Experimental Demonstration of Converged Inter/Intra Data Center Network Architecture

Payman Samadi*, Junjie Xu, Ke Wen, Hang Guan, Zhuo Li, Keren Bergman
Lightwave Research Laboratory, Department of Electrical Engineering, Columbia University, New York, NY 10027
**e-mail: ps2805@columbia.edu*

ABSTRACT

We present a novel converged inter/intra data center network architecture to enable on-demand rack-to-rack connectivity across data centers. The hardware architecture includes a bidirectional software-defined optical gateway that aggregates racks or pods on a conventional data center data plane and provides both east-west and north-south connectivity. The software architecture consist of two Software-Defined Networking (SDN) agents over the data center and transport network control planes that manages connection requests and finds the optimal routing and wavelength configuration from the available WDM channels. We present bulk data transfer and Virtual Machine (VM) migration on a testbed of three data centers.

Keywords: data center networks, software-defined networking, optical circuit switching, virtual machine migration.

1. INTRODUCTION

The increasing growth in big data generation and cloud-based services has escalated the deployment of mid-sized data centers. These data centers actively communicate in the optical transport network for services such as backup, Virtual Machine (VM) migration, video streaming and fault/disaster recovery. Companies that employ these data centers require operation reliability and fast scalability. Furthermore, studies show that optical transport networks are generally under-utilized and over-provisioned [1, 2]. One solution would be a converged inter/intra data center architecture using Software-defined Networking (SDN) and an application-driven approach to address these key requirements and potentially improve the link utilization in optical transport networks.

Optical links at data rates above 10 Gbps are now widely used in data centers. Researchers have proposed a hybrid architecture to offload larger traffic flows from the over-subscribed electronic packet switching to an optical circuit switching network [3, 4]. This approach requires complex methods to identify proper elephant flows that result in inefficient utilization of the optical links. All-optical data center networks [5] are also proposed by placing Top-of-Rack (ToR) switches in an optical ring and perform Wavelength Division Multiplexing (WDM) switching. Scaling such network is a challange since the network latency becomes non-deterministic and wavelengths are used inefficiently.

We propose a converged inter/intra data center architecture to 1) increase application reliability by operation distribution over multiple data centers, 2) scale out data center in distance to surpass scalability limits, 3) improve utilization of optical transport network links by an application driven-approach and finer granularity in managing the wavelength usage, 4) improve utilization of optical links in hybrid data center networks by routing the north-south traffic. We introduced the concept in [6]. In this work, we present the complete end-to-end system and experimentally evaluate it on a prototype testbed. Our design consist of a hardware architecture that enables on-demand cross data center rack-to-rack connectivity over the optical transport network. The architecture only requires commodity optical components. It also includes a SDN control plane that receives the connection requests and finds the optimal routing and wavelength configuration by an Integer Linear Programming (ILP). We experimentally evaluated the performance of the architecture on our testbed and achieved 142 ms control plane latency to create rack-to-rack connection between two data centers 25 km apart. We also demonstrate VM migration and bulk data transfer between three data centers as end-to-end applications.

2. ARCHITECTURE

Figure 1(a) demonstrates the converged inter/intra data center architecture over a mesh optical transport network. The hardware architecture consist of a software-defined bidirectional optical gateway [7] demonstrated in Figure 1(b) that provides both inter and intra data center connectivity. The optical gateway includes an Optical Space Switch (OSS) that aggregates ToR or aggregation switches of the data center. The OSS provides east-west point-to-point connectivity inside the data center. It also provides north-south connections through the add/drop WDM multiplexers. The optical gateway includes two Wavelength Selective Switch (WSS) that routes WDM channels to the adjacent gateways in the optical transport network. WDM signals are amplified after aggregation and power balanced by the WSS before entering the transport network.

