LETTER

A Dual-Port Access Structure of 3D Mesh-Based NoC*

Yuanyuan ZHANG†, Shijun LIN††, *Nonmembers*, Li SU†, Depeng JIN†a), *Members*, *and* Lieguang ZENG†, *Nonmember*

SUMMARY Since the length of wires between different layers, even between the top and bottom layers, is acceptably small in 3D mesh-based NoC (*three-Dimensional mesh-based Network on Chip*), a structure in which an IP (*Intelligence Property*) core in a certain layer directly connected to a proper router in another layer may efficiently decrease the average latency of messages and increase the maximum throughput. With this idea, in the paper, we introduce a dual-port access structure, in which each IP core except that in the bottom layer is connected to two routers in two adjacent layers, and, in particular, the IP core in the bottom layer can be directly connected to the proper router in the top layer. Furthermore, we derive the close form expression of the average number of hops of messages and also give the quantitative analysis of the performance when the dual-port access structure is used. All the analytical results reveal that the average number of hops is reduced and the system performance is improved, including a decrease of average latency and an increase of maximum throughput. Finally, the simulation results confirm our theoretical analysis and show the advantage of the proposed dual-port access structure with a relatively small increment of area overhead.

key words: *3D mesh-based NoC, dual-port access structure, latency, throughput, area overhead*

1. Introduction

The traditional bus-shared architecture tends to cause the bottleneck effect in the high-performance SoC (*Systems on Chip*). To solve this problem, NoC (*Network on Chip*) was proposed as a new interconnection architecture [1], [5]. Almost at the same time, 3D IC (*three-Dimensional Integrated Circuit*) emerged as an attractive option which can offer an opportunity to further improve the performance of an IC system [2]. Consequently, 3D NoC appears as a new approach in which the NoC architecture is merged with the 3D IC. As shown in Fig. 1, it is 3D mesh-based NoC in which the mesh architecture, i.e. a common architecture of NoC, is merged with the 3D IC. Moreover, it was proved that the 3D mesh-based architecture has a better performance than the 2D case [6]. Thus, further study on 3D mesh-based NoC becomes significant.

Manuscript received November 30, 2009.

Manuscript revised February 27, 2010.

†The authors are 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.

††The author is with the Department of Communication Engineering, Xiamen University, Xiamen 361005, China.

*This work is supported by the National High Technology Research and Development Program with No.2008AA01Z107.

a) E-mail: jindp@mail.tsinghua.edu.cn

DOI: 10.1587/transinf.E93.D.1987

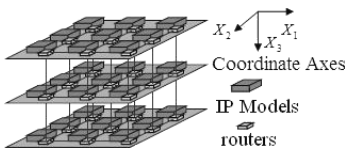

Fig. 1 Illustration of the 3D mesh-based NoC architecture.

In the paper, we focus our study on 3D mesh-based NoC associated with the through-via interconnection because, for such kind of 3D mesh-based NoC, the length of the through-via interconnection between layers ranges from 5 μm to 50 μm [4]. Thus, this length is much smaller than the intra-layer wiring length [3], i.e. approximately 1 mm which is more than twenty times of the former. Hence, in 3D mesh-based NoC, the physical distance between two marginal layers is acceptably small, making it feasible for an IP core in the bottom layer to be directly connected to one router in the top layer. With this property, we propose a dual-port access structure that the IP core in the bottom layer can be directly connected to the proper router in the top layer and other IP cores can be connected to two routers in adjacent layers. Finally, both the analytical and simulation results confirm the validity of this structure.

The rest of the paper is organized as follows. In Sect. 2.1, we introduce the dual-port access structure in 3D mesh-based NoC. Then, in Sect. 2.2, we derive the close form expression of the average number of hops and give the performance analysis. In Sect. 3, we explain the architecture of the corresponding NI (*Network Interface*) and explain the changes in the router and the IP core. In Sect. 4, we give the simulation results. Finally, we conclude the work in Sect. 5.

