UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Nanoélectronique et nanotechnologie**

Arrêté ministériel : 7 août 2006

Présentée par

Ogun TURKYILMAZ

Thèse dirigée par **Fabien Clermidy**

préparée au sein du **Laboratoire Intégration Silicium des Architectures Numériques**
et de l'**École Doctorale Electronique, Electrotechnique, Automatique et Traitement du Signal (EEATS)**

Emerging 3D Technologies for Efficient Implementation of FPGAs

Thèse soutenue publiquement le **28 novembre 2014**,
devant le jury composé de :

Régis Leveugle
Professeur, Univ. Grenoble Alpes, TIMA, Président
Wim Dehaene
Professeur, KU Leuven, Rapporteur
Marian Verhelst
Professeur, KU Leuven, Rapporteur
Ian O'Connor
Professeur, Ecole Centrale de Lyon, Rapporteur
Jean-Michel Portal
Professeur, Ecole Polytechnique Universitaire de Marseille, Examinateur
Olivier Lepape
NanoXplore, Examinateur
Fabien Clermidy
HDR, CEA-LETI, Directeur de thèse

Abstract

The ever increasing complexity of digital systems leads the reconfigurable architectures such as Field Programmable Gate Arrays (FPGA) to become highly demanded because of their in-field (re)programmability and low non-recurring engineering (NRE) costs. Reconfigurability is achieved with high number of point configuration memories which results in extreme application flexibility and, at the same time, significant overheads in area, performance, and power compared to Application Specific Integrated Circuits (ASIC) for the same functionality. In this thesis, we propose to design FPGAs with several 3D technologies for efficient FPGA circuits. First, we integrate resistive memory based blocks to reduce the routing wirelength and widen FPGA employability for low-power applications with non-volatile property. Among many technologies, we focus on Oxide Resistive Memory (OxRRAM) and Conductive Bridge Resistive Memory (CBRAM) devices by assessing unique properties of these technologies in circuit design. As another solution, we design a new FPGA with 3D monolithic integration (3DMI) by utilizing high-density interconnects. Starting from two layers with logic-on-memory approach, we examine various partitioning schemes with increased number of integrated active layers to reduce the routing complexity and increase logic density. Based on the obtained results, we demonstrate that multi-tier 3DMI is a strong alternative for future scaling.

Acknowledgments

This thesis has been the result of many disscussions and efforts from inspiring people. Hereby, I would like to acknowledge all of their contribution. Above all I am indebted to my thesis supervisor Fabien Clermidy. I had the most privilege to have worked with such an inspirational figure who had an admirable vision and great enthusiasm. He is a great mentor from whom I learned the most.

Even though a great effort has been spent to form this thesis, it would have no value without the approval of a great jury. First of all, I would like to express my gratitude to Prof. Ian O'Connor, Prof. Wim Dehaene, and Prof. Marian Verhelst for being the reporters and for their invaluable suggestions. I would also like to thank Prof. Régis Leveugle, Prof. Jean-Michel Portal and Olivier Lepape for accepting to be part of the thesis jury. It has been a pleasure to defend and discuss my thesis work in front of such a group of technical experts.

During this thesis research I worked in collaboration with many researchers from whom I had the chance to get first hand experience. I would like to acknowledge Perrine Batude for her expertise and enthusiasm in 3D Monolithic Integration, Maud Vinet for directing the research effort in 3D Technologies, Olivier Rozeau for his help on transistor modeling, Olivier Billoint and Sébastian Thuries for the discussions and brainstorming about 3D design, Olivier Thomas and Bastien Giraud for their help on memories, Gérald Cibrario for his rigorous work in the 3D PDK, Elisa Vianello and Marina Reyboz for the compact models of resistive memories, Haykel Ben-Jamaa for sharing his grand knowledge about nearly any technical subject as well as the Santhosh Onkaraiah, Hraziia, Natalija Jovanovic, Georgio Palma, Hossam Sarhan and Houcine Oucheikh with whom I exchanged many ideas and

technical discussions. I really appreciate their contribution which shaped my thesis research substantially.

As a part of the PhD program, I had the opportunity to spend three months in the LSI laboratory at EFPL. I would like to thank the LSI team for being very welcoming and accepting me as a fellow member. Especially, I would like to express my gratitude to Pierre-Emmanuel Gaillardon and Prof. Giovanni de Micheli for making this stay possible and, thus, contributing to my thesis research.

I feel privileged for the chance to conduct research in such a dynamic and enriching workplace as CEA where I met wonderful people. Especially, I would like thank the members of my laboratory: Marc Belleville, Jérome, Fredéric, Alexander, Yvain, Pascal, Ivan, Romain, Jean-Fredéric, Michel, Yves and especially Edith for taking the responsibility of coaching PhD students, my friends David, Meycene, Sébastien, Bartosz, Adam, Vincent Alex, Soundous, Thiago, Bilel, Guillaume, Marie-Sophie, Grégory for making the life more enjoyable. I would like to thank Catherine Bour for being such a helpful person during my administrative struggles. I also thank the laboratory secretaries Caroline and Armelle for being kind and patient even with my limited French. Last but not the least, I am grateful to my dear friends Yeter, Lionel, Onur, Eda, and Alp who were always there to help me through the times of distress.

I would like to thank Hugh Metras without whom my path would have never crossed with CEA as he is the person who invited me to the interview for this PhD position at a career fair in Boston. I would also like to take thank my previous advisor from Northeastern University Prof. Jong-bin Kim for motivating me towards this PhD and Prof. Gunar Schirner who was generous enough to share his experiences and was an inspiration to me to do research.

My most heartfelt thanks, appreciation, and gratitude go to my family. A day does not go by when I do not think about them. My father Nevzat, my mother Nuket and my sister Pinar have been the source of my motivation

and strength. Without their constant encouragement I would be able to get this far. I dedicate this dissertation to each and every one of my family for their unconditional love and generous support.

Ogun Turkyilmaz

Contents

List of Figures

LIST OF FIGURES

List of Tables

1

Introduction

Contents

1.1. The End of Scaling Era

Until now we relied on traditional scaling in order to increase performance and logic density. For more than four decades, according to Moore's law [19], the number of transistors doubled every two years. As the transistors operated faster, higher frequencies are reached. However, traditional scaling is facing fundamental limitations on manufacturing, performance, and power consumption figures.

The structure of conventional transistor imposes difficulties on manufacturing for further scaling. The complexity of lithography process grows with each advancing node. Thus, either new lithography technologies or more patterning steps are required which increase the production cost tremendously. Secondly, the tiniest feature of the transistors, the width of the gate dielectric, recently reached the size of several atoms, which imposes problems: higher dependence to the number of atoms which change due

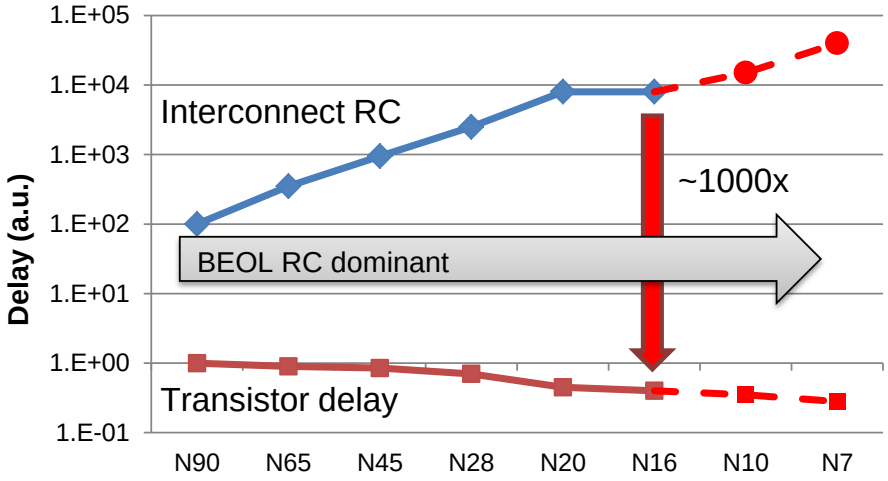


Figure 1.1: Transistor and interconnect delay scaling for future nodes. Adapted from [1].

to variations in manufacturing and lowered gate barrier which increases the leakage current.

Apart from manufacturing implications, the performance gains are shrinking due to the limits on interconnects. Even though the transistors are shrinking, the circuits cannot reach high performance without fast and dense interconnects. Unfortunately, the wires cannot be made fast and dense at the same time: wider cross-section increases the parasitic capacitance while reduced height or width increases electrical resistance. As demonstrated in Fig. 1.1, the gap between the gains of gate performance and interconnect delay widens as the technologies advance. Since the interconnect wires cannot keep up with that trend, they become one of the major bottlenecks in traditional scaling.

The inevitable growing in power consumption leaded to the end Dennard's theory [20]. Dennard stated that each advancing node should increase the performance by 40% while keeping the same power consumption. Recent trends show that scaling can only fulfill gains in either power efficiency or transistor density. The two components of power consumption leakage and dynamic are increasing significantly with each advancing technology node as depicted in Fig. 1.2. Previously, leakage consumption was assumed to be negligible, however as a result of scaling, it is becoming more significant because it reaches to the level of dynamic power. Dynamic power consumption increases as more devices switch at higher frequencies even though the supply voltage is reduced due to scaling. Furthermore, the severity of interconnects became more apparent recently

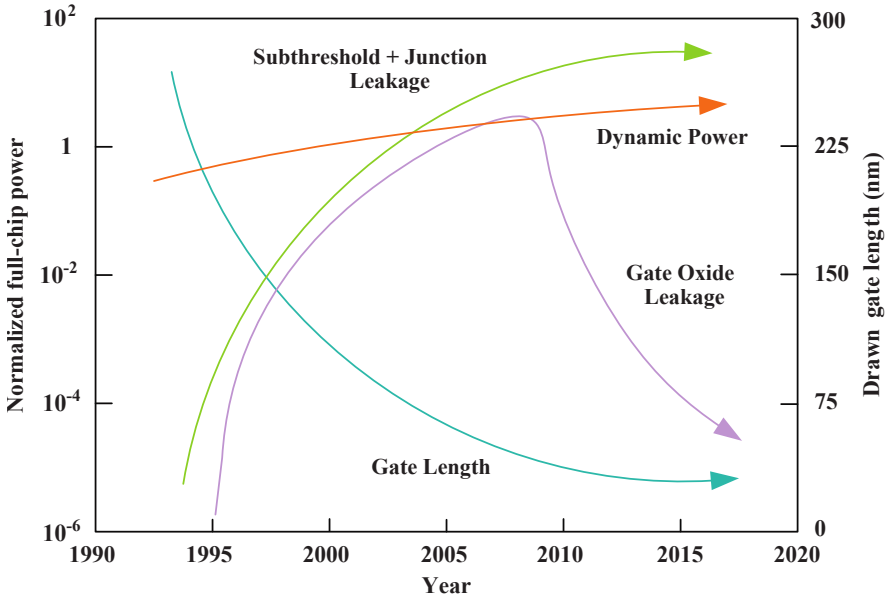


Figure 1.2: Total chip dynamic and static power dissipation trends based on ITRS2006 [2]. Gate leakage is improved significantly with High-k materials.

as they account more than 50% of the total consumption [21]. Since the wires are not shortening as expected with scaling, bigger transistor are needed to drive these wires which leads to higher power consumption. Consequently, power consumption is becoming the major bottleneck limiting the traditional scaling.

Due to the diminishing returns of scaling, we are now facing the so-called utilization wall. The designers aimed to keep the same power density while trusting scaling for increased logic density and higher frequency. However, the power density reached to highest limit that the chip could handle. Thus, acknowledging the end of frequency scaling, a new trend was born which proposed adding more cores working at lower frequencies. This allowed to add more functionality while still being lower than the power cap. Even though the frequency is lowered, more transistors are switching with each added core which push power consumption towards power cap. Namely, the utilization of transistors must be limited because only a number of transistor can be active at any given time due to the power constraint. Therefore, some of these transistors have to stay inactive which leads to the phenomena called dark silicon.

1.2. Emerging 3D Technologies

The limitations of traditional scaling are forcing us to find next scaling paradigms. Thanks to all the research efforts, we are now in reach of many possibilities. 3D technologies are receiving unprecedented attention for future circuits and they could create the aspired flexibility to address the demands. Fig. 1.3 depicts several promising 3D technologies as 3D integration and advanced memories. One of the 3D integration possibilities is with Through-Silicon Vias (TSV). TSV integration offers stacking of multiple wafers and it helps reducing the wirelength however they occupy a significant space for vertical connections. Thus, they only allow very limited number of connections between layers. In order to derive the maximum benefits from 3D integration, the technology must support a very high density of vertical interconnects with via dimensions compatible with device sizes. A new technology called 3D monolithic integration (3DMI) extends the limitations of TSVs by achieving 40x smaller inter-tier vias (Table 1.1). The unique feature of 3DMI is the sequential integration of active layers one after another on the same die. Thus, the layers are aligned with high precision which allows fabrication of inter-tier vias similar to regular vias. Apart from 3D integration, advanced memories can introduce increased functionality in the third dimension. Advanced memories bring substantial benefits to the conventional CMOS with the integration at the Back-End-of-Line (BEOL) and non-volatile operation possibilities. These memories can be fabricated between metal layers requiring no silicon area and store one bit of information. These emerging 3D technologies can address the challenges of existing technologies and be the next paradigm for future scaling.

Table 1.1: Comparison of 3D Integration interconnect size.

	Diameter(μm)	Pitch(μm)
TSV	2 - 4	4 - 8
3DMI	0.1	0.2
3DMI vs. TSV gain	20x - 40x	20x - 40x

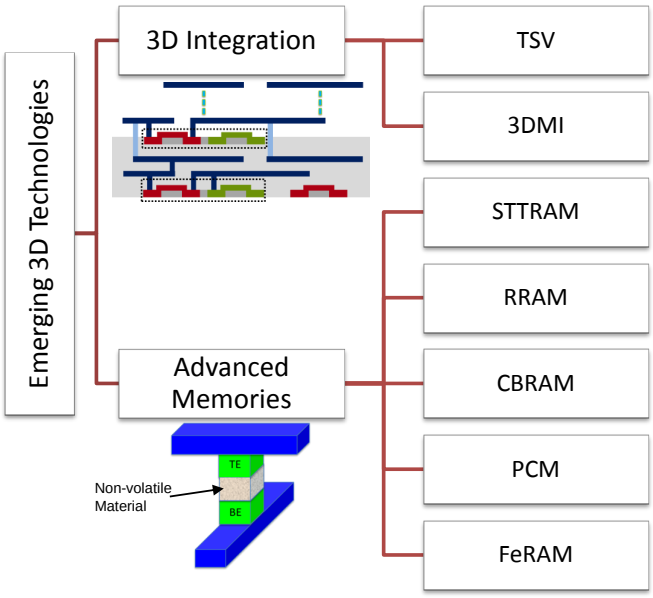


Figure 1.3: Emerging 3D technologies.

1.3. FPGA Architecture

As a computing solution Field-Programmable Gate Arrays (FPGA) are gaining increasing interest because of their easy in-field programmability, high flexibility, and reduced non-recurring engineering (NRE) costs. Compared to Application-Specific Integrated Circuits (ASIC) with very long design cycles especially considering increased design complexity due to high number of layout rules, FPGAs bring an efficient solution for managing this complexity. The advantages offered by FPGA circuits, on the other hand, are obtained by intensive utilization of configuration memory nodes and routing resource. Compared to ASICs, FPGAs use 40x more area with 4x slower performance and 12x more power consumption for the same functionality [22]. Standby power which only considers retaining the state (i.e. no clock switching) reaches two orders magnitude more than ASICs [9]. Consequently, even though FPGAs offer an innovative computing platform, the drawbacks prevent FPGAs from being utilized in low-power mobile applications.

Still FPGAs hold one of the most effective advantage over ASICs. Compared to FPGAs, ASICs have to use less expensive process nodes because the effort of moving to a more advanced node imposes high costs in terms of physical design and production.

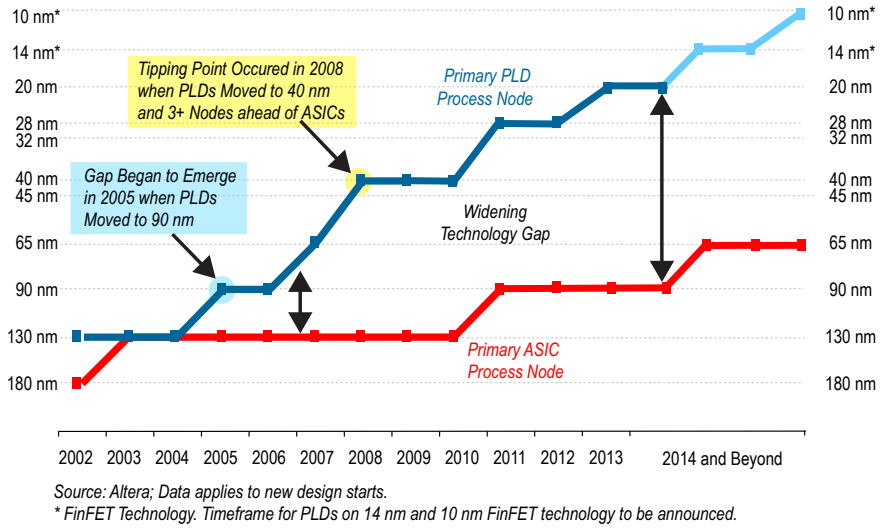


Figure 1.4: Programmable logic vs. ASIC for new designs in primary process nodes. [3]

Current FPGAs, on the other hand, have already reached 28nm soon to be on 20nm and smaller process technologies. Most new ASIC designs lag behind FPGAs with a two or three nodes as illustrated in Fig. 1.4 [3]. This trend shows that there are extreme demands on FPGAs and more aggressive solutions are needed. It is, therefore, safe to say FPGAs will be the first platform to take benefit from emerging technologies.

1.4. Research Contributions

In this thesis, the application of 3D technologies for efficient implementations in FPGAs are analyzed. Among different technologies, 3D monolithic integration and non-volatile memories are evaluated. The contributions of this work constitutes development of exploration framework and effective design solutions with emerging 3D technologies.

1.4.1. FPGA Experimental Evaluation Platform

The integration of emerging technologies to FPGAs cannot be accomplished without an evaluation platform. Even though the FPGA fabric is generic and simple, overall FPGA area, performance, and power metrics depend on the application that is mapped onto this fabric. Furthermore, the technological properties cannot be imposed by simple transistor parameters. Therefore, first an evaluation framework is established by a set of tools which takes an architecture definition of the FPGA fabric including

technological impacts and maps benchmarking circuits onto the fabric to obtain area, performance, and power consumption values. A methodology is proposed for accurate architectural and technological modeling. In this methodology, post layout parameters are extracted and the blocks are characterized in terms of area, delay and power. With these values an architectural definition is created which allows FPGA evaluation platform (Versatile Place and Route-VPR-based) to be used for fast evaluation of design level improvements on real Microelectronics Center of North Carolina (MCNC) benchmark circuits. This developed platform forms the basis for all explorations with targeted emerging technologies in this work.

1.4.2. RRAM-based NVFPGA

In this thesis, we focus on FPGA evaluation using Non-Volatile Memory (NVM)-based blocks with Oxide Random Access Memory (OxRAM) and Conductive Bridge RAM (CBRAM) technologies. Main contributions include the improvement assessment of design with each technology. Furthermore, new low-power application fields are targeted taking advantage of the Resistive RAM (RRAM) non-volatility.

First, novel RRAM-based architectures are characterized in terms of area and power consumption. With OxRAM, Non-Volatile Static Random Access Memory (NVSRAM) and Non-Volatile flip-flop (NVFF) circuits and with CBRAM, a voltage-divider based Non-Volatile memory Element (NVE) are analyzed through layout and power simulations. The characterization results for each RRAM-based block are reported.

In the next step, RRAM-based modules are integrated in the FPGA by replacing each corresponding module with the non-volatile counterpart. FPGA architecture models are created based on the characterized blocks. When all the Static Random Access Memories (SRAM) are replaced with NVSRAM or NVE, configuration-saving applications and when all FFs are replaced with NVFF, context-saving applications are targeted. Consequently, Non-Volatile FPGA (NVFPGA) operation is achieved and the FPGA employability is extended with new set of applications utilizing non-volatility.

Finally, a low power computing system is proposed using the designed NVFPGA. Since all volatile elements are replaced with their non-volatile counterparts, NVFPGA can be switched off when not in use to save the static power while still keeping the configuration information. For this functionality, power gating is implemented and system level impacts are reported. The designed system is suitable for Normally-off

Instantly-on applications which do intensive computation for a short amount of time and sleep during rest of the time. By carefully analyzing the applications and the impacts of the designed system, an activity ratio is defined for overall power savings.

1.4.3. 3D-FPGA with Monolithic Integration

In this thesis, several 3D-FPGA circuits with monolithic integration are designed and evaluated. This work constitutes contributions in design and partitioning aspects with 3DMI.

In order to get the highest benefits from this novel technology, full-custom circuits for 3D-FPGA are designed. Layouts of the corresponding blocks are created using 14nm 3D LETI-FDSOI PDK considering logic-on-memory approach. All the blocks are characterized with post-parasitic simulations for area, delay, and power consumption values. The blocks are then integrated to create a 3D-FPGA. The 3D-FPGA is compared to the 2D counterpart and the improvements are reported.

The designed 3D blocks establish a basis for further partitioning exploration. Several partitioning schemes are investigated considering implementations having more than 2 active layers up to 4 layers. Architecture definitions for each of the design are developed and using the evaluation platform, area and EDP gains are estimated. The analysis of the results shows that 3DMI provides higher gains than traditional scaling expectations.

1.5. Thesis Summary

The thesis is briefly explained as follows:

In chapter 2, the context of this research work is outlined. The first section, 2.1.1., starts with a summary of recent FPGA history. The next section, 2.1.2., details the FPGA architectures. The baseline FPGA architecture is formed with logic blocks and routing resources. In the next section, 2.2., the limitations of current FPGAs are identified. The limitations are categorized as configuration memory and routing resource. Emerging technologies are introduced as advanced memories and 3D integration in Section 2.3.. State-of-the-art implementations are compared and related works on integration of emerging technologies on FPGAs are presented in Section 2.3.3..

In chapter 3, the FPGA exploration framework and methodology for emerging technology adoption is presented. The chapter starts with an overview the FPGA CAD flow in Section 3.1.. It explains the properties of available open-source FPGA tools. In Section 3.2., the details of the FPGA exploration framework are discussed. Architecture definition, area, performance, and power modeling, and MCNC benchmark circuits are explained in the sections between 3.2.1. - 3.2.3.. The final section, 3.3., presents a methodology for fast adoption of emerging technologies through modifications for FPGA architecture definition.

In chapter 4, we target OxRAM and CBRAM technologies for efficient FPGA design. First, the motivation towards RRAM-based FPGAs is explained. In section 4.1., the circuits NVSRAM, NVFF, and NVE are characterized in terms of area, performance and power consumption including operating principle of each circuit. In the section 4.2., new applications fields are proposed as configuration-saving and context-saving. The modification needed for NVFPGA with the analyzed circuits using OxRAM and CBRAM technologies are explained. In section 4.3., a new computing scheme with low-power property is discussed. In order to meet the low-power goal, power-gating opportunities and the associated cost are explored in section 4.3.1.. System level implications of NVM-based low-power FPGA are determined for each technology in the final section, 4.3.2..

In chapter 5, 3D-FPGA with monolithic integration is proposed. 3D Monolithic Integration technology is detailed in section 5.1.. The Logic-on-Memory approach is explained in section 5.2.. The designed 3D blocks are included in section 5.2.2. and the performance of the blocks are compared to the 2D in section 5.2.3.. Using the VPR flow, the designed blocks are evaluated in FPGA. In section 5.2.4., 2D and 3D FPGAs are compared in terms of area, delay and power consumption. Section 5.3. discusses a multi-tier 3D FPGA approach. Starting from two up to four layers are evaluated in section 5.3.4.. Multi-tier FPGA results are compared to traditional scaling expectation in Section 5.4..

In chapter 6, all the results obtained from this work are summarized for global comparison. The contributions of the research work are highlighted and possible future extensions are presented.

2

Background and Motivation

Contents

FPGAs are under rapid growth and they are becoming one of most popular digital circuit implementation solution. FPGAs are reconfigurable devices which enable the use of the same silicon for different applications. They can fulfill high performance requirements and at the same time, lower the NRE costs and time-to-market. The programmable nature of FPGA make them very flexible while making them larger, slower, and more power consuming than ASICs. Thus, the FPGA adoption in embedded applications is very limited.

The advancements in process technology has lead to improvements of FPGA efficiency. As explained in the previous chapters, FPGAs gained benefits from advanced technologies faster than ASICs. It is observed that FPGAs lead ASICs with at least 1

process node ahead. In this perspective, emerging technologies will appear first in an FPGA integration. For the highest benefits, these technologies must be fully exploited. Especially, advanced memories with nonvolatile property and 3D integration are very promising for increased logic density, functionality, performance, and power efficiency in the FPGAs.

In this chapter, first an overview of reconfigurable devices and FPGAs is provided. The FPGA architecture is explained in detailed. In the following section, existing FPGA programming technologies are expressed. Based on the present FPGAs, the limitations due to configuration memory and routing resource are specified. Emerging 3D technologies, advanced memories and 3D integration are explained in detail. Related works of FPGA-based solutions with these technologies are discussed. The chapter is then concluded with insights for the thesis work.

2.1. FPGA Background

In computing systems, many different processing elements exist depending on programmability, cost, performance, and power consumption requirements. They span from General-Purpose Processors (GPP), reconfigurable systems, Application-Specific Integrated Circuits (ASIC) up to full-custom integrated circuits. Reconfigurable systems are becoming widely adopted in system design where Field Programmable Gate Arrays (FPGAs) find interest as a reconfigurable device. Several reconfigurable architectures have been presented in the literature [23].

Among other computing solutions, FPGAs are considered as a bridge between GPPs and ASICs. FPGAs are able to cover an extremely wide range of applications with varying trade-offs and efficiencies. The point multiplication of elliptic curve cryptography algorithm implementation example shows that FPGAs can compute the result 540x faster while operating at a clock frequency 40x slower than a GPP [24]. Moreover, according to Stitt et al., when critical software loops are moved to reconfigurable hardware, energy savings up to 70% can be achieved [25]. On the other hand, ASICs perform better than FPGAs in terms of area, speed and power. Kuon et al. quantify the difference between FPGA and ASIC designs [22]. They show that purely Look-Up Table(LUT)-based mapping requires 35x larger area with 4x higher delay and 14x more power consumption. It is also noted that with extensive usage of hard-implemented

blocks (ex. multipliers, accumulators, block memories) the area difference can narrow below 5x.

FPGAs bring the possibility to use the same circuit for many different applications thanks to reconfigurability. This eliminates the long design cycle required for ASIC development. Thus, FPGAs significantly reduce time-to-market and Non-Recurring Engineering (NRE) costs. Design of an ASIC can only be justified for applications requiring products in large quantities such as mobile market. For low-to-medium volume applications FPGAs are often preferred. Moreover, FPGAs always lead ASICs with at least 1 and often 3-4 technology nodes ensuring circuits with the most advanced technologies to reduce the performance gap compared to ASICs. Whereas the same technology used for FPGAs becomes highly costly for ASIC manufacturing. Consequently, FPGAs provide a viable solution for digital circuit applications with a trade-off between flexibility, performance and cost.

2.1.1. Recent History of FPGAs

The ability to configure the logic functionality of a circuit after fabrication, started to appear in the 1970s with read-only memory-based arrays called Programmable ROMs (PROM). In these one time programmable devices, the address lines serve as logic inputs and data lines as output. Thus, N address inputs can implement any N-input functions. PROMs became area inefficient because the required area increases exponentially with the number of inputs (N). Thus, Programmable Logic Arrays (PLA) are introduced specifically for logic functions. A two-level AND-OR plane including a programmable wired-AND plane, followed by a programmable-OR plane, can closely match the structure of common logic function. In a PLA, any input can be ANDed together in the AND plane corresponding to the product terms and OR plane can be configured to produce the sum of any AND plane output. Even though PLAs provided sufficient flexibility, the programmable planes were difficult to manufacture and the performance suffered from high propagation delays. Programmable-Array Logic (PAL) was introduced as a solution to address these problems. As opposed to the two programmable planes in PLAs, PALs feature a single programmable AND-plane followed by fixed-OR gates. Many variants of these circuits are developed for increased functional capabilities including sequential logic circuits with flip-flops connected after the fixed-OR gate outputs.

PAL-like structures continued to be seen in the market with higher capacities. As the area of the devices grows quickly with increasing number of inputs, manufacturing became difficult for traditional PAL-like architecture. Complex Programmable Logic Devices (CPLD) provided a solution with programmable interconnects. More sophisticated devices for digital hardware became apparent as the difficulty to integrate more devices increased significantly. Mask Programmable Gate Arrays (MPGAs) achieved the highest capacity general purpose logic chips. The wire routing is carried out according to user demands where the transistors are prefabricated. Although they are not user programmable, they created the motivation for Field Programmable Gate Arrays (FPGA).

Wahlstrom introduced the first Static Random Access Memory (SRAM)-based FPGA in 1967 [26]. By loading a bitstream to the SRAM cells, both the logic functionality and the interconnect configuration can be altered. Although a programmable architecture with the most flexibility is established, a large portion of the device has to be allocated for the SRAM configuration cells compared to ROM implementations. Thus, the development of a commercial product had to be postponed until the transistor cost became affordable. Today's classical FPGA was first announced by Xilinx in 1986 [27]. The device capacity and complexity had grown tremendously ever since.

2.1.2. FPGA Architecture and Hardware Structures

The conventional FPGA design consists of programmable logic blocks to implement logic functions, programmable routing to create interconnects between these functions and I/O blocks to make off-chip connections. The generalized view in Fig. 2.1 shows that the logic blocks are arranged in a two-dimensional array and are interconnected by routing resource. I/O blocks are placed at the periphery to connect external devices. FPGA architecture corresponds to tile-based implementation and by replicating the tile, FPGA fabric can extended. A tile is composed of a Logic Block (LB) and the components of Routing Resource (RR) which are Switch Box (SB) and Connection Box (CB).

In this section, an overview of the FPGA components are presented. The architectures of logic block and the routing resource are discussed including the implementations from the major FPGA vendors. The section ends with FPGA configuration techniques.

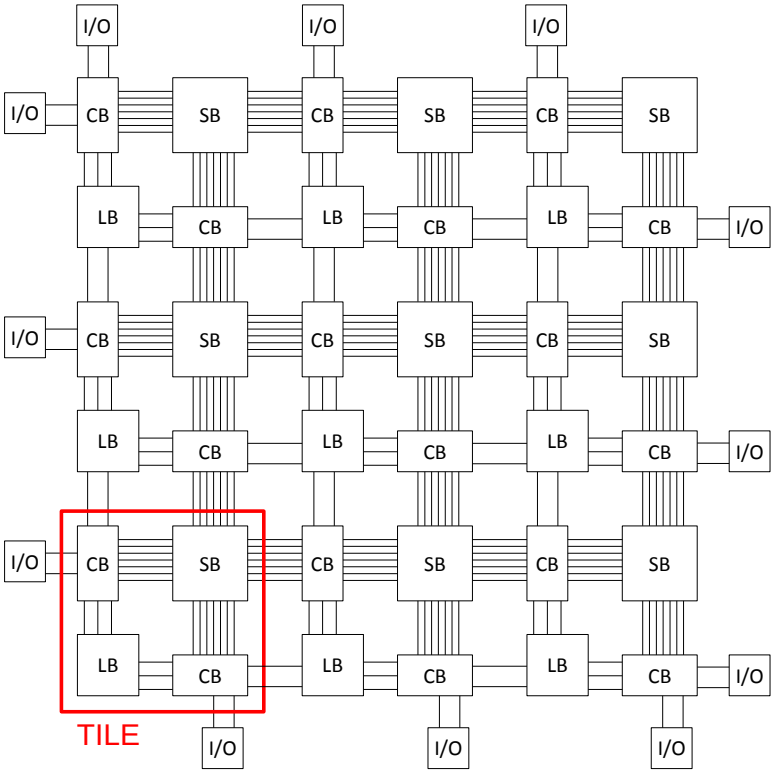


Figure 2.1: Island-style FPGA architecture.

2.1.2.1. Logic Blocks

LBs provide the basic computation capability and storage cells that are used in digital systems. The LB architecture can be built using various granularities. On one end, the logic block can be built using simple transistor pairs proposed by Crosspoint as in Fig. 2.2 [4]. With this very fine-grained approach, routing resources occupy very large area resulting in reduced area efficiency, low performance, and high power consumption. Conversely, an entire processor can build the LB. With this approach, the flexibility of the customizable hardware suffers and implementing smaller circuits results in area inefficiency. Several LB designs have been proposed which fit in between these extreme ends of granularity. In the literature, LBs built with transistors [4], NAND gates [28], multiplexers [29], Look-Up Tables (LUT) [27] and PAL-like planes [30] can be found. These architectures affect three key parameters of FPGAs: total area, critical path delay, and total power consumption.

Most commonly used LB is a LUT-based design. Basically, as in Fig. 2.3 the LB

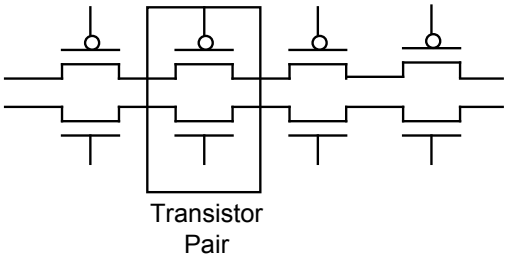


Figure 2.2: Transistor pair tiles in Crosspoint FPGA [4].

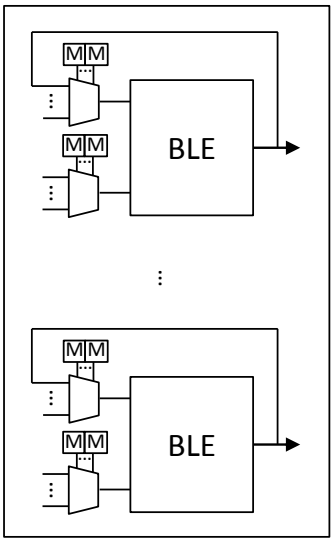


Figure 2.3: Logic Block (LB).

consists of Basic Logic Elements (BLE) and multiplexers (MUX) to choose the inputs to connect to the BLEs. A BLE (in Fig. 2.4) is composed of a LUT, a Flip-Flop (FF), and a 2-input multiplexer (MUX) to choose either the combinational output from LUT or the sequential output from FF. The LUT is a memory-based unit executes a logic function with several inputs and one output. The truth table of the function is stored in the SRAM cells and depending on the input, a value is selected using the MUX. Typically 4-input LUTs (LUT4) are integrated which require 2^4 SRAM cells as in Fig. 2.5, i.e. N-input LUT (LUTN) is built with 2^N SRAM cells.

In the recent FPGAs, the capabilities of the LB has increased significantly. The slice of a Xilinx Virtex-7 FPGA [5] is demonstrated in Fig. 2.6. Two interconnected slices build a CLB. The slice constitutes four 6-input LUTs as function generators, eight FFs as storage elements, and dedicated carry logic for fast add operations. The LUTs

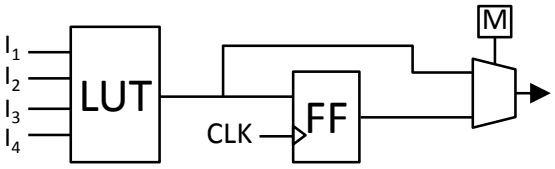


Figure 2.4: Basic Logic Element (BLE).

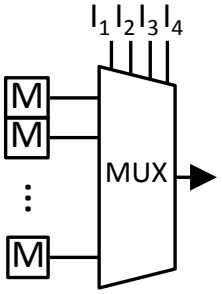


Figure 2.5: 4-input LUT (LUT4).

can be configured as one 6-input LUT or two 5-input LUTs as long as some inputs are shared among the two LUTs. Functions up to 8 inputs can be implemented using multiplexers to combine the output of two LUTs.

Altera uses the logic blocks called Logic Array Blocks (LABs). Several Adaptive Logic Modules (ALMs) are interconnected in the LAB. Fig. 2.7 depicts the Stratix-V ALMs [6] which has a variety of LUT-based resources that can be divided between two combinational Adaptive LUTs (ALUTs) and four registers. An ALM can implement various combinations of two functions up to eight inputs when some of the inputs are shared. An ALM also includes two dedicated full adders and the interconnections necessary to create a carry chain with the neighboring ALMs.

Conventionally, FPGAs were built with LUT4s because the LBs with LUT4 gives the best area-delay product [31]. Increasing the LUT size decreases the delay because of the reduced global routing complexity but results in increased overall FPGA area. Latest commercial FPGAs, on the other hand, employ larger LUTs. In these FPGAs, the inherent area cost of increasing the LUT size is reduced by the utilization of fracturable LUTs [32][33]. These LUTs can be configured into smaller sizes and packed together efficiently with a combination of shared inputs. As a result, the input muxing cost is greatly reduced and the utilization rate of LBs is improved. When the LUT4s are replaced with fracturable LUT6, Altera [32] claims 15% performance improvement

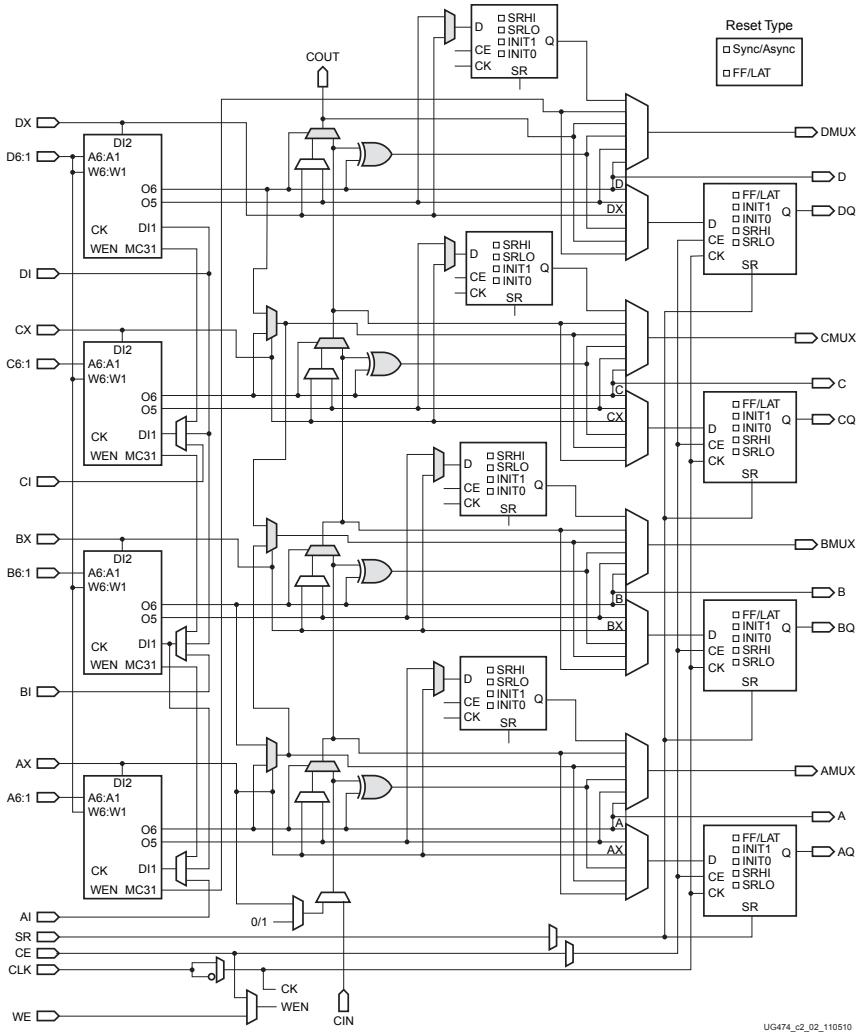


Figure 2.6: Xilinx Virtex-7 slice diagram [5].

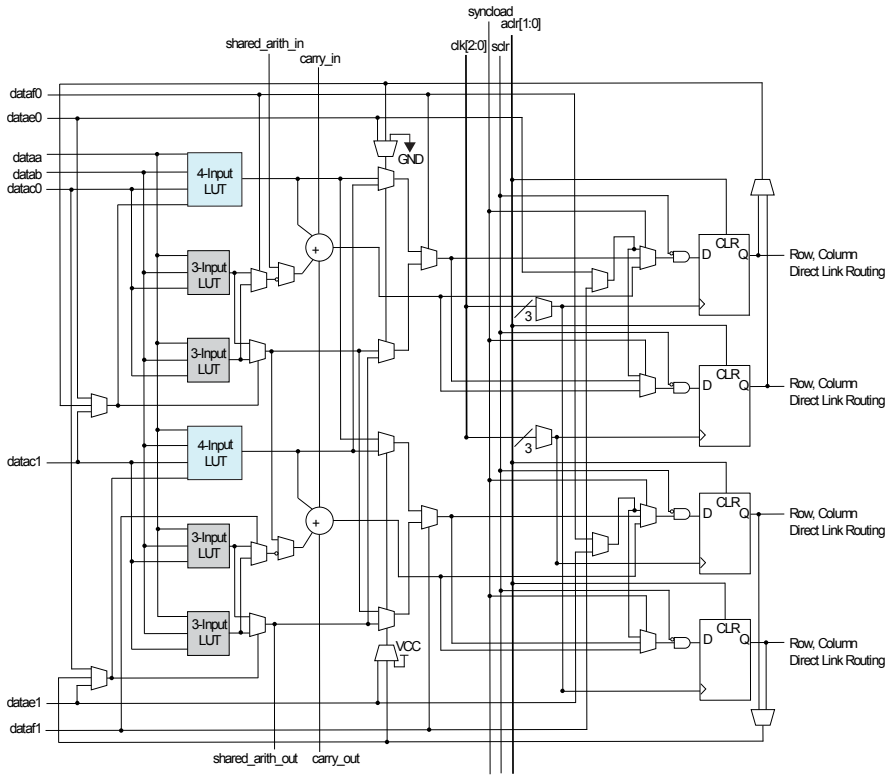


Figure 2.7: Altera Stratix-V ALM diagram [6].

with a 12% area overhead and Xilinx [33] shows 15-20% decreased power consumption.

2.1.2.2. Routing Resource

Programmable routing in the FPGA achieves connections between LBs and I/O blocks. Programmable switches containing SRAM cells connect wires and creates the desired routing. Some FPGA applications are calculation dominated requiring more local connections and other types are I/O dominated requiring long wires to the I/O blocks. The routing resource must be flexible enough to accommodate the application while fulfilling the design constraints. Additionally, in the design, there may be signals like clock and reset which require dedicated interconnect network apart from the routing resource.

Most commonly, FPGAs are designed with island-style routing where the evenly distributed routing resources surround the LB placed in two dimensional mesh structure. LBs have connections to the routing resource on all four sides through wires which form the channel. The detailed routing structure is shown in Fig. 2.8. The channel width is fixed during fabrication. In Fig. 2.8 the channel width, w, is fixed as 4.

In the routing fabric, a variety of segment lengths are included. Depending on the wire length, the segment can make a connection to all the LBs if it is a short wire or it can span more than one LB if it is a long wire. A long wire can be extended to the width of FPGA. Including various segment length, increases the area efficiency. In Fig. 2.9, a routing architecture example includes length 1, 2, and 4 tracks.

A LB input is connected to channel wire segments through the input Connection Block (CB) switches and LB output is connected to the channel through the output CB. A number of channel wire can connect to the LB input defining the input connection block flexibility, $F_{c,in}$ and the number of wires connected to the LB output defines the output connection block flexibility, $F_{c,out}$. In Fig. 2.8, only 2 tracks out of 4 are connected to the LB input, meaning that the $F_{c,in} = 2/4$ and the LB output is only connected to 1 track, making $F_{c,out} = 1/4$.

A Switch Block (SB) creates the connections of wire segments at the channel intersections where the horizontal and vertical channels meet. The number of possible connections between adjacent segments define the switch block flexibility, F_s. In Fig. 2.8, the switch box can make 3 possible connections making $F_s = 3$. Depending on the

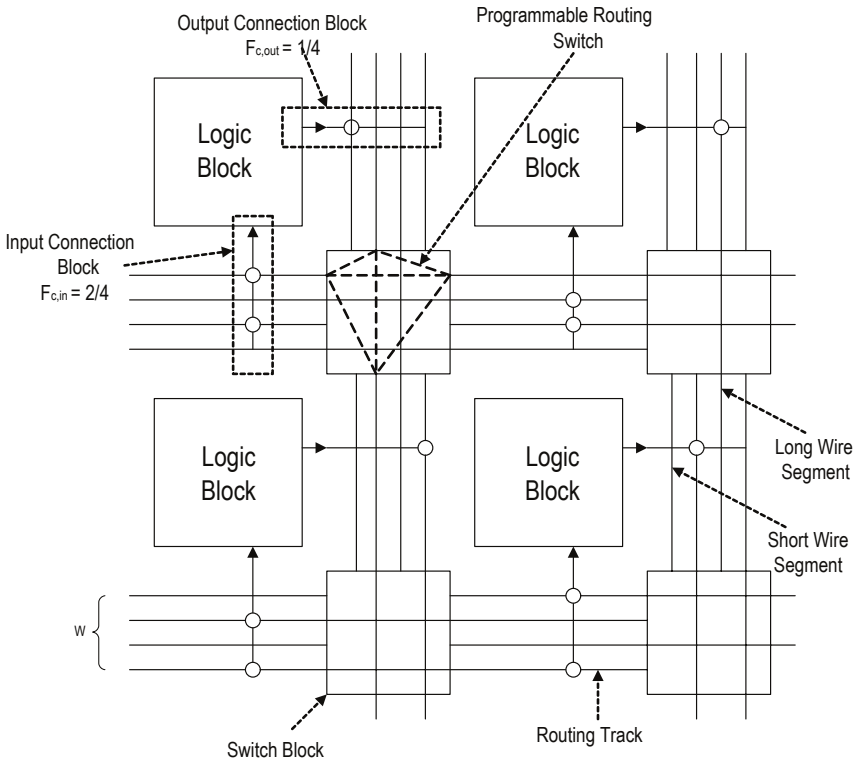


Figure 2.8: Island-style FPGA detailed routing architecture [7].

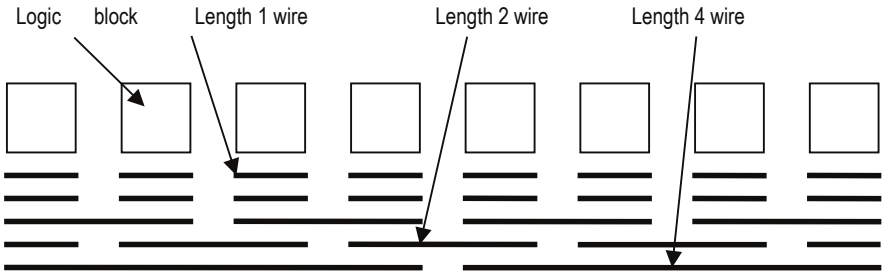


Figure 2.9: Channel segment distribution.

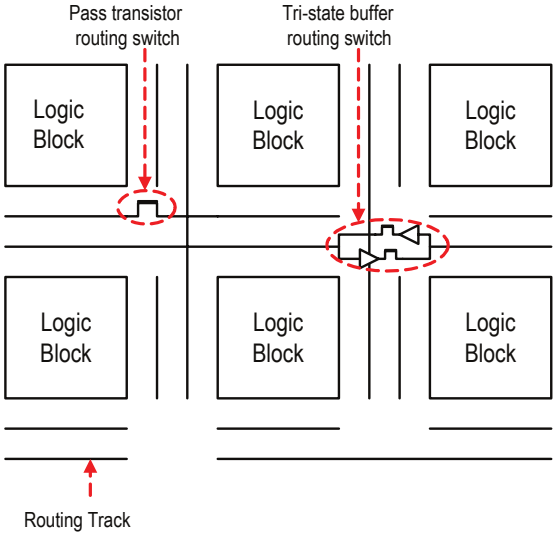


Figure 2.10: Bidirectional routing switches.

connection styles different switch types have been proposed [34]. Universal, disjoint, and Wilton have been the most popular types.

Apart from the connection pattern, the type of the switches that are used in the SB affects the routing architecture performance. Many FPGA architectures have been developed using pass transistors and tri-state buffers as routing switches [35]. These switches support bidirectional wire segments and each segment can be driven by multiple switches. Fig. 2.10 illustrates the bidirectional routing architecture. The use of bidirectional segments can leave up to 50% of the total switches unused because only one tri-state buffer can be programmed for each segment. A unidirectional approach is now widely used instead, which contains wire segments driven in a single direction. As depicted in Fig. 2.11, multiplexers are used to connect the tracks. Memory cells are connected to the select signals of the multiplexers as shown in Fig. 2.12. The unidirectional approach while reducing the flexibility of the individual routing segments, was found to be advantageous for both area and performance reasons. The area reduction is achieved mainly due to the increased utilization rate. Delay is improved because less buffers are utilized and the driving strength of the multiplexers are higher than tri-state buffers.

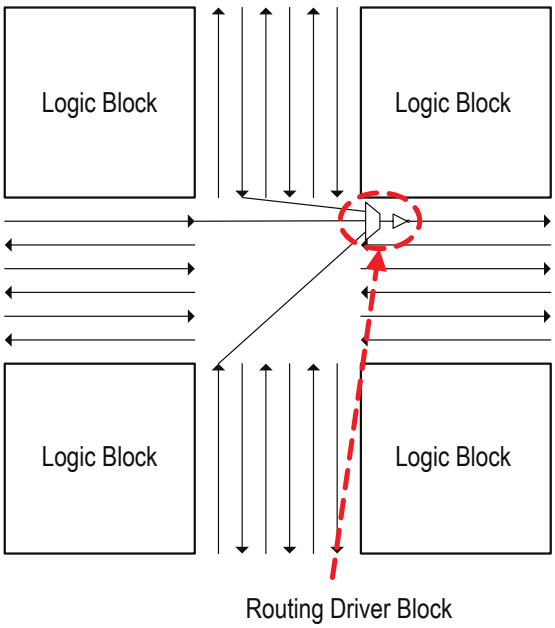


Figure 2.11: Unidirectional routing switches.

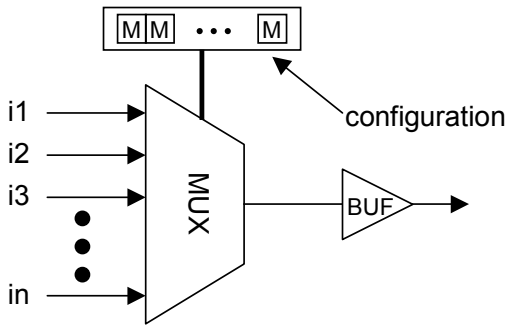


Figure 2.12: Unidirectional MUX routing switch.

2.1.3. FPGA Configuration Techniques

One of the main defining features of FPGAs is its ability to be configured according to the application needs by the end user. In this section, we will review available approaches and technologies to FPGA programmability.

Each configurable element in an FPGA includes 1 bit of storage. The programmable locations maintain the contents of logic block and the connectivity of the routing fabric. Configuration of the FPGA is achieved by programming the storage bits connected to these programmable locations. For the LUTs, the memories are filled with 1s and 0s based on the desired function. For the routing fabric, programming connects and disconnects the switches along the wires.

