ALPHA: A hybrid topology for memory-centric network

Jiaxiang Li¹, Huaxi Gu^{1a)}, Shixiong Qi¹, Haoran Wang²,
and Kang Wang¹

¹ State Key Laboratory of ISN, Xidian University, Xi'an, China

² School of Electronic Engineering, Xidian University, Xi'an, China

a) hxgu@xidian.edu.cn

Abstract: Memory system, as the basic constituent of the computing system, has a inevitable impact on the system performance. Traditional bus-based memory access system cannot fulfill the performance requirement of the future high-performance computing system due to its limitation on energy-efficiency. Memory-centric network (MCN) is considered as a promising candidate for future system interconnect. To improve the energy-efficiency of memory access system by utilizing MCN, the topology of MCN needs to be carefully designed, as interconnection links take a huge part of power consumption. In this letter, we propose a new memory network topology named ALPHA. ALPHA is a 2-D topology. As ALPHA employs different types of interconnection links in the X-dimension and the Y-dimension, respectively, it can maintain high throughput while using fewer links. Besides we redesign the switch of HMC for its implementation in ALPHA. We make a comparison between ALPHA and four popular topologies. The simulation results show that ALPHA greatly increases throughput and decreases latency.

Keywords: topology, memory-centric network, hybrid memory cube

Classification: Integrated circuits

References

- [1] M. Zia, *et al.*: “Chip-to-chip interconnect integration technologies,” IEICE Electron. Express **13** (2016) 20162001 (DOI: [10.1587/elex.13.20162001](https://doi.org/10.1587/elex.13.20162001)).
- [2] K. Wang, *et al.*: “On the design of a 3D optical interconnected memory system,” IEICE Electron. Express **11** (2014) 20140664 (DOI: [10.1587/elex.11.20140664](https://doi.org/10.1587/elex.11.20140664)).
- [3] G. Kim, *et al.*: “Memory-centric system interconnect design with hybrid memory cubes,” PACT (2013) 145 (DOI: [10.1109/PACT.2013.6618812](https://doi.org/10.1109/PACT.2013.6618812)).
- [4] X. Jian, *et al.*: “Understanding and optimizing power consumption in memory networks,” HPCA (2017) 229 (DOI: [10.1109/HPCA.2017.60](https://doi.org/10.1109/HPCA.2017.60)).
- [5] S. Qi, *et al.*: “Testudo: A low latency and high-efficient memory-centric network using optical interconnect,” GLOBECOM (2017) (DOI: [10.1109/GLOCOM.2017.8254776](https://doi.org/10.1109/GLOCOM.2017.8254776)).
- [6] Hybrid Memory Cube Specification 2.1, Micron (2015).
- [7] D. Fujiki, *et al.*: “Randomizing packet memory networks for low-latency processor-memory communication,” PDP (2016) 168 (DOI: [10.1109/PDP.2016.18](https://doi.org/10.1109/PDP.2016.18)).
- [8] X. Yu, *et al.*: “Thor: A scalable hybrid switching architecture for data centers,”

- IEEE Trans. Commun. **66** (2018) 4653 (DOI: [10.1109/TCOMM.2018.2831200](https://doi.org/10.1109/TCOMM.2018.2831200)).
- [9] J. Kim, *et al.*: “Flattened butterfly topology for on-chip networks,” MICRO (2007) 172 (DOI: [10.1109/MICRO.2007.29](https://doi.org/10.1109/MICRO.2007.29)).
- [10] J. Kim, *et al.*: “Technology-driven, highly-scalable dragonfly topology,” ISCA (2008) 77 (DOI: [10.1109/ISCA.2008.19](https://doi.org/10.1109/ISCA.2008.19)).

1 Introduction

The memory wall problem is increasingly severe over the past two decades, caused by the gap between memory access bandwidth and processor computing capacity. As the number of cores increasing, the gap tends to be widened. Traditional memory interconnection such as bus-based architecture cannot meet the performance requirement of future memory access system [1, 2]. In order to further improve the memory system performance in post-moore law period, memory-centric network (MCN) has been proposed [3, 4, 5]. MCN is considered as a promising way to improve the memory access bandwidth in an energy-efficient way. Meanwhile, with the emerging of new DRAM devices, such as Hybrid Memory Cube (HMC) [6], the implementation of MCN tends to be possible. Memory bandwidth provided by a single HMC can be up to 320 GB/s, and the logic layer can provide packet routing function, which makes the MCN design practical [7].

