

Optical Networks

Series Editor: Biswanath Mukherjee

Christoforos Kachris

Keren Bergman

Ioannis Tomkos *Editors*

Optical Interconnects for Future Data Center Networks



Springer

Optical Interconnects for Future Data Center Networks

Optical Networks

Series Editor: Biswanath Mukherjee
University of California, Davis
Davis, CA

Broadband Access Networks: Technologies and Deployments
Abdallah Shami, Martin Maier, and Chadi Assi (Eds.)
ISBN 978-0-387-92130-3

Traffic Grooming for Optical Networks: Foundations, Techniques, and Frontiers
Rudra Dutta, Ahmed E. Kamal, and George N. Rouskas (Eds.)
ISBN 978-0-387-74517-6

Optical Network Design and Planning
Jane M. Simmons
ISBN 978-0-387-76475-7

Quality of Service in Optical Burst Switched Networks
Kee Chaing Chua, Mohan Gurusamy, Yong Liu, and Minh Hoang Phung
ISBN 978-0-387-34160-6

Optical WDM Networks
Biswanath Mukherjee
ISBN 978-0-387-29055-3

Traffic Grooming in Optical WDM Mesh Networks
Keyao Zhu, Hongyue Zhu, and Biswanath Mukherjee
ISBN 978-0-387-25432-6

Survivable Optical WDM Networks
Canhui (Sam) Ou and Biswanath Mukherjee
ISBN 978-0-387-24498-3

Optical Burst Switched Networks
Jason P. Jue and Vinod M. Vokkarane
ISBN 978-0-387-23756-5

For further volumes:
<http://www.springer.com/series/6976>

Christoforos Kachris • Keren Bergman
Ioannis Tomkos
Editors

Optical Interconnects for Future Data Center Networks

 Springer

Editors

Christoforos Kachris
Athens Information Technology
Networks and Optical
Communications Lab
Peania, Greece

Keren Bergman
Columbia University
Department of Electrical Engineering
New York, USA

Ioannis Tomkos
Athens Information Technology
Networks and Optical
Communications Lab
Peania, Greece

ISSN 1935-3839

ISBN 978-1-4614-4629-3

DOI 10.1007/978-1-4614-4630-9

Springer New York Heidelberg Dordrecht London

ISSN 1935-3847 (electronic)

ISBN 978-1-4614-4630-9 (eBook)

Library of Congress Control Number: 2012948542

© Springer Science+Business Media New York 2013

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Foreword

Recent advances in cloud computing and emerging web applications have created the need for more powerful data centers. These data centers need high-bandwidth interconnects that can sustain the heavy communication requirements between the servers in the data centers. Data center networks based on electronic packet switches will consume excessive power in order to satisfy the required communication bandwidth of future data centers. Optical interconnects have gained attention recently as a promising solution offering high throughput, low latency, and reduced energy consumption compared to current networks based on commodity switches.

This book presents the most recent and promising solutions that have been presented in the domain of the optical interconnects for data centers. First, it presents the requirements of future data center networks and how optical interconnects could support the data center scaling by providing the required bandwidth. The requirements of future data centers are provided for major data center owners and operators. The majority of the book presents most of the data center network architectures that are based on optical interconnects proposed by leaders in industry and academe. This book should be a valuable source of comprehensive information for researchers, professors, as well as network and computer engineers who are interested in high-performance data center networks and optical interconnects.

Department of Computer Science,
University of California, Davis

Biswanath Mukherjee

Preface

The rise of cloud computing and web applications such as streaming video and social networks has created the need for more powerful data centers that can sustain much higher bandwidth than a few years ago. This trend requires high-performance interconnects for the data center networks that can sustain this bandwidth inside the data centers without consuming excessive energy.

In the last few years many researchers have pointed out the limitations of the current networks and have proposed as a viable solution data center networks based on optical interconnects. Data center networks based on optical interconnects are an interdisciplinary field which encompasses such fields as computer networks, computer architecture, hardware design, optical networks, and optical devices. Therefore it required a wide and in-depth knowledge of the above fields. This book collects the most recent and innovative optical interconnects for data center networks that have been presented in the research community by universities and industries in the last years.

This book is a valuable reference book for researchers, students, professors, and engineers who are interested in the domain of high-performance interconnects and data center networks. In addition this book will provide researchers and engineers who are working on high-performance interconnects invaluable insights into the benefits and advantages of optical interconnects and how they can be a promising alternative for the future data center networks.

Finally, we would like to thank all the authors who provided such interesting and valuable contributions and helped towards the realization of this book.

Peania, Greece
New York, NY
Peania, Greece

Christoforos Kachris
Keren Bergman
Ioannis Tomkos

Contents

Part I Introduction to Data Center Networks

- 1 Introduction to Optical Interconnects in Data Centers** 3
Christoforos Kachris, Keren Bergman, and Ioannis Tomkos

Part II Optical Interconnects in Data Center Networks

- 2 Optical Interconnects for Scale-Out Data Centers** 17
Hong Liu, Ryohei Urata, and Amin Vahdat
- 3 Optical Interconnects in Next Generation Data Centers:
An End to End View** 31
Madeleine Glick
- 4 Simulation and Performance Analysis of Data Intensive
and Workload Intensive Cloud Computing Data Centers** 47
Dzmitry Kliazovich, Pascal Bouvry, and Samee Ullah Khan

Part III Optical Interconnects Architectures

- 5 The Role of Photonics in Future Datacenter Networks** 67
Al Davis, Norman P. Jouppi, Moray McLaren,
Naveen Muralimanohar, Robert S. Schreiber, Nathan Binkert,
and Jung-Ho Ahn
- 6 All-Optical Networks: A System's Perspective** 95
Nikolaos Chrysos, Jens Hofrichter, Folkert Horst,
Bert Offrein, and Cyriel Minkenberg
- 7 A High-Speed MIMO OFDM Flexible Bandwidth Data
Center Network** 119
Philip N. Ji, D. Qian, K. Kanonakis, Christoforos Kachris,
and Ioannis Tomkos

8 A Petabit Bufferless Optical Switch for Data Center Networks	135
Kang Xi, Yu-Hsiang Kao, and H. Jonathan Chao	
9 Optically Interconnected High Performance Data Centers	155
Keren Bergman and Howard Wang	
Index	169

Contributors

Al Davis HP Labs, Palo Alto, CA, USA, ald@hp.com

Amin Vahdat Google Inc., Mountain View, CA 94043, UC San Diego, USA
hongliu@google.com

Bert Offrein IBM Research Zurich, Säumerstrasse 4, Rüschlikon, Switzerland,
ofb@zurich.ibm.com

Christoforos Kachris Athens Information Technology, Peania, Athens, Greece,
kachris@ait.edu.gr

Cyriel Minkenberg IBM Research Zurich, Säumerstrasse 4, Rüschlikon,
Switzerland, sil@zurich.ibm.com

Dayou Qian NEC Laboratories America, Inc., 4 Independence Way, Princeton,
NJ 08540, USA dqian@nec-labs.com

Dzmitry Kliazovich University of Luxembourg, 6 rue Coudenhove Kalergi,
Luxembourg, Dzmitry.Kliazovich@uni.lu

Folkert Horst IBM Research Zurich, Säumerstrasse 4, Rüschlikon,
Switzerland, fho@zurich.ibm.com

H. Jonathan Chao Polytechnic Institute of New York University, Brooklyn,
NY, USA, chao@poly.edu

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

Howard Wang Department of Electrical Engineering, Columbia University,
New York, NY, USA, howard@ee.columbia.edu

Ioannis Tomkos Athens Information Technology, Peania, Athens, Greece,
itom@ait.edu.gr

Jens Hofrichter IBM Research Zurich, Säumerstrasse 4, Rüschlikon,
Switzerland, jho@zurich.ibm.com

Kang Xi Polytechnic Institute of New York University, Brooklyn, NY, USA,
kxi@poly.edu

Keren Bergman Department of Electrical Engineering, Columbia University,
New York, NY, USA, bergman@ee.columbia.edu

Konstantinos Kanonakis Athens Information Technology, Athens, Greece
kkan@ait.edu.gr

Madeleine Glick APIC Corporation, Culver City, CA, USA, glick@apichip.com

Moray McLaren HP Labs, Bristol, UK, moray.mclaren@hp.com

Naveen Muralimanohar HP Labs, Palo Alto, CA, USA,
naveen.muralimanohar@hp.com

Nikolaos Chrysos IBM Research Zurich, Säumerstrasse 4, Rüschlikon,
Switzerland, cry@zurich.ibm.com

Norman P. Jouppi HP Labs, Palo Alto, CA, USA, norm.jouppi@hp.com

Pascal Bouvry University of Luxembourg, 6 rue Coudenhove Kalergi,
Luxembourg, pascal.bouvry@uni.lu

Philip N. Ji NEC Laboratories America, Inc., Princeton, NJ, USA,
pji@nec-labs.com

Robert S. Schreiber HP Labs, Palo Alto, CA, USA, rob.schreiber@hp.com

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

Samee Ullah Khan North Dakota State University, Fargo, ND, USA,
samee.khan@ndsu.edu

Yu-Hsiang Kao Polytechnic Institute of New York University, Brooklyn, NY,
USA, ykao01@students.poly.edu

Acronyms

ADC	Analog-to-Digital Converter
AOC	Active Optical Cable
ASIC	Application Specific Integrated Circuit
AWGR	Arrayed Waveguide Grating Routing
BPSK	Binary Phase Shift Keying
BTE	Bit-Transport Energy
CAGR	Compound Annual Growth Rate
CAWG	Cyclic Arrayed Waveguide Grating
CMOS	Complementary Metal-Oxide Semiconductor
DCN	Data Center Network
DDR	Double Data Rate
DFB	Distributed Feedback Laser
DML	Directly Modulation Laser
DSP	Digital Signal Processing
DWDM	Dense Wavelength Division Multiplexing
EARB	Electrical Arbiter
EPS	Electrical Packet Switch
FEC	Forward Error Correction
FFT	Fast Fourier Technology
FIFO	First In-First Out
FPGA	Field-Programmable Gate Array
FXC	Fiber Cross-Connect
GHS	Greenhouse Gases
HOL	Head-of-Line
HPC	High-Performance Computing
HVAC	Heating, Ventilating and Air-Conditioning
IT	Information Technology
ITRS	International Technology Roadmap for Semiconductors
LAN	Local Area Network
MAN	Metropolitan Area Network
MEMS	Micro-Electro-Mechanical Systems

MIMO	Multiple-Input, Multiple-Output
MMF	Multi-Mode Fiber
MtCO ₂	Metric Tonne Carbon Dioxide
NIC	Network Interface Card
OFDM	Orthogonal Frequency-Division Multiplexing
PD	Photo Detector
PIC	Photonic Integrated Circuit
PON	Passive Optical Network
QAM	Quadrature Amplitude Modulation
QoS	Quality of Service
QPSK	Quadrature Phase Shift Keying
QSFP	Quad Form-factor Plug-in
ROADM	Reconfigurable Optical Add Drop Multiplexer
SaaS	Software-As-A-Service
SCC	Single-chip Cloud Computer
SDM	Space Division Multiplexing
SDN	Software Defined Network
SERDES	Serializer-Deserializer
SFP	Small Form-factor Plug-in
SOA	Semiconductor Optical Amplifier
SMF	Single-Mode Fiber
TCP	Transmission Control Protocol
ToR	Top of Rack Switch
VCSEL	Vertical Cavity Surface Emitting Laser
WAN	Wide Area Network
WDM	Wavelength Division Multiplexing
WSC	Warehouse Scale Computers
WSS	Wavelength Selective Switch
WXC	Wavelength Cross-Connect

Part I
Introduction to Data Center Networks

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.

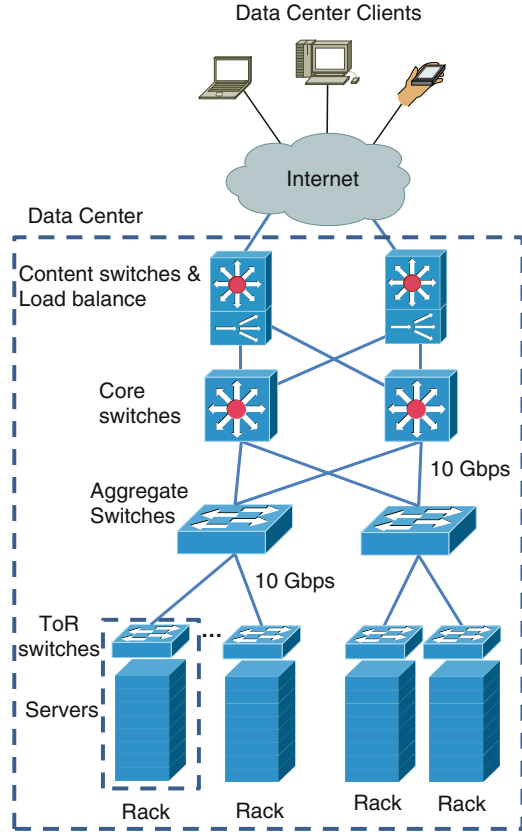
1.2 Architecture of Data Center Networks

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

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 constrained 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 grows 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

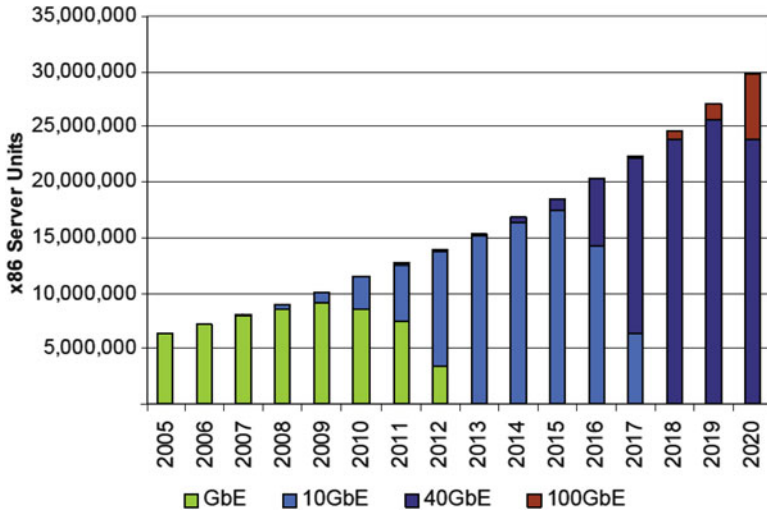


Fig. 1.2 Server Datarate forecast by ethernet connection type, Source: Intel and Broadcom, 2007

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.





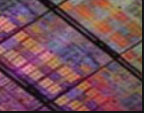
Network type	MAN & WAN	LAN	System	Board	Chip
	Metro & long haul	Campus, Enterprises	Intra-rack Inter-rack	Chip-to-chip	On-chip
					
Distance	Multi-km	10 – 300 m	0.3 – 10 m	0.01 – 0.3 m	<2 cm
Adoption of optical	Since 80s	Since 90s	Since late 00s	After 2012	After 2012
Type of Connectivity	All-optical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

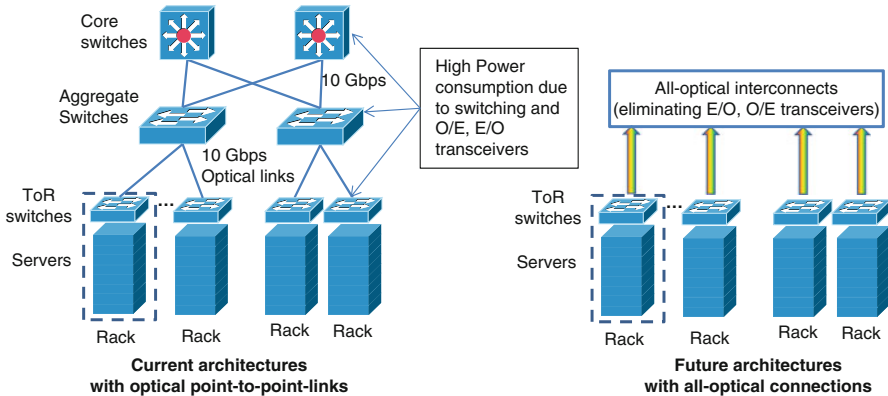


Fig. 1.4 Point-to-point vs. all-optical interconnects

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 Choice 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

Part II
Optical Interconnects in Data Center
Networks

Chapter 2

Optical Interconnects for Scale-Out Data Centers

Hong Liu, Ryohei Urata, and Amin Vahdat

2.1 Introduction

An increasing fraction of computing and data storage is migrating to a planetary cloud of warehouse-scale datacenters [1]. While substantial traffic will continue to flow between users and these datacenters across the Internet, the vast majority of overall data communication is taking place within the datacenter [2]. For example, a datacenter with 100,000+ servers, each capable of 10 Gb/s of bandwidth, would require an internal network with 1 Petabits/s of aggregate bandwidth to support full-bandwidth communication among all servers. While seemingly outlandish, the technology, both on the software [3] and hardware [4,5] sides, is available today.

However, leveraging existing datacenter topologies, switching and interconnect technologies makes it difficult and costly to realize such scale and performance. The bandwidth and power efficiency must scale accordingly to meet the growth of large datacenter networks.

Optics plays a critical role in delivering the potential of datacenter networks and addressing the above challenges. However, fully realizing this potential requires a rethinking of the optical technology components traditionally used for telecom; optimizations must specifically target deployment within a datacenter environment. In this paper, we present an overview of current datacenter network deployments, the role played by optics in this environment, and opportunities and requirements

H. Liu (✉) • R. Urata
Google Inc., Mountain View, CA 94043, USA
e-mail: hongliu@google.com; ryohei@google.com

A. Vahdat
Google Inc., Mountain View, CA 94043, UC San Diego, USA
e-mail: vahdat@google.com

for developing variants of existing technologies specifically targeting large-scale deployment in the datacenter, and emerging optical technologies that could further accelerate the ability of the datacenter to scale.

2.2 Datacenter Network Architecture

We begin by exploring some of the communication and network requirements in emerging large-scale datacenters. The first question is the target scale. 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 and latency locality. The second question is the total computing and communication capacity required by a target application. Consider social networking as an example. Their sites must essentially store and replicate all user-generated content across a cluster worth of machines. The network requirements supporting such services are also significant. For each external request, many hundreds or even thousands of servers must be contacted in parallel to satisfy the request. The last question is the degree that individual servers are multiplexed across applications and properties. For instance, a portal such as Yahoo! may host hundreds of individual user-facing services along with a similar number of internal applications to support bulk data processing, index generation, ads placement, and general business support.

While no hard data is available in answering these questions, on balance we posit a trend to increasing compute densities in datacenters certainly at the level of tens of thousands of servers. It is of course possible to partition individual applications to run on dedicated machines with a dedicated interconnect, resulting in smaller-scale networks. However, the incremental cost of scaling the network will ideally be modest [6], and the flexibility benefits of both shifting computation dynamically and supporting ever-larger applications are large.

Figure 2.1 shows the architecture of typical datacenter networks using a traditional scale-up approach. Individual racks house tens of servers, which connect to a top-of-rack (ToR) switch via copper or optical links. ToR switches then connect to an access switch layer via optical transceivers. If each TOR employs u uplinks, then the network as a whole can support u access switches within a single cluster, as ToRs typically connect to several switches in parallel. The port count c of each access switch then determines the total number of ToRs that may be supported. If each ToR employs d downlinks to hosts, then the network for each cluster could scale to $c \cdot d \cdot u$ total ports (with an oversubscription ratio of $d : c$ at the ToR). If the scale of this two-stage architecture, often limited by the radix of switch silicon [7], is insufficient, then additional layers may be added to the hierarchy [5] to create an aggregation layer, at the cost of increased latency and larger overhead for internal network connectivity. To connect multiple clusters, Layer-3 cluster routers (CR) are employed at the top of the datacenter fabric. Ideally, a fully meshed networking fabric that connects every server to every other server in a datacenter provides

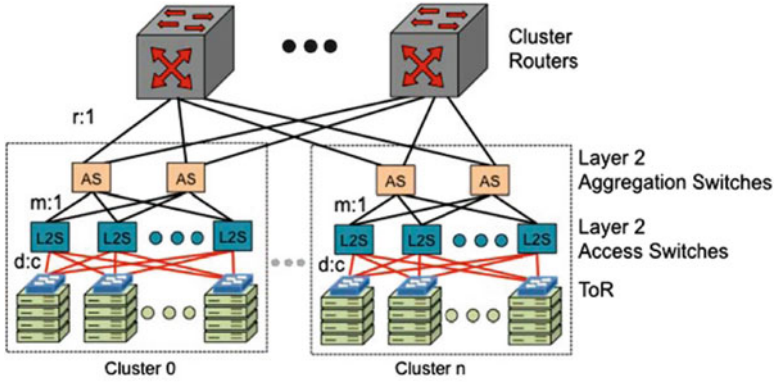


Fig. 2.1 Traditional hierarchical datacenter using scale-up model: ToR layer oversubscription ratio d:c, aggregation layer oversubscription ratio m:1. The cluster router layer is highly oversubscribed with ratio r:1

full bisectional bandwidth, easier programming and better utilization of server computation efficiency. However, such a design would be prohibitively expensive, and oversubscription is often applied at every layer. When a system cannot support the bandwidth needs, new hardware with higher capacity may be purchased to build a larger core (scale-up approach).

While scale-up network fabrics can be cost-effective and easier to set up, particularly for small to medium-size datacenters, they require large up-front investment in more expensive and highly reliable large capacity hardware. In particular, the switches and routers higher in the hierarchy handle more traffic, becoming prohibitively expensive with the efforts needed to increase availability. Further, their inability to scale beyond the limits of a current deployment makes them less attractive for large-scale datacenters.

Over the past decade, with the advances in merchant switch silicon [5] and software defined network (SDN) control plane (<http://www.openflow.org/>) [8], the scale-out model has replaced the scale-up model as the basis for delivering large-scale computing and storage platforms [6, 9].

Figure 2.2 shows a scale-out datacenter architecture that employs an array of small pods composed of identical switches, built with merchant switch silicon, to create the large-scale, non-blocking, networking fabric. The access layer could be a traditional ToR switch performing L2 switch function, or transparently aggregated server links connecting to the aggregation switches. There is full bisectional bandwidth with extensive path diversity within the pod and among the pods.

The scale-out datacenter offers many advantages for building large-scale datacenters: (1) Agility: The networking bandwidth can be allocated for different applications in a modular fashion; (2) Scalability: With its modular approach, we can add computing and storage capacity on an as-needed basis. The datacenter fabric can scale while delivering constant cost per port and per bit/sec of bisection bandwidth; (3) Accessibility: With no bandwidth fragmentation and oversubscription among a large fungible pool of servers, the computation power of each server

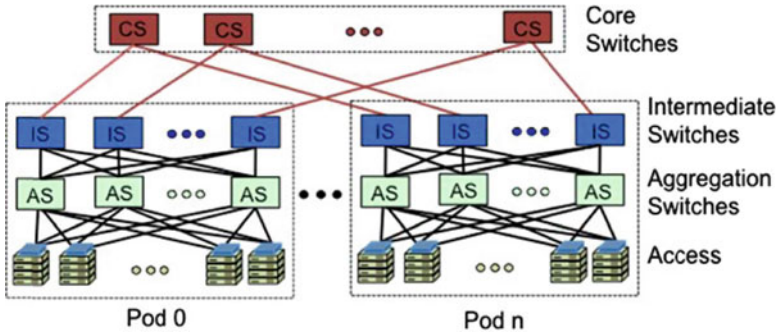


Fig. 2.2 Emerging datacenter using scale-out model. Non-blocking networking fabric is formed with full bisectional bandwidth at each layer

can be widely accessible; (4) Reliability: With extensive path diversity, the network fabric degrades gracefully under failure; (5) Manageability: With a software-defined control plane, hundreds of thousands of servers are managed as one single computer. Petabytes of data can be moved and managed under a single distributed system and one global namespace.

Scale-out datacenters also pose many technical and deployment challenges at and beyond Petabyte scale. While the software and management is beyond the scope of this paper, there are many limitations with existing technologies, including: (1) Management: the number of electrical packet switches (EPS) would substantially complicate management and overall operating expenses; (2) Cost: the cost of fiber cables and optical transceivers would dominate the overall cost of the network fabric; (3) Power: the power of optical transceivers would limit the port density as bandwidth scales; and (4) Cabling complexity: millions of meters of fiber would be required to interconnect large scale-out datacenters, presenting an extremely daunting deployment and operational overhead.

2.3 Enabling Optical Technologies

Optics has already played a critical role, mostly as interconnect media, in delivering on the potential of the datacenter network. Various emerging optical technologies are candidates to address the above technical and operational challenges of scale-out networking and improve the performance and efficiency of large datacenters.

Figure 2.3 shows an example future datacenter network employing WDM transceivers as a first-class entity for modular datacenters [4, 5]. For connections to and from the pods to core switches, traditional parallel optical transceivers are replaced with integrated WDM transceivers (e.g., 40G, 100G, and 400G) to aggregate electrical channels with a common destination over a single strand of fiber. To optimize power efficiency, the interconnect bandwidth between pods can be dynamically tuned to match the required network bandwidth.

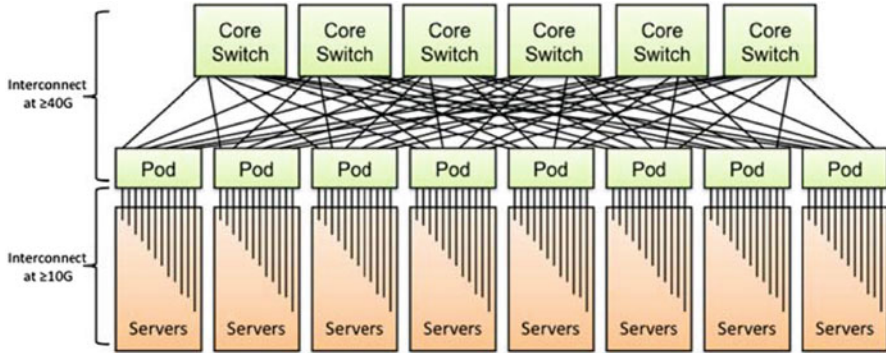


Fig. 2.3 An example future datacenter. A combination of high-radix EPSes along with high-bandwidth optical interconnect will be needed to scale the performance and efficiency of the datacenter

Within the fabric created by the EPSes (Pods and core switches), interconnect technologies will similarly be at higher speeds ($\geq 40\text{G}$) while needing to maintain fixed cost, size, and power per bandwidth. In this application area, integrated multi-core fiber transceivers could provide a very efficient way to scale the bandwidth.

For intra-rack communication, photonics will potentially replace traditional copper interconnects as data rates rise ($\geq 10\text{G}$). At 10Gb/s speeds and beyond, passive and active copper cables are impractical to use beyond a few meters of reach because of their bulky size, high loss at high data rates, and high required power consumption. The emergence of cheap, short-reach optics using IC type of optical packaging (e.g., Light Peak), could change the equation in the datacenter. In the next few years, we will see commodity network interface cards (NICs) with cost-effective $n \times 10\text{G}$ optical interfaces. In addition, the switch silicon will also have native PHY and accept 10G serial connections to further reduce cost and power.

In the following sections, we give an overview of the current state of optical interconnect technologies for the datacenter. We then describe future requirements and potential directions forward, with the end goal of achieving a flexible, energy-efficient, cost-effective datacenter network with bandwidth towards Exabyte scale.

2.3.1 Bandwidth and Scalability of Optical Interconnects

Optical interconnects, with reach between 10m to 2km , are of utmost importance for datacenters. Regardless of implementing a scale-up or scale-out approach, there is a constant, ever-growing demand for increasing aggregate interconnect bandwidth within the datacenter.

To meet the server and network bandwidth growth as outlined in Fig. 2.4, new optical technologies, at the device level, modulation to lane multiplexing

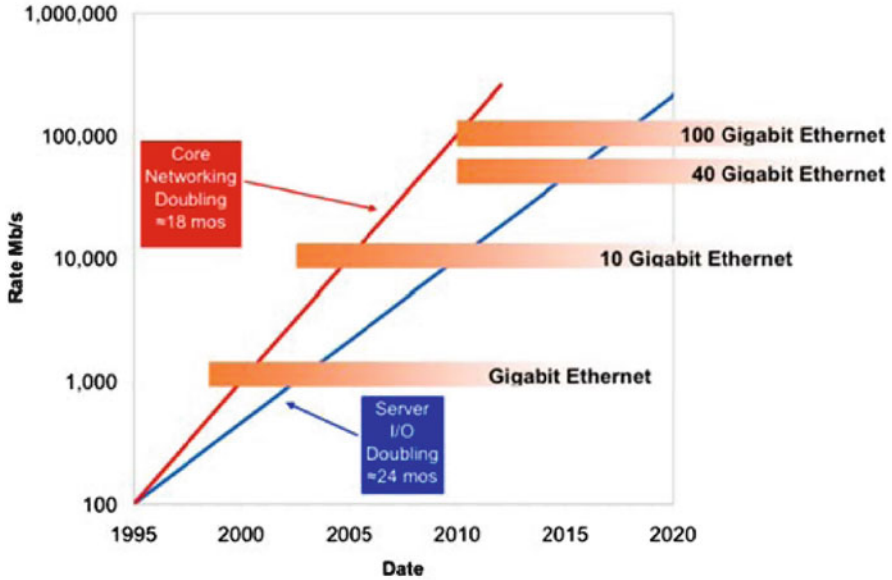


Fig. 2.4 Bandwidth trend for networking and server I/O (from [10])

and photonic packaging, are all required to scale data rate, power, cost, and space/density. Within this optimization, the choice of fiber (single vs. multimode) also needs to be carefully considered along with the corresponding compatible technologies.

2.3.1.1 High-Speed VCSEL, DFB, and Silicon Photonics

Low-power, inexpensive vertical cavity surface emitting lasers (VCSELs) and multimode fiber (MMF) already play a critical role for communication at 10 Gb/s within the datacenter. Although significant progress has been made in higher speed VCSEL using alternative materials [11], overcoming the reliability and yield hurdles to scale VCSELs significantly beyond 10 Gb/s link speed has thus far proven difficult. Further, traditional VCSELs coupled with MMF have a limited reach-bandwidth product due to modal dispersion. At 10 Gb/s, the associated reach is then insufficient to cross a single datacenter building. This maximum reach shrinks rapidly with higher data rates (Fig. 2.5).

Higher-power, more expensive distributed feedback (DFB) lasers and single mode fiber (SMF) are often used in the datacenter to cover the reach beyond 300 m at 10 Gb/s. As we scale from 10 to 25G per lane, DFB lasers employing more novel quaternary materials (InGaAlAs/InP, with a larger band offset) can give better high temperature performance at higher speeds. Novel DFB laser structures, such

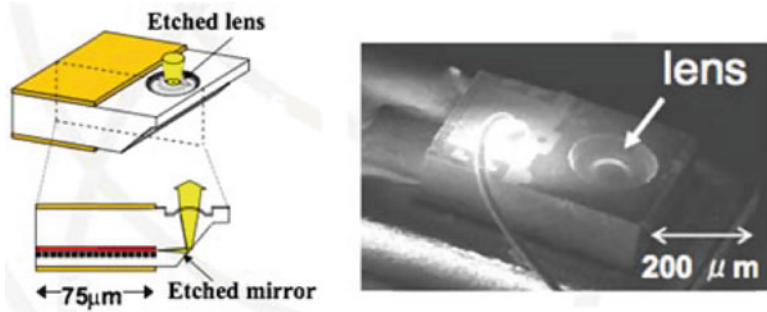


Fig. 2.5 Lensed-integrated surface emitting DFB laser (from [13])

as short cavity [12] and lensed-integrated surface emitting [13] DFB lasers, have also been demonstrated. These approaches provide higher device bandwidth and a narrower spectrum in comparison with their VCSEL-based counterparts to increase interconnect bandwidth and reach while maintaining low power consumption and cost.

In the past decade, significant advances have been made in silicon photonics to address the energy efficiency and cost of traditional optical transceiver using III–V compound materials. Silicon, although not the material of choice for semiconductor lasers due to its indirect band gap, has good thermal conductivity, transparency at the traditional telecom wavelengths, low noise for avalanche multiplication (from high electron/hole impact ionization ratio), and most importantly, allows the leveraging of the silicon CMOS fabrication/process developed for the electronics industry. Silicon photodetectors are the oldest and perhaps best understood silicon photonic device. For wavelengths below 1,000 nm, silicon is a low cost and highly efficient photodetector. Silicon has also demonstrated low loss waveguides for wavelengths above 1,000 nm, thus allowing the creation of higher functionality waveguide-based devices as well as chip-level interconnection of various components (photonic integrated circuits (PICs)). Other recent advances in the main building blocks of silicon photonics include: high efficiency Germanium photodetectors [14], high-speed silicon modulators with extremely small switching energy [15], and Germanium/silicon lasers [16]. The intimate integration of electronics with photonics allows the realization of higher bandwidth at lower power, giving silicon photonics the potential to improve datacenter flexibility, energy efficiency, and cost, contingent on overcoming various packaging and integration hurdles.

2.3.1.2 Multiplexing

Through fundamental device improvements described above, optical link speed is increased to align with the electrical switch I/Os. In addition, methods for increasing interconnect bandwidth through multiplexing must be utilized.

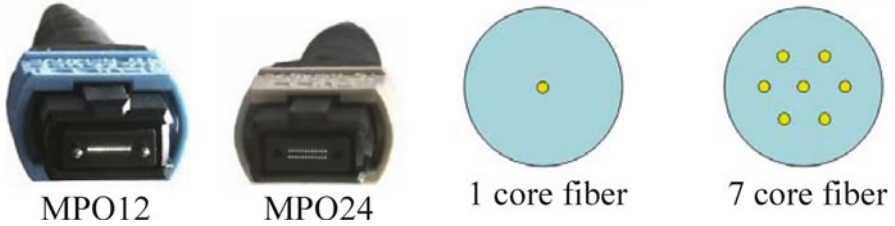


Fig. 2.6 Spacing division multiplexing with (a) parallel ribbon fiber cable (b) multi-core fiber cable

Space division multiplexing (SDM) and wavelength division multiplexing (WDM), taking advantage of the nature parallelism of data lanes in computer architecture and switch silicon, are two widely employed multiplexing techniques in the datacenter. There are other multiplexing techniques, such as optical orthogonal frequency division multiplexing (O-OFDM), multilevel or advanced modulation, which could also scale the bandwidth and capacity of a single fiber. However, these approaches all require a gear box to perform signal encoding, ASICs for DSP and/or A/D, D/A converters for analog-to-digital signal conversion, which incurs large power consumption penalty and could be cost prohibitive for datacenter applications.

Space Division Multiplexing

One natural way of scaling bandwidth is dedicating one fiber per lane along with parallel laser and photodetector arrays. Parallel optical transceivers using ribbon fiber and MPO connectors (Fig. 2.6a) is widely deployed within datacenter and HPC environments. However, the MPO connector and ribbon fiber can incur a significant portion of the entire datacenter network cost [4]. Scaling bandwidth through parallelism in this manner can also lead to an unmanageable volume and size in the fiber infrastructure. Thus, when longer reach interconnects are required, this approach becomes obsolete.

Beyond space division multiplexing using parallel ribbon cables, there has been recent growing interest in developing multi-core fiber (MCF) technology for long-haul transmission in telecom [17]. This field and the associated components developed for it may also be leveraged for the datacenter to extend the application area and lifetime of the space division multiplexing approach [18, 19]. Within a single MCF, multiple cores share one cladding as shown in Fig. 2.6b. Using a grating coupler, MCF can be terminated directly using regular LC connection to laser and photodetector arrays [20]. The interconnect density is thus increased by placing more bandwidth within a single strand of fiber cable.

Wavelength Division Multiplexing

Wavelength division multiplexing has been widely employed in metro and long-haul transmission, allowing the telecom industry to gracefully scale bandwidth over the past several decades. It is clear WDM will need to make its way from these traditional telecom application areas to the short reach datacenter interconnect area. To reduce the cabling overhead described above, to scale to ever-increasing link bandwidth, spectrally efficient optics needs to be employed in next-generation datacenter transceivers [4]. However, meeting datacenter economies and scale requires WDM performance without an associated explosion in power and cost, as outlined below:

- *Cost*: In traditional, telecom applications, the approach was, and still is to a large degree, spend more cost at the link end points to maximize the throughput of precious long distance fiber links, as evidenced by the activity in research and development on coherent transmission devices and systems over the past several years. Within the datacenter, fiber resources are much more abundant and cheap. Thus, the transceiver cost must be aggressively reduced so as not to dominate the cost of the datacenter interconnect fabric.
- *Power consumption*: Transceivers with large power consumption present thermal challenges and limit EPS chassis density. In the datacenter, non-retimed, uncooled solutions are preferred. Photonic integrated circuits (PIC), low-threshold lasers with better temperature stability (e.g., quantum dot laser [21]) and silicon photonic modulators with low switching energy hold promise for further reducing power.
- *Optical link budget*: Datacenter transceivers must account for multi-building span reaching 2 km and optical loss from patch panels. For large-scale deployments, additional link budget is also needed for operational simplicity to provide coverage of high loss links at the tail end of the distribution.
- *Bandwidth and speed*: The photonics highway must align seamlessly with the electrical switch fabric in bandwidth and speed. Today 10G, $4 \times 10\text{G}$ LR4 and $10 \times 10\text{G}$ LR10 provide cost-effective and power-efficient WDM transceiver solutions. Moving forward, further integration in the transceiver to align with the bandwidth and speed from the switch silicon I/O speed will be necessary, with the availability of $n \times 10\text{G}$, or $n \times 25\text{G}$ native electrical link speeds.
- *Spectral efficiency*: There will continue to be a tension between spectral efficiency, power consumption, path diversity and cabling complexity. For the intra-building network, a rich-mesh topology is desirable; hence, lower spectral efficiency can be traded for lower power, cheaper transceiver cost, and a richer network fabric. While at higher aggregation layers or the inter-building network, bandwidth is more concentrated over point-to-point links and dark fiber is expensive to procure; hence, DWDM with higher spectral efficiency is preferred.

Table 2.1 Comparing SMF and MMF

	SMF cable		MMF cable	
	Cost	Volume	Cost	Volume
10GE	1×	1×	2×	1×
40GE (4 × 10Gb/s)	1×	1×	6×	2.25×
100GE (4 × 25Gb/s)	1×	1×	12×	2.25×
400GE (16 × 25Gb/s)	1×	1×	30×	4×

2.3.1.3 Fiber

Optical fiber is fast becoming a dominant transmission media for modern datacenters. The vast number of interconnections for scale-out networks drives the need for compact cabling solution.

At 10G, rack-to-rack communication in the datacenter and high-performance computing environments have traditionally been the realm of VCSEL-based transmitters, and multi-mode fiber (MMF) primarily due to their low transceiver cost.

However, with the rising cost, bandwidth and reach limitation (approximately 10 Gb/s, several hundreds of meters) of these MMF-based interconnections, moving to single mode fiber (SMF)-based interconnections for even the shorter, rack-to-rack distances provides significant benefits [4]. Due to its simple structure and its prevalence for decades in the telecommunications industry, SMF is a low-cost, commodity technology. A single strand of fiber can support tens (to hundreds) of terabits per second of bandwidth. These high bandwidths per SMF are obtained not by a single transmitter–receiver pair, but by a number of pairs, each operating on a separate wavelength of light contained in the same fiber through WDM, as described in the previous section.

As a result of these characteristics, SMF-based interconnects provide a number of advantages over MMF-based interconnects within the datacenter, contrary to the conventional viewpoint. As listed in Table 2.1, there is a large saving in cable cost and volume through multiple generations of networking fabric when the bandwidth scales from 10GE, 40GE/100GE to 400GE. Thus, there is both a CapEx and OpEx advantage. The fiber is installed once for a particular interconnect speed. Subsequent increases in speed only require adding wavelength channels, with the same fiber infrastructure remaining in place. Fiber thus becomes a static part of the facility and requires only a one-time installation, similar to the electrical power distribution network. Considering the large number of fibers and time and cost to install them, this represents a huge cost saving. In addition, scalability in interconnect bandwidth is greatly enhanced as wavelengths in the same fiber are increased for higher speeds, and not the number of parallel fibers, as would be required in an MMF interconnection. The maximum reach of the interconnection is also significantly increased, along with reduction of fiber count and patch panel space.

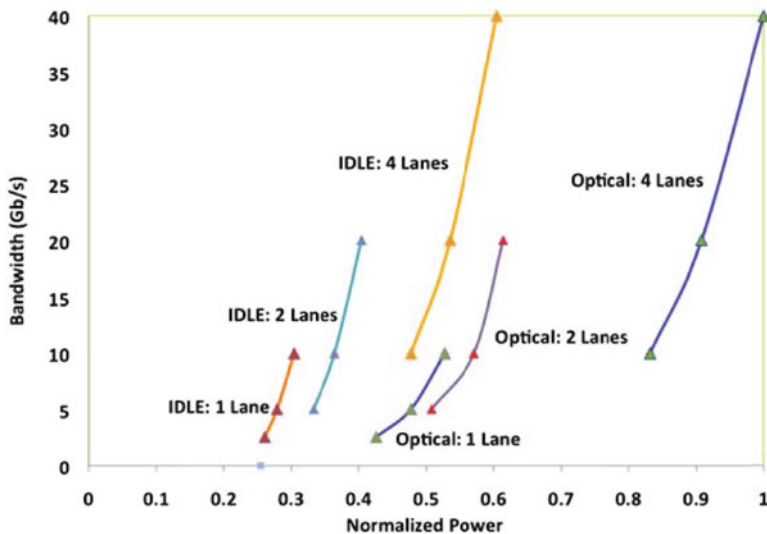


Fig. 2.7 Power and bandwidth dynamics for optical links for 4 lanes with data rate from 2.5 to 10 Gb/s per lane

2.3.2 Energy Proportionality of Optical Interconnects

Traditional hierarchical datacenter networks consume little power, relative to servers, because of the high degree of over-subscription at each layer and low utilization of servers. However, for scale-out networking, because of a substantial increase in cluster bisectional bandwidth and better utilization of servers, the networking power, which was less than 12%, could now become a significant portion of overall datacenter power [22].

Besides using low power optical transceivers for the datacenter, further improvement of network power efficiency can be achieved by making communication more energy-proportional to the amount of data being transmitted.

Optical interconnects and associated high speed serializer/deserializer (SerDes) have a large dynamic range in power and delivered bandwidth. Figure 2.7 illustrates the normalized dynamic range of an off-the-shelf switch chip available today, where it is possible to manually adjust the link data rates accordingly. The maximum link rate of 40 Gb/s is obtained with four lanes running at 10 Gb/s each. The dynamic range of this particular chip is 64% in terms of power, and $16\times$ in terms of performance. Therefore it is possible to operate the link with fewer lanes and at a lower data rate to reduce the power consumption of the optical links. This allows efficient network power consumption by making communication more energy-proportional, to the amount of data being transmitted.

Both Infiniband and Ethernet allow links to be configured for a specified speed and width, although the reactivation time of the link can vary from several

nanoseconds to several microseconds. For example, when the link rate changes by 10 Gb/s, 20 Gb/s, and 40 Gb/s, with all four lanes on, the chip simply changes the receiving Clock Data Recovery (CDR) bandwidth and relocks the CDR. Since most SerDes today uses digital CDRs at the receive path, the locking process for receiving data at a different data rates is fast, $\approx 50\text{--}100\text{ns}$ for the typical to worst case. Adding and removing lanes offers more energy saving, but the process is relatively slow compared to link rate changes, within a few microseconds.

Although optical links are already capable of having their performance and power tuned, in current networks and switches, variable link speed is typically something that must be manually configured. With the advances in software defined networking (SDN), the link speed can be dynamically configured on-the-fly to tailor bandwidth (and power) based upon real-time network utilization and traffic demand [23]. In doing so, the power efficiency of scale-out networks can be made more energy proportional without fundamentally changing the performance of the network.

2.4 Conclusions

Optics has already had a significant impact on the datacenter. However, we are at the cusp of a transformation of datacenter network architecture fueled by emerging optical technologies and components. These along with other yet undeveloped optical technologies will be critical in fueling the ever-growing demand for performance and bandwidth of the global compute infrastructure.

References

1. Barroso L et al (2009) The Datacenter as a Computer - an introduction to the design of warehouse-scale machines, May 2009
2. Lam CF et al (2010) Fiber optic communication technologies: what's needed for datacenter network operations. IEEE Comm
3. Niranjan R et al PortLand: A scalable fault-tolerant layer 2 datacenter network fabric. In: ACM SIGCOMM '09
4. Liu H et al Scaling optical interconnects in datacenter networks. In: 18th IEEE HotInterconnects'10, pp 113–116
5. Al-Fares M et al A scalable, commodity, datacenter network architecture. In: ACM SIGCOMM'10
6. Vahdat A et al (2010) Scale-out networking in the datacenter. IEEE Micro 29–41
7. Kim J et al Microarchitecture of a high-Radix router. In: ISCA'05
8. Farrington N et al Data center switch architecture in the age of merchant silicon. In: 17th IEEE HotInterconnects'09
9. Greenberg A et al VL2: a scalable and flexible data center network. In: SIGCOMM'10
10. HSSG IEEE 802 An overview: the next generation of ethernet. <http://www.ieee802.org/3/hssg/public/nov07/>
11. Anan T et al (2008) High-speed 1.1 – μm -range InGaAs VCSELs. In: OFC

12. Fukamachi T et al (2009) 95°C uncooled and high power 25-Gbps direct modulation of InGaAlAs ridge waveguide DFB laser. In: ECOC
13. Shinoda K et al (2010) Monolithic lens integration to 25-Gb/s 1.3- μm surface-emitting DFB laser for short-reach data links. In: OECC
14. Vivien L et al (2008) 42 GHz p.i.n Germanium photodetector integrated in a silicon-on-insulator waveguide. *Opt Express* 16
15. Liu A et al (2008) Recent development in a high-speed silicon optical modulator based on reverse-biased pn diode in a silicon waveguide. *Semicond Sci Technol* 23
16. Liu J et al (2010) Ge-on-Si laser operating at room temperature. *Opt Lett* 35(5):679–681
17. Hayashi T et al (2011) Ultra-low-crosstalk multi-core fiber feasible to ultra-long-haul transmission. In: OFC/NFOEC
18. Zhu B et al (2010) 7 x10-Gb/s multicore multimode fiber transmissions for parallel optical data links. In: ECOC
19. Lee BG (2010) 120-Gb/s 100-m transmission in a single multicore multimode fiber containing six cores interfaced with a matching VCSEL array. In: Photonics Society Summer Topical
20. Doerr CR et al (2011) Silicon photonics core-, wavelength-, and polarization-diversity receiver. *IEEE Photonics Technol Lett* 23(9)
21. Bimberg D (2007) Semiconductor quantum dots: genesis – the excitonic zoo – novel devices for future applications. In: Kaminow I et al (eds) *Optical fiber telecommunications - V*, chap 2, vol A. Academic, New York
22. Abts D et al (2010) Energy proportional datacenter networks. In: *Proceedings of the International Symposium on Computer Architecture*
23. Das S et al (2011) Application-aware aggregation and traffic engineering in a converged packet-circuit network. In: OFC

Chapter 3

Optical Interconnects in Next Generation Data Centers: An End to End View

Madeleine Glick

3.1 Introduction

Optics has enormous potential for high bandwidth transmission with experimental results achieving transmission over 100 Tb/s in a single mode fiber [1]. However, until recently commercial implementation of optical links for data centers computer networks has been hampered by relatively high power consumption and high cost in addition to the barriers inhibiting the adoption of new technology. Recently we have seen advances with the adoption of active optical cables and plans to incorporate vertical cavity laser modules in supercomputers [2–4]. These solutions had been considered too costly for the commodity-based data center but the requirements for higher bandwidth are changing these perceptions towards a more favorable view of the incorporation of optical solutions. Although low cost is still a primary metric for the data center, increasing data rates are making optical transmission more advantageous in terms of cost/bit. Meeting the challenge of incorporating optics in the data center network can be facilitated by considering end-to-end solutions in collaboration with application software and network engineers. Power consumption of a subsystem can be determined; however, it is most relevant in relation to the whole. It is important to ask how a reduction in power (or other metric such as latency) of the optical subsystem gets reflected in a reduction of the power consumption of the system as a whole. Implementing optical switches in a network is made more challenging by the lack of optical random access memory. Viable solutions are most likely to be the result of collaboration between jointly developed optical subsystems and novel scheduling and routing algorithms [5–10].

M. Glick (✉)
APIC Corporation, Culver City, CA 90230, USA
e-mail: glick@apichip.com

Consequently, the functioning of the optical system is intimately related to other aspects of the network, requiring a heuristic assessment of the system as a whole rather than the individual optical subsystem.