There are several configuration techniques in the literature with significant effects on the programmable architecture. Most popular approaches for FPGA programmability are discussed as follows.

2.1.3.1. SRAM

The most widely accepted method for storing the configuration information by the FPGA vendors is the SRAM based programming technology. SRAM-based solution has received attention because it provides re-programmability with fast configuration in a well-known standard CMOS process technology. In these FPGAs, static memory cells are distributed throughout the FPGA to provide configurability.

Traditionally 6T are used for the SRAM functionality. However, in FPGAs the configuration is constantly connected to the succeeding programmed transistor. Thus, one of the access transistors becomes obsolete which leads to the adoption of 5T SRAM cells in FPGAs. The fig. 2.13 shows a 5T-SRAM cell.

SRAM cells are inherently volatile. Since the configuration information is lost at each power down, the SRAM cells have to be configured at each start up. Hence, external devices are required to permanently store the configuration data. At the power on, the configuration has to be transferred to the FPGA which requires very long cycles and consumes a significant amount of energy. Thus, these external devices not only adds extra cost and area overhead to FPGA but also requires long boot-up time and high power.

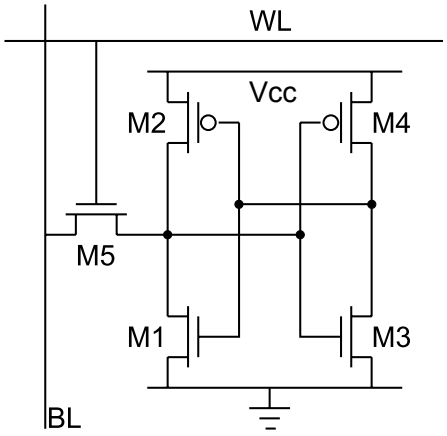


Figure 2.13: 5T-SRAM cell for configuration node in FPGA.

2.1.3.2. Flash

Instead of the SRAM-based configuration nodes, flash or EEPROM-based programming technology can be integrated in the FPGA. This solution offers non-volatile operation. Thus, the FPGA remains configured with user-defined logic when the power is cut off and it does not require extra storage or devices. Further, Flash memory cells can be realized with fewer transistors compared to SRAM cell. However, flash solution is facing several drawbacks: 1)additional masks due to non-standard CMOS process and complex CMOS co-integration. 2) higher voltage requirement for write operation and slow access time. 3) finite number of writes and scaling limitations.

2.1.3.3. Antifuse

A third approach is the use of antifuse for programmability. In this technology, a metal-based link is integrated which is normally open when unconnected and when a high current is applied, the link is melted to form an electrical connection. The main advantage stems from the low area and zero static power consumption as no transistors are required for the connection. Also, the capacitance and the resistance is significantly reduced which results in very low propagation delays. Since the link has undergone a physical transformation, it offers a non-volatile operation. However, because of this transformation, the links cannot be reprogrammed. Thus, FPGAs based on this technology are considered One-Time Programmable (OTP). Furthermore, the

fabrication requires non-standard CMOS process. As a result, these antifuse FPGAs have limited exposure in the market.

2.2. FPGA Limitations

As explained previously SRAM-based FPGAs are considered as the main market driver and therefore, in this section, the limitations for these FPGAs are discussed in detail.

2.2.1. FPGA Limitations due to Configuration Memory

One of the main feature of FPGAs is the reconfigurability and memory cells are used to grant this feature. Lin et al. quantified the memory overhead of the FPGAs based on a Xilinx Virtex-4 FPGA [8]. In Fig. 2.14, authors show that the memories occupy nearly half of the FPGA area. This area overhead makes wirelenghts longer than that of ASICs resulting in higher capacitive loads. Thus, it affects not only the area but also the delay and dynamic power consumption of the FPGA circuit. Another major limitation of the configuration memory is the contribution to the leakage power consumption. Tuan et. al. examined the leakage consumption in Xilinx Spartan-3 FPGA designed in 90nm node [9]. The breakdown in Fig. 2.15a of leakage power based on the blocks in Spartan 3 shows that the configuration memory has a major effect on the overall FPGA leakage consumption by 38%.

Even though the number of SRAM cells are reduced with architectural improvements such as gate-based LBs [36] and hard-wired SBs [37], for reconfigurability feature, the memory cells will always be a significant part of FPGA area. Additionally, since these memory cells are used for node configuration, they are distributed on the entire FPGA making it very difficult to use strict SRAM design rules in large SRAM macros. Moreover, the volatile nature of SRAMs highly limit power management possibilities. Consequently, SRAM configuration memory creates a large overhead in total area, performance and power consumption.

2.2.2. FPGA Limitations due to Routing Resource

The dominance of interconnect in FPGA is majorly due to the routing flexibility in order to increase routability of the targeted applications. The routing resource area

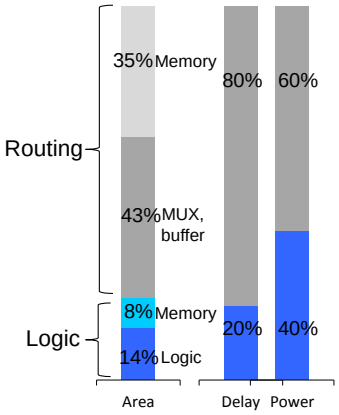


Figure 2.14: Area, delay and power breakdown of different components in Xilinx Virtex-4 [8].

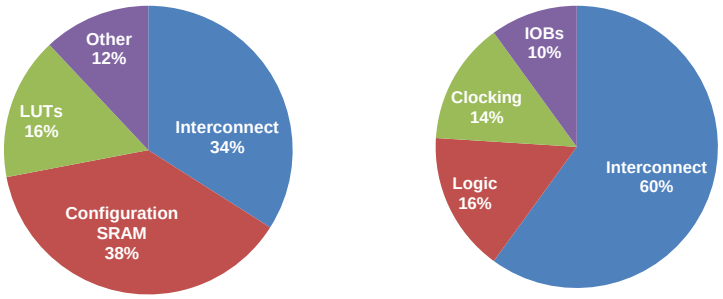


Figure 2.15: (a)Detailed breakdown of leakage power consumption of Xilinx Spartan-3 [9]. (b)Detailed breakdown of dynamic power consumption of Xilinx Virtex-II [10].

correspond to more than 75% of the total area which is composed of pre-fabricated wire segments. Each wire segment includes used and unused switches which contribute to the switching capacitance in logic transition. Such periphery does not exist in ASICs. Thus, there is a direct effect of the routing resource on the delay and dynamic power consumption of the FPGA. The Fig. 2.14 shows that routing resource constitutes 80% of critical path delay. The work of Shang et al. [10] demonstrates that in FPGAs, 60% of total power consumption are consumed by the interconnect in Fig. 2.15b which differs from that of ASICs where clock networks are usually the major contributor of dynamic power consumption [38]. Moreover, the switches contribute to the leakage consumption significantly, following after the configuration memory, by 34% as illustrated in Fig.

2.15a. Therefore, the routing resource, while bringing the benefits of flexibility of FPGAs, effects the area, delay, and power consumption of FPGA heavily.

2.3. Emerging Technologies

In this section, several emerging technologies are presented. The main focus in on advanced memories and 3D integration. The advantages and disadvantages are discussed including recent applications.

2.3.1. Advanced Memories

2.3.1.1. Spin-Torque Transfer RAM (STT-RAM)

Although magnetic memories have existed in a large variety since the earliest magnetoresistance based memories with ferromagnetic films [39][40], the development of Magnetic Tunnel Junction (MTJ) material with high Tunneling Magneto-Resistance (TMR) offered efficient memory integration [41][42]. Typically, an MTJ includes an oxide barrier (e.g. MgO) between two ferromagnetic (FM) layers. The relative directions of magnetization in FMs determine the MTJ resistance. When the layers are magnetized in different directions, i.e. polarization in anti-parallel, the MTJ has high resistance and when the layers magnetized in the same directions, i.e. polarization in parallel, the MTJ has low resistance. During the write operation, a bit of information is stored using two perpendicular lines to create the necessary magnetization at the crosspoint. A large current is required for this operation resulting in large access transistor to guarantee enough driving current. Thermally Assisted Switching (TAS) method is proposed in order to reduce the write current [43]. TAS-MRAM increases the selectivity, scalability and thermal stability [44]. Recently, spin-torque transfer random access memories (STT-RAM) took attention in the literature due to the advantages it provides over traditional MRAM devices [45]. The schematic of the memory cell is presented in Fig. 2.16. Rather than an indirect current to generate a magnetic field as in traditional MRAMs, STT-RAM uses a spin-polarized current through the MTJ to achieve switching. A positive current flow on the bottom terminal programs the memory in parallel state (low resistance) and a negative current flow leads to anti-parallel state (high resistance). Hence, write operation is established with smaller current which

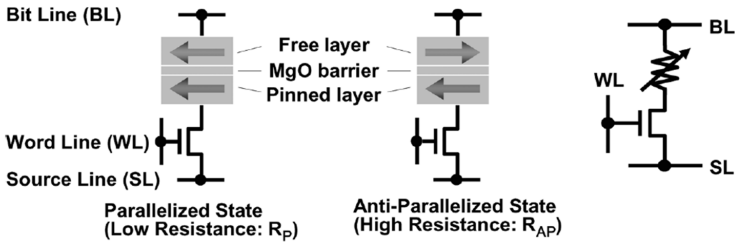


Figure 2.16: STT-RAM memory cell with 1T/1MTJ structure. Data is stored in the form of magnetization in the MTJ. a) The parallelized state exhibits low resistance to represent the logic 0. b) Anti-parallelized state exhibits high-resistance to represents the logic 1. c) Memory cell circuit connection between bit line (BL), source line (SL) and word line (WL).

relaxes the large area requirement of the access transistor. Thus, STT-RAM brings an added benefit of reduced area due to smaller access transistor and simpler design.

STT-RAM is one of the preferred MRAM technology and several circuits have already been developed as DRAM replacements. Sony demonstrated 4Kb Spin-RAM in $0.18\mu m$ with 10^{12} endurance [46]. Hitachi fabricated a 2Mb [47] and 32Mb SPRAM circuits [48]. Finally, Toshiba released 64Mb memory [49]. Beyond storage, STT-RAM found interest in logic applications. Several Non-volatile Flip-Flop (NVFF) circuits are analyzed in [50]. A full-adder with STT-RAMs is designed in [51]. Recently, NEC presented a non-volatile microcontroller in 90nm [52].

In summary MRAM technology is one of the candidates for non-volatile memory applications. Estimated STT-MRAM endurance reaches 10^{15} as in [53] however highest reported value is 4×10^{12} [54]. As scaling continues further in sub 50nm regime, the power density is expected to increase which puts constraint on the write energy [55]. The reported resistive states are low with R_{low} : $2K\Omega$ and R_{high} : $4K\Omega$ [56]. Moreover as stated in [57], the performance of STT-RAM is very susceptible to ambient temperature and, thus, it brings challenges for high performance applications.

2.3.1.2. Phase-Change Memory (PCRAM)

A Phase-Change Memory (PCRAM) cell stores the bit information using the large resistance difference between crystalline (low resistance state) and amorphous (high-resistance state). The cell is composed of thin-film chalcogenide material (typically

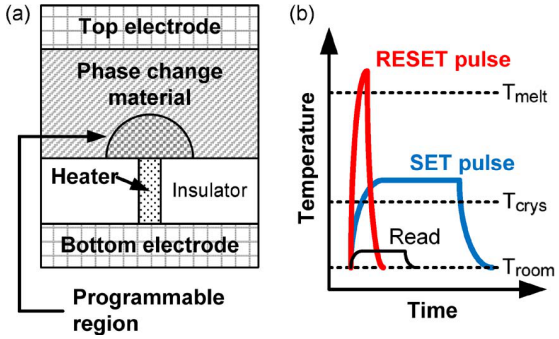


Figure 2.17: (a) Phase-change memory cell schematic. When electrical current flows between the top electrode and the bottom electrode, the heater affects the boundary in the phase-change material to form high/low resistance states. (b)The device is programmed and read by electrical pulses which change the temperature accordingly.(Adapted from [11].)

$Ge_2Sb_2Te_5$ - GST) layer in contact with a metallic heater. Fig. 2.17 shows the cross-section of the PCRAM device. When a large electrical current pulse is applied for a time period, the GST material is first melted and then quickly cooled-down forming a highly resistive amorphous region at the boundary of the heater. In this case the cell is said to be reset. In order to set the cell back into crystalline form, GST is heated beyond the crystallization temperature and then cooled-down slowly for sufficient time to crystallize.

Depending on the materials and cell geometry, varied resistance windows, endurance programming current, and durations can be achieved [11]. Up to five orders of magnitude (10^5) resistivity contrast between high and low resistances is observed [11]. Set operation can be achieved in 200ns with $60\mu A$ and reset in 20ns with $200\mu A$ [58]. Recently, an endurance value of 10^{10} has been reported [59].

One of the major challenges in PCRAM is the requirement of large reset current which puts limitation on scaling beyond 20nm [60] and creates thermal disturbs among adjacent bits. The addition of an inter-facial HfO_2 layer can reduce the current by 80% [61] and up to 10x reduction can be achieved with SiO_2 layer [62].

Main PCM applications include DRAM replacement. An 8 Mb memory in 130nm technology is reported in [59]. Micron released 1Gb memory in 45nm [60]. Samsung fabricated a 8Gb memory in 20nm with $4F^2$ cells [63]. Several techniques are proposed to reduce the number of writes due to limited PCM endurance [64] especially for cache

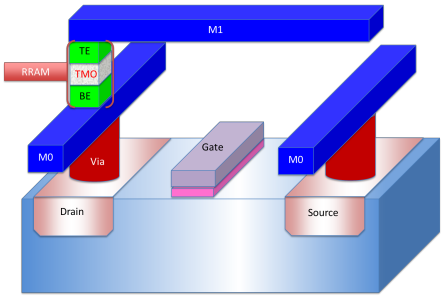


Figure 2.18: RRAM memory cell cross-section with select transistor.

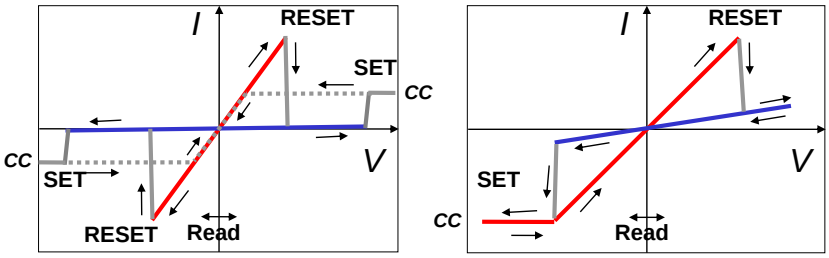


Figure 2.19: RRAM I-V characteristics for a) unipolar devices and b) bipolar devices.(Adapted from [12].)

applications.

2.3.1.3. Resistive Memory (RRAM)

Various terms have been used for devices exhibiting resistive properties by switching from high resistive state to low resistive state by flowing a current through the device such as RRAM, ReRAM, OxRAM, OxRRAM. The device structure is constructed by a simple Metal-Insulator-Metal (MIM) where an oxide material is placed in between two metal electrodes as shown in Fig. 2.18. Several materials have been researched which exhibit resistive switching behaviors. Most extensively studied metal oxide materials include HfO_2, Al_2O_3, TaO_2 and TiO_2 [12].

Two major categories can be identified depending on the switching properties: unipolar and bipolar as explained in Fig. 2.19. In unipolar memories, the amplitude of the voltage determines the switching direction. In bipolar RRAMs, the switching depends on the polarity of the applied voltage. Classically, switching from high-resistance state (HRS) to low-resistance state (LRS) refers to set and LRS to HRS switching refers to reset.

During set operation, high electric field causes the oxygen atoms to drift and creates oxygen vacancies leading to a Conductive Filament (CF). Usually the device goes through a forming process at a higher voltage than set voltage. Although several explanations have been discussed for the understanding of reset operation, a unified mechanism for both unipolar/bipolar modes is proposed [65]. For devices switching in unipolar mode, increased temperature due to the current flowing through, dissolves the oxygen ions either in the bulk to combine with the oxygen vacancies or in the metal boundary to oxidize the metal precipitates. Thus, a larger reset current is necessary. However, in bipolar mode, the inter-facial oxide layer created at the anode presents a significant diffusion barrier which requires a reverse electric field because pure thermal diffusion cannot overcome the barrier.

Since the high forming voltage imposes difficulties in circuit design, forming free devices have been proposed. HfO_2 is shown to operate without forming due to the reduced oxide layer thickness at 3nm [66].

RRAMs received interest because of back-end-of-line (BEOL) compatibility and reduced programming current requirement. Compact modeling took effort for fast circuit simulations [67]. Several logic applications have been proposed [68] with OxRAM cell having 100ns switching time and 10^8 endurance. Samsung reported the highest OxRAM endurance of 10^{12} with 10ns switching duration. Panasonic released a 8Mb OxRAM in $0.18\mu m$ technology [69]. Recently, Toshiba and Sandisk fabricated a 32 Gb OxRAM in 24nm tehnology [70].

2.3.1.4. Conductive-Bridge Memory (CBRAM)

CBRAMs, also known as Programmable Metallization Cell (PMC) [71] or NanoBridge [72], are one of the emerging technologies for Non-Volatile Memories (NVMs). It is a two terminal device composed of an active anode and an inert cathode layer where a solid electrolyte is embedded in between. CBRAM switching mechanism is based on polarity dependent electrochemical deposition and removal of metal in the thin solid state electrolyte film. The device switching mechanism is illustrated in Fig. 2.20. When a positive voltage is applied to the anode electrode, CBRAM goes into a fast program operation where the Ag ions are carried into chalcogenide material as a result of a redox reaction. The device is said to be set and a Low Resistance State (LRS) is achieved due to the stable conductive bridge between the electrodes. In order to reset,

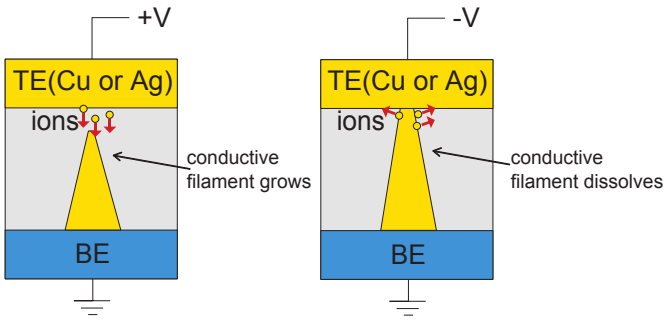


Figure 2.20: CBRAM device switching mechanism for SET and RESET operations. (Adapted from [13].)

a reversed voltage is applied, i.e. a positive bias at the cathode electrode. Since the conductive bridge is dissolved, the metal ion concentration is reduced. The electrodes are disconnected and a high resistance state (HRS) is achieved.

Several works target the application of CBRAMs in the literature. NEC demonstrated a novel 1T1R CBRAM cell [72] with a very low on resistance of 50Ω to be included in FPGA. A 1T2R memory cell for FPGA configuration memory is proposed and different material stacks are compared to achieve reduced leakage current and footprint [73]. Furthermore, the cell is characterized under various voltage levels and it is shown that the cell can be configured with distinctive resistance values (R_{ON}:4kΩ - 10kΩ, R_{OFF}:10^6Ω - 10^8Ω). Various materials have been considered when constructing the stack in different process nodes resulting in different resistance levels and endurances [74]. ALTIS fabricated a 2Mb CBRAM memory core in 90nm technology with $10^{11}\Omega$ R_{off} resistance [75]. Taking advantage of high resistance window and well-separated resistance values, Multi-Level Capability (MLC) of CBRAM has been investigated for R_{ON} between $10^3\Omega$ - $10^6\Omega$ with programming currents ranging from 100nA to $100\mu A$ [76]. Using the MLC concept, ALTIS fabricated a 4Mb CBRAM memory in 90nm with $4F^2$ 1T1R cells programmable under $1.4\mu s$. Recently, the crossbar concept is revisited with CBRAM-based cells using the high resistance advantage of CBRAMs to overcome the sneak-path problems and increase selectivity [77].

2.3.1.5. Ferroelectric Ram (FRAM)

FRAM cell is composed of ferroelectric capacitor, similar to DRAM (1T1C) to store the logic value of either 1 or 0 in the form of positive or negative polarization

charge states of ferroelectric dielectrics made of peroskite films such as $PbZr_xTi_{1-x}O_3$ (PZT). These cells exhibit non-volatile property however there are drawbacks due to low-compatibility with standard CMOS and large cell footprint. Toshiba recently fabricated a 128Mb chain FRAM memory to reach 25% low operation energy with high read/write bandwidth of 1.6GB/s [78]. Koga et al. proposed an FPGA in which the configuration SRAM cells are replaced with a FRAM-based Non-Volatile Flip-Flop (NVFF) [79]. Through the use of power gating a low stand-by power mode is established with a power-on duration of 1ms. Since the NVFFs are 9.6x bigger than a regular FF, the performance of the designed FPGA decreases by 1.8x. A work from Lien et al. demonstrates NVM/CMOS hybrid chip using FRAM-like nonvolatile memories and transistor stacking with 3D sequential integration [80].

2.3.1.6. Discussion

Reconfigurability in FPGAs brings the main advantage but at the same time it is the main drawback. The configuration memory cells occupy almost half of the chip, imposing deterioration on performance and power consumption due to the increased routing wirelengths and complexity. Configuration memories are, thus, the main target of emerging memory technologies in FPGAs. A comparison of emerging memories is illustrated in Table 2.1.

PCRAM, RRAM, and CBRAM cells provide area advantages due to the achievable $4F^2$ memory cell footprint. MRAM cells have larger footprints which results in less area improvement. Smaller cell area means that the memory overhead is less in FPGA which not only decreases total area but increases performance and power efficiencies.

Endurance is a metric which indicates how many times the memory cell can be written. Normally, in FPGAs, the bitstream is created once for an application. The bitstream can then be stored in the non-volatile memories which avoids consequent write operations if the same bitstream is preserved. If the FPGA executes with many different applications and bitstreams, a high endurance value is desired. Even though the highest demonstrated endurance is 10^{12} for MRAMs, 10^{15} values are expected. PCRAMs and CBRAM under perform in endurance with 10^6. RRAM endurance can reach similar levels as that of MRAM with 10^{12} cycles.

Table 2.1: Comparison of emerging memories

	SRAM	DRAM	Flash	PCM	RRAM	CBRAM	STTRAM	FRAM
Cell size (F^2)	140	4 - 6	4 - 10	4	4	4	8 - 20	12-22
Access time (W/R)	70-200ps	10ns	10ms/8-15ns	50-100ns/10-12ns	<1ns	<1ns	10-35ns	10-65ns/20-40ns
Write energy	0.5fJ	4fJ	100pJ	6pJ	115fJ	1pJ	2.5pJ	30fJ
Endurance	10^{16}	10^{16}	10^5	10^9	10^9	10^6	$10^{12} - 10^{15}$	$10^{14} - 10^{15}$
$R_{low}/R_{high}(\Omega)$				10K/2M	1K/100K	10K/10G	2K/4K	

In terms of resistance values, CBRAMs reach the highest R_{off} with low R_{on} ensuring the largest resistance window. Large resistive window increases design possibilities as it improves selectivity when integrated in a circuit.

When several write operations are requested to the memory, write energy plays a significant role in the total power consumption. Thus, the write energy consumption should be analyzed considering the target application. Compared to CBRAMs and RRAMs, MRAMs and PCRAMs have higher power consumptions during programing stage. It should be also noted that thermal effects due to the programming in MRAM and PCRAM might present reliability degradation on the surrounding transistors.

In comparison to other technologies, CBRAM and RRAM present small footprint, low write current, and thermal stability required for dense CMOS integration. Furthermore, RRAM reaches high endurance values and CBRAM provides large resistance window. Considering these properties, FPGAs can take significant benefits from RRAM and CBRAM integration.

2.3.2. 3D Integration

With exponentially increasing cost of new technology nodes, 3D integration becomes an appealing solution for "low cost" scaling. 3DICs provide performance benefits as improved packing density, noise immunity, optimized total power consumption due to decreased wire length/lower capacitance, increased performance, and more functionalities [14]. Theoretically, introducing the third dimension brings the possibility to decrease the wirelength by a factor $N_{tiers}^{1/2}$, delay by N_{tiers} and power by $N_{tiers}^{1/2}$ where N_{tiers} is the number integrated tiers [81]. Previously, several fabrication methods have been proposed for 3D integration based on design trade-offs such as wire bonded integration, microbump and through via [81]. Through Silicon Vias (TSV) [82] have potential due to the higher interconnect density it provides. On the other hand, recently developed 3D Monolithic Integration (3DMI) [83] technology achieves the highest interconnect density.

In TSV-based technologies, chips are processed with TSVs and integrated into 3D stacks either at the wafer or die level [84]. The dies or wafers are thinned to typically 50 - 100 μm before bonding. Depending on the top layer orientation during the process, the assembly can be face-to-face if the top of the second layer faces the top of the bottom layer and face-to-back if the bottom of the second layer faces the top of the bottom

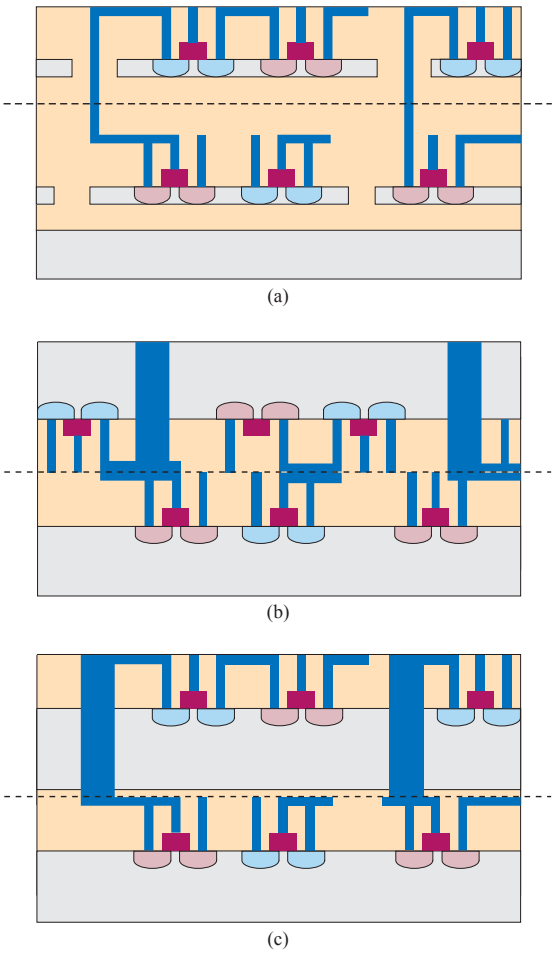


Figure 2.21: 3DIC assembly diagrams. a) SOI-based face-to-back process. b) face-to-face bonding. c) face-to-back process with deep vias formed between layers. (Adapted from [14].)

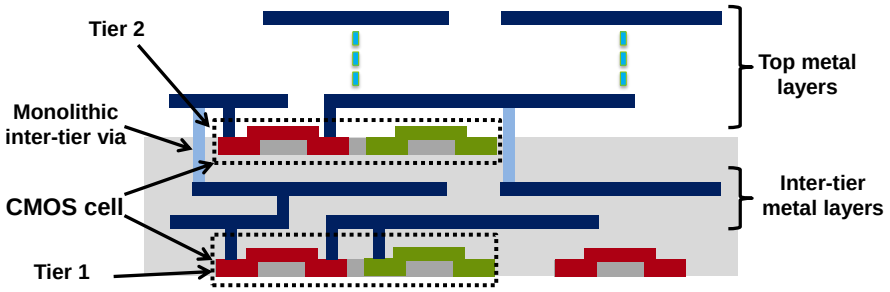


Figure 2.22: Cross-sectional view of 3D monolithic integration. Inter-tier vias are fabricated as traditional vias ensuring very small footprint and high interconnect density.

layer as demonstrated in Fig. 2.21 [14]. Following the bonding process Si is etched to form the TSVs connecting the two layers. Face-to-face bonding option requires creation of high-density Cu-Cu bonded links as well as deep TSVs to bring the signals out of the package. On the other hand, face-to-back relaxes the requirements for alignment as it has the largest via dimensions and thus lowest via density. The process of TSV-based assembly imposes difficulties in wafer thinning, alignment, bonding, TSV fabrication and thermal dissipation. In order to overcome the wafer handling problems during the bonding process wafers cannot be thinned below 50 μm increasing the demand for high-aspect ratio TSVs for smaller via surface. At the same time, large alignment tolerance restricts having smaller vias. Typically, the alignment achieved is 0.5 - 1μm resulting in a TSV size with 2 - 4μm.

Recently developed 3D Monolithic Integration consists of sequential fabrication of active layers on the same die. Fig. 2.22 shows the cross-sectional view of 3DMI. Since the inter-tier vias are fabricated same as regular vias between metals, the process alignment is only limited with lithography. An alignment tolerance of 10nm is achievable resulting in an inter-tier via diameter of 100nm [85]. Thus, high-granularity and less capacitive vertical interconnects are obtained with 3DMI.

Table 2.2: 3DMI VS TSV VERTICAL CONNECTION COMPARISON

	Alignment(μm)	Diameter(μm)	Pitch(μm)	Minimum depth(μm)
TSV	0.5 - 1	2 - 4	4 - 8	20 - 50
3DMI	0.01	0.1	0.2	0.1
3DMI vs. TSV gain	50x - 100x	20x - 40x	20x - 40x	200x - 500x

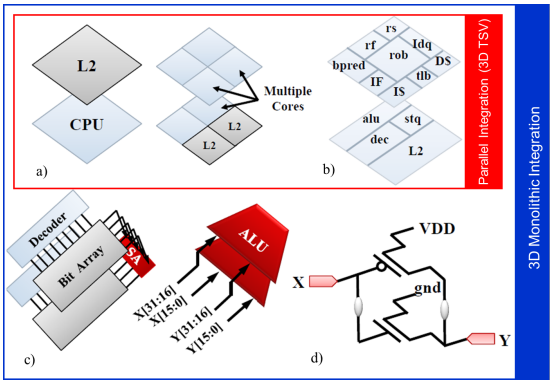


Figure 2.23: 3D partitioning for circuits from coarse to fine grain [15] a)Core-on-core b) Functional unit block c) Gate-on-gate d) Transistor-on-transistor.

Technological properties of TSV-based integration and 3DMI are summarized in Table 2.2. The size of the vertical vias in the technologies determines the achievable granularity. The TSV pitch of $4\mu m$ yields an interconnect density of $\sim$ 60k TSVs/mm^2 whereas in 3DMI $\sim$ 1.2M inter-tier vias/mm^2 can be fabricated. Consequently, the number of TSVs between two layers can only support coarse/medium grain partitioning. On the contrary, very fine grain partitioning is achievable with 3DMI. Fig. 2.23 shows partitioning on different granularity levels. At coarse grain level, core-on-core type integration is more suitable due to the reduced number of required vertical interconnects. In this case 3D TSVs provide efficient integration. As the granularity is increased, ex. gate-on-gate integration, high number of vertical interconnects is required. Since TSVs occupy large area due to alignment limitations, 3DMI becomes an ideal choice for 3D circuits with high density interconnects.

In circuit design, TSVs can provide performance advantages by reducing the wire lengths and placing the functional units closer. In [15], a 1mm long wire is replaced with a TSV and the delay is reduced from 225ps to 8ps. A number of works focused on circuit level gains with TSVs. Samsung [86] demonstrated a TSV-based four layer 8Gb DRAM memory with 50% less standby and 25% less active power while reaching an I/O speed more than 1600Mb/s. In [87], 3D DRAM stacking on Pentium 4 core increases the performance by 15% with 15% less power consumption with an 14°C elevated operating temperature.

Compared to TSV-based circuits, there are few works related to design with 3DMI. Jung et al. fabricated their SRAM design with single-crystal thin-film-based process

[85]. Thomas et al. demonstrated an SRAM design in 45nm with placing PMOS and NMOS transistors seperately resulting in 20% area gain [88]. Golshani et al. fabricated monolithically integrated silicon layers of SRAM and image sensors [89]. In [90], various design tradeoffs in 3DMI are studied and they are compared with TSV-based integration. The power benefits of 3DMI are discussed in [91]. Several design techniques have been developed in order to design 3D circuits with existing 2D standard cell libraries [92][93].

2.3.3. FPGA with Emerging Technologies

In the previous section, various emerging technologies have been reviewed. The main focus is on the emerging nonvolatile memories and 3D integration. In this section, related works on FPGAs with emerging technologies are presented as Non-Volatile Memory (NVM)-based FPGAs and 3D-FPGAs.

2.3.3.1. NVM-based FPGAs

Reconfigurable circuits such as FPGAs can be designed with nonvolatile STT-RAM-based memories [94][95][96]. STT-RAM based–FPGA logic circuits are presented targeting runtime reconfiguration and multicontext configuration [97]. Only LUT and routing switches are proposed and compared with other MRAM technologies such as TAS-MRAM and FIMS-MRAM. STT-RAM shows the highest benefits compared to other MRAM technologies and reaches up to 50% smaller LUT compared to SRAM-based counterpart. Paul et. al. proposed CMOS and STT-RAM hybrid approach for FPGA design with gains 11% in area, 11% in delay and 16% in power consumption [94].

A number of works focused on PCRAM-based FPGAs. Simulations are carried out to prove the performance gains when the PCMs are integrated in FPGAs [98][99][58]. In [98], a novel PCM-based switchbox is presented where the pass transistors are replaced with PCRAMs. In [99], PCRAMs and TSVs are integrated and 2-layer FPGA is simulated. Unfortunately, the result is very optimistic because the PCRAM access transistor is not taken into account in the area model. Moreover, as scaling continues the TSV overhead will become higher which decreases the area benefits. In [58], a PCRAM-based Non-Volatile Static Random Access Memory (NVSRAM) is proposed

and integrated in the FPGA for multi-context functionality. A low leakage operation is achieved, however area and delay evaluations are not carried out. Recently, a non-volatile look-up table is fabricated with IBM 90nm CMOS technology [100].

Due to the reduced footprint of memory cell, FPGAs can benefit highly from RRAM integration. Cong et al. proposed a RRAM-based FPGA where the interconnect MUXs are replaced with RRAM-based cells [101]. Only routing resource is updated with this approach. Results show major improvements in the routing resource with 96% less area, 55% higher performance and 79% lowered power consumption. Even though significant improvements are gained, sneak-path effects must be considered as it degrades selectivity. Chen et al. proposed a RRAM-based LUT which uses a crossbar-based design [102]. The LUT4 design grants 24% less area with 26% reduced delay. Sneak-path effects and power consumption including additional circuit are not analyzed. Recently, Liauw et al. fabricated a RRAM-based FPGA in which the SRAM-based configuration memory nodes are replaced with RRAM-based memories [103]. The adoption of RRAM-based memory node reduced the FPGA area by 40% and EDP by 28%.

2.3.3.2. 3D-FPGAs

The benefits of 3D integration are especially greater for designing FPGAs compared to other logic circuit applications, since FPGAs suffer from communication overhead: the interconnect delay and power are the main bottlenecks compared to ASIC alternatives. With 3D, the wirelenghts can be extremely reduced which results in increased performance with lowered power consumption. Furthermore, the replicated architecture of FPGAs ensure fast adoption of 3D technologies and efficient scalability.

In the literature a number of works has shown interest in the FPGA with TSV integration. In this case, multiple layers of FPGA tiles can be stacked. In [104], different 3D switch box topologies are surveyed. Since more complex switch boxes are required, the improvement of delay is limited to 10% when two layers are considered. There are several works which focus on decreasing the number of TSVs in order to reduce area and delay [105][106]. A novel approach placing the I/O and logic resources in different layers improves the maximum frequency by 26% [107]. Finally, a tool called TPR is developed for placement and routing of FPGA benchmarks with TSVs [108]. Hamada et al. designed an FPGA using face-to-face integration with 30% less critical path delay on a 48% smaller footprint [109] where the technological properties

of TSV and microbumps are not included in the evaluation. Therefore, the research for TSV-based FPGAs has been carried out in different levels and perspectives.

Silicon interposers are proposed as a low cost solution node between 2D and 3D. FPGA vendor Xilinx has released an FPGA fabricated on interposer [110]. In this work, the TSV aspect ratio is reduced to 10 with increased yield. Several slices, which correspond to a group of FPGA tiles, are integrated on the passive interposer. As a result, the fabricated slice circuit surface is reduced which increases the yield of each slice.

Recently, a number of works focused on monolithically integrated FPGAs. Naito et al. demonstrated the first 3D FPGA design with monolithic integration in 90nm by placing TFT-based configuration SRAM cells in the BEOL [111]. Due to the large footprint of the SRAMs with TFTs, an underoptimized tier utilization is expected. In [8], authors show improvements depending on several different stacking scenarios of FPGA blocks. The blocks are separated based on the area estimation and evaluations up to 3 layers are carried out with technological assumptions. Assuming smaller area in 3 layer stack, the results show improvements of 69% area, 41% delay and 41% power reduction. In [112], a switchbox with memory and logic separation is presented different benchmarks are evaluated on the FPGA. The overall effect of the proposed 3D SB on the FPGA is estimated as reduced delay by 22% and area by 21%.

2.4. Conclusion and Work Positioning

FPGAs have received remarkable attention in the digital circuit domain due to their reconfigurable, high-performance, and low-cost solutions. However, compared to ASICs, the reconfigurability comes with overheads in area, performance and power consumption. It is worth noting that FPGAs are designed by replicating the tile-based units which offers a scalable implementation. FPGA design is also accompanied by efficient tools which can define the FPGA architecture in detail. These properties create a very efficient platform for advanced technologies. Industry trends have already shown that FPGAs are usually the first applications for new technologies and, thus, well-suited for fast adoption of emerging technologies.

Emerging memories provide a unique feature of non-volatility in a very small form factor with CMOS-compatible integration. Several nonvolatile memory technologies

have been surveyed and it is observed that, the properties of these memories vary in a wide range. When designing with these memories, it is extremely important consider the device properties and, then, evaluate FPGA level implications. CBRAM and OxRAM technologies offer promising results small footprint, low write current, and thermal stability required for dense CMOS integration. Furthermore, with non-volatility, FPGAs can support a very efficient power reduction mechanism. Conventional FPGAs offer very limited power management. The main limitation stems from the volatile memories. Even though significant leakage reductions can be achieved with technological improvements, until the power supply is cut off, the leakage cannot be completely reduced to zero and when power gating is applied, configuration memories lose their information. Therefore, with non-volatile memories, FPGA power consumption can be efficiently reduced.

Logic design in the third dimension addresses the difficulties of traditional CMOS scaling. FPGAs in this case can highly benefit due to their regular architecture. As explained before, TSV and 3DMI are the candidate technologies for 3DIC design. Previous works with TSV integration shows limited improvements on FPGA while the only mainstream application is the yield improvement with the inclusion of interposers. It can concluded that FPGAs clearly need high-granular 3D interconnects due to their communication dominated nature which cannot be fulfilled by the available TSV integration. Due to the very-high granular interconnects proposed by 3DMI, FPGAs can gain more advantages.

3

FPGA Evaluation with Emerging Technologies

Contents

Compared to ASICs, FPGAs benefit from the most advanced technologies. When a real industrial FPGA is designed, circuit designers will spend months carefully tuning and trading off aspects in the circuit design targeting the new technology. It is clearly not possible to spend such manual effort for every potential architecture and technology.

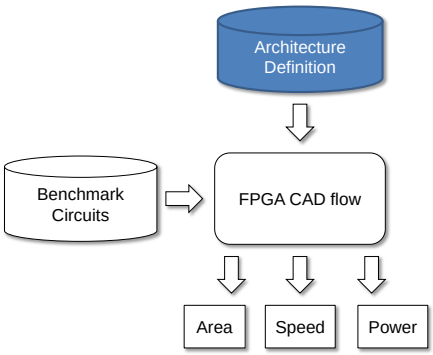


Figure 3.1: Architecture exploration empirical FPGA flow [16].

When emerging technologies are targeted, it is mostly impossible to explore all the design possibilities by designing and optimizing with these technologies and measuring the performance metrics from the fabricated FPGA circuits. Therefore, a fast, reliable, and flexible FPGA evaluation environment is indispensable.

FPGA vendors use their own environments for the development of the FPGAs. The tools in these environments are optimized for the specific architectures offered by the vendors. Therefore, it is not possible to evaluate different architectures with these tools. In the academia, some research effort has been spent to develop a toolflow which uses a parameterized FPGA fabric. Commonly, an empirical approach is employed to study and explore different architectures. In this approach, application benchmark circuits are mapped into the FPGA fabric to determine the area, speed, and power benefits of the developed FPGA architecture as illustrated in Fig. 3.1.

FPGA Computer-Aided Design (CAD) flow in Fig. 3.1 fills the gap between high-level application and the low-level implementation. The flow has to map the functionality into a predetermined architecture of logic and routing resources. The CAD tools, which are used to configure the programmable logic and routing switches of the FPGA, have a significant impact on speed, area and power of an application. Hence, the tools have to closely follow the low-level implementation properties in order to optimize the application mapping.

The architecture definition includes all architectural properties and technology dependent parameters. The flow in Fig. 3.1 was used for the exploration of architectural modifications such as LUT size, cluster size, channel width etc. Beyond that, the architecture definition also provides enough flexibility for evaluation with technology

dependent details such as the delay of the connection box, the resistance and the capacitance of the switches, the leakage current of the memory cell etc. Thus, by modifying the architecture definition, not only architectural modifications but also technological improvements can be imposed in the FPGA evaluation.

Several emerging technologies are presented in the Section 2.3.. Before designing FPGAs with these technologies, an evaluation platform has to be established. Therefore, in this chapter, first an FPGA CAD flow is assembled. This flow produces the design comparison metrics; area, delay, and power, at the output of each evaluation run. In order to evaluate different architectures, architecture definition, which is an input to the flow, has to be generated accurately. Thus, a methodology is developed for creating the architecture definition for FPGA evaluation using emerging technologies. Consequently, with this evaluation platform, FPGA fabric can be designed with emerging technologies and several benchmark circuits can be evaluated with this fabric to assess the potential improvements from these technologies.

In this chapter, first section describes the typical FPGA CAD flow. The next section presents the FPGA experimental evaluation framework, which is used in this thesis. The architecture definition, modeling for area, performance, and power, and MCNC benchmark circuits are also explained. The final section provides the development of architecture definition necessary for the adoption of emerging technologies.

3.1. CAD for FPGAs

Usually, the functionality of FPGA is defined at a higher level of abstraction, generally with a hardware descriptive language (such as VHDL or Verilog). Computer-aided design (CAD) tools transform this high-level descriptions into configuration bitstreams. Circuits are, then, mapped into FPGAs using the bitstreams which contain the state information of each configuration point defining the connections in routing resource and the logic functionality in Logic Blocks (LB). FPGA CAD flow consists of sequential execution of several tools. A typical CAD flow is illustrated in Fig. 3.2 and described in the following sections.

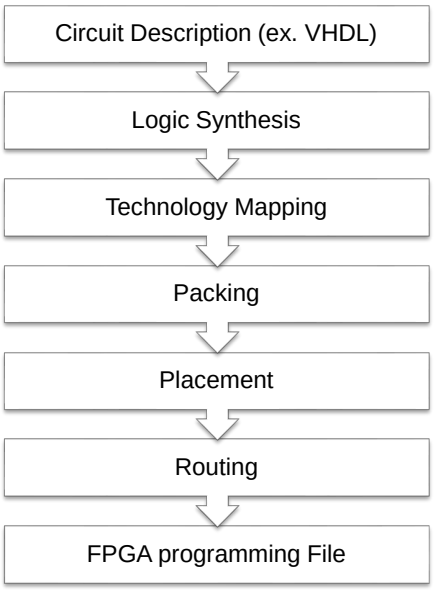


Figure 3.2: Typical FPGA CAD flow.

3.1.1. Front-end Synthesis

FPGA CAD flow starts with the front-end synthesis of the application being mapped. At this step, the high-level description (ex. VHDL) in register-transfer-level (RTL) is translated into a netlist of technology-independent logic functions in terms of generic truth tables including how they are connected to each other. For academic research, ODIN II [113] is the most commonly utilized synthesis tool. ODIN II is an open source HDL elaboration environment that converts Verilog HDL designs and creates a flattened netlist in the form of Berkeley Logic Interchange Format (BLIF) consisting of I/Os, logic gates, flip-flops and hard circuits such as multipliers.

3.1.2. Technology Mapping

The resulting netlist from front-end synthesis is a circuit definition using Boolean gates, flip-flops and connecting wires. Before mapping, various technology-independent techniques are applied to optimize boolean network. Removing redundant gates and simplifying netlist help to reduce area utilization and switching nets. After this optimization step, for a given netlist, technology mapping can be expressed as finding a network of cells using the cell library associated with the target technology. In the case

of FPGA logic, the cell library is composed of Look-Up Tables (LUT) and Flip-Flops (FF). Therefore, technology mapping in FPGAs refers to transforming the Boolean netlist into LUTs. Each cell is then implemented with a k-input LUT (K-LUT). Since a K-LUT can implement any K-input function, the mapping task of a Boolean network to a LUT is simply choosing a set of K-feasible cuts that include all the nodes in the network. Depending on the objectives, technology mapping algorithms can optimize the design for logic depth, area and power. SIS [114], FlowMAP [115] and ABC [116] tools have been proposed to address the technology mapping problem. The cut algorithm in ABC produces much lower literal count.

3.1.3. Packing

In typical FPGAs a hierarchical structure is used for logic blocks. In the first stage, logic functions are mapped into the K-LUTs, which, along with the corresponding FF, compose Basic Logic Elements (BLE). In the second stage, N LUTs are grouped together to form clusters. In this phase of CAD flow, clusters are created using the technology mapped BLIF netlist consisting of LUTs and FFs. Clustering tool groups several LUTs and the associated registers into one logic block, considering the limitations such as the number of LUTs contained in a logic block, the size of the LUT, independent number of input signals and clocks associated with the logic block. The main optimization goals are to cluster connected LUTs together to minimize the routing complexity between logic blocks and to reduce the number of logic blocks by utilizing the maximum capacity of each logic block. Clustering problem is mainly dividing the netlist into several pieces each of which containing minimum number of interconnections while respecting constraints such as maximum partition size.

Packing communicating BLEs together improves routability and critical path delay significantly [117]. Specifically, through clustering most of the global connections are converted into local connections. The efficiency of clustering depends also on LUT size. Thus, a trade-off can be defined using cluster and the LUT size. When too many BLEs and large LUTs are clustered, the area efficiency decreases because the CLBs cannot be fully utilized and the critical path suffers because of the increase area. In the case of small or no clusters and small LUTs, the delay is dominated by the global connections. It has been proven that for highest efficiency in area and delay, 4-6 input LUTs are included in the BLEs with 4-10 BLEs in the LBs [117].

Clustering algorithms can be categorized mainly into three approaches: top-down [118], depth-optimal [119] and bottom-up [120] [121]. In FPGA CAD flow, bottom-up is the preferred approach due to its fast and simple execution because they only consider local connectivity and can satisfy the LB pin constraints.

The VPack algorithm is proposed using the bottom-up approach [121]. In this algorithm, the clusters are built sequentially one at a time. For each cluster, an attraction function ise used to select a seed LB from the available set of LBs and new LBs related to the seed LB are selected until the cluster is full or all the cluster inputs are exhausted. A timing-driven version of VPack is the T-VPack algorithm [7]. The clustering algorithm is identical to VPack whereas the attaction function selects the seed LBs which are on the critical path rather than the LBs with the most used inputs as in VPack. Other algorithms include RPack [122], T-RPack [122], iRAC [123], multi-level clustering [124] and simultaneous mapping-clustering [125].

3.1.4. Placement

Placement determines the best on-chip location for every resources (ex. LB) based on the netlist produced by the clustering algorithm. The optimization goals are to reduce wire length by placing connected logic blocks close together (wire length-driven placement), and sometimes to balance the wiring density between blocks (routability-driven placement) or to minimize the critical path delays (timing driven placement).

Placement approaches can be categorized by three classes: min-cut (partitioning-based) [126], analytic [127], and simulated annealing [128]. Among all approaches simulated annealing is preferred because its algorithm is more adaptable to new optimization goals and architectural modifications. The algorithm in simulated annealing starts with an initial random placement of logic blocks to available locations in FPGA. Pairs of blocks are randomly selected and swapped repeatedly. At the end of each swap, the change in cost is calculated based on the new location. If the cost decreases, the new location is always accepted and the block is moved. If the cost increases, the block may or may not be moved depending on the improvement in the further steps. The output of placement stage is the detailed floorplan with an accepted location of all the resources.

Several tools are proposed for the placement of blocks in FPGAs. In general, the main objective is to minimize the wirelength between blocks. In all the tools, simulated

annealing is selected as the placement algorithm. Betz et al. proposed VPlace tool as a part of Versatile Place and Route (VPR) tool [7]. VPlace attemps to minimize the amount of the interconnect required by placing the blocks that on the same net close together. The cost function is calculated based on wire-length estimation between placed blocks.

Since the amount of routing in FPGAs is limited, the placement tool can optimize not just the wirelength but also the routability. Ebeling et al. developed a placer for Triptych FPGA [129]. Apart from the term for wirelength optimization, similar to VPlace, the cost function of the placer includes another term which monitors the fraction of logic blocks in a local area that are being used.

Marquardt et al. proposed the tool called T-VPlace, which is part VPR tool, as a timing-driven placement algorithm [130]. The main objective of T-VPlace is to minimize total wirelength and critical path delay. The cost function of T-VPlace has therefore two components: wiring cost, which is the sum of all dimensions of all nets and timing cost, which is the sum of connection delays weighted based on how close they are to the critical path.

3.1.5. Routing

In the final stage of FPGA CAD flow, all the nets are assigned to the prefabricated configurable routing resources in a way that only one net corresponds to one routing resource. The main objective is to minimize the critical path delay while avoiding congestion. Choosing direct connections and balancing the usage of routing wires ensure congestion-free routing. Delay minimization is achieved by routing the high-criticality nets first in the case of contention for a given resource.

Routing algorithms are categorized with two approaches: two-step routing [131] [132] which performs detailed routing after global routing and combined global-detailed routing [7] [133] which is performed in a single step. In FPGAs, typically, two-step approach is not preferred due to the difficulty in detailed routing as a result of the limited flexibility of the routing fabric. In single-step approach, during global routing, nets are assigned between logic block pin and routing channel, and during detailed routing, wire segments in a routing channel are assigned to nets. The routing of a circuit is generally represented by a directed graph, where each wire source node must

be connected to its sink nodes without any overlapping and often, directed search algorithm is included for efficient analysis of routing graph.

VPR [134] router is the most commonly used router in academic research. The routing algorithm is based on PathFinder [133]. In the router, an iterative routing algorithm is adopted which routes nets using minimum costs even though the solution may lead to congestion in some routing channels. Next the cost function is adjusted to increase the penalty of routing through congested resources. This process continues until a congestion free routing is achieved. The cost function includes a delay term and a congestion term. Thus, the cost function in VPR can optimize either the routability or performance. In routability-driven router, the cost function minimizes wirelength by limiting each net to use the fewest number of routing resources possible. In timing-driven router, the cost function assigns timing critical nets to use the fastest routing resource.

