SRAM Power Optimization for Emerging Processes
Lucas Calderin, Greg LaCaille

Abstract—As CMOS devices scale, SRAM arrays are requiring higher operating voltages to compensate for increased leakage and device variation, resulting in increased power dissipation. New device structures have been proposed, aimed at reducing leakage relative to drive current, but processing capability to obtain high yields is still in development. We propose a method of relating the desired supply voltage to requirements on the physical variations within the device. From this we would obtain insight into the feasible limit of power consumption levels for different technologies based on fabrication capabilities.

I. INTRODUCTION

As the minimum feature size of transistors continues to shrink, the density of SRAM is expected to scale as well. If bulk CMOS devices are scaled with constant electric field, then the corresponding reduction in \( V_{dd} \) will decrease dynamic power dissipation. However, in deep sub-micron processes, further reductions in \( V_{th} \) would result in substantial leakage currents, especially when short channel effects increase the sub-threshold swing of the device. This practical limit on \( V_{th} \) has shifted scaling trends toward constant voltage scaling, which will not see the same reduction in dynamic power consumption. Additionally, the variation across devices has also begun to increase as technologies continue to scale. This variation may lead to an increase in \( V_{min} \), which is the required \( V_{dd} \) to maintain a reasonable yield [1]. As power consumption becomes an increasingly important design constraint, this trend of increasing supply voltage presents a potential road block in the scaling of SRAM.

Many circuit based strategies to reduce \( V_{min} \) and improve the performance of SRAM in deep sub-micron process have been investigated [1]. However, circuit design approaches add significant complexity, consume valuable area, and cannot address the fundamental issue of leakage current preventing \( V_{th} \) from scaling. As an alternative to bulk CMOS, other gated devices that would allow for reduced supply voltages are currently being explored. Currently, many manufacturers are moving toward fully depleted devices, such as Ultra Thin Body (UTB) Silicon-on-Insulator (SOI) and FinFETs, which offer improved leakage performance by reducing short channel effects and allow for more aggressive scaling [2]. Carbon Nanotube (CNT) FETs are an emerging technology which also show both an improvement in short channel effects as well as increased Ion due to ballistic carrier transport [3]. Finally, band to band tunnel FETs (TFETs) suffer from poor on current, but are capable of achieving sub-threshold swings far below the fundamental 60mV/dec limit inherent for most gated devices [4]. All of these devices would improve leakage performance, but the key to a successful implementation of them will be their susceptibility to process variations.

We plan to explore the potential of these four devices for future use in high density, low power SRAM. In this paper we first provide a description of the fundamental operation and tradeoffs associated with each of these alternative devices. This is followed by an evaluation of practical methods of simulating designs that incorporate them. Next, we discuss modern SRAM design, focusing on optimization of power consumption and area while maintaining suitable yields. Finally we discuss our plans for evaluating the manufacturing variability requirements of these devices for implementation in SRAM.

For the final project, we plan to follow this paper with evaluations of these different devices in both 6T and 8T configurations. Since these devices are all in early stages of development, a wide range of device structure and performance exist. Trying to estimate the exact trends for the future of these devices would prove infeasible and produce little value. However, it may be much easier to estimate change in performance for a typical device based on the deviation of few key features such as feature size and doping. Using these estimates of performance dependence on manufacturing parameters we plan to accurately infer the manufacturing requirements to produce a reasonable SRAM yield, while only requiring rough estimates of the \( I_d \) vs \( V_{gs} \) curves for the devices.

