Enabling Internet-Of-Things Using Low-Power Circuit Design and Automation Techniques

by

Shourya Gupta

A dissertation presented to

The Faculty of the School of Engineering and Applied Science

University of Virginia

In partial fulfillment of the requirements for the degree

Doctor of Philosophy in Electrical Engineering

May 2023

APPROVAL SHEET

This

is submitted in partial fulfillment of the requirements
for the degree of

Author:

Advisor:

Advisor:

Committee Member:

Committee Member:

Committee Member:

Committee Member:

Committee Member:

Committee Member:

Accepted for the School of Engineering and Applied Science:

Jennifer L. West, School of Engineering and Applied Science

Abstract

Wireless sensor nodes for Internet-of-Things (IoT) applications collect information in a variety of consumer, commercial, and industrial applications such as wearable electronics, healthcare monitoring, smart homes etc. These aim to dramatically improve our quality of life and productivity. However, the proliferation of such IoT networks has been hindered by the increase in cost of deployment due to limited operational lifetimes of its nodes. This is because it becomes prohibitively expensive to change the batteries of dead nodes in ever growing networks. In addition, since IoT applications differ greatly in the type of sensing and data collection and consequently their circuit implementation and power budgets, there are challenges related to design complexity, manual engineering design, time-to-market constraints, and the requirement for number of functional blocks. Therefore, there is a strong need to develop solutions that allow fast and cost-effective generation of low power circuits that can run on ambient energy, thereby reducing, or eliminating their need for a battery. This research aims to enable self-powered operation in a larger number of IoT applications by investigating and modelling integrated circuit design techniques for maximum power optimization in memory circuits, creating ultra-low power memory circuits, and enabling rapid generation of analog and mixed signal circuits using digitally synthesizable unit-cell-based approach.

In this work, we aim to model the minimum operating voltage (VMIN) and the various design variables in an SRAM for rapid designing and power optimization. In particular, the statistics of design variables in the critical path of the SRAM read and write operation that greatly affect the minimum operating voltage (VMIN) are studied and modelled. A statistical model is developed that provides quick (~15 sec) estimation of failure probability and the corresponding VMIN for a given SRAM design with low error (<6%). The model is also used to create a dataset (with ~160K unique SRAM design points) in 20 hours, to observe the effect of various SRAM design variables, quantize their importance and determine inter-variable correlation.

Next, we design a new memory bit-cell that uses cross-coupled Dynamic Logic Suppression (DLS) logic inverters to significantly reduce the leakage power of the memory to enable lower power budget for an SoC or IoT node. The new bit-cell allows to retain pW to nW power budget at higher supply voltages without depending on aggressive supply voltage scaling to retain higher performance. The design and performance of the DLS inverter pair is analyzed from this context to ensure reliable operation. Two SRAM implementations (2KB and 6KB) are

developed for use with the DLS bit-cell, and are fabricated in a 65nm test chips. The 2KB DLS SRAM consumes 52pW (at 0.3V) to 132pW (at 0.9V) and the 6KB version consumes 618pW (at 0.3V) and 10.1nW (at 0.9V), thereby enabling nW operation for IoT nodes. Another SRAM is implemented using Scalable Dynamic Logic Suppression (SDLS) logic inverters to enable high performance during access mode, while simultaneously retaining low power in stand-by mode. Simulation results show performance ranging from 3.5KHz (at 0.3V) to up to 10MHz (0.7V) with a leakage power of 1.8nW to 23nW respectively.

Finally, to address the ever-growing need for automation in analog circuit design and integration to meet modern-day IoT SoC requirements, we propose methodologies to synthesize correct-by-construction RTL descriptions for both analog and mixed signal circuits using a unit-cell-based approach. We apply these methods to SRAM and Low Dropout Voltage Regulator (LDO) as proof of concept. Several prototypes are implemented in 65nm bulk planar CMOS and 12nm FinFET technologies. A large reduction is observed in the manual effort from several weeks/months to just a few hours with minimal loss of performance compared to manual design efforts.

Acknowledgements

I would like to express my sincere gratitude to my PhD advisor, Ben Calhoun, for his unwavering support, guidance, and encouragement throughout graduate school. Without your mentorship and expertise, this research would not have been possible. You have guided me through tough situations, both personal and professional. You always pushed me to strive for excellence and always challenge myself. Your guidance has allowed me to develop skills that I will benefit from for the rest of my life. I would also like to thank the members of my dissertation committee: Prof. Steven M. Bowers, Prof. N. Scott Barker, Prof. Nikhil Shukla, and Prof. Brad Campbell, for their valuable feedback and constructive criticism that helped shape this work.

I want to thank Prof. Steven Bowers for his guidance and teachings. Your analog classes were always very helpful and you made it fun and interesting to learn. I would also like to thank Prof. N. Scott Barker. Your RF class was my first class at UVA and I still think of it as one of my most intellectually enlightening and favourite classes. You always taught difficult concepts with ease and made them very interesting. I am also very grateful to Terry Tigner for her personal support and administrative help she provided me during my years at grad school. I could always come to you for any help and you made everything better.

I would like to thank Prof. Kirti Gupta and Prof. Neeta Pandey. You taught me things in such a lucid and fun way. It helped spark an interest for research in me. It was with your guidance, that I thought about going into research and pursue grad school. Without your guidance, teachings, and encouragement, I couldn't have accomplished what I have.

There are many people at UVA who contributed greatly to my life at UVA. I would like to thank Sumanth Kamineni with whom I worked a lot on the IDEA project throughout my PhD. Our frequent discussions about work and personal life helped me grow and improve. I am grateful to Chien-hen Chen for his friendship, guidance, and support throughout all the time we spent together. I would like to thank Xinjian Liu, Natalie Ownby, and Katy Flynn. ISSCC 2020 was so much fun with you guys. Xinjian, you have been a great friend. Technical discussions with you have always been fun and interesting. I am so grateful to have a friend like you. I would also like to thank my seniors Daniel Truesdell, Shuo Li, Jacob Brieholz, and Anjana Dissanayake. I could always come to you guys for your expertise and help whenever I was completely stuck. I would also like to thank Rishika

Agarwala. Our trips to the library and book store were always so much fun. You have helped and guided me so much and I am really grateful to have a friend like you. I would also like to thank Dave Wentzloff, Yaswanth Cherivirala, and Kyumin Kwon. Our biannual meetups at IDEA meetings were always so fun and interesting.

There are many past and current students who have interacted with me and helped me in many ways: Ningxi Liu, Henry Bishop, Nick Napoli, Peng Wang, Peter Le, Suprio Bhattacharya, Omar Faruqe, Daehyun Lee, Anjali Agrawal, Akiyoshi Tanaka, Nugaira Gahan Mim. I would also like to thank Yaobin Zhang. You have been a dear friend and our outings together were always so much fun. I would like to thank Yu Pan. Our interactions made my days in Link Lab happier and fun.

Outside of UVA, there are many people who I would like to thank. First and foremost, my Dada-ji. You always encouraged me to study hard and strive for excellence. You have always been so proud of me and I dedicate this achievement to you. To my Dadi, I wish you were here to see me achieve this milestone. I am sure you would have been super proud and happy for me. To my mummy and papa, I am eternally grateful to you. You have provided me immense support and encouragement. Without you, I would not have been able to become what I am today. You have inspired me to work hard, always better myself, and be kind and helpful to others. I want to thank my sister Muskan. You supported me in ways nobody else did. Our endless conversations about even the most mundane things were so funny and uplifting. I am so far away from home. I wish I could see you more often and be there for you. I love you so much. I would also like to thank Sagar and Nisha for their support during my PhD. You always helped and supported me and for that I am grateful. I would also like to thank Rohit, Rachit, Bua, and Foofa Ji for their support and love that they provided me throughout my PhD.

To Bade-papa, Badi-mummy, Papa, Mummy, and Muskan

List Of Figures

Fig. 2.14. Interaction of bit-line discharge and sense-amplifier-enable results in a near-gaussian distribution across majority of the small signal sensing operation. The same analysis also shows that change in sense amplifier strobe signal does not have a very strong effect on the resulting distribution. This trend will be similar irrespective of region of operation since resultant distribution will always spread in both *X* and *Y* directions.

Fig. 2.15. Comparison of read-access distribution using MC simulations and proposed method in super-threshold and subthreshold regions.

Fig. 2.16. Comparison of skew and kurtosis evaluated using model and MC simulations. Evaluation depicts a deviation of less than 0.6 skew and kurtosis in the worst case.

Fig. 2.17. (a) Correlation between various SRAM design variables and failure probability (b) Dynamic behavior of correlation between frequency and failure probability as a function of separation.

Fig. 2.18. Importance (absolute correlation) of various SRAM design variables in descending order.

Fig. 2.19. Comparison of various yield prediction methods based on speed and error.

Fig. 2.20. The Log-Transformation approach can be used to transform the skewed distribution of the write-access operation into a normal distribution to easily calculate the failure probability. The histograms of the linear inverse transformation distribution are shown in (a) super-threshold & (b) sub-threshold region of operation, with the corresponding probability plots in (c) & (d). The histogram of the proposed transformation model is shown in (e) & (f), with the corresponding probability plots in (g) & (h). (i) Value of zeta variable computed using 100K MC sims. (j) Trend of Write-Access failure probability and corresponding VMIN with varying supply voltage.

Fig. 2.21. Determining write-access failure probability by modelling it using Noncentral F distribution. Normalized Write Access-Time variation with respect to change in V_T for each transistor in super-threshold region (a) when X=0 (b) X=1 and in subthreshold region (c) when X=1 (d)X=0. Comparison of write-access operation distributions using proposed modelling technique and MC sims in (e) Super-threshold region (f) Sub-threshold region of operation, with the corresponding probability plots in (g) & (h).

Fig. 2.22. (a) Comparison of write-access operation distributions using sensitivity analysis and MC sims in 12nm FinFET process (1M iterations) (b) The relative impact and contribution of each transistor in the 6T bit-cell on the write access operation performance in super-threshold and subthreshold region of operation. The contribution of the pull-up transistor on the hold ‘one’ (and write ‘zero’) side of the bit-cell increases in subthreshold region. (c) Correlation between threshold voltage of transistor and write-access time for each transistor in the 6T bit-cell.

Fig. 2.23. Comparison of (a) Static write margin distribution using model and MC sims. at V_{DD} = 0.4V (100K iterations) (b) write-access distribution with respect to frequency at V_{DD} = 0.4V. Results show about 1.2% error in computation of contention-limited write-access failures.

Contents

1. Introduction

1.1. Motivation

Fig. 1.1. Internet-of-Things (IoT) spans across a large variety of applications.

The Internet-of-Things (IoT) has grown significantly over the past few years leading to new applications in wearable electronics, healthcare monitoring, smart homes etc. as shown in Fig 1.1. The IoT is growing at a rapid pace and is projected to reach billions of nodes as shown in Fig. 1.2. These nodes will be deployed in a variety of consumer, commercial, and industrial spaces to facilitate the collection and exchange of information for generating valuable insights and feedback. However, several challenges stand in the way of the sustained growth of the IoT. The nodes within a network are usually powered by a battery. As such, as the network grows in size, there is an ever-increasing cost of replacing the batteries. This severely limits the scope and size of the IoT. Solutions to this problem have ranged from reducing the power consumption of the circuits to enabling the nodes to harvest energy from its environment like solar or thermal. Recent developments in harvesting technologies offer new sources of harvesting and can potentially unlock more IoT applications. For example, harvesting energy from sources like WiFi can provide nW-levels of power [5]. Ambient humidity can provide nWs of power using nanometer-scale protein wires [6]. Other sources like the thermal motion of graphene have been shown to provide pW to nW levels of power [7]. However, to make use of these sources, and open the pathway to a new gamut of IoT applications, there must be innovation to reduce the power consumption down to harvestable levels to enable battery-free operation as shown in Fig. 1.3.

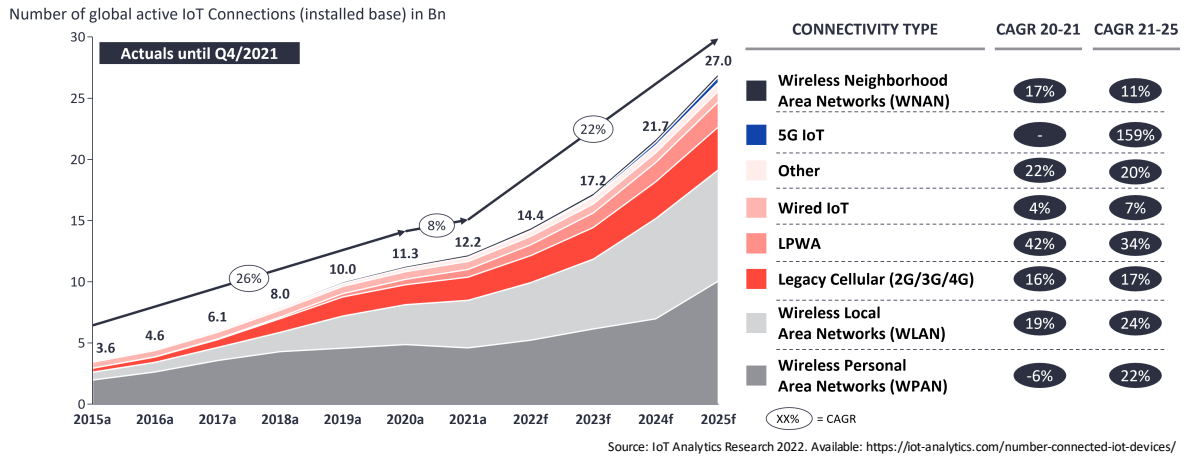


Fig. 1.2. Internet-of-Things (IoT) is expected to reach >27 billion devices by 2025.

In addition to the power budgeting challenges of the IoT, there are challenges related to design complexity, manual engineering design, time-to-market constraints, and the number of functional blocks. Due to the varying application space, the hardware implementation can also vary greatly, both in terms of complexity and power level as shown in Fig. 1.3. Moreover, wireless connectivity, standards compliance, sensing modalities, and energy harvesting modalities can also add to the complexity of the construction of the hardware. To address these issues, there needs to be innovation in how we approach the generation of such nodes to enable low cost and rapid growth of the IoT.

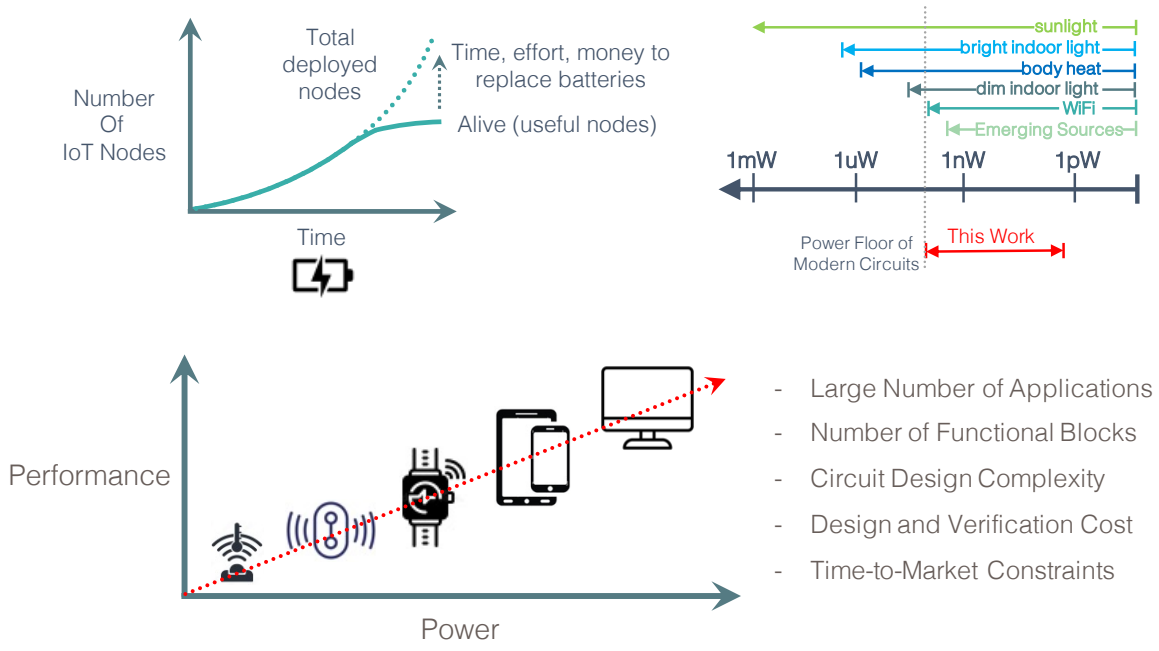


Fig. 1.3. Batteries can limit the sustained growth of IoT networks due to increased cost of replacement. Lowering power can increase the number of applications where self-powered operation can achieved. Design and verification costs vary according to application requirements can affect cost and deployment.

1.2. Thesis Statement and Contributions

The Internet-of-Things (IoT) nodes that can harvest energy from their environments can theoretically work forever, thereby greatly extending their practical operational lifetimes. This allows for IoT nodes to support increased growth and sustainability of the IoT. The amount of power that can be harvested from the ambient environment is often intermittent and scarce depending on the application. Thus, decreasing the power consumption of the IoT node will increase the number of applications where it can achieve self-powered operation. This dissertation proposal demonstrates circuit modelling and design techniques applicable to Static Random Access Memory (SRAM) that enable state-of-the-art for ultra-low power memory design to lower the power barrier to IoT. In addition, since IoT applications differ greatly in the type of sensing and data collection and consequently their circuit implementation and power budgets, there are challenges related to design complexity, manual engineering design, time-to-market constraints, and number of functional blocks. Thus, we demonstrate design automation techniques that help to enable rapid generation of analog and mixed signal circuits suitable for the various type of applications.

Previous works have relied on aggressively scaling the supply voltage to reduce the leakage power, but at the cost of reduced performance. The leakage floor of SRAMs can be lowered to pW level without relying mainly on voltage scaling, thereby simultaneously retaining high performance and low power at higher operating voltages, and allowing to increase the number of applications where an IoT node can achieve self-powered operation.

In addition, using circuit design automation techniques to create mixed-signal and analog circuits can bring down the design time for such circuits from years/months down to days/hours, resulting in decreased costs, improved time-to-market, and higher quality and reliability.

1.2.1. SRAM Dynamic VMIN Modelling

Random variations in nano-scale Static Random-Access Memories (SRAM) pose a major challenge to achieving design robustness due to their large effect on bit-cell and array characteristics. The worst-case VT mismatch, combined with the increased sensitivity of current in the subthreshold region, greatly affects the minimum operating voltage (VMIN) and yield of the memory. Since the yield and the directly related VMIN parameter determine the extent of voltage supply scaling, their accurate estimation is important for maximizing energy and performance savings. To achieve >95% yield for a 10 Mbit memory, a failure probability of less than 1e-9 should be reached. To ascertain this probability, more than 1e11 samples would be required, which is not practically possible. Additionally, the SRAM design space is a multidimensional one, with variables having interdependent trade-offs and varying levels of correlation with design feasibility. In this work, we propose an analytical model that evaluates the read and write access failure probability and the corresponding VMIN in a short amount of time (few seconds) and low error (<10%), thereby providing a huge speedup than previous techniques (up to 100,000X). We also provide accurate statistical estimation and impact of various design variables in the SRAM's multidimensional design space.

1.2.2. Ultra-Low Power Memory Design

The power consumption of SRAM is similar to that of a digital circuit. The conventional memory circuits contain two cross-coupled inverters in every bit-cell which can add up to hundreds of thousands of inverters, with only a small fraction of them being switched during an access operation. Memories for low power IoT

applications usually have a much lower activity factor than their counterparts in high processing work stations or servers. Thus, the power consumption such memory circuits with low activity factor is easily dominated by subthreshold leakage. To reduce the power of the memory to enable lower power budget for an SoC or node, a new memory bit-cell is designed that uses cross-coupled Dynamic Logic Suppression (DLS) logic inverters to significantly the leakage. The design and performance of the DLS inverter is reanalyzed from this context to ensure reliable operation. A full SRAM implementation is developed for use with the DLS bit-cell, and is fabricated in a 65nm test chip. In addition, a Scalable Dynamic Logic Suppression (SDLS) logic-based SRAM is prototyped in 65nm to enable higher performance during a write access operation.

1.2.3. Analog and Mixed Signal Circuit Design Automation

As IoT is maturing, its implementation will require billions of hardware sensor nodes, and will be deployed in a variety of consumer, commercial, and industrial spaces to facilitate the collection and exchange of information for generating valuable insights and feedback. All these applications differ in the type of sensing and data collection and consequently their circuit implementation and power budgets. Additionally, time-to-market constraints have become tighter, design complexity has increased and more functional blocks (in number and variety) are being integrated into SoCs. These challenges often translate to increased manual engineering efforts and non-recurring engineering (NRE) costs. Thus, there is an ever-growing need for automation in analog circuit design, validation, and integration to meet modern-day IoT SoC requirements. In this work, we propose methodologies to automatically synthesize correct-by-construction RTL descriptions for both analog and mixed signal circuits. We apply these to SRAM and Low Dropout Voltage Regulator (LDO) as proof of concept.

2. SRAM Dynamic VMIN Modelling

2.1. SRAM Background

SRAM is a type of semiconductor memory that is commonly used in microprocessors, digital signal processors (DSPs), and other high-speed digital applications. It is a volatile memory, meaning that it requires a continuous supply of power to maintain its stored data. Unlike DRAM (Dynamic Random Access Memory), SRAM does not need to be periodically refreshed, which makes it faster and more power-efficient. It is widely used as a cache memory in microprocessors and other high-speed digital applications due to its high-speed performance and low power consumption. Additionally, SRAM is used as a buffer memory in communication systems, as well as in high-performance graphics cards and gaming consoles.

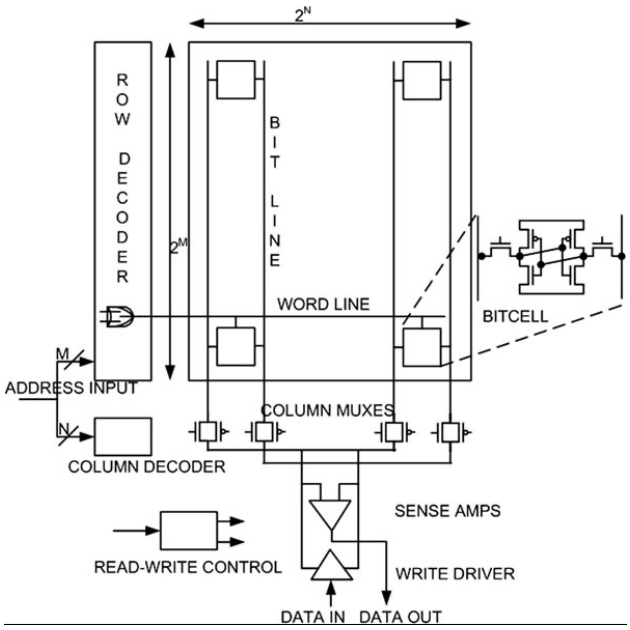


Fig. 2.1. Architecture of an SRAM array [8].

An SRAM cache is made up of an array of memory bit-cells, which are bistable and can store one bit of information as shown in Fig. 2.1. The bit-cells are connected to peripheral circuitries, such as address decoders, sense amplifiers, write drivers, and bit-line pre-charge circuits, which enable reading from and writing into the array. The memory array typically consists of 2^n words of 2^m bits each, and an SRAM array is made up of millions of identical bit-cells.

Optimizing the SRAM bit-cell designs for a target application is an active area of research because small improvements in reliability, performance, and static power consumption can have a significant impact on the entire processor or SoC product. In high-performance processors, operating speed and bit-cell area are the primary concerns for high-density caches while maintaining adequate reliability. However, in energy-constrained applications like sensor nodes or medical implants, energy efficiency and reliability are the main issues.

Each memory bit-cell can store a single bit of information (shown in Fig. 2.2), and they share a common word-line and bit-line pairs in each column of the SRAM array. The dimensions of each SRAM array are limited by its electrical characteristics, such as the capacitances and resistances of the bit-lines and word-lines used to access the bit-cells. To meet the bit and word line capacitance requirement, every row of the memory contains 2^k words after folding, so the array is physically organized as 2^{n-k} rows and 2^{m+k} columns.

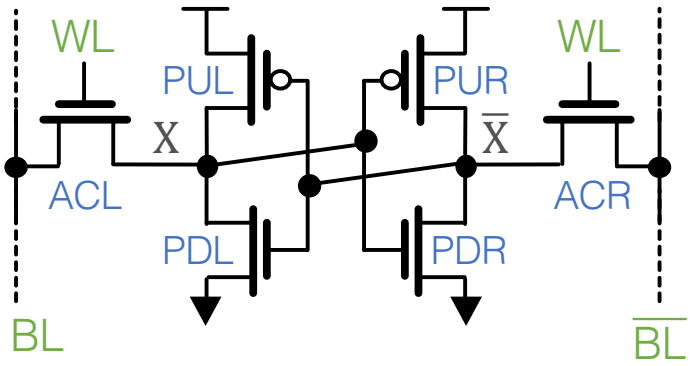


Fig. 2.2. Circuit schematic for the conventional 6T cell.

To randomly address each bit-cell, the appropriate word-line and bit-line pairs are activated by the row and column decoders, respectively. For large-sized memories, the memory can be folded into multiple blocks with a limited number of rows and columns to meet the electrical characteristics requirements. The SRAM bit-cell is the fundamental component of the SRAM array, capable of storing a single bit of information. It offers non-destructive read operation, write capability, and data storage as long as it has power. The standard six-transistor (6T) SRAM bit-cell comprises two cross-coupled inverters and two access transistors that are linked to each data storage node. The inverter pair forms a latch that holds binary information, and the true and complementary versions of the binary data are stored in the storage nodes. The access transistors enable access to data storage

nodes during read and write operations and provide isolation from neighboring circuits during hold state. During read and write operations, the bit-cells are accessed horizontally by asserting the word-line. When the word-line of a row is activated, all the memory bit-cells in the selected row become "active" and ready for read and write operations. To decode m word-lines, log2m address bits are required. The SRAM bit-cell has three modes of operation: read, write, and standby, which correspond to the states of reading, writing, and data retention, respectively.

During a read operation, the bit line (BL) and its complement ($B\bar{L}$) are precharged to a certain voltage level (usually halfway between Vdd and Gnd). The word line (WL) is then activated to enable the access transistor of the selected bit cell. If the bit stored in the cell is a logic 0, the access transistor will be in the off state, and the BL voltage will remain at the precharge level. On the other hand, if the bit stored in the cell is a logic 1, the access transistor will be turned on, and the BL voltage will be pulled down to a lower level (typically close to Gnd) due to the cross-coupled inverters configuration.

After the bit line voltage has stabilized, it is sensed by the sense amplifier, which amplifies the voltage difference between BL and $B\bar{L}$ and generates an output signal that represents the value stored in the selected bit cell. The sense amplifier is designed to detect and amplify this small voltage difference, and convert it into a large and stable output voltage that represents the stored data in the memory cell. The sense amplifier operates in two stages: precharge and amplification. During the precharge stage, the sense amplifier circuit is reset to its initial state by discharging the differential inputs and holding the output at a reference voltage level. This prepares the sense amplifier to receive the voltage difference on the bit lines. During the amplification stage, the differential inputs of the sense amplifier are connected to the bit lines, and the voltage difference is allowed to propagate to the sense amplifier inputs. The voltage difference is amplified by the differential amplifier circuit, and the output voltage is latched into the output buffer. The output voltage represents the stored data in the selected memory cell and can be used by the external circuitry.

The sense amplifier also includes a feedback mechanism that helps stabilize the output voltage and reduce the effects of noise and other factors in the system. This feedback mechanism is typically implemented using a

capacitor and a transistor that form a feedback loop between the output and the differential inputs of the sense amplifier.

During a write operation, the BL and $B\overline{L}$ lines are precharged to the same voltage level as in the read operation. The WL is then activated to turn on the access transistor of the selected bit cell. To write a logic 0 into the cell, the BL voltage is forced to a lower level (typically close to Gnd), which turns on the pull-up transistor and turns off the pull-down transistor in the cross-coupled inverters configuration. This sets the output to a logic 0 state, which is then latched into the cell by the inverters. To write a logic 1 into the cell, the $B\overline{L}$ voltage is forced to a lower level, which turns on the pull-down transistor and turns off the pull-up transistor in the cross-coupled inverters configuration. This sets the output to a logic 1 state, which is then latched into the cell by the inverters. Once the write operation is complete, the WL is deactivated and the BL and $B\overline{L}$ lines are precharged again to their original voltage levels. This completes the write cycle.

2.1.1. SRAM Scaling

Embedded SRAM has become an integral part of modern microprocessors and SoCs, with caches occupying a significant portion of the chip area. This trend is expected to continue due to the need for higher performance, lower power, and higher integration. To achieve higher memory density, memory bit-cells are scaled down with each technology node as shown in Fig. 2.3, resulting in smaller bit-cell sizes and increased vulnerability to process variations. In advanced CMOS technology nodes, process variations significantly impact SRAM functionality as supply voltage is reduced, leading to a decrease in robustness. Local random variations, such as line edge roughness, have a strong effect on SRAM operation. Vth variation for SRAM devices increases significantly with scaling, which presents a major challenge for SRAM design. The impact of process variations on SRAM is much stronger than on random logic and is the predominant factor in yield loss in advanced technology nodes.

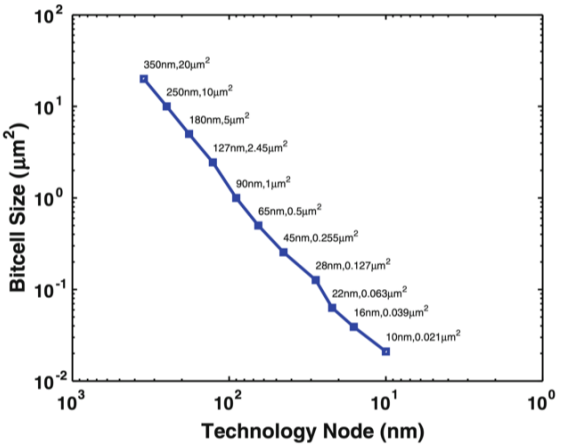


Fig. 2.3. SRAM bit-cell scaling trend across technology nodes [3].

As technology advances and SRAM bit-cells are scaled down, the traditional hard fails due to defect density decrease. However, the reduction in bit-cell size also leads to an increase in process variations, which becomes the dominant cause of bit-cell failure. This increase in SRAM failures can have a significant impact on overall product yield due to the high memory densities on chip. Additionally, the sensitivity of stability failures to supply voltage can limit lower VDD operation. The four main parametric failure mechanisms in SRAM are

1. read access failure,
2. read stability or read disturb failure,

3. write failure,
4. hold or retention fail.

These failures are considered parametric because they affect memory operation under specific conditions and can be recovered at higher supply voltages. Therefore, these failure mechanisms become a limiting factor for SRAM supply voltage scaling.

2.2. Introduction to Modelling

Random variations in nano-scale Static Random Access Memories (SRAM) pose a major challenge to achieving design robustness due to their large effect on bit-cell and array characteristics [10]-[12]. These variations include device threshold voltage (V_T) mismatch due to random dopant fluctuations (RDF) and line edge roughness (LER) [13]. The device V_T mismatch in deep sub-micrometer technologies is greatest in minimum sized devices, which are often used in SRAMs [14]. The worst-case V_T mismatch, combined with the increased sensitivity of current in the subthreshold region, greatly affects the minimum operating voltage (VMIN) and yield of the memory. Since the yield and the directly related VMIN parameter determine the extent of voltage supply scaling, their accurate estimation is important for maximizing energy and performance savings. Monte-Carlo (MC) simulation is a well-known approach to determine the worst-case VMIN for a given memory. However, memory arrays can require millions of MC simulations, which is prohibitively expensive. Additionally, the SRAM design space is a multidimensional one, with variables having interdependent trade-offs and varying levels of correlation with design feasibility. This includes the SRAM read critical path which includes the largest number of design variables. Therefore, it becomes challenging and time consuming to arrive at an optimized design solution. Many analytical and semi-analytical approaches have previously been proposed to determine the VMIN and yield of the memory, which we describe in the next section. While some of these approaches greatly reduce the simulation time over conventional MC simulations to help determine design feasibility more quickly, they do not help to resolve the design space or quantify the statistical importance of underlying design variables.

In this work, we propose an analytical model that evaluates the read-access failure probability and the corresponding VMIN. The model takes key design variables into account, including supply voltage, temperature, process variations, and array design parameters including bit-cell sizing, read current, bit-line capacitance (number of rows), word-line rise time (number of columns), sense amplifier strobe timing, bit-line leakage, and sense amplifier offset voltage. The method can complete a design evaluation within a few seconds with small error (<6%).

Here, we present the following:

1. A new analytical time-based relationship describing the average bit-line discharge rate and its corresponding distribution.

2. For the first time, analytically describing the read access operation using Modified Bessel function of the second kind, which is then approximated with an asymmetrical gaussian distribution with finitely limited skew and kurtosis.
3. These mathematical developments help the model to compute the VMIN and failure probability very fast (~15 sec) and with low error (<6%)
4. The model is then used to create a dataset (with ~160K unique SRAM design points) in 20 hours, which otherwise would have taken >100 years to generate with MC simulations. The dataset is then used to observe the effect of SRAM design variables, quantize their importance and determine inter-variable correlation.
5. An SRAM designer who is accustomed to using MC simulation (or a faster equivalent tool) would be able to supplement their design approach by using this model to analyze the timing distribution of various components in the SRAM read critical path, target the most impactful design variables to save design time and effort, and co-optimize iteratively for speed, area, and power using a yield-aware approach.

For especially long tailed distributions such as the dynamic write-access operation, accurate determination of the tail is crucial to determining the failure point. A few analytical approaches have previously been proposed to determine the dynamic write-access distribution. However, they have limited success in doing so, especially in sub-threshold region of operation, due to the approximations considered in fitting the distribution which we discuss in the next Section. In this work, we provide an analytical transformation model to determine the failure probability of the write-access operation, that works well in both super-threshold and sub-threshold regions of operation. Additionally, since the write-access operation distribution resembles the noncentral F distribution, we also provide an analytical solution to determine its tail. In more recent advanced FinFET technologies, the gate work function variations impact the V_T variations significantly, which is why we present a modified sensitivity analysis-based method to determine the distribution of the write-access operation in advanced technologies. We also present a methodology to determine the contention-limited write-access failures which occur irrespective of pulse width. These proposed alternative analytical methods offer a fast approach with reasonably low error to determine the write-access operation failure threshold and yield in a given SRAM design process in both sub-threshold and super-threshold regions of operation. Since in advanced FinFET technologies, assist circuits play a crucial role in ensuring the continued scaling of SRAMs, we also compare various write assist techniques based on their effect on performance across various regions of operation.

2.2.1. SRAM Yield Determination Approaches

The most straightforward approach to determine the yield of SRAMs is Monte Carlo simulations. In the conventional Monte-Carlo approach, we try to find yield for the metric of interest $f(x)$ with the random variable being x. If y_{limit} is threshold for the performance metric, then the pass or fail function $I(x)$ is defined as

$$I(x) = I(y_i > y_{limit}) = \begin{cases} 1 & if\ \ y_i > y_{limit}, \\ 0 & if\ \ y_i \leq y_{limit} \end{cases} \tag{2.1}$$

And the probability of failure P_f can be defined as

$$P_f = P(y_i > y_{limit}) = \int_{-\infty}^{\infty} I(x)\, R(x)(dx) \tag{2.2}$$

where, $R(x)$ is the probability density function of the random variable x (e.g. V_T).

Since the distribution of $I(x)$ is generally unknown, a large number of samples are needed to be generated corresponding to the random variable. To obtain an estimate with $(1-\varepsilon)\,100\%$ accuracy and with $(1-\delta)\,100\%$ confidence, the required number of samples $N(\varepsilon, \delta)$ is given by [15]

$$N(\varepsilon, \delta) \approx \frac{log\left(\frac{1}{\delta}\right)}{\varepsilon^2 P_f} \tag{2.3}$$

To achieve >95% yield for a 10 Mbit memory, a failure probability of less than 1e-9 should be reached. To ascertain this probability with 95% confidence interval and 10% error, more than 1e11 samples would be required, which is not practically possible. In SRAM circuit design, where the performance metric depends on multiple variables, determining $I(x)$ becomes even more challenging due to the large multi-dimensional design space. Therefore, there is a need to develop alternate methods to verify SRAM design yield.

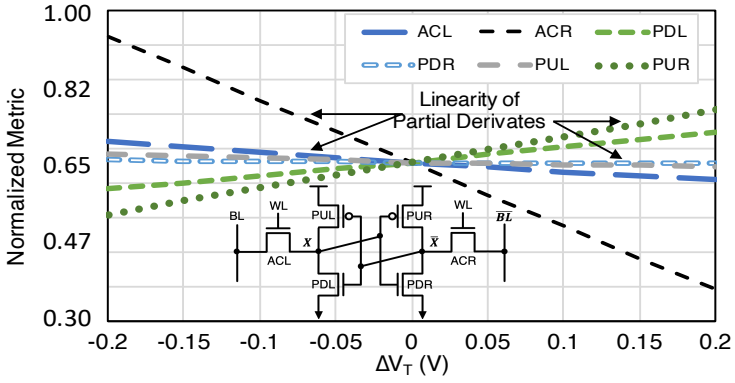


Fig. 2.4. Partial derivatives with respect to V_T using Sensitivity Analysis for a 6T SRAM cell show linearity but this may not always be true.

Some of the alternate methods aim to reduce the simulation time for determining yield by analyzing the impact of process variations on the SRAM. The work in [16]-[20] use Sensitivity Analysis to estimate the SRAM failure probability and yield. This method simplifies the simulation significantly because only $(N+1)$ number of partial derivatives with respect to V_T are needed to be evaluated to estimate the sensitivities (N is the number of independent variables). The partial derivatives for the six transistor (6T) SRAM cell are shown in Fig. 2.4 for a bulk 32nm process. These aid to calculate the mean ($\mu_{f(x)}$) and standard deviation ($\sigma_{f(x)}$) of the distribution as

$$\mu_{f(x)} \approx f(\mu_x) + \sum_{i=1}^{n} \left(\frac{1}{2} \frac{\partial^2 f(\mu_x)}{\partial x_i^2} \right) \sigma_{x_i}^2 \tag{2.4}$$

$$\sigma_{f(x)}^2 \approx \sum_{i=1}^{n} \left[\left(\frac{\partial f(\mu_x)}{\partial x_i} \right) \sigma_{x_i} \right]^2 \tag{2.5}$$

However, this method can be quite inaccurate, because the Taylor expansion can yield errors in approximations away from the nominal point. The sensitivity of the metric with respect to V_T can be highly non-linear in some processes and for some design points, leading to large inaccuracies. Additionally, applying this method to large circuits can be unwieldy, due to the large number of variables involved.

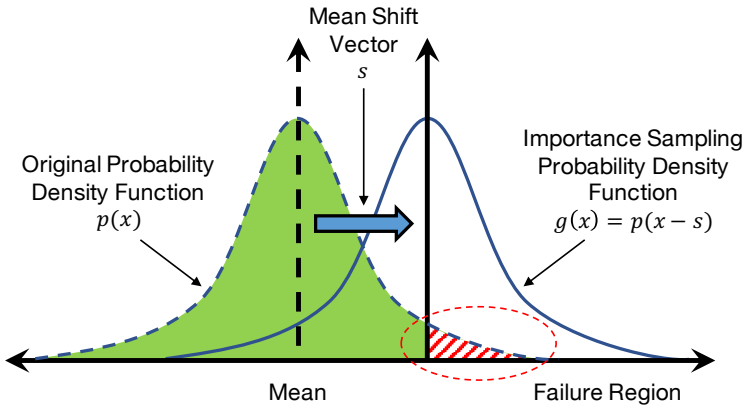


Fig. 2.5. Importance Sampling using the Mean-Shift approach [21].

Another method that aids to reduce the simulation time is Importance Sampling (IS). In one variation of this method, known as Mean-Shift IS, samples are generated away from the mean where failures are much more likely to occur as opposed to the mean of the distribution, where usually no failures occur [21] (shown in Fig. 2.5). The Mean-Shift IS, in which the center of the original distribution $p(x)$ with zero mean and standard deviation σ_j is shifted by a shift-vector $s = (s_1, \ldots, s_M)$, is represented as

$$g(x) = \prod_{j=1}^{M} \frac{1}{\sqrt{2\pi}\sigma_j} exp\left(-\frac{(x_j - s_j)^2}{2\sigma_j^2}\right) \tag{2.6}$$

Then the probability estimated using IS becomes

$$P_{IS} = \frac{1}{N}\sum_{i=1}^{n} I(x_i).w(x_i) \quad (x_i \in G(x)) \tag{2.7}$$

where, the weight function $w(x)$ is represented as

$$w(x_i) = \frac{p(x)}{g(x)} = exp\left(-\sum_{j=1}^{M} \frac{s_j(2x_j - s_j)}{2\sigma_j^2}\right) \tag{2.8}$$

The main disadvantage of this approach is the ambiguity in determining the shift-vector. This is because it is difficult to estimate where the failure region might lie. Additionally, the search region might be too wide and therefore, difficult to explore with a few number of samples. In another IS method [22], a mixture of distributions $g_\lambda(x)$ is used to model the shifted density function.

$$g_\lambda(x) = \lambda_1 R(x) + \lambda_1 U(x) + (1 - \lambda_1 - \lambda_2) R(x - s_j) \tag{2.9}$$

where, $0 < \lambda_1 + \lambda_2 < 1$. This method enables efficient sampling without leaving any non-sampled regions in the event of outliers. Another IS approach improves over mean-shift IS by using norm minimization to reduce the variance [15]. Still, the overall efficiency of all Importance Sampling methods depends on the shift-vector because sampling of the modified distribution function must occur where maximum number of failure points are likely to occur. This makes it hard to implement IS based methods to assess the yield of SRAMs.

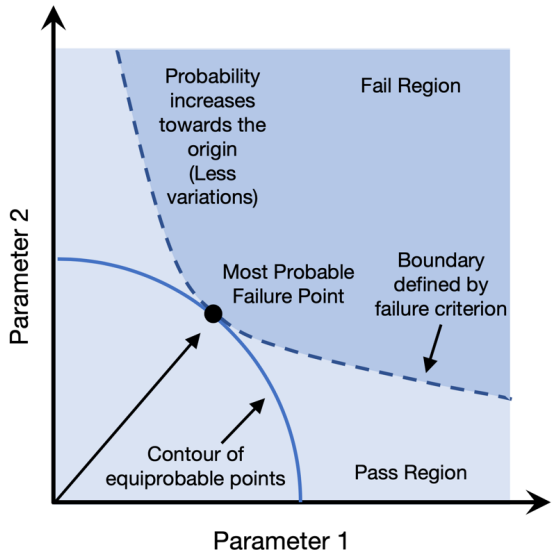


Fig. 2.6. Optimization problem in 2D-space illustrating the Most Probable Failure Point [23].

Most Probable Failure Point (MPFP) is another method that is used to evaluate the yield of SRAMs [23]. In this method, the failure probability determination is treated as a process of optimization as shown in Fig. 2.6. It aims to find the worst-case variations which maximize the failure probability P_{fail}.

$$P_{fail} = \prod_{i=1}^{6} P\left(\Delta V_{t_i} > k_i \sigma_{V_{t_i}}\right) \qquad (2.10)$$

where, k_i represents the threshold voltage variation for the SRAM bit-cell's transistor's V_T with respect to the standard deviation at the most probable failure point. In this approach, the search region is divided into a six-dimensional space (assuming 6T SRAM bit-cell), with sixty-four regions. The search is then performed in only those regions where failures are more likely to occur. Although this method is applicable to large number of cases, even where the failure region might not be known, this brute-force approach can quickly become unwieldy as the number of variables increase.