3.2. Experimental FPGA Evaluation Framework

A toolflow for FPGA CAD can be assembled using the previously discussed tools. The experimental process used in typical VPR-based flow is illustrated in Fig. 3.3. All the tools included are open-source which allows evaluation on various different architectures unlike the commercial tools which are targeted for specific architectures.

The toolflow takes a benchmark circuit and an FPGA architecture definition to implement the design and evaluate on a specified FPGA. ODIN is used as the front-end synthesis tool to convert high-level description (ex. VHDL) into flattened BLIF netlist. ABC tool takes the BLIF netlist and maps all the logic circuits into LUTs. Since ABC does not connect the clock signals to the FFs in the BLEs, a latch binding script is executed to allow evaluation with sequential circuits. ACE2.0 [135] performs the switching activity estimation which is necessary for power evaluation. A modified version of T-VPack [136] enables to cluster considering switching activities. VPR5 with Power Extension [137] accomplishes placement and routing of the circuit and extracts estimations of area, delay, and power.

VPR5 [138] offers several advantages in the evaluation framework. Similar to previous VPR, timing-driven placement and routing algorithms are included. Furthermore,

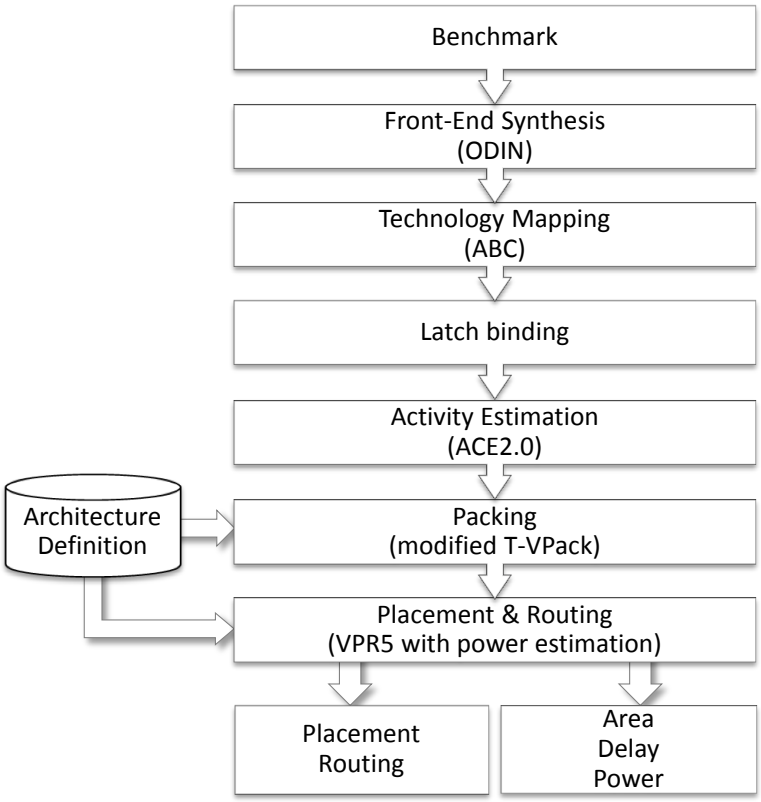


Figure 3.3: VPR5 with power estimation toolflow.

starting from this version VPR, a highly expressive architecture definition can be executed which enables utilization of a varied range of architectural and technological properties. VPR5 also introduces heterogeneity with which hard-blocks such as block memories or multipliers can be included in the FPGA architecture. It also presents the routing possibility with single-driver multiplexers. Currently the most common method for programmable interconnects in FPGA routing tracks is to employ a MUX-based switches for every track as opposed to the multiple tri-state or pass transistor drivers used in previous generations of FPGA. Thus, the toolflow is capable of evaluating recent FPGA architectures.

The remaining of this section discusses the architecture definition for design under evaluation, the estimation approach for area, delay, and power, and the MCNC benchmark circuits.

3.2.1. Architecture Definition

In general, tools cannot directly take the transistor size definitions of the FPGA fabric for correct evaluation. Typically, an architecture definition file is included in the evaluation toolflows. This file provides the freedom to define various FPGA architectures using parameters for LB organization, CB connectivity, and SB topologies, and technological characterization such as transistor capacitance, resistance, and area. For the toolflow defined in section 3.2., the architecture file can be defined as in 3.4. The device section defines the generic FPGA fabric such as LB area and SB topology. In the following parts, the routing resource properties are defined such as switch capacitance and segment properties. In the architecture, a simple timing model such as the delays of the switches in SB and CB are also included. For power calculations, parameters such as transistor parasitics, supply voltage level, and memory leakage are given.

3.2.2. Area, Delay, and Power Estimation

Area, delay, and power estimations of the FPGA circuit are obtained with the toolflow in Fig. 3.3. After generating an FPGA architecture with the definition discussed previously, VPR5 first places and routes the circuit and then extracts area, delay, and power based on the models implemented in VPR5.

VPR models area by accumulating the area of all the transistors in the FPGA including the routing resource, LBs, clock network, and configuration memory. The area approximation is achieved using Minimum Transistor Equivalents (MTE) metric as defined in [7], which calculates the layout area occupied by a minimum sized transistor plus the minimum spacing.

The delay is modeled after routing is completed. Based on the architecture definition, using the intrinsic delay of LBs and routing resource components, a directed graph is constructed. The capacitance and resistance of each node are then extracted based on the final routing. Thus, an RC equivalent is assigned to each node and the delay estimate of the critical path is calculated by Elmore delay model [7].

For power consumption modeling dynamic, short circuit, and leakage power have to be estimated. The dynamic power component depends on the activity of the node. The switching activity of the node implies how of the associated capacitance of node charges and discharges which directly affect power dissipation. For switching activity

ARCHITECTURE FILE

```
<device>
  <sizing R_minW_nmos="9107" R_minW_pmos="18214" ipin_mux_trans_size="1"/>
  <timing C_ipin_cblock="1.3e-15" T_ipin_cblock="2.52394e-10"/><area grid_logic_tile_area="61992"/>
  <switch_block type="wilton" fs="3"/>
</device>
<switchlist>
  <switch type="mux" name="normal" R="1750" Cin="20e-18" Cout="20e-18"  Tdel="0" buf_size="" mux_trans_size="1"/>
</switchlist>
<segmentlist>
  <segment type="unidir" length="1" freq="4" Rmetal="241.72" Cmetal="8.20654e-5">
  <sb type="pattern">1 1</sb><cb type="pattern"> 1 </cb><mux name="normal" />
  </segment>
</segmentlist>
<typelist>
  <type name=".clb"><subblocks max_subblocks="10" max_subblock_inputs="4">
   <timing>
   <T_comb><trow>7.83494e-10</trow></T_comb>
  <T_seq_in><trow>6.5e-11</trow></T_seq_in>
  <T_seq_out><trow>3.50254e-10</trow></T_seq_out>
  </timing>
  </subblocks>
  <fc_in type="frac">1</fc_in><fc_out type="full"></fc_out>
  <pinlocations><loc side="left">0 4 8 12 16 20 24 28 32 </loc>
  </pinlocations>
  <timing>
  <tedge type="T_sblk_opin_to_sblk_ipin">5.111e-10</tedge><tedge type="T_fb_ipin_to_sblk_ipin">5.111e-10</tedge>
  <tedge type="T_sblk_opin_to_fb_opin">0</tedge>
  </timing></type>
</typelist>
<power>
  <Nmos Vth="0.423" CJ="5E-4" CJSW="5E-10" CJSWG="3e-10" CGDO="15E-11"  COX="19e-3" EC="4e6"/>
  <Pmos Vth="0.365" CJ="5E-4"  CJSW="5E-10" CJSWG="3e-10" CGDO="15E-11"  COX="18e-3" EC="5e6"/>
  <poly Cpoly="1E-10" poly_extension="0.18e-6"/>
  <min_transistor_size length="65e-9" width="130e-9"/>
  <Vdd>1.2</Vdd><Vswing>1.2</Vswing><Vgs_for_leakage>0.12</Vgs_for_leakage>
  <SRAM_leakage>0.28e-09</SRAM_leakage>
  <short_circuit_power_percentage>0.1</short_circuit_power_percentage>
</power>
```

Figure 3.4: Architecture definition file for VPR5.

estimation the tool called ACE2.0 [135] is included in the framework as shown in Fig. 3.3. ACE2.0 takes the technology-mapped netlist and determines the switching activity of each node in the circuit. Since the activity estimation is carried out before clustering, a tool is necessary to match the switching activities with clustered architecture. For this reason, the T-VPack tool from [130] is modified to consider the activity information [136]. The modified T-VPack determines the switching activities of the nets between clusters, the global clock lines, and the input/output of the LBs.

As a part of VPR5 with Power Extension [137], Poon et al. integrated a flexible power model to estimate dynamic, short circuit and leakage power for island style

Table 3.1: Properties of MCNC 20 largest benchmark suite.

Benchmark	BLEs	FFs	I/Os	Type	Operation
alu4	1592		14/8	Compute-bound	Combinational
apex2	2072		38/3	Compute-bound	Combinational
apex4	1299		9/19	Compute-bound	Combinational
bigkey	2039	224	4/6	IO-bound	Sequential
clma	8383	33	75/69	Compute-bound	Sequential
des	2155		29/106	IO-bound	Combinational
diffeq	1875	305	229/197	Compute-bound	Sequential
dsip	1619	224	38/304	IO-bound	Sequential
elliptic	4183	194	41/35	IO-bound	Sequential
ex1010	4682		256/245	Compute-bound	Combinational
ex5p	1162		64/39	Compute-bound	Combinational
frisc	5869	886	229/197	Compute-bound	Sequential
misex3	1477		131/114	Compute-bound	Combinational
pdc	5353		10/10	Compute-bound	Combinational
s298	2429	14	8/63	Compute-bound	Sequential
s38417	8108	1462	20/116	Compute-bound	Sequential
s38584.1	6920	1260	16/46	IO-bound	Sequential
seq	1911		14/14	Compute-bound	Combinational
spla	3865		16/40	Compute-bound	Combinational
tseng	1401	385	52/122	IO-bound	Sequential

FPGA architectures [136]. The operating frequency of the application is determined with the critical path delay of the circuit under evaluation and the dynamic power is, then, calculated assuming that the circuit is operating with this frequency. Dynamic power dissipation is calculated based on the signal activities and associated node capacitances that are extracted in VPR5 tool. The short circuit power consumption is defined as a percentage of total power consumption which is 10% by default and can be modified by the user. Subthreshold leakage consumption model is used for the calculation of leakage power.

3.2.3. Benchmarks

Since FPGAs are programmable devices, performance, area, or power efficiency of an architecture cannot be measured until the designed FPGA is programmed with the desired application. Therefore, the performance of a particular FPGA implementation is experimentally measured using the 20 largest MCNC benchmarks [139]. The included benchmarks (Table 3.1) range from computation intensive to routing intensive and small area to large area to observe the effects of each architectural modification.

3.3. Methodology for Emerging Technology Evaluation

Even though FPGA fabric is a simple functional unit, several design choices can be imposed in the architectural development. Especially, when emerging technologies are employed, accurate modeling of these architectural and technological parameters is crucial. It is expected that several runs have to be executed iteratively for architecture optimization and design space exploration. Therefore, fast, flexible, technology and architecture-aware exploration platform is necessary.

3.3.1. FPGA Evaluation Platform

For the adoption of emerging technologies, the conventional design approach is not sufficient. In general, a standard cell library is constructed and then verified for logic design. These cells are provide varying driving capabilities and other properties such as low-power or high-performance. The cells are then used in an automated place and route tool. This approach is optimized for advanced technologies where the tools and libraries are mature enough for reliable designs. However, during design with emerging technologies, this level of maturity cannot be attained for design exploration. Since these technologies are not mature enough, it is sometimes difficult and most of time impossible to find the required tools and the standard cell libraries.

In order to accomplish an FPGA design a partial standard cell library might suffice because FPGA building blocks are not very complex and the entire FPGA design can be accomplished with few cells. However, depending on the targeted technology new tools might be included in the automated CAD flow such as a 3D place and route tool to design 3D circuits. Thus, a certain level of abstraction is necessary. In this thesis, a FPGA evaluation methodology for emerging technologies is proposed as shown in Fig.

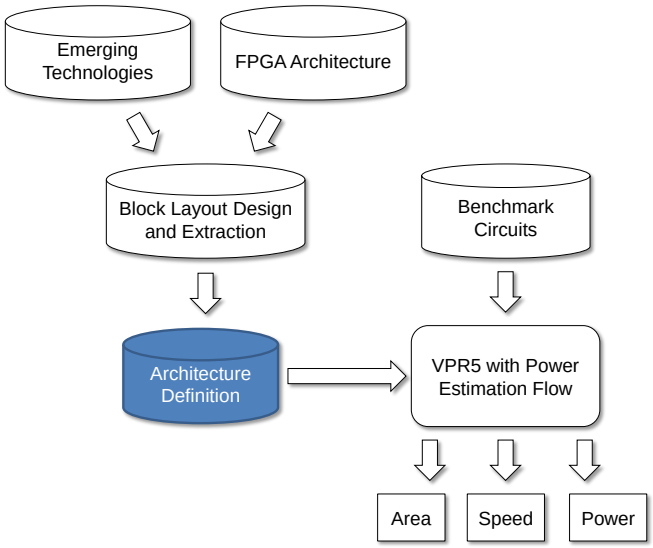


Figure 3.5: FPGA Exploration platform with emerging technologies.

3.5. Depending on the maturity of the technology, cells are designed and characterized either at the device or block level. With this characterization, FPGA architecture definition is created. This definition is used as an input to the VPR-based CAD flow for performance, area, and power evaluations.

3.3.2. Architecture Definition Development for Emerging Technologies

With the framework defined previously in Section 3.2., the impacts of LUT, cluster size, SB topology, or transistor parasitics can be observed rapidly. In these cases, the FPGA architecture is evaluated based on parameterized assumptions where the layout parasitics can not be effectively examined. Thus, in this thesis, the methodology in Fig. 3.6 is proposed for the development of the architecture definition with emerging technologies.

With the methodology, architecture definition is built using FPGA building blocks. As described in Section 2.1., FPGAs are tile-based circuits and one tile is constructed with one Logic Block (LB), two Connection Boxes (CB), and one Switch Box(SB). These building blocks are established with configuration memories, multiplexers, and buffers. Hence, once these elementary blocks are designed in the layout and included in the design library, all the FPGA related blocks can be designed quickly. The de-

sign process is, therefore, shortened substantially because only few blocks need to be designed on layout rather than the entire FPGA.

After the layout is completed, design and technology related parameters are extracted for architecture definition development. The parameters span from transistors, to configuration memories and blocks. First transistor related parameters are extracted. Transistors parasitics capacitances are necessary for the calculation of dynamic power. Transistor dimensions are used for the area estimation of the FPGA circuit. With these parameters, technological details of transistors are introduced in the architecture definition. Memory properties refer to the configuration memories in the FPGA. Cell area allows the estimation of the configuration memory area of the FPGA. Leakage power refers to the leakage power consumption of one memory cell and it is included in the calculation of total leakage of FPGA circuit. In the next step, FPGA building blocks are characterized. FPGA blocks are designed with full-custom design rules. Using the layout dimensions of the blocks, area related parameters are calculated. The blocks are then simulated considering the parasitic extraction. I/O capacitance, resistance and delay of the blocks and segments are carefully measured. With this methodology, FPGA architectures can be evaluated following closely the technological impact and achieving very rapid adoption of emerging technologies.

3.3.3. Memory Cell Area Modeling

Normally, VPR assumes in the source code, a 6T SRAM cell implemented with minimum size transistors. This memory transistor count is a part of the FPGA area model and used for the area estimation of the configuration memory in the FPGA blocks which is then used to estimate the routing wirelength. Hence, it not only affects the area but also routing complexity which is directly related to performance and power consumption due to the resistance and capacitance implications. In different technologies and topologies, however, SRAM might consume less or more area than a 6T estimation. Thus, we modified the VPR tool and the architecture file to be able to change the area of the memory cell. The *trans_SRAM* variable in the source code is added in the architecture file as shown in Fig. 3.7.

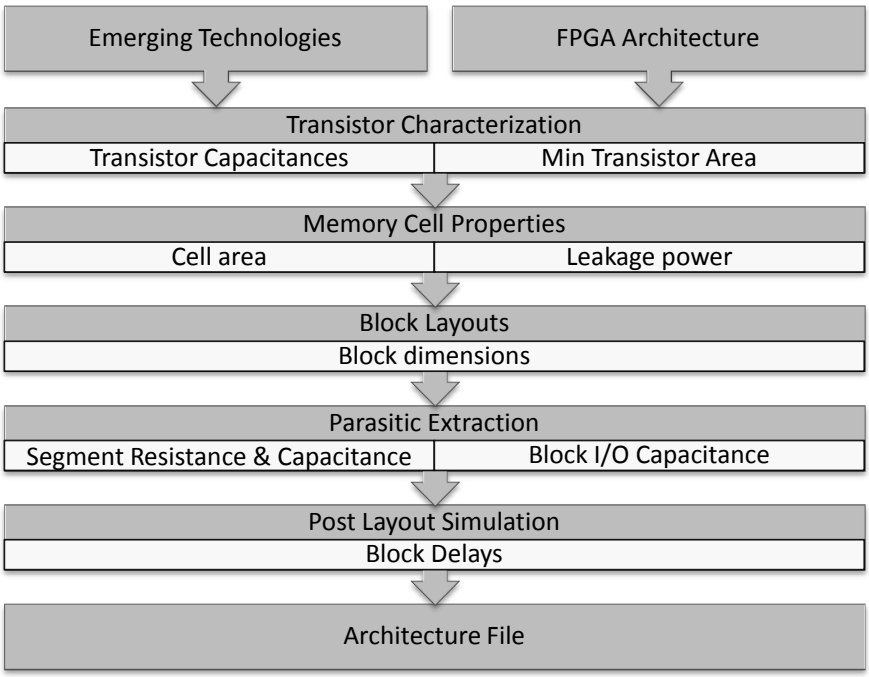


Figure 3.6: Methodology for architecture definition creation with emerging technologies.

```
<device>
  <trans_sram trans_sram_bit= " 6" />
</device>
```

Figure 3.7: Parameterized memory cell area.

3.4. Conclusion

FPGAs are the first computing units that benefit from the emerging technologies. In order to estimate the improvements achieved from these technologies, an evaluation framework is constructed in this chapter. The framework consists of a complete set of tools for the evaluation of benchmark circuits. Technological and architectural properties are expressed in an architecture definition which constitutes the FPGA fabric. The framework offers a fast, flexible environment to estimate area, performance, and power benefits of the designed FPGA.

As surveyed in Chapter 2, emerging technologies have impacts on various levels in the design process. For the assessment of the these impacts using the evaluation framework, a methodology is proposed. This methodology lowers the design effort with emerging technologies by introducing an abstraction level. Only FPGA building blocks

are characterized and the properties are included in the architecture definition for the evaluation using the framework.

The architecture definition methodology and the experimental framework from this chapter allow estimating the impacts of emerging technologies. This platform will be the basis for the evaluation of targeted emerging technologies in the following chapters.

4

Non-volatile FPGA with Resistive Memories

Contents

In order to support the main benefit of FPGAs which is the reconfigurability, a high number of Configuration Random Access Memory (CRAM) cells is necessary. As discussed in Section 2.1., almost half (43%) [8] of the FPGA area is allocated for CRAM with Static Random Access Memory (SRAM) being the preferred choice. The inclusion of these memories not only increases the total area of the FPGA but also extends the length of routing wires which deteriorates performance and power consumption. Apart from area, these memories also contribute to the leakage power consumption heavily.

Power gating can be incorporated for zero leakage but configuration memories lose the configuration information because they are volatile. Flash-based solutions can be integrated but CMOS integration difficulties, cost, and scaling limitations prevent Flash integration for large scale usage. Therefore, compact and non-volatile memories are of quest for area-efficient, high performance, and low power FPGAs.

Resistive memory technologies can mitigate the overheads due to configuration memories in FPGAs. A selection of emerging memories is surveyed in Section 2.3.. In this thesis Oxide Random Access Memory (OxRAM) and Conductive Bridge Random Access Memory (CBRAM) are the targeted technologies. These memories offer promising opportunities with inherent non-volatile operation capability and small footprint with CMOS Back-End-of-Line (BEOL) compatibility. Furthermore, with OxRAM high endurance, not only the configuration information but also operation results in the registers can be stored. CBRAM technology offers very high resistance windows which allows to design very compact memories.

Non-volatility can be introduced in the FPGA using Non-Volatile Memories (NVM)-based circuit elements. For the quest of non-volatile FPGA all the volatile elements have to be replaced. Mainly the memory is used for configuration storage (CRAM) as explained previously. Apart from CRAMs, the registers, Flip-Flops (FF) in the Basic Logic Elements (BLE), are also volatile elements of FPGA. A FF stores 1-bit result of the computation completed in the corresponding Look-Up Table (LUT). For the exploration with non-volatile elements, OxRAM and CBRAM based solutions are analyzed in FPGAs.

For the replacement of CRAMs, first, OxRAM-based non-volatile SRAM (NVSRAM) cells are integrated in FPGAs by replacing the SRAM counterparts. An NVSRAM operates exactly as an SRAM with the extended capability due to the non-volatile devices. Second, CBRAM-based 1T2R cell is used as CRAM by replacing the SRAM memories. This cell takes advantage from the high resistance window offered by CBRAM which allows to design a very compact cell.

With the integration of NVM in FPGA, new application fields can be targeted. Since SRAMs lose the information when the power is cut-off due to their volatile nature, the configuration information (bitstream) has to be transferred into the FPGA during wake-up. Thus, an external non-volatile memory (ex. Flash) is dedicated to

store the bitstream. With the integration non-volatile memories, the configuration information can be kept locally in the FPGA. Thus, configuration-saving applications can be accommodated with the designed FPGA.

For the replacement of FFs, an OxRAM-based non-volatile FF (NVFF) is introduced in the FPGA for non-volatile context storage. OxRAM elements store the information when necessary and the NVFF operates exactly as a FF. For some applications previously computed data are required for new calculations, as most of the information is contained in the evolution of monitored data. In this case, the previously calculated information (context) has to be stored in an external memory before the FPGA power is cut off and it has to be fetched when FPGA is turned to continue calculations. With the integration of NVFFs, the context can be locally stored in the NVFFs and the communication with the external memory can be avoided. Thus, context-saving applications can be accommodated with the designed FPGA. Furthermore, when OxRAM-based NVSRAM and NVFF are integrated together in the FPGA, configuration and context-saving applications can targeted.

When SRAM configuration memory cells are replaced with the NVM counterparts, significant area improvements can be achieved. The area reduction of configuration memories decreases the wirelength of routing wires which also improves the wiring capacitance. In order to observe the benefits, compact CBRAM-based 1T2R cell is integrated in the FPGA by replacing all the SRAM nodes. With this memory node, the FPGA area can be reduced by 33%. The reduction of area leads to less capacitance and resistance in the routing wires which results in decreases delay and power consumption by 34% and 23% respectively.

With the development of Internet-of-Things (IoT), embedded applications are emerging with varied specifications. Some of these applications can be categorized as "Normally-off, Instantly-on". These applications require short highly intensive computing phase in between of long idle periods. As explained in Chapter 2, FPGAs can fulfill the requirement for intensive computing however FPGAs have high overhead in power consumption which prevent FPGA from being employed in battery operated applications. Taking advantage of the non-volatility, the proposed FPGAs can store the information locally in the FPGAs and power gating can be applied to cut-off the power during idle periods. The FPGA can then be switched on rapidly for computations before switching off. In this case, a zero leakage state can be achieved during the idle periods and the

overall power consumption can be reduced. For applications having low activity, the power consumption can be reduced by 90%.

In this chapter, first the OxRAM and CBRAM-based elementary non-volatile circuits are explained. The area and power consumption characteristics are extracted. In the following section, non-volatile FPGA circuits are designed using the elementary cells for configuration and context-saving applications. Area, delay, and power consumption figures of the FPGA circuits are extracted using the flow in Chapter 3. Normally-off, Instantly-on computing is discussed in the following section. Power-gating implementation is explained for the utilization in FPGA and the power gains of the designed non-volatile FPGAs are evaluated.

4.1. RRAM-based Elementary Circuits

In this section, several basic elements are analyzed with OxRAM and CBRAM technologies. The design and operation of the cells are explained in detail. The elementary circuits can be used in other applications such as in ASICs if non-volatility is desired. In this thesis, the main objective to have non-volatile FPGAs. Therefore, the cells are characterized in area and power consumption targeting FPGA usage.

4.1.1. Non-volatile SRAM (NVSRAM)

4.1.1.1. Memory Cell Architecture

NVSRAM circuit is constructed by the hybrid combination of a typical SRAM cell and RRAM devices. The NVSRAM cell used for this work is designed by ISEP. The cell architecture and operating principle are proposed in [17]. Two OxRAM devices (R1, R2) are used per SRAM cell (M1-M6) in order to obtain nonvolatile operation. The OxRAM devices are connected to the data nodes of the SRAM cell in order to store the logical information of the SRAM cell in the event of a power-down. The OxRAMs (R1, R2) are accessed by the control transistors: CM1 and CM2 (Fig. 4.1).

Depending on the information at the SRAM data nodes, each OxRAM device is put either to a low or a high resistance state. By convention, the set of OxRAM corresponds to low resistance state and the reset of OxRAM corresponds to high resistance state. The 22nm FDSOI technology from CEA-LETI and HfO_2-based OxRAM compact model [140] are used to design the NVSRAM cell.

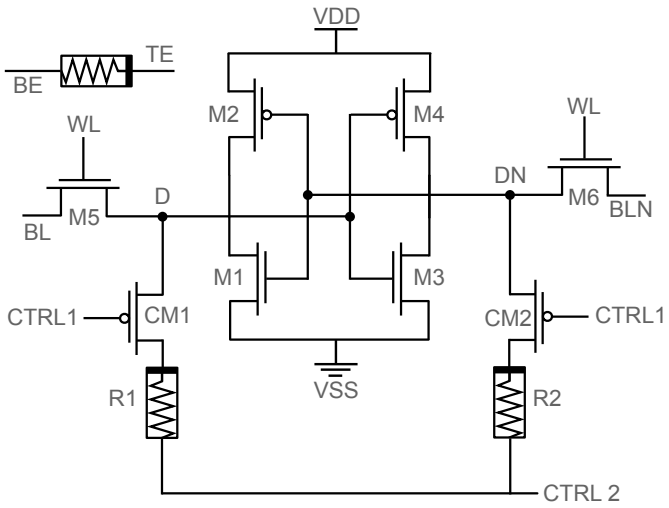


Figure 4.1: NVSRAM schematic of 8T2R architecture [17].

4.1.1.2. Operating Principle

One complete power-down/power-up cycle involves the following sequence: Normal SRAM operation – Reset – Store – Power-down – Power-up – Restore as depicted in Fig4.2. Table 4.4 summarizes the signals conditions of control signals for all operations involving the use of OxRAM devices.

For reset, appropriate signals are applied to the control lines, CTRL1 and CTRL2, to obtain a higher potential at the bottom electrode with respect to the top electrode of the RRAM devices. This operation is achieved by turning on CM1 and CM2 and providing a voltage of 1.5V to CTRL2. Depending on the logic value at the data nodes of the SRAM cell, either R1 or R2 will be reset; i.e. when D= '0' (DN= '1'), the negative potential drop across R1 will reset R1 and when D= '1' (DN= '0'), R2 will be reset.

For information storing, the control transistors are turned on and CTRL2 is set to 0V. A positive potential drop across one of the RRAM devices will set the corresponding RRAM device and the same potential at the top and bottom electrode of the other RRAM will make the corresponding RRAM stay at the same resistance state. In this way, the logic values at the data nodes (D= '1'/ '0', DN= '0'/ '1') are stored into the RRAMs as '1'/ '0' (R1= SET/RESET) and '0'/ '1' (R2= RESET/SET). The inherent non-volatile characteristic of RRAMs ensures the retention of this information when a power-down occurs.

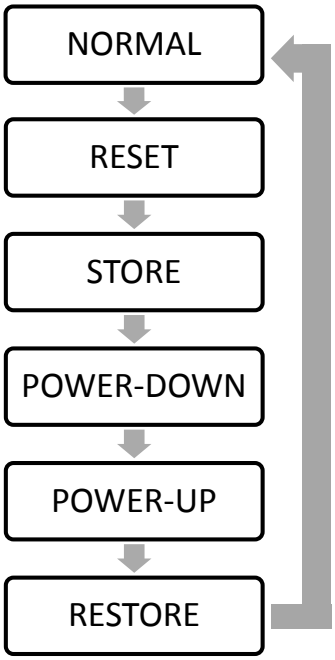


Figure 4.2: NVSRAM operation cycle: Normal SRAM operation – Reset – Store – Power-down – Power-up – Restore.

During the power-down, all the control signals are pulled down to VSS (0V). Thus, a state with no leakage consumption is achieved. At power-up, before restoring the information from the RRAMs, VDD is put to 1V and at the same time VSS is pulled up to VDD to have D=DN='1'. This prevents the flipping of the RRAMs during the RESTORE operation.

For restoring, CM1 and CM2 are turned on and simultaneously, 1V is applied to

Table 4.1: NVSRAM signal conditions during different operation phases.

	VDD(V)	VSS(V)	CTRL1(V)	CTRL2(V)
NORMAL	1	0	0	0
RESET	1	0	0	1.5
STORE	1	0	0	0
POWER-DOWN	0	0	0	0
POWER-UP	1	1	1	1
RESTORE	1	0	0	1

CTRL2. With respect to CTRL1 and CTRL2, VSS is lowered to 0V after a delay of 5ns. If R1 is low, node D is maintained at logic '1' but if R2 is high, node DN discharges through M3. In this way, the values are restored at the data nodes.

Normal SRAM operation can be performed right after the restore operation without having to reset the RRAM at low resistance state. SRAM operation is not influenced by the addition of the control transistors and RRAMs. In fact, CM1, and CM2 isolate the RRAMs preventing the degradation of the SRAM cell stability during normal SRAM mode. Thus, the control transistors do not affect the normal SRAM operation because they remain switched off after the restore phase is completed.

4.1.1.3. NVSRAM Cell Characterization

In order to evaluate system level impacts of NVSRAM integration in FPGA, power dissipation for each operation and total area of an NVSRAM cell are extracted. All the simulations are performed using Eldo simulator from Mentor Graphics and the results are validated on 22nm FDSOI process design kit (PDK) developed at CEA-LETI. A bipolar OxRAM model, calibrated on experimental results obtained on HfO_2-based devices, is included for simulations [140]. NVSRAM cell is optimized to achieve data store and restore operations at 20ns with low operating voltage of 1V. During power-down phase, a duration of 3.8μs is necessary to discharge all the capacitances.

For a fair comparison of area between SRAM and NVSRAM, the OxRAM devices and control transistors are removed to obtain a regular SRAM cell. The total area is calculated for the NVSRAM and SRAM by normalizing the transistor dimensions to minimum size transistor. Table 4.2 shows that NVSRAM is 32% larger than SRAM. The area overhead stems from the additional control transistors. Since the OxRAM devices are manufactured at the BEOL between metal layers, no area impact from OxRAM devices is observed.

Table 4.2: Area comparison of SRAM and NVSRAM cells

	Normalized area(x min. size transistor area)	Area(μm^2)
SRAM	11.5	4.10
NVSRAM	15.25	5.44

As explained before, for a single ON/OFF cycle, NVSRAM has to go through the following stages: reset, store, power-down, power-up, and restore as discussed previously. Using Eldo, an NVSRAM cell including the OxRAM model is simulated and, then, average power consumption for each operation is extracted with corresponding duration. Table 4.3 shows the results. For the completion of single ON/OFF cycle, NVSRAM dissipates 17.9 fJ. This value assumes that new information has to be stored in the OxRAM devices each time before powering down. The total cycle does not need to be followed if the stored information in OxRAMs is the same as the information in the SRAM part during normal operation.

When the NVSRAMs are integrated in the FPGA as configuration storage, it is not required to follow all the steps in the ON/OFF cycle. Specifically, the configuration bitstream does not change once it is generated and transferred to the FPGA. Since the same information is retrieved when restoring, it is not necessary to reset and store repeatedly. Hence, only power down and restore operations are taken into account. Considering this characteristic, the total required power for one ON/OFF cycle is reduced to 0.78 fJ per unit cell as shown in Table 4.3.

4.1.2. Non-volatile Flip-Flop (NVFF)

4.1.2.1. Flip-Flop Architecture

A Non-volatile Flip Flop (NVFF) based on RRAM is necessary when not only configuration, but also data values must be restored in an ON/OFF application. The NVFF cell architecture and operating principle are proposed in collaboration of LETI

Table 4.3: Power Data for general NVSRAM operation

	P_{avg}(nW)	Duration(ns)	Energy(fJ)
Reset	425	20	8.5
Store	434	20	8.7
Power down	0.06	3800	0.23
Restore	28	20	0.56
Energy for general operation including on/off (fJ)			17.9
Energy for FPGA configuration memory on/off (fJ)			0.78

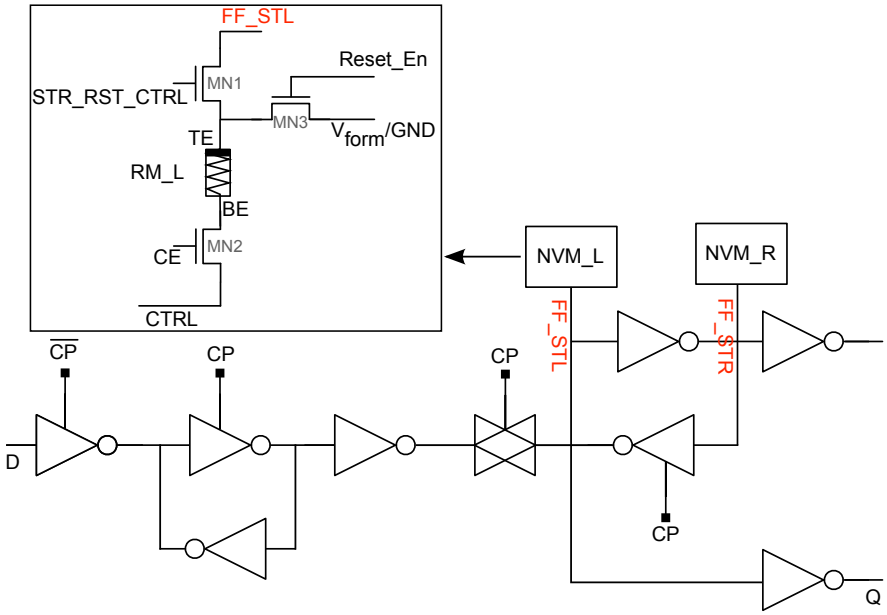


Figure 4.3: NVFF architecture with non-volatile block based on RRAM, the RRAMs store the slave state in NVM_L and NVM_R blocks.

and IM2NP [18]. The NVFF architecture is depicted in Fig.4.3. It consists of NVM_-L and NVM_R blocks as add-ons to the slave part of classical flip-flop architecture. Each NVM (NVM_L, NVM_R) block consists of three transistors and an OxRAM device. The transistors allow programming control of OxRAM and, also, generate the required compliance current on OxRAM. The various control signals utilized to achieve the functionality are shown in Fig.4.3.

4.1.2.2. Operating Principle

The entire sequence of operations that is followed to faithfully store and restore the state of the flip-flop is shown in Fig. 4.4.

In the normal operating mode of the flip-flop, the input data is latched in the master stage at active low clock signal and subsequently pushed to the slave part to be available at the output for following clock signal. During this phase all the control signals (STR_-RST_CTRL, CE, RESET_ENABLE, CTRL) governing the NVM blocks are inactive and, hence, the NVM blocks are isolated. The RRAMs remain in High Resistive State (HRS).

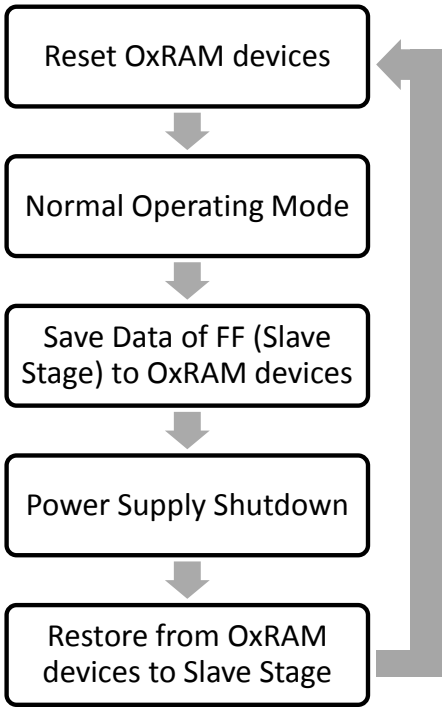


Figure 4.4: Control signal and state transition of NVFF.

Store operation is triggered by the control signals STR_RST_CTRL = '1', CE = '1', and CTRL = '0' guiding the logic state present in the slave latch to be stored in RRAM. For example, when slave has a '1', the RRAM in the NVM_R is SET (LRS), since a positive bias is applied between its terminals. The NVM_L block remains unchanged (HRS) since there is no potential drop between the RRAM_L terminals is 0V. In essence, the NVM_R with RRAM_R in LRS and NVM_L with RRAM_L in HRS form the logic state of '1'. Once RRAMs are programmed the flip-flop and, hence, the FPGA can be powered-down completely owing to the information stored in

Table 4.4: NVSRAM signal conditions during different operation phases.

	VDD(V)	STR_RST_CTRL(V)	RESET_ENABLE(V)	CE(V)	CTRL(V)
RESET	1	0	1	1	1
NORMAL	1	0	0	0	0
STORE	1	1	0	1	0
POWER-DOWN	0	0	0	0	0
RESTORE	1	1	0	1	1

non-volatile memories.

The restore is initiated by signals STR_RST_CTRL = '1', CE = '1', CTRL = '1', and RESET_ENABLE = '0'. The logic is restored on the slave latch for resuming normal operation with the VDD restoration. This is, in fact, a read operation on the RRAM. The presence of the complementary logic in NVM_L and NVM_R assists the acceleration of logic restoration process. This is followed by resetting the RRAMs to HRS using STR_RST_CTRL='0', CE = '1', CTRL= '1', and RESET_ENABLE = '1', which puts a negative bias on the RRAM terminals.

4.1.2.3. NVFF Cell Characterization

In order to evaluate system level impacts of NVFF integration in FPGA, power dissipation for each operation and total area of an NVFF cell are extracted. All the simulations are performed using Eldo simulator from Mentor Graphics and the results are validated on 22nm FDSOI process design kit (PDK) developed at CEA-LETI. The same OxRAM model used for NVSRAM is included for simulations [140]. In order to observe the total overhead imposed by the nonvolatile functionality, normal FF and NVFF are compared. As described previously, the nonvolatile block consists of 3 transistors and 1 OxRAM for each complementary node, adding up to 6 transistors and 2 OxRAMs in total for 1 NVFF. The total area is calculated for the FF and NVFF by normalizing the transistor dimensions in each cell to minimum size transistor. It can be observed from Table 4.5 that NVFF has 26% larger area than a regular FF. The area overhead stems from the additional control transistors. Since the OxRAM devices are manufactured at the BEOL between metal layers, no area impact from OxRAM devices is observed.

Similar to NVSRAM, NVFF has to follow reset, store, power-down, and power-up/restore stages in one ON/OFF cycle. The average power consumption for each

Table 4.5: Area comparison of FF and NVFF cells

	Normalized area(x min. size transistor area)	Area(μm^2)
FF	78.2	27.9
NVFF	99.2	35.4

operation is extracted from the simulations carried out using Eldo as shown in Table 4.6.

Contrary to NVSRAM integration in FPGA, NVFFs are used for context-saving purposes. Specifically, NVFFs will be utilized to store the results of the calculations carried out in the LUTs. Thus, before power-down, the information stored in the NVFF needs to be updated for no information loss. As a result, it is necessary to do the reset and store operations before power-down and restore. Considering this operation scheme, each NVFF consumes 11.4 fJ for one ON/OFF cycle as shown in Table 4.6.

4.1.3. NVE-based Design

4.1.3.1. Elementary Non-volatile 1T2R Memory Element(NVE)

A resistive divider node is a very attractive solution for storing information because of the reduced footprint. The memory cell called Non-Volatile Element (NVE) in Fig. 4.5 is proposed in collaboration between LETI and IM2NP [141]. The cell only requires one transistor (1T) which is used for programming of the cell and two resistive devices (2R) in complementary resistance states (one in High Resistance State (HRS) and one in Low Resistance State (LRS)). Since the resistive devices are fabricated at the BEOL, the area of the memory cell is limited with the area of the access transistor. Thus, compared to an SRAM cell, the memory area can be reduced significantly. However, even though there is an area advantage, the design of resistive divider imposes restrictions on the memory technology. Since two resistive elements are connected between source and ground, there is a continuous flow of leakage current. When all the configuration

Table 4.6: Power data for general NVFF operation

	P_{avg}(nW)	Duration(ns)	Energy(fJ)
Reset	180	20	3.6
Store	180	20	3.6
Power down	0.419	3800	1.6
Restore	131	20	2.6
Energy for general operation including on/off (fJ)			11.4

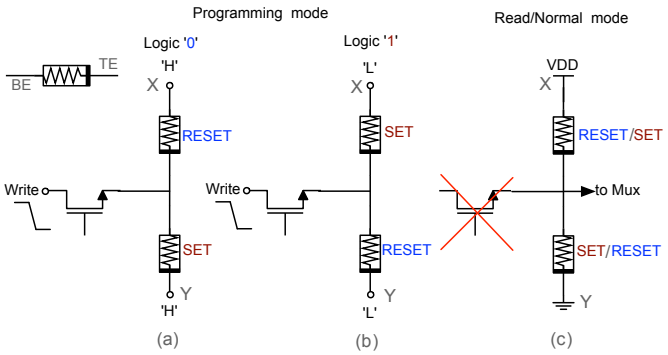


Figure 4.5: NVE circuit scheme for programming and reading of CBRAM cells in voltage divider configuration.

memories are considered in FPGAs, the leakage current becomes reasonably high. Another implication is that when sufficient voltage difference is attained on the resistive device programmed to LRS, the resistance value might flip leading to information loss. Therefore, in order to limit the leakage current and retention implications, a large HRS and a large resistance window (the ratio between LRS and HRS) are required.

As explained in Section 2.3., CBRAM offers a very large HRS of $10^{10}G\Omega$ and wide on/off state resistance window of 10^6. Thus, with CBRAM technology, reliable operation of resistive divider node can be achieved. The structure including one programming transistor and two CBRAM devices (1T2R) stores one bit of information. It can be integrated in the FPGAs to replace SRAM based configuration memory (CRAM) with NVE-based counterpart. Due to the compact layout NVE leads to reduction in FPGA silicon area in comparison to the traditional SRAM. Taking advantage of large HRS, the leakage consumption of the cell is in the range of SRAM.

4.1.3.2. Operating Principle

The (non-volatile element) NVE consisting of a transistor and two CBRAM cells is programmed by first selecting the cell using the selection transistor and then applying the appropriate voltages to the terminals across the CBRAM devices. The conditions for storing 0 and 1 to the cell are illustrated in Fig.4.5 . For storing 0, the top CBRAM should be reset and the bottom CBRAM should be set, which is achieved by applying logic 1 to the the X Y nodes and logic 0 to the write signal. For storing 1, the previous condition is reversed. Namely, logic 0 is applied to the X Y nodes and a logic 1 to the

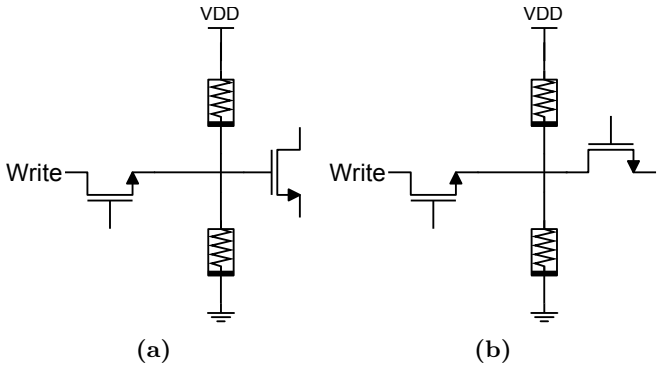


Figure 4.6: (a)NVE configuration node connection in the FPGA switches.(b)NVE configuration node connection in the FPGA LUTs.

write signal. When programming of the cell is concluded, the selection transistor is unselected by applying 0 to the gate and X Y nodes are regarded as VDD and GND terminals during normal operation.

4.1.3.3. NVE-based blocks

In traditional FPGAs, SRAMs are used as the configuration memory (CRAM) to store either the logic functionality of the LB or the routing selection of the RR (routing resource). SRAM-based configuration cells in the FPGA can be replaced with CBRAM-based NVE cells. There are two possibilities for the integration depending how the configuration cell is connected to the circuit. First, NVLUT is introduced by replacing the SRAM cells in the LUT. In this case, NVE is connected to the input of the LUT MUX as in Fig. 4.6b and the NVE voltage divider acts as a resistive load to the critical path. Second, new blocks for routing resource (CB and SB) are designed by replacing the SRAM cells in the blocks with NVEs as in Fig. 4.6a. Since the NVE is connected to the select signals (to the gate of the transistor) of the MUXs, there is no resistive loading to the critical path. In all the cases, significant area reduction is expected due to the compact layout of NVE.

4.1.3.4. NVE Cell Characterization

Table 4.7 shows the area and leakage comparison of NVE to SRAM. The surface areas of the 6T SRAM and NVE are extracted from circuit layouts designed in 130nm

CMOS Bulk technology. The programming transistor shown in Fig. 4.5 is sized based on the required compliance current of the CBRAM devices. The results show that the NVE cell reduces the cell surface by 2.6x. Due to the serially connected resistors in NVE, the leakage current always flows between the power lines. Since the resistors are programmed in complementary values, the leakage current is directly proportional to the R_{OFF} value. Considering an R_{OFF} of 10GΩ forms a leakage current of 150pA at 1.5V due to the HfO_2/GeS_2 technology.

When new information is written to the NVE, the resistors have to reset and store the new value. If an ON/OFF operation follows the the write operation, the total energy consumption is the sum of reset, store, power-down, and restore. Using Eldo, an NVE cell including the CBRAM model is simulated and, then, average energy consumption for each operation is extracted with corresponding duration. Table 4.8 shows the results. For the completion of single write operation including ON/OFF cycle, NVE dissipates 3.42 pJ. This value assumes that new information has to be stored in the CBRAM devices each time before powering down.

When the NVEs are integrated in the FPGA as configuration storage, it is not required to follow all the steps in the ON/OFF cycle. Specifically, the configuration bitstream does not change once it is generated and transferred to the FPGA. Since the same information is retrieved when restoring, it is not necessary to reset and store repeatedly. Hence, only power down and restore operations are taken into account. Considering this characteristic, the total required power for one ON/OFF cycle is reduced to 1 fJ per unit cell as shown in Table 4.8.

4.2. Towards Non-volatile FPGA

As explained in Section 2.1.2., all the blocks in the FPGA are constructed with configurable nodes. These nodes, conventionally being SRAM, need to be programmed with configuration information in order to map the correct application functionality.

Table 4.7: Area and leakage comparison between 6T SRAM and NVE

130nm Bulk	Cell	Area(μm^2)	Leakage current(pA)
SRAM	6T	11.28	25
CBRAM	1T2R	4.00	150

Hence, a bitstream has to be created based on the functionality and transferred into FPGA. Usually, an external non-volatile memory (ex. Flash) stores the bitstream and transfers it to the FPGA upon start-up because the volatile SRAMs lose configuration data when the power source is cut off. With this scheme several cycles are required to load the bitstream and due to the external communication, a high level of power is wasted. Moreover, application security cannot be guaranteed because the external memory is accessible for external attacks and the content might be duplicated which creates vulnerability for valuable intellectual property.

In FPGAs FFs are included in the BLEs. These FFs act as registers which store the result (context) of the operations. Some applications require computations based on temporal evolution. In this case, previously calculated data has to be present in the FPGA. In order not to lose context information, all the calculated results must be transferred to an external storage (ex. Flash) and then retrieved at start-up. The communication to the external unit requires very long cycles and consumes a high level of power. The data in the external memory can also be duplicated by unauthorized access which creates vulnerability as the data might contain sensitive information.

In order to eliminate external storage, Flash cells can be integrated in the FPGA [142]. FPGA vendors such as Actel [143], Lattice [144], and finally Xilinx [145] have announced their flash FPGAs which enclose a hybrid SRAM+Flash architecture where the configuration data is stored in Flash memory. However, the integration of Flash is costly in terms of CMOS process and offers slow operation speeds with a large footprint [146]. Moreover, commercial products only consider configuration data for Flash storage not the context which imposes implications on Flash integration as Flash

Table 4.8: Power data for general NVE operation

	P_{avg}(nW)	Duration(ns)	Energy(fJ)
Reset	1170	500	585
Store	5670	500	2835
Power down	900	0.5	0.45
Restore	1200	0.5	0.6
Energy for general operation including on/off (pJ)			3.42
Energy for FPGA configuration memory on/off (fJ)			1.05

can only withstand 10^4-10^6 write cycles, which is not suitable for frequent context storing.

Resistive memories represent new opportunities and application fields due to their inherent non-volatile properties. In this section, we introduce Non-Volatile FPGAs (NVFPGA) and compare the performance, area, and power figures with SRAM-based FPGA. Two RRAM technologies are targeted and design level benefits unique to these technologies are exploited. First, SRAM-based Configuration Memories (CRAM) in FPGA are replaced with OxRAM-based NVSRAMs and CBRAM-based NVEs for configuration saving applications. Secondly, traditional FFs are replaced with OxRAM-based NVFFs for context-saving applications.

4.2.1. Evaluation on Applications Requiring Configuration Saving

In this section, NVFPGAs with configuration saving property are designed using OxRAM and CBRAM-based memories. Since all the configuration nodes are replaced with non-volatile counterparts, the bitstream can be kept locally in the FPGA. With these FPGAs, there is no need for an external memory to save the bitstream. Thus, following benefits are obtained: 1) The start-up period can be significantly shortened. The FPGAs can be turned on instantly. 2) There is no power consumption due to the bitstream transfer. 3) Bitstream security is granted since there is no external memory.

4.2.1.1. OxRAM-based NVFPGA

In the OxRAM-based FPGA all the configuration memory nodes are replaced with NVSRAM cells. The VPR flow in Chapter 3 is used for evaluation. In order to consider the effects of 22nm LETI-FDSOI process, the architecture definition in VPR is modified accordingly and then changes due to the NVSRAM cell are included. The memory area parameter defined in Section 3.3. is updated with the area estimation calculated in Table 4.2. Due to the two additional transistors in NVSRAM, there is an increase of the routing wire length between LBs. Thus, beyond the direct increase of the memory area, there exists a routing congestion due to the extra transistors. As a result, the addition of two transistors affects all the performance metrics i.e. critical path delay, total area and total power consumption.

