2015 The 5th International Workshop on Computer Science and Engineering
doi: 10.18178/wcse.2015.04.115

Low Latency Mesh-based Hierarchical Network-on-Chip Structure

Gang Jian[1,a], Guo-dong Han[1,b], Yu-han Zhou[1,c]

[1]National Digital Switching System Engineering & Technological R&D Center

Zhengzhou, 450002, China

[a]lyonardo@163.com, [b]hgd@mail.ndsc.com.cn, [c]1300611739@qq.com

Keywords: network-on-chip, Hierarchical interconnection, parameterized design, clustered network.

Abstract. With the integration of cores increasing, on chip network (NoC) latency and throughput get worse in traditional structures. This paper proposed a novel low latency hierarchical mesh-based network-on-chip (PHNoC) structure which uses three parameters to describe hierarchical topology for the size changing design, and three types of base clusters to construct multilevel structure. Experimental results demonstrated that the proposed structure had lower latency and higher throughput than traditional 2D mesh and conventional hierarchical NoC in different size systems, and the larger size the better performance it improved.

Introduction

International Semiconductor Technology Roadmap (ITRS) predicts that by 2025 the physical size of the circuit will be narrowed down to 8nm [1], which drives the number of processors, ASIPs and other modules integrated in the Chip multiprocessors (CMP) will be continuously increasing. Network-on-chip (NoC) provides end-to-end communication infrastructure for multicore systems, but needs more scalable and lower average network latency. How to improve the size of NoC while ensure good latency and throughput performance is hotspot study. The right topology for large size of CMP is fatal important.

Conventional 2D topologies have limited path diversity and scalability lead to the bad latency and throughput performance when system nodes increased. 3D topology provides rich path diversity and lower network diameter, but high structural complexity and communication bottleneck of TSV [2]. Considering the complexity and resource constraints of Network-on-Chip design, the hierarchical interconnected NoC structure is a feasible solution.

Related work

Hierarchical NoC topology consists of local network and global network. Local network is the full connected PEs and global network interconnects the local networks by attaching upper rapid switching networks. The far-end messages can go through upper network to reduce the average latency. Hierarchical structure is also used in urban transportation such as the overpass, which reduces the waiting times of red lights and total traffic latency. Hierarchical interconnection structures can significantly reduce the hop count and provide better scalability, which is suitable for large size of CMP system [2].

Clustered structure, which determines the size of cluster and the concentration degree of upper-layer router, is regarded as the base of hierarchical topology. Winter M et al. [3] classified the clustered structures into two categories: real cluster structure and additional highway connections (AHC) structure. Real cluster structure (Fig.1a) 3x3 PEs is organized into one cluster and is

completely separated from other clusters. While the AHC structure, (Fig.1b) clusters are fully connected with each other.

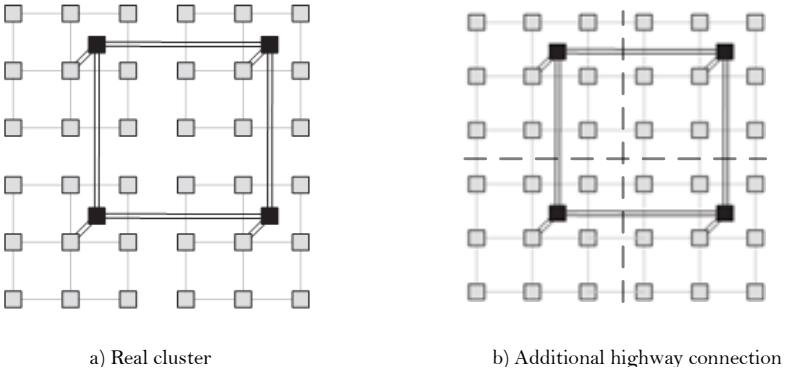


Fig.1: Two Types of Clustered Mesh Structure

GiGaNoC, BusMesh, Hrbird Ring/MeshNoC [4] and Concentrated Mesh (CMesh) [5] adopt the real cluster structure, which makes good use of communication locality and reduces the network diameter. However, the poor connectivity of real cluster is the main drawback which leads to heavy load pressure and congestion of upper-layer routers.

