UCSB

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

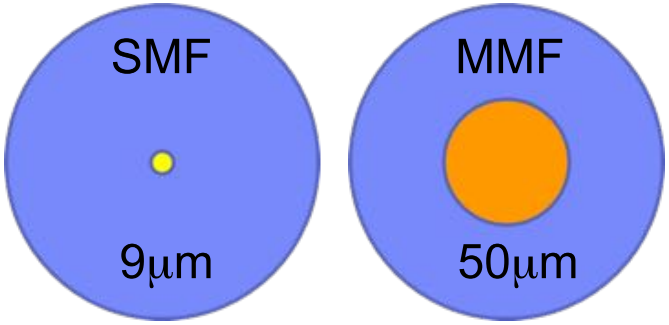
Data Centers

Datacom (100’s of meters)

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

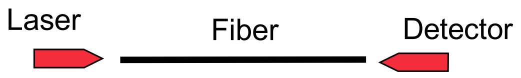
Relative core size:

MMF >30X SMF



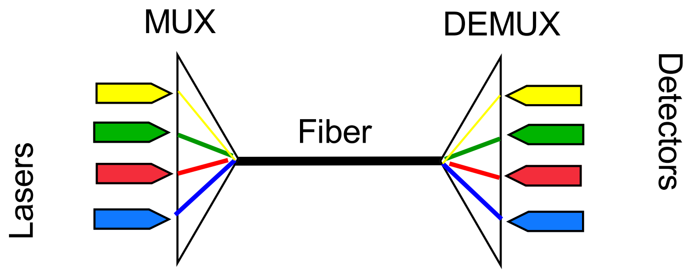
TDM = Time Division Multiplexing

Single optical channel,
Electronic Mux/Demuxing



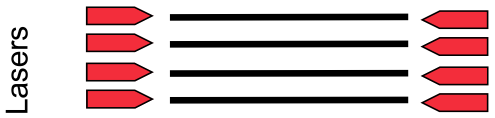
WDM = Wavelength Division Multiplexing

Single optical channel
data carried on separate λ’s



SDM = Space Division Multiplexing

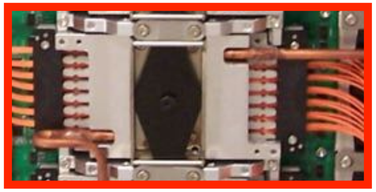
Parallel fiber channels,
No Mux/Demuxing

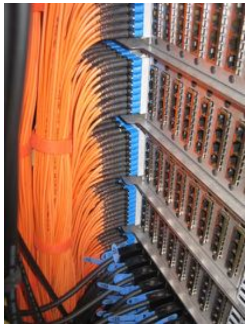


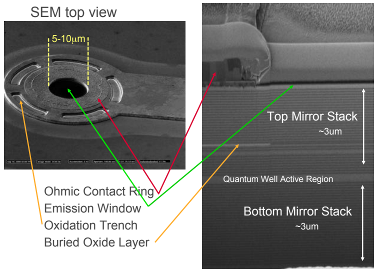
Technologies for Data Center Interconnects

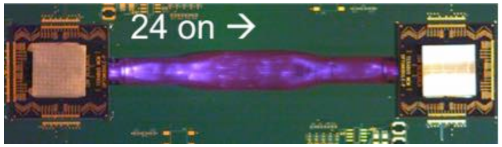
VCSELs (Vertical Cavity Surface Emitting Lasers)

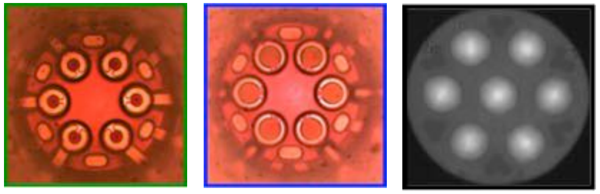
Multi-mode fiber (MMF), Polymer Waveguides, Multicore fiber





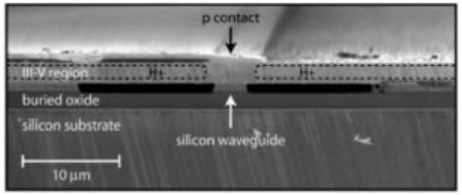




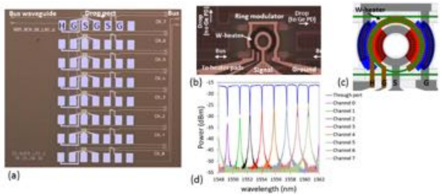


Si Photonics

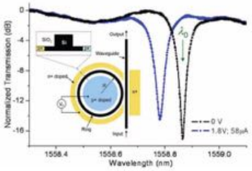
Single-mode fiber (SMF), Wavelength Division Multiplexing (WDM)



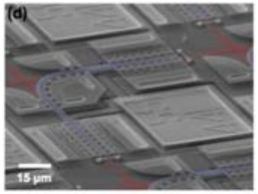
UCSB



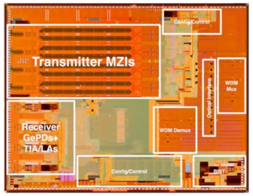
IMEC/Ghent



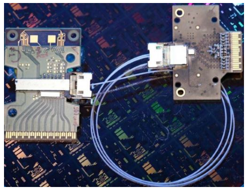
Cornell/Columbia



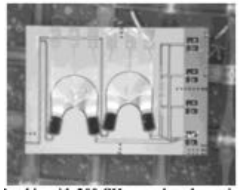
UC Berkeley



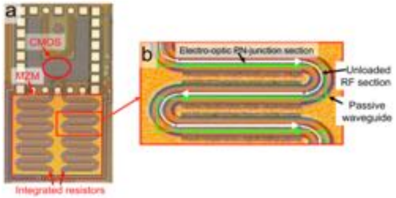
Luxtera



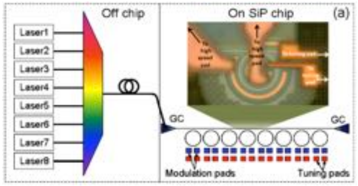
Intel



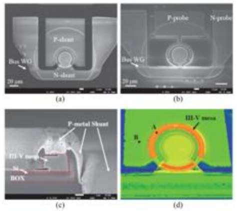
Aurrion



IBM

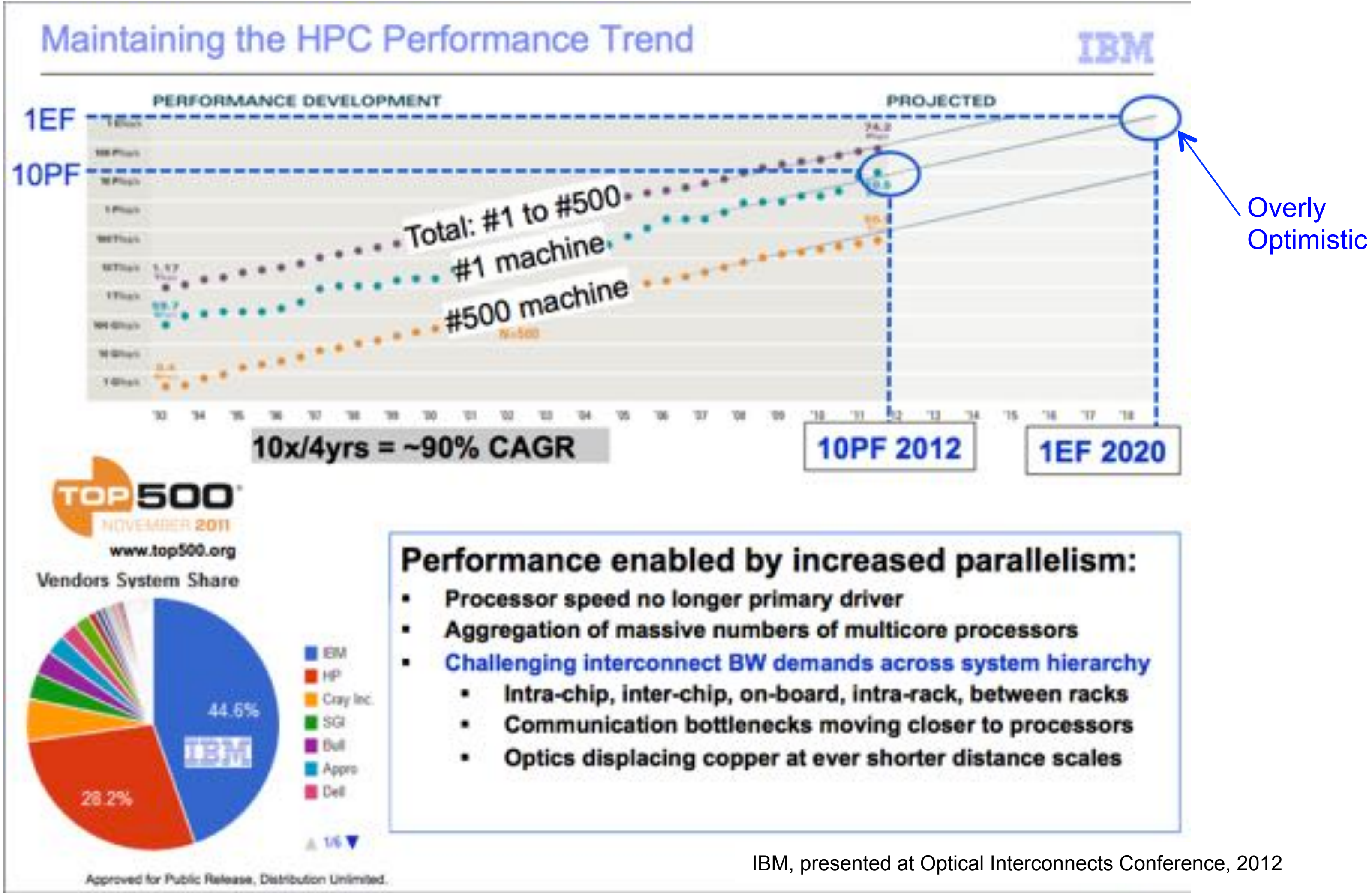


Oracle



HP

The Old Days: 2012



Maintaining the HPC Performance Trend

Performance enabled by increased parallelism:

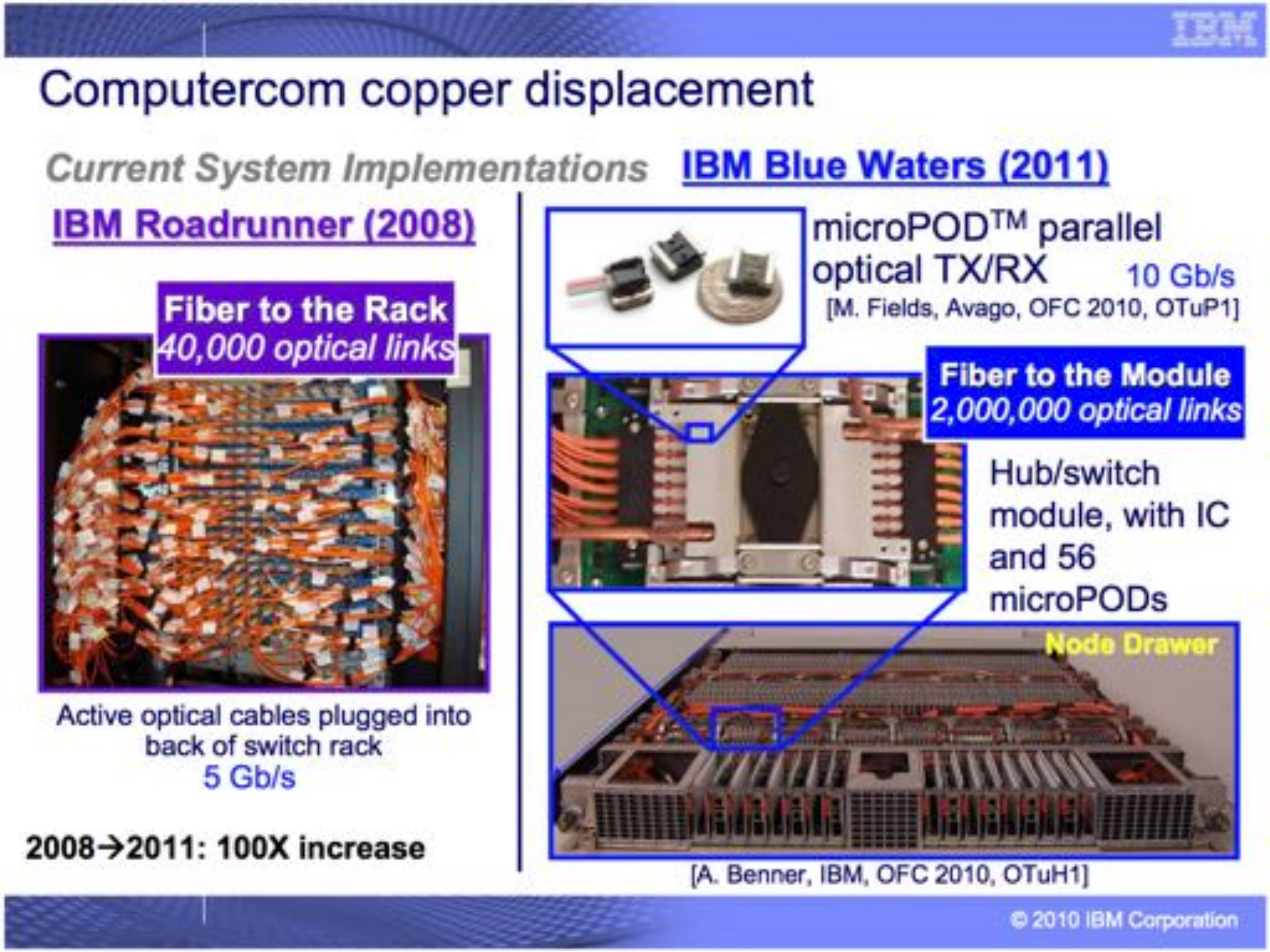
- Processor speed no longer primary driver
- Aggregation of massive numbers of multicore processors
- Challenging interconnect BW demands across system hierarchy
  - Intra-chip, inter-chip, on-board, intra-rack, between racks
  - Communication bottlenecks moving closer to processors
  - Optics displacing copper at ever shorter distance scales

Approved for Public Release, Distribution Unlimited.

IBM, presented at Optical Interconnects Conference, 2012

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

The Promise of A Machine with 2,000,000 VCSELs



IBM

Computercom copper displacement

Current System Implementations **IBM Blue Waters (2011)**

IBM Roadrunner (2008)

Fiber to the Rack
40,000 optical links

Active optical cables plugged into back of switch rack
5 Gb/s

2008→2011: 100X increase

microPOD™ parallel optical TX/RX 10 Gb/s
[M. Fields, Avago, OFC 2010, OTuP1]

Fiber to the Module
2,000,000 optical links

Hub/switch module, with IC and 56 microPODs

[A. Benner, IBM, OFC 2010, OTuH1]

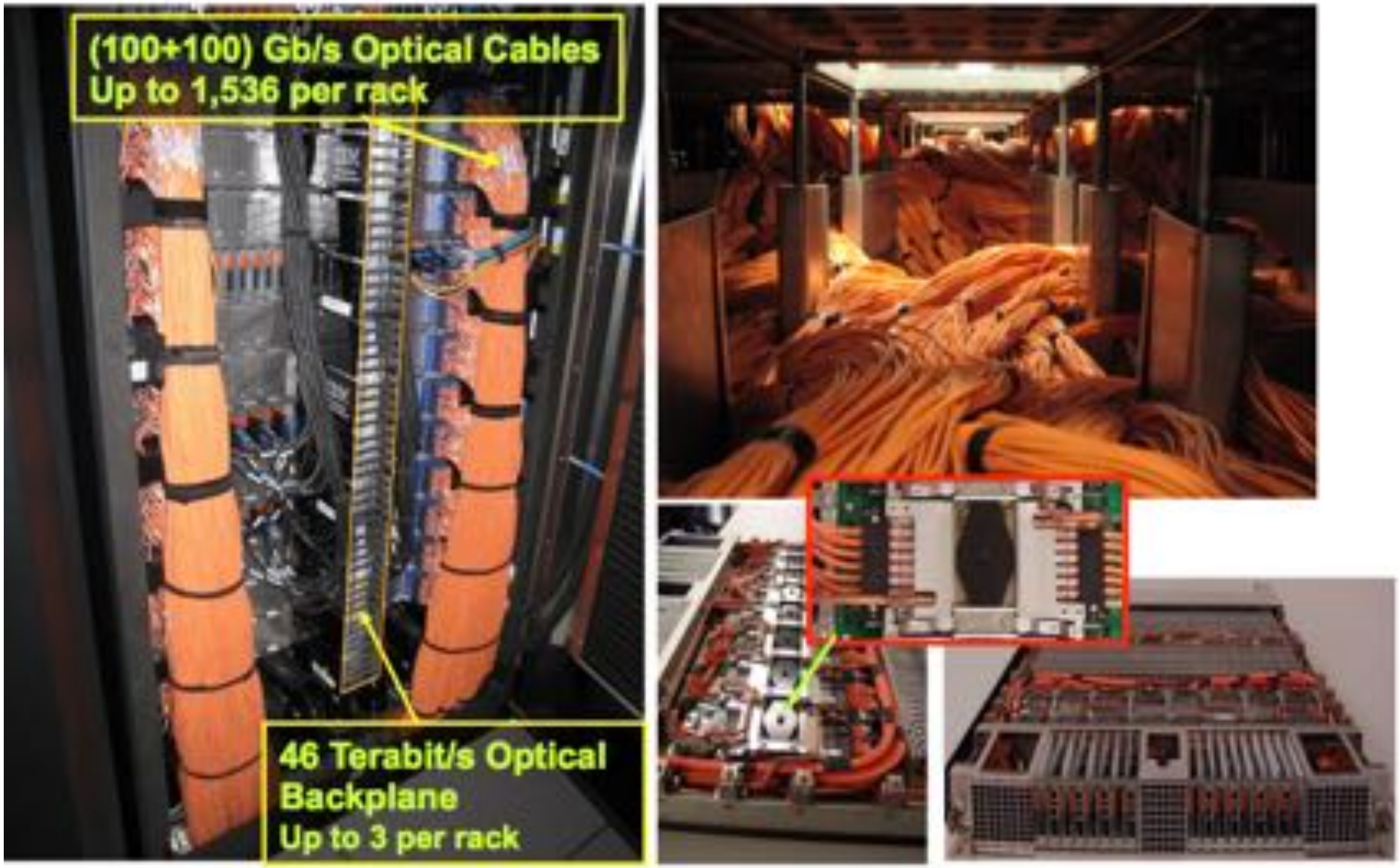
IBM, public presentation 2010

- A. Benner, "Optical Interconnect Opportunities in Supercomputers and High End Computing," OFC Tutorial 2012.

Fiber Backplane Density Pushed to the Limit

IBM Power 775: Pushing the Limits

IBM



IBM, public presentation 2010

Need more BW/fiber: WDM, multicore fiber

- A. Benner, “Optical Interconnect Opportunities in Supercomputers and High End Computing,” OFC Tutorial 2012.

