Fully programmable and scalable optical switching fabric for petabyte data center

Zhonghua Zhu,[1] Shan Zhong,[1,*] Li Chen,[2] and Kai Chen[2]

[1]*CoAdna Photonics Inc. 733 Palomar Ave. Sunnyvale, CA, 94085, USA*
[2]*HKUST, Clearwater Bay, Hong Kong, China*
**shanz@coadna,com*

Abstract: We present a converged EPS and OCS switching fabric for data center networks (DCNs) based on a distributed optical switching architecture leveraging both WDM & SDM technologies. The architecture is topology adaptive, well suited to dynamic and diverse *-cast traffic patterns. Compared to a typical folded-Clos network, the new architecture is more readily scalable to future multi-Petabyte data centers with 1000 + racks while providing a higher link bandwidth, reducing transceiver count by 50%, and improving cabling efficiency by more than 90%.

OCIS codes: (060.4250) Networks; (060.4253) Networks, circuit-switched; (060.4258) Networks, network topology; (060.4255) Networks, multicast; (060.6718) Switching, circuit.

References and links

1. Cisco Networks white paper, "Cisco global cloud index: forecast and methodology, 2013-2018." (Cisco Systems, 2013), *http://www.cisco.com/c/en/us/solutions/collateral/service-provider/global-cloud-index-gci/Cloud_Index_White_Paper.html*
2. IEEE 802.3 25 Gb/s Ethernet Study Group, *http://www.ieee802.org/3/25GSG/*
3. M. Al-Fares, A. Loukissas, and A. Vahdat, "A scalable, commodity data center network architecture," in *Proceeding of ACM SIGCOMM,* 18 (2008).
4. S. Peng, R. Nejabati, B. Guo, Y. Shu, G. Zervas, S. Spadaro, A. Pages, and D. Simeonidou, "Enabling multi-tenancy in hybrid optical packet/circuit switched data center networks," in *Proceedings of ECOC* (2014), paper Tu1.6.4.
5. G. Porter, R. Strong, N. Farrington, A. Forencich, P.-C. Sun, T. Rosing, Y. Fainman, G. Papen, and A. Vahdat, "Integrating microsecond circuit switching into the data center," in *Proceedings of SIGCOMM* (ACM, 2013), pp. 447–458.
6. G. Wang, D. G. Andersen, M. Kaminsky, K. Papagiannaki, T. S. E. Ng, M. Kozuch, and M. Ryan, "c-Through: part-time optics in data centers," in *Proceeding of SIGCOMM* (ACM, 2010), pp. 327–338.
7. K. Chen, A. Singla, A. Singh, K. Ramachandran, L. Xu, Y. P. Zhang, X. Wen and Y. Chen, "OSA: an optical switching architecture for data center with unprecedented flexibility," Networking, IEEE/ACM Trans. **22**(2), 498–511 (2013).
8. J. Kim, W. J. Dally, and D. Abts, "Flattened butterfly: a cost-efficient topology for high-radix networks," in *Proceeding of ISCA* (2007), pp. 126–137.
9. J. Kim, W. J. Dally, S. Scott, and D. Abts, "Technology-driven, highly-scalable dragonfly topology," in *Proceedings of the 35th International Symposium on Computer Architecture (ISCA-35)* (2008), pp. 194–205.
10. Z. Zhu and S. Zhong, "Scalable and topology adaptive intra-data center networking enabled by wavelength selective switching," in *Proceedings of OFC/NFOFC* (2014), paper Th2A.
11. Y. Ishii, M. Wakamiya, A. Kanagawa, K. Hadama, J. Yamaguchi, and Y. Kawajiri, "MEMS-based 1×43 wavelength-selective switch with flat passband," in *Proceeding of ECOC* (2009) PDP.
12. J. Perry, A. Ousterhout, H. Balakrishnan, D. Shah, and H. Fugal, "Fastpass: a centralized "zero-queue" datacenter network," in *Proceeding of Sigcomm* (ACM, 2014) pp. 307–318.
13. T. Benson, A. Anand, A. Akella, and M. Zhang, "The case for fine-grained traffic engineering in data centers," in *Proceeding of INM/WREN* (USENIX, 2010), pp. 2
14. S. Kandula, J. Padhye, and P. Bahl, "Flyways to de-congest data center networks," in *Proceeding of HotNets* (2009).
15. S. Cheung, T. Su, K. Okamoto, and S. J. B. Yoo, "Ultra-compact silicon photonic 512 x 512 25 GHz arrayed waveguide grating router," IEEE J. Sel. Top. Quantum Phys. **20**, 1 (2014).
16. M. A. Popovi, T. Barwicz, M. S. Dahlem, and F. W. Gan, "Tunable, fourth-order silicon micro-ring-resonator add-drop filters," in *Proceeding of ECOC* (2007).
17. Z. Wang, W. Chen, Z. Zhu, and Y. J. Chen, "Design of wavelength-selective switch Using micro-ring resonators," in *Proceedings of IPRA* (2005), paper IWE2.

#226742 - $15.00 USD Received 11 Nov 2014; revised 10 Jan 2015; accepted 14 Jan 2015; published 5 Feb 2015

18. L. Chen, K. Chen, H. Wang, Z. Zhu, M. Yu, G. Wang, S. Zhong, P. X. Gao, and C. M. Qiao, "Mega-switch: a massive-port optical switch architecture for large scale data-intensive communications in data center Networks, " manuscript submitted to Sigcomm 2015.
19. A. Azad, M. Halappanavar, S. Rajamanickam, E. G. Boman, A. Khan, and A. Pothen, "Multithreaded algorithms for maximum matching in bipartite graphs," in *Proceeding of Parallel & Distributed Processing Symposium (IPDPS)* (IEEE, 2012), pp. 860–872.
20. N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, and L. Peterson, "Open-flow: enabling innovation in campus networks," in *Proceeding of ACM CCR* (2008).
21. OpenDaylight website, *http://www.opendaylight.org/*.
22. M. Technologies, "Freebase Wikipedia extraction (WEX)," (2010), *http://download.freebase.com/wex/*.
23. P. Skomoroch, "Wikipedia traffic statistics dataset," (2009), *http://aws.amazon.com/datasets/2596*.
24. J. C. Mogul and L. Popa, "What we talk about when we talk about cloud network performance," in Proceeding of ACM SIGCOMM Comput. Commun. Rev. **42**(5), 44–48 (2012).
25. N. Farrington, G. Porter, S. Radhakrishnan, H. H. Bazzaz, V. Subramanya, Y. Fainman, G. Papen, and A. Vahdat, "Helios: a hybrid electrical/optical switch architecture for modular data centers," in *Proceeding of SIGCOMM* (ACM, 2010), pp. 339–350.
26. Y. Liu, X. Gao, B. Wong, and S. Keshav, "Quartz: a new design element for low-latency DCNs," in *Proceeding SIGCOMM* (ACM, 2014), pp. 283–294.
27. Y. Peng, K. Chen, G. Wang, W. Bai, Z. Ma, and L. Gu, "Hadoop-watch: a first step towards comprehensive traffic forecasting in cloud computing," in *Proceeding of INFOCOM* (IEEE, 2014) pp. 19–27.
28. M. Chowdhury, M. Zaharia, J. Ma, M. I. Jordan, and I. Stoica, "Managing data transfers in computer clusters with orchestra," in *Proceeding of SIGCOMM* (ACM, 2011), pp. 98–109.
29. M. Chowdhury, Y. Zhong, and I. Stoica, "Efficient coflow scheduling with Varys," in *Proceeding of SIGCOMM* (ACM, 2014), pp. 443–454.
30. X. W. Lin, W. Hu, X. K. Hu, X. Liang, Y. Chen, H. Q. Cui, G. Zhu, J. N. Li, V. Chigrinov, and Y. Q. Lu, "Fast response dual-frequency liquid crystal switch with photo-patterned alignments," Opt. Lett. **37**(17), 3627–3629 (2012).
31. M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M. J. Franklin, S. Shenker, and I. Stoica, "Resilient distributed datasets: a fault tolerant abstraction for in-memory cluster computing," in *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation* (USENIX Association, 2012), pp. 2

1. Introduction

Data centers networks (DCNs) are approaching Petabyte/s scale [1]. As DCNs and their underlying applications are scaling up the challenge lies in designing an upgradeable network architecture scalable in both size (number of network nodes or racks from hundreds to 2k +) and bandwidth (from current 10Gbaud single lane rate to 25Gbaud [2] for emerging 100Gb/s or 400Gb/s links) while reducing the cost and power consumption per byte transmitted. Multi-tier electrical switch architectures (e.g., folded Clos or fat-tree [3]) encounter difficulties in driving the cost down when scaling up as they require a large number of high-radix and high performance switches, transceivers and fibers at each layer as well as high operational power. As a result, the data center community has resorted to optical networking to lower the cost, power, and network complexity while achieving high bandwidth and network flexibility in DCNs.

