21. Evolving Requirements and Trends in Datacenter Networks

Hong Liu, Ryohei Urata, Xiang Zhou, Amin Vahdat

In this manuscript, we present an overview of Google's datacenter network, which has led and defined the industry over the past few decades. Starting inside the datacenter, we cover all aspects, from networking/topology to key hardware components of interconnect and switching, traffic/throughput, and energy usage/efficiency for intra-datacenter networks. Likewise, we discuss topology and interconnect for inter-datacenter networks. With particular focus on optical interconnect, we also discuss future technology directions for scaling bandwidth through a combination of higher baud rates, wavelength-division multiplexing (WDM), coherent communication (polarization multiplexing, I/Q modulation), and space-division multiplexing (SDM), along with the corresponding trade-offs between these various dimensions and how these trade-offs are adjusted at different length scales. Although questions remain on the exact implementation to be adopted in the future, one thing is clear: the evolution of datacenter networks and the underlying technologies have been and will remain a critical driver for enabling new compute capabilities in the cloud.

Over the past decade, the datacenter has become the technology enabler for web-based applications, with prominent examples being web search, social media applications, and enterprise software (email/docs/storage). With the user interface often being a thin client device (stateless mobile/laptop device), the actual running of the application is performed in a remote datacenter. Starting from humble beginnings several decades ago [21.1], Google's compute infrastructure has been dramatically improved on all axes, from the cooling and power infrastructure and associated efficiency (power usage effectiveness, PUE), to the underlying hardware of servers, storage, and networking. In fact, the datacenter has quickly evolved from just racks of servers to a unified, global computer consisting of *copies* of massive datacenters interconnected throughout the world to deliver various services with low latency and high reliability, in a synchronized/consistent fashion. Google has thus built a global computing infrastructure and, in turn, developed and incorporated transformative datacenter networking technologies in order to keep pace with the increasing number and bandwidth demands of these applications.

The importance of the datacenter will be further enhanced as the cloud IaaS (infrastructure-as-a-service) growth migrates an increasing share of all compute to the cloud. Cloud-based platforms enjoy the same benefits as the aforementioned computing for web apps in terms of scalability, accessibility, and reliability. Without the concerns of running and maintaining an information technology (IT) infrastructure, users have immediate access to a single machine or thousands of machines to quickly and easily scale and handle increasing user workloads/services whenever the need arises. All user data reside in the cloud and are accessible anywhere in the world as long as there is a network connection. Lastly, the data are automatically backed up with multiple copies dispersed throughout the world,

B. Mukherjee et al. (Eds.), *Springer Handbook of Optical Networks*, Springer Handbooks, https://doi.org/10.1007/978-3-030-16250-4_21

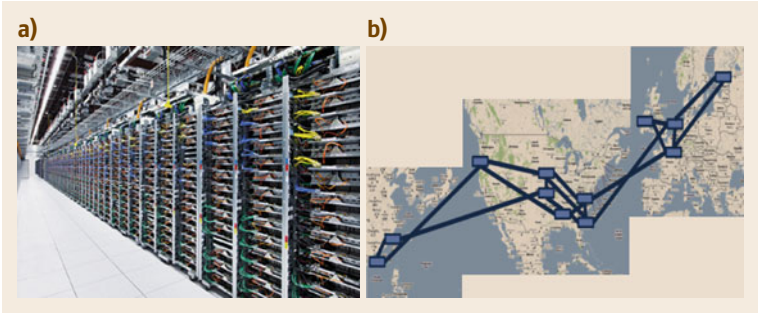

Fig. 21.1a,b Google's datacenter **(a)** showing rows of racks which house hundreds of thousands of servers inside the building (http://www.google.com/about/datacenters/) and **(b)** global B4 inter-datacenter wide area network (WAN) (after [21.2])

being fault-tolerant to local hardware or facility failure.

All of these benefits require a reliable and scalable network infrastructure. Datacenter networks are typically separated into two categories: (1) intra-datacenter networks used to connect machines/servers within the same building or same campus, and (2) inter-datacenter networks that interconnect multiple datacenters. As a result of their role/function, the intra- versus inter-type networks have different requirements regarding topology, reliability, and bandwidth capacity. Intra-datacenter networks employ a massively parallel fabric with rich path diversity for scalability and load balancing. Inter-datacenter networks are more point-to-point, with much higher capacity per link and interconnect having much longer reaches/distances. Figure 21.2 provides a high-level view of datacenter interconnects, divided into four segments based on reach and corresponding technology adopted:

1. The intra-datacenter interconnect for link distances up to 1 km within the same building (mix of copper and optics)
2. The intra-campus network, which interconnects clusters housed in different buildings within a 2 km campus neighborhood
3. The point-to-point metro edge access, which provides connections between datacenters and the global backbone network or POPs (point-of-presence), with a link distance typically less than 80 km; and
4. The long-haul and subsea backbone transport interconnecting datacenters throughout the world, with up to thousands of kilometers of reach.

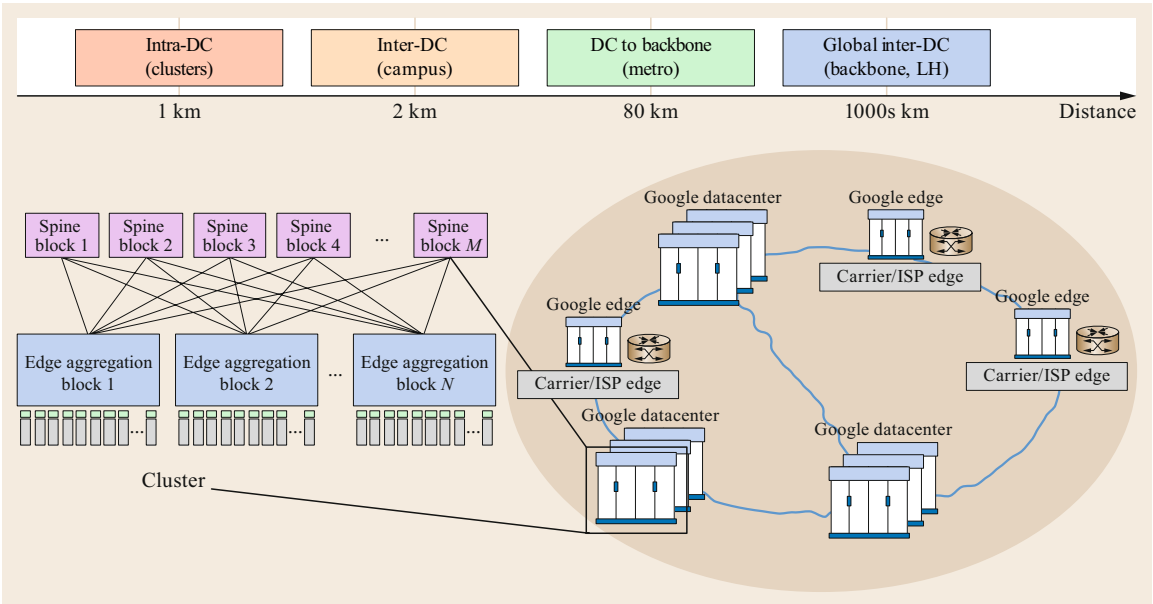

Fig. 21.2 A high-level view of interconnection types for the global datacenter network

21.1 Intra-Datacenter Network

The datacenter network at the building scale is typically a scale-out type of network that is wide and flat (such as a fat-tree network) through a richly interconnected and meshed network. This network interconnects thousands of commodity-class servers and storage devices beneath it. Leveraging a similar approach to scale-out computing with the implementation of arrays of commodity servers for increasing aggregate compute, scale-out datacenter networks are constructed using the following design principles: multiple stages of meshed switching blocks to maximize scale-out and nonblocking bandwidth provisioning with small failure domains; the use of low-cost, commodity merchant switch silicon for faster product cycles to take full advantage of Moore's law; and centralized control software with routing and management protocols tailored to intra-datacenter requirements [21.1]. A diagram of an intra-datacenter network example is shown in Fig. 21.3a,b [21.1, 3]. Machines are housed in racks, which connect up to top-of-rack (ToR) switches. There are multiple paths from the ToR to edge switches (often referred to as leaf switches) for path redundancy and fault tolerance. The edge switches may consist of multiple switches interconnected in a tightly coupled meshed fashion to constitute a nonblocking switching unit. These internal switches may consist of single switch silicon application-specific integrated circuits (ASICs) or an array of ASICs interconnected with a mesh topology. With its large radix, the edge switches can then be fanned out to all spine switches to create a network which maximizes bisection bandwidth (bisection bandwidth is the minimum possible bandwidth of the network when the fabric is bisected at all possible tiers).

An ideal datacenter network should be nonblocking to allow flexible placement of compute jobs among machines, flat with as few tiers as possible to reduce latency and cost, and wide enough to support all compute nodes with predictable latency from machine to machine. The two most important characteristics of intra-datacenter networks are scalability and cost/power efficiency, which are influenced by all aspects of the datacenter network design: fabric topology, physical layout, switch hardware, interconnect selection, network routing, and management control. With the anticipated growth of the cloud, the scaling and efficiency aspects will continue to influence future network designs. For example, a datacenter with $> 100\,000$ servers, each with 100 Gb/s of bandwidth allocated, would require an internal network with 10 Pb/s of aggregate bandwidth to support full-bandwidth communication among all servers. On the other hand, in order to maintain energy and space efficiency, as the bandwidth and number of servers scale up, the datacenter network needs to scale within the same footprint and cost.

There are many considerations in building a large-scale datacenter network, including but not limited to cost (capital expenditure and operating expenses; CapEx and OpEx) and performance. The key technologies are the fabric topology chosen, the switch silicon ASIC, and the interconnect to implement the fabric, each of which will be discussed below.

