LOW POWER NANOMETER FPGA DESIGN TECHNIQUES AT THE DEVICE AND CITCUIT LEVELS

By

Osama Ahmed Mohamed Ahmed Abdelkader

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE in
Electronics engineering

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
September - 2016

LOW POWER NANOMETER FPGA DESIGN TECHNIQUES AT THE DEVICE AND CITCUIT LEVELS

By
Osama Ahmed Mohamed Ahmed Abdelkader

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE in
Electronics engineering

Under the Supervision of

Ahmed M. Soliman

Professor of electronics
Electronics and communications
Faculty of Engineering, Cairo University

Hassan Mostafa

Assistant Professor
Electronics and communications
Faculty of Engineering, Cairo University

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
September - 2016

LOW POWER NANOMETER FPGA DESIGN TECHNIQUES AT THE DEVICE AND CITCUIT LEVELS

By
Osama Ahmed Mohamed Ahmed Abdelkader

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE in
Electronics engineering

Approved by the
Examining Committee

_____
Prof. Dr. First S. Name, External Examiner

_____
Prof. Dr. Second E. Name, Internal Examiner

_____
Prof. Dr. Third E. Name, Thesis Main Advisor

_____
Prof. Dr. Fourth E. Name, Member

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
September – 2016

Engineer's Name: Osama Ahmed Mohamed Ahmed
Date of Birth: 21/12/1989
Nationality: Egyptian
E-mail: Osama_abdelkader@mentor.com
Phone: +201114140997
Address: 2 Nahda st – Hawamdia - Giza

Registration Date: 1/1/2014
Awarding Date: …./…./……..
Degree: Master of Science
Department: Electronics

Supervisors:

Prof. Ahmed M.Soliman
Dr. Hassan Mostafa

Examiners:

Prof. …… (External examiner)
Prof. …… (Internal examiner)
Porf. …… (Thesis main advisor)
Porf. …… (Member)

Title of Thesis:

Low power nanometer FPGA design techniques at the device and circuit levels

Key Words:
FPGA; DTMOS; FinFET; Charge recycling; Low power

Summary:

We studied in this work replacing the conventional MOSFETs in FPGA with DTMOS and evaluated FPGA cluster and MUXs based on performance, power, and energy. DTMOS shows better energy for cluster (FPGA logic block), and MUXs (FPGA routing). We also studied using FinFET for technology nodes from 20nm down to 7nm, and also studied the impact of environmental variations on FinFET FPGAs metrics. Finally, we present a novel technique for FPGA power reduction on circuit level, the multiple charge recycling technique shows power saving by 32% in SPICE simulations.

Acknowledgments

First, thanks God for everything that he has presented to me in life. I would like to express my appreciation and gratitude to the people who supported and encouraged me during my MSc study. Without their help I would not be able to complete this work.

I would like to express my sincerest appreciation and gratitude to my advisor Prof. Ahmed Soliman, for his continual guidance, support, motivating suggestions and encouragement during my research. His endless energy and enthusiasm in research has motivated all of his students including me. I wish him all the best. Also, I would like to express my gratitude to Dr. Hassan Mostafa for his important advices and comments for my thesis. It was an honor to have Dr. Hamdy Abdelhamid in my thesis work. I am grateful for their comments and objective directions to complete the master‘s thesis work.

Last but not least, I would like to extend my gratitude to the people who have helped and inspired me during my research. My heartfelt appreciation goes toward my family and my dearest parents. They were always supportive whenever it was a hard time for me.

Table of Contents

List of Tables

List of Figures

Abstract

Field-Programmable Gate Arrays (FPGAs) have become one of the key digital circuit implementation media over last decades. The importance of FPGAs comes from their architecture, which consists of programmable logic functionality blocks and programmable interconnects. This nature of FPGA has a terrific impact on the quality of the final product's performance, area, and power consumption. There are many techniques to make FPGAs more energy efficient. The different techniques can be categorized to: device, circuit, system, architecture, and computer-aided design (CAD). Device techniques refer to the usage of new emerging low-power process technologies offered by the semiconductor manufacturers, and new devices materials and structures. Circuit techniques refer to transistor level implementations of logic and routing resources. System techniques refer to high level techniques such as dynamic voltage and frequency control, power gating for unused resources, and dynamic reconfiguration. Architecture techniques refer to functionality of logic blocks, memory, and I/Os resources and the connectivity between these resources. Finally, CAD techniques refer to improvements added to the tools used to configure FPGAs to consider power consumption. In this work, we target introducing new design techniques to lower FPGAs power at device and circuit levels.

First, we studied using dynamic threshold MOSFET (DTMOS) in FPGA logic blocks and showed that DTMOS can be used as a good candidate for designing ultra-low power FPGAs. The study also covered DTMOS MUXs as MUXs are the main part in FPGA routing fabric.

Following that, we studied the implementation of FPGA using FinFET instead of CMOS to explore future technologies impact on FPGAs energy, also environmental variations are covered in the study since the variations in nanoscale technologies cannot be neglected. We used predictive technology models (PTM model cards) for 20nm down to 7nm technologies to explore performance, power, and energy of FPGA components at each technology node and the impact of variations on the trends of these metrics, We also obtained the optimum supply voltage from the energy point of view for a variety of flip-flops topologies which can be used in today's FPGAs at each technology node. We evaluated FPGA cluster metrics using SPICE simulations based on three benchmark circuits: 2-bit adder, 4-bit NAND, and cascaded chain.

Finally, we present a novel technique to recycle charges in FPGAs interconnects on multiple stages, The new technique exhibits more power saving than single stage charge recycling, We analyzed the new technique from theory of work through circuits details to CAD tools support, The study also covers the area overhead of this technique for a set of benchmarks in the versatile placement and routing (VPR) tool which is widely used in academic researches of FPGA.

Chapter 1 : Introduction

1.1. Field-Programmable Gate Arrays

Field-programmable gate arrays (FPGAs) are programmable devices that user can configure to implement any desired digital circuit. The popularity of FPGAs has grown rapidly since they introduced in the mid-80s, and today, they account for more than half of the three billion dollar programmable logic industry. Modern FPGAs can implement electronic circuits with millions of logic gates which operate at speed of hundreds of megahertz.

FPGA is a chip that consists of two dimensional array of pre-fabricated programmable logic blocks that can be connected through configurable interconnections (routing channels). SRAM cells in the FPGA logic blocks define the desired function to be implemented for each block, and the connectivity between those blocks. FPGA is configured to implement a given circuit in a matter of seconds and can be re-programmed any number of times. Custom ASICs are the primary competitor to FPGAs, and they require weeks or months for fabrication. Hence, a key advantage held by FPGAs over ASICs is that FPGAs reduce “time-to-market", which is crucial in the development of new electronic products.

The rapid expansion of the programmable logic market has been driven by a number of factors. Perhaps most important is that as technology scales, the costs associated with building a custom ASIC rise drastically. For example, in 90nm process technology, the cost of mask sets alone is over a million US dollars [1]. Such costs make design mistakes extremely costly, as they necessitate the creation of new mask sets and impose lengthy delays. Comprehensive and rigorous design verification is a mandatory part of custom ASIC design. In FPGAs, the requirement to “get it right the first time" is less critical, since mistakes, once identified, can be taken care of quickly and cheaply by re-programming the device.

Coupled with the high cost of ASIC fabrication, the CAD tools required to design an ASIC cost anywhere from hundreds of thousands to millions of dollars [2]. In contrast, FPGA vendor tools are typically provided free-of-charge by the FPGA vendors to their best customers, and third-party FPGA CAD tools, such as Synplicity, cost only tens of thousands of dollars. Initially, FPGAs were used only in low-volume production applications or for prototyping circuits that were eventually to be implemented as custom ASICs. However, the volume threshold at which FPGAs are cost-effective versus ASICs has advanced to a point such that modern FPGAs are cost-effective for all but high volume applications.

One of the drawbacks of FPGAs is that they are less area-efficient and also slower than custom ASICs. This characteristic has been the motivation for nearly two decades of academic and industrial research on FPGA CAD and architecture. The result has been a narrowing of the gap between ASICs and FPGAs from the area and speed viewpoints. Today, FPGAs are a viable alternative to custom ASICs and can be used in applications with speed and size requirements that previously, could only be met by ASICs.

1.2. Technology scaling and emerging devices

Traditional CMOS scaling faces challenges due to material and process technology limits. Obstacles for scaling planar devices to sub-32nm gate lengths include short-channel effects (SCE), sub-threshold leakage and gate dielectric leakage. FinFET is a multi-gate three-dimensional transistor structure [3]. This increases the control of the gate on the channel, and reduces drain induced barrier lowering, enhances sub-threshold swing [4-7]. In addition, the near-ideal sub-threshold characteristics raise the potential of FinFET to be used in near-threshold supply circuits, which dissipates an order of magnitude less energy compared to regular strong inversion circuits that operate with the super-threshold supply voltage, these properties makes it the potential candidate to substitute CMOS to keep on technology scaling. Compared to conventional transistor, the DIBL and sub-threshold swing are improved by using the double gate structure. As a depleted-substrate transistor [8], FinFET can overcome the continue scaling obstacles [9]. Recently Intel [10] announced its 22nm FinFET process that will be used for the next generation processor. IBM is also spending a lot of R&D efforts in FinFET [11].

1.3. Motivation

Nowadays the demanding on low power designs is increasing rapidly, reducing energy is not required only for battery life, but also to reduce cooling cost. FPGAs are spreading widely in data centers, mobile devices, and many other fields due to their flexibility and low non-recurring cost compared to ASIC and custom designs. As in ASIC domain the design requires months to fabricate the first device and costs millions of dollars in CAD tools for the different design flow steps, and engineers work for multiple of years. Moreover, the fabrications cost and time, furthermore as technology advances the ASIC design becomes more difficult and expensive due to reliability requirements. On the other hand, FPGA can be programmed in few seconds, and can be easily re-programmed for many times which makes it a favorable solution for prototyping and debugging and drives most digital design starts toward FPGA implementation. Moreover, the recent FPGAs can be partially and dynamically re-programmed. However, the flexibility of FPGA involves significant area overhead and makes it slower than ASIC and consuming more power. To implement a certain logic circuit in FPGA, it needs more transistors compared to ASIC which makes power consumption for a logic gate is higher in FPGA [12-13]. In general, FPGA power efficiency is ill-reputed area which ASIC is superior to it [14]. Power dissipation is a limiting factor for FPGAs to continue replacing ASICs [15].

Despite the fact that FPGAs are weaker than ASIC from power perspective, the research work to reduce their power just started recently and FPGAs vendors give the power reduction high consideration besides the performance and area. Though area and speed have been the main research focus to date, power is likely to be a key consideration in the design of future FPGAs for the reasons outlined below.

The leakage power increases with technology scaling which has a large implications for FPGAs since they have a huge number of transistors for flexibility, and large portion of FPGAs is unused even when implementing large systems. Thus, the need to reduce and manage the leakage power in FPGAs is amplified compared to other design platforms.

Since FPGA power consumption stands in its way of spreading, reducing the power consumption of FPGAs is expected to spread their usage in new several other fields. In addition, it's mandatory for FPGAs to break into low energy ASIC market. Previously, portable applications have been inaccessible to FPGA vendors due to the tight power budget.

1.4. Thesis Contributions

This thesis focuses on two overarching themes:

1. Reducing FPGA power consumption on device level
2. Reducing FPGA power consumption on circuit level

With respect to these themes, a number of different contributions are made, as summarized below.

Chapter 3 considers reducing leakage power dissipation in FPGAs through using dynamic threshold MOSFET (DTMOS) [16] instead of the conventional MOSFET, and shows that replacing traditional MOSFET with DTMOS in both logic and routing fabric in FPGA leads to reducing FPGA power and energy. The study considered different MUXs sizes and FPGA logic cluster.

Exploration for the future of FPGA also considered by studying using FinFET instead of CMOS for future technologies, we used PTM models [17] for low standby power for technology nodes: 20nm, 16nm, 14nm, 10nm, and 7nm in our study. We also studied the impact of environmental variations such as hot carrier injection (HCI), bias temperature inversion (BTI), and redundant doping fluctuations (RDF) on the performance, power, and energy for separate components of FPGA since the variations in those technologies scope have a considerable impact on digital designs metrics. This work has been published in [18]. We also provide the optimum supply voltage for different flip-flops topologies that can be used inside FPGA from power delay product point of view. This work has been published in [19].

In addition, we built FinFET FPGA cluster and studied the above metrics while programming the cluster to work as: 2-bit adder, 4-bit NAND, and cascaded chain.

Chapter 4 presents circuit-level techniques for reducing power dissipation in FPGA, a detailed description of using multiple charge recycling technique from theory to circuit to CAD application is provided with power savings value and area overhead for a set of benchmarks circuits.

1.5. Organization of the thesis

The remaining of this thesis is organized as follows. Chapter 2 provides a detailed survey of the previous related studies. Chapter 3 provides details of device level techniques to reduce FPGA power. Chapter 4 provides details of FPGA power reduction techniques at circuit level. Finally, conclusions and discussions concludes the thesis results and recommendations.

Chapter 2 : Literature Review

2.1. Introduction

This chapter presents the background material that forms the basis for the research presented in later chapters. Section 2.2 gives an overview of FPGA architecture and hardware structures, highlighting the features of two state-of-the-art commercial FPGAs. Section 2.3 discusses the power dissipation breakdown (static and dynamic) in FPGA with two commercial FPGAs examples. Section 2.4 surveys the recent literature on FPGA power optimization. Section 2.5 summarize the chapter.

2.2. FPGA architecture and hardware structures

This section presents an overview of FPGA architecture and hardware structures using two recently-developed commercial FPGAs as example cases: the Xilinx Virtex-7 FPGA [20] and the Altera Stratix-V FPGA [21]. FPGA consists of two dimensional array of programmable logic blocks which can be connected through a configurable interconnection fabric. Figure 2.1 shows a view of island style FPGA. Routing channels lie between rows and columns of logic clusters and contain pre-fabricated routing tracks. Look-up-table (LUT) is the base element to implement logic functions, and each logic block contains flip-flops to implement sequential logic. A K-input look-up-table (K-LUT) is a memory that can implement any logic function with up to K inputs. A simplified generic FPGA logic block is provided in Figure 2.2. Basic logic element (BLE) consists of a K-LUT with a flip-flop. A programmable multiplexer allows the flip-flop to be bypassed. Figure 2.3 shows the internal details of a 4-LUT. 16 SRAM cells hold the truth table for the logic function implemented by the LUT. The LUT inputs, labeled f0-f3, select a particular SRAM cell whose content is passed to the LUT output.

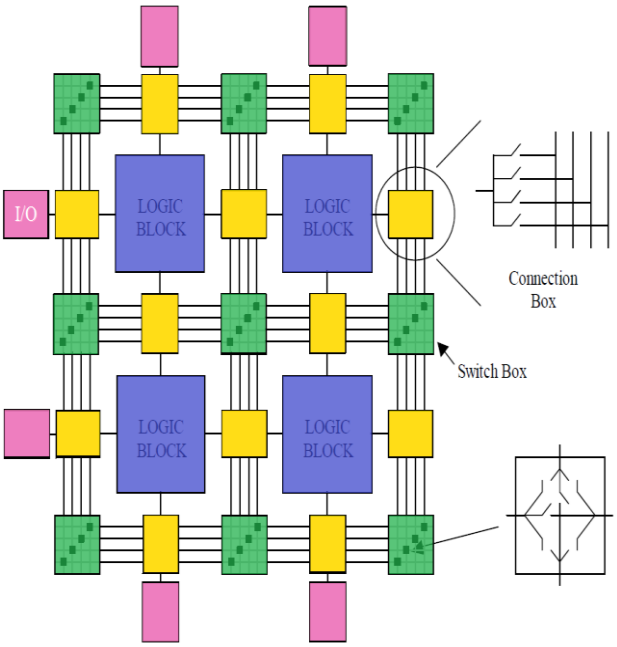


Figure 2.1: FPGA architecture

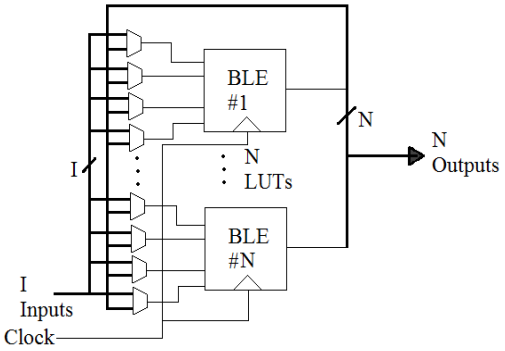


Figure 2.2: FPGA cluster

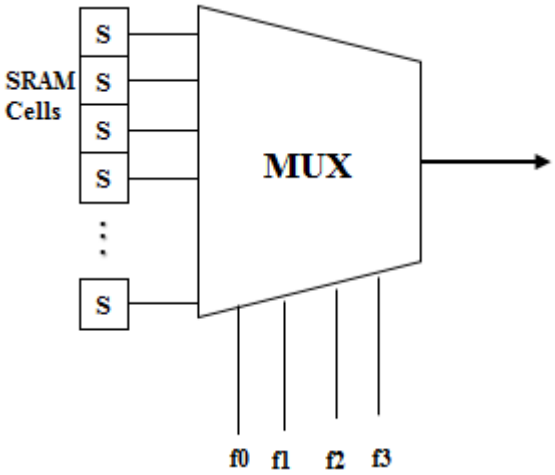


Figure 2.3: 4-LUT

The logic blocks in commercial FPGAs are more complex than that of Figure 2.2 and contain clusters of LUTs and flip-flops. Figure 2.4 shows the logic blocks in Virtex-7 and Stratix-V. A Virtex-7 logic block is referred to as a Configurable Logic Block (CLB) and it contains two SLICEs, where each SLICE has four 6-LUTs, eight flip-flops, as well as arithmetic and other dedicated circuitry. A Stratix-V logic block, referred to as a Logic Array Block (LAB), is shown in Figure 2.5. A LAB has ten adaptive logic modules (ALMs), various carry chains circuits, control signals, shared arithmetic chains, a local interconnect, and register chain connection lines. ALMs in the same LAB transfer signals between them through the local interconnect. A LAB can drive the local interconnect of its left/right neighbors through the direct link interconnect. The output of the ALM register can be transferred to the adjacent ALM register in the LAB through register chain connections. An ALM has a group of different LUT-based resources that can be divided between two adaptive LUTs (ALUTs), and four registers. An ALM can implement different combinations of two logic functions with up to eight inputs for the two ALUTs. This allows an ALM to be backward compatible with four inputs LUT architectures. An ALM can also implement any six inputs logic function and some seven inputs functions. Furthermore, each ALM has four programmable registers, two dedicated full adders, a carry chain, a register chain, and a shared arithmetic chain. Dedicated resources allow ALM to implement arithmetic logic functions efficiently. An ALM drives all types of interconnects: column, row, local, register chain, carry chain, direct link, and shared arithmetic chain. Figure 2.6 provides a block diagram of Stratix-V ALM.

Note that in addition to the LUT-based logic blocks described here, commercial FPGAs contain other hardware blocks including block RAMs, multipliers, and DSP blocks [20-21]. Typically, such blocks are placed at regular locations throughout the two dimensional FPGA fabric. Furthermore, commercial FPGAs have programmable I/O blocks, capable of operating according to a variety of different signaling standards.

Connections between logic blocks in an FPGA are formed using a programmable interconnection network, having variable length segments and programmable routing switches. Figure 2.7 shows a common FPGA routing switch. It consists of a multiplexer, a buffer, and SRAM configuration bits. Within an FPGA, the switch's multiplexer inputs, labeled i_0-i_n, can be connected to routing conductors or to the outputs of a logic block. The output of a buffer can be connected to a routing segment or to an input of a logic block. FPGA’s interconnections can be programmed though SRAM cells, labeled “n ctrl signals" in Figure 2.7. The SRAM cell contents control which input signal is selected to be driven through the buffer.

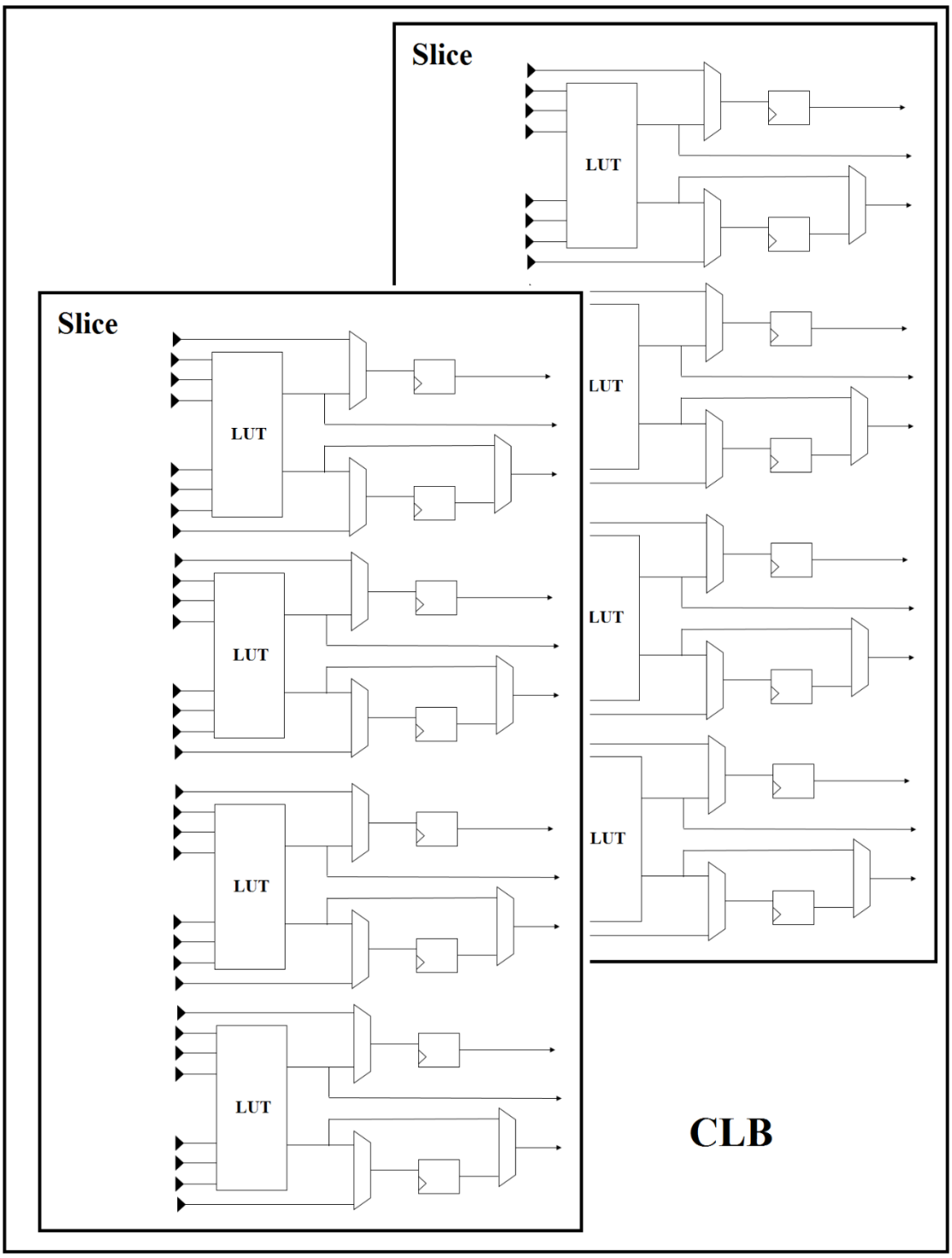


Figure 2.4: Virtex-7 FPGA Cluster

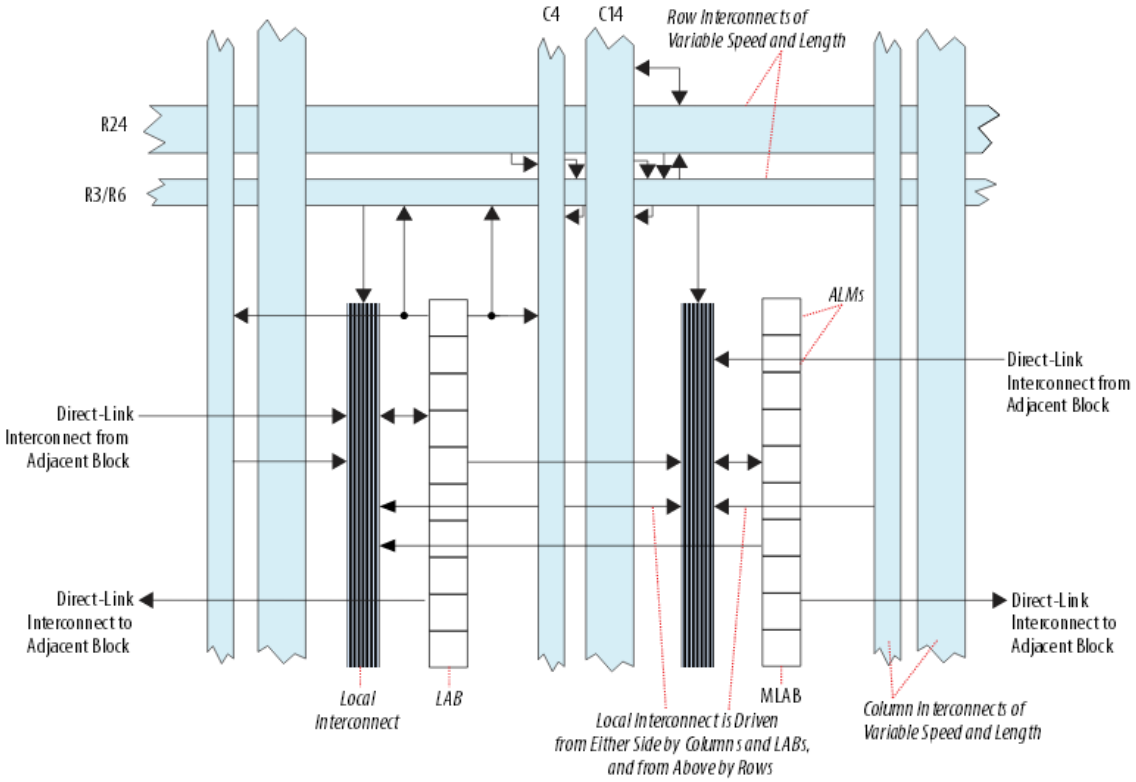


Figure 2.5: Stratix-V FPGA logic array block

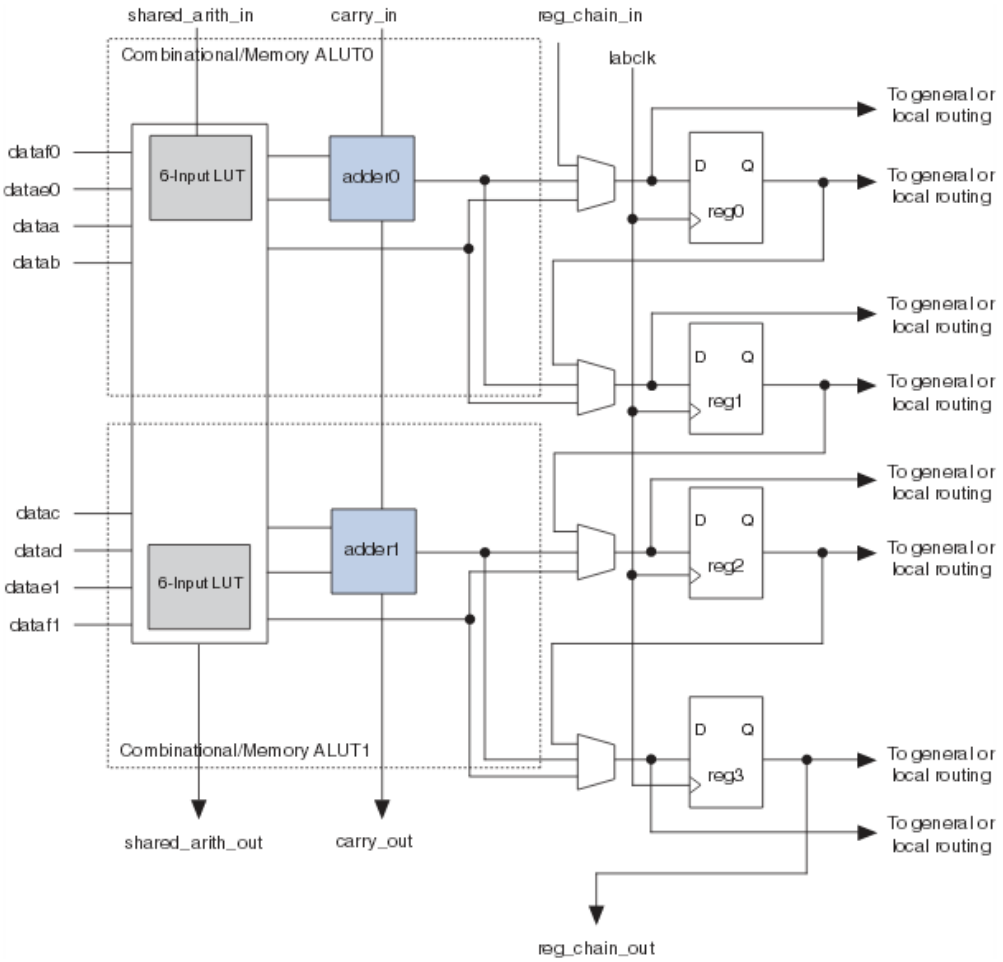


Figure 2.6: Stratix-V FPGA adaptive logic module

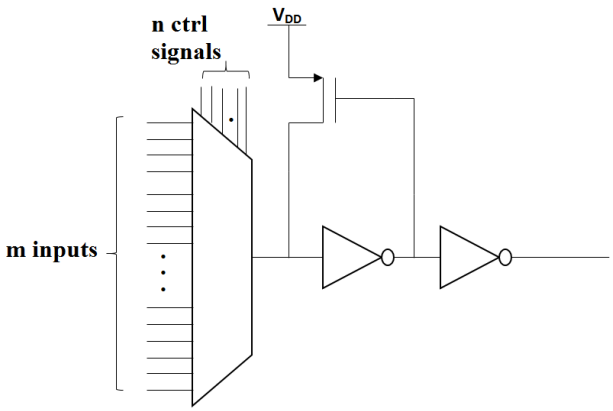


Figure 2.7: FPGA switch buffer

Combined, a routing switch and the conductor it drives are referred to as a routing resource. The connectivity pattern between logic blocks and routing, as well as the length and connectivity of routing conductors constitute the FPGA's routing architecture. Routing resources in FPGAs can be classified to two types; segmented routing resources, and dedicated routing resources. Segmented routing resources provide connections between the logic blocks. As depicted in Figure 2.7, the segmented prefabricated wires are allocated in channels between connection blocks to provide

configurable connections between them, switch blocks, and logic blocks. The number of segments in a channel is usually denoted by W.

Logic block's I/Os can be connected to segmented routing wires on each of its four sides using connection blocks. Connection block flexibility (Fc) is defined as the number of wires in a channel that a logic block pin can connect to. Moreover, the switch box adds more flexibility by providing programmable connections between horizontal/vertical channels. Switch block flexibility (Fs) is defined as the number of segments that an incoming wire can connect. Segment length is defined as the number of logic blocks that can be spanned by a routing wire. Modern FPGAs use a variety of conductors of different lengths (segments) to achieve the optimum performance (delay, and routability). Dedicated routes are used for global signals like reset and clock, thus reducing the skew. In addition, some FPGAs have PLLs and Delay-Locked Loops (DLLs) to reduce skew. Modern FPGAs have the flexibility to provide different clock domains inside the FPGA to enable asynchronous designs.

Basically FPGA offer "local" routing resources for connections within a CLB or a LAB. DIRECT resources that connect a CLB to its eight neighbors (including the diagonal neighbors). DOUBLE and HEX resources run horizontally and vertically and span two and six CLB tiles, respectively. LONG resources span 24 or more CLB tiles.

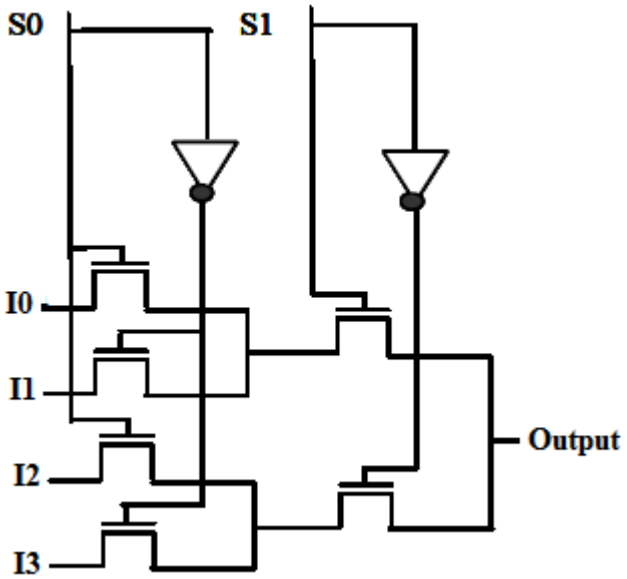


Figure 2.8: Encoded multiplexer

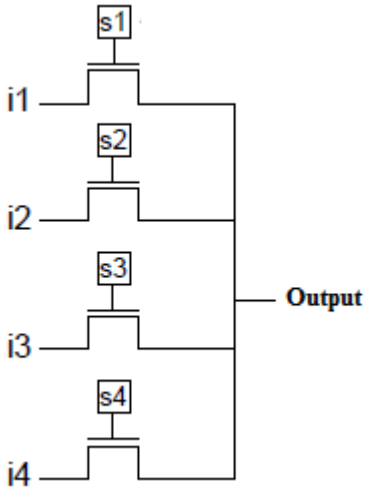


Figure 2.9: Decoded multiplexer

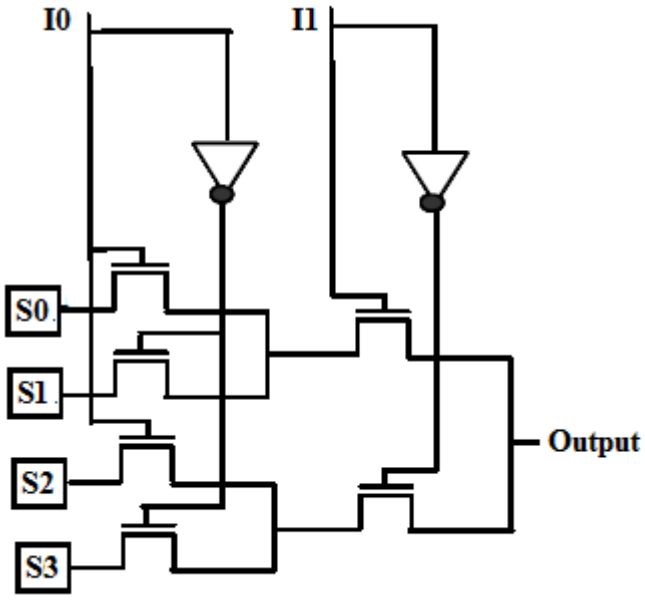


Figure 2.10: Multiplexer inside look-up-table

Given the discussion so far, the reader will appreciate that the multiplexer is perhaps the most important circuit element in an FPGA, since it is used extensively throughout the interconnect and is also used to build LUTs. It is therefore worthwhile to review this structure in some detail. Multiplexer are implemented using pass transistor logic in FPGAs [22]. Figures 2.8 and 2.9 depict multiplexers, as they would be deployed in a routing switch. Transmission gates are not used generally to implement

FPGAs multiplexers due to their larger capacitance and area [26]. Figures 2.8 and 2.9 show two implementations of a 4:1 multiplexer. Figure 2.8 shows a “decoded" multiplexer, which needs four configuration SRAM cells to be used in a FPGA routing switch. Input-to-output paths through the decoded multiplexer consist of only one transistor. Figure 2.9 shows an “encoded" multiplexer that requires only two configuration SRAM cells, though has larger delay as its input-to-output paths has two series transistors. A combination of the two types in Figures 2.8 and 2.9 can be used in large multiplexers which allows trading off between area and delay. In a LUT, the LUT inputs drive multiplexer select signals; SRAM cells containing the truth table of the LUT's logic function attach to the multiplexer's inputs. A multiplexer in a two-input LUT is shown in Figure 2.10.

2.3. Power dissipation in FPGA

2.3.1. Dynamic power

Many recent papers have considered breaking down dynamic power consumption in FPGAs [27-29]. [29] analyzed power consumption in the Xilinx Spartan-3 commercial FPGA. Results are summarized in Figure 2.11. Interconnect, logic, and clock load were found to account for 62%, 19%, and 19% of Spartan-3 dynamic power respectively. Another similar breakdown was observed in [27]. Clock network usually is the major source of power dissipation in ASICs [30], however FPGA power breakdown is different than in ASIC where interconnects is the major source of dissipation. FPGA Interconnects has the dominant dynamic power due to their composition, which consist of prefabricated wire segments, with used and unused switches for each wire segment. Such attached switches are not present in custom ASICs, and they contribute to the capacitance that must be charged/discharged in a logic transition. Moreover, configuration SRAM cells form a significant fraction of FPGA’s total area. For example, [31] suggests that SRAM configuration cells takes more than 40% of FPGA's logic block area. Such area overhead makes wirelengths in FPGAs longer than wirelengths in ASICs. FPGAs interconnects present high capacitive loads which makes it the dominant source of dynamic power dissipation.

