

Silicon Photonics for Extreme Scale Computing — Challenges & Opportunities

Keren Bergman

Department of Electrical Engineering

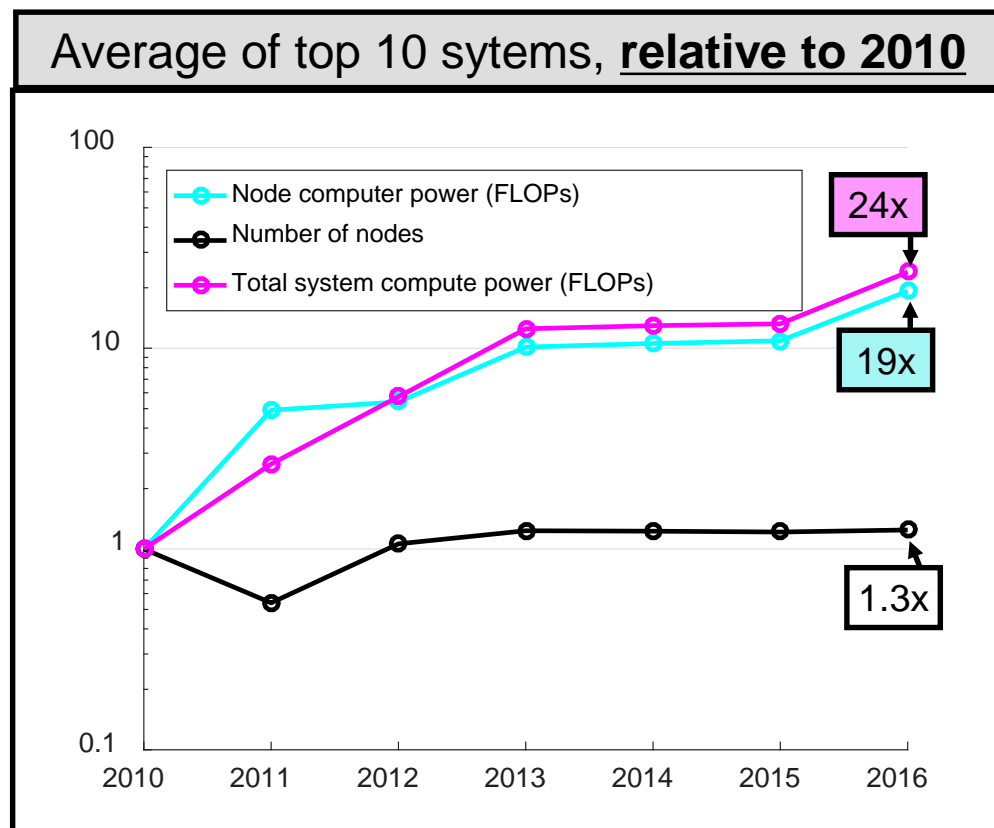
Lightwave Research Lab, Columbia University

New York, NY, USA



Trends in extreme HPC

- Evolution of the top10 in the last six years:
 - Average total compute power:
 - 0.86 PFlops → 21 PFlops
 - ~24x increase
 - Average nodal compute power:
 - 31GFlops → 600GFlops
 - ~19x increase
 - Average number of nodes
 - 28k → 35k
 - ~1.3x increase

