Partial-LastZ: An Optimized Hybridization Technique for 3D NoC Architecture Enabling Adaptive Inter-Layer Communication

Amir-Mohammad Rahmani[1,2], Pasi Liljeberg[1], Juha Plosila[1],
Ka Lok Man[3,4,5], Youngmin Kim[6], and Hannu Tenhunen[1]
[1]Embedded Computer and Electronic Systems Lab., Department of IT, University of Turku, Finland
[2]Turku Centre for Computer Science (TUCS), Finland [3]Xi'an Jiaotong-Liverpool University, China
[4]ASIC LAB, Myongji University, South Korea [5]Baltic Institute of Advanced Technology, Lithuania
[6]UNIST (Ulsan National Institute of Science and Technology), South Korea
Email: {amir.rahmani, pasi.liljeberg, juha.plosila, hannu.tenhunen}@utu.fi,
ka.man@xjtlu.edu.cn, youngmin@unist.ac.kr

***Abstract*—Three-dimensional (3D) integration offers greater device integration, reduced signal delay and reduced interconnect power. It also provides greater design flexibility by allowing heterogeneous integration. Stacked mesh 3D NoC architecture was proposed to take advantage of the intrinsic capability of reducing the wire length in 3D ICs. However, this architecture still exacerbates the on-chip power density and router cost. In this paper, we propose a novel hybridization scheme for inter-layer communication using efficient 5-input routers to enhance the overall system power, performance, and area characteristics of the existing Hybrid NoC-Bus 3D mesh architecture. By defining a rule for routing algorithms called *LastZ*, the proposed area-efficient architecture decreases the overall average hop count of a NoC-based system compared to the existing architectures. We further improve this design by proposing partial-*LastZ*-based 3D NoC-bus hybrid architecture to provide adaptivity for implementing congestion-aware and fault-tolerant inter-layer routing algorithms. Extensive quantitative experiments demonstrate up to 16% performance improvement compared to the full *LastZ*-based 3D NoC-bus hybrid architecture and around 20% area reduction compared to the typical hybrid NoC-Bus 3D mesh architecture.**

***Index Terms*—3D Networks-on-Chip; Routing Algorithm; 3D ICs; NoC-Bus Hybridization**

I. Introduction

Network-on-Chip (NoC) is a general concept, proposed for complex on-chip communications because of scalability, better throughput and reduced power consumption [1]. However, increasing the number of cores over a 2D plane is not efficient enough due to long interconnects. The advent of 3D silicon integration technology has opened a new horizon for new on-chip interconnect design innovations. In 3D integration technologies, multiple layers of active devices are stacked above each other and vertically interconnected using Through-Silicon Vias (TSVs) [2]. The comparison of 2D and 3D NoC architectures show that, 3D NoCs deliver better system performance with significantly lower energy per packet, as compared to the 2D implementations due to increased package density and shorter wires [3].

The straightforward extension of popular planar 2D NoC structure is 3D Symmetric NoC created by simply adding two additional physical ports to each router; one for Up and one for Down [4]. Despite simplicity, this architecture has two major inherent drawbacks. Firstly, it does not exploit the beneficial feature of a negligible inter-wafer distance in 3D chips, because in this architecture, inter-layer and intra-layer hops are indistinguishable. Secondly, a considerably larger crossbar is required as a result of two extra ports [5].

The stacked mesh 3D NoC (Hybrid NoC-Bus 3D mesh) architecture presented in [6] is a hybrid architecture between the packet switched network and the bus architecture to overcome the mentioned 3D Symmetric NoC challenges. It integrates the multiple layers of 2D mesh networks by connecting them with a bus spanning the entire vertical distance of the chip. As the inter-layer distance for 3D ICs is small, the bus length will also be smaller. This makes the bus suitable for inter-layer communication in vertical direction. By using the stacked mesh architecture, six-port router is required instead of seven ports for typical 3D NoC router and vertical communication is just one hop away to any destination layer. However, the additional input port still imposes considerable extra logic to a NoC router, especially when complex routers with support of Virtual Channels (VCs) and load management schemes are required.