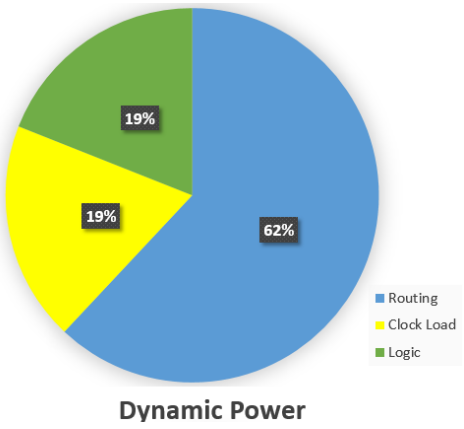


Figure 2.11: Dynamic power breakdown in Spartan-3 FPGA [29]

2.3.2. Leakage power

In comparison with dynamic power dissipation, relatively little has been published on FPGA leakage power. One of the few studies was published by Tuan and Lai in [32], and examined leakage in the Xilinx Spartan-3 FPGA, a 90nm commercial FPGA [33]. Figure 2.12 shows the breakdown of leakage in a Spartan-3 CLB, which is similar to the Virtex-4 CLB. Leakage is dominated by that consumed power in the interconnect, configuration SRAM cells, and to a lesser extent, LUTs. Combined, these structures account for 88% of total leakage. As pointed out in [32], the contents of an FPGA's configuration SRAM cells change only during the FPGA's configuration phase. Configuration is normally done once at power-up. Therefore, FPGA's SRAM configuration speed is not critical, since it does not have impact on the speed of the implemented circuit on FPGA. The SRAM cells can be slowed down and their leakage can be reduced or eliminated using previously-published low leakage memory techniques, such as those in [34], or by using high-V_{TH} for memory cells or using long channel transistors. Leakage power didn't have a big consideration in the design of Spartan-3. If SRAM configuration leakage were eliminated to zero, the Spartan-3 LUTs and interconnect would account for 26% and 55% of total leakage respectively. Note that FPGA design only uses a portion of its resources which is not the case in ASICs. Both used and unused portions of FPGA contribute to leakage power. To be sure, [32] suggests that up to 45% of leakage in Spartan-3 is "unused" leakage (assuming reasonable device utilization). Notably, today's commercial FPGAs do not yet offer support for a low leakage sleep mode for unused regions.

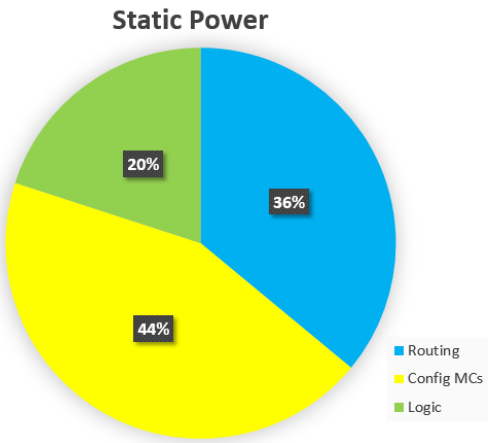


Figure 2.12: Static power breakdown in Spartan-3 FPGA [29]

2.4. Related work

FPGAs power reduction techniques have been explored at all levels of design: CAD, architecture, circuits, and devices. CAD techniques basically adapt the core algorithm's cost function to target low-power operation. Architectural level techniques evaluate FPGA power as a function of standard parameters: number of LUTs per logic block, LUT size, arrangements of logic blocks, segments lengths, etc.

Device level: some recent research works studied replacing CMOS transistors with carbon nanotubes and tunnel FETs [35], other researchers studied using resistive

random access memory (RRAM) in FPGA instead of the conventional static random access memory (SRAM). Since SRAMs are used in FPGA in many parts: configuration RAMs, and LUTs, the impact of any saving of such replacement on the overall power saving in FPGA will be significant. In this work we studied using FinFET FPGA cluster and MUXs instead of CMOS FPGA for future technology nodes (20nm, 16nm, 14nm, 10nm, and 7nm) to explore the impact of that replacement on FPGA metrics like performance, power, and energy. We also studied the impact of process variations on these metrics for FPGA components: MUXs, and flip-flops separately and suggest the optimum supply operating point from energy perspective for a set flip-flop topologies. Moreover, we studied the process variations on FinFET critical path (Ring ocillator) to explore their impact on future technologies.

In addition, several studies targeted replacing CMOS in FPGA's MUXs with dynamic threshold MOSFET (DTMOS) [16] and showed that DTMOS is a good candidate for low power applications. In this work we are proposing using DTMOS in FPGA logic blocks and routing fabric for ultra-low power applications and show that DTMOS configurable logic block (CLB) consumes less power than equivalent CMOS CLB.

Architecture and circuit level: Several attempts were conducted to determine the optimum architecture parameters from energy perspective, such as number of inputs of LUT (K), and number of LUTs per cluster (N) since changing each parameter has tradeoffs of area, performance and power. For instance, increasing K or N increases interconnects and routing inside CLB and hence increases the dynamic power inside CLB, on the other hand it reduces the required CLBs to implement the design. Several researches have been shown that k = 4 minimizes area and leakage energy [36].

Interconnect Architecture: some researcher studied the impact of switchbox configuration and segments length on FPGA power [37], results show that using segment length = 1 is the most energy efficient. In addition, using unidirectional versus bidirectional switches [38].

Dynamic Voltage Scaling: Dynamic voltage scaling (DVS) saves power by lowering the supply voltage, but at the cost of increasing the delay. DVS is useful in scenarios where a design needs to operate at a target frequency. In such cases, the supply voltage can be reduced to a certain point where the desired frequency is still achieved, minimizing the wasted energy on slack. The exact value of V_{DD} can be different between chips due to variation and can change over time due to environmental variation; hence, an on-chip control circuit with feedback of the delay is used to adjust the supply. DVS in FPGAs was examined by [39]. They use a design-specific measurement circuit that tracks the delay of a design's critical path to provide feedback to the voltage controller. Through this technique energy savings of 4-54% can be observed.

Power Gating: Leakage power forms a big part of FPGA power consumption, and it is increasing with technology scaling [32]. FPGAs have significant area overhead due to programmability| hence, large portions of an FPGA are often unused. Instead of leaving these unused parts to be idle and dissipate leakage power, power gating is preferable. In power gating a large, high Vth sleep transistor is inserted between the power supply and the block to be gated. The high threshold ensures that leakage through the sleep transistor will be negligible. A control bit that can be set at configuration time or runtime controls the sleep transistor gate. Many researches have studies different points of granularity of power gating in FPGAs. Power gating at the gate level is performed in Calhoun et al. [40]; larger power gating blocks are used in Gayasen et al. [41], where gate off regions of four CLBs at a time, and that shows

leakage power savings around 20%. In addition, that coarse-grained power gating accompanied with improved placement achieves results equivalent to fine-grained power gating. Rahman et al. [42] mention that the best results can be achieved by using a combination of fine and coarse grained power gating.

Dual supply voltage: To use sleep transistors to select from different supply voltages instead of using them to disable/enable logic blocks only. Because not all paths in a circuit are critical, only elements on critical paths need to be placed in high V_{DD} regions to ensure fast operation; all other block can be on a low V_{DD} region to save power. Dual V_{DD} design has been studied extensively in the FPGA literature [43-44], typically achieving approximately 50% leakage power savings.

Dual threshold voltage and Body Biasing: Dual Vth FPGA defines low and high threshold regions at fabrication. High threshold voltage reduces leakage power but at cost of the delay. Body biased FPGAs are very similar, using embedded circuitry to change the body to the source voltage for regions at configuration time which effectively changes V_{th}. A lot of work has been performed to use dual V_{th} techniques in FPGAs [45-46] with applying body biasing in commercial FPGAs [47]. Similar effort has been introduced while mapping circuits to dual V_{th}/body biased/dual supply architectures. Critical paths should be placed in low threshold blocks to keep on the speed performance, and non-critical paths with sufficient timing slack can be place in high V_{th} regions to reduce leakage energy. Block granularity [45] and body bias voltage selection [46] are important factors for power reduction. Previous work of V_{th} and V_{DD} selection was at level of the mapped design, however, it is also useful to use different threshold for different circuits in the FPGA architecture according to each circuit role. Notably, high Vth devices can reduce the configuration SRAM bit leakage significantly without impacting area or delay [29]. Configuration bits are a good candidate for high Vth transistors because they constitute a large fraction of FPGA area and are always active and leaking. Moreover, configuration bits are configures only once at configuration time. Increasing configuration SRAM Vth has been shown to reduce leakage energy by 43% for a particular implementation [29]. Today's commercial FPGAs are fabricated with three different effective threshold voltages to reduce leakage [48].

Low Swing Interconnect: low swing interconnect segment consists of a driver, a wire operating at low voltage, and receiver operating at nominal voltages. The driver takes a full swing input and converts it to a low swing interconnect signal and the receiver reverts it back. With this technique the amount of dynamic energy dissipated in interconnect segments can be reduced significantly; for an FPGA, interconnect energy can be reduced by a factor of 2 [49].

Sub-threshold Design: Some studies show that minimal energy/operation is achieved when V_{DD} is set below the threshold voltage. Ryan et al. fabricated an FPGA designed for sub-threshold operation, demonstrating a very significant 22x energy reduction relative to a conventional FPGA at full V_{DD} [50]. The design used dual V_{DD}, low swing interconnect to reduce energy and improve delay.

Other several works targeted optimizing FPGA power from CAD tools on each step in CAD flow for FPGAs. CAD Technology Mapping: reduces dynamic power by minimizing the inter-LUTs connections switching activity, or reduces the total number of these connections thus the overall capacitance and limits duplications. Clustering: operates similar to low power technology mapping. Instead of grouping high activity gate to gate connections within the same LUT, the target is to place high activity connections between LUTs within the same CLB to utilize the more energy efficient

local interconnect. Placement: placement algorithms reduce power by attempting to place CLBs that have high activity connections between them to be close. Many researchers developed placers with modified cost functions to consider switching activity. Routing: PathFinder routing algorithm in CAD tools can be modified by modifying the cost function to consider net activity and capacitance.

Table 2.1: Previous FPGA power reduction techniques

Technique	Targeted power	Level
Technology Mapping	Total	CAD
Clustering	Total	
LUT input transformation	Static	
Placement	Total	
Routing	Total	
Glitch reduction	Dynamic	
Logic block architecture	Total	Architecture - Circuit
Interconnect architecture	Total	
Dynamic voltage scaling	Total	
Power gating	Static	
Dual V_{DD}	Total	
Dual V_{th}	Static	
Body biasing	Static	
Low swing signaling	Total	
Sub-threshold operation	Total	
Bus encoding and grouping	Total	
Transistors stacking	Static	
Double-edge flip-flop triggering	Dynamic	
Polarity selection	Dynamic	
Charge recycling	Total	
Interconnect capacitance optimization	Static	
Pulsed signaling		
Tri oxide	Static	Device
Low-k dielectric	Total	
Carbon Nanotubes	Total	
Tunnel FET	Total	
RRAM	Total	

2.5. Summary

Table 2.1 summarizes low-power FPGA techniques, the level at which they operate, and the type of power reduction. An important note is that the experimental parameters for each technique explored in prior work is very different, so the savings achieved per technique may not be directly comparable.

Chapter 3 : Device Level Power Reduction Techniques

3.1. Introduction

FPGA power optimization can be amplified by using ultra low power devices such as Tunnel-FET, FinFET, and other multi-gate devices. Those devices are able to work with ultra-low voltage which would be efficient way to reduce FPGA power consumption. In this chapter we are going to discuss about latest advanced devices which would be used to implement FPGA. FPGA circuit design will have to cope with improved leakage power and large process variations. Replacing MOSFEETs with FinFET transistors is expected to reduce the leakage power drastically. We are going to discuss the impact of using such devices on FPGA performance. Furthermore we study the usage of those circuits in FPGA for low power consumption.

Leakage power is a major part of the power consumed in nowadays FPGA or ASIC designs, hence reducing it has a significant impact on the total power consumption. This can be done at the device level by tying the body and the gate of MOSFET to form a dynamic threshold MOSFET (DTMOS). DTMOS shows ideal 60 mV/Dec sub-threshold swing, its threshold voltage is dropped as the gate voltage is raised which results in higher current drive compared to the traditional MOSFET. While it keeps high threshold voltage when it's off, hence it has low leakage. The remainder of the chapter is organized as the following: Section 3.2 discusses replacing CMOS transistors in FPGA with DTMOS and the impact of this replacement on FPGA performance metrics and shows DTMOS FPGA SPICE simulation results. Using FinFETs in FPGA and how this can reduce FPGA power/enery, and analysis of process variations impact on FinFET FPGA components various metrics accompanied with SPICE simulation results are presented in Section 3.3. Section 3.4 summarizes the chapter.

3.2. DTMOS

Dynamic threshold MOS (DTMOS) is introduced to reduce the power dissipation and improves the performance of FPGA interconnects. Some studies showed that DTMOS can reduce the FPGA interconnect power and enhance the power delay product [51-52], for instance [51] shows an average improvement of 23.35% in power delay product (PDP) of simple switch (NMOS pass transistor) and an average 32.83% enhancement in Virtex –II FPGA routing interconnects PDP can be obtained by replacing NMOS in pass switch with DTMOS. Moreover, [52] showed that by sizing DTMOS transistors properly for augmenting fixed reference voltage transistor DTMOS and augmenting DTMOS an improvement of 11.19% and 12.32% in delay, and 8.26% and 8.29% in optimum PDP for Virtex-4 90 nm FPGA can be obtained. DTMOS mainly consists of conventional CMOS with a connection from transistor gate to the body terminal. Prior studies were targeting different MUXs at 90 nm technology node

and 65nm PTM models, in this work we are studying DTMOS for more power reduction in logic blocks of FPGA using 65nm commercial technology node.

Figure 3.1 shows different DTMOS configurations, which are as the following: (a) Basic DTMOS: - It consists of NMOS transistor with the body terminal connected to the gate terminal. The gate voltage cannot exceed the diode cut-in voltage otherwise a large current would flow in the forward-biased body to source and body to drain diodes. To overcome above limitation of DTMOS some adoptions have been introduced to the basic circuit, which are as follows.

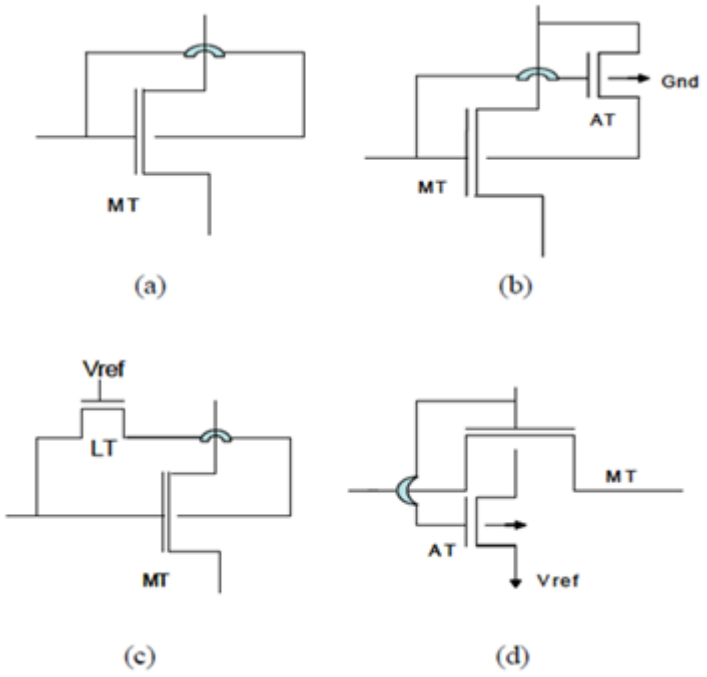


Figure 3.1: DTMOS configurations [51]

(b) *DTMOS with augmenting transistor*: - This scheme has a main transistor (MT), and augmenting transistor (AT). Drain and gate terminals of both (MT) and (AT) are shorted, so it is not applicable to share the same augmented transistor between many main transistors.

(c) *DTMOS with limiting transistor*: - This scheme has MT and a limiting transistor (LT) the gate of limiting transistor is connected to a reference voltage (Vref). During active mode, MT threshold voltage is reduced by a magnitude of (Vref - Vt). The drawback of this scheme is that the LT is always active due to the reference voltage on the transistor gate, which increases the tunneling between the gate and the oxide when the switch is inactive therefore this scheme has the highest standby leakage among the other schemes.

(d) *DTMOS with augmenting fixed reference voltage transistor*: - This scheme has MT and AT with fixed reference voltage (Vref) applied to AT's drain. MT's body bais equals (Vref - Vt) while it is ON, since only the two transistors gates are connected,

hence it is feasible to share a single AT with many main transistors which is preferable in routing based on multiplexer switches in which a select line drives a large tree of NMOS pass transistors. This scheme decreases the area overhead by a large margin. It can be further divided as follows.

SMSA- (MT and AT have standard threshold 'SVt') this switch consumes low power, its delay is also low, but due to SVt of both transistors its leakage consumption is highest.

SMHA- (MT has standard threshold 'SVt' and AT has high threshold 'HVt' voltage) this switch's delay is slightly higher than SMSA delay since AT has a delay in biasing the body because of 'HVt', but its leakage is less than SMSA switch.

HMSA- (MT transistor has high threshold 'HVt', and the AT transistor has standard threshold 'SVt' voltage) this switch's delay is higher than SMHA due to MT high threshold voltage but this scheme has a comparable PDP to SMHA scheme.

HMHA- (MT and AT have high threshold voltage 'HVt') accordingly, this switch's delay and PDP are the highest. PDP is used as the metric to get the optimum switch scheme from above DTMOS switches.

Figure 3.2 illustrates the simulation setup used and Figures 3.3 to 3.7 show the different FPGA routing MUXs configuration, in Table 3.1, covered in prior study [51] and the corresponding power delay product result for each.

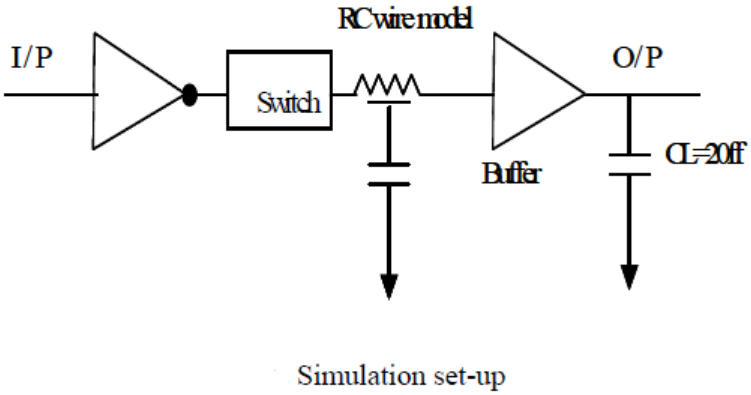


Figure 3.2: Simulation setup [51]

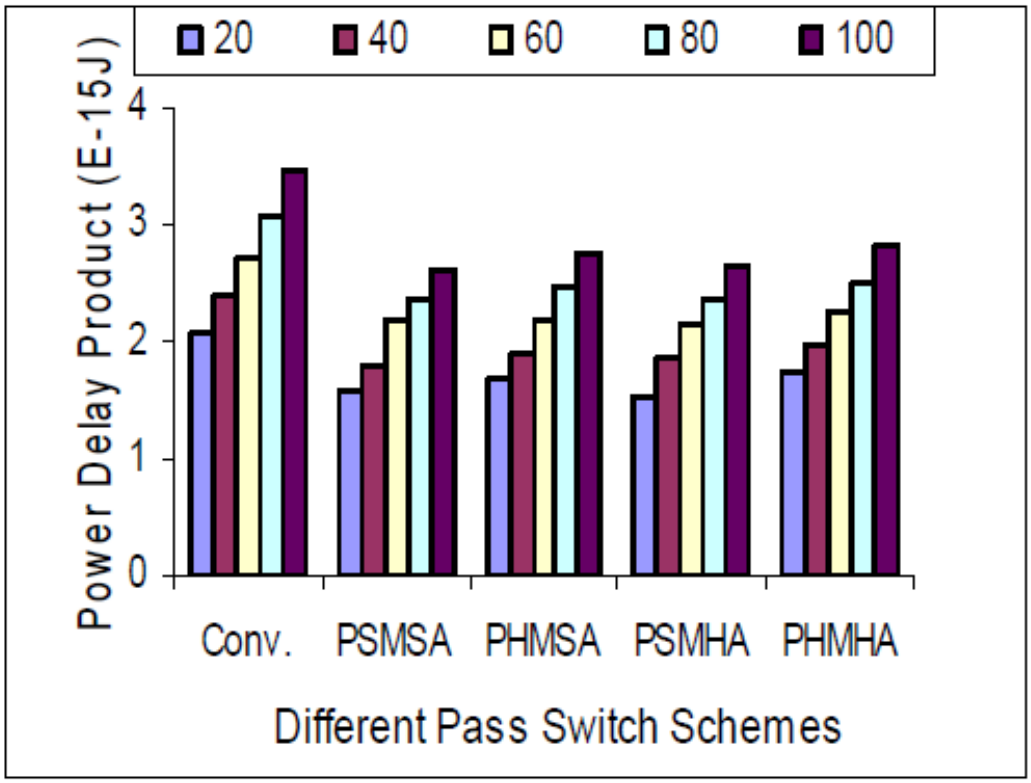


Figure 3.3: PDP of pass switch vs. DTMOS schemes [51]

Table 3.1: Major interconnects in the switch box [51]

Circuit	Description
IMUX	30:1 MUX and buffer
OMUX	24:1 MUX and buffer
DOUBLE	16:1 MUX and buffer
HEX	12:1 MUX and buffer
LONG	n:1 MUX and buffer

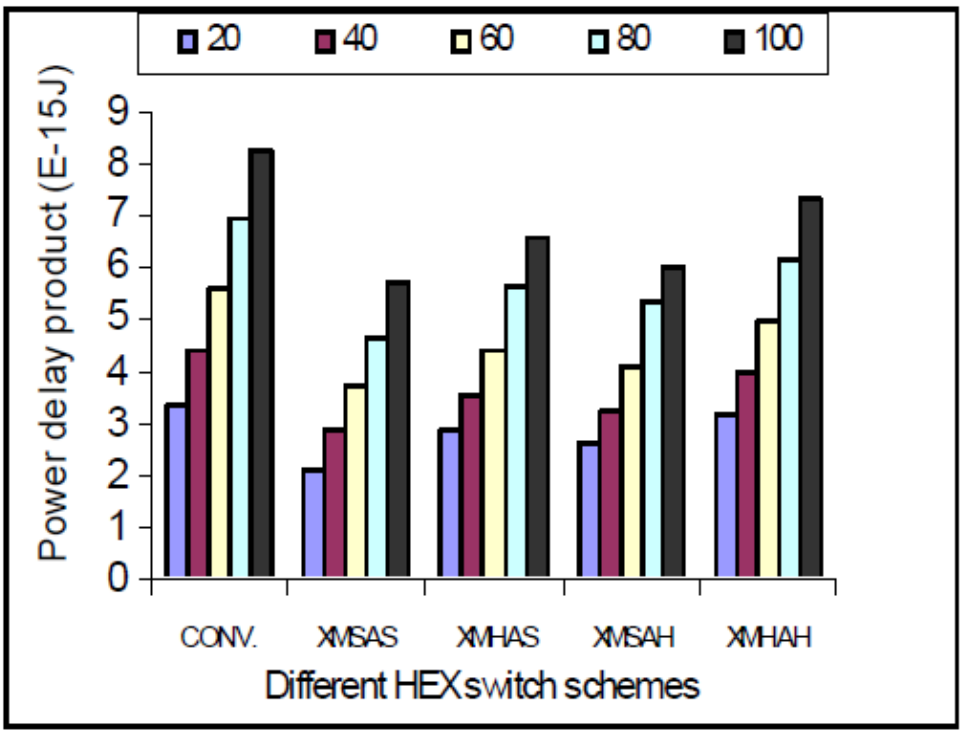


Figure 3.4: PDP vs. HEX switch schemes [51]

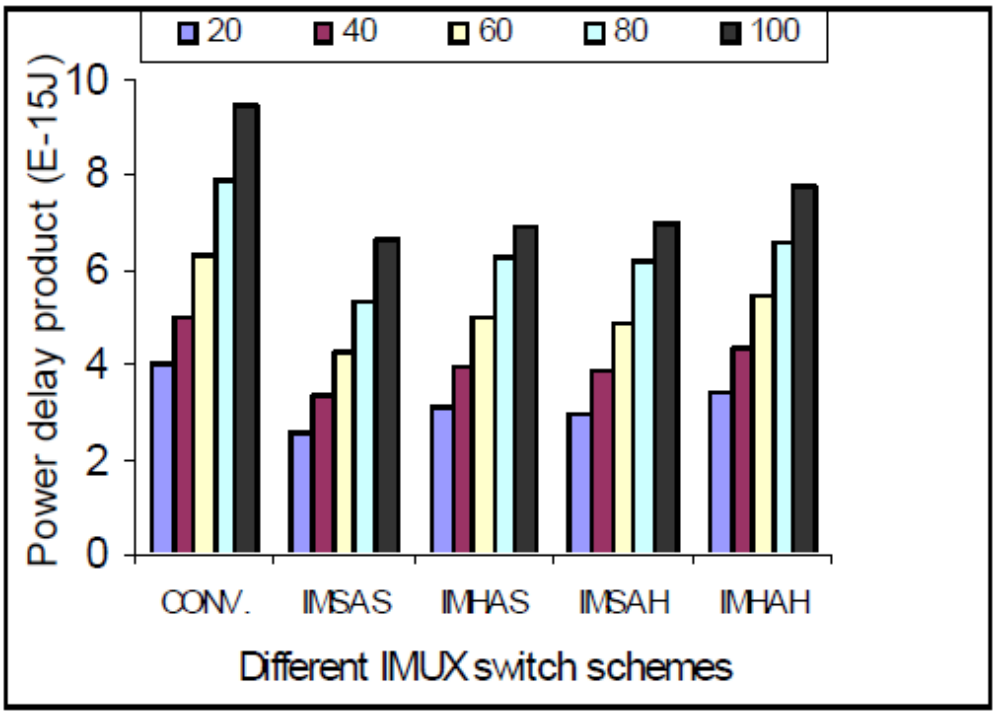


Figure 3.5: PDP vs. IMUX switch schemes [51]

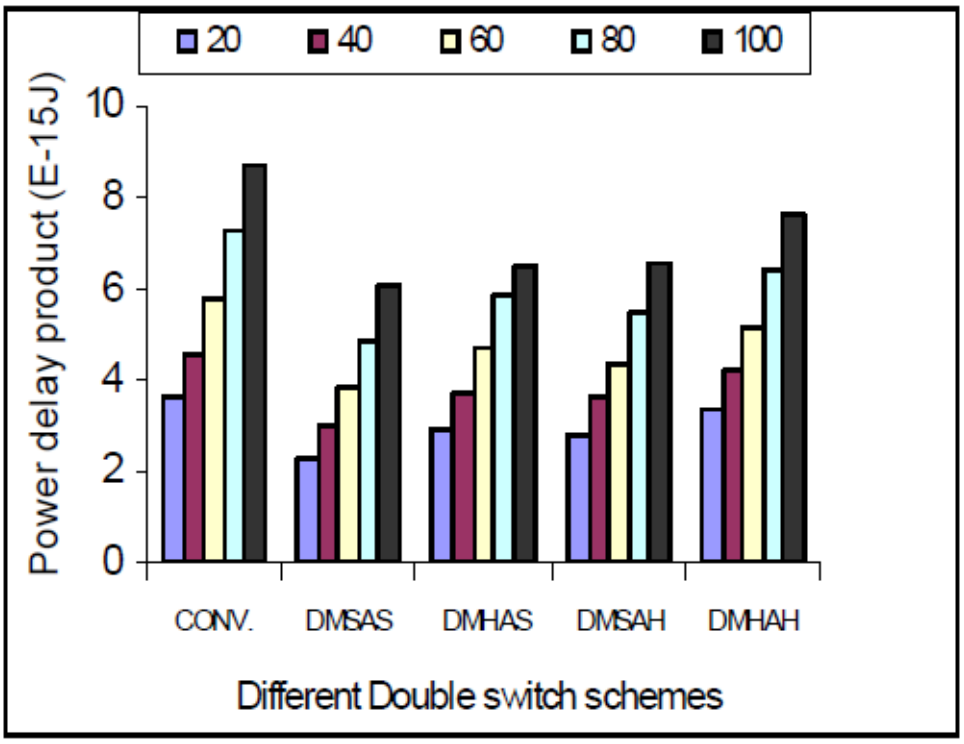


Figure 3.6: PDP vs. double switch schemes [51]

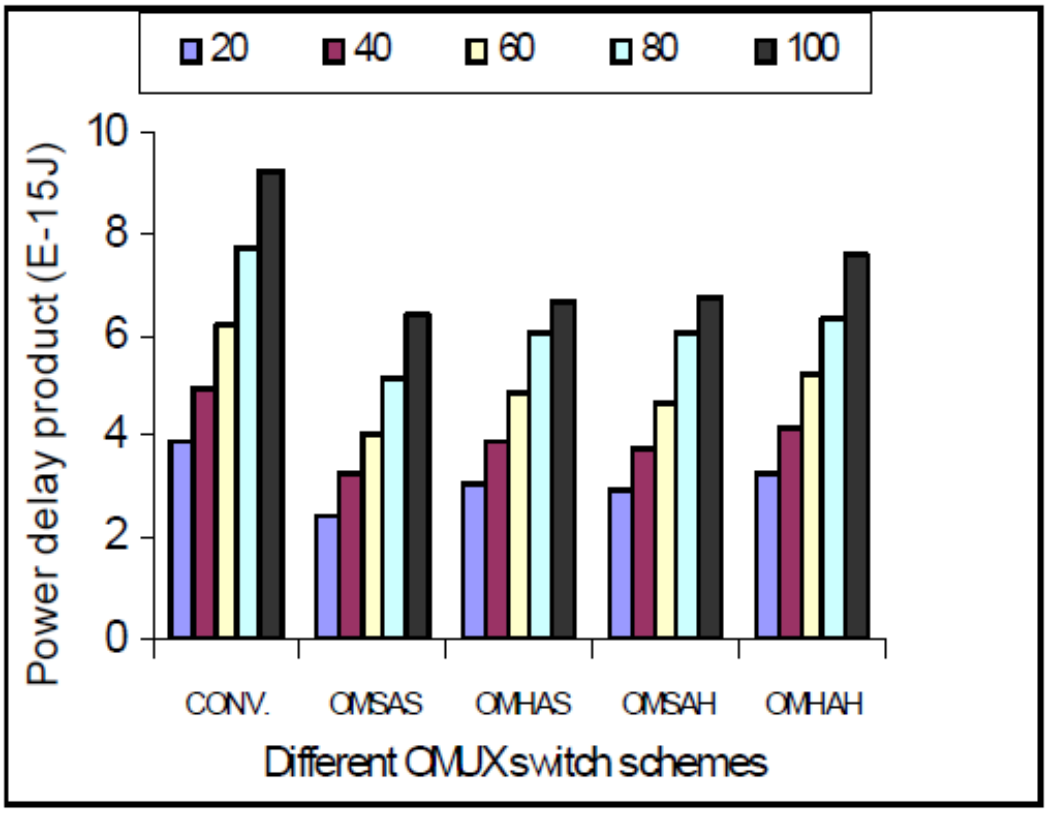


Figure 3.7: PDP vs. OMUX switch schemes [51]

The average reduction using these schemes (with respect to traditional interconnect) are 33.27%, 31.80%, 32.8 and 33.38% for HEX, IMUX, DOUBLE and OMUX interconnects respectively using 65nm Berkeley PTM technology.

Figures 3.8 to 3.10 show the delay, power, and power delay product of 4-input MUX at 90nm with different transistors sizing. Results showed up 12.32% and 11.19% in delay, also 8.29% and 8.26% in PDP can be achieved for augmenting transistor DTMOS and augmenting fixed reference voltage transistor DTMOS respectively [52].

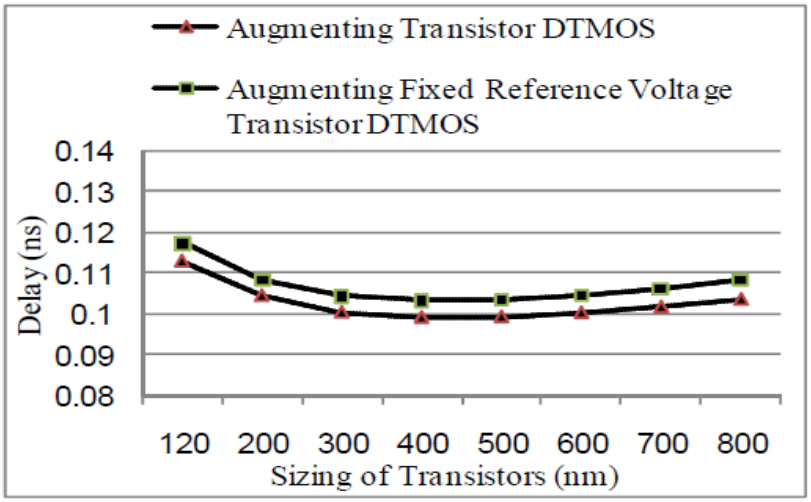


Figure 3.8: Delay vs. size of Augmented and Augmented with fixed reference DTMOS [52]

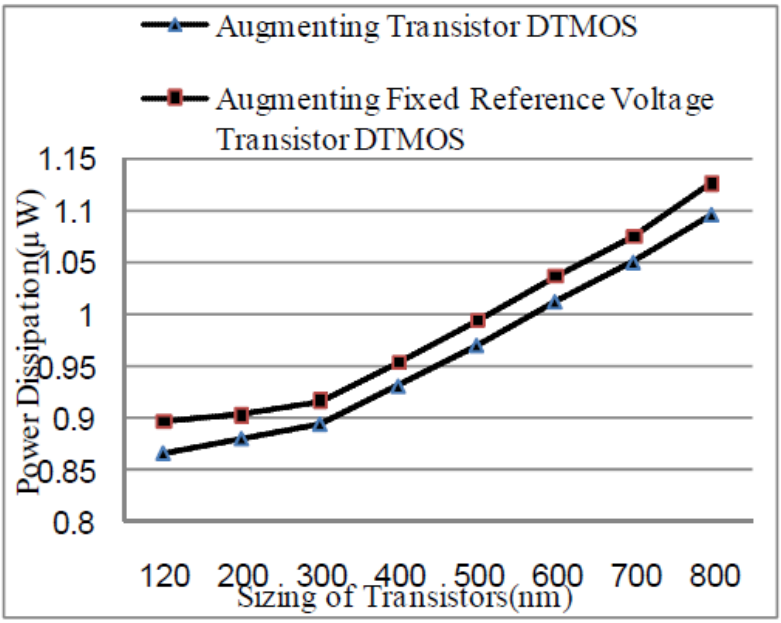


Figure 3.9: Power vs. size of Augmented and Augmented with fixed reference DTMOS [52]

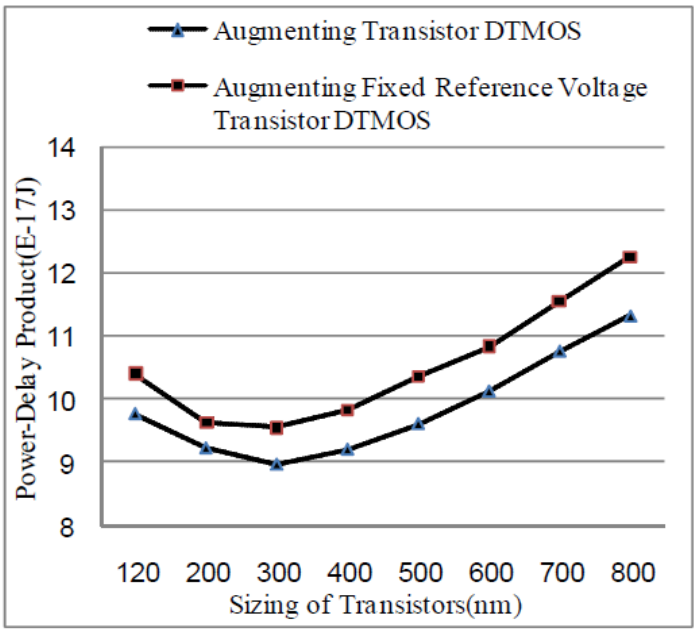


Figure 3.10: PDP vs. size of Augmented and Augmented with fixed reference DTMOS [52]

In this work we investigated on delay, power, and power delay product for 4:1 MUX and 16:1MUX for commercial 65nm process using Cadence virtuoso and Spectre simulator. DTMOS MUX consumes less power than CMOS as shown in Figures 3.11 to 3.14. For instance, 16:1 DTMOS MUX power dissipation is 9.4 times lower than equivalent CMOS MUX, however delay is 6.1 times CMOS delay, thus PDP is reduced by 35%.

