Chapter 1
Introduction to Optical Interconnects in Data Centers

Christoforos Kachris, Keren Bergman, and Ioannis Tomkos

1.1 Introduction

Over the last few years, the exponential increase of the Internet traffic, mainly driven from emerging applications like streaming video, social networking and cloud computing has created the need for more powerful warehouse data centers. These data centers are based on thousands of high performance servers interconnected with high performance switches. The applications that are hosted in the data center servers (e.g., cloud computing applications, search engines, etc.) are extremely data-intensive and require high interaction between the servers in the data center. This interaction creates the need for high bandwidth and low latency communication networks between the servers in the data centers. Furthermore, these data centers must comply with low power consumption requirements in order to reduce the total operating cost.

1.2 Architecture of Data Center Networks

Figure 1.1 shows the high level block diagram of a typical data center. A data center consists of multiple racks hosting the servers (e.g. web, application, or database servers) connected through the data center interconnection network. When a request is issued by a user, then a packet is forwarded through the Internet to the front end of

Ch. Kachris (✉) • I. Tomkos
Athens Information Technology, Athens, Greece
e-mail: kachris@ait.edu.gr; itom@ait.edu.gr

K. Bergman
Department of Electrical Engineering
Columbia University, New York, NY, USA
e-mail: bergman@ee.columbia.edu

C. Kachris et al. (eds.), *Optical Interconnects for Future Data Center Networks*,
Optical Networks, DOI 10.1007/978-1-4614-4630-9_1,

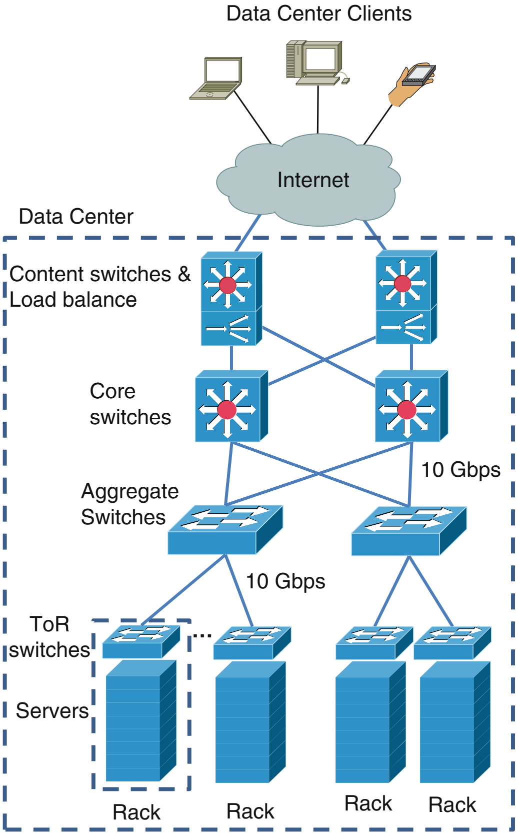


Fig. 1.1 Architecture of current data center network

the data center. In the front end, the content switches and the load balance devices are used to route the request to the appropriate server. A request may require the communication of this server with many other servers. For example, a simple web search request may require the communication and synchronization between many web, application, and database servers.

Most of the current data centers are based on commodity switches for the interconnection network. The network is usually a canonical fat-tree 2-Tier or 3-Tier architecture as it is depicted in Fig. 1.1 [7]. The servers (usually up to 48 in the form of blades) are accommodated into racks and are connected through a *Top-of-the-Rack* Switch (ToR) using 1 Gbps links. These ToR switches are further inter-connected through *aggregate* switches using 10 Gbps links in a tree topology. In the 3-Tier topologies (shown in the figure) one more level is applied in which the aggregate switches are connected in a fat-tree topology using the *core* switches either at 10 Gbps or 100 Gbps links (using a bundle of 10 Gbps links). The main advantage of this architecture is that it can be scaled easily and that it is fault tolerant (e.g., a ToR switch is usually connected to 2 or more aggregate switches).

