LOW-POWER HYBRID TFET-CMOS MEMORY

A Thesis

Submitted to the Faculty

of

Purdue University

by

Anoop Gopinath

In Partial Fulfillment of the

Requirements for the Degree

of

Master of Science in Electrical and Computer Engineering

May 2018

Purdue University

Indianapolis, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Maher E. Rizkalla, Chair
    Department of Electrical and Computer Engineering
Dr. Zina Ben Miled
    Department of Electrical and Computer Engineering
Dr. Lauren A. Christopher
    Department of Electrical and Computer Engineering

Approved by:
    Dr. Brian S. King
        Head of the Graduate Program

ACKNOWLEDGMENTS

I would like to thank my thesis advisers, Dr. Maher E. Rizkalla and Dr. Zina Ben Miled for their insightful and continued advice and input during the course of this work. I remain grateful to Dr. Trond Ytterdal of Norwegian University of Science and Technology (NTNU) for his recommendations and Notre Dame University for providing the Tunnel FET models. I would also like to thank Dr. Lauren A. Christopher, member of my thesis committee, for her feedback, and the Electrical and Computer Engineering department at Indiana University-Purdue University Indianapolis (IUPUI) for their support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

SYMBOLS

μ_n	Mobility of Electrons
ϵ_{OX}	Permitivity of the Oxide Layer
$Addr$	Address
B	Branch Effort
BL	Bit Line
BL_B	Bit Line Bar
C_{in}	Carry-in
C_{out}	Carry-out
C_{OX}	Capacitance Per Unit Area of the Oxide Layer
$Clk1$	Clock 1
$Clk2$	Clock 2
$Clk3$	Clock 3
$Clk4$	Clock 4
F	Effort Delay
G	Logical Effort
gm	Transconductance
I_{DS}	Drain to Source Current
I_{GS}	Gate to Source Current
$Inout$	Input-Output Pin
L	Channel Length
P	Parasitic Delay
Phi_1	Clock with 6ns Pulse Width
Phi_2B	Clock with 4ns Pulse Width and 1ns Delay
Phi_3	Clock with 6ns Pulse Width

Phi_4B	Clock with 4ns Pulse Width and 1ns Delay
Rd_en	Read Enable
$Sense\_en$	Enable for the Sense Amplifier
t_{OX}	Thickness of the Oxide Layer
V_{BS}	Body to Source Voltage
V_{DS}	Drain to Source Voltage
V_{GS}	Gate to Source Voltage
V_{th}	Threshold Voltage
W	Channel Width
WL	World Line
WR_en	Write Enable

ABBREVIATIONS

ADE	Analog Design Environment
BJT	Bipolar Junction Transistor
CMFB	Common Mode Feedback
CPU	Central Processing Unit
D-FF	D-Flip Flop
DIBL	Drain Induced Barrier Lowering
FA	Full Adder
FUG	Unity Gain Frequency
GaN	Gallium Nitride
GIDL	Gate Induced Drain Leakage
GPU	Graphic Processing Unit
MOSFET	Metal Oxide Semiconductor Field Effect Transistor
mV	Millivolt
nm	Nanometer
NMOS	N-type Metal Oxide Semiconductor
ns	Nanosecond
OTA	Operational Transconductance Amplifier
PDN	Pull-Down Network
PMOS	P-type Metal Oxide Semiconductor
ps	Picosecond
PUN	Pull-Up Network
SRAM	Static Random Access Memory
TB	Test Bench
TFET	Tunnel Field Effect Transistor

ABSTRACT

Gopinath, Anoop. M.S.E.C.E., Purdue University, May 2018. Low-Power Hybrid TFET-CMOS Memory. Major Professor: Maher E. Rizkalla.

The power consumption and the switching speed of the current CMOS technology have reached their limits. In contrast, architecture design within computer systems are continuously seeking more performance and efficiency. Advanced technologies that optimize the power consumption and switching speed may help deliver this efficiency.

Indeed, beyond CMOS technology may be a viable approach to meeting the ever increasing need for low-power design. These technology includes devices such as Tunnel Field Effect Transistor (TFET), Graphene based devices such as GFET and GRNFET and FinFET. However, the low cross-sectional area of the channel associated with smaller technology nodes brings with it the challenges associated with leakage current below the threshold. Mitigating these challenges with devices such as TFETs may allow higher levels of integration, faster switching speed and lower power consumption.

This thesis investigates the use of Gallium Nitride (GaN) TFET devices at 20nm for memory cells. These cells can be used in the L1 data cache of the Graphic Processing Units (GPU) thereby minimizing the static power and the dynamic power within these memory systems. The TFET technology was chosen since it has a low subthreshold slope of nearly 30mV/decade. This enables the TFET-based cells to function with a 0.6V supply voltage leading to reduced dynamic power consumption and leakage current when compared to the current CMOS technology.

The results suggest that there are benefits in pursuing an integrated TFET-based technology for Very Large Scale Integrated Circuit (VLSI) design. These benefits are demonstrated using simulation at the schematic-level using Cadence Virtuoso.

1. INTRODUCTION

1.1 Motivation

Stalls are due to various reasons in GPUs. These include cache misses, a busy pipeline, instructions dependencies and data dependency [1]. In order to support instruction-level parallelism, modern microprocessors use a pipelined architecture. In addition, computer systems utilize efficient pre-fetching mechanisms based in the principle of locality, branch prediction, out of order execution capabilities and high memory bandwidth in order to reduce stalls. The issue still remains because the processor speed is orders of magnitude faster than the memory speed. [2]. While the processor speed is not increasing as fast as once predicted by Moore's Law [3], the need for bridging the gap between processor speed and memory speed remains an active area of research.

Current research is focusing on processing in memory in addition to improving data fetch latency in traditional memory architecture [3]. In both cases, extra hardware including memory and processing units are needed at the L1, L2 or L3 cache levels in order to process data while the main processor is busy. The memory associated with these caches remains in an off-state, or a hold state in the SRAM cells while the units are not being accessed. Therefore, these memories will leak current and consume static power even if they are not being accessed. For these reasons, memories with low static and dynamic power consumption are needed.

Several power analysis and reduction techniques have been employed. For example, PowerRed estimates the power consumed by GPUs by using a framework that uses area-based models, RTL-based empirical models for measuring dynamic and leakage current, and analytical power models for the interconnects [4]. The Integrated Power and Performance prediction model (IPP) [5] allows the selection of an optimum number of cores for maximum efficiency by taking a GPU core and analyzing the performance and power consumption simultaneously. IPP predicts the performance/watt of GPGPU applications by using an empirical runtime power prediction model that chooses the optimum number of cores based on the application. GPUWatch [6] is an architectural power model based on a bottom-up methodology that allows cycle-level power estimates. In summary, the above research efforts focus on first developing analytical power models for GPUs then employing these models in order to estimate the power consumed, the performance of current architectures as well as to suggest enhancements.

1.2 Methodology

In this thesis, an integrated model with TFET devices and CMOS devices is proposed. The goal is to minimize static power as well as dynamic power in comparison to a CMOS only system. The analysis of the integrated model is conducted using Cadence Virtuoso Schematic L-Editing, Analog Design Environment and Spectre Circuit Simulator tools using the 20nm GaN TFET model [7] and the 45nm CMOS free PDK device obtained from North Carolina State University [8]. The experiment is confined to the schematic stage as the GaN TFET devices do not have layout cells or design rules which are required for fabrication, and post post-Silicon validation [7].

The designs are based on two technologies: one with CMOS arrays and one with CMOS-TFET arrays. The proposed design includes a 4-bit CMOS-based Carry Ripple Adder which act as the ALU, 4x4 CMOS-based Flip-Flops which act as register files, a 8x4 combination of CMOS-based SRAM array and TFET SRAM array which act as the L1 data cache. This design is implemented and functionally verified. Moreover, the leakage current and static power of the SRAM array is are measured and compared to a design with CMOS only 8x4 SRAM array-based L1 cache. Leakage current is evaluated for different configurations with varying sizes 8x4, 16x4, 32x4 to 64x4.

The proposed design can be adopted for upper level memories for GPUs. The TFET-based array can be used in the event of an off-chip memory access or stall. This off-chip access is simulated by disabling the read for the SRAM array rows which contain the data that is requested by the Central Processing Unit (CPU).

While the CMOS array stalls, the TFET array is used by the ALU to compute an addition of two 4-bit operands. Essentially, when the CMOS array is active, the TFET array is switched OFF and conversely, when the CMOS array is stalled the TFET array is switched ON, to compute data.

The rest of this thesis is structured as follows: Chapter 2 discusses the sources of leakage current that contribute to static power consumed and related research efforts in this area. Chapter 3 describes the TFET device, their operation and their properties. Chapter 4 focus on the GPU architecture and memory hierarchy and Chapter 5 includes the design of some basic digital and an analog circuits using the GaN TFET model. The leakage current of these circuits are also analyzed in Chapter 5. The focus of Chapter 6 is on the design of a CMOS-based L1 data cache, an integrated TFET-CMOS L1 data cache, CMOS-based registers designed from D-Flip Flops and a 4-bit carry ripple adder.

2. RELATED WORK

Current research into VLSI design at the sub-micron technology level is facing issues due to leakage current, supply voltage and a subthreshold swing of CMOS devices. The leakage current is the unwanted current when the transistor is switched-off. Supply voltage scaling helps with reducing dynamic power consumption. The subthreshold swing shows the switching speed of the transistor at the linear operating region of the transistor and the leakage in the subthreshold region of the transistor.

Researchers have pursued various approaches to address the above issues including using nanotechnology materials and devices, seeking higher levels of integration, higher switching speed, and minimum power consumption. FinFET devices were pursued in order to integrate larger number of transistors on a chip at lower operating voltages. The advantages of FinFETs include low body effects, the need for a single additional mask during fabrication, lower leakage current and lower threshold voltage. The disadvantage of these devices include the corner effect and the parasitic source/drain resistance which lead to lower drive current [9]. Since the level of integration, the power consumption, and the switching speed are related to the device geometry and the ability to use interconnects with very small cross-section, nanotechnology materials (e.g., graphene or carbon nanotubes) were pursued. These materials can be incorporated into the devices to provide much shorter interconnect materials and very short channels.

2.1 Leakage Current

As the channel length between the source and the drain of the device reduces with technology scaling, the role of the gate of the transistor as a voltage controlled switch degrades. This is due to the tunneling of the electrons even when the gate voltage is at a logic zero or off. Primarily, there are three leakage components: subthreshold leakage, gate leakage and diode/junction leakage [10]. Together, these components cause significant amount of electrons to leak through to the ground from the transistors which are switched-off in VLSI chips.

Subthreshold leakage is the leaking of electrons from the source of the transistor to the body terminal of the transistor due to Drain Induced Barrier Lowering (DIBL) [10]. When the voltage across gate and source (VGS) is below the threshold voltage (Vth) of an NMOS transistor, if the drain of the transistor is subjected to high positive voltage, it creates a depletion region around the drain. This is due to the increased positive polarity of the drain diffusion, causing the p-type body, where the majority carrier are holes, to repel away from the drain. At the same time, a lateral electric field is created from the drain to the source given by the following equation:

$$E = V_{DS}/L \tag{2.1}$$

where, V_{DS} is the drain to source voltage and L is the channel length of the NMOS transistor. This lateral electric field causes the barrier between the source diffusion region and the body to lower allowing electrons to tunnel through to the body [10].

Leakage is also associated with the gate terminal of the transistor. As transistor scaling continues, not only is the length, diffusion regions, body and gate terminal smaller, but also so is the oxide layer between the gate terminal and the body that isolates and protects the gate terminal from electrons moving from and to the gate. The oxide layer allows the channel to be formed under it as it is a good insulator and does not allow electrons to transfer through it. However, the current oxide layer thickness is in the region of a few layers of atoms, which increases the probability of electrons tunneling through the gate terminal [10].

The PN junctions between the diffusion regions and the substrate act as reverse biased PM-junction diodes. Reverse biased diodes do not conduct current. In a traditional NMOS transistor, the p-type body and the two n-well regions (source and drain) form an NMOS transistor. This creates two reverse biased diodes: one between the p-type body and the source and the other between the p-type body and the drain. The p-type body is the anode or the positive terminal and the n-well source and drain diffusion regions are the cathode or negative terminal. These PN-junction diodes will remain reverse biased as long as the body terminal is grounded for the NMOS transistor. Therefore, they ensure that no significant voltage flows through the body to forward bias and turn-on the diodes. However, the reverse biased diodes also conduct some current from the diffusion regions to the body when the transistor is switched-off.

Other methods such as halo doping and body biasing are used to increase the threshold voltage to reduce subthreshold leakage [10].

Gate Induced Drain Lowering (GIDL) occurs when the gate overlaps with the drain and source diffusion region. Some transistors have this structure with an overlapped gate to better control the transistor. However, when a negative voltage is applied at the gate, the overlap portion of the gate over the n-type drain in an NMOS transistor attracts the minority holes in the drain under the overlap region. This causes a P-P short between the drain and the p-type body, and creates a pathway a pathway for electrons to leak through to the body from the drain. This component also adds to the total off-current of the transistor [10].

Leakage current can be reduced by increasing the threshold voltage through the stack effect, by the use of sleep transistors or by controlling the input vectors. At the system level, parallelism and pipelining can further reduce leakage current as they promote hardware reuse thus reducing the off-time for the components and as a consequence reducing the leakage current [10].

Several methods have been employed to reduce leakage current in CMOS circuits. The stacked effect reduces the leakage in circuits by a factor of 10 [10]. Stacking two transistors in a series reduces the leakage current because the leakage for this configuration is a function of the input pattern and the number of transistors. Threshold voltage, V_{DS} and V_{GS} have an exponential effect on the subthreshold leakage [11]. Subthreshold leakage can be controlled if the bias conditions of the device are adjusted in order to control the threshold voltage. The input patterns of each gate affects the subthreshold as well as the gate leakage current. Reducing V_{GS} exponentially reduces the subthreshold leakage and one method to achieve this is by source biasing to reduce the source voltage of the transistor [12]. This can also result in the increase of V_{BS}, known as body biasing which can lead to body effect.

The body effect results in an increase in charge required at the gate to invert the channel which in turn increases the threshold voltage and decreases the leakage current [10]. The V_{GS} and the V_{BS} of the upper transistor is negative. Conceptually this will result in an increase of threshold voltage and in turn a reduction of leakage current of the upper transistor. This phenomenon is known as the stack effect [13] [14]. The stack effect induces increased delay.

Sleep transistors can be used to cut down the power to a logic block when it is inactive or in sleep mode by using PMOS-based sleep circuitry [15]. PMOS transistors are designed to modulate the V_{DD} to the logic block in the sleep mode by using an enable signal. When this enable is logic low, the sleep transistors are switched-on, thus providing the actual V_{DD} to the logic block. However, when the enable is logic high, the PMOS sleep transistors are switched-off, thus completely cutting-off supply voltage to the block [10]. The sleep transistors need to be carefully sized in order to avoid leakage current leaking through these transistors which can accidentally switch-ON the logic block in the sleep mode.

Another popular method for reducing leakage current consists of using a minimum leakage vector (MLV) [11]. The leakage current from a circuit varies depending on the combination of transistors in the circuit that are switched-on. For example, in [11], it was shown that the leakage current for a 3-input NAND gate in a particular process is smallest when the three input combination is 1-0-0 and highest when 1-1-1. To ascertain the lowest leakage from a circuit, this method could be used. When the input vector with the lowest leakage is established, it can be used to trigger the sleep mode. These vectors can be added to the circuit by using multiplexers.

2.2 Supply Voltage

The supply voltages used for CMOS integrated circuit design has been dropping due to transistor scaling. Currently, it has a value of about 0.7V. Reducing this voltage is important in order to reduce dynamic power consumption. However, this reduction remains a challenge because of its direct influence on the threshold voltage and switching speed. Dynamic voltage and frequency scaling cuts-off voltage and clock supply to blocks that are in sleep mode, or in switched-off mode. The method also supplies different supply voltage to different IP blocks depending on the need of each block. Multivoltage level shifters can be used to supply different voltages to different blocks in order to reduce the switching power. Dynamic power has two components: switching power and short circuit power. Short circuit power is the voltage that is directly shorted from the power supply to the ground when both the pull-up and pull-down network are switched-on partially during switching. Typically, this is very small small. The switching power is governed by the following equation:

$$P_{switching} = \alpha \times C \times F \times V_{DD}^2 \tag{2.2}$$

where, α is the switching activity factor. It represents the probability that a signal switches in a cycle. C is the output drain diffusion capacitance. It is the primary cause of delay at the output of a circuit. F is the frequency of the clock, and V_{DD}, is the supply voltage [16].

Clock gating is a popular method to reduce switching power in sequential circuits. It consists of cutting down the clock to blocks in sleep mode [10]. Clocks have a very high switching activity factor α. This activity factor is 0.5 for clocks as clocks switch every cycle [10]. Clock gating consists of using an AND gate and a latch, where one of the input to the AND gate is the original clock and the other input is the output of the latch. The latch takes in as its input an enable signal and the original clock's inverted input. When the latch is enabled, it supplies a logic one to the AND gate, which switch-on the ADN gate, and supply the clock to the block.

Switching power can be reduced by reducing the capacitance C of the circuit. This capacitance is the output diffusion capacitance of the circuit also known as parasitic capacitance. It contributes to the delay at the output node of the circuit. Parasitic capacitance depends on the diffusion area, perimeter of the source and the drain diffusion regions of a transistor. It is a function of the depth and doping levels of the diffusion region and the voltage. The parasitic capacitance can be reduced by using smaller transistors with smaller parasitic capacitance [10]. Furthermore, these capacitances can be reduced via physical design practices such as shared or merged diffusion, or by reducing the number of stages (at the expense of increasing the delay).

2.3 Subthreshold Slope

The MOSFET I-V characteristics show that traditional devices have a slope of around 60mV/decade. This indicates the switching speed of the device. The speed at which a device turns-on and off is dependent on the supply voltage applied. The current in the subthreshold region of a transistor drops by a power of 10 when the voltage applied at the gate drops by 60mV/decade at room temperature (i.e., 27 °C). Moreover, the current in the subthreshold region increases exponentially. Therefore, it is more advantageous to have a smaller subthreshold slope in order to reduce the current in subthreshold region. Compared to traditional CMOS devices, GaN TFETs can provide attractive subthreshold swing which is close to 30mV/decade.

3. TUNNEL FIELD EFFECT TRANSISTOR

TFET is a relatively new type of transistor which is similar to the popular CMOS transistor but with its source and drain doped of opposite type. TFETs are becoming increasingly popular in micro and nanoelectronic applications due to their considerably low static leakage when the transistor is turned-off compared to the CMOS transistors. As a result of this low leakage property, TFETs are being explored in various areas where power optimization is a primary concern including memory cells and non-critical logic paths.

CMOS transistors have been popular since the 1970s. However, they are known to consume energy while switching between ON and OFF operation states. CMOS are similar to the Bipolar Junction Transistors (BJTs) in that they both control the flow of current by raising and lowering energy barriers. The electrons flow from the source of the transistor to the drain through the channel formed in the bulk material substrate at the appropriate gate bias. The electrons flow through the conduction band and the holes flow through the lower energy valence band.

Electrons are given energy to allow them to transcend the energy gap and move into the conduction band, a strategy to lower the energy barrier. Lowering the energy barrier depends on the voltage required to switch the transistors ON and OFF. With technology scaling, voltage scaling is becoming a focus area in the chip industry.

Another issue that is being widely researched is the static leakage current in CMOS transistors. Leakage current is the unwanted current in the transistor channel even when the transistor is switched-off. This is the current in the transistor below the threshold voltage. This region is called the subthreshold region. The subthreshold slope for MOSFETs is 60mV/decade whereas, it is 30mV/decade for GaN TFETs. The presence of leakage current suggests the need for alternatives to CMOS-based circuits, particularly for very low supply voltage electronic design. While the source and drain of TFETs are doped of the opposite type, they have a similar structure to that of CMOS transistors. Figures 3.1 and 3.2 show the cross-section of a traditional NMOS transistor and an NTFET transistor, respectively.

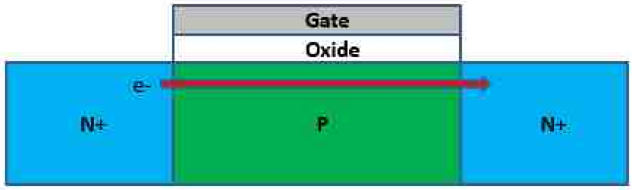

Fig. 3.1. Cross-section of an NMOS Transistor.

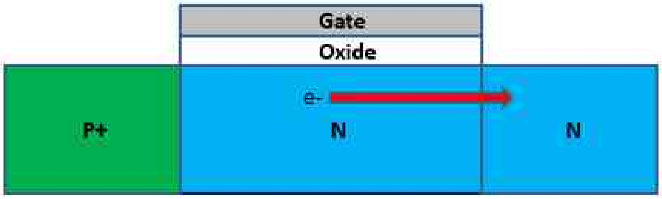

Fig. 3.2. Cross-section of a NTFET.

At the nano-electronic scale, where low power design is an overriding factor, the emerging TFETs may provide a solution to leakage current. Not only do GaN TFETs work with a lower input voltage, they can turn the unwanted leakage current to and advantage. TFETs use the band-to-band tunneling of electrons from the source to the drain. As opposed to raising or lowering the energy barrier between the source and drain in CMOS devices, TFETs control the probability of the electrons tunneling through barrier based on the electrical thickness of the barrier. [17]

TFETs were first simulated and shown to hold a subthreshold swing closer to 40mV/decade at IBM in 2004 [18]. TFETs are similar to Metal Oxide Semiconductor Field Effect Transistors (MOSFETs) in that they have a source and drain. They are different in that their source and drain are oppositely doped. This opposite doping essentially making a p-i-n or n-i-p combination reminiscent of PN-junction diodes. The TFETs facilitate the Band-To-Band-Tunneling (BTBT) of electrons through a barrier instead of the traditional thermionic emissions of electrons over a barrier in MOSFETs [18]. Figure 3.3 shows the I_{DS} vs. V_{GS} characteristics in the subthreshold region of an NTFET.

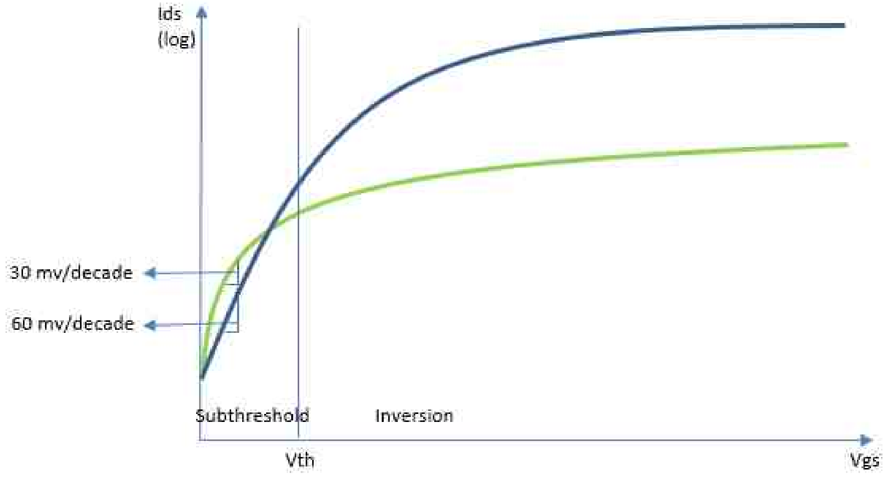

Fig. 3.3. Subthreshold Slope of a NTFET.

As previously mentioned, in TFETs, the source and the drain have opposite type doping. Thus, if the source is doped n-type, then the drain is doped p-type. Unlike in CMOS, where electrons flow from the source to the drain via a channel that is formed at the appropriate gate voltage, in TFETs, band-to-band tunneling of electrons results in the flow of electrons from the source, through channel to the drain. Figure 3.4 shows the alignment of the bands resulting in a TFET ON and OFF states.

In MOSFETs, electrons and holes either move through the conduction band or the valence band all the way from the source to the drain. But in TFETs, at the appropriate gate bias, the conduction band of the source and valence band of the intrinsic region align, and thus allow the electrons to flow through, resulting in current

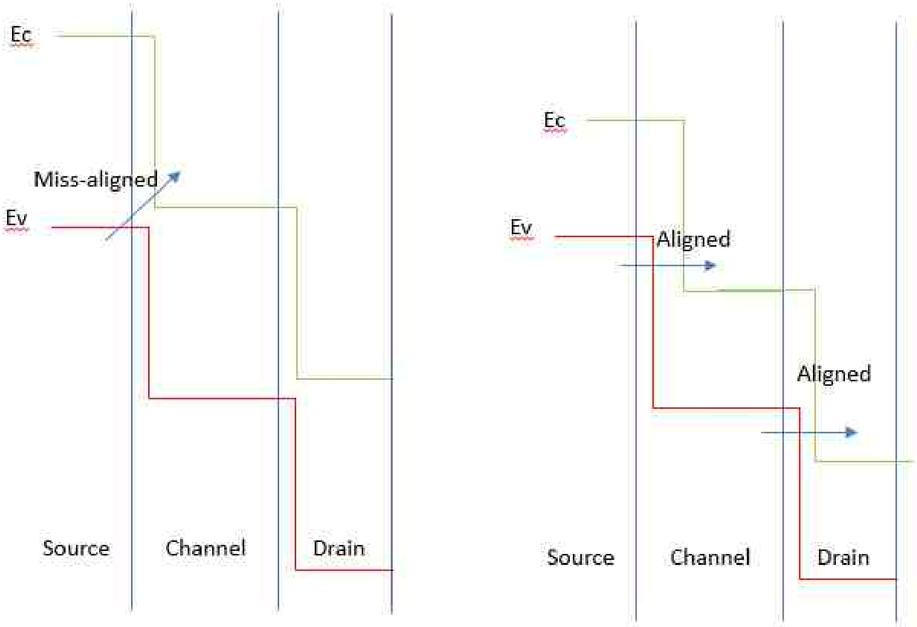

Fig. 3.4. Band-To-Band-Tunneling in TFETs.

flow. Since the primary mechanism behind the tunneling is the alignment of the conduction and the valence band, voltage applied to the gate should just be enough to create or remove an overlap of the two bands in order to switch ON and OFF the transistor.