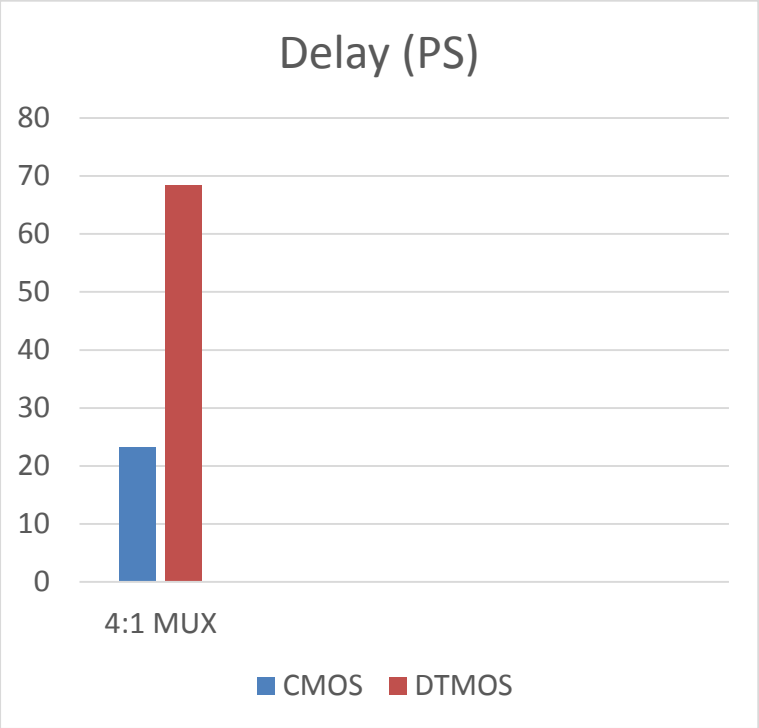


Figure 3.11: 4-1 MUX delay DTMOS vs. CMOS

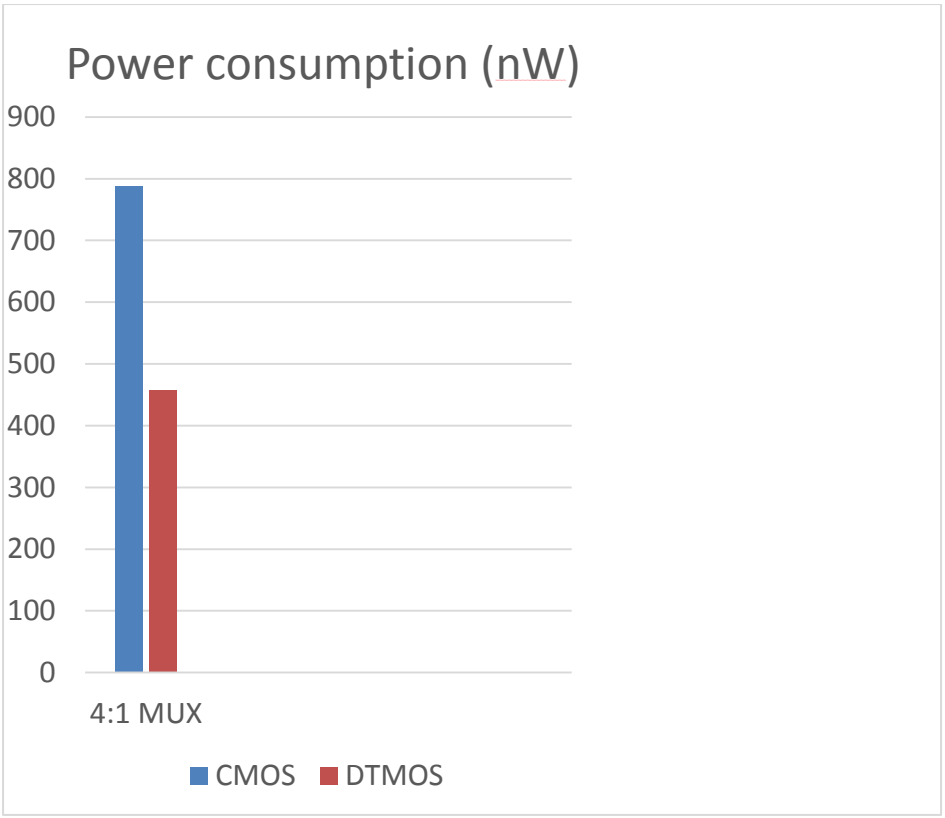


Figure 3.12: 4-1 MUX power DTMOS vs. CMOS

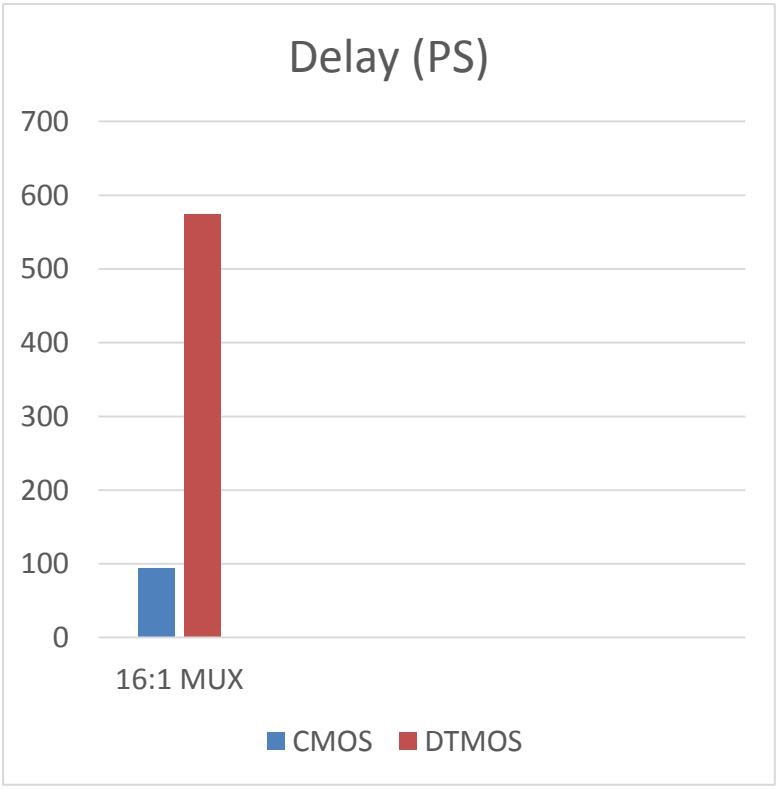


Figure 3.13: 16-1 MUX delay DTMOS vs. CMOS

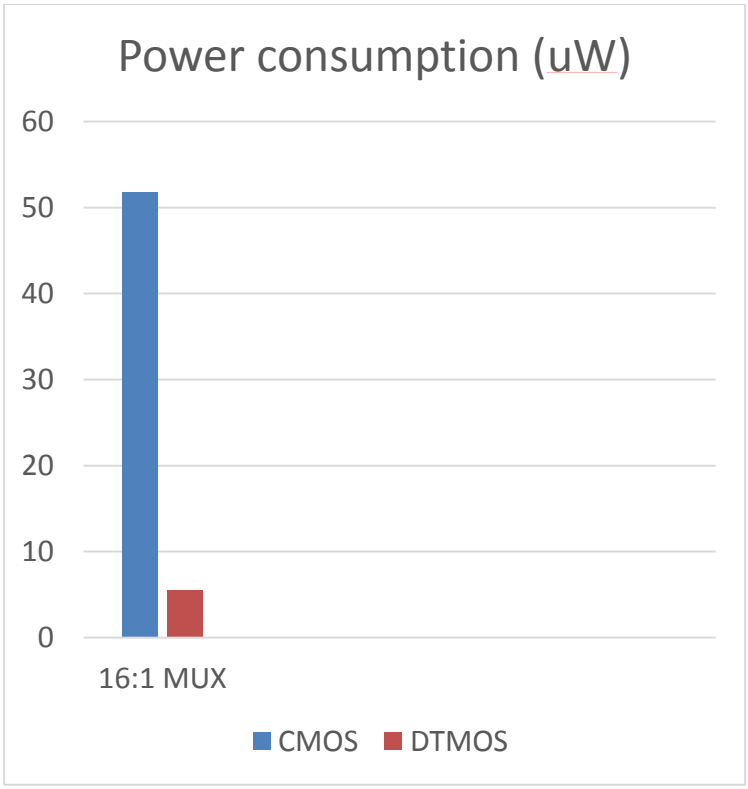


Figure 3.14: 16-1 MUX power DTMOS vs. CMOS

3.2.1. DTMOS FPGA structure

In addition to using DTMOS in FPGA interconnects, we studied using DTMOS in FPGA logic blocks. The simulated architecture is shown in Figure 2.2.

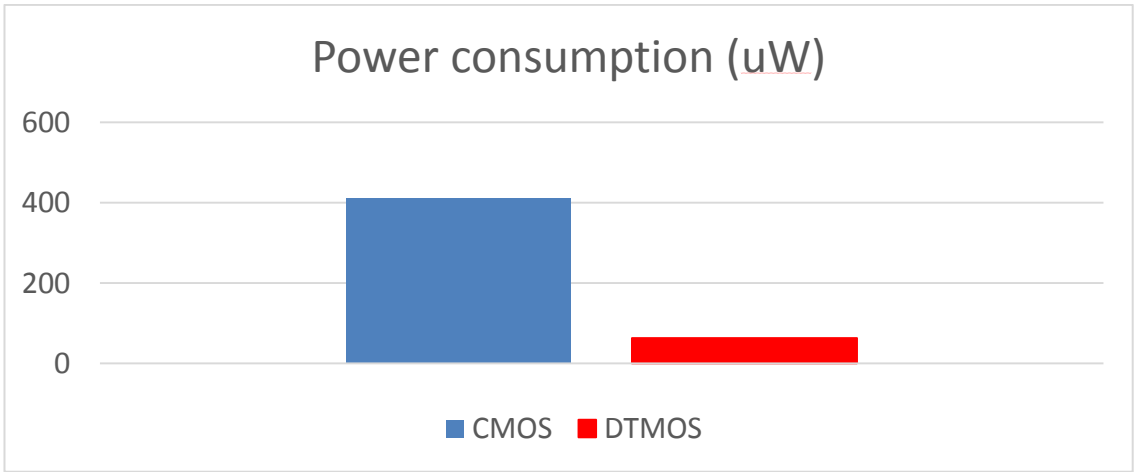


Figure 3.15: CLB power DTMOS vs. CMOS

DTMOS FPGA cluster reduces power reduction by 85% compared to equivalent CMOS cluster, and power delay product (energy) improvement by 3% while configuring the cluster to implement 4-inputs NAND gate as a benchmark as described in Figure 3.15.

3.3. FinFET

FinFET has been introduced as the promising alternatice for the traditional bulk CMOS device for nanoscale technologies, due to its extraordinary characteristics such as high ON to OFF current ratio, reduced short channel effects, relative immunity to gate line-edge roughness, and enhanced channel controllability. Furthermore, the semi ideal sub-threshold behavior gives an indication to the potential application of FinFET circuits in near-threshold supply circuits, which dissipates an order of magnitude less energy compared to regular strong inversion circuits that operate with the super-threshold supply voltage. Figure 3.16 shows some of Multi-gate devices like double-gate MOSFET, tri-gate, and gate-all-around (GAA) FET.

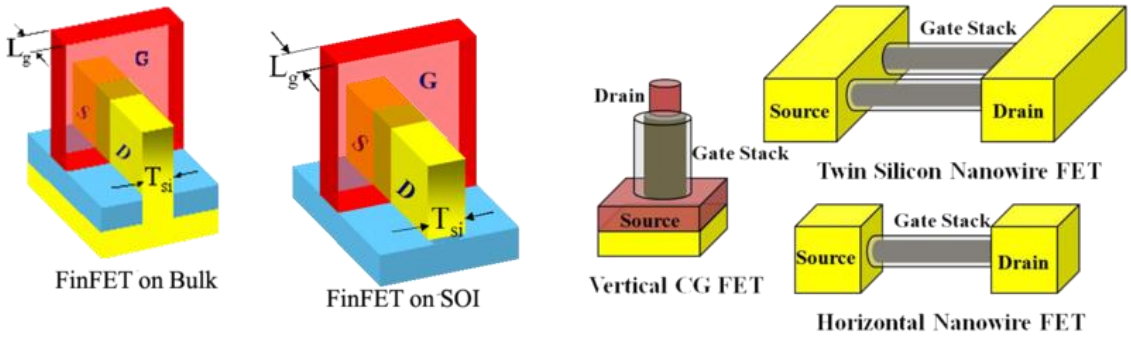


Figure 3.16: Multi-gate transistors

3.3.1. Predictive technology models

Predictive technology models (PTM) are very critical for early stage co-optimization of design-technology and circuit design research. Predictive technology models are developed by ASU university with co-operation with intel [17] based on MOSFET scaling theory, the 2011 ITRS roadmap and early stage silicon data from published results and verified against those data.

Figure 3.17 shows saturation drive current of PTM-MG models normalized per effective width (Weff) for a constant off-current (Ioff=0.1nA/um for LSTP and 100nA/um for HP) compared to On/OFF currents in ITRS specifications. The PTM-MG LSTP devices follow the ITRS LSTP trend but are shifted to be slightly stronger

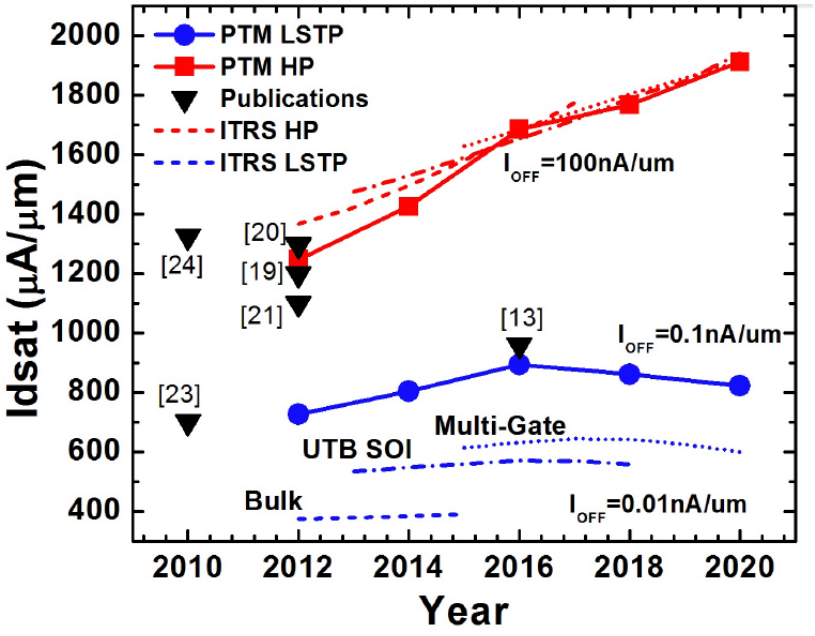


Figure 3.17: Saturation current of PTM vs ITRS [17]

Figure 3.18 shows the top and cross-sectional view of a Fin-FET. The dimensions are labeled using the corresponding BSIM-CMG model parameters. The parameters used in PTM-MG development are listed in Table 3.2. The behavior of a FinFET device is most sensitive to the primary parameters, technology specifications and physical parameters. The secondary parameters are useful to fine-tune a fit to the complete current-voltage characteristics or capture secondary effects.

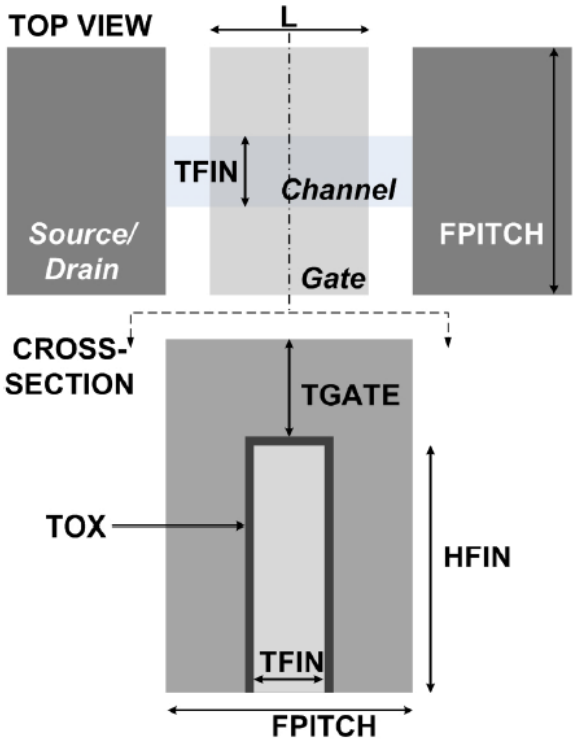


Figure 3.18: FinFET cross section [17]

Table 3.2: PTM-MG models parameters [17]

Primary parameters	
L	Gate length
TFIN	Fin thickness
HFIN	Fin height
FPITCH	Fin pitch
Technology parameters	
EOT	Equivalent oxide thickness
V_{DD}	Supply voltage
R_{DS}	Source/drain resistance
Secondary parameters	
PHIG	Gate work function
NBODY	Channel doping
CDSC	SD-channel coupling
E_{ta0}	DIBL coefficient
Physical parameters	
μ_0	Low field mobility
V_{sat}	Saturation velocity

Figure 3.19 shows a flowchart describing PTM-MG model development. The details of each step are described in [53].

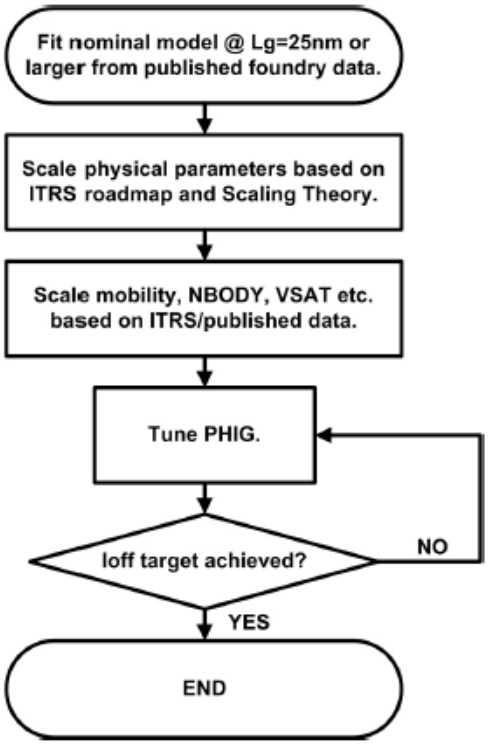


Figure 3.19: PTM models generation flow [17]

PTM-MG used the published results from foundries such as Intel, TSMC, and IBM [54-57] to extract the fitting PTM parameters such as DIBL, sub-threshold slope by fine-tuning both primary parameters (Gate length, Fin thickness, Fin height, and Fin pitch) and secondary parameters (Gate work function, channel doping, source-drain channel coupling, and DIBL coefficient) [53] to match on-current and off-current of the published results.

For future technologies (Beyond 14nm) PTM-MG model cards are developed using ITRS as a reference. The off-current for 14nm technology node and below is expected to be (I_{off}=0.01nA/um for LSTP and 100nA/um for HP) according to ITRS trends [58]. The difference between ITRS off-current and PTM off-current impact on transmission gate flip-flop (TG-FF) metrics is evaluated and plotted in Figures 3.20 to 3.22. We used Cadence virtuoso with Spectre simulator for all simulations.

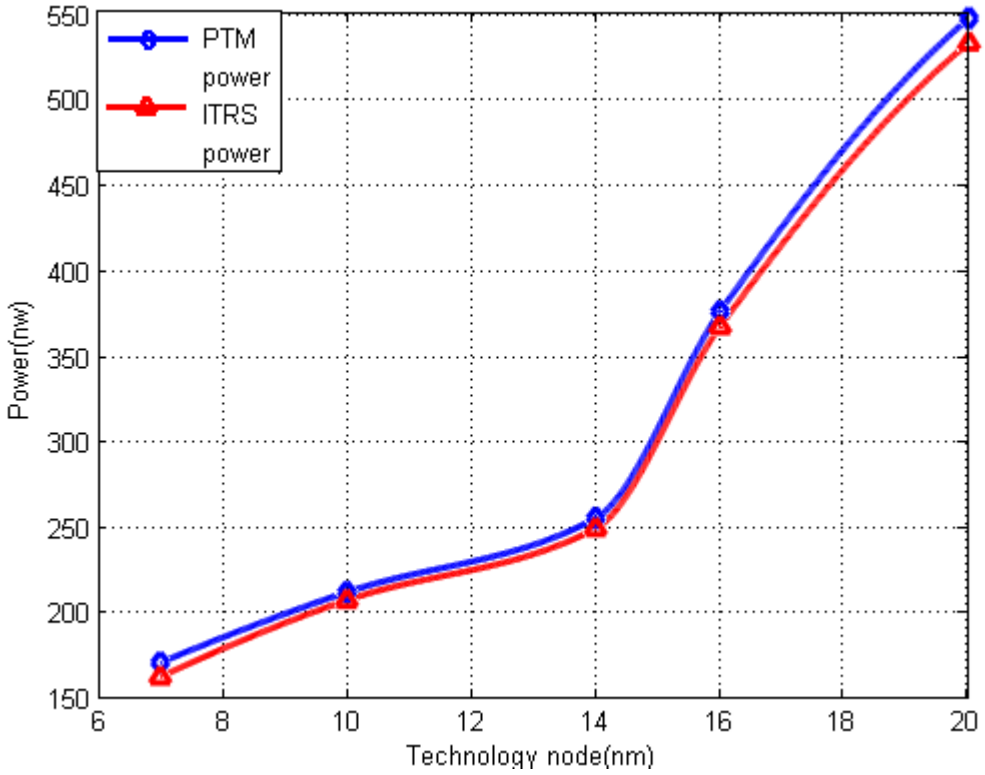


Figure 3.20: The difference between ITRS off-current and PTM off-current impact on TG-FF power [19]

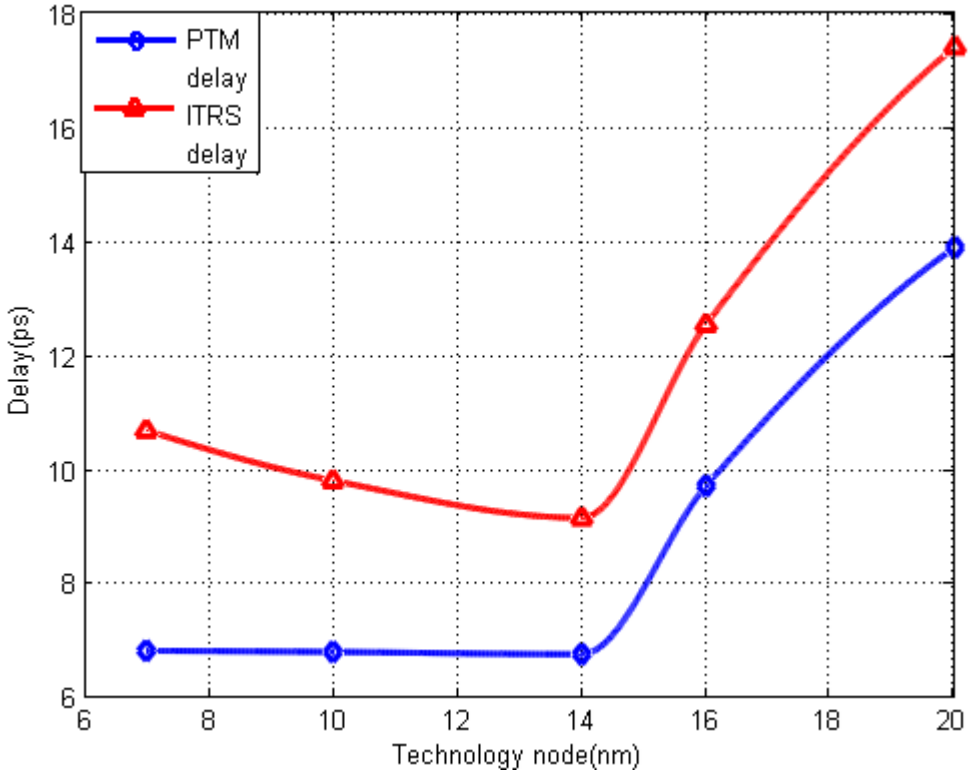


Figure 3.21: The difference between ITRS off-current and PTM off-current impact on TG-FF delay [19]

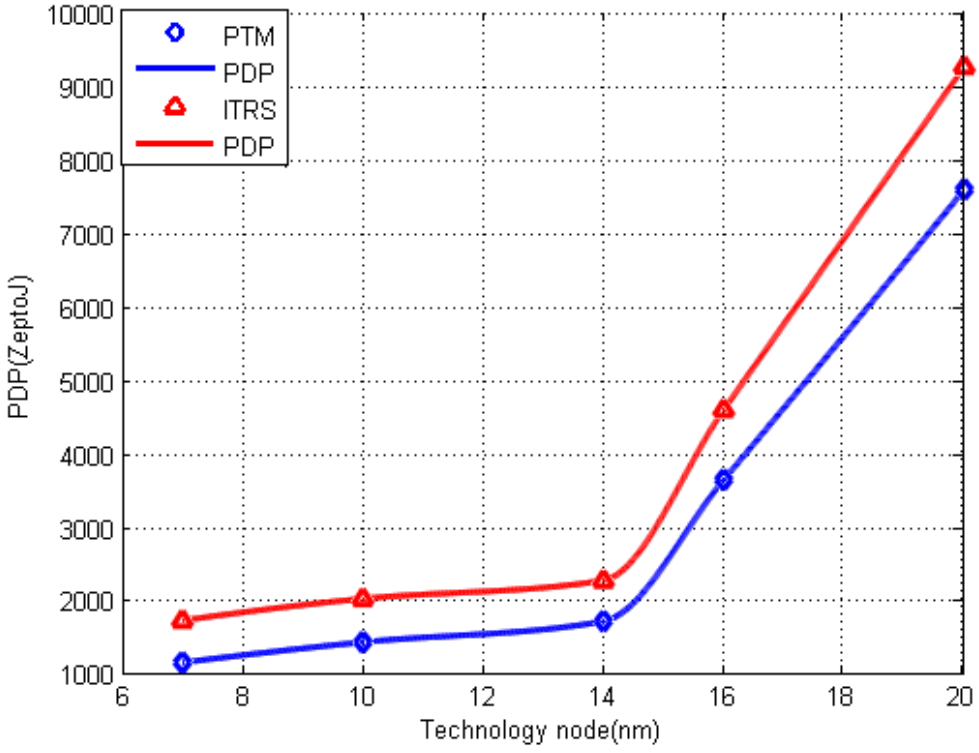


Figure 3.22: The difference between ITRS off-current and PTM off-current impact on TG-FF PDP

This means that simulation results using nominal PTM-MG parameters have slight deviation from fabricated devices with ITRS off-current. For instance, 7nm PTM TG-FF has power with 5% deviation from similar device with ITRS off-current.

3.3.2. FinFET FPGA components

3.3.2.1. Critical path simulation

Multiplexer is the main building block In FPGAs, since it's used in both logic blocks and routing switches [59], and it is essential building block for datapaths.

Reliability is one of the important design challenges for sub-micron technologies is. Since decreasing supply voltages, shrinking geometries, increasing clock frequencies, and increasing density of circuits all have a great impact on reliability. Variations in random dopant fluctuation (RDF), critical dimensions (CD), voltage variations, bias temperature instability (BTI), temperature variations, and hot carrier injection (HCI) have direct effects on device threshold voltage [60-62].

Some researchers studied designing high performance low power multiplexers [63], and benchmarking 7nm FinFET designs using predictive technology models [64]. In this work, the performance as well as the power dissipation of critical path (RO) are evaluated alongside with variations which results from fabrication factors mentioned above within a certain range of threshold voltage variations at different technologies (20nm, 16nm, 14nm, 10nm, and 7nm).

We used PTM models for low-standby power devices (LSTP) [17] that are based on BSIM-CMG as multi-gate devices (Tri-gate FinFET) for 20 nm down to 7 nm technology nodes. We adopted scaling strategy according to PTM models, that involves: channel length (L), supply voltage (V_{DD}) scaling, fin height (H_{fin}), and fin thickness (T_{fin}). For tri-gate FinFET, channel width:

$$W = N_{fin}(2H_{fin} + T_{fin}) \qquad (1)$$

The nominal threshold voltage is extracted for each technology node by using the first derivative of Ids-Vgs characteristic curve as highlighted in Figure 3.23. We considered threshold voltage variations in a range of ±18% of the nominal value with 6% step. The parameters of the simulated FinFET device are tabulated in Table 3.3.

As technology advances, the nominal threshold voltage decreases due to supply voltage decrease to satisfy technology requirements of performance and low power. The studied ring oscillator consists of 16 to 1 pass transistor logic multiplexer cascaded with three stages of logic gates (2-inputs NAND, 2-inputs NOR, and INV) as shown in Figure 3.24.

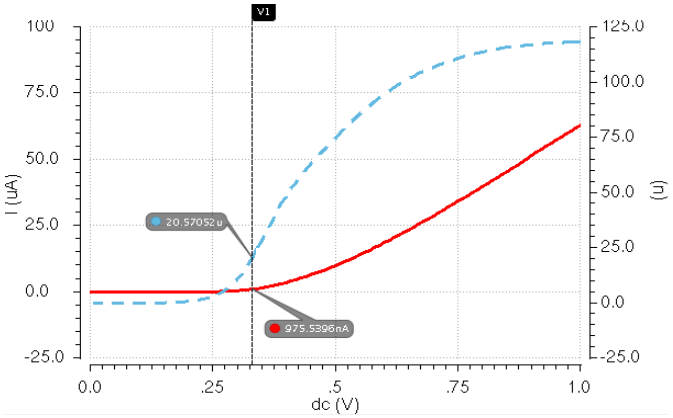


Figure 3.23: Extraction of Threshold voltage, the dashed line is the first derivative of the drain current, the solid line is the drain current

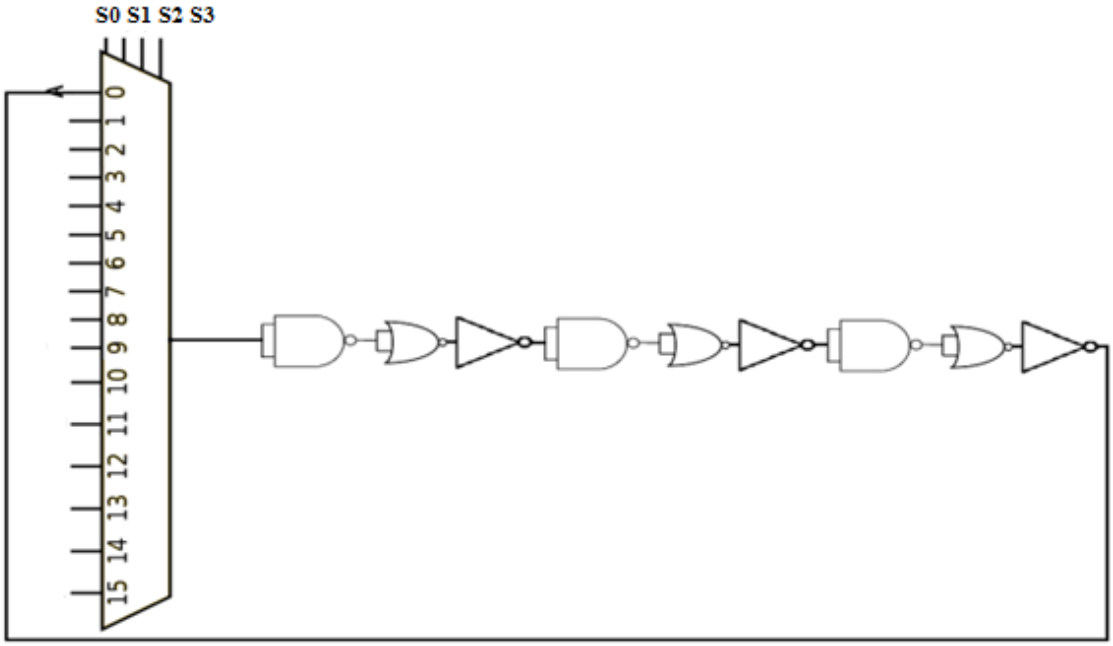


Figure 3.24: Critical path simulation setup [18]

HCI, and BTI affect the device threshold voltage, typically they shift the threshold voltage within ±18% of its nominal value. Temperature, RDF, and variations in CD also shift the device threshold voltage.

The RO is sized with minimal dimensions. Nominal supply voltages, and channel lengths are set according to PTM low-standby power models and are reported in Table 3.3.

Table 3.3: The simulated device parameters

Device	**TG-FinFET**				
L(nm)	20	16	14	10	7
T_{fin}	15	12	10	8	6.5
H_{fin}	28	26	23	21	18
N_{fin}	1	1	1	1	1
V_{DD} (V)	0.9	0.85	0.8	0.75	0.7

Figures 3.25 and 3.26 show that the speed performance of the RO is enhanced with technology scaling, however it degrades beyond 14nm, however using a higher V_{DD} would lead to enhanced performance at the cost of reduced power. Using high mobility channel [65] and/or using gate-all-around (GAA) nanowires [66] options have the potential to enhance device scaling in that time frame, and are important areas for future work. For instance, the time period at the nominal threshold value at 7nm is higher than its value at 14nm technology by 18%, however power is reduced by 1.76.

Performance sensitivity, which is defined as the Δfrequency/nominal frequency, increases with technology scaling. For instance, the time period corresponding to +18% threshold variation at 7nm is 1.05 times the nominal value, however it's 0.6 times the nominal value at 20nm technology.

Performance trends are also improved as temperature increases and degraded as threshold voltage increases. For instance, the time period at 7nm and 120° is lower than at 27° by 0.45.

Trends of power consumption are enhanced with technology scaling as illustrated in Figures 3.27 and 3.28 due to supply voltage scaling as well as other technology scaling effects. For instance, the nominal threshold voltage power at 7nm is lower than its value at 20nm by 0.43. In addition, power sensitivity, which is defined as the Δpower/nominal power, has a proportional relationship with technology scaling. For instance, +18% threshold variation power at 7nm is reduced by 0.52 of its nominal value, however corresponding power at 20nm technology is reduced by 0.41. Power consumption trends are enhanced with threshold voltage increase, but degraded with temperature increase. For instance, the power at 120° at 7nm is 1.2 times 27° power.

Since the power consumption and the speed always have a trade-off, power delay product is used as a metric. Observing PDP trends with technology scaling, it's is continuously improving (decreasing) with technology scaling as shown in Figures 3.29 and 3.30. For instance, nominal threshold PDP at 7nm is lower than at 20nm technology by 0.3.

Trends of PDP are improved as threshold voltage increase, but degraded as temperature increase. For instance, the PDP at 120° at 7nm is higher than at 27° by 0.2.

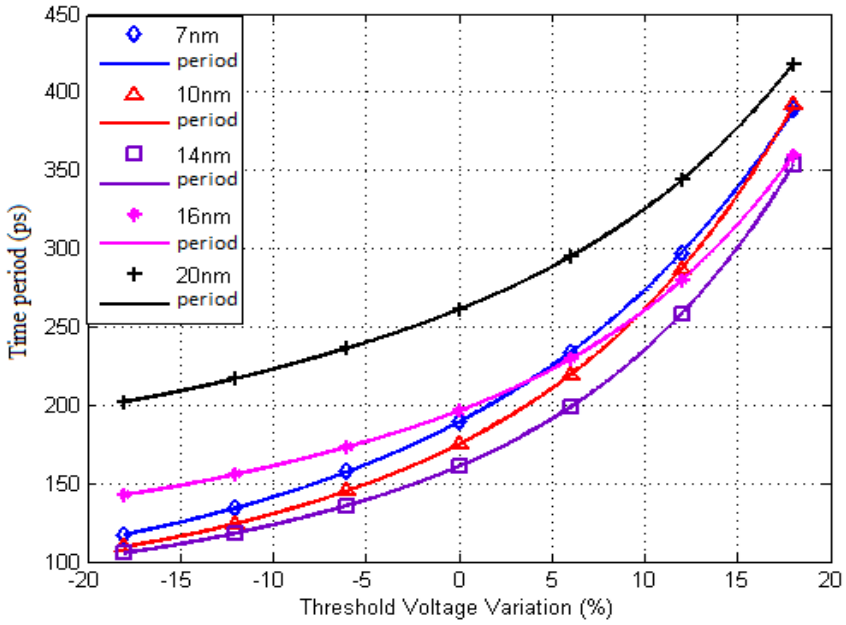


Figure 3.25: RO time period vs. Threshold voltage [18]

Ring oscillator frequency is a mandatory metric in evaluating the performance of the critical path in digital designs. It is obvious that 14nm technology exhibits the optimum performance because of its large saturation current, beyond 14 nm technologies performance degrades which leads to searching for alternative device scaling options such as using high mobility channel [65] and/or using gate-all-around (GAA) nanowires [66]. Despite expecting that the current per unit width to increase as technology advances, however RO current is decreasing due to the adopted scaling strategies to minimize SCEs, since scaling T_{fin}, and H_{fin}, and reducing the effective channel width. The time period increases with the threshold voltage increase due to driving current limiting.

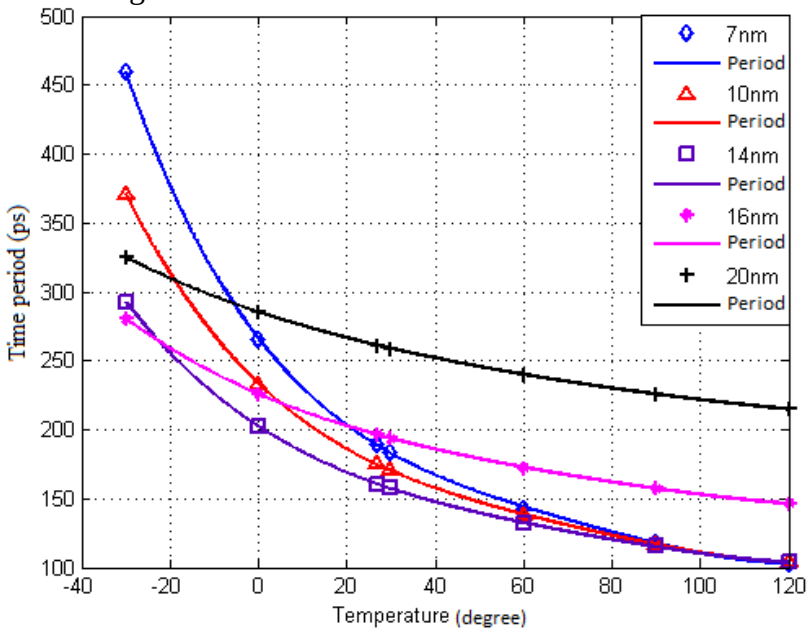


Figure 3.26: RO time period vs. Temperature [18]

Ring oscillator driving current increases as temperature increases, hence the time period is decreased. For instance, the time period at 120° at 7nm is less than at 27° by 0.45.

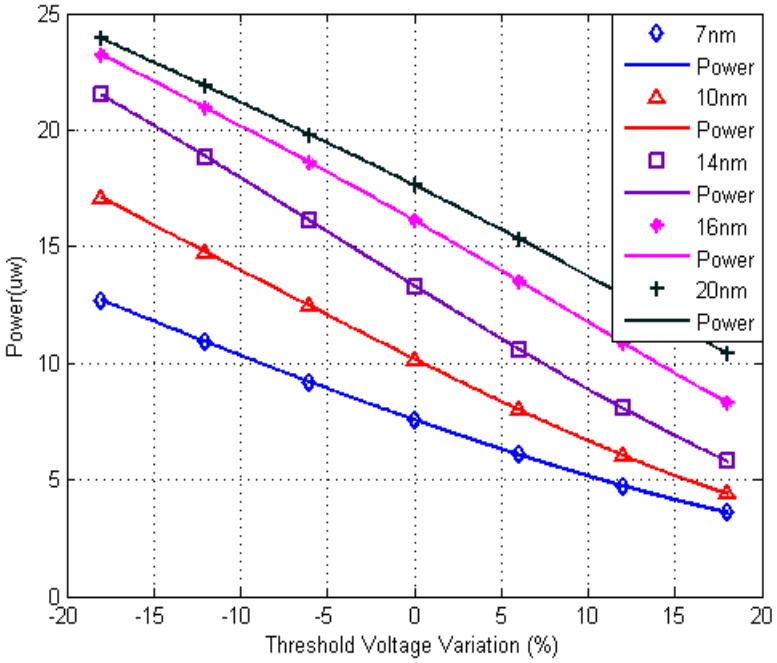


Figure 3.27: RO power vs. Threshold voltage [18]

The demand on Low power designs increased dramatically nowadays since mobile devices have a limited budget of energy and also to reduce the cost of cooling. From our study, ring oscillator power dissipation is decreased with technology scaling. For instance, the nominal threshold power at 7nm is lower than at 20nm by 0.43.

