

HyperDC: A re-arrangeable non-blocking data center networks topology

Xuefang Zhu

Wuxi High Electromechanical
Technology Vocational School
Wuxi, People's Republic of China
wxzxf688@hotmail.com

Li Peng

Jiangsu Key Laboratory of
IOT Application Technology
Taihu University of Wuxi
Wuxi, People's Republic of China
pengli@jiangnan.edu.cn

Abstract—In this paper, an re-arrangeable topology of data center networks is presented which is called HyperDC. Combining the merits of k -dilated, k -bristled and k -ary hypercube topology, this work introduces the definition of HyperDC topology in the view of graph theory; Using the mathematical induction, the properties of HyperDC is analyzed especially the part of re-arrangeable non-blocking property, which benefits line speed transmission of flow and simplification of cache mechanism in data center networks. Furthermore, the optimal embedment of many other topologies into HyperDC is realized, which also benefits the realization of a large number of parallel algorithms and stimulations of many other topologies in HyperDC; Finally, Compare with several different topologies, the property of re-arrangeable non-blocking in the aspects of latency, link complexity, switch complexity and scalability, which proves superior features in HyperDC topology

Keywords—HyperDC, mathematical induction, labeling strategy, data center networks, re-arrangeable topology;

I. INTRODUCTION

The available fat tree topology and other relative topologies of data center networks (DCNs) are facing with the challenge of complicated network environment. In the situation of rapid development of timely flow requirements and networked applications, DCNs have a strict requirement for adopting the distributed switching strategy to realize the higher parallel computing capacity. On the one hand, due to the explosive growth of network flow, data center networks require the higher bandwidth and throughput capacity. On the other hand, for satisfying the demand of various services, DCNs need to provide the service of ‘*-cast’, such as unicast and multi-cast. In addition, DCNs also need to consider properties of quick transfer of service, energy saving, excellent scalability, fault tolerance and so on [1], [2].

As we know, the DCNs topology develops from fat tree basically that is an indirect topology. These topologies own the multi-path non-blocking property, but properties of latency, cost, and scalability can't satisfy the requirement of enormous networks scale. The utilization of so many middle switches hinders the scalability of the large-scale DCNs and the timely networked application [3]. Therefore, many researchers turn to study topologies of hypercube and hypermesh that are the direct connection structures. Due to no middle switches, every switch element (SE) of the topology can carry one end terminal (ET) or more, which can improve the utilization factor of

network equipment. Because of no redundant middle SEs, the diameter of these topologies is smaller and the time of transmitting the data is less [4], [5], [6], [7].

With the hardware development of the high radix switch, per switch owns more ports under the condition that the output port bandwidth doesn't change [8], [9]. Under current craft circumstance of production, the switch can have 256 even 512 ports with the bandwidth of 10G. Extensive application of the high radix switch provides equipment foundation for the implementation of multi-dimensional direct bristled topologies in DCNs.

The emergence of optical technologies and equipment motivates the development of DCNs [10], [11]. With the application of wavelength division multiplexing (WDM), per fiber can transfer hundreds of light channels, which provides the technology foundation for the realization of hyperedge topology and its related topologies. The development of passive star coupler (PSC) and arrayed waveguide grating router (AWGR) brings optical SE into reality in the optical topology. Miniaturization and micro-miniaturization of optical transceiver and optical modulator provides a method of mutual transformation of optical signals and electrical signals. Optical equipment not only decrease the complication of topologies, but also reduce the energy consumption of DCNs due to the property of low power consumption of light.

The innovations of this paper reflect in several following aspects: (1) Combining the merits of the present direct topologies such as k -dilated k -bristled hypercube and k -ary hypermesh and then modifying and improving the original topologies to put forward a new topology called HyperDC; (2) Transforming the HyperDC topology into the multistage interconnection networks (MINs) topology and proving its re-arrangeable non-blocking property by the mathematical induction; (3) Summarizing the embedment property of HyperDC and realizing the optimal embedment of other topologies into it topology by using the labeling strategy.