Another key difference between TFETs and MOSFETs is the direction of the current flow. While MOSFETs are bi-directional devices that allow current to flow from source to drain and from drain to source, TFETs are uni-directional devices. The current flow in TFETs is from drain to source but not from drain to source.

Initially TFETs were fabricated using Silicon and Germanium. However, they failed to achieve the necessary drive current to rival CMOS at a smaller supply voltage. This led to a shift to III-V materials in order to create homo-junction and hetero-junction TFETs with smaller and more direct tunnel barrier. The homo-junction TFETs are devices which are manufactured using a single semiconductor material for the source, channel and the drain. The material used for fabrication is in the family of

Indium-Gallium-Arsenide. A pocket material with a narrower bandgap can be added to the boundary between source and channel in order to create a tunnel barrier at the source and increase the tunneling probability of the TFETs.

For hetero-junction TFETs, two separate semiconductors are used to engineer a tunnel barrier between the source and the channel. The material used for the source is different from the material used for the channel and the drain. Indium-Phosphorous is commonly used for the source and Indium-Gallium-Arsenic for the channel and the drain in the case of a type-I heterojunction TFET. A type III heterojunction TFET can be made using Gallium-Antimony for the source, and Indium-Arsenic for the channel and drain [19].

4. GRAPHIC PROCESSING UNIT

Graphic Processing Units (GPUs) consist of a large number of cores, allowing thousands of threads to execute in parallel. The primary objective of GPUs is to accelerate graphic processing. However, due to their inherent parallel architecture, more general purpose computations such as deep learning, analytics, processing of large data sets have taken advantage of the processing power of GPUs [20]. In general, GPUs transfer data from the memory associated with the CPU to their local memory through a PCI data bus as shown in Figure 4.1.

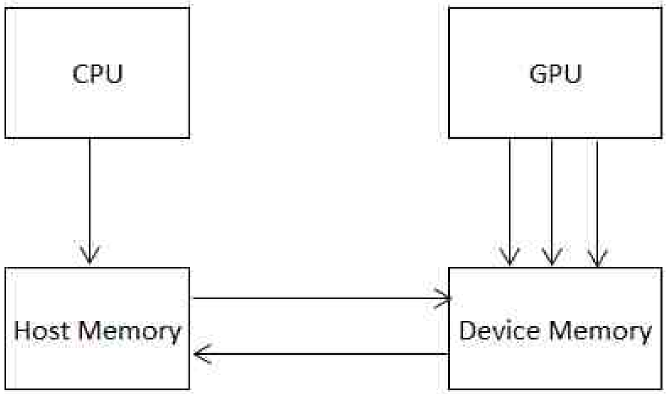

Fig. 4.1. CPU-GPU Communication.

The memory hierarchy in GPUs includes Texture memory, Global memory, caches and registers. The texture memory is the lowest level and the registers are the highest level of the hierarchy. GPUs execute a single program (kernel) on all the cores in parallel. The kernels consist of multiple threads where each thread is executed on a single core. To effectively manage threads, they are grouped into blocks and the thread blocks in turn are grouped into grids as shown in Figure 4.2.

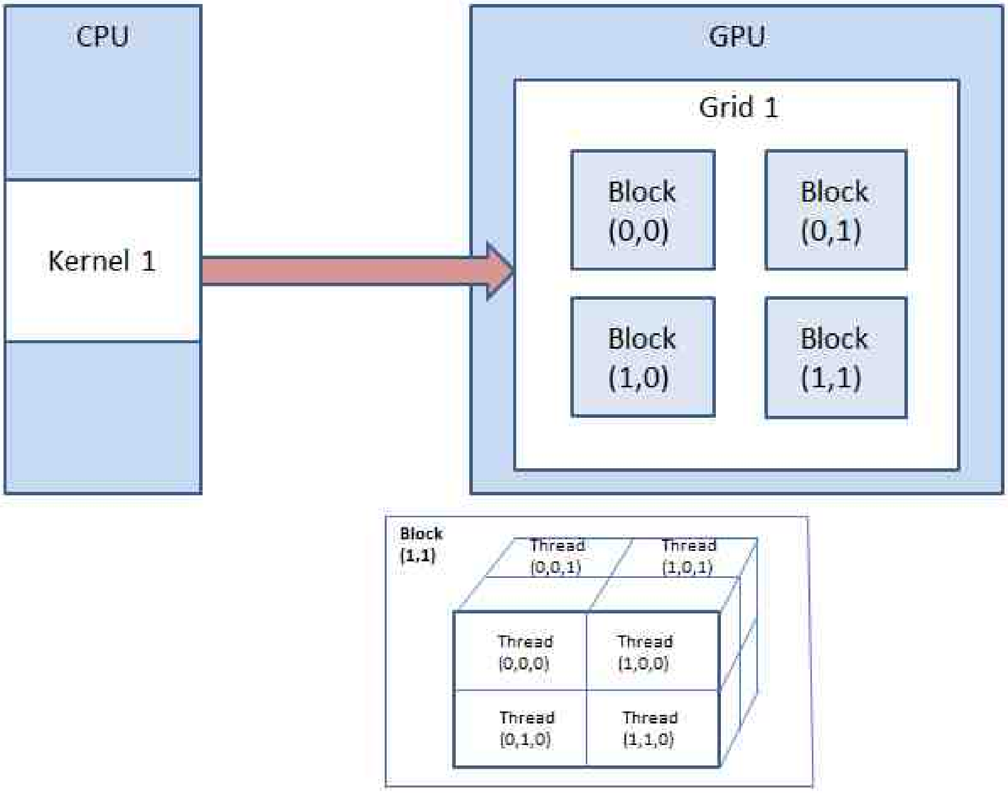

Fig. 4.2. Grids, Blocks, Threads in GPUs.

Running the same sequence of instructions on different data streams on GPUs requires large number of memory accesses. These memory accesses are initiated by each core in order to migrate data from memory to the register file as shown in Figure 4.3. Each thread executes on a core and has access to a dedicated register file. If a data miss occurs resulting in an off-chip memory access request, the rest of the threads in the block are likely to also request an off-chip memory access as all threads in a block execute the same sequence of instructions. While modern GPUs have large bandwidths and very strong pre-fetching mechanism to avoid misses, the bandwidth is still constrained when compared to the number of off-chip memory access requests issued by the threads. This is especially true for memory intensive benchmark programs [21], and these memory stalls can degrade performance.

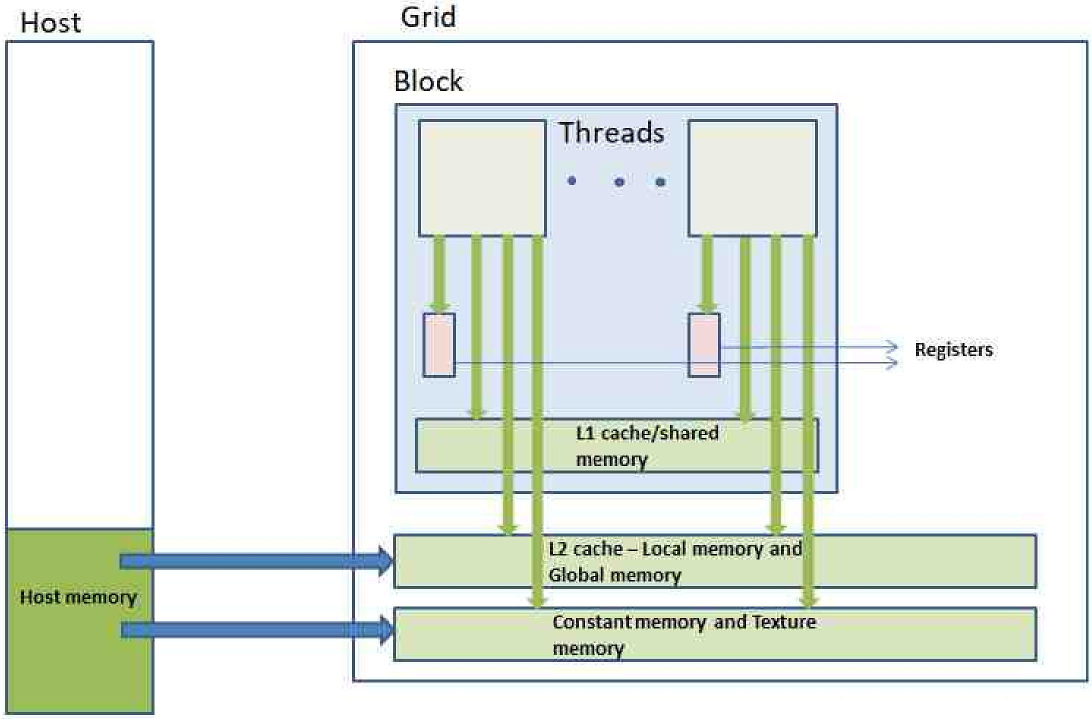
Host
Grid
Block
Threads
Registers
L1 cache/shared memory
L2 cache – Local memory and Global memory
Constant memory and Texture memory
Host memory

Fig. 4.3. GPU Memory Hierarchy.

Incorporating large number of cores on a single chip also raises both dynamic and static power consumption issues. Power consumption issues due to the decreasing feature sizes of the various components within a GPU architecture. Chip manufacturers attempt to circumvent the limitations imposed by Moores Law on traditional CMOS devices by using smaller transistors. However, smaller transistor sizes lead to smaller channel length which increases static power consumption when the transistors are switched-off. GPUs use multiple cores in parallel in order to improve speed and functionalities. However, this also comes at the expense of an increase in power consumption [20].

5. GAN TFET CIRCUITS

This chapter introduces digital and analog circuits using the Gallium-Nitride (GaN) TFET and compares the leakage current and static power of these designs against CMOS-based technology nodes. A new performance metric (Figure of Merit - FOM) is proposed as the basis for comparison for both the digital and analog cells. This chapter also presents current versus voltage plots (i.e., I_{DS} vs. V_{DS} and I_{DS} vs. V_{GS}) for the GaN NTFET and PTFET. More specifically, transient response characteristics and transfer curves are analyzed for a basic unskewed inverter, two-input NAND (NAND2), and two-input NOR (NOR2). In addition, an operational transconductance amplifier (OTA) designed with GaN TFET is compared to various CMOS-based counterparts.

In order to obtain a fair baseline for comparison, the gate width to channel length ratio was kept the same across all technology nodes. The circuits were designed in Cadence Virtuoso, design parameters were set in the Analog Design Environment (ADE), and the simulation was performed using the Spectre Circuit Simulator tool. It is to be noted that there is no body terminal for GaN TFET devices. Therefore, factors such as body effect or variable threshold voltage are not considered.

TFETs are uni-directional devices. The current for n-channel device flows from the drain to the source, but not from the source to the drain in normal operating mode [17]. Therefore, our analysis of leakage current focuses on the subthreshold and gate leakage and does not extend to junction leakage [7].

The remainder of this chapter is organized as follows: Section 5.1 discusses devices characteristics. Section 5.2 focuses on schematic design and output voltage curves for commonly used digital and analog cells and Section 5.3 presents the static power-dependent FOM results.

5.1 GaN TFET Device Characteristics

This section presents the DC characteristics of the GaN PTFET and NTFET. The voltage transfer curves for the GaN NTFET device are shown in Figures 5.1 and 5.2. Figure 5.1 shows the I_{DS} vs. V_{DS} for the device at V_{GS} values of 0.15, 0.3, 0.45 and 0.6V. Figure 5.2 shows the simulation results of I_{DS} vs. V_{GS} at V_{DS} value of 0.6V. The threshold voltage (V_{th}) for the GaN NTFET from the curves was found to be approximately 0.15V. The GaN PTFET device was designed by reversing the direction of the current of the NTFET device [7]. The magnitude of the threshold voltage for the PTFET was also found to be approximately 0.15V.

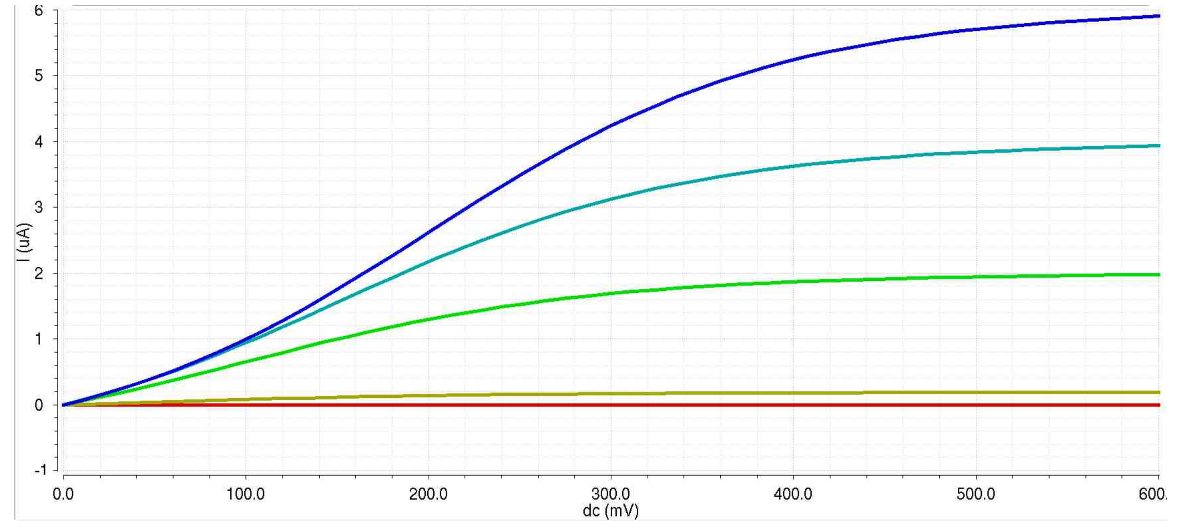

Fig. 5.1. NTFET I_{DS} vs. V_{DS}.

5.2 GaN TFET Circuit Design

The GaN TFET devices characterized in the previous section were used to design a basic inverter, two-input NAND gate (NAND2), two-input NOR gate (NOR2), as well as an operational transconductance amplifier (OTA). For comparison purposes, equivalent CMOS-based circuits with 45nm, 100nm, 150nm and 280nm gate lengths

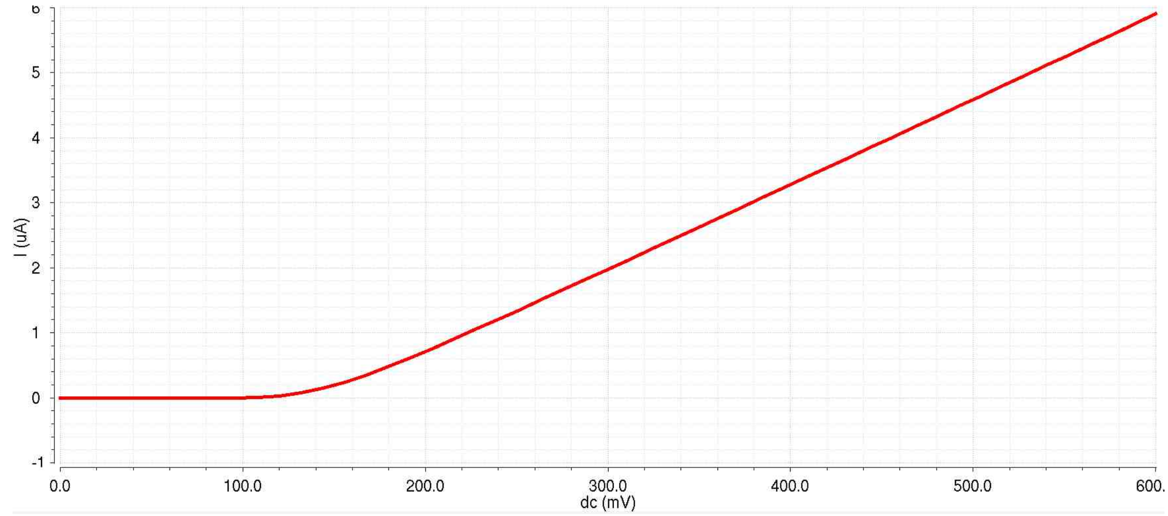

Fig. 5.2. NTFET I_{DS} vs. V_{GS}.

were also designed. In order to measure the leakage current and static power, the design parameters were set in the ADE in Cadence Virtuosos, before running a transient response and a DC response for the digital circuits. In the case of the OTA, an AC analysis was also performed in order to obtain the gain and unity-gain frequency.

For consistency purposes and in order to enable cross-technology analysis, the gate width to length (W/L) ratio was set to 5:1 across all technology nodes since the leakage current decreases as the channel length increases. Moreover, the FOM metric is proposed to compare different designs. In the case of the digital circuits, FOM is the product of the static power measured via Cadence Virtuoso and the delay which is manually calculated using the linear delay model [10]. In case of the analog cell (OTA), FOM is defined in Equation 5.1:

$$FOM = P/(A0 \times FUG) \tag{5.1}$$

where, P is the static power consumption of the cell, $A0$ is the DC gain, and FUG is the unity gain frequency.

The inverter is sized PTFETs: NTFETs at 2:1. The NAND2 is sized PTFETs: NTFETs at 2:2 and the NOR2 is sized PTFETs: NTFETs at 4:1. All are unskewed cells [10] with equal rise and fall times. For a faster rise time, the pull-down network NTFETs can be made half the size of the unskewed cells. These cells are called high-skewed gates [10]. Conversely, for a faster fall time, the pull-up network PTFETs can be sized at half the size of the unskewed cell [10]. These cells are called lo-skewed gates [10].

5.2.1 Inverter

A TFET inverter or NOT gate inverts the input and is similar to CMOS-based inverters. It consist of a p-type device as the pull-up network and an N-type device as the pull-down network. For an inverter, when the input V_{IN} is 0V, the PTFET is switched-on, the NTFET is switched-off and the output V_{OUT} is driven to V_{DD} through the PTFET, thus inverting a logic 0 to a logic 1. When the input V_{IN} is 0.6V, the PTFET is switched-off, the NTFET is switched-on and the output is discharged to the value at Gnd (0). Figure 5.3 shows the schematic of the GaN TFET inverter and Figure 5.4 shows the transient response of the cell in Cadence Virtuoso with a simulation time of 100ns. This simulation follows the expected output pattern.

5.2.2 NAND Gate

The GaN TFET NAND2 gate consist of 4 transistors: two PTFETs in parallel connected to the drain of the top of the two NTFETs in series. The NAND2 generates an output of 0.6V when either one of the PTFET devices or both PTFETs are switched-on and generates the output 0V only when both NTFET devices in series are switched-on. Figure 5.5 shows the schematic of an unskewed NAND2 gate. Figure 5.6 shows the corresponding transient response of the gate where the output is simulated at 100ns time intervals using Cadence Virtuoso.

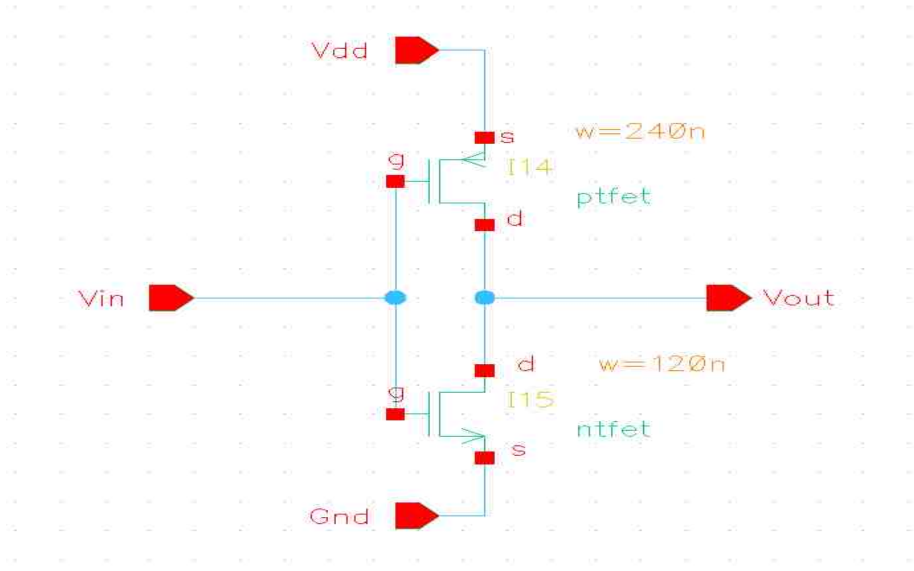

Fig. 5.3. GaN TFET Inverter Schematic.

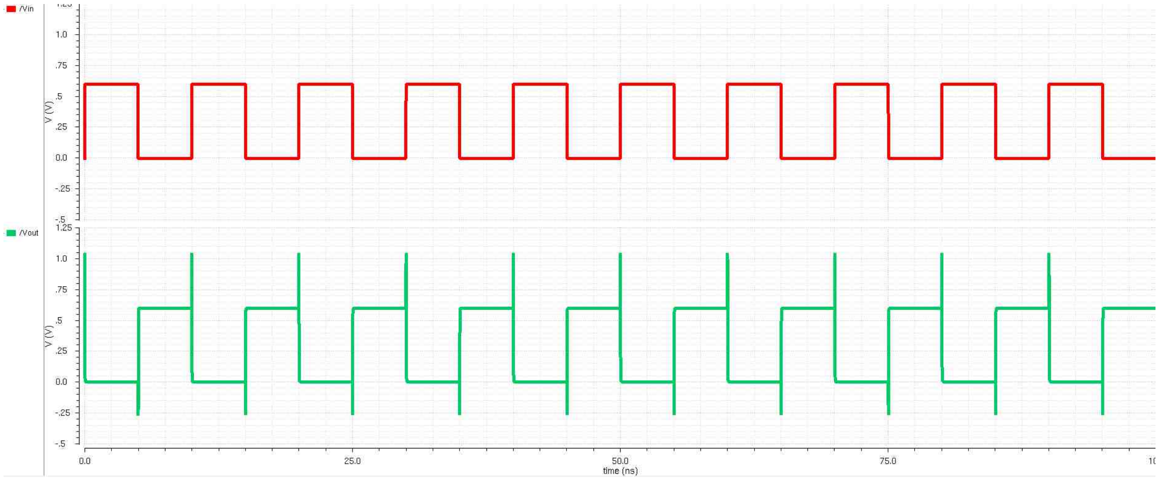

Fig. 5.4. GaN TFET Inverter Transient Response.

5.2.3 AND Gate

The GaN TFET AND2 gate is designed by connecting the output of an GaN TFET NAND2 to an GaN TFET inverter. The AND2 generates an output of 0.6V only when both of the PTFET devices are switched-on and generates an output of

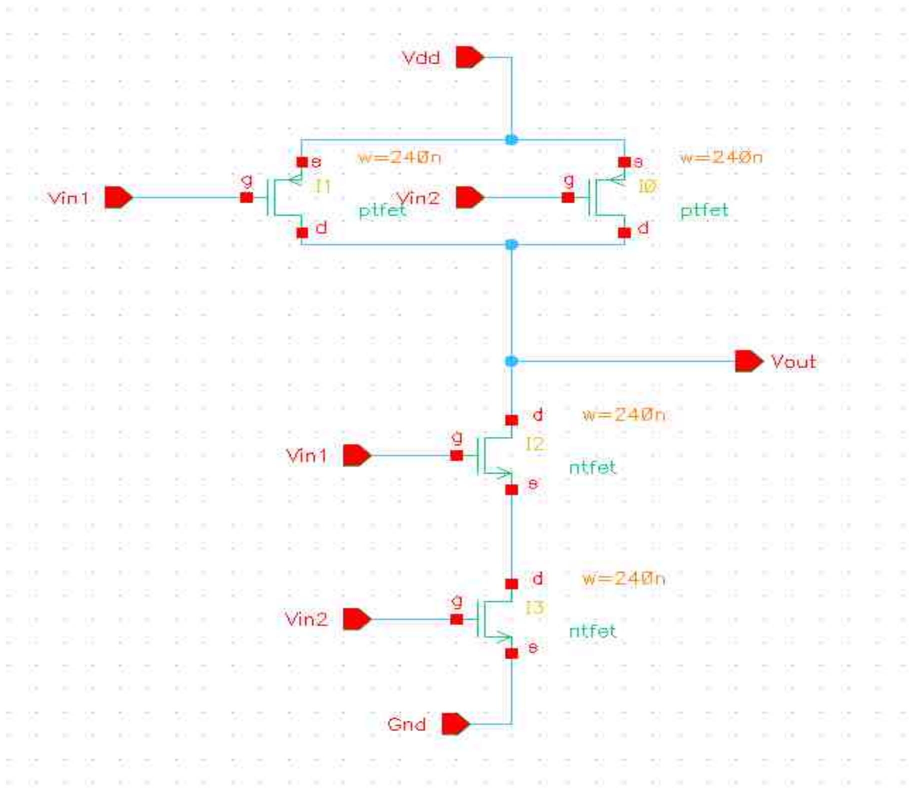

Fig. 5.5. GaN TFET NAND Schematic.

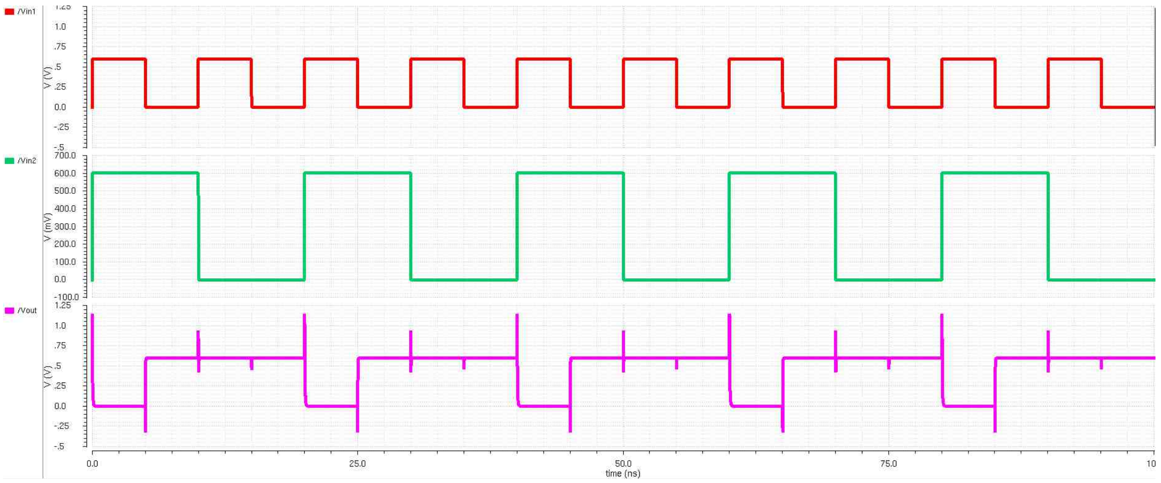

Fig. 5.6. GaN TFET NAND Transient Response.

