150 COLUMBIA | ENGINEERING

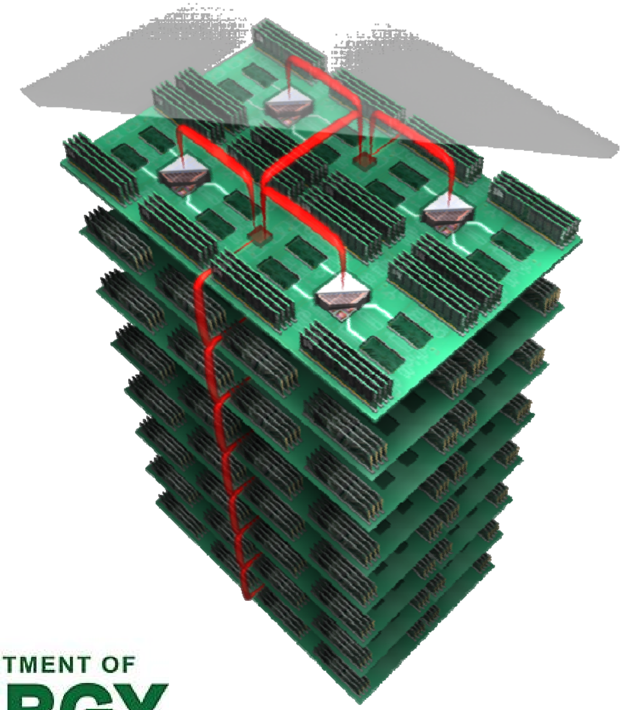
The Fu Foundation School of Engineering and Applied Science

1864–2014

Scalable Computing Systems with Optically Enabled Data Movement

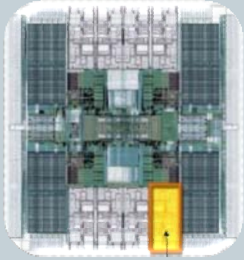
Keren Bergman

Lightwave Research Laboratory,
Columbia University

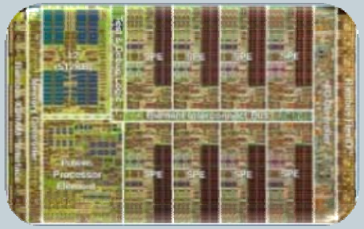


Computation to Communications Bound

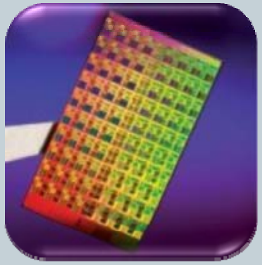
Computing platforms with increased **parallelism** at all scales:



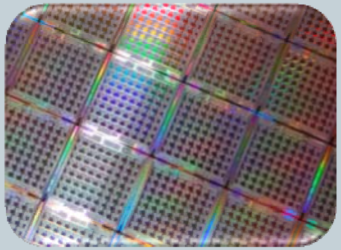
Sun Niagara
8 cores
2005



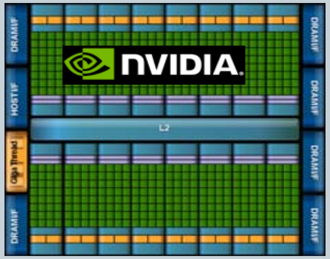
Sony/Toshiba/IBM Cell
9 cores
2006



Intel Polaris
80 cores
2007



Tilera TILE-Gx100
100 cores
2009



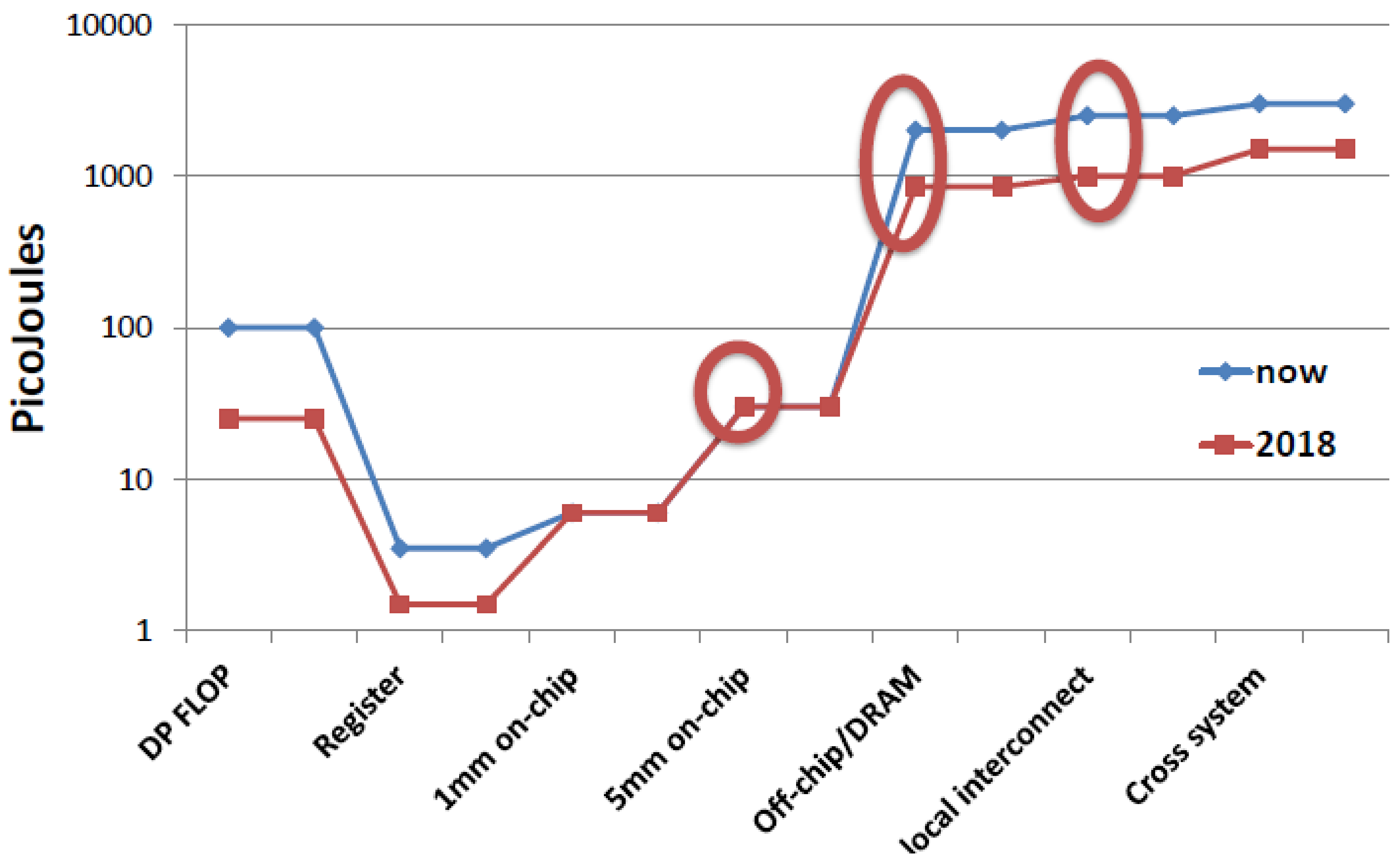
NVIDIA Fermi
512 cores
2012

Handheld
System-on-Chip

Embedded Systems

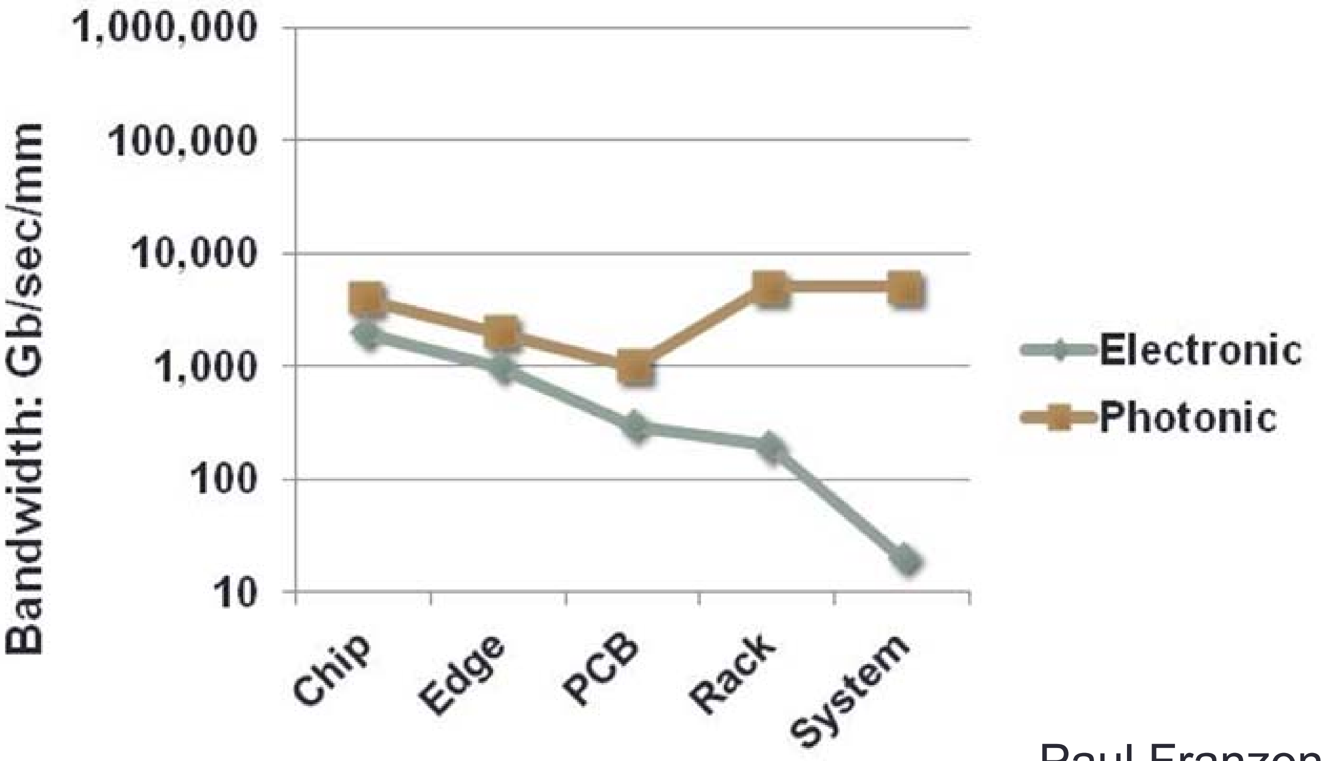
Data Centers

Data Movement Dominates– Energy



John Shalf, LBL

Data movement bandwidth taper challenge



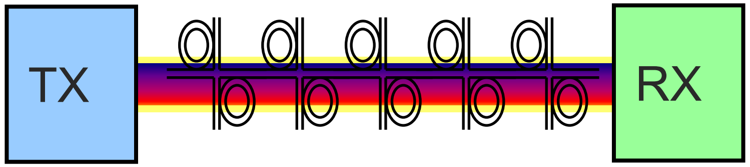
Paul Franzon, NCSU

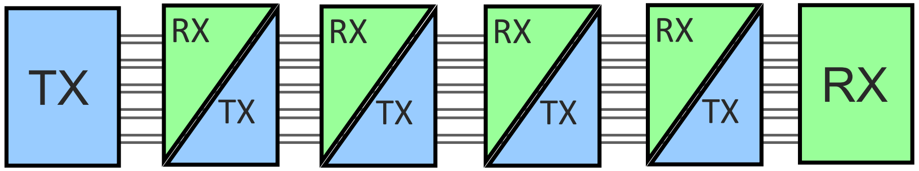
Bandwidth taper for conventional electronic interconnect reduces by:

➢***2 orders of magnitude***

➢as data propagates from the chip, across the die and the system racks

Photonic Interconnects for Computing Platforms: Change the Rules for Bandwidth-per-Watt





PHOTONICS

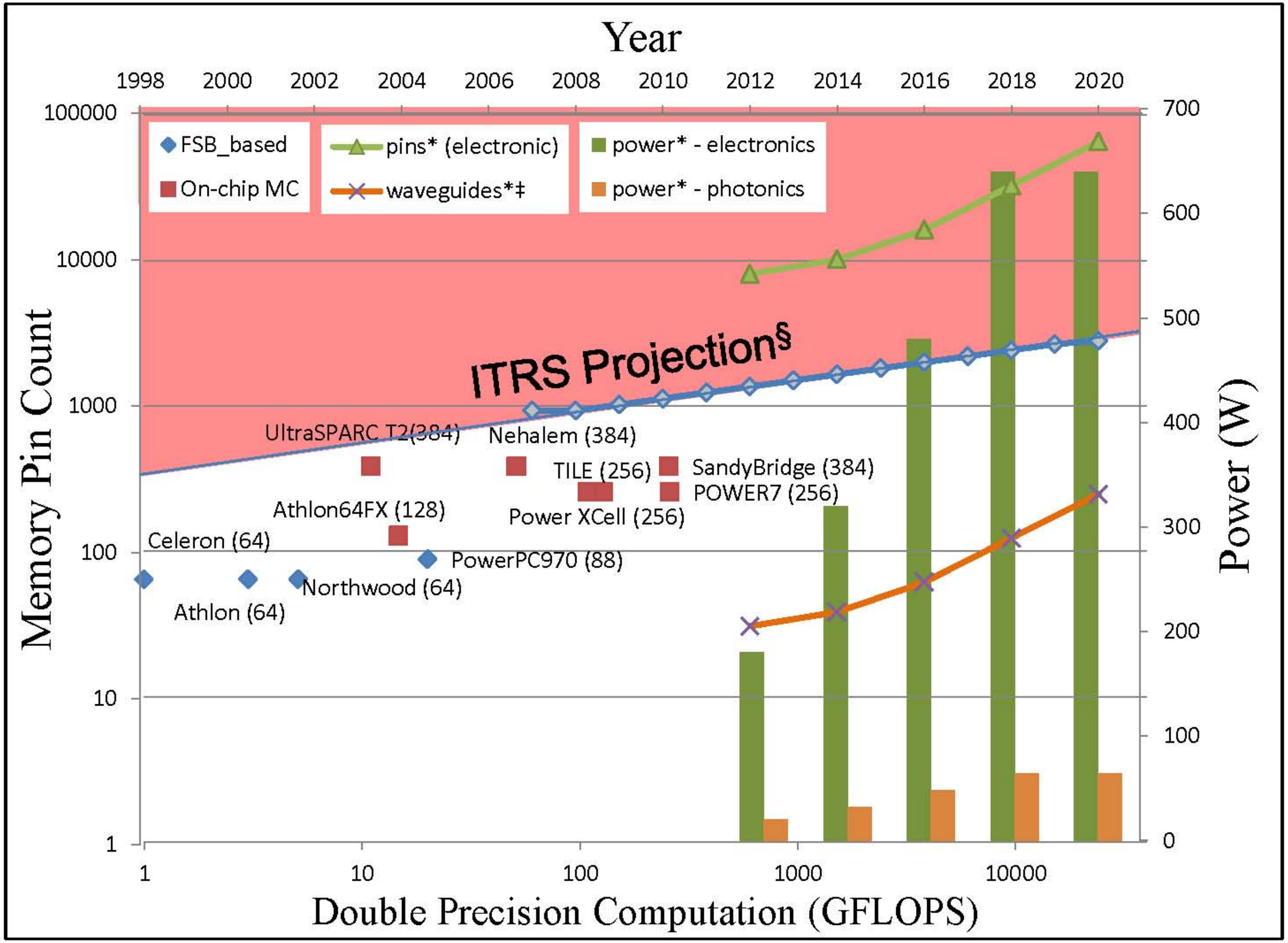
- Modulate/receive data stream once per communication event
- *Wavelength Parallelism :*
  - Broadband switch routes entire multi-wavelength stream
  - High I/O bandwidth density
- Distance Independence
  - Off-chip BW ≈ on-chip BW for nearly same power

ELECTRONICS

- Buffer, receive, and re-transmit at every repeater/router
- *Space Parallelism :*
  - Each bus lane routed independently ($P \propto N_{LANES}$)
  - Low I/O bandwidth density
- Off-chip BW requires much more power than on-chip BW

In the context of computing – Photonic communication can be fully exploited only by rethinking how to leverage its unique data-movement capabilities to realize new system architectures

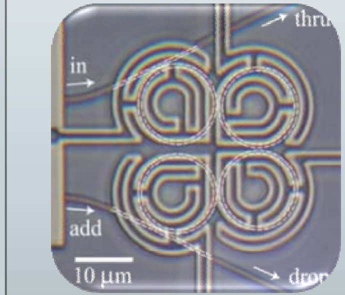
Photonic Interconnectivity delivers scalability – energy and bandwidth density



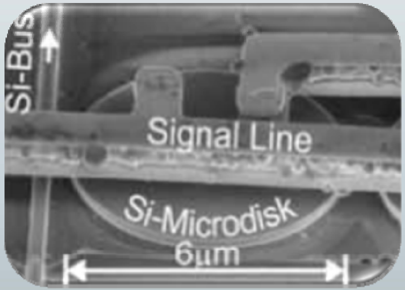
Silicon Photonics

Silicon-on-insulator (SOI) platform photonic building blocks:

High index contrast enables high confinement, low-loss propagation, virtually lossless bending

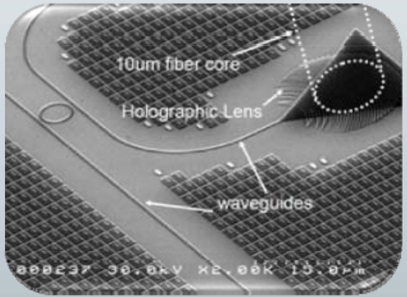


MIT

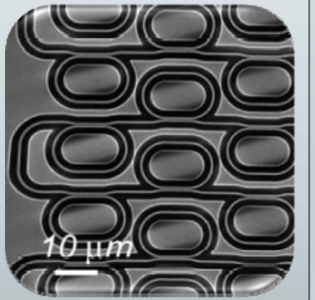


Sandia

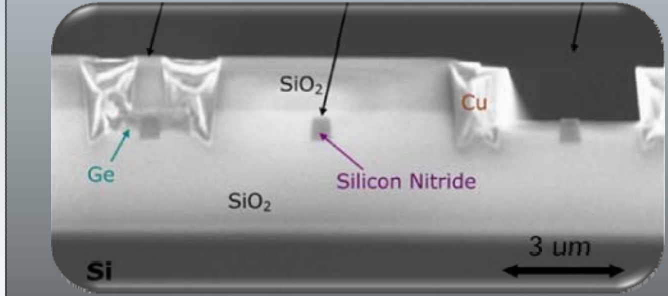
Ghent



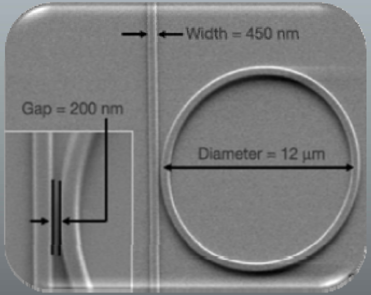
Luxtera



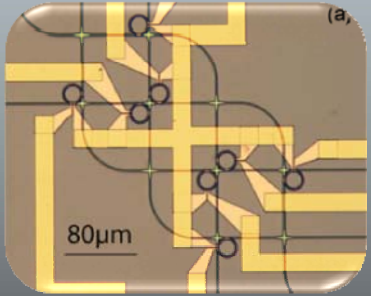
IBM



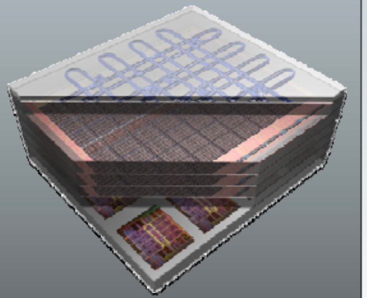
Intel



Cornell



Cornell/Columbia

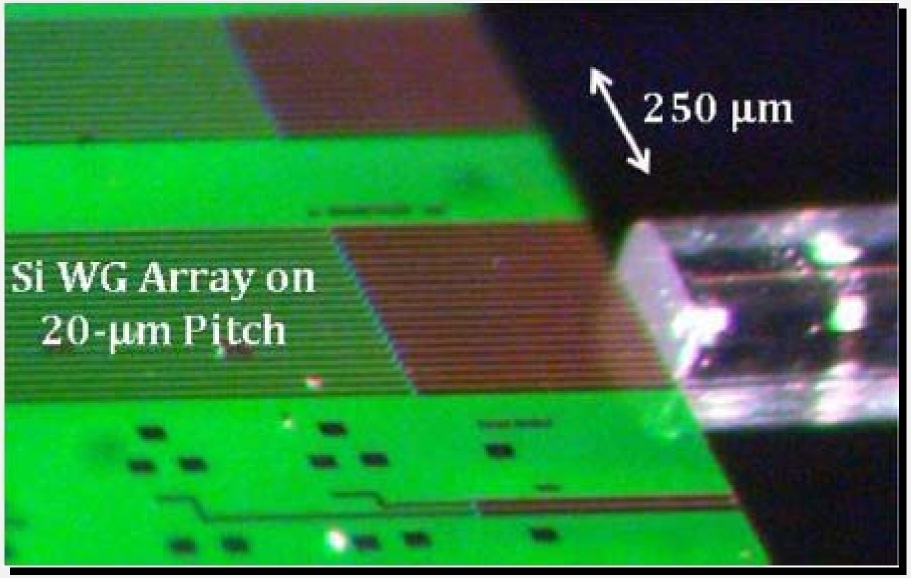


Columbia

Silicon Photonic Interconnects in Computing

- Silicon photonics:
- Off-chip BW = On-chip BW for nearly same **power**.
- Dense WDM = extreme **bandwidth density** - I/O bottlenecks
- Broadband switch routes entire multi-wavelength stream.

- Bandwidth density: ~ 2 Tbps/20 µm pitch at chip's edge.



