Towards all optical DCI networks

Ginni Khanna[1,2], **Shengxiang Zhu**[1,3], **Mark Filer**[4], **Christos Gkantsidis**[1], **Francesca Parmigiani**[1], **Thomas Karagiannis**[1]

[1]*Microsoft Research Cambridge, UK.* [2]*TU Munich, Germany.* [3]*University of Arizona, USA.* [4]*Microsoft Redmond, USA*

thomkar@microsoft.com

Abstract: We propose and experimentally demonstrate an all-optical architecture for data center interconnect networks with reconfiguration times of a few seconds. Filtering and amplification transient effects have minimal impact on BER performance.

OCIS codes: (060.0060) Fiber optics and optical communications; (220.4830) Systems design

1. Introduction

As the demand for cloud services keeps increasing following exponential growth rates, cloud providers massively expand their cloud footprints around the globe. The epicenter of this cloud infrastructure expansion is the regional Data-Center Interconnect (DCI) network which provides connectivity across tens of data-centers located within metropolitan distances with the aim of providing the illusion of a massive distributed data-center.

Fig. 1a shows a typical example of a DCI connectivity model. Regional Network Gateways (RNGs) interconnect all Data Centers (DCs) in a region and perform layer-3 switching for all DC-to-DC regional traffic, and RNG-to-DC fiber distances are limited to 60 km [1]. This is a nice hierarchical model that allows DCs to be spread across a contained metropolitan area, and facilitates fast scalability and growth; connecting a new DC in the region is straightforward without affecting other DCs. To achieve this however, RNGs host a series of high-end electrical chassis switches that provide full connectivity across DC endpoints, with significant requirements in terms of the number of switch ports, space and power. For example, for a region of 10 DCs with 16 fiber-pairs each and 100 GHz 40-channel DWDM system per fiber, RNGs need to host a total of 12,800 transceivers! This number is expected to increase by more than a third as DCI systems move to the 400ZR ecosystem (75 GHz grid) [2]. Indeed, transceivers represent one of the major cost components of DCI networks.

In this paper, we examine the potential for all-optical DCI networks, thus removing the need for transceivers in the RNGs. Replacing electrical switches at the RNGs with optical ones reduces the overall DCI cost as well as power and space requirements. To achieve this, dynamic end-to-end optical paths are setup between DC endpoints depending on traffic demands through a centralized software controller. Fig. 1b presents the proposed architecture which is amenable to the high port count required at the RNGs.

Emulating a 4-DC region with varying DC-to-DC distances up to 120 km [1], we demonstrate that fast optical reconfiguration of DC-to-DC paths is feasible with minimal impact on the signal performance and reconfiguration overhead. When 128 200G 16 QAM channels in the C-band are used, our results show that i) even when reconfiguring all but one channels on a fiber, the surviving channel bit error ratio (BER) measurements are minimally impacted; and ii) individual channels take less than 200 msec to recover the signal after being switched. These results are distinctly linked to DCI networks. The number of amplifiers affected by reconfiguration is limited (to 2), so that their power excursions are a non-issue, in contrast to observations in wide area networks [3]. Similarly, the extra optical filtering penalty introduced [4, 5] is minimal even for tight examined grids of 37.5 GHz.

2. Network architecture for optical DCI

Following today's practices [1], we adopt the notion of a deployment stamp that is identical at all fiber end-points. This simplifies deployment and management, and ensures that the optical line system for the optical RNG is exactly the same as for the electrical one, minimizing transition overhead to the new architecture. All DCs connect

