Feasibility Study on Topology Malleable Data Center Networks (DCN) Using Optical Switching Technologies

Ankit Singla[*], Atul Singh, Kishore Ramachandran, Lei Xu[**], and Yueping Zhang
[*]*Department of Computer Science, University of Illinois, Urbana–Champaign, IL, USA*
NEC Laboratories America, Inc., 4 Independence Way, Princeton, NJ, USA
[**]*Email: leixu@nec-labs.com*

Abstract: We recently presented a highly advantageous DCN architecture that supports on-demand, run-time topology malleability using WDM and optical space/wavelength switching technologies. We further investigate some implementation-related issues including scalability, hop-by-hop latency and optical power budget.

OCIS codes: (060.4258) Networks, network topology (060.6718) Switching, circuit

1. Introduction

With increasing demand from both conventional (e.g., email, messaging, multimedia and banking) and emerging (e.g., cloud computing) services, data center networks are facing challenges in supporting higher bandwidth and computing capacity while achieving lower power consumption and better manageability. Many new data center network (DCN) architectures have been proposed to improve the scalability, bandwidth, and fault tolerance [1-3]. However, these designs, based on the all-to-all electrical connectivity paradigm, suffer from high cabling complexity and power consumption. Further, they require expensive re-wiring to support higher bandwidth between servers (e.g. when migrating from 10Gbps to 40 Gbps). By utilizing optical communications and switching elements, new DCN architectures have been proposed that achieve significantly improved networking performance and lower power consumption [4-6] compared to conventional DCN architectures.

While *optical packet switching* is the ideal technology to support high-bandwidth interconnections, the associated performance and cost related issues create barriers for their practical deployments. This motivates the use of *optical circuit switching*-based devices and systems that have been generally commercialized and can support a large number of channels and high bandwidth WDM communications. In our recent work [6], we show a new DCN architecture (called *Proteus*), which uses WDM and optical wavelength/space switching to achieve run-time network topology reconfigurability according to the traffic demand dynamics, and supports higher bandwidth, simplified cabling, good scalability and network flexibility. In this paper, we further investigate some implementation-related issues for the *Proteus* DCN architecture, including scalability, latency of hop-to-hop communications, and optical power budget.

2. *Proteus* DCN architecture

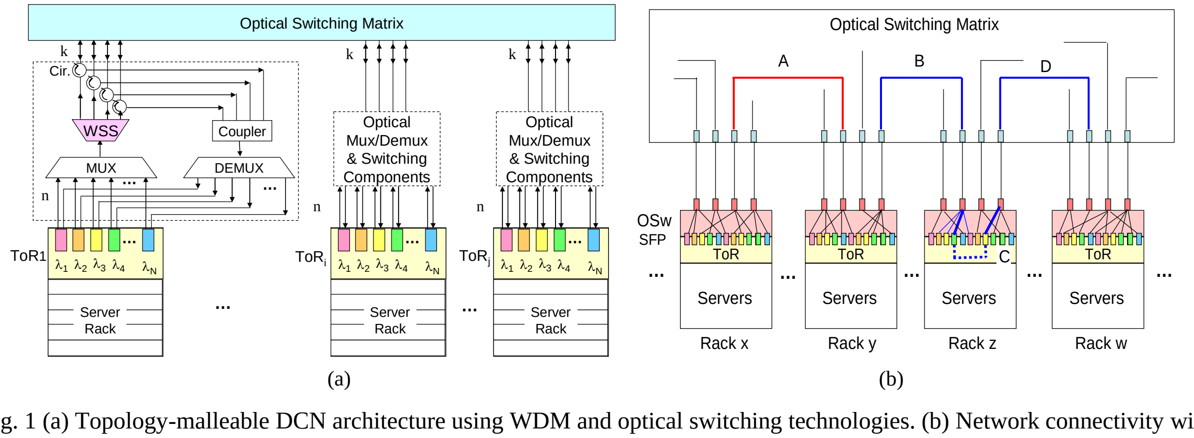


Fig. 1 (a) Topology-malleable DCN architecture using WDM and optical switching technologies. (b) Network connectivity with direct (link A) and hop-by-hop (link B-C-D) communications. MUX: optical wavelength multiplexer, DEMUX: optical wavelength demultiplexer; WSS: wavelength selective switch; Cir: optical circulator.

While a DCN can be over-provisioned (using all-to-all electrical connectivity) to satisfy wide range of traffic patterns, it may be cheaper and more energy-efficient if the DCN could adapt or be malleable to traffic, i.e.

providing on-demand, high-bandwidth connectivity between any reasonably sized subset of servers. Using malleability as the central property, we designed the *Proteus* DCN architecture in [6].

