International Journal of Electrical, Electronic and Communication Sciences
ISSN: 2517-9438
Vol:4, No:9, 2010

Dual-Link Hierarchical Cluster-Based Interconnect Architecture for 3D Network on Chip

Guang Sun, Yong Li, Yuanyuan Zhang, Shijun Lin, Li Su, Depeng Jin and Lieguang zeng

Abstract—Network on Chip (NoC) has emerged as a promising on chip communication infrastructure. Three Dimensional Integrate Circuit (3D IC) provides small interconnection length between layers and the interconnect scalability in the third dimension, which can further improve the performance of NoC. Therefore, in this paper, a hierarchical cluster-based interconnect architecture is merged with the 3D IC. This interconnect architecture significantly reduces the number of long wires. Since this architecture only has approximately a quarter of routers in 3D mesh-based architecture, the average number of hops is smaller, which leads to lower latency and higher throughput. Moreover, smaller number of routers decreases the area overhead. Meanwhile, some dual links are inserted into the bottlenecks of communication to improve the performance of NoC. Simulation results demonstrate our theoretical analysis and show the advantages of our proposed architecture in latency, throughput and area, when compared with 3D mesh-based architecture.

Keywords—Network on Chip (NoC), interconnect architecture, performance, area, Three Dimensional Integrate Circuit (3D IC).

I. Introduction

WITH the development of the semiconductor technology, large quantities of transistors are available on a single chip, which allows designers to integrate numerous processors together with large amounts of embedded memory [1][2]. In order to alleviate the complex communication issues which occur as the number of on-chip components increases, Network on Chip (NoC) architecture has been recently proposed as a promising communication paradigm to replace global interconnects [3][4][5]. NoC provides lower power consumption and better performance, flexibility and scalability compared to previous solutions for on-chip communication [3][6][7].

The ability of the network to efficiently disseminate information depends largely on the underlying topology architecture [3]. The simplicity and regularity of grid structures make design approaches based on such a modular topology (e.g., mesh and torus) very attractive [3]. High radix networks like the flattened butterfly [8] reduce latency and power by reducing the number of intermediate routers. However, they increase the number of long wires. Three Dimensional Integrate Circuit (3D IC) emerges as an attractive option, for the reduction of interconnection length and the added interconnect scalability in the third dimension offer an opportunity to further improve the performance of NoC. As show in Fig.1, 3D mesh-based NoC provides the interconnect scalability in the third dimension. Moreover, the length of the through-via interconnection between layers ranges from $5\ \mu m$ to $50\ \mu m$ [3], which is much smaller than the intra-layer wiring length.

Guang Sun is with State Key Laboratory on Microwave and Digital Communications, Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China, Email: sung08@mails.thu.edu.cn, jindp@mail.tsinghua.edu.cn.

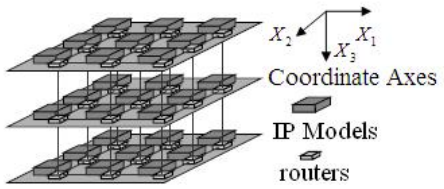


Fig. 1. The interconnect architecture of the 3D mesh-based NoC.

In this paper, a hierarchical cluster-based interconnect architecture is merged with the 3D IC. Compared with 3D mesh-based topology, this topology reduces the number of intermediate routers and leads to lower latency and area. Moreover, the added interconnect scalability in the third dimension and the small interconnection length between layers reduce the number of long wires. Meanwhile, in order to improve the throughput of the network, we insert some dual links into the bottlenecks of communication.

The rest of this paper is organized as follows. Section II describes our proposed interconnect architecture. Section III gives the performance analysis. In Section IV we show the simulation results compared with 3D mesh-based topology. Finally, in Section V we draw conclusions.

II. Proposed Interconnect Architecture

In this section, we first describe the hierarchical cluster-based topology for 2D NoC, and then expand it to 3D NoC.

A. Hierarchical Cluster-Based Topology for 2D NoC

The hierarchical cluster-based interconnect architecture for 2D NoC is showed in Fig. 2. Each local router is connected with four IPs and meanwhile each local router is connected to higher hierarchy router, namely global router. Thus, the communications from local router to global router become the bottlenecks of the whole network, and we insert dual links to improve the throughput. This topology contains 4^n Intelligence Properties (IPs) and approximately 4^{n-1} routers, where n is the number of the hierarchies. Therefore, there are small number of intermediate routers in this topology. However, the interconnect length from the local router to the global router gets longer when the number of IPs increasing.