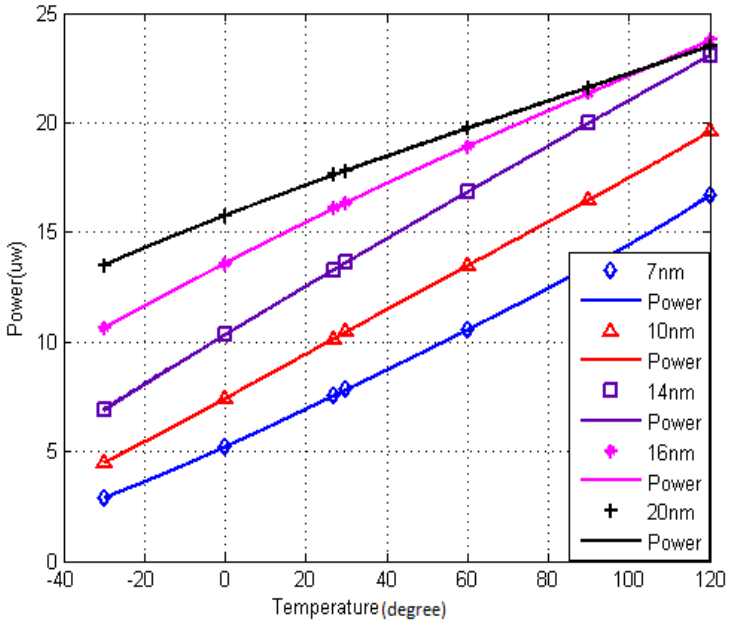


Figure 3.28: RO power vs. Temperature [18]

Ring oscillator power dissipation has proportional relationship with temperature, since the dissipated power of RO is increased as driving current increases. For instance, the power at 120° at 7nm is higher than at 27° by 20%.

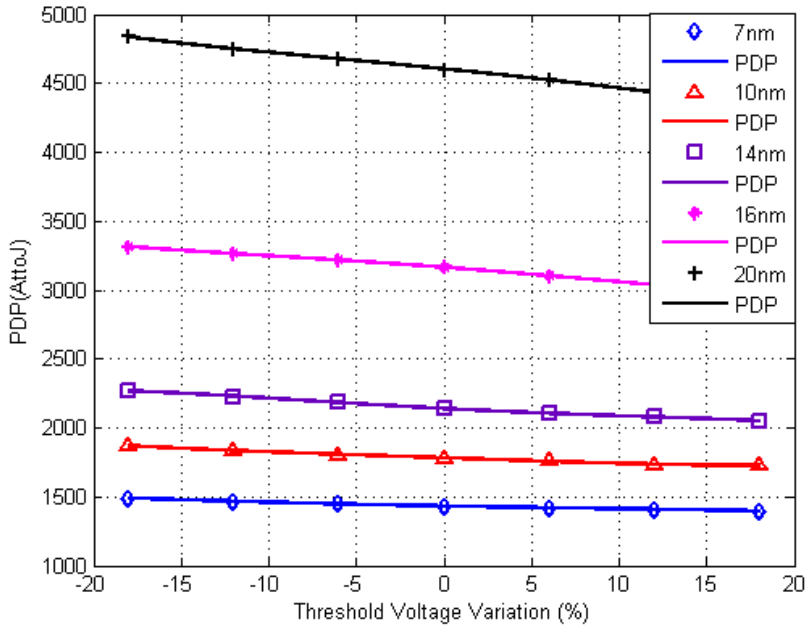


Figure 3.29: RO PDP vs. Threshold voltage [18]

In addition, PDP trends of RO are decreasing with technology scaling continuously. For instance, nominal threshold PDP at 7nm is less than at 20nm technology by 0.3.

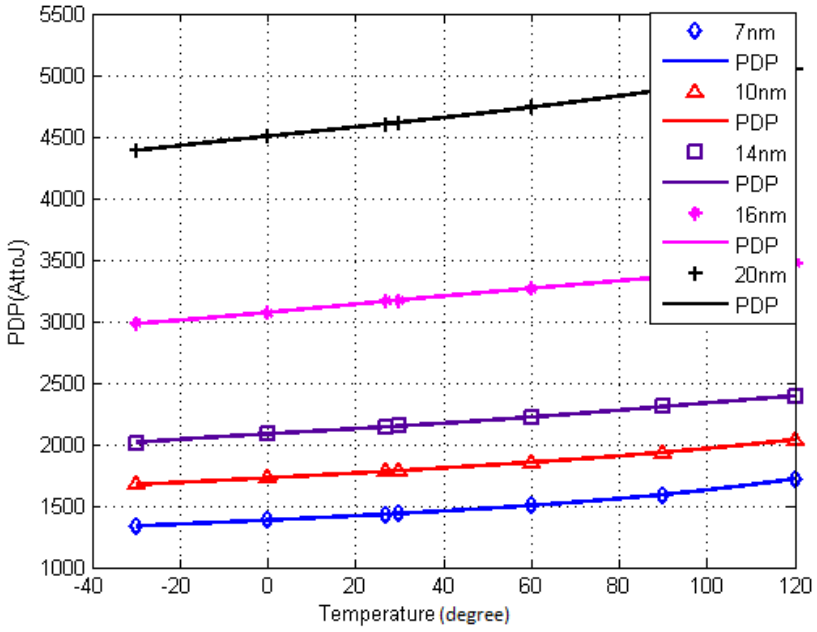


Figure 3.30: RO PDP vs. Temperature [18]

PDP of RO has increases as temperature increases, because of power dominance. Also PDP sensitivity has a proportional relationship with technology scaling. For instance, the PDP at 120° at 7nm is higher than the nominal value by 0.2, however the PDP at 20nm technology at same temperature is increased by only 0.1.

From this study, we can observe enhancement of performance with technology scaling till 14nm technology node. Power consumption, and PDP decrease with technology scaling.

Threshold voltage increase limits the power consumption, however it has negative effect on the performance, but the PDP is enhanced as a total. Temperature increase increases the power consumption, and the PDP, however the performance is enhanced due to the increase of driving current.

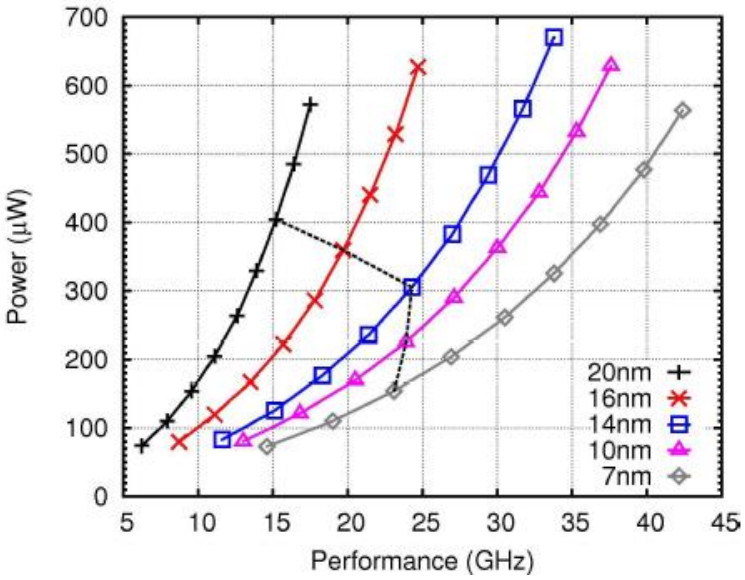


Figure 3.31: PTM models power vs. performance [17]

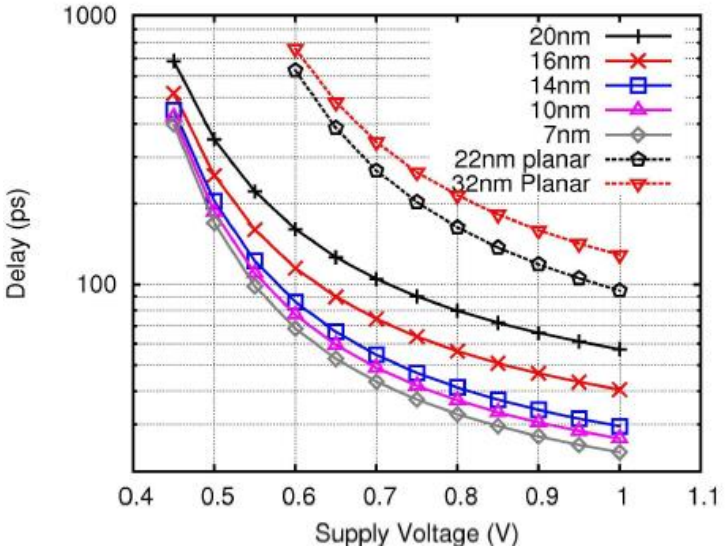


Figure 3.32: PTM models delay vs. supply voltage [17]

Performance of RO is degraded with technology scaling beyond 14nm due to supply scaling at these technology nodes as shown in Figure 3.31, as with the same supply voltage for all technology nodes the performance trend is enhanced even beyond 14nm as illustrated in Figure 3.32. Simulation results in Figures 3.33 to 3.38 illustrate the increase of sensitivity with technology scaling. For instance at 7nm node, the RO delay/power/energy change response to temperature increase is higher than the corresponding change at 20nm technology node.

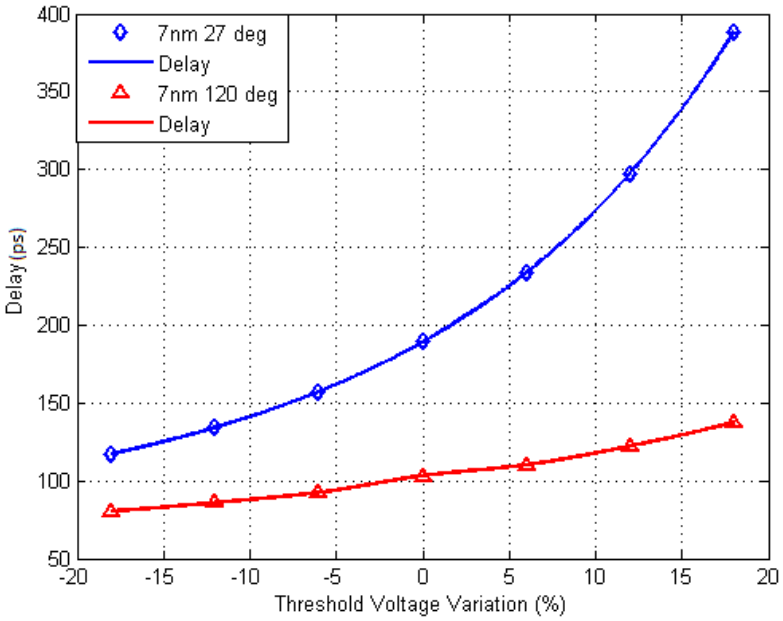


Figure 3.33: 7nm delay vs. threshold voltage at 27° and at 120°

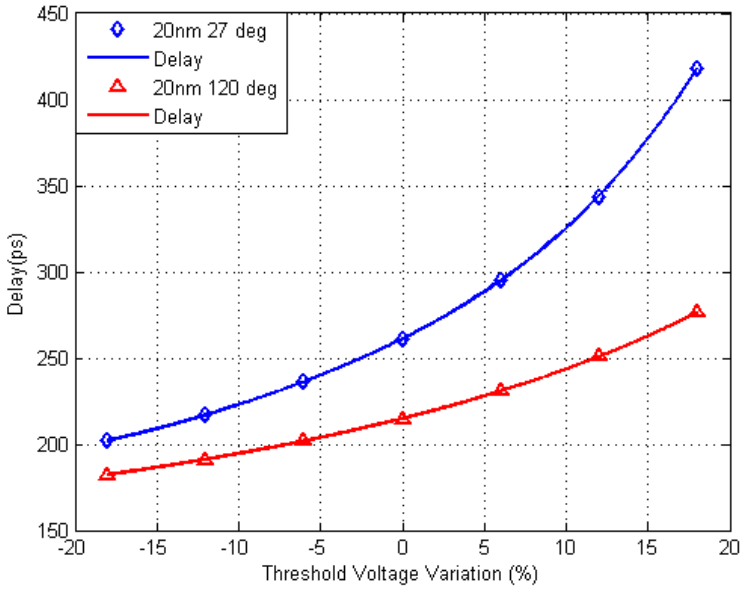


Figure 3.34: 20nm delay vs. threshold voltage at 27° and at 120°

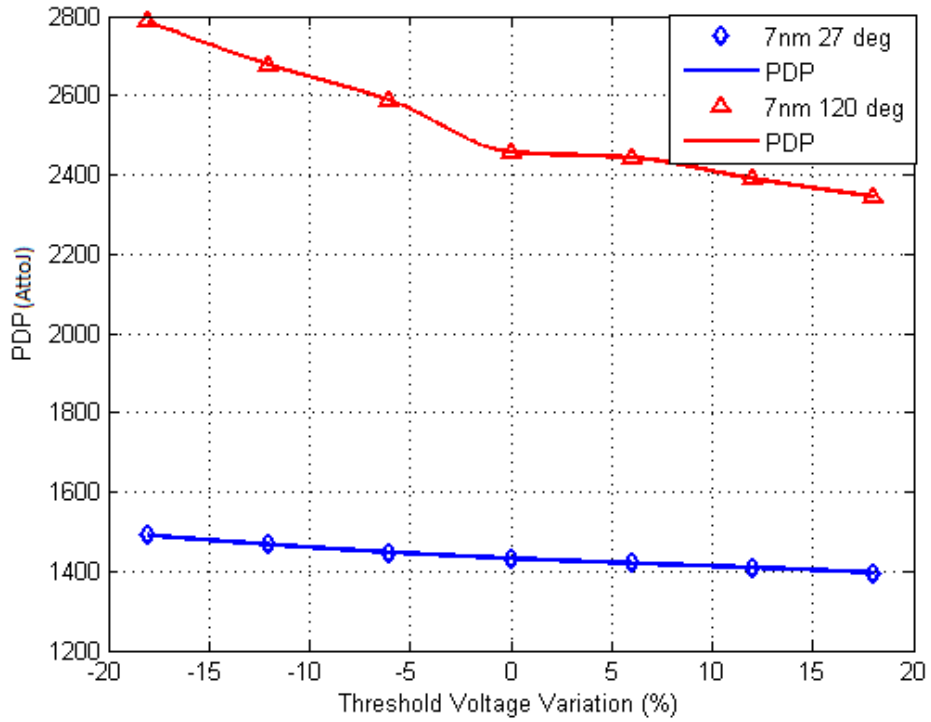


Figure 3.35: 7nm PDP vs. threshold voltage at 27° and at 120°

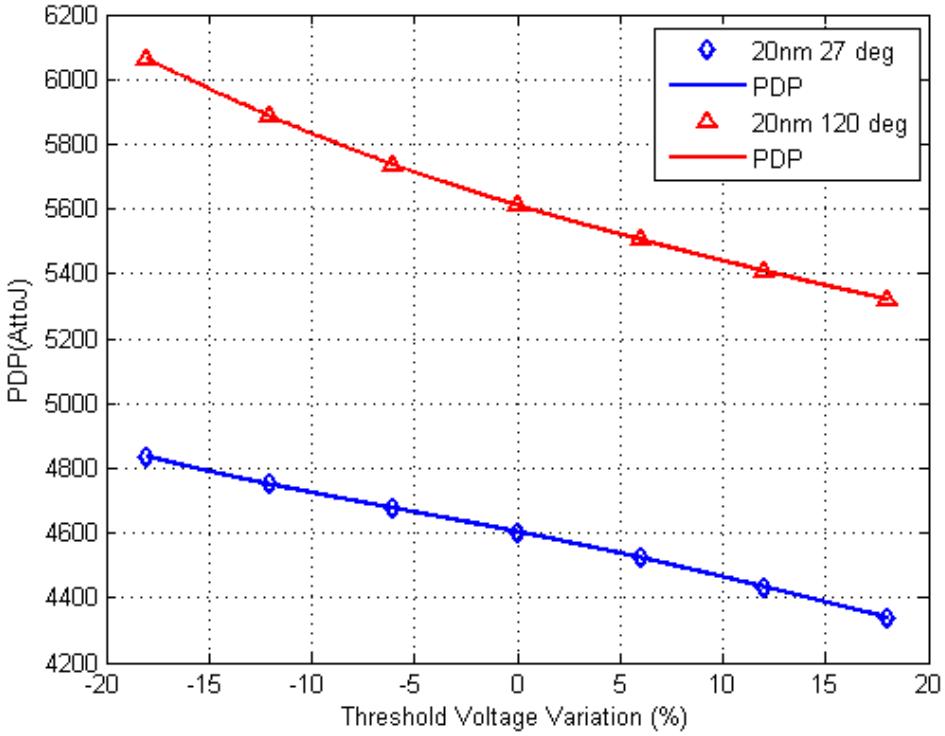


Figure 3.36: 20nm PDP vs. threshold voltage at 27° and at 120°

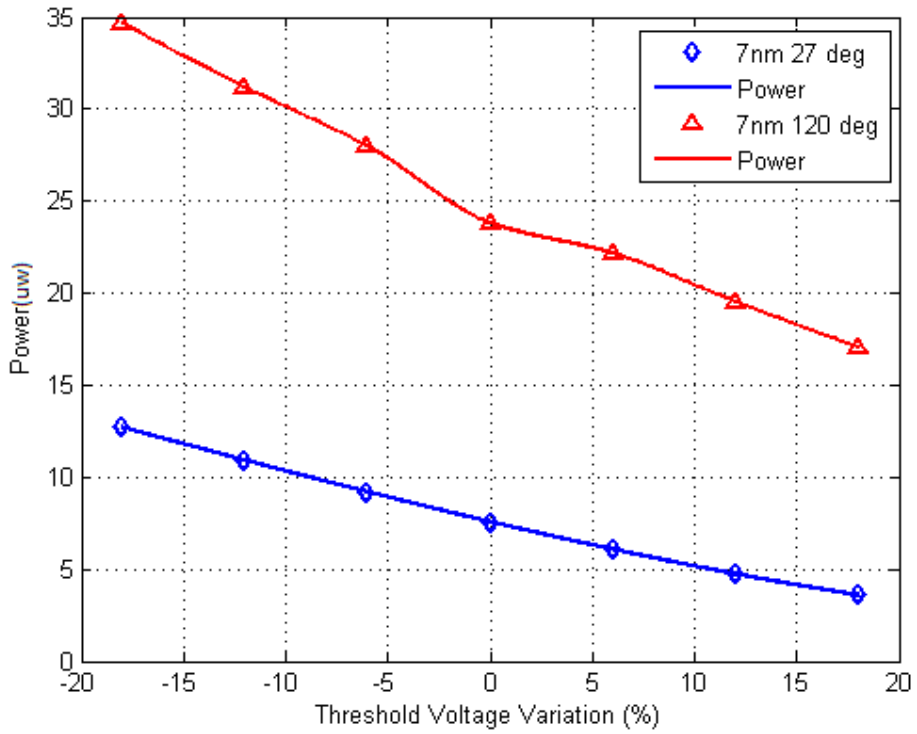


Figure 3.37: 7nm power vs. threshold voltage at 27° and at 120°

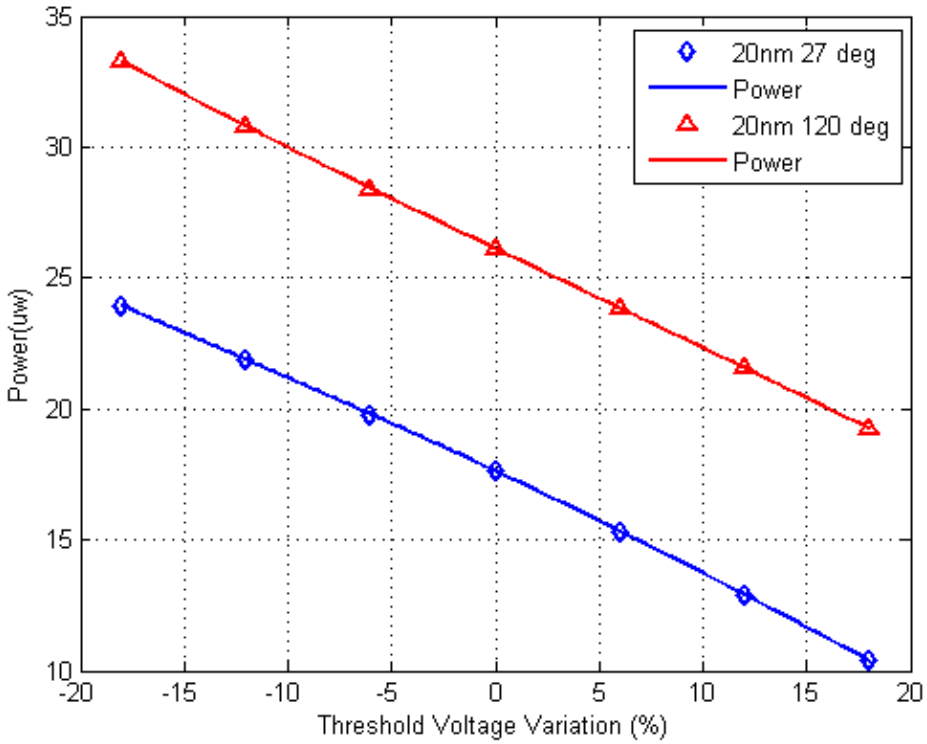


Figure 3.38: 20nm power vs. threshold voltage at 27° and at 120°

We can conclude that performance of FinFET based ring oscillator is evaluated using predictive technology models for low standby power with technology scaling. Threshold voltage and temperature variations impact on performance metrics is illustrated. The results show enhancement of the performance with technology scaling, however beyond the 14nm node it degrades as a result of scaling other device parameters besides channel length.

The study also illustrated that, power consumption is reduced with technology scaling, also it showed an improvement in trends of the power delay product, however the sensitivity of the power and frequency to threshold and temperature variations is increased with technology scaling. Threshold voltage increase has a positive impact on PDP while temperature increase has a negative one.

3.3.2.2. FinFET flip-flops

Latches and Flip-flops are typically used as elements for data storage, they are mandatory blocks for sequential logic circuits, and digital circuits [67]. Flip-flop is an essential part of programmable logic devices (PLD), field programmable gate array (FPGA), and system on chip (SoC). Flip-flops also can be used for synchronization purposes.

Some studies have analyzed PTM circuits with technology scaling [53] [64] [68-69]. For instance, a simulation study for PTM ring oscillator and basic logic gates is discussed [64]. Other studies have discussed analysis of process variations impact on Flip-flops. For instance, analysis of process variation impact on CMOS Flip-flops soft error rate is discussed [70]. In this work, we report supply voltage impact on four FinFET based Flip-flops topologies performance and power at different technology nodes starting from 20 nm down to 7 nm, and we obtain the optimum supply voltage from energy prespective at each technology node as the optimum supply voltage is used by the industry to optimize logic and memory circuitry designs.

Transmission gate Flip-flop in Figure 3.39 is simulated using device parameters listed in Table 3.3. Power delay product of transmission gate Flip-flop versus supply voltage at each technology node is drawn in Figure 3.40.

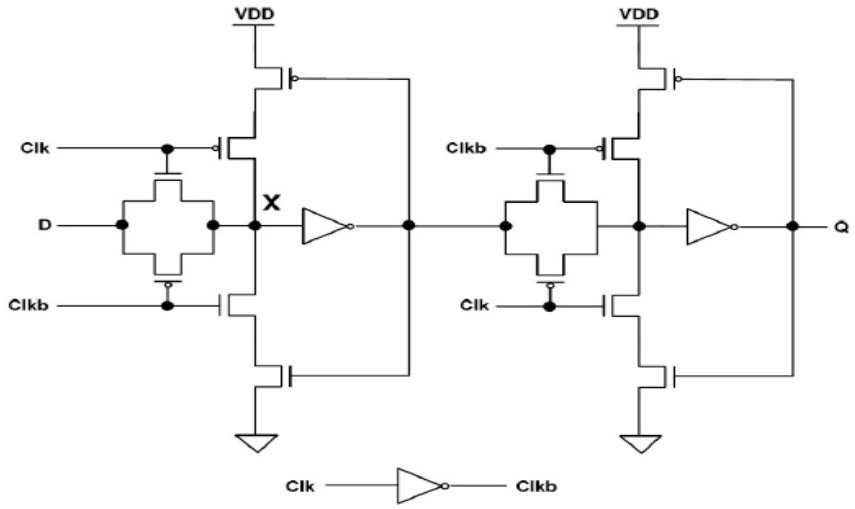


Figure 3.39: Transmission gate Flip-Flop

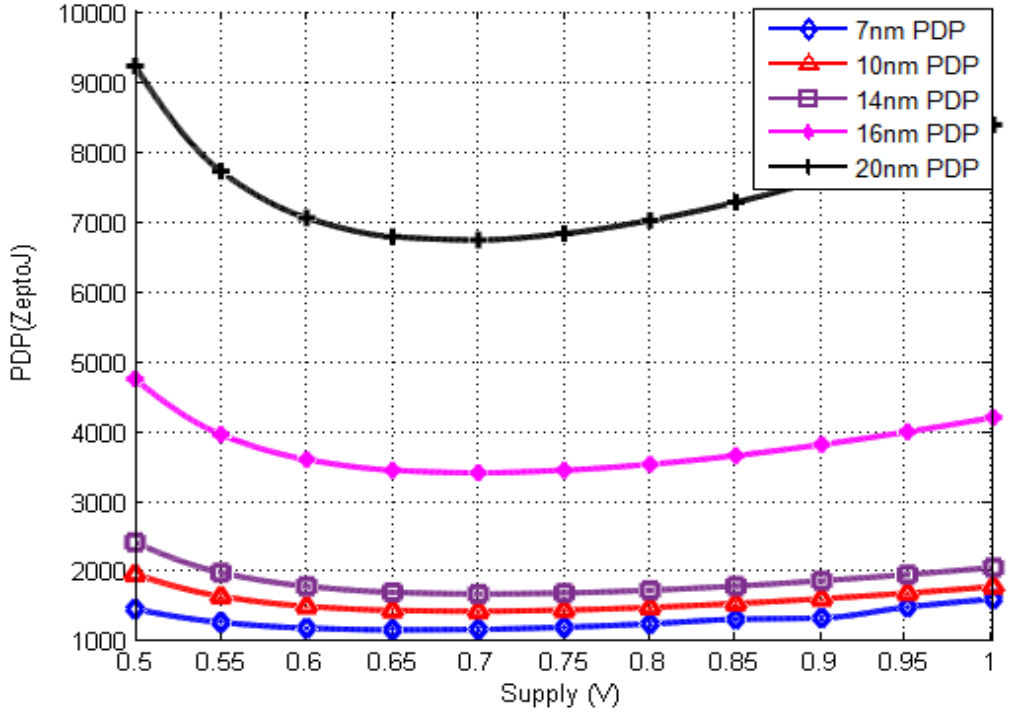


Figure 3.40: TG-FF power delay product vs. supply voltage for 20nm to 7nm nodes [19]

From Figure 3.40, the transmission gate Flip-flop optimum (minimum) power delay product (PDP) value at different technology nodes from 20nm down to 10nm occurs at 0.7V supply voltage. However, for the 7nm technology node at 0.65V supply voltage optimum power delay product is achieved. Figure 3.40 also shows that PDP trends of TG FF are improved with technology scaling.

Sense Amplifier Flip-flop shown in Figure 3.41 is simulated using device parameters in Table 3.3 (with N_{fin} = 2 for pmos). Power delay product of SA Flip-flop versus supply voltage at each technology node is illustrated in Figure 3.42.

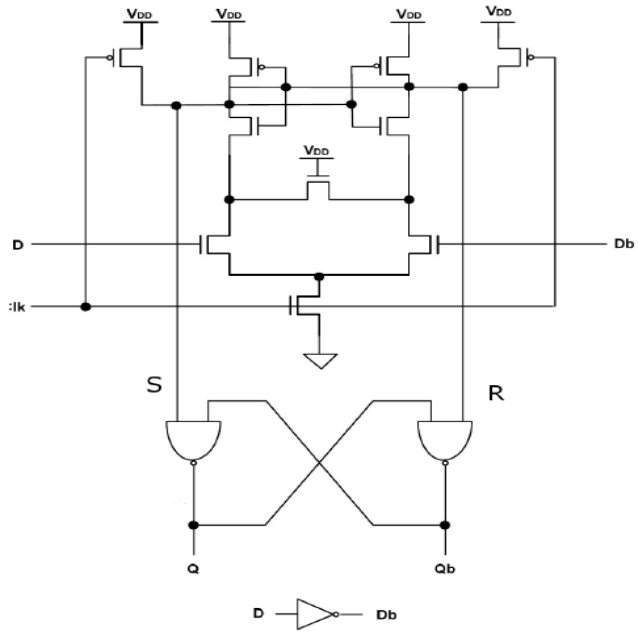


Figure 3.41: Sense Amplifier Flip-Flop

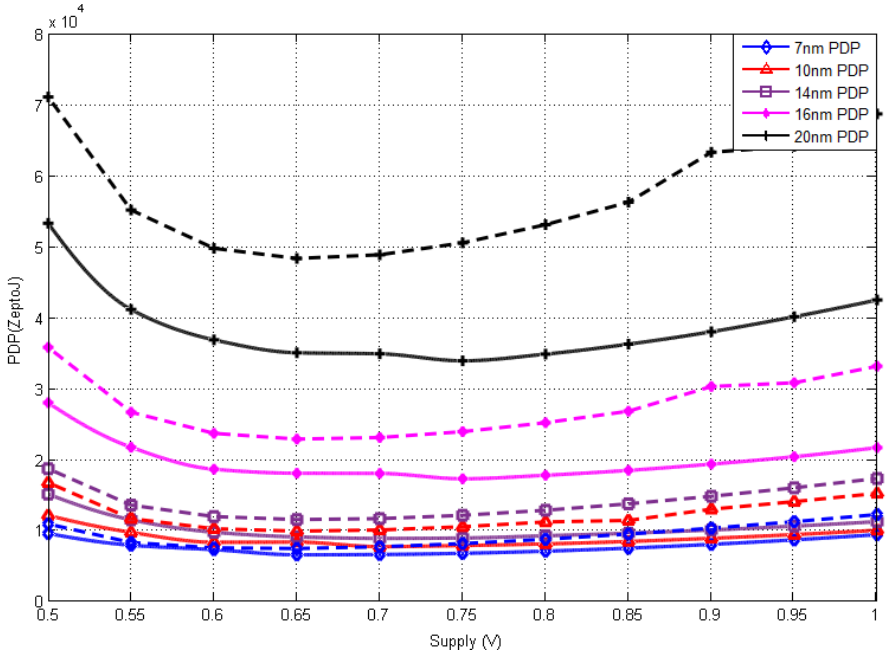


Figure 3.42: SA-FF and SD-FF power delay product vs. supply voltage fror 20nm to 7nm nodes (the solid lines are for SA-FF, and the dotted lines are for SD) [19]

From Figure 3.42, sense amplifier Flip-flop optimum power delay product for technology nodes from 20nm to 16nm occurs at 0.75V supply voltage, however for technology nodes 14nm and 10nm it occurs at 0.7V supply, and for 7nm technology node it occurs at 0.65V supply. From the Figure we also can obtain that PDP trends of SA FF are enhanced with technology scaling.

Semi Dynamic Flip-flop shown in Figure 3.43 is simulated using device parameters in Table 3.4. Power delay product of SD Flip-flop versus supply voltage at each technology node is illustrated in Figure 3.42.

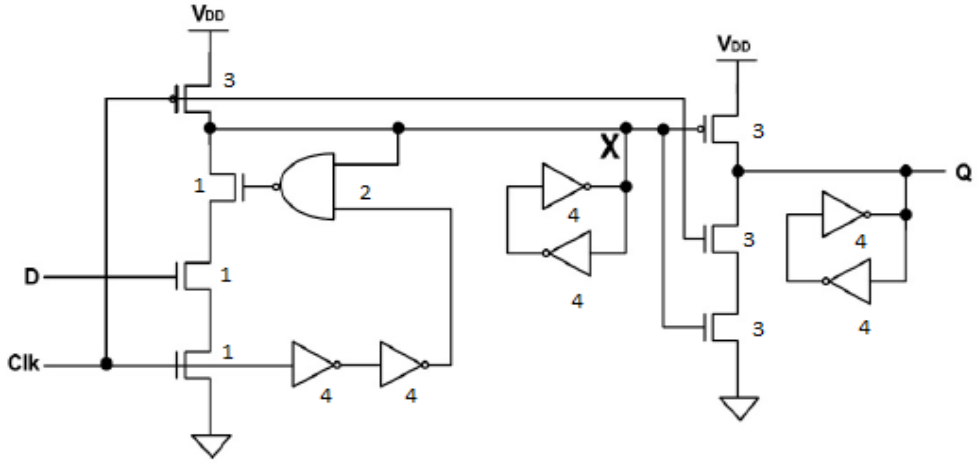


Figure 3.43: Semi Dynamic Flip-Flop

Table 3.4: The simulated device parameters of SD-FF

Device	**TG-FinFET**				
L(nm)	20	16	14	10	7
H_{fin}	28	26	23	21	18
N_{fin} (1)	25	22	22	18	15
N_{fin} (2)	7	6	6	5	4
N_{fin} (3)	5	4	4	3	3
N_{fin} (4)	1	1	1	1	1

Observing SD Flip-flop energy, semi dynamic Flip-flop optimum (minimum) supply voltage from power delay product (energy) perspective for technology nodes from 20nm and 7nm occurs at 0.65V. SD Flip-flop PDP trends also decrease with technology scaling.

We evaluated Tri-gate FinFET based Flip-flops performance according to many factors and metrics such as:

Critical charge (Qcritical) is defined as the minimum value of the collected charge (Qcollected) at storage node of flip-flop that can flip its logic, hence Qcritical be used to measure the vulnerability of flip-flop to soft errors. Qcritical can be modeled as a measurement metric of the SER for the different flip-flops topologies. The recombination of those collected charges results in a current pulse with very short duration which probably cause soft errors [70].

The critical charge is calculated at all nodes of each flip-flop for the 1-to-0 flip and the 0-to-1 flip at the output node. Then, the node that has the smallest critical charge is selected as the most susceptible node to soft errors (Node X in Figure 3.39, and Figure 3.43, and "S" in Figure 3.41). Soft Errors Rate (SER) of the different four Flip-flops types at the nominal supply voltage of each technology node is reported in Table 3.5.

Table 3.5: Soft errors rate (In coulombs)

Tech. node	20nm	16nm	14nm	10nm	7nm
Nominal supply	0.9V	0.85V	0.8V	0.75V	0.7V
TG	1.25f	1.25f	1.25f	1f	0.75f
SA	0.25f	0.25f	0.25f	0.25f	0.25f
SD	2.5f	2f	2f	1.75f	1.25f

From Table 3.5, the sense amplifier Flip-flop is the most vulnerable type to soft errors, while semi dynamic (SD) Flip-flop is the least vulnerable one to soft errors. Also technology scaling impact on SER of FinFET exhibits a similar trend to CMOS technology (SER is decreased with technology scaling in both of CMOS and FinFET).

Delay (Clk-Q) is key metric for evaluating the performance of Flip-flops. Our study shows that Flip-flops performance is enhanced with increasing the supply voltage, for 7nm TG Flip-flop increasing the supply voltage from 0.5V to 1V the performance is enhanced by 3.14 of its value at 0.5V.

The power is decreasing continuously with scaling down the technology as a result of shrinking the channel length and the scaling of the supply voltage.

The study also shows that Flip-flops power dissipation is increased with increasing the supply voltage, for 7nm TG Flip-flop the supply voltage from 0.5V to 1V the power dissipation at 1V supply voltage is 4.55 times its value at 0.5V.

By observing PDP trends with technology scaling, we obtained the optimum supply for each Flip-flop topology with technology scaling from 20nm to 7nm. The study also illustrates that PDP of each Flip-flop topology is improved with technology scaling.

In this section we recommend each Flip-flop for a specific application according to the obtained simulation results (Not listed) such as:

Semi dynamic (SD) Flip-flop is the fastest one of the four types, also it has negative setup time, so it is very good choice for high performance systems (within available power budget), however it is the most power consuming and has hold time.

Transmission gate (TG) Flip-flop is the least power consuming compared to the other Flip-flops, it has positive setup time and small clock to output delay and it has the minimum number of transistors, however its Clock load is high.

Sense Amplifier (SA) Flip-flop has a very useful feature of monotonous transitions at its outputs, which drives fast domino logic; however its rise and fall times degrade speed, and also cause glitches in successive logic stages, which increases total power consumption. SA-FF is considered as the most vulnerable topology to soft errors because of its small flipping time [70]. While SD-FF is considered as the least vulnerable one.

The performance of FinFET Flip-flops is evaluated with technology scaling. Supply voltage effects on performance metrics are illustrated. The results show that, TG, SA, and SD Flip-flops have better performance and power with technology scaling, also the optimum supply voltage from energy saving perspective for each technology node is reported. The study also shows SER values for each Flip-flop topology with technology scaling.

Some design insights and Flip-flop choice recommendation are obtained and reported. For instance, SD FF is the optimum topology for high performance designs. From power consumption point of view, TG-FF is the optimum once since it is the least power consuming Flip-flop among the three discussed topologies.

Threshold voltage increase has a positive impact on trends of TG-FF PDP, while temperature increase has a negative impact. For instance, the PDP at +18% increase of its nominal threshold voltage value at 7nm is less than the nominal value PDP by 0.18.

Energy of Flip-flop is improved with temperature increase. For instance, at 16nm, PDP at 120° is lower than PDP at room temperature value by 0.065.

We analyzed the impact of temperature and threshold voltage variations on TG-FF as shown in Figures 3.44 to 3.49. Trends of PDP are improved with threshold voltage increase and degraded with temperature increase, for instance, the PDP at +18% increase of its nominal threshold voltage value at 7nm is lower than PDP at the nominal value by 0.18. Energy of Flip-flop is improved with temperature increase. For instance, at 16nm, PDP at 120° is lower than PDP at room temperature value by 0.065.