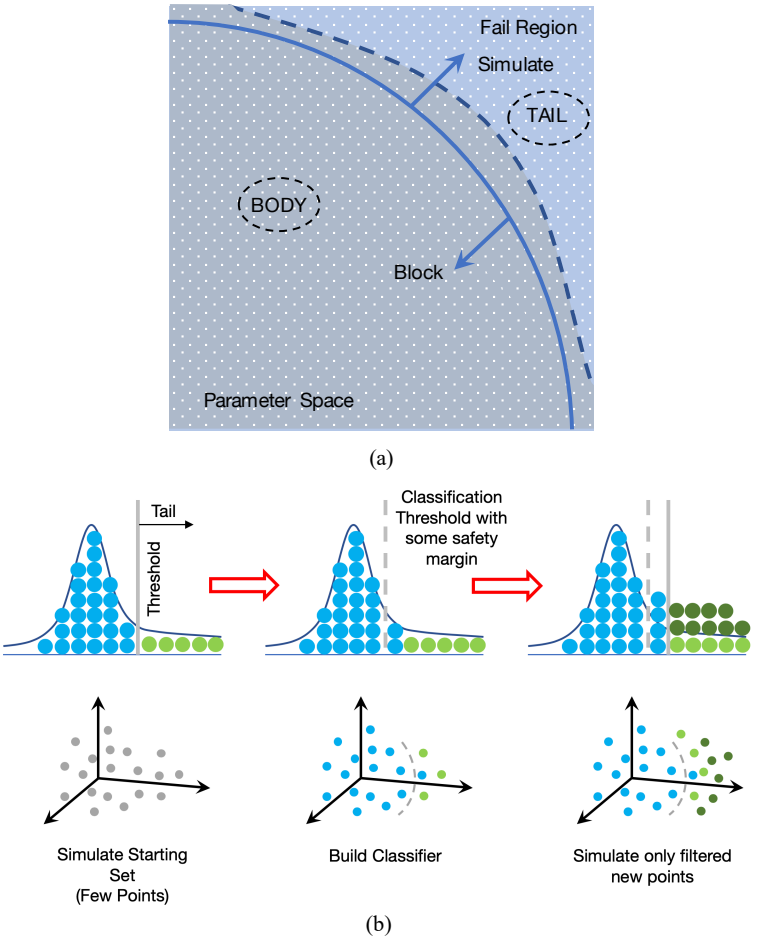


Fig. 2.7. (a) Illustration of the Statistical Blockade (SB), showing the body and tail in the parameter space. The region inside the body, marked by the classifier solid line is blocked [24] (b) Steps to perform SB analysis [24].

Another method that is used to quickly estimate the yield of SRAMs is Statistical Blockade. In this approach, an initial sampling using MC or other sampling methods is performed to build a classifier for the metric of interest as shown in Fig. 2.7 [24]. Only points that are beyond the classifier threshold are simulated, and all other points are blocked. This allows a huge speed-up of simulation by only simulating points which are more likely to fail. In an improved version called the recursive statistical blockade, the search starts with a lower threshold classifier, which is then used to estimate a higher threshold classifier multiple times until the target threshold classifier is reached [25]. This method reduces the simulation time for larger memories where the regular statistical blockade can become unwieldy. Although, the statistical blockade method enables a huge speedup over conventional MC simulations, it can still require up to sixty hours to determine the yield for a given design [19].

Some methods reduce simulation time by modelling the behavior of the SRAM [26], [27]. However, these

methods can still require up to several hundred thousand MC simulations for pre-characterization and evaluation. Another method [13] also models the SRAM behavior, but it has been shown in [16] that the analytical method presented underestimates the failure probability.

2.2.2. SRAM Read Access

Read-access time is defined as the time required to generate a potential difference between the two bit-lines (e.g. 100mV). If more time is elapsed to generate this voltage difference than the given word-line pulse width, then the SA might not be able to evaluate the correct data, thereby resulting in a read-access failure. The conventional method for determining the read-access failure probability P_A can be expressed as [18], [20], [28]

$$P_A = Prob\left(T_A > T_{WL}^A\right) \tag{2.11}$$

where, T_A is the read-access time and T_{WL}^A is the word-line pulse width. T_A can be evaluated using

$$T_A = \int_{V_{DD}}^{V_{DD}-\Delta V_{BL}} \frac{C_{BL} dV_{BL}}{I_{BL}} \tag{2.12}$$

where, V_{BL} is the bit-line voltage, C_{BL} is the bit-line capacitance, and I_{BL} is the read access current. It has previously been established that the read-access time T_A does not follow a normal distribution, but $1/T_A$ does [18]. Therefore, the read access failure probability can be expressed as [18]

$$P_A = Prob\left(\frac{1}{T_A} < \frac{1}{T_{WL}^A}\right) = \Phi\left(\frac{\left(\frac{1}{T_A}\right)_{nom} - \frac{1}{T_{WL}^A}}{\sigma_A}\right) \tag{2.13}$$

Here Φ represents the standard normal cumulative density function, T_A is the access time and, T_{WL}^A is the word-line pulse width. However, the conventional approach fails to consider many of the failure mechanisms that affect the read-access operation. We briefly discuss these mechanisms below.

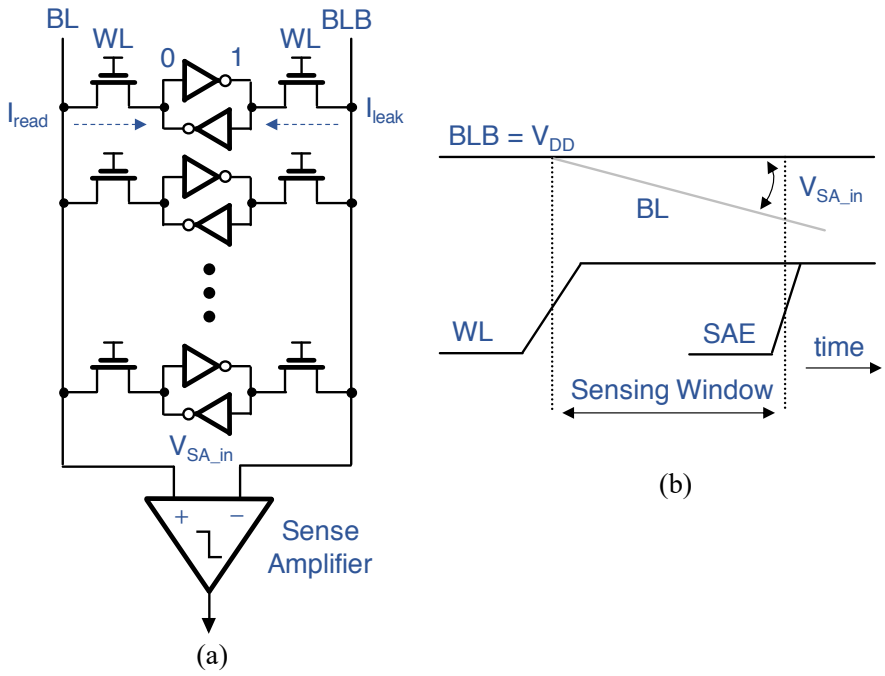


Fig. 2.8. (a) Schematic of the SRAM read-accessed column (b) Timing Diagram for the SRAM read-access operation [27].

The method described above considers a pre-defined bit-line differential voltage threshold point and ignores the sense amplifier offset distribution. Therefore, this approach considers an arbitrary worst-case point, which leads to overdesign and loss of performance. It also does not consider the negative effect of the bit-line leakage current which reduces the effective read current for a bit-cell. For a column with N bit-cells, the effective read current I_{eff} is

$$I_{eff} = I_{read} - \sum_{i=1}^{N-1} I_{off-PG_i} \tag{2.14}$$

$$I_{eff} = I_{read} - (N-1)\mu_{I_{off-PG}} \tag{2.15}$$

$$I_{eff} \approx I_{read} - (N-1)I_{off-PG}\left(1 + \frac{ln^2(10)}{2}\left(\frac{\sigma_{V_{TH}}}{S}\right)^2\right) \tag{2.16}$$

where I_{read} is the bit-cell read current, I_{off-PG} is the access transistor leakage current, $\sigma_{V_{TH}}$ is the standard deviation of threshold voltage and, S is the subthreshold slope. The effect of bit-line leakage is especially great in near-threshold and sub-threshold regions of operation where the $I_{on/off}$ ratio is severely degraded.

The above method also does not consider the sensing window, which is determined by the time elapsed

between the word-line enable and sense amplifier strobe enable. This window of time determines the total time available to develop a differential voltage on the bit-lines, as opposed to the word-line pulse width indicated in the method above. The read-accessed column and the sensing window are shown in Fig. 2.8. Small changes in the sensing window can greatly affect the read-access performance. The timing variations in both the word-line and sense amplifier strobe signal can greatly alter the sensing window, which is why it becomes imperative to consider sensing window variations when assessing the read-access failure probability. Therefore, the above method fails to capture many of the read-access failure mechanisms, thereby resulting in an underestimation of the failure probability. This results in large errors as shown in Fig. 2.9.

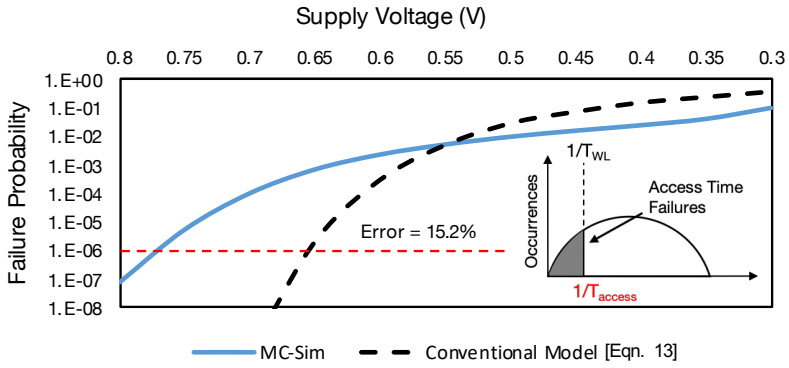


Fig. 2.9. Comparison between the failure probability evaluated using Monte-Carlo simulations and the conventional access method highlights the large error.

2.2.3. SRAM Write Access

Write failure is caused when an SRAM cell is unable to reach desired value in the time duration of the clock pulse width. Therefore, the write failure probability can be expressed as

$$P_W = Prob\left(T_W < T_{WL}^W\right) \tag{2.17}$$

where T_W is the time required to pull down the node storing 'one' and T_{WL}^W is the write word line (WL) pulse width [20],[29]. T_W cannot be easily approximated because its distribution has a tail as shown in Fig. 2.10. In the following section, we briefly discuss previous analytical approaches for determining the distribution of the write access operation and then describe the proposed modelling techniques to determine the tail of the write access distribution.

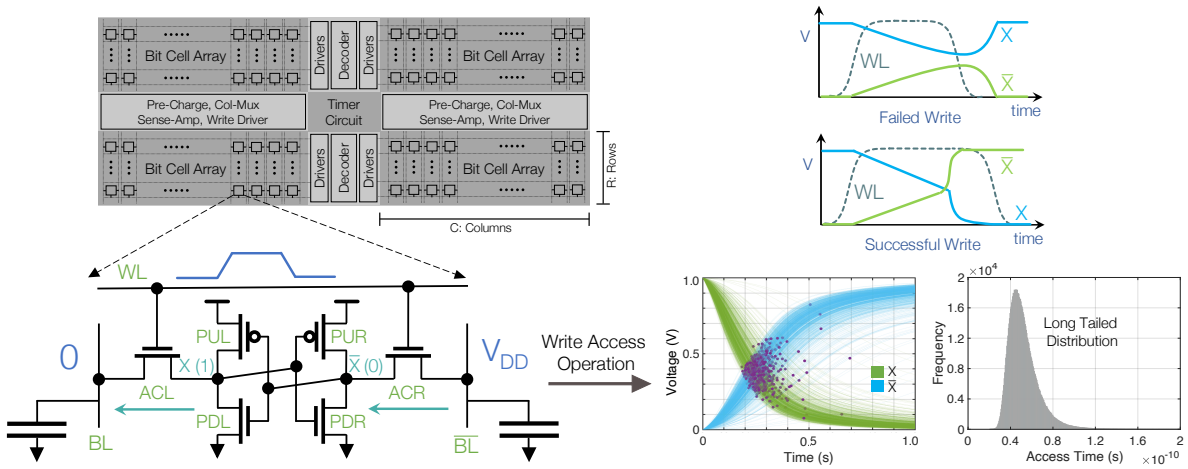


Fig. 2.10. Schematic of an SRAM bit-cell in write mode, depicting the corresponding dynamic write-access operation. Its distribution has a long tail which requires a large Monte-Carlo simulation for accurate determination and is difficult to model analytically.

2.3. Dynamic Read VMIN

2.3.1. Read-Access Model Description and Analysis

The variation in threshold voltage due to Random Dopant Fluctuations ($\sigma_{VT,RDF}$), transistor length variations ($\sigma_{VT,L}$), Random Telegraphic Noise ($\sigma_{VT,RTN}$), and other sources of variability ($\sigma_{VT,Other}$), which affect stability and performance of the cell can be modelled as given in [30].

$$\sigma_{VT} = \sqrt{\sigma^2_{VT,RDF} + \sigma^2_{VT,L} + \sigma^2_{VT,RTN} + \sigma^2_{VT,Other}} \tag{2.18}$$

For a given technology with given minimum transistor sizing (W_{min} and L_{min}), the deviation in threshold voltage ($\sigma_{V_{t_i}}$) for any transistor can be calculated by using Pelgrom's Law [31]. However, advanced technologies exhibit deviation from Pelgrom's Law. Therefore, to accurately model V_T variations, we use modified Pelgrom's Law [32], [33], which is given as

$$\sigma_{V_{t_i}} = \sigma_{VT} \times \sqrt{\frac{L_{min} W_{min}}{(W)^{\alpha}(L)^{\beta}}} \tag{2.19}$$

where α and β are technology constants.

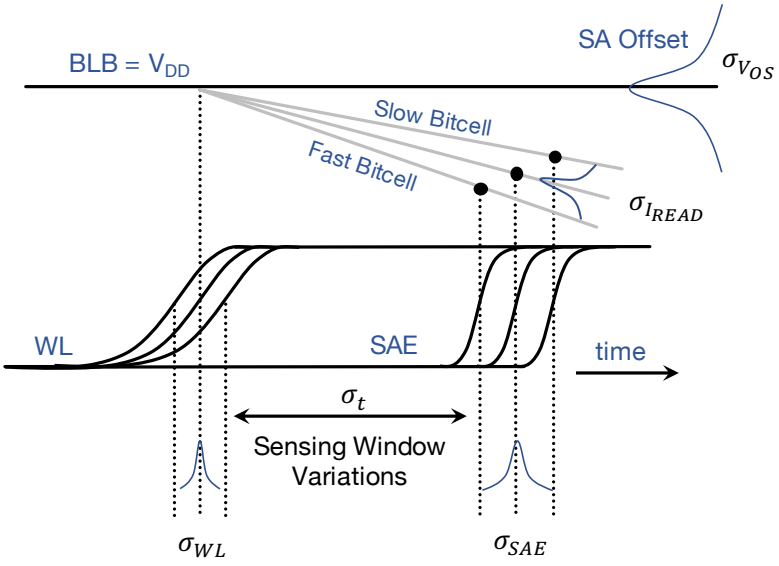


Fig. 2.11. Sources of variation that affect the read-access operation [27].

The sources of variation which affect the read access operation are shown in Fig. 2.11. Process variations in the logic circuitry path of the word-line signal cause deviations in timing, which change the time after which the

bit-line starts to discharge, thereby affecting read-access performance. The word-line logic path timing variations can be analyzed by modelling it as a chain of inverters. Let μ_{t_d} and σ_{t_d} be the mean and standard deviation, respectively, for the delay of a minimum sized inverter. For a chain of inverters, the standard deviation of delay grows as the square root of the number of stages [34]. If the word-line logic path is modelled as a chain of q inverters, then the distribution for the delay can be expressed as $Z_{WL} \sim \mathcal{N}\left(\mu_{t_{WL}}, q{\sigma_{t_d}}^2\right)$. This distribution can be scaled accordingly with change in inverter sizing [32]. Similarly, the Sense-Amplifier strobe signal (SAE) can be modelled as a chain of r inverters. Then the distribution for it can be expressed as $Z_{SAE} \sim \mathcal{N}\left(\mu_{t_{SAE}}, r{\sigma_{t_d}}^2\right)$. Therefore, the amount of time elapsed between word-line enable and SAE enable can be modelled as

$$Z_t \sim \mathcal{N}(\mu_t, {\sigma_t}^2) \tag{2.20}$$

where

$$\mu_t = \mu_{t_{SAE}} - \mu_{t_{WL}} = r\left(\mu_{t_d}\right) - q\left(\mu_{t_d}\right) = (r-q)\mu_{t_d} \tag{2.21}$$

$$\begin{aligned} {\sigma_t}^2 &= {\sigma_{t_{WL}}}^2 + {\sigma_{t_{SAE}}}^2 = \left[\sqrt{q}\left(\sigma_{t_d}\right)\right]^2 + \left[\sqrt{r}\left(\sigma_{t_d}\right)\right]^2 \\ &= (q+r){\sigma_{t_d}}^2 \end{aligned} \tag{2.22}$$

Alternatively, for a singular inverter chain of length r that is tapped at different locations to generate the SAE (at r^{th} inverter) and word-line (at q^{th} inverter; $q < r$) timing signals, the distribution can be modelled as

$$\mu_t = \mu_{t_{SAE}} - \mu_{t_{WL}} = r\left(\mu_{t_d}\right) - q\left(\mu_{t_d}\right) = (r-q)\mu_{t_d} \tag{2.23}$$

$${\sigma_t}^2 = (r-q){\sigma_{t_d}}^2 \tag{2.24}$$

The above expressions indicate that the mean of the elapsed time depends on the difference between the number of inverters in both paths, and the standard deviation depends on the number of inverters. This means that the uncertainty in timing can be quite large, thereby worsening the read-access yield. This effect is shown in Fig. 2.12 (a), where increasing the inverter chain length results in greater deviation, which dampens its intended positive effect. The read access failure probability as a function of length of the sense-amp-enable is shown in Fig. 2.12 (b). As seen in Fig. 2.12 (b), the failure probability decreases slowly with increase in inverter chain length, indicating that sense amplifier strobe signal timing does not have a very strong impact on yield. A

large change in the inverter chain length is therefore required to achieve a given yield threshold. The analysis shown in the figure can thus be very useful to precisely ascertain the sense amplifier strobe signal timing to meet specific yield targets corresponding to various memory sizes.

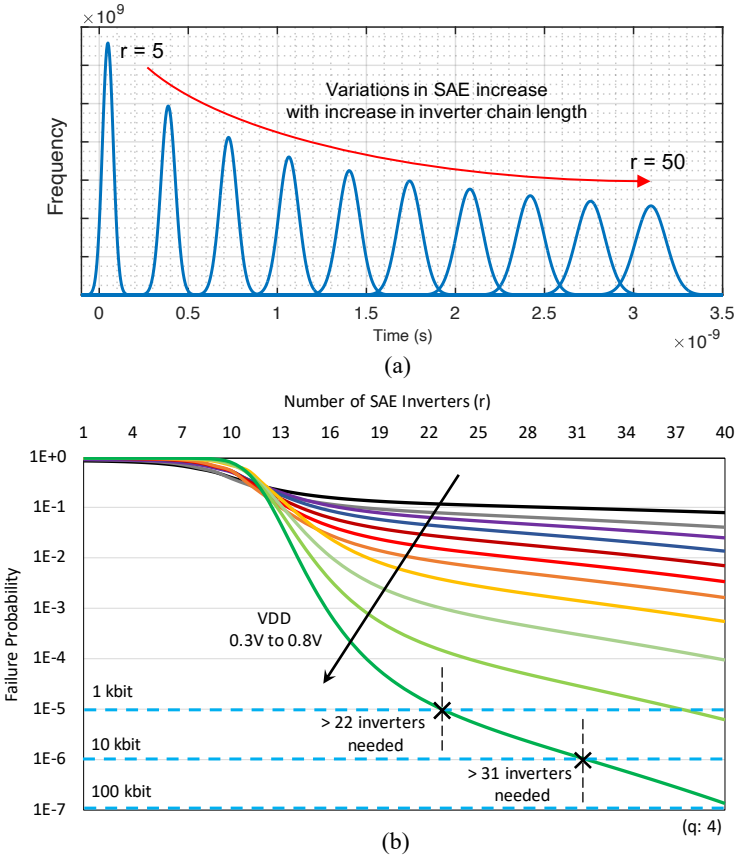


Fig. 2.12. Sense-Amp-Enable distribution with varying inverter chain length (b) Read Access failure prob. (produced using model) as a function of the number of SAE inverters across supply voltage.

Another method that can be used to generate the SAE signal is the replica-bit line [35]. In this technique, the sense amplifier enable signal is generated using replica bit-line capacitance and pre-tied bit-cells. The number of bit-cells on the replica bit-line define its discharge time and consequently the SAE timing. The timing for the replica bit-line can be modelled using the principles of the sampling distribution as

$$\mu_t = \frac{\mu_{t_{cell}}}{N} \tag{2.25}$$

$$\sigma_t = \frac{\sigma_{t_{cell}}}{\sqrt{N}} \tag{2.26}$$

where, t_{cell} is the replica bit-line discharge time for a single *ON* bit-cell on the replica bit-line and N is the

number of bit-cells that are turned on. These relationships depict exactly the opposite trend in variability in comparison to inverter chain-based techniques in which the variability increased with increase in number of elements. Another interesting observation to note about the replica bit-line technique is that its resultant timing distribution will always tend towards a gaussian distribution irrespective of region of operation due to the Central Limit Theorem. This not only makes it easier to model, but also impedes the far-off outliers as in heavy long tailed distributions.

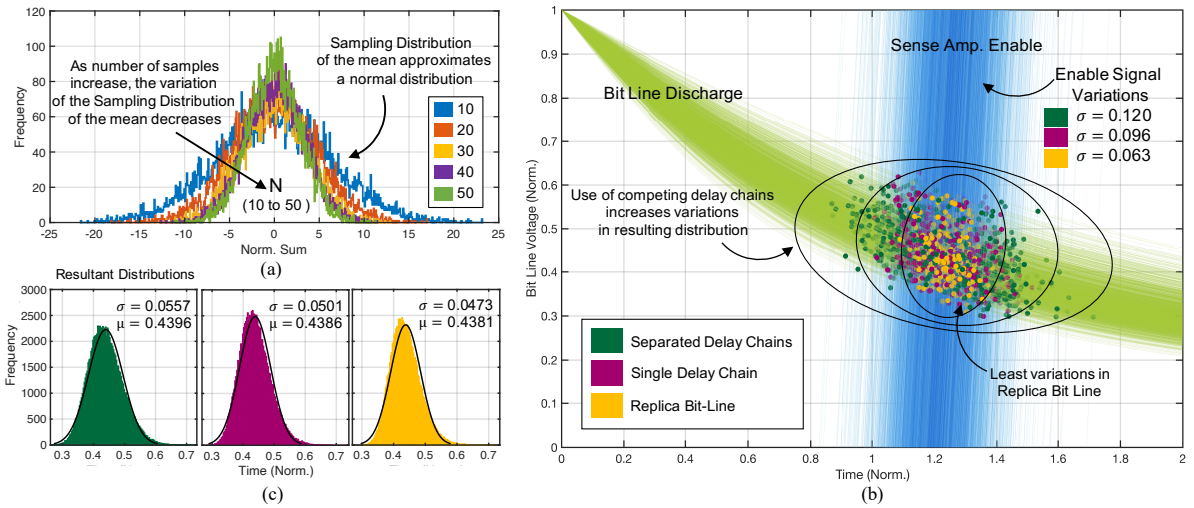


Fig. 2.13. (a) Variations in sampling (SAE timing) distribution of the mean with increase in number of samples (*ON* bit-cells) (b) Variations in SAE signal timing using various techniques of generation (c) Resultant distribution of bit-line discharge and SAE for various techniques.

The replica bit-line can either be constructed using a short fractional bit-line with very few *ON* bit-cells or using a full array length bit-line with larger number of *ON* bit-cells. Fig. 2.13 (a) shows the effect of increase in number of *ON* bit-cells in a replica bit-line on the variations. As the number of *ON* cells increase per bit-line, the variations in the timing decrease. As such, it would be desirable to have a long replica bit-line with a large capacitive load and a large number of *ON* bit-cells to minimize the variations. However, this will also increase the area and power. Fig. 2.13 (b) shows the effect of using a long bit-line and a large number of *ON* bit-cells on the SAE timing. As seen in Fig. 2.13 (b), the replica bit-line technique nearly halved the variations in the enable signal in comparison to other inverter chain-based methods, suggesting its viability in timing sensitive circuits. However, despite this large improvement, the resultant distribution of bit-line discharge and SAE sees only a modest improvement due to no change in the variations of the bit-line discharge of the accessed bit-line as seen in Fig. 2.13 (c). The overall improvement in timing variations using replica bit-line technique is then observed to be about 15%.

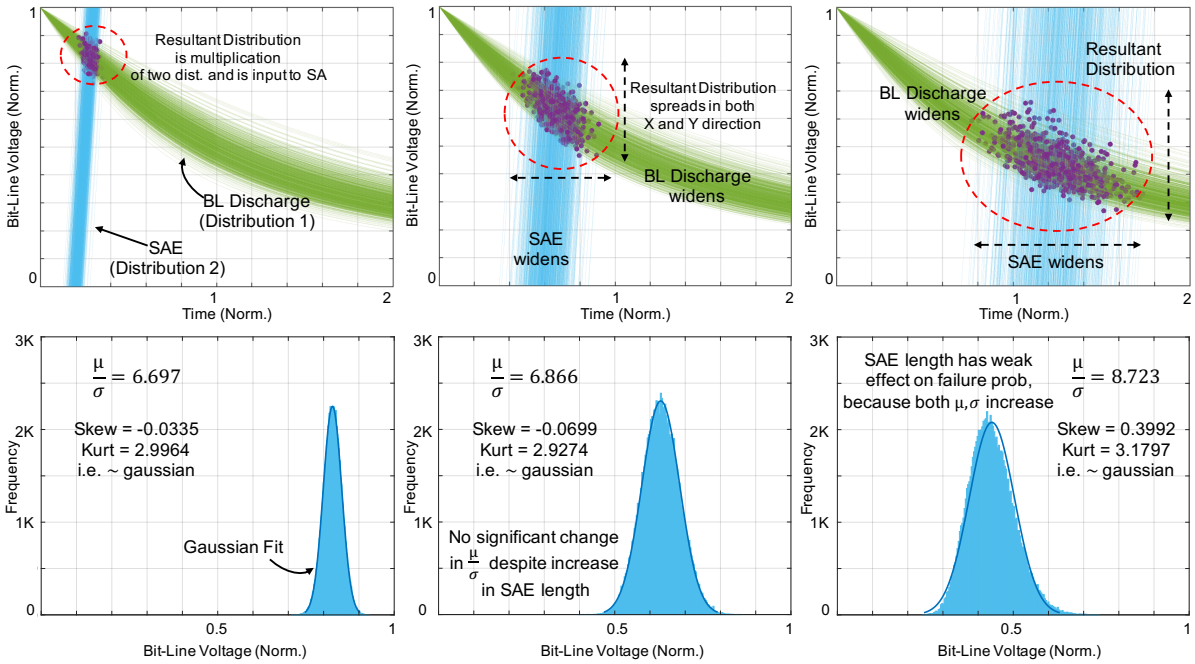


Fig. 2.14. Interaction of bit-line discharge and sense-amplifier-enable results in a near-gaussian distribution across majority of the small signal sensing operation. The same analysis also shows that change in sense amplifier strobe signal does not have a very strong effect on the resulting distribution. This trend will be similar irrespective of region of operation since resultant distribution will always spread in both *X* and *Y* directions.

The read performance depends on the variations in bit-cell read current (I_{read}). Since the bit-line discharge rate (V_r) depends on the read current, the statistical distribution of rate of bit-line discharge will follow the distribution of read current [28]. This can be expressed as

$$\left.\frac{\sigma}{\mu}\right|_{V_r} = \left.\frac{\sigma}{\mu}\right|_{I_{read}} \tag{2.27}$$

The bit-line discharge rate can be defined as the change in bit-line voltage per unit time. Its distribution is calculated based on the supply voltage, array design variables, temperature, and process variations. Although the bit-line discharge rate is nearly constant at the beginning of the read operation, it quickly falls as the bit-line voltage reduces further. The average bit-line discharge rate (α_t) can then be approximated by the following derived relationship.

$$\alpha_t = \int_0^{\mu_t} \frac{V_{DD}}{\mu_t}\left[\frac{d}{dx}\left(1 - \frac{tan^{-1}\left(\left(\frac{\Delta V_{BL}}{\Delta t}\right)x\right)}{\lim\limits_{x\to\infty} tan^{-1}\left(\left(\frac{\Delta V_{BL}}{\Delta t}\right)x\right)}\right)\right](dx) \tag{2.28}$$

Here, $(\Delta V_{BL}/\Delta t)$ represents the initial constant slope of the bit-line discharge voltage. Consequently, we can

derive the approximated distribution of the bit-line discharge rate $Z_{V_r} \sim \mathcal{N}\left(\mu_{V_r}, \sigma_{V_r}^2\right)$ as

$$\mu_{V_r} = |\alpha_t| \left(1 - \frac{(N-1)\,\mu_{I_{off-PG}}}{\mu_{I_{read}}}\right) \tag{2.29}$$

$$\sigma_{V_r} = \left| V_{DD} \frac{d}{dx} \left(1 - \frac{tan^{-1}\left(\left(\frac{\Delta V_{BL}}{\Delta t}\right) x\right)}{\lim\limits_{x \to \infty} tan^{-1}\left(\left(\frac{\Delta V_{BL}}{\Delta t}\right) x\right)} \right) \right|_{x\,=\,\mu_t} \left| \left(\frac{\sigma_{I_{read}}}{\mu_{I_{read}}}\right) \right. \tag{2.30}$$

where the read current $Z_{I_{read}} \sim \mathcal{N}\left(\mu_{I_{read}}, \sigma_{I_{read}}^2\right)$ follows

$$\mu_{I_{read}} = I_{read} + \sum_{i\,=\,1}^{n} \left(\frac{1}{2} \frac{\partial^2 I_{read}}{\partial V_{t_i}^2}\right) \sigma_{V_{t_i}}^2 + \sum_{k\,=\,1}^{n} \sum_{\substack{i\,=\,1 \\ i \neq k}}^{n} \frac{\partial^2 I_{read}}{\partial V_{t_i} \partial V_{t_k}}\, r(i,k)\, \sigma_{V_{t_i}} \sigma_{V_{t_k}} \tag{2.31}$$

$$\sigma_{I_{read}}^2 = \sum_{i\,=\,1}^{n} \left[\left(\frac{\partial I_{read}}{\partial V_{t_i}}\right) \sigma_{V_{t_i}} \right]^2$$

$$+ 2 \sum_{k\,=\,1}^{n} \sum_{\substack{i\,=\,1 \\ i \neq k}}^{n} \left(\frac{\partial I_{read}}{\partial V_{t_i}}\right) \left(\frac{\partial I_{read}}{\partial V_{t_k}}\right) r(i,k) \sigma_{V_{t_i}} \sigma_{V_{t_k}} \tag{2.32}$$

Here, $r(i,k)$ is the correlation coefficient and V_{t_i} represents the threshold voltage of the i^{th} transistor.

The read-access failure probability is the probability of the voltage differential developed between the bit-lines being less than the Sense Amplifier offset (V_{OS}). This can be expressed as

$$P_{FAIL} = Prob\left\{V_{SA_{in}} < V_{OS}\right\} = Prob\{(V_r \cdot t) < V_{OS}\}$$

$$= Prob\{\, V_r{}' < V_{OS}\} \tag{2.33}$$

Here, both bit-line discharge rate V_r and time t have been assumed to have a gaussian distribution. The distribution for the product of two gaussian variables with zero mean can be expressed as [36]

$$P_{XY}(u) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{e^{-\frac{x^2}{(2\sigma_x^2)}}}{\sigma_x \sqrt{2\pi}} \frac{e^{-\frac{y^2}{(2\sigma_y^2)}}}{\sigma_y \sqrt{2\pi}} \delta(xy - u)\, dxdy \tag{2.34}$$

$$P_{XY}(u) = \frac{K_0\left(\frac{|u|}{\sigma_x\sigma_y}\right)}{\pi\sigma_x\sigma_y} \tag{2.35}$$

where $\delta(x)$is a delta function and $K_n(Z)$ is the modified Bessel function of the second kind. Similarly, for two variables with non-zero mean, the distribution can be expressed as

$$P_{XY}(z) = \frac{1}{\pi}K_0(\bar{z}) \tag{2.36}$$

Where

$$\bar{z} = \left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) \tag{2.37}$$

To solve these equations, we calculate the first two moments of $Q = XY$ (In context of SRAM, Q represents V_{SA_in}), and then find a distribution whose parameters match the moments of Q. We shall derive the moment-generating function for Q, and show that Q can be approximated by a normal curve. We previously showed (in Fig. 2.14) how this distribution is nearly normal using data based on SRAM functional behavior. Here, we mathematically derive this approximation and quantify the limits of these assumptions. The moment-generating function for $Q = XY$ can be written as

$$M_Q(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(e^{-\frac{(x-\mu_x)^2}{2\sigma_x^2}-\frac{(x-\mu_y)^2}{2\sigma_y^2}}\right)e^{xyt}\,dxdy \tag{2.38}$$

$$M_Q(t) = \frac{exp\left\{\frac{t\mu_x\mu_y + \frac{1}{2}\left(\mu_y^2\sigma_x^2 + \mu_x^2\sigma_y^2\right)t^2}{1 - t^2\sigma_x^2\,\sigma_y^2}\right\}}{\sqrt{1 - t^2\sigma_x^2\,\sigma_y^2}} \tag{2.39}$$

Defining the variables $\delta_x = \frac{\mu_x}{\sigma_x}$ and $\delta_y = \frac{\mu_y}{\sigma_y}$, and rewriting the moment-generating function as

$$M_Q(t) = \frac{exp\left\{\frac{t\mu_x\mu_y + \left(t\delta_y^2\mu_x\mu_y + \delta_x^2\left(2\delta_y^2 + t\mu_x\mu_y\right)\right)}{2\delta_x^2\delta_y^2 - 2t^2\mu_x^2\,\mu_y^2}\right\}}{\sqrt{1 - \frac{t^2\mu_x^2\,\mu_y^2}{\delta_x^2\delta_y^2}}} \tag{2.40}$$

Although the product of two normal variables is not normally distributed, the limit of the moment-generating function is normally distributed [37]. If δ tends to increase, the moment-generating function tends to

$$M_Q(t) = exp\left\{t\mu_x\mu_y + \frac{1}{2}\left(\mu_x^2\sigma_y^2 + \mu_y^2\sigma_x^2\right)t^2\right\} \qquad (2.41)$$

The corresponding first four moments can then be written as

$$E(Q) = \mu_{V_r}{}' = \mu_{V_r}\mu_t \qquad (2.42)$$

$$V(Q) = \sigma_{V_r}^2{}' = \mu_{V_r}{}^2\sigma_t^2 + \mu_t^2\sigma_{V_r}{}^2 + \sigma_{V_r}{}^2\sigma_t^2 = \left(1 + \delta_{V_r}{}^2 + \delta_t^2\right)\sigma_{V_r}{}^2\sigma_t^2 \qquad (2.43)$$

$$\xi_3(Q) = \frac{6\delta_{V_r}\delta_t\sigma_{V_r}{}^3\sigma_t^3}{\left(\left(1 + \delta_{V_r}{}^2 + \delta_t^2\right)\sigma_{V_r}{}^2\sigma_t^2\right)^{\frac{3}{2}}} \qquad (2.44)$$

$$\xi_4(Q) = \frac{6\sigma_{V_r}{}^4\sigma_t^4\left\{2\left(\delta_{V_r}{}^2 + \delta_t^2\right) + 1\right\}}{\left(\left(1 + \delta_{V_r}{}^2 + \delta_t^2\right)\sigma_{V_r}{}^2\sigma_t^2\right)^2} \qquad (2.45)$$

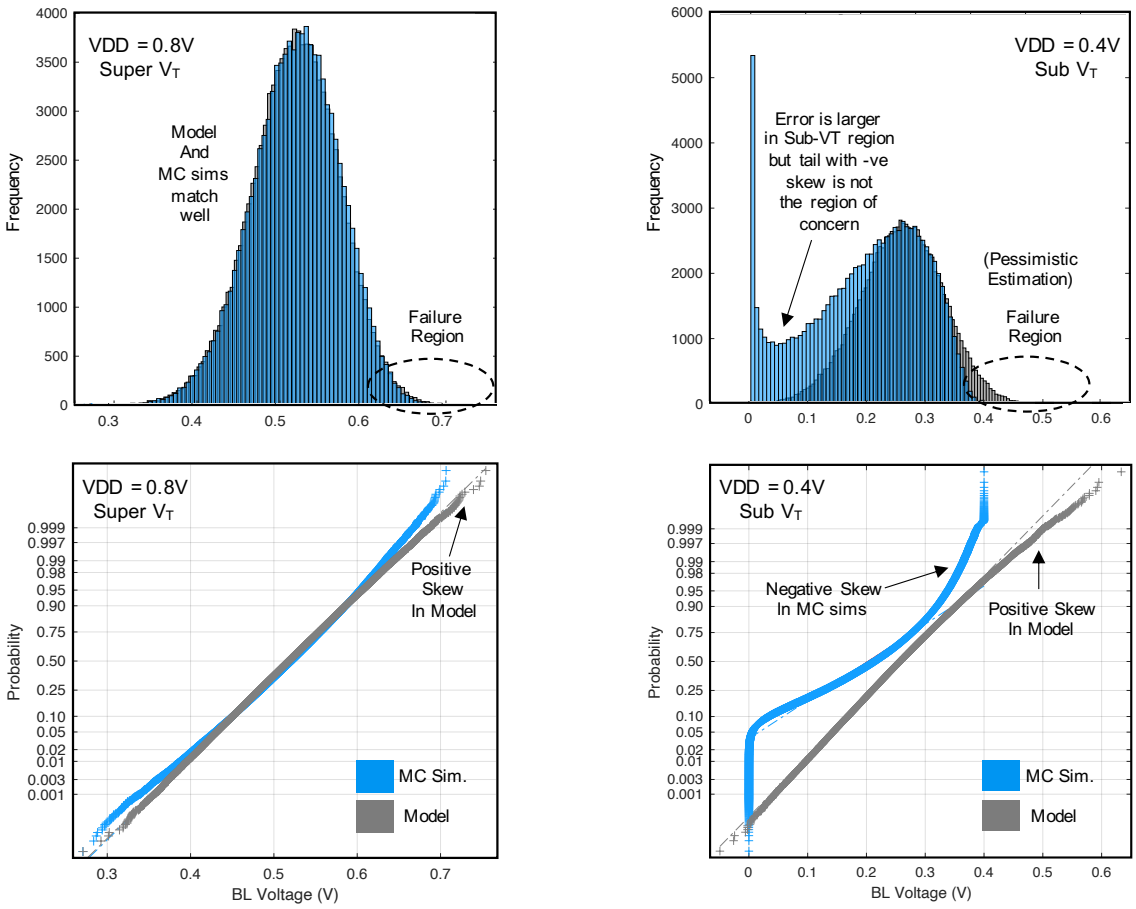


Fig. 2.15. Comparison of read-access distribution using MC simulations and proposed method in super-threshold and subthreshold regions.

The moments obtained in eqn. (2.42)-(2.45) represent the distribution of the resultant bit-line voltage (V_r'). This resultant voltage is input to the Sense Amplifier and should be less than its offset (V_{OS}) for a successful read. The model is evaluated and shown in Fig. 2.15. As seen in Fig. 2.15, the model shows near normal behavior in the super-threshold region with very little error in comparison with distributions obtained from MC simulations. The error increases in the sub-threshold region, with the model predicting the failure pessimistically. The normal probability plot shows a deviation in the right tail of about 8% when comparing the model and MC simulations in sub-threshold region. Despite this deviation, the model can provide insightful results as shown later in this.

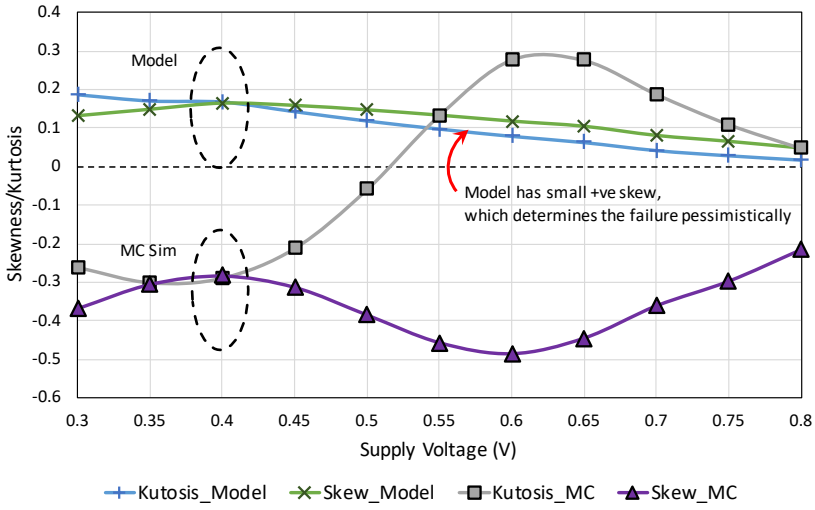


Fig. 2.16. Comparison of skew and kurtosis evaluated using model and MC simulations. Evaluation depicts a deviation of less than 0.6 skew and kurtosis in the worst case.

The skewness and kurtosis of the resulting distribution depend on the value of δ. For small δ, the skewness becomes large but is always $\leq \frac{2\sqrt{3}}{3}$. The excess or kurtosis is always ≤ 6. As $\delta \to \infty$, the skewness tends to zero. As shown in Fig. 2.16, the skewness ranges from 0.04 to 0.16 and the excess or kurtosis ranges from 0.02 to 0.18. These results suggest that the normal approximation for the product of variables is very close. The deviation from values obtained from MC simulations is also small (<0.6).

