{2}e^{-\frac{w^2}{2}}. \tag{2.12}$$

Since $\mathrm{var}(Z_i) = 2$,

$$\frac{1}{\sqrt{\pi}t}\left(1 - \frac{2}{t^2}\right)e^{-\frac{t^2}{4}} < \Pr\left(Z_1 \ge t\right) \ . \tag{2.13}$$

Now the intersection probability $\Pr\left(Z_1 \ge t \bigcap Z_2 \ge t\right)$ is examined. Conditioning on $Y = y$, the two events $Z_1 \ge t$ and $Z_2 \ge t$ are independent. Therefore the joint probability can be

written as:

$$\begin{aligned}\Pr\left(Z_1 \geq t \bigcap Z_2 \geq t\right) &= \int_{-\infty}^{\infty} f_Y(y) \left[\Pr\left(X_1 + y \geq t\right)\right]^2 dy \\ &\leq \Pr\left(Y \leq 0\right) \left[\Pr\left(X_1 \geq t\right)\right]^2 + \\ &\quad \int_0^t f_Y(y) \left[\Pr\left(X_1 + y \geq t\right)\right]^2 dy + \Pr\left(Y \geq t\right). \end{aligned} \tag{2.14}$$

The inequality above follows from partitioning the region of integration into $(-\infty, 0)$, $(0, t)$, and (t, ∞). For the first and third terms, the maximum value for $\Pr\left(X_1 + y \geq t\right)$ is substituted and $f_Y(y)$ is integrated. The first term is bounded by $\frac{1}{8} \exp -t^2$ and the third term is bounded by $\frac{1}{2} \exp -\frac{t^2}{2}$. The middle term can be bounded through bounds on the integrand:

$$\begin{aligned}&\int_0^t f_Y(y) \left[\Pr\left(X_1 + y \geq t\right)\right]^2 dy \\ &\leq \int_0^t \frac{1}{2} e^{-\frac{y^2}{2}} \frac{1}{4} e^{-(t-y)^2} \leq \frac{t}{8} \cdot \max_{y \in (0,t)} e^{-\frac{y^2}{2}} e^{-(t-y)^2} \ . \end{aligned} \tag{2.15}$$

From elementary calculus, the right hand side bound evaluates to $\frac{t}{8} e^{-\frac{t^2}{3}}$. This bound on the middle term decays the slowest, and therefore the ratio $\Pr\left(Z_1 \geq t\right) / \Pr\left(Z_1 \geq t \bigcap Z_2 \geq t\right)$ grows at least as fast as $\frac{1}{t^2} e^{\frac{t^2}{12}}$, which grows arbitrarily large, causing the right-hand side in Eq. (2.11) to go to zero and in turn verifies the limit in Eq. (2.10).

Chapter 3

SRAM Design for Voltage Scaling

There is a need for large embedded memory that operates over a wide range of supply voltage compatible with the limits of static CMOS logic that also minimizes standby power. Fig. 3.1 illustrates the SRAM voltage scaling challenge by plotting the reported 6T SRAM die-level or macro-level minimum operating voltage, V_{min}, versus technology node for the past five years. As semiconductor process technology continues to scale, the increased impact of process variation inhibits the reduction of supply voltage. Yet, the reduction of supply voltage has several benefits [36, 37]: it relaxes the thermal constraints on power dissipation in enterprise applications; it improves energy efficiency in battery operated applications; and it preserves the reliability of deeply scaled transistors.

As a solution, this chapter presents circuit solutions to voltage scaling in SRAM for both active operation and standby mode, embodied in a voltage scalable single-supply 8T SRAM with no dynamic voltage assists that minimizes both the area and standby power. It addresses the design challenges related to area efficiency and process variation with three contributions: 1) an AC coupled sense amplifier (ACSA) that operates down to a power supply ultimately limited by the worst-case bit line on/off current ratio; 2) area efficient, regenerative driving of long data lines to permit bidirectional signaling on a single metal 4 wiring track on the memory cell column pitch; 3) a data retention voltage (DRV) sensor to aggressively reduce supply voltage to the limit set by local mismatch, without corrupting

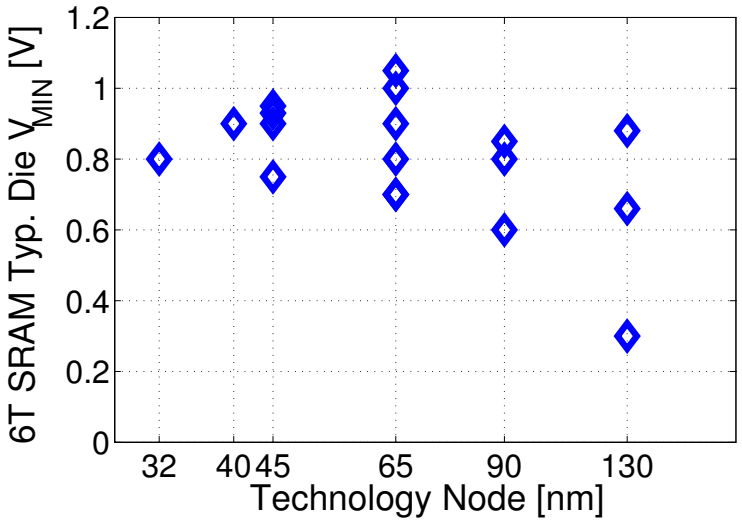


Figure 3.1: The SRAM voltage scaling barrier: reported 6T SRAM V_{min} versus technology node (source: ISSCC & VLSI 2004—2009). Array sizes range from 64kb to 153Mb.

the contents of the memory.

The remainder of this chapter discusses in detail a 512kb 8T SRAM macro realization in 45nm SOI CMOS. In section 3.1 the proposed sensing scheme is introduced, compared to conventional methods, and characterized by physical measurement. In section 3.2 follows an exposition of the DRV sensor along with experimental corroboration from both the sensor and functional array. Finally, section 3.3 provides a commentary on future directions and summarizes the key characteristics of the prototype as a viable candidate for SRAM voltage scaling.

3.1 Proposed Sensing Scheme for Voltage Scaling

Conventional sensing approaches fall into two categories. For density, 6T memory cells are placed on a long bitline of 128 to 256 cells, Fig. 3.2(a). When accessed, the small differential swing is converted to a logic-level by a sense amplifier at a preset strobe time. For performance, 6T or 8T memory cells are partitioned hierarchically among short single-ended bitlines. As a result, the weak and variable cell needs only to discharge a small capacitance, which is buffered to a global bitline pull-down device through a static or dynamic NAND

gate, Fig. 3.2(b). The proposed sensing scheme targets the benefit of density with long bitlines and small signal sensing in the context of a single-ended hierarchical read path, enabling the coverage of 4096 rows before the data must be latched thereby lowering the overall latency, Fig. 3.2(c).

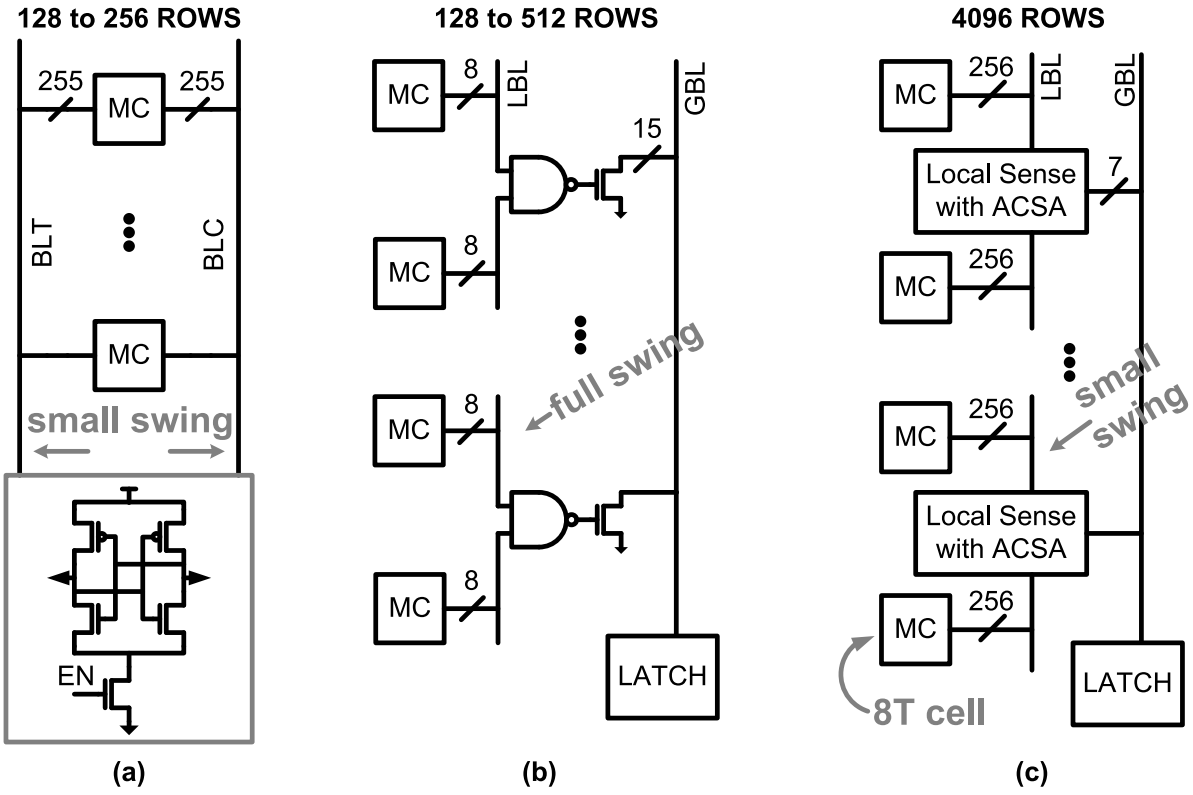


Figure 3.2: Illustration of (a) the conventional small swing sensing scheme, (b) the conventional hierarchical read path with full swing sensing, and (c) the proposed hierarchical small swing read path supporting a significantly larger number of rows.

This approach and the choice of the 8T bitcell sets the design challenges. A good small swing single-ended sense amplifier is required. In addition, the area of the sensing network is critical because it cannot be amortized over multiple columns. 8T memory cells belonging to different words must not be interleaved in order to avoid dummy read disturbance on unselected cells during a write operation. Finally, as the supply voltage is reduced, the read path must tolerate the increased impact of process variation.

3.1.1 The ACSA for local sensing

Shown in Fig. 3.3 is the column path for a 256x128 half-bank of 32kb. The ACSA supports a local bitline of 256 8T memory cells. When the data in the accessed memory cell Q_n is high, the assertion of RWL_n initiates a discharge on the local bitline, LBL. PMOS M4 detects this discharge and rapidly charges the low capacitance dynamic output from $0V$ to VDD. Otherwise, if Q_n is low, the output remains stable at $0V$.

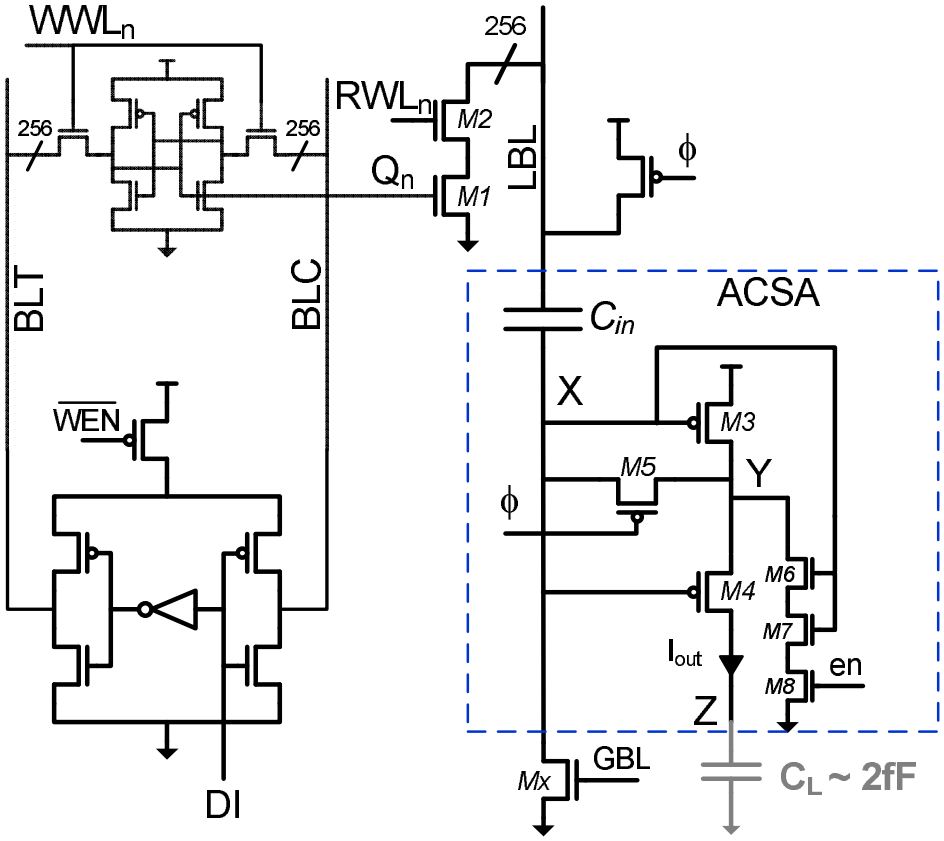


Figure 3.3: Schematic of the local column path in a 32kb half-bank. The local output node Z connects to a buffering inverter and another ACSA, corresponding to approximately $2fF$ of load.

The salient feature of the ACSA is the thick oxide MOS coupling capacitor. It stores a PMOS threshold voltage drop in series with the bitline signal while maintaining a coupling ratio close to 1 and occupying $4\mu m^2$. Equalization PMOS M5 generates the voltage drop across the capacitor by diode-connecting the amplifying PMOS M3 while keeping the sensing

PMOS M4 in cutoff with $0V$ of V_{GS} to ensure a stable output. Finally, the weak NMOS pull-down stack M6–M8 biases M3 close to a level of current that can quickly charge the internal node Y and sustain the charging of the dynamic output Z. To save static power, the pull down network is disabled when not reading. To provide additional margin and further save static power in unselected half-banks during read, the pull down network is enabled only during the evaluate phase in the read cycle for 15 out of 16 half-banks. As a result, the non-bitcell static power during a read operation is dominated by the ACSAs in only one half-bank.

The simulation waveforms at a supply of $0.6V$ in Fig. 3.4 illustrate the amplification when reading 1 and stability when reading 0 of the ACSA. During precharge ($\phi = 0$), the internal nodes X and Y are initialized to roughly $250mV$ below VDD. Upon exiting precharge, RWL is asserted and when sensing 1, a discharge on the local bitline couples to the internal node X. This causes a rapid rise of the internal node Y as a result of amplification from device M3. After $100mV$ of signal development, the rapid separation between X and Y cause device M4 to charge the local output node Z high followed by a rise on the global bitline (GBL). When sensing 0, the RWL is asserted and the local bitline droops at a slower rate. Therefore, the separation in X and Y takes longer, and no false pulse on the global bitline appears within $2ns$.

The ACSA is a dynamic circuit and insight into its operation can be obtained by examining the transfer characteristics from input voltage at node X to output current at node Z plotted in Fig. 3.5(a). First considering just a single PMOS device, as employed in the conventional domino read path [38], the performance benefit of dynamic sensing is made clear. The subthreshold characteristic of the transistor separates a "1" from a "0" with a one hundred-fold increase in output current over an input range of only $200mV$. The ACSA exhibits an even steeper transfer characteristic by stacking device M4 in series with the inverting amplifier formed by PMOS M3 and its associated pull-down stack. Therefore, when sensing a "1," less than $100mV$ of signal traces a steep four-orders of magnitude rise in output current. When sensing a "0," the coupling of ϕ through C_{gd} and charge injection,

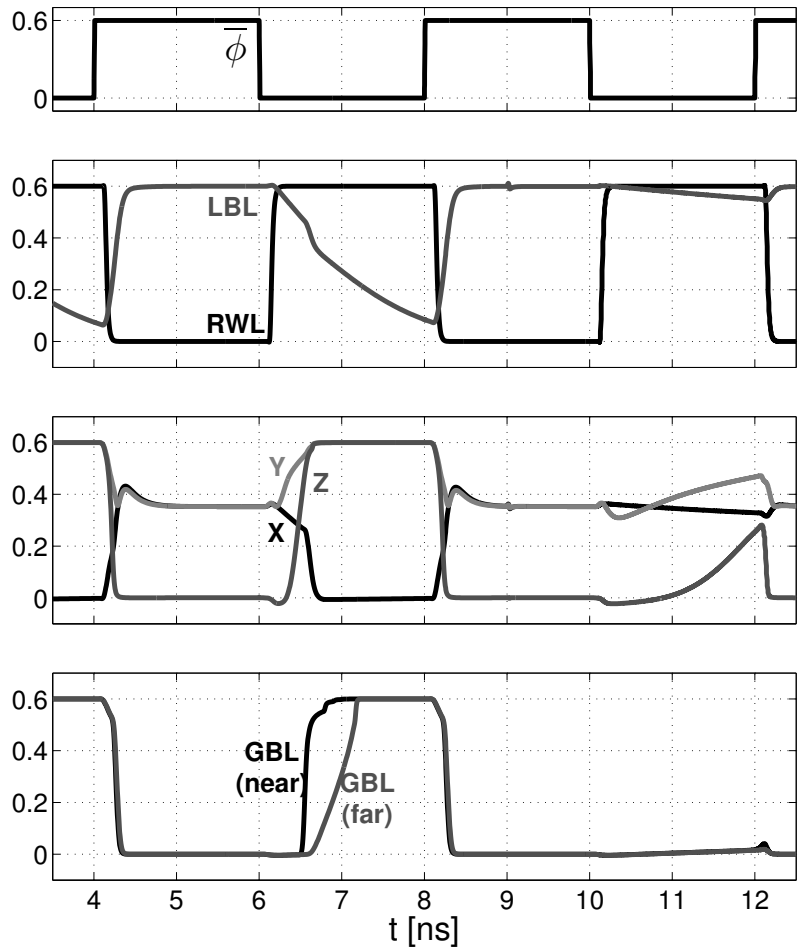


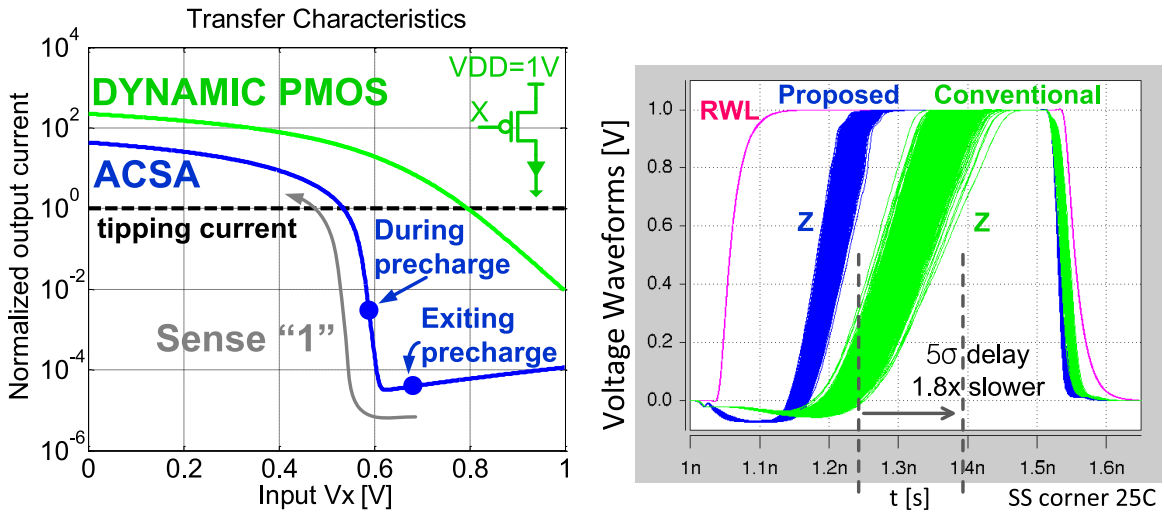
Figure 3.4: ACSA operational waveforms at 0.6V. Read "1" followed by read "0" (FF 85C).

traces the steep current versus voltage transfer characteristic backwards by $30mV$ to $50mV$ (depending on VDD) to guarantee a very low initial output current.

The performance improvement of the ACSA extends from the nominal sense to the worst-case realization under local mismatch variation. By storing the variable trip point of the inverting amplifier based on M3 during precharge, the output device M4 is guaranteed to begin in cutoff and its gate to source voltage will develop according to:

$$V_{SG4} = \Delta V_{LBL} \cdot \frac{C_{in}}{C_{in} + C_{parx}}(1 + g_{m3}r_o) \approx \Delta V_{LBL} \cdot (1 + g_{m3}r_o). \tag{3.1}$$

Therefore, the required nominal gate-to-source overdrive plus worst-case local threshold volt-



(a) Comparison of static transfer characteristics (TT 25C)

(b) Comparison of transient characteristics (SS 25C)

Figure 3.5: Comparison of the proposed ACSA to the conventional domino read path based on dynamic PMOS local evaluation networks for a long bitline of 256 cells.

age deviation are both suppressed by factor of $1 + g_{m3}r_o \approx 5$ with respect to the required bitline signal. The term C_{in} represents the input coupling capacitance and C_{parx} represents the parasitic capacitance to AC ground at the internal node X. The capacitive divider ratio in Eq. 3.1 is expected to be 80% to 90%. The resulting performance benefit is illustrated by the waveform plots in Fig. 3.5(b). The 5σ delay at a slow corner of the ACSA is 1.8x less than the delay of the conventional technique based on sensing with a dynamic PMOS inverter.

Offset compensation provides dual benefits of reducing delay and lowering V_{min}. Without consideration of the sensing network there is a fundamental limit related to the worst-case "on" (Fig. 3.6(a)) to worst-case "off" (Fig. 3.6(b)) current ratio of a long bitline of 256 cells. In order to distinguish the two data states in a dynamic fashion (no keeper on the bitline), there must exist a sampling window between the arrival time of the slowest "true 1" (t_{true}) and the fastest "false 1" (t_{false}). In Fig. 3.7, Monte Carlo simulation of these delay histograms are shown for both sensing the bitline through a PMOS and sensing the bitline through the proposed ACSA. At 1V (Fig. 3.7(a)), the PMOS exhibits slower performance

and wider spread in delay, but robust sampling windows exist for both sensing schemes. At 0.55V (Fig. 3.7(b)), the PMOS t_{true} distribution overlaps with the t_{false} distribution. However, the bitline on/off current variation still exhibits a separation, and the ACSA is able to successfully distinguish this separation as illustrated by its histograms of t_{true} and t_{false}. This fully dynamic scheme offers superior performance compared to an approach that eliminates the "false 1" with a keeper [39], shifting the design challenge from keeper sizing to developing a timer circuit for the SRAM array that can sample the data at the correct instant. Such a timing circuit must be employed in conjunction with the proposed ACSA, whose benefit is the widening of the dynamic sampling window.

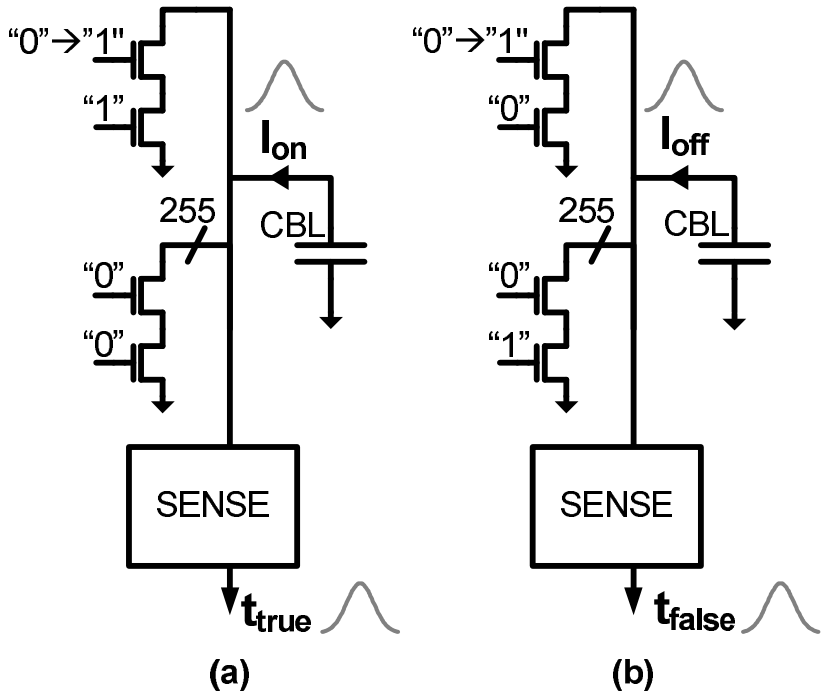


Figure 3.6: Shown is (a) the schematic and worst-case data state for an "on" bitline with an associated delay for a "true 1," and (b) the schematic and worst-case data state for an "off" bitline with an associated delay for a "false 1."

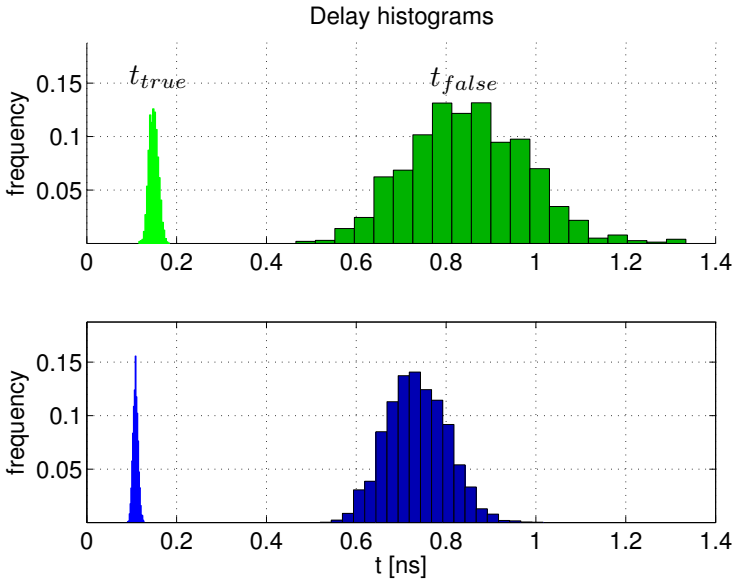
At process conditions of FF 85C, Monte Carlo simulation results in Fig. 3.8(a) show that the (5σ) ratio of bitline current degrades to 1 at $0.45V$, resulting in zero voltage difference between the weakest "1" bitline and the strongest "0" bitline. A sensing network that itself suffers from local variation, cannot achieve this fundamental limit to supply voltage. The

spread in the trip point of the sensing network will require a greater separation in the worst-case on to off bitline current. Also shown in the plot are simulation results for sensing with a dynamic PMOS. The ratio of t_{false} to t_{true} is significantly degraded and reaches 1 around 0.7V. On the other hand, the offset-compensated ACSA preserves a wider separation between worst-case t_{false} and t_{true}, degrading to a ratio of 1 around 0.5V, only $50mV$ above the ultimate limit set by the bitline current ratio. The change in minimum supply voltage across temperature and bitline length of the ACSA based read path is characterized in Fig. 3.8(b) at a constant array size of 512kb. The y-axis shows V_{unity}, defined as the supply voltage at which the sampling window between t_{true} and t_{false} disappears. Reducing the bitline length enables a lower supply voltage by reducing the impact of leakage from "off" cells. Because of leakage, the worst-case temperature and required supply voltage constrain the longest possible bitline, and, therefore, the area efficiency of the memory.

