A Cost and Performance Analytical Model for Large-scale On-chip Interconnection Networks

Kurihara, Takanori

(出版者 / Publisher)
法政大学大学院情報科学研究科

(雑誌名 / Journal or Publication Title)
法政大学大学院紀要. 情報科学研究科編 / 法政大学大学院紀要. 情報科学研究科編

(巻 / Volume)
12

(開始ページ / Start Page)
1

(終了ページ / End Page)
6

(発行年 / Year)
2017-03-31

(URL)
https://doi.org/10.15002/00014404

A Cost and Performance Analytical Model for Large-scale On-chip Interconnection Networks

Takanori Kurihara*
Graduate School of CIS, Hosei University
Email: 15t0007@cis.k.hosei.ac.jp

***Abstract*—As an interconnection topology, two-dimensional mesh is widely used in the design of the network-on-chip (NoC) for integrating dozens of processing elements (PEs). However, as the progress of IC technology, it becomes possible to integrate a large-scale system on a chip that contains more than one thousand PEs. In such a case, mesh topology will deteriorate performance due to the increase of communication time among PEs. This research investigates topologies and IC layout schemes of the mesh, torus, hypercube, and metacube, for achieving good cost-performance. We propose an analytical model for evaluating cost-performance ratio by considering NoC's topology and layout. The model is parameterized with the node degree, the network diameter or the average number of hops between any two nodes, the total number of routers, the router complexity, the number of PEs connected to a router, the connection width between routers, the cost ratios of the router section and the link section, the total number of PEs, and the total length of links. This model is helpful for us to find out the optimal topology and layout for NoC with a given network size. It was found that when the network size is small, mesh has a better cost-performance than others; as the network size increases, torus and hypercube outperform mesh; and metacube has the best cost-performance among them.**

1. Introduction

Network-on-chip (NoC) is a subset of system-on-chip (SoC) that provides communication fabric among many cores or processing elements (PEs) in a single chip. A number of research studies have shown the feasibility and advantages of NoC over traditional bus-based architecture [1], especially for designing a large-scale SoC that contains a large number of PEs.

Conventional interconnection networks for constructing large-scale supercomputers consist of routers and links. A router is usually implemented with a crossbar switch which has a number of ports for exchanging data or messages among routers. Links typically are high-speed cables that connect among ports of routers by following a certain topology. Because the cables are the ones of off-chip components, supercomputers can be implemented with complex interconnection networks based on high-dimensional topologies. Research studies of the conventional interconnection networks, aiming at the achievement of low cost and high performance, focus on the node degree and the network diameter which are major determining factors of cost and performance of the network. Achieving low cost and high performance meant to reduce the node degree and shorten the diameter.

Different from conventional interconnection networks, the NoC implements links with on-chip wires in the integrated circuit (IC) design. Due to the limitation of the IC manufacturing process, NoC faces difficulty to implement high dimensional networks. Multidimensional topologies must be projected onto a two-dimensional (2D) plane. Although 3D layout using Through Silicon Vias is expected to be available in future, we consider 2D layout in this research. To achieve low cost and high performance of NoC, in addition to node degree and network diameter, we must consider the IC layout, the length of link, and the area ratio of routers and wires to the total components on chip which contains also the PEs.

At the current circumstances, two-dimensional mesh is a popular topology used for the design of NoC because of its very simple structure and ease of on-chip implementation [2], [3], [4]. TILE64 [2] has 64 switches (routers) with an 8×8 array layout. The routers are interconnected with the mesh topology. Each router has also two additional ports to connect a processor and a cache, respectively. Intel developed a teraflops processor chip [3], [4] that contains 80 cores and 80 five-port crossbar routers. These are connected in an on-chip 8×10 2D mesh. Among the five ports, four ports are used to implement the 2D mesh topology and one port is used to connect a core to the router.

Usually, the NoC is implemented as an array (lattice) of tiles. A tile consists of a router and one or more PEs. Routers and links (wires) constitute the network which enables the communications among PEs. In the network following mesh topology, every routers is connected one or several PEs. In such a case, a tile in NoC can be considered as a node, a term in the interconnection networks. For some other topologies, tree and fat-tree for examples, some routers may have not direct-connected PEs. In the fat-tree implementations, the PEs may be attached to only the lowest-level routers (leaves). Such a network is not symmetric and hardly to be implemented on a chip. In this research, except for mesh, we consider the symmetric topologies of torus, hypercube, and metacube.

