ON ENHANCING THE PERFORMANCE OF BUFFERLESS NETWORK-ON-CHIP

By

Mohamed Assem Abd ElMohsen 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
Computer Engineering

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2016

ON ENHANCING THE PERFORMANCE OF BUFFERLESS NETWORK-ON-CHIP

By
Mohamed Assem Abd ElMohsen 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
Computer Engineering

Under the Supervision of

Dr. Hatem M. El-Boghdadi

Professor
Computer Engineering Department
Faculty of Engineering, Cairo University

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2016

ON ENHANCING THE PERFORMANCE OF BUFFERLESS NETWORK-ON-CHIP

By
Mohamed Assem Abd ElMohsen 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
Computer Engineering

Approved by the
Examining Committee

_____
Dr. Hatem M. El-Boghdadi, Thesis Main Advisor
- Professor at the Faculty of Engineering, Cairo University

_____
Dr. Amr G. Wassal, Internal Examiner
- Associate Professor at the Faculty of Engineering, Cairo University

_____
Prof. Dr. Mohammad Z. Abdel Majeed, External Examiner
- Professor at the Faculty of Engineering, Al-Azhar University

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2016

Engineer's Name:	Mohamed Assem Abd ElMohsen Ibrahim
Date of Birth:	07/07/1988
Nationality:	Egyptian
E-mail:	mohamedassem@eng.cu.edu.eg
Phone:	+2 01007846437
Address:	407N, Pyramids Gardens Giza, Egypt

Registration Date:	01/10/2010
Awarding Date:	…./…./2016
Degree:	Master of Science
Department:	Computer Engineering
Supervisors:	Dr. Hatem Mahmoud El-Boghdadi
Examiners:	Prof. Dr. Mohammad Zaki Abdel-Majeed (External examiner) - Professor at the Faculty of Engineering, Al-Azhar University Dr. Amr Jalal El-Deen Wassal (Internal examiner) Dr. Hatem Mahmoud El-Boghdadi (Thesis main advisor)

Title of Thesis:
On Enhancing the Performance of Bufferless Network-on-Chip

Key Words:
Bufferless Network-on-Chip; Selection Function; Maximum Flexibility; Ranking Policies; Congestion Management;

Summary:

With the arrival of chip multiprocessor systems, Network-on-Chip (NoC) has started to form the backbone of communication within a microprocessor chip. However, unfortunately, the performance of NoC is bounded by the limited power and area budgets. Bufferless NoC has emerged as a solution to reduce power and area. Bufferless NoC eliminates the buffers used for routing and/or flow control and handle contention using packet dropping or packet deflection. In this thesis, we focus on enhancing the performance (latency and deflection count) of deflection-based bufferless NoC running latency-sensitive applications.

First, we present an analytical study for the traffic in bufferless NoC under the Maximum Flexibility (MaxFlex) selection function with different step sizes. We also provide an experimental study under MaxFlex. Simulation results show that with large values of step size, the latency could be reduced by 97% over using Straight Line selection function. The proposed analysis explains the outperforming experimental results.

Then we propose different flit ranking policies that focus on decreasing the deflection count of the flits. Simulation results show that the proposed ranking policies can reduce the latency by up to 58% compared to Oldest First policy.

Finally, we consider relaxing the effect of congestion in bufferless NoC under high injection rate. We propose two approaches for congestion prevention. The first considers running applications on NoC with extra nodes. The second considers dividing a certain load into a sequence of lighter loads. Simulation results show that the proposed approaches enhance the latency by up to 61% in addition to operating at higher injections rates.

Acknowledgments

بسم الله الرحمن الرحيم

"سُبْحَانَكَ لَا عِلْمَ لَنَا إِلَّا مَا عَلَّمْتَنَا ۖ إِنَّكَ أَنتَ الْعَلِيمُ الْحَكِيمُ﴿٣٢﴾" البقرة.
“Glory to You (O Lord), we have no knowledge except what you have taught us. Indeed, it is You who is the knowing, the wise (32)” Al-baqarah

" وَمَا تَوْفِيقِي إِلَّا بِاللَّـهِ ۚ عَلَيْهِ تَوَكَّلْتُ وَإِلَيْهِ أُنِيبُ ﴿٨٨﴾" هود.
“And my success is not but through Allah. Upon him I have relied and to Him I return (88)” Hood.

I would like to express my sincere gratitude to my advisor, Dr. Hatem El-Boghdadi, for his huge support, patience and immense knowledge. His guidance helped me in all the time of working on this thesis. I could not have imagined having a better advisor and mentor for my masters study.

Also, I take this opportunity to express gratitude to all of the Computer Engineering department members for their help and support.

I would like to thank my family for their encouragement, support, and attention without which I would never have made it to the end.

Last but not the least; I would like to thank my wife, Yousra, for being always there for me and for her support and kindness. My uttermost gratitude goes to Allah that I met her in such critical point in my life.

