JOURNAL OF
Optical Communications
and Networking

INTREPID program: technology and architecture for next-generation, energy-efficient, hyper-scale data centers [Invited]

Adel A. M. Saleh,[1,*] Katharine E. Schmidtke,[2] Robert J. Stone,[2]
James F. Buckwalter,[1] Larry A. Coldren,[1] and Clint L. Schow[1]

[1]Department of Electrical and Computer Engineering, University of California, Santa Barbara, Santa Barbara, California 93106, USA
[2]Facebook, Menlo Park, California 94025, USA
*Corresponding author: AdelSaleh@ece.ucsb.edu

Received 14 July 2021; revised 30 September 2021; accepted 3 October 2021; published 29 October 2021 (Doc. ID 437858)

The INTREPID program is developing power-efficient coherent optics for package-level integration with future switch integrated circuits as a path to realizing higher-radix switches for flatter networks. The link architecture is underpinned by coherent quadrature phase-shift keying (QPSK) polarization-multiplex transceivers at 200 Gb/s per λ, further enhanced with wavelength division multiplexing (WDM) to enable energy-efficient 800 or 1600 Gb/s inter-switch fiber connections. The technology is compatible with conventional three-level data center designs as well as a two-level data center design introduced here, which includes an added layer of passive, arrayed waveguide grating routers (AWGRs) or WDM circuit switches to further improve the cost, energy efficiency, and latency of the network.

https://doi.org/10.1364/JOCN.437858

1. MOTIVATION

Data centers have become a key component of the world's information infrastructure and play an ever-increasing role in storing, processing, and routing the data that we rely upon in our personal and professional lives. Indeed, the number of Internet users is projected to exceed 5 billion within the next several years [1]. Data center traffic is now measured in 10's of zetabytes (10^{21}), with intra-data-center traffic making up >70% of the total and increasing at a >23% compound annual growth rate (CAGR) [2]. Consequently, in order to improve overall data center productivity and efficiency, a key focus must be placed on maximizing the bandwidth and efficiency of intra-data-center communications. Today these interconnects, which are generally accepted to be limited to distances under 2 km, are implemented as concatenations of electrical-optical-electrical (E-O-E) interconnections that suffer degraded efficiency due to requiring multiple 50 Ω high-speed electrical interfaces to switching application-specific integrated circuits (ASICs). The electrical input/output (I/O) cells in these switching chips must be designed for worst-case electrical channels and therefore can consume of the order of 50% of the total power for the switches.

The INTREPID project, part of the Advanced Research Projects Agency–Energy (ARPA-E) ENergy-efficient Lightwave Integrated Technology Enabling Networks that Enhance Dataprocessing (ENLITENED) program, was launched in 2017 as a collaboration between University of California, Santa Barbara (UCSB) and Facebook [3] and is still ongoing. The project focus is twofold: (1) developing a technology platform to integrate efficient high-speed photonic interfaces directly into chip packages and (2) exploring new network architectures that incorporate photonic routing and/or switching that are made possible by the expanded optical link budgets enabled by analog coherent link architectures. The efficiency targets of the co-packaged optical interfaces are aggressive, scaling to sub-pJ/bit for multimode (MM) vertical-cavity surface-emitting laser (VCSEL)-based short-reach server links and to less than 10 pJ/bit for single-mode (SM) analog coherent data-center-scale interconnects. Achieving these targets will enable highly integrated solutions for the 102 Tb/s switch generation, and beyond that can potentially offer substantial expansions in switch radix with simultaneous improvements in efficiency compared to aggressive projections of conventional module-based transceiver technology. Such large, highly efficient switches can enable flatter networks with higher bandwidth to improve the overall efficiency of data centers of all scales.

Our focus is on the integration of photonic I/O with electrical switch cores since the network switches are the points of highest bandwidth concentration and where efficient photonic I/O can have the greatest impact. Fat-tree networks, schematically depicted in Fig. 1(a), are the workhorse topology for data center networks due to their superior performance and scaling properties that fundamentally depend on switch radix,

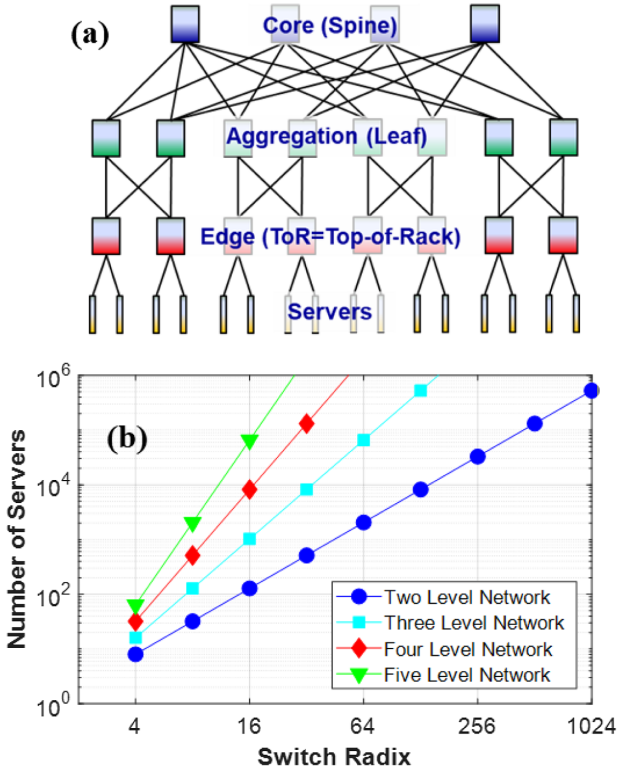

Fig. 1. (a) Illustration of a fat-tree network and (b) scaling properties: number of connected servers as a function of switch radix.

as shown in Fig. 1(b). The top of rack (ToR) switches require two types of interconnects: short distance (<3 m) for server connections and longer distance (<2 km) for connections to switches in the next level of the hierarchy. For the longer fabric links above the ToR, the use of a SM fiber is essentially a requirement due to its substantial advantages in operational management, cost, and support for bandwidth scaling through wavelength division multiplexing (WDM).

2. CO-PACKAGING FOR HIGHER RADIX SWITCHES

Our approach is conceptually illustrated in Fig. 2, representing what is often called "co-packaged optics (CPO)." Co-packaging has become a significant focus for the field over the last several years, and, an industry group led by Microsoft and Facebook, the CPO Collaboration, was recently launched with the goal of open development and broad commercial adoption of switch packages with integrated optical I/O [4]. By bringing all high-speed data on and off packages optically, instead of through conventional electrical interfaces that rely on ball grid array (BGA) or land grid array (LGA) connectors, the primary packaging bottleneck that limits the bandwidth and efficiency of today's systems is overcome. Integrating photonic interfaces into switch chip packages enables electrical connections at chip-scale pitch (e.g., C4 at ~130 μm) instead of package-scale pitch (e.g., BGA/LGA at ~1 mm). The electrical paths between the photonics and electronics are minimized, potentially enabling a >10× improvement in the efficiency of the ASIC to photonic I/O electrical links,

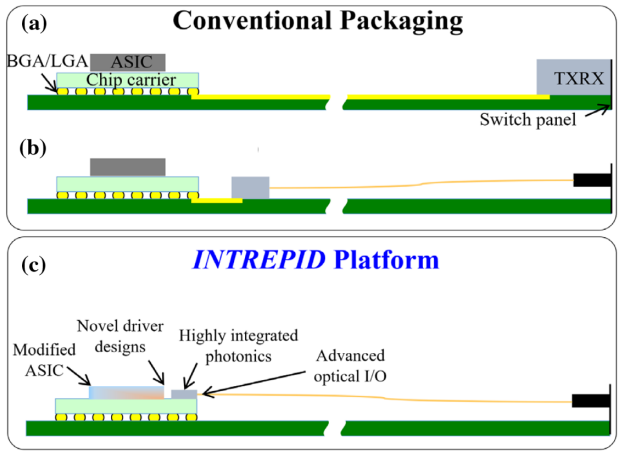

Fig. 2. Conceptual illustration of conventional optical module packaging in switches: (a) pluggable optics: pluggable transceiver (TXRX) modules, (b) on-board optical modules, and (c) optics in chip package: the integrated platform under development.

with a concurrent enhancement of bandwidth density of up to ~60×. General trends of electrical chip I/O show a direct dependence on channel loss, with 30 dB of channel loss (a typical target for general purpose electrical I/O) degrading efficiency by 10−20× [5]. Conversely, the short interconnects within a chip package have low channel loss and therefore can be designed for maximum efficiency. A 2 cm electrical interconnect demonstrated an efficiency of 1.4 pJ/bit, >14× more efficient compared to typical general purpose I/O cells that consume ~20 pJ/bit [6].

Figure 3 presents a conceptual view of the modular photonic integration platform we envision applied to a ToR switch, encompassing the co-design of interface circuitry to the digital switch core, the I/O bridge, electronic/photonic interposers, and SM and MM photonics with array fiber coupling.

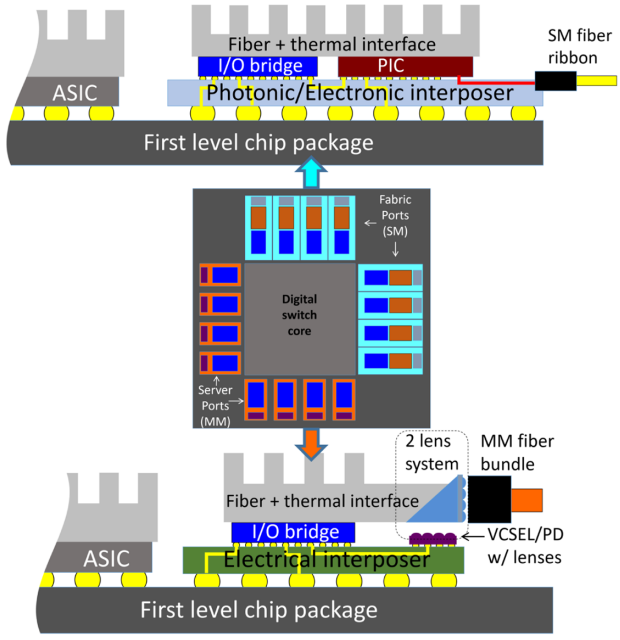

Fig. 3. Implementation concepts for integrating optics into first-level chip packages.

