Cairo University

DESIGN EXPLORATION FOR NETWORK ON CHIP BASED FPGAS: 2D AND 3D TILES TO ROUTER INTERFACE

By

Alaa Salaheldin Gomaa Ibrahim

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

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2018

DESIGN EXPLORATION FOR NETWORK ON CHIP BASED FPGAS: 2D AND 3D TILES TO ROUTER INTERFACE

By
Alaa Salaheldin Gomaa Ibrahim

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

Under the Supervision of

Prof. Dr. Ahmed M. Soliman	Dr. Hassan Mostafa
Professor	Assistant Professor
Electronics and Communications Engineering Department Faculty of Engineering, Cairo University	Electronics and Communications Engineering Department Faculty of Engineering, Cairo University

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2018

DESIGN EXPLORATION FOR NETWORK ON CHIP BASED FPGAS: 2D AND 3D TILES TO ROUTER INTERFACE

By

Alaa Salaheldin Gomaa Ibrahim

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

Approved by the
Examining Committee

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

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

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

____
Prof. Dr. Fourth E. Name, Member

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2018

Engineer's Name:	Alaa Salaheldin Gomaa Ibrahim
Date of Birth:	18/03/1989
Nationality:	Egyptian
E-mail:	alaa.salaheldin89@gmail.com
Phone:	+201065896629
Address:	13rd Abo Ordia st, Ezbet Fahmy Al Basatin, Cairo 11742, Egypt
Registration Date:	01/10/2012
Awarding Date:	…./…./2018
Degree:	Master of Science
Department:	Electronics and Communications Engineering

Supervisors:

Prof. Dr. Ahmed M. Soliman
Dr. Hassan Mostafa

Examiners:

Prof. …………………	(External examiner)
Prof. …………………	(Internal examiner)
Prof. Prof. Dr. Ahmed M. Soliman	(Thesis advisor)
Prof. Dr. Hassan Mostafa	(Thesis advisor)

Title of Thesis:

Design Exploration for Network on Chip Based FPGAs: 2D and 3D Tiles to Router Interface

Key Words:
Network on Chip; Fields Programmable Gate Array; Router Interface

Summary:

This thesis explores how to adapt and use Networks-on-Chips for designing the next-generation FPGAs, a literature survey of existing Networks-on-Chip designs is presented. Then a comparative review between three NoC routers is provided; the comparison is held in the context of area and operating frequency, the comparison results show that increasing number of router ports affects the area, power and frequency of the network significantly. For that, the Codec is introduced; it is a tile to router interface used to connect more tiles or modules to the network without increasing the router port count. A comparison is held between two 2D networks, with and without Codec. Finally, the effects of adding Codec to 3D-NoCs are investigated.

Acknowledgments

Alhamdulillah, I praise and thank Allah for giving me the strength and courage to complete this thesis.

I would like to thank Prof. Ahmed M. Soliman and Dr. Hassan Mostafa for their help, support and patience.

Table of Contents

ACKNOWLEDGMENTS I

TABLE OF CONTENTS II

LIST OF TABLES IV

LIST OF FIGURES V

NOMENCLATURE VII

ABSTRACT VIII

CHAPTER 1 : INTRODUCTION 1

1.1. OVERVIEW AND MOTIVATION 1
1.2. CONTRIBUTIONS 2
1.3. ORGANIZATION OF THE THESIS 2

CHAPTER 2 : LITERATURE REVIEW 3

2.1. INTRODUCTION 3
2.2. FPGA VS. ASIC 3
2.2.1. Non-Recurring Engineering costs 3
2.2.2. Unit cost 4
2.2.3. Time to market 5
2.2.4. Scalability and configurability 5
2.2.5. Development cycle 5
2.2.6. Summary 6
2.3. NOC OVERVIEW 7
2.3.1. Why Choosing the NoC Approach 7
2.4. A CLOSER LOOK AT THE NOC ARCHITECTURE 11
2.4.1. Macro-Architecture view 12
2.4.1.1. Topology 12
2.4.1.2. Number of network nodes 12
2.4.1.3. Virtual channels 13
2.4.1.4. Routing algorithm and Flow control 13
2.4.1.5. Realization method 13
2.4.2. Micro-Architecture view 14
2.5. PREVIOUS WORKS 15
2.5.1. NoCem 15
2.5.2. PNoC 16
2.5.3. Dual Crossbar Router 17
2.5.4. HW NoC 19
2.5.5. SOTA 20
2.5.6. CONNECT 21
2.5.7. Split and Merge PS 23
2.5.8. FLNR 25
2.5.9. RROCN 28

2.6. COMPARATIVE REVIEW 29
2.6.1. Comparison workflow 29
2.6.2. Frequency 30
2.6.3. LUTs usage 32
2.6.4. Registers usage 34
2.7. SUMMARY AND FUTURE WORK 36

CHAPTER 3 : CODEC, TILES TO ROUTER INTERFACE 38

3.1. INTRODUCTION 38
3.2. MODELING AND SIMULATION 39
3.2.1. Simulation results 41
3.3. IMPLEMENTATION AND NETWORK SYNTHESIS 41
3.3.1. The Codec module 41
3.3.2. Network configurations 42
3.4. COMPARISON RESULTS 42
3.4.1. Frequency 43
3.4.2. Logic utilization 44
3.4.3. Power consumption 46
3.5. SUMMARY 48

CHAPTER 4 3D-NOC DESIGN EXPLORATION WITH CODEC 49

4.1. INTRODUCTION 49
4.2. 3D-NOC IN LITERATURE 49
4.3. 3D-NOCET AS AN EXPLORATION TOOL 50
4.3.1. Introduction 50
4.3.2. Tool updates 51
4.4. COMPARISON SETUP AND RESULTS FOR FULL-MESH TOPOLOGY 52
4.4.1. Logic utilization 53
4.4.1.1. Vertical complexity 53
4.4.1.2. Network complexity 54
4.4.2. Frequency 56
4.4.2.1. Vertical complexity 56
4.4.2.2. Network complexity 56
4.4.3. Power consumption 57
4.4.3.1. Vertical complexity 57
4.4.3.2. Network complexity 58
4.5. COMPARISON SETUP AND RESULTS FOR RING TOPOLOGY 58
4.5.1. Logic and memory utilization 59
4.5.2. Power consumption 61
4.6. SUMMARY 63

SUMMARY AND CONCLUSION 64

REFERENCES 65

List of Tables

Table 2.1: Area, Power and Operating frequency cost [10] 10
Table 2.2: PNoC router Implementation Results [15] 17
Table 2.3: Configurable Router for Embedded NoC synthesis results for FPGA and ASIC [16] 19
Table 2.4: Synthesis Results for CONNECT and SOTA Mesh Network [18] 22
Table 2.5: Map & Post-PAR report for Split-Merge and CONNECT on XC6VLX240T-1 [19] 24
Table 2.6: FLNR performance and area comparison with some previous NoC routers [20] 28
Table 3.1: Simulink models comparison 41
Table 3.2: Network B CpR and TpC configurations 42

List of Figures

Figure 2.1: ASIC NRE and mask costs for different technology nodes [3] 4
Figure 2.2: Unit cost for FPGA vs. ASIC [4] 4
Figure 2.3: Typical development cycle for FPGA and ASIC [5] 6
Figure 2.4: Comparison summary for FPGA vs. ASIC [6] 7
Figure 2.5: Basic architecture of an FPGA [7] 8
Figure 2.6: Basic NoC architecture for a mesh topology [8] 9
Figure 2.7: Relative wire delay in ASIC implementation [9] 10
Figure 2.8: High-level overview of the NoC research exploration [11] 11
Figure 2.9: A sixteen node NoC in a 4x4 Mesh topology [12] 12
Figure 2.10: SOTA router architecture [13] 14
Figure 2.11: FPGA routing and logic power consumption [14] 15
Figure 2.12: PNoC router block diagram [15] 16
Figure 2.13: Configurable Router for Embedded NoC block diagram [16] 18
Figure 2.14: Configurable Router for Embedded NoC FPGA resource utilization breakdown [16] 19
Figure 2.15: Hard and Soft NI Shell [17] 20
Figure 2.16: SOTA architecture [13] 21
Figure 2.17: CONNECT router architecture [18] 22
Figure 2.18: Split-Merge architecture [19] 23
Figure 2.19: Packet Format of CONNECT Network [19] 24
Figure 2.20: Packet Format of Split-Merge Network [19] 24
Figure 2.21: Cycle comparison between CONNECT and Split-Merge on uniform random traffic on an 8x8 mesh with eight flit packets [19] 25
Figure 2.22: FLNR packet format [20] 26
Figure 2.23: FLNR block diagram [20] 26
Figure 2.24: Synthesis results for FLNR [20] 27
Figure 2.25: FLNR performance and area comparison with previous routers [20] 28
Figure 2.26: RRCON router block diagram [24] 28
Figure 2.27: RRCON crossbar architecture [24] 29
Figure 2.28: Frequency vs. buffer depth 30
Figure 2.29: Frequency vs. data width 31
Figure 2.30: Frequency vs. VCs 31
Figure 2.31: LUTs utilization vs. buffer depth 32
Figure 2.32: LUTs utilization vs. data width 33
Figure 2.33: LUTs utilization vs. VCs 34
Figure 2.34: Registers utilization vs. buffer depth 34
Figure 2.35: Registers utilization vs. data width 35
Figure 2.36: Registers utilization vs. VCs 35
Figure 3.1: Metal layers in modern ASIC devices [9] 38
Figure 3.2: 6-port router in network A 39
Figure 3.3: 3-port router in network B 40
Figure 3.4: Router core of a 3-port router 40
Figure 3.5: Codec Simulink model 41
Figure 3.6: Maximum operating frequency comparison between network A and network B with 1CpR 43

Figure 3.7: Maximum operating frequency comparison between different network B configurations 44
Figure 3.8: Logic utilization comparison between network A and network B with 1CpR 45
Figure 3.9: Logic utilization comparison between different network B configurations 46
Figure 3.10: Power consumption comparison between network A and network B with 1CpR 47
Figure 3.11: Power consumption comparison between different network B configurations 47
Figure 4.1: 3D-NOCET GUI without modification 51
Figure 4.2: Updated 3D-NOCET 52
Figure 4.3: LUTs utilization for different numbers of Tiers 53
Figure 4.4: Registers utilization for different numbers of Tiers 54
Figure 4.5: LUTs utilization for different numbers of Tiles per Tier 55
Figure 4.6: Registers utilization for different numbers of Tiles per Tier 55
Figure 4.7: Maximum operating frequency for different numbers of Tiers 56
Figure 4.8: Maximum operating frequency for different numbers of Tiles per Tier 57
Figure 4.9: Power consumption for different numbers of Tiers 58
Figure 4.10: Power consumption for different numbers of Tiles per Tier 58
Figure 4.11: LUTs utilization for Ring 59
Figure 4.12: LUTs utilization for RingWithCodec 60
Figure 4.13: Memory utilization for Ring 60
Figure 4.14: Memory utilization for RingWithCodec 61
Figure 4.15: Power consumption for Ring 62
Figure 4.16: Power consumption for RingWithCodec 62

Nomenclature

ASIC	Application Specific Integrated Circuits
BRAM	Block RAM
BSV	Bluespec System Verilog
CLB	Configurable Logic Blocks
DOR	Dimension Ordered Routing
DRAM	Distributed RAM
DSP	Digital Signal Processing
FF	Flip Flop
FPGA	Field Programmable Gate Arrays
IP	Intellectual Property
LUT	Look Up Table
NI	Network Interface
NoC	Network on Chip
NRE	Non Recurring Engineering
PAR	Place and Route
PCF	Physical Constraints File
PDR	Partial Dynamic Reconfiguration
QoS	Quality of Service
RLOC	Relative Location Constraints
SAMQ	Statically Allocated Multi Queue
SoC	Systems on Chip
VC	Virtual Channel

Abstract

Due to the continuous demand for larger and more powerful chips, new blocks are added contentiously to System on Chips (SoCs), such as embedded processors, digital signal processors (DSPs), peripheral interfaces and embedded memory blocks. As the system complexity increases, the negative impact of its routing fabric increases as well. Bus-based and point-to-point interconnects become bottlenecks as they are unable to meet the system requirements. In general, they are not suitable for large systems as their performance degrades when used to connect many blocks. In addition, these interconnects normally include very long wires (global wires) to connect all parts of the chip and these global wires contribute heavily to the increased area and power consumption of the routing fabric.

Field programmable gate arrays (FPGAs) are like SoCs, new blocks and components are continuously added to their architecture in order to meet the increased demand of today's applications. With the increased number of components, the interconnect fabric starts gradually to use Network on Chips (NoCs) to overcome the problems of conventional point-to-point and bus-based interconnects. NoC consists of a network of routers connected with short links, for an FPGA block or tile to connect to another one, it only has to send its data to the nearest router instead of using global wires.

A review for several NoC designs is provided to get an idea about the current research state in this topic. The review is conducted in the context of contributions, architecture, implementation and future work. Then a comparison is held between three NoC routers to analyze the effect of changing the number of Virtual Channels (VCs), flit data width and buffer depth on the consumed area (LUTs and registers) and operating frequency. The comparison shows that the NoC architecture affects the area and maximum operating frequency of the system significantly.

As a result of the mentioned comparison, it is found that one drawback of using NoC is that increasing the router port count affects the area, power and frequency of the system significantly. In order to overcome this problem and to make the NoC approach useful in designing the next generation of FPGAs, a concentrator module or a Codec is proposed to connect between routers and multiple Tiles (FPGA basic building block). Codec reduces the effect of increasing tile count on the area, power and frequency of the routing network.

In order to evaluate the effect of using Codec, a comparison is held between two networks with the same topology and size, one uses routers only and the other uses routers and Codec modules. The comparison is held in the context of area, power and maximum operating frequency. The comparison results show that the area of the Codec network is only 15% compared to the routers only network, its power consumption is 50% less, and operates with 2.5x higher frequency.

Finally, as the three-dimensional integrated circuits technology (3D-IC) is increasingly adopted to cop up with the application demands, the effect of adding Codec to 3D-NoC systems is also investigated.

Chapter 1 : Introduction

1.1. Overview and Motivation

FPGAs (Field Programmable Gate Arrays) are used increasingly in today's applications because of their low development cost, fast design cycle, configurability and short time to market. On the other hand, ASICs (Application Specific Integrated Circuits) have long design cycle, poor configurability and require high development effort. These strong points of the FPGA made it an appropriate candidate for most research and industry applications. However, these advantages come at a significant cost in delay, area and power consumption caused mostly by their programmable routing fabric.

An FPGA mainly consists of three components. Processing elements (PEs), storage elements (SEs) and a complex programmable routing fabric. PEs are programmable logic blocks that perform logic calculations, for example, look-up tables (LUTs) with a fixed configuration of logic gates. SEs are memory blocks placed across the chip area; they are used to store data or algorithm states. The programmable routing fabric is a massive network of wires, multiplexers and bus-based interconnects; all used to connect PEs, SEs and IPs (Intellectual Property cores).

Due to the continuous demand for more powerful and larger chips, new blocks are added to the FPGA architecture, such as Digital Signal Processing (DSP) blocks and embedded processors. As the system complexity increases, the negative impact of the routing fabric increases as well. Bus-based interconnects, such as ARM's AMBA [1] and IBM's CoreConnect [2], become bottlenecks since they are unable to meet the system requirements. In general, they are not suitable for large systems as their performance degrades if used to connect many blocks. In addition, these interconnects include very long wires (global wires) that connect all parts of the chip, these global wires contribute heavily to the increased area and power consumption of the routing fabric.

Network on Chip (NoC) comes as a promising solution for the conventional interconnects problems. NoC has the benefits of independent implementation and optimization of nodes, simplified and customized architecture per application, support for multiple topologies and options, reduced area and power consumption, scalability and increased operating frequency.

Using the NoC approach instead of depending on long interconnect wires solves the conventional interconnect problems because NoC uses high-speed optimized lanes to transfer packets between the routers, and these routers interface with the main application blocks through a configurable number of input/output ports solving most of the problems introduced by long and medium-size routing wires.

Correspondingly, the NoC approach is the right choice as an interconnect fabric for the next generation FPGA. On the other hand, the problems of integrating NoC into the FPGA architecture should be investigated and solved which has been addressed in this research work.

1.2. Contributions

- Review of different NoC designs; especially their architectures and performance measurement results.
- Comparative review of three NoC routers; this comparison helps to determine which parameters or sub-modules need to be optimized to better adapt NoC for FPGA integration.
- Introduce Codec as a solution to the increased router port count problem; a comparison is held between two 2D networks, with and without Codec.
- Investigate the impact of integrating Codec into 3D-NoC.

1.3. Organization of the thesis

The following sections of the thesis are organized as follows. In Chapter 2, a survey is provided for several NoC routers followed by a comparative review of three of these routers. In Chapter 3, the modeling and simulation of Codec are provided, then a comparison between two 2D networks is held to show the impact of using Codec. In Chapter 4, the impact of adding Codec to 3D-NoC is investigated. Finally, a discussion and conclusion chapter.

Chapter 2 : Literature Review

2.1. Introduction

In this chapter, FPGAs are compared to ASICs with respect to non-recurring engineering cost, unit cost, time to market, scalability, configurability and development cycle. Then an introduction to NoC is provided that especially highlights the importance of NoC for FPGA. Then a literature review of various NoCs is presented; the review shows their contribution, architecture, implementation, performance measurement results and future work. Finally, a comparison between three NoC routers is held; the results give design guidelines and recommendations to help choose the appropriate NoC according to system requirements.

2.2. FPGA vs. ASIC

In general, FPGAs outperform ASICs due to their configurability, programmability and scalability. Unlike ASICs, which are designed to implement a specific function, FPGAs can be programmed to implement different digital functions and their function can be changed in a matter of seconds. In addition, FPGA has a relatively short design cycle compared to ASIC as no need for physical manufacturing.

These advantages of FPGA come at the cost of a larger area, higher power consumption and lower operating frequency. Although ASICs have better performance, FPGAs' market share is increasing because of their flexibility and shorter development and deployment cycles.

2.2.1. Non-Recurring Engineering costs

The design of ASIC chips requires going through a long expensive process that at least includes the costs of engineering teams for design and layout, software licenses for EDA tools, masks production and finally extensive testing. These costs vary with the target manufacturing technology and with the complexity of the chip itself. Figure 2.1 shows that the NRE costs of an ASIC design increase with technology advances.

On the other hand, most of the mentioned costs are excluded for an FPGA design. In most cases, a complete FPGA design only requires buying an FPGA chip and a compilation tool.

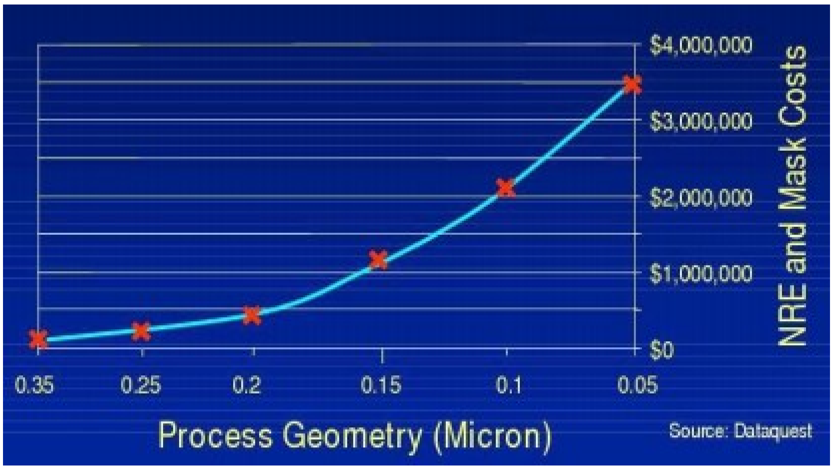


Figure 2.1: ASIC NRE and mask costs for different technology nodes [3]

2.2.2. Unit cost

Despite its high NRE costs, the unit cost of ASICs is lower than the unit cost of FPGAs when used for high volume production. In Figure 2.2, it is shown that the target volume production is an essential factor to determine which approach should be used to reduce the total cost. In addition, the same figure shows that the increased initial costs of new technology nodes are in favor of FPGA since such costs affect the ASIC approach significantly.

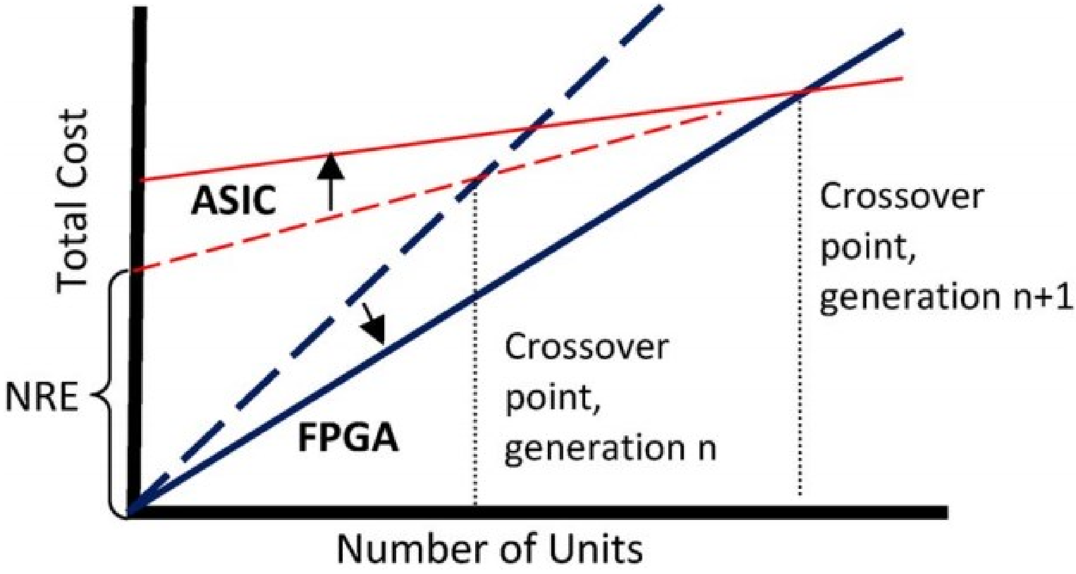


Figure 2.2: Unit cost for FPGA vs. ASIC [4]

2.2.3. Time to market

The long time-to-market is one of the bottlenecks facing the development of ASIC; decreasing the process feature length brings deep submicron effects that need longer time for mitigation and testing. On the other hand, introducing a new feature into FPGA might initially take a long time. However, this time is still less than implementing the feature using ASIC. In addition, once the feature is implemented on an FPGA, it can be deployed in most cases by a software upgrade without the need for hardware changes.

2.2.4. Scalability and configurability

Once manufactured, an ASIC chip cannot be reconfigured because its internal modules and interconnects are fixed. On the other hand, FPGA can be reconfigured with a new design in a matter of seconds thanks to its programmable building blocks and interconnects; this makes FPGA much more scalable, configurable and flexible compared to ASIC.

2.2.5. Development cycle

FPGA has a shorter development cycle compared to ASIC. For an FPGA, synthesis, timing analysis, placement and routing can be handled by the vendor software and the results are very close to an actual running system. Having a short design cycle enables early system integration and testing; which leads to early detection and investigation of possible issues.

As shown in Figure 2.3, ASIC designs need at least few months for the semiconductor foundry to produce first samples. In addition, a time-consuming floor planning and verification tasks need to be done efficiently in order to reduce the chip malfunctioning risks.