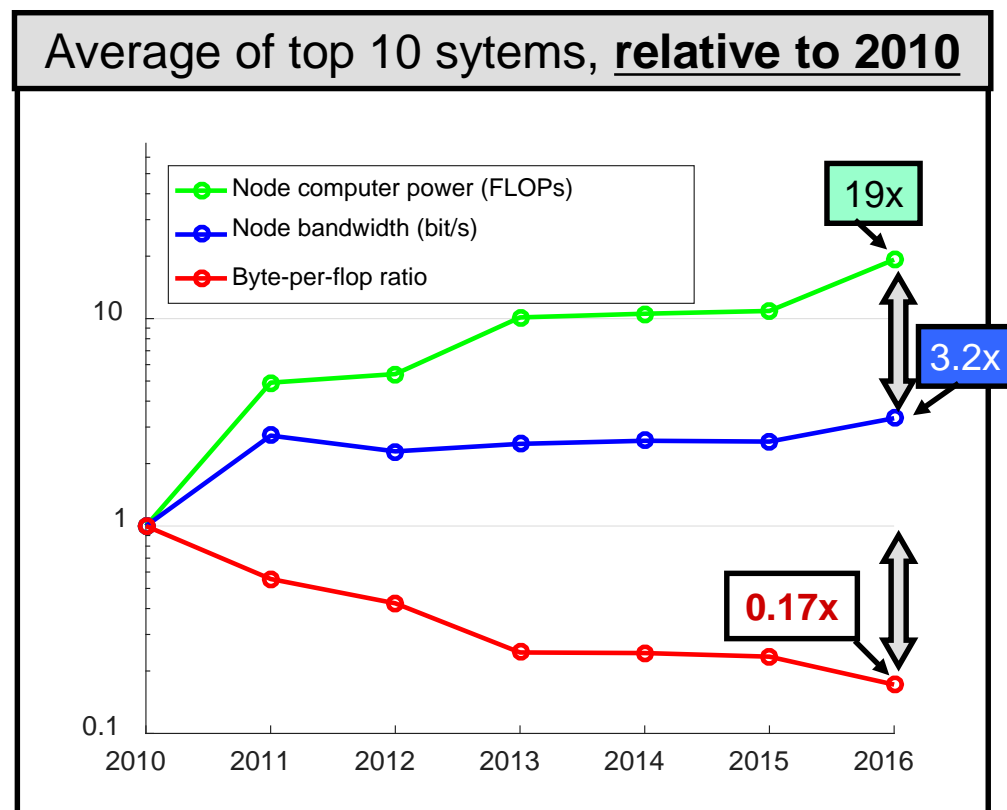


[top500.org, S. Rumley, et al. Optical Interconnects for Extreme Scale Computing Systems, Elsevier PARCO 64, 2017]

→ Node compute power main contributor to performance growth

Interconnect trends

- Top 10 average node level evolutions:
 - Average node compute power:
 - 31GFlops → 600GFlops
 - ~19x increase
 - Average bandwidth available per node
 - 2.7GB/s → 7.8GB/s
 - ~3.2x increase
 - Average byte-per-flop ratio
 - 0.06 B/Flop → 0.01 B/Flop
 - ~6x **decrease**
 - **Sunway TaihuLight (#1) shows 0.004 B/Flop !!**



[top500.org, S. Rumley, et al. Optical Interconnects for Extreme Scale Computing Systems, Elsevier PARCO 64, 2017]

→ Growing gap in interconnect bandwidth



Exascale interconnects – power and cost constraints

- **Real Exascale goal: reaching an Exaflop...**
 - ...while satisfying constraints (20MW, \$200M)
 - ...with reasonably useful applications

$$\begin{aligned}
 & 1.25 \text{ ExaFLOP} \\
 \times & 0.01 \text{ B/FLOP} \\
 & = 125 \text{ Pb/s injection BW} \\
 \times & 4 \text{ hops} \\
 & = 500 \text{ Pb/s installed BW}
 \end{aligned}$$

- Assume 15% of \$ budget for interconnect:

$$- 15\% \times \$200\text{M} / 500 \text{ Pb/s} = 6 \text{ ¢/Gb/s}$$

– Bi-directional links must thus be sold for ~10 ¢/Gb/s

- Today:

optical	10\$/Gb/s
electrical	0.1-1 \$/Gb/s

- Assume 15% of power budget for interconnect:

$$\begin{aligned}
 - 15\% \times 20\text{MW} / 125 \text{ Pb/s} &= 24 \text{ mW/Gb/s} = 24 \text{ pJ/bit} \\
 &= \text{budget for communicating a bit end-to-end}
 \end{aligned}$$