As the progress of IC technology, it becomes possible to integrate a large-scale system on a chip that may contains more than ten-thousand cores. In such a case, mesh topology will deteriorate performance due to the increase of communication time among PEs. It is needed to evaluate the on-chip implementations of other topologies and their cost-performance.

The main contributions of this research are to investigate the on-chip network topologies and their layout schemes of mesh, torus, hypercube, and metacube, and propose an analytical model for evaluating the cost-performance of on-chip interconnection networks based

* Supervisor: Prof. Yamin Li

on the topological properties, architecture, and IC layout characteristics. By using this model, we can design a large-scale high performance SoC at low cost.

2. Cost and Performance Metrics for NoC

When designing NoC, some parameters concerning the IC layout must be considered. In the traditional interconnection networks of supercomputers, the feature of the routing between two nodes is affected by the number of hops, not the physical distance. However, in the NoC, links between routers are implemented on the silicon. They have an impact on the area and power consumption of the chip. In this section, we describe the metrics that affect the cost and performance of NoC, including the common properties of supercomputers.

2.1. Topological Properties

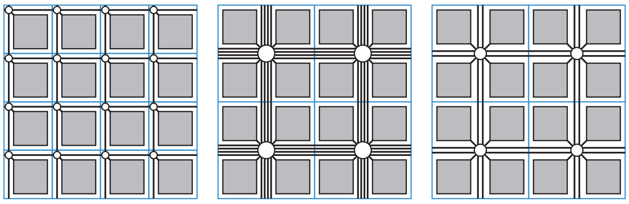
Topology is the logical arrangement of nodes and links. It greatly affects the cost and the performance of the interconnection networks. The following properties depends on topology and they affects the cost and the performance of both supercomputers and NoCs.

2.1.1. Node Degree (d). Node degree d is defined as the number of links connected to a node. Smaller or fixed d is a preferable characteristic. This is because the cost of the router increases non-linearly as d increases. Actually, the number of ports of the router is not equal to d because the router may have more ports for connecting PEs.

2.1.2. Network Diameter (D). Network diameter D is defined as the maximum number of passed through links of the shortest path between any two nodes in the network. The worst communication time between two nodes is proportional to D. Large D enlarges the communication delay and hence worsens the performance.

2.1.3. Average Hops ($\widetilde{D}$). Average hops $\widetilde{D}$ is defined as the average number of links of the shortest path between any two nodes. It is also known as the average distance. In order to distinguish the logical distance and physical distance, we use the term of average hops in this research. Like D, $\widetilde{D}$ affects the communication delay. $\widetilde{D}$ is more feasible value for estimating the communication delay at actual workloads. For this reason, we will adopt both D and $\widetilde{D}$ as the properties for the communication time estimation.

2.1.4. Total Number of Routers (R). Usually, N is used for denoting the number of nodes in a system. If every node has only one router, it equals the number of routers. But in some research papers, N is used for denoting the number of PEs. Several PEs may be connected to a router. Meanwhile, in the supercomputer field, a node may contain multiple routers. For the clarity, in this research, we use R to denote the total number of routers and let each tile contain only one router.



(a) $p = 1, t = 1$ (b) $p = 4, t = 1$ (c) $p = 4, t = 1/2$

Figure 1. Router-to-router connection width adjusted by thickness t

2.2. Architectural Properties

In the pure topological thinking, there are no meaning for physical properties such as the area of components; but we must consider such matters when we design a NoC. Here we describe such properties concerning the architecture of the network and limitation of 2D layout.

2.2.1. Router Complexity (λ). Router complexity λ is a property for the router cost estimation. A router consists of various components such as the crossbar switch, FIFO buffers for virtual channels, and the routing control logic. We must configure the order of the total cost according to the router architecture. The router complexity λ is an exponential factor for the router cost adjusting and we let it be in the range from 1.0 to 2.0 because the cost of a router's component increases linearly or quadratically with the number of ports.

2.2.2. The Number of PEs Connected to a Router (p). As we mentioned before, each router can connect several PEs directly. We let p be the number of PEs connected to a router. For the network with a given number of PEs, a large p will reduce R (the total number of routers) but meanwhile the cost of each router worsens because the number of ports in a router increases.