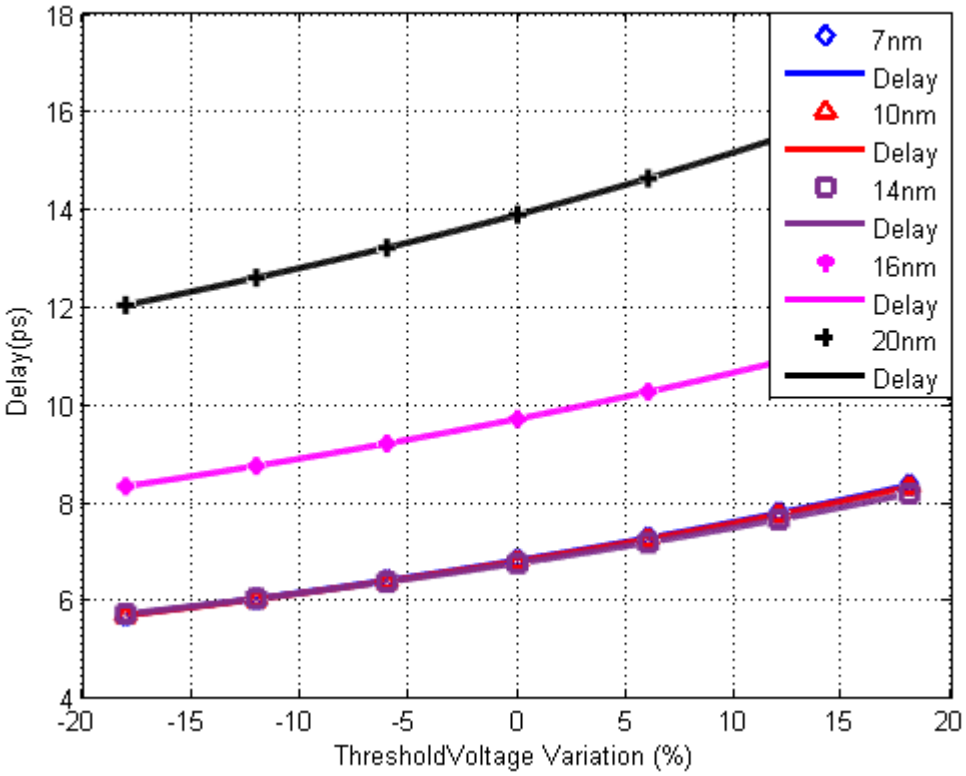


Figure 3.44: TG-FF Delay vs. Threshold voltage

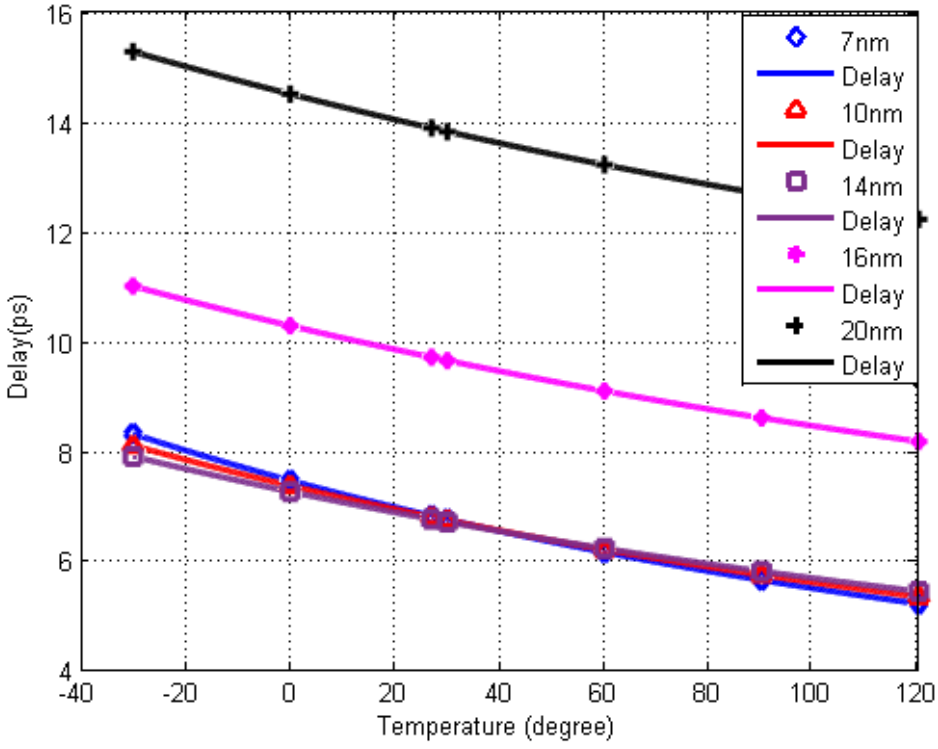


Figure 3.45: TG-FF Delay vs. Temperature

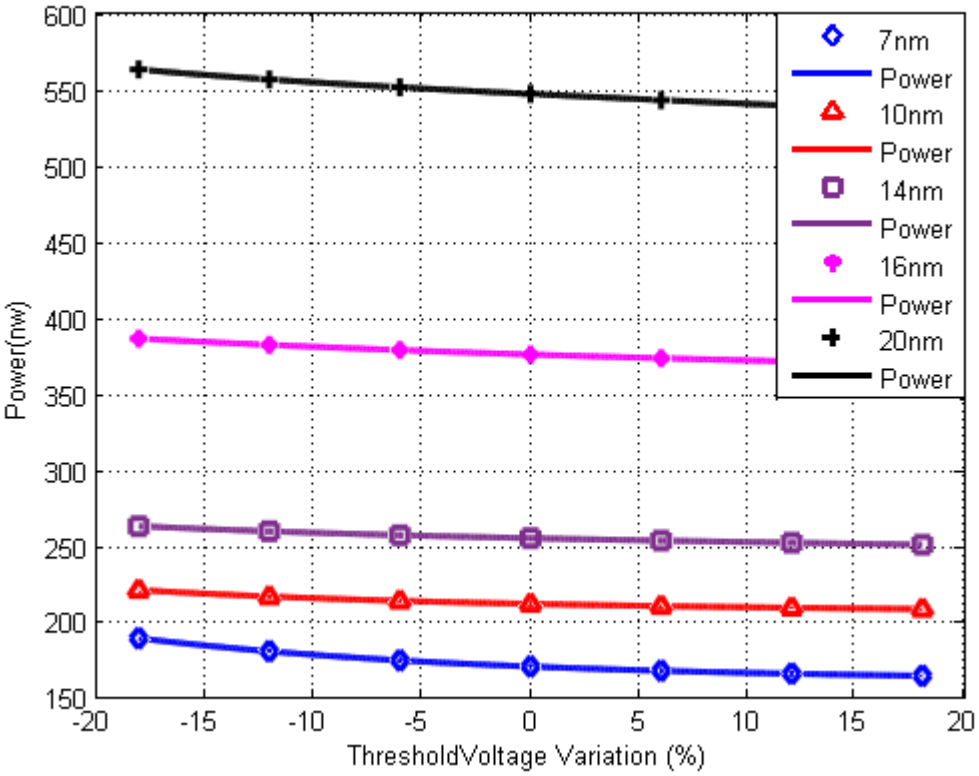


Figure 3.46: TG-FF Power vs. Threshold voltage

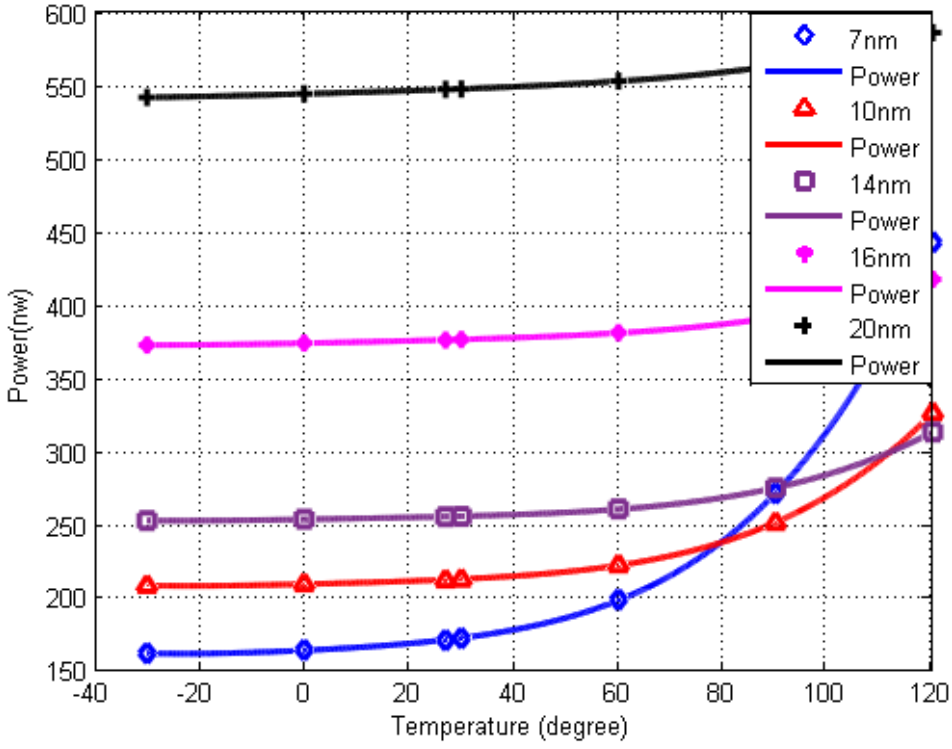


Figure 3.47: TG-FF Power vs. Threshold voltage

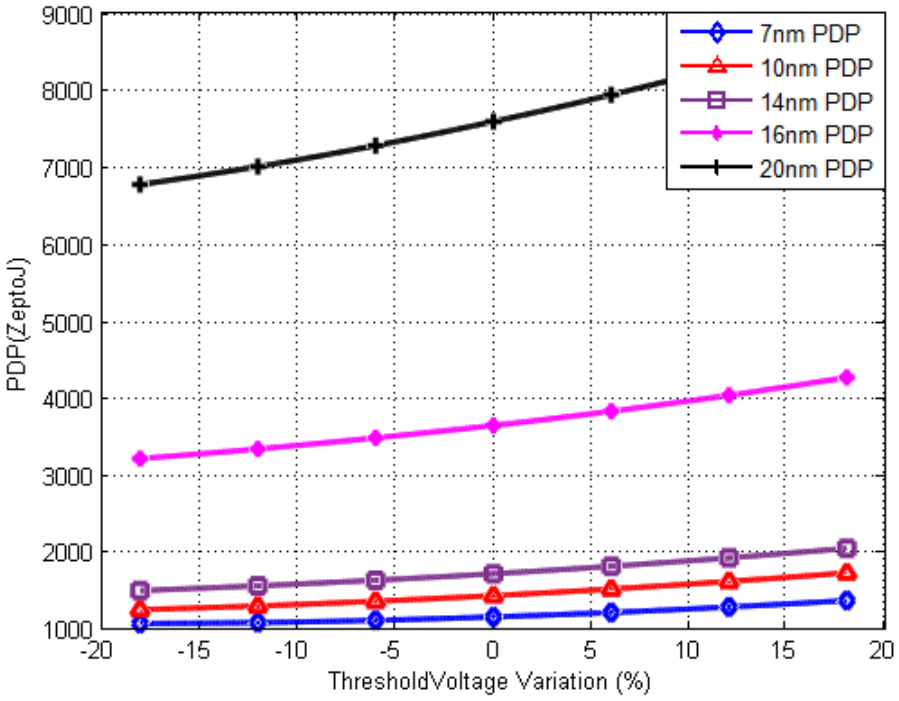


Figure 3.48: TG-FF PDP vs. Threshold voltage

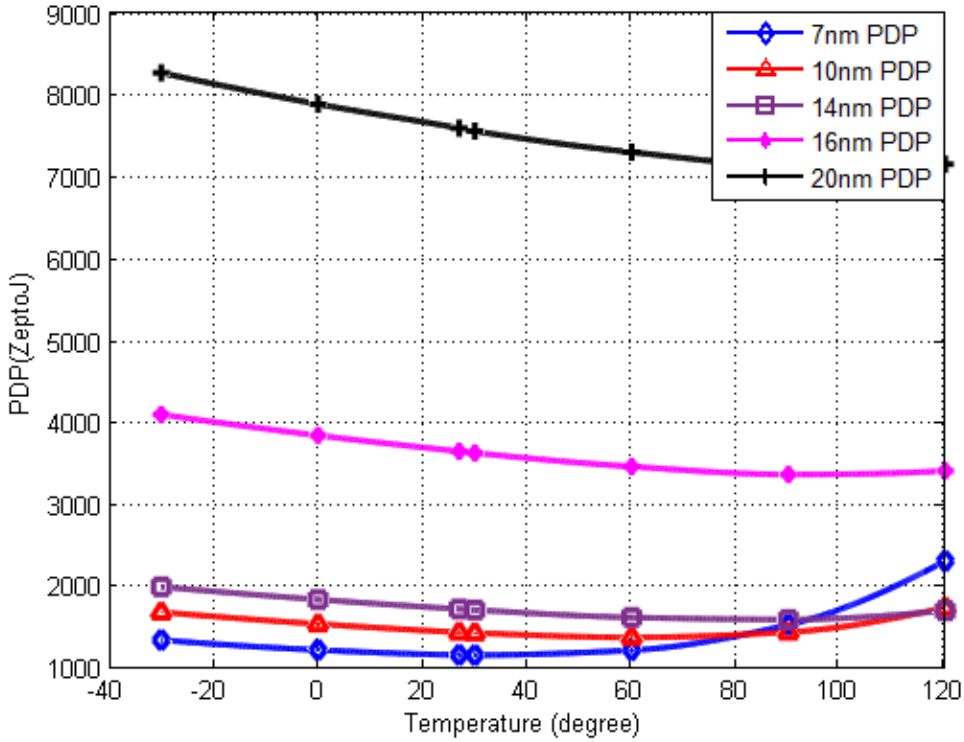


Figure 3.49: TG-FF PDP vs. Temperature

Clocked CMOS Flip-Flop shown in Figure 3.50 is simulated using device parameters in Table 3.3. Figures 3.51 to 3.56 show the delay, power, and PDP of Clocked CMOS Flip-Flop. This Flip-Flop is insensitive to overlap since the rise and fall times of the clock edges (clock slew) are sufficiently small.

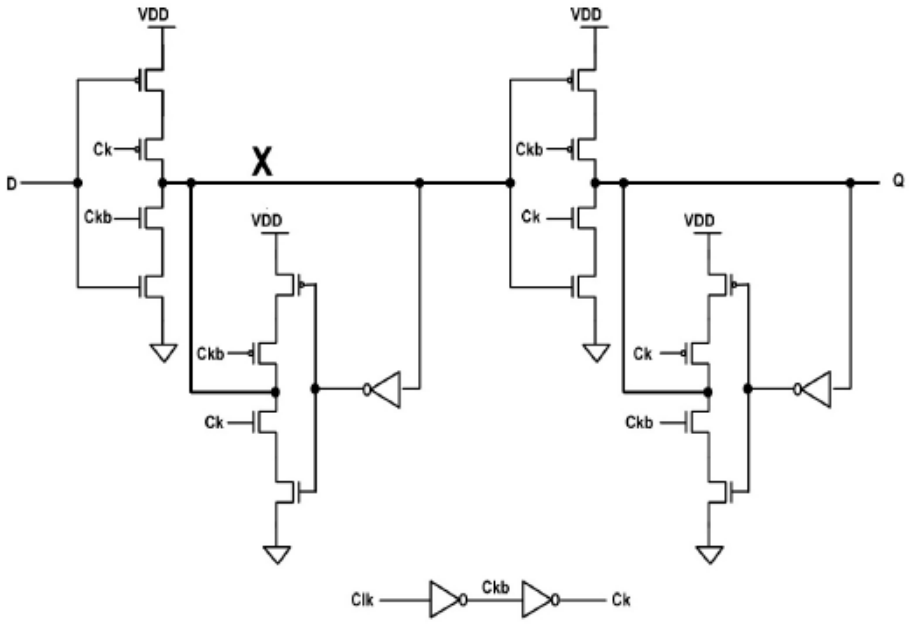


Figure 3.50: Clocked CMOS flip-flop

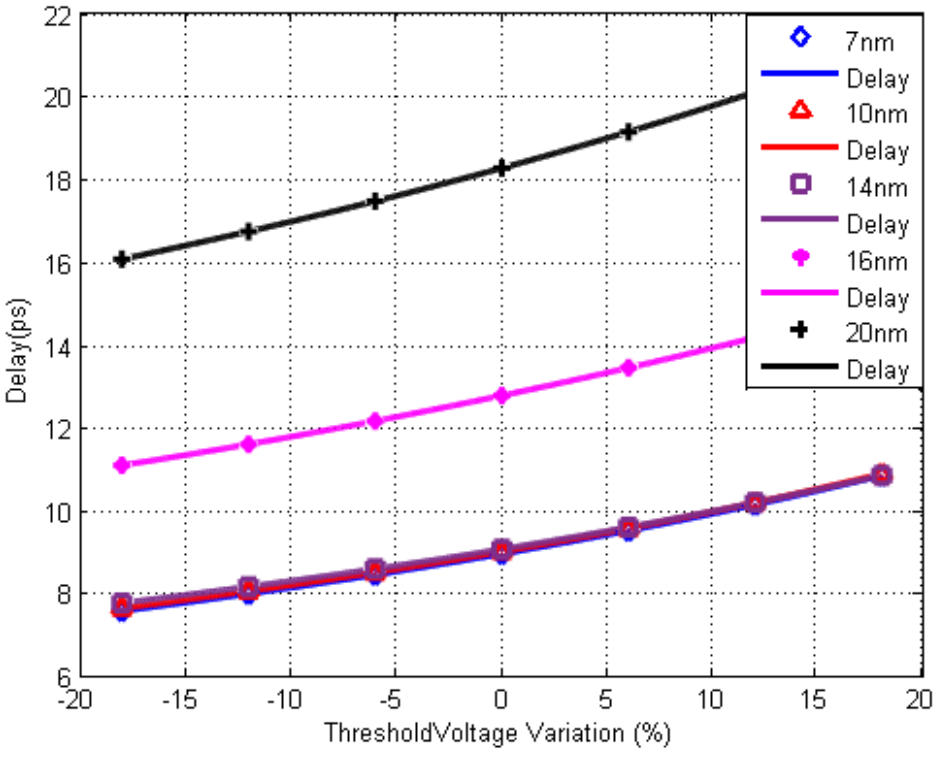


Figure 3.51: C2MOS-FF Delay vs. Threshold voltage

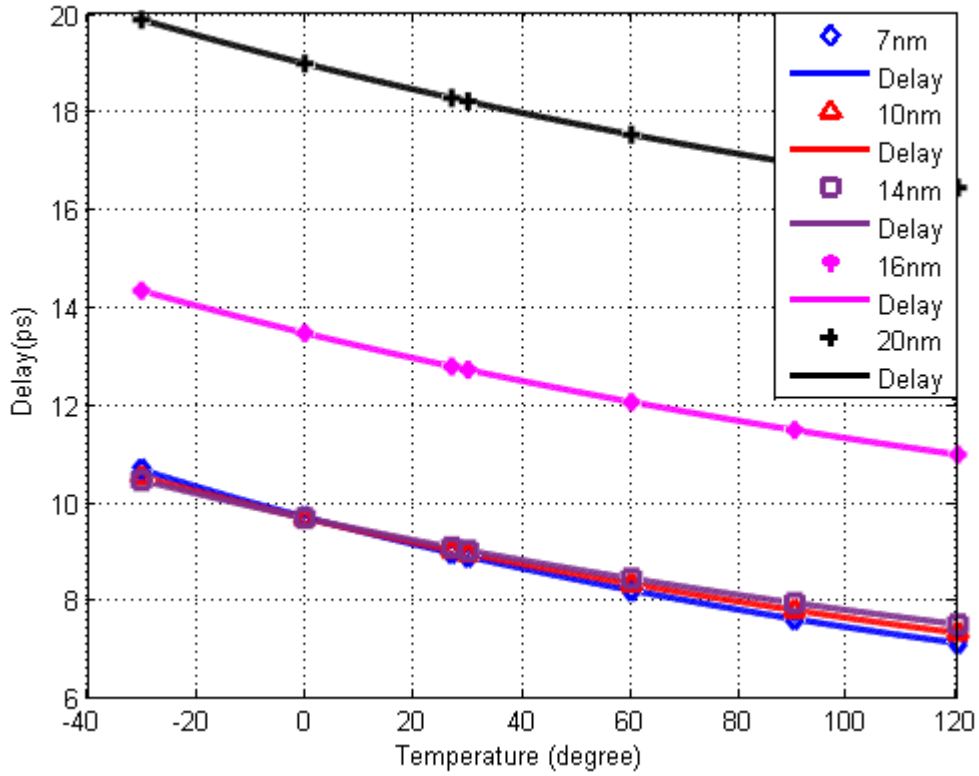


Figure 3.52: C2MOS-FF Delay vs. Temperature

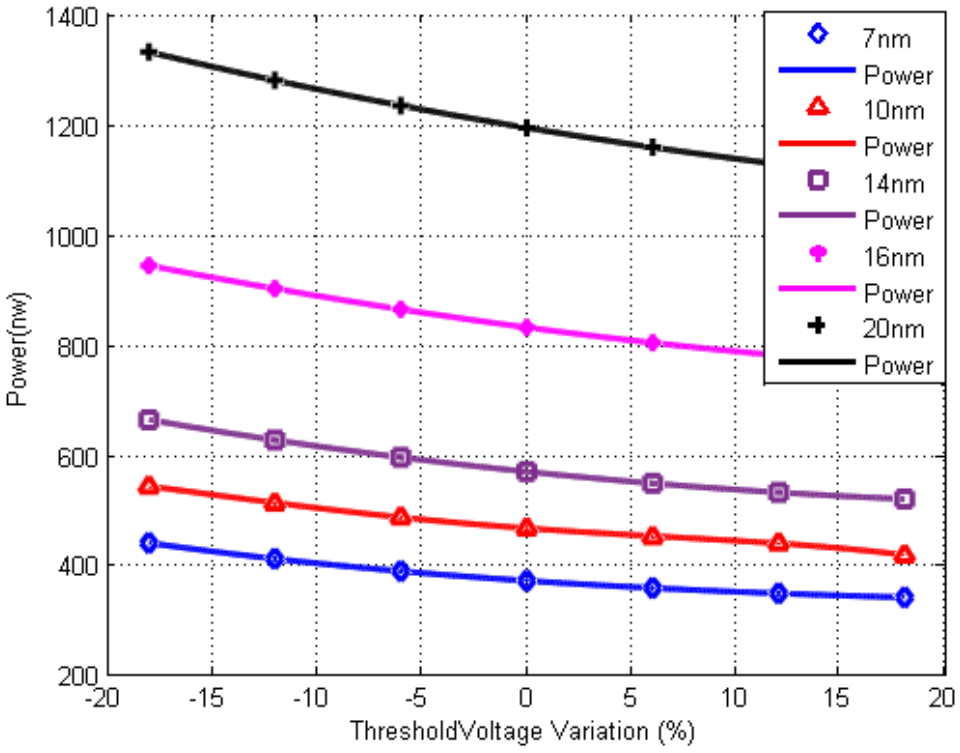


Figure 3.53: C2MOS-FF Power vs. Threshold voltage

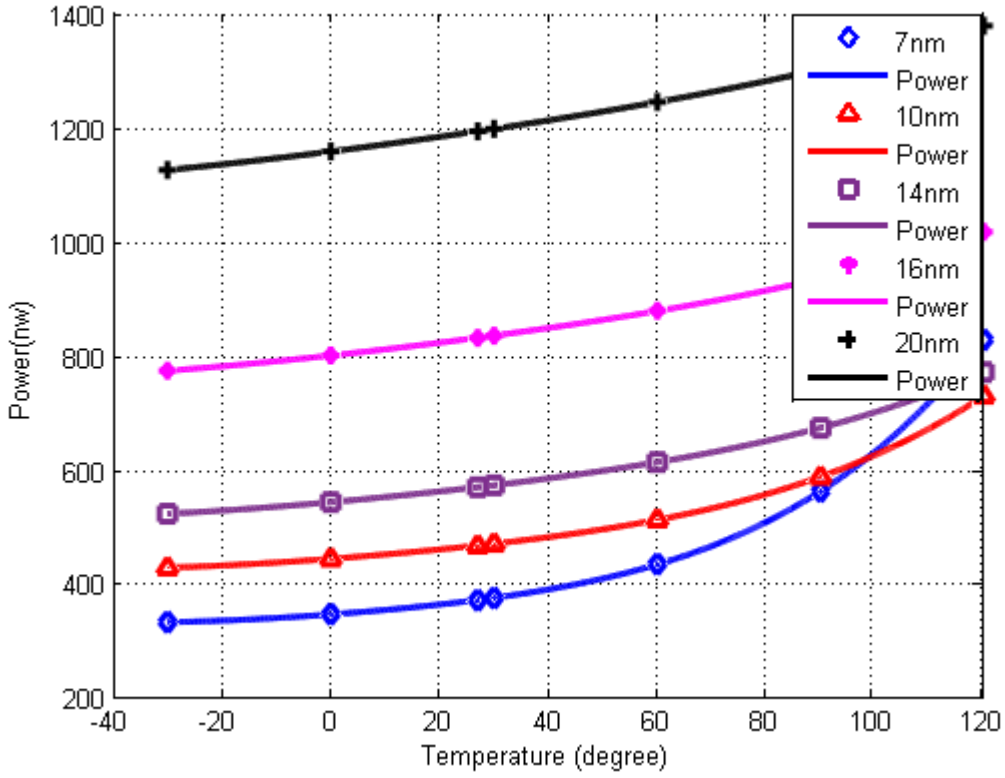


Figure 3.54: C2MOS-FF Power vs. Temperature

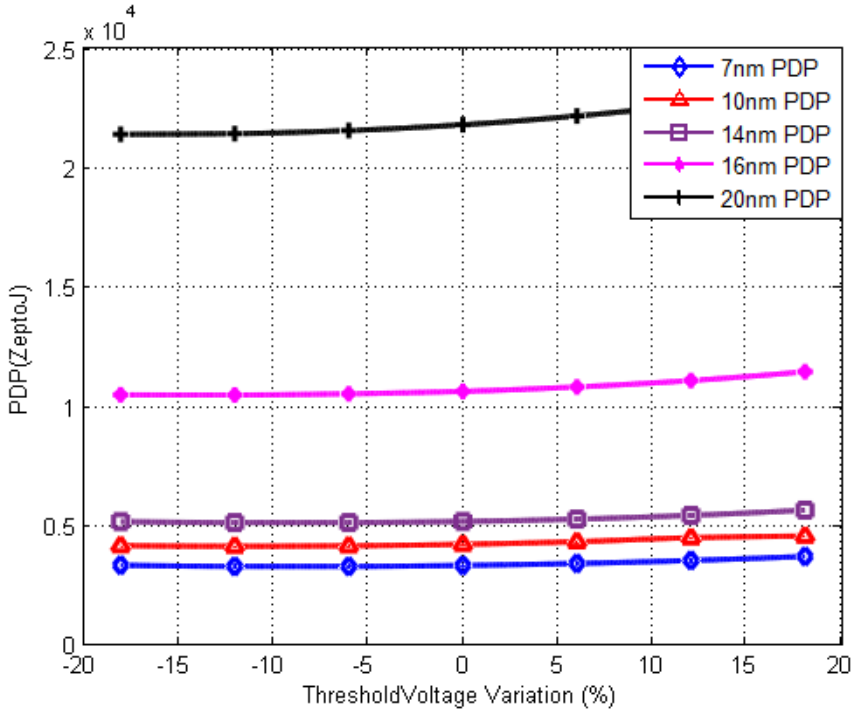


Figure 3.55: C2MOS-FF PDP vs. Threshold voltage

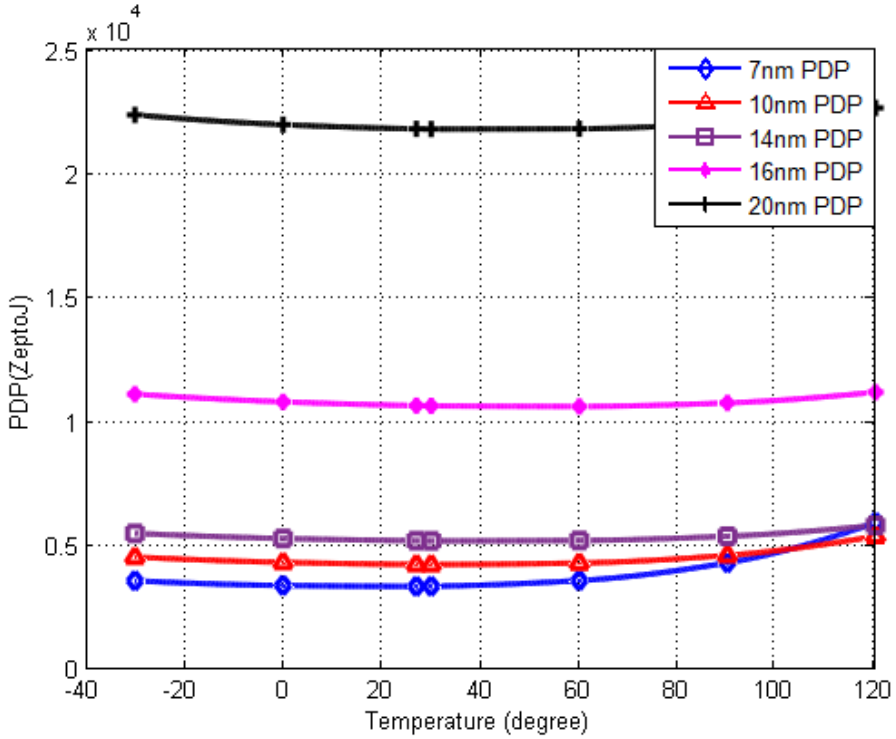


Figure 3.56: C2MOS-FF PDP vs. Temperature

Sense Amplifier Flip-Flop consists of sense amplifier in the first stage and slave set-reset (SR) latch in the second stage. Sense Amplifier Flip-Flop shown in Figure 3.41 is simulated using device parameters in Table 3.3 (N_{fin} = 2 for pmos). Figures 3.57 to 3.62 show the delay, power, and PDP of Sense Amplifier Flip-Flop.