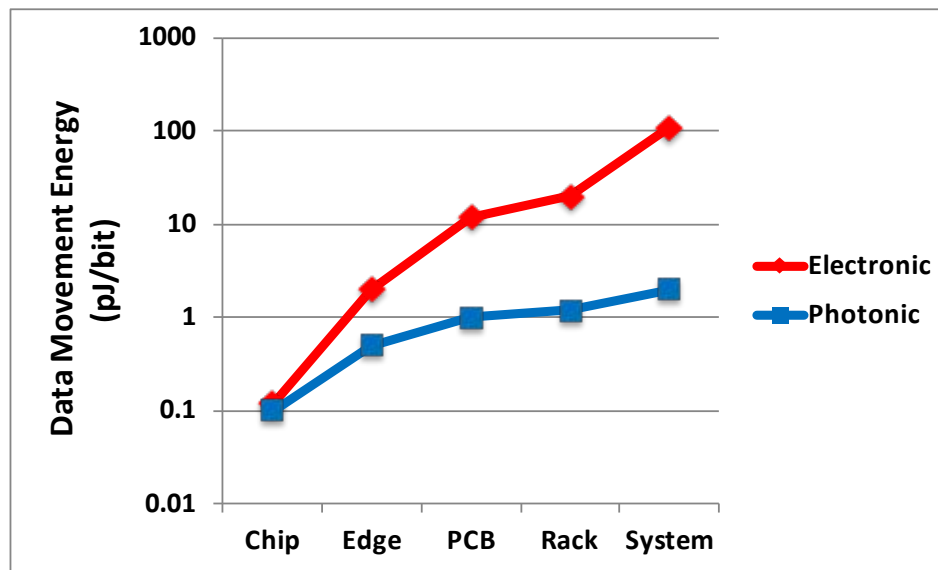
→ 6 pJ/bit per hop

→ 4 pJ/bit for switching today ~20 pJ/bit

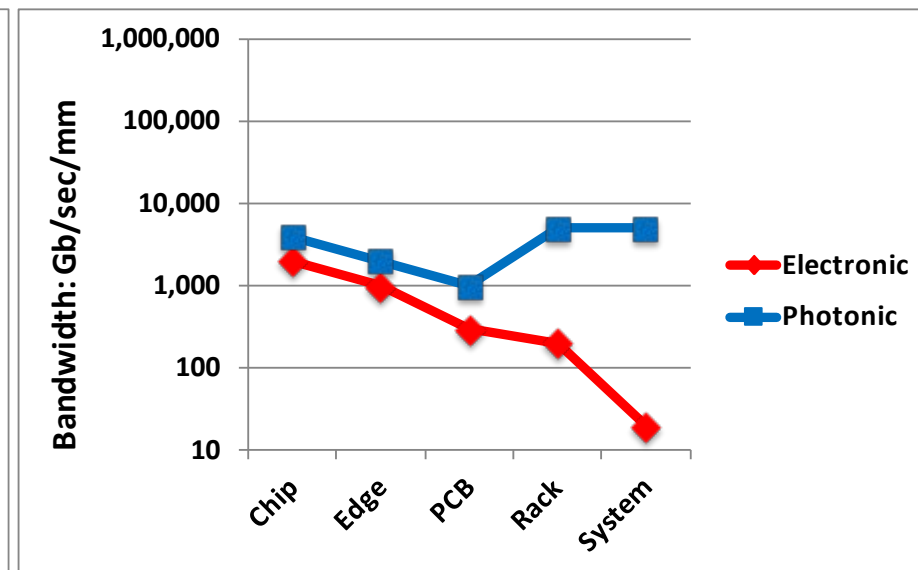
→ 2 pJ/bit for transmission today ~10 pJ/bit (elec) ~20 pJ/bit (optical)

The Photonic Opportunity for Data Movement

- ❑ Energy efficient, low-latency, high-bandwidth *data interconnectivity* is the core challenge to continued scalability across computing platforms
- ❑ Energy consumption completely dominated by costs of data movement
- ❑ Bandwidth taper from chip to system forces extreme locality

