Topological Impact on Latency and Throughput:
2D versus 3D NoC Comparison

Yan Ghidini[1], Thais Webber[2], Edson Moreno[1], Ivan Quadros[1], Rubem Fagundes[2], César Marcon[1]
[1]PPGCC – Programa de Pós-Graduação em Ciência da Computação
[2]PPGEE - Programa de Pós-Graduação em Engenharia Elétrica
PUCRS - Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, Brazil
{yan.ghidini, thais.webber, ivan.quadros}@acad.pucrs.br, {edson.moreno, rubem.fagundes, cesar.marcon}@pucrs.br

Abstract—**NoC has emerged as an efficient communication infrastructure to fulfill the heavy communication requirements of several applications, which are implemented on MPSoC target architectures. 2D NoCs are natural choices of communication infrastructure for the majority of actual chip fabrication technologies. However, wire delay and power consumption are dramatically increasing even when using this kind of topology. In this sense, 3D NoC emerges as an improvement of 2D NoC aiming to reduce the length and number of global interconnections. This work explores architectural impacts of 2D and 3D NoC topologies on latency, throughput and network occupancy. We show that, in average, 3D topologies minimize 30% the application latency and increase 56% the packets throughput, when compared to 2D topologies. In addition, the paper explores the influence of the buffer length on communication architecture latency and on application latency, highlighting that when applying an appropriate buffer length the application latency is reduced up to 3.4 times for 2D topologies and 2.3 times for 3D topologies.**

Keywords-component; Throughput, latency, 2D NoC, 3D NoC

I. Introduction

Planar network-on-chip (NoC), or 2D NoC, is an efficient communication infrastructure for multiprocessor system-on-chip (MPSoC) architectures, containing dozens or hundreds of processing elements (PEs), due to its high scalability and parallelism. 3D NoCs emerged as a natural evolution of 2D NoCs, reducing the number of hops that packets must pass through, and consequently, decreasing the network latency [1]. It offers an opportunity to continue the performance improvements using CMOS technology, with smaller form factor, higher integration density, and support for the realization of mixed-technology chips [2]. The 3D technology results in smaller footprint in each layer and shorter vertical wires that are implemented using Through Silicon Vias (TSVs) across the layers. Vertically stacked dies with TSVs together with NoCs have been proposed as powerful solutions to tackle the on-chip communication problem [3].

The major advantage of 3D ICs is the considerable reduction in the length of global interconnections, resulting in an increase in the performance and a decrease in the power consumption and area of wire limited circuits [4]. Moreover, 3D integrated systems hold promises to significantly reduce latency and power consumption, improve area and communication performance while increasing bandwidth, mainly due to their compact geometry. The geometric distance between PEs grows very differently in 2D and 3D structures with the number of PEs. Since for global and long distance communication the geometric distance translates linearly to latency, it can be expected to minimize the communication latency by 50% [5].

978-1-4673-2608-7/12/$31.00 ©2012 IEEE

The main contribution of this paper is to analyze the architectural impact of 2D and 3D NoC topologies on network and application latencies and data throughput. We compare a 3D NoC architecture called Lasio, which was implemented in the context of this work, with one Hermes 2D NoC [6], evaluating latency and throughput on different traffic scenarios, such as all-to-all and complements. Our evaluation shows the significant impact of buffer length on application latency for 2D and 3D topologies, which cannot be neglected in NoCs design exploration.

This paper is organized as follows. Section 2 presents the 3D NoC Lasio architecture implemented for this work. Section 3 shows experimental setup applied to analyze the latency and throughput results. Section 4 describes the experimental results in detail, and Section 5 presents the conclusion of the paper.

II. LASIO 3D NOC ARCHITECTURE

For this work a 3D NoC architecture called Lasio was designed. The Lasio characteristics are described in the subsequent sections.

A. Topology

The way the routers and PEs are interconnected defines the topology of the network. The topology of a three-dimensional NoC may vary according to the design requirements [7]. In the Lasio NoC design it was used 3D mesh topology to facilitate the placement of routers and PEs, as well as the routing channels between routers, simplifying the routing algorithm implemented in the control logic. In the 3D architecture developed in this work, the hops between layers (Z coordinate) and the hops within each layer (coordinates X and Y) have the same cost.