0V when either of the NTFET devices in series are switched-on. Figure 5.7 shows the schematic of an unskewed AND2 gate and Figure 5.8 shows the corresponding transient response of the gate covering all four input combinations. The resulting output pattern is as expected for a logic AND2 gate.

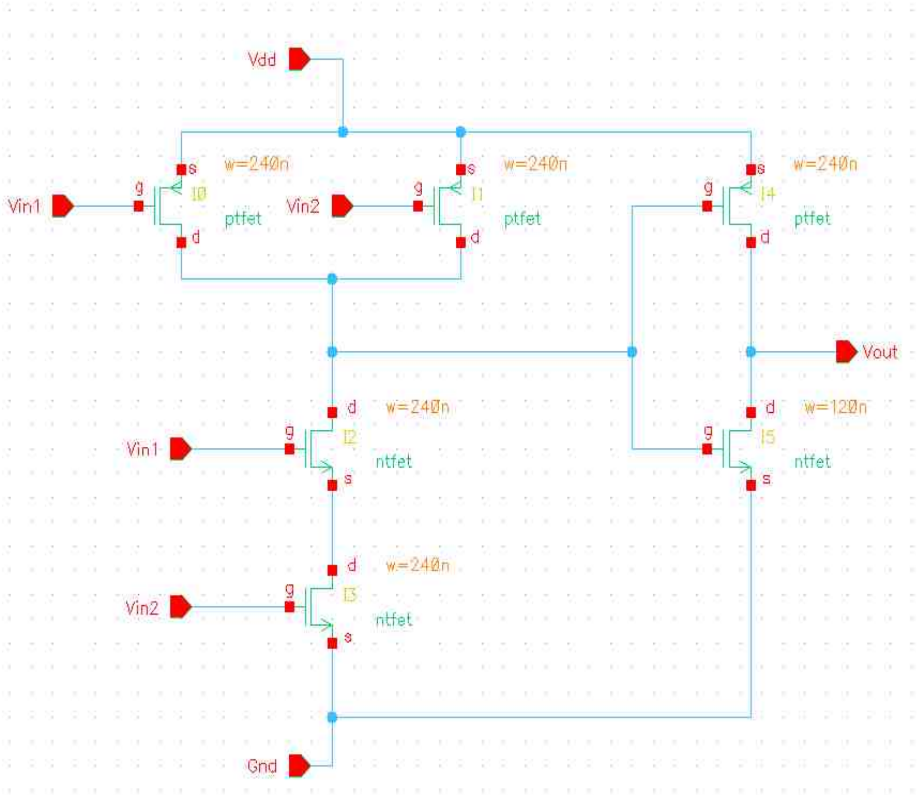

Fig. 5.7. GaN TFET AND Schematic.

5.2.4 NOR Gate

The GaN NOR2 gate consist of two NTFETs in parallel, connected via their gates to the drain of the bottom of two PTFETs in series. The NOR2 gate generates an output of 0.6V when both inputs to the gate are 0.6V thereby switching-on the PTFETs in series, and allowing the output node V_{OUT} to be driven to V_{DD}. When either one of the NTFET is switched-on, the NOR2 gate generates an output of 0V.

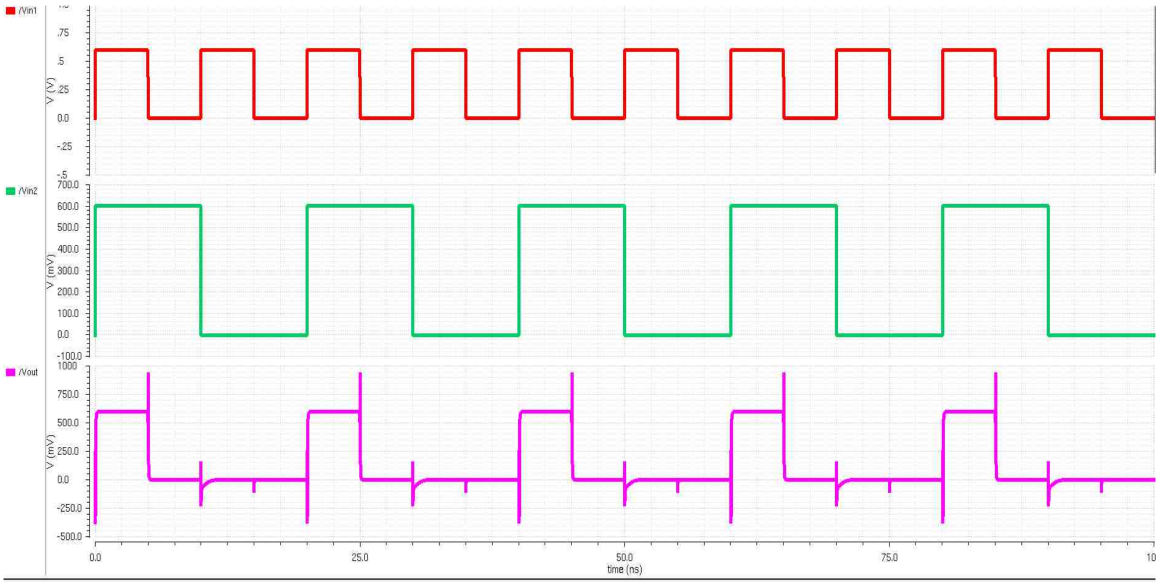

Fig. 5.8. GaN TFET AND Transient Response.

Figure 5.9 and Figure 5.10 show the schematic of the NOR2 gate and its transient response, respectively. The output pattern is as expected for a two-input logic NOR gate.

5.2.5 OR Gate

The GaN TFET OR2 gate is designed by connecting a GaN TFET NOR2 gate to an GaN TFET inverter. The OR2 gate generates an output of logic 0 when both of the inputs to the gate are logic 0 thereby switching-on the PTFETs in series. This drives the intermediate node between the NOR2 and the inverter high, thus activating the NTFET of the inverter, which in turn drives the output to a logic 0. The OR2 gate generates a logic 1 when either of the inputs to the gate is high. This switches-off the PTFET series combination and switching-on one or both of the NTFETs, thus allowing the intermediate node to be discharged to a logic 0. This logic 0 switches-on the PTFET of the inverter, thus driving the output of the OR2 gate to a logic 1.

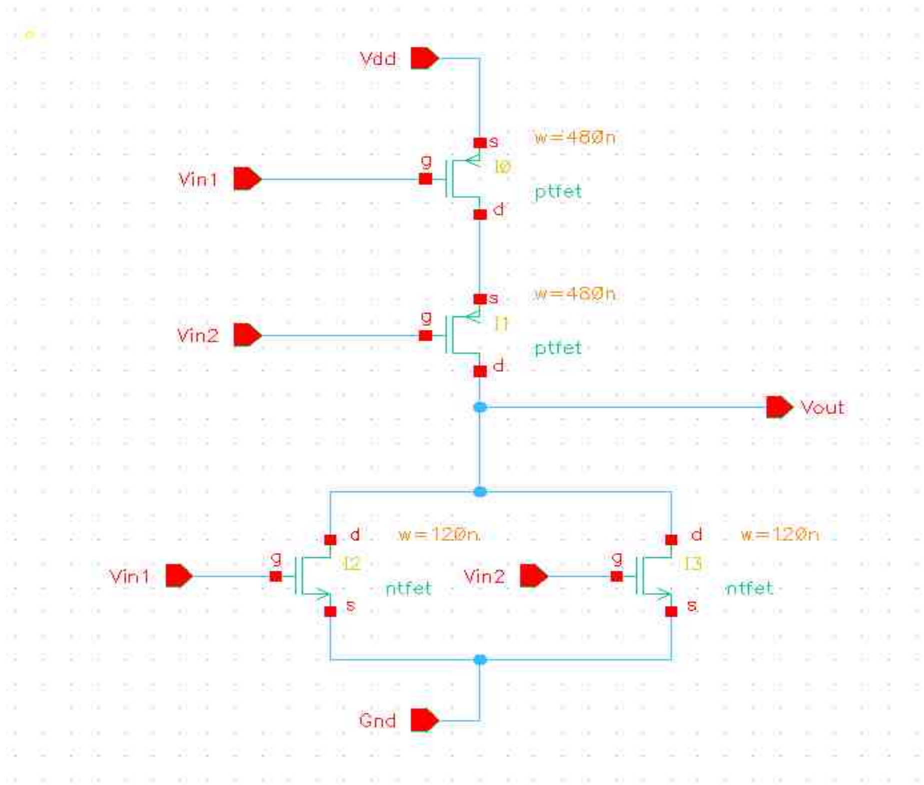

Fig. 5.9. GaN TFET NOR Schematic.

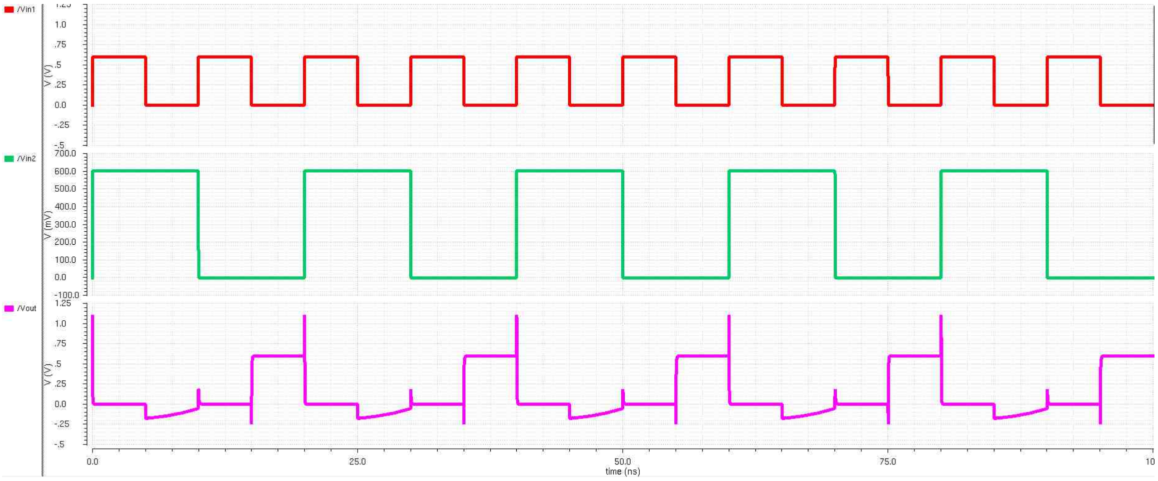

Fig. 5.10. GaN TFET NOR Transient Response.

Figure 5.11 and Figure 5.12 show the schematic of the NOR2 gate and its transient response, respectively. This response is as expected for the gate and corresponds to the logic of a two-input OR gate.

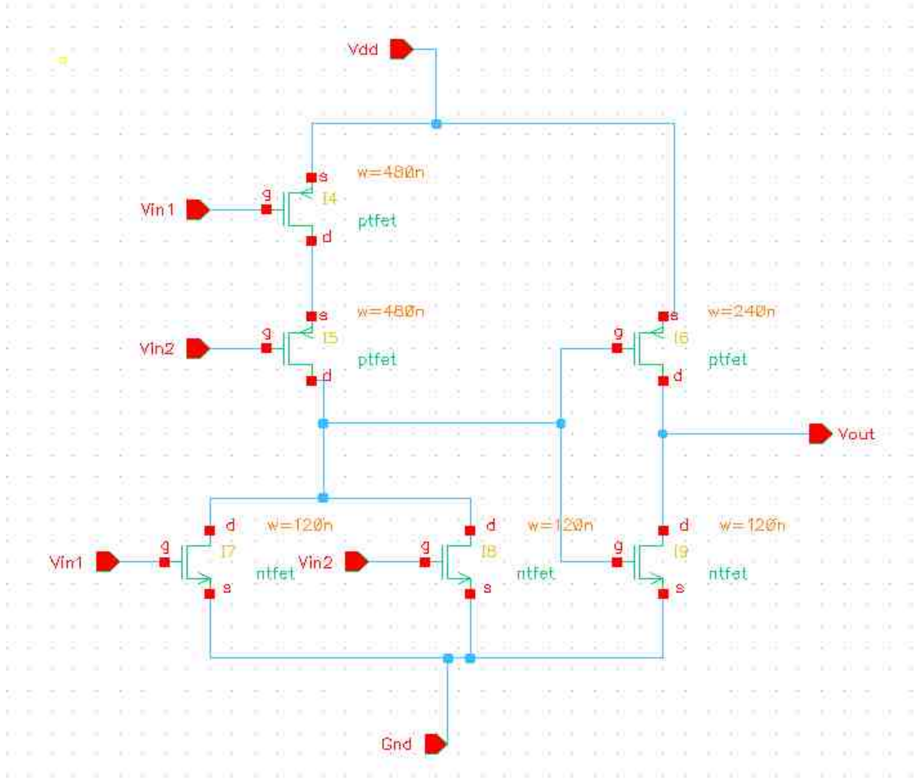

Fig. 5.11. GaN TFET OR Schematic.

5.2.6 Operational Transconductance Amplifier

Using the GaN TFET model, an Operational Transconductance Amplifier (OTA) was designed. The OTA produces an output current based on the differential input voltage. It acts as a voltage controlled current source where the output current can be controlled through the input current I_{bias} to the amplifier. The output of the OTA is given by the following equation:

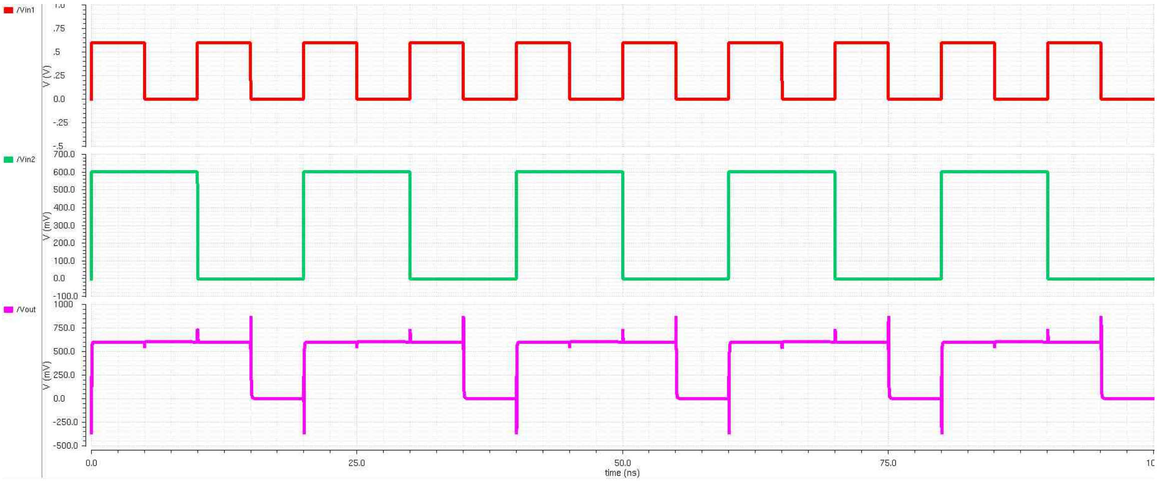

Fig. 5.12. GaN TFET OR Transient Response.

$$Out = (InP - InN) \times gm \tag{5.2}$$

where, InP and InN are the input voltages, gm is the transconductance gain of the amplifier, and Out is the output current of the amplifier which is controlled by I_{bias}.

The OTA's functionality, magnitude and phase response is determined by the use of a test bench, custom AC voltage source, and Common Mode Feedback (CMFB) using resistors. Figure 5.13 shows the schematic of an ideal common mode feedback circuit. Figure 5.12 shows the schematic of the OTA and Figure 5.14 shows the test bench used for the simulation of the OTA. The AC test output curves (magnitude and phase) of the OTA are included in Figure 5.15. The unity gain frequency (FUG) and the static power were analyzed by using ADE in Cadence Virtuoso (i.e., AC and DC analysis.)

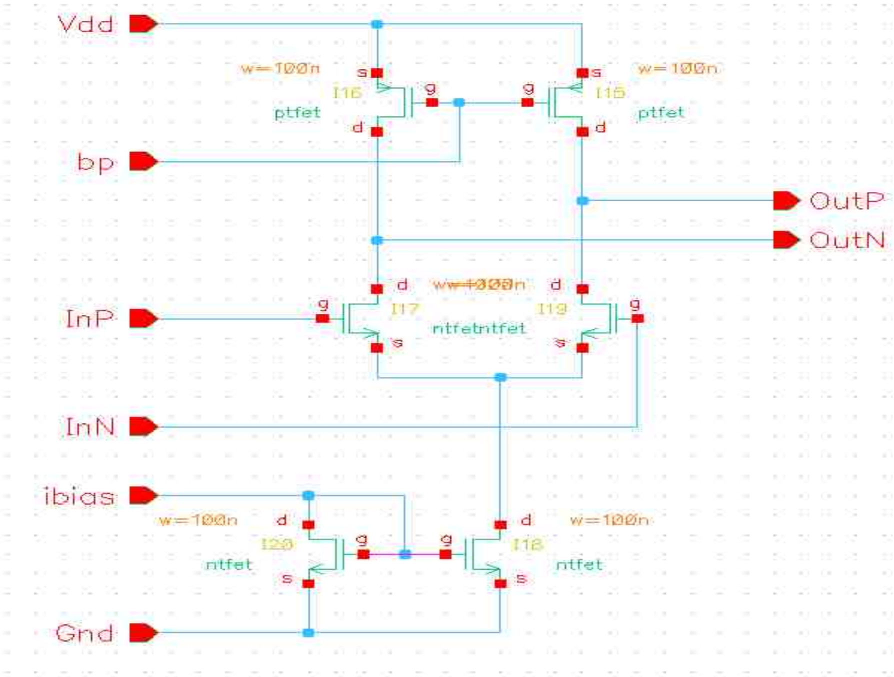

Fig. 5.13. GaN TFET Operational Transconductance Amplifier.

5.2.7 Results and Discussion

A figure of merit (FOM) calculated as the product of static power and delay was adopted. The delay was calculated by using the linear delay model [10] given by the following equation:

$$delay = N \times F^{1/2} + P \tag{5.3}$$

where, N is the number of stages, P is the parasitic delay and F is the effort delay. The latter is defined as:

$$F = G \times H \times B \tag{5.4}$$

where, G is the logical effort, H is the fan-out or electrical effort, and B is the branch effort.

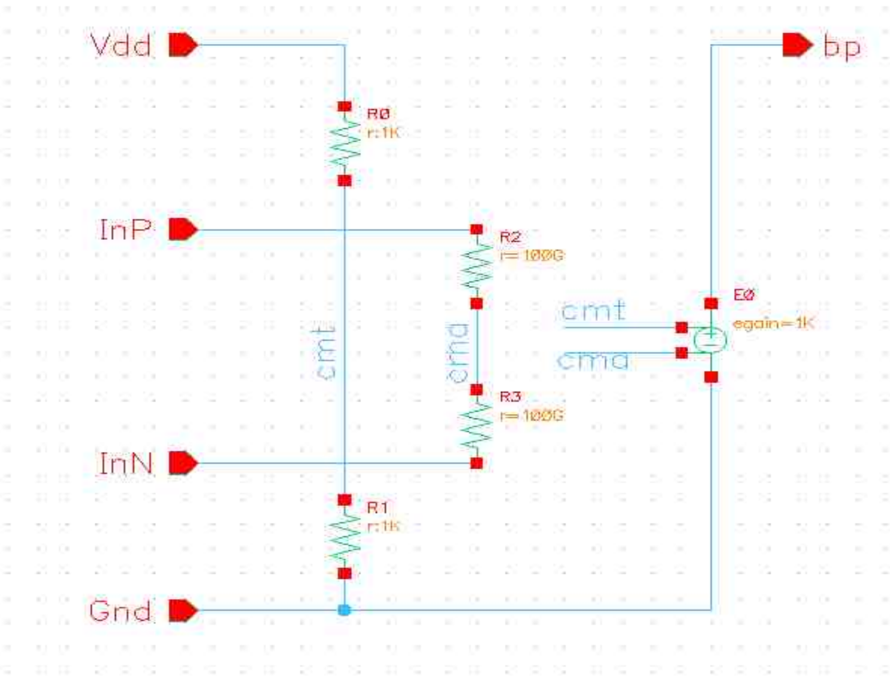

Fig. 5.14. GaN TFET Common Mode Feedback.

The logical effort G is the ratio of input gate capacitance of the circuit and the input gate capacitance of an inverter that can drive the same output current [10]. For an unskewed inverter with PTFET width equal 2 and NTFET width of 1, G=1. In the case of an unskewed NAND2 with PTFETs and NTFETs sized at width of 2, G=4/3. For an unskewed NOR2 with PTFETs sized at 4 and NTFETs sized at 1, the value of G=5/3. The electrical effort H is the ratio of the output capacitance to the input capacitance of the circuit. Assuming an input capacitance of 1 unit, the electrical effort H for and inverter, a NAND2 and a NOR2 is 3, 6 and 6, respectively. There are no branches associated with a single cell. Therefore, the branching effort B for all three cells is set to 1. The parasitic delay is the ratio of the output drain diffusion capacitance of a circuit to the output drain diffusion capacitance of an inverter that can deliver the same output current [10]. For the unskewed inverter, NAND2 and NOR2 gates, the parasitic delay is 1, 2 and 2, respectively.

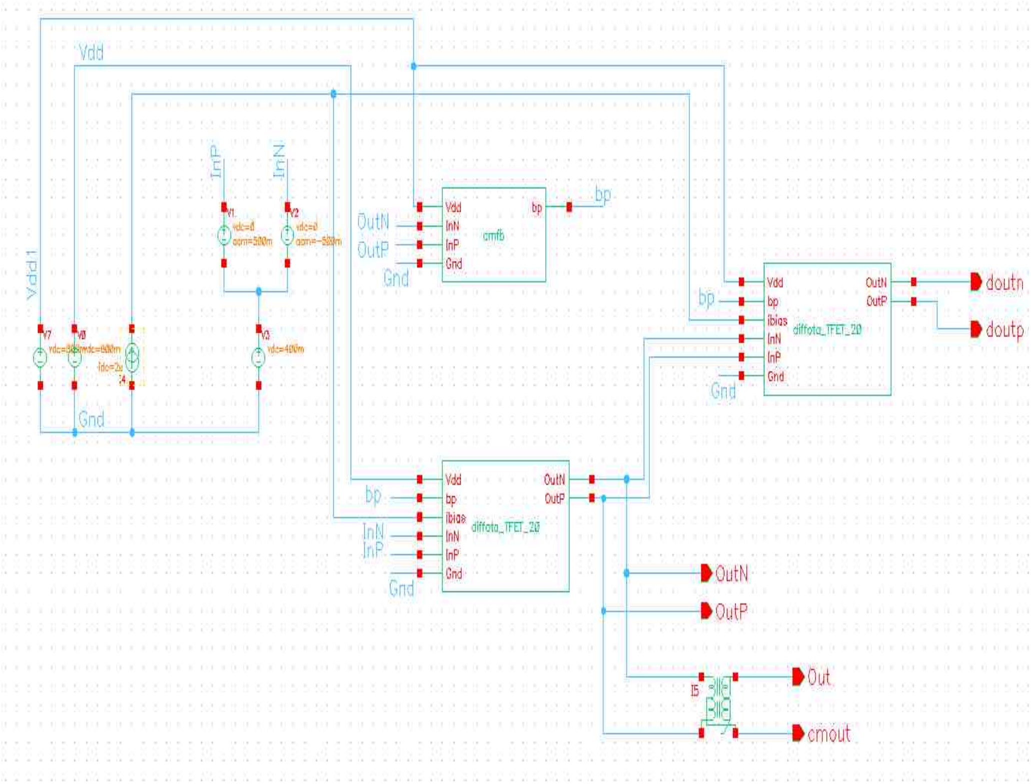

Fig. 5.15. Test Bench for simulating OTA.

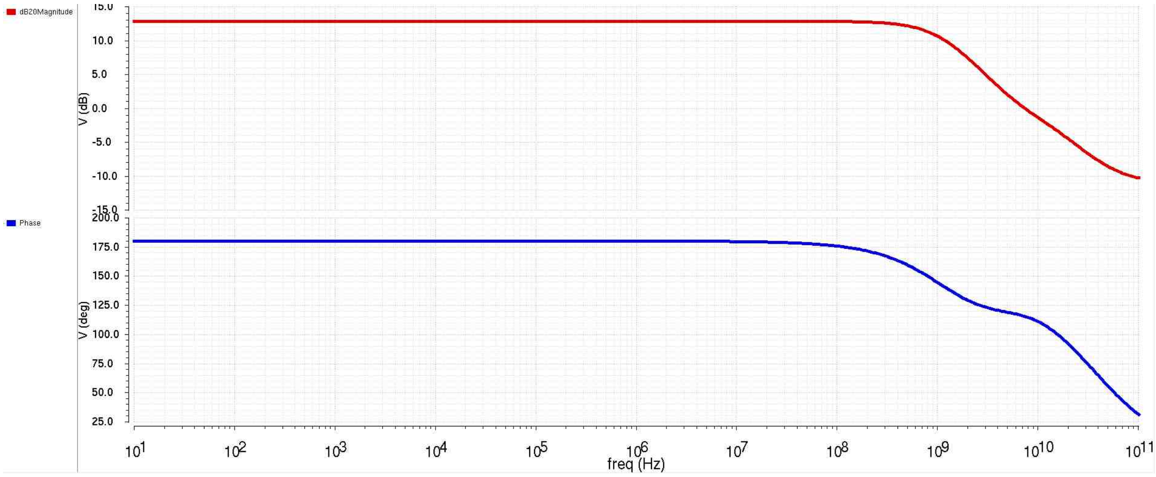

Fig. 5.16. Magnitude and Phase Response of the OTA.