The read-access failure probability can be calculated as

$$P_{FAIL} = Prob\{V_r{}' < V_{OS}\} \tag{2.46}$$

$$\text{If } Z = V_{OS} - V_r{}', \quad \text{then } Z \sim \mathcal{N}\left(\mu_{V_{OS}} - \mu_{V_r{}'}, \sigma^2_{V_r{}'} + \sigma^2_{V_{OS}}\right)$$

$$P(Z > 0) = \int_0^{\infty} \frac{1}{\sqrt{2\pi\left(\sigma^2_{V_r{}'} + \sigma^2_{V_{OS}}\right)}} e^{\left(\frac{-\left(z+\mu_{V_r{}'}-\mu_{V_{OS}}\right)^2}{2\left(\sigma^2_{V_r{}'}+\sigma^2_{V_{OS}}\right)}\right)}(dz) \tag{2.47}$$

With $t = \frac{z+\mu'_{V_r}-\mu_{V_{OS}}}{\sqrt{2\left(\sigma'_{V_r}{}^2+\sigma^2_{V_{OS}}\right)}}$ and using the complimentary error function,

$$erfc(x) = \frac{2}{\sqrt{\pi}}\int_x^{\infty} e^{-t^2}\,(dt) \tag{2.48}$$

We get,

$$P(Z > 0) = P_{FAIL} = \frac{1}{2} erfc\left(\frac{\mu'_{V_r} - \mu_{V_{OS}}}{\sqrt{2\left(\sigma^2_{V_r'} + \sigma^2_{V_{OS}}\right)}}\right) \quad (2.49)$$

where the moments μ'_{V_r} and $\sigma^2_{V_r'}$ are shown in Eqn. (2.42) and (2.43) respectively. The yield for a given memory size (N number of cells) can then be expressed as

$$Yield = (1 - P_{FAIL})^N \quad (2.50)$$

2.3.2. Read-Access Dataset-Based Dimensional Analysis

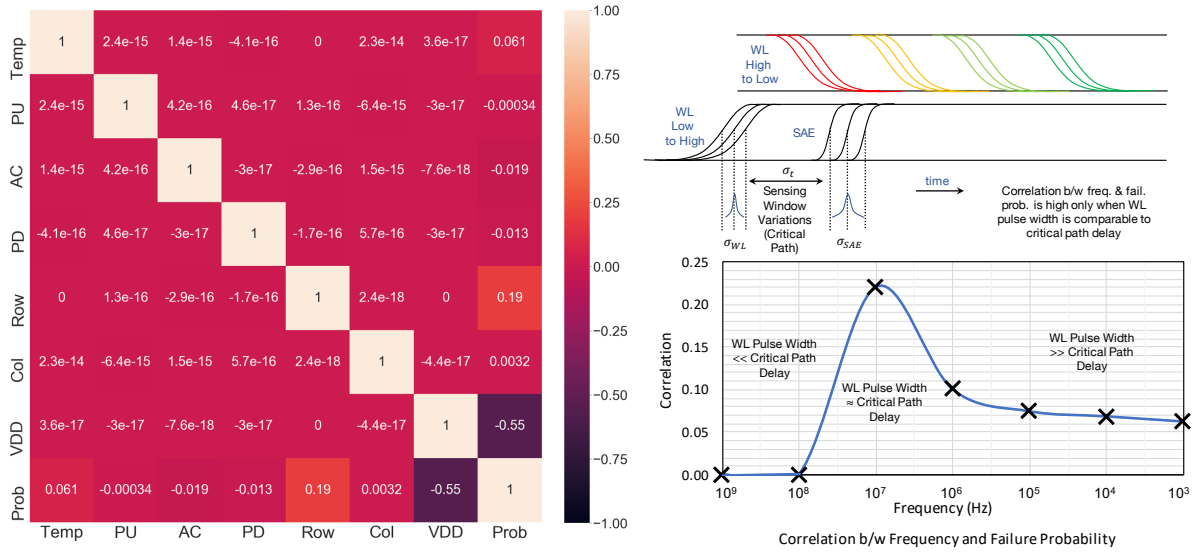
To analyze the multidimensional SRAM design space, the model is used to create a dataset with nearly 160K unique SRAM design datapoints in a bulk 65nm CMOS technology. Each datapoint is a set of values of design variables for a given design and the corresponding failure probability. All design variables are swept across a wide range to generate the dataset. To evaluate each datapoint using MC simulations would require ~5.6 hours (assuming 100K runs), which would equate to >100 years for a dataset of this size. In comparison, the model is able to generate the dataset in about 20 hours as shown in Table II. The results of the dataset are used to observe the effect of SRAM design variables, quantize their importance, and determine inter-variable correlation. The results are summarized in a correlation-matrix in Fig. 2.17 (a) and arranged in descending order of importance in Fig. 2.18.

The frequency is the only variable which spans several orders of magnitude, and thus, its effect is analyzed separately across several regions in accordance to its relative magnitude with respect to the critical path delay. Its correlation with failure probability is analyzed across three regions, when clock pulse width << critical path delay, pulse width ≈ critical path delay, and pulse width >> critical path delay, as shown in Fig. 2.17 (b). The results indicate that there is weak correlation between frequency and failure probability when pulse width >> critical path delay, suggesting the bottleneck in such a case could be the design variables involved in the critical path. The correlation peaks when the pulse width ≈ critical path delay and then falls rapidly when pulse width is reduced further. This analysis can also explain why some design points can show little to no decrease in failure rate even when the pulse width is increased indefinitely. In such a case, the unoptimized critical path variable(s) might be causing a high dynamic failure rate irrespective of the frequency.

TABLE I
SUMMARY OF EVALUATION

S. No.	Metric	Capacity	P_{FAIL} for > 95% yield	Method	Case I (r = 20)	Case II (r = 30)	Case III (r = 40)
1	VMIN (mV)	10Kbit	~1e-6	MC	855	809	776
				Prop. Method	851	805	771
		100Kbit	~1e-7	MC	881	836	803
				Prop. Method	876	832	795
		10Mbit	~1e-9	MC	922	882	846
				Prop. Method	916	874	838
2	Percentage Error	-	-	Min	0.47	0.48	0.64
				Max	0.65	0.91	1.00
				Mean	0.56	0.63	0.87
3	Time to Evaluate	-	-	Prop. Method (sec)	14.68	14.58	14.61
				MC (hours) (~1.3M runs)	72.3	72.5	72.6

r: Number of inverters in the Sense-Amplifier Enable strobe signal



(a) (b)

Fig. 2.17. (a) Correlation between various SRAM design variables and failure probability (b) Dynamic behavior of correlation between frequency and failure probability as a function of separation.

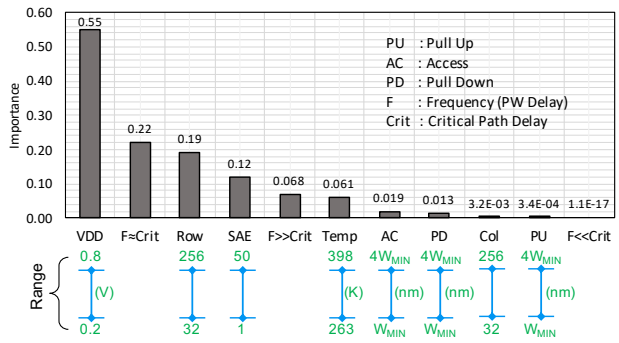


TABLE II
SUMMARY OF DATASET-BASED DESIGN SPACE ANALYSIS

Method	MC Sim.	This Work
Time to Compute	102 years	20 hours

Number of data points:159,746
Assuming each data point requires 100K MC runs
Number of CPU threads for MC eval. = 16
Number of CPU threads for Model = 1

Fig. 2.18. Importance (absolute correlation) of various SRAM design variables in descending order.

The read-access failure probability given by (2.49) has been evaluated and compared against results from Monte-Carlo simulations. The MC sims consider the entire read path including variations in the peripheral circuitry. The resulting distributions from MC sims are imported into MATLAB and then used to evaluate the final failure probability. Three cases with varying sense amplifier strobe timing has been considered as shown in Table I. The VMIN is calculated for various capacities (10Kbit to 10Mbit) and their corresponding failure rate. As seen in Table I, the time required to evaluate the VMIN is less than 15 seconds in all cases, with low error.

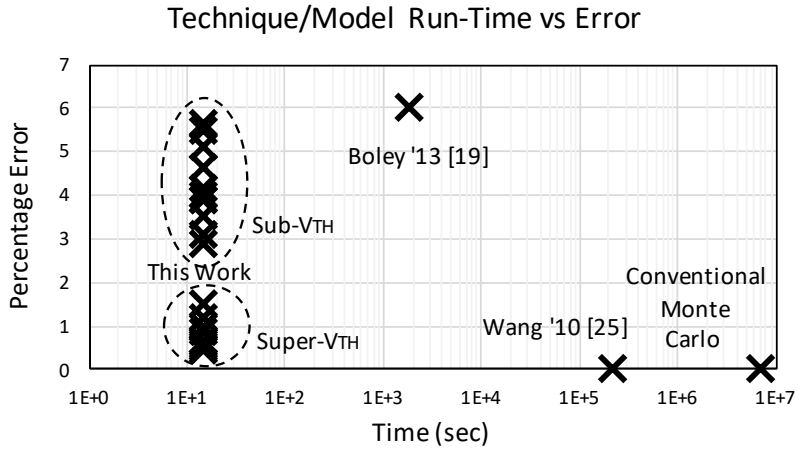


Fig. 2.19. Comparison of various yield prediction methods based on speed and error.

The comparison of various yield prediction methods based on speed and percentage error is shown in Fig. 15. Based on this comparison, including the ones shown in Table I and II, we can observe the method's convenience and effectiveness for SRAM design evaluation and exploration.

2.4. Dynamic Write VMIN

2.4.1. Distribution Transformation Approach

The work presented in [20] shows that the tail of write-access operation closely resembles the noncentral *F* distribution. The authors in [19] use sensitivity analysis to estimate the tail for write-access, but report an error of 6.83%. The authors in [18] show that by performing a linear inverse transformation on the write-access distribution, it can be transformed into a gaussian distribution, which can then be used to easily estimate the failure probability by calculating the mean ($\mu_{f(x)}$) and standard deviation ($\sigma_{f(x)}$) of the distribution as

$$\mu_{f(x)} \approx f(\mu_x) \qquad (2.51)$$

$$\sigma^2_{f(x)} \approx \sum_{i=1}^{n} \left[\left(\frac{\partial f(\mu_x)}{\partial x_i} \right) \sigma_{x_i} \right]^2 \qquad (2.52)$$

$$P_{FAIL} = Prob\left(\frac{1}{T_W} < \frac{1}{T^W_{WL}} \right) = \Phi \left(\frac{\left(\frac{1}{T_W} \right)_{nom} - \frac{1}{T^W_{WL}}}{\sigma_W} \right) \qquad (2.53)$$

Here Φ represents the standard normal cumulative density function, T_W is the write-access time and, T^W_{WL} is the word-line pulse width. The write-access operation distribution and its long tail nature is shown in Fig. 2.10. The inverse transformation of the same distribution is shown in Fig. 2.20 (a). As seen in Fig. 2.20 (a), the inverse of access time approximates the gaussian distribution. This is because the linear inverse transformation of delay makes it vary as $\propto (V_{DD} - V_T)$, i.e., linearly with change in V_T. However, this is only true for super-threshold region of operation. In sub-threshold region, the delay has an exponential dependence and a linear transformation does not yield a gaussian distribution as shown in Fig. 2.20 (b).

To transform the delay values (D) in the write-access distribution such that we can obtain a gaussian distribution in the transformed domain, we apply a transformation T as

$$T: D \rightarrow \frac{1}{K} \left| ln(|D|)^{\zeta} \right| \qquad (2.54)$$

where K is an integer constant and ζ is a fit parameter such that the skewness of the transformed distribution approximates to zero. This is because the third moment of a gaussian distribution is zero.

$$skewness\left[\left(T: f(x) \right) \right] \approx 0 \qquad (2.55)$$

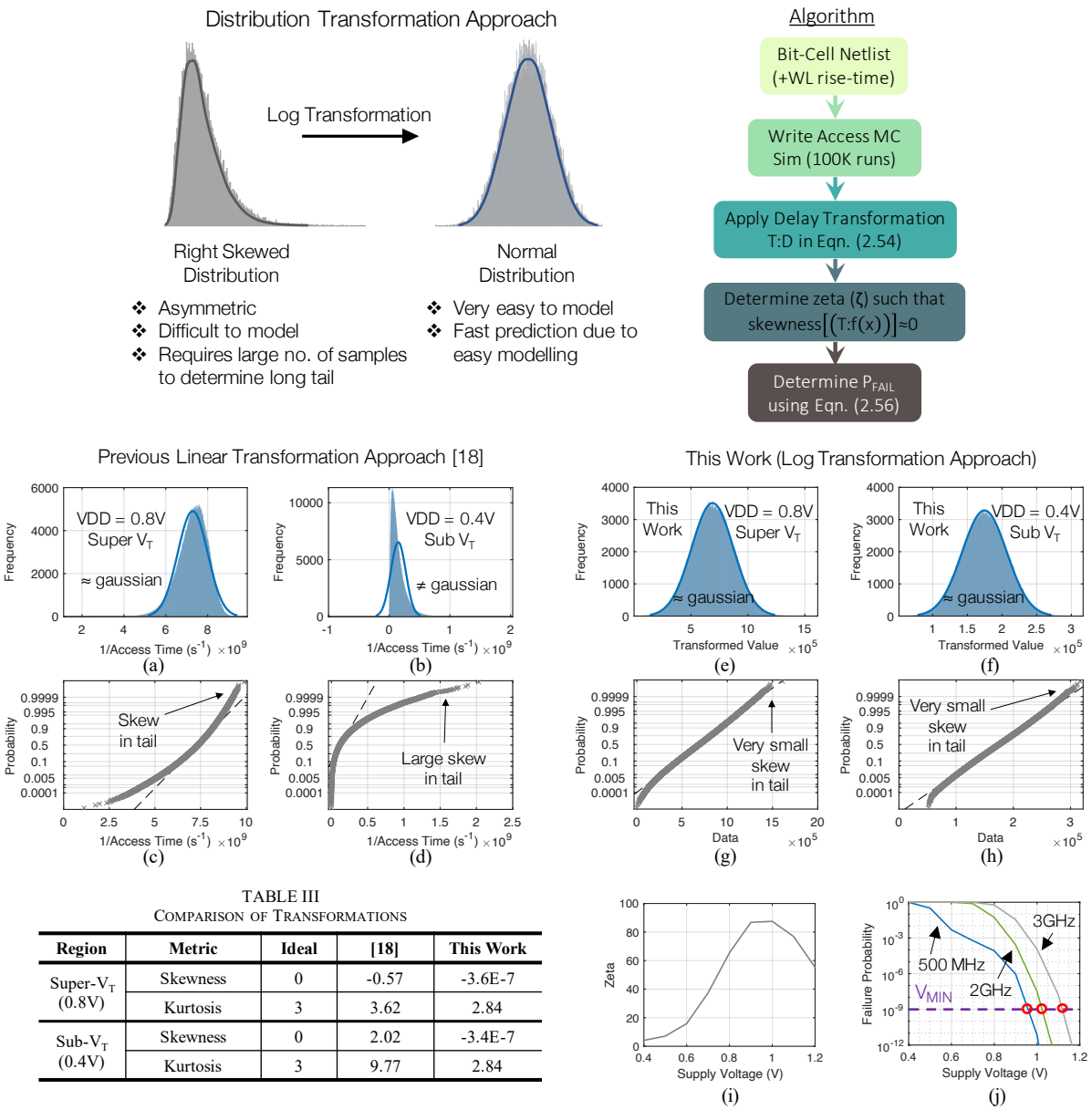


TABLE III
COMPARISON OF TRANSFORMATIONS

Region	Metric	Ideal	[18]	This Work
Super-V_T (0.8V)	Skewness	0	-0.57	-3.6E-7
	Kurtosis	3	3.62	2.84
Sub-V_T (0.4V)	Skewness	0	2.02	-3.4E-7
	Kurtosis	3	9.77	2.84

Fig. 2.20. The Log-Transformation approach can be used to transform the skewed distribution of the write-access operation into a normal distribution to easily calculate the failure probability. The histograms of the linear inverse transformation [18] distribution are shown in (a) super-threshold & (b) sub-threshold region of operation, with the corresponding probability plots in (c) & (d). The histogram of the proposed transformation model is shown in (e) & (f), with the corresponding probability plots in (g) & (h). (i) Value of zeta variable computed using 100K MC sims. (j) Trend of Write-Access failure probability and corresponding VMIN with varying supply voltage.

An initial MC simulation of 100K runs is performed and then used to estimate the value of ζ using (2.54) and (2.55). This process is repeated for any circuit or technology each time during the evaluation of the transformation and corresponding distribution. The value of ζ as a function of V_{DD} is shown in Fig. 2.20 (i).

The write-access delay values of 100K runs are then transformed using (5), after which the failure probability is quickly calculated as

$$P_{FAIL} = Prob\left(\frac{1}{K}\left|ln(|T_W|)^{\zeta}\right| < \frac{1}{K}\left|ln\left(\left|T_{WL}^{W}\right|\right)^{\zeta}\right|\right)$$

$$= \Phi\left(\frac{\frac{1}{K}\left|ln(|T_W|)^{\zeta}\right|_{\mu} - \frac{1}{K}\left|ln\left(\left|T_{WL}^{W}\right|\right)^{\zeta}\right|}{\frac{1}{K}\left|ln(|T_W|)^{\zeta}\right|_{\sigma}}\right) \qquad (2.56)$$

The transformed distribution, along with the linear transformation is shown in Fig. 2.20 (e). While the inverse transformation is a good fit for super-threshold region, it fails to work in the subthreshold region. Whereas the proposed transformation modelling allows linearity (i.e., gaussian nature) and very little skew in both regions of operation. The failure probability was calculated using (2.56) for different access frequencies and is shown in Fig. 2.20 (j). The write-access VMIN is calculated when the failure probability reaches 10E-9. Table III shows the comparison of transformation moments for various analytical methods and compares them with the ideal case. The closer the transformation moments are to the ideal case, the lower we can expect the error to be. The error in calculation and time to compute are summarized in Table IV, with the process steps for evaluation in Fig. 2.20.

2.4.2. Non-Central F Distribution Approach

Now we describe how to analytically estimate the tail of the distribution of the write-access operation in an SRAM by fitting a noncentral F distribution to it. This is accomplished by first estimating the moments of the distribution assuming V_{TH} as the variable and then mapping them on to a noncentral F distribution. The variation in threshold voltage due to Random Dopant Fluctuations ($\sigma_{VT,RDF}$), transistor length variations ($\sigma_{VT,L}$), Random Telegraphic Noise ($\sigma_{VT,RTN}$), and other sources of variability ($\sigma_{VT,Other}$), which affect stability and performance of the cell can be modelled as

$$\sigma_{V_T} = \sqrt{\sigma_{VT,RDF}^2 + \sigma_{VT,L}^2 + \sigma_{VT,RTN}^2 + \sigma_{VT,Other}^2} \qquad (2.57)$$

For a given technology with given minimum transistor sizing W_{MIN} and L_{MIN}, the deviation in threshold voltage ($\sigma_{V_{t_i}}$) for any transistor can be calculated by using Pelgrom's Law. Advanced technologies exhibit deviation from Pelgrom's Law, which are modelled as

$$\sigma_{V_{T_i}} = \sigma_{V_T} \times \sqrt{\frac{W_{MIN}L_{MIN}}{(W)^{\alpha}(L)^{\beta}}} \qquad (2.58)$$

where α and β are technology constants. The moments associated with V_T can then be represented as

$$\mu_{V_T}\big|_1 = \mu_{V_T} = E(V_T) \tag{2.59}$$

$$\mu_{V_T}\big|_2 = \sigma_{V_T}^2 = E\big(V_T - \mu_{V_T}\big)^2 \tag{2.60}$$

$$\mu_{V_T}\big|_3 = E\big(V_T - \mu_{V_T}\big)^3 \tag{2.61}$$

$$\mu_{V_T}\big|_4 = E\big(V_T - \mu_{V_T}\big)^4 \tag{2.62}$$

If $x_1, x_2, \dots x_n$ are independent random variables with mean η_i and variance σ_i, then the function $y = f(x_1, x_2, \dots x_n)$ can be expressed using the multivariable Taylor series expansion as

$f(x_1, \dots x_n)$

$$= f(\eta_1, \dots \eta_n) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}\bigg|_{\eta_1, \dots \eta_n} (x_i - \eta_i) + r(x_1, \dots x_n) \tag{2.63}$$

where $r(x_1, x_2 \dots x_n)$ is the higher order term. Applying this result to V_T and including the first few terms in $r(x_1, x_2 \dots x_n)$, $f(x_1, x_2, \dots x_n)$ can be approximated to

$$f(V_T) \cong f\big(\mu_{V_T}\big) + \big(V_T - \mu_{V_T}\big)\frac{\partial f}{\partial V_T} + \frac{1}{2}\big(V_T - \mu_{V_T}\big)\frac{\partial^2 f}{\partial V_T^2} \tag{2.64}$$

$$\mu_f\big|_1 = E[f(V_T)] \cong f\big(\mu_{V_T}\big) + E\big(V_T - \mu_{V_T}\big)^2 \frac{\partial^2 f}{\partial V_T^2} \tag{2.65}$$

Considering the third and fourth order moments of V_T, a few more terms in $r(x_1, x_2 \dots x_n)$ can be included to evaluate the variance of f as

$$\mu_f\big|_2 = \sigma_{V_T}^2 = Var\big(f(V_T)\big) = E\big[f(V_T) - \mu_f\big]^2 \tag{2.66}$$

$$\cong \sigma_{V_T}^2 \left(\frac{\partial f}{\partial V_T}\right)^2 - \frac{1}{4}\sigma_{V_T}^2 \left(\frac{\partial^2 f}{\partial V_T^2}\right)^2 + E\big(V_T - \mu_{V_T}\big)^3 \frac{\partial f}{\partial V_T}\frac{\partial^2 f}{\partial V_T^2}$$

$$+ \frac{1}{4}E\big(V_T - \mu_{V_T}\big)^4 \left(\frac{\partial^2 f}{\partial V_T^2}\right)^2 \tag{2.67}$$

For calculating the third moment, we first calculate $E[f(V_T)^2]$ and $E[f(V_T)^3]$ as

$$E[f(V_T)^2] = E\left[\left\{f\big(\mu_{V_T}\big) + \big(V_T - \mu_{V_T}\big)\frac{\partial f}{\partial V_T}\right\}^2\right] \tag{2.68}$$

$$= f\big(\mu_{V_T}\big)^2 + E\big(V_T - \mu_{V_T}\big)^2 \left(\frac{\partial f}{\partial V_T}\right)^2 \tag{2.69}$$

$$E[f(V_T)^3] = E\left[\left\{f(\mu_{V_T}) + (V_T - \mu_{V_T})\frac{\partial f}{\partial V_T}\right\}^3\right] \tag{2.70}$$

$$= f(\mu_{V_T})^3 + E(V_T - \mu_{V_T})^3\left(\frac{\partial f}{\partial V_T}\right)^3$$

$$+3f(\mu_{V_T})E(V_T - \mu_{V_T})^2\left(\frac{\partial f}{\partial V_T}\right)^2 \tag{2.71}$$

Using eqn. (2.68) and (2.70), the third moment can be calculated as

$$\mu_f\big|_3 = E[f(V_T) - \mu_f]^3$$

$$= E[f(V_T)^3] - 3E[f(V_T)]E[f(V_T)^2] + 2E[f(V_T)]^3 \tag{2.72}$$

$$\mu_f\big|_3 = 3f(\mu_{V_T})^2 E(V_T - \mu_{V_T})^2\frac{\partial^2 f}{\partial V_T^2} + E(V_T - \mu_{V_T})^3\left(\frac{\partial f}{\partial V_T}\right)^3$$

$$+\left[E(V_T - \mu_{V_T})^2\right]^2\frac{\partial^2 f}{\partial V_T^2}$$

$$\times\left[6f(\mu_{V_T})\frac{\partial^2 f}{\partial V_T^2} + 2E(V_T - \mu_{V_T})^2\left(\frac{\partial^2 f}{\partial V_T^2}\right)^2 - 3\left(\frac{\partial f}{\partial V_T}\right)^2\right] \tag{2.73}$$

Applying these results to the six transistors of the SRAM cell, we can rewrite the moments as

$$\mu_f\big|_1 = f(\mu_{V_T}) + \frac{1}{2}\sum_{i=1}^{6}\left(\frac{\partial^2 f}{\partial V_T^2}\right)\sigma_{V_{T_i}}^2 \tag{2.74}$$

$$\mu_f\big|_2 = \sum_{i=1}^{6}\left[\left(\frac{\partial f}{\partial V_{T_i}}\right)\sigma_{V_{T_i}}\right]^2 - \frac{1}{4}\sum_{i=1}^{6}\left[\left(\frac{\partial^2 f}{\partial V_{T_i}^2}\right)\sigma_{V_{T_i}}\right]^2$$

$$+\mu_{V_T}\big|_3\sum_{i=1}^{6}\left[\left(\frac{\partial f}{\partial V_{T_i}}\right)\left(\frac{\partial^2 f}{\partial V_{T_i}^2}\right)\right] + \frac{1}{4}\mu_{V_T}\big|_4\sum_{i=1}^{6}\left[\left(\frac{\partial^2 f}{\partial V_{T_i}^2}\right)^2\right] \tag{2.75}$$

$$\mu_f\big|_3 = 3f(\mu_{V_T})^2\sum_{i=1}^{6}\left(\frac{\partial^2 f}{\partial V_{T_i}^2}\right)\sigma_{V_{T_i}}^2 + \mu_{V_T}\big|_3\sum_{i=1}^{6}\left(\frac{\partial f}{\partial V_{T_i}}\right)^3 + 6f(\mu_{V_T})\sum_{i=1}^{6}\left(\frac{\partial^2 f}{\partial V_{T_i}^2}\right)^2\left\{\sigma_{V_{T_i}}^2\right\}^2$$

$$+2\sum_{i=1}^{6}\left(\frac{\partial^2 f}{\partial V_{T_i}^2}\right)^3\left\{\sigma_{V_{T_i}}^2\right\}^3 - 3\sum_{i=1}^{6}\left(\frac{\partial f}{\partial V_{T_i}}\right)^2\left(\frac{\partial^2 f}{\partial V_{T_i}^2}\right)\left\{\sigma_{V_{T_i}}^2\right\}^2 \tag{2.76}$$

The moments are evaluated using the differential coefficients extracted from Fig. 2.21 (a)-(d) by curve fitting a polynomial whose degree matches the order of the differential coefficient. These are then be used to calculate

the parameters of the F distribution (n_1, n_2, λ) by equating (2.74) to (2.76) with (A3) to (A5) respectively. Solving these, we get the values of n_1, n_2, λ, which are then used to obtain a noncentral F distribution $ncf(n_1, n_2, \lambda)$. Representing the skewness and kurtosis values of this distribution as $skew(ncf)$ and $kurt(ncf)$ respectively, the final noncentral F distribution for the write-access operation can be represented by the following moments.

$$Write - Access = \left\{\mu_f\big|_1, \mu_f\big|_2, skew(ncf), kurt(ncf)\right\} \tag{2.77}$$

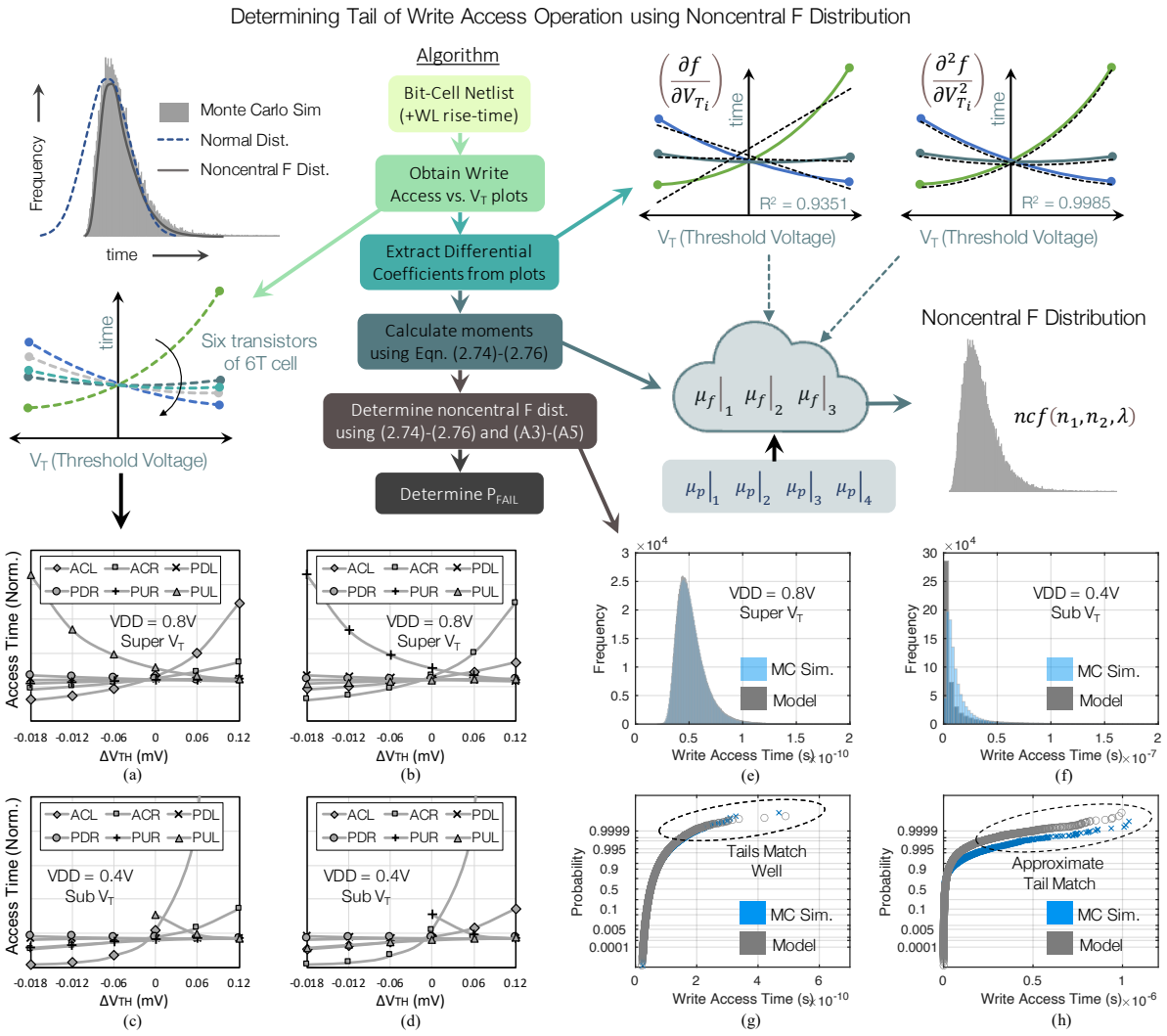


Fig. 2.21. Determining write-access failure probability by modelling it using Noncentral F distribution. Normalized Write Access-Time variation with respect to change in V_T for each transistor in super-threshold region (a) when $X = 1$ (b) $X = 0$ and in subthreshold region (c) when $X = 1$ (d) $X = 0$. Comparison of write-access operation distributions using proposed modelling technique and MC sims in (e) Super-threshold region (f) Sub-threshold region of operation, with the corresponding probability plots in (g) & (h).

The distribution has been evaluated using eqn. (2.77) in both sub-threshold and super-threshold regions of operation using a commercial 65nm bulk technology and shown in Fig. 2.21 (e) and (f) respectively. The

distribution matches well in super-threshold region with very low mismatch in probabilities. The means of the distributions don't match exactly in the subthreshold region, but the tails match approximately, with a difference of ~4E-5 in probabilities as shown in Fig. 2.21 (d). The error in evaluation and the time to compute are summarized in Table IV. The process steps for performing the analysis are shown in Fig. 2.21.

2.4.3. Modelling in Advanced FinFET Technologies

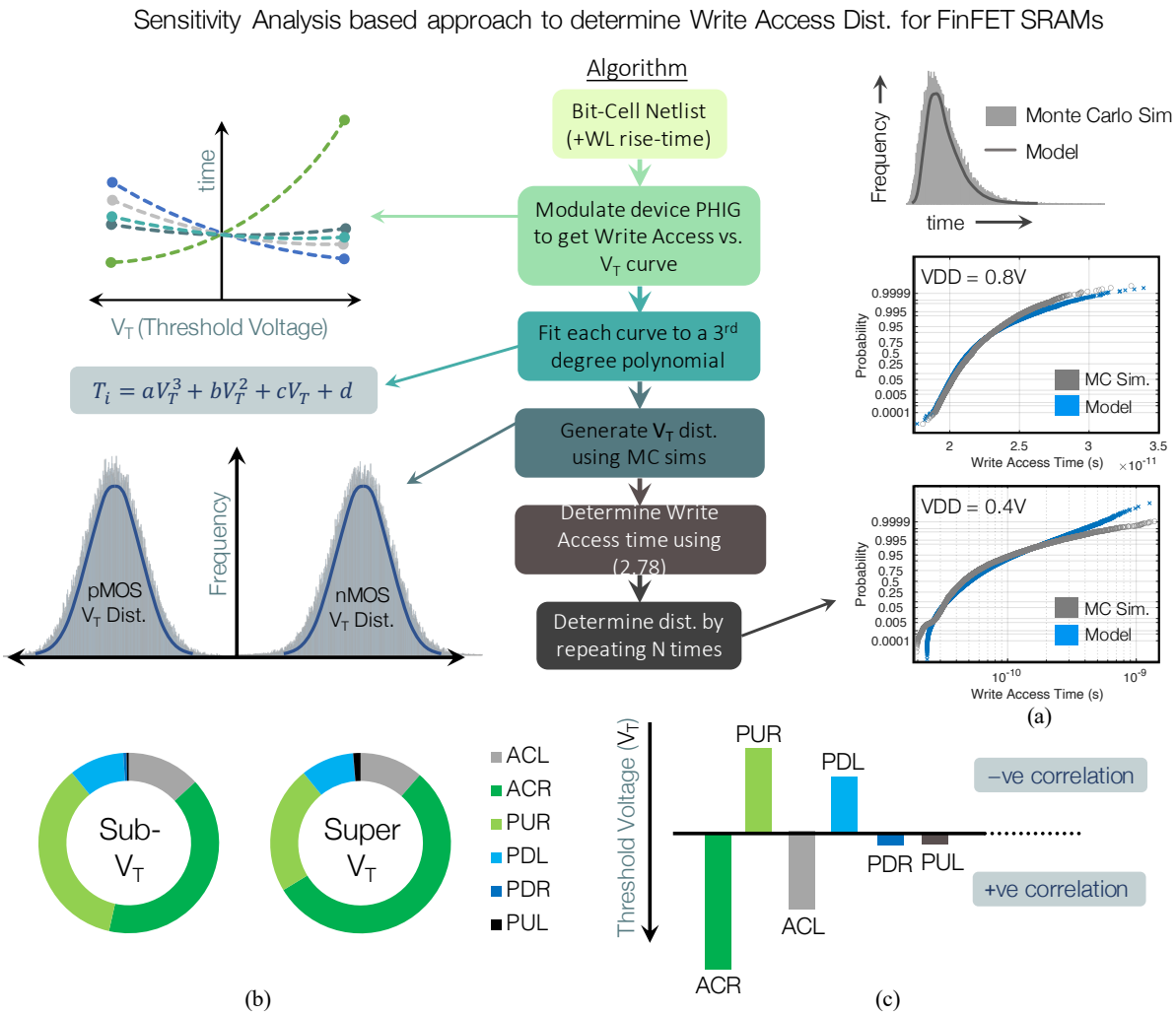


Fig. 2.22. (a) Comparison of write-access operation distributions using sensitivity analysis and MC sims in 12nm FinFET process (1M iterations) (b) The relative impact and contribution of each transistor in the 6T bit-cell on the write access operation performance in super-threshold and subthreshold region of operation. The contribution of the pull-up transistor on the hold 'one' (and write 'zero') side of the bit-cell increases in subthreshold region. (c) Correlation between threshold voltage of transistor and write-access time for each transistor in the 6T bit-cell.

In comparison to planar bulk technologies, FinFET devices provide higher performance with improved short-channel effects, subthreshold slope, drive current, and mismatch [38]. However, the implementation and development of these advanced devices involves major technical challenges, including an increase in V_T

variations in sub-30nm process technologies due to LER, RDF, and work function variations (WFV) [39],[40]. These can seriously degrade the V_T mismatch in various circuit blocks of the Integrated Circuit (IC). Additionally, it has been estimated that the magnitude of WFV-induced V_T variations is larger than that induced by either LER or RDF in sub-30-nm process technology [41],[42].

In this work, to assess the dynamic write-access operation in FinFET based SRAMs, we extend sensitivity analysis in [19] to observe the effect of V_T variations due to WFV. We perform the simulations on a commercial 12nm FinFET technology. The SRAM bit-cell transistor fins are sized as 1:1:2 for pull-up, access, and pull-down respectively. The V_T for both nMOS and pMOS devices is modulated by changing the metal gate work function (PHIG) in the BSIM-CMG model. With change in V_T, the dynamic write-access time is calculated and used to generate access time (T_i) versus V_T curves corresponding to each i^{th} transistor in the bit-cell [19]. A third-degree polynomial is fit to each curve as

$$T_i = aV_T^3 + bV_T^2 + cV_T + d \tag{2.78}$$

The offset write-access time (T_{OFFSET_i}) for each transistor is then calculated by subtracting the nominal write-access time ($T_{nominal}$) from T_i. As opposed to bulk planar technologies, the V_T variable is not explicitly described in the device model. As such, information about the V_T cannot be directly extracted from the model file and extrapolated using Pelgrom's Law.

TABLE IV
SUMMARY OF ERROR IN WRITE-ACCESS VMIN AND EVALUATION TIME

Method	Analytical	Error (Super-V_{TH})	Error (Sub-V_{TH})	Time to Evaluate
MC sim. (6 sigma)	No	-	-	Months*
MC sim. (1M runs)	No	-	-	10 hrs.
Statistical Blockade [25]	No	<1%	<1%	60 hrs.
Sensitivity Analysis (SA) [19]	Yes	4.5%	6.83%	32 min.
Linear Transformation (LT) [18]	Yes	<2%	<15%	1 hr.
This Work (Log Transformation)	Yes	<2%	<4%	1 hr.
This Work (Noncentral F-Dist.)	Yes	<3%	<7%	10 min.
This Work (SA)	Yes	<3%	<7%	23 min.

* Estimated from [25]

In this work, an initial simulation of 100K MC runs is performed for all the devices in the bit-cell to generate the V_T distributions. The V_T data samples are then plugged into Eqn. (29), to calculate the dynamic write-access time as

$$T_{WRITE-ACCESS} = T_{nominal} + \sum_{i=1}^{n} T_{OFFSET_i} \tag{2.79}$$

where n is the number of transistors in the bit-cell. This calculation in (2.79) is repeated N times depending on the desired sample size to generate the final dynamic write-access operation distribution. The distribution has been evaluated (1M iterations) using Eqn. (2.79) and compared with MC sims in Fig. 2.22. As seen in Fig. 2.22, the proposed analysis matches well with MC sims, including in the tail. The process steps for performing the analysis are shown in Fig. 2.22, with the runtime and error in evaluation summarized in Table IV.

2.4.4. Contention Limited Write-Access Failures

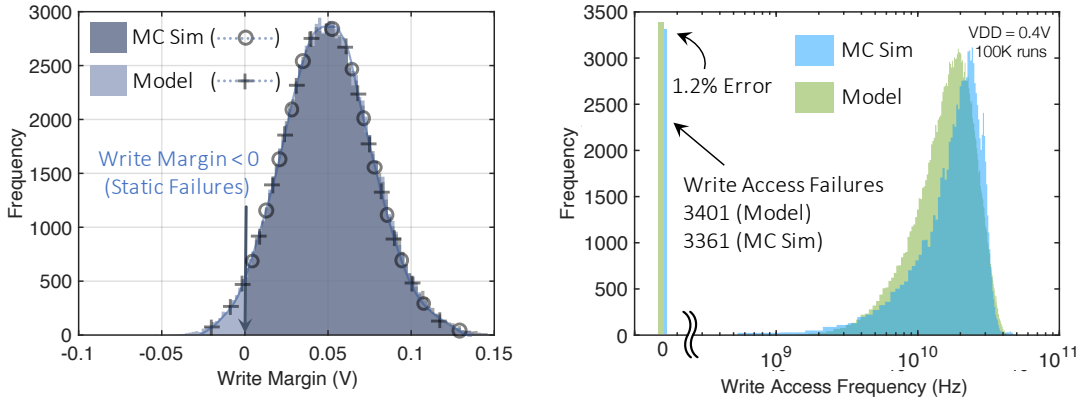


Fig. 2.23. Comparison of (a) Static write margin distribution using model and MC sims. at V_{DD} = 0.4V (100K iterations) (b) write-access distribution with respect to frequency at V_{DD} = 0.4V. Results show about 1.2% error in computation of contention-limited write-access failures.

In this section, the method to calculate the contention-limited write-access failures is presented. These failures are static in nature and occur irrespective of word-line pulse width. Since these failures occur due to insufficiency of write margin (WM) in the bit-cell, they can be modelled using the static write margin equations as

$$\mu_{WM} = WM + \sum_{i=1}^{n}\left(\frac{1}{2}\frac{\partial^2 WM}{\partial V_{t_i}^2}\right)\sigma_{V_{t_i}}^2 + \sum_{k=1}^{n}\sum_{\substack{i=1 \\ i\neq k}}^{n}\frac{\partial^2 WM}{\partial V_{t_i}\partial V_{t_k}}\, r(i,k)\,\sigma_{V_{t_i}}\sigma_{V_{t_k}} \qquad (2.80)$$

$$\sigma_{WM}^2 = \sum_{i=1}^{n}\left[\left(\frac{\partial WM}{\partial V_{t_i}}\right)\sigma_{V_{t_i}}\right]^2$$

$$+ 2\sum_{k=1}^{n}\sum_{\substack{i=1 \\ i\neq k}}^{n}\left(\frac{\partial WM}{\partial V_{t_i}}\right)\left(\frac{\partial WM}{\partial V_{t_k}}\right) r(i,k)\sigma_{V_{t_i}}\sigma_{V_{t_k}} \qquad (2.81)$$

Here, $r(i,k)$ is the correlation coefficient and V_{t_i} represents the threshold voltage of the i^{th} transistor. The total number of contention-limited write-access failures are then calculated using the cumulative distribution function as

$$Total\ Failures = \frac{N}{2}\left[1 + erf\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right] \qquad (2.82)$$

where *N* is the total number of samples. The write margin distribution was computed for a 6T bit-cell with fin ratio 1:1:2 at 0.4V using both MC simulations and model equations and plotted in Fig. 2.23 (a). As seen in Fig. (2.23), the distribution fits the gaussian curve and both MC and model distributions match well. The area under the curve from negative infinity to zero write margin represents the total number of samples which will fail statically. In Fig. 2.23 (b), the write-access distribution was plotted for the same bit-cell using the model presented in Fig. 2.22 and compared against MC simulations. The contention-limited write-access failures calculated using (2.82) were also added to the same plot to represent the entirety of the write-access distribution. As seen in Fig. 2.23 (b), the model and MC simulations match well even at low operating voltages (0.4V), with about 1.2% error in the number of contention-limited write-access failures.

2.4.5. Effect of Assists on Write Access

As SRAMs have continued to scale, they have required the adoption of assist circuits to ensure scaling. This adoption becomes necessary because the quantized nature of transistor fins leaves little room for VMIN or performance adjustment, lest the designer gives up array area efficiency. Therefore, it has become imperative to determine which assist technique is most suited for the maximal improvement in VMIN or performance depending on the application use case. In this section, we investigate the effect of various write-assist techniques on the dynamic write performance of 6T bit-cells (1:1:2 fin ratio) from V_{DD} = 0.4V to V_{DD} = 0.8V using the 12nm technology. The nominal supply voltage is 0.8V and the threshold voltages for nFET and pFET are about 350mV and -350mV respectively. The analyses are performed using MC simulations considering assist circuit voltages as a percentage of the supply voltage (10% and 20%) and compared against MC simulations with no assists in Fig. 2.24. The assist circuits include Word-Line Boosting (WLB), V_{DD} collapse (VDDU), and Negative Bit-Line (NBL) [43]. The Write-Margin [17],[44],[45] has also been calculated and summarized for varying bit-cell fin ratios and write-assist techniques in Table V.

