Reliability In The Face of Variability in Nanometer Embedded Memories

Shrikanth Ganapathy

Department d'Arquitectura de Computadors

Universitat Politècnica de Catalunya

A THESIS SUBMITTED IN FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

March, 2014

Reliability In The Face of Variability in Nanometer Embedded Memories

Shrikanth Ganapathy

Advisors:

Ramon Canal Corretger

Universitat Politećnica de Catalunya

Antonio González Colás

Universitat Politećnica de Catalunya & Intel Barcelona Research Center

Antonio Rubio Solá

Universitat Politećnica de Catalunya

Abstract

Constant feature miniaturization has been a principal component in the advancement of CMOS technology. As we continue to scale device dimensions further, the advantages are slowly diminishing as stringent constraints are being set on processor power and performance. In light of this, device characteristics such as supply and threshold voltage are scaled aggressively to achieve significant power reduction across successive process generations. However, as we continue to scale, manufacturing processes are becoming less deterministic inducing variations severely inhibited by process control limitations. Such variations have a strong negative impact on the operational margin of the processor affecting key metrics across the lifetime of the chip. This is worsened by lowering of reliability margins where regular operation of the system is characterised by frequent failures and errors. Therefore, it has become a serious challenge to design for high-performance and low-power in the presence of parametric variability.

In this thesis, we have investigated the impact of parametric variations on the behaviour of one performance-critical processor structure - embedded memories. As variations manifest as a spread in power and performance, as a first step, we propose a novel modeling methodology that helps evaluate the impact of circuit-level optimizations on architecture-level design choices. Choices made at the design-stage ensure conflicting requirements from higher-level are decoupled. We then complement such design-time optimizations with a run-time mechanism that takes advantage of adaptive body-biasing to lower power whilst improving performance in the presence of variability. Our proposal uses a novel fully-digital variation tracking hardware using embedded DRAM (eDRAM) cells to monitor run-time changes in cache latency and leakage. A special fine-grain body-bias generator uses the measurements to generate an optimal body-bias that is needed to meet the required yield targets. A novel variation-tolerant and soft-error hardened eDRAM cell is also proposed as

an alternate candidate for replacing existing SRAM-based designs in latency critical memory structures. In the ultra low-power domain where reliable operation is limited by the minimum voltage of operation ($V_{dd_{min}}$), we analyse the impact of failures on cache functional margin and functional yield. Towards this end, we have developed a fully automated tool (INFORMER) capable of estimating memory-wide metrics such as power, performance and yield accurately and rapidly. Using the developed tool, we then evaluate the effectiveness of a new class of hybrid techniques in improving cache yield through failure prevention and correction. Having a holistic perspective of memory-wide metrics helps us arrive at design-choices optimized simultaneously for multiple metrics needed for maintaining lifetime requirements.

Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank my advisors Prof.Ramon Canal, Prof.Antonio González and Prof.Antonio Rubio for providing me this wonderful opportunity to do a PhD under their tutelage. Ramon over these last six years has been a great mentor and an even better friend who has helped me successfully navigate through the rough seas of a PhD. Ramon has always insisted on keeping things simple and focused. I hope someday I can emulate the great qualities as an advisor he has to offer. Antonio González has been very supportive of my research endeavors and provided me the much needed freedom to work on a problem of my choice. One of the most valuable lessons Antonio taught me was that communicating research ideas in a lucid manner is of paramount importance to a good researcher. Antonio Rubio has been one of the most affable teachers I have studied under. I could approach him for scientific advise at all times and he would readily spend hours together working on the problem. For all the motivation, encouragement and enlightenment, I cannot thank my advisors enough. I would also like to thank Jaume Abella (Barcelona Super Computing Center), Serkan Ozdemir (Intel Barcelona Research Center) and Francesc Moll (UPC) for their valuable comments and suggestions through the course of preparing this thesis.

My time here at UPC was made all the more pleasant in part due to the many friends and colleagues who in one way or the other have contributed to the betterment of my life here. I would like to thank all the students who have worked at D6-113 over the course of last six years. I particularly enjoyed the numerous coffee breaks where we indulged in conversations ranging from fifth century Indian ideologies to twenty first century internet of things. I would like to thank all members of the ARCO group for instilling in me the importance of teamwork and providing important feedback after the group's seminar sessions. I would like to thank Dr.Dan Alexandrescu and Enrico Costenaro from iRoC technologies for having me as an intern in their organization. Dan

and Enrico have been very positive about my research and were instrumental in helping me receive the 2012 Intel Doctoral Student Honor Programme fellowship. They were also kind enough to let me access their servers and other resources for a sufficient period after my internship. I would also like to thank all personnel from LCAC and ADMAC who have offloaded from me all issues pertaining to non-technical aspects of my PhD.

Outside life at UPC, I've had the privilege of making friends with a number of people who have made my life all the more enjoyable. I thank Vinoth, Karthi, Aswin, Karishma, Rahul, Adi, Vinay, Martha, Sudhanshu, Prithvi, Rammy, CK and Maddy who through the years have managed to put up with all my tantrums and still enjoy my company. Back in India, Venkat, Varun, Yogi, Sats, Vrusha and Sujatha who I now know for more than a decade have always supported me in all my endeavors and have made my vacations back home extremely enjoyable. Thank you!

I would like to thank my undergraduate mentor Prof.Venkateswaran for having faith in me and providing me an opportunity to study at WARFT. Not only a great mentor, he has been a cornerstone in my professional development.

I would not have dared to begin this journey if not for my family. I am forever indebted to my parents for showering me with all the unconditional love and affection and quite literally standing by my side in times of crisis. My father has always valued the importance of hard work and perseverance while my mother often reminds me of how important it is to think and stay positive. Today, I think their combined efforts have brought to fruition a better researcher in me. I know I can always count on my brother and sister-in-law during rough times and I am grateful to them for that. My adorable niece has been a bundle of joy always re-energizing me after a hard day's work. Thank you amma, appa, anna & manni!

Finally, I would like express my gratitude to the almighty for providing me the strength and blessings to complete this journey.

Dedicated to
Akriti, Anna, Manni, Amma & Appa

Intentionally left blank

1

Introduction

In the last 40 years, computer architects have obediently followed Moore's law by designing microprocessors that scale performance at a rate equal to the growth of number of devices on a chip. This phenomena, made possible by the tremendous advancements in manufacturing technology, has today enabled billion transistor integration, paving the way for even faster processors with lower power consumption. Technology scaling has thus provided a stable platform for constant innovation that allows us to design complex processors requiring integration of increased amount of functionality. We, as consumers of such innovations are already enjoying the benefits as evinced by our heavy reliance on one or more number of modern day electronic gadgets.

To be able to sustain this pace of innovation, it is imperative that Moore's law needs to continue being a major driver for semiconductor companies. Moore's law mainly deals with scaling of resources quantitatively and does not delve into the attributes of changing device characteristics. In the present day context, this observation is of extreme importance as we scale to deep sub-micron technologies where semiconductor manufacturing is the biggest bottleneck that is mitigating the advantages of continuous device scaling. This is because, the ability to exert complete control and maintain uniformity over the

manufacturing process is reducing with every subsequent technology node. As a result, imperfect fabrication conditions introduce variations in process parameters that make the chip behave differently from their design specifications. These variations in parameter values between intended design and obtained chip manifest as a reduction in yield, increase in idle power and/or considerable reduction in performance. Overcoming such pitfalls by adopting a variation-aware design paradigm requires developing holistic design principles that involve leveraging information obtained from pre-silicon characterization through extensive design-space exploration and employing it to develop run-time response techniques using on-chip mechanisms for variability compensation.

1.1 Impact of Spatio - Temporal Variations

1.1.1 Power/Performance Variability

Variation in key transistor parameters like threshold voltage, channel length and width manifest as a wide spread in circuit characteristics such as delay and leakage. While certain chips run faster than the desired specifications, there are others that fail to meet the expected operating frequencies. The slower chips have to be discarded resulting in a yield loss or rely on *post-manufacturing binning* to be sold at a lower cost. It was shown in [16], the highest impact on maximum operating frequency F_{max} is due to within-die systematic variations. Assuming a 3σ variation of 20% in channel length in 50nm technology node, a generation of performance gain can be lost. However, further results have also indicated that the spread in F_{max} tends to average out with increasing logic depth. Dynamic power (switching power) is quadratically proportional to the supply voltage and supply voltage scaling is an effective method for power savings. However, with technology scaling, the rate of voltage scaling has diminished due to the stringent requirements set on performance (directly proportional to voltage). Designers have rather relied on aggressive generational (process generation) reduction in threshold voltage for improving performance with every subsequent technology node. One of the main challenges of such an approach is coping with static power (leakage) which increases exponentially with reducing threshold voltage. In the era of dark silicon where it is understood that large amounts of on-chip components will be under-utilised and remain in their idle-state, leakage power can become quite significant in the context of overall system power. As evinced by previous researchers in [32], either under- or over-designing processors will

never suffice as it can result in reduced performance or rejection of good-dies in lieu of not meeting power budgets. In such cases, it is important to develop accurate models that can capture the non-trivial impact of variations on parameters such as delay and power. This can ensure designers prevent either under- or over-estimation of system parameters and help them make important design decisions in the early-stages of the design life-cycle.

1.1.2 Soft Failures and Errors

Dynamic voltage scaling (DVS) techniques can help the system toggle between high-performance and low-power modes by frequently modulating the supply voltage. Lower supply voltage modes resulting in lower power consumption are generally exercised during periods of minimum activity or when parts of the processor are in the standby mode. As the voltage is lowered further, the analog distance between logic '0' and logic '1' is reduced tremendously. In other words, the margin differentiating a logic '0' and logic '1' is very minimum. This margin can be very easily disturbed by providing a small amount of noise. Process variations and external factors are very much capable of inducing such a noise during regular operation of the processor. The stability is disturbed causing intermittent failures (or errors) killing the functionality of the underlying logic. Such disturbances result in unintentional flipping of logic values and frequent flipping can stall the system in certain scenarios. Voltage-noise dependent stability concerns can result in two types of reliability problems

• *Parametric Failures*: Random variations in device parameters can lower the noise margin of the transistor by making them weak (high threshold voltage) reducing the overall drive current. The amount of current needed for a transition is dependent on threshold voltage value and for transistors with -$\sigma_{V_{th}}$, increase in current flow can upset the state of the transistor causing a failure. These types of intermittent failures are known as *parametric failures* that are largely dependent on parameter values affecting key electrical characteristics of the transistor.

• *Soft Errors*: Radio-active particles generated either by cosmic rays or impurities (secondary particles) in the package can interact with the substrate and generate electron-hole pairs. This process creates sufficient energy (noise) that can disturb the stability of the component leading to an error. Such errors are intermittent and dependent heavily on operating environment. The term *soft* indicates they are transient in nature as opposed to hard faults.

1.1.3 Lifetime Reliability

The previous two challenges highlighted concerns mostly arising from imperfect fabrication conditions resulting in static variations of process parameters. It is assumed that any system without mechanical moving parts will never degrade temporally and can guarantee consistent functionality across its lifetime. However, under heavy workloads and continuous stress, the underlying silicon material is subjected to physical changes that can alter the fundamental behaviour of transistor. This stress can be attributed to con-

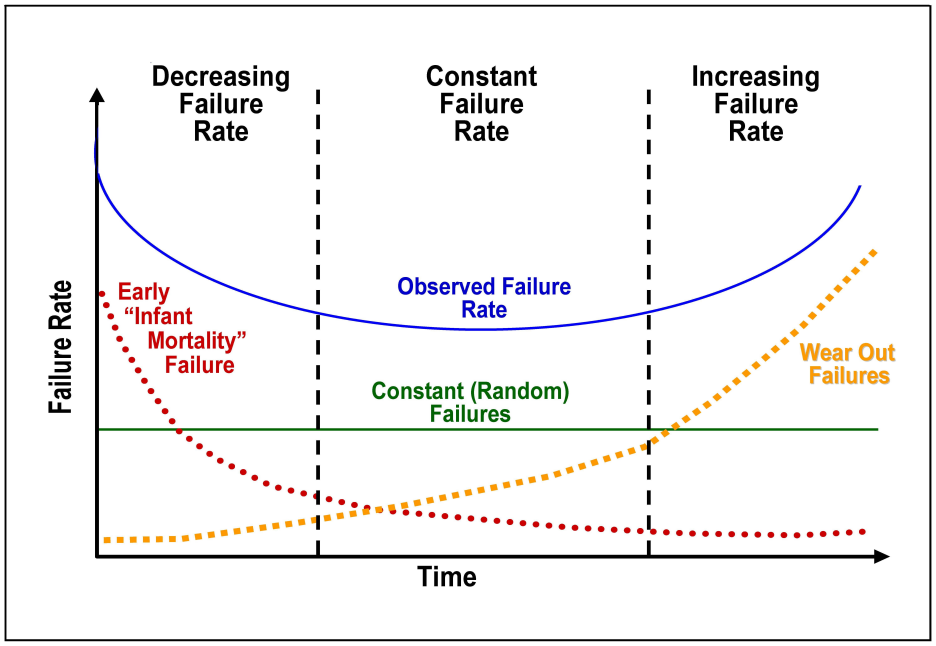


Figure 1.1: Bathtub Curve: Failure rate across operational lifetime

stant scaling of device dimensions that has resulted in reduction of gate oxide thickness leading to increased electric field densities across the device. *Negative bias temperature instability* (NBTI) is one such reliability mechanism that manifests as an increase in the threshold voltage of PMOS transistors across the lifetime. Such temporal variations in the threshold voltage translate directly to a spread in power and performance temporally. Increase in delay (reduced performance) can cause timing violations exacerbating NBTI-dependent parametric failures. Figure 1.1 shows the failure rate of any product across its lifetime. The trend in failure rate can be figuratively described by the bathtub curve. During the initial period of operation, also called *'infant mortality period'*, the rate of failure is high and reduces gradually with time. The high failure rate is due do

failures arising from imperfect manufacturing, defects in the material, bugs introduced during design-stages and some other failures that are accelerated by post-manufacturing testing processes such as *burn-in*. With reduction in failure rate, a plateaued behaviour is observed during which the failure rate is constant. Random errors and failures caused by external sources such as cosmic-rays and cross-talk induced noise maintain this constant failure rate. Aggressive degradation and ageing increase the failure rate towards the end of the product life-cycle. Such failures are called *wear-out* failures and drastically reduce product reliability.

1.2 Nanometer Embedded Memory Design

As postulated by the F_{max} theory, structures with the most number of parallel and shallow critical paths are the most affected by parametric variations. Embedded memories such as register files, shared and private caches are example structures that exhibit the above characteristics. Due to lower area overhead compared to the performance they offer, on-chip memories are growing in size with every new process generation. It is strongly expected that in the following generations, more than 90% of silicon real-estate will be covered with embedded memories. This is primarily driven by factors such as need for higher performance, lower power and most importantly higher density (tighter integration).

As memories are a very important component from an area point of view, it thus becomes extremely challenging to optimize chip yield keeping in mind the effects of spatial variations of process parameters and temporal variations of temperature and voltage. Modern day caches rely on high-performance, low-power six transistor static random access memory(6T-SRAM) cells for storing data. As shown in figure 1.2, as we move from one process generation to next, while there is moderate reduction of approximately 30% in feature sizes (gate-pitch), SRAM bitcells are subjected to restrictive design rules that necessitate atleast 50% reduction in area as we move from one technology to another. This enables integration of as much as 152Mbit/cm^2 in state of the art 45nm technology. Given these area constraints, memory cells are typically designed using minimum geometry transistors that are highly susceptible to intrinsic device level variations. This can be attributed to device dimensions (using minimum geometry transistors) where the effects of random dopant fluctuations (RDF) and line-edge roughness (LER) are more pronounced [107]. With aggressive scaling, due to atomic level fluctuations (intrinsic

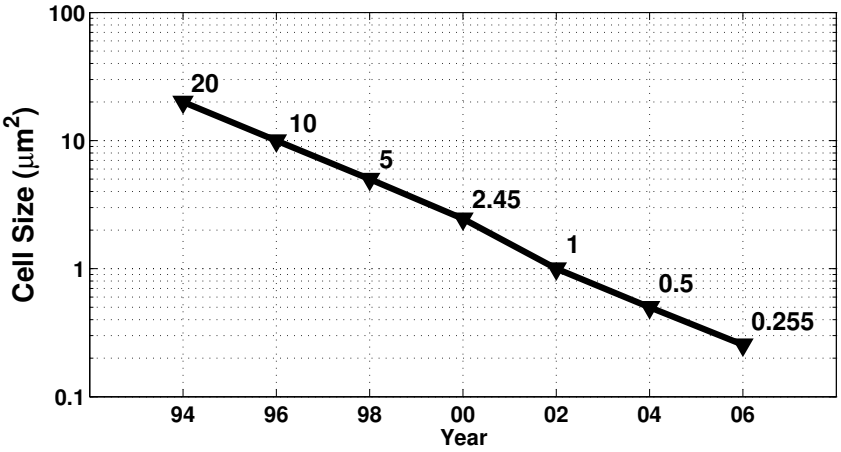


Figure 1.2: SRAM cell area scaling from 350nm to 45nm across 7 process generations [3, 113]

fluctuations), even adjacent SRAM cells do not exhibit uniform characteristics resulting in erratic runtime behaviour of similarly designed structures. In addition to influencing the power and performance characteristics of the memory negatively, variations can affect strongly the functionality of the memory cell making them especially susceptible to large number of failure mechanisms (see previous section).

As shown in figure 1.3, the amount of $V_{cc_{min}}$ (defined as the minimum voltage of operation that guarantees reliable functioning of the system) influenced failures increases by orders of magnitude with every subsequent process generation. Hard failures due to manufacturing defects or mechanisms like NBTI are shown to have reduced with scaling. This can be attributed to the improvements in the underlying manufacturing processes. The failures induced in memories concern chip designers the most as they can influence a large number of design choices at different levels of abstraction [18]. For example, the $V_{cc_{min}}$ of the cache which decides the $V_{cc_{min}}$ of the whole processor is dependent on the highest $V_{cc_{min}}$ among all cells in all arrays. This would mean that, under the effects of random variations where failures are distributed, a single SRAM cell could potentially affect the functional yield of the whole processor if no repair mechanisms are in place. Reduction in functional yield translates directly to a reduction in total cache addressable space. Such a phenomenon can have a serious detrimental effect on system-wide metrics such IPC and performance/watt.

Moving away from SRAM technology in light of such scalability challenges, semiconductor companies have recently started embracing embedded-DRAM (eDRAM) tech-

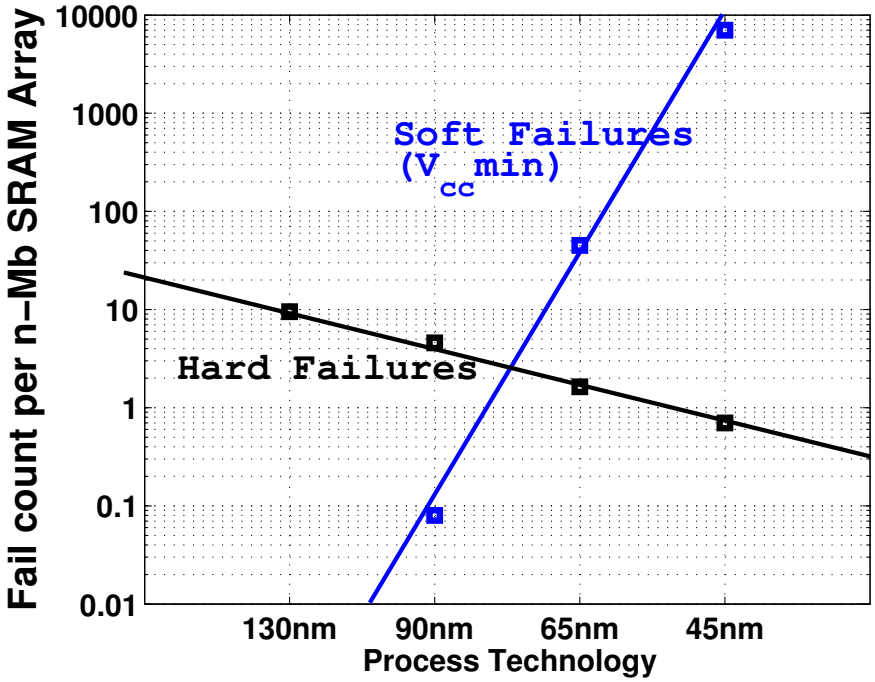


Figure 1.3: Hard and soft failure trends with SRAM scaling [86].

nology for on-chip memories that has recently led to their adoption in many commercial products. 1T1C eDRAM cells are the most widely used but because of their high latency and reads being destructive, they cannot be used in latency-sensitive components like L1 caches. On the contrary, basic three-transistor (3T)-eDRAM and its derivatives in addition to offering access speeds on par with regular SRAM, provide non-destructive reads and are capable of driving large-loaded bitlines suitable for operation in low-voltage caches. One of the major advantages of replacing regular SRAMs with eDRAM is the ability to seamless integrate eDRAM designs by virtue of their logic compatibility. As such, no additional process steps are required and eDRAM cells can be easily realised using standard-cell libraries. However, a major cause of concern with all DRAM technology is the data retention capability which diminishes over time. This particular characteristic makes eDRAM technology extremely unsuitable for L1 caches where the data needs to be held for a sufficiently large period without being destroyed. Also, unlike regular capacitor-based DRAMS, a refresh operation (write-back) in eDRAM-based memories is quite expensive from a performance and power perspective. For high-activity factor component like L1 caches, frequent refresh operations in addition to increasing the dynamic power overhead, can reduce the total available bandwidth as the read ports are blocked

during a write-data-back (refresh) operation. Further, as retention time can be in the order of seconds, for a pulse generated by a particle strike (cosmic rays), the window where it can manifest as a bit-flip is very wide, making the cell highly susceptible to soft-errors during hold mode. Therefore, by enhancing the retention time, not only can the cell guarantee fast reads for a larger number of accesses, the time period during which the total charge held by the storage node is much higher than the critical charge (Q_{crit}) improving temporal soft-error tolerance.

1.3 Main Contributions

This thesis focusses on addressing the problem of parametric variability in embedded SRAM- and eDRAM-based memories. The techniques proposed to combat variations include a combination of design-specific and run-time optimizations to trade-off reliability for other system-wide metrics.

1.3.1 Model-based Energy-Delay Variation Prediction

In order to optimize memory design under the effects of variability, it is important to model global figures of merit such as performance and power for early-stage estimation. Choices made at the design-stage ensure conflicting requirements from higher-levels are decoupled. A multivariate regression based modelling technique for estimating energy/delay has been proposed. The models rely extensively on circuit-level simulations for gathering empirical data. The multi-dimensional regression problem is then simplified into lesser dimensions by a combination of regression and clustering. By generating a polynomial best-fit for the output data, the slope is optimized till the error between simulated data and model generated data is minimum. The median error between generated and simulated data for delay and energy was 1.75% & 0.8% respectively. When combined with architecture-level specifics, the models can then guide large-scale design space exploration.

In relation to this topic, two papers have been published.

- *MODEST: A Model For Energy Estimation under Spatio - Temporal Variability* [42]
  S.Ganapathy, R.Canal, A.Gonzalez & A.Rubio
  International Symposium on Low Power Electronic Design ***(ISLPED'10)***
- *Circuit Propagation Delay Estimation through Multivariate Regression-Based Modeling*

under Spatio-Temporal Variability [41]
S.Ganapathy, R.Canal, A.Gonzalez & A.Rubio
Design, Automation & Test in Europe Conference ***(DATE'10)***

1.3.2 Post-Silicon Adaptivity Using Hardware Monitoring

In order to take full-advantage of constant feature minimization, it is important to provide runtime support to complement design-level optimizations. Post-silicon adaptivity involves detecting changes in low-level circuit parameters (delay & leakage currents) post-manufacturing using on-chip canary structures and providing recovery circuits for effective repair. We have proposed a novel three-transistor one-diode (3T1D)DRAM based on-chip sensor for obtaining run-time latency/leakage profiles of memories. The profiles are stored in a lookup table that is referenced by dynamic fine-grain body-bias generator to generate an optimal body-bias that trades-off leakage power for cache latency . Our technique reduces leakage energy consumption and improves access latency of the cache on an average by 20% & 18% respectively.

In relation to this topic, one paper has been published.

- *Dynamic Fine-Grain Body Biasing of Caches with Latency and Leakage 3T1D-based Monitors* [43]
  S.Ganapathy, R.Canal, A.Gonzalez & A.Rubio
  International Conference on Computer Design ***(ICCD'11)***

1.3.3 Soft-Error Hardened Embedded 4T-DRAM Cell

We propose a a novel 4T-based eDRAM cell that when compared to a similar sized eDRAM cell has higher tolerance to process variations and soft-errors. We replace the gated-diode in a 3T1D (3-Transistor 1-Diode) cell with a NMOS pass transistor to suppress sub-threshold leakage which in turn improves the retention time. This is shown to improve the retention by 2.04X on average. The only downside though is that the absence of gated diode reduces gate-overdrive making the cell slower. The resulting increase in write access time is 3% is small enough to not have a system-wide performance degradation. Another function of the pass transistor is to reduce the total sensitive-area exposed to neutron strikes. With a long-channel length pass-transistor, the amount of time needed by the generated pulse to traverse through is long enough to generate only a

small glitch and not cause a bit-flip due to neutron-induced energization. The soft-error rate which is measured in terms of failures-in-time (FIT) is reduced by 36% on average.

In relation to this topic, one paper has been published.

- *A Novel Variation-Tolerant 4T-DRAM with Enhanced Soft-Error Tolerance* [44]
  S.Ganapathy, R.Canal, D.Alexandrescu, E.Costenaro, A.Gonzalez & A.Rubio
  International Conference on Computer Design ***(ICCD'12)***

1.3.4 Parametric Yield Enhancement Using Hybrid Techniques

The effectiveness of combining *failure-prevention* and *failure-reduction* techniques is thoroughly investigated in this study. Proactive read/write assist mechanisms that operate at the sub-array level reduce the failure probability of the SRAM-cell by modifying the electrical characteristics of a transistor at run-time. This is effected using techniques such as body-biasing or wordline boosting that either improve read or write noise margins. Assist techniques can help lower $V_{cc_{min}}$ whilst maintaining a failure rate observable at nominal voltages. This can enable large-scale power reduction. Reactive mechanisms such as error-correcting codes (ECC) and redundancy can then be used to recover from any persistent failures due to lowering of $V_{cc_{min}}$. While proactive techniques can help improve functional margin (lowering power), reactive techniques improve functional yield (total addressable space). Contrary to the notion that the two class of techniques are mutually exclusive, we show that the hybrid schemes offer better quality-energy-area trade-offs when compared to their standalone configurations.

In relation to this topic, one paper has been published.

- *Effectiveness of Hybrid Recovery Techniques on Parametric Failures* [45]
  S.Ganapathy, R.Canal, A.Gonzalez & A.Rubio
  International Symposium on Quality Electronic Design ***(ISQED'13)***

1.3.5 INFORMER: A Tool for Memory Robustness Analysis

Adopting a variation-aware design paradigm requires a holistic perspective of memory-wide metrics such as yield, power and performance. However, accurate estimation of such metrics is largely dependent on circuit implementation styles, technology parameters and architecture-level specifics. In line with the requirements, we designed a prototype tool (*INFORMER*) that helps high-level designers estimate memory reliability metrics rapidly

and accurately for a given SRAM cell design, technology, topology, working environment and memory architecture. The tool relies on accurate circuit-level simulations to capture multiple failure mechanisms (ageing, soft-errors and parametric failures) and helps couple low-level statistics with higher-level design choices. Additionally, to estimate the failure probability of rare events (low probability) where traditional Monte-Carlo based sampling techniques suffer, we propose a novel algorithm that leverages SRAM transistor dimensions to determine regions of higher failure probability. In these regions, we then determine the *most probable failure point* around which norm-minimization based importance sampling is exercised. The technique achieves near-SPICE like accuracy while improving simulation time by orders of magnitude.

In relation to this topic, a paper has been published.

- *INFORMER: An Integrated Framework for Early-Stage Memory Robustness Analysis* **S.Ganapathy**, R.Canal, D.Alexandrescu, E.Costenaro, A.Gonzalez & A.Rubio Design, Automation & Test in Europe Conference ***(DATE'14)***

2

Background and Related Work

2.1 Overview

In this chapter, we present an overview of several topics needed to better understand the fundamentals of variation-tolerant design. Section 2.3 broadly discusses different sources of variation and how they are characterized. This section covers variations arising from manufacturing processes, fluctuations in environmental parameters (supply voltage and temperature) and finally temporal variations due to degradation. In section 2.4, techniques to cope with variability ranging from design-level optimizations to run-time adaptation using circuit and architecture mechanisms are discussed. Section 2.5 discusses the impact of variations on embedded memories specifically. Techniques to lower power/performance/yield variation using novel cell topologies, post-silicon adaptivity and cache reconfiguration are discussed here.

2.2 Process Variations

Process variations are statistically defined by the difference in parameter values between intended design and the obtained product. They are primarily caused by imperfect manufacturing conditions due to challenges imposed by the lack of complete control over the process.

2.3 Sources of Variations

Spatial variations of process parameters affect the electrical characteristics of both transistors and the underlying interconnect fabric. They are broadly classified into statistical and systematic variations. Statistical variations arising from unfavourable conditions either during wafer processing steps such as oxidation, diffusion, etching, ion implantation, chemical- mechanical polishing (CMP), electroplating or annealing can cause variation in parameters such as impurity concentrations and oxide thickness. Further, limitations imposed by photo-lithography in advanced process nodes, makes it extremely cumbersome to draw smaller lines. In such scenarios, issues such as lens abberations cause a further deviation in the $\frac{W}{L}$ ratio of transistor and interconnect dimensions. Broadly speaking, spatial variations of process parameters can be divided into two classes:

- **Die-to-Die (D2D):** Mainly caused due to discrepancies in lithography, CMP and RTA, die-level variations are global inter-die variations that cause a difference across die(s) without affecting devices within a die.

- **Within-Die (WID):** Characterised by both systematic and random WID variations. Systematic variations are proximity and layout geometry dependent wherein variations in certain components cause a variation in a nearby similarly designed component. As the distance increases, the correlation in variation reduces subjecting on-chip components to accidental heterogeneity with differing power and performance profiles. Intrinsic device level fluctuations caused due to effects such as random dopant fluctuations and line edge roughness induce discrepancy in electrical characteristics even across adjacent transistors.

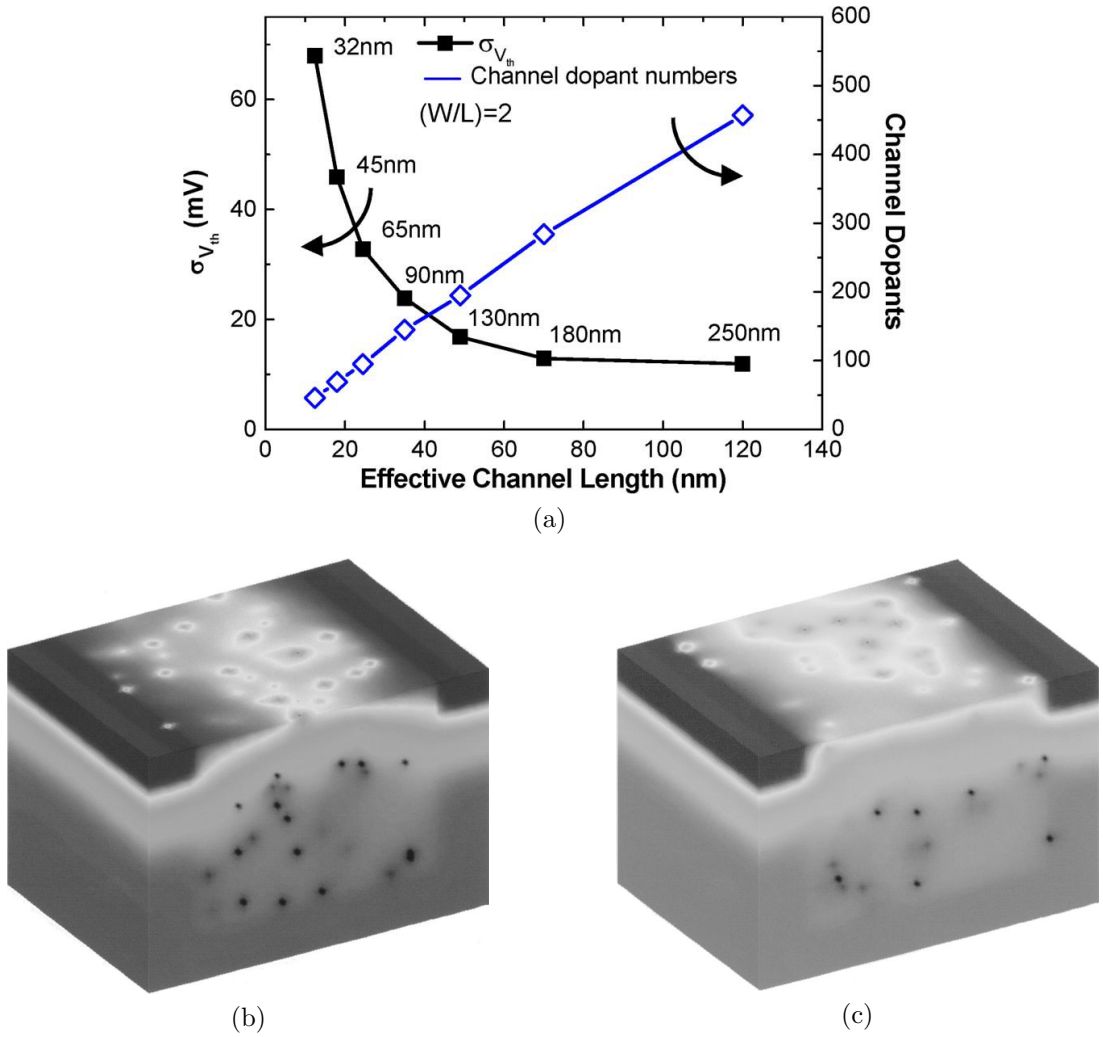
2.3.1 Random Dopant Fluctuations (RDF)

This type of variations are primarily caused due to the random placement of dopant atoms in the channel. The effect is more pronounced with scaling of technology where the total number of dopant atoms needed for implantation reduces with subsequent technology nodes. This problem of RDF has been well documented over the last three decades and has been predicted to be a major challenge for controlling device performance. Due to the random nature of this phenomena, the threshold voltage (V_{th}) of the transistor undergoes significant variation. This is because the intrinsic value of V_{th} is dependent on the charge of the ionized dopants in the depletion region. The standard deviation of V_{th} follows the inverse square law of the device area. In other words, with scaling of technology, σV_{th} dependent on RDF increases for transistors with smaller area. The variation in V_{th} due to RDF has been demonstrated to follow a Gaussian distribution with its standard deviation derived as,

$$\sigma V_{th} = (\sqrt[4]{2q^3\epsilon_{Si}N_a\phi_B}).\frac{T_{ox}}{\epsilon_{ox}}.\frac{1}{\sqrt{3WL}} \tag{2.1}$$

where q represents electron charge, ϵ_{Si} and ϵ_{ox} are permittivity of silicon and gate oxide, N_a is the channel dopant concentration, ϕ_B is the difference between Fermi level and intrinsic level, T_{ox} is the gate oxide thickness, W and L are the channel width and length of the transistor, respectively [3, 84].

The trend in reduction in the total number of dopant atoms when reducing device dimensions is shown in figure 2.1a. It is evident that with reduction in total number of dopant atoms with subsequent process nodes, the increase in σV_{th} is significant. Figures 2.1b & 2.1c show the cross-section of two-identical MOSFETs each with 170 dopant atoms. However, because of the difference in placement of the dopant atoms along the channel, the device in figure 2.1b has a V_{th} of 0.78V when compared to the device in figure 2.1c that has a threshold of 0.56V. This is primarily because of the fact that a minimum number of dopants in device (b) in the middle of the channel spaced equally along the channel width block the free flow of current increasing the threshold voltage [9]. As RDF is inversely proportional to the device area, SRAM cells constructed with minimum geometry transistors are intrinsically the most susceptible to this type of variation.



(a)

(b) (c)

Figure 2.1: (a) Scaling trend of V_{th} variance due to random dopant fluctuations (RDF) [114] (b) MOSFET with $V_{th} = 0.78V$ (170 dopant atoms) (c) MOSFET with $V_{th} = 0.56V$ (170 dopant atoms) [9]

2.3.2 Line-Edge Roughness

Line-edge roughness is mainly caused by the change in the shape of the gate along the channel width direction [114]. This roughness in the edge of the gate is caused by the inherent characteristics of the materials forming the gate and additional process steps such as etching and imperfection in lithography. The impact of this phenomenon is more pronounced at technologies below 50nm where stringent constraints set on the device performance require much higher level of control over the gate-length. The impact of LER induced variation on V_{th} follows Gaussian distribution and is inversely proportional to the gate width of the transistor [13]. The impact of LER when changing the device dimension from W_1 to W_2 on σV_{th} is given by the following equation,

$$\sigma_{V_{th}|W_2} = \sqrt{W_1/W_2}\ \sigma_{V_{th}|W_1} \tag{2.2}$$

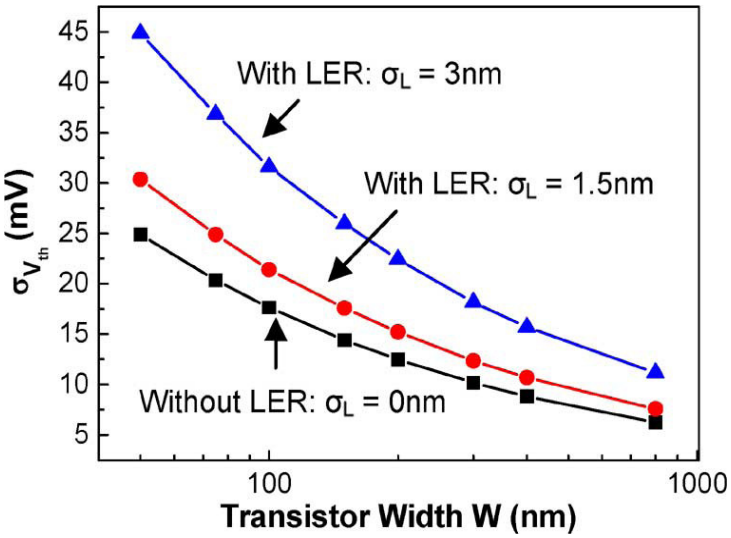


Figure 2.2: Impact of Line-Edge Roughness on V_{th} as a function of transistor width [114]

Figure 2.2 shows the impact of LER on V_{th} fluctuation with scaling of transistor widths. As explained in [114], the variance of this phenomena does not reduce with technology scaling despite improvements to the underlying manufacturing technology. As a result, the problem can become quite critical for devices such as memory cells that are extremely susceptible to device-level V_{th} mismatch.