Based on the above parameters, the delays for the inverter, NAND2 and NOR2 gates were calculated to be 4, 10 and 12 units, respectively. For the digital circuits, V_{DD} was set to 0.6V and for the analog circuits, V_{DD} was set to 0.8V. The delay components are summarized in Table 5.1 and the FOM for the digital cells is given by equation 5.5.

$$FOM = P_{STATIC} \times delay \tag{5.5}$$

Table 5.1.
Linear Delay Components.

Circuit	G	H	B	F	P	**delay**
Inverter	1	3	1	3	1	4
NAND2	4/3	6	1	8	2	10
NOR2	5/3	6	1	10	2	12

The results in tables 5.2, 5.3 and 5.4 show that not only is the gate width to length ratio is an important factor in the value of the leakage current, but so is the threshold voltage for the inverter, NAND2 and NOR2 gates. The CMOS devices with gate length of 100nm but lower threshold voltage have a higher leakage current when compared to other CMOS-based technology nodes for all three logic gates.

The GaN TFETs have the lowest threshold voltage among all the devices that were analyzed in tables 5.2, 5.3 and 5.4. Moreover, the leakage current is higher for the GaN TFETs than for most of the CMOS-based circuits. The exception is the 100nm CMOS device which has a higher threshold voltage than the GaN TFETs. The supply voltage used for the simulation of all of the above circuits is 0.6V in order to allow for comparison with the GaN TFET model. This voltage is lower than the threshold voltage for the 45nm technology node.

Table 5.5 shows the figure of merit (FOM) for different OTA designs. The FOM for this particular cell is the static power of the TB_diffora divided by the product of

Table 5.2.
GaN Inverter Results.

Device	**W/L (nm)**	**P-Type V_{th} (V)**	**N-Type V_{th} (V)**	I_{OFF} **(pA)**	P_{STATIC} **(pW)**	**FOM** (10^{-12})
TFET GaN	100/20	-0.15	0.15	432	259	1040
CMOS 45nm	225/45	-0.56	0.61	6.97	4.18	16.7
CMOS 100nm	500/100	-0.27	0.20	11800	7070	28300
CMOS 150nm	750/150	-0.49	0.53	3.01	1.81	7.24
CMOS 280nm	1440/280	-0.57	0.49	6.01	0.361	1.44

$A0$ and the unity gain frequency FUG as shown by equation 5.6. Table 5.5 indicates that the TFET-based cell has the lowest ratio of static power to the product of $A0$ and FUG.

$$FOM = P_{STATIC}/(A0 \times FUG) \tag{5.6}$$

The analysis of GaN TFET-based digital and analog circuits presented in this thesis shows that there are advantages to these circuits compared to CMOS-based circuits. They have lower leakage current when compared to CMOS-based devices with similar threshold voltages. In addition, the GaN TFET device used in this thesis has a very low threshold voltage of 0.15V, and as such, the leakage current was higher compared to CMOS-based circuits with higher threshold voltage. Lower threshold voltages have an added advantage for circuits that are in the critical paths of a system. Indeed, as the threshold voltage decreases, current increases when the

Table 5.3.
GaN NAND2 Results.

Device	**W/L (nm)**	**P-Type** V_{th} **(V)**	**N-Type** V_{th} **(V)**	I_{OFF} **(pA)**	P_{STATIC} **(pW)**	**FOM** (10^{-12})
TFET GaN	100/20	-0.15	0.15	752	451	4510
CMOS 45nm	225/45	-0.56	0.61	7.36	4.42	44.2
CMOS 100nm	500/100	-0.27	0.20	8320	5030	50300
CMOS 150nm	750/150	-0.49	0.53	27.7	16.6	166
CMOS 280nm	1440/280	-0.57	0.49	1.20	0.722	7.22

transistor is switched-on. However, when the transistor is switched-off, leakage current also increases [10]. Since the GaN TFETs have lower leakage, using this device for the design of components in the critical path should be considered. Another application area for GaN TFETs is low-power memory design. This is particularly true for SRAM cells [22], where the 30mV/decade switching speed becomes an attractive feature because of the resulting lower static power consumption by the cell.

5.3 8T-SRAM Cell

SRAM is predominantly used in integrated circuits design in part due to its fast access times (i.e., higher speed) and high reliability. SRAM-based memory is primarily used in the upper levels of the memory (e.g., registers and caches) which require high access rates. The SRAM stores data which is most frequently accessed by the

Table 5.4.
GaN NOR2 Results.

Device	**W/L (nm)**	**P-Type** V_{th} **(V)**	**N-Type** V_{th} **(V)**	I_{OFF} **(pA)**	P_{STATIC} **(pW)**	**FOM** (10^{-12})
TFET GaN	100/20	-0.15	0.15	863	518	6220
CMOS 45nm	225/45	-0.56	0.61	13.9	8.36	100
CMOS 100nm	500/100	-0.27	0.20	23600	14100	169000
CMOS 150nm	750/150	-0.49	0.53	31.5	18.9	227
CMOS 280nm	1440/280	-0.57	0.49	1.67	1.00	12.0

processor. As such, the speed and the reliability of the cell is vital for data delivery to the processing unit. SRAM cells are volatile memory which means that they will retain a single bit of data as long as power supply is switched-on. Typically, there are four types of SRAM cells which use 4, 6, 8, 10 transistors [23]. In this thesis, an 8-transistor design is used. The 8T-SRAM cell is designed using two sets of access transistors on either side of the cell instead of one access transistor. The access transistor provides read/write access to the cell from the bit lines. Two access transistors are used because the TFETs are uni-directional devices [18], and thus one access transistor provides read access to the cell, and the second access transistor provides write access to the cell. The idea is to write a single bit of data to the cell on cycle 1, hold the data in the cell on cycle 2, and read the data from the cell on cycle 3 in a synchronous fashion. In order to measure the leakage current and the static power, the word line (WL) is first switched-ON and data is written to the cell. This action

Table 5.5.
Operational Transconductance Amplifier Results.

Device	**W/L (nm)**	**P-ype V_{th} (V)**	N-Type V_{th} (V)	P_{STATIC} **(nW)**	**A0**	**FUG (GHz)**	**FOM** (10^{-15})
TFET GaN	100/20	-0.15	0.15	114	8.148	3.861	3620
CMOS 45nm	225/45	-0.56	0.61	150	8.992	0.874	0.0191
CMOS 100nm	500/100	-0.27	0.20	682	3.871	0.826	0.213
CMOS 150nm	750/150	-0.49	0.53	152	11.19	0.324	0.0419
CMOS 280nm	1440/280	-0.57	0.49	464	8.987	0.040	1.28

constitutes writing a single bit of data to the cell. Next, the WL is turned-OFF resulting in the data being stored in the cell, a state which is called the hold state of the cell. The leakage current and static power is analyzed in the Analog Design Environment (ADE) in Cadence Virtuoso while the cell holds a single bit of data. If the WL is raised again after the bit lines are kept at a pre-charged value, the data stored in the cell is transferred to the bit lines in an operating mode called the read. The cell needs to be carefully sized in order to demonstrate the read, hold and write of data. When the storage node Q stores the data value 1, the storage node Q_bar stores the complement of the data value 1, which is 0. This means that the Pull-down network (PDN) NTFET associated with Q_bar is turned-on and the Pull-up network (PUN) PTFET associated with Q_bar is turned-off. At the same time, the PDN NTFET associated with Q is turned-off and the PUN PTFET associated with Q is turned-on. This persists as long as the WL is turned-off and there is no interruption in the power supply to the cell through the voltage source V_{DD}. The leakage current and the static power of the cell are thus concerned with this combination of inverters that hold the data in the cell.

5.3.1 SRAM Bitcell

The unit cell is the most important component of an SRAM-based array as it determines the overall area of large arrays. The unit cell consists of 8 transistors in total; two inverters banked by four access/pass transistors. The inverter consists of a PTFET as the PUN and an NTFET as the PDN. The PTFET device was designed by reversing the direction of the current flow of the NTFET device designed by Notre Dame University [7]. The body terminal is not shown in the symbol as it is grounded [7]. As mentioned above, there are two access transistors on both sides of the unit cell that provide access to and from the cell. On each side, two inward facing NTFET access transistors are used: one with the source connected to the bit line and the drain connected to the storage node (for writing to the cell), and another with the drain connected to the bit line and the source connected to the storage node (for reading from the cell). The reason two inward facing access transistors are used to access the unit cell is because TFETs are uni-directional devices [24]. This means that current passes from source to drain, but not from drain to source in a normal operating state. Hence, one of the two access transistors allows data to be written to the cell, while the second access transistor on the same side of the unit cell allows data to be read from the cell [22].

Logically, the SRAM unit cell consists of two storage nodes with one node Q storing a single bit (e.g., logic 1) while its complement (e.g., logic 0) is stored in the other storage node Q_bar. Figure 5.17 shows the Cadence Virtuoso capture of an 8T-SRAM cell designed using GaN TFET devices. Figure 5.18 shows the Cadence Virtuoso capture of a 6T-SRAM cell designed using 45nm CMOS devices.

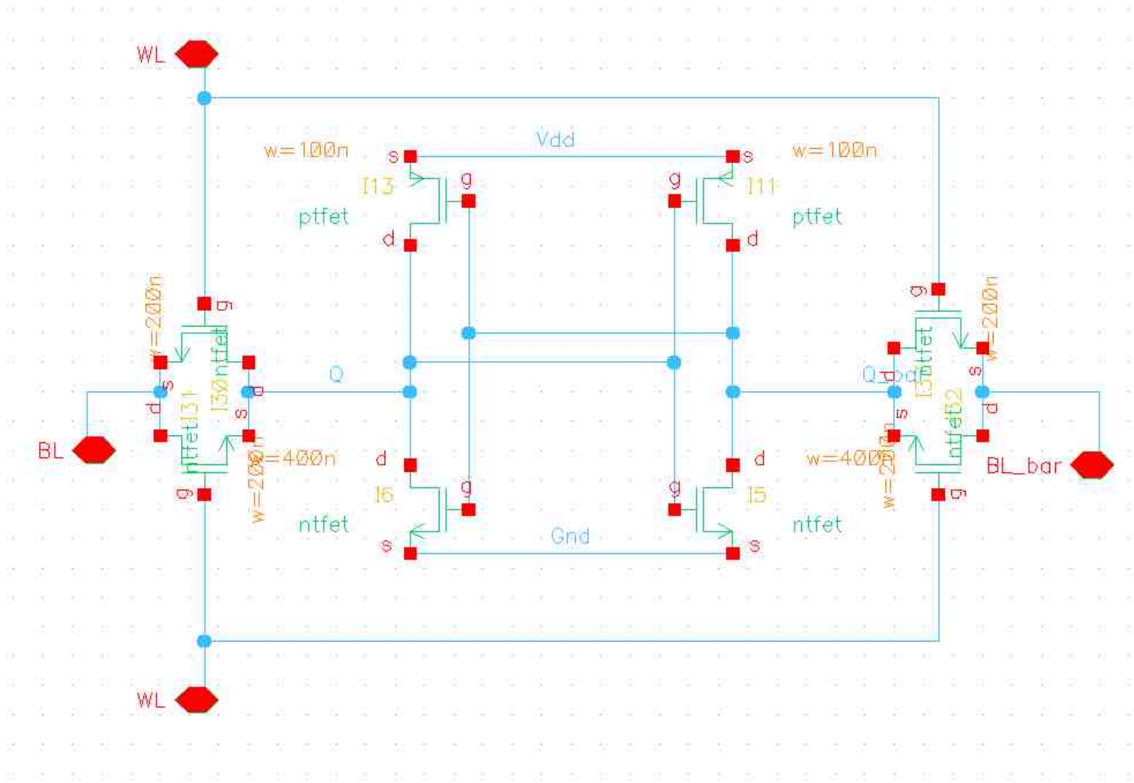

Fig. 5.17. TFET 8T-SRAM cell.

The schematic of the 8T bit-cell was designed in Cadence Virtuoso. The hold, read and write operations of the cell were verified using a transient analysis in ADE. The design of the SRAM cell and its associated peripheral circuitry is constrained to the schematic level design because Notre Dame University's GaN TFET Verilog-A model only supports circuit design [7]. The library does not include the cell layouts or the design rules needed to support the physical design. These are necessary for fabrication, and post fabrication validation. These activities are part of the future work and will be completed when the physical design becomes available.

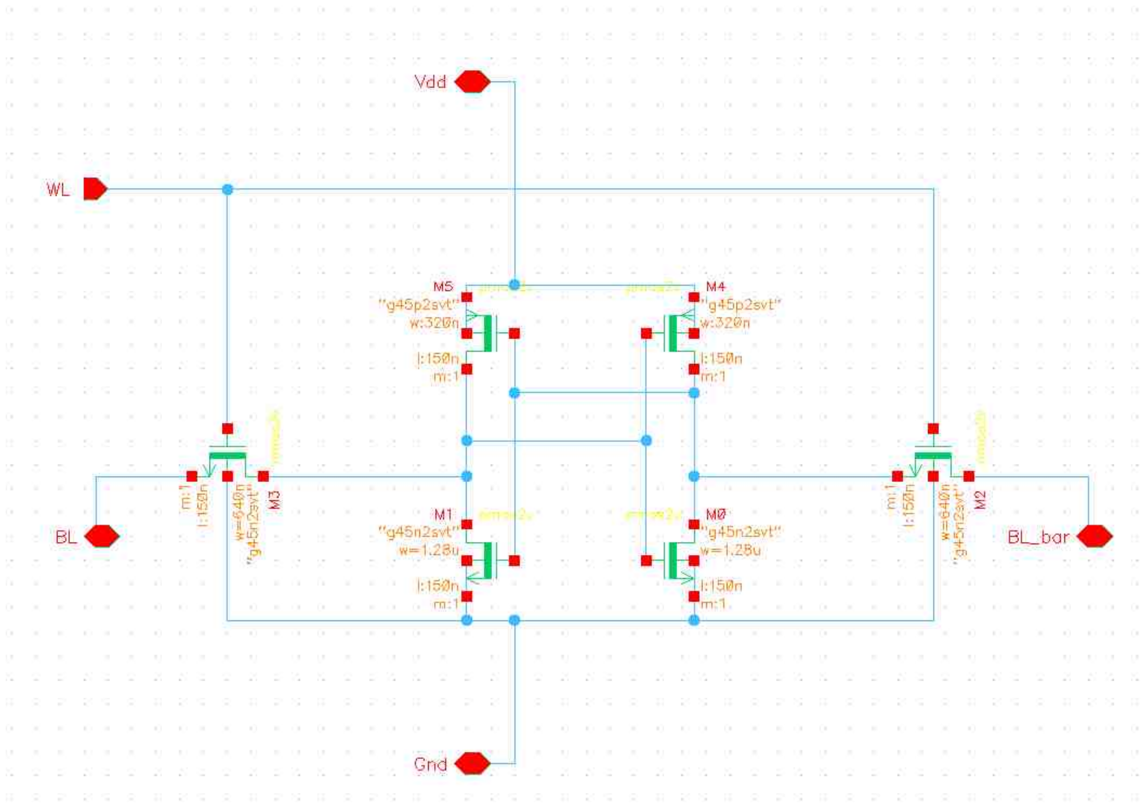

Fig. 5.18. CMOS 6T-SRAM cell.

5.3.2 Clocks

The write, hold and read to/from multiple cells require synchronization. For this purpose, the design uses two clocks Phi_1 and Phi_2 with duty cycles 60 % and 40 %, respectively. Two non-overlapping clocks can be designed using the Vpulse voltage source in Cadence Virtuoso. Clock Phi_1 is set to a clock cycle of 10ns with a pulse width of 4ns and a delay of 1ns. Clock Phi_2 is used exclusively for pre-charge. It has a clock cycle of 10ns and clock pulse width of 6ns. In this design, the compliment of clock Phi_2, clock Phi_2B, is used along with clock Phi_1 to demonstrate synchronous write and read operations. The two clocks are shown in Figure 5.19.

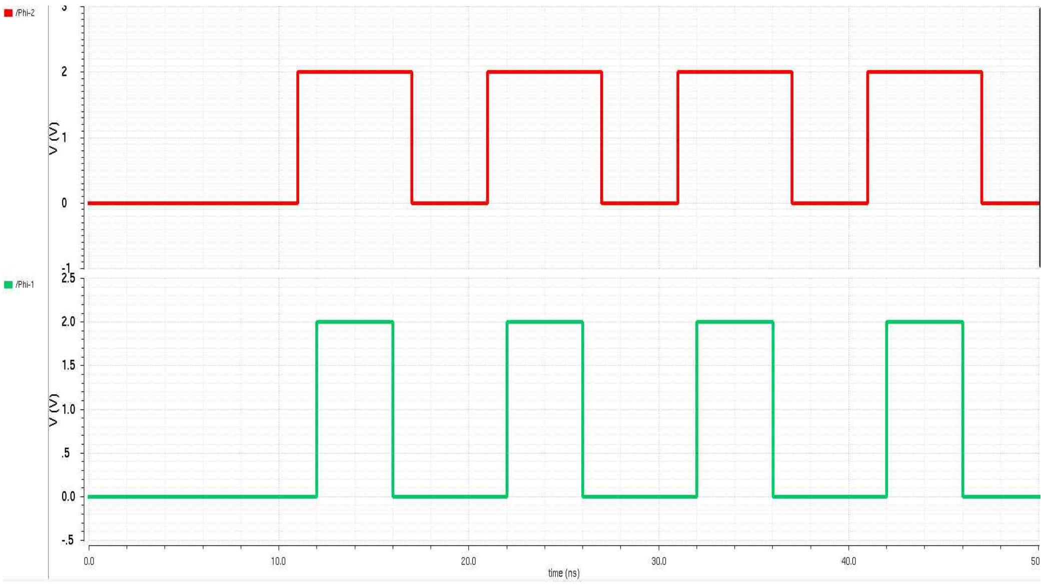

Fig. 5.19. Clocks Phi_2B and Phi_1.

5.3.3 Pre-charge Circuitry

Prior to reading from or writing to the cell, the bit lines are pre-charged to the value of V_{DD} (or $V_{DD}/2$) [16]. The bit lines are pre-charged prior to every read operation in order to avoid a read error. If a memory cell storing a logic 1 is read first followed by read of a memory cell storing logic 0, the bit line would remain high as a result of the first read for certain amount of time. This is due to the capacitance of the bit lines. In large memory arrays, the bit lines capacitance will be very large and the discharging time for the bit lines is long. Waiting for the bit line capacitance to discharge to a logic 0 before the next read operation can be issued increases the time of an SRAM read operation. In fact, if the second read operation is performed prior to the discharge of the bit lines, it can lead to the incorrect read of a logic 1 instead of logic 0. Figure 5.20 shows the pre-charge circuitry used for the bit lines using PTFET transistors. This circuitry is used to drive the bit lines to V_{DD} when clock Phi_2B is at logic 0.

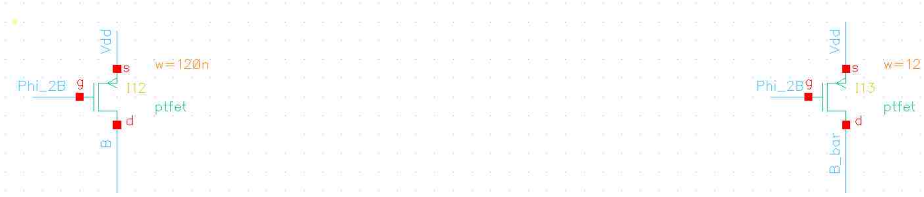

Fig. 5.20. Pre-charge Circuitry.

5.3.4 Row Address Decoder

In addition to the read/write circuitry, a buffer designed by using the AND gates shown earlier in Figure 5.7 is used for the read and write operations. These buffers are also used at the address lines (Addr) to assert the word line, and hence they act as the row decoder. When the address corresponding to a particular cell is AND-gated with the clock Phi_1, the word line corresponding to that particular cell is raised, allowing access to or from the cell.

For the write circuitry, the buffers are used to write data to the cell when the clock Phi_1 and the write enable (WR_en) are both held high. For the read circuitry, the clock Phi_1 and read enable (Rd_en) are held high in order to transfer the value stored in the cell to the bit lines. Only one word line in an entire column should be raised as the cells in the same column share the same bit lines.

5.3.5 Column Address Decoder

Multiplexers are used to read data from different columns when asymmetrical arrays are used (e.g.: 8x4 arrays). To read each bit of a word, the output of the sense amplifier associated with each column are used as the input signals for the multiplexer. The selector signals of the multiplexer select the data from different column addresses. Multiplexers can be designed using tristate inverters and transmission gates. Figure 5.21 shows the schematic of a transmission gate-based 2x1 multiplexer.

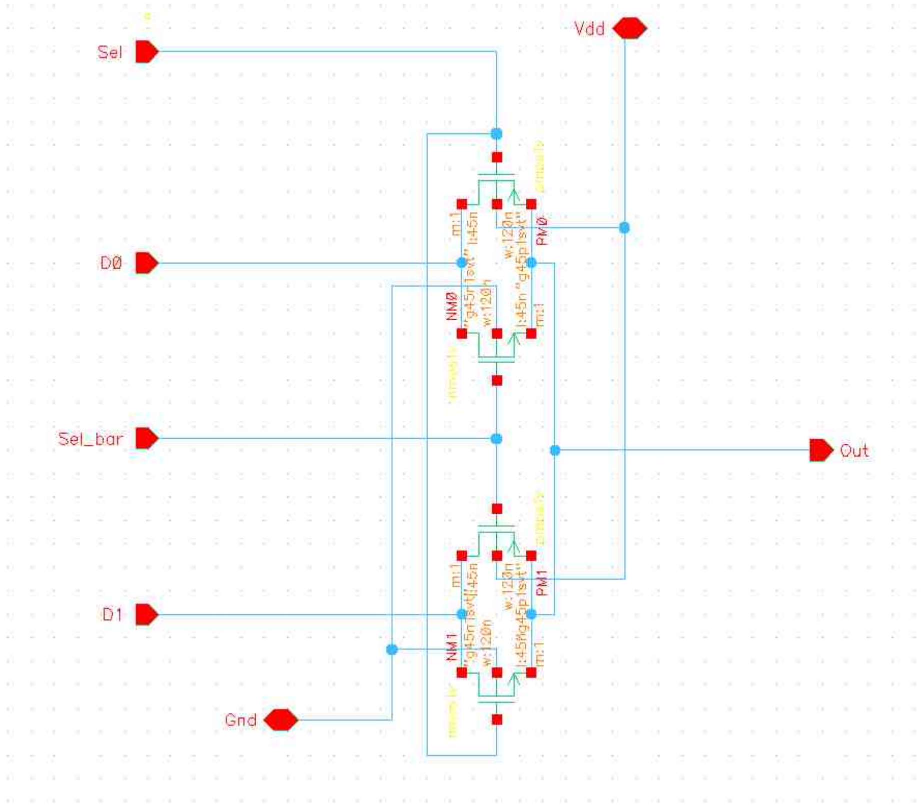

Fig. 5.21. CMOS 2x1 Multiplexer.

5.3.6 Sense Amplifier

In large memory arrays, the bit line capacitance is very large due to the increased wire capacitance and the added capacitance of the access transistors on the same column. Therefore, during the read operation, the bit lines take a long time to drop from their pre-charged value to a logic zero. To circumvent the slow discharge of the bit lines, a sense amplifier can be used as a read assist. The sense amplifier used in the design is a differential voltage sense amplifier (Figure 5.22). This amplifier senses a change in voltages in the two bit lines, BL ($In1$) and BL_bar ($In2$) and amplifies this signal. It is important to consider the sense amplifier as an analog circuit. The output of the sense amplifier is given by the following equation:

$$Sense\_out = (In1 - In2) \times A \tag{5.7}$$

where A is the gain of the amplifier and $In1$ and $In2$ are the voltages from the bit lines BL and BL_bar.

The sense amplifier allows for a faster transfer of the value read from the cell to the bit lines during the read operation faster. However, the NTFET transistors, which take in the bit line voltages from the column need to be adequately sized and should have high threshold voltage tolerance in order to sense the bit line values rapidly.

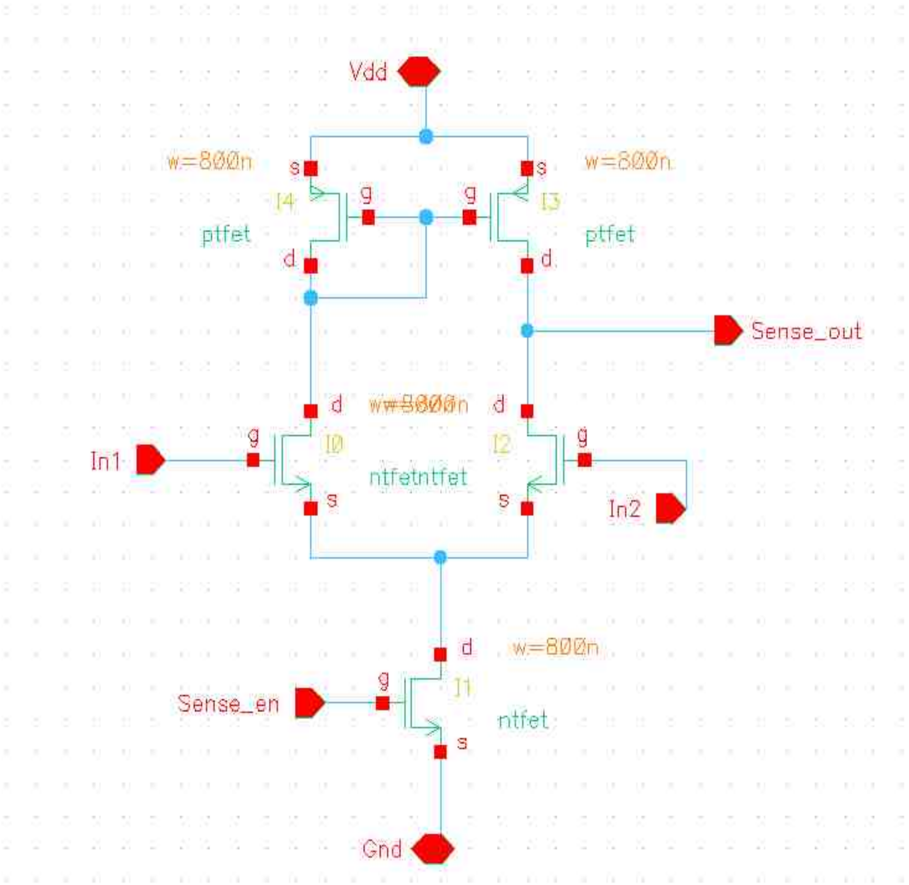

Fig. 5.22. TFET Differential Voltage Sense Amplifier.