However, the main drawback of these architectures is the high power consumption of the ToR, aggregate and core switches, and the high number of links that are required. The high power consumption of the switches is mainly caused by the power consumed by the Optical-to-Electrical (O-E) and E-O transceivers and the electronic switch fabrics (crossbar switches, SRAM-based buffers, etc.).

Another problem of the current data center networks is the latency introduced due to multiple store-and-forward processing. When a packet travels from one server to another through the ToR, the aggregate and the core switch, it experiences significant queuing and processing delay in each switch. As the data centers continue to increase to face the emerging web applications and cloud computing, more efficient interconnection schemes are required that can provide high throughput, low latency, and reduced energy consumption. While there are several research efforts that try to increase the required bandwidth of the data centers that are based on commodity switches (e.g., using modified TCP or Ethernet enhancements), the overall improvements are constraints by the bottlenecks of the current technology.

1.3 Network Traffic Characteristics

In order to design a high performance network for a data center, a clear understanding of the data center traffic characteristics is required. This section presents the main features of the network traffic in the data centers and discusses how these features affect the design of the optical networks. There are several research papers that have investigated the data center traffic such as the ones presented by Microsoft Research [2, 3, 12]. The data centers can be classified in three classes: university campus data centers, private enterprise data centers, and cloud-computing data centers. In some cases there are some common traffic characteristics (e.g., average packet size) in all data centers while other characteristics (e.g., applications and traffic flow) are quite different between the data center categories. The results presented in these papers are based on measurement of real data centers. The main empirical findings of these studies are the followings:

- *Applications:* The applications that are running on the data centers depend on the data center category. In campus data centers the majority of the traffic is HTTP traffic. On the other hand, in private data centers and in data centers used for cloud computing the traffic is dominated by HTTP, HTTPS, LDAP, and DataBase (e.g., MapReduce) traffic.
- *Traffic flow locality:* A traffic flow is specified as an established link (usually TCP) between two servers. The traffic flow locality describes if the traffic generated by the servers in a rack is directed to the same rack (intra-rack traffic) or if it directed to other racks (inter-rack traffic). According to these studies the traffic flow ratio for inter-rack traffic fluctuates from 10 to 80% depending on the application. Specifically, in data centers used by educational organization and private enterprises the ratio of intra-rack traffic ranges from 10 to 40%. On the

other hand, in data centers that are used for cloud computing the majority of the traffic is intra-rack communication (up to 80%). The operators in these systems locate the servers, which usually exchange high traffic between each other, into the same rack. The traffic flow locality affects significantly the design of the network topology. In cases of high inter-rack communication traffic, high-speed networks are required between the racks while low-cost commodity switches can be used inside the rack. Therefore, in these cases an efficient optical network could provide the required bandwidth demand between the racks while low cost electronic switches can be utilized for intra-rack communication.

- *Traffic flow size and duration:* A traffic flow is defined as an active connection between 2 or more servers. Most traffic flow sizes in the data center are considerably small (i.e., less than 10KB) and a significant fraction of these flows last under a few hundreds of milliseconds. The duration of a traffic flow can affect significantly the design of the optical topology. If the traffic flow lasts several seconds, then an optical device with high reconfiguration time can sustain the reconfiguration overhead in order to provide higher bandwidth.
- *Concurrent traffic flows:* The number of concurrent traffic flows per server is also very important to the design of the network topology. If the number of concurrent flows can be supported by the number of optical connections, then optical networks can provide significant advantage over the networks based on electronic switches. The average number of concurrent flows is around 10 per server in the majority of the data centers.
- *Packet size:* The packet size in data centers exhibit a bimodal pattern with most packet sizes clustering around 200 and 1,400 bytes. This is due to the fact that the packets are either small control packets or are parts of large files that are fragmented to the maximum packet size of the Ethernet networks (1,550 bytes).
- *Link utilization:* According to these studies, in all kinds of data centers the link utilization inside the rack and in the aggregate level is quite low, while the utilization on the core level is quite high. Inside the rack the preferable data rate links are 1 Gbps (in some cases each rack server hosts 2 or more 1 Gbps links), while in the aggregate and in the core network, 10 Gbps are usually deployed. The link utilization shows that higher bandwidth links are required especially in the core network, while the current 1 Gbps Ethernet networks inside the rack can sustain the future network demands.

