NETWORK-ON-CHIP PERFORMANCE EVALUATION ON FPGA: A HARDWARE/SOFTWARE CORE-BASED SOLUTION

Miklécio Bezerra da Costa, Ivan Saraiva Silva

PPGC - Instituto de Informática - UFRGS - miklecio.costa@inf.ufrgs.br
Departamento de Informática e Matemática Aplicada - UFRN - ivan@dimap.ufrn.br

ABSTRACT

The increase of transistor integration on a chip capability has allowed the development of MP-SoCs (Multi-Processor System-on-Chip) to solve the problem of the growing complexity on applications. Based on that, the concept of NoC (Network-on-Chip) came up, being considered more adequate to the purposes of MP-SoC. Many NoC projects have been developed in the last years, what has created the need of evaluating the performance of these mechanisms. This paper presents systems based on traffic generator cores for NoC performance evaluation. The cores were implemented in the hardware description language VHDL, integrated with the NoC SoCIN and with some Altera's components, like Avalon bus and Nios II/e processor, synthesized and prototyped on a FPGA board. Software applications implemented in the programming language C configured the cores in order to obey some traffic patterns. These applications collected latency and throughput results measured by the cores on the FPGA and generated SoCIN evaluations. The results showed that these systems compose a consistent and flexible methodology for Network-on-Chip performance evaluation.

1. INTRODUCTION

The increase of transistor integration on a chip capability has allowed the development of MP-SoCs (Multi-Processor System-on-Chip). MP-SoCs have been used to solve the problem of the growing complexity on applications, because the traditional mono-processor systems present a theoretical physical limit in their processing capability [9].

Based on that, the concept of NoC (Network-on-Chip) came up, a mechanism of module interconnection on MP-SoC. NoC is considered more adequate to the purposes of a MP-SoC when compared to mechanisms like dedicated point-to-point hires, shared bus or bus hierarchy, because of characteristics like high scalability and reusability [1].

Many NoC projects have been developed, trying to attend this growing demand. These projects differ from each other because of the large number of possible NoC implementations. For that reason, some performance measures are used to evaluate the quality of those NoC projects in order to decide which one attends the requisites of a MP-SoC. The most important measures to evaluate an interconnection network, like a NoC, are latency and throughput [4].

This paper proposes two systems based on traffic generator cores for Network-on-Chip performance evaluation: one uses the measure latency and the other uses the throughput. In both systems, the cores are coupled on NoC routers and communicate with a software system, composed by one Nios II/e processor and one memory. The developed methodology includes the system synthesis and prototyping on a FPGA board.

This paper is divided in the following sections: section 2 presents some related works on NoC performance evaluation; section 3 presents a general view about NoC; section 4 presents some concepts involved in NoC performance evaluation that are used in this project; section 5 describes both systems proposed in this paper; section 6 presents the results obtained by the use of these systems on the NoC SoCIN; and finally, section 7 presents the conclusion.

2. RELATED WORKS

In the recent years, some works have been developed to evaluate NoCs performance. They combine bit level accuracy and flexibility, implementing on FPGA (Field Programmable Gate Arrays) many kinds of NoCs, traffics and performance measures.

In [5], a NoC emulation framework is proposed. It implements traffic generators and traffic receptors, which are coupled to the Network-on-Chip through network interfaces. The platform is controlled by the PowerPC embedded onto the FPGA board. It uses the FPGA device Xilinx Virtex 2 Pro v20 working at 50 MHz. The generated traffic may be stochastic or based on input traffic traces collected from real-life applications.

The method proposed on [11] emphasizes the speed and accuracy of FPGA implementations and allows evaluating large Networks-on-Chip on a single FPGA. One at a time, the NoC routers have their behavior simulated. This sequential simulation performs a system cycle of N regular cycles, where N is the number of routers. In this case, the platform is controlled by the processor ARM9 at a frequency of 86 MHz on the FPGA device Xilinx Virtex-II 8000.

3. NETWORK-ON-CHIP

A Network-on-Chip is formed by a set of routers and point-to-point links that interconnect the cores of an integrated system [12], as illustrated in Figure 1. It's a concept inherited from the traditional computer

interconnection networks.

The messages transmitted among the system cores are divided into packets. The packets may yet be divided into flits (flow control unit). The routers are responsible for delivering the packets of core communication, forwarding them to each router toward the destination.