2.3.3 Channel Length Variation

Variations or perturbations during the lithography phase are known to affect a large number of electrical characteristics. Variation in transistor channel length or critical dimension (CD) is the most critical from a performance perspective [87] This is primarily because of the fact that effective channel length being the smallest implemented on-die feature, it is the most susceptible to intra-die variation resulting from Optical Proximity Errors (OPE). For short-channel devices with +ve standard deviation, the resulting reduction in drive current increases propagation delay tremendously ($\approx C_L V_{DD}/2I_D$). This effect can be better explained by the phenomena better known as *Drain induced barrier lowering* (DIBL) where in technologies below 45nm, in short-channel devices, the threshold voltage is strongly dependent on the channel length. Using techniques such as forward body-biasing, the V_{th} roll-off can be reduced lowering the influence of channel length variation on V_{th}. However, this incurs significant overhead in leakage power.

2.3.4 Environmental Variations

2.3.4.1 Thermal variations

In-line with the consequences of Moore's law, on-chip power density has been observed to double approximately every three years. It is expected that this phenomenon will exacerbate with the introduction of new materials and constant feature minimization. As the dissipated energy is converted into waste heat, the corresponding increase in thermal density is creating new challenges in reliability and manufacturing costs [100]. While techniques like supply voltage scaling can help reduce the overall on-die temperature, it impacts speed negatively lowering processor wide performance. As enumerated in [13], spatial variation of temperature on-die is dependent on the following factors

- *Correlation with temperature profile of adjacent blocks:* Depending upon the switching activity and power demands of a particular block, the thermal characteristics of adjacent areas are influenced significantly. Frequent temperature shoot-ups can cause functionality problems resulting in failures that can permanently damage the underlying silicon.
- *Material dependence:* The rate at which the heat spreads is dependent on the thermal conductivity of the material used. Bulk-CMOS technology in comparison to silicon-on-insulator (SOI) helps lower temperature levels as the heat generated is spread into

the substrate and the interconnects. SOI technology, in contrast has lower levels of thermal conductivity of the oxide material making the interconnects better conductors of heat. By making the wafers thinner, resistance of the material is reduced moving the areas of heat generation closer to the heat spreading packaging.

- *Cooling and packaging:* Traditional designs incorporate either heat pipes or water-cooling to absorb most of the heat that is dissipated from the package. Heat pipes with their increased surface area help remove or move heat to an area with sufficient air-flow. To reduce cost and lower area foot-print in mobile devices such as laptops, packaging decisions often involve omitting the heat spreader plate and heat sink and instead use heat pipes and other mechanisms that reduce the weight and size of the sink.
- *Activity factor:* Run-time variations in the workload of the processor temporally influence both thermal as well as supply voltage variations. Dynamic thermal management (DTM) techniques often rely on on-die hardware sensors to build real-time thermal profiles and enforce cross-layer optimizations to reduce temperature in a staggering manner. Figure 2.3 shows the thermal image of microprocessor with hotspots of temperature as high as 120°C. As the power densities continue to increase with every subsequent process generation, the disparity in performance of blocks across a die will continue to widen.

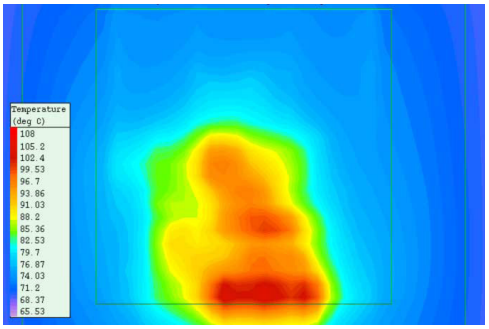
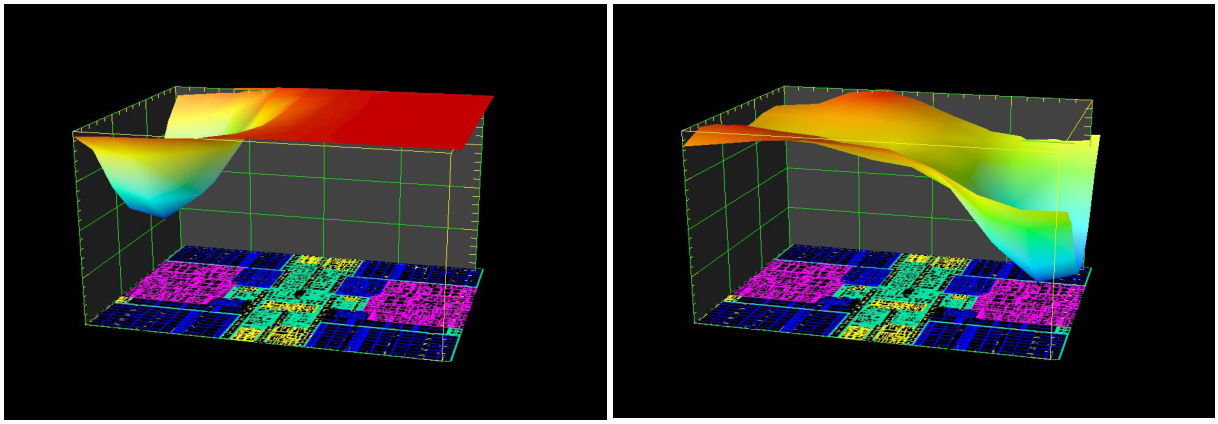


Figure 2.3: Within die temperature variations [15]

2.3.4.2 Supply voltage variations

This type of variation occur when there is a rapid change in the load (demand) within a short time interval. Due to the parasitic inductance in the power-delivery subsystem, voltage ripples are generated along the supply lines. This is fundamentally defined as the *Ldi/dt effect*. Due to the stringent constraints set on power consumption, mechanisms



(a) (b)

Figure 2.4: Voltage droop in chip multiprocessors where physical cores share the same power grid (a) Voltage droop in a core as it turns on increasing the load (b) coupled voltage droop in another block as it is switched on after a few ns delay [40].

like clock-gating, power-gating and idle/sleep modes that require fine-grained control of the run-time supply voltage make it difficult to prevent these types of variations. In chip multiprocessor architecture (CMP), depending upon the core utilization patterns and activity interaction between cores, the inter-core voltage fluctuation can be quite significant [48]. Figures 2.4a and 2.4b depict the phenomena of voltage droop in one block generating a new voltage droop in another block at a different instant of time. Initially, as shown in figure 2.4a a voltage droop is observed at a block, say Core 1. During this period, another block at the far right corner (say Core 2) is still in the idle mode. A few nano seconds later (4ns for the instruction stream to start on Core 2), as depicted in figure 2.4b, core 2 is switched on after a delay and experiences worse voltage droop compared to Core 1. This is primarily because of the coupling effect where the droop generated in core 1 is coupled to core 2 increasing the supply voltage noise. While it is very cumbersome to predict such run-time behaviour early in the design stage, sufficient guard-banding can mitigate the negative impact to a certain extent. The impact in such scenarios can be minimized by using on-chip regulators and decoupling capacitors [13].

2.3.4.3 Bias Temperature Instability (BTI)

Negative BTI (NBTI) occurs when a negative gate voltage (V_{gs} = -V_{dd}) is applied for a sufficient long period at high temperatures. It results in increased V_{th} with time degrading PMOS performance across the lifetime of the processor. NBTI experienced by PMOS

transistors results from the generation of interface traps along the channel and gate dielectric interface. Due to lower number of holes along the channel in NMOS devices, they do not suffer from NBTI.

NBTI occurs in two phases, namely *stress* and *recovery*. Periods of *stress* are caused due to the generation of interface traps as and when *Si-H* bonds are broken under the influence of high electric fields and elevated temperatures. The dangling bonds created by the separation of hydrogen-terminated trivalent silicon bonds (Si_3-Si-H) create traps at the interface causing hydrogen to diffuse into the gate-oxide. This results in a degradation of the V_{th} of PMOS devices. The annealing process of *recovery* is made possible by the application of V_{dd} to the gate that temporarily inhibits further generation of interface traps. As a result, the dissociated hydrogen bonds return to the interface to join with the broken silicon bonds to partially recover degraded threshold voltage.

Table 2.1: Reactive-Diffusion model based threshold voltage shift

	$\Delta\lvert V_{th}\rvert$ under NBTI
Stress	$\sqrt{K_v^2.(t-t_0)^{0.5} + \Delta V_{TH0}} + \delta_v$
Recovery	$(\Delta V_{TH0} - \delta_v).\lfloor 1 - \sqrt{\eta(t-t_0)/t} \rfloor$
K_v	$A.t_{ox}.\sqrt{C_{OX}(V_{GS}-V_{TH})}.$ $[1 - V_{DS}/\alpha(V_{GS}-V_{TH})].$ $exp(E_{OX}/E_0).exp(E_a/kT)$

Traditionally, the process of *stress* and *recovery* is modelled through the well established *Reaction-Diffusion* (R-D) theory [7, 73, 108]. Table 2.1 shows the different parameters influencing the amount threshold voltage shift during stress and recovery phases. It should be observable that the change in the threshold voltage (ΔV_{th}) is directly dependent on the total time under stress ($\sim t^{0.25}$) for a given supply voltage and temperature. By modulating the supply voltage, the overall power-density of the chip can be controlled which in-turn regulates temperature. Temperature has an exponential influence on NBTI *stress* thereby escalating the degradation at high temperatures (heavy-workload or high supply voltage). The degradation exhibited by NBTI is "front-loaded" in nature wherein the rate of degradation reduces with time [20]. By accounting for the supply voltage-speed dependence, the total time under stress can be considerably manipulated lowering the rate of ageing at nominal voltages. However, as demonstrated in [20], the returns on dynamic voltage scaling diminish due to the "front-loaded" nature of degradation and it cannot be used to extend the lifetime of the processor. Since the device is only under stress when a negative-bias ($V_{GS} = -V_{dd}|'0'$) is applied, the duty cycle (β) is an important parameter that can influence ΔV_{th} significantly [89]. Figure

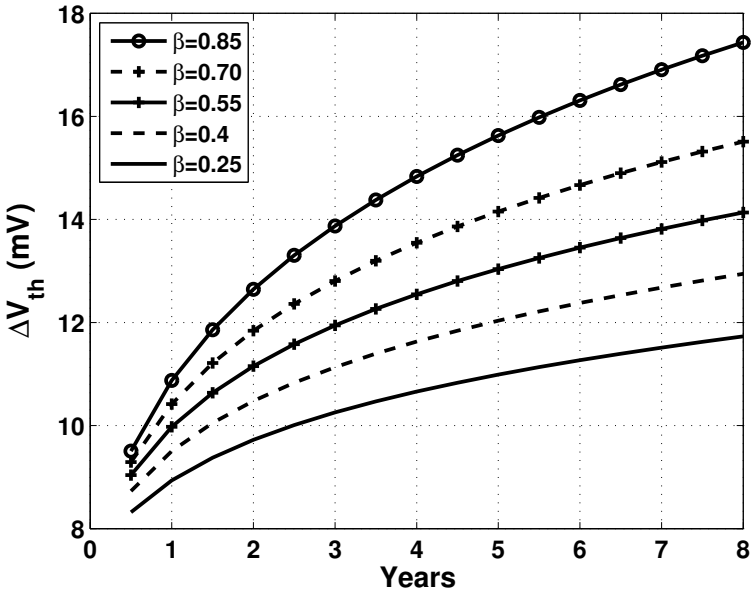


Figure 2.5: ΔV_{th} shift (under NBTI) in 32nm PMOS device operating at 85°C

2.5 shows the impact of duty cycle (β) variation on the ΔV_{th} for a 32nm PTM PMOS device operating at 85°C [1]. Duty cycle is defined as the ratio of time period between the time the device is under stress (*logic 0*) and recovery (*logic 1*). It can be observed that the rate of degradation (slope of ΔV_{th}) is higher for a device with larger duty cycle. Higher shifts in ΔV_{th} require higher margin of V_{dd} or temperature tuning to ensure sufficient recovery. Unlike other design specifications, duty cycle cannot be tuned arbitrarily and the effectiveness of duty-cycle modulation needs to be studied together with other metrics such as component utilization factor and input vector probability.

2.4 Coping with Variability

2.4.1 Modelling and Optimization using CAD

2.4.1.1 Propagation Delay Calculation

In order to provide a way to optimize performance parameters such as circuit propagation delay across a wide range of process parameters (random variations), statistical design is used extensively. While they can help cope with static variations arising from manufacturing processes, to cope with run-time variations of environmental factors needs on-chip mechanism that can provide necessary support. The first step towards designing fast

and reliable circuits is development of accurate timing models than capture performance variability under the impact of variations. This helps better understand the trade-offs between performance and functional correctness. Timing verification is used to determine the range the operational frequencies that can guarantee a certain yield. In this context, it is important to define two types of violations that are needed to verify the design from a timing perspective.

- *Hold-time violation:* This type of error is more prevalent in edge-triggered latches where the input is not held (or stable) for a sufficient period after the clock transitions. As a result, if the inputs to the latch change when the the edge is triggered, there is certain probability the output is in-deterministic for a short period of time.
- *Setup-time violation:* If the input signal arrives at a time before the clock has transitioned, then it results in a setup-time violation.

The arrival time of the input signal can vary depending upon several factors such as manufacturing variation, voltage, temperature and the implementation of a particular component. To compute the maximum and minimum bounds on the arrival times dependent on such factors, static timing analysis (STA) is one such technique that computes the propagation delay at specific process corners [12]. STA builds a look-up table of arrival and propagation delays of each component in the standard cell library for worst, nominal and best corners. One major disadvantage with STA is that it assumes perfect tracking where all the gates that make up a circuit share the same corner. Such an approach can be overly pessimistic or optimistic depending upon the choice of the corner. It is very optimistic if all gates are assumed to be in the fast-corner or extremely pessimistic if they are assumed to be in the slow-corner.

In-lieu of such disadvantages, statistical static timing analysis (SSTA) is an improved technique that builds a probability density function of the delay distribution of the circuit over a larger number of input variations [13]. Over here, the summation operation of individual gate delays in STA is replaced with convolution in SSTA [3]. Therefore, the problem of SSTA is defined as follows: given the probability distribution of the input parameters, determine the probability distribution of circuit delay.

2.4.1.2 Energy and Power Estimation

As we move from one process technology to another, we achieve a moderate reduction in switching energy but leakage due to sub-threshold and gate-oxide currents -in turn- grows exponentially caused by the reduced threshold voltage and increasing dependence

on temperature. Further, switching speed has decreased as the ratio of V_{dd}/V_{th} is reduced limiting the current driving capabilities. While supply voltage scaling has helped -to a great extent- in reducing the switching energy, in order to maintain robustness and reliable operation, the generational reduction in V_{th} has led to a significant increase in static energy (leakage). This would mean that estimation of idle energy is equally important as estimation of active energy. Energy estimation is possible through analytical, empirical and a combination of both methods [35, 36, 67, 74, 90]. Most designers prefer not to use empirical modelling techniques as they require complete specifications which are seldom available early in the design stage. Simulation based empirical approaches rely on an approximate design and assume a random set of input vectors for energy estimation. Monte-Carlo approaches mitigate the complexity to an extent for large designs by simulating a random input vector and determine the energy consumption per activity or power consumption within a specified time interval. On the other hand, analytical techniques rely heavily on theoretical approaches and most of the times, do not scale well with future technologies. Liang *et al.* [67] has demonstrated a hybrid analytical-empirical technique that takes the flexibility of analytical modelling while maintaining the accuracy and robustness of empirical solutions. Due to the very high dependence of sub-threshold and gate-leakage components of power on temperature and threshold, any high level approach targeting estimation of energy without the consideration of temperature and supply voltage variation is -nowadays-not a justified attempt. Any new model should ensure that it can be applied for design space exploration of energy-constrained and delay-critical designs.

2.4.1.3 Statistical Optimization

The most primitive technique towards moderating power and delay of switches is to modulate gate dimensions [81, 96, 99]. By increasing the length of the transistor, while leakage power is reduced, gate delay increases substantially. On the other hand, enhancing the width while improves the gate drive leading to faster transition times, also increases the current flow increasing power dissipation. The threshold voltage of transistors can be modulated either at design-time (dual-V_{th}) or at run-time (adaptive body biasing) for trading-off power for performance [69]. As postulated in the F_{max} theory, by reducing the depth of the pipeline, the impact of WID variations on delay can be reduced. In [31], the technique of unbalanced pipelines is proposed where different stages of the pipeline are designed by trading-off yield for area and power. The cumulative yield of the overall

pipeline defined as the product of the yield of each stage can be optimized by a combination of under- and over-designing specific stages independently saving on power and area.

2.4.2 Circuit Techniques

Post-silicon adaptive circuit-level designs involve measuring low-level transistor parameters like leakage currents and propagation delays post-manufacturing and provide on-chip mechanisms that alter one or more number of control parameters to ensure both power and performance satisfy desired targets. Body biasing is one such technique where the body-source potential is modulated to control the threshold value of transistor. In forward body biasing (FBB), a large positive voltage ensures that circuits are faster but at the cost of increasing leakage. In reverse body biasing, negative bias voltage increases the threshold making circuits slower and also less leakier. Tolerance to variations can be increased by utilizing both RBB and FBB mutually-exclusively and this is called Adaptive Body Biasing (ABB)[106]. The super-linear relationship of dynamic & leakage power with supply voltages can be exploited to control power consumption by adaptively scaling the supply voltage (ASV). When compared to ABB, ASV offers less area overhead. Recent works have proposed an ASV technique to lower the power consumption in processor pipelines [64]. However, such techniques are not applicable to caches where timing is critical. As the threshold voltage, like frequency, can control both power (leakage) and delay, dual-threshold designs have been proposed [6]. While all gates that lie on the critical path are assigned low-threshold for performance reasons, gates in the non-critical path are assigned high-V_{th} for reducing sub-threshold leakage. Unlike, other techniques, this does not require additional circuitry and enables high performance and low leakage simultaneously. Such a technique cannot be implemented in regular caches as all paths are considered to be critical and assigning low-V_{th} to all gates would result in tremendous increase in leakage power. As temporal variations of temperature and power are known to inhibit frequency scaling, dynamic fine grain body-biasing (DFGBB) utilizes feedback from critical path replicas to gauge the impact of variations on delay and adjusts the bias voltage frequently to account for dynamic variations [104].

Razor is a well-established technique towards improving safety margins reliably when scaling supply voltage without resulting in a failure [14]. The technique relies on an approximate critical path emulator that inserts *shadow* latches in addition to *master* latches between pipeline stages. The clock input to the *shadow latch* is a delayed version

of the system clock. With supply voltage scaling, as the delay increases, there is a certain probability (leading to a failure) that the values latched by the *master* and *shadow* flip-flops are different. In the event of an error, the output values of all *shadow* latches are XOR'ed to generate a single error signal. Recovery is initiated by rolling-back to the correct value using pipeline flushing and the voltage is scaled till the number of errors generated can be tolerated by the system.

2.4.3 Architecture Adaptation

At the system level, the impact of process variations is reduced by using guard banding techniques that typically involve running all cores at the frequency of the slowest core. This can have profound effects on the performance as all cores do not operate at the maximum allowable frequency. In [37], selective core-allocation is enforced for parallel workloads by varying the number of active cores. The optimal number of active cores is derived analytically by optimizing the power/performance ratio. In a similar fashion, variation-aware dynamic voltage frequency scaling (DVFS) has been utilized by selective mapping of different applications to different cores based on power/performance disparity across cores [49]. It was found that per-core voltage/frequency islands (VFI) are not viable given the area and routing complexity when compared to chip-wide VFI. In other proposals, across-core frequency distribution is minimized by using modified thread-to-processor mappings [11, 52, 63]. Tiware *et.al* propose *ReCycle* which relies on borrowing time between adjacent pipelines to improve the overall effective clock frequency [105]. An orthogonal approach to reducing inter-core frequency disparity is to adopt a globally asynchronous and locally synchronous (GALS) system design approach [95]. To enable synchronization at the synchronous-asynchronous interfaces, additional complexity is involved in the design. Further, the overall clock frequency is dependent on the slowest synchronous units thereby inhibiting faster units from operating at higher frequencies.

2.5 Variation-Tolerant Embedded Memory Design

2.5.1 Novel Cell Topologies

Alternatives to 6T-SRAM based memories have been researched diligently for want of increased memory density and lower vulnerability to variations. One such proposal is the

three-transistor one-diode (3T1D) DRAM cell proposed by Luk *et al*[71]. The capacitor-less DRAM cell stores the data using a gated diode that is tied to the read-wordline. The 3T1D unlike regular 1T1C eDRAM memory provides non-destructive reads and access speeds comparable to that of standard SRAM cells. When compared to regular SRAM cells, the transistors of the 3T1D can be asymmetrical in strength. This has a 2 fold advantage: Primarily, process variations causing device mismatch are likely to cause less failures to the cell. Secondly, it improves the overall stability making it radiation hardened against soft-errors. Liang *et al.* have proposed a 3T1D based cache architecture that relies on the fact that data stored in first level caches is transient and the time to the last cache reference to a given block (before it is erased) is well within the retention time of a 3T1D cell [66].

Failures in 6T-SRAM are characterized into three types - *Access Time Failure*, *Read and Write Stability Failure*. They result in either very slow access speeds or erroneous output caused as a result of bit-flipping. It is assumed that the primary sources of such failures is variation due to random fluctuations in the threshold voltages of the six individual transistors that constitute the cell [5]. The reduction in threshold voltage variation due to random dopant fluctuations can be reduced by designing SRAM cells with larger transistors. This is because the standard deviation (σV_{th}) is directly proportional to the inverse of the device area [103]. With reducing feature sizes, the β & γ ratio (proportional to cell size) of 6T-SRAM cells are increased to improve read and write margins. In addition to increasing cell area, the operational voltage is still restrictive. This is of particular concern in dynamic voltage schemes where wide range of power scaling is expected. As a result, 8T-SRAM cells are preferred over regular 6T-SRAM(s) in first level caches of modern day processors [39]. In addition to improving read/write stability, such designs provide separate read/write ports and allow for the operation at ultra low voltages [21, 76]

2.5.2 Device and Circuit

Halo doping profiles of transistors can be modified at design-time for increasing threshold-voltage leading to lower leakage power. On a similar vein, SRAM cells are optimized for High-V_{th} during design-time and at run-time forward body biasing is employed to reduce access latency. Results indicate a 64% leakage reduction when compared to state of the art techniques without significant performance loss [60]. Cache performance can also be improved by reducing the latency (at a line-level granularity) by selectively boosting word-

lines of failing rows [83]. Though such techniques incur dynamic power overhead, they enhance the parametric yield even under worst-case process variations. Recent proposals have suggested that post-silicon adaptivity can be used effectively to improve SRAM yield and also reduce power consumption significantly [25, 46, 98]. Post-silicon adaptivity typically involves measuring low-level circuit parameters (delay & leakage currents) post-manufacturing and providing on-chip mechanisms to enforce circuit-level optimizations to ensure yield targets are met.

In the ultra low-power domain that is constrained by the minimum voltage of operation ($V_{dd_{min}}$), caches are susceptible to a large number of transient or soft failures due to voltage noise. Assist methods that rely on modulating one or more voltage sources (V_{dd},V_{ss},V_{bias} & V_{wl}) connected to the SRAM cell improve read/write noise margins adequately [75, 83, 86]. Then, there are also other techniques that vary the pulse-widths of one or more enable signals that govern the regular access to the cell [58, 59]. While most target improving the overall cell failure probability, there are techniques which specifically target either read or write margins. Reactive techniques such as ECC and redundancy operate on failing lines and improve the overall functional yield (total addressable cache space). In [61, 116], both ECC and redundancy are used to evaluate the extent of parametric failures. It was shown that by optimizing the cell for reduced area, the impact of the consequent increase in failure probability can be mitigated by using either redundancy or ECC or both. In [5], block rearrangement technique is enforced to map faulty cells onto redundant columns and disabling faulty columns at the microarchitectural level. The technique improves the overall yield by 61%. ECC is used to correct failures that are generated by lowering Vdd_{min} in [110].

2.5.3 Architecture reconfiguration

Effective cache resizing mechanisms to reduce failures include *a)* Identification of faulty blocks using BIST circuitry and forcing the column multiplexer to remap and select a different block [5]. *b)* Utilising PADed decoders for rearranging pairs of physically adjacent high-latency blocks across logically adjacent sets thereby reducing the spread of physically adjoint low-latency sets across logical sets [79, 97]. *c)* Using effective cache-set prediction to identify optimum number of sets required for an application and then selectively turning-off variation affected blocks ensuring atleast one block per set is active [53]. *d)* Selectively turning off cache-ways that violate latency/leakage targets with minimal performance degradation [82]. This is based on *Selective Cache Ways* technique

[8]. In [30], Das *et al.* have proposed a novel framework for analysing cache yield that is aware of both process variations and revenue. Also a new cache redundancy scheme called *substitute cache* that replicates data from cache lines affected by process variation has been proposed. The only priority for replicating cache words in the redundant cache is lines with very high latency.

There are several other works in this area that have been published. In this chapter, we have discussed only the most relevant works to the research conducted in this thesis.

3

Energy-Delay Modeling and Optimization: CAD Approach

3.1 Overview

With every process generation, the problem of variability in physical parameters and environmental conditions poses a great challenge to the design of fast and reliable circuits. Increasing interaction of low-level process parameters with processor power and performance prompts the need for development of architectural level models designed for simultaneous co-exploration of circuit-centric optimizations. Towards this end, in this chapter, we propose two independent techniques to model propagation delay and energy variation in SRAM memories under the impact of spatio-temporal variations. These models can then be leveraged for prediction of the behaviour of such memories under dynamic operating conditions. The proposed models are validated against data obtained through rigorous SPICE-level simulations of the memory critical path. The remainder of the chapter is organised as follows: in section 3.2, we motivate the need for a new modelling approach. Section 3.3 discusses the multi-level spatial grid model used for

capturing process variation. Sections 3.4 - 3.6 discuss the proposed path-based delay modelling technique and how it is extended to estimate delay variations in the critical path of a memory pipeline. In section 3.7, we extend the proposed delay model to study the impact of two circuit-optimizations on delay variability. In a similar vein, sections 3.8 - 3.4 discuss a new technique to estimate energy variations in memories. The results of the technique coupled with a use-case are presented in section 3.10. Finally, the concluding remarks are presented in section 3.12.

3.2 Introduction

For us to understand the impact of variability on circuit behaviour, we need to see its effect on the 2 most important variables - power and performance [80]. While power has been a major concern for a very long time from an energy efficiency point of view, there is very little knowledge about the effect on delay due to combined effects of spatio -and- temporal variations. The importance of such metrics has prompted chip designers to build tools (models) with different levels of abstraction capable of performing design space exploration of such power and performance-bound designs throughout the design flow. These tools provide enough flexibility to rapidly explore different designs with sufficient accuracy.

In this chapter we propose two independent modelling methodologies that help designers identify power and performance bottlenecks in the memory pipeline and allow for the incorporation of circuit-centric optimizations for design space exploration at the architectural level.

3.3 Capturing Spatial Correlations

Spatial variations can be assumed to be the summation of Die-to-Die (D2D), Within-Die (WID) Systematic and Within-Die Random components. D2D variations are known to affect devices within the same die similarly. Systematic variations are deterministic in nature. Random variations on the other hand, are caused due to dopant fluctuations and line edge roughness[80]. Their effect is modelled by assuming a finite parameter variation. Sarangi *et al.* [93] have modelled the correlation in parameter values between 2 points as a cubic function of the distance. It is an analytical model and the results have been vali-

dated against empirical data. Agarwal et al.[4] have proposed a multi-level QUADTREE based modelling approach that reverse engineers the empirical residual modelling proposed by Stine et al.[102]. We begin with [4] as the base for our model and extend it based on the following assumptions:

1) Correlation between spatial points is a function of similarity of structures also. Structures having similar layout geometries will be affected in a similar fashion [29]. For example, columns of replicated SRAM cells will have similar parameter distribution.

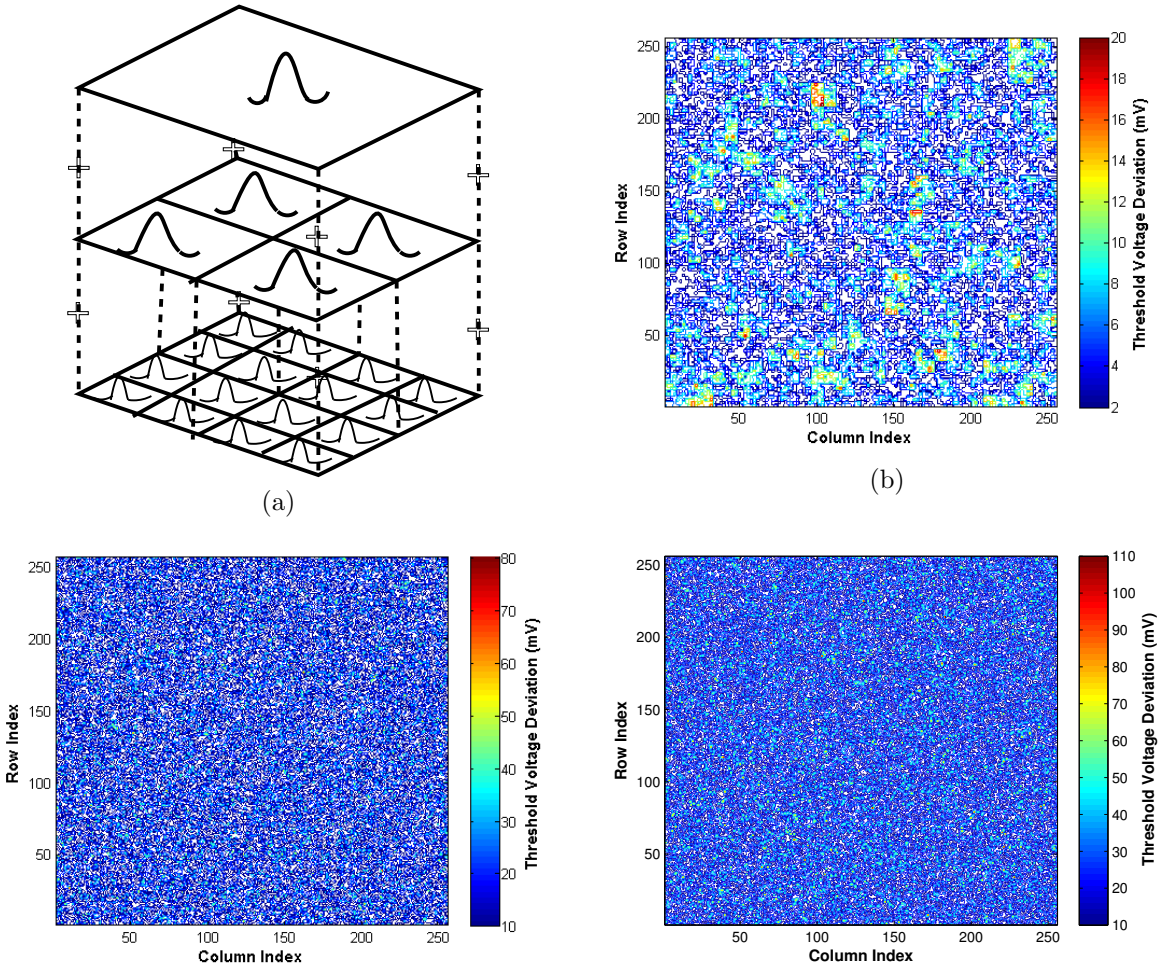
2) Physical parameters taken into account are V_{th} and L_{eff} are assumed to be Gaussian distributed. Variation in other parameters can be assumed as a function of the variation of above listed parameters.

3) Worst case deviation of parameter values is rare. By virtue of the additive nature of the model, realistic predictions about the distribution of parameter values can be made.

As shown in figure 3.1a, variations in process parameters are modelled using a multi-level hierarchy of spatial variation maps. The bottom-most level represents the floor-plan of the die that is partitioned into multiple grids. The value of a process parameter $P_{i,j}$ in a grid in the last level at co-ordinates i,j is given by:

$$\begin{aligned} P_{i,j} &= P_{nom} + \delta P_{variation} \\ &= P_{nom} + \delta P_{inter-die} + \delta P_{intra-die} \\ &= P_{nom} + \delta P_{inter-die} + \delta P_{systematic_i} \\ &+ \delta P_{random_{i,j}} \end{aligned} \tag{3.1}$$

All points within a grid are assumed to have a similar parameter distribution (or perfect correlation). This way, distance based correlation is achieved by virtue of the model and layout based correlation by virtue of the specification of the model. For 256x256 sub-array, the number of levels in the model is 9 where we use the top most level for inter-die variability and the bottom most for intra-die random variations. The remaining 7 intermediate levels share the total correlated variation among themselves. For each sub-array, we generate 7 such maps where 6 of them correspond to the random variation maps of the six-transistors (6T-SRAM) and the remaining one is shared among all the other transistors. Figures 3.1b, 3.1c and 3.1d show sample generated maps for correlated variation, independent random variations and overall variations derived using the multi-level spatial grid model respectively.



(a) (b) (c) (d)

Figure 3.1: (a) Multi-level spatial grid model to capture correlated variations dependent on proximity (b) Sample variation map for 256x256 memory array only with correlated variations (c) Sample variation map for 256x256 array only with random variation (d) Overall variation map with summation of D2D, WID-Systematic and WID-random variations

3.4 Path-Aware Delay Modeling

3.4.1 Transistor Specific Delay Modeling

Commercial CAD tools perform timing analysis of gates using lookup tables built during library characterization. For every combination of input slew and output loading, the delay is obtained as a function of the gate length [29]. The methodology as shown in Figure 3.2(a) assumes that all transistors within the gate have similar L_{eff} and V_{th} values. This assumption leads to a single distribution for delay as a function of varying parameters. Also, the impact of delay due to temperature variations on delay is higher when compared to other physical parameters. In reality, what happens is that gates have multiple distributions depending on their current state as shown in Figure 3.2(b). With

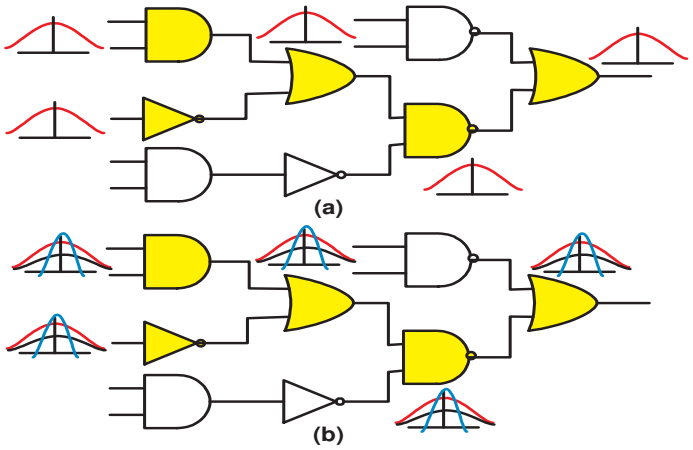


Figure 3.2: Example schematic implementing a path-based delay estimation methodology

increase in temperature, the probability density function broadens and the number of samples tending towards the mean decreases with more number of samples towards right hand side of the mean.

Consider the case of 2 NOT gates connected in series. For any input, both pull-up and pull-down network of adjacent gates would actively take part in the output transition. This model would only assume a variation in the inputs and it will not account for individual variations across the transistors that constitute the path between the input and output. If we are able to determine a path for every possible input and model the parameters of every transistor along the path as a random Gaussian function, then this would result in an ideal delay model that is aware of the characteristics of every transistor

that makes up the path. Nevertheless, the complexity of generating such a model would be very high. Considering the example of 2 NOT gates in series for the case where the input is 1, we extend the input based gate delay model of [112] to a path aware propagation-delay model using Equation 3.2,

$$\begin{aligned} D &= D_{P1} + D_{N2} \\ &= (D^{pmos}_{nominal} + \delta D_{ST_{P1}}) + (D^{nmos}_{nominal} + \delta D_{ST_{N2}}) && (3.2) \\ \delta D_{ST_{P1}} &= m_{leff} * \delta^{P1}_{leff} + m_{vth} * \delta^{P1}_{vth} + m_{temp} * \delta^{P1}_{temp} && (3.3) \\ \delta D_{ST_{N2}} &= m_{leff} * \delta^{N2}_{leff} + m_{vth} * \delta^{N2}_{vth} + m_{temp} * \delta^{N2}_{temp} && (3.4) \end{aligned}$$

where δD_{ST} is the variation in delay due to ST variations. Here, δD_{ST} is composed of variations due to spatial variations of V_{th}, L_{eff} and temporal variations of temperature. Assuming a linear dependence between spatial parameters and delay [27], we can derive a more generalized equation for any path i composed of $2j$ transistors given by the equation 3.5,[1]

$$\begin{aligned} D_i &= \sum_{a=1}^{j} (D^{pmos}_{nominal} + m^{pmos}_{leff} * \delta^{p(a)}_{leff} + m^{pmos}_{vth} * \delta^{p(a)}_{vth} + m^{pmos}_{temp} * \delta^{p(a)}_{temp}) \\ &+ \sum_{a=1}^{j} (D^{nmos}_{nominal} + m^{nmos}_{leff} * \delta^{n(a)}_{leff} + m^{nmos}_{vth} * \delta^{n}_{vth}(a) + m^{pmos}_{temp} * \delta^{n(a)}_{temp}) && (3.5) \end{aligned}$$

Equation 3.5 can be modified to equation 3.6 for same temperature devices and a nominal delay determined at design stage,

$$\begin{aligned} D_i &= \sum_{a=1}^{j} (m^{pmos}_{leff} * \delta^{p(a)}_{leff} + m^{pmos}_{vth} * \delta^{p(a)}_{vth}) + \sum_{a=1}^{j} (m^{nmos}_{leff} * \delta^{n(a)}_{leff} + m^{nmos}_{vth} * \delta^{n(a)}_{vth}) \\ &+ j * (D^{pmos}_{nominal} + D^{nmos}_{nominal}) + m_{temp} * \delta_{temp} && (3.6) \end{aligned}$$

$$m_{temp} = j * (\delta^{p(a)}_{temp} + \delta^{n(a)}_{temp}) \quad (3.7)$$

Solving Equation 3.6 for solution of the form Y=Xb+ϵ and estimating b=YX^{-1} we get,

[1] m^z_y represents the slope of parameter y for device of type z
$\delta^{p(z)}_y$ represents the delay deviation due to parameter y for the z^{th} pmos device
$\delta^{n(z)}_y$ represents the delay deviation due to parameter y for the z^{th} nmos device

$$
\begin{aligned}
D_i \;=\; & \begin{vmatrix} \delta_{leff}^{p(1)} & \delta_{vth}^{p(1)} & \delta_{leff}^{n(1)} & \delta_{vth}^{n(1)} \\ \delta_{leff}^{p(2)} & \delta_{vth}^{p(2)} & \delta_{leff}^{n(2)} & \delta_{vth}^{n(2)} \\ \vdots & \vdots & \vdots & \vdots \\ \delta_{leff}^{p(j)} & \delta_{vth}^{p(j)} & \delta_{leff}^{n(j)} & \delta_{vth}^{n(j)} \end{vmatrix} \times \begin{vmatrix} m_{leff}^{pmos} \\ m_{vth}^{pmos} \\ m_{leff}^{nmos} \\ m_{vth}^{nmos} \end{vmatrix} \\
& + \; j * (D_{nominal}^{pmos} + D_{nominal}^{nmos}) + m_{temp} * \delta_{temp}
\end{aligned} \tag{3.8}
$$

Solving Equation 3.8 would result in the slope values that closely follow the delay distribution. As in [29], we present the results of analysis only for one path (the other paths are assumed to behave similarly). In other words, the results presented show only one path derived from a given set of inputs; which is the case for regular structures such as memory structures; as well as for any balanced (i.e. similar delay paths for different inputs) design.