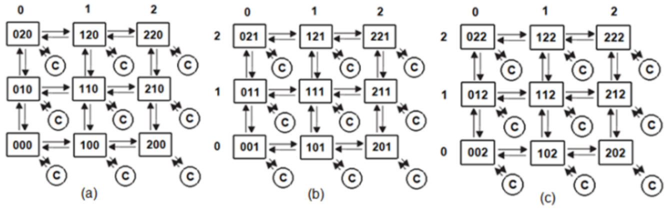

Figure 1. Example of 3x3x3 Lasio NoC - 'C' means PE; numbers are X,Y,Z router coordinates - (a) bottom layer (b) middle layer e (c) top layer.

Figure 1 illustrates the router address of Lasio according to X, Y and Z coordinates. For example, a router addressed by '123' has the value 1 for the X coordinate, 2 for the Y coordinate and 3 for the Z coordinate. In this figure, the three stacked layers are detailed separately. Layer (a) is the bottom layer (coordinate Z=0); layer (b) is the middle layer (coordinate Z=1) and layer (c) is the top layer (coordinate Z=2).

B. Routing, Arbitration and Switching

The Lasio router contains control logic responsible for routing and arbitration, and seven bi-directional ports, as illustrated in Figure 2: East, West, North, South, Local, Top and Bottom. Five are dedicated to connections within each layer (Local, North, South, East and West) and two other ports (Top and Bottom) ensure communication between adjacent layers. The Local port establishes a communication between the router and its corresponding PE, while the remaining ports are connected to neighboring routers. Each port has a buffer with configurable size for temporary data storage.

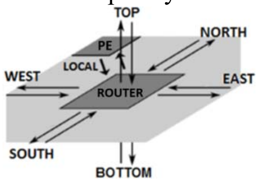

Figure 2. Connections of a Lasio NoC central router.

The packets employed in the transmissions are composed by a target address field, a size field, and a payload. Size field is the number of flits (minimal logic unit used for flow control) contained in the payload. Figure 3 shows the basic composition of a Lasio NoC packet.

1º flit	2º flit	3º flit	. . .	(Size + 2)º flit
Target address	Size	data	. . .	data
header		payload		

Figure 3. Lasio NoC packet.

Lasio implements XYZ routing algorithm, which is an extension of XY routing algorithm exploited on 2D NoCs. This routing algorithm is deadlock free and also enables a small implementation area. Hence, from the source router to the target router, packets are routed firstly in X, secondly in Y and thirdly in Z coordinates, respectively, passing through several buffers and ports. If the chosen output port is busy, all subsequent flits are blocked in the input buffer, and the request remains active until the connection with the port is established. When the port is free, the connection between the input port and output port is established.

TABLE I. EXAMPLE OF LASIO NOC SWITCHING TABLE

	Port number: Port name						
	0: East	1: West	2: North	3: South	4: Local	5: Bottom	6: Top
free	0	1	0	0	1	1	1
in	-	2	3	-	0	-	-
out	4	-	1	2	-	-	-

There are three vectors called in, out and free that are located in a switching table and are explored during the execution of the routing algorithm. The free vector is used to indicate the availability of a given output port, meaning if the port is transmitting a packet (i.e. busy, logical value '0') or not (i.e. free, logical value '1'). When there is a transmission request from an input port, the routing policy tries to find an available output port. In this case, all three vectors are updated. Otherwise, the input packet remains contained in the input buffer. The free position related to the output port is set as busy while the in and out vectors are interlaced. The in vector indicates to which output port the packet is being routed, while out indicates to which input port the packet is coming from. An example is depicted in Table 1.

It is important to remark that all router ports at Lasio NoC are bidirectional. Observing the North port (port number 2) on Table 1, it is noticeable that its output port is set as busy (observing the free line) because it is transmitting a packet from West input port (observing the number at the out line). At the same time, the input port is assigned to the South port (observing the number at the in line). When the packet transmission is finished, the free vector is updated.

The arbiter uses a dynamic rotating policy that prioritizes the packet routing on the input port. In other words, the arbitration is implemented using Round Robin algorithm. This method ensures that all incoming requests are processed, preventing starvation phenomenon. The arbitration logic takes four clock cycles to address a routing request. This additional time is necessary for the execution of the routing algorithm. If the routing algorithm cannot establish a connection to the desired output port, the input port requires to the arbiter a new routing request.

Furthermore, Lasio NoC implements wormhole switching method because of advantages as: (i) the need for smaller buffers for storing data, and (ii) low-latency communication.