5.3.7 Operating Modes

The SRAM cell has essentially three operating modes: read, hold and write. To write a single bit of data into the cell, the bit lines are first pre-charged to the value of V_DD. Assume that logic 1 is stored in the storage node Q and thus its complement logic 0 is stored in storage node Q_bar. In Cadence Virtuoso, the Initial Condition feature in the ADE can be used to initialize the nodes to a particular voltage. To write logic 0 to Q, WR_en is asserted and a logic 1 is supplied through the voltage source (Data). The bit line BL is set to logic 0 from its pre-charged value. The bit line BL_bar is kept at its pre-charged value. Then if the WL corresponding to the cell is raised by asserting the address (Addr) associated with the cell to a logic high, the logic 0 from the bit line BL will be written to the storage node Q and a logic 1 is written to Q_bar through the inverter combination in the cell.

If a logic 1 needs to be written to the cell, assuming Q currently holds a logic 0 and Q_bar a logic 1: Both bit lines are pre-charged, the bit line BL_bar is set to ground and the bit line BL is kept at its pre-charged value. This will write a logic 0 to Q_bar and a logic 1 to Q.

The bit lines need to be grounded in order to supply them with logic 0s since the access transistors and the NTFET PUN devices pass a strong 0 and a weak 1. In addition for the read process, the PUNs are sized to avoid a read upset. The access transistors can be made of PMOS devices. However, the width of the p-type access transistors is larger than the n-type devices. This is because the majority carriers in p-type devices are holes, and holes have half the mobility of electrons [10]. Hence, the p-type devices need to be made twice as wide as n-type devices in order to pass the same amount of current. If the access P-type devices are made twice as wide as NMOS access devices, the PUN PTFET and PDN NTFET also needs to be made larger. This will increase the overall area of the unit cell. Since the ultimate goal is to make the unit cell as small as possible, NTFET devices are more suitable for the access logic.

The cell is in a hold mode when the WL that provides access to the unit cell is at a logic 0. The WL is attached to the gates of the access transistors which turn-on/off the NTFET access transistors. If there is relevant data already stored in the cell, this data will be preserved as long as the supply voltage is maintained and the two inverters sustain the data stored in the storage nodes. Typically, the cell is most stable when holding data.

For the read operation, assume first that logic 1 is stored in the storage node Q and a logic 0 is stored in the storage node Q_bar. Similar to the write operation: the bit lines are first pre-charged to V_{DD}, the read enable (RD_en) is held high, and the WL is raised by asserting the corresponding address of the cell, which will switch-on the access transistors. The logic 0 stored in Q_bar will be transferred to the bit line BL_bar while the bit line BL will be held at V_{DD}. The sense amplifier circuitry is not necessary for reading a single bit. However, in large arrays, all the unit cells in the same column share the same bit lines. This increases the capacitance of the bit lines, and hence, logic 0 will not be immediately transferred to the bit line. Since it is desirable to read the data stored in the cell as fast as possible, sense amplifiers are used to sense the smallest change in the bit line voltage values. They then amplify this change to produce a logic 0 or logic 1 at the output.

When a read operation is performed, there is a chance of writing logic 1 accidentally into the storage node; an event called as a read upset. To avoid this event, the transistors need to be carefully sized. Generally, the PUN PTFET are sized the smallest, the PDN NTFET sized the largest, and the access transistors sized medium [16]. For an error free read and write operation, the associated circuitry needs to be sized as well as the SRAM cells. Generally, the ratio of NTFET: PTFET: Access transistors are sized at gate widths 1: 4: 2 or 1: 4: 3 depending on the technology node and the fabrication process. In the proposed design, the sizing ratio of the PTFET: NTFET: Access transistors are set to 1: 3: 2.

5.3.8 8T-SRAM Simulation

In order to verify the logic design of the SRAM cell, the circuit was designed in the Cadence Virtuoso Schematic L-Editing environment using a custom library of TFETs cells derived from the Notre Dame University's GaN TFET device model [10]. The design was then verified using the Spectre simulation tool after setting the design conditions in the Analog Design Environment (ADE) by using custom voltage sources and a transient response analysis. The rise time and the fall time for the input signals were set to 10ps and the circuit was sized using the ratios mentioned earlier. The schematic design of the setup that was used to verify the design is provided in Figure 5.23.

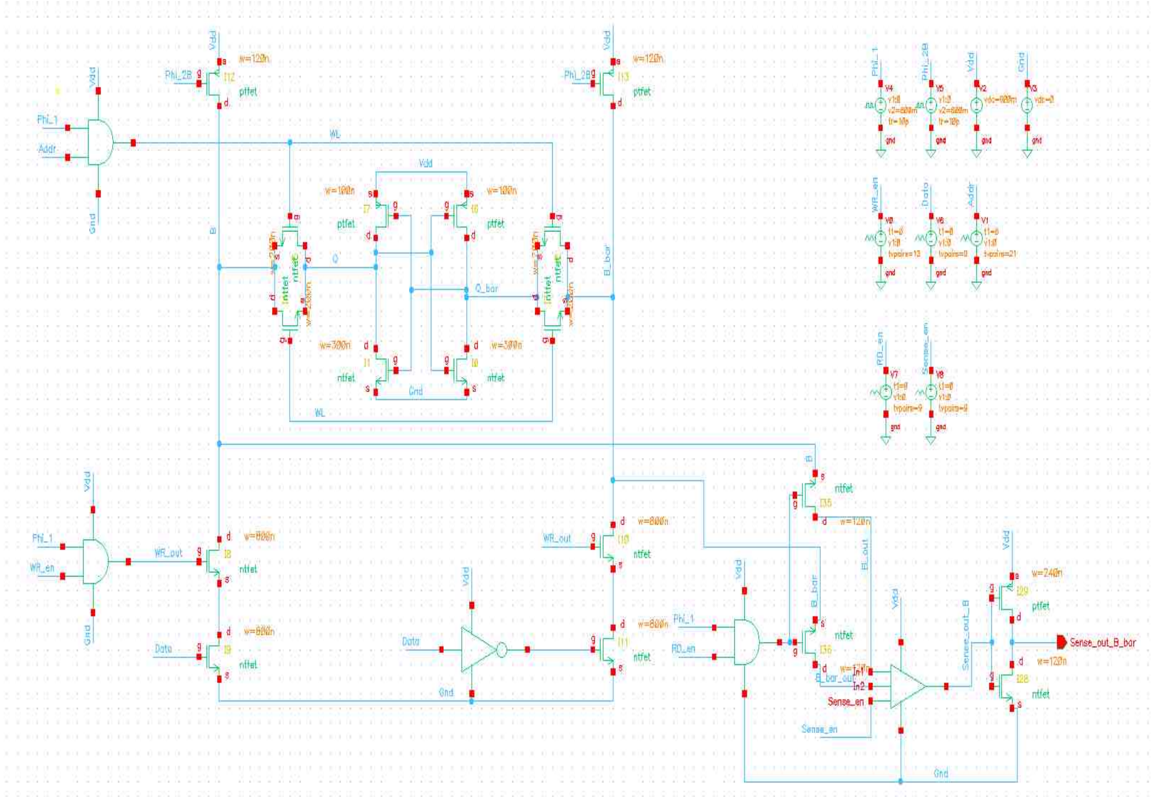

Fig. 5.23. Simulation of an 8T-SRAM Bitcell.

The simulation uses clocks Phi_2B and Phi_1 with clock phases 60 % and 40 %, respectively. The clock period for Phi_2B was set to 4ns over a clock pulse width of 10ns. The clock period for Phi_1 was set to 6ns over a clock pulse of 10ns. Introducing a 1ns delay in Phi_1 makes these two clocks non-overlapping.

Assume that logic 1 is stored in Q and logic 0 is stored in Q-bar. The V_{DD} for the design was set to 0.6V. After pre-charging the two bit lines to 0.6V by turning-on the two pre-charge PTFETs, the first step is to write logic 0 to the unit cell storage node Q, and then hold the data for a clock cycle before reading the data in the following cycle. It is important to ensure that writing and reading are not performed on the same cycle. Therefore, WR_en is given as an input to the AND gate (buffer) with the clock Phi_1. When the data source is at logic 1 along with WR_en enabled, the two NTFETs that drive the bit line B to the ground are turned-on and the bit line B is set to 0V. An inverter inverts the logic 1 of the input Data to logic 0 which turns-off the bottom NTFET in the series that drive the bit line B_bar.

The next step is to turn on the WL by asserting the address Addr to 0.6V and passing the clock Phi_1 as input to the buffer. In an array, each address belongs to a different unit cell. Therefore, writing must be performed in a synchronous fashion on different cycles for each cell in order to avoid incorrect writes. Once the logic 0 has been written to Q from the bit line B, the inverter associated with Q_bar supplies 0.6V to the node Q_bar via V_{DD}. Subsequently, the 0V stored in node Q is re-enforced by the inverter associated with Q from Gnd. The timing diagram for writing a logic 1 to Q (and thus a logic 0 to Q_bar) is shown in Figure 5.24.

By turning-off the WL, and disabling the WR_en and RD_en, the value will be held in the unit cell. The data being held in the cell can be observed in cycle 2 of Figure 5.24.

To read from the cell, the Addr corresponding to the unit cell is turned-on. This will raise the WL and drive the logic 0 back to the bit line B. However, the bit line B_bar will remain at logic 1 due to the pre-charge. By switching-on the RD_en and the Sense_en to turn-on the sense amplifier, the read operation can be performed next.

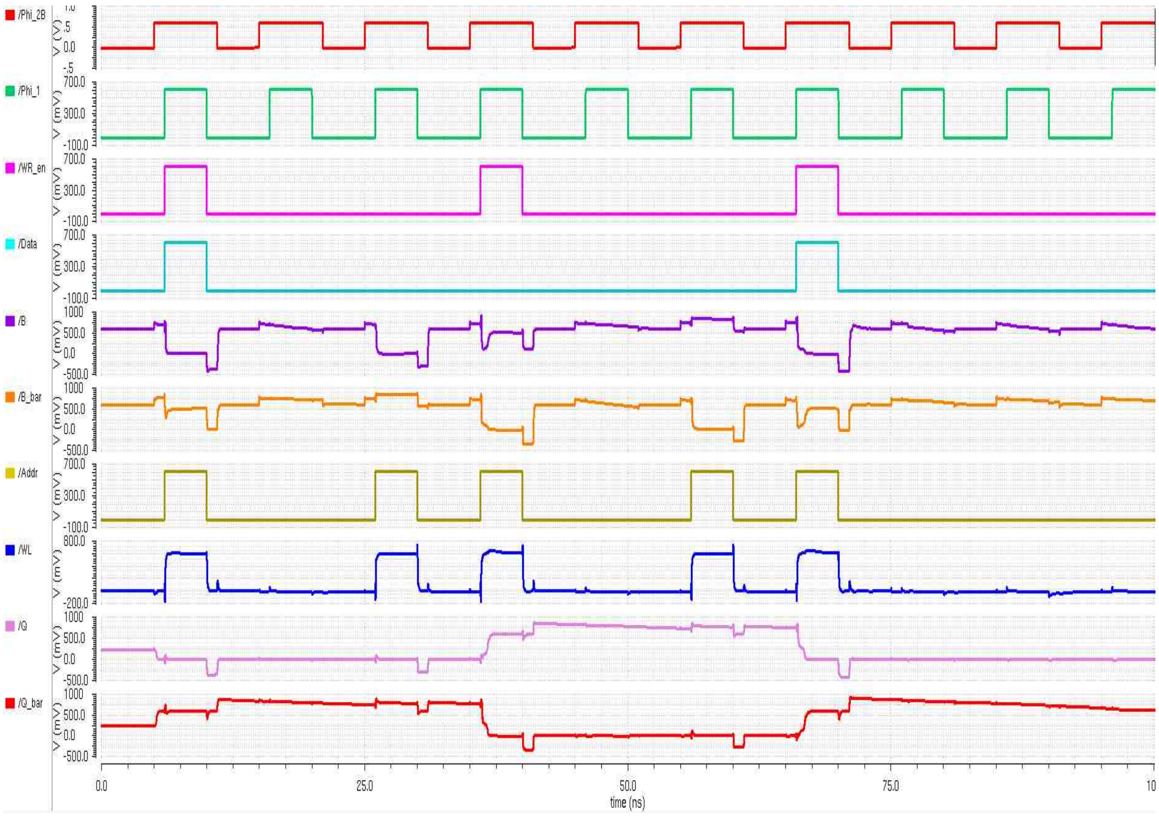

Fig. 5.24. Write Simulation.

While reading, it is important to turn-OFF the WR_en so as not to accidentally write a value to the cell while the read is being performed. The simulation waveform for reading from the cell is shown in Figure 5.25 and on cycle 3.

5.3.9 Static Noise Margin (SNM)

The Static Noise Margin (SNM) of an SRAM cell represents the stability of the cell. It shows the ability of the cell to resist to the flipping of the value stored in the cell due to external voltage noise. It is represented by the voltage transfer characteristics of the first inverter and the inverse voltage transfer characteristics of the second inverter in the cell [25]. The diagonal of the largest square that can be drawn within the lobe

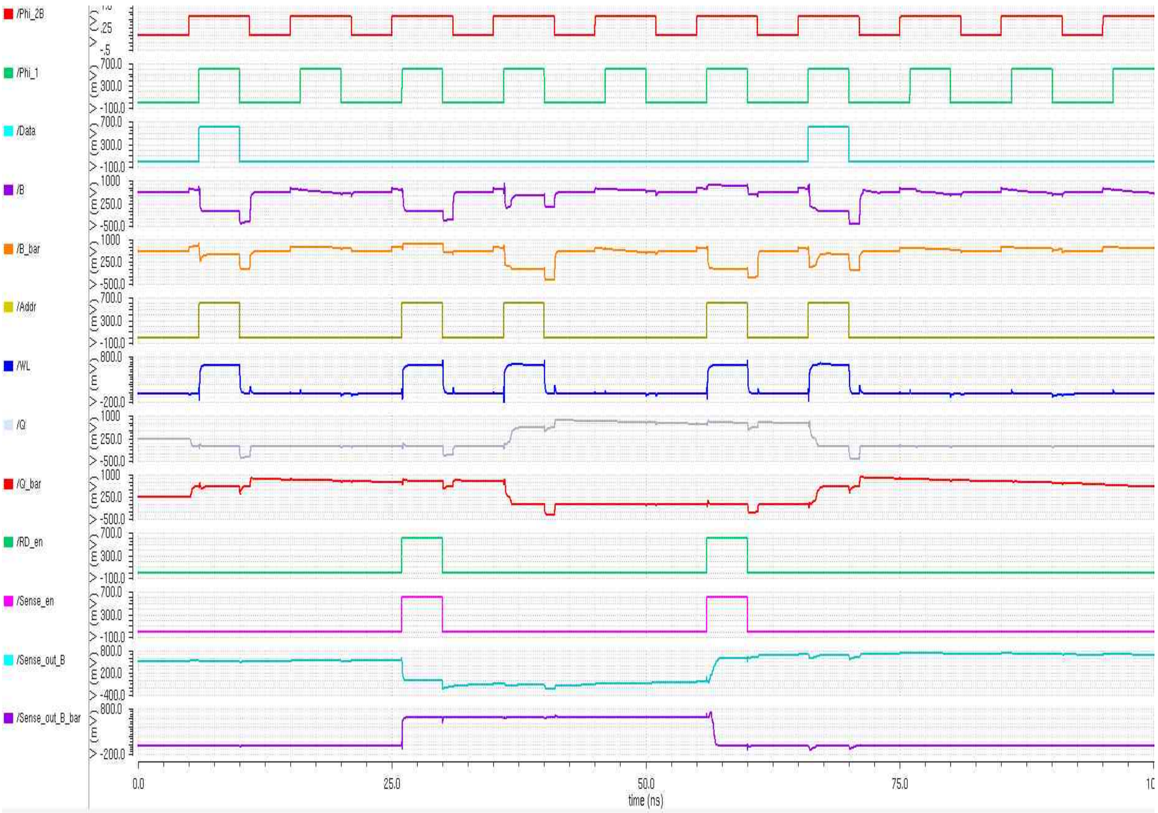

Fig. 5.25. Read Simulation.

of the butterfly curve is the SNM of the cell. The larger the SNM, the more stable the cell. Large noise margin help avoid the cell from flipping values due to internal changes in voltage fluctuations in V_{DD} or deu to external changes in voltage values of the bit lines [25]. The SNM is different during different operating modes of the cell. The write SNM is the minimum voltage that needs to be applied to the bit lines in order to flip the state of the cell. To calculate the SNM of a write operation, the storage node Q was initialized to 0.6V, the bit line B was set to 0V and the WL was asserted, allowing for the value in the bit line B to be written to node Q of the cell [10]. Figure 5.26 shows the setup used to obtain the write SNM. Figure 5.26 and Figure 5.28 shows the butterfly curve during a simulated write operation using CMOS 45nm and 20nm GaN TFET devices, respectively.

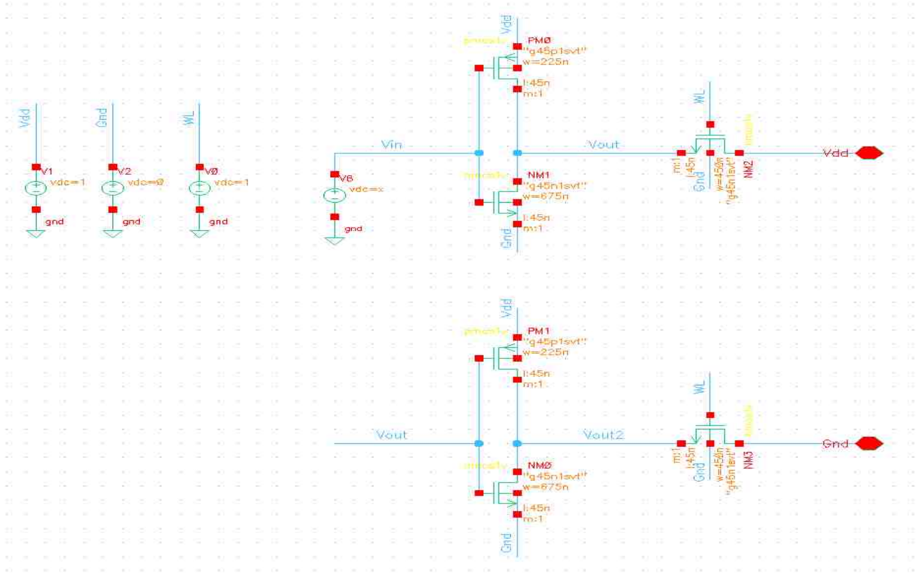

Fig. 5.26. Write SNM Simulation.

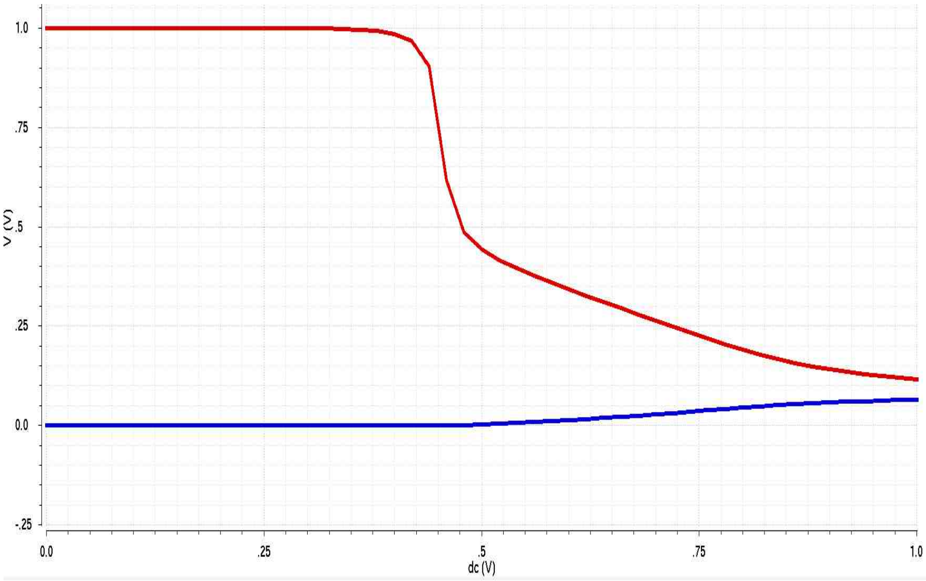

Fig. 5.27. CMOS Write SNM.

In order to investigate the hold SNM of the cell, the WL is turned-off and the voltage transfer characteristics of inverter one and the inverse voltage transfer characteristics of inverter two are plotted, after supplying 0.6V to the storage node Q.

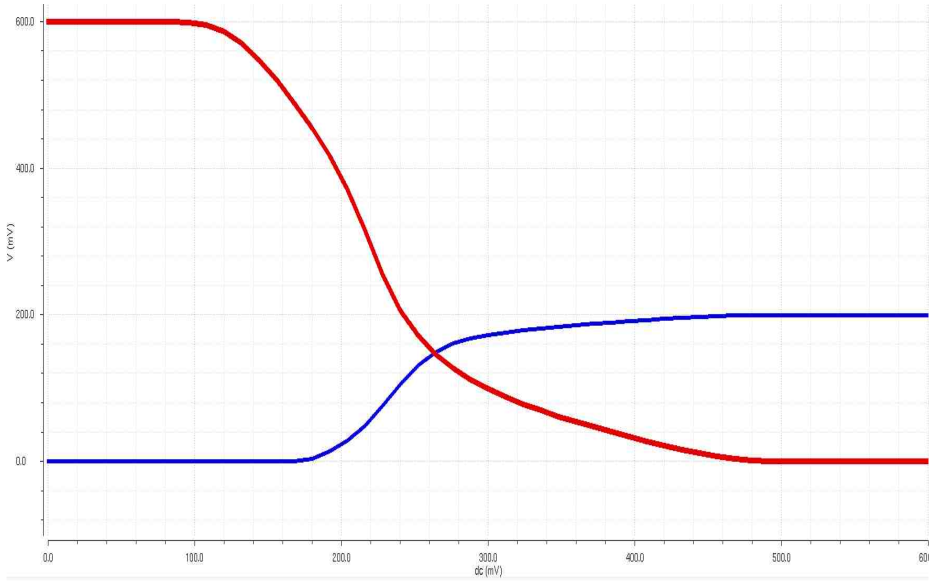

Fig. 5.28. TFET Write SNM.

Figure 5.29 shows the setup used to obtain the write SNM. Figure 5.30 and Figure 5.31 shows the butterfly curve during a simulated write operation using CMOS 45nm and 20nm GaN TFET devices, respectively.

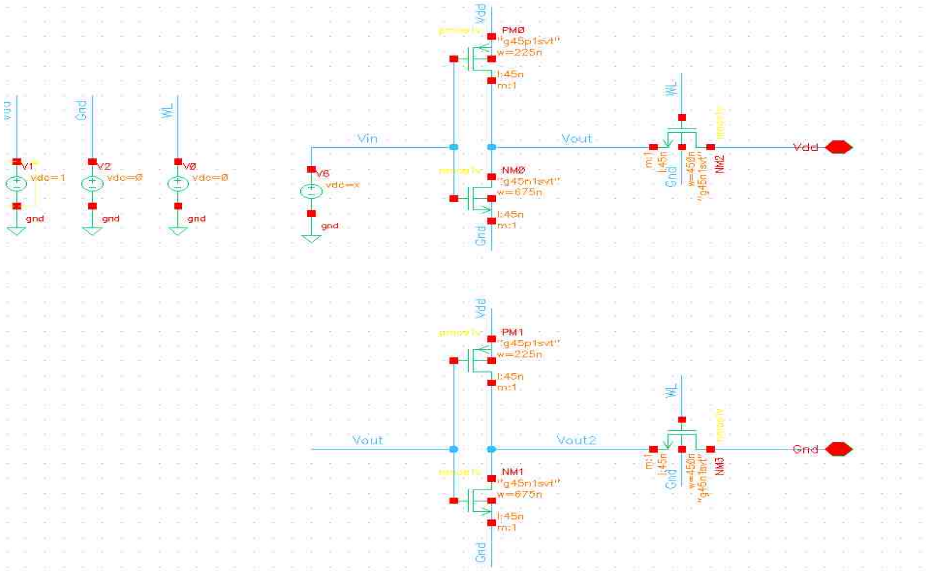

Fig. 5.29. Hold SNM Simulation.

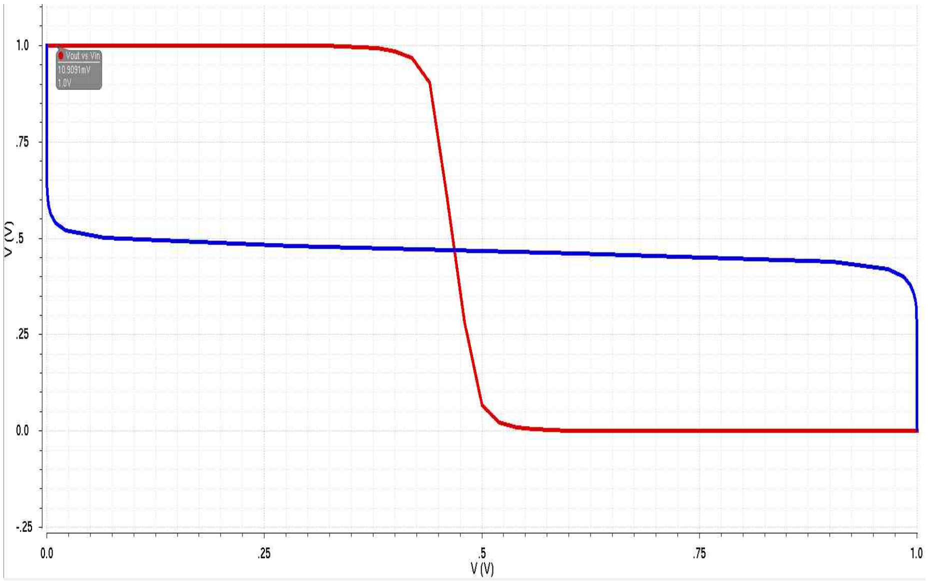

Fig. 5.30. CMOS Hold SNM.