There are basically 2 kinds of optical approaches, one is to operate the optical switching at packet level or rely on TDM to deal with volatile DCN traffic [4,5]. However, this approach requires not only fast optical switching, but also complicated optical label processing or a synchronized network environment, which jeopardizes the scalability of the system and leaves it less competitive against electrical solutions. The second category focuses on the flow switching and overlay optical channel switch (OCS) above electrical packet switch (EPS) as a hybrid system [6,7]. With the progress of application driven networking and software defined networking (SDN), the overlay OCS resource can efficiently bypass long background traffic, reduces the end-to-end latency and improves the over-all network efficiency. This approach builds on a mature optical technology that is able to drive costs down and thus becomes a practical alternative to electrical solutions. However, the existing optical approaches [6,7] typically are based on large centralized optical switching engines and rely on a hierarchical structure to support 1k + racks connectivity, this limits the system scale and also poses single-point failure reliability concerns.

#226742 - $15.00 USD Received 11 Nov 2014; revised 10 Jan 2015; accepted 14 Jan 2015; published 5 Feb 2015

On the other hand, folded Clos or fat-tree is not the only topology that fits for future high performance cloud computing applications including big data analytics. In facts, new architectures and topologies, such as flattened butterfly and dragonfly [8,9], were proposed not long ago to enable next generation large scale DCN but call for new switching technologies to bring them into reality. In this paper, we introduce such a distributed switching platform which is scalable to connect 1k + racks and 100k + ports. While supporting the flattened butterfly topology, the proposed architecture is fully flat that matches the physical two dimensional configurations of DCNs. The design goals for this system include:

1) A distributed optical/electrical hybrid switching system that reduces or preferably eliminates the core-switches.

2) Use of an OCS system rather than optical packet (OPS) or optical burst (OBS) switches due to cost concerns of the latter technologies and make it feasible by leveraging the commercial maturity of the former.

3) Utilize DWDM technology but limited to low cost DWDM transceivers and single span transmission (e.g. avoiding EDFA cascading).

4) Support diverse connectivity patterns (denoted as *-cast) including unicast, multicast, in-cast and all-cast by utilizing a broadcast and select architecture.

5) Minimize the risk of single point failure with passive optical signal by-pass.

6) Enable a pay-as-you-grow economic model.

The rest of this paper is organized as follows: In section 2, we introduce a generic optical switching infrastructure, the optical virtual switch (OvS) network, which meets these design goals. The key idea of this optical switching platform is the combination of Dense Wavelength Division Multiplexing (DWDM) and Space Division Multiplexing (SDM). Our proposed embodiment utilizes an Nx1 Wavelength Selective Switch (WSS); however, alternative implementations by wavelength selective filters are also presented. In section 3 we discuss the scalability of the proposed system and also present a wavelength switching algorithm that supports a large scale system with improved failure tolerance. In section 4, we present a concept proof implementation of OvS which using a centralized control plane and the OpenFlow protocol. We include test results of both physical and networking layers to demonstrate the clear advantages of OvS enabled networks for future high performance cloud computing and big data applications. Sections 5 and 6 discuss advantages and issues, proposed future directions, and conclusions.

2. WDM/SDM enabled optical switching fabric

2.1 Basic design

A high radix switch structure is usually required to enable the full mesh, non-blocking connection in a large scale flattened butterfly or dragonfly [8-10] network. For example, Fig. 1(a) shows a desired switching network with flattened butterfly topology, where each square block in the Fig. 1(a) represents a switching node. Since optical wavelengths switching technology allows the switching in wavelength domain, it enables a cost effective way to achieve a high-radix switching node to support above architectures. In this paper, we introduce such an optical wavelength switching system that is able to support over 100k ports (users) while achieving the design goals outlined in section 1.

#226742 - $15.00 USD Received 11 Nov 2014; revised 10 Jan 2015; accepted 14 Jan 2015; published 5 Feb 2015

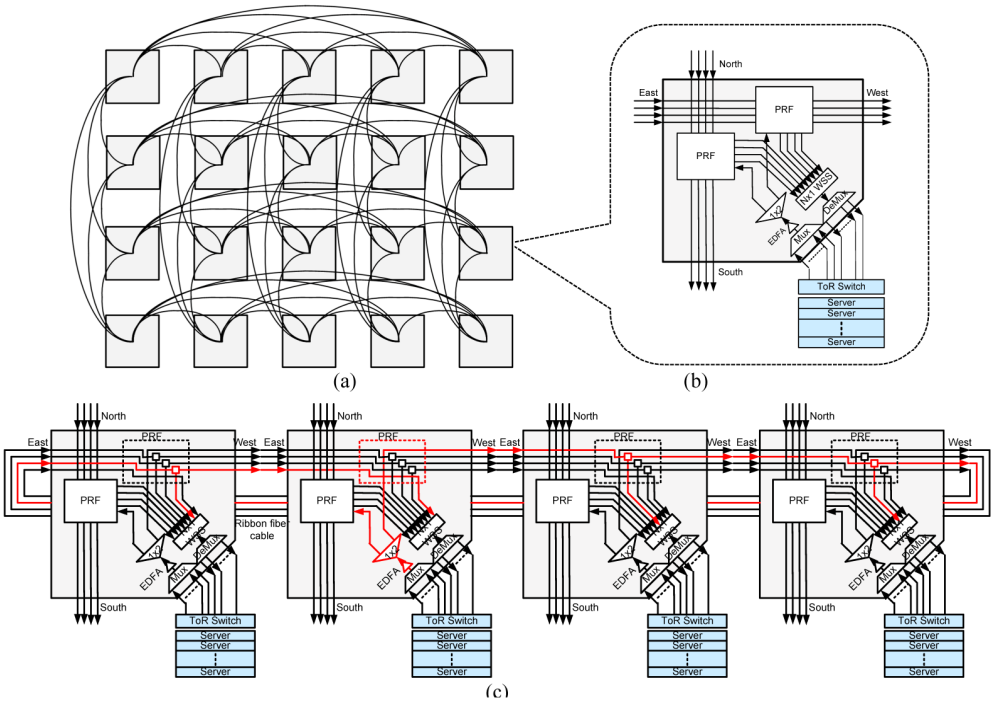


Fig. 1. OvS architecture: (a) Full-mesh connection of a flattened butterfly network; (b) Optical schematic design of OvS (2D); (c) Achieving full mesh connection among 4 nodes/racks using distributed WXC design.

As shown in Fig. 1(b), the building block of our proposed switching system, called optical virtual switch (OvS) box, is an add-on box that tends to provide full-mesh connections along both east–west and north–south directions. Fig. 1(c) illustrates how the OvS boxes are connected and what is the principle to achieve full-mesh connections, where:

1) Each ToR switch will have its uplink ports equipped with a set of DWDM transceivers that are interfaced through a DWDM Multiplexer/De-multiplexer (MUX/DMUX) pair.

2) The MUXed DWDM signals are boosted through an EDFA and then optically broadcast to all other nodes through a series of passive tap couplers organized as passive routing fabric (PRF) (see detail in Fig. 1(c)).

3) At the receiver end, a wavelength selective component, for example a wavelength selective switch (WSS), dynamically selects a combination of DWDM signals coming from other nodes/racks and passes them to its associated ToR switch.

4) The ToR switch performs traffic-aggregation as well as wavelength/port selection for the whole switching system to achieve end-to-end connection.

Focusing on one dimension, Fig. 1(c) shows a 4-node OvS network that interconnects 4 top of rack (ToR) switches. The red trace in the figure represents an example of the DWDM optical path from one node to all other nodes. DWDM optical signals are dropped and continued at each node along the path. WSS at each node is used to select the right wavelength channels to TOR. Thus a publishing and subscribing network is formed by these optical taps which assure a non-blocking switching system. Full-mesh connections among all 4 nodes/racks are achieved by overlaying 4 identical optical paths aforementioned. This design is originally based on the wavelength cross-connection (WXC) system in long-haul

#226742 - $15.00 USD Received 11 Nov 2014; revised 10 Jan 2015; accepted 14 Jan 2015; published 5 Feb 2015

and metro optical transport networks. Intrinsically, with an Nx1 WSS switching M wavelengths, one WXC can route M non-blocking channels to N ToR nodes. Thus, if all M uplink ports of the ToR switch are equipped with DWDM transceivers as shown in Fig. 1(b) and 1(c), we can create a full-mesh network among N ToRs with MxNxN possible interconnection configurations.