C. Flow Control

Lasio NoC utilizes credit-based flow control that is an optimized communication mechanism, since it may consume few clock cycles to perform a flit transmission, which enables low communication latencies. This method utilizes FIFO buffers with customizable size at the receiver input, and a return line to the transmitter informing if there is available space in this buffer. This information can be interpreted by the transmitter as a credit and it just sends data if a credit is available. This approach does not allow packets to be discarded because a transmission between routers starts only after verifying if data will be received. This verification is performed as follows: the receiver sends to the transmitter a signal information indicating credits availability and the transmitter sends data only when there is credit available.

III. EXPERIMENTAL SETUP

A. Traffic Scenarios

All requirements evaluated here are strongly dependent of task mapping into PEs. Therefore, we choose two synthetic traffic scenarios (*all-to-all* and *complement*) whose symmetry enables to perform independent task mapping evaluations.

All-to-all

In this traffic pattern, all nodes send the same quantity of data (uniform packet load) in a deterministic way to all others nodes, except to itself. Firstly, all nodes simultaneously send a packet to node 0, and then all nodes send packets to node 1, and so on. This model of traffic is used to cover as many traffic and blocking situations as possible, since a large number of packets are traveling on the network simultaneously. Despite of all-to-all traffic scenario not be the best approach when dealing with real applications, since the communication destinations change frequently, such traffic scenario allows to meet easily any shortcomings in the communications infrastructure [8].

Complement

In this scenario, packets are generated and injected into the network simultaneously and the router located in the first position inside the NoC sends packets to the router located in the last position inside the NoC; the router located in the second position inside the NoC sends packets to the router located in the next-to-last position, and so on. Basically each network router sends packets to the router located at its complementary position. Figure 4 illustrates an example of the complement traffic scenario.

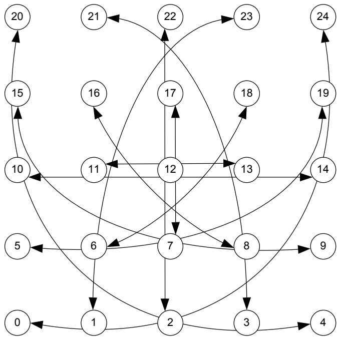

Figure 4. Complement traffic scenario - 5x5 2D mesh NoC [8].

B. Performance Evaluation

In this paper we analyze packet latency, throughput, and network occupancy as performance evaluations parameters which were achieved by RTL simulations of both NoC.

Latency analysis

Packets latency metric can be observed in different ways, as illustrated in Figure 5. The communication latencies presented here are not limited to the packets transmission delay into the NoC, but they also consider the packets that are delayed to enter into the NoC. Following we present some explanations to the differentiation of both transmission latencies, according to the distribution of packets injection and packets reception.

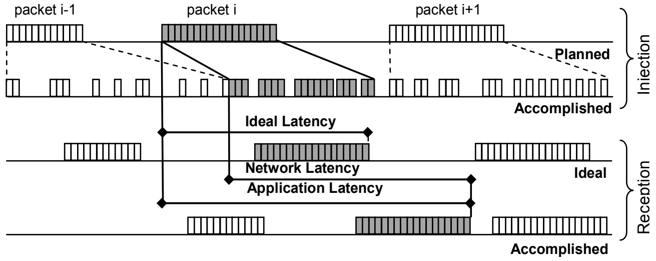

Figure 5. Communication latency metrics [8].

The planned injection is the moment that a packet is able to be injected into the NoC. In our simulation scenarios, all packets are specified in an input text file with its planned injection time. Accomplished injection considers the exact insertion timing of a packet into the NoC, which may be different from that defined in planned injection, due to the occurrence of contentions. The ideal reception represents the estimated time of packets delivery. The accomplished reception shows the real delivery time of packets at their destinations. Figure 5 shows distributions of such injection and reception scenarios. The ideal latency is the minimum number of clock cycles that a packet needs to reach its destination. This value is obtained from the difference between the planned injection time of the packet and the expected delivery time of the same packet.

In this paper, the concept of network latency is related to the transmitting delay of a packet from source to destination, which can be influenced by other packets competition for NoC resources (e.g. channels and storage queues). On the other hand, application latency expresses the time spent between the moment a packet is created by the application and the moment the packet is consumed by the target node. Application latency illustrates the most important impact on the ideal performance of a communication, since it is computed as the difference between the planned injection time of packets and their exact delivery moment at destination [8]. Both network and application latencies are assumed as performance metrics for comparison in this work.