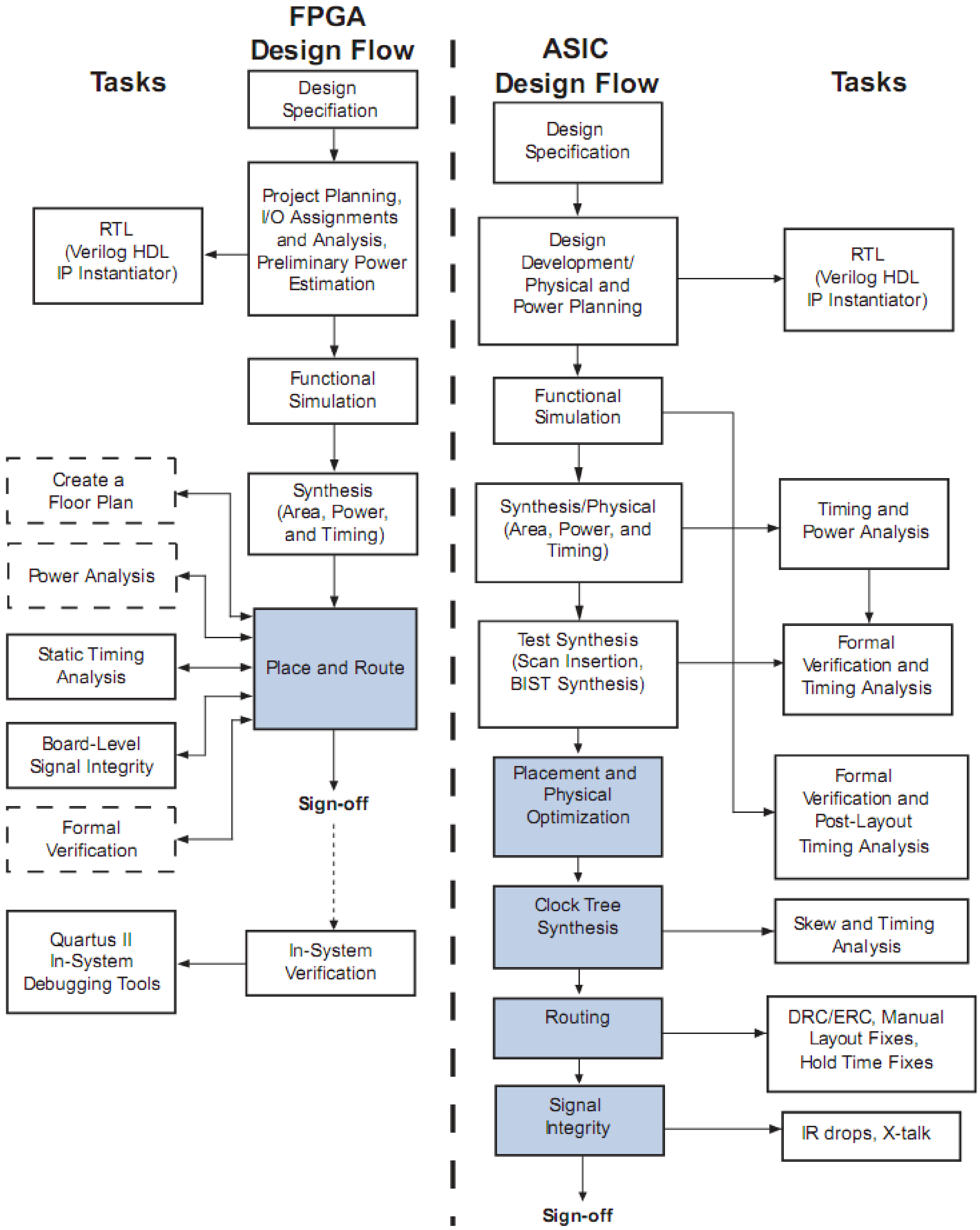
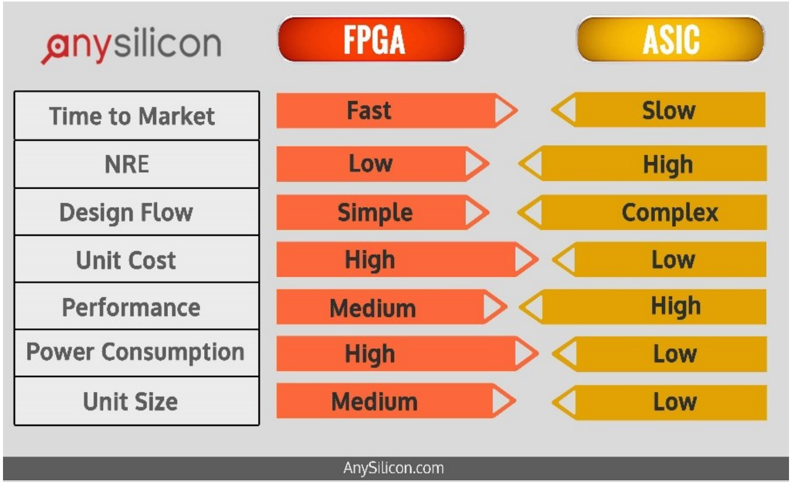

Tasks
FPGA Design Flow
ASIC Design Flow
Tasks
Design Specifiation
Project Planning, I/O Assignments and Analysis, Preliminary Power Estimation
RTL (Verilog HDL IP Instantiator)
Functional Simulation
Synthesis (Area, Power, and Timing)
Create a Floor Plan
Power Analysis
Static Timing Analysis
Board-Level Signal Integrity
Formal Verification
Place and Route
Sign-off
Quartus II In-System Debugging Tools
In-System Verification
Design Specification
Design Development/ Physical and Power Planning
RTL (Verilog HDL IP Instantiator)
Functional Simulation
Synthesis/Physical (Area, Power, and Timing)
Timing and Power Analysis
Test Synthesis (Scan Insertion, BIST Synthesis)
Formal Verification and Timing Analysis
Placement and Physical Optimization
Formal Verification and Post-Layout Timing Analysis
Clock Tree Synthesis
Skew and Timing Analysis
Routing
DRC/ERC, Manual Layout Fixes, Hold Time Fixes
Signal Integrity
IR drops, X-talk
Sign-off

Figure 2.3: Typical development cycle for FPGA and ASIC [5]

2.2.6. Summary

Figure 2.4 shows the differences between FPGAs and ASICs. In short, FPGA has a faster time to market as no chip production is needed neither an extensive hardware verification. NRE of ASIC is higher. The design flow of an FPGA implementation is more straightforward and faster. The unit cost of an FPGA chip is relatively higher than of ASIC in case of low volume production. The performance of an ASIC unit is always higher. The power consumption of FPGA is higher mostly due to its routing fabric and finally, the unit size of FPGA is bigger since ASIC is optimized for the size, area and power needs of a specific application.



anysilicon	FPGA	ASIC
Time to Market	Fast	Slow
NRE	Low	High
Design Flow	Simple	Complex
Unit Cost	High	Low
Performance	Medium	High
Power Consumption	High	Low
Unit Size	Medium	Low

AnySilicon.com

Figure 2.4: Comparison summary for FPGA vs. ASIC [6]

2.3. NoC Overview

2.3.1. Why Choosing the NoC Approach

Figure 2.5 and Figure 2.6 show the simple architecture of FPGAs and NoCs respectively. FPGAs consist of logic blocks used as gates or registers, I/O pins to interface with the outer world and interconnect fabric that includes switch blocks (MUXs and SRAMs) and wires (segmented, non-segmented, short, medium and long wires). In the following, some problems of conventional interconnect fabric are listed:

- Wire-speed does not scale with technology advances: Reducing the feature length of a technology node reduces the resistance and the active current of the transistors. However, it cannot do the same for a wire; both wire capacitance and resistance are increased with the technology advances leading to lower operating frequencies. Figure 2.7 shows the relative delay impact of process technology advances for gate delay, local and global wires.
- Large area: Interconnect wires come in different lengths; short wires are used to connect local nodes or to connect logic blocks that are close, medium wires are used to connect not too close and not too far logic blocks and finally long or global wires are used to traverse the chip and to connect between very far logic blocks. As shown in figure 2.5, it is inevitable to use global wires since some applications utilize a large number of I/O pins; for this scenario, placement and routing software configures the interconnect fabric to use global wires in order to connect different partitions.
- Large power consumption: Due to the continuous reduction of the transistor dimensions with each technology advance; the dynamic power consumption of logic blocks is decreasing. However, the dynamic power consumption of routing resources is considerably increasing. A switching box contains a large number of MUXs, wires and SRAMs to make it flexible and configurable as much as possible; this comes with the cost of a larger area and more power consumption.

- Slow compilation: In addition to the compilation effort needed to configure the LUTs, the vendor tool has to do some extra effort to configure the switching blocks in order to connect design partitions correctly and most efficiently. In addition, after the placement and routing are completed, a time analysis task is run to assure that the routing layout does not violate any timing constraints.

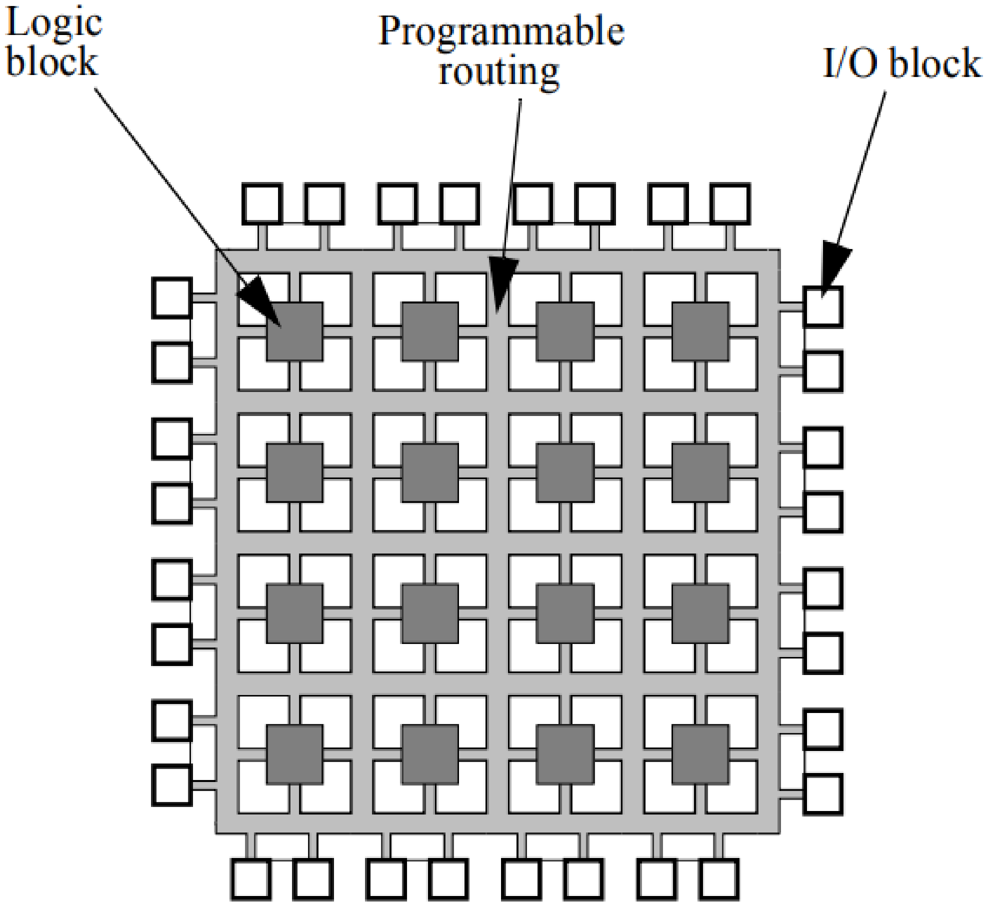


Figure 2.5: Basic architecture of an FPGA [7]

Like FPGAs, NoCs consist of a network of nodes and wiring resources to connect between the nodes. However, the difference is that a NoC node contains a router which is responsible for receiving and buffering packets until they are routed to their correct destination. The user logic blocks need to be connected only to one router port to reach any other logic block in the network. Using this approach, and assuming that the delay introduced due to multi-hop processing is minimized, there is no need to use long or global wires.

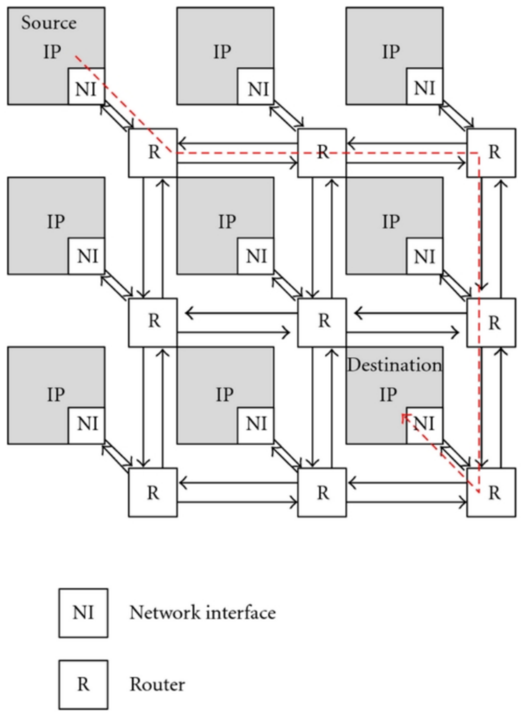


Figure 2.6: Basic NoC architecture for a mesh topology [8]

Figure 2.6 shows an example of NoC in a mesh topology, the network consists of routers interconnected with short and high-speed wires; each router provides a connection port locally to an IP or a logic block. In the following, some advantages of using the NoC approach are listed:

- More power and area efficiency: Interconnect fabric of FPGA is reduced to few routers, short links between routers and network interface adapters to connect logic blocks.
- Higher operating frequencies: Using shorter wires to connect routers solves the problem of long or global wires that lowers the maximum operation frequency of an FPGA design.
- NoC links are re-usable: A link between two routers is shared for multiple source-destination pairs and it is not dedicated to a single pair.
- Customization per application: The number of routers, topology, buffer depth, router interconnecting wires and routing algorithm can be optimized per application.
- Allows partial reconfiguration: Reprogramming of a single node doesn't affect the overall operation of the network as long as the time for this reprogramming is taken into consideration.
- Higher level abstraction: Network and application blocks are seen as two independent entities; that means that the design and implementation of both are done independently. In addition, the effort for the final integration is not significant as each part is verified alone before integration.

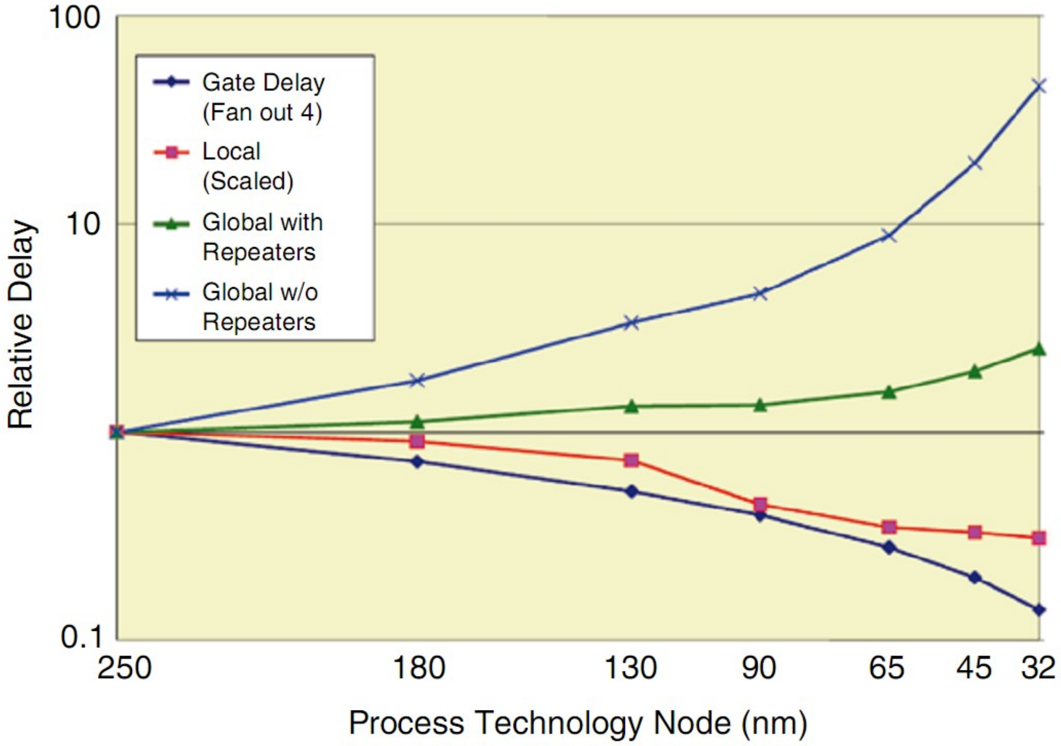


Figure 2.7: Relative wire delay in ASIC implementation [9]

In [10], the authors show the natural NoC scalability advantages. In addition, they compare the NoC approach with three alternative interconnect architectures; a non-segmented shared bus, a segmented bus and a point-to-point interconnect similar to that of an FPGA interconnect. An analytical expression is derived for the area, power and operating frequency for each architecture. As shown in Table 2.1 which summaries the findings for this study (the symbol (n) represent the number of modules or nodes of the system), NoC outperforms all alternative architectures in the context of the area, power and operating frequency.

Table 2.1: Area, Power and Operating frequency cost [10]

Arch	*Total Area*	*Power Dissipation*	*Operating Frequency*
NS-Bus	$O\left(n^3\sqrt{n}\right)$	$O\left(n\sqrt{n}\right)$	$O\left(\frac{1}{n^2}\right)$
S-Bus	$O\left(n^2\sqrt{n}\right)$	$O\left(n\sqrt{n}\right)$	$O\left(\frac{1}{n}\right)$
NoC	$O(n)$	$O(n)$	$O(1)$
PTP	$O\left(n^2\sqrt{n}\right)$	$O\left(n\sqrt{n}\right)$	$O\left(\frac{1}{n}\right)$

2.4. A Closer look at the NoC Architecture

The research of NoC usually divides the design space into two parts; Macro-Architecture and Micro-Architecture. As shown in Figure 2.8, Macro-Architecture exploration looks at the system as a whole; i.e., which topology and routing algorithm are used. The other view, Micro-Architecture exploration investigates the individual hardware components of the system and tries to adapt and optimize each component according to the system requirements.

As the NoC router is considered to be the main component of the system, an investigation on router design parameters and performance is discussed in this chapter. In addition, a mixture between NIC and topology investigation is discussed in Chapter three and four of this thesis.

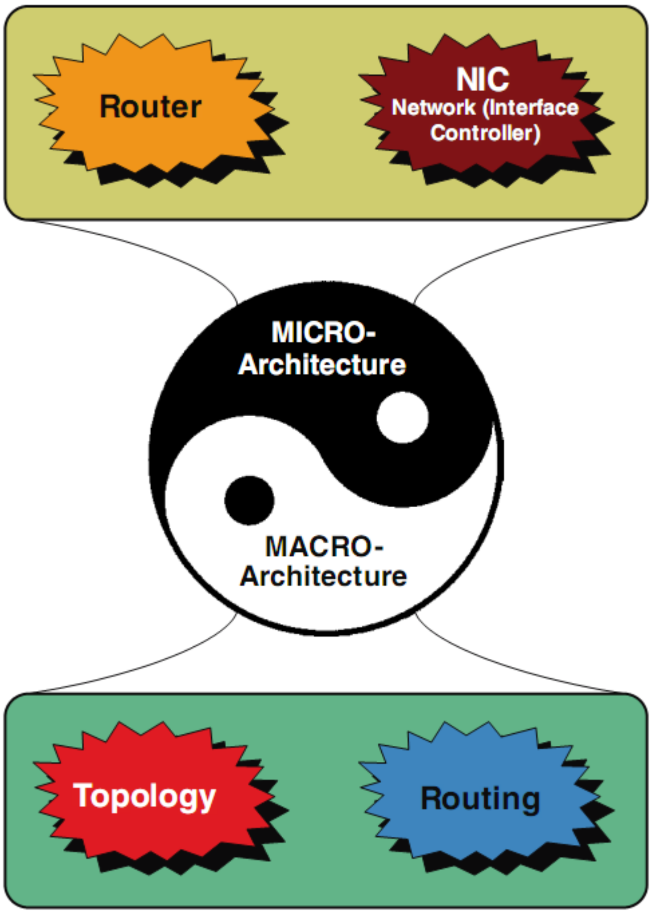


Figure 2.8: High-level overview of the NoC research exploration [11]

2.4.1. Macro-Architecture view

2.4.1.1. Topology

Network topology determines how the network components are physically connected to each other and it is directly related to design tradeoff between latency and area. A simple 4x4 mesh topology is shown in Figure 2.9, where each internal router has five ports, four to connect with its neighbor routers and one to interface with a local logical block or an IP. Other topologies exist as well, for example, tori, ring, full-mesh and cubes. Sometimes an irregular topology is used to fit a specific traffic load or application best.

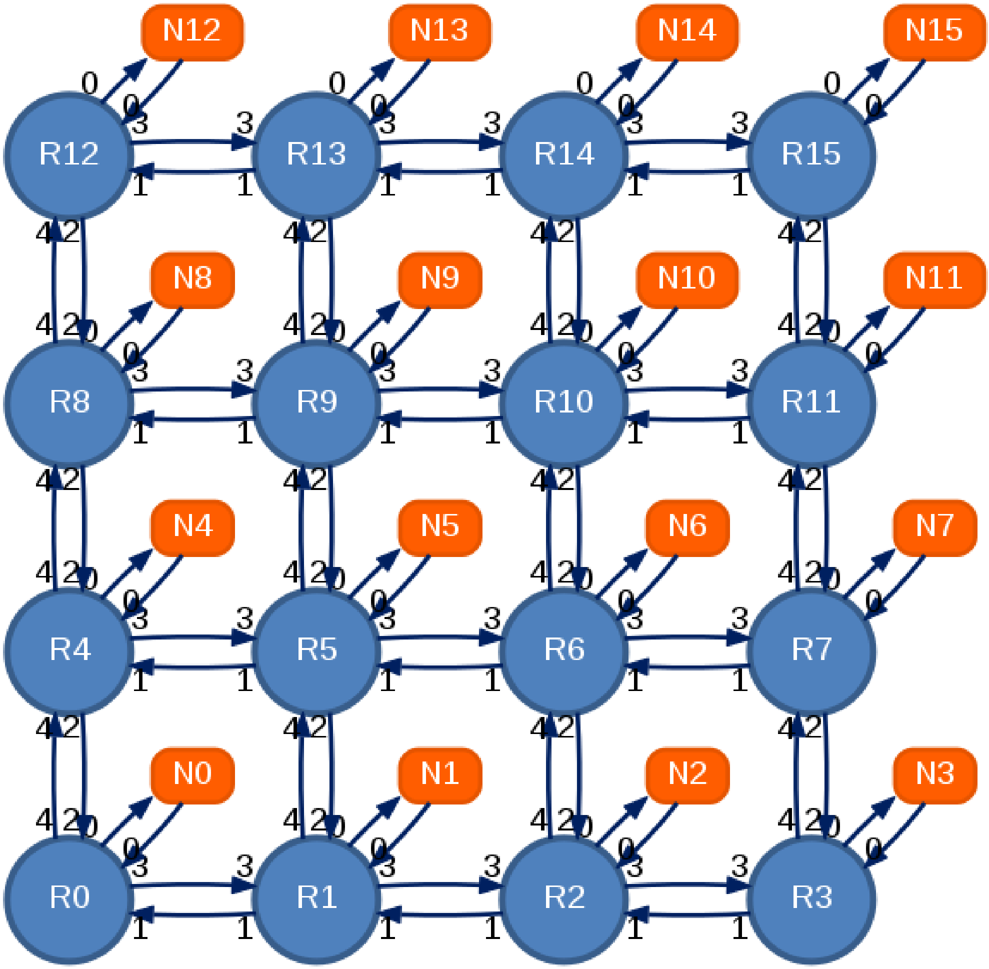


Figure 2.9: A sixteen node NoC in a 4x4 Mesh topology [12]

2.4.1.2. Number of network nodes

The number of network routers or nodes affects the network latency, throughput and area directly.

2.4.1.3. Virtual channels

A virtual channel (VC) divides physical channels between two routers into a set of logical separated channels in order to increase link utilization and improve performance. In addition, VCs add the support of quality of service (QoS) features to the NoC approach, in which a specific set of VCs are prioritized over other VCs. The benefits of adding VCs come with the cost of more area and power; with each added VC, input and output buffers are used to handle the traffic for this VC and this increases the size and complexity of the VC allocators.

2.4.1.4. Routing algorithm and Flow control

Since every link between two routers is shared and used to transfer packets between multiple source-destination pairs, the problems of network congestion, traffic blocking, deadlock and starvation are introduced. A suitable routing algorithm is essential to avoid or reduce the occurrence of these problems. Upon reception of an incoming packet or flit, each router has to determine the best route to deliver it to its destination.

Deterministic routing algorithms use known and pre-calculated routing tables to make the routing decision, a routing table is usually stored in all routers. For example, in XY routing algorithm, the packets are routed first to the X dimension until they reach a router that is located in the same Y coordinates as the destination router.

Adaptive routing algorithms take more area and more complex than deterministic algorithms. However, they address the mentioned problems more effectively. The routing algorithm of a router is not constant and it changes with traffic load, neighbor health status; this gives the network more flexibility and reliability.

Flow control is used to transfer a packet from a router to its neighbor regardless of its final destination and it is responsible for buffer resource allocation needed to transfer packets from one router to another. A control mechanism is needed for that process to avoid packet drops caused by buffer overflows, overruns and underruns. Two flow control techniques are discussed in the following:

- ON-OFF flow control: Each router checks the resources available for all ports and VCs, if it is below a defined threshold it sends an OFF signal to its neighbors. When a neighbor router receives such signal, it should not send a new packet until it is set back to ON.
- Credit-based flow control: Each router signals its resource availability to its neighbors as counters. The router itself updates these counters as soon as a packet enters or leaves one of its buffers; with this method, the neighbor routers know when it is valid to send new packets to this router.

2.4.1.5. Realization method

NoC can be integrated as a hard or soft component. As a hard component, it is implemented as fixed silicon circuits. On the other hand, as a soft component, it can be reconfigured which gives it more flexibility. Hard components are more area and power efficient and reach higher performance compared to soft components. In general, the realization method of a NoC component has to take into consideration the trade-off between flexibility and area, power and performance.

2.4.2. Micro-Architecture view

Figure 2.10 shows the micro-architecture of the NoC router introduced in [13] and it also represents the architecture of a typical NoC router.

Before dividing it into subcomponents, a simple scenario is explained to describe how a router behaves when it receives a packet:

1. Store the received packet in input buffers: If the packet cannot be forwarded immediately, it is stored in FIFO buffers at the router input ports until further processing.
2. Route calculation: According to the defined routing algorithm of the network, the router selects a suitable output port and a set of possible output VCs for each packet. This step needs to be done only once for the header flit of a received packet.
3. VC allocation: A limited set of VCs of the output ports are shared with all input ports' VCs. In this step, the virtual channel allocator assigns the available output VCs to the input port packets. This is done only once per packet for the header flit.
4. Switch or Crossbar allocator: After a packet is assigned to an output port and VC, in order to traverse from the input port to the output port, the switch allocator first resolves the conflicts between flits having the same target output port then it schedules a time slot for each one of them to access the destination output port.
5. Crossbar traversal: A flit traverses the router crossbar if a grant is received from the switch allocator. After successful traversing to an output port, the flit is ready to move on to the network.

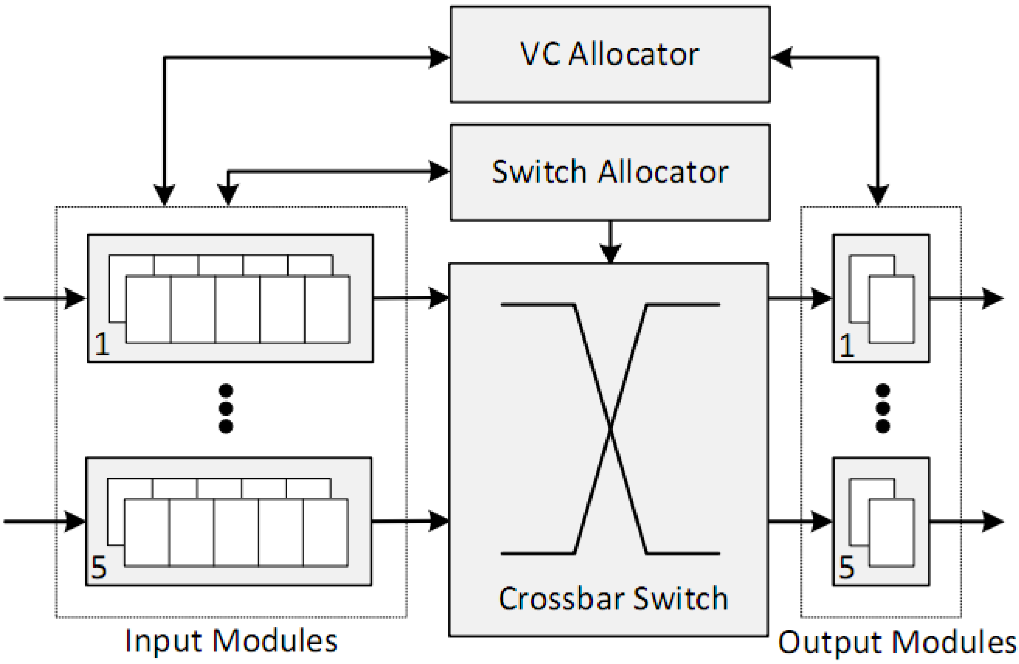


Figure 2.10: SOTA router architecture [13]

From the previous list and Figure 2.10, it is clear that a NoC router is divided into these main subcomponents:

- Input-output modules: The primary function of the input module is to buffer input flits until routing and allocation calculations are complete; each packet is stored in a different part of the buffer according to its VC identifier.
- Virtual channel allocator: Acts as a route computation logic that calculates to which VC the packet is assigned.
- Switch allocator: Resolves the conflicts between input flits having the same output port as a destination and assigns time slots to access the crossbar.
- Crossbar: Connecting all input ports with output ports.