Optics in HPC: IBM Sequoia

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

SEQUOIA
LAWRENCE LIVERMORE NATIONAL LABORATORY
Science in the National Interest

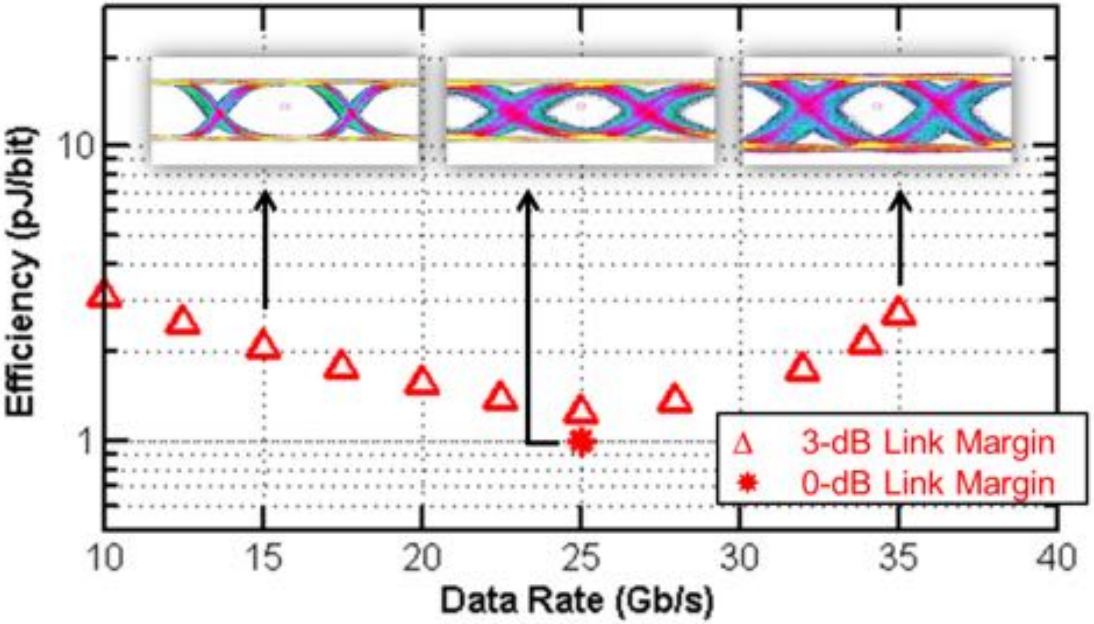
620,000 VCSEL links

IBM, presented at IPC, 2013

- HPC requires technologies optimized for short reach ~50m

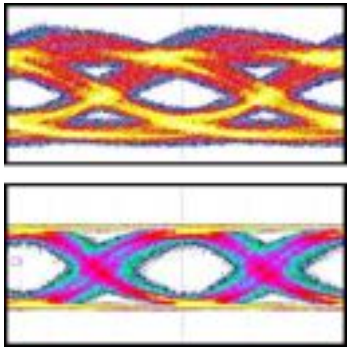
VCSELs: Efficient and Fast

32-nm CMOS-Driven Link



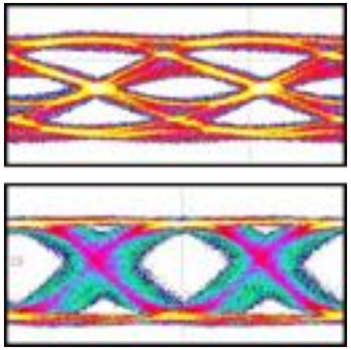
25 Gb/s, 24mW

35 Gb/s, 95mW



TX out

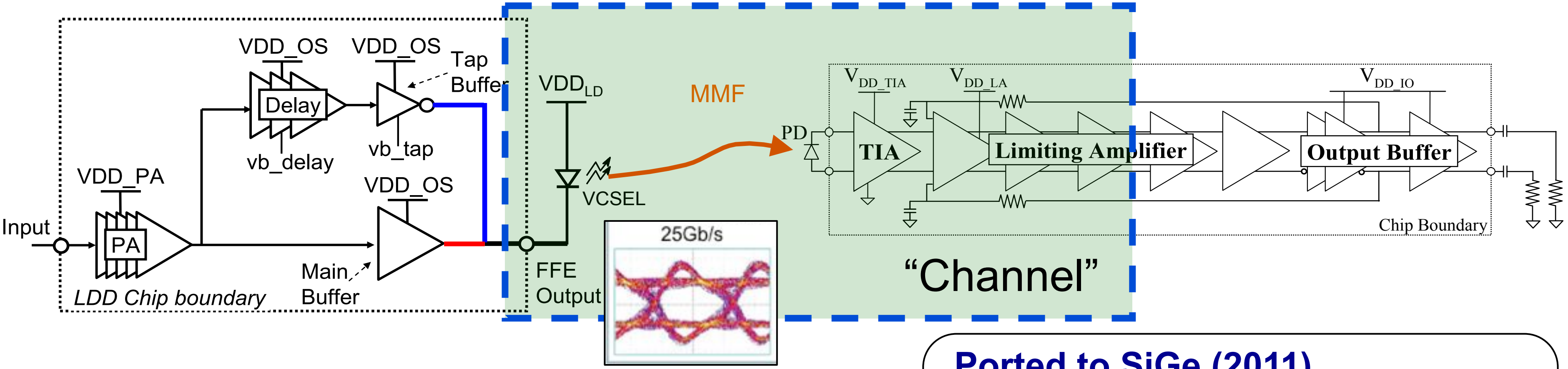
RX out



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

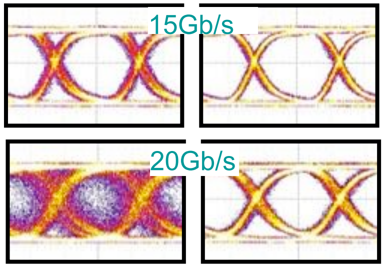
- J. E. Proesel *et al.*, "35-Gb/s VCSEL-based optical link using 32-nm SOI CMOS circuits," *OFC 2013*, Paper OM2H2, Mar. 2013.

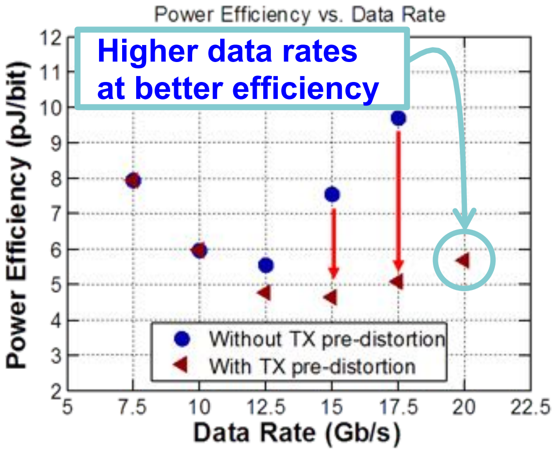
Rethinking Equalization: Optimizing the Performance of Complete Links



First Implemented in CMOS (2011)

No Equalization | With Equalization





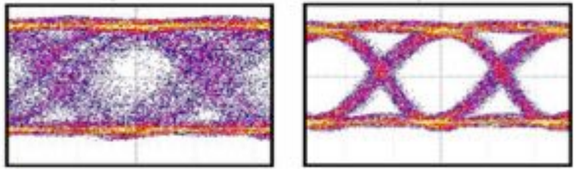
Higher data rates at better efficiency

Ported to SiGe (2011)

First 40 Gb/s VCSEL Links (2012)

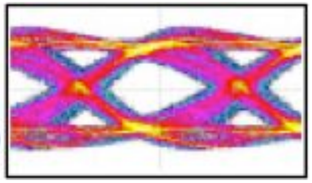
No Equalization | With Equalization

Link Output (40 Gb/s)



71 Gb/s (2015)

Link Output (71 Gb/s)



Next opportunity: Si photonic WDM links

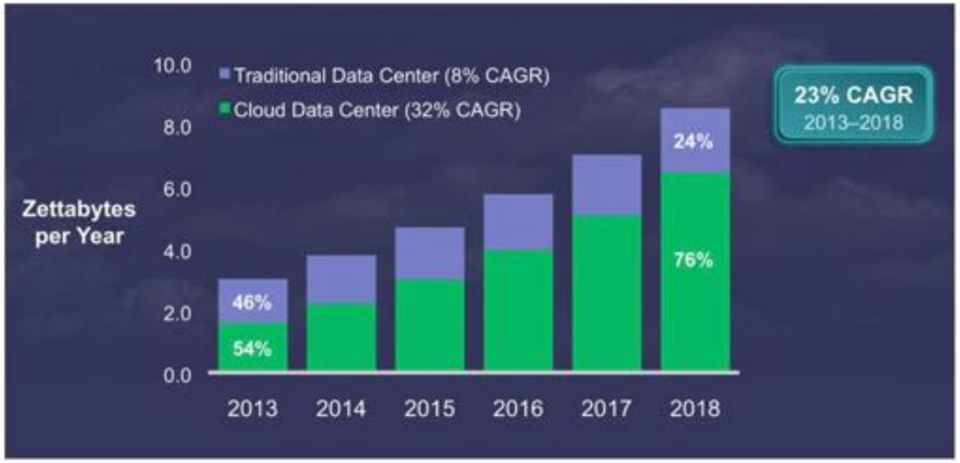
- A.V. Rylyakov *et al.*, "Transmitter Pre-Distortion for Simultaneous Improvements in Bit-Rate, Sensitivity, Jitter, and Power Efficiency in 20 Gb/s CMOS-driven VCSEL Links, *JLT* 2012.
- A. V. Rylyakov *et al.*, "A 40-Gb/s, 850-nm VCSEL-based full optical link," *OFC* 2012.
- D. Kuchta *et al.*, "A 71-Gb/s NRZ Modulated 850-nm VCSEL-Based Optical Link," *PTL,* March 2015.

And Then The Cloud Rolled In

View of Goleta Beach from UCSB, source www.theinertia.com

Growth in Cloud Data Centers

Figure 3. Total Data Center Traffic Growth

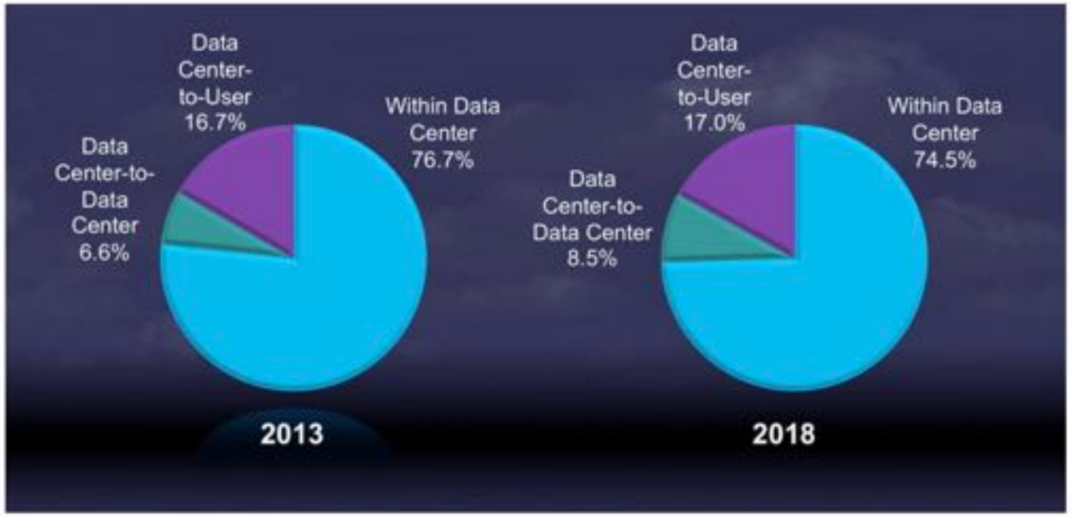


Source: Cisco Global Cloud Index, 2013–2018

Cloud is the growth market

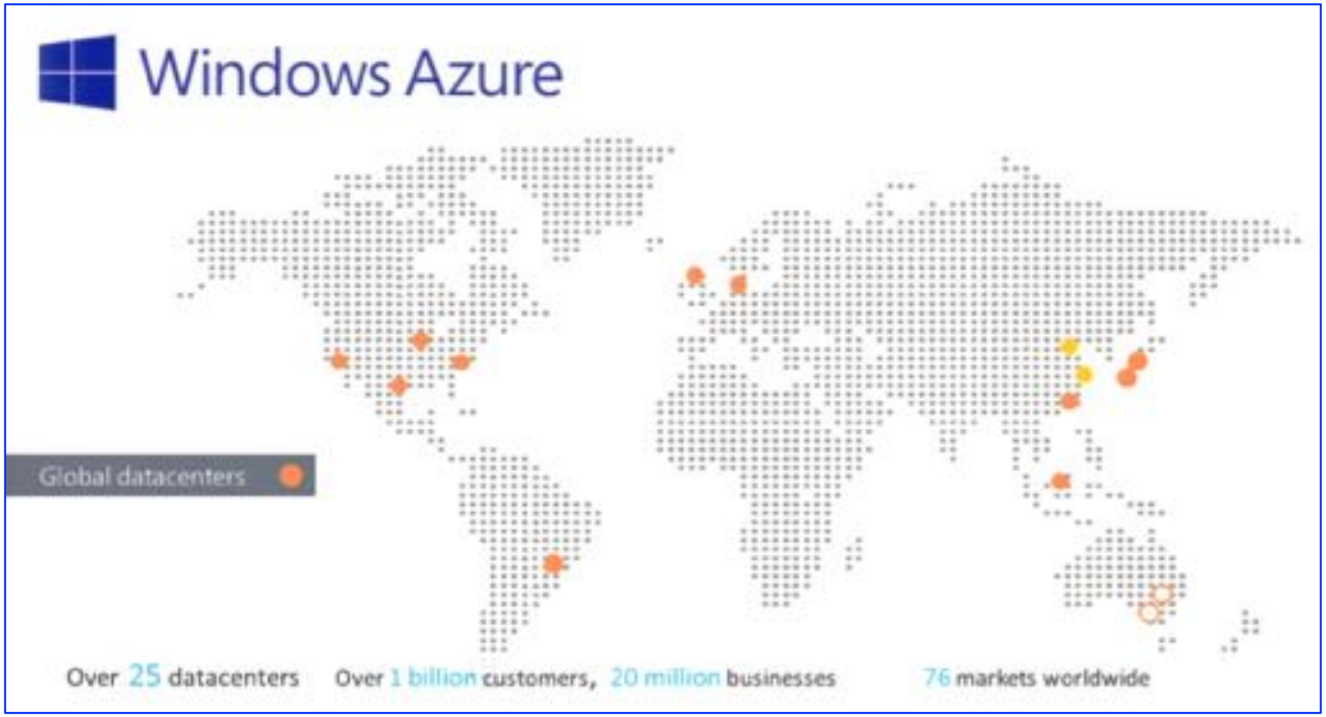
75% of traffic is within the data center

Figure 2. Global Data Center Traffic by Destination



Source: Cisco Global Cloud Index, 2013–2018

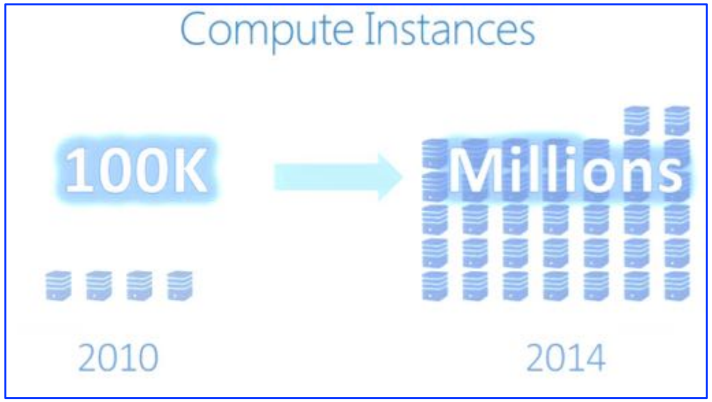
Huge Growth in the Cloud (Microsoft)



Windows Azure

Global datacenters ●

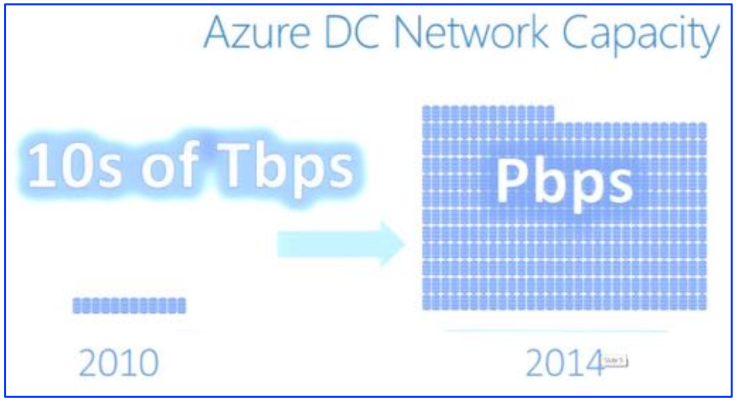
Over 25 datacenters Over 1 billion customers, 20 million businesses 76 markets worldwide



Compute Instances

100K → Millions

2010 2014



Azure DC Network Capacity

10s of Tbps → Pbps

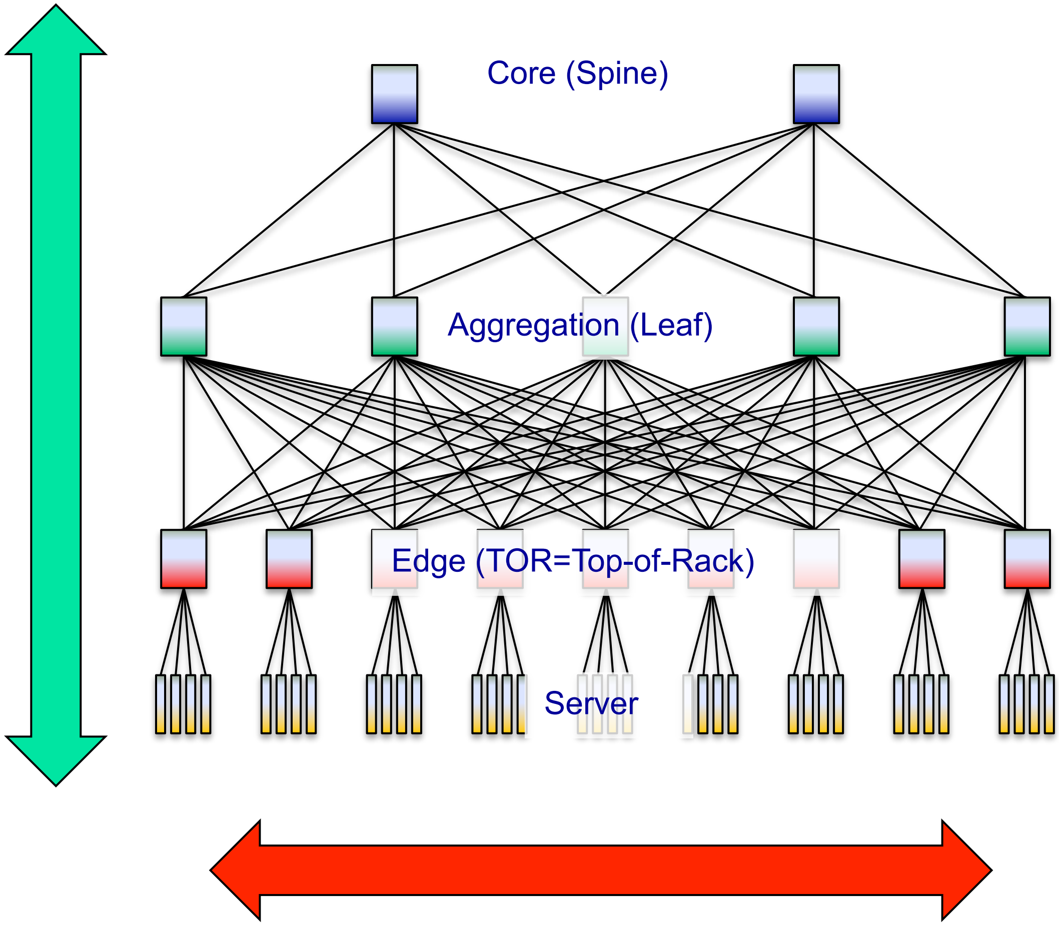
2010 2014

Source: D. Maltz, Microsoft, OFC 2014

Traditional Data Center Network Hierarchy

Traditional hierarchical networks grew out of campus/WAN installations

Good for North-South traffic

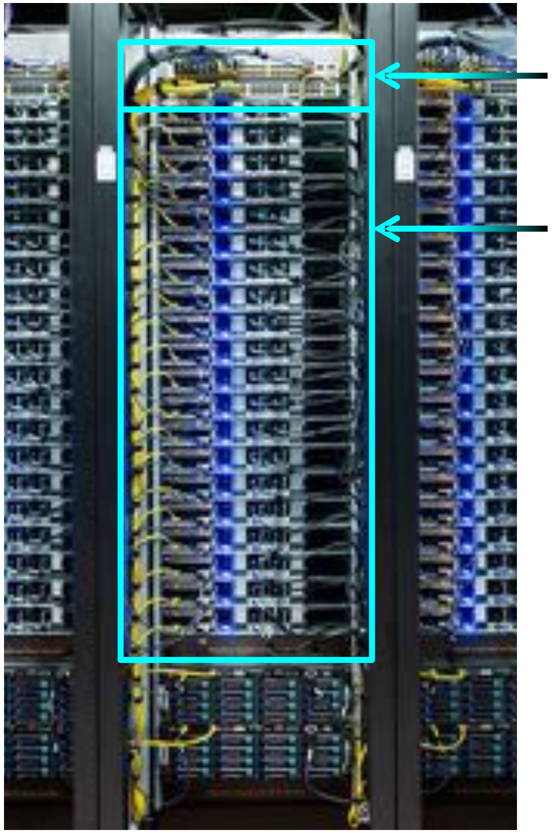


Bad for East-West traffic (75% of data center traffic)

See L. A. Barroso and U. Hölzle, "*The Datacenter as a Computer—An Introduction to the Design of Warehouse-Scale Machines,*"