Throughput analysis

Another parameter considered to evaluate NoCs performance was the packet throughput, computed by the amount of packets that leave the network in a given period of time. The packet throughput can be influenced by the buffer length, traffic scenario and other NoC resources competitions.

Network Occupancy

The network occupancy is the percentage of packets that are in the network in a given period of time. It is estimated, therefore, the percentage related to resources occupancy in NoC communication. In [9] the authors observed that an average occupancy above a threshold value (e.g. 80% of full buffer) and high rates of packets reception can reveal a difficulty in the packet transmission and congestion occurrence. In [8] the author concludes that the network occupancy increase leads to higher communication latency, contributing to network congestion.

IV. EXPERIMENTAL RESULTS

The task of defining a generic model for network performance evaluation is not that simple. The behavior of a NoC has considerable differences from one architecture to another, and from one application to another. This section presents the evaluation of network and application latencies, and network throughput and occupancy. All these parameters are evaluated and compared in both NoC architectures: Lasio 3D NoC and Hermes 2D NoC [6]. Both NoCs have buffer length of 2n flits, where n ranges from 1 to 10. In such comparisons, all-to-all (4032 packets) and complement (4096 packets) traffic scenarios were used with injection rate of 800 Mbps. The experiments resulted in 10 measurements for each traffic scenario, in each NoC topology.

To provide a fair comparison, both NoCs contains 64 tiles and routers in a square and cubic format: Hermes 2D NoC is 8x8 mesh and Lasio 3D NoC is 4x4x4 mesh. However, the majority of design parameters is the same and fixed for all experiments: packet size equal to 5 flits, each flit is 16 bits width; one virtual channel, and credit based control flow (where one flit can be transmitted in each clock cycle).

A. Network and Application Latencies

Performance evaluation is performed in this paper considering NoCs as "black boxes" with input and output rates. Therefore, the data for analysis are collected in the external network interfaces. When a packet reaches its destination router, the elapsed time is saved along with the packet data in an output file. Hence, the data packets generated in the NoC traffic scenarios for both NoCs are saved allowing the latency calculation for each individual packet. Each packet injected into the NoC contains 5 flits with the following contents:

1. The address of the packet destination router;
2. The size of the packet payload;
3. The address of the packet source router;
4. The input time of the packet in the network;
5. The elapsed time from the beginning of the simulation.

Figure 6 shows the influence of buffer length in NoC latency versus application latency for both 2D and 3D NoCs architectures with all-to-all traffic scenario. It is possible to notice that the increase of buffer length implies the decrease of application latency. However, with greater buffers more packets are injected into the NoC compromising the network latency. This phenomenon can be explained by the fact that the increase of buffers length enables a better packets distribution into the network, reducing contention and the payload flits reaching their target routers.

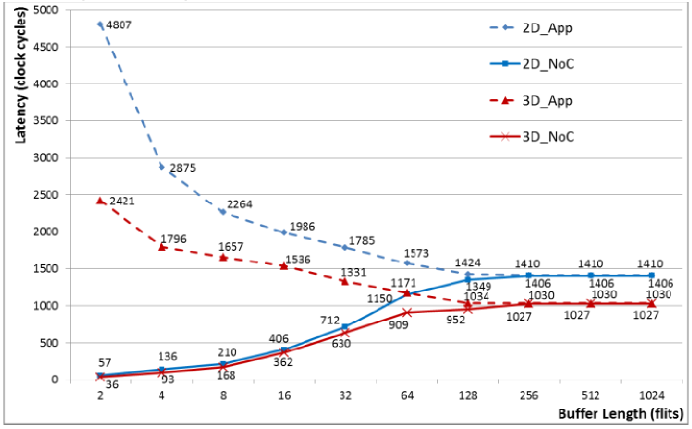

Figure 6. Network latency versus application latency.

The average packet size, packet injection rate, NoC topology, NoC size and NoC geometry influence on the "ideal buffer depth" for a given application considering its behavior. Remark that from 128-flit buffer, the difference between application and NoC latencies tends to zero in both 2D and 3D NoCs, which means that the latencies are no longer influenced by buffer length, and the use of bigger buffers is not justified. However, when applying an appropriate buffer length the application latency is reduced up to 3.4 times for 2D topologies (e.g. 2421/1034) and 2.3 times for 3D (e.g. 4807/1424).