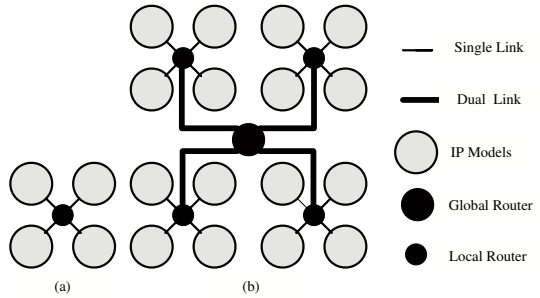


Fig. 2. The Dual-Link Hierarchical Cluster-Based Topology for 2D NoC with different number of IPs. (a)4 IPs (b) 16 IPs

B. Hierarchical Cluster-Based Topology for 3D NoC

Fig. 3 shows the hierarchical cluster-based interconnect architecture for 3D NoC. In this topology, each local router not only connects to the global router in the intra-layer, but also connects to the directly upper and below local routers in the adjacent layers. Moreover, we insert dual links into these connections due to the bottlenecks of communication in the network. Different with the situation in 2D NoC, the small interconnection length between layers in 3D NoC increases the interconnect scalability in the third dimension, reduces the number of long wires and improves the performance of network. The added interconnect scalability in the third dimension leads to small number of IPs in each layer (usually 16 IPs). Therefore, the length of long wires is acceptant. Meanwhile, in order to reduce the impact of long wires on performance, repeaters and pipeline registers can be inserted into the long wires.

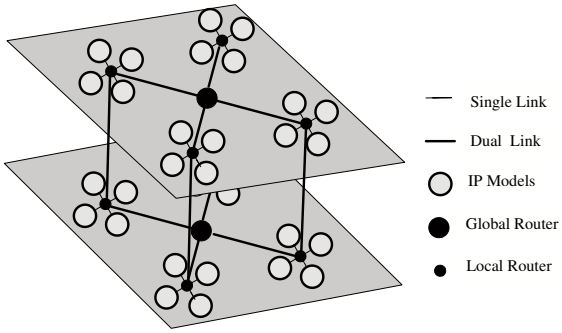


Fig. 3. The Dual-Link Hierarchical Cluster-Based Topology for 3D NoC

III. Performance Analysis

With the knowledge of the proposed interconnect architecture, next, we will analyze its performance. First, we calculate and compare the average number of hops in our proposed topology and 3D mesh-based topology. The decrease of the average number of hops usually leads to a decrease of average latency and an increase of throughput. And then we discuss the impact of multi-link on performance (e.g. latency and throughput) and area.

A. Average Number of Hops

Intuitively, our proposed topology has smaller average number of hops than 3D mesh-based topology, because our topology only has approximately a quarter of routers in 3D mesh-based topology. Next, we will calculate and compare the average number of hops in the two topologies in which each layer has 16 IPs (4×4).

In our discussion, it is assumed that the destination addresses of the generated messages are uniformly distributed across all of the IP cores and meanwhile each IP core doesn't send messages to itself.

In our proposed topology, the close form expression of the average number of hops, denoted by H_{our}, is given by,

$$H_{our} = \frac{16k^2 + 72k - 16}{3(16k - 1)} \tag{1}$$

where k is the number of layers in 3D NoC. The proof of equation (1) is given in the Appendix A.

In 3D mesh-based topology, the average number of hops, denoted by $H_{3D-mesh}$, is given by [9]

$$H_{3D-mesh} = \frac{16k^2 + 120k - 16}{3(16k - 1)} \tag{2}$$

Compared with 3D mesh-based topology, the average number of hops in our topology is smaller, which is showed in Fig.4. Moreover, the decrement, denoted by $H_{3Dmesh-our}$, is given by

$$H_{3Dmesh-our} = H_{3D-mesh} - H_{our} = \frac{16k}{16k - 1} \tag{3}$$

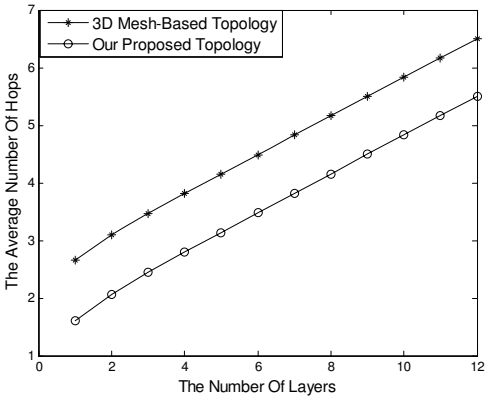


Fig. 4. The average number of hops in our proposed topology and 3D mesh-based topology