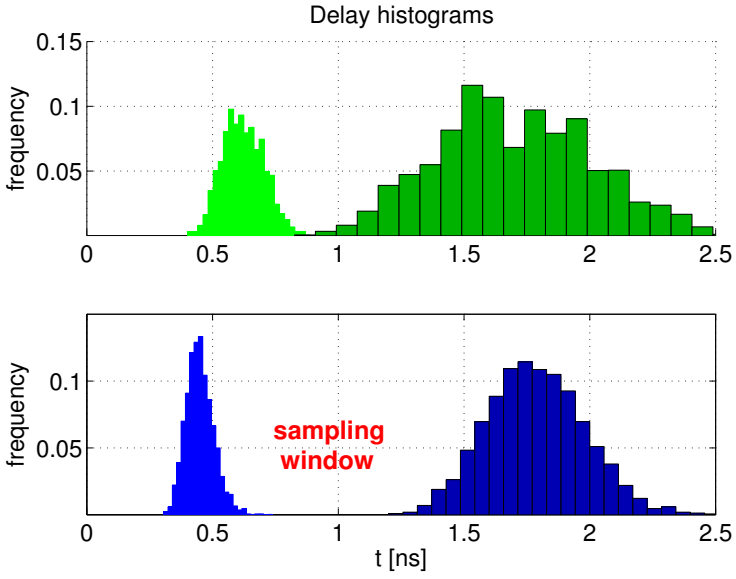
Small-signal, single-ended sensing with offset compensation has also been employed in the works listed in Table 3.1. By AC coupling, this work avoids the unconditional full V_T swing on the bitlines and the requirement for interlock between wordline access and precharge found in [40]. By exploiting the sharp cutoff characteristic of a PMOS stack M3-M4 to separate 1 from 0 (Fig. 3.5(a)), this work avoids the 2 capacitors, 13 transistors, and power penalty related to regenerative feedback in [41]. Similarly, the proposed ACSA is more easily power-gated and has fewer transistors and capacitors than the scheme in [42], while also avoiding the need to distribute an explicit reference that must recharge input capacitors on a cycle-by-cycle basis. Hence, 128 parallel ACSAs can be laid out on the bitline pitch, satisfying the non-interleaved column constraint of low-voltage 8T SRAM to avoid half-selected write operation.

Table 3.1: Related offset-compensation sense amplifiers for memory

Publication	Approach
[40]	Precharges CAM match line through the sense amplifier input device to automatically set a switching threshold that is adjusted to sense amplifier mismatch
[42]	Cascades two common-source amplifiers biased at their trip point during precharge.
[41]	Cascades two AC-coupled inverters biased in their high gain region. Regenerative feedback is introduced with a feedback NMOS.
This work	Dynamic sensing (to minimize static power) with an AC coupled input, during precharge the PMOS threshold voltage drop is stored on the capacitor, and sensing with only PMOS devices facilitates voltage scaling

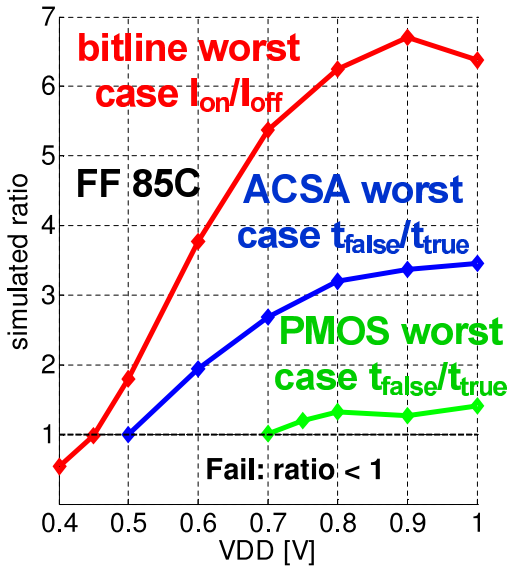


(a) $VDD = 1V$

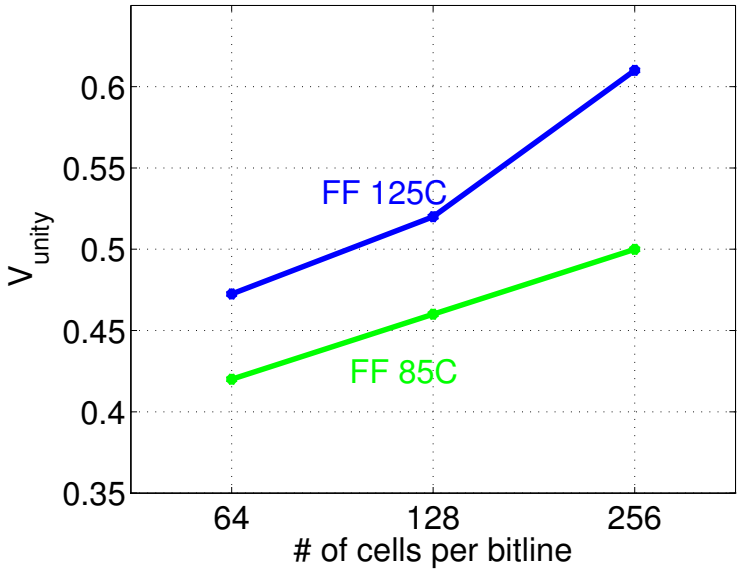


(b) $VDD = 0.55V$

Figure 3.7: Delay histograms of "true 1" and "false 1" for local sensing with a dynamic PMOS (upper green plot) and sensing with the ACSA (lower blue plot) at FF 85C



(a) Sensing margin in the context of the fundamental bitline on/off limit FF 85C



(b) Simulation of V_{unity}, the supply voltage at which the sampling window disappears, for the ACSA across bitline length and temperature (FF corner).

Figure 3.8: Monte Carlo simulation of sensing margin at 5σ, corresponding to 90% yield of the 512kb memory.

3.1.2 The Regenerative Global Bitline Scheme

After local sensing, data is forwarded across eight 64kb banks in the read cycle via the regenerative global bitline scheme (RGBS). Shown in Fig. 3.9 (a) is the local evaluation network based on two ACSAs supporting a total of 512 bits. The dynamic outputs of the two ACSAs are shorted and buffered to the global bitline through the pull-up device Mgp. When a "1" is sensed on either the upper or lower bitline, device Mgp turns on, charging the global bitline from $0V$ to VDD.

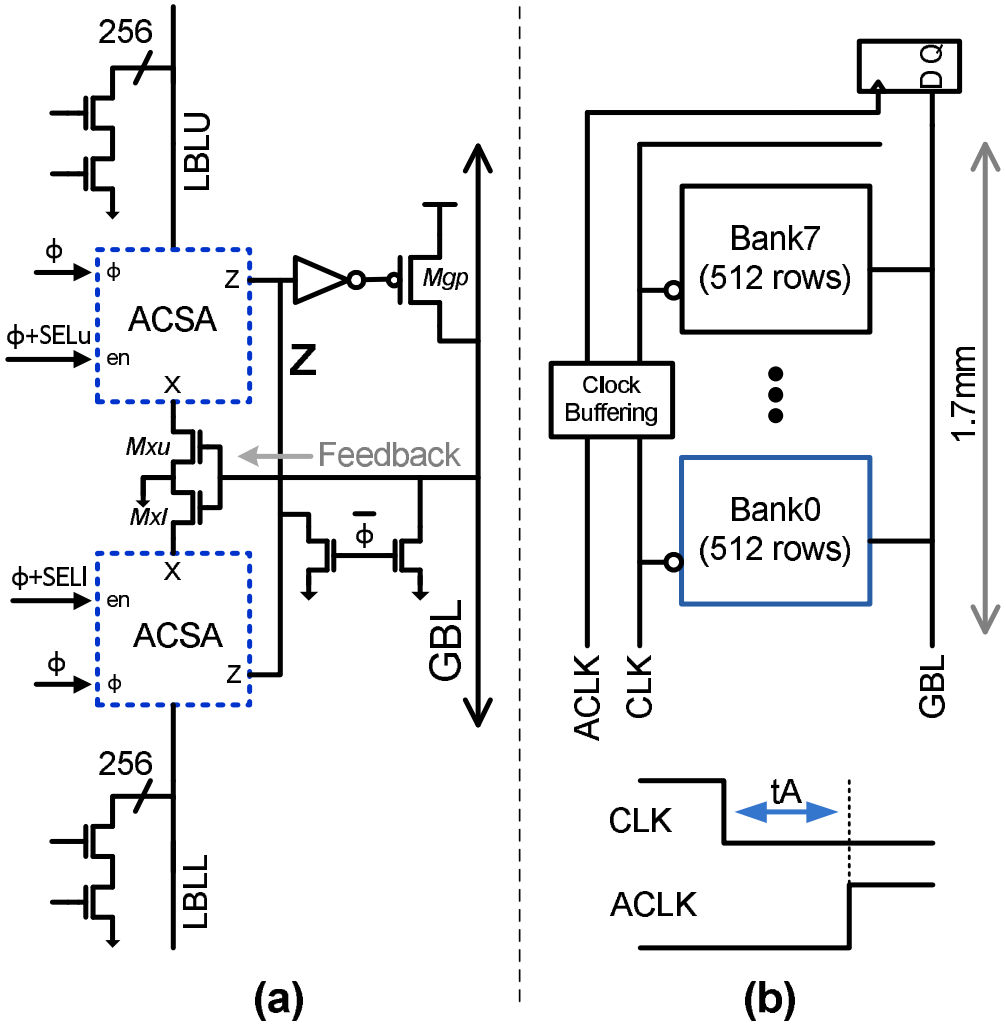


Figure 3.9: Shown is (a) the ACSA based local evaluation network for the regenerative global bitline scheme and (b) the full global read path consisting of 8 banks containing 512 rows each, along with the associated timing signals for access time measurement.

As the global bitline charges up, it need only reach the threshold voltage of feedback devices Mxu and Mxl $220\mu m$ away at the adjacent bank, at which point the ACSAs in the adjacent bank trigger the turn-on of their associated pull-up device Mgp. This process repeats until the data arrives at the edge of the macro $1.7mm$ away. This cascaded operation is illustrated by the simulation waveforms in Fig. 3.10. The assertion of RWL produces a rise on node Z at bank 0, causing the GBL near bank 0 to sharply rise. The GBL near bank 1 rises with a first order RC response until it triggers the ACSAs in bank 1 causing a sharp rise from the turn-on of device Mgp in bank 1. In this manner all banks work together to drive the global bitline.

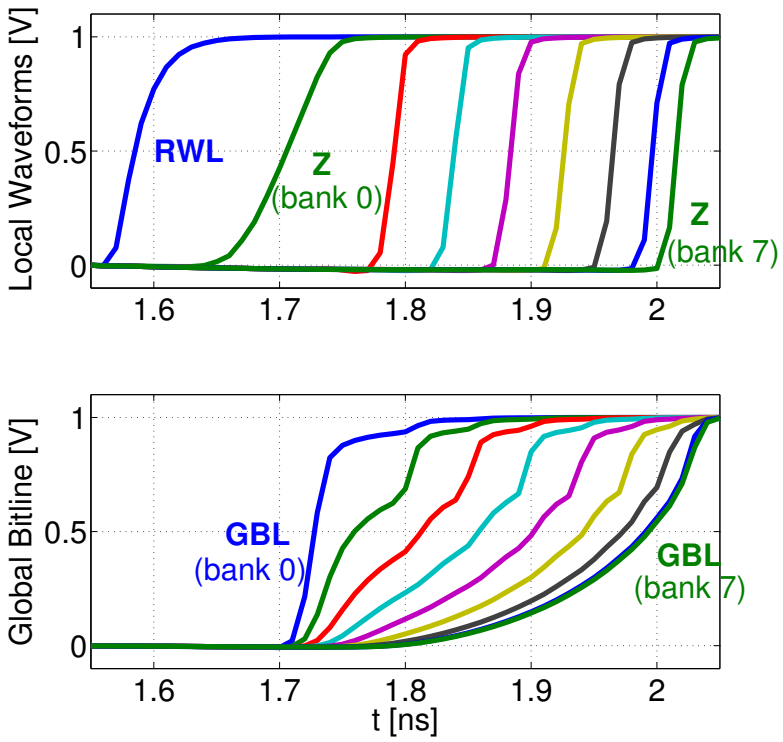


Figure 3.10: Simulation waveforms (TT 25C) illustrate the cascaded operation of the regenerative global bitline scheme in which all banks work together to drive the global bitline (GBL).

The simulated delay of the RGBS is 40% faster than the conventional technique of pulling down the global bitline with a single NMOS device. The RGBS scales linearly with distance,

preserving sharp transitions and performance comparable to a buffering daisy chain [43]; yet the RGBS avoids cross-over currents and enables bidirectional signaling—preventing routing congestion of the 128b word—through a single metal 4 track on the memory cell pitch. The full read path benefits from CMOS scaling by avoiding differential sense amplifiers, which are sensitive to mismatch, and variable timing control signals [41, 44], especially in SOI technology that is susceptible to history-dependent floating body effects [45].

3.1.3 Read Path Characteristics and Measured Performance

The $2.0mm \times 0.35mm$ test site contains a 512kb macro comprised of 8 banks of 64kb. Each bank, shown on the bottom left in Fig. 3.11 is $223\mu m$ tall, and within each bank (shown on the bottom right) the read circuits are $14\mu m$ tall and replicated for each column of memory cells. There are 2048 coupling capacitors on the read path that work in concert when accessing a 128 bit word. The realized macro-level bit density is $1.19\text{Mb}/mm^2$ with a projected bit density of at least $1.39\text{Mb}/mm^2$ with an optimized row periphery in this $45nm$ technology.

The access times in Fig. 3.12 equal the minimum separation between the falling edge of CLK and the rising edge of ACLK as depicted in Fig. 3.9(b). Each measurement corresponds to 100% pass of all 512kb under the write and read back of alternating checkerboard and blanket data patterns. Two measurements below 0.65V require partial turn-on of a bleeder PMOS device on the bitlines to compensate for fast process corner leakage. At $1.2V$ the access time is $400ps$. Down to $0.8V$ the access time slows to $1ns$ and more steeply increases to $3.4ns$ at $0.57V$. Beyond this point, the slowest "1" takes longer than the fastest "false 1," resulting in a few failing bits. Measured leakage power scales down 9.4x over this range.

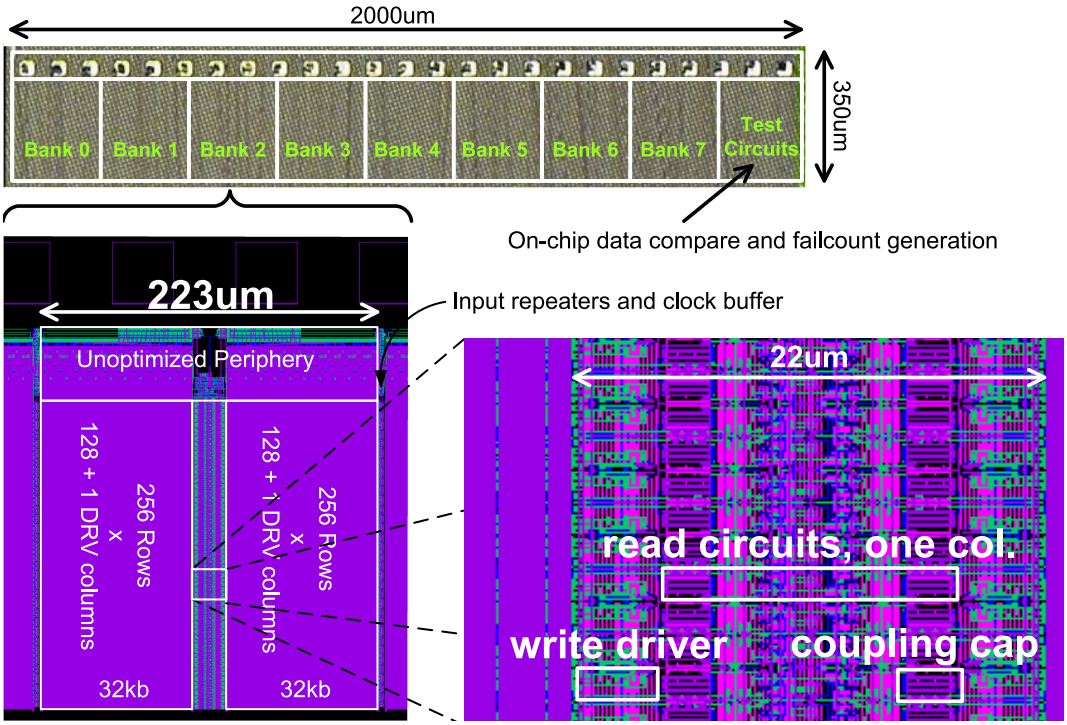


Figure 3.11: Photo of the 512kb SRAM Macro test site along with layout snapshot of the subarray and read circuits.

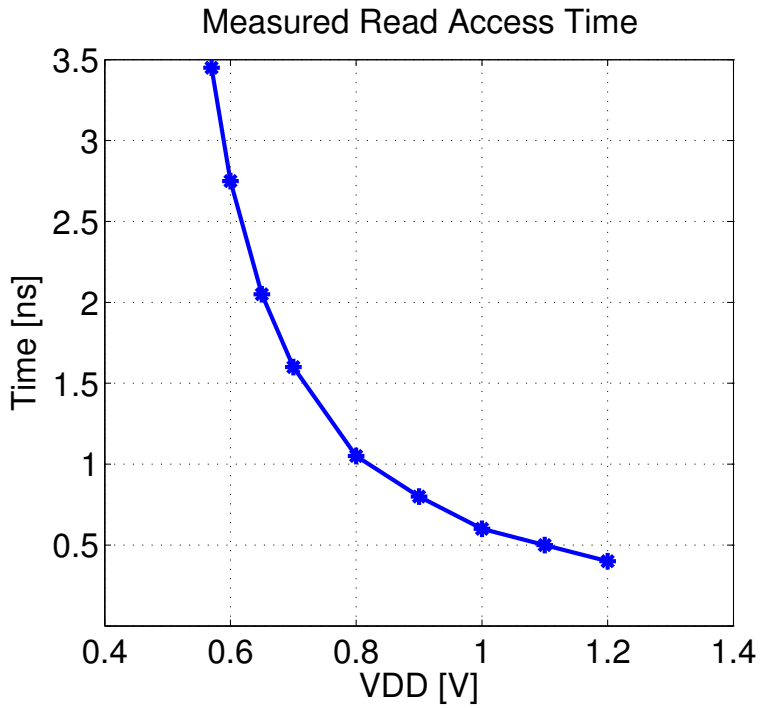
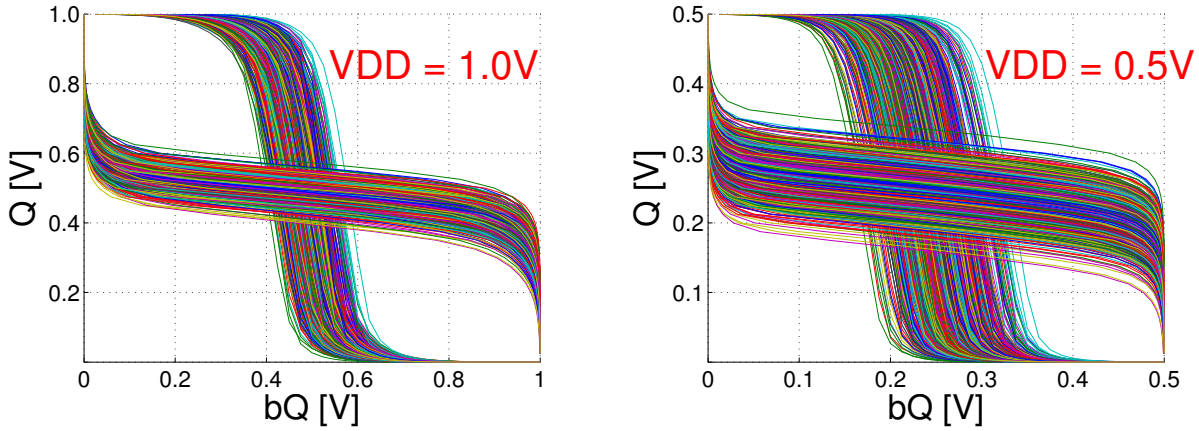


Figure 3.12: Measured read access time versus supply voltage. Two measurements below 0.65V require partial turn-on of a bleeder PMOS device on the bitlines to compensate for fast process corner leakage.

3.2 The Data-Retention-Voltage Sensor

For idle banks, lowering the supply to the data-retention-voltage enables dramatic reduction of standby power. However this limit is uncertain as it results from the extremes of local mismatch variation. The degradation in retention characteristics versus supply voltage is illustrated going from VDD = 1.0V in Fig. 3.13(a) to VDD = 0.5V in Fig. 3.13(b).



(a) Retention butterfly characteristics at 1.0V

(b) Retention butterfly characteristics at 0.5V

Figure 3.13: The reduction of supply voltage degrades retention stability until failure occurs at an uncertain limit determined by the extremes of local mismatch variation.

It turns out that the DRV of the particular test site measured, for its particular process corner and set of operating conditions, lies between $0.375V$ and $0.4V$. If this can be determined without perturbing the functional array, leakage power can be reduced by 29x going from $1.2V$ to $0.4V$, Fig. 3.14. However, in general the minimum supply voltage is determined by the one memory cell out of 524,288 with the largest DRV, determined by local mismatch variation and its functional relation to the static noise margin (SNM) of the memory cells [46]. This relation changes with process corner, temperature, and end-of-life degradation.

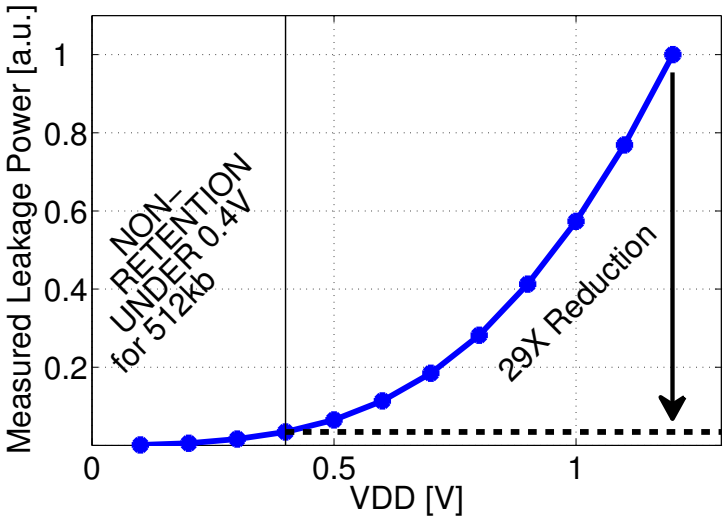


Figure 3.14: Measured standby leakage power.

3.2.1 Background

Because of the uncertainty related to the DRV, current approaches conservatively place PMOS [47] or NMOS [48] diodes in series with the SRAM power supply to reduce the SRAM array leakage. There is an increasing emphasis on accurately setting the DRV—with programmable diodes [49] or, most recently, active regulation [50] of the array standby supply voltage—since both the DRV and the diodes vary with global process corner. Nevertheless, conventional approaches must insert excessive margin into the target retention supply level that bounds the dependency on process corner, temperature, and end-of-life conditions.

Prior work recognizes the importance of tracking the DRV. The approach in [51], biases the array supply to twice the threshold voltage of the bitcell devices, $2 \times \max(V_{Tn}, |V_{Tp}|)$, as observed from on-chip replica cell transistors. This solution has the benefits of tracking global variation and achieving greater leakage reduction than a series diode. Although this level of supply voltage is sufficient to preserve data, it is not a necessary condition. Further reduction is possible. Another approach proposes "canary" cells [52], which are significantly altered memory cells that must be calibrated against a known DRV distribution before they can be interpreted as a premature indication of retention failure. The primary challenge remains: how to efficiently determine the DRV on-chip without corrupting the contents of

the functional array?

Analyzing such a problem in simulation is straightforward with the Monte Carlo method, a generally applicable and easy to use approach to evaluating failure probability. A rule of thumb for the required number of simulation runs for a high confidence estimate of a failure probability, p, is given by:

$$N_{MC} = \frac{100}{p} \ ,$$

which was discussed in Eq. 2.1 of Chapter 2. Failures of interest for this work's 512kb array are 10^{-5} to 10^{-7}, requiring a prohibitively large number of simulations from 10^7 to 10^9 runs. Because the failure is observed by measuring how long one must wait to encounter a failing realization, the process takes an extremely long time. In the realm of hardware, this same approach is taken by building an extremely large number of SRAM cells in a yield learning vehicle. Once again failure is observed by counting how many good cells can be built before the first bad one is encountered.

Because of this barrier, accelerated simulation techniques have been developed to more efficiently utilize computational resources in the evaluation of failure probability [35, 14, 5, 10]. When a memory cell fails it will usually fail under its most likely manifestation. For this reason, the simulation approach described in Chapter 2 focuses on identifying the worst-case point [27] in the space of variation parameters that corresponds to the dominant failure mechanism. This approach employs statistical sampling to search for this point in combination with Importance Sampling simulation, a type of variance reduction algorithm, to estimate SRAM failure. Run time speedups of 1000x to 10,000x have been demonstrated.

3.2.2 DRV sensor operation

To meet this challenge in the context of the 512kb array, one must make a 5σ measurement with 2σ-worth of samples or fewer. Techniques from the simulation realm will be adopted in the hardware realm to more efficiently utilize resources. The sampling of process variation present in 256 memory cells will be employed to explore the parameter space of threshold variation to identify the dominant retention failure mechanism on chip.

The overview of the DRV sensor is given in Fig. 3.15. It consists of a single column of 256 memory cells in each half-bank, incurring an area overhead of less than 2%. Unlike cells in the functional array, the sensor cells contain split VDDand VSS wiring, providing a means to skew and program without disturbing the functional array. Ultimately, the algorithm aims to recover the functional relationship between threshold voltage variation and static noise margin:

$$\mathrm{SNM}(\boldsymbol{\Delta V}) = 1 + c_1 \Delta V_{T1} + c_2 \Delta V_{T2} + \ \ldots \ + c_6 \Delta V_{T6}$$

Accuracy of this description is needed only in the neighborhood of the worst-case point. The generation of skewed supplies and the processing of the DRV sensor data, including matrix inversion, are implemented off-chip in this work but can be implemented on-chip in future work.