Figure 7 demonstrates 3D NoC gains percentage over 2D NoC observing network and application latencies for all-to-all traffic scenario. Again, it is clear that there is no latency gain from 128-flits buffer. An explanation for this behavior is the 128-flit buffer does not reach its full capacity with the packet amount used in this experiment.

One of the 3D NoC strengths compared to 2D NoC is that this architecture allows reducing the number of hops that packets must pass through and, consequently, decreasing the network and application latencies, mainly due to the vertical connections between adjacent layers. For instance, experiments presented here pointed out, in average, 25% and 30% of latency reduction for network and for application, respectively.

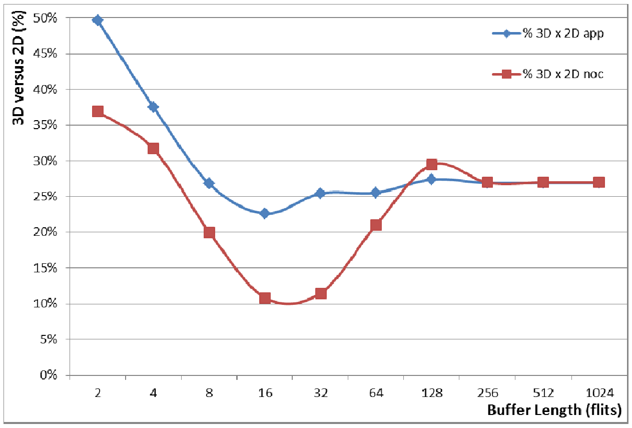

Figure 7. 3D NoC gains on network and application latencies.

B. *Throughput Analysis*

The purpose of this section is to evaluate and compare the throughput in both NoCs topologies (3D and 2D) according to the traffic scenarios implemented.

Figure 8 shows the average throughput (in flits per clock cycle) for both traffic scenarios. Examining the graphics note that Lasio 3D NoC has superior throughput when compared to the 2D NoC regardless traffic scenario and buffer length. It is also noticeable that throughput directly increases with buffer length until a given buffer length limit, which depends on traffic scenario and packet size. For instance, in the all-to-all traffic scenario with 5-flit packet size, the results have presented higher throughput until 128-flit buffer.

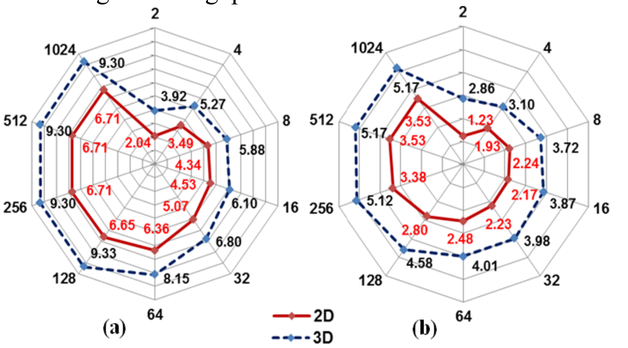

Figure 8. Average throughput for 2D and 3D NoCs – (a) All-to-all traffic; (b) Complement traffic

Figure 9 shows the 3D NoC gains percentage over the 2D NoC from the average throughput perspective, for the two traffic scenarios and all buffer length implemented.

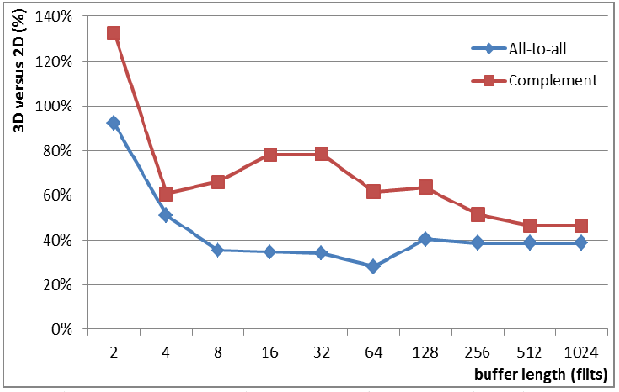

Figure 9. 3D NoC gain percentage for average throughput.

It is observable that the better buffer distribution and the increase of communication channels enables greater throughput to 3D NoCs. In average, experiments here presented pointed out gains of 56% when comparing the throughput on 3D NoC over 2D NoC (43.2% for all-to-all traffic scenario and 68.6% for complement traffic scenario). For smaller buffers this gain is more expressive. It happens because small buffers for proportionally large packets imply more routers occupied to transmit a single packet, and consequently, more contention of other packets. This problem is minimized in 3D NoC since, in average, it has more communication paths and less hops.

