Elastic Optical Networking in the Microsoft Cloud

Mark Filer, Jamie Gaudette, Monia Ghobadi, Ratul Mahajan, Tom Issenhuth, Buddy Klinkers, and Jeff Cox

***Abstract*—To keep pace with the tremendous bandwidth growth in cloud networking, web-scale providers, such as Microsoft, have been quick to adopt elastic features of modern optical networks. In particular, colorless flexible-grid reconfigurable optical add–drop multiplexers, bandwidth-variable transceivers, and the ability to choose a variety of optical source types are integral for cloud network operators to improve network efficiency while supporting a variety of service types. We take an in-depth look at Microsoft's deployed network infrastructure and discuss the impact of elasticity on network capacity and flexibility. As a proof-of-concept, a new elastic open line system (OLS), in which the line system components and the signal sources are disaggregated, was assembled in a laboratory environment, and 4000 km of propagation over primarily nonzero dispersion-shifted fiber using multiple source types is demonstrated. Finally, the long-term goal of unifying the control plane of the OLS, DWDM signal sources, routers, and Ethernet switches under a single software-defined network controller is briefly addressed.**

***Index Terms*—Data center networking; Elastic optical networks; Optical coherent transceiver; Optical fiber communication; Software-defined networking.**

I. Introduction

This paper examines advances in elastic optical networking (EON) technologies and considers how these developments may be leveraged in Microsoft's long-haul data center network architecture. Microsoft's currently deployed intercity network is based largely on traditional fixed-grid technology, and work is under way to adopt EON features and capabilities into the next generation of deployable long-haul solutions. The advent of colorless, flexible-grid architectures and bandwidth-variable transceivers (BVTs), and the availability of coherent optical sources in a variety of form factors and platforms, provide opportunities to maximize capacity, spectral utilization, and spectral efficiency like never before. In addition, developments brought about by the capabilities of software-defined networking (SDN) for cloud network providers will enable the full range of benefits of network elasticity to be utilized.

Manuscript received January 21, 2016; revised April 7, 2016; accepted April 19, 2016; published May 25, 2016 (Doc. ID 257847).

M. Filer (e-mail: mark.filer@microsoft.com), J. Gaudette, T. Issenhuth, B. Klinkers, and J. Cox are with Azure Networking, Microsoft Corporation, Redmond, Washington 98052-6399, USA.

M. Ghobadi and R. Mahajan are with Microsoft Research (MSR), Microsoft Corporation, Redmond, Washington 98052-6399, USA.

http://dx.doi.org/10.1364/JOCN.8.000A45

The paper is organized as follows. In Section II, we begin by taking a high-level look at advances in line system and source technologies that enable elastic features in optical networks, with particular attention paid to those that are most relevant to cloud network operators such as Microsoft. Specifically, we discuss colorless, flexible-grid reconfigurable optical add–drop multiplexer (ROADM) architectures, disaggregation of the photonic level, and bandwidth-variable features of modern transceivers. In Section III we take a detailed look at some of the drivers in Microsoft's network for rapid adoption of elastic networking features. In particular, we present data summarizing the physical fiber makeup of Microsoft's North American backbone network, including fiber type and span length distributions. We then disclose the results of a three-month study of network performance in the form of logged Q-factor/signal-to-noise ratio (SNR) data for all currently deployed coherent transceivers in the network. Finally, we look at the gains that may be brought about by incorporating BVTs into the network. We follow this with Section IV, which presents results from a lab trial of an elastic open line system (OLS), in which the line system and optical sources are disaggregated. The elastic OLS utilizes colorless add/drop and 20 deg flexible-grid ROADMs. The coherent sources, which reside on a layer 2/3 modular switch linecard and utilize CFP2-ACO (analog coherent optical) pluggable optical modules, are transported through the OLS over 4000 km with Nyquist channel spacing. Finally, Section V touches on Microsoft's approach in incorporating the optical line system and sources into its SDN controller and the associated challenges that lie ahead.

II. Elastic Optical Networking Enablers

Several recent developments in optical networking technology have allowed for the rapid adoption of elastic features in the Microsoft network. Advances in line system components and subsystems, disaggregation at the photonic layer, and the advent of BVT devices have all coalesced to give the network operator far more flexibility than ever before. We present a brief summary of the key elastic enablers below, which are of primary relevance to the Microsoft ecosystem.

A. Line System Technology

One of the most significant advances in DWDM system technology over the last 7–8 years has been the advent of

flexible-grid capable ROADMs [1–4]. The core underlying component that enables full bandwidth flexibility in ROADMs is the wavelength-selective switch (WSS) [5], which allows the switching of incoming wavelengths from an input port to any of N output ports. Some of the first iterations of WSSs in the marketplace were based on micro-electro-mechanical mirrors (MEMS) technology and were channelized by design due to physical limitations of the MEMS mirrors themselves [6,7]. Later iterations, based on digital light processing (DLP) [8,9] and (to a larger degree) liquid crystal on silicon (LCoS) [10,11], were able to circumvent this channelization limitation due to the very fine granularity of the pixels that make up the devices. The current generation of DLP- and LCoS-based WSSs is capable of directing arbitrarily wide segments of continuous spectrum from any input port to any output port with resolutions down to 6.25 GHz [11,12]. This flexibility comes only at the cost of slightly reduced port isolation when compared to MEMS-based devices [13], but it has been shown that this does not present much of a compromise in practice [14].

By introducing this form of elasticity into WSS devices, networks can be designed to fully utilize spectrum that would otherwise be unused in the more traditional fixed-grid configuration. Additionally, flexible-grid WSSs allow for the possibility of dynamic reconfiguration of network wavelength routing, reducing the operational burden on the service provider. Lastly, it enables the line system itself to be on a refresh cycle that far outlives that of the optical sources due to the future-proofing that flexibility provides for future modulation types.

Along with the WSS, the added combination $M \times N$ optical switches and colorless multiplexers (with optical amplifiers that are, by nature, colorless) enable ROADM topologies that are both directionless and contentionless [15,16], which adds another level of elasticity to the photonic layer. These features are outside the scope of most cloud providers' network requirements due to the point-to-point, static nature of distributed data center networks. For full treatment of ROADM architectures, see [3,4,17].

B. Photonic Layer Disaggregation

The web-scale cloud networking environment that Microsoft operates in has increasingly imposed a paradigm shift onto traditional optical transmission system suppliers in several ways.

First, the requirements and feature sets of the optical line systems and sources are greatly relaxed over those required by "tier-1" telco providers, making for a less complex line system design. In Microsoft's case, the following features typically found in telco providers' optical networks can be omitted:

- electrical OTU switching,
- optical layer restoration,
- sub-line-rate aggregation and grooming,
- optical "bandwidth on demand,"
- true mesh connectivity.

Microsoft's intercity network requirements can be fully addressed by a line system with a streamlined feature set optimized for coherent transmission. The network is made up largely of point-to-point segments connecting geographical regions, so directionless and contentionless ROADM capabilities are generally not in scope.