2. The Dual-Port Access Structure

2.1 Introduction to the Dual-Port Access Structure

As illustrated in Fig. 2 (a), for the dual-port access structure, the IP cores in the middle layer are connected to the two adjacent routers, respectively in the middle and bottom layers, and meanwhile, the IP cores in the bottom layer are connected to the two routers, respectively in the bottom and top layers. Such kind of connection does not exist in the traditional one-port access structure, as shown in Fig. 2 (b).

It can be inferred that the dual-port access structure will decrease the average number of hops of messages, if

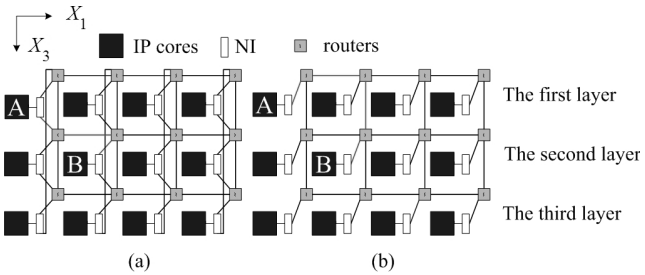

Fig. 2 The dual-port access structure in (a) and the one-port case in (b).

the shortest path routing strategy is used, because, for an IP core, this new access structure creates another short-length connection to the router in the adjacent layer. For example, one message is generated by the IP core A and its destination is the IP core B. For the dual-port access structure shown in Fig. 2 (a), it is observed that only one hop is required. However, for the traditional one-port access structure, the number of hops is two, as shown in Fig. 2 (b).

In addition, the dual-port access structure has another potential advantage. That is, it can mitigate the message congestion at the only one router in the traditional one-port access structure, since two adjacent routers could be chosen currently for an IP core in the dual-port case which decreases the probability of a high congestion level with only one router.

2.2 Performance Analysis

With the knowledge of the dual-port access structure, next, we will analyze its performance. In our discussion, it is assumed that the destination addresses of the generated messages are uniformly distributed across all of the IP cores in the 3D mesh-based NoC and meanwhile each IP core doesn't send messages to itself. Upon these assumptions, the following proposition gives the average number of hops of messages, when the dual-port access structure is used.

Proposition 1: In the 3D mesh-based NoC with the dual-port access structure, the close form expression of the average number of hops, denoted by $H^{(dp)}_{X_1X_2X_3}$, is given by

$$H^{(dp)}_{X_1X_2X_3} = \frac{X_3(X_1+X_2)(X_1X_2-1)}{3(X_1X_2X_3-1)} + \frac{(X_3-1)X_1X_2}{(X_1X_2X_3-1)} \times \frac{\left(\frac{X_3}{3}-1\right)(X_3-1)(X_3-2)+A}{X_3(X_3-1)} \quad (1)$$

where X_1, X_2 and X_3 are the numbers of IP cores in the corresponding dimensions, as shown in Fig. 1. X_3 is also the number of layers in the 3D mesh-based NoC. Symbol A takes the following expressions.

$$A = \begin{cases} \left(\frac{X_3-1}{2}-1\right)(X_3-1), & X_3 \text{ is odd} \quad (2) \\ \left(\frac{X_3}{2}-1\right)\frac{X_3}{2}, & X_3 \text{ is even} \quad (2') \end{cases}$$

The proof of *Proposition 1* is given in the Appendix.

Next, to compare the performance of our proposed structure with the traditional one-port access structure, we first review the average number of hops in 3D mesh-based NoC with the traditional one-port access structure given in [6], denoted by $H^{(op)}_{X_1X_2X_3}$.

$$H^{(op)}_{X_1X_2X_3} = \frac{X_3(X_1+X_2)(X_1X_2-1)}{3(X_1X_2X_3-1)} + \frac{(X_3^2-1)X_1X_2}{3(X_1X_2X_3-1)} \quad (3)$$

Associated with the average number of hops given by (1), i.e. $H^{(dp)}_{X_1X_2X_3}$, for the dual-port access structure, we get the following proposition.