Although that the qualitative characteristics of the network traffic in the data center remains the same, the amount of network traffic inside the data centers is growing rapidly. Larger data centers are required that can sustain the vast amount of network traffic from the end users due to emerging web applications (e.g., cloud computing) and due to higher data rates that access networks provide.

The amount of network traffic inside the data center growths not only due to larger data centers but also due to higher-performance servers. As more and more processing cores are integrated into a single chip, the communication requirements between servers in the data centers will keep increasing significantly [19]. According to Amdahl's Law for every 1 MHz of processing power we need 1MB of memory and 1Mbps I/O. If we target the current data center servers that have 4 processors

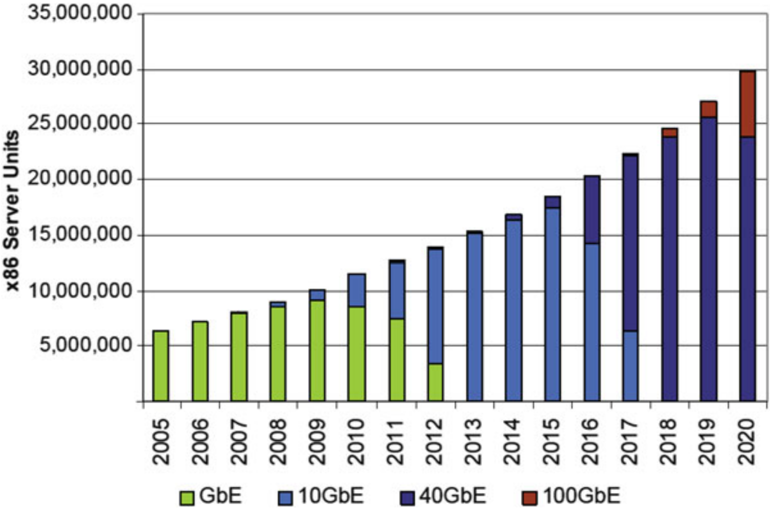


Fig. 1.2 Server Datarate forecast by ethernet connection type, Source: Intel and Broadcom, 2007

running at 2.5 GHz, and each processor has 4 cores the total I/O bandwidth is 40 Gbps per server [22]. If we assume a data center with 100k servers, the total bandwidth requirements is 4 Pbps.

To face this overwhelming growth of bandwidth, service providers around the globe are racing to transform their networks by adopting higher bandwidth links. Analysts estimate a compound annual growth rate (CAGR) of more than 170% from 2011 to 2016 for 100G Ethernet ports as service providers rush to meet this demand [10].

Figure 1.2 depicts the forecast for the server data-rates inside the data centers [6]. As it is shown, while in 2012 only a small portion of the servers is using 40G Ethernet transceivers, it is estimated that by 2017 the majority of the Ethernet transceivers will be based on 40G modules. Therefore, high performance switches will be required consuming vast amount of energy for the E-O and O-E conversion of the transceivers and the switching in the electronic domain. It is clear that if the data rates continue to increase exponentially novel requirements will be required that will be able to sustain the high data rates with minimum latency and minimum power consumption.

1.4 Power consumption requirements

A main concern in the design and deployment of a data centers is the power consumption. Many data consume a tremendous amount of electricity; some consume the equivalent of nearly 180,000 homes [8]. Greenpeace's Make IT Green report [14] estimates that the global demand for electricity from data centers was

Table 1.1 Performance, BW requirements, and power consumption bound for future systems, Source: IBM [1]

Year	Peak performance (10×/4 years)	Bandwidth requirements (20×/4 years)	Power consumption bound (2×/4 years)
2012	10 PF	1 PB/s	5 MW
2016	100 PF	20 PB/s	10 MW
2020	1,000 PF	400 PB/s	20 MW

around 330bn kWh in 2007 (almost the same amount of electricity consumed by UK [8]). This demand in power consumption demand is projected to more than triple by 2020 (more than 1,000bn kWh). According to some estimates [17], the power consumption of the data centers in the USA in 2006 was 1.5% of the total energy consumed at a cost of more than $4.5B.

