Opportunities and Challenges for Photonics in Next Generation Data Centers

Clint Schow
Department of Electrical and Computer Engineering
The University of California Santa Barbara
schow@ece.ucsb.edu
Outline

• Background: Interconnects and Technologies

• Optics In Today’s Data Centers
  – Point-to-point interconnects
  – Multiple technologies for multiple purposes

• Opportunities
  – Photonic I/O
  – Photonic routing and switching

• Path Forward: Large-Scale Electronic/Photonic Integration

• Closing Thoughts
Historically Two Fiber Optics Camps: Datacom and Telecom

Telecom (10’s – 1000’s of km)
- Expensive to install fiber over long distances
  - Single-mode fiber (SMF)
  - Wavelength Division Multiplexing (WDM)
- Cost of transceivers a secondary concern
- Performance is the primary objective

Data Centers

Datacom (100’s of meters)
- Cost of everything (transceivers, fibers, connectors) is the biggest factor
  - Multi-mode fiber (MMF)
  - Transceivers are commodities

Relative core size:

MMF >30X SMF

<table>
<thead>
<tr>
<th>SMF</th>
<th>MMF</th>
</tr>
</thead>
<tbody>
<tr>
<td>9µm</td>
<td>50µm</td>
</tr>
</tbody>
</table>
Technologies for Data Center Interconnects

**VCSELs** (Vertical Cavity Surface Emitting Lasers)
Multi-mode fiber (MMF), Polymer Waveguides, Multicore fiber

**Si Photonics**
Single-mode fiber (SMF), Wavelength Division Multiplexing (WDM)

October 22, 2015
C. Schow, UCSB
The Old Days: 2012

Trickle-Down: HPC drives development of highest performance components that are later picked up by commercial servers

IBM, presented at Optical Interconnects Conference, 2012
The Promise of A Machine with 2,000,000 VCSELs

Fiber Backplane Density Pushed to the Limit

IBM Power 775: Pushing the Limits

Need more BW/fiber: WDM, multicore fiber

Optics in HPC: IBM Sequoia

96 IBM Blue Gene/Q Racks
20.013 Pflops Peak ... 1.572M Compute Cores ... ~8MW ... 2026 Mflops/Watt

• HPC requires technologies optimized for short reach ~50m
VCSELs: Efficient and Fast

32-nm CMOS-Driven Link

Wall-plug efficiency: 1pJ/bit at 25 Gb/s

25 Gb/s, 24mW

35 Gb/s, 95mW

TX out

RX out

Rethinking Equalization: Optimizing the Performance of Complete Links

First Implemented in CMOS (2011)


Next opportunity: Si photonic WDM links

Ported to SiGe (2011)

First 40 Gb/s VCSEL Links (2012)

Higher data rates at better efficiency

No Equalization

With Equalization

Link Output (40 Gb/s)

71 Gb/s (2015)
And Then The Cloud Rolled In
Growth in Cloud Data Centers

Cloud is the growth market

75% of traffic is within the data center
Huge Growth in the Cloud (Microsoft)

Source: D. Maltz, Microsoft, OFC 2014
Traditional hierarchical networks grew out of campus/WAN installations.

Good for North-South traffic

Bad for East-West traffic (75% of data center traffic)

See L. A. Barroso and U. Hölzle, “The Datacenter as a Computer—An Introduction to the Design of Warehouse-Scale Machines,”
Data Center Hardware

Lots of Racks

Top of Rack (ToR) switch

Servers

Photos of Facebook data centers found on Google images
Cloud Data Centers (Microsoft)

WDM identified as path to lower cost if transceivers are cheap

Source: D. Maltz, Microsoft, OFC 2014
Largest Data Centers, 2015

Source: Computer Business Review (cbronline.com)

1. **Range International Information Group** (Langfang, China)
   Area: 6,300,000 Sq. Ft.

2. **Switch SuperNAP** (Nevada, USA)
   Area: 3,500,000 million Sq. Ft.