C. Network Occupancy

This section aims to evaluate in how much the buffer length and the NoC topology implies on buffers occupation.

Figure 10 contains four graphics showing network and application latencies of all packets for all-to-all traffic scenario with the latencies sorted in ascending order. This figure shows that:

- Buffer occupancy of 2D NoCs has a similar behavior compared to the one found on 3D NoCs, except by their magnitude. In fact, 3D NoC, as previously stated, always allows to minimize latency for the same buffer length - the sum of all packets latency shows a latency reduction, in average, 33.4% on network and 51.5% on application;
- For small buffers, just a few quantity of packets has high network latency (e.g. for 2-flit buffer of 3D NoC, 1% of packets has latency between 150 and 430 clock cycles, and 90% has latency slower than 67 clock cycles). On the other hand, greater buffers minimize this critical variation and the packets latency distribution is practically linear. The correspondent application latency has also a linear latency distribution, independent of buffers length.

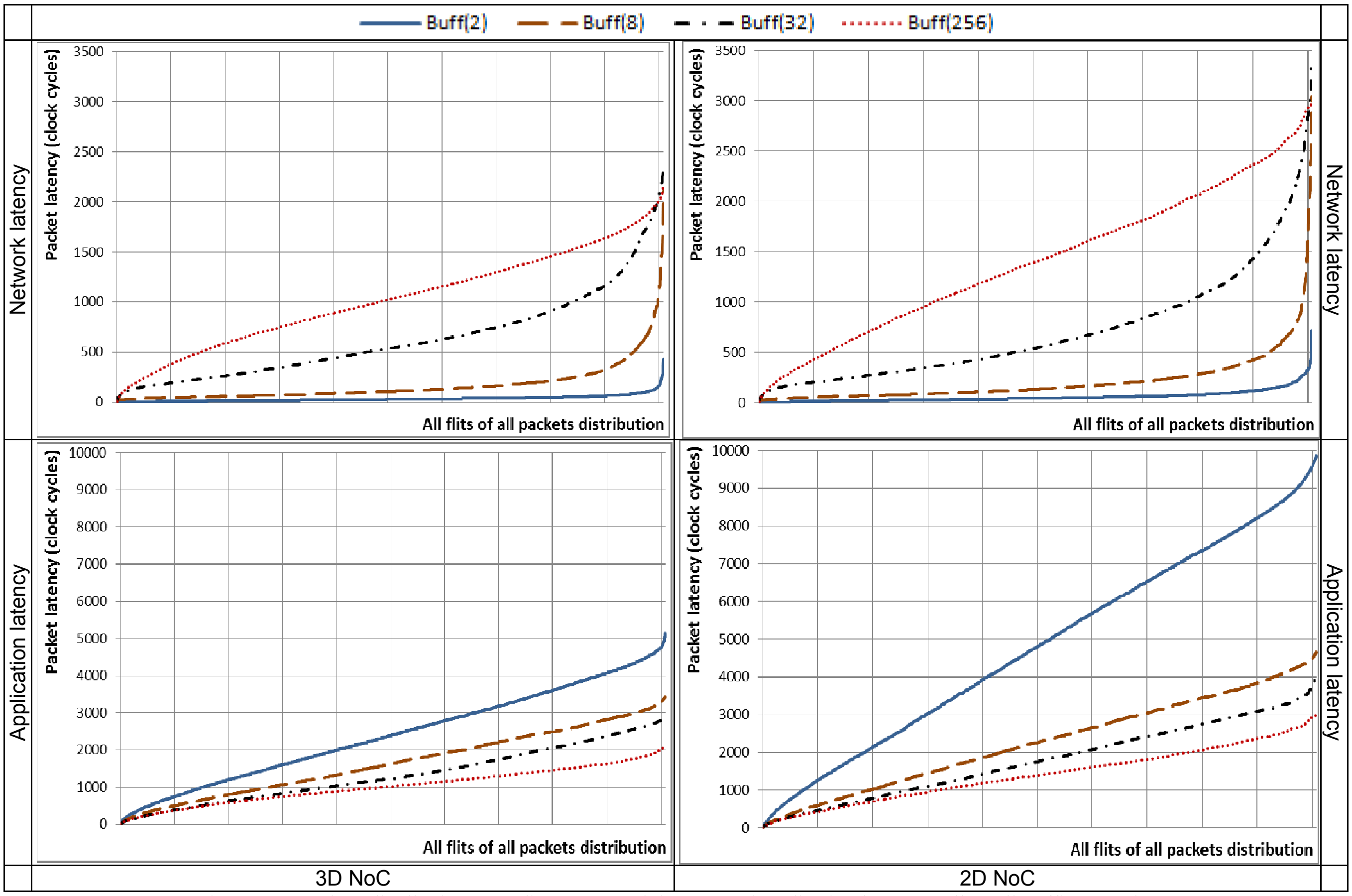

Figure 10. Packets distribution according to network latency and application latency – all-to-all traffic scenario.

Figure 11 illustrates four graphics for the analysis of the percentage rate and the number of flits remaining inside the network (occupying buffers and links), during all simulation period. The graphics were captured using all-to-all traffic scenario and four different buffer lengths, comparing both NoC topologies. Besides, for all configurations the same quantity of data was injected into the NoC.

Analyzing the graphics it is possible to determine that when a larger buffer is used, shorter is the time period in which the network is occupied, which means that all packets are going to reach their destiny before. This behavior is noticed in both Lasio and Hermes NoCs.

Both 3D and 2D NoCs differ on the time in which the last flit will be delivered at its destination router, as well as the percentage rate of the network occupancy. In the 3D NoC with buffer length equal to 256-flit positions, the last flit is delivered after 2500 clock cycles, while in the 2D NoC the last flit is only delivered after 3300 clock cycles. This means that in the 3D NoC the last flit is delivered approximately 800 clock cycles before. The 3D NoC also reaches better results when smaller buffers are used, e.g. in the 3D NoC with buffer length equals to 4-flits positions, the transmission ends approximately 2000 clock cycles before. Another advantage of 3D NoC is that the network occupancy for small buffers reaches approximately 72% at most, while in the 2D NoC reaches around 92%. This can be explained by the two extra ports that increase 3D throughput reducing the NoC occupation. However, these numbers will decrease as the buffer length increases.