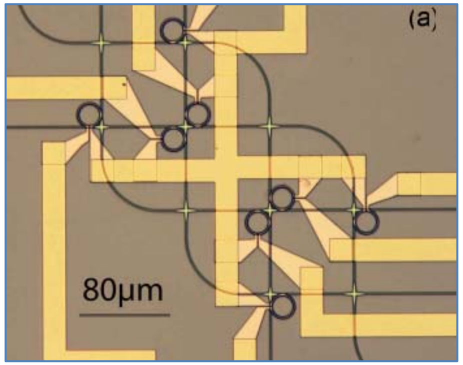
[F. Doany *et al.*, *JLT 29 (4) (2011)*]

Silicon Photonics – Optical Interconnection Networks

- **Silicon as core material**
  - High refractive index and high contrast – sub micron cross-section dimensions, smallest bend radius.
- **Small footprint devices**
  - 10 µm – 1 mm scale compared to cm-level scale for telecom components
- **Low power consumption**
  - Can reach <1 pJ/bit per full point to point link
- **Aggressive WDM platform**
  - Bandwidth densities 1-2Tb/s per pin
- **Silicon wafer-level CMOS processing**
  - Integration
  - Mass production, price
  - Compatibility with CMOS fabs, CMOS electronics

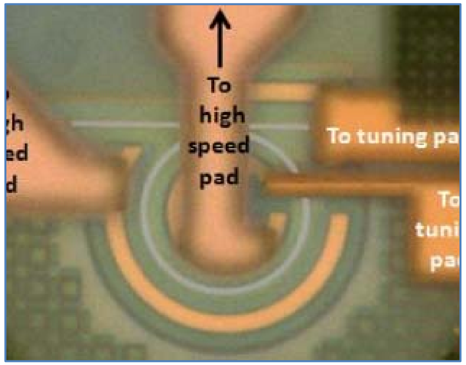
Silicon Microring/Microdisk Based Devices

Switching



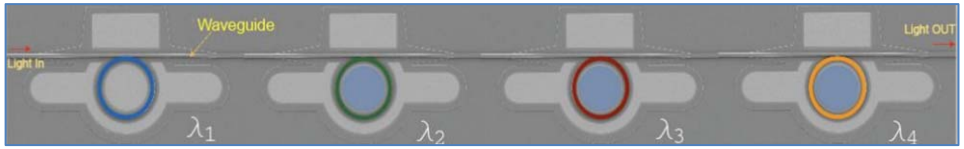
(Cornell/Columbia)

Modulation



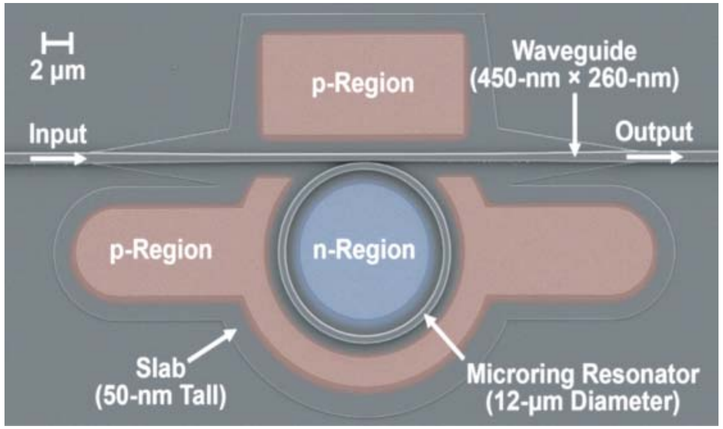
(Oracle)

WDM Modulation & Demultiplexing

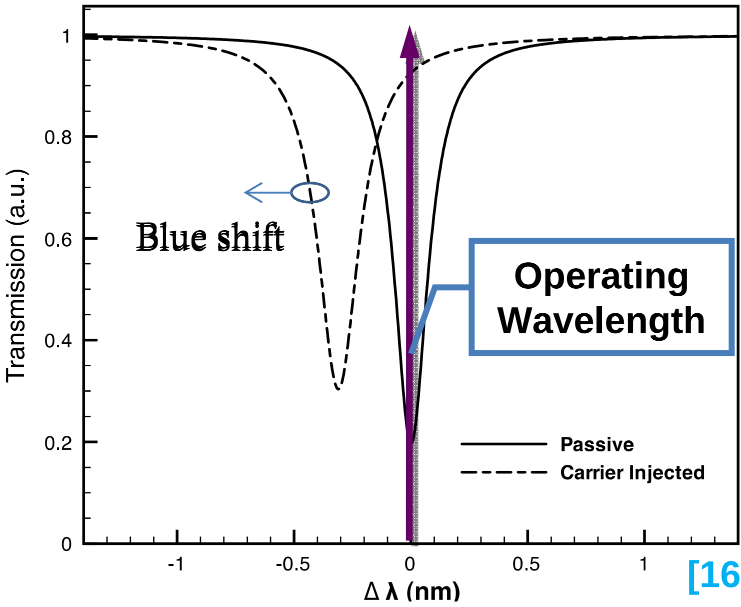


Active / Tunable Micoring Devices

- P/N-doping of silicon - diodes for carrier injection (p-i-n) or depletion (p-n).
- OOK modulator can be based on small resonance shifts.
- Power dissipation ∝ device volume → fJ/bit.
- Integrated local heaters allow thermal stabilization.
- Functionalities: modulators (up to 40 Gbps), WDM mux / demux, filters.



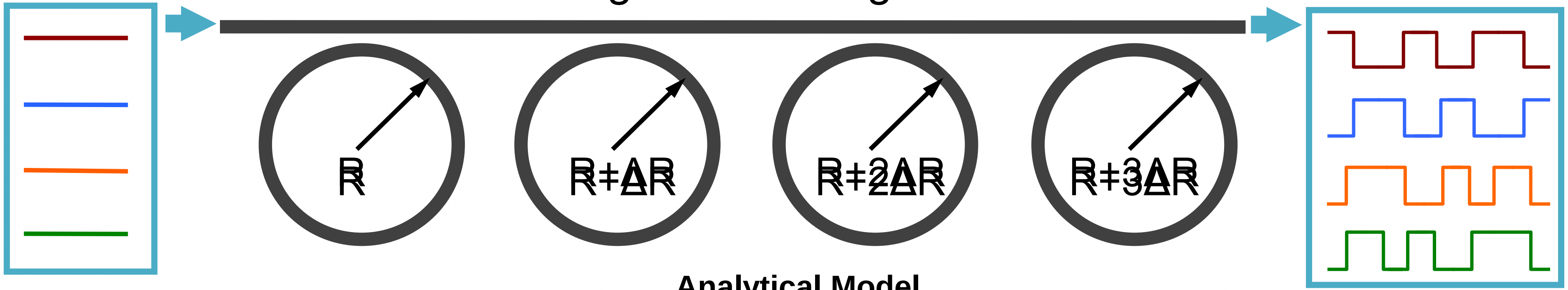
p-i-n

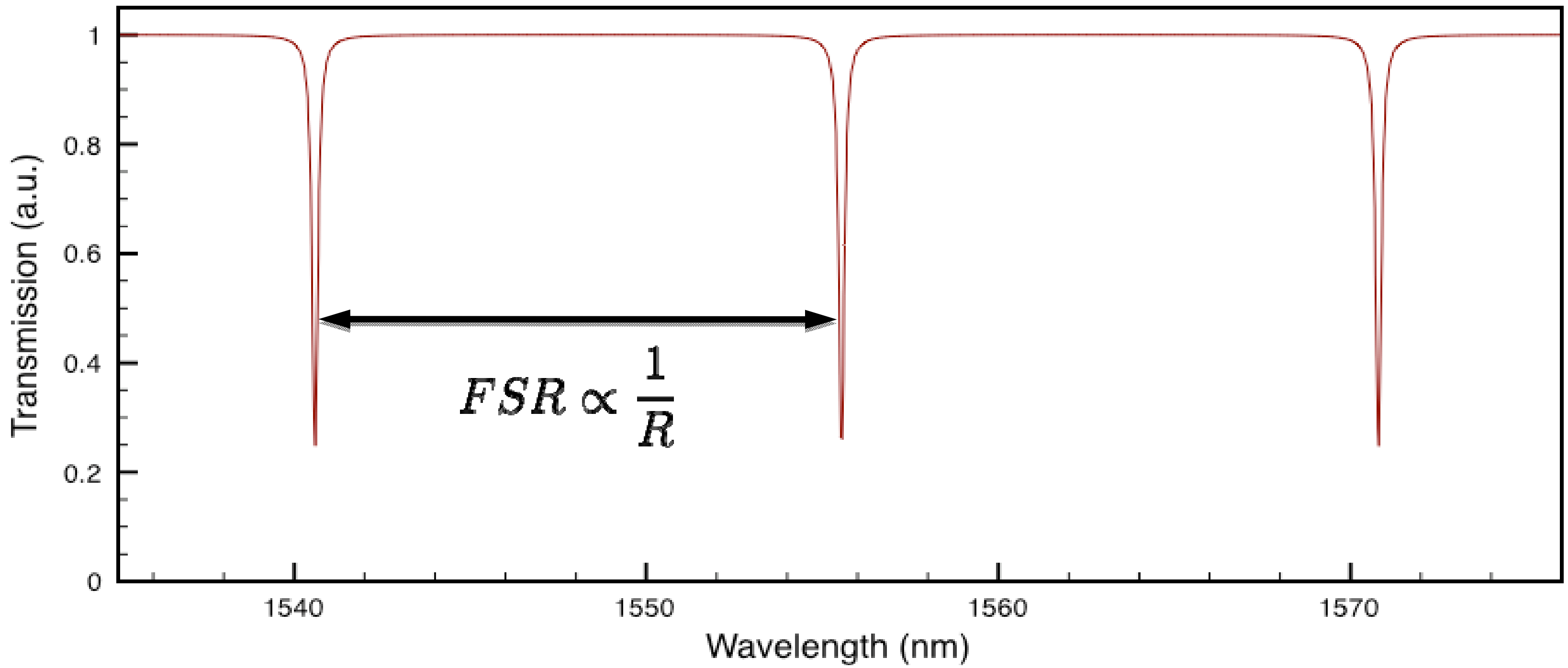


[16]

Microring-Based Comm. Links

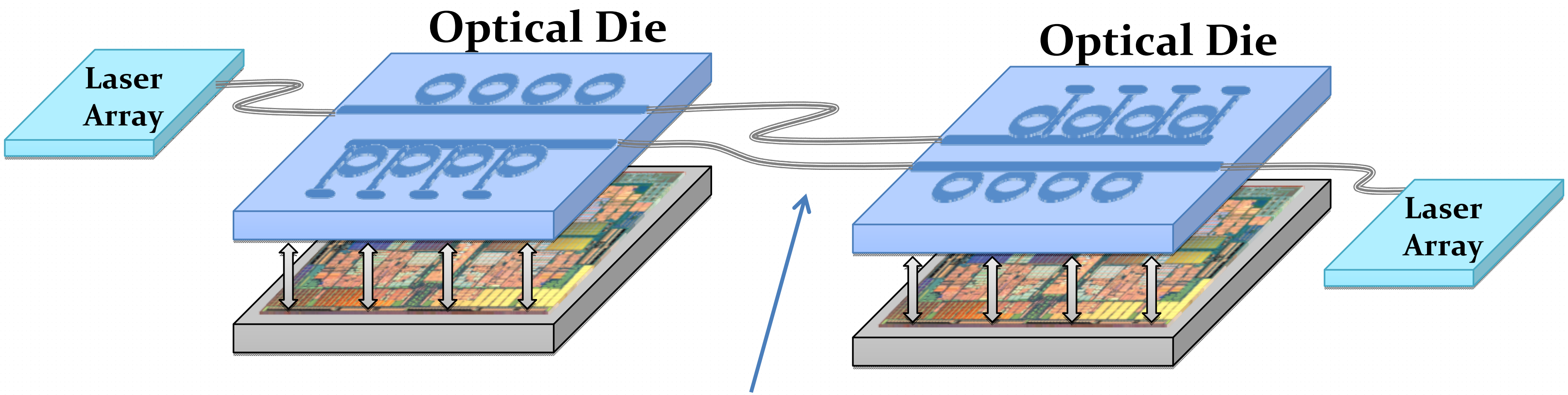
Wavelength selectivity inherently supports WDM configuration with a single bus waveguide.