3. **DuPont Fabros Technology** (Virginia, USA)
   Area: 1,600,000 million Sq. Ft.

4. **Utah Data Centre** (Utah, USA)
   Area: 1,500,000 million Sq. Ft.

5. **Microsoft Data Centre** (Iowa, USA)
   Area: 1,200,000 Sq. Ft.

6. **Lakeside Technology Centre** (Chicago, USA)
   Area: 1,100,000 Sq. Ft.

7. **Tulip Data Centre** (Bangalore, India)
   Area: 1,000,000 Sq. Ft.

8. **QTS Metro Data Centre** (Atlanta, USA)
   Area: 990,000 Sq. Ft.

9. **Next Generation Data Europe** (Wales, UK)
   Area: 750,000 Sq. Ft.

10. **NAP of the Americas** (Miami, USA)
    Area: 750,000 Sq. Ft.

Huge facilities need lots of longer-distance links: 2km becoming the magic number for data centers

October 22, 2015
C. Schow, UCSB
Toward More Connected Data Centers

1) Flatten the Network:
   Higher radix (larger port count) switches
   Electrical or Optical Cores

2) Change the Network:
   Photonic switching
Photonic I/O: Optics to the Chip
Shared Visions: Photonically Connected Chips

Today: Electrical Chip Packaging

IC packages with course BGA/LGA electrical connectors
- Poor scalability, signal integrity
- Reduced system performance
- Reduced system efficiency

Future: Photonic Packages

IC’s: one or more can be 3-D stacks
High-density, ultra-low power photonics
Multi-chip package (BGA, LGA)

Photonic integration must provide more I/O bandwidth at better efficiency
Limitations of Electrical Switches

Mellanox

- **Switch IB**
- **Spectrum**

### Ordering Part Number | Description | Typical Power
--- | --- | ---
MT52236A0-FDCR-E | Switch-IB, 36 Port EDR InfiniBand Switch IC | 83W

<table>
<thead>
<tr>
<th>Ordering Part Number</th>
<th>Description</th>
<th>Typical Power</th>
</tr>
</thead>
</table>
| MT52132A0-FCCR-C | Spectrum, 32 Port Ethernet 100GbE Switch IC (RoHS R6) | 135W

Source: mellanox.com

Broadcom

**Tomahawk**

- 128 x 25Gb/s SerDes = 32 100GbE ports
- 7 Billion transistors

Two primary limitations
- Power (mostly electrical I/O)
- Density

Source: broadcom.com
Density Limits Imposed by Packages, Not ICs

Challenge: Limited chip radix

- Do we have enough package area for all the SERDES needed?
- 2015 – 150 SERDES (3000mm²)
- 2017 – 200/250 SERDES (5000mm²)
- 2020 – end of the road?

Courtesy of M. Laor, Compass Networks
Higher BW Density: Demands Integration at (First-Level) Chip Package

Cross-sectional view of chip module on board

- Electrical Packaging is mature
  - # Signal Pins & BW/pin not increasing at rate of silicon
  - All-electrical packages will not have enough pins or per-pin BW
- Chokepoint is at the module-to-circuit board connection

Electrical I/O: Burning Power to Overcome Packaging Limitations

Electrical Link Example: Backplane

- Backplane: even higher loss than most chip-to-chip links (extra connectors, longer distance). Very hard to scale to higher data rates.
- Why keep pushing it then?
- Ease of use: plug in a new card into the existing legacy backplane and get a speed boost.
- Could be a very cost-effective solution, but power dissipation is an issue, especially at higher data rates.