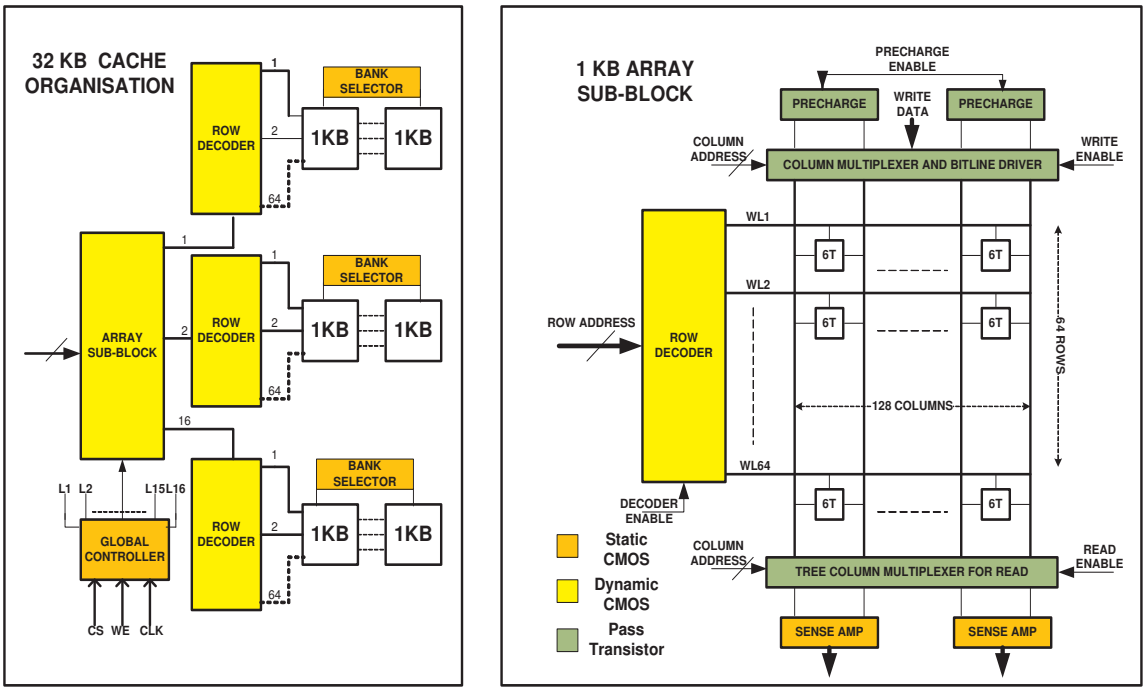


Figure 3.3: 32KB memory array design using sub-blocking for energy minimization

3.5 Memory Array Model

A cache as shown in Figure 3.3 was designed and implemented in HSPICE. It is a 32KB memory block (similar to D-Cache) that takes advantage of array sub-blocking to reduce the impact of variations [65]. Every sub-block is decoded using a array sub-block decoder and a row decoder decodes the row within the decoded sub-block. The global and local controller generate timing signals for the following: address generation, precharging and read/write enable. Wires are more resilient to variations when compared to logic and memory structures and hence we assume it to be a fixed amount of the obtained access time. As for access time calculation, we have adopted a model similar to the one implemented in CACTI [111].

$$\begin{aligned} T_{write} &= T_{control} + T_{arraysub-blockdecoder} + T_{addressdecoder} \\ &+ T_{wordlinedriver} + T_{bitlinedriver} \end{aligned} \tag{3.9}$$

$$\begin{aligned} T_{read} &= T_{control} + T_{array-sub-block-decoder} + T_{address-decoder} \\ &+ T_{wordline-driver} + T_{sram-transfer} + T_{sense-amplifier} \\ &+ +T_{column-multiplexer} \end{aligned} \tag{3.10}$$

It is a state of the art tool to determine the absolute access time of caches based on size, technology and supply voltage characteristics. While it is capable of calculating access times for any given temperature, it does not take into consideration the joint impact of spatial and temporal variations. Moreover, we are interested only in the calculation of normalized propagation delay which would give a better idea about worst case performance under spatio-temporal variations. Extending the proposed propagation delay model to calculate the access time would be a cumbersome process due to the existence of very long paths from the control circuitry to the cell(/sense amplifier) for write(/read) access measurements. Looking at the model for R/W access in Equation 3.9 and 3.10, we find that the access time is the summation of individual segment delays where each segment delay is a dependent variable. This analysis is a specific form of the Multivariate Analysis of Variance (MANOVA) [28] where the interaction between the independent variables of every segment (L_{eff},V_{th} and temperature) with its dependent variable (segment delay) is established and the individual contribution to the final dependent variable (access time) is determined. There are two main advantages of breaking the logic path

into smaller segments. Firstly, it would greatly reduce the number of paths and secondly, the estimation of segment delay would reduce the error in calculation of the overall delay. This would also expose those variables that behave randomly and provide scope for control mechanisms. Algorithm 1 summarises the proposed flow for deriving the normalised delay model based on user supplied input netlist and distribution of transistor parameters.

Algorithm 1: Compute Delay-Model

```
Input: Input Memory Netlist and Parameter Distribution
Output: Variation Aware Normalised-Delay Model
while Netlist Confirms with Floorplan do
    while NumberofSegments ≠ 0 do
        for Every Monte Carlo Run do
            for Every Parameter in the List do
                P = P_nom + δP_variation;
                Print P;
            Perform Simulation;
            Measure Propagation Delay;
            Normalize Value to smallest Delay
        for Every Temperature Range do
            Build Lookup Table for Parameter Values and Propagation Delay;
            Fit the Input Parameters to Normalized Output Delay using Model;
            Extract the Slopes of the Model;
        Build Lookup Tables for Slope Values and Temperature Values;
        Fit Input Temperature to Output Slopes using Linear Regression;
        Print Model Parameters;
    NumberofSegments = NumberofSegments − 1
```

3.6 Experimental Results

The cache is simulated on the HSPICE simulator using 16nm Predictive Technology Model [1, 19]. The parameter deviations at each level of granularity is specified in Table 3.1. In order to provide scope for a broad design space, the estimates made in table 5.1 are large enough to enclose any reasonable design point. From [93], we know that the standard deviations of both systematic and random components are equal and σ/μ of L_{eff} is strongly correlated to the systematic component of V_{th}. For every temperature

Table 3.1: Parameter Deviation

Parameter	σ_{D2D}	$\sigma_{WID_{systematic}}$	$\sigma_{WID_{random}}$
V_{th}(nmos,pmos)	±3%	±6.4%	±6.4%
L_{eff}	±3%	±3.2%	±3.2%
Temperature	30°C-110°C(Step 10°C)		

interval we run 500 Monte-Carlo samples. This is sufficient enough to validate the notion that propagation delay is linearly dependent on the selected input parameters. The Monte-Carlo samples generated for each temperature interval is constant across the entire range. The access times are normalized to the fastest cache at 30°C so as to facilitate the estimation of worst, best and average case behaviour. The simulations were performed on a machine with a dual-core processor running at near 3GHz with 4GB of main memory.

3.6.1 Simulation Results

Figures 3.4a-3.4f show the error box-plots between our model output and the HSPICE simulation output. The error is computed for every individual segment present in the access time model. The advantage of such a segment based methodology was to test the model over different types of structures (static, dynamic and pass). In five cases, the median error computed is less than 5% with the lowest being 1.8%. In the specific case of the address decoder, the model performance is degraded as the median error is of the order of 14%. This can be attributed to the fact that the address decoder built with dynamic CMOS gates drives a large load inside the sub-block. Thus, the effects of ST variation will be most felt by this structure and the non-linear behaviour is very tough to be captured by such a linear model. However, due to a segment based model computation, the overall error in the model can be reduced substantially with higher number of Monte-Carlo runs. This study particularly focusses on using the model for rapid design space exploration and does not delve deep into deriving near-accurate estimates. There is no relation between the increasing temperature and the model error rate. At higher temperatures due to the very high dependence of delay on temperature, access time failures [5] occur very frequently. It results due to the increase in read/write times. Thus the number of successful samples generated for post-simulation analysis will decrease with every temperature increment. This phenomena can be particularly observed in the case of sense amplifier & column multiplexers which drive large loads with minimal number of devices.

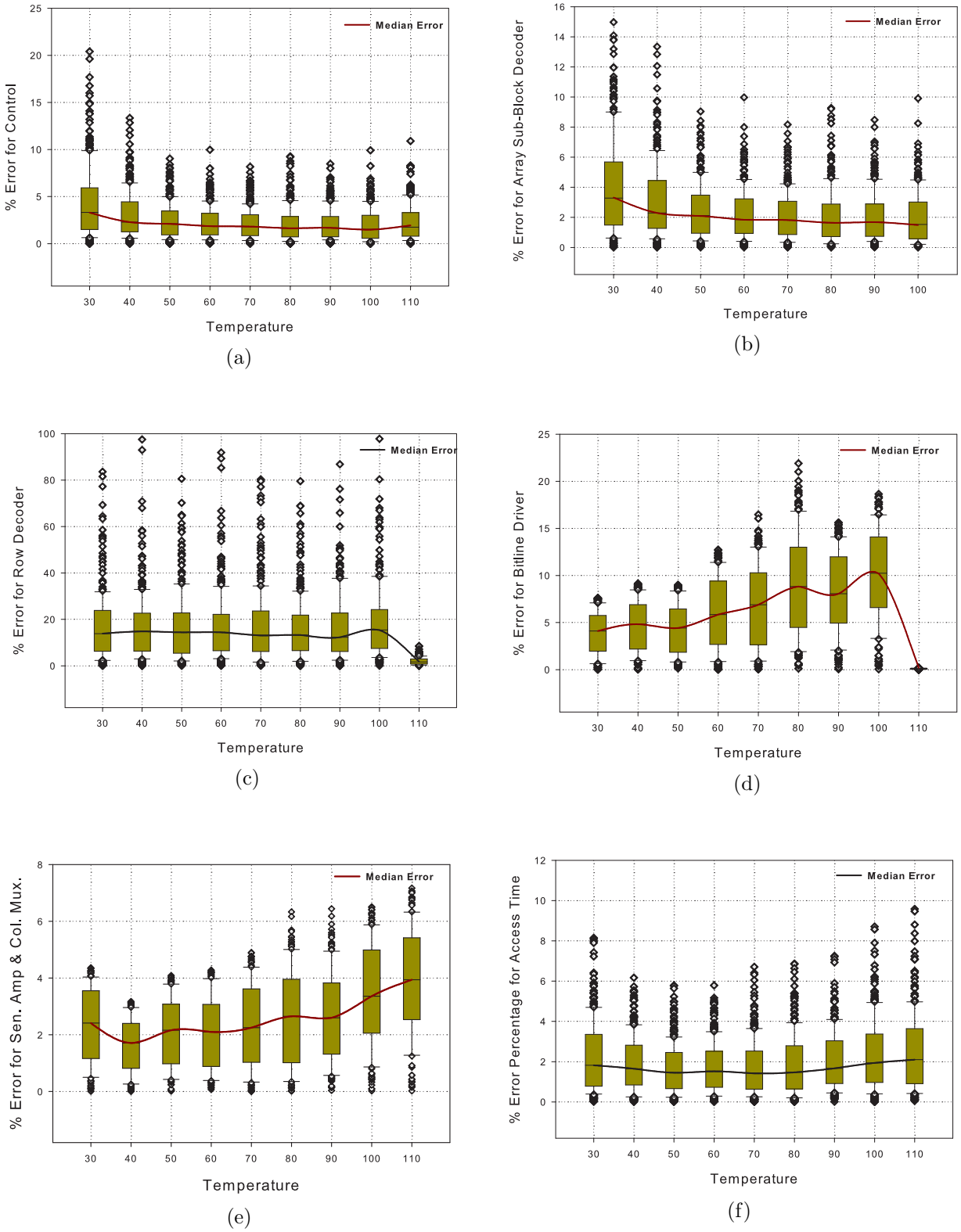


Figure 3.4: Error in Calculation of Propagation Delay between Proposed Scheme and HSPICE for (a) Control Circuitry (b) Array Sub-Block Decoder (c) Row Decoder (d) Bitline Driver (e) Sense Amplifier and Column Multiplexer (f) Overall access time

Table 3.2 presents the simulation time for HSPICE and the model. The model achieves very high speed-ups with increasing number of samples. For instance, when running a decent amount of samples (over 500 simulations), the speed-up of the model is of the order of $1x10^5$. The model generated for this study is based on a specific memory critical path circuit. However, using the aforementioned methodology, a similar model for other circuits can also be constructed.

Table 3.2: Simulation time speedup of the model compared to HSPICE simulations

Monte Carlo Runs	HSPICE Execution Time(s)	Model Execution Times(s)	Speed Up [$x10^3$]
50	7105	0.245	29
100	14960	0.312	47.9
200	30304	0.378	80.2
500	72243	0.557	129.7
1000	143757	0.978	146.9
2000	289260	1.762	164.2

3.7 Use Cases of Proposed Model

3.7.1 Simultaneous Impact of Temperature and D2D Variation

This section uses the model derived and validated in the previous sections to study the effect on the access time of a 32KB direct-mapped cache memory under process and temperature variations. To conduct the study, we took a sample of 500 dies. Figure 3.5 shows the normalized access time for the fastest, median and slowest chip for each temperature. At the same time, we also plot HSPICE simulations on those chips so that the precision of the model can be seen for each configuration and across temperatures. In the case of best and worst case performance, the model results are the equivalents of the simulator output. For average performance, the average over the entire set of available output samples was calculated for both simulator and model outputs. In accordance with the linear dependence of delay with temperature, the access time increases sharply beyond temperatures of 60°C. The combined effect of process and temperature variations degrades significantly the performance of the cache studied which warns us of the negative effects of variations. Plus, using the proposed model, the study just needed a fraction of the time (compared to HSPICE) to be completed. This speed and preciseness may allow this model to be a good technique to be incorporated into early stage design tools as well as run-time performance prediction or risk prevention tools.

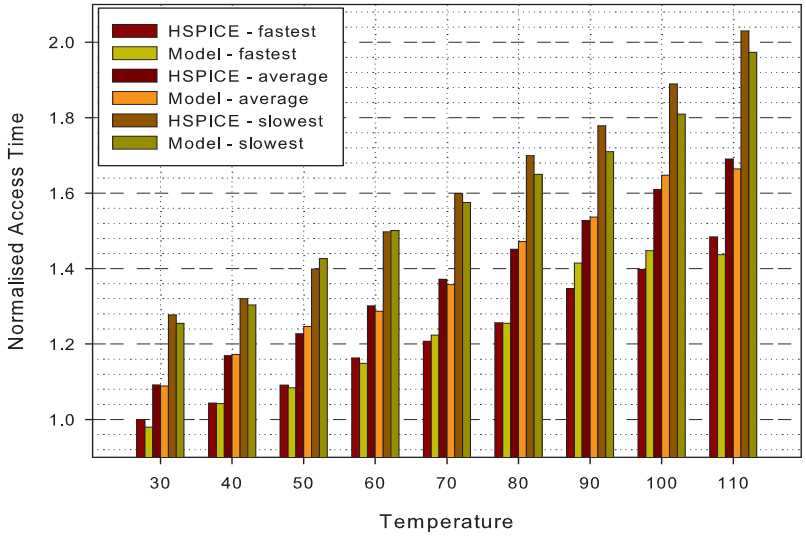


Figure 3.5: Variation of normalized access time across multiple dies and temperatures.

3.7.2 Memory Critical Path Dual-V_{th} Assignment

Though the model is capable of considering the impact of parameter variations on circuit propagation delays; it can also be used to evaluate, for instance, the cache behaviour under Dual-V_{th} assignments (where only V_{th} varies) [69]. While reducing the V_{th} decreases propagation delay, it increases leakage power. In [5], failures due to variation of threshold voltage within the 6T SRAM cell is discussed. Hence, we run our model (and HSPICE simulations to compare the results) on a cache where the peripheral circuitry has a lower V_{th} than that of the drivers and the memory array. As shown in Figure 3.6, the write access time is calculated for a 5% and 10% reduction in V_{th} of the peripheral circuitry. This study assumes a single σ/μ for the entire peripheral circuitry. This is similar to assuming a corner case situation when all the components have similar distribution. While at low temperatures the benefits of reducing the threshold are minimal, at higher temperatures for a 10% reduction in V_{th}, the delay is reduced by nearly 18%. As in the previous study, the median error is below 2% and the speedup achieved for generating each of the nine-sets is in the order of 1.3x10^3.

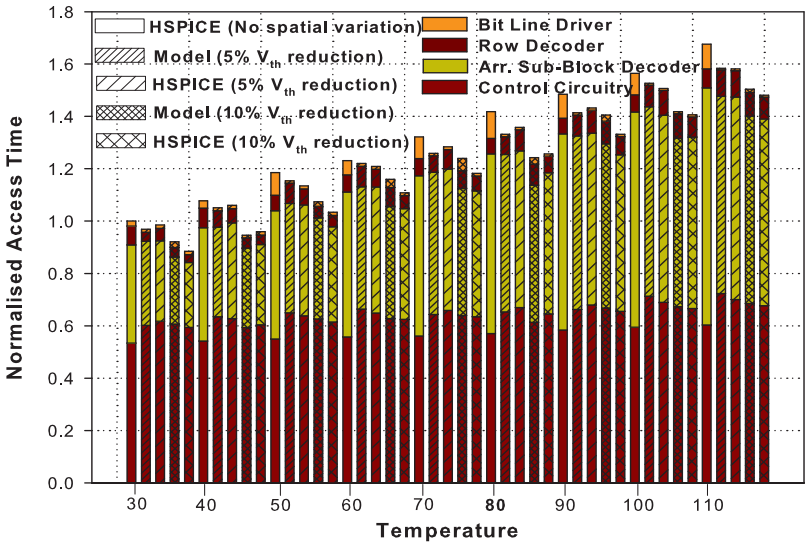


Figure 3.6: Impact of Dual V_{th} Assignment on Cache Access Time

3.8 Energy Estimation in Memories

3.8.1 General Characteristics of a Suitable Model

By having knowledge of the drawbacks of existing models a priori, it is possible to develop one with minimal limitations by careful analysis of the requirements early in the design stage itself. Satisfiable model properties are:

1) **Modest behavior** : The model has to be simple to implement and yet reach accuracy levels close to empirical data. The multi dimension regression problem is simplified by decomposing into lesser dimensions by a combination of a regression and clustering. This can be performed by modelling the dependency of each parameters on the final energy estimates.

2) **Leakage mechanism** : As shown in figure 3.7a, the model should account for different components of static energy namely gate and sub-threshold leakage. During any read/write operation(r/w), the application of a gate voltage to all the cells in the decoded row will contribute significantly to the gate leakage. Even the inactive cells in that row will leak due to the presence of this voltage on the wordline. Similarly, a high on the bitline will affect all cells in a given column irrespective of the selected row due to effect of sub-threshold leakage. For the benefit of the reader, we term the dissipation due to

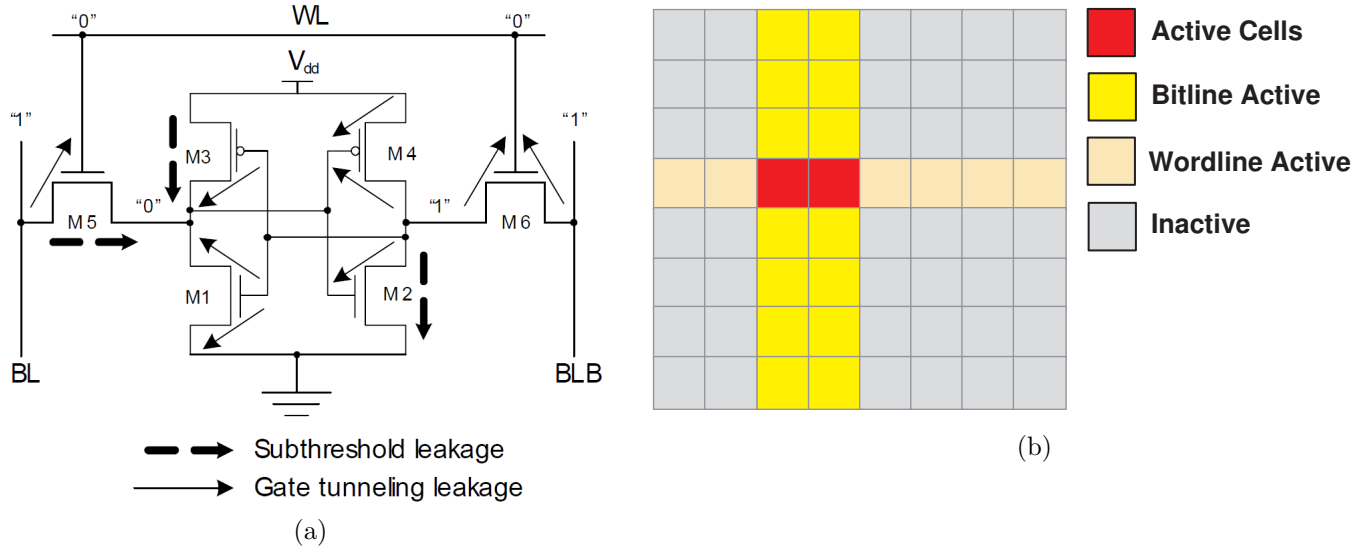
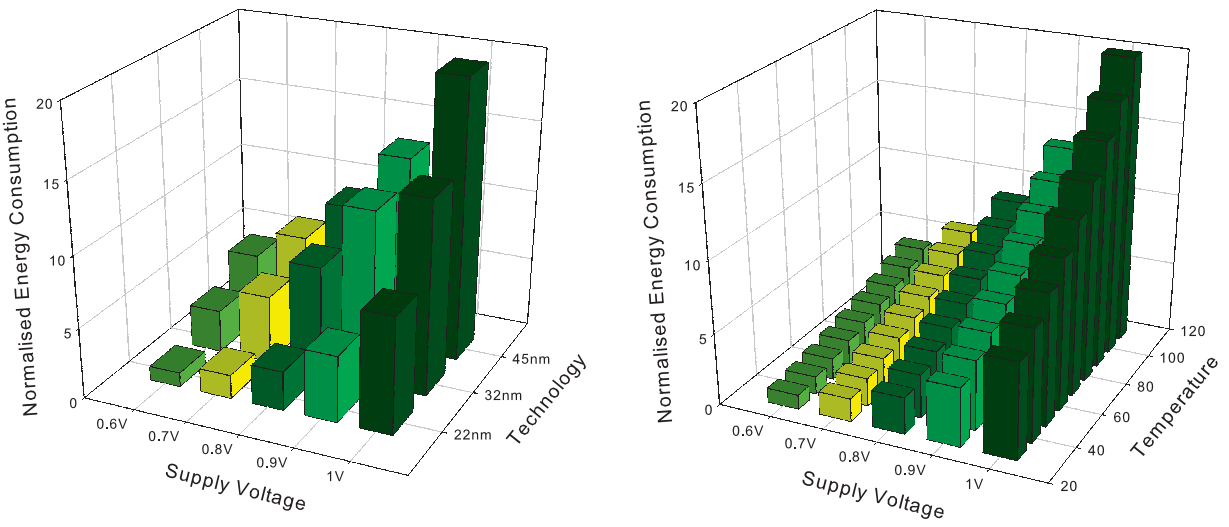


Figure 3.7: Leakage power dissipation in SRAM cells (a) Subthreshold and gate tunnelling leakage in 6T-SRAM (b) Sub-array power dissipation layout

gate-oxide leakage in a given row as *wordline-active* and the leakage due to sub-threshold in 'n' columns as *bitline-active* where 'n' is r/w width. Figure 3.7b shows the different components of energy dissipation during read/write accesses.

3) **Temporal variations** : Often, it is not possible to evaluate the benefits of multiple designs due to the complexity involved in the design implementation. For example, in order to evaluate the thermal stability of a given circuit, the production of faults can be accelerated by enhancing the operating conditions such as temperature and studying the impact of such variations on reliability and performance. As shown in Figure 3.8a, we evaluate the energy consumption for a 32nm memory array design (32KB) with varying supply voltage and temperature, normalized to the value at 0.6V and 30°C. It should be noted in this case that a significant portion of energy consumption is due to leakage and not accounting for temperature variations results in underestimation of leakage energy. While, procedures like burn-in and post-fabrication tuning have the same functionality, they require prototype chips which are seldom available at early design stages. As shown in figure 3.8b, we plot the energy estimates for different technologies for different supply voltages, normalized to the value at 0.6V of 22nm technology node. It is evident from the plot that the potential benefits of scaling supply voltage is a large reduction in the energy consumption and with reducing feature sizes the reduction is greater. Analytical



(a) (b)

Figure 3.8: Normalised memory array energy consumption across a range of supply voltages (a) Comparison across three process generations (45nm, 32nm and 22nm) at 30°C(b) Comparison across a range of temperatures from 20°C-110°C in steps of 20°C in 22nm node.

models can then be used for capturing the impact of scaling of device dimensions and other electrical characteristics.

3.9 MODEST: Model for Energy-Estimation under Spatio-Temporal Variations

In this section, we discuss the modelling technique used for estimation of dynamic and static energy components in a cache. The total energy required for any operation

(read,write or precharge) can be derived as,

$$\begin{aligned} Energy_{total} &= Energy_{dynamic} + Energy_{static} \qquad (3.11)\\ &= (Energy_{switching} + Energy_{shortcircuit} \\ &+ Energy_{glitch}) + (Energy_{leakage} \\ &+ Energy_{pseudo-nmos}) \end{aligned}$$

(3.12)

$$\begin{aligned} Energy_{total} &= V_{dd}\left[\int_{starttransition}^{endidle} I_{supply}\,dt + \int_{time@Val=V_{tn}}^{time@Val=V_{dd}-|V_{tp}|} I_{device}\,dt \right. \\ &+ \left. \int_{startglitch}^{endglitch} I_{input}\,dt\right] \qquad (3.13)\\ &= V_{dd}\left[\left(\int_{starttransition}^{endtransition} I_{supply}\,dt + \int_{startidle}^{endidle} I_{supply}\,dt\right)\right. \\ &+ \left. \int_{startglitch}^{endglitch} I_{input}\,dt + \int_{time@Val=V_{tn}}^{time@Val=V_{dd}-|V_{tp}|} I_{device}\,dt\right] \qquad (3.14) \end{aligned}$$

Based on the premise that energy due to glitches has a minimal impact, we ignore this component of energy. Since we do not adopt a pseudo-nmos design style, it is safe to ignore DC-Standby components as well. Also, switching energy can be estimated accurately only through capacitance extraction. Circuit level simulators like HSPICE do not provide a direct method to estimate switching energy. We compute the integral of the current through the supply over the entire time period and over the time period there is zero-activity (idle period) and we subtract the latter from the former to compute the switching energy. Such a technique has been envisaged in [33]. This method also provides means of computing the short-circuit and static energy components accurately.

$$\begin{aligned} Energy_{dynamic} &= V_{dd}\left[\int_{starttransition}^{endidle} I_{supply}\,dt - \int_{startidle}^{endidle} I_{supply}\,dt\right] \\ &+ V_{dd}\int_{time@Val=V_{tn}}^{time@Val=V_{dd}-|V_{tp}|} I_{device}\,dt \qquad (3.15)\\ Energy_{static} &= V_{dd}\int_{startidle}^{endidle} I_{supply}\,dt \qquad (3.16) \end{aligned}$$

Once the total static energy is estimated, the models proposed in [47] for estimating the different components of static energy is used. The memory array design implemented in [41] is used in this study as well. It is a 32KB array with multiple 1KB data blocks. As we estimate only the energy on a per-activity basis, it is assumed that only one row (within an array sub-block) is active and all other components are inactive (standby). Each of these blocks is organized into 128 columns and 64 rows. Energy is estimated for a period of write followed by precharge and then read. The control circuitry is responsible for generating and synchronizing local and global timing signals. The array and row decoders are designed with dynamic CMOS, the column multiplexer is a pass-transistor based tree design and a differential sense amplifier capable of amplifying very low voltage swings is designed. The total energy for any of the three operations (r/w/p) is derived as the summation of the dynamic & static energy of different blocks at different instants of time as given in equation 3.19.

$$
\begin{aligned}
Cache_{Energy} &= Energy_{active-blk} + Energy_{inact-blk} && (3.17)\\
&= E_{Periphery} + E_{active-cells} \\
&+ E_{wline-active} + E_{bline-active} + E_{control} \\
&+ E_{inactive-cells} + E_{inactive-block} && (3.18)\\
&= (E_{precharge} + E_{column-mux} + E_{Driver} \\
&+ E_{bank-selector}) + E_{active-cells} \\
&+ E_{wline-active} + E_{bline-active} + E_{control} \\
&+ E_{inactive-cells} + Energy_{inactive-block} && (3.19)
\end{aligned}
$$

While size of the array sub-block is fixed, we can generalize the model for any given size by extending the proposal of [67]. We estimate the energy consumption for a cache having an array sub-block with m rows, n columns and p as the read/write width as shown in equation 3.20.

$$
\begin{aligned}
Cache_{Energy} &= E_{precharge} + E_{column-mux} + E_{Driver} \\
&+ E_{Bank-Selector}) + p * E_{(active/cell)} \\
&+ (n-p) * E_{(wordline-active/cell)} \\
&+ (m-1) * p * E_{(bitline-active/cell)} \\
&+ E_{control} + (m-1)(n-p)E_{(inactive/cell)} \\
&+ Energy_{inactive-block} * f(m,n) && (3.20)
\end{aligned}
$$

We simulate the entire cache at 30°C with zero-spatial variation and estimate the energy consumption of each block (only one sample simulated at ambient temperature). As shown in Table 3.3, it can be seen that a considerable portion of the energy is consumed by unused blocks. There are no r/w access to unused blocks and the energy consumed

Component	**Type**	**Consumption(%)**
Control	Static&Dynamic	0.018
Block Decoder	Static&Dynamic	0.044
Row Decoder	Static&Dynamic	0.093
Wordline Active	Static	0.195
Active Cells	Static&Dynamic	0.246
Periphery	Static&Dynamic	0.298
Bitline Active	Static	0.084
Inactive Cells	Static	2.660
Inactive Blocks	Static	95.570

Table 3.3: Energy breakdown of each component in the memory block

corresponds to the static energy consumption. The dynamic energy required for completing the 3 operations is 0.4% and the rest of the energy consumed corresponds to the static energy of the active and inactive blocks. It should be noted that we estimate only the energy for these operations and not the power which is dependent on the activity factor and access patterns. For memory blocks with high utilization factor, dynamic energy consumption will be much higher (realistic scenario). Using the two-cell bitline and wordline models, we can estimate the energy of a cache of any given size. However, the scalability would be limited only to the size of the cache and would not take into account variations of physical parameters and changing operating conditions. In order to estimate the energy in a more realistic scenario, we adopt the quadtree based multi-level spatial-grid technique for capturing spatial variations [4]. The specifics of the model implemented and the total number of samples generated have been previously discussed in this chapter.

Under the effects of the spatio-temporal variations, energy can be modeled as

$$Energy_{variation} = Energy_{nominal} + \delta Energy \tag{3.21}$$

$$\delta Energy \simeq f(\vec{x}, V_{dd}, T) \tag{3.22}$$

where $\vec{x}$ is a vector of spatial parameters, V_{dd} is supply voltage and T, temperature. It is also known that each of these variables is independent and can be varied irrespective

of the value of other parameters. The measured difference in energy due to ST variation, $\delta Energy$, can be approximated using a second order best fit curve as,

$$\delta Energy \simeq \alpha + \sum_{i=1}^{n} x_i * \beta_i + \sum_{i=1}^{n} x_i^2 * \gamma_i \tag{3.23}$$

where α, β & γ are coefficients of the fitted equation. Then Equation 3.23 can be used to estimate the energy deviation due to ST variations. However, estimation of energy at the block level has a major disadvantage. Depending upon the row and column address, accessed and unaccessed cells in the memory array occupy multiple grids at different locations at the bottommost level in the quadtree. Each of these grids have different parameter value. In order to estimate the total energy of the memory array, we need to take into account multiple distributions of the same parameter at different locations on the grid. The process of selecting the right set of parameters by *main effect analysis* is pivotal for the model performance. We exercise the entire procedure by estimating the deviation due to the independent variation of each of the parameters [62]. The difference in energy due to simultaneous deviation of $\vec{x}$, V_{dd} & T from nominal values can be approximated as

$$\delta Energy \simeq f(\vec{x}, V_{dd}, T) - f(\vec{x}_0, V_{dd_{nom}}, T_{nom}) \tag{3.24}$$

Now, by varying each of spatial and temporal parameters separately, Equation 3.24 can be further approximated as,

$$\begin{aligned} \delta Energy \simeq\ & [f((\vec{x}_0, V_{dd_{nom}}, T) - f(\vec{x}_0, V_{dd_{nom}}, T_{nom})] \\ +\ & [f(\vec{x}_0, V_{dd}, T_{nom}) - f(\vec{x}_0, V_{dd_{nom}}, T_{nom})] \\ +\ & [f(\vec{x}, V_{dd_{nom}}, T_{nom}) - f(\vec{x}_0, V_{dd_{nom}}, T_{nom})] \end{aligned} \tag{3.25}$$

The advantage of reducing the total dimensions of the regression equation is to achieve better control over the system. The above approximation enables us to select a minimum number of parameters that have maximum effect on the energy deviation. In other words, by varying few parameters we can observe maximum changes.

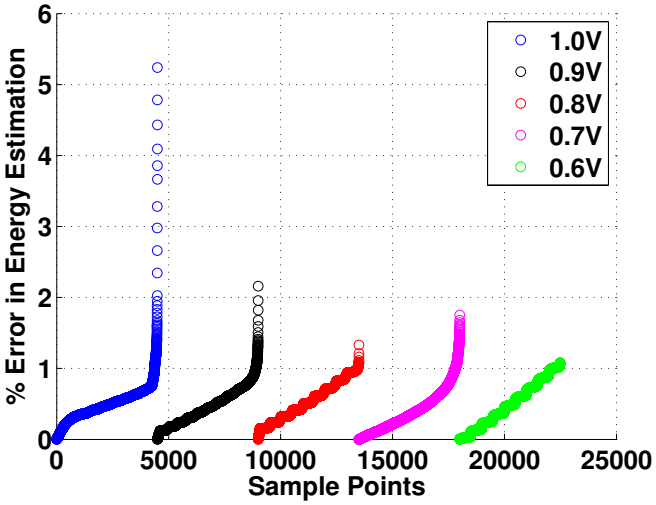


Figure 3.9: Percentage error in estimation of memory array energy between HSPICE simulations and MODEST calculations. For the sake of clarity, the data points (estimation error) have been arranged in increasing order for each voltage level.

3.10 Experimental Results

A 32KB memory array designed with 16nm PTM is simulated using HSPICE on a dedicated server with Dual Core processor running at 3.06GHz with 4GB memory. As explained in [93], it is assumed that the variances of intra-die systematic and random variation to be equal. We set σ for systematic and random variations of V_{th} as 6.4%. Further, in deep sub-micron technologies there is a strong correlation in values of L_{eff} and V_{th}. The systematic and random variations for L_{eff} is derived as 3.2%. The inter-die variations of both parameters is set to an offset value of 3%. While some researchers may argue that the amount of within-die variation in regular structures like memory is far lesser, it is safe to assume that MODEST will work well within the bounds of the total parameter deviation irrespective of the individual inter-die and within-die distributions. We simulate at temperatures from 30°C to 110°C with 10°C steps. As energy has very high dependence on supply voltage, a scaling from 1V to 0.6V is exercised. In order to build a minimum error model, 500 Monte-Carlo samples for each temperature range for 9 ranges of temperature and 5 voltage levels are simulated. The total number of samples generated is 22,500. When comparing with other models having similar functionality [24, 67, 74], error rate of MODEST is by far the least. The computed average median error over the entire set is near 0.5% with a maximum observable error of 7.8%.

3.10.1 Supply Voltage Reduction Analysis

Once the model is defined and verified, we conducted an evaluation of the effects on supply voltage reduction as well as energy-delay study of the 32KB cache under spatio-temporal variations. Figure 3.10 shows the benefits of supply voltage reduction. The

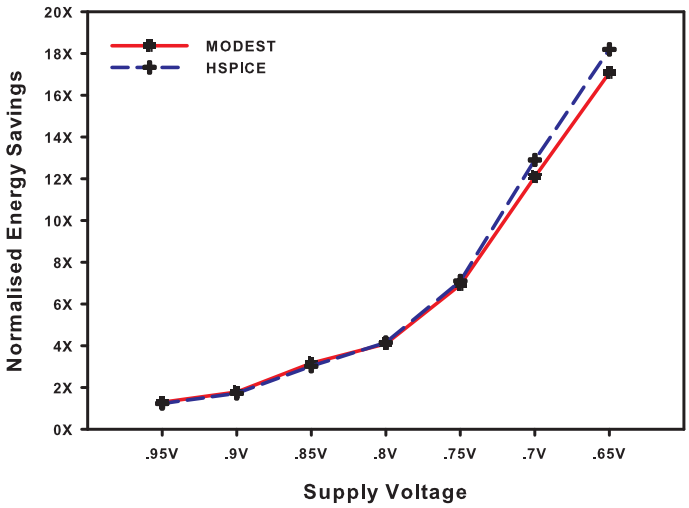


Figure 3.10: Comparison between HSPICE and MODEST for estimates of energy savings in a 32KB memory block with supply voltage scaling.

cache -in this case- was simulated with zero spatial variations and maintained at 80°C to isolate the effect of supply voltage reduction from other factors. The energy estimated was normalized to the value at 1V. Since reducing supply voltage has potential performance loss with reducing (V_{dd} - V_{th}), we run MODEST simultaneously with the delay model proposed in [41]. For a fixed threshold, the overall increase in delay from 1V to 0.6V is of the order of 2.5X. While a performance loss of such extent is intolerable, the aim of the study is to arrive at an optimal delay and energy configuration by means of careful supply voltage selection.