Dense WDM Microring Link Design

Design driven by "best possible" single-waveguide optical link in terms of BW density and energy efficiency



Tx Array:

- Si or SiN bus WG
- Inverse-taper edge couplers
- Depletion-mode microring modulators

SM fiber:

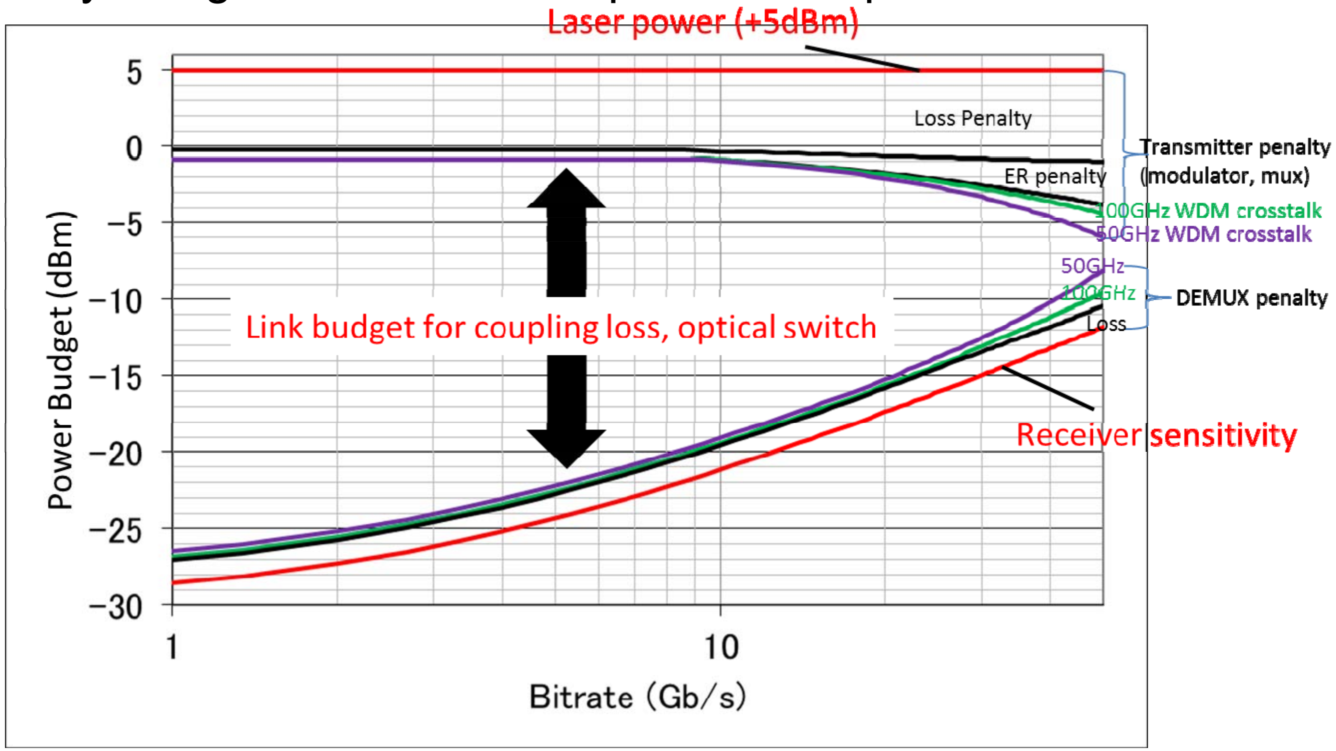
- PM
- Negligible loss up to 1 km

Rx Array:

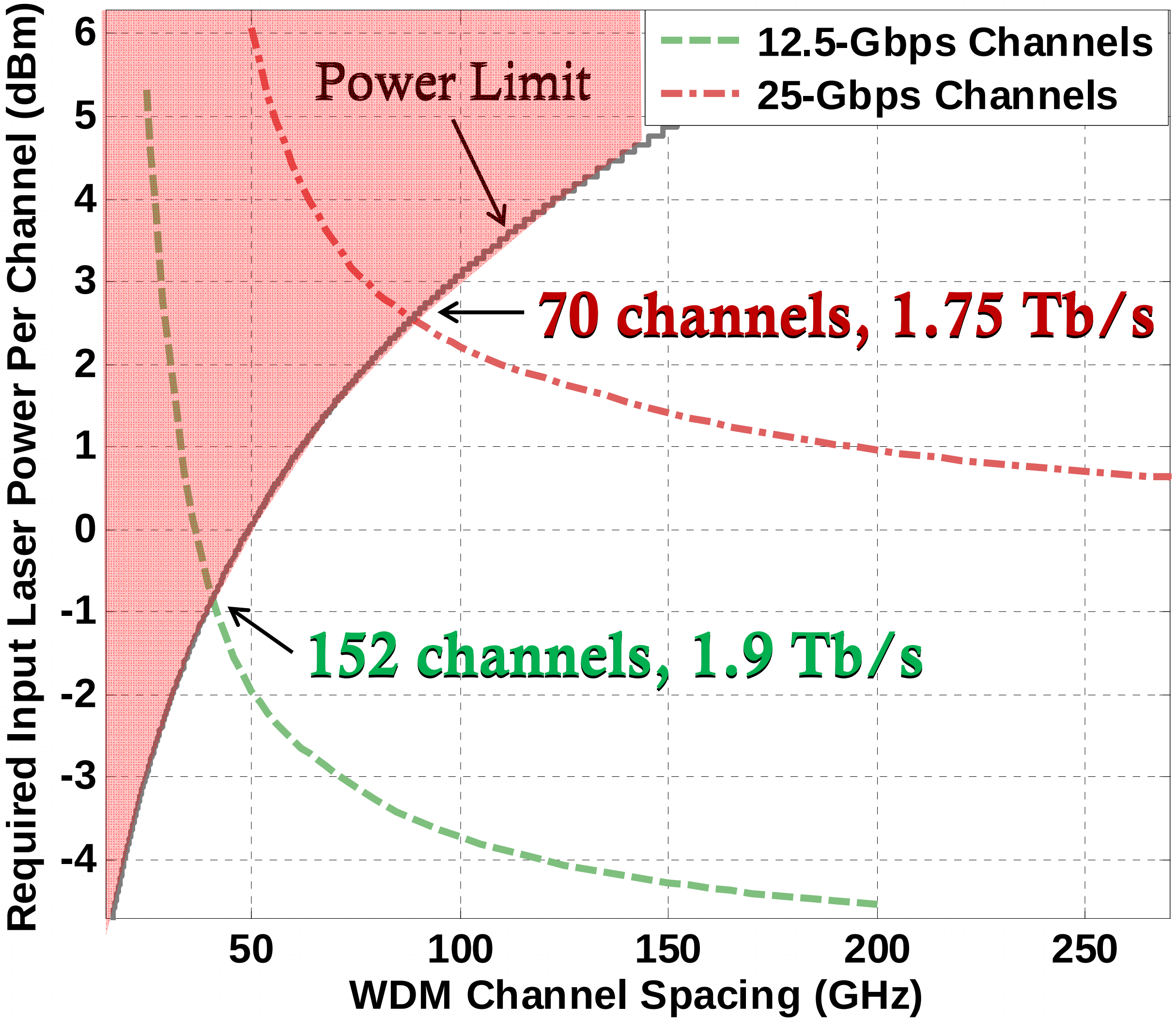
- Thermally tuned microring filters
- Ge PD on drop ports

Analysis of a Microring-Based Optical I/O Link

- Approach:
  - Account for all mechanisms involved in power penalty and loss
  - Analyze expected performance and scalability of microring-based SiPh links
  - Identify key parameters / devices that need further improvement.
  - Identify design trade-offs and optimal work points for the link.



Optical Power Budget



- Power per channel inversely proportional to channel spacing.
- 20-dBm power limit determines achievable BW.
- 12.5 Gb/s test case more scalable. Mainly because of receiver sensitivity.

Link Power Efficiency

Examine for a 1.55-Tb/s aggregate BW work point:

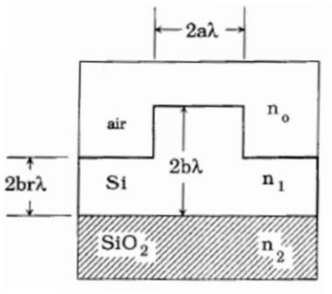
- 12.5 Gb/s rate, 50-GHz spacing, 124 channels.
- 25 Gb/s rate, 100-GHz spacing, 62 channels.

	12.5 Gb/s Modulation	25-Gb/s Modulation
Microring modulation	**0.01 pJ/bit**	**0.01 pJ/bit**
Modulation driver	**0.1 pJ/bit**	**0.3 pJ/bit**
Modulator thermal stabilization	**0.11 pJ/bit**	**0.06 pJ/bit**
Demux thermal stabilization	**0.11 pJ/bit**	**0.06 pJ/bit**
PD and receiver circuitry	**0.4 pJ/bit**	**1 pJ/bit**
Laser source (wall-plug efficiency)	**5.56 pJ/bit @1% efficiency** **0.56 pJ/bit @10% efficiency**	**7 pJ/bit @1% efficiency** **0.7 pJ/bit @10% efficiency**
Electronic data transmission to and from optical module	**1 pJ/bit**	**2 pJ/bit**
Overall with 10% laser wall-plug efficiency	**2.3 pJ/b**	**4.1 pJ/b**

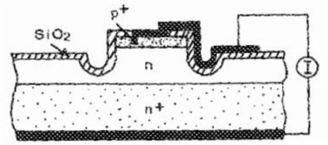
Path to Commercialization: Silicon Photonic Technology

Fundamental Discoveries

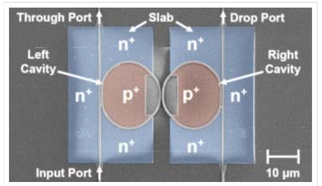
- low-loss, single-mode waveguiding



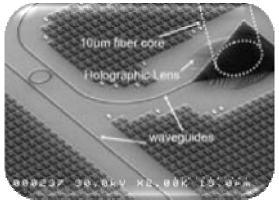
- optical coupling
- optical modulation via carrier injection



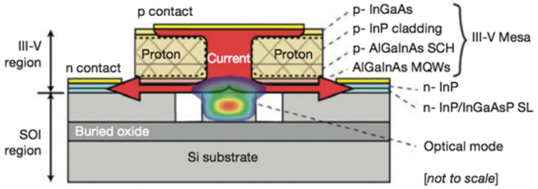
Introduction of Innovative Devices and Processes



- high-speed microring modulators and switches
- arrayed waveguide gratings

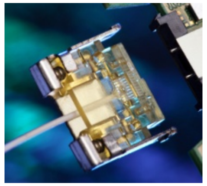


- Germanium photodetectors
- ultra low-loss waveguides and crossings
- hybrid silicon lasers