> 20dB loss @ 5GHz

Tyco 16” Backplane Channel Response

Courtesy A. Rylyakov, OFC Short Course #357
Integrating photonics into the most expensive, constrained, and challenging environment in the system

- Cost
- Reliability
- Thermal
- Power delivery

Move from bulky optics located far away
To highly-integrated optics close to the logic chips
From Discrete Modules to Integrated Chip I/O (Luxtera)

High-Speed Optical Interconnect Evolution II

<table>
<thead>
<tr>
<th>CONTEMPORARY – Today</th>
<th>EMERGING – 2014/15</th>
<th>STRATEGIC DIRECTION – 2018+</th>
</tr>
</thead>
</table>

- Traditional MSA compliant pluggable modules and AOCs on card edge
- Considerable SI issues (electrical connectors, long traces on host PCBA) require re-timers.
- Front panel interconnect density limited by module size (physical implementation + module power dissipation)

- Embedded optical transceivers located closer around ASIC
- Shorter traces on PCB alleviate SI issues
- Optical fibers bring IOs to optical connectors on front panel
- Front panel interconnect density limited by size optical connectors
- Very high reliability/quality required

- Optical transceivers co-packaged w/ ASIC
- Minimized electrical interconnect eliminates SI issues
- Optical fibers bring IOs to optical connectors on front panel
- Lowest system power dissipation
- Highest front panel density and smallest potential system form factor
- Very high reliability required

Slide courtesy of P. De Dobbelaeere, Luxtera
Highly-Integrated Photonic I/O to Improve Power Efficiency (Luxtera)

Reducing system power dissipation by integration

- Faceplate
- Embedded
- MCM
- CMOS
- CMOS - Future

Power Consumption (μJ/bit)

Optical I/O
Module I/O
Host I/O
Serializer

Slide courtesy of P. De Dobbelaeere, Luxtera
A macrochip with WDM links

- Silicon lattice engineered with waveguides carries CPUs & DRAMs
- Bridges convert electrons to photons & couple to waveguides in lattice
- A high-bandwidth fully connected point-to-point optical network connects the sites
Hybrid Integration Platform for Large-Scale Photonic I/O

Electronic and Photonic Co-design

Integration

Packaging and System Impact

Courtesy of A. Krishnamoorthy, Oracle
32 nm SOI CMOS-Driven Link (Aurrion/IBM)

Current practice: Low integration level, inefficient 50 Ω interfaces

Maximum efficiency: directly drive the EAM

- 3 pJ/bit at 30 Gb/s (not including laser power)
- No measured penalty for 10km transmission at 25 Gb/s


October 22, 2015
C. Schow, UCSB
Electronic and photonic chip integration

CMOS IC (IBM) → EAM bias decoupling capacitors → Photonic IC (Aurrion) → Transmitter output

Demonstrated in hardware

CMOS IC (IBM) → Photonic IC (Aurrion) →

Wall-Plug WDM links (Aurrion/IBM)

Low modulation power $\rightarrow$ directly driving EAM, no 50$\Omega$ interfaces


October 22, 2015

C. Schow, UCSB
Photonic Switching
What do we mean by “fast” optical switching?

**ms-scale**

Mice flows over packet switch, elephant flows over OCS

**Promise:**
- Lower cost/power, fewer cables, software control (SDN)

**Challenges:**
- Scalability of software scheduler
- Slow reconfiguration time may limit applications

**µs-scale**

OCS at first switch level, hybrid but much faster then 3D-MEMS

**Promise:**
- Reconfigure network at flow-level, based on workloads
- Hardware control (FPGA).

**Challenges:**
- Custom NICs
- Scalability: switch radix

**ns-scale**

All-optical switching, electronic buffering at end points

**Promise:**
- Switching times ~ packet durations
- More power-efficient than electrical networks of equal BW

**Challenges:**
- Switch hardware and fast synchronizing links
- Scalability: switch radix: losses, fast control plane, flow control

---

Examples: Mordia/REACTOR (UCSD)
Hybrid Electrical/Optical Networks

- **Circuit switching**
  - Decouple line rate from speed of control plane
  - Used for persistent high-data rate traffic – must be scheduled

- **Packet switching**
  - Handles ‘tail’ of traffic demand
  - Can correct for errors in the circuit schedule