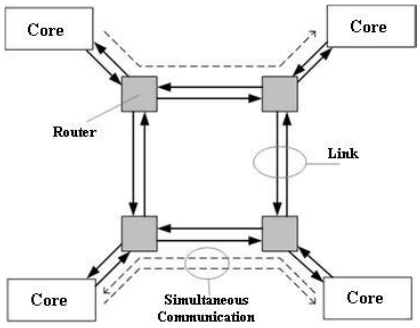

Figure 1 – A Network-on-Chip model

A NoC project is defined by how it implements a set of characteristics, such as: topology, routing, arbitration, switching, memorization and flow control. The diversity of implementations of these characteristics has allowed the development of several NoC designs, what intensifies the relevance of evaluating such projects performance. The definitions of these NoC characteristics are detailed in [2], [6], [4] and [3].

4. NOC PERFORMANCE EVALUATION

The performance evaluation of an interconnection network, such as a NoC, is based on two main measures: latency and throughput [4].

Latency is the interval of time between the start of the message transmission and the reception of it on the target node. However latency is a vague concept [4]. It may be the time between the injection of the packet's header into the router and the reception of its last information or it may count the time just on the routers, not on the nodes. The latency is measured in clock cycles or another time unit.

Throughput is the quantity of information transmitted on the network per time unit [3]. It may be measured in number of bits transmitted on each pair source-target or in percentage of this link utilization. In this last case, a communication link works on its maximum capability if it is used 100% of the time [4].

In order to collect these values, the literature describes some traffic patterns which define pairs source-target in the network for generation of packets, for example: Complement, Bit Reversal, Butterfly, Matrix Transpose and Perfect Shuffle. These patterns consider the data exchanges that parallel applications usually perform [7] [8] [10].

5. SYSTEM ARCHITECTURE

This paper proposes two systems based on cores for NoC performance evaluation: one measures the latency and the other measures the throughput. When coupled on the NoC routers, these cores return values of network performance. The systems were constructed to integrate hardware and software on a FPGA device, which allows fast and accurate evaluations.

Initially, both systems were implemented together, but the silicon costs of the full system were too high for the FPGA device used in this paper. Therefore, the systems were separated in order to ensure low silicon costs.

5.1. Latency-based System

The system that measures latency is illustrated in Figure 2, in which C represents the core. The hardware of the system communicates with a software application executed on the Nios II/e processor, through the Avalon bus. A memory module coupled on this bus and controlled by the Nios II/e processor completes the system. The software Altera Nios II IDE provides a graphic interface for the user to run the evaluation and analyze the results.

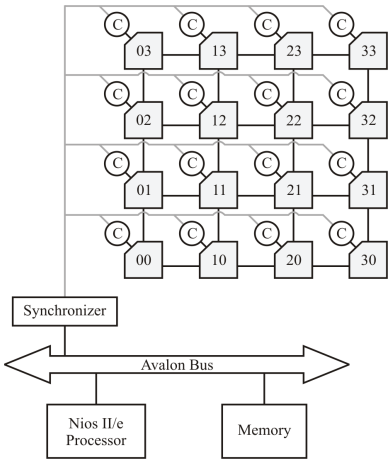

Figure 2 – Latency-based system architecture

Each core has a sender, a receiver and a timer module in order to count the number of cycles spent between the start and the end of a packet transmission. The application executed on the Nios II/e processor communicates with the cores through a synchronizer, which warranties a parallel configuration of the cores and does not make use of the NoC links for that.

The application sends configuration parameters to the cores, such as pairs source-target, size of the packets, injection rate and number of packets collected per result. Then the configured cores generate traffic on the NoC. The latency results measured by the cores are sent back to the Nios II/e and collected by the software application. These data transferences must be adapted to NoC's packet format and addresses, such as shown in Figure 3, where the packets were adapted to the NoC used in the results section of this paper. In this figure, the 4 types of packets used in the system are detailed: a) receiver

configuration, b) sender configuration, c) generated traffic and d) latency result.

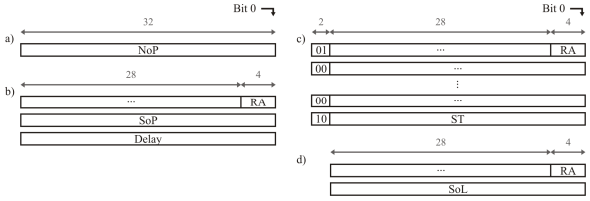

Figure 3 – Packets' format in the latency-based system

The receiver configuration packet must be sent by the application to define the number of packets (NoP) that the cores must receive before returning the latency result. Then, the application must transmit one sender configuration packet to each core sequentially to define the target receiver address (RA), the size of the packet (SoP) generated and the delay between each packet injection. The generated traffic packets are automatically sent to the receiver address, carrying the sender time (ST) to calculate the latency. Finally, when the core receives NoP packets, the latency result packet is sent to the processor to return the sum of latencies (SoL) accumulated.