3.10.2 Energy-Delay Analysis

Figure 3.11 shows the impact of spatio-temperature variations on energy variation. The values are normalized to the most energy efficient cache at 30°C operating at 1V. Figure 3.11 shows that difference between minimum and maximum energy consuming caches maintained at the same temperature is around 12X. The difference for the same chip at the

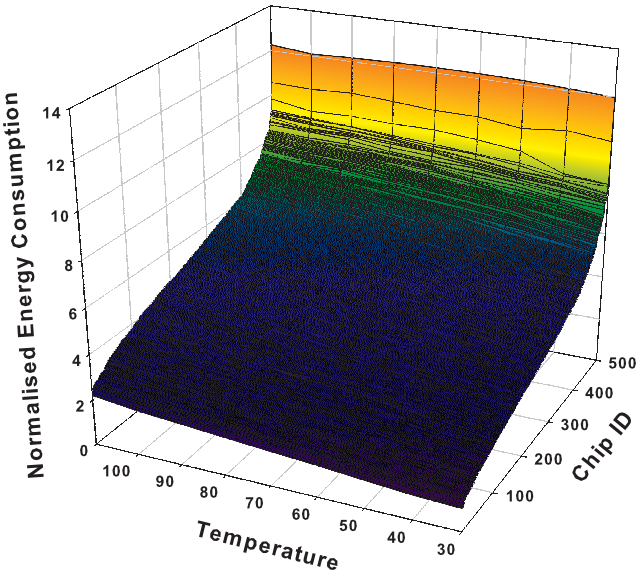


Figure 3.11: Energy variations of a 32KB memory block under the impact of spatial and temperature variations.

ends of the temperature range is around 2.3X. By running MODEST and the delay model simultaneously, we also observe that in the presence of temperature variations the point of convergence of energy and delay need not necessarily correspond to the optimum supply voltage. This is because the energy deviation due to temperature variations is an order of magnitude greater than that of delay. This means that in the absence of spatial variations, the supply voltage for optimum energy-delay product is 0.6V ($30°C$) which is one case that is known intuitively. From Figure 3.12, when considering only temporal variations, it is clear that voltage scaling is a promising technique for energy reduction but at the cost of reduced performance. We also know that temperature variations are very much dependent on hotspot generation and computed workload. Temperature reduction is possible iteratively through dynamic voltage and frequency scaling schemes which reduce the dynamic power but once again at the cost of performance. On the other hand, it is known that leakage is highly dependent on threshold voltage and delay dependent on supply voltage - threshold difference. Dual-V_{th} assignment [101] has been proposed for leakage reduction in energy-constrained areas and improving the delay of delay-critical paths. However due to reducing supply-threshold margins, variation of one (V_{dd} or V_{th}) without the other (V_{th} or V_{dd}) would lead to serious reliability concerns. Figure 3.13 shows the variation of energy and delay due to spatial variations. The values are normalized to

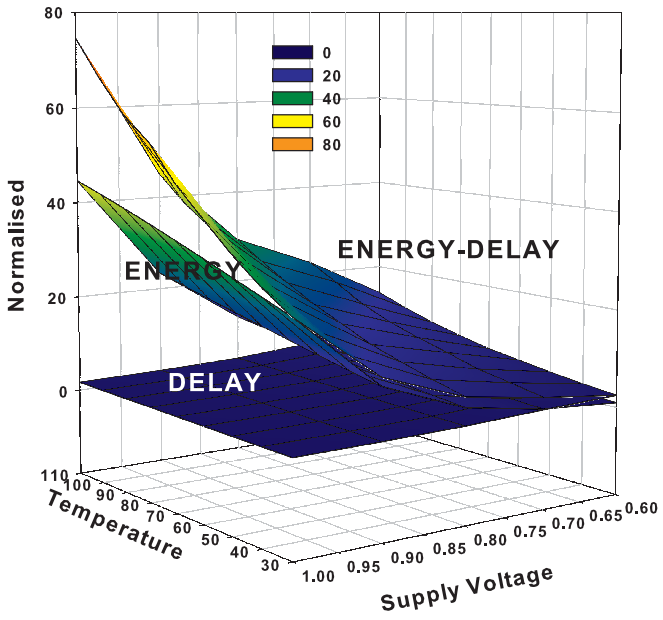


Figure 3.12: Energy-delay variation of a 32KB memory block under the impact of temperature and supply voltage variations.

that of a chip with zero spatial-variation at 30°C. Comparing figure 3.11 and figure 3.13, it can be seen there is a 3.4X difference between the maximum energy consuming dies in both cases which only means there is still enough scope for energy savings under ST variations without significant performance loss. For a fixed supply voltage, reducing V_{th} also decreases delay which is also an observed trend. On the left-hand side of the zero-variation (delay) line are chips with maximum delay and minimum energy. These have higher V_{th} and higher L_{eff} due to spatial variations from manufacturing. In-between the zero-variation(energy) and zero-variation(delay) is the region with ideal delay and ideal energy chips. Intuitively, it can be understood that they have lower L_{eff} in delay critical paths and higher V_{th} in energy-constrained areas and vice-versa. We would like to call this region(dark-grey) as the window of reclamation (WOR). Reclamation of maximum-energy chips can be achieved by minimization of standby supply voltage and active supply voltage assignment [23, 88]. Reclamation of maximum-delay chips can be achieved by increasing the V_{th} of delay-critical paths [101]. This way the optimized energy-delay of all chips subjected to spatial variability can be made closer to the chips with zero-variability.

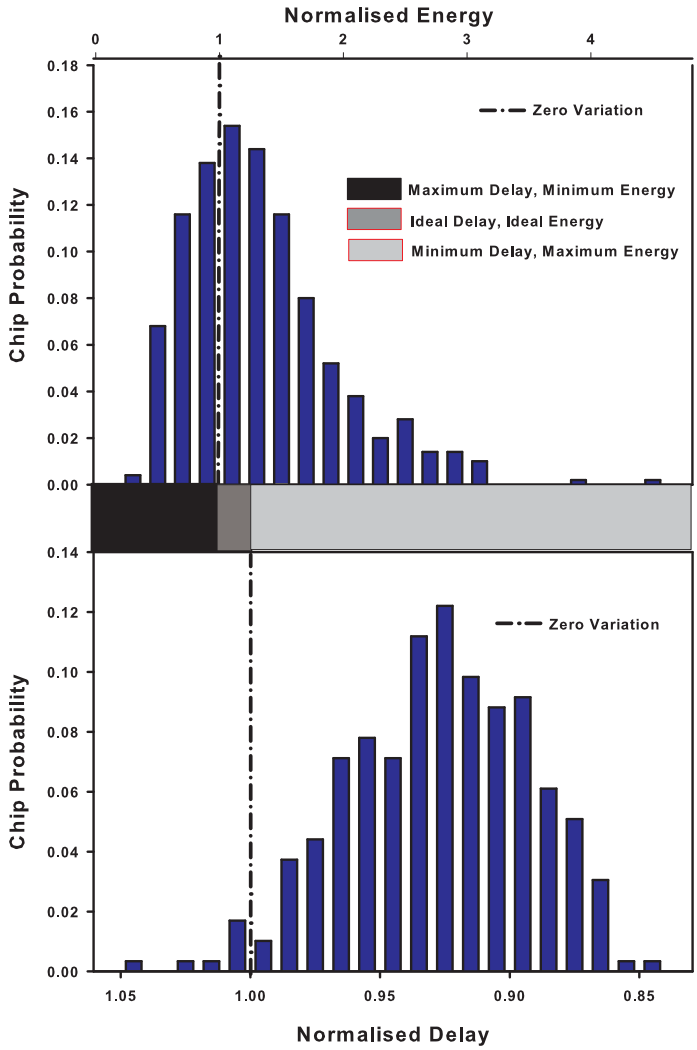


Figure 3.13: Distribution of energy-delay variations of a 32KB memory block under the impact of spatial variations

3.11 Use Case: Fast Per-Chip Energy Optimization through selective Dual V_{th} assignment

In this section, we present a case study describing the wide-usage of the model. We show how through the fast feedback of the model we are able to determine the appropriate V_{th} value for the delay critical paths and the appropriate supply voltage for unused blocks so that the overall energy consumption is reduced at its maximum. This is a study of the potential of the technique when combined with the fast predictions of the model, implementation details are out of the scope of this work. To begin with, we pick a chip

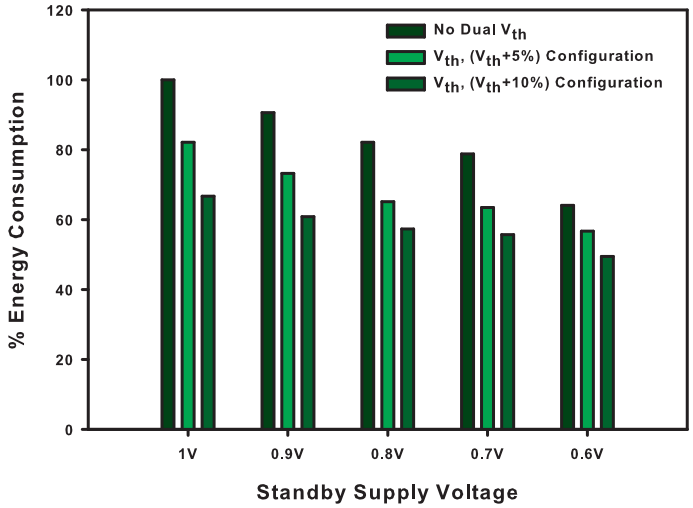


Figure 3.14: Studying the impact of simultaneous dual-V_{th} & standby supply voltage minimization optimization by using MODEST.

with maximum V_{th} and L_{eff} variation (most-leaky) maintained at 80°C and try to bring it into the WOR iteratively through Dual V_{th} assignment and supply voltage minimization. Given an access to the cache, Figure 3.14 shows the energy costs, we decrease the V_{th} of the blocks that may be part of the critical (delay) path. Simultaneously, we reduce the standby supply voltage of unaccessed blocks. The 2 optimization steps have been performed mathematically using the model and verified against implementation in HSPICE. The median error is on the order of the one reported for the model. Reducing the standby voltage of unused blocks by itself reduces the energy by 35% in the best case. Further reduction in energy is possible by increasing the V_{th} of energy-constrained blocks. In the case of 5% increase (V_{th}), it yields a further reduction of 7% in energy, totalling 43% reduction. Increasing the threshold by 10% yields 50.8% reduction in overall energy. When considering the case of V_{th}, V_{th}+10% configuration, the reduction in energy consumption across the range of supply voltages is minimal. Thus for performance and reliability critical applications it is better to scale the standby supply voltage to 80% of the operating voltage. MODEST achieved a speed-up of 750X over HSPICE for performing the dual-optimization with 3 sets of 5 voltage levels.

3.12 Summary

In this chapter, we have discussed an efficient methodology to capture the effects of spatio-temporal variations on propagation delay and energy consumption. The models based on simulation and interpolation are developed and validated against data obtained from simulations with HSPICE. They provide enough scope for specification of technological constraints and can be integrated into early-stage design tools with provisions for circuit-level optimizations as well as analysis. Finally, we have discussed three use-cases of the proposed models to lower power consumption whilst improving performance for variation intolerant memory designs.

4

Dynamic Fine-Grain Body-Biasing (DFGBB)

4.1 Overview

While designers have cognizance of corner-case scenarios at design time, chips are seldom designed for worst-case owing to reduced performance. Dynamic variations of temperature and power to a great extent are dependent on operating conditions and environment. Therefore, it is important to complement design-time optimizations with run-time support that is much needed for calibration of processor structures post-manufacturing. This entails developing on-chip hardware mechanisms that can track variations at run-time and enable cross-layer optimizations based on measured power/performance profiles. In this chapter, we propose a fully digital variation tracking hardware using embedded DRAM cells to monitor changes in cache access latency and leakage. This hardware is interfaced with a dynamic fine-grain body-bias generator that generates an optimal body-bias (reverse or forward) to trade-off power for performance at run-time. The remainder of the chapter is organized as follows: the following section 4.2 discusses the need for a novel post-silicon adaptive mechanism tailored for embedded memories. We also discuss the pitfalls and shortcomings of existing hardware monitoring mechanisms. In section 4.3, we

discuss in detail about the functioning of a regular 3T1D eDRAM cell. We show that this type of memory cell is the right candidate to be used as a canary structure for monitoring variations. Sections 4.4 and 4.5 discuss the hardware discretization architecture and the novel DFGBB mechanism. The experimental results are presented in section 4.6 and the concluding remarks in section 4.7.

4.2 Motivation and Background

Increasing power densities is a major cause of concern for high performance/low power designs. Power is dissipated in the form of heat leading to increased heat densities. With the increase in temperature, leakage power increases exponentially. A recent study has shown that operating temperature of chips can be as high as 90 °C and in some cases as high as 120 °C [68]. These frequent variations in temperature and power can permanently damage the underlying silicon when considering reliability mechanisms such as electromigration, thermal cycling & stress migration [17]. While design-level techniques can be used to meet certain constraints by improving tolerance to process variations, it is impossible to design considering worst-case temperature or power (in-turn leakage & delay) conditions owing to reduced yield and revenues. As a result, relying only on design-level techniques to minimize the impact of process variations will result in conservative designs that reduce the overall tolerance to dynamic variations.

These dynamic variations to a great extent are dependent on the operating conditions and environmental factors. As their frequency of occurrence and nature of existence is very random, it becomes virtually impossible to monitor and measure them. Designing systems with reduced guard bands would greatly improve performance but at the cost of reliable operation. Reliable operation can then be restored at the circuit level by making the system aware of the static variations and also sensing dynamic variations at regular intervals. Emergency conditions like voltage droops or thermal viruses can then be mitigated at run-time by tracking thermal/power profiles and enabling response mechanisms. Such a scheme lacks the ability to monitor high-frequency temperature variations due to the time-delay in sensing on-chip temperatures and off-chip regulation [92]. Further, fixed-point one time calibration (process variations) does not account for the effect of degradation and ageing. Most of the sensors implemented on-chip are very big and require special processes for components like thermistors or platinum resistors [85]. Such special requirements often make their usage near-to-impossible for regular designs.

Recent proposals [25, 83, 98] have suggested that post-silicon adaptivity can be used effectively to improve SRAM yield and also reduce power consumption significantly. Post-silicon adaptivity typically involves measuring low-level circuit parameters (delay & leakage currents) post-manufacturing and provide on-chip mechanisms to enforce circuit-level optimizations to ensure yield targets are met. Body biasing (BB) is one such technique. The threshold voltage of the transistor which is dependent on the body-source potential is modulated to improve performance or reduce power (leakage). In forward body biasing (FBB), the application of a positive bias voltage reduces threshold voltage making transistors faster at the cost of increasing leakage. In reverse body biasing (RBB), a negative voltage increases the threshold making transistors slower and also less leakier. Tolerance to process variations can be improved by utilizing both RBB and FBB and this is called adaptive body biasing (ABB). Based on delay measurements obtained at manufacturing time, bias voltages (either FB or RB) are set permanently for the lifetime of the chip. While this greatly reduces the impact of spatial variations (die-to-die), susceptibility to temporal variations increases. Thus conventional techniques employing ABB result not only in heterogeneous performance across time but across circuit structures as well. Also, techniques for ABB target only performance improvements and do not account for the effects of leakage variation. In order to reap maximum benefits would require fine-grain control of circuit structures by measuring both latency and leakage local to that particular block at run-time and applying an optimal body-bias that trades off power for performance. This is called dynamic fine-grain body biasing (DFGBB) [104]. In essence, this is a 2 step mechanism that requires a sensor like unit (to measure the latency/leakage) to be interfaced with a body-bias control unit for generating an optimal bias voltage based on the measurements.

Adhering to the demands discussed above, the chapter makes the following contributions,

1.) As a first step, we present a novel three-transistor one-diode (3T1D) DRAM-based latency/leakage measurement hardware specially targeted towards memory structures such as register files & caches. By embedding a 3T1D into a regular sram array, we show that each read (or write) to the 3T1D cell will suffer almost the same variation on access power and latency when compared to any sram cell in that array (since it will use the same periphery circuits and the physical variations will be almost identical between the cells due to their proximity). The retention time and access time of the 3T1D are measured to determine the effects of process variation on leakage and latency respectively.

Because of the transient nature of both latency and leakage, the mechanism behaves well in tracking temporal variations.

2.) The measurement hardware is then interfaced with a modified version of the lookup table based adaptive FBB generator [26]. In addition, a hybrid charge pumping circuit is used for generating the negative bias required for RBB [54]. By exploiting the unique access patterns that caches exhibit, active arrays are forward biased while inactive/unused arrays are reverse biased in a very speed effective manner. Not only does this offer enhanced access speeds (forward biasing), tremendous leakage power reduction is made possible by reverse biasing multiple unused arrays. Using a gradient sensing mechanism negates the need for a separate analog-to-digital improving speed tremendously. Further, as both the sensing and regulation is performed on-die, the transition latency between obtaining run-time measurements of the monitor and generating an optimal body-bias is minimal.

4.3 Three Transistor One-Diode (3T1D) eDRAM

Alternatives to 6T/8T SRAM based memories have been researched diligently for want of increased memory density and lower vulnerability to variations. One such proposal is the 3T1D cell proposed by Luk *et al.* [72]. As shown in figure 5.1, the capacitor-less DRAM cell stores the data using a gated diode that is tied to the read-wordline. The 3T1D unlike 1T DRAM memory provides non-destructive reads and access speeds comparable to that of standard SRAM cells. When compared to regular SRAM cells, the transistors of the 3T1D can be asymmetrical in strength. This has a 2 fold advantage: Primarily, process variations causing device mismatch are likely to cause less failures to the cell [70]. Secondly, it improves the overall stability making it radiation hardened. Data is written into the cell by raising the write-wordline high and charging the bitline. The voltage level corresponding to a value '1' at the storage node is largely dependent on the strength (V_{th}) of transistor T1. The voltage level at the storage node is degraded and roughly about $0.6V_{dd}$. T1 with large V_{th} would further degrade the voltage resulting in lower retention time. This can be avoided by increasing the threshold of the write driver [72]. The read operation is initiated by precharging (to V_{dd}) the read-bitline and strobing the read-wordline. The retention or storage time of the cell can further be increased by holding the read-wordline at a negative voltage during idle state. Due to boosting by T2, the value at the storage node increases temporarily close to the value of write-voltage. As

the only path for sub-threshold leakage is through read-wordline tied to the gated diode, it can be reduced by holding the read-wordline at negative voltage (say -0.2V) when the cell is in its idle state (no access). This has shown to increase the retention rate by as much as 40X [72].

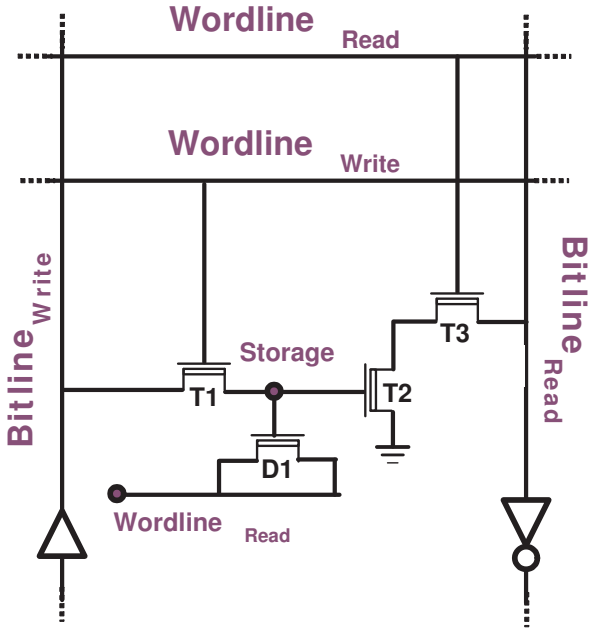


Figure 4.1: Schematic of the 3T1D eDRAM memory cell

4.3.1 Retention and access time

For the sake of comparative study, a 3T1D and 8T-sram are embedded into the same array along the same line as shown in figure 4.2. The sram cell is a 1R/1W ported cell found in register files. With minor modification to the column periphery, the 3T1D can be embedded into a conventional 6T-sram (dual-ended) based array. The access latency of both cells is measured independently and normalized to value at 30 °C. Looking at figure 4.3, only at high temperatures, the 8T cell is more prone to performance loss when compared to the 3T1D. As opposed to regular 8T cells, the 3T1D are designed for single ended sensing. This combined with T2's boosting action provides very high read speeds even at high temperatures. This validates the fact that both 6T and 3T1D have similar access latency and in some sense mimic each other's functional behavior. While the 8T cell could already be used to measure access latency, 3T1D provides an extra measurable parameter called retention time. The retention time of the 3T1D is defined

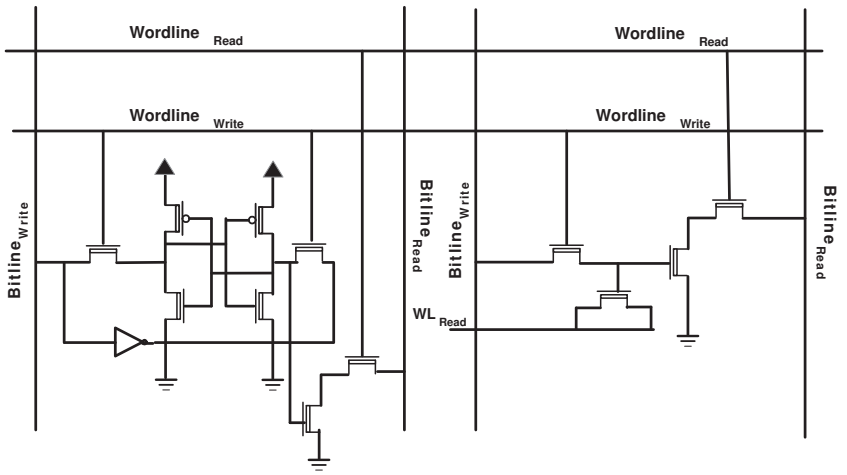


Figure 4.2: 3T1D-DRAM embedded with a 8T-SRAM

as the time taken for the voltage at the storage node to decay past $V_{dd}/4$. In [56], it is shown that the leakage through the cell is directly proportional to the retention (decay) time of the cell. In other words as leakage increases, the retention time decreases and vice-versa. Figure 4.4 shows the measured retention time for 500 cache samples simulated for spatial-temporal variability. The retention time is normalized to the lowest retention time at 110 °C. The samples are organized in order of reducing magnitude of retention time. Retention time with zero-variability at 30 °C is found to be 9.3μs. Under the presence of process variations, operating at 30 °C, the retention can be as high as 34.2μs or as low as 5μs. Because of the exponential relationship between leakage and temperature, the retention time can be as low 980ns at 110 °C under worst case process variations. It should be clear from the above argument that both access and retention time of the 3T1D are an important figure of merit that can be measured to reflect the SRAM's latency and leakage power variation under the effects of spatio-temporal variability.

4.3.2 Simulation Parameters

We interleave the 3T1D cells in the memory arrays to use them as latency and leakage sensors as we will explain in short. The arrays keep the original SRAM cells for program execution. Each array is organized into 128 columns by 64 rows with a 32 bit read-out. Due to area constraints, the decoders are designed with dynamic cmos and column multiplexer is tree-like design. 500 samples of the cache are simulated on HSPICE with 45nm PTM [1]. We modify only the periphery of one column to accommodate single-

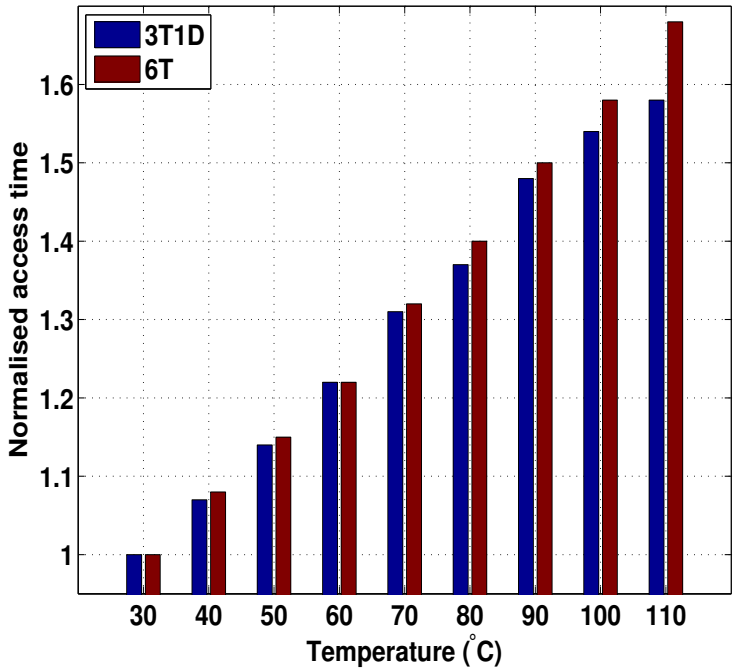


Figure 4.3: Comparison of the normalized access times of 3T1D eDRAM and 6T SRAM cell across a range of temperatures.

ended 3T1D-dram cell in a regular dual-ended 6T-sram array. The area associated with a single 3T1D cell for 45nm technology is approximately $0.45\mu m^2$ [66]. Assuming there is one 3T1D cell per row of sram then the associated area & energy overhead is estimated as 0.31% and 0.78% respectively. Because random variations are known to affect sram cells more than systematic variations and there is no definite way of tracking random variations, 1 3T1D per whole array is more than sufficient. We use this configuration for the remainder of the study. The σ for systematic and random variation of V_{th} is 6.4%. The systematic and random variations of L_{eff} is derived as 3.2%. Inter-die variations of both parameters is set to an offset value of 3%. While measuring temperature directly as a variable parameter is not within the scope of this work, the gradient exhibited in measured access latency & leakage across temperatures can be used to detect temperature variations.

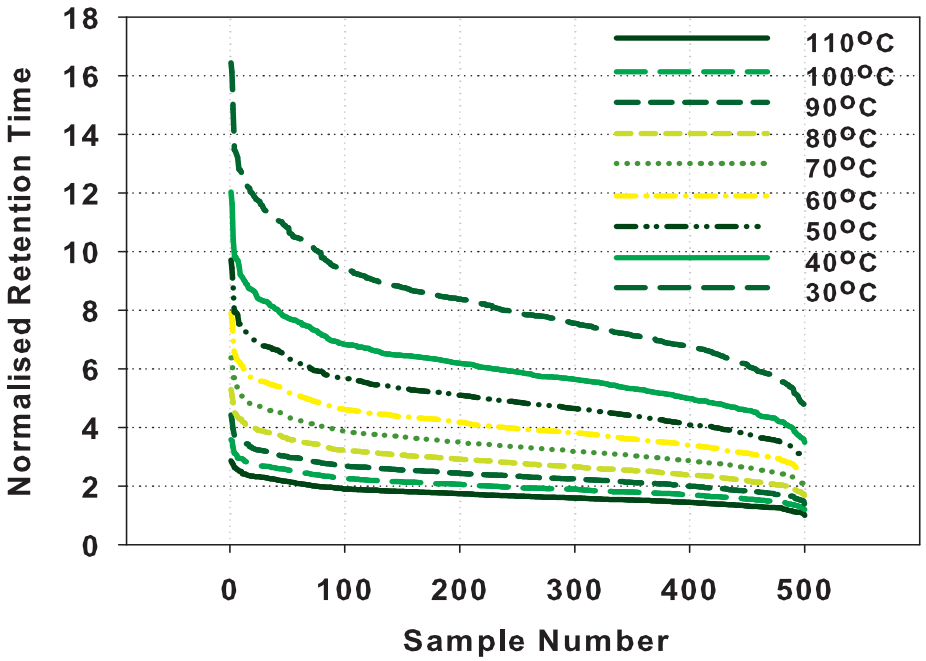


Figure 4.4: Comparison of the retention times of a 3T1D cell across a range of temperatures under the impact of spatial variations of process parameters

4.4 Latency/Leakage Measurement

4.4.1 Run-Time Classification of Memory Arrays

The purpose of classifying cache arrays based on latency/leakage profiles at run-time is very much like calibration. Calibration enables to rectify any deviations that arise out of manufacturing or during the lifetime of the chip (i.e. degradation). From a statistical standpoint, these deviations can be on either side of the mean value obtained at design time. Most often run-time circuit level optimizations like body & source biasing, supply voltage minimization that have been proposed for leakage minimization, are enforced without this available information [15, 69]. Such optimizations resulting from holistic procedures have been enforced across varying chips yielding non-uniform benefits. For any optimization that needs to extract maximum benefits using the available leakage/latency measurements, the granularity of the classification has to be very fine as shown in figure 4.5 (classification & measurement are used in a interchangeable fashion). A very high access time and low retention translates directly to high access latency and significantly high leakage power. This is one of the non-ideal cases that we would like to avoid at

any cost. For the sake of simplicity, we would like to call each discrete combination of measured leakage and latency as a *bin*. The nomenclature used (min,low,high,max) is specific to our scheme and is not representative of the actual degree of separation. It

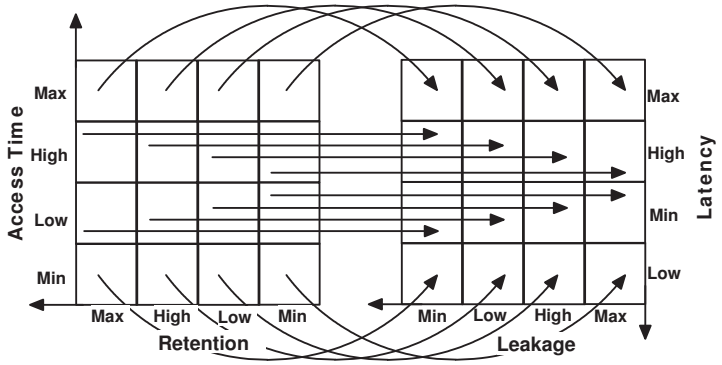


Figure 4.5: Discrete Classification based on Latency/Leakage

is well established that both latency and leakage are transient and by generating this table-based data, on-die registers can be frequently updated with this information to be made available for cross-layer optimizations. By making the latency/leakage bounds more tight during classification, circuit optimizations can have more fine-grain control by having better cognizance of the power/performance status of each array.

4.4.2 Discretization Architecture

As temperature has a more observable effect on the retention time when compared to access time, we begin with the classification based on leakage. Theoretically, to measure the retention time, we could use a simple delay-to-pulse circuitry to count the number of cycles the voltage corresponding to a '1' at the storage node of the 3T1D takes to decay by sending a continuous stream read requests one after another and waiting till the voltage degrades completely. This has few disadvantages. Firstly, it was earlier shown that the retention time is of the order of μs. This means that it would require hundreds of thousands of cycles for a counter with a low pulse-width clock to complete the operation. Response mechanisms typically are expected to have very fast response in the order of thousands of cycles. As a simple rule of thumb, faster the response, greater the benefits. Further, lower the leakage, higher the retention and so longer the time it takes it complete the operation. This is furthered by the decaying of the voltage at the storage node which makes subsequent read accesses slower. During decaying action of the diode

in the forward biased mode, initial period of decay is very fast. With further reduction in the voltage at the storage node, the decaying speed reduces as a result of increasing diode resistance. Thus it is sufficient to measure the drop in voltage for the first few hundred nano seconds rather than having to wait for the voltage at the storage node to decay completely. The decaying behavior of the storage node is replicated at the output of the sense amplifier. The measuring hardware just has to convert the time the output of the sense amplifier is high into something measurable on-chip. Any scheme that involves a delay-to-pulse circuitry can generate a clock cycle for every period that the output of the sense amplifier is held high [22]. Our proposed leakage-bin classification architecture

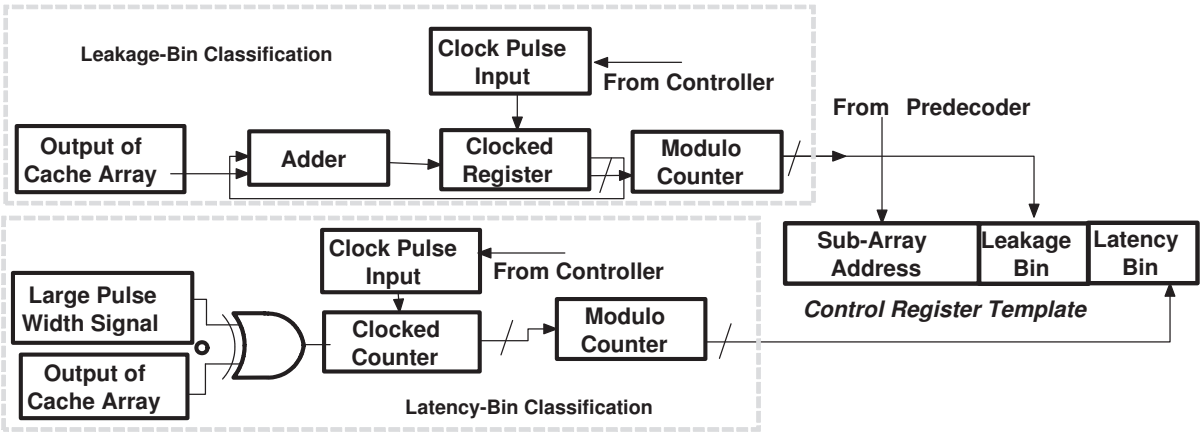


Figure 4.6: Hardware Based Leakage/Latency Bin Classification based on Retention/Access Time

is shown in Figure 4.6. The output of the cache array (sense amplifier) is linked to an adder which has the feedback of a clocked register. The register is clocked at a frequency bounded by the pulse width of the minimum difference between any 2 adjacent bins. The total number of bins to be used for classification can be decided at design-time. With reducing number of bins, the bounds of leakage (minimum and maximum value) within which an array is placed into a bin is well spread. In other words, with minimum number of bins for classification, those arrays that have very different leakage profiles (or retention times) have every chance of being placed in the same bin. The bin selection procedure is initiated by writing a 1 to a 3T1D cell and signaling a read access and constantly strobing the read-wordline high. The output of the sense-amplifier after a given period begins to decay. As long as the output of the sense-amplifier is high enough to signal a 1, the adder increments the value of register by a 1 at the clock rate. The register is incremented at a predetermined frequency whose clock period is low enough to make sure adjacent bins exhibit a difference of at least 1 cycle as shown in figure 4.7. It is clearly observable from

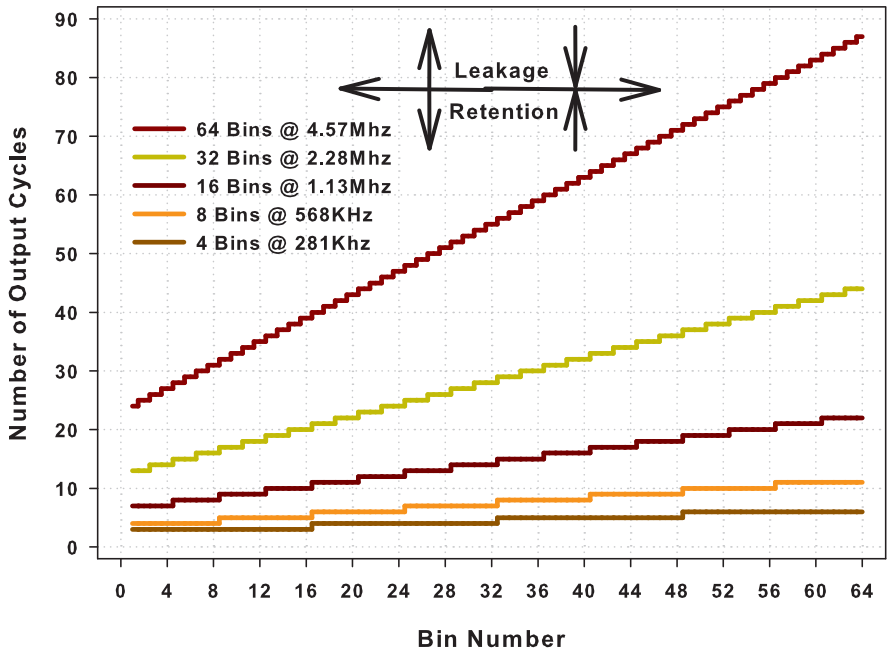


Figure 4.7: Number of Output Cycles Vs Power Bin Number. (The number of output cycles (modulo) target number of bins) is the value that is written into the control register

figure 4.7 that the cycle count increases with the bin number. This is in direct relation to the fact that reducing leakage along bin number corresponds to increasing retention times which is reflected by the increase in cycle count along the x-axis. It can be seen that the clock frequency used for 64-bin classification is 16X higher than the frequency used for 4-bin classification. This exhibits a linear relationship between the number of bins to be used and the frequency of the clock. Some researchers may argue that it is not a viable option to have a frequency divider for bin-classification purposes. The simplest solution would be to design for a frequency that would cater to the maximum number of bins, for instance 64. For classification based on lesser number of bins, say 8, grouping is performed by placement of arrays into bins which are multiples of 8. This is made possible by using a modulo counter. This way in a 8-bin classification using 64-bins, bin-8 (in 64-bin) would represent bin-1 (in 8-bin) and bin-16 represents bin-2 and so on. This can be seen in figure 4.7 where for a 4-bin classification using 64-bins, all the arrays that have their bin-number lower than 16 exhibit 1 cycle output (lowest retention/highest leakage), and all those between 16-32 have 2 cycles output and so on.

In order to classify based on latency, we determine the time to read access the 3T1D cell (corresponding to the critical path delay). Under the impact of spatial variability, for a set of 500 samples, the access times have been found to vary between 14-18% when maintained at a fixed ambient-temperature. This translates to a difference of about 400ps between the slowest and fastest arrays. In effect, the separation between adjacent bins can be as low as 6ps. As a result, for 64-bin classification, even multi-GHz frequencies cannot produce a clock whose period is 6ps. Thus for multi-MHz frequencies, the maximum target-number of bins is 4. The procedure to measure delay in terms of cycle count is identical to the proposal in [22]. A signal with a very large pulse width is XOR'ed with the output of the cache array. The clocked counter starts incrementing on enabling the control signal to initiate reading a '1' from the 3T1D. As long as the output of the sense-amplifier is 0 and large-pulse width signal high, the counter is incremented for every cycle of the input clock. As soon as the output of the sense-amplifier reaches a high, the counting stops.