Integration and Commercialization

Hybrid platforms



intel

Monolithic CMOS Integration

IBM

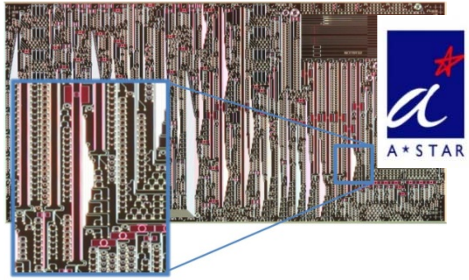
ORACLE

Transceivers for Datacom

APIC CORPORATION

LUXTERA NANOPHOTONIC INTEGRATED CIRCUITS

Foundry Services



Institute of Microelectronics A*STAR

imec

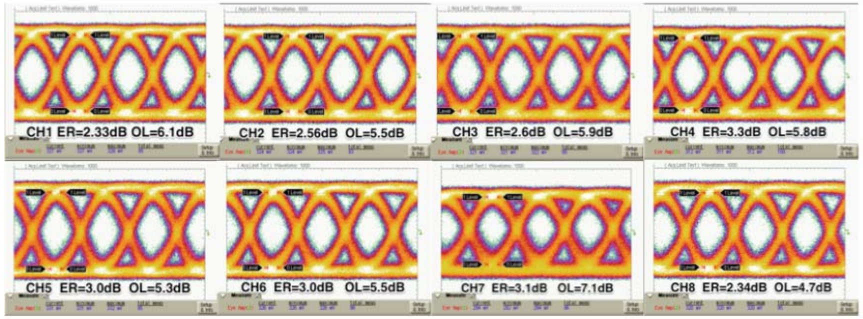
1990s 2000s 2010 2013 +

Columbia LRL Demonstrations

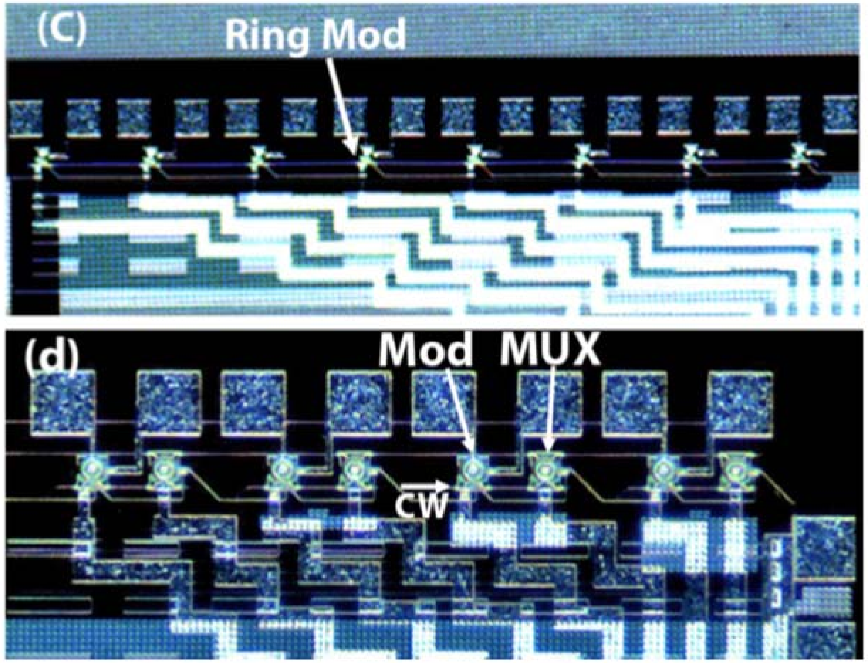
320 Gb/s WDM transmitters based on silicon microrings

- ❑8-channel WDM transmitter based on conventional common-bus architecture
- ❑32fJ/bit modulation power efficiency, less than 0.04 mm^2 chip area
- ❑highest aggregated data rate achieved in silicon transmitters

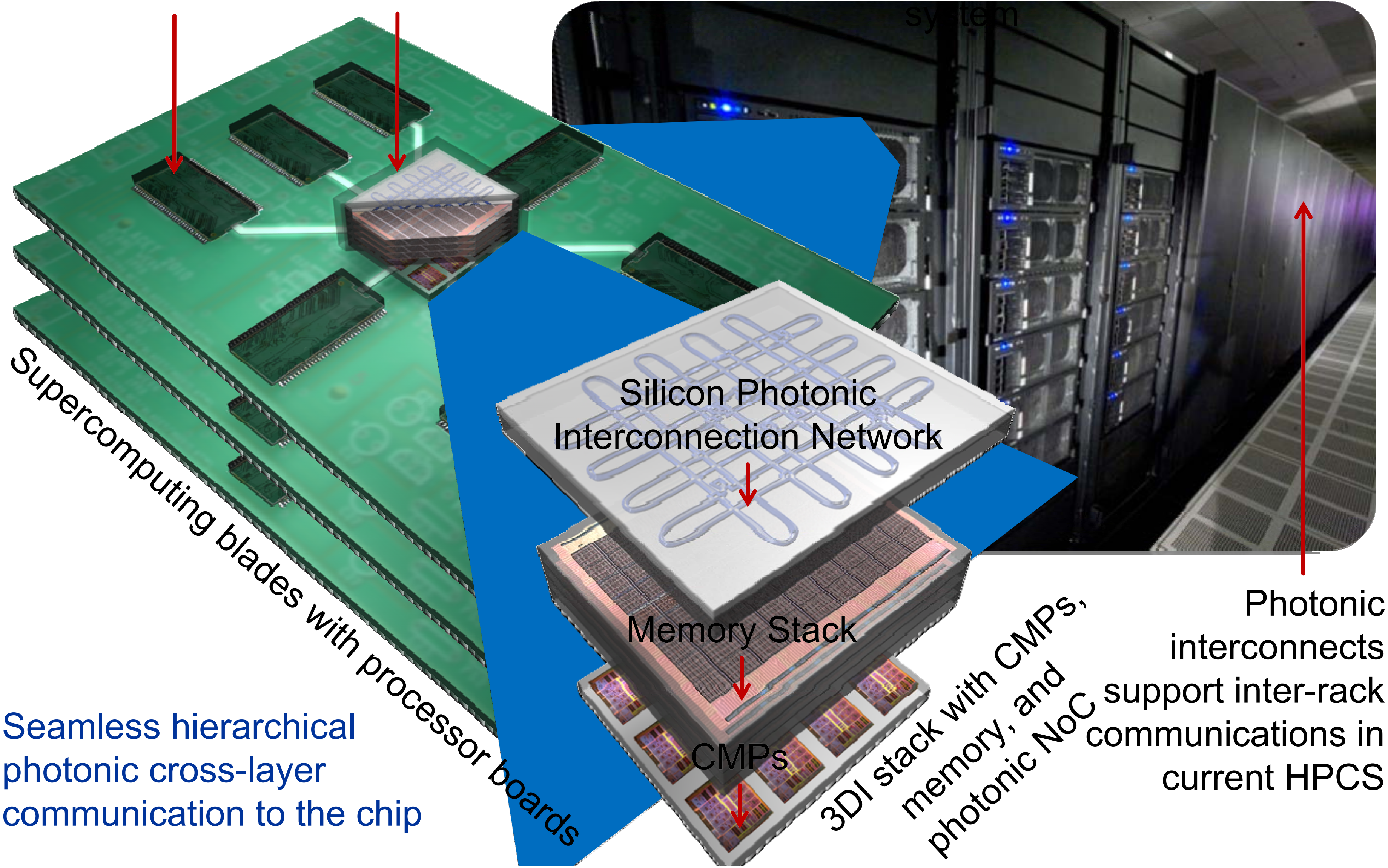
8-ring transmitter spectra



40Gb/s eye diagrams



Silicon Photonics for Exascale Computing



DRAM

CMPs 3DI Stack

Exaflop-scale high-performance computing system

Silicon Photonic Interconnection Network

Memory Stack

CMPs

Supercomputing blades with processor boards

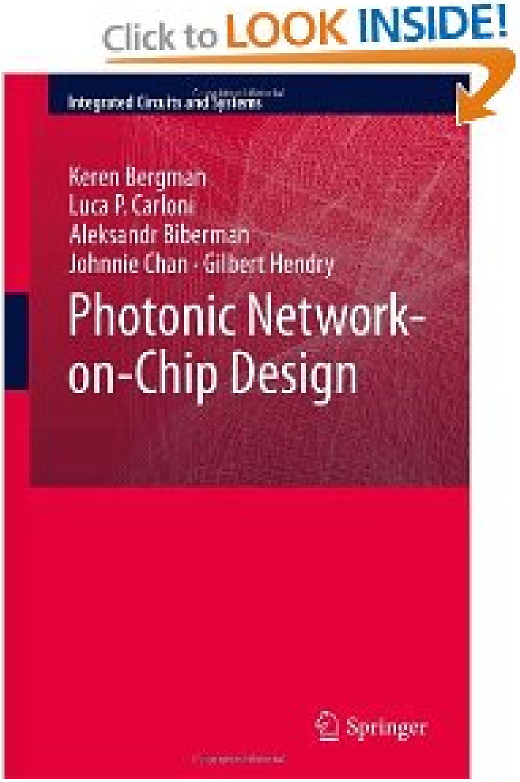
3DI stack with CMPs, memory, and photonic NoC

Photonic interconnects support inter-rack communications in current HPCS

Seamless hierarchical photonic cross-layer communication to the chip

Silicon Photonics based systems design

Photonic Network-on-Chip Design
Keren Bergman, Luca Carloni, Aleksandr Biberman,
Johnnie Chan, and Gilbert Hendry
Series: Integrated Circuits and Systems, Vol. 68,
Springer Science + Business Media New York 2014



Click to LOOK INSIDE!

Integrated Circuits and Systems

Keren Bergman
Luca P. Carloni
Aleksandr Biberman
Johnnie Chan · Gilbert Hendry

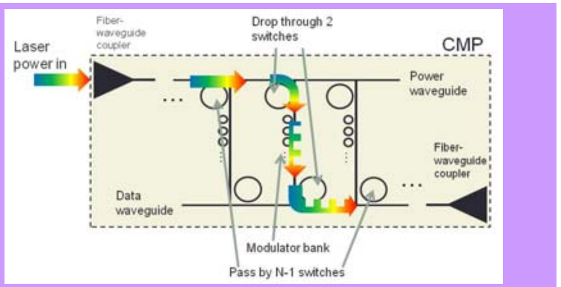
Photonic Network-on-Chip Design

Springer

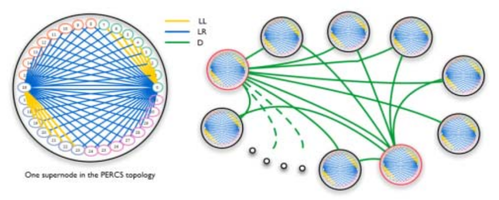
Photonic-Enabled Systems: Multi-Level Co-Design

PhoenixSim: Design, Modeling and Simulation Environment

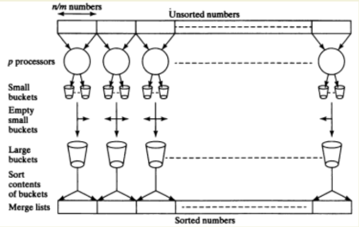
- **Physical link layer:**
  - SiP components modeling
  - Link bandwidth maximization
  - Optical power budget validation



- **Network layer**
  - Optical data flow, switching, routing protocols
  - Network performance analysis

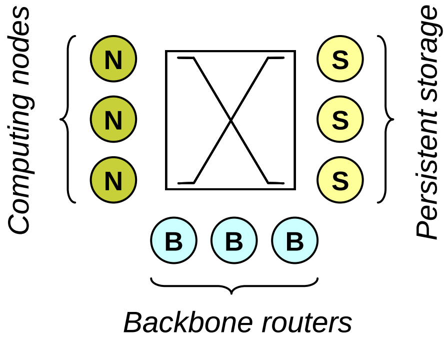


- **Application layer**
  - BW and data flow application mapping
  - Optically enabled algorithm re-design
  - Large scale application simulation