The results from the VPR5 flow can be found for the SRAM and NVSRAM-based FPGA on Fig. 4.7, 4.8, and 4.9. Delay and area are increased on an average by 7% and

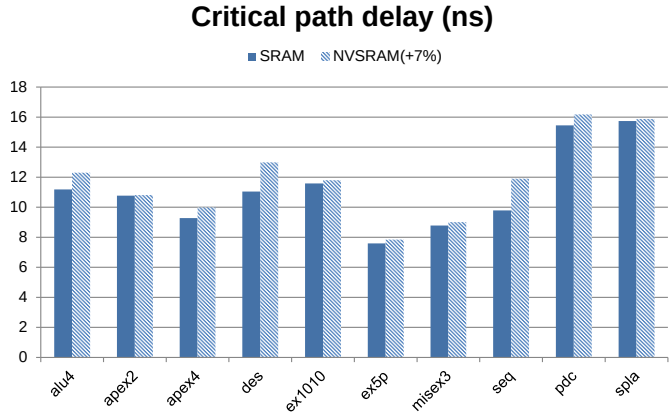


Figure 4.7: Critical path delay of FPGA benchmark circuits for SRAM and NVSRAM integration. The results show an increase in delay on average by 7%.

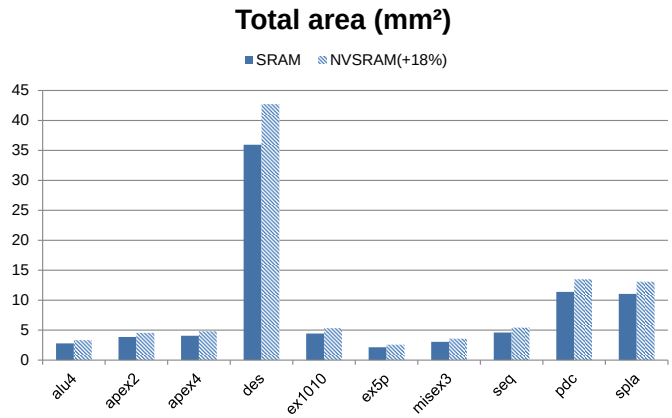


Figure 4.8: Total area of FPGA benchmark circuits for SRAM and NVSRAM integration. The results show an increase in delay on average by 18%.

18% respectively in NVSRAM implementation. Total power consumption is affected by a 2% increase which is due to the effect of more complex routing due to NVSRAM area overhead.

4.2.1.2. CBRAM-based NVFPGA

CBRAM technology is available in 130nm CMOS process. Using the values in Section 4.1.3., the architecture definition in VPR (from Chapter 3) is modified considering 130nm technological parameters and corresponding NVE cell properties. The evaluation on FPGA is carried out in two phases: first the effects of NVLUT are observed by replacing the SRAM-based LUTs with their CBRAM-based NVLUT counterparts and

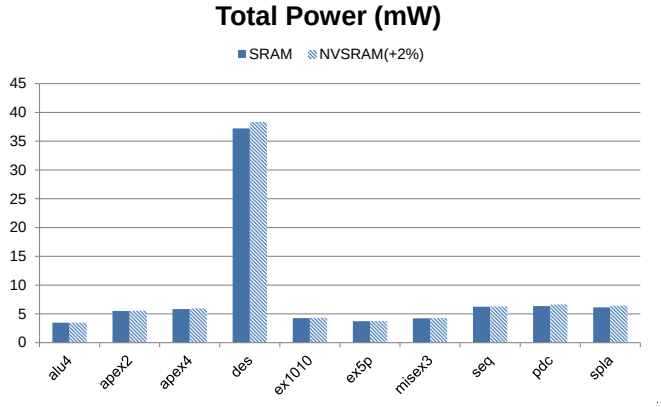


Figure 4.9: Total power consumption of FPGA benchmark circuits for SRAM and NVSRAM integration. The results show an increase in delay on average by 2%.

keeping the rest of the SRAM memory cells. After the NVLUT evaluation, the effects of NVE integration on the entire FPGA are compared with SRAM-based FPGA. For the integration of NVLUT, it is necessary modify the power model in the VPR tool because the leakage profile of the NVE cells are different than that of SRAM cells. In the modified power model, the memory cells are first identified and then the power calculation is updated with the NVE leakage consumption value. Since the memory cells in the routing resources are not affected, SRAM memories are preserved and the leakage consumption value is provided through the architecture definition of VPR. For the evaluation of NVE integration on the entire FPGA, since NVE cells replace SRAM cells globally, the NVE leakage consumption is provided in the architecture definition. Therefore, the SRAM power model in VPR is preserved and only the architecture definiton is updated.

Figs. 4.10, 4.11, 4.12 compare the CBRAM-based FPGA with SRAM-based counterpart. Taking advantage of the compact footprint of the NVE compared to 6T-SRAM cell, the total area of the CBRAM-based FPGA is reduced on average by 5% with NVLUT integration while the NVFPGA achieves 33% better area efficiency. As a result of smaller area, routing wires shorten in length reducing the routing complexity leading to decreased critical path delay and power consumption. NVLUT integration reduces delay by 24% and the NVFPGA reaches 34% reduction. Finally, power consumption is decreased by 18% with NVLUT integration and in NVFPGA 23% lower power is achieved.

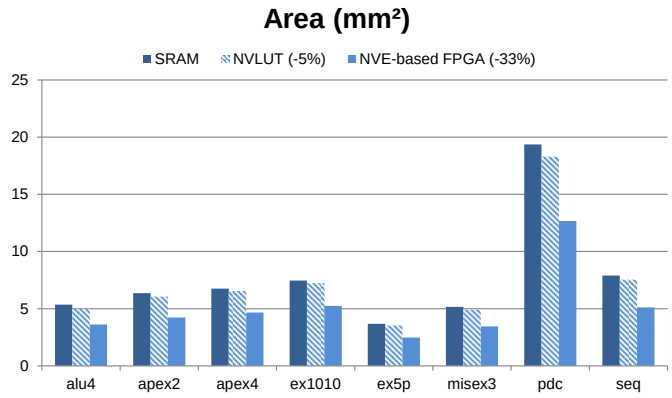


Figure 4.10: Total area of FPGA benchmark circuits for SRAM and CBRAM integration. Reduced area values are achieved by 5% with NVLUT and 33% with NVFPGA.

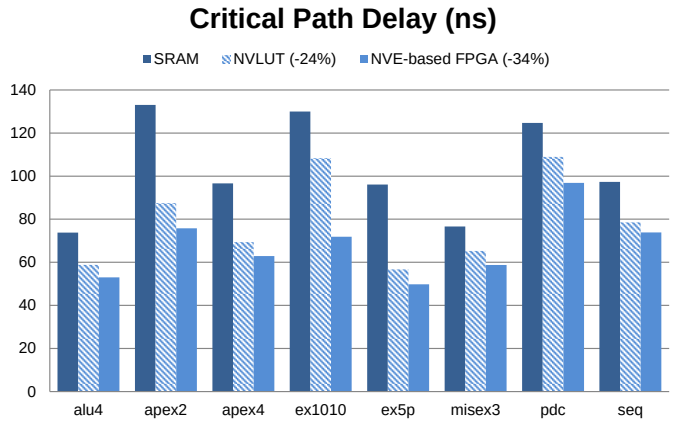


Figure 4.11: Critical path delay of FPGA benchmark circuits for SRAM and CBRAM integration. Reduced critical path delays are achieved by 24% with NVLUT and 34% with NVFPGA.

4.2.1.3. Discussion

By replacing the SRAM in the configuration memory with non-volatile memories, FPGAs with non-volatile property are obtained. Two different technologies and memory designs are analyzed in the FPGA circuit level. By replacing the SRAM-based CRAM with non-volatile counterpart, performance, power consumption, and area (PPA) figures are extracted. It is observed that several trade-offs can be observed with different technologies: 1) PPA: Taking advantage of the compact memory node provided by CBRAM, CBRAM-based FPGA achieves increased performance, reduced area and power consumption. NVSRAM cell has a substantial area overhead due to the

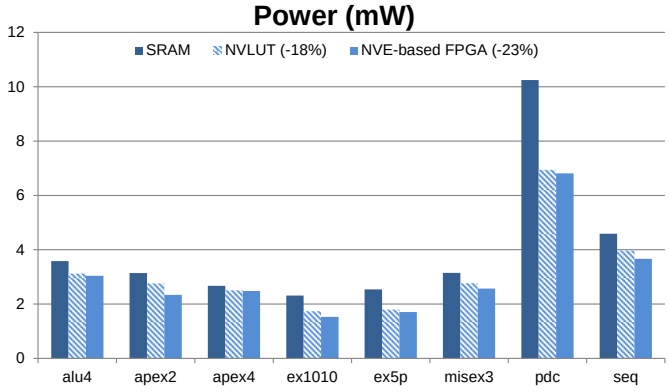


Figure 4.12: Power consumption of FPGA benchmark circuits for SRAM and CBRAM integration. Reduced critical path delays are achieved by 18% with NVLUT and 23% with NVFPGA.

additional control transistor. This overhead results in longer wirelength which increases delay and power consumption. 2) Leakage: The leakage consumption of the NVE, due to the direct path between the source and ground lines, is still significantly higher than that of SRAM. High HRS value is crucial for the limitation of this leakage current. In NVSRAM, on the other hand, since the resistive devices are connected separate nodes, the leakage is similar to the value in SRAM. 3) Write Energy: The write energy of NVE is higher than that of NVSRAM. Normally, the bitstream is stored in the non-volatile memories in the FPGA once and then retrieved from the same memories at each start-up. However, when different applications are considered, the bitstream of each application has to be transferred to non-volatile memories which consumes the write energy. If a configuration-saving FPGA is used for many different applications, the write energy should be considered. 4) Endurance: As discussed in Table 2.1, CBRAM has considerably smaller endurance (10^6) compared to OxRAM (10^{12}). If the FPGA fabric is reprogrammed with many different applications, endurance might limit FPGA utilization.

4.2.2. Evaluation on Applications Requiring Context and Configuration Saving

Some of the applications targeted in the FPGA may require computation based on temporal evolution. Namely, some functions may require previous values to be included in the calculation. As discussed previously, it is possible to retain the configuration in

the NVSRAMs after a sleep period. However, computational results that are stored in the FFs cannot be preserved when switched-off. Therefore, non-volatile elements must be included to save the contents of registers. In the designed OxRAM-based FPGA, SRAM configuration nodes are replaced with NVSRAMs as explained in Section 4.2.1.1. and NVFFs are integrated in place of the regular FFs in the LBs. Therefore, configuration and context saving feature is established.

Similar to OxRAM-based NVSRAM, 22nm LETI-FDSOI process is considered for NVFF integration. The architecture definition created in Section 4.2.1.1. including NVSRAM changes is updated to reflect the modifications imposed by NVFF integration and, thus, context-saving feature is introduced. The created architecture definition is used with VPR5 toolflow explained in Chapter 3. For the evaluation of NVFF integration, sequential circuits from MCNC benchmarks (explained in Section 3.2.3.). Previously, in the configuration saving applications, combinational circuits were used in which the FFs in the LBs are bypassed by the MUXs. In order to observe the impact of NVFF, sequential circuits from MCNC benchmarks are assessed where the FFs are activated.

The results from VPR5 evaluation flow can be found in Figures 4.13 – 4.15 for critical path, area, and power metrics. For sequential benchmarks, NVSRAM integration presents similar overhead as the combinational ones resulting in 6%, 17%, and 1,5% increase in delay, area and power respectively. Proceeding with the NVFF integration results 3%, 1%, and 1% delay, area and, power overhead respectively. Since the NVFF has a larger area than a FF due to the extra 6 transistors as described previously, the routing resource becomes more complex and, hence, there exists an influence on delay and power parameters. Moreover, the NVFF has a slightly higher Clk-to-Q delay than a regular FF, due to the additional capacitance from the control transistor, which affects the final critical path delay metric. The overall effect is smaller compared to NVSRAM integration because the number of employed FFs is relatively smaller than that of SRAMs.

4.2.3. Optimization of Resistance States for NVFPGA

As explained previously, the technologies, OxRAM and CBRAM, offer different properties which affect the design of the memory architectures. In the NVSRAM as shown in Fig. 4.1, the nonvolatile devices are connected to the data nodes of SRAM

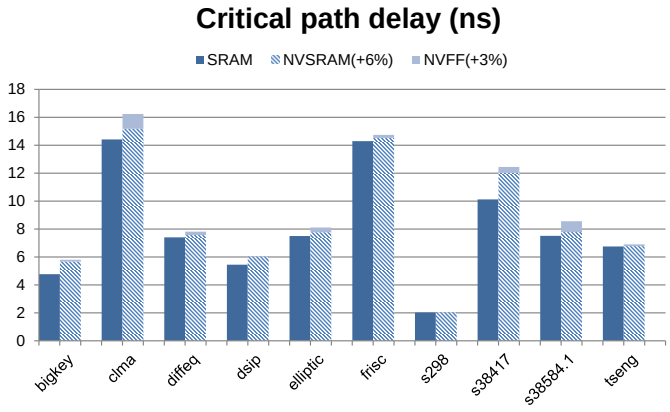


Figure 4.13: Critical path delay of FPGA benchmark circuits for SRAM, NVSRAM and NVFF integration. The results show an increase in delay on average by 6% in NVSRAM and 3% in NVFF implementations.

cell. Since the access transistor of the RRAM devices are switched off during normal operation, circuit performance is not affected by the resistive devices and the leakage current flowing on the RRAMs is limited to the subthreshold leakage of the access transistors. This leakage can be further reduced with the use of high-V_t transistor for the access transistor. Due to the CBRAM-based NVE structure, on the other hand, there are significant impacts in the leakage consumption and critical path delay.

As explained before, the NVE is either connected to the gate of the transistors in the switches (Fig. 4.6a) or the source of the transistors in the LUTs (Fig. 4.6b). When it is in the switches, it has no impact on the performance but in the case of LUTs, it affects the delay of the LUTs. As explained before, the top and bottom resistive devices are programmed in complementary values, i.e. at any time one of them is programmed with high resistance (R_{OFF}) and the other with low (R_{ON}). The R_{ON} of the NVE determines how long it takes to charge or discharge the configuration node of the LUTs. Thus, the lower the R_{ON} the higher is the operating speed.

In the NVE structure during normal operation, regardless of where it is connected for configuration, there is a direct current path between the supply and ground sources. Considering the complementary states of the resistances, a resistance value of $R_{on} + Roff$ lies between the supply and ground voltage sources. The leakage current becomes;

$$I_{leakage} = \frac{V_{dd} - V_{ss}}{R_{ON} + R_{OFF}} \tag{4.1}$$

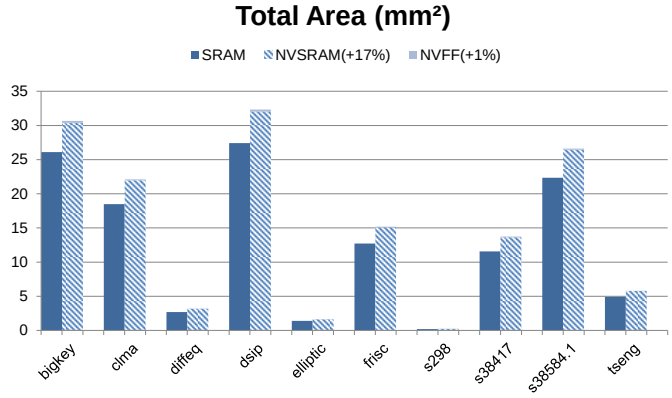


Figure 4.14: Total area of FPGA benchmark circuits for SRAM, NVSRAM and NVFF integration. The results show an increase in delay on average by 17% in NVSRAM and 1% in NVFF implementations.

Assuming that $R_{OFF} \gg R_{ON}$ and $V_{ss} = 0$, the leakage equation simplifies to;

$$I_{leakage} \approx \frac{V_{dd}}{R_{OFF}} \tag{4.2}$$

The impact of R_{ON} on delay is investigated based on the LB delay. Using evaluation platform in Chapter 3, several architecture files are created. Benchmark circuits are evaluated with varied memory resistance values and average critical path delay for each resistance value is reported. Fig. 4.16 shows the resistance impact on the critical path delay. For increased performance, the trend from the results confirms the expectation that the resistance should be lowered for increased performance. It can also be noticed that even with very high resistance of 100kΩ, the performance of NVE-based FPGA is still 23% better than the SRAM counterpart. This can be explained with the area advantage of NVE-based implementation which is still higher than the diminishing effect of high resistance.

R_{OFF} impact on leakage current is examined by using the equation in 4.2. Table 4.9 shows the resistance pair examples reported in the literature with different technologies and corresponding leakage current values in NVE configuration assuming $V_{dd} = 1.5V$. Depending on the technology, the achievable R_{OFF} value changes significantly and the effect of the R_{OFF} can be observed clearly as it is the dominating factor. The highest R_{OFF} is offered by CBRAM technology as 10GΩ with a leakage as low as 15nA per NVE cell.

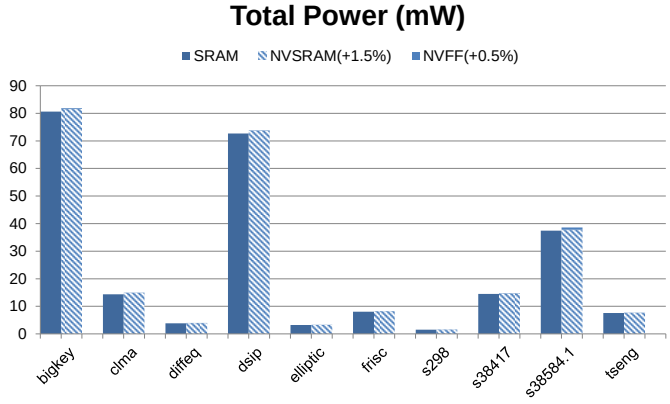


Figure 4.15: Total power consumption of FPGA benchmark circuits for SRAM, NVSRAM and NVFF integration. The results show an increase in delay on average by 1.5% in NVSRAM and 0.5% in NVFF implementations.

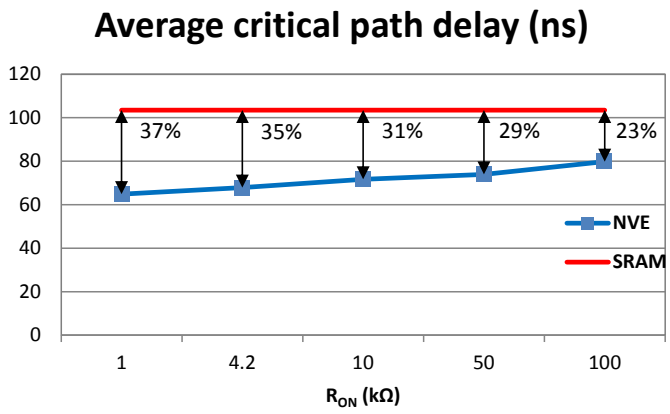


Figure 4.16: R_{ON} impact on FPGA critical path. Gain is reduced with increasing resistance value.

Several R_{OFF} values are analyzed in order to observe the impact on FPGA power consumption. Leakage currents are calculated for R_{OFF} values between 1MΩ and 10GΩ. Several architecture files are created with these values for VPR evaluation. For each R_{OFF} value, total power consumption values, including dynamic and leakage, are extracted and they are normalized to the 10GΩ case. Fig. 4.17 depicts the obtained results. As expected when the R_{OFF} is reduced, the leakage consumption increases and for lower R_{OFF} values, leakage consumption becomes the dominant factor in total power consumption. Starting from 10GΩ, at 1GΩ total power consumption remains without significant change however at 1MΩ, power consumption increases by 60x.

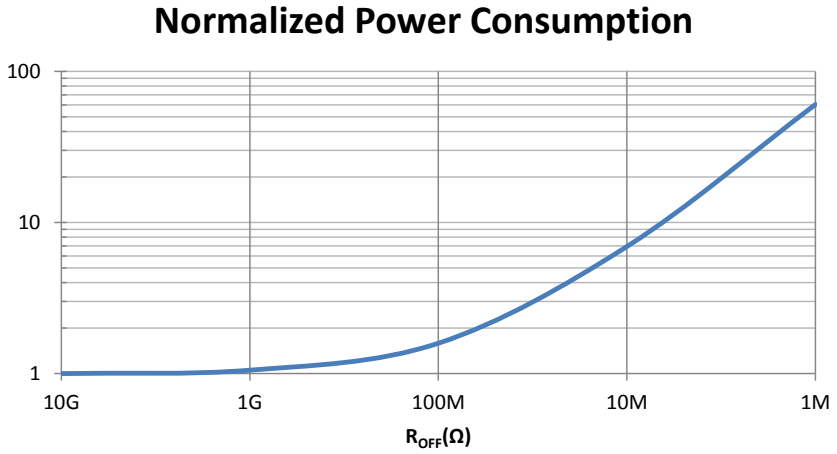


Figure 4.17: R_{OFF} impact on FPGA total power consumption. Power consumption increase with reduced resistance value.

This study examines several trade-offs of the resistance states over critical path delay and power consumption values. Depending on the application requirement and power budget, the resistance values and the targeted technology can be selected. Thus, with this study, the process engineers can get a feedback for the selection of material stack of the NVE device to fulfill the FPGA requirements.

4.3. Normally-OFF Instantly-ON Computing

Many new embedded applications can be characterized as "Normally Off, Instantly On". Video surveillance, smart grids distributed sensors or healthcare monitoring systems are some examples of such applications. They share a common feature: a long idle period followed by a short highly intensive computing phase. A small circuitry can be reserved in the chip to generate the wake-up signal by observing the events with an appropriate sampling frequency. When incoming burst of events happens, wake-up

Table 4.9: Resistance state values reported in the literature.

	OxRAM[147]	CBRAM[73]	PCRAM[148]
R_{ON}	1kΩ	4.2kΩ	10kΩ
R_{OFF}	100kΩ	10GΩ	2MΩ
$I_{leakage}$	15μA	0.15nA	0.75μA

process should be completed as quickly as possible in order to avoid missing important information after the arrival of the wake-up signal. For example, a disruptive event on power grids must be quickly analyzed by reacting in few milliseconds. Similarly, a heart attack can be anticipated by analyzing electrical signals after an abnormal event. Especially, avoiding false detections is particularly challenging due to the high number of measurements that needs to be processed in a few microseconds after a wake-up event.

The solutions are generally embedded, frequently battery-powered, and distributed on large areas. Their computing performance requirements tend to increase with higher complexity of applications (e.g. aggression detection for video surveillance). On the other hand, these systems being pervasive in our environment and, thus, more and more numerous, static power consumption must be minimized during sleep periods, while keeping the dynamic power consumption as low as possible.

General Purpose Processors (GPP) are not efficient for these computing intensive applications as they offer reduced power efficiency and require long boot sequences when switching into 'ON' states [149]. Application Specific Integrated Circuits (ASIC) or System-On-Chip (SoC) are power efficient but are expensive to develop and not flexible enough to address a wide range of applications. Field Programmable Gate Arrays (FPGA) offer a good compromise between flexibility and power efficiency, as they are flexible enough to accommodate a large number of applications and have reduced dynamic power consumption compared to GPP [150]. Moreover, they can be designed using high-end technologies to provide the required performance. However, as the supply voltage decreases and the feature size becomes smaller, the leakage current of FPGA increases very rapidly [151]. As a result, the leakage is the major source of power consumption and reduce FPGA field of applications.

Several techniques have been proposed to achieve low-power FPGAs. Meng et al. studied the use of drowsy modes to reduce leakage in the embedded memory [152] and proposed a method that is applicable to the block memories in the FPGA. However, since the configuration memory is distributed, the granularity of the control circuitry imposes a high area overhead. Lewis et al. proposed power management schemes featuring body biasing [153] while Li et al. evaluated the effects of dual-V_{dd} and dual-V_t fabrics on power consumption [154]. These solutions require technology dependent adjustments and it is still not possible to have full power down mode.

In order to improve total power consumption, power-gating technique is massively employed [155]. It can help decrease the leakage current by a factor of more than 1000; however, in the case of FPGA, all the information contained in SRAM memories is lost when switched off. To solve this issue, FPGAs can be associated with external non-volatile Flash memories to store their context. In such a case, restoring a context after a power-off stage is accomplished by serially loading a bitstream to all the configuration cells of the FPGA. This process is quite long and can take up to hundreds of milliseconds for a large FPGA. Moreover, for calculations requiring a history of data, information stored in the registers is lost after an ON/OFF cycle. For such applications, the results must be stored externally and they need to be restored after every power-off operation. Hence, due to external storage communication overhead, switching-off FPGA is not compatible with fast wake-up and requires very long idle periods. As explained in Section 4.2.1., Flash memories can be integrated in FPGAs to reduce wake-up period. Lattice reports that when Flash memory is integrated in the FPGA, the transfer of configuration from Flash to SRAM takes 1ms. Even though, the wake-up time is significantly reduced, it is still very high for instant-on feature. Consequently, current FPGA structures are not compatible with "Normally Off, Instantly ON" application constraints.

4.3.1. Power-gating Implementation

4.3.1.1. Power Overhead - Duty Cycle Relation

In an application where FPGA is switched into sleep mode, the dissipated leakage power will be conserved. However, it is necessary to take into account the power overhead due to RRAM integration and power switch implementation. In Fig. 4.18, P_1 and P_2 refer to anticipated power consumption of SRAM and RRAM-based architectures, P_L is leakage power consumption of SRAM configuration, t_0 and t_1 are the ON and standby durations respectively. In an activity when the consumed leakage energy during t_1 exceeds the energy overhead due to the power overhead ($P_{OH} = P_2 - P_1$) spent during t_0, the leakage energy for a smaller duty cycle will be conserved.

Initially, as in Eqn. 4.3, a ratio between t_1 and t_0 can be defined when the leakage

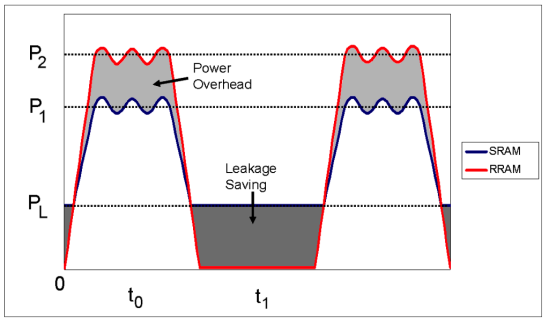


Figure 4.18: Conceptual view of potential gain and power overhead of SRAM and RRAM-based implementations. During standby mode, the circuit with SRAM consumes leakage power, whereas it is possible to reduce consumption in the same mode to zero with RRAM integration.

energy in t_1 duration is equal to the energy spent during t_0 due to the power overhead:

$$\int_{t}^{t+t_0} P_2 - P_1 \, \mathrm{d}t = \int_{t+t_0}^{t+t_0+t_1} P_L \, \mathrm{d}t \tag{4.3}$$

$$\frac{t_1}{t_0} = \frac{P_2 - P_1}{P_L} \tag{4.4}$$

The ratio given by Eqn. 4.4 defines the minimum t_1 in terms of t_0 when the energy overhead is in equilibrium with the gained leakage energy. It is possible to relate this ratio to duty cycle, which is defined as the ratio of 'ON' duration to the total period. Specifically, in an on/off application, break-even time (BET) indicates the duty cycle when the energy overhead is equal to the leakage energy. In this application, it is calculated based on the following equation:

$$BET = \frac{t_0}{t_0 + t_1} = \frac{1}{1 + \left(\dfrac{P_2 - P_1}{P_L}\right)} \tag{4.5}$$

Eqn. 4.5 defines the break-even time (BET). When the actual duty cycle of the application is smaller than the BET, it is possible to save leakage power and reduce the total power consumption. For applications having a duty cycle larger that BET, total power gains are observed. The gain can be defined as follows:

$$\begin{aligned} Power\,Gain(\%) &= \frac{P_L\, t_1 - (|P_2 - P_1|)t_0}{P_1 t_0 + P_L t_1}\, 100 \\ &= \frac{P_L - (|P_2 - P_1|)D}{P_1\, D + P_L}\, 100 \end{aligned} \tag{4.6}$$

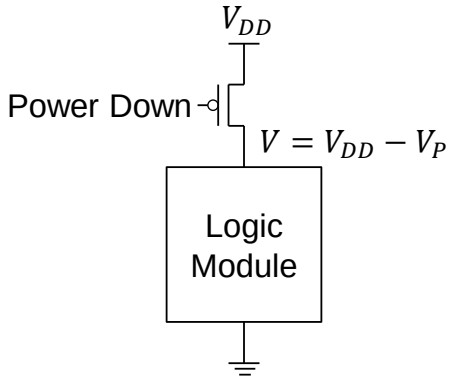


Figure 4.19: Power gating switch utilization for power down mode.

where $(|P_2 - P_1|)$ denotes the total power consumption difference between RRAM and SRAM-based FPGA implementations, (t_1, t_2) the on/off durations, $D = t_0/t_1$ the duty cycle, and P_L SRAM configuration leakage power consumption.

4.3.1.2. Power Gating Cost

If a logic circuitry is considered in which a PMOS switch, as in Fig. 4.19, provides the power gating functionality, the absolute source voltage available for the logic operation is decreased due to the voltage drop on the switch transistor. If the highest available VDD is not increased to compensate the effect, the lowered source voltage will affect the performance of the logic circuit. The voltage drop is directly proportional to the size of the switch. If the area of the switch is increased, its resistance will be decreased, thus, the voltage drop will be decreased and the performance will suffer less. Therefore, it is possible to define a trade-off between switch size and performance of the circuit, i.e. the maximum frequency of the circuit.

Assuming a switch, with an area of 'A_s', driving a load of area 'A', and providing current 'I', the trade-off between area and performance can be defined as [156]:

$$\Delta f = \alpha \frac{I}{A} \frac{A_s}{g_{on}} \frac{V_{DD}}{V_{DD} - V_{TH}} \frac{1}{\Delta A} \tag{4.7}$$

where g_{on} is the conductivity of the switch, α is the velocity saturation index, V_{DD} is the supply voltage, and V_{TH} is the threshold voltage of the switch. Δf and ΔA refer to the change in frequency and area. With this equation, it is possible to observe the effect of switch strength on operating frequency.

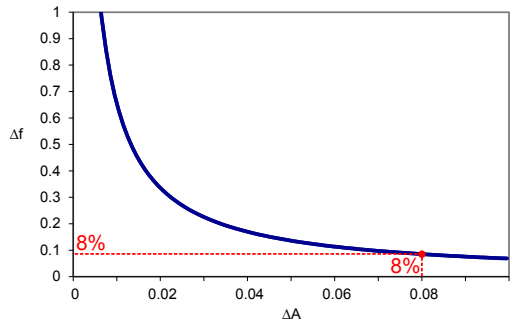


Figure 4.20: Operating frequency/area trade-off in power gating application with 22nm FDSOI technology.

In the trade-off defined in 4.7, I/A and g_{on}/A_s provide area independent ratios. In other words, as defined in [156], a module with area of 'A' and consuming current 'I', will have the same trade-off as another module which has K times bigger area and K times higher current consumption. In the NVSRAM cells, write current is critical in order to obtain successful operation. Hence, for the sizing of the power switches, write current of the NVSRAM cells are considered as the worst case. Consequently, one NVSRAM cell is simulated in Eldo and the required typical current is obtained as 12μA.

Finally, the performance/area trade-off is calculated for 22nm technology using Eqn. 4.7 as shown on Fig. 4.20. From the figure, it can be observed that the rate of change of area-frequency trade-off decreases as the area increases. After 8% of area overhead, increasing the area does not reduce the frequency overhead significantly. Therefore, 8% area overhead is considered which results in 8% frequency overhead for the power gating application in the RRAM-based FPGA.

4.3.2. Normally-OFF Instantly-ON FPGA

In Normally-off Instantly-on applications, the computation unit is powered off when it is not in use and it is turned on very fast to continue computing. Due to the volatile memories and registers, these applications cannot be addressed in traditional FPGAs. The proposed FPGAs in Section 4.2., on the other hand, can fulfill the requirement of these applications. Moreover, with the integration of non-volatile memories, the leakage consumption can be preserved by introducing a zero-leakage state. In this section

Table 4.10: Break-Even Times (BET) for OxRAM-based Configuration-Saving Applications

Circuit	P_{OH} (mW)	$P_L(\mu W)$	BET (%)
alu4	0.03	24.97	48.24
apex2	0.04	37.90	50.64
apex4	0.11	46.66	30.46
des	1.12	994.48	47.01
ex1010	0.04	39.89	47.08
ex5p	0.04	24.63	36.17
misex3	0.04	37.12	46.66
seq	0.05	50.22	48.06
pdc	0.28	129.26	31.36
spla	0.29	120.07	29.62
		Avg.	42.26

previously designed NVFPGAs are analyzed in Normally-off Instantly-on computing based on the requirement of application activities for increased power gain.

4.3.2.1. OxRAM-based Normally-OFF Instantly-ON FPGA for Configuration-Saving Applications

When a power on/off application is considered, with the use of non-volatility, leakage power is gained during sleep mode. As mentioned earlier, Break-Even Time (BET) indicates the activity ratio necessary for power savings. In order to observe the power gain resulting from switching-off the FPGA, the BETs for different benchmarks are calculated using Eqn. 4.5. Table 4.10 shows the power overhead (P_{OH}), leakage power of SRAM configuration (P_L) and the BET for each of the benchmark circuits. The BET values range between 30% and 51% with an average of 42%. If the targeted FPGA application has an activity time smaller than 42%, the leakage power is reduced and the overall power consumption is decreased.

The duty cycle of the application has a direct influence on the conserved power levels. Fig. 4.21 shows that for the on/off application, if the duty cycle is much lower than the BET, the power gain increases rapidly. The zero crossing of the curves gives the BETs of the circuits. Starting from 42%, if the duty cycle is reduced to 10% and

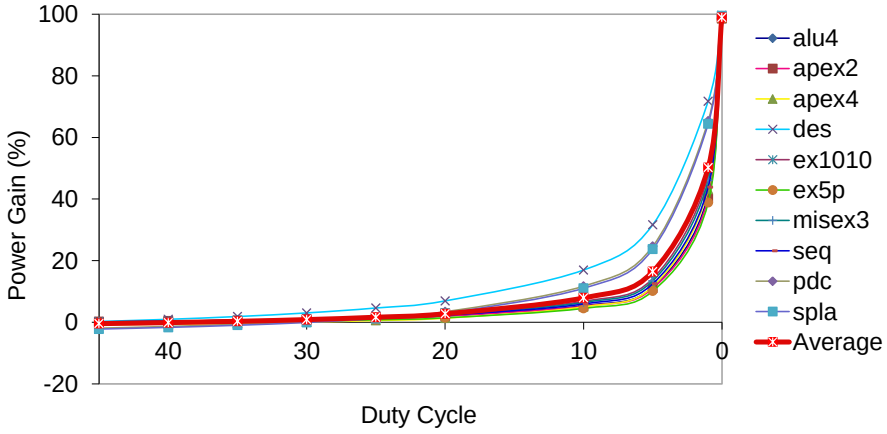


Figure 4.21: Power gain depending on duty-cycle values for OxRAM-based FPGA with configuration saving applications. Considering 1% ON time, gained power reaches 50% on average.

5%, power gains of 7% and 16% can be obtained respectively. For a specific application where the circuit is effective for 1% of the time, power gain reaches 50% on average. If the on time is further reduced, higher power gain can be attained.

4.3.2.2. CBRAM-based Normally-OFF Instantly-ON FPGA for Configuration-Saving Applications

As explained previously in Section 4.2.1.2., small footprint of the NVE brings area efficient design opportunities in the FPGA. In 4.2.1.2., it has been demonstrated that due to reduced routing congestion as a result of the decreased total area figure, the total power consumption was reduced by 23%. Since there is already a positive power gain at full activity (100% duty cycle), BET values, in definition, do not exist when Normally-Off Instantly-On Applications are considered. Depending on the activity of the circuit, the power gain can be further increased by conserving the leakage consumption which is wasted during inactive periods. Using Eqn. 4.6, power gain values are calculated based on circuit activity. Fig.4.22 shows that for applications having 50% duty cycle, the total power consumption can be reduced by 44%. Power gain reaches more than 97% when the application is active for 1%.

By replacing the SRAM in the configuration memory with non-volatile memories, FPGAs with non-volatile property are obtained. Taking advantage of the compact memory node provided by CBRAM, in comparison to OxRAM-based FPGA, CBRAM-

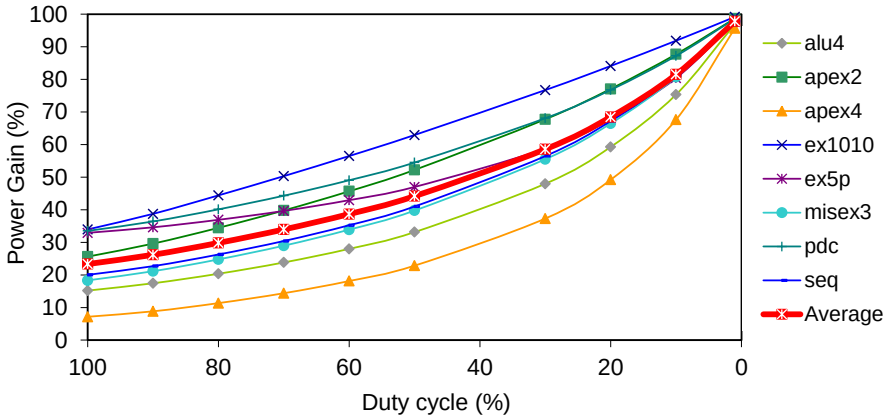


Figure 4.22: Power gain depending on duty-cycle values for CBRAM-based FPGA with configuration saving applications. Considering 1% ON time, gained power reaches 97% on average.

based FPGA achieves better performance, area, and power consumption figures. Due to the implications of area overhead in NVSRAM, OxRAM-based FPGA has less power gain while fulfilling the instant-on requirement with 20ns wake-up time. The simple structure of CBRAM-based NVE ensures faster wake-up operation than NVSRAM as the NVSRAM requires a sequence to restore the values to the SRAM. However, when voltage supply is switched on, NVE is restored instantly in 0.5ns. Applications having at most 42% duty cycle can benefit from the OxRAM-based FPGA. On the other hand, CBRAM-based FPGA offers power gain for any application even for full-time activity (100% duty cycle).

4.3.2.3. OxRAM-based Normally-OFF Instantly-ON FPGA for Configuration and Context-Saving Applications

With the previously assessed FPGAs, only configuration can be stored in the FPGA for Normally-off Instantly-on applications. However, for applications requiring temporal evolution, the calculated results in the FFs are lost when the FPGA power is cut-off. With the integration of NVSRAM and NVFF, the OxRAM-based FPGA designed in Section 4.2.2. accomplishes both configuration and context saving property which is a requirement for fast power on/off cycles.

For the evaluation of power gain in Normally-off Instantly-on applications, the results from Section 4.2.2. are analyzed. After extracting the leakage power and power

Table 4.11: Break-Even Times (BET) for OxRAM-based Configuration and Context-Saving Applications

Circuit	P_{OH} (mW)	$P_L(\mu W)$	BET (%)
bigkey	0.96	721.30	42.95
clma	0.55	181.96	24.92
diffeq	0.09	21.93	19.28
dsip	0.99	727.92	42.26
elliptic	0.02	10.81	38.71
frisc	0.16	117.57	43.04
s298	0.01	3.47	24.66
s38417	0.13	109.74	45.14
s38584.1	0.34	325.27	48.89
tseng	0.09	95.02	52.69
		Avg.	38.25

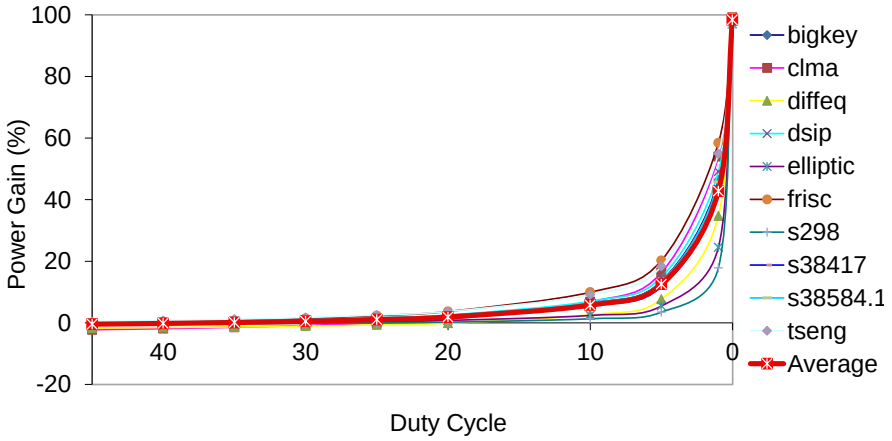


Figure 4.23: Power gain depending on duty-cycle values for OxRAM-based FPGA with configuration and context-saving application. Considering 1% ON time, more than 40% power gain can be achieved.

overhead of the sequential benchmarks including the NVSRAM and NVFF integrations, the BET values are calculated for the ON/OFF application. It is assumed that perfect clock gating is already present in the SRAM-based FPGA; i.e. there exists no performance overhead due to the extra control logic, which is necessary to prune the clock tree. Consequently, compared to a volatile FPGA, our solution provides a massive reduction of total power consumption. Table 4.11 show the calculated BETs using Eqn. 4.5. The BET values range between 20% and 53% with an average of 38%. Specifically, depending on the application, if the duty cycle is at most 38%, total power consumption can be reduced due to the conserved leakage power. After the BETs are obtained, the power consumption for several duty cycles are calculated with Eqn. 4.6 to observe their effect on power saving. Fig. 4.23 shows that the gained power increases rapidly for lower duty cycles because more leakage power is conserved. Starting from 38% reduced duty cycles of 10% and 5% guarantee 6% and 13% power gain, respectively. For applications remaining inactive for most of the operation period, activating the FPGA for 1% of the time and powering down for the rest, provide more than 40% power gain.

4.4. Conclusion

In this section, several NVM-based architectures have been discussed. Two emerging memory technologies are considered: OxRAM and CBRAM. With OxRAM, NVSRAM and NVFF cells are analyzed. These cells enable non-volatile functionality to regular SRAM and FF designs. With CBRAM NVE and NVLUT are analyzed. It is observed that the technological properties have a significant impact on the elementary circuit architecture. The high R_{off} granted by CBRAM gives the opportunity to design the NVE in a very compact surface. The circuits are, then, characterized in terms of area and power consumption. They are integrated in the FPGA by replacing the corresponding blocks and several NVFPGA fabrics are designed. FPGA employability is increased with the utilization of non-volatility. NVE-based blocks and NVSRAM-based blocks allow storing the bitstream on the FPGA which accomplishes the configuration saving feature. NVFF integration enables storing the computation results in FPGA which establishes context-saving property. Moreover, with non-volatility, low-power FPGA functionality can be achieved. For this purpose, a very efficient power management

scheme, power gating, is integrated by taking advantage of non-volatility. It has been demonstrated that the power consumption can be reduced up to 97% for Normally-off Instantly-on applications with the proposed NVFPGAs. The results from this chapter are published in [141][157][158][159][160].

5

3D-FPGA with Monolithic Integration

Contents

3D Monolithic Integration (3DMI) brings many opportunities with the possibility of high density interconnects. 3DMI shortens the interconnect wires by allowing placement of cells on the critical path close to each other and increases the device density by integrating more devices on the same footprint. Therefore, migration to 3D addresses both the interconnect domination issue and logic density enhancement need in modern ICs.

In the past, FPGAs have always relied on advanced technologies for better performance. The regular tile-based architecture of FPGAs allows fast adoption of advanced nodes. In this case, 3DMI can bring unprecedented advantages to FPGAs. The main problem to be addressed with 3DMI design is the utilization rate of stacked layers due to partitioning for optimized performance. It is expected in the current 3DMI technology that only 2 active layers are integrated because of the complex fabrication process. However, we can also imagine the integration of more than 2 layers as the technology matures. Therefore, careful partitioning has to be analyzed considering multi-tiers to gain the highest benefit from 3DMI.

In this chapter, several 3D FPGA blocks are designed using 3DMI technology. As explained in 2.2.1., the main bottleneck of the FPGAs is the high utilization of configuration memories. Since these memories are distributed, they require a large number of interconnects. Taking advantage of high-density interconnects in 3DMI, logic-on-memory technique is adopted for the design of 3D cells. 2D FPGA cells are also included in the design library for comparison and design space exploration. Starting from 2 layers up to 4 layers, several 3D-FPGA circuits are analyzed and evaluated with varying partitioning schemes. Based on the results, 3DMI demonstrates to be very efficient in area, performance and power improvements for FPGAs. Furthermore, the gains from stacked layers with 3DMI outperform the expected scaling gains, which proves that 3DMI can be an alternative for future scaling.

The chapter is organized as follows: Section 5.1. briefly explains the 3DMI fabrication steps. In Section 5.2., the logic-on-memory approach is explained, 3D blocks are designed and characterized, and a 3D-FPGA is assessed with the proposed approach. In Section 5.3., multi-tier FPGAs are designed and evaluated. Section 5.4. compares traditional scaling and 3DMI results. Concluding remarks are given in Section 5.5..

5.1. 3DMI Technology

3DMI allows sequential fabrication of several active layers on the same die. The fabrication is achieved with the following steps (Fig. 5.1): fabrication of bottom transistor, top film deposition, and the realization of top transistor.

In the first step, the transistors are fabricated on the bottom layer and standard high temperature spike anneal at 1050°C is applied for dopant activation [161]. Due

to the implications imposed by the integration of the second active layer, the thermal robustness of the bottom layer has to be guaranteed. The thermal budget for subsequent steps is limited by the stability of the salicidation step. In order to avoid additional dopant diffusion or interfacial oxide growth, top FETs must be processed under 600°C. Considering this low temperature limit, an approach is established with incorporation of platinum together with flourine and tungsten implantation to enable stable Ni-based salicide [162].

High quality active layer has to be deposited for better performance of the top transistor. Low temperature molecular bonding process (at 200°C) allows obtaining pristine crystalline quality and accurate thickness control. Before bonding, a thin Inter Layer Dielectric (ILD) is deposited and planarized on the top of the bottom transistors. At this step, the intermediate metal layers are realized. Typically, the integration of 2 metal layers is foreseeable. Molecular bonding is, then, performed for full transfer of monocristalline Silicon layer. Silicon layer thickness can be as thin as 10nm and the ILD can be reduced down to 23nm [163].

The top transistor thermal budget is limited to 600°C in order to retain the performance of the bottom layer transistor. The most critical thermal budget is the dopant activation step. The Solid Phase Epitaxy Regrowth (SPER) allows replacing the high temperature spike dopant activation anneal. Currently, SPER is the most mature way which can reduce the temperature for dopant activation below 600°C. In this technique, the dopant activation occurs during the recristallization of an amorphized semiconductor region.

In the last step, 3D contacts are connected and the BEOL metal layers are integrated. Several metal layers are realized same as in 2D BEOL. For contact realization, a single lithography can be used between top and bottom transistors. In this case, a highly-selective etch is used to open contacts reaching down to the bottom layer without passing through the upper layer [164]. Consequently, equal lithography alignment performance for top and bottoms is obtained which grants no additional enclosure ensuring high density vertical interconnects. The pitch of the vertical vias can be as small as 100nm.

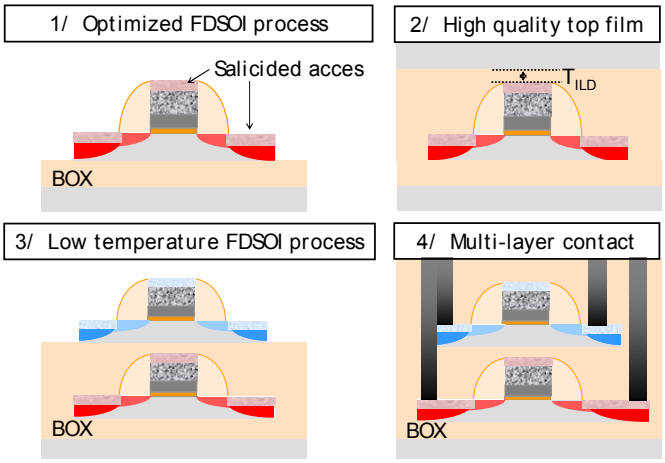


Figure 5.1: Description of the process flow enabling to achieve stable performance bottom FET, high quality top substrate, high performance top FET with 600°C process and 3D contacts realization.

5.2. 3DFPGA with Logic-on-Memory Approach

5.2.1. 3D Design Implications

One of the main focus with 3DMI is the partitioning planning while taking advantage from high density interconnects. In order to exploit these benefits, recent works focus on transistor-on-transistor [90] and gate-on-gate 3D integration [92][93]. Transistor-on-transistor approach improves the total area, critical path delay and power. However, well balanced NMOS and PMOS transistor footprints are necessary in order to reach highest benefits. As will be discussed in the next section, a large portion of the FPGA is built with NMOS-only MUXs. Therefore, for the proposed FPGA, gate level partitioning is preferred.

Technological and cost-related challenges must be taken into consideration during design with 3DMI. Fabrication of the bottom layer is classical but the top transistors, even if performed in a cold process, imply technological difficulties on the intermediate metal layers. Moreover, due to cost, the total number of available intermediate metal layers must be limited. Typically, integration of one or two metal layers is a reasonable technological target. Additionally, 3D connections are performed by vias which can be fabricated with a low pitch. Here we assume a pitch between vias of 100nm given by realistic technological rules. This figure shows that density of 3D connections in 3DMI

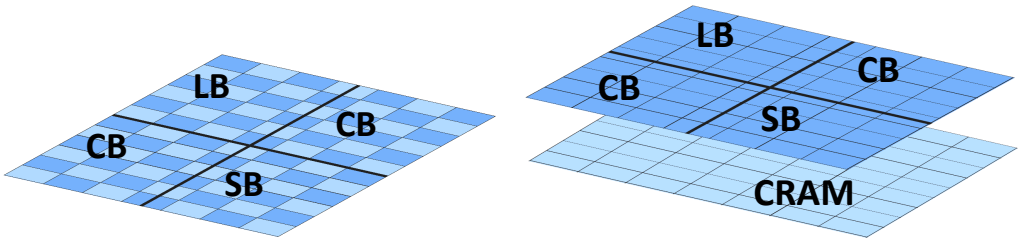


Figure 5.2: Logic-on-Memory approach

is at least of one order of magnitude higher than its 3D TSV counterpart.

As described in Section 2.2.1., almost half of the FPGA is made of memories. In consideration of this characteristic, we determine the 3DMI partitioning as follows: The bottom layer contains SRAM cells while the top, computing and routing resources. For keeping a good global performance, SRAM cells must be entirely integrated on the bottom layer which leads to the choice of two intermediate metal layers. A second target is on the design side: a good equilibrium between the two layers must be reached while retaining modularity and scalability capacities of FPGA for performance and low area purposes. To fulfill the first requirement, a coarse grain top and bottom layer co-design must be carried out, while the second constraint leads us to keep the classical FPGA partitioning for design optimization, i.e. LB, SB and CB.

Logic-on-memory approach is depicted in Fig. 5.2 in which the distributed CRAM in 2D tile is placed into bottom layer in the 3D tile. Logic-on-memory partitioning scheme helps attaining following gains: 1) CRAM is kept close to configuration nodes which reduces the demand on routing wires. 2) Since all the memory cells are in the bottom layer, the cell layout can be optimized with strict memory design rules for decreased footprint and high-V_t cells can be utilized without area penalty for further leakage reductions. Also, the programming circuitry can be integrated in the bottom layer which decreases the routing complexity due to programming wires. 3) Since all the routing resource is placed on the top layer, no critical path signal crosses between layers which allows avoiding the effect of inter-tier vias.