Second, since advances such as colorless, flexible-grid ROADMs have led to a potentially longer shelf life for the line system hardware, it is likely that the line system can remain viable through several technology refreshes of the coherent sources that sit on the ends. The line system could perhaps be used for two, three, or more generations of coherent source technology.

Lastly, because all of the traffic in such networks is packet-based, some simplifications can be made over traditional designs, e.g., the elimination of gray optical interfaces and/or the need for demarcation between layer 1 and layer 0.

This has led Microsoft to pursue the OLS concept [18], which is a disaggregation of the photonic layer in which the DWDM optical transceivers/transponders are decoupled from the DWDM line system (Fig. 1). (This is similar in concept to work others have done in the ITU-T's G.698.2 "black link" recommendation; see, for example, [19]). An important implication of such an approach, which we consider to be another dimension of "elasticity" for the Microsoft network, is the ability to procure coherent optical sources from multiple vendors and platforms. Three primary categories these fall into are the following:

1) traditional transponder linecards with gray client-side optics;
2) more recent high-density "data center interconnect" (DCI) sources with reduced feature set (no OTN support, minimized sheet metal, and reduced alarming capabilities), but still with gray client optics [20];
3) layer 2/3 modular linecards with embedded coherent DSP chips and pluggable analog coherent optical (ACO) modules.

To properly handle the variety of coherent optical source options, the line system must be able to efficiently manage alien wavelengths and allow for full add/drop of arbitrarily shaped channels or super-channels at terminal sites.

Disaggregating the photonic layer forces more intense marketplace competition among coherent source and line system providers, as well as giving the cloud provider flexibility (elasticity) to select which optical source platform best suits their needs for a given application. It also enables the path toward true SDN control in the optical domain (Fig. 1), with northbound interfaces from the line system network monitoring system (NMS), and eventually from the network elements (NEs) themselves, to the network orchestration layer—typically developed in-house by the cloud provider. This aspect is addressed in more detail in Section V.

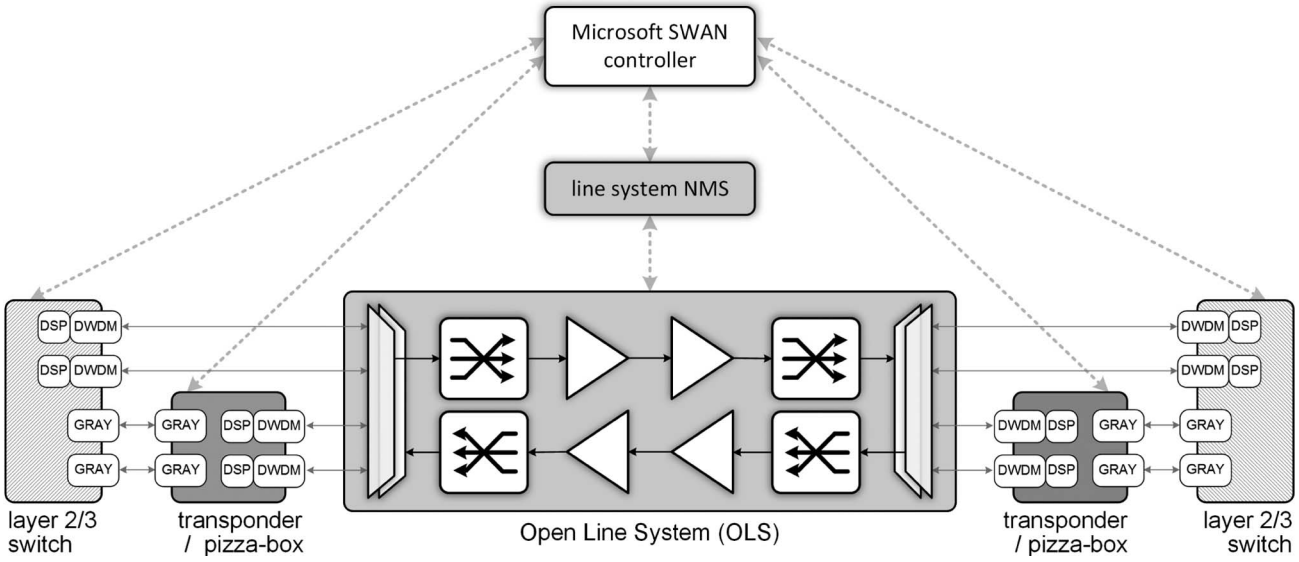

Fig. 1. Open line system (OLS) concept showing disaggregated photonic layer with multiple coherent source platforms (traditional transponder, high-density "pizza-box," and layer 2/3 switch with embedded coherent optics). All layers of the ecosystem tie into one overarching SDN controller.

C. Bandwidth-Variable Transceivers

A third aspect of EONs that Microsoft is keen to exploit is BVTs, sometimes also referred to as "software-defined optics." Such BVTs trade physical layer performance for network capacity by using variable coded modulation, rate-adaptive FEC, parity coding, or other means [21–23]. Several examples of this technology are generally available in the marketplace today, including units that offer variable modulation formats such as BPSK, QPSK, 8QAM, and 16QAM [24–26] (Fig. 2), and varying levels of FEC coding gain (with associated increases in physical bandwidth overhead). Future advances in BVT technology will include variable baud rates, time-domain hybrid modulation schemes [27,28], and shaped constellation modulation (i.e., variable constellation alphabet) [22,29–31].

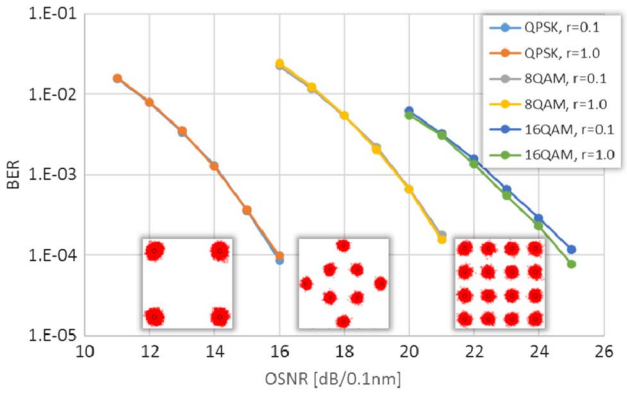

Fig. 2. Example of BVT: QPSK/8QAM/16QAM constellations and performance for different RRC roll-off factors at 32 GBaud.

III. Elasticity Drivers in Microsoft's Network

The cost of acquiring and lighting additional fiber pairs relative to the low cost of using the existing infrastructure more efficiently is a primary driver toward adopting EON capabilities in the Microsoft network. Microsoft's existing backbone uses traditional fixed-grid technology, so there are obvious benefits to be had in spectral utilization by moving to a system that is flexible-grid capable. However, the gains offered by BVTs are more difficult to quantify without a detailed study of the existing infrastructure.