2.2.3. Thickness (t). Thickness t is a coefficient for adjusting the multiplicity of the network in case there are multiple PEs connected to a router. We let $t \leq 1$. For a given p and a given d, we design a router with p internal ports and d external ports. Then we use $p \times t$ such routers to connect p PEs with multiplexers and demultiplexers. In such a case, the cost of the routers becomes $pt(d + p)^{\lambda}$. In order to decrease the cost, we can let $t < 1$. Figure 1 shows three configurations of the tile. Figure 1a shows the simplest case of $p = 1$ and $t = 1$. Figure 1b shows the case of $p = 4$ and $t = 1$, like what a fat-tree does. And Figure 1c shows the case of $p = 4$ and $t = 1/2$, like what a thin-tree does.

2.2.4. Cost Ratios of the Router Section and Link Section (α). The total cost of the network is the sum of the costs of the router section and the link section, and it will be affected significantly by the dominant section [5]. The result in [6] shows that the link section significantly dominates at both area and power. Meanwhile in another study of [3], the router section dominates, and the link consumes 17% of the communication power. The cost balance depends on the chip architecture, and also it will change in the future. Therefore, we let α be the cost ratio of the router section to the sum of the router and link sections with $0 < \alpha < 1$. For example, $\alpha = 0.6$ means that the router section and the link section dominant 60% and 40%, respectively, of the total cost of the network.

2.2.5. Unused or Reserved Ports for Connecting Off-chip Components. The actual NoC must have external interfaces for connecting with off-chip memory modules, input/output controllers, and/or other processor chips. Mesh network has unused ports on outer boundary routers. These free ports can be used as the external interfaces. Meanwhile, the symmetric networks such as torus and hypercube do not have the unused port; it is not appropriate to compare these networks to the mesh with the same number of routers. In this research, for these symmetric networks, we remove PEs from outer boundary tiles and reserve vacant ports for external interfaces. For example, in order to have the same PE number of a 4×4 mesh, we can adopt a 6×6 torus in which no PEs are connected to outer boundary routers. In this case, R, D, and the total length of the links are increased.

2.2.6. Total Number of PEs (P). As mentioned repeatedly, a router connects p PEs in the interior tiles. On the other hand, for each topology excluding mesh, we assume that all routers in the boundaries of 2D layout have no PEs connected. Thus, the total number of PEs of the system P is obtained from the product of p and the number of interior routers.

2.3. Layout Properties

Except for mesh, wires connecting routers may be different in lengths. Layout properties concern the length and the cost estimation of the links.

2.3.1. Unit Length and Unit Cost of Links. We define the unit length of links as the length of the shortest link: the distance between two adjacent routers which has the minimum value. It is equal to the side length of the tile, i.e., the square root of the tile size. Because the tile size is mainly affected by the area of the processing section, we estimate the cost of links approximately using the number of PEs in a tile. That is, the unit link cost is $\sqrt{p}$. We assume that all tiles are the same size squares.

2.3.2. Total Length of All Links (L). The area of links occupying to the entire chip is proportional to the total link length. That is, the total link length linearly affects the cost of the link section. In the logical aspect, the total length of links is equivalent to the total number of links in network. But in NoC implementations, it is not true. If a link connects two routers which are not adjacent, the link length is larger than the unit length. We use the symbol L to represent the total length of all links.

2.3.3. Maximum Link Length. The maximum link length is the length of the longest link among all the links. Excessively long links may cause that messages cannot be transferred in one clock cycle due to the wiring delay; or it causes the reduction of the network clock frequency. Wiring delay depends on manufacturing process and other physical or electrical properties. In this research, we make an important assumption: The delay of the longest wire is less than the cycle time of the PEs. That is, messages can be transmitted along all wires in one clock cycle.

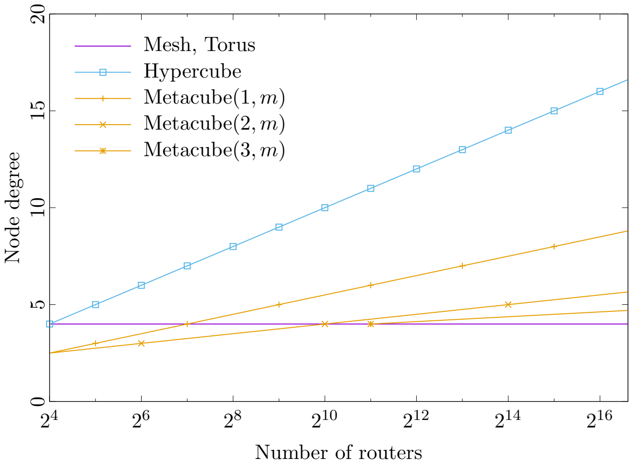


Figure 2. Node degrees of interconnection networks

3. Topologies and their Link Projection

Here we introduce briefly some topologies that are possible to be implemented with NoC. We describe their projections and layout schemes on 2D IC. Also, the total length of the links L and the average hops $\widetilde{D}$ for each topology will be calculated.