2.5. Previous Works

2.5.1. NoCem

NoCem is a NoC emulation tool. G. Schelle and D. Grunwald [14] propose it with configurable network topology, channel FIFO depth, data width and packet length. To guarantee the flexible integration with other tools, it provides a common external interface.

Figure 2.11 shows NoCem architecture components, which are:

- VC: Each physical channel has a number of VCs that divide it into multiple lanes, which leads to higher throughput.
- Node Arbitration: It handles VC and switch allocations so that all incoming and outgoing transactions are capable of taking the proper arbitration decisions. Flit-reservation algorithm is used for flow control.
- Node Switch: It is an all-to-all multiplexer. This module is responsible for allowing multiple simultaneous paths of communication.

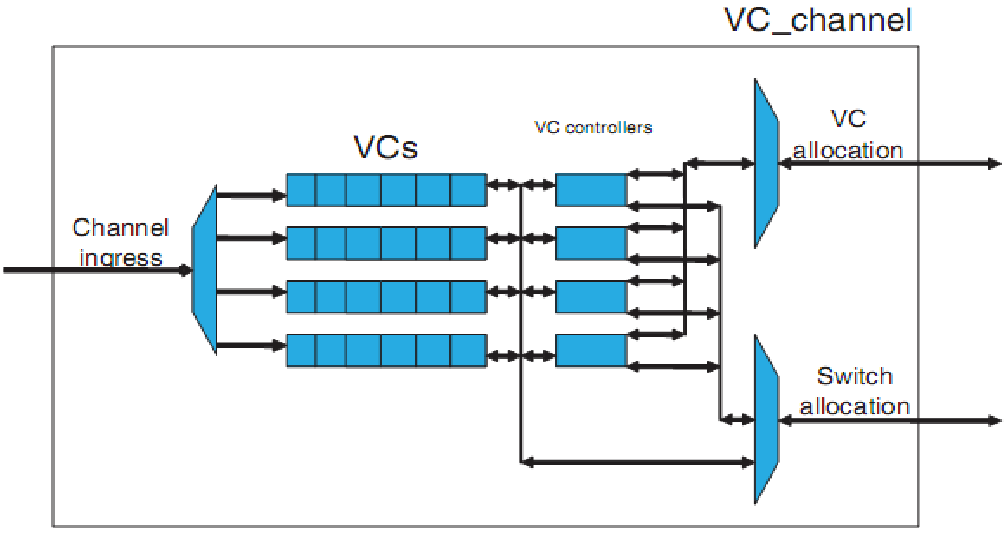


Figure 2.11: FPGA routing and logic power consumption [14]

The main parameters of the NoCem architecture are data width, network topology, channel FIFO depth, and packet length.

Using a Virtex-II Pro Xilinx FPGA, NoCem is implemented and tested. In [14], it is compared with a simple NoC that does not support VCs, has buffers with single word capacity per channel and it includes a simple switch. The comparison is held for three applications; a cryptographic accelerator, a synthetic benchmarking application and an 802.11 transmitter. The comparisons for the cryptographic accelerator and synthetic benchmarking applications show that using complex NoC does not always give better performance. On the other hand, VC implementation is very efficient for data flow applications demonstrated by the 802.11 transmitter.

2.5.2. PNoC

C. Hilton and B. Nelson [15] introduce an FPGA-embedded circuit switched NoC. It is configured with different topologies and data paths. In addition, it has standard network interfaces and simple network protocols.

PNoC consists of a group of subsets; each subset contains a router that applies circuit switching between multiple nodes. Each node connects to a single router by a router port interface. The main components of PNoC router are shown in Figure 2.12.

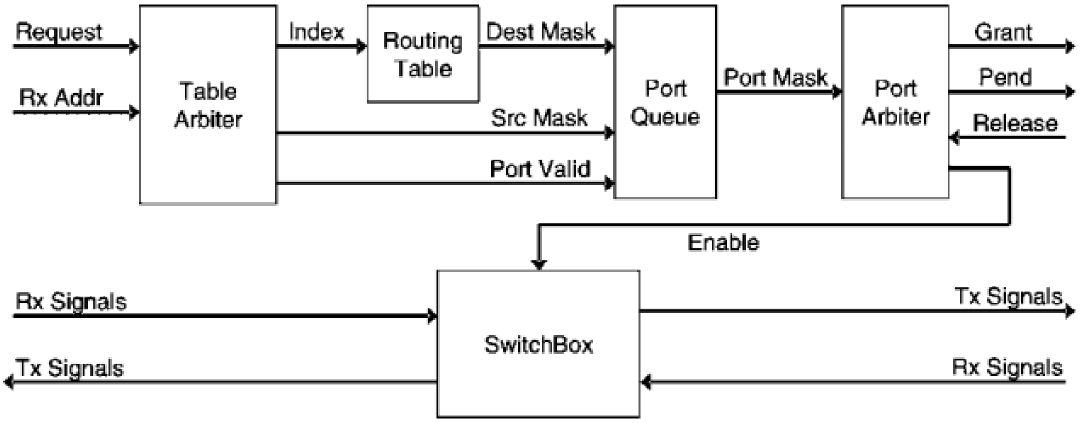


Figure 2.12: PNoC router block diagram [15]

PNoC components functionalities are as follows:

- Table arbiter: It receives multiple connection requests and schedules access to the routing table. In addition, it manages the routing table update requests.
- Routing table: It receives the required module address and uses it as an index that points to candidate ports.
- Port queue: It keeps the order of connection requests.
- Port arbiter: When the destination is free, the port arbiter establishes the desired connection and updates the signals that represent the status of connected ports for the flow control mechanism.
- Switch box: It forms the actual connections between modules.

One main difference between PNoC and the other architectures is that PNoC excludes the central crossbar (which consumes large area that affects the performance remarkably). Instead, it defines the connections by using distributed routers across the system; and sets up the router parameters which are the number of ports, data width and buffer depth.

Partial dynamic reconfiguration is taken into consideration in PNoC design. In case of adding a new module to the system, its local router is notified; which updates the routing table of the system. The same behavior is used when a module is removed.

Xilinx Virtex-II Pro FPGA (xcv2p30-7) is used to implement PNoC blocks. Table 2.2 shows the area and speed results for multiple configurations; different numbers of ports and different port data widths. One block RAM (BRAM) is used to implement the routing table. Note that the area of the routing table and the node interface hardware are not included in the results.

Table 2.2: PNoC router Implementation Results [15]

Number of ports	Data width	Area (Slices)	Frequency (MHz)
2	8	83	160
4	8	249	151
8	8	1113	138
2	32	131	145
4	32	366	138
8	32	1305	126

An image bit-serialization example is used to evaluate PNoC and two different bus-based implementations. The example uses an algorithm that quantizes grayscale image pixels to binary black and white values by computing median values at three hierarchical levels, then it uses them as quantization thresholds. Results show that; for concurrent data transfers applications, the performance of PNoC is similar to direct interconnects.

2.5.3. Dual Crossbar Router

R. Pau and N. Manjikian [16] attempt to implement a configurable router for an embedded network on chip using dual crossbar instead of one full crossbar. The router is implemented as a hard router to reduce the area.

The router has five bi-directional ports; a local port is used to establish the connection with associated node elements. On the other hand, the other four ports are used for different network topologies. The router uses a deterministic XY routing algorithm in which the first crossbar handles the X direction routing while the second crossbar handles the Y direction routing.

The router uses two 3x3 crossbars instead of one 5x5 crossbar; each one contains three bi-directional connections: Local, Left, and Right as shown in Figure 2.13.

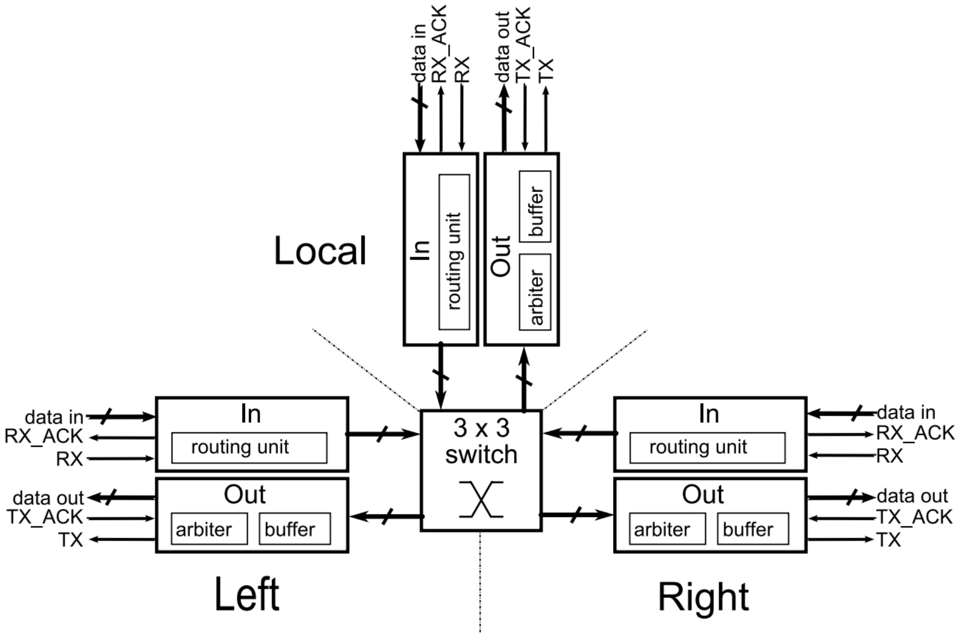


Figure 2.13: Configurable Router for Embedded NoC block diagram [16]

Routing of the packets is made as follows:

- Outgoing packets from the node element that is attached to the router pass locally through the first crossbar.
- Incoming packets that arrive through the North/South ports are switched directly to the attached node.
- Incoming packets that arrive through the East/West ports should first be switched to the second crossbar to reach the required node.

The router uses handshaking signals on each port to indicate the reception of a new packet from the neighbor routers.

The implementation is done on Altera Stratix FPGA using Altera Quartus v6.1 and ASIC TSMC 0.18 micrometer, Synopsys Design Compiler V-2004.06-SP1 and Cadence First Encounter v4.10.

The comparison between the dual crossbar and full crossbar with different interconnection widths is shown in Figure 2.14.

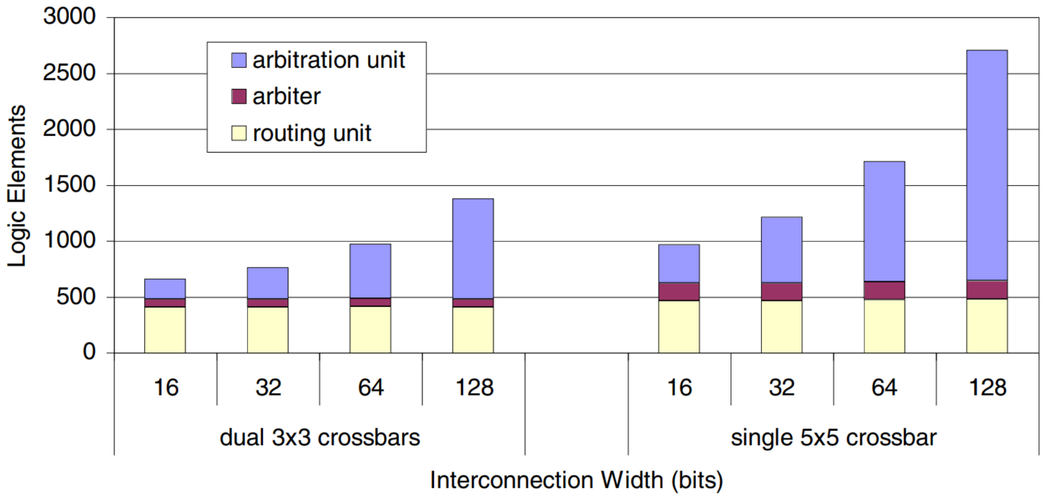


Figure 2.14: Configurable Router for Embedded NoC FPGA resource utilization breakdown [16]

The above results show that the dual crossbar is more area efficient due to the usage of fewer logic elements. However, it slows down the circuit as shown in Table 2.3.

Table 2.3: Configurable Router for Embedded NoC synthesis results for FPGA and ASIC [16]

	Altera Stratix	ASIC
Logic area reduction	24%	22%
Average operating frequency	123 MHz	340 MHz
Operating frequency reduction	19%	4%

2.5.4. HW NoC

K. Goossens, M. Bennebroek3, J. Y. Hur and M. A. Wahlah [17] compare HW NoC design to the conventional soft FPGA NoC. It is found that HW NoC has a better area, bandwidth and performance with a factor of 150 or more over the soft NoC.

NoC routers usually contain two components; routers that handle traversing packets in the network and network interfaces (NI) that translate the packets coming from/to NoC clients. A network interface is either a kernel or a shell. Kernels and shells are either hard or soft. One IP is attached to one or more NIs, such as functional IO as shown in Figure 2.15.

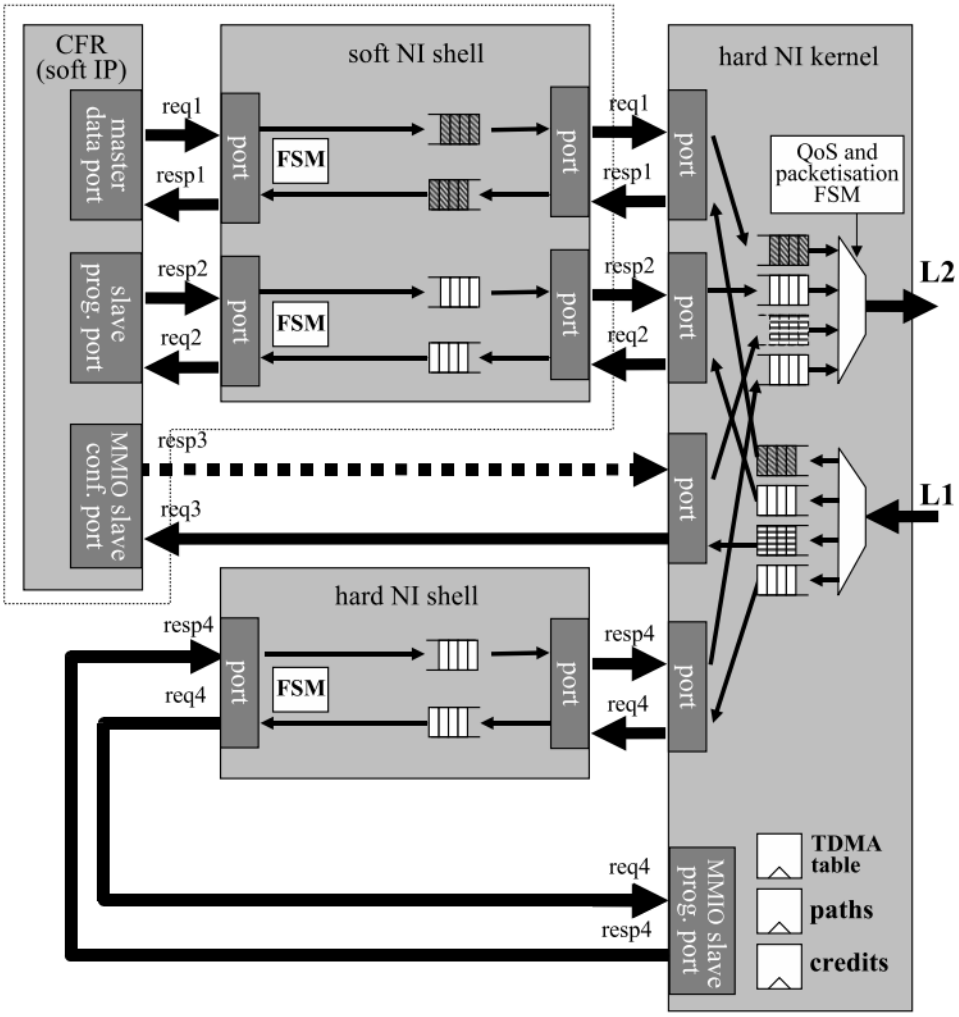


Figure 2.15: Hard and Soft NI Shell [17]

NoC routers are best implemented as hard due to large FPGA to ASIC overhead ratio of their arbiters and allocators.

The NI shell is soft for two reasons; first, the port protocol depends on the application IP which is different from one application to another. Second, the channel FIFO depth depends on the required bandwidth and latency which also differs from one application to another.

2.5.5. SOTA

Input buffers in SOTA [13] are implemented using dual-ported memory elements and they are organized as statically allocated multi-queue (SAMQ) so that the memory is shared between all VCs equally. Flit width and memory width have the same size to guarantee that writing and reading flits fit in one clock cycle. Each flit is routed in two phases using Valiant's routing algorithm to improve loading balance. First, the flit is routed to an intermediate node then it is routed to its destination.

A dimension-ordered routing algorithm is applied in each phase using two or three stages depending on whether the speculative switch allocation is successful or not. Flits are transferred from input nodes to output nodes via crossbar which is a 4x4 multiplexer. SOTA architecture is shown in Figure 2.16.

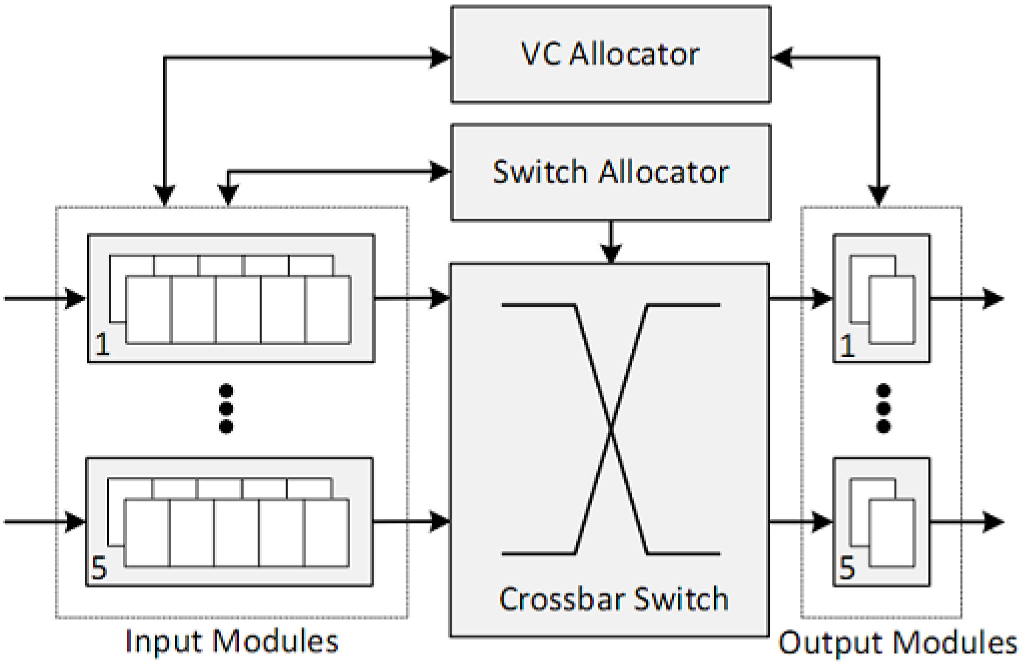


Figure 2.16: SOTA architecture [13]

2.5.6. CONNECT

In [18, 12], the authors introduced a soft router designed for FPGAs; CONNECT adds new features, such as virtual link and peak flow control. It maximizes routing resources utilization by using wider buses between routers.

It is an open source configurable RTL-based router designed for FPGA. Its architecture is shown in Figure 2.17.

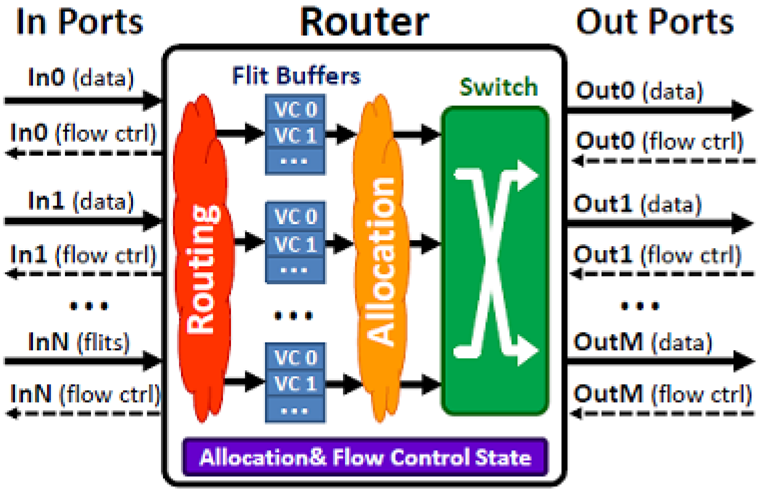


Figure 2.17: CONNECT router architecture [18]

Data is packetized while passing through the network; each packet is divided into multiple flits which include routing information along with the original packet data.

CONNECT supports two flow control mechanisms; credit-based flow control and a similar mechanism to the ON-OFF algorithm called peak flow control.

Four separable input-output allocation algorithms are supported in CONNECT. The router is configured with a different set of parameters which are the number of virtual channels, input ports and output ports, buffer depth, flit data width, network topology and flow control algorithms.

In addition to prioritizing flits using flow control credits, CONNECT introduces virtual links to guarantee that once a port starts receiving flits of a packet, it finishes before starting to handle another packet.

CONNECT is implemented using Bluespec System Verilog (BSV) with a design methodology that makes it flexible.

In [18], CONNECT is compared with SOTA [13] using Xilinx Virtex-6 LX240T and LX760 FPGAs. Regarding LUTs usage, CONNECT routers save about fifty percentages of equivalent SOTA routers as shown in Table 2.4.

Table 2.4: Synthesis Results for CONNECT and SOTA Mesh Network [18]

	Xilinx LX240T		Xilinx LX760	
4x4 Mesh w/ 4VCs	**%LUTs**	**MHz**	**%LUTs**	**MHz**
SOTA (32-bit)	36%	158	12%	181
CONNECT_32 (32-bit)	15%	101	5%	113
CONNECT_128 (128-bit)	36%	98	12%	113

2.5.7. Split and Merge PS

Y. Huan and A. DeHon [19] were interested in analyzing NoCs that are designed to target FPGA rather than ASIC. Their study compared two designs; the first design is CONNECT and the second is Split-Merged Packet Switched (PS) NoC which is shown in Figure 2.18. Their analysis results show that for different benchmarks, Split-Merged PS gives about three times higher frequency and throughput compared to CONNECT, but with the cost of using more area.

CONNECT uses only one single stage pipelining to reduce the effect of long wires delay. On the other hand, multiple stage pipelining is used in Split-Merge PS to get better results in performance and throughput.

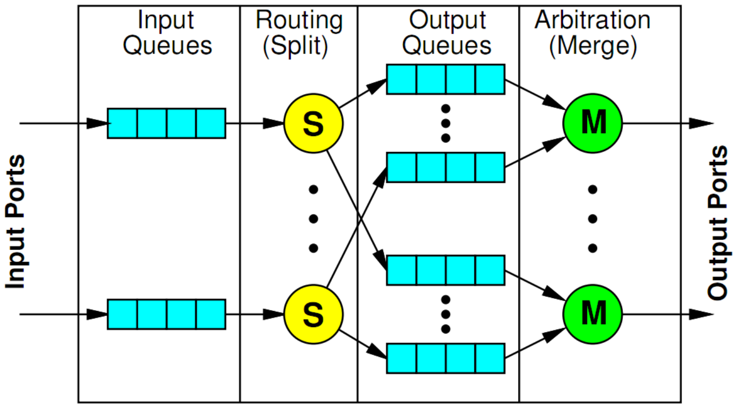


Figure 2.18: Split-Merge architecture [19]

The components of a Split-Merge router are as follow:

- Buffers: Implemented by shift registers as FIFO queues.
- Split primitive: Detects the flit header and routes input packets to the proper output port.
- Merge primitive: Receives and reconstructs packets coming from different input ports to a specific output port and sends them to that port.
- Flow control: Valid/backup pressure flow control is used, which is very similar to the peak flow control used in CONNECT.
- Routing algorithm: Two deadlock free algorithms are used:
  - Dimension ordered routing (DOR): Routes the packet along the X side then the Y dimension. However, this introduces long routes in some cases.
  - West-side first (WSF) routing: Offers more flexibility to avoid long routes in case of local congestion.

Using Xilinx Virtex 6 FPGA (XC240T-1), Split-Merge is compared with CONNECT. Mesh topology is used with flit width of 32 bits and buffer depth of 16. CONNECT is configured by peak flow control rather than credit-based flow control since peak flow control is similar to back pressure flow control used in Split-Merge. In addition, virtual links are activated in CONNECT to give the same functionality of Split-Merge.

According to the packet format of CONNECT and Split-Merge in Figure 2.19 and Figure 2.20 respectively, CONNECT adds 10 bits over Split-Merge for routing information, so a Split-Merge switch is tested with 42 bits channel width besides the 32 bits to reach a direct comparison with CONNECT.