Topology, as the crucial building component of the network like data center network [8], has a decisive impact on several network parameters, such as bisection bandwidth, network diameter, and average hop count. A well-designed topology can observably decrease transmission delay while maintaining considerable energy-efficiency. However, typical topologies such as Mesh, Torus, Flattened-Butterfly [9], and Dragonfly [10] are mainly designed for networks that take little consideration of the trade-off between performance and link number, which also have less compatibility for MCN. Meanwhile, links in memory-centric network are mainly off-chip, those links contribute a lot to energy consumption [3, 4]. Thus, it is important to reduce the number of links in the MCN. Additionally, as the feature of MCN, placing processors at 4 corners will cause the average hop count of MCN is higher than traditional interconnects [3]. The topology of the MCN needs to be specially designed with less link number and low average hop count in order to reduce power consumption.

In this letter, we propose a new topology for MCN called ALPHA. ALPHA is a 2-D topology but is connected in different ways in the X-dimension and the Y-dimension, considering the traffic flow in MCN. This enables ALPHA to have high throughput while using less energy-intensive off-chip links.

2 Architecture

2.1 Topology

ALPHA is a hybrid topology specially designed for MCN. Like Mesh, Torus, and Flattened-Butterfly, ALPHA is organized in a 2-D way (Fig. 1). In topologies like

Fig. 1. ALPHA architecture

Mesh, Torus and Flattened-Butterfly, the nodes in X-dimension and Y-dimension are always interconnected in the same way. In Dragonfly, nodes are connected in the same way in every group. However, to adjust the traffic flow, ALPHA is connected differently in X-dimension and Y-dimension. In MCN, CPUs are usually placed at 4 corners [3]. While using XY routing algorithm, there will be heavier contention in the X-dimension. Thus, the fully interconnection links, which derives from the Flattened-Butterfly, are employed in the first and the last column in X-dimension of ALPHA. Adversely, in the Y-dimension, there will be less traffic flow compared with the X-dimension. Thus, employing the fully interconnection links will decrease the utilization of transmission bandwidth and lead to the waste of network resources. Besides, as the fully interconnection links will increase the number of links, power consumption of Y-dimension will be brought to a higher level. Thus, the nodes in the Y-dimension are connected by utilizing a ring, which derives from Torus. Interconnecting the nodes in a ring-like link will decrease the hop count of the topology. Besides, in MCN, communication between CPUs is switched by HMCs. Nodes at 4 corners should be connected directly. Organizing the link in the Y-dimension as a ring will help to decrease the CPU communication latency, compared to the point-to-point links in Mesh.

Fig. 2. 22×22 switch for ALPHA

2.2 Switch architecture

In HMC, there is a switch integrated in the logic layer. I/O ports and vault controllers are connected by this switch, which composes an intra-network inside HMC. Each HMC contains 8 I/O ports and 16 vault controllers, so the switch is equivalent to a 24×24 electrical switch. However, with the number of port increasing, the design of switch becomes more complex. Meanwhile, ALPHA only require 22 ports in the switch. Hence we adopt a 22×22 crossbar to replace the original switch (Fig. 2). Considering the size of switch and the limited connectivity, we assume a concentrated mesh and 6-way concentration referred to [3].

3 Simulation and evaluation

Fig. 3. (a) Throughput under uniform random traffic. (b) Average memory access latency under uniform random traffic. (c) Link number of different topologies in 4×4 MCN. (d) Average hop count of different topologies in 4×4 MCN. (e) Power of different topologies in 4×4 MCN. (f) TPL of different topologies in 4×4 MCN

In this section, we simulate ALPHA in garnet2.0. To show the high performance and high efficiency of ALPHA, we choose Mesh, Torus, Dragonfly and Flattened-Butterfly as comparisons. All 5 topologies consist of 16 nodes. Link speed is set to 10 Gbps. The response latency of HMC is set to 170 cycles. And the packet size is 8 bit for control packet, 72 bit for data packet. We use the uniform random traffic

pattern to simulate the performance of all 5 topologies. And we use the link model in [3] to compare the power consumption of all 5 topologies when they reach the saturation point.

