Optically Cross-Braced Hypercube: a Reconfigurable Physical Layer for Interconnects and Server-Centric Datacenters

Henggang Cui[1,2], Danielle Rasooly[2], Moises R. N. Ribeiro[2,3], and Leonid Kazovsky[2]
1-TsinghuaUniversity (China), 2-Stanford University (USA), 3-Federal University of Espirito Santo (Brazil)
cuihenggang@gmail.com, drasooly@stanford.edu, moises@ele.ufes.br, l.kazovsky@stanford.edu

Abstract: Our proposal is to gradually deploy 2x2 optical switches to hypercubes planes in order to decrease about 15% of transit traffic processing for bidirectional physical connections and over 20% forwarding traffic in unidirectional links.

OCIS codes: (060.4250) Networks; (200.4650) Optical interconnects.

1. Introduction

Datacenter network architectures based on commodity equipment can be divided into network-centric designs and server-centric designs [1]. In network-centric approaches, servers are interconnected by a hierarchy cluster of high-end switches with large number of ports [2]. In server-centric networks, servers are not only computing units but also routing nodes that actively participate in packet forwarding and load balancing. Server-centric approaches are more scalable and may significantly reduce cost, so that they are good candidates for building the huge future data centers. However, the burden of multi-hop forwarding on server computation resources and their interfaces bandwidth and the increased latency are challenge for server-centric networks yet to be addressed (e.g., up to 90% of CPU utilization in [1] is taken by traffic forwarding). A similar problem is faced by architectures for microprocessor interconnects since a significant part of processing power maybe consumed in the handling of transit traffic [3].

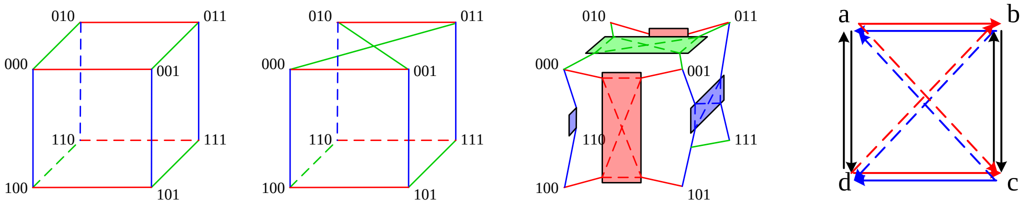
Optical switching is a promising solution both to server-centric datacenter and interconnects since it provides shortcuts across multi-hop networks [4]. However, large optical switches to meet that end are not only expensive; they also have other issues such optical crosstalk. It is likely that an optical solution to be successful in the network-centric datacenters and microprocessor interconnect arenas should try and avoid the intrinsic downsides of the network-centric architecture. In other words, a more distributed solution must be sought for building the reconfigurable physical layer needed for reducing the negative impact of multi-hop routing. The novelty of our proposal is to present an optical cross-bracing reinforcement to the interconnection architecture of server-centric datacenters and microprocessor interconnects but without changes in node degree. The case investigated in this paper looks at how simple 2x2 optical switches combined with the traditional hypercube topology can create short cuts to the heavy flows in the network could be adaptively create and thus reduce their impact on the intermediary nodes and the total traffic in the network without affecting the original hypercube routing scheme.

2. Hypercubes and Optical Cross Bracing

Hypercube is a widely studied network topology that offers benefits such as small diameter, high connectivity, symmetry, simple control and routing, and fault tolerance. It is the underlying architetural of diverse recent proposals, such as [1], [5], and [6]. A hypercube network of dimension n connects up to 2^n nodes (each of which can thus be identified uniquely with n bits), using a physical connection between two nodes if and only if their n-bit addresses differ in exactly one bit position. Each node links to those nodes with Hamming distance of 1, and to reach those with Hamming distance larger than 1, they need to use other nodes as intermediary nodes. Our proposal is to dynamically reconfigure connections in order to reduce the average number of hops for each packet but without affecting the degree of those nodes and without imposing modifications in the original routing strategy of hypercubes. The aim is to keep the servers (or microprocessors) totally agnostic about the underlying physical topology so that cross-layer communication for routing purposes would not be necessary. This isolation between topological layers enables optical switching to be gracefully deployed.

As an illustrative case, a (hyper)cube is presented in Figure 1(a). If most of the traffic that goes through link 000-010 would also go through link 010-011, it would be better to rewire the original green link to connect node 000 and 011 directly as illustrated in Figure 1(b). After the rewiring, node 000 would still send the packets through the green link, because it is not aware of the rewiring. However, these packets would not reach the intermediary node 010, but go directly to the destination node 011, and the other little traffic that should not reach 011 would return to

010 through link 011-010 by taking advantage the fault tolerance of hypercube routing. We can use six 2x2 optical switches to achieve this dynamic rewiring as illustrated in Figure 1(c).