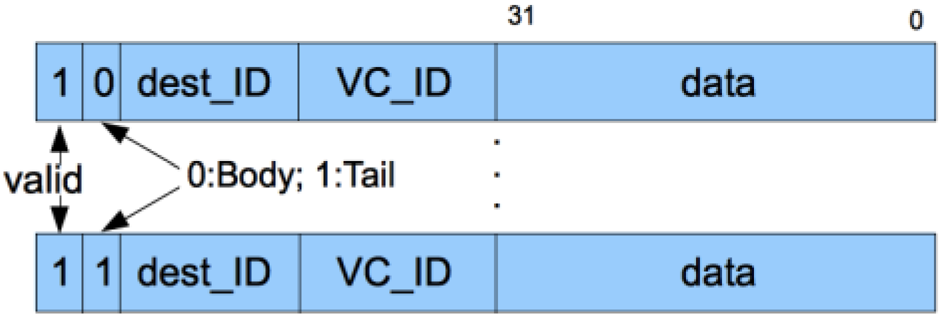


Figure 2.19: Packet Format of CONNECT Network [19]

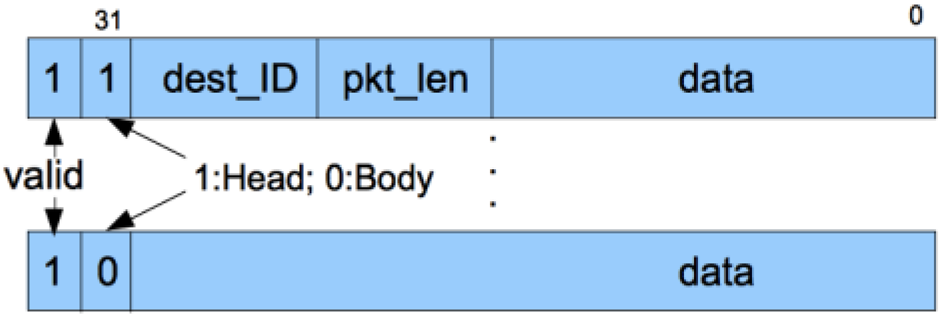


Figure 2.20: Packet Format of Split-Merge Network [19]

Results in Table 2.5 show that Split-Merge has the advantage of higher speed, but with the cost of more area consumption.

Table 2.5: Map & Post-PAR report for Split-Merge and CONNECT on XC6VLX240T-1 [19]

		Area			Timing		
			Logic	Mem.	Constrain	Cycle	Freq.
		Regs	(LUTs)		(ns)	(ns)	(MHz)
CONNECT	2VCs; 32bit	635	1396	166	9.0	9.6	104
1 clock	4VCs; 32bit	1265	1926	288	10.0	10.9	92
split-merge	DOR; 32bit	541	1449	336	4.5	4.5	220
1 pipe	DOR; 42bit	641	1686	462	4.5	4.6	219
(2 clocks)	WSF; 32bit	579	1839	400	4.6	4.6	217
	WSF; 42bit	679	2139	550	4.6	4.6	216

split-merge 2 pipes (4 clocks)	DOR; 32bit	1262	1157	336	3.3	3.3	303
	DOR; 42bit	1572	1302	462	5.0	5.0	201
	WSF; 32bit	1454	1491	400	3.3	3.4	298
	WSF; 42bit	1804	1666	550	4.7	4.7	213

Simulation results in Figure 2.21 show that under low congestion, CONNECT works with lower average delay. On the other hand, Split-Merge achieves higher performance under congested traffic.

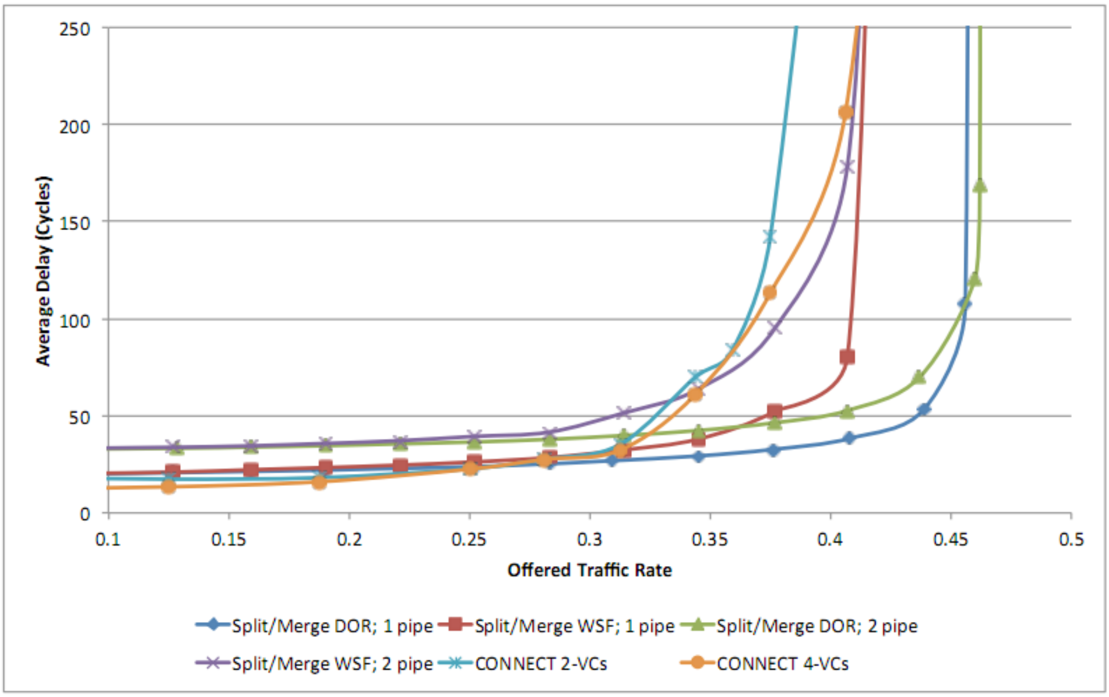


Figure 2.21: Cycle comparison between CONNECT and Split-Merge on uniform random traffic on an 8x8 mesh with eight flit packets [19]

2.5.8. FLNR

A. Imbewa and M. A. S. Khalid [20] introduced a fast lightweight NoC router designed for FPGAs with the objectives of minimizing resource consumption and improving the performance.

The packet has been modified to minimize the control fields by removing the control fields from its body and removing the tail flit as shown in Figure 2.22. This approach yields the reduction of FIFO width, buffer area and power consumption.

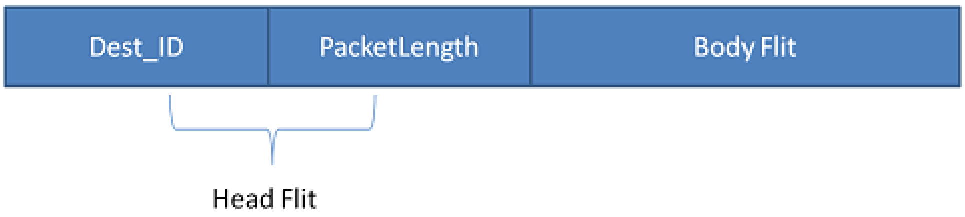


Figure 2.22: FLNR packet format [20]

In FLNR design, the router decision time is only one clock cycle; it takes one clock cycle to write the body flits since credit-based flow control is used.

As shown in Figure 2.23, each router is connected to its neighbor routers (North, East, South, and West) and to the local IP core as well.

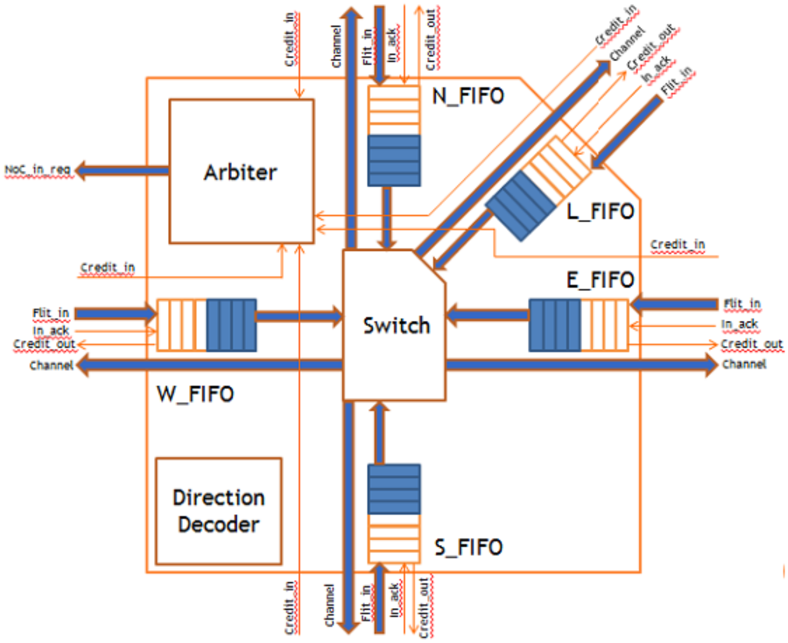


Figure 2.23: FLNR block diagram [20]

FLNR components and their functionalists:

- Arbiter: It receives the notifications (flit headers that contain packet information including destination address) coming from input ports and routes to north, east, south, west or local direction using round robin arbitration. In addition, it detects the head flit and payload end.
- Direction decoder: It receives the destination address of the packet and calculates the routing directions using XY routing (the cheapest schema to have deadlock-free network).
- FIFO depth: The minimum depth is the number of possible flits that are stored during routing decision time. If there is no blocking, only two buffers (one for head flit, one for body flit) are enough to get the minimum latency.

- Switch: Finally the switch assigns the incoming packets from input ports to available channels. The switch is a five five–to-one multiplexers that support all possible connections between input and output buffers.

FLNR is implemented on Altera Stratix II EP2S15F672I4 FPGA. The synthesis results for FLNR with three hops and buffer size of eight flits are shown in Figure 2.24.

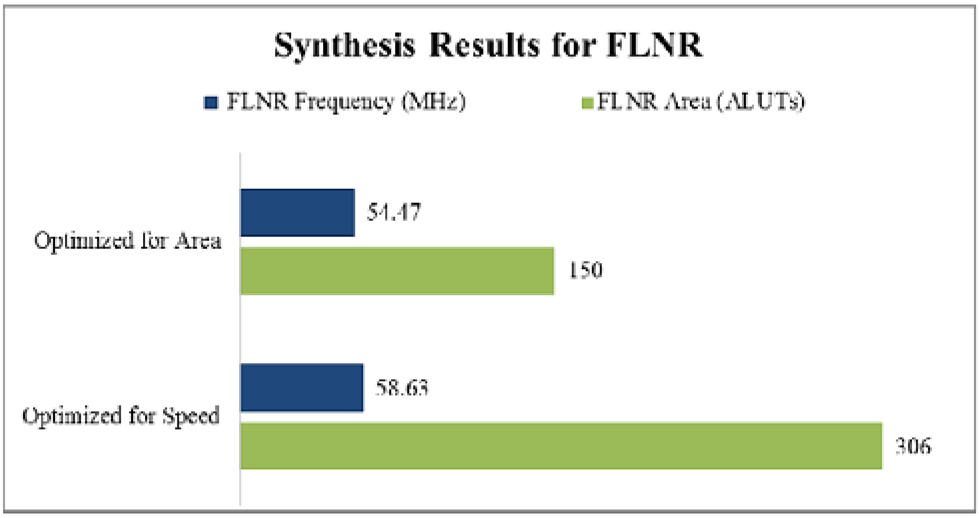


Figure 2.24: Synthesis results for FLNR [20]

The comparison with other routers (HERMES [21], ICN [22] and Bartic [23]) is made by calculating the port bandwidth (maximum throughput) for each design, then calculating the best case latency based on the same case study. Figure 2.25 and Table 2.6 give the comparison results; FLNR significantly outperforms the other routers with a lower area, latency and higher frequency. Furthermore, the number of clock cycles consumed to finish the routing decision (R_d) is only one cycle.

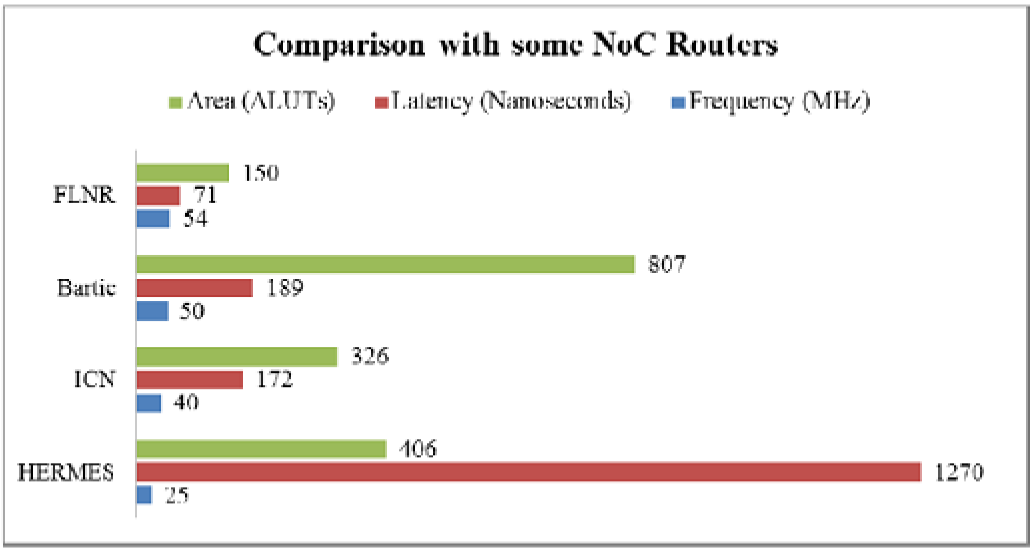


Figure 2.25: FLNR performance and area comparison with previous routers [20]

Table 2.6: FLNR performance and area comparison with some previous NoC routers [20]

Design	Flit size	Flit/Cycle	Slices	Frequency (MHz)
HERMES [21]	8	0.5	406	25
ICN [22]	16	0.5	326	40
Batric [23]	16	1	807	50
FLNR	8	1	150	54

2.5.9. RROCN

HY. Luo, SJ. Wei, and DH. Guo [24] introduced an on-chip network with regular reconfigurable topology (RROCN) which contains both routed network and shared bus. The network disables and bypasses the unwanted nodes; this leads to a suitable throughput and power consumption for application with different bandwidth demands. The primary goal of RROCN is to provide a reconfigurable suitable NoC with low cost.

RROCN architecture consists of several nodes; each one contains a router, a CPU core is attached to the network through the local port of the router while the peripherals are located around the network which gives an NxN 2D mesh topology as the largest topology that RROCN constructs with different MxH shapes but should be less than N.

The main components of RROCN router are shown in Figure 2.26.

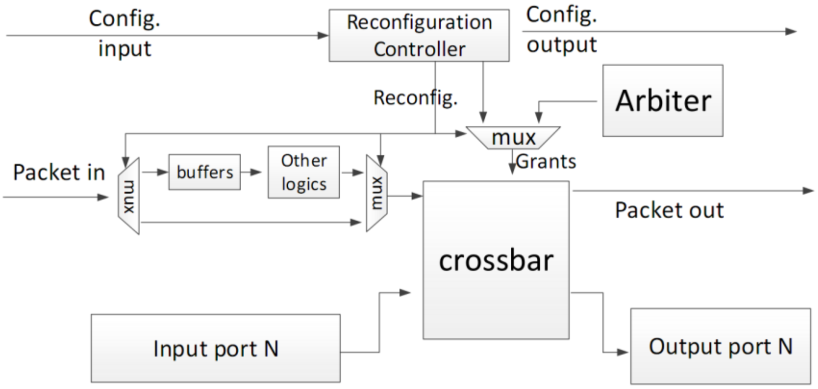


Figure 2.26: RRCON router block diagram [24]

PRCON components functionalities:

- Reconfiguration controller: Configures the crossbar and the multiplexers using the information received from the previous router; after that it generates new configuration information which is passed to the next router.
- Crossbar: Responsible for connecting input ports to output ports. It consists of five ports; one for local port and the others are for the processor and the peripheral group as shown in Figure 2.27.
- Arbiter: Handles only the requests from the peripherals group and constructs the connections using priorities included in the configuration information.

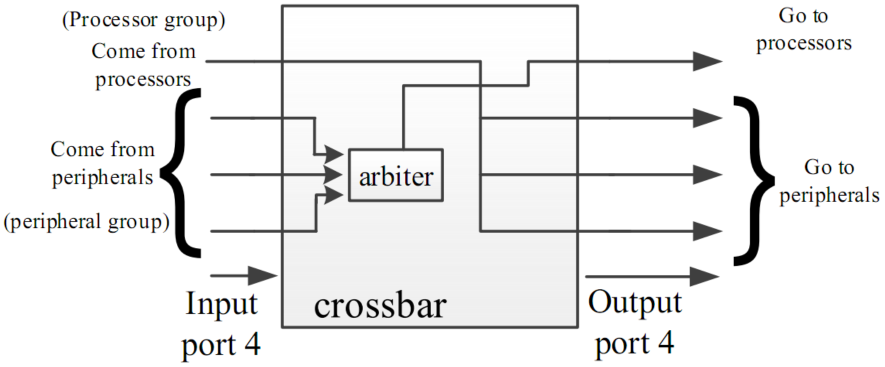


Figure 2.27: RRCON crossbar architecture [24]

The reconfiguration process starts at the runtime from the processor by first selecting an original node to be the starting point of the network, and then the configuration information spreads inside the network to reach each node using reconfiguration controllers in each node by using YX constructive algorithm. After constructing the network, a modified self-adaptive XY routing algorithm is used.

2.6. Comparative Review

In this section, three open-source NoC designs from [13, 14, 18] are used to make a comparison. The comparison investigates the effects of changing the buffer depth, data width and number of VCs on both the maximum operating frequency and FPGA resource utilization. The comparison results helps to select the suitable NoC parameters according to system requirements.

2.6.1. Comparison workflow

The comparison is held between the three architectures across different numbers of VCs, data width and buffer depth to analyze their effects on frequency, LUTs and registers usage. A fixed 4x4 mesh topology is used for the comparison, in which each router has five input/output ports.

Xilinx ISE v14.4 is used as a synthesis tool and Virtex6 XC6VLX240T FPGA as a target. During synthesis, RAM extraction option is disabled to guarantee fairness among all three routers as it is noticed that their on-chip memory utilization differs.

2.6.2. Frequency

For buffer depth; increasing buffer depth improves the overall network performance and reduces the congestions. Consequently, this adds extra logic that decreases the operating frequency slightly. Figure 2.28 shows that NoCem has the highest operating frequency and at the same time, it is the most sensitive router to buffer depth changes.

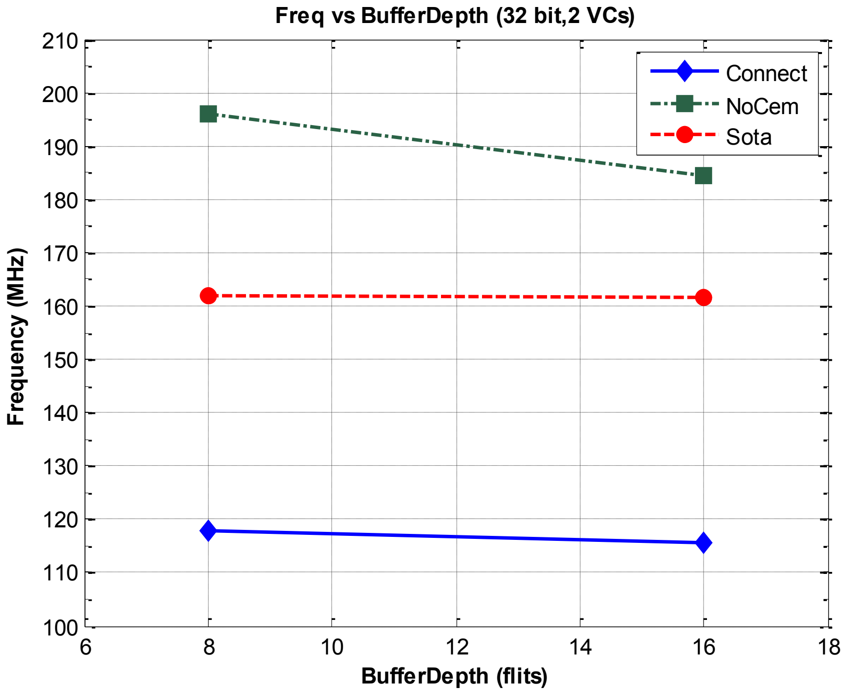


Figure 2.28: Frequency vs. buffer depth

Data width increase does not have a high impact on the operating frequency of the three routers as it does not affect the allocators or arbiters. As shown in Figure 2.29, CONNECT is the most sensitive router to this parameter, whereas SOTA operating frequency is almost fixed. NoCem has the highest operating frequency for all data width values.

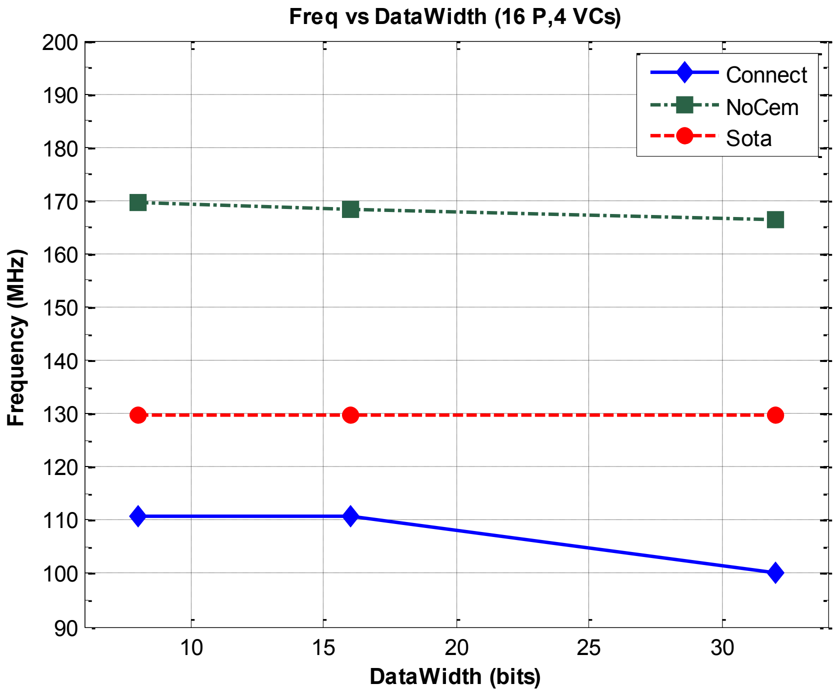


Figure 2.29: Frequency vs. data width

For changing the number of VCs; as shown in Figure 2.30, increasing VCs decreases the operating frequency for all routers because adding VCs leads to more combinational delays due to the extra logic introduced in allocators and arbiters. NoCem has the highest operating frequency. However, it supports only up to four VC. SOTA is the most sensitive router to VCs increase.

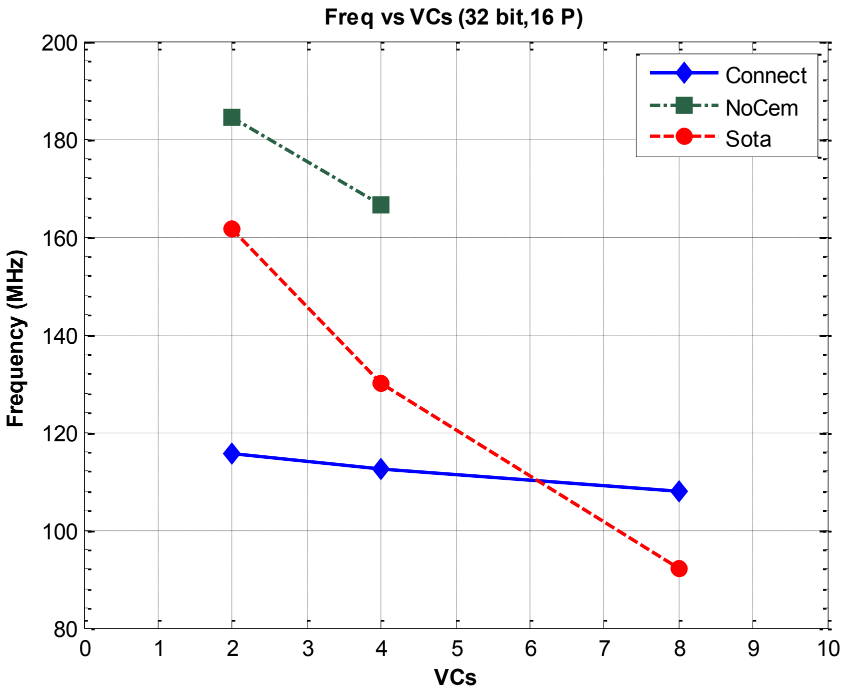


Figure 2.30: Frequency vs. VCs

2.6.3. LUTs usage

For buffer depth; as shown in Figure 2.31 and for almost all buffer depths, SOTA consumes the least amount of LUTs, whereas NoCem consumes the most.

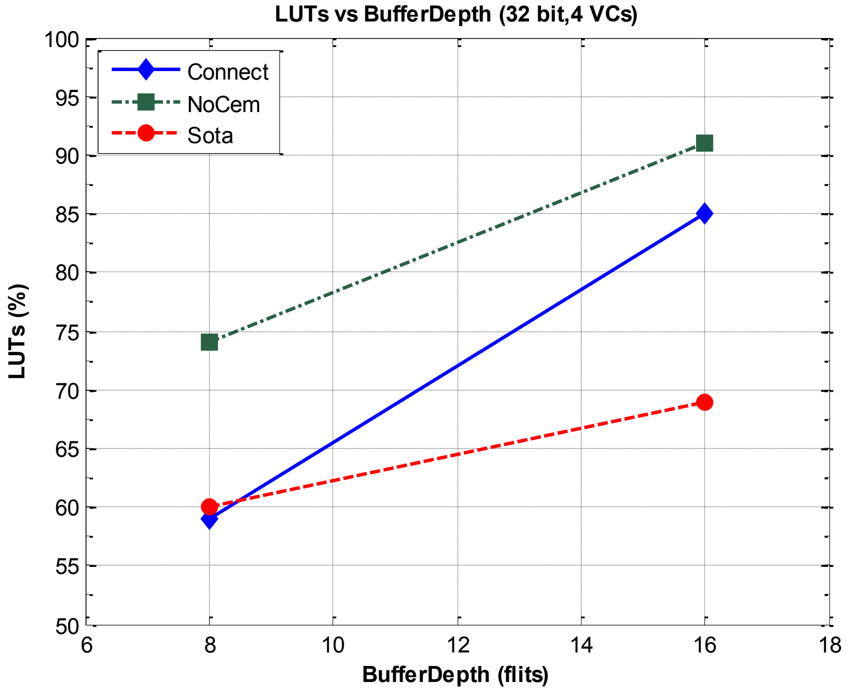


Figure 2.31: LUTs utilization vs. buffer depth

For data width; as shown in Figure 2.32, for 8 and 16 bits data width, NoCem is the most efficient in LUTs consumption, whereas it consumes the most significant number of LUTs for 32-bits data width.