21.1.1 Fabric Topology

To connect tens of thousands of nodes, there are a variety of network topologies which can be chosen: flat-

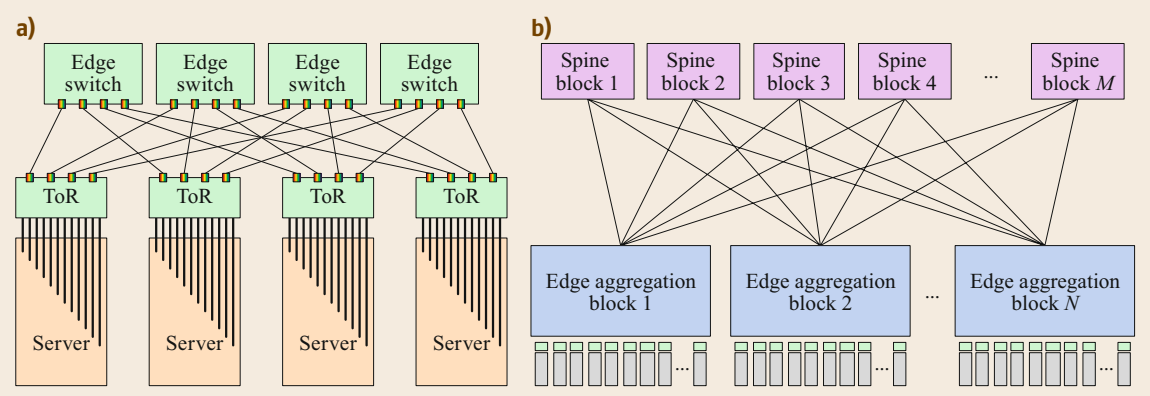

Fig. 21.3a,b Diagram of scale-out type of datacenter network. A tiered architecture is used with **(a)** fan-out from ToR to edge switches with multiple connections, **(b)** from edge aggregation block to spine blocks with all-to-all connections

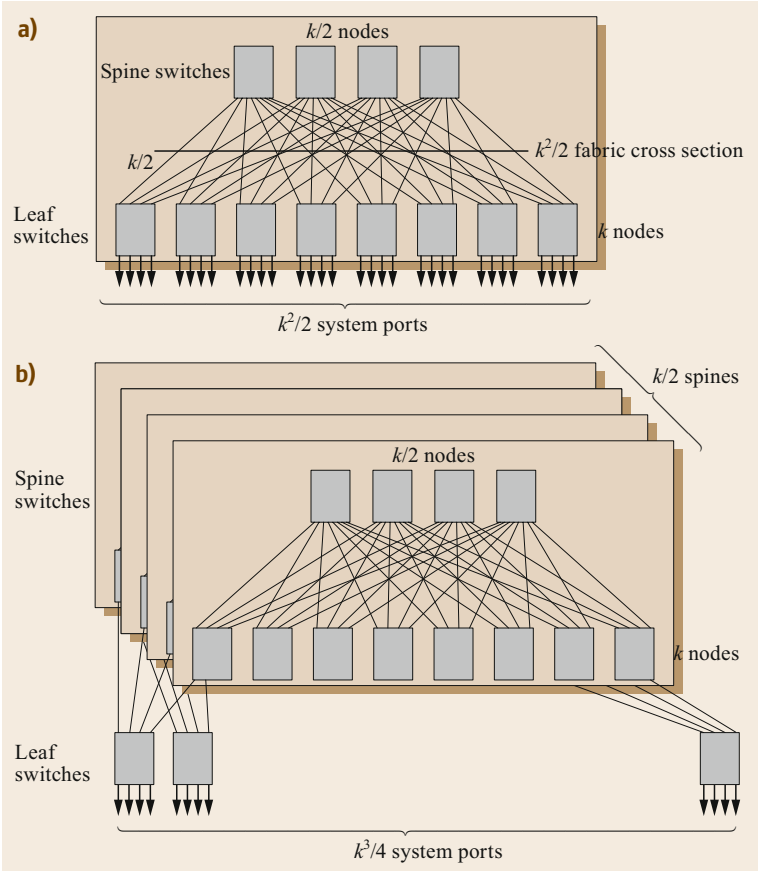

Fig. 21.4a,b Diagram of Clos fabric built with k-port switches showing **(a)** two-stage Clos, **(b)** three-stage Clos

tened butterfly [21.4], B-cube [21.5], Clos [21.6], and others [21.7]. The topology choice is often a trade-off between software routing complexity, physical link efficiency/number, and fabric scalability. For example, flattened butterfly trades off fewer physical links but requires adaptive routing to load-balance arbitrary traffic patterns, whereas a Clos/fat tree has multiple physical paths and simpler routing to handle arbitrary traffic patterns.

As described in Fig. 21.3, Clos topology is most commonly used for large-scale datacenters to support full-bisection bandwidth and graceful fault tolerance. This architecture enables the use of identical switching elements with smaller radix (i.e., the switch ASIC) to form a multistage Clos for a very large-scale network fabric that would be impossible to achieve with monolithic single-chip or single-chassis technologies [21.8]. Figure 21.4a shows a two-stage Clos constructed with identical k-port switches with all-to-all spine-to-leaf switch connections. The leaf switches have k nodes of k-port switches, with $k/2$ ports of each switch connecting to system ports, and $k/2$ ports fanning out to $k/2$ spine switch nodes. The k ports of each spine switch fan out to all leaf switches (k nodes)

$$\text{Total bisection bandwidth} = \text{number of spine links} \times \text{bandwidth per link} \times 2$$

$$\text{Total number of connected system ports} = (k^2)/2\,.$$

To further increase the bisection bandwidth and total number of system ports (or hosts) that can be interconnected, the two-stage fat tree can be extended to a three-stage fat tree (Fig. 21.4b) with $k/2$ port connections to each lower-stage switch. Layers can be continually added in this manner to increase the bandwidth of the network at the expense of latency and cost. Ideally, a fully meshed networking fabric that connects every system port in a datacenter provides full bisection bandwidth (same amount of bandwidth between any two nodes), which leads to easier programming (no consideration of the underlying network infrastructure and where bandwidth constraints lie) and better utilization of server compute capability. However, such a design would be prohibitively expensive, and modest oversubscription (i.e., bandwidth provisioned is less

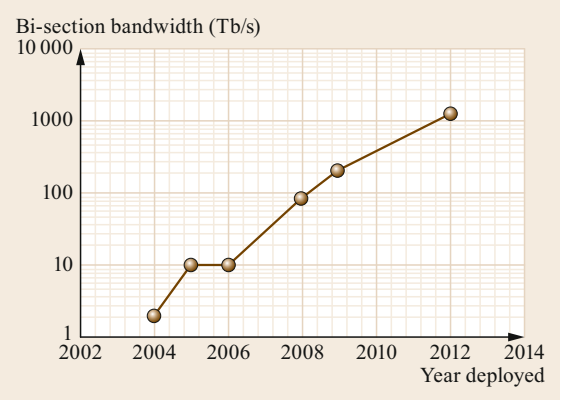

Fig. 21.5 Evolution of bisection bandwidth inside Google's datacenter

than maximum possible peak/demand bandwidth with fewer upward-facing than downward-facing links) is often applied at some layer/tier to reduce CapEx.

Figure 21.5 shows the evolution of Google's datacenter network fabric bandwidth over time. Leveraging the Clos topology and the increase of switch ASIC bandwidths, a three-orders-of-magnitude increase in bandwidth was achieved over roughly a decade, with the Jupiter network capable of 1.3 Pb/s of bisection bandwidth at maximum scale (after [21.1]).

21.1.2 Switch Silicon

Driving sustained increases in bandwidth, merchant switch silicon has become a key building block for datacenter fabrics, as it enables faster time to market and quicker product refresh cycles at much lower cost than with customized ASICs. With the continuous scaling of technologies for signaling speed, the increase in the number of signal pins, and power scaling from Moore's law, the capacity of high-speed switch chips has seen an impressive tenfold increase over the past 10 years. The evolution of the bandwidth available from a single switch chip is shown in Fig. 21.6. A single switch chip today can offer > 12 Tb/s of switching capacity, and > 25 Tb/s switching capacity in the near future using a 7 nm complementary metal–oxide semiconductor (CMOS) technology node. The radix (number of ports) of the switch chip has also increased from 16 to 256 in the past 10 years. The higher switch capacity and larger radix enable large-scale and more efficient datacenter fabrics with > Pb/s bisection bandwidth, using fewer stages and external interconnections.

However, switch silicon bandwidth may become constrained by the saturation of package pin count and slowing power consumption improvements of the I/O (SerDes interface of each port), due to Moore's law slowdown. While opportunities exist to improve the energy efficiency of each channel with advances in CMOS technology, the package pin count scales more slowly than I/O speed and power, per the International Technology Roadmap for Semiconductors (ITRS) projection, thus slowing the aggregate switch bandwidth growth.