Simulations show that the WLB assist brings about the most in VMIN reduction in bit-cells with fin ratios 1:1:1 and 1:1:2, whereas the NBL assist is best at reducing the VMIN in the 1:2:2 bit-cell. The WLB assist has the greatest overall impact on write-access performance and the VDDU and NBL assist techniques have the modest impact across operating voltages and amount of assist voltage applied. The distribution is long tailed

when no assist is applied, meaning that the outliers take an exceptionally long time to write. The outliers can differ by as much as two orders of magnitude as shown in Fig. 2.24 (c) and (d). The assist circuits greatly impact these outliers and transform the distribution to more gaussian-like in super-threshold region of operation. In near-threshold region of operation, the tail is impacted even more so, thereby greatly shortening the tail. Even though the overall performance of the WLB technique is the best, it has the worst row half-select stability issue because it exacerbates the static read noise margin problem during a pseudo-read operation in half selected cells in interleaved memories. While a careful circuit implementation can be used to somewhat mitigate this effect, the trade-off between the read-stability and write-ability remains delicate. The NBL technique doesn't impact the row half-select stability, but it impacts the stability of column half selected cells. However, the work in [43] suggests that the probability of half-select stability issues is very unlikely for smaller (<30%) NBL assist values. As such, NBL assist can offer the best overall trade-off between write-access performance improvement and half-select stability issues. The VDDU technique can impact both row and column half-select stability of other cells depending on the designer's implementation. It is up to the designer to decide and trade-off various read and write assist techniques to balance both performance and stability across the entire array of cells.

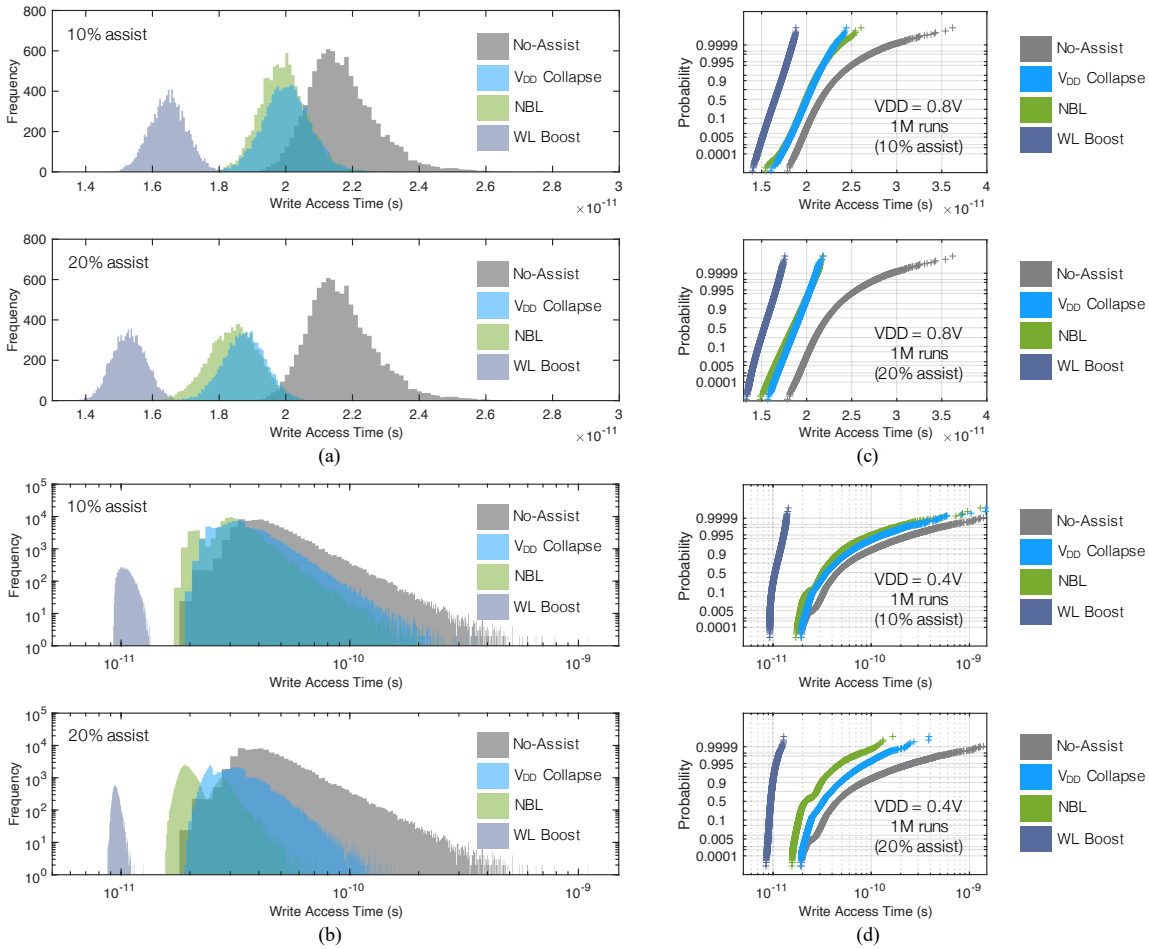


Fig. 2.24. Performance comparison of write-access operation histograms with various write assist techniques at (a) V_{DD} = 0.8V (b) V_{DD} = 0.4V and write-access operation probability plots at (c) V_{DD} = 0.8V (d) V_{DD} = 0.4V in 12nm FinFET process.

TABLE V
SUMMARY OF 6σ WRITE-MARGIN VMIN USING VARIOUS ASSIST TECHNIQUES

Bit-Cell Fin Ratio	No Assist (mV)		V_{DD} Collapse (mV)		Negative BL (mV)		WL Boost (mV)	
	MC	Analytical	MC	Analytical	MC	Analytical	MC	Analytical
1:1:1	794.5	808.8	676.6	687.4	690.0	700.3	664.2	674.9
1:1:2	829.1	838.2	699.1	706.1	716.8	724.7	692.3	701.3
1:2:2	674.4	680.5	593.1	599.0	550.6	557.1	565.5	571.6

MC simulation (100K runs) is performed with 20% assists.

2.5. Conclusion

In this work, we discussed the major mechanisms which affect the SRAM read-access operation and the common evaluation techniques which help to analyze them. The benefits and issues of various such analytical and semi-analytical techniques were discussed. To evaluate the read-access failure probability and the corresponding VMIN, we presented a fast analytical model which investigates key SRAM read-access components and analytically models their behavior in the small signal sensing region of operation. This model takes into account several variables, such as the supply voltage, temperature, process variations and, array design variables i.e. bit-cell sizing, read current, bit-line capacitance (number of rows), word-line rise time (number of columns), sense amplifier strobe timing, bit-line leakage, and sense amplifier offset voltage. Simulations in a commercial bulk 65nm technology showed that the proposed method is able to evaluate the failure probability within a few seconds (~15 sec) with small error. With this gain in speed over other evaluation methods, the model is used to evaluate about 160K different SRAM designs. The results of this evaluation were used to analyze the multidimensional SRAM design space and determine the importance of various design variables. This analysis also provided insightful results about the effect of operating frequency and sense-amplifier strobe timing on read access failure probability.

Thus, the proposed read access model can be very useful for SRAM designers to quickly calculate design feasibility and analyze the design space to optimize power, area, and speed.

In this work, we also showed how the write-access operation in an SRAM has a skewed and long tailed distribution which sets the failure threshold for the dynamic write operation. Analytical approaches which use transformation and sensitivity analysis are viable methodologies to evaluate the tail. While previous analytical approaches are successfully able to trade off speed and accuracy in comparison to MC simulations, they are only able to do so in super-threshold region of operation and introduce large errors in subthreshold region. We presented an analytical methodology to estimate the tail of the write-access operation using a log transformation model which works well in all regions of operation. We also provided an analytical solution to the tail of the write-access operation distribution which is known to closely resemble the noncentral F distribution. Furthermore, we presented a sensitivity analysis-based evaluation method to determine the write-access operation distribution in advanced FinFET technologies. The dynamic write-access failures can also include

failures which are independent of pulse-width and are caused by insufficiency of static write margin. Therefore, we presented a methodology to extend all these methods to include the evaluation of contention-limited write-access failures as well. Since in advanced technologies, assist circuits play a crucial role in enabling performance and stability, we evaluated and compared various write assist techniques for different bit-cell fin ratios and regions of operation.

Appendix

Moments of the Non-Central F Distribution

Let $(X_1, X_2, \dots, X_i, \dots, X_k)$ be k independent, normally distributed with means μ_i and unit variances. Then the random variable $\sum_{i=1}^{k} X_i^2$ is distributed according to the noncentral chi-squared (χ^2) distribution. It has two parameters: k specifies the number of degrees of freedom and λ (also called the non-centrality parameter) is related to the mean of the random variables X_i by $\lambda = \sum_{i=1}^{k} \mu_i^2$. Then, if $X: \chi_{n_1}^2(\lambda_1)$ and $Y: \chi_{n_2}^2(\lambda_2)$ are two independently distributed noncentral chi-squared variables with n_1 and n_2 degrees of freedom respectively, then the doubly noncentral F distribution can be defined as $F = (\chi_{n_1}^2/n_1)/(\chi_{n_2}^2/n_2)$ [46]. With $\lambda_1 \neq 0, \lambda_2 = 0$, the distribution is called the singly noncentral F distribution and its probability density function is defined as

$$p(x) = e^{-\lambda/2+(\lambda n_1 x)/[2(n_2+n_1 x)]}\, {n_1}^{n_1/2}\, {n_2}^{n_2/2}\, x^{n_1/2-1}\, (n_2 + n_1 x)^{-(n_1+n_2)/2}$$

$$\frac{\Gamma\left(\frac{1}{2}n_1\right)\Gamma\left(1+\frac{1}{2}n_2\right) L_{n_2/2}^{n_1/2-1}\left(-\frac{\lambda n_1 x}{2(n_2+n_1 x)}\right)}{B\left(\frac{1}{2}n_1, \frac{1}{2}n_2\right)\Gamma\left(\frac{1}{2}(n_1+n_2)\right)} \tag{A1}$$

$$= \frac{1}{B\left(\frac{1}{2}n_1, \frac{1}{2}n_2\right)} e^{-\lambda/2} {n_1}^{n_1/2}\, {n_2}^{n_2/2} x^{n_1/2-1}\, (n_2 + n_1 x)^{-(n_1+n_2)/2}\, {}_1F_1\left(\frac{1}{2}(n_2+n_1); \frac{1}{2}n_1; \frac{\lambda n_1 x}{2(n_2+n_1 x)}\right) \tag{A2}$$

where $\Gamma(z)$ is the Gamma function, $B(\alpha, \beta)$ is the Beta function, $L_m^n(z)$ is a generalized Laguerre polynomial, and ${}_pF_q$ is a Hypergeometric Function. The first few central moments are then evaluated as

$$\mu_p\big|_1 = \mu = \frac{n_2(\lambda + n_1)}{n_1(n_2-2)} \tag{A3}$$

$$\mu_p\big|_2 = \sigma^2 = \frac{2n_2^2[\lambda^2 + 2(n_1+n_2-2)\lambda + n_1(n_1+n_2-2)]}{n_1^2(n_2-2)^2(n_2-4)} \tag{A4}$$

$$\mu_p\big|_3 = \frac{8n_2^3\begin{bmatrix} 2\lambda^3 + 6(n_1+n_2-2)\lambda^2 \\ +3(n_1+n_2-2)(2n_1+n_2-2)\lambda \\ +n_1(n_1+n_2-2)(2n_1+n_2-2)\end{bmatrix}}{n_1^3(n_2-2)^3(n_2-4)(n_2-6)} \tag{A5}$$

$$\mu_p\big|_4 = \frac{12n_2^4\begin{bmatrix} (n_2+10)\lambda^4 + 4(n_2+10)(n_1+n_2-2)\lambda^3 \\ +2(3n_1+2n_2-4)(n_2+10)(n_1+n_2-2)\lambda^2 \\ +8(n_1+n_2-2)(3n_1+n_2-2)(2n_1+n_2-2)\lambda \\ +4(n_1+n_2-2)^2(n_2-2)(n_1+2)\lambda \\ +2n_1(n_1+n_2-2)(3n_1+n_2-2)(2n_1+n_2-2) \\ +n_1(n_1+n_2-2)^2(n_2-2)(n_1+2)\end{bmatrix}}{n_1^4(n_2-2)^4(n_2-4)(n_2-6)(n_2-8)} \tag{A6}$$

3. Ultra-Low Power Memory Design

3.1. Motivation and Prior Art

Sensor node applications like medical and wearables require higher lifetime and miniaturization. Such nodes try to trade-off form factor and lifetime due to energy harvesting and battery size limits. Moreover, sensor nodes for these applications typically spend most of their time in standby, only waking up periodically to sense and store data in an on-chip memory until it needs to be processed or transmitted. Due to their significant idle time, the energy consumption of these nodes is dominated by leakage power. Fig. 3.1 shows some typical examples of sensor nodes and their power breakdown by sub-component type. As seen in Fig. 3.1, the memory can contribute greatly to the total power budget and constraint the node lifetime. In many of the IoT applications, there is a limited availability of harvestable energy. For e.g., environmental sources (moonlight, moisture), radio frequency sources (GSM, 3G, WiFi), and biological sources (glucose, O2, endocochlear) have pW to nW levels of harvestable energy [47]-[52]. In addition, size constraints for many medical or wearable applications also limits the total amount of harvestable energy through the transducer. With typical memory sizes of at least a few KBs, the power budget of these nodes would be unsustainable with pW/bit leakage power. As such, power down to fW/bit is required for sustained node operation in these applications.

To improve system energy and lifetime, memories for sensor nodes have used a multitude of techniques to reduce leakage power. Scaling of supply voltage to operate into subthreshold region for a linear reduction in standby power has been a popular choice but is severely limited by yield sensitivities to process-voltage-temperature (PVT) variations in deep sub-micrometer region. But achieving subthreshold operation is a challenge in itself. It is greatly affected by VTH (threshold voltage) mismatch due to random dopant fluctuations (RDF) and line edge roughness (LER), thereby limiting memory yields [53]. Sensitivity to voltage and temperature variations is also exacerbated in this region. In addition, the device mismatch in deep sub-micrometer technologies is greatest in minimum sized devices, which is often the case with SRAMs [54]. To enable voltage scaling, many bit-cell designs, architectures, circuit techniques, and device types have been proposed. On the device and technology level, there have been implementations of custom DRAM-based superSRAM cells [55], custom device-based XLL-SRAM [56], and Silicon-on-Thin-Box (SOTB) SRAM [57]. These implementations allowed reduction in leakage down to the high fW/bit or low pW/bit range. Although

these SRAMs can achieve low leakage power, they are often implemented in older technologies and in other cases, very specialized and often expensive technologies. As such, their use can prohibit integrated system scaling. At the bit-cell level, the ten-transistor custom bit-cell design allowed supply scaling down to 0.2V, leading to 1.65pW/bit [58]. Another custom bit-cell design in an FDSOI technology allowed to reduce the power down to 7.4pW/bit [59]. Other implementations employ a variety of circuit techniques like low-V_{DD} operation, power gating, and back-biasing [60]. Although these can allow fW/bit level power consumption, they are often implemented in older technologies which inherently have very low leakage.

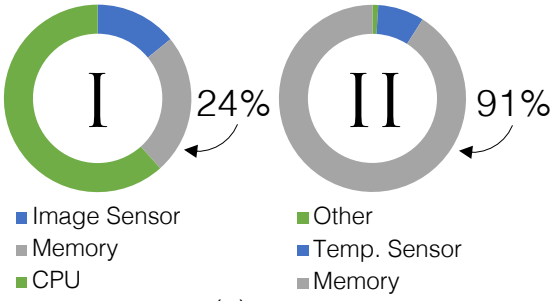


Fig. 3.1. Energy consumption by functional unit in example sensor nodes [63].

Additionally, the efficacy of such circuit techniques can be limited or have diminishing returns as one transitions to newer technology nodes. Non-Volatile memories do not consume standby leakage power but have high write energy [61], that leads to a prohibitive amount of average power consumption for many IoT applications.

3.2. Dynamic Leakage Suppression (DLS) Logic

Conventional SRAMs implement the six transistor (6T) bit-cell structure shown in Fig. 2.2. This structure allows high density, bit-interleaving and fast differential sensing but suffers from half-select stability, read-disturb stability, and conflicting read and write sizing. The maximum noise margin in conventional 6T SRAMs is limited by the inherent VDD/2 limit of the transfer characteristic of an inverter. This problem is exacerbated by the weak effect of device sizing on noise margins in subthreshold region of operation [53], [64]. These inherent issues with the 6T cell's topology prevent aggressive voltage scaling to reduce leakage power consumption. As such, new circuit techniques have been proposed that allow to maintain functional robustness at low supply voltages. By aggressively scaling the supply voltage, these memory designs are then able to reduce leakage power which scales linearly with VDD. However, this approach allows for limited reduction in leakage power because there is only so much room for supply voltage reduction when these designs are already operating at low voltages in near or sub-threshold regions of operation. In this work, we aim to reduce the leakage power primarily by suppressing leakage current instead of relying mainly on scaling the supply voltage. We achieve this by using Dynamic Leakage Suppression (DLS) logic [62],[65]-[70].

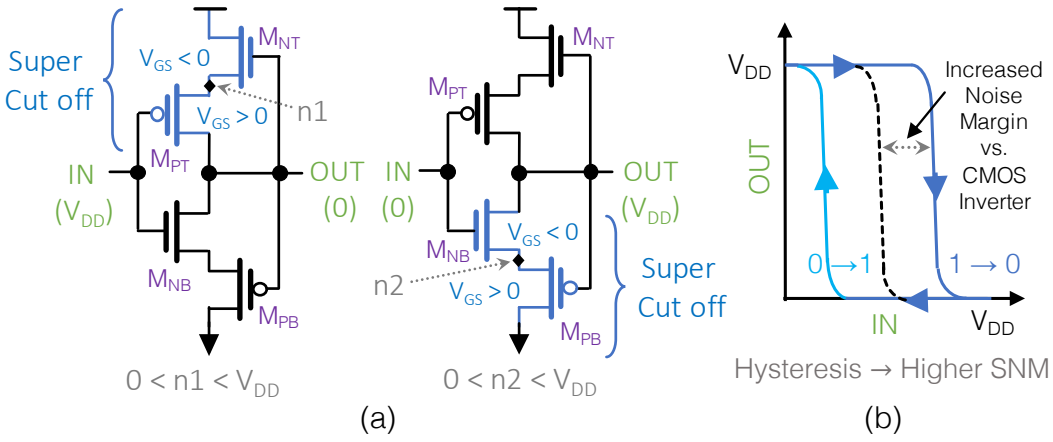


Fig. 3.2. (a) Circuit schematic of the DLS Inverter demonstrating the super-cut-off mechanism. (b) Hysteresis in DLS inverter.

Fig. 3.2 (a) shows an inverter made using DLS logic for both low and high input voltages. The output voltage of the inverter is fed-back to the bottom pMOS and the top nMOS, thereby putting all leaking transistors in a super cut-off state. When IN = 0, the leakage current flows through the pull-down logic. Since the gate of the pMOS in the pull-down network is connected to a high OUT voltage, the internal node n2 settles to a voltage

approximately half of V_{DD}, thereby making both transistors in the pull-down network develop a negative VGS i.e., super cut-off. Conversely, the same super cut-off mechanism occurs in both transistors in the pull-up network when IN = V_{DD}. This super cut-off mechanism allows for an ultra-low leakage state. Measurement results show that the leakage power of a DLS inverter is up to 6790x lower than the conventional CMOS inverter in super-threshold region and up to 787x lower in sub-threshold region as shown in Fig. 3.3 (a).

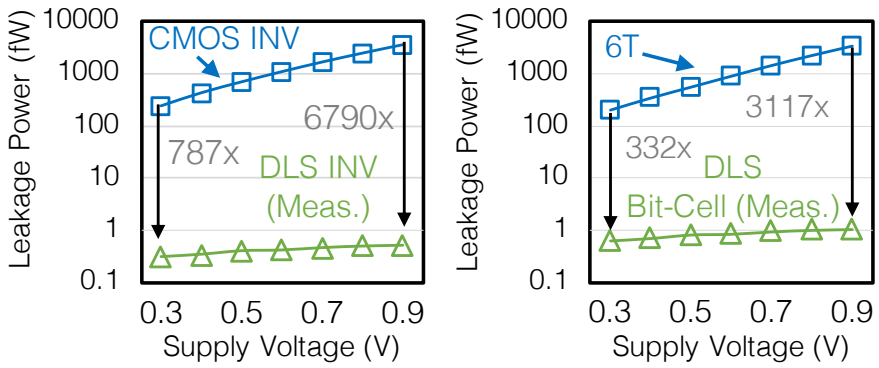


Fig. 3.3. Leakage Power comparison between conventional (a) CMOS and DLS inverter and (b) 6T and DLS bit-cell.

During the dynamic operation of DLS logic, the output node transitions using the leakage currents of the top and bottom transistors, which are in initially in super cut-off (nMOS for the rising transition and pMOS for the falling output transitions). As the IN node goes from 0V to V_{DD}, M_{NB} switches from super cut-off to weak-inversion and starts to equalize the voltage of n2 and OUT. This makes M_{PB} go from super cut-off to regular cut-off point. As M_{NB} pulls up the n2 node, OUT is also simultaneously discharged. This causes both M_{NT} and M_{PT} to go into super cut-off, thereby greatly reducing the leakage from the supply voltage to the OUT node. Whereas, M_{PB} being no longer in super cut-off, continuously acts to pull down the OUT node to low logic, and consequently suppresses the leakage through the header and accelerates the discharge of the OUT node. Due to this super cut-off feedback effect, DLS logic naturally has different rising and falling switch points, resulting in hysteresis and a large increase in static noise margin in comparison to a regular CMOS inverter [69] as shown in Fig. 3.2(b).

3.3. DLS SRAM

3.3.1. Architecture of the Proposed Bit-cell

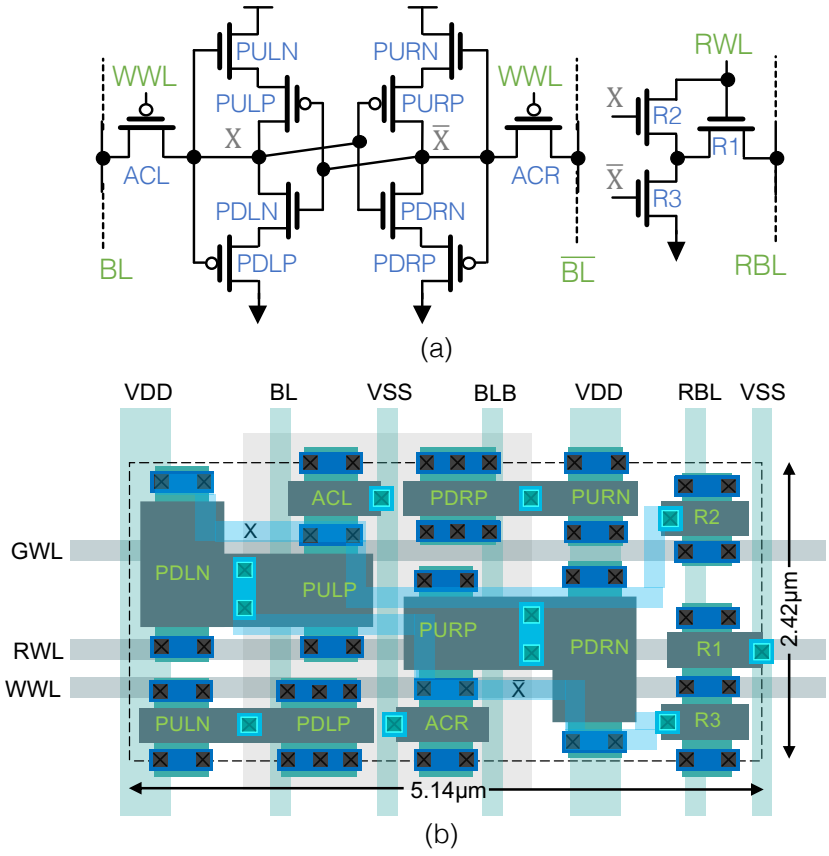


Fig. 3.4. (a) Circuit schematic of the Proposed Dynamic Leakage Suppression (DLS) Logic based 13T bit-cell with an isolated read port and pMOS access transistors. (b) Layout of the proposed bit-cell.

The proposed bit-cell uses the same cross-coupled inverter structure and access transistors as the conventional 6T bit-cell to store a single bit of information. The conventional CMOS inverters are replaced by DLS logic style-based inverters to reduce the leakage power. The read and write ports are decoupled to enable contention free single-ended reads as shown in Fig. 3.4. The proposed bit-cell is implemented using IO devices with dimensions to ensure sufficient write-read ability and noise margin. The IO devices have thick gate oxide and high VTH to reduce subthreshold leakage and prevent the onset of gate leakage at high V_{DD}. This translates to up to 3117x and 787x lower leakage power consumption than the conventional 6T bit-cell in super-threshold region and subthreshold region respectively as shown in Fig. 3.3 (b). The bit-cell can be prone to data

disturbances due to leakage currents from other un-accessed bit-cells on the same column as shown in Fig. 3.5 (a). To circumvent this, the nMOS based access transistors are replaced by pMOS transistors (shown in Fig. 3.5 (b)) and their gate voltage is boosted to super-cut-off them, thereby reducing leakage current and preventing inadvertent writes. This is explained in more detail in the next sub-section.

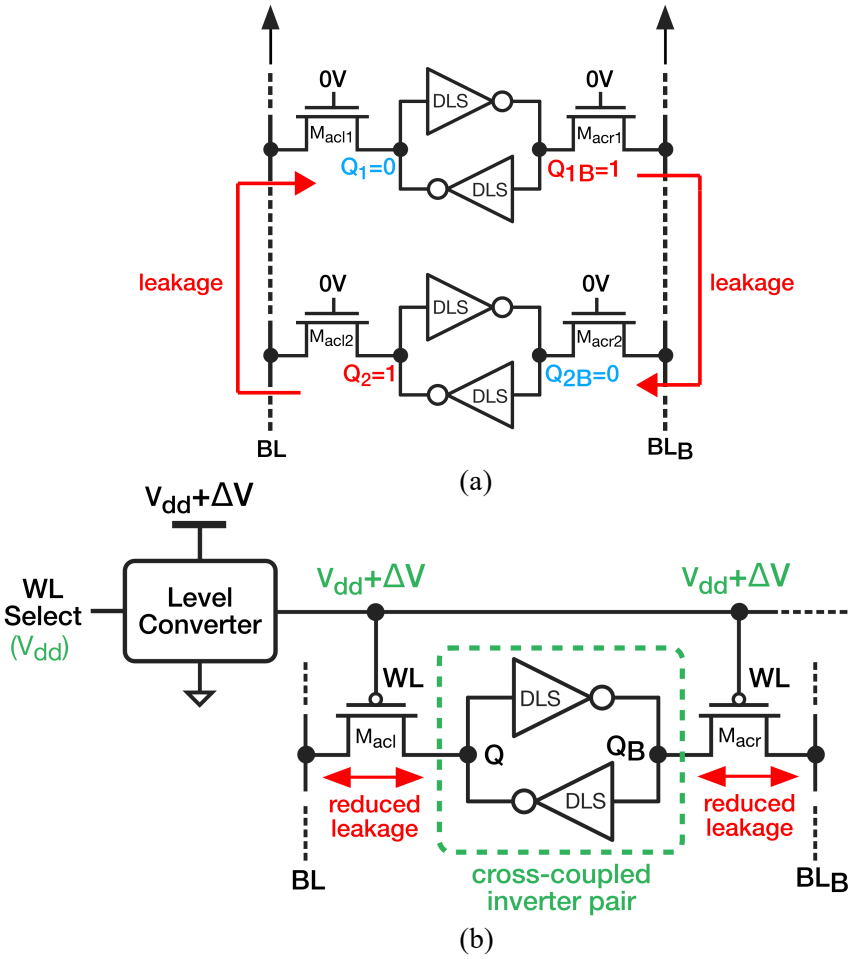


Fig. 3.5. (a) Circuit schematic of the proposed bit-cell with nMOS access transistors shows BL leakage disturbances. (b) Use of pMOS access transistors with boosted gate voltage allows to reduce BL leakage and prevent inadvertent disturbances.

The large savings in power consumption come at the cost of increased area. Nevertheless, special care was taken to minimize the area footprint and reduce the layout induced performance and yield issues. The bit-cell layout was constructed in a rectangular fashion with its length longer than its width. This allowed to reduce the effective cell width and reduce the bit-line capacitance at the expense of longer word lines, while simultaneously providing minimal overall area. The cross coupled inverter pair was laid out symmetrically in a thin form. This

layout strategy allows to minimize bends in the layout to avoid mask misalignment. Dummy transistors were added to the edge of the array to avoid boundary related anisotropic photolithographic issues due to discontinuities in layout structures, as is common in most SRAM macros. The proposed bit-cell occupies an area of 12.44 μm^2 as shown in Fig. 3.4 (b). The conventional 6T bit-cell occupies an area of 0.57 μm^2 [71]. In addition to the greater number of transistors used, a large area penalty is also due to the minimum device sizing requirements of IO devices in this PDK. In technologies older than 65nm, smaller transistors (such as High-V_{TH}) with naturally thick gate oxides may be used for a much smaller penalty. The sizing of all transistors in the proposed bit-cell is shown in Table VI.

TABLE VI
PROPOSED BIT-CELL DIMENSIONS

S. No.	Transistor	Width (nm)	Length (nm)
1	PULN, PURN	400	280
2	PULP, PURP	400	600
3	PDLN, PDRN	400	1025
4	PDLP, PDRP	600	280
5	ACL, ACR	400	280
6	R1, R2, R3	400	280

3.3.2. Architecture of the Read Port

When operating in near and sub-threshold region, the ION/IOFF current ratio is severely degraded and it becomes increasingly difficult to implement greater number of cells on a single column. As the number of cells increase, the combined pass-gate leakage becomes comparable to the read current, thereby making it difficult for the sense amplifier to correctly evaluate the read bit-line voltage level. Furthermore, the data stored in the cell also affects the read bit-line leakage, thereby making the off-state read bit-line leakage current to fluctuate highly. This is exacerbated at ultra-low voltages, where the worst-case data pattern can lead to the RBL voltage level of ‘zero’ becoming greater than the RBL voltage level of ‘one’ as shown in Fig. 3.6 in the conventional read port. It is desirable to have full swing sensing to make ‘zero’ and ‘one’ voltage levels discernable.

In the proposed bit-cell, we implement a single-ended nMOS-only based read port that greatly reduces data-dependent leakage, thereby allowing for data independent ION/IOFF ratio. This in turn greatly improves the read bit-line swing and sensing margin [72]-[74].

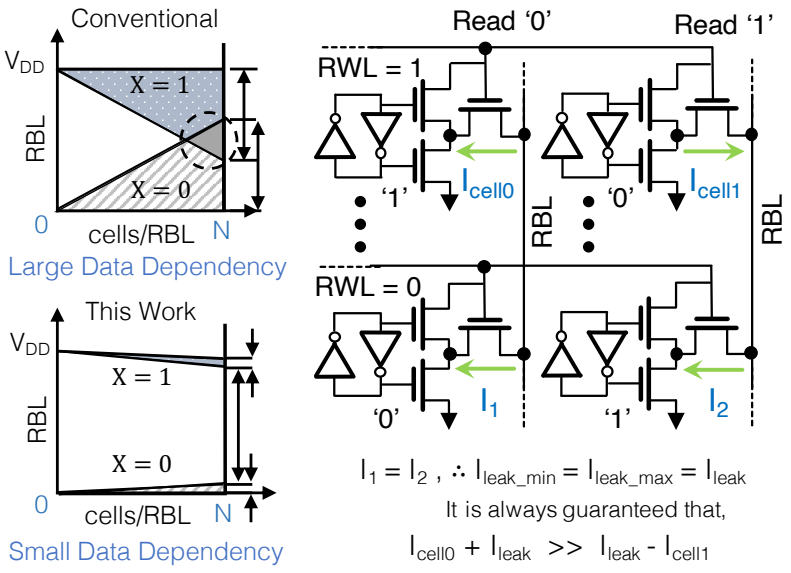


Fig. 3.6. Read Bit-Line leakage scenario in Conventional read port and the read port in the proposed DLS bit-cell.

Fig. 3.6 shows how the implemented nMOS only read port allows for data independent leakage. As seen in Fig. 3.6, the magnitude of base-line I_{leak} becomes equal in both read 'zero' and read 'one' case. This helps to maintain the required difference in magnitude between accessed-cell current in both cases. As such, we observe small data dependency and thus, a significant effective RBL swing can be observed as shown in Fig. 3.6. This is not possible in the case of conventional isolated read port sensing, because of the large dependence of leakage current on the data pattern.

The effective RBL swing, as a percentage of VDD, and with respect to varying voltage, temperature and data pattern is shown for this work and conventional read port in Fig. 3.7. The following three cases have been considered when measuring the RBL swing -

1. All cells in the column store 'zero'.
2. All cells in the column store 'one'.
3. 'One' and 'zero' are distributed equally in the column.

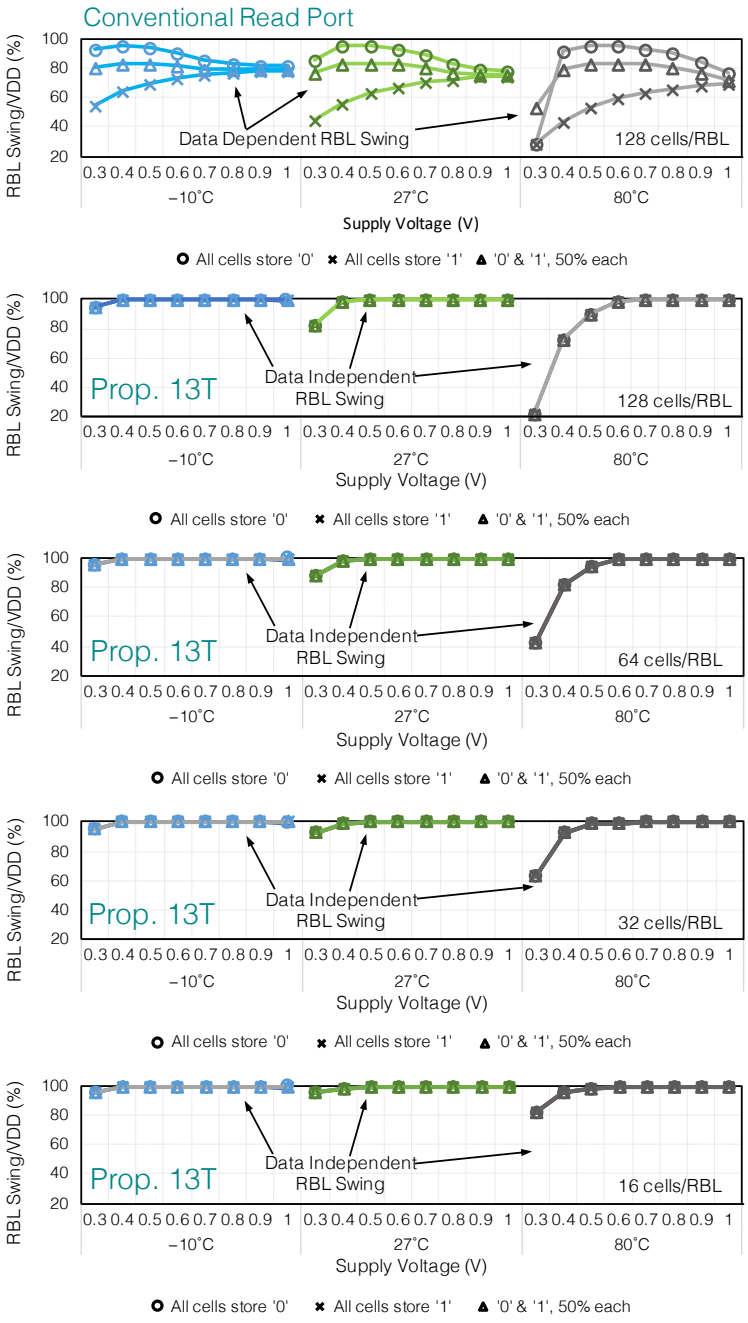


Fig. 3.7. Comparison of Effective Read Bit-Line swing in conventional SRAM read port and proposed 13T with varying supply voltage, temperature, and data pattern.

As seen in Fig. 3.7, the conventional read port's effective RBL swing is low and varies greatly according to the data pattern. The nMOS-only read port has a data-independent leakage path in it, which leads to a data-independent RBL swing, thereby improving performance. With decrease in number of bit-cells, the RBL swing improves further, especially at higher temperatures where leakage can be an issue at low supply voltages.

3.3.3. Noise Margin Analysis

Due to the hysteresis of the DLS inverters in the proposed bit-cell, they naturally exhibit much higher noise margin than the conventional 6T bit-cell. This allows for the supply voltage to be scaled for energy savings. The minimum value to which the cell's supply voltage can be lowered and still have it retain the data without failure is called the Data Retention Voltage (DRV). The DRV distribution for the proposed cell and the conventional 6T cell is shown in Fig. 3.8 (a). As seen in Fig. 3.8 (a), the DRV distribution for the proposed cell is much tighter with about 140mV savings in the 3 sigma DRV point. This increased resilience to process variations is due to its increased noise margin as shown in the Butterfly curve in Fig. 3.8 (b). The noise margin is the side length of the largest square that can be embedded in the butterfly curve. If the static noise margin falls below the thermal voltage (kT = 26mV at 300K), the bit cell data may be corrupted due to thermal noise. The proposed cell with DLS inverters has a very wide lobe in the curve indicating high noise margin. Note that when the high-level node is flipped to the other inverter, the widened lobe is also moved to the flipped operating-point-side, retaining its high noise margin. This is in contrast to the 6T cell whose noise margin is always determined by the smaller lobe in its asymmetric butterfly curve [75]. The proposed cell exhibits a 4.4x improvement over the conventional 6T cell in terms of the largest square that can fit inside the butterfly curve. Both the DRV distributions and the butterfly curves have been plotted using 100K Monte-Carlo simulations. Since SRAMs have a large number of bit-cells, the arrays can require millions of MC simulations, which is prohibitively expensive. Therefore, to account for the rare sigma events and calculate the failure rate accordingly, we evaluate the failure probability as

$$P_{fail} = \int_{-\infty}^{-\frac{\mu_{HNM}}{\sigma_{HNM}}} \frac{1}{\sqrt{2\pi}} exp\left(\frac{-x^2}{2}\right) dx \qquad (3.1)$$

where HM is the hold noise margin. The hold failure probability with varying supply voltage for different cells has been calculated and shown in Fig. 3.8 (c). The hold VMIN is determined at the 3σ failure probability for smaller capacity memories (i.e., P_{HNM}-Fail = 10^{-5}). The proposed DLS bit-cell has a much larger area than the conventional 6T bit-cell, therefore it is only fair to compare the 6T cell with the proposed cell under iso-area conditions. Since any resizing along the vertical direction will cause large increase in bit line capacitance and power, the 6T cell has been resized laterally for iso-area analysis (cell named as 6T Iso-Area). The hold failure

probability is plotted in Fig. 3.8 (c). Since the transistors in the 6T Iso-Area cell have been up-scaled, the standard deviation in transistor variations goes down considerably, thereby greatly improving the failure probability. It achieves a hold VMIN of 291mV. The proposed cell achieves a hold VMIN of 332mV compared to 448mV for the conventional non-iso sized 6T cell.

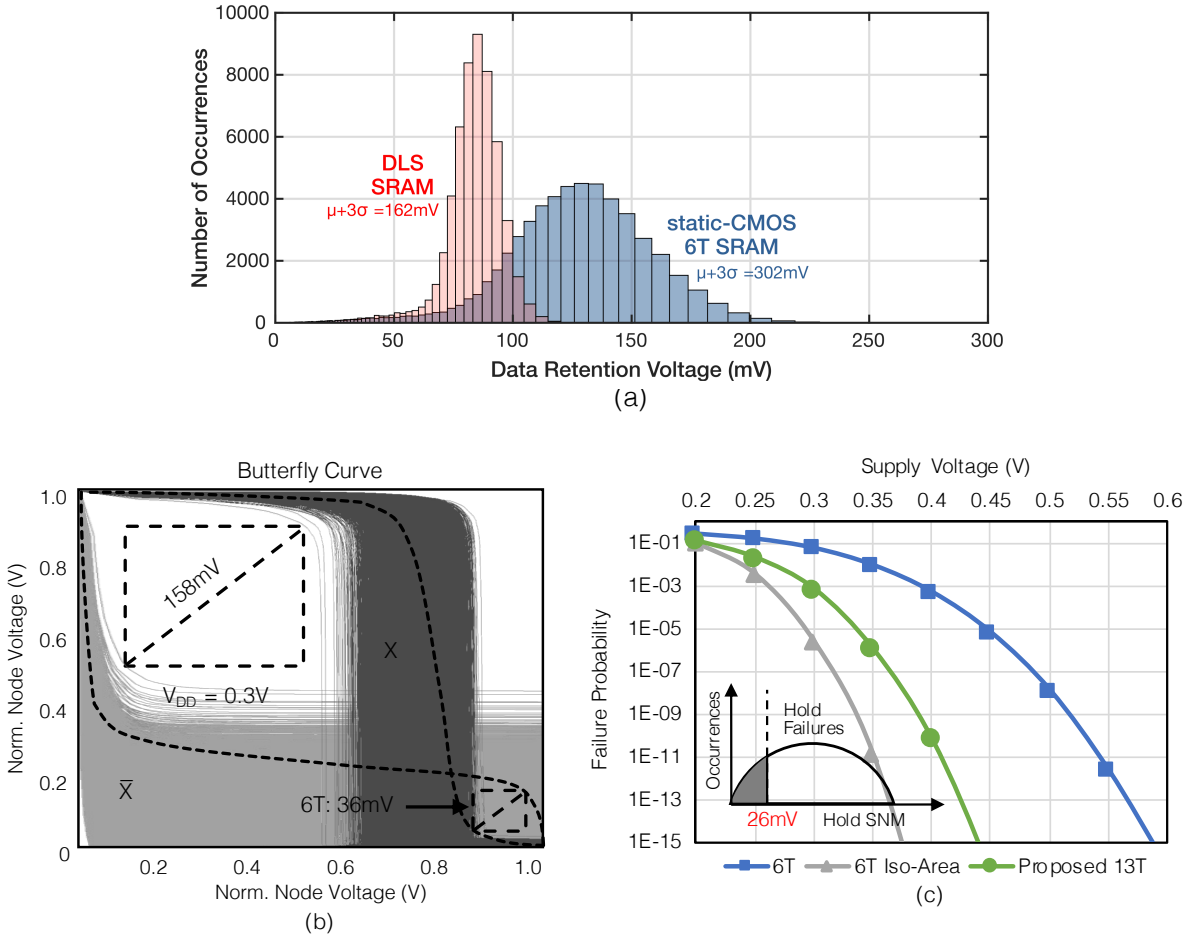


Fig. 3.8. (a) Data Retention Voltage (DRV) for 6T SRAM and Proposed DLS SRAM using a 100K Monte-Carlo simulation (b) Hold '0' Butterfly curve (V_{DD} = 0.3V) showing 4x improvement in hold noise margin (c) Hold Failure Probability with varying supply voltage.