II. ALTERNATIVE DEVICES

All of the devices analyzed in this work are planar devices consisting of a semiconducting channel controlled by a capacitive gate structure. The voltage applied to the gate controls the carrier distribution by altering the potential between the Fermi level in the channel and the intrinsic Fermi level of the substrate. The initial value of this potential can be controlled via the gate work function, allowing \( V_{th} \) to be tuned to a desired value with out significantly altering the \( I_d \) vs \( (V_{gs} - V_{th}) \) curve [5]. The tuning of \( V_{th} \) impacts the leakage power of the circuit by changing the ratio of \( I_{on} \) to \( I_{off} \) for a given \( V_{dd} \). The value of the sub-threshold swing for a device will affect the achievable values \( I_{on}/I_{off} \). All of the following devices attempt to improve the \( I_{on}/I_{off} \) ratio relative to bulk CMOS.

A. Fully Depleted CMOS

Fully-Depleted (FD) devices, like FinFETs and UTBSOI are currently the main choices for next-generation logic and SRAM technology in leading-edge nodes. They are able to achieve good \( I_{on}/I_{off} \) ratios by having subthreshold swings only slightly higher than the theoretical limit of 60mV/decade. This is possible due to the gate having substantial control over the entire channel, rather than just the surface [2]. To enable this, UTBSOI devices rely on an extremely thin doped substrate, where the depletion region extends the entire way to the insulating block beneath. Alternatively, FinFET and multi-gate devices use gate contacts which surround multiple sides of the channel, causing the \( C_{gate-chan} \) to be dominant over \( C_{drain-chan} \).
improving gate control and thereby increasing subthreshold swing. Furthermore, because the channel is fully depleted, the body is essentially floating and can assume any voltage. This eliminates the body effect, which slows conventional MOSFET devices down and can cause systematic $V_{th}$ variation [6].

Another major advantage that Fully Depleted devices have over traditional MOSFETs is their use of an intrinsic body. In current MOSFETs, the substrate is highly doped in order to reduce the depletion region width. It is ideal for channels to have a very thin depletion region, for this permits the gate to have good control across the entire channel. Depletion region shrinking done using a highly-doped body increases subthreshold swing and reduces Drain-Induced Barrier Lowering (DIBL). Unfortunately, body doping also affects the $V_{th}$ of the devices. In even the most precise fabrication process, there will be significant variation in doping profiles and concentrations throughout the silicon. This leads to large $V_{th}$ variations for MOSFETs within a die [2].

In Fully-Depleted devices, a thin depletion region is instead created by limiting body depth. The extremely shallow (~1/4 channel length) bodies of Fully-Depleted devices restricts the thickness of the depletion region, rendering strong doping unnecessary. Using an undoped silicon body, threshold mismatch through substrate doping fluctuation is nearly eliminated. Additionally, the mobility of intrinsic silicon is orders of magnitude higher than doped silicon, allowing higher current density. Moreover, due to the uniformity of the depletion region over the entire channel, $V_{th}$ roll-off is minimized, even for sub 10nm lengths. Unfortunately, with an undoped channel, one needs to change the work-function of the metal gate to change the threshold voltage of a FD device. This process is different from traditional doping techniques and introduces significant $V_{th}$ variation [2].

One disadvantage of FD devices is the extra parasitic capacitances associated with the structures. The incredibly thin body of UTBSOI devices requires shallow drain and source contacts, but this also increases contact resistance considerably. To reduce the resistance, the source and drain rise high above the body, similar to the gates of modern devices, creating large capacitive coupling between the contacts and the gate. Similar mechanisms appear in the FinFET, with capacitive coupling between the fin and the source/drain contacts [2].