Toward that end, we performed both measured and simulated studies on our North American (N.A.) backbone. For the measured study, we polled the Q-factor of all 100G PM-QPSK channels in the network (thousands of linecards) over a period of three months to get a sense of statistical performance variation. For the simulation study, offline simulations were performed in cooperation with our transport suppliers over segments of our fiber infrastructure where the performance of 8QAM and 16QAM modulation formats was modeled in order to determine nonlinear penalties. Then, we further examined the possibility of improving the granularity from the 50 Gb/s of QPSK/8QAM/16QAM to 25 Gb/s or less in order to see whether the additional complexity of improved granularity is offset by realized gains. This section is, in part, a summary of the results published in [32].

A. Microsoft N.A. Backbone Fiber Summary

Before delving into results, it is helpful to highlight some details of the physical fiber infrastructure that makes up the Microsoft backbone. The deployed fiber is primarily composed of ITU-T G.655 nonzero dispersion-shifted fiber

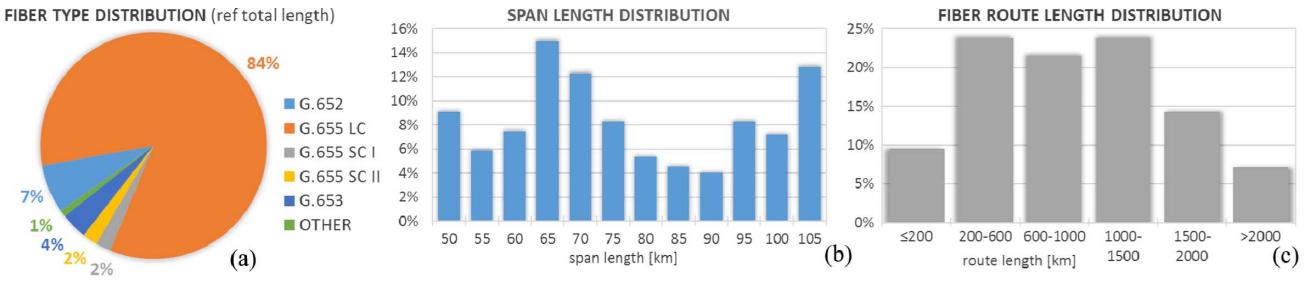

Fig. 3. Microsoft North American backbone fiber information: (a) distribution of fiber types found in the network (LC, large core; SC, small core). (b) Distribution of individual span lengths throughout all network routes. (c) Route length distribution.

(NZ-DSF) fiber, with a relatively small proportion of G.652 standard SMF fiber (SSMF). Figure 3(a) shows the mix of fiber types found in the backbone, with percentage values indicating the total length of that fiber type relative to the total fiber in the network. As can be seen, the network is largely dominated by ITU G.655 NZ-DSF fibers (over 90%), with only 7% G.652 SSMF. This makes the network quite challenging to design for from a nonlinear standpoint, and implies that measured nonlinear penalties should be relatively high.

Figure 3(b) shows the distribution of individual fiber span lengths over all of the N.A. backbone routes. The relatively high concentration of fiber spans longer than 90 km would suggest a sizable infrastructure of nodes utilizing hybrid Raman-EDFA amplification.

Finally, Fig. 3(c) shows the range of route lengths that make up the N.A. backbone arranged in relevant route length bins. As can be seen, routes carrying DWDM traffic between data center regions range from sub-100-km to more than 2500 km.

B. Capacity Gain With Line-Rate Granularity

The polling of Q-factor from Microsoft's existing N.A. backbone network was performed over a period of three months, from February to April 2015, for all of the 100G PM-QPSK linecards deployed at the time. The linecards had a baud rate of 31.8 GBaud. At the time the data was taken, several segments of the network still had inline dispersion compensating modules (DCMs) in order to support legacy technologies. Polling samples were in consecutive 15 min bins, with minimum, maximum, and average values extracted over each 15 min window. Average values are used in the analysis because, for the vast majority of samples, the difference in min, max, and average was negligible.

The observed Q-factor measurements are converted to received electrical SNR in order to plot the cumulative distribution function (CDF) of all the samples over the three-month period. The received SNR is defined as E_S/N_0, where E_S is the average symbol energy, and N_0 is the double-sided noise power spectral density (PSD) [33]. The measured Q-factors include all linear and nonlinear propagation penalties. These values should be conservative because they include samples from the network segments that still had inline DCMs, where nonlinearities are higher than fully uncompensated (i.e., purely coherent) links.

1) QPSK, 8QAM, and 16QAM: In order to estimate the gains to be had by BVTs, we calculated the electrical SNR limits of 8QAM and 16QAM modulation formats. For these calculations, we assumed transceivers operating on a flexible grid at 32 Gbaud, with a FEC limit of 3.77e-2 (or equivalently, a Q_{dB} of 5.0), and a minimum OSNR at the FEC limit of 10 dB for QPSK, 14 dB for 8QAM, and 17 dB for 16QAM. For 8QAM and 16QAM, nonlinear propagation penalties were simulated to be between 0 and 2.2 dB and 0 and 1.5 dB of SNR (respectively), depending on segment fiber type and distance. The smaller penalty for 16QAM is most likely explained by the shorter propagation distances supported relative to 8QAM.

The black CDF series in Fig. 4 shows the distribution of SNRs observed from the 100G DP-QPSK signals in the network over the three-month period. The shaded segments of the figure show the computed SNR limits for each modulation format, with the widths spanning the range of simulated propagation penalties. By using these SNR limits, it is estimated that 8QAM could address between 78% and 99% of the samples from the network, and 16QAM could address between 12% and 43% of them, where the range

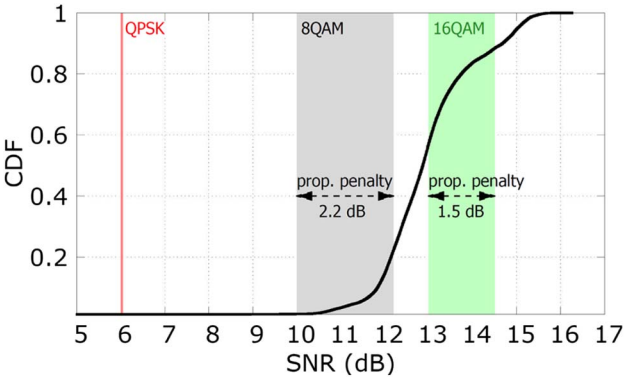

Fig. 4. CDF of SNRs observed over three months on existing 100G PM-QPSK infrastructure (black series) overlaid with SNR requirements for PM-8QAM and PM-16QAM.

in percent addressable corresponds to the range of propagation penalty for each modulation format. The net capacity gain amounts to between 45% and 70% by using 8QAM and 16QAM, where possible.