CHMesh [6] adopts AHC structure. Its underlying mesh is made up of 2x2 clusters and the concentration of upper layer router is four. Messages will be routed through upper layer when the distance of layer is shorter than that of base layer. AHC structure brings CHmesh better network connectivity and capacity which leads to 20% decrease of latency and 10% more throughput than CMesh. However, the concentration degree is still high which results in high complexity and congestion. In addition, the packages are not controllable whether to cross layers, which lead to unbalanced traffic load and poor utilization of network resources.

What's more, the throughput of CHMesh and CMesh decreases as network size increases. This is due to the fixed layers and concentration degree when system size grows, which leads to the change of traffic characteristics (especially the locality characteristics). Then, the unbalanced traffic load leads to large latency and bad throughput [7, 8].

Parameterized multilevel mesh-based NoC

Considering the former disadvantages, this paper put forward parameterized hierarchical NoC (PHNoC), including multiple layers, clustered interconnections and structure parameters. There are several improvements: making use of AHC structure, decreasing the concentration of routers, controlling cross-layer conditions and the parameterized topology description.

PHNoC parameters. Parameterized description of topology can facilitate the design of different size of NoC and the exploration of its design space. We designed three topology parameters.

- layer count, total number of layers.
- cluster-type in layer , ranging from 1,2,3.
- latency-threshold in each layer .

The "cluster-type" parameter chooses the type of basic cluster for each layer in hierarchical topology. Basic cluster structure determines the cluster size and concentration of routers, and is the cell of hierarchical mesh topology. Cluster size should not exceed 4x4 to prevent side-effects and the concentration of upper router should not be larger than 4, so that can reduce its complexity and max traffic load [3]. Therefore, three types of basic clusters (Fig.2) are designed, in which the first

type consists of a 2x2 cluster concentrated to upper router with one link; the second type consists of four 2x2 sub-clusters concentrated to upper router with four links; the third type consists of a 4x4 cluster concentrated to upper router with one link.

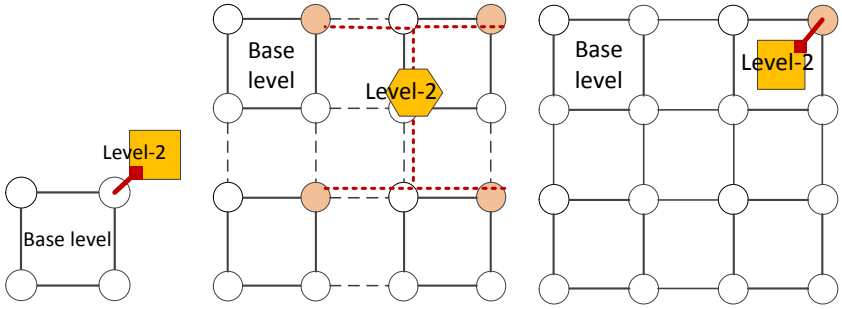


Fig.2: Three Types of Basic Cluster

The "latency-threshold" parameter controls the traffic load in each layer. One uniform threshold is kept by routers in each layer, represents the upper limit of message Manhattan distance in each layer, and determines the highest reachable layer of each message. The rational choice of latency-threshold should be made so as to balance the traffic load in each layer and improve resource utilization.

PHNoC Structure design. The base of PHNoC is fully-connected PE network, the upper (except the top) layers are routers interconnected by three types of basic clusters. Routers in the same layer have the same latency-threshold used for controlling traffic.

For the sake of good latency and throughput performance, we simulated various groups of structure parameters to explore PHNoC design space. Base on multilayer and the simulation experiences, we designed three typical topologies for 4x4, 8x8, 16x16 system respectively showed in Fig.3a,b,c.