The power consumption inside the data center is distributed in the following way: the servers consume around 40% of the total IT power, storage up to 37% and the network devices consume around 23% of the total IT power [24]. And as the total power consumption of IT devices in the data centers continues to increase rapidly, so does the power consumption of the HVAC equipment (Heating-Ventilation and Air-Conditioning) to keep steady the temperature of the data center site. Therefore, the reduction in the power consumption of the network devices has a significant impact on the overall power consumption of the data center site. According to a study from Berk-Tek, saving 1W from the IT equipment results in cumulative saving of about 2.84 W in total power consumption [9]. Therefore, a reduction on the power consumption of the interconnection network will have a major impact on the overall power consumption of the data center.

The power consumption of the data centers has also a major impact on the environment. In 2007, data centers accounted for 14% of the total ICT greenhouse gases (GHG) emissions (ICT sector is responsible for 2% of global GHG emissions), and it is expected to grow up to 18% by 2020 [20]. The global data center footprint in greenhouse gases emissions was 116 Metric Tonne Carbon Dioxide ($MtCO_2e$) in 2007 and this is expected to more than double by 2020 to 257 $MtCO_2e$, making it the fastest-growing contributor to the ICT sectors carbon footprint.

Table 1.1 shows the projections for performance, bandwidth requirements, and power consumption for the future high performance systems [16],[21]. Note that while the peak performance will continue to increase rapidly, the budget for the total allowable power consumption that can be afforded by the data center is increasing in a much slower rate (2× every 4 years) due to several thermal dissipation issues.

Table 1.2 depicts the power consumption requirements for the future high performance parallel systems like data centers. In this table it is assumed that the data center network consumes only 10% of the total power consumption. Based on this numbers, we have to reduce the power consumption to only 5 mW/Gbps in 2016 (the bandwidth requirements are in terms of bidirectional traffic). Therefore, novel schemes have to be developed to achieve the power processing requirements of the future data center networks.

Table 1.2 Performance and power consumption requirements for the interconnects, Source: IBM [1]

Year	Bandwidth requirements (20×/4 years)	Network power consumption	Power consumption requirement
2012	1 PB/s	0.5 MW	25 mW/Gbps
2016	20 PB/s	2 MW	5 mW/Gbps
2020	400 PB/s	8 MW	1 mW/Gbps

1.5 The Rise of the Optical Interconnects

In order to face this increased communication bandwidth demand and the power consumption in the data centers, new interconnection schemes must be developed that can provide high throughput, reduced latency, and low power consumption. Optical networks have been widely used in the last years in the long-haul telecommunication networks, providing high throughput, low latency, and low power consumption. Figure 1.3 depicts the adoption of the optical links in different network topologies. In the case of WAN and MAN, the optical fibers were adopted in late 1980s in order to sustain the high bandwidth and latency requirements of the rising global Internet traffic. Firstly, optical fibers were adopted in the domain of LAN networks and later they were adopted for the interconnection of the data center racks. However, in all cases the optical fibers can be used either only for point-to-point links or for all-optical networks (i.e., transparent networks).

The optical telecommunication networks (WAN's and MAN's) have evolved from traditional opaque networks toward all-optical networks. In opaque networks, the optical signal carrying traffic undergoes an optical-electronic-optical (OEO) conversion at every routing node. But as the size of opaque networks increases, network designers had to face several issues such as higher cost, heat dissipation, power consumption, and operation and maintenance cost. On the other hand, all-optical networks provide higher bandwidth, reduced power consumption, and reduced operation cost using optical cross-connects and reconfigurable optical add/drop multiplexers (ROADM) [18].