B. Carbon-Nanotube FETs

CNTFETs are promising devices which transport carriers though single-walled or multi-walled carbon tubes, rather than silicon. They are created by embedding nanotubes in intrinsic silicon under a high-k gate oxide, and with either ohmic contacts or Schottky Barriers (SB) at the drain and source. It has been shown that the mobilities of electrons and holes within CNTs are nearly equal and are up to an order of magnitude higher than state of the art N and P type doped silicon. This is cause for CNTFETs’ excellent intrinsic delay ($C/V$) figures, enabling faster operation for the same dynamic power. While both ohmic contact and Schottky Barrier devices can be fabricated, we will limit our investigations to ohmic contact CNTFETs due to their MOSFET-like properties and superior $I_{on}$ as compared to SB CNTs [3]. Theoretically, the FO4 delay of CNTFET inverters can be up to 5x lower than CMOS for the 32nm process node with a 0.9V supply. This cannot be achieved with a single CNT, but 8 in parallel, since the capacitance load per CNT is very small (5-10x lower than MOS gate to channel and fringe capacitances), while the drain and source contact resistances dominate the $R_{on}$ [7]. Additionally, as technology nodes continue to scale, CNTFETs with 5nm CNT pitch sustain a 5x FO4 improvement over CMOS into the 10nm region and beyond [3].

Unfortunately, CNTFETs require advances in manufacturing technology before they can truly compete with other leading-edge technologies. The FO4 delay is a strong function of CNT pitch, and currently 5nm pitch is far from being available. Additionally, with current manufacturing technology, 8-30% of the CNTFETs created on a wafer are metallic, rather than semiconducting, causing “always-on” paths in the gate. Such tubes can be destroyed on a chip using electrical burning, but even if this process is performed perfectly, there will still be $I_{on}$ variation due to the random number of metallic tubes present in the pull-up and pull-down networks of gates. Additionally, CNTs can be misaligned while being grown, causing shifts in the logical operation of a gate, as well as creating “always-on” paths. Work has been done to create layouts which are immune to these imperfections at a reasonable (10-20%) energy/area cost [8].

C. Tunneling FETs

Unlike conventional MOSFETs, where an inverted channel region is formed to allow conduction, TFETs operate by creating a large enough potential between the source and channel that carriers can tunnel from the conduction band of one region to the valence band of the other. Figure 1 from [9] shows a band diagram for a potential device. The major advantage of TFETs is that they are capable of obtaining sub-threshold swings lower than the 60 mv/dec fundamental limit seen in channel inversion based devices, which allows for large $I_{on}/I_{off}$ ratios at very low supply voltages. Although the off state current for TFETs is very good, $I_{on}$ tends to be lower than in a similarly sized MOSFET. Because of this trade-off TFETs are expected to outperform MOSFETs in speeds <0.5GHz with higher speed operation still in need of advancement [4].

![Figure 1: TFET Band Diagram (a) off state, (b) on state](image)
The use of TFETs in SRAM has been analyzed in several studies. Because the source and drain of TFETs are inherently different, they cannot act as pass transistors restricting them from use in 6T SRAM designs. Techniques for TFET SRAM cells include standard 8T, 7T cells that use an active low wordline tied to the source of the read buffer transistor, and 6T hybrid SRAM cells where the access devices are conventional MOSFETs [10].

III. SIMULATION TECHNIQUES

Several options for simulating circuits with experimental devices exist. Each of these methods has varying degrees of physical and empirical modeling, which typically leads to tradeoffs between accuracy and computing time. Additionally, flexibility will be a factor when trying to benchmark distinctly different technologies against each other.

TCAD simulators, like MEDICI, can very accurately estimate device performance due to the physical laws governing its operation. They have the ability to simulate multi-device structures, such as inverters or other simple logic blocks, but due to the high fidelity of the results, simulation time grows very quickly with the number of elements. Additionally, large levels of detail such as doping profiles are required to produce meaningful results.

SPICE is used extensively as a middle ground between device simulators and behavioral models, with many models available online. While SPICE does offer a good tradeoff between accuracy and computing time, it requires a high level of expertise to adapt the models, making it not ideal for a predictive study that covers several technologies.

VERILOG-A enables the simulation of large scale circuits efficiently by using behavioral models for support circuitry, and lookup tables for the alternative device blocks. It provides very short simulation times, but lacks any kind of connection to the original device physics. The accuracy of the simulation is completely dependent on the efficacy of the simulated tables.