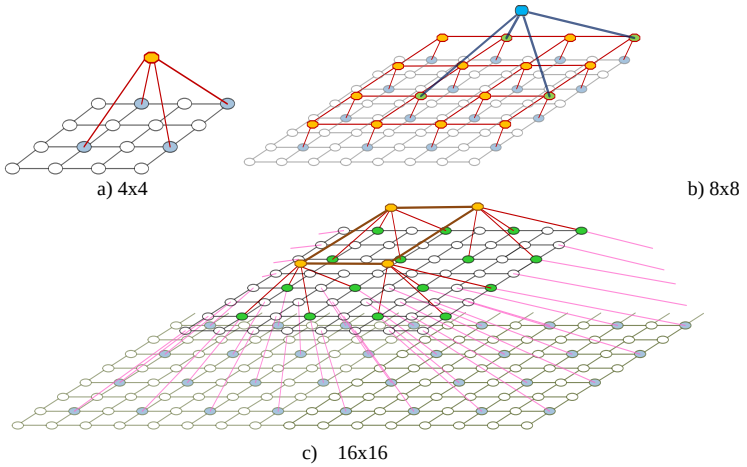


Fig.3: Three Size of PHNoC Topology

Table 1 shows the specific parameters. The 4x4 size system has two layers while 8x8 and 16x16 systems have three layers. In addition, the latency-threshold of 16x16 is larger than 8x8 systems.

Table 1: PHNoC Parameters in Three Sizes

Size of the System	**Parameters**		
	Ln	Cl_i	Ld_i
4x4	2	1	3
8x8	3	[1,2]	[7,8]
16x16	3	[1,2]	[8,17]

As for routing algorithm, we modified dimension order XY routing for PHNOC. XY routing is used separately in each layer, while the northeast node is set as the only access for cross-layer transmission so as to keep deadlock free [3].

PHNoC Structure analysis

Table 2: PHNoC Structure Analysis

Topology	System size	Network diameter	Average network distance	Router degree	Total router number	Links ratio	Area ratio
Mesh	16	6	2.5	4,5	16	1	1
	64	14	5.25		64	1	1.2
	256	30	10.625		256	4.29	0.86
CMesh	16	4	1.35	8	4	16.8	0.75
	64	8	2.75		16	1.18	1.05
	256	16	6.745		64	1.18	1.10
CHMesh	16	5	2.375	4,5,8	17	1.05	1.13
	64	8	4.15		68	1.28	1.04
	256	12	10.035		272	1.32	1.20
PHNoC	16	5	2.375	4,5	17	1.36	1.25
	64	10	4.065	4,5,6	81		
	256	12	9.25		324		

Table 2 shows the structure characteristic of PHNoC, Mesh, CMesh and CHMesh.

The analyses show that the network diameter of PHNoC is much smaller than Mesh and 25% larger than that of CHMesh in 64 size system. However, due to multiple-layers and latency-threshold the average network distance is smaller than CHMesh in all three size systems.

In addition, PHNoC has smaller node degree than that of CMesh and CHmesh. So, the traffic load per router in PHNoC shall be less than that of CMesh and CHMesh, from which we can deduce that PHNoC has the higher saturation point rate and better throughput than CMesh and CHMesh.

PHNoC resource overhead analysis. A quadratic functional relationship exists between router area overhead and its node degree, described in (1), where the parameters A, B, C can be derived from Orion [9]. Then, we calculate the area overhead of router with different degree and their number respectively. Last two columns in table 2 show the normalized links and area comparison of PHNoC, CMesh, CHMesh with Mesh.

$$S_{router}(Port) = AgPort^2 + BgPort + C \quad (1)$$

The maximum wiring overhead of PHNoC is 36% and 17.5% more than that of Mesh and CHMesh respectively in the same system size, and is far less than that of CMesh. The maximum area overhead of PHNoC is 25% and 10% more than that of Mesh and CHMesh respectively in the same system size. The node degree of PHNoC is small which indicates the complexity of router is small.

Experiments

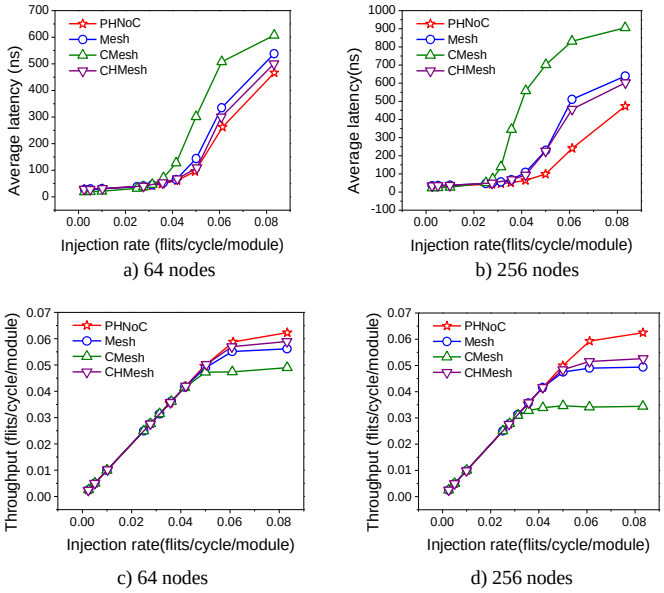
Simulation experiments are implemented with HNOCS, an open source NoC simulator [10]. Rent, Uniform Random, Bit Complement, Reversal, Transpose traffic pattern were used. Rent traffic pattern is based on Rent rule which is closer to practical application [8].