PhoenixSim Suite

From a general data-center description



Computing nodes

N N N

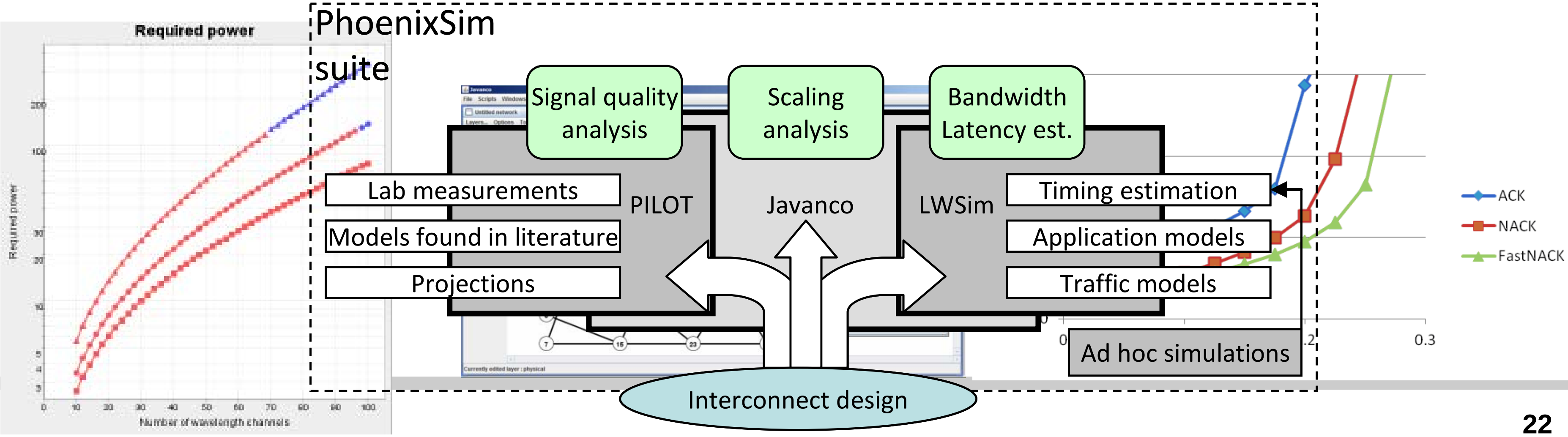
S S S

B B B

Persistent storage

Backbone routers

- Network structure level: Javanco
  - Topology construction and visualization, data-structures for other tools
- System-level physical layer modeling: PILOT
  - Study of individual component impact on the signal
  - Component parameter optimization for higher bandwidth
- Response to traffic and application demands: LWSim
  - Packets/connection dynamics, protocols, queues, contention



PhoenixSim suite

Signal quality analysis

Scaling analysis

Bandwidth Latency est.

Lab measurements

Models found in literature

Projections

PILOT

Javanco

LWSim

Timing estimation

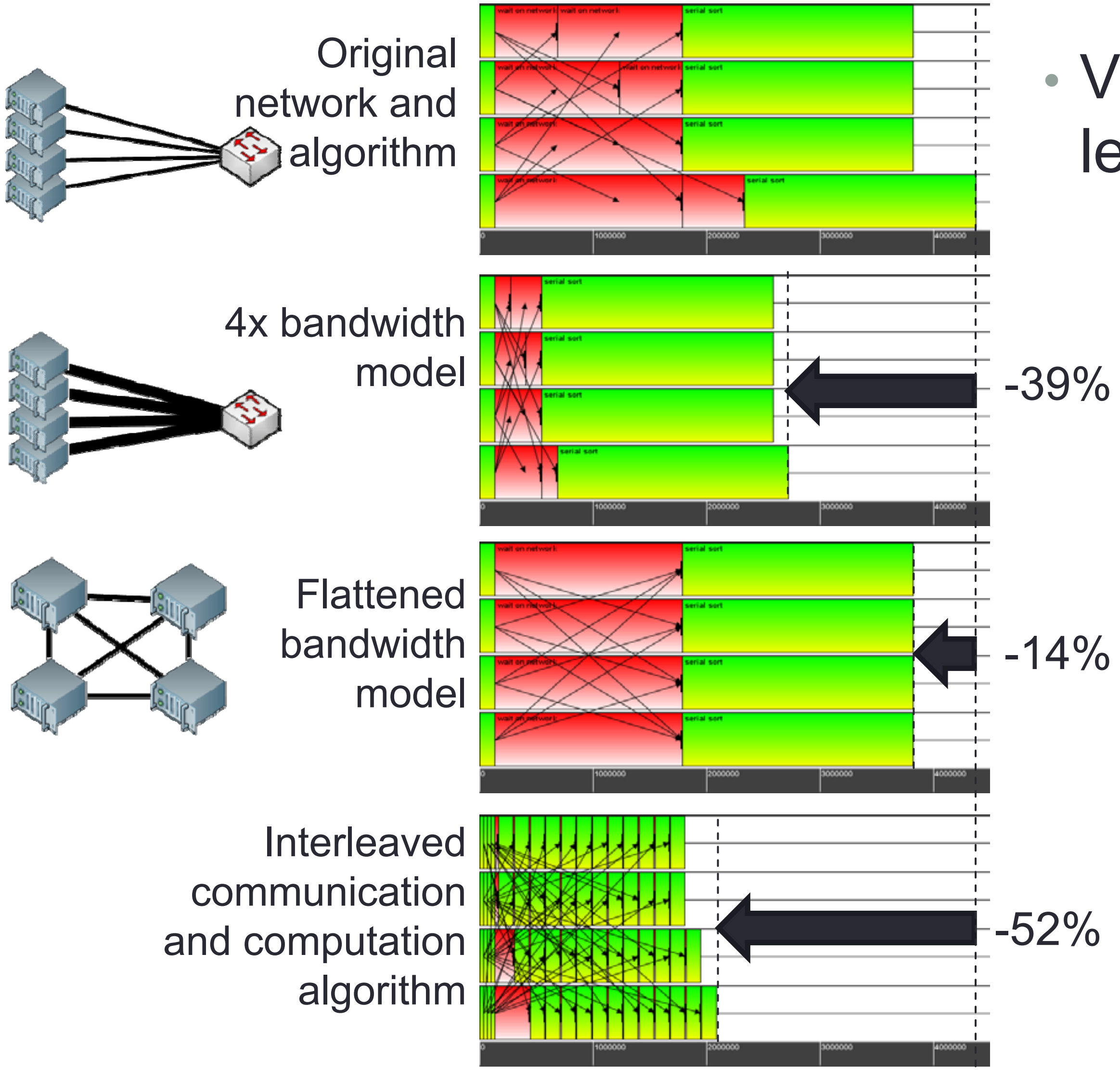
Application models

Traffic models

Ad hoc simulations

Interconnect design

Connecting with the Applications



Original network and algorithm

4x bandwidth model

-39%

Flattened bandwidth model

-14%

Interleaved communication and computation algorithm

-52%

- Various opportunities to leverage optics:
  - More bandwidth

    With more wavelengths and higher rates
  - More concurrency

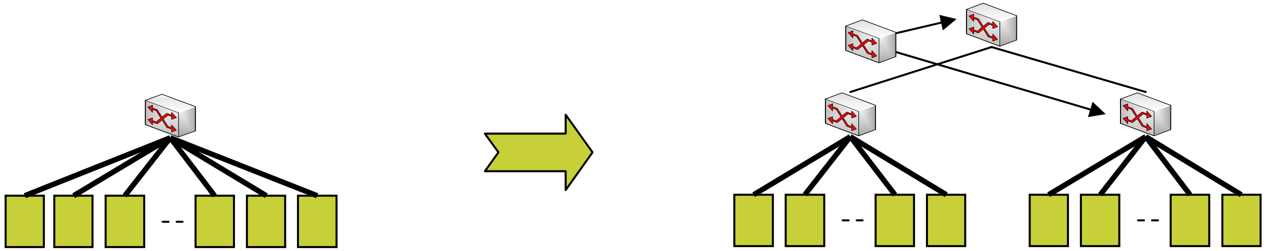
    Thanks to distance-independent links
  - More agility

    With all-optical processing

    Interleaving

Co-design – an example

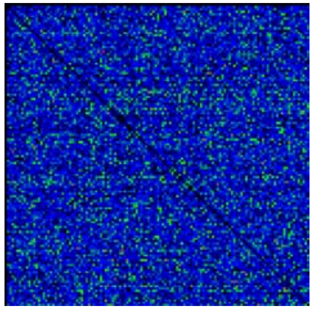
1. Replace central switch architecture by distributed network

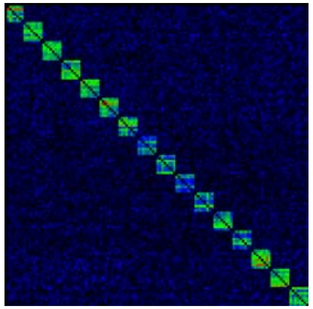


2. Make the application topology aware

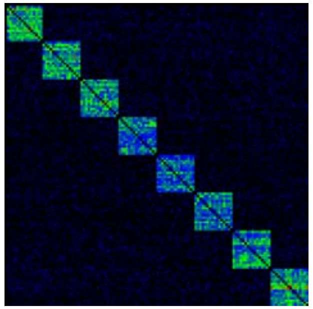
Measured traffic (by simulation)
for unaware application