Arrange this paper using the following structure. The second section, introducing two kinds of topologies that own great properties which are k -dilated k -bristled hypercube topology and k -ary hypermesh topology; The third section, through concluding and summarizing the present topologies, putting forward HyperDC topology and describing some relative properties about this topology; the fourth section, taking advantage of mathematical induction

to innovatively prove re-arrangeable non-blocking property of the HyperDC topology by transforming HyperDC topology into MINstopology; The fifth section, analyzing the embed-ment property of HyperDC and taking advantage of labeling strategy to realize the optimal embedment of other topologies into HyperDC topology; The sixth section, comparing with many other excellent non-blocking topologies in properties of latency, cost and scalability; The seventh section, about the conclusion of this paper.

II. BACKGROUNDS

This section mainly introduces k -dilated k -bristled hypercube topology and k -ary hypermesh topology.

A. K -dilated k -bristled hypercube

Hypercube topology [11] has already been applied widely into parallel computing system because of the small diameter and low cost as a typical orthogonal topology. Researchers have already presented an oblivious routing algorithm to prove the non-blocking property of hypercube. Despite having many merits of hypercube, one SE can only connect an ET, which can result in large cost for large network size. Researchers have put forward k -dilated k -bristled hypercube (DBHC) to solve this problem. Every SE of this topology can connect k ETs and every interconnection link should be dilated by a factor of k considering the requirement of non-blocking property. The structure of 16 ETs is shown in Fig 1 DBHC has extended scalability and reduced hops compared to traditional hypercube.

B. K -ary hypermesh topology

Hypermesh [10] is also an orthogonal multi-dimensional topology, but every dimension of this topology has k SEs unlike hypercube that it only have 2 SEs in the same dimension. Moreover, it simplifies the original topology and lower the complication of connections by replacing all links in the same dimension with a hyperedge. A hyperedge can contain one communication signal at the same time and every SE can only carry one ET, so hypermesh topology possess no non-blocking property. The structure of 16 ETs is shown in Fig.2.

III. HYPERDC

This section defines a new topology called HyperDC based on the second section. Consider non-blocking property, simplification of connection complication and scalability of the same dimension, the definition of k -ary HyperDC topology is as follows.

Definition 1. Figure $G = (V, E)$ denotes an n dimensional orthogonal indirect figure, $V = V(G)$ denotes the collection of SEs and $E = E(G)$ denotes the collection of hyperedges. The integer address of every SE of figure G can be denoted by $0, 1, \dots, |V|$, $|V|$ denotes the number of SEs. The integer address of every SE can be written by the format of n dimensional k -ary coordinate: $(x_n, x_{n-1}, \dots, x_1)$, $x_i = 0, 1, \dots, k-1$. Per hyperedge of figure G connects k SEs of the same dimension that per SE

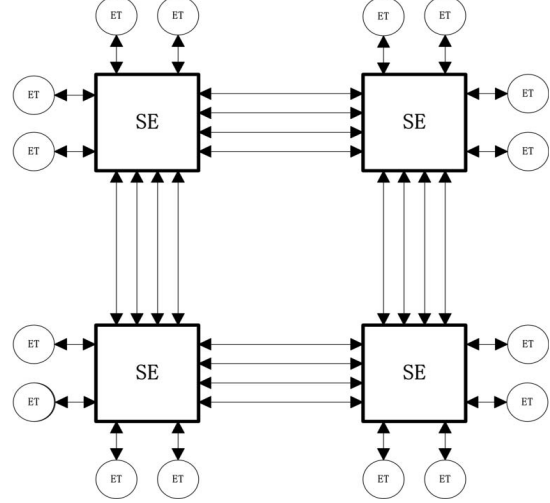


Figure 1. 4-dilated 4-bristled hypercube.

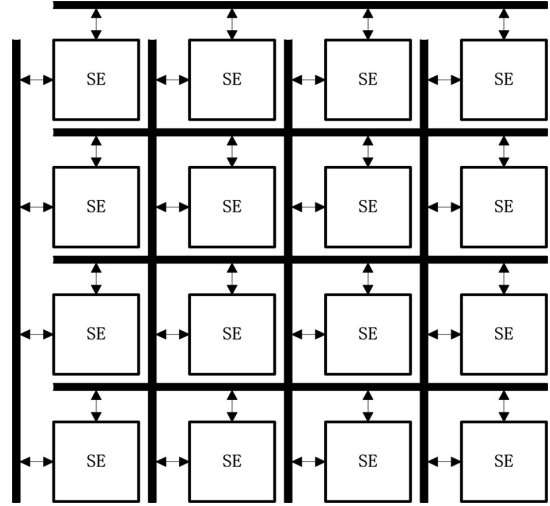


Figure 2. 4-ary hypermesh.