Data Center Hardware

Racks



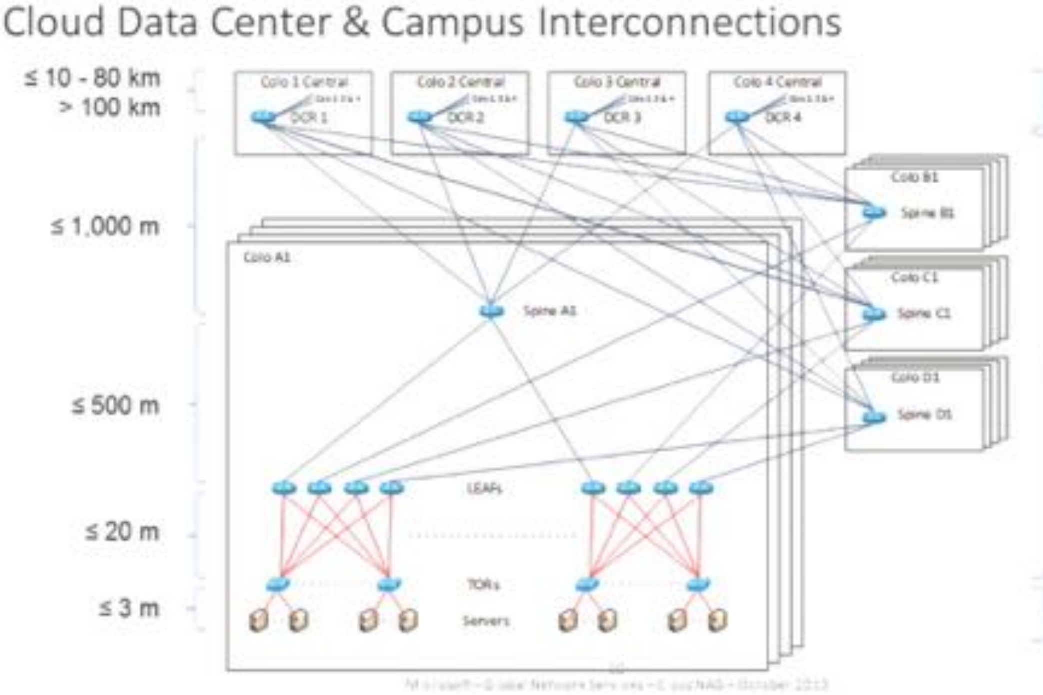
Top of Rack (ToR) switch

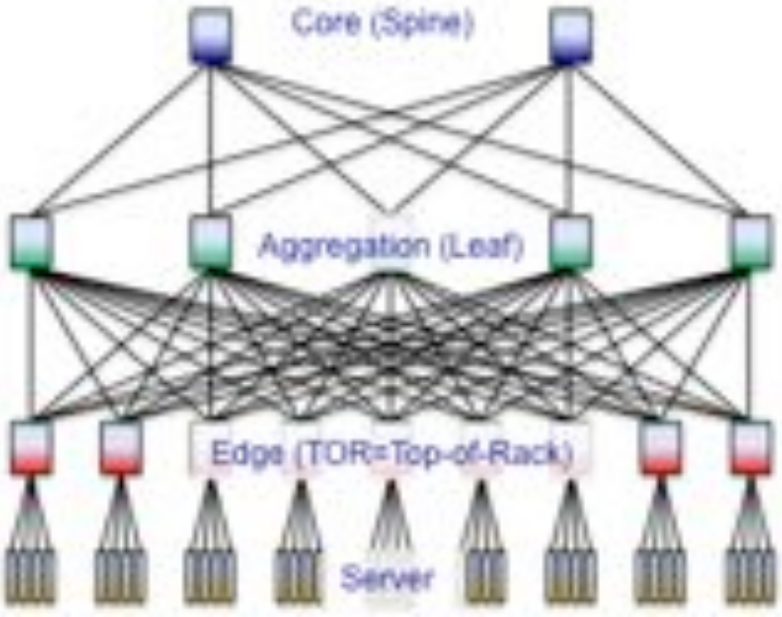
Servers

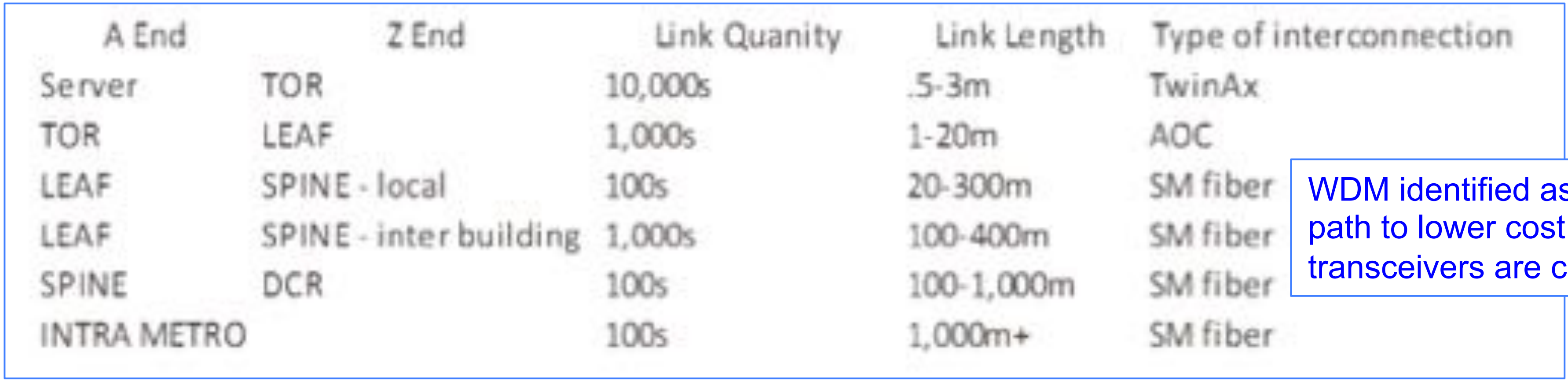
Lots of Racks

Photos of Facebook data centers found on Google images

Cloud Data Centers (Microsoft)







A End	Z End	Link Quanity	Link Length	Type of interconnection
Server	TOR	10,000s	.5-3m	TwinAx
TOR	LEAF	1,000s	1-20m	AOC
LEAF	SPINE - local	100s	20-300m	SM fiber
LEAF	SPINE - inter building	1,000s	100-400m	SM fiber
SPINE	DCR	100s	100-1,000m	SM fiber
INTRA METRO		100s	1,000m+	SM fiber

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.

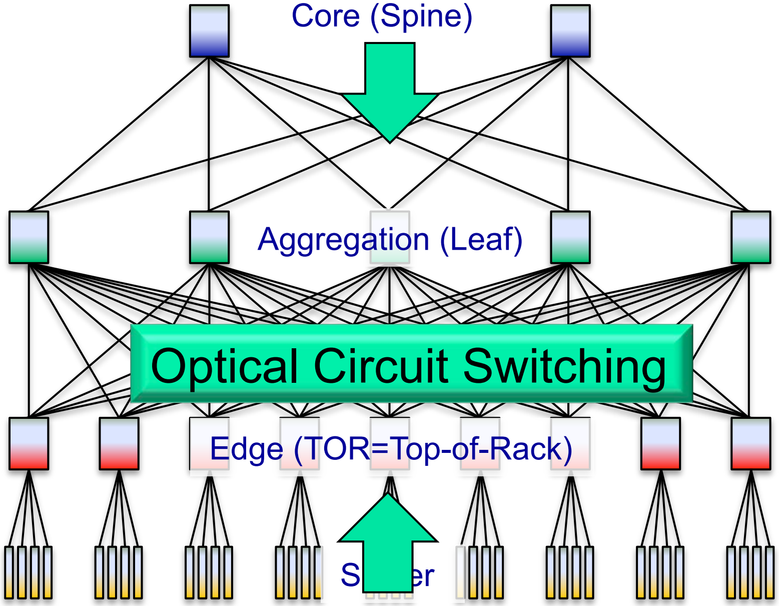
10. **NAP of the Americas** (Miami, USA)
Area: 750,000 Sq. Ft.

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

UCSB

Toward More Connected Data Centers

Number of ports per switch	#Nodes Connected Two Level	#Nodes Connected Three Level	#Nodes Connected Four Level	#Nodes Connected Five Level
8	32	128	512	2048
16	128	1024	8192	65536
32	512	8192	131072	2.10E+06
64	2048	65536	2.10E+06	6.71E+07
128	8192	524288	3.36E+07	2.15E+09
256	32768	4.19E+06	5.37E+08	6.87E+10
512	131072	3.36E+07	8.59E+09	2.20E+12
1024	524288	2.68E+08	1.37E+11	7.04E+13



1) Flatten the Network:

Higher radix (larger port count) switches

Electrical or Optical Cores

2) Change the Network:

Photonic switching

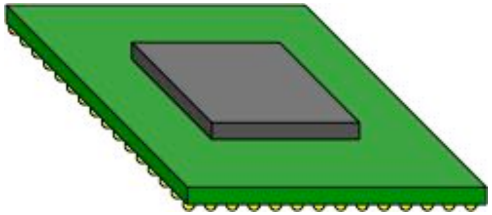
Photonic I/O:
Optics to the Chip

Shared Visions: Photonically Connected Chips

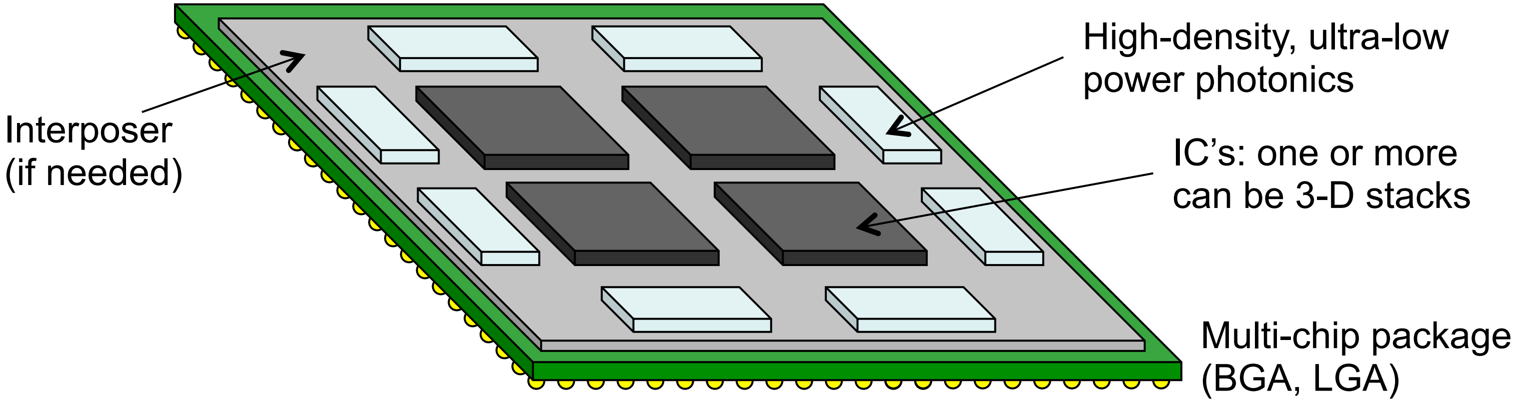
Today: Electrical Chip Packaging

IC packages with course BGA/LGA electrical connectors

- Poor scalability, signal integrity
- Reduced system performance
- Reduced system efficiency



Future: Photonic Packages



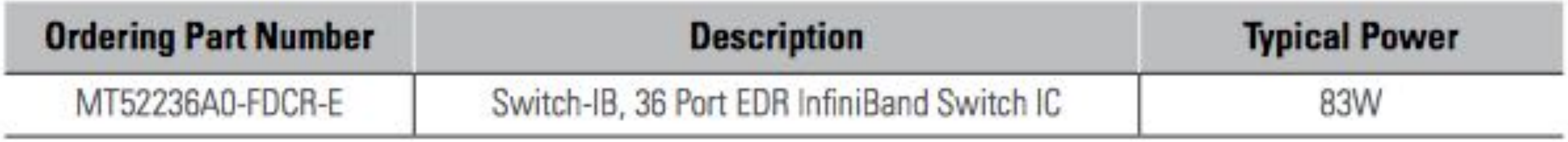
Photonic integration must provide more I/O bandwidth at better efficiency

Limitations of Electrical Switches

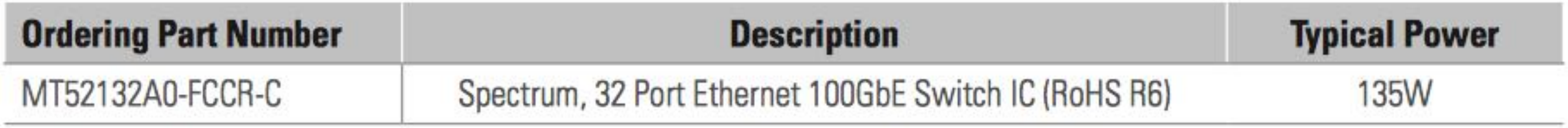
Mellanox

SwitchIB

Spectrum™



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



Ordering Part Number	Description	Typical Power
MT52132A0-FCCR-C	Spectrum, 32 Port Ethernet 100GbE Switch IC (RoHS R6)	135W

Source: mellanox.com

Broadcom

Source: broadcom.com

Tomahawk

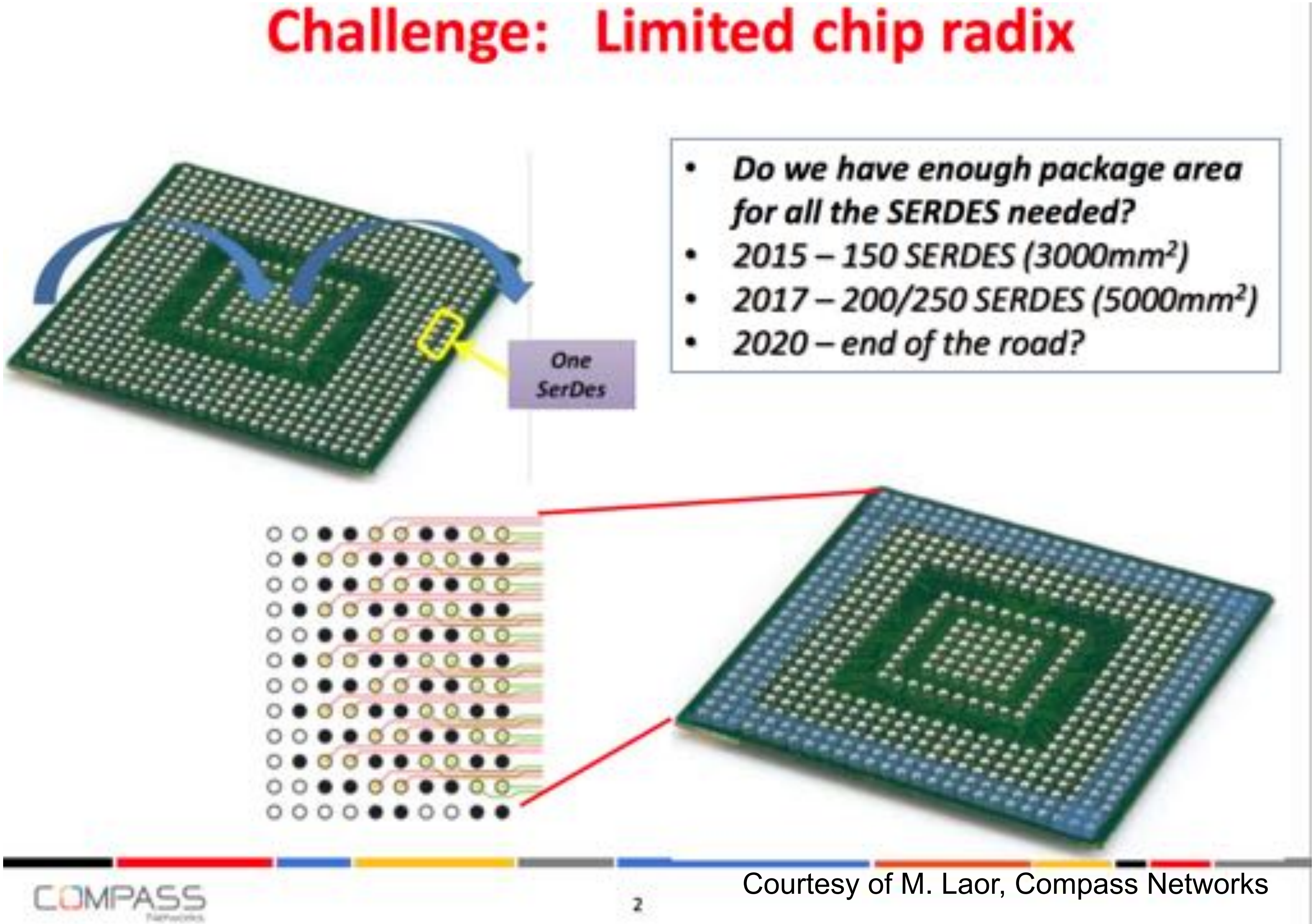
128 x 25Gb/s SerDes = 32 100GbE ports

7 Billion transistors

Two primary limitations

- **Power (mostly electrical I/O)**
- **Density**

Density Limits Imposed by Packages, Not ICs



Challenge: Limited chip radix

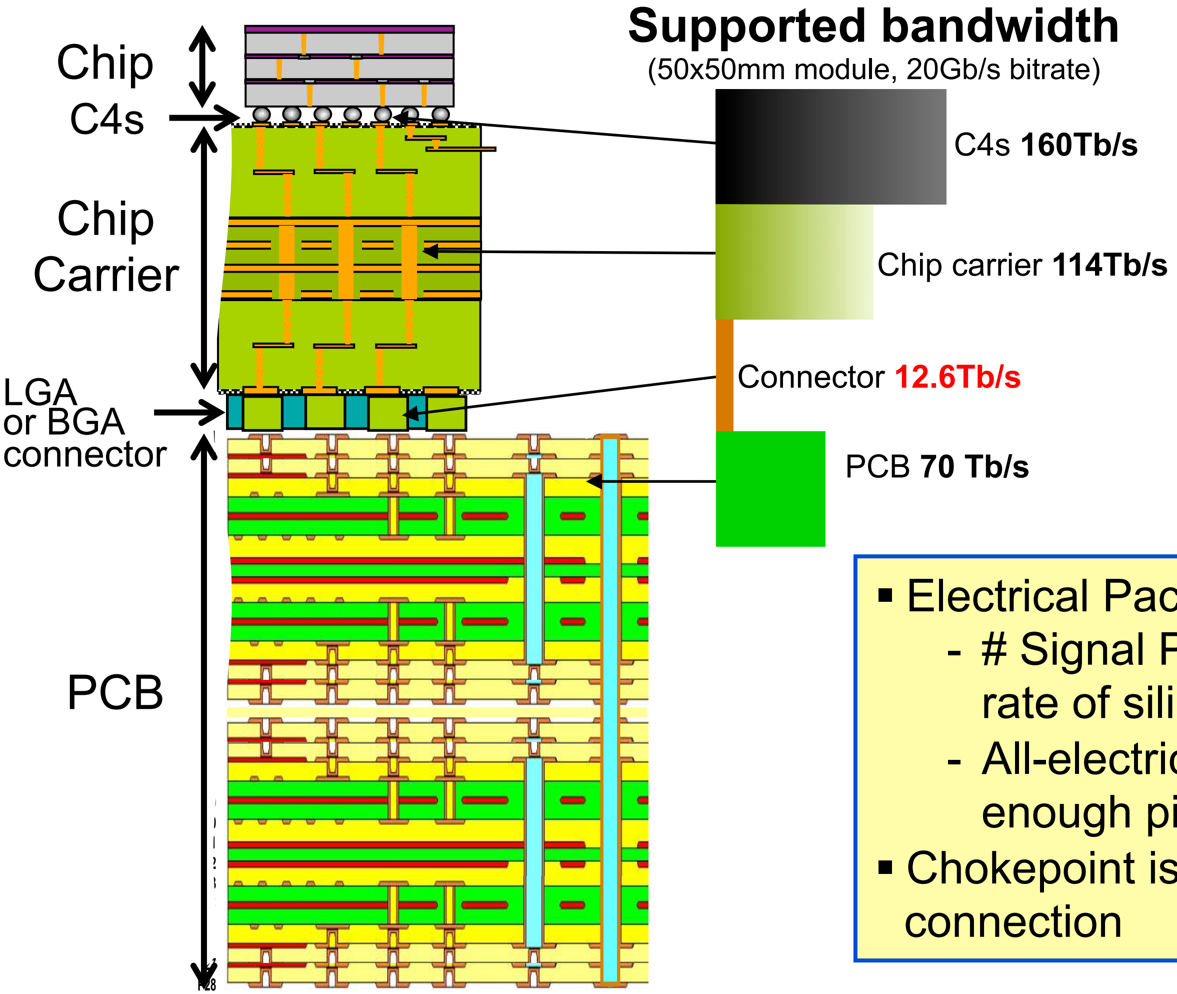
- *Do we have enough package area for all the SERDES needed?*
- *2015 – 150 SERDES (3000mm²)*
- *2017 – 200/250 SERDES (5000mm²)*
- *2020 – end of the road?*