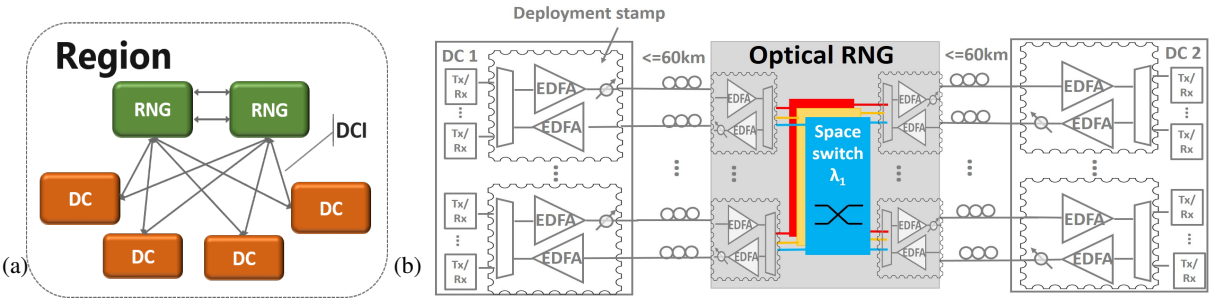

Fig. 1: (a) Current electrical RNG DCI model; (b) Line system and architecture for optical RNG.

as before to the now "optical" RNG (Fig. 1b) and re-routing happens in the optical domain only. This is justified by the realization that the DC-to-DC traffic patterns are quite stable for hours making the DCI particularly attractive for dynamic reconfigurable optical paths; existing optical switches take a few tens of msec to reconfigure. The only effective difference compared to the electrical architecture is the lack of signal regeneration at the RNG.

This has an important performance implication –optical components, such as amplifiers and mux/demux filters are doubled in-between transceivers to guarantee wavelength granularity, from two (electrical) to four (optical). Similarly, transmission distances also double to 120 km. Yet, only two amplifiers exist after the reconfiguration point, and only these would experience any power excursions. We show that these effects do not impact performance, thus making the optical RNG architecture feasible in terms of signal properties.

Optical attenuators (OAs) after the amplifiers on the transmission side ensure equal power levels for signals reaching the RNG despite varying-length fiber spans across DCs. Importantly, this design does not require any power adaptation at the amplifiers or OAs during or after reconfiguration. The only component that our software controller needs to adapt to provide reconfigurability are the mappings of input-to-output ports of optical space switches that provide the re-routing functionality at the optical RNG.

Following our previous region example for 400ZR (64 channels), an all-to-all space switch would need a clearly unattainable number of ports for all wavelengths (10,240 input/output ports). Fortunately, not all combinations are feasible due to the wavelength constraints, as we only need to shuffle the same wavelength across input and output fibers. Fig. 1b shows our design, where we introduce a number of optical space switches equal to the number of channels per fiber; the port count of each space switch is then equal to the number of total fibers reaching the RNG (one channel per space switch), 160 input/output ports in this example. Such space switches are available today.

3. Experimental setup and results

The experimental setup (Fig. 2) mirrors a scaled-down optical RNG topology (Fig. 1b). Two dual polarization (DP) 200 Gbit/s 16QAM optical signals are generated by commercially available real-time coherent transceivers, Acacia AC200, to produce 2^{31} pseudo random bit sequences. They are spectrally shaped with a root-raised cosine and 0.2 roll-off factor. An amplified spontaneous emission (ASE) source generates DWDM (dummy) channels which are then split and multiplexed with the signals in two separate single mode fibres (SMFs) via two wavelength selective switches (WSSs). The carrier wavelengths can be tuned within the C-band and their values are identical so they can be dynamically re-routed to different DCs. At the receiver side, the optical-to-electrical converted signal is fed to the application-specific integrated circuit (ASIC)'s analogue to digital conversion for further processing by the ASIC's digital signal processing, which includes signal recovery, polarization mode dispersion and chromatic dispersion compensation, before forward error decoding (FEC) decoding. Pre-FEC BER measurements are taken every 10 msec and the received powers are kept within the range of the receiver's optimal performance.

Each signal is amplified and launched in two transmission lines of different lengths, 20 km and 60 km SMFs, respectively, together with ASE sources to emulate two 37.5 GHz fully loaded C-band lines. Two extra DP 200 Gbit/s 16QAM optical signals, used as dummy channels only, are located at both side of one of the two channels under test, instead of ASE at those wavelengths, to investigate their effect (Fig. 2). All EDFAs are operating at constant gain, followed by fixed attenuators. In our optical RNG (Fig. 2) one space switch is used to switch the wavelength of the signals, while, for simplicity, all the other channels are routed to the second one to achieve the same functionality. The space switches have a loss of less than 1 dB. The reshuffled fully loaded lines are then separately routed to two amplified transmission lines of different lengths, 35 km and 60 km SMFs, respectively. This way we emulate varying span lengths with DC-to-DC distances ranging from 55 km to 120 km.