Measured traffic (by simulation)
for a topology aware application

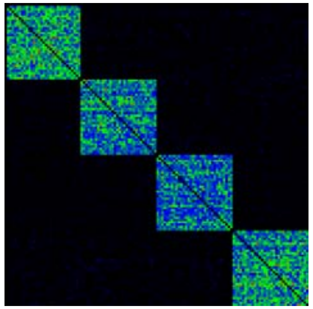




8 nodes per cluster



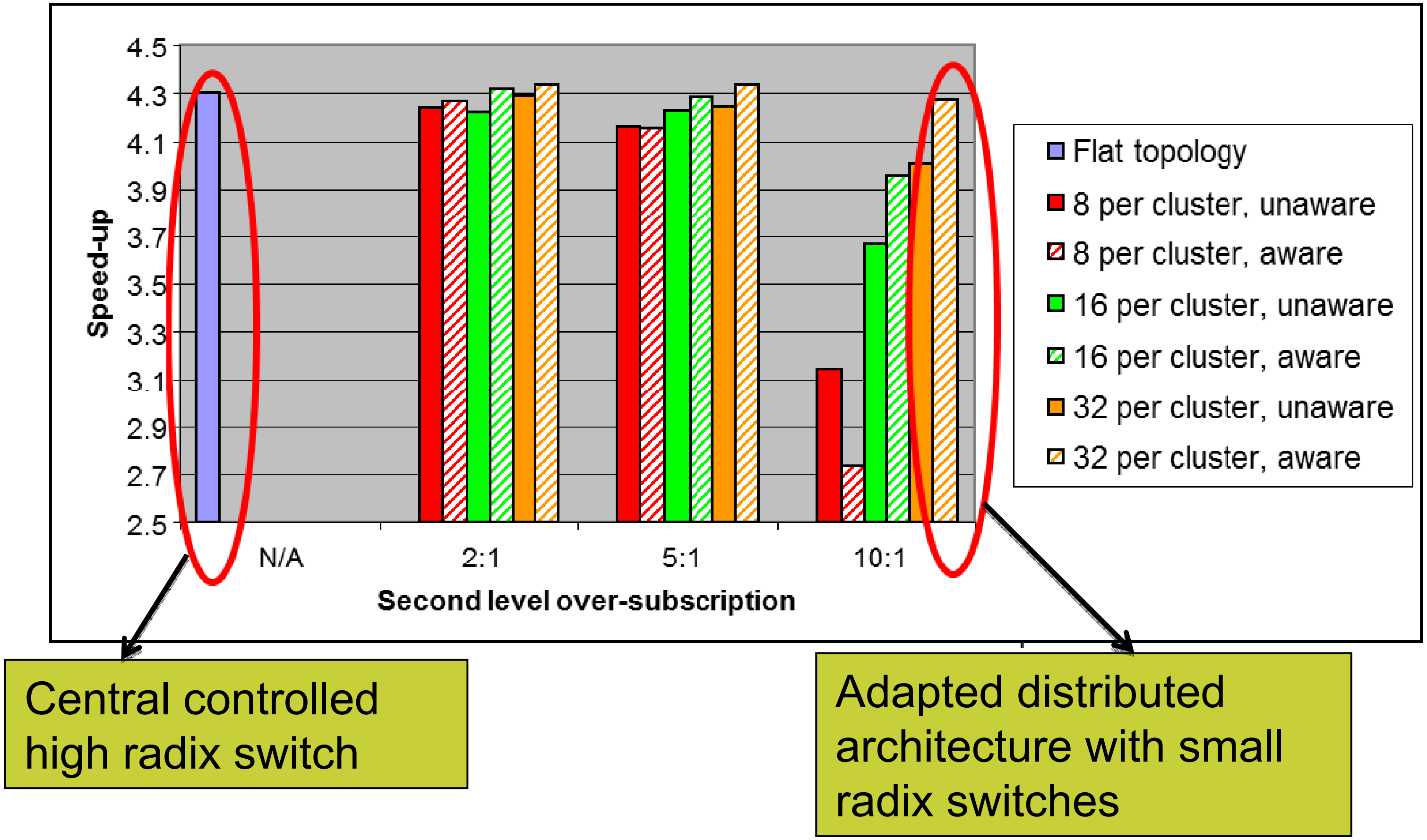
16 nodes per cluster



32 nodes per cluster

Co-design – initial simulation results

- nearly the same speed-up is achieved as with ideal central switch
  - Using smaller radixes and with a non full bisectional bandwidth



Central controlled high radix switch

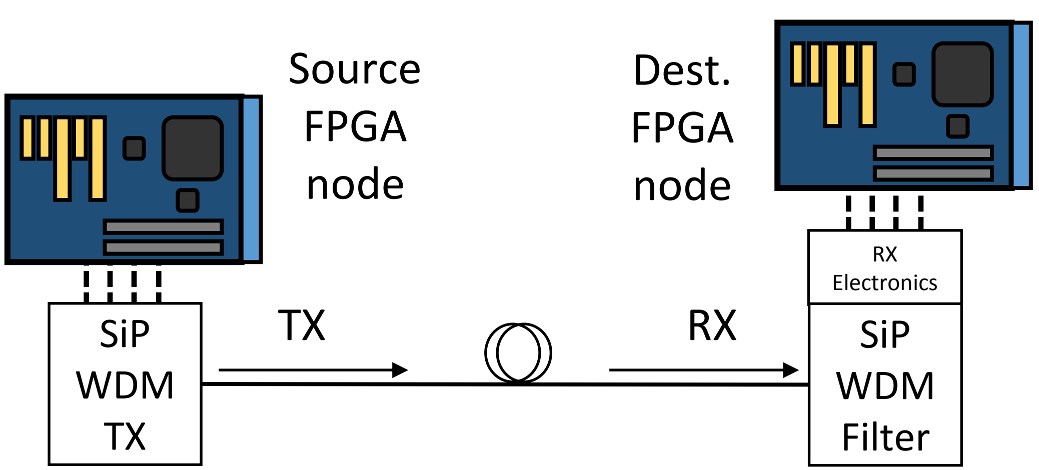
Adapted distributed architecture with small radix switches

Putting it all together…

FPGA Programmable SiP Interconnected Networking Platform

Interconnected System
Optical Network Interface: O-NIC Link Negotiation and Control

Subsystem 1: Link Negotiation



SiP Link Initialization & Maintenance

-Insertion of SiP TX and RX

-Compatible with packet and circuit-based communications

SiP Link O-NIC Operation

-Measurable PHY negotiation characteristics

- Clock and data locking (no distributed clock)
- Data synchronization
- Data delivery statistics (link up-time / packet loss)

-Programmable node emulation in firmware

- CPU, memory, hardware accelerators
- Measure performance with SiP connectivity

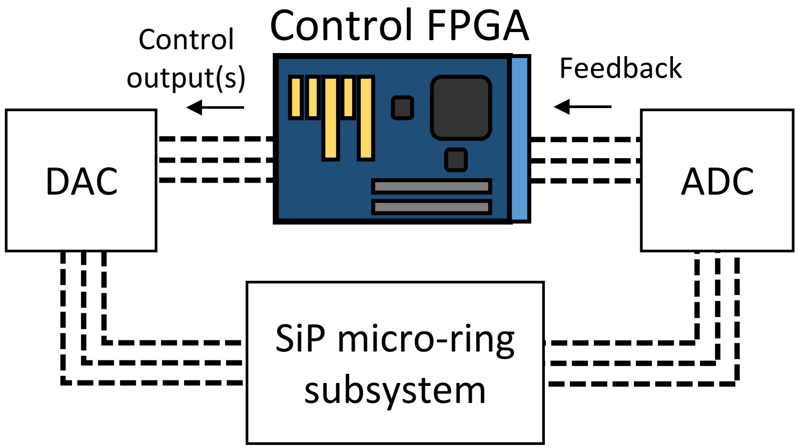
Micron

ALTERA
MEASURABLE ADVANTAGE™

FPGA-Controlled Silicon Photonic Interconnected System

Subsystem Thermal Control and Operation

Subsystem 2: SiP Device Control



Initialization and Stabilization of SiP

-Electrical feedback:

- generated by microring subsystem

-In-waveguide power monitoring PDs

-Applied dithering signal

-Error signal generation for locking

Stabilized Operation of SiP Microring Subsystems

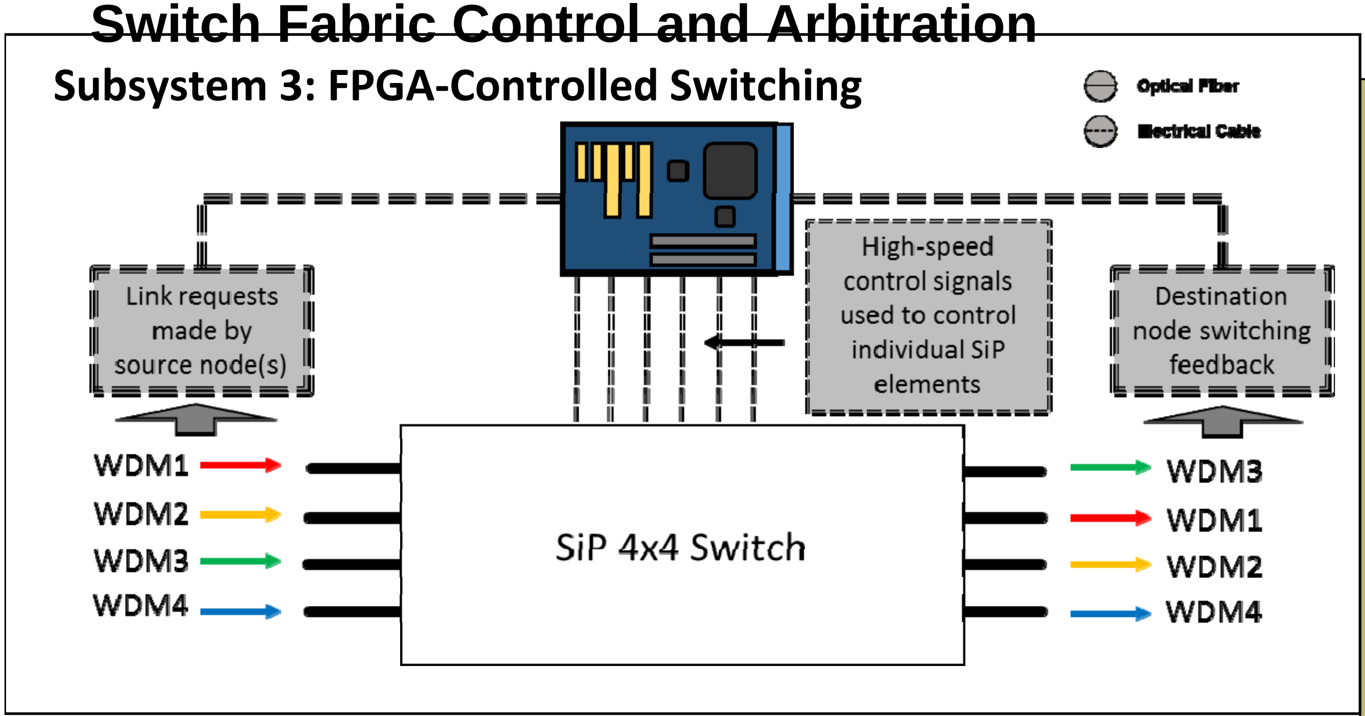
-Analysis and maintenance using state-based FPGA logic

-Analog-to-digital and digital-to-analog conversion is critical

- High-speed sampling compatible with nanosecond rise times

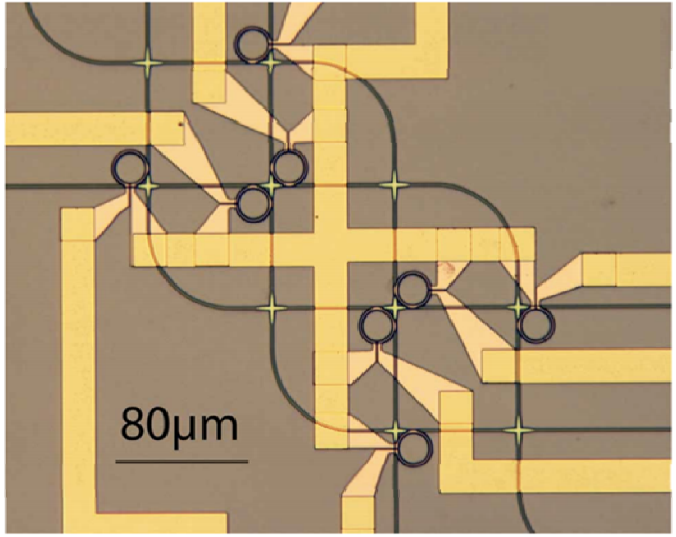
FPGA-Controlled Silicon Photonic Interconnected System