can be denoted by k -ary coordinate which is only different in one coordinate. Every SE of the same dimension possess $2d$ channels, d is between 1 and k . The $2d$ channels should connect the d different SEs in this dimension, in other words, per 2 channels of per one SE of this dimension can connect another SEs. As shown in Fig.3 and Fig.4.

From the view of definition 1, we can get some results. When $d = 1$, per SE can only carry 2 channel and the 2 channels can only communicate with k SEs of the same dimension simultaneously. The result corresponds to the definition about 2-dual hypermesh in the paper of Szymanski[14]. When $d = k$, per SE possesses $2k$ channels and these channels can communicate with k SEs of the same dimension simultaneously. So it owns $2k^2$ channels when k SEs communicate with each other. According to the above definition, we can conclude some properties of

2-dilated d -way k -ary bristled HyperDC.

Property 1. HyperDC topology has k^n SEs and per SE can carry d ETs, so this topology sums up dk^n ETs.

Property 2. The hyperedge of this topology can connect k SEs in the same dimension and have k^{n-1} hyperedges, so n -dimensional HyperDC topology sums up nk^{n-1} hyperedge.

Property 3. Because every SE possesses $2d$ channels in one dimension and every channel is of full duplex, the topology owns dnk^n channels.

Property 4. Diameter of the topology is n .

Property 5. Only consider the structure of topology not the physical realization and don't compute the number of links between every ET and SE, the degree of per SE is $2dn$.

Property 6. From the point of k -ary, the distance of any two SEs of HyperDC topology equals to their hamming distance.

Property 7. According to the definition and graph theory, n -dimensional HyperDC can be denoted by $k(n-1)$ -dimensional HyperDC recursively, in another word, $k(n-1)$ -dimensional HyperDC can be connected by a hyperedge.

Property 8. Divided by the different dimension, n -dimensional HyperDC can be divided into $k(n-1)$ -dimensional HyperDC with n possibilities. Moreover, $k(n-1)$ -dimensional HyperDC every kind of division are symmetrical.

For example, as Fig.3 and Fig.4, HyperDC topology has 16 switch elements, 64 ETs, 8 hyperedges and 128 channels. The degree of every switch element in the figure is 16, maximum hamming distance of any two nodes is 2 and the topology consists 4 1-dimensional HyperDC.

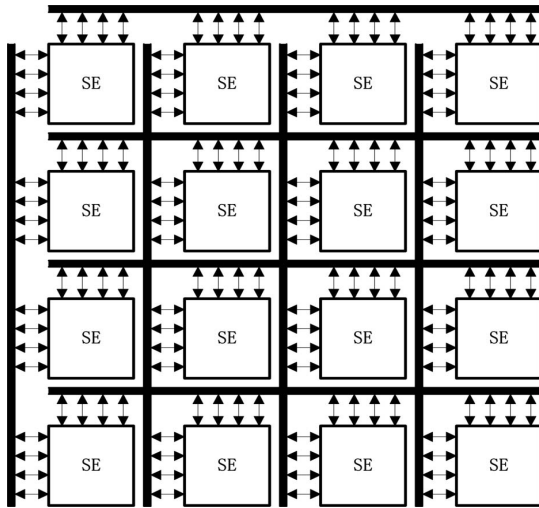


Figure 3. 2-dilated 4-way 4-ary bristled HyperDC.

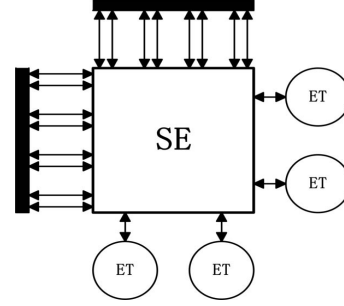


Figure 4. A unit in the HyperDC.