5.2.2. 3D-FPGA Blocks

All the blocks in the FPGA are designed using the 3D 14nm FDSOI process. First the layouts of 2D blocks are prepared and then by placing the Configuration Memory

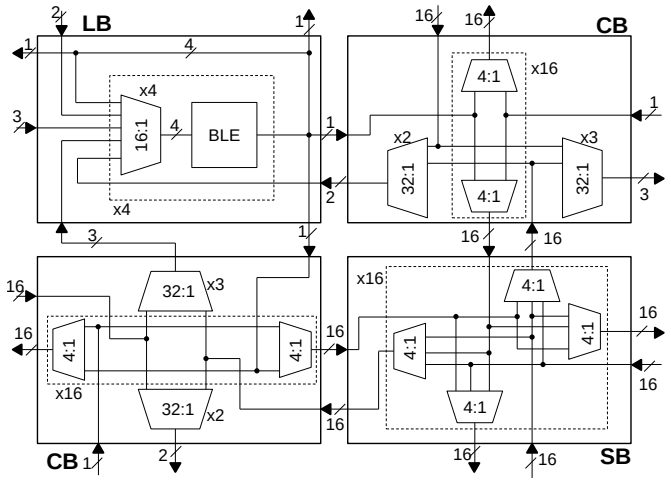


Figure 5.3: Tile block diagram.

(CRAM) cells into the bottom layer, 3D blocks are obtained. All of the remaining logic such as multiplexers and buffers, are placed into the top layer.

Fig. 5.3 shows the block diagram of the tile design assumption for the 2D and 3D FPGAs. LB is constructed with a cluster of 4 BLEs which contain 4-input LUTs (LUT4). Thus, N(number of BLEs in LB)=4 and K(number of input to LUT)=4. Each LUT4 input requires one 16-input MUX (MUX16) for selection from LB inputs. The optimized input number to the LB (I) is achieved with $2 \times N + 2 = 10$ [7]. This configuration gives the highest area efficiency [165]. Unidirectional routing is preferred because it achieves higher area and performance efficiency [137]. The channel width is fixed to 32 for the routability of MCNC benchmarks. The CB connects the tracks to LB input via 32-input MUXs (MUX32) and the LB output to tracks via 4-input MUXs (MUX4). The SB supports 32 channel width and is constructed with MUX4-based switchpoints (SP).

5.2.2.1. 3D MUX4

In the designed FPGA multiplexers are used to create unidirectional routing blocks. The main component used in the CB and SB is the 4-input MUX with 2 memory cells. Each memory cell integrated with the MUX is a traditional 6T SRAM. The MUX is designed with NMOS-only pass gates with 6 transistors for minimized footprint. Buffers with PMOS keeper [166] are added after the MUX to improve the signal levels. Fig.5.4

shows the designed 2D and 3D cells. The 14nm 3D process allows dense placement of inter-tier vias as close as 100nm (Fig.5.4). As a result of the stacked NMOS placement and the small footprint of the inter-tier vias in the MUX, a very compact layout is achieved. With this partitioning, the MUX4, SRAM, and the buffer occupy equal area and therefore a balanced area between the top and bottom layers is established which is one of the major challenges in 3D design. Compared to the 2D cell (in Fig. A.27e), the 3D MUX4 achieves a very small area

5.2.2.2. 3D LUT

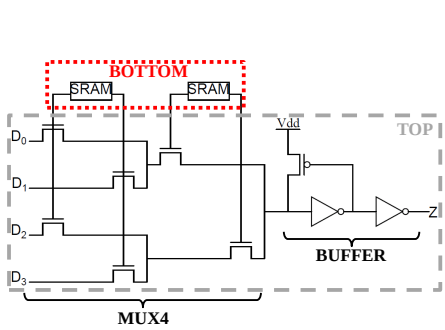
LBs are constructed with a cluster 4-input LUTs (LUT4). The LUT4 contains 16 SRAM cells, 16-input MUX, input, and output buffers as shown in Fig. 5.5. The memory and logic layers are designed and optimized separately. SRAM cells are placed in 4x4 configuration in order to keep close proximity to the configuration nodes. Even though the area of the memory layer is larger than that of the MUX, the difference is compensated when the input selection multiplexers, which use fewer memory cells, are added to construct the BLEs.

5.2.2.3. 3D SB

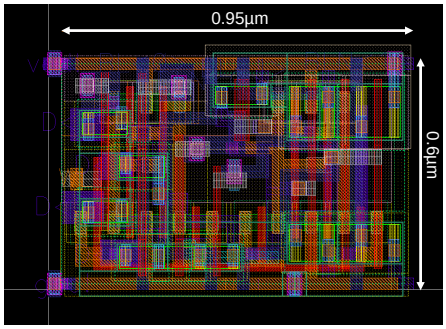
A switch point for unidirectional routing is designed as shown based on 3D MUX4 (Fig. 5.4). Each Switch Point (SP) includes four 3D MUX4. Fig. 5.6 shows the designed 3D SP. The number of inputs to the SPs is limited to 3, i.e. the switchbox flexibility: $F_S = 3$. Switch points are placed in 4x4 configuration which allows keeping the same height of the CB. In the designed 3D SB, there are 16 3D SPs which support a channel width of 32. The SB has the same connectivity scheme of a disjoint SB which is used in Xilinx XC4000 family FPGAs [167].

5.2.2.4. 3D CB

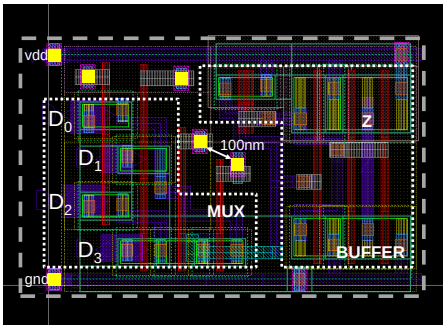
The connection box (CB) is composed of two different blocks: one block to connect the output of LB to the channels (CB_{out}) and one block to connect the channels to the input of LB (CB_{in}). CB_{in} is designed with 32-input MUXs. LB input number (I) is set to 10 and the inputs are distributed along each side of the LB as 3 inputs in the south, 2 in the north, 3 in the east, and 2 in the west. Thus, each CB_{in} between 2 LBs



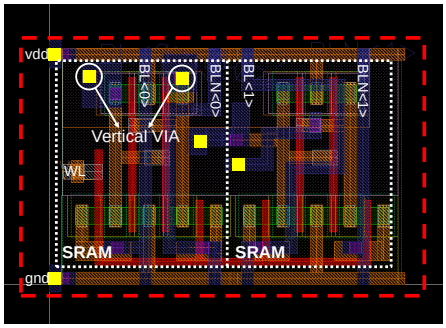
(a) Schematic view.



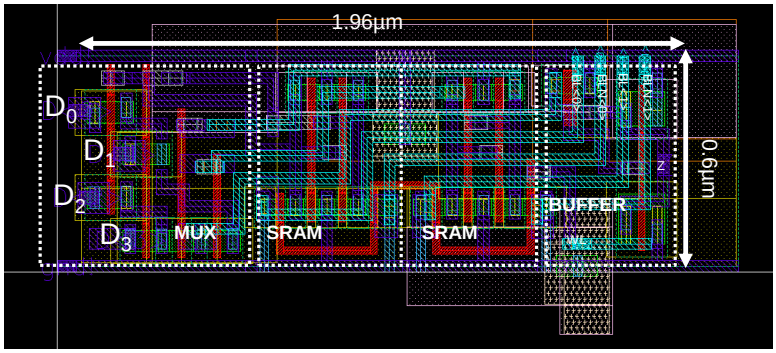
(b) 3D layout view.



(c) Top layer view.

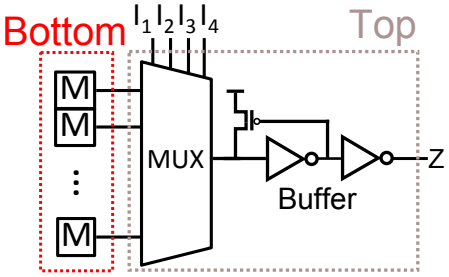


(d) Bottom layer view.

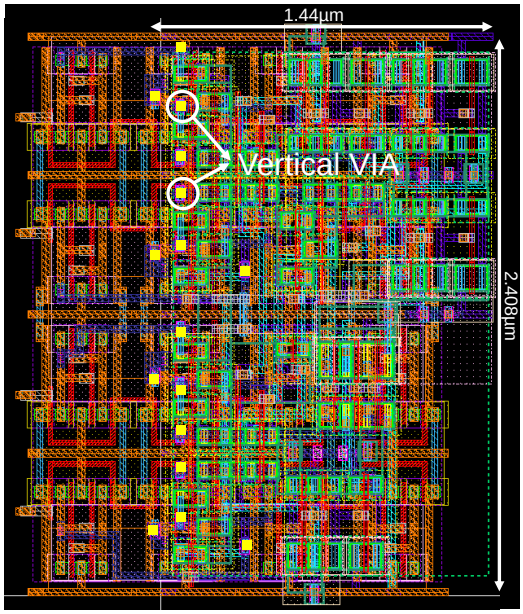


(e) 2D layout view.

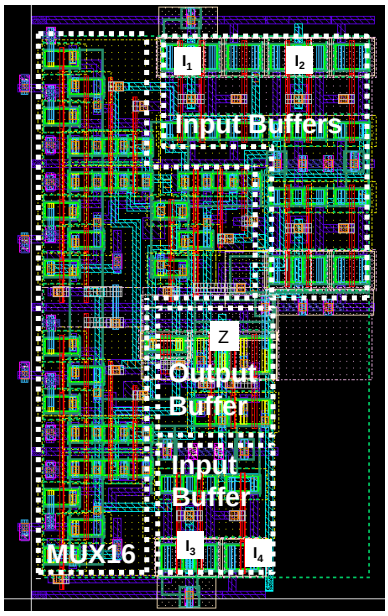
Figure 5.4: MUX4 full view



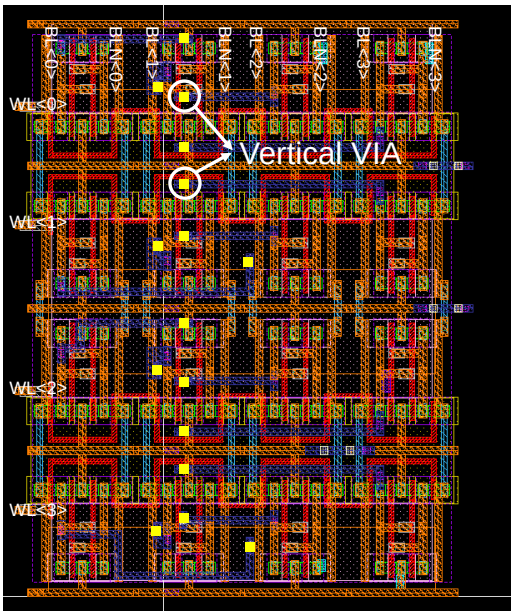
(a) Schematic view.



(b) 3D layout view.

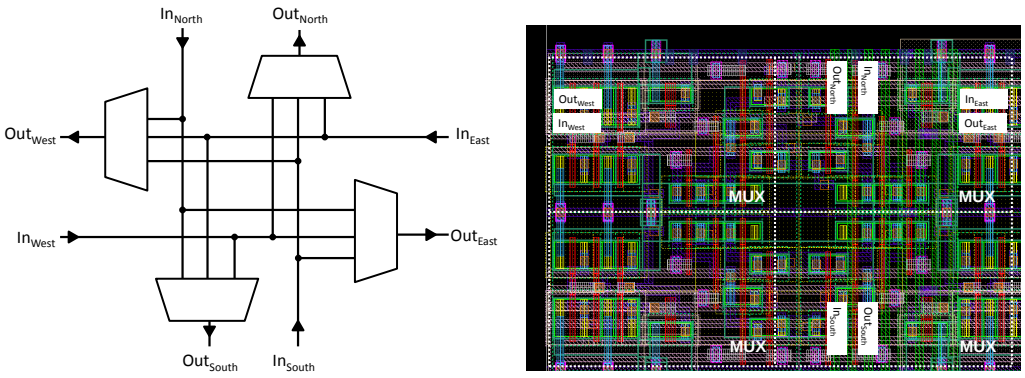


(c) Top layer view.



(d) Bottom layer view.

Figure 5.5: 3D LUT4 design



(a) SP schematic view. **(b)** 3D layout view.

Figure 5.6: 3D SP design for SB.

connects 5 inputs. Therefore, CB_{in} includes five MUX32s. For CB_{out}, each channel requires one 3D MUX4. The MUX4s are placed with 4x8 configuration to conform with the CB_{in} height.

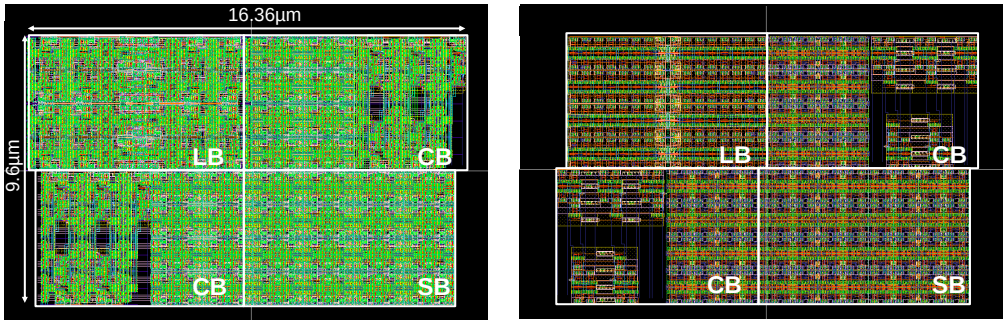
5.2.2.5. 3D TILE

Using the previously designed blocks, FPGA tiles are designed. As explained in Section 2.1.2., FPGA is comprised of replicated tiles. The tile contains a LB, a SB, and two CBs that interface the LBs and SBs.

Two different tiles are constructed with the 2D and 3D blocks. Fig. 5.7a and 5.7c show the 3D and 2D tiles respectively. The layout of the tiles are similar because the conventional FPGA architecture is preserved. The height of the tiles are kept the same because same logic and routing structure is adopted for the two tiles. Fig.5.7b shows the memory layer of the 3D tile. It can be observed that logic-on-memory partitioning achieves high area utilization between the two layers.

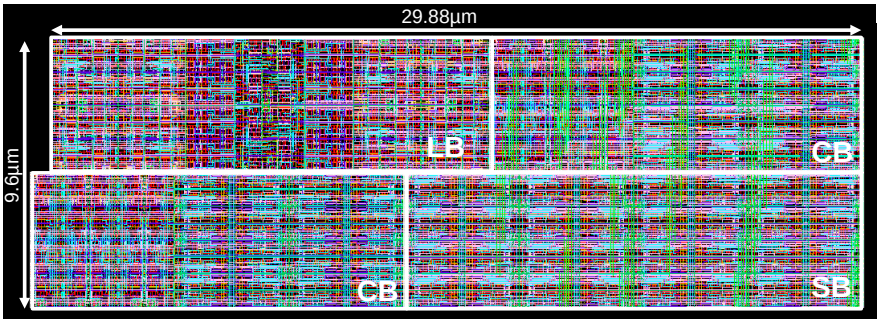
5.2.3. Performance Comparison of 2D and 3D Blocks

For each 3D block described in the previous section, the corresponding 2D blocks are also designed in the layout. Post-layout parasitic extraction is carried out for each cell and all the parasitics are back annotated. The extracted netlist with parasitics is simulated with ELDO and, delay and power metrics are measured. Dynamic power values assume an activity of 2GHz.



(a) 3D tile.

(b) 3D tile bottom layer view.



(c) 2D tile.

Figure 5.7: 2D and 3D FPGA tiles.

Table 5.1 shows the performance metrics for the 2D and 3D MUX4s. A gain of 51%, 14% and 12% in area, delay and power can be achieved. More than 50% of area reduction is achieved because the MUX4 and buffer shapes enable closer placement. Since the area is significantly reduced, the routing complexity is lowered. As a result, the output is generated faster with decreased power consumption. Furthermore, since the CRAM cells are connected to the selection input of the MUXs, the CRAMs are not on the critical path. Therefore, no effect from the inter-tier vias is observed on the performance.

Table 5.1: MUX4 Performance

MUX4	Area(μm^2)	Delay(ps)	Power(μW)
2D	1.18	28.48	2.75
3D	0.57	24.35	2.39
3D vs. 2D gain(%)	51	14	12

Table 5.2 shows that the LUT4 area can be reduced by 56% but the delay and power are increased by 1% and 3%. In the 3D LUT4 although the routing is simplified, the additional capacitance on the critical path introduced by the vertical inter-tier vias connection increases the delay and the power consumption.

Table 5.2: LUT4 Performance

LUT4	Area(μm^2)	Delay(ps)	Power(μW)
2D	7.93	60.17	7.10
3D	3.47	60.98	7.37
3D vs. 2D gain(%)	56	-1.3	-3.8

Table 5.3 shows the performance metrics for the 2D and 3D CBs. It can be observed that the area is lowered by 39%, delay by 5% and power by 4%. The decreased area reduction in the CB is due to the 3D MUX32s since they only have 5 SRAM cells which are placed in the bottom layer. Thus, the area cannot be reduced as much as it is for the MUX4s.

Table 5.4 shows the performance figures for the SB. Since the SB is constructed with replicated MUX4 cells, similar results are expected for the SB figures. Compared

Table 5.3: Connection Block Performance

CB	Area(μm^2)	Delay(ps)	Power(μW)
2D	61.8	98.4	7.2
3D	37.1	92.6	6.9
3D vs. 2D gain(%)	40	5.9	4.6

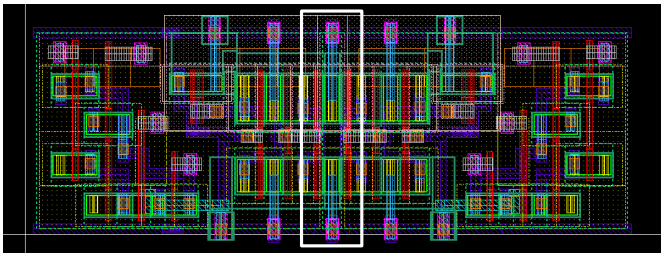


Figure 5.8: The supply connections designated in white box are overlapped due active area sharing for reduced total area.

to the 2D SB, 3D SB provides improvements by 55% reduction for area, 12% for delay and 5% for power. The area gain is more than that of MUX4 because the layout can be optimized by active area sharing which enables overlapping source/drain connections if they are connected to the same signal. Fig. 5.8 shows that the active area connecting to the same supply lines are overlapped which reduces total area. Delay and power metrics improves with less gains than MUX4 due to the increased capacitance and length of wires.

Table 5.4: Switch Box Performance

SB	Area(μm^2)	Delay(ps)	Power(μW)
2D	79.3	30.6	2.8
3D	65.3	26.7	2.7
3D vs. 2D gain(%)	55	12	5.3

Table 5.5 shows the performance figures for the 3D tile. 2D tile occupies an area of $287.8\mu m^2$. On the other hand, the 3D tile requires an area of $157.6\mu m^2$. The placement of memory cells to the bottom layer grants area reduction of 47%. The power and delay metrics are decreased by 6.2% and 2.4% respectively due to the reduced wirelength and

overall capacitance. Despite the the higher capacitance 3D vias which are introduced in 3D LUT, overall improvements are still observed.

Table 5.5: Tile Performance

Tile	Area(μm^2)	Delay(ps)	Power(μW)
2D	287.8	240.3	19.9
3D	151.6	225.4	19.4
3D vs. 2D gain(%)	47	6.2	2.4

5.2.4. Evaluation on 3D-FPGA with Logic-on-Memory Approach

In order to correctly evaluate the changes in the planar and 3D circuits, first the results from post layout simulations are extracted. Using the flow in Section 3.3., the architecture files for 2D and 3D FPGAs have been created. The timing model is updated with the wirelength and capacitance reductions. The area model is updated with the parameterized memory area in Fig. 3.7 by allocating no area for the configuration memory since the 3D architecture is constructed with logic-on-memory approach.

The architecture parameters of the 3D-FPGA for VPR experiments are shown in Table 5.6. The parameters are set to match the previously designed FPGA blocks. A set of MCNC benchmark circuits are evaluated with this configuration.

Table 5.6: 3DFPGA Architecture parameter

Parameter	Description	Value
K	Number of inputs per LUT	4
N	Cluster size (BLEs per LB)	4
I	Inputs per LB	10
$F_{C,in}$	CB input flexibility	1
$F_{C,out}$	CB output flexibility	1
F_S	SB flexibility	3
W	Channel width	32
L	Segment length	1

It can be observed in Fig. 5.9 that the area is reduced by 55%. The area benefits of 3D Monolithic Integration can be described as follows: First, the memory is completely

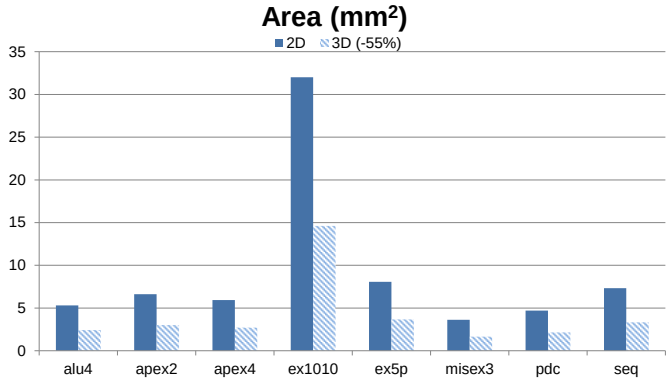


Figure 5.9: Area of FPGA benchmark circuits for 2D and 3D architectures. Area can be reduced by 55% on average when designed in 3D.

removed from the logic layer. Second, due to the high density of vertical connections, replacing the memory on the bottom layer does not impose any routing congestion. Especially, the use of metal layers between the top and bottom layers enables very flexible memory placement while keeping high proximity to the logic layer.

The Energy-Delay Product (EDP), as shown in Fig. 5.10, can be reduced by 47% with the proposed 3D FPGA. The improvement in the EDP is twofold: The intrinsic delays of the blocks are reduced with 3D integration due to simplified internal routing and the wirelength between blocks decreases leading to lowered wiring capacitances. In LUT4, however, even though the area is decreased, as well as the internal routing, the intrinsic delay becomes higher because of the vertical interconnects which represent additional loading to the critical path and power consumption (Table 5.2). After the VPR run, the results show that there is a significant reduction in the EDP. The main reason is, even though the 3D LUT has increased delay, shorter wire length in the global routing between blocks due to lowered area overcomes this effect and the EDP is further improved.

5.3. Multi-tier 3DFPGA

In the previous section, all the corresponding blocks for 2D and 3D FPGA are designed and the 3D-FPGA with logic-on-memory approach is evaluated. Using the library of 2D and 3D blocks and considering different partitioning granularities, multi-tier FPGA exploration can be carried out. In this section, several blocks partitioning

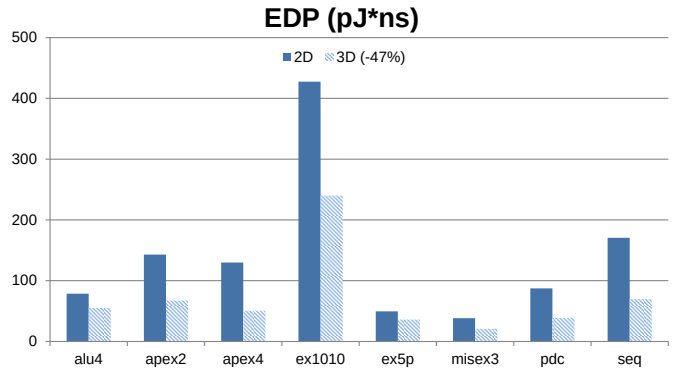


Figure 5.10: EDP of FPGA benchmark circuits for 2D and 3D architectures. EDP can be reduced by 47% on average when designed in 3D.

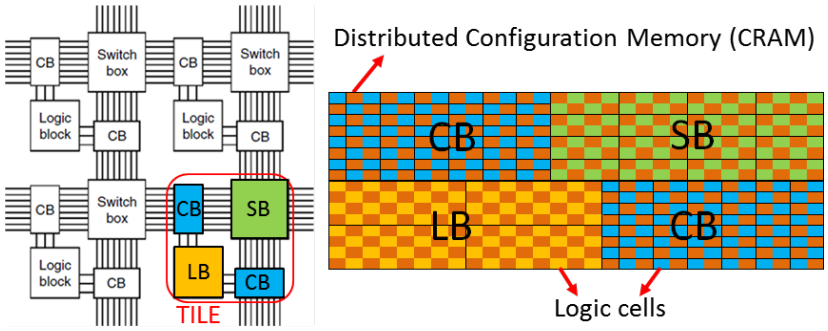


Figure 5.11: Island style FPGA a) Highlighted FPGA tile. b) 2D layout based view of FPGA tile in 14nm. Distributed CRAM and logic cells are highlighted.

schemes are analyzed considering 2, 3, and 4 active layers integration. Fig. 5.11 depicts the baseline 2D FPGA TILE designed in Section 5.2.2.5..

5.3.1. 2-Tier FPGA Stack

Two different schemes are considered for two layer integration (Fig. 5.12). The first scheme (2L_1) refers to the FPGA designed in Section 5.2.4. in which the blocks are separated in logic-on-memory approach. This FPGA is considered for comparison with other partitioning schemes. In the second scheme (2L_2), 2D layout is separated into two tiers, with block level partitioning. The LB and one CB are placed in one layer and the remaining CB and SB are placed into the other. The effect of 3D vias is observed due to the crossing of the critical path signal into the second layer. Nevertheless, the wirelength is significantly reduced compared to the 2D layout.

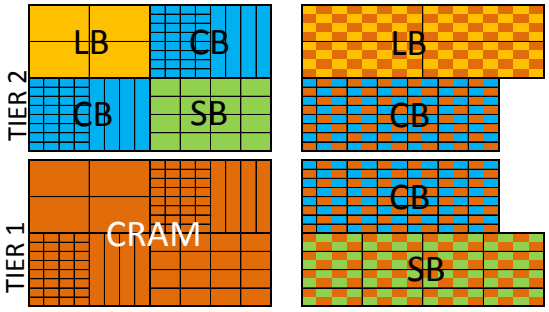


Figure 5.12: FPGA design in two-tiers: a) logic-on-memory approach (2L_1): configuration memory (CRAM) on the bottom and logic on the top tier. b) block-level partitioning (2L_2): 2D blocks are separated between two tiers.

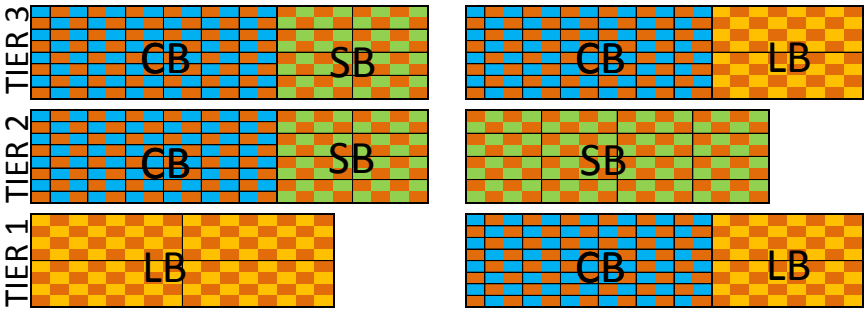


Figure 5.13: Block level FPGA partitioning in 3 tiers: a) 3L_1 b) 3L_2.

5.3.2. 3-Tier FPGA Stack

For an implementation including three active layers, two different schemes are analyzed (Fig. 5.13). In the first scheme (3L_1), the LB is kept in one layer. The CBs are placed in separate layers and the SB is divided into two layers. Since the LB and CBs are in different layers, the effect of 3D vias is observed on the local routing between CB and LB. The effect of 3D via on the global routing is minimized because the partial SBs are kept on the same layer as CBs. In the second scheme (3L_2), the LB is separated into two (2x BLEs each) layers. CBs are put close to the BLEs and the SB is kept as one block in middle layer. The internal delay of LB is increased because of the additional 3D vias. Since SB is in the middle layer, global routing is carried out in one layer and the effect of 3D vias is minimized. When three layers are considered, logic-on-memory approach cannot be applied because it does not significantly improve area results compared to 2-tier: logic-on-memory requires even number of tiers due to inherent symmetric area partitioning.

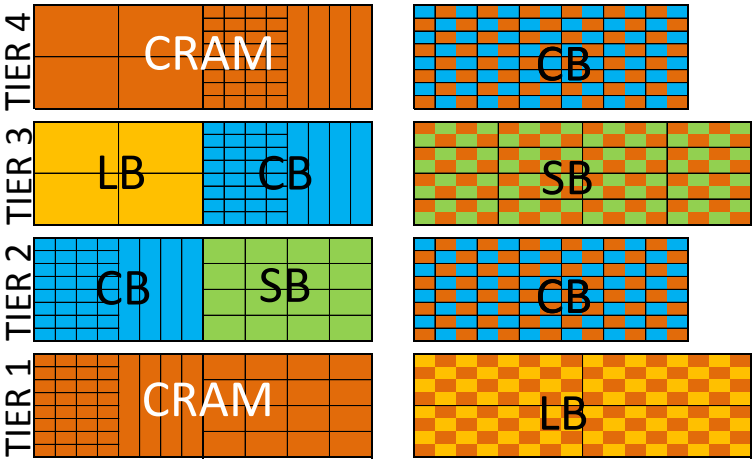


Figure 5.14: Block level FPGA partitioning in 4 tiers: a) 4L_1 b) 4L_2.

5.3.3. 4-Tier FPGA Stack

For an implementation including four active layers, two different schemes are analyzed (Fig. 5.14). In the first scheme (4L_1), the 2L_1 tile is splitted into two tiers in which the LB and one CB are on one layer and the other CB and SB are on the other layer. The remaining two layers consist of the associated CRAMs. In the second scheme (4L_2), the LB and SB are put into different layers. The CBs are placed close to the SB. In this way the effect of 3D vias on global routing is minimized. On the other hand, there is significant effect of the vias on the local communication between CBs and LB.

5.3.4. Multi-tier FPGA Performance Evaluation

For the evaluation of the proposed stacking schemes, several benchmarks are analyzed in VPR5 with Power Extension flow 3.3.. Separate VPR architecture files are created for each of the partitioning schemes. The 2D and 3D logic-on-memory cell layouts in 14nm using 3D add-on are characterized. In order to observe the realistic effects of the schemes, wirelength and capacitive loading of the wires are modeled based on the area improvements. Table 5.7 shows the results of several FPGA benchmark circuit evaluations considering multiple tiers with 3D monolithic integration. The final average values of all schemes are normalized to 2D for comparison.

Starting from 2-tier schemes, compared to 2D, significant area improvements are

observed as the number of tiers is increased. The area is reduced up to 55%, 69% and 77% in comparison to 2D for 2, 3, and 4-tier implementations respectively.

The results for EDP show that the reduction in area improves the EDP substantially. According to the results two different types of improvements are observed: 1) reductions due to shorter wirelength and less capacitive loading and 2) reduction of gate internal delay and capacitive loading due to logic-on-memory approach. When two layers are considered, EDP can be reduced by 47% and 26% for the different schemes in comparison to 2D. In 2L_2, the partitioning is performed using 2D blocks and in 2L_1, logic-on-memory approach is applied. Specifically, in 2L_2 reductions stem purely from the shorter wirelength. In 2L_1, however, the delay is reduced not only by shorter wirelength but also by smaller internal delay of the blocks as a result of the logic-on-memory approach. In three and four layer solutions, it is possible to observe the effect of decreased wirelength. In three layers, up to 40% EDP reduction is observed depending on the scheme and in four layers, up to 66% compared to 2D. These results show that the improvement in delay is proportional to the reduction in wirelength however, redesign of blocks (eg. with logic-on-memory approach) can further improve EDP.

5.4. 3DMI Impact on Scaling

In this section, the best results obtained by multi-tier FPGA partitioning exploration are compared to the scaling expectations. The reason is that when the number of integrated layers with 3DMI are increased, logic density, performance, and power efficiency values are improved. Traditional scaling defined by Moore's law also aims to make these values better with each advancing node. The comparison between 3DMI multi-tier stacking and traditional scaling demonstrates what can be achieved from 3DMI for future implementations.

Recent ITRS 2013 report [168] concludes that the transistor count is expected to continue scaling by 1.6x per node. Similarly in 3DMI, as more layers are stacked, the logic density increases. Expected ITRS trend and the highest 3DMI area improvements from previous results are illustrated in Fig. 5.15. The designed 2L FPGA in 14nm decreases the area by 27% more than traditional 10nm scaling can achieve. 3L

Table 5.7: Area and EDP results of FPGA benchmark circuits based on multi-tier design

	Area (mm^2)							*EDP* (*pJ*ns*)						
	2D	***2L_1***	***2L_2***	***3L_1***	***3L_2***	***4L_1***	***4L_2***	***2D***	***2L_1***	***2L_2***	***3L_1***	***3L_2***	***4L_1***	***4L_2***
alu4	5.3	2.4	2.6	1.6	1.6	1.2	1.3	78.6	55.1	67	60	54.1	32.2	49
apex2	6.6	3	3.2	2	2	1.5	1.6	143	67.2	91.2	83.9	71.2	47.6	64.5
apex4	6	2.7	2.9	1.8	1.8	1.4	1.4	129.9	50.5	64.6	53.1	48.4	27.9	43.5
des	32	14.6	15.6	9.8	9.9	7.4	7.8	427.5	240.1	334.8	307	272.3	162.1	259.8
ex1010	8.1	3.7	3.9	2.5	2.5	1.9	2	49.5	36.1	46.7	39.7	40.2	19.2	30.5
ex5p	3.6	1.6	1.8	1.1	1.1	0.8	0.9	38.4	20.7	29.2	25.9	25.5	15.1	21.6
misex3	4.7	2.1	2.3	1.4	1.4	1.1	1.1	87.4	38.6	58.9	52.1	46.2	28.3	40.5
seq	7.3	3.3	3.6	2.2	2.3	1.7	1.8	170.5	69.7	135.1	111.9	101.4	51	90.4
Normalized to 2D (avg.)	**100**	**45.4**	**48.6**	**30.4**	**30.7**	**23.1**	**24.3**	**100**	**53**	**74.3**	**65**	**59.6**	**34.8**	**52.4**

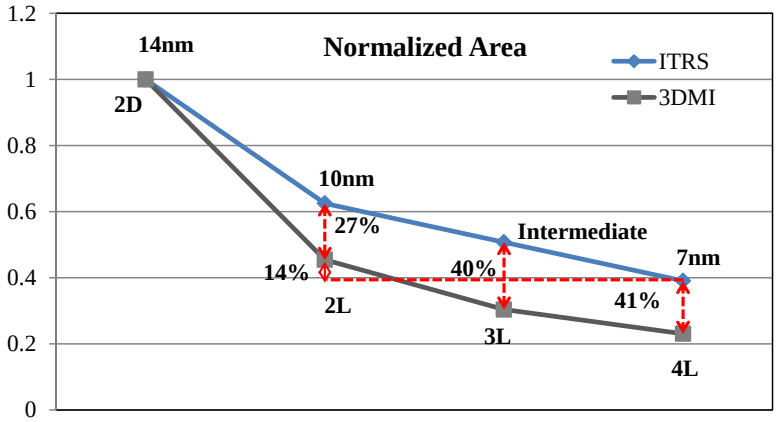


Figure 5.15: Projection of area improvement for future technology nodes based on ITRS roadmap and gain from multi-tier 3DMI approach.

integration in 3DMI can be comparable to an intermediate node which provides 40% improved area figure. Finally, 4L FPGA integration in 14nm exceeds the gains in 7nm by 41%.

It is also interesting to notice in Fig. 5.15 that the extrapolated 2D FPGA in 7nm node and the 2L 3D FPGA exhibit similar area gains with only 14% difference. Since the difficulties of traditional scaling affect the gains per each node advancement, it is expected that this difference between 2D and 3D case will be narrowing down for the new nodes. Thus, the area gain with 2-tier case might attain the same level as advancing two nodes with scaling in the future.

Based on the ITRS report [169], the EDP improvements, up to 7nm node, are calculated. The ITRS trend and the highest 3DMI EDP improvements from the designed FPGAs are illustrated in Fig. 5.16. It can be observed that for the 2L implementation in 14nm not only reaches 10nm node performance but also further improves it by 24%. Since logic-on-memory approach is not utilized in 3L, the relative improvement is only by 6%. This result shows the importance of partitioning as the gain is strongly corrolated to the design when using 3DMI technology. The 3L case can be an intermediate node which targets low area (due to the improvements in Fig. 5.15) while still keeping similar performance to the next scaling advancement. The 4L implementation in 14nm surpasses the 7nm improvements by 29%.

It is also worth noting in Fig. 5.16 that the extrapolated 2D FPGA in 7nm node and the 2L 3D FPGA demonstrate similar EDP gains with only 7% difference. Considering

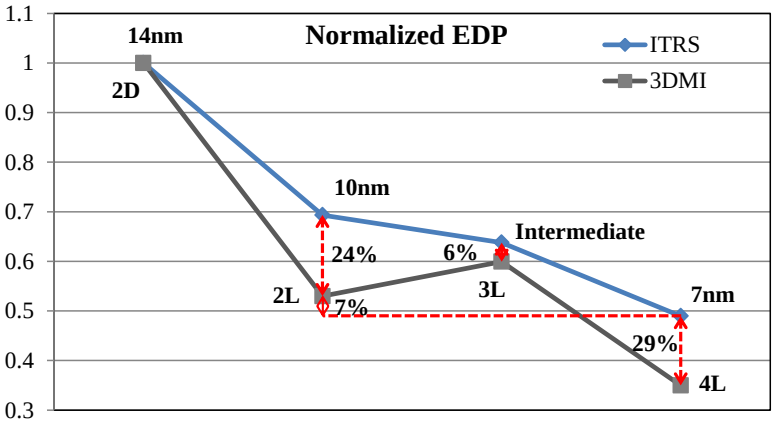


Figure 5.16: Projection of EDP improvement for future technology nodes based on ITRS roadmap and gain from multi-tier 3DMI approach.

the difficulties in scaling, this difference might disappear in the near future. Thus, the 2-tier 3D-FPGA in 14nm can reach the same efficiency of a 2D-FPGA in 7nm. This shows that the integration with 2 tiers might be an alternative to not only one technology node but two nodes advancement.

Consequently, multi-tier 3DMI shows a very efficient way of improving the logic density and EDP. When the number of integrated layers is doubled, the gains not only reach the scaling improvements of one node but also outperform them. Furthermore, considering the reduced gains of scaling trends, doubling the number of layers might exhibit the same gains as two nodes advancement with scaling. Therefore, multi-tier 3DMI is a very strong alternative for future scaling.

5.5. Conclusion

In this chapter, the details about 3DMI technology are explained and the advantages that it can bring to the future 3D FPGA designs are discussed. 3DMI offers a very unique design opportunity due to the high density inter-tier connections while integrating multiple active layers on the same die. FPGAs can benefit from these dense connections to separate the logic and configuration memories which are distributed and consume a large portion of FPGA surface. 3DMI also enables optimizing the layers individually which allows placement of low leakae devices for memories on one layer and high performance for logic on the other one.

Initially, taking advantage of 3DMI, all the FPGA blocks are designed by separating the memory and logic cells in 2 layers with logic-on-memory approach. All the designs are realized with 3D 14nm FDSOI process. By benefiting from high density inter-tier connections, the design effort is put for compact layouts and no area overhead is observed due to the inter-tier vias. It is proven that 3DMI reduces the total area and the routing complexity significantly. Moreover, the performance and power consumption metrics are improved substantially due to decreased wire lengths and capacitances.

Several 3D FPGA designs are analyzed using the designed blocks. First, an FPGA including 2 active layers is designed and evaluated with FPGA benchmark circuits. The designed FPGA blocks are, then, utilized considering several partitioning schemes for evaluation of 3D FPGAs up to 4 active layers. The 3D FPGA with 4 layers presents significant improvements in terms of area (77%) and EDP (66%) in comparison to the 2D counterpart.

The results show that each stacked active layer with 3DMI improves logic density and EDP. When compared to the extrapolated results of 14nm 2D FPGA into 10nm and 7nm nodes based on ITRS figures, stacking layers with 3DMI in 14nm outperforms the expected scaling improvements. Namely, adding more active layers while keeping the same technology can be an alternative to the traditional scaling. The results from this chapter are published in [170][171][172][173][174][175].

6

Conclusion & Perspectives

Contents

In this thesis, we are motivated by the opportunities of emerging 3D technologies from which the FPGAs can take advantage. Available technologies are vast and they have advantages on different levels. Hence, these technologies should not only be assessed but also exploited in order to find the unique benefits. In this work, we have proposed an evaluation framework for emerging technologies applied to FPGAs. Using the FPGA evaluation framework, several emerging 3D technologies (OxRAM, CBRAM, and 3DMI) have been assessed. The results show that circuit footprint can be significantly reduced with 3D technologies mainly because some of the circuit functionality can be integrated in the third dimension. Moreover, the simplified routing enables higher performance and low power consumption. Apart from that, the inclusion of nonvolatility leads to greater power savings with added functionality. All these results show that FPGAs are one of the leading computing platforms which offers many opportunities for improvement and now we can rethink how the FPGAs will be employed in future computing systems.

In this concluding chapter, results are discussed for final comparison of gains achieved with these emerging technologies. The contributions are also summarized and future

perspectives are provided for further continuation of this research.

Table 6.1 summarizes the gains accomplished with different 3D technologies. The depicted results are expressed in comparison to the conventional FPGA implementation in the corresponding technology node. Area and Energy-Delay Product (EDP) are chosen as the global comparison metrics. Area refers to the silicon footprint of the implementation. EDP refers to both performance and energy efficiencies as it can be interpreted as improved performance in the same energy level or improved energy considering the same operating frequency.

Circuit footprint can be significantly reduced with 3D technologies mainly because some of the circuit functionality can be integrated in the third dimension. In the FPGA, almost half of the chip area is dedicated for the configuration nodes which is one of the main limitations of the FPGA. With Resistive RAM(RRAM)-based integration, resistive devices can be fabricated between metal layers and only the control transistors consume silicon area. The CBRAM-based compact memory cell achieves 33% less area when it replaces the SRAM cells in an FPGA. Highest area gains can be achieved with 3DMI taking advantage of very compact inter-level via contacts. Considering a logic-on-memory partitioning in 2 tiers, area can be reduced by 55%. With integration of more layers, the area can be reduced by 77% for the case of four layers.

Since FPGA fabric is also dominated by routing resource, the reduction in area leads to reduced wirelength. This enables reduction in routing capacitance which significantly improves circuit performance and power consumption. CBRAM-based FPGA improves the EDP by 66%. With 3DMI, 2-tier integration reduces the EDP by 47% and 4-tier achieves 66% better EDP. Even though more area gain is achieved with 3DMI, the EDP is improved less compared CBRAM-based implementation. One reason could be the inter-tier via capacitance overhead in LUTs in 3DMI which affects performance and power consumption.

One of the major benefits of RRAM integration is the non-volatility. This property enables saving configuration bitstream and context results in the registers. Thus, external Flash memories become redundant. Moreover, with non-volatility, FPGAs can target Normally-off Instantly-on applications which can save significant power consumption. Depending on the application activity, OxRAM-based implementation reduces the EDP by 42% for configuration saving applications and 32% for configuration and context saving applications. CBRAM-based FPGA, on the other hand, achieves

Table 6.1: FPGA improvements gained by emerging 3D technologies in comparison to traditional SRAM-based FPGAs in respective technology node

Technology	Implementation	Area(%)	EDP(%)	Feature
OxRAM - 22nm	NVSRAM	+18	+16	Configuration saving
			-42	Normally-off Instantly-on
	NVSRAM+NVFF	+18	+21	Configuration and Context saving
			-32	Normally-off Instantly-on
CBRAM - 130nm	NVE(1T2R)	-33	-66	Configuration saving
			-98	Normally-off Instantly-on
3DMI - 14nm	2-tier	-55	-47	Logic-on Memory
	3-tier	-69	-40	Block-level partitioning
	4-tier	-77	-66	Logic-on-memory and block-level partitioning

up to 98% better EDP. Compared to CBRAM, the implementation in OxRAM has substantial area overhead which reduces the EDP efficiency. Nevertheless, OxRAM enables increased functionality and low EDP operation with non-volatility which could be of higher importance than the area overhead for battery-operated applications.

6.1. Contributions

After summarizing the motivation for emerging technologies in the introduction, general background of FPGAs and emerging technologies are explained in Chapter 2. For the analysis with emerging technologies an experimental FPGA framework and an evaluation methodology is proposed in Chapter 3. Using the framework, RRAM-based elementary circuits are evaluated in Chapter 4 with a low-power goal exploiting non-volatile property. In Chapter 5, several FPGAs are designed with 3DMI and analyzed with various partitioning schemes considering multiple layer stacking.

Since this thesis aims to find efficient FPGA implementations using emerging 3D technologies, in Chapter 3 we developed an evaluation framework and proposed a methodology to translate properties of emerging technologies into FPGA fabrics. First, an experimental framework is constructed which includes a complete set of tools for exploration with emerging technologies on FPGAs. This VPR-based toolflow allows fast assessment of wide range of technologies with broad application specifications. The toolflow is intended for generic, replicated fabric of FPGAs and the evaluation is carried out with an architecture definition which contains FPGA fabric related parameters.

Since different technologies have impacts on various levels, a generic methodology is proposed in order to convert technological impacts imposed by different technologies into specific parameters of FPGA fabric. With this methodology, FPGA architecture definitions are created for the assessment of emerging technologies using the evaluation framework. Thus, the platform including the evaluation framework and architecture definition methodology creates the basis for the evaluations in emerging 3D technologies in the rest of the thesis.

In Chapter 4, we explored the benefits gained from non-volatile Resistive RAMs (RRAM). We focused on OxRAM and CBRAM-based implementations in 22nm FD-SOI and 130nm BULK technology nodes which are available in LETI respectively. First, a very compact CBRAM cell is integrated in the FPGA as replacement for SRAM configuration node. The memory area overhead, which is 43% in SRAM-based FPGAs, is reduced significantly. Overall FPGA area is reduced by 33% with CBRAM-based cell. With smaller area, shorter routing wires with smaller capacitance suffice which lead to reduction in critical path delay and power consumption. With OxRAM, non-volatile configuration memories and registers are integrated in the FPGAs. Since the cells are larger than their volatile counterparts, there are overheads in area, delay, and power metrics. For both technologies, the non-volatile behavior brings increased functionalities to the FPGA which enables FPGA utilization in applications that are otherwise unexplored. By replacing the traditional SRAM and registers with OxRAM and CBRAM counterparts, configuration saving and configuration/context saving applications are targeted respectively. Furthermore, when FPGAs are used for Normally-off Instantly-on computing, power consumption can be significantly optimized by switching off the FPGA when not in use. Therefore, RRAM technologies are analyzed considering their advantages and disadvantages with a focus on increased functionality due to non-volatility.

In Chapter 5, we explored the benefits of 3D Monolithic Integration (3DMI) in FPGAs. Taking advantage of the very high density inter-tier connections due to compact 3D via size, several partitioning possibilities are analyzed starting from 2 and up to 4 active layers. Using the 14nm 3D LETI-FDSOI PDK, all the FPGA blocks are designed in 2D and 3D views on layout. The 3D blocks consider logic-on-memory approach and all the blocks are optimized for small area. The designed blocks are first simulated for functionality and then characterized for area, delay, and power consumption after

parasitics extraction. Using the parameters from the designed blocks, a 2-tier FPGA is assessed with the framework defined in Chapter 3. The placement of the memory cells on the bottom tier results in a very compact FPGA with 55% area reduction which leads to shorter less capacitive routing wires increasing the EDP efficiency by 47%. We also explored different partitioning schemes on multi-layer stacks up to 4 active layers. With multi-stacking FPGA area and EDP can be reduced by 77% and 66% respectively. The obtained results are encouraging towards defining a new scaling parameter which states that rather than reducing the transistor size which is becoming more and more difficult to justify economically and technologically with each advancing node, active layers can be stacked to achieve increased logic density, performance, and power efficiency.

6.2. Future Works

6.2.1. Towards 3DNVFPGA: Merging RRAM and 3D Integration

In this thesis emerging 3D technologies, RRAM and 3DMI, are analyzed for future adoption possibilities in FPGAs. The benefits from these technologies are estimated using the established FPGA evaluation framework. The results show that RRAM integration and active layer stacking with 3DMI accomplish significant benefits in terms of performance, power, and area (PPA). In the thesis work, these technologies are studied individually but as a short term future work, an FPGA design can be envisioned using both of these technologies. Recently, Zhang et al. fabricated an FPGA with low temperature 3DMI process and RRAM integration for the purpose of proof concept [176]. In this FPGA RRAM-based cells are integrated to replace the configuration memories and in the second active layer, PMOS transistors are integrated. PMOS-only second layer limits partitioning possibilities to only transistor-on-transistor approach. In the new FPGA, the memory cells can be replaced with a cell similar to the previously examined NVE cell in Chapter 4. The rest of the logic and routing blocks can be partitioned using 3DMI. In this case, finding an efficient partitioning scheme is important because block separation should consider the layout designs of the blocks as well as the area of the control transistor in the NVE. The evaluation framework defined in Chapter 3 can be used for fast exploration of several partitioning schemes. The designed FPGA

also supports Normally-off Instantly-on computing defined in Chapter 4 with which the power consumption can be further optimized.

6.2.2. Thermal impacts of FPGA designs with 3DMI

One of the main concerns in the design and manufacturing of 3D circuits is the heat dissipation [104]. Stacking multiple layers and increased logic density lead to higher thermal density which is becoming more difficult to dissipate. The created hotspots impose challenges in packaging but also the circuit performance is linked to the ambient temperature. In order to address these concerns, in the 3DMI technology, the thermal model must be created based on the material stack including the thermal resistivity of the vias. FPGA blocks designed in Chapter 5 can be used towards circuit specific thermal evaluation. The effect of inter-tier vias on thermal dissipation can be also evaluated to find the best placement of vias for optimized thermal dissipation.

6.2.3. Reliable designs with SiNWFETs

Extreme scaling and increased operation frequencies lead the devices to deeper sub-micron levels. Hence, noise margins significantly shorten and circuits become severely susceptible to soft errors. In addition to these trends, the ever increasing complexity of the systems requires more efficient solutions for fault tolerance. In this context, integration of mature CMOS and emerging technologies might lead to unreliable systems. Technological variabilities during the fabrication process due to the lack of maturity might highly impact the devices properties. This implies that these impacts should be expected during the design step where architectural solutions can target reliable operation with unreliable devices.

In order to demonstrate reliable design opportunities with emerging technologies, we designed a self-checking adder with Silicon Nanowire FETs (SiNWFET) in collaboration between LETI and EPFL [177]. SiNWFETs present good channel control properties and limited fabrication complexity. They have strong arguments thanks to their classical CMOS material compatibility and they exhibit ambipolar behavior, i.e., both n and p-type conduction, that can be controlled by the use of an extra gate. Recently, very efficient implementations of digital circuits, e.g., XOR, have been demonstrated using this technology and more specifically its ambipolar property [178]. These XORs lead to the design of very compact full adder designs. In the designed self-checking

adder, SiNWFETs reduce the overhead arising from the application of self-checking. Therefore, compared to CMOS design style up to 56% smaller and 62% faster adders can be achieved with reliable operation.