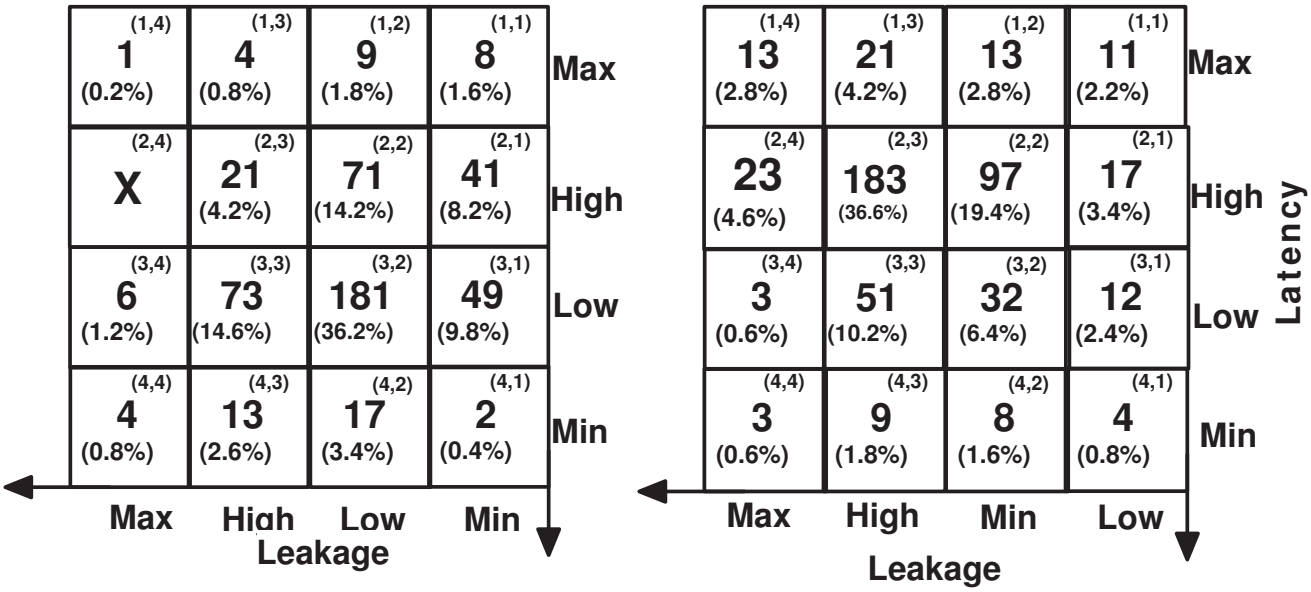


Figure 4.8: (a) Power/Performance Binning using Shadow Cells at ambient temperature (b) Power/Performance Binning using Shadow Cells at 110°C temperature

For a fixed supply voltage of 1V and 500 cache samples, the classification was performed for 4 binning levels of latency and leakage. A cache is placed into a respective bin after measuring the retention and access time of the 3T1D embedded in each array and picking the slowest and leakiest. It is strange that no cache has been placed in the high latency, maximum leakage bin as shown in Figure 4.8a. This phenomenon is characteristic to our single frequency grouped-levels binning methodology. As the 4-bins have been approximated by scaling the 64-bin classification, the bounds of each bin are loose resulting in misplacement of high latency, maximum leakage caches across the 3 immediate neighboring bins along its cartesian co-ordinates. By re-running the simulations adjusting the input-pulse frequency specific to 4-bin classification, a considerable number of caches were categorized into the high latency, maximum leakage bin. Assuming we consider all caches that have latency and leakage greater than *high* as yield loss, hardly 50% of caches are accepted. Further the presented yield estimates in figure 4.8a hold true only when the cache is operating at nominal temperatures. Common phenomena such as sudden temperature shoot-ups can result in the performance going from high to low and in some cases to minimum. The problem is compounded by increasing leakage with temperature. It is clearly observable from Figure 4.8a that number of caches placed in the low-latency low-leakage bin at 30 °C shifts diametrically to the high-latency high-leakage bin at 110 °C as shown in figure 4.8b. This results in yield going from bad to worse. From the above results it is clear that, we have been successful in translating the logical relation presented in figure 4.5 to a hardware based methodology. In the next section, we will discuss as to how we can exploit these available measurements to improve the overall cache yield.

4.5 Applying Fine Grain Body Biasing

It was shown in [60] that reduction in leakage power is possible by optimizing the 6T-sram cell at design-time for high V_{th} and applying a large forward body bias at run-time to compensate for the increased latency. In other words, the array is purposefully designed for high latency (and low leakage) and is made to run faster during operation. This would mean that a large FBB is applied irrespective of whether the array meets the required timing or not. From a statistical standpoint, both latency/leakage can be on either side of target design value as a result of process variations. Hence no forward biasing is required for those arrays that already meet both leakage and latency targets.

It is this very non-determinism that we would like to exploit in order to generate optimal bias voltages dependent on the actual latency/leakage measured. The only downside of body biasing is that it requires separate n-wells of whole arrays to be isolated from each other to improve immunity to substrate biasing. Modern day triple well processes offer this option at an increased area overhead. Techniques to improve immunity to substrate noise include - providing low overhead control circuits to bias wells individually [46] or routing bias lines through upper layer metals [60].

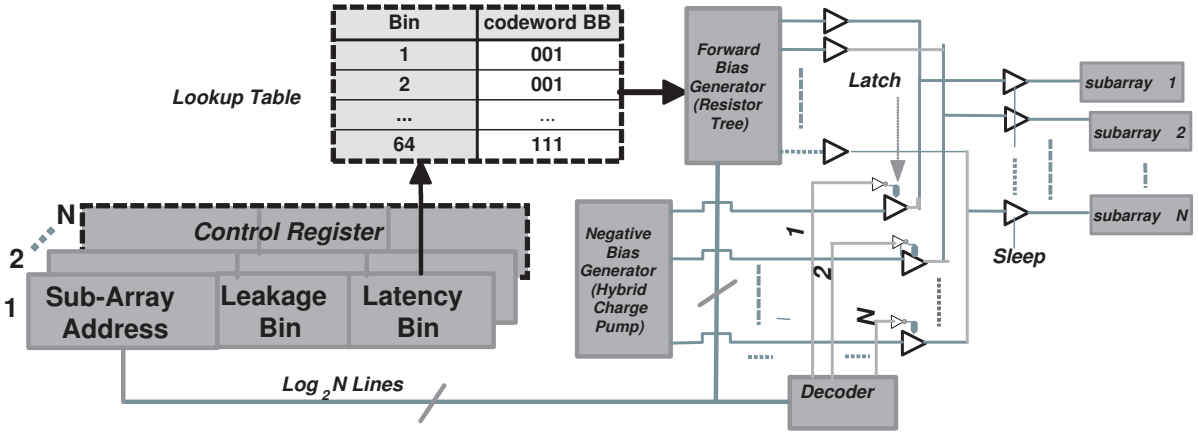


Figure 4.9: Fine-Grain Body Bias Generator for Caches

We propose to use a modified version of the lookup table-based adaptive forward body biasing mechanism[26]. A global decoder inside a cache receives the address of the block to be accessed from the address buffer. Without loss of generality, we assume that the DFGBB generator receives this address at the same time. The index bits corresponding to the array address are decoded and obtained prior to the access. As shown in figure 4.9, the latency bin of the to-be accessed array is sent to the Look-up Table (LUT) by comparing the address received from the global decoder to the address field of all control registers. The latency bin is referenced inside the LUT to obtain a code-word that is sent to the FBB generator. This LUT is defined at design time and can be stored in a small on-chip EEPROM. This code-word corresponds to the lowest forward bias voltage that ensures the desired cache access latency is met. The code-words in the figure are only indicative and do not represent the actual bias voltages. The FBB generator consists of four components - decoder, level shifter, demultiplexer & resistor tree (placed in same order). The resistor tree is used for generating the forward bias voltages. The resistor tree consists of a series of transistors connected together acting as a potential divider. The number of transistors divide the range (V_{ddhigh}-V_{ddlow}) into intermediate voltages.

In our case, we assume a maximum range of 500mV. We have used 20 transistors in our design, and connected switches to the 4th,8th,12th and 16th transistors to generate intermediate voltages of 0.1,0.2,0.3 & 0.4V respectively. The decoders are used to select the correct combination of switches to generate the appropriate FBB. The generated bias voltages are routed to the correct array using the demultiplexer.

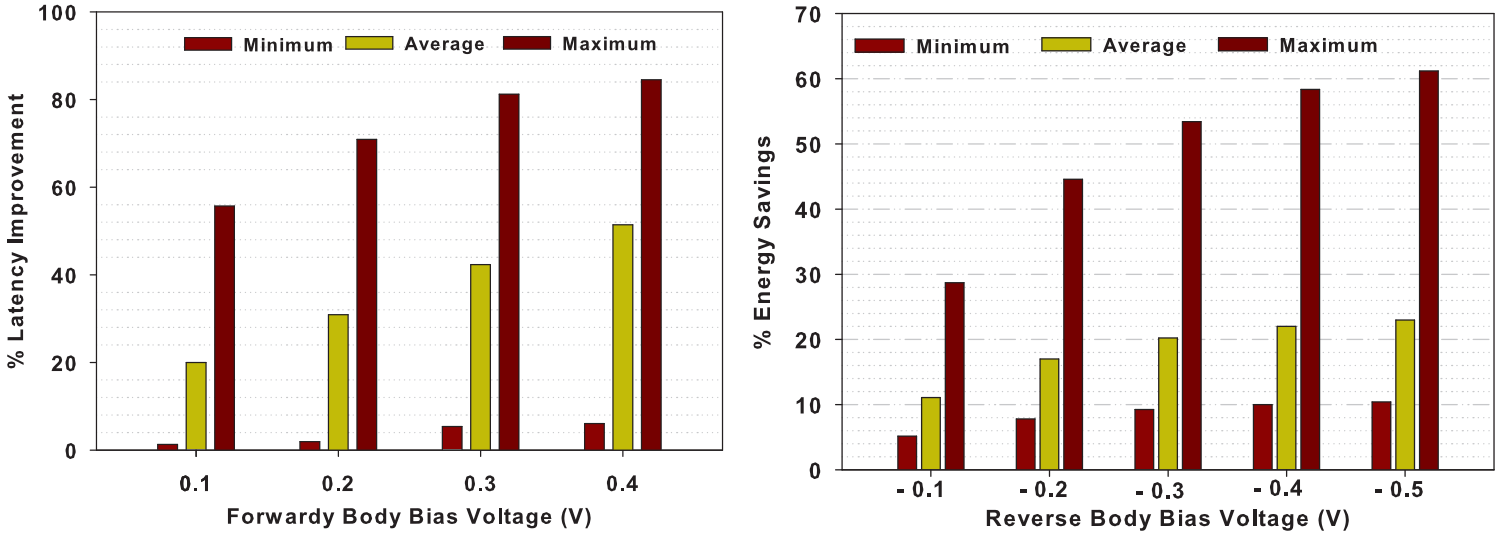
Each of the N arrays require 3 amplifiers - one each for FBB & RBB to boost the body voltage to a level sufficient to bias the entire array and one for enabling/disabling sleep mode. A hybrid charge pump is used to generate negative bias for RB biasing inactive arrays. The amplifiers for routing the RBB voltage are enabled based on the address of the array to be accessed. The address of the to-be accessed array is decoded and all the output lines of the decoder act as the enable signal for the RBB amplifiers. Only one of output lines is high (corresponding to the array to be accessed) and the remaining are low. Thus by inverting these lines, the RBB amplifiers are enabled. This ensures that the FBB and RBB amplifiers corresponding to a given array operate in a mutually exclusive manner reducing the transition latency (from RBB to FBB and vice-versa)tremendously. It was shown in [60] that if an array is accessed in a given cycle then it is likely to be accessed in the immediate next cycle and those that are idle are expected to remain idle for a considerable amount of time. This phenomena called temporal locality of reference, can be exploited to forward bias those arrays that are currently being accessed and reverse bias those that are idle. It also eliminates the need to regenerate the same FBB voltage on per-cycle basis by constantly referencing the LUT. As a result, RBB generator needs to be aware of the idle arrays for a large number of cycles. Because it receives the address of the to-be accessed array only once, the state of idle arrays needs to be stored. An extra latch is provided to store the state of the inactive array for enabling/disabling RBB mode. In addition to hiding the transition latency in a very time-effective manner, the transition energy involved in switching between RBB and FBB is also reduced significantly.

4.6 Experimental Results

In accordance with the analysis presented in [104], BB voltages range from a minimum RBB of -500mV to a maximum FBB of 400mV. On a per-access basis, only one array is active and the remaining are inactive. This is the closest representation of the actual architectural state of the entire cache.

4.6.1 Leakage & Latency Reduction

Looking at figures 4.10a and 4.10b, both leakage & latency are a very strong function of the bias voltages. Energy is calculated after accounting for the energy consumed by bias generators and the energy lost due to active mode forward biasing. The minimum and maximum values correspond to the lowest & highest improvements obtained for one array among all arrays (WID) of a cache across all samples (D2D). The average is the lowest of the arithmetic mean obtained for all arrays of a cache (WID) across all samples (D2D). It can be seen that the minimum average savings in energy is 12% (-0.1V) and the maximum is 24% (-0.5V). For -0.3V it is 20% and the improvement in energy savings is minimal for voltages above. This is because process variations are known to affect multiple transistor parameters (threshold, oxide thickness, effective channel length) which in-turn affect leakage and threshold voltage is the only parameter that can be dynamically altered with body biasing. By providing a LUT based RBB generator, the leakage-bin field can be used to determine appropriate reverse bias voltages for further energy reduction. Looking at the results of latency improvements in figure 4.10a, it can



(a) (b)

Figure 4.10: (a) Percentage Energy Savings as a function of the reverse body voltage & (b) Percentage Latency Improvement as a function of forward body voltage. The bars represent savings when compared to ZBB

be seen that there is a large discrepancy between minimum and maximum values. This is because, only the SRAM array is BB'ed and the latency is calculated for the entire access path constituting predecoders, row decoder, column multiplexer, sense amplifiers, wordline and write drivers that are not BB'ed. Techniques like dual V_{dd} & dual V_{th} can then be employed to reduce the impact of process variations on periphery [42, 51]. Because our mechanism can alter the forward body voltage based on the measured latency, we can expect maximum latency reduction even under worst-case process variations.

4.6.2 Evaluating Yield

Heuristics for estimating parametric yield suggest that caches which fail to meet the latency constraint (maximum allowed access latency under process variations) can be considered as yield loss. In sub-65nm designs, as leakage can play a very important role, it was shown that in addition to considering a latency cut-off, caches that consume leakage power greater than 3μ may also be rejected [82]. Adopting the above heuristics,

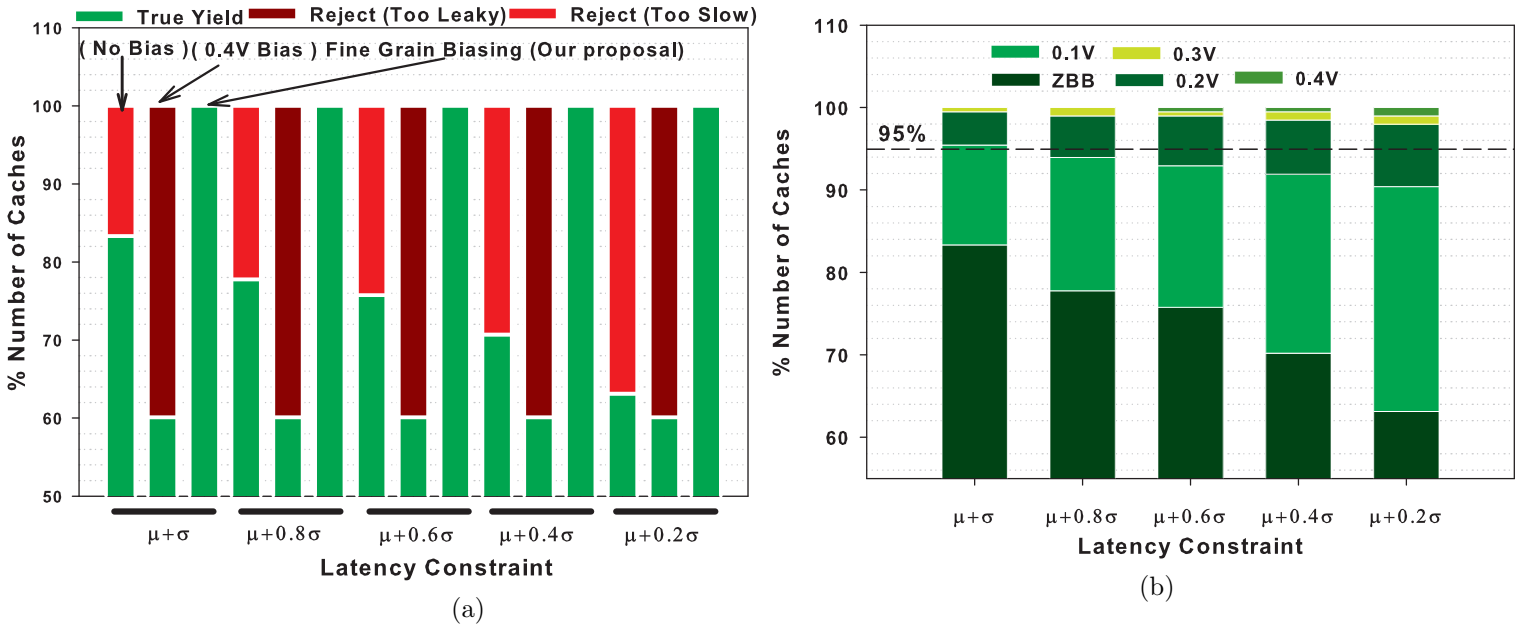


Figure 4.11: (a) Yield estimated for ZBB, constant FBB and DFGBB as a function of latency constraints & (b) Number and amount of FB Voltages required for 100% yield.

we determine the parametric yield for three different cases - zero body biasing (ZBB), only forward body biasing active arrays with one voltage [60] & our proposal - fine-grain body biasing each array. We assume that all idle arrays are reverse biased at -0.3V in our proposal. The yield is determined for multiple latency constraints and for one leakage constraint of 3μ. A cache is considered functionally unworthy (yield loss) if more than 3 arrays fail to meet the constraints. It can be seen from figure 4.11a, that under no body biasing, the yield reduces from 82% to 60% for tighter latency constraints. The yield loss is only as a result of arrays failing to meet latency constraints and not because of leakage constraints. The yield for forward biasing with one voltage is constant at 60% in all cases. While all arrays clear the latency cut-off because of lowering the threshold, some arrays fail to meet the leakage cut-off resulting in yield loss. For all cases of latency constraint, the yield is more than 98% in our case. This is mainly because of 2 factors. Unlike adaptive body biasing where we decide to either use RBB or FBB, here we use both in a time shared manner. The selected forward bias voltage is the minimum voltage for which the latency cut-off is met as shown in figure 4.11b. This is to ensure that active leakage power is further reduced. For the case when latency constraint is $\mu+\sigma$, 82% of caches do not need any bias. By providing a bias generator with just 1 voltage of 0.1V, the yield can be significantly improved to 95%. With further increase in the number of available bias voltages, the increase in yield is minimal. The yield is actually not a function of the number of bias voltages but a function of the minimum & maximum bias voltage that ensure all arrays meet both latency & leakage targets. By increasing the number of available voltages (by reducing the intermediate steps), more fine-grain control can be achieved. For high performance designs, the latency constraints are between $\mu+0.2\sigma$ & $\mu+0.4\sigma$ and it is clearly evident that there are caches that require both 0.3V and 0.4V FB voltages.

4.7 Summary

In this chapter, we have primarily shown the importance of complementing design-level solutions with run-time optimizations. By measuring the time to access & retention time of the 3T1D, it was shown that the sram arrays can be classified based on run-time leakage/latency measurements. Then a lookup table based adaptive fine grain body biasing mechanism utilizes this measurement, to generate an optimal bias. While active arrays are forward biased to improve performance, inactive arrays are reverse bias to

reduce leakage. The experimental results show that our technique on average improves access latency & reduces leakage energy by 18% & 23% respectively. The adaptability to temporal changes ensures cache performance & power consumption over the lifetime of the chip is consistent.

5

Retention Enhancement and Improved Radiation Tolerance In Embedded DRAM

5.1 Overview

In the last chapter, we introduced embedded DRAM (eDRAM) cells that can be considered as a potential alternative to current day SRAM solutions. 3T1D-DRAM solutions can be used in L1 caches by modifying existing architecture-level access policies to suit the behavioural pattern exhibited by such structures. However, one major concern is the large spread in retention time across cells that eventually mitigate the various advantages they offer over regular SRAM cells. In this chapter, we propose a novel variation-tolerant 4T-DRAM cell that offers enhanced retention when compared to similar alternatives. We then detail the importance of single-event upsets in such cells and investigate the radiation-tolerance of our cell compared to 3T1D cells. The remainder of the chapter is organized as follows: section 5.2 details the need for an alternative to SRAM based memories and the shortcomings of existing eDRAM variants. We then revisit the 3T1D operation from a retention time perspective in section 5.3. In section 5.4, the novel 4T

DRAM is introduced and studied in detail. The impact of retention time and access time of both cells under variations is studied in section 5.5. In section 5.6, we present an analysis on the effect of single event upsets due to neutron strikes in eDRAM cells. The summary is presented in section 5.7.

5.2 Motivation and Background

Given the tight area-constraints of on-chip memories, it becomes a daunting task to design 6T-SRAM cells optimized for a wide operating range. 8T/10T cells are alternative design styles where read/write (r/w) paths are decoupled allowing designers to focus on r/w margins separately thereby improving reliability. While on one hand cell writability and readability is significantly improved, a new phenomenon known as *half-select problem* becomes inevitable. Columns of un-selected cells experience a mild-disturbance during r/w accesses increasing the probability of bit-flips. Such effects can be negated by either designing memory arrays with hierarchical wordlines that isolate accessed columns from unaccessed ones or employing complex write-back schemes where a read operation is always followed by a write-data-back operation [59]. The consequent improvement in parametric yield comes at a cost of increased dynamic power consumption or reduced memory area-efficiency. Memory area-efficiency can be defined as the factor of memory-array area excluding circuits used only for improving memory-array yield.

Moving away from SRAM technology in light of technology scaling issues, semiconductor companies have started embracing eDRAM technology (for on-chip storage) for use in many commercial products [10]. 1T1C eDRAM cells are the most widely used but because of their high latency and destructive reads, they cannot be used in latency-sensitive components like L1 caches. On the contrary, basic 3T-eDRAM and its derivatives in addition to offering access speeds on par with regular SRAM, provide non-destructive reads and are capable of driving large-loaded bitlines suitable for operation in low-voltage caches [115]. As the data stored in L1 caches is highly transient, when using eDRAM over SRAM, the need to refresh the cells can be completely eliminated if the data can be held until its last reference before it is evicted. Previous studies have highlighted that data residing in a L1 cache-line is accessed no later than 20000 processor-cycles after it was first written [66]. However, the two major disadvantages of using 3T-based (3-Transistors) eDRAM over SRAM are : *a.)*Read access times incrementally increase with reducing storage-node voltage and beyond a certain point in time is no longer comparable with SRAM speeds.

*b.)*As retention time can be in the order of μ seconds, for a pulse generated by a particle strike, the window where it can manifest as a bit-flip is very wide, making the cell highly susceptible to soft-errors during hold mode. By increasing the retention time, not only can the cell guarantee fast reads for a larger number of accesses, the time period during which the charge-equivalent of the voltage at storage node is near $Q_{critical}$ is reduced relatively enhancing soft-error tolerance also.

In this chapter, we present a novel 4T-based eDRAM cell that when compared to a similar sized eDRAM cell has higher tolerance to process variations and soft-errors. We replace the gated-diode in a 3T1D (3-Transistor 1-Diode) cell with a NMOS pass transistor to suppress sub-threshold leakage which in turn improves the retention time. This has shown to improve the retention by 2.04X on-average. The only downside though is that the absence of the gated diode reduces gate-overdrive making the cell slower. The resulting degradation in access time time is 3%. This increase can be accounted for at design-time and in most-cases can be accommodated within the read-access latency. Another function of the pass transistor is to reduce the total sensitive-area exposed to neutron strikes. With a long-channel length pass-transistor, the amount of time needed by the generated pulse to traverse through is long enough to generate only a small glitch and not a bit-flip itself. As we will show later, the soft-error rate which is measured in terms of failures-in-time (FIT) is reduced by 36% on average.

5.3 Revisiting 3T1D Operation: Retention Time Perspective

Figure 5.1 shows a regular 3T1D-eDRAM cell. Pass transistors T1 and T3 provide separate write and read ports respectively. Such a multi-ported cell in addition to improving noise margins, helps the cell cope with device mismatches due to variations because of its asymmetrical design. The gates of transistors T2 and D1 represent the storage node. A value is written into the cell by strobing $Wordline_{write}$ and raising $Bitline_{write}$. A read operation is initiated by pre-charging $Bitline_{read}$ to V_{dd} and then raising $Wordline_{read}$. We use a domino sense-amplifier consisting of a series of inverters to drive the output data. Unlike SRAM cells, where the bi-stable loop of inverters ensure that voltage at the storage node is always near V_{dd} or Gnd, the voltage after a write-operation at the storage node is always degraded (near $0.6V_{dd}$ for a '1'). As a result, the consequent reduction in read-current impacts read-access times negatively. However, an auxiliary function of

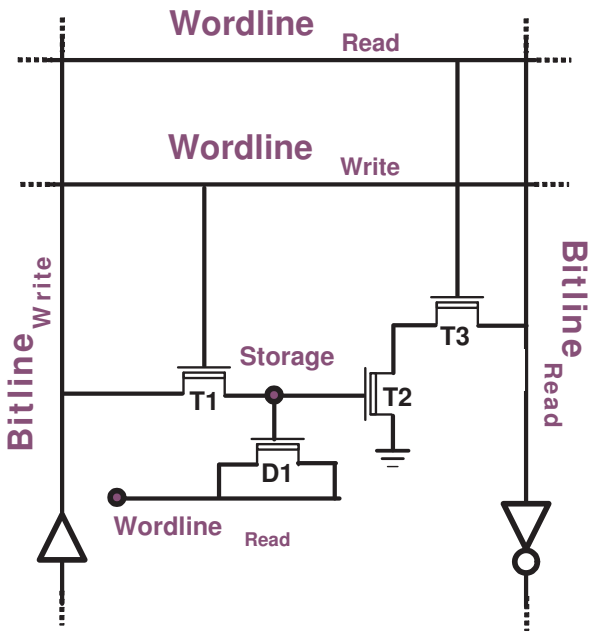


Figure 5.1: 3T1D-DRAM cell schematic

the gated-diode D1 is to speed-up read-accesses by boosting the voltage at storage-node during a read. At any given point in time, the total charge stored in the gated-diode is given by (V_{high} - V_{th})*C_{diode} where V_{high} is the initial-voltage (after a write) corresponding to a '1' and C_{diode} is the internal capacitance of the diode that reaches a maximum value when $V_{gate-source_{diode}}$ >(V_{th}+150mV). Here $V_{gate-source_{diode}} = V_{high}$ when the data is first written. When the cell is read by precharging $Bitline_{read}$ and enabling $Wordline_{read}$ (also connected to the source & drain of D1), extra amount of charge is pumped into the cell through D1 to raise the voltage at the gated-diode to a level sufficient enough to guarantee fast reads. At the end of the ready cycle, when $Wordline_{read}$ is disabled, as the voltage at the source of the diode starts reducing, the charge lost during read operation is returned back to *storage-node* helping to restore the voltage at the gate to the state before the read operation resulting in non-destructive read accesses.

Despite the fact that reads being non-destructive helps to lower dynamic power by reducing refresh rates, the gradual reduction in stored-charge makes the cell slower with subsequent accesses and beyond a point in time, the speed of the cell is no longer comparable to that of a regular SRAM as shown in figure 5.2. In this context, we define the *retention time* of the cell as the total time elapsed after a write beyond which the cell is slower than a SRAM. In order to design an eDRAM cell capable of operating at near-SRAM speeds, the expected access-time is of the order of 320-350ps (45nm Technology)

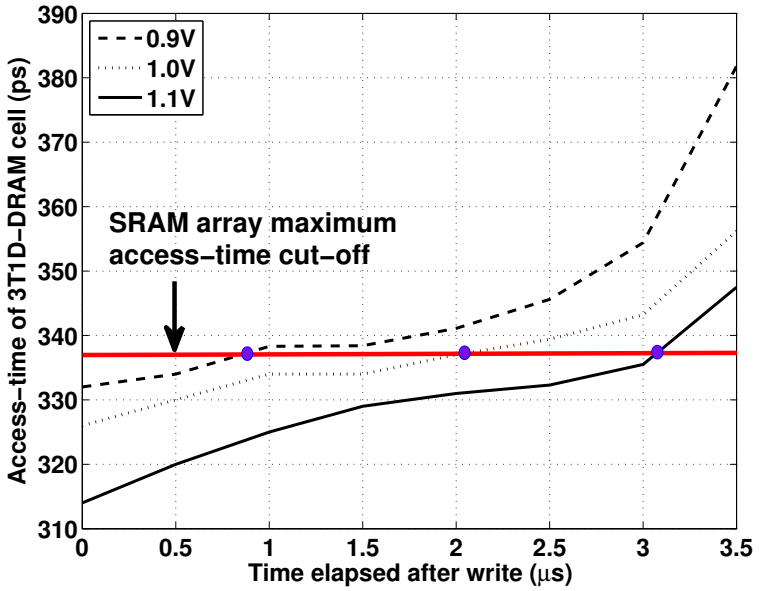


Figure 5.2: Increasing 3T1D read access latency with time after a write access.

[66]. We can see that with reducing supply voltage, the time period for which the cell can be accessed at SRAM speeds also reduces considerably. As a result, even if the cell can hold the data for a much longer period, the data needs to be refreshed in-order to guarantee SRAM-like access speeds. As the gain of the gated-diode is dependent both on the charge held and V_{high}, it should be noted that boosted voltage-level itself reduces with time making it harder for T2 to be switched on. Contrary to the notion that at low-voltages where leakage is minimum and thus retention is maximum, over here V_{high} is lower and hence the factor of gated-diode drive strength is an important factor in the access-speed of the cell.

Improving retention time of the cell has a two-fold advantage: *a.)* The cell can be operated at near-SRAM speeds for a larger period of time. *b.)* High retention time reduces refresh rate which in-turn helps minimize dynamic power consumption. In the next-section, we will discuss a novel implementation of the cell that achieves all of the above desired characteristics.

5.4 4T-DRAM Cell

When modifying the 3T1D-DRAM cell to function without the gated-diode, the access-latency of the new cell is only slightly higher because the diode boosts only a fraction of the total node capacitance. In figure 5.3, we show a new 4T-DRAM cell that does not suffer from the primitive leakage issues found in the 3T1D-DRAM. The gated-diode is replaced with a NMOS pass-transistor that serves to decouple the leakage paths present in a conventional 3T1D cell. Similar to a 3T1D-DRAM, a write operation is initiated by raising both $Wordline_{Write}$ and $Control_{refresh}$ and then raising $bitline_{write}$. Under the effects of process variations, the threshold voltage of T1 can be severely degraded and leakage currents in the hold mode can potentially destroy the data by discharging into the $bitline_{write}$. The threshold of T1 can be made large enough only to such extent that

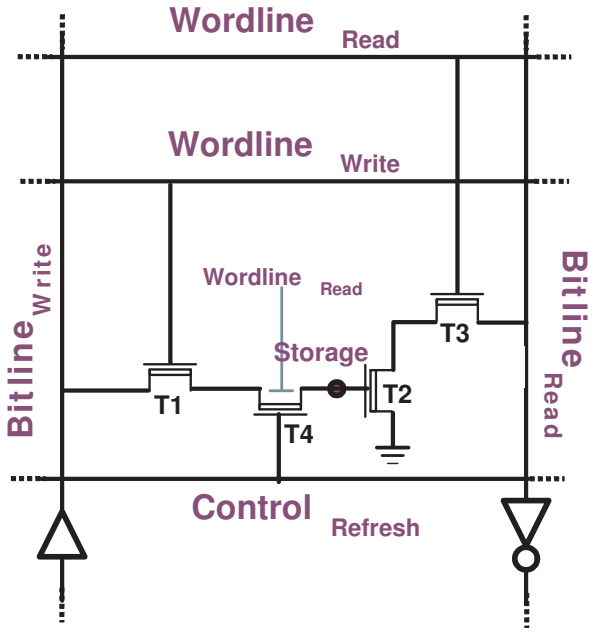


Figure 5.3: 4T-DRAM Cell Schematic

it does not affect the write latency negatively. Also, by introducing an extra transistor, the voltage-level at the storage node when a '1' is written is degraded further due to the threshold of transistor T4. This results in lower V_{high} that could potentially worsen the retention time. However, as we will later-explain, we circumvent this problem using conventional mechanisms like body-biasing. Read operation is initiated by precharging of the $bitline_{read}$ and enabling both $Control_{refresh}$ and $Wordline_{read}$.

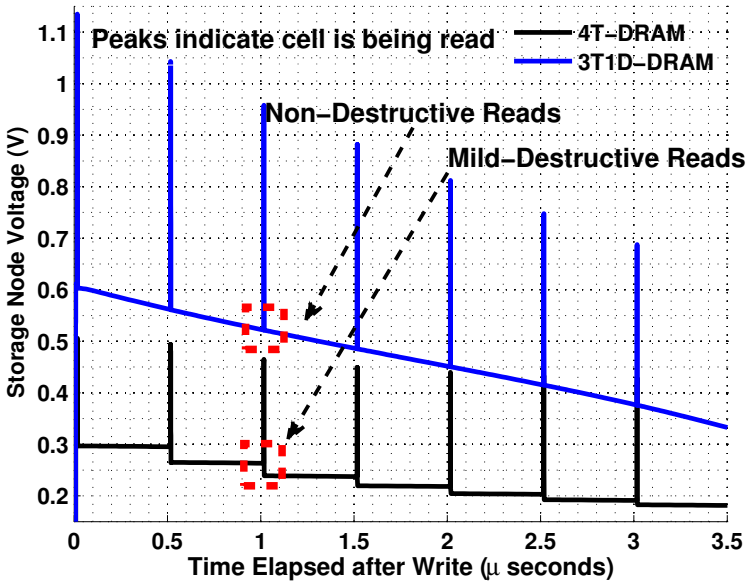


Figure 5.4: Storage node voltage of 3T1D-DRAM and 4T-DRAM during read and hold mode.

Over here, we read the cells every 0.5μseconds after the initial write. The peaks in the figure represent the boosted voltage during read accesses. We can notice two main deficiencies with the modified 4T-DRAM cell. Firstly, the initial node voltage (at time = 0) is smaller than that of 3T1D-DRAM by a factor equal to the threshold of T4. As a result, the gain (ratio of boosted/non-boosted voltage) in a 4T when compared to 3T1D is 1.43 as opposed to 1.88. In a 3T1D, the diode boosts only the charge stored internally which reduces with increasing threshold. Hence, the effectiveness of the diode in boosting reduces with increasing threshold. Increasing the threshold of T4 reduces leakage by orders of magnitude improving retention times but also increases the time for a write access. However, architecture-level simulations have shown that write accesses in caches are not as critical as reads and therefore minimal increase in write-time does not impact system-wide performance. Modern day manufacturing processes allow co-existence of low-V_{th} and high-V_{th} devices using advanced body doping profiles or gate work-function engineering [50]. Large threshold devices also have much better V_t roll-off resulting in lower amount of threshold deviation due to process variations. Increasing only the threshold of T4 by 75mV, we were able to notice a 89% increase in the retention time. From the figure 5.4, it should also be noted that the slope of decay is higher for 3T1D when compared to 4T. This is because, T4 completely suppresses sub-threshold leakage by being super-cut-off and reverse-biased in the hold-state which enables the

storage node to stay at a particular voltage level for a longer amount of time.

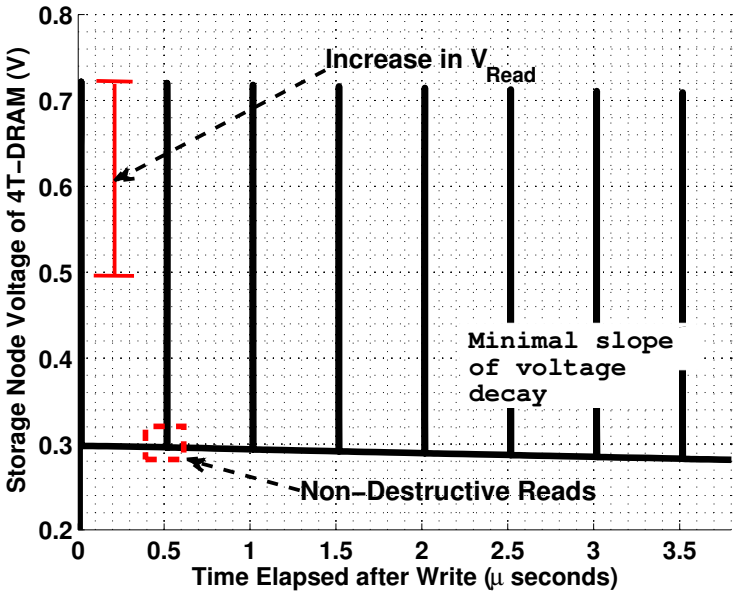


Figure 5.5: Storage node Voltage of 4T-DRAM with body-biasing

The other deficiency in a 4T results due to a fixed amount of charge being lost after every read-access. The consequent drop in the voltage decreases with subsequent accesses. We circumvent this problem by forward body-biasing (FBB) T4 using the $Wordline_{read}$ line as bias control line. Under such scenario, the behaviour of the storage-node voltage during read and hold mode is demonstrated in figure 5.5. During write and hold state, the behaviour of T4 is unaltered because it is biased at zero voltage (switching-off $Wordline_{read}$). Upon enabling the $Wordline_{read}$ signal, the device is forward-biased reducing the threshold of the device making it faster. Note here that $Wordline_{read}$ is used only as an enable signal and T4 is body-biased from a separate source. Therefore, wordline drivers need not be designed to drive large capacitance. Because we bias all the cells sharing the same $wordline_{read}$ irrespective of whether they are selected or not, the extra charge flowing into the cell should be only significant enough to compensate for the charge lost after a read and insignificant enough to ensure that it does not flip the state of a cell storing a '0'. It can also be observed from the figure that voltage at the storage node during a read increases by a factor of 35% when compared to the non-biased 4T DRAM cell. In addition to improving the overall retention time, the reduction in slope of decay helps to access the cell at SRAM speeds for a larger period of time. The advantage of applying forward body-biasing to a device with high threshold is that it allows to re-

duce the gate-length by a significant percentage when compared to a low-threshold device without body-biasing while maintaining the same drive current. Thus for a smaller sized T4 (hence smaller cell area), the same performance and retention time can be achieved. The applied bias voltage is kept minimal to ensure source-body junction currents are limited. Our simulations indicated that FBB'ing at 25-40mV was sufficient enough to improve the retention time and also ensure there are no bit-flips (0→1). The need for separate wells (triple-well) in FBB mode can be eliminated by using minimum-overhead control structures (that are immune to substrate biasing) for isolating the supplies of the well and bodies [46]. Also, the existence of substrate resistance increases the time to body-bias transistor T4. By enabling the control structures during the address decoding process, the transition latency involved in enabling/disabling the bias lines can be eliminated and T4 can be biased well in advance of driving $Wordline_{read}$.

5.5 Impact of Process Variations

In this section, we compare the cell access and retention times of 4T and 3T1D under the impact of process variations.