Switches for the higher levels of the network will integrate only SM photonics for the reasons discussed above.

3. ENERGY-EFFICIENT ANALOG COHERENT LINKS

For interconnects above the ToR/end of row (EoR) tier, we are developing low-cost, low-power coherent WDM photonic interconnects purpose-built for the longer fabric links required in the data center. Tailoring the transceiver to data center requirements requires optimization for a different set of metrics compared to current long-haul and metro coherent technology, specifically: (1) low-power consumption, (2) expanded link budgets, (3) low cost, and (4) low latency. Future scalability to higher data rates is possible through higher-order modulation formats, polarization modulation, and additional wavelengths. For data centers, the significantly larger link budget is a key advantage for coherent links, enabling reduced link power (lower required source laser power), lower cost (relaxed alignment tolerances/device specifications), and novel network architectures that incorporate all-optical routing/switching. Expansions of link budget of the order of 20 dB are possible [7], and our analysis shows that link budgets of 13 dB can be achieved with wall-plug link efficiencies better than 5 pJ/bit [8]. This level of tolerance to link loss allows for the incorporation of an arrayed waveguide grating router (AWGR) or active photonic switching layer without requiring complex and costly integrated optical gain in such components. Furthermore, the high selectivity offered by coherent reception significantly reduces the optical crosstalk requirements between channels for photonic routing/switching devices.

The INTREPID analog coherent links under development drastically reduce power and complexity compared to current digital coherent technology that relies heavily on digital signal processing (DSP) to compensate for chromatic dispersion (CD), polarization mode dispersion (PMD), and nonlinear effects in dense WDM (DWDM) links. The INTREPID links operate in the O-band near the zero-dispersion wavelength for standard SM fiber (1264–1338 nm), meaning CD and PMD will not have to be compensated, as they contribute negligible performance penalties for links up to 2 km. Furthermore, to eliminate the need for inefficient high-resolution analog to digital converters (ADCs) and DSP-based carrier recovery, we are developing optical phase locked loops (OPLLs) that lock and track the phase, frequency, and polarization of the receiver local oscillator (LO) to the incoming signal [9]. Highly integrated OPLLs have been demonstrated to enable robust and high-performance "analog coherent" receivers that operate with very low uncorrected bit error rate (BER, $<10-12$) and do not rely upon costly, high-power ADCs and DSPs [10,11]. The analog coherent receivers we are developing are optimized for power efficiency through photonic device and circuit co-design, choice of modulation format (quadrature phase-shift keying, QPSK), and close integration of electronics and photonics to minimize loop delays and maximize noise tolerance.

The baseline link architecture for the analog coherent links targets 200 Gbps/λ, achieved through QPSK modulation (2 bits per symbol) at 56 Gbd, with polarization multiplexing to achieve an additional factor of two in bandwidth per

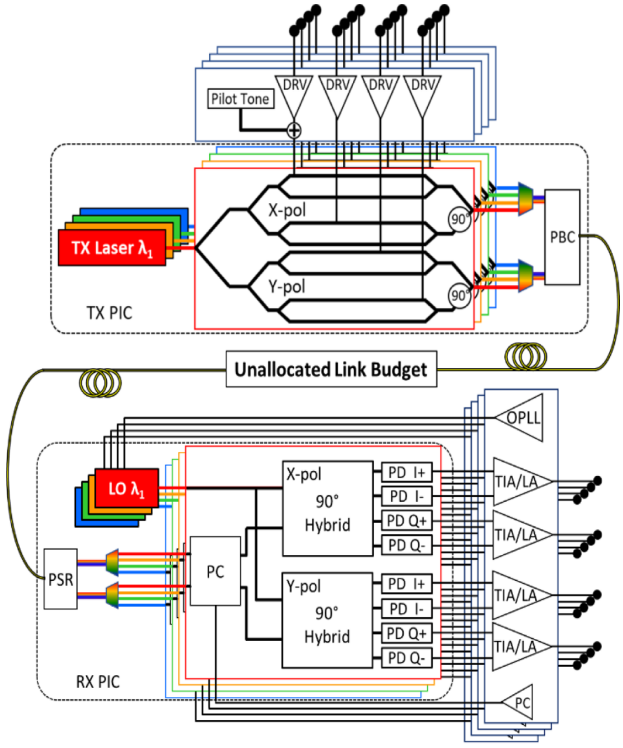

Fig. 4. High-level schematic of 4λ PM-QPSK analog coherent link architecture. DRV, modulator driver; LO, local oscillator; PBC, polarization beam combiner; TIA/LA, transimpedance amplifier/limiting amplifier; PSR, polarization splitter/rotator; PD I/PD Q, photodiode in/quadrature phase; PC, polarization controller; OPLL, optical phase locked loop controller.

wavelength. Scaling to higher bandwidths is supported by adding additional wavelength channels, between 4 and 8, to provide solutions for 800G and 1.6T links. A block diagram of a 4λ analog coherent link is presented in Fig. 4. A comprehensive simulation framework has been developed to model and optimize the energy efficiency of the link architecture. The model includes all of the required components, including driver and receiver plus OPLL circuitry, source and LO lasers, polarization multiplexing and control structures, optical 90° hybrids, and high-speed photodetectors. Details of the simulation framework can be found in [8], and we project that the analog coherent links can support link budgets of 13 dB operating at an uncorrected BER of 10^{-12}, with a wall-plug energy efficiency of ∼5 pJ/bit. Forward error correction (FEC), such as the common KR4-FEC (BER $< 2.1 \times 10^{-5}$), is widely used for data center links. The operating point of the analog coherent links can be tuned to achieve even better energy efficiency if FEC is utilized.

First-generation functional prototypes of the key hardware components have been demonstrated, including InP and SiP coherent receiver photonic integrated circuits (PICs), high-speed drivers and receivers, and transmitter PICs [12,13]. Full coherent receivers consisting of PICs integrated with electrical amplifier integrated circuits (ICs) have been demonstrated to operate at 80 Gb/s [14] and 100 Gb/s [15], with the latter result exhibiting an efficiency of <1 pJ/bit. Other notable results include monolithically integrated optical receivers

operating at 50 Gb/s [16,17], optical transmitters operating at 50 Gb/s [18], a novel architecture to implement feed-forward equalization in the optical or electrical domains [19], and a transimpedance amplifier achieving a record data rate of 108 Gb/s [20].

4. EFFICIENT VCSEL LINKS FOR SERVER CONNECTIONS

For server links, VCSEL technology provides a viable path to low-cost, short-distance links with sub-pJ/bit efficiency. VCSEL links have demonstrated the best wall-plug efficiency of any high-speed optical links [21] and have achieved data rates that were previously thought unattainable [22]. VCSEL links, implemented as active optical cables, are currently ubiquitous for 100G ToR-to-aggregation layer connections [23]. Due to their simplicity, efficiency, and low cost, VCSEL links have the potential to displace copper interconnections within the rack between servers and ToR switches. The VCSEL links we are developing can serve these <3 m applications and can also support migration from ToR to EoR switches (<30 m) if demanded by the evolution of data center architectures (see Section 5).

The majority of the efforts for hardware development in the INTREPID program are devoted to developing low-power coherent links, as these are a missing piece of technology that has not been demonstrated and that can make a significant impact on data center networks. VCSEL links, on the other hand, are continuing to be developed and advanced by multiple groups and companies, including another program funded under ARPA-E ENLITENED, MOTION, led by IBM [24]. The VCSEL links developed under INTREPID utilize proven equalization techniques [25] to realize single-pJ/bit full-link efficiencies at data rates of 50 Gb/s and above [13]. Novel implementations of transmitter equalizers have yielded low-power operation at high data rates, including a <3 pJ/bit VCSEL driver operating up to 52 Gb/s [26] and a full optical link operating up to 50 Gb/s at an efficiency of 9.5 pJ/bit [27].

5. DATA CENTER NETWORK ARCHITECTURE

A. Alternative Data Center Designs

As mentioned earlier, one of the primary goals of the INTREPID project is to utilize CPO to enable the use of large electronic packet switches to flatten the data center (i.e., to reduce the number of switching levels for a given number of servers). The largest switch size available to date has a throughput of 25.6 Tb/s [28], which is expected to double about every two years. Thus, we will target future switch sizes of 51.2 and 102.4 Tb/s. Furthermore, in our design examples, we will consider hyper-scale data centers supporting of the order of hundreds of thousands of servers, each of a bit rate of 50 or 100 Gb/s.

1. *Conventional, Three-Level Data Center Design (Design Type 1)*

One can achieve the above objectives using a traditional three-level folded-Clos (fat-tree) data center design, which is depicted in Fig. 5. This design, which we will refer to as *Design Type 1*, employs large electronic switches in both the top switching level (spine) and the intermediate switching-level (aggregation) layers and smaller ToR switches in the bottom level. Each ToR switch supports one rack of servers, as indicated in the figure. The *module* shown in the figure, which is often called *server pod* in the literature, represents a grouping of switches and server racks that repeats across the data center and is connected to each of the spine switches.

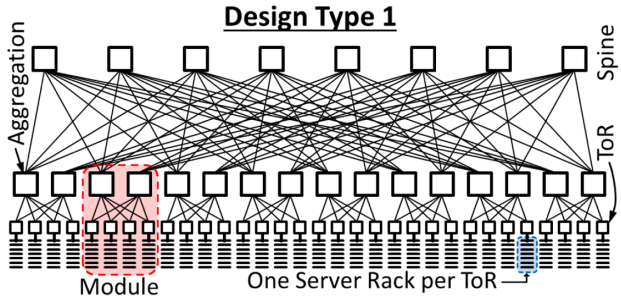

Fig. 5. Conventional, three-level, folded-Clos (fat-tree) data center utilizing large spine and aggregation switches of the same size and smaller ToR switches.