One SerDes
COMPASS
Networks
2

Courtesy of M. Laor, Compass Networks

Higher BW Density: Demands Integration at (First-Level) Chip Package

Cross-sectional view of chip module on board

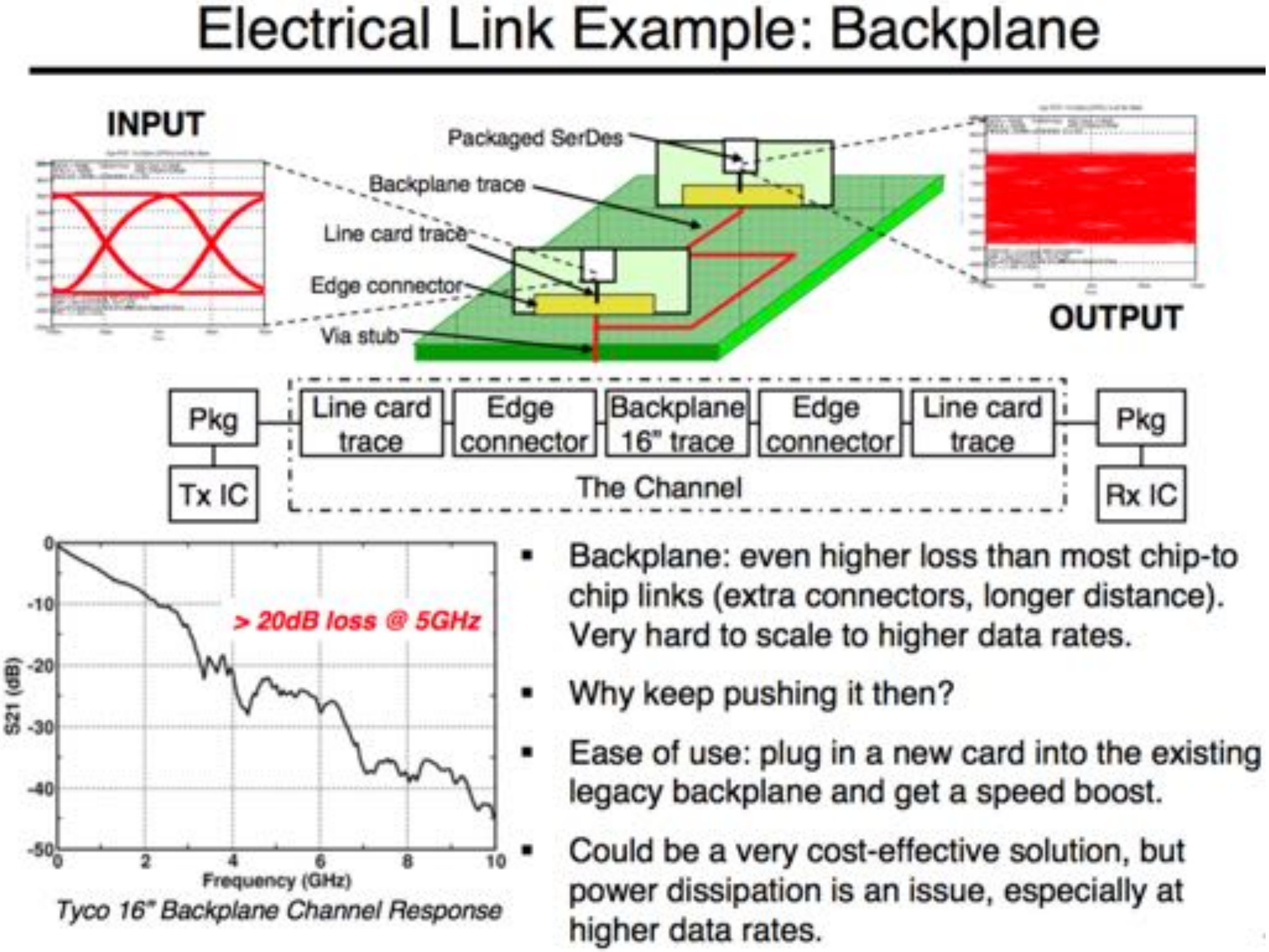


Courtesy of M. Ritter, IBM

- 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

- D. Kam et al., “Is 25 Gb/s On-Board Signaling Viable?, *IEEE Adv. Packag.*, 2009

Electrical I/O: Burning Power to Overcome Packaging Limitations



Electrical Link Example: Backplane

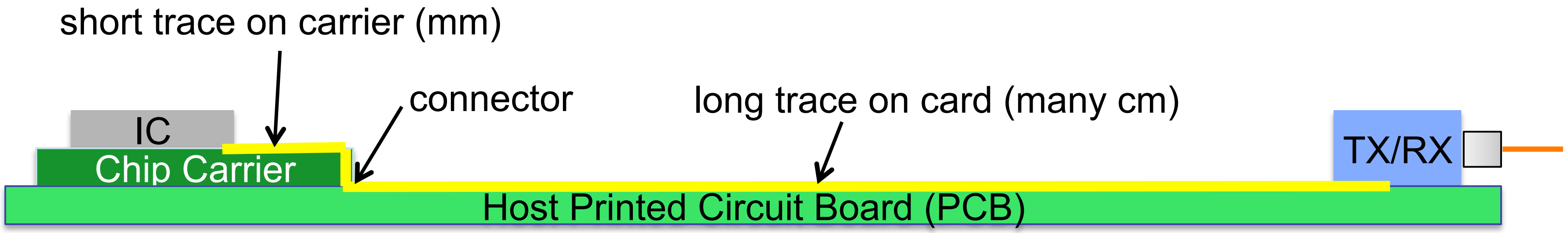
Tyco 16" Backplane Channel Response

- 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.

Courtesy A. Rylyakov, OFC Short Course #357

UCSB

Integration to Maximize BW and Efficiency

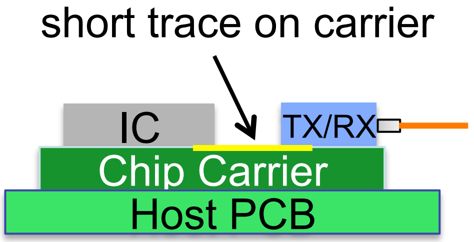

short trace on carrier (mm)
connector
long trace on card (many cm)
IC
Chip Carrier
TX/RX
Host Printed Circuit Board (PCB)

Move from bulky optics located far away

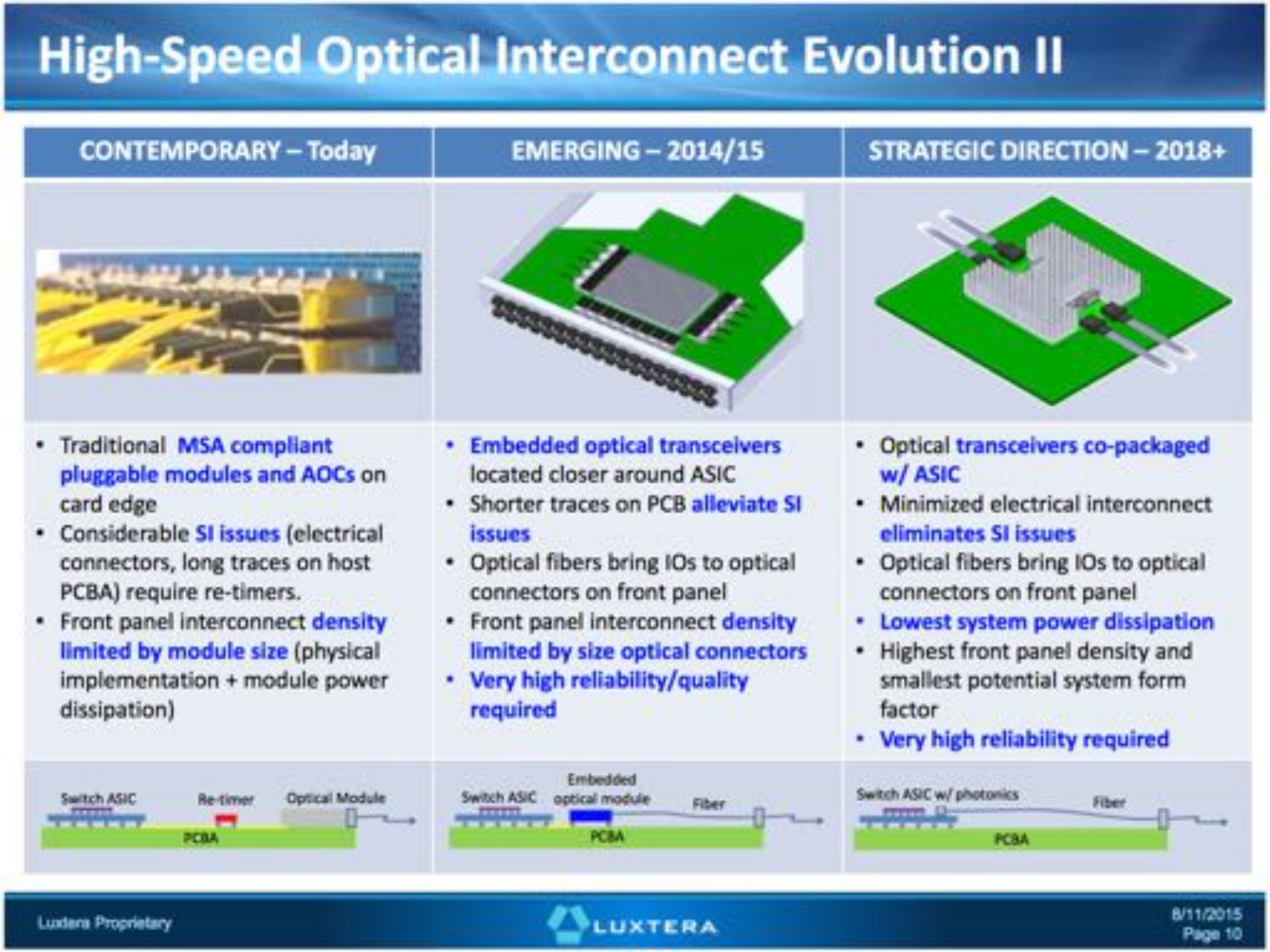
To highly-integrated optics close to the logic chips

Integrating photonics into the most expensive, constrained, and challenging environment in the system

- **Cost**
- **Reliability**
- **Thermal**
- **Power delivery**



From Discrete Modules to Integrated Chip I/O (Luxtera)



High-Speed Optical Interconnect Evolution II

CONTEMPORARY – Today

- 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)

EMERGING – 2014/15

- **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**

STRATEGIC DIRECTION – 2018+

- 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**

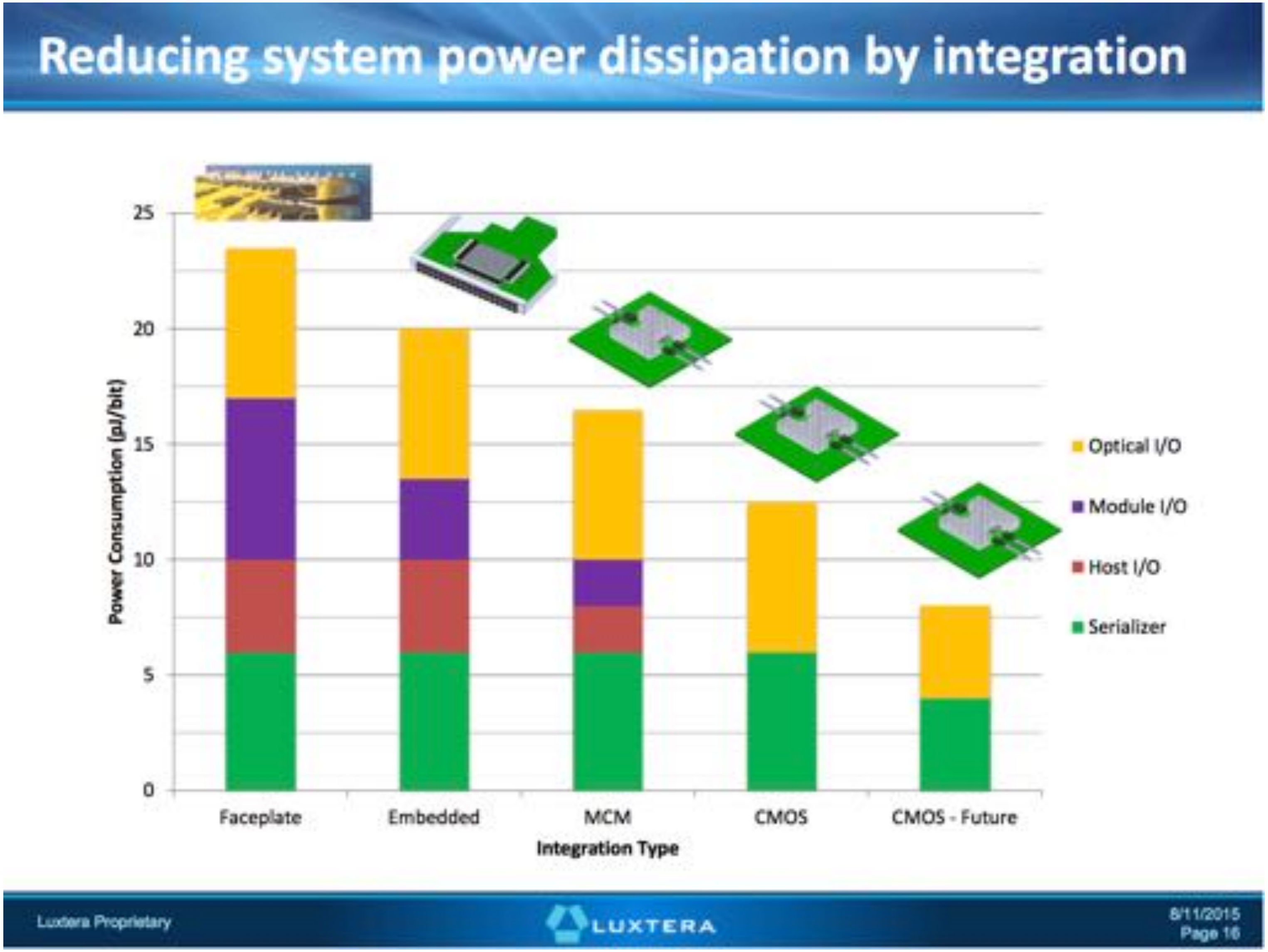
Luxtera Proprietary

LUXTERA

8/11/2015
Page 10

Slide courtesy of P. De Dobbelaere, Luxtera

Highly-Integrated Photonic I/O to Improve Power Efficiency (Luxtera)



Reducing system power dissipation by integration

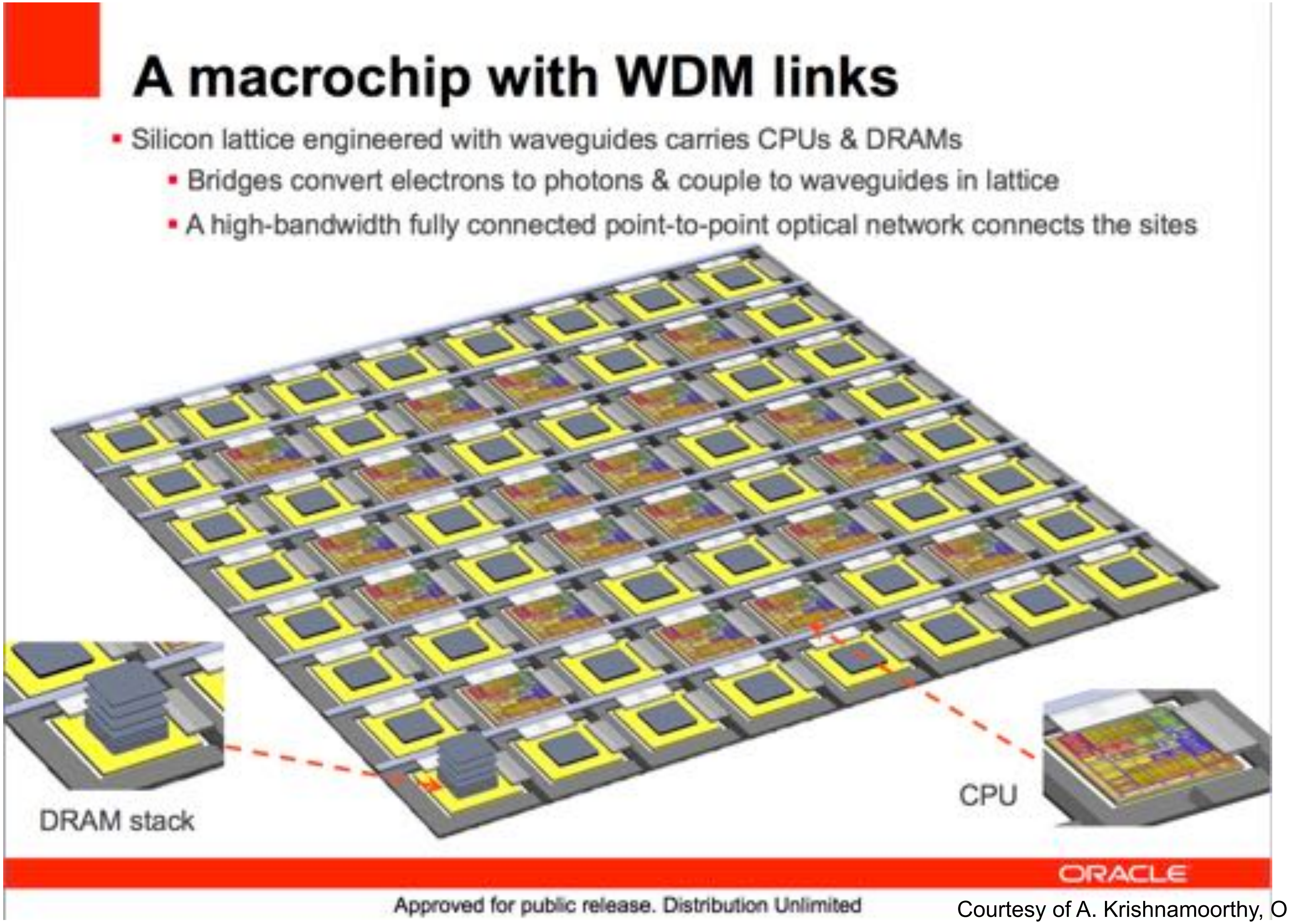
Luxtera Proprietary

LUXTERA

8/11/2015
Page 16

Slide courtesy of P. De Dobbelaere, Luxtera

Toward All-Photonic WDM I/O (Oracle)



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

DRAM stack
CPU

ORACLE

Approved for public release. Distribution Unlimited

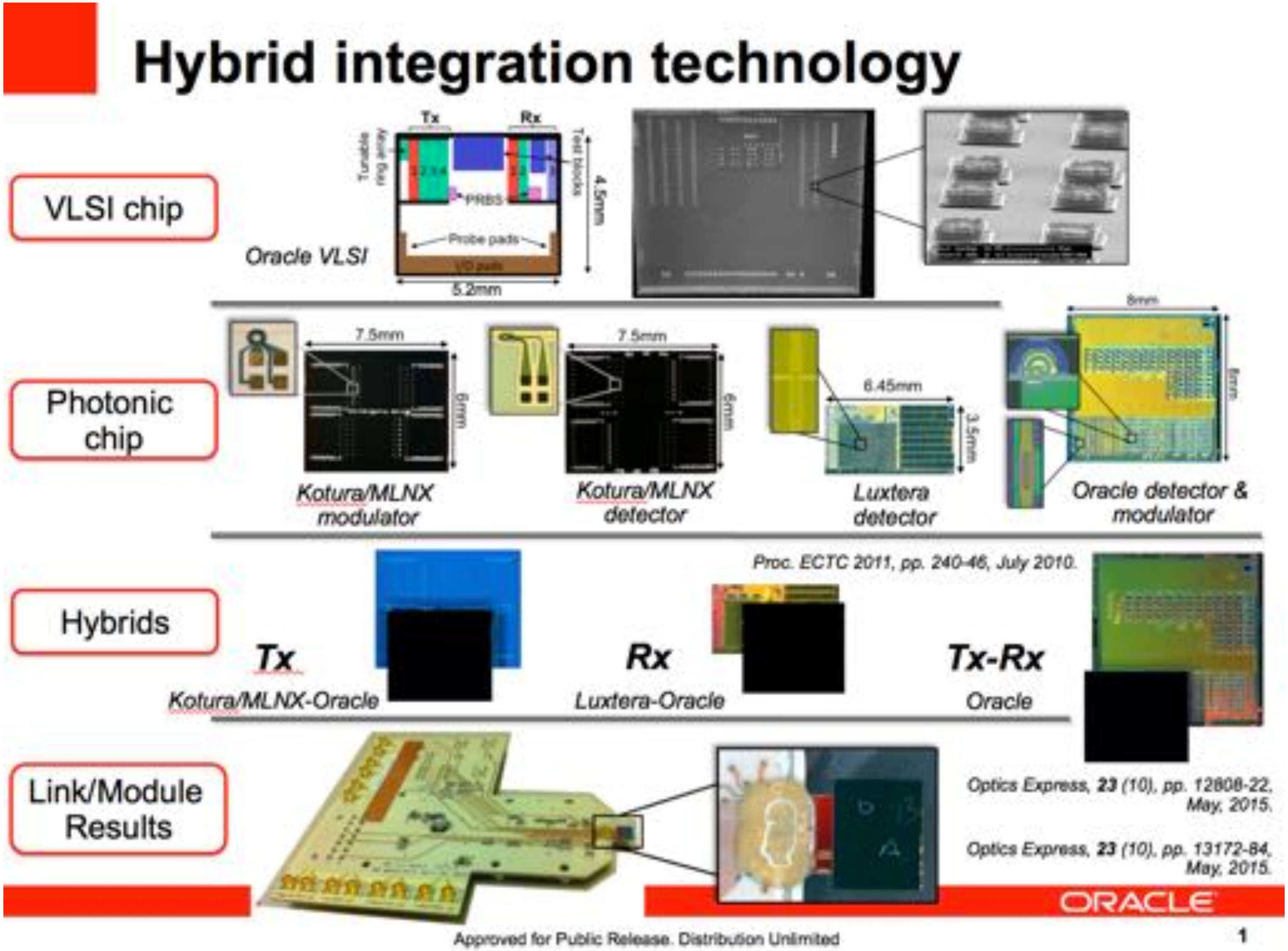
Courtesy of A. Krishnamoorthy, Oracle

Hybrid Integration Platform for Large-Scale Photonic I/O

Electronic and Photonic Co-design

Integration

Packaging and System Impact



Hybrid integration technology

VLSI chip

Photonic chip

Hybrids

Link/Module Results

Tx
Rx
Tunable ring array
Test blocks
PRBS
Probe pads
I/O pads
4.5mm
5.2mm
Oracle VLSI
7.5mm
6mm
Kotura/MLNX modulator
7.5mm
6mm
Kotura/MLNX detector
6.45mm
3.5mm
Luxtera detector
8mm
8mm
Oracle detector & modulator
Tx
Kotura/MLNX-Oracle
Rx
Luxtera-Oracle
Tx-Rx
Oracle
ORACLE

Proc. ECTC 2011, pp. 240-46, July 2010.