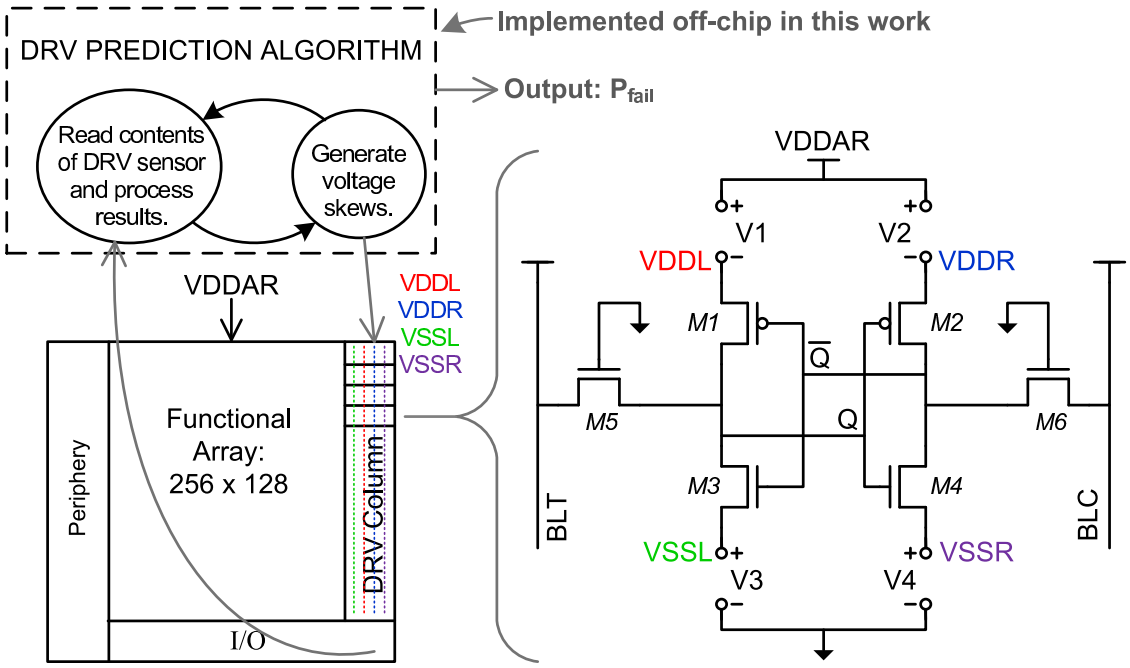
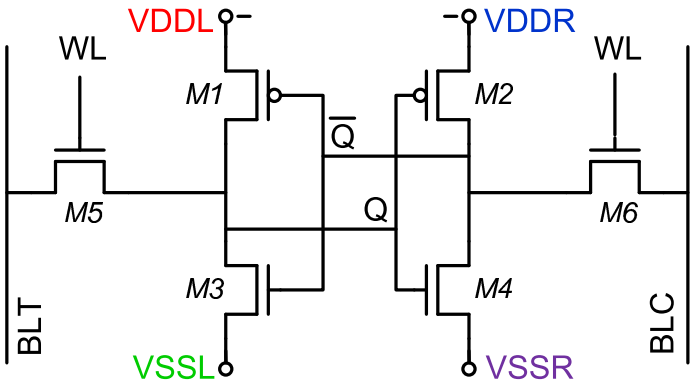


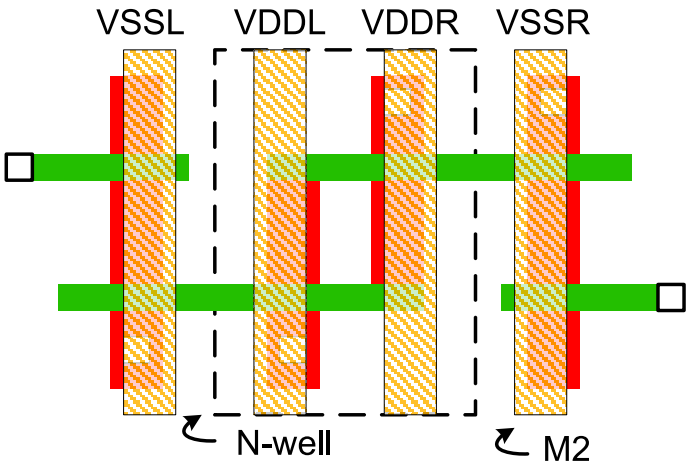
Figure 3.15: Overview of the DRV sensor. The DRV sensor cells are 8T cells with the same size as the functional 8T cells. The read path is identical to the functional 8T cells. The 2T read stack is not shown for clarity.

The sensor cells must match the functional SRAM cells. The schematic in Fig. 3.16(a) illustrates the splitting of supply wiring. The corresponding layout in Fig. 3.16(b) shows that

the single difference with respect to the functional cell is the splitting of a fat M2 VDDline into VDDL and VDDR. The VSS lines are naturally split by the bitcell layout. Therefore, the identical layout of the sensor cell transistors will track the random-dopant-fluctuation mismatch in the functional cells. The actual DRV sensor cells are 8T cells with transistor layouts identical to the functional 8T cells. The 2T read stack does not interfere with the routing of the skewed supplies to the 6T storage portion.



(a) Schematic of DRV sensor cell



(b) Layout of DRV sensor cell

Figure 3.16: The transistor layout in the DRV sensor cell is identical to the transistor layout in the functional bitcell. The DRV sensor cell is an 8T cell, the 2T read stack which does not interfere with split supply wiring is not shown for clarity.

The application of voltage skews to the sensor cell emulates an effective shift in threshold voltage since the drain current depends on $(V_{GS} - V_T)$ and the voltage skews directly add to or detract from V_{GS}. For the data state of Q = 0, this transformation is obtained from the

matrix:

$$\mathbf{T}_{0\rightarrow1} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix}$$

and the column vector of voltage skews:

$$\boldsymbol{V} = [\ V1 \quad V2 \quad V3 \quad V4\ ]^T.$$

The effective threshold voltage change, $\boldsymbol{\Delta V}_{T\text{eff}}$, is obtained from the product $\mathbf{T}_{0\rightarrow1}\boldsymbol{V}$. The columns of $\mathbf{T}_{0\rightarrow1}$ correspond to the applied voltage skews, and the rows of $\mathbf{T}_{0\rightarrow1}$ correspond to the resulting threshold voltage deviations of transistors M1-M6 (both the skews and transistors are labeled in Fig. 3.15). As an example, consider a skew, $\boldsymbol{V}^*$, in which the only non-zero component is $V2 = 100mV$:

$$\boldsymbol{V}^* = [\ 0 \quad 100mV \quad 0 \quad 0\ ]^T$$

Because the gate-to-source overdrive of M2 is degraded, its threshold voltage is effectively $100mV$ higher.[1] Also, since M2 is "on" for the given data state, the $100mV$ drop propagates to the gate of M3 making it also $100mV$ weaker. For the same reason, both M1 and M6 are $100mV$ stronger. Hence,

$$\boldsymbol{\Delta V}^*_{T\text{eff}} = [\ -100mV \quad 100mV \quad 100mV \quad 0 \quad 0 \quad -100mV\ ]^T.$$

The above vector corresponds to the second column of $\mathbf{T}_{0\rightarrow1}$. Tracing through the effect of the other voltage sources on the gate-to-source overdrive of the bitcell transistors, one can fully determine $\mathbf{T}_{0\rightarrow1}$, the transformation from supply voltage skew to threshold voltage

[1]For the analysis related to the DRV sensor, a positive V_T convention is employed for PMOS devices.

skew. Therefore, the resulting SNM, under a voltage skew, $\boldsymbol{V}$, is:

$$\begin{aligned} \text{SNM} &= 1 + \boldsymbol{c}^T \boldsymbol{\Delta V}_{T\text{eff}} \\ \text{SNM} &= 1 + \boldsymbol{c}^T \mathbf{T}_{0\to 1} \boldsymbol{V} \end{aligned}$$

Given a 1σ measurement of V_T variation in memory cell devices,[2] the DRV sensor is skewed in multiple directions to reconstruct the SNM as a linear function of V_T. The chosen voltage skews are:

$$\begin{aligned} \boldsymbol{V}^a &= \tfrac{k_a}{\sqrt{2}} \begin{bmatrix} 0 & 1 & 0 & -1 \end{bmatrix}^T \\ \boldsymbol{V}^b &= k_b \begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix}^T \\ \boldsymbol{V}^c &= k_c \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix}^T. \end{aligned}$$

where k_a, k_b, k_c parametrize the magnitudes. The algorithm in Fig. 3.17 searches for the value of these magnitudes that collapse the nominal SNM, as observed by 50% failure in the 256 memory cells, for a given supply voltage level under test. Specifically, the cells are programmed to "0" and failure is observed by counting the number of sensor cells that flip to "1." The sensor cells are read through through the same ACSA based read path as functional cells. The sensor cells are written to "0" by collapsing VDDL to $0V$. It is important that the sensor cells are not written through the standard write path, because it would require the assertion of WWL and therefore disturb the functional memory cells in the same row. Both the read and write operations are executed when the sensor cells receive nominal supply voltage conditions (e.g. VDDL and VDDR at $0.9V$ for read, VDDR at $0.9V$ for write, and VSSL and VSSR at $0V$). These magnitudes of the skews are labeled k_a, k_b, k_c and the projections of two of the skews are illustrated in Fig. 3.18(a).

By simplifying coefficients (assuming M5 and M6 do not influence retention and M3 and M2 have similar influence on the SNM), a reduction of dimensionality is obtained:

$$\begin{bmatrix} c_1 & c_2 & c_3 & c_4 & c_5 & c_6 \end{bmatrix} = \begin{bmatrix} c_1 & c_2 & c_4 \end{bmatrix} \mathbf{R}$$

[2] An example approach to obtaining the standard deviation of SRAM transistor threshold voltage can be found in [53].

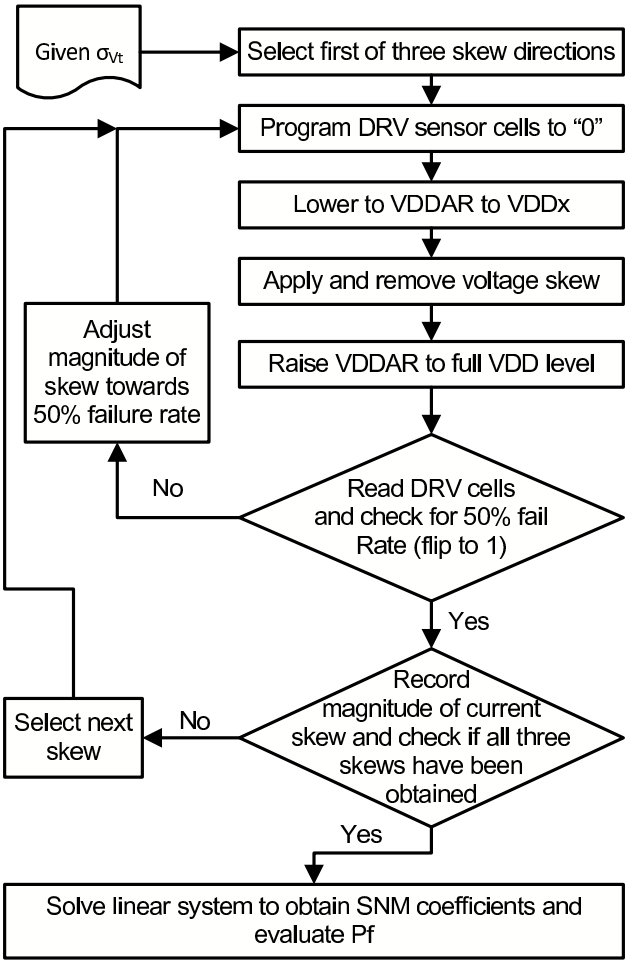


Figure 3.17: DRV sensor algorithm

where

$$\mathbf{R} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

With the observed values of $\boldsymbol{V}^a, \boldsymbol{V}^b, \boldsymbol{V}^c$, the following expression gives the SNM coefficients corresponding to the particular global corner and power supply level during measurement:

$$\begin{bmatrix} c_1 \\ c_2 \\ c_4 \end{bmatrix} = - \left(\begin{bmatrix} \boldsymbol{V}^{aT} \\ \boldsymbol{V}^{bT} \\ \boldsymbol{V}^{cT} \end{bmatrix} \mathbf{T}_{0\to1}^{T} R^T \right)^{-1} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

Finally, the resulting coefficients are combined with σ_{V_T} to estimate failure

$$P_f = \phi\left(-\frac{1}{\sqrt{(c_1\sigma_1)^2 + (c_2\sigma_2)^2 + \ldots + (c_6\sigma_6)^2}}\right)$$

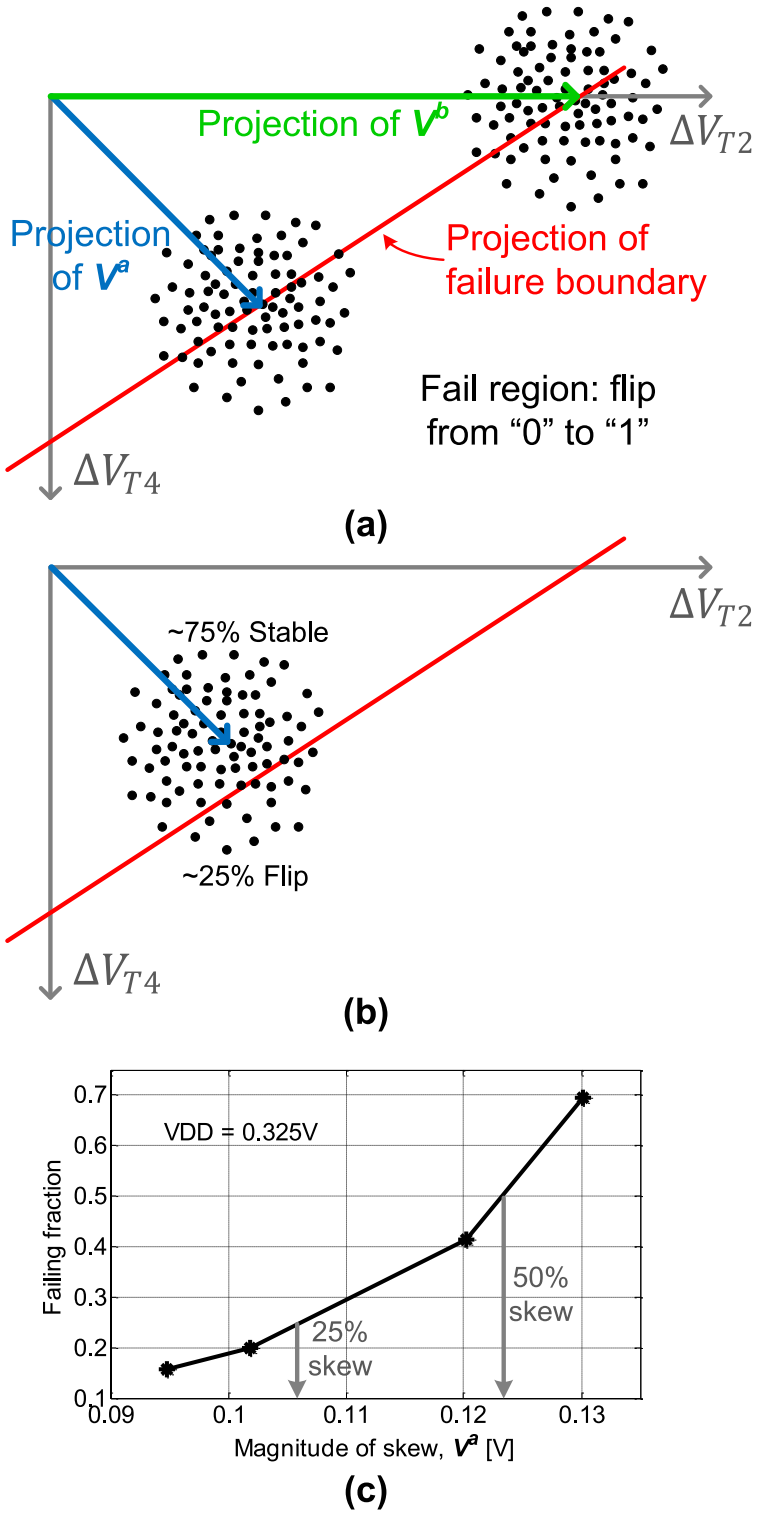


Figure 3.18: Shown is (a) two DRV skews projected on a 2D space with a mismatch sampling cloud of size 256, (b) a reduced magnitude vector for conservative estimation, and (c) an example measurement for skew $\boldsymbol{V}^a$ at VDDAR $= 0.325V$.

3.2.3 Measurement results

The retention failure rate versus supply voltage is first observed on the functional array by writing the entire array to "0" at an operational VDD(e.g. $0.8V$), then reducing the array to the retention level under measurement, then restoring the array to an operational VDD, and finally reading out the array to count the number of bits that flipped. In this experiment, all voltage adjustments and skews are quasi-statically applied at a rate of $200mV$/minute, and furthermore the skews are generated by incrementing one source at a time by $1mV$. The black line in Fig. 3.19 shows that at $0.4V$, no bits fail in the functional array, and at $0.375V$, four bits fail. At $0.16V$, over 80% of the bits are still good. This wide transition poorly modeled by the normal CDF illustrates the challenging statistical question that must be answered.

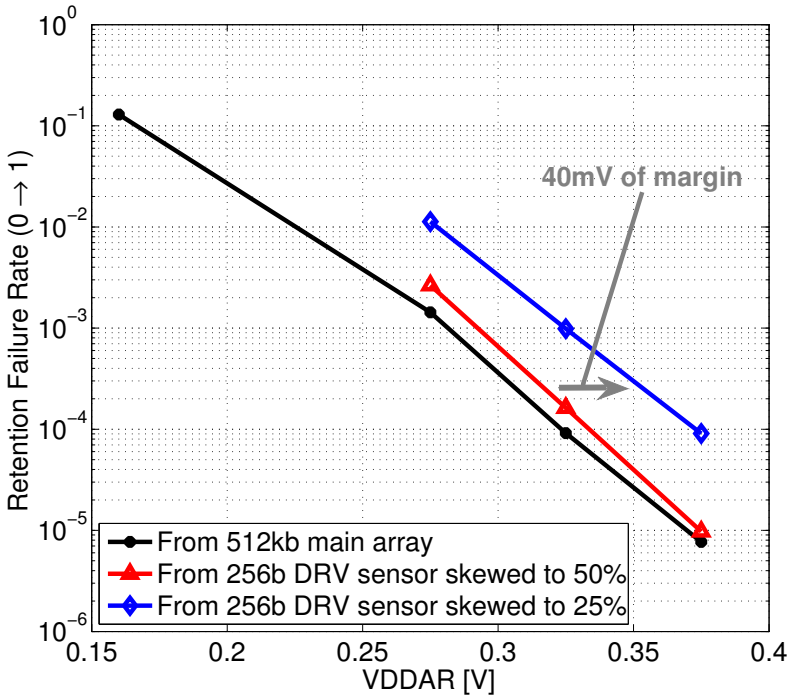


Figure 3.19: DRV Sensor measurement results on one 512kb test chip.

Next, the DRV sensor algorithm is run at $0.275V, 0.325V,$ and $0.375V$ with σ_{V_T} of the bitcell transistors taken from the technology design manual. The red line shows the results

of the DRV sensor algorithm. A reasonable matching of failure probability is attained. Drawing a horizontal line at a given failure probability such as 10^{-5} results in less than $10mV$ of error in the predicted data retention voltage.

Furthermore, margin can be inserted into the algorithm by replacing the failure criterion of 50% with 25% as illustrated in Fig. 3.18(b). As a result of the spherically symmetric sampling in the threshold voltage domain, the sensitivity coefficients of the static noise margin will be consistently overestimated, producing a measured margin of approximately $40mV$ in the data retention voltage. Each data point from the DRV algorithm comes from the observation of the three skew magnitudes k_a, k_b, k_c. Fig. 3.18(c) gives the measurement data for $\boldsymbol{V}^a$ at $0.325V$. Executing the inner loop of the algorithm samples the relation between the fraction of flipping sensor cells versus skew magnitude. When sufficient data is gathered, interpolation of the points gives the magnitude for the 50% and 25% points.

Therefore, the DRV sensor algorithm reveals the limit of standby supply voltage while guaranteeing negligible risk of losing data in real-time embedded operation. This technique is relevant to state-of-the-art embedded SRAM that requires the retention voltage as an input to standby power regulation circuits [54].

3.3 Concluding Remarks on SRAM Voltage Scaling

This prototype has revealed that AC coupling is viable on a finer scale in advanced CMOS technology as the gap between intrinsic device capacitance and interconnect capacitance continues to grow. Secondly, the analysis of the bitcell in isolation cannot reveal the voltage scaling limits of the overall memory. Variation tolerant sensing networks are critical to avoid further degradation. Finally, the extreme statistical fluctuation of performance metrics can be efficiently predicted on-chip by recovering the functional relationship between the process variation parameter and performance metric.

As future work, timing circuits need to be developed for capturing the dynamic operating window of the read path. In addition, the theoretical basis of the DRV sensor can be augmented to account for DIBL and body bias (in bulk technologies) while also being

extended to other failure mechanisms such as read instability and other circuits such as sense amplifiers and flip-flops.

Summarized in Table 3.2 and Fig. 3.20 are the features of the 512kb Macro organized into 4096 words of 128b. Each half-bank contains its own column of DRV sensor cells for fine resolution observation of the DRV. A sensing network of $20.9\mu m^2$ supports the read-out of 512b. The access time scales from $400ps$ to $3.4ns$ from $1.2V$ to $0.57V$. The 64kb bank presents a building block that can be tiled to form high density, low-voltage last level cache in scaled CMOS technology.

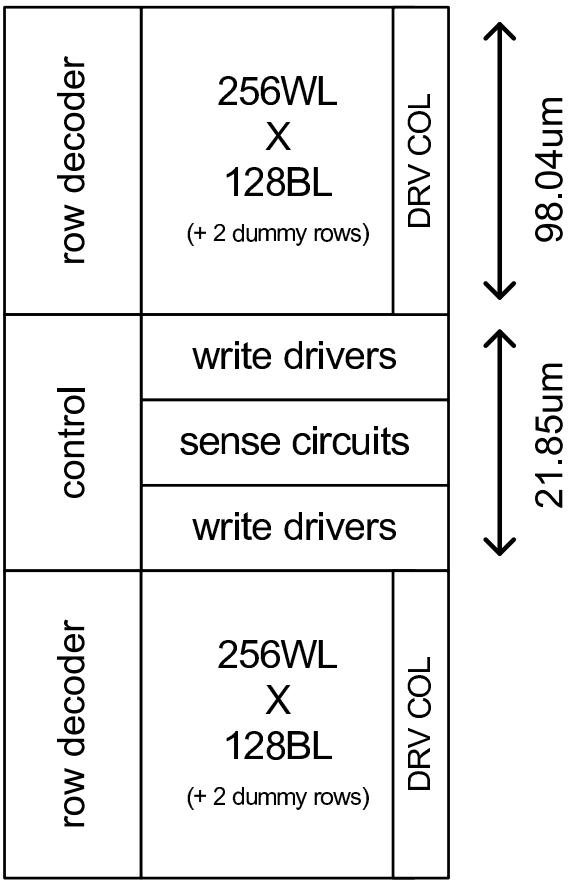


Figure 3.20: 64kb bank structure

Table 3.2: Summary of test chip characteristics

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

Chapter 4

Time-to-Digital Sensing for FRAM

Non volatile memory has the essential feature of consuming zero standby power while retaining data. Within each of the non-volatile memory technologies in Table 4.1, there exists a trade-off among cost, performance, endurance, and energy consumption. The capability of the sensing circuits to distinguish 1 from 0 across all the statistically fluctuating bits in a given memory chip constrains this trade-off.

Table 4.1: Comparison of Representative Non-volatile Memory Designs

	NAND Flash	NOR Flash	PRAM	FRAM	MRAM
Density (Mb/mm^2)	216	20.6	5.86	1.47	0.502
CMOS node	32 nm	65 nm	90 nm	150 nm	130 nm
Timescale of bitcell	1.89 ms (prog.)	70 ns (read) 435 μs (prog.)	78 ns (read) 430 ns (write)	70 ns (read)	32 ns (read) 70 ns (write)
Endurance	$\approx 10^5$	$\approx 10^5$	$> 10^5$	$> 10^{13}$	$> 10^{15}$
Energy per written bit	1.4 nJ (prog.)	—	9.8 nJ	50 pJ	253 pJ
Ref.	[55]	[56]	[57]	[58]	[59]

* The calculation for energy per written bit is taken from the product of total active power divided by random cycle time and number of IOs. This is not to be confused with other methods that evaluate the active power divided by burst-mode IO bandwidth as in [60].

In order to achieve the full potential of these memory technologies, the cells must be

read with circuits that exhibit low mismatch and operate at low voltage while using the transistors of deeply scaled CMOS. Area efficiency must also be preserved. The work in this chapter begins with an observation about the time scale of the memory cell—whether it is the RC delay of a narrow flash WL, or the polarization time of a ferroelectric capacitor, or the duration of annealing in a phase change memory element. This time scale is significantly longer than the fanout-of-four delay of the underlying CMOS logic gates.

One circuit noted for low voltage operation and performance improvement with technology scaling is the time-to-digital converter (TDC) [61]. This work presents an embedded FRAM prototype that uses this type of circuit to sense with high resolution for the purpose of reducing the access energy in the non-volatile RAM and storage of ultra low power SoC's (with 16 to 32 bit data paths) that target a mere 10 to 100 picojoules of energy consumption per clock cycle [62, 63, 64]. Although NAND Flash, the predominant non-volatile memory technology, offers the ultimate in cost for low-power portable electronics, it forces the user to pay a significant penalty in endurance, performance, and access-energy. Many applications such as implantable medical devices require significantly lower access energy non-volatile memory to deliver longer battery lifetime and richer functionality, necessitating the exploration of an alternative technology.

One such promising technology is low-voltage FRAM. Presently, the minimum required operating range of FRAM memory designs, using the the most viable PZT ferroelectric capacitor process technology, is around $1.3 \sim 1.5$ V [65, 66]. This range needs to be reduced for more advanced technology nodes because smaller-geometry transistors must operate at lower supply voltage. Also, in presently available technologies, lowering the operating voltage can dramatically reduce the energy for accessing data because ferroelectric memories contain significant switched capacitance due to long bitlines, long and boosted wordlines, global IO routing, and even the large capacitance bitcell itself which must be toggled a second time after a destructive read operation. The ultimate voltage limit of FRAM depends on both technology properties and circuit design characteristics. Some technology properties include compatibility with CMOS processing, dielectric thickness engineering, leakage,

dynamics of polarization, and density of polarization. Some circuit design characteristics include variation-tolerant design, neighbor cell disturbance, low-voltage peripheral circuits, and reference generation. Hence, low-voltage FRAM remains an active area of investigation.