Let T be the throughput of a spine or an aggregation switch, which will be referred to as the *large switch*, and let τ be the throughput of a ToR switch. Let R be the bit rate per inter-switch fiber link, which is assumed to employ SM fibers. The ToR switches are connected from below to servers via links of a bit rate of σ per server. The radix (i.e., the number of fiber ports) of a packaged large switch is $N = T/R$. Moreover, one can define the *effective* radix of a ToR switch as $M = \tau/R$.

Each spine switch has all of its N fiber ports (each at a bit rate of R) directed downward, and (assuming no *over-subscription*, which will be considered in Section 5.B) each aggregation switch has $N/2$ of its fiber ports directed upward and $N/2$ directed downward. Moreover, each ToR switch has $M/2$ fiber ports directed upward (each at a bit rate of R), and $(M/2) \times (R/\sigma)$ ports (each at a bit rate of σ) directed downward to the servers. (Note that the reason we call M the *effective* radix of a ToR switch is because this would have been the radix if all of its ports were of the same bit rate of R.)

It follows that Design Type 1 has N modules, each containing $N/2$ ToR switches, each supporting one rack of $(M/2) \times (R/\sigma)$ servers, for a total number of servers of $z_1 = (MN^2/4) \times (R/\sigma)$, which can also be written as $z_1 = \tau T^2/(4R^2\sigma)$.

Plots of the number of servers versus the bit rate, R, per inter-switch fiber link (or, equivalently, per integrated switch port) are given in Figs. 6 and 7 for various values of the ToR switch size, τ. The two figures, respectively, correspond to large switch sizes of $T = 51.2$ and 102.4 Tb/s and for server bit rates of $\sigma = 50$ and 100 Gb/s. In each figure, the scale at the top represents the radix, N, of the corresponding integrated switch. The three design points represented by the triangle, circle, and square in each of these figures will be used later for comparisons with other designs.

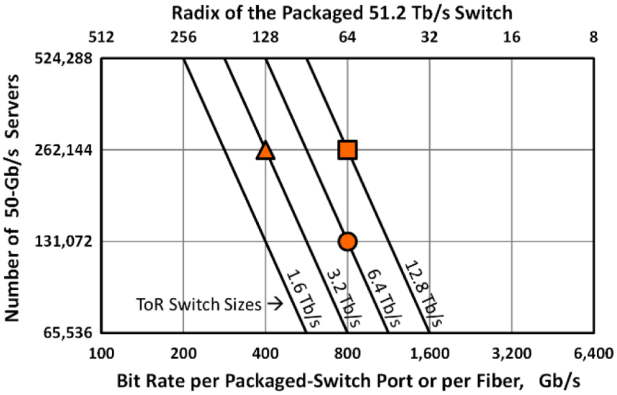

Fig. 6. Number of 50 Gb/s servers versus bit rate per inter-switch fiber link for data center Design Type 1 with a large switch size of 51.2 Tb/s for various ToR switch sizes.

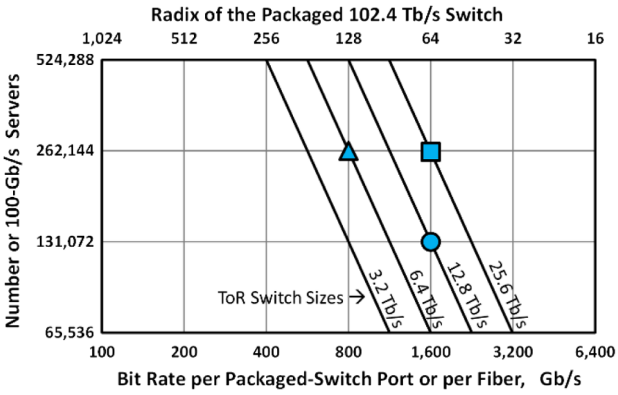

Fig. 7. Number of 100 Gb/s servers versus bit rate per inter-switch fiber link for data center Design Type 1 with a large switch size of 102.4 Tb/s for various ToR switch sizes.

2. EoR-Based, Two-Level Data Center Design (Design Type 1.5)

One can further flatten the data center by reducing the number of electronic switching levels from three to only two. To support the same number of servers, however, this requires both (1) using large switches for both the top and bottom switching layers and (2) reducing the bit rate of the inter-switch fiber links, thus increasing the switch radix. This design, which will be referred to as *Design Type 1.5*, is depicted in Fig. 8. As shown in the figure, the large bottom switches are called EoR switches. Because of its large size (compared to a ToR switch), each EoR switch supports a row of multiple racks of servers, not just one. Note that a module in Design Type 1.5 consists of only one EoR switch and its associated multiple server racks.

Let the throughput of each of the spine and EoR switches be T (which is the same as the throughput of the large switches in Design Type 1), and let σ be the bit rate per server. Furthermore, let R' be the bit rate of the inter-switch fiber links, and let $N' = T/R'$ be the corresponding radix of a spine switch, which is the same as the effective radix of an EoR switch. In general, $R' \ll R$ and $N' \gg N$, but this will be made more precise shortly.

Following the same analysis used in the above design, and assuming no oversubscription, one can show that

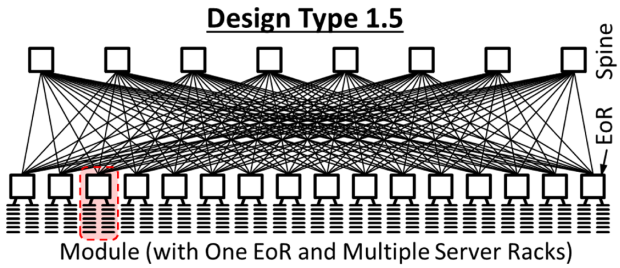

Fig. 8. Flat data center design with only two levels of large electronic switches of the same size.

Design Type 1.5 has N' EoR switches, each supporting $(N'/2) \times (R'/\sigma)$ servers (placed in multiple racks), for a total number of servers of $z_{1.5} = (N'^2/2) \times (R'/\sigma)$, which can also be written as $z_{1.5} = T^2/2R'\sigma$.

Using the above formulas and noting that $T = NR = N'R'$, one can show that Design Types 1 and 1.5 will have the same number of servers, i.e., $z_1 = z_{1.5}$, if $R' = R/(M/2)$ and, thus, $N' = N \times (M/2)$, where, as mentioned before, $M = \tau/R$ is the effective radix of a ToR switch in Design Type 1. One can get some numerical examples by considering the three design points of Design Type 1 represented by the shaded shapes in Figs. 6 and 7. The corresponding values of the effective ToR radixes are given by $M = 8$ for the circle and triangle design points and $M = 16$ for the square design point.

Thus, for Design Type 1.5 to have the same number of servers as Design Type 1, it follows that Design Type 1.5 requires $M/2 =$ four or eight times the number of fibers required in Design Type 1. This makes Design Type 1.5 not desirable from a practical point of view. On the other hand, because of its flat, two-level design, it has the advantage of lower cost, latency, and energy consumption because of the elimination of an electronic switching level and its associated transceivers.

3. EoR/AWGR-Based, Two-Level Data Center Design (Design Type 2)

We now introduce a novel data center design, which will be referred to as *Design Type 2*, that retains all the performance advantages of Design Type 1.5, while requiring the same number of interconnect fibers as in Design Type 1. The new design is compatible with the INTREPID transceiver technology of the inter-switch links. As mentioned in Section 3, this technology is based on the use of WDM and polarization-multiplexed, analog QPSK modulation with coherent reception (which results in $\mu = 4$ bits per symbol). Our current goal is to have $\nu = 4$ wavelengths per fiber with a modulation symbol rate of $\rho = 50$ GBaud. In this case, the bit rate per wavelength is $r = \mu\rho = 200$ Gb/s, and the bit rate per fiber is $R = \nu r = 800$ Gb/s. In the future, we plan to double R to 1600 Gb/s, without changing the modulation format, by either doubling ρ to 100 GBaud (which results in $r = 400$ Gb/s per wavelength) or doubling ν to eight wavelengths per fiber.

Besides being highly energy-efficient, this modulation/reception technique also yields a link budget of more than 10 dB [8]. This large link budget, combined with WDM, enables the realization of the novel data center architecture of

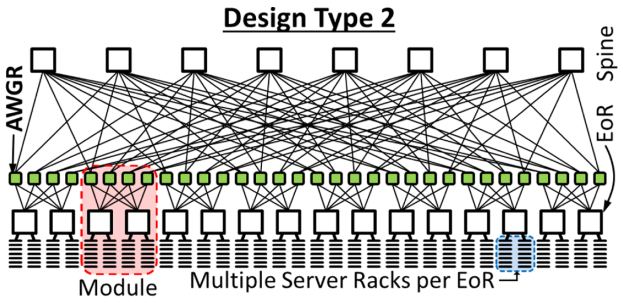

Fig. 9. Novel data center architecture utilizing two levels of large (spine and EoR) electronic switches of the same size, interconnected with WDM fibers, with an added layer of AWGRs (small, shaded boxes).

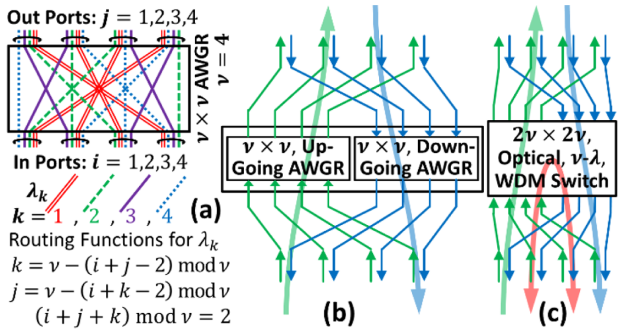

Fig. 10. (a) Routing pattern of a $\nu \times \nu$ AWGR. (b) Connecting a pair of AWGRs to handle the up- and down-going traffic. (c) A $2\nu \times 2\nu$ WDM circuit switch may replace each AWGR pair in the future to provide wavelength-level circuit switching flexibility (see Section 5.D.2).