Design with SiNWFETs represent promising opportunities due to their reduced transistor count and increased functionality. For future work, the same concept of self-checking property can be applied to other arithmetic blocks such as multipliers. The designed blocks can be included in the datapath of a processor in ASIC or DSP. Additionally, in FPGAs, heterogeneity is an increasing trend for future FPGAs as more dedicated units are included as a part of circuit. SiNWFET-based blocks can be integrated as standalone units for reliable FPGA operation. The exploration framework defined in Chapter 3, can be used for the estimation of impacts on area, performance, and power consumption as it supports evaluation with heterogeneous blocks.

List of Publications

Journal Publications

O. Turkyilmaz, S. Onkaraiah, M. Reyboz, F. Clermidy, Hraziia, C. Anghel, J.-M. Portal, M. Bocquet, "RRAM-based FPGA for Normally Off, Instantly On applications", *Journal of Parallel and Distributed Computing*, 2013.

L. Brunet, P. Batude, F. Fournel, L. Benaissa, C. Fenouillet-Beranger, L. Pasini, F. Deprat, B. Previtali, F. Ponthenier, A. Seignard, C. Euvrard-Colnat, M. Rivoire, P. Besson, C. Arvet, E. Beche, O. Rozeau, O. Billoint, **O. Turkyilmaz**, F. Clermidy, T. Signamarcheix, and M. Vinet, "Direct Bonding: A Key Enabler for 3D Monolithic Integration", *The Electrochemical Society (ECS) Transactions*, 2014.

Conference Publications

O. Turkyilmaz, G. Cibrario, O. Rozeau, P. Batude, and F. Clermidy, "3D FPGA using high-density interconnect Monolithic Integration", *Design, Automation and Test in Europe Conference and Exhibition (DATE)*, 2014.

O. Turkyilmaz, F. Clermidy, L. G. Amarù, P.-E. Gaillardon, and G. De Micheli, "Self-Checking Ripple-Carry Adder with Ambipolar Silicon NanoWire FET", *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2013.

O. Turkyilmaz, S. Onkaraiah, M. Reyboz, F. Clermidy, Hraziia, C. Anghel, J.-M. Portal, and M. Bocquet, "RRAM-based FPGA for Normally Off, Instantly On Applications", *IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH)*, 2012.

E. Vianello, O. Thomas, G. Molas, **O. Turkyilmaz**, N. Jovanovic, D. Garbin, G.

Palma, M. Alayan, C. Nguyen, J. Coignus, B. Giraud, T. Benoist, M. Reyboz, A. Toffoli, C. Charpin, F. Clermidy, and L. Perniola "Resistive Memories for Ultra-Low-Power embedded computing design", *IEEE International Electron Devices Meeting (IEDM)*, 2014.

F. Clermidy, N. Jovanovic, S. Onkaraiah, H. Oucheikh, O. Thomas, **O. Turkyilmaz**, E. Vianello, J.M. Portal, and M. Bocquet, "Resistive memories: Which applications?", *Design, Automation and Test in Europe Conference and Exhibition (DATE)*, 2014.

M. Vinet, P. Batude, C. Fenouillet-Beranger, F. Clermidy, L. Brunet, O. Rozeau, JM Hartmann, O. Billoint, G. Cibrario , B. Previtali, C. Tabone, B. Sklenard, **O. Turkyilmaz**, F. Ponthenier, N. Rambal, MP. Samson, F. Deprat, V. Lu, L. Pasini, J-E. Michallet, and O. Faynot, "Monolithic 3D Integration: a powerful alternative to classical 2D Scaling", *IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*, 2014.

F. Clermidy, **O. Turkyilmaz**, O. Billoint, P.-E. Gaillardon, "3D technologies for reconfigurable architectures", *International Symposium on Reconfigurable and Communication-Centric Systems-on-Chip (ReCoSoC)*, 2014.

G. Cibrario, M. Gary, F. Gays, K. Azizi-Mourier, O. Billoint, **O. Turkyilmaz**, and O. Rozeau, "A High-Level Design Rule Library Adressing CMOS and Heterogeneous Technologies", *International Conference on IC Design and Technology (ICICDT)*, 2014.

P. Batude, B. Sklenard, C. Fenouillet-Beranger, B. Previtali, C. Tabone, O. Rozeau, O. Billoint, **O. Turkyilmaz**, H. Sarhan, S. Thuries, G. Cibrario, L. Brunet, F. Deprat, J-E. Michallet, F. Clermidy and M. Vinet, "3D sequential integration opportunities and technology optimization", *IEEE International Interconnect Technology Conference (IITC)*, 2014.

S. Onkaraiah, **O. Turkyilmaz**, M. Reyboz, F. Clermidy, J.-M. Portal, and C. Muller, "An Hybrid CBRAM/CMOS Look-Up-Table structure for improving performance efficiency of Field-Programmable-Gate-Array", *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2013.

Posters & Presentations

O. Turkyilmaz and F. Clermidy, "Using 3D technologies to reduce power consumption of FPGAs", *EDAA/ACM SIGDA PhD Forum DATE*, 2014.

O. Turkyilmaz and F. Clermidy, "Using Novel Technologies to Reduce Power Consumption in FPGAs", *FETCH Winter School*, 2013.

Bibliography

[1] G. Yeap. Smart mobile socs driving the semiconductor industry: Technology trend, challenges and opportunities. In *Electron Devices Meeting (IEDM), 2013 IEEE International*, pages 1.3.1–1.3.8, Dec 2013. doi: 10.1109/IEDM.2013.6724540. xii, xvi, 2, 158

[2] ITRS. Semiconductor industry association, 2006. URL `http://public.itrs.net/`. xii, xvi, 3, 159

[3] Altera. Expect a breakthrough advantage in next- generation fpgas, 2013. xii, xvi, 6, 161

[4] D. Marple and L. Cooke. An mpga compatible fpga architecture. In *Custom Integrated Circuits Conference, 1992., Proceedings of the IEEE 1992*, pages 4.2.1–4.2.4, May 1992. doi: 10.1109/CICC.1992.591107. xii, 15, 16

[5] Xilinx. 7 series fpgas configurable logic block ug474 (v1.6), 2014. URL `http://www.xilinx.com/support/documentation/user_guides/ug474_7Series_CLB.pdf`. xii, 16, 18

[6] Altera. Logic array blocks and adaptive logic modules in stratix-v devices (2014.01.10), 2014. URL `http://www.altera.com/literature/hb/stratix-v/stx5_51002.pdf`. xii, 17, 19

[7] Vaughn Betz, Jonathan Rose, and Alexander Marquardt, editors. *Architecture and CAD for Deep-Submicron FPGAs*. Kluwer Academic Publishers, Norwell, MA, USA, 1999. ISBN 0792384601. xii, 21, 50, 51, 54, 106

[8] Mingjie Lin, A. El Gamal, Yi-Chang Lu, and Simon Wong. Performance benefits of monolithically stacked 3-d fpga. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, 26(2):216–229, Feb 2007. ISSN 0278-0070. doi: 10.1109/TCAD.2006.887920. xii, 26, 27, 42, 63, 167, 178

[9] T. Tuan and B. Lai. Leakage power analysis of a 90nm fpga. In *Custom Integrated Circuits Conference, 2003. Proceedings of the IEEE 2003*, pages 57–60, Sept 2003. doi: 10.1109/CICC.2003.1249359. xii, 5, 26, 27, 160

[10] Li Shang, Alireza S. Kaviani, and Kusuma Bathala. Dynamic power consumption in virtex™-ii fpga family. In *Proceedings of the 2002 ACM/SIGDA Tenth International Symposium on Field-programmable Gate Arrays*, FPGA '02, pages 157–164, New York, NY, USA, 2002. ACM. ISBN 1-58113-452-5. doi: 10.1145/503048.503072. URL http://doi.acm.org/10.1145/503048.503072. xii, 27

[11] H. S P Wong, S. Raoux, SangBum Kim, Jiale Liang, John P. Reifenberg, B. Rajendran, Mehdi Asheghi, and Kenneth E. Goodson. Phase change memory. *Proceedings of the IEEE*, 98(12):2201–2227, Dec 2010. ISSN 0018-9219. doi: 10.1109/JPROC.2010.2070050. xiii, 30

[12] H. S P Wong, Heng-Yuan Lee, Shimeng Yu, Yu-Sheng Chen, Yi Wu, Pang-Shiu Chen, Byoungil Lee, F.T. Chen, and Ming-Jinn Tsai. Metal oxide rram. *Proceedings of the IEEE*, 100(6):1951–1970, June 2012. ISSN 0018-9219. doi: 10.1109/JPROC.2012.2190369. xiii, 31

[13] Shimeng Yu and H. S P Wong. Compact modeling of conducting-bridge random-access memory (cbram). *Electron Devices, IEEE Transactions on*, 58(5):1352–1360, May 2011. ISSN 0018-9383. doi: 10.1109/TED.2011.2116120. xiii, 33

[14] A.W. Topol, D.C.La Tulipe, L. Shi, D.J. Frank, K. Bernstein, S.E. Steen, A. Kumar, G.U. Singco, A.M. Young, K.W. Guarini, and M. Ieong. Three-dimensional integrated circuits. *IBM Journal of Research and Development*, 50(4.5):491–506, July 2006. ISSN 0018-8646. doi: 10.1147/rd.504.0491. xiii, 36, 37, 38

[15] Gabriel H. Loh, Yuan Xie, and Bryan Black. Processor design in 3d die-stacking technologies. *Micro, IEEE*, 27(3):31–48, May 2007. ISSN 0272-1732. doi: 10.1109/MM.2007.59. xiii, 39

[16] Ian Kuon, Russell Tessier, and Jonathan Rose. Fpga architecture: Survey and challenges. *Found. Trends Electron. Des. Autom.*, 2(2):135–253, February 2008. ISSN 1551-3939. doi: 10.1561/1000000005. URL http://dx.doi.org/10.1561/1000000005. xiii, 46

[17] Hraziia, Adam Makosiej, Giorgio Palma, Jean-Michel Portal, Marc Bocquet, Olivier Thomas, Fabien Clermidy, Marina Reyboz, Santhosh Onkaraiah, Christophe Muller, Damien Deleruyelle, Andrei Vladimirescu, Amara Amara, and Costin Anghel. Operation and stability analysis of bipolar oxrram-based non-volatile 8t2r {SRAM} as solution for information back-up. *Solid-State Electronics*, 90(0):99 – 106, 2013. ISSN 0038-1101. doi: http://dx.doi.org/10.1016/j.sse.2013.02.045. URL http://www.sciencedirect.com/science/article/pii/S0038110113001068. Selected papers from {EUROSOI} 2012. xiv, xvi, 66, 67, 169

[18] S. Onkaraiah, M. Reyboz, F. Clermidy, J. Portal, M. Bocquet, C. Muller, H. Hraziia, C. Anghel, and A Amara. Bipolar reram based non-volatile flip-flops for low-power

architectures. In *New Circuits and Systems Conference (NEWCAS), 2012 IEEE 10th International*, pages 417–420, June 2012. doi: 10.1109/NEWCAS.2012.6329045. xvii, 71, 170, 175

[19] Gordon E. Moore. Cramming more components onto integrated circuits. *Electronics*, 38 (8), April 1965. 1, 157

[20] R.H. Dennard, F.H. Gaensslen, V.L. Rideout, E. Bassous, and AR. LeBlanc. Design of ion-implanted mosfet's with very small physical dimensions. *Solid-State Circuits, IEEE Journal of*, 9(5):256–268, Oct 1974. ISSN 0018-9200. doi: 10.1109/JSSC.1974.1050511. 2, 158

[21] Nir Magen, Avinoam Kolodny, Uri Weiser, and Nachum Shamir. Interconnect-power dissipation in a microprocessor. In *Proceedings of the 2004 International Workshop on System Level Interconnect Prediction*, SLIP '04, pages 7–13, New York, NY, USA, 2004. ACM. ISBN 1-58113-818-0. doi: 10.1145/966747.966750. URL `http://doi.acm.org/10.1145/966747.966750`. 3, 158

[22] Ian Kuon and J. Rose. Measuring the gap between fpgas and asics. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, 26(2):203–215, Feb 2007. ISSN 0278-0070. doi: 10.1109/TCAD.2006.884574. 5, 12, 160

[23] R. Hartenstein. A decade of reconfigurable computing: a visionary retrospective. In *Design, Automation and Test in Europe, 2001. Conference and Exhibition 2001. Proceedings*, pages 642–649, 2001. doi: 10.1109/DATE.2001.915091. 12

[24] Nicolas Telle, Wayne Luk, and RayC.C. Cheung. Customising hardware designs for elliptic curve cryptography. In AndyD. Pimentel and Stamatis Vassiliadis, editors, *Computer Systems: Architectures, Modeling, and Simulation*, volume 3133 of *Lecture Notes in Computer Science*, pages 274–283. Springer Berlin Heidelberg, 2004. ISBN 978-3-540-22377-1. doi: 10.1007/978-3-540-27776-7_29. URL `http://dx.doi.org/10.1007/978-3-540-27776-7_29`. 12

[25] Greg Stitt, Frank Vahid, and Shawn Nematbakhsh. Energy savings and speedups from partitioning critical software loops to hardware in embedded systems. *ACM Trans. Embed. Comput. Syst.*, 3(1):218–232, February 2004. ISSN 1539-9087. doi: 10.1145/972627.972637. URL `http://doi.acm.org/10.1145/972627.972637`. 12

[26] S. E. Wahlstrom. Programmable logic arrays, cheaper by the millions. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, 40:90–95, Dec. 1967. 14

[27] W. Carter, K. Duong, R. H. Freeman, H. Hsieh, J. Y. Ja, J. E. Mahoney, L. T. Ngo, and S. L. Sze. A user programmable reconfiguration gate array. In *Proceedings of the IEEE Custom Integrated Circuits Conference*, pages 233–235, 1986. 14, 15

[28] Plessey Semiconductor. Era60100 preliminary data sheet, 1989. 15

[29] AE. Gamal. Two-dimensional stochastic model for interconnections in master slice integrated circuits. *Circuits and Systems, IEEE Transactions on*, 28(2):127–138, Feb 1981. ISSN 0098-4094. doi: 10.1109/TCS.1981.1084958. 15

[30] S.C. Wong, H.C. So, J.H. Ou, and J. Costello. A 5000-gate cmos epld with multiple logic and interconnect arrays. In *Custom Integrated Circuits Conference, 1989., Proceedings of the IEEE 1989*, pages 5.8/1–5.8/4, May 1989. doi: 10.1109/CICC.1989.56697. 15

[31] Satwant Singh, Jonathan Rose, Paul Chow, and David Lewis. The effect of logic block architecture on fpga performance. *IEEE Journal of Solid-State Circuits*, 27:281–287, 1992. 17

[32] M. Hutton, D. Lewis, B. Pedersen, J. Schleicher, R. Yuan, G. Baeckler, A. Lee, R. Saini, , and H. Kim. Fracturable fpga logic elements, cp-01006-1.0, 2006. URL `http://www.altera.com/literature/cp/cp-01006.pdf`. 17

[33] Matt Klein. Xilinx - power consumption at 40 and 45 nm, wp298 (v1.0), 2009. URL `www.xilinx.com/support/documentation/white_papers/wp298.pdf`. 17, 20

[34] Steven Joseph Edward Wilton. *Architectures and Algorithms for Field-programmable Gate Arrays with Embedded Memory*. PhD thesis, Toronto, Ont., Canada, Canada, 1997. AAINQ28082. 22

[35] G. Lemieux, E. Lee, M. Tom, and A Yu. Directional and single-driver wires in fpga interconnect. In *Field-Programmable Technology, 2004. Proceedings. 2004 IEEE International Conference on*, pages 41–48, Dec 2004. doi: 10.1109/FPT.2004.1393249. 22

[36] Man-Ho Ho, Yan-Qing Ai, T.C.-P. Chau, S.C.L. Yuen, Chiu-Sing Choy, P.H.W. Leong, and Kong-Pang Pun. Architecture and design flow for a highly efficient structured asic. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, 21(3):424–433, March 2013. ISSN 1063-8210. doi: 10.1109/TVLSI.2012.2190478. 26

[37] Satish Sivaswamy, Gang Wang, Cristinel Ababei, Kia Bazargan, Ryan Kastner, and Eli Bozorgzadeh. Harp: Hard-wired routing pattern fpgas. In *Proceedings of the 2005 ACM/SIGDA 13th International Symposium on Field-programmable Gate Arrays*, FPGA '05, pages 21–29, New York, NY, USA, 2005. ACM. ISBN 1-59593-029-9. doi: 10.1145/1046192.1046196. URL `http://doi.acm.org/10.1145/1046192.1046196`. 26

[38] Gary K. Yeap. *Practical Low Power Digital VLSI Design*. Kluwer Academic Publishers, Norwell, MA, USA, 1998. ISBN 0-7923-8009-6. 27

[39] L.J. Schwee, P. E. Hunter, K. A. Restorff, and M. T. Shephard. The concept and initial studies of a crosstie random access memory (cram). *Journal of Applied Physics*, 53(3): 2762–2764, Mar 1982. ISSN 0021-8979. doi: 10.1063/1.330958. 28

[40] A.V. Pohm, J. S T Huang, J.M. Daughton, D.R. Krahn, and V. Mehra. The design of a one megabit non-volatile m-r memory chip using 1.5 times;5 mu;m cells. *Magnetics, IEEE Transactions on*, 24(6):3117–3119, Nov 1988. ISSN 0018-9464. doi: 10.1109/20.92353. 28

[41] J. S. Moodera, Lisa R. Kinder, Terrilyn M. Wong, and R. Meservey. Large magnetoresistance at room temperature in ferromagnetic thin film tunnel junctions. *Phys. Rev. Lett.*, 74:3273–3276, Apr 1995. doi: 10.1103/PhysRevLett.74.3273. URL `http://link.aps.org/doi/10.1103/PhysRevLett.74.3273`. 28

[42] S. Tehrani, J.M. Slaughter, E. Chen, M. Durlam, J. Shi, and M. DeHerren. Progress and outlook for mram technology. *Magnetics, IEEE Transactions on*, 35(5):2814–2819, Sep 1999. ISSN 0018-9464. doi: 10.1109/20.800991. 28

[43] R. H. Koch, G. Grinstein, G. A. Keefe, Yu Lu, P. L. Trouilloud, W. J. Gallagher, and S. S. P. Parkin. Thermally assisted magnetization reversal in submicron-sized magnetic thin films. *Phys. Rev. Lett.*, 84:5419–5422, Jun 2000. doi: 10.1103/PhysRevLett.84.5419. URL `http://link.aps.org/doi/10.1103/PhysRevLett.84.5419`. 28

[44] I L Prejbeanu, M Kerekes, R C Sousa, H Sibuet, O Redon, B Dieny, and J P Nozières. Thermally assisted mram. *Journal of Physics: Condensed Matter*, 19(16):165218, 2007. URL `http://stacks.iop.org/0953-8984/19/i=16/a=165218`. 28

[45] J.C. Slonczewski. Current-driven excitation of magnetic multilayers. *Journal of Magnetism and Magnetic Materials*, 159(1–2):L1 – L7, 1996. ISSN 0304-8853. doi: http://dx.doi.org/10.1016/0304-8853(96)00062-5. URL `http://www.sciencedirect.com/science/article/pii/0304885396000625`. 28

[46] M. Hosomi, H. Yamagishi, T. Yamamoto, K. Bessho, Y. Higo, K. Yamane, H. Yamada, M. Shoji, H. Hachino, C. Fukumoto, H. Nagao, and H. Kano. A novel nonvolatile memory with spin torque transfer magnetization switching: spin-ram. In *Electron Devices Meeting, 2005. IEDM Technical Digest. IEEE International*, pages 459–462, Dec 2005. doi: 10.1109/IEDM.2005.1609379. 29

[47] T. Kawahara, R. Takemura, K. Miura, J. Hayakawa, S. Ikeda, Young Min Lee, R. Sasaki, Y. Goto, K. Ito, T. Meguro, F. Matsukura, Hiromasa Takahashi, Hideyuki Matsuoka, and H. Ohno. 2 mb spram (spin-transfer torque ram) with bit-by-bit bi-directional current write and parallelizing-direction current read. *Solid-State Circuits, IEEE Journal of*, 43(1):109–120, Jan 2008. ISSN 0018-9200. doi: 10.1109/JSSC.2007.909751. 29

[48] R. Takemura, T. Kawahara, K. Miura, H. Yamamoto, J. Hayakawa, N. Matsuzaki, K. Ono, M. Yamanouchi, K. Ito, Hiromasa Takahashi, S. Ikeda, H. Hasegawa, Hideyuki Matsuoka, and H. Ohno. A 32-mb spram with 2t1r memory cell, localized bi-directional write driver and '1'/'0' dual-array equalized reference scheme. *Solid-State Circuits, IEEE Journal of*, 45(4):869–879, April 2010. ISSN 0018-9200. doi: 10.1109/JSSC.2010.2040120. 29

[49] K. Tsuchida, T. Inaba, K. Fujita, Y. Ueda, T. Shimizu, Y. Asao, T. Kajiyama, M. Iwayama, K. Sugiura, S. Ikegawa, T. Kishi, T. Kai, M. Amano, N. Shimomura, H. Yoda, and Y. Watanabe. A 64mb mram with clamped-reference and adequate-reference schemes. In *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2010 IEEE International*, pages 258–259, Feb 2010. doi: 10.1109/ISSCC.2010.5433948. 29

[50] Taehui Na, Kyungho Ryu, Jisu Kim, S.H. Kang, and Seong-Ook Jung. A comparative study of stt-mtj based non-volatile flip-flops. In *Circuits and Systems (ISCAS), 2013 IEEE International Symposium on*, pages 109–112, May 2013. doi: 10.1109/ISCAS.2013.6571794. 29

[51] E. Deng, Yue Zhang, J.-O. Klein, D. Ravelsona, C. Chappert, and Weisheng Zhao. Low power magnetic full-adder based on spin transfer torque mram. *Magnetics, IEEE Transactions on*, 49(9):4982–4987, Sept 2013. ISSN 0018-9464. doi: 10.1109/TMAG.2013.2245911. 29

[52] N. Sakimura, Y. Tsuji, R. Nebashi, H. Honjo, A. Morioka, K. Ishihara, K. Kinoshita, S. Fukami, S. Miura, N. Kasai, T. Endoh, H. Ohno, T. Hanyu, and T. Sugibayashi. 10.5 a 90nm 20mhz fully nonvolatile microcontroller for standby-power-critical applications. In *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2014 IEEE International*, pages 184–185, Feb 2014. doi: 10.1109/ISSCC.2014.6757392. 29

[53] S. Tehrani, Jon M. Slaughter, M. DeHerrera, B.N. Engel, N.D. Rizzo, J. Salter, M. Durlam, R.W. Dave, J. Janesky, B. Butcher, K. Smith, and G. Grynkewich. Magnetoresistive random access memory using magnetic tunnel junctions. *Proceedings of the IEEE*, 91(5):703–714, May 2003. ISSN 0018-9219. doi: 10.1109/JPROC.2003.811804. 29

[54] Y. Huai. Spin-transfer torque MRAM (STT-MRAM): Challenges and prospects. *AAPPS Bulletin*, 18(6), 2008. 29

[55] Haiwen Xi, J. Stricklin, Hai Li, Yiran Chen, Xiaobin Wang, Yuankai Zheng, Zheng Gao, and M.X. Tang. Spin transfer torque memory with thermal assist mechanism: A case study. *Magnetics, IEEE Transactions on*, 46(3):860–865, March 2010. ISSN 0018-9464. doi: 10.1109/TMAG.2009.2033674. 29

[56] E. Chen, D. Apalkov, Z. Diao, A. Driskill-Smith, D. Druist, D. Lottis, V. Nikitin, X. Tang, S. Watts, S. Wang, S.A. Wolf, A. W. Ghosh, J.W. Lu, S. J. Poon, M. Stan, W.H. Butler, S. Gupta, C. K A Mewes, T. Mewes, and P.B. Visscher. Advances and future prospects of spin-transfer torque random access memory. *Magnetics, IEEE Transactions on*, 46(6):1873–1878, June 2010. ISSN 0018-9464. doi: 10.1109/TMAG.2010.2042041. 29

[57] Xiuyuan Bi, Hai Li, and Xiaobin Wang. Stt-ram cell design considering cmos and mtj temperature dependence. *Magnetics, IEEE Transactions on*, 48(11):3821–3824, Nov 2012. ISSN 0018-9464. doi: 10.1109/TMAG.2012.2200469. 29

[58] K. Huang, Y. Ha, R. Zhao, A. Kumar, and Y. Lian. A low active leakage and high reliability phase change memory (pcm) based non-volatile fpga storage element. *Circuits and Systems I: Regular Papers, IEEE Transactions on*, PP(99):1–1, 2014. ISSN 1549-8328. doi: 10.1109/TCSI.2014.2312499. 30, 40

[59] Daolin Cai, Houpeng Chen, Qian Wang, Yifeng Chen, Zhitang Song, Guanping Wu, and Songlin Feng. An 8-mb phase-change random access memory chip based on a resistor-on-via-stacked-plug storage cell. *Electron Device Letters, IEEE*, 33(9):1270–1272, Sept 2012. ISSN 0741-3106. doi: 10.1109/LED.2012.2204952. 30

[60] Livio Baldi and Gurtej Sandhu. Emerging memories. In *Solid-State Device Research Conference (ESSDERC), 2013 Proceedings of the European*, pages 30–36, Sept 2013. doi: 10.1109/ESSDERC.2013.6818813. 30

[61] Q. Hubert, C. Jahan, A. Toffoli, V. Delaye, D. Lafond, H. Grampeix, and B. De Salvo. Detailed Analysis of the Role of Thin HfO_2 Interfacial Layer in $Ge_2Sb_2Te_5$ Based PCM. *Electron Devices, IEEE Transactions on*, 60(7):2268–2275, July 2013. ISSN 0018-9383. doi: 10.1109/TED.2013.2264323. 30

[62] N. Kanán, A. Faraclas, N. Williams, H. Silva, and A. Gokirmak. Computational analysis of rupture-oxide phase-change memory cells. *Electron Devices, IEEE Transactions on*, 60 (5):1649–1655, May 2013. ISSN 0018-9383. doi: 10.1109/TED.2013.2255130. 30

[63] Youngdon Choi, Ickhyun Song, Mu-Hui Park, Hoeju Chung, Sanghoan Chang, Beakhyoung Cho, Jinyoung Kim, Younghoon Oh, Duckmin Kwon, Jung Sunwoo, Junho Shin, Yoohwan Rho, Changsoo Lee, Min-Gu Kang, Jaeyun Lee, Yongjin Kwon, Soehee Kim, Jaehwan Kim, Yong-Jun Lee, Qi Wang, Sooho Cha, Sujin Ahn, H. Horii, Jaewook Lee, Kisung Kim, Hansung Joo, Kwangjin Lee, Yeong-Taek Lee, Jeihwan Yoo, and G. Jeong. A 20nm 1.8v 8gb pram with 40mb/s program bandwidth. In *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2012 IEEE International*, pages 46–48, Feb 2012. doi: 10.1109/ISSCC.2012.6176872. 30

[64] Shimin Chen, Phillip B Gibbons, and Suman Nath. Rethinking database algorithms for phase change memory. In *CIDR*, pages 21–31, 2011. 30

[65] Shimeng Yu and H. S P Wong. A phenomenological model for the reset mechanism of metal oxide rram. *Electron Device Letters, IEEE*, 31(12):1455–1457, Dec 2010. ISSN 0741-3106. doi: 10.1109/LED.2010.2078794. 32

[66] H.Y. Lee, P.S. Chen, T. Y Wu, Y.S. Chen, C.C. Wang, P.J. Tzeng, C. H Lin, F. Chen, C.H. Lien, and M. J Tsai. Low power and high speed bipolar switching with a thin reactive ti buffer layer in robust hfo2 based rram. In *Electron Devices Meeting, 2008. IEDM 2008. IEEE International*, pages 1–4, Dec 2008. doi: 10.1109/IEDM.2008.4796677. 32

[67] M. Bocquet, D. Deleruyelle, H. Aziza, C. Muller, J.-M. Portal, T. Cabout, and E. Jalaguier. Robust compact model for bipolar oxide-based resistive switching memories. *Electron Devices, IEEE Transactions on*, 61(3):674–681, March 2014. ISSN 0018-9383. doi: 10.1109/TED.2013.2296793. 32

[68] E. Vianello, O. Thomas, M. Harrand, S. Onkaraiah, T. Cabout, B. Traore, T. Diokh, H. Oucheikh, L. Perniola, G. Molas, P. Blaise, J.F. Nodin, E. Jalaguier, and B. De Salvo. Back-end 3d integration of hfo2-based rrams for low-voltage advanced ic digital design. In *IC Design Technology (ICICDT), 2013 International Conference on*, pages 235–238, May 2013. doi: 10.1109/ICICDT.2013.6563344. 32

[69] A. Kawahara, R. Azuma, Y. Ikeda, K. Kawai, Y. Katoh, K. Tanabe, T. Nakamura, Y. Sumimoto, N. Yamada, N. Nakai, S. Sakamoto, Y. Hayakawa, K. Tsuji, S. Yoneda, A. Himeno, K. Origasa, K. Shimakawa, T. Takagi, T. Mikawa, and K. Aono. An 8mb multi-layered cross-point reram macro with 443mb/s write throughput. In *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2012 IEEE International*, pages 432–434, Feb 2012. doi: 10.1109/ISSCC.2012.6177078. 32

[70] Tz yi Liu, Tian Hong Yan, R. Scheuerlein, Yingchang Chen, J.K. Lee, G. Balakrishnan, G. Yee, H. Zhang, A. Yap, J. Ouyang, T. Sasaki, S. Addepalli, A. Al-Shamma, Chin-Yu Chen, M. Gupta, G. Hilton, S. Joshi, A. Kathuria, V. Lai, D. Masiwal, M. Matsumoto, A. Nigam, A. Pai, J. Pakhale, Chang Hua Siau, Xiaoxia Wu, R. Yin, Liping Peng, Jang Yong Kang, S. Huynh, Huijuan Wang, N. Nagel, Y. Tanaka, M. Higashitani, T. Minvielle, C. Gorla, T. Tsukamoto, T. Yamaguchi, M. Okajima, T. Okamura, S. Takase, T. Hara, H. Inoue, L. Fasoli, M. Mofidi, R. Shrivastava, and K. Quader. A 130.7mm2 2-layer 32gb reram memory device in 24nm technology. In *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2013 IEEE International*, pages 210–211, Feb 2013. doi: 10.1109/ISSCC.2013.6487703. 32

[71] R. Symanczyk, M. Balakrishnan, C. Gopalan, T. Happ, M. Kozicki, M. Kund, T. Mikolajick, M. Mitkova, M. Park, C. Pinnow, J. Robertson, and K. Ufert. Electrical characterization of solid state ionic memory elements. In *Proc. Non-Volatile Memory Technology Symp.*, pages 17–1–17–6, 2003. 32

[72] T. Sakamoto, N. Banno, Noriyuki Iguchi, H. Kawaura, H. Sunamura, S. Fujieda, Kazuya Terabe, Tsuyoshi Hasegawa, and Masakazu Aono. A ta_2o_5 solid-electrolyte switch with improved reliability. In *VLSI Technology, 2007 IEEE Symposium on*, pages 38–39, June 2007. doi: 10.1109/VLSIT.2007.4339718. 32, 33

[73] G. Palma, E. Vianello, O. Thomas, M. Suri, S. Onkaraiah, A. Toffoli, C. Carabasse, M. Bernard, A. Roule, O. Pirrotta, G. Molas, and B. De Salvo. Interface engineering of ag-ges_2 -based conductive bridge ram for reconfigurable logic applications. *Electron Devices, IEEE Transactions on*, 61(3):793–800, March 2014. ISSN 0018-9383. doi: 10.1109/TED.2014.2301694. 33, 88

[74] Ilia Valov, Rainer Waser, John R Jameson, and Michael N Kozicki. Electrochemical metallization memories—fundamentals, applications, prospects. *Nanotechnology*, 22(25): 254003, 2011. URL `http://stacks.iop.org/0957-4484/22/i=25/a=254003`. 33

[75] S. Dietrich, M. Angerbauer, M. Ivanov, D. Gogl, H. Hoenigschmid, M. Kund, C. Liaw, M. Markert, R. Symanczyk, L. Altimime, S. Bournat, and G. Mueller. A nonvolatile 2-mbit cbram memory core featuring advanced read and program control. *Solid-State Circuits, IEEE Journal of*, 42(4):839–845, April 2007. ISSN 0018-9200. doi: 10.1109/JSSC.2007.892207. 33

[76] M. Kund, G. Beitel, C.-U. Pinnow, T. Rohr, J. Schumann, R. Symanczyk, K.-D. Ufert, and G. Muller. Conductive bridging ram (cbram): an emerging non-volatile memory technology scalable to sub 20nm. In *Electron Devices Meeting, 2005. IEDM Technical Digest. IEEE International*, pages 754–757, Dec 2005. doi: 10.1109/IEDM.2005.1609463. 33

[77] Yuhao Wang, Hao Yu, and Wei Zhang. Nonvolatile cbram-crossbar-based 3-d-integrated hybrid memory for data retention. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, 22(5):957–970, May 2014. ISSN 1063-8210. doi: 10.1109/TVLSI.2013.2265754. 33

[78] D. Takashima, Yasushi Nagadomi, K. Hatsuda, Y. Watanabe, and S. Fujii. A 128 mb chain feram and system design for hdd application and enhanced hdd performance. *Solid-State Circuits, IEEE Journal of*, 46(2):530–536, Feb 2011. ISSN 0018-9200. doi: 10.1109/JSSC.2010.2091324. 34

[79] M. Koga, M. Iida, M. Amagasaki, Y. Ichida, M. Saji, J. Iida, and T. Sueyoshi. First prototype of a genuine power-gatable reconfigurable logic chip with feram cells. In *Field Programmable Logic and Applications (FPL), 2010 International Conference on*, pages 298–303, Aug 2010. doi: 10.1109/FPL.2010.67. 34

[80] Yu-Chung Lien, Jia-Min Shieh, Wen-Hsien Huang, Wei-Shang Hsieh, Cheng-Hui Tu, Chieh Wang, Chang-Hong Shen, Tung-Huan Chou, Min-Cheng Chen, J.Y. Huang, Ci-Ling Pan, Yin-Chieh Lai, Chenming Hu, and Fu-Liang Yang. 3d ferroelectric-like nvm/cmos hybrid chip by sub-400c sequential layered integration. In *Electron Devices Meeting (IEDM), 2012 IEEE International*, pages 33.6.1–33.6.4, Dec 2012. doi: 10.1109/IEDM.2012.6479160. 34

[81] W.R. Davis, J. Wilson, S. Mick, J. Xu, Hao Hua, C. Mineo, A.M. Sule, M. Steer, and P.D. Franzon. Demystifying 3d ics: the pros and cons of going vertical. *Design Test of Computers, IEEE*, 22(6):498–510, Nov 2005. ISSN 0740-7475. doi: 10.1109/MDT.2005.136. 36

[82] M. Koyanagi, T. Fukushima, and T. Tanaka. High-density through silicon vias for 3-d lsis. *Proceedings of the IEEE*, 97(1):49–59, Jan 2009. ISSN 0018-9219. doi: 10.1109/JPROC.2008.2007463. 36

[83] P. Batude, M. Vinet, A. Pouydebasque, C. Le Royer, B. Previtali, C. Tabone, J. Hartmann, L. Sanchez, L. Baud, V. Carron, A. Toffoli, F. Allain, V. Mazzocchi, D. Lafond, S. Deleonibus, and O. Faynot. 3d monolithic integration. In *Circuits and Systems (ISCAS), 2011 IEEE International Symposium on*, pages 2233–2236, May 2011. doi: 10.1109/ISCAS.2011.5938045. 36

[84] T. Karnik, D. Somasekhar, and S. Borkar. Microprocessor system applications and challenges for through-silicon-via-based three-dimensional integration. *Computers Digital Techniques, IET*, 5(3):205–212, May 2011. ISSN 1751-8601. doi: 10.1049/iet-cdt.2009.0126. 36

[85] Soon-Moon Jung, Youngseop Rah, Taehong Ha, Hanbyung Park, Chulsoon Chang, Seungchul Lee, Jongho Yun, Wonsuk Cho, Hoon Lim, Jaikyun Park, Jaehun Jeong, Byoungkeun Son, Jaehoon Jang, Bonghyun Choi, Hoosung Cho, and Kinam Kim. Highly cost effective and high performance 65nm s3 (stacked single-crystal si) sram technology with 25f2, 0.16um2 cell and doubly stacked sstft cell transistors for ultra high density and high speed applications. In *VLSI Technology, 2005. Digest of Technical Papers. 2005 Symposium on*, pages 220–221, June 2005. doi: 10.1109/.2005.1469275. 38, 40

[86] Uksong Kang, Hoe-Ju Chung, Seongmoo Heo, Soon-Hong Ahn, Hoon Lee, Soo-Ho Cha, Jaesung Ahn, DukMin Kwon, Jin-Ho Kim, Jae-Wook Lee, Han-Sung Joo, Woo-Seop Kim, Hyun-Kyung Kim, Eun-Mi Lee, So-Ra Kim, Keum-Hee Ma, Dong-Hyun Jang, Nam-Seog Kim, Man-Sik Choi, Sae-Jang Oh, Jung-Bae Lee, Tae-Kyung Jung, Jei-Hwan Yoo, and Changhyun Kim. 8gb 3d ddr3 dram using through-silicon-via technology. In *Solid-State Circuits Conference - Digest of Technical Papers, 2009. ISSCC 2009. IEEE International*, pages 130–131,131a, Feb 2009. doi: 10.1109/ISSCC.2009.4977342. 39

[87] B. Black, M. Annavaram, N. Brekelbaum, J. DeVale, Lei Jiang, G.H. Loh, D. McCauley, P. Morrow, D.W. Nelson, D. Pantuso, P. Reed, J. Rupley, Sadasivan Shankar, J. Shen, and C. Webb. Die stacking (3d) microarchitecture. In *Microarchitecture, 2006. MICRO-39. 39th Annual IEEE/ACM International Symposium on*, pages 469–479, Dec 2006. doi: 10.1109/MICRO.2006.18. 39

[88] O. Thomas, M. Vinet, O. Rozeau, P. Batude, and A. Valentian. Compact 6t sram cell with robust read/write stabilizing design in 45nm monolithic 3d ic technology. In *IC Design and Technology, 2009. ICICDT '09. IEEE International Conference on*, pages 195–198, May 2009. doi: 10.1109/ICICDT.2009.5166294. 40

[89] J. Derakhshandeh, N. Golshani, R. Ishihara, M.R. Tajari Mofrad, M. Robertson, T. Morrison, and C. I M Beenakker. Monolithic 3-d integration of sram and image sensor using

two layers of single-grain silicon. *Electron Devices, IEEE Transactions on*, 58(11):3954–3961, Nov 2011. ISSN 0018-9383. doi: 10.1109/TED.2011.2163720. 40

[90] Chang Liu and Sung-Kyu Lim. A design tradeoff study with monolithic 3d integration. In *Quality Electronic Design (ISQED), 2012 13th International Symposium on*, pages 529–536, March 2012. doi: 10.1109/ISQED.2012.6187545. 40, 104

[91] Young-Joon Lee, D. Limbrick, and Sung Kyu Lim. Power benefit study for ultra-high density transistor-level monolithic 3d ics. In *Design Automation Conference (DAC), 2013 50th ACM / EDAC / IEEE*, pages 1–10, May 2013. 40

[92] S. Bobba, A Chakraborty, O. Thomas, P. Batude, T. Ernst, O. Faynot, D.Z. Pan, and G. De Micheli. Celoncel: Effective design technique for 3-d monolithic integration targeting high performance integrated circuits. In *Design Automation Conference (ASP-DAC), 2011 16th Asia and South Pacific*, pages 336–343, Jan 2011. doi: 10.1109/ASPDAC.2011.5722210. 40, 104

[93] H. Sarhan, S. Thuries, O. Billoint, and F. Clermidy. 3dcob: A new design approach for monolithic 3d integrated circuits. In *Design Automation Conference (ASP-DAC), 2014 19th Asia and South Pacific*, pages 79–84, Jan 2014. doi: 10.1109/ASPDAC.2014.6742870. 40, 104

[94] S. Paul, S. Mukhopadhyay, and S. Bhunia. A circuit and architecture codesign approach for a hybrid cmos sttram nonvolatile fpga. *Nanotechnology, IEEE Transactions on*, 10 (3):385–394, May 2011. ISSN 1536-125X. doi: 10.1109/TNANO.2010.2041555. 40

[95] W. Zhao, E. Belhaire, Q. Mistral, E. Nicolle, T. Devolder, and C. Chappert. Integration of spin-ram technology in fpga circuits. In *Solid-State and Integrated Circuit Technology, 2006. ICSICT '06. 8th International Conference on*, pages 799–802, Oct 2006. doi: 10.1109/ICSICT.2006.306511. 40

[96] Y. Guillemenet, L. Torres, and G. Sassatelli. Non-volatile run-time field-programmable gate arrays structures using thermally assisted switching magnetic random access memories. *Computers Digital Techniques, IET*, 4(3):211–226, May 2010. ISSN 1751-8601. doi: 10.1049/iet-cdt.2009.0019. 40

[97] Weisheng Zhao, Eric Belhaire, Claude Chappert, and Pascale Mazoyer. Spin transfer torque (stt)-mram–based runtime reconfiguration fpga circuit. *ACM Trans. Embed. Comput. Syst.*, 9(2):14:1–14:16, October 2009. ISSN 1539-9087. doi: 10.1145/1596543.1596548. URL `http://doi.acm.org/10.1145/1596543.1596548`. 40

[98] P.-E. Gaillardon, M.H. Ben-Jamaa, G.B. Beneventi, F. Clermidy, and L. Perniola. Emerging memory technologies for reconfigurable routing in fpga architecture. In *Electronics, Circuits, and Systems (ICECS), 2010 17th IEEE International Conference on*, pages 62–65, Dec 2010. doi: 10.1109/ICECS.2010.5724454. 40

[99] Yibo Chen, Jishen Zhao, and Yuan Xie. 3d-nonfar: Three-dimensional non-volatile fpga architecture using phase change memory. In *Low-Power Electronics and Design (ISLPED), 2010 ACM/IEEE International Symposium on*, pages 55–60, Aug 2010. 40

[100] C. Y Wen, J. Li, S. Kim, M. Breitwisch, C. Lam, J. Paramesh, and L.T. Pileggi. A non-volatile look-up table design using pcm (phase-change memory) cells. In *VLSI Circuits (VLSIC), 2011 Symposium on*, pages 302–303, June 2011. 41

[101] J. Cong and Bingjun Xiao. Fpga-rpi: A novel fpga architecture with rram-based programmable interconnects. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, 22(4):864–877, April 2014. ISSN 1063-8210. doi: 10.1109/TVLSI.2013.2259512. 41

[102] Yi-Chung Chen, Hai Li, and Wei Zhang. A novel peripheral circuit for rram-based lut. In *Circuits and Systems (ISCAS), 2012 IEEE International Symposium on*, pages 1811–1814, May 2012. doi: 10.1109/ISCAS.2012.6271619. 41

[103] Young Yang Liauw, Zhiping Zhang, Wanki Kim, AE. Gamal, and S.S. Wong. Nonvolatile 3d-fpga with monolithically stacked rram-based configuration memory. In *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2012 IEEE International*, pages 406–408, Feb 2012. doi: 10.1109/ISSCC.2012.6177067. 41

[104] A. Gayasen, V. Narayanan, M. Kandemir, and Arifur Rahman. Designing a 3-d fpga: Switch box architecture and thermal issues. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, 16(7):882–893, July 2008. ISSN 1063-8210. doi: 10.1109/TVLSI.2008.2000456. 41, 130

[105] V. Pangracious, H. Mehrez, and Z. Marakchi. Architecture level tsv count minimization methodology for 3d tree-based fpga. In *Cool Chips XVI (COOL Chips), 2013 IEEE*, pages 1–3, April 2013. doi: 10.1109/CoolChips.2013.6547925. 41

[106] Kostas Siozios, Vasilis F. Pavlidis, and Dimitrios Soudris. A novel framework for exploring 3-d fpgas with heterogeneous interconnect fabric. *ACM Trans. Reconfigurable Technol. Syst.*, 5(1):4:1–4:23, March 2012. ISSN 1936-7406. doi: 10.1145/2133352.2133356. URL `http://doi.acm.org/10.1145/2133352.2133356`. 41

[107] Harry Sidiropoulos, Kostas Siozios, and Dimitrios Soudris. A novel 3-d {FPGA} architecture targeting communication intensive applications. *Journal of Systems Architecture*, 60 (1):32 – 39, 2014. ISSN 1383-7621. doi: http://dx.doi.org/10.1016/j.sysarc.2013.09.012. URL `http://www.sciencedirect.com/science/article/pii/S1383762113002609`. 41

[108] C. Ababei, Y. Feng, B. Goplen, Hushrav Mogal, Tianpei Zhang, K. Bazargan, and S. Sapatnekar. Placement and routing in 3d integrated circuits. *Design Test of Computers, IEEE*, 22(6):520–531, Nov 2005. ISSN 0740-7475. doi: 10.1109/MDT.2005.150. 41

[109] T. Hamada, Qian Zhao, M. Amagasaki, M. Iida, M. Kuga, and T. Sueyoshi. Three-dimensional stacking fpga architecture using face-to-face integration. In *Very Large Scale Integration (VLSI-SoC), 2013 IFIP/IEEE 21st International Conference on*, pages 192–197, Oct 2013. doi: 10.1109/VLSI-SoC.2013.6673274. 41

[110] B. Banijamali, S. Ramalingam, K. Nagarajan, and R. Chaware. Advanced reliability study of tsv interposers and interconnects for the 28nm technology fpga. In *Electronic Components and Technology Conference (ECTC), 2011 IEEE 61st*, pages 285–290, May 2011. doi: 10.1109/ECTC.2011.5898527. 42

[111] T. Naito, T. Ishida, T. Onoduka, M. Nishigoori, T. Nakayama, Y. Ueno, Y. Ishimoto, A. Suzuki, W. Chung, R. Madurawe, S. Wu, S. Ikeda, and H. Oyamatsu. World's first monolithic 3d-fpga with tft sram over 90nm 9 layer cu cmos. In *VLSI Technology (VLSIT), 2010 Symposium on*, pages 219–220, June 2010. doi: 10.1109/VLSIT.2010.5556234. 42

[112] P. Batude, T. Ernst, J. Arcamone, G. Arndt, P. Coudrain, and P.-E. Gaillardon. 3-d sequential integration: A key enabling technology for heterogeneous co-integration of new function with cmos. *Emerging and Selected Topics in Circuits and Systems, IEEE Journal on*, 2(4):714–722, Dec 2012. ISSN 2156-3357. doi: 10.1109/JETCAS.2012.2223593. 42

[113] P. Jamieson, K.B. Kent, F. Gharibian, and L. Shannon. Odin ii - an open-source verilog hdl synthesis tool for cad research. In *Field-Programmable Custom Computing Machines (FCCM), 2010 18th IEEE Annual International Symposium on*, pages 149–156, May 2010. doi: 10.1109/FCCM.2010.31. 48

[114] E.M. Sentovich, K.J. Singh, L. Lavagno, C. Moon, R. Murgai, A. Saldanha, H. Savoj, P.R. Stephan, Robert K. Brayton, and Alberto L. Sangiovanni-Vincentelli. Sis: A system for sequential circuit synthesis. Technical Report UCB/ERL M92/41, EECS Department, University of California, Berkeley, 1992. URL `http://www.eecs.berkeley.edu/Pubs/TechRpts/1992/2010.html`. 49

[115] J. Cong and Y. Ding. Flowmap: an optimal technology mapping algorithm for delay optimization in lookup-table based fpga designs. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, 13(1):1–12, Jan 1994. ISSN 0278-0070. doi: 10.1109/43.273754. 49

[116] BERKELEY LOGIC SYNTHESIS. Abc: a system for sequential synthesis and verification, release 70930, 2007. URL `http://www.eecs.berkeley.edu/~alanmi/abc/`. 49

[117] E. Ahmed and J. Rose. The effect of lut and cluster size on deep-submicron fpga performance and density. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, 12(3):288–298, March 2004. ISSN 1063-8210. doi: 10.1109/TVLSI.2004.824300. 49

[118] L.W. Hagen and A.B. Kahng. Combining problem reduction and adaptive multistart: a new technique for superior iterative partitioning. *Computer-Aided Design of Integrated*

Circuits and Systems, IEEE Transactions on, 16(7):709–717, Jul 1997. ISSN 0278-0070. doi: 10.1109/43.644032. 50

[119] Mehrdad Eslami Dehkordi and S.D. Brown. The effect of cluster packing and node duplication control in delay driven clustering. In *Field-Programmable Technology, 2002. (FPT). Proceedings. 2002 IEEE International Conference on*, pages 227–233, Dec 2002. doi: 10.1109/FPT.2002.1188686. 50

[120] J. Cong and Sung-Kyu Lim. Edge separability-based circuit clustering with application to multilevel circuit partitioning. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, 23(3):346–357, March 2004. ISSN 0278-0070. doi: 10.1109/TCAD.2004.823353. 50

[121] Alexander (Sandy) Marquardt, Vaughn Betz, and Jonathan Rose. Using cluster-based logic blocks and timing-driven packing to improve fpga speed and density. In *Proceedings of the 1999 ACM/SIGDA Seventh International Symposium on Field Programmable Gate Arrays*, FPGA '99, pages 37–46, New York, NY, USA, 1999. ACM. ISBN 1-58113-088-0. doi: 10.1145/296399.296426. URL `http://doi.acm.org/10.1145/296399.296426`. 50

[122] E. Bozorgzadeh, S. Ogrenci-Memik, and M. Sarrafzadeh. Rpack: routability-driven packing for cluster-based fpgas. In *Design Automation Conference, 2001. Proceedings of the ASP-DAC 2001. Asia and South Pacific*, pages 629–634, 2001. doi: 10.1109/ASPDAC.2001.913379. 50

[123] Amit Singh and Malgorzata Marek-Sadowska. Efficient circuit clustering for area and power reduction in fpgas. In *Proceedings of the 2002 ACM/SIGDA Tenth International Symposium on Field-programmable Gate Arrays*, FPGA '02, pages 59–66, New York, NY, USA, 2002. ACM. ISBN 1-58113-452-5. doi: 10.1145/503048.503058. URL `http://doi.acm.org/10.1145/503048.503058`. 50

[124] J. Cong and M. Romesis. Performance-driven multi-level clustering with application to hierarchical fpga mapping. In *Design Automation Conference, 2001. Proceedings*, pages 389–394, 2001. doi: 10.1109/DAC.2001.156171. 50

[125] J.Y. Lin, Deming Chen, and J. Cong. Optimal simultaneous mapping and clustering for fpga delay optimization. In *Design Automation Conference, 2006 43rd ACM/IEEE*, pages 472–477, 2006. doi: 10.1109/DAC.2006.229262. 50

[126] Dennis J.-H. Huang and Andrew B. Kahng. Partitioning-based standard-cell global placement with an exact objective. In *Proceedings of the 1997 International Symposium on Physical Design*, ISPD '97, pages 18–25, New York, NY, USA, 1997. ACM. ISBN 0-89791-927-0. doi: 10.1145/267665.267674. URL `http://doi.acm.org/10.1145/267665.267674`. 50