There have already been significant developments to help understand the voltage limit of FRAM. The work in [67] achieves 0.9 V operation with a 2.7 μm^2 cell containing a strontium bismuth tantalate (SBT) ferroelectric capacitor. It presents a low variance reference cell that tracks data imprinting and employs a bitline layout that shields neighboring cells from each other. So as the signal margin shrinks with lower supply voltage, the reference remains centered between the two data states. The work in [66] demonstrates a memory with a 0.72 μm^2 cell in a more CMOS-process-compatible lead zirconium titanate (PZT) capacitor technology that functions down to 1.3 V. To enable low-voltage operation, peripheral circuits actively drive neighbor bitlines to couple additional voltage bias across the sensed capacitor. Because the ferroelectric capacitor exhibits a hysteresis in charge versus voltage, developing a larger voltage across the cell ensures that maximum charge is extracted.

By employing a time-to-digital sensing scheme, this work departs significantly from the classic DRAM read path. It eliminates clocked sense amplifiers that convert a small signal on a high impedance bitline directly into a logic 1 or 0 at a specific strobe timing. Other alternatives to the DRAM read path have been explored for FRAM. The design in [68] uses a sense amplifier based on cross-coupled inverters with an unconventional precharge scheme that avoids the toggling of the cell capacitor to reduce access time. The work in [69] recognizes that a low-impedance bitline can extract more charge from the ferroelectric capacitor, so it includes a pre-amplifier to establish a low impedance bitline, at the cost of area (24 transistors and 14 coupling capacitors) and additional timing signals. Neither of these alternatives to the DRAM style read path mitigate the offset inherent to the cross-coupled inverter sense amplifier that results from local device mismatch.

This paper presents a low-voltage FRAM design that compensates the offset of the sensing circuitry so that the reference value for the 1T1C cell can be accurately compared against the bitcell signal. It is first observed that, at low power supply voltage, longer bitlines maximize

the diminishing bitcell signal. In this sense, placing many (1024) cells on a common bitline best utilizes the smaller VDD-limited voltage range to span the entire width of the cell hysteresis. Next, a description of how the cell is read by a time-to-digital converter without compromising area efficiency is presented. Finally, characterization of a 1Mb prototype, containing a 0.71 μm^2 cell with a 70 nm PZT ferroelectric capacitor, shows operation down to 1.0 V.

4.1 FRAM for Low-Access-Energy Non-Volatile RAM

4.1.1 The 1T1C FRAM Cell

The 1T1C FRAM cell in Fig. 4.1 based on the technology in [65] contains a ferroelectric capacitor in series with an access transistor. The ferroelectric capacitor exhibits a hysteresis in charge versus voltage shown in Fig. 4.2(a). When writing a 0, the selected cell's WL is raised at least one NMOS threshold voltage above VDD and the local plateline (LPL) is driven to VDD while the bitline (BL) on the other end of the cell is driven to ground. This is indicated by the dot in the upper-right quadrant of the hysteresis plot, provided that VDD is beyond the level, known as the coercive voltage, required to saturate the hysteresis. When writing a 1, the opposite condition is applied, represented by the dot in the lower-left quadrant. During retention, the applied bias is removed and a charge separation remains on the capacitor that depends on the previous write polarity.

This information can be extracted from the cell by an after-pulse read operation, which is illustrated in the hysteresis trajectories of Figs. 4.2(b)-(c). When sensing the memory cell, the BL is precharged to ground and left floating. Then the LPL is pulsed across the series combination of the memory cell and the floating bitline capacitance. This sequence is depicted by the waveforms in Fig. 4.3(a). For the case of reading a 1 (Fig. 4.2(b)), the hysteresis is traced from point X to Y during the rising edge of the local plateline and then from point Y to Z during the falling edge. The dashed loadline of the bitline capacitor constrains this trajectory and generates a voltage V_1 associated with data 1 that corresponds to the

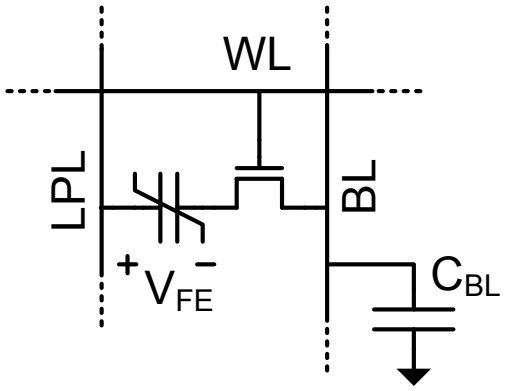


Figure 4.1: Schematic of the 1T1C bitcell with the local plateline (LPL) routed in parallel to the bitline (BL).

vertical separation between points X and Z. At lower supply voltage, the signal development on the bitline capacitance prevents the ferroelectric capacitance from experiencing its full coercive voltage even though the write operation still enables the saturation of the hysteresis. Hence the read operation limits the voltage scaling of the memory.

It is also important to note that the read operation is destructive, and that the correctly sensed data must be written back to the memory cell. When reading a 0, the same bitline precharging and local plateline sequence is applied. However, the trajectory originates at the data "0" point on the hysteresis in Fig. 4.2(c) and consequently traces the non-hysteretic linear capacitance. Therefore, the resulting voltage V_0 is close to 0 V for this after-pulse sequence (in which LPL returns to ground). The difference between V_1 and V_0 is much less than VDD so the small voltages on the bitline are converted to logic levels by comparison with a midpoint reference and amplification with a sense amplifier. It is also expected that the after-pulse sequence (as opposed to on-pulse sensing with LPL still high) cancels a component of signal dispersion related to the non-switching capacitance of the cell at the expense of a longer read cycle [70]. Finally, switching dynamics slow drastically at low enough voltage bias in ferroelectric capacitors, and the timescale of writing can pose a practical barrier to low voltage operation [71]. However, significant switched polarization was observed down to $\sim 0.75V$ on a $2\mu s$ timescale for the specific ferroelectric capacitor technology of this chip [65].

Due to process variation and material properties, the local fluctuation in signal from one ferroelectric capacitor to another degrades the observed voltage difference between 1 and 0. For the single-ended 1T1C cell, the V_1 distribution and V_0 distribution must not overlap so that a global reference can correctly separate the two data states. Offsets from the sensing circuitry impose an additional requirement on the separation of these distributions. Shown in Fig. 4.3(b) are histograms of V_0 and V_1 for a conventional bitline length of 256 cells and a tenuous power supply of 1.2 V. Also shown is the Gaussian fit of the offset of a typically sized sense amplifier (with an ideal and centered reference voltage) for a conventional FRAM design of approximately 60% array efficiency. The Monte Carlo simulation of a bitcell with a $70nm$ thick, $0.44\mu m^2$ PZT capacitor (model details can be found in [72], full bitcell specifications can be found in [73]) illustrates that a 1Mb chip will have less than 1% yield (assuming no redundancy or ECC) under these conditions.[1]

Shown on the top axes in Fig. 4.4, is the statistical voltage signal varied across bitline lengths. The voltage signal is obtained from the difference of the 5.6σ points of a Gaussian fit of the V_1 and V_0 distributions. For a modest supply voltage of 1.35 V, the conventional design point maximizes the voltage signal at 256 cells per bitline. For shorter bitlines the capacitive divider between bitcell and bitline capacitance presents too little bias across the ferroelectric capacitor to extract enough remnant charge. For long bitlines the large capacitance simply attenuates the voltage.

The bottom plot in Fig. 4.4 shows how larger bitline capacitance faithfully extracts more charge because the cell experiences larger bias. As the supply voltage is lowered, this extraction of charge becomes increasingly important and motivates longer bitlines if a lower voltage signal can be sensed. In this work, the bitcell charge is converted to a delay and sensed with a 5bit time-to-digital converter. Because the improved resolution permits operation with a smaller separation between data states, the memory chip supports long 1024 cell bitlines, preserves area efficiency, and scales in supply voltage.

[1]The chip yield is numerically evaluated by sampling the three Gaussian random variables for V_1, V_0, and sense amplifier offset in proportion to the number of cells per bitline.

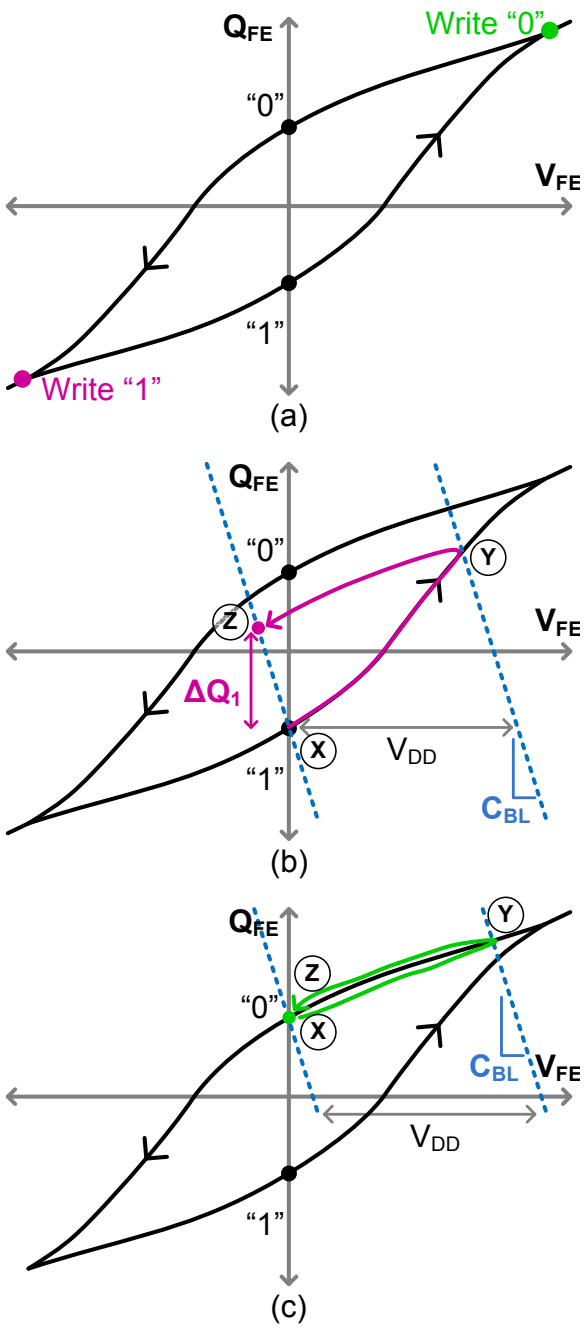
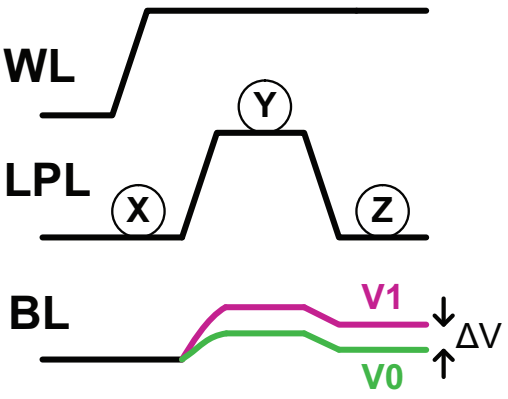
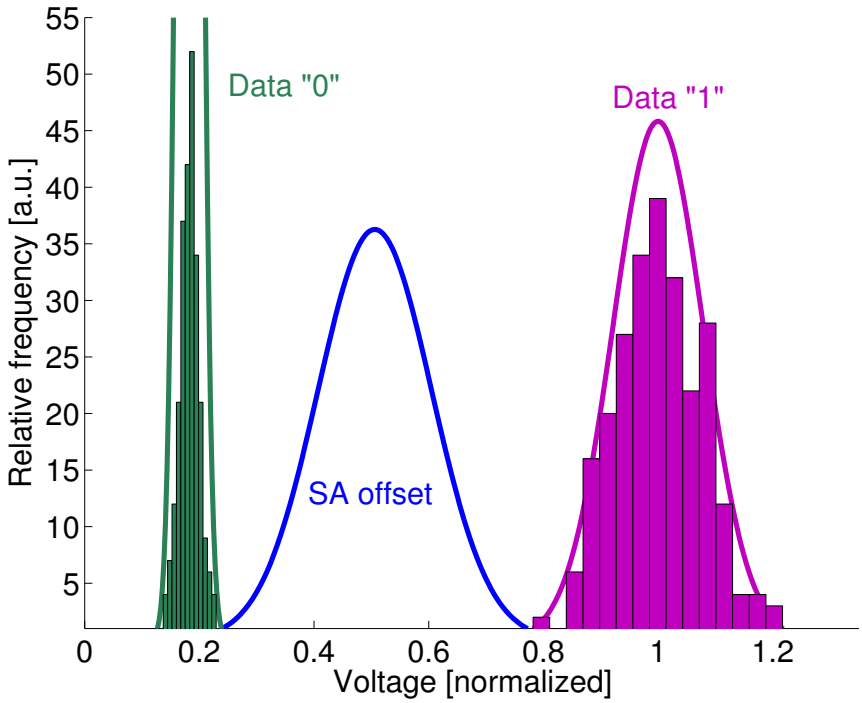


Figure 4.2: The charge versus voltage hysteresis of ferroelectric capacitor illustrating (a) write operation, (b) read 1 operation, and (c) read 0 operation.



(a)



(b)

Figure 4.3: Shown is (a) the bitcell operating sequence and (b) simulated histograms of the statistical bitcell signal with an overlay of Gaussian fits for a bitline of 256 cells and a tenuous supply voltage of 1.2 V.

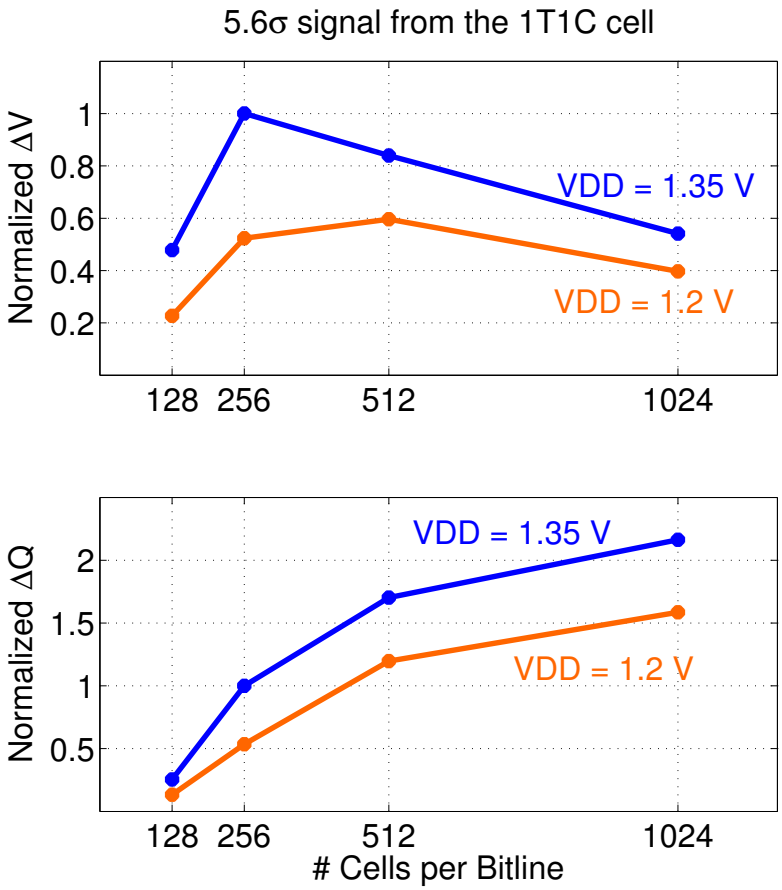


Figure 4.4: The 5.6 σ voltage signal and charge signal from the 1T1C cell over various bitline lengths and supply voltages.

4.1.2 The TDC-based Read Path

Fig. 4.5 introduces the TDC based read path, and the accompanying timing diagram in Fig. 4.6 describes the sequence of input signals and chip behaviors. The read path begins with a long bitline of 1024 1T1C cells with platelines routed in parallel to bitlines. At the end of the bitlines, a current source and comparator connect to bitlines in groups of 16. The purpose of this circuit is to convert the bitline level to a delay. Next, the bitcell signal, represented by the the delay of the rail-to-rail comparator output, multiplexes through a daisy chain of OR gates across a total of 128 columns, with 64 bitlines above and 64 bitlines below. Finally, the signal delay is processed by a 5 bit subtraction TDC, which measures both a signal delay and a reference delay and then performs subtraction to determine the cell data to be 1 or 0.

Within the 5 bit TDC, there are 32 slices. The schematic of one slice is shown in Fig. 4.7(a). Each slice contains a current-starved delay element with dynamic output to cut off cross-over current, a minimum area 9T dynamic register (Fig. 4.7(b)), and a feedback NOR gate that embeds subtraction within the TDC, avoiding the need for a 32 bit adder outside the TDC. Because the continuous bitline voltage is locally converted to a continuous delay, the OR gates serve as an effective mechanism to deliver the signal from a very small 1T1C cell to a large, high resolution sensing circuit that occupies 110 μm^2. Not only does the highly interleaved nature of the design amortize the area of the sensing circuit across several columns, but it also eliminates the neighbor bitline coupling noise encountered in DRAM-style designs [74, 70] by grounding 63 bitlines between every active bitline.

Figs. 4.8(a)-(b) show the schematic of the cascode current source and comparator. The rudimentary 9T wide-output-range OTA provides a continuous-time comparator with simple biasing, low voltage operation, and low power. The comparator is biased to a level of current that preserves a modest gain (7.1 to 8.5) and keeps the systematic offset from both gain error and crossing delay under $15mV$. Although the comparator itself has modest gain, the AND gate sharpens the timing edge of detection. More importantly, the mismatch of the comparator results in a local variation of offset across the 64 comparators in the 1Mb chip.

Fundamentally, the TDC sensing scheme needs only an inverter to convert a bitline ramp to a delay interval[2], but using a comparator instead lowers the required range of bitline voltage swing. Therefore, as will be seen during the discussion of measurement, it is necessary to ramp the bitline up to approximately $150mV$—covering the bitcell voltage signal, plus the finite transition width of the OTA, and the spread of the comparator offset—instead of VDD/2. This reduced range permits the capability of 32 TDC slices to measure the comparator offset to within a resolution of $5mV$. Because the TDC sensing circuits measure both the offset of this comparator and the bitline signal, the offset can be be canceled and its distribution need not fit within the bitline voltage distributions. This behavior will become apparent in the following description of the TDC sensing operation.

4.1.3 TDC Read Operation

First, consider the read 0 operation, which is illustrated in Fig. 4.9. To begin, the local plateline is pulsed to extract the cell charge onto the bitline, which is connected to the master bitline MBL through a transmission gate. In this case, the resulting bitline level is low for data 0. Then, an externally supplied start pulse activates the ramping of the bitline and simultaneously propagates through the TDC slices. When the comparator detects that the bitline has crossed V_{ref1}, it sends a stop signal to the TDC through the OR daisy chain, capturing the input chain of 1's followed by 0's. This measurement is illustrated in the first two snapshots of the TDC inputs (Z[0 : 31]) and outputs(TQ[0 : 31]) in Fig. 4.9(b). The location of this transition, which is processed by each slice's NOR gate and stored in the TDC registers (third snapshot), represents the delay of a 0 plus any offsets associated with the comparator, current source, and timing skews from the daisy chain. The first rising edge of the stop signal also triggers the discharge of the MBL node to ground in preparation for the second delay measurement.

Another externally supplied start pulse triggers this second delay measurement. The same bitline is ramped up with the same current source. The same comparator detects the crossing

[2]If an inverter is used, the mechanism to deliver the reference between 0 and 1 must be different from what is described herein. Namely, the reference charge must be delivered directly to the bitline.

with respect to a different voltage V_{ref2}. As before, the time of the comparator detection will be digitized by the TDC. This second measurement (fourth snapshot in Fig. 4.9(b)) represents the reference delay plus the same offsets as described before. When the TDC registers update, the NOR operation will cause the TDC slices with zero at both the input and output to go high. This comparison (fifth snapshot) accomplishes subtraction, canceling the offset, and indicating that signal delay was longer than reference delay.

The complimentary case of read 1 is shown in Fig. 4.10. At the beginning of the read cycle, the pulsing of the local plateline produces a higher voltage on the bitline. Consequently, it takes less time for the current source to ramp the bitline up to V_{ref1} during the first delay measurement (second snapshot in Fig. 4.10(b)). Therefore, a shorter delay of data 1 is stored in the TDC registers (third snapshot). When the reference delay—which remains the same because it corresponds to bitline ramping from ground to V_{ref2}—is measured, the chain of 1's at the inputs to the TDC overlaps with the chain of 1's at the outputs (fourth snapshot). This overlap causes all outputs of the TDC to update to 0, indicating that the signal delay of 1 was shorter than the reference delay. As before, subtraction is accomplished with logic gates instead of a differential pair. To determine the final result the outputs of the TDC need to be processed by a simple OR operation and the sensed data is written back to the memory cell.

The simulations of the read cycles employ a quasi-static ferroelectric capacitor model and exhibit a cycle time of 217 ns. This time scale reflects a minimum possible cycle time at 1.0V, given the two-stage TDC sensing scheme. As discussed later, actual chip measurement at 1.0 V reveals a 730 ns cycle time which is largely due to the actual dynamics of ferroelectric polarization at 1.0 V. For nominal conditions at room temperature, the minimum delay of 32 TDC slices is 9.0 ns at 1.5 V and 23.7 ns at 1.0V. Twice this time gives a coarse estimate of access time penalty compared to conventional sensing. If there exist significant timing skews in the read path (from the OR daisy chain for example), or if it takes longer to ramp the bitline up to V_{ref1}, one may choose to further slow down the TDC chain by reducing the bias on VCTL of Fig. 4.7(a). If the ferroelectric material polarization takes longer than this

time scale, then the penalty for TDC sensing becomes tolerable.

The principle of charge sensing permits widening of operating margin by slowing down: the current source charging rate can be reduced while maintaining the TDC delay resolution. A fixed charge difference between data 1 and data 0 can be expanded to a longer delay separation between t_0 and t_1. As a practical constraint, the spread in comparator voltage offset—which converts into time offset—can overwhelm the dynamic range of the TDC circuit, leaving no range for the bitcell signal. Too low of a BL charging current adds to the spread in time offset because of fluctuation across current sources. Nevertheless, the TDC circuit can efficiently achieve an expanded dynamic range with a counter-based structure for MSB data while using the high resolution delay line for the LSB data [75] at the expense of longer access time. This sensing scheme still has fundamental limits related to parasitic current paths from unselected bitcells overpowering very low charging currents and thermal noise—appearing as varying comparator decision time and accumulated delay deviation in the chain of TDC slices—producing random errors between signal delay and reference delay. These last two sources of error are not expected to be significant for bitcell signal levels larger than $1mV$ (from kT/C_{BL}).

4.1.4 Chip Architecture

The architecture of the 1Mb FRAM prototype is illustrated in Fig. 4.11(a). The chip is organized into 8 slices of 128 kb, each with one bit data input and one TDC to process one bit of data output. The vertical spine in the middle of the chip contains a level-shifting row driver for the elevated wordline voltage (0.7 V higher than VDD). This region also contains a decoding structure to support a hierarchical plateline configuration. Local plateline driver supplies are routed horizontally to activate one of eight plateline groups (the one common to the active WL) per bitline so that the delay, energy, and LPL-BL coupling of the vertical plateline is tolerable with 128 cells per local plateline group. A variety of FRAM designs have amortized the area of plateline driving circuits across multiple cells, most of which are unselected, [74, 70] but care must be taken so that a significant voltage does not develop

across unselected capacitors. The unwanted voltage drop comes from both a parasitic divider between ferroelectric capacitor and junction capacitance and reverse-bias junction diode current at high temperature. For this chip's technology, the dominant source of disturbance comes from parasitic junction capacitance as in [70].

The core area is $1.12mm^2$, resulting in a overall bit density of $0.936Mb/mm^2$. In this architecture, one million 1T1C cells multiplex to eight large 110 μm^2 5 bit TDCs, while preserving a memory array efficiency of 66.4%.

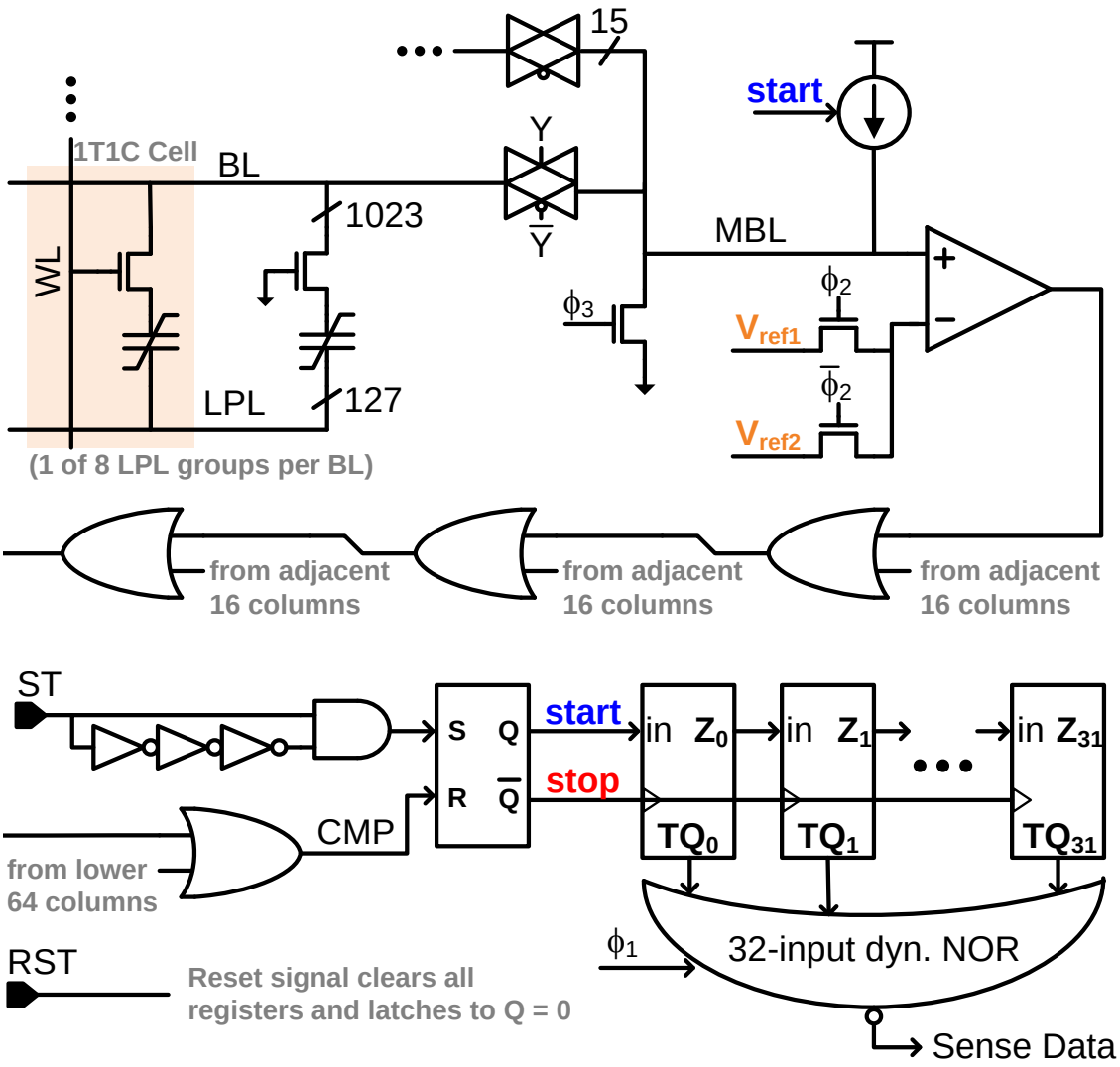


Figure 4.5: Shown is the simplified schematic of the time-to-digital read path. Not shown is decoding logic for the column select signal Y, toggle flip flops for ϕ_2 and ϕ_1, and an additional SR latch for ϕ_3, and write circuits. All ϕ signals derive from the input signal ST and comparator output CMP. Other signals not shown are VCTL of the TDC slices (see Fig. 4.7), bias voltages for the comparator and current source, and comparator power and output gating signals (see Fig. 4.8).