The following presents the interrelation and research opportunities of high bandwidth applications, microprocessor advances, and interconnect research. This brief review cannot cover many of the research avenues being pursued to improve the capabilities and efficiency of the data center; it focuses on areas that relate to high bandwidth requirements relating to the interconnection network. We start with an overview of the three main forces driving innovation in the data center, enormous increases in traffic to and from and within the data center, advances in multiprocessors, and efforts to reduce energy consumption. We then review approaches to overcoming these using optical interconnects.

3.2 The Data Center

3.2.1 *Data Centers and Cloud Computing*

Data centers come in many forms and sizes. Cisco defines them as follows: “Data Centers house critical computing resources in controlled environments and under centralized management, which enable enterprises to operate around the clock or according to their business needs. These computing resources include mainframes; web and application servers; file and print servers; messaging servers; application software and the operating systems that run them; storage subsystems; and the network infrastructure, whether IP or storage-area network (SAN). Applications range from internal financial and human resources to external e-commerce and business-to-business applications” [11].

The definition of the data center in general is much broader than the warehouse scale version, containing tens of thousands of computers often appearing in the news headlines [12, 13]. In [12], the authors make the case that the warehouse scale data center or computer differs significantly from a large data center. Much of the application, middleware and system software is built in house and these centers run a smaller number of very large applications. They are controlled by a single organization and focus on cost efficiency. It is these largest scale centers, belonging to single organizations, that often require and drive technology innovation to achieve the required performance and become the machines enabling advanced innovation in applications

One of the innovations leading to more traffic in the larger data centers is Cloud Computing. In [14] this is defined as “Cloud Computing refers to both the applications delivered as services over the Internet and the hardware and systems software in the datacenters that provide those services. The services themselves have long been referred to as Software as a Service (SaaS). The datacenter hardware

and software is what we will call a Cloud. When a Cloud is made available in a pay-as-you-go manner to the general public, we call it a Public Cloud; the service being sold is Utility Computing. We use the term Private Cloud to refer to internal datacenters of a business or other organization, not made available to the general public. Thus, Cloud Computing is the sum of SaaS and Utility Computing, but does not include Private Clouds. People can be users or providers of SaaS, or users or providers of Utility Computing.” The use of the cloud is growing dramatically. A Cisco forecast on global cloud computing predicts that global cloud IP traffic will increase by 12 times in the next 5 years and will be over one third of all data center traffic by 2015 (http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns1175/Cloud_Index_White_Paper.pdf).

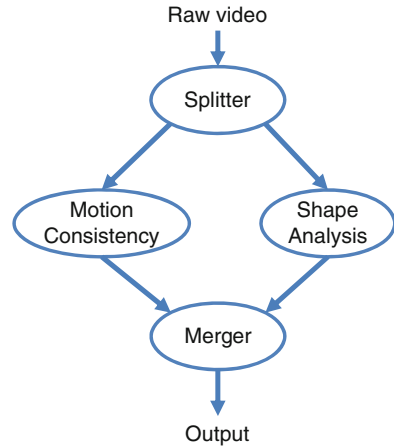
3.2.2 Applications

Recent increases in the usage and data rates of video, satellite imagery, peer-to-peer, and storage have significantly driven growth in Internet traffic [15]. We would like to have a better understanding of how new applications are affecting traffic to and within the data center in order to better understand how optical solutions might be applied to meet the challenges being faced. Apart from the sheer increase in usage due to video streaming, applications such as medical scans, virtual reality modeling and physics simulations are obtaining and storing more and more data and using these data sets for more complex manipulation. In addition, sensors are collecting and analyzing more and more information in our surroundings. The advances being enabled by the ever-increasing processing power of new generation multicore microprocessors. These applications are yielding huge quantities of data that must be processed on the fly and/or stored for later processing. The world is becoming *data rich*. Researchers are looking to find the best ways to handle and manipulate this data in order to drive further advances in many fields including mobile computing, personal media, machine learning, robotics, and other applications [16].

The application or a portion of it may be more heavily based on using processing cores for computation or for communicating stored information. For example, earthquake prediction and scientific computations which are often run on dedicated supercomputers have a phase which is more communication intensive as the stored data is transferred to compute nodes and a phase that is more compute intensive as the task is divided among processing cores to carry out the computation. The Reduce portion of a MapReduce [17] is dominated by the communication of results between the cores.

As a specific example, consider real-time event recognition in video. There are significant efforts to identify and localize objects and events within video data for intelligence and surveillance applications [18, 19]. Rather than analyzing a specific individual frame or scene, the goal of event detection is to identify and localize

Fig. 3.1 Event detection data flow. Thick (thin) arrows denote high (low) bandwidth links



specified patterns in video across space and time, for example, a person waving his or her hand. Real-world actions occur often in crowded, dynamic environments and it is often difficult to separate the actor from the background. For real-time detection of multiple events, such as hand waving, running, and cell phone usage, the video stream is replicated for task parallelism to different nodes on the computer cluster, massively increasing the amount of data being transmitted.

Applications which employ computer vision are very computationally demanding have specific latency requirements in interactive settings and often have time-varying and data-dependent execution characteristics. They generally have characteristics that make them amenable to parallel execution.

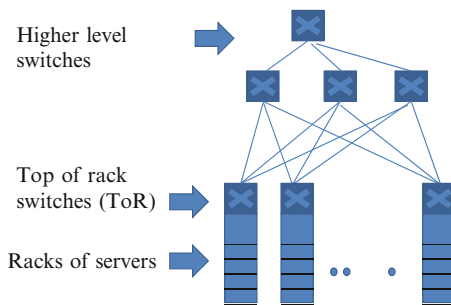
Figure 3.1 shows a decomposition of event detection tasks. Incoming video is replicated to two different analysis modules, the results of which are sent to a merging task that determines if an event is recognized.

The data communication requirements between tasks are sharply different: the channels transmitting video data require much higher bandwidth than those transmitting results of the analysis tasks.

The amounts of data to analyze quickly get extremely large. For standard NTSC video, with 640×480 pixels $\times 3$ bytes (for 24 bit color) $\times 30$ frames/s = 27,648,000 or 27.6 Mbytes/s. If one goes to high definition, the numbers become $1,920 \times 1,080 \times 3 \times 60$ frames/s = 373,248,000 or approximately 373.3 Mbytes/s. These numbers are for one camera. In a surveillance situation there would be 10s to 100s of cameras over fairly large areas (e.g., airports). Compression and sophisticated algorithms can be used to reduce the data rates (the compression for Mpeg is approximately 100 for HD and 20–40 for standard video). But this would not completely solve the problem especially as usage increases.

Due to the parallelization of the compute task especially for a real-time response, many cores are used simultaneously. For object recognition, hundreds to thousands of cores will be required.

Fig. 3.2 Schematic of data center based on tree architecture. East–West traffic is traffic or communication between nodes or servers on the different racks. This traffic is increasing due to new applications



3.2.3 Advances in Microprocessors

The new applications described above are being enabled by advances in the capabilities of the new multicore microprocessors. The trend to multi- and many-core architectures with shared memory and storage brings new capabilities to computing and new bandwidth demands for the interconnect [20, 21].

At the processor level, there is a communications bottleneck, for multicore CPU to CPU communications and also the well-known bottleneck from CPU to memory. The requirements on interconnect bandwidth are constantly increasing. Despite research advances in copper interconnects, the increasing transceiver complexity is becoming limited by increased signal error [22–24] and power consumption constraints.

Trends show that the CPU to memory link will require a bandwidth of greater than 200 GB/s by 2015 [25]. Optical interconnects offer an alternative option to enable higher bandwidth communication, scalability, and added design flexibility.

3.2.4 Network Bottleneck

As discussed above, new applications are creating an increasing demand for bandwidth. Many of these applications from scientific computing to web search and MapReduce require substantial intra-cluster bandwidth. Intra-cluster data center bandwidth, known as east–west traffic, is increasing even faster than the increase of traffic into the data center. In 2011 the ratio was approximately 4:1 in Microsoft data centers [26]. As data centers and their applications continue to grow, scaling the capacity of the network fabric to the ideal all-to-all communication becomes more and more challenging.

In a classic data center, tree architecture network design (Fig. 3.2), there is more bandwidth available within a rack and more within an aggregation router than between racks. The network is over-subscribed. Although enabling massive storage and computer power from commodity or relatively low-cost processors, this architecture is more suitable to tasks where most high bandwidth communication

is between nearby nodes. The parallelization programs should therefore be aware of the physical layout of nodes for efficient parallelization and execution. These placement restrictions make the already over-constrained workload placement problem even more difficult. In addition, as we have described above, many applications are using more and more cores, increasing the complexity of this approach. In addition, to take full advantage of virtualization [27] it would be beneficial to reduce the constraint of and dependence on placement of compute tasks and stored data [28]. Therefore, despite advances in multicore processors, offering increasing processing capabilities, the network performance and the incorporation of new applications may be limited by the inter-connection network [29]. As James Hamilton of Amazon says, “We are allowing the network to constrain optimization of the most valuable assets” (<http://perspectives.mvdirona.com/2010/10/31/DatacenterNetworksAreInMyWay.aspx>).

3.2.5 Energy Efficiency and Energy Proportionality

There is a growing awareness that power consumption of computer networks cannot increase at the same rate as in the past, from the perspective of both social responsibility and cost [24, 30, 31]. In 2006, servers and data centers were estimated to consume 1.5% of the electricity in the USA (61 billion kilowatt-hours (kWh)), doubling from the year 2000 and implying that with no changes in power consumption trends this could double again by 2011 as more and more data is being stored in data centers for online use, leading to an increase in the number of data centers and the large number of servers in each along with the required network and cooling equipment. This increase may not have been as much as expected due to the economic downturn [32–34]. Locations for data farms are being chosen on the basis of the cost of electricity. Google, for example, has built a data center along the Columbia River Gorge for its lower cost electricity. Increases in energy consumption are projected even when including cloud computing [14, 35] and virtualization [27] both of which cut down the number of servers and therefore total energy consumed [36]. These trends are motivating considerable effort into improving energy efficiency in the data center.

Apart from the sheer cost of the energy bill, energy consumption has become a very public social issue. The relative value of “dematerialization,” moving bits instead of atoms, is not completely obvious. A Times of London report claiming that two Google searches require as much energy as boiling a kettle of water (<http://www.technewsworld.com/rsstory/65794.html>) resulted in a flurry of comments, explanations, and clarifications (<http://googleblog.blogspot.com/2009/01/powering-google-search.html>) ending with Google pointing out that a typical user’s Google searching for a year would produce the same amount of CO₂ as a single load of washing and a table of values showing that 15,000 searches is equivalent to the energy required to make a single cheeseburger (<http://googleblog.blogspot.com/2009/05/energy-and-internet.html>). It is not clear how any of these

claims and counter-claims shed much light on the issue. It does, however, indicate a desire and effort to relate day-to-day Internet activities to other day-to-day activities.

The energy consumption discussion has gone beyond the issue of merely reducing the amount of energy used. Apart from the comprehensive and widely cited studies by Koomey on Worldwide Electricity Used in Data Centers [33, 34], Greenpeace has also published a report, “How dirty is your Data?” (<http://www.greenpeace.org/international/Global/international/publications/climate/2011/Cool%20IT/dirty-data-report-greenpeace.pdf>) distinguishing between efficient IT and green IT. They express the concern that making data centers more energy and cost efficient will encourage more usage and not reduce energy consumption as a whole. They rate major companies on the environmental and transparency policies concerning their data centers.

From perhaps a more technical viewpoint, we see that companies have found many ways to improve energy efficiency over the last few years. The concept of the power usage effectiveness, PUE, has become more widespread. The PUE is defined as the total facility power divided by the IT equipment power and is therefore a measure of how efficiently a data center uses its power, with $PUE = 1.0$ being the ideal. Google reports quarterly on PUE performance, which is steadily reducing and currently approximately 1.2 (<http://www.google.com/about/datacenters/inside/efficiency/power-usage.html>). Some of the energy saving techniques are listed on the web site. At Facebook’s data center in Pineville the cold aisle is at 81F and hot air from the servers is used to heat the offices. They have changed the dimensions of their servers to 1.5U in order to promote better airflow. They have announced that the PUE of this data center has achieved a remarkable 1.08 [37].

In “The Case for Energy Proportional Computing” [38] Barroso and Holze point out that studies of average CPU utilization have found that servers were rarely completely idle and seldom operate at maximum utilization. The consequence of this is that servers spend most of their time in the lowest energy efficiency regime. Their claim that energy proportional computing could potentially double the efficiency of a typical server has spawned considerable activity towards this goal. One should note that 100% utilization is not necessarily a desirable goal, as that leaves the system in a state too close to the edge of poor performance. In addition, turning off relatively idle servers is also not as obvious a solution as it might seem, as data is distributed among the servers and idle time is often used for necessary background tasks. In [39] the authors take this concept further and advocate for energy proportional data center *networks*. They point out that the trends towards less oversubscription and an increase in bisection bandwidth will require additional switching and network capabilities, and therefore network power will become a much more significant portion of the energy budget. The major keys to the energy proportional network are topology (the authors propose using the flattened butterfly) and optimal use of high bandwidth links. The authors introduce the concept of *dynamic topologies* for a dynamically changing, energy proportional network.

3.3 Optical Interconnects

3.3.1 System Level Interconnection Networks

The massive increase in traffic in the data center, along with new applications, advances in semiconductor capabilities, and the need to reduce energy consumption have all led to a consensus that change is needed in the data center architecture. Many research groups in industry and universities are engaged in efforts to find scalable solutions to improve data center performance while reducing energy consumption. There are efforts in software, electronics and photonics, and combinations of these. Some are looking into nearer term solutions with off-the-shelf components while others depend on device research, primarily silicon photonics.

Several research groups have proposed modified electronic networks to improve the bisection bandwidth of the data center while maintaining the use of low-cost commodity hardware; this has been called scaling out in comparison with scaling up to higher cost, higher bandwidth equipment. The scale out solutions, however, have a high cost in wiring and switching complexity [40] and are less appealing if we would like to find a solution that is scalable to future generations of data centers, making these solutions perhaps more feasible for the short term [41–45]. Pursuing an idea for hybrid electrical/optical networks originally proposed for supercomputing applications [46] several groups almost simultaneously proposed extrapolations of this concept for the data center [6, 7, 47]. The basic idea being that full bisection bandwidth is not a requirement for improved performance and that a few high bandwidth pipes at the higher levels of the tree could relieve congestion (Fig. 3.3). In addition, if the higher bandwidth requirements were based on latency tolerant, long-lived flows, the high bandwidth links could be built from commercially available optical links and optical MEMs switch technology. By using the circuit-based optical switch, these networks become not only hybrid electrical/optical but also hybrid packet/circuit networks. Reference [47] provides information on the MEMs reconfiguration times and considers applications for financial institutions.

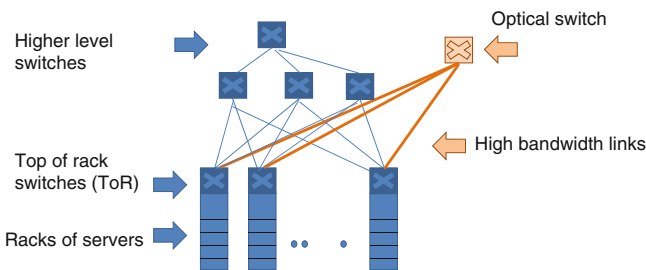


Fig. 3.3 Schematic of a hybrid optical/electrical (circuit/packet) network. Based on the traditional tree architecture (blue, thin lines), the hybrid network adds an optical circuit switch connected by high bandwidth optical links (orange, thick lines) for selected high bandwidth transfers

Helios [6] and c-through [8,48] differ primarily in their implementation of traffic estimation and buffering. It was recognized at the outset that an advantage would be dependent on the traffic characteristics of the data center network and suitable application aware interfaces. A recent joint publication [49] reviews these projects and their limitations. These hybrid optical/electrical solutions were pursued using off-the-shelf equipment. Some network limitations derived from the off-the-shelf nature of the components used, for example timing constraints [50] and are not fundamental to the optical devices. Recognizing limitations in the MEMs switching reconfiguration time and scalability work has been pursued using semiconductor amplifiers as hybrid packet/circuit switches [51]. NEC has proposed Proteus which adds scalability through additional wavelength selective switching [52]. Reference [49] in reviewing the hybrid electrical optical experiments concludes that there are software challenges that need to be addressed. Information regarding the temporal and spatial heterogeneity of data center traffic and the analysis of application needs requires further effort and information to support the dynamic switch scheduling proposed in these architectures. They propose solutions for the control framework and the use of OpenFlow to resolve the challenges. These hybrid proposals have added value in that they have brought new concepts and potential solutions to the attention of those outside of the photonics community raising awareness of the possibility of using optics in computer networks.

3.3.2 *Optical Networks on Chip*

The networks discussed above focus on solving the traffic bottlenecks in the conventional tree architecture, primarily by modifying the tree architecture itself using commercially available or near to available off-the-shelf equipment. As briefly described above, there is also a bandwidth pressure on the network from below, at the microprocessor level. As the number of processing cores per chip grows, an efficient high bandwidth interconnection network becomes essential. Silicon photonics-based optical interconnects, leveraging the capacity and transparency of optics and fabricated in high volume CMOS compatible foundries, form the foundation of a vision solving communications bottlenecks. For many years researchers have recognized that the relatively high cost of adoption of photonics in computer systems might be overcome if the photonic components could be made in fabrication environments compatible with silicon-based electronics [53]. This section very briefly outlines components and some highlights of this burgeoning research direction.

The ingredients for an optical network on chip have been investigated and several systems proposed. Starting with the waveguides we have seen a steady improvement in quality and loss characteristics [54]. Waveguide loss characteristics depend on the geometry and fabrication technology [53, 55–57]. Reference [53] reports a very low-loss hybrid silicon waveguide circuit consisting of straight rib sections with a propagation loss of 0.272 ± 0.012 dB/cm and compact photonic wire

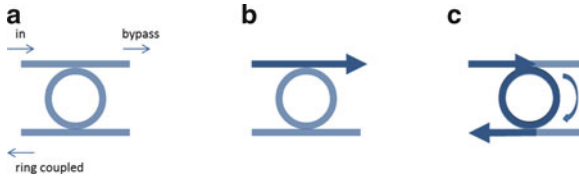


Fig. 3.4 (a) Schematic of basic ring resonator design and function (b) When the transmission wavelength does not fall within the resonance of the ring (the circumference of the ring is not an integer multiple of the wavelength of the light), the light passes through to the bypass output port, (c) When the wavelength falls within the resonance band of the ring (the circumference of the ring is an integer multiple of the wavelength of the light), light is coupled from the input waveguide into the ring and then to the ring coupled drop port

bends of $5\ \mu\text{m}$ radius with a loss of $0.0273 \pm 0.0004\ \text{dB}/90^\circ$ bend. In [56] Oracle and Kotura present low loss shallow-ridge silicon waveguides with an average propagation loss of $0.274\ \text{dB}/\text{cm}$ in the C-band. New etchless techniques are also being proposed [57].

A high speed modulator is a key component of the optical link. Significant progress has been made in advancing the silicon photonics optical modulator based on both silicon Mach Zehnder modulators and electronically tuned ring resonators (Fig. 3.4) [58, 59].

Many groups are exploring technologies to reduce power consumption and increase bandwidth and fabrication tolerances. An all-silicon optical modulator using a CMOS compatible fabrication process with a data rate of $40\ \text{Gb}/\text{s}$ and extinction ratio up to approximately $6.5\ \text{dB}$ for TE and TM polarizations is demonstrated in [60]. Intel demonstrated a high-speed silicon optical modulator based on the free carrier plasma dispersion effect based on carrier depletion of a pn diode embedded inside a silicon-on-insulator waveguide. A travelling-wave design was used resulting in a $3\ \text{dB}$ bandwidth of $\sim 30\ \text{GHz}$ and data transmission up to $40\ \text{Gb}/\text{s}$ [61].

Low power consumption is a critical requirement of silicon photonics and the modulator in particular. Many groups have explored ways to reduce power consumption [62–65]. Oracle demonstrated standard ring resonators with driver circuits with $<100\ \text{fJ}/\text{b}$ [62] Reference [65] reviews vertical junction microdisk modulators and their potential for ultra-low power consumption, demonstrating the first sub- $100\ \text{fJ}/\text{bit}$ silicon modulators.

Spectrally aligned networks of such ring resonator modulators and filters are being proposed for photonic on-chip interconnection networks [66, 67]. Broadband switches have also been proposed and demonstrated [68, 69]. Reference [68] reported the fabrication and experimental verification of a multiwavelength high-speed 2×2 silicon photonic switch for ultrahigh-bandwidth message routing in optical on-chip networks. The structure employed two microring resonators in order to implement the bar and cross states of the switch.

An important aspect of this research, particularly for networks such as those proposed with thousands of ring resonators, is the effort to achieve low power tuning and trimming of the rings. Several methods are under consideration including joule heating and the use of an overlay of thermally compensating materials [70–74].

Germanium is the preferred choice for the photodetector in the silicon CMOS-based link, [75–78] Germanium-based photodetectors can be monolithically integrated with silicon and compatible with CMOS fabrication technology. Reference [75] demonstrated waveguide integrated germanium detectors a low capacitance of 2.4 fF and directly recorded impulse response at 8.8 p. Intel [76] demonstrated < 1 fF capacitance and 0.9A/W, with a slightly higher impulse response of 12.5 ps.

A remaining challenge is the choice of a light source. It is well known that, because it is an indirect band gap material and despite extensive efforts [79–84], an efficient manufacturable silicon laser has not yet been achieved. Some have chosen to bypass this challenge by using an off-chip light source. An off-chip light source already exists commercially at fairly low cost and is more easily serviceable and replaceable. The power consumption of the off-chip laser, although part of the whole system, does not contribute to the heat dissipation challenge of the integrated chip. On the other hand, the off-chip laser introduces additional packaging and alignment challenges with the need for on-chip distribution. An efficient on-chip source would not require coupling and could be packaged more compactly and would make more efficient use of the optical power. Of course here the challenge is the development of the new laser. This laser should be suitable for high volume manufacture to maintain low cost of the silicon photonic circuit. Currently the most promising sources are the hybrid laser being developed by Intel and UCSB [82] and the germanium laser being developed by MIT and APIC [83, 84].

The above discussion shows that the elements comprising a silicon photonic network on chip have been demonstrated in research laboratories and that several network architectures have been proposed. While work continues to improve device performance and power consumption, effort is now focusing on developing and demonstrating solutions that will be most manufacturable in terms of cost, yield, and compatibility with standard CMOS processes.

3.4 Conclusions

In the last few years the data center has undergone extraordinary change, increasing influence on our lives. At the same time, at the processor level, advances and trends towards increasing numbers of cores are putting more pressure on the processor to processor and processor to memory interconnects. The immense communication bandwidth requirements are driving research groups and industries to seek solutions to these challenges by using the transmission capabilities of optics. As described above, in this brief and necessarily incomplete summary, results from system level research for the data center down to intra chip interconnects show that optics can offer solutions and scalability for future generations, although there are many

technical challenges to resolve. Optical switch fabrics can add to high speed low latency capabilities; however, novel routing and scheduling algorithms must be developed to demonstrate their benefits to throughput and performance. Tremendous advances are ongoing in silicon photonic components, with development underway to demonstrate that they are economical, power efficient, and manufacturable. Optics is poised to make a significant impact in the data center, in the first instance, to overcome current bottlenecks and, in the longer term, to enable new architectures and applications.

Acknowledgements The author would like to thank many colleagues and collaborators who have contributed to these ideas through research collaborations and discussions, special thanks to Keren Bergman, Robert Killey, Lily Mummert, Phil Watts, and Kevin Williams.

References

1. Qian D, Huang M-F, Ip E, Huang Y-K, Shao Y, J Hu, Wang T (2012) High capacity/spectral efficiency 101.7-Tb/s WDM transmission using PDM-128QAM-OFDM Over 165-km SSMF within C- and L-bands. *J Lightwave Technol* 30(10):1540–1548
2. Kash JA, Benner A, Doany FE, Kuchta D, Lee BG, Pepeljugoski P, Schares L, Schow C, Taubenblatt M (2011) Optical interconnects in future servers. In: *Optical fiber communication conference*, Paper OWQ1
3. Benner AF, Ignatowski M, Kash JA, Kuchta DM, Ritter MB (2005) Exploitation of optical interconnects in future server architectures. *IBM J Res Dev* 49(4/5):755
4. Schow C, Doany F, Kash J (2010) Get on the optical bus. *IEEE Spectrum* 47(9):32–56
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*, pp 178–181, August 2008
6. Farrington N, Porter G, Radhakrishnan S, Bazzaz HH, Subramanya V, Fainman Y (2010) Helios: a hybrid electrical/optical switch architecture for modular data centers. *ACM SIGCOMM Comp Comm Rev* 40(4):339–350
7. Glick M, Andersen DG, Kaminsky M, Mummert L (2009) Dynamically reconfigurable optical links for high-bandwidth data center networks. In: *Optical Fiber Communication Conference, OFC 2009*, pp 1–3
8. Wang G, Andersen DG, Kaminsky M, Papagiannaki K, Ng TS, Kozuch M, Ryan M (2010) c-Through: part-time optics in data centers. *ACM SIGCOMM Comp Comm Rev* 40(4):327–338
9. Petracca M, Lee BG, Bergman K, Carloni LP (2009) Photonic NoCs: system-level design exploration. *IEEE Micro* 29(4):74–85
10. Batten C, Joshi A, Orcutt J, Khilo A, Moss B, Holzwarth CW, Popovic MA, Li H, Smith HI, Hoyt JL, Kartner FX, Ram RJ, Stojanovic V, Asanovic K (2009) Building many-core processor-to-DRAM networks with monolithic CMOS silicon photonics. *IEEE Micro* 29(4)
11. Arregoces M, Portolani M (2003) *Data center fundamentals*. Data Center Fundamentals Cisco Press. ISBN: 1587050234
12. Hoelzle U, Barroso LA (2009) *The datacenter as a computer: an introduction to the design of warehouse-scale machines (synthesis lectures on computer architecture)*. Morgan and Claypool Publishers. (<http://www.morganclaypool.com/doi/pdf/10.2200/S00193ED1V01Y200905CAC006>). ISBN: 159829556X
13. Katz RH (2009) Tech Titans building boom. *IEEE Spectrum* 46(2):40–54

14. Armbrust M et al Above the clouds; A Berkeley view of cloud computing. <http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009--28.pdf>. Accessed June 2012
15. Netflix now biggest source of internet traffic in North America. http://www.huffingtonpost.com/2011/05/17/biggest-source-of-us-inte_n_863474.html. Accessed June 2012
16. Kozuch M, Campbell J, Glick M and Pillai P (2010) Cloud computing on rich data. *Intel Technol J* 14(1). <http://www.intel.com/technology/itj/2010/v14i1/index.htm>
17. Dean J, Ghemawat S (2008) MapReduce: simplified data processing on large clusters. In: *Communications of the ACM - 50th anniversary issue*, vol 51, no 1. ACM , New York, pp 107–113
18. Ke Y, Sukthankar R, Hebert M (2007) Event detection in crowded videos. In: *Proceedings of International Conference on Computer Vision*, 2007, pp 1–8
19. Leininger B A next-generation system enables persistent surveillance of wide areas. <http://spie.org/x23645.xml>. Accessed July 2012
20. Vangal S et al (2007) An 80-tile 1.28 TFLOPS network-on-chip in 65 nm CMOS. In: *Intl. solid state circuits conference*, Feb 2007, pp 98–100
21. Patterson D (2010) The trouble with multicore. *IEEE Spectrum* 47(7):28–32
22. Young IA, Mohammed E, Liao JTS, Kern AM, Palermo S, Block BA, Reshotko MR, Chang PLD (2010) Optical I/O technology for tera-scale computing. *IEEE J Solid-State Circ* 45(1):235–248
23. Balamurugan G, Casper B, Jaussi JE, Mansuri M, O'Mahony F, Kennedy J (2009) Modeling and analysis of high-speed I/O links. *IEEE Trans Adv Packaging* 32(2):237–247
24. Miller DAB (2009) Device requirements for optical interconnects to silicon chips. *Proc IEEE* 97(7):1166–1185
25. Young IA, Mohammed E, Liao JTS, Kern AM, Palermo S, Block BA, Reshotko MR, Chang PLD (2010) Optical technology for energy efficient I/O in high performance computing. *IEEE Comm Mag* 48(10):184–191
26. Gill P, Greenberg A, Jain N, Nagappan N (2011) Understanding network failures in data centers: measurement, analysis, and implications. *ACM Sigcomm* 41(4):350–361
27. Barham P et al (2003) Xen and the art of virtualization. In: *ACM SIGOPS operating systems review archive*, vol 37, no 5 (table of contents SOSP '03), pp 164–177
28. Al-Fares M, Radhakrishnan S, Raghavan B, Huang N, Vahdat A (2010) Hedera: dynamic flow scheduling for data center networks. In: *USENIX NSDI, NSDI'10 Proceedings of the 7th USENIX conference on Networked systems design and implementation*, April 2010 pp 19–20 2010
29. Kandula S, Sengupta S, Greenberg A, Patel P, Chaiken R The nature of data center traffic: measurements & analysis. In: *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, 04–06 November 2009, Chicago, IL, USA
30. Glick M, Benlachar Y, Killey RI (2009) Performance and power consumption of digital signal processing based transceivers for optical interconnect applications. In: *11th International Conference on Transparent Optical Networks, ICTON 2009*, pp 1–4
31. Barroso LA, Dean J, Holzle U (2003) Web search for a planet: the Google cluster architecture. *IEEE Micro* 23(2):22–28
32. Report to Congress on Server and Data Center eEnergy Efficiency, Public Law 109–431" US Environmental Protection Agency ENERGY STAR Program. http://www.energystar.gov/ia/partners/prod_development/downloads/EPA_Datacenter_Report_Congress_Final1.pdf. Accessed 2 August 2007
33. Koomey JG Estimating total power consumption by servers in the U.S. and the world. <http://sites.amd.com/de/Documents/svrpwrusecompletefinal.pdf>. Accessed May 2012
34. Koomey JG Growth in data center electricity usage 2005 to 2010. <http://www.migrationsolutions.co.uk/Content/Uploads/koomeydatacenterlectuse2011.pdf>. Accessed July 2012
35. Weiss A (2007) netWorker. Vol. 11:Issue 4.
36. Tucker R International Workshop on the Cloud/Grid/Utility Computing over Optical Networks OFC/NFOEC 2009, <http://www.cse.buffalo.edu/Cloud/>. Accessed August 2012

37. Frachtenberg E, Heydari A, Li H, Michael A, Na J, Nisbet A, Sarti P (2011) High efficiency server design. In: Proceedings of the 24th IEEE/ACM international conference on high performance computing, networking, storage and analysis (SC) Seattle, WA, November 2011. Facebook server room tour <http://www.youtube.com/watch?v=nhOo1ZtrH8c&feature=g-hist&context=G2a51b55AHT0RQGAABAA>
38. Barroso LA, Holzle U (2007) The case for energy-proportional computing. *IEEE Comp* 40:12
39. Abts D, Marty MR, Wells PM, Klausler P, Liu H (2010) Energy proportional datacenter networks. In: International Symposium on Computer Architecture, ACM (2010), pp 338–347
40. Farrington N, Rubow E, Vahdat A Data center switch architecture in the age of merchant silicon. In: 7th IEEE Symposium on High Performance Interconnects, pp 93–102
41. Al-Fares M, Loukissas A, Vahdat A (2008) A scalable, commodity, data center network architecture. In: Proceedings of ACM SIGCOMM, Seattle, WA, Aug 2008
42. Greenberg A, Jain N, Kandula S, Kim C, Lahiri P, Maltz D, Patel P, Sengupta S (2009) VL2: A scalable and flexible data center network. In: Proceedings of ACM SIGCOMM, Barcelona, Spain, Aug 2009
43. Guo C, Wu H, Tan K, Shi L, Zhang Y, Lu S (2008) DCell: a scalable and fault-tolerant network structure for data centers. In: Proceedings of ACM SIGCOMM, Seattle, WA, Aug 2008
44. Guo C, Lu G, Li D, Wu H, Zhang X, Shi Y, Tian C, Zhang Y, Lu S (2009) BCube: a high performance, server-centric network architecture for modular data centers. In: Proceedings of ACM SIGCOMM, Barcelona, Spain, Aug 2009
45. Mysore RN, Pamboris A, Farrington N, Huang N, Miri P, Radhakrishnan S, Subramanya V, Vahdat A (2009) Portland: a scalable fault-tolerant layer2 data center network fabric. In: Proceedings of ACM SIGCOMM, Barcelona, Spain, Aug 2009
46. Barker KJ et al On the feasibility of optical circuit switching for high performance computing systems. In: Proceedings of the ACM/IEEE SC 2005 Conference on Supercomputing, pp 16
47. Schares L, Zhang XJ, Wagle R, Rajan D, Selo P, Chang SP, Giles J, Hildrum K, Kuchta D, Wolf J, Schenfeld E (2009) A reconfigurable interconnect fabric with optical circuit switch and software optimizer for streamcomputing systems. In: Conference on Optical Fiber Communication, OFC 2009, pp 1–3
48. Wang G, Andersen DG, Kaminsky M, Kozuch M, Ng TSE, Papagiannaki K, Glick M, Mummert L Your data center is a router: the case for reconfigurable optical circuit switched paths. In: ACM HotNets'09
49. Bazzaz HH, Tewari M, Wang G, Porter G, Ng TSE, Andersen DG, Kaminsky M, Kozuch MA, Vahdat A (2011) Switching the optical divide: Fundamental challenges for hybrid electrical/optical datacenter networks. In: Proceedings of the 2nd ACM Symposium on Cloud Computing, pp 30
50. Farrington N, Fainman Y, Liu H, Papen G, Vahdat A (2011) Hardware requirement for optical circuit switched data center networks. In: Optical fiber conference (OFC/NFOEC'11), Mar 2011
51. Wang H, Garg AS, Bergman K, Glick M Design and demonstration of an all-optical hybrid packet and circuit switched network platform for next generation data centers. In: Conference on Optical Fiber Communication (OFC), 2010 (OFC/NFOEC), pp 1–3
52. Singla A, Singh A, Ramachandran K, Xu L, Zhang Y (2010) Proteus: a topology malleable data center network. In: ACM HotNets, Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks, Article no. 8
53. Soref R (2006) The past, present, and future of silicon photonics. *IEEE J Sel Top Quant Electron* 12(6):1678–1687
54. Selvaraja SK, Bogaerts W, Dumon P, Van Thourhout D, Baets RG (2010) Subnanometer linewidth uniformity in silicon nanophotonic waveguide devices using CMOS fabrication technology. *IEEE J Sel Top Quant Electron* 16(1):316–324
55. Selvaraja SK, Bogaerts W, Absil P, Thourhout DV, Baets R (2010) Record low-loss hybrid rib/wire waveguides for silicon photonic circuits. In: 7th International conference on Group IV Photonics, pp 1–3

56. Dong P, Qian W, Liao S, Liang H, Kung C-C, Feng N-N, Shafiiha R, Fong J, Feng D, Krishnamoorthy AV, Asghari M (2010) Low loss silicon waveguides for application of optical interconnects. In: Photonics society summer topical meeting series, IEEE, 19–21 July 2010, pp 191–192
57. Cardenas J, Poitras C, Robinson J, Preston K, Chen L, Lipson M (2009) Lowloss etchless silicon photonic waveguides. *Opt Express* 17(6):4752–4757
58. Lipson M (2006) Compact electro-optic modulators on a Silicon chip. *J Sel Top Quant Electron* 12:1520
59. Marris-Morini D, Vivien L, Rasigade G, Fedeli J-M, Cassan E, Le Roux X, Crozat P, Maine S, Lupu A, Lyan P, Rivallin P, Halbawax M, Laval S (2009) Recent progress in high-speed Silicon-based optical modulators. *Proc IEEE* 97(7):1199–1215
60. Gardes FY, Thomson DJ, Emerson NG, Reed GT (2011) 40 Gb/s silicon photonics modulator for TE and TM polarisations. *Opt Express* 19(12):11804–11814
61. Liao L, Liu A, Rubin D, Basak J, Chetrit Y, Nguyen H, Cohen R, Izhaky N, Paniccia M (2007) 40 Gbit/s silicon optical modulator for high speed applications. *Electron Lett* 43(22):1196–1197
62. Zheng X, Liu F, Lexau J, Patil D, Li G, Luo Y, Thacker H, Shubin I, Yao J, Raj K, Ho R, Cunningham JE, Krishnamoorthy AV (2011) Ultra-low power arrayed CMOS Silicon photonic transceivers for an 80 Gbps WDM optical link. In: Optical fiber communication conference (OFC 2011), Paper PDPA
63. Rosenberg JC, Green WM, Assefa S, Barwicz T, Yang M, Shank SM, Vlasov YA (2011) Low-power 30 Gbps silicon microring modulator. In: CLEO- laser applications photonic applications, OSA Tech. Dig, Baltimore, MD, 2011, Paper PDPB9
64. Miller DAB (2012) Energy consumption in optical modulators for interconnects. *Opt Express* 20(S2):A293
65. Watts MR, Zortman WA, Trotter DC, Young RW, Lentine AL (2011) Vertical junction silicon microdisk modulators and switches. *Opt Express* 19(22):21989–22003
66. Vantrease D, Schreiber R, Monchiero M, McLaren M, Jouppi NP, Fiorentino M, Davis A, Binkert N, Beausoleil RG, Ahn JH (2008) Corona: System implications of emerging nanophotonic technology. In: Proceedings of the 35th international symposium on computer architecture, Beijing, China, June 2008
67. Joshi A, Batten C, Kwon Y-J, Beamer S, Shamim I, Asanovic K, Stojanovic V (2009) Silicon-photonic crosstalk networks for global on-chip communication. In: Proceedings of the 2009 3rd ACM/IEEE international symposium on networks-on-chip, pp 124–133, 10–13 May 2009
68. Lee BG, Biberman A, Sherwood-Droz N, Poitras CB, Lipson M, Bergman K (2009) High-speed 2×2 switch for multiwavelength silicon-photonic networks-on-chip. *J Lightwave Technol* 27(14):2900–2907
69. Yang M, Green WMJ, Assefa S, Van Campenhout J, Lee BG, Jahnes CV, Doany FE, Schow CL, Kash JA, Vlasov Y (2011) A Non-blocking 4×4 electro-optic silicon switch for on-chip photonic networks. *Opt Express* 19(1):47–54
70. Zortman WA, Lentine AL, Trotter DC, Watts MR (2011) Low-voltage differentially-signaled modulators. *Opt Express* 19(27):26017–26026
71. DeRose CT, Watts MR, Trotter DC, Luck DL, Nielson GN, Young RW Silicon microring modulator with integrated heater and temperature sensor for thermal control. In: Conference on lasers and electro-optics, OSA Technical Digest (CD), Paper CThJ3. Optical Society of America, 2010
72. Teng J, Dumon P, Bogaerts W, Zhang H, Jian X, Han X, Zhao M, Morthier G, Baets R (2009) Athermal silicon-on-insulator ring resonators by overlaying a polymer cladding on narrowed waveguides. *Opt Express* 17:14627–14633
73. Raghunathan V, Ye WN, J Hu, Izuhara T, Michel J, Kimerling L (2010) Athermal operation of Silicon waveguides: spectral, second order and footprint. *Optics Express* 18(17):17631–17639
74. Guha B, Kyotoku BB, Lipson M (2010) CMOS-compatible athermal silicon microring resonators. *Opt Express* 18(4):3487–3493

75. Chen L, Lipson M (2009) Ultra-low capacitance and high speed germanium photodetectors on silicon. *Opt Express* 17(10):7901–7906
76. Reshotko MR, Block BA, Jin B, Chang P (2008) Waveguide coupled Ge-on-oxide photodetectors for integrated optical links. In: 5th IEEE international conference on group IV photonics, 2008, pp 182–184
77. Feng N-N, Dong P, Zheng D, Liao S, Liang H, Shafiiha R, Feng D, Li G, Cunningham JE, Krishnamoorthy AV, Asghari M (2010) Vertical p-i-n germanium photodetector with high external responsivity integrated with large core Si waveguides. *Opt Express* 18(1):96–101
78. Ahn D, Hong C-Y, Liu J, Giziewicz W, Beals M, Kimerling LC, Michel J, Chen J, Kärtner FX (2007) High performance, waveguide integrated Ge photodetectors. *Opt Express* 15(7):3916–3921
79. Pavesi L, Lockwood DJ (2004) *Silicon photonics*. Springer, New York
80. Rong H et al (2005) A continuous-wave Raman silicon laser. *Nature* 433:725–728
81. Boyraz O, Jalali B (2004) Demonstration of a silicon Raman laser. *Opt Express* 12:5269
82. Fang AW, Park H, Cohen O, Jones R, Paniccia M, Bowers JE (2006) Electrically pumped hybrid AlGaInAs-silicon evanescent laser. *Opt Express* 14:9203–9210
83. Sun X, Liu J, Kimerling LC, Michel J (2010) Toward a germanium laser for integrated silicon photonics. *IEEE Sel Top Quant Electron* 16:124–131
84. Michel J, Camacho-Aguilera RE, Gai Y, Patel N, Bessette JT, Romagnoli M, Dutt R, Kimerling L An electrically pumped Ge on Si laser. In: OFC 2012 PDP5A.6

Chapter 4

Simulation and Performance Analysis of Data Intensive and Workload Intensive Cloud Computing Data Centers

Dzmitry Kliazovich, Pascal Bouvry, and Samee Ullah Khan

4.1 Introduction

Data centers are becoming increasingly popular for the provisioning of computing resources. The cost and operational expenses of data centers have skyrocketed with the increase in computing capacity [1]. Energy consumption is a growing concern for data center operators. It is becoming one of the main entries on a data center operational expenses (OPEX) bill [2,3]. The Gartner Group estimates energy consumptions to account for up to 10% of the current OPEX, and this estimate is projected to rise to 50% in the next few years [4]. However, computing-based energy consumption is not the only power-related portion of the OPEX bill. High power consumption generates heat and requires an accompanying cooling system that costs in a range of \$2–\$5 million per year for classical data centers [5]. Failure to keep data center temperatures within operational ranges drastically decreases hardware reliability and may potentially violate the service level agreement (SLA) with the customers.

From the perspective of energy efficiency, a cloud computing data center can be defined as a pool of computing and communication resources organized in the way to transform the received power into computing or data transfer work to satisfy user demands. The first power saving solutions focused on making the data center hardware components power efficient. Technologies, such as dynamic voltage and frequency scaling (DVFS), and dynamic power management (DPM) [6], were

D. Kliazovich (✉) • P. Bouvry
University of Luxembourg, 6 rue Coudenhove Kalergi, Luxembourg
e-mail: Dzmitry.Kliazovich@uni.lu; pascal.bouvry@uni.lu.

S.U. Khan
North Dakota State University, Fargo, ND 58108-6050, USA
e-mail: samee.khan@ndsu.edu

extensively studied and widely deployed. Because the aforementioned techniques rely on power-down and power-off methodologies, the efficiency of these techniques is at best limited. In fact, an idle server may consume about two-thirds of its peak load [7].

Because the workload of a data center fluctuates on a weekly (and in some case on hourly) basis, it is a common practice to overprovision computing and communicational resources to accommodate the peak load. In fact, the average load accounts only for 30% of data center resources [8]. This allows putting the rest of the 70% of the resources into a sleep mode for most of the time. However, achieving the above requires central coordination and energy-aware workload scheduling techniques. Typical energy-aware scheduling solutions attempt to: (a) concentrate the workload in a minimum set of the computing resources and (b) maximize the amount of resource can be put into sleep mode [9].

Most of the current state-of-the-art research on energy efficiency has predominantly focused on the optimization of the processing elements. However, as recorded in earlier research, more than 30% of the total computing energy is consumed by the communication links, switching and aggregation elements. Similar to the case of processing components, energy consumption of the communication fabric can be reduced by scaling down the communication speeds and cutting operational frequency along with the input voltage for the transceivers and switching elements [10]. However, slowing the communicational fabric down should be performed carefully and based on the demands of user applications. Otherwise, such a procedure may result in a bottleneck, thereby limiting the overall system performance. A number of studies demonstrate that often a simple optimization of the data center architecture and energy-aware scheduling of the workloads may lead to significant energy savings. Reference [11] demonstrates energy savings of up to 75% that can be achieved by traffic management and workload consolidation techniques.

In this chapter, we survey power-saving techniques implemented at both component and system levels. In energy efficiency optimization we focus on both computing and communication fabrics. As the system level, energy-efficient network-aware scheduling solutions are presented. Finally a simulation environment, named GreenCloud, for advanced energy-aware studies of cloud computing data centers in realistic setups is presented. GreenCloud is developed as an extension of a packet-level network simulator ns-2 [12]. Unlike few existing cloud computing simulators such as CloudSim [13] or MDCCSim [14], GreenCloud extracts, aggregates, and makes information about the energy consumed by computing and communication elements of the data center available in an unprecedented fashion. In particular, a special focus is devoted to accurately capture communication patterns of currently deployed and future data center architectures.

4.2 Simulating Energy-Efficient Data Centers

In this section, we present the main aspects of design of energy-efficient data centers, survey the most prominent architectures, and describe power-saving techniques implemented by individual data center components.

4.2.1 *Energy Efficiency*

Only a part of the energy consumed by the data center gets delivered to the computing servers directly. A major portion of the energy is utilized to maintain interconnection links and network equipment operations. The rest of the electricity is wasted in the power distribution system, dissipates as heat energy, and used up by air-conditioning systems. In light of the above discussion, we distinguish three energy consumption components: (a) computing energy, (b) communicational energy, and (c) the energy component related to the physical infrastructure of a data center.

The efficiency of a data center can be defined in terms of the performance delivered per watt, which may be quantified by the following two metrics: (a) Power Usage Effectiveness (PUE) and (b) Data Center Infrastructure Efficiency (DCiE) [15, 16]. Both PUE and DCiE describe which portion of the totally consumed energy gets delivered to the computing servers.

4.2.2 *Data Center Architectures*

Three-tier trees of hosts and switches form the most widely used data center architecture [17]. It (see Fig. 4.1) consists of the core tier at the root of the tree, the aggregation tier that is responsible for routing, and the access tier that holds the pool of computing servers (or hosts). Earlier data centers used two-tier architectures with no aggregation tier. However, such data centers, depending on the type of switches used and per-host bandwidth requirements, could typically support not more than 5,000 hosts. Given the pool of servers in today's data centers that are of the order of 100,000 hosts [11] and the requirement to keep layer-2 switches in the access network, a three-tiered design becomes the most appropriate option.

Although 10 Gigabit Ethernet (GE) transceivers are commercially available, in a three-tiered architecture the computing servers (grouped in racks) are interconnected using 1 GE links. This is due to the fact that the 10 GE transceivers: (a) are too expensive and (b) probably offer more capacity than needed for connecting computing servers. In current data centers, rack connectivity is achieved with inexpensive Top-of-Rack (ToR) switches. A typical ToR switch shares two 10 GE uplinks with 48 GE links that interconnect computing servers within a rack.

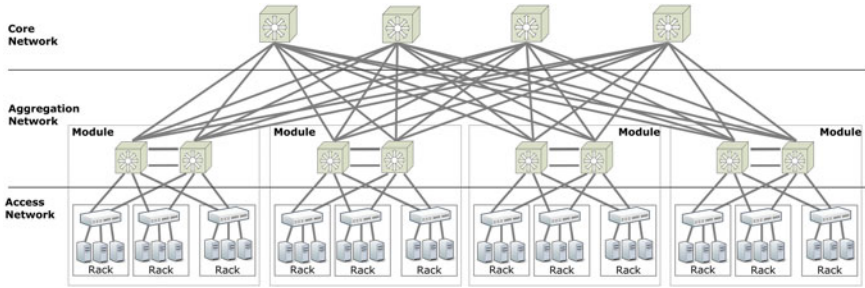


Fig. 4.1 Three-tier data center architecture

The difference between the downlink and the uplink capacities of a switch defines its oversubscription ratio, which in the aforementioned case is equal to $48/20 = 2.4 : 1$. Therefore, under full load, only 416 Mb/s will remain available to each of the individual servers out of their 1 GE links.

At the higher layers of hierarchy, the racks are arranged in modules (see Fig. 4.1) with a pair of aggregation switches servicing the module connectivity. Typical oversubscription ratios for these aggregation switches are around 1.5:1, which further reduces the available bandwidth for the individual computing servers to 277 Mbps.