Currently the optical technology is utilized in data centers only for point-to-point links in the same way as point-to-point optical links were used in older telecommunication networks (opaque networks). These links are based on low cost multi-mode fibers (MMF) for short-reach communication. These MMF links are used for the connections of the switches using fiber-based Small Form-factor Pluggable transceivers (SFP for 1 Gbps and SFP+ for 10 Gbps) displacing the copper-based cables. In the near future higher bandwidth transceivers are going to be adopted (for 40 Gbps and 100 Gbps Ethernet) such as 4 × 10 Gbps QSFP modules with four 10 Gbps parallel optical channels and CXP modules with 12 parallel 10 Gbps channels. The main drawback in this case is that power hungry electrical-to-optical (E/O) and optical-to-electrical (O/E) transceivers are required since the switching is performed using electronic packet switches.

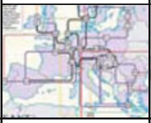
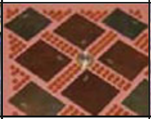
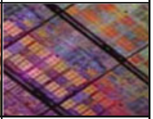
Network type	MAN & WAN	LAN	System	Board	Chip
	Metro & long haul	Campus, Enterprises	Intra-rack Inter-rack	Chip-to-chip	On-chip
					
Distance	Multi-km	10 – 300 m	0.3 – 10 m	0.01 – 0.3 m	<2 cm
Adoption of optical	Since 80s	Since 90s	Since late 00s	After 2012	After 2012
Type of Connectivity	All-optical	Point-to-point and All-optical	Point-to-point	Point-to-point	Point-to-point & all-optical

Fig. 1.3 Optical networks evolution, Source: IBM [1]

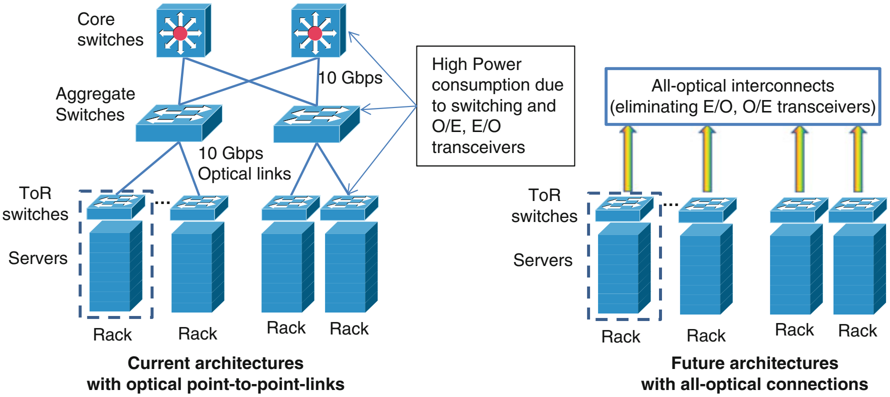

Core switches
Aggregate Switches
10 Gbps
High Power consumption due to switching and O/E, E/O transceivers
10 Gbps Optical links
ToR switches
Servers
Rack
Rack
Rack
Rack
Current architectures with optical point-to-point-links
All-optical interconnects (eliminating E/O, O/E transceivers)
ToR switches
Servers
Rack
Rack
Rack
Rack
Future architectures with all-optical connections

Fig. 1.4 Point-to-point vs. all-optical interconncets

Current telecommunication networks are using transparent optical networks in which the switching is performed at the optical domain to face the high communication bandwidth. Similarly, as the traffic requirements in data centers are increasing to Tbps, all-optical interconnects (in which the switching is performed at the optical domain as it is depicted in Fig. 1.4) could provide a viable solution to these systems eliminating the electrical switches and the E-O and O-E transceivers. These system based on all-optical interconnects could meet the high bandwidth requirements while decreasing significantly the power consumption [4, 5, 11, 15]. According to a study from IBM the replacement of copper-based links with VCSEL-based optical interconnects can reduce the power consumption from 8.3 MW to 1.4 MW [1]. This reduction in total power consumption of a data center by using optical interconnects can saves more than $150M operating cost over 10 years.