To use any of these tools for accurate prediction of device parameters would require significant time and high levels of expertise in device modeling. By relaxing the requirements on accuracy of absolute performance values and instead focusing on deviations in performance we can use a TCAD simulator to estimate the impact of physical parameters on device performance. We can then use existing spice models in conjunction with scaling approximations for estimates of $I_d$ curves for Verilog models.

IV. SRAM DESIGN

Modern SRAM designs must optimize performance while remaining aware of the impacts of process variation on cell functionality. For an SRAM cell to function properly Read, Write, and Access margins must be met over a wide value of process variation. The strength of the NMOS pull down device ($N_{pd}$), NMOS access device ($N_{ax}$), and the PMOS pull up device ($P_{pu}$) all affect these margins.

During the read $N_{pd}$ must be strong enough prevent $N_{ax}$ from raising the voltage at the node containing a 0 above the transition point of the inverter and $P_{pu}$ must be strong enough to keep the inverter transition point at a reasonable level. This read stability benefits from the fact that there is a finite time that the access transistor is on. The combined strength of $N_{ax}$ and $N_{pd}$ must also be large enough to produce the required voltage difference at the sense amp within the allotted time. During the write $N_{ax}$ must be strong enough to overpower $P_{pu}$ and force a 0 into the cell. This write ability suffers from the fact that there is a finite time that the access transistor is on.

Random dopant fluctuations can cause significant changes in $V_{th}$ of two devices in close proximity to each other, which will alter the desired drive strength of the devices, changing the margins of the cell. The random variation in margins of operation from cell to cell requires that the entire array be operated so that only does the average cell have functional margins but that cells with performance several standard deviations out will also have enough margin to function [11]. Many methods for evaluating these margins and effects of process variation on them have been proposed. In [11] an analytical approach to calculate the voltage at the storage node and the transition voltage of the inverters is used in conjunction with probability distribution functions to predict the failure probability of a cell. The results in this work agree well with Monte Carlo simulations of the same cells.

As devices scale and process variation increases the standard deviation of the margins, SRAM yield can be improved through circuit techniques that either reduce the failure probability of cells or make the array more tolerant of single cell failures.

Circuit techniques to reduce the failure probability of a single SRAM cell focus on breaking the tradeoffs associated with Read Margins and Write Margins. The main failure point of an SRAM cell stems from the fact that $N_{ax}$ is used for both reading and writing, and the sizing required for good a Read Margin reduces the Write Margin. One way to break this tradeoff is to have separate wordlines and bitlines for reading and writing. This then requires two extra transistors to control the read bitline, leading to the 8-T SRAM cell [12]. With reads and writes not linked, the write transistors can be strong, reducing write time and increasing Write Margin, without disrupting the Read Margin.

There are still other circuit design options to separate read and write margin tradeoffs by creating different conditions within the SRAM cell for the read and write cycles. One technique is multi-voltage SRAM, which dynamically changes SRAM cell voltages (wordlines, bitlines, inverter supplies) for optimal performance, depending on the operation being executed. This approach can be implemented in a number of ways, both with and without an extra supply [12]. On the array level, strategies such as error-correcting codes, column redundancy, or selling economy versions of chips with a smaller fraction of the total array functional can be can increase overall yield [12].

V. PROCESS REQUIREMENTS POWER MINIMIZATION

In [5], the energy delay trade-off for several device types is analyzed for use in logic gates, as seen in Figure 2. In this
analysis, a $V_{dd}$, $V_{th}$, and width for each device is found which minimizes power given a noise margin over a range of different delay requirements. This is done by looking at the current state of the art in each device then projecting device performance to the 10nm node. However, the devices that these projections are based on differ greatly in actual dimensions (30nm to 5um). Additionally, the lower limit of $V_{dd}$ in these curves is set by the deviation in $V_{th}$, yet no explanation on method for projecting to 10nm or inferring the variation is provided. This calls into question the amount of uncertainty in actual energy-delay tradeoffs.