To achieve a uniform broadcasting loss across all the nodes the tap ratio of the optical tap coupler tree takes into account the in/out connector losses as well as the broadcasting losses along the distributed broadcasting route. This network of by-passing tap couplers defines a unified passive routing fabric (PRF) module for each node, as highlighted in a red box in Fig. 1(c). Taking advantage of this uniform passive routing structure, we then use only a simple ribbon fiber cable to connect adjacent nodes in a ring form network thereby achieving a meshed fiber connection physically among all the nodes. This OvS along with the PRF and ribbon fiber cables then constitute the building blocks for a fully scalable, low-cost, fault-tolerant DCN.

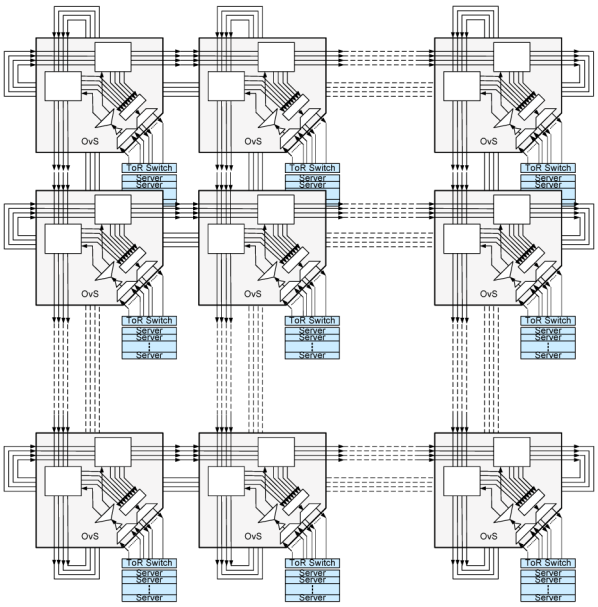


Fig. 2. An OvS network under 2-D configurations, which is connected through ribbon fiber cable.

To scale up the network to 1k + nodes, the basic OvS network is extended to 2 dimensions (see Fig. 2) which also reflects the physical placement of server racks as 2D array in data centers. This forms a 2D flattened butterfly network [8,9]. Accordingly, the OvS box for 2D arrays is designed to support both east–west and north–south ribbon cable connections as shown in Fig. 1(b). With today's technology, the N of an Nx1 WSS can be as high as 32 [11] at reasonable cost. A 33x33-array, 2-fly network is then achievable to connect up to 1089 racks. By leveraging the standard 50GHz wavelength spacing of DWDM technology, each fiber can potentially carry close to 100 wavelengths at 10 Gbs in the C-band, so the 2D OvS network can support 100k + ports and push the bisection BW to ~***1Pb/s*** with the existing low-cost technology. Although the topology provides plenty of path diversity, the node distance between any two racks can be kept to no more than 2 to maintain low latency anywhere. Also, the proposed symmetric network topology allows simple addressing thus reducing the computational overhead for routing calculations.

Although the 2D OvS platform achieves a full-mesh connection within each dimension, each OvS only needs to connect to its direct neighbor. The cabling complexity is O(N) where

#226742 - $15.00 USD Received 11 Nov 2014; revised 10 Jan 2015; accepted 14 Jan 2015; published 5 Feb 2015

N is the rack number. We only need 2178 optical cables to interconnect 1089 racks. Basically, long cables are avoided by introducing a folded torus cabling system, which greatly simplifies the cabling management and system upgrading.

Big data applications impose heavy bandwidth demands with diverse communication patterns (denoted as *-cast) that mix together unicast, multicast, in-cast, and all-to-all-cast traffics. For example, Hadoop and Spark requires in-cast traffic delivery during the shuffle stage of MapReduce, but requires multicast for data replication, parallel database join operation, as well as data dissemination in virtual machine (VM) provisioning. In OvS network, *-cast is supported. This is achieved by the efficient optical multicast/broadcast through optical splitters. Since DWDM is supported along each of the fibers, any pair of nodes can generate a one-to-one link using one or a group of wavelengths. For each wavelength channel, the OvS design is inherently a publish-and-subscribe type network. By default, we achieve an all-to-all-cast network. For a unicast connection, the traffic is only selected at its destination node. For multicast or broadcast connection, the OvS can be selectively configured to accept or block the signals at each node. Figure 3(a) illustrates an optical multicast case from node A to nodes B, D and E using the same wavelength (blue line) that originated from node A. In 2D OvS networks *-cast diverse traffic pattern can still be supported. If the multicast signals need to cross into the other dimension, the signals can be electrically relayed to the next dimension at the conjunction node (s)/rack(s).

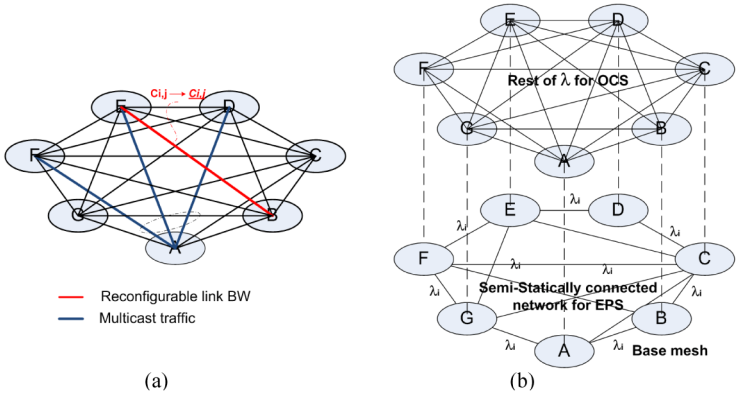


Fig. 3. OvS networking: (a) OvS supports *-cast traffic and reconfigurable link BW; (b) OvS to integrate EPS (base mesh) and OCS.

Different from the network architecture in [6] where the EPS and OCS fabrics remain separate, the OvS network overlay EPS fabric above OCS fabric. The network resources are managed seamlessly and thus the network efficiency is maintained: some wavelengths on some links are assigned to form a base connected graph (called a base-mesh in Fig. 3(b)) among all nodes for packet switching network. Short and volatile traffic are tagged and sent to the base-mesh. The links' BW and network topology of the base mesh can be further semi-statically optimized to adapt to traffic affinity where long steady traffic is assigned to use separate resources from the base-mesh by the control plane, thus allowing the base-mesh EPSs to maintain low-buffering [12] and achieve nearly cut-through packet switching instead of store & forward. Studies [13,14] have shown the data center background traffic is stable on the order of seconds. In particular, it is noted in [14] that 60% of ToR-pairs see less than 20% change in traffic demand between 1.6 to 2.2 seconds on average. The OCS design is well adapted for long steady traffic patterns thus the existing OCS with millisecond switching time is able to accommodate the slow changing background traffic. For more volatile traffic, the base-mesh mentioned above can be used to absorb the traffic volatility with additional wavelength paths mapped to different priority queues at EPS.

#226742 - $15.00 USD Received 11 Nov 2014; revised 10 Jan 2015; accepted 14 Jan 2015; published 5 Feb 2015

The OvS network is also tolerant to network failures. There are 2 types of common failures: node failure and link failure. When there is a node failure in the OvS network, for example: an OvS loses its power, it only affects the traffic that is to/from the local node/rack, but won't affect the connections among the rest of the network as PRF is passive and all optical signals will pass through transparently. Thus the failure is isolated to the local node until the failed OvS is replaced. To mitigate the impact from link (cable) failure, the PRF in OvS is designed to broadcast signals in both directions (west and east or north and south) within each dimension. Even if there is a cable cut or disconnection, the two OvSes directly linked to the failure point are still able to reach the rest of the network, although losing N fiber links over total N*(N + 1)/2 fiber links on the specific network ring. The centralized controller can quickly select a rerouting path within the mesh network and minimize the impact on live traffic.

2.2 Alternative implementations of wavelength selective switching

Section 2.1 describes a distributed optical switching fabric using the Nx1WSS as the key switching element. The wavelength selective function can be implemented by other optical configurations to achieve same functionalities.