Proposition 2: Compared with the traditional one-port access structure, the average number of hops of messages in 3D mesh-based NoC with the dual-port access structure decreases. Moreover, the decrement, denoted by H_{op-dp}, is given by (4).

$$H_{op-dp} = H^{(op)}_{X_1X_2X_3} - H^{(dp)}_{X_1X_2X_3} = \frac{(X_3-1)X_1X_2}{X_1X_2X_3-1} \times B \quad (4)$$

where the symbol B takes the following expressions.

$$B = \begin{cases} \frac{3X_3-1}{2X_3}, & \text{if } X_3 \text{ is odd} \quad (5) \\ \frac{(7(X_3-1)^2+1)}{4X_3(X_3-1)}, & \text{if } X_3 \text{ is even} \quad (5') \end{cases}$$

The proof of *Proposition 2* is straightforward. That is, according to (1) and (3), we can directly get (4), i.e. the expression of the number decrement of hops H_{op-dp}. Since $X_1, X_2, X_3 > 1$ holds, $H_{op-dp} > 0$ can be guaranteed. Hence, the average number of hops with the dual-port access structure is less than that with the traditional one-port access structure.

Furthermore, according to (1) and (3), we derive that the average numbers of hops with the dual-port access structure and the traditional one-port case are 24/13 and 36/13, respectively, based on the parameters in the standard model given in the simulation Sect. 4. Thus, for the dual-port access structure, the average number of hops decreases by 33.3%. Since the number decrease of the hops usually leads to a decrease of average latency and an increase of throughput, the dual-port access structure provides a nicer performance. Nevertheless, it requires more area overhead mainly cost by an additional router port and NI port.

3. Implementation of the Dual-Port Access Structure

To implement the dual-port access structure, we make some changes to the buffers in IP cores. More specifically, to avoid the messages transferred to different NI ports blocking each other, we divide the buffers in IP cores into two parts so as to store the messages transferred to different NI ports, respectively. Since the buffers are simply parted to two subareas, so no more overhead is introduced by the change of

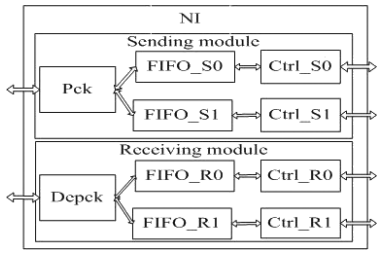

Fig. 3 The architecture of NI.

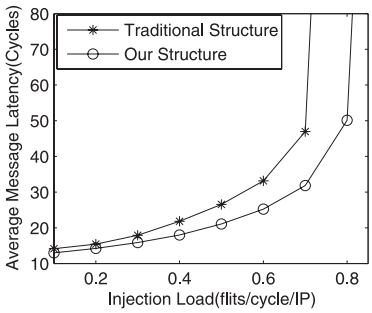

Fig. 4 Comparison of the average latency versus injection load.

Table 1 Comparison of the throughput and area overhead.

	Maximum throughput	Area overhead
Traditional structure	0.74 flits/cycle/IP	720954 LEs
Our structure	0.89 flits/cycle/IP	845262 LEs

buffers. Besides, the number of router ports and NI ports are both increased by one. The architecture of the NI is shown in Fig. 3. Pck and Depck modules implement the capsulation and decapsulation operations of messages. Ctrl_S0, Ctrl_S1, Ctrl_R0, Ctrl_R1 modules use the round-robin arbitration to select queued messages to handle. FIFOs can store messages in several separate queues corresponding to the virtual channels in routers. Pck and Depck are shared by the two ports and FIFOs can implement clock domain crossings.

4. Performance Evaluation

We set up a standard simulation model as follows:

1. The topology is 3D mesh-based NoC with 27 IP cores ($3 \times 3 \times 3$). Fixed-length Messages are broken into 9 flits, and each flit is 64 bits wide.
2. IP cores independently generate messages and follow a Poisson process. Moreover message destinations are uniformly distributed across IP cores.
3. 4 virtual channels per physical channel are used and each buffer can store 4 flits at most.
4. The channels are used according to wormhole switching and shortest path routing.
5. Buffer in the source IP core has infinite capacity.
6. The clock in IP cores is much faster than that in routers so messages in the source IP core can be transferred to FIFOs in the NI as soon as there is space and vice visa.
7. There are 4 queues in FIFOs of NIs and every queue can store 3 flits at most.