We first study the effects of the increased number of filters and of the amplifiers' transients in Fig. 3. For the filtering characterization, we modify our setup in Fig. 2 to a chain of WSSs and amplifiers to compensate for their losses where needed. The channel under test is set at 1550 nm, but similar results are expected for any wavelength

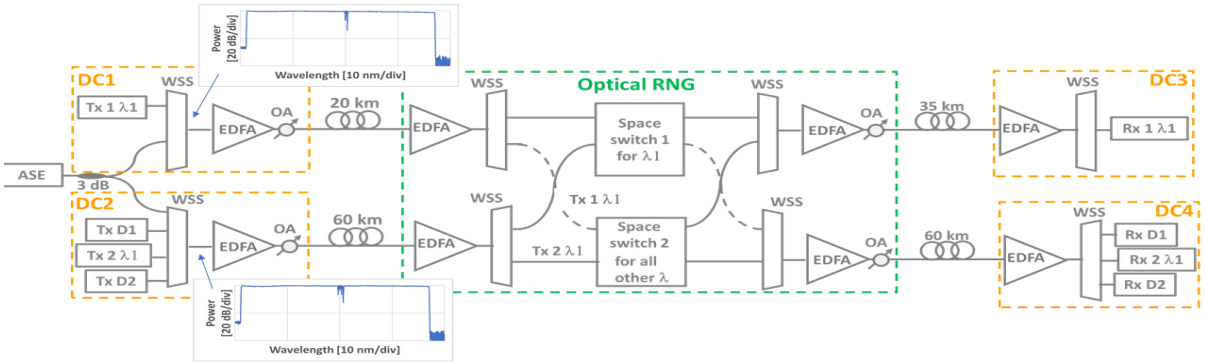

Fig. 2: Experimental setup. TX: transmitter, RX: receiver, EDFA: erbium-doped fiber amplifier. Insets: fully loaded spectra.

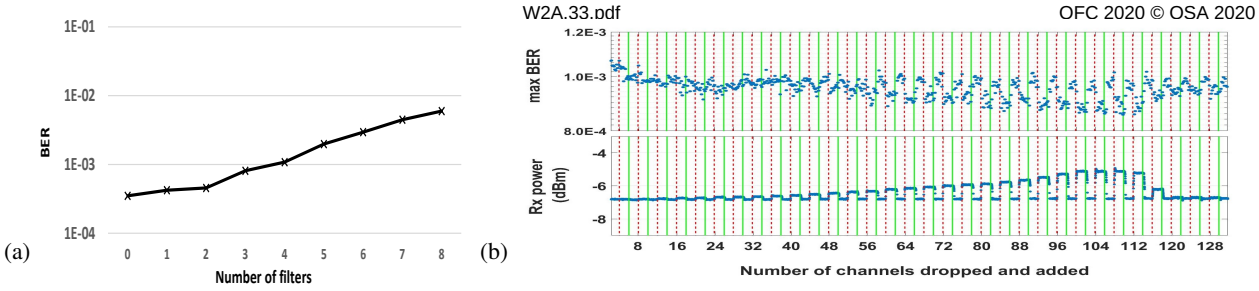

Fig. 3: (a) BER versus number of filters. (b) Maximum BER (over 1 sec interval) and receive power at the transceiver as increasing number of channels are being removed and added. Green (red dashed) lines identify the add (drop) times.