5.2. Throughput-based System

Structurally the system that measures throughput is formed by the same components as the latency-based system. The adaptations were restricted to the core, the synchronizer and the interface between core and router. It's a different type of measure, so the core and the synchronizer behave in a different way. In the core-router interface four ports were inserted to count the number of packets transmitted on the four router links, as illustrated in Figure 4.

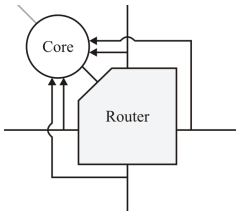

Figure 4 – Alteration for the throughput-based system architecture

A counter module was inserted into the core to count the throughput on the links at the same time that the packets are sent and received. The system packets for the throughput evaluation of the NoC used in the results section of this paper are illustrated in Figure 5: a) counter configuration, b) sender configuration, c) generated traffic and d) throughput result. In the counter configuration packet, CT (counter time) defines the interval of time spent during the measurement. The throughput result packet returns to the processor the number of packets counted on north (NNoP), east (ENoP), south (SNoP) and west (WNoP) link.

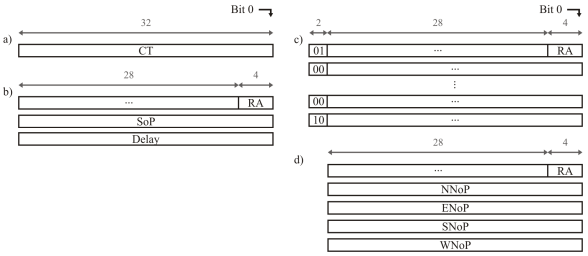

Figure 5 – Packets' format in the throughput-based system

6. RESULTS

In order to obtain experimental results and validate the systems developed, both of them were applied to the NoC SoCIN, a very well recognized and documented interconnection network that implements the router ParIS [13]. The NoC SoCIN has 2-D grid (or mesh) topology and wormhole packet-based switching. Other aspects are parameterizable. In this project, a NoC of dimensions 4x4 was configured to use deterministic (XY) routing, handshake flow control, dynamic (round roubin) arbitration and FIFO memorization on the input ports.

The implementation in the hardware description language VHDL of each system was integrated to the SoCIN implementation, also described in VHDL. Then the complete systems were synthesized and prototyped on the Altera DE2 FPGA board, specifically the device EP2C35F672C6 of the family Cyclone II, obtaining fast and accurate evaluations. Table 1 shows the silicon costs of both systems.

Table 1 – Silicon costs of the systems

Component	Logic Elements	
	Latency-based	Throughput-based
Core	615	795
Synchronizer	1821	2186
Processor	926	886
3-port router	613	190
4-port router	940	299
5-port router	1208	417
Whole system	26130 (79%)	20212 (61%)

The core and the synchronizer are more complex in the throughput-based system because it measures more results per core, four instead of only one like in the latency-based system. However the routers are smaller because the generated traffic does not use the data channel, obtaining a whole system with a reduction of 5918 logic elements when compared to the latency-based system.

The maximum operational frequency of the latency-based system is 71.4 MHz and the one of the throughput-based system is 91.6 MHz. Based on this, the results were calculated using the FPGA board oscillator at 50 MHz as clock source.

With the systems prototyped on FPGA, software applications implemented in the programming language C and executed on the Nios II/e processor injected

configuration packets into the systems, defining parameters for measuring latency and throughput and selecting one of a set of programmed traffic patterns: Complement, Bit Reversal, Butterfly, Matrix Transpose and Perfect Shuffle. The results were collected by the same applications.

In Figure 6, a graphic of average latency per core, as the interval of time between the start of packet transmission and its reception on that core, is illustrated for the traffic pattern Complement. As expected from the traffic pattern Complement, the cores located in the center of the network, that is, where the source core is nearer the target core, present lower latency values when compared to the peripheral ones.