Optics Express, **23** (10), pp. 12808-22, May, 2015.

Optics Express, **23** (10), pp. 13172-84, May, 2015.

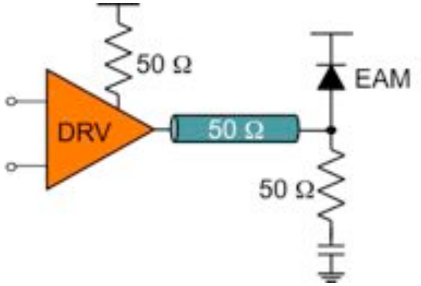
Approved for Public Release. Distribution Unlimited

1

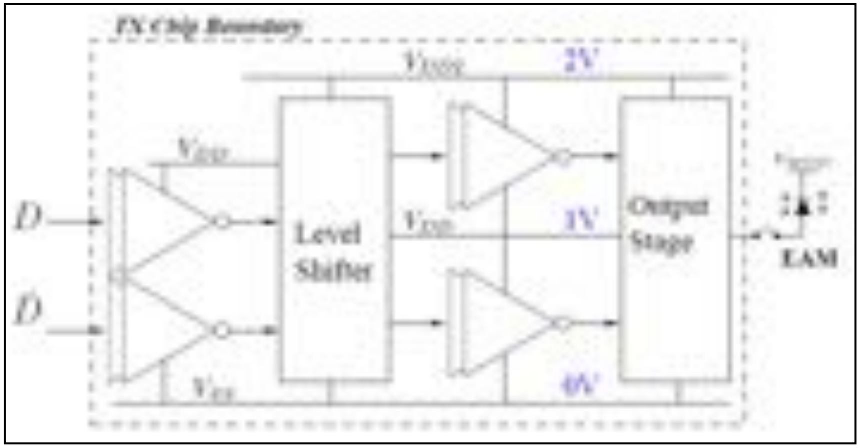
Courtesy of A. Krishnamoorthy, Oracle

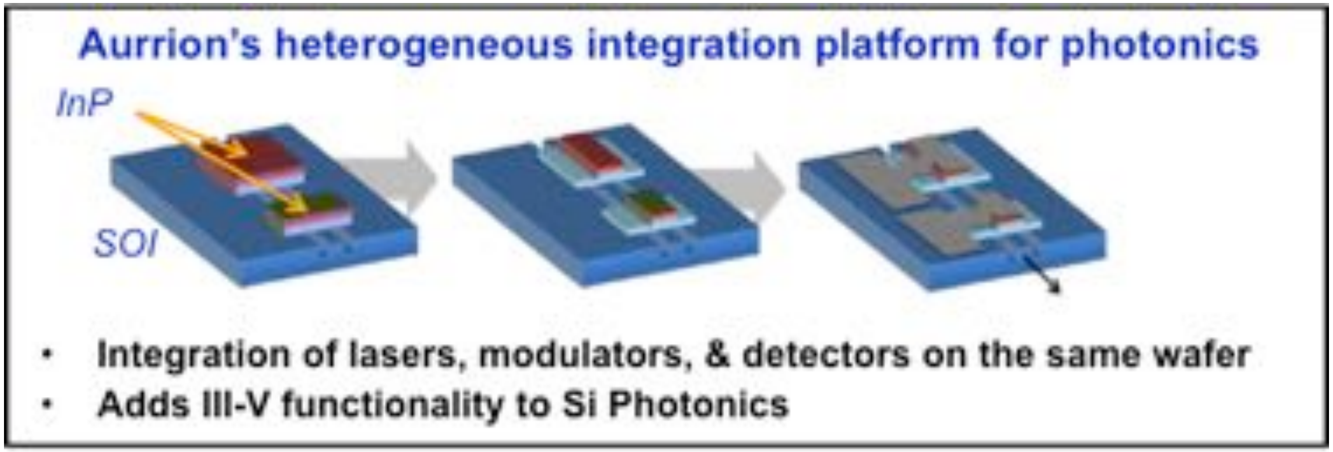
32 nm SOI CMOS-Driven Link (Aurrion/IBM)

Current practice:
Low integration level,
inefficient 50 Ω interfaces



Maximum efficiency: directly drive the EAM



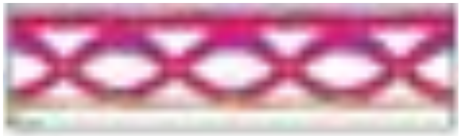


Aurrion's heterogeneous integration platform for photonics

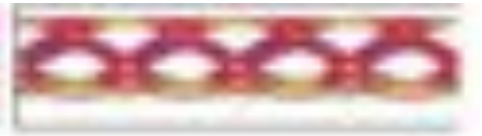
- Integration of lasers, modulators, & detectors on the same wafer
- Adds III-V functionality to Si Photonics

30 Gb/s

TX out



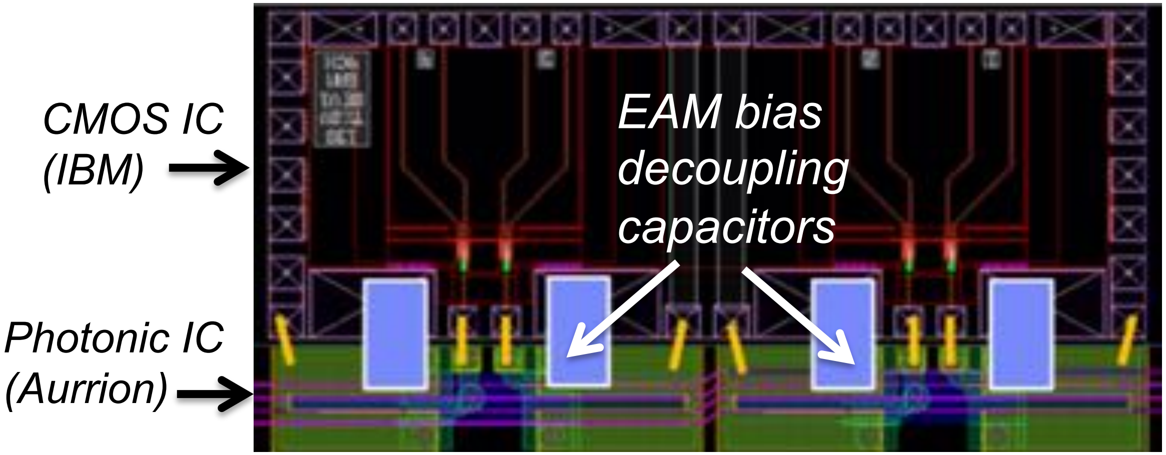
RX out



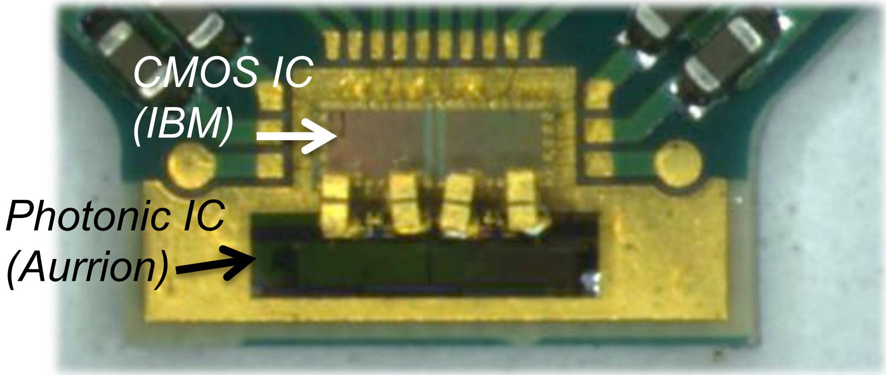
- **3 pJ/bit at 30 Gb/s (not including laser power)**
- **No measured penalty for 10km transmission at 25 Gb/s**
  - N. Dupuis *et al.*, "30Gbps Optical Link Utilizing Heterogeneously Integrated III-V/Si Photonics and CMOS Circuits," *OFC 2014* (post deadline).
  - N. Dupuis *et al.*, "30Gbps Optical Link Combining Heterogeneously Integrated III-V/Si Photonics with 32nm CMOS Circuits," *JLT*, 2015.

Circuit/Photonics/Packaging Co-Design

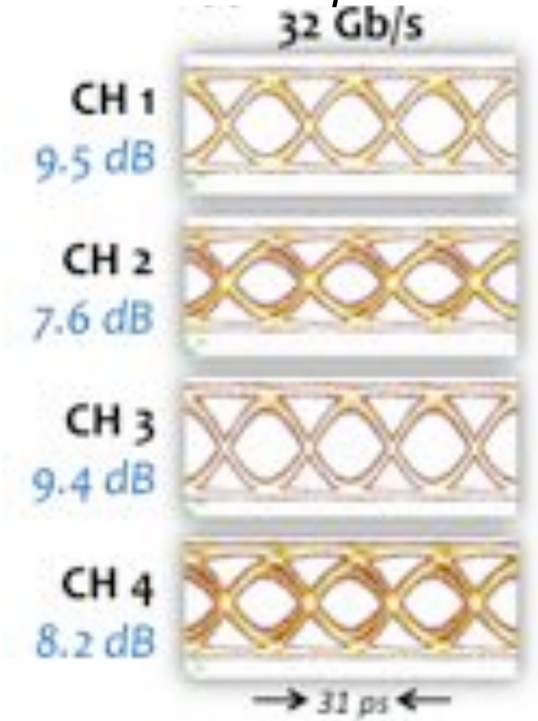
Electronic and photonic chip integration



Demonstrated in hardware

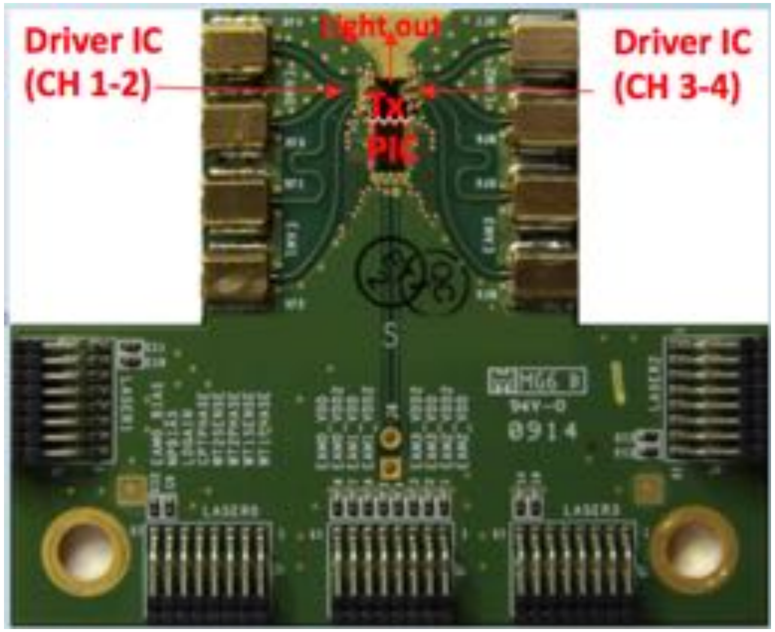


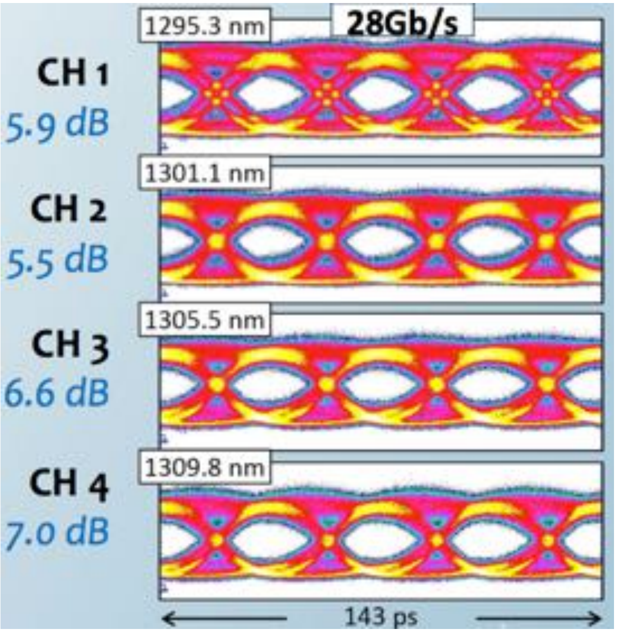
Transmitter output

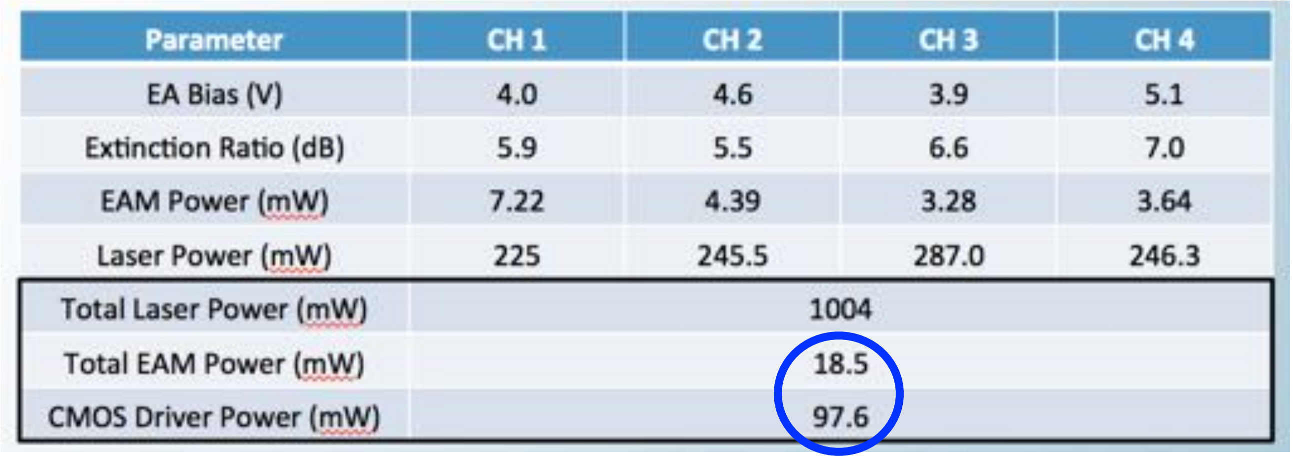


- B. G. Lee *et al.*, “A WDM-Compatible 4 × 32-Gb/s CMOS-Driven Electro-Absorption Modulator Array,” *OFC* 2015.

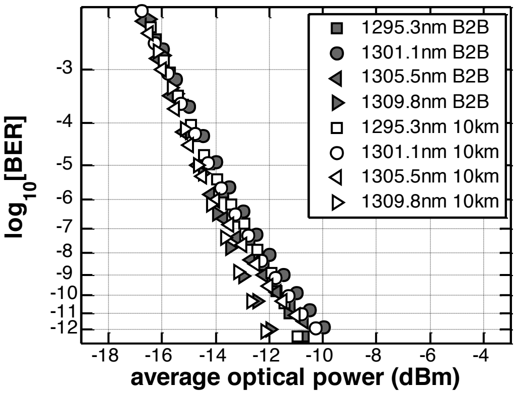
Wall-Plug WDM links (Aurrion/IBM)







Parameter	CH 1	CH 2	CH 3	CH 4
EA Bias (V)	4.0	4.6	3.9	5.1
Extinction Ratio (dB)	5.9	5.5	6.6	7.0
EAM Power (mW)	7.22	4.39	3.28	3.64
Laser Power (mW)	225	245.5	287.0	246.3
Total Laser Power (mW)	1004			
Total EAM Power (mW)	18.5			
CMOS Driver Power (mW)	97.6			

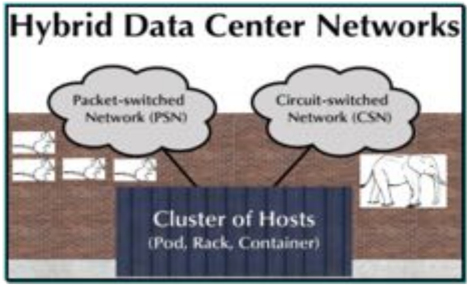


Low modulation power → directly driving EAM, no 50Ω interfaces

- A. Ramaswamy *et al.*, “A WDM 4x28Gbps Integrated Silicon Photonic Transmitter driven by 32nm CMOS driver ICs,” *OFC 2015* (post deadline).

Photonic Switching

What do we mean by ***"fast"*** optical switching?


Hybrid Data Center Networks
Packet-switched Network (PSN)
Circuit-switched Network (CSN)
Cluster of Hosts
(Pod, Rack, Container)

ms-scale

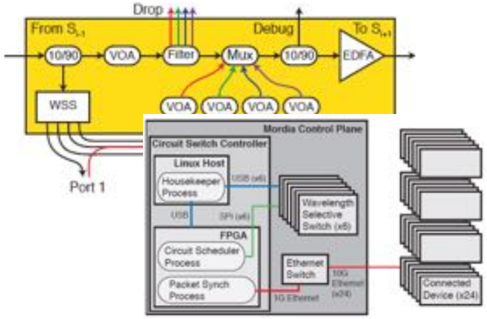
Mice flows over packet switch, elephant flows over OCS

Promise:

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

Challenges:

- Scalability of software scheduler
- Slow reconfiguration time may limit applications



Examples: Mordia/ REACTOR (UCSD)

µs-scale

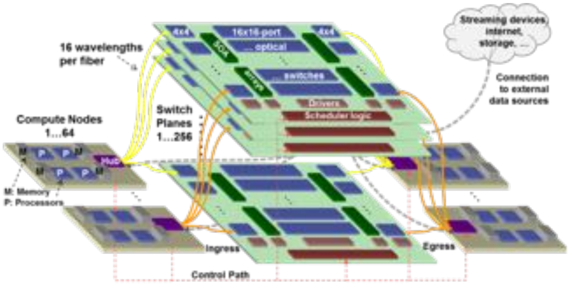
OCS at first switch level, hybrid but much faster then 3D-MEMS

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:

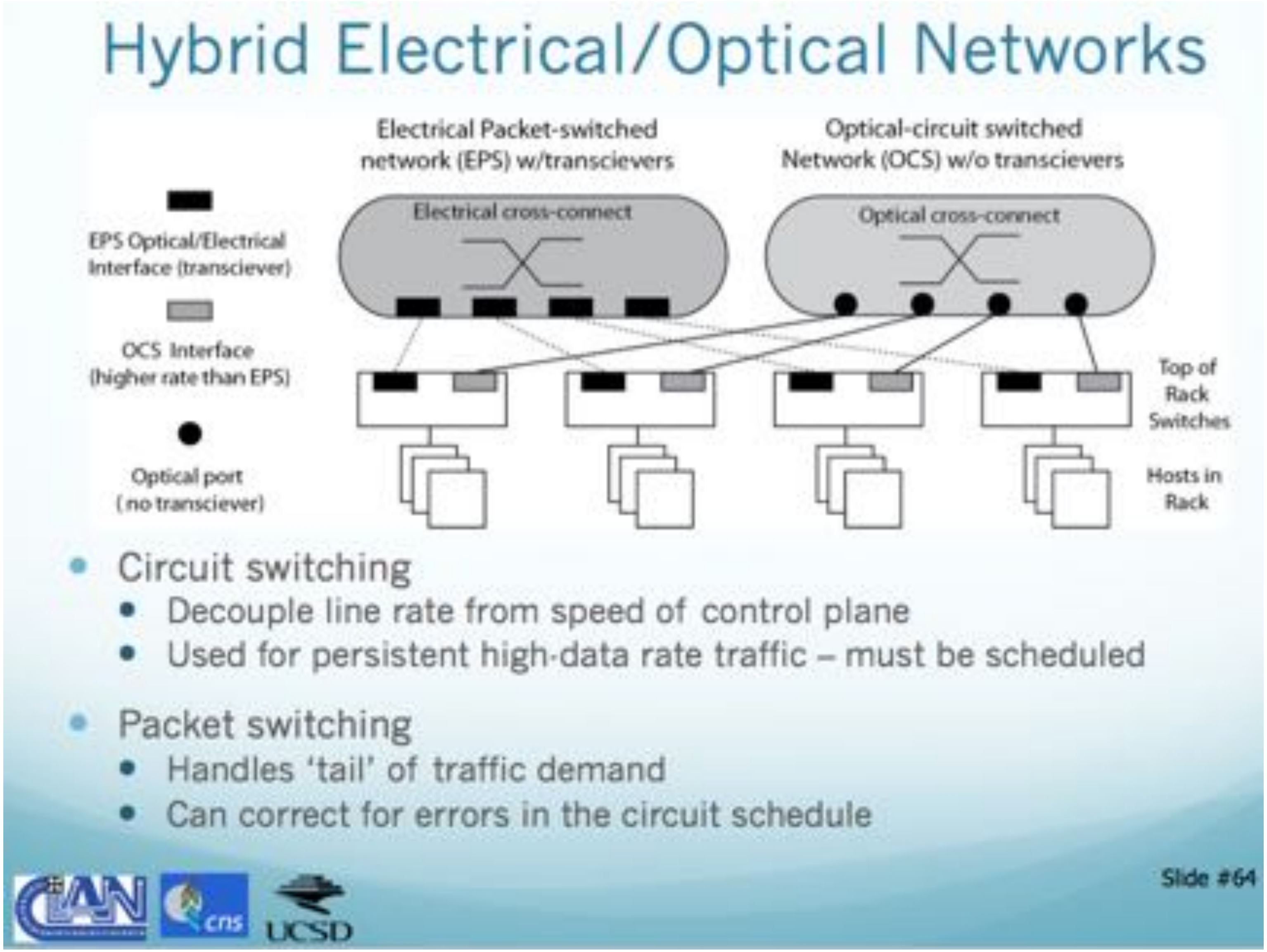
- Switching times ~ packet durations
- More power-efficient than electrical networks of equal BW