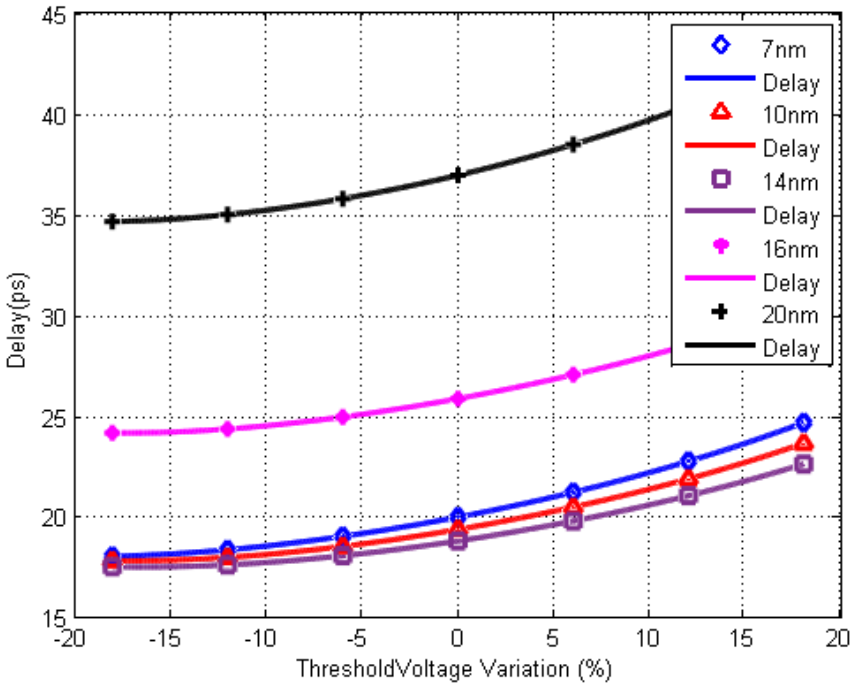


Figure 3.57: SA-FF Delay vs. Threshold voltage

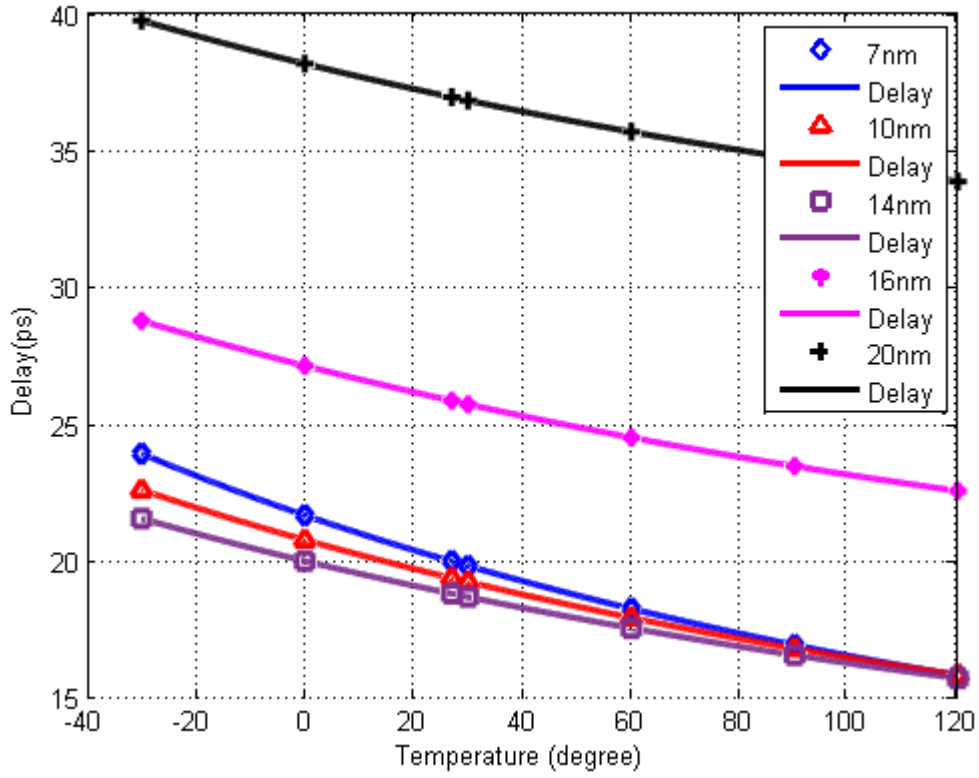


Figure 3.58: SA-FF Delay vs. Temperature

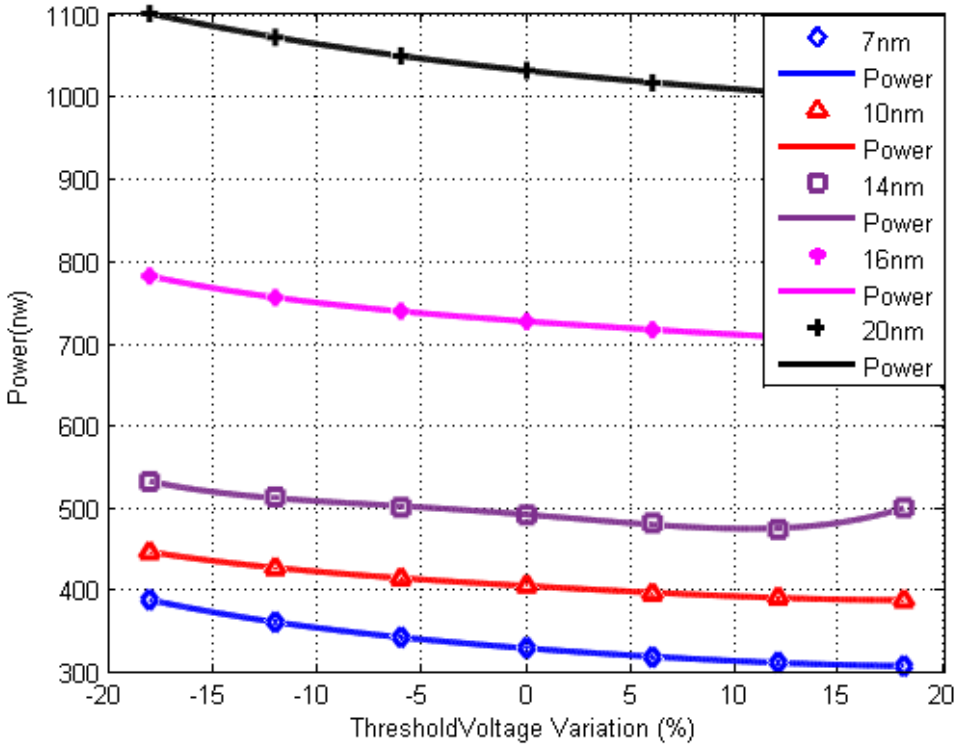


Figure 3.59: SA-FF Power vs. Threshold voltage

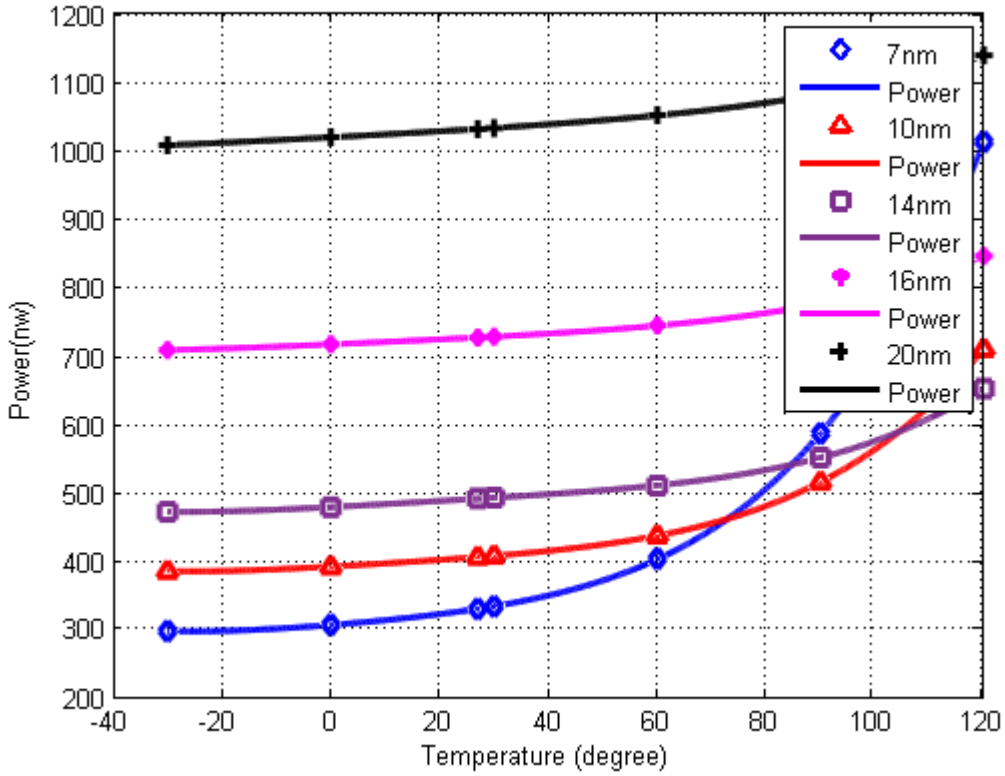


Figure 3.60: SA-FF Power vs. Temperature

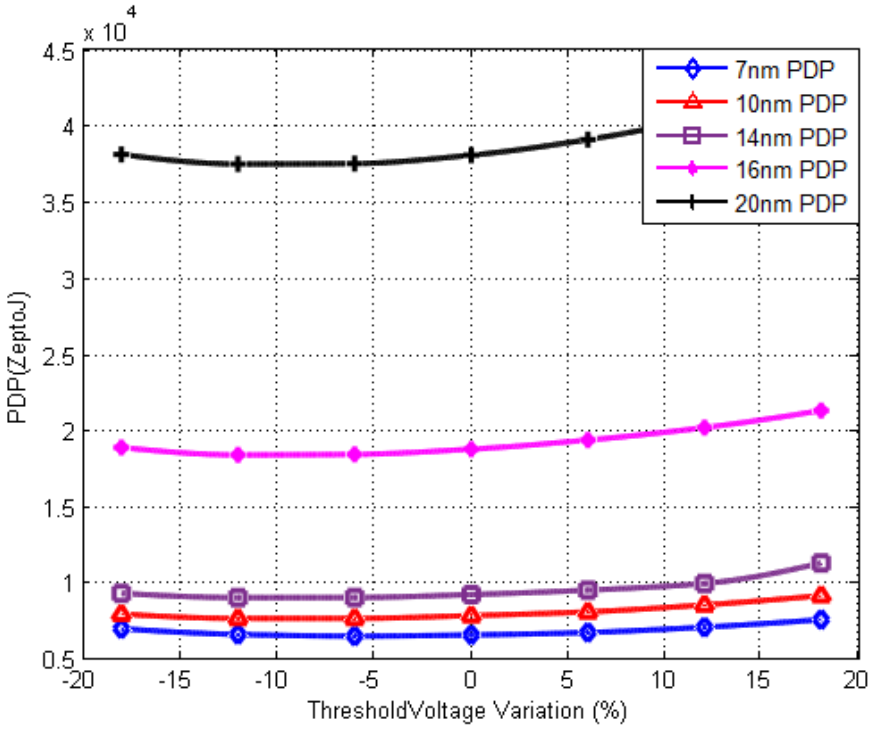


Figure 3.61: SA-FF PDP vs. Threshold voltage

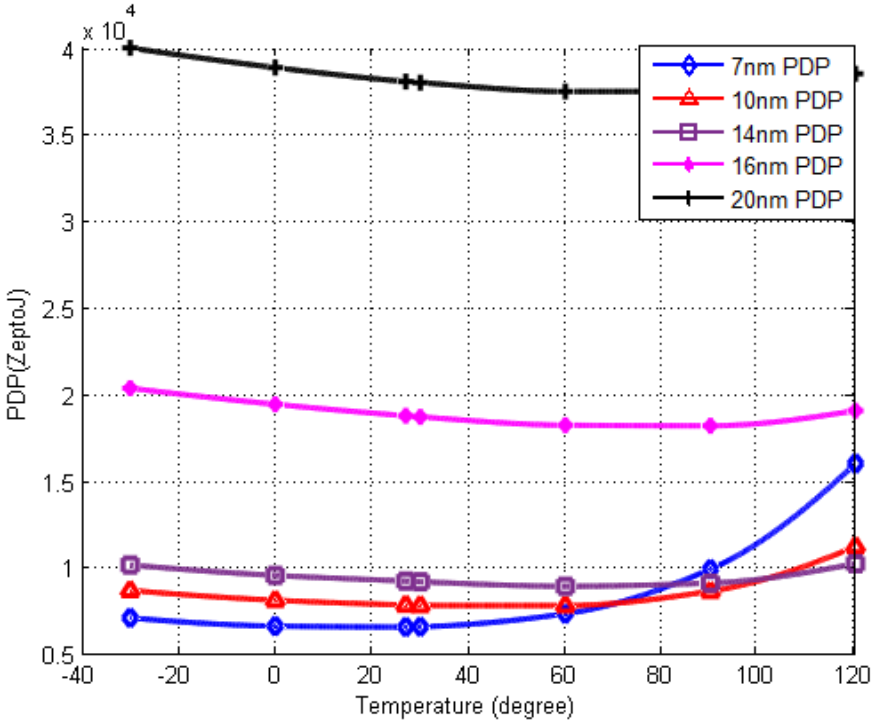


Figure 3.62: SA-FF PDP vs. Temperature

Semi Dynamic Flip-Flop shown in Figure 3.43 is simulated using device parameters in Table 3.4. Figures 3.63 to 3.68 show the delay, power, and PDP of Semi Dynamic Flip-Flop. The Flip-Flop consists of a dynamic front-end and a static backend, that is why it is semi dynamic circuit.

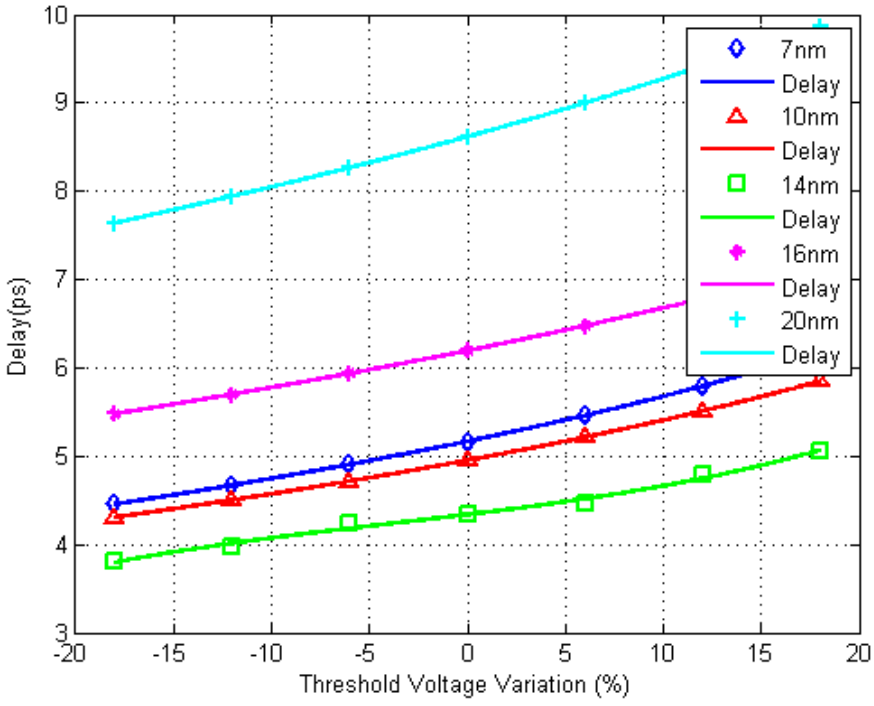


Figure 3.63: SD-FF Delay vs. Threshold voltage

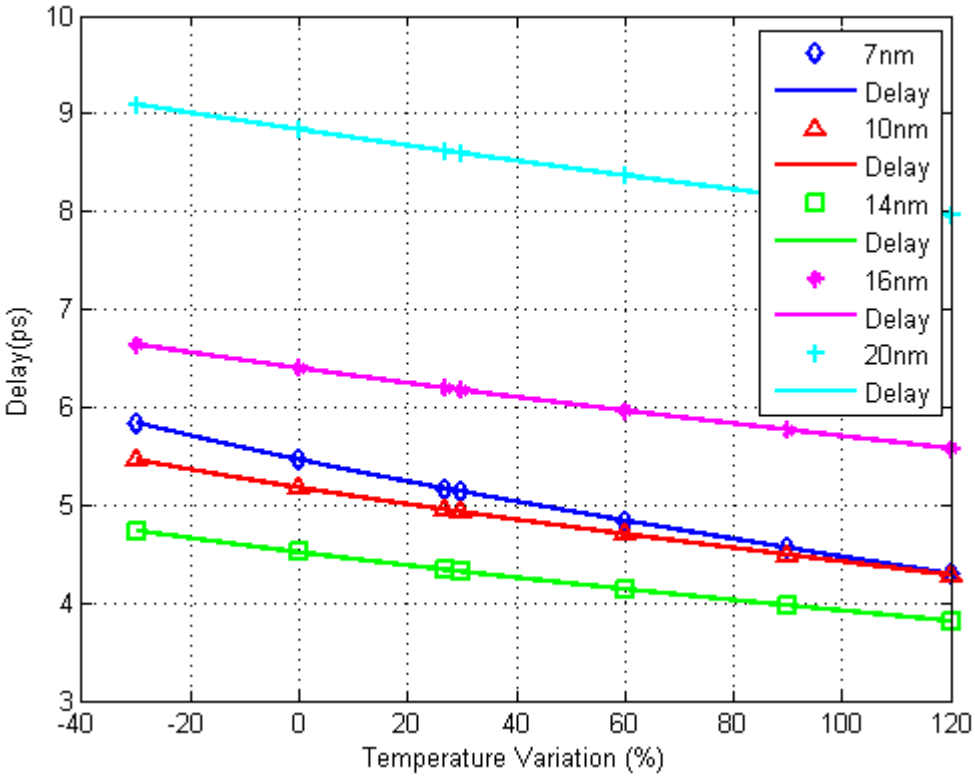


Figure 3.64: SD-FF Delay vs. Temperature

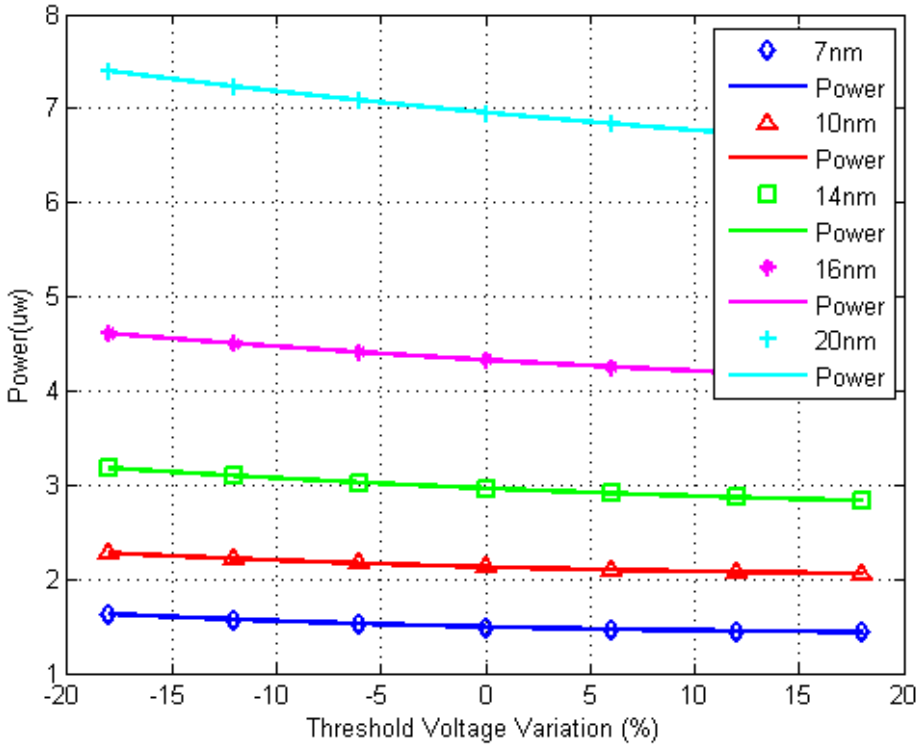


Figure 3.65: SD-FF Power vs. Threshold voltage

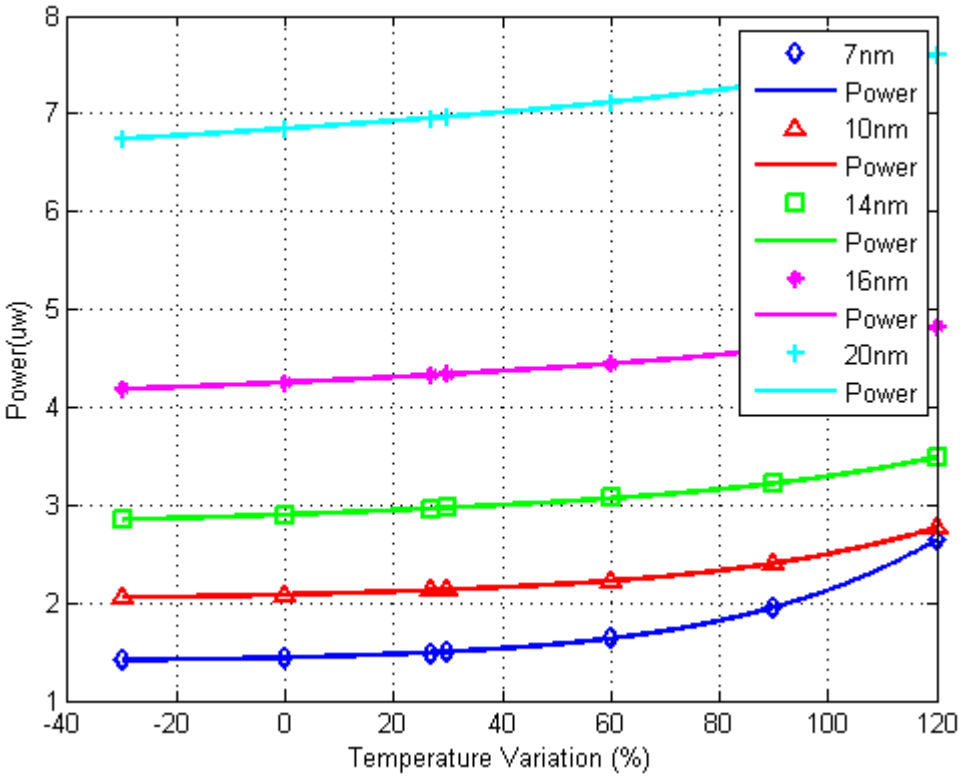


Figure 3.66: SD-FF Power vs. Temperature

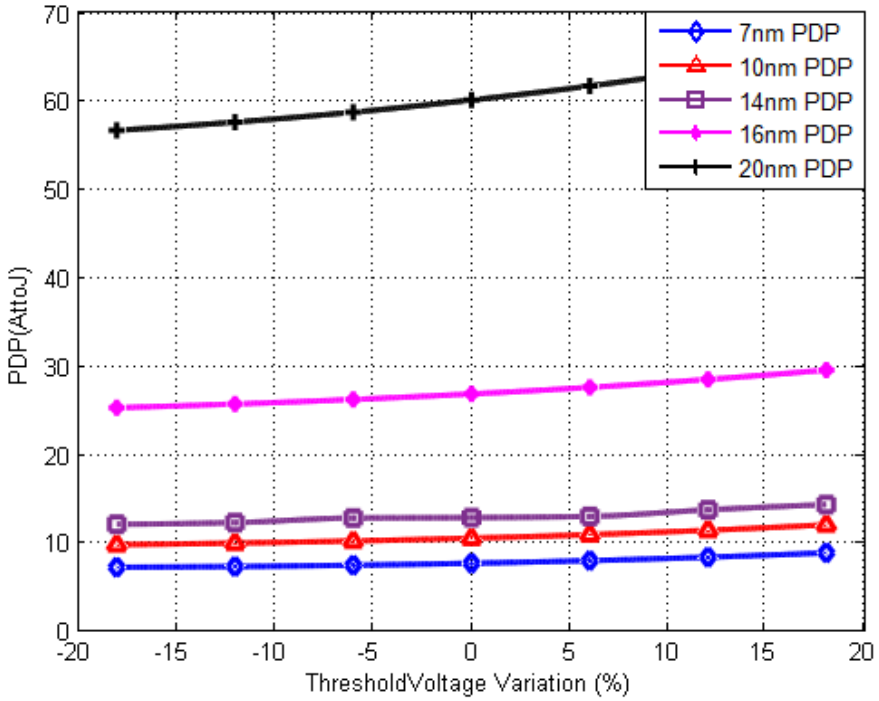


Figure 3.67: SD-FF PDP vs. Threshold voltage

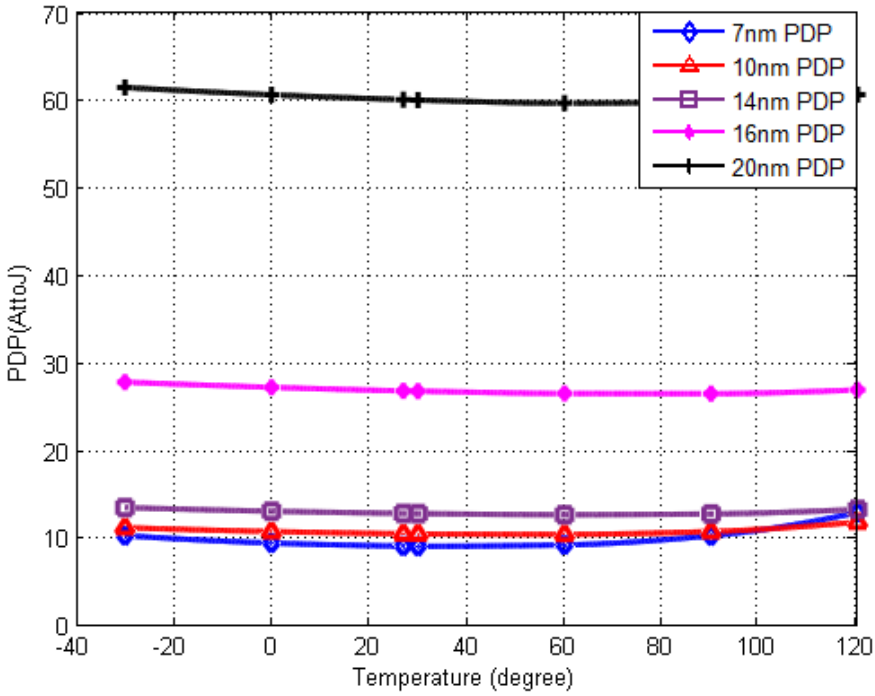


Figure 3.68: SD-FF PDP vs. Temperature

From this work, we figured out that Semi dynamic (SD) flip-flop is the fastest one of the four types. Also it has negative setup time, so it's very good choice for high performance systems (within available power budget), however it's the most power consuming and has hold time.

Compared to the other flip-flops, transmission gate (TG) flip-flop is the least power consuming type. It has positive setup time and small clock to output delay. It has also the minimum number of transistors compared to other three types, but it has high clock load though.

Clocked CMOS flip-flop has small clock load, achieved by the local clock buffering, also it's robust to clock slope variation due to the local clock buffering, however, it is slower than TG flip-flop.

The PDP sensitivity (variation) increases with technology scaling in flip-flops, this can be illustrated in Figure 3.49, Figure 3.56, Figure 3.62, and Figure 3.68 where 7nm technology node has a high rise in PDP value at high temperature values (power is the dominant factor of this increase).

3.3.3. FinFET FPGA cluster

Adder and NAND benchmarks

Two benchmarking circuits are simulated (2-bit adder, and 4-bit NAND) with technology scaling from 20nm technology node to 7nm.

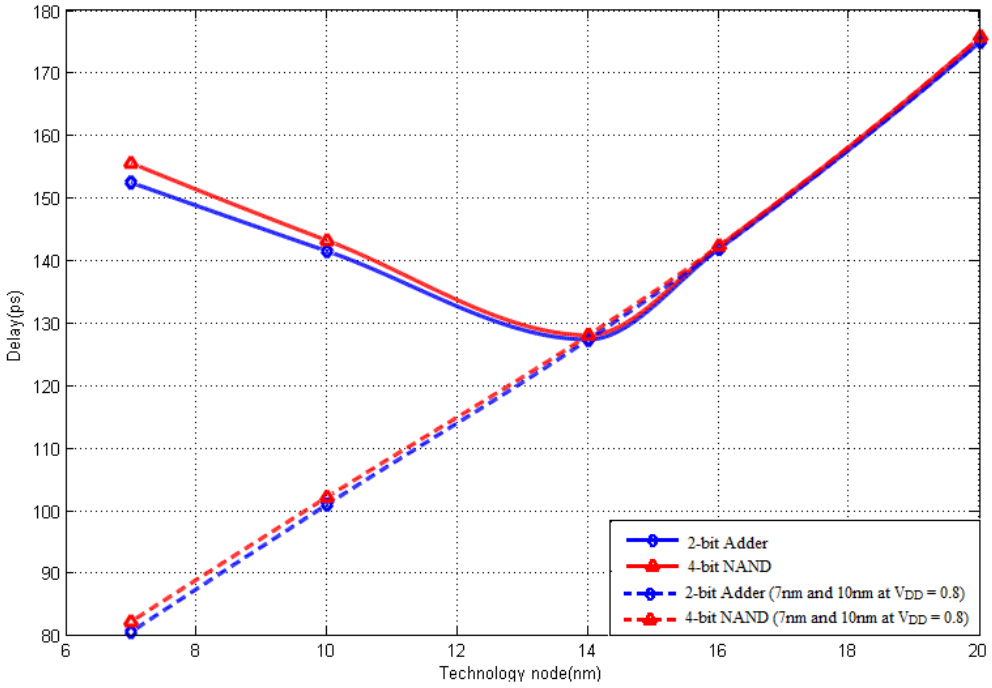


Figure 3.69: Delay of 2-bit adder, and 4-bit NAND circuits for 20nm to 7nm nodes, the dashed line are for 7nm and 10nm at V_{DD} = 0.8

Delay trend is enhanced with technology scaling. However, beyond 14nm technology node performance is degraded. However using a higher V_{DD} would lead to better performance, this will be at the cost of power reduction as presented in Figure 3.69. Device scaling options such as using high mobility channel [65] and/or using gate-all-around (GAA) nanowires [66] have the potential to enhance device scaling in this time frame.

Using supply voltage of 0.8V keeps on performance enhancement trend with technology scaling at 10nm and 7nm technologies. For instance, 7nm 2-bit adder delay at 0.8V supply is 80.645 ps while it is 152.35 ps at the nominal supply voltage at this technology node (V_{DD} = 0.7V).

Observing power consumption trends in Figure 3.70, 2-bit adder consumes more power than 4-bit NAND as its switching factor is greater than NAND one. Also, power trends indicate an improvement with technology scaling till 10nm. Since SRAM's in FPGA LUT is configured once at FPGA programming phase, leakage power is the dominant source of the average power dissipation. As leakage power increases with technology scaling, SRAM's leakage power affects the overall average power significantly at 7nm which leads to power dissipation increase at this technology node.

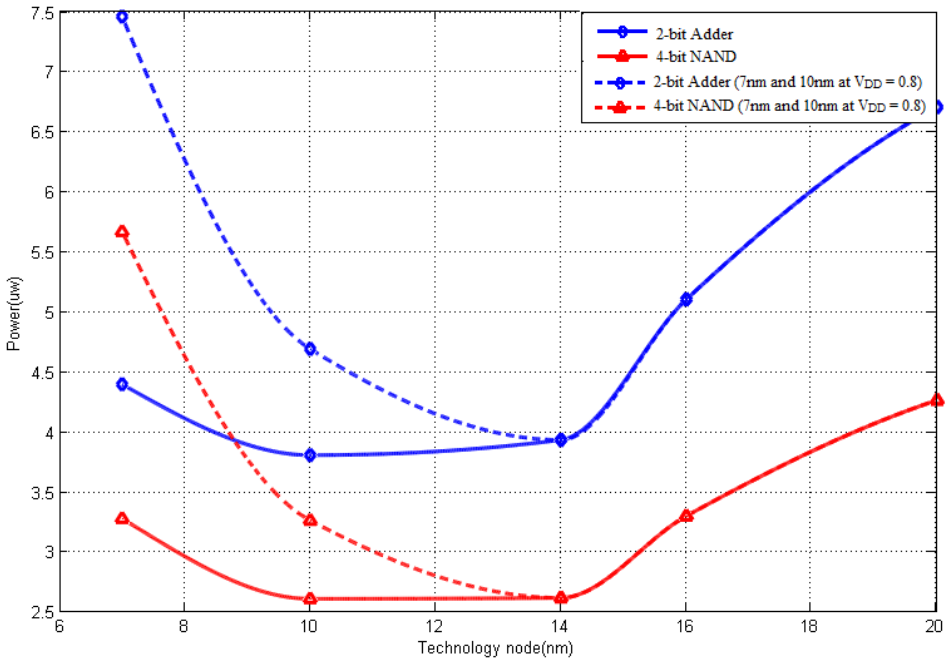


Figure 3.70: Power consumption of 2-bit adder, and 4-bit NAND circuits for 20nm to 7nm nodes, the dashed line are for 7nm and 10nm at V_{DD} = 0.8

The higher supply voltage keeps on performance improvement with technology scaling but this at cost of power reduction at 10nm and 7nm technologies as discussed earlier. For instance, 7nm 2-bit adder power at 0.8V supply is 7.4496 uw while it is 4.3932 uw at the nominal supply voltage at this technology node (V_{DD} = 0.7V).

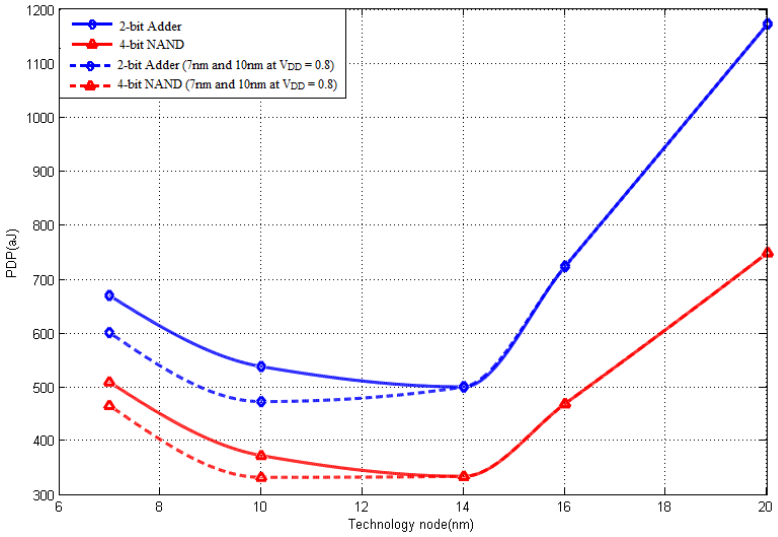


Figure 3.71: PDP of 2-bit adder, and 4-bit NAND circuits for 20nm to 7nm nodes, the dashed line are for 7nm and 10nm at V_{DD} = 0.8

PDP is a key metric in evaluating any digital circuit as it indicates the energy consumption and hence battery life for portable devices. PDP trends in Figure 3.71 also indicate improvement of energy consumption with technology scaling from 20nm down to 14nm.

While using higher supply voltage (V_{DD} = 0.8V in this case) increases power consumption at 10nm and 7nm technologies, the overall PDP is enhanced. For instance, 7nm 2-bit adder PDP at 0.8V supply is 600.773 aJ while it is 669.304 aJ at the nominal supply voltage at this technology node (V_{DD} = 0.7V) which is equivalent to 10.24% energy reduction.

Cascaded flip-flops chain benchmark

Cascaded flip-flops chain consists of three cascaded flip-flops path, it is formed by driving one of first BLE inputs and connecting its output to one of the inputs of the second BLE and second BLE output to one of the inputs of the third BLE. Simulations are done at 200 MHz frequency with phase difference 400ps from FPGA cluster inputs.

Delay, power consumption, and PDP trends with technology scaling of the benchmark circuit are presented below in Figures 3.72 to 3.74.

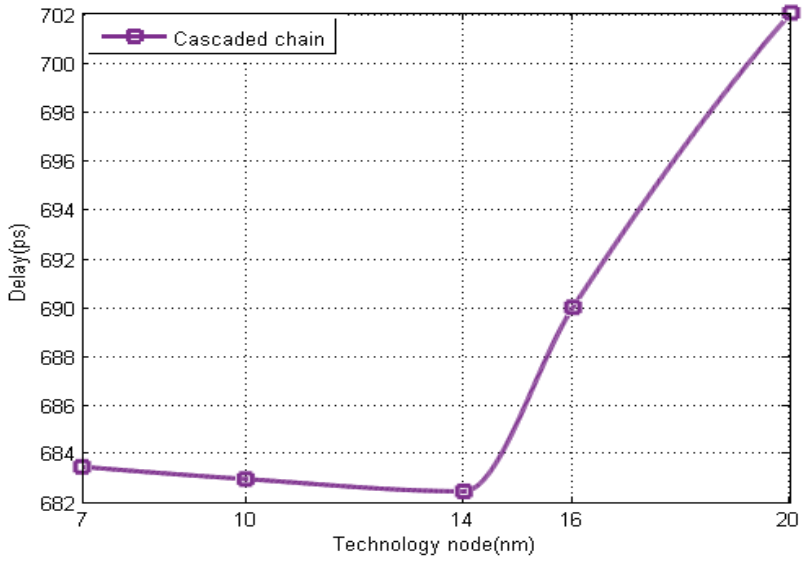


Figure 3.72: Delay of cascaded chain circuit for 20nm to 7nm nodes

The performance of Cascaded flip-flops chain is predicted to be worse than adder and NAND circuits, as flip-flops are triggering on clock edges. Monitoring performance with technology scaling, it has the same trend (enhanced from 20nm down to 14nm), for instance, 14nm technology node has a speed 3% higher than 20nm speed.

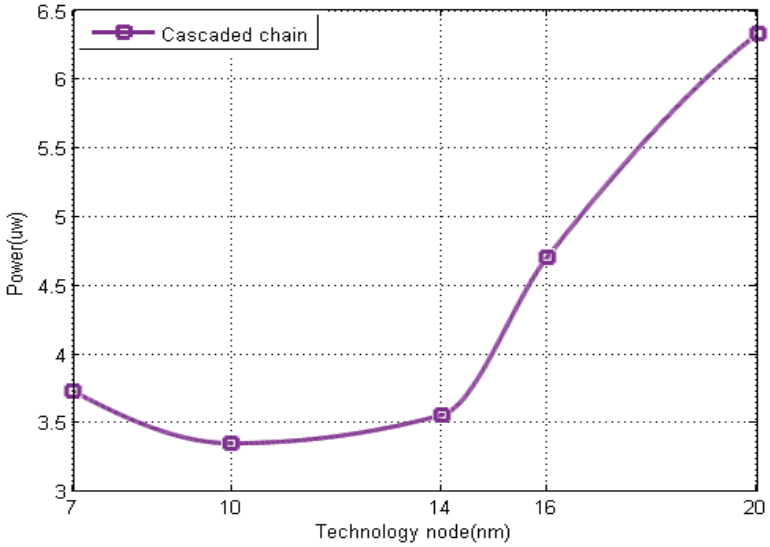


Figure 3.73: Power consumption of cascaded chain circuit for 20nm to 7nm nodes

Power consumption trend also is reduced with technology scaling as a result of supply voltage scaling with technology.

Cascaded flip-flops chain's PDP trend has its optimum value at 10nm technology node, however, 14nm technology node has a better performance, 10nm node is less power consuming than 14nm node. PDP also is improved with technology scaling.

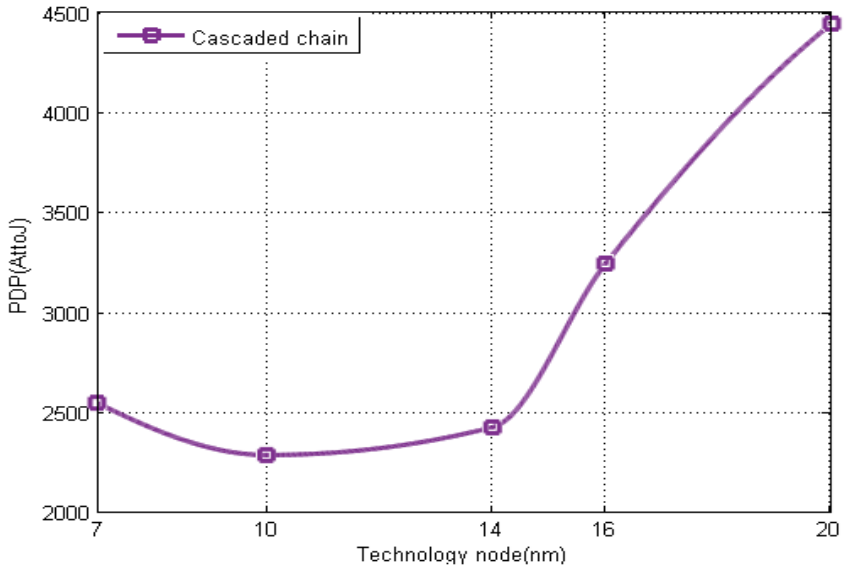


Figure 3.74: PDP of cascaded chain circuit for 20nm to 7nm nodes

FinFET-Based FPGA cluster's performance evaluation

We evaluated Tri-gate FinFET-Based FPGA cluster's performance based on metrics such as:

Operations Delay

Delay is an essential parameter in evaluating the performance of any digital circuit. Observing its trend with the technology scaling, the delay is decreasing with scaling down the technology continuously as a result of shrinking the channel length despite the scaling of the supply voltage which usually leads to degradation in the delay. FPGA cluster's performance is enhanced with technology scaling. For instance, 7nm 2-bit adder circuit speed (performance) is 15% higher its value at 20nm.

Power consumption

Power dissipation is the major metric for low power designs. Recently there has been a huge of interest in low-power devices and design techniques. The power dissipation` is continuously decreasing with scaling down the technology as a result of shrinking the channel length and the scaling of the supply voltage. For instance, 7nm cascaded flip-flop chain circuit power consumption is reduced by 41% from its value at 20nm.

Power Delay Product

As the power and delay always have a trade-off, PDP product is an important key metric in circuit's evaluation. PDP is enhanced with technology scaling from 20nm to 14nm. For instance, 7nm 2-bit adder circuit PDP is reduced by 43% from its value at 20nm.

Some design insights based on nominal simulations

Power consumption of the simulated FPGA cluster is decreased with technology scaling from 20nm down to 10nm, however, it's increased at 7nm due to the large static power of SRAMs at that technology node.

Cluster speed is increased with technology scaling starting from 20nm down to 14nm but it's degraded beyond 14nm. While using higher V_{DD} would lead to better performance, this will be at the cost of power reduction.

PDP is reduced with technology scaling from 20nm down to 14nm technology node which makes it necessary to looking for alternative scaling options such as using high mobility channel [65] and/or gate-all-around (GAA) nanowires [66] to keep on technology scaling beyond 14nm technology node.

3.4. Summary

Designing ultra-low power FPGAs requires looking for new devices to use instead of the conventional CMOS MOSFETs, multi-gate devices and DTMOS have low leakage current and can operate with low voltage supply which leads to large power reduction in the FPGA. We studied using DTMOS and FinFET in FPGA implementation. DTMOS offers power reduction by 85% more than equivalent CMOS FPGA cluster configuring it as NAND gate. FinFET provides large power reduction and better performance compared to CMOS. Some design insights are drawn for FPGA designers in sub 20nm technologies era.

Chapter 4 : Circuit Level Power Reduction Techniques

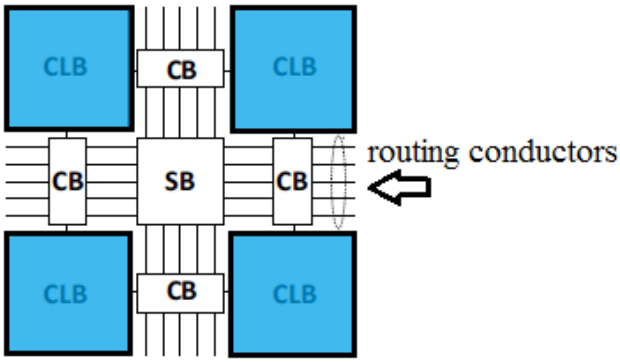
4.1. Introduction

Interconnect plays a dominant role in dynamic, and static (leakage) power dissipation of FPGAs. In comparison with custom ASICs, FPGA interconnect presents a high capacitive load, due to the presence of lengthy pre-fabricated wire segments and the programmable routing switches attached to each wire. Dynamic power scales in direct proportion to amount of capacitance switched in a logic transition. Leakage power, on the other hand, is proportional to total transistor width and interconnect comprises roughly 2/3 of an FPGA's total silicon area [25]. The influence of interconnect on overall FPGA power implies that any future low-power FPGA must include a low-power interconnection fabric. This chapter presents a novel FPGA routing switch design that reduce dynamic power dissipation.