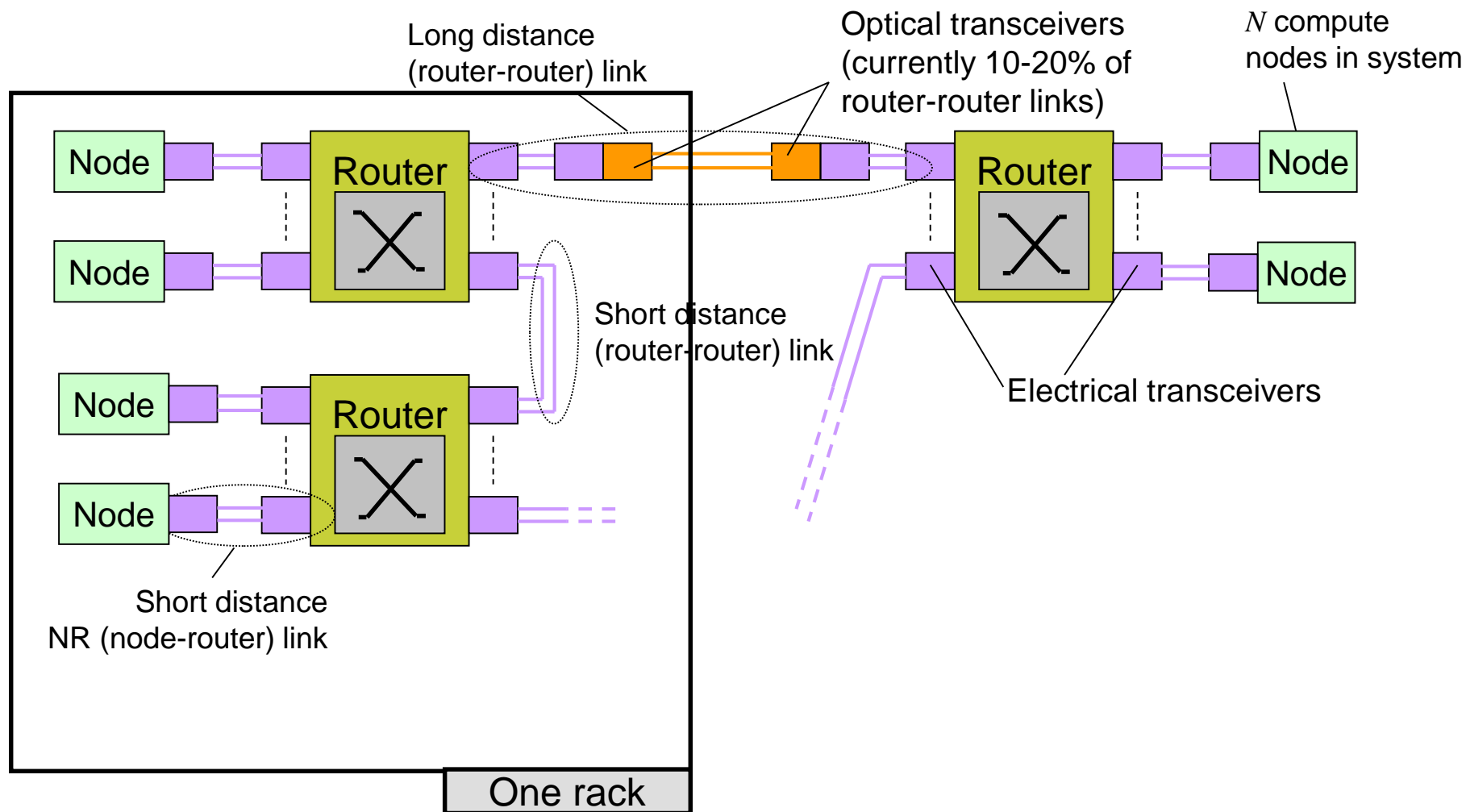


Reduce Energy Consumption

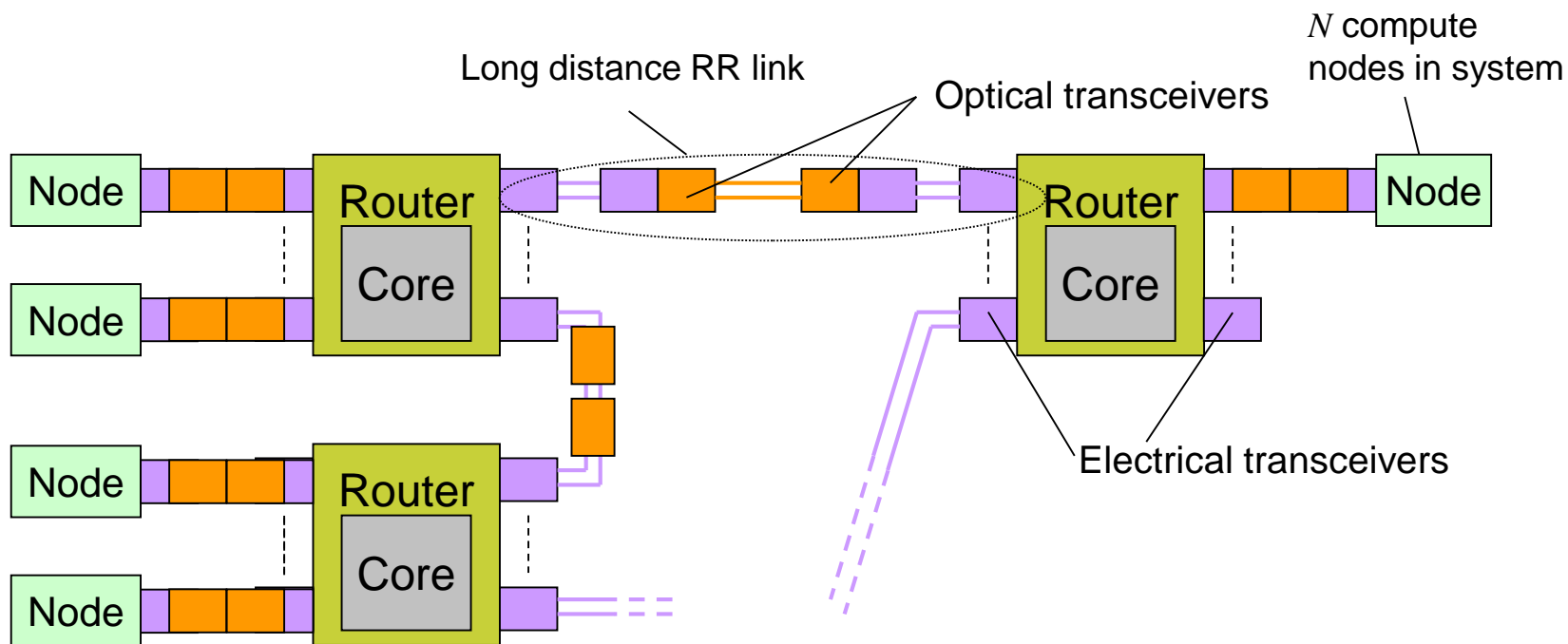


Eliminate Bandwidth Taper

Default interconnect architecture



Exascale optical interconnect ?