5.5.1 Simulation Parameters

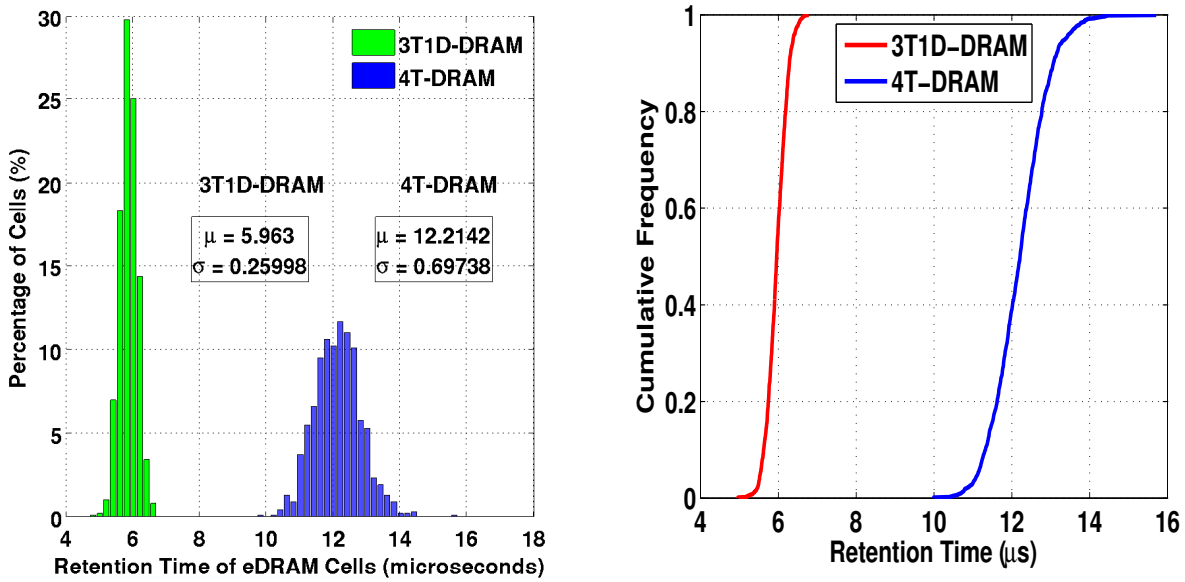
Both eDRAM cells are designed in 45nm HP PTM technology [1]. The transistor dimensions are scaled from $\text{Length}_{T1} = 4\lambda$, $\text{Width}_{T1} = 3\lambda$, $\text{Length}_{T2} = 2\lambda$, $\text{Width}_{T2} = 16\lambda$, $\text{Length}_{T3} = 2\lambda$, $\text{Width}_{T1} = 4\lambda$, $\text{Length}_{D1} = 8\lambda$ and $\text{Width}_{D1} = 20\lambda$ where $2\lambda =$ 45nm. Transistor dimensions (listed previously) of the 3T1D cell are scaled-up (by 30% approximately) to account for the area-overhead due to metal track of the refresh line and body contacts in the 4T cell. As the maximum speed of a memory array is dependent on the ability of the slowest cell to create a minimum bit-differential on the worst sense-amplifier, we simulate a 256X256 eDRAM (both 3T1D and 4T) array and measure the access-time of the cell with the lowest retention to extract corner-case scenarios. The values for different sources of variation for the 2 process parameters considered are provided in table 5.1.

Table 5.1: Parameter Deviation

Parameter	D2D	Systematic	Random
V_{th}	±3%	±6.4%	±6.4%
L_{eff}	±3%	±3.2%	±3.2%

5.5.2 Variation in Access and Retention Times

Figure 5.6a shows the distribution of retention time of both cells under process variations. Notice here that retention time corresponds to the total period after a write where the access-time of both cells are comparable to SRAM array speeds. The retention time of the 4T on-average is 2.04X higher than that of 3T1D. In a 3T1D cell majority of the charge is stored in transistor T2. In order to improve the read speed, read-access transistor T3 is sized relatively smaller when compared to D1 and T1. In figure 5.7, we have shown



(a) (b)

Figure 5.6: (a) Retention time-variation of 4T and 3T1D under process variability (b) Cumulative distribution function of the retention times of 4T and 3T1D under process variability.

the influence of the read-path on cell retention time. It is clearly observable that the retention of 3T1D is heavily influenced by gate-length variation of T3 while the retention

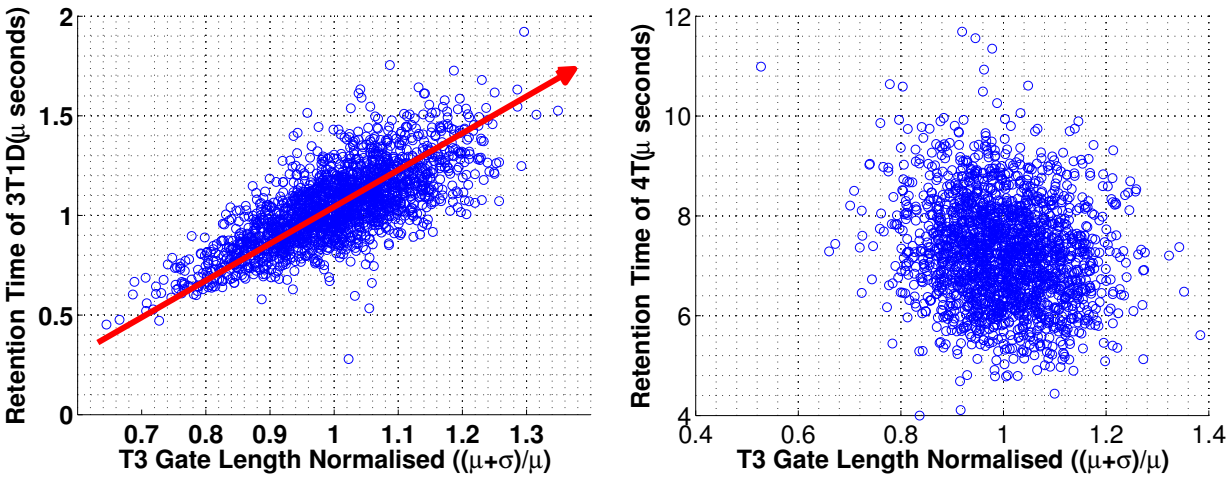


Figure 5.7: Influence of gate-length variation of T3 on retention time. (left) 3T1D (right) 4T

of 4T is least influenced. Our results indicate that the low retention in 3T1D is primarily due to sub-threshold leakage due to a weak read access-transistor resulting in majority of the charge flowing into the bitlines. In addition, $V_{gate-source}$ of T2 in hold state is well above V_{th}, results in large gate-leakage currents further reducing retention period. Transistor models in this particular PTM card do not exploit advanced leakage reduction mechanisms like High-K gate dielectrics. Hence the effects of gate-leakage are not trivial. In the case of 4T, the threshold of T4 ensures $V_{gate-source}$ of T2 is just near V_{th} reducing gate-leakage by 3-6 orders of magnitude. Also, the $V_{gate-source}$ of control transistor T4 in hold mode is at 0, suppressing other sources of junction-currents further. The higher standard deviation in the retention time of the 4T is not of much concern as it was earlier discussed that for L1 caches the data needs to be held only for 20000 processor cycles (approximately 6μs at maximum frequency) and our results indicate that the 4T in the worst-case holds the data for more than 9.8μs.

The access-time is mainly dependent on storage node voltage at the time of access and the read path drive strength [38]. The read-path driving capability is mainly dependent on the threshold of transistor T2 and T3. By lowering the threshold, the speed can be considerably improved. However, with reducing threshold the decaying of storage node voltage is faster resulting in slower read accesses. In figure 5.8, results for read-access time variation are presented. It was shown in [115] that the dependence of cell access time on transistor parameters varies temporally. For example, when T3 has higher

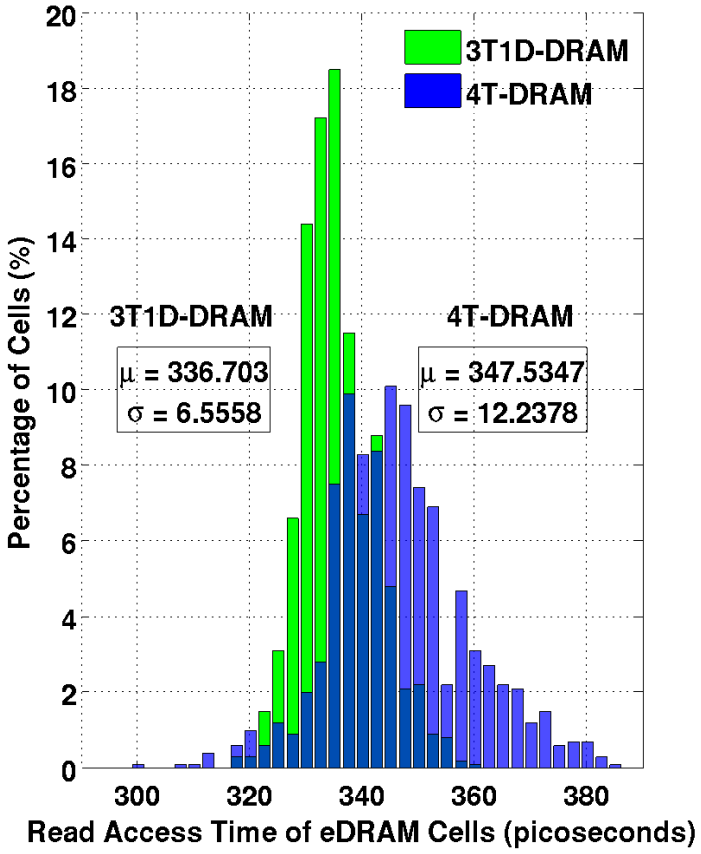


Figure 5.8: Read access time-variation of 4T and 3T1D under process variability

threshold, during the initial period of decay during which the cell is accessed the most number of times, the drop in voltage is insignificant enough to increase cell-access time. When accessing at a later point, the effect of cell leakage due to channel length variation will be more crucial because with reducing storage node voltage, voltage-gain reduces, diminishing the ability to strongly pull-down T2. From the figure we observe that the average access time of 3T1D cell is better by 11ps (3.2%) on average. This is mainly because of the voltage-gain achieved by the gated diode and initial V_{high}. Nevertheless, the access-time of both cells is around the expected latency cut-off of 340ps. While the 3T1D is faster than the 4T cell, it should be noted that with lower retention time, the need for frequent dynamic refreshes increases and during this period the cell cannot be accessed. On the contrary, the 4T can be accessed for a much longer period at SRAM speeds without the need for frequent refresh operations.

5.5.3 Power consumption comparison

The dynamic and leakage power of a 3T1D cell (nominal-case variations) were found to be 47.8mW and 13.2mW respectively. On the other hand, for 4T the dynamic and leakage power is 39.7mW and 7.9mW respectively. It is interesting to note that despite employing forward-body biasing in a 4T, there is a near 20.4% and 67% reduction in dynamic and leakage power respectively. High retention time of 4T which reduces the need for frequent data refreshes tremendously and hence lower dynamic power. Whereas in the case of a 3T1D, the refresh rate was 2.2X higher. Note that biasing as such does not alter the basic functioning of the cell because it pumps only the lost charge back into the cell. When choosing not to use biasing, we were able to match the retention time of 4T with that of 3T1D without significant overhead in the read-access time. It also negates the need for extra body contacts and improves the density of the 4T cell.

5.6 Soft-Error Tolerance

We have seen in the earlier section that the amount of charge held in the cell during standby reduces with time. This means that the amount of charge necessary to flip the state of the cell also reduces increasing vulnerability to soft-errors. Soft-errors occur when high energy particles (e.g. atmospheric neutrons for ground and aeronautical devices, heavy ions and protons for space applications) strike sensitive regions of a transistor depositing enough energy to flip the state of the cell. Simultaneously, the radioactive decaying of impurities present in on-chip interconnects and packaging will emit secondary α-particles with sufficient energy also capable of producing soft-errors. In the context of eDRAM cells, soft-error susceptibility is primarily dependent on write voltage, storage-node voltage at the time of strike and bitline capacitance [38]. We consider only the impact during hold mode as the time period during which the cell is susceptible to soft-errors during an access dependent on the transfer of charge from/to the cell is negligible when compared to the bitline capacitance. For the sake of brevity, only neutron soft-error characteristics of 4T-DRAM are discussed. The soft-error rate of both cells is estimated using the iRoC TFIT simulator [2]. The tool relies on a technological process characterization database allowing the generation and evaluation of any transient currents that may be induced by single events. The cell response is then evaluated w.r.t.. to these events in a given working environment, allowing the tool to evaluate the failure-in-time (number of failures/10^9 operating hours) rate for all the sensitive nodes within

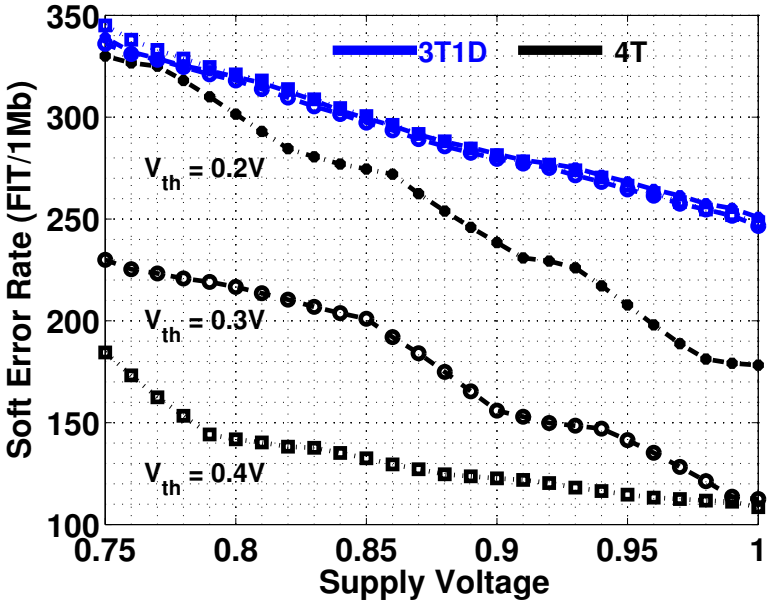


Figure 5.9: Soft-Error rate under varying supply

the cell for a set of large number of input parameters like neutron-flux, supply voltage, threshold, pulse-width etc. In our simulations we used a 45-nm generic CMOS database in conjunction with 45nm HP PTM models. Figure 5.9 shows the soft-error rate (FIT/1Mb) for 3T1D and 4T under varying supply voltages and for 3 values of threshold voltage. We initially notice that the soft-error rate of 4T is the same as that of 3T1D when threshold is 0.2V and at supply voltages less than 0.8V. With reducing supply and threshold, $Q_{critical}$ also reduces substantially. In the hold mode, when a particle strikes at the storage node, the capacitance at the gate of T2 should be large enough for the node to be not to be flipped. In a 3T1D, as the diode D1 itself contributes only a percentage of the charge it holds, soft-error susceptibility is maximum in this scenario. The measured FIT rate is the average of FITs of all the sensitive nodes inside the cell. The 4T in contrast to 3T1D, has two sensitive nodes (storage node and the node joining T1 and T4). Henceforth we will refer to this node as non-storage node. When considering high-threshold voltages, for a particle strike at the non-storage (during hold-mode), the generated pulse travelling through T4 into storage-node is attenuated because the control transistor is switched off disconnecting the storage node from the point of strike (non-storage node). As a result, it generates only a small glitch and not an upset itself.

In figure 5.10, the effects of process variations on soft-error rate is shown. The FIT rate of 4T is lower than a 3T1D by 1.56X on average. In the case of 3T1D, with reducing

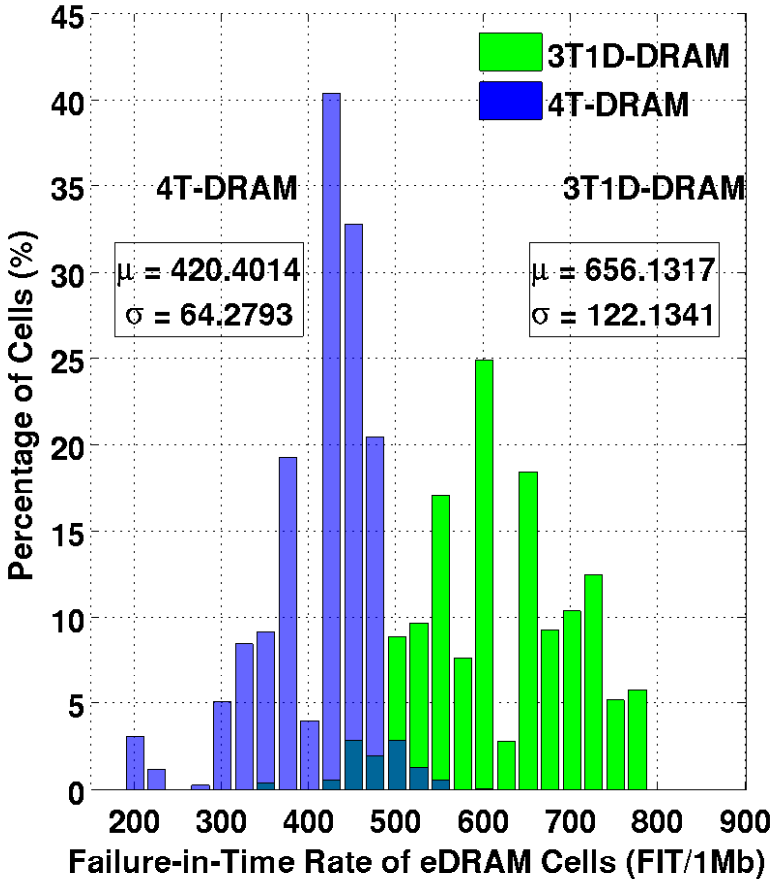


Figure 5.10: Soft-Error rate under process variations

gate-length of transistor T2, the $Q_{critical}$ is reduced because of a strong pull-down action. With increase in length, the gate capacitance increases improving $Q_{critical}$. The highest impact on $Q_{critical}$ in a 4T is due to the threshold of transistor T4. Increasing the V_{high} makes the cell stay well above the $Q_{critical}$ for a longer period. When the threshold of T4 reduces, a larger amount of current can travel through T4 increasing the susceptibility to bit-flips due to particle strikes. Thus the probability of a strike on non-storage node resulting in an upset increases in such a scenario.

5.6.1 Multiple Bit Upsets (MBU)

Another common issue with particle strikes is the energization of adjacent bitcells when neutron-induced ions pass through silicon. While 2-bit cell upsets are more common, higher-order cell upsets are largely dependent particle incident angle. Using TFIT, for

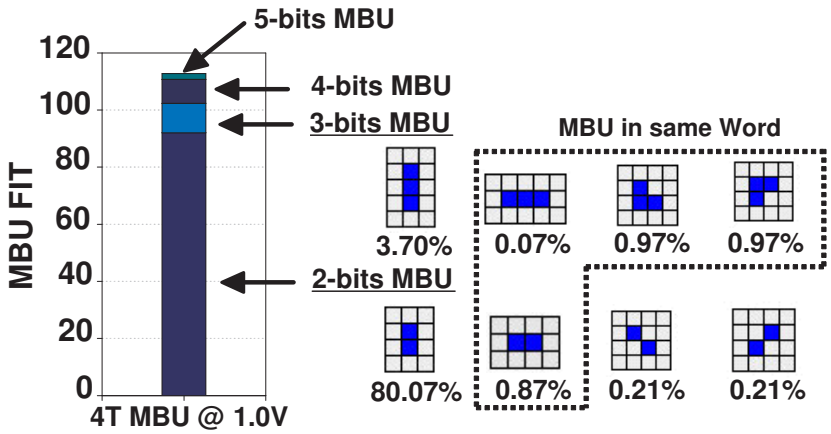


Figure 5.11: Multiple Bit Upsets in 4T

a set of random data pattern we measure the impact of MBU on the overall FIT rate of 4T as shown in figure 5.11. We see that most of the MBU (80.07%) are 2-bit upsets in the vertical direction. These errors can be corrected by simple single-bit ECC codes. However, errors in the lateral direction need more complex higher-order ECC codes. Otherwise, techniques like column-interleaving (where adjacent columns hold data from different words) can be used to minimize impact on system failure. In table 5.2, we

Cell	VDD [V]	w/o ECC [FIT]	w/ECC [FIT]	MBU Reduction
3T1D	0.9	279.30	55.08	80.28%
	1.0	269.25	53.59	80.1%
	1.1	239.57	49.04	80.53%
	1.2	227.94	43.54	80.90%
	1.3	188.404	36.09	80.85%
4T	0.9	129.4	26.2	76.77%
	1.0	112.8	21.32	80.1%
	1.1	42.48	5.08	89.04%
	1.2	36.16	3.6	90.04%
	1.3	2.2	0.16	92.77%

Table 5.2: MBU Comparison between 3T1D and 4T

compare the MBU FIT of 3T1D and 4T with and without ECC. The relative MBU tolerance of both types of cells is very high which is evident from the fact that when applying ECC, most errors are corrected. This means that the percentage of errors in the lateral direction is minimal. When comparing the MBU FIT (w/o ECC) of both cells, 4T's FIT is lower by a minimum of 46% and maximum of 99.99% observed at 0.9V

and 1.3V respectively. Employing ECC, 4T's FIT is lower by a minimum of 61% and a maximum of 99.996% observed at 0.9V and 1.3V respectively.

5.7 Summary

In this chapter, we have presented a novel 4T-DRAM cell that is both variation-tolerant and hardened against soft-errors. We have thoroughly investigated the sources of inefficiencies in a 3T1D eDRAM cell and propose to replace the gated diode with a NMOS pass-transistor that controls the leakage path. It was shown that the control transistor by suppressing leakage currents can help the cell achieve better retention time. Finally, we show that the cell is more tolerant to soft-errors by making the cell stay above the $Q_{critical}$ for a larger period of time with large retention times.

6

Hybrid Techniques to Enhance Parametric Yield

6.1 Overview

While new devices and structures are extremely promising in terms of robustness to process induced variations, and offering better leakage power profiles; their widespread induction into mainstream products is inhibited by the prohibitive costs involved in investments in new foundries, development of new design rules that may or may not scale with future nodes and most importantly the economics governing chip yield. It was shown that among all factors affecting yield, nearly 50% yield losses result from parametric variations [82]. In this chapter, we observe that large-scale yield enhancement is achievable through failure prevention and correction. We first develop a tool to be able to rapidly explore a wide choice of robust memory designs with minimum available specifications about the memory design and architecture. Using this tool, the effectiveness of a new class of hybrid techniques in improving cache yield is studied. We also show using a combination of failure prevention and correction techniques offer better quality-energy-area

trade-offs when compared to their standalone configurations. The remainder of the chapter is organized as follows: section 6.2 discusses prior yield improvement techniques and argues the need for an alternate design platform for adopting a variation-aware design paradigm. Section 6.3 introduces and discusses in detail parametric failures in SRAM cells. We also propose a novel methodology for fast estimation of failure probability under the impact of parametric failures in this section. Then, we show how this methodology is seamlessly integrated into a novel design framework (INFORMER) in section 6.4. Three different use cases of the proposed framework are presented in section 6.5. In the second part of the chapter, starting with section 6.6, we introduce proactive read/write assist techniques for failure reduction. In the following section 6.7, two reactive techniques are studied. The joint impact of proactive and reactive techniques as hybrid techniques is studied in section 6.8. The concluding remarks are presented in section 6.9.

6.2 Motivation and Background

Process variations make designing SRAM memories extremely cumbersome not only because they produce a large spread in power and performance characteristics but because they also make the cells weak inducing failures caused from lowering of noise margins. Process variation induced failures known as - *parametric failures* are of particular concern to chip designers as they can influence a large number of design choices at different levels of abstraction [18]. For example, the Vdd_{min} of the cache which decides the Vdd_{min} of the whole processor is dependent on the highest Vdd_{min} among all cells in all arrays. This would mean that, under the effects of random variations where failures are distributed, a single SRAM cell could potentially affect the functional yield of the whole processor if no proactive/reactive techniques are in place [110].

Prior yield improvement techniques proposed in the literature use special assist circuits to modulate one or more voltage sources governing regular operation of the cell [83, 86]. Then there are also other techniques that vary the pulse-widths of one or more enable signals that govern the regular access to the cell [58]. These improve the failure probability of the 6T-SRAM by enhancing read/write margins thereby improving ***functional margin*** of the memory cell. However, the consequent improvement in parametric yield comes at the cost of increased dynamic power consumption or reduced memory area-efficiency. In the context of memory yield, assist methods can be thought of as *proactive techniques* that can help prevent failures by lowering of failure probability and ensuring failures are

a rare occurring phenomena. Worst-case scenarios where failures will still be persistent, require *reactive techniques* that can help the system cope with those remaining failures or completely eradicate them for complete fault-free operation. Error-correcting codes (ECC) and redundancy are two examples of *reactive techniques* that can be leveraged at run-time to improve the ***functional yield*** of the memory system albeit incurring area, performance and power overheads.

An important component of analyzing and evaluating new memory architectures that exploit proactive or reactive techniques is the need to be able to explore a wide array of design choices rapidly and accurately. This can be crucial for design teams as it can have a potential impact on revenues by influencing product turn-around times. Recent research focused in this direction has highlighted the pressing need for tools and design flows capable of evaluating global metrics such as yield, power and performance to enable making optimal design choices. In the context of robust embedded memory design, such tools can help designers across different levels of abstraction have a holistic perspective of system-wide metrics with minimum specifications available about the memory architecture, technology, process characterization, periphery topology, bitcell design *etc.* Choices made early in the design stage ensure conflicting requirements from higher-levels are decoupled.

In this chapter, as a first step, we propose a novel tool, *INFORMER (Integrated Framework for Early-Stage Memory Robustness Analysis)* that encompasses statistical SPICE-level simulations of the memory macro and back-of-the-envelope analytical models to estimate memory reliability under the impact of parametric failures. We present a novel algorithm for estimating the failure probability of SRAM cells by leveraging transistor dimensions. The implementation achieves near-SPICE like accuracy while improving simulation time by orders of magnitude. In the second step, using *INFORMER*, the effectiveness of a new class of hybrid techniques in improving cache yield through failure prevention and correction is evaluated. Proactive read/write assist techniques like body-biasing (BB) and wordline boosting (WLB) when combined with reactive techniques like ECC and redundancy are shown to offer better quality-energy-area trade-off when compared to their standalone configurations. We will show that best solution is not to choose between reactive or proactive techniques but to carefully analyse their associated trade-off and implement a system where they can co-exist by complementing each other.

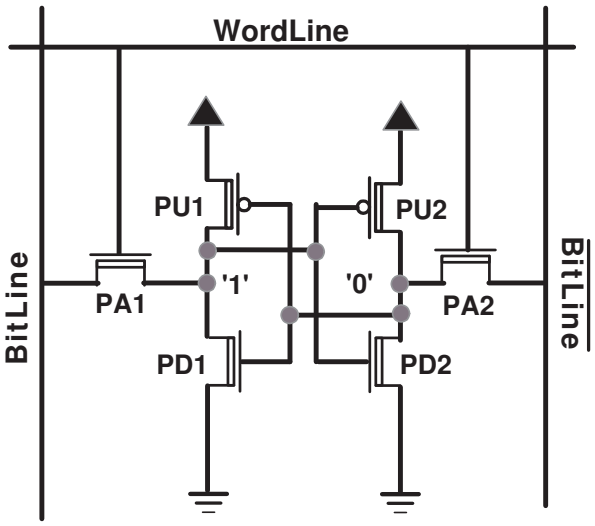


Figure 6.1: 6T-SRAM Cell Schematic

6.3 Parametric Failures in 6T-SRAM Cells

As shown in figure 6.1, a 6T-SRAM cell consists of a pair of inverters connected in a positive feedback loop creating a bi-stable circuit which allows for the storage of complementary values in the input/output nodes of the inverters. Strobing the word-line provides a read/write path into the cell through the access transistors PA1 and PA2. By raising the bit-lines high/low (low/high) and strobing the word-line, a successful write operation is completed. Prior to a read access, the bitlines are pre-charged to V_{dd} and the wordline is enabled again. A sense-amplifier is enabled to amplify the developed bitline differential to signal either a 0 or 1. As a result of parametric variations in the threshold voltage of each of the 6Ts, a SRAM cell can fail in one or more of the following ways :

- *Read Stability Failure* (RF) When reading a node storing a '1', the opposite node storing a '0' can flip as a result of induced noise. This noise is primarily due to the transfer of charge developed on the bitlines (post pre-charge) on to the pull-down creating a voltage divide between the pass transistors and the pull-down. This results in a temporary voltage surge at the node storing a '0'. If this voltage is greater than the trip point of the left-inverter (PU1-PD1), then the cell flips during the read mode. By reducing the wordline voltage or increasing the threshold of the pass transistor, the read disturbance can be significantly reduced. The consequent reduction in read current manifests as an increase in cell access time. The mechanism behind such failures is shown in figure 6.2.

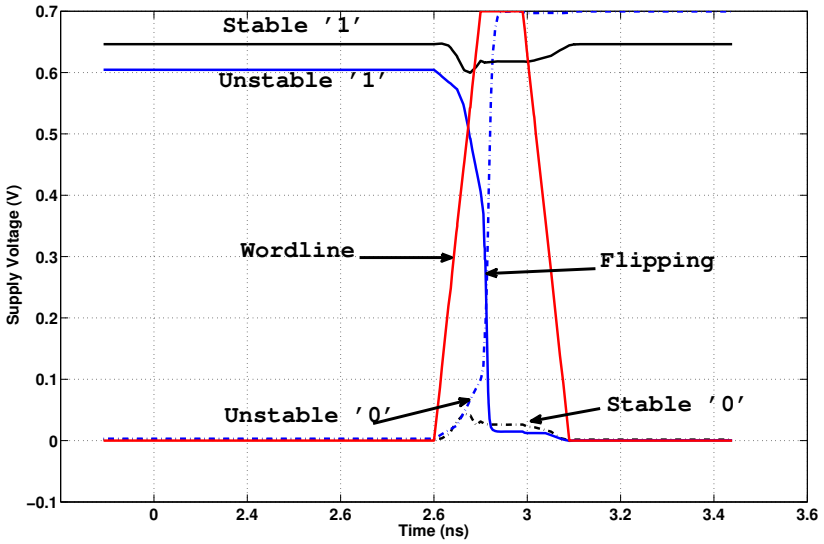


Figure 6.2: Mechanism of unstable and stable read operations leading to a failed/successful read accesses

- *Write Stability Failure*(WF) This mode of failure occurs when a node storing a '1' cannot be pulled below the trip-point of the opposite inverter in the time the wordline is high. In other words, writing a '0' into a node storing a '1' does not result in a flip of the values. By making the access transistor stronger than the pull-up or by increasing the pulse-width of wordline enable signal, these failures can be reduced. The mechanism behind such failures is shown in figure 6.3.
- Read Access Failure(AF) If a cell fails to produce a bit differential greater than the delta of the sense-amplifier (SA) in the time the SA enable (SAE) signal is high, then it results in an access failure. Increasing the pulse-width of the wordline helps reduce these failures. However, it also creates the opportunity for an accidental write operation flipping the state of the cells.
- Hold Failure(HF) With reducing supply, the voltage difference between a '1' and '0' reduces. As a result, in the standby mode, when reducing the supply below a certain level, the contents of the cell are lost. For caches that do not employ any aggressive failure mitigation mechanism, the Vdd_{min} of the array is calculated based on the worst-case cell within the array [61]. Maximum leakage reduction in the standby mode entails a minimum $V_{dd_{min}}$. Such a pessimistic approach, in addition to narrowing the window for possible leakage reduction, also reduces the gap between Vdd_{min} & Vdd_{max} leading to device reliability issues. The mechanism behind such failures is shown in figure 6.4.

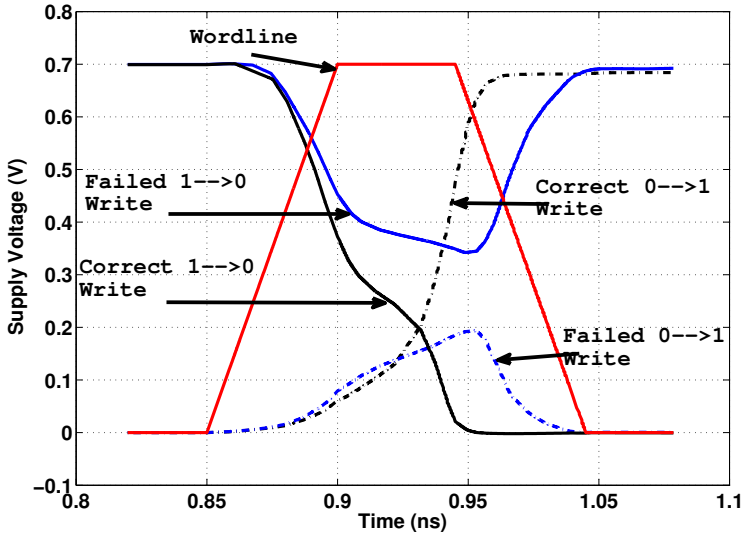


Figure 6.3: Mechanism of unstable and stable write operations leading to a failed/successful write accesses

At low-variation corners operating at high-V_{dd}, access failures and hold failures dominate over write and read failures. This is because the relative small width of PA transistors cause a large deviation in the V_t of the transistor and if the Vt_{PA} is larger than the nominal V_t, then the access time increases resulting in a failure. Similarly, when the storage node is not tightly held to either V_{dd} or V_{ss} during hold mode, the increase in leakage can cause voltage to degrade further and eventually the cell loses its contents. These type of failures increase by orders of magnitude at lower supply voltages, when the trip point of the inverters also reduces significantly. While at low-voltages the leakage factor is not important, because of parametric variations, if PU becomes weak, then voltage stored at the node gets reduced further exacerbating this type of failures. As the failure probability increases, the number of failures for any given cache size also increase. This phenomenon is reflected in last-level caches which typically require cells with 10^{-11} failure probability for achieving 99.9% yield [116]. Read failures are dependent on the relative strengths of the PA and PD transistors. At high V_t corners, if Vt_{PA} is low and Vt_{PD} very high, then large currents discharge from the bitlines on to the storage node accumulating charges and/or the storage node becomes too weak to discharge through the bitlines and this causes a failure. On the other hand, when PA becomes weaker when compared to PU, the time to discharge through the bitlines increases and cause write

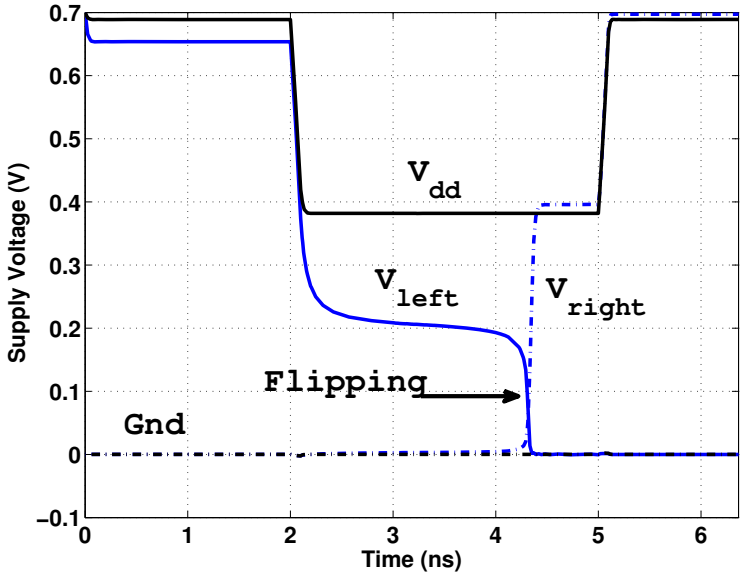


Figure 6.4: Mechanism of unstable and stable standby mode leading to a failed/successful retention of value.

failures. It is for this reason that the cell transistors are sized properly to ensure that the time to write is lesser than the time the wordline is high [77].

6.3.1 Rapid Failure Probability Estimation

At nominal voltages, the failure probability is extremely low owing to sufficient noise margins requiring exhaustive number of simulations for estimating failure probability. As a result, estimating the failure probability in minimum time requires special approximation techniques. It is possible to minimize the large number of simulations by reducing the search space by determining *apriori* the failure region of maximum likelihood. The accuracy of measurements in such cases is largely dependent on the technique used. *Most probable failure point* analysis is one such technique that can be leveraged to improve the efficiency and accuracy of traditional monte-carlo based methods [57]. The task of failure probability estimation can be simplified into the problem of finding in the variation space the most probable point of occurrence for which the SRAM cell fails. In a regular six-transistor (6T) cell, random variations in the threshold voltage are primarily responsible for parametric failures. The number of *directions* in which the threshold variations of all six transistors can vary within the variation space is $2^6 = 64$. Previous published work assumes that for each of the four failure mechanisms, there is only one direction where

failures are dominant. However, our simulations have demonstrated that the *length* and *width* of the six constituent transistors can greatly influence the failure probability as they modify read/write parameters, cell leakage and propagation delays making it harder to determine accurately the failure probability accurately when limiting the search space to just one direction.

Algorithm 2: *Leveraged* MPFP analysis

Input: Cell Dimensions (Symmetric), σV_{th}, $P_{fail} = 0$
Output: failure probability
for *Each Failure Mechanism* **do**
  $Dim = \{L_{PU}, L_{PD}, L_{NA}, W_{PU}, W_{PD}, W_{NA}\}_{normalised}$ $Failure_{Directions} = \bigcup_{i=1}^{|Dim|} Direction^{i}_{lookup}$, i$\in Dim$
foreach $Failure_{Direction}$ **do**
  determine *failure* and *no-failure* boundary ;
  while *Simulation number* $<$ *limit* **do**
    generate ΔV_{th_1} ... ΔV_{th_6} (within boundaries);
    run *Simulation* ;
    if *Failure criteria == True* **then**
      $P_{failure} = \prod_{i=1}^{6} P(\Delta V_{th_i})$
  $P_{fail} = P_{fail} + max(P_{failure})$

We introduce a new technique called *leveraged* MPFP that calculates the MPFP for more than one direction based on cell dimensions. The concept of using cell dimensions to trade-off estimation accuracy for simulation speed is based on the notion that with increase in cell dimensions, improved noise margins lower failure probability. This will enable us to tune the estimation technique so as to reduce/increase the number of required simulations dependent on cell dimensions. The basic methodology of l MPFP is shown in algorithm 2. For each direction (out of 64) where a failure event is likely to occur, the technique searches for the MPFP based on threshold variability. We improve the computational complexity of the search algorithm by first using a 1-D approximation by generating the ΔV_{th} values in equal normalised variations and determining the boundary between a failure event and a no-failure event. Within the region between the two boundaries, the traditional MPFP technique is executed to determine the combination of ΔV_{th} values of the six transistors that maximizes the failure probability. The accuracy of our proposed technique is largely dependent on the increments of ΔV_{th} values used to determine the two boundaries. Assuming a 3σ variation, we were able to determine the

boundaries with acceptable levels of accuracy within 100 increments. As for determining the failure directions with maximum likelihood, we use data obtained *apriori* by running exhaustive simulations for a wide range of cell sizes and using a look-up table to reference based on normalized λ values of transistor dimensions.