978-1-4673-7879-6/15/$31.00 ©2015 IEEE

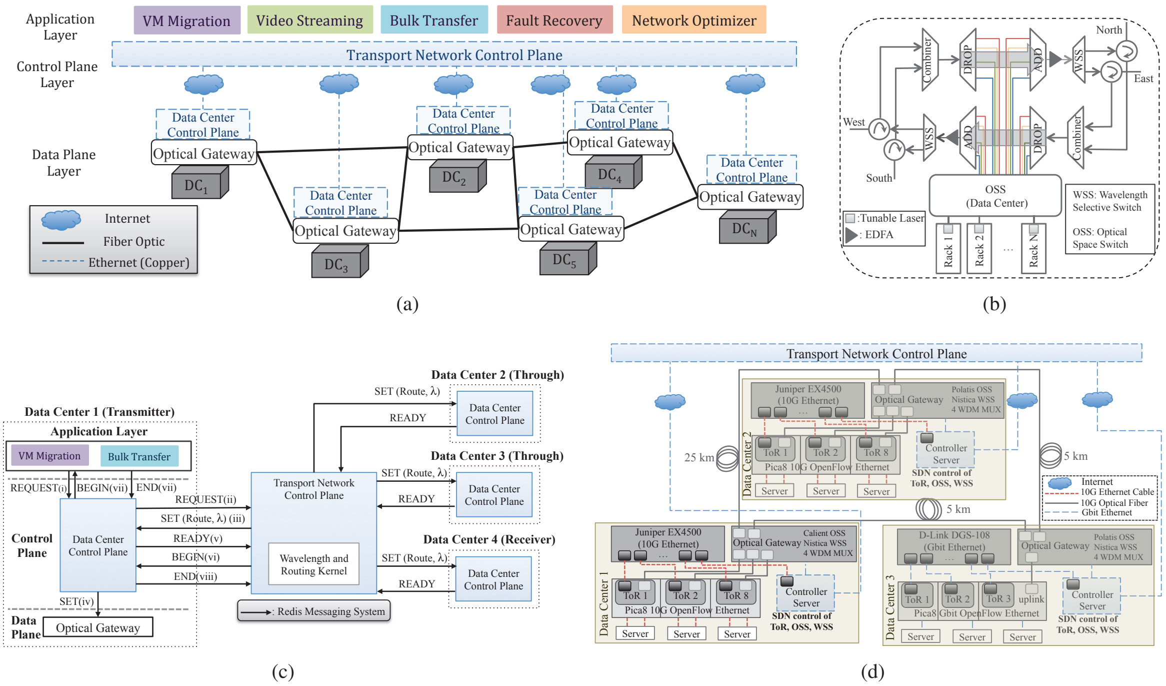


Figure 1. (a) Converged inter/intra data center network, (b) Optical gateway architecture, (c) Software architecture workflow, (d) Testbed setup.

The software architecture consist of two SDN agents that integrate with the data center and the transport network control planes, respectively. Figure 1(c) demonstrates the workflow to create a connection. The application sends the connection REQUEST to the data center control plane (i) consisting of the transmitter and the receiver IP addresses and the service type. The request is forwarded (ii) to the transport network control plane that first determines the respective data centers using IP addresses and then computes the optimal routing and wavelength configuration. Next, the configuration is sent with a SET command (iii) to data centers involved in the connection. Once each data center's optical gateway (iv) finishes the required configuration, a READY (v) command is sent to the transport network control plane, which will then notify the transmitter data center to start (vi) the service. At last, the data center control plane sends a BEGIN (vii) command to the application.

Routing and wavelength assignment in the transport network is implemented with a *first fit* algorithm. The algorithm first finds the shortest path between the requested source and destination, and then searches for a wavelength that is available on all edges of the path. The search considers a lower-indexed wavelength prior to a higher-indexed one, and the first available wavelength will be selected. If no available wavelength is found, the algorithm searches for the next shortest path and repeats the same wavelength search procedure as described above. As a result, the algorithm packs all the wavelengths in use towards the lower wavelength space, so that the remaining wavelength space will have a higher probability of being available when future requests arrive. Furthermore, since shorter paths are preferred, power consumption and wavelength occupation of the lightpaths are reduced.

3. IMPLEMENTATION

We build a three-node data center testbed using commodity optical and electronic components to evaluate our design (Figure 1(d)). Each data center is equipped with 3 ToRs connected to one server in a hybrid architecture. Each server is equipped with a 10G Network Controller, an Intel Xeon 6-core processor and 24 GB of RAM. The electronic packet switching is provided by a L2/L3 Ethernet switch (Juniper EX-4500, D-Link DGS-108) and the optical circuit switching by an OSS (Calient S320, Polatis 10 and 16). The OSS is also part of the optical gateway (Nistica WSS and 1:8 WDM Mux/Demux) that provides north-south connectivity. Each data center is connected to the other two with either 25 or 5 km of SMF. The optical transceivers are DWDM 10G SFP+ ZR modules, providing 24 dB optical link budget. Each data center is equipped with a SDN controller server and there is a separate server for the optical transport network controller. We used Redis messaging system [8] as the north-bound API for the communication between the

Table 1. Delay of control plane components on the testbed.