The cell's write ability is determined by calculating its write margin, measured in Volts. A higher write margin translates to ease of writing, while a lower write margin translates to a harder write operation. A well-balanced write margin is necessary to ensure required write ability while simultaneously safeguarding the cell against unwanted writes [76], [77]. The write ability quickly degrades at low supply voltages due to degradation of required transistor strength ratio to maintain sufficient write margin.

The write margin has been determined in sub-threshold region of operation (V_{DD} = 0.3V) using a 1M Monte Carlo (MC) simulation and plotted in Fig. 3.9 (a). As shown in the figure, the write margin is close to half of

V_{DD} indicating close to ideal trade-off between write ability and robustness. On the other hand, the conventional 6T cell has a negative write margin at V_{DD} = 0.3V and only exhibits a positive three sigma write margin around 600mV [78]. The write failure probability is evaluated as

$$P_{fail} = \int_{-\infty}^{-\frac{\mu_{WM}}{\sigma_{WM}}} \frac{1}{\sqrt{2\pi}} exp\left(\frac{-x^2}{2}\right) dx \qquad (3.2)$$

The write VMIN is determined at the 3σ failure probability (i.e., $P_{WM\text{-}Fail} = 10^{-5}$). The failure probability is plotted in Fig. 3.9 (b). The proposed bit-cell achieves a write VMIN of 278mV.

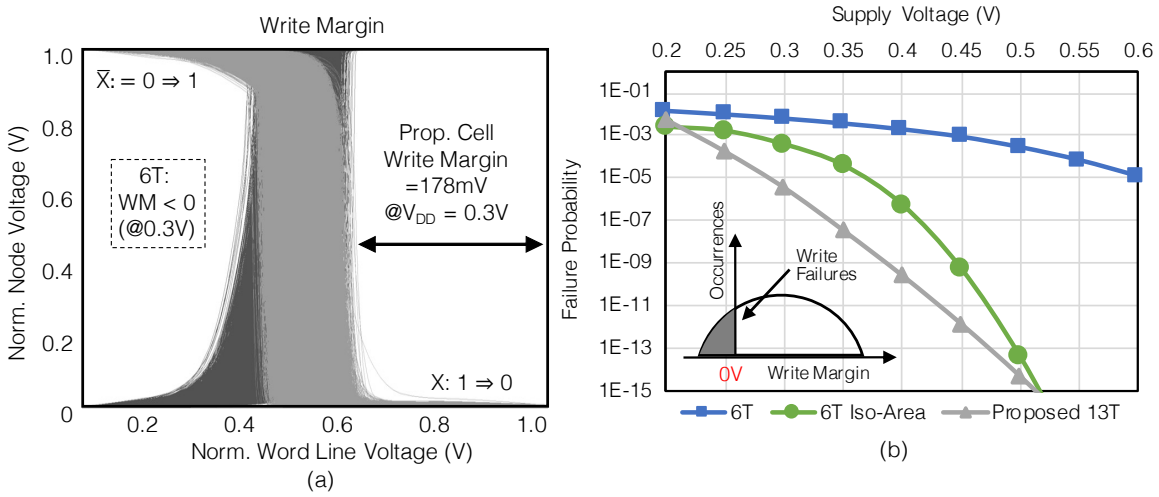


Fig. 3.9. Write margin (write ability) for the proposed cell (1M Monte-Carlo Runs). (b) Write Failure Probability with varying supply voltage.

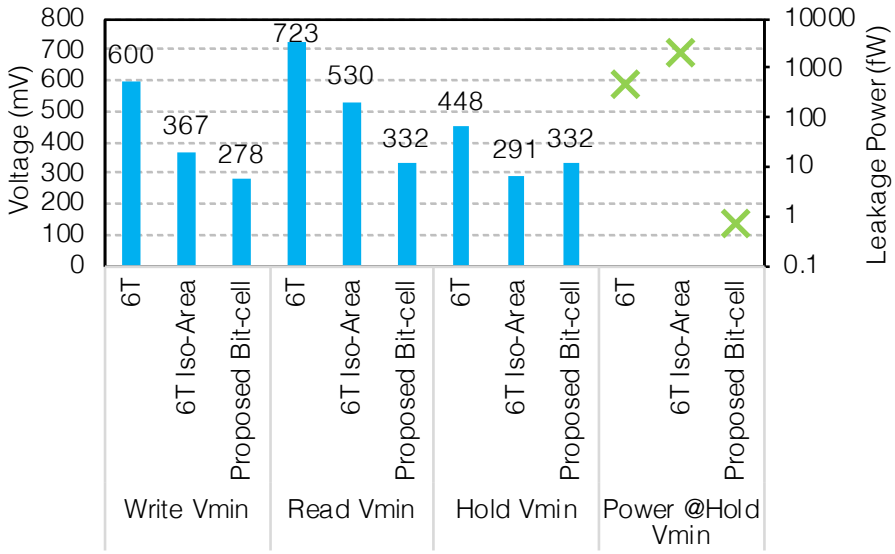


Fig. 3.10. VMIN and leakage power comparison between 6T, 6T Iso-Area, and proposed DLS bit-cell.

The read, write, and hold VMIN of the 6T, 6T Iso-Area, and the proposed bit-cell along with their leakage power at their respective hold VMIN are summarized in Fig. 3.10. The leakage power of proposed DLS bit-cell is 2941X lower than that of 6T iso-area cell.

3.3.4. Word-Line Overdrive Technique

Since the proposed cell uses DLS inverters in place of regular CMOS inverters, it is susceptible to unwanted writes i.e., data corruption due to leakage currents from the bit-line, thereby warranting the need of a leakage suppression technique. One way to accomplish this is by boosting its gate voltage (word line voltage) to super cut-off the access pMOS transistor. Here, we determine the minimum amount of boosting required such that any changes on the bit-line do not cause any noise margin violations, causing disturbances in unselected cells. To determine this, we measured the write margin for the bit-cell as a function of word line voltage using a 100k MC simulation and plotted it in Fig 3.11. As shown in the figure, when the word line voltage is equivalent to logic high (V_{DD} = 0.3V), there is non-zero write margin, thereby making all cells susceptible to unwanted write disturbances. As the word line voltage is increased further, the write margin falls off quickly. With a word line voltage of >=520mV, all cells have zero write margin, indicating their robustness. This implies that when the word line is turned off (voltage level high), then its minimum voltage should be 520mV i.e., a minimum word line boost of 220mV.

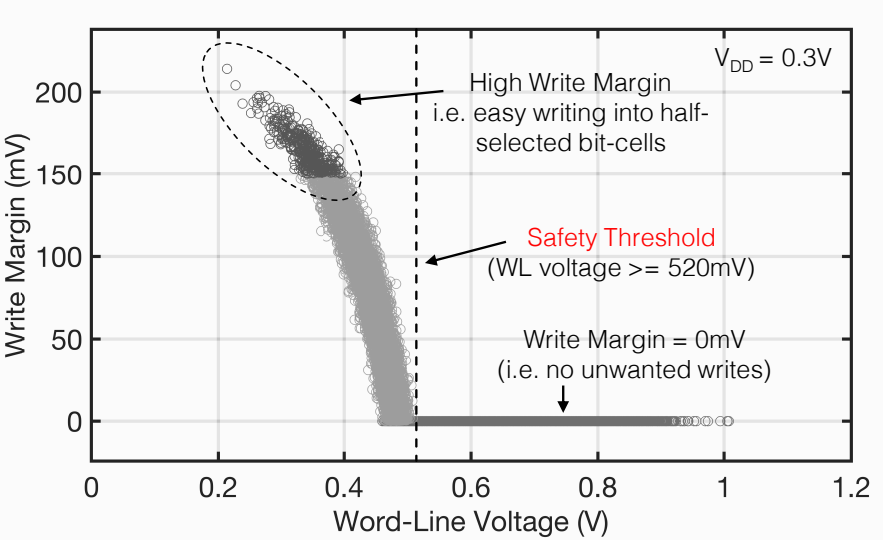


Fig. 3.11. Write margin (bit-cell write ability) vs. varying word-line voltage shows the minimum word-line voltage required to prevent inadvertent writes in column write half-select mode (100K Monte-Carlo).

The ultra-low leakage of the bit-cell causes it to be susceptible to data corruption via BL disturbances and

leakage currents. This BL leakage current from one bit-cell can flow into another bit-cell via the BL. So, the word-line is boosted to a higher voltage VDDH (V_{DD} + 0.3V) to super cut-off the pMOS access transistors as shown in Fig. 3.12. The word-line is boosted for each row using a level converter. The word-line level converter is constructed using DLS logic style (Fig. 3.12) and uses low V_{TH} transistors in the header and footer of the inverters and high V_{TH} transistors in middle. The DLS based implementation of the level converter reduces the leakage power by 2700x (down to a minimum power of just 100fW) compared to conventional differential cascode level converters to maintain a pW-level power budget [79]. The leakage power has been plotted with varying supply voltage in Fig. 3.13 (b). The transient waveform for the level shifter has been plotted in Fig. 3.13 (b). Since the level shifter solely relies on leakage currents to function without feedback implementation, it has slightly degraded high and low logic levels. The transient waveform shows that the level shifter output is fairly stable even under process variations. This is because of the high noise margins of the DLS-based structure. The boosted WL voltage output of the level converter is then fed to a strong buffer capable of driving the large word line load. This technique allows to reduce subthreshold leakage through the pMOS access transistors by up to 1000x compared to conventional nMOS access transistors, thereby maintaining zero write margin and preventing unwanted writes into the bit-cell. The higher supply voltage need not be regulated and clean and can therefore be generated using a Switched Capacitor Voltage Regulator (SCVR) with minimal power overhead, or another readily available on-chip supply may be used.

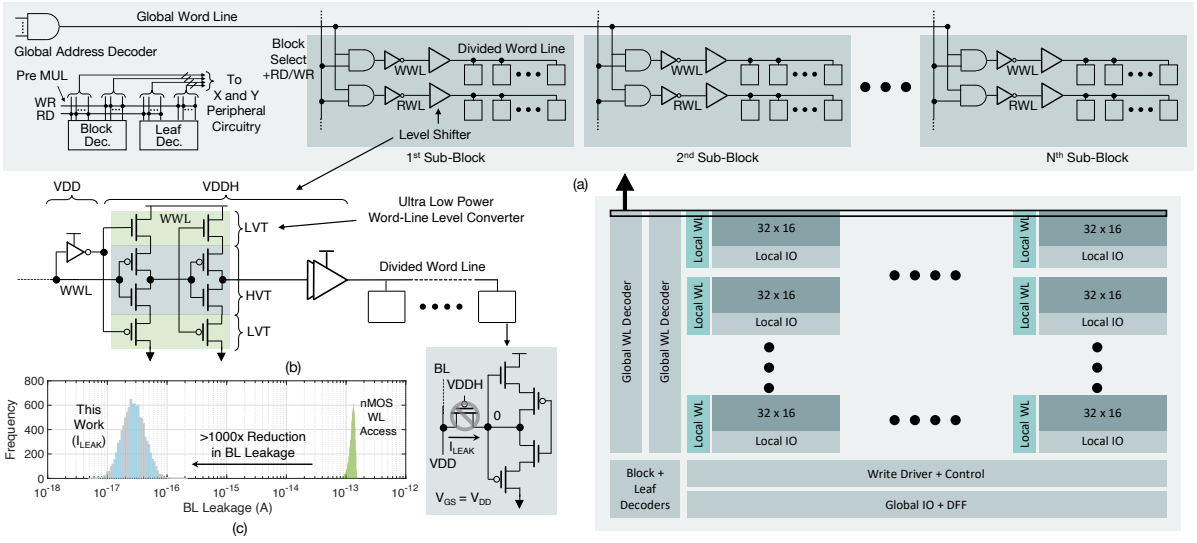


Fig. 3.12. (a) Schematic of a Row Slice and the (b) bit-line leakage suppression technique in the proposed memory architecture. (c) Bit-line leakage comparison in conventional nMOS-based and over-driven pMOS-based access transistors.

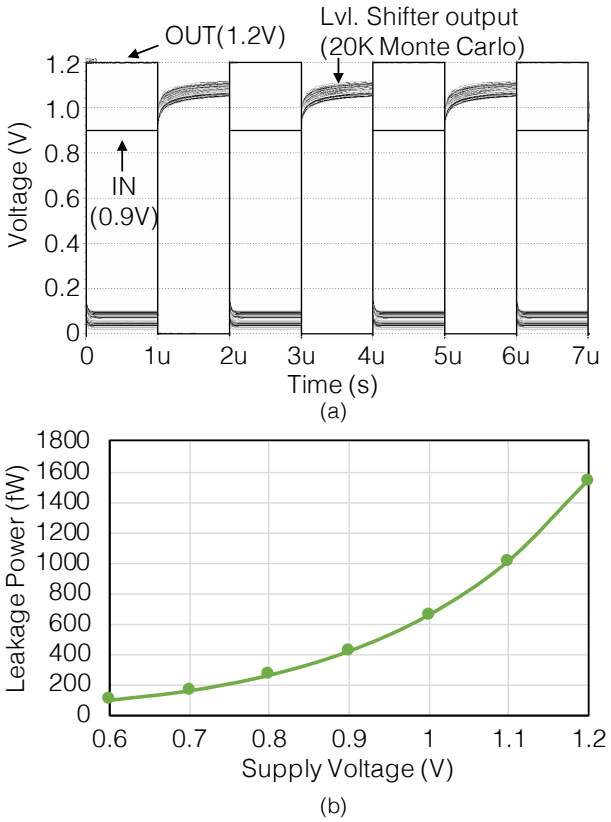


Fig. 3.13. (a) Transient waveform (20K Monte Carlo) for the proposed ultra-low power level shifter. (b) Leakage power of the level shifter with varying supply voltage.

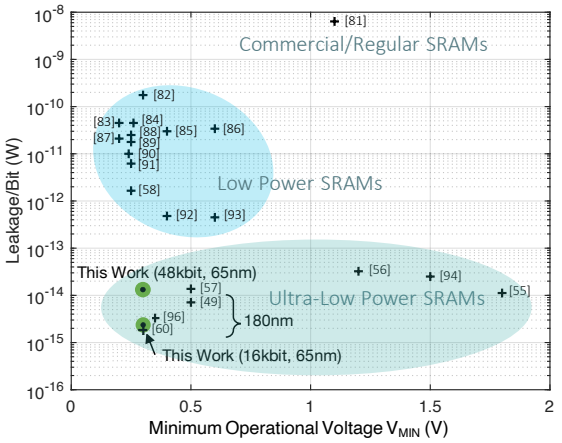
3.4. Test Chip Implementation and Results

3.4.1. Architecture of the SRAM Macro

The block diagram of the memory is shown in Fig 3.12. Since the cell is prone to pseudo-read-disturb issue, the array was constructed in a non-interleaved architecture without column-select circuitry. The 16kb array comprises thirty-two 0.5kb sub-blocks each with 16 cells per column and a 32-bit word size. The 48kb memory comprises of three 16kb banks and each bank is accessed using address decoders. Selective precharge using BS (Block-Select) was used to enable the precharge of bit-lines of only the accessed block to reduce active power consumption. Four metal layers were used to route the supply voltage (V_{DD}), V_{SS}, and the bit-lines vertically and the local and global word-lines horizontally as shown in Fig. 3.4 (b). Since the array capacity is relatively small (16kb-48kbit), the Divided-Word-line-Decoding (DWD) scheme [80] was implemented. The global word line (GWL) was routed across all banks and combined with the local block select to generate local word lines. The local and global word-line signals are negative-edge synchronized with the clock signal. The bit-lines are

precharged at the beginning of each write operation, after which the data is loaded onto them before the enabling of WWL. Similarly, all the local bit-lines of the accessed block are precharged during the first half of the read clock cycle. The RWL is enabled during the second half of the clock to allow the RBL to develop conditionally. The voltage level on the RBL is detected by the local sense amplifier (skewed inverter buffers), which is then used to evaluate the sub-global RBL. Tri-state buffers are used to mux sub-global RBL to global RBL. The output flip-flops capture and produce logical output according to global RBL level.

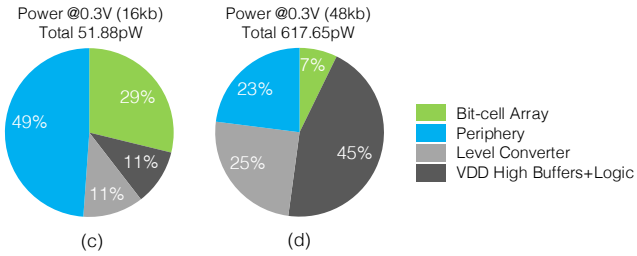
3.4.2. Measurement Results



Metric	JSSC'09 [95]	JSSC'13 [96]	ISCAS'11 [60]	This Work	
Technology	180nm	180nm	180nm	65nm	65nm
Bit-Cell	14T	10T	10T	13T	13T
Bit-Cell Area (μm²)	40	17.48	17.48	12.44	12.44
Access Frequency	100KHz @0.5V	3.5KHz @0.35V	1KHz @0.3V	6.9KHz @0.4V	0.03KHz @0.4V
Write Frequency	Async	Multi-Cycle	Multi-Cycle	3.6KHz @0.4V	0.03KHz @0.4V
V_{MIN} (V)	0.5V	0.35V	0.3V	0.3V	0.3V
Best DRV (mV)	-	-	-	115	126
Leak./bit@V_{MIN} (fW)	7.1	3.3	1.85	2.74*	12.56
SRAM Leakage (pW)	22	80.53	-	44.88	617.65
Capacity	1kb	24kb	24kb	16kb	48kb

*The bit-cell leakage represents 0.61fW out of the total 2.74fW/bit @0.3V. And 0.525fW @0.2V.

(b)



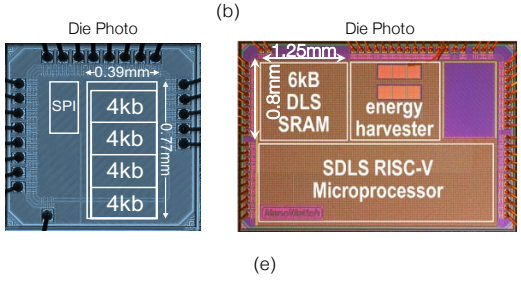


Fig. 3.14. Comparison of proposed memory with other state-of-the-art works with respect to (a) leakage/bit (b) memory performance metrics. Power breakdown for the (c) 16kb and (d) 48kb SRAM at 0.3V. (e) Die photo of the 16kb and the 48kbit DLS SRAM.

To measure the performance of the memory with the proposed bit-cell and bit-line leakage suppression technique, we implemented a standalone 16kb array in a 65nm bulk planar low power CMOS technology. The die photo of the 16kb SRAM is shown in Fig. 3.14 (e). Another 48kb macro is integrated in an SoC to showcase a real use case. Its die photo is shown in Fig. 3.14 (e). The 48kb macro uses an integrated memory controller, power supply regulators and other digital IO and control logic to function and communicate with the on-chip processor.

The prototype chips were wire bonded with a 64-pin Pin Gated Array (PGA) ceramic package with gold bond

wires and ball bond type. It was mounted on a printed circuit board using a 10x10 PGA Gold Through Hole Socket. Sub-Miniature version A (SMA) connectors were mounted on the PCB for power supplies. The PCB was placed in a Tenny JR temperature chamber which controlled the environmental temperature of the chip in the measurement. An IO-3200 Pattern Generator Logic Analyzer was used to interface digitally to the memory. All power consumption measurement in this work were taken with a Keithley 6430 sub-fA Source Meter with high-isolation shielded co-axial cables.

The performance metrics and leakage power of 10 chips are measured across voltage and temperature to validate design robustness. These results are compiled and compared against the state-of-the-art as shown in Fig. 3.14.

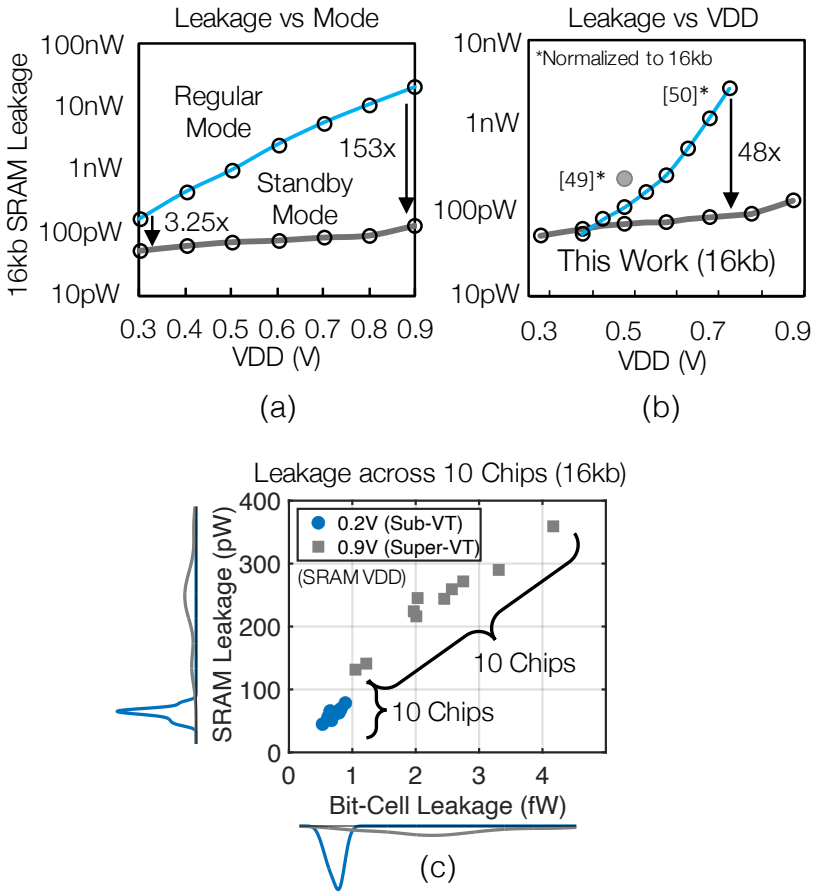


Fig. 3.15 (a) Leakage of the proposed 16kb SRAM as a function of V_{DD} in memory access mode (regular mode) and stand-by mode. (b) 16kb SRAM leakage as a function of V_{DD} compared against other state-of-art. (c) Leakage measured for SRAM and bit-cell across ten 16kb chips in both sub-threshold (at DRV point) and super-threshold regions of operation.

As seen in the Fig. 3.14 (a), the proposed memory lies in the ultra-low power regime and has the lowest leakage/bit recorded at 525aW (at 0.2V). When the memory is in access mode (read or write), it is operating in

regular mode and consumes leakage with respect to varying supply voltage as shown in Fig. 3.15 (a). It takes a single clock cycle to put the memory into access mode, after which the memory can be written to or read from. As soon as the access clock cycle is complete, the memory automatically goes into stand-by mode in the next clock cycle and power gates all peripheral circuitry besides the bit-cell array and word-line converters. This allows up to 153x reduction in the leakage power consumption, resulting in an ultra-low leakage of 44.88pW at 0.2V and up to 131.5pW at 0.9V. Across 10 chips, the leakage power of the entire 16kb SRAM is up to 360pW at the highest operating voltage as shown in Fig. 3.15 (c). The leakage in DLS logic is not a strong function of V_{DD}. This is because, with increase in supply voltage, the negative V_{GS} in the super cut-off transistors also increases. As such, the proposed memory exhibits low leakage even at higher operational voltages. Whereas, the leakage in other state-of-the-art works rises exponentially as they go from subthreshold to super-threshold region of operation. This translates to up to 48 times reduction in leakage power as shown in Fig. 3.15 (b). Extrapolation of leakage power trends predicts >400X reduction in leakage power compared to previous state-of-art. The SRAM is able to work up to several MHz in super-threshold region as shown in Fig. 3.16 (a). This combination of pico-watt to nano-watt power and low MHz performance makes it a very ideal solution for most low-power IoT applications. The robustness of the SRAM is demonstrated by measuring its Data Retention Voltage across 10 chips as shown in Fig. 3.17 (a). Measurement results from 10 chips show a best-case chip data retention voltage (DRV) of 115mV, 10-chip average DRV of 156.5mV and full functionality across temperature from 0°C to 60°C. The energy-per-access is measured to be 1.33fJ at the lowest operating voltage (for access), with an average of 2.36fJ across 10 chips as shown in Fig. 3.17 (b).

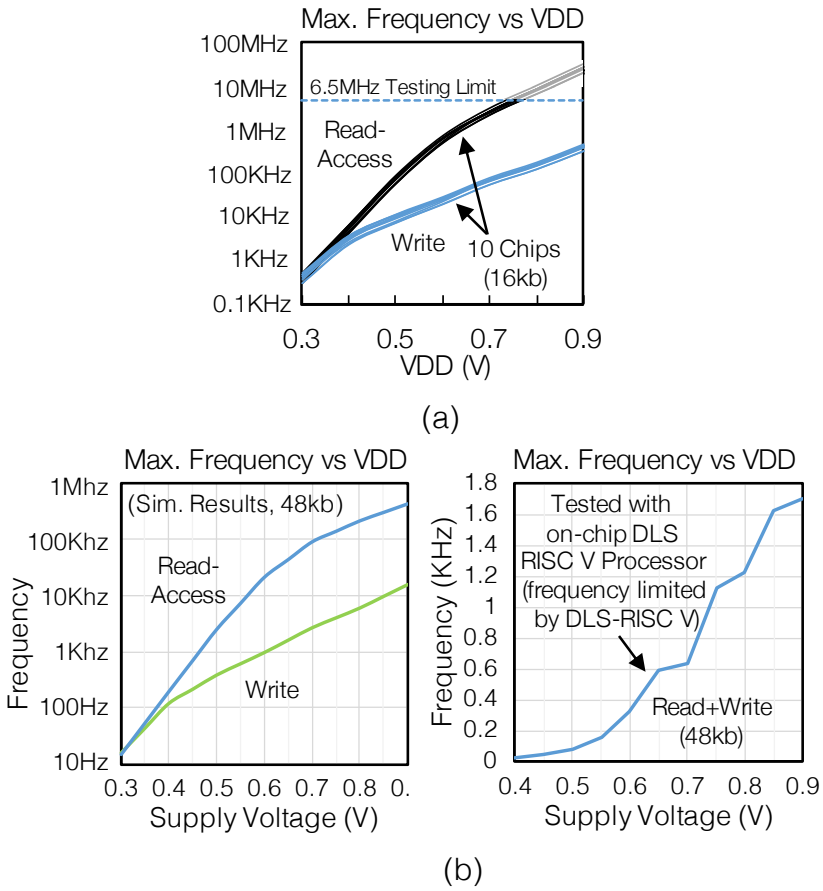
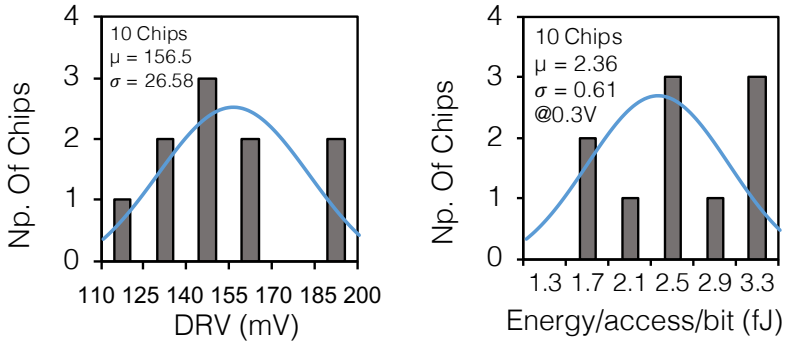


Fig. 3.16 Maximum operating frequency for the (a) standalone 16kb SRAM (b) 48kb SRAM integrated in an SoC.



(a) (b)

Fig. 3.17. Measurement results for (a) Data Retention Voltage (DRV) across ten 16kb chips (b) Energy/access/bit with varying supply voltage (c) Energy/access/bit at a supply voltage of 0.3V across ten 16kb chips.

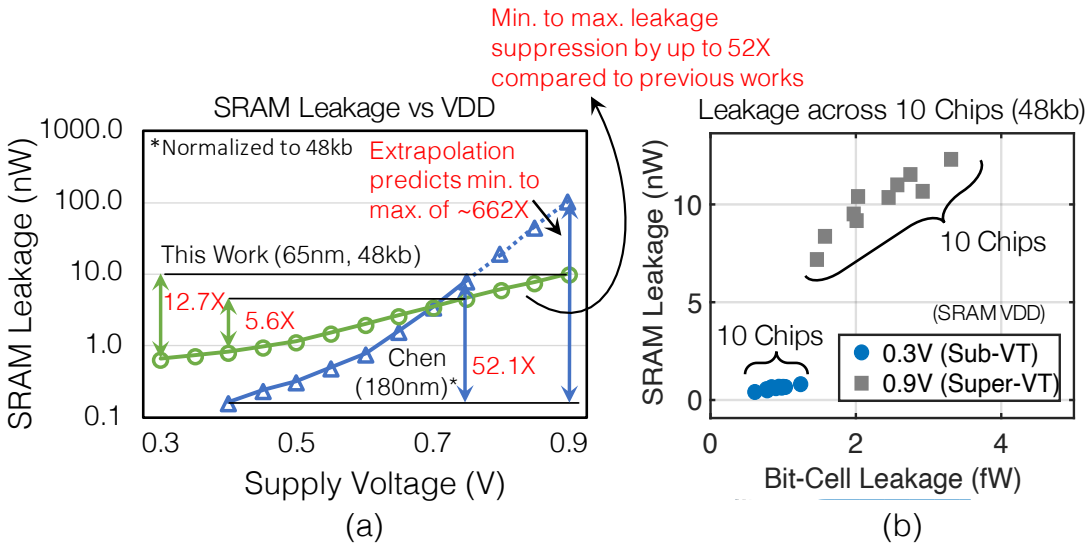


Fig. 3.18 (a) 48kb SRAM leakage as a function of V_{DD} and compared against other state-of-art (b) Leakage measured across ten 48kb chips in both sub-threshold and super-threshold regions of operation.

Measurement results for the 48kb SRAM, as seen in the Fig. 3.14 (a), show that the proposed memory lies in the ultra-low power regime and has a leakage of 12.56fW/bit at 0.3V. The proposed 48kb SRAM has an ultra-low 10 chip average leakage of 617.65pW (12.56fW/bit) at 0.3V and 10.1nW (205.6fW/bit) at 0.9V as shown in Fig. 3.18 (a). This translates to up to 52 times suppression in min. to max. leakage. The proposed SRAM has a measured max. frequency of 1.7KHz. The testing frequency is limited by the on-chip DLS-based RISC V Processor. Simulations show that the memory can run at up to 416KHz at 0.9V as shown in Fig. 3.18 (b).

3.5. Scalable DLS Memory

Modern IoT sensor nodes can monitor various physiological and environmental signals, such as temperature, humidity, and motion. Since these signals are low bandwidth and require Hz-KHz rate processing, they can be powered directly from harvested energy, making them compact and low-maintenance. However, the available energy can fluctuate unpredictably, causing the harvested power to drop down to the nW level.

In such scenarios, DLS (Dynamic Leakage Suppression) logic offers several benefits in low-power circuit design. DLS logic reduces leakage current during standby mode, resulting in lower power consumption compared to conventional static CMOS logic. This enables longer battery life and lower operating costs. DLS logic is designed to operate at ultra-low power levels, in the sub-nW range, which makes it well-suited for battery-less and energy-harvesting applications where power is limited. DLS logic can also be easily integrated into existing CMOS design flows, making it a cost-effective and efficient solution for low-power circuit design.

While DLS logic offers several advantages for ultra-low power applications, there are also some potential drawbacks to consider. DLS logic has limited performance and can only operate at relatively low frequencies in the Hz-range. This means that it may not be suitable for applications that require high-speed processing.

This can be true for DLS based SRAMs as well. The implementation of the DLS SRAM in previous section had a lower power in fW/bit range. But this came at the cost of reduced performance (few Hz-KHz). On the other hand, conventional 6T memories can run at giga-hertz frequency but they consume a lot of power (up to hundreds of pW/bit). Low power variants of conventional CMOS SRAMs can have high performance but still consume several pW/bit in leakage. Therefore, there is a need to bridge the gap between traditional low power memories (pW /bit) and ultra-low power (fW/bit) memories. In this work, we present a Scalable DLS (SDLS) SRAM that can scale its performance to KHz to MHz range when required but still have sub-pW/bit leakage power.

3.5.1. SDLS SRAM Architecture

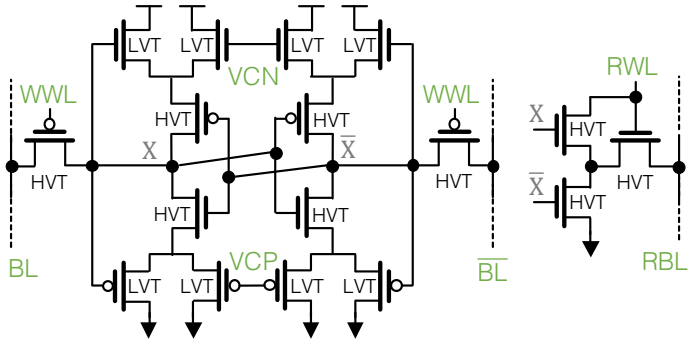


Fig. 3.19. Circuit Schematic of the proposed SDLS 17T SRAM bit-cell.

The circuit schematic of the proposed SDLS-based bit-cell is shown in Fig. 3.19. Two additional bypass FETs are added in the header and footer of each of the DLS inverters. Toggling of these FETs switches the inverters to CMOS mode, thereby temporarily increasing performance at the cost of power. The header nMOS transistor in the DLS inverter is shorted using another nMOS using the VCN signal, while the pMOS in the footer is shorted to VSS using another pMOS transistor using the VCP signal. The increased performance allows to speed up the feedback mechanism of the cross-coupled inverter pair. This allows to dramatically improve the write access speeds which was a performance limitation of the regular DLS-based SRAM. When the write operation is completed, the bypass transistors are turned back off, thereby returning to the low power mode. During the read operation, the bypass transistors are not turned on because the read port has been decoupled and is accessed separately. The functioning of the read port in the SDLS-based bit-cell is the same as the previous DLS-based bit-cell.

By adding four more transistors, the area of the bit-cell would increase. However, this is offset by implementing HVT and LVT transistors instead of IO devices. This increases the leakage power when compared to the IO-transistor based bit-cell shown in the previous DLS-based SRAM version. The layout of the proposed SDLS bit-cell is shown in Fig. 3.20. In the SDLS-based bit-cell, the local and global word-lines are routed horizontally and the write and read bit-lines are routed vertically. The power lines are routed on M4 to reduce wire resistance as much as possible. The layout is created in a rectangular manner with the word-line being longer than the bit-lines. This is done to reduce the bit-line capacitance to improve access performance. As seen

in the figure, the area of the proposed SDLS-based bit-cell is 4.3368μm^2, which is approximately 3.8X times that of the conventional 6T bit-cell. It also achieves an area reduction of 65% when compared to the regular DLS logic-based bit-cell implemented using IO devices.

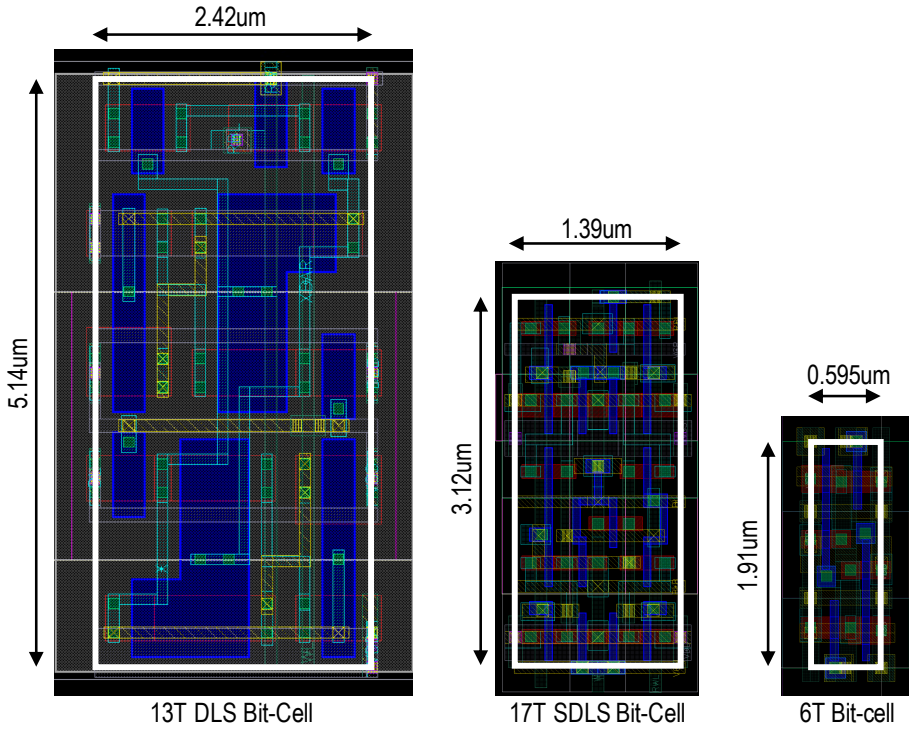


Fig. 3.20. Layout comparison of the DLS based 13T bit-cell, SDLS based 17T bit-cell, and the conventional 6T bit-cell.

The block diagram of the SDLS memory's architecture floor plan is shown in Fig. 3.21. As seen in the figure, the 16kbit memory is broken down in four sub-banks of 4kbit each. Each sub-bank is 32 bit-cell wide and 128 bit-cell long. The sub-banks are accessed using the divided word-line scheme because of its small capacity. As shown in the previous section, the bit-line leakage is reduced by implementing a word-line boosting technique. This technique uses a level shifter to boost the word-line voltage, thereby boosting the gate voltage of the pMOS access transistor. This allows to super cut-off the access transistors by creating a negative V_{GS} on its terminals and consequently reduce bit-line leakage considerably.

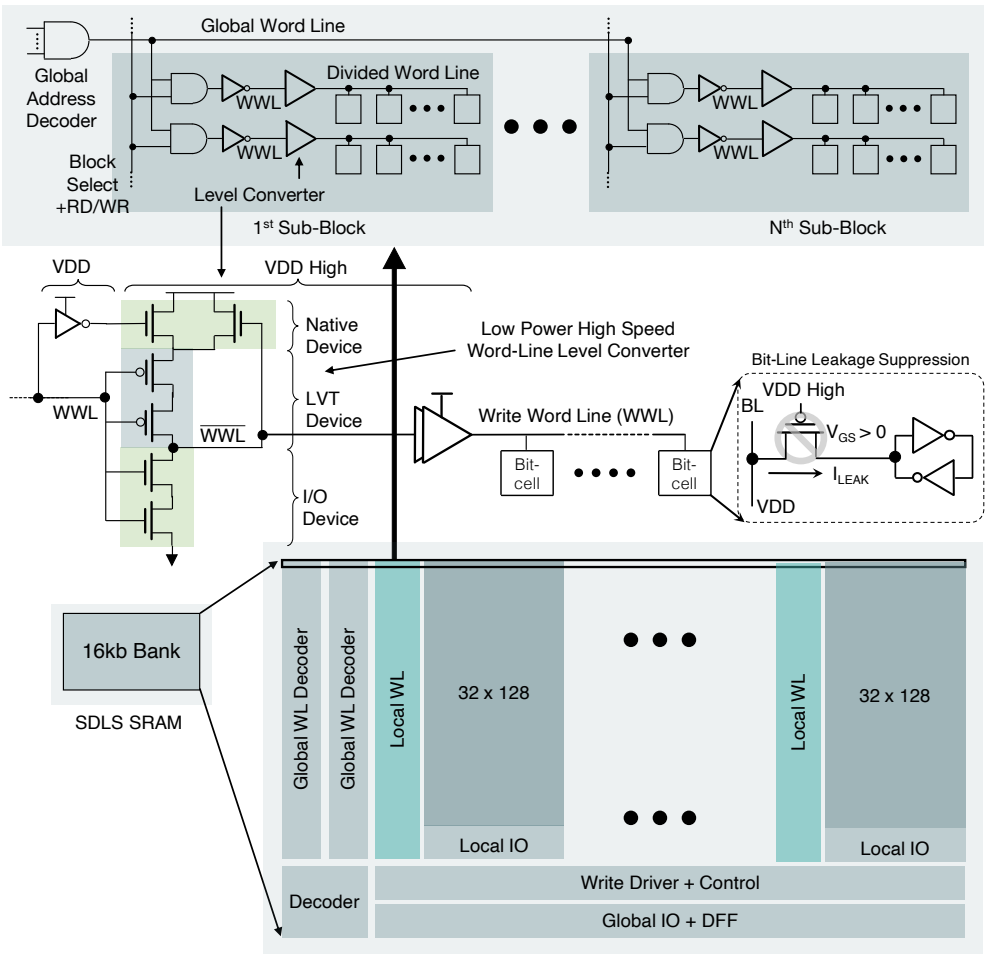


Fig. 3.21. Block diagram of the architecture of the SDLS SRAM with the schematic of the row-slice shown.

The level shifter that was implemented in the DLS logic-based SRAM in the previous section consumed a very low leakage of 100fW. That level of leakage power was necessary to maintain the pW budget of the SRAM. This was achieved at the cost of low performance (up to several hundred kilo-hertz). For the SDLS SRAM, mega-hertz range performance is needed. As such, an improved version of the level shifter is required. Additionally, the previous version could only up-convert up to 400mV. In this version of the SRAM, the nominal supply voltage is 1.2V and the peripheral VDD is 0.6V. In other words, the amount of required up-conversion is 600mV. As such, the level shifter should be able to perform this across process-voltage-temperature variations. The circuit schematic of the proposed level shifter is shown in Fig. 3.22. It requires complementary input signals. The header consists of Native nMOS devices. The output of the level shifter is fed back to the gate of the header nMOS transistor. The pull-up and pull-down network in the level shifter are made using low-threshold (LVT) and high-threshold (HVT) transistors respectively.

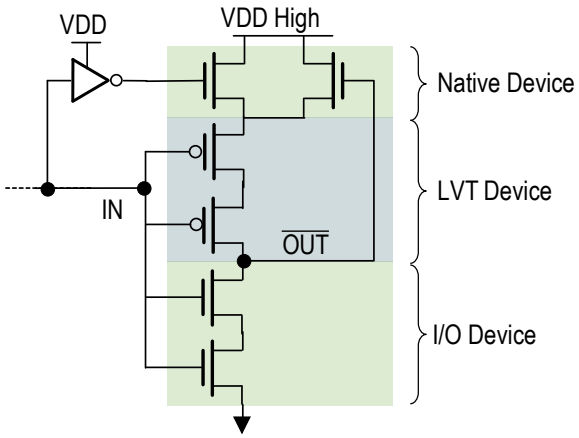


Fig. 3.22. Circuit Schematic of the low power Level Shifter.