21.1.3 Intra-Datacenter Interconnect

Inside the datacenter, one of the most notable characteristics of the network is the large amount of fan-out. A much larger number of transceivers and interconnects are thus required to implement the topology, which mo-

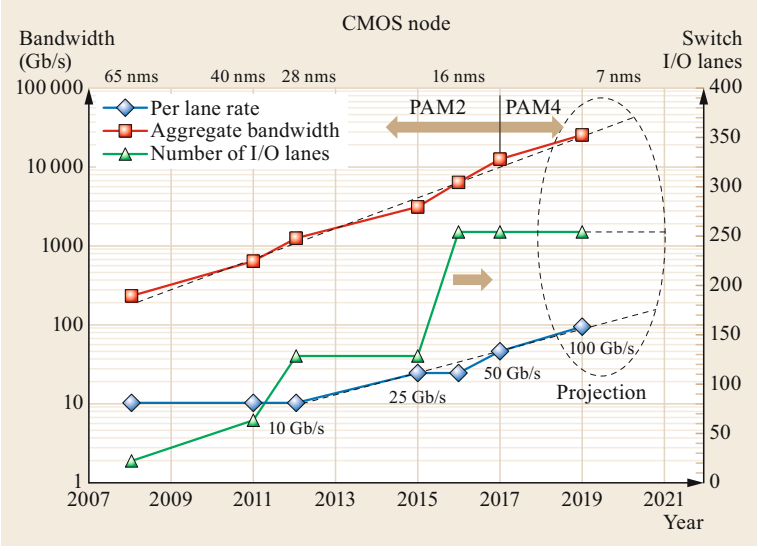

Fig. 21.6 Switch I/O bandwidth and technology scaling trend

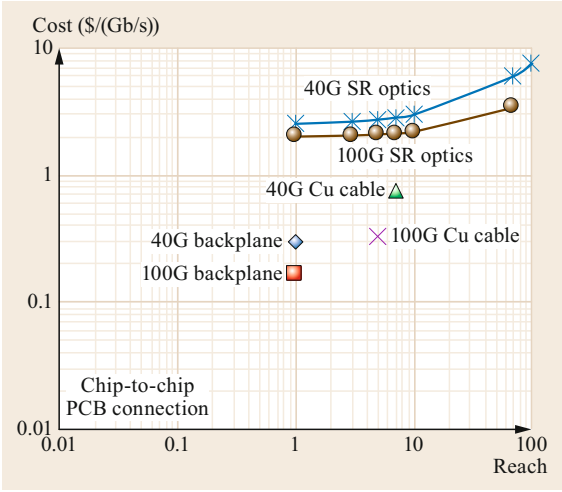

Fig. 21.7 Cost and reach comparison for electrical and optical interconnection for 40G and 100G

tivates a strong focus on improving cost, power, and density/size, at a variety of reaches.

Figure 21.7 shows the cost comparison for electrical and VCSEL-based short-reach optics at 40G (4×10 Gb/s) and 100G (4×25 Gb/s). Printed circuit board (PCB) traces are used for the switch I/O connections between chip-to-chip, chip-to-module, and backplane interconnects up to 1 m reach. Although there has been much anticipation and debate regarding optical intra-rack connections, for the data rates shown in Fig. 21.7, copper direct attach cable (DAC) still seems to be the most cost-effective and power-efficient interconnect for intra-rack connections up to 3 m at 25 Gb/s. Vertical-cavity surface-emitting laser (VCSEL)- and multimode fiber (MMF)-based short-reach technologies have shown the best overall link cost (transceiver cost plus fiber cost) for up to 25 Gb/s per lane speed and for reaches up to 100 m.

VCSEL- and MMF-based technologies have been widely deployed in Google datacenters since 2007 [21.1]. This has tipped the scales away from copper interconnects, which were bulky and very difficult to deploy at the number and lane speed scales required for inter-rack connection. Over the past 10 years, the performance of VCSEL arrays has improved to extend the reach to 100 m at 40 Gb/s (4×10 Gb/s) over OM3 fiber and 100 m at 100 Gb/s (4×25 Gb/s) over OM4 fiber. Beyond 50 Gb/s lane speed, due to the performance limitations on bandwidth and reach, it may be challenging for VCSELs to scale for 100 Gb/s PAM4 without advanced digital signal processing (DSP).

The interconnections between the edge aggregation switches and the spine switches form the core of the network fabric, which can span an entire mega-datacenter building or campus. Long-reach optical transceivers and single-mode fiber (SMF) transmission technologies must be used to achieve the required bandwidth reach and cabling efficiency. In 2012, Google achieved the first large-scale deployment of wavelength-division multiplexing (WDM)/single-mode fiber (SMF)-based interconnects inside the datacenters with 40 GbE quad (4-channel) small form-factor pluggable (QSFP) form-factor transceivers [21.1]. Using wavelength-division multiplexing and photonic integration, the solution can scale to higher data rates of 400 Gb/s, with four wavelengths each running at 100 Gb/s PAM4 modulation, and longer reach (> 2 km).

Next Generation: Intra-Rack Interconnect

For traditional on-off keying (OOK) modulation, beyond 50 Gb/s lane speed, the reach of signal transmission over copper is limited by the large loss at high frequency. Pulse-amplitude modulation (PAM) signaling overcomes the loss limitation at the expense of signal-to-noise ratio (SNR) penalty due to multilevel signaling. Lower-loss PCB materials and DSP-powered PAM4 (four-level PAM) SerDes may be needed to enable low-cost PCB interconnects. Because of the higher power consumption of SerDes I/O required to compensate for the physical impairments (loss, inter-symbol interference (ISI), reflection, etc.), achieving a PCB backplane trace of 1 m, which is typically needed for interconnecting line cards, is challenging. Although

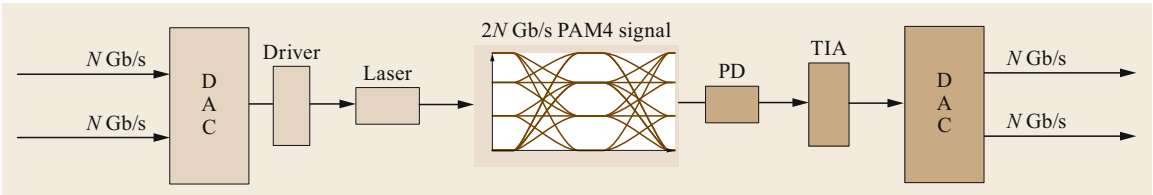

Fig. 21.8 Pulse-amplitude modulation-based link (specifically PAM4, four levels). Leveraging of technological developments for the line side (long-haul), in the form of high-speed ADC (analog-to-digital converter) and DAC (digital-to-analog converter) design, allows increased bandwidth with the same baud rate and number of optical components

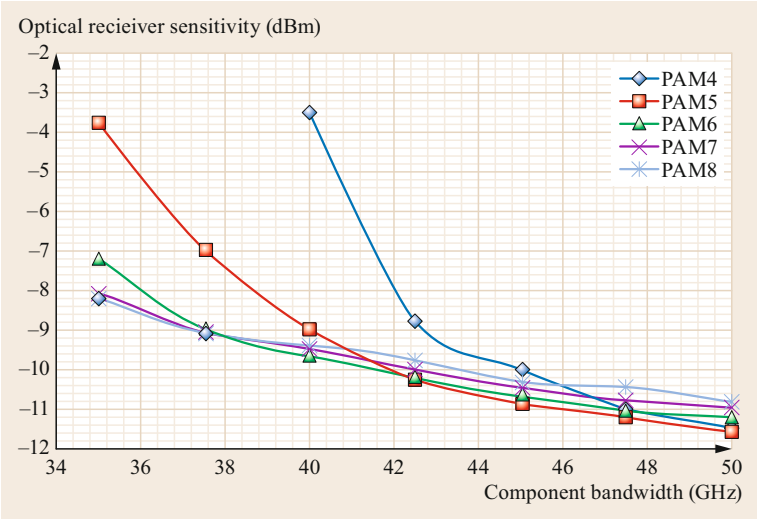

Fig. 21.9 Achievable optical receiver sensitivity using different PAMs with a net throughput of 200 Gb/s for a pragmatic short-reach optical system (thermal noise and bandwidth limited) (after [21.9])

high-speed I/O can scale beyond 100 Gb/s, the electrical reach may be too short even for chip-to-module connection with good energy efficiency. This may necessitate the closer integration of optics to electronics.

Next Generation: Inter-Rack Interconnect

The next-generation electrical I/O will be PAM4-based, with linear interfaces for the optical transmitter and receiver. Technologies developed for modern coherent systems used for dense wavelength-division multiplexing (DWDM) line-side optics (such as linear laser drivers (LD) and linear transimpedance amplifiers (TIAs) at the front end and high-speed analog-to-digital converter (ADC) and digital-to-analog converter (DAC)) form the necessary basis of next-gen high-speed datacenter transceivers. In addition, high-performance DSP capability has emerged for datacenter applications in order to relax the bandwidth and linearity requirements for optical components and to compensate for other link impairments [21.10]. Compared with long-haul transmission, the power consumption of the DSP functions could be significantly reduced for the much shorter-reach datacenter applications with the elimination of chromatic/polarization mode dispersion (CD/PMD) and polarization/phase recovery to meet the power and density requirements. Another potential area of development is low-latency and low-overhead forward error correction (FEC) specifically targeted for datacenter applications.

To scale the performance of direct-detection PAM signaling beyond 100G PAM4, PAM-N technology may be extended by a recently proposed FlexPAM concept [21.9], where irregular PAM (such as PAM5) with flexible and fine spectral efficiency (SE) granularity can be realized by using a single chip with a common DSP architecture. The need for FlexPAM with fine SE granularity can be clearly seen from Fig. 21.9, which shows the achievable receiver sensitivity of different PAMs versus component bandwidth (by using Shannon mutual information theory) to achieve 200 Gb/s throughput for a pragmatic externally modulated short-reach optical system. A dramatic improvement in performance can be achieved with a slight increase in the modulation bandwidth efficiency for a bandwidth-limited system. For components with achievable bandwidth of 40 GHz, the receiver sensitivity can be improved by 5.5 dB using irregular PAM5 compared with regular PAM4. Although using regular PAM8 can achieve similar receiver sensitivity, PAM8 has a 5–6 dB higher link penalty than PAM5, due to level-dependent impairments such as multiple-path interference (MPI) [21.9]. Such a technique may be useful for scaling of short-reach systems to beyond 1 Tb/s.