Component	**Delay (ms)**
Northbound API (Redis)	(i) – (vii) in Figure 1(c) : 7 × 0.650 = 4.55
Optical Gateway	125
Routing and Wavelength Assignment Algorithm	10
Data center controller code	1.7
Transport network controller code	1
Total	**142.25**

application and the control plane layers in each data center and also for the communication between the data center and the transport network controllers. The southbound APIs of the optical gateway are TL1 commands for the Polatis and Calient optical switches and in-house developed C code for the Nistica WSS.

4. EVALUATIONS

We evaluated the performance of control plane implementation by measuring the execution delay of different control plane components including the the wavelength and routing assignment algorithm on the testbed. We also measured the link speeds and demonstrated bulk data transfer and VM migration as end-to-end applications.

4.1 Testbed Results

In the first set of experiments, we measured the latency of control plane components. Table 1 shows the results measured on a server equipped with two Intel Xeon E5-2403 4-core processors and 24 GB of RAM. The total delay consist of the northbound API, the controller code for the data center and transport network control plane, the optical gateway reconfiguration time and the wavelength and routing assignment algorithm. For Redis, we measured the average latency for transmitting 100 messages of 20 bytes between the controllers on the campus internet. The optical gateway latency consist of the OSS (25 ms) and WSS (125 ms) switching time that are configured in parallel. The wavelength and routing assignment algorithm had 5 – 10 ms latency when the shortest path was free on the network.

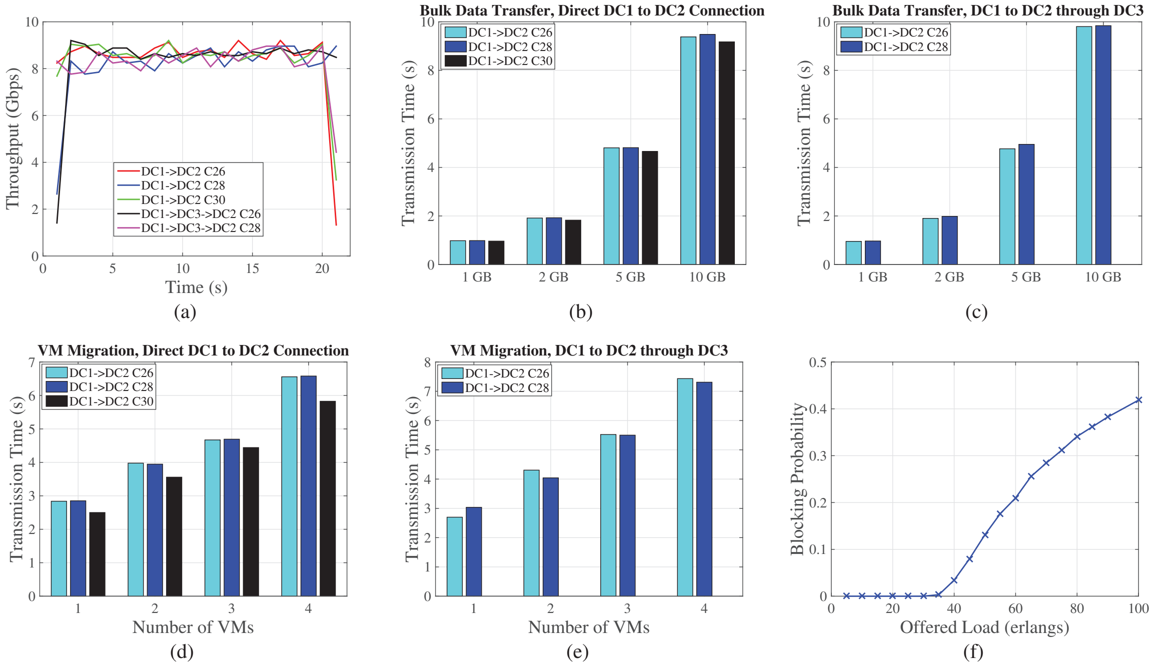
For the end-to-end evaluations, we considered two scenarios of i) direct DC1 to DC2 connection and ii) indirect connection through DC3 in the testbed (Figure 1(d)). We measured the throughput of the established links between servers on the racks of DC1 and DC2 using iperf that is a common network performance measurement tool. The setup supports C26, C28 and C30 of 50 GHz DWDM ITU Grid with 10 Gbps SFP+ transceivers. In scenario (i), 3 connections from racks 1 (C26), 2 (C28), and 3 (C30) of DC1 are made to racks 1, 2, and 3 of DC2 and in scenario (ii) racks 1 (C26) and 2 (C28) of DC1 and DC2 are connected passing through DC3 optical gateway. Figure 2(a) demonstrates the results with average 8.85 Gbps throughput for all five connections.