Design Type 2, which is depicted in Fig. 9. As shown in the figure, this is a folded-Clos architecture with only two electronic switching levels (spine and EoR), with all switches having the same large size (as was the case in Design Type 1.5) and with a layer of passive, $\nu \times \nu$ AWGRs inserted in the fiber links between the two switching levels. (This architecture resembles that described in [29].) The bit rate per fiber in this design is the same as that used in Design Type 1. Moreover, the same WDM-based transceivers are employed here. The function of the AWGRs is to statically demultiplex, shuffle, then re-multiplex the wavelengths in the various fibers such that the electronic switches in the two switching levels surrounding the AWGRs will be inter-connected in a folded-Clos pattern at a *single-wavelength* level. For example, for a fiber bit rate of $R = 800$ Gb/s, with $\nu = 4$ wavelengths per fiber, the bit rate of each connection becomes $r = R/\nu = 200$ Gb/s. In effect, this design has identical connectivity and, hence, also identical performance advantages as Design Type 1.5 (with $R' = r$), while having the same number of fibers as in Design Type 1. Also, as in Design Type 1.5, each EoR switch supports a row of multiple racks of servers, not just one.

Note that Design Type 2 does not require tunable transceivers.

Each shaded box in Fig. 9 actually consists of a pair of AWGRs, one for the up-going traffic and the other for the down-going traffic. Figure 10(a) shows the wavelength routing pattern of a typical AWGR, and Fig. 10(b) shows how a pair of AWGRs is to be connected to the up-going and down-going fibers.

Let T be the throughput of each of the spine and the EoR switches, R be the bit rate of the inter-switch fiber links, ν be the number of wavelengths per fiber, $r = R/\nu$ be the bit rate per wavelength, σ be the bit rate per server, and $N = T/R$ be the radix of a spine switch, which is the same as the *effective* radix of an EoR switch.

Each spine switch has all of its N fiber ports directed downward. Each AWGR has ν fiber ports directed upward and ν fiber ports directed downward. Moreover (assuming no oversubscription), each EoR switch has $N/2$ fiber ports directed upward and $(N/2) \times (R/\sigma)$ ports (each at a bit rate of σ) directed downward to the servers.

It follows that Design Type 2 has N modules, each containing ν EoR switches, each supporting $(N/2) \times (R/\sigma)$ servers (placed in multiple racks), for a total number of servers of $z_2 = \nu(N^2/2) \times (R/\sigma)$, which can also be written as $z_2 = T^2/2r\sigma$, which is independent of R or ν. For Design Types 1 and 2 to have the same number of servers, i.e., $z_1 = z_2$, one must have $\nu = M/2$, where M is the effective radix of a ToR switch in Design Type 1. This condition can also be written as $2\nu R = \tau$; i.e., the throughput of an AWGR pair in Design Type 2 is equal to the throughput of a ToR switch in Design Type 1.

Plots of the number of servers versus the bit rate, R, per inter-switch fiber link (or, equivalently, per integrated switch port) are given in Figs. 11 and 12 for various values of the number of wavelengths per fiber, ν. The two figures, respectively, correspond to switch sizes of $T = 51.2$ and 102.4 Tb/s and for server bit rates of $\sigma = 50$ and 100 Gb/s. In each figure, the scale on the right shows the bit rate per wavelength, r, and the top scale represents the radix, N, of the corresponding integrated switch. The three design points in each of these figures represented by the triangle, circle, and square give the same data center designs (in terms of the number of servers and the bit rate per fiber) as those given in Figs. 6 and 7, respectively, for Design Type 1. More specifically, in each of the four figures, the circle design point corresponds to ($N = 64$, $M = 8$, $\nu = 4$, and $z = 131{,}072$ servers), the triangle design point corresponds to ($N = 128$, $M = 8$, $\nu = 4$, and $z = 262{,}144$ servers), and the square design point corresponds to ($N = 64$, $M = 16$, $\nu = 8$, and $z = 262{,}144$ servers).

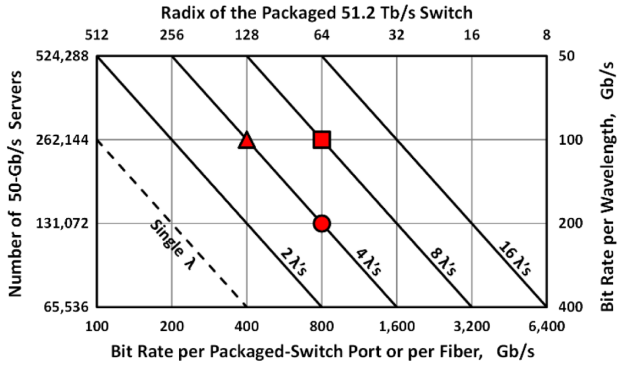

Fig. 11. Number of 50 Gb/s servers versus bit rate per inter-switch fiber link for Design Type 2 with a switch size of 51.2 Tb/s and for various values of the number of λ's per fiber.

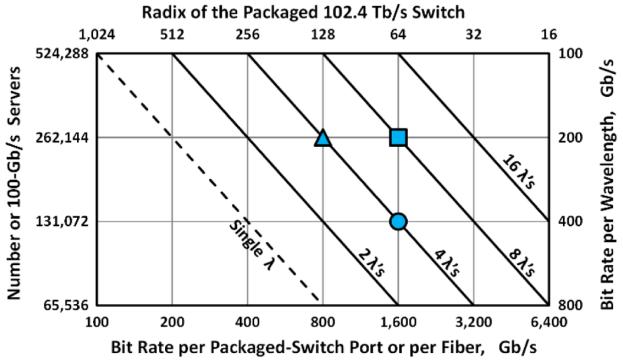

Fig. 12. Number of 100 Gb/s servers versus bit rate per inter-switch fiber link for Design Type 2 with a switch size of 102.4 Tb/s and for various values of the number of λ's per fiber.

Each of the dashed lines in Figs. 11 and 12 represents the case of a single-wavelength design, i.e., $\nu = 1$. Mathematically, this represents the limiting case of having Design Type 2 with a single wavelength and 1×1 AWGRs, which implies that there are no AWGRs. In this case, Design Type 2 reduces to Design Type 1.5 with $R' = r$ and $N' = T/r$.

B. Comparing the Traditional, Three-Level Data Center Design Type 1 and the EoR/AWGR-Based, Two-Level Data Center Design Type 2

1. Oversubscription Ratio

Before comparing the two designs, we will generalize the results for arbitrary values of the *oversubscription ratio*, which we will denote by Ω. This is defined for a given electronic switching layer as the ratio of the bandwidth below the layer to that above the layer. In general, $\Omega \geq 1$. As is often done in practice, we will assume that oversubscription occurs only in the bottom switching layer, i.e., the ToR switches in Design Type 1 and the EoR switches in Design Type 2. Designs with $\Omega = 1$, which have been considered so far, imply that the ports of each switch are arranged such that the bandwidth above and below the switch are equal. This has the advantage of eliminating packet blocking, thus improving the latency. On the other hand, designs with $\Omega > 1$ imply that the bandwidth above the switch is smaller than that below the switch. This can result in an appreciable decrease in the number of the switches required in the interconnect network, as well as a corresponding increase in the number of supported servers. However, it can lead to some packet blocking. The resulting latency penalty may not be a problem if the servers in the data center are not fully utilized. While this is generally not desirable, it is often the case in practice because of unpredictable workload variations.

2. Summary of Design Formulas

Table 1 summarizes the design formulas for data center Design Types 1 and 2 as a function of Ω. An implicit assumption in the table is that the bit rate per fiber, R, which is the same for Design Types 1 and 2, is generated by ν wavelengths, each of bit rate $r = R/\nu$, as would be the case when using the INTREPID WDM transceiver technology. (It should be noted that, in general, since Design Type 1 is based on using direct point-to-point fiber links between corresponding switch ports, any other modulation format, WDM or not, can be used as long as the total bit rate per fiber is R.)

Table 1. Design Formulas for Data Center Design Types 1 and 2

Design	Inter-Switch-Fiber Bit Rate	Electronic Switch Type (or AWGR)	Throughput	Radix[a] of Switch (or AWGR)	Number of Up Ports	Number of Down Ports[b]	Number of Switches (or AWGRs)	Total Number of Inter-Switch Transceivers	Total Number of Servers[c]
Design Type 1	R	Spine	T	$N = T/R$	0	N	$\frac{MN/2}{1+\Omega}$	$\frac{2MN^2}{1+\Omega}$	$(MN^2/2) \times \left(\frac{R}{\sigma}\right) \times \left(\frac{\Omega}{1+\Omega}\right)$
		Aggregation			$N/2$	$N/2$	$\frac{MN}{1+\Omega}$		
		ToR	τ	$M = \tau/R$	$\frac{M}{1+\Omega}$	$\frac{(R/\sigma)M\Omega}{1+\Omega}$	$N^2/2$		
Design Type 2	$R = \nu r$	Spine	T	$N = T/R$	0	N	$\frac{\nu N}{1+\Omega}$	$\frac{2\nu N^2}{1+\Omega}$	$\nu N^2 \times \left(\frac{R}{\sigma}\right) \times \left(\frac{\Omega}{1+\Omega}\right)$
		EoR			$\frac{N}{1+\Omega}$	$\frac{(R/\sigma)N\Omega}{1+\Omega}$	νN		
		AWGR	$2\nu R$	2ν	ν	ν	$\frac{N^2}{1+\Omega}$		

Defining the Rest of the Variables: Number of λ per Fiber $= \nu$, Bit rate per $\lambda = r$, Oversubscription Ratio $= \Omega \geq 1$, Bit Rate per Server $= \sigma$

note: [a] For ToR and EoR switches, this represents an *effective* switch radix. [b] For ToR and EoR switches, this is the number of server ports. [c] When $\nu = M/2$, the two designs will have the same number of servers.