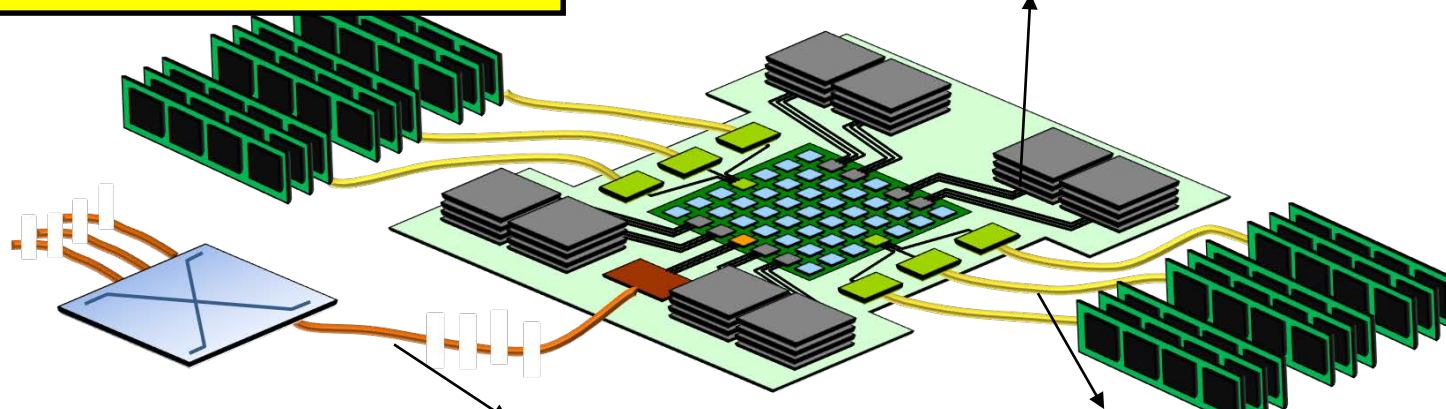


- Requirements for that to happen:
 - Divide cost by 1.5 orders of magnitude at least
 - Improve energy-efficiency by one order of magnitude at least

Exascale supercomputing node

**Compute power:
From 10 to 30 Teraflop (TF)**

Near memory bandwidth:
 $10\text{-}30 \text{ TF} \times 8\text{bit} \times 0.5\text{B/F} = 40 - 120\text{Tb/s}$



Ideal case:
Back to 0.01B/F to
ensure well fed nodes

Interconnect bandwidth:
0.01 B/F \rightarrow 0.8 – 2.4 Tb/s

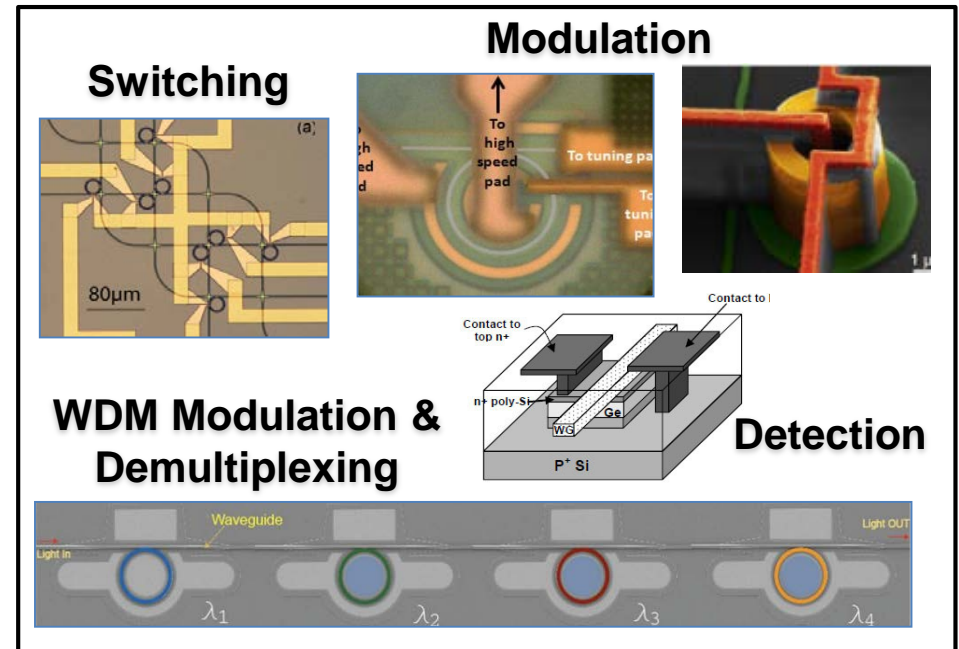
Bulk memory bandwidth:
0.1 B/F \rightarrow 8 - 24 Tb/s

- Consider 50k nodes
 - Total injection bandwidth $\sim 100 \text{ Pb/s}$
 - 4 ZB per year at 1% utilization
 - Total cumulated unidirectional bandwidth: $\sim 500 \text{ Pb/s}$