The bandwidth between the core and aggregation networks is distributed using a multi-path routing technology, such as the equal cost multi-path (ECMP) routing [18]. The ECMP technique performs a per-flow load balancing, which differentiates the flows by computing a hash function on the incoming packet headers. For a three-tiered architecture, the maximum number of allowable ECMP paths bounds the total number of core switches to eight. Such a bound also limits the deliverable bandwidth to the aggregation switches. This limitation will be waved with the (commercial) availability of 100 GE links, standardized in June 2010 [19].

But how the data center architecture will look like in the future? The most promising trend in to follow a modular design. Traditional racks of servers will be replaced with standard shipping containers hosting 10 times as many servers as conventional data center in the same volume [20]. Each container is optimized for power consumption. It integrates a combined water and air cooling system and implements optimized networking solutions. These containers, being easy to ship, can become plug-and-play modules in future roof-less data center facilities [21]. Their current PUE is in the order of 1.2 [22] while the average PUE for the industry is between 1.8 and 2.0 [1] depending on the reporting source. Some skeptics addressing the problem of individual component failures and the overhead of shipping the whole container back to the manufacturer. This can be addressed by packing even more servers into self-contained container solutions requiring no operational maintenance [23]. Whenever an individual component fails the whole container can continue operation with only minor degradation in computing capacity. To make it a reality, each container as well as the data center itself

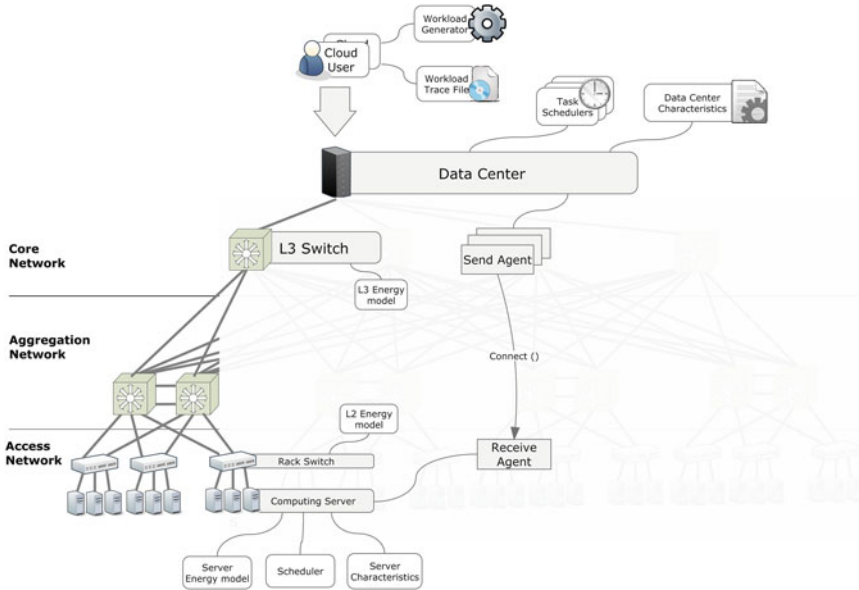


Fig. 4.2 Architecture of GreenCloud simulator

should follow a distributed design approach. But current data center architectures are completely hierarchical. This way, for example, a failure in the rack switch can disable all servers in the rack. A failure of the core or aggregation switches may degrade operation or even disable a large number of racks. Therefore, fat-tree architectures will be replaced with distributed approaches like DCell [24], BCube [25], FiConn [26], or DPillar [27] in future data centers.

4.2.3 Simulator Structure

In this section we introduce GreenCloud simulator which offers fine-grained simulation of modern cloud computing environments focusing on data center communications and energy efficiency. GreenCloud is an extension to the network simulator ns-2 [12]. It offers users a detailed fine-grained modeling of the energy consumed by the elements of the data center, such as servers, switches, and links. Moreover, GreenCloud offers a thorough investigation of workload distributions. Furthermore, a specific focus is devoted on the packet-level simulations of communications in the data center infrastructure, which provide the finest-grain control and is not present in any cloud computing simulation environment. Reference [28] provides more details on the GreenCloud simulator. Figure 4.2 presents the structure of the GreenCloud extension mapped onto the three-tier data center architecture.

4.2.4 Hardware Components and Energy Models

Computing servers are the staple of a data center that are responsible for task execution. In GreenCloud, the server components implement single core nodes that have a preset on a processing power limit in MIPS (million instructions per second) or FLOPS (floating point operations per second), associated size of the memory/storage resources, and contain different task scheduling mechanisms ranging from the simple round-robin to the sophisticated DVFS and DNS approaches.

The servers are arranged into racks with a ToR switch connecting it to the access part of the network. The power model followed by server components depends on CPU utilization. As reported in [2] and [7] an idle server consumes about two-thirds of its peak load consumption. This is due to the fact that servers must constantly manage memory modules, disks, I/O resources, and other peripherals. Moreover, the power consumption increases with the level of CPU load linearly. As a result, the aforementioned model allows implementation of power saving in a centralized scheduler that can provision consolidation of workloads in a minimum possible amount of the computing servers.

1. Another option for power management is dynamic voltage/frequency scaling (DVFS) [10], which introduces a trade-off between computing performance and the energy consumed by the server. The DVFS is based on the fact that switching power in a chip decreases proportionally to $V^2 \times f$, where V is the voltage and f is the switching frequency. Moreover, voltage reduction requires frequency downshift. This implies a cubic relationship from f in the CPU power consumption. Note that server components, such as bus, memory, and disks do not depend on the CPU frequency. Therefore, the power consumption of an average server (see Fig. 4.3) can be expressed as follows [29]:

$$P = P_{\text{fixed}} + P_f \times f^3, \quad (4.1)$$

where P_{fixed} accounts for the portion of the consumed power which does not scale with the operating frequency f , while P_f is a frequency-dependent CPU power consumption.

Network switches and links form the interconnection fabric that delivers workloads to any of the computing servers for execution in a timely manner. The interconnection of switches and servers requires different cabling solutions depending on the supported bandwidth, physical and quality characteristics of the link. The quality of signal transmission in a given cable determines a trade-off between the transmission rate and the link distance, which are the factors defining the cost and energy consumption of the transceivers.

The twisted pair is the most commonly used medium for Ethernet networks that allows organizing Gigabit Ethernet (GE) transmissions for up to 100 m with the transceiver power consumed of around 0.4 W or 10 GE links for up to 30 m with the transceiver power of 6 W. The twisted pair cabling is a low cost solution. However, for the organization of 10 GE links it is common to use optical multimode fibers.

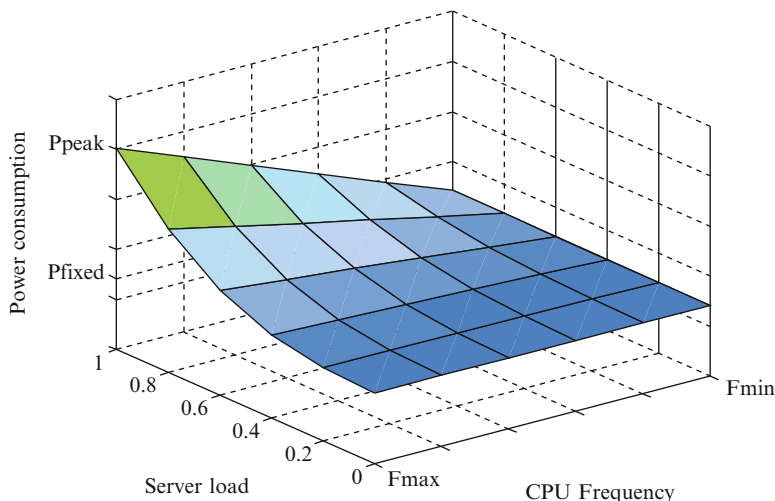


Fig. 4.3 Computing server power consumption

The multimode fibers allow transmissions for up to 300 m with the transceiver power of 1 W [30]. On the other hand the fact that multimode fibers cost almost 50 times of the twisted pair cost motivates the trend to limit the usage of 10 GE links to the core and aggregation networks as spending for the networking infrastructure may top 10%–20% of the overall data center budget [31].

The number of switches installed depends on the implemented data center architecture. However, as the computing servers are usually arranged into racks the most common switch in a data center is ToR switch. The ToR switch is typically placed at the top unit of the rack unit (1RU) to reduce the amount of cables and the heat produced. The ToR switches can support either gigabit (GE) or 10 gigabit (10 GE) speeds. However, taking into account that 10 GE switches are more expensive, current capacity limitation of aggregation and core networks gigabit rates are more common for racks.

Similar to the computing servers early power optimization proposals for interconnection network were based on DVS links [10]. The DVS introduced a control element at each port of the switch that depending on the traffic pattern and current levels of link utilization could downgrade the transmission rate. Due to the comparability requirements only few standard link transmission rates are allowed, such as for GE links 10 Mbps, 100 Mbps, and 1 Gbps are the only options.

On the other hand, the power efficiency of DVS links is limited as only a portion (3%–15%) of the consumed power which scales linearly with the link rate. As demonstrated by the experiments in [32] the energy consumed by a switch and all its transceivers can be defined as:

$$P_{\text{switch}} = P_{\text{chassis}} + n_{\text{linecards}} \times P_{\text{linecard}} + \sum_{i=0}^R n_{\text{ports},r} \times P_r, \quad (4.2)$$

where P_{chassis} is related to the power consumed by the switch hardware, P_{linecard} is the power consumed by any active network line card, P_r corresponds to the power consumed by a port (transceiver) running at the rate r . In Eq. (4.2), only the last component appears to be dependent on the link rate while other components, such as P_{chassis} and P_{linecard} remain fixed for all the duration of switch operation. Therefore, P_{chassis} and P_{linecard} can be avoided by turning the switch hardware off or putting it into sleep mode.

4.2.5 Jobs and Workloads

Workloads are the objects designed for universal modeling of various cloud user services. In grid computing the workloads are typically modeled a sequence of jobs that can be divided into a set of tasks. The tasks can be dependent requiring an output from other tasks to start execution or be independent. Moreover, due to the nature of grid computing applications (biological, financial, or climate modeling) the number of jobs available prevail the number of computing resources available. While the main goal is in minimization of the time required for the computing of all jobs which may take weeks or months the individual jobs do not have a strict completion deadline.

In cloud computing, incoming requests are typically generated for such applications like web browsing, instant messaging, or various content delivery applications. The jobs tend to be more independent, less computationally intensive, but have a strict completion deadline specified in SLA. To cover the vast majority of cloud computing applications, we define three types of jobs:

- *Computationally Intensive Workloads (CIWs)* model high-performance computing (HPC) applications aiming at solving advanced computational problems. CIWs load computing servers considerably, but require almost no data transfers in the interconnection network of the data center. The process of CIW energy-efficient scheduling should focus on the server power consumption footprint trying to group the workloads at the minimum set of servers as well as to route the traffic produced using a minimum set of routes. There is no danger of network congestion due to the low data transfer requirements, and putting the most of the switches into the sleep mode will ensure the lowest power of the data center network.
- *Data-Intensive Workloads (DIWs)*, on the contrary, produce almost no load at the computing servers, but require heavy data transfers. DIWs aim to model such applications like video file sharing where each simple user request turns into a video streaming process. As a result, the interconnection network and not the computing capacity becomes a bottleneck of the data center for DIWs. Ideally, there should be a continuous feedback from network switches to the central

workload scheduler. Based on such a feedback, the scheduler will distribute the workloads taking current congestion levels of the communication links. It will avoid sending workloads over congested links even if certain server's computing capacity will allow accommodating the workload. Such scheduling policy will balance the traffic in the data center network and reduce average time required for a task delivery from the core switches to the computing servers.

- *Balanced Workloads (BWs)* aim to model the applications having both computing and data transfer requirements. BWs load the computing servers and communication links proportionally. With this type of workloads the average load on the servers is proportional to the average load of the data center network. BWs can model such applications as geographic information systems which require both large graphical data transfers and heavy processing. Scheduling of BWs should account for both servers' load and the load of the interconnection network.

The execution of each workload object requires a successful completion of its two main components: (a) computing and (b) communicational. The computing component defines the amount of computing that has to be executed before a given deadline on a time scale. The deadline aims at introducing Quality of Service (QoS) constraints specified in SLA. The communicational component of the workload defines the amount and the size of data transfers that must be performed prior, during, and after the workload execution. It is composed of three parts: (a) the size of the workload, (b) the size of internal, and (c) the size of external to the data center communications. The size of the workload defines the number of bytes that after being divided into IP packets is required to be transmitted from the core switches to the computing servers before a workload execution can be initiated. The size of external communications defines the amount of data required to be transmitted outside the data center network at the moment of task completion and corresponds to the task execution result. The size of internal to the data center communications defines the amount of data to be exchanged with another workload that can be executed at the same or a different server. This way the workload interdependencies are modeled. In fact, internal communication in the data center can account for as much as 70% of total data transmitted [11].

Figure 4.4 captures energy consumption measured in a DVFS- and DNS-enabled data center running different types of workloads. An efficient and effective methodology to optimize energy consumption of interdependent workloads is to analyze the workload communication requirements at the moment of scheduling and perform a coupled placement of these interdependent workloads—a co-scheduling approach. The co-scheduling approach will reduce the number of links/switches involved into communication patterns.

Figure 4.5 shows a typical distribution of energy consumption between data center components obtained via simulations.

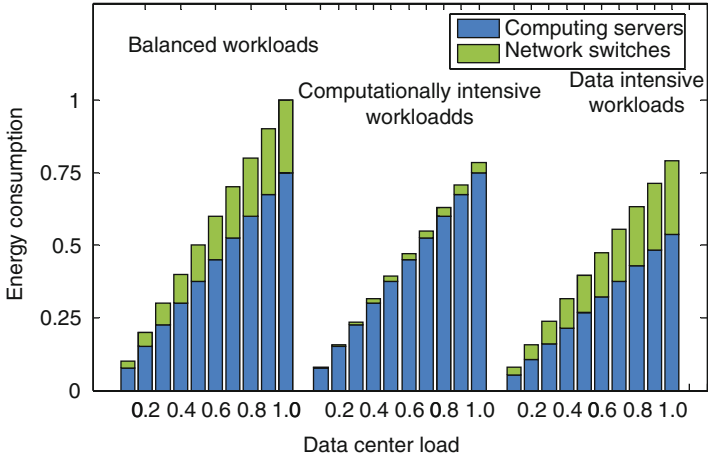


Fig. 4.4 Energy consumption for different types of workloads

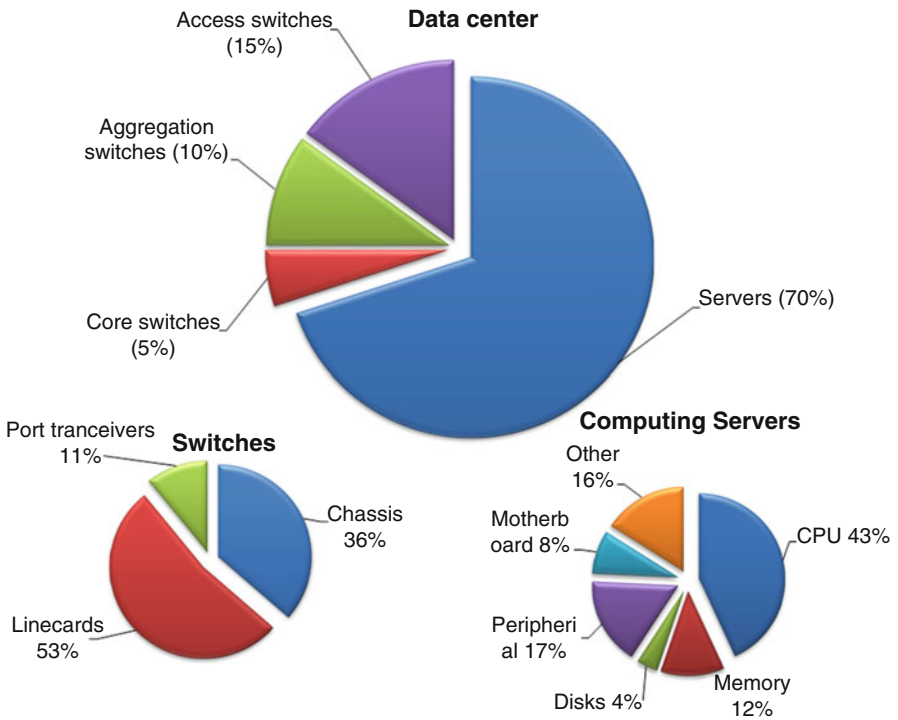


Fig. 4.5 Distribution of energy consumption in data center

4.3 Energy-Efficient Scheduling

4.3.1 Network Congestion

Utilizing a communication fabric in data centers entails the concept of running multiple types of traffic (LAN, SAN, or IPC) on a single Ethernet-based medium [33]. On one side, the Ethernet technology is cheap, easy to deploy, and relatively simple to manage, on the other side, the Ethernet hardware is less powerful and provisions for small buffering capacity. A typical buffer size in an Ethernet network is in the order of 100s of KB. However, a typical buffer size of an Internet router is in the order of 100s of MB [34]. Small buffers and the mix of high-bandwidth traffic are the main reasons for network congestion.

Any of the data center switches may become congested either in the uplink direction or in the downlink direction or both. In the downlink direction, the congestion occurs when individual ingress link capacities overcome individual egress link capacities. In the uplink direction, the mismatch in bandwidth is primarily due to the bandwidth oversubscription ratio, which occurs when the combined capacity of server ports overcomes a switch's aggregate uplink capacity.

Congestion (or hotspots) may severely affect the ability of a data center network to transport data. Currently, the Data Center Bridging Task Group (IEEE 802.1) [35] is specifying layer-2 solutions for congestion control, termed IEEE 802.1Qau specifications. The IEEE 802.1Qau specifications introduce a feedback loop between data center switches for signaling congestion. Such a feedback allows overloaded switches to hold off heavy senders from sending with the congestion notification signal. Such a technique may avoid congestion-related losses and keep the data center network utilization high. However, it does not address the root of the problem as it is much more efficient to assign data-intensive jobs to different computing servers in the way that jobs avoid sharing common communication paths. To benefit from such spatial separation in the three-tiered architecture (see Fig. 4.1), the jobs must be distributed among the computing servers in proportion to their communication requirements. Data-intensive jobs, like ones generated by video sharing applications, produce a constant bit-stream directed to the end-user as well as communicate with other jobs running in the data center. However, such a methodology contradicts the objectives of energy-efficient scheduling, which tries to concentrate all of the active workloads on a minimum set of servers and involve minimum number of communication resources. This trade-off between energy efficiency, data center network congestion, and performance of individual jobs is resolved using a unified scheduling metric presented in the subsequent section.

4.3.2 The DENS Methodology

The DENS methodology minimizes the total energy consumption of a data center by selecting the best-fit computing resources for job execution based on the load level

and communication potential of data center components. The communicational potential is defined as the amount of end-to-end bandwidth provided to individual servers or group of servers by the data center architecture. Contrary to traditional scheduling solutions [36] that model data centers as a homogeneous pool of computing servers, the DENS methodology develops a hierarchical model consistent with the state-of-the-art data center topologies. For a three-tier data center, DENS metric M is defined as a weighted combination of server-level f_s , rack-level f_r , and module-level f_m functions:

$$M = \alpha \times f_s + \beta \times f_r + \gamma \times f_m, \quad (4.3)$$

where α , β , and γ are weighted coefficients that define the impact of the corresponding components (servers, racks, and/or modules) on the metric behavior. Higher values of α favor the selection of highly loaded servers in lightly loaded racks. Higher values of β will prioritize computationally loaded racks with low network traffic activity. Higher values of γ favor selection of lightly loaded modules. The γ parameter is an important design variable for job consolidation in data centers. Taking into account that $\alpha + \beta + \gamma$ must equal unity, the values of $\alpha = 0.7$, $\beta = 0.2$, and $\gamma = 0.1$ are selected experimentally to provide a good balance in the evaluated three-tier data center topology. The details of the selection process are presented in [37].

The factor related to the choice of computing servers combines the server load $L_s(l)$ and its communication potential $Q_r(q)$ that corresponds to the fair share of the uplink resources on the ToR switch. This relationship is given as:

$$f_s(l, q) = L_s(l) \times \frac{Q_r(q)^\phi}{\delta_r}, \quad (4.4)$$

where $L_s(l)$ is a factor depending on the load of the individual servers l , $Q_r(q)$ defines the load at the rack uplink by analyzing the congestion level in the switch's outgoing queue q , δ_r is a bandwidth over provisioning factor at the rack switch, and ϕ is a coefficient defining the proportion between $L_s(l)$ and $Q_r(q)$ in the metric. Given that both $L_s(l)$ and $Q_r(q)$ must be within the range $[0, 1]$ higher ϕ values will decrease the importance of the traffic-related component $Q_r(q)$. Similar to the case of computing servers, which was encapsulated in Eq. (4.4), the factors affecting racks and modules can be formulated as:

$$f_r(l, q) = L_r(l) \times \frac{Q_m(q)^\phi}{\delta_m} = \frac{Q_m(q)^\phi}{\delta_m} \times \frac{1}{n} \sum_{i=1}^n L_s(l), \quad (4.5)$$

$$f_m(l) = L_m(l) = \frac{1}{k} \sum_{j=0}^k L_r(l), \quad (4.6)$$

where $L_r(l)$ is a rack load obtained as a normalized sum of all individual server loads in the rack, $L_m(l)$ is a module load obtained as a normalized sum of all of

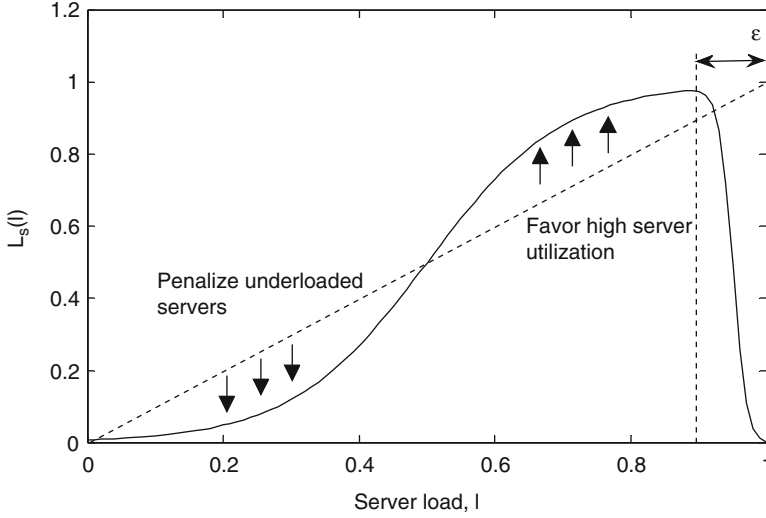


Fig. 4.6 Computing server selection by DENS metric

the rack loads in this module, n and k are the number of servers in a rack and the number of racks in a module respectively, $Q_m(q)$ is proportional to the traffic load at the module ingress switches, and δ_m stands for the bandwidth over provisioning factor at the module switches. It should be noted that the module-level factor f_m includes only a load-related component l . This is due to the fact that all the modules are connected to the same core switches and share the same bandwidth using ECMP multi-path balancing technology.

The fact that an idle server consumes energy that is almost two-thirds of its peak consumption [7] suggests that an energy-efficient scheduler must consolidate data center jobs on a minimum possible set of computing servers. On the other hand, keeping servers constantly running at peak loads may decrease hardware reliability and consequently affect the job execution deadlines [38]. To address the aforementioned issues, we define the DENS load factor as a sum of two sigmoid functions:

$$L_s(l) = \frac{1}{1 + e^{-10(l - \frac{1}{2})}} - \frac{1}{1 + e^{-\frac{10}{\epsilon}(l - (1 - \frac{\epsilon}{2}))}}. \quad (4.7)$$

The first component of Eq. (4.7) defines the shape of the main sigmoid, while the second component is a penalizing function aimed at the convergence towards the maximum server load value (see Fig. 4.6). The parameter ϵ defines the size and the incline of this falling slope. The server load l is within the range $[0, 1]$. For the tasks having deterministic computing load the server load can be computed as the sum of computing loads of all of the running tasks. Alternatively, for the tasks with predefined completion deadline, the server load l can be expressed as the minimum amount of computational resource required from the server to complete all the tasks right-in-time.

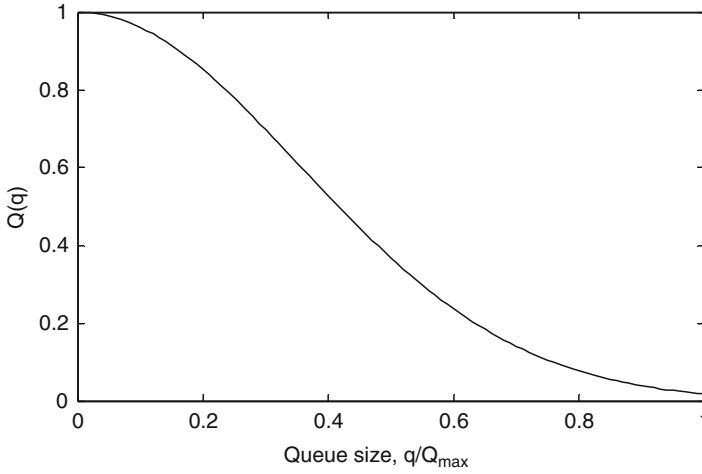


Fig. 4.7 Queue selection by DENS metric

Being assigned into racks, computing servers share the same ToR switch their uplink communication demands. However, defining a portion of this bandwidth used by a given server or a flow at the gigabit speeds during runtime is a computationally expensive task. To circumvent the aforementioned undesirable characteristic, both Eqs. (4.4) and (4.5) include a component that is dependent on the occupancy level of the outgoing queue $Q(q)$ at the switch and scales with the bandwidth over provisioning factor δ .

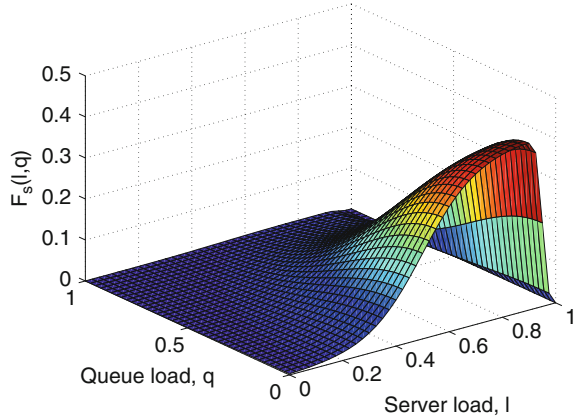
Instead of relying on the absolute size of the queue, the occupancy level q is scaled with the total size of the queue Q_{\max} within the range $[0, 1]$. The range corresponds to none and full buffer occupancy. By relying on buffer occupancy, the DENS metric reacts to the growing congestion in racks or modules rather than transmission rate variations. To satisfy the aforementioned behavior, $Q(q)$ is defined using inverse Weibull cumulative distribution function:

$$Q(q) = e^{-\left(\frac{2q}{Q_{\max}}\right)^2}. \quad (4.8)$$

The obtained function, illustrated in Fig. 4.7, favors empty queues and penalizes fully loaded queues. Being scaled with the bandwidth over provisioning factor δ in Eqs. (4.4) and (4.5) it favors the symmetry in the combined uplink and downlink bandwidth capacities for switches when congestion level is low. However, as congestion grows and buffers overflow, the bandwidth mismatch becomes irrelevant and immeasurable.

Figure 4.8 presents the combined $f_s(l, q)$ as defined in Eq. (4.4). The obtained bell-shaped function favors selection of servers with the load level above average located in racks with the minimum or no congestion. Reference [37] provides more details about DENS metrics and its performance in different operation scenarios.

Fig. 4.8 Server selection by DENS metric according to its load and communicational potential



4.4 Conclusions

The cost and operating expenses of data centers are becoming a growing concern as cloud computing industry is booming. The challenge of energy efficiency allows maintaining the same data center performance while the level of energy consumption is reduced. This can not only significantly reduce costs of operating the IT equipment and cooling but also increase server density enlarging the capacity of existing data center facilities.

To understand the optimization space we surveyed energy consumption models of computing servers, network switches, and communication links. Thereafter, main techniques for energy efficiency, like DVFS or dynamic shut-down, are studied at both the component and system levels. It is demonstrated that approaches for centralized coordination and scheduling are required to achieve satisfactory optimization levels. Such coordination should combine traditional scheduling approaches with the awareness of the state of communication equipment and network traffic footprints. Furthermore, the characteristics of the incoming workloads must be taken into account. Currently, GreenCloud simulator and presented energy-aware scheduling approaches are being extended to cover scenarios which include geographically distributed data centers and renewable sources of energy.

References

1. Brown R, Chan P, Eto J, Jarvis S, Koomey J, Masanet E, Nordman B, Sartor D, Shehabi A, Stanley J, Tschudi B (2007) Report to congress on server and data center energy efficiency: Public law 109–431. Lawrence Berkeley National Laboratory. 1–130. Available at http://www.energystar.gov/ia/partners/prod_development/downloads/EPA_Datacenter_Report_Congress_Final1.pdf
2. Fan X, Weber W-D, Barroso LA (2007) Power provisioning for a warehouse-sized computer. In: ACM international symposium on computer architecture, San Diego, CA, June 2007

3. Raghavendra R, Ranganathan P, Talwar V, Wang Z, Zhu X (2008) No “Power” struggles: coordinated multi-level power management for the data center. In: SIGOPS Oper. Syst. Rev. 42(2): 48–59
4. Gartner Group. Available at: <http://www.gartner.com/>, Accessed Aug 2012
5. Moore J, Chase J, Ranganathan P, Sharma R (2005) Making scheduling “Cool”: temperature-aware workload placement in data centers. In: USENIX annual technical conference (ATEC '05). USENIX Association, Berkeley, CA, USA, pp 5–5
6. Horvath T, Abdelzaher T, Skadron K, Liu X (2007) Dynamic voltage scaling in multitier web servers with end-to-end delay control. *IEEE Trans Comp* 56(4):444–458
7. Chen G, He W, Liu J, Nath S, Rigas L, Xiao L, Zhao F (2008) Energy-aware server provisioning and load dispatching for connection-intensive internet services. In: The 5th USENIX symposium on networked systems design and implementation, Berkeley, CA, USA
8. Liu J, Zhao F, Liu X, He W (2009) Challenges towards elastic power management in internet data centers. In: Proceedings of the 2nd international workshop on cyber-physical systems (WCPS 2009), in conjunction with ICDCS 2009, Montreal, QC, Canada, June 2009
9. Li B, Li J, Huai J, Wo T, Li Q, Zhong L (2009) EnaCloud: An energy-saving application live placement approach for cloud computing environments. In: IEEE international conference on cloud computing, Bangalore, India
10. Shang L, Peh L-S, Jha NK (2003) Dynamic voltage scaling with links for power optimization of interconnection networks. In: Proceedings of the 9th international symposium on high-performance computer architecture (HPCA '03). IEEE Computer Society, Washington, DC, USA, pp 91–102.
11. Mahadevan P, Sharma P, Banerjee S, Ranganathan P (2009) Energy aware network operations. In: Proceedings of the 28th IEEE international conference on Computer Communications Workshops (INFOCOM'09). IEEE Press, Piscataway, NJ, USA, pp 25–30.
12. The Network Simulator ns-2. Available at: <http://www.isi.edu/nsnam/ns/>, Accessed Aug 2012
13. Buyya R, Ranjan R, Calheiros RN (2009) Modeling and simulation of scalable cloud computing environments and the CloudSim toolkit: challenges and opportunities. In: Proceedings of the 7th high performance computing and simulation conference, Leipzig, Germany
14. Lim S-H, Sharma B, Nam G, Kim EK, Das CR (2009) MDCCSim: a multi-tier data center simulation, platform. In: IEEE international conference on cluster computing and workshops (CLUSTER). pp 1–9
15. Rawson A, Pflueger J, Cader T (2008) Green grid data center power efficiency metrics: PUE and DCIE. The Green Grid White Paper #6
16. Wang L, Khan SU (2011) Review of performance metrics for green data centers: a taxonomy study. *The Journal of Supercomputing*. Springer US, pp 1–18
17. Cisco Data Center Infrastructure 2.5 Design Guide (2010) Cisco press, March 2010
18. Thaler D, Hopps C (2000) Multipath issues in unicast and multicast nexthop selection. Internet Engineering Task Force. Request for Comments 2991, November 2000. Available at <http://tools.ietf.org/html/rfc2991>
19. IEEE Standard for Information technology-Telecommunications and information exchange between systems-Local and metropolitan area networks-Specific requirements Part 3: Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications Amendment 4: Media Access Control Parameters, Physical Layers and Management Parameters for 40 Gb/s and 100 Gb/s Operation,” IEEE Std 802.3ba-2010 (2010) (Amendment to IEEE Standard 802.3-2008), pp 1–457
20. Christesen S Data center containers. Available at <http://www.datacentermap.com/blog/datacenter-container-55.html>., Accessed Aug 2012
21. Katz RH (2009) Tech Titans building boom. *IEEE Spectrum* 46(2):40–54
22. Worthen B (2011) Data centers boom. Wall Street Journal. Available at <http://online.wsj.com/article/SB10001424052748704336504576259180354987332.html>
23. Next generation data center infrastructure. CGI White Paper, 2010
24. Guo C, Wu H, Tan K, Shiy L, Zhang Y, Luz S (2008) DCell: a scalable and fault-tolerant network structure for data centers. In: ACM SIGCOMM, Seattle, Washington, DC, USA

25. Guo C, Lu G, Li D, Wu H, Zhang X, Shi Y, Tian C, Zhang Y, Lu S (2009) BCube: a high performance, server-centric network architecture for modular data centers. In: ACM SIGCOMM, Barcelona, Spain, 2009
26. Li D, Guo C, Wu H, Tan K, Zhang Y, Lu S (2009) FiConn: using backup port for server interconnection in data centers. In: IEEE INFOCOM 2009, pp 2276–2285
27. Liao Y, Yin D, Gao L (2010) DPillar: scalable dual-port server interconnection for data center networks. In: 2010 Proceedings of 19th International Conference on computer communications and networks (ICCCN), pp 1–6
28. Kliazovich D, Bouvry P, Khan SU (2010) GreenCloud: a packet-level simulator of energy-aware cloud computing data centers. *The Journal of Supercomputing*, pp 1–21
29. Chen Y, Das A, Qin W, Sivasubramanian A, Wang Q, Gautam N (2005) Managing server energy and operational costs in hosting centers. In: Proceedings of the ACM SIGMETRICS international conference on Measurement and modeling of computer systems. ACM, New York, pp 303–314
30. Farrington N, Rubow E, Vahdat A (2009) Data center switch architecture in the age of merchant silicon. In Proceedings of the 17th IEEE symposium on high performance interconnects (HOTI '09). IEEE Computer Society, Washington, DC, USA, pp 93–102
31. Greenberg A, Lahiri P, Maltz DA, Patel P, Sengupta S (2008) Towards a next generation data center architecture: scalability and commoditization. In: Proceedings of the ACM workshop on programmable routers for extensible services of tomorrow, Seattle, WA, USA
32. Mahadevan P, Sharma P, Banerjee S, Ranganathan P (2009) A power benchmarking framework for network devices. In: Proceedings of the 8th international IFIP-TC 6 networking conference, Aachen, Germany 2009
33. Garrison S, Oliva V, Lee G, Hays R (2008) Data center bridging, Ethernet Alliance. Available at <http://www.ethernetalliance.org/wp-content/uploads/2011/10/Data-Center-Bridging1.pdf>
34. Alizadeh M, Atikoglu B, Kabbani A, Lakshmikantha A, Pan R, Prabhakar B, Seaman M (2008) Data center transport mechanisms: Congestion control theory and IEEE standardization. In: Annual Allerton conference on communication, control, and computing, pp 1270–1277.
35. IEEE 802.1 Data Center Bridging Task Group. Available at: <http://www.ieee802.org/1/pages/dcbridges.html>, Accessed Aug 2012
36. Song Y, Wang H, Li Y, Feng B, Sun Y (2009) Multi-tiered on-demand resource scheduling for VM-based data center. In: IEEE/ACM international symposium on cluster computing and the grid (CCGRID), pp 148–155
37. Kliazovich D, Bouvry P, Khan SU (2011) DENS: Data center energy-efficient network-aware scheduling. *Cluster Computing*, Springer US, pp 1–11.
38. Koppurapu C (2002) Load balancing servers, firewalls, and caches. Wiley, New York

Part III
Optical Interconnects Architectures

Chapter 5

The Role of Photonics in Future Datacenter Networks

Al Davis, Norman P. Jouppi, Moray McLaren, Naveen Muralimanohar, Robert S. Schreiber, Nathan Binkert, and Jung-Ho Ahn

5.1 Introduction

Over the past decade, the very nature of our computing and information infrastructure has gone through a dramatic change. Besides the normal near-exponential requirement for more of everything, there are some new driving factors. The reach and bandwidth of the Internet has rapidly expanded, and this increase has been amplified by the near ubiquitous reach of cellular telephone networks. As a result, the most common information endpoint for most users is a mobile device such as a smart phone, tablet, or laptop. By themselves these devices are useful. Now that they are connected to the Internet, they have spawned a wide range of new information-centric activities such as streaming video, social networking, satellite maps, and cloud computing. Even the word “Google” is no longer just the name of a company, but is commonly used as a verb associated with rapidly searching massive amounts of data and returning pointers to hopefully relevant results.

The change is in evidence at the corporate and consumer level as well. Shopping is no longer restricted to physical presence, but can be done virtually anywhere

A. Davis (✉) • N.P. Jouppi • N. Muralimanohar • R.S. Schreiber
HP Labs, Palo Alto, CA, USA
e-mail: ald@hp.com; norm.jouppi@hp.com; naveen.muralimanohar@hp.com;
rob.schreiber@hp.com

M. McLaren
HP Labs, Bristol, UK
e-mail: moray.mclaren@hp.com

N. Binkert
nou data, Palo Alto, CA, USA
e-mail: nate@binkert.org

J.H. Ahn
Seoul National University, Seoul, Korea
e-mail: gajh@snu.ac.kr

via the Internet. Point of sale transactions, whether they are via the Internet or a conventional store are tracked and analyzed to predict anything from what should be advertised to a particular consumer to information that guides corporations to invent new products and services. The amount of data available to large corporations is enormous, and the need to analyze that data for a business advantage involves an equally staggering mix of computational power, storage capacity, and communication.

The result is that the bulk of the processing and storage action has moved from the information endpoint to powerful and centralized warehouses comprising massive storage and computational resources—e.g., the datacenter. Given the economic advantages of large-scale installations, this centralization theme is just the beginning of what will likely continue. We use the term datacenter in this chapter loosely since datacenters vary significantly in terms of their size and the nature of their constituent components. At the high-end, high-performance computing (HPC) installations utilize the fastest and most powerful components. Low-end private enterprise datacenters are harder to characterize since they may employ high- or low-performance components or even a mixture of both. The middle ground is even harder to place in some sort of taxonomy since these datacenters are very cost sensitive and therefore use the best *bang for the buck* components while servicing a much larger user base than either the enterprise or HPC installations. In terms of their size, this *middle tier* may well equal or even exceed the size of supercomputer facilities, e.g., the warehouse scale computers (WSCs) employed by Google, Yahoo, Twitter, Facebook, etc.

Each datacenter class is optimized for various metrics. For supercomputers, it is all about performance and this involves not just only computational power but also network performance which includes minimizing latency and maximizing bandwidth. For commercial WSCs, maximizing raw performance is less important than high availability and throughput per dollar, since they service an enormous number of concurrent requests. It is noteworthy that these high-volume request servers have a strange computation vs. communication ratio. Each request may well involve a massive amount of data but the computational requirements may be small in comparison. One can safely conclude that for any particular large-scale WSC there is a need for a massive level of data communication, computation, or both. Other observations support this premise. Astfalk [6] points out that:

1. For every byte written or read to/from a disk, 10 KB is transmitted over the datacenter network.
2. For every byte transmitted over the Internet to/from the datacenter, 1 GB of data is moved through the datacenter network.
3. The server growth rate is 7% per annum. Note that in the US Environmental report to Congress [1], the rate was 17% but they did not account for the high levels of virtualization and the attendant server consolidation that is in play today.
4. Storage requirement growth is 52% per year. In 2007, 5 exabytes of additional datacenter storage was required which is 10,000 times the size of the entire printed data contained in the Library of Congress.

5. Internet traffic was 6.5 exabytes per month in 2007 and the annual growth rate is expected to be 56%.
6. The number of Internet nodes in 2007 indicated a sustained annual growth rate in the number of Internet nodes to be 27%. Note that given the current popularity of cell phones and tablets, this number is likely very conservative.

The goal is to enable the design of more energy-efficient future datacenters and there is a clear conclusion that can be drawn from this data. Namely that it is more important to improve the bandwidth and energy-efficiency of the network infrastructure than it is to improve the energy and performance of the processor socket. In part, this is because the latter will happen anyway due to the merchant semiconductor industry's efforts to improve semiconductor processes and improve socket architectures. The former is a bigger challenge since as integrated circuit feature sizes become smaller, transistors scale nicely in terms of area, switching speed, and energy consumption. Unfortunately both on-chip wires and off-socket I/Os do not scale nearly as well.

WSC-scale datacenters are also very expensive and the total capital expense (CapEx) of the building, cooling, power infrastructure, networking, storage, and compute servers varies between \$150M to over \$1B today. The amortization schedules vary. The building, cooling, and power distribution costs are typically amortized over a 10-year period, networking equipment over 4 years, and servers, memory, and storage are considered to have a 3-year lifetime. Operational expense (OpEx) includes personnel costs as well as energy costs and the 3–4 year OpEx expense often equals the initial amortized CapEx expense.

Hoelzle and Barroso [15] report that the total power for a 2007 Google WSC can be broken down into the following components: 33% for servers, 25% for power distribution and cooling, 15% for electrical utility losses, and 15% for the network. They also argue that while networking is not the largest power component, networking and systems innovation is the key to getting the most out of the WSC investment. We argue that for WSC-scale datacenters, the network is in fact the critical component and the next question to answer is what should be done to improve these networks for future datacenters.

Fundamentally there are two technologies used today to transport data: (1) electrical and (2) optical. The telecommunications industry has long recognized the advantages of optical communication in terms of modulation rate, latency, bandwidth, and bit-transport energy (BTE) for long-haul communication. We note that the definition of *long* should be modulation rate dependent. Electrical communication faces a number of serious problems as modulation rates exceed a few GHz and transport lengths exceed a few mm. Namely:

1. The amount of power required to drive an unrepeat wire is fundamentally linear with the length of the wire and delay is quadratic with length [14].
2. Wire delay can be reduced to nearly linear with appropriate repeater spacing but at the cost of increased power.
3. Signal integrity for high-speed signals is a serious problem for off-socket communication.

4. Both the number of signal I/Os on a package and the per-pin modulation rate are increasing much more slowly than the number of transistors that can be placed on the die [5]. This creates a serious edge bandwidth communication constraint.
5. The use of high-speed SerDes I/Os is one way to increase the pin bandwidth, but these circuits are very power intensive and their power consumption does not scale well with process improvements.

Optical/photonic signaling has rather different properties:

1. Due to the low loss rates of the waveguides or fibers, the active power consumed by photonics is dominated by the end-points where electrical to optical (EO), or optical to electronic (OE) domain conversion happens. Hence optical active power consumption is essentially independent of path length for the distances of interest in the datacenter. There is an idle-power problem, however, that will be discussed subsequently.
2. Bandwidth is not strictly dependent upon the modulation rate since multiple wavelengths of light can be used on the same waveguide/fiber, an option that is not practical for high-speed electronic signaling.
3. The signal integrity problem is significantly reduced in the optical domain.

The obvious conclusion is that we should abandon electrical signaling in favor of photonics. Photonics should win in most of the critical datacenter metrics, namely bandwidth and power. However, this conclusion would be premature; cost cannot be ignored and the designers of better electronic systems are not standing still. The cost of all-electrical communication structures is well known since this technology has been in play for a long time. Photonic device structures are relatively new and we don't have a similar ability to reliably predict the future, primarily because volume manufacturing of these devices on modern fab lines has yet to happen. Cost for integrated devices is heavily dependent on both volume and manufacturing efficiency. All of the requisite devices necessary to support photonic communication have been fabricated and tested in lab environments, but it is safe to say that the gap between lab device demonstration and cost-effective, high-yield manufacturing is a big step.

Modern datacenters have already embraced photonic signaling for long distances. Electronic signaling is still the norm inside the rack but today's datacenter cables are rapidly transitioning from electrical to photonic. These *Active Optical Cables* (AOCs) have the OE and EO engines embedded in the connector. AOCs are significantly lighter, have a tighter bend radius, and are more energy efficient than the electrical alternative. They are, however, more expensive. The primary reason that they are increasingly becoming the *right choice* is that the higher CapEx expense is offset by the reduced energy/OpEx cost. The scaling argument also favors this trend since components become cheaper and energy will be more expensive in the future.

One way to look at what might be referred to as the *optical invasion* is to note that the long haul telecom and Internet backbone is already photonic. The next longer distance is the cable lengths inside the datacenter and they are rapidly

becoming photonic with the use of AOCs. The open question is when and with what technology will the next longest paths convert to photonics. The next longest paths in order of decreasing length from an interconnect perspective are backplanes, server or router board traces, and finally the individual switch or processor chips. In May, 2011 Hewlett Packard demonstrated an all-optical passive backplane in a router cabinet [34]. The same technology can be easily applied to other backplane applications.

The reader will note that so far, little has been said about latency. It is the other key datacenter metric since everybody cares about how fast they can get an answer to a search or large database query. For cases where computational needs are light, the key performance component is often network latency. It is a common misconception that photons in a waveguide travel faster than electrons in a good copper trace or cable. This is just not true in all cases. The fact is that optical signal propagation is the speed of light in a vacuum (c) divided by the group refractive index of the waveguide. Currently on-chip waveguides and electrical signals both achieve approximately $1/3c$. For intra-datacenter cables it is a different story: optical transmission achieves about $2/3c$ and electrical signaling achieves about $1/3c$. For circuit boards using hollow metal waveguides the optical option wins again since the group refractive index is very close to 1.0. One can safely conclude that for on-chip signals, it is not about latency but the winning technology will be based on other factors such as energy consumption. Once the signal is off-chip, photonic latency simply wins but delay through the OE and EO engines and cost must be considered.

Total packet latency in a datacenter can be defined as the time it takes for a packet to traverse the network from the sender to the receiver node (a.k.a. *terminal*). Total packet latency is the sum of all of the path latencies and all of the switch latencies encountered along the route. A packet that travels over N paths will pass through $N - 1$ switches. The value of N for any given packet will vary depending on the amount of locality that can be exploited in an application's communication pattern, the topology of the network, the routing algorithm, and the size of the network. However, when it comes to typical case latency in a WSC-scale datacenter network, path latency is a very small part of total latency. Total latency is dominated by the switch latency which includes delays due to buffering, routing algorithm complexity, arbitration, flow control, switch traversal, and the load congestion for a particular switch egress port. Note that these delays are incurred at every switch in the network and hence these delays are multiplied by the hop count.

The best way to reduce hop-count is to increase the radix of the switches. Increased switch radix means less switches for a network of a given size and therefore a reduced CapEx cost. Reduced hop-count and fewer switches also lead to reduced power consumption as well as reduced latency. For all-electrical switches, there is a fundamental trade-off due to the poor scaling of both signal pins and per-pin bandwidth. Namely one could choose to utilize more pins per port which results in a lower radix but with higher bandwidth per port. The other option is to use fewer pins per port which would increase the radix but the bandwidth of each port would suffer. Photonics may lead to a better option, namely the bandwidth advantage due

to wave division multiplexing, and the tighter signal packaging density of optics means that high-radix switches are feasible without a corresponding degradation of port bandwidth.

The remainder of this chapter is devoted to exploring where photonics will be most useful in the creation of high-radix switches for future datacenters. Given the the cost uncertainty, the primary focus will be on the very-large scale and high-performance datacenter target where higher priced components are more palatable.

5.2 Background

High-end system performance is expected to grow by three orders of magnitude, from petascale to exascale, by 2020. The Moore's law scaling of semiconductor technology will not, by itself, meet this need; to close the gap, there will be more processing and storage components. A recent study [23] shows that an exascale system will likely have 100,000 computational nodes. The increasing scale and performance will put tremendous pressure on the network, which is rapidly becoming both a power and a performance bottleneck [24]. High-radix network switches [17] are attractive since increasing the radix reduces the number of switches required for a given system size and the number of hops a packet must travel from source to destination. High-radix switches can be connected hierarchically (in topologies such as folded Clos networks [18]), directly (in a flattened butterfly or HyperX topology [2, 19]), or in a hybrid manner [20].

The chip I/O bandwidth and chip power budget are the two key limits to boosting radix. Our goal is to assess whether electronics or photonics will be better suited to overcome these limits in future switches. In order to make this assessment, we need guideposts. For electronics, we use the ITRS [30]. Since photonics has no published roadmap, we develop one as described in Sect. 5.3 and use it in a performance and power comparison between electronics and photonics.