3. Numerical Comparisons between Design Types 1 and 2

Numerical comparisons between Design Types 1 and 2 are given in Table 2 for the design scenarios represented by the circle, triangle, and square design points in Figs. 6, 7, 11, and 12. The results are presented for both 51.2 and 102.4 Tb/s switch sizes and for two representative values of the oversubscription ration, namely, $\Omega = 1{:}1$ and $\Omega = 3{:}1$. (Design Type 1.5 is not included in the comparisons because, as mentioned above, it does not represent a practical design on its own since it requires a large number of fibers.)

In all design scenarios, note that the large number of ToR switches required in Design Type 1 are eliminated in Design

Table 2. Comparing Different Scenarios of Design Types 1 and 2

Switch Size		Design Parameters	Design Point in Figs. 6, 7, 11, and 12					
			Circle		Triangle		Square	
T = 51.2 Tb/s		ToR Switch Size, τ, Tb/s	6.4		3.2		12.8	
		Server Bit Rate, σ, Gb/s	50		50		50	
		Bit Rate per λ, r, Gb/s	200		100		100	
		Fiber Bit Rate, R, Gb/s	800		400		800	
T = 102.4 Tb/s		ToR Switch Size, τ, Tb/s	12.8		6.4		25.6	
		Server Bit Rate, σ, Gb/s	100		100		100	
		Bit Rate per λ, r, Gb/s	400		200		200	
		Fiber Bit Rate, R, Gb/s	1,600		800		1,600	
Both Switch Sizes		# of λs per Fiber, ν	4		4		8	
		Radix of Large Switch, N	64		128		64	
		Radix of ToR Switch, M	8	--	8	--	16	--
Data Center Design Type →			1	2	1	2	1	2
Oversubscription Ratio, Ω	1:1	# of Transceivers	32,768	16,384	131,072	65,536	65,536	32,768
		# of ToR Switches	2,048	--	8,192	--	2,048	--
		# of Large Switches	384	384	768	768	768	768
		# of $\nu \times \nu$ AWGRs	--	2,048	--	8,192	--	2,048
		# of Servers	**131,072**		**262,144**		**262,144**	
	3:1	# of Transceivers	16,384	8,192	65,536	32,768	32,768	16,384
		# of ToR Switches	2,048	--	8,192	--	2,048	--
		# of Large Switches	192	320	384	640	384	640
		# of $\nu \times \nu$ AWGRs	--	1,024	--	4,096	--	1,024
		# of Servers	**196,608**		**393,216**		**393,216**	

Type 2 and replaced by an equal or smaller number of AWGRs. This is quite advantageous, since the cost of an AWGR is much less than that of a ToR switch. Moreover, the AWGRs eliminate the latency and power consumption associated with the ToR switches.

Note also the dramatic 50% reduction in the number of required transceivers in Design Type 2 compared to that in Design Type 1, which results in further reduction of cost and power consumption.

In all cases for $\Omega = 1{:}1$, the total required number of large switches is the same for Design Types 1 and 2. Thus, this does not affect the comparison. On the other hand, for $\Omega = 3{:}1$, the total required number of large switches in Design Type 1 is 60% of that required in Design Type 2. This will result in increased cost and power consumption for Design Type 2 associated with this part of the interconnect network. However, this increase will be more than offset by the corresponding elimination of the ToR switches and the dramatic reduction in the number of transceivers.

4. Considerations for ToR- versus EoR-Based Architectures

There are important differences between ToR-based designs, e.g., the conventional Design Type 1, and EoR-based designs, e.g., Design Type 2. (Additional considerations for using AWGRs in Design Type 2 will be discussed in Section 5.D.1.) Because of its relatively small size, a ToR switch supports only one rack of servers. Thus, the length of the links between the ToR switch and any server within its rack is of the order of a meter. Thus, these links have typically been copper-based. On the other hand, an EoR switch supports multiple racks of servers. In this case, the length of the links between the EoR switch and its servers can be of the order of several meters. Thus, these links should be realized using fiber-optic technology, e.g., based on VCSELs and MM fibers as described in Section 4.

Another important consideration is that a ToR switch failure will disable only one rack of servers, which is tolerable, while an EoR switch failure will disable multiple racks of servers, which might not be acceptable. A good way to mitigate this, which is depicted in Fig. 13, is to use double redundancy by homing each server to two different EoR switches within the same module. Ideally, both connections would be used during the non-failed state to provide high-bandwidth server connectivity. Then, upon failure of one EoR switch, the servers connected to it would still be connected to the rest of the system at half the bit rate, instead of being totally disconnected, as would be the case with no redundancy. A totally different option is to implement a one-for-N protection scheme, which would require the use of optical protection switches and somewhat longer fiber runs between the EoR switches and the servers. To reduce the number of fibers and to enable the longer fiber runs in this case, one can use coarse WDM (CWDM) and SM fibers in the links to the server racks instead of using VCSELs and MM fibers.

C. Disaggregated Data Centers

In a legacy data center, each server has its own storage, memory, processor, accelerator, etc. Depending on the overall workload,

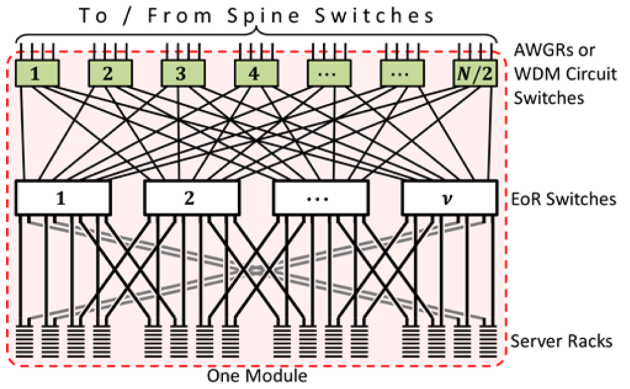

Fig. 13. Illustration of server double-homing redundancy to protect against a failure of an EoR switch within a module in Design Type 2.

these resources may not be used efficiently, i.e., they may be oversubscribed in one server while underutilized in another. The concept of a disaggregated data center involves placing a large part of these resources in a common location in the data center outside the servers, then sharing them among the servers [30]. The sharing results in increased utilization of the resources, leading to savings in cost and energy. But, to avoid bottlenecks between the servers and the shared resources, the connectivity between them needs to have high bandwidth and low latency [31]. Figure 14 shows a disaggregated data center based on Design Type 1, where some of the original server modules are replaced by various shared resources. A high-performance computing (HPC) cluster is included among the shared resources, which provides computationally intense functions such as artificial intelligence and machine learning (AI/ML). One of the desirable characteristics of this disaggregated design is that the EoR switches run across the entire data center, providing uniform interfaces to the servers and to the shared resources. Note that the path from a server to the shared resources involves passing through five electrical packet switches, and similarly in the reverse direction, which is likely to introduce a level of latency that may be too high for some applications.

To reduce latency and increase the bandwidth, various interesting, disaggregated data center architectures utilizing optical switches have been proposed in [32,33]. Here, we present a

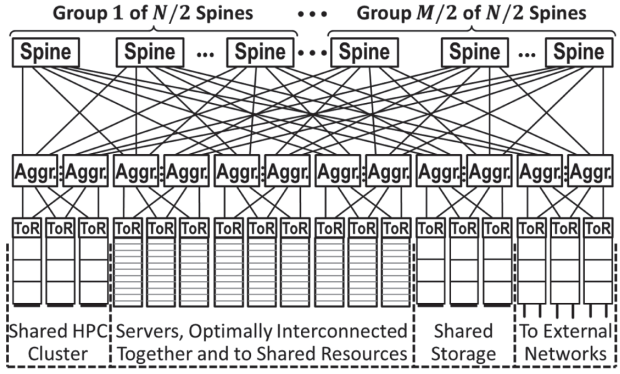

Fig. 14. Conventional disaggregated data center architecture based on Design Type 1.

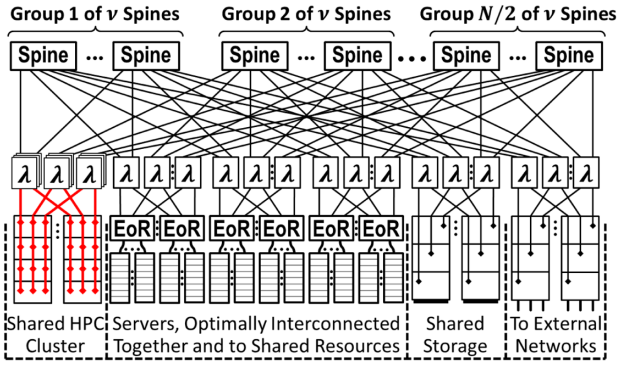

Fig. 15. Disaggregated data center based on Design Type 2. The λ boxes represent optical wavelength routing devices (AWGRs) or optical WDM circuit switches.

modified, scalable, low-latency architecture that is based on our Design Type 2. The new architecture is shown in Fig. 15.

Each of the λ boxes shown in Fig. 15 represents either a pair of $\nu \times \nu$ optical routing devices (AWGRs) or an optical $2\nu \times 2\nu$ WDM circuit switch (see Fig. 10). Note that the λ boxes run across the entire data center and provide uniform all-optical interfaces to the shared resources and to the EoR switches supporting the servers. The path from a server to the shared resources here passes through only two electrical packet switches, which will clearly reduce latency in comparison to the architecture of Fig. 14, where such a path passes through five electrical switches.

Considerations for and comparisons between using AWGRs or WDM circuit switches in data center architectures based on Design Type 2 are discussed next.