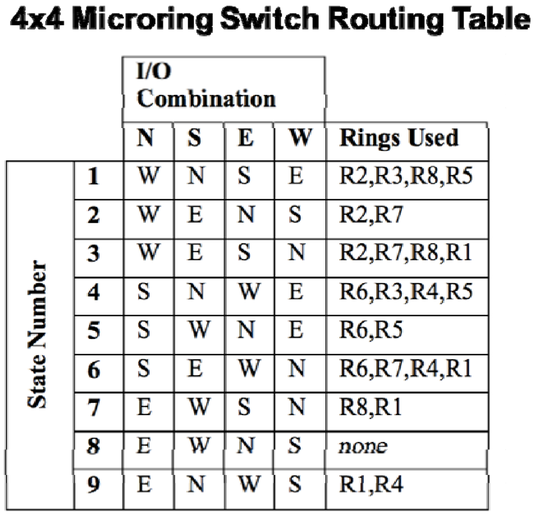
Switch Fabric Control and Arbitration



Subsystem 3: FPGA-Controlled Switching

Optical Fiber
Electrical Cable
Link requests made by source node(s)
High-speed control signals used to control individual SiP elements
Destination node switching feedback
WDM1
WDM2
WDM3
WDM4
SiP 4x4 Switch
WDM3
WDM1
WDM2
WDM4





4x4 Microring Switch Routing Table

		I/O Combination				
		N	S	E	W	Rings Used
State Number	1	W	N	S	E	R2,R3,R8,R5
	2	W	E	N	S	R2,R7
	3	W	E	S	N	R2,R7,R8,R1
	4	S	N	W	E	R6,R3,R4,R5
	5	S	W	N	E	R6,R5
	6	S	E	W	N	R6,R7,R4,R1
	7	E	W	S	N	R8,R1
	8	E	W	N	S	*none*
	9	E	N	W	S	R1,R4

[N. Sherwood-Droz *et. al.*, *Optics Express*, 2008]

WDM Switching Fabric
High-Speed FPGA Arbitration

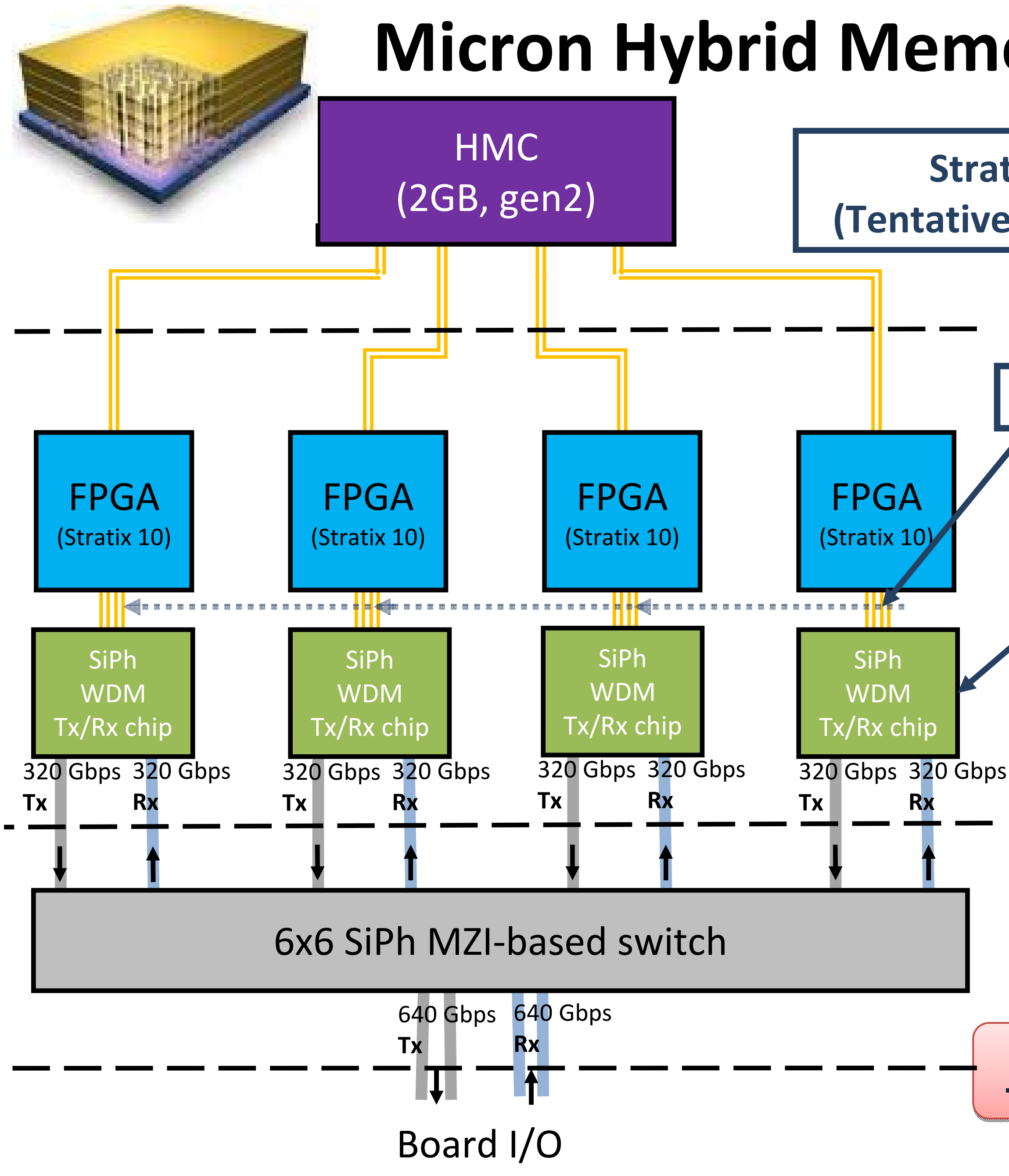
High-speed electrical control signals via FPGA
- State-based control of SiP
- Circuit requests and ACKs performed out-of-band

Broadband and wavelength-selective switching

Programmable switch arbitration protocols in firmware

Measurable switch arbitration impact on system performance (latency characteristics)

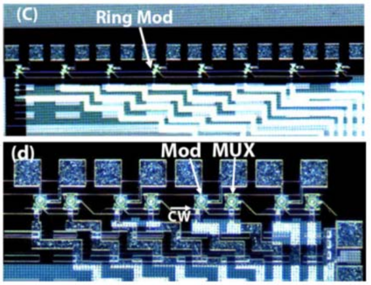
Silicon Photonic Interconnected Micron Hybrid Memory Cube



Stratix 10 FPGAs –
(Tentative release date 2015)

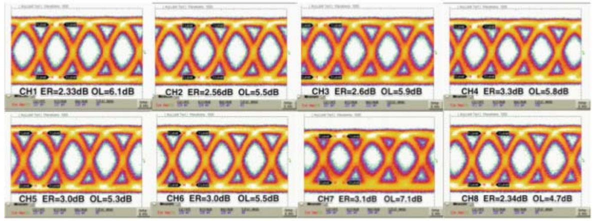
1.28 Tbps bisectional bandwidth

8 bidir. lanes @ 40 Gbps per FPGA



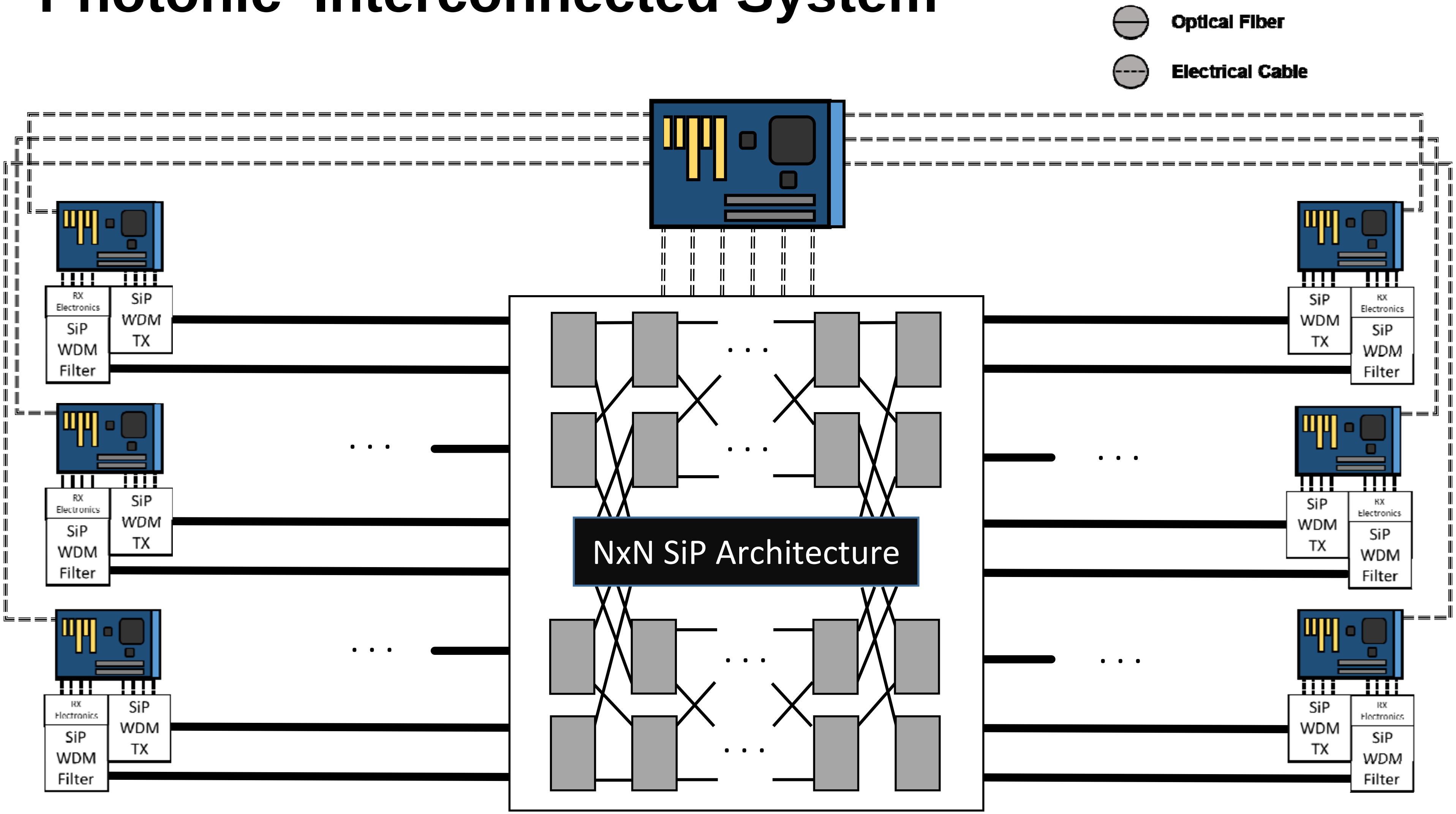
8 WDM CH SiPh Chip (OPSIS)

8 X 40Gb/s eye diagrams



1.28 Tbps bisectional bandwidth

Scalability of an FPGA-Controlled Silicon Photonic Interconnected System



scaled up to multi-node (4 nodes currently, 8 nodes), bi-directional FPGA-programmable SiP Interconnection Network Platform

Silicon Photonic for Exascale: Paths Forward

- Data movement rather than computation is the key challenge
- Silicon photonic technologies – commercial ecosystem
  - Links + switching required for full optical interconnection networks
- Energy – Si photonics can get to 1 Tb/s per pin at 1 pJ/bit system wide
- *Photonic switching* is central technology to realizing optical interconnection network that is beyond ‘wire replacement’
  - Uniquely optical - routing extreme bandwidth with minimal energy
- Optical network architectures are fundamentally different, circuit switched, plus optical functions
- *Holistic co-design:* of software-architecture-interconnect to realize performance and energy efficiency
- Create new truly photonic-enabled architectures

150 COLUMBIA | ENGINEERING

The Fu Foundation School of Engineering and Applied Science

1864–2014