In electronic switches, increasing radix to reduce latency while maintaining per-port bandwidth will be hard because of chip-edge bandwidth: the ITRS predicts only modest growth in per-pin bandwidth and pin count over the next decade. For example, Cray's YARC is a high-radix, high-performance, single-chip switch [29], with 768 pins shared by 64 bi-directional ports, giving an aggregate bandwidth of 2.4 Tb/s. Each port has three input and three output data signals, but the use of differential signaling, necessary to improve high-speed signaling reliability, means that 12 pins are required in total. High-speed SERDES can help by increasing the signaling rate, but this reduces the power budget available for the actual switching function. In YARC, high-speed differential SERDES consume approximately half the chip power (Parker (2010) Personal communication).

Emerging silicon nanophotonics technology [21, 22, 25, 35, 36] may solve the pin bandwidth problem. Waveguides or fibers can be coupled directly onto on-chip waveguides, eliminating electrical data pins. While the signaling rate is comparable to that of electrical pins, high-bandwidth per waveguide can be achieved with dense

wavelength division multiplexing (DWDM), in which up to 64 wavelengths of light constitute independent communication channels. Because of DWDM, a high-radix photonic switch will have fewer off-chip fiber connections than pins in a comparable electronic switch. Furthermore, over a long path, an inter-switch cable or a circuit board trace, the energy cost to send a bit of information is lower in optics than in electronics. At datacenter scale, the bit transport energy (BTE) of photonic communication is nearly independent of path length. Electrical BTE grows linearly for unrepeated wires and worse than linearly if repeaters are used to improve latency and signal integrity.

The next scaling limit will be power in the on-switch-chip electrical interconnect. Again, an all-electrical solution will not work. But unlike the I/O limit, the right answer is not an all-photonic solution; it is a reasonable hybrid of long-distance photonics with short-distance electronics.

On-chip global wires are increasingly slow and power hungry [14]. Global wire geometry is not scaling at the same rate as transistor geometry. To minimize fall-through latency, YARC uses repeated wires in global data and control paths. Many intermediate buffers and wires are required to support YARC's over-provisioned intra-switch bandwidth.

Photonic BTE is low, and is length independent on-chip as well as off-chip. But there are other issues. Optical modulators and receivers require constant tuning even when not being used (more in Sect. 5.3.2) resulting in static power not present in plain electrical wires. This static tuning power requirement implies that photonics are most energy-efficient if the optical links are heavily utilized. Electrical signaling over small distances can have lower BTE and be faster than optical signaling, partly due to endpoint EO/OE domain conversion power. The distance at which optics becomes preferable will change with shrinking feature size, because electrical wires and optics scale differently. Thus a short-range electronic, long-range optical design seems reasonable, although the cross-over point for short vs. long will certainly depend on technology improvements that are difficult to determine precisely at the moment.

5.3 Electronic and Photonic Roadmaps

High-performance switches are not manufactured in the same volume as processors; they are relegated to older fabs. YARC, a standard cell ASIC, was fabricated in a 90 nm fab, and custom microprocessors were then fabricated in a 65 nm process [29]. Microprocessors are now fabricated in 32 nm CMOS technologies; ASICs remain at least a generation behind. We therefore focus on the 45, 32, and 22 nm CMOS technology steps.

We describe electrical and photonic I/O roadmaps. These help define the design space for high-radix switches. The electrical I/O roadmap is based on the 2009 ITRS [30]. It provides the roadmap for most switch components, but does not predict I/O power. We supplement it with SERDES power predictions based on recently

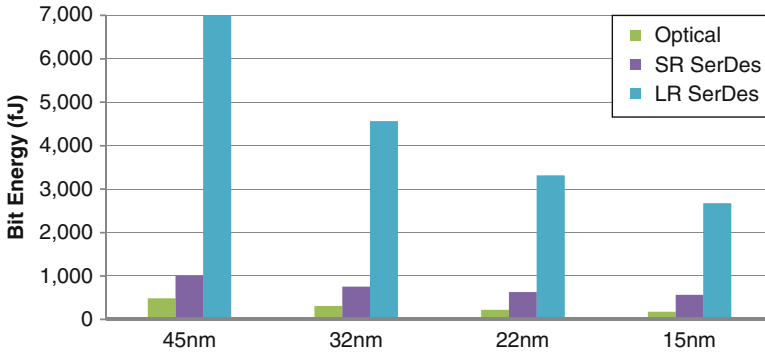


Fig. 5.1 I/O energy per bit scaling

published results. Although the impact of technologies such as photonics is being considered by the ITRS, there is no industry roadmap at the present time. We make a first attempt to create a photonic roadmap, based on recent literature as well as our own laboratory efforts.

The benefits of photonics are compelling, but technology challenges remain before it can be deployed. Laboratory device demonstrations have been performed; waveguides, modulators, and detectors have been built and tested [9], but the ability to cheaply and reliably manufacture hundreds to millions of these devices on the same substrate has not yet been demonstrated.

5.3.1 Electrical I/O Roadmap

The ITRS is concerned primarily with the “short reach” or SR-SERDES, with trace lengths of a few centimeters, used for processor to main memory interconnects. Recently a number of low-power SR-SERDES have been demonstrated [12, 28]. In switch applications, “long reach” or LR-SERDES are generally required to drive a path of up to one meter of PC board trace with at least two backplane connectors in the path. SR-SERDES use less power than LR-SERDES, but they require some form of external transceiver or buffer to drive longer transmission paths. Although switch-chip power in this arrangement decreases, the overall system power grows.

Historical data show that SERDES power scales by roughly 20% per year [28]. Not all components of SERDES power will continue to scale at this rate. The external loads (impedances of off-chip wires) are not changing, and the output drive power cannot be expected to improve. Our power model for SR-SERDES and LR-SERDES takes the current state-of-the-art BTE value as its starting point. We assume that the power of the transmitter output stage remains constant, and the balance of the energy will scale according to the ITRS roadmap. The predicted BTE values for both types are shown Fig. 5.1.

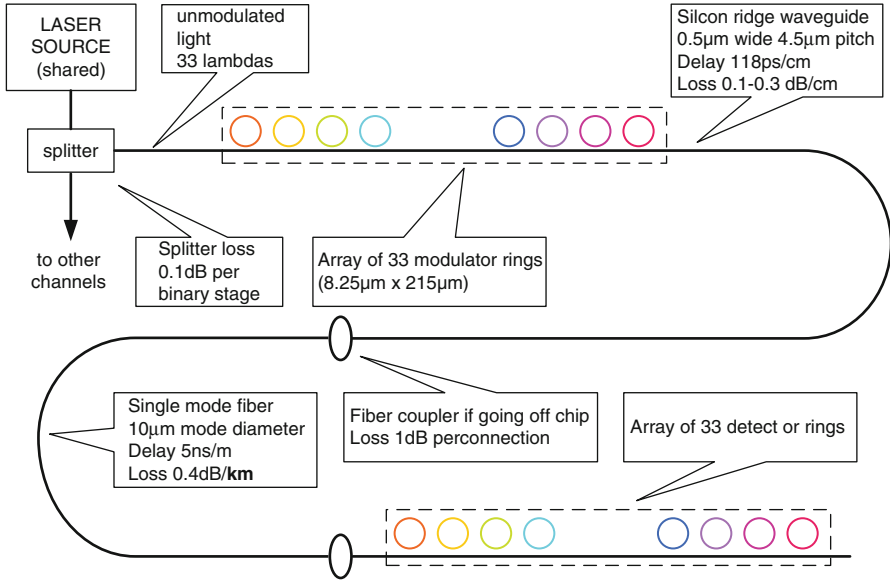


Fig. 5.2 Interchip point-to-point DWDM link

5.3.2 Photonic Roadmap

External transceivers cannot overcome the chip-edge bandwidth wall. An integrated technology can, by bringing light directly onto the chip. Integrated CMOS photonics, where all the components for communication with the exception of an external laser power source are integrated in a CMOS compatible process, have been demonstrated using indirect modulation [4]. However, the Mach-Zehnder modulators used in these systems are impractical for systems requiring many optical channels due to their large area and relatively high BTE.

Compact, power-efficient modulators based on resonant structures have been demonstrated [9]. Our proposed technology uses silicon ring resonators, similar to the devices described by [3]. A ring can be used as a modulator, as a wavelength-specific switch, or as a drop filter. Rings have the additional advantage of being wavelength specific, allowing DWDM (dense wavelength division multiplexed) transmitters to be created. Rings, together with silicon ridge waveguides for on chip connectivity, waveguide-integrated germanium detectors, and grating couplers for external connectivity, constitute a complete set of components required for communications. All components can be manufactured on a common silicon substrate with the optical source being provided by an off-chip laser.

Figure 5.2 depicts a complete DWDM photonic link. An external mode-locked laser provides light as a “comb” of equally spaced wavelengths. An array of ring resonators in one-to-one correspondence with the wavelength comb modulates a

Fig. 5.3 Optical losses for 2 cm of waveguide and 10 m of fiber

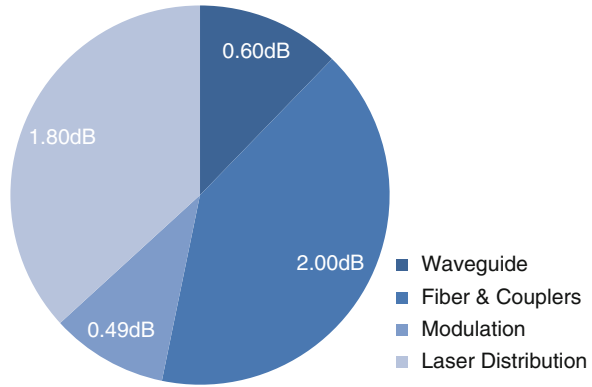
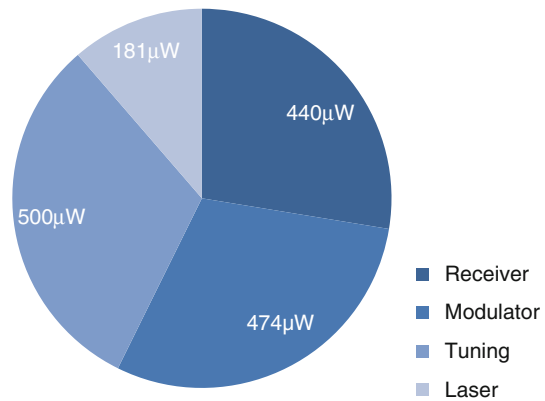


Fig. 5.4 Point-to-point power, 22 nm process node



signal on the passing light. That light is transmitted through a waveguide, into fiber via a coupler, and back into another waveguide on a different chip, and into another array of ring resonators for detection. This link can be used for both inter-chip communication via the single mode fiber or for intra-chip communication if that fiber and the related couplers are removed.

Power and losses for a complete inter-chip DWDM photonic link consisting of 2 cm of waveguide and 10 m of fiber are illustrated respectively in Figs. 5.3 and 5.4. We calculate the laser electrical power from the required receiver optical input power, the total path loss including optical power distribution, and the laser efficiency. Receiver electronic power was simulated using HSPICE to model the transimpedance amplifier and limiting amplifiers. Modulation power was estimated from the measured circuit parameters of ring resonators assuming a modulation rate of 10 Gb/s at each process step. The final component is the thermal tuning power. Since all the power terms except modulation are independent of link activity, link power is not strictly proportional to usage. High-speed differential electronic links exhibit a similar lack of proportionality. Since they must be constantly active sending idle frames when real data is not being transmitted.

5.4 Switch Microarchitecture

A scalable switch microarchitecture is used in a design space exploration to compare the photonic and electrical alternatives for a high-radix switch-chip. For electrical systems, this is accomplished by increased chip pin-count and/or improved SERDES speeds. For photonic interconnects this is enabled by the availability of more wavelengths for the optical links. Constrained by the limits of the electrical and photonic roadmaps, we investigate switches of radix 64, 100, and 144, of each in three process generations. The decision to use an N^2 square number of ports was motivated by the desire to maintain an $N \times N$ array of subswitches in the all-electronic switch case. We view feasible designs as falling within ITRS package limitations, consuming less than 140 W, and fitting within an 18x18 mm die. Higher power switches are possible, but would require significantly more expensive liquid conductive cooling. We view designs between 140 and 150 W as cautionary and designs greater than 150 W as infeasible. The die size is based on a floorplan that accounts for port interconnect pitch, input and output buffer capacity, photonic element pitch, port tile logic, and optical arbitration waveguides or electrical arbitration logic.

Datacenter switches typically conform to Ethernet style packet sizes, and vary in length from jumbo packets, commonly 9,000 or more bytes, to the smallest 64 B size. For simulation purposes, we vary the packet size in multiples of 64 B, where the multiplier varies between 1 and 144. In both electronic and photonic designs, we provide buffers at both the input and output ports. Input buffers are 32, 64, and 128 KB, respectively, for the 45, 32, and 22 nm feature sizes. This $2\times$ scaling tracks the $2\times$ scaling projection of additional wavelengths. The output buffer is sized at 10 KB to accommodate an entire jumbo packet. The output buffer can also be increased in size to support link-level retry, but we are not modeling failure rates and link-level retry in this work since this chapter's focus is concerned with the design options for a single switch.

For optical I/Os, we allow one input fiber and one output fiber per port, and hence the per-port bandwidths over the three process generations are 80 Gbps, 160 Gbps, and 320 Gbps. Flow control is done on a per-packet basis. The worst case inter-switch link in our model is 10 m, and flow control must account for the round trip latency on the link plus the response time on either end. Table 5.1 shows the worst case number of bits that could be in flight, and the buffers are sized accordingly. Our simulations and power estimation models focus on datapath and arbitration resources. The remaining details of the various tile resources are shown in Table 5.1. We assume a 5 GHz electrical component clock based on ITRS [30] and drive the optical links in DDR fashion at 10 Gbps.

Table 5.1 Radix independent resource parameters

General	Process	nm	45	32	22	
	System clock	GHz	5			
Link characteristics	Port bandwidth	Gbps	80	160	320	
	Max link length	m	10			
	In flight data	Bytes	1,107	2,214	4,428	
Optical link parameters	Data wavelengths		8	16	32	
	Optical data rate	Gbps	10			
Electronic link parameters	SERDES per channel bandwidth	Gbps	10	20	32	
	SERDES channels per port		8	8	10	
	Bit energy (LR_SERDES)	fJ/bit	7,000	4,560	3,311	
	SERDES TDP/port	mW	560	730	1,060	
	Electronic I/O pins/port		32	32	40	
	Buffers	Input buffer size	kB	32	64	64
		Header queue entries		64	128	256
Input buffer read width		bits	32	64	128	
Input buffer write width		bits	16	32	64	
Flit size		Bytes	64			
Packet size		Flits	1–144			
	Output buffer size	Bytes	9,216			

5.4.1 Electronic Switch Architecture

A simple switch consists of three primary components: input buffers to store incoming messages; a crossbar to transmit the messages to the appropriate output port; and an arbiter to allocate resources and resolve conflicts. Since the latency of all three components increases with radix and size, scaling them directly to a high radix will either reduce the operating frequency or the switch throughput. Where a simple FIFO structure is used for the input buffers, a packet at the head of the buffer waiting for a busy output port will block subsequent packets from progressing even if their destination is free. This phenomenon, called head-of-line (HOL) blocking, limits the throughput of a simple crossbar switch to around 60% under uniform random traffic [16]. To address the latency problem, YARC splits crossbar traversal into three stages; 1-to-8 broadcasting (or demultiplexing), 8×8 subswitch traversal stage, and 8-to-1 multiplexing. Buffers are inserted between stages to alleviate HOL blocking by buffering packets according to destination. A fully buffered crossbar with a dedicated buffer at every crosspoint can handle loads close to 100% of capacity. This significantly increases buffering, which grows as the square of port count. The YARC architecture reduces the buffer size requirements by partitioning the crossbar into multiple subswitches.

Figure 5.5 shows the organization of a distributed high-radix switch similar to YARC. The switch uses a single repeated tile with one instance for each bidirectional port. The tiles are organized as an M row by N column array; hence, there are $M \times N$ ports. Each tile consists of an input buffer, an N input to M

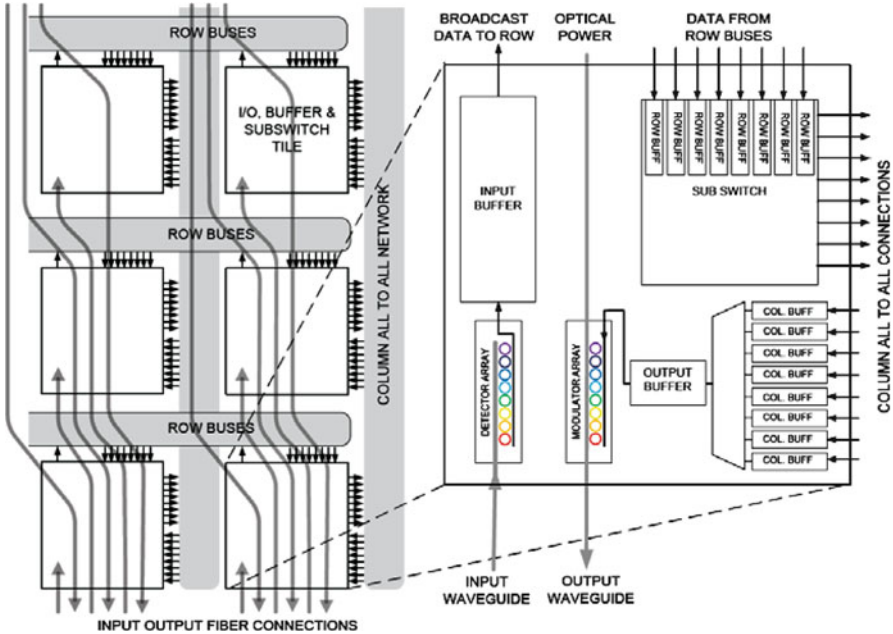


Fig. 5.5 Array of electronic switch tiles and waveguides. Photonic I/O is incorporated into the tile

output subswitch, an M input multiplexer, and an output buffer. Every subswitch has buffers at its inputs called row buffers. Every multiplexer has buffers at its inputs called column buffers. The size of these intermediate buffers is critical to avoiding HOL blocking. Packets flow from the input SERDES to the input buffer and are then sent (via a broadcast message) along the row bus to the tile that is in the same column as the output port. Note that on average, the N input buffers along a row will send one phit per cycle to each subswitch. Hence, the average load in a subswitch is only $100/N\%$. Once a phit reaches a subswitch, the first stage arbitration maps it to the tile of the correct output port. Within each column, the subswitches and output multiplexors are fully (all-to-all) connected. A second stage arbitration picks packets from the column buffers and sends them to the output buffer. This arrangement means that arbitration is local to a tile and is limited to N inputs for the first stage and M inputs for the second stage. For electronic switch datapaths, we scale the input port bandwidth based on the roadmap we discussed in Sect. 5.3. The size of the subswitches, column, and row resources scale as the square root of the port count. For optical I/O, the output modulators and output detectors are assumed to be integrated with the tile in order to eliminate long wires and use the optical waveguides as an additional low-loss routing layer. For electronic I/O, the high-speed SERDES are placed around the periphery of the chip to provide a more controlled analog environment.

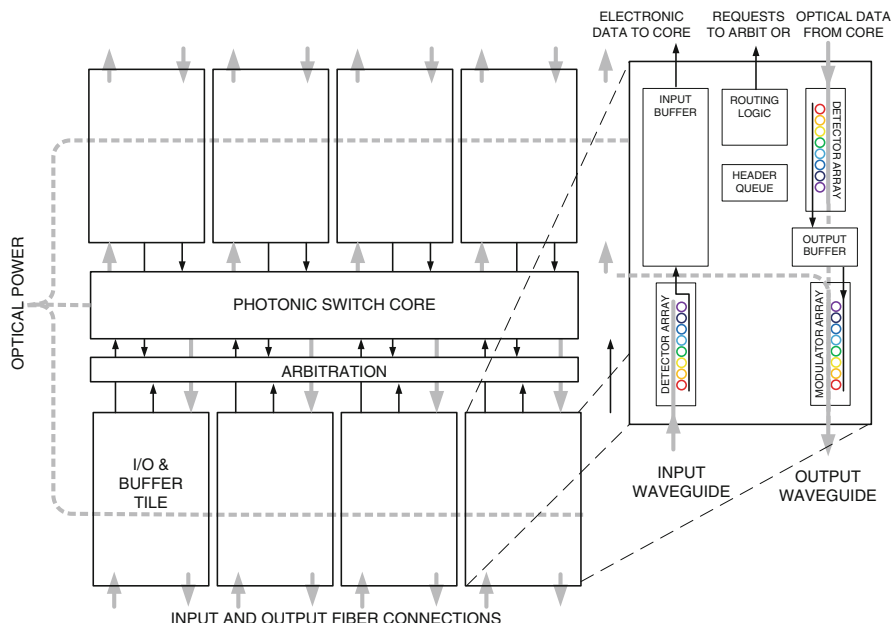


Fig. 5.6 Tile placement for an optical switch core. A switch core with a high aspect ratio is used to exploit the low-loss of the photonic interconnect

5.4.2 Optical Switch Architecture

In the optical switch architecture, we return to a simple single-level switch using an optical crossbar. This choice is motivated by the high static power of optical interconnects. YARC over-provisions wires to interconnect subswitches; they are underutilized. This is not a power-efficient way of using an optical interconnect due to the static tuning power requirement.

We exploit the low propagation loss of optical waveguides to build an optical crossbar that spans the chip more power efficiently than an electronic crossbar. HOL blocking is addressed by using a flexible input buffer structure, and an arbitration algorithm that considers multiple requests from each input. The optical switch architecture is shown in Fig. 5.6, with multiple I/O tiles surrounding a high-aspect-ratio optical crossbar. The I/O tile consists of a unified input buffer, output buffer, input header queue, and request generation logic.

Packets arriving on the input fiber are immediately converted into the electronic domain and stored in the input buffer. A separate header FIFO contains the routing information for every packet in the input buffer. The first eight elements of the header FIFO are visible to the request generation logic, which generates up to eight requests to the central arbiter. When a grant is received for one of the requests, the input buffer sends the relevant packet to the switch core and frees the buffer space.

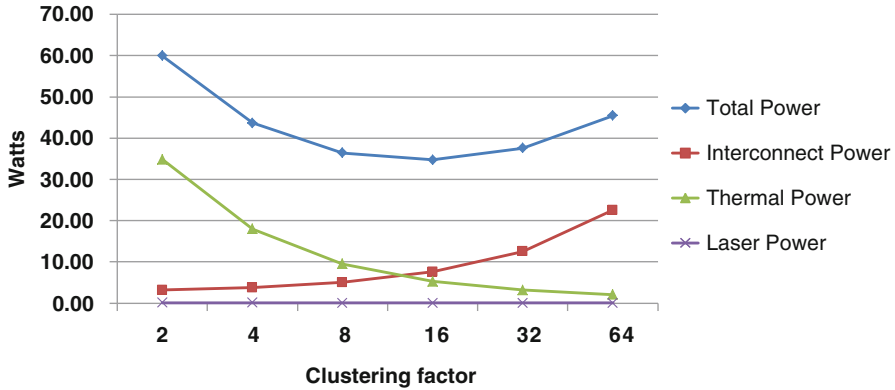


Fig. 5.7 12 input, 4 output crossbar with 4 way clustering

The input buffer has sufficient bandwidth to transfer two packets to the crossbar at a time. Since the input buffer is not a FIFO, buffer space management is more complex. The crossbar operates at double the external link bandwidth, which allows the input port to “catch up” when output port contention occurs. Since the crossbar bandwidth is twice the external port bandwidth, output ports require sufficient buffering to accommodate at least one maximum-sized packet.

5.4.3 Optical Crossbar

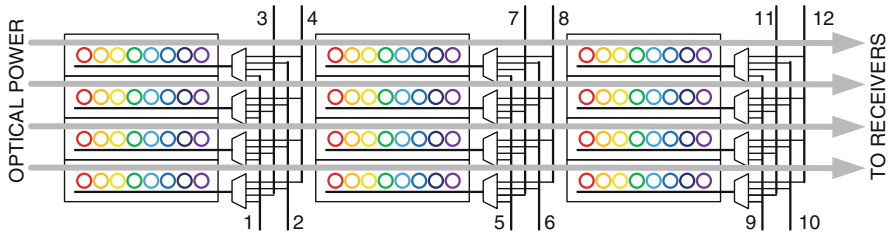
A crossbar is a two-dimensional structure that broadcasts in one dimension and arbitrates in the other. In our optical crossbar, a waveguide is associated with each output port. Input port requests are granted by the arbitration structure so that at any given time only one bank of modulators will be actively driving any given waveguide. In this *channel per destination* approach [33], each receiver ring must always actively listen to its associated waveguide.

The tuning power for this approach scales linearly with the number of inputs, as inactive modulator arrays must be kept at a known off-frequency position to avoid interference. Multiple crossbar inputs may share a set of modulators, without impacting crossbar performance, since only one set of modulators is ever active at a time. We refer to this as *clustering*, and use this technique to minimize the number of ring resonators per waveguide.

The optical crossbar in Fig. 5.7 shows the optical modulators shared by two pairs of inputs, one pair on each side of the optical switch, for a clustering factor of 4. Each waveguide of the 12-port switch therefore requires only three sets of modulators. The clustering factor can be adjusted to share the modulators between any number of adjacent tiles without impacting the throughput of the switch, but higher clustering factors require additional intra-cluster electrical interconnect.

Table 5.2 Components of optical loss

Component name	dB
Waveguide single mode (per cm)	1
Waveguide multi mode (per cm)	0.1
Adjacent ring insertion loss	0.017
Ring scattering loss	0.001
Off-chip coupling loss	1
Nonideal beam-splitter loss [13]	0.1

**Fig. 5.8** Varying clustering factor, radix 64 switch in 22 nm technology

The large number of rings per waveguide in the photonic crossbar means that ring related losses are more significant than for point-to-point links. Every ring induces some scattering loss, and idle, off-resonance modulator rings add loss due to adjacent partial coupling. Clustering reduces both of these loss factors. The components of loss are listed in Table 5.2. For the largest switch configuration studied the worst case path loss is 7.7dB.

Figure 5.8 shows the power savings that can be obtained by sharing the optical modulators. Initially, power drops due to the static power saved by reduced ring count. Beyond the minimum (cluster factor = 16), power grows due to the long wires in the cluster.

5.4.4 Thermal Tuning of Rings

A ring is resonant with a wavelength when its circumference is an integer multiple of this wavelength. Manufacturing variability and thermal expansion of the silicon make it necessary to add per-ring, active temperature control to align one of the resonant frequencies of the ring with one of the wavelengths of the laser-generated comb. Watts et al. demonstrated this using Joule heating elements embedded in or near the rings [35].

Complete tuning flexibility for a single ring would require sufficient heating power to move the ring across a wide wavelength range. A more efficient design can minimize the thermal tuning power. One idea is to use an extended array of equally spaced rings (see the top of Fig. 5.9). Tuning only needs to put the ring on

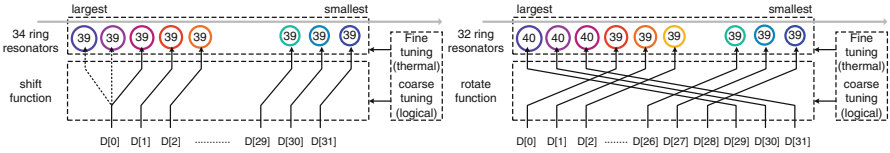


Fig. 5.9 Coarse tuning methods to minimize heating power: (*left*) additional rings; and (*right*) using a higher mode

the closest wavelength. By adding rings to extend the array, combined with a shift function between the rings and the electronic signals, the heating power required to tune between adjacent frequencies can be dramatically reduced.

A ring has multiple modes of resonance and is said to be resonant in mode M when the effective ring path length is M times the wavelength. To avoid the added power and area costs of additional rings, with a similar reduction in the maximum required heating power, we can design the geometry of the ring array such that the resonant frequency of the $M + 1$ th mode of the largest ring is one wavelength comb “tooth” to the low wavelength side of the shortest comb wavelength (right side Fig. 5.9). The number inside the colored ring represents the resonant mode of the ring; thus $D[0]$ is always connected to the longest (reddest) wavelength, and $D[31]$ to the shortest. The use of two modes in all rings gives a logical tuning range that is almost equivalent to the ring’s full free spectral range, which is the frequency range between two adjacent resonant modes.

Our photonic scaling assumptions are as follows. The geometry of the rings does not scale with process improvements since ring size and resonant frequency are coupled. We assume that the modulation frequency will remain constant across generations, a consequence of the use of charge injection as the mechanism for modulation. Modulation speed in this case is limited by the carrier recombination time of the rings. A relatively low modulation rate has the advantage that simple source synchronous clocking can be used. This requires an additional clock wavelength but allows simple, low power receiver clocking when compared to high-speed SERDES. We use a single added wavelength for the forwarded clock, along with groups of 8, 16, and 32 data wavelengths, respectively, for the three process steps that are considered.

5.4.5 Arbitration

Our photonic crossbar design requires a high-speed, low-power arbiter. To better utilize the internal switch bandwidth, we performed a novel design space study using uniform random traffic to quantify the benefit that would result from increasing the number of requests and grants available for each input port. We found that

allowing 8 requests and 2 grants per port improved internal bandwidth utilization by approximately 30% on average for all radices and packet sizes. This choice allows an input port to concurrently send two packets to different output ports.

We employ two forms of electrical arbitration. The electrical arbiter (YARB) for the electrical baseline datapath is an exact replica of the distributed YARC arbitration scheme. Since our goal is to evaluate the best arbitration choice for the photonic datapath, the electrical arbiter (EARB) implementation for a photonic datapath departs from the YARC model in order to more closely mimic the optical arbitration scheme. We employ the parallel-prefix tree arbitration design of [11]. This approach is similar to parallel-prefix adder design, where the trick is to realize that carry propagate and kill are similar to a prioritized grant propagate and kill. The EARB contains k -tiles for each radix k configuration. Each tile is logically prioritized in a mod- k ring, where the highest priority grantee for the next selection is just after the current grantee in ring order. This provides a fairness guarantee similar to round-robin scheduling.

The EARB is centralized and pipelined, but there is little doubt that additional improvements to our current version can be found. In particular area, speed, and power improvements are likely possible with more rigorous attention to critical path timing and transistor sizing issues. Layout can be improved to reduce wire delays. Finally our current scheme uses one prefix-tree arbiter for each output port and each arbiter returns a single grant to the winning requester. Hence it is possible for an input port to receive more than two grants. When this happens, logic at the input port will select which grants are to be rejected by dropping the associated request lines. The result is that the eventual grantee will wait longer than necessary due to the extra round trip delays between the input port and arbiter.

Sending a minimum-sized packet takes eight clocks. The most important aspect of any arbitration scheme is to have a round trip delay that is less than the packet transmission time. Our EARB design is optimized for delay, although we note that the dominant delay is due to the long electrical request and grant wires. Our EARB tile takes less than one 200 ps cycle for all process steps and radices. The worst case EARB request to grant time is seven clocks. The EARB power has a negligible impact on total switch power and in the worst case (radix 144, 45 nm) the arbiter requires 52 pJ/operation. For 22 nm the 144 radix power is 25.7 pJ/op.

Optical arbitration uses a separate set of arbitration waveguides where a particular wavelength on an arbitration waveguide is associated with a particular egress port in the switch. We employ the *token channel* arbitration scheme proposed by [33]. The optical arbitration round trip time is also less than eight clocks and the arbitration power has a negligible impact on total switch power. We conclude that there is no substantial difference between EARB and optical token channel arbitration and that either will be suitable through the 22 nm process step. Since the dominant delay component of EARB is the long request and grant wires, which grow with each new process step, we believe that in the long run optical arbitration may prove to be the winner.

Table 5.3 I/O and package constraints

	Ports	64	100	144
All optical generations	Max die size (mm)	18.1		
	Fibers per side (250 μm)	72		
	Fibers per side (125 μm)	144		
	Fibers required	128	200	288
	Fiber sides (250 μm)	2	3	4
	Fiber sides (125 μm)	1	2	2
	Port bandwidth	80 Gbps		
45 nm	SERDES rate	10 Gbps		
	Available SERDES pairs	600		
	Pairs required	512	800	1,152
	Port bandwidth	160 Gbps		
32 nm	SERDES rate	20 Gbps		
	Available SERDES pairs	625		
	Pairs required	512	800	1,152
	Port bandwidth	320 Gbps		
22 nm	SERDES rate	32 Gbps		
	Available SERDES pairs	750		
	Pairs required	640	1,000	1,440

5.4.6 Packaging Constraints

We evaluated the feasibility of all the switch variants against the constraints of the ITRS roadmap for packaging and interconnect.

Table 5.3 shows the electrical and photonic I/O resources that will be required for our choice of I/O models in all three process generations. The key conclusion is that the only feasible design for an all-electrical system capable of port bandwidths of 80 Gbps is radix 64. However even with today's 250 micron fiber packaging pitch, all of the optical I/O designs are feasible using fibers on four sides of the device. Using 125 micron pitch fiber packaging all the optical connectivity can be achieved on two sides. Even given the optimistic ITRS provisioning of high-speed differential pairs, there just aren't enough to support 100 and 144 port electronic designs with the requisite port bandwidth due to packaging limitations. From a packaging perspective, the trend is clear; increasing the switch radix over the radix-64 YARC while significantly increasing bandwidth requires optical I/Os. Since power and performance are equally critical in determining feasibility, they will be discussed next.

5.5 Experimental Setup

We estimate performance using the M5 simulator [7]. New modules were created for each of our design points. Module interactions are modeled at packet granularity. The optical model accounts for the propagation delay of light in waveguides in order to accurately quantify communication and arbitration delay.

We use CACTI 6.5 [27] to model the electronic switch and the electronic components of the photonic switch. The photonic model includes an analytic model of optical losses, input laser power, and thermal tuning power. For both, we model in detail the dominant components of the datapath, such as input and output buffers, crossbars, row and column buffers, arbiters, and the vertical/horizontal buses. Other logic modules such as the Link Control Block (LCB) [29] and statistical counters contribute to the total power, but their contribution is negligible.

In the YARC model, to calculate peak power we assume a 100% load on input and output buffers. Although each subswitch can be fully loaded, the aggregate load on all of the subswitches is inherently limited by the switch's bandwidth. For example, in a switch with n subswitches handling uniform traffic, the mean load on each subswitch is no greater than $100/\sqrt{(n)}\%$, even when the switch is operating at full load. Similarly, the number of bytes transferred in horizontal and vertical buses is also limited to the aggregate I/O bandwidth.

5.6 Results

Our initial experiments compare the performance and power of the optical full crossbar with a YARC style electronic crossbar for a range of switch sizes and traffic types. Overall, the performance results show that a YARC style electronic crossbar can perform as well as an optical crossbar, but as the radix and port bandwidth increase the power consumed by the electronic crossbar becomes prohibitive. Finally, we present power results for large networks based on the various switches that we have modeled.

5.6.1 Performance Results

Both switches do well on most traffic patterns, except for some contrived patterns where YARC performs poorly. Once the switch radix is large, the performance variation due to switch radix is minimal, making the performance results for all three radices roughly equivalent. The performance results also don't change appreciably at the different technology nodes. With an optical datapath, both electrical and optical arbitration schemes provide roughly the same performance because the electrical scheme is fast enough for our data points. The main benefit of the higher radix switches comes at the system level, where hop-count, switch power and cost are reduced.

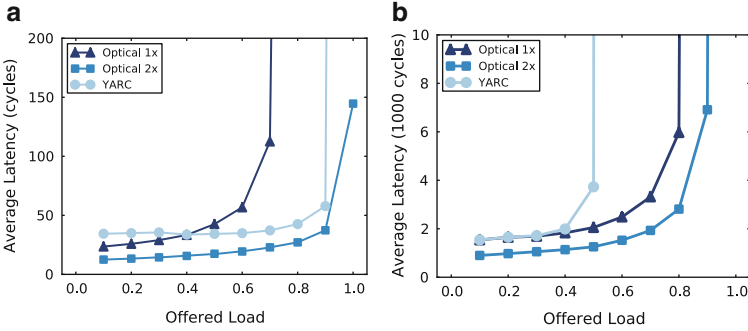


Fig. 5.10 (a) Uniform random traffic, 64 byte packets, and 22nm technology. The 1x and 2x refer to the internal speedup of the optical switch. (b) Uniform random traffic, 9216 byte packets, and 22nm technology. The 1x and 2x refer to the internal speedup of the optical switch

Figure 5.10 shows the performance for uniform random traffic with 64 byte packets across three switch configurations at the 22 nm technology node. The performance of the optical crossbar with and without speedup brackets the YARC design. The optical crossbar, without speedup, is performance limited by its inability to catch up when an input is unable to send to a particular output due to contention. Though YARC also doesn't have internal speedup, the column wires, being independent resources, in effect give the output half of the switch significant speedup. With very large input buffers, the YARC design is easily able to keep its row buffers filled and thus output contention never propagates back to the input stage. The increase in latency with the applied load is almost identical for both approaches reflecting the fact that although the YARC is a multistage design, the use of minimal shared internal resources means that it performs as well as a full crossbar.

Figure 5.10b shows the performance for jumbo packets. With jumbo packets, there are two problems with the YARC design which prevent high throughput. First, the row buffers are too small to store an entire packet, so congestion at the output causes the packet to trail back through the switch and the row bus and results in HOL blocking. Since we are targeting switches for Ethernet networks, flits cannot be interleaved because packets must be single units. We can fix this HOL blocking by providing credit-based flow control from input to output, but even with zero-latency flow control this doesn't improve the load that the switch can handle because the switch is unable to keep the column buffers full, thus losing its ability to catch up when there is output contention. The optical crossbar without internal speedup does better with large packets because the duration of output contention is short compared to the duration of packet transmission (i.e., a failed arbitration might cause a few cycle loss of bandwidth whereas the data transmission takes hundreds of cycles).

Table 5.4 Switch core power in watts

Generation	Port BW	Core type	Radix		
			64	100	144
45 nm	80 Gbps	Electronic	41.8	72.7	120.7
		Optical	13.2	17.4	31.9
32 nm	160 Gbps	Electronic	38.0	65.9	109.0
		Optical	22.9	27.7	50.9
22 nm	320 Gbps	Electronic	52.4	91.9	153.8
		Optical	34.2	41.3	76.3

5.6.2 Power Results

Table 5.4 compares the peak power for optical and electronic switch cores for various switch sizes and technology generations. It is clear that across all technology nodes, optical cores consume less power. In many cases the electrical switch power is very high, so that even if we break the pin barrier with optical off-chip interconnects, it is not feasible to build high-bandwidth, high-radix electric switches without incurring exorbitant cooling costs.

Compared to electrical switch cores, optical core power increases more slowly with radix. In electrical switches, the buffered crossbar design is a key to enabling high throughput. But its complexity grows quadratically with radix, leading to high power consumption. The row/column interconnects, consisting of fast repeated wires switching at high frequency, contribute heavily to the total power in electrical switches. Optical switch cores overcome both these problems by leveraging the superior characteristics of optical interconnect and our arbitration scheme. The proposed 8-request, 2-grant scheme is able to achieve high-throughput without intermediate buffers. The optical crossbar is effective in reducing the communication overhead. The only optical component that scales nonlinearly is the laser power (due to the loss in the link), but its contribution to the total power is minimal. The clustering technique helps keep the laser power contribution low even for high radices by reducing the number of optical rings required.

Table 5.5 shows the total power including I/O for all configurations. For high port count, devices with electronic I/O become impractical. Across the design space, electronic switch cores are considered feasible if the total power consumed is within 140 W. Beyond this threshold, more expensive conductive liquid cooling is required. Hence for high port count designs, the optical switch core has a considerable power advantage. Packaging requirements make the case even stronger for photonics.

Figure 5.11 shows the per-bit energy for large-scale HyperX networks [2] for a range of switch components in the 22nm generation. This shows a double advantage of photonic I/O in both reducing power and enabling higher radix switches; switches of greater than 64 ports with electronic I/O exceed practical device power limits and packaging constraints. The combination of greater radix and lower component power leads to a factor-of-three savings in interconnect power for large networks using photonic I/O. A further $2\times$ power savings can be realized by exploiting

Table 5.5 Overall switch power including I/O in watts

Generation	Port BW	Switch		Radix		
		core	I/O	64	100	144
45 nm	80 Gbps	E	E	77.6	128.7	201.4
		E	O	44.1	76.3	125.9
		O	O	15.5	21.0	37.0
32 nm	160 Gbps	E	E	89.7	146.7	225.3
		E	O	40.9	70.4	115.5
		O	O	25.8	32.2	57.5
22 nm	320 Gbps	E	E	135.3	221.5	340.4
		E	O	56.3	98.0	162.6
		O	O	38.1	47.4	85.1

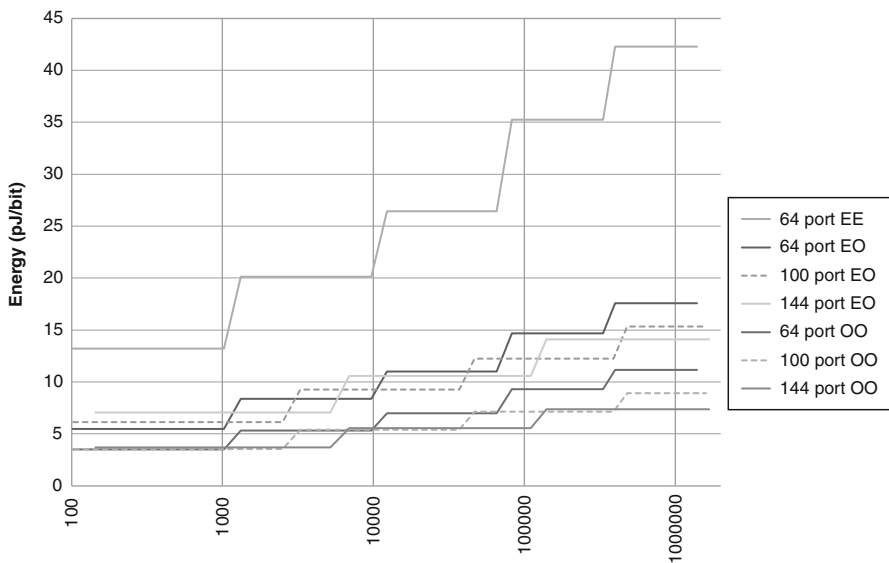


Fig. 5.11 Switch throughput comparison

photonics for the switch core. When photonics is applied in our channel per destination approach, the tuning power of idle modulator rings becomes the most significant power overhead.

5.7 Related Work

Single-chip CMOS high-radix Ethernet switches with up to 64 ports have recently become available [8, 10]. A significant fraction of the silicon area and power consumption in these devices is associated with the complexity of Ethernet routing.

In this work we assume a simplified, compact addressing scheme to avoid the need for content addressable memories for routing tables in a sparse address space. In a multistage network used for Ethernet traffic, the function of translating between standards-based addressing schemes and the compact scheme is required only at the ingress side of the network. This saves power on inter-switch transfers and enables larger switches to be constructed due to the lower routing overhead.

Recent work has studied the design challenges of building high-radix single chip switches. Mora et al. [26] propose partitioning the crossbar in half to improve scalability. We follow [17] by using a deeper hierarchical structure to construct electronic switch cores. A more detailed discussion on the implementation of the YARC switch is contained in [29].

The state of the art for CMOS integrated photonics today is limited to simple transceiver devices [4]. Krishnamoorthy et al. [24] demonstrate some of the component technologies for larger-scale integrated CMOS photonics in chip-to-chip applications. However this work is focused on the use of photonics to build photonic enabled macrochips, rather than components for use in data center networks. The use of integrated photonics for intra-chip communication is the subject of much current research. Shacham et al. [31] propose an on-chip optical network for core-to-core communication. In this case, the switching function is optical circuit switching with an optical path being established between the communicating cores. While this can be more power efficient for long transfers, it is less efficient for heavy short-packet loads.

5.8 Conclusions

In this chapter, the authors argue that integrated CMOS photonic I/O provides the capability to scale the radix of routers beyond the electrical pin and power limitations of projected CMOS technology. A number of conclusions can be drawn from the data presented in this chapter. Once optical I/O's are used to break the pin barrier, the next bottleneck will be global on-chip wires for switch radices greater than 64. This can be addressed using a flat optical crossbar. By leveraging high-bandwidth optical waveguides to provide significant internal speedup, and by using an arbitration scheme that allows eight requests for each ingress port but which is only allowed to accept up to two grants. This scheme overcomes the HOL blocking problem.

The photonic crossbar might not be heavily loaded which could result in excess tuning power. To reduce this inherent high static power problem, a clustered approach can be employed to balance the use of optics and electrical wires. The architecture presented here restricts the use of buffers to just input and output ports, and this makes it feasible to size them adequately to handle jumbo packets that would be needed to be compatible with Ethernet switches. In addition, the lack of intermediate buffers reduces the number of EO and OE domain conversions which improves both latency and power consumption. The analysis shows that photonics

can reduce the system power in several ways. With the adoption of optical I/Os, power can be reduced by up to 52%. The use of an optical datapath provides another 47% reduction in power for a radix 144 switch in a 22 nm process. Clustering to allow ring sharing by multiple ports, the power can be further reduced by 41% in a radix 64 switch.

Exploiting both the power savings of photonics and the reduction in component count resulting from the ability to build higher radix switches gives a factor of six improvement in the energy per bit for a 100,000 port photonic switch network compared to an all electrical implementation.

As datacenters grow in both scale and performance, the interconnection network will become increasingly critical. Given the advantages of photonic communication, it is clear that photonics will continue to be a key technology to deploy in future datacenters. It makes sense to use photonics where it will be the most effective. At the time of this writing (2012), copper inter-rack cabling is being replaced by active optical cables to improve latency on the cable, and to reduce the energy consumed by communication. The next step is to replace copper intra-rack cables or backplane traces with optical waveguides. Both of these options have already been demonstrated. As intra-rack signaling becomes photonic, the OE and EO conversion engines will move from inter-rack AOCs to inside the rack. The next step is to put waveguides on the long data traces on the server blade. There are several ways to do this. Tan et al. [32] provide a nice study of this option. The next step is to move photonics onto the die. From a datacenter perspective, the most helpful choice for on-die photonics is the switch-chip that is the foundation for the datacenter's interconnection network. This chapter has described how much photonics will help in terms of energy efficiency and the ability to build much higher-radix switch-chips.