(a) Original interconnection. (b) Rearranged connections. (c) Optical switches fro rearrangement. (d) Traffic on a given plane.

Fig. 1. Optical cross-bracing of a (hyper)cube.

For a general case of *n*-dimension hypercube, we define the *k*-th bit of the address (from left to right) as dimension *k*, the links that connecting nodes with their *k*-th bits different as dimension k links, the planes formed by two dimension *k* links and two dimension *k*+1 links as dimension *k* planes. Here *k* is from 0 to *n*-1, and *k*+1 equals 0 if *k* equals *n*-1. We will deploy an optical switch on each plane formed by dimension *k* links and dimension *k*+1 links by connecting the two dimension *k* links to the optical switch. There would be *n* x $2^n/4$ optical switches in total. In case of bidirectional links, this number is obviously doubled. This partially reconfigurable scheme may take far less optical switches than a centralized optical switch. In addition, signal regeneration is performed in each hop and there is no cascading of optical switches (multi-stage) to build a large switching matrix. This feature enables low-end (e.g., high insertion loss and high crosstalk) optical switches to be used.

2.1 Controlling the Optical Switches

A controller should decide, on plane-to-plane basis, about the state (i.e., either *cross* or *bar*) the optical switches should be at a given moment in time. The goal is to reduce transit traffic by using information collected about the flows sizes (or bandwidth demand) in the network. However, in order not to exceed the fault tolerance of hypercube routing mechanism for the other traffic not benefited by the rearrangement, constraints must be imposed on the state space of the optical switches. Not only loop-free routing should be ensured, but also the convergence for the controlling algorithm as multiple planes on *cross* state may cause route flapping. The first point is to decide whether a plane qualify to be place on *cross* state. For a certain dimension *k* plane formed by nodes {a, b, c, d}, dimension *k* unidirectional links {a⟶b, d⟶c, b⟶a, c⟶d}, and dimension *k*+1 unidirectional links {b⟶c, a⟶d, c⟶b, d⟶a}, which is illustrated in Figure 1(d), we define $F^{(0)}_{(i,j)}$ as the amount of traffic that goes through link (*i*,*j*) in the original hypercube (all optical switches on bar state), and $F^{(0)}_{(i,j),(j,l)}$ as the amount of traffic that goes through link (*i*,*j*) and link (*j*,*l*) sequentially. Transit traffic would be reduced by switching the link a⟶b and d⟶c when condition (1) is fulfilled and transit traffic would be reduced by switching the link b⟶a and c⟶d when condition (2) is met.

$$2F^{(0)}_{(a,b),(b,c)} + 2F^{(0)}_{(d,c),(c,b)} - F^{(0)}_{(a,b)} - F^{(0)}_{(d,c)} > 0 \quad (1)$$

$$2F^{(0)}_{(b,a),(a,d)} + 2F^{(0)}_{(c,d),(d,a)} - F^{(0)}_{(b,a)} - F^{(0)}_{(c,d)} > 0 \quad (2)$$

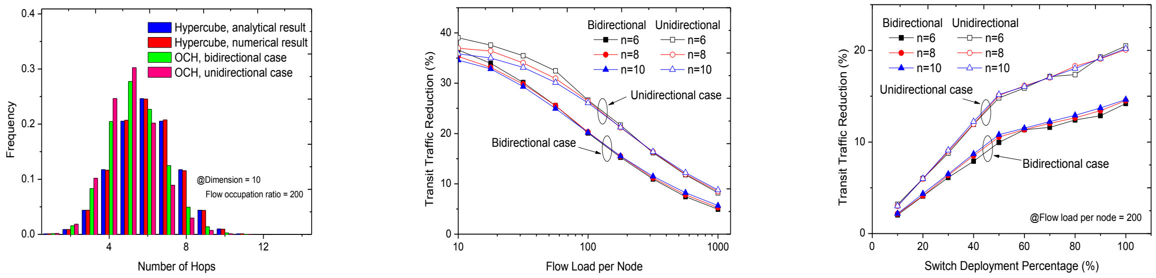
However, we cannot set to *cross* state all the planes that could yield transit traffic reduction, so we have to design an algorithm to maximize the total reduction of transit traffic in the network. For ensuring both loop-free routing and algorithm convergence, we have imposed the following constraint: each plane that meets the condition described in (1) and (2) would prevent all its neighbor planes to operate on *cross* state. This problem can be modeled as nodes in an auxiliary graph, and the other planes that could not be switched due to the switching of that plane as its neighbors (each node would have at most 4 neighbors for both bidirectional case and unidirectional case). Thus, the optimization problem of minimizing the total transit traffic can be abstracted to the so-called maximum weighted independent set problem, which is NP-hard. Fortunately, there is a fast approximate algorithm called GWMIN [7] to solve this problem within reasonable running time.