Fig. 1 (a) shows the *Proteus* DCN architecture, which includes three units: the server racks including the top-of-rack (ToR) switch and *n* WDM small form-factor pluggable (SFP) transceivers; the optical multiplexing/demultiplexing and switching unit; and the optical switching matrix (OSM) (e.g. micro-electro-mechanical systems/MEMS). In the current design, the number of WDM SFP transceivers is equal to the number of servers on the rack (usually 20-80 servers per rack). In the transmitting side, the multiplexed WDM signals from a ToR are grouped into *k* fiber ports by a *1×k* wavelength selective switch (WSS). All the server racks are connected to the optical switching matrix (OSM) through the *k* fiber ports from WSS. Here optical circulators are optional, which can support bidirectional communication of each port of the OSM. At the receiving side, WDM signals in *k* fibers are combined using an optical coupler (or WSS) and demultiplexed for receiving by the SFP transceivers on the ToR switch.

With this design, each ToR can communicate simultaneously with any *k* other ToRs, which implies that OSM reconfiguration allows the construction of all possible k-regular ToR graphs. Through WSS re-configuration at runtime, the capacity of each of the *k* links can be varied in 0 to *n* times the bit rate of the SFP transceivers. Note that the *Proteus* DCN design includes a centralized network manager (NM) that obtains the traffic matrix from the top-of-rack (ToR) switches, calculates appropriate configurations, and pushes them to the optical switching matrix, WSS, and ToRs [6]. The NM should also avoid the wavelength contention when WDM signals from *k* fiber ports are combined with the optical coupler in the receiving side.

Given a connected ToR graph, direct connection and hop-by-hop communications are two ways to achieve network connectivity, as shown in Fig. 1 (b). To reach ToRs not directly connected to it through the OSM, a ToR uses one of its *k* connections. The first hop ToR receives the transmission over fiber, converts it to electrical signals, reads the packet header, and retransmits it towards the destination. For efficient use of the network bandwidth, the NM should manage the topology such that high volume traffic uses the minimal number of hops. This requires direct, out-of-band connections between the NM and OSM/WSS/ToRs.

The next challenge is to identify the optimal topology given a traffic demand. At a high level, the problem can be formulated as a mixed integer linear program (MILP) and heuristic approach can be used to seek optimal/near-optimal solutions [6].

3. Implementation-related issues

3.1 Network Scalability

The number of servers that the *Proteus* DCN design can support is decided by the port count of the OSM, *k* (the number of WSS fiber switching ports), and *n* (the number of WDM SFP transceivers on each ToR, which is assumed to be equal to the number of servers per server rack). Here we define the over subscription factor (OSF), which is decided by *n/k*. Fig. 2 shows the number of servers supported by an OSM with a certain port count. Please note that OSF is the slope the lines in Fig. 2. With larger OSF factors, more WDM SFP transceivers are sharing the WSS fiber switching ports, and therefore, more servers can be supported with the same OSM. Larger OSF factors typically mean lower network implementation cost. With smaller OSF factors, fewer servers can be supported by the same OSM, but the network flexibility in terms of supporting different traffic patterns increases.

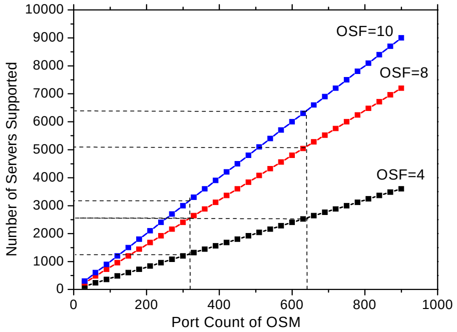


Fig. 2. Number of servers supported by different sizes of OSM. OSF: over-subscription factor, which is the slope of the line.

Table 1. Link power budget estimation

MUX and DEMUX (40 Channel Gaussian type, e.g. JDSU)	2.5 dB ×2
WSS (100GHz, 1×4)	~ 7 dB
Optical coupler (1×4) / WSS	~ 7 dB
Optical space switch (MEMS)	~ 3 dB
Optical Circulator (e.g. JDSU)	0.6 dB ×2
Fiber connector/splicing loss (total)	~ 1 dB
Optical fiber insertion loss (2 km standard single mode fiber)	< 1 dB
Total loss	**~25 dB**