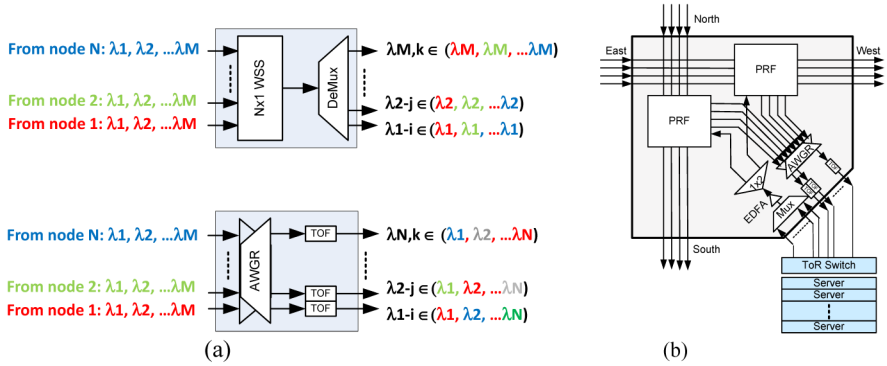


Fig. 4. AWGR based design: (a) AWGR based wavelength selective scheme vs. Nx1 WSS based design; (b) OvS implementation using AWGR based wavelength selective switching scheme.

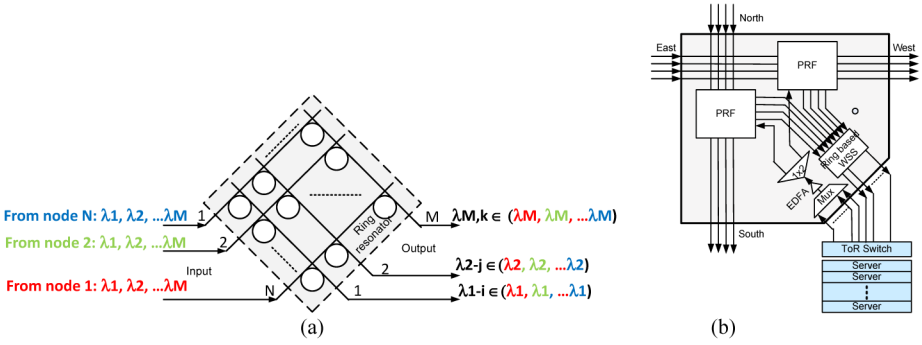


Fig. 5. Ring-resonator based design: (a) Wavelength selective switch design based on ring-resonators; (b) OvS implementation using ring resonator based wavelength selective switching scheme.

As an example, the wavelength selective function can be achieved through two steps: first a wavelength shuffling and then wavelength filtering. In Fig. 4(a), an AWGR is used to shuffle wavelength signals. Optical tunable filters are used to select the designated wavelengths. Then the combination of Nx1 WSS and 1xM DeMux are replaced by an NxN

#226742 - $15.00 USD Received 11 Nov 2014; revised 10 Jan 2015; accepted 14 Jan 2015; published 5 Feb 2015

AWGR plus an N-channel tunable optical filter (TOF) array. Here we assume N = M. As the DWDM signals coming from different nodes are shuffled through the AWGR, the TOF can perform the similar wavelength selection function as regular WSS though the wavelength channel plan is different than when using an Nx1 WSS (see Fig. 4(a)). Figure 4(b) shows the OvS implementation using AWGR and TOF. Though 512*512 AWGRs have been reported [15], the practical choice is to choose 48x48 AWGR and assume 48 DWDM transceivers per rack at 100GHz channel spacing to support 48 uplinks from the ToR. Thus, the whole system is also scale-able to support ~2k + racks. Comparing with the WSS based approach, the AWGR approach uses a low-cost PLC AWGR device for wavelength shuffling and has wider selections on TOF array in terms of cost, filtering performance and tuning speed to meet different system level requirements. In addition, the AWGR and TOFs have the potential to be monolithically integrated by silicon photonics to greatly simplify the packaging and further save costs.

Theoretically, the wavelength contention occurs in above two designs, either using Nx1 WSS based or NxN AWGR based, thus the WXC part in above OvS networks are both re-arrangable non-blocking networks. A fully functional N by N wavelength selective device (WSD) can be used to realize a strictly non-blocking network. One example of a fully functional WSD is a switching matrix structure based on ring-resonators as shown in Fig. 5(a) [16,17]. Under the simplest switching configuration, each ring-resonator filter is only used as a switch-on/switch-off filter for a specific wavelength. For example, ring-resonator filters along the line of output 1 are only responsible for picking lambda 1 from inputs 1 to N. Figure 5(b) shows the schematic drawing of OvS implementation using such kind WSD. It`s worthwhile to mention that a tunable transmitter is not needed to achieve strictly non-blocking functionality in this configuration.

3. OvS network performance study

3.1 2D OvS network throughput analysis

We conduct simulations in Matlab to understand the scalability of the circuit switching in the 2D OvS network. We first setup a network with 32x32 nodes. Each node represents a channel switching node. A centralized traffic generator creates connection requests as if a centralized SDN controller receives the demands and dispatches the requests. In the simulation, we assume the traffic requests are uniformly random: the node pair is selected with equal probability from the pool. We record the wavelength usage on each node. In the simulation, we assume the connection time is equal for each request. The resource is released immediately after the connection is finished. We used the greedy algorithm for wavelength routing assignment. The first shortest path is always used. We defined the total connections (TC) as total bidirectional links required for any to any connection among all users at a time. For example, if a network has 32x32 nodes and each node is requested to accommodate 24 users, the bi-directional TC is 32x32x24/2 = 12288. We calculated the total network blocking rate (TB) as the number of successfully assigned connection divided by TC. We repeated the above process in a Monte-Carlo simulation to transverse all the possible connection combinations and RWA assignments to get a statistical view. Figure 6(a) illustrates the modeled throughput of a 32x32 2-D OvS network with up to 48 wavelengths available as the ToR uplink interface to accommodate 24 users per node. When the system is fully loaded, the TB rate is around 8%. TB rate is negligible when the system load is below 90%. The number of transceivers required in a 2D OvS network to achieve low TB is in general the number the TC multiplied by 2. This is understandable because the cross dimensional traffic needs to occupy 2 wavelength channels and most connections are cross dimensional under the assumption of the simulation. An example of wavelength usage in a fully loaded OvS network is shown in Fig. 6(b): it's an example of network usage while network load is full: there are total 1024 nodes. The result indicates that the wavelength utilization rate is reasonably

#226742 - $15.00 USD Received 11 Nov 2014; revised 10 Jan 2015; accepted 14 Jan 2015; published 5 Feb 2015

uniform. There are no local hot spot because there is no centralized switching point that causes a traffic bottleneck, while we accommodated 12288 bidirectional connections.

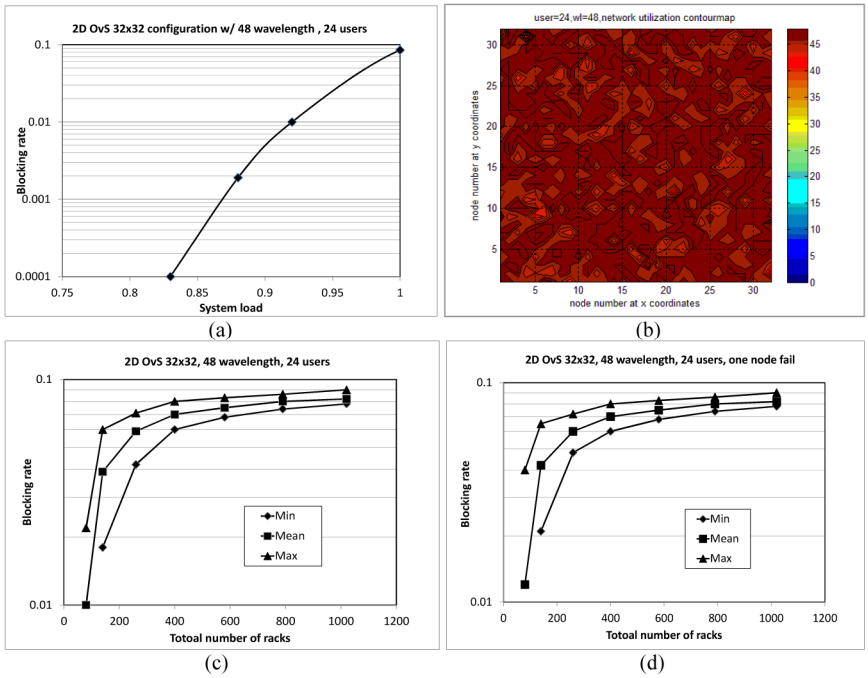


Fig. 6. Throughput study of a 32x32 OvS network: (a) Blocking rate under different traffic load (or called system load, network load) in terms of required circuit paths; (b) Wavelength usage, the color bar indicates the percent utilization; (c) Blocking rate and its max-min range vs. different rack numbers; (d) The same as (c) except with one failed node.