4. Results

An *ad-hoc* simulator was implemented using GWMIN to accommodate different scenarios of flows for diverse *n*-dimensional optically-braced hypercubes. The network loading is represented by the ratio of the number of total flows to the number of nodes in the network. According to the experimental data provided by [2], we assume that 1% of the total flows are heavy flows and the base-10 logarithms of their sizes follow normal distribution N(8,0.32). We generate traffic randomly according to this distribution between a pair of equally probable nodes and uniformly select a route (out of *n* routes) from the original hypercube for traffic balancing purposes. Therefore, the network is loaded as uniformly as possible to provide conservative comparisons for the benefit of optically cross bracing. The transit traffic reduction measures the ratio of optically cross-braced approach to original hypercube requires from its nodes. In order to validate the simulator, Figure 2(a) shows the statistic of hop distribution under a typical flow occupation ratio of 200 flows per node [3]. The blue bars are the theoretical hop distribution for random flows in the hypercube, while the red and green ones are simulation results. We find that the red and blue ones match well. After using optical switches to cross-brace the hypercube and GWMIN to decide the best configuration of optical elements on *cross* state, results show that (green bars for bidirectional case, and pink bars for unidirectional case), the packets in the network see, on average, fewer hops between a pair of nodes, which is the reason why our design could significantly reduce transit traffic processing in intermediate nodes and also latency to end user. Note, however, that there is no transparent lightpath and packets are still regenerated in every hop.

For evaluating the effect of loading in different network configurations, Figure 2(b) gives the result of transit traffic reduction under different flow occupation and hypercube dimensions for both unidirectional and bidirectional links. We find that if the flow occupation ratio is the same, the transit traffic saving would be very close in all hypercube dimensions. Our proposal would have transit traffic reductions between 34% and 40% under smaller flow occupation ratios such as 10 flows per node. The traffic in the network would more likely to be unbalanced in these situations due to the flow distribution itself, despite the fact that traffic balancing across multipath is performed when routing across hypercube. Using a typical loading of 200 flows per node, our design would give about 15% transit traffic reduction in bidirectional case and over 20% in unidirectional case. Therefore, by optically cross-bracing hypercubes we can add an extra degree of freedom in traffic balancing for interconnects.

Figure 2(c) brings the result of partial deployment, i.e., when only a part of the optical switches are added to the original hypercube topology, at 200 flows per node. The transit traffic is reduced at a steep rate as the optical switches are gradually deployed at the beginning, and, when around half of the planes are already cross braced, the incremental benefit of deploying an extra switch diminishes. This outcome suggests that partial deployment should be considered in cost analysis of such topologies as a means of maximizing investment.



(a) Transit traffic saving (b) Hop distribution (c) Switch partial deployment

Fig. 2 Optically cross-braced Hypercube performance evaluation.

5. Conclusions

We presented a case for optical 2x2 switches inexpensively bringing transparent switching to traditional hypercube network, so that we could reduce the impact of multi-hop forwarding of server-centric datacenter networks and optical interconnect architectures. Our proposal can save about 15% transit traffic for bidirectional case and over 20% transit traffic for unidirectional case. Economically wise partial deployment schemes are also possible in our approach as routing is left completely unaware of the physical topology. In addition, there are extra benefits in protection/restoration brought by incorporating optical switches that will be investigated in future studies.

6. References

[1] C. Guo, H.Wu, K. Tan, L. Shiy, Y. Zhang, and S. Lu. “BCube: A High Performance, Server-centric Network Architecture for Modular Data Centers”. SIGCOMM, 2009.

[2] A. Greenberg et al. “VL2: A Scalable and Flexible Data Center Network”. SIGCOMM, 2009.

[3] Madeleine Glick, "Optical Interconnects in Next Generation Data Centers: An End to End View," hoti, pp.178-181, 2008 16th IEEE Symposium on High Performance Interconnects, 2008.

[4] Guohui Wang, David G. Andersen, Michael Kaminsky, Konstantina Papagiannaki, T. S. Eugene Ng, Michael Kozuch, and Michael Ryan. “c-Through: Part-time Optics in Data Centers.” SIGCOMM, 2010.

[5] Ahmed Louri and Hongki Sung. “An Optical Multi-Mesh Hypercube: A Scalable Optical Interconnection Network for Massively Parallel Computing”. Journal of Light Wave Technology, Vol. 12, No. 4, April 1994.

[6] Toshikazu Sakano and Shuto Yamamoto. “Scalable Photonic Interconnection Network With Multiple-Layer Configuration for Warehouse-Scale Networks”. Journal of Optical Communication Networks, Vol. 3, No. 8, August 2011.

[7] Shuichi Sakai, Mitsunori Togasaki, and Koichi Yamazaki. “A Note on Greedy Algorithms for the Maximum Weighted Independent Set Problem.” Discrete Applied Mathematics 126 (2003) 313 – 322.