References

1. U.S. Environmental Protection Agency ENERGY STAR Program (2007) Report to Congress on Server and Data Center Energy Efficiency Public Law 109–431 Washington D.C., USA
2. Ahn J, Binkert N, Davis A, McLaren M, Schreiber RS (2009) HyperX: topology, routing, and packaging of efficient large-scale networks, Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis, Portland, Oregon
3. Ahn J, Fiorentino M, Beausoleil R, Binkert N, Davis A, Fattal D, Jouppi N, McLaren M, Santori C, Schreiber R, Spillane S, Vantrease D, Xu Q (2009) Devices and architectures for photonic chip-scale integration. *Appl Phys A: Mater Sci Process* 95(4):989–997
4. Analui B, Guckenberger D, Kucharski D, Narasimha A (2006) A fully integrated 20-Gb/s optoelectronic transceiver implemented in a standard 0.13 micron CMOS SOI technology. *IEEE J Solid-State Circ* 41(25):2945–2955
5. Association SI (2009) International technology roadmap for semiconductors. <http://www.itrs.net/>
6. Astfalk G (2009) Why Optical Data Communications and Why Now? *Appl Phys A* 95:933–940
7. Binkert NL, Dreslinski RG, Hsu LR, Lim KT, Saidi AG, Reinhardt SK (2006) The M5 Simulator: modeling networked systems. *IEEE Micro* 26(4):52–60
8. Broadcom (2010) BCM56840 series high capacity StrataXGS® Ethernet switch series. <http://www.broadcom.com/products/Switching/Data-Center/BCM56840-Series>

9. Chen L, Preston K, Manipatruni S, Lipson M (2009) Integrated GHz silicon photonic interconnect with micrometer-scale modulators and detectors. *Opt Express* 17(17):15248–15256
10. Cummings U (2006) FocalPoint: a low-latency, high-bandwidth Ethernet switch chip. In: *Hot Chips 18*
11. Dimitrakopoulos G, Galanopoulos K (2008) Fast arbiters for on-chip network switches. In: *International conference on computer design*, pp 664–670
12. Fukuda K, Yamashita H, Ono G, Nemoto R, Suzuki E, Takemoto T, Yuki F, Saito T (2010) A 12.3mW 12.5Gb/s complete transceiver in 65nm CMOS. In: *ISSCC*, pp 368–369
13. Hewlett SJ, Love JD, Steblina VV (1996) Analysis and design of highly broad-band, planar evanescent couplers. *Opt Quant Electron* 28:71–81. URL <http://dx.doi.org/10.1007/BF00578552>, 10.1007/BF00578552
14. Ho R (2003) *On-Chip Wires: Scaling and Efficiency*. PhD thesis, Stanford University
15. 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
16. Karol M, Hluchyj M, Morgan S (1987) Input versus output queueing on a space-division packet switch. *IEEE Trans Comm* 35(12):1347 – 1356. DOI 10.1109/TCOM.1987.1096719
17. Kim J, Dally WJ, Towles B, Gupta AK (2005) Microarchitecture of a High-Radix Router. In *ISCA '05: Proceedings of the 32nd annual international symposium on computer architecture*, IEEE Computer Society, pp 420–431
18. Kim J, Dally WJ, Abts D (2006) Adaptive Routing in High-Radix Clos Network. In: *SC'06*
19. Kim J, Dally WJ, Abts D (2007) Flattened butterfly: A cost-efficient topology for high-radix networks, Proceedings of the 34th annual international symposium on Computer architecture, San Diego, California, USA doi)10.1145/1250662.1250679
20. Kim J, Dally WJ, Scott S, Abts D (2008) Technology-Driven, Highly-Scalable Dragonfly Topology, Proceedings of the 35th International Symposium on Computer Architecture, Beijing, China, pp 77–88 doi)10.1109/ISCA.2008.19
21. Kirman N, Kirman M, Dokania RK, Martinez JF, Apsel AB, Watkins MA, Albonese DH (2006) Leveraging optical technology in future bus-based chip multiprocessors. In: *MICRO 39 Proceedings of the 39th Annual IEEE/ACM International Symposium on Microarchitecture* pp 492–503
22. Koch BR, Fang AW, Cohen O, Bowers JE (2007) Mode-locked silicon evanescent lasers. *Opt Express* 15(18):11225
23. Kogge (editor) PM (2008) *Exascale computing study: technology challenges in achieving exascale systems*. Tech. Rep. TR-2008-13, University of Notre Dame
24. Krishnamoorthy A, Ho R, Zheng X, Schwetman H, Lexau J, Koka P, Li G, Shubin I, Cunningham J (2009) The integration of silicon photonics and vlsi electronics for computing systems. In: *International conference on photonics in switching, 2009. PS '09*, pp 1 –4. DOI 10.1109/PS.2009.5307781
25. Lipson M (2005) Guiding, modulating, and emitting light on silicon—challenges and opportunities. *J Lightwave Technol* 23(12):4222–4238
26. Mora G, Flich J, Duato J, López P, Baydal E, Lysne O (2006) Towards an efficient switch architecture for high-radix switches. In *ANCS '06: Proceedings of the 2006 ACM/IEEE symposium on Architecture for networking and communications systems*, New York, NY, USA, ACM, 2006, pp. 11–20.
27. Muralimanohar N, Balasubramanian R, Jouppi N (2007) Optimizing NUCA organizations and wiring alternatives for large caches with CACTI 6.0. In *Proceedings of the 40th International Symposium on Microarchitecture (MICRO-40)*
28. Palmer R, Poulton J, Dally WJ, Eyles J, Fuller AM, Greer T, Horowitz M, Kellam M, Quan F, Zarkeshvarl F (2007) *Solid-State Circuits Conference. ISSCC 2007, Digest of Technical Papers*. IEEE International, San Francisco, CA 440–614
29. Scott S, Abts D, Kim J, Dally WJ (2006) The black widow high-radix Clos network. *Proceedings ISCA '06 Proceedings of the 33rd annual international symposium on Computer Architecture*, IEEE Computer Society, Washington, DC, USA pp 16–28

30. Semiconductor Industries Association (2009 Edition) International technology roadmap for semiconductors. <http://www.itrs.net>
31. Shacham A, Bergman K, Carloni LP (2007) On the design of a photonic network-on-chip. In: First International Symposium on Digital Object Identifier, NOCS, pp 53–64
32. Tan MR, Rosenberg P, Yeo JS, McLaren M, Mathai S, Morris T, Kuo HP, Straznicky J, Jouppi NP, Wang SY (2009) A high-speed optical multidrop bus for computer interconnections. *IEEE Micro* 29(4):62–73
33. Vantrease D, Binkert N, Schreiber RS, Lipasti MH (2009) Light speed arbitration and flow control for nanophotonic interconnects. In: MICRO-42. 42nd Annual IEEE/ACM International Symposium on MICRO-42, pp 304–315
34. Warren D (2011) HP Optical Backplane Demonstration, InterOp. <Http://www.youtube.com/watch?v=dILsG8C6qVE>
35. Watts MR, Zortman WA, Trotter DC, Nielson GN, Luck DL, Young RW (2009) Adiabatic resonant microrings (ARMs) with directly integrated thermal microphotonics. In: Lasers and Electro-Optics, 2009 Conference on Quantum electronics and Laser Science Conference, pp 1–2
36. Xu Q, Schmidt B, Pradhan S, Lipson M (2005) Micrometre-scale silicon electro-optic modulator. *Nature* 435:325–327

Chapter 6

All-Optical Networks: A System’s Perspective

Nikolaos Chrysos, Jens Hofrichter, Folkert Horst, Bert Offrein,
and Cyriel Minkenberg

6.1 Introduction

Computer interconnection networks for commercial data centers (DC) as well as scientific high-performance computing (HPC) installations are at the core of a profound shift from the conventional *computing-centric* (i.e., CPU-centric) system view to a more *communication-centric* (i.e., network-centric) view. The trend in computing in general has been towards *increasing parallelism* at all levels: more threads per core, more cores per CPU, more CPUs per node, more nodes per machine, more machines per cloud. The full potential of all these levels of parallelism can only be unleashed if the communication infrastructure enables it.

In this context “interconnection network” mainly refers to various types of *system area networks* (as opposed to wide area networks), which comprise (a) local area networks (LANs) such as Ethernet, (b) storage area networks (SANs), (c) clustering or inter-process communication (IPC) networks, including standardized or proprietary offerings, and (d) I/O expansion networks, such as PCI Express (PCIe).

There is a strong drive to improve the efficiency, flexibility, manageability, and total cost of ownership of interconnection networks by consolidating the different types of traffic (LAN, SAN, IPC) on a common, shared physical infrastructure

Part of this chapter was published previously in Proc. IEEE 12th International Conference on High Performance Switching and Routing (HPSR), Dallas, TX (2010).

IBM, the IBM logo, and *ibm.com* are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. Other product and service names might be trademarks of IBM or other companies.

N. Chrysos (✉) • J. Hofrichter • F. Horst • B. Offrein • C. Minkenberg
IBM Research – Zurich, Säumerstrasse 4, Rüschlikon, CH-8803, Switzerland
e-mail: cry@zurich.ibm.com; jho@zurich.ibm.com; fho@zurich.ibm.com; ofb@zurich.ibm.com; sil@zurich.ibm.com

as well as by virtualization of network resources. Such a converged, virtualized interconnect implies a more comprehensive and stringent set of requirements. We briefly summarize a few key requirements here, some of which differ significantly from those found in, for example, IP routers.

- *Scalability*: The interconnect architecture should enable scaling to thousands of nodes in a cost-efficient manner. Extension of an existing network to support a larger node count should be possible in an incremental fashion, i.e. without having to replace a disproportionate amount of the installed hardware.
- *Reliability*: DC and HPC networks have very demanding loss rate requirements, because the higher layers are typically very intolerant with respect to packet losses. To achieve the objective of close to 100% reliability, suitable link-layer coding techniques in combination with end-to-end retransmission policies are needed. Moreover, all existing DC interconnects employ some form of link-level flow control to avoid unnecessary losses due to buffer overflows.
- *Latency*: End-to-end latency is a crucial factor in the overall system performance, as it directly determines the penalty in performing a remote operation, for instance accessing a remote memory location, passing an MPI message, or performing a cache line update. With increasing system sizes, the latency component purely due to the interconnect grows significantly. This latency comprises the network interfaces at the send (ingress) and receive (egress) sides, the network switches, and all interconnecting cables or fibers. Moreover, the latency has a pure *time-of-flight* component, i.e. the time it takes to traverse an otherwise idle network, and a *queuing* component. Typical end-to-end latency requirements are in the low single-digit microsecond range.
- *Data rate and throughput*: With the advent of 10-Gigabit Ethernet (10GE) and quad-data-rate (QDR) InfiniBand, 10 Gb/s has become the minimum data rate to be supported. The effective throughput should naturally be as close to 100% as possible, especially as convergence and virtualization are both expected to drive mean utilization higher. Factors that reduce throughput include various overheads (line coding, protocol headers, segmentation) and other inefficiencies (because of topological issues, routing, scheduling, queuing, contention, congestion, acknowledgements, retransmissions).

6.1.1 Optical Transmission

Optical transmission offers three major benefits in computer interconnects compared to electrical transmission:

Power: Optical fibers and waveguides exhibit very low loss compared to electrical wires, thus requiring much less power to cover a given distance. Moreover, as most systems are already using optics to connect racks, additional power can be saved by

implementing the switching nodes in an all-optical fashion, thereby eliminating the need to perform O/E/O conversions at every hop.

Data rate: Optical fibers and waveguides are basically data-rate transparent. Aggregate fiber bandwidth exceeds that of copper cables by orders of magnitude.

Density: Wavelength division multiplexing (WDM), which has no electrical equivalent, can considerably increase bandwidth density [8]. In addition, waveguide pitches are significantly smaller than wire pitches.

6.1.2 Architectural Considerations

Although there has been an enormous amount of work on all-optical interconnects [2, 12, 15–17, 22], protocols, and devices, there is still no common consensus on the high-level organization of these networks. This work attempts to formulate such a generic architecture that is derived from lessons in current electronic networks, as well as on the basic premises that:

- (1) Optical buffering will need time to mature, although significant progress, e.g. using slow light, has been made recently [3].
- (2) The integration density of optical devices is still orders of magnitude removed from that of electronics.

In the following, we describe a set of “best practices” for all-optical interconnects that have driven the design of our architecture as described in Sect. 6.2. We believe that some of these practices apply beyond the specific architectures that we propose.

6.1.2.1 Time-Division-Multiplexing

Scalable multistage networks rely either on complicated scheduling schemes [5], or on the existence of relatively large buffers inside the network [6]. Both of these alternatives contradict the premises stated earlier regarding the capabilities of optical devices and networks in the foreseeable future.

On the other hand, time-division-multiplexing (TDM) scheduling can provide high throughput in certain types of networks, e.g. those based on multistage interconnection networks (MINs), without requiring any buffers to do so. At the same time, orchestrating packet injections according to a fixed, TDM schedule is scalable to large port counts, as it relies on simple, distributed control units, requiring no coordination other than a global periodic signal (clock). Effectively, TDM scheduling is very attractive for all-optical interconnects [13, 21].

There is, however, a significant cost to be paid for these benefits: the fixed TDM schedule imposes packet delays that are intolerable for packet-switched networks, especially those employed for interprocessor communications.

6.1.2.2 End-to-End Reliability

There is a long list of possible hazards in an all-optical, multistage interconnect, e.g. power penalties when the optical signal is passing through switches and waveguide crossings, just to mention a few [18]. In traditional networks, link-level error detection and recovery schemes may be used to deal with such random errors. As implementing such protocols in the optical domain is not feasible at present, we have to rely on end-to-end reliable delivery schemes, implemented by network interfaces (adapters) in the electronic domain.

Besides improving robustness to errors, an end-to-end reliability scheme can support a viable solution to the high-latency issue of TDM-operated networks, and thus broaden the applicability that these networks have. In particular, a possible method to avoid the TDM latency at low loads is to allow speculative/eager packets to override the TDM schedule. In a distributed, uncoordinated setup, breaking the TDM rules may result in collisions among eager packets, or between eager and prescheduled ones. TDM scheduling avoids collisions among prescheduled packets, under the premise that they all progress synchronously in lock-step; if one of them is delayed, e.g. because of an eager packet getting in its way, then prescheduled packets may collide with one another. Thus, a better way to resolve contention is to drop eager packets, when required, and allow prescheduled ones to move forward so as to sustain throughput. As we show here, by means of a responsive, low-delay, end-to-end reliability scheme, we can recover the lost eager packets, and sustain reasonable mean packet delays.

6.1.2.3 Minimally-Sized Switching Nodes

As the switch complexity (in terms of data path and control plane) scales super-linearly with the number of switch ports, small-radix switches have a definite advantage. In this work, we assume the minimum switch radix of two, although the results qualitatively apply to larger radices also. Note that switching nodes need not have any buffers if the network conveys prescheduled (TDM) packets only. But in view of the requirement for a speculative mode, the capacity to store some packets in case of collision can improve performance. Therefore, we assume a minimal buffering capacity per output link, which for a 2×2 switching node can be implemented using a couple of fiber delay lines per switch output [3].

6.1.2.4 Multistage Interconnection Networks

Based on our objective to keep the switch radix small while allowing cost-effective scalability to large node counts, we turned to logarithmic unidirectional multistage interconnection networks (MINs). These topologies are sometimes referred to collectively as *Delta*, *Banyan*, or k -ary n -fly networks and comprise $n = \log_2 N$ stages, each stage having N/k switches with k ports per switch. They support $N = k^n$

end nodes, have a diameter equal to $n - 1$, and have a bisection ratio equal to one. Therefore, such a topology can scale to very large node counts using a small switch radix without compromising on bisectional bandwidth and with good scaling of the diameter ($\log_k(N)$ instead of $\sqrt[n]{N}$ for k -ary n -meshes). Moreover, these networks are all *self-routing*: to reach a destination node with address $d_{n-1}d_{n-2}\cdots d_1d_0$, with $0 \leq d_i < k$, a switch in stage $n - i$ routes to output port d_i , the stages being numbered in increasing order from left to right. This makes the routing algorithm easy to implement: the most significant digit of the destination node identifier serves as the local routing tag for the first stage, the second-most significant digit for the second stage, etc.

The basic MIN topology cannot support arbitrary non-uniform traffic. However, one can use it to implement more expensive and more efficient networks, such as the Beneš topology or the load-balanced switch [4], which can provide full-throughput for any admissible arrival pattern.

6.1.3 Previous Work

TDM scheduling schemes have their roots in telephony (circuit-switching) networks, and have reappeared recently, in various ways, in packet-switched networks [1, 4, 10, 14, 18, 21]. We employ TDM scheduling in our architecture to address the inherent performance limitations of multistage networks with no or limited internal buffers [7, 9, 20]. Another alternative that we examined is dynamic scheduling, which also reduces the degree of blocking in multistage fabrics with small internal buffers, but requires larger buffers and more complicated control schemes [6].

The idea to drop eager packets when they collide with prescheduled ones, by taking advantage of reliable delivery retransmissions, was first presented in [11]. Our proposal extends the work of [11], by (a) applying it to TDM scheduling, and (b) considering end-to-end retransmissions in a *multistage* network. In addition, [11] uses both positive ACKs and negative NACKs that are sent out-of-band on special lossless channels. We, on the other hand, consider only positive ACKs¹ which are routed in-band, through the data network, and can therefore be lost. In the absence of NACKs, we handle packet drops by means of retransmission timeouts at ingress adapters. These issues make it harder for speculation to work as intended. One particular difficulty is that in-band ACKs can nearly double the network utilization, thus reducing the network capacity that is required for successful speculations. We address this issue by dynamically adapting the retransmission timeout periods so as to increase the success of piggybacking and reduce ACK overhead.

¹As NACKs should be generated within the optical fabric, their implementation would not be feasible in the proposed architecture.

6.1.4 Contents

In Sect. 6.2, we describe the network topology and the scheduling and flow control schemes that we use. We also introduce the concepts of prescheduled and speculative (eager) injections, as well as the rules that govern their coexistence in the network. To ensure reliable, in-order delivery, we use a bandwidth-efficient end-to-end reliable delivery scheme based on selective retry, as described in Sect. 6.3. We evaluate the latency-throughput characteristics of the proposed architecture in Sect. 6.4, and conclude this study in Sect. 6.5.

6.2 Network Architecture

The system comprises end nodes (e.g., compute nodes, servers, storage), which are attached to the interconnect via network interfaces (adapters) that incorporate electronic buffering and control, along with E/O and O/E conversions to interface with the network. Every adapter comprises an ingress path (end node to network) and an egress path (network to end node). At its ingress path, an adapter accepts incoming packets, stores them in virtual output queues (VOQs), and converts them to the optical domain before injecting them into the network. At its egress path, an adapter receives optical packets from the network, converts them to the electrical domain, performs re-ordering and reassembly when required, and forwards them to the attached end node. We consider synchronous slotted time for all nodes, where a time slot equals the duration of a fixed-size packet on a network link.

The network itself provides an all-optical data path end-to-end, i.e. without any O/E/O conversions along the way. Among the k -ary n -fly networks, we selected the Omega network (see Fig. 6.1), which is characterized by having a perfect shuffle permutation before each network stage; the connections from the last stage to the end nodes do not produce a permutation.

As mentioned in Sect. 6.1.2.3, we consider switching elements that contain a couple of FDLs per output (shown in Fig. 6.2) in order to handle speculative injections more efficiently. To prevent buffer overflow, every switch conveys *stop* & *go* flow control to its upstream nodes, individually controlling access to each FDL. In order to route these flow control signals in unidirectional MINs, a backward communication channel needs to be present between any pair of forward-connected switches.

We performed a preliminary study on the gate-level implementation of the control unit for such a switch. We found that output link arbitration and assertion of the backpressure signals require a few tens of two-input logical gates, and a few one-bit latches for storing the backpressure and scheduling (round-robin, next-to-serve) state for each output. Such two-input gates and one-bit latches may be realized using novel arrangements of optical devices implementing digital photonic logic; however, one may also implement them using conventional CMOS technology.

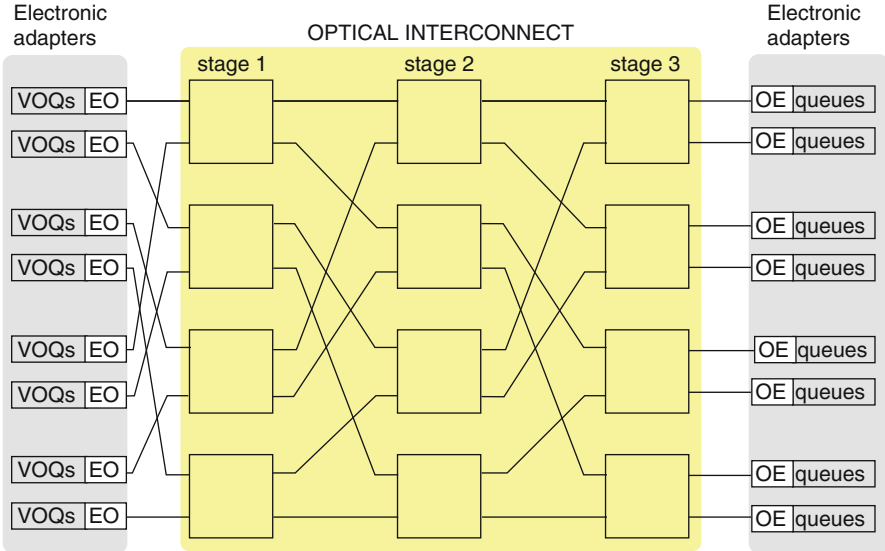


Fig. 6.1 An 8x8 Omega network

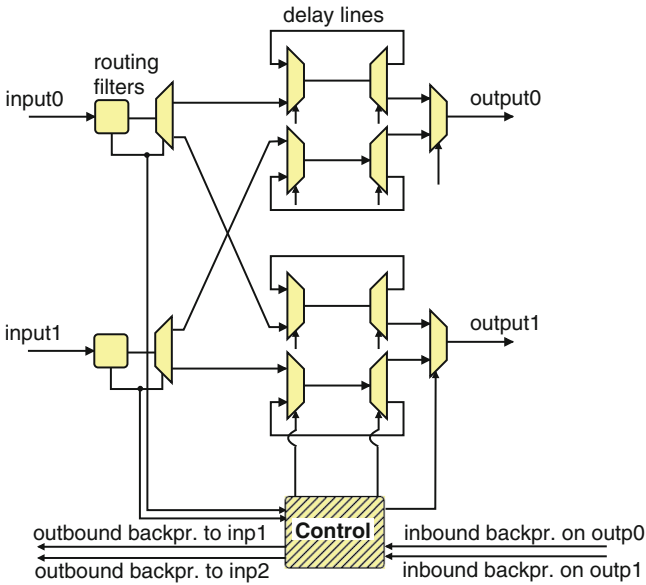


Fig. 6.2 A 2x2 switching element, with one delay line per input/output pair

Figure 6.3a depicts the mean packet delay of this lossless buffered Omega network (baseline), for uniform, random traffic, as a function of offered load, for

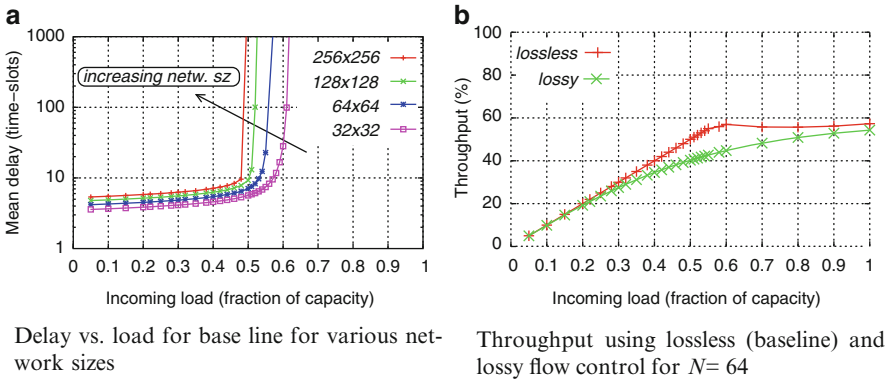


Fig. 6.3 Throughput and delay performance of Omega networks

different network sizes.² The maximum throughput for $N = 32$ is approximately 60%, and declines to 50% for $N = 256$ — N denotes the network size. Hence, in a lossless network, once the incoming load exceeds the maximum throughput of the system, packets queue up in VOQs.

This effect gives us the incentive to present the performance of an optical network that ignores flow control signals. In such a *lossy* network, packets are never blocked and are dropped when they arrive at a full buffer. Figure 6.3b compares the throughput of the lossy network to the baseline. The figure shows that ignoring the flow control signals leads to even worse throughput than conforming to them. Note that in this experiment, we do not recover dropped packets.

6.2.1 TDM Prescheduled Packet Injections

Prescheduling the injections according to a TDM schedule works as follows: At every time slot, a set of prescheduled flows (i.e. VOQs) is eligible to inject a packet. In each set, each source or destination node appears exactly once. Thus, each set is essentially a full bipartite graph matching and can be characterized as a permutation of the nodes. When a prescheduled VOQ is non-empty, the corresponding source adapter injects the head-of-line (HOL) packet from this VOQ; otherwise, the source adapter stays idle.

To ensure conflict-free routing of prescheduled packets in the Omega network, we exploit the property that an Omega network can route, without conflicts, any permutation wherein destinations are monotonically increasing in modulo- N arithmetic [10]. Figure 6.4 depicts four conflict-free matchings for $N = 4$;

²Simulation models and parameters are described in Sect. 6.4.

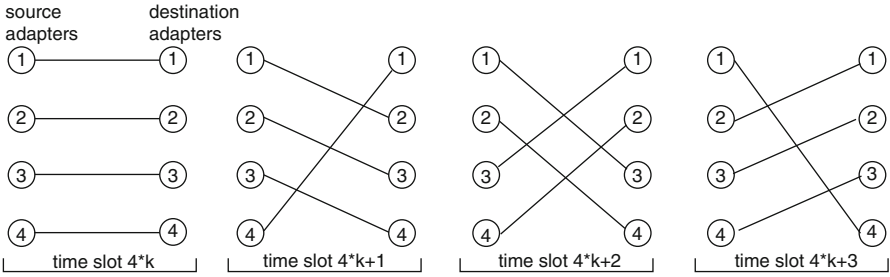


Fig. 6.4 Prescheduled flows for 4x4 network

these are four differently shifted instances of the identity permutation. Such static round-robin-based patterns have been proposed for use in crossbar schedulers [21], Clos networks [19], and load-balanced switches [4].

Given the above set of N routable matchings, we must consider the order in which to apply them. We consider a fixed, repeated sequence of matchings, visiting every source-destination pair once every N time slots. This sequence is suitable for uniform traffic, where the flow arrival rate never exceeds $1/N$.³ The selected sequence of matchings for $N = 4$ is presented in Fig. 6.4. Generalizing for arbitrary network size N , at time slot t , source i is matched to destination $(t + i) \bmod N$.

Figure 6.5 shows that for a 64×64 Omega network the saturation throughput is 58% *without* prescheduling and 100% *with* prescheduling. On the downside, prescheduling increases the delay at low loads by about 31 time slots. It follows that, at low loads, for networks of size N that use prescheduling, the maximum delay equals $N-1$ time slots, and the average delay $\sim \frac{N}{2}$ time slots. In fact, for i.i.d. uniformly distributed Bernoulli packet arrivals, at an arbitrary load of $\lambda < 1$, the average packet delay can be estimated by [23]:

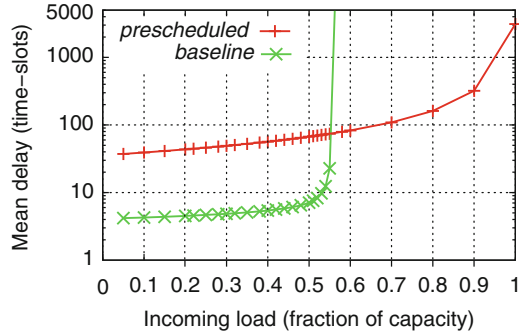
$$E(d) = \frac{N-1}{2} + \frac{(N-1) \times \lambda}{2 \times (1-\lambda)} \tag{6.1}$$

6.2.2 Speculative Packet Injections

We can remove the large delay overhead of prescheduled transmissions by allowing adapters to speculatively inject packets from non-empty VOQs at low loads, without having to wait for the designated prescheduled time slots. Speculative injections

³We restrict this work to uniform traffic, as Omega networks are inherently incapable of routing all possible non-uniform traffic patterns because of internal conflicts. In principle, other matching sequences are possible, and the applied matching sequence could be changed dynamically.

Fig. 6.5 Packet delay vs. load for a 64×64 Omega network with and without prescheduling; the latter system corresponds to the baseline



also allow persistent flows to reach full bandwidth, when they do not face internal contention inside the network. We will refer to packets injected in this way as *eager*.

When the load is low, we expect that almost all packets will be eager. As the load increases and approaches the saturation load of the network, flow control will eventually start blocking eager injections, pushing packets back to their VOQs. Therefore, adapters will start finding their prescheduled VOQs being non-empty, and more injections will be prescheduled. As eager and prescheduled packets can coexist simultaneously in the network, we need to consider how they should interact.

When contention occurs between prescheduled and eager packets, the prescheduled packet always takes precedence to preserve the conflict-free property among prescheduled flows: forcing a prescheduled packet to wait, might lead to a conflict with another prescheduled packet belonging to a different matching in the next time slot. Similarly, if prescheduled and eager packets were subject to non-discriminate flow control, a buffered eager packet could prevent progress of a prescheduled packet. We eliminate such blocking by changing the flow control rules to allow prescheduled packets to overwrite buffered eager packets. When a buffer is held by an eager packet, flow control blocks eager but not prescheduled packets: a prescheduled packet can proceed and overwrite the eager one in the buffer. On the other hand, a prescheduled packet blocks *all* packets, whether eager or not, i.e. prescheduled packets are never overwritten. For this reason, the flow control signal from each buffer carries an extra bit to indicate whether it is exerted by an eager or a prescheduled packet.

Prescheduled packets have strictly higher priority than eager in adapters and switches *and* ignore stop signals caused by eager ones. The main cost is that eager packets can now be dropped in the network. In addition, an eager packet may have to wait indefinitely, if prescheduled packets continuously utilize the output it is targeting. These deficiencies are covered by the reliable delivery mechanism in Sect. 6.3.

When a prescheduled packet that has just overwritten an eager one in a network buffer departs from that buffer, the corresponding flow control signal unblocks the link, thus possibly triggering the injection of new eager packets waiting in the upstream adapter. At high loads, these eager packets are very likely to be

dropped. To prevent such unfortunate injections, a source adapter with a backlog of threshold TH_{pre} or more packets cannot have speculative injections. This threshold will effectively block eager injections at high loads, when packets queue up at adapters.

Setting the threshold TH_{pre} too low will diminish eager injections at low loads, thus diminishing the latency-reducing benefit of speculation. On the other hand, setting the threshold too high may allow too many speculations at loads close to or beyond the saturation point of the baseline system, causing many packet drops and therefore reduced throughput.

6.3 End-to-End Reliable Packet Delivery

We designed an end-to-end, selective-retry reliable delivery scheme in order to examine the performance of speculative injections. We chose selective-retry, and not the cheaper alternative of go-back- N , in order not to flood the network with successfully delivered when some eager packet is dropped. Nevertheless, we tried to keep the protocol simple, and to minimize the hardware components that are required for its implementation. In this section we outline its operation, along with some interesting features that we have incorporated, and refer the reader to [24] for additional issues on reliable packet delivery.

Figure 6.6 depicts the basic steps performed by end-point adapters. To limit the number of pending packets (i.e., those waiting for an ACK) per destination, the source adapter maintains two variables: the next-packet sequence ID, which is appended to every packet injected, and the anticipated ACK ID, which is the ID of the oldest (earliest sent) pending packet: if their difference is W or more, the source

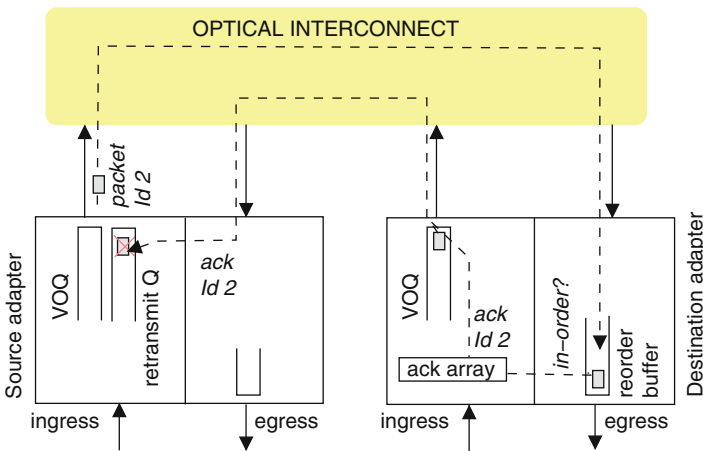


Fig. 6.6 Routing of ACK messages

cannot inject new packets towards the corresponding destination. Every source adapter needs a buffer space for up to W unacknowledged packets per destination, and every destination adapter needs a re-order buffer space for up to W packets per source. Packets are numbered in modulo $2W$ arithmetic.

The source adapter appends a sequence number to the header of every injected packet. Sequence numbers are defined per flow, and are used to identify lost, out-of-order, and duplicate packets. The source adapter also keeps a copy of each packet while waiting for the corresponding ACK message. If this ACK does not arrive in a predefined amount of time, a timeout will expire, causing the retransmission of the packet.

The destination adapter generates ACK messages that carry the ID of the latest in-order packet.⁴ ACKs are first forwarded to the ingress path of the destination adapter, where they are stored in an ACK array with per-flow entries, waiting to be routed through the network to the egress path of their source adapter. ACKs may either be sent as standalone messages or be piggybacked on payload packets that are routed along the reverse path, i.e. from destination to source.

We never retransmit a packet in eager mode, ascertaining that retransmissions are successful in most of the cases. To cover accidental losses of prescheduled packets that may be caused by device or link errors, prescheduled packets also wait for acknowledgment. Hence, if an ACK for a retransmitted packet is lost or excessively delayed, we may have a second retransmission of the same packet.

In addition, in single-path fabrics, prescheduled-only retransmissions eliminate starvation: if an eager packet waits in the network for too long, the source adapter will time out and retransmit. The prescheduled retransmission will follow the route of the starving packet, overwrite it, and reach the destination. In multi-path topologies, one may use a timer, and drop a packet that has been waiting for too long.

6.3.1 *New Injections vs. Retransmissions vs. Standalone ACKs*

The injection policy from input VOQs is as follows⁵: at time slot t , source adapter i selects the prescheduled $((t + i) \bmod N)$ flow, if this flow has a normal packet, i.e. a packet that has not yet been injected, or a packet waiting for retransmission. Otherwise, if the aggregate backlog B in the adapter is below threshold TH_{pre} , then the adapter polls flows searching for an eager injection. Note that backlog B is computed as the number of normal packets in VOQs, plus the number of packets waiting for retransmission.⁶

⁴This is equal to the next-awaited packet ID “minus 1” $\bmod 2W$.

⁵This is the injection policy that we use in our computer simulations. A slightly modified version of it is presented in Sect. 6.4.4.

⁶Not all unacked packets are included in B , but only those that the adapter has to retransmit.

Within a given flow, three kinds of packets may be candidates for injection: (a) packets in the VOQ of that flow, (b) packets waiting for retransmission, and (c) standalone ACK messages for the reverse flow. When a flow is selected, the adapter has to choose which packet to inject from that flow. Highest priority is assigned to retransmissions (only as prescheduled injections). The HOL packet of the corresponding VOQ is considered next, and, finally, standalone ACK messages are considered. The latter can be sent either in prescheduled or in eager mode, whichever comes first.⁷

6.3.2 Reducing ACK Overhead

In-band ACKs consume network bandwidth. To make things worse, we have to consider equal sizes for standalone ACKs and payload packets because the network operates in a time-slotted fashion. Without piggybacking, there is one ACK message for every payload packet, effectively doubling the network load. Piggybacking can reduce ACK bandwidth, but depends on the presence of traffic in the reverse direction. When load is low, we expect a lot of standalone ACKs. Even if the network has the capacity to route these, the increase in network load due to standalone ACKs may reduce the chances for eager transmissions. To ameliorate this effect, we use cumulative ACKs to reduce the volume of ACK IDs that need to be conveyed from destination to source, and we defer the generation of standalone ACKs to the benefit of piggybacking [24].

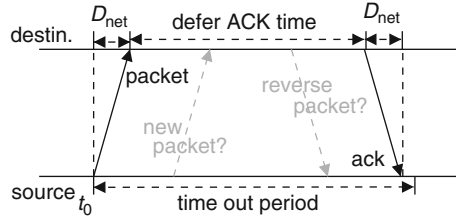
6.3.2.1 Format and Semantics of ACK Messages

By means of an ACK, a destination communicates the ID y of the last in-order packet correctly received from a given source. Assume that when the source receives this ACK, it anticipates ACK ID x . If $(y - x) \bmod 2W \geq W$, then the ACK is considered a duplicate. Otherwise, the source is apprised that the destination has correctly received all packets in the interval $[x, y]$. Thus, with ack combining, communicating a single packet ID may acknowledge up to W packets [24]. Note that the combining of the per-flow ACKs takes place while these wait in the ACK array.

With ack combining alone, the source adapter is unaware of any out-of-order arrivals at the destination. In single-path topologies, like the Omega network, we can perform the following optimization. Every ACK message, in addition to the ID y of the last in-order packet, conveys the ID z of the first (if any) out-of-order packet received by the destination. This allows the source to expedite the retransmission of all pending packets in the interval $[y + 1, z - 1]$, without having to wait for a timeout.

⁷ACKs have the lowest priority, because any injected packet from source s to destination d piggybacks ACK information for the reverse flow $d \rightarrow s$.

Fig. 6.7 The destination deferring sending a standalone ACK message



6.3.2.2 Generation of Standalone ACK Messages

The destination piggybacks ACK information for a given flow $s \rightarrow d$ on *all* reverse packets $d \rightarrow s$. In this way, if an ACK message is dropped inside the fabric, it can be automatically recovered by a subsequent reverse packet. On the other hand, this method may produce numerous duplicate ACKs, but as these are piggybacked, they do not affect the network load.

An insightful observation is that the destination can safely wait for a duration that is linked with the timeout period at the source node before generating a standalone ACK [24]. The benefit is that, while waiting, reverse traffic may arrive and convey the ACK message. Waiting also favors combining several ACKs from packets of the same flow that arrive at the destination within the timeout window into a single acknowledgment.

As shown in Fig. 6.7, assume that a source adapter injects packet p at time t_0 . The source adapter will retransmit packet p at time $t_0 + T_{TO}$, where T_{TO} is the timeout period. Assuming that D_{net} is the delay that packets undergo between the two adapters, in either direction, the destination can safely defer sending the ACK message until time $t_0 + T_{TO} - 2 \cdot D_{net}$. In this way, the destination waits as much as possible to profit from piggybacking and ACK combining, while also making sure that the source adapter will receive the ACK before the respective timeout expires. Thus, we can defer sending a standalone ACK for time T_{defer} , where

$$T_{defer} = T_{TO} - 2 \cdot D_{net}. \quad (6.2)$$

Obviously, with this method, a larger timeout period increases the deferral time and tends to decrease the ACK overhead. However, increasing the timeout also increases the delay of discovering and retransmitting lost packets. Furthermore, a larger timeout increases the window of in-flight packets and requires larger buffers at the end points.

We use an adaptive ACK generation delay to balance this trade-off. Each adapter estimates the mean time between successive packet receptions from each connection, and defers the generation of standalone ACK messages for a longer or shorter time depending on the activity of the reverse connection. It also informs the corresponding source adapter of the new deferral period so that the latter can update its timeout period using (6.2).

In particular, consider adapters A and B , which communicate via connections c ($A \rightarrow B$) and c' ($B \rightarrow A$). Algorithm *Adaptive Timeout*, below, describes how adapter B updates T_{TO} and then T_{defer} (using (6.2)) for connection c . We denote by μ the mean inter-arrival time of packets belonging to c' at adapter B , and consider that the timeout period of connection c can range within $[T_{\text{min}}, T_{\text{max}}]$. Note that k is a constant scaling factor applied to μ .

Algorithm 1 ADAPTIVE TIMEOUT

```

if ( $k \times \mu + 2 \times D_{\text{net}} < T_{\text{min}}$ ) then
   $T_{\text{TO}} = T_{\text{min}}$ ;
else
  if ( $k \times \mu + 2 \times D_{\text{net}} > T_{\text{max}}$ ) then
     $T_{\text{TO}} = T_{\text{max}}$ ;
  else
     $T_{\text{TO}} = k \times \mu + 2 \times D_{\text{net}}$ ;
  end if
end if

```

Note that μ is updated every time, t , that we receive a new packet from connection c' using the method of *exponential moving averages*. Assuming that the last packet from c' was received at time t_{prev} , we update μ when we receive a new packet at time t , as:

$$\mu := \alpha \times (t - t_{\text{prev}}) + (1 - \alpha) \times \mu. \quad (6.3)$$

Parameter α is a constant $\in [0, 1]$. If a reverse connection suddenly becomes inactive, μ may inaccurately retain a small value. Therefore, if before updating the timeout at time t , the adapter finds that the $t - t_{\text{prev}} \gg \mu$, it sets T_{TO} equal to T_{max} .

6.3.3 Retransmission Timers

The source adapter needs to program a timeout event every time it injects a new packet. Doing so typically requires maintaining an excessive number of timers, as each adapter may have up to $N \times W$ packets pending. Here we use only N timers per adapter, essentially one timer per VOQ.

A timer is idle when there is no packet pending towards the corresponding destination. When a first packet is injected, the timer is programmed to expire after a timeout period, and the ID of the injected packet is stored in a variable to bind the pending timeout event with that particular packet. If another packet from the same connection is injected while the timer is still programmed, no new timeout event will be programmed. When the timer expires or the packet that is currently associated with it receives an ACK message, the adapter searches for a next pending packet, if any, from the same connection, to associate the timer with. The timer

is then programmed to expire after a new timeout period. This scheme saves a considerable amount of resources at the expense of timeout accuracy. In the worse case, the timeout for a packet may be delayed for $W - 1$ timeout periods, if $W - 1$ “earlier” packets from the same connection receive a timeout first.

6.4 Evaluation Using Computer Simulations

We implemented computer simulation models for the proposed systems using an event-driver simulator. In our simulations, we assume a finite buffer space for 4,096 packets at each adapter, which is shared among all VOQs. The window size is set to 128 packets. Thus, every adapter needs an additional buffer space for $128 \cdot N$ packets at its ingress path to store pending packets, and another $128 \cdot N$ packets buffer at its egress path to store out-of-order packets. We use random (Bernoulli) traffic, uniformly distributed across all outputs. Unless otherwise noticed, TH_{pre} is set to 24 packets, and $N = 128$ (7 stages). Packets experience a fixed delay of approximately half a time slot at every switch or adapter node, resulting in a time-of-flight of 4 time slots for $N = 128$. The timeout period T_{TO} is allowed to range from 100 to 1,000 time slots, and parameters k and α which pertain to its dynamic adaptation are set to 5 and 0.9, respectively. The estimate of the network delay, D_{net} , is set to 30 time slots.

6.4.1 Impact of Retransmissions

Figure 6.8a,b plot the average packet delay and throughput curves of five system variants:

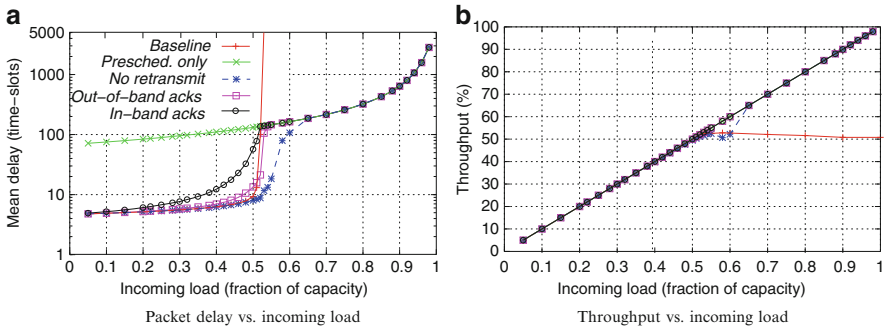


Fig. 6.8 Delay and throughput performance of various systems; $N = 128, TH_{pre} = 24$

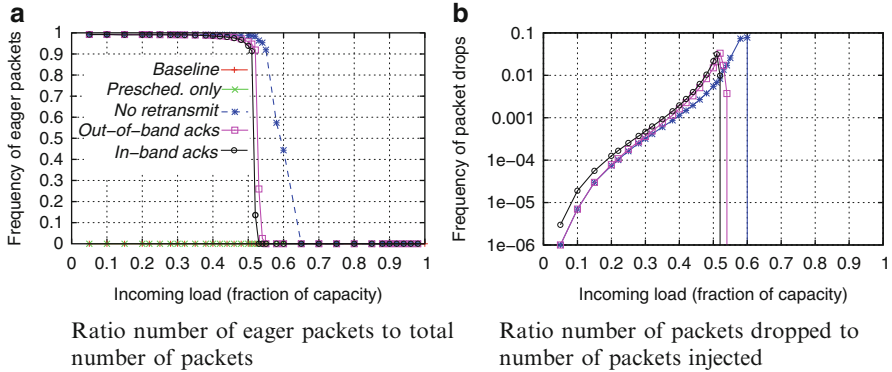


Fig. 6.9 Frequency of eager packets, and frequency of packet drops, in various systems; $N = 128$, $TH_{pre} = 24$

1. *Baseline*: Lossless Omega network without prescheduling or eager transmissions,
2. *Prescheduled only*: Prescheduled but no eager transmissions,
3. *Hybrid without retransmissions*: Combines prescheduled with eager injections, but no retransmissions,
4. *Hybrid with out-of-band ACKs*: As above, but dropped packets are retransmitted and ACKs travel out-of-band,
5. *Hybrid with in-band ACKs*: As above, but ACKs are sent in-band, sharing bandwidth with data packets.

System 1 cannot sustain high loads but has small delays when traffic is light, whereas system 2 exhibits the opposite behavior. Systems 3–5 deliver small delays when possible and toggle to prescheduled transmissions when required to achieve high throughput. The systems differ on the characteristics of this transition.

Figure 6.9a presents the ratio of eager packets to total packets injected for the same systems as in Fig. 6.8. Without reliable delivery, eager injections predominate for a wider load range, as no bandwidth is consumed by retransmissions or ACKs. However, as seen in Fig. 6.8b, when eager and prescheduled packets coexist at intermediate loads, not retransmitting the eager packets that get dropped leads to loss of throughput. Finally, Fig. 6.9b plots the frequency of packet drops in the network.

When ACKs are conveyed out-of-band, the transition to prescheduled transmissions is ideal: the corresponding delay plot in Fig. 6.8a tracks the baseline delay up to the saturation point, and the prescheduled-only delay afterwards. With in-band ACKs, the performance is very similar. The only notable difference is the increased packet delay just before toggling to prescheduled transmissions, caused by the delay overheads in recovering occasionally dropped packets.

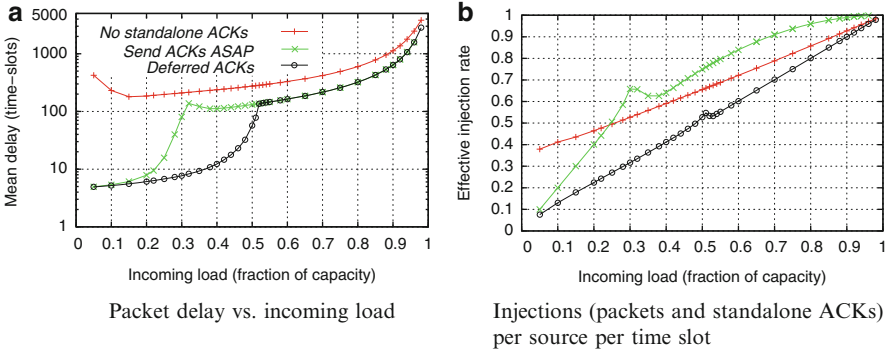


Fig. 6.10 Comparing standalone ACK generation strategies; $TH_{pre} = 24$, $N = 128$

6.4.2 ACK Overhead

Figure 6.10 compares the following ACK-forwarding strategies: (1) piggyback all ACKs (no standalone ACKs), (2) send standalone ACKs as soon as possible (ASAP), i.e. without using a timer, and (3) use deferred ACKs, using an adaptive timer to trigger sending of the standalone ACK if no piggybacking opportunity appeared in the meantime. Figure 6.10a shows that using only piggybacked ACKs is the worst option, because reverse packets offering piggybacking opportunities may not appear soon enough, causing sources to needlessly retransmit packets. In the load range from 0.05 to 0.15, delay actually *decreases* as load increases, because the increased contention is more than compensated by the increased rate of piggybacking and the resulting reduction in spurious retransmissions.

This is validated by Fig. 6.10b, which shows that if standalone ACKs are not allowed, sources inject packets at an effective rate that is much higher than the input load, indicating that multiple timeout-induced spurious retransmissions of the same packet may occur before the corresponding ACK is piggybacked. Effectively, because of the increased load, sources toggle to prescheduled transmissions.

When standalone ACKs are allowed, spurious retransmissions diminish, but the ACK overhead may induce a shift to the prescheduled mode. Figure 6.10b depicts this overhead. When generating ACK messages immediately, the effective injection rate can be nearly doubled. For loads above 0.25, this strategy causes the highest overall injection rate.

The scheme using deferred ACKs performs best by far. It drastically reduces the overhead due to standalone ACKs, while eliminating spurious retransmissions that can be caused by piggybacking delays. Using the proposed combination of eager and prescheduled injections, coupled with an optimized end-to-end retransmission and acknowledgment scheme, the performance of the low-cost Omega network can be enhanced dramatically.