We also simulate the scalability for a 2D OvS network. We assume the network is full loaded in the following discussions: There are total 12288 random connection requests for a 32x32 node network and each node has 24 users. Fig. 6(c) illustrates the modeled results for the OvS networks with different network size (number of racks). We assume random traffic requests through the network. Again, we select the shortest path always based on a first come and first served strategy, this results in some blocking as the algorithm is overly greedy and not optimal. In the same figure, we show the upper and lower bounds (3σ) as well as the mean of the blocking probability. Blocking rate can be high or low depending on the current traffic pattern. The variation is large while network size is small. This is because there is not much path diversity to adapt to different traffic patterns. As network size increases, the blocking probability and its max-min range converge because there are a plenty of paths/wavelengths for controllers to select. We note that the result shows the total network blocking rate saturates when we scale up the network from 200 racks to over 1000 racks. The mean total network blocking rate is below 8% for a 1024 rack system or beyond, compared with 5% for a 200 rack system. This indicates that the network throughput performance is reasonably scalable for the OvS configured in a 2D environment for 1k + racks.

We also studied the network blocking probability when there are node failures. During the failure event, the network performance, outside of the failing node, should not be affected. In the simulation, we randomly selected a pair of nodes to be offline when the network size ranges from 64 nodes to 1024 nodes. We calculated the normalized blocking probability by excluding the traffic requests originated from the failure nodes in the calculation. Comparing with Fig. 6(c), we show the normalized blocking probability is nearly same, thus concluding

#226742 - $15.00 USD Received 11 Nov 2014; revised 10 Jan 2015; accepted 14 Jan 2015; published 5 Feb 2015

that the OvS network blocking rate is not significantly affected by few node failures at any network scale.

3.2 Wavelength switching algorithm adaptive to petabytes DCN

Clearly, the proposed scale of the fabric necessitates algorithms with scalable wavelength assignment. The wavelength assignment problem can be formulated as a bipartite edge-coloring problem on a multi-graph. For a 1-D OvS ring, we are able to show that a distributed wavelength assignment algorithm (DWA) [18] satisfies arbitrary port-to-port communication patterns with available wavelengths.

Using DWA, each node measures its own traffic demand to other nodes, and makes requests for wavelengths in other nodes to satisfy the demand. The controller of the OvS maintains the wavelength availability of the entire ring in a table locally, and updates the table via messages in an out-of-band control network. DWA is based on the parallel Hopcroft-Karp (PHK) algorithm [19], which constructs layered graphs in parallel by simultaneous Breadth-First Search (BFS) from the unmatched vertices. The following properties of our wavelength assignment problem lead to a simplified version of PHK with small communication overhead: (1) the underlying graph is regular and fully connected, thus we only have one layered graph at any time; (2) BFS can be done via table look-up locally, as any two nodes are reachable within one hop. On a 33-node ring, we demonstrate that DWA can achieve the same spectral efficiency as any centralized algorithm, while reducing computational latency by 80% [18].

3.3 Comparison

Folded Clos (fat free) architecture is considered as a cost effective and scalable solution in data centers for basic electrical packet switching. We compare the performance of our proposed architecture with folded Clos networks.

Table 1. Performance comparison vs. folded-Clos

Proposed architecture (2D), no over subscription,10Gbps per server					
Total racks	*Servers (48 /rack)*	*WSS size*	*Bisection BW (Tb/s)*	*Number of transceivers*	*Number of ribbon cables*
81	*3888*	*1x8*	*69.12*	*7776*	*162*
289	*13872*	*1x16*	*261.12*	*27744*	*578*
1089	*52272*	*1x32*	*1013.76*	*104544*	*2178*
Folded Clos (2-tier or 3-tier), no over subsription,10Gbps per server					
Total racks	*Servers*	*Switch Radix /Dimension*	*Max Bisection BW (Tb/s)*	*Number of transceivers*	*Number of ribbon cables*
96	*4608*	*96/2-tier*	*92.16*	*9216*	*4608*
648	*11664*	*36/3-tier*	*233.28*	*46656*	*23328*
2048	*65536*	*64/3-tier*	*1310.72*	*262144*	*131072*

In the comparison shown in Table 1, we consider 3 different cases with data center size from 4k servers to 50k servers. Both the folded-Clos and the proposed OvS network are scalable and able to provide enough BW to support a DCN. However, our proposal needs many less transceivers and cables, even compared to a 2-tier folded-Clos network. The advantages become larger when comparing to 3-tier folded-Clos.

4. Implementation

4.1 Hardware/software implementation of OvS box

We built the OvS prototype design shown in Fig. 7 as a test bed for network characterization. Figure 7(a) shows the schematics of the prototype OvS with two PRFs. We used 8x1 WSSs in the OvS prototype, where 4 ports were assigned for east–west traffic and 4 ports for north–south traffic. On the box interface, 4 MPO-12 connectors were used for 2D inter-OvS

#226742 - $15.00 USD Received 11 Nov 2014; revised 10 Jan 2015; accepted 14 Jan 2015; published 5 Feb 2015

connections. With N = 4 for each dimension, this prototype OvS test bed supports up to a 5-node network along each dimension for a 5x5 = 25 nodes (racks) network evaluation. Figure 7(b) shows the design of PRF for this implementation. To make the network more resilient to cable-cut and reduce the excess loss along the broadcasting path due to coupler and connector loss, we implemented bi-directional design by broadcasting the DWDM signal in both directions, where only two tap couplers are required in this prototype design. We simply chose the splitting ratio of 50:50 for both tap couplers. To compensate component losses, a single stage EDFA was used to boost the DWDM signal to about 8dBm per channel. With estimated losses of 7dB for the 1x4 splitter, 3.5dB for 1x2 splitter, 4dB for WSS, 2.5dB for DeMux and 1dB for connector loss, the receiving power is about −9dBm~-12dBm. As shown in Fig. 7(c), the proto-OvS box is equipped with 16 client ports each mapping to a particular wavelength from 190.5THz to 193.5 THz at 200GHz channel spacing. Commercial DWDM SFP + ER transceivers are used in our testing.

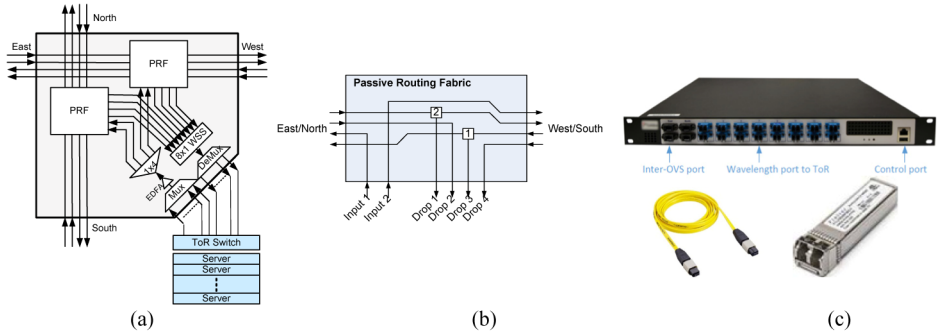


Fig. 7. Implementation of OvS prototype: (a) Schematics of prototype OvS using 8x1 WSS; (b) PRF design that can support bi-directional traffic; (c) OvS prototype photo, the ribbon cable and SFP + ER transceivers used.

On the front panel there is also a command line interface (CLI) port for local management and an Ethernet port for network management system (NMS) connection. Regarding the control plane, the OvS supports OpenFlow protocol [20] under SDN networking concept. Figure 8(a) illustrates the design of OvS node controller. The prototype OvS node controller uses an ARMv6k 700MHz CPU with 256MB memory, and runs multiple software functional components on an ARM version of Debian GNU/Linux. Each component runs at a distinctive software layer based on an abstracted information model and data model pertaining to each layer. The OVS software components are organized into four software abstraction layers: firmware layer, network element (NE) software layer, control layer, and interface layer. The firmware layer is the lowest layer of software in the OvS. It consists of device drivers which interact with hardware through low-level interfaces provided by each hardware components, for example serial interfaces. Each of the components in OvS has a device driver provided by the component manufacture and may speak only device specific languages. The NE software layer consists of software components that provide essential tools to support node fault management, accounting management, and security management of OvS as a telecom module, it also includes the resource management utility to translate the abstract configurations at control layer to physical configurations at each optical module. The control layer operates on top of the NE software layer. It is responsible for abstracting the physical resources on the NE software layer, e.g. modules and ports, into logical resources that are usable for path computation and virtualization. Thus DWA in Section 3.2 can be implemented in the control layer. At the interface layer, the OvS supports Openflow [20] v1.4 with the optical extension that allows an optical switch to be connected in a similar manner to that of traditional L2/L3 switches by a centralized controller such as OpenDayLight [21] as shown in

#226742 - $15.00 USD Received 11 Nov 2014; revised 10 Jan 2015; accepted 14 Jan 2015; published 5 Feb 2015

Fig. 8(b). The OvS also supports a client interface and web control interface (see Fig. 8(b)) thus the user can manage OvS locally or remotely.