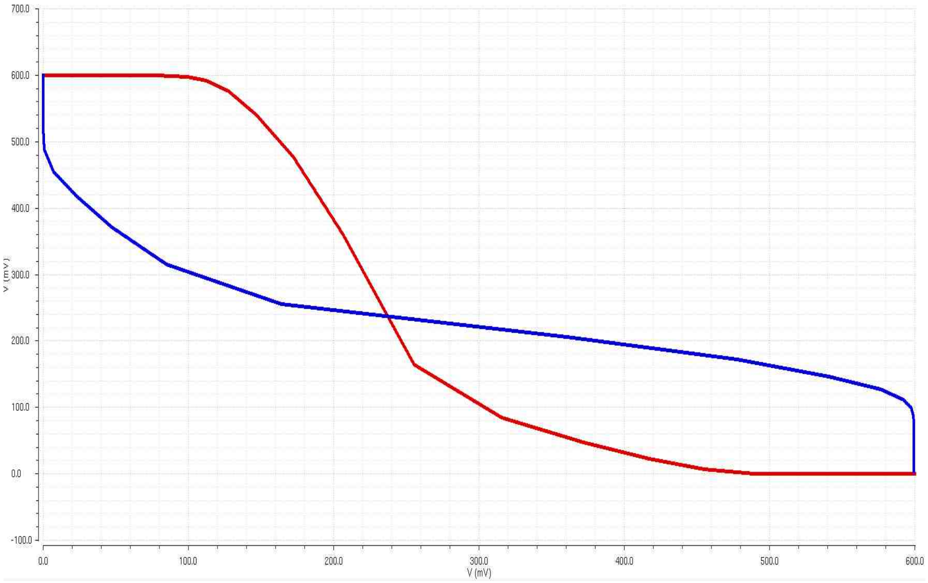

Fig. 5.31. TFET Hold SNM.

The SNM for a read operation is calculated with the WL asserted, node Q at logic 0 and node Q_bar at logic 1. Figure 5.32 show the setup used for obtaining the read

SNM. Figure 5.33 and Figure 5.34 shows the butterfly curve during a simulated write operation using CMOS 45nm and GaN TFET devices, respectively.

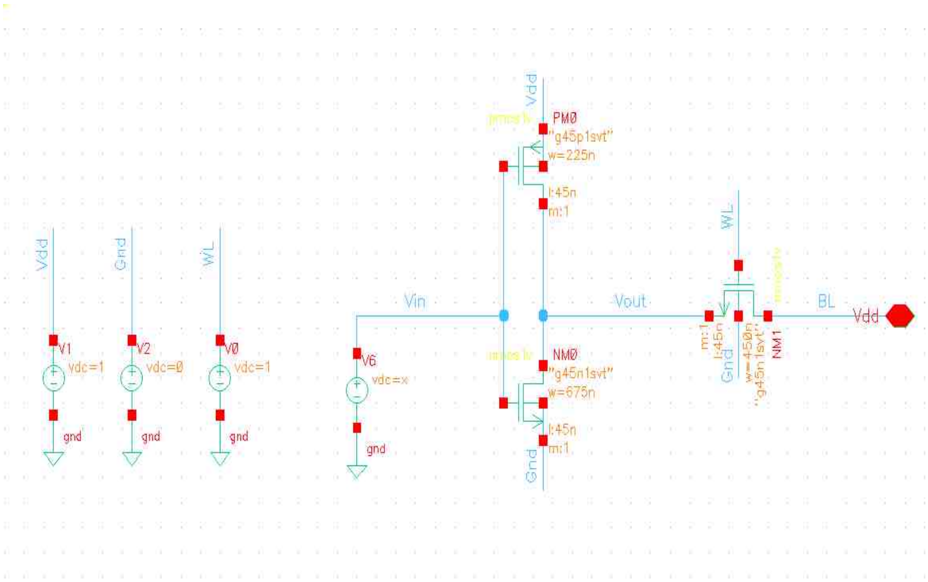

Fig. 5.32. Read SNM Simulation.

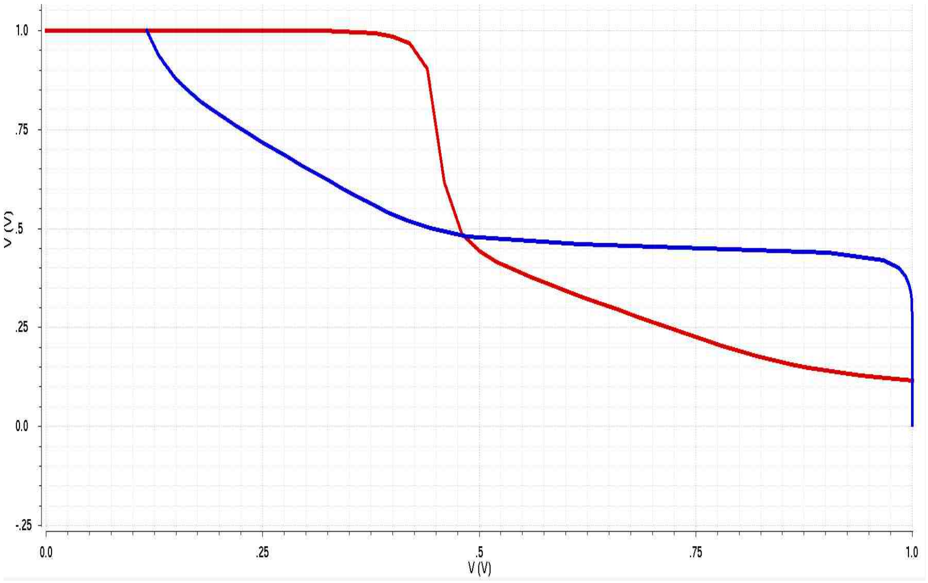

Fig. 5.33. CMOS Read SNM.

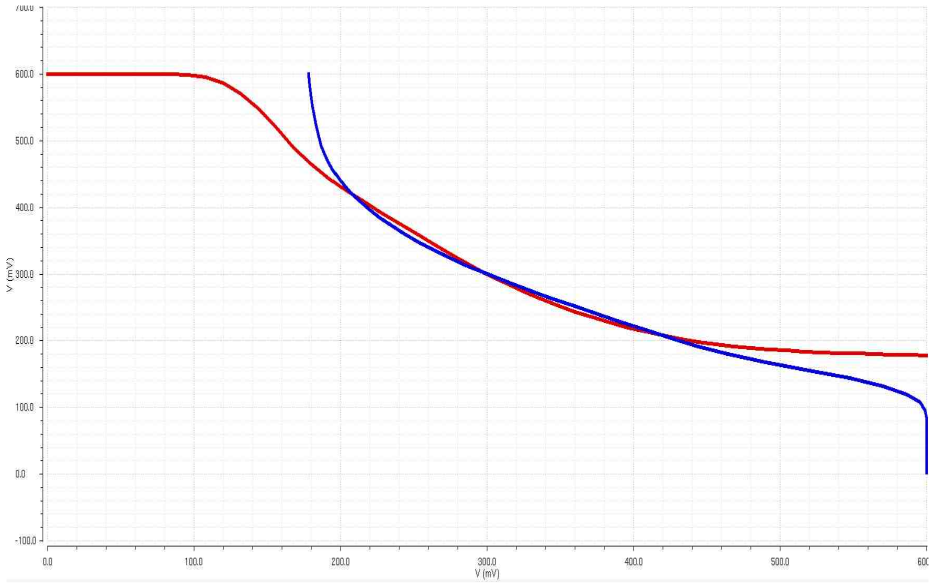

Fig. 5.34. TFET Read SNM.

5.3.10 Leakage Current and Static Power

As previously mentioned, leakage current is the unwanted current that leaks through the transistor when the transistor is switched-off. In order to measure the leakage current and the static power of the unit cell, the setup shown in Figure 5.35 is used. In this setup, the bit lines BL and BLB represented using Inout pins in Cadence Virtuoso and are set to logic 0. The WL is set to logic 0, thus turning-off all four access transistors given by instance numbers I4, I5, I6, I7.

First, node Q was given 0V and then 0.6V using the Node Set feature in the ADE in order to measure the leakage from the cell. When node Q is set to 0V, it turns-on the PTFET I2 and turns-off the NTFET I0, thereby supplying 0.6V to node Q_bar through the V_{DD}. This will, in turn, switch-on NTFET I1 and turn-off PTFET I3, thus reinforcing the value 0V in node Q which is driven to the Gnd. The leakage current is the current leaking through the transistors which are switched-off in this operating mode: NTFET I0 and PTFET I3. The static power is the product of V_{DD} and the leakage current.

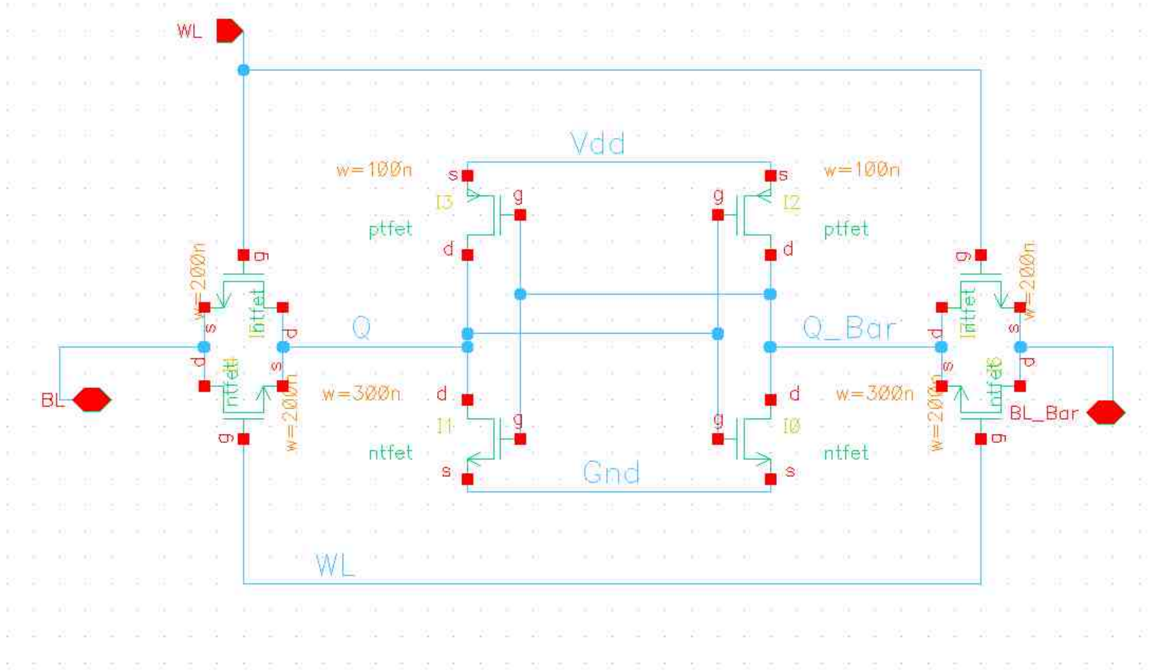

Fig. 5.35. TFET 8T-SRAM Cell Leakage Setup.

$$P_{STATIC} = V_{DD} * I_{OFF} \tag{5.8}$$

Next, node Q is initialized to 0.6V which switches-on NTFET I0 and switches-off PTFET I2, thereby supplying 0V to node Q_bar, which is driven to the Gnd. This will turn-off NTFET I1 and turn-on PTFET I3, reinforcing the value 0.6V driven from V_{DD} in Q. Table 5.6 below summarizes the leakage current and static power measured for the SRAM cells designed using different devices.

From Table 5.6, it can be concluded that the TFET-based cells have lower leakage current and therefore consume lower static power during the hold state when compared to SRAM cells designed using traditional CMOS devices. In order to offer a fair comparison, the channel width to channel length ratios across all devices at different technology nodes were set to the same value. This would in theory, allow for the same beta ratio for the devices. The current through the transistors in the

Table 5.6.
Leakage Current and Static Power in SRAM Bitcells.

Device	**L (nm)**	**W (nm)**			**Q (V)**	I_{OFF}	P_{STATIC}
		PUN	**PDN**	**Access**		**(nA)**	**(nW)**
TFET GaN	20	100	300	200	0	1.72	1.07
CMOS 45nm	45	225	675	450	0.6	30.2	18.1
CMOS 100nm	100	500	1500	1000	0	51.3	30.8

linear $I_{DS-Linear}$ and saturation I_{DS-SAT} regions of an N-type device is given by the following equations 5.7 and 5.8 respectively:

$$I_{DS-SAT} = \mu_n \times Cox \times W/L \times (V_{GS} - V_{th}) \times V_{DS} - V_{DS}^2/2 \qquad (5.9)$$

$$I_{DS-Linear} = \mu_n \times Cox \times W/2L \times (V_{GS} - V_{th})^2 \times (1 - \lambda \times V_{DS}) \qquad (5.10)$$

where I_{DS} is the drain to source current, μ_n is the mobility of the electrons for an N-type device, C_{OX} is the capacitance per unit area of the oxide layer given by ϵ_{OX}/t_{OX}. ϵ_{OX} is the permittivity of the oxide layer and t_{OX} is the thickness of the oxide layer. W is the channel width and L is the channel length of the transistor, V_{GS} is the gate to source voltage, V_{th} is the threshold voltage and V_{DS} is the drain to source voltage.

If the channel length L is increases, the channel width W needs to be decreased in order to drive the same amount of current. This also means that as the channel length increases, the current decreases and the resistance increases. Moreover, when the channel width increases, the current increases or the resistance decreases [10].

6. SYSTEM DESIGN

The SRAM cell described in the previous chapter is used to design two prototype systems. Both systems include an SRAM array-based L1 data cache, CMOS-based registers designed using D-Flip Flops, and a carry ripple adder designed from four single bit full adders. The L1 data cache of system one is designed using CMOS-based SRAM cells. The L1 data cache of system two is designed with CMOS and TFET-based SRAM arrays. In both systems, the size of the L1 data cache is 8x4. Moreover, this L1 data cache is partitioned into two 4x4 arrays, a left and a right array. The D-FF is described in Section 6.1. The carry ripple adder is described in Section 6.2. The functionality of the integrated system is included in Section 6.3. Leakage from the L1 data cache is analyzed in Section 6.4.

6.1 D-Flip Flop

The register of the prototype systems are designed using CMOS-based D-Flip Flops (D-FF) at 45nm technology. A D-FF consists of two D-latches. While latches are level sensitive, the combination of two D-latches connected together in series makes an edge sensitive D-FF. The first latch in the series is called the master, and the second latch the slave. When the clock to the D-FF goes high, data is captured by the master latch. When the clock to the D-FF goes low, the data is transferred from the master to the output via the slave. Figure 6.1 shows the TFET-based D-FF designed in Cadence Virtuoso. It is to be noted that the CMOS-based D-FF is used for register design.

The D-FF-based register requires four clock signals: Clk1, Clk2, Clk3, Clk4. These four clock signals are designed using custom voltage sources. Clk1 is used to transfer a 4-bit data from row 0 of the left 4x4 array in both systems, Clk2 is used to transfer

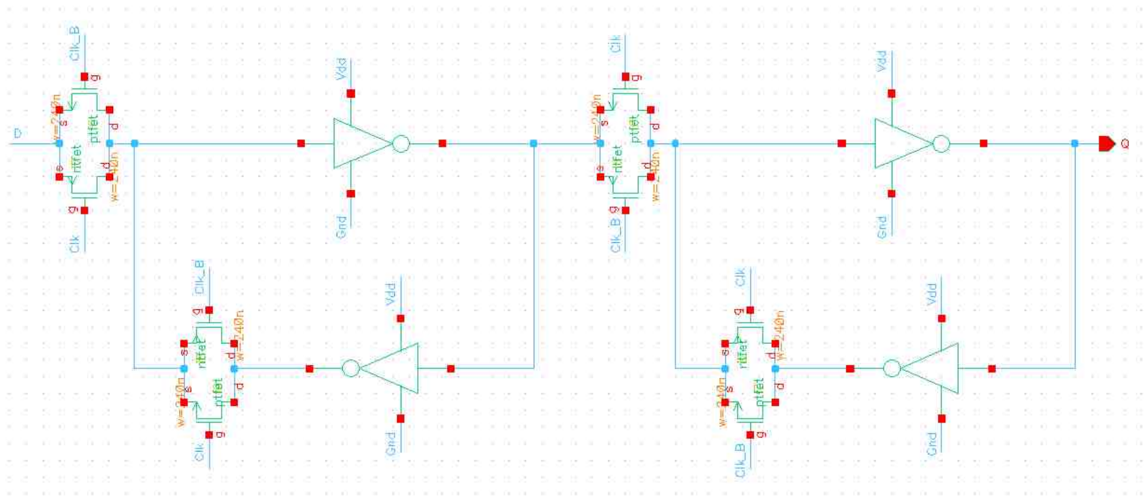

Fig. 6.1. TFET D-Flip Flop.

a 4-bit data from row 1 of the left 4x4 arrays in both systems, Clk3 is used transfer a 4-bit data from row 0 of the right 4x4 arrays in both systems, and Clk4 is used transfer 4-bit data from row 1 of the right 4x4 arrays in both systems. For example, in the CMOS-TFET SRAM array (system two), a 4-bit data 0001 from row 2 is transferred to the adder by Clk3 and a 4-bit data 1000 from row 3 is transferred to the adder by Clk4. Essentially, Clk3 and Clk4 act as the clock signal for the D-FF.

6.2 Carry Ripple Adder

A 4-bit carry ripple adder was designed using four 1-bit full adders in the CMOS technology. A 1-bit full adder has three inputs and produces two outputs. The inputs to a full adder are two single bits which are to be added, and a carry-in. The adder calculates the sum and the carry using the following logic equations:

$$Sum = A \oplus B \oplus C_{in} \tag{6.1}$$

$$C_{out} = A.B + A.C_{in} + B.C_{in} \tag{6.2}$$

Figures 6.2 and 6.3 shows the schematic of a CMOS-based XOR gate and a full adder which are used in the design of the carry ripple adder, designed using Cadence Virtuoso.

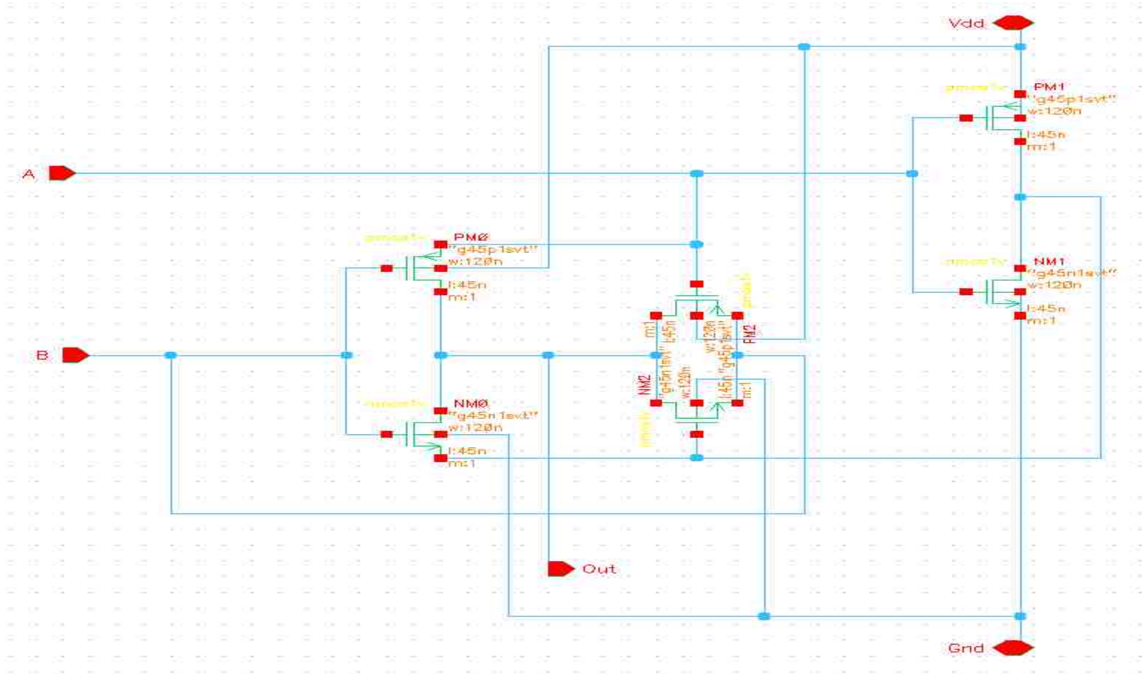

Fig. 6.2. CMOS XOR Gate

In the carry ripple adder, the carry out, (C_{out}), of the first full adder is connected to the input carry-n (c_{in}) of the second full adder, thus rippling the carry through to the next stage of the full adder. This process continues until the last full adder. The critical path delay of the carry ripple adder is given by the equation 6.3:

$$T_{delay} = T_{PG} + (N-1) \times T_{AO} + T_{XOR} \tag{6.3}$$

where, T_{PG} is the delay of the 1-bit propagate/generate gates, N is the number of stages in the adder, T_{AO} is the delay of the AND/OR gates and T_{XOR} is the delay of the XOR gates [10].

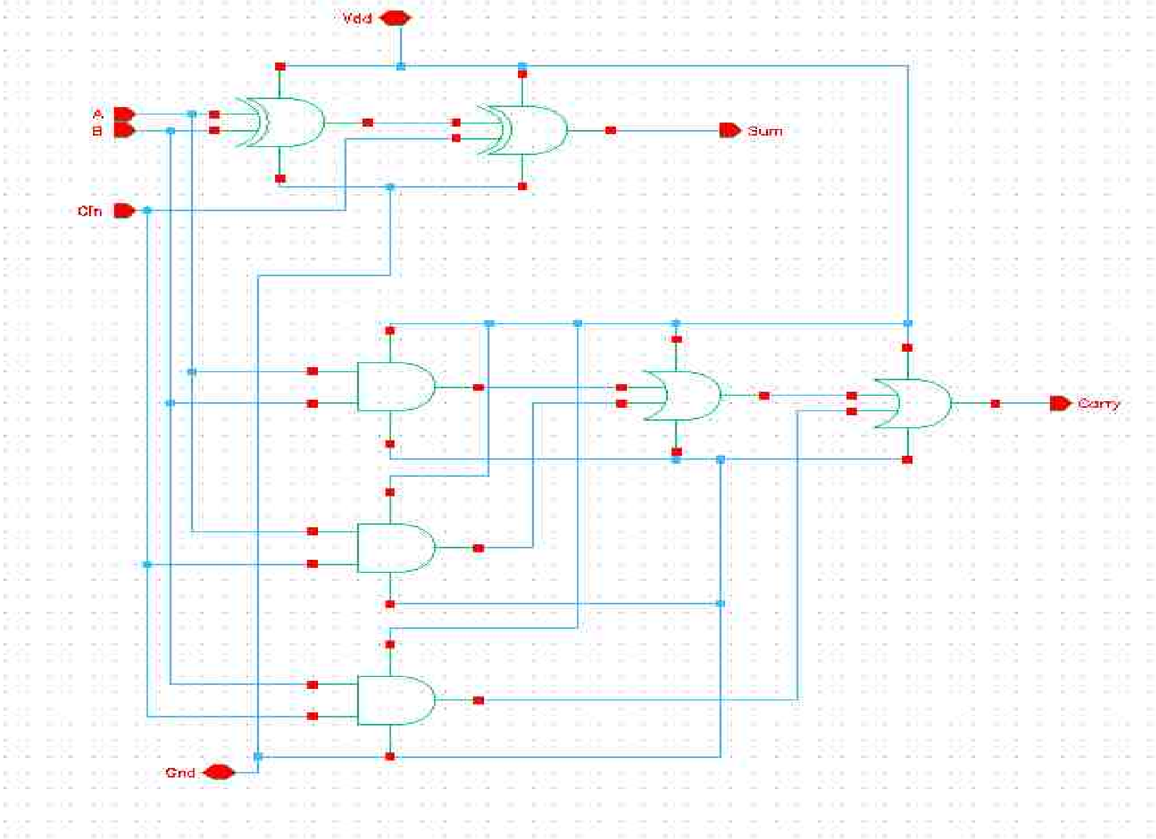

Fig. 6.3. CMOS Full Adder.

6.3 System Functionality

To show the functionality of the proposed design, we implement two systems as mentioned above: system one with CMOS-CMOS 8x4 SRAM array and system two with TFET-CMOS 8x4 SRAM array. A block diagram of the TFET-CMOS-SRAM array (system two) is shown in Figure 6.4.