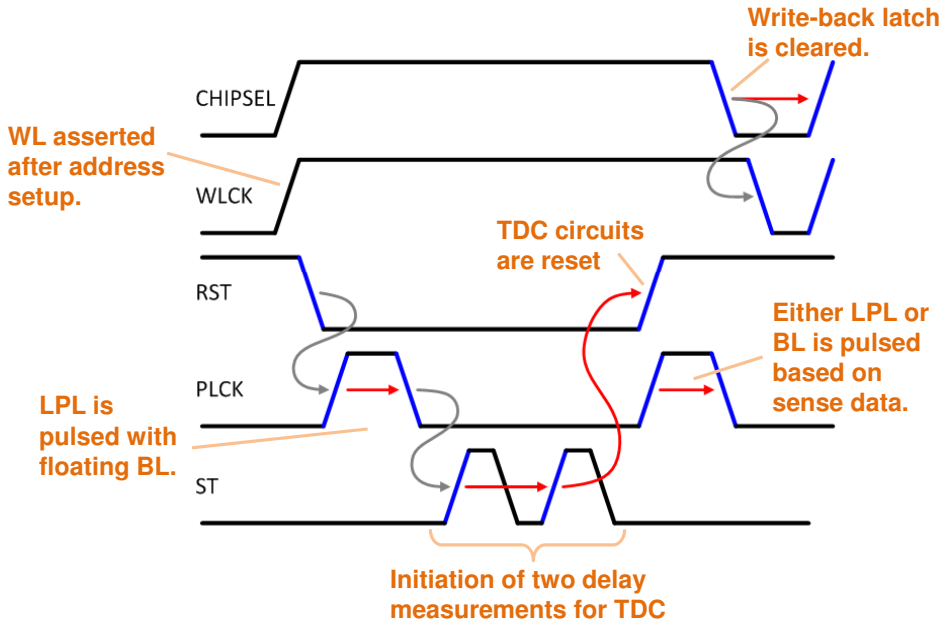
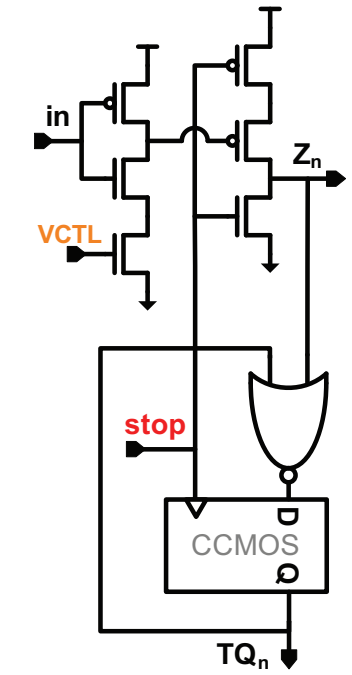
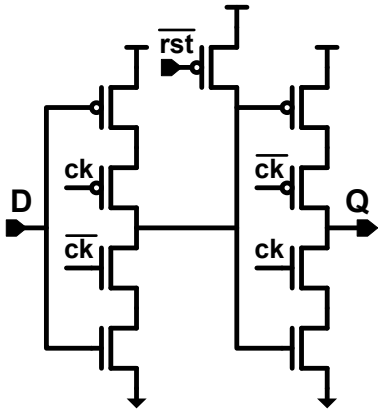


Figure 4.6: Timing diagram with description of chip behaviors

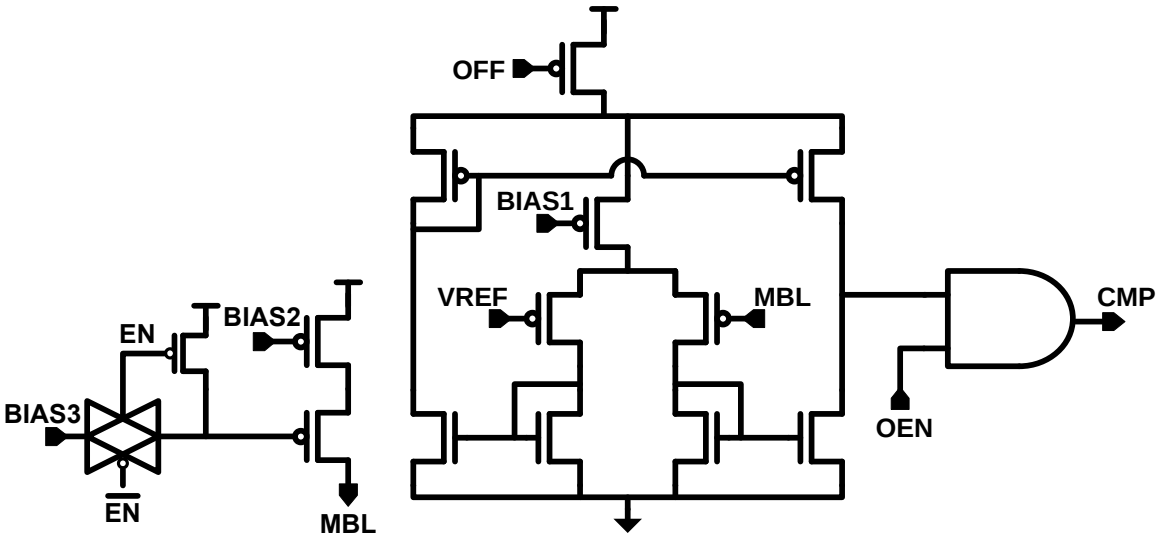


(a)



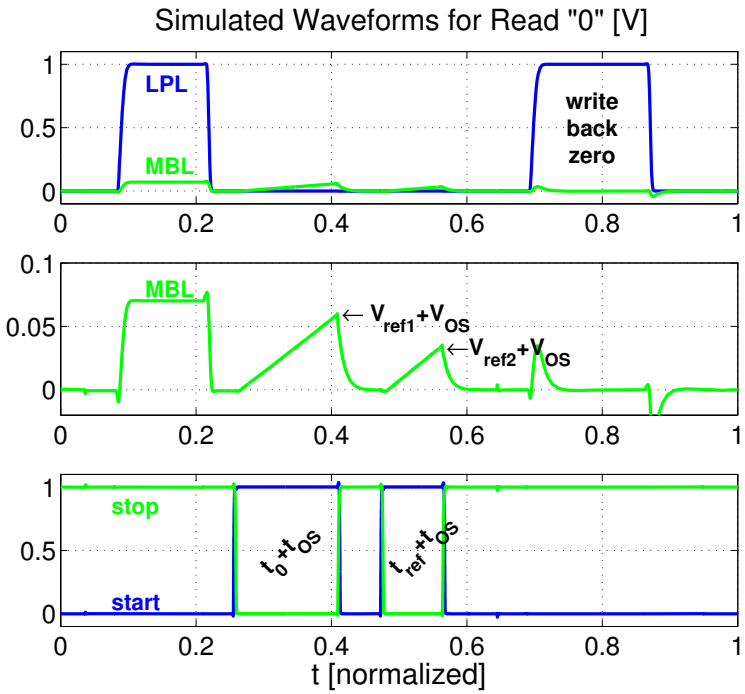
(b)

Figure 4.7: Shown is (a) the schematic of the TDC Slice and (b) the dynamic CCMOS register within the TDC slice.

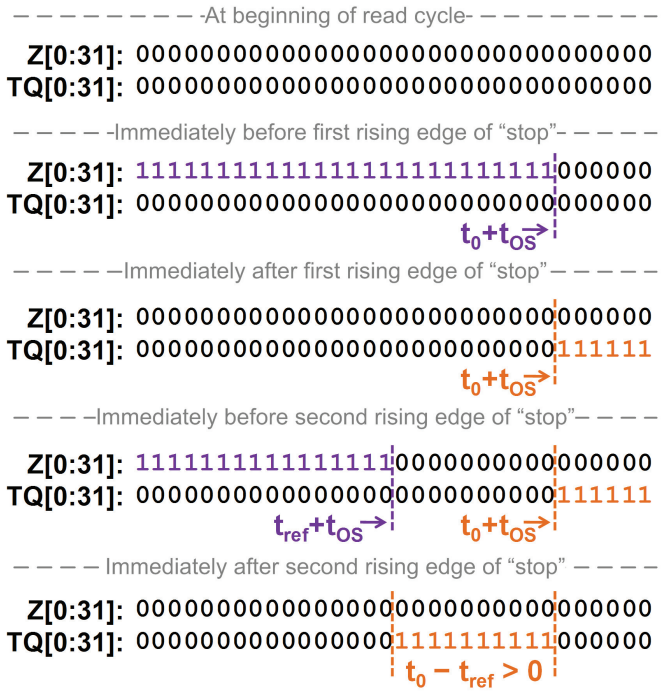


(a) (b)

Figure 4.8: Shown is (a) the current source and (b) comparator which support each group of 16 bitlines.

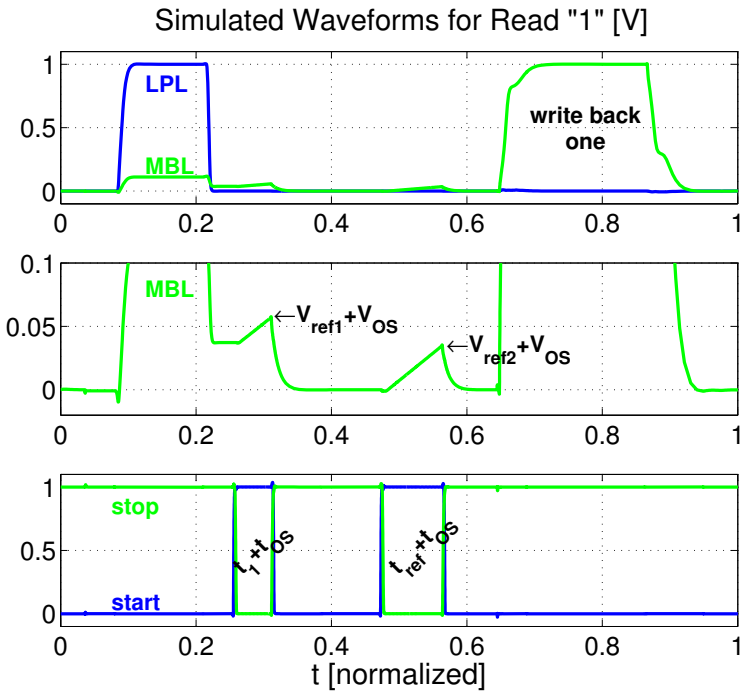


(a)



(b)

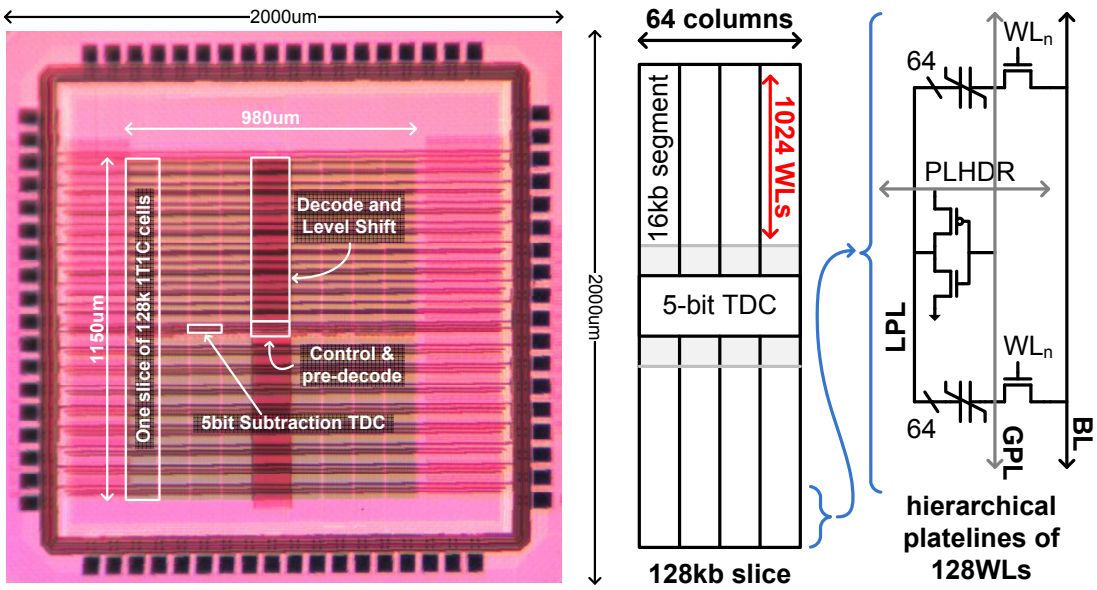
Figure 4.9: Simulated read "0" (a) waveforms and (b) snapshots of the TDC state. Normalized time of 1 corresponds to 217 ns.



(a)

(b)

Figure 4.10: Read "1" (a) waveforms and (b) snapshots of the TDC state. Normalized time of 1 corresponds to 217 ns.



(a) (b)

Figure 4.11: Shown is the (a) die photo with top-level chip architecture and (b) elaboration of the details of one of the eight 128 kb slices.

4.1.5 Chip Characterization

When characterizing the chip, voltage references and the timings in were explored under alternating checkerboard data patterns. Each of the edges associated with an arrow in Fig. 4.6 can be adjusted in this experiment with off-chip clocks to accommodate the different events in the read cycle with sufficient time. A midpoint reference that distinguishes the two data states is delivered through the voltages V_{ref1} and V_{ref2}. The quantity $(V_{ref1} - V_{ref2})C_{BL}$ corresponds to a reference charge which must be greater than a worst-case 0 and less than a worst-case 1 after capacitor relaxation.

For experimental flexibility, the reference charge in this work is delivered directly by off-chip voltages and so are the voltages for biasing the comparator and current source (BIAS1, BIAS2, BIAS3 in Fig. 4.8), and TDC delay (VCTL in Fig. 4.7). The test board in Fig. 4.12 generates the tunable references and address counting to extract the schmoo plot in Fig. 4.13, which shows the fraction of failing bits from a checkerboard pattern versus a sweep of the data reference. For low values of $V_{ref1} - V_{ref2}$, all 0's read out as false 1's (50% fail for a checkerboard) and for high values, all 1's read out as false 0's. Thus, the first transition down to zero fails reveals the tight distribution of the 0 data state, and the second transition from zero fails back up towards 50% fails reveals a significantly wider spread in 1's which is expected from the after-pulse sense operation. The $12mV$ wide floor of zero fails quantifies the additional noise that could be tolerated on the externally supplied reference voltages. Generating low-power, accurate, temperature-stable, and power-supply ripple immune references on-chip is a solved problem [76, 77, 78]. It is also anticipated that providing these references on-chip will widen the reference window of Fig. 4.13.

The data in Fig. 4.14 shows that time-to-digital sensing can scale the voltage operating point from a nominal VDD of 1.5 V to 1.0 V. The performance and energy are measured under conditions for which all bits in the 1Mb chip pass the write and read back of alternating checkerboard data patterns. Furthermore, the chip was powered down for 30 minutes at $27°C$ with VDD shorted to ground between the write and read back of both patterns to verify non-volatile operation. The read access energy scales by approximately a factor of 2 from

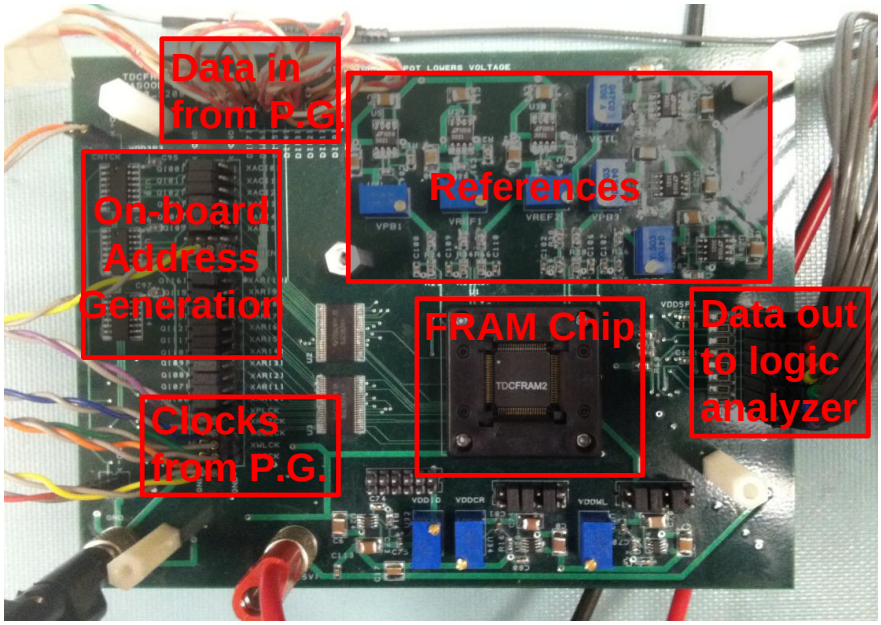


Figure 4.12: Photo of test board with address generation and reference tuning

19.2pJ to 9.8pJ per accessed bit. This energy scaling accompanies a trade-off in read cycle time going from 200 ns to 730 ns.

Also shown in Fig. 4.14(a) is the energy associated with a write cycle. This information captures the energy related to toggling the high capacitance nodes of the local plateline and bitline once in addition to the decoding energy. A read operation consumes the same energy as write plus additional energy from a second toggling of the local plateline, activity within the TDC, and static power from the continuous-time comparator. Simulation of the comparator with the settings used during testing, gives $43.5\mu W$ and $7.7\mu W$ of static power at $1.5V$ and $1.0V$ respectively. During a read cycle, the comparator is power gated off for 57% of the cycle, so the simulated energy consumption of the comparator is $3.7pJ$ (1.5 V) and $2.4pJ$ (1.0 V) for each accessed bit. For either write or read, the boosted WL did not consume a significant amount of energy because of its comparatively small switched capacitance. During idle mode, the 1Mb macro consumes leakage power of 251 nW (at 1.5 V, 27° C) to 95 nW (at 1.0 V, 27° C). This is 3,000 to 1,000 times less than the active read power, and there is no additional wake-up time needed before an access.

Summarized in Table 4.2 are the key physical and electrical features of the low-access-

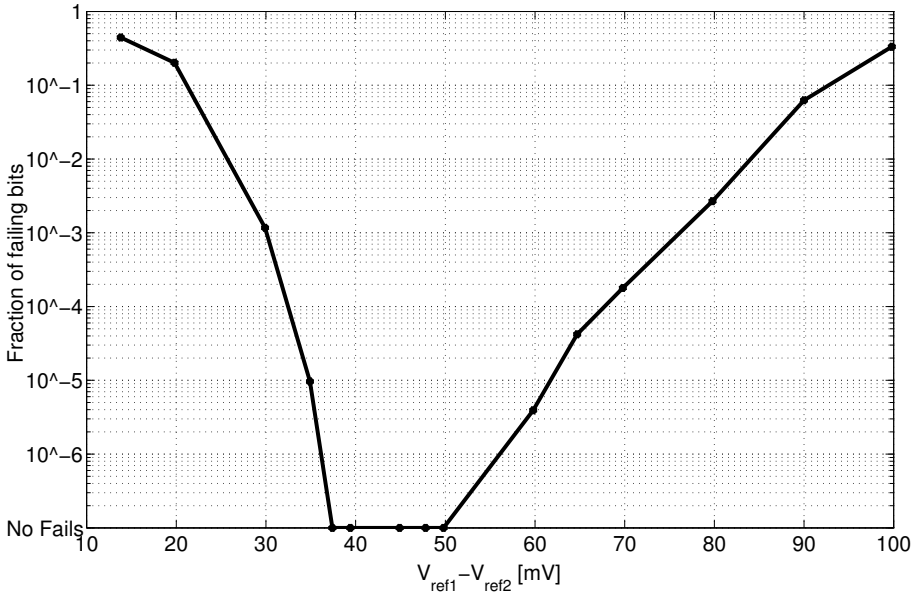


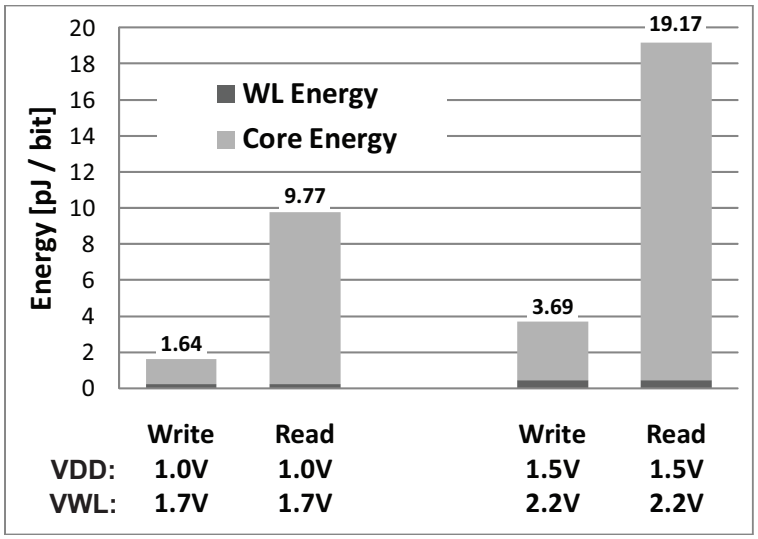
Figure 4.13: Shown is a sweep of fail count (checkerboard pattern) versus signal reference. V_{ref2} is fixed to 120 mV. Chip conditions are: 200 ns read cycle, VDD = 1.5 V, VCTL = 0.82 V (estimated ramp time of 12 ns), and estimated bitline charging current of 18 μA. A similar curve is observed for the complimentary checkerboard pattern.

energy FRAM chip.

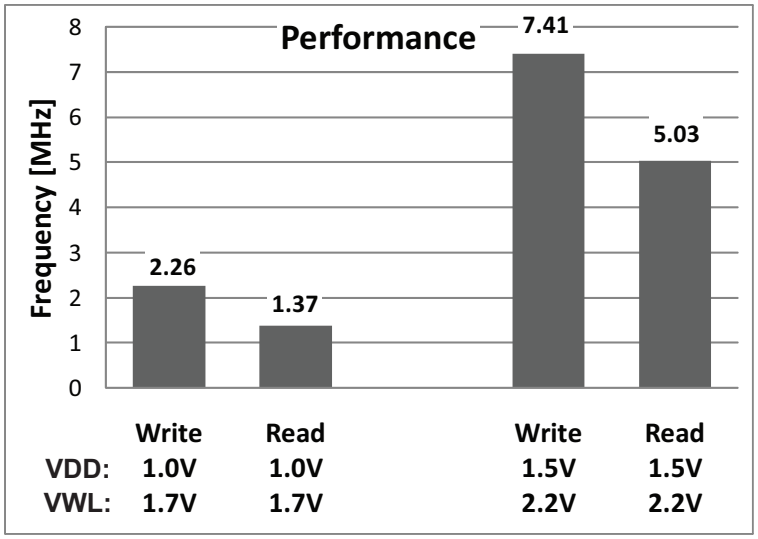
Table 4.2: 1Mb FRAM Chip Summary

Organization	128k words of 8 bits
Technology	130 nm LP CMOS
Array Efficiency	66.4%
Full Macro-level Density	0.936 Mb/mm^2
Operating Voltage (CORE / WL)	1.5 V / 2.2 V to 1.0 V / 1.7 V
Read Cycle Time	200 ns to 730 ns
Read Power	772 μW to 107 μW
Idle Power	251 nW to 95 nW
Standby Power	0 W

* Measurements in this table are at 27° C.



(a)



(b)

Figure 4.14: Measurement reveals the scaling of (a) energy with (b) accompanying trade-off in performance.

4.2 Concluding Remarks on TDC Sensing

A a 1Mb FRAM prototype targeting low-access-energy has been described. As supply-voltage is lowered to achieve this target, the statistical signal of the ferroelectric capacitor reduces. Conventional sense amplifiers based on clocked comparators exhibit a fluctuating offset that causes bits to fail before the bitcell signal completely disappears with lowering supply voltage. Using a time-to-digital converter instead provides a higher resolution circuit that captures both the signal and the offset of the sensing devices, permitting the reduction of overall sensing circuit variation. Representing the signal as time to process it with a TDC required architectural modifications to the cell array in order to accommodate the high degree of column multiplexing, and it also required the custom development of a delay-line TDC that embeds subtraction functionality. Measurement results verified the capability to scale the supply voltage and achieve a 2x reduction in access energy. The low-voltage circuits employed in this time-to-digital sensing approach ensures that the minimum supply voltage will be limited by the ferroelectric capacitor and not by the peripheral circuits. This sensing scheme also remains compatible with a variety other FRAM circuit developments such as the generation of an optimum midpoint reference to track capacitor imprint and global variation. It is also compatible with low-offset comparator design techniques to attain further margin.

Through the example of reducing FRAM energy consumption, the significance of time-to-digital sensing emerges. It efficiently compensates analog offset with digital circuits. It provides a sensing scheme that scales in voltage, minimizes static power, and gets better with technology scaling. And, representing the signal as a delay better serves the essential multiplexing operation of memory. Recalling that in many cases the time scale of the bitcell relative to a CMOS gate delay is large, time-to-digital sensing will enable a broad class of semiconductor memories to achieve their full potential.

Chapter 5

Non-volatile processing

This chapter will present the design and implementation of a custom standard cell register that can retain its data during power interruption. Furthermore a power management unit (PMU) accompanies the non-volatile D flip-flop (NVDFF) with the integration into a flow that takes arbitrary RTL and produces a circuit that never has to reboot, and hence has no distinction between on and off. The result of the proposed work is a method to produce a digital circuit (as specified by RTL) that automatically computes when power is available and retains its complete state for every internal node while power is unavailable. This desired operation is illustrated in the cartoon of Fig. 5.1. The circuit and design methodology presented herein is useful for energy harvester applications, low power portable electronics [8], and even enterprise applications that aggressively power gate idle portions of digital integrated circuits [79].