The project is synthesized in Stratix EP1S80F1508C5.

Now, we first clarify the definitions of the latency, throughput and area overhead discussed in our work. The latency refers to the length of time elapses measured by cycles 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

$$TP = \frac{Num \times Len}{Nc \times T} \tag{6}$$

where Num denotes the number of messages successfully arriving at their destination IP cores and Len denotes the message length measured by flits. Nc is the number of IP cores and T is the time elapses (in cycles) between the occurrence of the first message generation and the last message reception. Thus, the throughput is measured as the fraction of the maximum load that the network is capable of physically handling. Area overhead is the area required by routers and NIs.

Next, we show the performance. Figure 4 gives the comparison of the average latency and in Table 1, we compare the maximum throughput and area overhead.

It is observed in Fig. 4 that when we use the dual-port access structure, the average latency deceases by 32% at most, and in Table 1, the maximum throughput increases by 20% while the area overhead in logical elements (LEs) increases by 17% mainly caused by the number increment of ports at the routers and NIs.

5. Conclusion

In this paper, we proposed a dual-port access structure efficiently utilizing the property of 3D mesh-based NoC. We gave the theoretical analysis and evaluated its performance. The numerical results showed the validity of our proposed structure for the improved performance with a tolerable increment of area overhead. And this structure is robust because if one router an IP core connected to failed, it is still connected to the network through another router. This problem will be further discussed in my future work.

References

[1] W.J. Dally and B. Towles, "Route packets, not wires: On-chip interconnection networks," Proc. 38th Design Automation Conference (DAC01), vol.1, pp.684–689, 2001.

[2] W.R. Davis, J. Wilson, S. Mick, J. Xu, H. Hua, C. Mineo, A.M. Sule, M. Steer, and P.D. Franzon, "Demystifying 3D ICs: The pros and cons of going vertical," IEEE Des. Test Comput., vol.22, no.6, pp.498–511, 2005.

[3] S.B. Lee, S.W. Tam, I. Pefkianakis, S. Lu, M.F. Chang, C. Guo, G. Reinman, C. Peng, M. Naik, L. Zhang, and J. Cong, "A scalable micro wireless interconnect structure for CMPs," Proc. 15th annual international conference on Mobile computing and networking, pp.217–228.

ACM, 2009.
[4] F. Li, C. Nicopoulos, T. Richardson, Y. Xie, V. Narayanan, and M. Kandemir, "Design and management of 3D chip multiprocessors using network-in-memory," ACM SIGARCH Computer Architecture News, vol.34, no.2, p.141, 2006.
[5] 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 Trans. Comput., vol.54, no.8, pp.1025–1040, 2005.
[6] V.F. Pavlidis and E.G. Friedman, "3-D topologies for networks-on-chip," IEEE Trans. Very Large Scale Integr. (VLSI) Syst., vol.15, no.10, p.1081, 2007.

Appendix

Proof of Proposition 1: We can first calculate the average number of hops, denoted by $H_{X_1X_2}$ for both the dual-port and one-port cases, in the two dimensional plane scaled along X_1 and X_2 and then extend the derivation to the three dimensional case.