Challenges:

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

Adapted from L. Schares, IBM

Hybrid Networks



Hybrid Electrical/Optical Networks

Electrical Packet-switched network (EPS) w/transcievers
Optical-circuit switched Network (OCS) w/o transcievers
Electrical cross-connect
Optical cross-connect
EPS Optical/Electrical Interface (transciever)
OCS Interface (higher rate than EPS)
Optical port (no transciever)
Top of Rack Switches
Hosts in Rack

- 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

Slide #64

Courtesy of Prof. G. Papen and G. Porter, UCSD

Calient: 3-D MEMS OCS

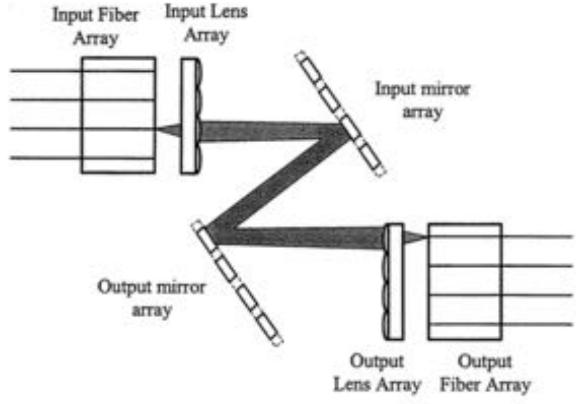
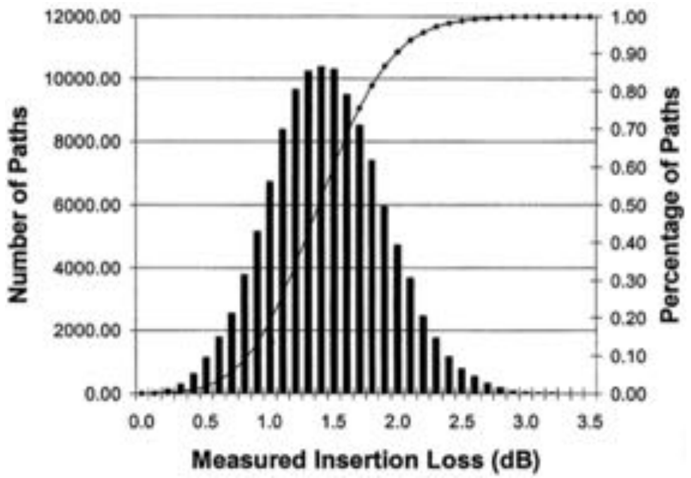


Fig. 3. Switch configuration using 3-D MEMS.



X. Zheng et al., "Three-dimensional MEMS photonic cross-connect switch design and performance," JSTQE, 2003.

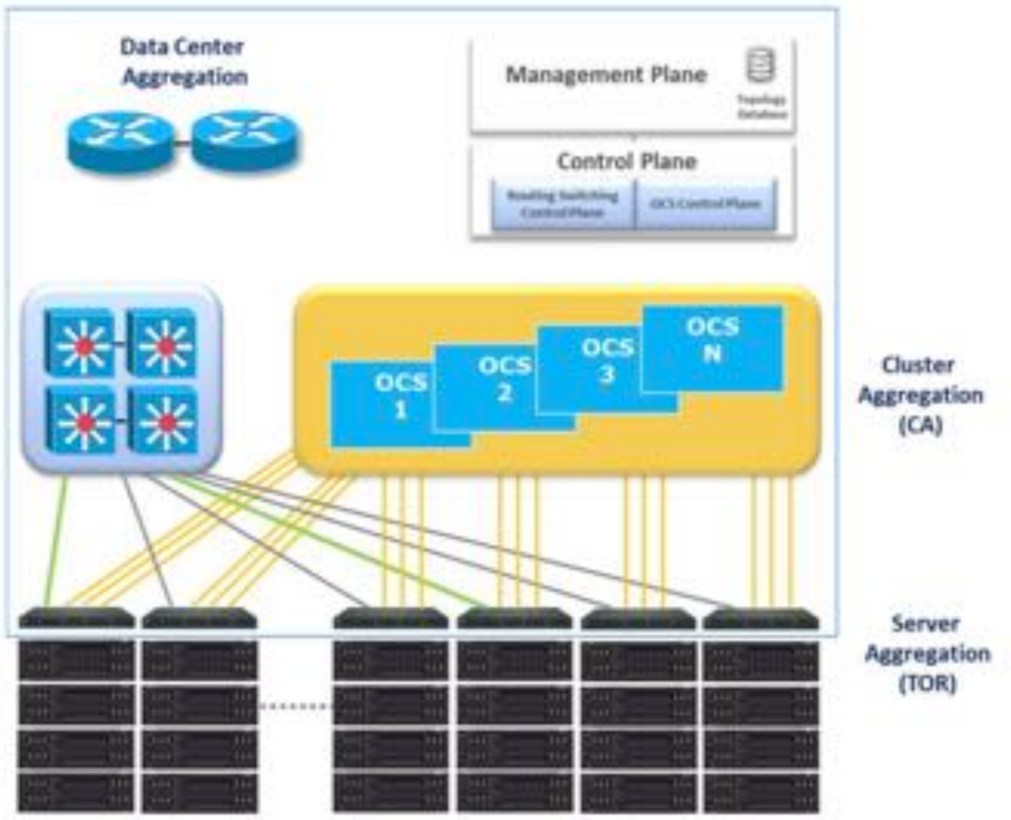
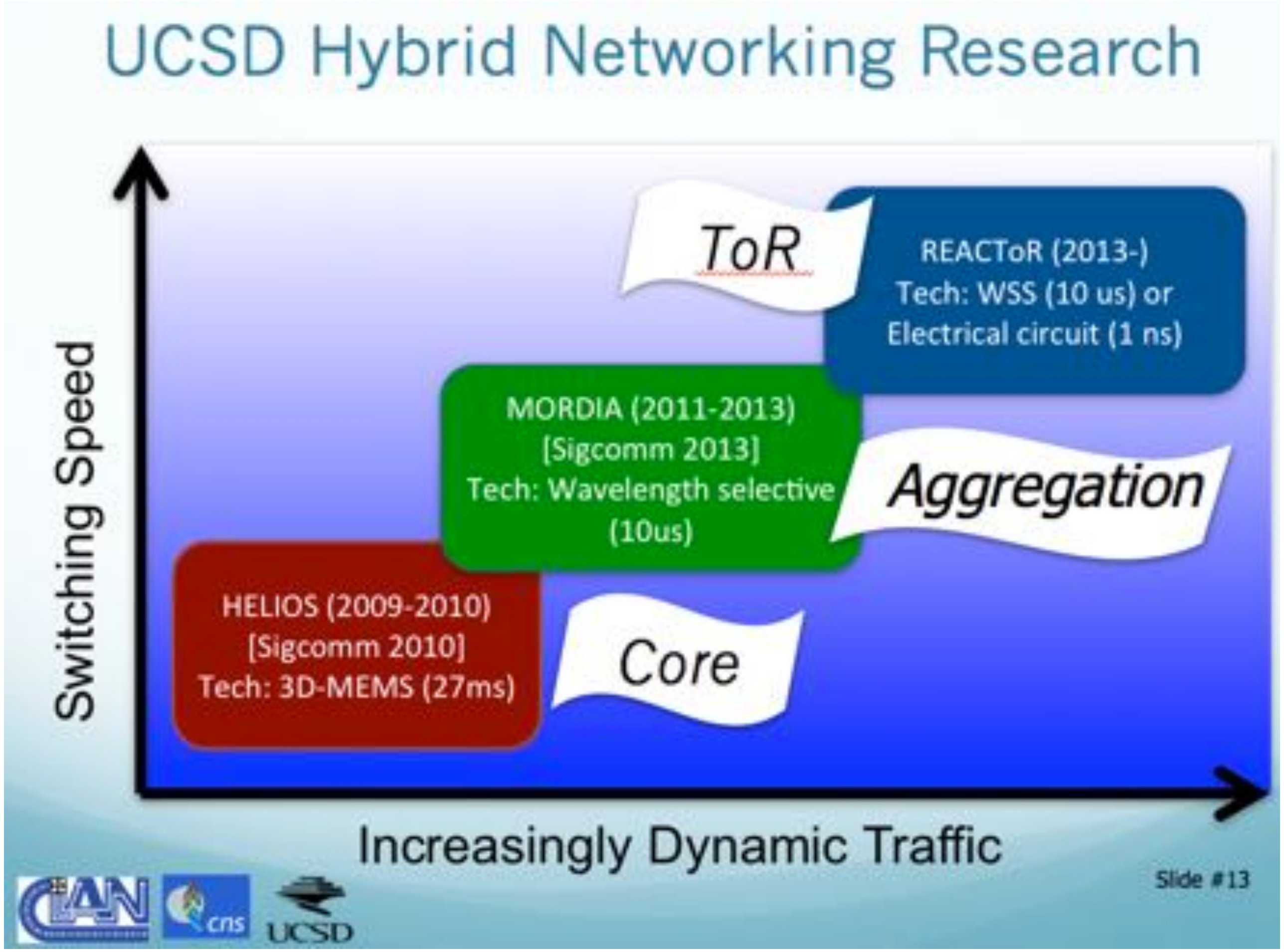


Figure 4: Hybrid Packet-OCS Datacenter Network Architecture

“The Software Defined Hybrid Packet Optical Datacenter Network” whitepaper available at www.calient.net

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

Photonic Switches in Data Centers



UCSD Hybrid Networking Research

Switching Speed

ToR

REACToR (2013-)
Tech: WSS (10 us) or
Electrical circuit (1 ns)

MORDIA (2011-2013)
[Sigcomm 2013]
Tech: Wavelength selective
(10us)

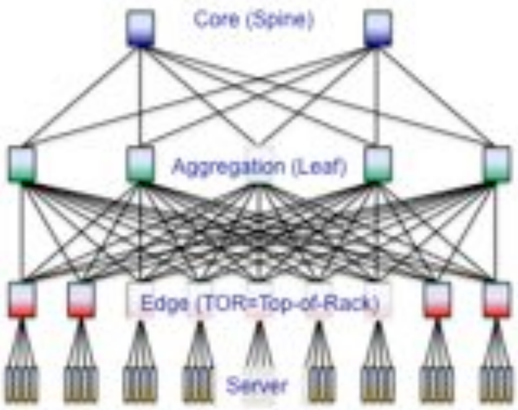
Aggregation

HELIOS (2009-2010)
[Sigcomm 2010]
Tech: 3D-MEMS (27ms)

Core

Increasingly Dynamic Traffic

Slide #13



ToR photonic switches:

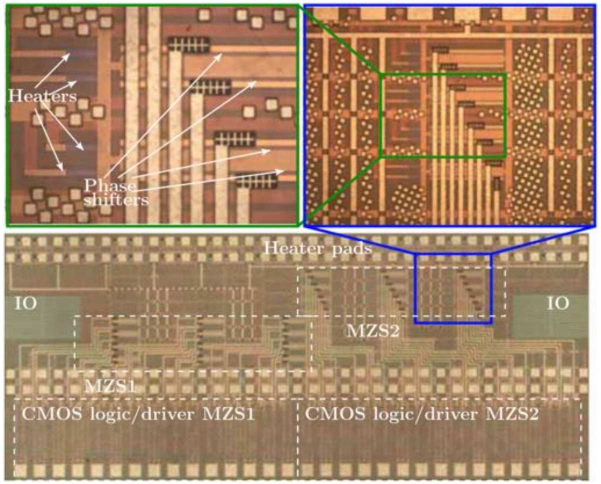
- Fast reconfiguration (dynamic traffic)
- High-radix
- Low cost

Courtesy of Prof. G. Papen and G. Porter, UCSD

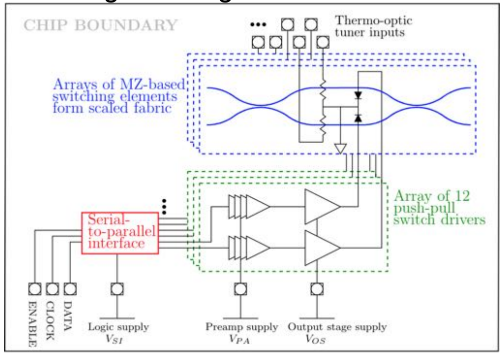
- N. Farrington *et al.,* "Helios: a hybrid electrical/optical switch architecture for modular data centers." *SIGCOMM*, 2011.
- R. Aguinaldo *et al*., "Energy-efficient, digitally-driven "fat pipe" silicon photonic circuit switch in the UCSD MORDIA data-center network." *CLEO* 2104.
- H. Liu *et al.*, "REACToR: A reconfigurable packet and circuit ToR switch," *Photonics Society Summer Topicals*, 2013.

Switch/Driver Integration (IBM)

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

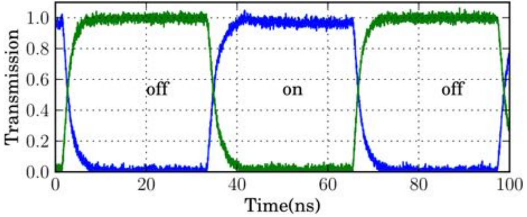


Integrated digital switch drivers


CHIP BOUNDARY
Thermo-optic tuner inputs
Arrays of MZ-based switching elements form scaled fabric
Array of 12 push-pull switch drivers
Serial-to-parallel interface
ENABLE
CLOCK
DATA
Logic supply V_{SI}
Preamp supply V_{PA}
Output stage supply V_{OS}

Fast reconfiguration: 4 ns

Transient response of 2x2 MZ switch

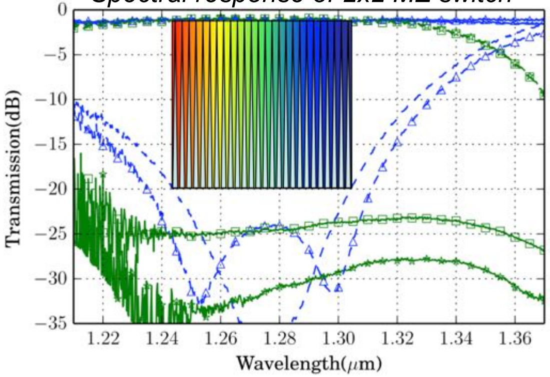


Broad spectral bandwidth:

Routing many wavelength channels

- <-20dB crosstalk over 60 nm BW
- 32 channels at 200GHz spacing

Spectral response of 2x2 MZ switch



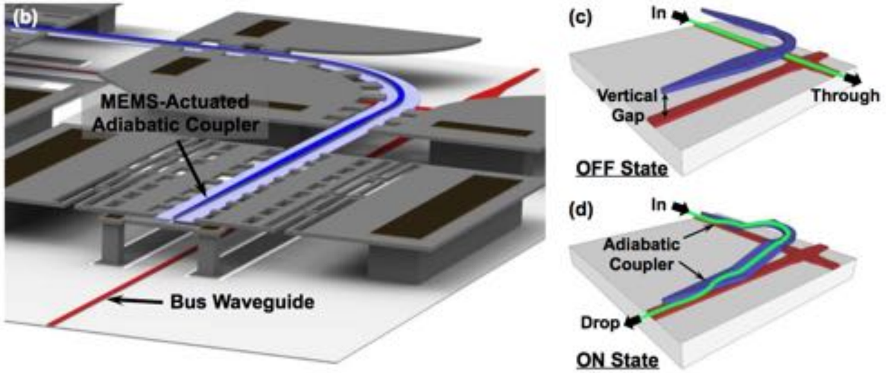
Losses too high, need significant feedback and control to manage crosstalk

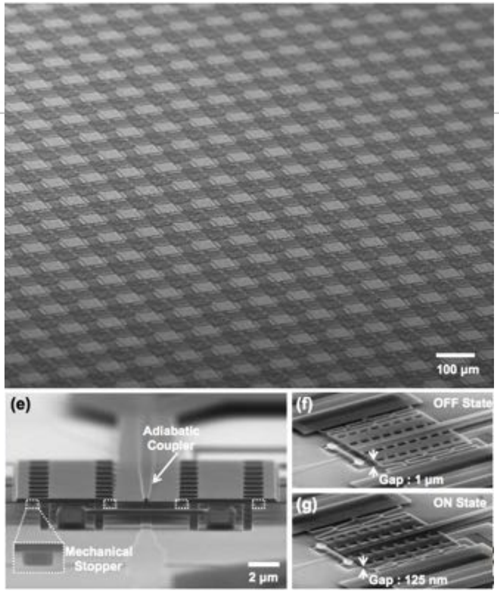
- **High level of electronic/photonic integration demanded**

- N. Dupuis *et al.*, " Modeling and Characterization of a Non-Blocking 4 × 4 Mach-Zehnder Silicon Photonic Switch Fabric," *JLT* 2015.
- N. Dupuis *et al.*, "Design and Fabrication of Low-Insertion-Loss and Low-Crosstalk Broadband 2 × 2 Mach–Zehnder Silicon Photonic Switches," *JLT* 2015.
- B. G. Lee *et al.*, "Monolithic Silicon Integration of Scaled Photonic Switch Fabrics, CMOS Logic, and Device Driver Circuits," *JLT* 2014.
- B. G. Lee *et al.*, "Silicon Photonic Switch Fabrics in Computer Communications Systems," *JLT* 2015.

High Port Count MEMS Switches (UC Berkeley)

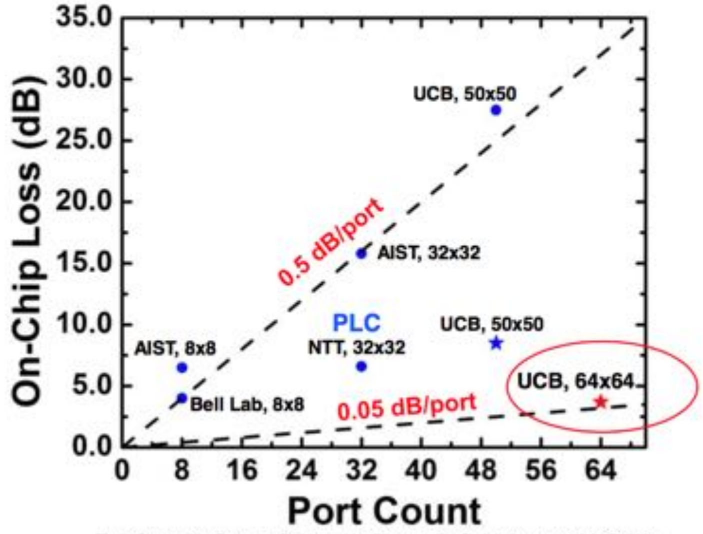
Detailed Schematic of Vertical Adiabatic Coupler Switch





Ming Wu (INOW) 29

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

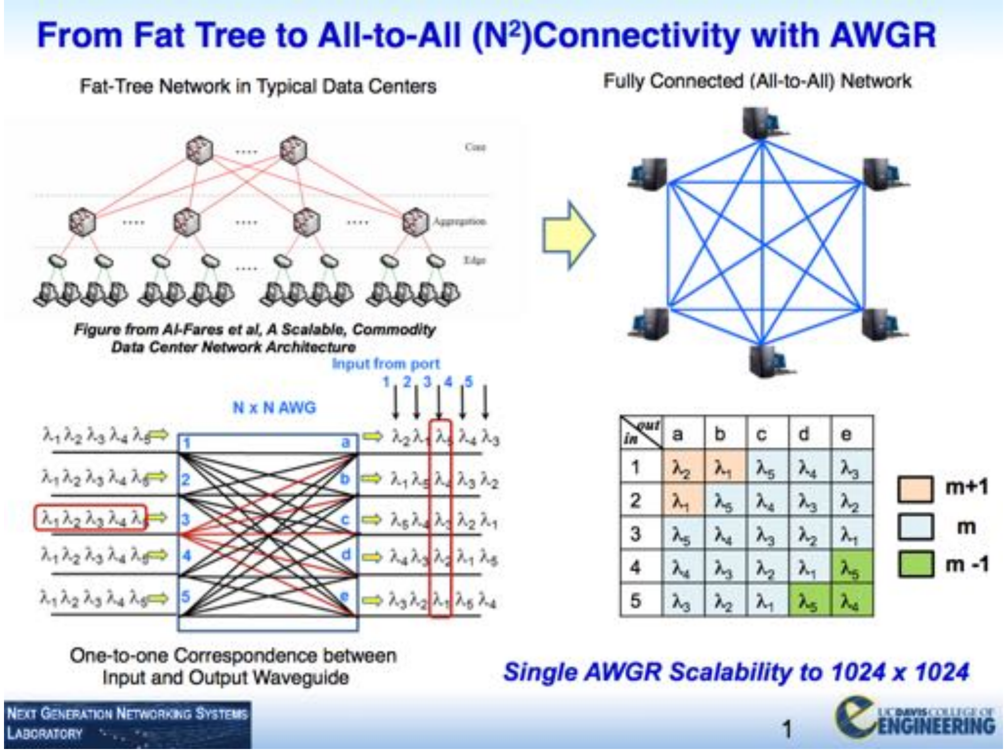


T. J. Seok, N. Quack, S. Han, W. Zhang, R. S. Muller, and M. C. Wu, "64x64 Low-Loss and Broadband Digital Silicon Photonic MEMS Switches," will be presented in ECOC 2015, 2015.