2) Sub-50-Gb/s Resolution: Next, we considered the gains possible using a finer granularity of 25 Gb/s between line rates of 100 and 250 Gb/s, which conveniently matches up with the proposed granularity outlined in the OIF FlexEthernet Implementation Agreement [34,35]. For 225 and 250 Gb/s, we assume SNR limits corresponding to 512SP-QAM and PM-32QAM, respectively [22]. We used a simplified assumption in estimating the SNR limits for 125 and 175 Gb/s by estimating that they lie halfway between QPSK and 8QAM and 8QAM and 16QAM, respectively. In reality, these values would depend on implementation, but for the purposes of this exercise, they are representative. The vertical lines in Fig. 5 indicate the approximate 25 Gb/s granularity SNR limits, overlaid again with the SNR CDF (black series) from Fig. 4. Figure 6 depicts the cumulative percentage of channels that can be addressed as a function of line rate, shown on the x axis. For reference, the previous 50 Gb/s granularity is shown in blue, and the finer 25 Gb/s granularity is shown in magenta. The plot shows that greater than 90% of samples can increase their capacity to 175 Gb/s or higher. The average capacity gain using 25 Gb/s steps is 86%, compared to the 70% gain previously determined using 8QAM and 16QAM. These results demonstrate that a 25 Gb/s granularity on optical line rate can offer 16% additional capacity gain (green hatched area of figure) over the coarser 50 Gb/s granularity of QPSK, 8QAM, and 16QAM.

To investigate the further possible gain, analysis was performed with 1 Gb/s line-rate granularity. This is represented with the red series in Fig. 6. Only an additional 13% gain is realized over the 25 Gb/s granularity case. Given the disproportionately high complexity required to achieve granularity of 1 Gb/s (both in line-side DSP and layer 2/3 devices), and the relatively modest 13% gain over 25 Gb/s granularity, we do not recommend finer than 25 Gb/s line-rate resolution.

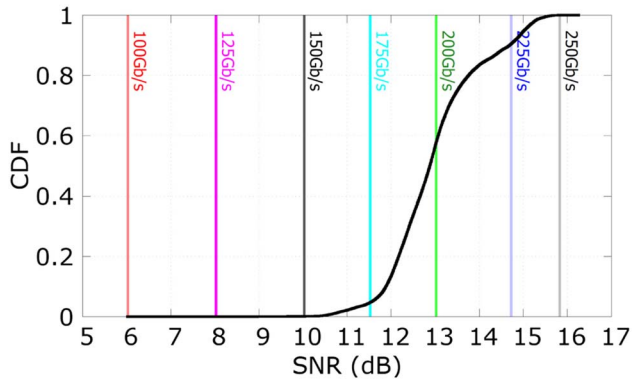

Fig. 5. Same CDF from Fig. 4 (black series) overlaid with SNR requirement line rates in 25 Gb/s step sizes (vertical lines).

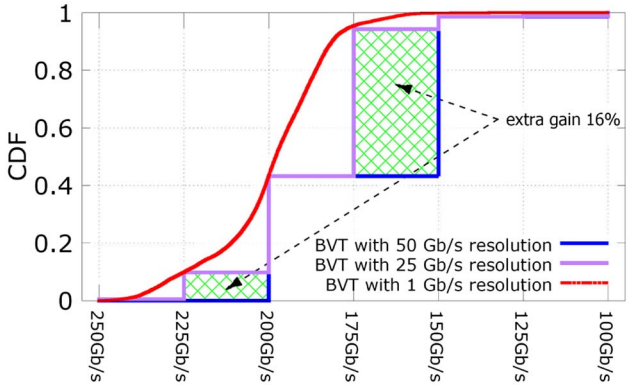

Fig. 6. CDF as function of line rate for 50 Gb/s (blue), 25 Gb/s (magenta), and 1 Gb/s (red) granularities. Hatched areas are realizable gain when going from 50 to 25 Gb/s granularity.

C. Q-Factor Variability Over Time

In order to understand how much a given channel's Q-factor might be expected to vary over time, we generated a CDF of Q-factors over channels across a representative sample of network segments over the three-month period. The CDFs of the per-channel Q value standard deviations are shown in Fig. 7. The results indicated that during "steady-state" network operation (e.g., maintenance and fiber cut events omitted), the standard deviation of the Q values was less than 0.2 dB for 95% of the samples.

D. Remarks

Based on the data examined, we estimate that substantial gains, near 70%, can be achieved using elastic BVTs with 50 Gb/s line-rate granularity, i.e., QPSK, 8QAM, and 16QAM. A further increase of 16%, for a total capacity gain of 86%, can be realized using BVTs with 25 Gb/s resolution. The relative lack of Q-factor variance over time would indicate that BVTs in cloud networks can likely be static entities, configured at start of life or after major network

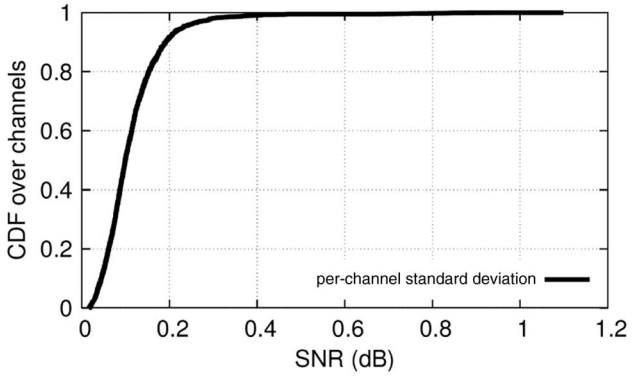

Fig. 7. CDF of the per-channel Q value standard deviations over a representative sample of network segments. Q-factor deviation is less than 0.2 dB for 95% of samples with the network in "steady-state" operation (e.g., omitting fiber cut or maintenance events).

maintenance events only, not needing to adapt dynamically to changing network conditions.

IV. Elastic OLS Proof-of-Concept

In this section we demonstrate in a laboratory environment the operation of an elastic OLS that makes use of the technologies referenced in Section II. This is accomplished using a commercially available DWDM line system operating in an open fashion, which is configured to transport Nyquist-shaped PM-QPSK signals at 32 GBaud over a mix of ITU-T G.652 and G.655 fibers. The optical sources reside on a layer 2/3 modular Ethernet switch card with embedded coherent ASICs and corresponding CFP2-ACO pluggable optical modules. Transmission over 4000 km of fiber with seven channels spaced 37.5 GHz apart is achieved with considerable OSNR margin. The results in this section are a summary of the results published in [36].

A. Open Line System

A line system and approximately 4000 km of fiber were procured and installed in the optical transmission lab at Microsoft's Redmond headquarters. The transmission testbed has been configured to emulate large portions of Microsoft's N.A. backbone network, with enough G.652 standard SMF (Corning SMF-28e LL) and G.655 NZ-DSF (Corning LEAF) to pass traffic over two bidirectional ~2000 km paths. A summary of the span fiber types, lengths, and associated amplification schemes is shown in Table I and Fig. 8 for the 2000 km forward path. To achieve a full 4000 km transmission, the system is looped back on itself after span 25.