B. Single Link Versus Multiple Links

One of the key differences between on chip and inter-chip interconnects is that there are more wire resources on chip, while inter-chip connections are normally limited by available chip IO pins [10]. One way is to increase the width of the links in NoC to improve performance [11]. we explore another option of increasing the number of links in the bottlenecks of

the communication. The difference between multi-link architecture and virtual channels is that virtual channels increase the number of buffers and utilization of links, while multi-link architecture increases the number of connecting links and then improves the bandwidth of communication. Therefore, inserting some links into the bottlenecks of the communication can improve the performance (e.g. latency and throughput) efficiently. However, multi-link architecture increases the area overhead. For the tradeoff between performance and area, we only insert some dual links in the bottlenecks of communication in NoC.

IV. Performance Evaluation

In order to evaluate the performance of our proposed interconnect architecture, we compare it with 3D mesh-based architecture. VHDL language is used to design our performance simulation platform, because it is believed that platform designed by hardware description language is more similar to realistic on-chip network [3].

We set up a standard simulation model as follows:

1) Both interconnect architectures have 32 IP cores ($4 \times 4 \times 2$).
2) Fixed-length messages are broken into 8 flits, and each flit is 32 bits wide.
3) IP cores independently generate messages and follow a Poisson process.
4) Message destinations are uniformly distributed across all the IP cores.
5) Each physical link has 4 virtual channels.
6) Wormhole switch and shortest path routing is used.
7) Buffers in the source IP core have infinite capacity.

The project is synthesized in Stratix EP1S80F1508C5.

First, we clarify the definitions of the latency, throughput and area overhead discussed in this section [12]. The latency, which is measured by cycles, is defined as the length of time elapses between the occurrence of the message header at the source IP core and the reception of the message tail at the destination IP core. The throughput, denoted by TP, is defined by

$$TP = \frac{M \times L}{I \times T} \tag{4}$$

where M is the number of messages that successfully arrive at their destination IPs. L denotes the message length measured by flits. I is the number of IP cores in NoC. T refers to the time (in cycles) from the occurrence of the first message generation and the last message reception. Therefore, throughput is measured as the maximal load that the network is capable of physically handling. Area overhead is the area required by all routers.

Next, we evaluate the performance of the two interconnect architectures. Fig.5 shows the comparison of the average latency. Table I compares the maximal throughput and area overhead.

It is observed in Fig.5 that compared with 3D mesh-based architecture, the average latency in our architecture decreases by 17% at most. In Table I, the maximal throughput increases by 8.5% and the area overhead in logical elements(LEs) decreases by 25% mainly caused by smaller number of routers.

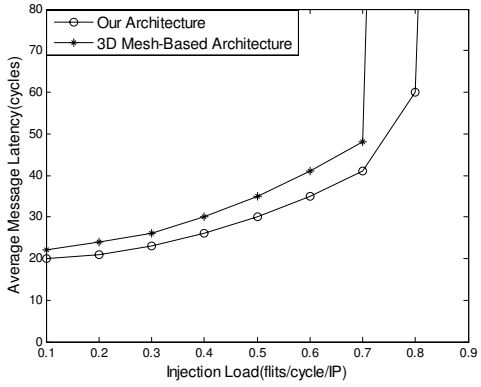


Fig. 5. Comparison of the average latency versus injection load

TABLE I
Comparison of the throughput and area overhead.

	Maximal throughput	***Area overhead***
3D mesh-based architecture	0.71 flits/cycle/IP	748715 LEs
Our architecture	0.77 flits/cycle/IP	561240 LEs

V. Conclusion

In this paper, we propose a dual-link hierarchical cluster-based interconnect architecture for 3D NoC. It only has approximately a quarter of routers in 3D mesh-based architecture, which results in low area overhead and small average number of hops. The decrease of the average number of hops leads to a decrease of average latency and an increase of throughput. Moreover, in order to improve performance, we insert dual links into some connections which are the bottlenecks of communication in the network. Simulation results show that compared with 3D mesh-based architecture, the average latency in our architecture decreases by 17%, the maximal throughput increases by 8.5% and the area overhead decreases by 25%.