3.1. Mesh

Mesh has a very simple structure like a lattice. The 2D mesh is just called mesh in this research, and we will not consider higher dimensional meshes. Mesh networks are not symmetric, the boundary and corner nodes have free port(s), and these ports can be used to connect with off-chip components. For the $k \times k$ mesh, the number of nodes is k^2 and diameter is $2(k-1)$. The degree of the mesh is fixed to four. The $k \times k$ mesh has $k-1$ links for one row or column, and there are k rows and k columns in the network. Thus the number of links of mesh is $2k(k-1)$. Because each link connects two adjacent nodes, the length of each link is 1 unit. Thus the total length of the links is $2k(k-1) \times 1 = 2k(k-1)$.

Figure 2 shows the node degree of the mesh. We plot it here for the comparison with other topologies which we will discuss late. Note that in the horizontal axis, we use the number of routers R, instead of the number of PEs P or the network size N because of the node degree is a topological parameter.

Figure 3 shows the diameter for the mesh. We can see that the diameter of the mesh increases quickly as R increases. Similarly, we also plot the diameters of other topologies for comparison.

Figure 4 shows the average hops of the mesh. We calculate it by dividing the sum of distances of all node pairs by the total number of pairs. The average hops of the mesh is approximately $2k/3$, about $1/3$ of its diameter. We also plot the average hops of other topologies for comparison.

3.2. Torus

The k-ary n-dimensional torus, or k-ary n-cube, is commonly used as the interconnection network for commercial supercomputers. In this topology, k nodes are

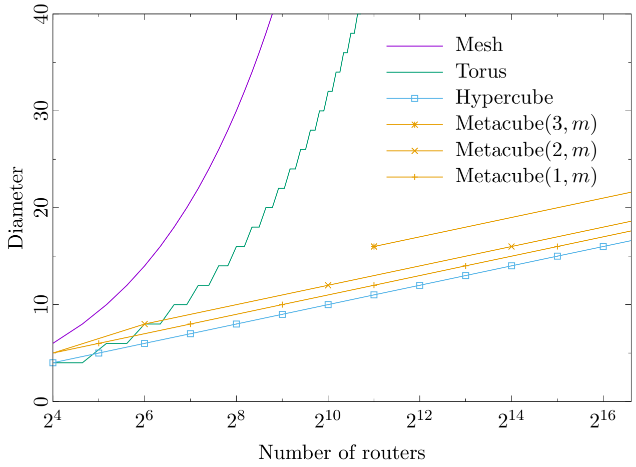


Figure 3. Diameters of interconnection networks

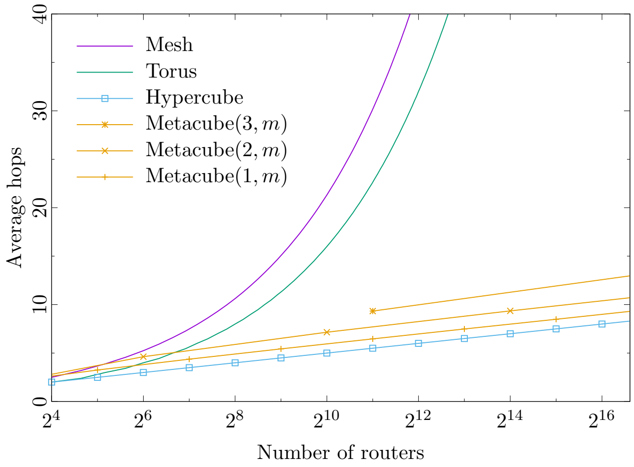


Figure 4. Average hops of interconnection networks

connected in a ring and each node belongs to n rings. Compared to the mesh network of same size, the torus has approximately half diameter, although the degrees of these two networks are equal. In particular, 2D torus ($n = 2$) can be obtained by adding wrap-around links that connect end-to-end nodes of 2D mesh network. Compared to the mesh with the same number of nodes k^2, the diameter 2D torus is reduced to $2 \lfloor k/2 \rfloor$; and the degree is also fixed to four, as shown in Figure 2 and Figure 3. The average hops $\widetilde{D}$ is $D/2$ if k is even, and otherwise it is in between $\widetilde{D}_{torus(k-1)}$ and $\widetilde{D}_{torus(k+1)}$. Figure 4 shows the average hops of the 2D torus that is approximately half of the diameter.