TABLE I
Fiber Type, Distance, and Amplification Scheme of Lab System

Span	Fiber	km	ps/nm	Amplifier
1	G.652 SMF	75	1275	EDFA
2	G.655 LEAF	75	375	EDFA
3	G.655 LEAF	75	375	EDFA
4	G.655 LEAF	110	550	EDFA + DRA
5	G.655 LEAF	70	350	EDFA
6	G.655 LEAF	65	325	EDFA
7	G.655 LEAF	100	500	EDFA + DRA
8	G.655 LEAF	65	325	EDFA
9	G.655 LEAF	75	375	EDFA
10	G.655 LEAF	70	350	EDFA
11	G.655 LEAF	70	350	EDFA
12	G.655 LEAF	75	375	EDFA
13	G.655 LEAF	65	325	EDFA
14	G.655 LEAF	90	450	EDFA + DRA
15	G.655 LEAF	65	325	EDFA
16	G.655 LEAF	70	350	EDFA
17	G.655 LEAF	100	500	EDFA + DRA
18	G.655 LEAF	75	375	EDFA
19	G.655 LEAF	75	375	EDFA
20	G.652 SMF	70	1190	EDFA
21	G.652 SMF	65	1105	EDFA
22	G.655 LEAF	75	375	EDFA
23	G.655 LEAF	75	375	EDFA
24	G.655 LEAF	120	600	EDFA + DRA
25	G.652 SMF	75	1275	EDFA
Total		*1945*	*13,145*	

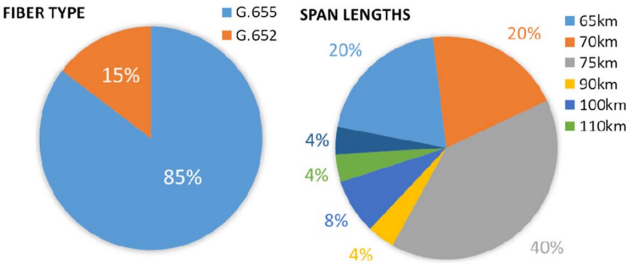

Fig. 8. Lab system fiber type and span length distributions.

The upper portion of Fig. 9 depicts the line system configuration, which utilizes a colorless architecture. There are colorless mux/demux and flexible-grid 20 deg ROADMs at network ingress and egress, and either 9 deg broadcast-and-select or 20 deg route-and-select flex-grid ROADMs after every fifth span, primarily for channel power balancing (these also allow for the option of future channel add/drop at these nodes). The spans range in length between 65 and 120 km, with one span in every five long enough to require backward-pumped Raman amplification in addition to EDFA (either every second or fourth span, shown by the dotted "R" modules in Fig. 9).

B. Coherent Linecard

The optical sources used in the testbed reside on a layer 2/3 modular switch card made by Arista (7500E-6CFPX, Fig. 10) for its 7500E platform. The linecard integrates the 100-GbE Ethernet layer 2/3 switch silicon with wire-speed MACsec encryption and DSP silicon supporting coherent PM-QPSK transmission over long-haul distances. Each linecard has six embedded coherent ASICs and corresponding CFP2-ACO pluggable optical modules. This is depicted in the gray lower box in Fig. 9; two linecards were used to host the seven DSPs and ACO modules, but shown conceptually as one in the figure.

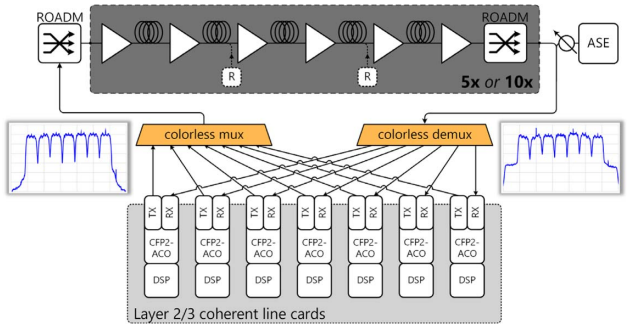

Fig. 9. 4000 km optical transmission testbed—elastic OLS with coherent layer 2/3 Ethernet linecard as source.

Fig. 10. Arista 7500E-series layer 2/3 six-port coherent DWDM linecard.

The analog coherent pluggable modules used were CFP2-ACO OIF class 2 variants (per OIF implementation agreement [37]) from multiple vendors, each vendor using a slightly different physical implementation. They are based on maturing III-V processing technologies that enable integration of the tunable laser source with the nested polarization-multiplexed IQ Mach–Zehnder modulator (MZM). Monolithic integration of high-power laser sources and low drive-voltage modulators offers a significant reduction in power consumption over traditional discrete implementations, which is required for the CFP2 standard form factor that houses this technology.

The coherent ASIC DSP chips are located on the linecard near the front panel. While this decoupled configuration provides operational ease in swapping out modules, and allows for vendor diversity, it comes at the cost of requiring precise compensation of frequency-dependent channel loss of linecard RF traces, CFP2-ACO connectors, and E/O and O/E transfer functions of the optical assembly [38]. In order to address this, the DSP egress-side 40-tap FIR filter coefficients are carefully calculated by time-domain convolution of the aggregate RF channels' inverse-S21 responses with a theoretical root-raised cosine (RRC) sampled impulse response. An 8-bit resolution DAC generates the final analog Nyquist-shaped waveform that feeds the CFP2-ACO linear drivers.

C. Measurements and Results

For this demonstration, two Arista linecards and seven early "alpha" CFP2-ACO units were employed to generate seven channels of 32 GBaud Nyquist-shaped PM-QPSK signals spaced at 37.5 GHz (ACO transmit spectrum Fig. 11). A RRC with a roll-off factor of 0.2 was used for all of the 37.5-GHz-spaced channels, as it empirically provided the best compromise between maximum spectral confinement and minimum BER penalty due to jitter that may exceed the DSP timing recovery tolerance. During calibration, no significant OSNR penalty was measured between Nyquist-shaped signals with 0.2 and 1.0 roll-off factors, confirming that RF channel loss precompensation was sufficient in the full [0,16] and [16,32] GHz frequency ranges. As seen in the optical spectrum (see Fig. 9 inset, and Fig. 11), carrier leakage of varying severity affected a few of the neighbor channels, suggesting suboptimal modulator bias control. Issues like these are expected with alpha units, and have been resolved in the commercially available final products. The measurements that follow were performed only on the center channel, which delivered the expected transmit spectrum and QPSK constellation.