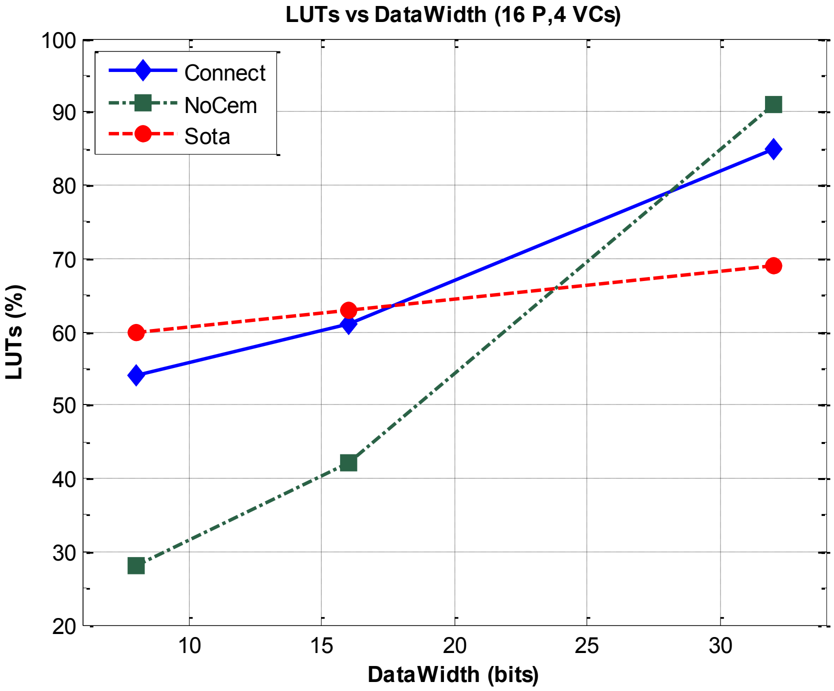


Figure 2.32: LUTs utilization vs. data width

For changing the number of VCs; adding VCs introduces more logic for routing computation which increases LUTs consumption. Figure 2.33 shows that NoCem consumes more LUTs than SOTA and CONNECT for all the number of VCs that it supports. CONNECT consumes the least amount of LUTs.

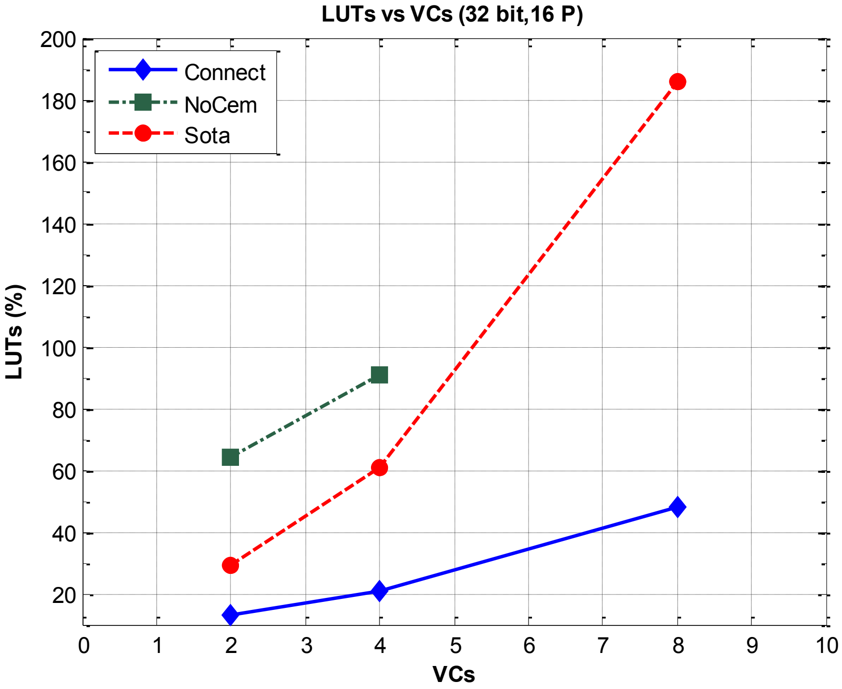


Figure 2.33: LUTs utilization vs. VCs

2.6.4. Registers usage

More memory elements are needed if the three parameters (buffer depth, data width and the number of VCs) are increased. As shown in Figures 2.34, 2.35 and 2.36, SOTA is the most efficient in registers utilization, whereas NoCem consumes the most significant number of registers.

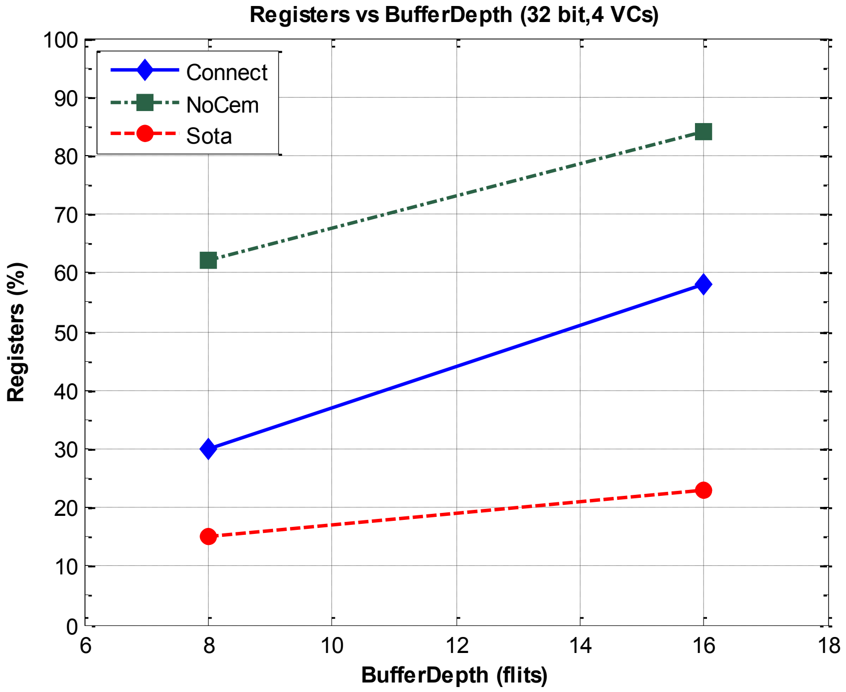


Figure 2.34: Registers utilization vs. buffer depth

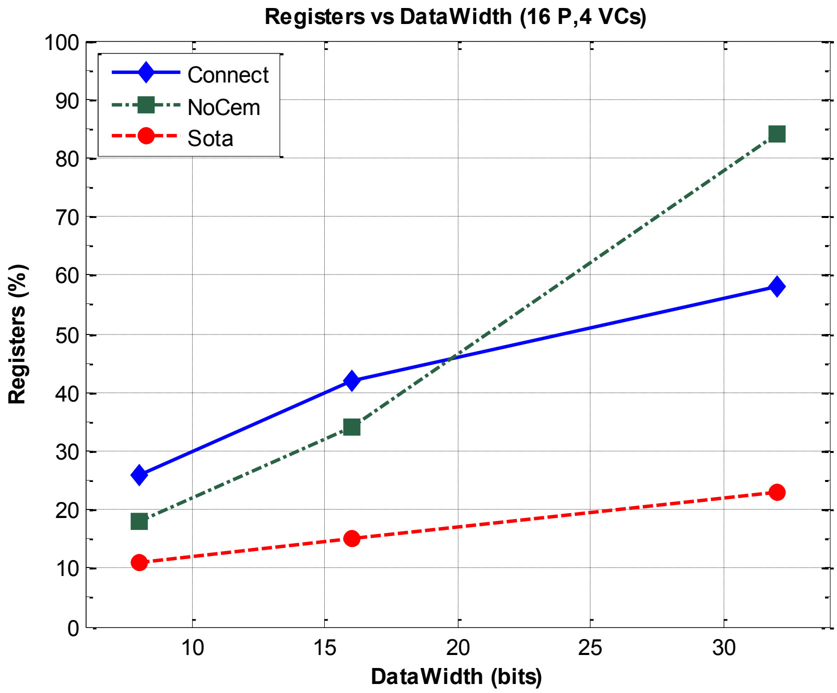


Figure 2.35: Registers utilization vs. data width

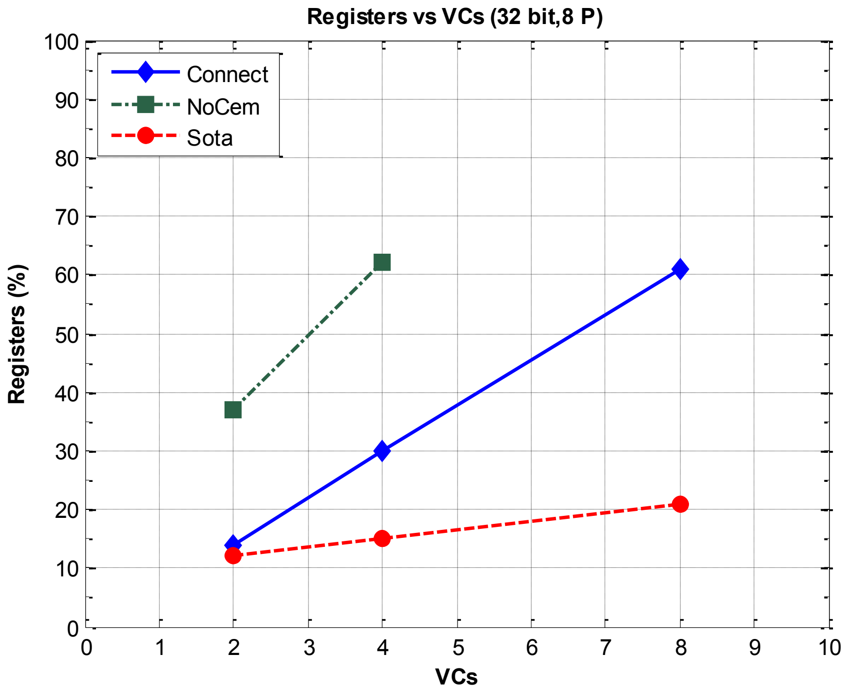


Figure 2.36: Registers utilization vs. VCs

2.7. Summary and future work

PNoC [15] is a circuit-switched approach applied to FPGA-based systems. It provides a flexible, lightweight and easy design. Its performance is similar to direct interconnects.

PNoC design is used for partial dynamic reconfiguration by updating the routing table with the added and removed modules. On the other hand, it is not suitable for applications subjected to conflicting flows since it is similar to a circuit-switched system; once its connections are established, no other modules can communicate.

Its future work includes:

- Use of multiple routers, topologies and subnets in a system.
- Perform a detailed comparison with packet-switched NoCs.
- Apply more tests to check its suitability for partial dynamic reconfiguration.

The configurable router in [16] provides flexibility in supporting a variety of network topologies with a simple three-bit input configuration. A dual crossbar arrangement has a lower area with some reduction in the operating frequency.

The router configuration can be improved by including:

- Virtual channels to achieve higher throughput under high traffic congestion.
- Using the concept of middle-buffering to achieve smaller designs and superior performance than output buffering.
- To use custom memory blocks for buffer implementation.

In [19], a detailed comparison between Split-Merged PS approach and CONNECT has been introduced using different sets of benchmarks. Results show that Split-Merged PS system reaches up to 300 MHz which is three times higher than CONNECT but with an increase in the area usage.

FLNR [20] is a NoC router for FPGA that minimizes the area and provides good performance by minimizing the control fields in the packets to decrease the buffer width. In addition, it decreases the routing decision time and can deliver one flit each one clock cycle. Future work is to implement a dual-clock wormhole router to forward the body flits at a higher frequency than the head flits.

In addition, the authors should consider comprising FLNR results with more recent NoC approaches, e.g., CONNECT and SOTA. FLNR is not included in the comparative review because its design is not open source.

RROCN [24] is proposed for chip-multiprocessors to achieve lower power consumption for a certain throughput. RROCN is evaluated with four specific reconfiguration topologies and compared with HCS network. RROCN is suitable for specific applications, for example, an application with specific throughput demand, the RROCN is configured with a topology that provides suitable throughput with less power consumption and lower zero-load latency, the same thing happens for an application that requires lower latency or less power consumption.

The reconfiguration process is used to compromise between throughput, latency and power consumption, or it can be used to optimize for one of them.

Future work is to improve the router design to include other network topologies other than mesh topology and to use virtual channels to increase the maximum throughput.

In addition to the literature review, a comparison is provided between three NoCs to investigate the effects of changing the buffer depth, data width and number of VCs on both the maximum operating frequency and FPGA resource utilization. The comparison results help to select the suitable NoC parameters according to system requirements:

- If the system requires a high operating frequency; NoCem is the best choice, it comes with the cost of more LUTs if used with a bigger buffer depth or a higher number of VCs.
- For networks with small numbers of VCs; CONNECT is the most efficient in LUTs consumption. On the other hand, it has the lowest operating frequency across all NoC parameters.
- If the target is to introduce QoS to the system (by increasing the number of VCs); SOTA is the most suitable router. As with the increased data width, buffer depth or VCs, it consumes the least amount of registers.
- If the target is to get a high data rate while using SOTA; it is more suitable to increase its data width since it doesn't support operating with high frequencies.

Chapter 3 : Codec, Tiles to Router Interface

3.1. Introduction

FPGAs are facing a big challenge, which is the area, power and delay overhead of its routing network. A medium-size FPGA design includes multiple IPs and modules that are placed at different locations across the FPGA; short wires are used to connect nodes within the same block or module and long wires (global wires) are used to connect modules far apart. Increasing the number of blocks in an FPGA requires more long wires which introduce large delay and consume more power. Figure 3.1 shows different metal layers in a modern ASIC device, in which global wires are thicker and wider leading to higher resistance and capacitance.

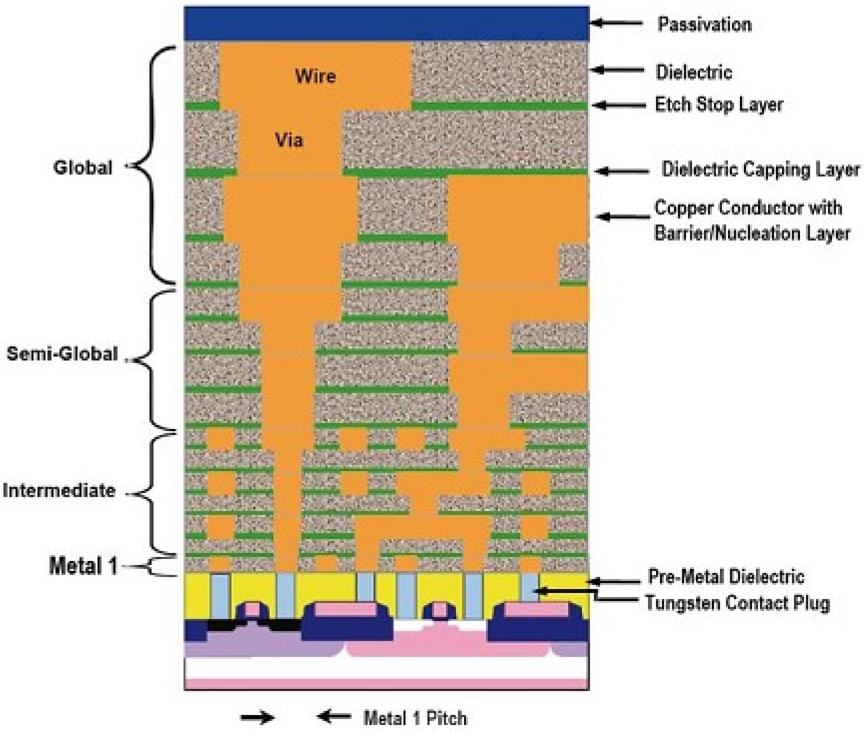


Figure 3.1: Metal layers in modern ASIC devices [9]

Network on chip (NoC) is widely used for complex SoCs as it overcomes the conventional interconnect problems of high power consumption and latency. NoC has the benefits of independent implementation, simplified customization, scalability and support for different network topologies. PNoC [15], SOTA [13] and Split-Merge [19] are NoC router examples.

NoC routers are implemented either as hard or soft, a hard router is more area and power efficient than a soft router. However, the latter is more flexible and configurable. Design guidelines for soft and hard routers are introduced in [25, 26], the tradeoffs between soft and hard routers are discussed in [27, 28].

Using the NoC approach instead of long interconnect wires solves some of the conventional interconnects problems; because NoC uses high-speed optimized lanes to transfer packets between routers. The routers interface with the application blocks through a configurable number of input/output ports solving most of the problems introduced by long and medium-size routing wires.

By applying the NoC approach to FPGAs, a major problem appears because of the large number of tiles that need to be connected to the network. As shown in [29], the area, delay and power consumption of the NoC routers are increasing significantly if the port count is increased. In order to overcome this problem and to make the NoC approach useful in designing the next generation of FPGAs, a tile to NoC router (or a *Codec*) is used to enable multiple tiles to share a single router port, given that the required rate for all multiplexed tiles does not exceed the maximum rate of a router port.

3.2. Modeling and simulation

A Simulink system-level model is built using SimEvents toolbox to measure the throughput difference between two networks, Network A which includes routers only, and Network B which includes routers and Codec modules. SimEvents provides a discrete-event simulation engine and component library for analyzing event-driven system models and optimizing performance characteristics such as latency, throughput, and packet loss.

A 2x2 mesh network with sixteen tiles is used for both networks; each router is connected to four tiles either by direct port connections or through a Codec. As network A does not use Codec, each router interfaces with four tiles and two neighbor routers (6-port router shown in Figure 3.2).

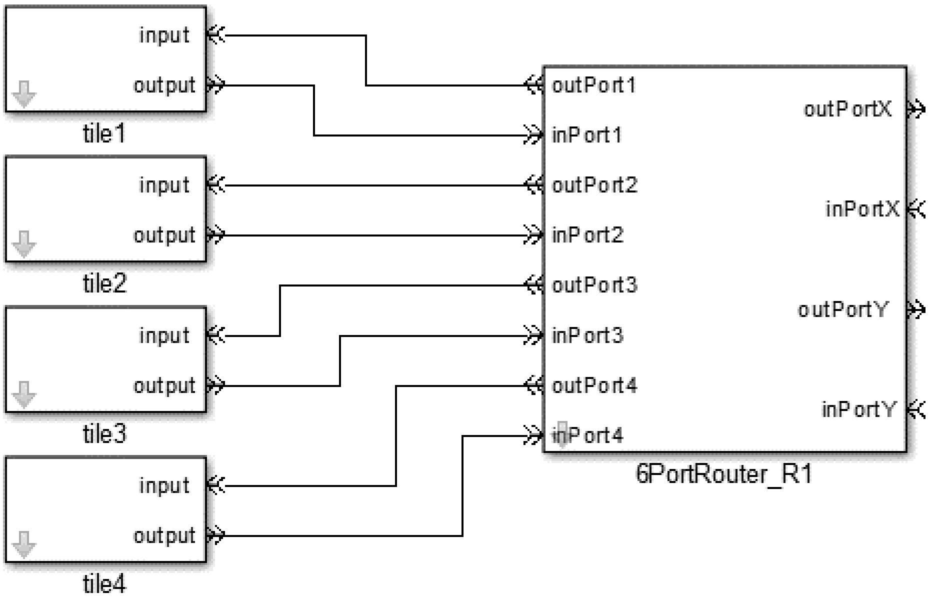


Figure 3.2: 6-port router in network A

In network B, each router uses a Codec module to interface with four tiles so the used routers are 3-port routers (shown in Figure 3.3). The packet length is increased by two bits in this case to handle the switching required from Codec to tiles.

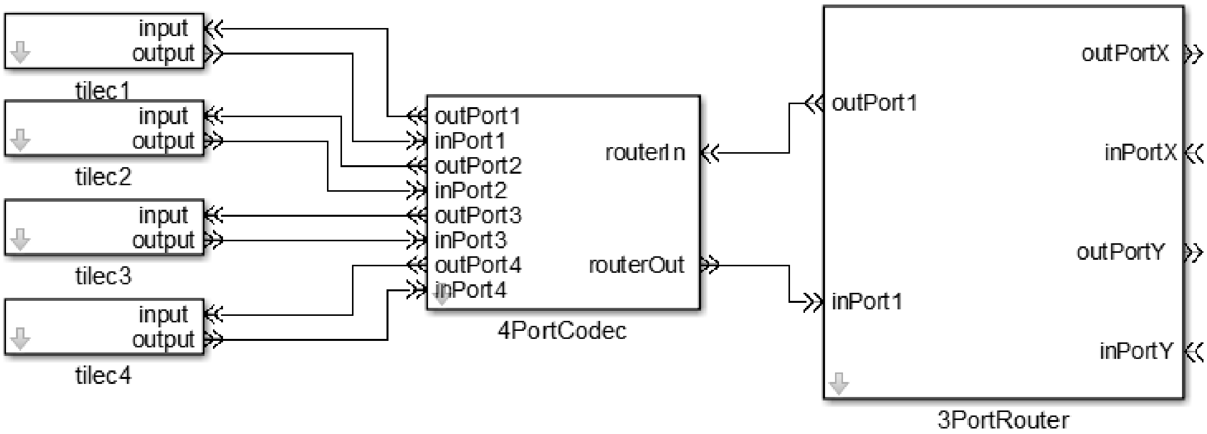


Figure 3.3: 3-port router in network B

A router consists of routing core and input/output queues. The routing core (shown in Figure 3.4) consists of routing logic and output switch; its routing logic is implemented as delay server to model the packet processing latency and a routing table to determine which output port the packet goes to. Packet processing time in a 6-port router is assumed to be longer than in a 3-port router.

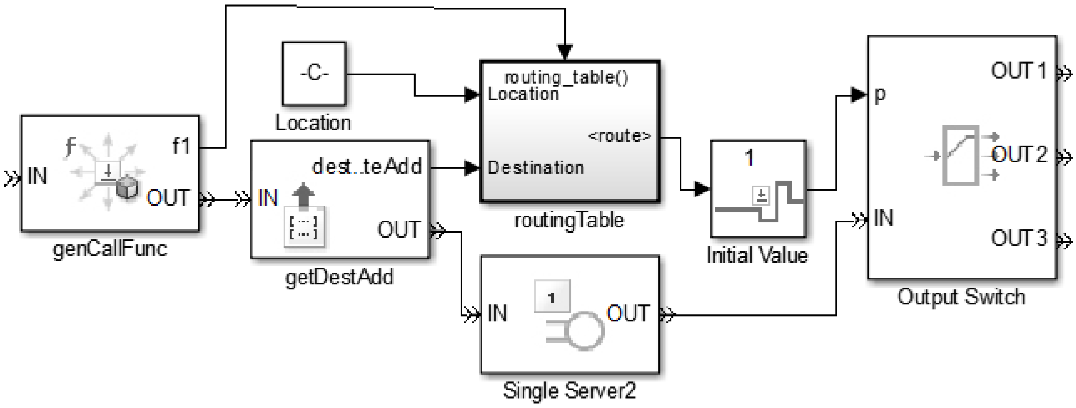


Figure 3.4: Router core of a 3-port router

Figure 3.5 shows a Codec, in the sending path, where Codec acts as multiplexer or path combiner and in the receiving path, it is similar to a regular router with shorter server time.

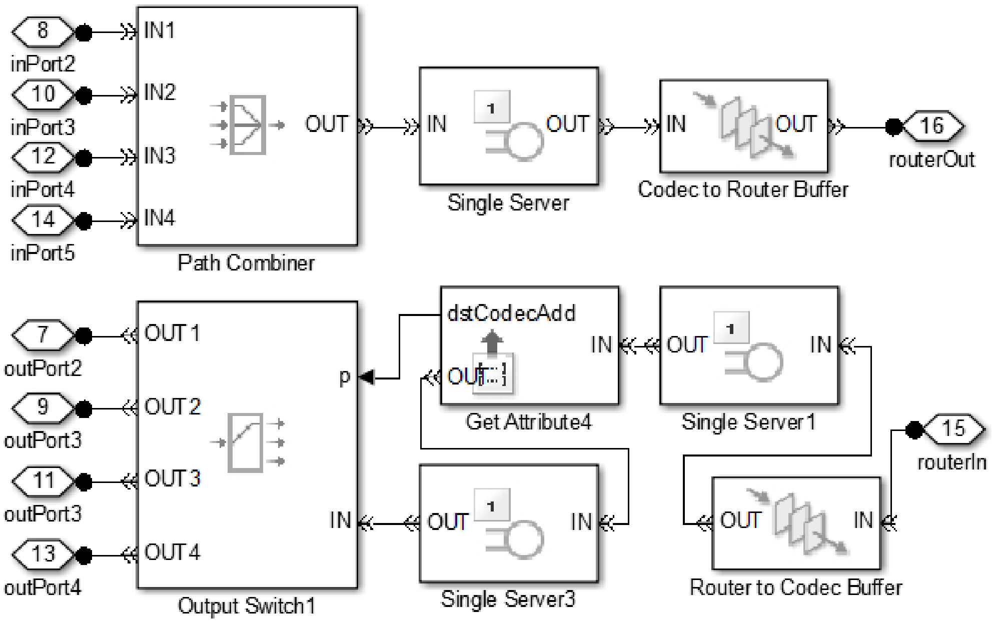


Figure 3.5: Codec Simulink model

3.2.1. Simulation results

The two network models are run for the same time with the following parameters; router server time is set to three for 3-port routers, six for 6-port routers and one for Codec modules and buffer depth is set to four packets. A uniform distribution packet generator is used at each tile output to simulate the existence of FPGA tiles.

After running the simulation, the number of received packets at each tile is counted. Table 3.1 shows the comparison between the numbers of received packets at some tiles in network A compared to network B. The total number of received packets is 281 packets in network A and 677 packets in network B. The throughput is enhanced by using Codec in network B.

Table 3.1: Simulink models comparison

Tile number	1	2	5	6	9	10	13	14	All tiles
Network A	18	19	12	17	21	17	16	18	281
Network B	54	44	43	41	47	39	45	39	677

3.3. Implementation and network synthesis

3.3.1. The Codec module

The sending path of Codec (from tile to a router) acts as a path combiner that rotates across all attached tiles and checks for available payloads to send.

As lower part of Figure 3.5, the receiving path is similar to a router; a packet is examined once received from a router and sent to a tile according to its destination address. Codec design is made modular to facilitate generating any network configuration with different tile widths and count. In addition, it inherits some parameter from CONNECT design for compatibility given that CONNECT router is dedicated to the embedded NoC in the next generation FPGA.