The number of links of 2D torus is $2k^2$. The length of each wrap-around link is $k - 1$ units. Therefore, the total length of the links is $2k^2 + 2k(k - 2) = 4k(k - 1)$, two times compared to mesh. Folded torus reduces the maximum link length, but the total length does not change with folding.

3.3. Hypercube

Hypercube (HC) was adopted widely in HPC systems. An n-dimensional hypercube (n-cube) has 2^n nodes. HC has very simple and interesting topological properties. Both of the node degree and the diameter are n; it grows logarithmically as the number of nodes. The average hops is $n/2$. We plot these in Figure 2, Figure 3, and Figure 4.

Figure 5 shows examples of 2D layout for the hypercube projections. There are many long links. An n-cube consists of two $(n - 1)$-cubes and additional links connecting corresponding nodes in two distinct $(n - 1)$-cubes. There are 2^{n-1} such links because each $(n - 1)$-cube has 2^{n-1} nodes. The length of these links is doubled when the dimension n is incremented by two. Thus we have the recursive equation for calculating the total length of links $L_{hc(n)}$ for n-cube. Not exactly, but we have an approximation $L(n) \approx \left(2^{n-1}(2^{\frac{n}{2}} - 1)\right) / \left(\sqrt{2} - 1\right)$.

3.4. Metacube

Metacube (MC) [7] has a hierarchical hypercube structure with two parameters: k and m. It is a symmetric network with small node degree and short diameter. The number of nodes of MC(k, m) is $2^{(2^k m)+k}$, the node degree is $m + k$, and the diameter is $2^k(m + 1)$.

Figure 2, Figure 3, and Figure 4 also show the node degree, diameter, and average hops, respectively, of the metacube. The three curves of the metacube are for MC$(1, k)$, MC$(2, k)$, and MC$(3, k)$, respectively. The average hops is approximately half of the diameter.

Some example layouts of the metacube are shown in Figure 6. The address of each node is expressed in binary notation and x in the notation means it may be a 0 or a 1. The least significant k-bits show the local address in k-cube (x for $k = 1$ or xx for $k = 2$ in the figure), the next 2^k bits show the ID of k-cubes for $m = 1$ (there are 2^{2^k} k-cubes), and so on. In the MC(k, m) with $m > 0$, all nodes in separate low-dimensional clusters (MC$(k, m-1)$) are connected. Thus we can obtain the total length of links hierarchically. The calculation of the total link length is complex for higher-order ($k > 3$) metacubes.

4. Cost-Performance Analytical Model

There are two major components in the NoCs: routers and links. Here, we estimate the total cost of each section as follows.

$$Cost_{routers} \propto (d + p)^{\lambda} R; \qquad Cost_{links} \propto \sqrt{p} L$$

where d is the node degree for the network, p is the number of PEs connected to a router directly, $(d + p)$ is the number of ports in a router including the ports for PEs, λ is router complexity, R is the number of routers, $\sqrt{p}$ is unit cost of links, and L is the total length of links on the entire network. As described before, the cost of each router can be estimated by the λ-th power of the number of ports. And the total cost of the link section is the product of the total length of links and the unit cost of links.

The total cost of the network is the sum of the costs of the router section and the link section. In order to reflect the impact of the influential section, we introduce a cost ratio of each section. Also, the connection width adjustment for the router with multiple PEs is considered. We define the total cost of NoC as follows.

$$Cost_{net} \propto \left(\alpha(d + p)^{\lambda} R + (1 - \alpha)\sqrt{p} L\right) tp$$