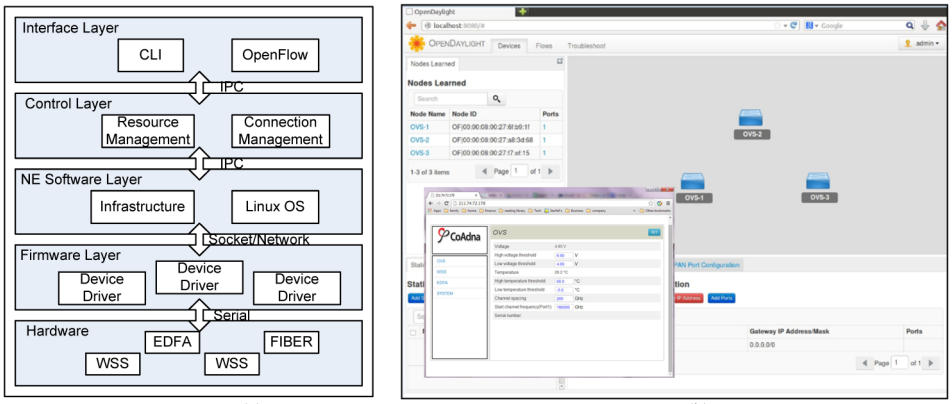


Fig. 8. Software implementation of OvS prototype: (a) OvS software components; (b) Snapshot of OpenDayLight GUI that communicates with OvS and OvS Web GUI.

4.2 Optical performance verification

As described in section 2, the OvS network is based on end-to-end optical links with no optical hopping, and aims to use low-end DWDM transceivers. To realize a 33-node 1D ring network, enabled by using a 32x1 WSS, it's critical to manage the link budget of the distributed broadcasting path along the 16 nodes in each direction. OSNR distribution and path loss uniformity are the performance indicators for such a network. Though the prototype OvS and test-bed (see Fig. 10) are mainly developed for networking study with only 5x5 network size, we did collect some statistical data on OSNR and path loss for larger network sizes. The path loss does not include Tx side component loss due to measurement simplicity.

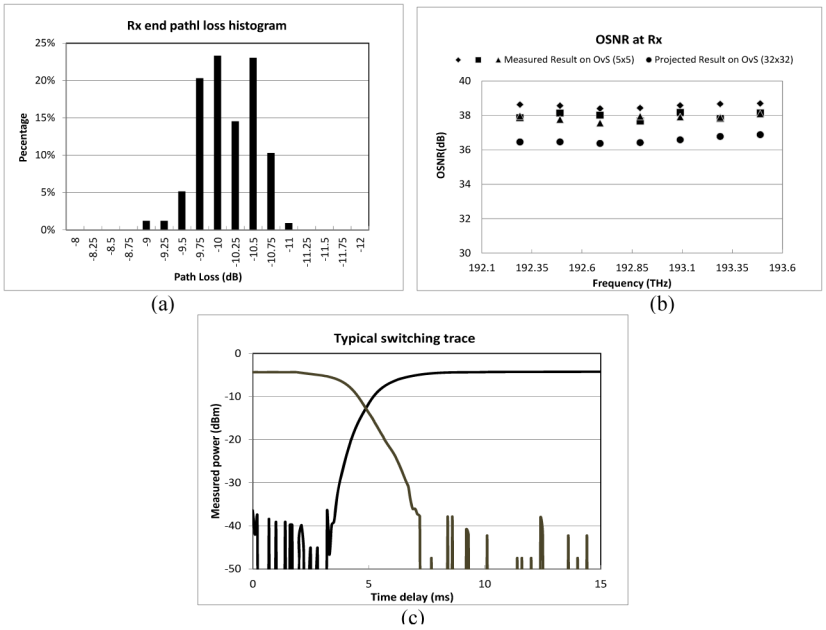


Fig. 9. Optical performance result: (a) Rx end path loss Histogram; (b) OSNR performance; (c) Typical optical switching latency.

#226742 - $15.00 USD Received 11 Nov 2014; revised 10 Jan 2015; accepted 14 Jan 2015; published 5 Feb 2015

On Fig. 9(a), we show the path loss histogram at the network endpoints at different wavelength output ports, through different paths and at different nodes. The distribution is within 2dB which meets the design targets. We also show the OSNR performance is generally better than 38dB across C-band in our prototype box (see Fig. 9(b)). Per our calculation, we expect OSNR to be no worse than 36dB for the OvS box to support 1089 racks. In the experiment we use DWDM SFP + ER from Innolight (TR-GXxxE-N00) with 8.5 + dB optical extinction ratio. Error free communications are achieved without FEC.

We also calibrated the optical switching speed of the prototype OvS as shown in Fig. 9(c) where an optical switching delay of around 7ms is observed. The monotonic (no overshoot or oscillations), predictable transition between states is a benefit of the digital LC switching technology of the WSS. The delay time is primarily limited by the wavelength switching element, but is on the order of the approximately 1ms response time of the control plane and is consistent with many DCN requirements while further delay time reductions are being developed.

4.3 Network implementation

Network performance characterization was conducted using the same test-bed outlined in section 4.1 using 5 OvS nodes along with 5 ToR EPSs from Pica8 (3295), each with 48 × 1 Gbps Ethernet (GbE) ports and 4 × 10 GbE ports (see Fig. 10). Each EPS was connected to one OvS via 3 × 10GbE links (also equipped with DWDM SFP + ER transceivers from Innolight). Because our EPS only has 4 10GbE ports (3 are fitted with 10 GbE optical transceivers), to achieve 1:1 oversubscription on the EPS we used 30 Dell PowerEdge R320 servers and 90 1GbE network interface cards (NICs) at the last hop. Every 10 servers were connected to 1 EPS. Each server had 3 NICs, which were all connected to the EPS via 1 Gbps links. Each server runs Debian 6.0 64-bit version (kernel 2.6.32-5).

First we evaluated the delay of the OvS network. In our test bed, the average port to port latency (RTT) within a single EPS is 95μs. The value is restricted by the performance of ToR and server. Comparatively, we achieve 78μs on a single OvS hop-link (connection within a degree) and achieve 114μs on a 2-OvS hop link (connection for 2D OvS network). The additional RTT overhead on a 2-OvS hop link was mainly from the 2nd ToR because the delay of the Pica8 ToR switch is quantified at the 20μs level. If similar EPS are used to construct a folded-Clos network, the RTT delay is expected to be at the 200μs level while a 3 layer fat-tree network is required to support the similar size of data center and there will be 5 hopping of EPSs on an end to end link.

Next, we tested the scaling performance of OvS network with a real application. We deployed Spark 1.0 with Oracle JDK 1.7.0_25 because Spark does cloud computation in memory and is much faster than Hadoop. Thus Spark can better exploit the high bandwidth feature from optics. We set up 4 scenarios:

1. Single ToR (1-D prototype with 1 nodes): 30 workers are connected to a single Top-of-Rack (ToR) EPS (Pronto-3295).
2. 1-D prototype with 2 ToRs (nodes): Each node is connected to 15 workers.
3. 1-D prototype with 3 ToRs (nodes): Each node is connected to 10 workers.
4. 2-tier tree with 3 ToRs: We use 4 EPSs to form a 2-tier tree with 3 ToRs and 1 aggregate EPS (Pronto-3780). Each ToR is connected to 10 workers.