3.3.2. Network configurations

Similar to the Simulink comparison mentioned previously in this chapter, an RTL comparison is held between two network models both have four routers in a 2x2 mesh topology. The first model is network A; it uses CONNECT routers only, three configurations of this model are built to interface with 16-tiles, 32-tiles and 54-tiles. A 64-tiles network could not be built as CONNECT generation tool [12] is limited to 16-port per router; two ports out of the 16 are used to connect with neighbor routers and 14 are left to interface with the tiles, which gives a total of 54 ports for the four routers inside the network.

The second model is network B; it uses Codec to interface with 16-tiles, 32-tiles and 64-tiles. For each group, three configurations are built for a different number of Codec modules per router, Table 3.2 illustrates the Codec network configurations; CpR is the Codecs per router and TpC is the tiles per Codec. For example, to build a 2x2 16-tiles configuration of network B, each router connects with one Codec (1 CpR) and each Codec connects to four tiles (4 TpC), or each router connects to two Codecs (2 CpR) and each Codec connects to two tiles (2 TpC).

Table 3.2: Network B CpR and TpC configurations

Configuration\Tiles	16 Tiles	32 Tiles	64 Tiles
1CpR	4TpC	8TpC	16TpC
2CpR	2TpC	4TpC	8TpC
4CpR	-	2TpC	4TpC

3.4. Comparison results

Altera Arria II GX EP2AGX260 FPGA is used as a target chip to compare synthesis results. It has 205200 combinational ALUTs, 102600 memory ALUTs and 692 IO pins. Quartus II 12.0 is used with ModelSim Altera Starter Edition for synthesis and RTL simulation.

The logic utilization values shown in the following figures are the sum of both consumed combinational and memory resources. Quartus PowerPlay Analyzer tool is used to estimate the consumed power.

3.4.1. Frequency

As shown in Figure 3.6, the maximum operating frequency of network A decreases with increasing the number of tiles. On the other hand, network B starts at higher frequency and decreases slightly as the number of tiles increases.

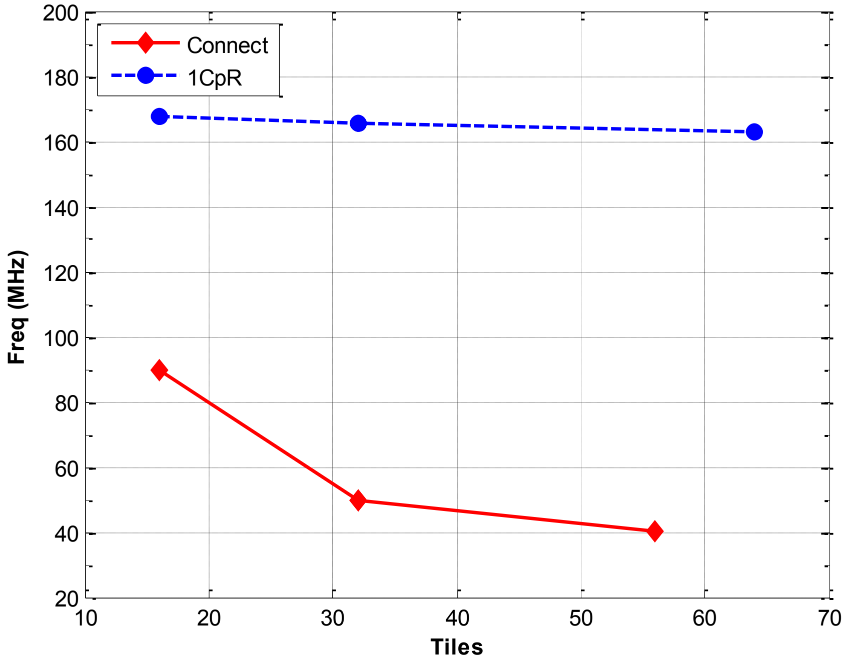


Figure 3.6: Maximum operating frequency comparison between network A and network B with 1CpR

The reason for the difference between network B configurations shown in Figure 3.7 is not the change in the Codec circuit. The reason for this difference is mostly the increased size of CONNECT routers. A 3-port CONNECT router used in 1CpR and a 4-port router is used in 2CpR network; the 4-port router occupies more area and operates with a lower frequency compared to the 3-port router.

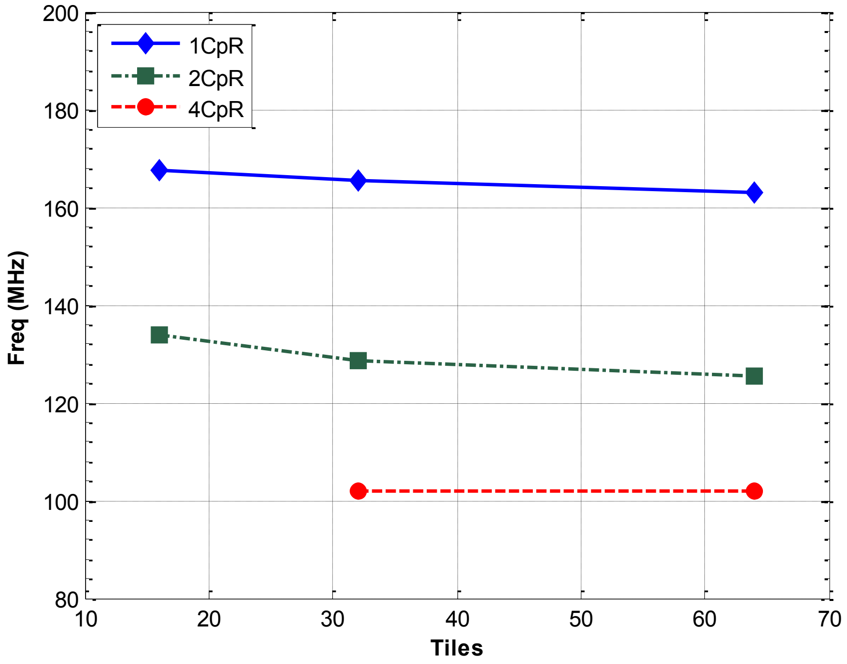


Figure 3.7: Maximum operating frequency comparison between different network B configurations

3.4.2. Logic utilization

As displayed in Figure 3.8, a 56-tiles network A uses 30% of the FPGA resources. In this network, each of the four routers has sixteen ports in order to be able to interface with 14 tiles. However, a 64-tiles network B with 1CpR uses at maximum 2% of the resources as each of the four routers has only three ports; two ports to interface with adjacent routers and one to interface with the Codec which connects to sixteen tiles.

The reason for this large logic utilization difference is that a 16-port CONNECT router consumes larger area than a 16-port Codec.

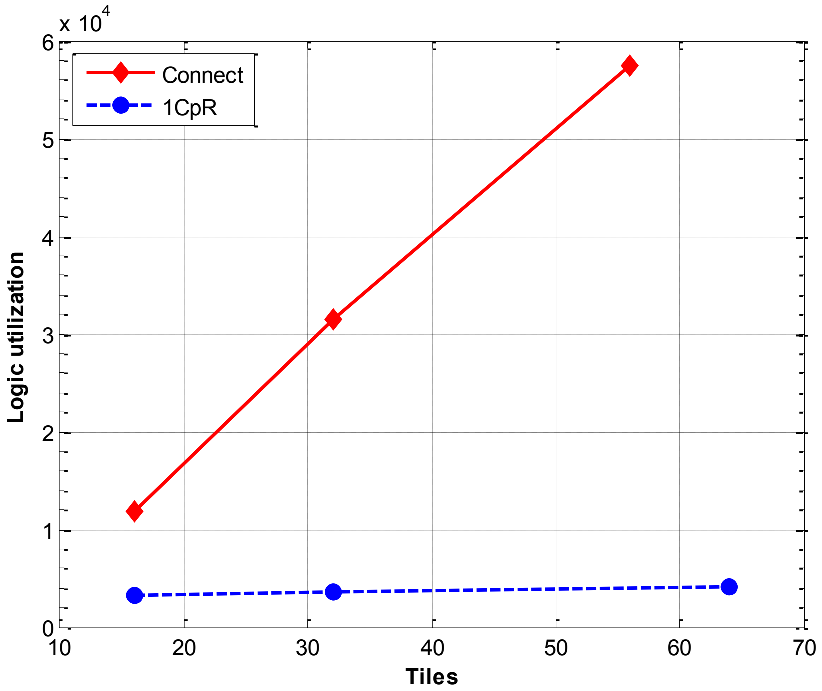


Figure 3.8: Logic utilization comparison between network A and network B with 1CpR

In Figure 3.9, a comparison between 1CpR, 2CpR and 4CpR configurations is shown. The 4CpR configuration consume the largest area because each router has six ports; two to connect with adjacent routers and four to connect with four Codec modules. 2CpR network configuration has a 4-port router and 1CpR has a 3-port router.

The number of Codec ports increases as the number of connected tiles increases; this explains the slight increase of logic utilization between different network B configurations.

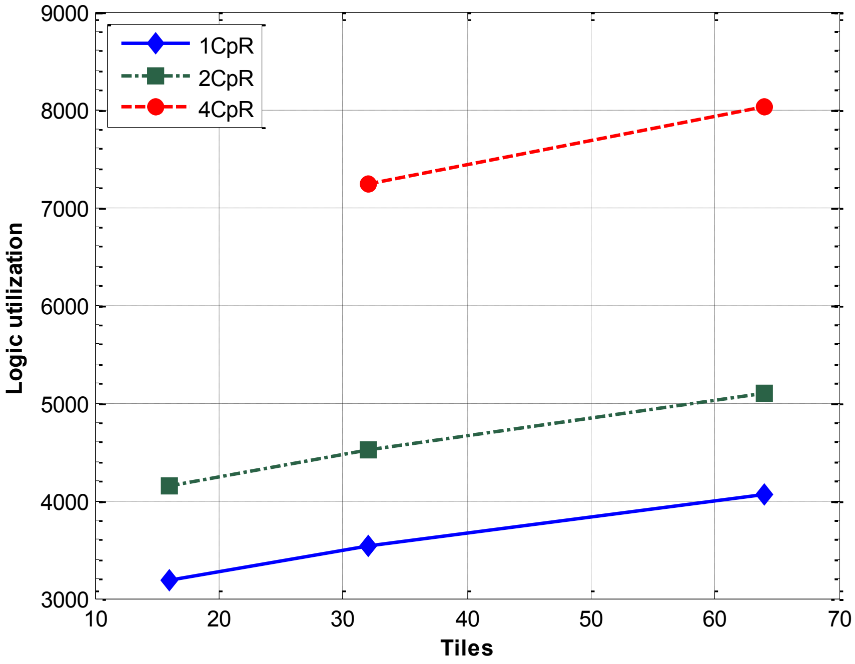


Figure 3.9: Logic utilization comparison between different network B configurations

3.4.3. Power consumption

As shown in Figure 3.10 and Figure 3.11, network A consumes more power than all network B configurations because of the large area consumed by CONNECT 6-port routers. In addition, it is shown that increasing the input/output port count affects the power consumption of network A more significantly than network B.

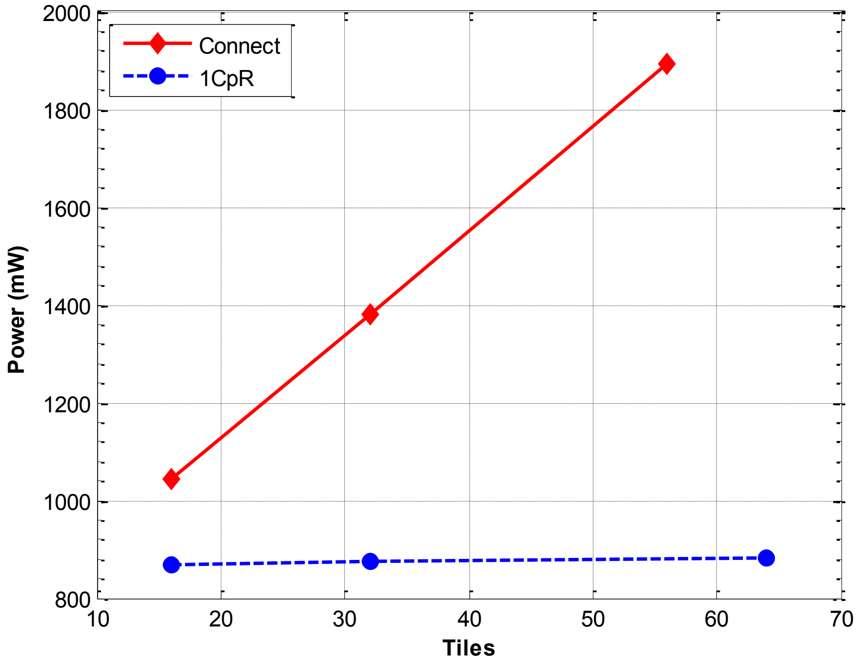


Figure 3.10: Power consumption comparison between network A and network B with 1CpR

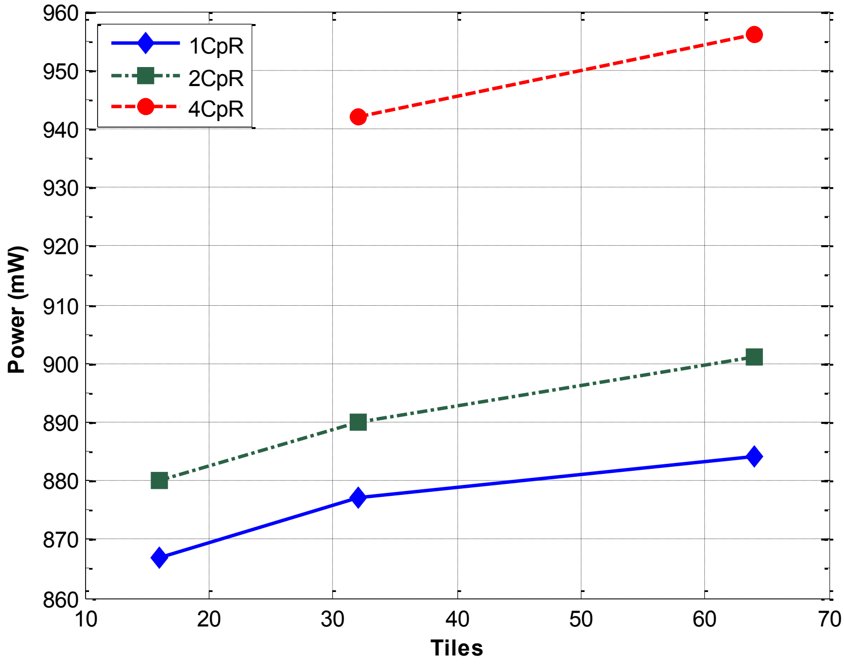


Figure 3.11: Power consumption comparison between different network B configurations

3.5. Summary

Hard NoC can be used in next-generation FPGAs to overcome the problems of conventional long wires. However, using NoC to connect the FPGA's large number of tiles and blocks causes a big problem, which is the effect of increasing the router input/output ports on its area, power and operating frequency.

Instead of increasing the router port count in order to interface with the increased number of tiles, a Codec module is used. It concatenates many tiles together and interfaces them with one router port. The Codec is a time division multiplexing block that is simpler than a NoC router. Therefore, its area and power does not scale significantly with the increased number of tiles and its frequency is not greatly decreased like a NoC router.

When comparing two 2x2 networks, one uses routers only and the other uses routers and Codecs, it is found that the routers and Codecs network takes less than 15% area, consumes less than 50% power of the routers only network and operates with 2.5x frequency.

Chapter 4 3D-NoC Design Exploration with Codec

4.1. Introduction

As illustrated in the second chapter of this thesis, long-interconnects stand in the way of creating power efficient and high-performance SoCs. Therefore, the challenges introduced by these long interconnects pushed the system designers to adopt new approaches such as 2D NoCs, which are discussed in chapter two and chapter three. Three-Dimensional IC integration or 3D is another approach, which is currently being adopted widely in the SoC design field.

3D-IC technology is a result of the increasing demand for dense and power efficient integrated solutions. For example, in a mobile device, the conventional way is to implement the digital and analog sub-systems on different chips; this approach requires using onboard connections to connect between the two chips. However, with the 3D integration approach, both analog and digital sub-systems can be integrated on different layers on the same chip. This leads to a significant area and power saving and better interconnect utilization.

Another important application for the 3D-IC technology is the development of many-cores systems which have put enormous constraints on the on-chip memory bandwidth.

Combining 3D with NoC is a natural solution to overcome the long interconnects problem and to reduce the number of hops/routers a packet should pass to reach a far target. The authors in [30] showed the advantages of using 3D-NoC architectures over 2D NoC especially for latency, throughput and consumed power and area. However, the challenges introduced by the 3D technology limit the integration between the two technologies [31].

Some of these challenges are related to the fabrication process of the 3D systems, such as vertical interconnect modeling, thermal management and power delivery. Other challenges are related to the integration of the NoC concept with 3D systems. For example, coming up with new network topologies and routing algorithms that fit 3D systems.

4.2. 3D-NoC in literature

2D-NoC has been investigated and explored intensively during last years. However, 3D-NoC is still considered as a new emerging technology. Most research is targeted to the development of 3D-NoC modeling and simulation tools. In [32], the authors developed an open source and generic NoC simulator using SystemC which is called Noxim. The tool does not support 3D-NoC natively but can be modified to mimic the behavior of 3D-NoC systems.

In [33], the authors implemented a 16-processor NoC-based 3D system that consists of two tiers in a mesh topology. In [34, 31], the authors explored the performance improvements and constraints for different 3D topologies. In [35], the authors designed a NoC router that exploits the vertical nature of 3D-NoCs.

In [36], the authors propose a distributed routing algorithm for vertically partially connected regular 2D topologies of different shapes and sizes.

In [37], the authors introduced two look-up table based routing algorithms for 3D-NoC. In [38], a virtual channel based routing algorithm is introduced to avoid deadlock in irregular networks and it uses a compact form of routing tables in order to minimize their overhead. In [39], the authors proposed a routing algorithm that depends on splitting the network into layers in order to provide deadlock and live-lock free operation. In [40], a routing algorithm optimized for power consumption and latency is introduced. Since mesh topology is the most used topology for 3D-NoCs, in [41, 42], the authors provide routing mechanisms designed especially for mesh topologies.

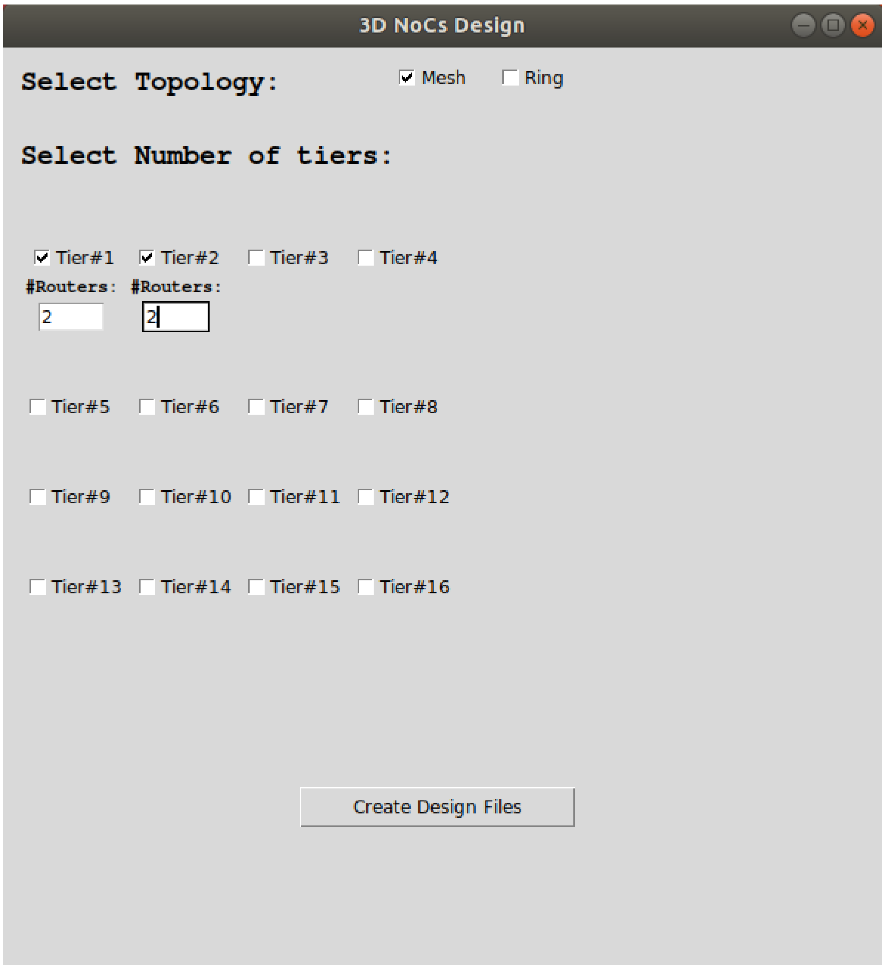
In [43], the authors developed a simulation tool called 3D-NOCET, the tool offers a generic and flexible solution to generate different 3D-NoC configurations. This tool could be used to do different performance evaluations according to the main network factors which are the number of tiers, number of routers per tier and the planar topology for the tiers.

4.3. 3D-NOCET as an Exploration Tool

4.3.1. Introduction

3D-NOCET tool supports full-mesh and ring as 2D topologies. The tool supports a maximum number of 16 tiers and 256 routers per tier. As shown in Figure 4.1, the tool provides a simple GUI (Graphical User Interface), by choosing the “Mesh”, “Tier #1” and “Tier #2” boxes, the tool is easily configured to generate 3D-NoC configuration with two tiers, full-mesh topology as 2D topology for each tier.

Behind its GUI, lays the automation infrastructure which consists of few scripts that generate the synthesizable SystemVerilog RTL code.



3D NoCs Design

Select Topology: ☑ Mesh □ Ring

Select Number of tiers:

☑ Tier#1 ☑ Tier#2 □ Tier#3 □ Tier#4

#Routers: 2 #Routers: 2

□ Tier#5 □ Tier#6 □ Tier#7 □ Tier#8

□ Tier#9 □ Tier#10 □ Tier#11 □ Tier#12

□ Tier#13 □ Tier#14 □ Tier#15 □ Tier#16

Figure 4.1: 3D-NOCET GUI without modification

4.3.2. Tool updates

The authors of 3D-NOCET tool have made it possible to extend the tool further to include more planar topologies; the original tool supports only full-mesh and ring topologies, a maximum number of 16 tiers and a maximum number of 256 routers per tier.

To study the effect of adding Codec to 3D-NoC networks, it is required to add the Codec block to the auto-generated RTL code. The updated tool, as shown in Figure 4.2, supports full-mesh and ring topologies, the number of tiles per Codec is set to four.

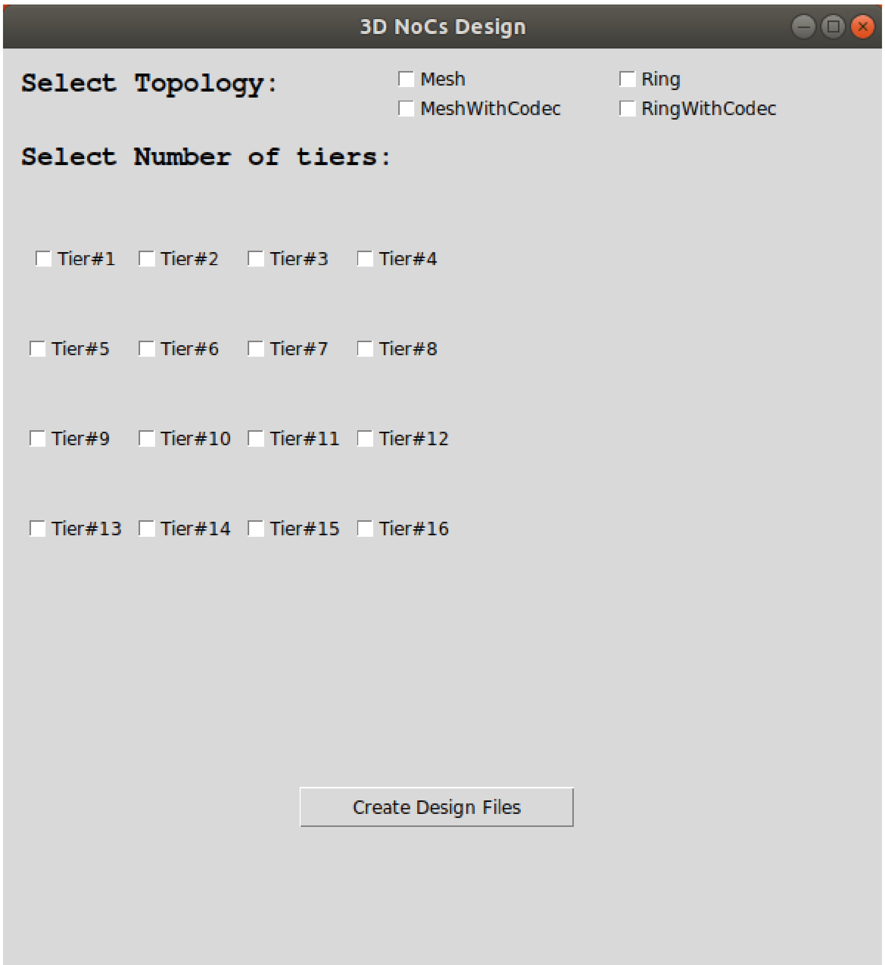

3D NoCs Design
Select Topology:
□ Mesh
□ Ring
□ MeshWithCodec
□ RingWithCodec
Select Number of tiers:
□ Tier#1 □ Tier#2 □ Tier#3 □ Tier#4
□ Tier#5 □ Tier#6 □ Tier#7 □ Tier#8
□ Tier#9 □ Tier#10 □ Tier#11 □ Tier#12
□ Tier#13 □ Tier#14 □ Tier#15 □ Tier#16
Create Design Files

Figure 4.2: Updated 3D-NOCET

4.4. Comparison setup and results for Full-Mesh topology

A comparison setup is created to study the performance differences between 3D-NoCs with and without Codec. The comparison methodology is done similar to the one in [43]. First, with respect to vertical complexity in which the impact of increasing the number of tiers is investigated. Second, the network complexity in which the 2D topology of one tier is investigated.

In this comparison, the full-mesh topology is used as the 2D topology for all tiers. In addition, one Codec per router (1CpR) and four tiles per Codec (4TpC) are used for all configurations. For investigating vertical complexity, a constant number of four tiles per tier is used (4TpT). For network complexity, a constant number of two tiers is used, the limitation of using only two tiers comes from the long compilation time required to do synthesis for bigger full-mesh topology.