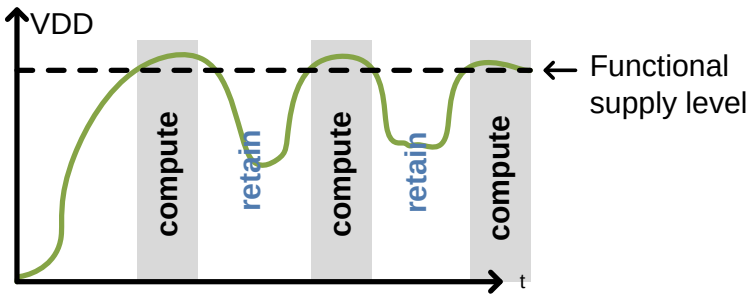


Figure 5.1: Cartoon of desired feature of non-volatile operation

5.1 Non-volatile DFF

The emergence of non-volatile memory technology alternatives to FLASH presents exciting new opportunities as discussed in the previous chapter. Because many of these technologies, including the ferroelectric capacitor (fecap) technology explored in this work, are compatible with a digital CMOS process, they can also be embedded into registers instead of being built into high density memory arrays. Prior to the availability of non-volatile technologies, retention registers had been developed to enable very low power standby modes for portions of a circuit that are power-gated or even entire chips [80]. Such a solution requires an always-on power rail that adds system-level complexity in the form of an extra power pin and a specification for a battery to always supply static current to the state elements.

For this reason, the register in [81] employs magnetic tunnel junctions as the state elements. By using the state-dependent resistance to skew the impedance to ground in a cross-coupled inverter pair, self timed operation is achieved. However, this design suffers from large currents ($1mA$) and in-turn large devices to support the write current requirements for the magnetic tunnel junctions. Another related work presents an non-volatile register that stores its state in ferroelectric capacitors [3], similar to the approach in this thesis. The schematic is shown in Fig. 5.2.

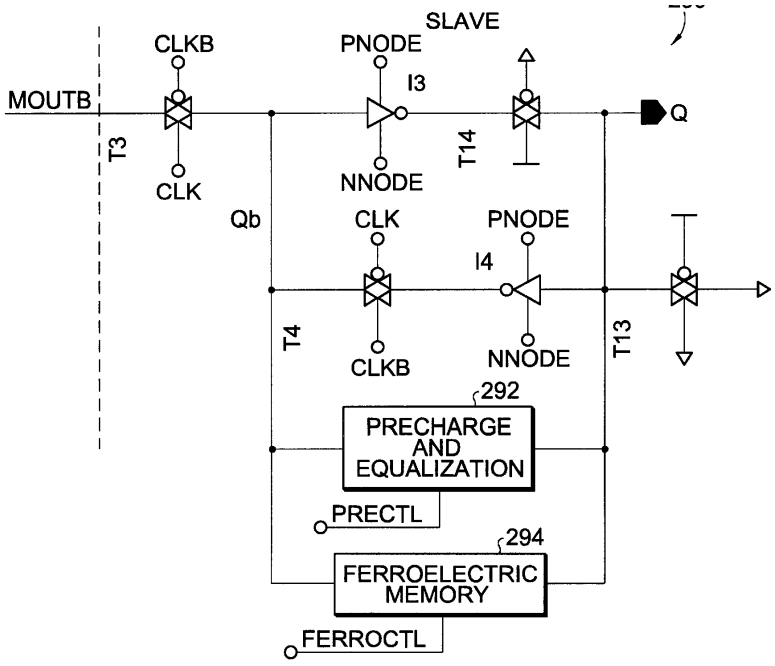
The most critical operation for the non-volatile register—also known as non-volatile D flip-flop (NVDFF)—is the ability to sense an analog signal stored in the hysteresis of two embedded ferroelectric capacitors (a technology described in the previous chapter) and place the result in the slave stage latch. The conventional NVDFF programs the ferroelectric capacitors by driving a positive or negative supply voltage across the capacitor terminals. During restore, the voltage divider of the two branches of series capacitors are alternately above and below $VDD/2$ when a voltage is pulsed across both ferroelectric capacitor dividers. After a pre-determined time, a cross-coupled inverter pair is enabled and amplifies the small signal to the logic state of the slave stage.

Shown in Fig. 5.3 is a simplified schematic of the proposed solution along with simulation waveforms during a restore operation. Two identical current sources charge up the ferro-

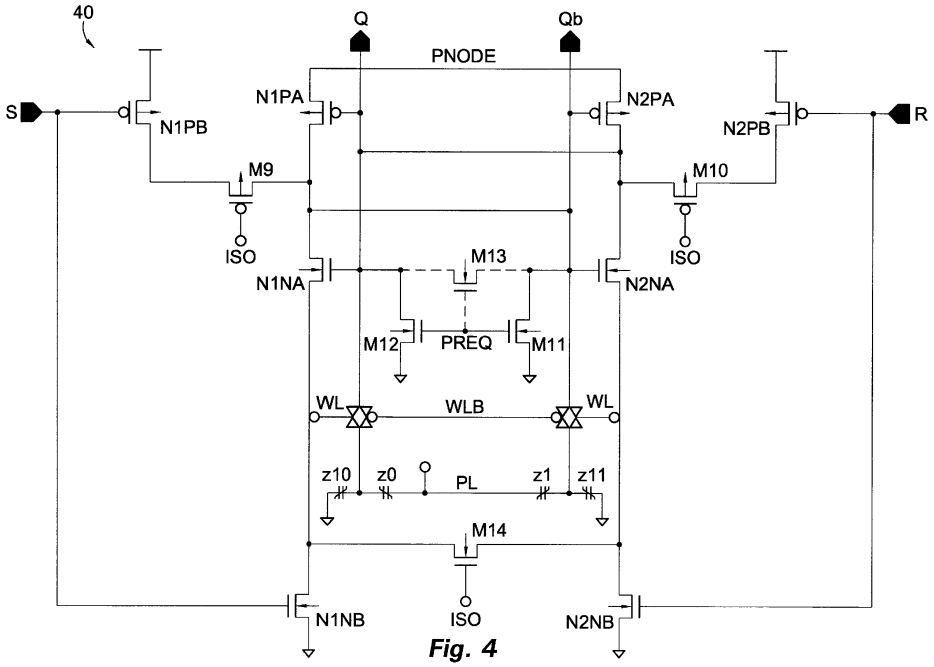
electric capacitors and the difference in charge due to the state of the capacitors develops a difference in voltage. The node with the higher voltage (in this case FET) will turn on the diode in series with the PMOS header device in the cross-coupled inverter pair. As soon as this path conducts, the internal node (QT) rises and cuts off the path from FEC to QC, latching the data with positive feedback. The internal latch node QT quickly rises without disrupting the continual accumulation of signal on FET because of the significant difference in capacitance. Namely, the small swing on the high capacitance nodes FET and FEC translates to a large swing on the smaller capacitance nodes QT and QC. This charge sharing technique is applied in contexts where an analog signal needs to be detected on a large capacitance (often times a consequence of large fan-in) node [82].

The simulation comes from a post-layout RC extracted netlist of the flip flop so the initial trajectory of QT and QC are not perfectly matched, but the placement of metal shapes has been constructed to match the coupling to both nodes so that almost all of the asymmetric voltage comes from the actual state of the ferroelectric capacitors.

The full schematic of the slave stage with ferroelectric interface circuits is shown in Fig. 5.4(a). The circuit exhibits four different modes based on the state of the two control signals PG and LD which are described in Table 5.1. As a central design objective, it must be ensured that there are no glitches on the ferroelectric capacitors that can corrupt the data. The transition between modes is therefore accompanied with the transition of only one of the two signals. The schematic of Fig. 5.4(a) shows an additional signal EQ which is conservatively added to the non-volatile state management scheme to clear voltages across the fecaps and prevent unwanted coupling to build up a parasitic voltage.

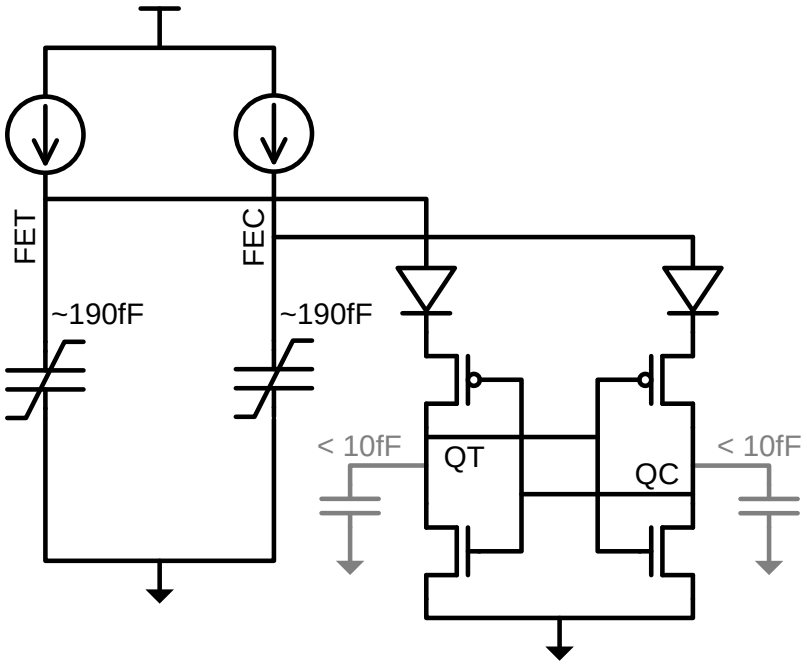


(a) Slave stage

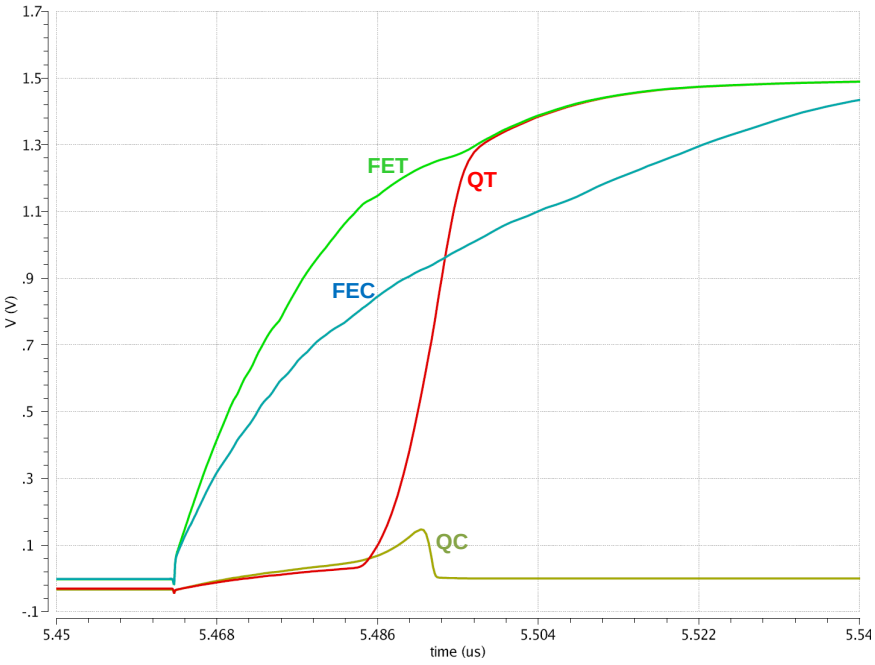


(b) Ferroelectric interface circuits

Figure 5.2: Schematic of conventional FRAM NVDFF from US Patent #6650158 [3].

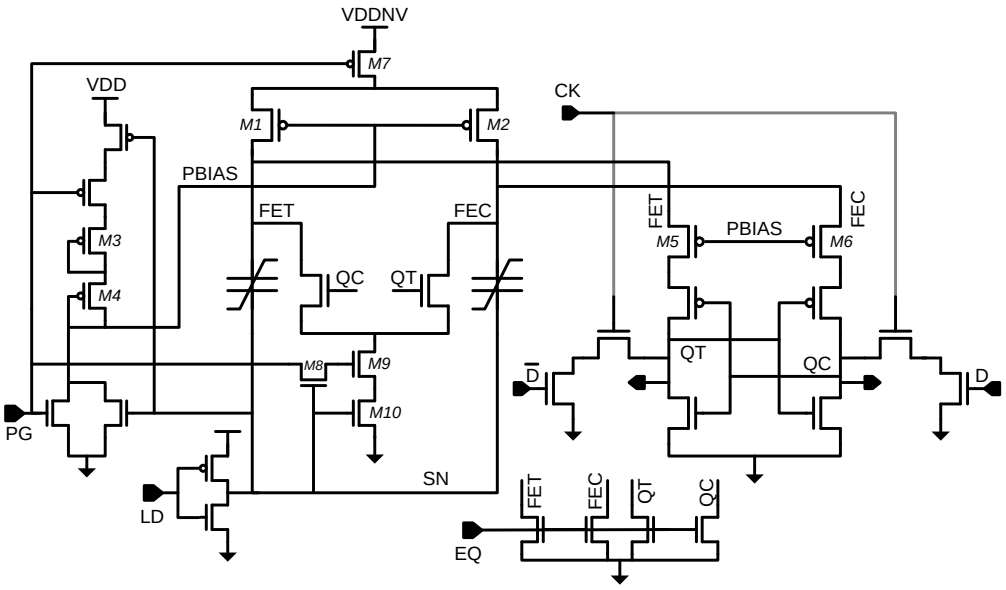


(a) Simplified schematic during restore

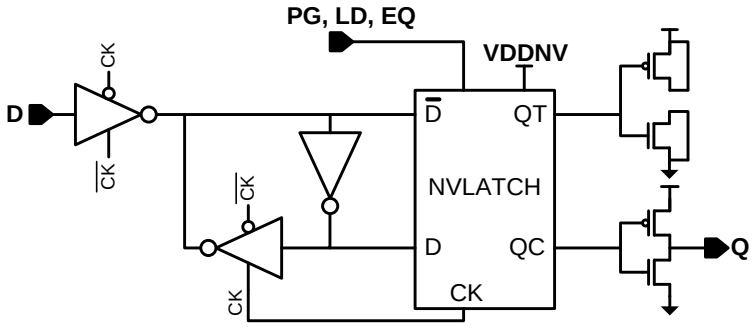


(b) Associated waveforms

Figure 5.3: Simplified description of restore operation



(a) NV Latch



(b) NV DFF

Figure 5.4: Schematic of proposed NVDFF

Table 5.1: State table of NVLATCH modes

PG	LD	Mode
1	1	Off
0	1	Restore
0	0	Active
1	0	Save

The two PMOS transistors connected to FET and FEC act as the current sources. They are sized as long-channel devices to set a low level of charging current and minimize offset. They are coarsely biased by two PMOS diode drops, ensuring that a functional bias point is set across all global process corners. The node PBIAS also sets the diode drop, or threshold at which the slave latch triggers. This is important because voltage bias across the capacitors is a critical issue for sensing dynamics.

Fig. 5.4(b) shows the additional devices needed to construct a full DFF from the circuit. Not shown are additional devices that implement a asynchronous active-low reset (by breaking feedback in the master stage and adding extra NMOS inputs to the slave stage jam latch). The detailed operation of the NVDFF across all modes is shown in Fig. 5.5. In this simulation, the NVDFF begins in the active operating mode and goes through a save, off, restore, active sequence.

Going step-by step, the operation of the NVDFF and internal behavior of the slave stage non-volatile latch is as follows:

Save (PG=1, LD=0, EQ=0). Entering this mode, PG goes high, and, with LD low, the pull-down NMOS stack turns on to write back a "0" to one of the two fecaps depending on the state of QT/QC.

Off (PG=1, LD=1, EQ=1). In this mode all internal nodes are cleared to ground through the equalization devices driven by EQ. PG being high cuts off the VDDNV rail, and LD being high grounds node SN. Under this condition it is safe for VDDNV to be discharged, and then the main system VDD can lose integrity too.

Restore (PG=0, LD=1, EQ=0). In this mode, the current mirror PMOS

devices connected to nodes FET and FEC are biased in saturation by two PMOS diodes in series with VDD. The PMOS diodes are enabled by PG going low. Also during this mode, both capacitors are written to "1" as their signal is sensed.

Active (PG=0, LD=0, EQ=0). In this mode, nodes FET and FEC act as virtual supply rails with the node PBIAS pulled to ground. The signal LD goes to "0" so that both ends of the ferroelectric capacitors are at 1.5V.

Probing the current into the register supply (VDDNV and VDD), it is observed that 1.3 pJ of energy is consumed during a round-trip save and restore operation. To design a system using the NVDFF to preserve data during *unpredictable* power interruption, this energy times the number of registers (on the order of $10nJ$ for a microcontroller) sets the requirement on the energy buffering at the system level.

This timing diagram also reveals the purpose of a secondary rail, VDDNV. This second rail is shut down before VDD loses integrity and this second rail is charged up after VDD establishes integrity (this integrity is monitored by another chip that is described in the next section). By ensuring that the path from either end of the fecaps through any transistor channel terminates at either ground or VDDNV, one can guarantee the fecaps will not be corrupted even if the control signals PG, LD, EQ lose their correct values during VDD interruption. For conservatism, the EQ signal clears the nodes FET, FEC, QT, QC to VSS after saving and before restoring.

The proposed work is compared against the conventional approach in Fig. 5.2 which is best thought of as a DFF with a one-bit memory adjacent to it. One significant issue is the sense amplifier timing of the one-bit memory. This approach requires the delicate handling of a race condition between ferroelectric signal and sense timing, which is hard to protect in the context of an arbitrary placed-and-routed digital circuit. Also, it has been shown that the dynamics of ferroelectric domain switching—the underlying source of signal—slows exponentially with reduced voltage bias [83]. Because available simulation models do not capture this internal delay mechanism, one must take care to design a non-volatile flip flop to interrogate the voltage of the capacitors only when significant bias has been developed.

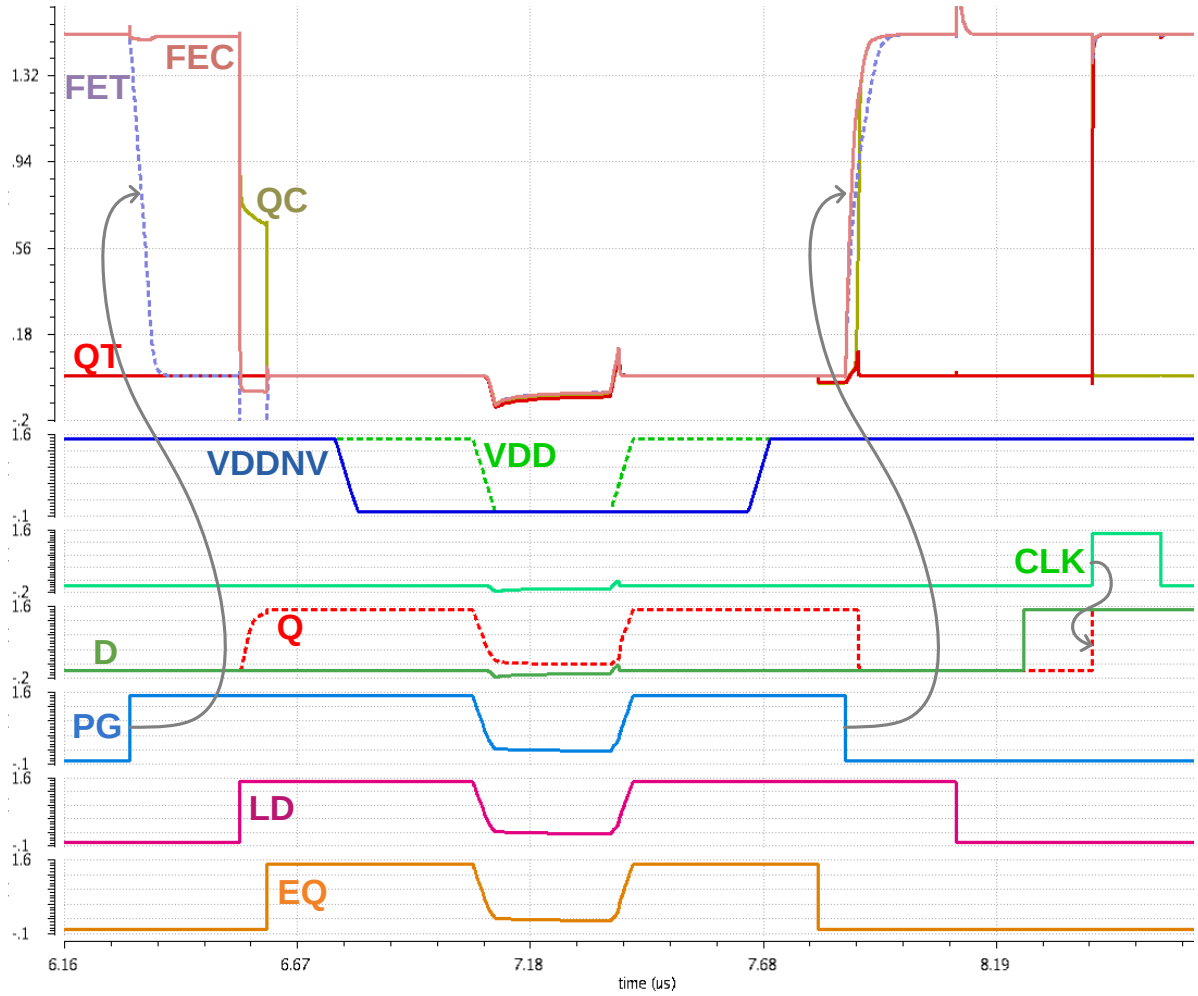


Figure 5.5: Timing diagram and operation waveforms of NVDFF

A third issue for integration is peak current. If the peak current exceeds the capability of the system decoupling capacitance, the functional VDD can collapse and corrupt the state of the PMU resulting in erroneous operation. For the proposed circuit, the pull-down network for write-back and the current source devices are sized to draw less than $10\mu A$ per register (as are the charging current sources during restore), so that a few thousand registers can be programmed or read in parallel.

Shown in Table 5.2 is a quantitative comparison of the proposed non-volatile slave latch (Fig. 5.4) and the conventional non-volatile slave latch (Fig. 5.2). The energy is not directly 2x less because during the sense operation the conventional approach does not apply a full

swing to all the capacitors. However, during write-back, a full swing must be applied to four capacitors which more than offsets the energy advantage of restore. Both designs are compared using identically sized capacitors, which are usually targeted to be minimum size capacitors. Although the energy comes from spice simulations including the sense and write transistors, the energy is dominated by how the ferroelectric capacitors are toggled. Both designs result in similar amount of differential voltage signal. One key limitation is voltage bias versus sensing time. Data in [83] shows that the electric field resulting from 0.6V in this process technology ([65]) will take over $10\mu s$ to develop 80% of the stored signal; whereas, it has been shown in [70] that operation at 1.2 V bias or higher extracts sufficient signal under $100ns$.

Table 5.2: Comparison with conventional approach (energy is normalized)

	conventional	proposed
Restore energy	0.16	0.38
Save energy	0.84	0.20
Total energy	1.00	0.58
Nominal differential signal (static model)	383 mV	309 mV
Voltage bias across switching cap	0.50 V — 0.62 V	1.10 V—1.30 V

The layout of the NVDFF is sketched out in Fig. 5.6. It is compatible with a standard cell library in a 130nm CMOS technology. The cell uses only one level of metal and occupies two standard cell rows, whose horizontal running rails automatically connect the power supply VDD (middle rail) and ground node VSS (outer rails).

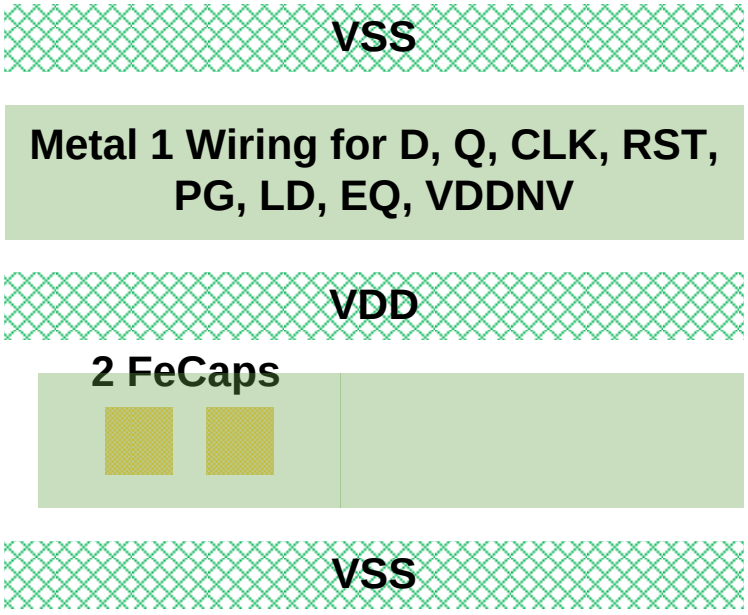


Figure 5.6: Cartoon of the layout of the 60.7μm^2 non-volatile D flip-flop (NVDFF) showing the first level of metal and ferroelectric capacitor shapes.

5.2 Integration of NVDFF into the design flow

Fig. 5.7 shows the complete expression of the non-volatile computing scheme. It contains (1) a digital circuit—synthesized from arbitrary RTL that is agnostic to non-volatile operation—with embedded NVDFFs, (2) a power-management unit that that stops or resumes computation based on energy availability, and (3) an energy harvester chip ([84]) that gives sufficient warning to shut down such that enough energy is available to save state.

To provide the correct stimulus for the NVDFF as described in the previous section, a PMU is implemented that exhibits the state transition diagram in Fig. 5.8. It locks itself into reset with a custom SR latch designed to set to zero upon power-up. The same SR latch is used to drive the VDDNV rail, so that during power interruption when the state of the PMU may perhaps become corrupted, the fecaps are still protected.

A digital design flow, outlined in Fig. 5.9 is implemented. The NVDFF timings are characterized to generate a *.lib file for the synthesis and back-end tool. Additionally a *.lef layout representation for the back-end tool is created. During synthesis, the tool is instructed to avoid all volatile DFFs and use only the NVDFF along with the logic gates

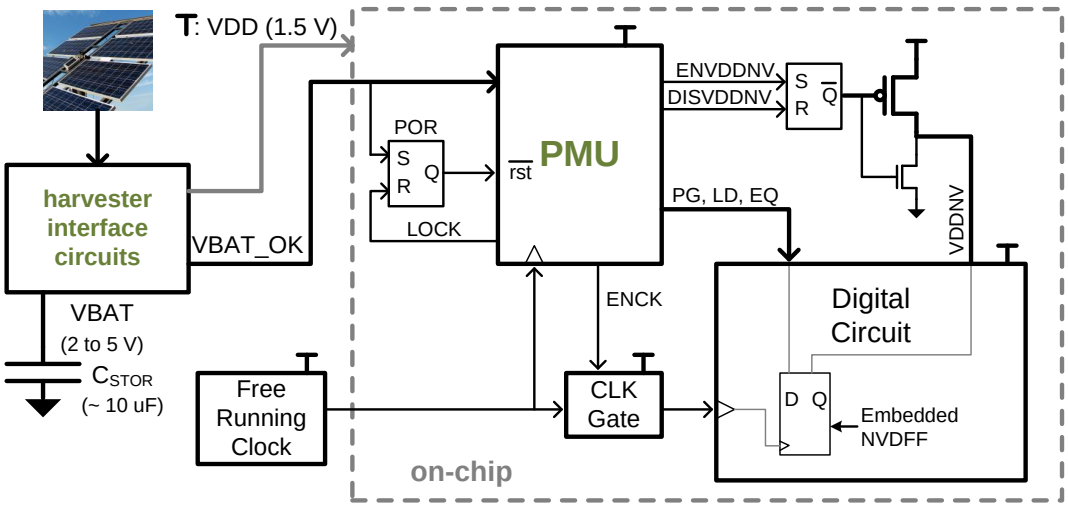


Figure 5.7: Block diagram of non-volatile power management scheme

in the standard cell library. During the back-end stage, extra buffer trees are synthesized so that the PMU can drive the PG, LD and EQ ports of all the NVDFFs in the digital circuit. These buffer trees have a simple constraint of maximum delay from input port to register to be equal to $10ns$. For the target clock period of $200ns$ and restore time of $50ns$, these timings for the buffer trees ensure robust operation from a PMU that can generate one edge transition per clock cycle for each non-volatile control signal. Furthermore, an extra global rail for VDDNV is implemented so that all NVDFFs have short connecting routes to the common VDDNV. In the actual chip, VDDNV is split into VDDNVT and VDDNVC (associated with FET and FEC) for debugging purposes. In the final layout, the NVDFF has both its volatile (D, CLK, Q, CLRZ) and non-volatile (PG, LD, EQ, VDDNT/VDDNVC) ports automatically routed by the tool.