“High port count (320), low insertion loss, low crosstalk, <50ms reconfiguration”

Photonic Switches in Data Centers

ToR photonic switches:
- Fast reconfiguration (dynamic traffic)
- High-radix
- Low cost


Courtesy of Prof. G. Papen and G. Porter, UCSD
Switch/Driver Integration (IBM)

Monolithically integrated switch +driver chip (IBM 90nm photonics-enabled CMOS)

Fast reconfiguration: 4 ns

Transmit response of 2x2 MZ switch

Broad spectral bandwidth:
Routing many wavelength channels
- <-20dB crosstalk over 60 nm BW
- 32 channels at 200GHz spacing

Spectral response of 2x2 MZ switch

Losses too high, need significant feedback and control to manage crosstalk
- High level of electronic/photonic integration demanded

- B. G. Lee et al., "Monolithic Silicon Integration of Scaled Photonic Switch Fabrics, CMOS Logic, and Device Driver Circuits," JLT 2014.

October 22, 2015
C. Schow, UCSB
High Port Count MEMS Switches (UC Berkeley)

Detailed Schematic of Vertical Adiabatic Coupler Switch

On-Chip Losses of NxN Si Photonic Switches (N ≥ 8)

Courtesy Professor M. Wu, UC Berkeley
Wavelength Routing: AWGR

From Fat Tree to All-to-All (N^2) Connectivity with AWGR

Fat-Tree Network in Typical Data Centers

Fully Connected (All-to-All) Network

One-to-one Correspondence between Input and Output Waveguide

Single AWGR Scalability to 1024 x 1024

200GHz 8x8 Si-LIONS Chip

Individual components on Si-LIONS

Passband of the 8x8 200GHz AWGR

Modulator Spectrum

Courtesy of Professor S. J. Ben Yoo, UC Davis
Integrating Optical Gain for Scalability

Hybrid Approach (IBM):


Solder attachment of SOA arrays to photonic carriers
Challenging Packaging

• Challenging fabrication and assembly
• Precise alignment required for each SOA chip

New Electronic Capabilities Needed for Photonic Switching

- Optical circuits configured through a switch or fabric to connect $TX_i$ with $RX_j$
- Switching time defines context of usage
- But, switching time is not all hardware reconfiguration

---

**Millisecond-scale**

- Hybrid (circuit + packet) networks
- Coarse reconfiguration
- Software control (SDN)
- Highly scalable (> 1000 ports)

*e.g. Calient: 3D-MEMS*  
[J. Opt. Netw. 6 (1) 19]

**Microsecond-scale**

- Hybrid networks
- Reconfigure at flow-level
- Hardware control (FPGA)
- Scalable (10’s to 100 ports)

*e.g. Finisar: LCOS*  
[OFC 2006, OTuF2]

**Nanosecond-scale**

- Reconfigure at packet granularity
- Quasi-packet switching with buffering at end points
- Limited scalability (10’s of ports)

*e.g. IBM: Photonics*  
[OFC 2013, PDP5C.3]

Courtesy of B. Lee, IBM
Switching Time = Switch Reconfiguration Time + Link Synchronization Time

\[
\text{Switching Time} = \text{Switch Reconfiguration Time} + \text{Link Synchronization Time}
\]

Receivers must adapt to abrupt changes in incoming data amplitude and phase

**Millisecond-scale**
- Hybrid (circuit + packet) networks
- Coarse reconfiguration
- Software control (SDN)
- Highly scalable (> 1000 ports)

**Microsecond-scale**
- Hybrid networks
- Reconfigure at flow-level
- Hardware control (FPGA)
- Scalable (10’s to 100 ports)

**Nanosecond-scale**
- Reconfigure at packet granularity
- Quasi-packet switching with buffering at end points
- Limited scalability (10’s of ports)