In this paper, an efficient hybridization scheme for inter-layer communication is presented in order to enhance the overall system power, performance, and area characteristics of the existing Hybrid NoC-Bus mesh architecture. By defining a rule for routing algorithms called *LastZ*, the proposed area-efficient architecture decreases the overall average hop count of a NoC-based system compared to the existing architectures. Based on the proposed hybridization scheme, we present a low-power and high-performance 3D NoC architecture which enables congestion-aware and fault-tolerant inter-layer communication.

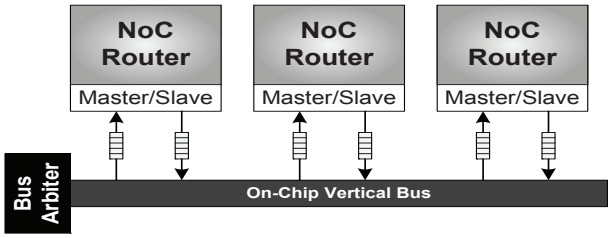

Fig. 1. Side view of the conventional 3D Hybrid NoC-Bus mesh architecture

II. LastZ Routing Policy

As discussed in Section I, a 6×6 router has many disadvantages over its 5×5 counterpart. For instance, based on the reported results in [5] for 90nm CMOS technology, a 6-port router incurs around 36% area and 20% power overheads compared to a 5-port one. On the other hand, the packet delay to cross the router increases because more input channels compete to get access to the target output port. The larger crossbar also increases the router critical path and thereby reduces the router maximum operating frequency.

Our investigation shows that a straightforward hybridization of two different communication media (i.e., NoC and Bus) without considering their intrinsical characteristics is not an efficient strategy. Fig. 1 shows the conventional hybridization style of a Hybrid NoC-Bus mesh-based system. In this architecture, stacked routers in different layers are connected to a vertical bus. The routers are able to serve as either a master or a slave depending on the arbitration decision. Surprisingly, as will be shown later, just by following one basic rule in routing algorithm policy, it is possible to remove one input port and substitute 6×6 routers with 5×6 ones.

A. LastZ Rule

In 3D NoC, the packet routing process is classified into two different categories: intra-layer routing and inter-layer routing. For the 3D Hybrid NoC-Bus mesh architecture, the intra-layer packet routing is multi-hop because traditional NoC architecture is utilized for communication. In multi-hop communication, packet routing plays a crucial role because there are many minimal or non-minimal paths to send a packet from a source to a destination. In contrast, for inter-layer communication, the 3D Hybrid NoC-Bus mesh architecture benefits from bus-based one-hop communication. In this work, we define a rule which we call *LastZ*.

Definition 1 (LastZ). A 3D routing algorithm is *LastZ*-based if the intra-layer routing process is completed before the inter-layer routing. In other words, in the *LastZ*-based routing algorithm, when a node N_{source} sends a flit to a node $N_{destination}$, the flit will first travel along the X or Y direction (statically or adaptively) in N_{source} dimension until $Flit_{xy}$=$Pillar_{xy}$, then it will traverse the last hop in the *Z* direction.

As will be shown later, this rule is astonishingly beneficial to improve system characteristics. It is noteworthy that this rule does not force significant limitations to a routing algorithm owing to the one-hop bus-based vertical communication. Many of recently proposed architectures benefit from routing algorithms being by default *LastZ*-based such as [7][8][9][10].