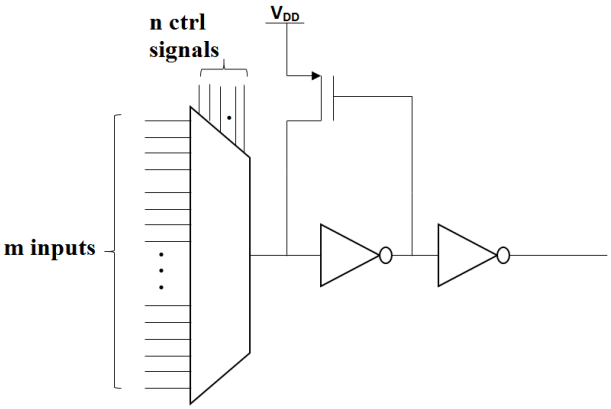
FPGA has an interesting property that it has many unused routing conductors, due to its routing flexibility, which can be used as reservoirs to the used conductors. Recycling charges through the unused reservoirs reduces the amount of charges needed to be drawn from the supply to charge the used conductors during rising transition. The remaining of the chapter is organized as the following: Section 4.2 shows a detailed description of FPGA routing fabric. Charge recycling idea is presented in Section 4.3. The proposed FPGA switch, analysis of maximum power saving than can be obtained from the multiple charge recycling technique, and SPICE simulation results are provided in Section 4.4. Section 4.5 describes application of the proposed FPGA switch in CAD tools. An experimental study of CAD tools application of the proposed technique using set of benchmarks is provided in Section 4.6. The summary of the chapter is drawn in Section 4.7.

4.2. FPGA Routing Fabric Hardware

Typically FPGA routing fabric consists of: connection boxes (CBs) to connect logic clusters (CLBs) to routing wires, switch boxes (SBs) which add flexibility of routing paths through the entire routing fabric. Island FPGA architecture (which is commonly used) is shown in Figure 4.1(a). Switch box has many buffered routing switches as shown in Figure 4.1(b). The circuit consists of m-inputs multiplexer, level restorer followed by an inverter which form a buffer, the PMOS transistor in the level restorer is used to retrieve logic ‘1’ since the MUX is implemented using NMOS pass transistors which has a poor logic ‘1’.



(a) FGPA top level architecture



(b) Routing switch (driver)

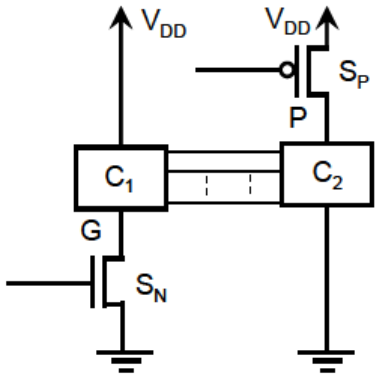
Figure 4.1: FPGA circuits structures

Since modern FPGAs have more flexible routing, the routing fabric power consumption is the dominant part of FPGA total power consumption as mentioned in [29] that it forms 62% of total power consumption. This fact encourages us to work more on reducing routing power which will has a large reflect on reducing the total FPGAs power.

4.3. Charge Recycling

Several previous studies conducted charge recycling in ASIC designs, usually in design of on-chip busses [71-73] as shown in Figures 4.2 to 4.5. The main idea of charge recycling is to store some the wasted charges during a transition of a signal from high to low on another conductor (friend), and to re-use some of them charges during a transition from low to high so that this amount of charges is saved from being drawn

from the supply. Figure 4.6(a) shows charge recovery phase, while a signal on load capacitor (CL) is falling, instead of connecting it to the ground directly, the load capacitor is disconnected from the supply and connected to a reservoir capacitor (CR) for a period of time so that they share the charges between them. After sharing the charges between them (each capacitor has a voltage of $V_{DD}/2$), the load capacitor is disconnected from the reservoir and connected to the ground to complete transition to logic '0'.



The conventional power gating structure using an NMOS or a PMOS sleep transistor for each circuit block.

Figure 4.2: Charge recycling in CMOS [71]

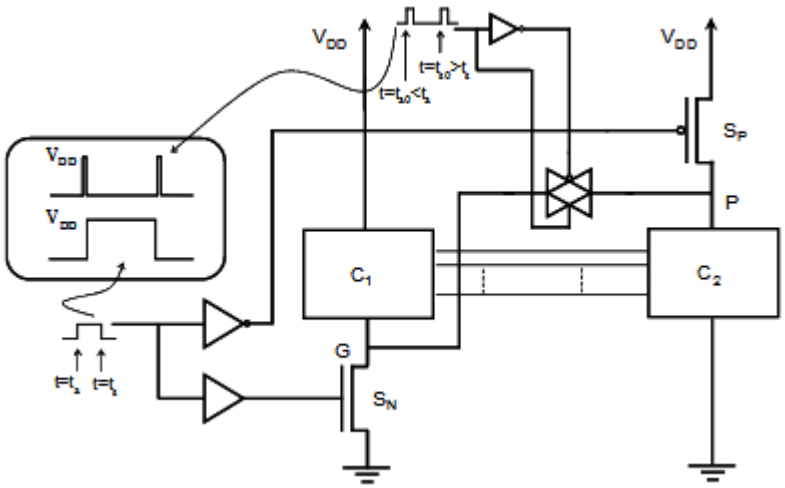


Figure 4.3: Charge recycling CMOS circuit [71]

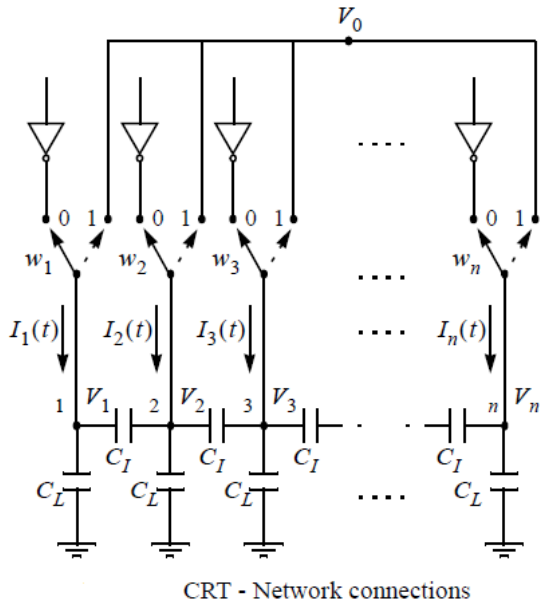


Figure 4.4: Bus charge recycling [72]

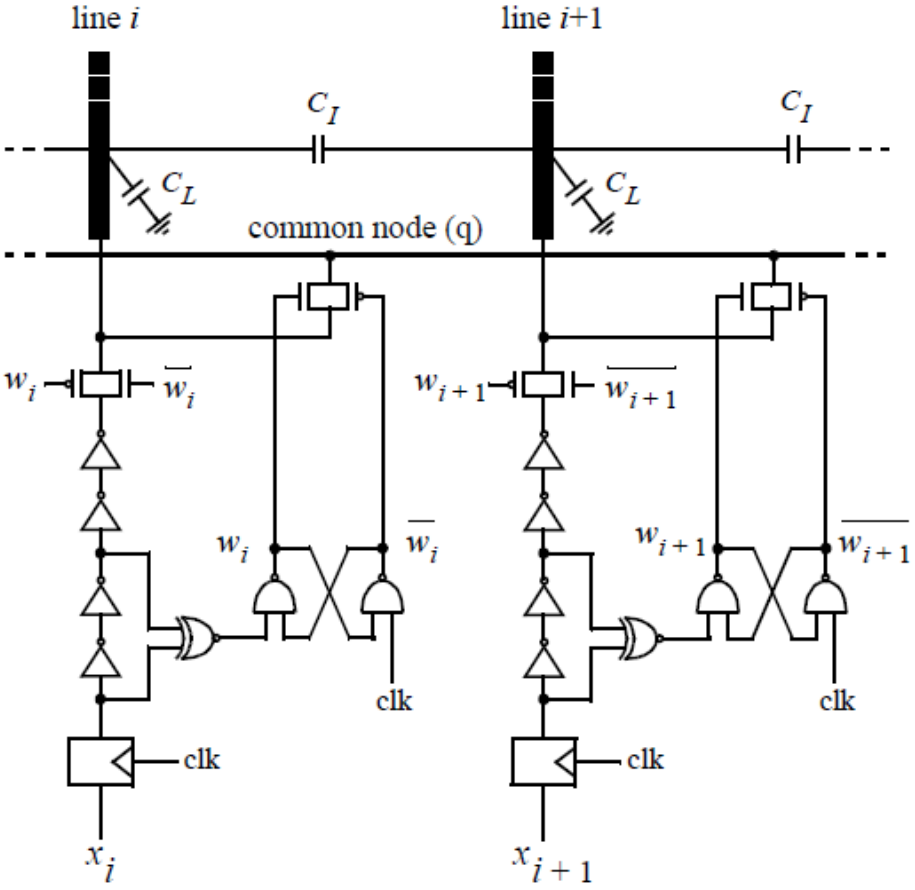
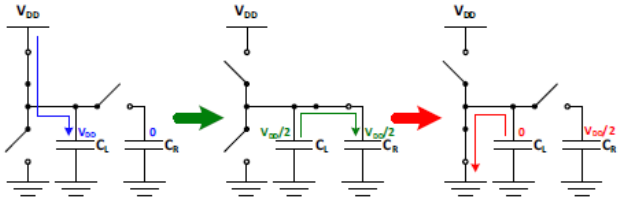
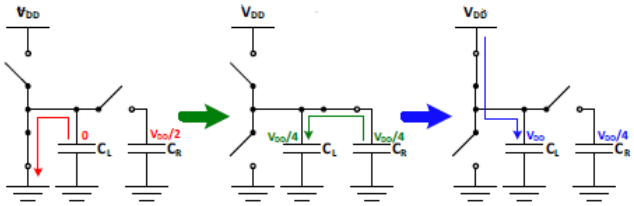


Figure 4.5: Bus charge recycling circuit [72]



(a). Charge recovery during falling signal transition

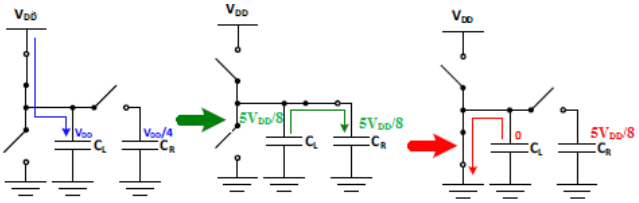


(b). Charge recycling during rising signal transition

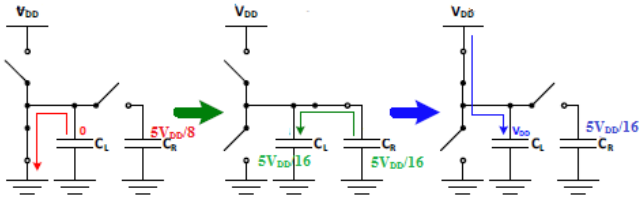
Figure 4.6: FPGA charge recycling [74]

Figure 4.6(b) shows charge recycling phase, while a signal on load capacitor (CL) is rising, instead of connecting it to the supply directly, the load capacitor is disconnected from the ground and connected to a reservoir capacitor (CR) for a sufficient time so that they share the charges (each capacitor has a voltage of $V_{DD}/4$). Then, the load capacitor is disconnected from the reservoir and connected to the supply to complete the full transition to logic '1', in this case the load will be charged with only $0.75V_{DD}$ and 25% of V_{DD} is saved.

The new proposed idea is to use multiple charge recycling phases on more than one reservoir sequentially. When a signal is going to make a falling transition followed by a rising one it go through a charge recovery + recycling (completes a cycle). Figure 4.7(a) shows the second cycle charge recovery phase where the reservoir has $V_{DD}/4$ as an initial value from first cycle above, and Figure 4.7(b) shows that the load capacitor saves $5V_{DD}/16$ from being drawn from the supply. On the other hand the proposed technique shows that $6V_{DD}/16$ can be saved, which implies 6% more saving, Figure 4.8 illustrates the first and second cycle's charge recovery and recycling phases.

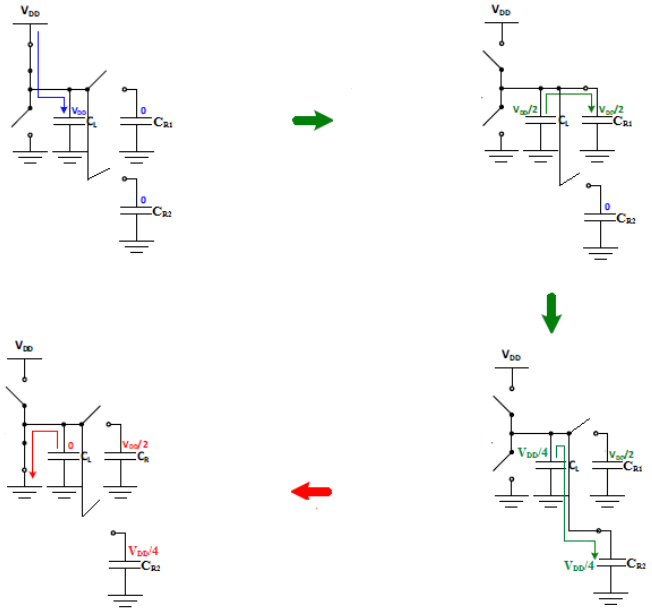


(a). Charge recovery of falling signal during second cycle

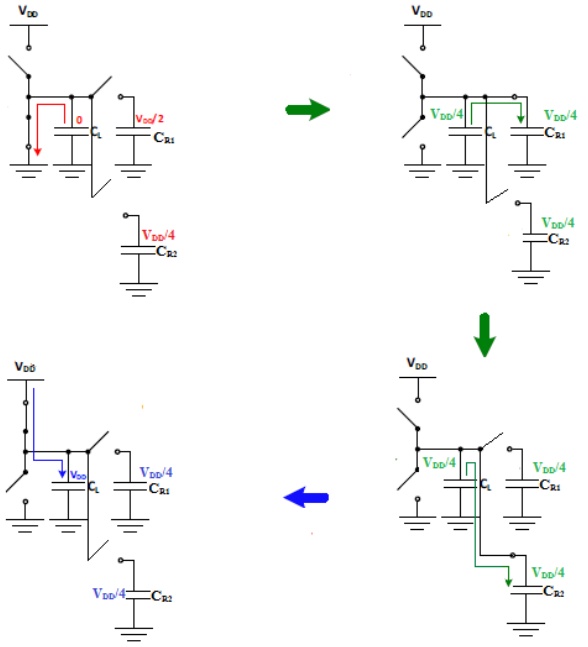


(b). Charge recycling of rising signal during second cycle

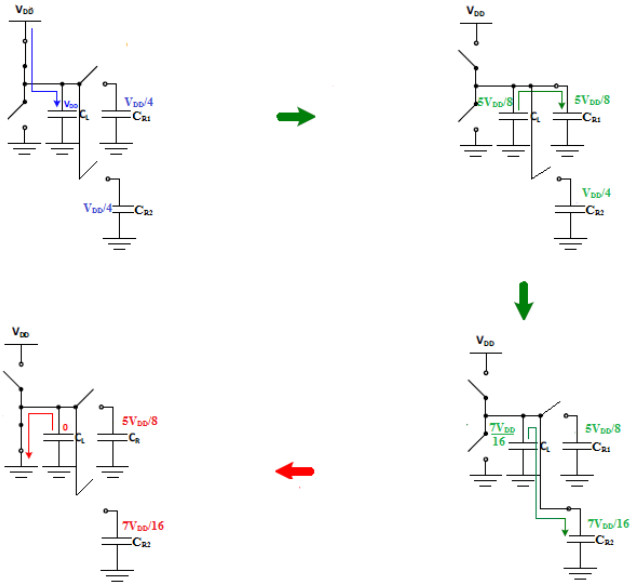
Figure 4.7: FPGA charge recycling second cycle [74]



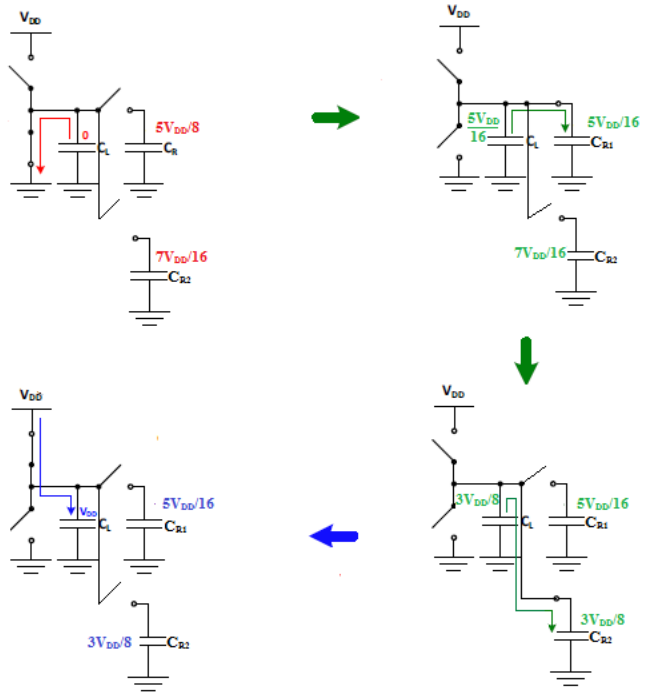
(a). Charge recovery of falling signal in multiple CR



(b). Charge recycling of rising signal in multiple CR



(c). Charge recovery of falling signal during second cycle in multiple CR

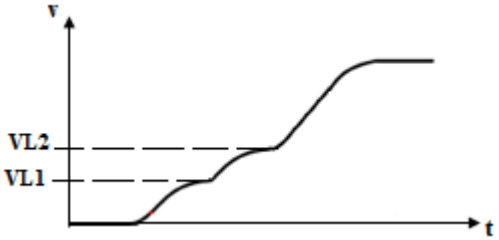


(d). Charge recycling of rising signal during second cycle in multiple CR

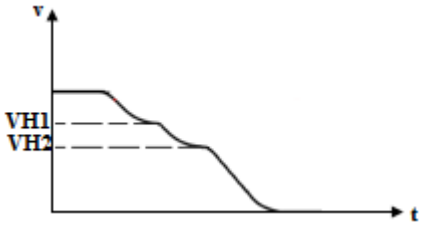
Figure 4.8: Multiple charge recycling phases

4.4. CR-Capable Interconnect Buffer

In charge recycling, when a signal is going to be discharged we store the charges from the signal using charge reservoirs and when that signal is going to rise again we reuse the stored charges from reservoirs before connecting the signal to V_{DD}. Unused routing conductors are used as reservoirs and thus we don't have to create new capacitors to be used as reservoirs. The maximum theoretically energy reduction is found to be 33 % for the CR technique [74]. Figure 4.9 illustrates multiple CR output signal waveforms during falling and rising transitions, the rising of such signal consists of three phases, the output is charged to intermediate value VL1 from charge sharing with the first reservoir, then it's charged to intermediate value VL2 from sharing with the second reservoir, and finally it's connected to the supply to be fully charged to V_{DD}. Similarly, during a falling transition, the output is discharged to intermediate value VH1 from sharing with the first reservoir, then it's discharged to intermediate value VH2 from sharing with the second reservoir and finally it's connected to the ground to be fully discharged.



(a). Waveform of a rising multiple CR signal



(b). Waveform of a falling multiple CR signal

Figure 4.9: Rising and falling signals in multiple CR

After several transitions, the intermediate values VL1, VL2, VH1, and VH2 will settle to constant values. Assuming equal capacitors (the load, and reservoirs) those intermediate values will be:

$$VH1 = \frac{V_{DD} + VL1}{2} \qquad (2)$$

$$VH2 = \frac{VH1 + VL2}{2} \qquad (3)$$

$$VL1 = \frac{VH1}{2} \qquad (4)$$

$$VL2 = \frac{VL1 + VH2}{2} \qquad (5)$$

Which implies approximately 44% power saving theoretically using our multiple CR technique.

CR capable FPGA interconnect buffer is designed in a way that the driving buffer is disabled at a transition for a period of time in which the load and reservoir capacitors are sharing charges and enabled after charge sharing to continue the falling or rising transition thus the supply doesn't have to draw the full V_{DD} to go to logic '1' as the load is already storing some charges from sharing with the reservoir. This behavior is implement using the CR buffer circuit which is illustrated in Figure 4.10 [74]. The CR buffer comprises of input stage which drives the input signal to the output and has two modes: conventional driver, and CR mode, and delay line circuit which generate a delayed version of the input signal, and charge sharing circuit to share the load and reservoir capacitors during the difference time between the input and the delayed version (the delay line time).

The two SRAM configuration cells shown in Figure 4.10 select the mode of the buffer to be in CR mode, normal mode or tri-stated mode. At a transition on VIN, the delay line generates a delayed version of it DLOUT. As there is a difference between the two signal (VIN, and DLout) the input stage is tri-stated and CR circuit is activated to allow sharing between the load and reservoir. After transition propagates, VIN, DLout signals are equal which disconnects the CR circuit and enables the input stage to continue the output transition to V_{DD} or GND.

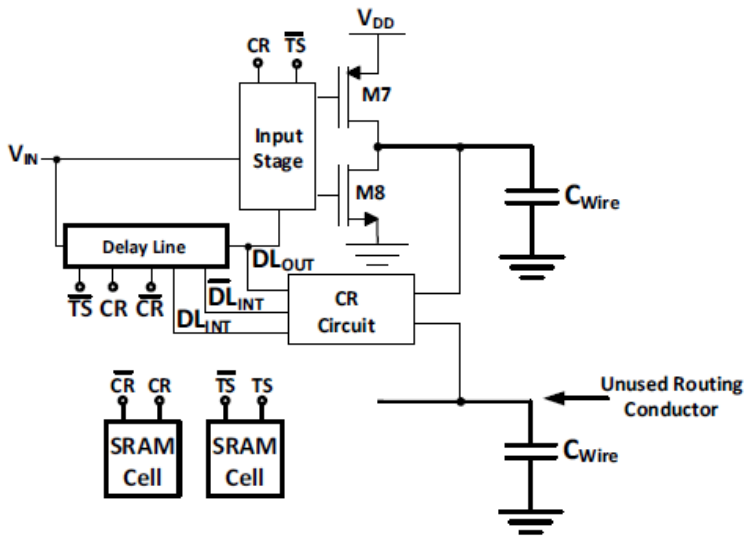


Figure 4.10: CR buffer circuit [74]

Our proposed buffer is shown in Figure 4.11, it consists of the reference buffer in [74] with additional delay line and CR circuits for the second reservoir. We exploit the pulsed charge recycling property here where the load capacitor is connected for sharing to the second reservoir after sharing with the first reservoir is done, this technique can

be used with more number of reservoir to share charges between the load and the reservoirs sequentially, the input stage is tri-stated during all reservoirs sharing and enabled after last one sharing. The second delay line circuit in the buffer consists mainly of four cascaded current-starved inverters to delay its input signal. The sub-circuits used in the buffer are shown in Figures 4.12 to 4.14.

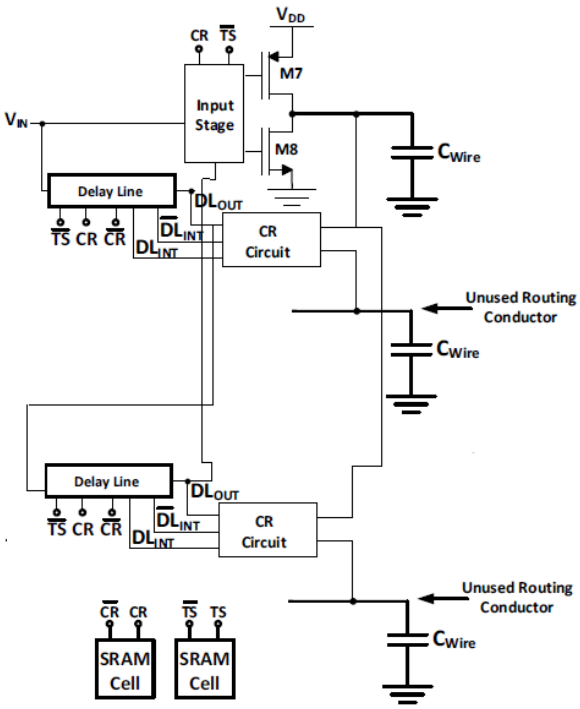


Figure 4.11: The proposed multiple CR buffer

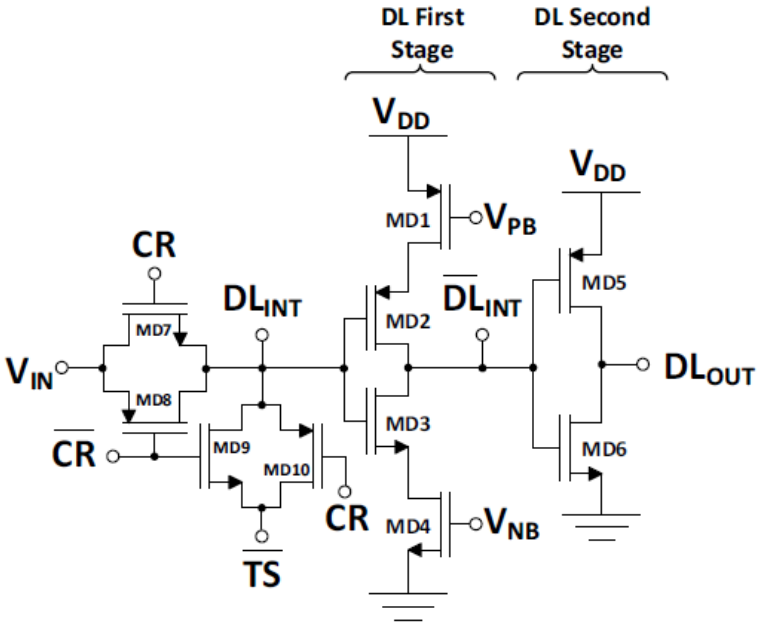


Figure 4.12: CR delay line [74]

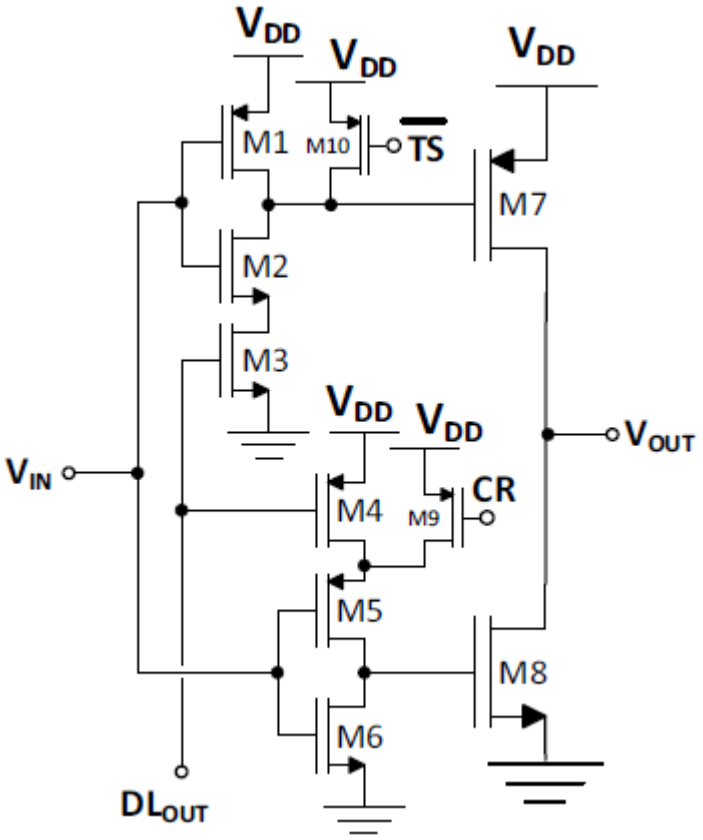


Figure 4.13: CR input stage [74]

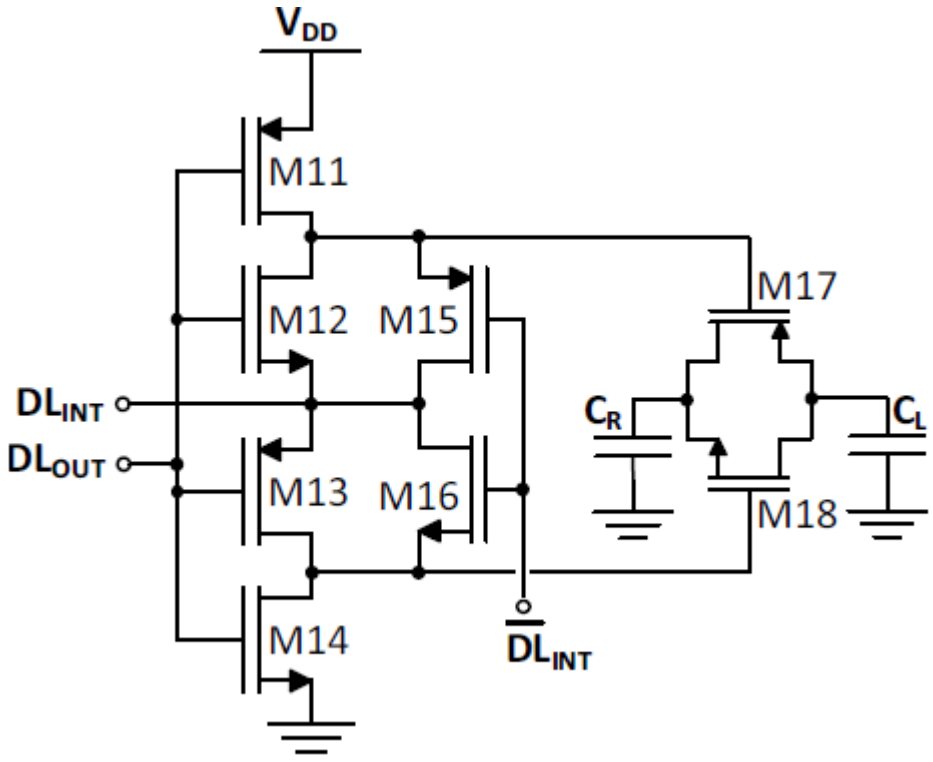


Figure 4.14: CR charge recycling sub-circuit [74]

We studied using many reservoirs sequentially as this technique will lead to more power saving. However, this saving is at the cost of performance as the load capacitor should waits for sharing with each reservoir before completing the transition. Figure 4.15 shows the idea of sequential charge recycling and how it can save more power. We also analyzed the amount of saving at each number of reservoir used in both cases: ideal case where the circuit control is done through ideal switches (overhead of control circuits is not counted), and actual case where the overhead of control circuits is considered as illustrated in Figure 4.16, and Figure 4.17 respectively.

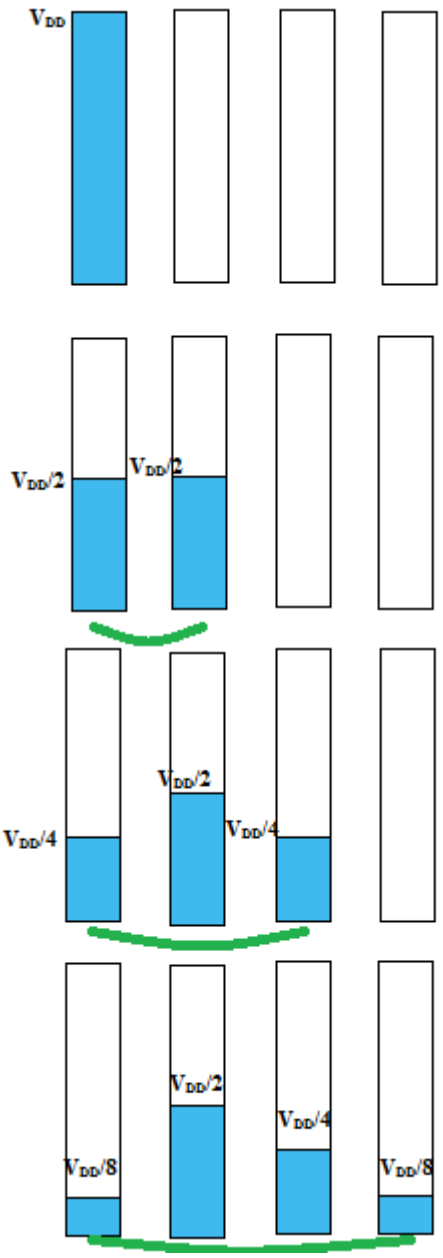


Figure 4.15: Multiple charge recycling idea

Figure 4.16 describes the ideal case where that the amount of power saving increases with increasing the number of reservoirs, but after using three reservoirs the increase of saving looks to be semi flat. Figure 4.17 describes the actual case where the power saving is enhanced with increasing the number of reservoirs from one reservoir to two but it is degraded for reservoirs more than two which is the optimum number of reservoirs to be used using the CR capable buffer circuit described above.

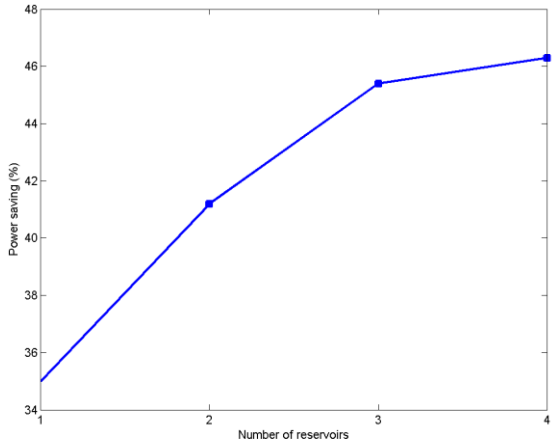


Figure 4.16: Ideal multiple charge recycling power saving

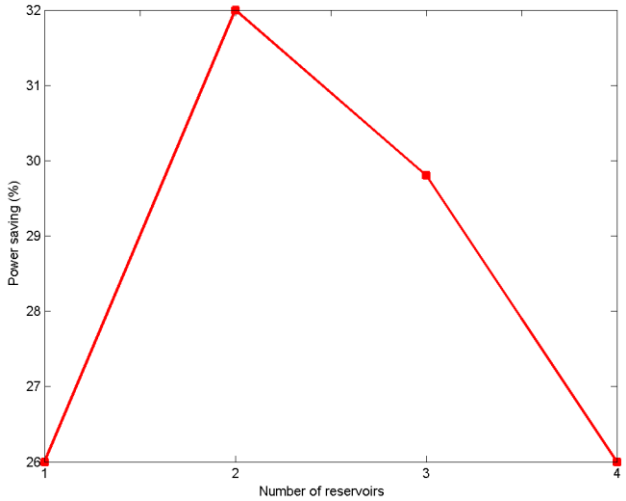


Figure 4.17: Actual multiple charge recycling power saving

SPICE simulation results of the charge recycling technique is done using HSPICE with 65nm commercial process with load capacitor of 200fF and 50MHz frequency [74], it shows that 26% of power saving can be achieved by this technique.

Multiple charge recycling technique shows 32% power saving using Spectre simulator and Cadence virtuoso tool with 65nm process and 50MHz frequency also, this power saving increase is obtained from charging recycling with the second reservoir as shown in the Figure 4.18 since the load capacitor output signal has two initial values obtained from recycling with each reservoir sequentially before the driver starts pumping the remaining charges to complete rising transition. The average delay of the proposed driver is 3.4 times the delay of the conventional driver.

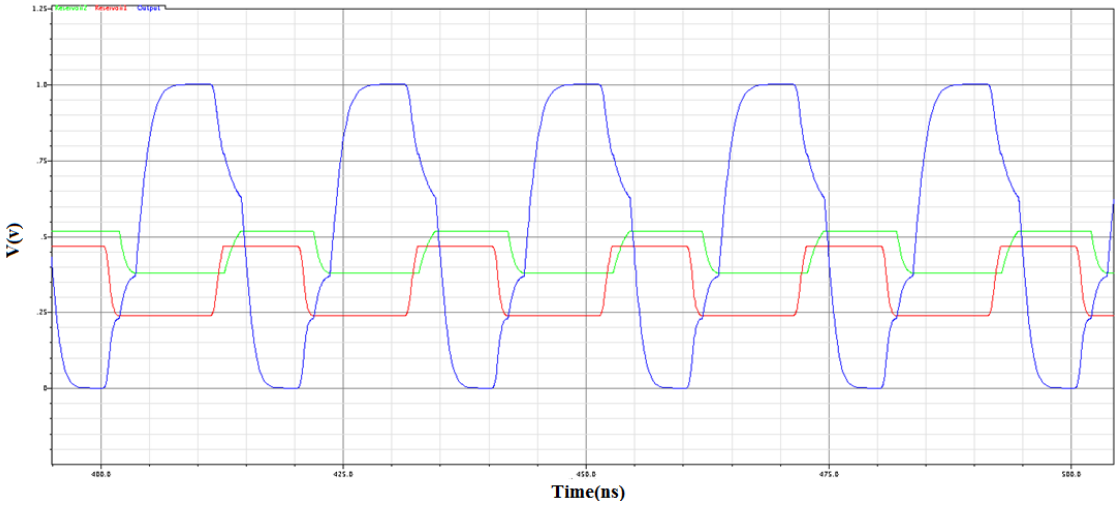


Figure 4.18: SPICE simulation waveforms of output and reservoirs. Blue for output, red for the first reservoir, and green for the second reservoir

4.5. Tool support

As multiple charge recycling technique is useful for power saving when unused conductors are exist adjacent to the used ones to be considered as reservoirs, an additional effort should be made during routing the application circuit to increase the opportunity that the signals which are non-timing critical and have high activity to be routed through CR capable switches and to have adjacent unused segments to be used as reservoirs.

The VPR router uses the PathFinder algorithm [75] which routes individual driver/load connections one at a time and uses a cost function to find a low cost path through the routing fabric from a driver to a load. The baseline cost function defined in VPR is:

$$Cost_n = (1 - Crit_i) \,.\, congs\_cost_n + Crit_i \,.\, delay\_cost_n$$

The cost function considers the routability for non-timing crictical signals and the delay for timing critical ones. Our modification to the VPR router is to consider also to put the CR preferable signals (signals with high activity and non-timing critical) in

multiple CR mode, and ensure that their reservoirs are kept unused, accordingly the modified cost function is implemented as:

$$Cost_n = (1 - Crit_i) \,.\, [congs\_cost_n + (1 - \alpha_i).\, (res1\_cost_n + res2\_cost_n) + \alpha_i \,.\, (PF \,.\, (res1\_occ_n + res2\_occ_n) + GF \,.\, not\_cr_n)] + Crit_i \,.\, delay\_cost_n$$

PF and GF are scalar tuning variables (empirically determined), and $res1\_cost_n$ is equal to $(1 - Crit_j) \cdot \alpha_j$, where j is the index of the connection currently occupying the first potential reservoir of node n, similarly $res2\_cost_n$ for the second reservoir. $res1\_occ_n$ is a binary variable which equals one in case the CR capable switch's first reservoir is used, and $res2\_occ_n$ for the second reservoir, and not_cr_n is a binary variable which equal one in case the signal to be routed doesn't use CR capable switch. Our target is to route nets with high activity and sufficient slack to use CR capable switches with unoccupied reservoirs since if any of the two reservoirs is occupied this will lead to power saving opportunity lose. Thus the second term in the modification penalizes such cases, while the first term penalizes the case where we use one of the two reservoirs of another CR preferable switch while we route the current net. The last term penalizes the case where the current net is CR preferable switch but routed through non CR capable switch.