Bit-error ratio measurements were performed, first back-to-back, then with the signals launched in the 2000 km forward path, and finally with the signals looped back through the reverse path to achieve the full 4000 km transmission. For each distance, launch power was swept over a range to find the optimum before making final BER versus OSNR measurements (with "optimum" referring to optimizing BER performance, not maximizing OSNR margin). Figure 12 shows Q-factor versus per-channel launch power into NZ-DSF (launch powers into the SSMF spans were nominally 3 dB higher than powers shown on the figure's x axis). A discernible impact from nonlinearity is observed when going from single-channel to multi-channel transmission for both distances, with a Q-penalty approaching 1 dB and a reduction in optimum launch power ranging from 1 to 2 dB.

Next, BER versus OSNR measurements were performed at the optimum launch powers for each case (Fig. 13). After 2000 km transmission, there is about 0.5 dB OSNR penalty at 1e-2 for both single- and multi-channel transmission, and for 4000 km, this penalty increases to nearly 1.3 dB. Even at 4000 km, the system operated with more than 3 dB of OSNR margin from the coherent DSP SD-FEC

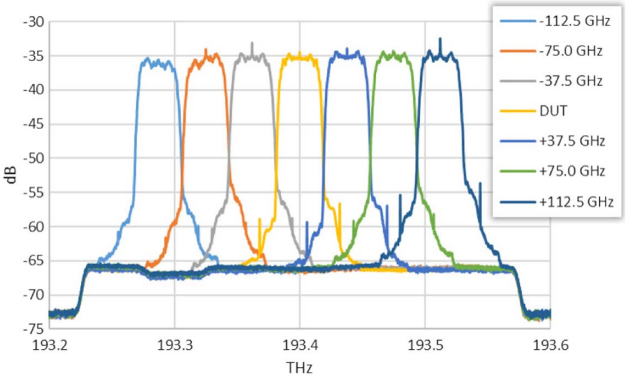

Fig. 11. Transmitted spectrum of Nyquist 100G PM-QPSK channels originating from a layer 2/3 Ethernet switch card with embedded coherent DSP and CFP2-ACO class II optics.

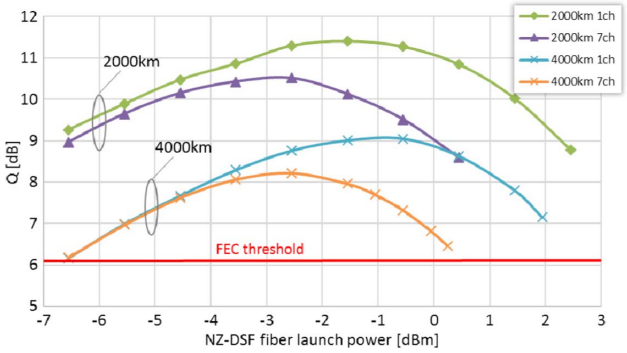

Fig. 12. Q versus launch power into NZ-DSF spans (power into SSMF spans was 3 dB lower than values shown) for Nyquist PM-QPSK signals over an elastic OLS testbed.

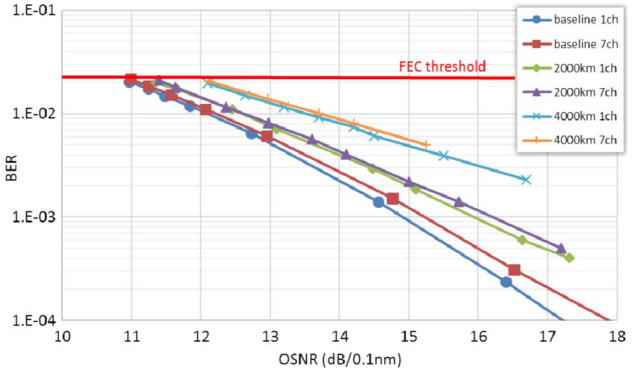

Fig. 13. BER versus OSNR at optimal launch powers for Nyquist PM-QPSK signals over an elastic OLS testbed.

threshold. Propagation with a fully filled 120-channel spectrum could be expected to generate additional OSNR penalty on the order of 1 ± 0.2 dB [39,40].

D. Remarks

For Microsoft, this demonstration was the first step in showing the feasibility of a truly elastic and flexible OLS. We were able to demonstrate a line system based on colorless, flexible-grid technology interoperating with alien sources residing in a layer 3 Ethernet switch platform, with no perceivable implementation penalty over a more traditional "locked" line system with vendor-specific transponders [41–43]. At the time of the measurements, we were not able to fully leverage the bandwidth-variable property of the coherent sources, but this testbed is permanent and these types of demonstrations will be done in the near future.

V. Software-Defined Networking

Microsoft's wide-area network (WAN) spans the globe, and is one of the larger networks in the Internet. A diverse set of services runs across the WAN, ranging from enterprise cloud applications and email to search and cloud storage. This service diversity leads to a diverse set of requirements on the WAN. For example, enterprise applications must experience near-perfect availability, high capacity, and the flexibility to change with unpredictable service adoption. On the other hand, internal storage replication across data centers can be treated with low priority, and although it produces a large volume of traffic, this can be scheduled and planned around the needs of the higher priority traffic. The combination of scale and application diversity presents significant operational challenges.

SDN enables centralized management and control of network infrastructure [44], and this ability has been leveraged by Microsoft to boost the efficiency of its WAN infrastructure [45]. Our current SDN approach, called SWAN (software-defined wide-area network), assumes that the optical layer is static and dynamically computes routing paths to drive high utilization on the WAN. In other words, it moves bits around to best fit into the fixed infrastructure. Examples include routing around failures, removing or inserting low priority traffic as necessary, and automatically turning up nodes after new infrastructure installation or new service activation. In operation, SWAN measures an observed state from the network and services, computes a desired network state, and automatically applies state changes to the network.

Elastic optics can enable further optimization of the network, through the use of BVTs and ROADMs. However, when considering the scale, diversity, and multi-vendor nature of our WAN, this could add enormous complexity to the operation of the network. To cope with the added complexity, today the elastic features are configured at start-of-life to a precalculated optimum and minimal adjustments are made, manually, as required. To extract the full potential of elastic optics, SWAN needs to be extended to include dynamic control of the optical layer. To achieve this, two challenges need to be addressed. The first is to develop a standard data model and control interface for the optical layer, as was done for the IP layer [46]. YANG, RESTCONF, and SNMP provide potential starting points for such a model. The physical models that define optical performance would also need to be made available, with standard outputs consumable by the controller to dynamically determine desired states. Once under operation, ideally, the line-side optics and photonic line hardware would be interoperable and interchangeable, even across vendors, so the network and the SDN controller can operate seamlessly through technology cycles, agnostic to vendor or hardware variant. The second challenge is computational. Elastic optics, combined with routing flexibility, offer a vast number of possibilities for configuring the network. Efficient optimization techniques are needed to quickly compute network configurations that deliver high performance based on current traffic demand.