[127] C. J. Alpert, T. F. Chan, A. B. Kahng, I. L. Markov, and P. Mulet. Faster minimization of linear wirelength for global placement. *Trans. Comp.-Aided Des. Integ. Cir. Sys.*, 17(1):3–13, November 2006. ISSN 0278-0070. doi: 10.1109/43.673628. URL `http://dx.doi.org/10.1109/43.673628`. 50

[128] C. Sechen and A Sangiovanni-Vincentelli. The timberwolf placement and routing package. *Solid-State Circuits, IEEE Journal of*, 20(2):510–522, April 1985. ISSN 0018-9200. doi: 10.1109/JSSC.1985.1052337. 50

[129] C. Ebeling, L. McMurchie, S.A Hauck, and S. Burns. Placement and routing tools for the triptych fpga. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, 3(4):473–482, Dec 1995. ISSN 1063-8210. doi: 10.1109/92.475966. 51

[130] Alexander Marquardt, Vaughn Betz, and Jonathan Rose. Timing-driven placement for fpgas. In *Proceedings of the 2000 ACM/SIGDA Eighth International Symposium on Field Programmable Gate Arrays*, FPGA '00, pages 203–213, New York, NY, USA, 2000. ACM. ISBN 1-58113-193-3. doi: 10.1145/329166.329208. URL `http://doi.acm.org/10.1145/329166.329208`. 51, 55

[131] Guy G. F. Lemieux, Stephen D. Brown, and Daniel Vranesic. On two-step routing for fpgas. In *Proceedings of the 1997 International Symposium on Physical Design*, ISPD '97, pages 60–66, New York, NY, USA, 1997. ACM. ISBN 0-89791-927-0. doi: 10.1145/267665.267682. URL `http://doi.acm.org/10.1145/267665.267682`. 51

[132] Yao-Wen Chang, S. Thakur, K. Zhua, and D. F. Wong. A new global routing algorithm for fpgas. In *Computer-Aided Design, 1994., IEEE/ACM International Conference on*, pages 356–361, Nov 1994. doi: 10.1109/ICCAD.1994.629817. 51

[133] Larry McMurchie and Carl Ebeling. Pathfinder: A negotiation-based performance-driven router for fpgas. In *Proceedings of the 1995 ACM Third International Symposium on Field-programmable Gate Arrays*, FPGA '95, pages 111–117, New York, NY, USA, 1995. ACM. ISBN 0-89791-743-X. doi: 10.1145/201310.201328. URL `http://doi.acm.org/10.1145/201310.201328`. 51, 52

[134] Vaughn Betz and Jonathan Rose. Vpr: A new packing, placement and routing tool for fpga research. In *Proceedings of the 7th International Workshop on Field-Programmable Logic and Applications*, FPL '97, pages 213–222, London, UK, UK, 1997. Springer-Verlag. ISBN 3-540-63465-7. URL `http://dl.acm.org/citation.cfm?id=647924.738755`. 52

[135] J. Lamoureux and S. J E Wilton. Activity estimation for field-programmable gate arrays. In *Field Programmable Logic and Applications, 2006. FPL '06. International Conference on*, pages 1–8, Aug 2006. doi: 10.1109/FPL.2006.311199. 52, 55, 163

[136] Kara K. W. Poon, Andy Yan, and Steven J. E. Wilton. A flexible power model for fpgas. In *Proceedings of the Reconfigurable Computing Is Going Mainstream, 12th International Conference on Field-Programmable Logic and Applications*, FPL '02, pages 312–321, London, UK, UK, 2002. Springer-Verlag. ISBN 3-540-44108-5. URL `http://dl.acm.org/citation.cfm?id=647929.740248`. 52, 55, 56, 163

[137] P. Jamieson, W. Luk, S. J E Wilton, and G.A Constantinides. An energy and power consumption analysis of fpga routing architectures. In *Field-Programmable Technology, 2009. FPT 2009. International Conference on*, pages 324–327, Dec 2009. doi: 10.1109/FPT.2009.5377675. 52, 55, 106, 163

[138] Jason Luu, Ian Kuon, Peter Jamieson, Ted Campbell, Andy Ye, Wei Mark Fang, and Jonathan Rose. Vpr 5.0: Fpga cad and architecture exploration tools with single-driver routing, heterogeneity and process scaling. In *Proceedings of the ACM/SIGDA International Symposium on Field Programmable Gate Arrays*, FPGA '09, pages 133–142, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-410-2. doi: 10.1145/1508128.1508150. URL `http://doi.acm.org/10.1145/1508128.1508150`. 52

[139] S. Yang. *Logic Synthesis and Optimization Benchmarks User Guide: Version 3.0*. Microelectronics Center of North Carolina (MCNC), 1991. URL `http://books.google.fr/books?id=7ruGuAAACAAJ`. 57

[140] C. Cagli, J. Buckley, V. Jousseaume, T. Cabout, A Salaun, H. Grampeix, J-F Nodin, H. Feldis, A Persico, J. Cluzel, P. Lorenzi, L. Massari, R. Rao, F. Irrera, F. Aussenac, C. Carabasse, M. Coue, P. Calka, E. Martinez, L. Perniola, P. Blaise, Z. Fang, Y. H. Yu, G. Ghibaudo, D. Deleruyelle, M. Bocquet, C. Muller, A Padovani, O. Pirrotta, L. Vandelli, L. Larcher, G. Reimbold, and B. De Salvo. Experimental and theoretical study of electrode effects in hfo2 based rram. In *Electron Devices Meeting (IEDM), 2011 IEEE International*, pages 28.7.1–28.7.4, Dec 2011. doi: 10.1109/IEDM.2011.6131634. 66, 69, 73

[141] S. Onkaraiah, O. Turkyilmaz, M. Reyboz, F. Clermidy, E. Vianello, J.-M. Portal, and C. Muller. A hybrid cbram/cmos look-up-table structure for improving performance efficiency of field-programmable-gate-array. In *Circuits and Systems (ISCAS), 2013 IEEE International Symposium on*, pages 2440–2443, May 2013. doi: 10.1109/ISCAS.2013.6572372. 74, 99, 169

[142] Kyung JoonHan, N. Chan, SungRae Kim, Ben Leung, V. Hecht, and B. Cronquist. A novel flash-based fpga technology with deep trench isolation. In *Non-Volatile Semiconductor Memory Workshop, 2007 22nd IEEE*, pages 32–33, Aug 2007. doi: 10.1109/NVSMW.2007.4290569. 78

[143] Actel. Actel igloo, 2011. URL `http://www.microsemi.com/products/fpga-soc/fpga/igloo-overview`. 78

[144] Lattice. Latticexp non-volatile fpga, 2011. URL http://www.latticesemi.com/products/maturedevices/xp/index.cfm. 78

[145] Xilinx. Xilinx spartan^{TM}-3an fpgas, 2011. URL http://public.itrs.net/Links/2013ITRS/2013Chapters/2013SysDrivers_Summary.pdf. 78

[146] ITRS. Executive summary, 2013. URL http://www.itrs.net/links/2013ITRS/2013Chapters/2013ExecutiveSummary.pdf. 78

[147] Y. S Chen, H.Y. Lee, P.S. Chen, P.Y. Gu, C.W. Chen, W.P. Lin, W.H. Liu, Y.Y. Hsu, S.S. Sheu, P.-C. Chiang, W-S Chen, F.T. Chen, C.H. Lien, and M. J Tsai. Highly scalable hafnium oxide memory with improvements of resistive distribution and read disturb immunity. In *Electron Devices Meeting (IEDM), 2009 IEEE International*, pages 1–4, Dec 2009. doi: 10.1109/IEDM.2009.5424411. 88

[148] G. Servalli. A 45nm generation phase change memory technology. In *Electron Devices Meeting (IEDM), 2009 IEEE International*, pages 1–4, Dec 2009. doi: 10.1109/IEDM.2009.5424409. 88

[149] Xiang Tian and Khaled Benkrid. High-performance quasi-monte carlo financial simulation: Fpga vs. gpp vs. gpu. *ACM Trans. Reconfigurable Technol. Syst.*, 3(4):26:1–26:22, November 2010. ISSN 1936-7406. doi: 10.1145/1862648.1862656. URL http://doi.acm.org/10.1145/1862648.1862656. 89

[150] Tobias G. Noll, Thorsten von Sydow, Bernd Neumann, Jochen Schleifer, Thomas Coenen, and Götz Kappen. Reconfigurable components for application-specific processor architectures. In Marco Platzner, Jürgen Teich, and Norbert Wehn, editors, *Dynamically Reconfigurable Systems*, pages 25–49. Springer Netherlands, 2010. ISBN 978-90-481-3484-7. doi: 10.1007/978-90-481-3485-4_2. URL http://dx.doi.org/10.1007/978-90-481-3485-4_2. 89

[151] B.S. Deepaksubramanyan and A. Nunez. Analysis of subthreshold leakage reduction in cmos digital circuits. In *Circuits and Systems, 2007. MWSCAS 2007. 50th Midwest Symposium on*, pages 1400–1404, Aug 2007. doi: 10.1109/MWSCAS.2007.4488809. 89

[152] Yan Meng, T. Sherwood, and R. Kastner. Leakage power reduction of embedded memories on fpgas through location assignment. In *Design Automation Conference, 2006 43rd ACM/IEEE*, pages 612–617, 2006. doi: 10.1109/DAC.2006.229306. 89

[153] David Lewis, Elias Ahmed, David Cashman, Tim Vanderhoek, Chris Lane, Andy Lee, and Philip Pan. Architectural enhancements in stratix-iii™ and stratix-iv™. In *Proceedings of the ACM/SIGDA International Symposium on Field Programmable Gate Arrays*, FPGA '09, pages 33–42, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-410-2. doi: 10.1145/1508128.1508135. URL http://doi.acm.org/10.1145/1508128.1508135. 89

[154] Fei Li, Yan Lin, Lei He, and Jason Cong. Low-power fpga using pre-defined dual-vdd/dual-vt fabrics. In *Proceedings of the 2004 ACM/SIGDA 12th International Symposium on Field Programmable Gate Arrays*, FPGA '04, pages 42–50, New York, NY, USA, 2004. ACM. ISBN 1-58113-829-6. doi: 10.1145/968280.968288. URL `http://doi.acm.org/10.1145/968280.968288`. 89

[155] MasudH. Chowdhury, Pervez Khaled, and Juliana Gjanci. An innovative power-gating technique for leakage and ground bounce control in system-on-a-chip (soc). volume 30, pages 89–105. SP Birkhäuser Verlag Boston, 2011. doi: 10.1007/s00034-010-9211-7. URL `http://dx.doi.org/10.1007/s00034-010-9211-7`. 90

[156] Jean-Michel Chabloz. *Globally-Ratiochronous, Locally-Synchronous Systems*. KTH Royal Institute of Technology, Stockholm, 2012. ISBN 978-91-7501-258-2. 92, 93

[157] F. Clermidy, N. Jovanovic, S. Onkaraiah, H. Oucheikh, O. Thomas, O. Turkyilmaz, E. Vianello, J.-M. Portal, and M. Bocquet. Resistive memories: Which applications? In *Design, Automation and Test in Europe Conference and Exhibition (DATE), 2014*, pages 1–6, March 2014. doi: 10.7873/DATE.2014.282. 99

[158] E. Vianello, O. Thomas, G. Molas, O. Turkyilmaz, N. Jovanovic, D. Garbin, G. Palma, M. Alayan, C. Nguyen, J. Coignus, B. Giraud, T. Benoist, M. Reyboz, A. Toffoli, C. Charpin, F. Clermidy, , and L. Perniola. Resistive memories for ultra-low-power embedded computing design. In *Electron Devices Meeting (IEDM), 2014 IEEE International*, Dec 2014. 99

[159] O. Turkyilmaz, S. Onkaraiah, M. Reyboz, F. Clermidy, C.A. Hraziia, J. Portal, and M. Bocquet. Rrambased fpga for normally off, instantly on applications. In *Nanoscale Architectures (NANOARCH), 2012 IEEE/ACM International Symposium on*, pages 101–108, July 2012. 99

[160] Ogun Turkyilmaz, Santhosh Onkaraiah, Marina Reyboz, Fabien Clermidy, Hraziia, Costin Anghel, Jean-Michel Portal, and Marc Bocquet. Rrambased fpga for normally off, instantly on applications. *Journal of Parallel and Distributed Computing*, 74(6):2441 – 2451, 2013. ISSN 0743-7315. doi: http://dx.doi.org/10.1016/j.jpdc.2013.08.003. URL `http://www.sciencedirect.com/science/article/pii/S0743731513001391`. Computing with Future Nanotechnology. 99

[161] P. Batude, M. Vinet, B. Previtali, C. Tabone, C. Xu, J. Mazurier, O. Weber, F. Andrieu, L. Tosti, L. Brevard, B. Sklenard, P. Coudrain, S. Bobba, H. Ben Jamaa, P. Gaillardon, A. Pouydebasque, O. Thomas, C. Le Royer, J. Hartmann, L. Sanchez, L. Baud, V. Carron, L. Clavelier, G. De Micheli, S. Deleonibus, O. Faynot, and T. Poiroux. Advances, challenges and opportunities in 3d cmos sequential integration. In *Electron Devices Meeting (IEDM), 2011 IEEE International*, pages 7.3.1–7.3.4, Dec 2011. doi: 10.1109/IEDM.2011.6131506. 102

[162] P. Batude, B. Sklenard, C. Xu, B. Previtali, B. De Salvo, and M. Vinet. Low temperature fdsoi devices, a key enabling technology for 3d sequential integration. In *VLSI Technology, Systems, and Applications (VLSI-TSA), 2013 International Symposium on*, pages 1–4, April 2013. doi: 10.1109/VLSI-TSA.2013.6545629. 103

[163] P. Batude, M. Vinet, C. Xu, B. Previtali, C. Tabone, C. Le Royer, L. Sanchez, L. Baud, L. Brunet, A Toffoli, F. Allain, D. Lafond, F. Aussenac, O. Thomas, T. Poiroux, and O. Faynot. Demonstration of low temperature 3d sequential fdsoi integration down to 50 nm gate length. In *VLSI Technology (VLSIT), 2011 Symposium on*, pages 158–159, June 2011. 103

[164] Perrine Batude, Maud Vinet, Arnaud Pouydebasque, Laurent Clavelier, Cyrille LeRoyer, Claude Tabone, Bernard Previtali, Loic Sanchez, Laurence Baud, Antonio Roman, Véronique Carron, Fabrice Nemouchi, Stéphane Pocas, Corine Comboroure, Vincent Mazzocchi, Helen Grampeix, François Aussenac, and Simon Deleonibus. Enabling 3d monolithic integration. *ECS Transactions*, 16(8):47–54, 2008. doi: 10.1149/1.2982853. URL `http://ecst.ecsdl.org/content/16/8/47.abstract`. 103

[165] Vaughn Betz and J. Rose. Cluster-based logic blocks for fpgas: area-efficiency vs. input sharing and size. In *Custom Integrated Circuits Conference, 1997., Proceedings of the IEEE 1997*, pages 551–554, May 1997. doi: 10.1109/CICC.1997.606687. 106

[166] K. Yano, Y. Sasaki, K. Rikino, and K. Seki. Top-down pass-transistor logic design. *Solid-State Circuits, IEEE Journal of*, 31(6):792–803, Jun 1996. ISSN 0018-9200. doi: 10.1109/4.509865. 106

[167] Xilinx. Xc4000e and xc4000x series field programmable gate arrays (version 1.6), 2009. URL `http://www.xilinx.com/support/documentation/data_sheets/4000.pdf`. 107

[168] ITRS. System driver summary, 2013. URL `http://public.itrs.net/Links/2013ITRS/2013Chapters/2013SysDrivers_Summary.pdf`. 119, 187

[169] ITRS. Process integration, devices, and structures, 2013. URL `http://www.public.itrs.net/Links/2013ITRS/2013Chapters/2013PIDS.pdf`. 121, 187

[170] Laurent Brunet, Perrine Batude, Frank Fournel, Lamine Benaissa, Claire Fenouillet-Beranger, Luca Pasini, Fabien Deprat, Bernard Previtali, Fabienne Ponthenier, Aurélien Seignard, Catherine Euvrard-Colnat, Maurice Rivoire, Pascal Besson, Christian Arvet, Elodie Beche, Olivier Rozeau, Olivier Billoint, Ogun Turkyilmaz, Fabien Clermidy, Thomas Signamarcheix, and Maud Vinet. (invited) direct bonding: A key enabler for 3d monolithic integration. *ECS Transactions*, 64(5):381–390, 2014. doi: 10.1149/06405.0381ecst. URL `http://ecst.ecsdl.org/content/64/5/381.abstract`. 123

[171] O. Turkyilmaz, G. Cibrario, O. Rozeau, P. Batude, and F. Clermidy. 3d fpga using high-density interconnect monolithic integration. In *Design, Automation and Test in Europe Conference and Exhibition (DATE), 2014*, pages 1–4, March 2014. doi: 10.7873/DATE.2014.351. 123

[172] M. Vinet, P. Batude, C. Fenouillet-Beranger, F. Clermidy, L. Brunet, O. Rozeau, JM Hartmann, O. Billoint, G. Cibrario, B. Previtali, C. Tabone, B. Sklenard, O. Turkyilmaz, F Ponthenier, N. Rambal, MP. Samson, F. Deprat, V. Lu, L. Pasini, J-E. Michallet, and O. Faynot. Monolithic 3d integration: a powerful alternative to classical 2d scaling. In *SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S), 2014 IEEE*, Oct 2014. 123

[173] F. Clermidy, O. Turkyilmaz, O. Billoint, and P.E. Gaillardon. 3d technologies for reconfigurable architectures. In *Reconfigurable and Communication-Centric Systems-on-Chip (ReCoSoC), 2014 9th International Symposium on*, pages 1–2, May 2014. doi: 10.1109/ReCoSoC.2014.6861337. 123

[174] G. Cibrario, M. Gary, F. Gays, K. Azizi-Mourier, O. Billoint, O. Turkyilmaz, and O. Rozeau. A high-level design rule library addressing cmos and heterogeneous technologies. In *IC Design Technology (ICICDT), 2014 IEEE International Conference on*, pages 1–4, May 2014. doi: 10.1109/ICICDT.2014.6838599. 123

[175] P. Batude, B. Sklenard, C. Fenouillet-Beranger, B. Previtali, C. Tabone, O. Rozeau, O. Billoint, O. Turkyilmaz, H. Sarhan, S. Thuries, G. Cibrario, L. Brunet, F. Deprat, J.-E. Michallet, F. Clermidy, and M. Vinet. 3d sequential integration opportunities and technology optimization. In *Interconnect Technology Conference / Advanced Metallization Conference (IITC/AMC), 2014 IEEE International*, pages 373–376, May 2014. doi: 10.1109/IITC.2014.6831837. 123

[176] Zhiping Zhang, Chien-Yu Chen, F. Crnogorac, Shu-Lu Chen, P.B. Griffin, R.F. Pease, J.D. Plummer, and S.S. Wong. Low-temperature monolithic three-layer 3-d process for fpga. *Electron Device Letters, IEEE*, 34(8):1044–1046, Aug 2013. ISSN 0741-3106. doi: 10.1109/LED.2013.2266111. 129

[177] O. Turkyilmaz, F. Clermidy, L.G. Amaru, P.-E. Gaillardon, and G. De Micheli. Self-checking ripple-carry adder with ambipolar silicon nanowire fet. In *Circuits and Systems (ISCAS), 2013 IEEE International Symposium on*, pages 2127–2130, May 2013. doi: 10.1109/ISCAS.2013.6572294. 130

[178] D. Sacchetto, Y. Leblebici, and G. De Micheli. Ambipolar gate-controllable sinw fets for configurable logic circuits with improved expressive capability. *Electron Device Letters, IEEE*, 33(2):143–145, Feb 2012. ISSN 0741-3106. doi: 10.1109/LED.2011.2174410. 130

A

Résumé en Français

A.1. Introduction

Jusqu'à présent nous nous sommes appuyés sur le scaling traditionnel afin d'augmenter la densité et la performance. Depuis plus de quatre décennies, conformément à la loi de Moore [19], le nombre de transistors double tous les deux ans. Comme les transistors fonctionnent plus rapidement, les fréquences plus élevées sont atteintes. Cependant, le scaling traditionnel se confronte aux limitations fondamentales avec la fabrication, la performance et la consommation d'énergétique.

La structure des transistors classiques impose des difficultés de fabrication avec des nœuds avancés. La complexité du processus de lithographie augmente avec chaque nœud avancé. Ainsi, des nouvelles technologies de lithographie ou plusieurs étapes de patterning sont nécessaires qui augmentent le coût de production considérablement. Deuxièmement, la partie plus petite du transistor, la largeur du diélectrique de grille, a récemment atteint la taille de plusieurs atomes, qui impose des problèmes suivants: la dépendance supérieur au nombre d'atomes qui changent en raison des variations dans la fabrication et la réduction de la barrière porte qui augmente les courants de fuites.

Outre implications de fabrication, les gains de performance diminuent en raison des limites sur les interconnexions. Même si les transistors diminuent, les circuits ne peuvent pas atteindre des performances élevées sans interconnexions rapides et denses. Fig. A.1 montre que l'écart entre les gains de la performance de grille et le délai d'interconnexion s'élargit avec technologies avancées. Car les fils d'interconnexions ne peuvent pas suivre cette tendance, ils deviennent l'un des principaux goulets d'étranglement

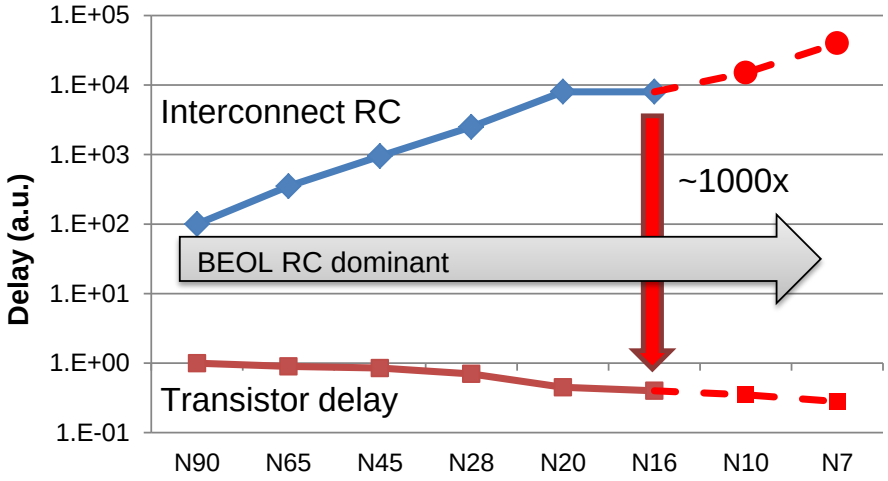


Figure A.1: Le scaling du transistor et du délai pour les nœuds d'interconnexion futures. Adapté de [1].

dans le scaling traditionnel.

La consommation croissante de la puissance inévitable a mené à la fin de la théorie de Dennard [20]. Dennard a déclaré que chaque avancement de noeud devrait augmenter la performance de 40% tout en conservant la même niveau de consommation d'énergie. Les tendances récentes montrent que scaling ne peut remplir soit un meilleur rendement énergétique ou supérieur densité de transistor. Les deux composantes de la consommation d'énergie; la fuite et la dynamique augmentent significativement avec chaque noeud comme représenté sur la Fig. A.2. Précédemment, la consommation de fuite a été supposée être négligeable, mais à la suite de scaling, elle est de plus en plus importante, car elle atteint le niveau de consommation dynamique. La consommation d'énergie dynamique augmente, lorsque plus de transistors s'allument avec des fréquences plus élevées, même si la tension d'alimentation se réduit en raison de scaling. En outre, la gravité d'interconnexions est récemment devenue plus apparente car elles représentent plus de 50% de la consommation totale [21]. Puisque les fils ne se raccourcissent pas comme prévu avec scaling, les transistors plus grand sont nécessaires pour conduire ces fils qui mènent à une plus grande consommation d'énergie. Par conséquent, la consommation d'énergie devient le principal obstacle limitant le scaling traditionnel.

Les limites de scaling traditionnel nous obligent à trouver des prochaines paradigmes de scaling. Grâce à tous les efforts de recherche, nous sommes maintenant à la portée

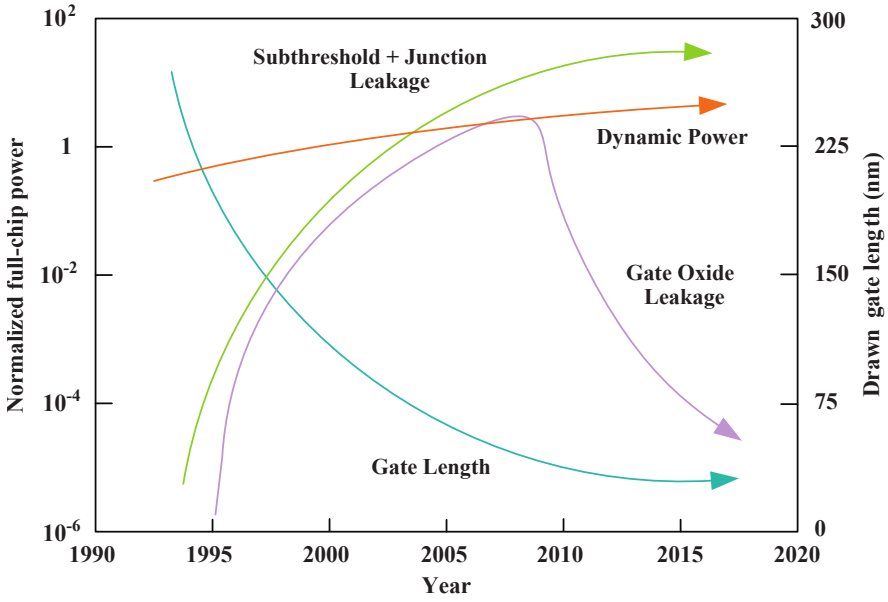


Figure A.2: Tendances de la consommation dynamique et statique de puce basé sur ITRS2006 [2]. La fuite de grille est nettement améliorée avec des matériaux High-K.

de nombreuses possibilités. Dans la section suivante, nous nous concentrons sur les technologies émergentes 3D.

A.1.1. Technologies 3D Emergentes

Technologies 3D reçoivent une attention sans précédent pour les circuits à venir et elles pourraient créer la flexibilité aspirait à répondre aux demandes. L'intégration 3D et les mémoires avancés sont parmi les technologies 3D les plus prometteurs. Une des possibilités d'intégration 3D est accompli avec Through Silicon Vias (TSV). L'intégration TSV offre l'empilement de plusieurs couches et elle aide à réduire la longueur de fils mais elle nécessite une surface significative pour les connexions verticales. Ainsi, elle ne permet qu'un nombre de connexions très limité entre les couches. Afin de tirer le maximum d'avantages de l'intégration 3D, la technologie doit supporter une très forte densité d'interconnexions verticales avec des dimensions de via compatibles avec la taille de l'appareil. Une nouvelle technologie appelée intégration monolithique 3D (3DMI) étend les limites de TSV en atteignant les connexions inter-niveaux 40x plus petites (Table A.1). La caractéristique unique de 3DMI est l'intégration séquentielle des couches actives l'une après l'autre sur la même puce. Ainsi, les couches sont alignées

avec une grande précision qui permet la fabrication des vias l'inter-niveaux similaires aux vias réguliers. En dehors de l'intégration 3D, les mémoires avancées (ex. mémoire résistive RRAM) peuvent introduire une fonctionnalité supplémentaire dans la troisième dimension. Les mémoires avancées apportent des avantages substantiels aux CMOS conventionnels avec l'intégration à Back-End-Of-Line (BEOL) et la possibilité d'exploitation non-volatile. Ces mémoires peuvent être fabriqués entre les couches métalliques qui ne nécessitent aucune surface de silicium. Ces nouvelles technologies 3D peuvent relever les défis des technologies existantes et être le prochain paradigme pour le scaling d'avenir.

A.1.2. Architecture du FPGA

Comme une solution de calcul, les FPGA gagnent l'intérêt croissant en raison de leurs facilités de programmation, de leurs grandes flexibilités, et de leurs coûts non récurrents l'ingénierie (NRE) réduits. Comparé à l'ASIC qui ont les cycles de conception très longues, en considérant la complexité de conception avec des nombre de règles de layout plus élevés, les FPGA apportent une solution efficace pour la gestion de cette demande. Les avantages offerts par des FPGA, d'autre part, sont obtenus par l'utilisation intensive de mémoires de configuration et de ressources de routage. Comparé à l'ASIC, les FPGA utilisent 40x plus de surface avec 4x plus lent performance et 12x plus de consommation d'énergie pour la même fonctionnalité [22]. Consommation en attente qui ne considère que retenir l'état de computation (ce est à dire pas de commutation d'horloge) atteint deux ordres de grandeur plus de l'ASIC [9]. Par conséquent, même si les FPGA offrent une plateforme informatique innovante, les inconvénients empêchent les FPGA d'être utilisés dans les applications mobiles de faible puissance.

Table A.1: Comparaison de la taille de l'interconnect 3D

	Diamètre(μm)	Emplacement(μm)
TSV	2 - 4	4 - 8
3DMI	0.1	0.2
Gain 3DMI vs. TSV	20x - 40x	20x - 40x

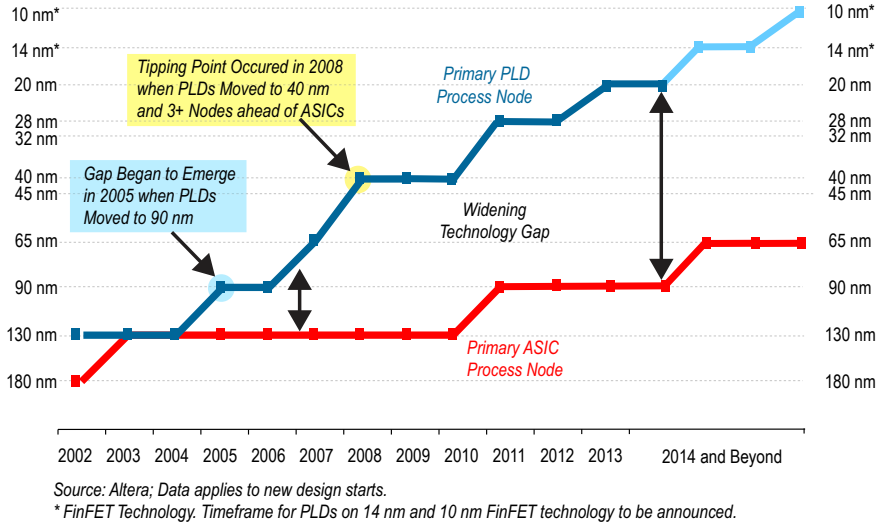


Figure A.3: Logique programmables vs. ASIC pour de nouvelles conceptions dans les processus primaires. [3]

Toujours les FPGA tiennent une de l'avantage le plus efficace sur l'ASIC. Comparé aux FPGA, les ASIC doivent utiliser les noeuds de processus moins coûteux parce que l'effort de se déplacer à un noeud plus avancé impose des coûts élevés en termes de conception physique et de la production. Les FPGA actuelles, par contre, ont déjà atteint 28nm qui seront bientôt sur 20nm et sur plus petits processus. La plupart des nouveaux modèles d'ASIC sont deux ou trois noeuds derrières les FPGA, comme illustré sur la figure. Fig. A.3 [3]. Cette tendance montre qu'il y a des demandes extrêmes sur FPGA et les solutions plus agressives sont nécessaires.

A.1.3. Résumé de la Thèse

Le manuscrit est construit en trois sections principales qui sont expliqués de la manière suivante suivante:

Dans la section A.2. , le cadre de l'exploration des FPGA et de la méthodologie pour l'adoption de nouvelles technologies sont présentés. Le chapitre commence par une synthèse des outils CAO FPGA. Les détails du cadre de l'exploration du FPGA sont discutées. Une méthodologie est présenté pour l'adoption rapide de nouvelles technologies grâce à des modifications de définition de l'architecture FPGA. Cette plateforme d'évaluation établit la base pour le reste de la thèse.

Dans le section A.3., les conceptions FPGA plus efficaces sont démontrées en utilisant des mémoires avancées. Premièrement, la motivation pour les FPGA avec des mémoires avancées s'explique. Les nouveaux domaines d'applications sont proposées comme l'application de sauvegarde de configuration et de contexte. La modification nécessaire pour NVFPGA avec des circuits utilisant des technologies OxRAM et CBRAM est expliqué. Un nouveau système de calcul avec la propriété basse consommation est présenté. Les implications de mémoire avancée au niveau du système sont déterminées pour chaque technologie dans la dernière section.

Dans le section A.4., la conception de 3D-FPGA avec l'intégration monolithique est proposée. Logic-sur-Mémoire approche est présentée et les blocs conçus avec cette approche sont évalués en utilisant de plateforme VPR. La concept de multi-niveaux FPGA est discutée et les FPGA à partir de deux jusqu'à quatre couches sont évalués. Enfin, les résultats du FPGA multi-niveaux sont comparés aux attentes de scaling traditionnel.

A.2. Evaluation du FPGA avec les Technologies Emergentes

Par rapport à l'ASIC, les FPGA bénéficient des technologies plus avancées. Losqu'un réel FPGA industriel est conçu, les concepteurs de circuits passeront mois pour ajuster soigneusement les aspects de la conception de circuits pour cibler la nouvelle technologie. Surtout, lorsque les technologies émergentes sont ciblés, il est impossible d'explorer toutes les possibilités de la conception, d'optimiser avec ces technologies et de mesurer les indicateurs de performance des circuits FPGA fabriqués. Par conséquent, un environnement d'évaluation du FPGA rapide, fiable et flexible est indispensable .

A.2.1. Cadre d'Evaluation du FPGA Expérimentale

Le processus expérimental utilisé dans le cadre basé de VPR est illustré à la Fig. A.4. Tous les outils inclus sont open-source qui permettent l'évaluation sur des différentes architectures.

Le toolflow prend un circuit de référence et une définition de l'architecture du FPGA pour implémenter la fonctionnalité et évaluer sur l'FPGA. ODIN est utilisé comme outil de synthèse frontal pour convertir la description de haut niveau (ex. VHDL) à une

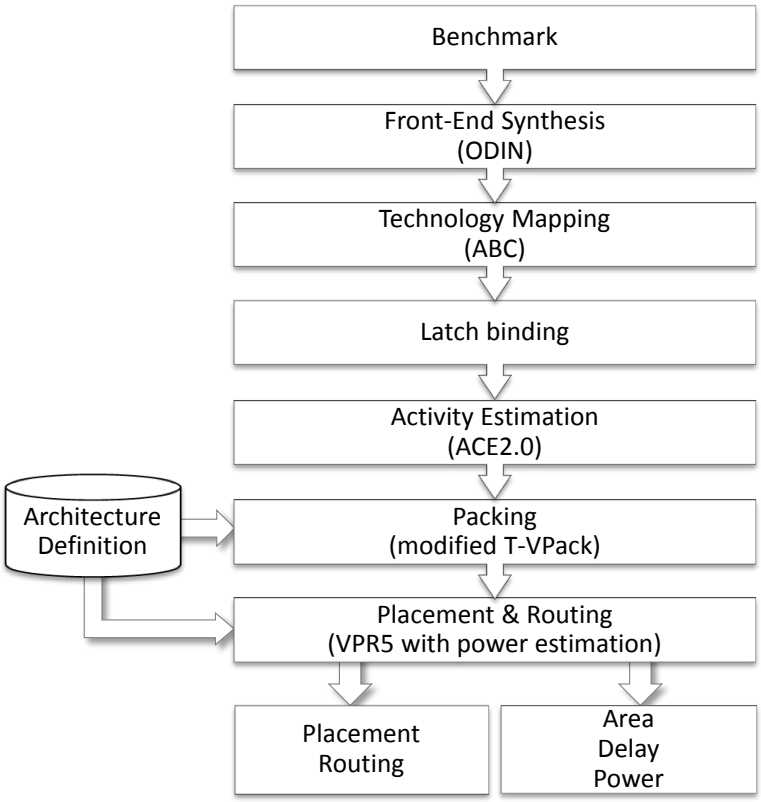


Figure A.4: VPR5 outil avec estimation de la puissance.

netlist BLIF. L'outil ABC prend la netlist BLIF et établit tous les circuits logiques en LUT. Puisque l'ABC ne se connecte pas les signaux d'horloge aux bascules, un script de liaison de verrou est exécuté pour permettre une évaluation des circuits séquentiels. ACE2.0 [135] effectue l'estimation de l'activité qui est nécessaire pour l'évaluation de consommation de puissance. Une version modifiée de T-VPACK [136] permet de grouper des blocs qui ont de même niveau d'activité. VPR5 avec Power Extension [137] accomplit le placement et le routage du circuit. Il aussi extrait les estimations de la surface, du délai, et de la puissance.

Le reste de cette section discute la méthodologie pour l'évaluation des technologies émergentes, y compris la plateforme d'évaluation FPGA, le développement de la définition de l'architecture, et la modélisation de la surface de cellule mémoire.

A.2.2. Méthodologie pour l'Evaluation des Technologies Emergentes

Même si, la structure du FPGA est une unité fonctionnelle simple, plusieurs choix de conception peuvent être imposées dans le développement architectural. Surtout, lorsque les nouvelles technologies sont employées, une modélisation précise de ces paramètres architecturaux et technologiques est crucial. Il est prévu que plusieurs essais doivent être exécutées de manière itérative pour optimiser l'architecture. Par conséquent, une plateforme d'exploration est nécessaire qui est rapide et flexible.

A.2.2.1. Plateforme d'Evaluation du FPGA

Pour adopter des nouvelles technologies, l'approche de conception classique n'est pas suffisant. En général, une bibliothèque de cellules standards est construite et ensuite vérifiée pour la conception de la logique. Ces cellules offrent divers capacités de conduite et d'autres propriétés telles que faible puissance ou haute performance. Les cellules sont ensuite utilisés dans un outil de placement et de routage automatisé. Cette approche est optimisée pour les technologies de pointe où les outils et les bibliothèques sont assez matures pour les conceptions fiables. Cependant, lors de la conception avec des technologies émergentes, ce niveau de maturité ne peut pas être atteint pour l'exploration de la conception.

Afin d'accomplir une conception FPGA, une bibliothèque de cellules partielle est suffisante car des blocs élémentaire du FPGA ne sont pas complexes et la conception du FPGA totale peut être accomplie avec quelques cellules. Toutefois, selon la technologie ciblée, nouveaux outils pourraient être inclus dans le cadre de CAO automatisé, ex. un outil de placement et de routage en 3D. Dans cette thèse, une méthodologie d'évaluation du FPGA pour les technologies émergentes est proposée (Fig. A.5). Selon la maturité de la technologie, les cellules sont conçues et caractérisés soit au niveau de l'appareil ou bloc. Avec cette caractérisation, la définition de l'architecture du FPGA est créée et utilisée comme une entrée à la CAO basée de VPR afin d'évaluer la performance, la surface, et la puissance.

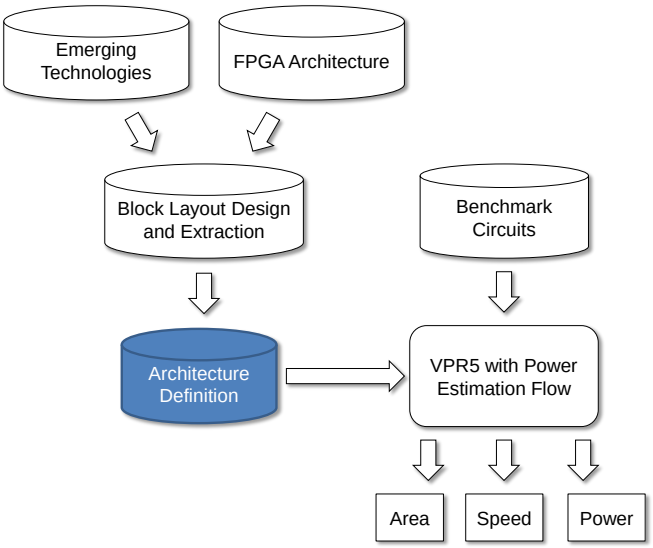


Figure A.5: Plateforme de l'exploration du FPGA pour les technologies émergentes.

A.2.2.2. Développement de la Définition Architecture pour des Technologies Emergentes

En utilisant du cadre d'évaluation FPGA, les impacts de LUT, la taille de cluster, ou SB topologie peuvent être observés rapidement. Dans ces cas, l'architecture du FPGA est évaluée sur la base d'hypothèses paramétrées où les parasites de circuit ne peuvent pas être examinés efficacement. Ainsi, la méthodologie de la Fig. A.6 est proposée pour le développement de la définition de l'architecture avec les technologies émergentes.

Dans cette méthodologie, la définition de l'architecture est construite en utilisant des blocs FPGA. Les FPGA sont conçus à base de tuiles et une tuile est construite avec un bloc logique (LB), deux boîtes de connexion (CB), et une boîte de commutation (SB). Ces blocs de base sont établis avec des mémoires de configuration, les multiplexeurs, et des buffers. Par conséquent, une fois que ces blocs élémentaires sont conçus en layout et inclus dans la bibliothèque de conception, tous les blocs du FPGA liés peuvent être conçues rapidement. Le processus de conception est, donc, considérablement simplifié car quelques blocs doivent être conçus sur la disposition plutôt que de l'ensemble du FPGA.

Après le layout se termine, les paramètres concernant la conception et la technologie sont extraites pour le développement de définition de l'architecture. Premièrement, les

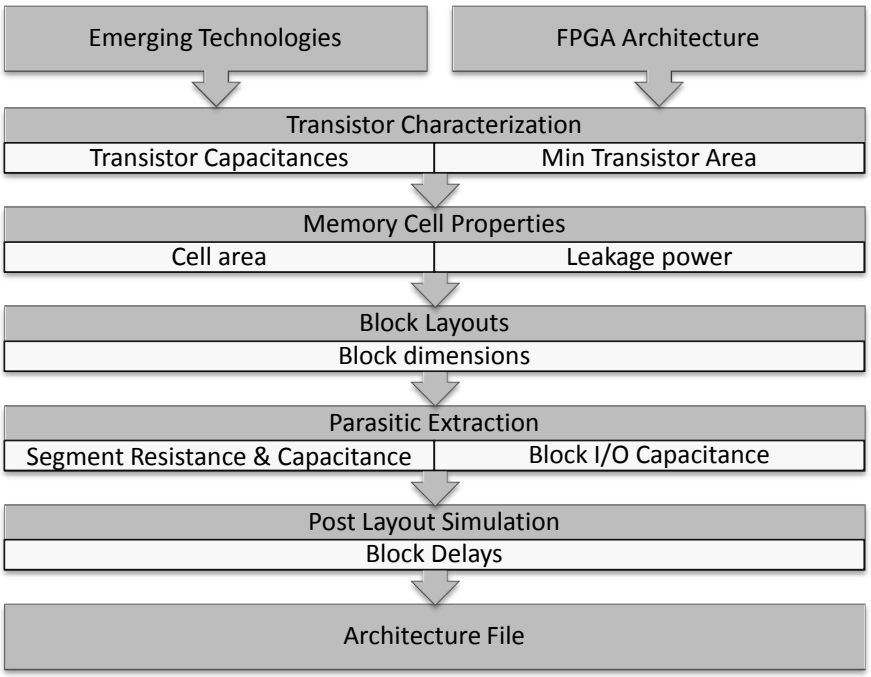


Figure A.6: Méthodologie de Création de la Définition de l'Architecture avec les Technologies Emergentes

paramètres à transistor sont extraits. La capacitance parasite de transistor est utilisée pour le calcul de la puissance dynamique. Les dimensions de transistor sont utilisés pour l'estimation de surface du circuit FPGA. Avec ces paramètres, les détails technologiques de transistors sont introduits dans la définition de l'architecture. Les propriétés de mémoire se réfèrent aux mémoires de configuration dans le FPGA. La surface cellulaire permet d'estimer la surface totale des mémoire de configuration du FPGA. La puissance de fuite se réfère à la consommation de puissance de fuite d'une cellule de mémoire et elle est incluse dans le calcul de fuite total du circuit FPGA. Dans l'étape suivante, les blocs du FPGA sont caractérisés. Les blocs sont conçus avec les règles de conception full-custom. En utilisant les dimensions de layout des blocs, les paramètres concernant la surface sont obtenus. Les blocs sont ensuite simulés en considerant de l'extraction de parasites. La capacitance I/O, la résistance et le délai des blocs et des segments sont mesurés. Avec cette méthodologie, les architectures du FPGA peuvent être évalués en suivant l'impact technologique et en adoptant rapidement des technologies émergentes.

```
<device>
  <trans_sram trans_sram_bit= " 6" />
</device>
```

Figure A.7: paramétrisation de la surface de cellule de mémoire.

A.2.2.3. Modélisation de la Surface de Cellules Mémoire

Normalement, VPR considère une cellule SRAM 6T avec des transistors de taille minimum pour la mémoire de configuration. Ce nombre de transistors de mémoire est une partie du modèle de surface du FPGA. Apart de la surface totale, la surface dédié aux mémoires de configuration a une forte influence sur la longueur des fils de routage. Par conséquent, elle n'affecte pas seulement la surface totale, mais aussi un routage complexité qui est directement liée à la performance et la consommation d'énergie en raison d'implications de la résistance et de la capacité. Dans différentes technologies et topologies, cependant, SRAM peut consommer une surface inférieure ou supérieure à l'estimation 6T. Ainsi, l'outil de VPR et les fichiers d'architecture sont modifiées pour pouvoir imposer la surface de la cellule de mémoire. La variable $trans\_SRAM$ dans le code source est ajouté dans le fichier de l'architecture comme représenté sur la Fig. A.7.

A.3. FPGA Non-Volatile avec Mémoires Resistives

Le principal avantage des FPGA est la reconfiguration, pour cela un nombre élevé des cellules de mémoire de configuration à accès aléatoire (CRAM) est nécessaire. Près de la moitié (43%) [8] de la surface d'un circuit est dédiée aux CRAM, parmi les différentes mémoires existantes, les mémoires SRAM restent la solution préféré. L'intégration de ces mémoires augmente non seulement la surface totale du FPGA, mais aussi la longueur des fils de routage qui détériore les performances et la consommation énergétique. En plus de la surface, ces mémoires impactent les courants de fuites de manière importante. La technique de power gating peut être implémentée pour éliminer les courants de fuites mais les mémoires de configuration perdent les informations de configuration car les mémoires sont volatiles. Les solutions à base de mémoire flash peuvent être intégrées mais les difficultés d'intégration CMOS (le coût et les limites de scaling) empêchent l'intégration de flash pour une utilisation à grande

échelle. Par conséquent, les mémoires compacts et non-volatiles sont prometteuses en termes de surface, de haute performance et de faible consommation pour les FPGA.

Les technologies de mémoire résistive (RRAM) peuvent atténuer les surcoûts de surface, de haute performance et de faible consommation. Dans cette thèse, les mémoires oxyde (OxRAM) et les mémoires pont conducteur (CBRAM) sont les technologies ciblées. Ces mémoires offrent des possibilités prometteuses avec la fonctionnalité non volatile intrinsèque et la surface compacte avec la compatibilité CMOS Back-End-of-Line (BEOL).

L'intégration de la mémoire résistive dans les FPGA apporte des opportunités pour de nouveaux domaines d'application. Lorsqu'elle remplace les mémoires SRAM, les informations de configuration peuvent être conservées localement dans le FPGA pendant la période de coupure de l'alimentation. Ainsi, l'application de sauvegarde de configuration peut être réalisée dans le FPGA conçu. De plus, lorsque des FFs sont remplacées, les résultats de calculs (contexte) peuvent être conservées pendant la période de coupure d'alimentation. Par conséquent, les applications de sauvegarde de contexte peuvent être fournies.

Des applications embarquées ayant des spécifications variées apparaissent dans le cadre de l'Internet des Objets. Certaines de ces applications peuvent être classées comme «Normalement-off, Instantanément-on". Ces applications nécessitent une courte phase de calculs très intensifs entre les longues périodes d'inactivité. Les FPGA ont une surcharge importante de consommation énergétique ce qui les empêchent d'être employés dans des applications fonctionnant avec des batteries. Le FPGA proposé prend avantage de la non-volatilité des mémoires, ainsi il peut stocker l'information locale et la technique de power gating peut être appliquée afin de couper l'alimentation pendant les périodes d'inactivité. Le FPGA peut ensuite être mis en marche rapidement pour les calculs avant de s'éteindre. Dans ce cas, les courants de fuites peuvent être éliminées pendant les périodes d'inactivité et la consommation globale d'énergie peut être réduite.

A.3.1. Evaluation du FPGA Non-Volatile

Dans cette section, les FPGA Non-Volatile (NVFPGAs) avec la propriété de sauvegarde de la configuration sont conçus en utilisant des mémoires à base d'OxRAM et de CBRAM. étant donné que tous les nœuds de configuration sont remplacés par leurs

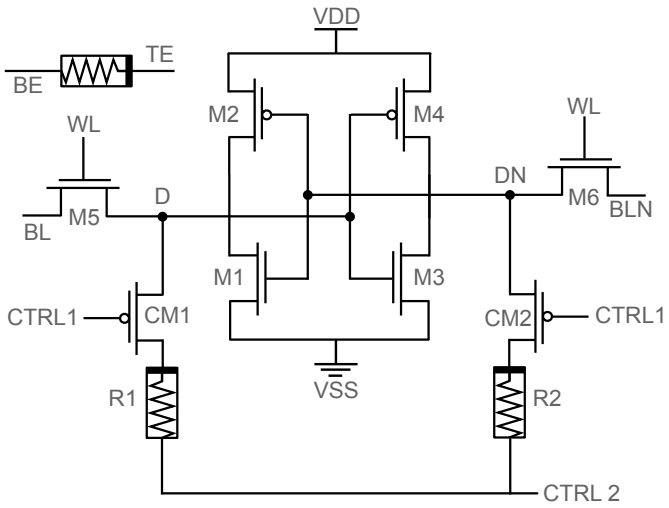


Figure A.8: NVSRAM schématique de l'architecture 8T2R [17].

équivalents non-volatiles, la mémoire externe n'est plus nécessaire pour enregistrer le bitstream qui peut être maintenu localement dans le FPGA. Ainsi, les avantages suivants sont obtenus: 1) La période de démarrage peut être considérablement raccourcie et, donc, le FPGA peut être activé immédiatement. 2) Il n'y a pas de consommation d'énergie en raison du transfert de bitstream. 3) La sécurité de bitstream est accordée puisqu'il n'y a pas de mémoire externe. Dans le FPGA à base d'OxRAM tous les nœuds de mémoire de configuration sont remplacés par des cellules de NVSRAM qui sont construits par une combinaison hybride de cellules SRAM et RRAM. La cellule NVSRAM (in Fig. A.8) utilisée dans ce travail est conçue par l'ISEP [17]. L'outil VPR de la section A.2.1. est utilisé pour l'évaluation du FPGA. Afin d'examiner les effets de la technologie 22nm LETI-FDSOI, la définition de l'architecture dans le VPR est modifiée, puis les modifications dues à la cellule NVSRAM sont inclues.