4.6. EXPERIMENTAL STUDY

To assess the merits of the proposed buffer circuit and CAD flow, we used the set of benchmark circuits packaged with VPR 7.0 [76]. Our baseline non-CR-capable architecture contains unidirectional wire segments which span four CLB tiles, and uses the Wilton switch block [77], and has logic blocks with ten 6-LUTs/FFs per CLB. We simulated all benchmark circuits on the baseline architecture to determine the minimum channel width (W_{min}). Then we used $W = 1.3 \times W_{min}$ in all simulations to reflect a medium stress. We computed signals switching activity using ACE switching activity estimator tool [78]. We made the following assumptions about the architecture in our study: (1) each routing segment is paired with two other routing segments and either can serve as the reservoir for the other, and (2) the paired routing conductors have the same start/end points but run in opposite directions. The power saving results correspond to 26% saving for CR switch, and 32% saving for multiple CR switch. The results for fully multiple CR capable switches architecture are detailed in Table 4.1. The results show 23.4% power saving for multiple CR technique, the power saving is diminished from 32% to 23.4% since not all the signals can be put in multiple CR mode, on the other hand, the CR reference technique shows 20.75% saving, we suspect that the difference between this result and the reference result is due to the difference between routing paring algorithm and the power model used.

We also studied the area overhead to implement such architecture, we used minimum width transistor as the area measurement unit since it's used in VPR as the area metric. The total area overhead of the proposed switch equals 89 minimum width

transistor, 77 minimum width transistor for CR circuits, delay lines, and input stage circuit, and 12 minimum width transistor for the two SRAMs used for mode selection. The proposed technique shows a routing area increase by 50%, while the total area is increased by 6.1% which means an increase by 3% ($\sqrt{1.061}$) in each x and y dimensions for square tile layout. Hence the power reduction (23.4%) exceeds the increase of wire capacitance (3% at most)

Table 4.1: Power reduction and area overhead of each benchmark circuit in fully populated with Multiple CR switches archirecture

Circuit	CR Pwr Red.	Multiple CR Pwr Red.	CR Routing Area Increase	Multiple CR Routing Area Increase
diffeq1	25%	28.2%	26%	50.9%
raygentop	22%	24%	25.3%	50%
sha	21%	24%	24.3%	48%
blob_merge	15%	18%	24.2%	47.9%
stereovision0	21%	23%	23.4%	46.3%
or1200	19%	22.3%	24.3%	48.1%
mkSMAdapter4B	22%	25%	24.6%	48.6%
boundtop	21%	23%	25%	49.5%
Geomean	**20.8%**	**23.4%**	**25.5%**	**50.4%**

4.7. Summary

Power dissipation in FPGAs is dominated by interconnection fabric consumption, which makes low-power interconnect a mandatory for future low-power FPGAs. In this chapter, we proposed a novel multiple charge recycling technique to reduce interconnections power by recycling charges using unused conductors exploiting the fact that many of FPGA conductors remain unused. The proposed design offers 32% power saving in SPICE simulations. We also proposed the modifications needed for CAD tools to support the multiple charge recycling switch, and run a set of benchmarks to evaluate power savings and area overhead, using VTR project benchmarks the proposed technique offers 23.4% power saving and 6.1% total area increase.

Discussion and Conclusions

Technology scaling trends imply a dramatic increase in leakage power and a steady increase in dynamic power with each successive process generation. Field-programmable gate arrays (FPGAs) require considerable hardware overhead to offer programmability, making them less power-efficient than custom ASICs for implementing a given logic circuit. The huge number of transistors on the largest FPGA chips suggest that the power trends associated with scaling may impact FPGAs more severely than custom ASICs. Despite this, until recently, the majority of published research on FPGA CAD and architecture, as well as the focus of the commercial vendors, has been on improving FPGA speed and density. Power management in FPGAs will be mandatory at the 65nm technology node and beyond to ensure correct functionality, provide high reliability, and to reduce packaging costs. Furthermore, lower power is needed if FPGAs are to be a viable alternative to ASICs in low-power applications, such as battery-powered electronics.

This dissertation has contributed new device level and circuit-level techniques for the optimization of FPGA power consumption:

Chapter 3 looked at device level techniques to reduce FPGA power consumption and considered two techniques. The first technique involves replacing traditional CMOS MOSFETs in FPGA cluster with FinFET devices using predictive technology models for multi-gate devices for technologies 20nm, 16nm, 14nm, 10nm, and 7nm. Each component of FPGA cluster is evaluated separately in terms of digital metrics: power, performance, and energy considering process variations like threshold voltage variations due to HCI and BTI, and temperature variations. This work has been published in [18]. Optimum supply voltage for a variety of flip-flops that can be used in FPGAs also are determined to reduce the overall FPGA energy. This work has been published in [19]. Overall evaluation of general FPGA cluster also is performed with set of benchmarks. The other techniques involves using dynamic threshold MOSFET (DTMOS) instead of conventional CMOS MOSFET in FPGA logic and routing fabric. DTMOS shows power reduction for all FPGA's components such as SRAM, MUXs, and flip-flops, but it degrades the performance of them, in terms of energy DTMOS MUXs shows better energy consumption compared to CMOS MUXs. For instance 16:1 DTMOS MUX has less energy than CMOS one by 35%. DTMOS FPGA cluster also compared with CMOS cluster and showed better power consumption, on average power reduced by 85% using 4-NAND benchmark circuit.

Chapter 4 targeted reduction FPGA power using circuit techniques. A novel proposed technique, multiple charge recycling, can save more dynamic power consumption in FPGA route fabric. The technique involves recycling charges from falling interconnects to be used by rising ones. The chapter goes through the idea of multiple charge recycling to maximum power reduction that can be achieved using this technique theoretically, then circuit details of the new proposed switch, and application of this technique in CAD tools. SPICE simulation of the proposed switch indicated power saving by 32% in and interconnect. VTR project is used to estimate the power

saving and area overhead due to application of this technique using set of benchmarks defined in the project and showed power reduction by 23.4%.

As extension to this work, studying DTMOS FPGA tile metrics, and using the different DTMOS configuration can be addressed as future work, also studying using other emerging new devices like RRAM and magnetic RAM in entire FPGA tile. On circuit level, combing multiple charge recycling switch with other related circuit and architecture level power reduction techniques is targeted as future work.

References

1. D. Lammers. “Chip Industry Growing, But ExpertsWonder For How Long". EE Times, December 2003.
2. M. Santarini. “Are Structured ASICs a Dead End for EDA?". EE Times, June 23 2003.
3. J.M. Hergenrother, D. Monroe, F.P. Klemens, A. Komblit, G.R. Weber, W.M. Mansfield, M.R. Baker, F.H. Baumann, K.J. Bolan, J.E. Bower, N.A. Ciampa, R.A. Cirelli, J.I. Colonell, D.J. Eaglesham, J. Frackoviak, H.J. Gossmann, M.L. Green, S.J. Hillenius, C.A. King, R.N. Kleiman, W.Y.C. Lai, J.T.-C. Lee, R.C. Liu, H.L. Maynard, M.D. Morris, S.-H. Oh, C.-S. Pai, C.S. Rafferty, J.M. Rosamilia, T.W. Sorsch, and H.-H. Vuong. The vertical replacement-gate (vrg) mosfet: a 50-nm vertical mosfet with lithography-independent gate length. In Electron Devices Meeting, pages 75 –78, 1999.
4. E. J. Nowak *et al*., "Turning silicon on its edge [double gate CMOS/FinFET technology]," in *IEEE Circuits and Devices Magazine*, pp. 20-31, Jan-Feb 2004.
5. A. Es-Sakhi and M. H. Chowdhury, "Analytical model to estimate the subthreshold swing of SOI FinFET," *Electronics, Circuits, and Systems (ICECS)*, pp. 52-55, 2013.
6. Tsu-Jae King. Finfets for nanoscale cmos digital integrated circuits. In IEEE/ACM International conference on Computer-aided design (ICCAD), pages 207–210, 2005.
7. Brian Swahn and Soha Hassoun. Gate sizing: finfets vs 32nm bulk mosfets. In ACM/IEEE Design Automation Conference (DAC), pages 528–531, 2006.
8. R. Chau, B. Doyle, J. Kavalieros, D. Barlage, A. Murthy, M. Doczy, R. Arghavani, and S. Datta. Advanced depleted-substrate transistors: Single-gate, double-gate and tri-gate. In International Conference on Solid-State Devices and Materials (SSDM), 2002.
9. Chenming Hu. New sub-20nm transistors – why and how. In ACM/IEEE Design Automation Conference (DAC), pages 460 –463, June 2011.
10. Intel. Intel announces new 22nm 3D Tri-gate transistors. Technical report, 2011.
11. P. L. Yang, T. B. Hook, P. J. Oldiges and B. B. Doris, "Vertical Slit FET at 7-nm Node and Beyond," in *IEEE Transactions on Electron Devices*, pp. 3327-3334, Aug. 2016.

12. V. George and J. Rabaey. Low-Energy FPGAs: Architecture and Design. Kluwer Academic Publishers, Boston, MA, 2001.

13. P. Zuchowski, C. Reynolds, R. Grupp, S. Davis, B. Cremen, and B. Troxel. “A Hybrid ASIC and FPGA Architecture". In: IEEE International Conference on Computer-Aided Design, pp. 187-194, San Jose, CA, 2002.

14. B. Fuller. “IC Vendors Lock Horns Over Design Approaches". EE Times, February 11 2004.

15. L. Stok and J. Cohn. “There is Life Left in ASICs". In: ACM/IEEE International Symposium on Physical Design, pp. 48-50, Monterey, CA, 2003.

16. Fariborz Assaderaghi ,Dennis Sinitsky, Stephen A.Parke, Jeffrey Bokor,Ping K. Ko, and Chenming Hu “ Dynamic Threshold –Voltage MOSFET (DTMOS) for Ultra-Low Voltage VLSI,” IEEE Transactions on Electron Devices, pp. 414-422, March 1997.

17. Predictive Technology Model (PTM). http://ptm.asu.edu/.

18. O. Abdelkader, H. Mostafa, H. Abdelhamid and A. Soliman, "The impact of FinFET technology scaling on critical path performance under process variations," Energy Aware Computing Systems & Applications (ICEAC), 2015 International Conference, pp. 1-4, Cairo, 2015.

19. O. Abdelkader, H. Mostafa, H. Abdelhamid and A. Soliman, "Impact of technology scaling on the minimum energy point for FinFET based flip-flops," 2015 IEEE International Conference on Electronics, Circuits, and Systems (ICECS), pp. 462-465, Cairo, 2015.

20. Virtex-7 FPGA Data Sheet. Xilinx, Inc., San Jose, CA, 2016.

21. Stratix-V FPGA Data Sheet. Altera Corp., San Jose, CA, 2016.

22. G. Lemieux and D. Lewis. “Circuit Design of Routing Switches". In: ACM/SIGDA International Symposium on Field Programmable Gate Arrays, pp. 19-28, Monterey, CA, 2002.

23. G. Lemieux. “Directional and Single-Driver Wires in FPGA Interconnect". In: IEEE International Conference on Field-Programmable Technology, pp. 41-48, Brisbane, Australia, 2004.

24. D. Lewis, V. Betz, D. Jefferson, A. Lee, C. Lane, P. Leventis, S. Marquardt, C. McClintock, B. Pedersen, G. Powell, S. Reddy, C. Wysocki, R. Cliff, and J. Rose."The Stratix Routing and Logic Architecture". In: ACM/SIGDA International Symposium on Field Programmable Gate Arrays, pp. 12-20, Monterey, CA, 2003.

25. A. Rahman and V. Polavarapuv. “Evaluation of Low-Leakage Design Techniques for Field-Programmable Gate Arrays". In: ACM/SIGDA International Symposium on Field Programmable Gate Arrays, pp. 23-30, Monterey, CA, 2004.

26. G. Lemieux. “Routing Architecture for Field-Programmable Gate Arrays". In: Ph.D. Thesis, Department of Electrical and Computer Engineering, University of Toronto, Toronto, Ontario, Canada, 2003.

27. K. Poon, A. Yan, and S. J. E. Wilton. “A Flexible Power Model for FPGAs". In: International Conference on Field-Programmable Logic and Applications, pp. 312-321, Montpellier, France, 2002.

28. F. Li, D. Chen, L. He, and J. Cong. “Architecture Evaluation for Power-Efficient FPGAs". In: ACM/SIGDA International Symposium on Field Programmable Gate Arrays, pp. 175-184, Monterey, CA, 2003.

29. T. Tuan, A. Rahman, S. Das, S. Trimberger, and S. Kao, “A 90-nm low-power FPGA for battery-powered applications,” IEEE Trans. on CAD, pp. 296–300, 2007.

30. G. Yeap. Practical Low Power Digital VLSI Design. Kluwer Academic Publishers, Boston, MA, 1998.

31. A. Ye, J. Rose, and D. Lewis. “Using Multi-Bit Logic Blocks and Automated Packing to Improve Field-Programmable Gate Array Density for Implementing Datapath Circuits". In: IEEE International Conference on Field-Programmable Technology, pp. 129-136, Brisbane, Australia, 2004.

32. T. Tuan and B. Lai. “Leakage Power Analysis of a 90nm FPGA". In: IEEE Custom Integrated Circuits Conference, pp. 57-60, San Jose, CA, 2003.

33. Spartan-3 FPGA Data Sheet. Xilinx, Inc., San Jose, CA, 2004.

34. C. Kim, J.-J. Kim, S. Mukhopadhyay, and K. Roy. “A Forward Body-Biased Low-Leakage SRAM Cache: Device and Architecture Considerations". In: ACM/IEEE International Symposium on Low-Power Electronics and Design, pp. 6-9, Seoul, Korea, 2003.

35. Pooran Singh, Santosh Kumar Vishvakarma, Device/Circuit/Architectural Techniques for Ultra-low Power FPGA Design, Microelectronics and Solid State Electronics, pp. 1-15, 2013.

36. Y. Lin and J. Cong. Power modeling and characteristics of field programmable gate arrays. IEEE Transactions on Computed-Aided Design for Integrated Circuits and Systems, November 2005.

37. K. Poon, S. Wilton, and A. Yan. A detailed power model for field-programmable gate arrays. ACM Transactions on Design Automation of Electronic Systems, 2005.

38. P. Jamieson, W. Luk, S. J. Wilton, and G. A. Constantinides. An energy and power consumption analysis of FPGA routing architectures. Proceedings of the International Conference on Field-Programmable Technology, pages 324-327, December 2009.

39. C. Chow, L. Tsui, P. Leong, W. Luk, and S. Wilton. Dynamic voltage scaling for commercial FPGAs. Proceedings of the International Conference on Field-Programmable Technology, pages 173-180, December 2005.

40. B. Calhoun, F. Honore, and A. Chandrakasan. Design methodology for fine-grained leakage control in MTCMOS. In Proceedings of the International Symposium on Low Power Electronics and Design, 2003.

41. A. Gayasen, Y. Tsai, N. Vijaykrishnan, M. Kandemir, M. J. Irwin, and T. Tuan. Reducing leakage energy in FPGAs using region-constrained placement. In Proceedings of the International Symposium on Field-Programmable Gate Arrays, pages 51-58, 2004.

42. A. Rahman, S. Das, T. Tuan, and S. Trimberger. Determination of Power Gating Granularity for FPGA Fabric. In Proceedings of the IEEE Custom Integrated Circuits Conference, pages 9-12, September 2006.

43. Y. Hu, Y. Lin, L. He, and T. Tuan. Physical synthesis for FPGA interconnect power reduction by dual-Vdd budgeting and retiming. ACM Transactions on Design Automation of Electronic Systems, 2008.

44. J. Anderson and F. Najm. Low-power programmable routing circuitry for FPGAs. In Proceedings of the International Conference on Computer-Aided Design, 2004.

45. M. Hioki, T. Kawanami, T. Tsutsumi, T. Nakagawa, T. Sekigawa, and H. Koike. Evaluation of granularity on threshold voltage control in flex power FPGA. In Proceedings of the International Conference on Field-Programmable Technology, pages 17-24, December 2006.

46. T. Kawanami, M. Hioki, Y. Matsumoto, T. Tsutsumi, T. Nakagawa, T. Sekigawa, and H. Koike. Optimal set of body bias voltages for an FPGA with field-programmable Vth components. Proceedings of the International Conference on Field-Programmable Technology, pages 329-332, December 2006.

47. D. Lewis, E. Ahmed, D. Cashman, T. Vanderhoek, C. Lane, A. Lee, and P. Pan. Architectural enhancements in Stratix-III and Stratix-IV. In Proceedings of the International Symposium on Field-Programmable Gate Arrays, pages 33-42. ACM, 2009.

48. M. Klein. The Virtex-4 Power Play. Xcell Journal, Spring 2005.

49. V. George, H. Zhang, and J. Rabaey. The design of a low energy FPGA. In Proceedings of the International Symposium on Low Power Electronics and Design, pages 188-193, 1999.

50. J. Ryan and B. Calhoun. A sub-threshold FPGA with low-swing dual-VDD interconnect in 90nm CMOS. In Proceedings of the IEEE Custom Integrated Circuits Conference, pages 1-4, September 2010.

51. Kureshi A.K. and Mohd. Hasan, "DTMOS Based Low Power High Speed Interconnects for FPGA," J. Computers, pp. 921-926, October 2009.

52. D. Kumar, P. Kumar and M. Pattanaik, "Performance Analysis of Dynamic Threshold MOS (DTMOS) Based 4-Input Multiplexer Switch for Low Power and High Speed FPGA Design", Symposium on Integrated Circuits and Systems Design, pp. 2-7, Sep. 2010.

53. S. Sinha, G. Yeric, V. Chandra, B. Cline, Y. Cao; "Exploring Sub-20nm FinFET design with Predictive Technology Models," in Design Automation Conference (DAC), 2012 49th ACM/EDA/IEEE, June 2012.

54. J. Kavalieros et al., "Tri-gate transistor architecture with high-k gate dielectrics, metal gates and strain engineering," in VLSIT, pp. 50 -51, 2006.

55. C. Wu et al., "High performance 22/20nm finfet cmos devices with advanced high-k/metal gate scheme," in IEDM, pp. 27.1.1 -27.1.4, dec. 2010.

56. C.-Y. Chang et al., "A 25-nm gate-length finfet transistor module for 32nm node," in IEDM, pp. 1 -4, dec. 2009.

57. T. Yamashita et al., "Sub-25nm finfet with advanced fin formation and short channel effect engineering," in VLSIT, pp. 14 -15, june 2011.

58. "International Technology Roadmap of Semiconductors," http://www.itrs.net/Links/2013ITRS/Home2013.htm, 2013.

59. P Metzgen and D Nancekievill, “Multiplexer Restructuring For FPGA Implementation Cost Reduction,” USA DAC 2005, pp.421-426, June 13-17, Anaheim, California, 2005.

60. Nunes, C.; Butzen, P.F.; Reis, A.I.; Ribas, R.P. "A methodology to evaluate the aging impact on flip-flops performance", Integrated Circuits and Systems Design (SBCCI), 2013 26th Symposium.

61. Rao, V.G.; Mahmoodi, H., "Analysis of reliability of flip-flops under transistor aging effects in nano-scale CMOS technology," Computer Design (ICCD), 2011 IEEE 29th International Conference on , pp.439,440, 9-12 Oct. 2011.

62. E. Takeda et al., “New hot-carrier injection and device degradation in submicron MOSFETs,” IEEE Proceedings I, Solid-State and Electron Devices, pp. 144–150, Jun. 1983.

63. A Dixit, S Khandelwal and S Akashe, “Power Optimization of 8:1 MUX using Transmission Gate Logic (TGL) with Power Gating Technique” International Journal of Computer Applications, August 2014.

64. S Sinha, B Cline, G Yeric, V Chandra, Y Cao, “Design benchmarking to 7nm with FinFET predictive technology models” Proceedings of the 2012 ACM/IEEE international symposium on Low power electronics and design.

65. K. C. Saraswat, C. O. Chui, D. Kim, T. Krishnamohan and A. Pethe, “High Mobility Materials and Novel Device Structures for High Performance Nanoscale MOSFETs”, IEEE Intl. Devices Meeting (IEDM), pp. 1-4, 2006.

66. C.-H. Park, S.-H. Lee, Y.-R. Kim, C.-K. Baek and Y.-H. Jeong, “Gate-All-Around Silicon Nanowire Field Effect Transistors”, IEEE Intl. Conf. Nanotechnology, pp.255-259, 2011.

67. Esmaeili, S.E.; Al-Khalili, A.J., "10 GHz throughput FinFET dual-edge triggered flip-flops," Electrical and Computer Engineering (CCECE), 2014 IEEE 27th Canadian Conference on , pp.1,4, 4-7 May 2014.

68. Z. Jakšić, R. Canal, Effects of FinFET Technology Scaling on 3T and 3T1D Cell Performance Under Process and Environmental Variations, 3rd Workshop on Workshop on Resilient Architectures, in conjunction with the 45th Annual IEEE/ACM International Symposium on Microarchitecture, Vancouver (Canada), December 2012.

69. Amat, E.; Almudever, C.G.; Aymerich, N.; Canal, R.; Rubio, A., "Impact of FinFET Technology Introduction in the 3T1D-DRAM Memory Cell," *Device and Materials Reliability, IEEE Transactions on*, pp.287,292, March 2013.

70. Mostafa, H., M. Anis, and M. Elmasry, "Comparative analysis of process variation impact on flip-flops soft error rate", Quality Electronic Design, 2009.

71. E. Pakbaznia, F. Fallah and M. Pedram, "Charge Recycling in Power-Gated CMOS Circuits," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, pp. 1798-1811, Oct. 2008.

72. P. Sotiriadis, T. Konstantakopoulos, and A. Chandrakasan, “Analysis and implementation of charge recycling for deep sub-micron buses,” in ACM/IEEE ISLPED, pp. 364–369, 2001.

73. K. Patel, W. Lee, and M. Pedram, “In-order pulsed charge recycling in off-chip data buses,” in ACM Great Lakes Symp. on VLSI, 2008.

74. S. Huda et al. Charge recycling for power reduction in FPGA interconnect. In IEEE FPL 2013.

75. L. McMurchie and C. Ebeling, “PathFinder: A negotiation-based performance-driven router for FPGAs,” in ACM FPGA, pp. 111–117, 1995.

76. J. Rose and et al., “The VTR project,” in ACM FPGA, pp. 77–86, 2012.

77. S. J. Wilton, “Architecture and algorithms for field programmable gate arrays with embedded memory,” Ph.D. dissertation, 1997.

78. J. Lamoureux and S. Wilton, “Activity estimation for field-programmable gate arrays,” in IEEE FPL, pp. 1–8, 2006.

Appendix A: Introduction to VTR

What is VPR?

Versatile place and route (VPR) is a CAD tool for FPGAs that maps an arbitrary (user specified) netlist (circuits, memories, etc) to an arbitrary (user specified) FPGA architecture.

VPR has two required and many optional parameters; it is invoked by entering:
$ vpr architecture.xml circuit_name[.blif] [-options]

- architecture.xml describes the architecture of the FPGA
- The first tag in all architecture files is the <architecture> tag. This tag contains all other tags
- The architecture tag contains seven other tags. They are <models>, <layout>, <device>, <switchlist>, <segmentlist>, <directlist> and < complexblocklist>.
- These information and furthermore can be found in VPR user manual.

History of VPR

- Developed by the University of Toronto.
- VPR for place and route and T-Vpack for packing (clustering).
- Combined to one tool (VPR).
- From VPR 6, it comes a part from VTR (Verilog to route) project.
- Current release is VPR 7.
- VTR contains VPR, ODINII, ABC, ACE, and other (utilities scripts, benchmarks, architectures, technology files).

Installing VPR

- Getting the source of VTR; either by downloading it from:
- http://www.eecg.toronto.edu/~vaughn/vpr/register.html
  OR cloning it from git repository (developers repository):
  $ git clone https://github.com/verilog-to-routing/vtr-verilog-to-routing
- Installing on Linux machine:
  Running “make” in the parent tree of the source code

- Installing on Windows machine:

- VPR parent directory has a visual studio project "VPR.vcxproj", To input the required command-line options, go to VPR Properties in the Project menu and select Configuration Properties then Debugging. Enter the architecture filename, circuit filename and any optional parameters you wish in the Command Arguments box.
- Using Cygwin if you prefer Linux based environment on Windows.

VTR workflow

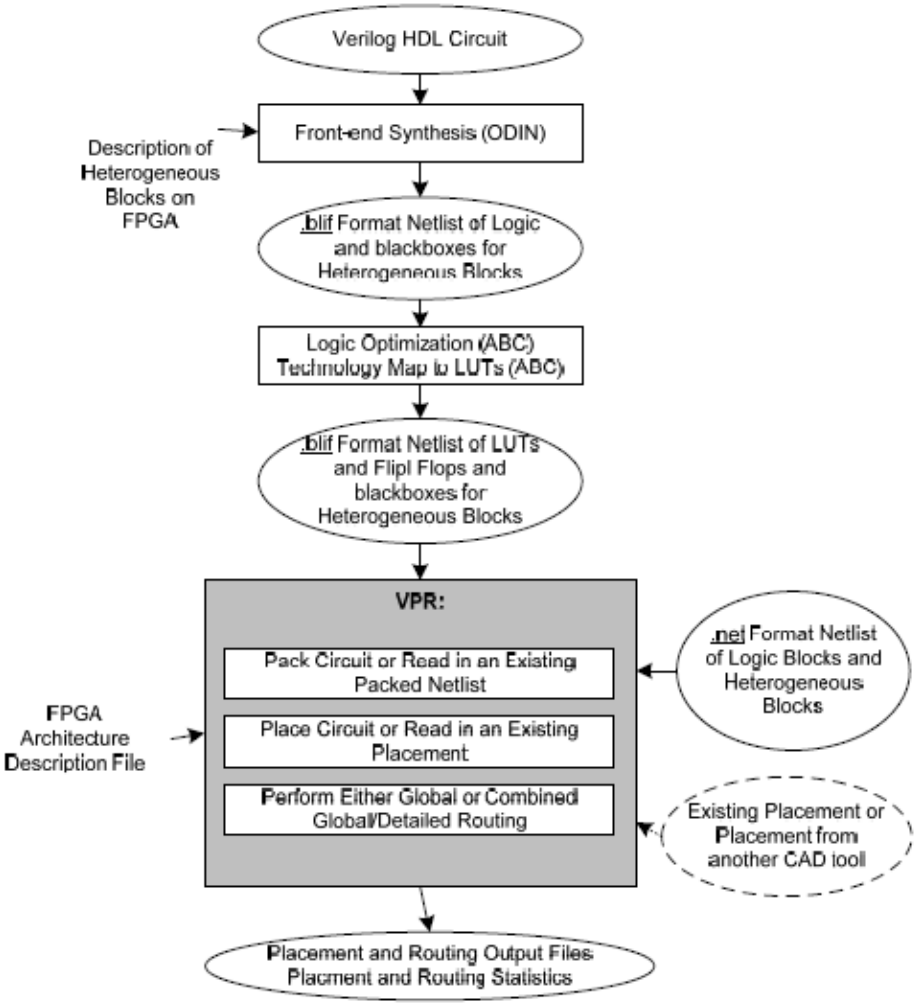


Figure A.1: VTR workflow

- Running basic flow:
- Architecture (vtr_release\vtr_flow/arch/k6_frac_N10_mem32K_40nm.xml): 6-bit LUT, 10 BLEs with FF, 32k Memory, 40nm technology .
- Circuit "benchmark"(vtr_release\vtr_flow\benchmarks\Verilog\ch_intrinsics.v)
- $ cd vtr_release/vtr_flow/tasks

- $ ../scripts/run_vtr_task.pl basic_flow
- $ ../scripts/parse_vtr_task.pl basic_flow
- Results are generated under: basic_flow/run<number>/
- To calculate power estimations in the results, add to basic_flow/config/config.txt:
- parse_file=vpr_power.txt, and "script_params= -power -cmos_tech vtr_release/vtr_flow/tech/PTM_<tech_node>/<tech_node>.xml

Anatomy of VTR

- abc_with_bb_support Logic optimization and Technology mapping
- ace2 Signal activity estimator
- blifexplorer .blif files explorer
- doc Documents: VPR user manual and power estimation manual
- libarchfpga Library to parse xml architecture files
- Makefile Makefile to build/install the tools
- ODIN_II HDL synthesizer
- pcre Regular expression library for C/C++ programs
- printhandler Generic message logging system for CAD tools
- quick_test Quick test to the tools after installation
- README.txt Readme file describing briefly how to install and run the tools
- run_quick_test.pl Perl script to run a quick test to the tools after installation
- run_reg_test.pl Perl script to run a regression
- spice Some SPICE primatives netlists
- vpr VPR tool
- vtr_flow Architectures, benchmarks, scripts to run, etc

- VTR_flow contains:
- arch FPGA architectures
- benchmarks Benchmarking circuits
- misc Misc files (currently contains a configuration xml for ODINII)
- parse Parsing configurations (pass requirements, running configs, qor)
- primitives.v Primitives Verilog descriptions (Mem, LUT, MUX, etc)
- Readme.txt Readme text (referring to VTR wiki for help)
- scripts Running and parsing perl scripts, and perl scipts of workflow
- sdc Synopsys design constraints files
- tasks Tasks (configs descriping circuit/architecture), and results
- tech Technology files (130nm, 45nm, 22nm in this release)

Hint: In this work we altered routing cost function (In <vtr_release>\vpr\SRC\route\route_timing.c) as mentioned above to consider the new multiple charge recycling technique.

الملخص

مصفوفة البوابات المنطقية القابلة للبرمجة اصبحت احدي مفاتيح التصميم الالكتروني الرقمي خلال العقود الاخيره. جزء كبير من تخطيط مصفوفة البوابات المنطقية القابلة للبرمجة يشمل قابليه البرمجة لكتل الوظائف المنطقيه والتوصيلات بينها. هذا التخطيط له تاثير كبير علي مدي جودة المنتج النهائي من حيث الاداء و مساحته و استهلاك الطاقة. هناك طرق عديدة لجعل مصفوفة البوابات المنطقية القابلة للبرمجة اكتر كفأة من حيث استهلاك الطاقة. هذه الطرق يمكن تقسيمها الي خمسه اقسام هي: عملية التصنيع, الدوائر الكهربيه, تخطيط المصفوفة, نظام المصفوفة ككل, و أدوات التصميم بمساعدة الحاسب الألي. طرق عمليه التصنيع تتضمن استخدام مواد خام و اشكال مختلفة لاجهزة الترانزستور وتكنولوجيات جديدة. طرق تصميم الدوائر الكهربيه تتضمن تكوين كتل الوظائف المنطقية والتوصيلات من الترانزستور.طرق تخطيط مصفوفة البوابات المنطقية القابلة للبرمجة تتضمن وظائف الكتل المنطقية و الذاكرة والمدخلات والمخرجات و التوصيل بين موارد المصفوفة. طرق نظام المصفوفة ككل يتضمن توفير استهلاك الطاقة علي مستوي اعلي مثل التحكم الديناميكي للفولت و اطفاء اجزاء من المصفوفة عندما تكون غير مستخدمة. واخيرا طرق التصميم بمساعدة الحاسب الألي وتتضمن التحسينات في ادوات برمجة المصفوفة لتقليل استهلاك الطاقة. في هذه الدراسة نحن نتطرق لتقليل استهلاك الطاقة في مصفوفة البوابات المنطقية القابلة للبرمجة علي مستوي جهاز الترانزستور والدوائر الكهربيه.

في البداية قمنا بدراسة استبدال الترانستور المستخدم في مصفوفة البوابات المنطقية القابلة للبرمجة بترانزستور ذو فولت حرج متغير ديناميكيا DTMOS

و اوضحنا ان هذا الاستبدال يعمل علي تقليل استهلاك الطاقة في مصفوفة البوابات المنطقية القابلة للبرمجة في كلا من كتل الوظائف المنطقية و التوصيلات.

ثم قمنا بانشاء مصفوفة البوابات المنطقية القابلة للبرمجة باستخدام جهاز FinFET بدلا من تقنية CMOS لاستكشاف تاثير ذلك على التكنولوجيات الجديدة, وايضا قمنا بتغطية تاثير التغيرات المحيطة عند هذه التكنولوجيات. وقد استخدمنا تصميمات تنبأيه للتكنولوجيا في المستقبل PTMs لاستكشاف اداء المصفوفة والقدرة و الطاقة مع التكنولوجيا عند مستوي 20nm حتي مستوي 7nm لاجزاء المصفوفة علي حده مع التغيرات الطارئة اثناء الصناعة. ايضا قدمنا مقترحا بتغذية الطاقة المثلي من ناحية الطاقة المستهلكة لمجموعة من ال flip-flops التي يمكن استخدمها في المصفوفة عند كل تكنولوجيا علي حده. وقد قمنا بتقييم اداء المصوفوفة باستخدام ادوات المحاكاة SPICE simulators اثناء برمجه المصفوفة لتعمل كاداة جمع او بوابة منطقية NAND او سلسلة منطقية cascaded chain.

واخيرا نقدم تقنية جديدة لتدوير الشحنات في وصلات مصفوفات البوابات المنطقية القابله للبرمجه, التقنية الجديدة اظهرت توفير اكتر للقدرة مقارنة بعمل سابق مشابه, وقد قمنا بتحليل التقنية الجديدة من مستوي النظرية مرورا بتفاصيل الدوائر الكهربيه المستخدمة وحتي كيفية تطبيقها باستخدام ادوات التصميم باستخدام الحاسب الألي, وتظهر الدراسة ايضا المساحة الزائدة المطلوبة و كمية الطاقة التي تم توفيرها باستخدام مجموعة من الموشرات benchmarks في اداة VPR

مهندس: أسامه أحمد محمد أحمد عبد القادر
تاريخ الميلاد: 1989\12\21
الجنسية: مصرى
تاريخ التسجيل: 2014\1\1
تاريخ المنح: ....\....\.....
القسم: الكترونيات
الدرجة: ماجستير
المشرفون:
ا.د. أحمد محمد سليمان
د. حسن مصطفي
الممتحنون:
أ.د..... (الممتحن الخارجي)
أ.د..... (الممتحن الداخلي)
أ.د أحمد محمد سليمان (المشرف الرئيسي)
د. حسن مصطفي (عضو)

عنوان الرسالة:

طرق تصميم مصفوفة بوابات منطقية قابلة للبرمجة قليلة القدرة علي مستوي الجهاز والدائرة

الكلمات الدالة:
مصفوفة البوابات المنطقية القابله للبرمجه, ترانزستور ذو جهد حرج متغير ديناميكيا, تقليل القدرة, تدوير الشحنات, ترانزستور ذو زعنفه

ملخص الرسالة:
درسنا في هذا العمل استبدال الترانزستور المستخدم في مصفوفة البوابات المنطقية القابله للبرمجه بترانزستور ذو جهد حرج متغير ديناميكيا DTMOS وقمنا بتقييم المصفوفة من حيث الاداء و القدرة والطاقة المستهلكة وقد اظهر هذا الاستبدال تحسنا في القدرة والطاقة المستهلكة في اجزاء المصفوفة وتحسنا في الطاقة المستهلكة في المصفوفة ككل. وايضا قما بدراسة استخدام جهاز FinFET لبناء المصفوفة وتقييمها مع اعتبار التغيرات الطارئة اثناء الصناعة في اجزاء مختلفة من المصفوفة. وتقدم الدراسة ايضا القيمة المثلي لجهد التغذية بالنسبة لمجموعة من القلابات flip-flops. واخيرا قدمنا مقترحا باستخدام تقنية تدوير الشحنات لتوفير طاقة المصفوفة علي مستوي الدوائر الكهربية والتي اظهرت توفير مايقرب من 32% من طاقة توصيلات المصفوفة باستخدام ادوات المحاكاة SPICE simulators وكيفية تطبيق هذا المقترح مع تقييمه باستخدام مجموعة من المؤشرات في ادوات التصميم باستخدام الحاسب الالي CAD tools.

طرق تصميم مصفوفة بوابات منطقية قابلة للبرمجة قليلة القدرة علي مستوي الجهاز والدائرة

اعداد
أسامه أحمد محمد أحمد عبد القادر

رسالة مقدمة إلى كلية الهندسة - جامعة القاهرة
كجزء من متطلبات الحصول على درجة ماجستير العلوم
في
الالكترونيات

يعتمد من لجنة الممتحنين:

الاستاذ الدكتور:	الممتحن الخارجي
الاستاذ الدكتور:	الممتحن الداخلي
الاستاذ الدكتور: أحمد محمد سليمان	المشرف الرئيسى
مدرس: حسن مصطفي	عضو

كليــة الهندســة - جامعــة القاهــرة
الجيـزة - جمهوريـة مصـر العربيــة

سبتمبر – 2016

طرق تصميم مصفوفة بوابات منطقية قابلة للبرمجة قليلة القدرة علي مستوي الجهاز والدائرة

اعداد
أسامه أحمد محمد أحمد عبد القادر

رسالة مقدمة إلى كلية الهندسة - جامعة القاهرة
كجزء من متطلبات الحصول على درجة ماجستير العلوم
في
الالكترونيات

تحت اشراف

أحمد محمد سليمان
أستاذ متفرغ / جامعة القاهرة

حسن مصطفي
مدرس/ جامعة القاهرة

كليــة الهندســة - جامعــة القاهــرة
الجيـزة - جمهوريـة مصـر العربيــة

سبتمبر – 2016

طرق تصميم مصفوفة بوابات منطقية قابلة للبرمجة قليلة القدرة علي مستوي الجهاز والدائرة

اعداد

أسامه أحمد محمد أحمد عبد القادر

رسالة مقدمة إلى كلية الهندسة - جامعة القاهرة
كجزء من متطلبات الحصول على درجة ماجستير العلوم
في
الالكترونيات

كليــة الهندســة - جامعــة القاهـرة
الجيـزة - جمهوريـة مصـر العربيــة

سبتمبر – 2016