IV. ANALYSIS OF REARRANGEABILITY

Because the rearrangeability property of MINs have been studied extensively, in this paper we transform HyperDC topology into MINstopology and analyze the rearrangeability property of MINstopology to understand it of HyperDC topology. Referring to the method of Choi [10], the steps are as follows.

(1) According to the number of dimensions of HyperDC topology, make n copies of SEs of HyperDC topology and arrange every copy as every stage of MIN-stopology. SEs of every stage of MINs topology are duplicated as input SE and output SE.

(2) According to connection type of HyperDC topology in every dimension, combine output SE of stage $i+1$ and input SE of stage i of MINstopology, $i = 1, 2, \dots, n-1$, and then connect every SE successively. Define the formed n -stage MINstopology into $G_{n,1}$.

(3) Define the output SE of stage 1 into axis of symmetry and make mirror graph of the G_n . Every ET of HyperDC topology can denote two channels in every dimension by this. Define the mirror graph into $G_{n,2}$.

Fig.3 and Fig.4 can be transformed into MINs topology by the above 3 steps, as shown in Fig.5. The following work is to prove the rearrangeability of the transformed MINstopology, in other words, to prove the rearrangeability of the original HyperDC topology. According to the recursion property 7 and symmetry property 8, we prove the rearrangeability property by the mathematical induction as follows.

Proof: According to the definition of HyperDC, per SE have $2d$ channels and every terminal can communicate with the other ETs in the same dimension by 2 channels of every dimension of n . So the whole topology has dk^n ETs and we can divide 1-permutation of dk^n ETs into d -permutation of k^n SEs, label as $p_0, p_1, \dots, p_i, \dots, p_{d-1}$, correspondingly. ■

(1) When $n = 1$, with the above transformed steps, the corresponding 1-dimensional HyperDC should be transformed into MINstopology, as shown in the Fig.6. In the Fig.6 every SE j is labeled correspondingly as $0, 1, \dots, k-1$. We can adopt the following method of routing. For any permutation p_i , the input SE j can route to any middle SE M_m , m is between 0 and $k-1$, than it has d possibilities. Then, the middle SE gets to

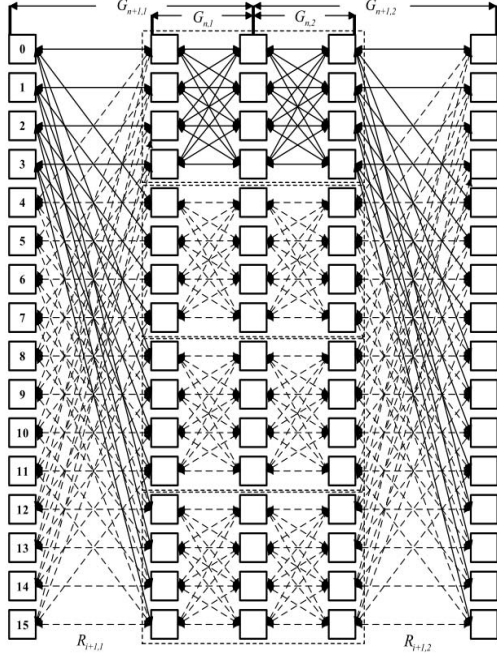


Figure 5. Transformed 2n MINs topology.

the output SE by the corresponding link of $G_{n,2}$, that it has d possibilities. In the process, connection roads of any SE j of any permutation p_i are non-blocking. Thus, the transformed 2 MINstopology is re-arrangeable, in other words, 1-dimensional HyperDC of k SEs is d -re-arrangeable and 1-dimensional Hy-perDC of dk ETs is 1-re-arrangeable, called rearrange-ability for short.

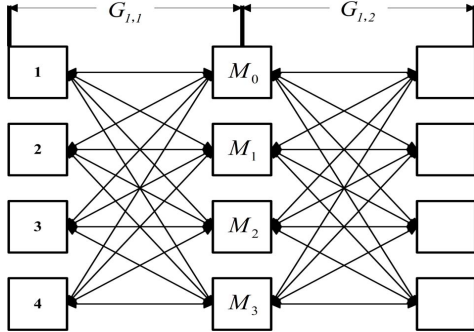


Figure 6. Transformed 2 MINs topology.