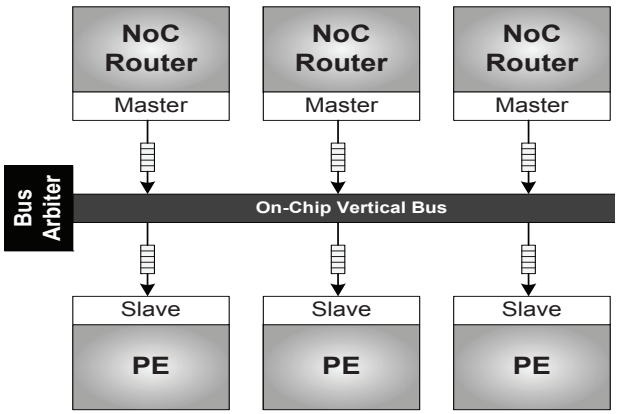

Fig. 2. Side view of the proposed 3D Hybrid NoC-Bus mesh architecture

III. LastZ-Based Hybridization Architecture

Based on the defined *LastZ* rule, assume that a packet, after complete intra-layer routing, has reached the destination pillar (vertical bus). In this case, it is obvious that one of the connected routers to the vertical bus is the last router (hop) for delivering the packet to the target processing element (PE). Based on the fact that the destination is already known, it is not wise to send the packet again to the respective router for the routing decision making. Instead, it is more efficient to deliver the packet to the connected PE directly.

The explained scenario is the motivation to propose a new hybridization scheme for connecting components to a vertical bus. The proposed architecture is shown in Fig. 2. As can be seen in the figure, it is practical to establish a more efficient inter-layer communication scheme without adding any extra workload and hardware to bus arbiters. Based on the proposed hybridization architecture, routers just serve as masters to initiate the transaction and PEs play the slave role and via the intermediate buffers directly receive their own packets. As can be seen in the figure, the intermediate input buffer which was used as an interface between a router and a bus, in this architecture connects a PE directly to the vertical bus. Bypassing routers enables a 3D NoC to utilize a 5×6 router instead of a larger 6-input port router.

IV. Partial-LastZ-Based 3D NoC Architecture

The *LastZ*-based hybridization offers many advantages such as low-cost and high-speed routers, fast intra-layer and inter-layer packet transmission, reduced power consumption, and high-throughput network. However, this architecture suffers from inability to support adaptivity for inter-layer communication. More precisely, in a fully *LastZ*-based network, the vertical hop must be taken as the last hop. This rule leads to a limitation that the routing adaptivity is restricted to the intra-layer routing. Therefore, if a totally adaptive routing algorithm