D. Contrasting the Use of AWGRs versus WDM Circuit Switches

As indicated by the results of Section 5.B.3, the enhanced economy and reduced power consumption of our novel data center Design Type 2 over those of the conventional Design Type 1 stem from including the layer of passive optical wavelength routing devices (AWGRs). Moreover, as mentioned earlier, Design Type 2 can be enhanced further by replacing the $\nu \times \nu$ AWGRs by $2\nu \times 2\nu$ WDM circuit switches (see Fig. 10). Here, we discuss various considerations for using these types of devices and contrast the differences between them.

1. Using AWGRs

Because of their wavelength dependence, there needs to be reasonable wavelength registration across each group of 2ν transceivers at the ports of the electronic switches connected to each of these devices. The degree of wavelength registration needs only to be sufficient for the various wavelengths to pass through the passbands of the AWGR. Generating wavelengths with this required stability across a large temperature operating range using uncooled lasers has been demonstrated [34]. This also requires the AWGRs to be reasonably athermal.

Because we use polarization multiplexing, another important requirement of the AWGRs is that they need to be polarization independent, or at least to have small polarization-dependent loss, so that they are compatible with the polarization control scheme that we are using [9]. Photonic fabrication technologies have been developed that are capable of achieving the above AWGR requirements of being athermal, polarization independent, and to have flat passbands (e.g., see [35]).

Because the AWGRs are passive devices, and because the associated transceivers are not tunable, Design Type 2 represents a conventional packet-switched network with *fixed* inter-switch connections (same as in Design Types 1 and 1.5). Hence, this AWGR-based design requires no changes in the underlying IP and/or Ethernet protocols.

2. Using WDM Circuit Switches

Using optical circuit switches to enhance the performance of data centers has been widely reported in the literature [36–43]. Here, we focus on architectures related to replacing the AWGRs in our Design Type 2 by WDM circuit switches in a similar way as initially suggested in [29,39,43]. The mode of operation that we are envisioning for the WDM circuit switches is to reconfigure them in a quasi-static regime (e.g., in seconds or longer) to respond to slowly changing types of workloads or computational scenarios. In this case, the added latency introduced by the reconfiguration process of the WDM switches will, on average, be negligible in comparison to other latencies in the system. Thus, in effect, since the path of a signal through a WDM switch is all-optical, its latency will be comparable to that of a static AWGR. But, the ability to reconfigure the WDM switches to match slow changes in the workload will increase utilization of the servers, thus reducing the system-wide latency and increasing the overall energy efficiency.

Note that the existing underlying Ethernet and IP protocols need not be changed, since the data center network reconfigurability that we require is sufficiently slow. On the other hand, if one wants to extend the vision to have the WDM switches respond to fast workload variations (e.g., in sub-milliseconds), then whole new protocols would be needed. This would be quite a challenging and costly task, which we are not currently considering.

Another important advantage of using the WDM switches, instead of the AWGRs, is that they can accommodate the variety of bandwidth requirements of the different types of end devices connected to a disaggregated data center. For example, consider the architecture of Fig. 15 with each of the fiber links having $\nu = 4$ wavelengths, with $r = 200$ Gb/s per wavelength, for a total of $R = 800$ Gb/s per fiber. If the λ boxes in the figure are AWGRs, the connectivity of all end devices would be at 200 Gb/s per port. If WDM switches are used instead, the connectivity of the end devices can be provisioned to achieve 200, 400, or 800 Gb/s per port. The high-bandwidth connectivity is quite desirable, especially in the latency-sensitive connections to the shared HPC cluster.

We are conducting various investigations under a different program on suitable types of WDM switches based on microring resonators [44–46]. Another promising type of WDM switches, which is based on microelectromechanical systems (MEMS) technology, has also been reported in

the literature [47]. All of these WDM switches, as well as other types reported in the literature, are not polarization-independent because they have a significant amount of polarization-dependent loss. Work is in progress on reducing the polarization dependence of these switches and, at the same time, on modifying the polarization control scheme that we are currently employing [9] to accommodate any residual polarization-dependent loss. Other schemes involving polarization diversity are also being considered.

3. Comparing Packet Latencies

As a demonstration of the difference in packet latency performance among various versions of Design Type 1 and Design Type 2 (with AWGRs or with WDM switches), consider a data center performing a computational task that involves multiple servers that are distributed across the data center. A reasonable measure of the packet latency performance is the estimated average number of electronic packet switches (which we will denote by $\bar{S}$) that need to be traversed to perform the computation as a function of the extent across the data center of the servers involved in the computation. (Note that not all of the servers in that range are necessarily involved in the computation, just a subset of them). The less $\bar{S}$ is, the less the expected latency is, and the better the computational performance is. Figure 16 shows plots representing this scenario for various types of data center designs.

For example, as shown in the figure for Design Type 1, if the servers involved in the computation are all within one rack, $\bar{S}$ would be exactly one (the ToR switch supporting the rack). If the two ends of the servers involved extend beyond one rack, but still within one module, then some of the interconnections among them need to go through an aggregation switch. Thus, $\bar{S}$ will increase towards three (two ToR switches and one aggregation switch). If the servers extend beyond one module, the spine switches will have to get involved, and $\bar{S}$ will increase toward five (two ToR switches, two aggregation switches, and one spine switch).

As shown in Fig. 16, in all cases of Design Type 2, if the servers involved are within an EoR domain, $\bar{S}$ would be exactly one (the EoR switch). This is of course a great improvement over Design Type 1 in that range, since an EoR switch covers multiple racks of servers, while a ToR switch covers just one rack.

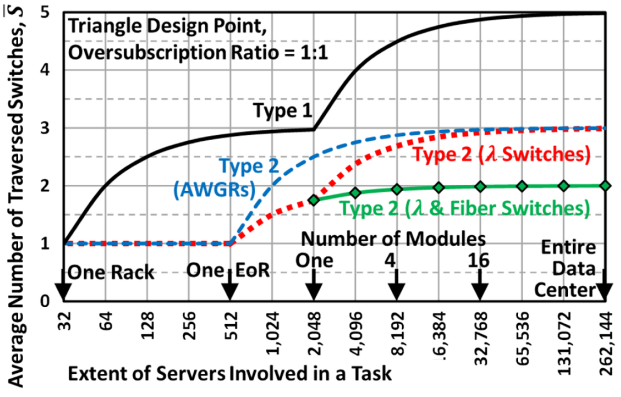

Fig. 16. Average number, $\bar{S}$, of traversed electronic packet switches to perform a computational task involving multiple servers versus the extent of the involved servers across the data center.

In the AWGR case, if the servers extend beyond the domain of one EoR, then the spine switches need to be involved, and $\bar{S}$ will increase toward three (two EoR switches and one spine switch), as the servers involved extend toward the whole data center. Note that the AWGRs cannot directly interconnect two EoRs, even within one module. On the other hand, if WDM circuit switches are used, then, as indicated in Fig. 10, they can loop back the signal and directly interconnect multiple EoRs within a module. But, beyond one module, the spine switches need to be involved, and $\bar{S}$ will increase toward three (two EoRs and one spine).

One can improve the performance further in the range beyond one module by enhancing the architecture by adding an array of fiber switches (e.g., MEMS switches) in a new layer between the WDM switches and spine switches (which is not shown in the figures). In this case, EoR switches can be directly interconnected across multiple modules. Thus, $\bar{S}$ would increase towards only two beyond one module. If we need to interconnect Q modules, each of the fiber switches needs to be connected to just one fiber port of each module, i.e., the fiber switch size would be $2Q \times 2Q$, and the number of such switches would be equal to the number of fiber ports at the top of each module, which is equal to $N/2$, where N is the effective radix of the EoR switch. Admittedly, this represents an added expenditure in the data center, but the added layer of fiber switches can potentially perform other useful functions in the data center such as upgrades, maintenance, and restoration. This subject is still under investigation.

6. SUMMARY AND OUTLOOK

We have summarized the technology and network architectural visions of the INTREPID project. The technology pursues the use of coherent QPSK, polarization-multiplex transceivers enhanced with WDM to enable energy-efficient 800 or 1600 Gb/s inter-switch fiber links. CPO is pursued for integrating the transceivers with next-generation 51.2 and 102.4 Tb/s electronic switching ASICs to enable the realization of future hyper-scale data centers that are flatter and more energy-efficient than current designs. The technology is compatible with conventional three-level data center designs as well as a newly introduced two-level data center design that includes an added layer of passive AWGRs or WDM circuit switches to further reduce cost, power consumption, and latency.

The second phase of INTERPID, which began in late 2020, focuses on robust demonstrations of analog coherent transceiver assemblies and a transition of the technology developed in the program to widespread commercial adoption [48]. The benefits of deploying WDM switches, possibly in combination with fiber switches, in the data center will also be further investigated and quantified to help to make the case for practical deployment of the novel data center architectures proposed here.

Funding. Advanced Research Projects Agency - Energy (DE-AR0000848).

Acknowledgment. The authors thank the ARPA-E team for management and guidance: James Zahler, Michael Haney, and John Qi. The authors greatly appreciate the technical contributions and rewarding collaborations with their colleagues at Facebook (James Stewart, Hans-Juergen Schmidtke,

Todd Hollmann, and Jimmy Williams) and UCSB (Hector Andrade, Takako Hirokawa, Junqian Liu, Aaron Maharry, Stephen Misak, Luis Valenzuela, and Yujie Xia). The information, data, or work presented herein was funded in part by the Advanced Research Projects Agency - Energy (ARPA-E), U.S. Department of Energy, (DE-AR0000848). The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Disclosures. The authors declare no conflicts of interest.