We simulated PHNoC in 16, 64 and 256 size systems with different traffic patterns, analyzed and compared the results of average latency, throughput and zero load latency with those of Mesh, Cmesh and CHMesh.

Fig.4 shows the average latency and throughput comparison under Rent traffic pattern in 16, 64 and 256 size systems. PHNoC has the minimum latency and the maximum saturation point injection

rate. In the 256 nodes system, the saturation point injection rates of PHNoC are 23% and 19.3% higher than that of Mesh and CHMesh respectively. Throughput of PHNoC is 18.57% and 12.97% more than that of Mesh and CHMesh respectively.

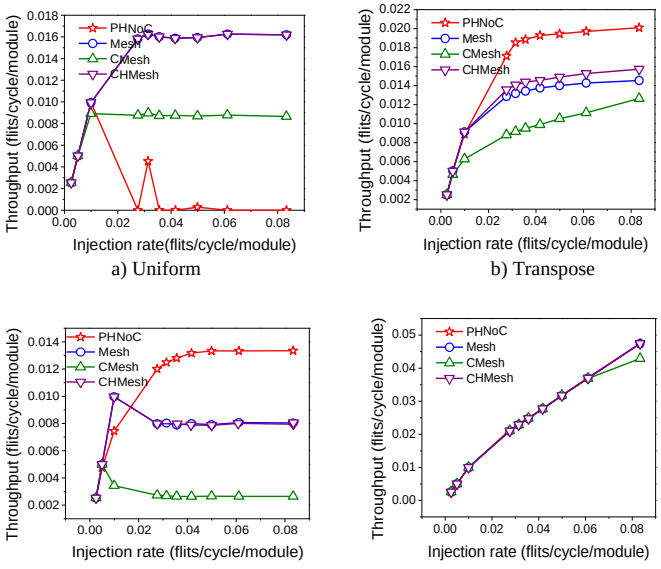
Fig.5 shows the throughput comparison under four compound traffic patterns. Under Transpose and Complement patterns, saturation injection rates of PHNoC are 77.9% and 65.6% higher than that of CHMesh and Mesh respectively. Under uniform pattern PHNoC has the worst throughput, and under complement pattern all the four topologies have low throughput. Because the destination of Complement pattern is calculated from the complement of source node number, which leads to large average distance, so as to uniform pattern. Under butterfly pattern all four topologies performed well, PHNoC showed no advantage.



a) 64 nodes b) 256 nodes

c) 64 nodes d) 256 nodes

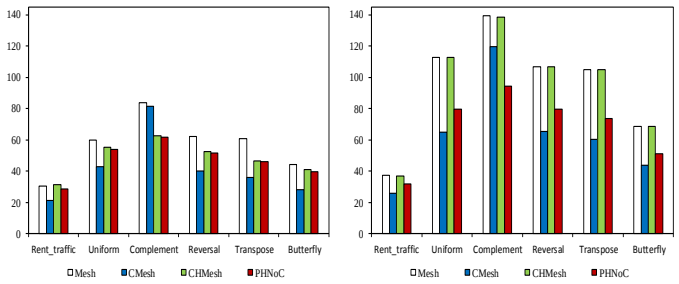
Fig 4: Average Latency & Throughput Comparison under Rent Traffic