(2) The next work is to prove the d -rearrangeability of $(n+1)$ -dimensional HyperDC of k^{n+1} SEs, in other words, $(n+1)$ -dimensional HyperDC of dk^{n+1} ETs is re-arrangeable. Firstly, assume that n -dimensional HyperDC of k^n SEs is d -re-arrangeable, $n \geq 1$, so the corresponding $2n$ -stage MINstopology is also d -re-arrangeable. According to the recursion property 7 of HyperDC, $(n+1)$ -dimensional HyperDC of k^{n+1} ETs should be transformed into the corresponding $2(n+1)$ -stage MINs

topology to analyze it easier. In other words, $2n$ -stage MINstopology should be added with two dimensions as shown in the Fig.5. In the figure, every input SE j can be labeled as $0, 1, \dots, k^{n+1} - 1$. We can adopt the following method of routing. For any per-mutation p_i , the input SE j in the stage 0 can route to any middle $2n$ -stage MINstopology. Then, the any middle $2n$ -stage MINstopology can route to the output SE of $(2n+1)$ -stage MINs topology. In this process, every input SE j can reach to any a middle $2n$ -stage MINstopology that has d possibilities and every middle $2n$ -stage MINstopology is d -re-arrangeable. From this, $2(n+1)$ -stage MIN-stopology can be d -re-arrangeable. In other words, n -dimensional HyperDC of k^n SEs is d -re-arrangeable, so $(n+1)$ -dimensional of Hy-perDC of dk^{n+1} ETs is re-arrangeable. So the result has already been proved strongly.

Through the above statement, we can know that every SE can choose d kind of permutations, so every SE can carry d ETs that can make it non-blocking. When $d = k$, the scale of topology is the largest and equal to k^{n+1} . The number of SEs shouldn't be restricted into the power of 2 or the integer multiple of 2 like the request in the paper of Szymanski in the process of statement and can achieve any value, which lower the restrictions of SEs of data center network and make convenience for the expansions of large data center networks topology.

V. ANALYSIS OF EMBEDMENT OF THE TOPOLOGIES

The embedment property of topology refers that the nodes and edges of guest graph $G(V_1, E_1)$ are mapped one-to-one to the nodes and edges of host graph $H(V_2, E_2)$ by the function b and d , as (1) $b(u) \rightarrow v$, $u \in V_1$, $v \in V_2$; (2) $d(u_1, u_2) \rightarrow (b(u_1), b(u_2))$, $(u_1, u_2) \in E_1$, $(b(u_1), b(u_2)) \in E_2$. Considering the embedment property of topology, the two main reasons are that: (1) Realize the algorithm of guest graph in host graph topology perfectly. (2) Great characters of host graph can optimize the algorithm embedded guest graph. Four characters of load, congestion, dilation and expansion can evaluate whether the function of topology is good or not. Load refers to the number of nodes that guest graph is embedded into host graph. Congestion refers to the number of edges that guest graph is embedded into host graph. Di-lation refers to the largest length that guest graph is embedded into host graph. Expansion refers to the rate of nodes that guest graph is embedded into host graph. Szymanski proved that in the 2^n hyper-cube the guest graph with the condition of dilation k , congestion C and load L can be embedded perfectly with the condition of dilation $\leq k$, congestion $C \log_{2N} / \log_{dN}$ and load L . Kim[16] put forward a kind of labeling strategy to make the n -dimensional ($n \geq 2$) mesh skeleton be embedded optimally into 2-dimensional hypermesh and finally generalize to n -dimensional hypermesh and not square hypermesh by the theory statement. Because HyperDC topology replaces one-to-one edge with a hyperedge and every SE can carry d ETs, the original definition of topology embedment doesn't correspond to the condition