REFERENCES

1. Cisco, "Cisco annual Internet report (2018–2023)," 2020, https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html.
2. Cisco, "Cisco global cloud index: forecast and methodology, 2016–2021," 2018, https://virtualization.network/Resources/Whitepapers/0b75cf2e-0c53-4891-918e-b542a5d364c5_white-paper-c11-738085.pdf.
3. C. L. Schow and K. E. Schmidtke, "INTREPID: developing power efficient analog coherent interconnects to transform data center networks," in *Optical Fiber Communication Conference (OFC)*, OSA Technical Digest (Optical Society of America, 2019), paper M4D.9.
4. "Co-packaged optics collaboration," http://www.copackagedoptics.com/.
5. ISSCC, "ISSCC 2020 trends," 2020, http://isscc.org/wp-content/uploads/sites/17/2020/03/isscc2020.press_kit_final.pdf.
6. T. O. Dickson, Y. Liu, S. V. Rylov, A. Agrawal, S. Kim, P.-H. Hsieh, J. F. Bulzacchelli, M. Ferriss, H. A. Ainspan, A. Rylyakov, B. D. Parker, M. P. Beakes, C. Baks, L. Shan, Y. Kwark, J. A. Tierno, and D. J. Friedman, "A 1.4 pJ/bit, power-scalable 16 × 12 Gb/s source-synchronous I/O with DFE receiver in 32 nm SOI CMOS technology," IEEE J. Solid-State Circuits **50**, 1917–1931 (2015).
7. G. P. Agrawal, *Fiber-Optic Communication Systems* (Wiley, 2012).
8. T. Hirokawa, S. Pinna, D. Hosseinzadeh, A. Maharry, H. Andrade, J. Liu, T. Meissner, S. Misak, G. Movaghar, L. A. Valenzuela, Y. Xia, S. Bhat, F. Gambini, J. Klamkin, A. A. M. Saleh, L. Coldren, J. F. Buckwalter, and C. L. Schow, "Analog coherent detection for energy efficient intra-data center links at 200 Gbps per wavelength," J. Lightwave Technol. **39**, 520–531 (2021).
9. J. K. Perin, A. Shastri, and J. M. Kahn, "Design of low-power DSP-free coherent receivers for data center links," J. Lightwave Technol. **35**, 4650–4662 (2017).
10. P. R. A. Binetti, M. Lu, E. J. Norberg, R. S. Guzzon, J. S. Parker, A. Sivananthan, A. Bhardwaj, L. A. Johansson, M. J. Rodwell, and L. A. Coldren, "Indium phosphide photonic integrated circuits for coherent optical links," IEEE J. Quantum Electron. **48**, 279–291 (2012).
11. M. Lu, H. Park, E. Bloch, A. Sivananthan, J. S. Parker, Z. Griffith, L. A. Johansson, M. J. W. Rodwell, and L. A. Coldren, "An integrated 40 Gbit/s optical Costas receiver," J. Lightwave Technol. **31**, 2244–2253 (2013).
12. C. L. Schow, A. A. M. Saleh, and K. E. Schmidtke, "Intrepid update: ARPA-E ENLITENED annual meeting," 2018, https://arpa-e.energy.gov/sites/default/files/Schow_ENLITENED2018.pdf.
13. C. L. Schow, A. A. M. Saleh, and K. E. Schmidtke, "INTREPID update: ARPA-E ENLITENED annual meeting," 2019, https://arpa-e.energy.gov/sites/default/files/UCSB_Schow_ENLITENED_Annual_Meeting.pdf.
14. Y. Xia, L. Valenzuela, A. Maharry, S. Pinna, S. Dwivedi, T. Hirokawa, J. Buckwalter, and C. Schow, "A fully integrated O-band coherent optical receiver operating up to 80 Gb/s," in *IEEE Photonics Conference (IPC)*, Virtual Conference (2021).
15. H. Andrade, Y. Xia, A. Maharry, L. Valenzuela, J. Buckwalter, and C. Schow, "50 GBaud QPSK 0.98 pJ/bit receiver in 45 nm CMOS and 90 nm silicon photonics," in *European Conference on Optical Communications (ECOC)*, Bordeaux, France (2021).
16. H. Andrade, T. Hirokawa, A. Maharry, A. Rylyakov, C. L. Schow, and J. F. Buckwalter, "Monolithically-integrated 50Gbps 2pJ/bit photoreceiver with Cherry-Hooper TIA in 250 nm BiCMOS technology," in *Optical Fiber Communication (OFC) Conference*, San Diego, California, March 2019.
17. H. Andrade, A. Maharry, T. Hirokawa, L. Valenzuela, S. Pinna, S. Simon, C. L. Schow, and J. F. Buckwalter, "Analysis and monolithic implementation of differential transimpedance amplifiers," J. Lightwave Technol. **38**, 4409–4418 (2020).
18. N. Hosseinzadeh, K. Fang, L. Valenzuela, C. L. Schow, and J. F. Buckwalter, "A 50-Gb/s optical transmitter based on co-design of a 45-nm CMOS SOI distributed driver and 90-nm silicon photonic Mach–Zehnder modulator," in *IEEE/MTT-S International Microwave Symposium (IMS)*, Online Conference, June 2020.
19. A. Maharry, H. Andrade, T. Hirokawa, J. F. Buckwalter, and C. L. Schow, "A novel architecture for a two-tap feed-forward optical or electrical domain equalizer using a differential element," in *IEEE Photonics Conference*, San Antonio, Texas, October 2019.
20. L. A. Valenzuela, A. Maharry, H. Andrade, C. L. Schow, and J. F. Buckwalter, "A 108-Gbps, 162-mW Cherry-Hooper transimpedance amplifier," in *IEEE BiCMOS and Compound Semiconductor Integrated Circuits and Technology Symposium*, Online Conference (2020).
21. J. E. Proesel, B. G. Lee, C. W. Baks, and C. L. Schow, "35-Gb/s VCSEL-based optical link using 32-nm SOI CMOS circuits," in *Optical Fiber Communications Conference*, Anaheim, California (2013).
22. D. M. Kuchta, A. V. Rylyakov, F. E. Doany, C. L. Schow, J. E. Proesel, C. W. Baks, P. Westbergh, J. S. Gustavsson, and A. Larsson, "A 71 Gb/s NRZ modulated 850 nm VCSEL-based optical link," IEEE Photon. Technol. Lett. **27**, 577–580 (2015).
23. M. Filer, "Opportunities for co-packaging in future microsoft data center networks," 2021, https://arpa-e.energy.gov/sites/default/files/2021-02/DAY2_Filer_ENLITENED_Phase2_Kickoff.pdf.
24. D. M. Kuchta, "Multi-wavelength optical transceivers integrated on node," 2019, https://arpa-e.energy.gov/sites/default/files/IBM_Kuchta_ENLITENED_Annual_Meeting.pdf.
25. A. V. Rylyakov, C. L. Schow, B. G. Lee, F. E. Doany, C. W. Baks, and J. A. Kash, "Transmitter pre-distortion for simultaneous improvements in bit-rate, sensitivity, jitter, and power efficiency in 20 Gb/s CMOS-driven VCSEL links," J. Lightwave Technol. **30**, 399–405 (2012).
26. L. A. Valenzuela, L. Andrade, N. Hosseinzadeh, A. Maharry, C. L. Schow, and J. F. Buckwalter, "A 2.85 pJ/bit, 52-Gbps NRZ VCSEL driver with two-tap feedforward equalization," in *IEEE/MTT-S International Microwave Symposium (IMS)*, Online Conference, June 2020.
27. A. Maharry, L. Valenzuela, H. Andrade, I. Kalifa, I. Cestier, M. Galanty, B. Atias, A. Sandomirsky, E. Mentovich, L. Coldren, J. Buckwalter, and C. L. Schow, "A 50 Gbps 9.5 pJ/bit VCSEL-based optical link," in *IEEE Photonics Conference (IPC)*, Virtual Conference (2021).
28. Broadcom, "Broadcom ships Tomahawk 4, industry's highest bandwidth Ethernet switch chip at 25.6 terabits per second," 2019, https://www.broadcom.com/company/news/product-releases/52756.
29. A. A. M. Saleh, "Scaling-out data centers using photonics technologies," in *Photonics in Switching Conference in Advanced Photonics for Communications*, San Diego, California, July 2014.
30. A. D. Papaioannou, R. Nejabati, and D. Simeonidou, "The benefits of a disaggregated data centre: a resource allocation approach," in *IEEE GLOBECOM* (2016).
31. P. X. Gao, A. Narayan, S. Karandikar, J. Carreira, S. Han, R. Agarwal, S. Ratnasamy, and S. Shenker, "Network requirements for resource disaggregation," in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI)* (2016).
32. R. Lin, Y. Cheng, M. De Andrade, L. Wosinska, and J. Chen, "Disaggregated data centers: challenges and trade-offs," IEEE Commun. Mag. **58**(2), 20–26 (2020).
33. Q. Cheng, Y. Huang, H. Yang, M. Bahadori, N. Abrams, X. Meng, M. Glick, Y. Liu, N. Hochberg, and K. Bergman, "Silicon photonic switch topologies and routing strategies for disaggregated data centers," IEEE J. Sel. Top. Quantum Electron. **26**, 8302010 (2020).
34. B. R. Koch, E. J. Norberg, B. Kim, J. Hutchinson, J.-H. Shin, G. Fish, and A. Fang, "Integrated silicon photonic laser source for telecom and datacom," in *Optical Fiber Communication Conference*