The comparison uses Altera Arria II GX FPGA (EP2AGX260) as a target, it is the same target chip used in the 2D comparison discussed in Chapter 3. Quartus II Version 12.0 Build 178 is used as synthesis, time analysis and power estimation tool. All designs are synthesized at a target frequency of 200MHz for routers and 50MHz for interfacing with tiles.

4.4.1. Logic utilization

4.4.1.1. Vertical complexity

As shown in Figure 4.3 and Figure 4.4, increasing the number of tiers in a 3D network increases the consumed FPGA resources significantly. This is due to the increased number of instantiated routers; without using Codec, one router is required per PE or SE or tile. In addition, the increased amount of wiring and routing resources used to connect routers to the network and to other entities.

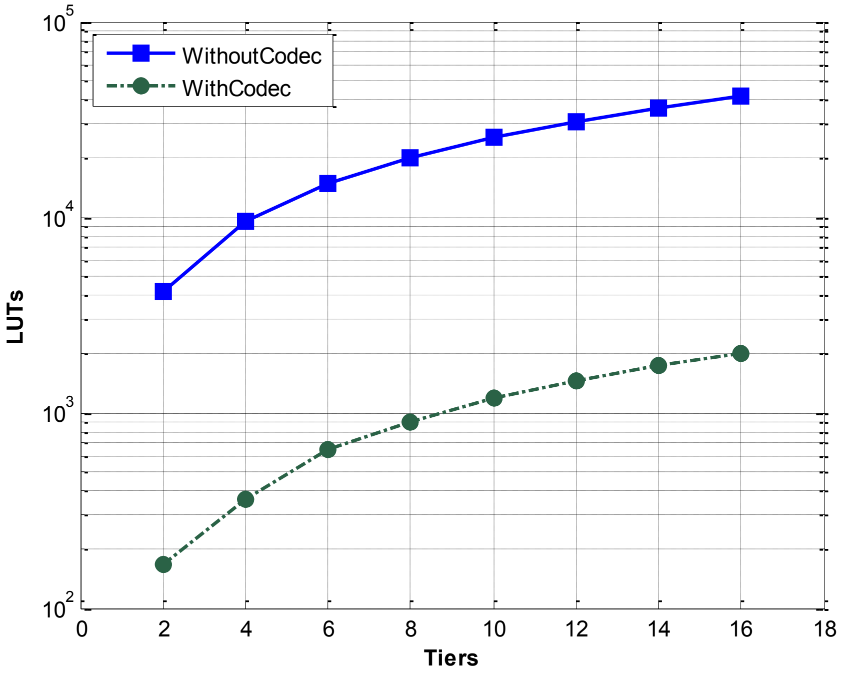


Figure 4.3: LUTs utilization for different numbers of Tiers

On the other hand, for 3D network configurations using Codec, every four tiles are sharing only one router port through a Codec module and this reduces the total number of instantiated routers to one-fourth. This leads to slightly increased FPGA resources due to extra wiring between tiles and Codecs and due to the logic consumption of the Codec modules themselves.

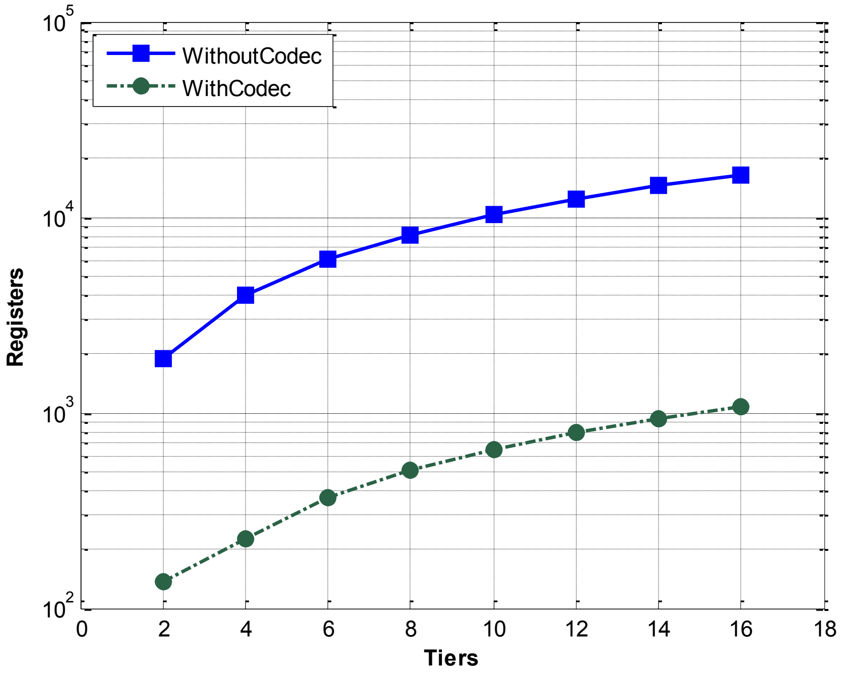


Figure 4.4: Registers utilization for different numbers of Tiers

4.4.1.2. Network complexity

Increasing network complexity means adding more routers or nodes in a single tier while maintaining the number of tiers constant (two tiers). As shown in Figure 4.5 and Figure 4.6, increasing the number of tiles per tier increases the consumed FPGA resources significantly. This is due to the increased number of router ports required to connect all routers in a full-mesh topology.

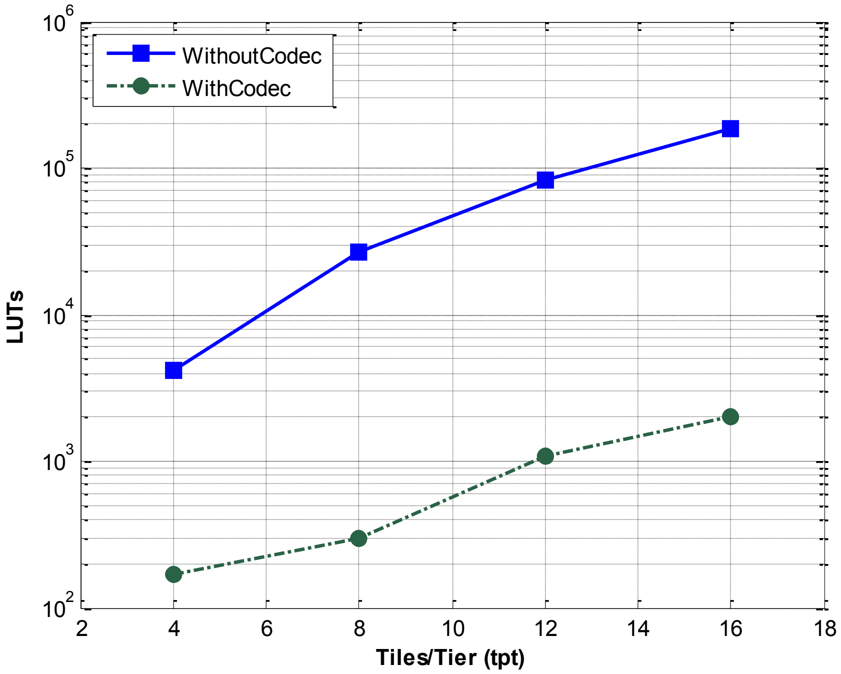


Figure 4.5: LUTs utilization for different numbers of Tiles per Tier

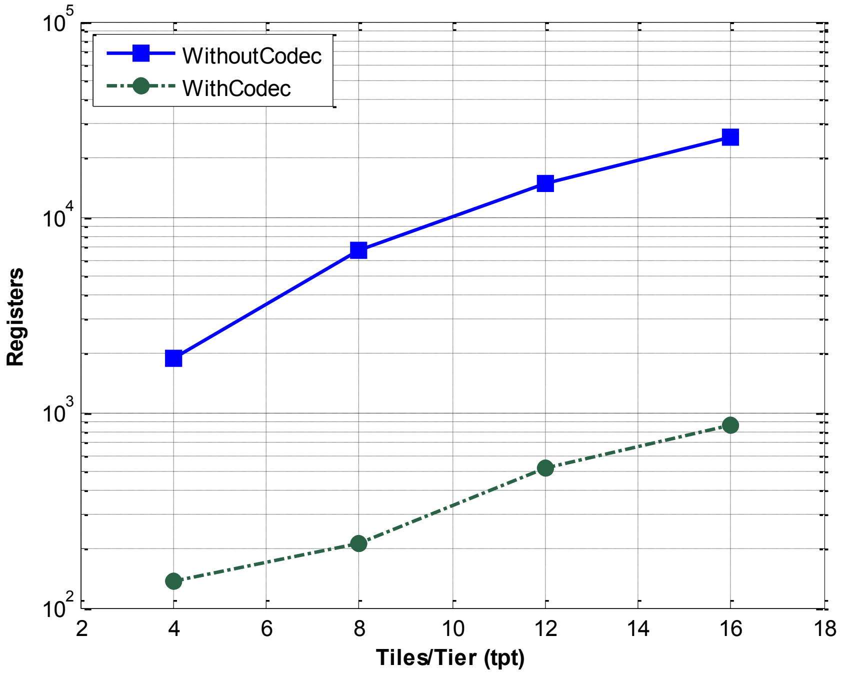


Figure 4.6: Registers utilization for different numbers of Tiles per Tier

4.4.2. Frequency

4.4.2.1. Vertical complexity

Increasing the network size in the vertical dimension (increasing number of tiers) with fixing the number of tiles per tier does not affect the complexity of arbiters and the switching logic. Hence, the maximum operation frequency is not affected significantly. As shown in Figure 4.7, the maximum operating frequency is not affected heavily by increasing the number of tiers.

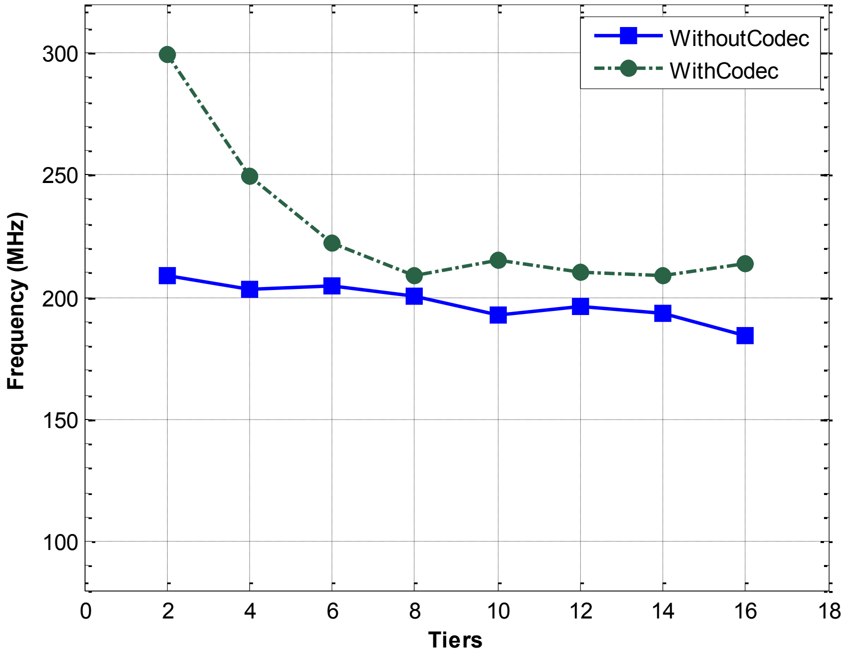


Figure 4.7: Maximum operating frequency for different numbers of Tiers

4.4.2.2. Network complexity

Increasing the number of tiles per tier in a full-mesh topology increases the router ports which leads to more complex arbiters and allocators, these two modules affects the maximum operating frequency significantly. As shown in Figure 4.8, for all network configurations, the Codec networks operate with a higher frequency that is at least 1.5x higher than the maximum frequency of regular networks.

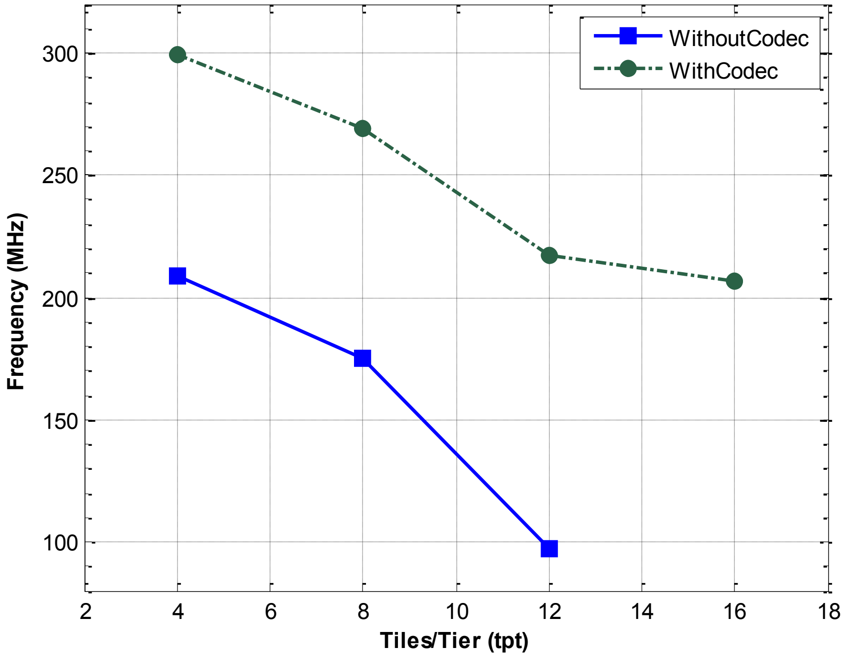


Figure 4.8: Maximum operating frequency for different numbers of Tiles per Tier

4.4.3. Power consumption

4.4.3.1. Vertical complexity

As shown in Figure 4.9, increasing the number of tiers in a 3D network increases the consumed power significantly. This is due to the increased number of instantiated routers.

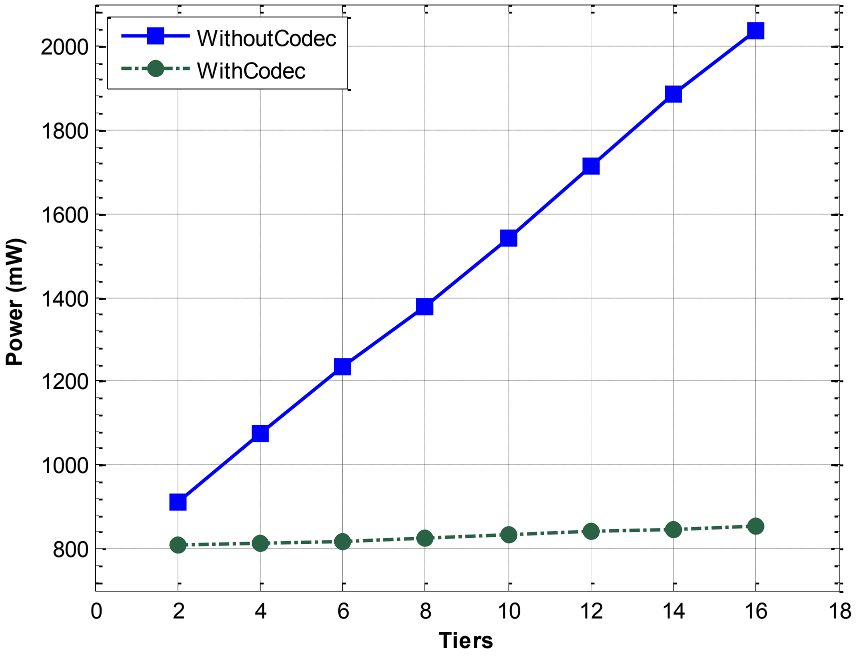


Figure 4.9: Power consumption for different numbers of Tiers

4.4.3.2. Network complexity

As shown in Figure 4.10, increasing the number of tiles per tiers increases the size of arbiters, allocators and input/output buffers leading to more static and dynamic power dissipation. On the other hand, Codec decreases the number of required routers compared to the regular network.

The value of the dynamic power dissipated for the 16tpt-without-Codec configuration is interpolated using the configurations of 12tpt and 8tpt. The reason for the interpolation is that the synthesis tool is not able to fit this network on the target FPGA.

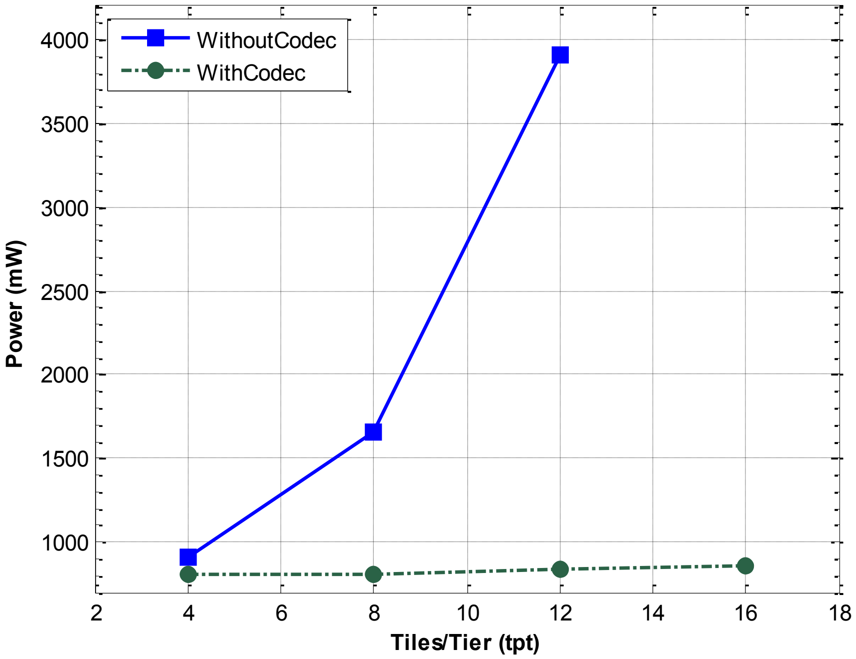


Figure 4.10: Power consumption for different numbers of Tiles per Tier

4.5. Comparison setup and results for Ring topology

The comparison for ring topology extends the results obtained from the previous full-mesh comparison. Unlike full-mesh topology, the router port count of ring topology does not depend on the number of routers per tier. Therefore, it enables exploring larger 3D-NoC configurations without increasing compilation time significantly. The comparison setup is created to explore networks with a higher number of tiers; networks with two to eight tiers are generated with a different number of tiles per tiers spanning four to sixteen.

In this comparison, the ring topology is used as the 2D topology for all tiers. One Codec per router (1CpR) and four tiles per Codec (4TpC) are used for all configurations. The maximum operating frequency is not considered in this comparison because it only changes slightly for ring topologies. The reason for that is that the port count of ring routers do not change with changing the network parameters, which are number of tiers and number of routers per tier.

4.5.1. Logic and memory utilization

As shown in Figure 4.11 and Figure 4.12, the logic utilization of ring topologies is significantly larger than that of ring-with-Codec topologies. Approximately larger by order of magnitude.

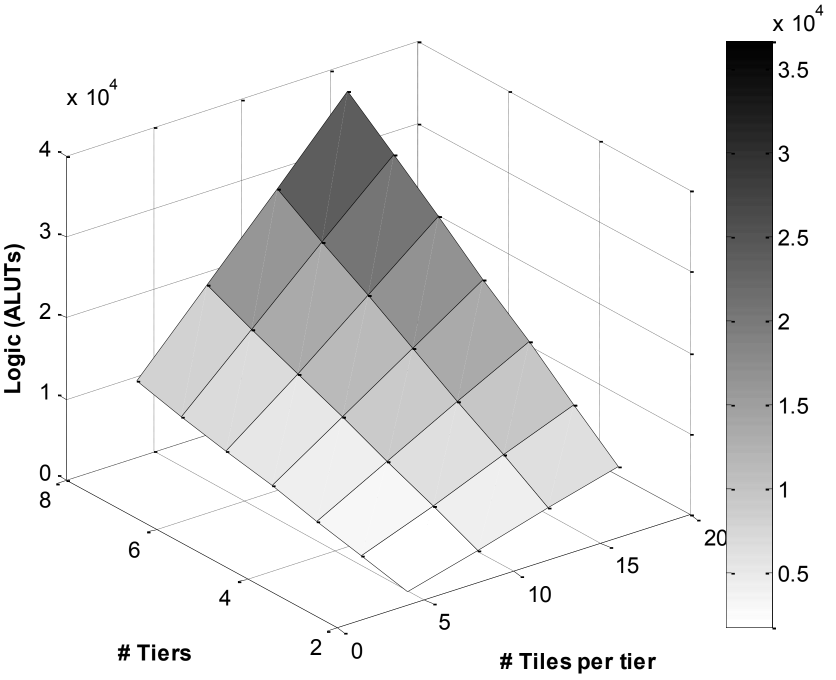


Figure 4.11: LUTs utilization for Ring

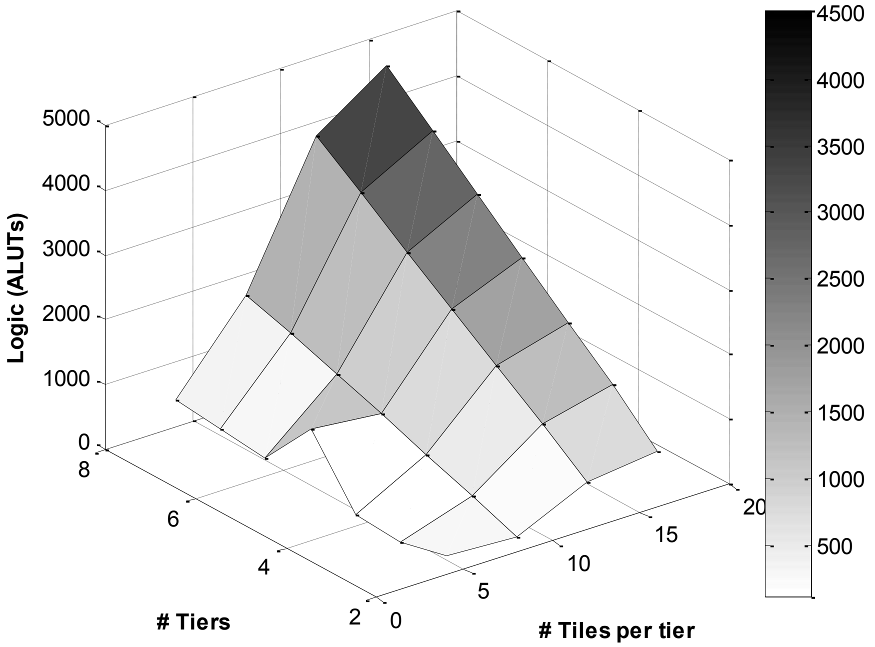


Figure 4.12: LUTs utilization for RingWithCodec

As shown in Figure 4.13 and Figure 4.14, memory consumption of ring topologies is also significantly larger than of that ring-with-Codec topology.

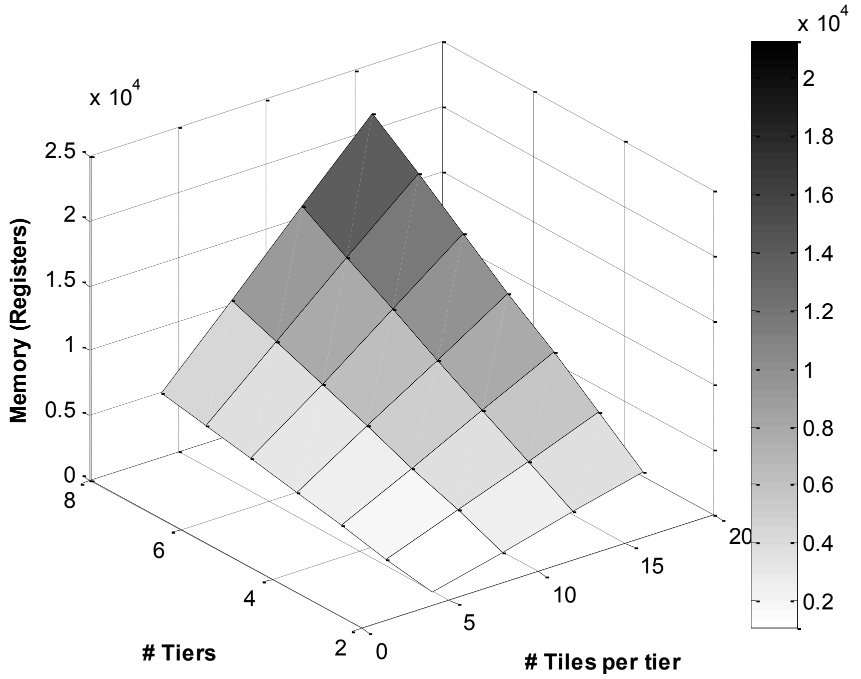


Figure 4.13: Memory utilization for Ring

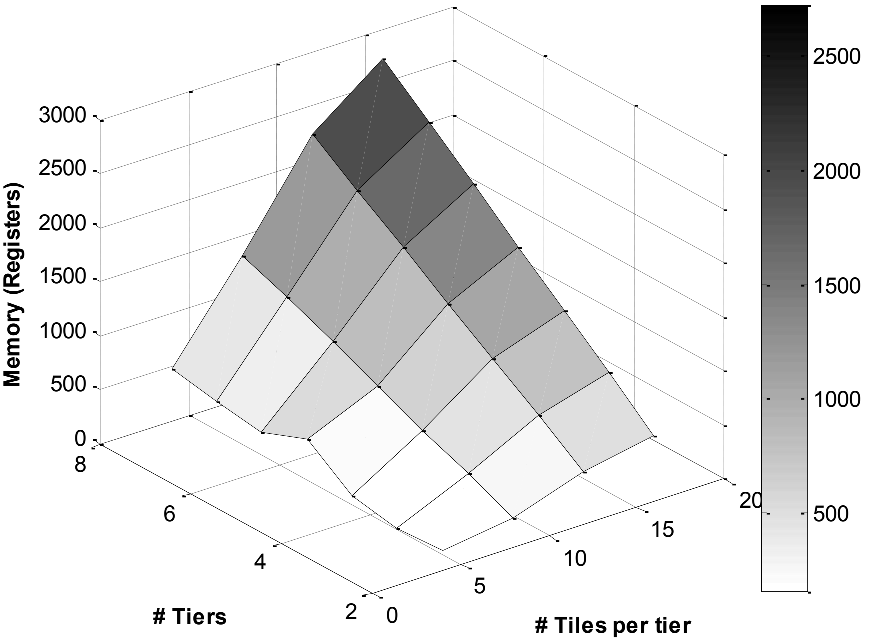


Figure 4.14: Memory utilization for RingWithCodec

4.5.2. Power consumption

As shown in Figure 4.15 and Figure 4.16, the power consumption of ring-only topology is higher than that of ring-with-Codec topology. For example, for the largest network (#Tiers = 8, #Tiles per tier = 16), the power consumption of ring-only network is approximately twice that of a ring-with-Codec network.