Appendix A
Proof of Equation (1)

Assume k is the number of layers in 3D NoC and each layer has 16 IPs. H_{ij} denotes the total hops from one source IP which is in layer i to destination IPs which are in layer j. Thus, we can get the results as follow:

$$H_{ii} = 0 \times 3 + 2 \times 12 = 24, \forall i \in [1, k] \tag{5}$$

$$\begin{aligned} H_{ij} &= |j - i| \times 16 + 0 \times 4 + 2 \times 12 \\ &= |j - i| \times 16 + 24, \forall i, j \in [1, k] \end{aligned} \tag{6}$$

The total hops from one source IP which is in layer i to destination IPs which are in all layers (the number is $16k-1$), denoted by T_i, is given by

$$T_i = \sum_{j=1}^{k} H_{ij} = H_{i1} + H_{i2} + ... + H_{ik} \tag{7}$$

Thus, we can get the results as follow:

$$\begin{aligned} T_1 =& H_{11} + H_{12} + H_{13} + ... + H_{1k} \\ =& 24 + (24 + 16 \times 1) + (24 + 16 \times 2) + ... \\ & + (24 + 16 \times (k-1)) \\ =& 24k + (1 + 2 + ... + (k-1)) \times 16 \end{aligned} \tag{8}$$

$$\begin{aligned} T_2 =& H_{21} + H_{22} + H_{23} + ... + H_{2k} \\ =& (24 + 16 \times 1) + 24 + (24 + 16 \times 1) + ... \\ & + (24 + 16 \times (k-2)) \\ =& 24k + (1 + 1 + 2 + ... + (k-2)) \times 16 \end{aligned} \tag{9}$$

... ...

$$\begin{aligned} T_{k-1} =& H_{(k-1)1} + H_{(k-1)2} + H_{(k-1)3} + ... + H_{(k-1)k} \\ =& (24 + 16 \times (k-2)) + (24 + 16 \times (k-3)) + \\ & (24 + 16 \times (k-4)) ... + (24 + 16 \times 1) \\ =& 24k + ((k-2) + (k-3) + (k-4) + ...1 + 1) \times 16 \end{aligned} \tag{10}$$

$$\begin{aligned} T_k =& H_{k1} + H_{k2} + H_{k3} + ... + H_{kk} \\ =& (24 + 16 \times (k-1)) + (24 + 16 \times (k-2)) + \\ & (24 + 16 \times (k-3)) ... + 24 \\ =& 24k + ((k-1) + (k-2) + (k-3) + ...1) \times 16 \end{aligned} \tag{11}$$

Next, we can calculate the average number of hops, denoted by H_{our}, in our architecture as follow:

$$\begin{aligned} H_{our} &= \sum_{i=1}^{k} T_i / (k \times (16k-1)) \\ &= (24k^2 + 16k(k-1)(\frac{k+1}{3})) / (k \times (16k-1)) \\ &= (16k^2 + 72k - 16) / (3 \times (16k-1)) \end{aligned} \tag{12}$$

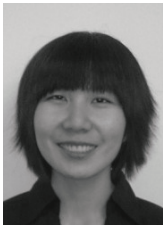
Until now, we have proved equation (1).