As the input voltage decreases, the pull-down nMOS get weaker. This increases the output voltage level slowly. Initially both native devices are in off-state. But then both the native nMOS devices start to turn on and very quickly become strong due to them being native devices (i.e. very low threshold voltage). At the same time, the LVT pMOS also start to conduct, thereby increasing the output voltage. As the output voltage level increases, the header nMOS native device starts to conduct even more strongly, thereby allowing a positive feedback mechanism. This allows to quickly develop a high logic voltage level. The transient performance for the proposed level shifter is shown in Fig. 3.23 using a 20K Monte-Carlo simulation. The input frequency is 10MHz and the high logic voltage level is 0.6V. As seen in the figure, the output voltage level is successfully able to reach 1.2V with low variability. All 20K samples pass at the input frequency of 10MHz.

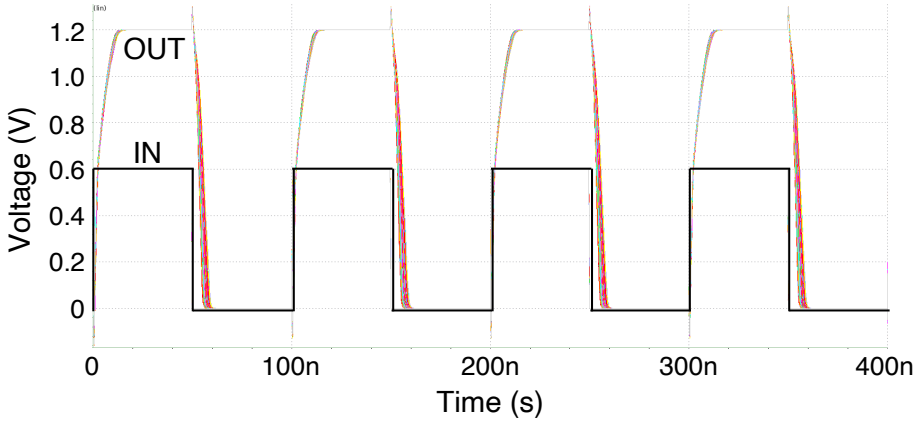


Fig. 3.23. Transient waveform for the proposed level shifter at 10MHz (20K Monte-Carlo simulation).

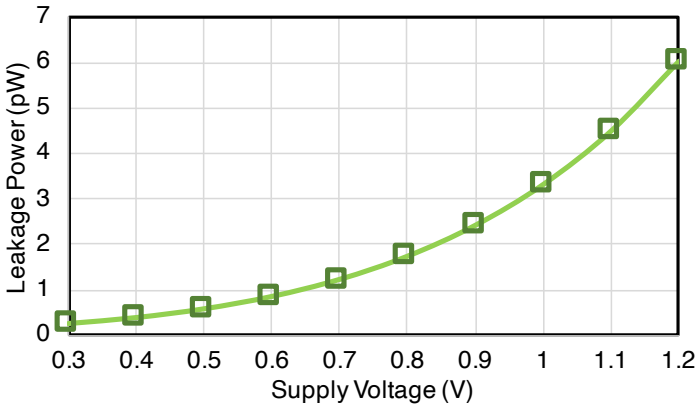


Fig. 3.24. Leakage power of the proposed level shifter with varying supply voltage.

The leakage power of the level shifter is plotted in Fig. 3.24 with varying supply voltage. As seen in the figure, the leakage varies from 262.5fW at 0.3V to a maximum leakage of 6pWat 1.2V. This indicates that the proposed level shifter will be able to maintain the memory's nW power budget and is therefore, suitable for the proposed SDLS Logic-based SRAM for both performance and power requirements.

3.5.2. Results

In this section, we discuss the performance metrics of the SDLS SRAM such as leakage power and operating voltage and frequency range. We also discuss how various transistor device types that can be used to construct the bit-cell and how that affects the leakage power. Fig. 3.25 shows the leakage power of the DLS SRAM, SDLS SRAM, and the conventional 6T SRAM with the maximum operating frequency across their respective operating voltage range. As seen in the figure, the DLS SRAM is suitable for applications where the performance requirements are in the range of a few hertz to a few kilo-hertz. A few examples can be where sensing and computation is required a few times per day such as soil pH sensing or intermittent temperature sensing. On the other end of the plot is the conventional 6T SRAM with operating frequencies up to a giga-hertz and several hundred nWs of leakage power. This is more suitable for applications with computational requirements and high capacities. In the middle of the plot, the SDLS SRAM bridges the performance and leakage design space between the DLS SRAM and the conventional 6T SRAM. It is more suitable for mid-range performance application requirements with few kilo-hertz to few mega-hertz frequency operating range.

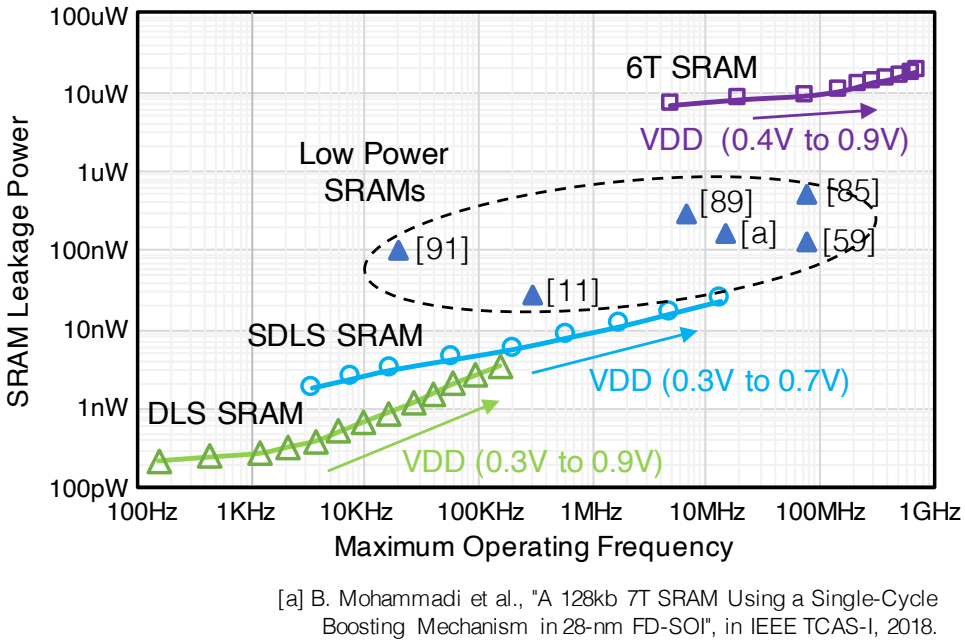


Fig. 3.25. Leakage Power and performance comparison of the DLS based 13T SRAM, SDLS based 17T SRAM, and the conventional 6T SRAM.

The leakage power per bit with varying supply voltage is shown for various kinds of SRAM in Fig. 3.26. The lowest leakage per bit is consumed by the IO device-based bit-cell. This is because of its thick gate oxide that reduces gate leakage and very high threshold voltage that reduces subthreshold leakage. The LVT-HVT based bit-cell uses LVT transistors for the header and footer and HVT transistors for the inner transistors in the DLS inverter. This allows for much lower leakage than the conventional 6T cell, especially at lower operating voltages where the 6T is unable to work due to insufficient noise margins. The LVT-LVT version of the bit-cell has a much higher leakage than all other DLS based variants. At lower supply voltages, the subthreshold leakage increases because of low voltage at intermediate nodes in the header and footer of the DLS inverter. At higher supply voltages, the intermediate node voltage rises and minimizes the flow of subthreshold leakage current. On the other hand, other leakage currents such as gate current decrease continuously with decreasing supply voltage. This creates a point for the LVT-LVT variant where the leakage becomes minimum. The HVT-HVT variant is susceptible to not being able to maintain the desired drive current ratios with respect to leakage currents. The parasitic leakage currents (such as gate leakage current) can overpower the intrinsic drive strength of the active

pull-up or pull-down network in the DLS inverter. As such, it is only able to work at lower voltages and consume higher leakage than the LVT-HVT and IO variants.

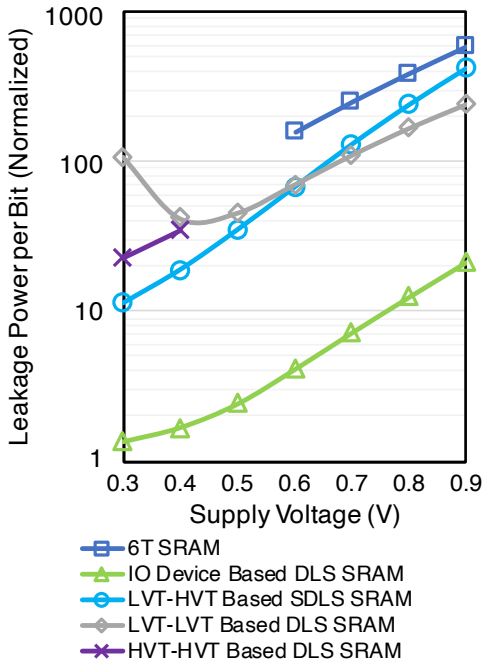


Fig. 3.26. Comparison of Leakage power/bit with varying supply voltage for various kinds of device types.

The SRAM area is an important metric because of the cost of the silicon, area and form factor requirements of the application, and the capacity requirement of the application. We show the layout area comparison of 16kbit versions of DLS SRAM, SDLS SRAM, and the 6T SRAM in Fig. 3.27. The DLS SRAM has the largest bit-cell in terms of area and thus consumes the highest amount of area of 0.3003mm^2. It is approximately 5X the area of the conventional 6T SRAM (0.06138mm^2). The SDLS SRAM has an area of 0.1229mm^2, which is approximately 2X the area of the 6T SRAM. It should be noted that the 6T SRAM in this comparison is constructed using a custom DRC-compliant bit-cell. A foundry bit-cell that uses pushed DRC rules will have a smaller area footprint and will increase the area penalty comparison of the proposed DLS and SDLS SRAMs.

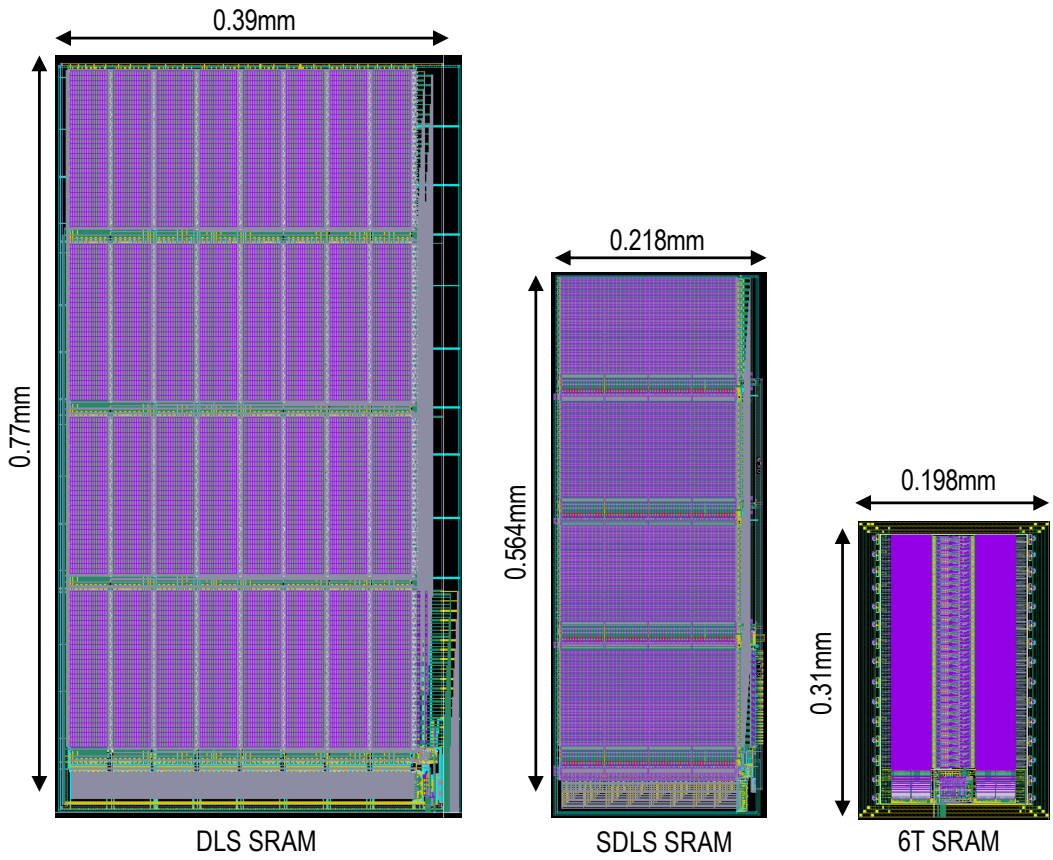


Fig. 3.27. Layout comparison between 16kbit versions of the DLS based 13T SRAM, SDLS based 17T SRAM, and the conventional 6T SRAM.

3.6. Conclusion

In this chapter, we proposed an ultra-low power leakage suppressing SRAM that reduces leakage even at higher supply voltages. The DLS-based bit-cell achieves an ultra-low leakage state of 525aW (at 0.2V) in the lowest state. An ultra-low power implementation of the bit-line leakage suppression technique using DLS based level shifters (~100fW) allows more than 1000X reduction in leakage resulting in design robustness and energy savings. The proposed memory is verified using silicon measurements in a 65nm bulk planar CMOS technology. Measurement results show up to 3.5x and 48x reduction in power in sub-threshold (0.2V) and super-threshold regions (0.9V) of operation respectively over previous state-of-art. Extrapolation of leakage trend of previous state-of-art with respect to supply voltage to 0.9V predicts more than 400X reduction in leakage, thereby making the proposed memory suitable for battery powered or self-harvesting IoT application nodes. Another 48kb ultra-low power SRAM in a bulk planar 65nm CMOS technology was integrated in an SoC to showcase a real use case. The proposed 48kb SRAM suppressed leakage down to 617.65pW at 0.3V and 10.1nW at 0.9V (up to 10X lower than previous state-of-art), resulting in an effective leakage/bit of 12.56fW and 205.6fW respectively, thereby making it suitable for energy constrained IoT applications. Extrapolation of leakage trend of previous state-of-art with respect to supply voltage to 0.9V predicts ~52x suppression in min. to max. leakage, thereby making the proposed memory suitable for battery powered or self-harvesting IoT application nodes even at higher supply voltages, where performance is more desirable. An SDLS based memory was also presented that aims to bridge the gap between ultra-low power memories and high-performance memories, making them suitable for mid-performance IoT nodes.

4. Mixed Signal and Analog Circuit Design Automation

4.1. Motivation and Prior Art

Circuit design is a complex process that requires specialized knowledge and expertise, as well as access to expensive software and hardware. The cost of designing circuits has risen significantly in recent years, making it increasingly difficult for individuals and small organizations to pursue new ideas and technologies. At the same time, as IoT is maturing, its implementation will require billions of hardware sensor nodes, and will be deployed in a variety of consumer, commercial, and industrial spaces to facilitate the collection and exchange of information for generating valuable insights and feedback. All these applications differ in the type of sensing and data collection and consequently their circuit implementation and power budgets. In addition, the proliferation of IoT devices has increased demand for low-power, low-cost, and highly integrated circuits. In this context, circuit design automation has become an essential tool for designers looking to keep pace with these trends and develop innovative new products.

Circuit design automation refers to the use of software tools and algorithms to automate certain aspects of the circuit design process. This can include tasks such as schematic capture, layout design, and routing. By automating these tasks, designers can reduce the time and resources required to develop a new circuit, as well as minimize the risk of errors and inconsistencies.

There are several key benefits to circuit design automation in the context of rising costs and IoT applications:

1. Cost Reduction: One of the main advantages of circuit design automation is that it can significantly reduce the cost of designing circuits. Automation tools can reduce the need for skilled engineers and expensive hardware, making circuit design more accessible and affordable for a wider range of individuals and organizations. This is particularly important in the context of IoT applications, where the demand for low-cost, low-power circuits is high.
2. Time-to-Market: Another important advantage of circuit design automation is that it can reduce the time-to-market for new products. By automating certain aspects of the design process, designers can develop and test new circuits more quickly, allowing them to bring new products to market faster. This is particularly

important in the fast-paced world of IoT, where competition is fierce and time-to-market can be a key differentiator.

3. Quality and Reliability: Circuit design automation can also improve the quality and reliability of circuits. By minimizing errors and inconsistencies, automation tools can help designers create circuits that are more robust and less prone to failure. This is particularly important in the context of IoT applications, where reliability is critical to the success of a product.
4. Integration: Circuit design automation can also enable greater integration between different components and systems. By automating tasks such as layout design and routing, designers can create highly integrated circuits that are optimized for performance and power consumption. This is particularly important in the context of IoT applications, where space and power are often at a premium.

Another benefit of circuit design automation in IoT is the ability to create custom and optimized circuits for specific applications. For example, IoT devices often require specific functions such as data processing, filtering, and communications. By using automated design tools, designers can quickly create custom circuits that meet these requirements, reducing the time and costs of development.

Furthermore, circuit design automation can help in optimizing power consumption in IoT devices, which is critical for battery-powered and energy-harvesting applications. By using automated power optimization tools, designers can identify and optimize power-hungry components, reduce power leakage, and design circuits that are optimized for power efficiency.

Automated testing is another important aspect of circuit design automation. It can help in reducing the time required for testing, increase reliability, and improve the quality of the circuit. In IoT applications, automated testing can ensure that the circuit can operate under varying environmental conditions and remain reliable even under extreme conditions.

Since circuit design is a complex process that requires significant expertise and resources, including skilled engineers, specialized software, and expensive hardware, the cost of designing circuits has become prohibitively high for many individuals and organizations, particularly those without access to substantial resources. As such,

the use of open-source tools in circuit design automation can lead to cost savings and faster development cycles. Open-source tools offer a wide range of functionalities and support community-based development, allowing designers to leverage the collective knowledge and expertise of the community. Additionally, open-source tools can allow a lower cost of ownership and provide access to a broader range of development resources.

Currently, the way circuit design engineers design circuits involve relying heavily on manual efforts, wherein they select and iterate on the architecture of the system, circuit topology, and device sizing. In addition, layout engineers must draw the layout and wire routing. While digital circuit design has well established automated flows that have allowed scaling to billions of transistors, analog and mixed-signal design methodologies are still mostly manual. Thus, there is a desire for developing analog design automation techniques to speed up the current manual design approach. Previous works have tried to solve these issues using various algorithmic and optimization techniques. The BAG tool [97] uses template-based procedural layout automation for easy layout adaptation. ALIGN [98] uses optimization-based layout generation with various routing algorithms to draw analog layouts similar to what humans would produce. OpenSAR [99] uses a combination of template-based and optimization-based layout generation for ADCs. However, these tools provide limited flexibility in circuit design and do not offer the same flexibility and scalability like digital synthesized circuits. OpenSerDes [100] and FASoC [101] can generate various analog circuit solutions, including PLLs, LDOs, SerDes, and temperature sensors by leveraging digital Automated Place-and-Route (APR) tools. However, these analog blocks are redesigned to use structures composed largely of digital components, which limits performance and circuit design flexibility.

4.2. SRAM Circuit Generation

One of the most common and easy ways to generate an SRAM is by using an SRAM compiler. An SRAM compiler is a software tool that generates memory arrays with specific sizes and configurations to meet the requirements of a particular application. Such compilers are commonly used in digital chip design to create on-chip memory arrays that can be used as caches, registers, or other storage elements. An SRAM compiler typically takes input parameters such as memory size, data width, and access time, and then generates the corresponding memory array along with the necessary control logic and interface circuitry. SRAM compilers are an important tool for chip designers as they allow them to quickly create customized memory arrays. This can help to reduce design time and improve overall chip performance and efficiency.

One of the primary problems with commercial memory compilers is their lack of flexibility. These compilers offer limited customization options, and designers are often forced to choose from pre-built memory architectures that may not fully meet their requirements. The lack of flexibility is a significant issue for designers who need to create unique memory blocks for specific applications. It is often challenging to optimize the performance of these memory blocks and meet power and area constraints. Commercial memory compilers can generate an SRAM for a given PDK but these are the outcome of a human-driven design effort for each PDK and cover a fixed design space that usually emphasizes high performance. Such limitations restrict compilers' usage for applications such as ultra-low-power systems, which often operate in the nW to μW space.

Another disadvantage of commercial memory compilers is their high cost. Many commercial memory compilers require a significant investment in licensing fees, which can add up to a substantial expense for semiconductor design companies. For small companies or start-ups, the high cost of commercial memory compilers can be a barrier to entry, preventing them from creating competitive products.

Additionally, commercial memory compilers may not provide optimal solutions for complex memory designs. While they may work well for simple memories, their automated design methods may not be well-suited for complex designs, which can lead to suboptimal performance or area utilization. This lack of optimization can cause significant design issues for the final product and impact its overall performance.

Furthermore, commercial memory compilers may also have limitations when it comes to design flexibility. The compilers may be limited in their ability to implement custom design requirements, which can be a significant issue when trying to create unique memory designs. In some cases, designers may need to create memory blocks that are not supported by commercial memory compilers. In such cases, they may need to resort to manual design methods, which can be time-consuming and error-prone.

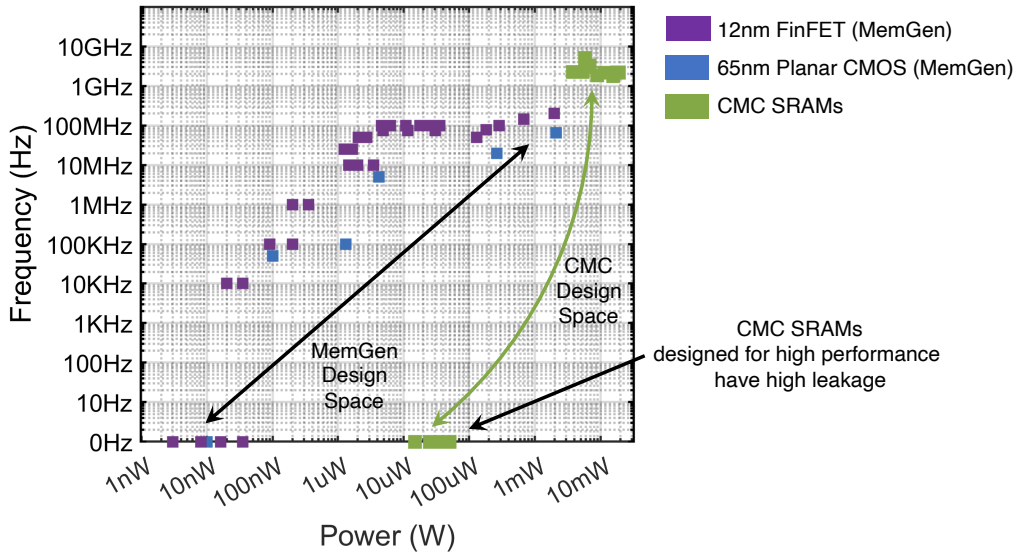
Another significant disadvantage of commercial memory compilers is the lack of support for new process technologies. As technology nodes become smaller and more complex, there is a growing need for memory compilers to support these new technologies. However, it can take a long time for commercial memory compilers to catch up with the latest technology nodes. This delay can cause significant issues for semiconductor design companies, which need to stay ahead of the curve to remain competitive.

Lastly, commercial memory compilers may not always be reliable. Despite rigorous testing, there may be bugs or errors that go undetected until later in the design process. This can lead to delays and additional costs associated with fixing the issue. Additionally, some commercial memory compilers may not be compatible with specific EDA tools or design flows, which can further complicate the design process.

In conclusion, commercial memory compilers offer a convenient solution for semiconductor designers looking to create memory blocks quickly and efficiently. However, there are several significant problems and disadvantages associated with their use, including limited flexibility, high costs, lack of optimization, limited design flexibility, lack of support for new process technologies, and unreliability.

4.2.1. Memory Generator (MemGen)

Static Random-Access Memories (SRAMs) form an integral part of System-on-Chips (SoCs), wireless sensor nodes and other Internet-of-Things (IoT) devices. The SRAM has a large multi-dimensional design space that includes various kinds of bit-cell designs, peripheral assist-circuit designs, operating voltages, and frequencies. Custom design of memories for any application in this broad design space is a tedious, iterative, and mostly manual process. Commercial memory compilers (CMCs) [102]-[104] have been used as an automated alternative but they have a lot of issues as discussed in the previous section. To address these issues and allow easy, autonomous, and versatile generation of memory macros, we present MemGen ("Memory Macro Generator"), an open-source memory macro generation framework that creates tapeout-ready integrated memories across a broad range of voltages, frequencies, and capacities. The new framework uses a template and cell-based design methodology and leverages the conventional digital flow to generate optimized memories based on high-level user intent. The proposed framework's template and cell-based design approach and its digital flow-based construction allows it to be highly modular, process-portable, and easily augmentable. The framework's novelty is demonstrated by generating multiple memories for various user intents in a planar 65nm and 12nm FinFET process. MemGen is also verified by fabricating multiple instances of 12nm and 65nm auto-generated memories.



(a)

Desired Features	This Work	CMCs [102]-[104]	Academic Reported Compilers		
			[105]	[106]	[107]
Open Source	Yes	No	No	Yes	Yes
Planar CMOS Support	Yes	Yes	Yes	Yes	Yes
FinFET Support	Yes	Yes	No	No	No
PDK Agnostic	Yes	No	No	Limited	Limited
Design Space Exploration & Optimization	Yes	No	No	No	No
Fabricable/Tapeout Ready	Yes	Yes	No	No	No
Silicon Verified	Yes	Yes	No	No	No

(b)

Fig. 4.1. (a) Target design space for MemGen (b) Comparison of MemGen with other compilers in terms of features and capability.

Fig. 4.1 (a) shows the comparison of the intended design and application space between MemGen, CMCs and other academic reported compilers [105]-[107]. Although CMCs offer many features, they have a narrower design space, thereby limiting their application space. Other academic compilers also offer a suite of various features, but they fall short on many other features as shown in the comparison in Fig. 4.1 (a). Memgen offers to generate memories in other regions of the design space where CMCs cannot be used, making them suitable for many other applications. Unlike other CMCs and academic compilers, MemGen can generate memories by performing device-circuit-architecture co-design. This is enabled by a closed-loop integrated flow that involves the translation of high-level user intent (e.g., voltage, frequency, and capacity) into an optimized SRAM layout through tightly coupled design-space exploration, optimization, and layout generation. Additionally, to the best of the authors' knowledge, MemGen is first open-source memory compiler to support advanced FinFET processes. The source code of the tool is available from https://github.com/idea-fasoc/fasoc/tree/master/generators/memory-gen.

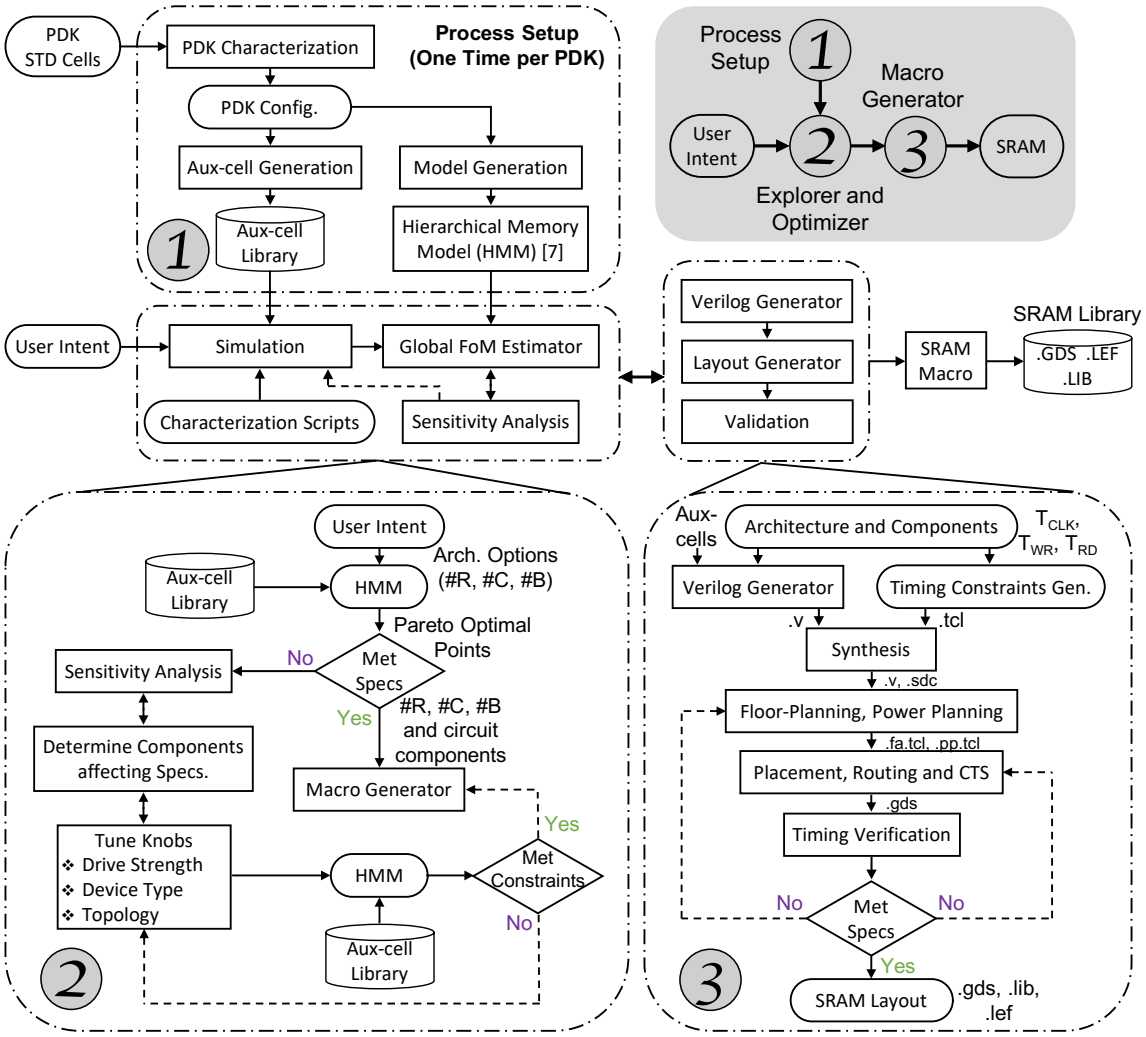


Fig. 4.2. MemGen Framework high-level overview and functioning.

Fig. 4.2 shows the high-level overview of the MemGen framework and the critical steps involved in generating a memory. To enable MemGen to be able to create memories on a specific technology, a one-time Process Design Kit (PDK) setup is required. The first step in the setup process is the PDK characterization, which involves running device-level simulations to extract information such as transistor behavior, bit-cell characteristics, metal parasitics, threshold voltage (VT), and FO4 delay using a template-based methodology which separates the technology dependent and independent aspects of the circuits. The second step involves the generation of aux-cells. Aux-cells are small SRAM peripheral circuits that extend the standard-cell library and provide specific analog functionality required for the memory operation. These aux-cells vary in terms of drive strengths, circuit topology, and VT. and include all files which are required as a part of the conventional synthesis

and Auto-Place-Route (APR) flow. The final step is the generation of PDK-specific SRAM Hierarchical Memory Model (HMM) which allows quick estimation of SRAM global FOMs energy(E) and delay(D), thereby circumventing the need of resource intensive and time-consuming complete circuit simulations [108].

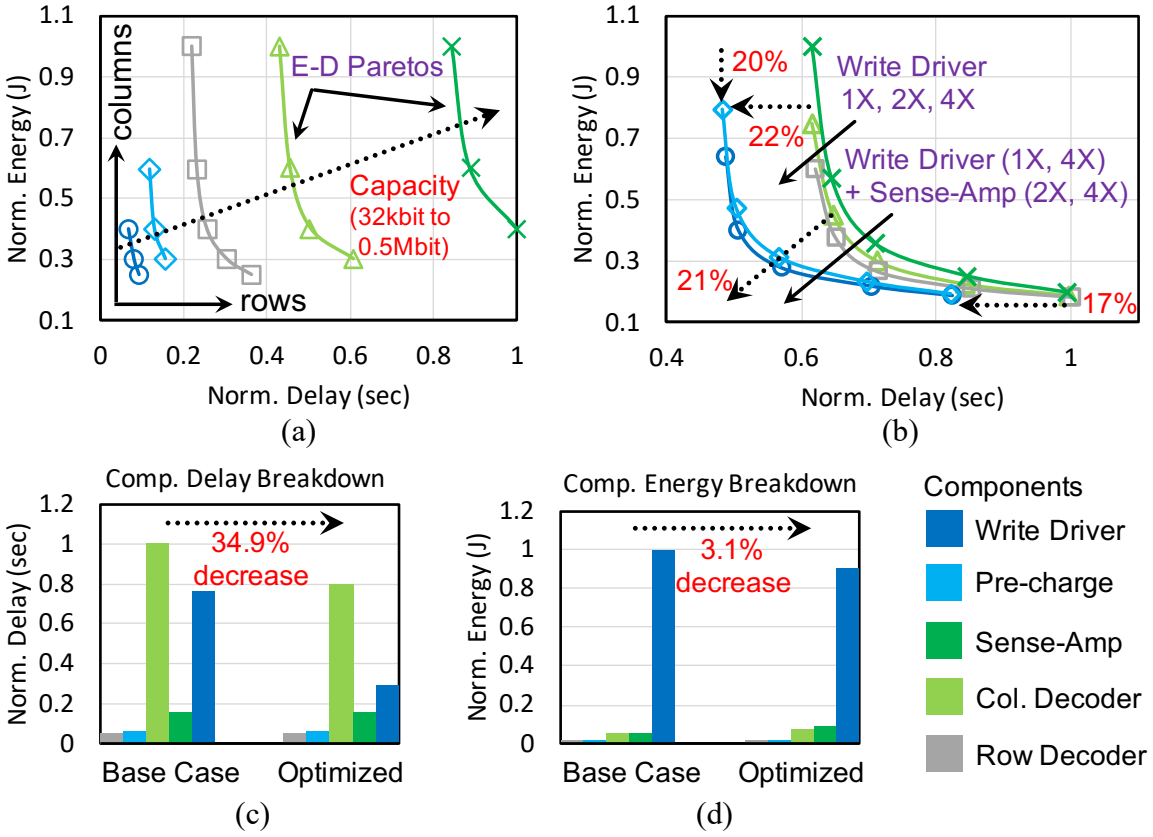


Fig. 4.3. (a) E and D pareto curves with varying capacity generated using HMM. (b) Pareto improvement using component level optimization. Component-wise breakdown of (c) D and (d) E.

To optimize the SRAM design for certain energy, delay, and area, which are in turn dependent on user intent, a weighted cost function $C(x)$ is calculated as

$$C(x) = W_E E(x) + W_A A(x) = W_E \sum_{i=1}^{c} e_i(x) + W_A \sum_{i=1}^{c} a_i(x) \tag{4.1}$$

where x is a vector of n optimization variables $x_1, x_2, \ldots, x_n$, $c \rightarrow$ number of sub-components, $E(x) \rightarrow$ energy-per-access and $A(x) \rightarrow$ SRAM area, $e_i(x) \rightarrow$ component's energy, $a_i(x) \rightarrow$ components' area, $W_E \rightarrow$ energy weight, $W_A \rightarrow$ area weight. The user may vary energy and area weights as per their application priorities.

MemGen considers the number of banks (B) and the number of rows (R) and columns (C) per bank at the architectural level and device sizing, device type (High-Low-Reg. VT) and component topology at the circuit level as optimization variables to minimize the cost function. The optimization process starts by generating a set of optimal FOMs tradeoff points for a given user intent using the HMM, as shown in Fig. 4.3. If any of the optimal points satisfy the user intent, the framework moves on to the macro generation phase. Otherwise, the framework proceeds with an iterative sensitivity analysis-based design space exploration by tuning circuit-level knobs and minimizing the cost function until the user intent is met. The design space can be extended by adding new architectures or new aux-cells to the aux-cell library using circuit templates, thereby making it modular and easily augmentable. The final step of generating the layout of the SRAM macro involves verilog and timing information generation as shown in Fig. 4.2. With architecture (R, C and B) and circuit decisions as inputs from the optimization phase, parameterized templates are used to generate timing constraints and a synthesizable bespoke structural verilog netlist. These are then passed through the standard digital flow to generate the Register Transfer Language (RTL) netlist. With the Register RTL netlist and the floor planning directives as inputs, the PNR step hierarchically creates the bit-cell array, row and column peripheral circuitry using an abutment process to build a bank. Several of these banks, along with the bank control logic and the memory controller are then placed and routed to form a complete multi-bank memory macro in a hierarchical tree-based architecture, as shown in Fig. 4.4 (a). The single clock cycle memory transactions for the input and output signals of the SRAM macro are shown in Fig. 4.4 (b). As part of the final design signoff process, MemGen performs a functional and performance check on the final output to ensure that the design operates according to the user's intent.

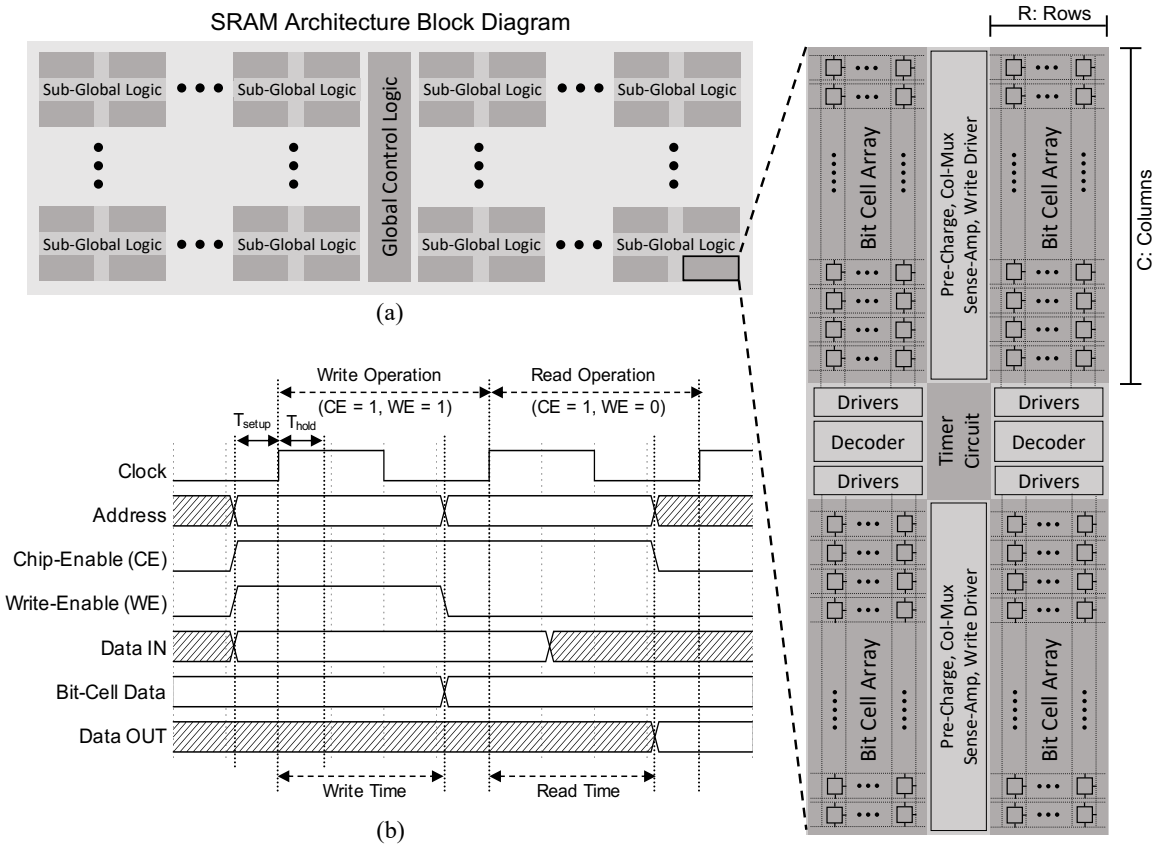
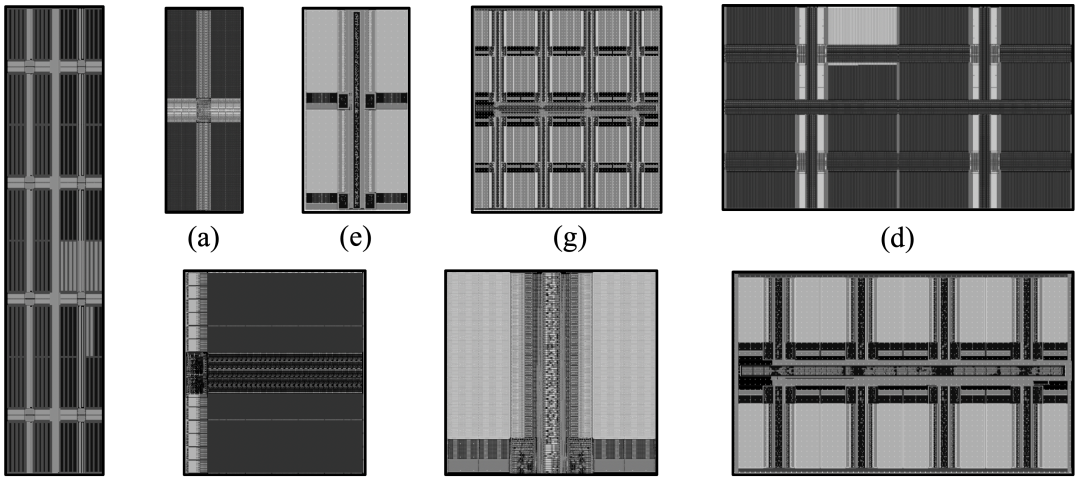


Fig. 4.4. (a) Block Diagram of the SRAM Architecture employed in the macro generation process (b) Timing Diagram of the SRAM.

Fig. 4.5 shows layouts of various SRAM macros auto-generated using MemGen for various user intents in a planar 65nm and a FinFET 12nm technology. As seen in the figure, MemGen is able to create memories for any given number of rows, columns, and bank sizes.



	65nm (Nominal V_{DD} = 1.2V)				12nm (Nominal V_{DD} = 0.8V)			
	(a)	**(b)**	**(c)**	**(d)**	**(e)**	**(f)**	**(g)**	**(h)**
Capacity (bit)	64K	64K	0.5M	0.5M	64K	64K	0.5M	0.5M
Area (mm^2)	0.1175	0.0895	1.045	0.8290	0.0109	0.00817	0.1152	0.0935
Rows	128	256	128	256	128	256	128	256
Columns	128	128	128	128	128	128	128	128
Banks	4	2	32	16	4	2	32	16
Energy (pJ)	13.1	15.6	62.4	78.3	3.49	4.64	27.89	37.13
Delay (ns)	8.3	9.7	12.2	15.3	1.31	2.48	2.40	3.57

65nm: Custom-made DRC compliant 6T bit-cell
12nm: Foundry 6T bit-cell

Fig. 4.5. Layouts of different SRAMs auto-generated using MemGen in planar 65nm and 12nm FinFET process.

4.2.2. Assist Circuit Features

In order to achieve optimal energy efficiency while meeting performance requirements, it is important for SRAM circuits to operate over a wide range of supply voltages. However, as semiconductor technologies continue to scale down, the variability of SRAM cells increases, limiting their ability to operate at low voltages compared to logic circuits. To overcome this limitation, assist techniques such as dynamic changes to the bit-cell operating characteristics, such as increasing the voltage of the word-line above the cell voltage, can be employed to reduce the minimum operational voltage (VMIN) for SRAM circuits. MemGen employs assist circuit features to reach the target supply voltage required by the user. It first calculates the SRAM VMIN and then proceeds to estimate the amount of assist circuit required to reach the target voltage. This could mean using larger capacitor or transistor arrays to change SRAM internal circuit voltages to desired values, such that the SRAM VMIN is met.