$$\begin{aligned} h_{X_1,X_2} &= \sum_{j=1}^{X_1X_2}\sum_{i=1}^{X_1X_2} h_{i,j} = \sum_{a=1}^{X_1}\sum_{b=1}^{X_2}\sum_{c=1}^{X_1}\sum_{d=1}^{X_2} h_{(a,b),(c,d)} \\ &= \sum_{a=1}^{X_1}\sum_{b=1}^{X_2}\sum_{c=1}^{X_1}\sum_{d=1}^{X_2} \left(h_{(a,b),(c,b)} + h_{(c,b),(c,d)}\right) \\ &= X_2^2 \sum_{a=1}^{X_1}\sum_{c=1}^{X_1} h_{a,c} + X_1^2 \sum_{b=1}^{X_2}\sum_{d=1}^{X_2} h_{b,d} \\ &= X_2^2 H_{X_1} X_1(X_1-1) + X_1^2 H_{X_2} X_2(X_2-1) \end{aligned} \quad (\text{A}\cdot 1)$$

where $h_{X_1X_2}$ denote the total number of hops in the dimentions X_1 and X_2. h_{ij} is the total number of hops between the IP cores i and j with the addresses (a, b) and (c, d), respectively. H_{X_1} and H_{X_2} denote the average number of hops in the dimensions X_1 and X_2 respectively. And from [6], we get $H_{X_i} = \frac{X_i+1}{3}$ $(i = 1, 2)$.

Since the difference of the two access structures is the addition of a connection between the adjacent layers only along the third dimension X_3 in the dual-port access structure, the total or average numbers of hops in the original two dimensions X_1 and X_2 for both the dual-port and one-port access structures are actually the same, i.e. $H^{(dp)}_{X_1X_2} = H^{(op)}_{X_1X_2}$, and given by

$$\begin{aligned} H^{(dp)}_{X_1X_2} = H^{(op)}_{X_1X_2} &= \frac{h_{X_1,X_2}}{X_1X_2(X_1X_2-1)} \\ &= \frac{X_2(X_1-1)}{X_1X_2-1} H_{X_1} + \frac{X_1(X_2-1)}{X_1X_2-1} H_{X_2} \end{aligned} \quad (\text{A}\cdot 2)$$

Based on the form of the result (A· 2) for the two dimensional case, we can extend it to the three dimensional case and get the following result directly.

$$H^{(dp)}_{X_1X_2X_3} = \frac{X_3(X_1X_2-1)}{X_1X_2X_3-1} H^{(dp)}_{X_1X_2} + \frac{X_1X_2(X_3-1)}{X_1X_2X_3-1} H^{(dp)}_{X_3} \quad (\text{A}\cdot 3)$$

Since the dual-port access structure changes the connection relationship in the third dimension X_3, the difference between the formula (A· 3) for the dual-port access structure and (3) for the one-port case given in [6] is $H^{(dp)}_{X_3}$ which is the mean value of the total number of hops in the third dimension X_3. It is given by

$$H^{(dp)}_{X_3} = \frac{\left(\frac{X_3}{3}-1\right)(X_3-1)(X_3-2)+A}{X_3(X_3-1)} \quad (\text{A}\cdot 4)$$

where the numerator denotes the total number of hops which is composed of two parts: *(i)* The item $\left(\frac{X_3}{3}-1\right)(X_3-1)(X_3-2)$ is the number of the hops between the IP cores in layers except the bottom layer. More specifically, from [6], the average number of hops in one dimensional mesh with X_3-1 IP cores is $\frac{X_3}{3}$. Thus, for the dual-port case in the third dimension X_3, the average number of hops between the IP cores in layers except the bottom layer is $\left(\frac{X_3}{3}-1\right)$. Moreover, the item $(X_3-1)(X_3-2)$ is the number of IP core pairs. *(ii)* The complement item A is the number of the hops between the IP core in the bottom layer and other X_3-1 IP cores in the other X_3-1 layers. In particular, when X_3 is odd, we get

$$A = \left(0 + \cdots + \left(\frac{X_3-1}{2}-1\right)\right)\times 2\times 2 = \frac{(X_3-3)(X_3-1)}{2} \quad (\text{A}\cdot 5)$$

otherwise, we get

$$\begin{aligned} A &= \left(\left(0 + \cdots + \left(\frac{X_3}{2}-2\right)\right)\times 2 + \frac{X_3}{2}-1\right)\times 2 \\ &= \left(\frac{X_3}{2}-1\right)\frac{X_3}{2} \end{aligned} \quad (\text{A}\cdot 6)$$

Until now, we have proved *Proposition 1*. □