*Courtesy of B. Lee, IBM*

---

e.g. Calient: 3D-MEMS
[J. Opt. Netw. 6 (1) 19]
e.g. Finisar: LCOS
[OFC 2006, OTuF2]
e.g. IBM: Photonics
[OFC 2013, PDP5C.3]
Complete Burst-Mode Receiver (31ns lock, 4 pJ/bit)

Fast photonic routing and switching fabrics must have companion electronics
• Needs more research in the community
Enables fine-grained power management

Path Forward: Large-Scale Electronic/Photonic Integration
**Si Photonics:** integrating photonic devices into Si platforms to leverage the huge Si electronics manufacturing infrastructure

- Very large scale integration
- Tight process control
- Wafer-level testing
- High-yield and low cost

**Advantages for the data center:**
- Single-mode $\rightarrow$ multi-kilometer links
- Wavelength-division multiplexing (WDM) $\rightarrow$ high BW/fiber
- High integration level $\rightarrow$ many devices needed for switching and high-speed interconnects
Step 1: Bond InP-dies on SOI waveguide

Step 2: Remove substrate

Step 3: Process lasers at wafer scale

Figures: IMEC/Ghent; Slide courtesy of Prof. J. Bowers, UCSB
Need for Many Functions: Heterogeneous Integration of 6 Photonic Platforms

GaAs

Silicon

LiNbO$_3$

InP

SiN/SiON/SiO$_2$

Ce:YIG

Isolator

Heck et al. JSTQE 2013

Top view

Input

B

Ce:YIG

October 22, 2015

C. Schow, UCSB
Large-scale photonic integration
• Switching
• Amplification
• Control

Hybrid III-V on Si SOAs Demonstrated in Multiple Platforms

UCSB\textsuperscript{1}, 2006

Ghent University-IMEC\textsuperscript{2}, 2012

III-V Lab\textsuperscript{3}, 2014

Large Scale Integration for Photonic I/O

Photonics must deliver system advantages
- More I/O bandwidth at less power for processors
- Larger port switches to enable flatter networks
- Higher Integration levels: processor, memory, network

Hundreds of photonic interfaces: all off-module high-speed I/O
- High BW/interface \(\rightarrow\) WDM
- Low cost \(\rightarrow\) compatibility with high-volume electronics manufacturing
- Low loss \(\rightarrow\) maximize link power efficiency by minimizing required laser power

Thousands of electrical interfaces: connecting electronics to photonics
- High density, high-speed, low-power
- New functions: rapid synchronization for low latency switching, power management
- Specialized chip I/O co-designed with and only for photonics, no general purpose cores

Reliability: components either don’t fail or can be spared (also for yield)

- Large-scale integration is fundamentally required
- Holistic design of photonics, electronics, packaging and assembly
Closing Thoughts

Potential: More highly connected systems

Photonic Switching
- Routing 10’s of Tb/s at sub pJ/bit power

Photonic I/O
- Higher radix electrical switches, flatter networks
- More processor/memory bandwidth
- Multiple photonic technologies for multiple purposes
  - VCSEL, PSM, WDM point-to-point, WDM switched

Challenge: Integrating large-scale electronics with large-scale photonics

Re-thinking manufacturability and supply chain
- Moving from electronic to photonic-centric packaging
- Electronics/Photonics/Package co-design required
- What are the highest-value components and what assembly flow makes sense
- Who does what?
Thank You!

schow@ece.ucsb.edu

Acknowledgements
IBM Research Colleagues:
C. Baks, R. Budd, F. Doany, N. Dupuis, D. Kuchta, B. Lee, J. Proesel,
P. Pepeljugoski, A. Rylyakov, L. Scharles, M. Taubenblatt, and many others...

Slides:
M. Tan (HP), P. De Dobbelare (Luxtera), D. Maltz (Microsoft),
M. Laor (Compass Networks), A. Krishnamoorthy (Oracle), G. Papen (UCSD),
M. Wu (UC Berkeley), S. J. B. Yoo, (UC Davis), J. Bowers (UCSB)