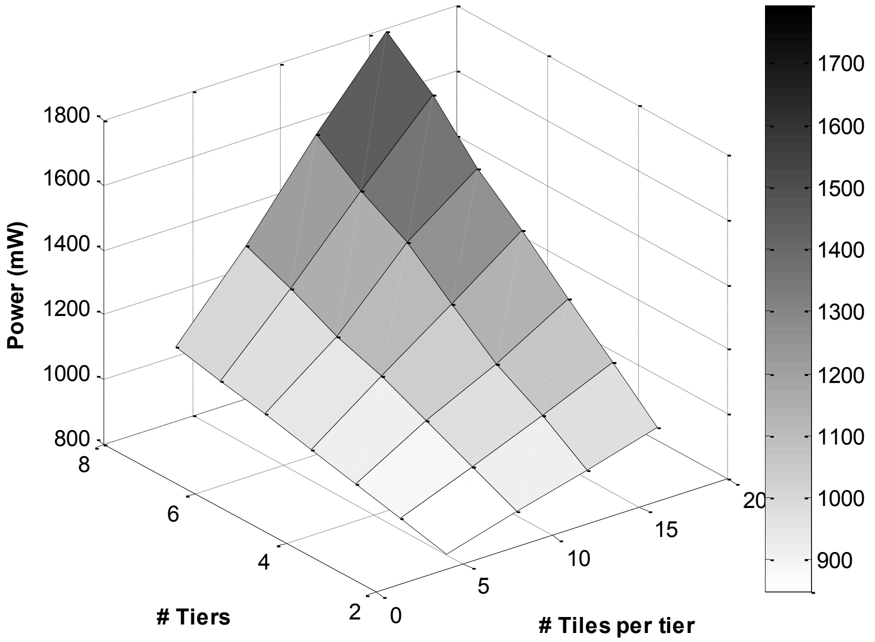


Figure 4.15: Power consumption for Ring

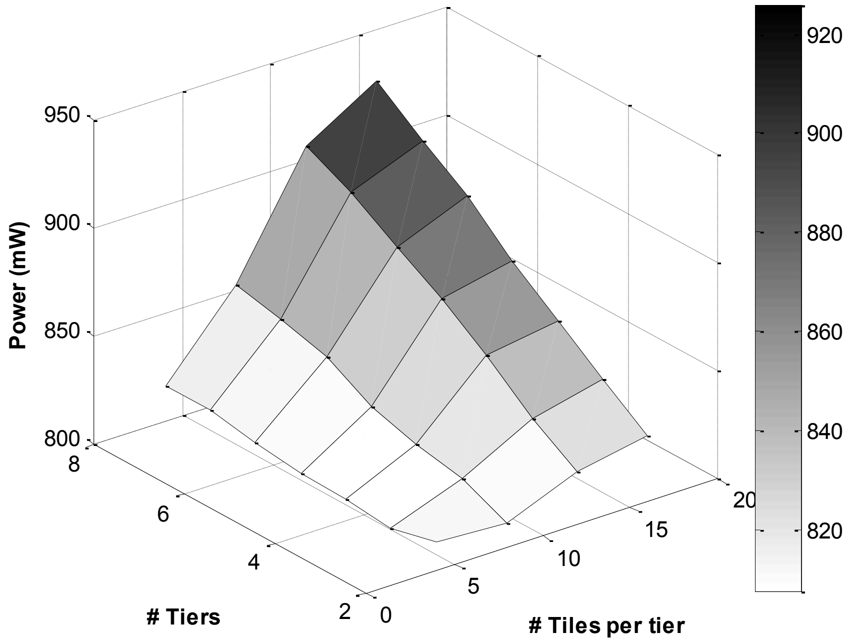


Figure 4.16: Power consumption for RingWithCodec

4.6. Summary

In this chapter, an introduction is provided that shows the advantages of using 3D technology with NoC to solve the interconnect problems. Then some previous 3D-NoC researches are presented, 3D-NOCET is one of these researches and it is chosen as an investigation and exploration tool.

The tool supports full-mesh and ring topologies, maximum of 16 tiers and 256 router per tier. The tool is updated in order to use it in further investigation and exploration of 3D-NoCs. The updated tool includes the integration of Codec to the auto-generated RTL code.

A comparison setup is created to study the performance differences between 3D-NoCs with and without Codec. The complexity of the 3D network is discussed with regard to two aspects. First, with respect to vertical complexity in which the impact of increasing the number of tiers is investigated. Second, the network complexity in which the 2D topology of one tier is investigated.

The comparison results for both full-mesh and ring topologies show that for the area, power and maximum operating frequency, 3D-NoC with Codec network outperforms the 3D-NoC only network.

Summary and Conclusion

Given the importance of FPGA platforms in today's market, this thesis explores the definition and solution of one of the issues facing the future development of FPGA.

In chapter two, first, a comparison between FPGA and ASIC is held in the context of unit cost, non-recurring engineering cost, development cycle, time to market, scalability and configurability, the importance of NoCs is highlighted especially for FPGA. Second, an introduction to NoC is given followed by a literature review for some NoC designs that are used in current research. The review studies their contributions, architectures, implementation, performance measurement results and future works. Finally, a comparison is held between three NoCs across different values of NoC parameters. The comparison results give design guidelines and recommendations to help choose the appropriate NoC according to system requirements.

In chapter three, first, Codec is introduced to come over one of the NoC problems, which is the performance degradation due to increasing input/output ports of NoC routers. Then a comparison between two 2x2 networks is held, one uses routers and Codec modules and the other uses routers only. It is found that the routers and Codecs network takes less than 15% area, consumes less than 50% power of the routers only network and operates with 2.5x frequency.

In chapter four, the effects of using Codec on 3D-NoC are investigated. First, the 3D-NOCET tool is updated to explore different 3D configurations. Second and similar to chapter three, a comparison between 3D-NoCs with and without Codec is held. The comparison results show that for the area, power and maximum operating frequency; the 3D-NoC network with Codec outperforms 3D-NoC only network.

The contributions of this thesis are briefly listed as follows:

- Review of different NoC designs, architectures and performance measurement results.
- Comparative review of three NoC routers to analyze their behavior while varying some parameters. This comparison helps to determine which parameters or sub-modules needs to be optimized to better adapt the NoC routers for the FPGA integration.
- Introduce Codec as a solution to the increased router port count problem. It is used to interface between FPGA tiles and NoC routers. A comparison is held between two 2D networks, with and without Codec.
- Investigate the impact of integrating Codec module into 3D-NoC

Mainly the results and conclusions of this thesis are published in [29, 44].

Recommendations for future work:

- Investigate more topologies with 3D-NOCET tool.
- Use a real NoC router in the generated RTL code instead of a simple one.
- Investigate latency and throughput performance Using network simulators (such as Noxim)

References

1. ARM, “Amba specification”, Technical report, Revision 2.0, 1999.
2. IBM Cooperation, “Coreconnect bus architecture”, Technical report, 1999.
3. R. Jayaraman, “(When) Will FPGAs Kill ASICs”, Proceedings of the 38th Design Automation Conference, pp. 321-322, 2001.
4. S. M. Trimberger, “Three Ages of FPGAs: A Retrospective on the First Thirty Years of FPGA Technology”, Proceedings of the IEEE, Vol. 103, Issue 3, 2015.
5. Altera Inc, “Standard Cell ASIC to FPGA Design Methodology and Guidelines”, Application Note 311, 2009.
6. Anysilicon, “FPGA vs ASIC, What to Choose”, https://anysilicon.com/fpga-vs-asic-choose, 2016.
7. A. Marquardt, V. Betz and J. Rose, “Speed and area tradeoffs in cluster-based FPGA architectures”, IEEE Transactions on Very Large Scale Integration (VLSI) Systems, Vol. 8, Issue 1, pp. 84-93, 2000.
8. W.-C. Tasi, Y.-C. Lan, Y.-H. Hu and S.-J. C, “Networks on chips: structure and design methodologies”, Journal of Electrical and Computer Engineering - Special issue on Networks-on-Chip: Architectures, Design Methodologies, and Case Studies, Vol. 2012, Article No. 2, 2012.
9. International Technology Roadmap for Semiconductors (ITRS), 2005 Edition, http:// http://www.itrs2.net, 2005.
10. E. Bolotin, I. Cidon, A. Kolodny and R. Ginosar, “Cost Considerations in Network on Chip”, Integration, the VLSI Journal, Vol. 38, pp. 19-42, 2004.
11. C. Nicopoulos, V. Narayanan and C. R. Das, “Network-on-Chip Architectures: A Holistic Design Exploration”, Lecture notes in Electrical Engineering Vol. 45, 2009.
12. M. K. Papamichael and J. C. Hoe, “CONNECT: CONfigurable NEtwork Creation Tool”, http://users.ece.cmu.edu/mpapamic/connect, 2012.
13. D. U. Becker, “Efficient Microarchitecture for Network-on-Chip Router”, Ph.D. dissertation, Stanford University, 2012.
14. G. Schelle and D. Grunwald, “Exploring FPGA network on chip implementations across various application and network loads”, International Conference on Field Programmable logic and applications, pp. 41–46, 2008.
15. C. Hilton and B. Nelson, “PNoC: A Flexible Circuit-Switched NoC for FPGA-based Systems”, IEEE Proceedings - Computers and Digital Techniques, Vol. 153, pp. 181-188, 2006.
16. R. Pau and N. Manjikian, “Implementation of a configurable router for embedded network-on-chip support in FPGAs”, M.Sc. Thesis, College of Engineering, Queen’s University, Kingston, Ontario, Canada, 2008.

17. K. Goossens, M. Bennebroek3, J. Y. Hur and M. A. Wahlah, “Hardwired Networks on Chip in FPGAs to Unify Functional and Configuration Interconnects”, NOCS '08 Proceedings of the Second ACM/IEEE International Symposium on Networks-on-Chip, pp. 45-54, 2008.

18. M. K. Papamichael and J. C. Hoe, “CONNECT: Re-Examining Conventional Wisdom for Designing NoCs in the Context of FPGAs”, 20th ACM/SIGDA International Symposium on FPGA, pp. 37–46, 2012.

19. Y. Huan1 and A. DeHon, “FPGA Optimized Packet-Switched NoC using Split and Merge Primitives”, IEEE International Conference on Field-Programmable Technology, pp. 47-52, 2012.

20. A. Imbewa and M. A. S. Khalid, “FLNR: A Fast Light-Weight NoC Router for FPGAs”, 56th International Midwest Symposium on Circuits and Systems (MWSCAS), pp. 445-448, 2013.

21. F. Moraes, A. Mello, L. Möller, L. Ost and N. Calazans, “Hermes: an Infrastructure for Low Area Overhead Packet-Switching Networks on Chip”, Integration, the VLSI Journal, Vol. 38, pp. 69-93, 2004.

22. T. Marescaux, T. Bartic, D. Verkest, S. Vernalde and R. Lauwereins, “Interconnection Networks Enable Fine-Grain Dynamic Multi-Tasking on FPGAs”, International Conference on Field-Programmable Logic and Applications, pp. 795-805, 2002.

23. T. A. Bartic, J.-Y. Mignolet, V. Nollet, T. Marescaux, D. Verkest, S. Vemalde and R. Lauwereins, “Highly Scalable Network on Chip for Reconfigurable Systems”, International Symposium on System-on-Chip, pp. 79-82, 2003.

24. H.-Y. Luo, S.-J. Wei, and D.-H. Guo, “RROCN: An on-chip network with regular reconfigurable topology for chip-multiprocessors”, Journal of computers (Taiwan), No.1, pp. 36-46, 2013.

25. N. Gamal, H. Fahmy, Y. Ismail and H. Mostafa, "Design Guidelines for Embedded NoCs on FPGAs", IEEE International Conference on Quality Electronic Design (ISQED’2016), Santa Clara, California, USA, IEEE, pp. 69-74, 2016.

26. N. Gamal, H. A. H. Fahmy, Y. Ismail, T. Ismail, M. Mohie-Eldin and H. Mostafa, "Design Guidelines for Soft Implementations to Embedded NoCs of FPGAs", International Design and Test Symposium (IDT 2016), Hammamet, Tunisia, IEEE, pp. 37-42, 2016.

27. M. S. Abdelfattah and V. Betz, “Design Tradeoffs for Hard and Soft FPGA-based Networks-on-Chip”, International Conference on Field-Programmable Technology, pp. 95–103, 2012

28. K. A. Helal, S. Attia, T. Ismail and H. Mostafa, "Comparative Review of NOCs in the Context of ASICs and FPGAs", International Symposium on Circuits and Systems (ISCAS 2015), Lisbon, Portugal, IEEE, pp. 1866-1869, 2015.

29. A. Salaheldin, K. Abdalah, N. Gamal and H. Mostafa, “Review of NoC-based FPGAs architectures”, International Conference on Energy Aware Computing Systems & Applications, pp.1-4, 2015.

30. B. Feero and P. Pande, “Networks-on-Chip in a Three-Dimensional Environment: A Performance Evaluation”, IEEE Transactions on Computers, pp. 32–45, 2009.

31. P. Leduca, F. de Crecy, M. Fayolle, B. Charlet, T. Enot, M. Zussy, B. Jones, J.-C. Barbe, N. Kernevez, N. Sillon, S. Maitrejean, and D. Louisa, “Challenges for 3D IC integration: Bonding Quality and Thermal Management”, IEEE International Interconnect Technology Conference, pp. 210–212, 2007.

32. V. Catania, A. Mineo, S. Monteleone, M. Palesi and D. Patti, “Cycle-Accurate Network on Chip Simulation with Noxim”, ACM Transactions on Modeling and Computer Simulation, Vol. 27, Issue 1, 2016.

33. M. H. Jabbar, D. Houzet and O. Hammami, “3D Multiprocessor with 3D NoC Architecture Based on Tezzaron Technology”, IEEE International 3D System Integration Conference (3DIC), p.1-5, 2011.

34. V. Pavlidis and E. Friedman, “3-D Topologies for Networks-on-Chip”, IEEE Transactions on Very Large Scale Integration (VLSI) Systems, pp. 1081–1090, 2007.

35. M. Bahmani, A. Sheibanyrad, F. Petrot, F. Dubois, and P. Durante, “A 3D-NoC Router Implementation Exploiting Vertically-Partially-Connected Topologies”, IEEE Computer Society Annual Symposium on VLSI (ISVLSI), pp. 9–14, 2012.

36. F. Dubois, A. Sheibanyrad, F. Pétrot, and M. Bahmani, “Elevator-First: A Deadlock-Free Distributed Routing Algorithm for Vertically Partially Connected 3D-NoCs”, IEEE Transactions on Computers, pp. 609–615, 2013.

37. F. Silla and J. Duato, “High-Performance Routing in Networks of Workstations with Irregular Topology”, IEEE Transactions on Parallel and Distributed Systems, pp. 699–719, 2000.

38. W. Jie and S. Li, “Deadlock-free Routing in Irregular Networks Using Prefix Routing Algorithm”, Tech. Rep., 1999.

39. O. Lysne, T. Skeie, S.-A. Reinemo, and I. Theiss, “Layered Routing in Irregular Networks”, IEEE Transactions on Parallel and Distributed Systems, pp. 51–65, 2006.

40. A. Ahmed and A. Abdallah, “LA-XYZ: Low Latency, High Throughput Look-Ahead Routing Algorithm for 3D Network-on-Chip (3D-NoC) architecture”, IEEE 6th International Symposium on Embedded Multicore SoCs (MCSoC), pp. 167–174, 2012.

41. K.-C. Chen, S.-Y. Lin, H.-S. Hung, and A. Wu, “Topology-Aware Adaptive Routing for Nonstationary Irregular Mesh in Throttled 3D NoC Systems”, IEEE Transactions on Parallel and Distributed Systems, pp. 2109–2120, 2013.

42. R. S. Ramanujam and B. Lin, “Randomized Partially-Minimal Routing on Three-Dimensional Mesh Networks”, IEEE Computer Architecture Letters, Vol. 7, Issue 2, pp. 37–40, 2008.

43. M. Beheiry, H. Mostafa, Y. Ismail, and A. M. Soliman, "3D-NOCET: A Tool for Implementing 3D-NoCs Based on the Direct-Elevator Algorithm", International Symposium on Quality Electronic Design (ISQED 2017)), Santa Clara, California, USA, IEEE, pp. 144-148, 2017.

44. A. Salaheldin, H. Mostafa, and A. M. Soliman, "A Codec, tiles to NoC Router Interface, for Next Generation FPGAs with Embedded NoCs", IEEE International Midwest Symposium on Circuits and Systems (MWSCAS 2017), Boston, MA, USA, pp. 1228-1231, 2017.

الملخص

نظرا للطلب المستمر لرقائق الكترونية أقوى وأكبر، يتم إضافة وحدات جديدة بصفة مستمرة إلى النظم على رقائق مثل المعالجات المدمجة ومعالجات الإشارات الرقمية وكتل الذاكرة. كلما زاد تعقيد النظام فإن التأثير السلبي لأنظمة التوصيل يزداد لأن سرعة اسلاك التوصيل لا تتطور مع تطور تكنولوجيا التصنيع وتصبح نظم التوصيل القائمة على نظام الحافلات او القائمة على التوصيل من نقطة إلى أخرى بمثابة عقبات امام تحقيق متطلبات النظم كلما زاد حجمها. عندما تستخدم في نظم كبيرة نسبيا فإن أدائها ينحدر لأنها تعتمد علي اسلاك طويلة للتوصيل بين جميع أجزاء الرقاقة، هذه الاسلاك الطويلة تؤثر بشدة في زيادة المساحة القدرة المستهلكين لشبكات التوصيل.

مصفوفات البوابات القابلة للبرمجة الميدانية مثلها كمثل النظم على الرقائق الالكترونية، يتم إضافة كتل ومكونات جديدة إلى هندستها حتى تستطيع تلبية الطلبات المتزايدة لتطبيقات اليوم. ومع ازدياد عدد مكوناتها يزداد أيضا اعتماد شبكات توصيلها على شبكات التوصيل على رقائق الكترونية وذلك للتغلب على مشاكل شبكات التوصيل المعتادة كشبكات التوصيل من نقطة الى نقطة والشبكات المعتمدة على نظام الحافلات. تتكون شبكات التوصيل على رقائق الكترونية من شبكة من الموجهات تتصل فيما بينها عن طريق اسلاك قصيرة نسبيا، كي تتصل احدى بلاطات مصفوفات البوابات القابلة للبرمجة الميدانية بأخرى فإن عليها فقط ان توصل البيانات التي تريد ارسالها الي أقرب موجه اليها بدلا من أن توصلها عن طريق اسلاك طويلة حيث يقوم الموجه بإدماج البيانات في حزمة وارسالها عن طريق الشبكة الي وجهتها النهائية.

يتم تقديم مراجعة للعديد من تصميمات شبكات التوصيل على رقائق الكترونية للحصول على نظرة عامة عن الابحاث الحالية في هذا الموضوع، تم إجراء المراجعة في سياق المساهمات والمعمارية والتنفيذ والعمل المستقبلي، ثم تم إجراء مقارنة بين ثلاثة أجهزة توجيه منهم لتحليل تأثير تغيير عدد القنوات الظاهرية وعرض البيانات وعمق المخازن على تردد التشغيل والمساحة المستهلكة وذلك للمساعدة على اختيار أفضل وفقا لمتطلبات النظم. تظهر المقارنة ان معمارية شبكات التوصيل على رقائق الكترونية تؤثر بشدة في المساحة والقدرة المستهلكة للنظام ككل.

نتيجة للمقارنة المذكورة فقد وجد أن أحد العوائق في استخدام شبكات التوصيل على رقائق الكترونية هو أن زيادة منافذ التوجيه الخاصة بها ستؤثر بشكل ملحوظ على المساحة والقدرة المستهلكة وأيضا على تردد النظام وتكرار النظام بشكل ملحوظ. ولكي تستفيد مصفوفات البوابات القابلة للبرمجة الميدانية من طريقة عمل شبكات التوصيل على رقائق الكترونية، يتعين على المرء إيجاد وسيلة لربط عدد كبير من الكتل دون زيادة منافذ الموجه، لذلك تم اقتراح استخدام وحدة مكثف أو كما يسمى كودك للربط بين أجهزة التوجيه والبلاط. سيؤدي استخدام برنامج الترميز هذا إلى تقليل تأثير زيادة عدد البلاط على المساحة والقدرة المستهلكة لشبكة التوجيه وعلي تردد تشغيل الشبكة أيضا.

من أجل تقييم تأثير استخدام الكودك تم إجراء مقارنة بين شبكتين بنفس الشكل والحجم، واحدة مع أجهزة التوجيه فقط والأخرى مع أجهزة التوجيه ووحدات الكودك. تم إجراء المقارنة في سياق المساحة والقدرة المستهلكة وتردد النظام. تظهر نتائج المقارنة أن الشبكة التي تستخدم كودك تستهلك مساحة 15٪ أقل وطاقة أقل بنسبة 50٪ من شبكة الموجهات فقط ويمكن أن تعمل بتردد أعلى بمقدار مرتين ونص.

وأخيرًا، مع ازدياد استخدام تقنية الدوائر المتكاملة ثلاثية الأبعاد بشكل متزايد للتأقلم مع الطلب اليوم، تم أيضًا دراسة تأثير إضافة وحدة الكودك الى أنظمة الدوائر المتكاملة ثلاثية الأبعاد المعتمدة على شبكات التوصيل على رقائق الكترونية.

مهندس: علاء صلاح الدين جمعه إبراهيم

تاريخ الميلاد: 1989/03/18

الجنسية: مصري

تاريخ التسجيل: 2012/10/01

تاريخ المنح:

القسم: هندسة الإلكترونيات والاتصالات الكهربية

الدرجة: ماجستير العلوم

المشرفون:

ا.د. أحمد محمد سليمان

د. حسن مصطفى حسن مصطفى

الممتحنون:

أ.د...................... (الممتحن الخارجي)

أ.د...................... (الممتحن الداخلي)

أ.د...................... (المشرف الرئيسي)

أ.د. .... (عضو)

عنوان الرسالة:

استكشاف تصميم مصفوفات البوابات القابلة للبرمجة الميدانية المعتمدة على شبكات التوصيل على رقائق الكترونية: رابط بين الموجه والبلاط للشبكات ثنائية وثلاثية الأبعاد

الكلمات الدالة:

مصفوفات البوابات القابلة للبرمجة الميدانية، شبكات التوصيل على رقائق الكترونية، رابط الموجه

ملخص الرسالة:

في هذه الرسالة ركزنا على كيفية تطوير واستخدام شبكات التوصيل على رقائق إلكترونية في الجيل القادم من مصفوفات البوابات القابلة للبرمجة الميدانية، تم تقديم مسح للأدبيات الموجودة التي تتناول شبكات التوصيل على الرقائق الإلكترونية، ثم تم تقديم مقارنة بين بعض من هذه الشبكات في نطاق المساحة المستهلكة والسرعة. تشير هذه المقارنة إلى أن زيادة عدد منافذ الموجه تؤثر على سرعة الشبكة ومساحتها والقدرة المستهلكة بشكل ملحوظ. لذلك قدمنا الكودك وهو رابط بين الموجه والبلاط يستخدم في ربط عدد أكبر من البلاط للشبكة دون زيادة عدد منافذ الموجه. تم عمل مقارنة بين شبكتين إحداهما تستخدم كودك وأخرى لا تستخدمه. وأخيرا درسنا تأثير إضافة الكودك الى الشبكات ثلاثية الأبعاد التي تستخدم شبكات على رقائق إلكترونية.

استكشاف تصميم مصفوفات البوابات القابلة للبرمجة الميدانية المعتمدة على شبكات التوصيل على رقائق الكترونية: رابط بين الموجه والبلاط للشبكات ثنائية وثلاثية الأبعاد

اعداد
علاء صلاح الدين جمعه إبراهيم

رسالة مقدمة إلى كلية الهندسة - جامعة القاهرة
كجزء من متطلبات الحصول على درجة
ماجيستير العلوم
في
هندسة الإلكترونيات والاتصالات الكهربية

يعتمد من لجنة الممتحنين:

الاستاذ الدكتور: الممتحن الخارجي

الاستاذ الدكتور: الممتحن الداخلي

الاستاذ الدكتور: المشرف الرئيسى

الاستاذ الدكتور: عضو

كليـة الهندســة - جامعــة القاهــرة
الجيـزة - جمهوريـة مصـر العربيــة
2018

استكشاف تصميم مصفوفات البوابات القابلة للبرمجة الميدانية المعتمدة على شبكات التوصيل على رقائق الكترونية: رابط بين الموجه والبلاط للشبكات ثنائية وثلاثية الأبعاد

اعداد
علاء صلاح الدين جمعه إبراهيم

رسالة مقدمة إلى كلية الهندسة - جامعة القاهرة
كجزء من متطلبات الحصول على درجة
ماجيستير العلوم
في
هندسة الإلكترونيات والاتصالات الكهربية

تحت اشراف

أ.د. أحمد محمد سليمان
أستاذ
قسم هندسة الإلكترونيات
والاتصالات الكهربية
كلية الهندسة – جامعة القاهرة

د. حسن مصطفى حسن مصطفى
مدرس
قسم هندسة الإلكترونيات
والاتصالات الكهربية
كلية الهندسة – جامعة القاهرة

كلية الهندسة - جامعة القاهرة
الجيزة - جمهورية مصر العربية
2018

استكشاف تصميم مصفوفات البوابات القابلة للبرمجة الميدانية المعتمدة على شبكات التوصيل على رقائق الكترونية: رابط بين الموجه والبلاط للشبكات ثنائية وثلاثية الأبعاد

اعداد
علاء صلاح الدين جمعه إبراهيم

رسالة مقدمة إلى كلية الهندسة - جامعة القاهرة
كجزء من متطلبات الحصول على درجة
ماجيستير العلوم
في
هندسة الإلكترونيات والاتصالات الكهربية

كلية الهندسة - جامعة القاهرة
الجيزة - جمهورية مصر العربية
2018