The simulations were performed for a range of threshold deviation values for a cell designed in 22nm HP-PTM operating at 0.7V with dimensions obtained from [1, 61]. The results are presented in figure 6.5. At low variation corners where failure probability is extremely low, both MPFP and LMPFP estimate failure probabilities that is an order of magnitude different from SPICE estimates. This is primarily because both techniques

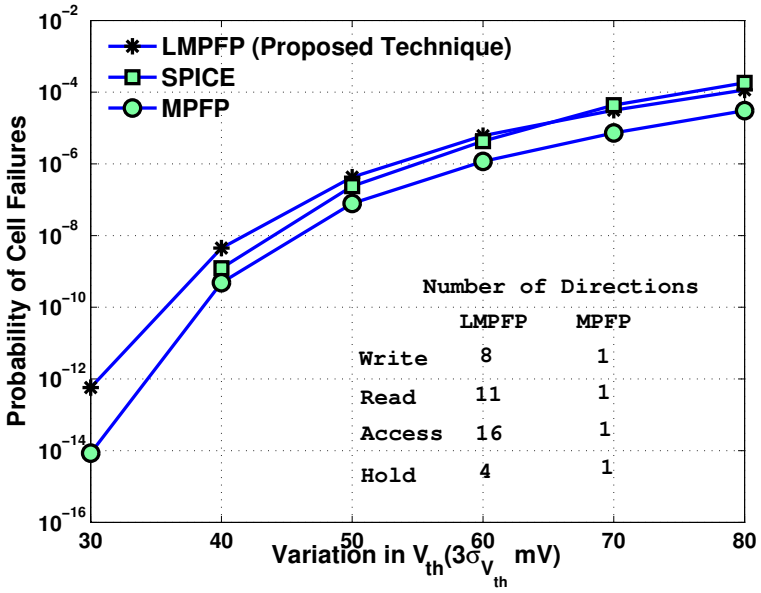


Figure 6.5: Failure probabilities under varying threshold voltage deviation

derive the probability density function (PDF) of only one point in the variation space with maximum probability. However, the presence of more than one failing point with near-identical probability can reduce the effectiveness of our technique. The number of directions where failures are more probable to occur as obtained by our technique is shown within the figure. As the threshold variation increases, LMPFP is able to perform on par with traditional Monte-Carlo simulations while MPFP still suffers from such primitive issues. For the case of failure probability $\approx 10^{-5}$, the simulations were run for more than 3 hours in HSPICE compared to the 31 seconds it took for our technique to complete achieving 350X speed-up. It should also be noted that there are several

other techniques such as importance sampling and probability collectives that build on top of MPFP to improve the accuracy of the measurement tremendously. Our technique in conjunction with the INFORMER framework is designed to help derive a trend in memory robustness when considering multiple design choices (across different levels of abstraction) simultaneously and is not designed to be a memory robustness sign-off tool.

6.4 INFORMER: An Integrated Framework for Early-Stage Memory Robustness Analysis

INFORMER has been designed as a generic tool capable of guiding designs across different process nodes. It has been written in *Python* on top of SPICE. The tool supports a wide variety of input parameters describing the different memory components with varying levels of detail. The primary inputs of the configuration file include :

- Technology specific transistor models - *SPICE* libraries
- Architecture-level memory specifics - number of rows and columns/array, number of redundancy columns, total size of memory, read-out width and ECC strength
- Bitcell specifications - maximum wordline pulse-width, SA enable time, SA offset, Hold voltage
- Parametric variability information - threshold, channel-length & width variability, Temperature range
- extra parameters needed for soft-error simulations like particle flux and pulse-width of strike

As shown in figure 6.6, at the heart of the system lies the engine that is completely responsible for all tasks ranging from characterizing SRAM cells to designing whole memory macros. The simulation and statistics generation is done in two phases. In the first phase, based on user inputs, a representative memory critical path is generated in a *spice* format. Either pre-written templates or user provided netlists can be used for this purpose. The netlists generally contain all components of the memory including write buffers, column and row decoders, multiplexers, precharging circuitry and sense amplifiers. The advantage of generating an approximate critical path from individual components is the ability

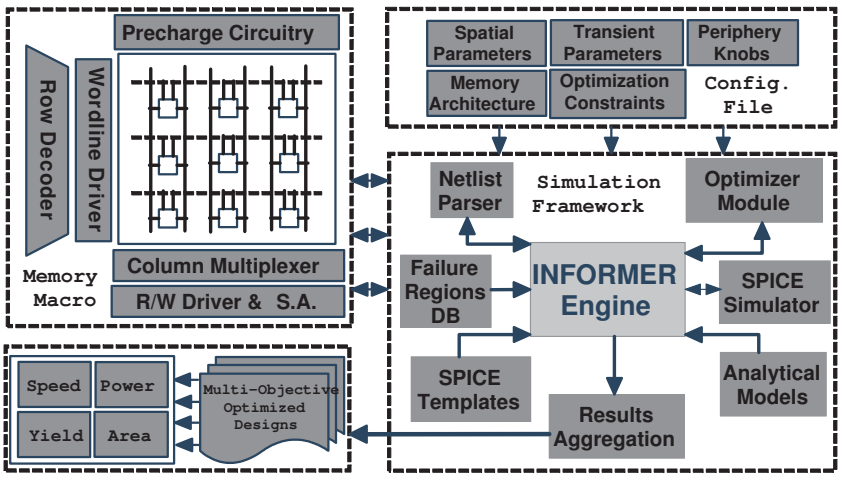


Figure 6.6: INFORMER design flow.

to port designs from one technology to another and rapidly estimate the memory-wide statistics to better understand the changes in the design. While the design of each peripheral circuitry can change, their input/output ports are fixed to ensure coherency in simulation results. Therefore, their interfaces and certain timing related specifications are static. Once the memory macro is generated, all the different failure criteria are evaluated based on input specifications. The influence of the peripheral circuitry on the failure probability is captured through the use of one or more *tuning-knobs* that control the timing and bias sources of the memory array. The *bias* knobs can typically control all voltage sources connected to the write, read and precharge lines while the *timing* knobs can limit read/write and sense amplifier enable times. We estimate the failure probabilities while tuning the design knobs like wordline pulse width, sense-amplifier offset, standby voltage and temporally varying parameter like temperature. The obtained results are presented in figure 6.7. The values are normalized to the minimum failure probability observed. Note that in some results the increase in failure probability is not necessarily an increase in the overall failure probability. Each failure mode exhibits different trends for different tuning parameters. We have highlighted only those failure modes that are the most affected by such design parameters.

In the second phase of simulation, the estimates obtained from the simulation are fed into black-box analytical models that provide accurate memory-wide statistics of power, yield and area overhead. We have also included an optimization layer that can take control of the simulations independently and guide design choices autonomously. For a minimum number of input constraints, it can optimize the cell specifications subject to limits set on the input parameters. Another unique feature of our tool is the seamless

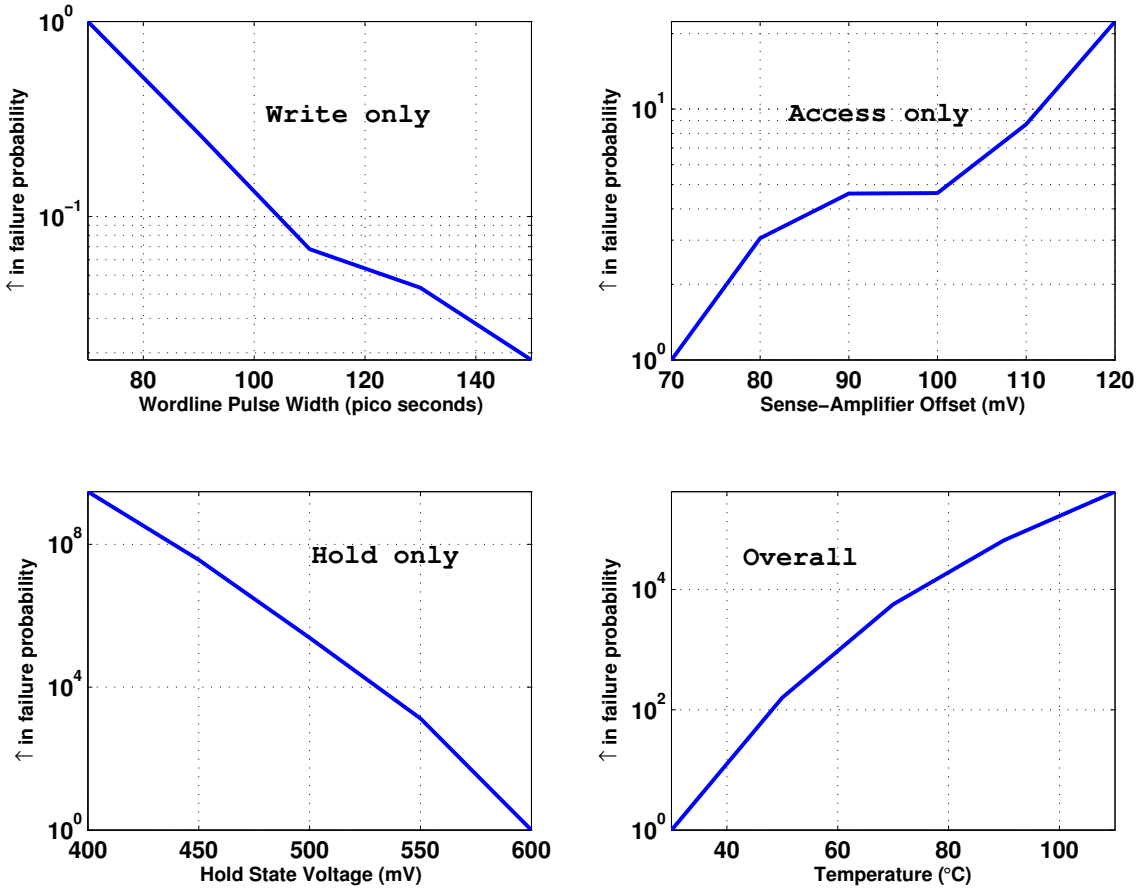


Figure 6.7: Estimated increase in failure probabilities as a function of varying control knobs. The values are normalized to the lowest failure probability.

integration with a state of the art soft-error rate estimation tool, TFIT [2]. We evaluate the tolerance of the cell to soft-errors through a technological process characterization database allowing the generation and evaluation of any transient currents that may be induced by single events. This database built from prior silicon test data is primarily a response model that tracks the effect of transient current pulses on the individual transistors of the cell. The cell response is then evaluated w.r.t. these events in a given working environment. In its current iteration, TFIT does not support *variability* specific control statements that many simulators accept. A custom wrapper-like *TFIT interface* interacts with TFIT and provides it with the necessary configuration files and a modified cell netlist with the necessary variation information injected.

All simulations have been performed on a 6-core machine running at 3.0 GHz and using 32GB of main memory. INFORMER supports multi-core processing and for each run of the input configuration, six operations (4 failure modes, soft-error, power/latency simulation) are performed in parallel to further enable rapid design space exploration.

6.5 Use Cases of INFORMER

6.5.1 Constraint based optimization

Figure 6.8 shows the percentage yield as a function of the spread in leakage power and cell area under the impact of spatial variations. The analysis was performed for a lot of 729 cells with different transistor dimensions. Each bubble in the figure represents the measured (X-μ)/σ value of leakage (and cell area) and the corresponding yield when using a full 32KB memory composed of such cells. The area estimates were derived from the layout-dependent formula proposed in [77]. It was observed that the 1-σ deviation of leakage and cell area is 51.2% and 5.4% respectively. From the figure, it should be clear that the yield by itself does not display any particular trend with respect to leakage and cell area. Thus the problem of cell sizing can be viewed as a non-linear optimization with a set of input constraints. We defined the constraint problem as, *minimize cell area* subject to *yield* $>$ *90%* and *std. dev. of leakage* $<$ *-1.2*σ. For this case, the optimized cell area was 3.2% lower than that of the mean close to the nominal design. When the constraint on *std. dev. of leakage* was $\leq$ 0, the cell was further optimized and the obtained area estimates were 8.25% lower than the mean.

6.5.2 Column redundancy limits on yield

Redundancy has been proposed as a low-cost technique to improve yield and reduce test cost by replacing defective (hard faults) rows/columns with redundant ones. More recently, redundancy has been used to improve the functional yield at low supply voltages where the functional margin is very poor [116]. In order to achieve maximum yield through the use redundancy, it is important to know exact fault locations *apriori*. That way, redundant columns (rows) can be allocated to the column (rows) with most number of faults with significant increase in yield. Figure 6.9 shows the estimated yield as a function of the number of redundant columns for five different cases of supply voltage.

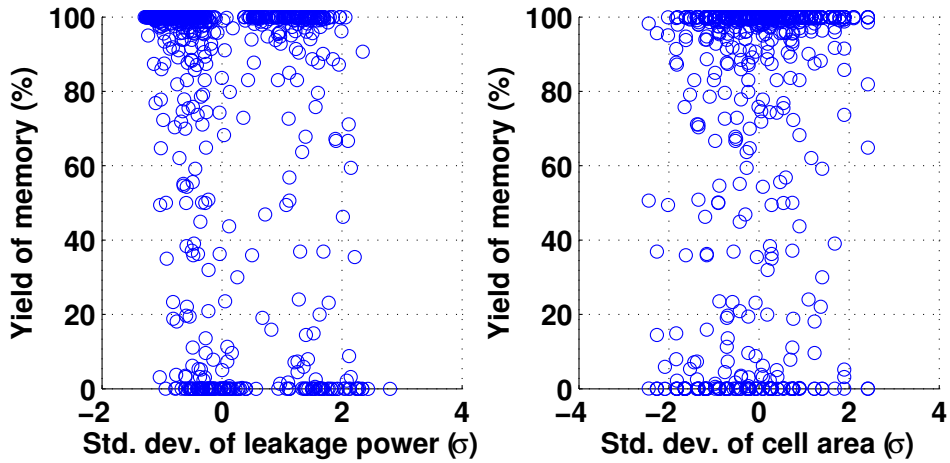


Figure 6.8: Impact of cell area and power on yield

The redundant columns are assumed to be hardened against parametric failures. It is well known that supply voltage scaling is the most effective technique for power savings. However, beyond a point it becomes a critical reliability constraint and no longer is energy efficient due to area and power-overheads of the recovery mechanisms (redundant columns). There is approximately a 33% reduction in power when lowering supply voltage from 700mV to 600mV and the area overhead needed to ensure 90% yield is nearly 47% which makes it a non-viable option for dense embedded memories. Instead, a small overhead in the cell area can be incurred to achieve lower failure probability and still be able to operate at lower voltages.

6.5.3 Variability aware soft-error rate estimation

As the soft-error rate is exponentially dependent on the critical charge, the impact of process variations on soft-error rate is not trivial. Previous studies have highlighted that gate-length variations has the most impact on the critical-charge and can be as high as 80% [34]. We use a 45nm cell designed in conjunction with a 45nm CMOS database. The choice of technology node for this study was prompted by the lack of accurate design files for advanced nodes. Figure 6.10 shows the distribution of the SER (FIT/1Mb) of memory designed using 3 different cells (C1, C2 and C3). The cell dimensions were obtained from [116] and scaled accordingly with C1 being the smallest and C3 the largest. C2 and C3 are 23% and 46% larger respectively compared to C1. We notice here that the spread is very small unlike leakage power or access latency. This is mainly because the effects

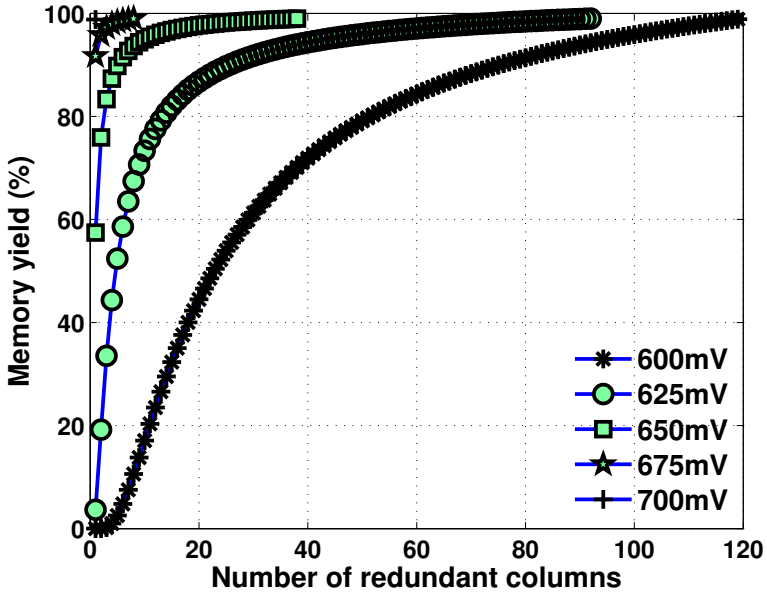


Figure 6.9: Yield w.r.t. the number of redundant columns in a 256x128 Array

of independent variation of channel length and width on the critical charge are quite different and they eventually cancel out each other. Further, as the soft-error phenomena is observed only in the hold states, the access transistors can be completely ignored in the analysis. It was shown in [116] that the probability of joint occurrence of a soft-error and parametric failure in the same memory word is extremely low. Therefore, it is sufficient if both types of errors are treated in an orthogonal manner.

In the next part of the chapter, using INFORMER, we analyse in detail the effectiveness of *body-biasing* **BB** and *wordline boosting* **WLB** on **failure prevention** in low-voltage SRAM cells. Under the effects of process variations where corner-case scenarios can still make a large number of cells to fail, we detail the importance of on-chip mechanisms like *error correction codes* **ECC** and **redundancy** on **failure reduction**. We then consider the impact of combining the two class of techniques on improving the overall parametric cache yield through failure prevention and reduction. Throughout the remainder of the chapter, when necessary, we highlight the energy/performance/ quality trade-offs associated with each independent technique and when combined.

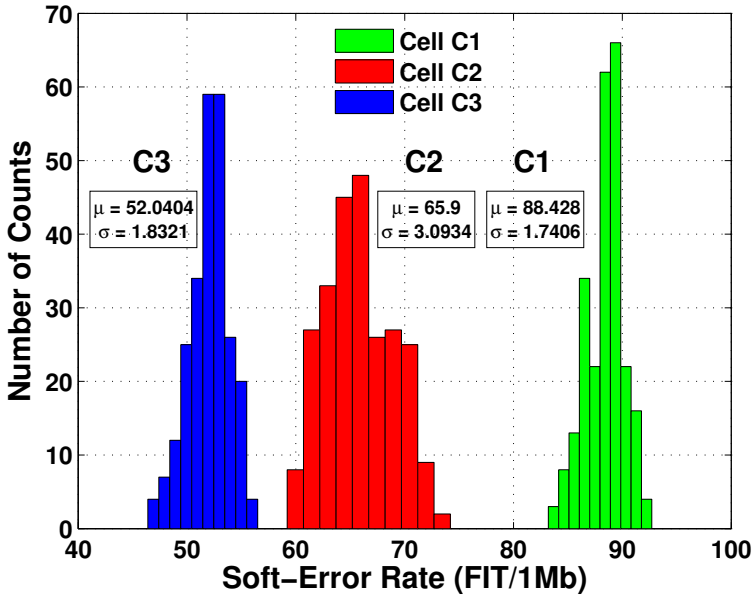


Figure 6.10: Soft-error rate (FIT/1Mb) of memory designed using differently sized SRAM cells under the impact of process variability.

6.6 Proactive R/W Assist Techniques

Assist techniques rely on modulating one or more voltage sources (V_{dd},V_{ss},V_{body} & V_{wl}) connected to the SRAM cell for improving failure probability [75, 83, 86]. There are also other techniques that vary the pulse-widths of one or more enable signals that govern the regular access to the cell [58, 59]. While most target improving the overall cell failure probability, there are techniques which specifically target either read or write margins. BB has shown to improve overall failure probability while WLB improves write margins and degrades read margin.

6.6.1 Adaptive Body Biasing

In body-biasing, by modulating the $V_{body-source}$ (V_{bs}) of the transistor, the threshold can be varied dynamically. BB has been proposed as an effective post-silicon technique to improve parametric yield. Forward BB (FBB with $+V_{bs}$) reduces delay by reducing threshold voltage and reverse BB (RBB with $-V_{bs}$) reduces leakage by increasing threshold. Adaptive BB (ABB) involves trading power for performance by applying an optimal bias based on threshold corner. In [78], ABB is used to improve the parametric yield of SRAM arrays by 8-25%. Figure 6.11 shows the impact of body biasing for a range

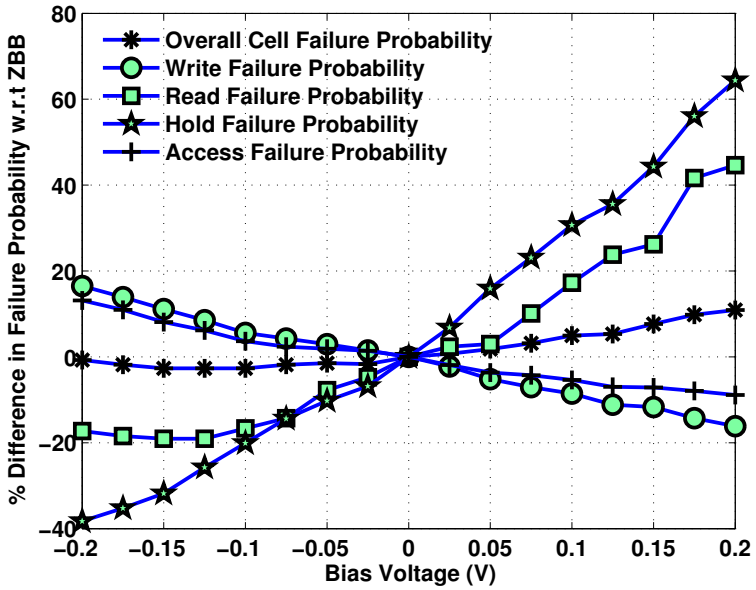


Figure 6.11: Impact of body biasing on cell failure probability

of BB values from a minimum RBB of -200mV to a maximum FBB of 200mV in steps of 25mV. We measure the percentage difference in failure probabilities with respect to zero BB (ZBB). BB is applied to all the transistors of the SRAM cell. The presented results are in line with those in [78]. With FBB, the consequent reduction in threshold voltage has multiple implications. a.) The *on-current* of the cell increases, speeding up the charging/discharging into/through the bitlines reducing the overall number of access and write failures. b.) However, with this increase, the amount of charge that flows into the storage node also increases creating a narrow window for potential flip during read mode. Similarly, with application of RBB (increase in threshold), the voltage surge (during a read) at the storage node reduces and also increases the trip-point thereby improving the overall read stability [78]. The cell becomes slower with RBB increasing both write and access failures. For the case of hold failures, both RBB and FBB have the same influence because they affect different cell parameters in different ways. The minimum reliable hold voltage is dependent on the relative strengths of the pull-up and pull-down transistors. The variation in minimum hold voltage reduces with increasing threshold voltage (as leakage reduces) [78]. As a result, with RBB the hold failures tend to reduce. At high-V_{th} corners, the impact of joint failures is noticeable and that can be observed with increase in read failures beyond -150mV where write failures are more pronounced. On the other hand, at low-Vt (-σ) corners the number of write and access

failures are minimum. At low-V_{th} corners, despite read failures being high, by adjusting the wordline pulse-width to be small, such failures can be avoided. This will ensure that despite increasing *on-current*, the time WL is high is low enough to not cause a read upset. The minimum pulse-width should be higher than the time taken to pull-down node storing a '1' during writing a '0'. The overall failure probability is not influenced much by BB because of the opposite effects of RBB and FBB with (Hold,Read) and (Write,Access) set of failures that eventually lead to one set canceling out the other.

6.6.2 Wordline Boosting

Boosting wordline gate voltage improves the write margin while reducing read margin. However, if carefully chosen, by suppressing the WL voltage at low-Vt corners where read currents are typically large, the overall cell failure probability can be improved. In [83], the technique of *selective wordline boosting* has been envisaged. Only the WL voltage of failing lines is boosted while the remaining are strobed at normal V_{dd}. Although the technique does result in an increase in leakage, it is marginal as only gate leakage of the pass transistors increases. While random variations cause failed cells to be distributed across the array, it is safe to assume that systematic variations will increase the probability of adjacent cells failing and techniques like *selective wordline boosting* can exploit this spatial locality to ensure minimal dynamic power overheads.

In figure 6.12 we evaluate the impact of boosting for different WL voltage (V_{WL}) on the failure probabilities of all failure mechanisms. We measure the difference in failure probability with respect to zero wordline boosting (ZWLB or when $V_{WL} = V_{dd}$). It is interesting to note that unlike ABB, where the optimal failure probability is always near ZBB, here with increase in V_{wl}, the failure probability reduces. This is because of the tremendous impact of this technique on reducing write failures and the subsequent reduction in joint failures. We see that for $V_{WL} > 1.2V_{dd}$, further reduction in overall failure probability is minimal because the impact of leakage dominates the impact of high read currents in this domain. As a result, the consequent reduction in write and access failures is offset by tremendous increase in read failures and hold failures.

While BB and WLB can help improve cell failure probability, under the impact of process variations their behaviour can be completely indeterministic without any post-silicon tuning. Therefore, it is imperative to provide solutions that can guarantee definite results under all scenarios. We discuss two such techniques in the following sections.

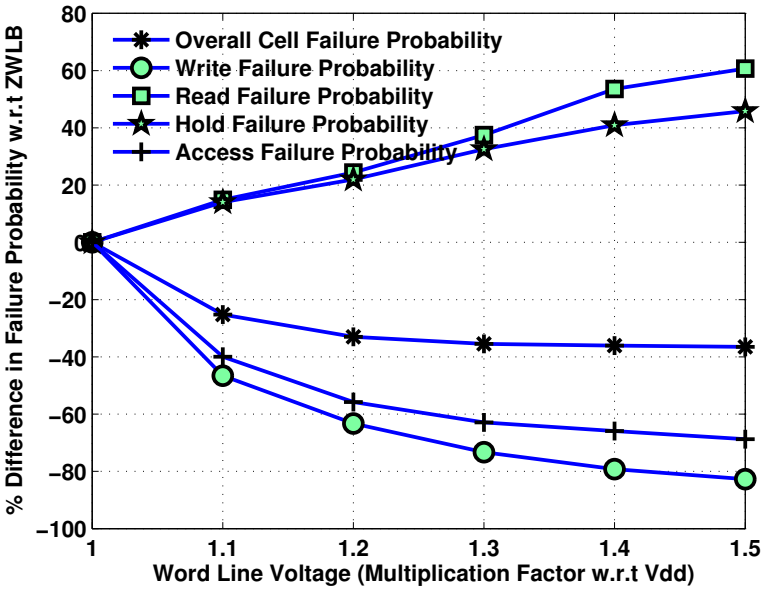


Figure 6.12: Impact of wordline boosting on cell failure probability

6.7 Reactive Techniques

6.7.1 Error-Correcting Codes (ECC)

In our analysis, we use the standard SEC/DED Hsiao codes because of the low complexity involved in implementation of encoding/decoding circuitry. The number of check bits required for any given word (N bits) is (log_2N+2). It was shown in [91] that, with reducing word size, the area overhead (check bits + circuitry) reduction is very less. Figure 6.13 shows the number of failures observed for varying word sizes. We have measured for three different scenarios - nominal, average and worst-case process variations. With increasing word size, the number of observed failures also increase. The amount of threshold voltage variation assumed for nominal, average and worst-case scenarios are 6%, 12% and 18% respectively. The amount of variation assumed is higher as we use 22nm technology for this study. For nominal case variations, we notice that at the most there are 2-failures per word for any given size. For average case scenario, for word size upto 128 bits, the maximum number of failures is 2-bits. However for 256-bit words, 79% of lines have 3 failures and 21% - 4 failures. Moving over to worst-case scenario, we notice that for 32-bit words, there are equal number of 1-bit and 2-bit failures. As ECC operates on worst-case lines, the system will have to be designed with either 2-bit correction capabilities

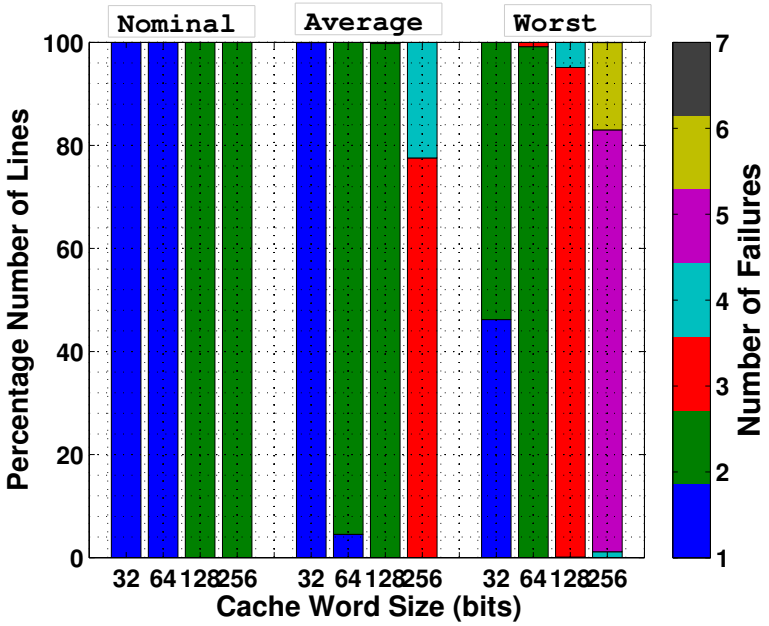


Figure 6.13: Number of Failures observed for varying word size

(DEC-TED codes) or partitioning the memory word into multiple segments. However, partitioning is not always successful as both failures can be observed in the same segment and only one can be corrected. Under worst case scenario, for 64 and 128 bit word sizes, only a small portion of the lines have greater than 3 failures. For 256-bit sizes, the system will have to be designed with TEC-QED codes which are far more complex to implement and incur huge area-overhead which is not permissible in L1 caches. Later we show that such conflicts can be avoided by using hybrid techniques.

6.7.2 Redundancy

Without loss of generality, it is assumed that the redundant columns are hardened against parametric failures. A redundancy multiplexer that is used to drive the bits in defective columns into the correct location in i/o lines can be programmed with *one time programmable*(OTP) fuses post-manufacturing [116]. We evaluate the fault coverage with redundancy for varying number of redundant columns/array as shown in figure 6.14. It can be seen that with just 2 redundant columns, more than 99.7% fault coverage can be achieved for nominal and average-case scenarios. However, for worst-case scenario, there is a steep drop in the coverage. In order to achieve required yield target, atleast 8 columns/array are required. This is because, unlike the previous two process-scenarios,

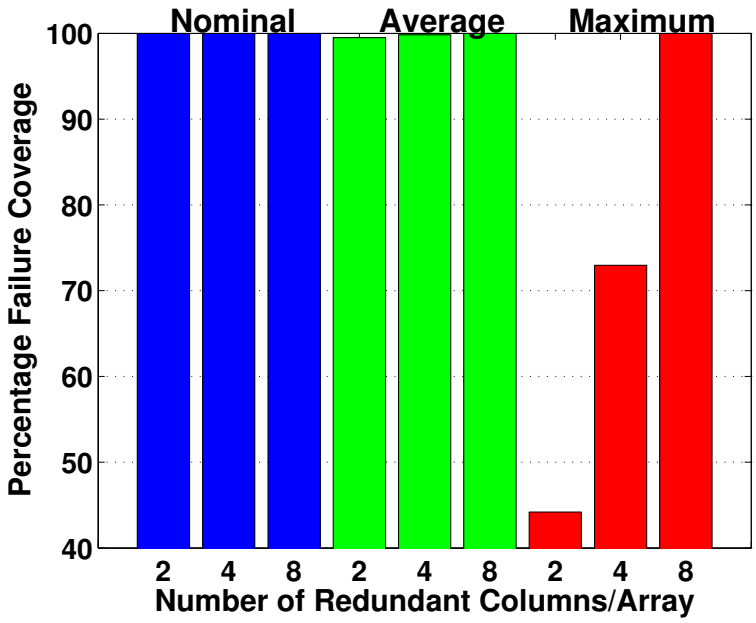


Figure 6.14: Impact of redundancy on fault coverage

the number of failures distributed in a single word varies and low-complexity redundancy can correct only one failure per row even if there is an extra unallocated column. Otherwise, expensive data steering mechanisms will have to be in place. Under the effects of random variations, as failures can be distributed across the array, it is hard to develop specific heuristics for allocating a defective column with maximum number of failures to a redundant column.

Orthogonally, researchers have also analyzed the trade-off between using larger cells (with lower failure probability) and the joint use of smaller cells with redundancy and ECC [116]. The area overhead due to ECC or redundancy is compensated for by using smaller footprint SRAM cells. The technique reduces area-overhead by as much as 27% with 90% yield while lowering $V_{dd_{min}}$ to 600mV. However, for technologies beyond 45nm, the area of an 6T-SRAM cell required to maintain sufficient noise margins is large enough to have paved the way for more area-frugal 8T-SRAM cells. In the next chapter, we discuss how 8T-SRAMs can be leveraged to improve noise margins both under the impact of parametric variations and NBTI.

6.8 Hybrid Yield Enhancement Techniques

First, we determine the extent to which failures can be reduced by combining WLB with the two reactive techniques (ECC and redundancy). We choose WLB over ABB for this particular study because of the increasing importance of substrate noise and resistance coupled with routing overheads due to body-biasing in advanced technologies.

Note here that in this particular implementation of ECC, only 1-bit can be corrected among 256-bits. Although this is a pessimistic assumption of the capabilities of ECC, level-1 caches unlike last level caches cannot afford a huge area penalty and we show that this is more than sufficient to achieve the desired yield.

$\mathbf{V}_{wl}$	**WLB**	**+2C**	**+4C**	**+2C+ECC**	**+4C+ECC**
$\mathbf{1.1V}_{dd}$	4.4	43.2	72	61.3	86.8
$\mathbf{1.2V}_{dd}$	78.6	96.4	98.6	98.7	99.9
$\mathbf{1.3V}_{dd}$	88.1	98.8	99.5	99.6	100
$\mathbf{1.4V}_{dd}$	91.6	99.3	99.8	99.8	100
$\mathbf{1.5V}_{dd}$	91.1	99.3	99.8	99.8	100

Table 6.1: Percentage reduction in failures for a combination of WLB and reactive mechanisms (C - Column Redundancy)

Looking at table 6.1, WLB by itself is responsible for reducing most failures. Therefore, redundancy and ECC will have to correct only the remaining failures. At V_{wl}=1.1V_{dd}, the reduction in failures increases from a mere 4.4% when employing only WLB to a massive 43.2% when combining WLB with 2-Column redundancy. So it is only intuitive to think that redundancy can improve the performance of WLB by 10X. However, this is not the case as redundancy is designed to fix only one failure in a segment (where a number of consecutive columns are grouped into one segment and the number of segments is equal to the number of columns in the array divided by the number of redundant columns). Moreover, we do not generate fault maps and we calculate only the distribution of failures for a given word size. Over here, we assume that it is possible to derive accurate heuristics for allocating most-failure prone segments to redundant columns to maximize failure reduction. With reduction in failure probabilities with WLB, the number of failures also reduce and there is only that much that redundancy can do to recover the remaining faulty columns. As shown in figure 6.15, for all combinations of proactive and reactive

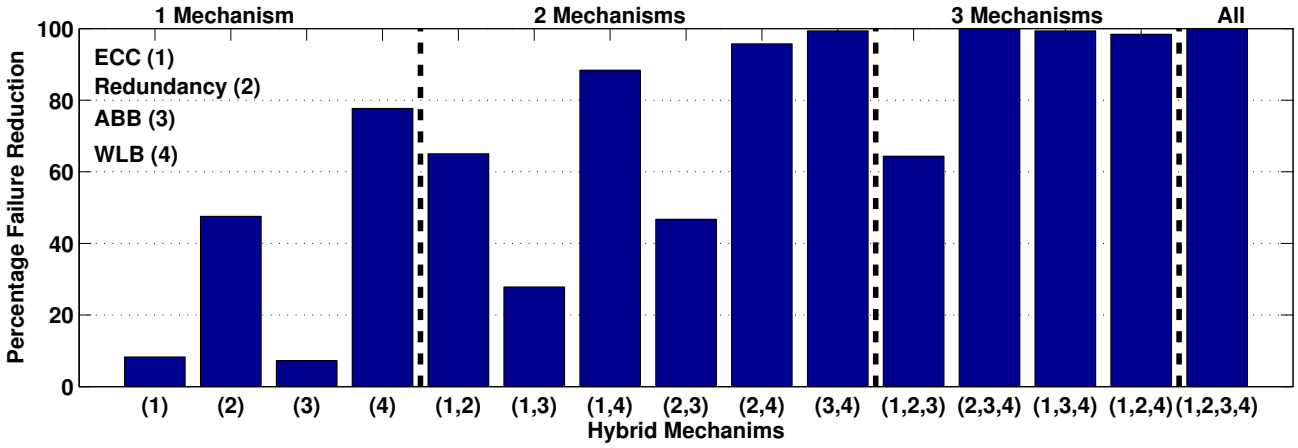
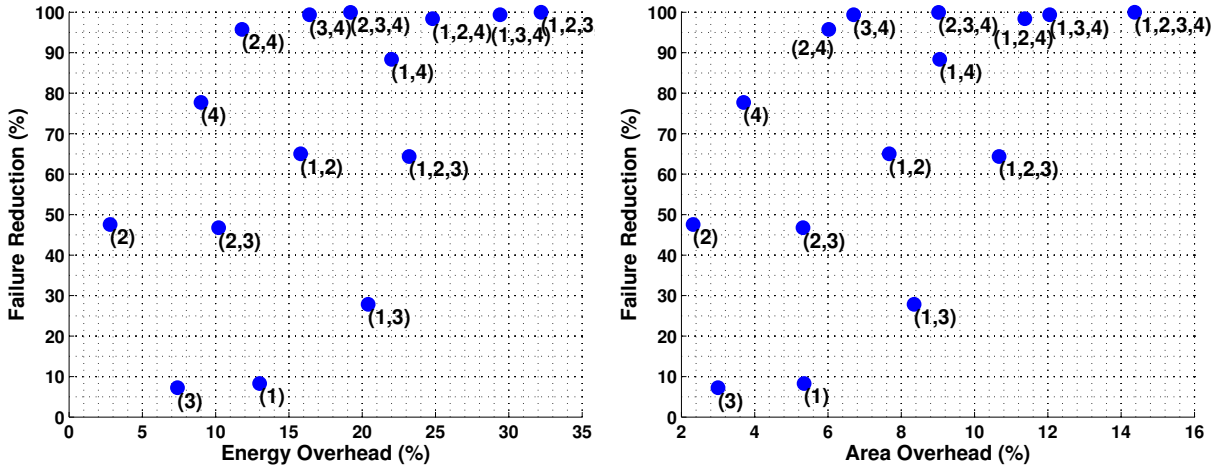


Figure 6.15: Hybrid yield enhancement techniques

techniques we estimate the failure reduction capabilities. For every 256X256 array, it is assumed that there are 2 redundant columns each operating on a 128-bit segment. The Wordline is boosted to $1.2V_{dd}$ and we employ RBB (for leakage reduction and improving read and hold margins) at -100mV. In order to avoid the problem of multiple failures in the same segment, we first allocate redundant columns and then use ECC if needed. When choosing between one of the four techniques to be used independently, WLB is the clear choice. On moving to a combination of 2 techniques, both (1,4) and (2,4) perform nearly as good as each other. However, (2,4) will be preferred over (1,4) as there is a certain extra delay in generating the check bits associated with ECC and this is critical for L1 caches especially. With redundancy, it is only a matter of block rearrangement performed post-manufacturing. In order to implement wordline boosting, extra power gating PMOS transistors that can select between V_{dd} & $V_{dd}H$ are required. From the figure 6.16b, the associated area-overhead is observed as 3.7% of the L1 cache (32KB) memory. Although the technique does result in an increase in leakage, it is marginal as only gate leakage of the pass transistors increases. The corresponding increase is 9% dynamic energy as shown in figure 6.16a. When we used ABB, the optimal bias was selected to maximize failure probability reduction for a given energy overhead. At high-V_{th} corners where there is good roll-off, it is better to use RBB for leakage power savings. Similarly, at low-V_{th} corners, further reduction in energy by applying RBB is limited by multiple short-channel effects, so we chose to use FBB. When using body biasing in memory arrays, a major concern is the presence of substrate noise. As a result, n-wells of whole arrays need to be isolated from each other to improve the immunity. The cost of implementing charge pumping/resistor tree circuits, routers and buffers is less than



(a) (b)

Figure 6.16: Improvement in parametric yield at the cost of increasing area and energy overheads when using hybrid techniques.