The use of coherent transmission to increase data rate per channel/wavelength has been previously discussed as another possibility for intra-datacenter interconnect beyond 100 Gb/s per channel [21.11]. The benefits of a coherent approach include the use of a single laser for multiple degrees of freedom (in-phase, quadrature, and polarization) and increased receiver sensitivity. The sensitivity advantage is shown in Fig. 21.10. Operating at an identical net symbol rate of 50 GBd, 480 Gb/s coherent PM-16QAM, with moderate 13 dBm local oscillator (LO) power, can achieve 15 dB better sensitivity than 112 Gb/s PAM4 with direct detection.

On the downside, the DSP functions required to extract the various degrees of freedom at the receiving end consume more power than with a direct-detection receiver. Therefore, it has traditionally been used only

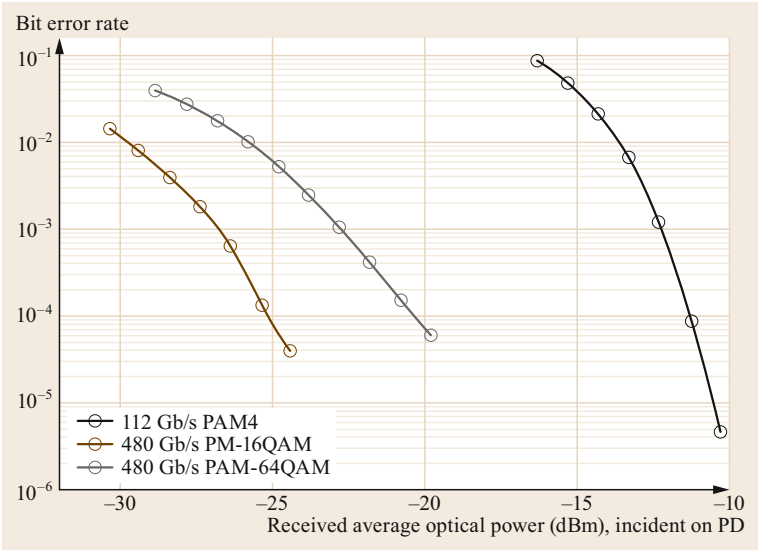

Fig. 21.10 Receiver power sensitivity comparison between coherent and direct detection

in long-haul and metro networks due to its implementation complexity and high power consumption. Other disadvantages are the increased requirements for the quality and control of components (modulator and laser), doubling the number of photodetectors, and test complexity. Finally, although not fundamental, the technology has not been developed for traditional datacenter wavelengths, affecting deployability. So far, disadvantages have outweighed advantages for intra-datacenter use of digital coherent technologies.

However, there have been enormous reductions in cost, power, and density with respect to coherent technologies for metro and long-haul transport systems over the last decade. Figure 21.11 shows power and linear density per gigabit per second as a function of time for datacenter transceivers and coherent transceivers (excludes external erbium-doped fiber amplifiers, EDFAs). Some estimates for target 200/400 Gb/s transceivers include 200 Gb/s (4×50 Gb/s PAM4) in QSFP for the datacenter, and 400 Gb/s in octal (8-channel) small form-factor pluggable (OSFP) for both datacenter and coherent (1×400 Gb/s) transceivers [21.12]. From these plots, it is clear that datacenter and coherent interconnect technologies are converging from a power and density perspective, with this trend expected to further accelerate with shorter coherent reaches (< 100 km, triangles in Fig. 21.11) shedding unnecessary DSP functions. The extension of coherent technologies to shorter reaches will likely enable aggregate interconnect cost reduction (inter- and intra-datacenter reaches) with shared volumes. In addition to relying on Moore's law for power reduction, low-voltage modulators and uncooled LOs need to be investigated to meet the power/density target for datacenter optics. For scaling beyond 1.6 Tb/s, coherent detection could be a more viable option than direct detection.

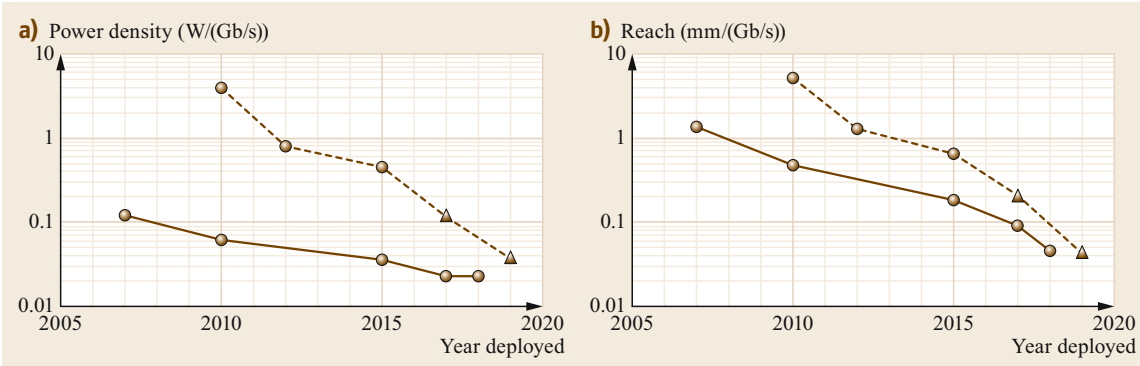

Fig. 21.11a,b Comparison of datacenter (*solid lines*) and coherent transceiver (*dashed lines*) power per Gb/s and linear density per Gb/s. *Triangles* indicate shorter reaches (< 100 km)

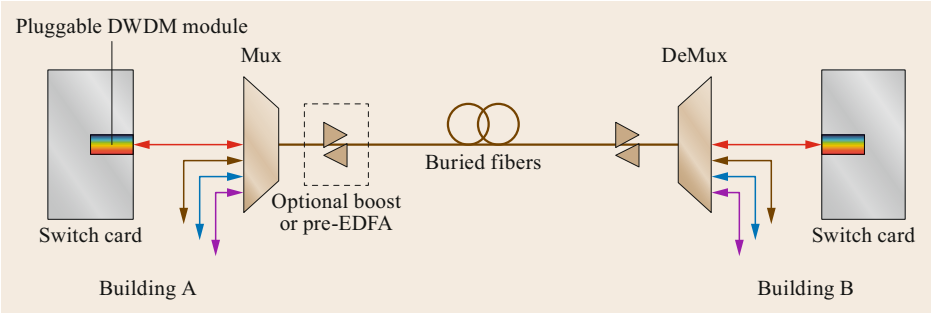

Fig. 21.12 Illustration of pluggable DWDM technique for intra-campus interconnection

Another option for increasing the aggregate data rate per interface is to increase the number of wavelength channels. With this approach, WDM integration will be critical, with three possible approaches: silicon photonics, monolithic III–V, or a hybrid solution. For silicon photonics, the main challenges are link budget, packaging, and optical and electrical integration. Grating and edge couplers with lower loss and improved alignment tolerance are essential to improving performance, yield, and cost. Electrical integration will minimize parasitics to reduce power and size. Silicon photonics may leverage the packaging technologies and techniques of traditional CMOS (f2 f, 2-D, 2.5-D, 3-D integration), giving it critical advantages. Monolithic III–V photonic integrated circuits need fundamental improvements in epitaxial growth and fabrication. Reduction and/or identification of defects early in the III–V process may be aided by data mining and better analytics.

Intra-Campus Interconnect

There are always trade-offs among spectral efficiency, power consumption, path diversity, and cabling complexity. For the intra-building network, a connection-rich mesh topology is desirable; hence, coarse wavelength-division multiplexing (CWDM) with lower spectral efficiency has been the technology of choice for achieving lower power, cheaper transceiver cost, and a richer network fabric. On the other hand, at higher aggregation layers such as the inter-building network, bandwidth is more concentrated over point-to-point links, and dark fiber is more expensive to deploy; hence, DWDM with higher spectral efficiency is preferred. As datacenter traffic continues to grow, the quantity of fibers required from building to building becomes difficult to scale and manage [21.13]. More spectrally efficient and pluggable DWDM technology is needed to scale the intra-campus bandwidth. The use of pluggable DWDM technology not only greatly reduces the number of fibers needed (e.g., a 100 GHz-spaced 48-channel DWDM system reduces the required number of fiber connections by a factor of 48), but also allows simpler fiber management. The higher optical link budget supported by the use of optical amplifiers may also enable new network-level functionalities such as larger campuses or intermediate switching nodes.

Density and cost are two key factors that determine the applicability of DWDM technologies for campus applications. Additionally, to facilitate deployment and inventory management, full-band wavelength tunability is highly desirable. DWDM technology requires stable wavelength grids, and this often necessitates temperature-stabilized (thermoelectric (TE)-cooled) packaging. Innovative laser chip and package designs are needed to bring down cost and power consumption, especially for full-band tunable lasers.

21.1.4 Throughput Requirements and Traffic Characteristics

Figure 21.13 shows the traffic generated by servers inside Google's datacenter, which has been doubling every 12–15 months [21.1]. This is much faster than the bandwidth growth of private backbone WANs, with 40–50% growth every 12 months [21.2].

Bandwidth provisioning to servers is a trade-off between performance (latency, energy) and cost efficiency. While various software techniques can be used to allocate resources more efficiently [21.14, 15] so as to extract the most value and efficiency of compute and storage resources, the networks that interconnect them must be correctly balanced. Otherwise, one or two resources would limit the value that could be extracted from other resources, or could cause some resources to be idle, and that would increase system costs. According to Amdahl's balanced system law, a system should have 1 Mb/s of I/O for every 1 MHz of computation in a parallel computing environment. Ideally, the network bandwidth would be slightly over-provisioned relative to both compute and storage to ensure there are no issues in efficiently connecting compute and storage.