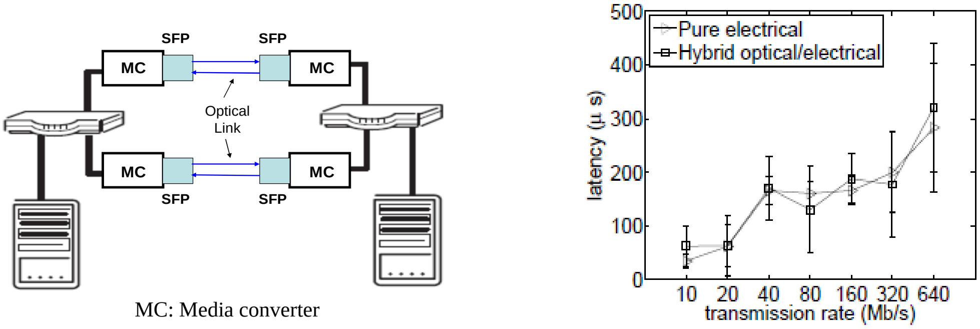
When the intended number of servers goes beyond what one OSM can handle at a required OSF factor, all the servers can be separated into different groups, and the different groups can be interconnected through OSMs. In a DCN with multiple OSM separated groups, the network may experience some performance limitations: (1) larger switching latency for the whole network reconfiguration and (2) bandwidth bottle neck when the OSM interconnection port count is small.

3.2 Optical power budget

Table 1 shows the estimated link power budget for the SFP transceivers, based on parameters from available optical components. With 3 dB extra optical power margin, the SFP transceivers should have a power budget of ~ 28 dB, which corresponds to commercial SFP units capable of transmission over 120 km standard single mode fibers.

3.3 Latency in hop-by-hop communications

At each hop, every packet experiences conversion from optics to electronics and then back to optics (OEO). The efficiency of such OEO conversion is essential to the feasibility of the *Proteus* DCN design. Figure 3 shows our initial results on the latency measurement on hop-by-hop routing. As illustrated in Fig. 3(a), our testbed is composed of four servers, each equipped with a single 2:33 GHz Intel Core 2 Duo processor, 4 GB RAM, and 3 GigE cards. Among these, two servers are configured as routers and connected by optical fibers. The conversion between optical and electrical signals is accomplished using a CTC Gigabit optical media converter (MC). The other two servers act as clients and generate network traffic at varying volumes. Furthermore, we create a routing loop by configuring the IP forwarding table of the routers. In each router, we deploy a netfilter kernel module and utilize the NF_IP_PRE_ROUTING hook to intercept all IP packets. We record the time lag between the instant when the packets first arrive in the network and when their TTL expires. This way, we are able to measure the multi-hop latency for OEO conversion and compare it with the baseline where all servers are directly connected using only electrical devices. In this experiment, we create a routing loop with four servers and four Gigabit SFP transceivers to measure the multi-hop latency. The measurement results shown in Figure 3(b) demonstrate that the OEO conversion generally does not incur additional switching latency, and that the overall hop-by-hop latency is within ~ millisecond level.



(a) (b)

Fig. 3. (a) Experimental setup for hop-to-hop testing with OEO conversion.
(b) Measured hop-to-hop transmission latency with SFP transceivers.

4. Discussions

Reconfigurable photonic networking technology is critical towards building future high-bandwidth, scalable and flexible DCNs using all-optical or hybrid optical/electrical switching [8]. Our *Proteus* DCN architecture leverages existing WDM and optical wavelength and space switching technologies to achieve run-time topology reconfigurability, and has preferable features including scalability, simple cabling and easier network bandwidth upgrade. Our future work includes incorporating fault tolerance, designing appropriate routing protocols, and performing experiments over a real prototype testbed.

5. References

[1] M. Al-Fares, A. Loukasass, and A. Vahdat, “A scalable, commodity data center network architecture,” in ACM SIGCOMM, 2008, pp. 63–74.
[2] A. Greenberg, J. R. Hamilton, N. Jain, et al, “VL2: a Scalable and Flexible Data Center Network,” in ACM SIGCOMM, 2009, pp. 51–62.
[3] R. N. Mysore, A. Pamboris, N. Farrington, et al, “PortLand: a Scalable Fault-Tolerant Layer 2 Data Center Network Fabric,” in ACM SIGCOMM, 2009, pp. 39–50.
[4] G. Wang, D. G. Andersen, M. Kaminsky, et al, “c-through: Part-time optics in data centers,” in ACM SIGCOMM, 2010.
[5] N. Farrington, G. Porter, S. Radhakrishnan, et al, “Helios: a hybrid electrical/optical switch architecture for modular data centers,” in ACM SIGCOMM, 2010.
[6] A. Singla, A. Singh, K. Ramachandran, et al, “Proteus: a topology malleable data center networks,” in ACM HotNets 2010.
[7] T. Benson, A. Anand, A. Akella, et al, ”the case for fine-grained traffic engineering in data-centers,” in USENIX INM/WREN, 2010.
[8] K. Bergman, “Optically Interconnected High-Performance Data Centers,” in ECOC 2010, paper Mo.2.D.1