MemGen currently supports two kinds of assist circuits. It supports Word-Line Underdrive (WLUD) circuit that helps to reduce read VMIN and Negative Bit-line (NBL) Assist Circuit to help reduce write VMIN. These are chosen because of their ability to reduce VMIN most effectively compared to other assist circuits. The circuit schematic for both assist is shown in Fig. 4.6.

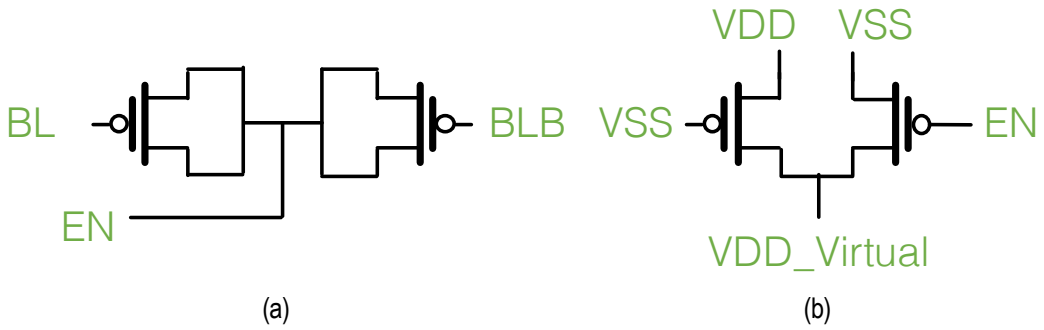
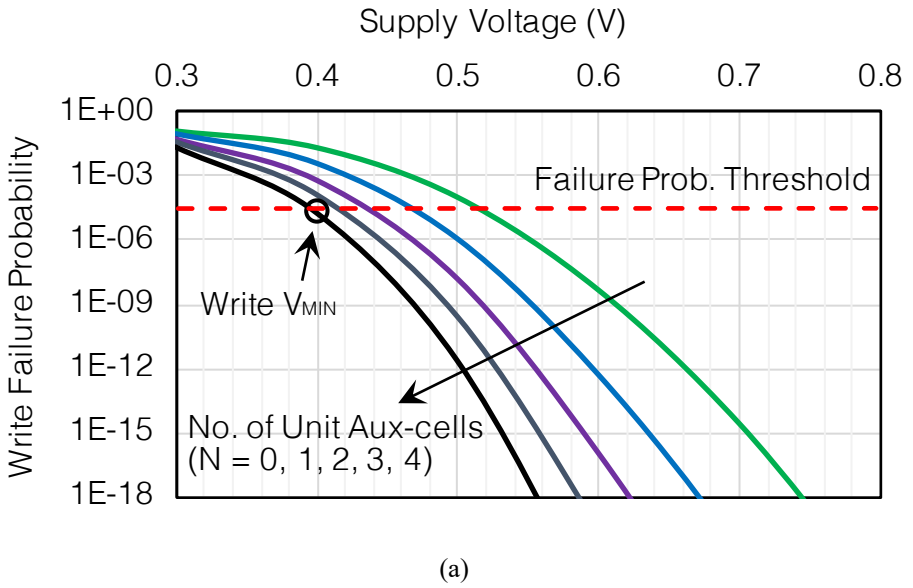
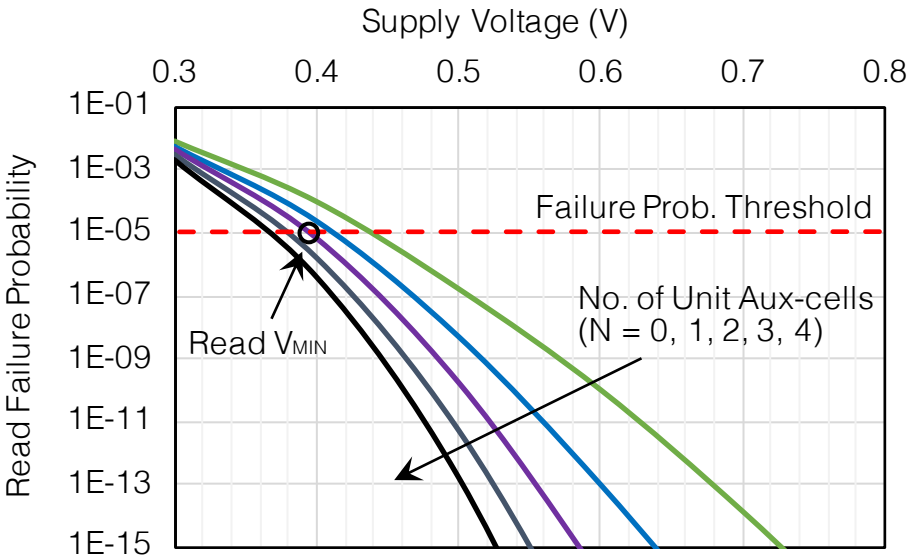


Fig. 4.6. Circuit schematic for (a) NBL write assist (b) WLUD read assist.

The NBL assist circuit is widely used and is generally able to reduce the VMIN the most because it simultaneously improves both the VGS and VDS of the access transistor in an SRAM bit-cell, thereby greatly improving its drive strength. Both these assist circuits are also simpler to implement in context with modularity and automated tiling requirements. The read and write failure probability with respect with varying supply voltage and amount of assist is shown in Fig. 4.7. As seen in the figure, the tool can estimate the VMIN without assist. It can then increment the amount of assist by increasing the number of the aux-cells and then recalculate the failure probability. When the VMIN is reached to the desired level, then the tool uses the corresponding number of aux-cells in the final design. In this case, the failure probability threshold is 1E-5 because of the small size of the memory (2KB). For larger memories (hundreds of KBs or several MBs), the failure probability threshold is 1E-9. The tool can accordingly account for this based on the user input.



(a)



(b)

Fig. 4.7. Failure probability and corresponding VMIN with varying supply voltages and different number of assist aux-cells for (a) Write operation (b) Read operation.

The layout of the different assist circuits is shown in Fig. 4.8. The capacitive array for the NBL circuit in the column periphery is shown on the left. These can be tiled in a modular way in each column based on the calculations shown previously. Similarly, the WLUD aux-cells for each row are tiled in the row periphery as shown on the right. They are tiled such that they share a common n-well and common virtual power routing. This allows area efficient integration into the row peripheral circuitry in the SRAM bank.

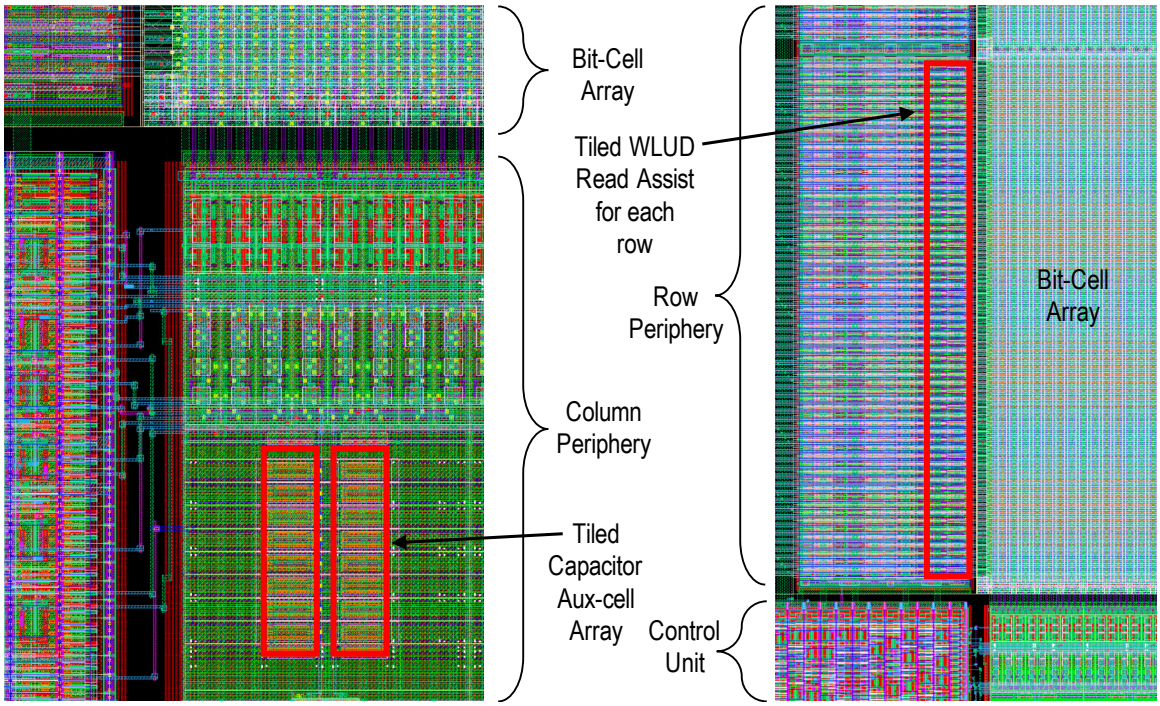


Fig. 4.8. Tiled layouts for the read and write assist aux-cells in the 12nm FinFET process.

4.2.3. Measurement Results

To experimentally verify the framework, we recorded measurement results from two different SRAM macros (64kbit and 128kbit) auto-generated for given user intents using MemGen in a planar 65nm CMOS technology. The first user intent is aimed at subthreshold operation and low frequency (50KHz). The second user intent is aimed at super-threshold operation and high frequency (50MHz). Fig. 4.9 shows the frequency and the power measurement results and the step-wise timing breakdown for the generation process of these memories using MemGen. As seen in Fig. 4.9, MemGen is able to achieve the desired performance and operating point for both user intents, with a runtime of ≤138 min in each case.

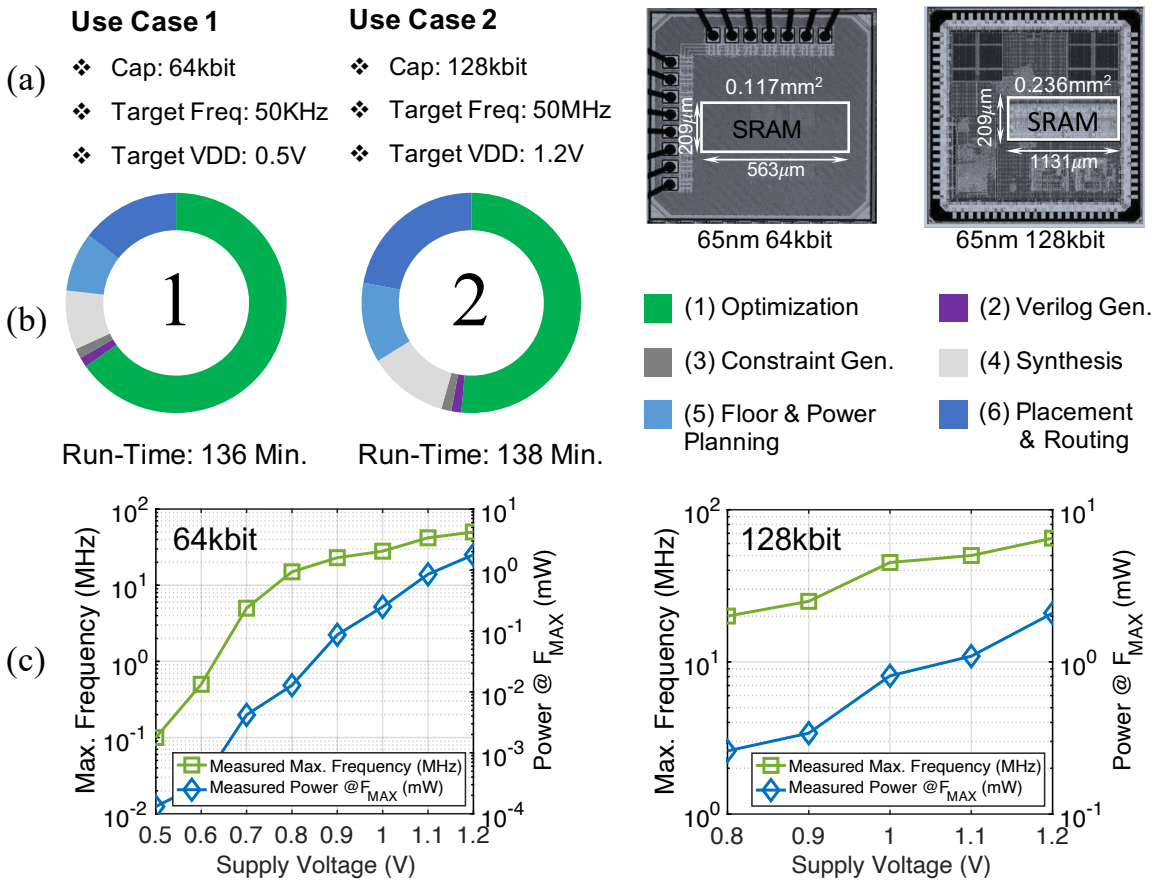


Fig. 4.9. (a) User intent (b) Framework run-time breakdown (c) Frequency and power measurement results.

To verify the tool in 12nm FinFET process, we implement a 64KB (0.5Mbit) design. The tool runs the optimization process and decides the architectural and design parameters. The resulting memory is constructed using eight 8KB banks and each 8KB bank is composed of four 2KB local banks (shown in Fig. 4.10 (a)). The tool then routes all the banks together using the digital synthesis flow to make sure the design meets timing requirements for desired frequency target. Measurement results show that the memory is fully functional from 0.8V down to 0.6V with a max. operating frequency of 200MHz and 50MHz respectively.

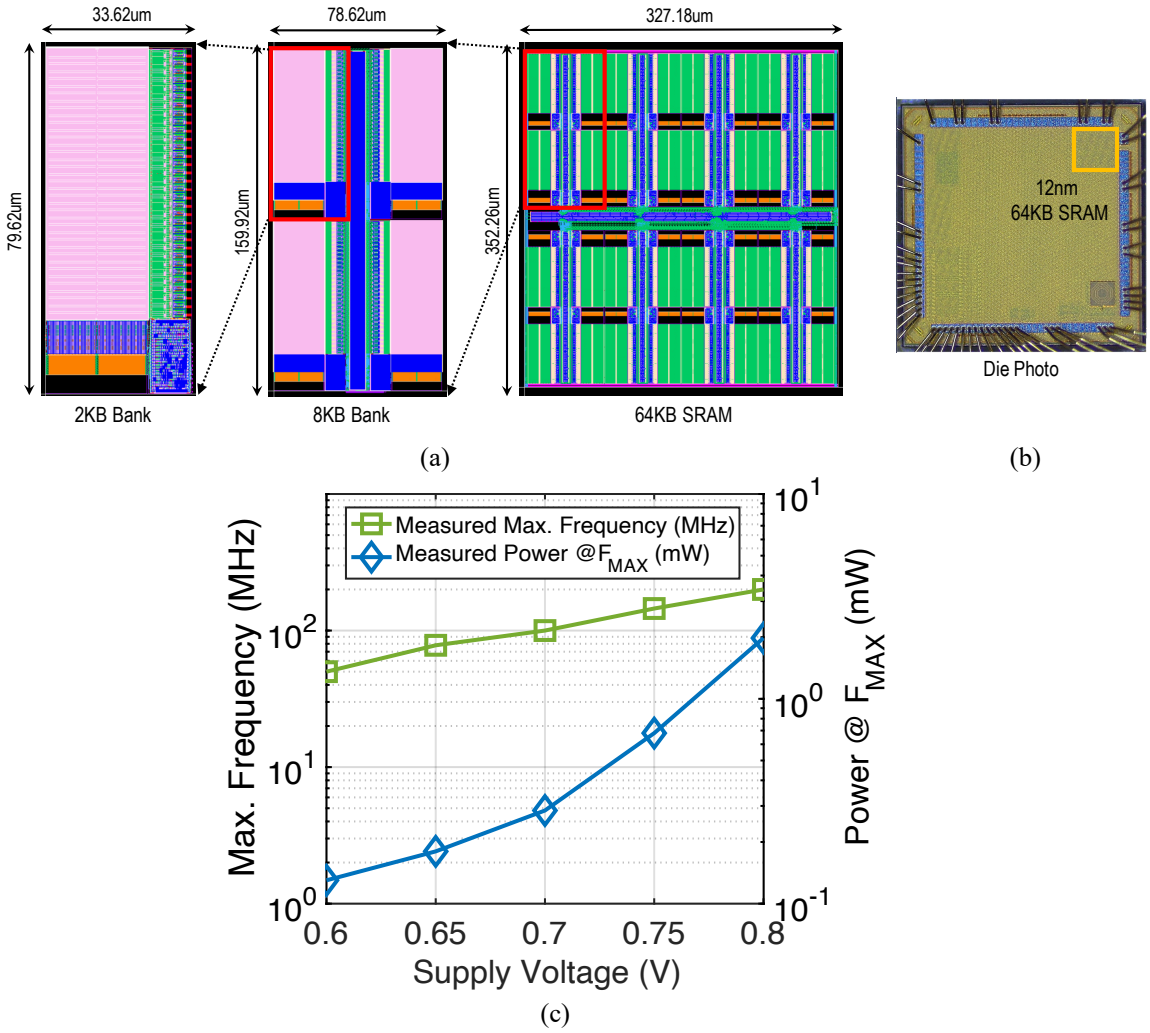


Fig. 4.10. (a) Layout of local bank hierarchy and 64KB SRAM macro. (b) Die photo (c) Measured power and frequency for the 12nm 64KB SRAM.

To verify the tool in for various read and write assist techniques across various row-column configurations, we implemented four 8KB SRAM macros in 12nm FinFET process. The desired goal for the minimum operating voltage was 0.4V. The tool was used to automatically calculate the amount of assist needed for each row-column configuration. The layout for each macro is shown in Fig. 4.11. Depending on the amount of assist needed for each architectural combination, several assist aux-cells were tiled together. The tool used the digital synthesis flow for closing the timing on the top level. With the Register RTL netlist and the floor planning directives as inputs, the PNR step hierarchically created the bit-cell array, row and column peripheral circuitry using an abutment process to build a bank. Several of these banks, along with the bank control logic and the memory controller are then placed and routed to form a complete multi-bank memory macro in a hierarchical tree-based

architecture, as shown in Fig. 4.11. The final design signoff process involved performing a functional and performance check on the final output to ensure that the design operates according to the user's intent.

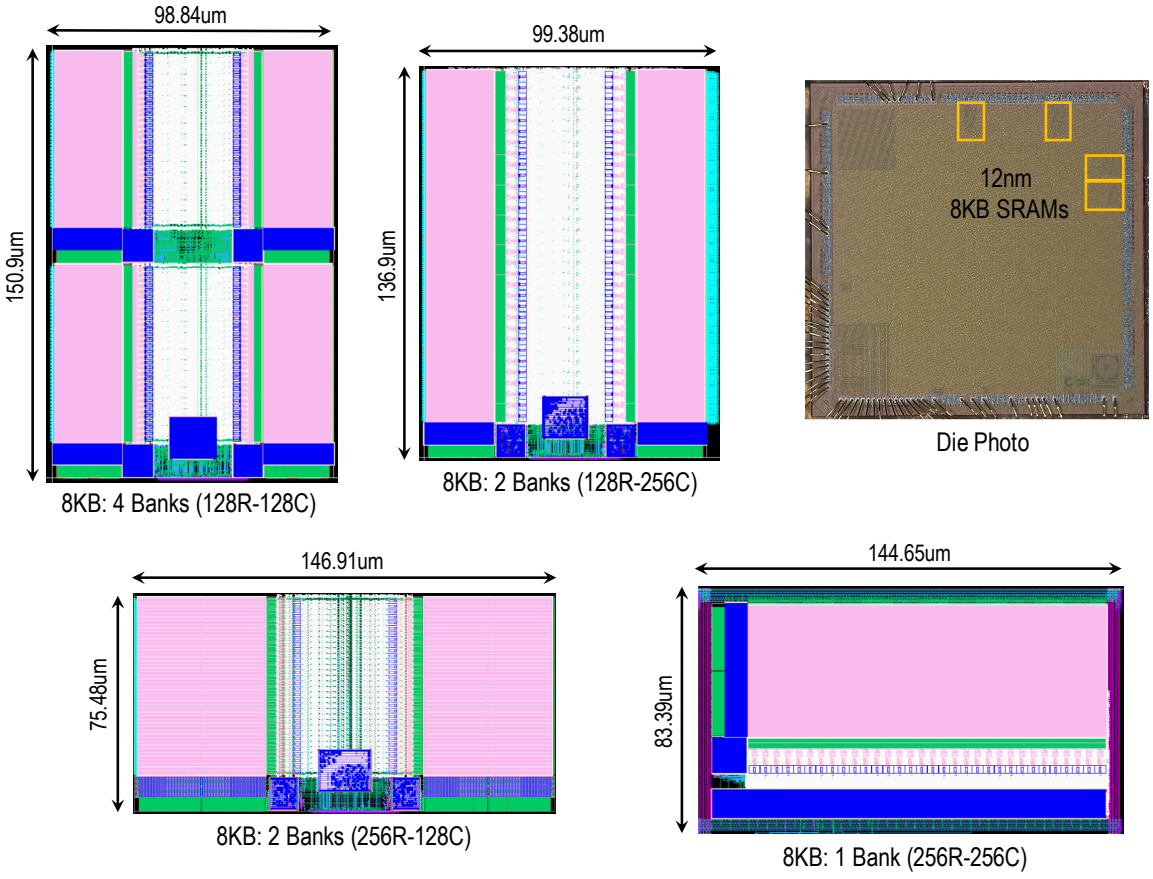


Fig. 4.11. Layouts of four 8KB SRAM macros with varying rows, columns, and banks in 12nm FinFET technology.

The layouts show various combinations of rows, columns, and banks. The four resulting combinations to verify the assists are:

1. 128 Rows, 128 Columns, 4 Banks
2. 128 Rows, 256 Columns, 2 Banks
3. 256 Rows, 128 Columns, 2 Banks
4. 256 Rows, 256 Columns, 1 Bank

The measurement results were recorded for all four SRAM macros across supply voltage and frequency. Fig. 4.12 shows the performance and power measurement results for the four macros. As seen in figure, all memories are fully functional down to 0.4V. Without the read and write assists, the memories work only down to 0.55V

(with the nominal voltage being 0.8V). When the write assist is enabled, the memories can work down till 0.45V. When the read assist is enabled, the memories work down to 0.4V. Note that the foundry bit-cell was used in all macros in the 12nm FinFET process and custom CDRC complaint bit-cell was used in the case of all 65nm SRAM macros.

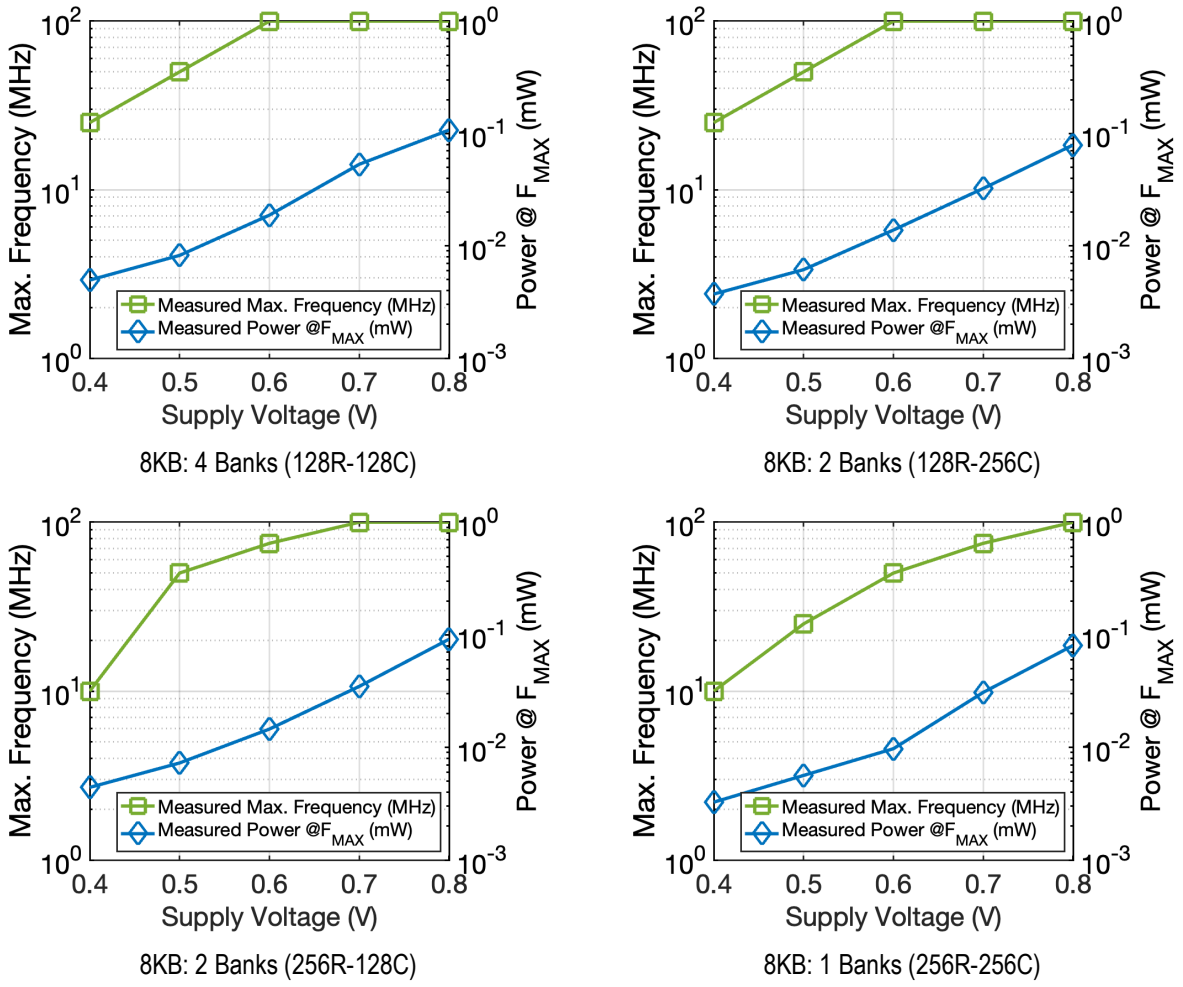


Fig. 4.12. Measured power and frequency for the four 8KB SRAM macros with varying rows, columns, and banks in 12nm FinFET process.

Capacity (b)	Col. Mux	Banks	CMC Area (mm^2)	MemGen WAssist Area (mm^2)	% Difference
16K	4	1	0.00252	0.0025	-0.862
32K	4	2	0.00451	0.005	10.88
64K	4	2	0.00758	0.00842	11.05
128K	4	4	0.0149	0.01684	12.34
256K	4	4	0.0275	0.0312	13.55
0.5M	8	8	0.0464	0.053	14.19

Fig. 4.13. Comparison of area between MemGen memories and CMC memories with varying capacities.

The area comparison between memories generated using MemGen and CMCs for various capacities is shown in the table in Fig. 4.13. As seen in the figure, the area penalty for MemGen generated memories is less than 15% in the worst case. For smaller memories, MemGen can save area because it is able to use appropriately designed peripheral circuitries for required application as opposed to using very high drive strength circuits in CMC memories and running them at lower frequencies. This highlights one of the main benefits of MemGen that allows us to generate memories designed specifically for the design application and user intent, leading to very optimized and efficient designs.

4.3. Low Dropout Voltage Regulator (LDO) Circuit Generation

One of the most used analog circuits in modern SoCs is the low dropout voltage regulator (LDO). While digital circuit implementations of the LDO, which replace analog elements with digital switches and controllers, have allowed easy generation using synthesis and APR tools [109]-[113] (as shown in Fig. 4.14), digital-analog-hybrid and analog LDOs are not able to benefit from this approach since they include analog elements in addition to digital cells. Analog LDOs still out-perform digital LDOs (DLDOs) in transient response and PSRR, stemming from the DLDOs' discontinuous operation, especially at low frequencies.

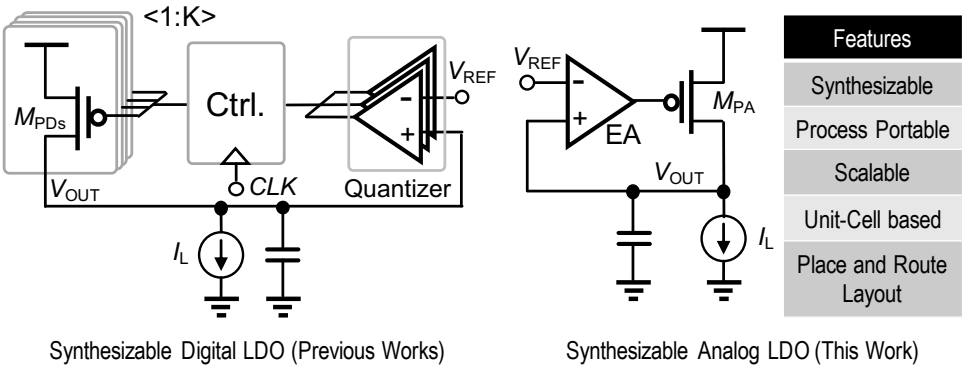


Fig. 4.14. This work targets solely the analog LDOs, allowing for the first time, fully synthesizable analog and hybrid LDOs.

In this work, we present three all-analog LDO designs at different design points that were generated using a synthesizable unit-cell based approach. We leverage the well-established digital synthesis-based tools to synthesize RTL descriptions of the analog LDOs that retain their analog circuits and topology, significantly cutting back on manual layout and verification efforts. Prior to this work, synthesizable LDOs had to use entirely digital topologies, but our approach to all-analog synthesizable LDOs can also pave the way to combine elements from both digital and analog designs to enable automatic generation of fully end-to-end synthesizable hybrid LDOs, thereby allowing reduced manual effort, improved scalability, and easier process portability.

4.3.1. Analog Circuit Generation Methodology

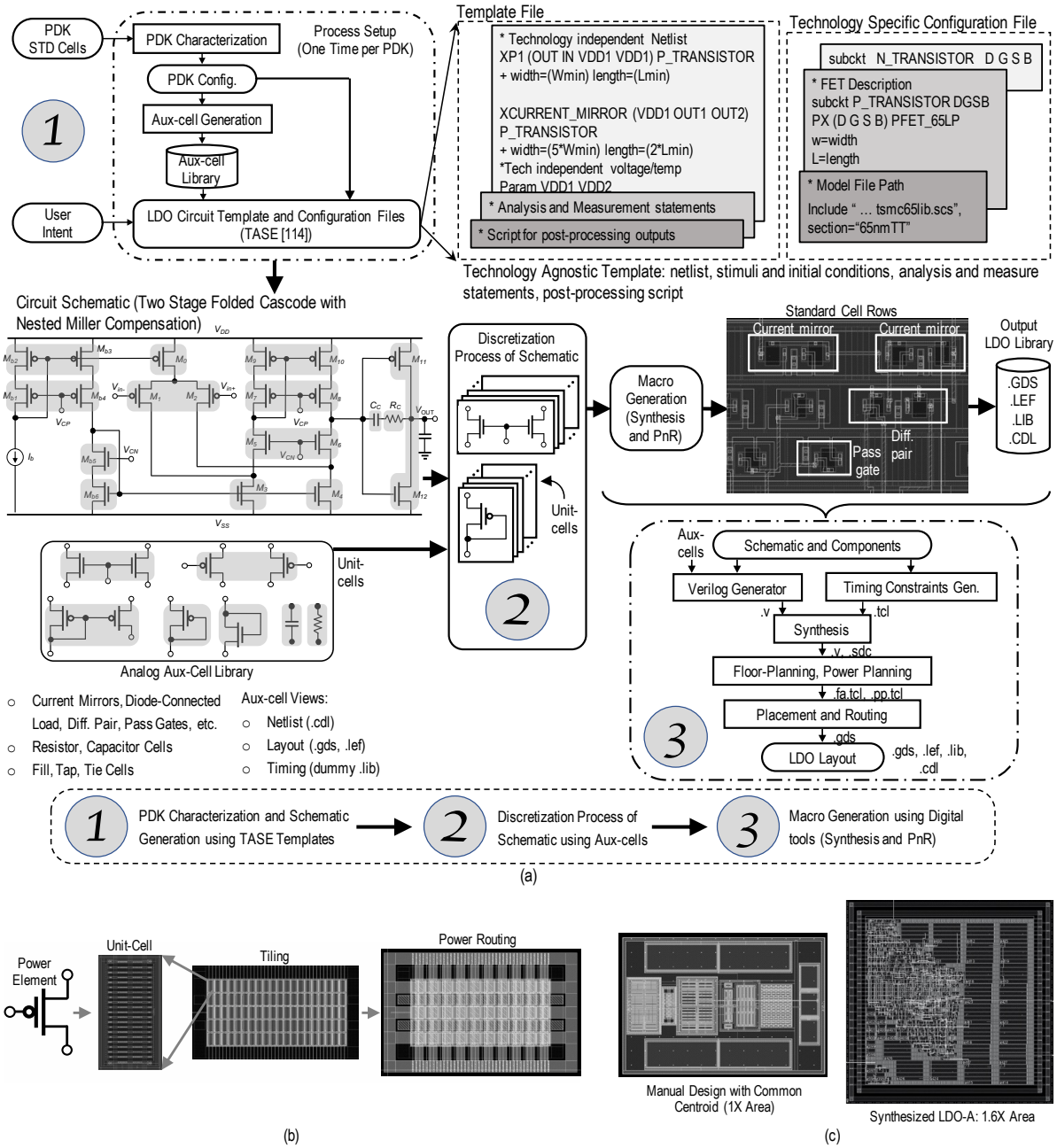


Fig. 4.15. (a) Digital-flow based generation methodology for analog LDOs using synthesizable unit-cell based approach. (b) Layout generation of power element using unit-cell based construction (c) Comparison between manually created common centroid layout and synthesized layout generated using unit aux-cells.

Fig. 4.15 (a) shows the flow diagram of the synthesizable methodology used to generate the LDOs. It involves the creation of aux-cells (auxiliary unit-cells), characterization scripts, and circuit design models and templates as a one-time effort per process design kit (PDK). Aux-cells in this work are small analog circuits that make up

the analog aux-cell library and provide specific analog functionality (unlike aux-cells in prior works that are digital in function, e.g., [101]). Examples include analog circuits like current mirrors, diode-connected loads, differential pairs, etc., passives like resistors and capacitors, and miscellaneous fill, tap, and tie cells as shown in Fig. 4.15 (a). Most aux-cells are similar in size to a D flip-flop and can be placed on standard cell rows but with ports that allow their connection by the APR tool into analog structures. The creation of aux-cells is simplified by using design templates in tandem with PDK characterization scripts. The templates capture the aux-cell's precise circuit behavior in a SPICE simulation. The characterization scripts operate on the PDK to derive technology-specific parameters (threshold voltage, metal parasitics etc.) required to set knobs within the templates. The aux-cell generation includes the netlist, layout, (dummy) timing library, and other files required to proceed with synthesis and APR. Presently, the layouts for the aux-cells are manually created as a one-time effort per PDKL, similar to how digital standard cell libraries are built, although, layout tools like ALIGN and BAG could be used for auto generation of aux-cell layouts.

The LDO generation begins by translating high-level user-intent into analog specifications that satisfy the user constraints. The circuit design is derived from the parameterized templates using TASE [114],[115] and circuit equations [116] as shown in Fig. 4.15 (a). These templates are technology agnostic and include information about the netlist, stimuli and initial conditions, measure and analysis statements, and post-processing scripts. The user intent along with simulation parameters (e.g. Monte Carlo seed), template parameters (e.g. device sizes), temperature and voltage, and model files form the configuration files. The circuit templates and configuration files are a one-time effort for each circuit. Once the circuit template has been created, the user can quickly run a whole suite of design (as shown in Fig. 4.16) and verification simulations using these technology agnostic circuit templates and design the LDO, thereby cutting back significantly on design time.

The next phase is the Verilog generation that leverages the schematics to produce a synthesizable Verilog description of the block that incorporate the analog aux-cells. Fig. 4.15 (a) shows the circuit diagram of our two-stage amplifier with Miller frequency compensation and pole-zero cancellation. The analog sub-components that make up the amplifier are also highlighted. Each analog sub-component is discretized into small aux-cells using the unit-cells from the analog standard cell library. These aux-cells are placed in series or parallel according to the schematic of the LDO and can be aggregated to vary the effective device width of a particular transistor or cell. Fig. 4.15 (a) shows the process of discretization of the analog circuit sub-components and its

APR. The Verilog is then passed on to a digital flow step to perform synthesis, APR, DRC, and LVS verification. The last step is a verification and reporting of the generated LDO. The full circuit goes through parasitic extraction, SPICE simulations, and other verification to generate performance numbers.

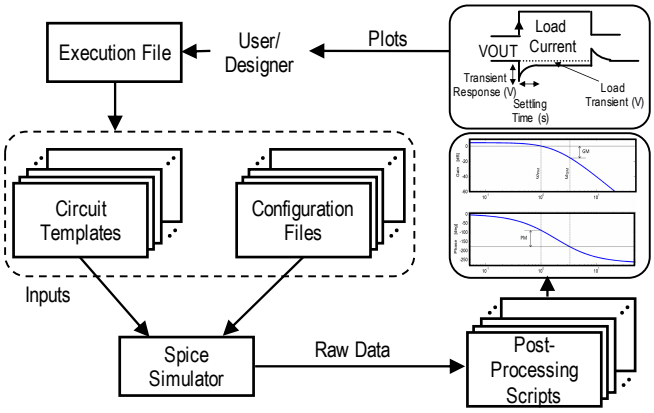


Fig. 4.16. User can quickly run a whole suite of simulations to design and verify LDOs using one-time generated technology agnostic circuit templates.

4.3.2. Performance Evaluation and Measurement Results

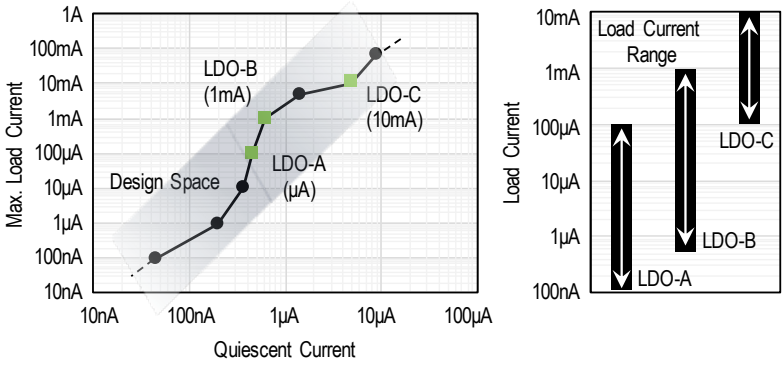


Fig. 4.17. Design space for synthesizable Analog LDOs and load current range for three design points (LDO-A, -B, -C) spanning a maximum load current range of 100X.

Fig. 4.17 shows the design space for various LDO designs. Three design points (LDO-A, -B, -C) spanning a maximum load current range of 100X were selected for measurement and verification purposes to showcase and prove the synthesizable analog unit-cell based approach. A manually drawn LDO-M, identical in schematic design to LDO-A was also generated and measured to compare it with the synthesized approach in terms of performance. The four LDO designs (LDO-A to -C and LDO-M) were fabricated in a 65nm LP process. Fig. 4.15 (c) shows the layout comparison between the manually created common centroid layout (LDO-M) and the synthesized version (LDO-A). Fig 4.18. shows the measurement results comparing the performance between the

manually created LDO-M and synthesized LDO-A. The increased interconnect from the auto-routing and the associated parasitics lead to some loss in transient response and power supply rejection ratio (PSRR), but the difference in performance is minimal.

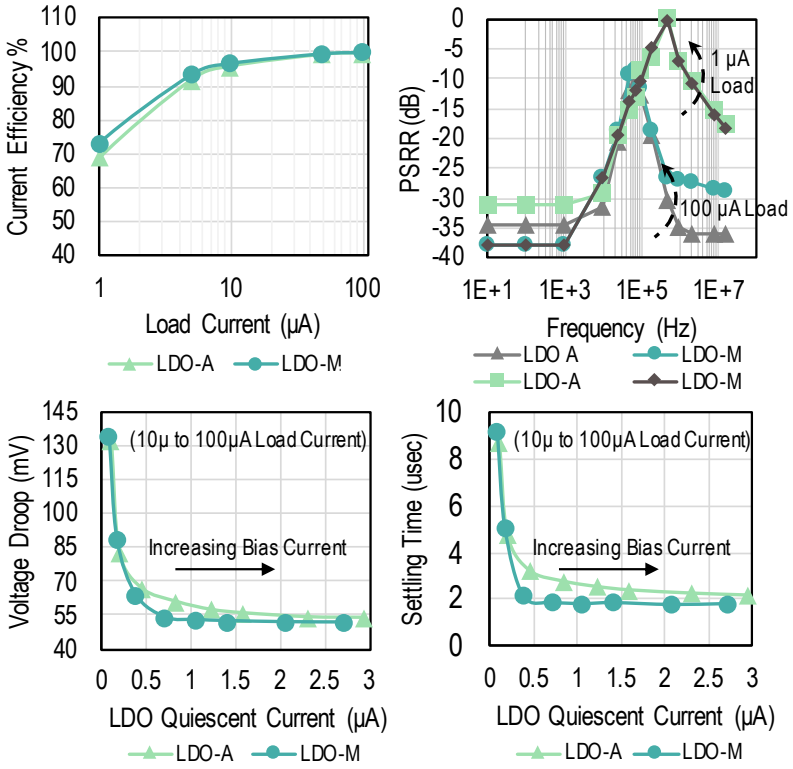
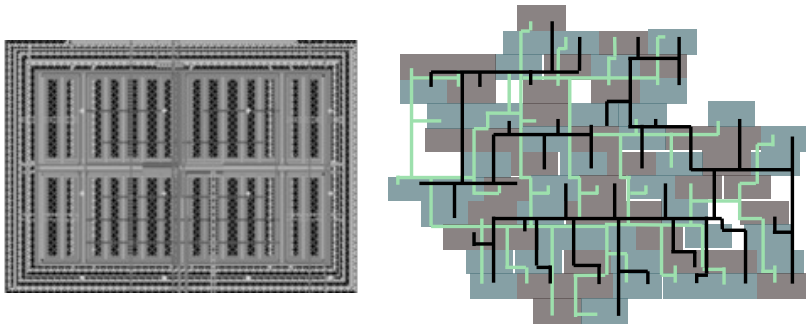
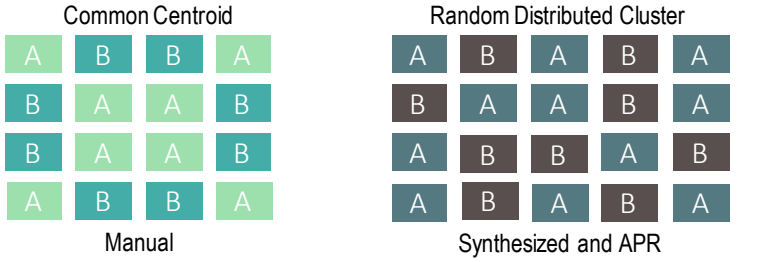


Fig. 4.18. Performance comparison between synthesized LDO-M and LDO-A for various metrics. LDO-M and LDO-A are manual and synthesizable versions of the same LDO design. (Load current step time:1ns)

It is desirable to have low input offset variability in LDOs, especially in precision circuit applications. Within-die variations affect devices differently based on their location on a chip, resulting in differential mismatch. Within-die systematic variations are often modeled by linear gradients, while random variations are modelled with distributions. Random variations have uncorrelated and spatially correlated components characterized by a correlation distance [117]. Fig. 4.19 shows the input offset variability comparison between a manually drawn common-centroid (CC) layout and our APR-based random-distributed cluster of unit-cells. The standard deviation for the CC version is 11.69mV, which is common for untrimmed, un-chopped LDOs. In comparison, the synthesized LDO-A reduces variability by 41.4% down to 6.85mV.