Unlike compute and storage, networking bandwidth is a distributed resource. Despite the availability of software techniques to improve the dynamic allocation of networking bandwidth at the datacenter scale [21.16–

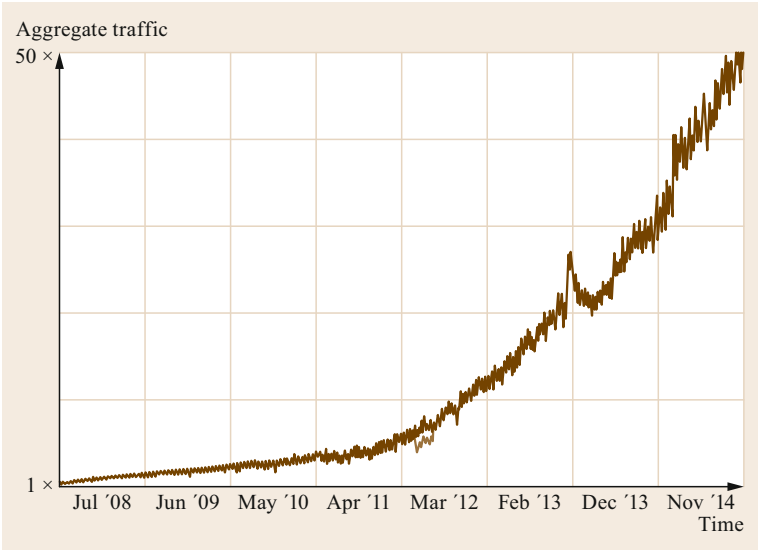

Fig. 21.13 Google's traffic growth generated by servers inside datacenters

18], reliable and scalable networking bandwidth sharing remains challenging even in network topologies with no oversubscription. Empirically, the bandwidth to the server can often be adjusted at the ToR by changing the oversubscription ratio and monitoring the utilization of the uplinks. For a rack of m servers, each capable of 100 Gb/s of burst bandwidth, if the ToR has $m \times 100$ Gb/s downlinks and $n \times 100$ Gb/s uplinks, the average bandwidth available to each server is $(n/m) \times 100$ Gb/s (with an oversubscription of $m : n$).

The total computing and communication capacity required for each target application also vary widely. Consider Google Search as an example: it often touches 50+ services and 1000+ servers with distributed data stored in memory or flash. Caching service (memkeyval) requires a large amount of networking bandwidth for copying of large files, while score and sort (websearch) consumes little network bandwidth [21.14]. With the adoption of flash today and other nonvolatile memory (NVM) technologies with large I/O capability in the future, designing a network to efficiently maximize the performance of compute and storage resources will remain a key challenge.

21.1.5 Energy Efficiency of Datacenter Network

With the unprecedented growth in users and traffic, datacenters continue to expand to house more machines and networking gear. The cost of power and its associ-

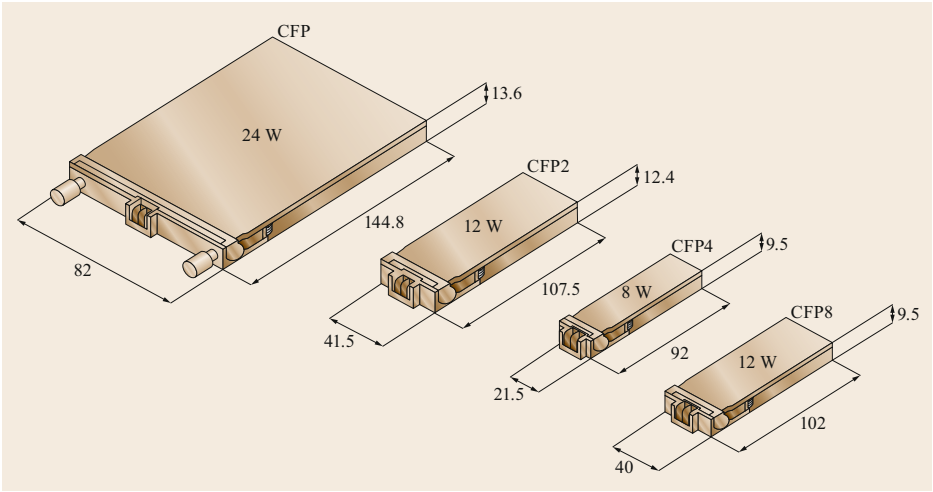

Fig. 21.14 Various form factors for 100 GbE pluggable transceivers and corresponding power envelope (www.cfp-msa.org)

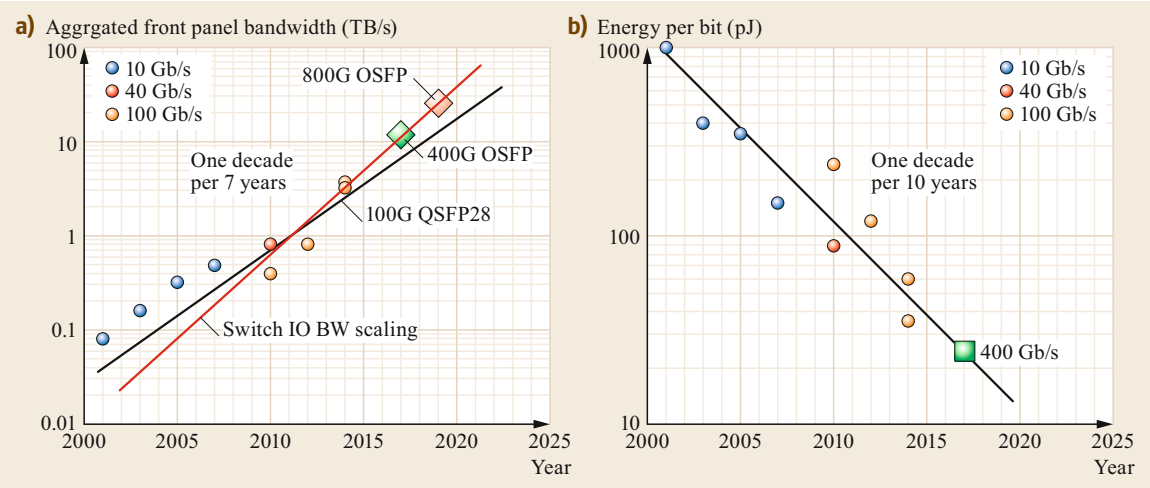

Fig. 21.15 **(a)** Front panel bandwidth of pluggable datacenter optical transceiver; **(b)** energy to transport a bit over 2 km intra-datacenter

ated delivery and cooling are becoming significant factors in the total expenditures of large-scale datacenters. Power consumption inside the datacenter is distributed in the following order [21.19]: servers consume around 33%, memory 30%, power delivery and building cooling 20%, disks 10%. Over the past 10 years, significant effort has been spent in improving energy efficiency, from the individual components to building-level design. Besides the improvement in cooling and power delivery systems, low-power hardware components and energy-proportional designs enable energy savings for processors [21.20], memory [21.21], storage devices, and networking systems [21.22].

The networking portion constitutes only 7% of the entire power envelope of the datacenter [21.19]. The portion of networking attributable to optics is a smaller subset of that power (half or less). Thus, a reduction in networking power consumption will not yield a dramatic improvement in overall datacenter power efficiency and corresponding reduction in CapEx and OpEx. However, the reduction of power in optical modules is still important for increasing front panel linear density and reducing the overall datacenter fabric footprint and cost. In practice, the switch line card/system's form factor is limited by the transceiver's power consumption and the electrical feeds to the transceiver. The size constraint on optics comes from the front panel linear density ((Gb/s)/mm) of the line card, where all the bandwidth of the switch ASIC(s) housed on the line card must be extracted via the front panel. Figure 21.14 shows some pluggable optical transceiver standard form factors for 100 GbE and the associated power consumption of each. Obviously, with the smaller form-factor solutions with small connector pin pitch, more bandwidth can be brought out per linear dimension of the line card. However, the small-form-factor optical transceiver often has a smaller power envelope because of a more limited surface area for heat extraction.

Figure 21.15a shows the switch I/O bandwidth trend line and the front panel bandwidth achieved for 10, 40, and 100 Gb/s. Pluggable optics technology saw a tenfold improvement over 7 years, with an increase in electrical I/O speed (from 10, 25, to 50 Gb/s) and a decrease in connector pin pitch from 0.8 mm for (40 Gb/s)/(100 Gb/s) QSFP to 0.6 mm for 400 Gb/s OSFP. Thus far, through the use of 400 Gb/s transceivers, there has been no scaling problem for datacenter pluggable optics to effectively match and extract the switch capacity.

Delving into what determines power at the component level can be quite complex, as the type of device and the design details give rise to a wide range of power numbers. Because of the analog nature of optical components, even with the advances in high-speed SerDes technologies, the power efficiency of datacenter optical transceivers improves only 10× every 10 years (Fig. 21.14b), which is slower than the growth of switch bandwidth capacity and front panel density. Improvements in thermal design for pluggable optics has helped to sustain bandwidth density improvements in the past. More aggressive thermal management and creative copackaging of electrical and optical devices will be needed to meet the more stringent bandwidth density requirements for port speed of 400 Gb/s and beyond.