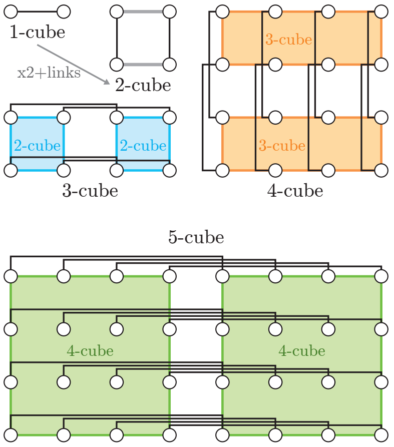


Figure 5. 2D projection of hypercube

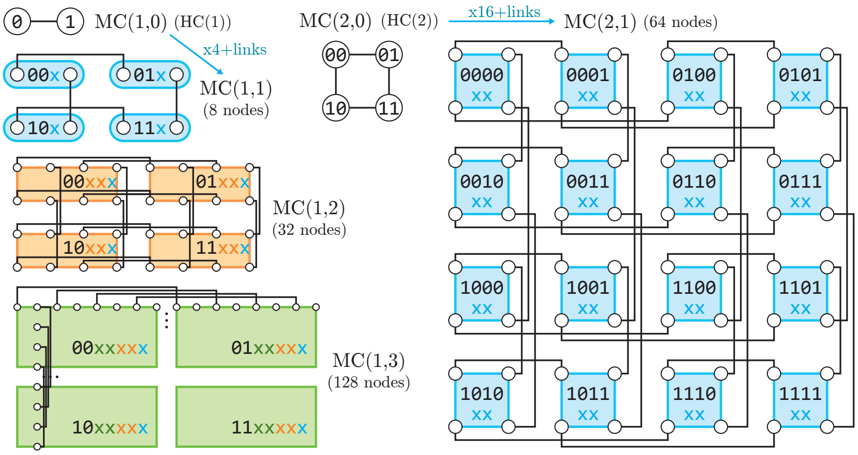


Figure 6. 2D projection of metacubes

where α is the cost ratio of the router section to the sum of router section and link section with $0 < \alpha < 1$ as we defined before, and t is the thickness. We adjust the connection width by tp with $0 < t \leq 1$, as discussed in the previous section.

Our model adopts the communication latency for evaluating the cost-performance ratio. As already mentioned, the communication latency is obtained from the network diameter or the average hop count between any two nodes. We assume the network performance is inversely proportional to the diameter or average hops. Conclusively, we define the performance of the network as below.

$$Performance_{net} \propto \frac{P}{D}; \quad \widetilde{Performance}_{net} \propto \frac{P}{\widetilde{D}}$$

where P is the number of PEs, D is the diameter of the network, and $\widetilde{D}$ is the average hops between any two routers.

The cost-performance is the cost over performance. The small value means low cost and high performance. We define the equations of the cost-performance as bellows. CP is the cost-performance of the network considering the network diameter, and $\widetilde{CP}$ is the another version considering the average hops.

$$CP = \left(\alpha(d+p)^{\lambda}R + (1-\alpha)\sqrt{p}L\right) tp\frac{D}{P}$$
$$\widetilde{CP} = \left(\alpha(d+p)^{\lambda}R + (1-\alpha)\sqrt{p}L\right) tp\frac{\widetilde{D}}{P}$$

The node degree d, the number of routers R, the total length of links L, the network diameter D, and the average hop count between any two nodes $\widetilde{D}$ are determined by the configuration of the target network. The cost ratio α, the router complexity λ, the number of PEs for one router p, and the connection thickness t are variables for adjusting the system.

When comparative evaluation, we need to compare the networks that have the same computing capability. Therefore we use the number of PEs (P) as the horizontal axis, not the number of routers. In addition, we also use a relative cost-performance (RCP [8]) for comparing with a baseline network. Relative means that the RCP is a relative CP of the target network to the baseline network. We use mesh as the baseline network. The numbers of PEs must be equal for the baseline network and the target network when calculating RCP. Let P_{targ} be the number of PEs of the target network. Then the baseline is determined as the $\sqrt{P_{targ}/p}$-ary 2D mesh. Thus, the RCP can be obtained as follows from the CP of baseline CP_{base}. Common configurations are used for the parameters α, λ, p, and t.

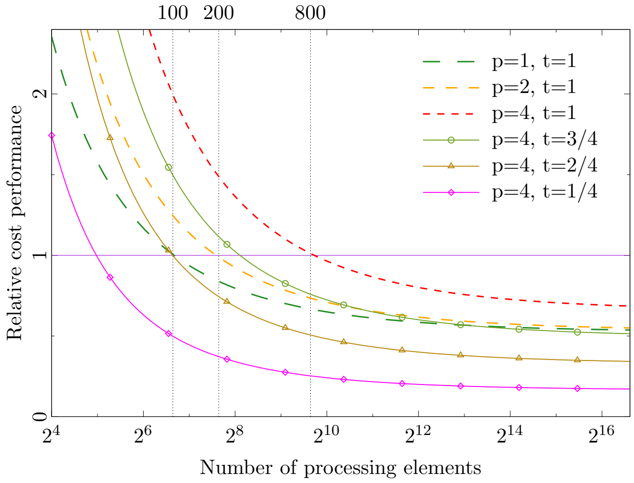


Figure 7. RCP comparison on p and t for torus with $\alpha = 0.6, \lambda = 2.0$

$$RCP = CP_{targ}/CP_{base}; \quad \widetilde{RCP} = \widetilde{CP}_{targ}/\widetilde{CP}_{base}$$