#226742 - $15.00 USD Received 11 Nov 2014; revised 10 Jan 2015; accepted 14 Jan 2015; published 5 Feb 2015

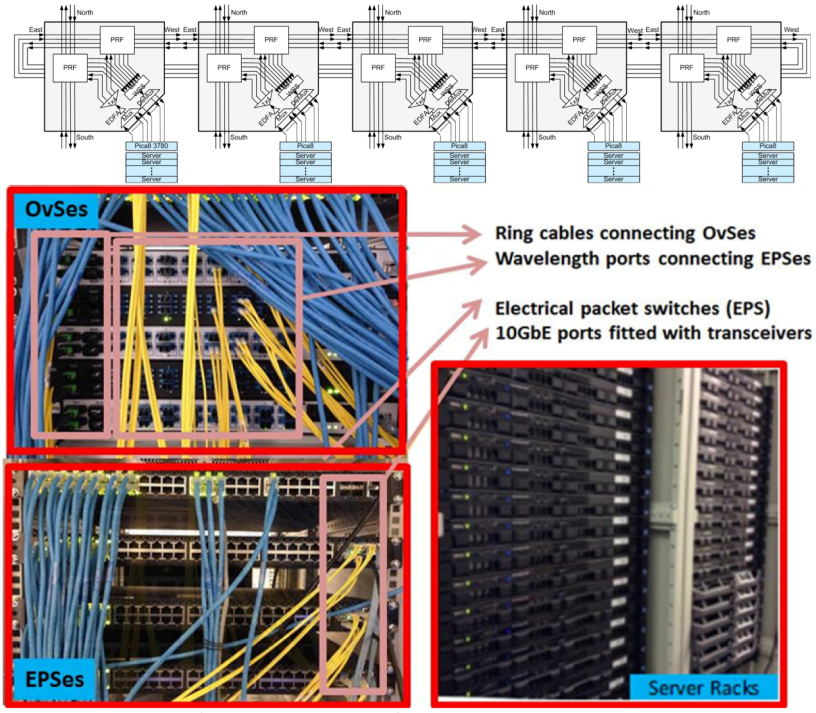


Fig. 10. Network test-bed of OvS platform with 5 OvSes inter-connected through ribbon fiber cable.

We use 3 benchmark applications on Spark:

1. WikipediaPageRank: a PageRank algorithm instance using Wikipedia entries as input. We process 13G and 26G Freebase-wiki-articles data sets [22] separately.
2. Spark K-means: K-Means clustering is a popular clustering algorithm that can be used to partition a data set into K clusters. The input data set is Wikipedia Page Traffic Statistics [23].
3. WordCount: It reads text files and counts words in them. We use a 20G and a 40G data set, which contain 3,560,179,980 words and 7,153,321,364 words respectively.

We ran these applications and increased the number of racks (nodes) in these experiments to compare the performance differences between the OvS networked system, a single ToR switch and a 2 tier fat-tree network in Fig. 11. These applications feature a large amount of small flows; and compared to high throughput applications, they are more sensitive to packet level metrics. We show that with the OvS the end-to-end latency (RTT) and packet loss rate do not degrade when more nodes (racks) are added, as both metrics are similar with 2 and 3 nodes. The completion times of different applications also do not increase significantly with more nodes. However, for 2-layer fat tree implementation, due to the extra hop on the aggregate switch, all metrics suffer: a 28%increase in RTT, 0.2% increase in packet loss rate, and an average 12.2% increase in completion time for 3 applications. The result shows our proposed architecture performs as if it's a single switch and outperforms fat tree network. It should be noted that although we compare to the 2 tier fat tree architecture it usually requires 3 tier fat tree architecture to scale up to support >10k ports.

Next, we demonstrate the adaptive link capacity function by leveraging SDN. We setup 2 servers under node B and Node C. Each server under node B simultaneously sends 30 TCP traffic to the corresponding server under node C to emulate the aggregation type traffic. Each

#226742 - $15.00 USD Received 11 Nov 2014; revised 10 Jan 2015; accepted 14 Jan 2015; published 5 Feb 2015

server is sending traffic at its full NIC speed. We create a lane aggregation group (LAG) with 2 wavelength links. At the beginning, the 2nd wavelength is turned off. Thus the link between node B and node C is congested. We show the aggregated TCP throughput from 30 TCP connections for both servers in Fig. 12(a). The net aggregated throughput fluctuates a lot as the TCP traffics compete for the link resource. During the congestion, the RTT jitter also increases and completion of each TCP flow varies due to TCP windowing thus jeopardizing the performance of applications. At the timestamp around 5s a new wavelength is added into the LAG while the centralized network controller makes a dynamic adjustment on the network resource. The logical link capacity is doubled to match the bisectional bandwidth at the server`s NIC. At that point we note that the TCP flows are transmitted at full speed with no throughput fluctuation.

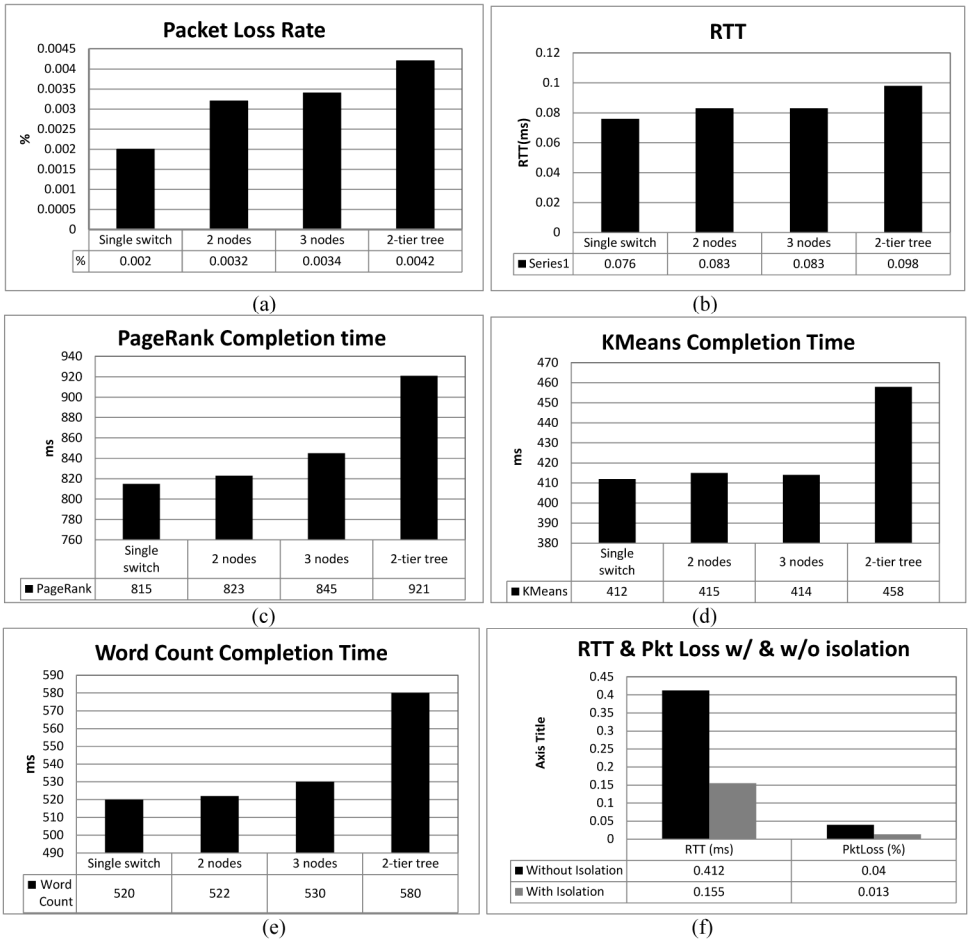
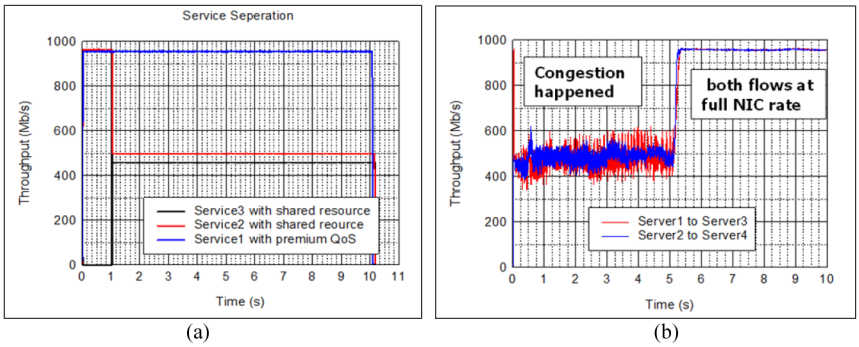


Fig. 11. Evaluation of OvS by applications: (a) Packet Loss Comparison; (b) RTT comparison; (c) PageRank completion time; (d) Kmeans Completion time; (e) Word count completion time; (f) Performance comparison with & without service separation. The 2 and 3 nodes are using OvS, the 2 tier tree is a typical 2 tier fat-tree network.