21.2 Inter-Datacenter Network

While economies of scale suggest that datacenters should be as large as possible, typically restricted by the amount of power available for the site, datacenters should also be distributed across the planet for fault tolerance, latency locality, and better user experience. The rapid adoption of cloud services also drives a new set of requirements and challenges for wide area networking (WAN) in terms of capacity, accessibility, and flexibility. To meet the two sets of requirements, Google has built two separate WAN networks: public back-end backbone (B2) and private back-end backbone (B4). Figure 21.16 shows Google's wide area datacenter network topology [21.22]. Parallel to the private back-end network is a user-facing public backbone (B2), which end users interconnect to through peering, transit, or direct connection to gain access to various cloud services. In order to support better user experience with lower latency, the fiber topology of the public backbone is meshier, with more stringent availability performance. The private back-end backbone (B4) provides connectivity among datacenters only, to support large-scale data copies, remote storage access, and end user data replication for reliability. The private back-end backbone network is usually architecturally simple, with point-to-point links, but requiring much higher bandwidth to support large-volume internal application traffic [21.2].

The inter-datacenter private back-end network employs scarce and expensive long-distance fiber for transmission. The public-facing transport network, in addition, contains many high-capacity metro transport links to interconnect with other carriers' networks. Metro transport networks also serve to connect carrier networks to edge cache systems used by content providers in order to improve content distribution experiences with low latency without burdening the expensive backbone transport network. In recent years, fast-growing bandwidth-intensive services such as YouTube and Netflix and other cloud applications are accelerating the deployment of edge cache and metro optical transport systems [21.22].

21.2.1 Interconnect Network Architecture Evolution for WAN

With traffic engineering or buffering at the edge, the aggregated WAN traffic entering and leaving datacenters can be controlled based on capacity availability [21.2]. Unlike fibers inside datacenters, long-haul transmission fibers are expensive and time-consuming to build or acquire. Datacenters are often located in remote areas, which drives the requirement for a range of reaches all the way from a few hundred kilometers to 6000 km. Given the space and power-hungry nature of regeneration nodes, the focus of inter-datacenter links is on high-capacity, high-spectral efficiency, ultra-long-haul transmission systems. Considerable hardware and soft-

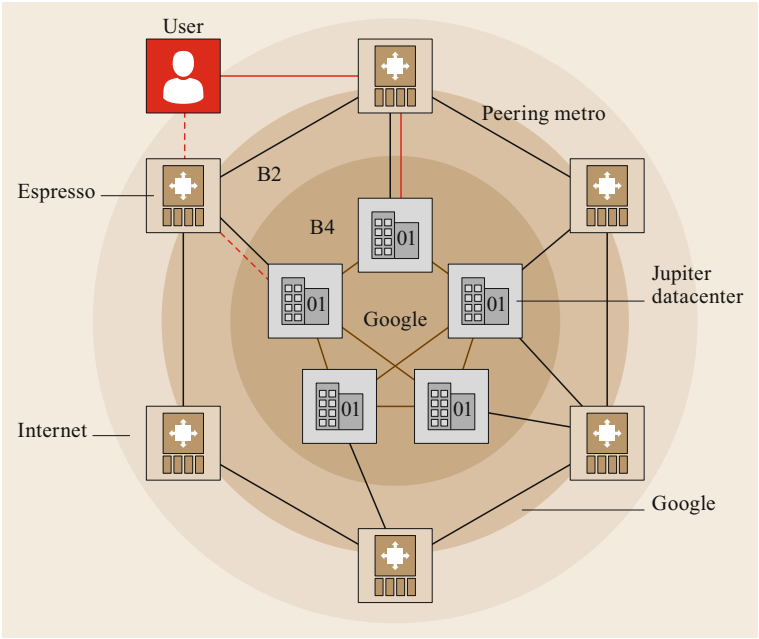

Fig. 21.16 Google's wide area network (WAN) architecture (after [21.23])

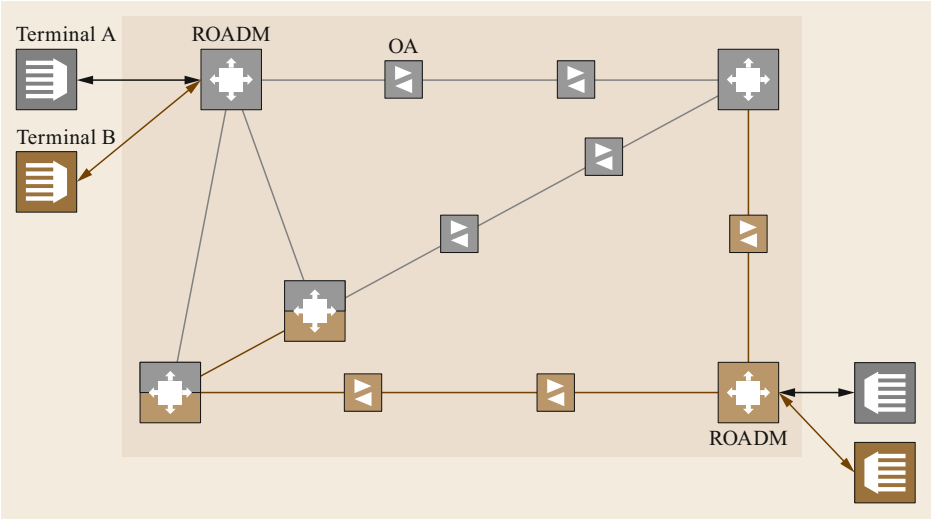

Fig. 21.17 Reconfigurable optical add-drop multiplexer (ROADM)-enabled mesh transport architecture (after [21.24]) for long-haul networks. (OA: optical amplifier)

ware [21.2] innovation and effort have been directed toward improving spectral efficiency and utilization of WAN links.

From a transport topology point of view, while metro networks (the network that connects our datacenters and peering POPs to our B2/B4 backbone networks) still use point-to-point link-based architectures, long-haul (LH) networks have evolved from traditional point-to-point link-based architectures to mesh architectures enabled by colorless, directionless, and contentionless (CDC) ROADM (reconfigurable optical add-drop multiplexer). With software control, CDC ROADMs can add-drop any wavelength from any direction to any transponder without contention in the add/drop path, thereby enabling an unconstrained dynamic network at the physical layer.

The CDC ROADM-based mesh network architecture has a number of advantages: (a) higher network utilization efficiency, (b) shorter reconfiguration time, and (c) operational simplicity:

(a) Network efficiency comes from a couple of areas. An arbitrary wavelength-routed mesh provides superior CapEx and OpEx performance relative to terminated point-to-point links. Secondly, CDC ROADM architecture enables bandwidth reconfiguration to meet the variation in network demand, unlike the traditional point-to-point network topology, where additional fibers and amplifiers must be deployed to transport those wavelengths or in the event of a bandwidth demand surge.
(b) CDC ROADM architecture enables optical layer reconfiguration around links under repair to reduce the mean time between failures, thus increasing the overall network availability. CDC-ROADM-enabled mesh networks could provide section-by-section (between ROADM nodes) $1:n$ share optical protection as well as dynamic mesh restoration (the capability to dynamically identify and activate new routes at the time of a failure). In particular, $1:n$ protection allows us to protect optical layer services without the need for reserving 50% of the network capacity, which is required with $1+1$ protection used in point-to-point-based networks.
(c) The use of CDC ROADM could enable end-to-end *touchless* provisioning and activation of network bandwidth without requiring technicians on-site to provision the connection of each circuit. Moreover, the sparing of transponders in the CDC add/drop node becomes a simple software command to activate a prepopulated spare instead of physically reseating a new card and reprovisioning it.

In addition to the introduction of CDC ROADM for a more flexible optical layer, disaggregation of key functional blocks is emerging as an important trend [21.24]. Traditionally, LH transport systems have been provided by system vendors as proprietary solutions, where the terminal optics (i.e., line card-based optical transponders) will work only with their proprietary line systems. The key functional blocks, such as ROADM and OA, can only be managed by their proprietary line management system, and have numerous control and equalization loops at the link level. Such an integrated solution has the advantage of well-defined transport performance, but it also has challenges in interoperability due to lack of common specifications, flexibility in vendor and technology selection, and

centralized network management and automation. The disaggregation of key transport functional blocks with a well-defined I/O interface addresses these problems. The first phase of disaggregation is to decouple terminal optics from the line system, which we have deployed for several years. In the second phase, ROADM disaggregation into key functional blocks provides a number of benefits, which include (a) modularity at the degree and link levels and the use of multivendor line systems, (b) faster insertion cycles for new subsystem building blocks, and (c) a greatly simplified control and management plane architecture by embodying modern software-defined networking (SDN) principles.

The ROADM technology has also been evolving from fixed grid (typically 50 GHz or 100G fixed DWDM grid) toward a flexible grid (with grid granularity reduced from 50 to 12.5 GHz). The introduction of flexible-grid ROADM could reduce the guard band between channels (e.g., 100 Gb/s quadrature phase-shift keying QPSK over 37.5 GHz channel spacing) and pack multiple channels together with minimal channel guard bands through the Nyquist pulse-shaping technique [21.25], improving spectral efficiency in long-haul transmission fiber.

21.2.2 LH Interconnect Bandwidth Scaling

Since long-haul fiber is a very scarce resource and time-consuming to deploy, in order to meet the increased bandwidth demands of private backbone WAN, the bandwidth capacity scaling of long-haul is achieved mainly by increasing the single-fiber capacity. Figure 21.18 shows an increase in single-fiber capacity from 3.2–40 Tb/s with the following techniques.

Dense Wavelength-Division Multiplexing (DWDM)

DWDM channels with 50 GHz spacing (with up to 96 channels over the extended C-band optical amplifier bandwidth) have been used for LH transport systems with 40–200 Gb/s per channel. Very tight laser wavelength control is required for DWDM systems with small wavelength spacing, because any wavelength drift will cause the system to experience higher loss and spectrum distortion from the filtering effects of Mux/Demux (multiplex/de-multiplex) (and cascaded ROADMs), as well as larger DWDM crosstalk from the neighboring channels. Wavelength control accuracy has improved from early DWDM systems with ±5 GHz accuracy, to current accuracy of ±2.5 GHz, in order to reduce the filtering and crosstalk penalties. Flexible-grid or gridless super-channels may be used to increase the optical spectrum utilization and thus the overall single-fiber capacity.