Input Offset	LDO-M (Manual)	LDO – A*	LDO – B*	LDO – C*
Mean (mV)	4.35	5.56	2.01	1.19
Sigma (mV)	11.69	6.85	8.03	4.13

(Measurement results from 10 Chips) *Synthesized
LDO-M and LDO-A are manual and synthesizable versions of the same LDO design

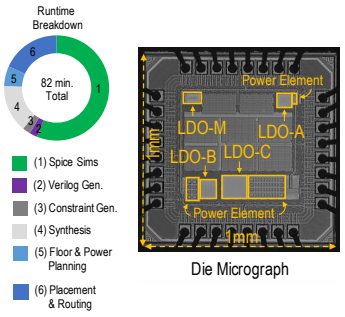
Fig. 4.19. Input Offset comparison between manually drawn common centroid layout and random distributed cluster generated by the Auto-Place-Route.

This Work

	LDO-M†	LDO-A*	LDO-B*	LDO-C*	[109] ISSCC'20	[110] ISSCC'20	[111] TCASII'20	[112] ESSCIRC'17	[113] ISSCC'18
Technology (nm)	65	65	65	65	28	10	28	130	65
Type	**Analog††**	**Analog Synthesize -able††**	**Analog Synthesize -able**	**Analog Synthesize -able**	Digital Synthesize-able	Digital Synthesize -able	Digital Synthesize -able	Digital Synthesize-able+Analog**	Digital
I_{LOAD} Max. (mA)	0.1	0.1	1.0	10	160 to 480	2740	4-6.5	15	100
V_{IN} (V)	1.2	1.2	1.2	1.2	0.5-1.0	0.7-1.05	0.5-1.0	1-1.2	0.6-1.2
V_{OUT} (V)	0.5-1.1	0.5-1.1	0.5-1.1	0.5-1.1	0.45-0.95	0.65-0.95	0.45-0.95	0.6-1.0	0.4-1.1
C_{LOAD} (pF)	10	10	10	10	4.11pF+7nF	534	100	--	0.04
I_Q	370nA	457nA	628nA	4.868µA	7.7µA to 241µA	21mA to 57mA	7.87µA to 20.1µA	300µA	0.1µA to 1.07µA
Peak Current Efficiency (%)	99.63	99.54	99.93	99.95	99.99	98.6	99.8	99.06	99.5
ΔV_{out}, Ts, @ΔI_{LOAD}	63.75mV, 2.14µs @90µA	69.85mV, 3.56µs @90µA	257.1mV, 9.8µs @900µA	336.2mV, 8.5µs @9mA	112mV, 1.4ns @430mA	200mV, -- @1.17A	92mV, 83ns @2mA	200mV, 0.5µs @15mA ***	108mV, 1.24us @50µA
PSRR (dB)(@1KHz)	-37.72	-34.42	-33.78	-32.78	--	--	-32 @10KHz	--	-
Area (mm^2)	0.0036	0.0057	0.0162	0.0396	0.049	0.126	0.0056	0.0875	0.0374
FOM_T	2.4ns	3.35ns	78.5ps	4.6ps	0.00041ps	--	66.8ps	--	1.38ps

† Manual *Synthesized
†† LDO-M and LDO-A are manual and synthesizable versions of the same LDO design
** Digital Synthesizable LDO (using digital standard cells only) with Analog Pass Transistor
$FOM_T = (C_{LOAD} * \Delta V_{out} * I_Q)/(\Delta I_{LOAD} * I_{LOAD}$ Max.) (Lower is better.)
*** Estimated from plot in paper

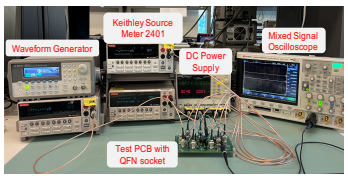
(a)



(b)

Die Micrograph

(c)



(d)

Fig. 4.20. (a) Comparison of synthesized analog LDOs with prior state-of-art synthesizable LDOs. (b) Runtime breakdown for the generation of the LDO (c) Die photo of the Synthesized Analog LDOs (d) Photograph of the testing bench setup and PCB.

The APR of unit cells in the synthesized version splits large transistors into many small unit cells and spatially distributes them over a large area, reducing variability due to gradients and spatially correlated randomness by averaging those affects across multiple distributed copies. The LDO-C, which has the largest area, reduces variability by up to 64.67%. We note that if correlation distance is reduced due to either the use of smaller transistors or larger discretization steps in unit cell sizing, the trend in variability can be expected to reverse in comparison to CC strategy.

Fig. 4.20 shows the performance comparison of the synthesizable analog LDOs with other state-of-art works. When comparing the synthesized analog LDO with the manually drawn version, key performance parameters like the current efficiency, the transient response, and the Figure-of-Merit show minimal deviation. In addition, the synthesized version allows for reduced input offset variability. As seen in Fig. 4.20, the synthesized LDOs achieve up to 99.95% peak current efficiency, a PSRR of -34.42dB, and a Figure-of-Merit of 4.6ps, which is comparable to other state-of-art LDOs.

4.4. Conclusion

In this chapter, we showed MemGen, a framework for the autonomous generation of the SRAMs across a wide range of design space that other compilers do not cover. It enables device-circuit-architecture co-design of memories. The framework is also open-source and PDK agnostic. It is highly modular, versatile, and easily augmentable by users to include more circuits in its component library and fit their requirements. We verified MemGen in both 65nm bulk planar CMOS and 12nm Fin FET technologies by fabricating several memories.

In this chapter, we proposed a digital flow-based approach to designing all-analog circuits that dramatically speeds the design and layout process while retaining the benefits of true analog topologies, and demonstrated its performance for three low dropout (LDO) regulators. Measurement results showed minimal loss in performance between manually generated LDO and its synthesized counterpart and showed up to 64.67% reduction in input offset. Using the synthesizable analog unit-cell based approach allowed us to significantly cut back on manual layout and verification efforts and improve turn-around-time and design scalability, pointing to an analog design approach in which components can be automatically optimized and implemented for each instance based on the precise context.

5. Conclusion

5.1. Summary of Contributions

5.1.1. SRAM Dynamic VMIN Modelling

- To evaluate the read and write access failure probability and the corresponding VMIN, we present fast analytical models which investigate key SRAM components and analytically models their behavior. This model takes into account several variables, such as the supply voltage, temperature, process variations and, array design variables i.e. bit-cell sizing, read current, bit-line capacitance (number of rows), word-line rise time (number of columns), sense amplifier strobe timing, bit-line leakage, and sense amplifier offset voltage. Simulations in a commercial bulk 65nm technology showed that the proposed method is able to evaluate the failure probability within a few seconds (~15 sec) with small error. This is up to 100,000X faster than previous methods, thereby helping to cut back on design time and verification. This analysis also provides insightful results about the effect of operating frequency and sense-amplifier strobe timing on read access failure probability.
- Thus, the proposed access models can be very useful for SRAM designers to quickly calculate design feasibility and analyze the design space to optimize power, area, and speed.

5.1.2. Ultra-Low Power SRAM Design

- A new Static Random-Access Memory (SRAM) bit-cell is proposed that leverages DLS inverters to form the cross-coupled inverter pair to reduce leakage power.
- An analysis of the operation of the DLS bit-cell is performed, and it is shown that the low intrinsic drive strength of the DLS bit-cell in combination with traditional peripheral circuits leaves it susceptible to data hold errors and read errors.
- A Word Line (WL) overdrive technique is proposed to reduce the leakage of the bit-cell access transistors to prevent data hold errors. To create the overdrive voltage, a DLS- based level converter circuit is designed that boosts the WL select signal to a voltage above the bit-cell V_{dd}. The level converter consumes very little power so as to not mitigate the power savings of the DLS bit-cell. The level converter is contributed by Daniel Truesdell.

- A full 2KB SRAM macro is designed for the DLS bit-cell and fabricated in 65-nm CMOS. Measurement results for the DLS SRAM chip show that the bit-cell achieves a leakage of 614aW, and the full macro consumes less than 200pW and is operable from 0°C to 60°C.
- A 6KB version of DLS SRAM is implemented into an SoC. Auto-sleep function is added to reduce power during stand-by. The SRAM is fully functional across 0.3V to 0.9V with 617.65pW at 0.3V and 10.1nW at 0.9V (up to 10X lower than previous state-of-art), resulting in an effective leakage/bit of 12.56fW and 205.6fW respectively, thereby making it suitable for energy constrained IoT applications.
- An SDLS Logic style SRAM is implemented to trade-off slightly higher leakage to achieve higher performance. A new low-leakage SDLS bit-cell is created for the memory. Simulations show that the macro achieves up to 10MHz operating frequency with 23.2nW leakage at 0.7V.
- A new DLS based implementation of the level converter allows to maintain nW budget and have high MHz range performance. The level converter is contributed by Nugaira Gahan Mim.

5.1.3. Mixed-Signal and Analog Circuit Design Automation

- A synthesizable approach to creating analog and mixed-signal circuits is proposed.
- This approach dramatically speeds the design and layout process to significantly cut back on manual layout and verification efforts and improve turn-around-time and design scalability, pointing to an analog design approach in which components can be automatically optimized and implemented for each instance based on the precise context.
- This approach is applied to SRAMs and three low dropout (LDO) regulators to serve as proof of concept.
- We show MemGen, a framework for the autonomous generation of the SRAMs across a wide range of design space that other compilers do not cover. It enables device-circuit-architecture co-design of memories. The framework is also open-source and PDK agnostic. It is highly modular, versatile, and easily augmentable by users to include more circuits in its component library and fit their requirements. We verify MemGen in both 65nm bulk planar CMOS and 12nm FinFET technologies by fabricating several memories. Assist circuits that allow to improve noise margins and performance of SRAMs are also implemented. The main framework of MemGen was developed and implemented by Sumanth Kamineni. The following are the contributions of my work:

    - Creation of SRAM aux-cell schematics and layouts for both version 1 and version 2 of MemGen for both 65nm and 12nm technologies.
    - Generation .LIB and .LEF files for aux-cells.
    - Simulating and verifying SRAM Control Blocks and top-level SRAM functional verification.
    - Design and verification of read (WLUD) and write assist (NBL) aux-cell (schematic and layouts).
- We also demonstrate the synthesizable approach to three low dropout (LDO) regulators. Template-based simulations allowed to cut back on design and verification time. Measurement results showed minimal loss in performance between manually generated LDO and its synthesized counterpart and showed up to 64.67% reduction in input offset.
- Using the synthesizable analog unit-cell based approach allowed us to significantly cut back on manual layout and verification efforts and improve turn-around-time and design scalability, pointing to an analog design approach in which components can be automatically optimized and implemented for each instance based on the precise context.

5.2. Future Work

The work in chapter three discussed the design and implementation of the SDLS Logic based ultra-low power SRAM design. This work presented measurement results for the DLS SRAMs and simulation results for the SDLS SRAM. For future work, the SDLS SRAM can be prototyped and tested to verify it using measurement results. The SDLS memory's performance is limited by the write driver's ability to function at higher voltages due to its simplistic implementation. In a future iteration of the work, the write driver can be improved to realize even higher performance gains This would allow the SRAM to expand the design space in which it can be used. For example, improving the performance of the write driver could potentially allow to push the frequency of the SDLS SRAM from sub-10MHz to more than 100MHz while simultaneously achieving low power, thereby widening the gamut of applications that could benefit from this improvement. All DLS and SDLS memories shown in this work are implemented without assist circuits. It would be very beneficial to improve performance even further for all such memories using assist circuits, especially during write operation since DLS based bit-

cells are limited by their ability to write in a fast manner.

The work in chapter four discussed the generation of the synthesizable analog and mixed-signal circuits. The memory generator MemGen supports only the 6T bit-cell. It would be interesting to see dual port bit-cell also be supported such as the conventional 8T bit-cell. In addition, application specific circuits such as the DLS and SDLS based bit-cells could also be added as a capability to the tool. Currently, the output of the APR process for the LDO analog circuit is a randomized placement of aux-cells. It would be beneficial to automate different arrangements of aux-cells for different sizes of devices to decrease the parasitics and improve the performance of the circuits even further.

5.3. Publications

5.3.1. Published Works

[1]. **S. Gupta**, B.H. Calhoun, "Dynamic Read VMIN and Yield Estimation for Nanoscale SRAMs," *IEEE Transactions on Circuits and Systems I: Regular Papers,* vol. 68, no. 3, pp. 1171–1182, *March* 2021.

[2]. **S. Gupta**, B.H. Calhoun, "Dynamic Write VMIN and Yield Estimation for Nanoscale SRAMs," *IEEE Transactions on Circuits and Systems I: Regular Papers,* (Submitted).

[3]. **S. Gupta,** D. S. Truesdell, B.H. Calhoun, "A 65nm 16kb SRAM with 131.5pW Leakage at 0.9V for Wireless IoT Sensor Nodes," *2020 IEEE Symposium on VLSI Circuits, Honolulu, HI, 2020*.

[4]. D. S. Truesdell, X. Liu, J. Breiholz, **S. Gupta**, S. Li**,** B.H. Calhoun, "NanoWattch: A Self-Powered 3-nW RISC-V SoC Operable from 160mV Photovoltaic Input with Integrated Temperature Sensing and Adaptive Performance Scaling," *2022 IEEE Symposium on VLSI Circuits, Honolulu, HI, 2022*.

[5]. T. Ajayi, Y. K. Cherivirala, K. Kwon, S. Kamineni, M. Saligane, M. Fayazi, **S. Gupta**, C.-H. Chen, D. Sylvester, D. Blaauw, R. Dreslinski Jr, B. Calhoun, D. D. Wentzloff, *"*Fully Autonomous Mixed Signal SoC Design & Layout Generation Platform*," 2020 IEEE Hot Chips 32 Symposium (HCS)*, August 2020.

[6]. T. Ajayi, S. Kamineni, Y. K. Cherivirala, M. Fayazi, K. Kwon, M. Saligane, **S. Gupta**, C. Chen, D. Sylvester, D., R. Dreslinski Jr, B. Calhoun, D. Wentzloff, "An Open-source Framework for Autonomous SoC Design with Analog Block Generation," *2020 IFIP/IEEE 28th International Conference on Very Large Scale Integration (VLSI-SoC),* pp. 141-146, Salt Lake City, USA, 2020.

[7]. S. Kamineni**, S. Gupta**, B.H. Calhoun, "MemGen: An Open-Source Framework for Autonomous Generation of Memory Macros," *IEEE Custom Integrated Circuits Conference (CICC), April* 2021.

5.3.2. Planned Works

1. A journal paper on the DLS SRAM 6KB version (2023).
2. A paper on the SDLS Logic-based SRAM (2023).
3. A paper on the Synthesized Analog LDO (2023).

References

[1] P. Newswire and Verified Market Research, “Internet of things (iot) market worth $1319.08 billion, globally, by 2026 at 25.68% cagr: Verified market research.”

[2] Everactive, “The battery problem: An infographic.”

[3] I. Insights, “Mcus sales to reach record-high annual revenues through 2022.”

[4] I. Insights, “Microcontrollers will regain growth after 2019 slump.”

[5] S. Kim, R. Vyas, J. Bito, K. Niotaki, A. Collado, A. Georgiadis, and M. M. Tentzeris, “Ambient rf energy-harvesting technologies for self-sustainable standalone wireless sensor platforms,” Proceedings of the IEEE, vol. 102, no. 11, pp. 1649–1666, 2014.

[6] X. Liu, H. Gao, J. E. Ward, X. Liu, B. Yin, T. Fu, J. Chen, D. R. Lovley, and J. Yao, “Power generation from ambient humidity using protein nanowires,” Nature, vol. 578, pp. 550–554, Feb 2020.

[7] P. M. Thibado, P. Kumar, S. Singh, M. Ruiz-Garcia, A. Lasanta, and L. L. Bonilla, “Fluctuation-induced current from freestanding graphene,” Phys. Rev. E, vol. 102, p. 042101, Oct 2020.

[8] Singh J, Pradhan DK, Mohanty SP (2013) Robust SRAM designs and analysis. Springer, New York, pp. 137–154.

[9] Mohamed H Abu-Rahma and Mohab Anis, Nanometer Variation-Tolerant SRAM, Springer, pp. 97-117, 2013.

[10] K. Cho, J. Park, T. W. Oh, and S.-O. Jung, “One-sided Schmitt-Trigger- based 9T SRAM cell for near-threshold operation,” IEEE Trans. Circuits Syst. I, Reg. Papers, vol. 67, no. 5, pp. 1551–1561, May 2020.

[11] Y.-C. Chien and J.-S. Wang, “A 0.2 V 32-kb 10T SRAM with 41 nW standby power for IoT applications,” IEEE Trans. Circuits Syst. I, Reg. Papers, vol. 65, no. 8, pp. 2443–2454, Aug. 2018.

[12] K. Shin, W. Choi, and J. Park, “Half-select free and bit-line sharing 9T SRAM for reliable supply voltage scaling,” IEEE Trans. Circuits Syst. I, Reg. Papers, vol. 64, no. 8, pp. 2036–2048, Aug. 2017.

[13] B. H. Calhoun and A. P. Chandrakasan, “Static noise margin variation for sub-threshold SRAM in 65-nm CMOS,” IEEE J. Solid-State Circuits, vol. 41, no. 7, pp. 1673–1679, Jul. 2006.

[14] A. Sheikholeslami, “Process variation and Pelgrom’s law,” IEEE Solid- State Circuits Mag., vol. 7, no. 1, pp. 8–9, Feb. 2015.

[15] L. Dolecek, M. Qazi, D. Shah, and A. Chandrakasan, “Breaking the simulation barrier: SRAM evaluation through norm minimization,” in Proc. IEEE/ACM Int. Conf. Comput.-Aided Design, Nov. 2008, pp. 322–329.

[16] R. Saeidi, M. Sharifkhani, and K. Hajsadeghi, “Statistical analysis of read static noise margin for near/sub-threshold SRAM cell,” IEEE Trans. Circuits Syst. I, Reg. Papers, vol. 61, no. 12, pp. 3386–3393, Dec. 2014.

[17] H. Makino et al., “Reexamination of SRAM cell write margin definitions in view of predicting the distribution,” IEEE Trans. Circuits Syst. II, Exp. Briefs, vol. 58, no. 4, pp. 230–234, Apr. 2011.

[18] K. Agarwal and S. Nassif, “Statistical analysis of SRAM cell stability,” in Proc. 43rd Annu. Conf. Design Autom. (DAC), 2006, pp. 57–62.

[19] J. Boley, V. Chandra, R. Aitken, and B. Calhoun, “Leveraging sensi- tivity analysis for fast, accurate estimation of SRAM dynamic write VMIN,” in Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE), 2013, pp. 1819–1824.

[20] S. Mukhopadhyay, H. Mahmoodi, and K. Roy, “Modeling of failure probability and statistical design of SRAM array for yield enhancement in nanoscaled CMOS,” IEEE Trans. Comput.-Aided Design Integr. Circuits Syst., vol. 24, no. 12, pp. 1859–1880, Dec. 2005.

[21] T. Date, S. Hagiwara, K. Masu, and T. Sato, “Robust importance sampling for efficient SRAM yield analysis,” in Proc. 11th Int. Symp. Qual. Electron. Design (ISQED), Mar. 2010, pp. 15–21.

[22] R. Kanj, R. Joshi, and S. Nassif, “Mixture importance sampling and its application to the analysis of SRAM designs in the presence of rare failure events,” in Proc. 43rd ACM/IEEE Design Autom. Conf., 2006, pp. 69–72.

[23] D. Khalil, M. Khellah, N.-S. Kim, Y. Ismail, T. Karnik, and V. K. De, “Accurate estimation of SRAM dynamic stability,” IEEE Trans. Very Large Scale Integr. (VLSI) Syst., vol. 16, no. 12, pp. 1639–1647, Dec. 2008.

[24] A. Singhee and R. A. Rutenbar, “Statistical blockade: A novel method for very fast Monte Carlo simulation of rare circuit events, and its application,” in Proc. Design, Autom., Test Eur., Apr. 2008, pp. 235–251.

[25] J. Wang, A. Singhee, R. A. Rutenbar, and B. H. Calhoun, “Two fast methods for estimating the minimum standby supply voltage for large SRAMs,” IEEE Trans. Comput.-Aided Design Integr. Circuits Syst., vol. 29, no. 12, pp. 1908–1920, Dec. 2010.

[26] J. Wang, A. Singhee, R. A. Rutenbar, and B. H. Calhoun, “Statis- tical modeling for the minimum standby supply voltage of a full SRAM array,” in Proc. 33rd Eur. Solid-State Circuits Conf. (ESSCIRC), Sep. 2007, pp. 400–403.

[27] M. H. Abu-Rahma, K. Chowdhury, J. Wang, Z. Chen, S. S. Yoon, and M. Anis, “A methodology for statistical estimation of read access yield in SRAMs,” in Proc. 45th Annu. Conf. Design Autom. DAC, Anaheim, CA, USA, 2008, pp. 205–210.

[28] J. P. Kulkarni and K. Roy, “Ultralow-voltage process-variation-tolerant Schmitt-Trigger-based SRAM design,” IEEE Trans. Very Large Scale Integr. (VLSI) Syst., vol. 20, no. 2, pp. 319–332, Feb. 2012.

[29] S. Gupta, K. Gupta, and N. Pandey, “Pentavariate VMIN analysis of a subthreshold 10T SRAM bit cell with variation tolerant write and divided bit-line read,” in IEEE Trans. Circuits Syst. I, Reg. Papers, vol. 65, no. 10, pp. 3326–3337, Oct. 2018.

[30] M. H. Abu-Rahma and M. Anis, “A statistical design-oriented delay vari- ation model accounting for within-die variations,” IEEE Trans. Comput.- Aided Design Integr. Circuits Syst., vol. 27, no. 11, pp. 1983–1995, Nov. 2008.

[31] M. J. M. Pelgrom, A. C. J. Duinmaijer, and A. P. G. Welbers, “Matching properties of MOS transistors,” IEEE J. Solid-State Circuits, vol. 24, no. 5, pp. 1433–1440, Oct. 1989.

[32] V. Wang, K. Agarwal, S. R. Nassif, K. J. Nowka, and D. Markovic, “A simplified design model for random process variability,” IEEE Trans. Semicond. Manuf., vol. 22, no. 1, pp. 12–21, Feb. 2009.

[33] C. Couso et al., “Dependence of MOSFETs threshold voltage variability on channel dimensions,” in Proc. Joint Int. Workshop Int. Conf. Ultimate Integr. Silicon (EUROSOI-ULIS), Apr. 2017, pp. 87–90.

[34] 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 Proc. Design, Autom. Test Eur., Munich, Germany, vol. 2, 2005, pp. 926–931.

[35] B. S. Amrutur and M. A. Horowitz, “A replica technique for wordline and sense control in low-power SRAM’s,” IEEE J. Solid-State Circuits, vol. 33, no. 8, pp. 1208–1219, Aug. 1998.

[36] C. C. Craig, “On the frequency function of xy,” Ann. Math. Statist., vol. 7, no. 1, pp. 1–15, Mar. 1936.

[37] A. Oliveira and A. Seijas-Macias, “An approach to distribution of the product of two normal variables,” Discussiones Mathematicae Probab. Statist., vol. 32, nos. 1–2, p. 87, 2012.

[38] D. Burnett, S. Parihar, H. Ramamurthy, and S. Balasubramanian, “FinFET SRAM design challenges,” in Proc. IEEE Int. Conf. Design Technol., Austin, TX, USA, May 2014, pp. 1–4.

[39] A. Asenov, “Simulation of statistical variability in nano MOSFETs,” in Proc. IEEE Symp. VLSI, Jun. 2007, pp. 86–87.

[40] H. Nam and C. Shin, “Study of high-k /metal-gate work function variation in FinFET: The modified RGG concept,” IEEE Electron Device Lett., vol. 34, no. 12, pp. 1560–1562, Dec. 2013.

[41] T. Matsukawa et al., “Comprehensive analysis of variability sources of FinFET characteristics,” in Proc. Symp. VLSI Technol., Honolulu, HI, USA, 2009, pp. 118–119.

[42] H. F. Dadgour, K. Endo, V. K. De, and K. Banerjee, “Grain-orientation induced work function variation in nanoscale metal-gate transistors— Part I: Modeling, analysis, and experimental validation,” IEEE Trans. Electron Devices, vol. 57, no. 10, pp. 2504–2514, Oct. 2010.

[43] B. Zimmer et al., “SRAM assist techniques for operation in a wide voltage range in 28-nm CMOS,” IEEE Trans. Circuits Syst. II, Exp. Briefs, vol. 59, no. 12, pp. 853–857, Dec. 2012.

[44] S. Gupta, K. Gupta, and N. Pandey, “A 32-nm subthreshold 7T SRAM bit cell with read assist,” IEEE Trans. Very Large Scale Integr. (VLSI) Syst., vol. 25, no. 12, pp. 3473–3483, Dec. 2017.

[45] S. Gupta, K. Gupta, B. H. Calhoun, and N. Pandey, “Low-power near- threshold 10T SRAM bit cells with enhanced data-independent read port leakage for array augmentation in 32-nm CMOS,” IEEE Trans. Circuits Syst. I, Reg. Papers, vol. 66, no. 3, pp. 978–988, Mar. 2019.

[46] R. Chattamvelli, “On the doubly noncentral f distribution,” Comput. Statist. Data Anal., vol. 20, no. 5, pp. 481–489, Nov. 1995.

[47] L. Lin, S. Jain, and M. Alioto, “Sub-nw microcontroller with dual-mode logic and self- startup for battery-indifferent sensor nodes,” IEEE Journal of Solid-State Circuits, pp. 1–1, 2020.

[48] X. Liu, H. Gao, J. E. Ward, X. Liu, B. Yin, T. Fu, J. Chen, D. R. Lovley, and J. Yao, “Power generation from ambient humidity using protein nanowires,” Nature, vol. 578, pp. 550–554, Feb 2020.

[49] S. Kim, R. Vyas, J. Bito, K. Niotaki, A. Collado, A. Georgiadis, and M. M. Tentzeris, “Ambient rf energy-harvesting technologies for self-sustainable standalone wireless sensor platforms,” Proceedings of the IEEE, vol. 102, no. 11, pp. 1649–1666, 2014.

[50] M. Piñuela, P.D. Mitcheson, and S.Lucyszyn, “Ambient rf energy harvesting in urban and semi-urban environments,” IEEE Transactions on Microwave Theory and Techniques, vol. 61, no. 7, pp. 2715–2726, 2013.

[51] S. Bandyopadhyay, P.P. Mercier, A.C. Lysaght, K.M. Stankovic, and A.P. Chandrakasan, “A 1.1 nw energy-harvesting system with 544 pw quiescent power for next-generation implants,” IEEE Journal of Solid-State Circuits, vol. 49, no. 12, pp. 2812–2824, 2014.

[52] B. J. Hansen, Y. Liu, R. Yang, and Z. L. Wang, “Hybrid nanogenerator for concurrently harvesting biomechanical and biochemical energy,” ACS Nano, vol. 4, no. 7, pp. 3647– 3652, 2010. PMID: 20507155.

[53] B. H. Calhoun and A.P. Chandrakasan, “Static noise margin variation for sub-threshold SRAM in 65-nm CMOS,” IEEE Journal of Solid-State Circuits, vol. 41, no. 7, pp. 1673–1679, July 2006.

[54] Ali Sheikholeslami, “Process Variation and Pelgrom’s Law,” IEEE Solid- State Circuits Magazine, vol. 7, no. 1, pp. 8–9, Feb. 2015.

[55] Y. Fujii et al., "Soft error free, low power and low cost superSRAM with 0.98 /spl mu/m/sup 2/ cell by utilizing existing 0.15 /spl mu/m-DRAM process," Digest of Technical Papers. 2004 Symposium on VLSI Technology, 2004., 2004, pp. 232-233.

[56] T. Fukuda et al., "13.4 A 7ns-access-time 25μW/MHz 128kb SRAM for low-power fast wake-up MCU in 65nm CMOS with 27fA/b retention current," 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2014, pp. 236-237.

[57] Y. Yamamoto et al., "Ultralow-voltage operation of Silicon-on-Thin-BOX (SOTB) 2Mbit SRAM down to 0.37 V utilizing adaptive back bias," 2013 Symposium on VLSI Circuits, 2013, pp. T212-T213.

[58] Y. Chien and J. Wang, "A 0.2 V 32-Kb 10T SRAM With 41 nW Standby Power for IoT Applications," in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 65, no. 8, pp. 2443-2454, Aug. 2018.

[59] T. Haine, D. Flandre and D. Bol, "8-T ULV SRAM macro in 28nm FDSOI with 7.4 pW/bit retention power and back-biased-scalable speed/energy trade-off," 2018 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S), 2018, pp. 1-3.

[60] D. Kim, G. Chen, M. Fojtik, M. Seok, D. Blaauw and D. Sylvester, "A 1.85fW/bit ultra low leakage 10T SRAM with speed compensation scheme," 2011 IEEE International Symposium of Circuits and Systems (ISCAS), 2011, pp. 69-72.

[61] D. S. Silveira, A. Mativi, M. S. Porto and S. Bampi, "Energy Savings with Non-Volatile Memory System for High Definition Video Encoders," 2019 17th IEEE International New Circuits and Systems Conference (NEWCAS), 2019, pp. 1-4.

[62] S. Gupta, D. S. Truesdell and B. H. Calhoun, "A 65nm 16kb SRAM with 131.5pW Leakage at 0.9V for Wireless IoT Sensor Nodes," 2020 IEEE Symposium on VLSI Circuits, 2020, pp. 1-2.

[63] Y. Lee, Y. Kim, D. Yoon, D. Blaauw and D. Sylvester, "Circuit and system design guidelines for ultra-low power sensor nodes," DAC Design Automation Conference 2012, 2012, pp. 1037-1042.

[64] S. Gupta, K. Gupta and N. Pandey, "Pentavariate VMIN Analysis of a Subthreshold 10T SRAM Bit Cell With Variation Tolerant Write and Divided Bit-Line Read," in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 65, no. 10, pp. 3326-3337, Oct. 2018.

[65] D. Bol, R. Ambroise, D. Flandre and J. Legat, "Building Ultra-Low-Power Low-Frequency Digital Circuits with High-Speed Devices," 2007 14th IEEE International Conference on Electronics, Circuits and Systems, 2007, pp. 1404-1407.

[66] David Bol, Julien De Vos, Renaud Ambroise, Denis Flandre, Jean-Didier Legat, “Building ultra-low-power high-temperature digital circuits in standard high-performance SOI technology,” Solid-State Electronics, Volume 52, Issue 12, Pages 1939-1945, 2008.

[67] D. Bol, J. De Vos, D. Flandre and J. -. Legat, "Ultra-low-power high-noise-margin logic with undoped FD SOI devices," 2009 IEEE International SOI Conference, 2009, pp. 1-2.

[68] D. Levacq, V. Dessard and D. Flandre, "Low Leakage SOI CMOS Static Memory Cell With Ultra-Low Power Diode," in IEEE Journal of Solid-State Circuits, vol. 42, no. 3, pp. 689-702, March 2007.

[69] W. Lim, I. Lee, D. Sylvester and D. Blaauw, "8.2 Batteryless Sub-nW Cortex-M0+ processor with dynamic leakage-suppression logic," 2015 IEEE International Solid-State Circuits Conference - (ISSCC) Digest of Technical Papers, 2015, pp. 1-3.

[70] D. S. Truesdell, J. Breiholz, S. Kamineni, N. Liu, A. Magyar and B. H. Calhoun, "A 6–140-nW 11 Hz–8.2-kHz DVFS RISC-V Microprocessor Using Scalable Dynamic Leakage-Suppression Logic," in IEEE Solid-State Circuits Letters, vol. 2, no. 8, pp. 57-60, Aug. 2019

[71] K. Zhang et al., "SRAM design on 65nm CMOS technology with integrated leakage reduction scheme," 2004 Symposium on VLSI Circuits. Digest of Technical Papers 2004, pp. 294-295.

[72] S. Gupta, K. Gupta, B. H. Calhoun and N. Pandey, "Low-Power Near-Threshold 10T SRAM Bit Cells With Enhanced Data-Independent Read Port Leakage for Array Augmentation in 32-nm CMOS," in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 66, no. 3, pp. 978-988, March 2019.

[73] B. Wang, T. Q. Nguyen, A. T. Do, J. Zhou, M. Je, and T. T. H. Kim, “Design of an ultra-low voltage 9T SRAM with equalized bitline leakage and CAM-assisted energy efficiency improvement,” IEEE Trans. Circuits Syst. I, Reg. Papers, vol. 62, no. 2, pp. 441–448, Feb. 2015.

[74] S. Gupta and B. H. Calhoun, "Dynamic Read VMIN and Yield Estimation for Nanoscale SRAMs," in IEEE Transactions on Circuits and Systems I: Regular Papers, 2020.

[75] S. Gupta, K. Gupta and N. Pandey, "A 32-nm Subthreshold 7T SRAM Bit Cell With Read Assist," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 25, no. 12, pp. 3473-3483, Dec. 2017

[76] S. Gupta and B. H. Calhoun, "Dynamic Write VMIN and Yield Estimation for Nanoscale SRAMs," in IEEE Transactions on Circuits and Systems I: Regular Papers, 2021.

[77] H. Makino et al., “Reexamination of SRAM cell write margin definitions in view of predicting the distribution,” IEEE Trans. Circuits Syst. II, Exp. Briefs, vol. 58, no. 4, pp. 230–234, Apr. 2011.

[78] J. P. Kulkarni and K. Roy, “Ultralow-Voltage Process-Variation-Tolerant Schmitt-Trigger-Based SRAM Design,” IEEE Transactions on Very Large-Scale Integration (VLSI) Systems, vol. 20, no. 2, pp. 319–332, Feb. 2012.

[79] D. S. Truesdell and B. H. Calhoun, "A Single-Supply 6-Transistor Voltage Level Converter Design Reaching 8.18-fJ/Transition at 0.3–1.2-V Range or 44-fW Leakage at 0.8–2.5-V Range," in IEEE Solid-State Circuits Letters, vol. 3, pp. 502-505, 2020.

[80] T. Hirose, “A 20-ns 4-Mb CMOS SRAM with hierarchical word decoding architecture,” IEEE Journal of Solid-State Circuits, vol. 25, no. 5, p. 1068–1074, Oct. 1990.

[81] Y. Ishii et al., “A 5.92-Mb/mm2 28-nm pseudo 2-read/write dual- port SRAM using double pumping circuitry,” in Proc. A-SSCC, 2016, pp. 17–20.

[82] M.-H. Chang, Y.-T. Chiu, and W. Hwang, “Design and iso-area VMIN analysis of 9T subthreshold SRAM with bit-interleaving scheme in 65-nm CMOS,” IEEE Trans. Circuits Syst. II, Exp. Briefs, vol. 59, no. 7, pp. 429–433, Jul. 2012.

[83] Tuan Do et al., “0.2 V 8T SRAM With PVT-Aware Bitline Sensing and Column-Based Data Randomization”, in IEEE JSSC, pp. 1487-1498, 2016.

[84] Chang et al., “A Sub-0.3 V Area-Efficient L-Shaped 7T SRAM With Read Bitline Swing Expansion Schemes Based on Boosted Read-Bitline, Asymmetric-VTH Read-Port, and Offset Cell VDD Biasing Techniques”, in IEEE JSSC, pp 2558-2569, 2013.

[85] T. Haine, Q. Nguyen, F. Stas, L. Moreau, D. Flandre and D. Bol, "An 80-MHz 0.4V ULV SRAM macro in 28nm FDSOI achieving 28-fJ/bit access energy with a ULP bitcell and on-chip adaptive back bias generation," ESSCIRC 2017 - 43rd IEEE European Solid State Circuits Conference, 2017, pp. 312-315.

[86] Fujiwara et al., “A 20nm 0.6V 2.1μW/MHz 128kb SRAM with no half select issue by interleave wordline and hierarchical bitline scheme”, in IEEE VLSIT, 2013.

[87] A.T. Do, Z. Lee, B. Wang, I.-J. Chang, and T.T. Kim,“0.2V 8T SRAM with improved bitline sensing using column-based data randomization,” in Proc. A-SSCC, 2014, pp. 141–144.

[88] R. Boumchedda et al., "1.45-fJ/bit Access Two-Port SRAM Interfacing a Synchronous/Asynchronous IoT Platform for Energy-Efficient Normally Off Applications," in IEEE Solid-State Circuits Letters, vol. 1, no. 9, pp. 186-189, Sept. 2018.

[89] Lutkemeier et al., “A 65 nm 32 b Subthreshold Processor With 9T Multi-Vt SRAM and Adaptive Supply Voltage Control”, in IEEE JSSC, pp. 8-19, 2012.

[90] Mohammadi et al., “A 128kb Single-bitline 8.4fJ/bit 90MHz at 0.3V 7T Sense-Amplifierless SRAM in 28nm FD-SOI”, in IEEE ESSCIRC ,pp 429-432, 2016.

[91] Sinangil et al., “A Reconfigurable 8T Ultra-Dynamic Voltage Scalable (U-DVS) SRAM in 65 nm CMOS”, in IEEE JSSC, pp. 3163-3173, 2009.

[92] Y. Yamamoto et al., "Ultralow-voltage operation of Silicon-on-Thin-BOX (SOTB) 2Mbit SRAM down to 0.37 V utilizing adaptive back bias," 2013 Symposium on VLSI Circuits, 2013, pp. T212-T213.

[93] R. Ranica et al., "FDSOI process/design full solutions for ultra low leakage, high speed and low voltage SRAMs," 2013 Symposium on VLSI Circuits, 2013, pp. T210-T211.

[94] K. Osada, Y. Saitoh, E. Ibe and K. Ishibashi, "16.7-fA/cell tunnel-leakage-suppressed 16-Mb SRAM for handling cosmic-ray-induced multierrors," in IEEE Journal of Solid-State Circuits, vol. 38, no. 11, pp. 1952-1957, Nov. 2003.

[95] S. Hanson et al., "A Low-Voltage Processor for Sensing Applications With Picowatt Standby Mode," in IEEE Journal of Solid-State Circuits, vol. 44, no. 4, pp. 1145-1155, April 2009.

[96] M. Fojtik et al., "A Millimeter-Scale Energy-Autonomous Sensor System With Stacked Battery and Solar Cells," in IEEE Journal of Solid-State Circuits, vol. 48, no. 3, pp. 801-813, March 2013.

[97] J. Crossley, A. Puggelli, H. Le et al., “Bag: A designer-oriented in- tegrated framework for the development of ams circuit generators,” in 2013 IEEE/ACM International Conference on Computer-Aided Design (ICCAD), 2013, pp. 74–81.

[98] K. Kunal, M. Madhusudan, A. K. Sharma et al., “Align – open-source analog layout automation from the ground up,” in 2019 56th ACM/IEEE Design Automation Conference (DAC), 2019, pp. 1–4.

[99] M. Liu, X. Tang, K. Zhu, H. Chen, N. Sun and D. Z. Pan, "1- and 80-MS/s SAR ADCs in 40-nm CMOS With End-to-End Compilation," in *IEEE Solid-State Circuits Letters*, vol. 5, pp. 292-295, 2022.

[100] G. Kumar, B. Chatterjee, and S. Sen, “OpenSerDes: an open source process-portable all-digital serial link,” in 2021 Design, Automation Test in Europe Conference Exhibition (DATE), 2021.

[101] T. Ajayi, S. Kamineni, Y. K. Cherivirala et al., “An open-source frame-work for autonomous soc design with analog block generation,” in 2020 IFIP/IEEE 28th International Conference on Very Large Scale Integration (VLSI-SOC), 2020, pp. 141–146

[102] https://developer.arm.com/ip-products/physical-ip/embedded-memory.

[103] http://www.globalfoundries.com/technology-solutions/asics.

[104] https://www.synopsys.com/dw/ipdir.php?ds=dwc_sram_memory_compilers.

[105] K. Chakraborty, et. al, IEEE TVLSI, Vol. 9, No. 2, pp. 352-364, April 2001.

[106] M. Guthaus, et. al, IEEE ICCAD, Austin, TX, pp. 1-6, Nov 2016.

[107] S. Ataei et. al,” IEEE ASYNC), Hirosaki, Japan, pp. 1-8, 2019.

[108] N. Liu et. al., ISVLSI, PA, USA, 2016, pp. 535-540.

[109] J. Oh, J. -E. Park, Y. -H. Hwang and D. -K. Jeong, "25.2 A 480mA Output-Capacitor-Free Synthesizable Digital LDO Using CMP- Triggered Oscillator and Droop Detector with 99.99% Current Efficiency, 1.3ns Response Time, and 9.8A/mm2 Current Density," 2020 IEEE ISSCC, 2020, pp. 382-384

[110] S. Bang et al., "25.1 A Fully Synthesizable Distributed and Scalable All-Digital LDO in 10nm CMOS," 2020 IEEE ISSCC, 2020, pp. 380-382.

[111]J. Oh, J. -E. Park and D. -K. Jeong, "A Highly Synthesizable 0.5-to-1.0-V Digital Low-Dropout Regulator With Adaptive Clocking and Incremental Regulation Scheme," in IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 67, no. 10, pp. 2174-2178, Oct. 2020.

[112]A. Fahmy, J. Liu, P. Terdal, R. Madler, R. Bashirullah and N. Maghari, "A synthesizable time-based LDO using digital standard cells and analog pass transistor," ESSCIRC 2017 - 43rd IEEE European Solid State Circuits Conference, 2017, pp. 271-274.

[113]S. Kundu, M. Liu, R. Wong, S. -J. Wen and C. H. Kim, "A fully integrated 40pF output capacitor beat-frequency-quantizer-based digital LDO with built-in adaptive sampling and active voltage positioning," 2018 IEEE International Solid - State Circuits Conference - (ISSCC), 2018, pp. 308-310.

[114]S. Nalam, M. Bhargava, K. Ringgenberg, K. Mai and B. H. Calhoun, "A Technology-Agnostic Simulation Environment (TASE) for iterative custom IC design across processes," 2009 IEEE International Conference on Computer Design, 2009, pp. 523-528.

[115]S. Kamineni, S. Gupta and B. H. Calhoun, "MemGen: An Open-Source Framework for Autonomous Generation of Memory Macros," 2021 IEEE Custom Integrated Circuits Conference (CICC), 2021, pp. 1-2.

[116]W. Qu, S. Singh, Y. Lee, Y. -S. Son and G. -H. Cho, "Design-Oriented Analysis for Miller Compensation and Its Application to Multistage Amplifier Design," in IEEE Journal of Solid-State Circuits, vol. 52, no. 2, pp. 517-527, Feb. 2017.

[117]A. K. Sharma et al., "Common-Centroid Layouts for Analog Circuits: Advantages and Limitations," 2021 Design, Automation & Test in Europe Conference & Exhibition (DATE), 2021, pp. 1224-1229.