Ming Wu (INOW) 31

Courtesy Professor M. Wu, UC Berkeley

Wavelength Routing: AWGR



From Fat Tree to All-to-All (N²)Connectivity with AWGR

Fat-Tree Network in Typical Data Centers
Core
Aggregation
Edge
Fully Connected (All-to-All) Network
Input from port
1 2 3 4 5
N x N AWG
$\lambda_1 \lambda_2 \lambda_3 \lambda_4 \lambda_5$
1
2
3
4
5
a
b
c
d
e
$\lambda_2 \lambda_1 \lambda_5 \lambda_4 \lambda_3$
$\lambda_1 \lambda_5 \lambda_4 \lambda_3 \lambda_2$
$\lambda_5 \lambda_4 \lambda_3 \lambda_2 \lambda_1$
$\lambda_4 \lambda_3 \lambda_2 \lambda_1 \lambda_5$
$\lambda_3 \lambda_2 \lambda_1 \lambda_5 \lambda_4$

Figure from Al-Fares et al, A Scalable, Commodity Data Center Network Architecture

One-to-one Correspondence between Input and Output Waveguide

in \ out	a	b	c	d	e
1	λ_2	λ_1	λ_5	λ_4	λ_3
2	λ_1	λ_5	λ_4	λ_3	λ_2
3	λ_5	λ_4	λ_3	λ_2	λ_1
4	λ_4	λ_3	λ_2	λ_1	λ_5
5	λ_3	λ_2	λ_1	λ_5	λ_4

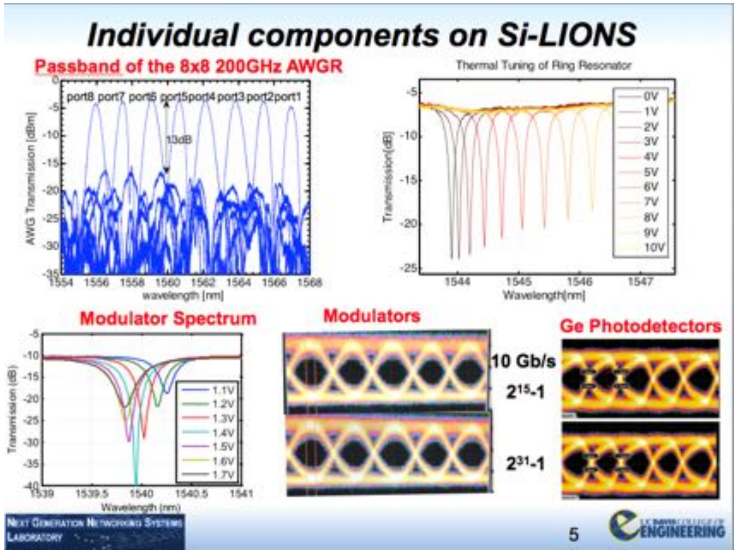
🟧 m+1
⬜ m
🟩 m -1

Single AWGR Scalability to 1024 x 1024

NEXT GENERATION NETWORKING SYSTEMS LABORATORY 1 UC DAVIS COLLEGE OF ENGINEERING

200GHz 8x8 Si-LIONS Chip

NEXT GENERATION NETWORKING SYSTEMS LABORATORY 4 UC DAVIS COLLEGE OF ENGINEERING

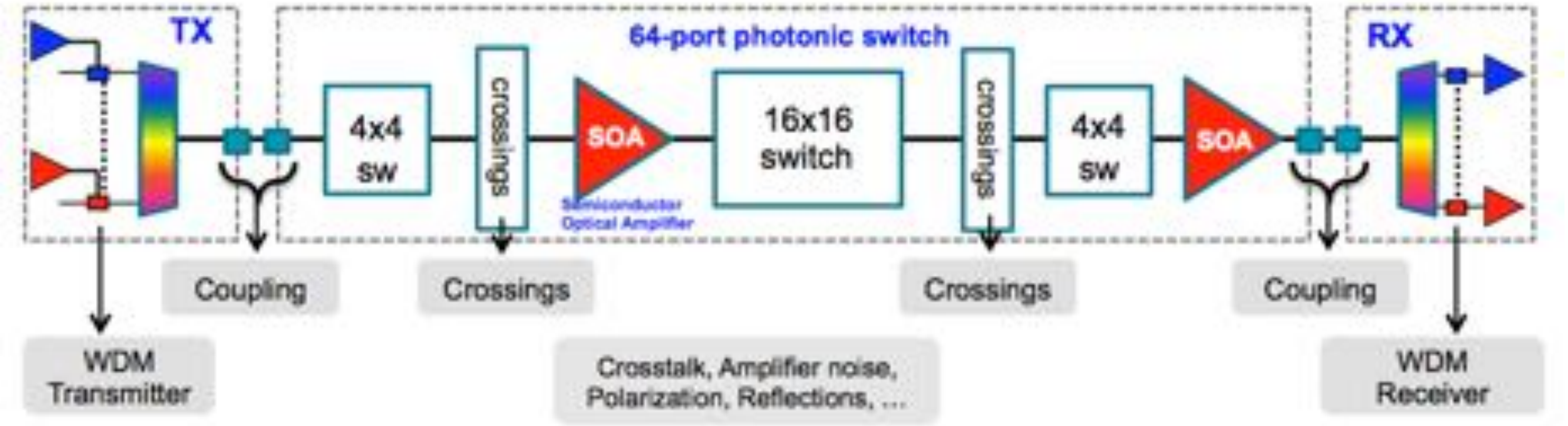


Individual components on Si-LIONS

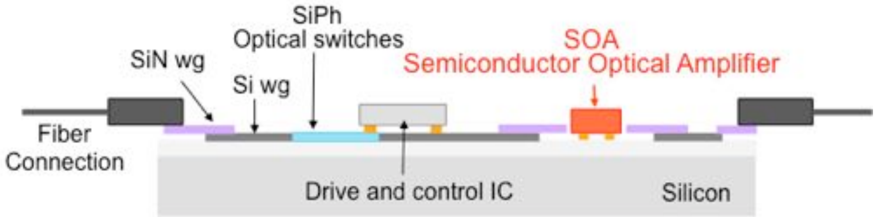
NEXT GENERATION NETWORKING SYSTEMS LABORATORY 5 UC DAVIS COLLEGE OF ENGINEERING

Courtesy of Professor S. J. Ben Yoo, UC Davis

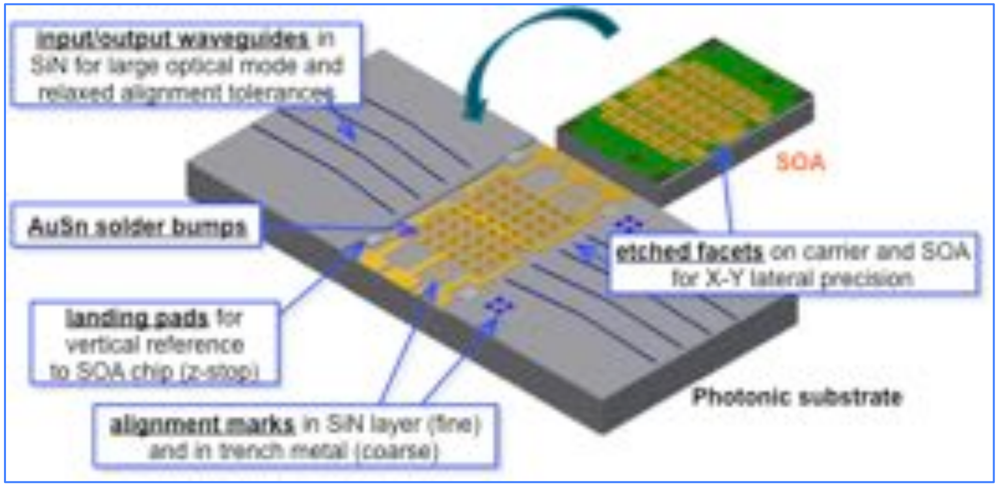
Integrating Optical Gain for Scalability



Hybrid Approach (IBM):



Solder attachment of SOA arrays to photonic carriers



input/output waveguides in SiN for large optical mode and relaxed alignment tolerances

AuSn solder bumps

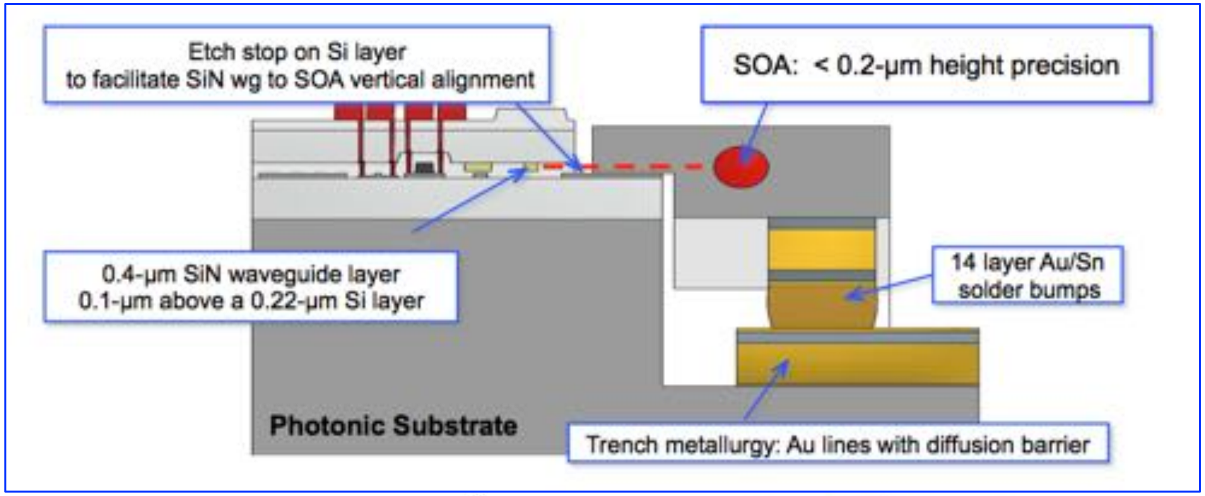
landing pads for vertical reference to SOA chip (z-stop)

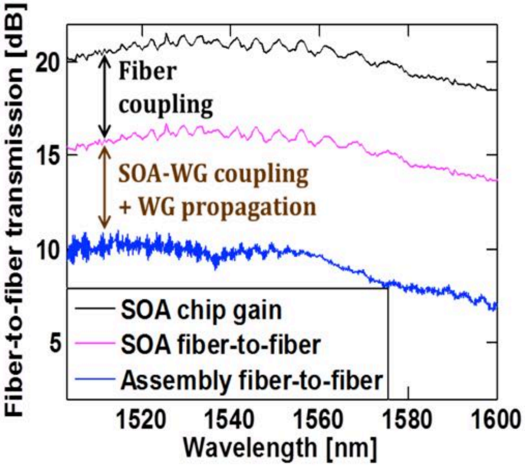
alignment marks in SiN layer (fine) and in trench metal (coarse)

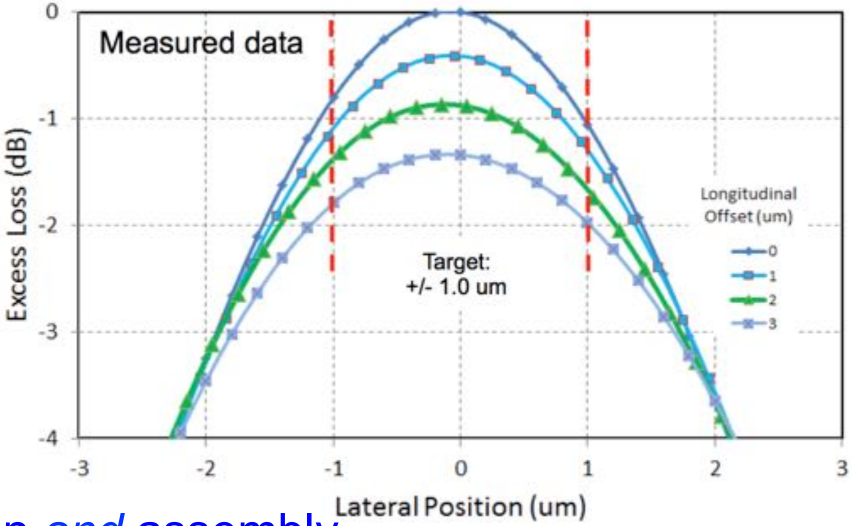
etched facets on carrier and SOA for X-Y lateral precision

- R. A. Budd *et al.*, “Semiconductor Optical Amplifier (SOA) Packaging for Scalable and Gain-Integrated Silicon Photonic Switching Platforms,” *ECTC 2015.*
- L. Schares *et al.*, “Etched-Facet Semiconductor Optical Amplifiers for Gain-Integrated Photonic Switch Fabrics,” *ECOC* 2015.

Challenging Packaging


Etch stop on Si layer
to facilitate SiN wg to SOA vertical alignment
SOA: < 0.2-µm height precision
0.4-µm SiN waveguide layer
0.1-µm above a 0.22-µm Si layer
14 layer Au/Sn
solder bumps
Photonic Substrate
Trench metallurgy: Au lines with diffusion barrier



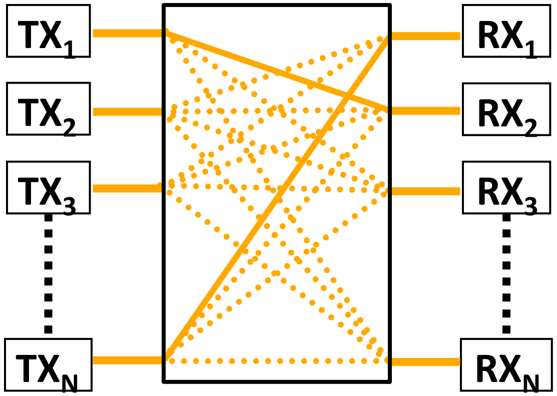


- Challenging fabrication *and* assembly
- Precise alignment required for *each* SOA chip

R. A. Budd *et al.*, “Semiconductor Optical Amplifier (SOA) Packaging for Scalable and Gain-Integrated Silicon Photonic Switching Platforms,” *ECTC 2015.*

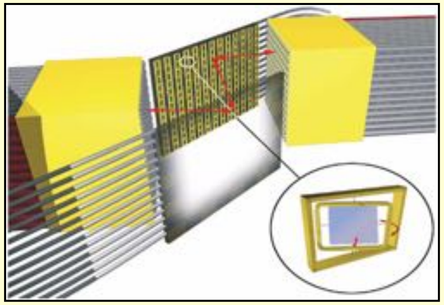
UCSB

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

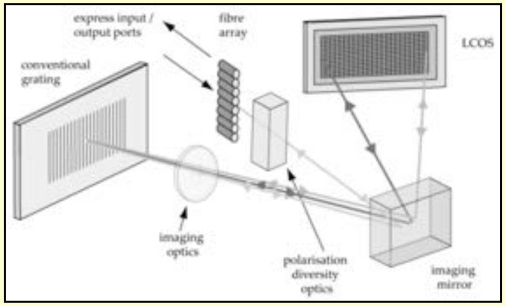


e.g. Calient: 3D-MEMS

[J. Opt. Netw. **6** (1) 19]

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

Microsecond-scale

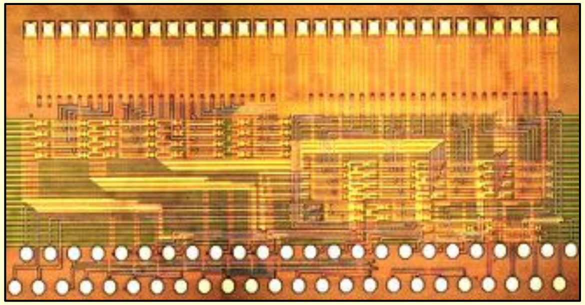


e.g. Finisar: LCOS

[OFC 2006, OTuF2]

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

Nanosecond-scale



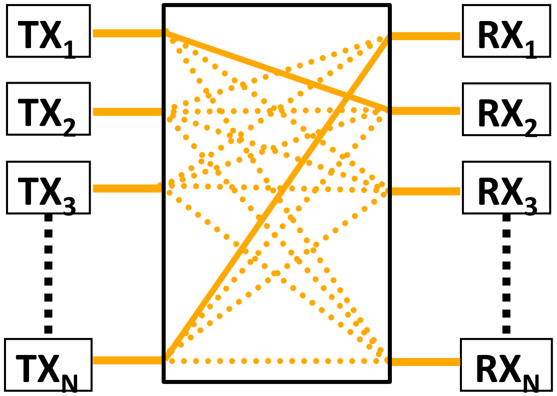
e.g. IBM: Photonics

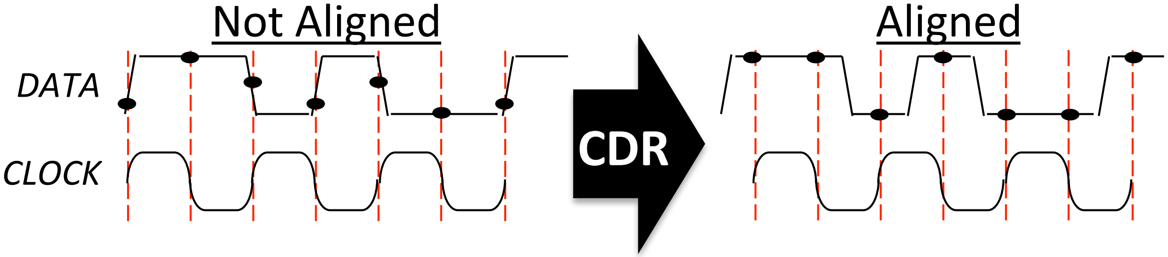
[OFC 2013, PDP5C.3]

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

Courtesy of B. Lee, IBM

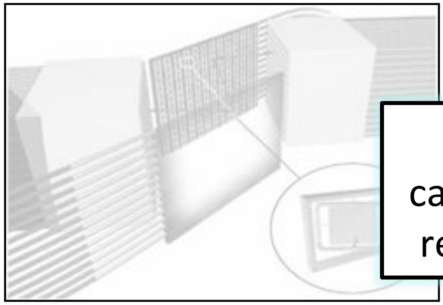
Switching Time = Switch Reconfiguration Time + Link Synchronization Time



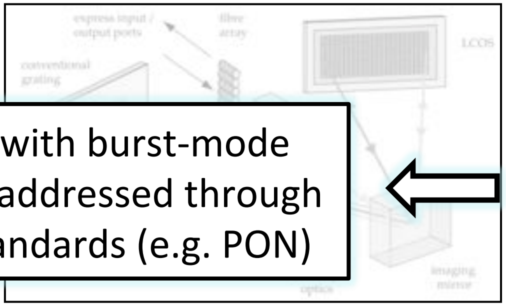


Receivers must adapt to abrupt changes in incoming data amplitude and phase

Millisecond-scale

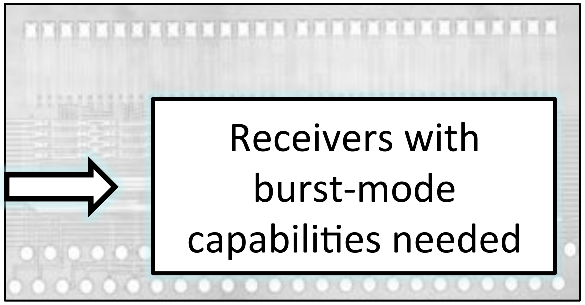


Microsecond-scale



Receivers with burst-mode capabilities addressed through relevant standards (e.g. PON)