Fig. 3(a) shows the throughput performance comparison. As Flattened-Butterfly utilizes the full interconnected links in both X-dimension and Y-dimension, it has the highest throughput among all the comparisons, which is 260 Gbps. The throughput of ALPHA is 240 Gbps, which is tightly closed to the throughput of Flattened-Butterfly and is equal to the throughput of Dragonfly. Additionally, the throughput of ALPHA is 2.3× and 1.7× better than the throughput of Mesh and Torus, respectively. The performance gains on the throughput of ALPHA benefits from the utilization of full interconnected links in X-dimension. With the use of full interconnected links in X-dimension, the network congestion in X-dimension can be well relieved when employing XY routing algorithm, which can avoid the downgrade in throughput performance even with less available links.

Fig. 3(b) shows the comparison on the average latency. When the offered load is low, Mesh and Torus have lower average latency. However, when the offered load exceeds 0.1, the average latency will increase drastically. ALPHA, Dragonfly and Flattened-Butterfly can keep a low contention rate even the offered load is relatively high owing to the adoption on full interconnected links in the topology. The simulation results show that the saturation point of ALPHA is 0.15. When the offered load is inferior to 0.15, the average latency of ALPHA can stay at fewer levels, which is close to the average latency of Flattened-Butterfly. Compared to Flattened-Butterfly, ALPHA employs less links, however, it has approximately the same saturation point.

Fig. 3(c) shows the comparison on the available links. The number of available links of ALPHA, Mesh, Torus and Dragonfly are around 30. Employing more links in the network will lead to extra power consumption, however, network congestion rate will increase with less available links in the network, which will decrease the network throughput and increase the packet transmission latency. Thus, the number of available links in the network should be reduced with least performance penalty. It can be seen from Fig. 3(c) that ALPHA have the least available links while keeping the high performance compared to Flattened-Butterfly. With nearly saturation offered load and network throughput, the number of available links in ALPHA is only 0.6× compared to Flattened-Butterfly.

Fig. 3(d) shows the comparison on average hop counts. Under 4×4 network configuration, the average hop counts of ALPHA, Mesh, Torus, Dragonfly, and Flattened-Butterfly all are around 3 with little difference. Flattened-Butterfly has the least hop counts among all the comparisons, which is 2.5. The hop counts of ALPHA is 2.75.

Fig. 3(e) shows the comparison on power consumption. The power consumption of Mesh, Torus, Dragonfly, Flattened-Butterfly and ALPHA is 11.64 W, 15.47 W, 14.6 W, 23.19 W and 13.6 W, respectively. Because of the high power-consumption off-chip links, the power consumption of the topology is closely related to the number of links. While having the best performance, Flattened-Butterfly also have a bigger power consumption than other topologies. With fewest links in the topology, Mesh consumes least power.

Throughput is a direct measurement of traffic workload. However, throughput cannot reflect the performance gain brought by each link. Additionally, due to the huge energy cost of off-chip links, the system power consumption will increase dramatically without a limitation on number of available links, even though more available links in the network will bring better throughput performance. Thus, we define a new measurement, Throughput per Link (TPL), to obtain accurate evaluation on network performance, which can be seen as a trade-off between available links and throughput. TPL is defined as $TPL = N_{\text{maximal-throughput}} / N_{\text{number-of-link}}$. It will increase with the rising of throughput, however, only the topology with less available links can achieve better TPL compared to the alternatives with nearly throughput performance. Besides, only when a topology has less available links and each link in this topology has good throughput performance, it can have high TPL. Fig. 3(f) shows the comparison on TPL under uniform traffic pattern. It can be seen that ALPHA has the best TPL, which is 1.92×, 1.9×, 1.06×, and 1.56× better than that of Mesh, Torus, Dragonfly, and Flattened-Butterfly, respectively. When adopting XY routing algorithm, most of the traffic collects in X-dimension with sparse traffic in Y-dimension. As ALPHA utilizes full interconnected links in X-dimension, the traffic contention in X-dimension can be highly reduced. Besides, it can be seen from the comparison between Mesh, Torus, Dragonfly and Flattened-Butterfly that there is a positive correlation between number of available links and throughput. The reason is less available links will lead to higher network contention rate, which decreases the TPL of those topologies with lower available links.

4 Conclusion

In this letter, we propose a new topology, ALPHA, designed for memory-centric network. We describe the topology in detail. We also redesign the switch architecture of HMC for ALPHA. Besides, we compare ALPHA with 4 typical topologies. The result shows that ALPHA can perform more effectively.

Acknowledgments

This work was supported by the National Science Foundation of China under Grants 61634004, and Grant 61472300, the Fundamental Research Funds for the Central Universities Grant No. JB180309 and No. JB170107, and the key research and development plan of Shaanxi province No. 2017ZDCXL-GY-05-01.