The CMOS-CMOS and TFET-CMOS SRAM arrays are the L1 data cache and can store eight 4-bit data. The data written into the L1 data cache is fetched to the upper level of memory (i.e., CMOS-based registers designed using D-FFs). The data being read from the L1 cache is stored in these registers before they are selected through a CMOS transmission gate multiplexer and added using a 4-bit CMOS-based

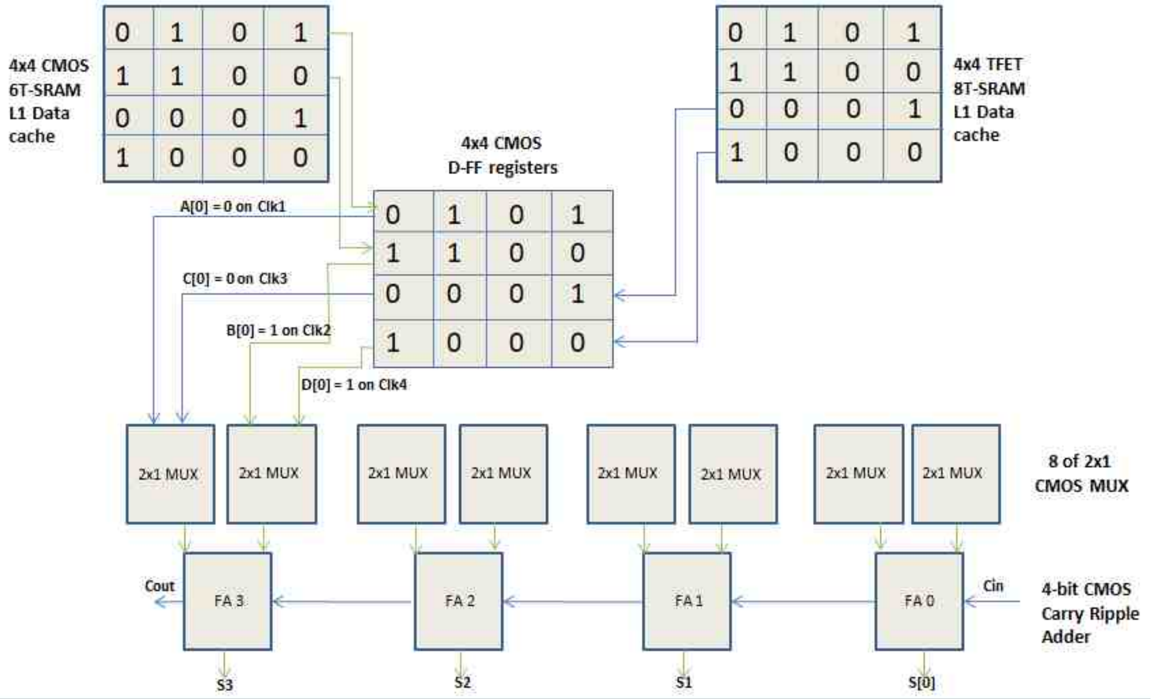

Fig. 6.4. TFET-CMOS Block Diagram.

carry ripple adder. The carry ripple adder is designed using four 1-bit full adders. The Cadence Virtuoso capture of the 8x4 CMOS-CMOS SRAM array with CMOS D-FFs and a CMOS carry ripple adder is shown Figure 6.5.

Custom Vpulse voltage sources are used to generate the data 0101, 1100, 0001 and 1000. This data is first written into rows 0 to 3 of the CMOS-CMOS SRAM array on four different clock cycles. The same data is simultaneously written into rows 4 to 7 for the CMOS-TFET SRAM array.

The SRAM array is partitioned into two 4x4 arrays in order to allow for a comparison between CMOS-CMOS and TFET-CMOS arrays. The CMOS-based arrays work with 1V PMOS and NMOS devices at the 45nm technology. The TFET-based arrays use 0.6V PTFET and NTFET at the GaN 20nm technology. If the arrays are combined as a single 8x4 array instead of two 4x4 arrays, this will result in timing issues during read and write operations. Moreover, if the design uses 8x4 arrays, the

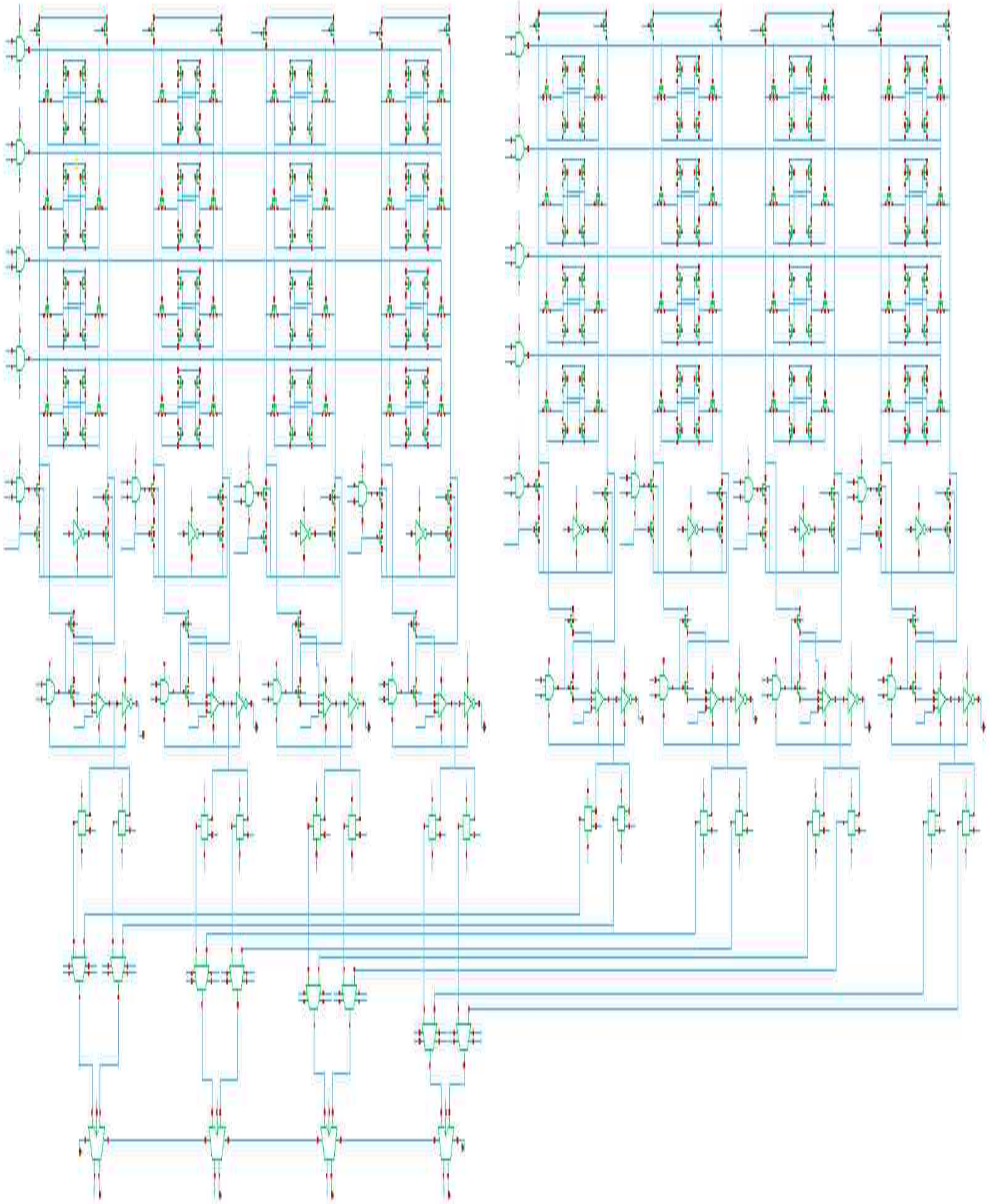

Fig. 6.5. CMOS-CMOS Ripple Adder (System One).

write and read operations need to be performed over 8 cycles each. This partitioning into two separate arrays allows the writing and reading to be performed over 4 clock cycles.

The same data sequence (i.e., 0101, 1100, 0001 and 1000) is used for all testing scenarios. Abdollahi et. al [11] showed that leakage current from a circuit varies depending on different combination of input to the circuit. Thus, using the same sequence of data allows consistent measurement of leakage current from the different SRAM arrays. The same data sequence was also used to measure leakage current for array sizes 8x4, 16x4, 32x4 and 64x4 for both CMOS-CMOS SRAM arrays and TFET-CMOS SRAM arrays.

To illustrate the functionality of system two, consider the block diagram of the system consisting of the CMOS-TFET SRAM array, CMOS D-FF registers and a CMOS carry ripple adder shown in Figure 6.4.

The timing diagrams in Figures 6.7, 6.8, 6.9 and 6.10 show the data 0101, 1100, 0001 and 1000 being written into the TFET-based array in system two, respectively. The TFET-based array writes: data 0101 into row 1 in nodes Q00, Q01, Q02, Q03 on clock cycle 1, data 1100 into row 2 in nodes Q10, Q11, Q12, Q13 on cycle 2, data 0001 into row 3 in nodes Q20, Q21, Q22, Q23 on cycle 3 and data 1000 into row 4 in nodes Q30 Q31, Q32, Q33 on cycle 4. The SRAM array in this case is designed using 20nm GaN TFET technology with 0.6V devices. It uses the clock Phi_2B for pre-charging the bit lines to 0.6V and the clock Phi_1 for writing and reading. This is the right array of system two.

Simultaneously the data sequence 0101, 110, 0001 and 1000 is written to the CMOS-CMOS array as follows: Data 0101 is written into row 1 in nodes P00, P01, P02, P03 on clock cycle 1, data 1100 is written into row 2 in nodes P10, P11, P12, P13 on cycle 2, data 0001 is written into row 3 in nodes P20, P21, P22, P23 on cycle 3 and data 1000 is written into row 4 in nodes P30 P31, P32, P33 on cycle 4. The SRAM array in this case is designed using 45nm CMOS technology with 1V devices.

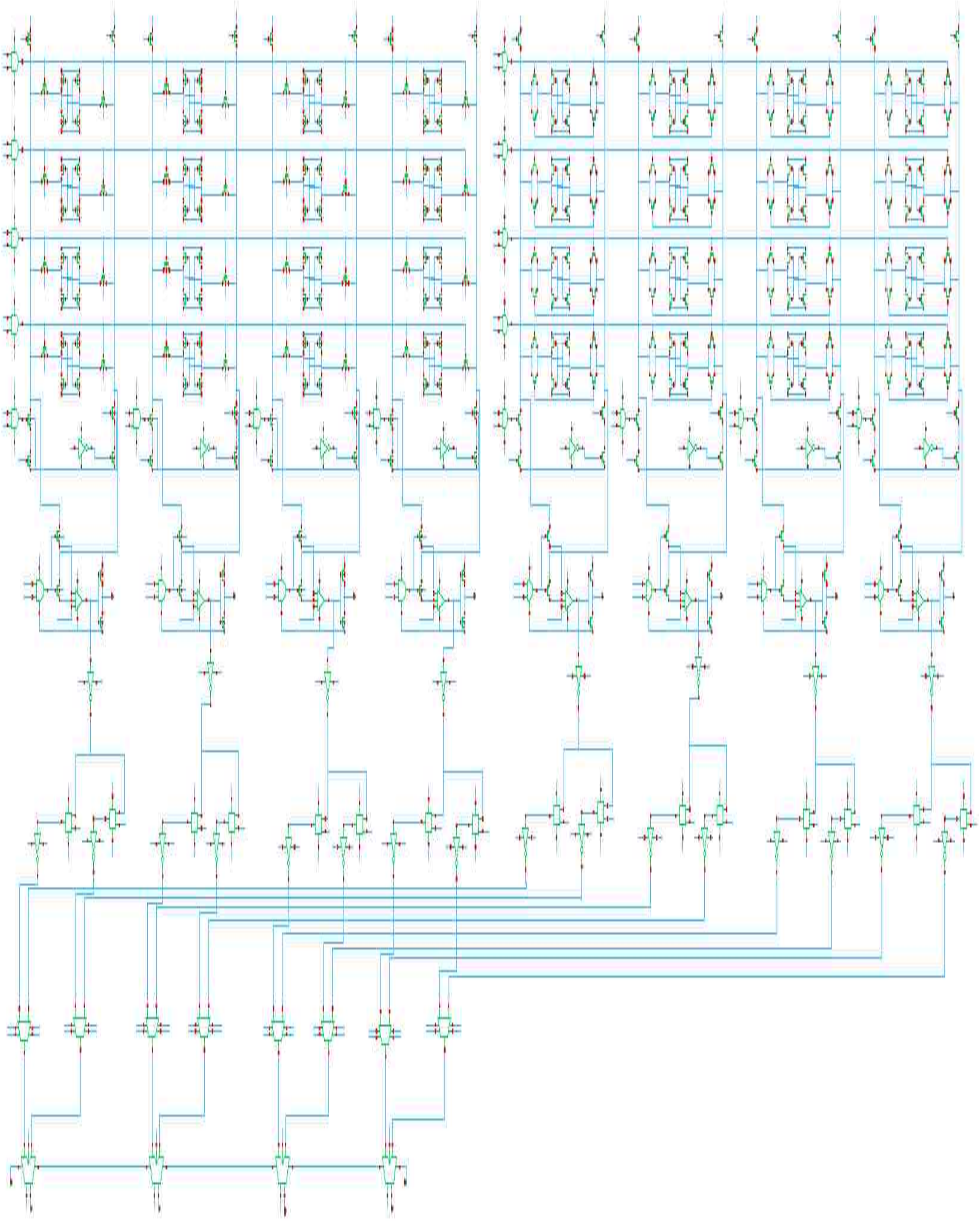

Fig. 6.6. TFET-CMOS Ripple Adder (System Two).

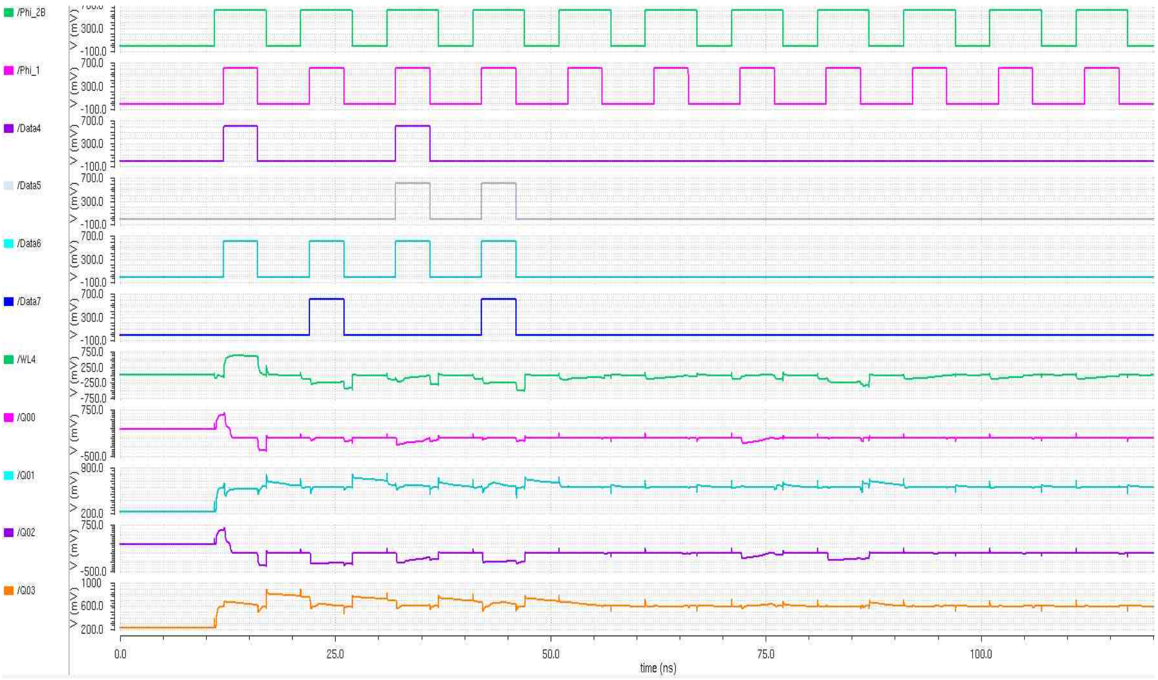

Fig. 6.7. TFET Row 0 Write.

It uses the clock Phi_4B for pre-charging the bit lines to 1V and Phi_3 for writing and reading. This is the left array of system two.

Once the write cycle for both arrays are complete, data is read from one row at a time via the sense amplifiers associated with each column. For example, data 0101 is read from row 0 of the CMOS left array on cycle 5 and data 1100 is read from the CMOS left array on cycle 6 as shown in Figure 6.11. At the same time, the sense amplifiers associated with the TFET array read data 0001 on cycle 7 and data 1000 on cycle 8 as shown in Figure 6.12. The read cycles for the TFET-CMOS SRAM array combination range from cycle 5 to cycle 8.

The data read from the arrays are stored in the D-FF registers. Data in these registers is then selected for addition through the clocks clk3 and clk4. Clk3 transfers the data 0101, through the signals A[0], A[1], A[2], A[3]. The data 1100 from the D-FF is also transferred through B[0], B[1], B[2] and B[3]. Figure 6.13 shows the data

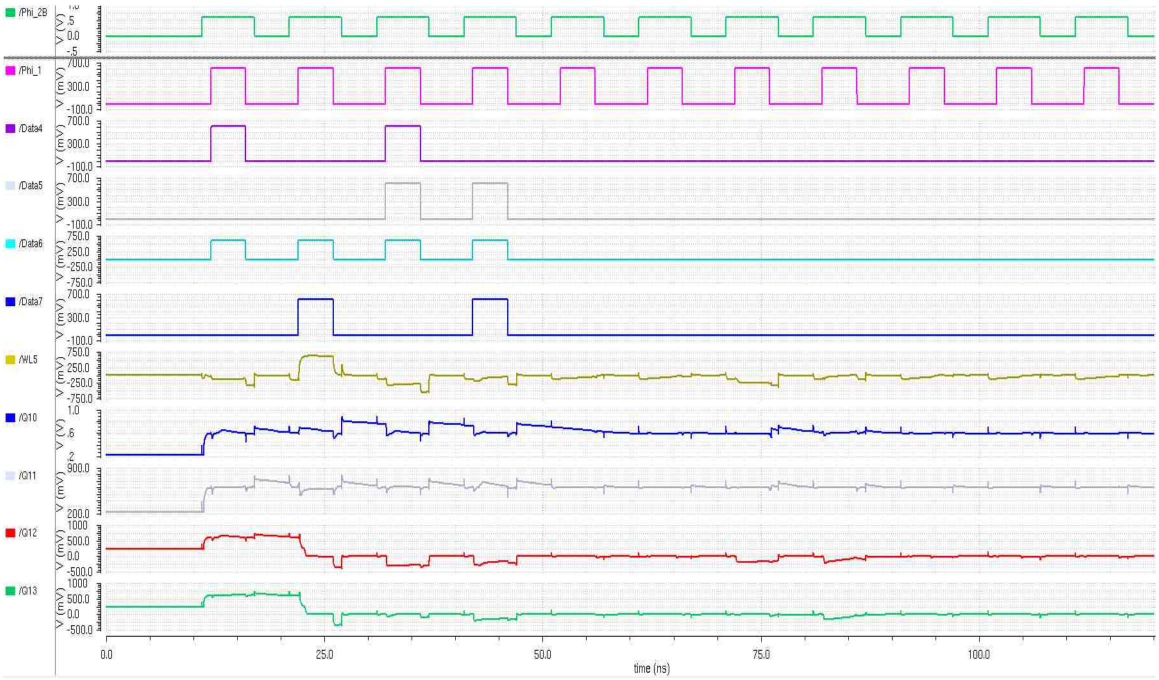

Fig. 6.8. TFET Row 1 Write.

being read from the CMOS-based D-FF registers which were originally read from the CMOS-CMOS SRAM array. Clock Clk4 then transfers the data 0001 from the registers which was originally fetched to the registers from the TFET SRAM array. This data transfer is represented by the output signals C[0], C[1], C[2] and C[3]. The transfer of the data 1000 from the TFET array is depicted by the signals D[0], D[1], D[2] and D[3]. This transfer is illustrated in Figure 6.14.

Multiplexers then selects the data that needs to be send to the carry ripple adder. The select signal (sel) is used to select data 0101 and 1100 (from the CMOS-CMOS SRAM array) on cycle 7 after the clocks Clk1 and Clk2 go low as shown in Figure 6.15. The carry ripple adder takes in as input the data from row 0 column 3 (single bit data 1) and row 1 column 3 (single bit data 0). The carry-in C_{in} for this stage is set to zero. These values are added using full adder 0 (FA 0) to provide a sum of 1 and a carry of 0 as depicted in the signal S[0] and C_{out}, respectively. The carry-out

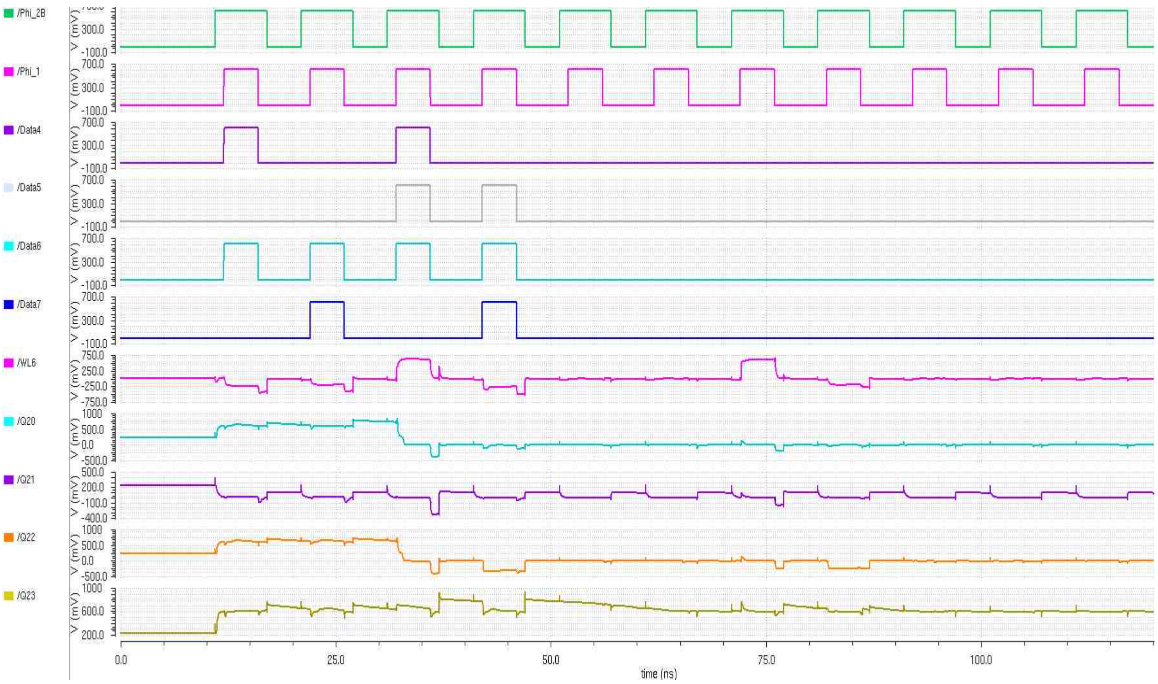

Fig. 6.9. TFET Row 2 Write.

C_{out} is then rippled to the next full adder (FA 1) as its carry-in. The data from row 0 column 2 (single bit data 0) and row 1 column 2 (single bit data 0) is added by the second stage full adder (FA 1) to obtain a sum of 0 and a carry of 0 as shown by signals S[1] and C_{out}, respectively. The ripple process continues and the sum S[2] and S[3] are generated by the third and fourth adders, respectively. This process results in obtaining a sum of 0001 and a carry of 1 when adding 0101 and 1100. For the TFET SRAM array, the data 0001 from row 2 and 1000 from row 3 are selected on cycle 8 of the clock Phi_3 after the clocks Clk3 and Clk4 go low, and are added together to obtain a sum of 1001 and a C_{out} of 0, following a similar execution pattern. This is illustrated in Figure 6.16.

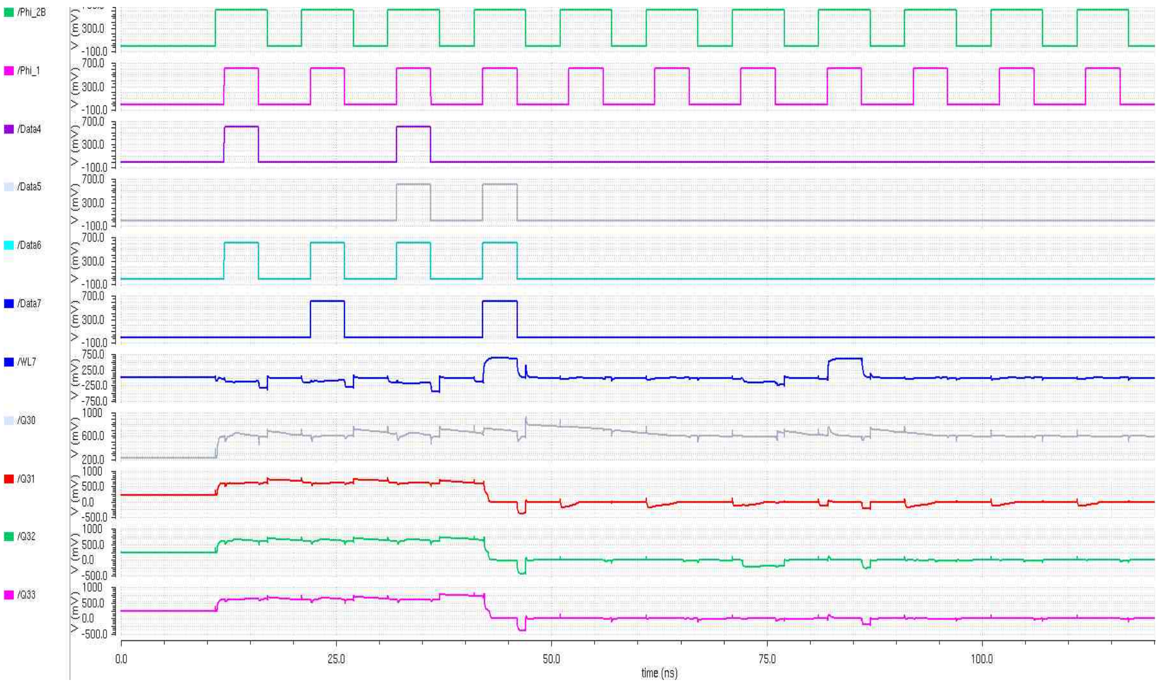

Fig. 6.10. TFET Row 3 Write.

6.4 Leakage Current in L1 Data Cache

The two arrays were expanded to measure the leakage current in the two systems (i.e., system one and system two). In addition to the 8x4 L1 data cache SRAM array. The leakage current was also calculated for arrays of sizes 16x4, 32x4 and 64x4. Table 6.2 summarizes the findings of this analysis.