Nanosecond-scale



Receivers with burst-mode capabilities needed

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

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

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

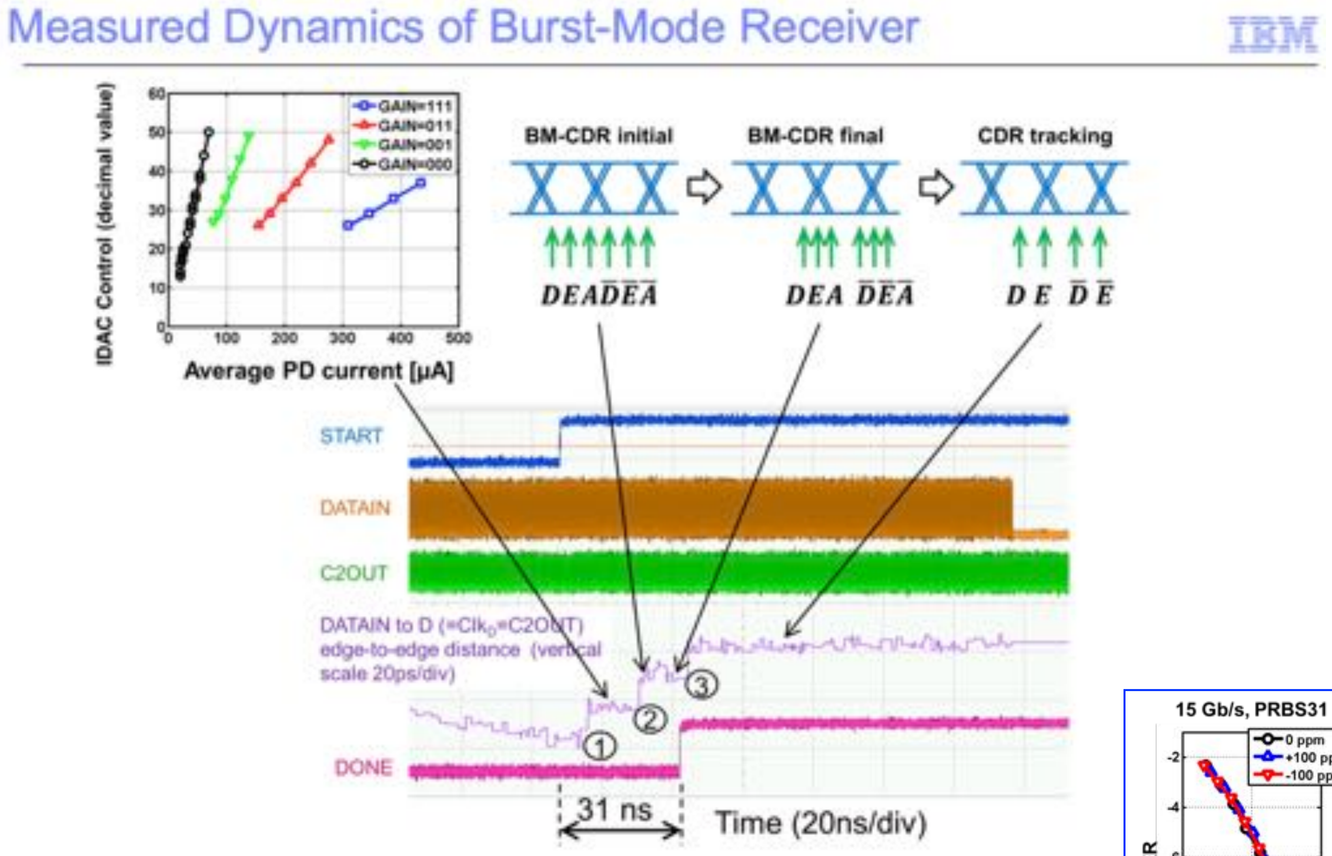
- Hybrid networks
- Reconfigure at flow-level
- Hardware control (FPGA)
- Scalable (10's to 100 ports)

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

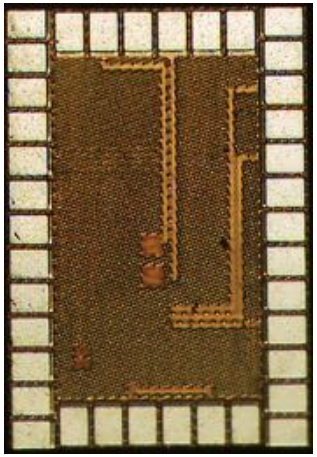
- Reconfigure at packet granularity
- Quasi-packet switching with buffering at end points
- Limited scalability (10's of ports)

Courtesy of B. Lee, IBM

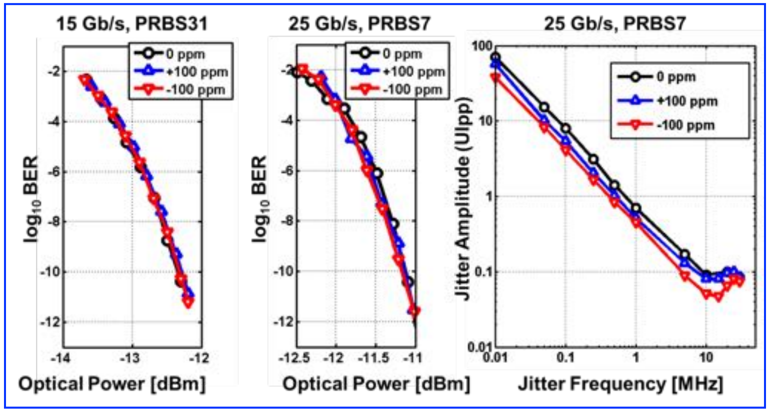
Complete Burst-Mode Receiver (31ns lock, 4 pJ/bit)



Measured Dynamics of Burst-Mode Receiver



32nm SOI CMOS
(1.3 mm × 0.9 mm)



Fast photonic routing and switching fabrics must have companion electronics

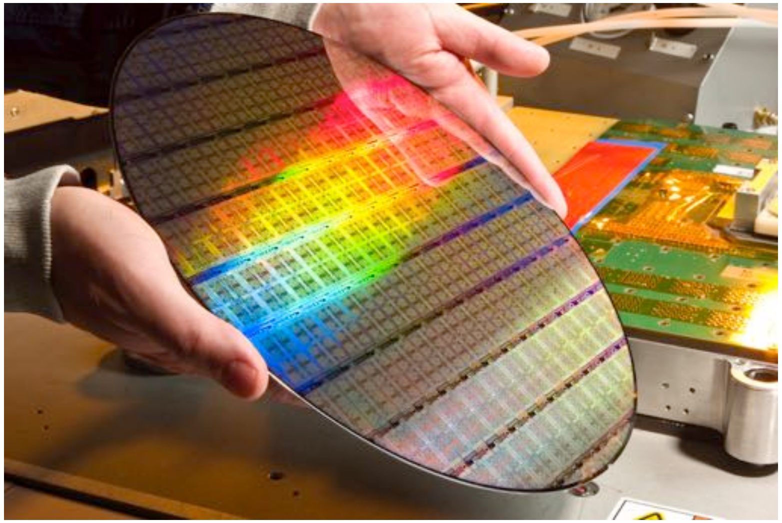
- **Needs more research in the community**

Enables fine-grained power management

- A. R. Rylyakov *et al.*, “A 25 Gb/s Burst-Mode Receiver for Low Latency Photonic Switch Networks ,” *OFC 2015.*
- A. R. Rylyakov *et al.*, “A 25 Gb/s Burst-Mode Receiver for Low Latency Photonic Switch Networks ,” *JSSC 2015.*

Path Forward: Large-Scale Electronic/Photonic Integration

Si Photonics to Enable Next Gen Data Centers



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

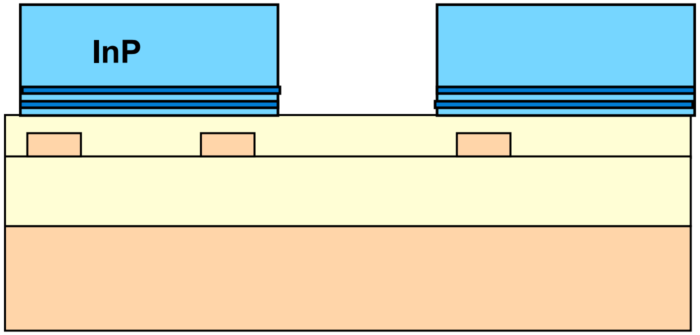
Advantages for the data center:

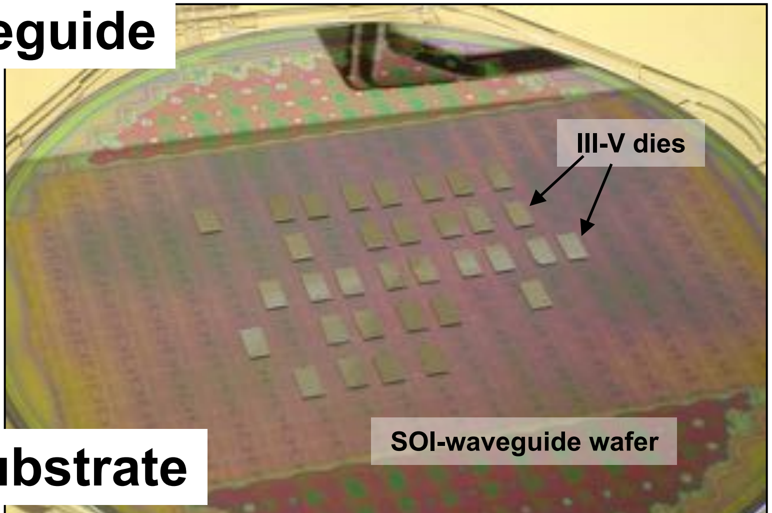
- Single-mode → multi-kilometer links
- Wavelength-division multiplexing (WDM) → high BW/fiber
- High integration level → many devices needed for switching and high-speed interconnects

Heterogeneous integration

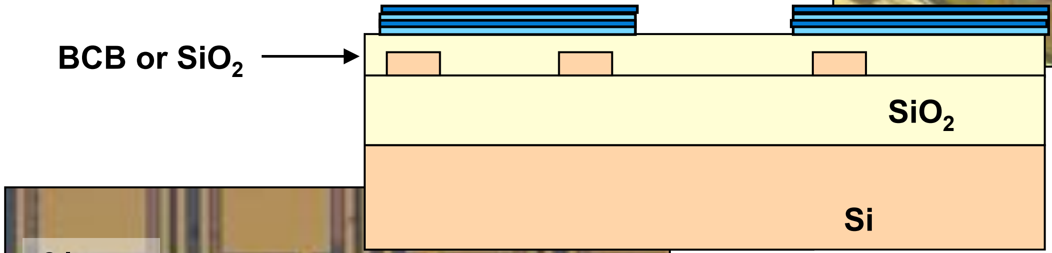
UCSB, Intel, HP, Ghent, IMEC,TIT, Caltech, NTT, Aurrion

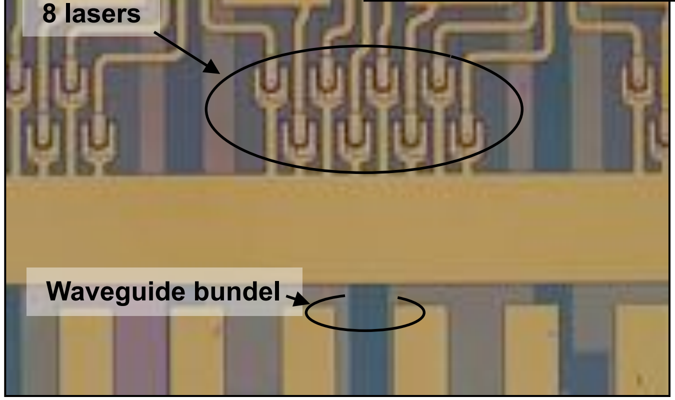
Step 1: Bond InP-dies on SOI waveguide



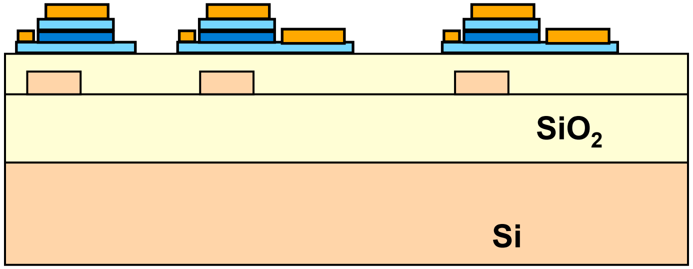


Step 2: Remove substrate





Step 3: Process lasers <u>at wafer scale</u>

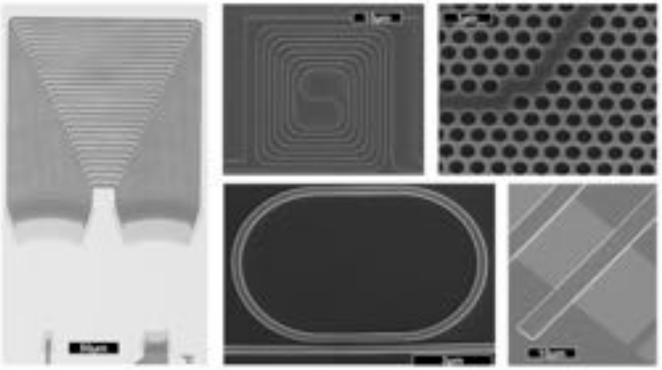


Figures: IMEC/Ghent; Slide courtesy of Prof. J. Bowers, UCSB

Need for Many Functions: Heterogeneous Integration of 6 Photonic Platforms

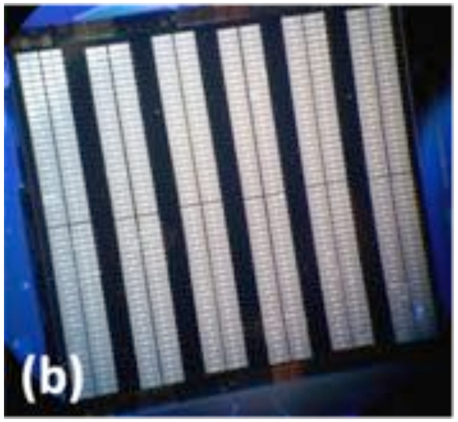
GaAs

Silicon

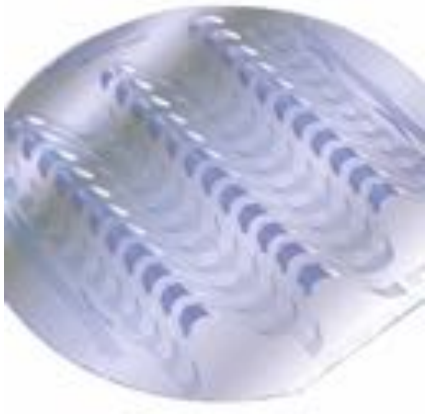


$LiNbO_3$

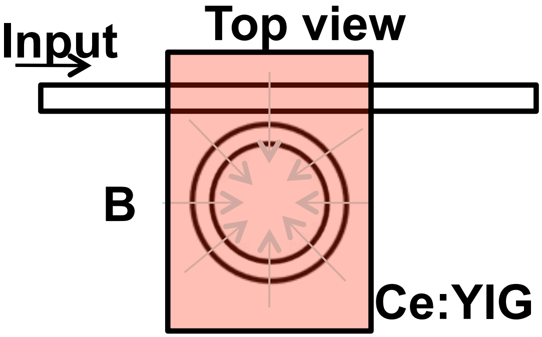
InP



SiN/SiON/SiO2



Ce:YIG Isolator



Heck et al. JSTQE 2013

Slide courtesy of Prof. J. Bowers, UCSB

Large Scale Integration for Switching

Large-scale photonic integration

- Switching
- Amplification
- Control

SOA
SOA
SOA
SOA
SOA
SOA

Hybrid III-V on Si SOAs Demonstrated in Multiple Platforms

UCSB[1], 2006

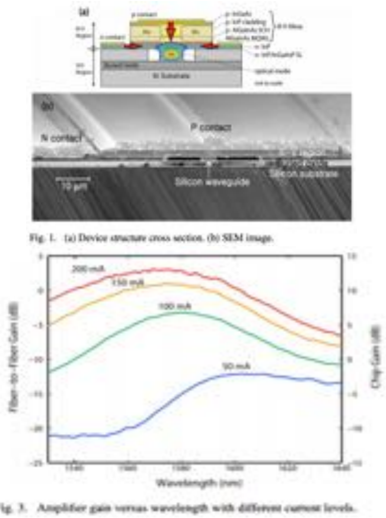
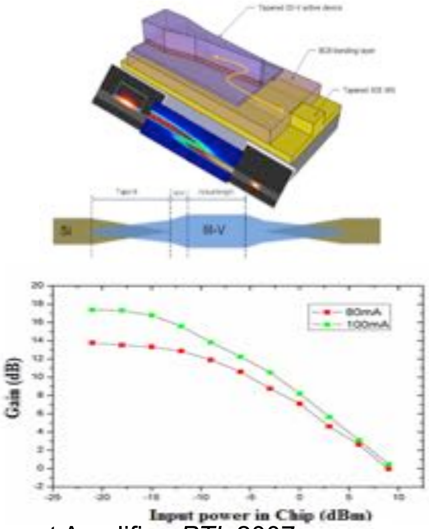


Fig. 1. (a) Device structure cross section. (b) SEM image.

Fig. 3. Amplifier gain versus wavelength with different current levels.

Ghent University-IMEC[2], 2012



III-V Lab[3], 2014

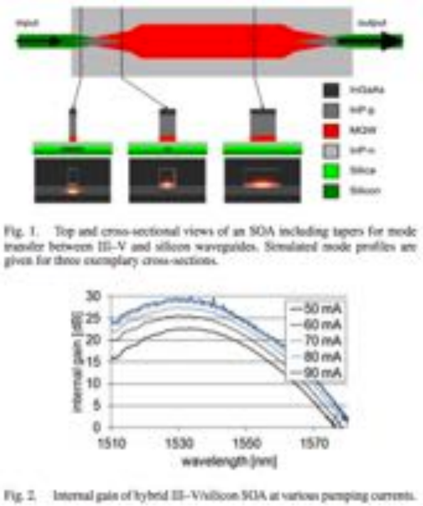
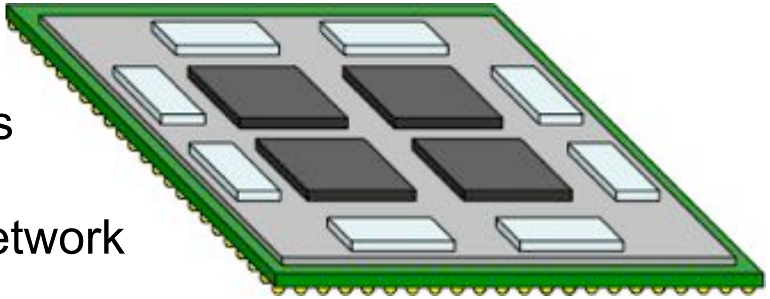


Fig. 1. Top and cross-sectional views of an SOA including tapers for mode transfer between III–V and silicon waveguides. Simulated mode profiles are given for three exemplary cross-sections.

Fig. 2. Internal gain of hybrid III–V/silicon SOA at various pumping currents.

1. H. Park *et al.*, “A Hybrid AlGaInAs-Silicon Evanescent Amplifier, *PTL* 2007.
2. S. Keyvaninia et al., “A Highly Efficient Electrically Pumped Optical Amplifier Integrated on a SOI Waveguide Circuit,” *Group IV Photonics* 2012.
3. G.-H. Duan *et al.*, “New Advances on Heterogeneous Integration of III-V on Silicon,” *JLT* 2015.

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 → WDM
- Low cost → compatibility with high-volume electronics manufacturing
- Low loss → maximize link power efficiency by minimizing required laser power

Thousands of electrical interfaces: connecting electronics to photonics

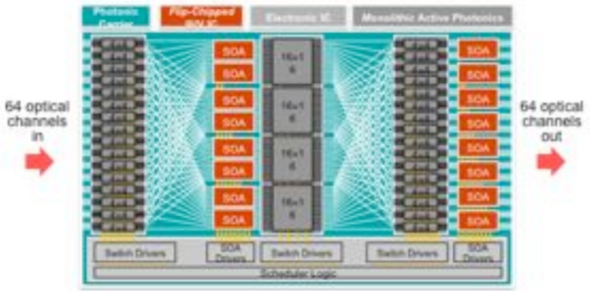
- High density, high-speed, low-power
- New functions: rapid synchronization for low latency switching, power management
- Specialized chip I/O co-designed with and only for photonics, no general purpose cores

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**

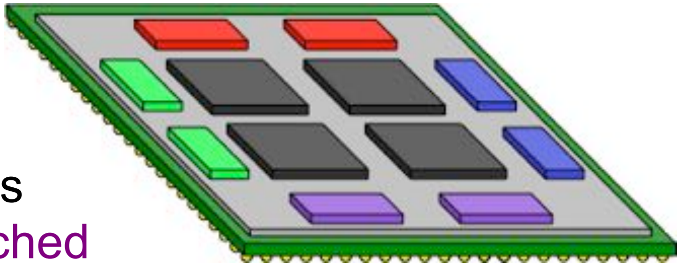
Closing Thoughts

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. Schares, M. Taubenblatt, and many others…

Slides:

M. Tan (HP), P. De Dobbelaere (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)