VI. Conclusion

In this paper, we have examined the implications of the EON paradigm on Microsoft's long-haul data center

network. Overall we can conclude that there are substantial efficiency gains to be had by incorporating features of modern EONs into the Microsoft network. In particular, we showed that with the addition of elastic transceiver technologies found in BVTs, the network capacity can be improved greatly (70%) with per-channel configurable QPSK/8QAM/16QAM modulation, which should be generally available in the marketplace at the time of publication. Capacity can be further improved (86% gain) using optical modulation with 25 Gb/s granularity, with many approaches to implementation discussed extensively in the literature. We also demonstrated an elastic OLS, which proved that we can maximize spectral utilization, operate with optical sources on a layer 2/3 switch card and pluggable ACO optics, and still achieve at least 4000 km transmission with Nyquist-spaced 100G PM-QPSK channels. In addition, the elastic features of such an OLS should ensure that the line system will live through multiple refresh cycles of the coherent sources, future-proofing the optical hardware infrastructure to the greatest degree possible.

There is still a considerable amount of work to be done before all benefits of the EON can be fully utilized. As discussed in the section on SDN, there are still challenges in properly modeling and controlling the plurality of optical source and line system options available in the marketplace today. Additionally, there must be a push toward standardization of a data model and control interface at the optical layer, and, in addition, a further push toward interoperability modes of the coherent sources (even at the expense of ultimate physical layer performance).

With the rapid explosion of online services, the supporting optical networks need to grow at rates never experienced before. We must continue to leverage new technological developments, such as those described in this paper, to maintain leadership in this space [47].

Acknowledgment

The authors thank Amar Phanishayee at Microsoft Research and Daniel Kilper at the University of Arizona for their support in analysis of the network polling data presented in Section III. Additionally, we extend thanks to Hacene Chaouch, Jonathan Chu, and Raju Kankipati, Arista Networks, for their hard work and support of the lab results presented in Section IV.

References

[1] P. Roorda and B. Collings, "Evolution to colorless and directionless ROADM architectures," in *Optical Fiber Communication Conf. and the Nat. Fiber Optic Engineers Conf.*, San Diego, CA, 2008, paper NWE2.

[2] A. Sahara, "Demonstration of colorless and directed/directionless ROADMs in router network," in *Optical Fiber Communication Conf. and the Nat. Fiber Optic Engineers Conf.*, San Diego, CA, 2009, paper NMD2.

[3] S. Tibuleac and M. Filer, "Transmission impairments in DWDM networks with reconfigurable optical add-drop multiplexers," *J. Lightwave Technol.*, vol. 28, pp. 557–598, 2010.

[4] S. Gringeri, B. Basch, V. Shukla, R. Egorov, and T. J. Xia, "Flexible architectures for optical transport nodes and networks," *IEEE Commun. Mag.*, vol. 48, no. 7, pp. 40–50, 2010.

[5] D. T. Neilson, C. Doerr, D. Marom, R. Ryf, and M. Earnshaw, "Wavelength selective switching for optical bandwidth management," *Bell Labs Tech. J.*, vol. 11, pp. 105–128, 2006.

[6] D. Marom, D. T. Neilson, D. S. Greywall, C.-S. Pai, N. R. Basavanhally, V. A. Aksyuk, D. O. Lopez, F. Pardo, M. E. Simon, Y. Low, P. Kolodner, and C. A. Bolle, "Wavelength-selective $1 \times K$ switches using free-space optics and MEMS micromirrors: Theory, design, and implementation," *J. Lightwave Technol.*, vol. 23, pp. 1620–1630, 2005.

[7] F. Heismann, "System requirements for WSS filter shape in cascaded ROADM networks," in *Optical Fiber Communication Conf. and the Nat. Fiber Optic Engineers Conf.*, San Diego, CA, 2010, paper OThR1.

[8] T. A. Strasser and J. L. Wagener, "Wavelength-selective switches for ROADM applications," *IEEE J. Sel. Top. Quantum Electron.*, vol. 16, pp. 1150–1157, 2010.

[9] T. A. Strasser and J. L. Wagener, "Programmable filtering devices in next generation ROADM networks," in *Optical Fiber Communication Conf. and the Nat. Fiber Optic Engineers Conf.*, Los Angeles, CA, 2012, paper OTh3D.4.

[10] G. Baxter, S. Frisken, D. Abakoumov, Z. Hao, I. Clarke, A. Bartos, and S. Poole, "Highly programmable wavelength selective switch based on liquid crystal on silicon switching elements," in *Optical Fiber Communication Conf. and the Nat. Fiber Optic Engineers Conf.*, Anaheim, CA, 2006, paper OTuF2.

[11] S. Frisken, S. B. Poole, and G. W. Baxter, "Wavelength-selective reconfiguration in transparent agile optical networks," *Proc. IEEE*, vol. 100, pp. 1056–1064, 2012.

[12] "Spectral grids for WDM applications: DWDM frequency grid," ITU-T Recommendation G.694.1, Feb. 2012.

[13] B. C. Collings, F. Heismann, and C. Reimer, "Dependence of the transmission impairment on the WSS port isolation spectral profile in 50 GHz ROADM networks with 43 Gb/s NRZ-ADPSK signals," in *Optical Fiber Communication Conf. and the Nat. Fiber Optic Engineers Conf.*, San Diego, CA, 2009, paper OThJ3.

[14] M. Filer and S. Tibuleac, "*N*-degree ROADM architecture comparison: Broadcast-and-select versus route-and-select in 120 Gb/s DP-QPSK transmission systems," in *Optical Fiber Communication Conf. and the Nat. Fiber Optic Engineers Conf.*, San Francisco, CA, 2014, paper Th1I.2.

[15] T. Zami, "High degree optical cross-connect based on multicast switch," in *Optical Fiber Communication Conf.*, San Francisco, CA, 2014, paper W2A.36.

[16] A. Malik, W. Wauford, P. Zhong, N. K. Goel, S. Hand, and M. Mitchell, "Implications of super-channels on colorless, directionless and contentionless (CDC) ROADM architectures," in *Optical Fiber Communication Conf.*, San Francisco, CA, 2014, paper W1C.4.

[17] T. Zami, "Current and future flexible wavelength routing cross-connects," *Bell Labs Tech. J.*, vol. 18, pp. 23–38, 2013.

[18] J. Cox, "SDN control of a coherent open line system," in *Optical Fiber Communication Conf.*, Los Angeles, CA, 2015, paper M3H.4.

[19] M. Gunkel, A. Mattheus, F. Wissel, A. Napoli, J. Pedro, N. Costa, T. Rahman, G. Meloni, F. Fresi, F. Cugini, N. Sambo, and M. Bohn, "Vendor-interoperable elastic optical interfaces: Standards, experiments, and challenges [Invited]," *J. Opt. Commun. Netw.*, vol. 7, pp. B184–B193, 2015.

[20] D. O'shea, "DCI boxes aren't just for metros anymore," Sept. 23, 2015 [Online]. Available: http://www.lightreading.com/data-center/data-center-interconnect-/dci-boxes-arent-just-for-metros-anymore/a/d-id/718343.