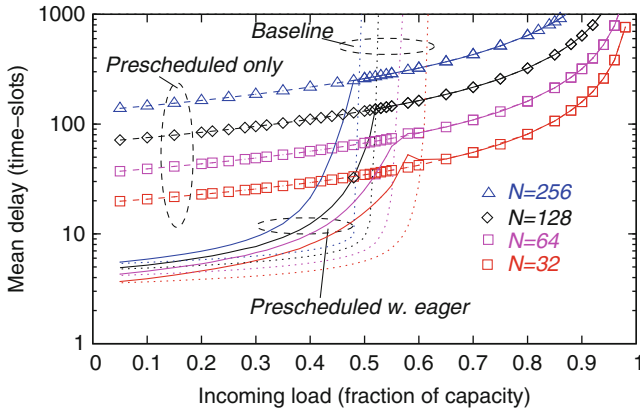


Fig. 6.11 Packet delay vs. incoming load for different network sizes

6.4.3 Different Network Sizes

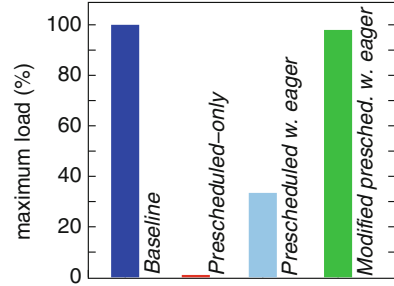
In Fig. 6.11 we plot the delay-throughput performance for different network sizes, ranging from 32×32 up to 256×256 . We used threshold values $TH_{pre} = 12, 24, 24,$ and 36 for $N = 32, 64, 128,$ and 256 , respectively. Additional results have shown that the optimal TH_{pre} increases with network size. Figure 6.11 demonstrates that for every network size the transition from eager to prescheduled injections is very close to the load point where the corresponding baseline network saturates, and thus scales well with N . Moreover, the absolute latency improvement increases with N .

We note here that although the mean packet latency is kept low, we have seen that some few packets undergo very large latencies, when they are dropped and have to be retransmitted. This holds true especially at the transition point when many eager coexist with many prescheduled packets. These large latencies must be attributed to the relatively large timeouts that we use in order to reduce the ACK overhead.

6.4.4 Permutation Traffic

Prescheduled injections perform well when traffic is uniform but are very inefficient when traffic is directed, as every connection is visited only once every N time slots. The speculative mode (eager injections) can correct this inefficiency. In the following experiment we set the window size equal to 512, and consider a permutation traffic pattern, in which each adapter i sends only to adapter $(i + 1) \bmod N$, at a varying load λ . Note that this permutation is monotonic and therefore routable through the Omega network.

Fig. 6.12 Maximum permissible load under permutation traffic for $N = 64$



In Fig. 6.12, we plot the maximum permissible load λ , i.e., the load for which the input backlogs remain bounded, for different system configurations. The baseline system achieves full throughput because the network experiences no contention with this monotonic permutation. On the other hand, the prescheduled-only system cannot sustain a load above $\frac{1}{N} \simeq 1.6\%$ for $N = 64$. The combination of prescheduled with eager can sustain a maximum load of around 33%.

Closer inspection revealed that the relatively low performance of prescheduled with eager is because of the injection policy. To sustain full line rate under this traffic pattern, the majority of injections have to exploit the eager mode. However, at full load, our injection policy uses the prescheduled mode once per N time slots. This should not be a problem, as payload data never conflict with each other, but in practice there is interference with the in-band ACK messages. In simulations, we observed cases where ACKs are dropped because of collisions with prescheduled packets, thus triggering retransmissions. Because the retransmissions use the prescheduled mode, they can overwrite additional ACKs, and hence induce more retransmissions, in a destructive positive-feedback loop. The backlogs that built up in this way bring about a sudden shift into the prescheduled mode for all loads above 33%.

The following modification to the injection policy remedies this performance issue. The basic idea is to avoid as many prescheduled injections as possible at low loads in order to prevent eager packets, and, most importantly for permutation traffic, eager acks, from being overwritten. As in Sect. 6.3.1, retransmissions always use the prescheduled mode and are given precedence. However, if a prescheduled retransmission is not possible, and $B < TH_{\text{pre}}$, then the adapter can inject only eager packets even if these come from prescheduled VOQs: in particular, the adapter first examines whether it can inject from the prescheduled VOQ in eager mode⁸; if the prescheduled VOQ is empty or not eligible, the adapter polls the remaining flows for an eager injection. As can be seen in Fig. 6.12, the modified injection policy sustains 98% of the line rate under this permutation traffic pattern. Additional results demonstrate that under uniform traffic it provides as good performance as our initial policy.

⁸We give precedence to the prescheduled flow, although the injected packet will be eager, to reduce the conflicts with prescheduled packets coming from other adapters.

6.5 Conclusions and Future Work

We have described the design of a high-throughput, low-latency, optical network. Its design was guided by the need to move complexity out of the network (optical switching nodes) and into the edges (electronic adapters) to enable an all-optical implementation of the switching nodes. We combine this simple, low-cost optical network with highly optimized end-to-end scheduling and transmission policies to meet the requirements for HPC and datacenter interconnects.

The proposed architecture is based on a minimum-cost banyan-class MIN using small-radix switching nodes, which employ shallow buffers and link-level flow control to perform distributed contention resolution, thereby eliminating the need for centralized control. Moreover, our approach exploits the property that an Omega network (and its topological equivalents) can route specific permutation patterns without conflicts to support prescheduled packet injections with 100% throughput, while at the same time reducing the latency penalty associated with prescheduling by allowing eager injections. Precedence rules govern the coexistence of eager and prescheduled traffic so that the former does not affect the latter, whereas an efficient end-to-end reliable delivery scheme deals with drops of eager packets.

We demonstrated that this approach combines low latency at low to medium utilization with high maximum throughput. The saturation throughput is about twice that of a buffered Omega network without prescheduled injections, whereas the low-load latency is about $N/2$ slots lower than for the same network with prescheduling but without eager injections. Thanks to speculation, we were also able to support directed traffic, comprising non-conflicting one-to-one connections, at line rate.

Additional insights that we collected while studying this drop-and-retransmit strategy highlight the potential benefits of using a separate medium for routing (out-of-band) the acknowledgments. If ACKs go in-band, as in this study, then timeouts must be large in order to reduce the ACK overhead, which is especially high in slotted systems. With such large timeouts, although we showed that the mean packet delay can be kept low, some packets may experience large delays if they have to be retransmitted. This may be undesired in delay-critical applications. The adaptive ACKs mechanism provides some means to keep the timeout short for connections that are intensive in both directions; in addition, one may also fix the maximum allowable timeout on a per connection basis. But in any case, a separate medium for ACKs would certainly perform best—it is a question of cost vs. performance.

So far, we have examined mostly uniform traffic, because an Omega network cannot efficiently handle non-uniform traffic. To cope with non-monotonic permutation patterns, a Batcher sorting network could be inserted in front of the Omega network, although this requires about $\frac{1}{2} \log^2(N)$ additional stages. Performance for non-uniform traffic could also be enhanced by adopting multi-path topology such as a Beneš network or a load-balanced Birkhoff-von Neumann switch.

Acknowledgements This research was supported by the European Union FP7-ICT project HISTORIC (“Heterogeneous InP on Silicon Technology for Optical Routing and logIC”) under contract no. 223876. The authors would like to thank Anne-Marie Cromack and Charlotte Bolliger for their help in preparing this manuscript.

References

1. Beldianu S, Rojas-Cessa R, Oki E, Zivara S (2009) Re-configurable parallel match evaluators applied to scheduling schemes for input-queued packet switches. In: Proceedings of IEEE ICCCN, San Francisco, CA, USA
2. Blumenthal DJ et al (2011) Integrated photonics for low-power packet networking. *IEEE J Sel Top Quant Electron* 17(2):458–471
3. Burmeister EF, Blumenthal DJ, Bowers JE (2008) A comparison of optical buffering technologies. *Optical Switching and Networking* 5(1):10–18
4. Chang CS, Lee DS, Jou YS (2002) Load-balanced Birkhoff-von Neumann switches, part I: one-stage buffering. *Comp Comm* 25(6):611–622
5. Chao HJ, Jing Z, Liew SY (2003) Matching algorithms for three-stage bufferless cros network switches. *IEEE Comm Mag* 41:46–54 (2003)
6. Chrysos N, Katevenis M (2006) Scheduling in non-blocking, buffered, three-stage switching fabrics. In: Proceedings of IEEE INFOCOM, Barcelona, Spain
7. Dias D, Jump JR (1981) Analysis and simulation of buffered delta networks. *IEEE Trans Comput* C-30(4):273–282
8. Germann R, Salemink HWM, Beyeler R, Bona GL, Horst F, Massarek I, Offrein BJ (2000) Silicon oxynitride layers for optical waveguide applications. *J Electrochem Soc* 147(6):2237–2241
9. Goke LR, Lipovski GJ (1973) Banyan networks for partitioning multiprocessor systems. In: Proceedings of ACM ISCA, New York, NY, USA, pp 21–28
10. Hui JH (1990) Switching and traffic theory for integrated broadband networks. Kluwer, Dordrecht
11. Iliadis I, Minkenberg C (2008) Performance of a speculative transmission scheme for arbitration latency reduction. *IEEE/ACM Trans Comp* 16(1):182–195
12. Iliadis I, Chrysos N, Minkenberg C (2007) Performance evaluation of the data vortex photonic switch. *IEEE J Sel Areas Comm* 25(S-6):20–35
13. Keslassy I, Chuang ST, Yu K, Miller D, Horowitz M, Solgaard O, McKeown N (2003) Scaling internet routers using optics. In: Proceedings of ACM SIGCOMM, ACM, Karlsruhe, pp 189–200
14. Liu J, Hung CK, Hamdi M, Tsui CY (2002) Stable round-robin scheduling algorithms for high-performance input queued switches. In: Proceedings of IEEE hot-interconnects (HOTI 2002), San Francisco, CA
15. Luijten, RP, Minkenberg C, Hemenway BR, Sauer M, Grzybowski R (2005) Viable optoelectronic HPC interconnect fabrics. In: Proceedings of supercomputing (SC). IEEE Computer Society, Washington, DC (2005)
16. Murdocca M (1989) Optical design of a digital switch. *Appl Opt* 28(13):2505–2517
17. Papadimitriou GI, Papazoglou C, Pomportsis AS (2003) Optical switching: switch fabrics, techniques, and architectures. *J Lightwave Technol* 21(2):384–405
18. Petracca M, Lee BG, Bergman K, Carloni LP (2009) Photonic nocs: system-level design exploration. *IEEE Micro* 29(4), 74–85 (2009)
19. Pun K, Hamdi M (2002) Distro: A distributed static round-robin scheduling algorithm for bufferless cros-network switches. In: Proceedings of IEEE GLOBECOM, Taipei, Taiwan, pp 2298–2302
20. Saha A, Wagh M (1990) Performance analysis of banyan networks based on buffers of various sizes. In: IEEE INFOCOM, San Francisco, CA, pp. 157–164

21. Scicchitano A, Bianco A, Giaccone P, Leonardi E, Schiattarella E (2007) Distributed scheduling in input queued switches. In: Proceedings of IEEE ICC, Glasgow, UK
22. Shacham A, Small BA, Liboiron-Ladouceur O, Bergman K (2005) A fully implemented 12 x 12 data vortex optical packet switching interconnection network. *J Lightwave Technol* 23(10):3066
23. Takagi H (1993) Queueing analysis. In: Discrete-time systems: a foundation of performance evaluation. Elsevier, Amsterdam
24. Tanenbaum AS (2002) Computer networks, 4th edn. Prentice Hall, NJ

Chapter 7

A High-Speed MIMO OFDM Flexible Bandwidth Data Center Network

Philip N. Ji, D. Qian, K. Kanonakis, Christoforos Kachris, and Ioannis Tomkos

7.1 Introduction

As the global Internet traffic is growing exponentially, the data centers, which host many Internet application servers, are also facing rapid increase in bandwidth demands. Due to emerging applications like cloud computing, next-generation data centers need to achieve low latency, high throughput, high flexibility, high re-source efficiency, low power consumption, and low cost. Furthermore, as more and more processing cores are integrated into a single chip, the communication requirements between racks in the data centers will keep increasing significantly. By integrating hundreds of cores into the same chip (e.g., Single-chip Cloud Computer-SCC [17]) we can achieve higher processing power in the data center racks. However these cores require a fast and low-latency interconnection scheme to communicate with the storage system and the other servers inside or outside of the rack.

Optical technology has been adopted in data center networks (DCN) due to its high bandwidth capacity. However, it is mainly used for point-to-point link, while the intra Data Center Network (DCN) interconnect is still based on electrical switching fabric, which has high power consumption and limited bandwidth capacity [13]. Currently, the power consumption of the data center networks accounts for 23% of the total IT power consumption [21]. However, due to the high communication requirements of the future data center networks, it is estimated that the data center networks will account for much higher percentages of the overall power

P.N. Ji (✉) • D. Qian

NEC Laboratories America, Inc., 4 Independence Way, Princeton, NJ 08540, USA

e-mail: pji@nec-labs.com; dqian@nec-labs.com

K. Kanonakis • Ch. Kachris • I. Tomkos

Athens Information Technology, Athens, Greece

e-mail: kkan@ait.edu.gr; kachris@ait.edu.gr; itom@ait.edu.gr

consumption [4]. Therefore it is expected that data center network may evolve to all-optical networks, similar to the telecommunication networks that have been evolved from opaque to transparent networks using all-optical switches.

In recent years, several hybrid optical/electrical or all-optical interconnect schemes for DCN have been proposed [9, 16, 19, 22, 23, 25]. Many of them rely on large-scale fiber cross-connect (FXC) [9, 19, 25] or multiple wavelength-selective switches (WSS) ([16], which are costly and have slow switching rate (at millisecond level). Having a large-scale FXC also present an undesirable single source-of-failure. A recent work in [23] uses silicon electro-optic mirroring WSS and semiconductor optical amplifier-based switch to achieve nanosecond scale switching, making all-optical packet level routing possible. However the key components are not commercially available and have low scalability. Other architectures use tunable wavelength converters [22, 25]. They are also costly and do not allow bandwidth resource sharing among the connections. Some of them also require electrical or optical buffer.

In this chapter we propose and experimentally demonstrate a novel all-optical DCN architecture that combines a passive cyclic arrayed waveguide grating (CAWG) core router with orthogonal frequency division multiplexing (OFDM) modulation and parallel signal detection (PSD) technologies. The architecture achieves fast (nanosecond speed), low latency, low power consumption multiple-input multiple-output (MIMO) switching while allowing fine granularity bandwidth sharing without requiring FXC, WSS, or tunable wavelength converter.

7.2 MIMO OFDM Flexible Bandwidth DCN Architecture

7.2.1 MIMO OFDM

A key technology for this DCN architecture is MIMO OFDM. OFDM is a modulation technology to achieve high spectral efficiency transmission by parallel transmission of spectrally overlapped, lower rate frequency-domain tributaries where the signals are mathematically orthogonal over one symbol period (Fig. 7.1). Originally applied for copper and wireless communications, OFDM technology has been adopted in optical communication network applications in the past few years as the high-speed digital signal processing and broadband DAC/ADC became feasible [1, 20]. Because of the advantages such as better tolerance to fiber dispersions and the ability to perform one-tap equalization at the frequency domain, OFDM has been demonstrated to be a good candidate for long distance transmission [2, 6, 14]. OFDM technology has also been proposed for optical access network to take advantage of its flexibility to share the spectrum among multiple users, such as in the application of OFDMA-PON (orthogonal frequency division multiple access passive optical network) [7]. The feasibility of using OFDM technology for data center application has been discussed in [3], but no actual network architecture has been proposed for the intra-data center network.

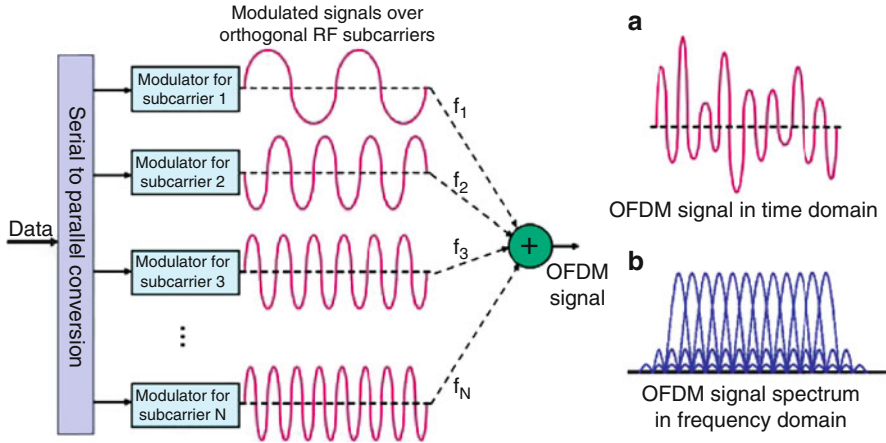


Fig. 7.1 OFDM signal generation

There are mainly two types of OFDM implementation in optical transmission. The first type is to generate the OFDM signal electrically and modulate the signal to an optical carrier [2, 6, 14]. This is referred to as the optical OFDM (O-OFDM). The receiver can use direct detection or coherent detection techniques. The second type is to generate the orthogonal subcarriers (or called tones) optically and then apply signal onto each subcarrier [10, 11]. This is called the all-optical OFDM (AO-OFDM).

The DCN architecture proposed in this chapter is based on O-OFDM implementation. It has network level MIMO operation because each rack can send OFDM signal to multiple destination racks simultaneously, and multiple racks can send the signal the same destination rack at the same time by modulation data on different OFDM subcarriers in the RF domain.

At each receiver, a common photo-detector (PD) can simultaneously detect multiple O-OFDM signals from many sources with different optical wavelengths, provided that there is no contention in OFDM subcarriers and the WDM wavelengths. This is referred to as the PSD technology [15] and has been demonstrated in OFDM WDM-based optical networks [12].

7.2.2 Cyclic Arrayed Waveguide Grating

The key optical component for the proposed DCN architecture is a CAWG. An $N \times N$ CAWG (also called an AWG router or a cyclic interleaver) is a passive optical multiplexer/demultiplexer that routes different wavelengths from N different input ports to N different output ports in a cyclic manner. Figure 7.2 illustrates the cyclic wavelength arrangement of an 8×8 CAWG. CAWG is usually constructed

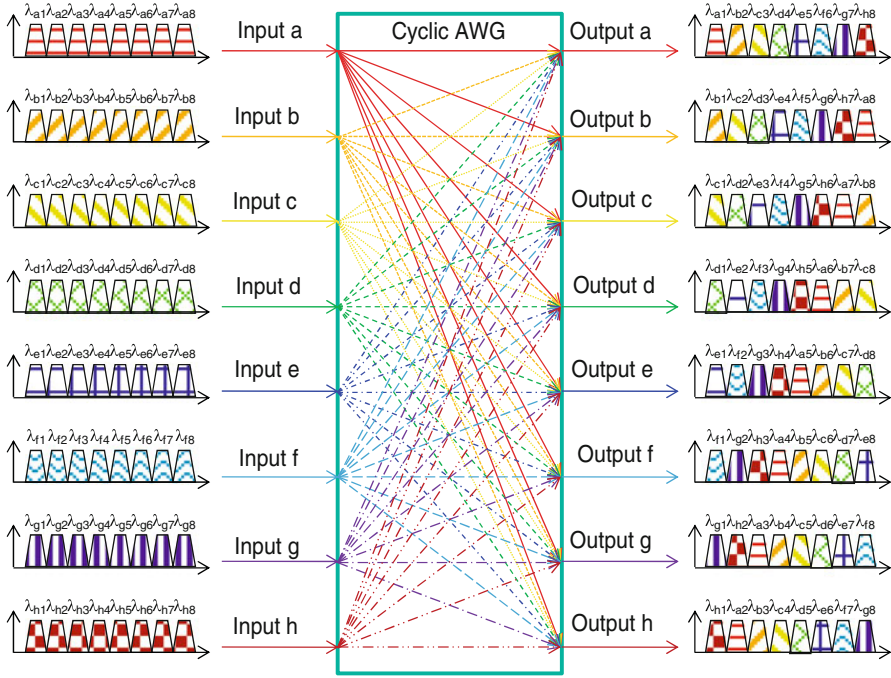


Fig. 7.2 Cyclic wavelength routing of a CAWG

using planar lightwave circuit technology. The cyclic wavelength arrangement characteristic avoids the wavelength contention and eliminates the need for large-scale core FXC or multiple WSS units. Several other DCN architectures also use CAWG as the core optical router [22, 25].

7.2.3 DCN Architecture

The schematic of the MIMO OFDM DCN architecture is illustrated in Fig. 7.3. It contains N racks, each accommodating multiple servers connected to a top-of-the-rack switch (ToR). Inter-rack communications are performed by inter-connecting these ToRs through the DCN.

The inter-rack signals at each rack are aggregated and sent to a transmitter, which contains an OFDM modulator that modulates the aggregated signals into K OFDM data streams with appropriate subcarrier assignments, where K is the number of destination racks that the signals from this source rack need to travel to, so $0 \leq K \leq N$. Different racks can have different K numbers. These OFDM data streams are converted to K WDM optical signals through an array of K directly modulation lasers (DMLs) or K sets of laser/modulator with different wavelengths.

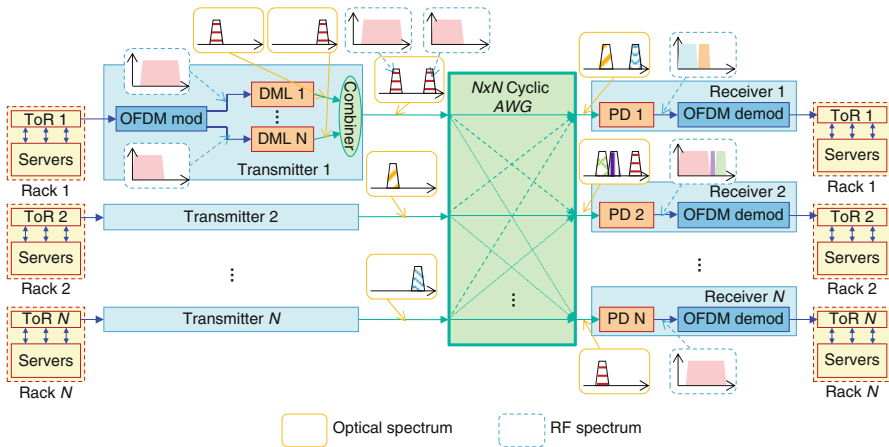


Fig. 7.3 Architecture of the MIMO OFDM DCN

If these lasers are the fixed wavelength type, N units will be needed since the signal from each ToR might be switched to any destination rack potentially.

If the number of the racks increases, it is not cost efficient to install N lasers at each transmitter, this is also not necessary because it is not likely that each rack needs to communicate with all other racks simultaneously. Therefore the N fixed wavelength lasers in the transmitter can be replaced with fewer tunable lasers.

These O-OFDM signals are then combined through a WDM combiner to form an OFDM-modulated WDM signal and sent to an $N \times N$ CAWG. Due to the cyclic non-blocking wavelength arrangement of the CAWG, the WDM channels are routed to the respective output ports for the destination racks. Each optical receiver receives one WDM channel from each input port. Through a centralized OFDM subcarrier allocation scheme, the WDM channels at each receiver do not have subcarrier contention, so that a single PD can receive all WDM channels simultaneously through the PSD technology. The received OFDM signal is then demodulated back to the original data format and sent to the appropriate servers through the destination ToR.

When a new switching state is required, the OFDM modulators execute the new subcarrier assignments determined by the centralized controller, and the respective lasers turns on and off to generate new OFDM WDM signals. Some servers in the DCN have constant large volume communication with other servers. It will be less efficient for them to go though the ToR before exiting the rack. In some cases, the large traffic volume from these servers might even congest the ToR. To serve these “super servers” more effectively, the MIMO OFDM DCN architecture can be extended to reserve dedicated OFDM WDM transmitters and dedicated CAWG ports for them. These servers can bypass the ToR and connect to the transmitters directly, as illustrated in Fig. 7.4.

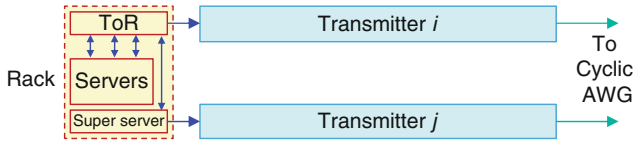


Fig. 7.4 ToR bypass for “super servers”

7.2.4 DCN Architecture Features

Comparing to other optical or hybrid DCN architectures proposed so far, this MIMO OFDM DCN architecture offers the following advantages:

MIMO switching: Conventional optical DCN architectures use optical circuit switching. Each rack can only talk to one rack at a time. It needs to wait for the current connection to complete before another connection can be established for the same rack. Due to the MIMO OFDM operation in this architecture, each rack can communicate with multiple racks simultaneously. Thus the waiting time is eliminated and high interconnect efficiency can be achieved.

Flexible bandwidth allocation and sharing: By flexibly selecting the number of subcarrier at each O-OFDM transmitter and splitting and sharing the total available subcarriers at each receiver among multiple sources, this architecture allows different bandwidth allocations for different source-destination pairs at the same time. And this allocation can be dynamically changed over time. Such feature is suitable for DCN application where there are frequent setting up and tearing down of connections and large fluctuation in bandwidth demands.

Fine granularity switching: Since the O-OFDM signal is generated electrically, the switching granularity is much finer than the current optical DCN technologies. For example, in direct optical point-to-point link, the granularity is one fiber; in regular WDM system, the granularity is one WDM channel, which typically carries 10 Gb/s to 40 Gb/s or 100 Gb/s data; in the AO-OFDM system, the granularity is one optically generated OFDM subcarrier, which is typically 10 Gb/s or higher. The switching granularity in the O-OFDM system is one electrically generated OFDM subcarrier, which is typically in the order of tens of Mb/s or less. Having finer granularity allows more flexible bandwidth allocation and more efficient spectrum utilization.

Flexible modulation format and data rate: OFDM modulation also provides the capability to change the modulation order to adjust the amount of data to be carried within the same subcarrier (or group of subcarriers). For example, the OFDM signal in each subcarrier can be modulated using BPSK, or QPSK, or 16QAM, or 64QAM, etc. This allows variable amount of data to be packed within the same subcarrier as these modulation formats encode different number of data bits in each symbol.

This can be used to solve the congestion issue at the destination racks. Within the same OFDM signal, different modulation formats can coexist, different subcarriers can use different modulation formats.

No guard band: In PSD-based OFDM system, guard bands are usually required between subcarrier groups from different sources due to the synchronization difficulty and impairments during transmission such as dispersion and OSNR degradation. However, such guard band is not required for the DCN application because the transmission distance is short (typically from tens of meters up to 2 km). This allows the maximum bandwidth utilization at each receiver.

Fast switching: This architecture performs optical circuit routing by turning respective lasers on and off. This can be achieved at sub-ns level, making this system feasible for packet level operation. If tunable lasers are used, the switching speed will then be determined by both the laser on/off speed and tuning speed, which can also be realized at ns level [8].

Low and uniform latency: All switched signals take exactly one hop (i.e., passing through the core optical router only once); therefore, the latency is very low and uniform. Furthermore, because this architecture uses MIMO operation and bandwidth sharing by OFDM subcarrier allocation, no optical or electrical buffer is required. Thus the latency can remain at low level.

Scalable: The key factor that determines the available scale of this DCN is the port count of the CAWG. A 400-channel AWG based on 6-inch silica waver and a 512-channel AWG based on 4-in silica wafer have been demonstrated more than a decade ago [18]. With the recent advancement in silicon photonics technology, even higher port count CAWG can be expected because silicon waveguide can achieve higher core/cladding index contrast and thus allow waveguide bending radius to be several order of magnitude lower than silica waveguide [24].

Simple control: While having the subcarrier contention restriction at each receiver, this architecture allows same OFDM subcarriers to be used by different OFDM signals generated from the same transmitter. Therefore the subcarrier allocation at each receiver can be considered independently, and thus greatly reduce the complexity of the subcarrier assignment problem compared to other bandwidth sharing networks.

Low power consumption: Because the core optical router in this architecture is completely passive and static, the optical components have lower power consumption compared with other optical DCN architectures that require switching through WSS or FXC. The heat dissipation is also lower.

Low cost: Since this architecture does not require FXC, WSS or tunable wavelength converter, the optical component cost is low. Having low power consumption and low heat dissipation also reduces the cooling requirement and the operation cost.

7.3 Experimental Demonstration of Flexible MIMO OFDM Optical Interconnect

The flexible MIMO interconnect capability of the proposed architecture is demonstrated on a lab testbed. The optical core router is an 8×8 CAWG with 100 GHz spacing. Each transmitter contains two tunable external cavity lasers with 10 GHz intensity modulators. The OFDM signals are generated by an arbitrary waveform generator. The signal has 1,200 subcarriers, each occupying 5 MHz bandwidth. Two scenarios with different wavelength and subcarrier assignments are tested (Table 7.1). Due to the cyclic wavelength routing arrangement in the CAWG, λ_3 and λ_6 from ToR1 are routed to ToR3 and ToR 6, respectively, while λ_4 and λ_7 from ToR2 are routed to ToR3 and ToR 6, respectively.

Different modulation formats, including QPSK, 16QAM, and 64QAM, are used for different OFDM subcarrier groups highlighted in black, producing per-subcarrier data rates of 10 Mb/s, 20 Mb/s, and 30 Mb/s, respectively. A 10 GHz bandwidth single-ended direct detection PD is used at each CAWG output to receive the WDM signal and convert it to electrical OFDM signal. This OFDM signal is then captured and digitized by a real-time oscilloscope and processed using an offline computer to recover the data from the OFDM signal. No optical amplification is used in this experiment.

The RF spectra of the OFDM signals are measured at each OFDM transmitter output each the receiver output for different tests (Fig.7.5). They show that different OFDM signals from the same transmitter can possess overlapping (such as subcarrier group A and subcarrier group B in Test 1) or non-contiguous subcarriers (such as subcarrier group E), and through the PSD technology each receiver can successfully detect and receive OFDM signals from multiple sources, provided that these OFDM subcarriers do not overlap. No guard band is required when

Table 7.1 Wavelength and subcarrier assignments in MIMO OFDM DCN experiment

Test	From ToR 1		From ToR 2	
	λ_3 193.5 THz 1549.32 nm (To ToR 3)	λ_6 193.2 THz 1551.72nm (To ToR 6)	λ_4 193.4 THz 1550.12 nm (To ToR 3)	λ_7 193.1 THz 1552.52 nm (To ToR 6)
1	SC 100-450 QPSK A	SC 300-450 64QAM C	SC 451-800 QPSK D	SC 850-950 QPSK F
2	SC 200-600 16QAM B		SC 50-150, 700-900 QPSK E	

SC: subcarriers

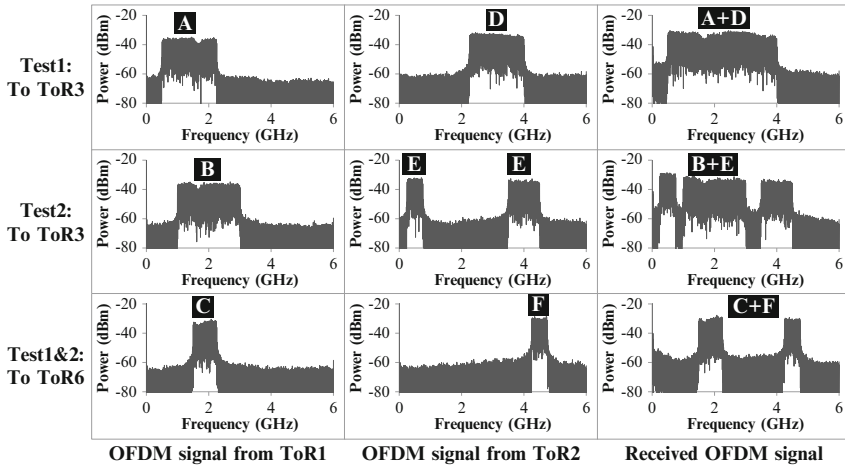


Fig. 7.5 Measured RF spectra of the OFDM signals at the transmitters and the receivers

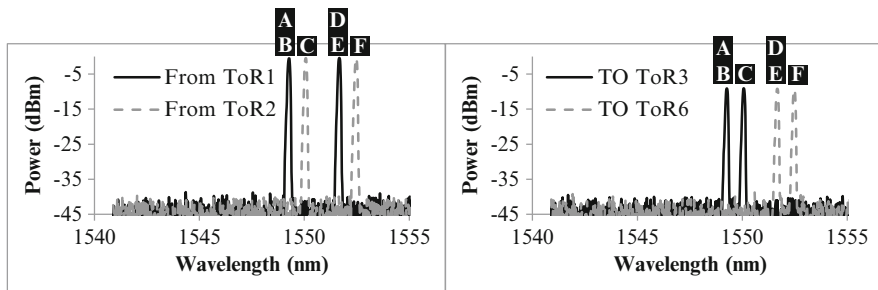


Fig. 7.6 Measured optical spectra at the inputs (top) and outputs (bottom) of the CAWG

assigning the OFDM subcarriers, as shown at the receiver of ToR 3 in Test 1. Besides allocating subcarriers, the centralized controller also balances the subcarrier powers among signals from different inputs.

The optical spectra of the WDM signals at the CAWG input ports from ToR 1 and ToR 2 transmitters and CAWG output ports to ToR 3 and ToR 6 receivers are shown on Fig. 7.6. They confirm the non-blocking cyclic routing operation of the CAWG.

The PSD receiver performance (solid symbols), represented by the bit error rates (BER) under different received optical power levels, are measured and compared with single channel receiver (hollow symbols) for different OFDM signals at each test (Fig. 7.7). While the absolute BER value varies with the modulation format and per-subcarrier power level, no significant degradation between single channel detection and PSD detection is observed in any tests, and OFDM signals with different modulation formats can all be detected successfully. This shows that PSD

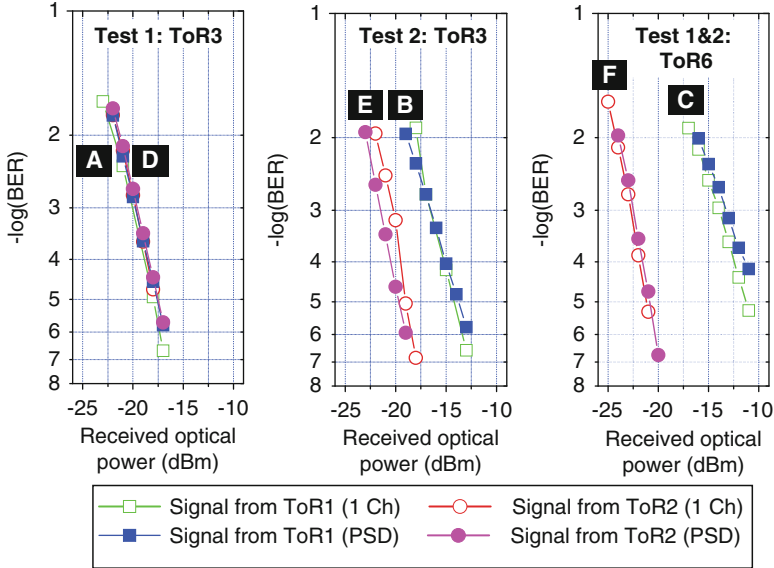


Fig. 7.7 Measured received signal performance

technology is feasible for receiving multiple OFDM WDM signals simultaneously using a single PD, and thus realizing MIMO switching.

7.4 Performance Evaluation

7.4.1 Simulation Model and Traffic Assumptions

In order to evaluate the performance of the proposed architecture, a custom simulation model was built using OPNET Modeler. The model simulated an OFDMA-based switch operating as described above, consisting of 8 ToRs. Each ToR receiver used 1,000 subcarriers, each offering a bit rate of 10 Mbps, hence the total ToR-ToR capacity was 10 Gbps. Traffic at each ToR was destined with equal probability to all other ToRs. Regarding traffic characteristics, studies performed on real data center networks [5] have indicated that packet arrivals are best modeled using long-tail distributions (e.g., Log-normal, Weibull). In the same work, it was

found that packet sizes approximated a trimodal distribution, with sizes of 64, 200, and 1,400 bytes appearing with probabilities of 0.05, 0.5, and 0.45, respectively. In that respect, for our simulation study traffic was simulated for each ToR pair by an ON-OFF source producing packets of the aforementioned sizes. Both the ON and OFF period durations, as well as the packet interarrivals during ON periods, followed the Log-normal distribution with a shape of 0.75. Different degrees of burstiness (characterized by a different average OFF/ON ratio) were tested. Finally, a buffer of 10 MB was dedicated to each source-destination ToR pair.

7.4.2 Subcarrier Allocation Algorithms

Several subcarrier allocation algorithms were considered to exploit the proposed architecture:

1. *Optimal Resource Utilization (ORU)*: Arriving packets from any ToR towards a specific destination ToR are transmitted in FIFO order using all available subcarriers each time. In other words this is in essence a time-based (rather than subcarrier-based) scheduling approach, employing OFDM technology for the transmission of each individual packet. It is assumed that time is continuous (i.e., no timeslots) and there are no guardbands, so that no bandwidth resources are wasted. Therefore, this scheme is expected to offer the minimum average packet delay. Note though that it is not realistic, since ToRs have to switch between states very fast, i.e. at the timescales of packet interarrivals (few μ s). Since this scheme is not feasible (and included here just for benchmarking purposes), below we describe ways of relaxing switching time requirements and at the same time exploiting the subcarrier domain for bandwidth arbitration.
2. *Fixed subcarrier allocation (FSA)*: In FSA, virtual transmission pipes of fixed bandwidth are created for each ToR pair by dedicating to each of them a number of subcarriers. Since in this work we considered uniform average traffic, each ToR pair was assigned 125 subcarriers (i.e., 1,000/8). Although this is a very simple solution, it cannot adapt to bursty traffic (which is the case in a real data center network) and is hence expected to lead to increased packet delay.
3. *Dynamic Subcarrier Allocation (DSA)*: We propose this scheme as a feasible approach to achieve assignment of subcarriers for each ToR pair according to the actual traffic needs. In that respect, DSA is executed periodically (the scheduling period is denoted as T) and tries to adjust the current assignment based on traffic information collected from all ToRs. More specifically, a ratio (denoted as f) of the available subcarriers is distributed equally to all ToR pairs in a fixed manner, while the rest are assigned each time in a weighted fashion, with the weights decided according to the traffic rate in each of them during the previous scheduling period.

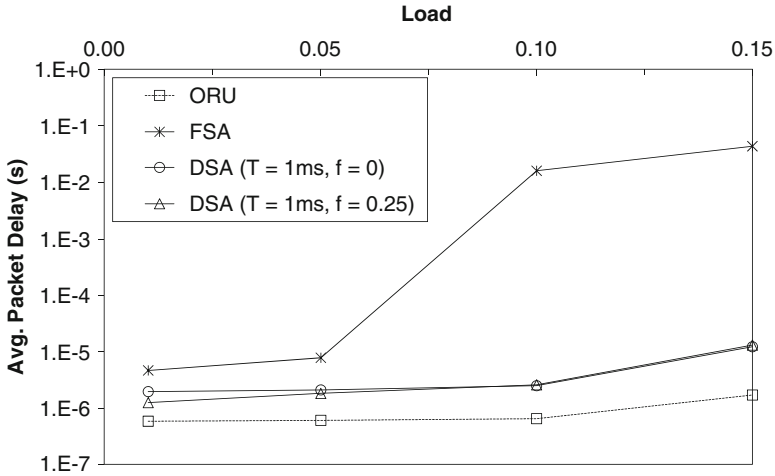


Fig. 7.8 Average packet delay comparison when OFF/ON = 10

Note that T should be selected carefully, since, on the one hand, it should be relatively short to avoid affecting QoS (i.e., inaccurate decisions should be corrected quickly) and, on the other hand, a very short T would impose challenges in the relevant electronics and could produce inaccurate measurements (depending on the exact traffic characteristics). As a final note, it should be mentioned that FSA can also be considered as a subset of DSA, with f equal to 1 and $T = \infty$ (since it is executed only once in the beginning).

7.4.3 Simulation Results

The performance of the proposed algorithms was evaluated for different traffic profiles and network loading conditions. More specifically, two general scenarios were considered; the first with the sources having an OFF/ON ratio of 10 and the second having OFF/ON = 50. The load values shown in the results indicate the average aggregate load as a ratio of the total available switch capacity.

Figure 7.8 shows the average packet delay when OFF/ON = 10. First of all it is clear that, as expected, ORU and FSA offer the best and worst performance, respectively. In particular, FSA proves to be completely inadequate, since it exceeds ORU by several orders of magnitude even at moderate loads. At the same time, the use of DSA (a value of 1 ms was used in these simulation experiments as a compromise, taking into account the issues listed above) offers delay values that are quite close to ORU. Moreover, the use of a hybrid FSA/DSA approach by means of a nonzero f value (0.25 here) seems to improve performance, though not significantly.

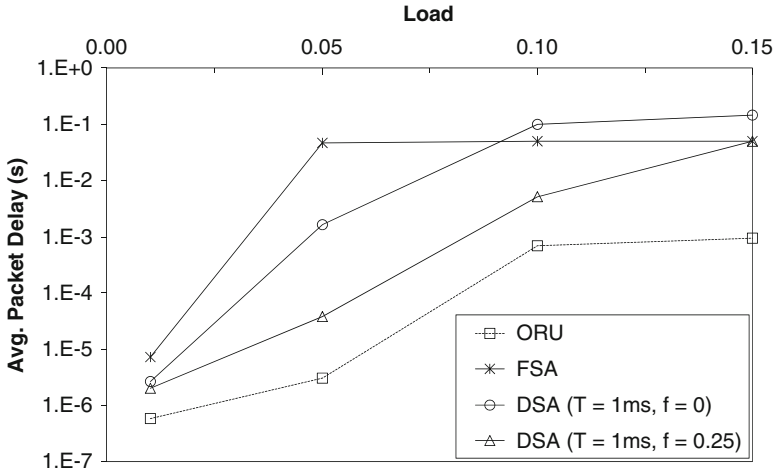


Fig. 7.9 Average packet delay comparison when OFF/ON = 50

Figure 7.9 depicts a comparison of the average packet delay performance for the same algorithms, however this time under more challenging traffic conditions (burstiness is increased to OFF/ON = 50). It is evident that the delay values for all algorithms are increased due to the more bursty traffic profile, while the conclusions drawn above for ORU and FSA still hold. Note also that the convergence of delay for FSA for higher loads (i.e., after 0.05) is only due to extensive buffer overflows, while the packet loss in those load regions exceeded 50% (the maximum loss observed for the rest of the algorithms was more than one order of magnitude lower). Therefore, comparisons with FSA are pointless and it should not be considered as a potential solution here. The delay performance of DSA is again between FSA and ORU. However, it is very interesting to point out that this time the hybrid DSA ($f = 0.25$) managed to reduce delay up to two orders of magnitude compared to the pure DSA ($f = 0$).

7.5 Conclusions

We propose a novel DCN architecture utilizing OFDM and PSD technologies. This architecture offers high switching speed, low and uniform latency, and low power consumption. We experimentally demonstrate the MIMO OFDM switching and fine granularity flexible bandwidth sharing features. We also develop efficient subcarrier allocation algorithms, which achieves high spectrum utilization at low computation complexity. Therefore this architecture is suitable for all-optical inter-rack and inter-server communication in next-generation DCN application.

References

1. Armstrong J (2008), OFDM: From Copper and Wireless to Optical, in Optical Fiber Communication Conference and Exposition and The National Fiber Optic Engineers Conference, OSA Technical Digest (CD) (Optical Society of America, 2008), paper OMM1.
2. Armstrong J (2009) OFDM for optical communications. *J. Lightwave Technol.* 27(3):189–204
3. Benlachar Y, Bouziane R, Killey RI, Berger CR, Milder P, Koutsoyannis R, Hoe JC, Pschel M, Glick M (2010) Optical OFDM for the data center. In: 12th International Conference on Transparent Optical Networks (ICTON), pp. 1–4, London, UK
4. 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.
5. 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
6. Djordjevic IB, Vasic B (2006) Orthogonal frequency division multiplexing for high-speed optical transmission. *Opt. Express* 14(9):3767–3775
7. Dual-polarization 2x2 IFFT/FFT optical signal processing for 100-Gb/s QPSK-PDM all-optical OFDM, May 2009
8. Engelsteadter JO, Roycroft B, Peters FH, Corbett B (2010) Fast wavelength switching in interleaved rear reflector laser. In: International Conference on Indium Phosphide & Related Materials (IPRM), pp. 1–3, Cork, Ireland
9. Farrington N, Porter G, Radhakrishnan S, Bazzaz HH, Subramanya V, Fainman Y, Papen G, Vahdat A (2010) Helios: a hybrid electrical/optical switch architecture for modular data centers. In: Proceedings of the ACM SIGCOMM 2010. ACM, New York, pp 339–350
10. Hillerkuss D, Schmogrow R, Schellinger T, Jordan M, Winter M, Huber G, Vallaitis T, Bonk R, Kleinow P, Frey F, Roeger M, Koenig S, Ludwig A, Marculescu A, Li J, Hoh M, Dreschmann M, Meyer J, Ben Ezra S, Narkiss N, Nebendahl B, Parmigiani F, Petropoulos P, Resan B, Oehler A, Weingarten K, Ellermeyer T, Lutz J, Moeller M, Huebner M, Becker J, Koos C, Freude W, Leuthold J (2011) 26 tbit s⁻¹ line-rate super-channel transmission utilizing all-optical fast Fourier transform processing. *Nat Photonics* 5(6):364–371, Geneva, Switzerland
11. Huang YK, Qian D, Saperstein RE, Ji PN, Cvijetic N, Xu L, Wang T (2009) Dual-polarization 22 IFFT/FFT optical signal processing for 100-Gb/s QPSK-PDM all-optical OFDM. In: Optical fiber communication conference and exposition and the national fiber optic engineers conference. Optical Society of America, San Diego, CA, USA, p OTuM4
12. Ji PN, Patel AN, Qian D, Jue JP, Hu J, Aono Y, Wang T (2011) Optical layer traffic grooming in flexible optical WDM (FWDM) networks. In: 37th European conference and exposition on optical communications. Optical Society of America, p We.10.P1.102
13. Kachris C, Tomkos I, A Survey on Optical Interconnects for Data Centers, *IEEE Communications Surveys and Tutorials*, doi:10.1109/SURV.2011.122111.00069
14. Lowery AJ, Du L, Armstrong J (2006) Orthogonal frequency division multiplexing for adaptive dispersion compensation in long haul wdm systems. In: Optical fiber communication conference and exposition and the national fiber optic engineers conference. Optical Society of America, p PDP39, Anaheim, CA, USA
15. Luo Y, Yu J, Hu J, Xu L, Ji PN, Wang T, Cvijetic M (2007) WDM passive optical network with parallel signal detection for video and data delivery. In: Optical fiber communication conference and exposition and the national fiber optic engineers conference. Optical Society of America, p OWS6, Anaheim, CA, USA
16. Singla A, Singh A, Ramachandran K, Xu L, Zhang Y (2010) Proteus: a topology malleable data center network. In: Proceedings of the ninth ACM SIGCOMM workshop on hot topics in networks, Hotnets '10. ACM, New York, pp 8:1–8:6
17. Single Chip Cloud Computing (SCC) Platform Overview. Intel White paper, 2011

18. Takada K, Abe M, Shibata M, Ishii M, Okamoto K (2001) Low-crosstalk 10-ghz-spaced 512 channel arrayed-waveguide grating multi/demultiplexer fabricated on a 4-in wafer, *IEEE Photonics Technology Letters*, 13(11):1182–1184
19. Wang G, Andersen DG, Kaminsky M, Papagiannaki K, Ng TE, Kozuch M, Ryan M (2010) c-through: Part-time optics in data centers. In: *Proceedings of the ACM SIGCOMM 2010 conference on SIGCOMM, SIGCOMM '10*. ACM, New York, pp 327–338
20. Weinstein SB (2009) The history of orthogonal frequency-division multiplexing. *Comm. Mag.* 47(11):26–35
21. Where does power go? GreenDataProject (2008). Available online at: <http://www.greendataproject.org>. Accessed date March 2012
22. Xia K, Kaob Y-H, Yangb M, Chao HJ (2010) Petabit optical switch for data center networks. Technical report, Polytechnic Institute of NYU
23. Xu L, Zhang W, Lira HLR, Lipson M, Bergman K (2011) A hybrid optical packet and wavelength selective switching platform for high-performance data center networks. *Opt Express* 19(24):24258–24267
24. Yamada H, Chu TCT, Ishida S, Arakawa Y (2006) Si photonic wire waveguide devices. In: *IEEE Journal of Selected Topics in Quantum Electronics*, 12(6):1371–1379
25. Ye X, Yin Y, Yoo SJB, Mejia P, Proietti R, Akella V (2010) DOS: a scalable optical switch for datacenters. In: *Proceedings of the 6th ACM/IEEE symposium on architectures for networking and communications systems, ANCS '10*. ACM, New York, pp 24:1–24:12

Chapter 8

A Petabit Bufferless Optical Switch for Data Center Networks

Kang Xi, Yu-Hsiang Kao, and H. Jonathan Chao

8.1 Introduction

Data centers are critical infrastructures of the Internet, providing data- and computing-intensive services for various types of applications. Data centers are the only platform that can support large-scale cloud computing applications, such as Microsoft Azure, Amazon Elastic Compute Cloud (EC2), Google search, and Facebook. With the rapid growth of Internet applications, data centers have witnessed growing demands for storage, computation power, and communication bandwidth. Today it is not uncommon for a data center to house tens of thousands of servers in a single facility. For instance, it is reported that a Google data center holds more than 45,000 servers [35]. Despite the large size, data centers keep growing at an exponential rate [19]. The short-term target is to host hundreds of thousands of servers in a single data center. For instance, Microsoft is building a data center that has the capacity for up to 300,000 servers [34]. Although Google does not reveal the total number of servers in its data centers, its vision goes up to ten million servers worldwide according to the design objectives of its new storage and computation system called Spanner [14].