of the embedment of Hy-perDC topology and we should expand the new concept of topology definition. New definition is that nodes and edges of guest graph $G(V_1, E_1)$ can be mapped into corresponding nodes and edges of host graph $H(V_2, E_2)$ by the function b . In other words, (1) $b(u) \rightarrow v, u \in V_1, v \in V_2$; (2) $d(u_1, u_2) \rightarrow (b(u_1), b(u_2)), (u_1, u_2) \in E_1, (b(u_1), b(u_2)) \in E_2$. The embedment of nodes and edges is not one-to-one and the k guest graph G can be embedded into the same position of host graph H at most, including the many-to-one mapping of nodes and edge. Consider the labeling strategy and realize the embedment of guest graph into HyperDC topology. Assuming the embedment of guest graph of M nodes into q -dimensional HyperDC topology of N^q SEs and demanding N is the integer multiple of M , we should demand that the amount of edges of guest graph is the integer multiple of q . Because every switch element can carry d ETs in Hy-perDC, q -dimensional HyperDC of N^q SEs sums up dN^q nodes. According to the relationship of the number of nodes, dN^{q-1} guest graphs can be embedded into HyperDC optimally, labeled as $G_{0,0}, G_{0,1}, \dots, G_{0,d-1}, G_{1,0}, G_{1,1}, \dots, G_{1,d-1}, \dots, G_{af-1,0}, G_{af-1,1}, \dots, G_{af-1,d-1}$, respectively. The steps of embedment are as follows.

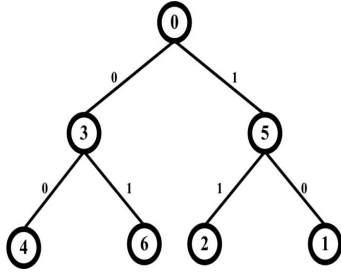


Figure 7. Guest treegraph.

- (1) Label nodes of $G_{i,j}$ as $0, 1, \dots, M-1, i = 0, 1, \dots, a^q-1, j = 0, 1, \dots, d-1$ randomly;
- (2) Label edge of $G_{i,j}$ as $0, 1, \dots, q-1$ averagely, in other word the number of every labeled edge is equal;
- (3) Have the per coordinate of k -ary coordinate of HyperDC mod M and then label them;
- (4) Put 0-labeled node of $G_{i,j}$ into the any position labeled 0 of HyperDC. The $G_{i,j}$ should be assigned into the same position labeled 0 when i is same in $G_{i,j}$;
- (5) Assign the another $1, 2, \dots, M-1$ nodes into the relative position combining the labeled dimension of edge and the next node follows the former node as a benchmark. So far, we complete the embedment of dN^{q-1} guest graphs of M node into Hy-perDC perfectly.

We consider guest graph as complete binary tree of 7 nodes and host graph as HyperDC of 196 nodes and realize the result of embedment of $G_{0,j}, j = 0, 1, \dots, d-1$, as shown in the Fig.7. Through the above process of embedment, the number of realizing guest graph embedded into host graph is $196d$ and the number of edges is $168d$. After being embedded, the length of every edge is 1 and the embedment rate of node is 1 to realize the optimal

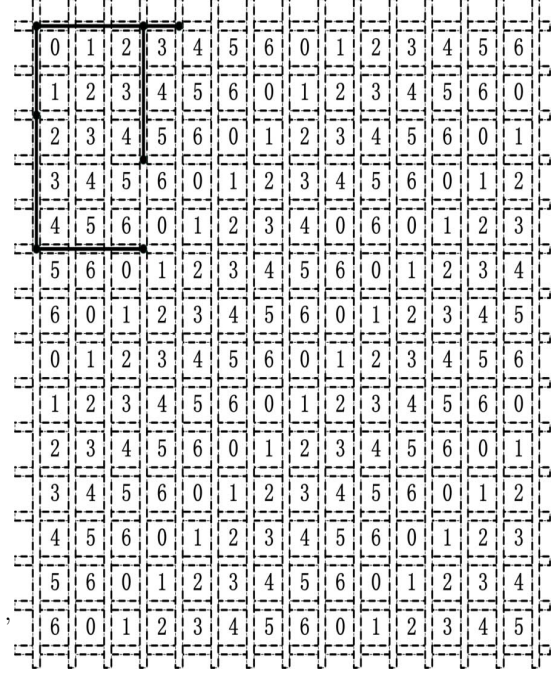


Figure 8. Embedment of guest graph into HyperDC.

index of load, congestion, dilation and expansion. This method belongs to optimal embedment.