According to a report, all-optical networks could provide in the future up to 75% energy savings in the data center networks [23]. Especially in large data centers used in enterprises the use of power-efficient, high bandwidth, and low latency interconnects is of paramount importance and there is significant interest in the deployment of optical interconnects in these data centers [13].

1.6 Structure of the Book

This book presents the most recent and most promising optical interconnects for data centers that have been presented recently by several universities, research centers, and industries. In this section we introduced the data center networks and we discussed the advantages of optical interconnects.

The second section of the book presents the communication requirements inside the data center and discuss the need for optical interconnects. Chapter 2, provided by one of the largest data center owners (Google), reviews the architecture of modern data center networks and their scaling challenges. Furthermore it presents the opportunities and needs for emerging optical technologies to support data center scaling. Chapter 3 provided by APIC Co. presents an end-to-end view of optical interconnects in next generation data centers. This chapter shows the interrelation and research opportunities of high bandwidth applications, microprocessor advances, and interconnect research. Finally, Chap. 4 presents the need for efficient and accurate simulation of energy-aware data center networks. This chapter presents a simulation environment that can be used for accurate simulation and efficient energy estimation of packet-level communications in realistic data center setups.

The third section of the book presents some of the most promising and innovative architectures based on optical interconnects that have been proposed recently. Some of the proposed schemes target current data centers and are usually based on readily available optical and electronic components. The main advantage of these schemes is that they can be adopted faster and usually the cost is quite low. However, most of these schemes cannot be easily scaled to the requirements of the future data center networks.

On the other hand, other schemes are targeting future data center networks that will have excessive requirements in terms of bandwidth and latency. These schemes are usually based on more advanced optical components that could be cost efficient in the near future. In any case all of the presented schemes have unique characteristics that can make them attractive for data center networks.

Chapter 5, provided by HP, focuses on the potential role of optical/photonic communication technology and the impact that this technology may have on future energy-efficient data centers. Furthermore, this chapter presents a scalable switch that is used in a design space exploration to compare the photonic and electrical alternatives for a high-radix switch-chip used in data centers.

Chapter 6, provided by IBM, presents an all-optical multi-stage data center network with distributed arbitration achieved through minimal per-node buffering. The proposed system can achieve low latency using a novel combination of deterministic (prescheduled) and speculative (eager) packet injections.

Chapter 7, from NEC, presents a novel data center network architecture based on cyclic arrayed waveguide grating device and multiple-input multiple-output (MIMO) orthogonal frequency division multiplexing (OFDM) technology. This architecture offers flexible bandwidth resource sharing at fine granularity, high speed switching, and low latency.

Chapter 8, from Polytechnic Institute of New York University, a novel optical architecture that includes interconnected arrayed waveguide grating routers (AWGRs) and tunable wavelength converters (TWCs). The proposed scheme achieves nanosecond-level reconfiguration overhead and provides Petabit switching capacity in the data center networks.

Finally, Chap. 9, provided by Columbia University, presents two network architectures explicitly designed to leverage the capacity and latency advantages of all-optical switching while utilizing unique system-level solutions to the photonic buffering and processing problems. The first architecture is based on the data vortex architecture and is comprised of simple 2×2 all-optical switching nodes. This architecture achieves ultra-high bandwidths and reduce routing complexity, while maintaining reduced packet latencies. The second architecture is called SPINet and is based on indirect multistage interconnection network (MIN) topology. This architecture exploits WDM to simplify the network design and provide very high bandwidths.