Modern data centers are different from traditional computer clusters in that a data center is not a simple collection of servers running many small, independent jobs. Instead, the servers work in collaborative ways to solve large-scale problems. This type of computing often requires extensive data exchange inside data centers. For instance, in a search using MapReduce, jobs are dispatched to many servers for parallel computing, the results are collected for post-processing to obtain the final results [15]. Storage servers need to periodically replicate data to multiple locations to achieve redundancy and load balancing. In data centers that allow dynamic

K. Xi (✉) • Y.-H. Kao • H.J. Chao
Polytechnic Institute of New York University, Brooklyn, New York 11201, USA
e-mail: kxi@poly.edu; ykao01@students.poly.edu; chao@poly.edu

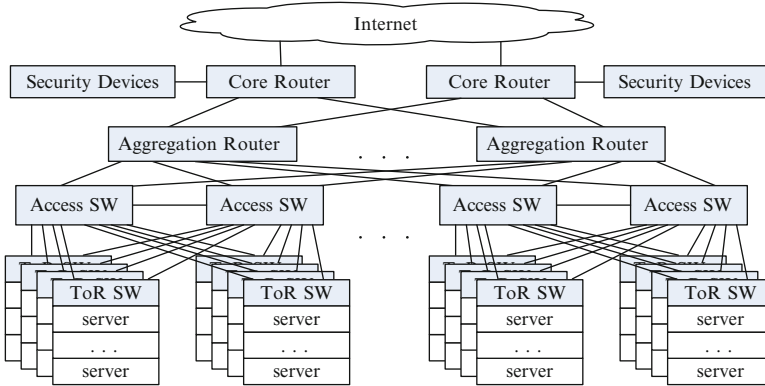


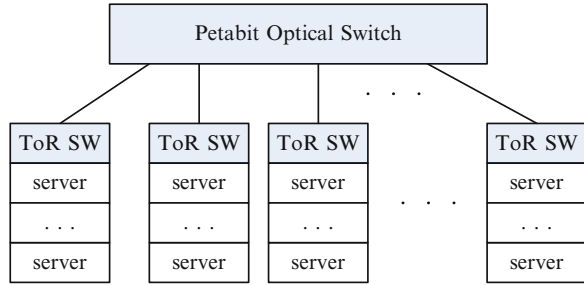
Fig. 8.1 Traditional data center network architecture

migration of virtual machines, system images need to be transferred between original and new servers whenever a migration is performed, generating huge amount of data exchange. To support such applications, data center networks need to provide high bandwidth and low latency with low complexity. However, when data centers scale up to host hundreds of thousands of servers and beyond, it becomes a great challenge to build an interconnection network with enormous bandwidth. For a data center hosting 300,000 servers each with two 1 Gb/s Ethernet interfaces, the required network bandwidth (without oversubscription) is 600 Tb/s. In contrast, the biggest router available on the market is Cisco's CRS-3 series, which offer a maximum bidirectional throughput of 322 Tb/s.

Today's data center networks have a multi-tiered architecture [11], as illustrated in Fig. 8.1. Servers on a rack are connected to one or two top-of-rack (ToR) switches. The ToR switches are then connected to access switches to form clusters. The access switches are interconnected by a small number of high-capacity aggregation switches, which are then interconnected by core switches. In such networks, the traffic concentrates toward the top, thus requires huge capacity at the aggregation and core switches. Currently people often rely on oversubscription at each layer to reduce the cost of switches. According to research in [18], the oversubscription ratio at the ToR switches is typically 1:5 to 1:20, while the overall ratio could reach 1:240.

The traditional architecture has several scalability problems. The first problem is the bandwidth bottleneck caused by oversubscription, which brings extra complexity to application design and deployment in that special consideration has to be taken to localize traffic [33]. The second problem is the long latency introduced by multiple hops when a packet traverses the aggregation and core switches to reach its destination. In particular, the delay grows significantly under heavy load and could harm delay-sensitive applications such as stock exchange [4]. The third problem is the wiring and control complexity, which grows super linearly when the data center scales. Furthermore, a lot of line cards are used between layers to carry transit traffic,

Fig. 8.2 Flat data center network using a single giant switch



but not revenue generating traffic. This is not very cost effective. Recently several designs have been proposed to address the above problems [2, 17–20, 24, 38, 41]. We discuss such designs in Sect. 8.6.

In this paper we propose to flatten data center networks by interconnecting all server racks through a single switch as shown in Fig. 8.2. Our target is to support 10,000 100 Gb/s ports in a single switch, providing Petabit/second switching capacity. The main contributions of our design include the following.

- We exploit recent advances in optics and combine them with the best features of electronics to design a high-capacity switch architecture that can scale up to a large number of high-speed ports while keeping the complexity and latency low.
- We design an optical interconnection scheme to reduce the wiring complexity of Clos network from $O(N)$ to $O(\sqrt{N})$, making it applicable to very large scale networks.
- We develop a practical and scalable packet scheduling algorithm that achieves high throughput with low complexity.
- We present the analysis of implementation issues to justify the feasibility of the design.

The rest of this paper is organized as follows: Section 8.2 provides the switch architecture. Section 8.3 presents the scheduling algorithm. Section 8.4 discusses the implementation issues. Section 8.5 gives performance evaluation results. Section 8.6 briefly surveys the related work. Finally, Sect. 8.7 concludes the paper.

8.2 Switch Architecture

The architecture of our design is illustrated in Fig. 8.3. The ToR switches are connected to the giant switch without any intermediate packet processing. Wavelength division multiplexing (WDM) is used to facilitate wiring. The switch fabric is a three-stage optical Clos network [9] including input modules (IMs), central modules (CMs), and output modules (OMs), where each module uses an arrayed waveguide grating router (AWGR) as the core. A prominent feature of the switch is that packets are buffered only at the line cards, while the IMs, CMs and OMs

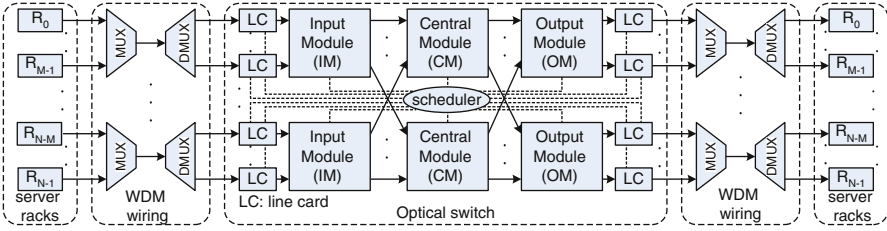


Fig. 8.3 Switch architecture

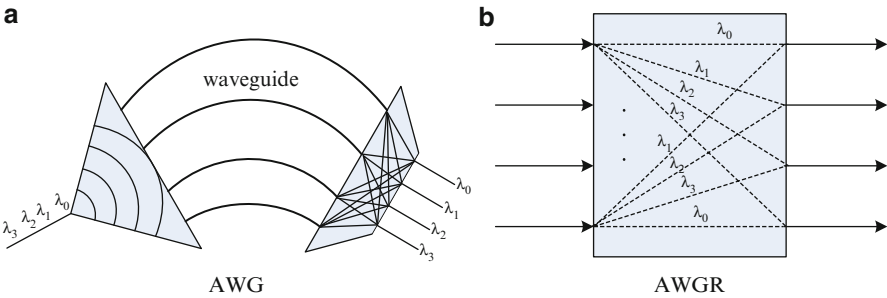


Fig. 8.4 Principle of AWG and AWGR

do not require buffers and fiber delay lines. This helps to reduce implementation complexity and to achieve low latency. We explain the details of the design in the following.

8.2.1 Optical Switch Fabric

Our design exploits two major optical devices: AWGRs and tunable wavelength converters (TWCs).

As shown in Fig. 8.4a, an arrayed waveguide grating (AWG) is a passive device that performs wavelength demultiplexing based on the principle of interference between light waves of different wavelengths. With waveguides engineered to specific lengths, the device creates desired phase shift for each wavelength. As a result, signals of a certain wavelength are directed to a specific output port. AWG can be used for WDM demultiplexing in one direction and multiplexing in the other direction. An $M \times M$ AWGR is an integration of M $1 \times M$ AWGs in a way that any output port receives a specific wavelength from only one input port in a cyclic way (Fig. 8.4b). The design ensures that input port i directs wavelength λ_m only to output port $(i+m) \bmod M$, where $i = 0, 1, \dots, M-1$. Therefore, to send a signal from input i to output j , the wavelength to use must be $\lambda_{M+j-i \bmod M}$.

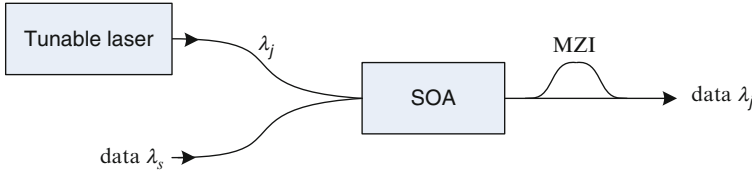


Fig. 8.5 Tunable wavelength converter

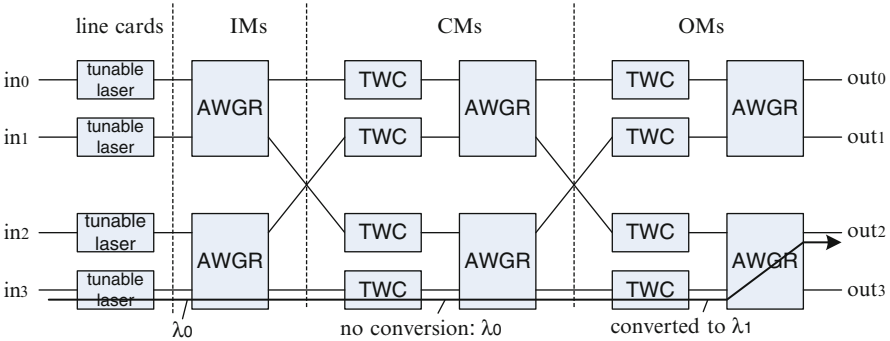


Fig. 8.6 A 4 × 4 switch fabric using 2 × 2 AWGRs and TWCs

AWGRs do not have good scalability due to limited number of ports they can support. In order to build a large-scale optical switch fabric, we need to interconnect multiple AWGRs to form a switch network. Due to the nature of AWG, the routing paths in such networks for a specific wavelength starting from an input port is fixed. To enable dynamic configuration of the switch fabric, we need to perform wavelength conversion using TWCs. A typical TWC is shown in Fig. 8.5. The input signal on wavelength λ_s is coupled onto λ_j that can be dynamically adjusted by tuning the tunable laser source. The conversion is performed by a semiconductor optical amplifier (SOA) followed by a Mach-Zehnder interferometer (MZI). The cross modulation in the SOA allows the data signal to change the gain and phase of λ_j . The MZI works as a filter to generate reshaped and clean pulses at λ_j . Technology advances have enabled wide range conversion of high bit rate signals. An SOA-MZI monolithic all-optical wavelength converter for full C-band operation has been reported in [37]. In [39] the authors show that wavelength conversion can be achieved with no extinction ratio degradation at 160 Gb/s. In a recent paper [1], all optical conversion was demonstrated to support 160 Gb/s in the full C-band.

In the Clos network, each IM/CM/OM includes an AWGR as the switch fabric. Each input port of the CMs and OMs has a TWC to control the routing path. The IMs do not need TWCs because the wavelength at their input ports can be adjusted by controlling the tunable laser source on the line cards. The example in Fig. 8.6 shows a 4 × 4 switch fabric using 2 × 2 AWGRs and a path from input 3 to output 2.

To establish the path, we configure the tunable devices so that the line card transmits at λ_0 , the TWC in the CM keeps λ_0 unchanged, and the TWC in the OM converts λ_0 to λ_1 . Amplifiers can be added between stages wherever necessary.

The Clos-based switch fabric has good scalability. If we use identical $M \times M$ AWGRs, the size of the switch is scaled to $M^2 \times M^2$. Currently 128×128 AWGRs are available [42], so it is feasible to reach our target at 10,000 ports. Building an optical switch fabric also helps to reduce the power consumption compared to electrical designs of the same throughput. It is worth noting that other fast reconfigurable optical switch modules can be used to substitute the AWGRs in our architecture with minor modifications of the scheduling algorithm.

8.2.2 Switch Configuration

Consider an $N \times N$ switch fabric using $M \times M$ AWGRs, we discuss the configuration of the tunable lasers and TWCs to set up a path from input port i to output port j through CM k , where $i, j = 0, 1, \dots, N - 1$ and $k = 0, 1, \dots, M - 1$. We explain how to select the correct wavelengths among $\lambda_0, \lambda_1, \dots, \lambda_{M-1}$.

1. Tunable laser on the line card: In the IM that input port i is connected to, the local input index is

$$i^* = i \bmod M. \quad (8.1)$$

The IM connects to CM k through its output port k . Therefore, the tunable laser at the input port should be tuned to transmit at wavelength

$$\lambda_{IM}(i, j, k) = \lambda_{M+k-i^* \bmod M}. \quad (8.2)$$

2. TWC in the CM: The index of the IM for this path is

$$I = \lfloor i/M \rfloor, \quad (8.3)$$

and the index of the OM is

$$J = \lfloor j/M \rfloor. \quad (8.4)$$

The task of CM k is to connect its input port I to its output port J , thus the corresponding TWC should be tuned to convert $\lambda_{IM}(i, j, k)$ to

$$\lambda_{CM}(i, j, k) = \lambda_{M+J-I \bmod M}. \quad (8.5)$$

3. TWC in the OM: CM k is connected to the k th input of OM J . The local output index on the OM corresponding to output j is

$$j^* = j \bmod M. \quad (8.6)$$

Therefore, the TWC at this OM is configured to convert $\lambda_{\text{CM}}(i, j, k)$ to

$$\lambda_{\text{OM}}(i, j, k) = \lambda_{M+j^*-k \bmod M}. \quad (8.7)$$

8.2.3 Frame-Based Switching

Although the optical switch fabric can be dynamically reconfigured, the configuration time is not negligible. Since the AWGRs do not need reconfiguration, the timing depends on the switching time of the TWCs and the tunable lasers on the line cards. Researchers in Bell Labs have demonstrated a monolithically integrated tunable wavelength converter with nanosecond level switching time. They also show that with proper electrical termination the tunable lasers can reach sub-nanosecond switching time [6]. Nonetheless, nanosecond-level switching time is still a considerable overhead that makes it impossible to perform per packet switching. The duration of a small Ethernet packet (64 bytes) on a 100 Gb/s link is only 5.12 ns.

We adopt frame-based switching in the data plane to reduce the impact of the switching time. Packets are assembled into fixed-size frames in the ingress of the line cards and disassembled at the egress of the line cards. In this design we set the frame size to 200 ns inside the switch fabric. For simplicity, we do not allow a packet being segmented to two frames. Between two consecutive frames we insert a guard time to allow switch fabric reconfiguration and frame alignment deviation.

8.2.4 Address Management

We provide a flat layer 2 address space for easy application deployment and virtual machine (VM) management. For example, a VM can be migrated to a new location without changing its IP address. The address management is performed by a centralized control plane. It is worth noting that centralized control has been adopted in several data center network designs since data centers are similar to autonomous systems and have small footprint [17, 18, 38].

The central controller performs address assignment and resolution as following.

- When a server (or VM) starts, the controller assigns it an IP address and creates a record with its MAC, IP, and rack index.
- The address resolution protocol (ARP) is modified (e.g., by modifying the hypervisor) to use the central controller. Each ARP request is sent to the controller using unicast, and the reply includes the MAC address and rack index of the destination. The rack index is embedded in the packet header and is used

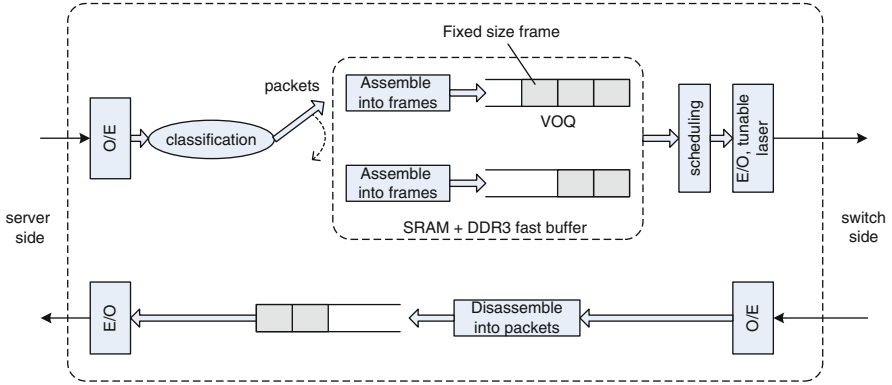


Fig. 8.7 Line card architecture

by the core switch to determine the output port. Note that the rack index can be converted to the corresponding output port, this design significantly reduces complexity since no table lookup is needed.

8.2.5 Line Card Design

The line card architecture is illustrated in Fig. 8.7. We focus on high-speed and low-complexity design and eliminate functions that are not necessary for data center network or can be distributed to other places.

At the ingress from the servers, a classification module assigns each packet to a VOQ. Note that the line card does not perform address lookup since each packet carries a label to identify the destination rack index, which is directly converted to the VOQ index. Packets belonging to the same VOQ are assembled into fixed-size frames. Given the huge size of the switch, it would introduce high complexity if a VOQ is maintained for each output port. In our design we create a VOQ for each OM. Although the per-OM VOQ scheme cannot completely avoid head-of-line (HOL) blocking, it reduces the number of queues and hardware complexity. Simulation shows that with a well-designed scheduling algorithm and a reasonable speedup, we can overcome HOL blocking to achieve near 100% throughput.

At the egress to the servers, frames are disassembled into packets, which are buffered briefly before being transmitted to the racks. Buffering is necessary because the switch fabric uses speedup to achieve high throughput. To avoid overflow and packet loss at the output buffer, backpressure is triggered once the queue length exceeds a predetermined threshold. The backpressure notifies the scheduler to suspend packet transmission to the signaling output port. To achieve high-speed buffering we use the widely adopted hybrid SRAM and DRAM architecture [23].

8.3 Scheduling Algorithm

8.3.1 Problem Statement

We consider an $N \times N$ switch consisting of $M \times M$ modules ($N = M^2$). The traffic demand is a binary matrix $\{d_{i,j}\}$, where $d_{i,j} = 1$ means at least one frame is waiting to be switched from input i to output j ($i, j = 0, \dots, N - 1$). The objective of scheduling is to find a bipartite match from the input ports to the output ports and assign a CM for each match such that the throughput is maximized. We use binary variable $s_{i,j}(k) = 1$ to denote that a match is found from input i to output j through CM k , where $k = 0, 1, \dots, M - 1$. The scheduling is formulated as a binary linear programming as follows.

maximize:

$$\sum_{k=0}^{M-1} \sum_{i,j=0}^{N-1} s_{i,j}(k) d_{i,j}, \quad (8.8)$$

subject to:

$$\sum_{k=0}^{M-1} \sum_{j=0}^{N-1} s_{i,j}(k) \leq 1, \quad \forall i = 0, 1, \dots, N - 1, \quad (8.9)$$

$$\sum_{k=0}^{M-1} \sum_{i=0}^{N-1} s_{i,j}(k) \leq 1, \quad \forall j = 0, 1, \dots, N - 1, \quad (8.10)$$

$$\sum_{i,j=0}^{N-1} s_{i,j}(k) \leq M, \quad \forall k = 0, 1, \dots, M - 1. \quad (8.11)$$

The objective function (8.8) is to maximize the number of input–output matches. Constraint (8.9) allows at most one connection from each input port. Constraint (8.10) allows at most one connection to each output port. Constraint (8.11) allows at most M connections through each CM.

Scheduling in input-queued switches has been researched extensively. Typical algorithms include parallel iterative matching (PIM) [3], iSLIP [31], dual round robin matching (DRRM) [8, 10, 27], longest queue first (LQF), oldest cell first (OCF) [32], etc. Routing in Clos network switches has also been studied and many algorithms have been proposed, such as m -matching [22], Euler partition algorithm [12], Karol's algorithm [9]. However, the above algorithms cannot be applied to very large switches (e.g., $10,000 \times 10,000$) due to implementation complexity

8.3.2 Frame Scheduling Algorithm

We develop an iterative frame scheduling algorithm that is practical and scalable. Corresponding to each IM, CM, and OM we have an scheduling module, which we call scheduler at IM (SIM), scheduler at CM (SCM) and scheduler at OM (SOM), respectively. The algorithm completes in H iterations as described below.

- *Request*: Each input port chooses H VOQs using round robin and sends the requests to the corresponding SIM.
- *Iteration*: Repeat the following steps for H cycles. In the h th cycle ($h = 1, 2, \dots, H$), we only consider the h th request from each input port.
 1. *Request filtering*: If an input port received a grant in the previous iterations, its requests are excluded from the subsequent iterations. If an SIM receives multiple requests competing for an output port, it randomly chooses one request.
 2. *CM assignment*: Each SIM randomly assigns an available CM to each request and sends the request to the corresponding SCM.
 3. *CM arbitration*: If a SCM receives multiple requests pointing to the same OM, it chooses one using round robin. The selected requests are sent to the corresponding SOMs.
 4. *OM arbitration*: For each output port, the corresponding SOM grants the first request using round robin if the output port did not grant any requests in the previous iterations and is not doing backpressure.

In our switch there are two contention places: at CMs and at output ports. We combine multiple approaches to resolve the contention. We introduce randomness to reduce the contention probability. We allow each input to generate H requests. (Note that generating N requests would be good but incurs high communication overhead.) We also employ multiple iterations and speedup. Our simulation shows that with three iterations and 1.6 speedup, the scheduling algorithm achieves nearly 100% throughput under various traffic distribution and switch sizes.

8.3.3 Packet Scheduler Design

Figure 8.8 shows an architecture to implement our proposed packet scheduling algorithm, which consists of three stages: (1) CM assignment at the SIM stage, (2) output link arbitration at the SCM stage, (3) output port arbitration at the SOM stage. A request consists of 15 bits, with the first bit indicating whether there is a request or not, and 14 bits for output port addresses.

At each SIM, a filter selects a request among those destined to the same output port. The winner is then assigned an available CM number randomly and then sent to the corresponding SCM. Each SIM, e.g., SIM_i , has an CM availability vector, denoted as A_i . The vector has M bits indicating CM's availability. A "1" means

Fig. 8.8 Packet scheduler with three-stage arbitration

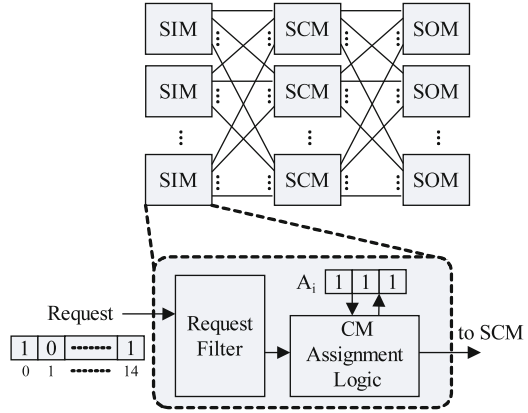
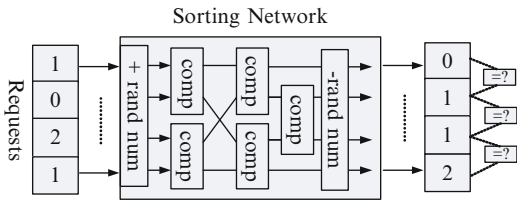


Fig. 8.9 Request filter structure



that the link to the CM is available and “0” unavailable. Detailed operation of the filter and CM assignment is described next. The output link arbitration at SCM and output port arbitration at SOM are performed based on round robin and can be easily implemented for the size of 128.

As shown in Fig. 8.9 the request filter can be implemented by sorting the requests based on the output port addresses and then picking the first request from each overlapped group. Sorting can be implemented using a Batcher sorting network with multi-stage comparators [5]. After being sorted, a request is invalidated if its destination is the same as its upper neighbor request.

Our algorithm selects a winner from the overlapped requests randomly to achieve fairness. The randomness is realized by appending a random number to each request before sorting and removing that number after sorting. For example, input ports 3 and 4 request output port 1, if their random numbers are 3 and 6, respectively, the values to be sorted become 10000011 and 100000110, making input port 3 the winner. The random numbers can be generated using simple thermal noise-based approaches [13].

The CM assignment logic is illustrated in Fig. 8.10. The input ports are sorted to place those with request (bit 1) at the beginning of the output vector. The CMs are sorted to place the available ones (bit 1) at the beginning of the output vector. By matching the two vectors we realize random CM assignment to the requests. The sequence of requests to this logic is already randomized in the request filter in Fig. 8.9, thus we don’t need to add randomness here.

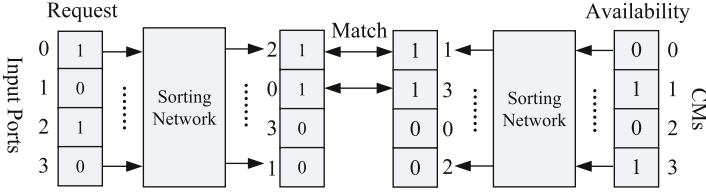


Fig. 8.10 CM assignment using two sorters

The packet scheduler has been designed with VHDL code and synthesized and analyzed by the Cadence Encounter RTL Compiler on the ST Microelectronics Company 65nm technology. The evaluation shows the latency is 21 ns for a 10,000 × 10,000 switch.

8.3.4 Multi-chip Scheduler and the Inter-chip Connections

It is not possible to build the packet scheduler on a single chip because the total I/O bandwidth is too high. We discuss practical multi-chip implementation in this section.

Based on our performance study, in order to achieve high throughput and low latency, the packet scheduler needs to use three iterations to arbitrate up to 3 different requests from each line card. In each cycle, each line card can send three requests (with 15 bits each) to the packet scheduler and receive a grant signal. Thus the link bandwidth from the line card to its SIM is 225 Mb/s (45 bits/200 ns). The link bandwidth between the stages of the packet scheduler is also 225 Mb/s. The bandwidth for carrying the acknowledge signals (grant/reject) on the reverse paths is negligible as compared to the bandwidth on the forwarding paths. The total bandwidth between the line card and the scheduler and between the stages for a 10,000-port packet scheduler is 2.25 Tb/s, which is too high for a single-chip scheduler. We study two methods to partition the packet scheduler into multiple chips with I/O bandwidth constraint for each chip.

The first method groups k SIMs into a chip, as shown in Fig. 8.11a, and does the same to SCMs and SOMs. This method needs three different types of chips to implement the scheduler and is not cost effective. The second method groups k SIMs, k SCMs and k SOMs in the same rows into a chip, as shown in Fig. 8.11b, and only requires one type of chip. The maximum number of links for each chip on the forwarding direction is:

$$L = M \times k + 2 \times (M - k) \times k. \tag{8.12}$$

As each SIM has M links connected to the line card and each chip has k SIMs, $M \times k$ is the number of links from the line card to the SIMs. On each chip, there are

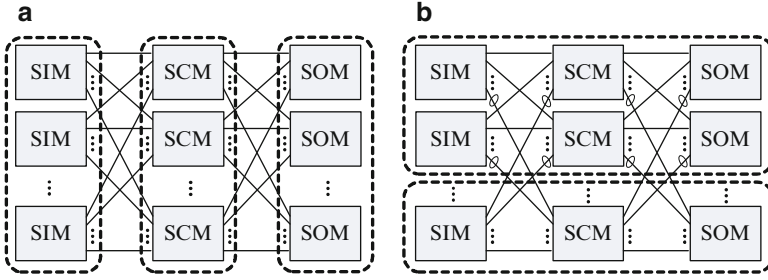


Fig. 8.11 Partitioning of three-stage packet scheduler

$(M - k) \times k$ inter-chip links (circled in Fig. 8.11b) between the SIMs and the SCMs, and the number of inter-chip links between the SCMs and SOMs is the same. Denote the rate of each line as r , the I/O bandwidth of each chip becomes

$$C_R = L \times r. \quad (8.13)$$

The second method achieves better reduction of wiring complexity. This is because some of the wires are on-chip connections. We want to maximize k to reduce the number of chips and the number of wires between chips. The current technology can achieve 470 Gb/s per single chip [30]. For a 10,000-port scheduler, we let $C_R < 470$ Gb/s and get $k = 7$. So the scheduler can be implemented using 15 chips. The bi-directional I/O bandwidth for each chip is 450 Gb/s, implemented with 10 Gb/s SERDES.

8.4 Implementation

8.4.1 Racks and Wiring

We adopt the multi-rack scheme to build the proposed large-scale switch. The switch fabric is placed on a single rack and the line cards are hosted on multiple racks.

Wiring in the switch fabric is a great challenge as there are M^2 interconnections between two stages and the wires from one module is spread to all the other modules, making it impossible to use WDM in a straightforward way. We reduce the number of connections from M^2 to M by using WDM and a single AWGR between two stages. The principle is illustrated in Fig. 8.12. The output ports of each IM is converted to a sequence of wavelengths. The wavelength pattern is fixed, thus only fixed wavelength converters (WCs) are needed. Now the M wavelengths from each IM can be multiplexed to a single fiber and transmitted to an $M \times M$ AWGR. Due to the cyclic property of AWGR, wavelengths from all the IMs are cross-connected to the output side and automatically multiplexed without conflict.

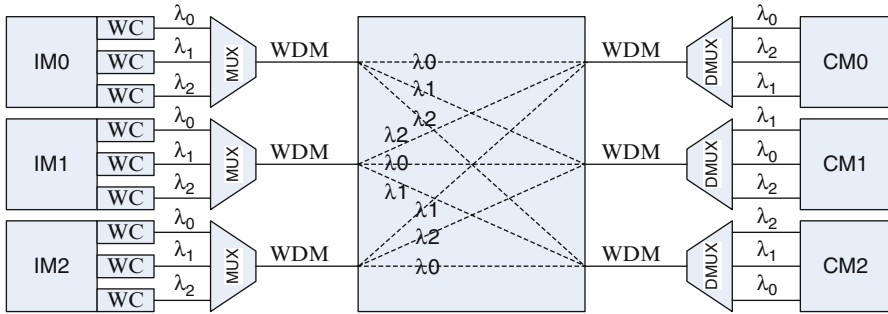


Fig. 8.12 Reduce wiring complexity between switching stages of the Clos network by using WDM and AWGR

The multiplexed signals are transmitted through M fibers to the CMs, where each fiber is demultiplexed to M signals.

We use WDM for interconnections between server racks and line card racks and between line card racks and the switch fabric rack. This can greatly reduce the wiring complexity.

8.4.2 Timing and Frame Alignment

The propagation of optical signals in fibers is about 5 ns per meter. With the frame length set to 200 ns, the delay between optical modules is not negligible. Since the optical switch fabric is bufferless, it would be ideal for the frames to be precisely aligned. We relax this strict constraint by inserting a guard time between consecutive frames and using configurable fiber delay lines to compensate the propagation delay difference. The guard time is also used to tolerate the switch reconfiguration time, which is in nanosecond level.

Fiber delay lines are installed at the input of each IM, CM, and OM so that frames are aligned. Each delay line is configured based on the corresponding propagation delay. Measurement of the propagation delay and the configuration are performed only once after system start or a rewire operation.

We place the scheduler and the optical switch fabric on the same rack. The distances from the line card racks to the switch fabric rack is bound by the frame length. This is because in one frame cycle (200 ns) we must complete the transmission of requests from the line cards to the scheduler, scheduling, and the transmission of grants from the scheduler to the line cards. Since the scheduling takes 21 ns for a $10,000 \times 10,000$ switch (refer Sect. 8.3.3), there are 179 ns for request/grant propagation. Thus the maximum distance between a line card rack to the switch fabric rack is about 179 m. By placing the line card racks and switch fabric properly, this is sufficient for typical data centers.

Table 8.1 Performance for various switch sizes

Number of switch ports	10,000	4,096	1,024
Throughput	99.6%	99.6%	99.6%
Average queue length, 80% load (frames)	1.53	1.52	1.51
Average queue length, 90% load (frames)	2.17	2.15	2.11
Average latency, 80% load (frames)	2.27	2.27	2.25
Average latency, 90% load (frames)	5.19	5.17	5.13

8.5 Performance Evaluation

We evaluate the performance of the design using our in-house frame-based simulator that is cycle-accurate. We study the throughput, queue length, and average frame latency under various switch sizes and traffic patterns. The results show that with internal speedup of 1.6 and three iterations the switch can achieve close to 100% throughput.

8.5.1 Scalability and Throughput

Table 8.1 shows the throughput, average queue length, and average latency under 80 and 90% load for various switch sizes under uniform traffic. All the switches can approach near 100% throughput. The average latency is only a little more than twice of a frame duration (200 ns) at 80% load. In particular, it shows that the performance is virtually independent of the switch size, which demonstrates excellent scalability of our design.

8.5.2 Delay Performance

Figure 8.13 shows the delay performance with switch size $1,024 \times 1,024$. We use the following traffic patterns.

- Uniform: Each input sends to all the outputs evenly.
- Transpose: Input i only sends to output $i + 1 \bmod N$.
- Non-uniform: 50% of the traffic is uniform, and 50% is transpose.
- Cluster: Input i sends to outputs $i + k \bmod N$ ($k = 1, \dots, 10$) evenly.

Figure 8.13 shows that the switch can approach near 100% throughput under various traffic. The performance under transpose traffic is the best because there is no output contention. On the other hand, the cluster pattern tends to create the most output contention. Nonetheless, the latency keeps to be very small until the load exceeds 95%.

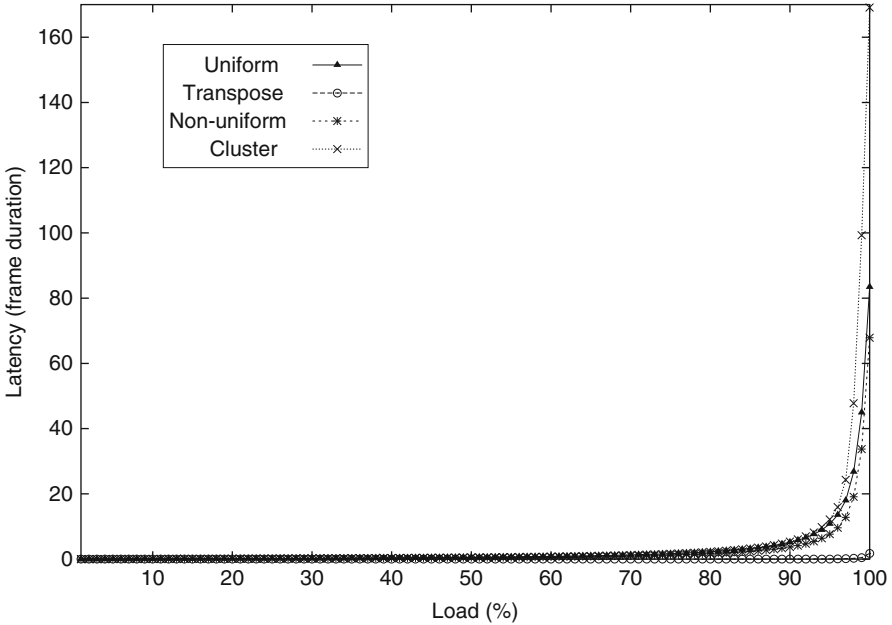


Fig. 8.13 Average latency (in frame duration) under various load

One may wonder if the performance can be further improved by using more iterations. Our simulation shows that while one and two iterations are not good enough, three iterations are sufficient (Fig. 8.14).

8.6 Related Work

With the rapid growth of data centers it has gained much attention to design novel data center architecture. One approach is to scale out data center networks using a large number of small commodity switches, the typical designs include Portland [38] and virtual layer 2 (VL2) [18]. Portland adopts a three-stage fat tree topology using k -port switches and can support non-blocking communication among $k^3/4$ host stations with $5k^2/4$ switches. In VL2, IP switches are interconnected as a Clos network to scale out the topology. Both designs separate names from locator by using a well-designed location-specific address space for routing in the network. Mapping between the actual addresses and the location-specific addresses is managed by a centralized controller. The wiring complexity can be reduced by packaging switch modules together [16]. We embrace the idea of scaling out topology using small switch modules. However, our design is different in that it adopts bufferless optical switch modules to improve performance and reduce complexity.

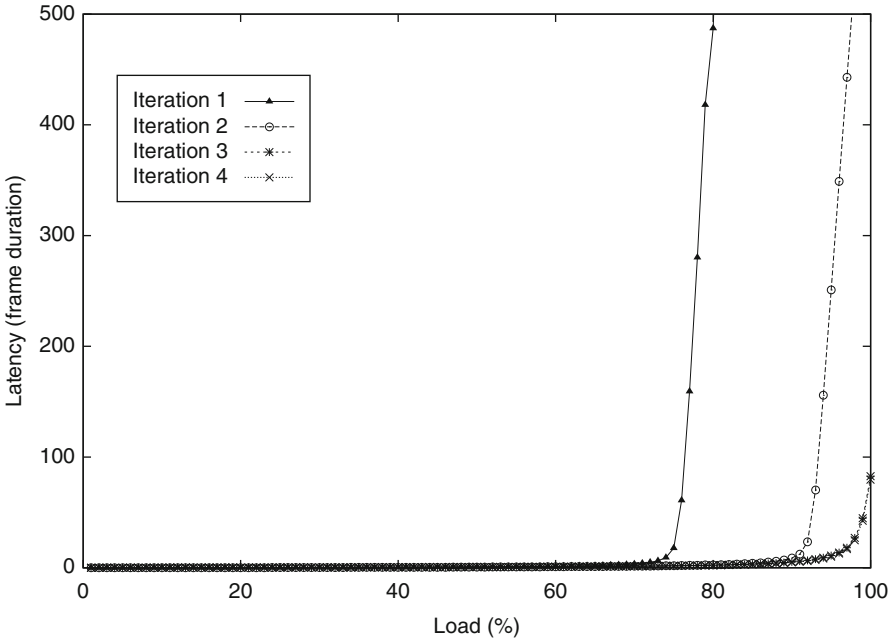


Fig. 8.14 Average latency (in frame duration) with various iterations

Another approach is server-centric where servers perform both computing and packet switching. Such designs include DCell [19], BCube [20], DPillar [28]. Using the proposed topologies a server-centric data center can scale to host hundreds of thousands of servers. Since general-purpose servers are not optimized for fast and reliable packet switching, special hardware and software processing is needed for this type of designs to provide performance guarantee in terms of end-to-end delay and network resilience.

Two recent work, called Helios [17] and HyPaC [41], propose to augment the packet-switched network with an optical circuit-switched network. Optical circuit switching has the advantages of high bandwidth, low cost, and low power consumption but suffers from considerable reconfiguration time. The essence of Helios and HyPaC is to examine the characteristics of traffic in real time and dynamically set up circuits to carry the suitable traffic. The scheduling has to be well designed to avoid frequent circuit reconfiguration and minimize the impact of reconfiguration time.

In the area of high-speed switch design, multi-stage multi-plane architecture using buffered switch modules is adopted by TrueWay switch [9] and Cisco CRS-1 router [40]. While this architecture is more scalable than single-stage input-queued switches, its scalability is limited by power consumption, multi-stage queueing delay, and hardware complexity. Load-balanced switches have a two-stage architecture where the first stage performs load balancing and the second

stage delivers packets to the destination [7, 25, 26]. The switch has low complexity because both stages perform periodic permutation without any scheduling. When scaling to many ports, the delay caused by fixed permutation and packet reordering would increase substantially, making it inappropriate for data center applications. In the area of optical packet switching, the major hurdle for all-optical switch is the lack of optical random access memory. Data Vortex is an optical switch that achieves high capacity using deflection routing and transparent wavelength switching [21]. Deflection routing is good at avoiding buffering but could cause significant delay when the switch scales to large size. The optical shared memory supercomputer interconnect system (OSMOSIS) has an SOA-based crossbar to support switching using broadcast-and-select [29, 36]. It also employs space- and wavelength-division multiplexing to increase the capacity. A 64-port, 40 Gb/s demonstration system has been successfully built. Like traditional crossbar switches, OSMOSIS's scalability is limited by its hardware complexity: $O(N^2)$. In addition, broadcasting is also limited by signal power when scaling to many ports.

8.7 Conclusions

We present an ultra large-scale switch to meet the requirements of data center networks for high capacity, low latency, and low complexity. The switch combines the best features of optics and electronics to reduce the complexity. We present a scalable and practical scheduling algorithm and verified its hardware implementation. We develop novel interconnection network to substantially reduce the wiring complexity. Simulation results show that the switch can reach very high throughput under various traffic patterns.

References

1. Akimoto R, Gozu S, Mozume T, Akita K, Cong G, Hasama T, Ishikawa H (2009) All-optical wavelength conversion at 160Gb/s by intersubband transition switches utilizing efficient XPM in InGaAs/AlAsSb coupled double quantum well. In: European conference on optical communication, pp 1–2, 20–24
2. Al-Fares M, Loukissas A, Vahdat A (2008) A scalable, commodity data center network architecture. In: SIGCOMM '08: Proceedings of the ACM SIGCOMM 2008 conference on data communication. ACM, New York, pp 63–74
3. Anderson TE, Owicki SS, Saxe JB, Thacker CP (1993) High speed switch scheduling for local area networks. *ACM Trans Comp Syst* 11:319–352
4. Bach A (2009) High Speed Networking and the race to zero. Keynote speech, 2009 IEEE Symposium on High Performance Interconnects. ISBN: 978-0-7695-3847-1
5. Batcher K (1968) Sorting networks and their applications. In: American Federation of Information Processing Societies conference proceedings, pp 307–314
6. Bernasconi P, Zhang L, Yang W, Sauer N, Buhl L, Sinsky J, Kang I, Chandrasekhar S, Neilson D (2006) Monolithically integrated 40-Gb/s switchable wavelength converter. *J Lightwave Technol* 24(1):71–76

7. Chang C-S, Lee D-S, Lien C-M (2001) Load balanced Birkhoff-von Neumann switches with resequencing. *SIGMETRICS Perform Eval Rev* 29(3):23–24
8. Chao H (2000) Saturn: a Terabit packet switch using dual round robin. *IEEE Comm Mag* 38(12):78–84
9. Chao HJ, Liu B (2007) High performance switches and routers. Wiley-IEEE Press. ISBN: 978-0-470-05367-6, Hoboken, New Jersey
10. Chao HJ, soo Park J (1998) Centralized contention resolution schemes for a large-capacity optical ATM switch. In: *Proceedings of IEEE ATM Workshop*, pp 11–16
11. Cisco (2007) Cisco Data Center infrastructure 2.5 design guide. Cisco Systems, Inc.
12. Cole R, Hopcroft J (1982) On edge coloring bipartite graph. *SIAM J Comput* 11(3):540–546
13. Danger JL, Guilley S, Hoogvorst P (2009) High speed true random number generator based on open loop structures in FPGAs. *Microelectron J* 40(11):1650–1656
14. Dean J (2009) Large-scale distributed systems at Google: current systems and future directions. In: *LADIS '09: ACM SIGOPS international workshop on large scale distributed systems and middleware*. Keynote speech, available online at www.cs.cornell.edu/projects/ladis2009/talks/dean-keynote-ladis2009.pdf, Accessed Sep 2012
15. Dean J, Ghemawat S (2008) MapReduce: simplified data processing on large clusters. *Comm ACM* 51(1):107–113
16. Farrington N, Rubow E, Vahdat A (2009) Data center switch architecture in the age of merchant silicon. In: *7th IEEE Symposium on High Performance Interconnects (HOTI)* pp 93–102
17. Farrington N, Porter G, Radhakrishnan S, Bazzaz HH, Subramanya V, Fainman Y, Papen G, Vahdat A (2010) Helios: A hybrid electrical/optical switch architecture for modular data centers. In: *SIGCOMM '10: Proceedings of the ACM SIGCOMM 2010 conference on data communication*. ACM, New York
18. Greenberg A, Hamilton JR, Jain N, Kandula S, Kim C, Lahiri P, Maltz DA, Patel P, Sengupta S (2009) VL2: a scalable and flexible data center network. In: *SIGCOMM '09: Proceedings of the ACM SIGCOMM 2009 conference on data communication*. ACM, New York, pp 51–62
19. Guo C, Wu H, Tan K, Shi L, Zhang Y, Lu S (2008) DCell: A scalable and fault-tolerant network structure for data centers. In: *SIGCOMM '08: Proceedings of the ACM SIGCOMM 2008 conference on data communication*. ACM, New York, pp 75–86
20. Guo C, Lu G, Li D, Wu H, Zhang X, Shi Y, Tian C, Zhang Y, Lu S (2009) BCube: a high performance, server-centric network architecture for modular data centers. In: *SIGCOMM '09: Proceedings of the ACM SIGCOMM 2009 conference on data communication*. ACM, New York, pp 63–74
21. Hawkins C, Small BA, Wills DS, Bergman K (2007) The data vortex, an all optical path multicomputer interconnection network. *IEEE Trans Parallel Distrib Syst* 18(3):409–420
22. Hopcroft J, Karp R (1973) An $n^{5/2}$ algorithm for maximum matchings in bipartite graphs. *SIAM J Comput* 2(4):225–231
23. Iyer S, Kompella R, McKeown N (2008) Designing packet buffers for router linecards. *IEEE/ACM Trans Networking* 16(3):705–717
24. Juniper (2010) Network fabrics for the modern data center. White Paper, Juniper Networks, Inc.
25. Keslassy I (2004) The load-balanced router. PhD thesis, Stanford University, Stanford, CA, USA. Adviser-Mckeown, Nick
26. Keslassy I, Chuang S-T, Yu K, Miller D, Horowitz M, Solgaard O, McKeown N (2003) Scaling internet routers using optics. In: *SIGCOMM '03: Proceedings of the ACM SIGCOMM 2003 conference on data communication*. ACM, New York, pp 189–200
27. Li Y, Panwar S, Chao H (2001) On the performance of a dual round-robin switch. In: *IEEE INFOCOM*, vol 3, pp 1688–1697
28. Liao Y, Yin D, Gao L (2010) DPillar: scalable dual-port server interconnection for data center networks. In: *IEEE International Conference on Computer Communications and Networks (ICCCN)*, pp 1–6
29. Luijten R, Grzybowski R (2009) The OSMOSIS optical packet switch for supercomputers. In: *Conference on Optical Fiber Communication OFC 2009*. pp 1–3

30. Mahony FO et al (2010) A 47times10 Gb/s 1.4 mW/(Gb/s) Parallel Interface in 45 nm CMOS. In: IEEE international solid-state circuits conference 45(12):2828–2837
31. McKeown N (1999) The iSLIP scheduling algorithm for input-queued switches. *IEEE/ACM Trans Networking* 7(2):188–201
32. McKeown N, Mekkittikul A, Anantharam V, Walrand J (1999) Achieving 100% throughput in an input-queued switch. *IEEE Trans Comm* 47(8):1260–1267
33. Meng X, Pappas V, Zhang L (2010) Improving the scalability of data center networks with traffic-aware virtual machine placement. In: IEEE INFOCOM, pp 1–9, 14–19
34. Miller R (2008) Microsoft: 300,000 servers in container farm. <http://www.datacenterknowledge.com/archives/2008/05/07/microsoft-300000-servers-in-container-farm>. Accessed May 2008
35. Miller R (2009) Who has the most web servers? <http://www.datacenterknowledge.com/archives/2009/05/14/whos-got-the-most-web-servers>. Accessed May 2009
36. Minkenber C, Abel F, Muller P, Krishnamurthy R, Gusat M, Dill P, Iliadis I, Luijten R, Hemenway R, Grzybowski R, Schiattarella E (2006) Designing a crossbar scheduler for HPC applications. *IEEE Micro* 26(3):58–71
37. Miyazaki Y, Miyahara T, Takagi K, Matsumoto K, Nishikawa S, Hatta T, Aoyagi T, Motoshima K (2006) Polarization-insensitive SOA-MZI monolithic all-optical wavelength converter for full C-band 40Gbps-NRZ operation. In: European conference on optical communication, pp 1–2, 24–28
38. Niranjan Mysore R, Pamboris A, Farrington N, Huang N, Miri P, Radhakrishnan S, Subramanya V, Vahdat A (2009) PortLand: a scalable fault-tolerant layer 2 data center network fabric. In: SIGCOMM '09: Proceedings of the ACM SIGCOMM 2009 conference on data communication. ACM, New York, pp 39–50
39. Pina J, Silva H, Monteiro P, Wang J, Freude W, Leuthold J (2007) Performance evaluation of wavelength conversion at 160 Gbit/s using XGM in quantum-dot semiconductor optical amplifiers in MZI configuration. In: Photonics in switching, 2007, pp 77–78, 19–22
40. Sudan R, Mukai W (1994) Introduction to the Cisco CRS-1 carrier routing system. Cisco Systems, Inc. White Paper
41. Wang G, Andersen DG, Kaminsky M, Papagiannaki K, Ng TSE, Kozuch M, Ryan M (2010) c-Through: part-time optics in data centers. In: SIGCOMM '10: Proceedings of the ACM SIGCOMM 2010 conference on data communication. ACM, New York
42. Xue F, Ben Yoo S (2004) High-capacity multiservice optical label switching for the next-generation Internet. *IEEE Comm Mag* 42(5):S16–S22

Chapter 9

Optically Interconnected High Performance Data Centers

Keren Bergman and Howard Wang

9.1 Introduction

Over the years, advances in optical technologies have enabled unprecedented data transmission capacities through the engineering and exploitation of a number of extremely advantageous physical properties inherent to photonic media. Wavelength division multiplexing (WDM), which is enabled by the enormous bandwidth of guided optics (nearly 32 THz in optical fiber [1]), represents a highly disruptive capability enabling information transmission across a single physical photonic channel at data rates many orders of magnitude greater than its copper-based counterpart, with demonstrated capacities exceeding 20 Tb/s [2] in single mode fiber. The characteristically low loss of optical media further enables extremely high bandwidth-distance and bandwidth-energy products, lifting previously unyielding architectural constraints dictated by the physical limitations of electronic interconnects [3]. Moreover, optical fiber media can sustain much smaller bending radii with significantly lower volume and weight, resulting in much more tenable and robust physical cabling.