It is necessary to consider the reserved ports for the torus, hypercube, and metacube. We implement reserved ports by removing PEs from outer boundary routers of the network. Thus the total number of PEs of the target network P_{targ} is obtained as follows:

$$P_{targ} = (R_{targ} - 4(\sqrt{R_{targ}} - 1))p$$

where R_{targ} is the total number of routers of the target network and p is the number of PEs in a tile.

Figure 7 shows the effects of p (the number of PEs for one router) and t (thickness of connection) on RCP of torus with $\alpha = 0.6$ and $\lambda = 2.0$. These curves are relative to the CP of mesh with $p = 1$ and $t = 1$. We also use the diameter of the torus $D = 2(k/2) = k$ instead of $2\lfloor k/2 \rfloor$. When p increases under $t = 1$, the required number of routers is reduced for the same number of PEs,

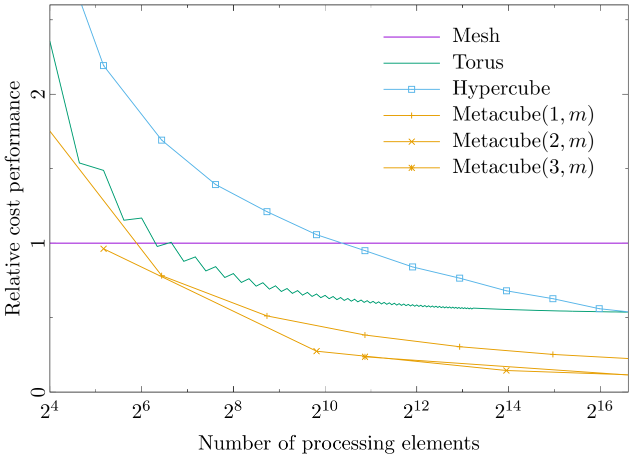


Figure 8. RCP in the range up to 10^5 PEs ($\lambda = 2$, $p = 1$, $\alpha = 0.6$, $t = 1$)

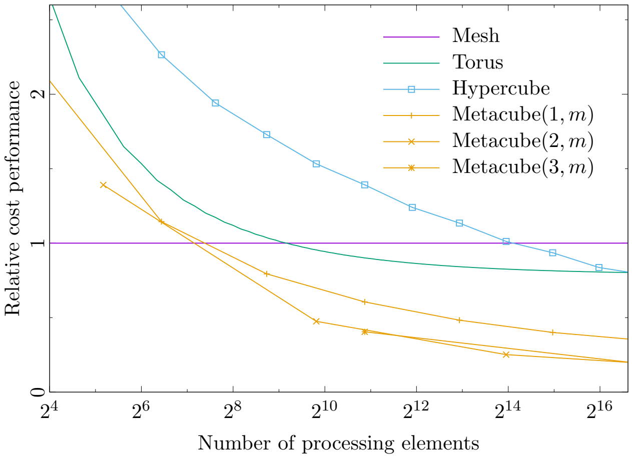


Figure 9. $\widetilde{RCP}$ in the range up to 10^5 PEs ($\lambda = 2$, $p = 1$, $\alpha = 0.6$, $t = 1$)

but the cost of the router, the unit cost of links, and the connection width increase. As a result, when comparing under the same number of PEs, the higher p brings a worse RCP (its value becomes larger). In this configuration, the torus with $p = 2$ has better cost-performance than mesh when the network has more than 200 PEs, and the torus with $p = 4$ is better than mesh if it has roughly over 800 PEs. If the communication traffic is sparse, we can cut down t and it will achieve the reduction of cost. As shown in Figure 7, the torus with $p = 4$, $t = 1/4$ achieves best cost-performance.