The mechanism to actually initiate system-wide save and restore exists inside the energy harvester interface circuitry, illustrated on the left hand side of Fig. 5.7. At the heart of the harvester interface is a chip that takes solar, thermal, or vibration energy and stores it as a voltage on a capacitor [84]. This voltage, V_{VBAT}, at node VBAT falls if the average power consumed by the system is higher than the average power harvested. Under such a circumstance, V_{VBAT} will eventually decrease to a point such that only enough energy to save system state remains. This condition can be detected as a programmable voltage threshold

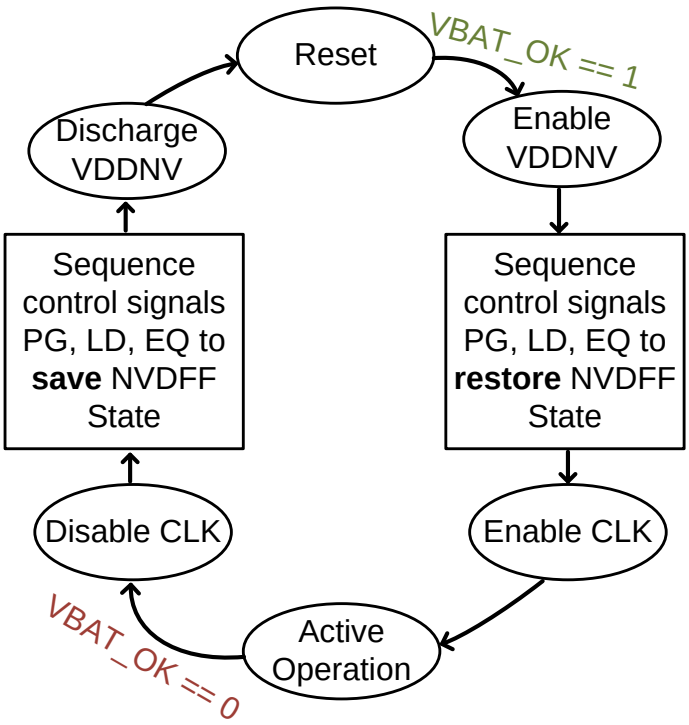


Figure 5.8: High-level diagram of finite-state machine for PMU controller

on V_{BAT} set by a resistor voltage divider, which will cause the harvester chip to transition the signal VBAT_OK from high to low. A separate resistor divider sets the threshold for the rising transition of VBAT_OK. By establishing hysteresis between the two transitions of VBAT_OK, one can ensure the system does not get stuck in a loop repeatedly turning on and off.

In order to correctly determine the settings for the two thresholds, one must know the amount of energy that is consumed by a save operation and a restore operation. For a digital circuit with embedded NVDFFs, the ferroelectric capacitors dominate the energy so the calculation is simply 1.3 pJ times the number of NVDFFs. Although there is a significant energy cost associated with saving and restoring, a scenario of battery-less operation requires a save operation whenever VBAT_OK goes low because the power source is unpredictable and there is no guarantee on the time until the next rising edge of VBAT_OK.

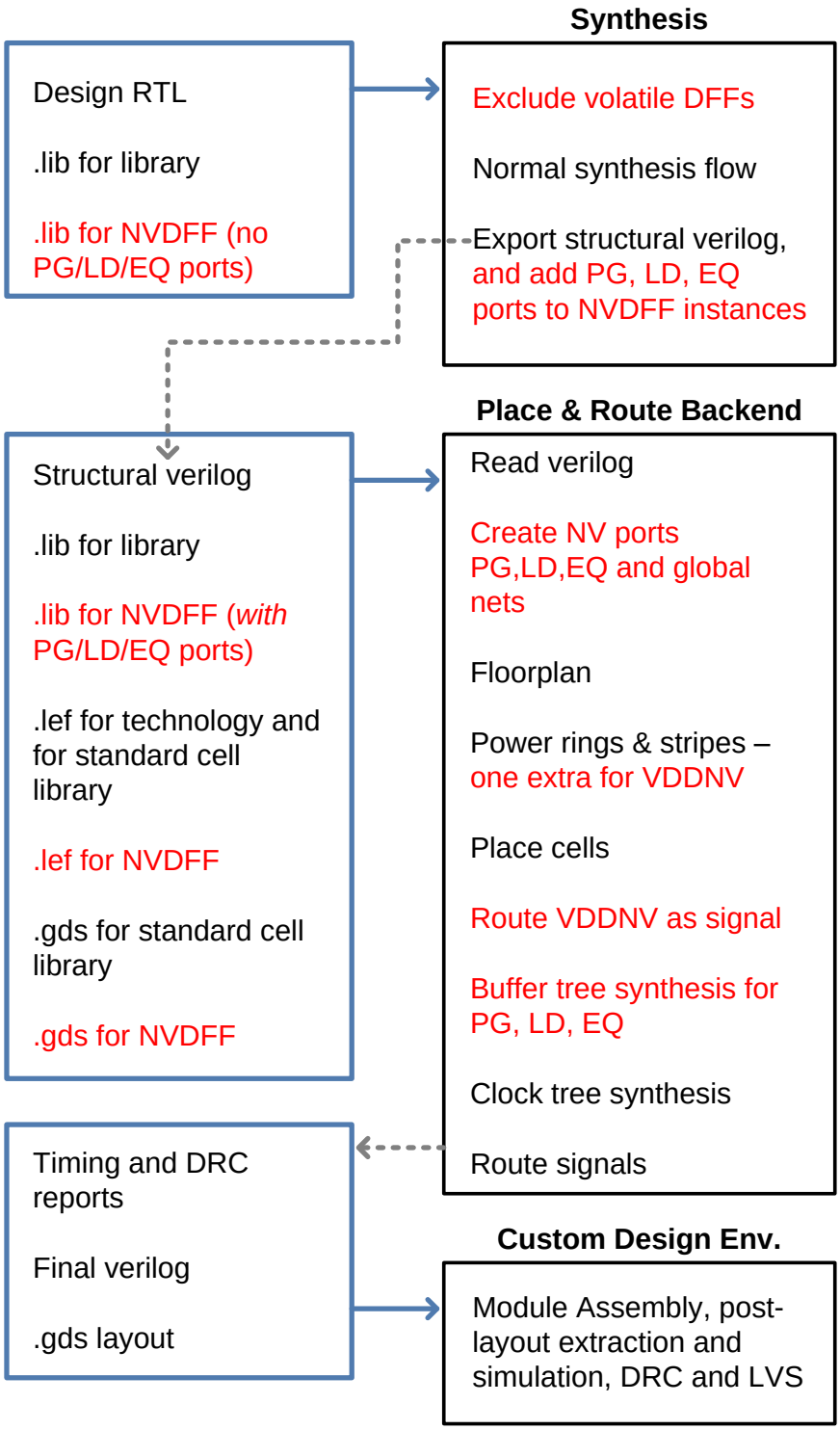


Figure 5.9: Outline of NVDFF modifications to the digital design flow (shown in red)

5.3 FIR Filter demonstrator chip

In order to gather data about the robustness, timing, and energy metrics of the NVDFF, it is important to establish a realistic context. Therefore, a FIR filter (shown in Fig. 5.10) is implemented to create a digital design environment for the NVDFF. The FIR filter has three taps, and buffers the input data for four cycles, implementing the relation:

$$y[n] = w_3 \cdot x[n-7] + w_2 \cdot x[n-6] + w_1 \cdot x[n-5]$$

The weighting factors w_i are programmable by a shift register of 24 bits (8 bits per weighting), and they default to $(w_1, w_2, w_3) = (53, -21, 89)$ upon the assertion of the active low reset signal. All arithmetic is signed two's complement and the bit width of the inputs is 8 bits, and the bit width of the output is 16 bits.

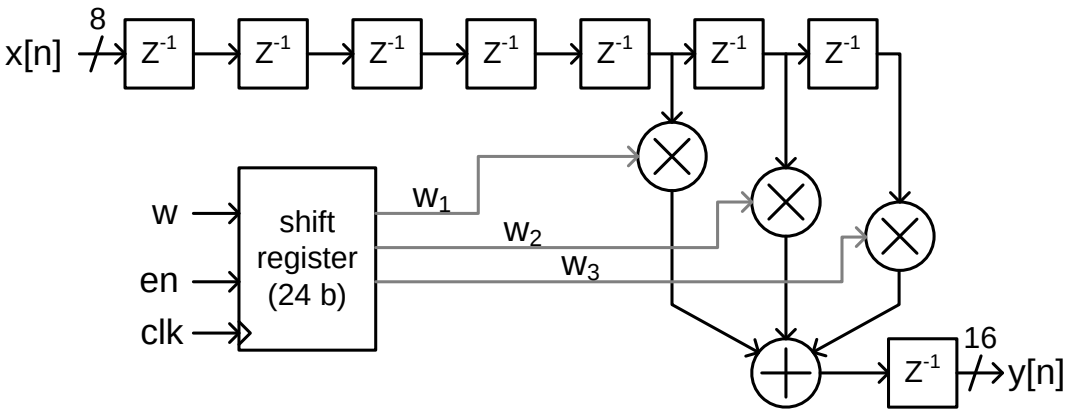
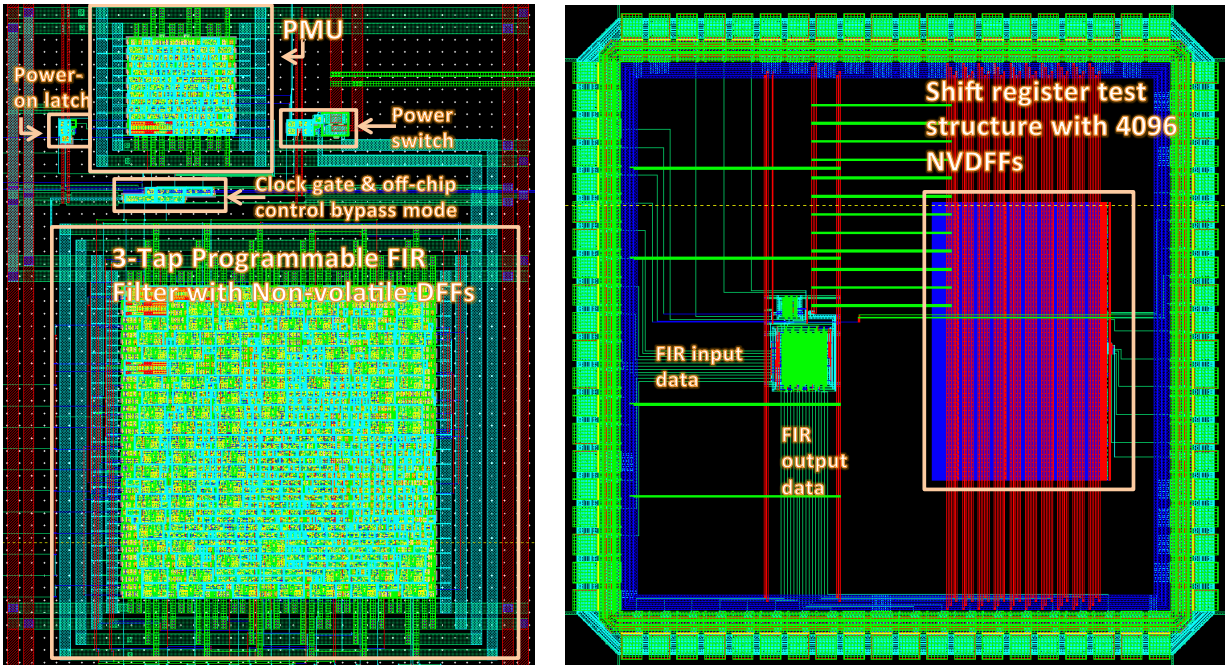


Figure 5.10: Block diagram of FIR test case

After the digital circuit (in this case FIR) and the PMU are implemented, they are assembled in a top level layout with three other pieces (1) clock-gating latch, (2) SR latch that resets to Q=0 during power-on, and (3) power switch for the quasi-power rail VDDNV. The final layout is shown in Fig. 5.11(a).

The area overhead of using a NVDFF instead of the normal DFF is evaluated based on the synthesis report of cells and their cumulative area in Table 5.3. This analysis assumes similar utilization and routing difficulty at the back-end stage.



(a) FIR module with PMU

(b) Full chip with NVDFF test structure

Figure 5.11: Layout view of non-volatile FIR demonstrator chip

For battery-less operation, it must be ensured that at least 125 pJ, based on 96 NVDFFs times 1.3 pJ of save and restore energy per NVDFF, is available before the harvester can no longer deliver a reliable 1.5 V power supply to the chip. For storage capacitance on the order of $10uF$ and a minimum functional V_{VBAT} of 2.5 V, very little excess voltage on VBAT is needed to ensure sufficient energy. The thresholds are chosen to be $100mV$ apart for robust separation (e.g. rising threshold of 2.7 V and falling threshold of 2.6 V). Even for a microcontroller with up to 10,000 NVDFFs, the difference between 2.6 V and 2.5 V on

Table 5.3: Evaluation of area overhead

Total standard cell area of non-volatile FIR	11020 μm^2
Area of gates	5194 μm^2
Area of NVDFFs	5826 μm^2
Area of equivalent number of DFFs	2184 μm^2
Total Area of equivalent volatile FIR	7378 μm^2
Overhead based on synthesis area report	49%

VBAT ensures sufficient energy to save state.

For other applications in which power-gating is part of a strategy to reduce the power of idle circuit blocks within a larger, actively operating circuit, the energy cost of save and restore determines a break-even time that influences the choice of power gating strategy. Break-even time is determined from:

$$T_{BE} = \frac{E_p}{P_{hi} - P_{lo}} \tag{5.1}$$

Where E_p is the energy penalty from toggling power switches—and in this case includes the NVDFF save and restore energy, and $P_{hi} - P_{lo}$ is the difference in the leakage power of the circuit when fully on and when power-gated. Because the NVDFFs have significant energy cost, they will significantly increase the required amount of "off" time to make power-gating worthwhile relative to an always-on retention register approach as in [80].

Using a basic rule of thumb that, for power-gated circuits, about 10% of the standard cell area is dedicated to a power switch (to ensure IR drop integrity) [85], along with the observed ratio of MOS gate area to standard cell area, the size of the power switch capacitance can be estimated. In the context of the FIR test case implementation, an increase in T_{BE} by a factor of 21.2X is estimated compared to an approach with an always-on retention register. The denominator of Eq. 5.1 does not differ significantly under both cases—in one case an always-on retention register consumes a non-dominant fraction of standby current relative to the rest of the power-gated domain, and in the other case a non-volatile DFF consumes zero standby current. The difference comes from the 21.2X change in energy penalty as shown in Table 5.4. Since, the conventional break-even time typically results in 10 clock cycles for CPU applications [85], the NVDFF approach would make sense for longer periods of planned power-gating closer to 200 clock cycles.

Returning to the battery-less demonstration of this work, the system integration of the PMU with the non-volatile digital circuit is verified with a transistor-level (with parasitic capacitance) simulation of the entire chip in Fig. 5.11(b). The simulations waveforms are in Fig. 5.12. In Fig. 5.12(a) the FIR filter is computing with coefficients (w1,w2,w3) of (87,-

Table 5.4: Estimate of break-even time overhead based on FIR test-case

Circuit component	Energy cost
Power switch for NVFIR	6.26 pJ
Fecaps in 96 NVDFFs	125 pJ
VDDNV switch and global rail	1.51 pJ
Energy penalty increase	21.2 X

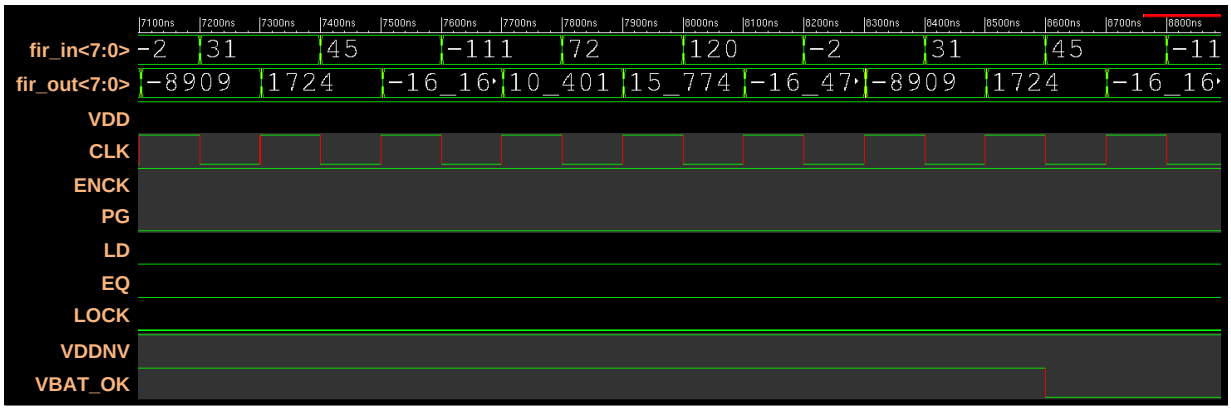
77,-98) and, towards the end, the harvester IC indicates that power is going to disappear with the fall of VBAT_OK. Then, the second set of waveforms in Fig. 5.12(b) shows how the PMU enters the appropriate sequence of commands by disabling the clock, then sequencing PG high, then LD high, then EQ high, discharging VDDNV, and finally locking itself into the reset state. Next the main power supply to VDD is lowered, removing all power sources to the chip. As a result the control signals are corrupted, however restoring the power supply keeps the PMU in its correct reset state and the chip is subsequently prepared to correctly sequence EQ low, then PG low, then LD low, and enable the clock to resume computation.

It is seen that after power is restored, the first three output values are consistent with the programmed coefficients (w1,w2,w3) of (87,-77,-98) and the samples prior to power loss. The subsequent samples remain consistent with the programmed coefficients (not the default coefficients) which verifies non-volatile operation. In this toy example, the power-up of the FIR filter avoids the 24 clock cycles that a volatile circuit would require to re-program the coefficients.

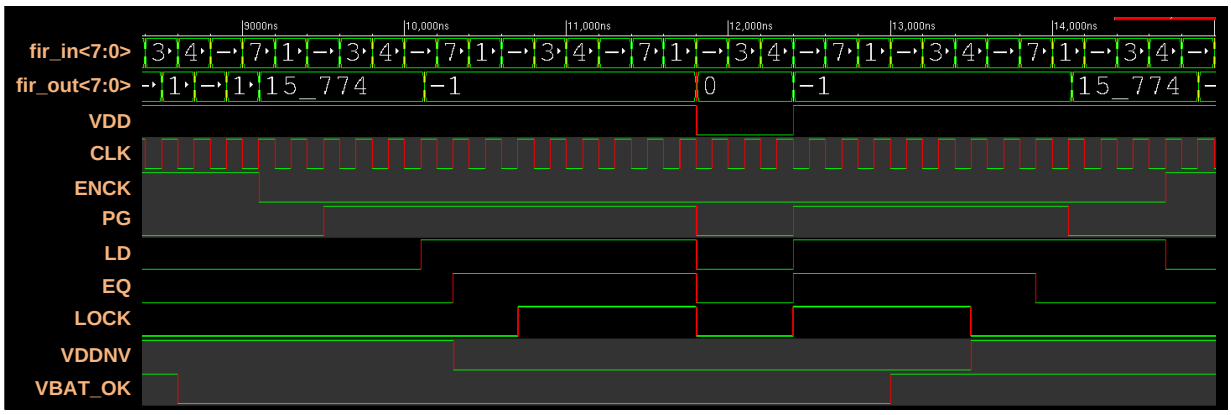
It can be ensured that there will always be enough energy to perform the save operation after VBAT_OK goes low by setting the appropriate hysteresis thresholds on the bq25504 energy harvester chip [84]. In a similar manner, it can be ensured that the chip is powered up and the PMU is in the known reset state with the clock gated before VBAT_OK goes high.

5.4 Concluding Remarks on Non-Volatile Processing

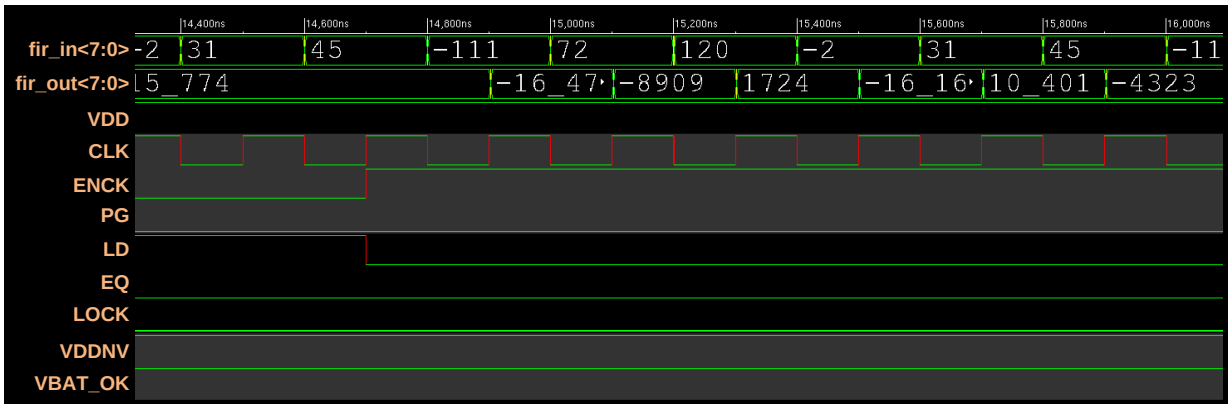
This chapter presented tools to take an arbitrary description of a positive-edge synchronous RTL design and produce a physical implementation that exhibits non-volatile operation. After carefully constructing an implementation of a non-volatile version of the standard cell library DFF, a design flow was developed to integrate the non-volatile DFF. By placing this NVDFF in the context of a programmable FIR filter, data can be gathered to validate the integration feasibility of the NVDFF into the harsh digital environmental aspects of signal coupling and random placement and routing of neighbor cells. Application feasibility will be demonstrated by successful operation of the PMU to automatically save and restore fecap state when appropriate.



(a) Before



(b) During



(c) After

Figure 5.12: Post-layout full-chip transistor-level simulation of power-interruption event (waveforms are viewed digitally)

Chapter 6

Conclusion

In this thesis, four specific areas of integrating embedded memory into low-power integrated circuits were explored. The contributions and conclusions are summarized as follows:

For statistical SRAM simulation:

Contributions:

- A numerical justification of the Loop Flattening simplifications
- A proof under additive Gaussian delays of Loop Flattening
- An illustration of the two-stage Spherical Importance sampling algorithm
- A comparison of Spherical Importance sampling with state-of-the-art statistical simulation techniques

Conclusions:

- The apparent structure of multiplexed paths differs under the statistical analysis of rare occurrences.
- Importance Sampling can be efficiently applied to the simulation of integrated circuits with non-Gaussian statistical sampling and a sequential focus on direction followed by magnitude to explore the space of variation parameters.

For low-voltage SRAM:

Contributions:

- The schematic and layout implementation of a single-ended SRAM read path based

on the AC coupled sense amplifier and regenerative global bitline scheme

- An algorithm to predict the the influence of transistor mismatch on the data retention voltage of an SRAM array

Conclusions:

- AC-coupling is viable on a finer scale in advanced CMOS technology.
- Variation-tolerant sensing networks permit operation closer to the fundamental voltage scaling limits of the bitcell.
- Statistical fluctuation can be predicted on-chip by recovering the functional relation between variation parameter and performance metric.

For Low-energy Non-volatile Memory:

Contributions:

- The schematic and operating details of a time-to-digital sensing scheme with a physical implementation example
- An analysis of fecap signal versus load conditions as the power supply V_{DD} is scaled down
- An architecture for multiplexing several FRAM columns to a single, large sensing network, facilitated by a hierarchical plateline scheme

Conclusions:

- Analog offset can be efficiently (in terms of area) compensated with digital circuits.
- Digital, time-based sensing schemes scale in voltage while consuming less static power and improving with technology scaling.
- Representing signal as a delay better serves the essential multiplexing operation of memory.

For Non-volatile Processing:

Contributions:

- The design of a DFF with ferroelectric capacitors that senses the state in a self-timed fashion and has glitch free transitions between its four modes: restore, active, save, and off.

- An energy harvester interface based on a PMU that manipulates fecap data and always resets to the correct state after power-on such that the fecaps are protected from corruption, also provided is a strategy to choose appropriate settings for turn-on and power-off conditions
- A design flow that puts the NVDFF into a digital circuit that is appropriately managed by the PMU

Conclusions:

- Sensing ferroelectric capacitors with a self-timed latch permits the creation of a non-volatile register that can be successfully integrated into a general-purpose digital design flow.
- An appropriately designed power management unit can interface with energy harvester circuits to automatically manage state in digital circuits that have embedded NVDFFs. This mechanism remains hidden from the designer of the processing circuits.

The work conducted in these four broad areas is motivated by low-power portable electronics for embedded systems. As discussed in the introduction of this thesis, the realm of human experience is surrounded by objects made intelligent by the embedding of integrated circuits that sense, process, communicate, and interact with people. To bring this trend to very extreme cases—inside bodies, or interspersed in office buildings by the thousands, practical low-power solutions for programmable systems are needed. Shown in Fig. 6.1 is the die photo of a microcontroller, which is the heart of any embedded system. Memory plays a key role throughout the chip. Low-access energy low-voltage SRAM can serve for the data memory and cache. Low-energy non-volatile memory can provide instruction memory and information storage. Even in the standard cell area of the core, the registers that store the intermediate results of computation and configuration information dictate how the entire system can be power-gated. In addition to embedded memory, the continual improvement of all aspects of embedded systems, extending to energy efficient processing [62], low-power communication, power management circuits [63], and silicon-technology advancement continues to open up new applications for embedded systems.