As a result, photonic media has recently seen appreciable penetration into large-scale cluster computing systems, where unprecedented growth in application scale and hardware parallelism has combined to impose increasingly intractable communications requirements on its underlying interconnection network. Given the immense number of computation and storage elements, performance of these systems is reliant upon the effective exchange of vast amounts of data among end nodes (e.g., processors, memory, and storage). Therefore interconnects capable of

K. Bergman (✉) • H. Wang (✉)
Department of Electrical Engineering, Columbia University, New York, NY 10027, USA
e-mail: bergman@ee.columbia.edu; howard@ee.columbia.edu

supporting high-bandwidth low-latency communication across the scale of these highly distributed machines have become a nearly ubiquitous requirement for large-scale systems [4].

Accordingly, system designers have begun to embrace optical interconnects in production large-scale systems in the form of point-to-point links [3, 5]. While point-to-point interconnects have received significant attention and acceptance commercially, they can only partially alleviate the burgeoning bandwidth and power constraints plaguing modern day systems. Conventional electronic switches are still required at the terminus of each optical link. As these switches scale in port count and capacity, they are reaching fundamental performance limits. Worse still, the power consumed by electronic switches is already prohibitively high and continues to grow super-linearly with port count and bandwidth.

Therefore, in order to effectively address the power, bandwidth, and latency requirements imposed by these systems, optically switched networks have been proposed as a possible solution. By providing end-to-end optical paths from the source to the destination, all-optical networks can forgo costly translations between the electronic and optical domains. Photonic switches operate on the principle of routing lightpaths. This is a fundamentally different operation than that of electronic switching, which must store and transmit each bit of information individually. By doing so, a conventional electronic switch dissipates energy with each bit transition, resulting in power consumption proportional to the bitrate of the information it carries. However, a lightpath through an optical switch ideally remains transparent to the information it carries, a critical characteristic known as bit rate transparency [6]. Consequently, unlike an electronic switch, the power consumed by an optical switch is independent of the bitrate of the information it is routing. Therefore, scalability to significantly higher transmission bandwidths (using techniques such as WDM) can be achieved, enabling extremely low power-per-unit bandwidths through all-optically switched interconnection networks.

While bit-rate transparency is an eminently advantageous property for the design of high-bandwidth, energy-efficient switches, there are other fundamental properties of optical technology that represent significant challenges toward the realization of all-optical switches. Depending on the design and technology employed in optical switches, signal impairment and distortion due to effects such as noise and optical nonlinearities must be carefully considered. More critically, inherent limitations of the optical medium give rise to two architectural challenges that must be addressed: namely, the lack of effective photonic memories and the extremely limited processing capabilities realizable in the optical domain. Electronic switches are heavily reliant upon random access memories (RAM) to buffer data while routing decisions are made and contention resolution is performed. Effective header parsing and processing in electronics is simply assumed. As no effective photonic equivalent of RAM or processing exists, critical functionalities such as contention resolution and header parsing will need to be addressed in a manner unique to optical switching, dictating photonic network designs that are similarly unique. In the following sections, we describe 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.

9.2 Data Vortex

The data vortex architecture [7], specifically designed to be implemented as an all-optical packet switched topology, is comprised of simple 2×2 all-optical switching nodes (Fig. 9.1). Each node utilizes two semiconductor optical amplifier (SOA) devices, which perform the switching operation. Given the wide gain-bandwidth of the SOAs, the network utilizes a multi-wavelength striped packet format (Fig. 9.2), with high bit-rate payload data segmented across multiple channels and low bit-rate addressing information encoded on dedicated wavelengths, one bit per wavelength per packet. Passive optical splitters and filters within the node extract the relevant routing information (a frame bit to denote the presence of a packet and a header bit to determine the switch’s configuration), which are subsequently detected by low-speed receivers. The SOA pair is controlled via high-speed electronic decision circuitry, and routes the packet based on the recovered header information.

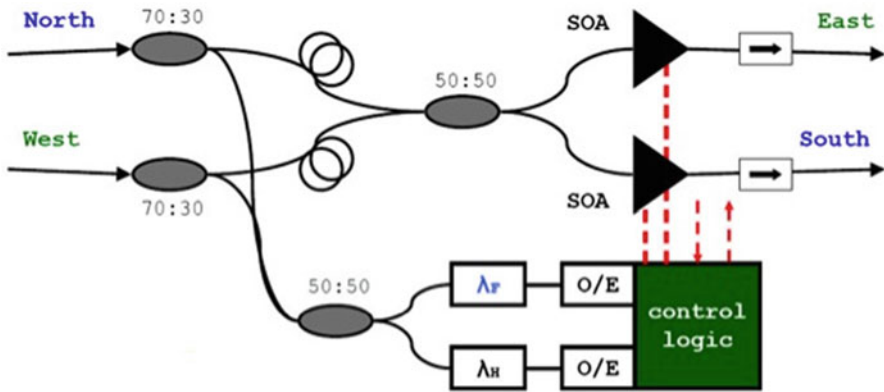


Fig. 9.1 (a) 2×2 data vortex switching node design

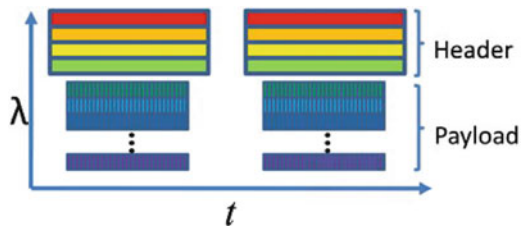


Fig. 9.2 Multi-wavelength striped packet format

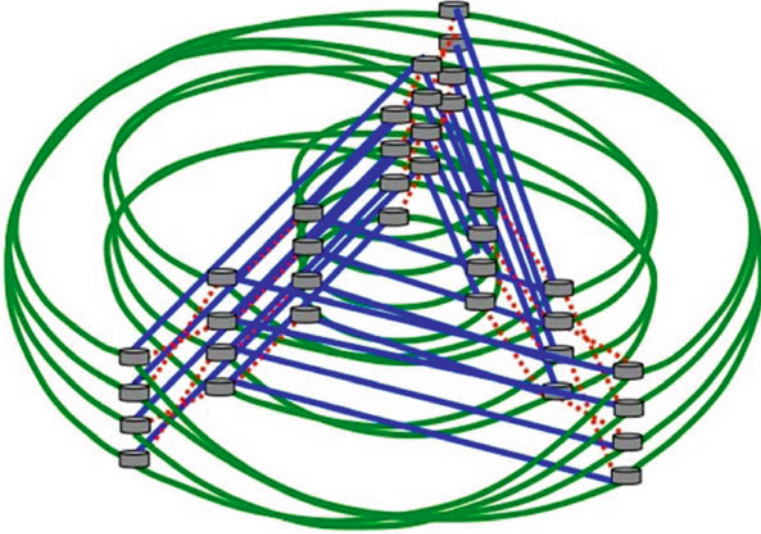


Fig. 9.3 Topology of a 12×12 data vortex all-optical packet switch consisting of $36 \times 2 \times 2$ switching nodes. Green lines represent deflection fibers while blue lines represent ingress fibers

In a data vortex topology, the 2×2 switching nodes are organized as concentric cylinders and addressed according to their location within the topology, represented by their cylinder, height, and angle (C, H, A) (Fig. 9.3). The switching elements are arranged in a fully connected, directed graph with terminal symmetry but not complete vertex symmetry. The single-packet routing nodes are wholly distributed and require no centralized arbitration. The topology is divided into C hierarchies or cylinders, which are analogous to the stages in a conventional banyan network (e.g., butterfly). The architecture also incorporates deflection routing, which is implemented at every node; deflection signal paths are placed only between different cylinders. Each cylinder (or stage) contains A nodes around its circumference and $H = 2^{C-1}$ nodes down its length. The topology contains a total of $A \times C \times H$ switching elements, or nodes, with $A \times H$ possible input terminal nodes and an equivalent number of possible output terminal nodes. The position of each node is conventionally given by the triplet (c, h, a) , where $0 \leq c \leq C-1$, $0 \leq h \leq H-1$, $0 \leq a \leq A-1$.

The switching nodes are interconnected using a set of ingress fibers, which connect nodes of the same height in adjacent cylinders, and deflection fibers, which connect nodes of different heights within the same cylinder. The ingress fibers are of the same length throughout the entire system, as are the deflection fibers. The deflection fibers' height crossing patterns direct packets through different height levels at each hop to enable banyan routing (e.g., butterfly, omega) to a desired height, and assist in balancing the load throughout the system, mitigating local congestion [8–11].

Incoming packets are injected into the nodes of the outermost cylinder and propagate within the system in a synchronous, time-slotted fashion. The conventional nomenclature illustrates packets routing to progressively higher numbered cylinders as moving inward toward the network outputs. During each timeslot, each node either processes a single packet or remains inactive. As a packet enters node (c, h, a) , the c th bit of the packet header is compared to c th most significant bit in the node's height coordinate (h). If the bits match, the packet ingresses to node $(c + 1, h, a + 1)$ through the node's south output. Otherwise, it is routed eastward within the same cylinder to node $(c, G_c(h), a + 1)$, where $G_c(h)$ defines a transformation which expresses the abovementioned height crossing patterns (for cylinder c) [10, 11]. Thus, packets progress to a higher cylinder only when the c th address bit matches, preserving the $c-1$ most significant bits. In this distributed scheme, a packet is routed to its destination height by decoding its address in a bitwise banyan manner. Moreover, all paths between nodes progress one angle dimension forward and either continue around the same cylinder while moving to a different height, or ingress to the next hierarchical cylinder at the same height. Deflection signals only connect nodes on adjacent cylinders with the same angular dimension; i.e. from $(c + 1, h, a)$ to a node at position $(c, G_{c+1}(h), a)$.

The paths within a cylinder differ depending upon the level c of the cylinder. The crossing or sorting pattern (i.e., the connections between height values defined by $G_c(h)$) of the outermost cylinder ($c = 0$) must guarantee that all paths cross from the upper half of the cylinder to the lower half of the cylinder; thus, the graph of the topology remains fully connected and the bitwise addressing scheme functions properly. Inner cylinders must also be divided into $2c$ fully connected and distinct subgraphs, depending upon the cylinder. Only the final level or cylinder ($c = C-1$) may contain connections between nodes of the same height. The cylindrical crossing must ensure that destinations can be addressed in a binary tree-like configuration, similar to other binary banyan networks.

Addressing within the data vortex architecture is entirely distributed and bitwise, similar to other banyan architectures: as a packet progresses inward, each successive bit of the binary address is matched to the destination. Each cylinder tests only one bit (except for the innermost one); half of the height values permit ingress for 1 values, and half for 0 values, arranged in a banyan binary tree configuration. Within a given cylinder c , nodes at all angles at a particular height (i.e., (c, h, a)) match the same $c + 1$ st significant bit value, while the paths guarantee preservation of the c most significant address bits. Thus, with each ingress to a successive cylinder, progressively more precision is guaranteed in the destination address. Finally, on the last cylinder $c = C-1$, each node in the angular dimension is assigned a least significant value in the destination address so that the packets circulate within that cylinder until a match is found for the last $\sim \log_2(A)$ bits (so-called angle-resolution addressing) [8].

The data vortex all-optical packet switch is the result of a unique effort towards developing a high-performance network architecture designed specifically for photonic media. The overall goals were to produce a practical architecture that leveraged wavelength division multiplexing (WDM) to attain ultra-high bandwidths

and reduce routing complexity, while maintaining minimal time-of-flight latencies by keeping packets in the optical domain and avoiding conventional buffering [7]. In keeping with these objectives, a functional prototype of a 12-port data vortex was implemented and demonstrated [8]. Physical layer scalability was analyzed and demonstrated in [12, 13], and further experimental studies of the optical dynamic range and packet format flexibility were performed [14, 15]. Sources of signal degradation in the data vortex were investigated in [16, 17] and data resynchronization and recovery was achieved using a source synchronous embedded clock in [18]. Extensible and transparent packet injection modules and optical packet buffers for the data vortex were presented in [19]. Alternative architectural implementations and performance optimizations were explored in [20–23].

9.3 SPINet

Based on an indirect multistage interconnection network (MIN) topology, SPINet (Scalable Photonic Integrated Network) [24], designed to be implemented using photonic integration technology, exploits WDM to simplify the network design and provide very high bandwidths. SPINet does not employ buffering, instead resolves contention by dropping contending messages. A novel physical-layer acknowledgment protocol provides immediate feedback, notifying the terminals whether their messages are accepted, and facilitates retransmissions when necessary in a manner resembling that in traditional multiple-access media.

A SPINet network is composed of 2×2 SOA-based bufferless photonic switching nodes [25, 26]. The specific topology can vary between implementations, and switching nodes of higher radices can be used if they are technologically available. The network is slotted and synchronous, and messages have a fixed duration. The minimal slot duration is determined by the round-trip propagation time of the optical signal from the compute nodes to the ports of the network. A slot time of 100 ns can, therefore, accommodate a propagation path of nearly 20 m.

A possible topology for SPINet is the Omega, an example of a binary banyan topology [4]. An $N_T \times N_T$ Omega network consists of $N_S = \log(N_T)$ identical stages. Each stage consists of a perfect-shuffle interconnection followed by $N_T/2$ switching elements, as Fig. 9.4a shows. In the Omega network, each switching node has four allowed states (straight, interchange, upper broadcast, and lower broadcast). In this implementation, we have modified the switching nodes by removing the broadcast states and introducing four new states (upper straight, upper interchange, lower straight, lower interchange). In these four states, the node passes data from only one input port to an output port and drops the data from the other port (see Fig. 9.4b).

At the beginning of each slot, any terminal may start transmitting a message, without a prior request or grant. The messages propagate in the fibers to the input ports of the network and are transparently forwarded to the switching nodes of the first stage. At every routing stage when the leading edges of the messages are

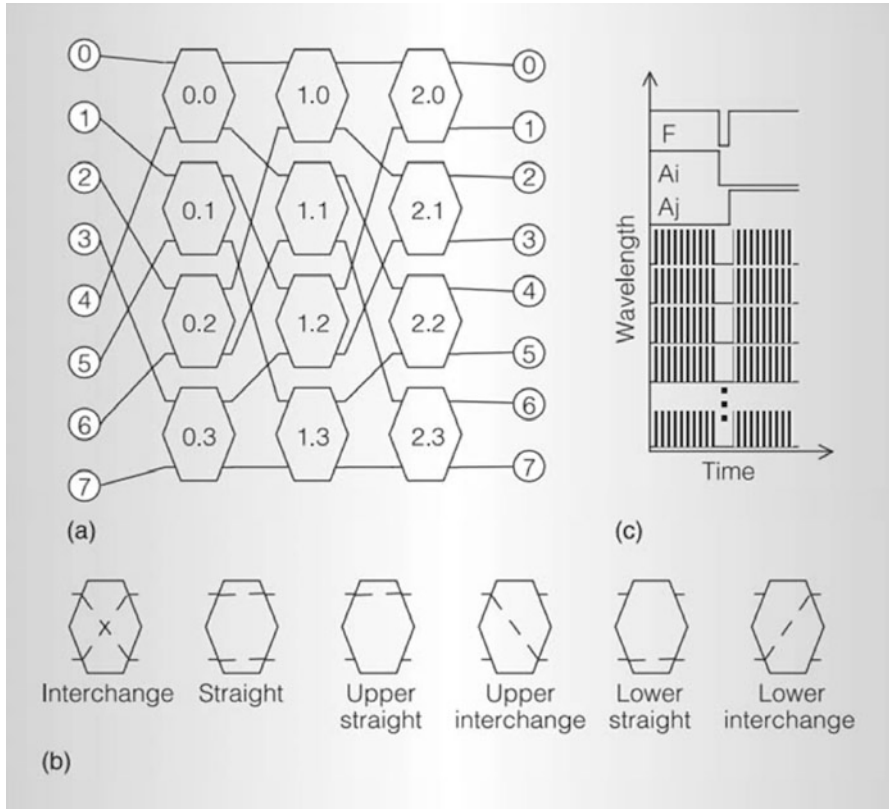


Fig. 9.4 An 8×8 Omega network (a); switching nodes' six states (b); and wavelength-parallel messages (c). Header bits and payload are encoded on dedicated wavelengths [25]

received from one or both input ports, a routing decision is made, and the messages continue to propagate to their requested output ports. In the case of output-port contention in a switching node, the network drops one of the contending messages. The choice of which message to drop can be random, alternating, or priority-based. Because the propagation delay through every stage is identical, all the leading edges of the transmitted messages reach all the nodes of each stage at the same time.

The nodes' switching states, as determined by leading edges, remain constant throughout the duration of the message, so the entire message follows the path acquired by the leading edge. Because the propagation delay through the network, which is ideally implemented via integrated photonics, is very short compared to the duration of the messages, the messages stretch across the PIC, effectively creating transparent lightpaths between the inputs and outputs. When the messages reach the output modules, they are transparently forwarded on the output fibers to the appropriate terminals; at the same time, the destination terminal generates an

acknowledgment optical pulse and sends it on the previously acquired lightpath in the opposite direction. Because the node's switching elements preserve their states and support bidirectional transmission, the source terminal receives the acknowledgment pulse, which serves as confirmation of the message's successful reception.

When the slot time is over, all terminals cease transmission simultaneously, the switching nodes reset their switching states, and the system is ready for a new slot. The slot duration is set to ensure that the acknowledgment pulses are received at the source terminals before the slot ends. Hence, before the beginning of the next slot, every terminal knows whether its message was accepted; when necessary, it can choose to immediately retransmit the message.

Leveraging ultralow-latency signal propagation through the network, SPINet eliminates the need for central scheduling, instead employing the distributed computing power of the switching nodes to produce an input–output match at every slot. This process of implicit arbitration enables scalability to large port counts without burdening a central arbiter with computations of complex maximal matches. Because SPINet uses blocking topologies to reduce hardware complexity, the network's utilization is lower than that of a traditional maximum-matched nonblocking network (as in switching fabrics for high-performance Internet routers). Techniques that exploit the properties of integrated photonics can increase utilization by adding a small number of stages.

SPINet uses the wavelength domain to facilitate a routing mechanism in the switching nodes that can instantly determine and execute the routing decision upon receiving the leading edges, without any additional information exchange between the switching nodes. The mechanism also maintains a constant switching state for the duration of the messages. The messages are constructed in a wavelength-parallel manner, similar to that used in the data vortex architecture, trading off a part of the enormous bandwidth of optical fibers to simplify the switching-node design. As Fig. 9.4c shows, the routing header and the message payload are encoded on separate wavelengths and are received concurrently by the nodes. The header consists of a frame bit that denotes the message's existence and a few address bits. Each header bit is encoded on a dedicated wavelength and remains constant throughout the message duration. When a binary network is used, a single address bit is required at every stage, and therefore the number of wavelengths required for address encoding is the number of routing stages in the network, or \log_2 of the number of ports. The switching nodes' routing decisions are based solely on the information extracted from the optical header, as encoded by the source. The switching nodes neither exchange additional information nor add any to the packet. The payload is encoded on multiple wavelengths at the input terminal, which segments it and modulates each segment on a different wavelength, using the rest of the switching band. A guard time is allocated before payload transmission, accommodating the SOA switching time, clock recovery in the payload receivers, and synchronization inaccuracies.

9.4 Networking Challenges in the Data Center

The aforementioned networks, in addition to other implementations of all-optical networks, are capable of the ultra-high bandwidths and low power densities necessary for enabling the continued scaling of large cluster computing systems. However, these systems represent a wide range of computing classes, ranging from highly specialized designs to commodity and cost-driven computing environments. For example, the increasing popularity of cloud-based services continues to drive the creation of larger and more powerful data centers. As these services scale in both number and size, applications oftentimes extend well beyond the boundaries of a single rack of servers. Moreover, the continued advancements in computational density enabled by increasing parallelism in contemporary microprocessors and chip multiprocessors (CMP) have resulted in substantial off-chip communication requirements. As a result, in a similar fashion to their supercomputing counterparts, the performance of modern data centers are becoming increasingly communication-bound, requiring upwards of hundreds of thousands of ports supporting petabits per second of aggregate bandwidth [27].

However, due to the superlinear costs associated with scaling the bandwidth and port density of conventional electronic switches, network oversubscription is common practice. Consequently, data-intensive computations become severely bottlenecked when information exchange between servers residing in separate racks is required. Unlike high-performance computing systems, the very nature of the data center as a pool of centralized computational resources gives rise to significant application heterogeneity. The resultant workload unpredictability produces significant traffic volatility, precluding the efficacy of static capacity engineering in these oversubscribed networks.

Energy efficiency has also emerged as a key figure-of-merit in data center design [28]. The power density of current electronic interconnects is already prohibitively high—on the order of hundreds of kilowatts—and continues to grow exponentially. As it stands, the power consumption of a single switch located in the higher network tiers can reach upwards of tens of kilowatts when also considering the dedicated cooling systems required. Moreover, measurements on current data center deployments have recorded average server utilization as low as 30% [29], indicating significant wasted energy due to idling hardware starved for data.

As a result, alleviating inter-rack communication bottlenecks has become a critical target in architecting next-generation data centers. The realization of a full bisection-bandwidth, “all-servers-equidistant” interconnection network will not only accelerate the execution of large-scale distributed applications, but also significantly reduce underutilization by providing sufficient network performance to ensure minimal idling of power-hungry compute elements. In addition, the increased connectivity between computing and storage resources located throughout the data center will yield more flexibility in virtualization, leading to further enhancements in energy efficiency.

Despite continued efforts from merchant silicon providers towards the development of application-specific integrated circuits (ASICs) for high-performance switches and routers, the sheer scale of the data center and the relentless demand from data-intensive applications for increased connectivity and bandwidth continues to necessitate oversubscription in hierarchical purely packet-switched electronic interconnection networks. While there have been significant efforts focused on architectural and algorithmic approaches towards improving the overall performance of data center networks [30, 31], these proposals are ultimately constrained by the fundamental limitations imposed by the underlying electronic technologies.

Recently, in the context of production data center environments, there have been a number of efforts exploring the viability of circuit-switched optics as a cost-effective means of providing significant inter-rack bandwidth. Helios [32] and C-Through [33] represent two data center network architectures proposing the use of micro-electro-mechanical system (MEMS)-based optical switches. By augmenting existing oversubscribed hierarchical electronic packet-switched (EPS) networks, each implementation realizes a hybrid electronic/optical architecture that leverages the respective advantages of each technology. These initial proposals have successfully demonstrated the potential for utilizing photonic technologies within the context of data center traffic to provide significantly increased network capacities while achieving reduced complexity, component cost, and power in comparison with conventional electronic network implementations. Another network architecture, called Proteus, combines both wavelength-selective switching and space switching to provide further granularity in the capacity of each optical link, varying between a few gigabits per second to a hundreds of gigabits per second on-demand [34].

While network traffic is characteristically unpredictable due to application heterogeneity, communication patterns where only a few top-of-the-rack (ToR) switches are tightly coupled with long, extended data flows have been observed in production data centers [35]. Therefore, the utility of the aforementioned architectures are reliant on the inherent stability of traffic patterns within such systems. Nevertheless, further bandwidth flexibility remains a key target for future data center networks as applications require even higher capacities with increased interconnectivity demands. When studied under the communication patterns imposed by a richer, more representative set of realistic applications, the efficacy of architectures utilizing purely commercial MEMS-based switches, which are limited to switching times on the order of milliseconds, becomes ambiguous [36].

9.5 Conclusions

Traditional supercomputers usually employ specialized top-of-the-line components and protocols developed specifically to support highly orchestrated distributed workloads that support massively parallel, long-running algorithms developed to solve complex scientific problems. As such, these applications impose stringent latency requirements on processor-to-processor and processor-to-memory transactions, which represent the major bottleneck in these highly specialized systems.

On the other hand, data centers, which are primarily deployed by enterprises and academic institutions, predominantly run general-purpose user-facing cloud-based applications and are largely composed of commodity components. The majority of the messages being passed across a data center consist of very short random transactions. However, there typically exist a small number of long extended flows, which account for a majority of the data being transmitted through the network. Furthermore, traffic at the edges of the network is often bursty and unpredictable, leading to localized traffic hotspots across the system that leads to network congestion and underutilization. While it is apparent that bandwidth restrictions result in significant performance degradation in data centers, the latency requirements of these systems are relatively relaxed in comparison with that of supercomputers.

Consequently, the application demands and traffic patterns characteristic of these systems vary widely, resulting in highly variegated network requirements. Therefore, in addition to the improved capacity, power consumption and bandwidth-distance product delivered by a photonic interconnect medium, capacity flexibility is a key requirement for enabling future optically interconnected high-performance data centers and supercomputers.

References

1. Agrawal GP(2002) Fiber-optic communication systems. Wiley, New York
2. Gnauck AH, Charlet G, Tran P, Winzer PJ, Doerr CR, Centanni JC, Burrows EC, Kawanishi T, Sakamoto T, Higuma K (2008) 25.6-Tb/s WDM transmission of polarization-multiplexed RZ-DQPSK signals. *IEEE J Lightwave Technol* 26:79–84
3. Benner AF, Ignatowski M, Kash JA, Kuchta DM, Ritter MB (2005) Exploitation of optical interconnects in future server architectures. *IBM J Res Dev* 49(4/5):755–775
4. Dally WJ, Towles B (2004) Principles and practices of interconnection networks. Morgan Kaufmann, San Francisco
5. Kash JA, Benner A, Doany FE, Kuchta D, Lee BG, Pepeljugoski P, Schares L, Schow C, Taubenblatt M (2011) Optical interconnects in future servers. In: Optical fiber communication conference, Paper OWQ1. <http://www.opticsinfobase.org/abstract.cfm?URI=OFC-2011-OWQ1>
6. Ramaswami R, Sivarajan KN (2002) Optical networks: a practical perspective, 2nd edn. Morgan Kaufmann, San Francisco
7. Liboiron-Ladouceur O, Shacham A, Small BA, Lee BG, Wang H, Lai CP, Biberman A, Bergman K (2008) The data vortex optical packet switched interconnection network. *J Lightwave Technol* 26 (13):1777–1789
8. Shacham A, Small BA, Liboiron-Ladouceur O, Bergman K (2005) A fully implemented 12×12 data vortex optical packet switching interconnection network. *J Lightwave Technol* 23(10):3066–3075
9. Yang Q, Bergman K, Hughes GD, Johnson FG (2001) WDM packet routing for high-capacity data networks. *J Lightwave Technol* 19(10):1420–1426
10. Yang Q, Bergman K (2002) Traffic control and WDM routing in the data vortex packet switch. *IEEE Photon Technol Lett* 14(2):236–238
11. Yang Q, Bergman K (2002) Performance of the data vortex switch architecture under nonuniform and bursty traffic. *J Lightwave Technol* 20(8):1242–1247

12. Liboiron-Ladouceur O, Small BA, Bergman K (2006) Physical layer scalability of a WDM optical packet interconnection network. *J Lightwave Technol* 24(1):262–270
13. Liboiron-Ladouceur O, Bergman K, Boroditsky M, Brodsky M (2006) Polarization-dependent gain in SOA-Based optical multistage interconnection networks. *IEEE J Lightwave Technol* 24(11):3959–3967
14. Small BA, Lee BG, Bergman K (2006) Flexibility of optical packet format in a complete 12×12 data vortex network. *IEEE Photon Technol Lett* 18(16):1693–1695
15. Small BA, Kato T, Bergman K (2005) Dynamic power consideration in a complete 12×12 optical packet switching fabric. *IEEE Photon Technol Lett* 17(11):2472–2474
16. Small BA, Bergman K (2005) Slot timing consideration in optical packet switching networks. *IEEE Photon Technol Lett* 17(11):2478–2480
17. Lee BG, Small BA, Bergman K (2006) Signal degradation through a 12×12 optical packet switching network. In: European conference on optical comm., We3.P.131, pp 1–2, 24–28. doi: 10.1109/ECOC.2006.4801324 <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4801324&isnumber=4800856>
18. Liboiron-Ladouceur O, Gray C, Keezer DC, Bergman K (2006) Bit-parallel message exchange and data recovery in optical packet switched interconnection networks. *IEEE Photon Technol Lett* 18(6):770–781
19. Shacham A, Small BA, Bergman K (2005) A wideband photonic packet injection control module for optical packet switching routers. *IEEE Photon Technol Lett* 17(12):2778–2780
20. Shacham A, Bergman K (2007) Optimizing the performance of a data vortex interconnection network. *J Opt Networking* 6(4):369–374
21. Liboiron-Ladouceur O, Bergman K (2006) Hybrid integration of a semiconductor optical amplifier for high throughput optical packet switched interconnection networks. *Proc SPIE* 6343–121, doi: 10.1117/12.708009
22. Liboiron-Ladouceur O, Bergman K (2006) Bistable switching node for optical packet switched networks. In: Proceedings 19th Annual Meeting of the IEEE Lasers and Electro-Optics Society (LEOS), 2006. Paper WW5, pp 631–632. <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4054342&isnumber=4054019>
23. Yang Q (2005) Improved performance using shortcut path routing within data vortex switch network. *Electron Lett* 41(22):1253–1254
24. Shacham A, Bergman K (2007) Building ultralow latency interconnection networks using photonic integration. *IEEE Micro* 27(4):6–20
25. Shacham A, Lee BG, Bergman K (2005) A scalable, self-routed, terabit capacity, photonic interconnection network. In: Proceedings of 13th Ann. IEEE Symp. High-Performance Interconnects (HOTI 05). IEEE CS Press, pp 147–150. doi: 10.1109/CONNECT.2005.6 <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1544590&isnumber=32970>
26. Shacham A, Lee BG, Bergman K (2005) A wideband, non-blocking, 2x2 switching node for a SPINet network. *IEEE Photonic Technol Lett* 17(12):2742–2744
27. Vahdat A, Al-Fares M, Farrington N, Mysore RN, Porter G, Radhakrishnan S (2010) Scale-out networking in the data center. *IEEE Micro* 30(4):29–41
28. Abts D, Marty MR, Wells PM, Klausler P, Liu H (2010) Energy proportional datacenter networks. In: Proceedings of 37th annual international symposium on computer architecture (ISCA'10), pp 338–347 ACM, New York, NY, USA <http://doi.acm.org/10.1145/1815961.1816004>
29. Meisner D, Gold BT, Wenisch TF (2009) PowerNap: eliminating server idle power. In: Proceedings of the 14th international conference on architectural support for programming languages and operating systems (ASPLOS'09), New York, NY, USA pp 205–216. <http://doi.acm.org/10.1145/1508244.1508269>
30. Al-Fares M et al (2008) A scalable, commodity data center network architecture. *SIGCOMM Comp Comm Rev* 38(4):63–74
31. Greenberg A et al (2009) VI2: a scalable and flexible data center network. *SIGCOMM Comp Comm Rev* 39(4):51–62

32. Farrington N, Porter G, Radhakrishnan S, Bazzaz HH, Subramanya V, Fainman Y, Papen G, Vahdat A (2010) Helios: a hybrid electrical/optical switch architecture for modular data centers. In: SIGCOMM '10 proceedings of the ACM SIGCOMM 2010 conference on SIGCOMM. ACM, New York, pp 339–350
33. Wang G, Andersen DG, Kaminsky M, Papagiannaki K, Ng TE, Kozuch M, Ryan M (2010) c-Through: part-time optics in data centers. In: SIGCOMM '10 proceedings of the ACM SIGCOMM 2010 conference on SIGCOMM. ACM, New York, pp 327–338
34. Singla A, Singh A, Ramachandran K, Xu L, Zhang Y (2010) Proteus: a topology malleable data center networks. In: Hotnets '10 proceedings of the ninth ACM SIGCOMM workshop on hot topics in networks. ACM, New York, article 8
35. 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, Barcelona, Spain, 21 August 2009. WREN '09. ACM, New York, pp 65–72
36. Bazzaz HH, Tewari M, Wang G, Porter G, Ng TSE, Andersen TG, Kaminsky M, Kozuch MA, Vahdat A (2011) Switching the optical divide: fundamental challenges for hybrid electrical/optical datacenter networks. In: Proceedings of SOCC'11: ACM symposium on cloud computing, Cascais, Portugal, Oct 2011

Index

A

ACK messages, 99
 adaptive timeout algorithm, 109
 format and semantics, 107
 pending packets, 105
 piggybacking, 107
 schematic representation, 105
 source adapter, 108
ACK overhead, 107–109, 112
Active optical cables (AOCs), 70
All-optical networks
 ACKs, 99
 architecture
 end-to-end reliability, 98
 minimally-sized switching nodes, 98
 multistage interconnection networks (MINs), 98–99
 time-division-multiplexing (TDM), 97
computer simulation models
 ACK overhead, 112
 impact of retransmissions, 110–111
 packet delay vs. incoming load, 113
 permutation traffic, 113–114
delay performance, 101–102
dynamic scheduling, 99
end-to-end reliable packet delivery
 ACK messages, 105–106
 ACK overhead reductions, 107–109
 new injections vs. retransmissions vs. standalone ACKs, 106–107
 retransmission timers, 109–110
 ingress path and egress path, 100
 Omega network, 100–101
 optical transmission, 96–97
 prescheduled and speculative injections, 100

 requirements, 96
 speculative packet injections, 103–105
 switching elements, 100–101
 TDM prescheduled packet injections, 102–103
 throughput performance, 102
All-optical OFDM (AO-OFDM), 121
Application of DCN, 33–34
Architecture
 block diagram, 3–4
 fat-tree 2-Tier/3-Tier, 4
 fault tolerant, 4
 high power consumption, 5
 limitations, 20
 multiple store-and-forward processing, 5
 social networking, 18
 using scale-out model, 19–20
 using traditional scale-up approach, 18–19
Arrayed waveguide grating (AWG), 138. *See* Cyclic arrayed waveguide grating (CAWG)
Arrayed waveguide grating router (AWGR), 138–139
Average packet delay performance
 OFF/ON = 10, 130
 OFF/ON = 50, 131

B

Balanced workloads (BWs), 55
Banyan network. *See* Omega network
BCube, 151
Bit-transport energy (BTE), 73–74
Butterfly, 37, 72, 158. *See also* Omega network

C

- CAWG. *See* Cyclic arrayed waveguide grating (CAWG)
- Cisco CRS-1 router, 151
- Clock data recovery (CDR), 28
- Cloud computing
 - application, 6
 - DENS methodology
 - communicational potential, 58
 - computing server selection, 59
 - load and communicational potential, 61
 - metric behavior, 58
 - queue selection, 59
 - energy efficiency, 49
 - GreenCloud simulator structure, 51
 - hardware components and energy models, 52–54
 - network congestion, 57
 - software as a service (SaaS), 32
 - three-tier
 - aggregation switches, 51
 - architecture, 49–50
 - gigabit Ethernet (GE) transceivers, 49
 - utility computing, 333
 - workloads
 - CIW, 54
 - communicational component, 55
 - computing component, 55
 - distribution of energy consumption, 55–56
 - DIW, 54–55
 - energy consumption for types of, 55–56
 - grid computing applications, 54
- Compound annual growth rate (CAGR), 7
- Computationally intensive workloads (CIWs), 54
- Computing servers, 52–53
- Cyclic arrayed waveguide grating (CAWG), 121–122

D

- Data intensive cloud computing data centers. *See* Cloud computing
- Data-intensive workloads (DIWs), 54–55
- Data vortex
 - ingression and deflection fibers, 158
 - multi-wavelength striped packet format, 157
 - petabit bufferless optical switch, 152
 - physical layer scalability, 160
 - sorting pattern, 159
 - switching node design, 157
 - topology, 158
- DCell, 151

Delta networks, 98–99

- DENS methodology
 - communicational potential, 58
 - computing server selection, 59
 - load and communicational potential, 61
 - metric behavior, 58
 - queue selection, 59
- DFB lasers. *See* Distributed feedback (DFB) lasers
- Distributed feedback (DFB) lasers, 22–23
- dOrthogonal frequency division multiple access, PSD, 121
- DPillar, 151
- Dynamic subcarrier allocation (DSA), 129
- Dynamic voltage/frequency scaling (DVFS), 52, 55–56

E

- Electrical arbiter (EARB), 84
- Electrical I/O roadmaps, 74
- Electronic switch architecture
 - arbitration, 79
 - components, 78
 - distributed high-radix switch, 78–79
 - head-of-line (HOL) blocking, 78
- End-to-end reliability
 - ACK messages, 105–106
 - ACK overhead reductions, 107–109
 - network architecture, 98
 - new injections vs. retransmissions vs. standalone ACKs, 106–107
 - retransmission timers, 109–110
- Energy efficiency and proportionality, DCN, 36–37

F

- Fiber cross-connect (FXC), 120, 122, 125
- Fixed subcarrier allocation (FSA), 129
- Frame-based switching, 141
- Frame scheduling algorithm, 144

G

- GreenCloud simulator structure, 51

H

- Head-of-line (HOL) packet, 102, 107
- High-performance computing (HPC), 95–96
- HyPaC, 151
- HyperX topology. *See* Butterfly

J

Jobs. *See* Workloads

K

k-ary n-fly networks, 98–99

L

Lensed-integrated surface emitting DFB laser, 23

Line card architecture, 142

M

Mach–Zehnder interferometer (MZI), 139

Microprocessors, 35

MIMO OFDM DCN

architecture, 122–123

CAWG, 121–122

features

control, 125

cost, 125

fast switching, 125

fine granularity switching, 124

flexible bandwidth allocation and sharing, 124

flexible modulation format and data rate, 124–125

low and uniform latency, 125

MIMO switching, 124

no guard band, 125

power consumption, 125

scalable, 125

modulation formats, 126

one-tap equalization, 120

optical core router, 126

optical spectra, WDM signals, 127

parallel signal detection (PSD), 121

performance evaluation

OFF/ON = 10, average packet delay, 130

OFF/ON = 50, average packet delay, 131

simulation model and traffic assumptions, 128–129

subcarrier allocation algorithms, 129–130

PSD receiver performance, 127–128

RF spectra, OFDM signals, 126–127

signal generation, 120–121

ToR bypass, 123–124

types, 121

wavelength and subcarrier assignments, 126

Multi-chip scheduler, 146–147

Multi-core fiber (MCF) cable, 24

Multi-mode fiber (MMF), 26

Multiple-input multiple-output (MIMO). *See* MIMO OFDM DCN

Multiple wavelength-selective switches (WSS). *See* Fiber cross-connect (FXC)

Multiplexing techniques

optical link speed, 23

optical orthogonal frequency division multiplexing (O-OFDM), 24

space division multiplexing (SDM), 24

wavelength division multiplexing (WDM), 25

Multistage interconnection networks (MINs), 98–99

Multi-tiered architecture, 136

N

Network bottleneck, 35–36

Networks on chip

germanium-based photodetectors, 41

off-chip laser, 41

ring resonator modulators and filters, 40

silicon photonics optical modulator, 40

waveguide loss characteristics, 39

Network switches and links, 52

Network traffic characteristics

applications, 5

compound annual growth rate (CAGR), 7

concurrent traffic flows, 6

link utilization, 6

packet size, 6

server datarate forecast, 7

traffic flow locality, 5–6

traffic flow size and duration, 6

O

OFDM. *See* MIMO OFDM DCN

Omega network, 100–101

header bits and payload, 161–162

mechanism, 162

switching nodes, six states, 160–161

wavelength-parallel messages, 160–161

Operational expenses (OPEX), 47

OPNET modeler, 128

Optical crossbar, 81–82

Optical fibers, 26

- Optical interconnects
 - application-specific integrated circuits (ASICs), 164
 - bandwidth and scalability
 - distributed feedback (DFB) lasers, 22–23
 - multiplexing, 23–25
 - optical fibers, 26
 - silicon photonics, 23
 - vertical cavity surface emitting lasers (VCSELs), 22
 - communication patterns, 164
 - data vortex architecture
 - ingression and deflection fibers, 158
 - multi-wavelength striped packet format, 157
 - physical layer scalability, 160
 - sorting pattern, 159
 - switching node design, 157
 - topology, 158
 - energy efficiency, 163
 - energy proportionality, 27–28
 - inter-rack communication bottlenecks, 163
 - network interface cards (NICs), 21
 - networks on chip
 - germanium-based photodetectors, 41
 - off-chip laser, 41
 - ring resonator modulators and filters, 40
 - silicon photonics optical modulator, 40
 - waveguide loss characteristics, 39
 - Proteus, 164
 - SPINet (*see* Scalable Photonic Integrated Network (SPINet))
 - system level networks
 - hybrid optical/electrical network, 38
 - scale out solutions, 38
 - timing constraints, 39
 - WDM transceivers, 20–21
 - Optical networks evolution, 9–10
 - Optical OFDM (O-OFDM), 121
 - Optical shared memory supercomputer
 - interconnect system (OSMOSIS), 152
 - Optical switch
 - architecture, 80–81
 - fabric
 - arrayed waveguide grating (AWG), 138
 - arrayed waveguide grating router (AWGR), 138–139
 - Clos-based switch fabric, 139
 - schematic representation, 137–138
 - 4 × 4 switch fabric, 139
 - tunable wavelength converter (TWC), 139
 - petabit bufferless (*see* Petabit bufferless optical switch)
 - Optical to electronic (OE) domain, 70
 - Optimal resource utilization (ORU), 129
 - Orthogonal frequency division multiple access (OFDM). *See* MIMO OFDM DCN
- P**
- Packet scheduler design
 - architecture with three-stage arbitration, 144–145
 - CM assignment logic, 145–146
 - request filter structure, 145
 - Parallel ribbon fiber cable, 24
 - Parallel signal detection (PSD), 121, 127
 - Petabit bufferless optical switch
 - architecture
 - address management, 141–142
 - arrayed waveguide grating (AWG), 138
 - arrayed waveguide grating router (AWGR), 138–139
 - Clos-based switch fabric, 139
 - configuration, 140–141
 - frame-based switching, 141
 - line card design, 142
 - schematic representation, 137–138
 - 4 × 4 switch fabric, 139
 - tunable wavelength converter (TWC), 139
 - data vortex, 152
 - flatten DCN through single switch, 137
 - Helios and HyPaC, 151
 - multi-tiered architecture, 136
 - optical shared memory supercomputer interconnect system (OSMOSIS), 152
 - performance evaluation
 - delay performance, 149–151
 - scalability and throughput, 149
 - Portland, 150
 - racks and wiring, 147–148
 - scalability problems, 136
 - scheduling algorithm
 - frame, 144
 - multi-chip scheduler and inter-chip connections, 146–147
 - packet scheduler design, 144–146
 - problem statement, 143
 - server-centric, 151
 - timing and frame alignment, 148
 - TrueWay switch and Cisco CRS-1 router, 151

- Photonic BTE, 73
 - Photonic roadmaps
 - interchip point-to-point DWDM link, 75
 - losses and power, 76
 - wavelength-specific switch, 75
 - Photonics role in DCN
 - active optical cables (AOC), 70
 - bit transport energy (BTE), 73
 - dense wavelength division multiplexing (DWDM), 72–73
 - electrical I/O roadmaps, 74
 - high-radix switches, 72
 - metrics, 68
 - modulation rates and transport lengths, 69
 - photonic roadmaps, 75–76
 - static tuning power, 73
 - switch architecture (*see* Switch microarchitecture)
 - total packet latency, 71
 - warehouse scale computers (WSCs), 68–69
 - Point-to-point vs. all-optical interconnects, 10
 - Portland, 150
 - Power consumption
 - distribution, 8
 - impact on environment, 8
 - and performance, bandwidth requirements, 8
 - requirements for interconnects, 8–9
 - Power usage effectiveness (PUE), 37
 - Prescheduled packet injections, 102–103
 - Private enterprise data centers, 5
 - PSD. *See* Parallel signal detection (PSD)
- R**
- Roadmaps
 - electrical I/O, 74
 - photonic
 - interchip point-to-point DWDM link, 75
 - losses and power, 76
 - wavelength-specific switch, 75
 - Role of photonics in DCN
 - active optical cables (AOC), 70
 - bit transport energy (BTE), 73
 - dense wavelength division multiplexing (DWDM), 72–73
 - electrical I/O roadmaps, 74
 - high-radix switches, 72
 - metrics, 68
 - modulation rates and transport lengths, 69
 - photonic roadmaps, 75–76
 - static tuning power, 73
 - switch architecture (*see* Switch microarchitecture)
 - total packet latency, 71
 - warehouse scale computers (WSCs), 68–69
- S**
- Scalable photonic integrated network (SPINet)
 - header bits and payload, 161–162
 - mechanism, 162
 - switching nodes, six states, 160–161
 - wavelength-parallel messages, 160–161
 - Server-centric DCN, 151
 - Silicon photonics, 23
 - Single mode fiber (SMF), 26
 - Software as a Service (SaaS), 32
 - Space division multiplexing (SDM), 24
 - Speculative packet injections, 103–105
 - SPINet. *See* Scalable photonic integrated network (SPINet)
 - Standalone ACK messages, 108–109, 112
 - Storage-area network, 32
 - Switch microarchitecture
 - arbitration, 83–84
 - electronic switch arbitration, 79
 - components, 78
 - distributed high-radix switch, 78–79
 - head-of-line (HOL) blocking, 78
 - optical crossbar, 81–82
 - optical switch, 80–81
 - packaging constraints, 85
 - radix independent resource parameters, 77–78
 - thermal tuning of rings, 82–83
- T**
- Thermal tuning of rings, 82–83
 - Three-tier DCN
 - aggregation switches, 51
 - architecture, 49–50
 - gigabit Ethernet (GE) transceivers, 49
 - Time-division-multiplexing (TDM), 97
 - Top-of-Rack (ToR) switch, 53
 - Total packet latency, 71
 - Traffic characteristics
 - applications, 5
 - compound annual growth rate (CAGR), 7
 - concurrent traffic flows, 6
 - link utilization, 6

Traffic characteristics (*cont.*)

- packet size, [6](#)
- server datarate forecast, [7](#)
- traffic flow locality, [5–6](#)
- traffic flow size and duration, [6](#)

TrueWay switch, [151](#)

Tunable wavelength converter (TWC),
[139](#)

Twisted pair, [52](#)

U

University campus data centers, [5](#)

Utility computing, [33](#)

V

VCSELS. *See* Vertical cavity surface emitting lasers (VCSELS)

Vertical cavity surface emitting lasers
(VCSELS), [22](#)

Virtual machine (VM), [141](#)

W

Warehouse scale computers (WSCs), [68–69](#)

Wavelength division multiplexing (WDM), [25](#)

Workload intensive cloud computing data
centers. *See* Cloud computing

Workloads

CIW, [54](#)

communicational component, [55](#)

computing component, [55](#)

distribution of energy consumption,
[55–56](#)

DIW, [54–55](#)

energy consumption for types of, [55–56](#)

grid computing applications, [54](#)