2-3% of die-area in 45nm [94]. Also, there is a certain transition latency incurred when switching from zero-bias state to a biased state due to the presence of substrate resistance.

When considering the combination of three techniques, the first thing we notice is that the addition of ABB to (1,2) does not result in any improvement. For worst-case process scenario where multiple failures in the same segment is expected, the addition of SEC/DED ECC to proactive techniques will not offer better robustness. It is therefore preferable to use redundancy in such cases. The incurring area overhead (technology agnostic estimation) for correcting one failure in every 32b when using ECC is 22.04% against 5.35% in correcting one in every 256 bits [91]. Since only the *read accesses* lies on the critical path of the microprocessor pipeline, the additional overhead in delay due to ECC is from the decoding circuitry. There is 40% increase in delay when moving from word sizes of 32b to 256b. This increase corresponds to approximately 7 gates-delay. While redundancy does not incur any additional overhead on delay, the extra area required for adding 8 column per array is less than 0.13% of die area. Due to the data steering mechanisms required to select the appropriate redundant columns, the additional energy overhead is estimated to be 2.8%. A major advantage of using reactive techniques like ECC and redundancy over read/write assist methods is the ability to recover from hard failures caused by manufacturing defects.

The benefits of using proactive techniques over reactive is that they can be enforced in a very fine-grain manner and they can help the system adapt to temporal variations over

time. However, reactive techniques are straight forward to implement with proven design methodologies and their in-situ behaviour is entirely deterministic. Thus it should be clear that the best solution is not to choose between reactive or proactive but to carefully analyse their associated trade-off and implement a system where they can co-exist by complementing each other.

6.9 Summary

To help designers evaluate memory metrics rapidly and accurately in a transparent manner, a novel integrated framework encompassing statistical SPICE-level simulations of the memory macro and back-of-the-envelope analytical models has been has been proposed in this chapter. Further, using the developed tool, we study the joint impact of failure prevention and correction techniques in improving cache parametric yield. The technique is shown to offer better quality-energy-area trade-offs when compared to their standalone configurations.

7

Conclusions and Future Work

7.1 Summary of Contributions

With continuous technology scaling, meeting power and performance budgets is severely inhibited by parametric variation of process and environmental parameters. Some of the most important problems include the exponential increase in leakage power, fundamental reduction in performance and increasing susceptibility to errors and failures. With every successive process generation, reliability is becoming a key design constraint that is preventing safe and correct operation of the underlying silicon. This is of particular concern for components like embedded memories that account for significant portion of silicon real-estate in modern day microprocessors. Current day solutions employing schemes like frequency binning or guard-banding are no longer effective as they can mitigate the advantages of device scaling. Further, from a performance and area perspective, the overheads are large enough to be considered viable for commercial designs. This thesis takes cognizance of such challenges and proposes several new techniques at different levels of design abstraction to counter the negative effects of parametric variations in embedded memories.

In Chapter 3, we proposed two new techniques to estimate propagation delay and energy of the memory critical path in the presence of spatio-temporal variations. Using simple circuit-level simulations of small blocks that constitute the memory path, we derive a statistical energy/delay model using a combination of clustering and regression. The advantage of using a hybrid analytical-empirical approach is to improve estimation accuracy whilst allowing the models to be scalable with increasing number of design parameters. The multivariate regression-based models are then employed to analyse few circuit optimizations like dual-V_{th} assignment and supply voltage minimization. Since embedded memories are a very important component, we strongly believe these models can be coupled with architectural models for early-stage design space exploration.

In Chapter 4, with the aim to complement design-level optimizations (proposed in chapter 3) with run-time support, we have designed a novel post-silicon tuning hardware based on dynamic fine-grain body-biasing (DFGBB). The hardware is composed of a eDRAM-based canary sensor that tracks array-wide latency and leakage variation at run-time. This is made possible by gradient sensing of the access and retention time of the eDRAM cell. The measurement then helps classify memory arrays at run-time based on latency and leakage profiles. Based on these profiles stored in LUTs, the DFGBB module generates an appropriate body-bias that trades-off leakage power for performance. The mechanism is able to improve access speeds while simultaneously reducing idle power by exploiting access patterns exhibited by cache memories (locality of reference).

Chapter 5 highlights two of the most important challenges, retention time and soft-error susceptibility, that is inhibiting widespread integration of eDRAM-based gain cell memories in data caches. Improving the retention time negates the need for frequent refreshes lowering dynamic power tremendously. Further, a large retention time helps the storage node stay above critical charge for a sufficiently large period before the data is evicted or considered *dead*. The proposed 4T-eDRAM variant improves the retention time by introducing an additional transistor that suppresses sub-threshold leakage by being super cut-off and reverse-biased in the hold-state which enables the storage node to stay at a particular voltage level for a longer amount of time. The additional write delay overhead due to the presence of the extra transistor is small enough to not incur system-wide performance penalty.

Finally in chapter 6, we propose a software framework to help design robust memories in future technologies. The tool relies on circuit-level characterization of failure mechanisms observed in SRAM memories and applies it on architecture-level specifics to gauge

the impact of low-level failures on higher-level metrics. Using the developed tool, the effectiveness of a new class of hybrid techniques in improving cache yield through failure prevention and correction is evaluated. Proactive read/write assist techniques like body-biasing (BB) and wordline boosting (WLB) when combined with reactive techniques like ECC and redundancy are shown to offer better quality-energy-area trade offs when compared to their standalone configurations. Proactive techniques can help lower $V_{dd_{min}}$ (improving functional margin) for significant power savings and reactive techniques ensure that the resulting large number of failures are corrected (improving functional yield).

7.2 Future Work

7.2.1 NBTI Tolerance in Caches (On-going research)

(Dis)graceful degradation of reliability due to NBTI in logic circuits can be offset to a large extent by carefully up-sizing transistors during the design stage [55]. However, such a technique cannot be applied to on-chip memories due to the stringent constraints set on area and power. Even at nominal voltages, NBTI can further reduce bitcell noise margins increasing the probability of a bit-flip leading to a failure. Such failures result in a reduction in cache functional yield (total addressable memory space) across the lifetime affecting key processor-wide metrics such as instructions/cycle (IPC) and performance/watt. As a part of on-going research effort, we are studying the impact of cache access patterns on lifetime performance and reliability degradation. Previous studies have highlighted that more than 75% of the time, logic bit value '0' is stored in the cells [109]. This is of great concern as it can imbalance the rate of degradation across the two halves of the symmetric SRAM structure. It is possible to ensure that both halves recover and degrade at the same rate by making sure both halves store the same logic values for the same amount of time. Towards, this end we are evaluating a novel proposal that combines a read-modify-write approach with invert operation to lower NBTI degradation and recover from column half-select failures. We leverage area and power frugal eDRAM cells to store the state information of the changing *bit patterns*. Initial results indicate our technique can be completely hidden from program execution wherein data modifications are always effected after a *write access* and before a *read access*.

7.2.2 Ultra Low Power Operation

As we embrace the near-threshold computing domain, novel techniques are needed to cope with multi-bit errors. Also, existing CAD algorithms to study the failure rate will have to be modified to take into account changing device characteristics as we scale the threshold voltage. Existing yield-aware cache architectures will not be scalable as they are targeted towards handling different types of errors and failures in an orthogonal manner. Cross-layer resiliency would help higher-level designers have better cognizance of underlying failure mechanisms and enable them to design more modular approaches that can tolerate the number of errors expected in this operational region.

7.2.3 Emerging Memory Technologies

With billion transistor processor designs, system designers are calling for much more tighter integration between on-chip SRAM and off-chip DRAM memories. As the need for memory capacity and bandwidth continues to increase at a rapid pace, current day DRAM solutions will not be able to provide the much needed performance growth whilst maintaining the benefits of technology scaling. In this regard, we would also like to study and propose feasible main memory alternatives based on emerging resistive memory technologies such as *Phase Change Memories* (PCM), *Magnetoresistive* RAM (MRAM) and *Memristors*. While current research has been targeted towards development of pertinent solutions for use in performance critical designs such as servers, we would like to study the applicability and usability of such memory technologies in emerging ultra-low power platforms such as wireless sensor nodes and energy harvesting systems.

In this dissertation, we have focused primarily on embedded memory structures and studied the impact of spatio-temporal variations on memory yield. Going forward, we would like to investigate the impact on other processor structures as well.

7.3 Publications

- *Circuit Propagation Delay Estimation through Multivariate Regression-Based Modeling under Spatio-Temporal Variability* [41]
  S.Ganapathy, R.Canal, A.Gonzalez & A.Rubio
  Design, Automation & Test in Europe Conference ***(DATE'10)***
- *MODEST: A Model For Energy Estimation under Spatio - Temporal Variability* [42]
  S.Ganapathy, R.Canal, A.Gonzalez & A.Rubio
  International Symposium on Low Power Electronic Design ***(ISLPED'10)***
- *Dynamic Fine-Grain Body Biasing of Caches with Latency and Leakage 3T1D-based Monitors* [43]
  S.Ganapathy, R.Canal, A.Gonzalez & A.Rubio
  International Conference on Computer Design ***(ICCD'11)***
- *A Novel Variation-Tolerant 4T-DRAM with Enhanced Soft-Error Tolerance* [44]
  S.Ganapathy, R.Canal, D.Alexandrescu, E.Costenaro, A.Gonzalez & A.Rubio
  International Conference on Computer Design ***(ICCD'12)***
- *Effectiveness of Hybrid Recovery Techniques on Parametric Failures* [45]
  S.Ganapathy, R.Canal, A.Gonzalez & A.Rubio
  International Symposium on Quality Electronic Design ***(ISQED'13)***
- *INFORMER: An Integrated Framework for Early-Stage Memory Robustness Analysis*
  S.Ganapathy, R.Canal, D.Alexandrescu, E.Costenaro, A.Gonzalez & A.Rubio
  Design, Automation & Test in Europe Conference ***(DATE'14)***

List of Abbreviations

Roman Symbols

ϵ_{ox} Permittivity of gate oxide

ϵ_{Si} Permittivity of silicon

σ_{D2D} Standard deviation of die-to-die variations

$\sigma_{WID_{random}}$ Standard deviation of within-die random variations

$\sigma_{WID_{systematic}}$ Standard deviation of within-die systematic variations

L_{eff} Transistor effective channel length

V_{th} Transistor threshold voltage

1T1C One-transistor one-capacitor

3T1D Three-transistor one-diode

ABB Adaptive body-biasing

BB Body-biasing

CD Critical dimension

D2D Die-to-die

DRAM Dynamic random access memory

DVS Dynamic voltage scaling

ECC Error-correcting codes

eDRAM embedded DRAM

F_{max} Maximum operating frequency

FBB Forward body-biasing

FIT Failures-in-time measured in number of failures/10^9 hours of operation

INFORMER Integrated framework for early-stage memory robustness analysis

L Transistor channel length

LER Line edge roughness

N_a Channel dopant concentration

NBTI Negative-bias temperature instability

NMOS N-type metal oxide semiconductor device

nT-SRAM n-Transistor SRAM

PMOS P-type metal oxide semiconductor device

Q_{crit} Minimum charge required to flip the state of the logic

RBB Reverse body-biasing

RDF Random dopant fluctuation

SRAM Static random access memory

SSTA Statistical static timing analysis

ST Spatio-temporal

STA Static timing analysis

T_{ox} Gate oxide thickness

$V_{cc_{min}}$ Minimum supply voltage guaranteeing reliable operation

V_{dd} Supply voltage

V_{gs} Gate-source voltage

W Transistor channel width

WID Within-die

WLB Wordline boosting

WOR Window of reclamation

References

[1] Predictive Technology Models, http://www.eas.asu.edu/ ptm. 21, 37, 61, 83, 100

[2] iROC TFIT tool, transistor level soft error analysis., 2012. URL `http://www.iroctech.com/pdf/TFIT-datasheet.pdf`. 87, 103

[3] Mohamed Hassan Abu-Rahma. *Design of Variation-Tolerant Circuits for Nanometer CMOS Technology: Circuits and Architecture Co-Design*. PhD thesis, University of Waterloo, 2008. viii, 6, 14, 22

[4] A. Agarwal, D. Blaauw, and V. Zolotov. Statistical timing analysis for intra-die process variations with spatial correlations. In *Proceedings of the International Conference on Computer Aided Design*, 2003. 31, 47

[5] A. Agarwal, B.C. Paul, H. Mahmoodi, Animesh Datta, and K. Roy. A process-tolerant cache architecture for improved yield in nanoscale technologies. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, 2005. 26, 27, 38, 41

[6] A. Agarwal, K. Kang, S. Bhunia, J.D. Gallagher, and K. Roy. Device-aware yield-centric dual-vt design under parameter variations in nanoscale technologies. *IEEE Transactions on VLSI Systems*, 2007. 24

[7] M.A. Alam, H. Kufluoglu, D. Varghese, and S. Mahapatra. A comprehensive model for pmos nbti degradation: Recent progress. *Microelectronics Reliability*, 2007. 20

[8] David H. Albonesi. Selective cache ways: on-demand cache resource allocation. In *Proceedings of the International Symposium on Microarchitecture*, 1999. 28

[9] A. Asenov. Random dopant induced threshold voltage lowering and fluctuations in sub-0.1 mu;m mosfet's: A 3-d ldquo;atomistic rdquo; simulation study. *Electron Devices, IEEE Transactions on*, 45, 1998. viii, 14, 15

[10] J.E. Barth, Jr., J.H. Dreibelbis, E.A. Nelson, D.L. Anand, G. Pomichter, P. Jakobsen, M.R. Nelms, J. Leach, and G.M. Belansek. Embedded dram design and architecture for the ibm 0.11-µm asic offering. *IBM Journal of Research and Development*, 2002. 76

[11] M. Bhadauria, V. Weaver, and S.A. Mckee. Accomodating diversity in cmps with heterogeneous frequencies. In *Proceedings of the International Conference on High Performance Embedded Architectures and Compilers*, 2009. 25

[12] J. Bhasker and R. Chadha. *Static timing analysis for nanometer designs: a practical approach*. Springer, 2009. 22

[13] S. Bhunia and S Mukhopadhyay. Low-power variation-tolerant design in nanometer silicon. *Springer Circuits and Systems*, 2011. 16, 17, 19, 22

[14] D. Blaauw, S. Kalaiselvan, K. Lai, W.H. Ma, S. Pant, C. Tokunaga, S. Das, and D. Bull. Razor ii: In situ error detection and correction for pvt and ser tolerance. In *Proceedings of the International Solid-State Circuits Conference*, 2008. 24

[15] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, and V. De. Parameter variations and impact on circuits and microarchitecture. In *Proceedings of the Design Automation Conference*, 2003. viii, 18, 63

[16] K.A. Bowman, S.G. Duvall, and J.D. Meindl. Impact of die-to-die and within-die parameter fluctuations on the maximum clock frequency distribution for gigascale integration. *Solid-State Circuits, IEEE Journal of*, 2002. 2

[17] D. Brooks. Power, thermal, and reliability modeling in nanometer-scale microprocessors. *Micro, IEEE*, 2007. 57

[18] B.H. Calhoun, Yu Cao, Xin Li, Ken Mai, L.T. Pileggi, R.A. Rutenbar, and Kenneth L. Shepard. Digital circuit design challenges and opportunities in the era of nanoscale cmos. 2008. 6, 93

[19] Y. Cao and Zhao. W. Predictive technology model for nano-cmos design exploration. In *Proceedings of the International Conference on Nano-Networks and Workshops*, 2006. 37

[20] T.B. Chan, J. Sartori, P. Gupta, and R. Kumar. On the efficacy of nbti mitigation techniques. In *Proceedings of the Design, Automation Test in Europe Conference and Exhibition*, 2011. 20

[21] 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. *Solid-State Circuits, IEEE Journal of*, 2008. 26

[22] P. Chen, C.C. Chen, C.C. Tsai, and W.F. Lu. A time-to-digital-converter-based cmos smart temperature sensor. *Solid-State Circuits, IEEE Journal of*, 2005. 65, 67

[23] X. Chen, Y. Wang, Y. Cao, Y. Ma, and H. Yang. Variation-aware supply voltage assignment for minimizing circuit degradation and leakage. In *Proceedings of the International Symposium on Low Power Electronics and Design*, 2009. 52

[24] M. Chinosi, R. Zafalon, and C. Guardiani. Automatic characterization and modeling of power consumption in static rams. In *Proceedings of the International Symposium on Low Power Electronic Design*, 1998. 49

[25] M. Cho, J. Schlessman, H. Mahmoodi, Ma. Wolf, and S. Mukhopadhyay. Postsilicon adaptation for low-power sram under process variation. *IEEE Design and Test*, 2010. 27, 58

[26] B. Choi and Y. Shin. Lookup table-based adaptive body biasing of multiple macros. In *ternational Symposium on Quality Electronic Design*, 2007. 59, 69

[27] K. Chopra, Cheng Zhuo, D. Blaauw, and D. Sylvester. Process variability-aware transient fault modeling and analysis. In *Proceedings of the International Conference on Computer-Aided Design*, 2008. 34

[28] E. Cinlar. Introduction to stochastic processes. 1975. 36

[29] B. Cline, K. Chopra, D. Blaauw, A. Torres, and S. Sundareswaran. Transistor-specific delay modeling for ssta. In *Proceedings of the Design, Automation and Test in Europe Conference*, 2008. 31, 33, 35

[30] A. Das, B. Ozisikyilmaz, S. Ozdemir, G. Memik, J. Zambreno, and A. Choudhary. Evaluating the effects of cache redundancy on profit. In *Proceedings of the International Symposium on Microarchitecture*, 2008. 28

[31] A. Datta, S. Bhunia, S. Mukhopadhyay, N. Banerjee, and K. Roy. Statistical modeling of pipeline delay and design of pipeline under process variation to enhance

yield in sub-100nm technologies. In *Proceedings of the Design, Automation and Test in Europe Conference*, 2005. 23

[32] A. Datta, S. Bhunia, J.H. Choi, S. Mukhopadhyay, and K. Roy. Speed binning aware design methodology to improve profit under parameter variations. In *Proceedings of the Asia South Pacific Design Automation Conference*, 2006. 2

[33] Jeffrey A. Davis. Power estimation in digital circuits. Technical report, Georgia Tech. 45

[34] Q. Ding, R. Luo, and Y. Xie. Impact of process variation on soft error vulnerability for nanometer vlsi circuits. In *Proceedings of the International Conference On ASIC*, 2005. 105

[35] M. Do, P. Larsson-Edefors, L. Bengtsson, and M. Drazdziulis. Capturing process-voltage-temperature (pvt) variations in architectural static power modeling for sram arrays. *Technical Report*. 23

[36] M. Do, P. Larsson-Edefors, and L. Bengtsson. Table-based total power consumption estimation of memory arrays for architects. In *Integrated Circuit and System Design. Power and Timing Modeling, Optimization and Simulation*, Lecture Notes in Computer Science, 2004. 23

[37] J. Donald and M. Martonosi. Power efficiency for variation-tolerant multicore processors. In *Proceedings of the International Symposium on Low power Electronics and Design*, 2006. 25

[38] Y-P Fang, B. Vaidyanathan, and A.S. Oates. Soft error rate cross-technology prediction on embedded dram. In *Proceedings of the International Reliability Physics Symposium*, 2009. 85, 87

[39] T. Fischer, S. Arekapudi, E. Busta, C. Dietz, M. Golden, S. Hilker, A. Horiuchi, K.A. Hurd, D. Johnson, H. McIntyre, S. Naffziger, J. Vinh, J. White, and K. Wilcox. Design solutions for the bulldozer 32nm soi 2-core processor module in an 8-core cpu. In *Proceedings of the International Solid-State Circuits Conference*, 2011. 26

[40] R. Franch, P. Restle, N. James, W. Huott, J. Friedrich, R. Dixon, S. Weitzel, K. van Goor, and G. Salem. On-chip timing uncertainty measurements on ibm microprocessors. In *Proceedings of the International Test Conference*, 2008. viii, 19

[41] S. Ganapathy, R. Canal, A. Gonzalez, and A. Rubio. Circuit propagation delay estimation through multivariate regression-based modeling under spatio-temporal variability. In *Proceedings of the Design, Automation Test in Europe Conference*, 2010. 9, 46, 50, 121

[42] S. Ganapathy, R. Canal, A. Gonzalez, and A. Rubio. Modest: a model for energy estimation under spatio-temporal variability. In *Proceedings of the International symposium on Low power Electronics and Design*, 2010. 8, 72, 121

[43] S. Ganapathy, R. Canal, A. Gonzalez, and A. Rubio. Dynamic fine-grain body biasing of caches with latency and leakage 3t1d-based monitors. In *Proceedings of the International Conference on Computer Design*, 2011. 9, 121

[44] S. Ganapathy, R. Canal, D. Alexandrescu, E. Costenaro, A. Gonzalez, and A. Rubio. A novel variation-tolerant 4t-dram cell with enhanced soft-error tolerance. In *Proceedings of the International Conference on Computer Design*, 2012. 10, 121

[45] S. Ganapathy, R. Canal, A. Gonzalez, and A. Rubio. Effectiveness of hybrid recovery techniques on parametric failures. In *Proceedings of the International Symposium on Quality Electronic Design*, 2013. 10, 121

[46] J. Gregg and T.W. Chen. Post silicon power/performance optimization in the presence of process variations using individual well-adaptive body biasing. *Very Large Scale Integration Systems, IEEE Transactions on*, 2007. 27, 69, 83

[47] R.X. Gu and M.I. Elmasry. Power dissipation analysis and optimization of deep submicron cmos digital circuits. *Solid-State Circuits, IEEE Journal of*, 31, 1996. 46

[48] M.S. Gupta, J.L. Oatley, R. Joseph, G.Y. Wei, and D.M. Brooks. Understanding voltage variations in chip multiprocessors using a distributed power-delivery network. In *Proceedings of the Design, Automation Test in Europe Conference and Exhibition*, 2007. 19

[49] S. Herbert and D. Marculescu. Variation-aware dynamic voltage/frequency scaling. In *Proceedings of the International Symposium on High Performance Computer Architecture*, 2009. 25

[50] A. Hokazono, S. Balasubramanian, K. Ishimaru, H. Ishiuchi, T.J.K. Liu, and C. Hu. Mosfet design for forward body biasing scheme. *Electron Device Letters, IEEE*, 2006. 81

[51] H. Homayoun and A. Veidenbaum. Reducing leakage power in peripheral circuits of l2 caches. In *Proceedings of the International Conference on Computer Design*, 2007. 72

[52] S. Hong, S. H. K. Narayanan, M. Kandemir, and Ö. Özturk. Process variation aware thread mapping for chip multiprocessors. In *Proceedings of the Conference on Design, Automation and Test in Europe*, 2009. 25

[53] M.A. Hussain and M. Mutyam. Block remap with turnoff: a variation-tolerant cache design technique. In *Proceedings of the Asia and South Pacific Design Automation Conference*, 2008. 27

[54] J.Y. Jeong, G.S. Kim, J.P. Son, W.J. Rim, and S.W. Kim. Body bias generator for leakage power reduction of low-voltage digital logic circuits. In *Proceedings of the International Conference on Integrated Circuit and System Design: Power and Timing Modeling, Optimization and Simulation*. 2006. 59

[55] K. Kang, S. Gangwal, S.P. Park, and K. Roy. Nbti induced performance degradation in logic and memory circuits: how effectively can we approach a reliability solution? In *Proceedings of the Asia and South Pacific Design Automation Conference*, 2008. 119

[56] S. Kaxiras and P. Xekalakis. 4t-decay sensors: A new class of small, fast, robust, and low-power, temperature/leakage sensors. In *Proceedings of the International Symposium on Low Power Electronics and Design*, 2004. 61

[57] D.E. Khalil, M. Khellah, N.S. Kim, Y. Ismail, T. Karnik, and V.K. De. Accurate estimation of sram dynamic stability. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, 2008. 98

[58] M. Khellah, Yibin Ye, Nam Sung Kim, D. Somasekhar, G. Pandya, A. Farhang, K. Zhang, C. Webb, and V. De. Wordline amp; bitline pulsing schemes for improving sram cell stability in low-vcc 65nm cmos designs. In *Proceedings of the Symposium on VLSI Circuits*, 2006. 27, 93, 107

[59] M.M. Khellah, A. Keshavarzi, D. Somasekhar, T. Karnik, and V. De. Read and write circuit assist techniques for improving vccmin of dense 6t sram cell. In *Proceedings of the International Conference on Integrated Circuit Design and Technology and Tutorial*, 2008. 27, 76, 107

[60] C.H. Kim, Jae-Joon Kim, S. Mukhopadhyay, and K. Roy. A forward body-biased low-leakage sram cache: device, circuit and architecture considerations. *Very Large Scale Integration Systems, IEEE Transactions on*, 2005. 26, 68, 69, 70, 73

[61] J. Kim, M. Mccartney, K. Mai, and B. Falsafi. Modeling sram failure rates to enable fast, dense, low-power caches. In *Workshop on Silicon Errors in Logic - System Effects*, 2009. 27, 96, 100

[62] S.V. Kumar, C.H. Kim, and S.S. Sapatnekar. Mathematically assisted adaptive body bias (abb) for temperature compensation in gigascale lsi systems. In *Proceedings of the Asia and South Pacific Conference on Design Automation*, 2006. 48

[63] J. Lee and N.S. Kim. Optimizing total power of many-core processors considering voltage scaling limit and process variations. In *Proceedings of the International Symposium on Low power Electronics and Design*, 2009. 25

[64] Hai Li, Yiran Chen, K. Roy, and Cheng-Kok Koh. Savs: a self-adaptive variable supply-voltage technique for process-tolerant and power-efficient multi-issue superscalar processor design. In *Proceedings of the Asia and South Pacific Design Automation Conference*, 2006. 24

[65] X. Liang and D. Brooks. Mitigating the impact of process variations on processor register files and execution units. In *Proceedings of the International Symposium on Microarchitecture*, 2006. 36

[66] X. Liang, R. Canal, Gu-Yeon Wei, and D. Brooks. Process variation tolerant 3t1d-based cache architectures. In *Proceedings of the International Symposium on Microarchitecture*, 2007. 26, 62, 76, 79

[67] X. Liang, K. Turgay, and D. Brooks. Architectural power models for sram and cam structures based on hybrid analytical/empirical techniques. In *Proceedings of the International Conference on Computer-aided Design*, 2007. 23, 46, 49

[68] W. Liao, F. Li, and L. He. Microarchitecture level power and thermal simulation considering temperature dependent leakage model. In *Proceedings of the International Symposium on Low Power Electronics and Design*, 2003. 57

[69] M. Liu, W. Wang, and M. Orshansky. Leakage power reduction by dual-vth designs under probabilistic analysis of vth variation. In *Proceedings of the International Symposium on Low Power Electronics and Design*, 2004. 23, 41, 63

[70] K. Lovin, B.C. Lee, X. Liang, D. Brooks, and G.Y. Wei. Empirical performance models for 3t1d memories. In *Proceedings of the International Conference on Computer Design*, 2009. 59

[71] W.K. Luk and R.H. Dennard. A novel dynamic memory cell with internal voltage gain. *Solid-State Circuits, IEEE Journal of*, 2005. 26

[72] W.K. Luk, J. Cai, R.H. Dennard, M.J. Immediato, and S.V. Kosonocky. A 3-transistor dram cell with gated diode for enhanced speed and retention time. In *Proceedings of the Symposium on VLSI Circuits*, 2006. 59, 60

[73] S. Mahapatra, P. Bharath Kumar, and M.A. Alam. Investigation and modeling of interface and bulk trap generation during negative bias temperature instability of p-mosfets. *Electron Devices, IEEE Transactions on*, 2004. 20

[74] M. Mamidipaka, K. Khouri, N. Dutt, and M Abadir. Idap: a tool for high level power estimation of custom array structures. In *Proceedings of the International Conference on Computer Aided Design*, 2003. 23, 49

[75] R.W. Mann, S. Nalam, J. Wang, and B.H. Calhoun. Limits of bias based assist methods in nano-scale 6t sram. In *Proceedings of the International Symposium on Quality Electronic Design*, 2010. 27, 107

[76] Y. Morita, H. Fujiwara, H. Noguchi, Y. Iguchi, K. Nii, H. Kawaguchi, and M. Yoshimoto. Area comparison between 6t and 8t sram cells in dual-vdd scheme and dvs scheme. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, 2007. 26

[77] S. Mukhopadhyay, H. Mahmoodi, and K. Roy. Modeling of failure probability and statistical design of sram array for yield enhancement in nanoscaled cmos. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, 2005. 98, 104

[78] S. Mukhopadhyay, H. Mahmoodi, and K. Roy. Reduction of parametric failures in sub-100-nm sram array using body bias. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, 2008. 107, 108

[79] M. Mutyam and V. Narayanan. Working with process variation aware caches. In *Proceedings of the Design, Automation and Test in Europe Conference and Exhibition*, 2007. 27

[80] S. Nassif, K. Bernstein, D.J. Frank, A. Gattiker, W. Haensch, B.L. Ji, E. Nowak, D. Pearson, and N.J. Rohrer. High performance cmos variability in the 65nm regime and beyond. In *Proceedings of the International Electron Devices Meeting*, 2007. 30

[81] O. Neiroukh and Xiaoyu Song. Improving the process-variation tolerance of digital circuits using gate sizing and statistical techniques. In *Proceedings of the Design, Automation and Test in Europe Conference*, 2005. 23

[82] S. Ozdemir, D. Sinha, G. Memik, J. Adams, and H. Zhou. Yield-aware cache architectures. In *Proceedings of the International Symposium on Microarchitecture*, 2006. 27, 72, 92

[83] Y. Pan, J. Kong, S. Ozdemir, G. Memik, and S.W. Chung. Selective wordline voltage boosting for caches to manage yield under process variations. In *Proceedings of the Design Automation Conference*, 2009. 27, 58, 93, 107, 109

[84] M.J.M. Pelgrom, Aad C J Duinmaijer, and A.P.G. Welbers. Matching properties of mos transistors. *Solid-State Circuits, IEEE Journal of*, 1989. 14

[85] M. A P Pertijs, K. A A Makinwa, and J.H. Huijsing. A cmos smart temperature sensor with a 3 sigma; inaccuracy of plusmn;0.1 deg;c from -55 deg;c to 125 deg;c. *Solid-State Circuits, IEEE Journal of*, 2005. 57

[86] H. Pilo, C. Barwin, G. Braceras, C. Browning, S. Lamphier, and F. Towler. An sram design in 65-nm technology node featuring read and write-assist circuits to expand operating voltage. *Solid-State Circuits, IEEE Journal of*, 2007. viii, 7, 27, 93, 107

[87] C J. Progler, Amir Borna, David Blaauw, and Pierre Sixt. Impact of lithography variability on statistical timing behavior. In *Proceedings of the SPIE Conference*, 2004. 17

[88] H. Qin, Y. Cao, D. Markovic, A. Vladimirescu, and J. Rabaey. Standby supply voltage minimization for deep sub-micron. *Elsevier Science Microelectronics Journal*, 2005. 52

[89] A. Ricketts, J. Singh, K. Ramakrishnan, N. Vijaykrishnan, and D.K. Pradhan. Investigating the impact of nbti on different power saving cache strategies. In *Proceedings of the Design, Automation Test in Europe Conference and Exhibition*, 2010. 20

[90] S. Rodriguez and B. Jacob. Energy/power breakdown of pipelined nanometer caches (90nm/65nm/45nm/32nm). In *Proceedings of the International Symposium on Low Power Electronics and Design*, 2006. 23

[91] D. Rossi, N. Timoncini, M. Spica, and C. Metra. Error correcting code analysis for cache memory high reliability and performance. In *Design, Automation Test in Europe Conference Exhibition (DATE), 2011*, 2011. 110, 115

[92] H. Sanchez, R. Philip, J. Alvarez, and G. Gerosa. A cmos temperature sensor for powerpc risc microprocessors. In *Proceedings of the Symposium on VLSI Circuits*, 1997. 57

[93] S.R. Sarangi, B. Greskamp, R. Teodorescu, J. Nakano, A. Tiwari, and J. Torrellas. Varius: A model of process variation and resulting timing errors for microarchitects. *Semiconductor Manufacturing, IEEE Transactions on*, 2008. 30, 37, 49

[94] A. Sathanur, A. Pullini, L. Benini, G. De Micheli, and E. Macii. Physically clustered forward body biasing for variability compensation in nanometer cmos design. In *Proceedings of the Design, Automation Test in Europe Conference and Exhibition*, 2009. 115

[95] G. Semeraro, D.H. Albonesi, S.G. Dropsho, G. Magklis, S. Dwarkadas, and M.L. Scott. Dynamic frequency and voltage control for a multiple clock domain microarchitecture. In *Proceedings of the International Symposium on Microarchitecture*, 2002. 25

[96] C. Seung-Hoon, B.C. Paul, and K. Roy. Novel sizing algorithm for yield improvement under process variation in nanometer technology. In *Proceedings of the Design Automation Conference*, 2004. 23

[97] P.S. Shirvani and E.J. McCluskey. Padded cache: A new fault-tolerance technique for cache memories. *Proceedings of the VLSI Test Symposium*, 1999. 27

[98] A.K. Singh, Ku He, C. Caramanis, and M. Orshansky. Mitigation of intra-array sram variability using adaptive voltage architecture. In *Proceedings of the International Conference on Computer-Aided Design*, 2009. 27, 58

[99] D. Sinha, N.V. Shenoy, and H. Zhou. Statistical gate sizing for timing yield optimization. In *Proceedings of the International Conference on Computer-Aided Design*, 2005. 23

[100] K. Skadron, M.R. Stan, K. Sankaranarayanan, W. Huang, S. Velusamy, and D. Tarjan. Temperature-aware microarchitecture: Modeling and implementation. *ACM Transactions on Architecture and Code Optimization*, 2004. 17

[101] A. Srivastava, D Sylvester, and D. Blaauw. Statistical optimization of leakage power considering process variations using dual-vth and sizing. In *Proceedings of the Design Automation Conference*, 2004. 51, 52

[102] B.E. Stine, D.S. Boning, and J.E. Chung. Analysis and decomposition of spatial variation in integrated circuit processes and devices. *Semiconductor Manufacturing, IEEE Transactions on*, 1997. 31

[103] P.A. Stolk, F.P. Widdershoven, and D.B.M. Klaassen. Modeling statistical dopant fluctuations in mos transistors. *Electron Devices, IEEE Transactions on*, 1998. 26

[104] R. Teodorescu, J. Nakano, A. Tiwari, and J. Torrellas. Mitigating parameter variation with dynamic fine-grain body biasing. In *Proceedings of the International Symposium on Microarchitecture*, 2007. 24, 58, 70

[105] A. Tiwari, S.R. Sarangi, and J. Torrellas. Recycle:: pipeline adaptation to tolerate process variation. In *Proceedings of the International Symposium on Computer Architecture*, 2007. 25

[106] J. Tschanz, S. Narendra, R. Nair, and V. De. Effectiveness of adaptive supply voltage and body bias for reducing impact of parameter variations in low power and high performance microprocessors. In *Proceedings of the Symposium on VLSI Circuits*, 2002. 24

[107] O.S. Unsal, J.W. Tschanz, K. Bowman, V. De, X. Vera, A. Gonzalez, and O. Ergin. Impact of parameter variations on circuits and microarchitecture. *Micro, IEEE*, 2006. 5

[108] R. Vattikonda, W. Wang, and Y. Cao. Modeling and minimization of pmos nbti effect for robust nanometer design. In *Proceedings of the Design Automation Conference*, 2006. 20

[109] S. Wang, G. Duan, C. Zheng, and T. Jin. Combating nbti-induced aging in data caches. In *Proceedings of the International Conference on Great Lakes Symposium on VLSI*, 2013. 119

[110] C. Wilkerson, H. Gao, A.R. Alameldeen, Z. Chishti, M. Khellah, and S.L. Lu. Trading off cache capacity for reliability to enable low voltage operation. In *Proceedings of the 35th Annual International Symposium on Computer Architecture*, ISCA '08, 2008. 27, 93

[111] S. J E Wilton and N.P. Jouppi. Cacti: an enhanced cache access and cycle time model. *Solid-State Circuits, IEEE Journal of*, 1996. 36

[112] Lin Xie, A. Davoodi, K.K. Saluja, and A. Sinkar. False path aware timing yield estimation under variability. In *Proceedings of the VLSI Test Symposium*, 2009. 34

[113] H. Yamauchi. Embedded sram trend in nano-scale cmos. In *Memory Technology, Design and Testing, IEEE International Workshop on*, 2007. viii, 6

[114] Y. Ye, T. Liu, Min Chen, S. Nassif, and Yu Cao. Statistical modeling and simulation of threshold variation under random dopant fluctuations and line-edge roughness. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, 19, 2011. viii, 15, 16

[115] W. Zhang, K.C. Chun, and C.H. Kim. Variation aware performance analysis of gain cell embedded drams. In *Proceedings of the International Symposium on Low-Power Electronics and Design*, 2010. 76, 85

[116] S.T. Zhou, S. Katariya, H. Ghasemi, S. Draper, and N.S. Kim. Minimizing total area of low-voltage sram arrays through joint optimization of cell size, redundancy, and ecc. In *Proceedings of the International Conference on Computer Design*, 2010. 27, 97, 104, 105, 106, 111, 112