VI. COMPARISON OF PROPERTIES

To evaluate the performance of HyperDC topology, we take into consideration properties of latency, link complexity, switch complexity and scalability compared to other non-blocking topologies. The non-blocking property of Benes topology as a kind of MINs has already been proved. According to the section 2.1, we can know that DBHC owns the non-blocking property. Thamarakuzhiput forward 2-dilated flattened butterfly (2DFB) topology based on flattened butterfly topology and stated a kind of conflict-free static routing schedule for its non-blocking property. Szymanski proved indirectly the non-blocking property of 2-dilated hypermesh (2DHM) in his paper. By comparing the Hy-perDC with several excellent topologies, the paper proves that HyperDC topology owns outstanding topology performances.

A. Latency

Latency property of topologies depends on diameter of topology. We set the number of ETs included in the all topologies as N and one SE can carry k ETs in DBHC, 2DFB and HyperDC topology. Thinking about Benes topology, as a kind of MINs topology, it needs a large amount of middle stage for transmitting the data.

Benes topology: $2\log_2(N) - 2$

DBHC topology: $\log_2(N/k)$

2DFB topology: $\log_k(N/k)$

2DHM topology: $\log_k(N)$

HyperDC topology: $\log_k(N/k)$

The diameter comparison is depicted in Fig.8. As we can know from this figure, Benes topology has the largest network diameter, 2DFB topology and HyperDC topology has the smallest diameter compared to other topologies.

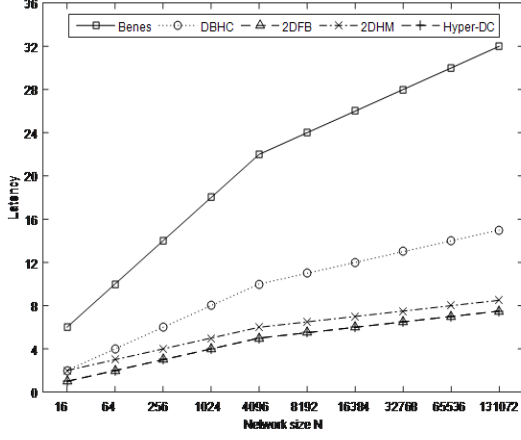


Figure 9. Relationship between latency and N.

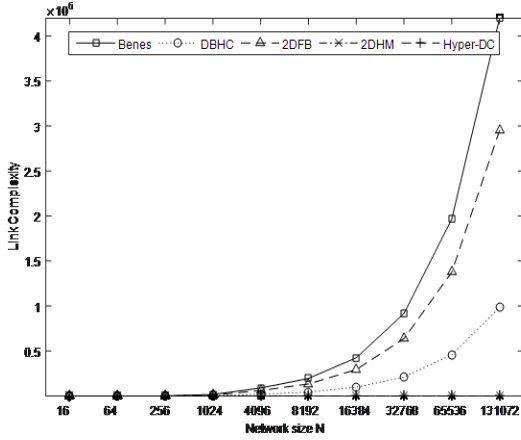


Figure 10. Relationship between link complexity and N.

B. Link complexity

Link complexity is defined as the total number of links. Link complexity measures the link cost and operational expense. The cost of a topology depends mainly on link complexity. With the same number of SEs, we have compared link complexity of the above five network topologies as shown in Fig.9. As we can observe, due to the use of bus lines, 2DHM topology and HyperDC topology owns the smallest link complexity when compared to the left topologies.

C. Switch Complexity

Switch complexity is defined as the total number of switch. It can measure the cost of network topology com-

paring with link complexity. The switch complexity comparison is shown below.

Benes topology: $N(2\log_2 N - 1)/2$

DBHC topology: N/k

2DFB topology: N/k

2DHM topology: N

HyperDC topology: N/k

As we can know from the above formulas, DBHC topology, 2DFB topology and HyperDC topology possess the same number of switches, which is less than Benes topology and 2DHM topology.

VII. SCALABILITY

Scalability property can measure the key index of carrying the number of ETs in the condition of the same dimension of topology, which signifies that a topology can support the bigger scale of servers in condition of same latency property. The excellent scalability property can be beneficial for building the large-scale data center networks and don't affect the other properties. According to the above parameter request and the scalability comparison is shown in Fig.10. We can observe from this picture, 2DFB topology and HyperDC topology has perfect scalability property compared to the other topologies. In conclusion, considering about the aspects of latency, link complexity, switch complexity and scalability, HyperDC topology have huge advantages compared to other topologies which have better properties. It can satisfy the development demand of the large data center networks in future.