As can be seen from Table 6.1, the hybrid combination of CMOS and TFET cells in system two has lower leakage current compared to the CMOS-only SRAM arrays. This was expected as we showed in Chapter 5 that TFET-based arrays have lower leakage compared to CMOS-based arrays. This reinforces the finding that TFETs can be used for memory design where low-power is more important than speed. While TFETs have attractive features in the subthreshold region, their current carrying capabilities in the linear and saturation region are not as strong as CMOS which account for the slower speed of TFET-based SRAM array. [7]

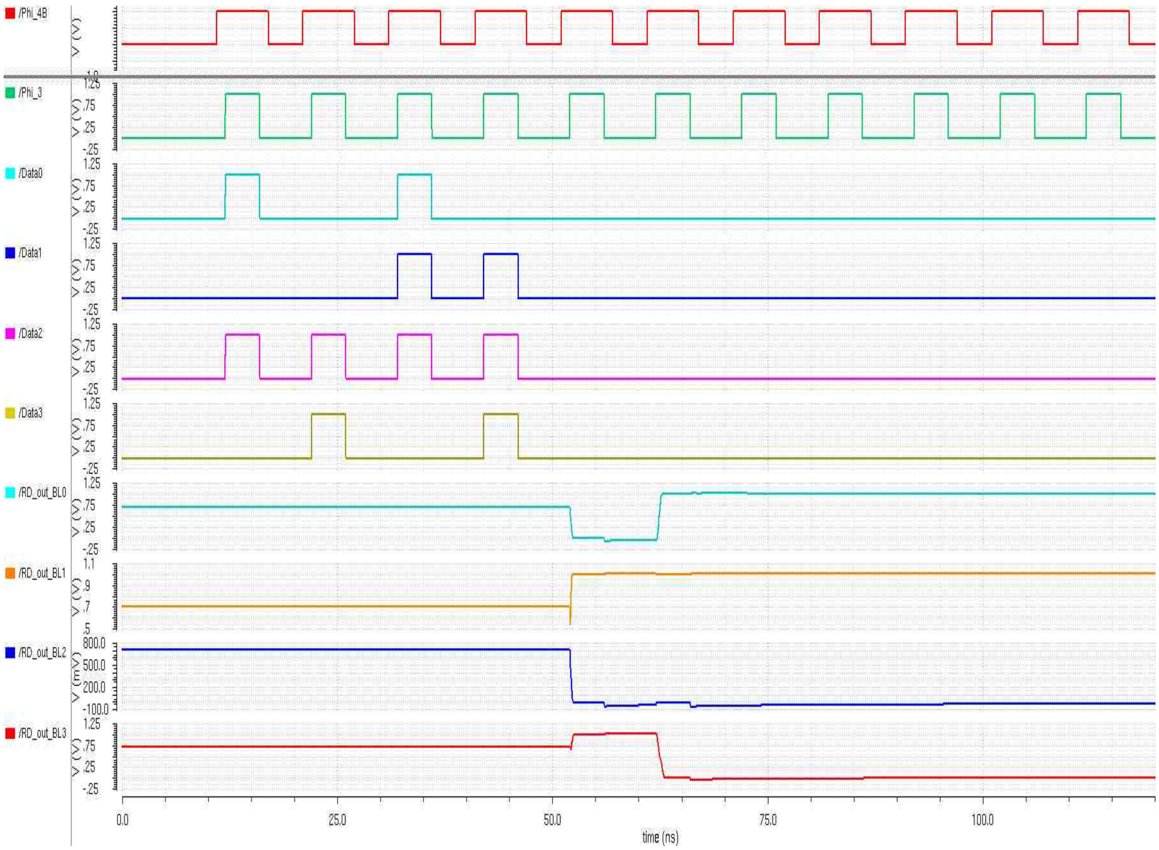

Fig. 6.11. System Two L1 Data Cache Left Array Read.

Table 6.1.
Leakage Current in L1 Data Cache.

Array size	**CMOS-CMOS SRAM Array (nA)**	**TFET-CMOS SRAM Array (nA)**
8x4	1.160	0.2307
16x4	2.327	0.4613
32x4	4.712	0.8803
64x4	1901	1.761

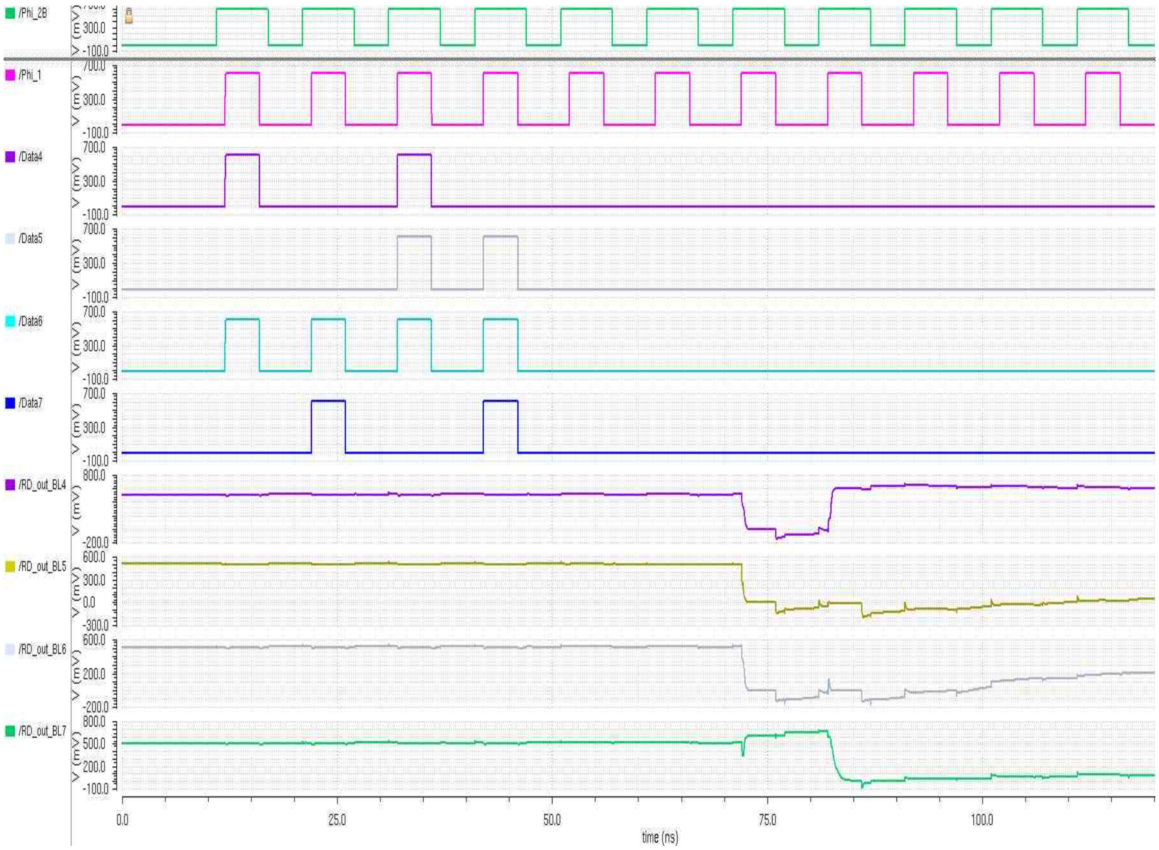

Fig. 6.12. System Two L1 Data Cache Right Array Read..

This chapter introduced two designs: system one with CMOS-CMOS L1 data cache and system two with TFTE-CMOS L1 data cache. Both systems have CMOS-based D-FF registers and a CMOS-based carry ripple adder. These systems perform a very basic addition by fetching data from the L1 data cache to the registers. They were used to compare the performance of TFETs and CMOS-based systems.

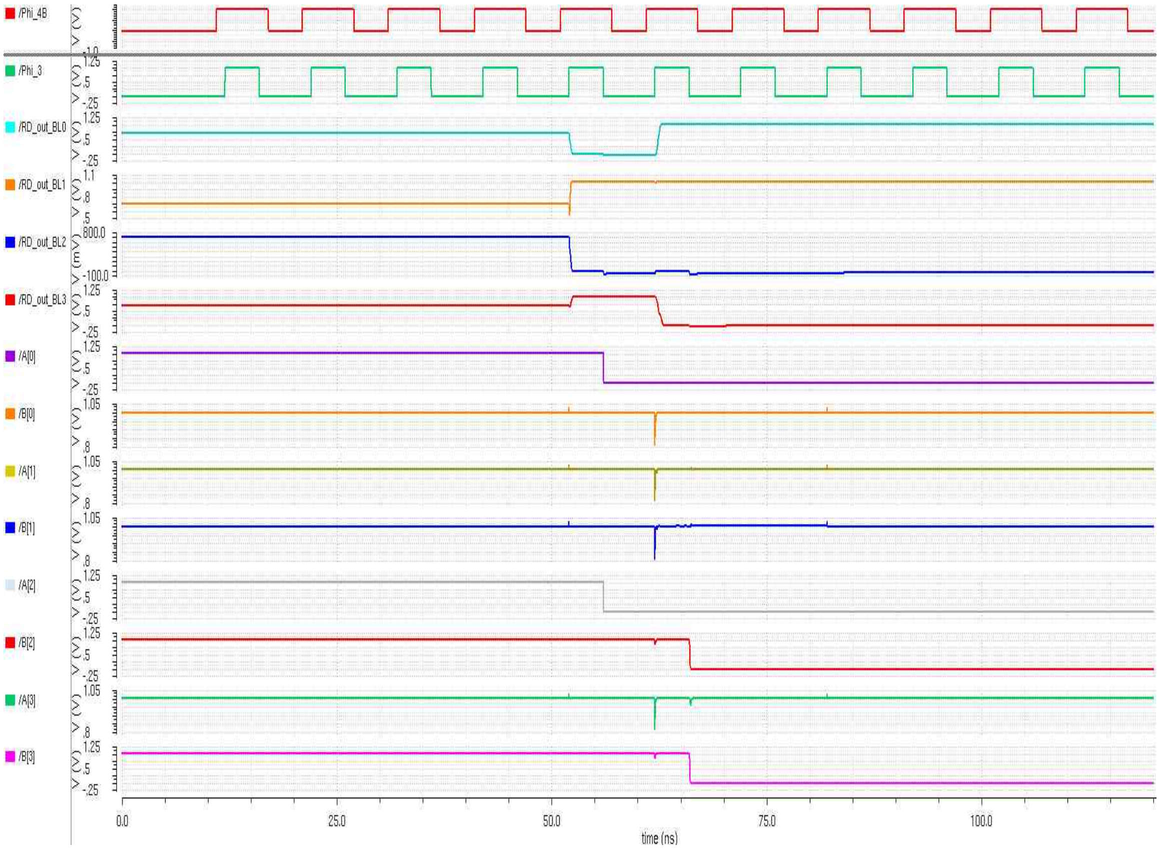

Fig. 6.13. CMOS Register Read of Data originally from Left Array of L1 Data Cache in System Two.

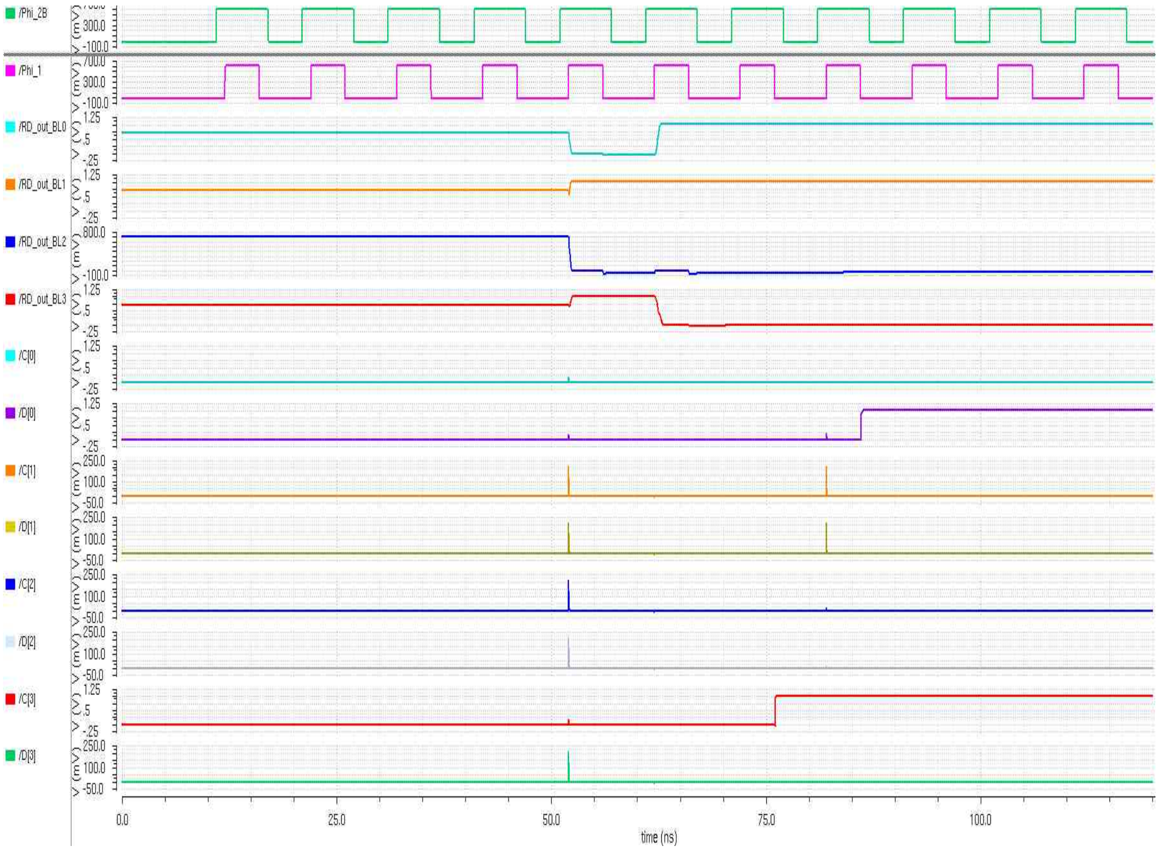

Fig. 6.14. CMOS Register Read of Data originally from Right Array of L1 Data Cache in System Two.

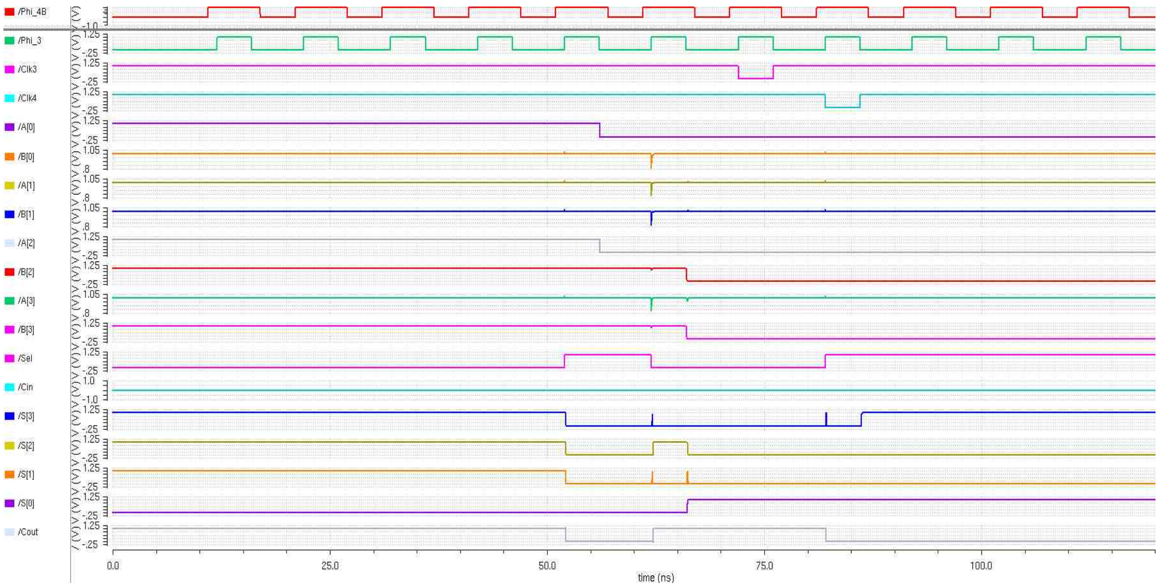

Fig. 6.15. Sum of Data originally from the Left Array in System Two.

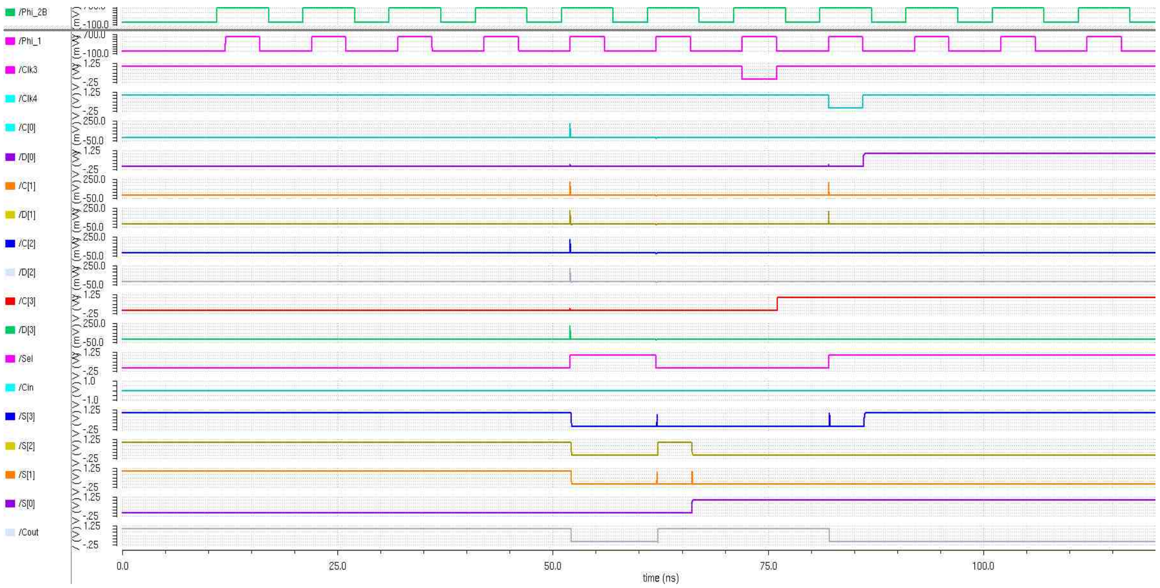

Fig. 6.16. Sum of Data originally from the Right Array in System Two.

7. FUTURE WORK

The designs in this thesis are limited to the schematic level. Future work, when layout cells and design rules can be obtained, will expanded to the physical level. The ability to conduct the analysis at the physical level may allow fabrication and post-silicon analysis which can in turn lead to more detailed assessments. The proposed design approach can also be investigated for lower level memories such the L2 cache or main memories. Theoretically, the longer the memory is switched-off, the more leakage current is dissipated from the memory arrays. Hence, using TFETs in lower level memories could be more advantageous than in upper-levels. In addition, the slower speed of TFETs could be less of an issue in lower-level memories when compared to upper-level memories.

The proposed design was limited to a single ALU function (i.e., a carry ripple adder). Future research should consider different adders and other ALU functions. This would entail a more detailed datapath control. Other extension include the implementation of other memory access functionalities such as branch prediction and pre-fetching.

8. SUMMARY

In this thesis, the successful use of GaN TFET N and P-channel devices in the design of digital and analog systems was demonstrated. The functionality of different system configurations was simulated in Cadence Virtuoso. The estimated power consumption was shown to be superior compared to the standard CMOS logics with a set figure of merit. Moreover, the simulation of the Operational Transconductance Amplifier (OTA) with current source models indicates that TFET devices can support the implementation of the next generation nano-electrical and mechanical systems (NEMS) within systems on chip.

A TFET-based 8 transistor SRAM cell schematic was designed, simulated, and verified. The Signal to Noise Ratio (SNM) of the write, hold and read operations was also measured. Moreover, the leakage current and static power from the cell was analyzed and compared to other CMOS cells. The leakage current and hence the static power consumed by the TFET SRAM cells indicates that this technology can be used in low-power design, for upper levels of memory such as caches. Another advantage of the GaN TFET model is its low operating voltage of about 0.6V. This low voltage can enable dynamic voltage scaling thereby resulting in more power saving when the memory is switched-off.

Two systems consisting of CMOS-CMOS L1 data cache and TFTE-CMOS L1 data cache were designed and functionally verified. The leakage current of the two systems was also compared. The results show that a hybrid combination of CMOS-based cells and TFET-based cells memory array are an adequate design solution for low static power consumption. The hybrid combination of a 64x4 array had a power consumption of roughly one thousand times less than CMOS-based cells.

REFERENCES

REFERENCES

[1] *Issue Efficiency.* NVIDIA Corporation, 2018 (accessed March 10, 2018). [Online]. Available: https://docs.nvidia.com/gameworks/content/developertools/issueefficiency.htm

[2] D. Patterson, T. Anderson, N. Cardwell, R. Fromm, K. Keeton, C. Kozyrakis, R. Thomas, and K. Yelick, "A case for intelligent ram," *IEEE Micro*, vol. 17, no. 2, pp. 34–44, March 1997.

[3] L. Barroso, M. Marty, D. Patterson, and P. Ranganathan, *Attack of the Killer Microseconds*, 2018 (accessed March 22, 2018). [Online]. Available: https://cacm.acm.org/magazines/2017/4/215032-attack-of-the-killer-microseconds/fulltext

[4] K. Ramani, A. Ibrahim, and D. Shimizu, "Powerred: A flexible modeling framework for power efficiency exploration in gpus," in *Proceedings of the Workshop on General Purpose Processing on GPUs, GPGPU*, vol. 7, 2007.

[5] S. Hong and H. Kim, "An integrated gpu power and performance model," in *Proceedings of the 37th Annual International Symposium on Computer Architecture*, ser. ISCA '10. New York, NY, USA: ACM, 2010, pp. 280–289.

[6] J. Leng, T. Hetherington, A. ElTantawy, S. Gilani, N. S. Kim, T. M. Aamodt, and V. J. Reddi, "Gpuwattch: Enabling energy optimizations in gpgpus," in *Proceedings of the 40th Annual International Symposium on Computer Architecture*, ser. ISCA '13. New York, NY, USA: ACM, 2013, pp. 487–498.

[7] H. Lu, T. Ytterdal, and A. C. Seabaugh, *Notre Dame TFET Model*, 2016 (accessed January 29, 2017). [Online]. Available: https://nanohub.org/publications/195/1

[8] R. W. Davis, P. D. Franzon, , Basavarajaiah, S. J. Bucher, M. S and, and I. Castellanos, "Freepdk: A free openaccess 45nm pdk and cell library for universities."

[9] M. Bhole, A. Kurude, and S. Pawar, "Finfets: Benefits, drawback and challenges," in *International Journal of Engineering Sciences and Research Technology*, November 2013.

[10] N. H. E. Weste and D. M. Harris, *CMOS VLSI Design: A circuits and Design Perspective*, 4th ed. Pearson, 2016.

[11] A. Abdollahi, F. Fallah, and M. Pedram, "Leakage current reduction in cmos vlsi circuits by input vector control," *IEEE Trans. Very Large Scale Integr. Syst.*, vol. 12, no. 2, pp. 140–154, February 2004.

[12] N. Saxena and S. Soni, "Leakage current reduction in cmos circuits using stacking effect," in *International Journal of Application or Innovation in Engineering and Management (IJAIEM)*, vol. 2, no. 11, November 2013.

[13] S. Narendra, S. Borkar, V. De, D. Antoniadis, and A. P. Chandrakasan, "Scaling of stack effect and its application for leakage reduction," in *Low Power Electronics and Design, International Symposium*, August 2001.

[14] L. Yuan and G. Qu, "A combined gate replacement and input vector control approach for leakage current reduction," in *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 14, no. 2, February 2006.

[15] A. P. Chandrakasan, W. J. Bowhill, and F. Fox, *Design of High-Performance Microprocessor Circuits*, 1st ed. Wiley-IEEE Press, 2000.

[16] J. M. Rabaey, A. P. Chandrakasan, and B. Nikolic, *Digital Integrated Circuits*, 3rd ed. Upper Saddle River, NJ, USA: Prentice Hall Press, 2008.

[17] A. C. Seabaugh and Q. Zhang, "Low-voltage tunnel transistors for beyond cmos logic," *Proceedings of the IEEE*, vol. 98, no. 12, pp. 2095–2110, December 2010.

[18] J. Appenzeller, Y. M. Lin, J. Knoch, and P. Avouris, "Band-to-band tunneling in carbon nanotube field-effect transistors," *Phys. Rev. Lett.*, vol. 93, p. 196805, November 2004.

[19] B. Romancyzk, *Fabrication and characterization of III-V tunnel field-effect transistors for low voltage logic applications*, Rochester Institute of Technology, 2013.

[20] *What is GPU-accelerated Computing*. NVIDIA Corporation, 2018 (accessed February 4, 2018). [Online]. Available: http://www.nvidia.com/object/what-is-gpu-computing.html

[21] S. W. Keckler, W. J. Dally, B. Khailany, M. Garland, and D. Glasco, "Gpus and the future of parallel computing," vol. 31, no. 5, September 2011, pp. 7–17.

[22] V. Saripalli, S. Datta, V. Narayanan, and J. P. Kulkarni, "Variation-tolerant ultra low-power heterojunction tunnel fet sram design," in *Proceedings of the 2011 IEEE/ACM International Symposium on Nanoscale Architectures*, ser. NANOARCH '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 45–52.

[23] J. P. Kulkarni, K. Kim, and K. Roy, "A 160 mv, fully differential, robust schmitt trigger based sub-threshold sram," in *Proceedings of the 2007 International Symposium on Low Power Electronics and Design*, ser. ISLPED '07. New York, NY, USA: ACM, 2007, pp. 171–176.

[24] U. E. Avci, D. H. Morris, and I. A. Young, "Tunnel field-effect transistors: Prospects and challenges," *IEEE Journal of the Electron Devices Society*, vol. 3, no. 3, pp. 88–95, May 2015.

[25] N. Rahman and B. P. Singh, "Static-noise-margin analysis of conventional 6t sram cell at 45nm technology," in *International Journal of Computer Applications*, vol. 66, 2013, p. 20.