Next, we demonstrate bulk data transfer between the servers of data centers 1 and 2 as an end-to-end application. The data sizes are 1, 2, 5, and 10 GB. We measured the transmission time that is the time to deliver the data excluding all connection overheads. Figures 2(b) and 2(c) shows the results for both scenarios (direct and indirect connections). Results confirm that the architecture had steady performance on all the connections for different data sizes and saturated the 10G data rate of the links. We also performed live VM migration between the servers of DC1 and DC2. The VMs configuration is 2 CPUs, 2 GB of RAM running Scientific Linux 7. We implemented live migrations by libvirt virtualization APIs. Figures 2(d) and 2(e) show transmission times for both direct and indirect scenarios with 1 to 4 VMs on each server. The system performed steady independent to the number of connections and VMs.

4.2 Simulation Results

We use simulation to numerically evaluate the performance of the routing and wavelength assignment algorithm at scale. The setup consists of 16 data centers each supporting n ($n = 1000$) racks, in a 4×4 meshed [topology] optical transport network with 50 available wavelengths on each link. Each rack requests cross data center connections according to a Poisson process with arrival rate p. The connections choose destinations randomly with uniform distribution among remote racks, and have an exponentially distributed holding time (i.e. application service time) of mean h. The simulation hence emulates a birth-death process on the entire network with an *offered load* of $E = nph$ (in erlangs), which is a product of the per-datacenter arrival rate and the holding time.

Figure 2(f) shows the blocking probability (y-axis) of the network under different offered loads (x-axis). The simulation changes the offered load by varying the per-rack request rate p against a fixed mean holding time of 10 min. As the result shows, when the load is below 30 erlangs, i.e. requests per data center per minute is less than 3, almost no request is blocked.



(a) (b) (c)

(d) (e) (f)

Figure 2. Experimental results: (a) Throughput of the links between DC1 and DC2 in scenarion (i) and (ii). (b, c) Transmission time of bulk data transfer between DC1 and DC2 for both scenarios and data sizes of 1, 2, 5, and 10 GB, (d, e) Transmission time of VM migration between DC1 and DC2 for both scenarios and 1–4 VMs on each server, Simulation result: (f) Blocking probability of the wavelength and routing assignment algorithm.

5. CONCLUSIONS

We presented a converged inter/intra data center architecture to enable on-demand cross data center rack-to-rack connectivity. It consist of a hardware system that is a bidirectional software-defined optical gateway and a software system that are SDN agents. We built a testbed, experimentally evaluated the architecture and also demonstrated end-to-end applications of bulk data transfer and VM migration. The main advantages of this architecture are (i) more efficient utilization of the optical transport network links by providing on-demand connections with wavelength granularity, (ii) increasing application reliability by enabling operation distribution over multiple data centers, and (iii) providing data center scalability in distance.

ACKNOWLEDGMENTS

This research was supported in part by CIAN NSF ERC under grant EEC-0812072 and NSF CNS-1423105. We would also like to thank Juniper Networks, Calient Technologies and Polatis for their generous donations to our testbed.

REFERENCES

[1] S. Jain, *et al.*, "B4: experience with a globally-deployed software defined wan", ACM SIGCOMM 2013.

[2] C. Hong, *et al.*, "Achieving high utilization with software-driven WAN", ACM SIGCOMM 2013.

[3] N. Farrington, *et al.*, "Helios: A hybrid electrical/optical switch architecture for modular data centers", ACM SIGCOMM 2010.

[4] G. Wang, *et al.*, "c-Through: Part-time optics in data centers", ACM SIGCOMM 2010.

[5] G. Porter, *et al.*, "Integrating microsecond circuit switching into the data center", ACM SIGCOMM 2013.

[6] P. Samadi, *et al.*, "Virtual Machine Migration over Optical Circuit Switching Network in a Converged Inter/Intra Data Center Architecture", Optical Fiber Communication (OFC), 2015.

[7] P. Samadi, *et al.*, "A Software-Defined Optical Gateway for Converged Inter/Intra Data Center Networks", Optical Interconnect (OI), 2015.

[8] S. Sanfilippo, P. Noordhuis, "Redis", http://redis.io.