[21] Y. Yoshida, A. Maruta, K.-I. Kitayama, M. Nishihara, T. Tanaka, T. Takahara, J. C. Rasmussen, N. Yoshikane, T. Tsuritani, I. Morita, S. Yan, Y. Shu, Y. Yan, R. Nejabati, G. Zervas, D. Simeonidou, R. Vilalta, R. Munoz, R. Casellas, R. Martinez, A. Aguado, V. Lopez, and J. Marhuenda, "SDN-based network orchestration of variable-capacity optical packet switching network over programmable flexi-grid elastic optical path network," *J. Lightwave Technol.*, vol. 33, pp. 609–617, 2015.

[22] J. K. Fischer, "Bandwidth-variable transceivers based on four-dimensional modulation formats," *J. Lightwave Technol.*, vol. 32, pp. 2886–2895, 2014.

[23] B. Teipen, M. H. Eiselt, K. Grobe, and J.-P. Elbers, "Adaptive data rates for flexible transceivers in optical networks," *J. Netw.*, vol. 7, pp. 776–782, 2012.

[24] K. Roberts and C. Laperle, "Flexible transceivers," in *Proc. European Conf. on Optical Communications*, Amsterdam, The Netherlands, 2012, paper We.3.A.3.

[25] M. Kuschnerov, "Flexi-rate optical interfaces go mainstream," June 12, 2015 [Online]. Available: http://www.lightwaveonline.com/articles/2015/06/flexi-rate-optical-interfaces-go-mainstream.html.

[26] "Acacia Communications announces the industry's first coherent flex-rate 400G 5 × 7 transceiver model," Mar. 22, 2015 [Online]. Available: http://acacia-inc.com/acacia-communications-announces-the-industrys-first-coherent-flex-rate-400g-5x7-transceiver-module.

[27] X. Zhou, L. E. Nelson, P. Magill, R. Isaac, B. Zhu, D. W. Peckham, P. I. Borel, and K. Carlson, "High spectral efficiency 400 Gb/s transmission using PDM time-domain hybrid 32–64 QAM and training-assisted carrier recovery," *J. Lightwave Technol.*, vol. 31, pp. 999–1005, 2013.

[28] Q. Zhuge, X. Xu, M. Morsy-Osman, M. Chagnon, M. Qiu, and D. V. Plant, "Time domain hybrid QAM based rate-adaptive optical transmissions using high speed DACs," in *Optical Fiber Communication Conf.*, Anaheim, CA, 2013, paper OTh4E.6.

[29] H. Bulow, T. Rahman, F. Buchali, W. Idler, and W. Kuebart, "Transmission of 4-D modulation formats at 28-Gbaud," in *Optical Fiber Communication Conf.*, Anaheim, CA, 2013, paper JW2A.39.

[30] M. Karlsson and E. Agrell, "Spectrally efficient four-dimensional modulation," in *Optical Fiber Communication Conf. and the Nat. Fiber Optic Engineers Conf.*, Los Angeles, CA, 2012, paper OTu2C.1.

[31] L. Beygi, E. Agrell, J. M. Kahn, and M. Karlsson, "Coded modulation for fiber-optic networks," *IEEE Signal Process. Mag.*, vol. 31, no. 2, pp. 93–103, 2014.

[32] M. Ghobadi, J. Gaudette, R. Mahajan, A. Phanishayee, B. Klinkers, and D. Kilper, "Evaluation of elastic modulation gains in Microsoft's optical backbone in North America," in *Optical Fiber Communication Conf.*, Anaheim, CA, 2016, paper M2J.2.

[33] A. Carena, V. Curri, G. Bosco, P. Poggiolini, and F. Forghieri, "Modeling of the impact of nonlinear propagation effects in uncompensated coherent transmission links," *J. Lightwave Technol.*, vol. 30, pp. 1524–1539, 2012.

[34] OIF Implementation Agreement for FlexEthernet, document oif2015.127.

[35] S. Hardy, "Ethernet speed tuning goal of OIF FlexEthernet project," Feb. 12, 2015 [Online]. Available: http://www.lightwaveonline.com/articles/2015/02/ethernet-speed-tuning-goal-of-oif-flexethernet-project.html.

[36] M. Filer, H. Chaouch, J. Chu, R. Kankipati, and T. Issenhuth, "Transmission of Nyquist-shaped 32 GBaud PM-QPSK over a production flex-grid open line system," in *Optical Fiber Communication Conf.*, Anaheim, CA, 2016, paper W4G.3.

[37] OIF Implementation Agreement for CFP2-Analogue Coherent Optics Module, document oif2014.006.

[38] T. Duthel, P. Hermann, T. Winkler von Mohrenfels, J. Whiteaway, and T. Kupfer, "Challenges with pluggable optical modules for coherent optical communication systems," in *Optical Fiber Communication Conf.*, San Francisco, CA, 2014, paper W3K.2.

[39] C. Xia, W. Schairer, A. Striegler, L. Rapp, M. Kuschnerov, J. F. Pina, and D. van den Borne, "Impact of channel count and PMD on polarization-multiplexed QPSK transmission," *J. Lightwave Technol.*, vol. 29, pp. 3223–3229, 2011.

[40] http://nlinwizard.eng.tau.ac.il/, retrieved Mar. 18, 2016.

[41] L. E. Nelson, G. Zhang, M. Birk, C. Skolnick, R. Isaac, Y. Pan, C. Rasmussen, G. Pendock, and B. Mikkelsen, "A robust real-time 100G transceiver with soft-decision forward error correction [Invited]," *J. Opt. Commun. Netw.*, vol. 4, pp. B131–B141, 2012.

[42] K. Roberts, F. S. Heng, M. Moyer, M. Hubbard, A. Sinclair, J. Gaudette, and C. Laperle, "High capacity transport—100G and beyond," *J. Lightwave Technol.*, vol. 33, pp. 563–578, 2015.

[43] http://www.cisco.com/c/en/us/products/collateral/optical-networking/network-convergence-system-2000-series/datasheet-c78-733699.html, retrieved Jan. 15, 2016.

[44] S. Gringeri, N. Bitar, and T. J. Xia, "Extending software defined network principles to include optical transport," *IEEE Commun. Mag.*, vol. 51, no. 3, pp. 32–40, 2013.

[45] C.-Y. Hong, S. Kandula, R. Mahajan, M. Zhang, V. Gill, M. Nanduri, and R. Wattenhofer, "Achieving high utilization with software-driven WAN," in *Proc. SIGCOMM*, Hong Kong, China, 2013.

[46] P. Sun, R. Mahajan, J. Rexford, L. Yuan, M. Zhang, and A. Arefin, "A network-state management service," in *Proc. SIGCOMM*, Chicago, IL, 2014.

[47] A. Bartels, J. R. Rymer, J. Staten, K. Kark, J. Clark, and D. Whittaker, "The public cloud is now in hypergrowth," Forrester Research Report, Apr. 24, 2014.