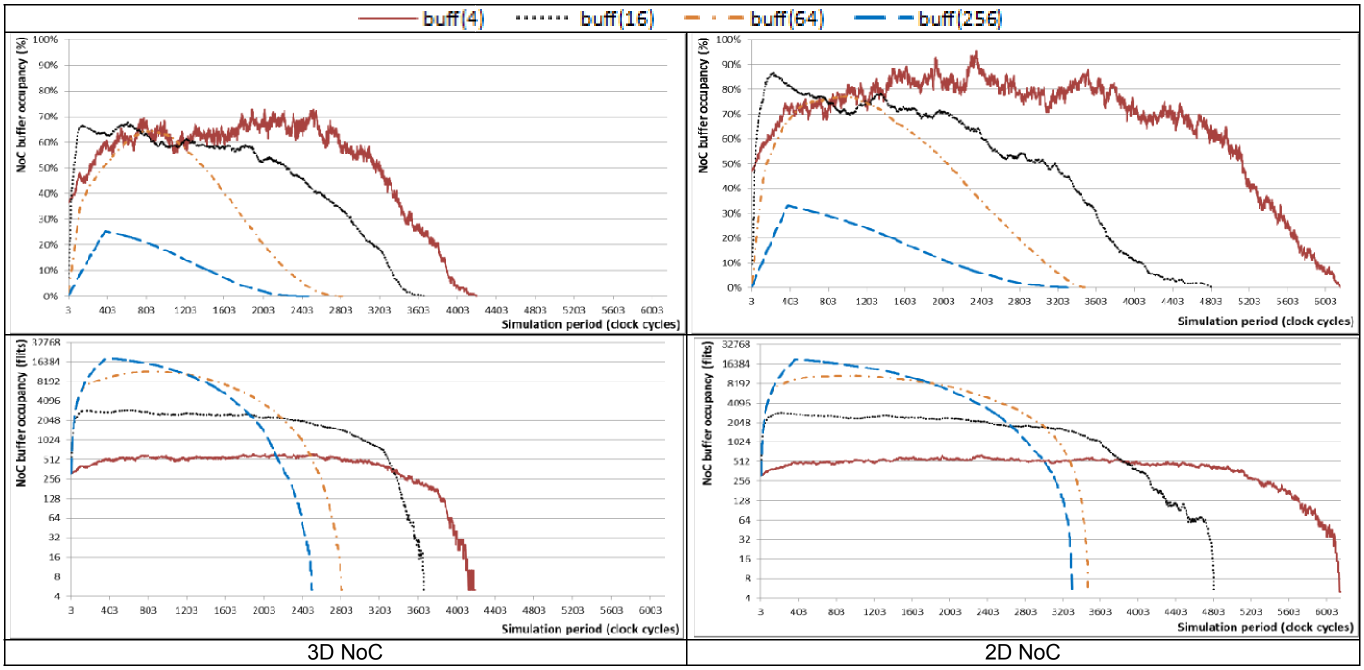

Figure 11. Buffer occupancy during the application execution - all-to-all traffic scenario.

V. CONCLUSION

The frequent increase in the number of PEs on SoCs requires more efficient communication infrastructures, which may be provided by a network with 3D technology. This work proposes 3D NoC (Lasio), whose implementation is based on 2D NoC (Hermes) [6]. Lasio implements all architectural features of Hermes, except the ones that have to be changed to perform the improvement from 2D to 3D technology. This procedure enabled a fair comparison between NoC topologies.

The paper explored comparisons of NoC and application latencies, throughput and NoC occupancy, according to several buffer depths and two traffic scenarios of 3D and 2D routers implementation enabling to evaluate architectural features.

For the selected set of experiments using all-to-all traffic scenario, 3D NoC implementation always minimized the latencies of the packets when compared to 2D NoC implementation. For instance, the average network latency minimization is 25% and the average application latency minimization is 30%. In addition, the paper explored the influence of the buffer length on network and on application latencies, showing that when applying an appropriate buffer length the application latency is reduced up to 3.4 and 2.3 times for the selected 2D and 3D topologies, respectively.

The 3D NoC also had better results than the 2D NoC according to throughput packets. Using all-to-all traffic scenario, the Lasio throughput increased approximately 43.2% and for complement traffic scenario the increase was around 68.6%. For such evaluations, all buffer lengths (2, 4, 8, 16, 32, 64, 128, 256, 512, 1024) implemented were taken into account. This throughput enables to finish the communication of 3D NoC much faster than the ones on 2D NoC, independently of the traffic scenario and the buffer length. Moreover, 