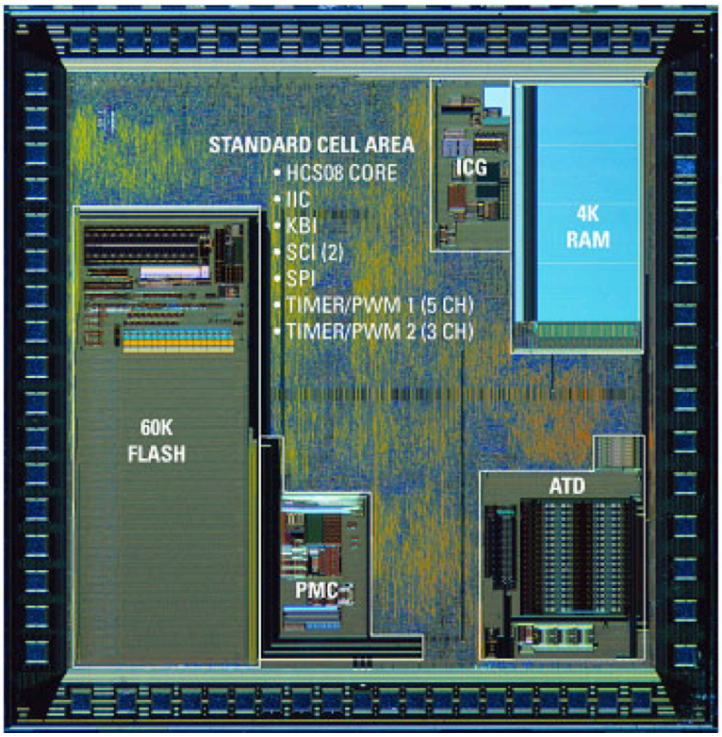


Figure 6.1: 8 bit Microcontroller die photo from [4]

6.1 Future Work

A natural extension of the work presented in this thesis is the integration of non-volatile memory into the entire memory hierarchy of a microcontroller—registers, data memory, and instruction memory. Such a device can serve as a central processing unit of a sensor embedded into an environment such as an office building. Such a device would eliminate the costly requirement that the sensor battery be regularly replaced because the system can power itself from unreliable sources such as ambient light or vibration energy. Given a system that has no distinction between "on" and "off" and never needs to reboot, embedded system designers can explore how to enrich the capability of distributed sensors to record, process,

and exchange information among a network of thousands or even millions of other nodes that collectively function as a storage medium. These questions have not been sufficiently explored because such a non-volatile processor has never existed.

As another direction, storage devices themselves are evolving into systems-on-chip with a variety of embedded memories integrated together. The availability of 3D stacking technology and through-silicon-vias will enable a heterogeneous layering of semiconductor technologies to create a complete storage system out of emerging non-volatile memory devices. Such a system will need a processor for data management, circuits for a high speed interface across die, and both dynamic and static memory for caching and buffering to improve the latency and bandwidth of access to data. Because no single memory device, or even memory architecture can satisfy the multiple distinct storage and data processing needs of low-power embedded systems, integrating the various types of memory discussed in this thesis will remain a central aspect of the design process.

Bibliography

[1] N. A. Kurd, S. Bhamidipati, C. Mozak, J. L. Miller, T. M. Wilson, M. Nemani, and M. Chowdhury, "Westmere: A family of 32nm IA processors," in *IEEE International Solid-State Circuits Conference*, 2010, pp. 96–97.

[2] K. Katayama, S. Hagiwara, H. Tsutsui, H. Ochi, and T. Sato, "Sequential importance sampling for low-probability and high-dimensional sram yield analysis," in *IEEE/ACM International Conference on Computer-Aided Design*, 2010, pp. 703–708.

[3] J. Eliason, "Ferroelectric non-volatile logic elements (patent)," Nov. 18, 2003 2003.

[4] J. Shandle, "More for less: Stable future for 8-bit microcontrollers," 5 August 2004 2004. [Online]. Available: http://www.eetimes.com/electronics-news/4196930/More-for-Less-Stable-Future-for-8-bit-Microcontrollers

[5] L. Dolecek, M. Qazi, D. Shah, and A. Chandrakasan, "Breaking the simulation barrier: SRAM evaluation through norm minimization," in *IEEE/ACM International Conference on Computer-Aided Design*, 2008, pp. 322–329.

[6] H. Kadota, J. Miyake, I. Okabayashi, T. Maeda, T. Okamoto, M. Nakajima, and K. Kagawa, "A 32-bit cmos microprocessor with on-chip cache and tlb," *IEEE Journal of Solid-State Circuits*, vol. 22, no. 5, pp. 800–807, 1987.

[7] Medtronic, *Medtronic Reveal Plus 9526 Insertable Loop Recorder System*, 2005, instruction manual.

[8] J. Kwong and A. P. Chandrakasan, "An energy-efficient biomedical signal processing platform," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 7, pp. 1742–1753, 2011.

[9] IBM, "Smarter buildings survey," Tech. Rep., 2010.

[10] M. Qazi, M. Tikekar, L. Dolecek, D. Shah, and A. P. Chandrakasan, "Loop Flattening & Spherical Sampling: Highly Efficient Model Reduction Techniques for SRAM Yield Analysis," in *Design Automation & Test in Europe*, March 2010, pp. 801–806.

[11] M. Qazi, K. Stawiasz, L. Chang, and A. P. Chandrakasan, "A 512kb 8T SRAM Macro Operating Down to 0.57V with an AC-Coupled Sense Amplifier and Embedded Data-Retention-Voltage Sensor in 45nm SOI CMOS," in *IEEE International Solid-State Circuits Conference*, 2010.

[12] M. Qazi, M. Clinton, S. Bartling, and A. P. Chandrakasan, "A low-voltage 1 mb fram in 0.13 m cmos featuring time-to-digital sensing for expanded operating margin," *JSSC*, vol. 47, no. 1, pp. 141–150, 2012.

[13] J. Pille, C. Adams, T. Christensen, S. R. Cottier, S. Ehrenreich, F. Kono, D. Nelson, O. Takahashi, S. Tokito, O. Torreiter, O. Wagner, and D. Wendel, "Implementation of the Cell Broadband Engine in 65 nm SOI Technology Featuring Dual Power Supply SRAM Arrays Supporting 6 GHz at 1.3 V," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 1, pp. 163–171, 2008.

[14] R. Kanj, R. Joshi, and S. Nassif, "Mixture importance sampling and its application to the analysis of SRAM designs in the presence of rare failure events," in *Design Automation Conference*, 2006, pp. 69–72.

[15] A. Singhee and R. A. Rutenbar, "Statistical blockade: Very fast statistical simulation and modeling of rare circuit events and its application to memory design," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 28, no. 8, pp. 1176–1189, 2009.

[16] J. Jaffari and M. Anis, "Adaptive sampling for efficient failure probability analysis of sram cells," in *IEEE/ACM International Conference on Computer-Aided Design*, 2009, pp. 623–630.

[17] M. H. Abu-Rahma, K. Chowdhury, J. Wang, Z. Chen, S. S. Yoon, and M. Anis, "A methodology for statistical estimation of read access yield in srams," in *ACM/IEEE Design Automation Conference*, 2008, pp. 205–210.

[18] R. Aitken and S. Idgunji, "Worst-case design and margin for embedded sram," in *DATE*, 2007.

[19] P. Zuber, P. Dobrovolny, and M. Miranda, "A holistic approach for statistical sram analysis," in *ACM/IEEE Design Automation Conference*, 2010, pp. 717–722.

[20] P. Zuber, M. Miranda, P. Dobrovolny, K. van der Zanden, and J.-H. Jung, "Statistical sram analysis for yield enhancement," in *Design, Automation & Test in Europe Conference & Exhibition*, 2010, pp. 57–62.

[21] D. Angluin and L. G. Valiant, "Fast probabilistic algorithms for hamiltonian circuits and matchings," in *Proceedings of the ninth annual ACM symposium on Theory of computing*, 1977, pp. 30–41.

[22] A. Agresti and B. A. Coull, "Approximate is better than "exact" for interval estimation of binomial proportions," *The American Statistician*, vol. 52, no. 2, pp. 119–126, 1998 1998.

[23] J. Bucklew, *Introduction to Rare Event Simulation*. Springer, 2004.

[24] J. Wang, S. Yaldiz, X. Li, and L. T. Pileggi, "Sram parametric failure analysis," in *ACM/IEEE Design Automation Conference*, 2009, pp. 496–501.

[25] C. Dong and X. Li, "Efficient sram failure rate prediction via gibbs sampling," in *ACM/EDAC/IEEE Design Automation Conference*, 2011, pp. 200–205.

[26] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. Springer, 1998.

[27] K. J. Antreich and H. E. Graeb, "Circuit optimization driven by worst-case distances," in *IEEE/ACM International Conference on Computer-Aided Design*, 1991, pp. 166–169.

[28] X. Du and W. Chen, "Towards a better understanding of modeling feasibility robustness in engineering design," *ASME Journal of Mechanical Design*, 2000.

[29] M. Tichy, *Applied Methods of Structural Reliability*. Boston: Kluwer Academic Publishers, 1993.

[30] L. Dolecek, M. Qazi, D. Shah, and A. Chandrakasan, "Breaking the simulation barrier: SRAM evaluation through norm minimization," in *ICCAD*, 2008, pp. 322–329.

[31] M. H. Abu-Rahma *et al.*, "A methodology for statistical estimation of read access yield in srams," in *DAC*, 2008, pp. 205–210.

[32] J. E. Stine *et al.*, "Freepdk: An open-source variation-aware design kit," in *ICMSE*, 2007.

[33] W. Zhao and Y. Cao, "New generation of predictive technology model for sub-45 nm early design exploration," *IEEE Trans. on Electron Devices*, vol. 53, no. 11, pp. 2816–2823, 2006.

[34] K. J. Kuhn, "Reducing Variation in Advanced Logic Technologies: Approaches to Process and Design for Manufacturability of Nanoscale CMOS," in *IEEE International Electron Devices Meeting*, 2007, pp. 471–474.

[35] A. Singhee and R. A. Rutenbar, "Statistical Blockade: A Novel Method for Very Fast Monte Carlo Simulation of Rare Circuit Events, and its Application," in *DATE*, 2007.

[36] D. J. Frank, W. Haensch, G. Shahidi, and H. Dokumaci, "Optimizing CMOS technology for maximum performance," *IBM Journal of Research and Development*, vol. 50, no. 4/5, p. 419, July/September 2006.

[37] R. K. Krishnamurthy, A. Alvandpour, S. Mathew, M. Anders, V. De, and S. Borkar, "Highperformance, lowpower, and leakagetolerance challenges for sub70nm microprocessor circuits," in *European Solid State Circuits Conference*, 2002, pp. 315–321.

[38] L. Chang, R. K. Montoye, Y. Nakamura, K. A. Batson, R. J. Eickemeyer, R. H. Dennard, W. Haensch, and D. Jamsek, "An 8T-SRAM for Variability Tolerance and Low-Voltage Operation in High-Performance Caches," *IEEE Journal of Solid State Circuits*, vol. 43, no. 4, pp. 956–963, 2008.

[39] A. Alvandpour, R. K. Krishnamurthy, K. Soumyanath, and S. Y. Borkar, "A sub-130-nm conditional keeper technique," *IEEE Journal of Solid-State Circuits*, vol. 37, no. 5, pp. 633–638, 2002.

[40] I. Arsovski and R. Wistort, "Self-referenced sense amplifier for across-chip-variation immune sensing in high-performance Content-Addressable Memories," in *IEEE Custom Integrated Circuits Conference*, 2006, pp. 453–456.

[41] N. Verma and A. P. Chandrakasan, "A High-Density 45nm SRAM Using Small-Signal Non-Strobed Regenerative Sensing," in *IEEE International Solid-State Circuits Conference*, 2008, pp. 380–621.

[42] A. Singh, M. Ciraula, D. Weiss, J. Wuu, P. Bauser, P. de Champs, H. Daghighian, D. Fisch, P. Graber, and M. Bron, "A 2ns-read-latency 4Mb embedded floating-body memory macro in 45nm SOI technology," in *IEEE International Solid State Circuits Conference*, 2009, pp. 460–461.

[43] K. Zhang, U. Bhattacharya, L. Ma, Y. Ng, B. Zheng, M. Bohr, and S. Thompson, "A fully synchronized, pipelined, and re-configurable 50 Mb SRAM on 90 nm CMOS technology for logic applications," in *Symposium on VLSI Circuits*, 2003, pp. 253–254.

[44] K. Zhang, K. Hose, V. De, and B. Senyk, "The scaling of data sensing schemes for high speed cache design in sub-0.18μM technologies," in *IEEE Symposium on VLSI Circuits*, 2000, pp. 226–227.

[45] H. Pilo, V. Ramadurai, G. Braceras, J. Gabric, S. Lamphier, and Y. Tan, "A 450ps Access-Time SRAM Macro in 45nm SOI Featuring a Two-Stage Sensing-Scheme and Dynamic Power Management," in *IEEE International Solid State Circuits Conference*, 2008, pp. 378–621.

[46] H. Qin, Y. Cao, D. Markovic, A. Vladimirescu, and J. Rabaey, "SRAM leakage suppression by minimizing standby supply voltage," in *International Symposium on Quality Electronic Design*, 2004, pp. 55–60.

[47] K. Nii, Y. Tenoh, T. Yoshizawa, S. Imaoka, Y. Tsukamoto, Y. Yamagami, T. Suzuki, A. Shibayama, H. Makino, and S. Iwade, "A 90nm low power 32 K-byte embedded SRAM with gate leakage suppression circuit for mobile applications," in *IEEE Symposium on VLSI Circuits*, 2003, pp. 247–250.

[48] M. Yamaoka, Y. Shinozaki, N. Maeda, Y. Shimazaki, K. Kato, S. Shimada, K. Yanagisawa, and K. Osadal, "A 300MHz 25uA/Mb leakage on-chip SRAM module featuring process-variation immunity and low-leakage-active mode for mobile-phone application processor," in *IEEE International Solid State Circuits Conference*, 2004, pp. 494–542 Vol.1.

[49] K. Zhang, U. Bhattacharya, Z. Chen, F. Hamzaoglu, D. Murray, N. Vallepalli, Y. Wang, B. Zheng, and M. Bohr, "SRAM design on 65nm CMOS technology with integrated leakage reduction scheme," in *IEEE Symposium on VLSI Circuits*, 2004, pp. 294–295.

[50] F. Hamzaoglu, K. Zhang, Y. Wang, H. J. Ann, U. Bhattacharya, Z. Chen, Y. Ng, A. Pavlov, K. Smits, and M. Bohr, "A 153Mb-SRAM Design with Dynamic Stability Enhancement and Leakage Reduction in 45nm High-K Metal-Gate CMOS Technology," in *IEEE International Solid State Circuits Conference*, 2008, pp. 376–621.

[51] Y. Takeyama, H. Otake, O. Hirabayashi, K. Kushida, and N. Otsuka, "A low leakage SRAM macro with replica cell biasing scheme," in *IEEE Symposium on VLSI Circuits*, 2005, pp. 166–167.

[52] J. Wang and B. H. Calhoun, "Canary Replica Feedback for Near-DRV Standby VDD Scaling in a 90nm SRAM," in *IEEE Custom Integrated Circuits Conference*, 2007, pp. 29–32.

[53] X. Deng, W. K. Loh, B. Pious, T. W. Houston, L. Liu, B. Khan, and D. Corum, "Characterization of bit transistors in a functional SRAM," in *IEEE Symposium on VLSI Circuits*, 2008, pp. 44–45.

[54] Y. Wang, U. Bhattacharya, F. Hamzaoglu, P. Kolar, Y. Ng, L. Wei, Y. Zhang, K. Zhang, and M. Bohr, "A 4.0 GHz 291Mb Voltage-Scalable SRAM Design in 32nm High-k Metal-Gate CMOS with Integrated Power Management," in *IEEE International Solid State Circuits Conference*, 2009.

[55] C. Lee, S.-K. Lee, S. Ahn, J. Lee, W. Park, Y. Cho, C. Jang, C. Yang, S. Chung, I.-S. Yun, B. Joo, B. Jeong, J. Kim, J. Kwon, H. Jin, Y. Noh, J. Ha, M. Sung, D. Choi, S. Kim, J. Choi, T. Jeon, H. Park, J.-S. Yang, and Y.-H. Koh, "A 32-Gb MLC NAND Flash Memory With Vth Endurance Enhancing Schemes in 32 nm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 1, p. 97, jan. 2011.

[56] C. Villa, D. Vimercati, S. Schippers, S. Polizzi, A. Scavuzzo, M. Perroni, M. Gaibotti, and M. L. Sali, "A 65 nm 1 Gb 2b/cell NOR Flash With 2.25 MB/s Program Throughput and 400 MB/s DDR Interface," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 1, p. 132, jan. 2008.

[57] K.-J. Lee, B.-H. Cho, W.-Y. Cho, S. Kang, B.-G. Choi, H.-R. Oh, C.-S. Lee, H.-J. Kim, J.-M. Park, Q. Wang, M.-H. Park, Y.-H. Ro, J.-Y. Choi, K.-S. Kim, Y.-R. Kim, I.-C. Shin, K.-W. Lim, H.-K. Cho, C.-H. Choi, W.-R. Chung, D.-E. Kim, Y.-J. Yoon, K.-S. Yu, G.-T. Jeong, H.-S. Jeong, C.-K. Kwak, C.-H. Kim, and K. Kim, "A 90 nm 1.8 V 512 Mb Diode-Switch PRAM With 266 MB/s Read Throughput," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 1, p. 150, jan. 2008.

[58] Y. M. Kang, H. J. Joo, J. H. Park, S. K. Kang, J. H. Kim, S. G. Oh, H. S. Kim, Y. J. Kang, J. Y. Jung, D. Y. Choi, E. S. Lee, S. Y. Lee, H. S. Jeong, and K. Kim, "World Smallest 0.34μm COB Cell 1T1C 64Mb FRAM with New Sensing Architecture and Highly Reliable MOCVD PZT Integration Technology," in *IEEE Symposium on VLSI Technology*, 0-0 2006, p. 124.

[59] T. Sugibayashi, N. Sakimura, T. Honda, K. Nagahara, K. Tsuji, H. Numata, S. Miura, K. Shimura, Y. Kato, S. Saito, Y. Fukumoto, H. Honjo, T. Suzuki, K. Suemitsu, T. Mukai, K. Mori, R. Nebashi, S. Fukami, N. Ohshima, H. Hada, N. Ishiwata, N. Kasai, and S. Tahara, "A 16-Mb Toggle MRAM With Burst Modes," *IEEE Journal of Solid-State Circuits*, vol. 42, no. 11, p. 2378, nov. 2007.

[60] D. Takashima, Y. Nagadomi, and T. Ozaki, "A 100MHz ladder FeRAM design with capacitance-coupled-bitline (CCB) cell," in *IEEE Symposium on VLSI Circuits*, june 2010, p. 227.

[61] S. Henzler, S. Koeppe, D. Lorenz, W. Kamp, R. Kuenemund, and D. Schmitt-Landsiedel, "A Local Passive Time Interpolation Concept for Variation-Tolerant High-Resolution Time-to-Digital Conversion," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 7, pp. 1666–1676, 2008.

[62] J. Kwong, Y. K. Ramadass, N. Verma, and A. P. Chandrakasan, "A 65 nm Sub-Vt Microcontroller With Integrated SRAM and Switched Capacitor DC-DC Converter," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 1, pp. 115–126, 2009.

[63] G. Chen, M. Fojtik, D. Kim, D. Fick, J. Park, M. Seok, M.-T. Chen, Z. Foo, D. Sylvester, and D. Blaauw, "Millimeter-scale nearly perpetual sensor system with stacked battery and solar cells," in *IEEE International Solid-State Circuits Conference*, 2010, pp. 288–289.

[64] M. Zwerg, A. Baumann, R. Kuhn, M. Arnold, R. Nerlich, M. Herzog, R. Ledwa, C. Sichert, V. Rzehak, P. Thanigai, and B. O. Eversmann, "An 82μA/MHz Micro-

controller with Embedded FeRAM for Energy-Harvesting Applications," in *IEEE International Solid-State Circuits Conference*, 2011, pp. 334–335.

[65] T. S. Moise, S. R. Summerfelt, H. McAdams, S. Aggarwal, K. R. Udayakumar, F. G. Celii, J. S. Martin, G. Xing, L. Hall, K. J. Taylor, T. Hurd, J. Rodriguez, K. Remack, M. D. Khan, K. Boku, G. Stacey, M. Yao, M. G. Albrecht, E. Zielinski, M. Thakre, S. Kuchimanchi, A. Thomas, B. McKee, J. Rickes, A. Wang, J. Grace, J. Fong, D. Lee, C. Pietrzyk, R. Lanham, S. R. Gilbert, D. Taylor, J. Amano, R. Bailey, F. Chu, G. Fox, S. Sun, and T. Davenport, "Demonstration of a 4 Mb, high density ferroelectric memory embedded within a 130 nm, 5 LM Cu/FSG logic process," in *IEEE International Electron Devices Meeting*, 2002, pp. 535–538.

[66] D. Takashima, H. Shiga, D. Hashimoto, T. Miyakawa, S. Shiratake, K. Hoya, R. Ogiwara, R. Takizawa, S. Doumae, R. Fukuda, Y. Watanabe, S. Fujii, T. Ozaki, H. Kanaya, S. Shuto, K. Yamakawa, I. Kunishima, T. Hamamoto, and A. Nitayama, "A scalable shield-bitline-overdrive technique for 1.3V Chain FeRAM," in *IEEE International Solid-State Circuits Conference*, 2010, pp. 262–263.

[67] K. Yamaoka, S. Iwanari, Y. Murakuki, H. Hirano, M. Sakagami, T. Nakakuma, T. Miki, and Y. Gohou, "A 0.9 V 1T1C SBT-based embedded non-volatile FeRAM with a reference voltage scheme and multi-layer shielded bit-line structure," in *IEEE International Solid-State Circuits Conference*, 2004, pp. 50–512 Vol.1.

[68] Y. Eslami, A. Sheikholeslami, S. Masui, T. Endo, and S. Kawashima, "A differential-capacitance read scheme for FeRAMs," in *IEEE Symposium on VLSI Circuits*, 2002, p. 298.

[69] S. Kawashima, T. Endo, T. Yamamoto, K. I. Nakabayashi, M. Nakazawa, K. Morita, and M. Aoki, "A bit-line GND sense technique for low-voltage operation FeRAM," in *IEEE Symposium on VLSI Circuits*, 2001, p. 127.

[70] H. P. McAdams, R. Acklin, T. Blake, X.-H. Du, J. Eliason, J. Fong, W. F. Kraus, D. Liu, S. Madan, T. Moise, S. Natarajan, N. Qian, Y. Qiu, K. A. Remack, J. Rodriguez, J. Roscher, A. Seshadri, and S. R. Summerfelt, "A 64-Mb Embedded FRAM Utilizing a 130-nm 5LM Cu/FSG Logic Process," *IEEE Journal of Solid-State Circuits*, vol. 39; 39, no. 4, pp. 667–677, 2004.

[71] I. Stolichnov, A. Tagantsev, N. Setter, J. S. Cross, and M. Tsukada, "Crossover between nucleation-controlled kinetics and domain wall motion kinetics of polarization reversal in ferroelectric films," *Applied Physics Letters*, 2003.

[72] A. G. Acosta, J. Rodriguez, B. Obradovic, S. Summerfelt, T. San, K. Green, T. Moise, and S. Krishnan, "Scaling reliability and modeling of ferroelectric capacitors," in *IEEE International Reliability Physics Symposium*, 2010, pp. 689–693.

[73] S. R. Summerfelt, T. S. Moise, K. R. Udayakumar, K. Boku, K. Remack, J. Rodriguez, J. Gertas, H. McAdams, S. Madan, J. Eliason, J. Groat, D. Kim, P. Staubs, M. Depner, and R. Bailey, "High-Density 8Mb 1T-1C Ferroelectric Random Access Memory Embedded Within a Low-Power 130 nm Logic Process," in *IEEE International Symposium on Applications of Ferroelectrics*, 2007, pp. 9–10.

[74] B.-G. Jeon, M.-K. Choi, Y. Song, and K. Kim, "A nonvolatile ferroelectric RAM with common plate folded bit-line cell and enhanced data sensing scheme," in *IEEE International Solid State Circuits Conference*, 2001, pp. 38–39, 426.

[75] D. Porat, "Review of sub-nanosecond time-interval measurements," *IEEE Transactions on Nuclear Science*, vol. 20, no. 5, p. 36, October 1973 1973.

[76] H. Tanaka, Y. Nakagome, J. Etoh, E. Yamasaki, M. Aoki, and K. Miyazawa, "Sub-1-μA dynamic reference voltage generator for battery-operated DRAMs," *IEEE Journal of Solid-State Circuits*, vol. 29, no. 4, pp. 448–453, 1994.

[77] K. N. Leung and P. K. T. Mok, "A CMOS voltage reference based on weighted ΔV_{GS} for CMOS low-dropout linear regulators," *IEEE Journal of Solid-State Circuits*, vol. 38, no. 1, pp. 146–150, 2003.

[78] G. DeVita and G. Iannaccone, "A Sub-1-V, 10 ppm / ° C, Nanopower Voltage Reference Generator," *IEEE Journal of Solid-State Circuits*, vol. 42, no. 7, pp. 1536–1542, 2007.

[79] T. Kawahara, "Scalable spin-transfer torque ram technology for normally-off computing," *IEEE Design & Test of Computer*, vol. 28, no. 1, pp. 52–63, 2011.

[80] S. Shigematsu, S. Mutoh, Y. Matsuya, Y. Tanabe, and J. Yamada, "A 1-v high-speed mtcmos circuit scheme for power-down application circuits," *IEEE Journal of Solid-State Circuits*, vol. 32, no. 6, pp. 861–869, 1997.

[81] N. Sakimura, T. Sugibayashi, R. Nebashi, and N. Kasai, "Nonvolatile magnetic flip-flop for standby-power-free socs," *IEEE Journal of Solid State Circuits*, vol. 44, no. 8, pp. 2244–2250, 2009.

[82] A. P. Chandrakasan, A. Burstein, and R. W. Brodersen, "A low-power chipset for a portable multimedia i/o terminal," *IEEE Journal of Solid-State Circuits*, vol. 29, no. 12, pp. 1415–1428, 1994.

[83] A. Gruverman, B. Rodriguez, C. Dehoff, J. Waldrep, A. Kingon, R. Nemanich, and J. Cross, "Direct studies of domain switching dynamics in thin film ferroelectric capacitors," *Applied Physics Letters*, vol. 87, no. 8, p. 082902, AUG 22 2005, pT: J; NR: 25; TC: 84; J9: APPL PHYS LETT; PG: 3; GA: 956OW; UT: WOS:000231310700039.

[84] K. Kadirvel, Y. Ramadass, U. Lyles, J. Carpenter, A. Chandrakasan, and B. Lum-Shue-Chan, "330 nA Energy-Harvesting Charger with Battery Management for Solar and Thermoelectric Energy Harvesting," in *IEEE International Solid-State Circuits Conference*, 2012, pp. 106–107.

[85] Z. Hu, A. Buyuktosunoglu, V. Srinivasan, V. Zyuban, H. Jacobson, and P. Bose, "Microarchitectural techniques for power gating of execution units," in *International Symposium on Low Power Electronics and Design*, 2004, pp. 32–37.