REFERENCES

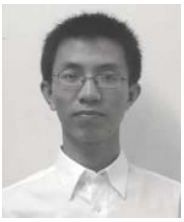
[1] Hu. Jingcao and R. Marculescu, *Energy-aware mapping for tile-based NoC architectures under performance constraints*, Proceedings of the ASP-DAC 2003, pp.233-239, 2003.
[2] L. Benini and G. Micheli, *Networks on chips: a new SoC paradigm*, IEEE Comput, vol.35, 2002.
[3] R. Marculescu, U.Y. Ogras, Peh. Li-Shiuan, N.E. Jerger, and Y. Hoskote, *Outstanding research problems in NoC design: system, microarchitecture, and circuit perspectives*, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol.28, pp. 3-21, 2009.
[4] C. Marcon and A. Borin, *Time and energy efficient mapping of embedded applications onto NoCs*, ASP-DAC 2005, pp.33-38, 2005.
[5] S. Murali, M. Coenen, A. Radulescu, K. Goossens and G. De Micheli, *A methodology for mapping multiple use-cases onto networks on chips*, Proc. Des. Autom. Test Eur. Conf 2006, pp.118-123, 2006.
[6] H. G. Lee, N. Chang, U. Y. Ogras and R. Marculescu, *On-chip communication architecture exploration: a quantitative evaluation of point-to-point, bus, and network-on-chip approaches*, ACM Trans. Des. Autom.Electron. Syst, vol.12, pp.20-40, 2007.
[7] M. Horowitz, R. Ho, and K. Mai, *The future of wires*, Proc. IEEE,vol.89, pp.490-504, 2001.
[8] J. Kim, W.J. Dally and D. Abts, *Flattened butterfly: a cost-efficient topology for high-radix networks*, Proceedings of the 34th annual international symposium on Computer architecture, pp.126-137,2007.
[9] V.F. Pavlidis and E.G. Friedman, *3-D topologies for networks-on-chip*, IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol.15, no.10, pp.1081-1090, 2007.
[10] Z. Yu, and B.M. Baas, *A Low-Area Multi-Link Interconnect Architecture for GALS Chip Multiprocessors*, IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol.18, no.5, pp.750-762, 2010.
[11] WJ. Dally and B. Towles, *Route packets, not wires: On-chip interconnection networks*, Design Automation Conference, 2001. Proceedings, pp.684–689, 2001.
[12] P.P. Pande, C. Grecu, M. Jones, A. Ivanov and R. Saleh, *Performance evaluation and design trade-offs for network-on-chip interconnect architectures*, IEEE Transactions on Computers, vol.54, no.8, pp.1025–1040, 2005.

Guang Sun received the B.S. from Xiamen University, Xiamen, China in 2008 in the department of electronics engineering. Now he is pursuing the Ph.D. degree in Tsinghua University. His research interests include on-chip network, message-passing multiprocessor and NoC architecture.

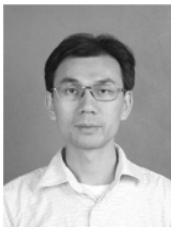
Yong Li received the B.S. from Huazhong University of Science and Technology, Wuhan, China in 2007 in the department of electronics and information engineering. Now he is pursuing the Ph.D. degree in Tsinghua University. His research interests include wireless networking, mobility management and future internet architecture



Yuanyuan Zhang received the B.S. from Shandong University, Jinan, China in 2007 in the department of electronics engineering. Now she is pursuing the Ph.D. degree in Tsinghua University. Her research interests include Network-on-Chip network, Multiprocessor System-on-Chip and Chip Multiprocessor.

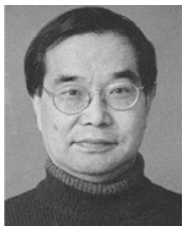


Shijun Lin received the B.S. degree from Xiamen University, Xiamen, China in 2005 and Ph.D. degree from Tsinghua University, Beijing, China in 2010. Now he is an assistant professor at Xiamen University. His research interests include On-Chip interconnect, high-speed networks and ASIC design.



Li Su received the B.S. degree from Nankai University, Tianjin, China, in 1999 and Ph.D. degree from Tsinghua University, Beijing, China in 2007 respectively both in electronics engineering. Now he is a research associate with Department of Electronic Engineering, Tsinghua University. His research interests include telecommunications, future internet architecture and on-chip network.

Depeng Jin received the B.S. and Ph.D. degrees from Tsinghua University, Beijing, China, in 1995 and 1999 respectively both in electronics engineering. Now he is an associate professor at Tsinghua University and vice chair of Department of Electronic Engineering. Dr. Jin was awarded National Scientific and Technological Innovation Prize (Second Class) in 2002. His research fields include telecommunications, high-speed networks, ASIC design and future internet architecture.



Lieguang Zeng received the B.S. and Ph.D. degrees from Tsinghua University, Beijing, China, in 1995 and 1999 respectively both in electronics engineering. Now he is an associate professor at Tsinghua University and vice chair of Department of Electronic Engineering. Dr. Jin was awarded National Scientific and Technological Innovation Prize (Second Class) in 2002. His research fields include telecommunications, high-speed networks, ASIC design and future internet architecture.