Four-Dimensional Modulation/Multiplexing

Each optical carrier has four degrees of freedom: X and Y-polarization, in-phase, and quadrature phase components. Polarization-multiplexed (PM) QPSK and PM M-QAM (quadrature amplitude modulation) such as PM-8QAM and PM-16QAM utilize all four dimensions of an optical carrier, and thus can increase the modulation spectral efficiency in terms of bits per symbol by a factor of 4 as compared with the single-polarization direct intensity modulation format with one dimension of an optical carrier only.

Digital Coherent Detection

Digital signal processing (DSP) has enabled coherent detection, where both polarization demultiplexing and

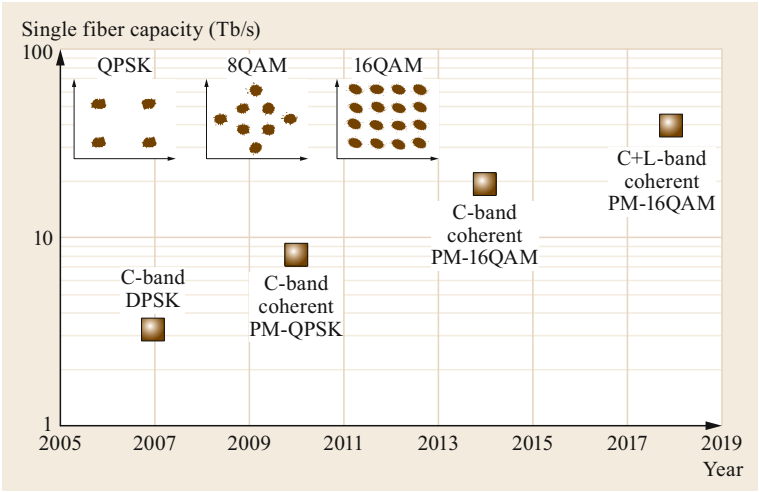

Fig. 21.18 LH network single-fiber capacity (with reach > 600 km) scalings. (PM: polarization-multiplexed, DPSK: differential phase-shift keying; QPSK: quadrature phase-shift keying; QAM: quadrature amplitude modulation)

carrier phase recovery are carried out in the digital domain by:

1. linearly mapping the incoming optical signal field into an electrical signal through a polarization- and phase-diverse coherent detection front end
2. digitizing the detected electrical signal; and
3. using a 2×2 multi-input and multi-output (MIMO) equalizer to recover the polarization and using a feed-forward-based phase estimation algorithm to track fast-changing optical phases.

Digital coherent detection not only encodes information into the full four dimensions of an optical carrier; it also enables electrical compensation of optical chromatic dispersion and polarization mode dispersion, which are the major impairments for high-speed optical transmission.

C+L Band

C-band (1528–1565 nm) and L-band (1570–1610 nm) are two communication bands that coincide with the third transmission window of silica fiber and the wavelength range of EDFAs. EDFAs can be made to operate at both C-band and L-band, and fiber loss in the L-band is still very low. When C-band capacity becomes insufficient, L-band is the next natural spectral region to double the single-fiber capacity.

Eventually, fiber nonlinear effects and OSNR (optical signal-to-noise ratio) requirements for higher-spectral-efficiency (i.e., higher bits per symbol) modulation formats [21.27] limit the single-fiber capacity. Figure 21.19 shows the estimated transmission reach limit for different higher-order modulation formats (with different spectral efficiency rates) at a fixed symbol rate of 40 GBd and fixed channel spacing of 50 GHz. The transmission reach rapidly decreases as the modulation spectral efficiency increases. Both hardware [21.26, 28, 29] and software techniques [21.2] designed to improve the resource utilization of network capacity have been investigated in recent years. Rate-adaptive optics [21.26], which optimizes the fiber capacity based on fiber route, can effectively increase network resource utilization: shorter routes using higher-SE (spectral efficiency) modulation formats (and thus higher capacity), and longer routes using lower-SE modulation formats (less capacity). Very fine-granularity rate-adaptive optical transceivers could be realized by using either the time-domain hybrid QAM technique [21.26] or the probabilistic constellation-shaping (PCS) technique [21.29]. The time-domain hybrid QAM has the advantage of simpler implementation, while the PCS offers better performance. As discussed earlier, the overall transport capacity can also be increased with Nyquist pulse-shaping-based super-channel techniques to increase spectrum utilization [21.25].

SDN has advanced greatly over the past 10 years. Not only does it enable the network-level programmability to make the above rate-adaptive optical transmission feasible, but the centralized optical control plane is also essential for allocating bandwidth among different services based on application priority. As a result, instead of over-provisioning and underutilization of bandwidth, the utilization of expensive long-haul links can be improved to near 100% [21.2].

Metro to LH: IP+DWDM to IPoDWDM

Traditional terminal optics-based transport technology has been used for metro and LH backbone networks. With this technique, the transport interface rate (i.e.,

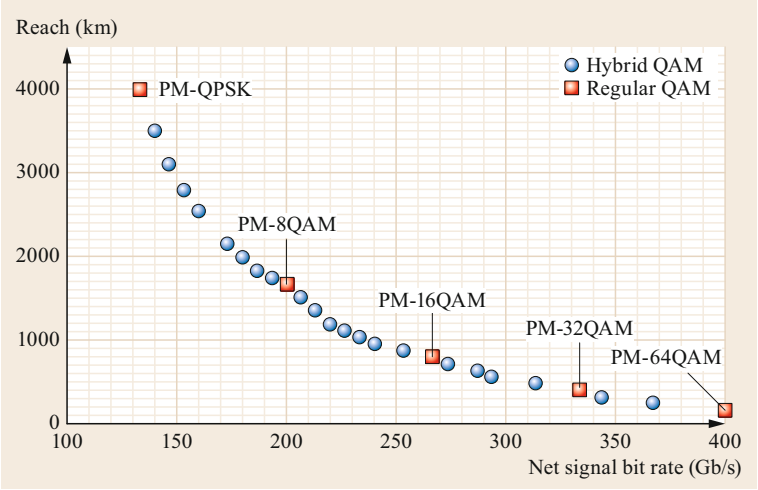

Fig. 21.19 Estimated transmission reach for different modulation formats all operating at 40 GBd symbol rate and 50 GHz channel spacing (after [21.26])

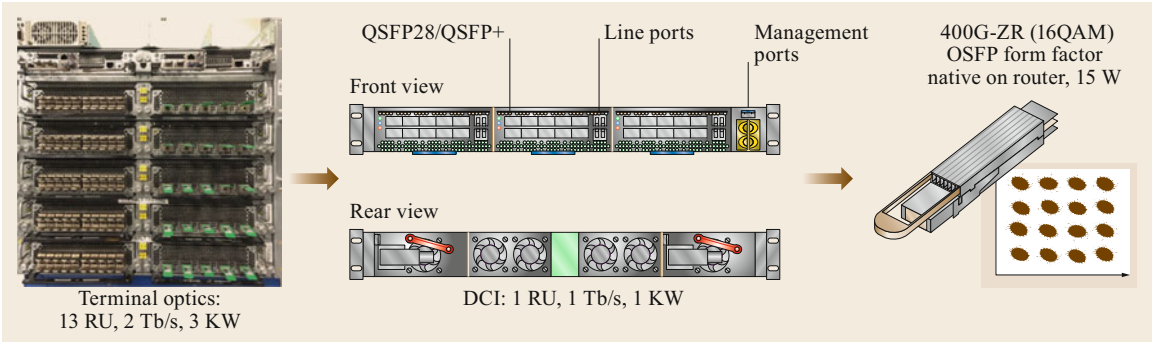

Fig. 21.20 Metro transport optics evolution (an example)

the line-side interface rate) is decoupled from the router interface rate (i.e., the client rate), with client optics interconnecting the router and the transport equipment. Such a technique was adopted mainly for two reasons: (1) the router interface rate is slower than the transport line-side interface rate, and (2) the router technology scales well with Moore's law, with much lower power consumption than transport optics. Thus the speed and density of router ports have evolved much faster than the transport optics. With the decoupling of the transport optics and router on separate systems, the router and transport optics can scale at their own pace to achieve the best overall economics. As the router interface rate increases to be comparable to DWDM line-side interface rates (100 Gb/s and beyond), not only is the cost of client optics greatly increased (relative to the line-side optics), but the role of traffic grooming using client optics is also reduced. On the other hand, with the advance of digital coherent detection, the transport technology evolution becomes largely dictated by the coherent DSP power, which has scaled well with Moore's law—the same as the routers. Therefore, starting from 100 Gb/s, a fundamental paradigm shift has taken place in the metro and LH networks: we are starting to move away from the traditional IP+DWDM architecture, toward a converged IP (internet protocol) over DWDM (IPoDWDM) architecture, by eliminating the client optics and putting the transport optics directly into the router interface ports. Such an IPoDWDM architecture not only saves power, cost, and space, but also simplifies network management and control.

Figure 21.20 shows the evolution of metro transport optics from traditional terminal optics to a lean datacenter connection interconnect (DCI) box with simplified management interface, and eventually to the pluggable 400G coherent module (i.e., the 400G-ZR, an Optical Internetworking Forum (OIF) standard which is under active development). As shown in the figure, traditional line card-based terminal optics occupies 13 rack units and consumes 3 kW of power for 2 Tb/s of transport capacity, whereas the 1 Tb/s DCI box occupies one rack unit and consumes 1 kW of power. When the 400G-ZR pluggable transceiver is available in 2019, its coherent optoelectronics is expected to consume < 15 W power and will be realized in an OSFP module [21.30], the same form factor that will be used for 400 Gb/s client optics. Such dramatic power and size reduction is made possible by the continuous power reduction with CMOS technology (40 to 28, to 14, to 7 nm) and the advancements in photonic integrated circuits (PICs), including both InP- and silicon photonics-based integration technologies.