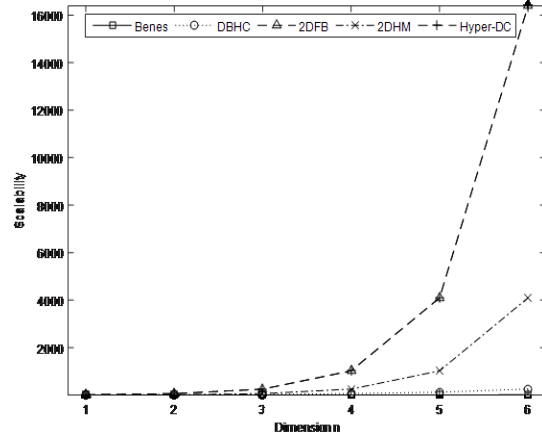


Figure 11. Relationship between scalability of n.

VIII. CONCLUSION

The paper synthesizes the advantages of the former direct topologies and removes the disadvantage and put forwards the HyperDC topology creatively. The topology is based on the hyperedge network topology and every switch element can carry several ETs to increase the scale of topology. HyperDC topology possesses the rearrangeable non-blocking property by doubling the number of channels, which is beneficial for linear transmission

of the flow and simplification of cache in data center networks and increase the speed of reaction. The paper takes advantage of mathematical induction to prove the rearrangeability non-blocking of HyperDC topology innovatively by transforming the direct topology into the indirect topology. By the analysis of the relative theory, the paper explains the embedment of HyperDC, which provides the conditions for realization of numerous parallel algorithms in HyperDC topology. These are all that we study now and the next work is that combining the requirement of load balance of data center networks topology and the flow control method of the flowlet for further optimizing the properties of HyperDC topology.

REFERENCES

- [1] Li D, Chen G H, Ren F Y, et al. Data Center Network Research Progress and Trends[J]. Chinese Journal of Computers, 2014, 2: 259-274.
- [2] Gang D, Zhenghu G, Hong W. Characteristics Research on Modern Data Center Network[J]. Journal of Computer Research and Development, 2014, 2: 017.
- [3] Niranjana Mysore R, Pamboris A, Farrington N, et al. Portland: a scalable fault-tolerant layer 2 data center network fabric[C]//ACM SIGCOMM Computer Communication Review. ACM, 2009, 39(4): 39-50.
- [4] Guo C, Wu H, Tan K, et al. Dcell: a scalable and fault-tolerant network structure for data centers[J]. ACM SIGCOMM Computer Communication Review, 2008, 38(4): 75-86.
- [5] Wu H, Lu G, Li D, et al. MDCube: a high performance network structure for modular data center interconnection[C]//Proceedings of the 5th international conference on Emerging networking experiments and technologies. ACM, 2009: 25-36.
- [6] Thamarakuzhi A, Chandy J A. 2-Dilated flattened butterfly: A non-blocking switching topology for high-radix networks[J]. Computer Communications, 2011, 34(15): 1822-1835.
- [7] Singla A, Godfrey P B, Kolla A. High throughput data center topology design[C]//11th USENIX Symposium on Networked Systems Design and Implementation (NSDI 14). 2014: 29-41.
- [8] Kim J, Dally W J, Towles B, et al. Microarchitecture of a high-radix router[C]//ACM SIGARCH Computer Architecture News. IEEE Computer Society, 2005, 33(2): 420-431.
- [9] Binkert N, Davis A, Jouppi N P, et al. The role of optics in future high-radix switch design[C]//Computer Architecture (ISCA), 2011 38th Annual International Symposium on. IEEE, 2011: 437-447.
- [10] Farrington N, Porter G, Radhakrishnan S, et al. Helios: a hybrid electrical/optical switch architecture for modular data centers[J]. ACM SIGCOMM Computer Communication Review, 2011, 41(4): 339-350.
- [11] Yin Y, Proietti R, Ye X, et al. LIONS: An AWGR-based low-latency optical switch for high-performance computing and data centers[J]. Selected Topics in Quantum Electronics, IEEE Journal of, 2013, 19(2): 3600409-3600409.