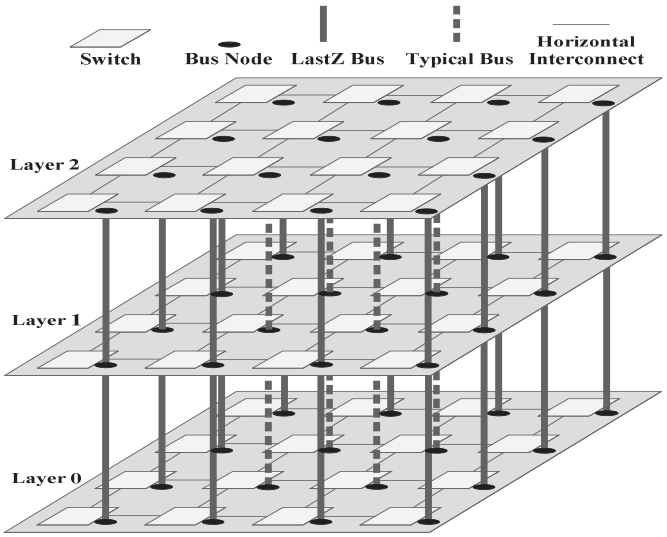

Fig. 3. Partial-*LastZ*-based 4×4×3 NoC architecture with 4 typical buses

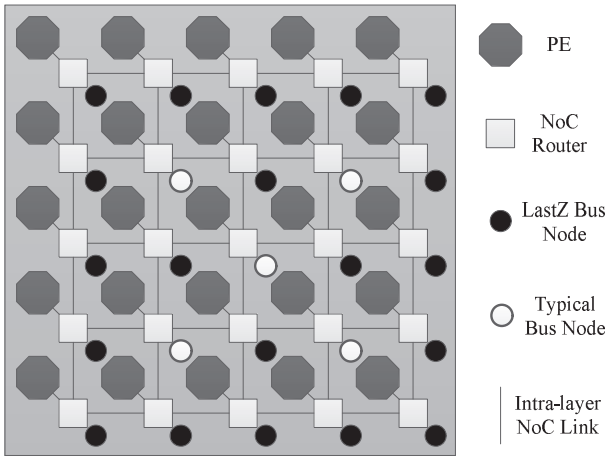

Fig. 4. Top view of 5×5×3 partial *LastZ*-based 3D NoC-bus hybrid architecture with 5 typical vertical buses

is desired in order to balance the load across all layers or to bypass a faulty vertical link, the *LastZ*-based routing will not be efficient. We address this issue by presenting a partial-*LastZ*-based 3D NoC architecture.

As shown in Fig. 3, the partial-*LastZ*-based architecture is the combination of the typical Hybrid NoC-Bus 3D mesh and the *LastZ*-Based 3D NoC architectures. In this architecture, a number of vertical buses are designated to follow the typical vertical bus architecture, while others are still based on the proposed low-cost *LastZ*-based hybridization scheme. Consequently, inter-layer networks consist of two types of routers: 6×6 routers connected to typical buses and 5×6 routers connected to *LastZ*-based buses. The partial-*LastZ*-based architecture has the advantage of adaptivity to handle congestions and faulty situations while enhancing the network characteristics in terms of performance, power consumption and area footprint.

In order to implement the routing mechanism for this architecture, the information of the typical bus nodes is stored in the network interface of each tile. The system uses the default *LastZ*-based routing algorithm in the normal situations. In the case of occurring congestion in particular layers or existence of faulty vertical buses, the information stored in the network interfaces can be used to find the closest alternative path to reach the destination layer.

V. Experimental Results

To demonstrate the better performance, power, and area characteristics of the proposed 3D NoC, a cycle-accurate NoC simulation environment was implemented in HDL. The full *LastZ*-based 3D NoC-bus hybrid architecture [11] and partial *LastZ*-based 3D NoC-bus hybrid architecture were analyzed for a synthetic traffic pattern. 5×5×3 meshes and packets with a length of eight-flits were used for the simulations. The on-chip network considered for experiment is formed by a typical state-of-the-art router structure including buffers, a routing unit, a switch allocator, VC allocators and a crossbar. For the full *LastZ*-based architecture, all the routers have 5 input and 6 output ports. For the partial *LastZ*-based architecture, we used the typical bus architecture with 6×6 routers for 5 vertical pillars while the other 20 pillars utilize the *LastZ*-based hybridization scheme as shown in Fig. 4. We chose the location of typical pillars in such a way that the maximum distance from each source node to the closet typical pillar is not more than 2 hops. In addition, since traffic congestion commonly occurs at the center of the mesh, the pillars being closer to the center are more suitable options to provide inter-layer adaptivity.

To perform the simulations under synthetic traffic profiles, an unbalanced traffic generation scenario was used. In this scenario, 40%, 35%, and 25% of the total network traffic is generated by the nodes located in Layer0, Layer1, and Layer2, respectively. Each node follows the uniform traffic pattern to distribute packets throughout the network. In the uniform traffic pattern, a node sends a packet to other nodes with an equal probability. The packet latencies were averaged over 50,000 packets. Latencies were not collected for the first 5,000 cycles to allow the network to stabilize. It was assumed the buffer size of each FIFO was eight flits, and the data width was set to 64 bits.