Figure 8 shows RCPs of mesh, torus, hypercube, and metacube in the range up to 100,000 PEs with $\lambda = 2.0$, $p = 1$, $\alpha = 0.6$, and $t = 1$. There are three curves for metacube MC(k, m), corresponding to $k = 1$, $k = 2$, and $k = 3$, respectively. In this configuration, RCP of torus reaches about half that of mesh as the number of PEs increases. Hypercube becomes better than mesh at about 1,300 PEs and becomes the same as the torus at about 100,000 PEs. We expect that such extra-large-scale NoCs can be realized in the next decade. From Figure 8, we can see that metacube has the best cost-performance among all the topologies for the large-scale NoC. This is because, compared to hypercube, metacube reduces node degree dramatically but the increase of the diameter is few. The disadvantage of metacube is that there are big node number gaps between configurations. For example, the metacube with 2^8 (256) PEs is not configurable. Anyway, in the configurable number of PEs, metacubes achieve high performance at low cost.

Figure 9 shows $\widetilde{RCP}$s of mesh, torus, hypercube, and metacube. Network configurations are the same as RCP: $\lambda = 2.0$, $p = 1$, $\alpha = 0.6$, and $t = 1$. When adopting the average hops for the performance estimation, curves have the same trend but change slowly than using the diameter. This is because the average hops of mesh is approximately $1/3$ of the diameter but those of other networks are almost $1/2$ of the diameter. In this configuration, the torus network overcomes the mesh at about 600 PEs. And also, metacubes achieve high performance at low cost.

5. Conclusion

We proposed an analytical model for evaluating the cost-performance ratio of the large-scale on-chip interconnection networks, and evaluated the cost-performance ratio of the torus, hypercube, and metacube relative to the mesh baseline network with various parameters. We found that when the network size is small, mesh has a better cost-performance than others; as the network size increases, torus and hypercube outperform mesh; and metacube has the best cost-performance among them. Our model considers the arrangement of outer boundary routers without connecting PEs to provide interfaces to the off-chip memory modules, input/output controllers, and/or other chips. The model and the evaluation results are helpful for designing large-scale high performance NoCs at low cost in the near future.

References

[1] J. Chen and P. Li, Cheng Gillard, “Network-on-chip (noc) topologies and performance: A review,” in *Newfoundland Electrical and Computer Engineering Conference*. IEEE Newfoundland and Labrador Section, 2011.

[2] S. Bell, B. Edwards, J. Amann, R. Conlin, K. Joyce, V. Leung, J. MacKay, M. Reif, L. Bao, J. Brown, M. Mattina, C. C. Miao, C. Ramey, D. Wentzlaff, W. Anderson, E. Berger, N. Fairbanks, D. Khan, F. Montenegro, J. Stickney, and J. Zook, “Tile64 - processor: A 64-core soc with mesh interconnect,” in *2008 IEEE International Solid-State Circuits Conference - Digest of Technical Papers*, Feb 2008, pp. 88–598.

[3] Y. Hoskote, S. Vangal, A. Singh, N. Borkar, and S. Borkar, “A 5-ghz mesh interconnect for a teraflops processor,” *IEEE Micro*, vol. 27, no. 5, pp. 51–61, Sept 2007.

[4] S. R. Vangal, J. Howard, G. Ruhl, S. Dighe, H. Wilson, J. Tschanz, D. Finan, A. Singh, T. Jacob, S. Jain, V. Erraguntla, C. Roberts, Y. Hoskote, N. Borkar, and S. Borkar, “An 80-tile sub-100-w teraflops processor in 65-nm cmos,” *IEEE Journal of Solid-State Circuits*, vol. 43, no. 1, pp. 29–41, Jan. 2008.

[5] W. J. Dally and B. Towles, “Route packets, not wires: on-chip interconnection networks,” in *Design Automation Conference, 2001. Proceedings*, 2001, pp. 684–689.

[6] D. Sanchez, G. Michelogiannakis, and C. Kozyrakis, “An analysis of on-chip interconnection networks for large-scale chip multiprocessors,” *ACM Transactions on Architecture and Code Optimization (TACO)*, vol. 7, no. 1, p. 4, 2010.

[7] Y. Li, S. Peng, and W. Chu, “Metacube — a versatile family of interconnection networks for extremely large-scale supercomputers’,” *The Journal of Supercomputing*, vol. 53, no. 2, pp. 329–351, August 2010.

[8] T. Kurihara and Y. Li, “A cost and performance analytical model for large-scale on-chip interconnection networks,” in *2016 Fourth International Symposium on Computing and Networking (CANDAR)*, Nov 2016, pp. 447–450.