References

1. Benner A (2012) Optical interconnect opportunities in supercomputers and high end computing. In: Optical fiber communication conference. OSA Technical Digest (Optical Society of America, 2012), paper OTu2B.4
2. Benson T, Anand A, Akella A, Zhang M (2009) Understanding data center traffic characteristics. In: Proceedings of the 1st ACM workshop on Research on enterprise networking. ACM, New York, pp 65–72
3. Benson T, Akella A, Maltz DA (2010) Network traffic characteristics of data centers in the wild. In: Proceedings of the 10th annual conference on Internet measurement (IMC). ACM, New York, pp 267–280
4. Davis A (2010) Photonics and future datacenter networks. In: HOT Chips, A symposium on high performance chips, Stanford, Invited tutorial (http://www.hotchips.org/wp-content/uploads/hc_archives/archive22/HC22.22.220-1-Davis-Photonics.pdf)
5. Glick M (2008) Optical interconnects in next generation data centers: an end to end view. In: Proceedings of the 2008 16th IEEE symposium on high performance interconnects. IEEE Computer Society, Washington, DC, pp 178–181
6. Hays R, Frasier H (2007) 40G Ethernet Market Potential. IEEE 802.3 HSSG Interim Meeting, April 2007 (http://www.ieee802.org/3/hssg/public/apr07/hays_01_0407.pdf)

7. Hoelzle U, Barroso, LA (2009) The datacenter as a computer: an introduction to the design of warehouse-scale machines, 1st edn. Morgan and Claypool Publishers. Mark D. Hill, University of Wisconsin, Madison. ISBN 9781598295566
8. How Clean is Your Cloud. Greenpeace Report, 2012
9. Huff L (2008) Berk-Tek: The Choise for Data Center Cabling. Berk-Tek Technology Summit 2008 (http://www.nexans.us/US/2008/DC_Cabling%20Best%20Practices_092808.pdf)
10. Infonetics Service Provider Router & Switch Forecast, 4Q11, 2011
11. Kachris C, Tomkos I (2011) A survey on optical interconnects for data centers. IEEE Communications Surveys and Tutorials, doi:10.1109/SURV.2011.122111.00069
12. Kandula S, Sengupta S, Greenberg A, Patel P, Chaiken R (2009) The nature of data center traffic: measurements & analysis. In: Proceedings of the 9th ACM SIGCOMM conference on internet measurement conference, IMC '09. ACM, New York, pp 202–208
13. Lee D (2011) Scaling networks in large data centers. In: Optical fiber communication conference. OSA Technical Digest (CD) (Optical Society of America, 2011), paper OWU1
14. Make IT Green: Cloud Computing and its Contribution to Climate Change. Greenpeace International, 2010
15. Minkenberg C (2010) The rise of the interconnects. In: HiPEAC interconnects cluster meeting, Barcelona, 2010
16. Pepeljugoski P, Kash J, Doany F, Kuchta D, Schares L, Schow C, Taubenblatt M, Offrein BJ, Benner A (2010) Low power and high density optical interconnects for future supercomputers. In: Optical fiber communication conference. OSA Technical Digest (CD) (Optical Society of America, 2010), paper OThX2
17. Report to Congress on Server and Data Center Energy Efficiency. U.S. Environmental Protection Agency, ENERGY STAR Program, 2007
18. Saleh AAM, Simmons JM (2012) All-optical networking: evolution, benefits, challenges, and future vision. Proceedings of the IEEE, 100(5):1105–1117
19. Schares L, Kuchta DM, Benner AF (2010) Optics in future data center networks. In: IEEE 18th Annual Symposium on High Performance Interconnects (HOTI), pp 104–108
20. SMART 2020: Enabling the low carbon economy in the information age. A report by The Climate Group on behalf of the Global eSustainability Initiative (GeSI), 2008
21. Taubenblatt MA, Kash JA, Taira Y (2009) Optical interconnects for high performance computing. In: Communications and photonics conference and exhibition (ACP), Asia pp 1–2
22. Vahdat A (2012) Delivering scale out data center networking with optics – why and how. In: Optical fiber communication conference. Optical Society of America, paper OTu1B.1
23. Vision and Roadmap: Routing Telecom and Data Centers Toward Efficient Energy Use. Vision and Roadmap Workshop on Routing Telecom and Data Centers, 2009
24. Where does power go? GreenDataProject (2008). Available online at: http://www.greendataproject.org. Accessed March 2012