For the full *LastZ*-based 3D NoC-bus hybrid architecture, *(DyXY)Z* [11][12] wormhole routing algorithm was used, while we utilized *(DyXY)Z(DyXY)* routing algorithm for the partial *LastZ*-based 3D NoC-bus hybrid architecture. Because this routing algorithm is not deadlock-free, we used routers with two virtual channels per input port. The packet latency versus average packet arrival rate for different architectures under uniform traffic profile are shown in Fig. 5. It can be observed for the mentioned scenario that the network with the partial *LastZ*-based architecture saturates at higher injection rates and always offers reduced average packet latency compared to the full *LastZ*-based 3D NoC-bus hybrid architecture. The reason being that, the partial *LastZ*-based architecture

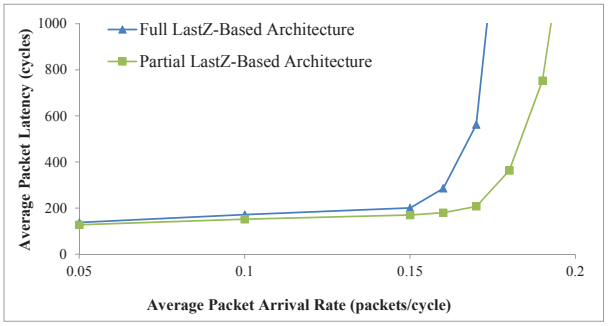

Fig. 5. Latency versus average packet arrival rate on a 5×5×3 mesh under uniform traffic profile

TABLE I
HARDWARE IMPLEMENTATION DETAILS

Component	**Area** (μm^2)
5×5 NoC Router Without VC	33678
5×6 *LastZ*-based NoC Router Without VC	35230
6×6 NoC-Bus Hybrid Router Without VC	42764
7×7 Symmetric NoC Router Without VC	58973
5×5 NoC Router With VC	70675
5×6 *LastZ*-based NoC Router With VC	73013
6×6 NoC-Bus Hybrid Router With VC	91951
7×7 Symmetric NoC Router With VC	122779

can balance the load distribution among all layers better than the full *LastZ*-based architecture due to the offered routing adaptivity.

The area of the different routers was computed once synthesized on CMOS 65nm *LPLVT STMicroelectronics* standard cells using Synopsys Design Compiler. To observe the area savings of more complex routers, we synthesized routers supporting virtual channels as well. For these routers, we set the number of virtual channels to 2. The layout area of a conventional 2D NoC router, the proposed *LastZ*-based NoC router, a conventional 3D NoC-Bus Hybrid router, a 3D Symmetric NoC router and the wrapper are listed in Table I. For all the routers, the data width and buffer depth were set to 32 bits and 8 slots, respectively. The figures given in the table reveal that compared to a conventional 3D NoC-Bus Hybrid router, the area savings for the proposed *LastZ-based* router is around 18% and 21% for without- and with-VC implementations, respectively. For more complex routers supporting a large number of VCs, complex VC management techniques [13][14], and wider data width, it is expected to have more area savings.

VI. Conclusion and Future Work

In this paper an efficient hybridization scheme was proposed to address the naive and straightforward hybridization between NoC and bus media in the 3D NoC-Bus Hybrid Mesh architecture. The hybridization mechanism benefiting from a rule called *LastZ*, enables low-cost inter-layer communication architecture. In order to provide routing adaptivity for the presented scheme, we proposed partial-*LastZ*-based 3D NoC-bus hybrid architecture. This architecture utilizes a combination of typical and *LastZ*-based bus architectures for inter-layer communication. Our extensive simulations showed that compared to the full *LastZ*-based 3D NoC-bus hybrid architecture, the partial-*LastZ*-based architecture achieves significant performance, and area improvements. In the future, our work will be extended by performing a comprehensive simulation to estimate the system power and measure NoC performance under realistic traces.

Acknowledgment

The authors wish to acknowledge the financial support by the Academy of Finland, Ulla Tuominen Foundation and Nokia Foundation during the course of this project.

References

[1] A. Jantsch and H. Tenhunen, *Networks on Chip*. Kluwer Academic Publishers, 2003.