The inter-tenant performance isolation [24] is one of the major metrics of cloud data centers. We show that our prototype can provide cross-EPS tenant performance isolation with ease at the physical layer using different wavelengths. We setup 2 servers at node B and 3 servers at node C. We assumed server 1 under node C is a service tenant with high quality of service (QoS) and is sending TCP traffic to server 1 under node B, while servers 2&3 are low

#226742 - $15.00 USD Received 11 Nov 2014; revised 10 Jan 2015; accepted 14 Jan 2015; published 5 Feb 2015

QoS tenants and both servers are sending TCP traffic to server 2 under node B. We configured a dedicated wavelength link for server 1 and a shared wavelength link for servers 2&3. We show the throughput result for each service in Fig. 12. At the beginning, both server1 and server 2 are able to send at full NIC speed. Once server 3 is added to send traffic we have to rate-limit the traffic from servers 2 and 3 to assure no significant packet loss during the transmission while traffic from server 1 is totally unaffected during the process. Even when we setup the experiment with no throughput congestion, we show both RTT and the packet loss rate are significantly reduced by 62.38% and 67.5% compared to without isolation in Fig. 11(f). Therefore, using different wavelengths, the OvS can easily and dynamically enforce strong performance guarantee for different tenants.



(a) (b)

Fig. 12. Throughput Results: (a) Service separation; (b) Adaptive bandwidth assignment.

5. Discussion

5.1 Comparison with related works

Recognizing the need for low-cost reconfigurable bandwidth throughout the network, researchers have recently proposed several designs incorporating OCS in DCN. C-Through [6] and Helios [25] have employed centralized optical channel switches to achieve a flattened network topology. Configuring a circuit in these two proposals is simpler than establishing a non-blocking path in a multi-tier Fat tree. However, both of these optical architectures suffer from constrained switching capacity and require multi-hop circuits to scale the system capacity. For example, OSA [7] enables arbitrary communication at any time by using hop-by-hop stitching of multiple optical circuits to provide all-to-all connectivity. OSA also introduces DWDM technology into the design. However, since each additional hop adds loss, there is a question of receiver sensitivity requirement due to the lack of optical amplification for these approaches. Mordia [5] introduces arbitrary port-to-port, non-blocking connectivity among all the ports by using TDM. However, the total system capacity is still restricted by single ring fiber which conveys 88 wavelengths at most. Thus the scalability in terms of port count is fundamentally restricted by the available non-interfering wavelengths. Mordia discusses scaling using multiple stacked ring fibers, e.g. with 8 rings it can scale to 704 ports. However, the stacked ring solution is blocking, and not all port-to-port mappings are possible. Recently, Quartz [26] emphasizes low latency, and is designed as part of DCN to provide low latency paths. Quartz is structured as a ring as well. However, the wavelength assignment in Quartz is fixed, whereas OvS can service variable bandwidth demands between any 2 ports.

In comparison to the other optical solutions, OvS focuses on supporting much larger scale communications in DCN, and emphasizes cost-effective scalability on both data and control planes. In Table 2, we compare the features among OvS and other pioneering OCS solutions. Recent solutions [5,7] adopt DWDM technology to increase link BW and achieve optical multicast by using EDFA to compensate the optical splitter loss. However, the commercial feasibility of DWDM optical solutions in DC relies on accommodating low end DWDM

#226742 - $15.00 USD Received 11 Nov 2014; revised 10 Jan 2015; accepted 14 Jan 2015; published 5 Feb 2015

transceivers while minimizing amplification. OvS is based on single hop optical connections which mean there is only one stage of EDFA from end to end on any link. As shown by the red trace in Fig. 1(b), node 1 communicates to all other nodes using only one stage of pre-amplification on the transmitter side and none on the receiver side. Because only single optical hopping is required in OvS system, it can maintain high OSNRs for all the DWDM channels through the whole system, which is the key factor in optical link-budget which determines the optical DWDM transceiver specifications.

Table 2. Comparison of different optical DCN architectures using OCS

	C- Through (2010)	*OSA (2012)*	*Mordia (2013)*	*Quatz (2014)*	*OvS*
Optical technology	*OCS*	*DWDM and OCS*	*DWDM and OCS*	*Static*	*DWDM and OCS*
Port Count Number	*single tier:320*	*single-tier:320 Multi-tiers:~2k*	*single ring:96 stacked rings: 1k*	*1K*	*1D: 1k~3k 2D: 100k*
Port to Port Connectivity	*unicast*	*unicast*	*unicast, multicast broadcast*	*unicast*	*unicast, multicast broadcast*
Design Complexity/Cost	*moderate*	*moderate*	*expensive*	*low*	*low*
Cabling	*complex*	*complex*	*complex*	*complex*	*simple*

5.2 Application driven networking

Overall, the OvS is designed to create a flattened, fully programmable interconnection layer among all the server racks, providing interconnection-on-demand or topology-on-demand capability for the upper application layer as part of the software defined infrastructure (SDI) solution.

To drive such a network, traffic demand estimation was originally proposed. For example, OpenFlow can be used to provide a snapshot of the overall traffic demand. Recently, people have started to forecast traffic demand from applications even before the traffic enters the network [27]. Application-aware network scheduling in data-intensive clusters has been proposed [28] for big-data analytic frameworks. As communication in data-parallel applications often involves a collection of parallel flows, new network scheduling was proposed to enable these data-intensive frameworks, such as coflow abstraction and Varys [29].

5.3 Device limitation and future work

We have emphasized throughout that one of the key design goals of the OvS platform is a cost competitive solution to target real future deployment: our unique one-hop-to-anywhere design allows a relatively low cost DWDM transceivers. Our digital Liquid Crystal (LC) based optical switching technology used in the WSS has proven to be cost-effective and reliable technology that meets this type application while supporting large port count (16 to 32 ports) and consuming extremely low power (2.5 W for the WSS used in the prototype).

In addition, other technologies continue to be evaluated that may further improve the integration level and provide a future cost-down path. To support more DWDM channels, the Planar Lightwave Circuits (PLC) AWG is a promising choice for Mux/DeMux. PLC may also be used to make the PRF as it has the potential to shrink the device size and cost, making the assembly much easier. During the discussion on alternate implementations we highlighted the PLC or silicon-photonics based ring resonator as another future candidate for TOF or WSS.

As measured in section 4.3, the switching speed of the current WSS is less than 10ms. Further reductions in switching delay will be beneficial for latency-sensitive flows and can help to avoid costly buffering of packets when the line rate goes up to 40/100/400Gbps. We note that the LC optical switching can be cost-effectively reduced to sub-ms [30] providing a

#226742 - $15.00 USD Received 11 Nov 2014; revised 10 Jan 2015; accepted 14 Jan 2015; published 5 Feb 2015

path for future improvement. The TOF mentioned in section 2.2 offers another option to further reduce the optical switching latency.

The OvS platform leverages the DWDM transmission technology. Since a number of DWDM transceivers need to be populated between each EPS and OvS add-on box, the cost of DWDM transceivers becomes very critical and could be a determining factor for the adoption of the OvS platform. Currently, the cost of a 10Gbps DWDM transceiver, i.e. DWDM SFP + ER is about 6 times of that of an SFP + SR. Clearly, EML based SFP + ER is overkill for 2km intra-data center networking. The cost of DWDM for intra-data center networking could come down quickly with a relaxed dispersion requirement. We are seeing many active development efforts on low cost DWDM transceivers that are targeting the DWDM applications in access (i.e. TWDM-PON) and wireless (i.e. mobile front-haul). The same technologies can be leveraged for OvS application. Recently, we see newly developed laser chips that can bring down the 10Gbps DWDM transceiver cost to the level of 2~3 times of an SFP + SR, which could trigger the adoption of DWDM technology widely into DCNs. Also, these emerging technologies support future modulation baud rate at 25Gbaudps to adapt to single lane 25GbE [2] for DCNs.

6. Summary

In summary, we proposed, built, and characterized a distributed optical switching fabric for DCN intra-data interconnection that can efficiently and cost-effectively support 1k + racks and is topology adaptive for dynamic and diverse traffic patterns. The proposed architecture provides a converged EPS/OCS to enable efficient and flexible resource management through SDN. Compared with multi-tiered optical solutions, our proposed solution significantly reduces the transceiver numbers needed and greatly reduces and simplifies the cabling thereby enabling a cost-effective practical application of optical networking to DCNs. We have prototyped and characterized the design and introduced the basic building element: OvS. We have carried out experiments of the OvS network based on the prototypes and demonstrated the potential for supporting scaling out applications with diverse traffic patterns by running realistic big data applications such as Spark [31].

Acknowledgments

We are thankful to Dr. Hudson Washburn for his valuable suggestions.

#226742 - $15.00 USD Received 11 Nov 2014; revised 10 Jan 2015; accepted 14 Jan 2015; published 5 Feb 2015