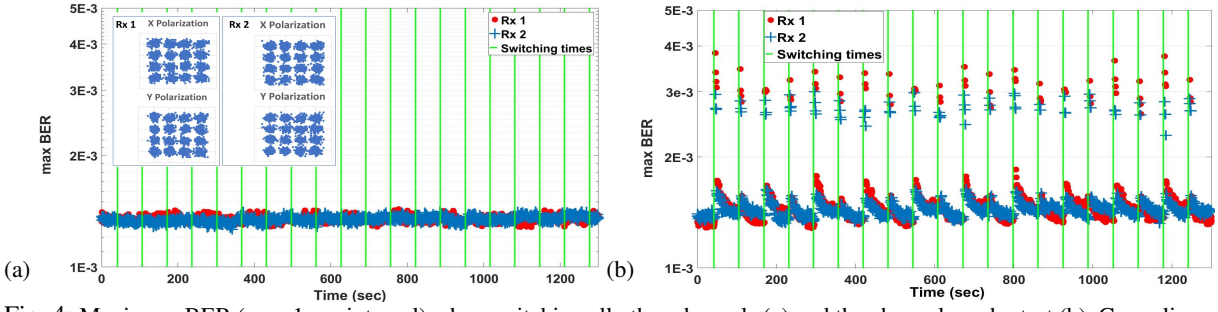

Fig. 4: Maximum BER (over 1 sec interval) when switching all other channels (a) and the channels under test (b). Green lines identify the switching times. Inset: corresponding constellation diagrams.

within the C-band. BER measurements are taken as we set each WSS from a non-filtering configuration to a filter band of 37.5 GHz centered at the signal carrier frequency. Fig. 3a shows minimal degradation as we go through 4 WSSs, the number required in our system. A BER degradation is observed for higher numbers of filters. To study the amplifier transients, we launch one fully loaded line through two cascaded amplifiers separated by an attenuator, set to match the amplifier's gain. Fig. 3b shows the BER of the surviving channel and its received power after the amplifiers as we drop and re-add other channels, to emulate the switching functionality, using a WSS. Negligible BER penalty can be seen even for the extreme case of all channels (but the one under test) dropped, while at the receiver the power increases roughly by 2 dB in the worst case.

We then study the performance of the whole system (Fig. 4). First, all channels (but the ones under test) are switched. Fig. 4a shows the maximum BERs of the surviving channels versus time; no BER degradations are observed during the switching time. Second, we examine the performance when only the channels under test are switched. Fig. 4b shows comparable BER across the two receivers. In all, it takes less than 200 msec for both receivers to recover the signals and less than 5 seconds to achieve stable BER values (comparable to the values before reconfiguration). Such switching times are acceptable given the stability of traffic. In real scenarios, our software controller will ensure that no live traffic is carried on the channels to be reconfigured. This is standard practice achieved through layer-3 traffic engineering and is feasible as DCIs are overprovisioned to account for failures. Overprovisioning and the existence of multiple fibers per DC can also ensure that our software controller has enough wavelengths to assign traffic to DC-to-DC paths without traffic being blocked in our architecture.

4. Conclusion

We have demonstrated the feasibility of an all-optical architecture for DCI networks, with filtering and amplification transient effects having minimal impact on signal performance. Two hundred milliseconds was enough for the signals to recover after reconfiguration (5 seconds to reach stable performance) and no BER degradations of test channels were observed when reconfiguring all the other channels in the line. An all optical network trades-off switching granularity in time and space (wavelength circuit reconfigurations vs. packet switched electrical network) to significantly reduce cost and power requirements. We believe DCI is the right space to achieve this.

References

1. M. Filer, J. Gaudette, Y. Yin, D. Billor, Z. Bakhtiari, and J. L. Cox, "Low-margin optical networking at cloud scale," J. Opt. Commun. Netw. **11**, C94–C108 (2019).
2. OIF, "400ZR," https://www.oiforum.com/technicalwork/hot-topics/400zr-2/.
3. A. S. Ahsan, C. Browning, M. S. Wang, K. Bergman, D. C. Kilper, and L. P. Barry, "Excursion-free dynamic wavelength switching in amplified optical networks," IEEE/OSA J. Opt. Commun. Netw. **7** (2015).
4. T. Zami, I. F. de Jauregui Ruiz, A. Ghazisaeidi, and B. Lavigne, "Growing impact of optical filtering in future wdm networks," in *Optical Fiber Communication Conference (OFC) 2019,* (OSA, 2019), p. M1A.6.
5. M. Filer and S. Tibuleac, "Cascaded ROADM Tolerance of mQAM optical signals employing nyquist shaping," in *2014 IEEE Photonics Conference,* (2014), pp. 268–269.