[2] R. Patti, "Three-Dimensional Integrated Circuits and the Future of System-on-Chip Designs," *Proceedings of the IEEE*, vol. 94, no. 6, pp. 1214–1224, 2006.

[3] A.-M. Rahmani, K. Latif, P. Liljeberg, J. Plosila, and H. Tenhunen, "Research and practices on 3D networks-on-chip architectures," in *Proceedings of the IEEE International NORCHIP Conference*, 2010, pp. 1–6.

[4] A.-M. Rahmani, K. Vaddina, K. Latif, P. Liljeberg, J. Plosila, and H. Tenhunen, "Generic Monitoring and Management Infrastructure for 3D NoC-Bus Hybrid Architectures," in *Proceedings of the Sixth IEEE/ACM International Symposium on Networks on Chip*, 2012, pp. 177–184.

[5] J. Kim, C. Nicopoulos, D. Park, R. Das, Y. Xie, V. Narayanan, M. S. Yousif, and C. R. Das, "A novel dimensionally-decomposed router for on-chip communication in 3D architectures," in *Proceedings of the ACM international symposium on Computer architecture*, 2007, pp. 138–149.

[6] F. Li, C. Nicopoulos, T. Richardson, X. Yuan, V. Narayanan, and M. Kandemir, "Design and Management of 3D Chip Multiprocessors Using Network-in-Memory," in *Proceedings of the 33rd International Symposium on Computer Architecture*, 2006, pp. 130–141.

[7] Y. Xu, Y. Du, B. Zhao, X. Zhou, Y. Zhang, and J. Yang, "A low-radix and low-diameter 3D interconnection network design," in *Proceedings of the IEEE 15th International Symposium on High Performance Computer Architecture*, 2009, pp. 30–42.

[8] Y. Qian, Z. Lu, and W. Dou, "From 2D to 3D NoCs: A case study on worst-case communication performance," in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design - Digest of Technical Papers*, 2009, pp. 555–562.

[9] A.-M. Rahmani, P. Liljeberg, J. Plosila, and H. Tenhunen, "BBVC-3D-NoC: An Efficient 3D NoC Architecture Using Bidirectional Bisynchronous Vertical Channels," in *Proceedings of the IEEE Computer Society Annual Symposium on VLSI*, 2010, pp. 452–453.

[10] J. Jiao, Y. Fu, T. Liu, H. Wang, X. Han, and J. Wang, "Performance analysis and optimization for homogenous multi-core system based on 3D Torus Network on Chip," in *Proceedings of the 8th IEEE International NEWCAS Conference*, 2010, pp. 313–316.

[11] A.-M. Rahmani, P. Liljeberg, J. Plosila, and H. Tenhunen, "Exploring a Low-Cost and Power-Efficient Hybridization Technique for 3D NoC-Bus Hybrid Architecture using LastZ-Based Routing Algorithms," *Journal of Low Power Electronics*, vol. 8, no. 4, 2012.

[12] M. Li, Q.-A. Zeng, and W.-B. Jone, "DyXY: a proximity congestion-aware deadlock-free dynamic routing method for network on chip," in *Proceedings of the 43rd annual Design Automation Conference*, 2006, pp. 849–852.

[13] C. Nicopoulos, D. Park, J. Kim, N. Vijaykrishnan, M. Yousif, and C. Das, "ViChaR: A Dynamic Virtual Channel Regulator for Network-on-Chip Routers," in *Proceedings of the 39th Annual IEEE/ACM International Symposium on Microarchitecture*, 2006, pp. 333–346.

[14] A.-M. Rahmani, M. Daneshtalab, A. Afzali-Kusha, and M. Pedram, "Forecasting-Based Dynamic Virtual Channel Management for Power Reduction in Network-on-Chips," *Journal of Low Power Electronics*, vol. 5, no. 3, pp. 385–395, 2009.