and Exposition/National Fiber Optics Engineers Conference (OFC/NFOEC) (2013).
35. A. J. Zilkie, P. S. Srinivasan, A. Trita, T. Schrans, G. Yu, J. Byrd, D. A. Nelson, K. Muth, D. Lerose, M. Alalusi, K. Masuda, M. Ziebell, H. Abediasl, J. J. Drake, G. Miller, H. Nykanen, E. Kho, Y. Liu, H. Liang, H. Yang, F. H. Peters, A. S. Nagra, and A. G. Rickman, "Multi-micron silicon photonics platform for highly manufacturable and versatile photonic integrated circuits," IEEE J. Sel. Top. Quantum Electron. **25**, 8200713 (2019).
36. 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 *SIGCOMM* (2010).
37. G. Porter, R. Strong, N. Farrington, A. Forencich, C.-S. Pang, T. Rosing, Y. Fainman, G. Papen, and A. Vahdat, "Integrating microsecond circuit switching into the data center," in *SIGCOMM* (2013).
38. H. Liu, F. Lu, A. Forencich, R. Kapoor, M. Tewari, G. M. Voelker, G. Papen, A. C. Snoeren, and G. Porter, "Circuit switching under the radar with REACToR," in *USENIX Symposium on Networked Systems Design and Implementation (NSDI)* (2014).
39. A. A. M. Saleh, A. S. P. Khope, J. E. Bowers, and R. C. Alferness, "Elastic WDM switching for scalable data center and HPC interconnect networks," in *OECC/Photonics in Switching Conference*, Niigata, Japan, July 2016.
40. M. Ghobadi, R. Mahajan, A. Phanishayee, N. Devanur, J. Kulkarni, G. Ranade, P.-A. Blanche, H. Rastegarfar, M. Glick, and D. Kilper, "ProjecToR: agile reconfigurable data center interconnect," in *SIGCOMM* (2016).
41. W. M. Mellette, R. McGuinness, A. Roy, A. Forencich, G. Papen, A. C. Snoeren, and G. Porter, "RotorNet: a scalable, low-complexity, optical datacenter network," in *SIGCOMM* (2017).
42. K. Bergman, J. Shalf, G. Michelogiannakis, S. Rumley, L. Dennison, and M. Ghobadi, "PINE: an energy efficient flexibly interconnected photonic data center architecture for extreme scalability," in *IEEE Optical Interconnects Conference* (2018).
43. R. Helkey, A. A. M. Saleh, J. Buckwalter, and J. E. Bowers, "High-performance photonic integrated circuits on silicon," IEEE J. Sel. Top. Quantum Electron. **25**, 8300215 (2019).
44. A. S. P. Khope, M. Saeidi, R. Yu, X. Wu, A. M. Netherton, Y. Liu, Z. Zhang, Y. Xia, G. Fleeman, A. Spott, S. Pinna, C. Schow, R. Helkey, L. Theogarajan, R. C. Alferness, A. A. M. Saleh, and J. E. Bowers, "Multi-wavelength selective crossbar switch," Opt. Express **27**, 5203–5216 (2019).
45. T. Hirokawa, A. Maharry, R. Helkey, J. E. Bowers, A. A. M. Saleh, and C. L. Schow, "Demonstration of a spectrally-partitioned 4x4 crossbar switch with 3 drops per cross-point," in *Optoelectronics and Communications Conference/International Conference Photonics in Switching and Computing (OECC/PSC)* (2019).
46. T. Hirokawa, M. Saeidi, S. Pillai, A. Nguyen-Le, L. Theogarajan, A. A. M. Saleh, and C. L. Schow, "A wavelength-selective multi-wavelength ring-assisted Mach–Zehnder interferometer switch," J. Lightwave Technol. **38**, 6292–6298 (2020).
47. T. J. Seok, J. Luo, Z. Huang, K. Kwon, J. Henriksson, J. Jacobs, L. Ochikubo, S. R. Muller, and C. M. Wu, "Silicon photonic wavelength cross-connect with integrated MEMS switching," APL Photon. **4**, 100803 (2019).
48. "INTREPID Phase 2," 2021, https://arpa-e.energy.gov/sites/default/files/2021-01/DAY1_Schow_ENLITENED_Phase2_Kickoff.pdf.

Adel A. M. Saleh (A'65–M'70–SM'76–F'87–LF'08) obtained a B.Sc. degree from Alexandria University, Egypt, and M.S. and Ph.D. degrees from MIT, Cambridge, MA, in 1963, 1967, and 1970, respectively, all in electrical engineering. He is currently a Research Professor with the Department of Electrical and Computer Engineering and the Institute for Energy Efficiency, University of California, Santa Barbara (UCSB), Santa Barbara, CA, USA, since October 2011, conducting research on optical networking and photonic technology for chip-scale to global-scale networking applications, with emphasis on data centers and high-performance computing. From 2005 to 2011, he was a DARPA Program Manager, where he initiated several research programs on advanced optical networking. From 1999 to 2004, he held leadership positions in the optical networking industry, including Corvis, the first company to commercialize national-scale all-optical networks. From 1970 to 1999, he was with Bell Labs/AT&T Labs Research conducting and leading research on optical and wireless networks. Between 1992 and 1999, he led the AT&T effort on several multi-million-dollar, cross-industry, DARPA-funded consortia that pioneered the vision of all-optical networking in backbone, regional, metro, and access networks. Dr. Saleh received the AT&T Bell Labs Distinguished Technical Staff Award for Sustained Achievement in 1985. He is a Fellow of Optica and a Life Fellow of the IEEE, has published more than 100 journal and conference papers, and has 25 issued patents.

Katharine E. Schmidtke received a B.Sc. degree in physics and mathematics from Keele University, UK, in 1989 and a Ph.D in laser physics and nonlinear optics from the University of Southampton in 1993. She went on to complete postdoctoral research in epitaxial growth of nonlinear optical materials at Stanford University, CA, USA. She has a 25-year career in the optical communications industry including roles at Finisar, JDS Uniphase, and New Focus. For the past 7 years she has worked at Facebook, Menlo Park, CA, where she has driven the technology strategy for data center optical interconnects. She is currently Director of Sourcing for ASICS and Custom Silicon focused on AI/ML applications. Dr. Schmidtke is a Fellow of Optica and has been an invited speaker and served on committees for numerous international conferences, including the Optical Fiber Communication Conference (OFC), the European Conference on Communications (ECOC), and the Optical Interconnects Conference.

Robert J. Stone received his B.Sc. from The University of Sheffield in 1993 and a D.Phil. from The University of Oxford in 1997 in physics. He currently holds the role of Technical Sourcing Manager at Facebook, focusing on next-generation high-speed interconnects. Prior to Facebook, Dr. Stone was a distinguished engineer at Broadcom within the switch architecture team, where he was responsible for the switch IO ecosystem and system design, actively representing Broadcom in a number of industry groups, including IEEE as well as MSA and industry organizations. He has over 20 years of industry experience bringing communications technologies to market and has held both technical and managerial positions at Broadcom, Intel, Infinera, Emcore, Skorpios Technologies, and Bandwidth 9.

James F. Buckwalter (SM'10) received a B.S.E.E. degree with honors in electrical engineering from the California Institute of Technology (Caltech), Pasadena, in 1999; an M.S. degree from the University of California, Santa Barbara (UCSB), in 2001; and a Ph.D. degree in electrical engineering from Caltech in 2006. He is currently a Professor of Electrical and Computer Engineering with UCSB. From 1999 to 2000, he was a Research Scientist with Telcordia Technologies. During Summer 2004, he was with the IBM T. J. Watson Research Center, Yorktown Heights, NY. In July 2006, he joined the faculty of the University of California, San Diego (UCSD) as an Assistant Professor and was promoted to Associate Professor in 2012. He joined UCSB in 2014 as a Full Professor. Dr. Buckwalter was the recipient of a 2004 IBM Ph.D. Fellowship, 2007 Defense Advanced Research Projects Agency (DARPA) Young Faculty Award, 2011 NSF CAREER Award, and 2015 IEEE MTT-S Young Engineer Award. He is a senior member of the IEEE. He has published more than 200 conference and journal papers and advised more than 20 Ph.D. students.

Larry A. Coldren (S'67–M'72–SM'77–F'82–LF'12) received a B.S. in electrical engineering and a B.A. in physics from Bucknell University and joined Bell Laboratories in 1968. Under Bell Lab's support he then attended Stanford University and received M.S. and Ph.D. degrees in electrical engineering in 1969 and 1972, respectively. Following 13 years in the research area with Bell Laboratories, he joined the ECE Department of the University of California, Santa Barbara (UCSB) in 1984. In 1986 he was a founding member of the Materials Department. He became the Fred Kavli Professor of Optoelectronics and Sensors in 1999. From 2009 to 2011, he was acting Dean of the College of Engineering, and in 2017 he became

Professor Emeritus and a Distinguished Research Professor. In 1990, he co-founded Optical Concepts, later acquired as Gore Photonics, to develop novel vertical-cavity surface-emitting laser (VCSEL) technology, and, in 1998, he cofounded Agility Communications, later acquired by JDSU (now Lumentum), to develop widely tunable integrated transmitters and transponders. At UCSB, he has worked on multiple-section widely tunable lasers and efficient VCSELs. He continues to research high-performance InP-based photonic integrated circuits and high-speed, high-efficiency VCSELs for various applications. Prof. Coldren has authored or coauthored over 1000 journal and conference papers, 8 book chapters, a widely used textbook, and 63 issued patents. He is a fellow of IEEE, Optica, IEE, and the National Academy of Inventors, as well as a member of the National Academy of Engineering. He has been a recipient of the 2004 John Tyndall, the 2009 Aron Kressel, the 2014 David Sarnoff, the 2015 IPRM, and the 2017 Nick Holonyak, Jr., Awards.

Clint L. Schow (SM'10–F'18) received B.S., M.S., and Ph.D. degrees from the University of Texas at Austin in 1994, 1997, and 1999, respectively. After positions at IBM and Agility Communications, he spent more than a decade at the IBM T. J. Watson Research Center in Yorktown Heights, NY, as a Research Staff Member and Manager of the Optical Link and System Design group. In 2015, he joined the faculty of the University of California, Santa Barbara. He has led international R&D programs spanning chip-to-chip optical links; VCSEL and Si photonic transceivers; nanophotonic switches; and new system architectures enabled by high-bandwidth, low-latency photonic networks. Dr. Schow has been an invited speaker and served on committees for numerous international conferences, including roles as General Chair for the Optical Fiber Communications Conference (OFC), the Optical Interconnects Conference, and the Photonics in Switching Conference. He is a Fellow of Optica and the IEEE, has published more than 200 journal and conference articles, and has 33 issued patents.