Table of Contents

List of Tables

List of Figures

Abstract

Network-on-Chip (NoC) is commonly used to connect different computing components. With the arrival of chip multiprocessor systems, NoC has started to form the backbone of communication between cores and memory within a microprocessor chip. Although NoC has started to form the backbone of communication between cores, the performance of such interconnection network is bounded by the limited power and area budgets. Bufferless NoC has emerged as a solution to reduce power and area. Bufferless NoC eliminates the buffers used for routing or flow control and handle contention using packet dropping or packet deflection.

We focus on enhancing the performance (in particular, packet latency and deflection count) of deflection-based bufferless NoC running latency-sensitive applications. We divide the work to focus on three aspects of NoC. First, we focus on selecting an output port for the outgoing packet. After that, we shift our focus to ranking the flits in order to select which one to serve first. Finally, we investigate relaxing the effect of congestion under high injection rate.

In the first part, we study the effect of Maximum Flexibility selection function (MaxFlex) on 2D bufferless meshes when a fixed or a variable step size is used. The selection function selects an output channel from a set of channels supplied by the routing function. MaxFlex is a well-known selection function that tries to maximize the number of routing choices as a packet approaches its destination. We investigate the distribution of packets through the NoC via increasing and/or varying the used step size as improving the distribution leads to better utilization and thus better performance. Simulation results show that using a larger step size can enhance the performance by up to 95% compared to using Straight Line selection function. Also, the results show that using variable step size enhances the performance compared to fixed step size by up to 29 %.

Concerning the second part, we devise and evaluate different flit ranking policies. A flit ranking policy chooses which flit should be served first, thus it determines which flit can select an output port first. In this work, we propose novel ranking policies that take the deflection behavior of the bufferless NoC into account. Via the experimental study, we compare these policies to the Oldest First (OF) ranking policy. Simulation results show that the performance of the proposed policies excels over fixed step size MaxFlex with OF as ranking policy by up to 58%.

Finally, we focus on congestion prevention for bufferless NoC running latency-sensitive applications. NoC congestion is one of the main roadblocks that prevent the bufferless NoC to operate under high injection rates. Thus, by relaxing the congestion, bufferless NoCs can approach the performance of buffered NoCs but without the extra cost of using buffers (power and area). To address this problem, we propose prevention mechanisms that target the deflection count of the flits. The proposed approaches aim to give more space for the flits to roam leading to fewer deflections which directly affects the overall packet latency. Via simulation, we show that the proposed approaches enhance the packet latency by 61% compared to fixed step size MaxFlex.

Chapter 1 : Introduction

In the last few years, there is an industry wide switch to many-core and multi-core systems. In such systems, the performance of the communication system is very critical to the performance of the whole system.

Network-on-Chip (NoC) has emerged as a solution for the limitations in the traditional communications approaches (e.g. buses) especially after the tremendous increase in the number of the communicating modules within a single silicon chip [1,2]. NoC is a group of switches connecting homogeneous or heterogeneous nodes in a multiple point-to-point fashion [3,4]. NoC switches forward the data to/from the nodes/switches over links equipped with input and output buffers.

Buffered NoCs became the de facto approach for communication between cores within chip as they are more scalable, reliable, and predictable. Buffered NoCs were shown to consume significant power and chip area. For instance, in the Intel Teraflops chip and the MIT RAW chip, NoC fabric consumes around 30% and 36% power respectively [5,6]. Focusing on a single NoC switch, a considerable fraction of power and area is used by the internal buffers of the switch. In [7,8], the buffers within a single switch consume around 37% power and 80% area. In addition to being heavy power and area consumers, buffered NoCs are more complex to design as they require extra handlers for packets placement and buffer overflow.

Bufferless NoC has emerged as a solution to decrease power and area requirements [9,10,11,12]. Bufferless NoC eliminates the buffers used within switches; which has a direct impact on power and area. In contrast to the traditional buffered NoC; when two packets compete for the same output port, the allocator either drops or deflects (misroute) the losing packet instead of buffering it. Dropped packet should be retransmitted again. On the other hand, deflected packet follows a non-productive port. Due to the hazards accompanying the dropping mechanism such as handling positive (ACK)/negative (NACK) acknowledgement (NACK buffers [9], NACK network [11]), storing the packet within the source node (extra storage), and retransmission (increase the total network load), in this thesis, we adapt the deflection approach.

Even though bufferless NoCs have their advantages regarding area and power consumption, they have their own problems. Eliminating buffers helps in decreasing the chip area and limiting the consumed power, but at the same time, the flits have no place to reside in case of port contention which leads to dropping or deflecting the flits. This dropping/deflecting mechanism results in increasing the NoC traffic volume which in turn consumes link bandwidth.

Both mechanisms under low to medium rates lightly affect the performance (packet latency and deflection count) leading to a performance approaching buffered NoCs. On the other hand, under high injection rates, the number of packets increases leading to more contention, as a result, using bufferless NoCs leads to reducing the total available bandwidth (as a result of increasing the traffic volume due to retransmitting the flits or deflecting the flits away from their destination) which eventually leads to a performance worse than buffered NoCs. Thus, bufferless NoC is shown generally to function efficiently under moderate loads and smaller NoC sizes [10].

In this thesis, we study several aspects of bufferless NoC to serve latency-sensitive applications. In other words, we aim to operate latency-sensitive applications on