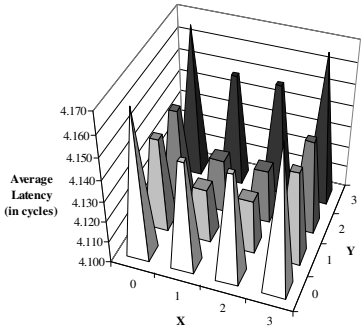

Figure 6 - Average latency per core for the Complement

In Figure 7, the throughput on each SoCIN bidirectional link as a percentage of utilization is illustrated for the traffic pattern Complement. The highest throughput values are highlighted and located exactly on the links with the biggest access disputes, according to the pattern Complement.

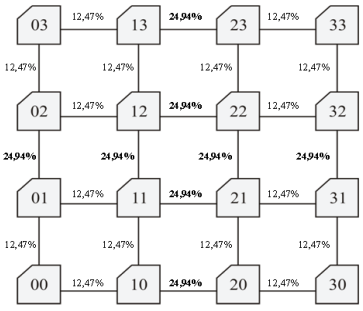

Figure 7 - Throughput per link for the Complement

The same analysis was performed for the other traffic patterns. It was observed that the behavior measured always obeys the traffic pattern applied. The latency result is always related to the distance between the pairs source-target in the pattern. And the throughput result is always related to the network congestion, determined by the static paths of the generated packets in the traffic pattern.

7. CONCLUSION

This paper presented two proposals of systems based on traffic generator cores for Network-on-Chip performance evaluation: one measures latency and the other measures throughput. The experiments were performed using a FPGA device, which provided fast and accurate evaluations. Both systems were applied to the NoC SoCIN in order to verify the consistence of the results. It was concluded that the systems are consistent, because the results revealed the traffic patterns' tendency of behavior.

The utilization of both systems allows an important type of NoC evaluation, because the measures latency and throughput are the most important for interconnection mechanisms performance evaluation. The methodology presented in this paper may be applied to any type of NoC that implements the handshake flow control, adapting only the format of system packets to the NoC packet format. Future works include improvements on silicon costs of the systems and the implementation of others flow control mechanisms, allowing the evaluation of more NoC designs, including larger ones.

8. REFERENCES

[1] Benini, L.; Micheli, G. D. "Networks on Chips: A New SoC Paradigm". IEEE Computer Society Press: 2002.

[2] Carara, E. A. "Uma Exploração Arquitetural de Redes Intra-chip com Topologia Malha e Modo de Chaveamento Wormhole". Trabalho de Conclusão II, PUCRS, Porto Alegre, 2004.

[3] Dally, W.; Towles, B. "Principles and Practices of Interconnection Networks". Morgan Kaufmann Publishers Inc.: 2003.

[4] Duato, J.; Yalamanchili, S.; Ni, L. "Interconnection Networks". Elsevier Science, 2002.

[5] Genko, N. et al. "A Complete Network-On-Chip Emulation Framework". In: Design, Automation and Test in Europe (DATE), 2005.

[6] Glass, C. J.; Ni, L. M. "The Turn Model for Adaptive Routing". In: International Symposium on Computer Architecture (ISCA'92). ACM: 1992.

[7] Kim, J. H.; Chien, A. A. "An Evaluation of the Planar/Adaptive Routing". In: IEEE Symposium on Parallel and Distributed Processing, 1992.

[8] Leighton, F. T. "Introduction to Parallel Algorithms and Architectures: Arrays, Trees, Hypercubes". Morgan Kauffmann, San Francisco, 1992.

[9] Loghi, M. et al. "Analyzing On-Chip Communication in a MPSoC Environment". In: Design, Automation and Test in Europe (DATE), 2004.

[10] Miller, P. R. "Efficient Communications for Fine-Grain Distributed Computers". Ph.D. Dissertation, Southampton University, U.K., 1991.

[11] Wolkotte, P. T.; Hölzenspies, K. F.; Smit, J. M. "Fast, accurate and detailed NoC simulations". In: International Symposium on Networks-on-Chip (NoCS), 2007.

[12] Zeferino, C. A. "Redes-em-Chip: arquiteturas e modelos para avaliação de área e desempenho". Tese de Doutorado, UFRGS, Porto Alegre, 2003.

[13] Zeferino, C. A.; Santo, F. G. M. E.; Susin, A. A. "ParIS: A Parameterizable Interconnect Switch for Networks-on-Chip". In: Symposium on Integrated Circuits and Systems Design (SBCCI), 2004, Porto de Galinhas. ACM Press, 2004.