Les résultats de l'outil VPR pour la SRAM et le FPGA à base de NVSRAM sont affichés sur les Fig. A.9, A.10 et A.11. Le délai, la surface et la consommation d'énergie sont augmentés en moyenne de 7%, 18% et 2% respectivement dans l'implémentation basée sur la NVSRAM en raison de sa surface.

En utilisant de la technologie CBRAM, une cellule de mémoire appelée l'élément non volatile (NVE) [141], (sur Fig. A.12a), est intégrée dans le FPGA. Pour l'évaluation au niveau du FPGA, la définition de l'architecture dans l'outil VPR est modifiée avec les paramètres technologiques 130 nm et les propriétés de cellules NVE. L'évaluation

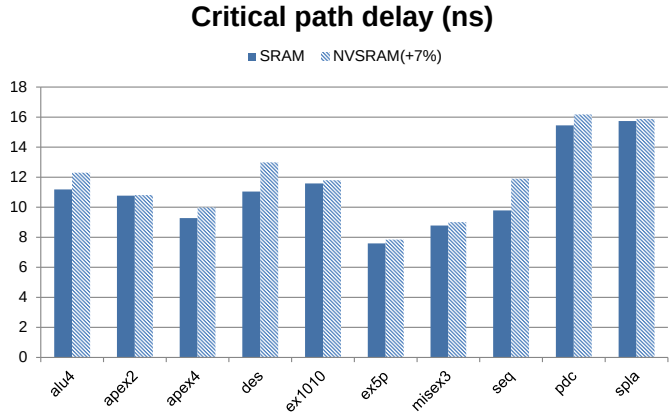


Figure A.9: Le délai de chemin critique des circuits de référence FPGA pour l'intégration FPGA avec SRAM et NVSRAM. Les résultats montrent une augmentation de délai en moyenne de 7%.

sur FPGA est effectuée en deux phases: d'abord les effets de NVLUT sont observés en remplaçant les LUT à base de SRAM avec leurs équivalents à base de CBRAM Non-Volatile Look-Up Table (NVLUT) et en gardant le reste des cellules de mémoires SRAM . Après l'évaluation de NVLUT, les effets de l'intégration NVE sur l'ensemble du FPGA sont comparés aux FPGA à base de SRAM.

Les Fig. A.13, A.14, A.15 comparent le FPGA basé de CBRAM avec le FPGA basé de SRAM. Profitant de l'empreinte compacte de la NVE par rapport à la cellule de SRAM, la superficie totale de la base du FPGA-CBRAM est réduite en moyenne de 5% avec l'intégration NVLUT tandis que le NVFPGA réduit la surface de 33%. En raison de sa plus petite surface, les fils de routage rétrécissent diminuant ainsi le délai du chemin critique et la consommation énergétique. L'intégration NVLUT réduit le délai de 24% et le NVFPGA atteint la réduction de 34% . Enfin, la consommation d'énergie est réduite de 18% avec l'intégration NVLUT et de 23% pour le NVFPGA.

Une Flip-Flop non-volatile (NVFF) basée sur une RRAM est nécessaire après l'intégration de la NVSRAM afin de rétablir la configuration et le contexte dans une application ON / OFF. L'architecture de la cellule NVFF, représentée sur la Fig. A.16 est proposée en collaboration entre le LETI et l'IM2NP [18].

Les modifications, similaires à l'intégration de NVSRAM, sont appliquées à la définition de l'architecture en considérant les cellules NVFF et la technologie 22nm LETI-FDSOI. Les résultats de l'évaluation de VPR5 sont visibles sur les Fig. A.17, A.18 et

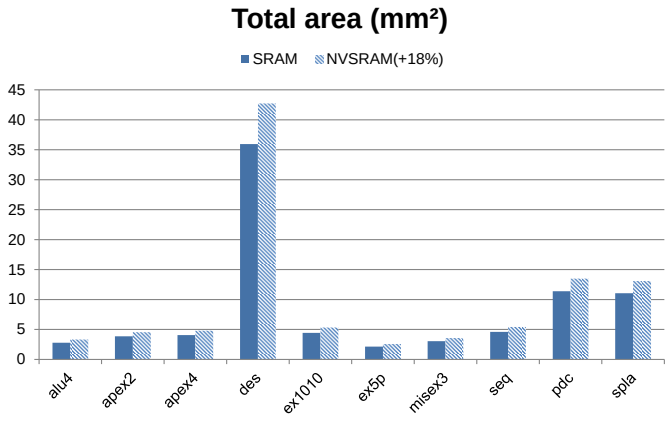


Figure A.10: La surface totale des circuits de référence pour l'intégration FPGA avec SRAM et NVSRAM. Les résultats montrent une augmentation de surface en moyenne de 18%.

A.19 pour les mesures de chemin critique, de surface et de consommation énergétique. L'intégration NVSRAM présente des augmentations de 6%, 17%, et de 1,5% de délai, de surface et de consommation énergétique respectivement en raison de la surface de la cellule NVSRAM. L'intégration NVFF mène à des augmentations de 3%, 1% et 1% de délai, de surface et de consommation énergétique respectivement en raison de la surface de la NVFF.

A.3.2. Optimisation des Niveaux de Résistance pour le NVFPGA

Cette section étudie l'impact des états de résistance de NVE sur le délai et la consommation. Les dispositifs résistifs haut et bas de NVE sont programmés dans un manière complémentaires. Donc, à tout moment l'un d'eux est programmé avec une haute résistance (R_{OFF}) et l'autre avec une faible (R_{ON}) résistance. Le R_{ON} de la NVE détermine combien de temps il faut pour charger ou décharger le noeud de configuration des LUTs. Ainsi, la faible résistance R_{ON}, améliore les performances. .

Dans le structure NVE pendant le fonctionnement normal, il y a un chemin de courant direct entre les sources d'alimentation. Compte tenu des états complémentaires des résistances, une valeur de résistance de $R_{ON} + R_{OFF}$ se situe entre les sources d'alimentation. Le courant de fuite est, par conséquent;

$$I_{leakage} = \frac{V_{dd} - V_{ss}}{R_{ON} + R_{OFF}} \approx \frac{V_{dd}}{R_{OFF}} \tag{A.1}$$

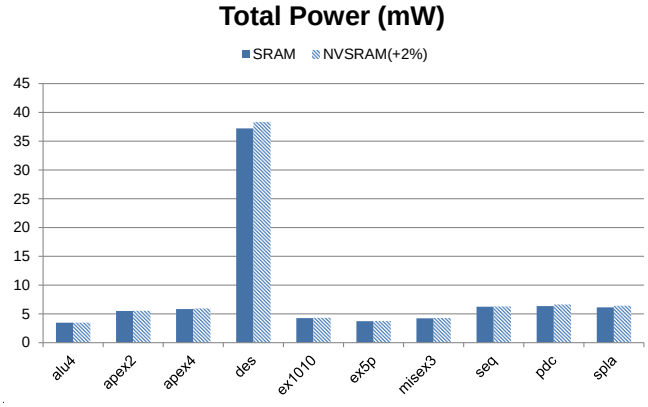


Figure A.11: La consommation totale d'énergie des circuits de référence pour l'intégration FPGA avec SRAM et NVSRAM. Les résultats montrent une augmentation de consommation en moyenne de 2%.

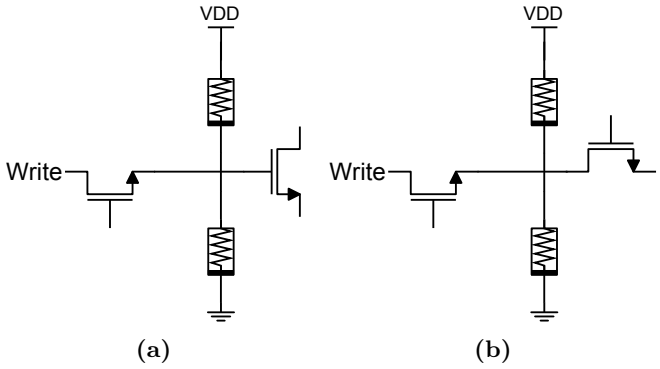


Figure A.12: (a) la connexion de noeud de configuration NVE dans les commutateurs FPGA (b) la connexion de noeud de configuration NVE dans LUT FPGA

L'impact de R_{ON} sur le délai est examiné sur la base du délai LB. En utilisant la plateforme d'évaluation VPR, plusieurs fichiers d'architecture sont créés avec des valeurs de résistance de mémoire variées et la moyenne délai du chemin critique pour chaque valeur de résistance est rapporté. Fig. A.20 montre l'impact de la résistance sur le délai du chemin critique. Pour une performance accrue, la tendance des résultats confirme l'espoir que la résistance devrait être abaissée. Il peut également être remarqué que, même à très haute résistance de 100k *Omega*, la performance de base du FPGA-NVE est encore 23% de mieux que l'équivalent de SRAM. Ceci peut être expliqué avec l'avantage de la surface gagnée de l'implémentation à base de NVE qui est encore plus élevé que l'effet de diminution de haute résistance.

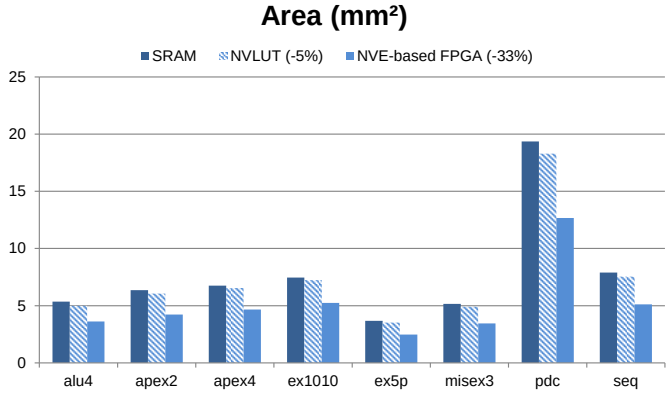


Figure A.13: La surface totale des circuits de référence pour FPGA SRAM et l'intégration CBRAM. La surface réduite de 5% avec NVLUT et de 33% avec NVFPGA.

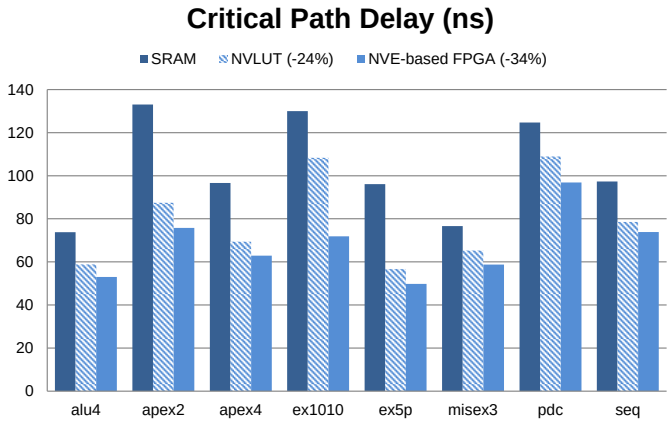


Figure A.14: Le délai du chemin critique des circuits de référence pour les FPGA SRAM et l'intégration CBRAM. Le délai réduit de 24% avec NVLUT et de 34% avec NVFPGA.

Plusieurs valeurs de R_{OFF} sont analysées afin d'observer l'impact sur la consommation d'énergie du FPGA. Les courants de fuite sont calculés pour les valeur de R_{OFF} entre 1MΩ et 10GΩ. Plusieurs fichier d'architecture sont créés avec ces valeurs pour l'évaluation VPR et la consommation totale, y compris la dynamique et la fuite, sont extraites et elles sont normalisées avec le cas 10GΩ. Fig. A.21 représente les résultats obtenus. Comme prévu lorsque la R_{OFF} est réduite, la consommation de fuite augmente et pour les valeurs de R_{OFF} inférieures , la consommation fuite devient le facteur dominant dans la consommation totale. A partir de 10GΩ, à 1GΩ la consommation d'énergie totale reste sans changer notablement, cependant, à 1MΩ, la consommation d'énergie augmente de 60x.

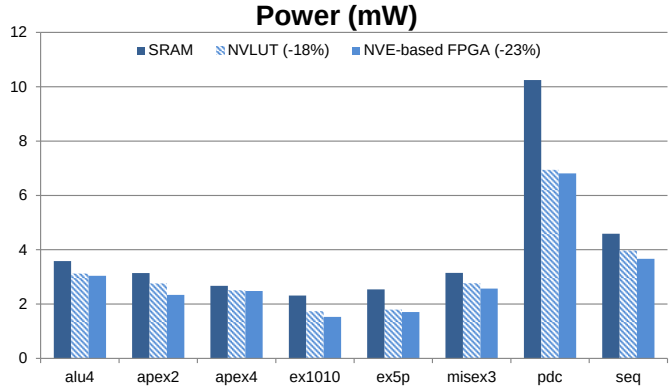


Figure A.15: La consommation des circuits de référence pour les FPGA SRAM et l'intégration CBRAM. Les délais du chemin critique réduits sont de 18% avec NVLUT et 23% avec NVFPGA

A.3.3. Normalement-OFF Instantanément-ON FPGA

Dans les applications normalement-off instantanément-on, l'unité de calcul est éteinte quand elle n'est pas utilisée et elle est activée très rapidement pour continuer les calculs. En raison des mémoires volatiles et les registres, ces applications ne peuvent pas être traitées dans les FPGA traditionnels. Les NVFPGAs proposées dans ce travail, d'autre part, peuvent satisfaire à l'exigence de ces applications. De plus, avec l'intégration des mémoires non volatiles, la consommation de fuite peut être conservée par l'introduction d'un état zéro-fuite. Dans cette section NVFPGAs conçus antérieurement sont analysés dans le régime Normalement-off instantanément-on.

A.3.3.1. Relation entre Consommation et Cycle d'Utilisation

Dans une application où le FPGA non-volatile est mis en mode veille, la puissance de fuite, qui est normalement dissipée, sera conservée. Cependant, il faut prendre en compte de l'impact de l'intégration de RRAM sur la consommation. Dans la Fig. A.22, P_1 et P_2 réfèrent aux consommation d'énergie prévue des architectures à base de SRAM et de RRAM, P_L est la consommation d'énergie fuite de l'architecture à base de SRAM, t_0 and t_1 sont les durées d'ON et d'attente respectivement. Dans une activité où l'énergie de fuite consommée pendant t_1 dépasse la surcharge de l'énérgie en raison de la surcharge de puissance ($P_{OH} = P_2 - P_1$) dépensée pendant t_0, l'énergie de fuite pour un cycle d'utilisation plus petit sera conservée.

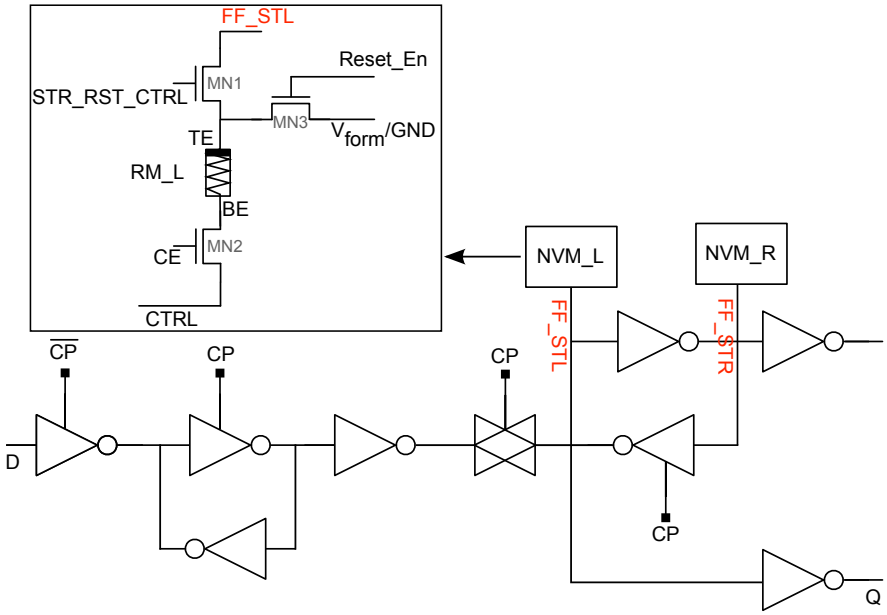


Figure A.16: L'architecture NVFF avec le bloc non-volatile basée sur RRAM, les RRAMS conservent l'état d'esclave dans NVM_L et NVM_R [18].

Initialement, un ratio entre t_1 et t_0 peut être définie lorsque l'énergie de fuite pendant t_1 est égale à l'énergie dépensée au cours de t_0 en raison de la surcharge de puissance:

$$\int_{t}^{t+t_0} P_2 - P_1 \, \mathrm{d}t = \int_{t+t_0}^{t+t_0+t_1} P_L \, \mathrm{d}t \tag{A.2}$$

$$\frac{t_1}{t_0} = \frac{P_2 - P_1}{P_L} \tag{A.3}$$

Le ratio donné par l'équation Eqn. A.3 définit le minimum de t_1 en termes de t_0 lorsque les frais généraux de l'énergie sont en équilibre avec l'énergie de fuite gagnée. Il est possible de relier ce ratio au cycle d'utilisation, qui est défini comme le rapport de la durée ON à la période totale. Plus précisément, dans une application on / off, le temps de rentabilité (BET) indique le cycle d'utilisation lorsque la surcharge de l'énergie est égale à l'énergie de fuite. Dans cette application, BET est calculé sur la base de l'équation suivante:

$$BET = \frac{t_0}{t_0 + t_1} = \frac{1}{1 + \left(\frac{P_2 - P_1}{P_L}\right)} \tag{A.4}$$

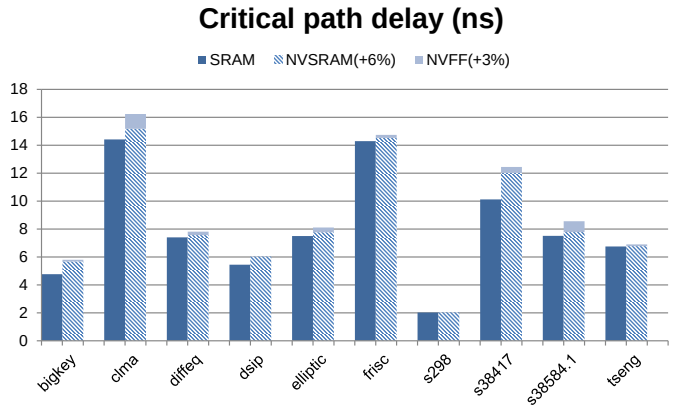


Figure A.17: Le délai du chemin critique des circuits de référence pour FPGA avec SRAM, NVSRAM et NVFF. Les résultats montrent une augmentation du délai en moyenne de 6% et 3% dans les implémentations de NVSRAM et NVFF.

Eqn. A.4 définit le temps de rentabilité (BET). Lorsque le cycle d'utilisation réelle de l'application est inférieure au BET, il est possible d'économiser de l'énergie de fuite et de réduire la consommation d'énergie totale.

A.3.3.2. Evaluation du FPGA Normalement-OFF Instantanément-ON

Le cycle d'utilisation de l'application a une influence direct sur le niveaux de puissance conservée. Fig. A.23 montre que pour la on/off application, si le cycle d'utilisation est beaucoup plus faible que le BET, le gain de puissance augmente rapidement. Le passage à zéro des courbes donne les BET des circuits. A partir de 42%, si le cycle d'utilisation est réduit à 10% et 5%, les gains de puissance de 7% et 16% peuvent être obtenues respectivement. Pour une application particulière où le circuit est actif pour 1% du temps, le gain de puissance atteint 50% en moyenne.

Il a été démontré dans la section A.3. qu'en raison de réduction de la congestion de routage avec l'intégration NVE, la consommation totale d'énergie a été réduite de 23%. Comme il est déjà un gain de puissance positif à pleine activité (cycle d'utilisation de 100%), le BET, de définition, n'existe plus. En fonction de l'activité du circuit, le gain d'énergie peut être encore accrue par la conservation de la consommation de fuite qui est perdu pendant les périodes d'inactivités. Les valeurs de gain de puissance sont calculés en fonction de l'activité du circuit. Fig.A.24 montre que pour les applications ayant un cycle d'utilisation 50%, la consommation de puissance totale peut être réduite

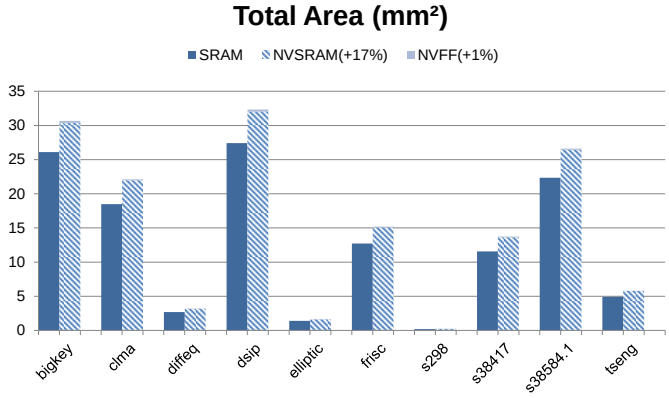


Figure A.18: La surface totale des circuits de référence pour FPGA avec SRAM, NVSRAM et NVFF. Les résultats montrent une augmentation de délai en moyenne de 17% et 1% dans les implémentations de NVSRAM et NVFF.

de 44%. Gain de puissance atteint plus de 97% lorsque l'application est actif pour 1% du temps.

Après les BETs sont obtenus, la consommation d'énergie pendant plusieurs cycles d'utilisation sont calculées pour observer leurs effets sur l'économie d'énergie. Fig. A.25 montre que la puissance gagnée augmente rapidement pour les cycles de fonctionnement plus faibles que le BET parce que plus de la puissance de fuite est conservée. A partir de 38% la réduction des cycles d'utilisation de 10% et 5% garantie de 6% et un gain de puissance de 13%, respectivement. Pour les applications restants inactif pour la plupart de la période d'exploitation, l'activation du FPGA pour 1% du temps, fournit plus de 40% gain de puissance.

A.4. 3D-FPGA avec l'Integration Monolithique

Le principal inconvénient du FPGA est le taux élevé d'utilisation des mémoires de configuration et étant donné que ces mémoires sont distribuées, il faut avoir beaucoup d'interconnexion. Avec l'intégration 3D monolithique (3DMI), l'impact de ces mémoires peut être réduite. Dans cette section, plusieurs blocs de 3D-FPGA sont conçus en utilisant la technologie 3DMI. Profitant d'interconnexions à haute densité de 3DMI, une technique logique-sur-mémoire est adoptée pour la conception de cellules 3D. Les cellules 2D FPGA sont également inclues dans la bibliothèque de conception pour la comparaison et l'exploration de conception. à partir de 2 couches jusqu'à 4 couches,

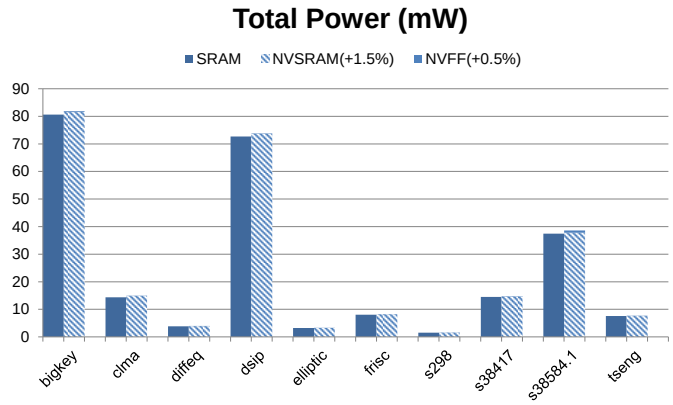


Figure A.19: La puissance totale des circuits de référence pour FPGA avec SRAM, NVSRAM et NVFF. Les résultats montrent une augmentation de délai en moyenne de 1,5% et 0,5% dans les implémentations de NVSRAM et NVFF.

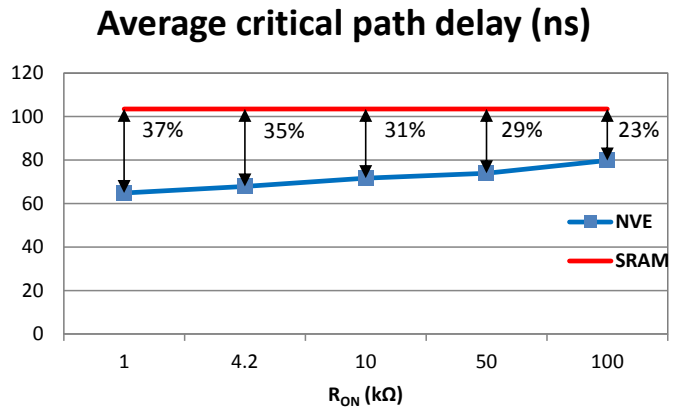


Figure A.20: R_{ON} impact sur le chemin critique. Le gain est réduit avec la valeur croissante de la résistance.

plusieurs circuits de 3D-FPGA sont analysés et évalués avec les partitionnements variés. Basé sur les résultats, 3DMI démontre être très efficace en termes de surface, de performance et de consommation énergétique pour les FPGA. En outre, les gains de couches empilées avec 3DMI surpassent les gains de scaling traditionnel attendus, ce qui prouve que 3DMI peut être une alternative pour le scaling de l'avenir.

A.4.1. 3D-FPGA à l'Approche Logic-sur-Mémoire

Comme décrit dans [8] près de la moitié de la surface d'un FPGA est occupée par les mémoires. En tenant compte de cette caractéristique, nous proposons le partition-

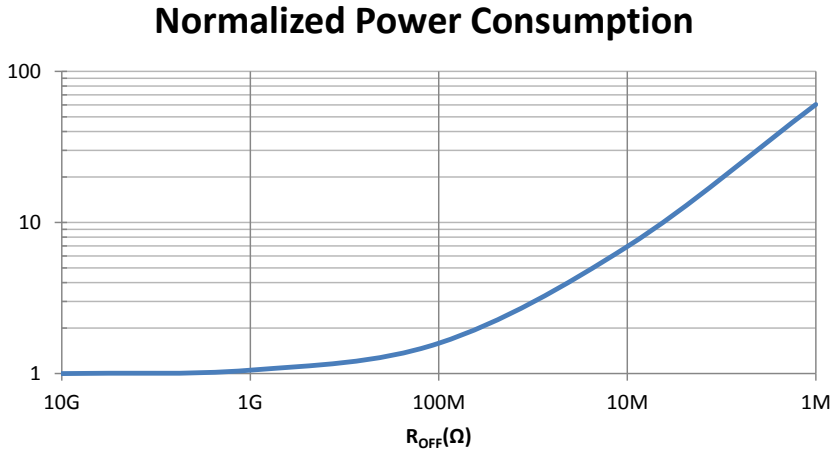


Figure A.21: R_{OFF} impact sur FPGA consommation totale d'énergie. La consommation augmente lorsque la valeur de résistance réduit.

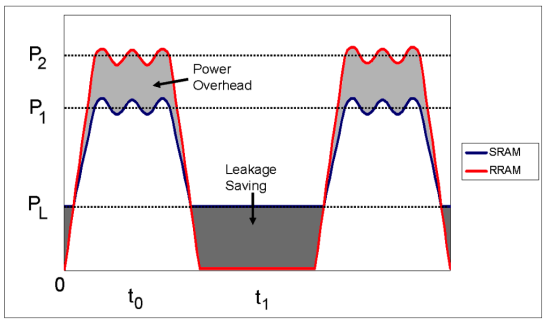


Figure A.22: Vue conceptuel du potentiel gain et consommation de l'énérgie des implémentations basée de SRAM et de RRAM. En mode d'attente, le circuit avec SRAM consomme de l'énergie de fuite, alors qu'il est possible de réduire la consommation dans le même mode à zéro avec RRAM intégration.

nement 3DMI de la manière suivante : La couche basse contient des cellules SRAM tandis que la couche haute contient les ressources de calculs et de routage. Pour conserver une bonne performance globale, les cellules SRAM doivent être entièrement intégrées sur la couche basse qui mène à la sélection de deux couches métalliques intermédiaires. Une deuxième cible concerne la conception: un équilibre de surface entre les deux couches doit être atteint tout en conservant les capacités de modularité et d'extensibilité. Pour remplir la première condition, la co-conception de la couche haute et la couche basse doit être effectuée à gros grain, alors que la deuxième contrainte nous conduit à maintenir le partitionnement du FPGA classique, c'est-à-dire une tuile

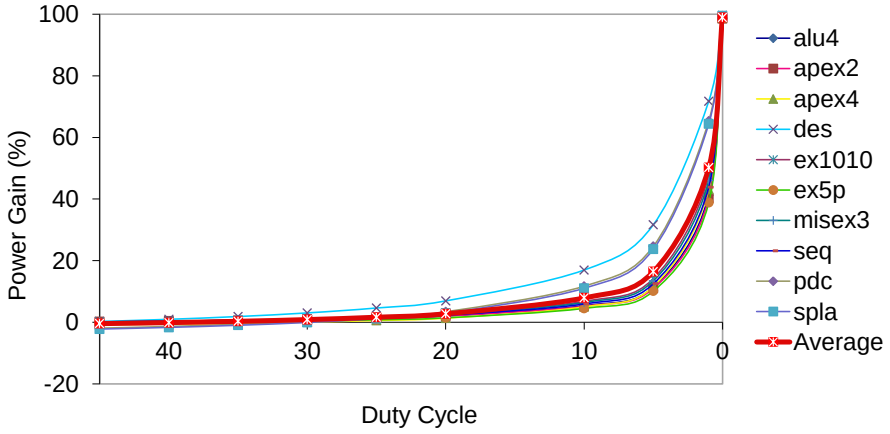


Figure A.23: Le gain de puissance en fonction de cycle d'utilisation pour le FPGA à base d'OxRAM avec les applications de sauvegarde de configuration. Considérant la durée ON 1%, la consommation gagnée atteint 50% en moyenne.

de SB,CB et LB.

L'approache logic-sur-mémoire est représentée sur la Fig. A.26 dans lequel la CRAM distribuées dans la tuile 2D est placée dans la couche basse de la tuile 3D. Le partitionnement avec Logic-sur-mémoire aide à réaliser les gains suivants: 1) Les CRAM sont maintenues près des nœuds de configuration permettant ainsi de réduire la complexité sur des fils de routage. 2) Comme toutes les cellules de mémoire sont dans la couche basse, la disposition de la cellule peut être optimisée avec les règles de conception de mémoires afin de réduire la surface et les cellules haute-V_t peuvent être utilisées sans affecter la surface et de réduire les courants de fuites. En outre, le circuit de programmation peut être intégré dans la couche basse ce qui diminue la complexité de routage causée par les fils de programmation. 3) Comme toutes les ressources de routage sont placées sur la couche haute, aucun signal de chemin critique traverse les couches permettant ainsi d'éviter l'effet de vias inter-niveaux.

Pour la conception de blocs du FPGA 3D, le processus 14nm FDSOI est utilisé. D'abord, le layout des blocs 2D sont préparés puis en plaçant la mémoire de configuration (CRAM) dans la couche basse, on obtient des blocs 3D. Toute la logique restante tels que des multiplexeurs et des buffers, sont placés dans la couche haute. La Fig.A.27 montre la conception 2D et 3D de MUX4 comme exemple.

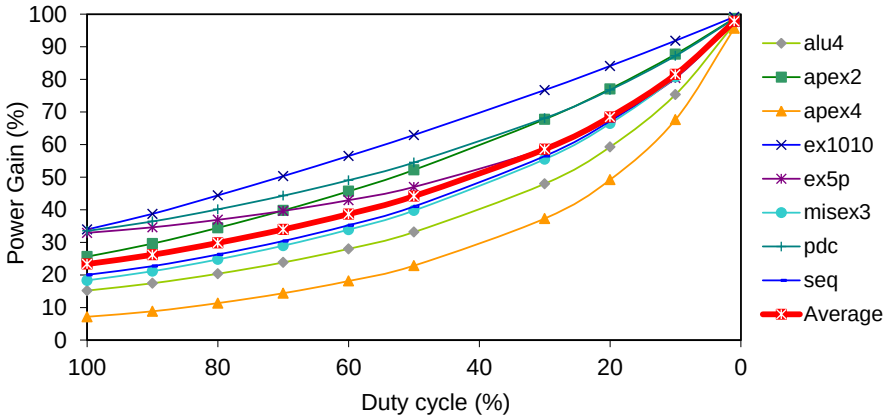


Figure A.24: Le gain de puissance en fonction de cycle d'utilisation pour le FPGA à base de CBRAM avec les applications de sauvegarde de configuration. Considérant la durée ON 1%, la consommation gagnée atteint 97% en moyenne.

A.4.2. Evaluation de 3D-FPGA à l'Approche Logic-sur-Mémoire

Pour l'évaluation FPGA, les résultats de simulations post-layout sont extraits tout d'abord. En utilisant la plateforme d'évaluation de VPR, les fichiers d'architecture pour les FPGA 2D et 3D sont créés. Le modèle de timing est mise à jour avec les changements de capacitance et de longueurs des fils. Le modèle de surface est mis à jour avec le paramètre de surface de mémoire en n'allouant pas de surface pour la mémoire de configuration puisque toutes les mémoires sont mises en couche basse.

Après l'exécution de VPR, il est observé dans la Fig. A.28 que la surface est réduite de 55%. Les avantages de surface de l'intégration 3D monolithique peuvent être décrits de la manière suivante : tout d'abord, la mémoire est complètement retirée de la couche logique. Deuxièmement, en raison de la forte densité de connexions verticales, le remplacement de la mémoire sur la couche basse n'impose aucun encombrement de routage. En particulier, l'utilisation des métaux entre les couches haute et basse permet le placement de mémoire très flexible tout en gardant la proximité de la couche logique.

Le produit énergie-délai (EDP), illustré sur la Fig. A.29, est réduit de 47% avec le 3D-FPGA proposé. L'amélioration de l'EDP est due à deux facteurs : Les délais intrinsèques des blocs sont réduits avec l'intégration 3D en raison de routage interne simplifiée et la longueur de fils de routage entre les blocs diminue ce qui conduit à des capacitances de routage réduites.

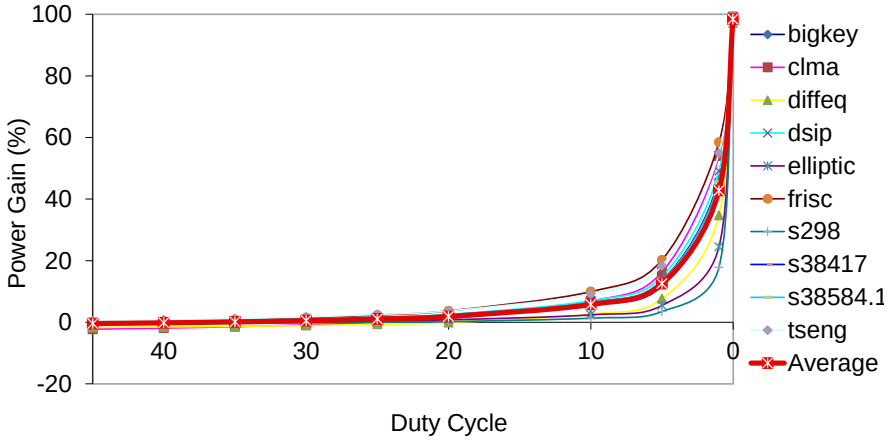


Figure A.25: Le gain de puissance en fonction de cycle d'utilisation pour le FPGA à base d'OxRAM avec les applications de sauvegarde de configuration et de contexte. Considérant la durée ON 1%, la consommation gagnée atteint plus de 40% en moyenne.

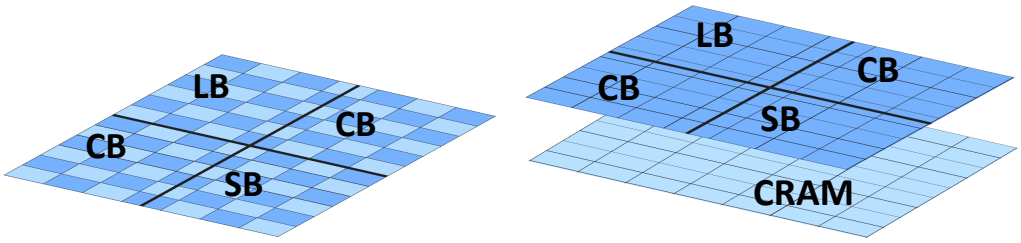
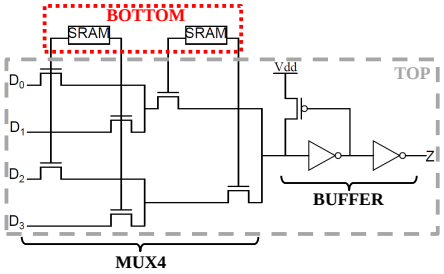
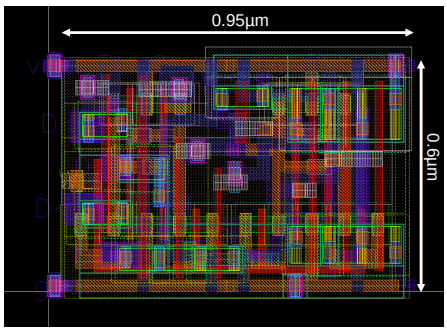


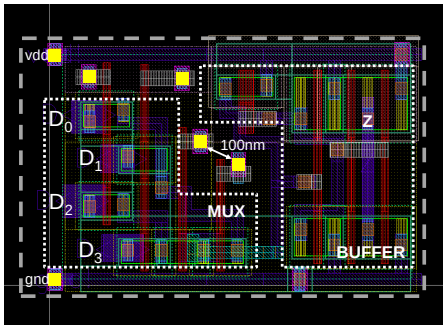
Figure A.26: L'approache Logic-sur-mémoire.



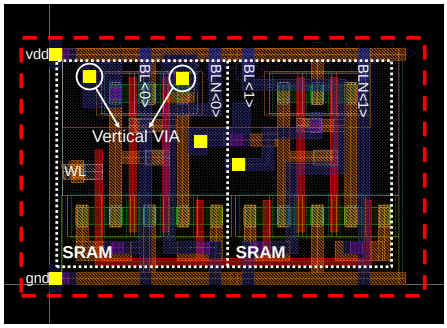
(a) Vue schematique.



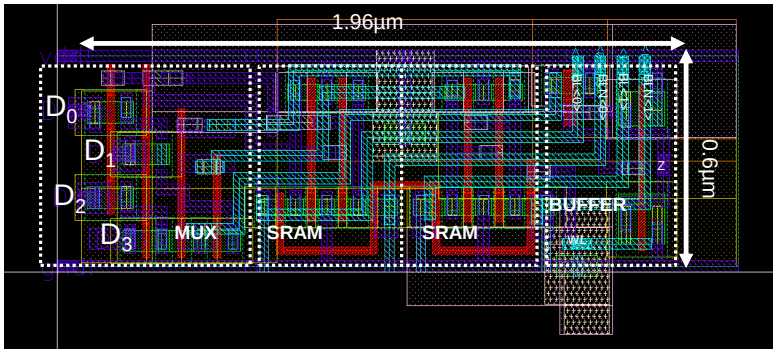
(b) Vue layout 3D.



(c) Vue couche haute.



(d) Vue couche basse.



(e) 2D layout view.

Figure A.27: MUX4 en 2D et 3D.

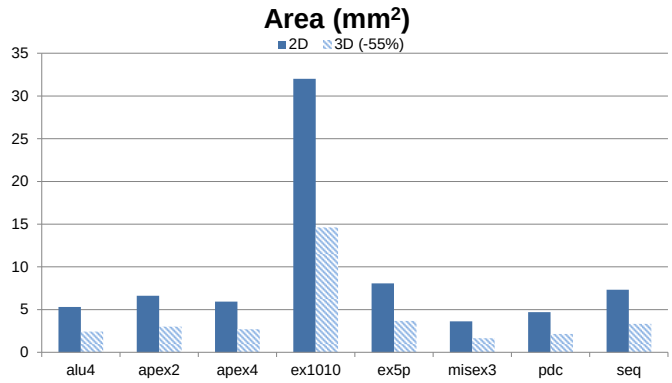


Figure A.28: La surface des circuits de référence FPGA pour 2D et 3D architectures. La surface peut être réduite de 55% en moyenne pour les blocs 3D.

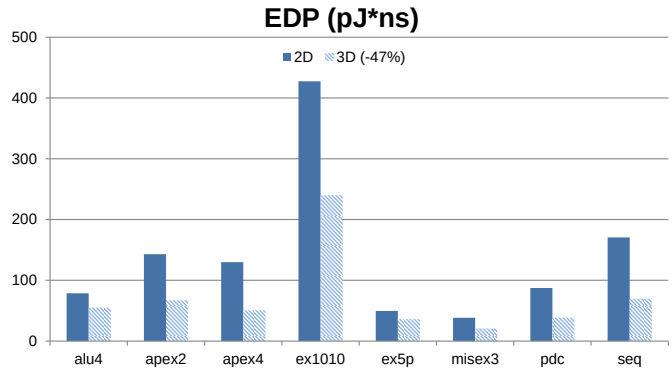


Figure A.29: EDP des circuits de référence FPGA pour 2D et 3D EDP peut être réduite de 47% en moyenne avec les blocs 3D.

A.4.3. 3D-FPGA Multi-Niveaux

Dans la section précédente, tous les blocs correspondants aux FPGA 2D et 3D sont conçus et le FPGA 3D est évaluée avec l'approche logique-sur-mémoire. En utilisant la bibliothèque de blocs 2D et 3D et en tenant compte des différentes granularités de partitionnement, l'exploration FPGA multi-niveaux est effectuée. Dans cette section, plusieurs blocs de partitionnement sont analysés en envisagent 2, 3 et 4 couches actives.

A.4.4. évaluation de 3D-FPGA Multi-Niveaux

Les fichiers d'architecture VPR distincts sont créés pour chaque manière de partitionnement. Afin d'observer les effets réelles de chaque manière de partitionnement, la longueur des fils de routage et la charge capacitive des fils sont modélisées à partir des améliorations de la surface. Le tableau A.2 montre les résultats des évaluations de circuits FPGA de référence envisagent plusieurs niveaux avec 3DMI. Les valeurs moyennes de tous les systèmes finaux sont normalisées à l'aide des valeurs moyennes du 2D pour pouvoir comparer.

Á partir de deux couches, par rapport au 2D, des améliorations de la surface significatives sont observées lorsque le nombre de niveaux est augmenté. La surface est réduite jusqu'à 55%, 69% et 77% par rapport au 2D pour les implémentations de niveaux 2, 3, et 4 respectivement.

Les résultats de l'EDP montrent que la réduction de la surface permet d'améliorer sensiblement l' EDP. Selon les résultats, deux différents types d'améliorations sont observées : 1) les réductions dues à la longueur plus courte de fils de routage et à la charge capacitive moins élevée 2) les réductions du délai interne de la grille et de la charge capacitive dues à l'approche logique-sur-mémoire. Lorsque deux couches sont considérées, l'EDP peut être réduite de 47% et 26% par rapport au 2D. Dans les solutions de trois et quatre couches, il est possible d'observer l'effet de la simplification de routage. Dans trois couches, jusqu'à 40% de réduction d'EDP est observée et dans quatre couches, jusqu'à 66% par rapport au 2D. Ces résultats montrent que l'amélioration de délai est proportionnelle à la réduction de routage néanmoins, la reconception de blocs peut encore améliorer l'EDP (ex. l'approche logique-sur-mémoire).

Table A.2: Area and EDP results of FPGA benchmark circuits based on multi-tier design

	Area (mm^2)							***EDP (pJ*ns)***						
	2D	***2L_1***	***2L_2***	***3L_1***	***3L_2***	***4L_1***	***4L_2***	***2D***	***2L_1***	***2L_2***	***3L_1***	***3L_2***	***4L_1***	***4L_2***
alu4	5.3	2.4	2.6	1.6	1.6	1.2	1.3	78.6	55.1	67	60	54.1	32.2	49
apex2	6.6	3	3.2	2	2	1.5	1.6	143	67.2	91.2	83.9	71.2	47.6	64.5
apex4	6	2.7	2.9	1.8	1.8	1.4	1.4	129.9	50.5	64.6	53.1	48.4	27.9	43.5
des	32	14.6	15.6	9.8	9.9	7.4	7.8	427.5	240.1	334.8	307	272.3	162.1	259.8
ex1010	8.1	3.7	3.9	2.5	2.5	1.9	2	49.5	36.1	46.7	39.7	40.2	19.2	30.5
ex5p	3.6	1.6	1.8	1.1	1.1	0.8	0.9	38.4	20.7	29.2	25.9	25.5	15.1	21.6
misex3	4.7	2.1	2.3	1.4	1.4	1.1	1.1	87.4	38.6	58.9	52.1	46.2	28.3	40.5
seq	7.3	3.3	3.6	2.2	2.3	1.7	1.8	170.5	69.7	135.1	111.9	101.4	51	90.4
Normalized to 2D (avg.)	**100**	**45.4**	**48.6**	**30.4**	**30.7**	**23.1**	**24.3**	**100**	**53**	**74.3**	**65**	**59.6**	**34.8**	**52.4**

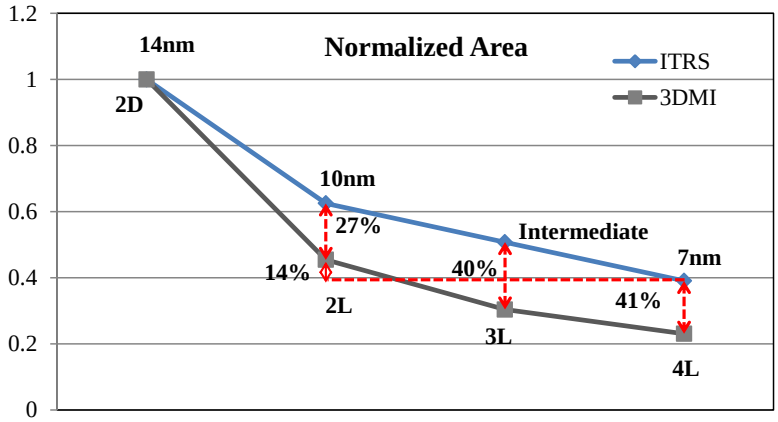


Figure A.30: Projection de l'amélioration de la surface pour les futurs nœuds technologiques basées sur l'ITRS 2013 et le gain de multi-niveaux approche 3DMI.

A.4.5. 3DMI Impact sur CMOS Scaling

Dans cette section, les meilleurs résultats obtenus par multi-niveaux FPGA sont comparés aux attentes de scaling. Lorsque le nombre de couches intégrées avec 3DMI est augmenté, la densité de la logique, de la performance et de l'efficacité énergétique sont améliorées. Le scaling traditionnel définit par la loi de Moore vise également à rendre ces valeurs meilleurs avec chaque avancement de nœud. La comparaison entre le FPGA multi-niveaux et le scaling traditionnel démontre ce qui peut être réalisé avec 3DMI pour les implémentations futures.

Le rapport récent de l'ITRS 2013 [168] conclut que le nombre de transistors devrait augmenter de 1,6 fois pour chaque nœud technologique. De même, dans 3DMI, comme plusieurs couches sont empilées, la densité de logique augmente. La tendance de l'ITRS et les meilleurs résultats de la surface du FPGA multi-niveaux sont illustrés sur la Fig. A.30. Le FPGA de 2L en 14nm diminue la surface de 27% de plus que le scaling en 10 nm. L'intégration dans 3L 3DMI peut être comparable à un nœud intermédiaire qui diminue de 40% la surface. Enfin, l'intégration 4L FPGA en 14nm augmente le gain en surface de 41% par rapport au scaling en 7nm.

Basé sur le rapport de l'ITRS [169], les améliorations EDP, jusqu'au nœud technologique 7 nm, sont calculées. La tendance de l'ITRS et les meilleurs résultats de l'EDP des FPGA multi-niveaux sont illustrés sur la Fig. A.31. Il est possible d'observer que l'implémentation 2L en 14nm arrive à atteindre les mêmes performances que le nœud de 10 nm, mais aussi elle améliore encore les performances de 24%. Puisque l'

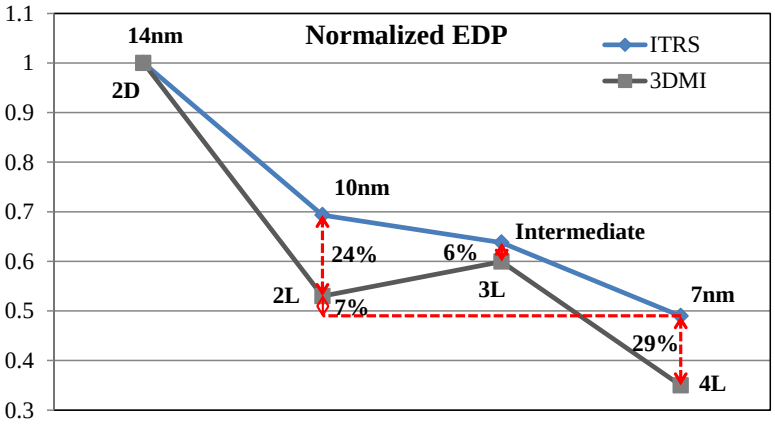


Figure A.31: Projection de l'amélioration de l'EDP pour les futurs noeuds technologiques basées sur ITRS 2013 et le gain de multi-niveaux approche 3DMI.

approche logique-sur-mémoire n'est pas utilisée dans 3L, l'amélioration relative est que de 6%. Ce résultat montre l'importance du partitionnement parce que le gain est fortement corrélé à la conception pour l'utilisation de la technologie 3DMI. Le cas 3L peut être un nœud intermédiaire qui vise une surface basse (voire la Fig. A.30) tout en gardant des performances similaires au prochain avancement de scaling. L'implémentation de 4L en 14nm dépasse les améliorations de 7 nm de 29%.

Par conséquent, plusieurs niveaux 3DMI montre un moyen très efficace d'améliorer la densité logique et l'EDP. Lorsque le nombre de couches intégrées est doublé, les gains atteignent non seulement les améliorations de scaling, mais les surpassent également. En outre, compte tenu des gains réduits de scaling, doubler le nombre de couches pourraient présenter les mêmes gains que deux noeuds d'avancement avec scaling. Par conséquent, plusieurs niveaux 3DMI est un très forte scaling alternative pour l'avenir.

A.5. Conclusion

Dans cette thèse, nous sommes motivés par les possibilités des technologies 3D émergents desquelles le FPGA peut profiter. Les technologies disponibles sont vastes et ils ont des avantages sur différents niveaux. Par conséquent, ces technologies ne devraient pas seulement être évaluées mais aussi être exploitées afin de trouver les avantages uniques. Dans ce travail, nous avons proposé un cadre d'évaluation pour les technologies émergentes appliquées aux FPGA. En utilisant du cadre d'évaluation du FPGA,

plusieurs technologies 3D émergents (OxRAM, CBRAM et 3DMI) ont été évalués. Les résultats montrent que l'empreinte de circuit peut être considérablement réduit avec les technologies 3D principalement parce que certaines des fonctionnalités de circuit peut être intégré dans la troisième dimension. En outre, le routage simplifié permet de meilleures performances et faible consommation d'énergie. De plus, l'inclusion de la non-volatilité conduit à de plus grandes économies d'énergie avec des fonctionnalités supplémentaires. Tous ces résultats montrent que les FPGA sont l'une des principales plates-formes informatiques qui offrent de nombreuses possibilités d'amélioration et maintenant nous pouvons repenser la façon dont le FPGA sera employé dans les futurs systèmes de calcul.