21.3 Conclusion

The evolution of the datacenter network and its photonic technologies over the past decade has been driven by tremendous advances in hardware and software technologies, both within (intra-datacenter) and between (inter-datacenter) networks. The tenets of SDN will continue to have a major impact by enabling scalable, dynamic networks with better programmability and efficiency. The line-side technologies (wavelength-division multiplexing, digital signal processing) have migrated from traditional telecom applications to help drive the bandwidth growth of datacenter applications. Moving forward, the continuing growth of datacenters and cloud applications will become the key technology driver for more aggressive reductions in power, cost, and density through intimate integration of electronics and photonics.

References

21.1 A. Singh, J. Ong, A. Agarwal, G. Anderson, A. Armistead, R. Bannon, S. Boving, G. Desai, B. Felderman, P. Germano, A. Kanagala, H. Liu, J. Provost, J. Simmons, E. Tanda, J. Wanderer, U. Hölzle, S. Stuart, A. Vahdat: Jupiter rising: A decade of clos topologies and centralized control in Google's datacenter network, Commun. ACM **59**(9), 88–97 (2016)

21.2 S. Jain, A. Kumar, S. Mandal, J. Ong, L. Poutievski, A. Singh, S. Venkata, J. Wanderer, J. Zhou, M. Zhu, J. Zolla, U. Hölzle, S. Stuart, A. Vahdat: B4: Experience with a globally deployed software defined WAN. In: *Proc. ACM SIGCOMM* (2013), https://doi.org/10.1145/2486001.2486019

21.3 A. Vahdat, H. Liu, X. Zhao, C.J. Johnson: The emerging optical data center. In: *Proc. Opt. Fiber Conf.* (2011), https://doi.org/10.1364/OFC.2011.OTuH2

21.4 J. Kim, W.J. Dally, D. Abts: Flattened butterfly: A cost-efficient topology for high-radix networks. In: *Proc. 34th Ann. Int. Symp. Comput. Archit.* (2007), https://doi.org/10.1145/1250662.1250679

21.5 C. Guo, G. Lu, D. Li, H. Wu, X. Zhang, Y. Shi, C. Tian, Y. Zhang, S. Lu: BCube: A high performance, server-centric network architecture for modular data centers. In: *Proc. ACM SIGCOMM* (2009), https://doi.org/10.1145/1592568.1592577

21.6 C. Clos: A study of nonblocking switching networks, Bell Syst. Tech. J. **32**(2), 406–424 (1953)

21.7 W. Dally, B.P. Towles: *Principles and Practices of Interconnection Networks* (Morgan Kaufmann, New York 2004)

21.8 M. Al-Fares, A. Loukissas, A. Vahdat: A scalable, commodity, data center network architecture. In: *Proc. ACM SIGCOMM* (2008), https://doi.org/10.1145/1402946.1402967

21.9 X. Zhou, H. Liu: Constellation shaping: Can it be useful for datacenter reach communication. In: *Eur. Conf. Opt. Commun.* (2017), Paper WS4

21.10 R. Urata, H. Liu: Datacenter interconnect and networking: Present state to future challenges. In: *Opt. Interconn. Conf.* (2016)

21.11 X. Zhou, H. Liu: Pluggable DWDM: Considerations for campus and metro DCI applications. In: *Eur. Conf. Opt. Commun.* (2016), Paper WS3

21.12 R. Urata, H. Liu, X. Zhou, A. Vahdat: Datacenter interconnect and networking: From evolution to holistic revolution. In: *Opt. Fiber Conf.* (2017), Paper W3G.1

21.13 U. Hölzle: Ubiquitous cloud requires a revolution in optics. In: *Opt. Fiber Commun. Conf.* (2017), Plenary talk

21.14 D. Lo, L. Cheng, R. Govindaraju, P. Ranganathan, C. Kozyrakis: Heracles: Improving resource efficiency at scale. In: *Proc. 42nd Ann. Int. Symp. Comput. Archit.* (2015), https://doi.org/10.1145/2749469.2749475

21.15 V. Jeyakumar, A. Kabbani, J. Mogul, A.Vahdat: Flexible network bandwidth and latency provisioning in the datacenter, arXiv:1405.0631 [cs.NI] (2014)

21.16 A. Shieh, S. Kandula, A. Greenberg, C. Kim, B. Saha: Sharing the data center network. In: *Proc. 8th USENIX Conf. Netw. Syst. Des. Implement. (NSDI)* (2011) pp. 309–322

21.17 H. Rodrigues, J.R. Santos, Y. Turner, P. Soares, D. Guedes: Gatekeeper: Supporting bandwidth guarantees for multitenant datacenter networks. In: *Proc. USENIX WIOV* (2011) p. 6

21.18 T. Benson, A. Akella, D.A. Maltz: Network traffic characteristics of data centers in the wild. In: *Proc. 10th ACM SIGCOMM Conf. Internet Meas. (IMC'10)* (2010), https://doi.org/10.1145/1879141.1879175

21.19 L. Barroso, U. Hölzle: *The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines* (Morgan Claypool, San Rafael 2009)

21.20 L.A. Barroso, U. Hölzle: The case for energy-proportional computing, Computer **40**(12), 33–37 (2007)

21.21 K.T. Mallad, F.A. Nothaft, K. Periyathambi, B.C. Lee, C. Kozyrakis, M. Horowitz: Towards energy-proportional datacenter memory with mobile DRAM. In: *39th Ann. Int. Symp. Comput. Archit. (ISCA)* (2012), https://doi.org/10.1109/ISCA.2012.6237004

21.22 D. Abts, M.R. Marty, P.M. Wells, P. Klausler, H. Liu: Energy proportional datacenter networks. In: *Proc. 37th Ann. Int. Symp. Comput. Archit. (ISCA)* (2010), https://doi.org/10.1145/1815961.1816004

21.23 K.K. Yap, M. Motiwala, J. Rahe, S. Padgett, M. Holliman, G. Baldus, M. Hines, T. Kim, A. Narayanan, A. Jain, V. Lin, C. Rice, B. Rogan, A. Singh, B. Tanaka, M. Verma, P. Sood, M. Tariq, M. Tierney, D. Trumic, V. Valancius, C. Ying, M. Kallahalla, B. Koley, A. Vahdat: Taking the edge off with espresso: scale, reliability and programmability for global Internet peering. In: *ACM SIGCOMM Conf.* (2017), https://doi.org/10.1145/3098822.3098854

21.24 V. Vusirikala: SDN enabled programmable, dynamic optical layer. In: *Eur. Conf. Opt. Commun.* (2017), Paper M.PL 1

21.25 G. Bosco, V. Curri, A. Carena, P. Poggiolini, F. Forghieri: On the performance of Nyquist-WDM terabit superchannels based on PM-BPSK, PMQPSK,PM-8QAM or PM-16QAM subcarriers, J. Lightwave Technol. **29**(1), 53–61 (2011)

21.26 X. Zhou, L.E. Nelson, P. Magill: Rate-adaptable optics for next generation long-haul transport networks, IEEE Commun. Mag. **51**(3), 41–49 (2013)

21.27 R.J. Essiambre, R.W. Tkach: Capacity trends and limits of optical communication networks, Proc. IEEE **100**(5), 1035 (2012)

21.28 P.J. Winzer: Spatial multiplexing in fiber optics: The 10x scaling of metro/core capacities, Bell Labs Tech. J. **19**, 22 (2014)

21.29 X. Zhou, C. Xie: *Enabling Technologies for High Spectral-efficiency Coherent Optical Communication Networks* (Wiley, New York 2016)

21.30 OSFP MSA Group: http://osfpmsa.org/

Hong Liu

Google Inc.
Mountain View, CA, USA
hongliu@google.com

Hong Liu is a Distinguished Engineer at Google Technical Infrastructure, where she is involved in the system architecture and interconnect for a large-scale computing platform. Her research interests include interconnection networks, high-speed signaling, optical access, and metro design. Prior to joining Google, Hong was a Member of Technical Staff at Juniper Networks. Hong received her PhD in Electrical Engineering from Stanford University.

Ryohei Urata

Google Inc.
Mountain View, CA, USA
ryohei@google.com

Ryohei Urata is currently a member of technical staff in the Platforms Optics Group, where he is responsible for datacenter optical technologies and corresponding roadmaps. Prior to joining Google, he was a research specialist at NTT Photonics Laboratories, Japan. He has over 125 patents, publications, and presentations in the areas of optical interconnects, switching, and networking. He received his MS and PhD degrees in Electrical Engineering from Stanford University.

Xiang Zhou

Google Inc.
Mountain View, CA, USA
xzhou@google.com

Xiang Zhou is currently with Google Datacenter optics group, leading 400Gb/s and beyond optical interconnect technologies and roadmap development. He received his PhD from BUPT (1999). He has extensive publications, including several record-setting results. He is the author of several book chapters and the holder of over 40 US patents. He has served on the Editorial Board of Optics Express and on the Program Committee of a variety of technical conferences.

Amin Vahdat

Google Inc.
Mountain View, CA, USA
vahdat@google.com

Amin Vahdat is a Google Fellow and Technical Lead for networking at Google. He has contributed to Google's data center, wide area, edge/CDN, and cloud networking infrastructure. Vahdat received his PhD from UC Berkeley in Computer Science, is an ACM Fellow and a past recipient of the NSF CAREER award, the Alfred P. Sloan Fellowship, and the Duke University David and Janet Vaughn Teaching Award.