a) Uniform b) Transpose

c) Bit Complement d) Butterfly

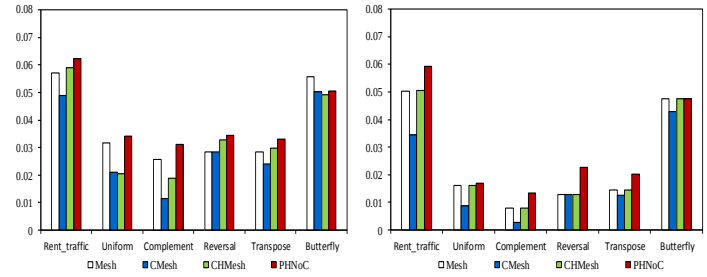
Fig 5: Throughput Comparison under Four Compound Traffic Patterns



a) 64 nodes b) 256 nodes

Fig. 6. Zero Load Latency Comparison of 64 and 256 Size System

From Fig.6 we can see the PHNoC has smaller zero load latency than Mesh and CHMesh under all the traffic except Complement. Especially in 256 size system, the zero load latency of PHNoC is 18.76% and 17.31% less than that of Mesh and CHMesh respectively under Rent traffic pattern. Due to zero load CMesh showed its advantage of high concentration structure and had the lowest zero load latency.



a) 64 nodes b) 256 nodes

Fig.7. Maximum Throughput Comparison of 64 and 256 Size System

From Fig.7 we can see that PHNoC had significant improvements in maximum throughput, and the larger system size was the higher throughput it got. Compared with CHMesh, PHNOC improved maximum throughput by 75.5% and 38.3% under Reversal and Transpose patterns respectively.

Conclusions

This paper proposed parameterized hierarchical PHNoC to achieve low latency in different size of system (especially large-scale systems). Using three types of basic-clusters to construct multilevel and brings better network connectivity, while the latency-threshold can balance the traffic load in layers. Experiments showed that PHNoC improved performance a lot and the larger system size was the better performance it got. However, PHNoC has wiring and layout complexity; further study will be the optimized structure parameters and its implementation difficulties.

Acknowledgements

The research work was supported by National High Technology Research and Development Program of China (863 Program) “The fifth-generation mobile communication system research and development (first-stage)”, under Grant No. 2014AA01A704.

References

[1] Wilson, Linda. "International Technology Roadmap for Semiconductors (ITRS)." (2013).
[2] Kim, John, Kiyoung Choi, and Gabriel Loh. "Exploiting new Interconnect technologies in on-chip communication." Emerging and Selected Topics in Circuits and Systems, IEEE Journal on 2.2 (2012): 124-136.
[3] Winter, Markus, Steffen Prusseit, and P. F. Gerhard. "Hierarchical routing architectures in clustered 2D-mesh networks-on-chip." SoC Design Conference (ISOCC), 2010 International. IEEE, 2010.
[4] Wanas, Mohamed A., M. A. Abd El Ghany, and Klaus Hofmann. "Hybrid Mesh-Ring wireless NoC for multi-core system." Design and Diagnostics of Electronic Circuits & Systems (DDECS), 2013 IEEE 16th International Symposium on. IEEE, 2013.
[5] Balfour, James, and William J. Dally. "Design tradeoffs for tiled CMP on-chip networks." Proceedings of the 20th annual international conference on Supercomputing. ACM, 2006.
[6] Kong, Feng, et al. "A novel mesh-based hierarchical topology for network-on-chip." Software Engineering and Service Science (ICSESS), 2014 IEEE 5th International Conference on. IEEE, 2014.
[7] Qian, Zhiliang, et al. "Performance evaluation of multicore systems: From traffic analysis to latency predictions (Embedded tutorial)." Computer-Aided Design (ICCAD), 2013 IEEE/ACM International Conference on. IEEE, 2013.
[8] Heirman, Wim, et al. "Rent's rule and parallel programs: characterizing network traffic behavior." Proceedings of the 2008 international workshop on System level interconnect prediction. ACM, 2008.
[9] Kahng, Andrew B., et al. "ORION 2.0: a fast and accurate NoC power and area model for early-stage design space exploration." Proceedings of the conference on Design, Automation and Test in Europe. European Design and Automation Association, 2009.
[10] Ben-Itzhak, Yaniv, Eitan Zahavi, and Israel Cidon. "HNOCS: modular open-source simulator for heterogeneous NoCs." Embedded Computer Systems (SAMOS), 2012 International Conference on. IEEE, 2012.