![Energy Delay Tradeoff](image)

Figure 2: (a) Energy Delay Tradeoff (b) Nominal $V_{dd}$ vs. $V_{th}$

In SRAM since the required supply voltage is often large simply to satisfy the read/write margins over the expected device variation, delay optimization may not be a primary concern at typical clock speeds. In the context of the energy delay tradeoff curve, this means that for SRAM, the asymptotic value of energy per switch has the potential to be more important than the delay at which the device starts to show a tradeoff. Due to these issues we propose an investigation of the process requirements for several devices as a function of supply voltage.

To perform this analysis on a given device, we will come up with a rough estimate the $I_D$ vs $V_{gs}$ curve, which determines the $I_{on}/I_{off}$ ratio for a given $V_{dd}$ and $V_{th}$. Then we will attempt to model how variation of doping and physical dimensions will alter the values of $I_{on}$ and $I_{off}$. This will allow us to determine the tolerable level of physical variation within a device that would allow for an acceptable yield of the SRAM array at that supply and threshold. We will use an approach similar to \cite{11} for evaluating the relationship between device performance variation and yield. At a given supply we plan to iterate over several $V_{th}$ values to determine the threshold resulting in the least stringent process requirements. We will work with a fixed size of the SRAM Array and analyze both 6T and 8T cells. For the TFET we will analyze the 8T and 7T configuration since 6T is not possible using only TFETs.

VI. Conclusion

In this paper, four leading-edge alternative devices for high $I_{on}/I_{off}$ are introduced, while the inherent advantages and drawbacks of each are considered separately. Due to the extreme differences in manufacturing and operation between these devices, a comprehensive and general comparison technique is needed. We propose a method which focuses primarily on the physical origins of $V_{th}$, $I_{on}$, and $I_{off}$ variation in each technology. While there are large differences in delay among the options, because we focus on SRAM design, speed does not necessarily take priority. The primary goal is to minimize energy consumption while ensuring high read/write margin ratings. This allows us to easily estimate yield in a variety of SRAM implementations.

The strength of our analysis method stems from a focus on the relative changes in device performance caused by manufacturing variations. This approach is far more robust than comparing absolute parameters of the devices with respect to one another, due to their dissimilarity. The initial TCAD simulations extract dependencies of device parameters on doping, length, and other sources of manufacturing variability. To characterize a single-bit cell, behavioral models are used which are based on TCAD simulations. For the overall performance of an array, statistical yield estimates are derived from the behavioral simulations. Together, these techniques provide intuitive design knowledge about a diverse range of devices, regardless of the physical complexity of their operation.

REFERENCES

\cite{1} Kiyoo Itoh, B.S., "Adaptive circuits for the 0.5-V nanoscale CMOS."
\cite{2} Chang, L., et al., "Extremely scaled silicon nano-CMOS devices."
\cite{3} Deng, Jie., and Hon-Sum Philip Wong. "Device Modeling and Circuit Performance Evaluation for Nanoscale Devices: Silicon Technology Beyond 45 Nm Node and Carbon Nanotube Field Effect Transistors."
\cite{4} Sung Hwan Kim, et al., "Germanium-source tunnel field effect transistors with record high ION/IOFF."
\cite{5} Lan Wei; Oh, S.; Wong, H. -S P, "Performance benchmarks for Si, III–V, TFET, and carbon nanotube FET - re-thinking the technology assessment methodology for complementary logic applications."
\cite{9} Qin Zhang; Tian Fang; Huili Xing; Seabaugh, A.; Jena, Debdeep, "Graphene Nanoribbon Tunnel Transistors."
\cite{10} Yin-Nien Chen, et al., "Design and Analysis of Robust Tunneling FET SRAM."
\cite{11} Mukhopadhyay, S.; Mahmoodi, H.; Roy, K., "Modeling of failure probability and statistical design of SRAM array for yield enhancement in nanoscaled CMOS."
\cite{12} Nikolic,B., “EE241 Lecture Notes” 2013.