3D integrated systems hold promises to significantly reduce cost factors (power consumption and area usage) and improve communication performance due to the considerable reduction in the length of global interconnections. In general, 3D NoC structures provide better performance on several aspects when compared to 2D NoC architectures.

ACKNOWLEDGMENT

The author Dr. Thais Webber receives grant from CAPES-Brazilian Ministry of Education (PNPD project 058792/2010).

REFERENCES

[1] V. Pavlidis, E. G. Friedman. **3-D Topologies for Networks-on-Chip**. *IEEE Transactions on VLSI Systems*, v.15, n.10, pp. 1081-1090, 2007.

[2] L. Carloni, P. Pande, X. Yuan. **Networks-on-Chip in Emerging Interconnect Paradigms: Advantages and Challenges**. *International Symposium on Networks-on-Chip*, pp. 93-102, 2009.

[3] R. J. Gutmann et al. **Three-dimensional (3D) ICs: A technology platform for integrated systems and opportunities for new polymeric adhesives**. *Conference on Polymers and Adhesives in Microelectronics and Photonics*, pp. 173–180, 2001.

[4] V. Pavlidis, E. Friedman. **Interconnect delay minimization through interlayer via placement in 3-D ICs**. *ACM GLS VLSI*, pp. 20–25, 2005.

[5] A. Sheibanyrad, F. Pétrot, A. Jantsch. **3D Integration for NoC-based SoC Architectures**. *Springer*, pp. 278, 2011.

[6] F. Moraes et al. **HERMES: an infrastructure for low area overhead packet-switching networks on chip**. *The VLSI Journal Integration*, v.38, n.1, pp. 69–93, 2004.

[7] A. B. Ahmed et al. **Architecture and Design of Efficient Network-on-Chip (3D NoC) for Custom Multicore SoC**. *Conf. on Broadband, Wireless Computing, Communication and Application*, pp. 67-73, 2010.

[8] E. Moreno et al. **Arbitration and Routing Impact on NoC Design.** *International Symp. on Rapid System Prototyping,* pp.193-198, 2011.

[9] U. Ogras, R. Marculescu. **Prediction-Based Flow Control for Network-on-Chip Traffic**. *DAC*, pp. 839-844, 2006.