**Requirements
for next-
generation
interconnects!**

Silicon Photonics: all the parts

- Silicon as core material
 - High refractive index; high contrast; sub micron cross-section, small bend radius.
- Small footprint devices
 - 10 μm – 1 mm scale compared to cm-level scale for telecom
- Low power consumption
 - Can reach <1 pJ/bit per link
- Aggressive WDM platform
 - Bandwidth densities 1-2Tb/s pin IO

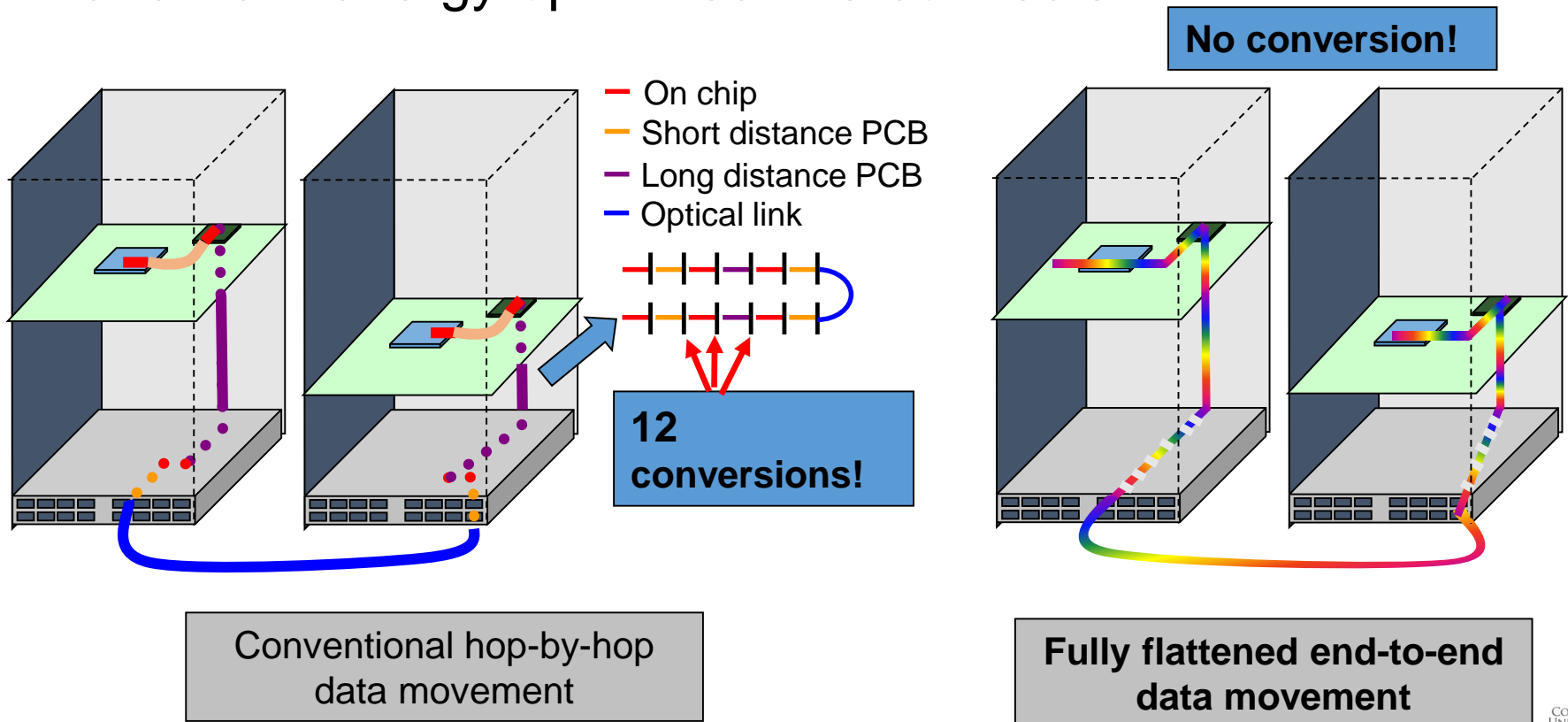


- Silicon wafer-scale CMOS
 - Integration, density scaling
 - CMOS fabrication tools
 - 2.5D and 3D platforms

Photonic Computing Architectures: Beyond Wires

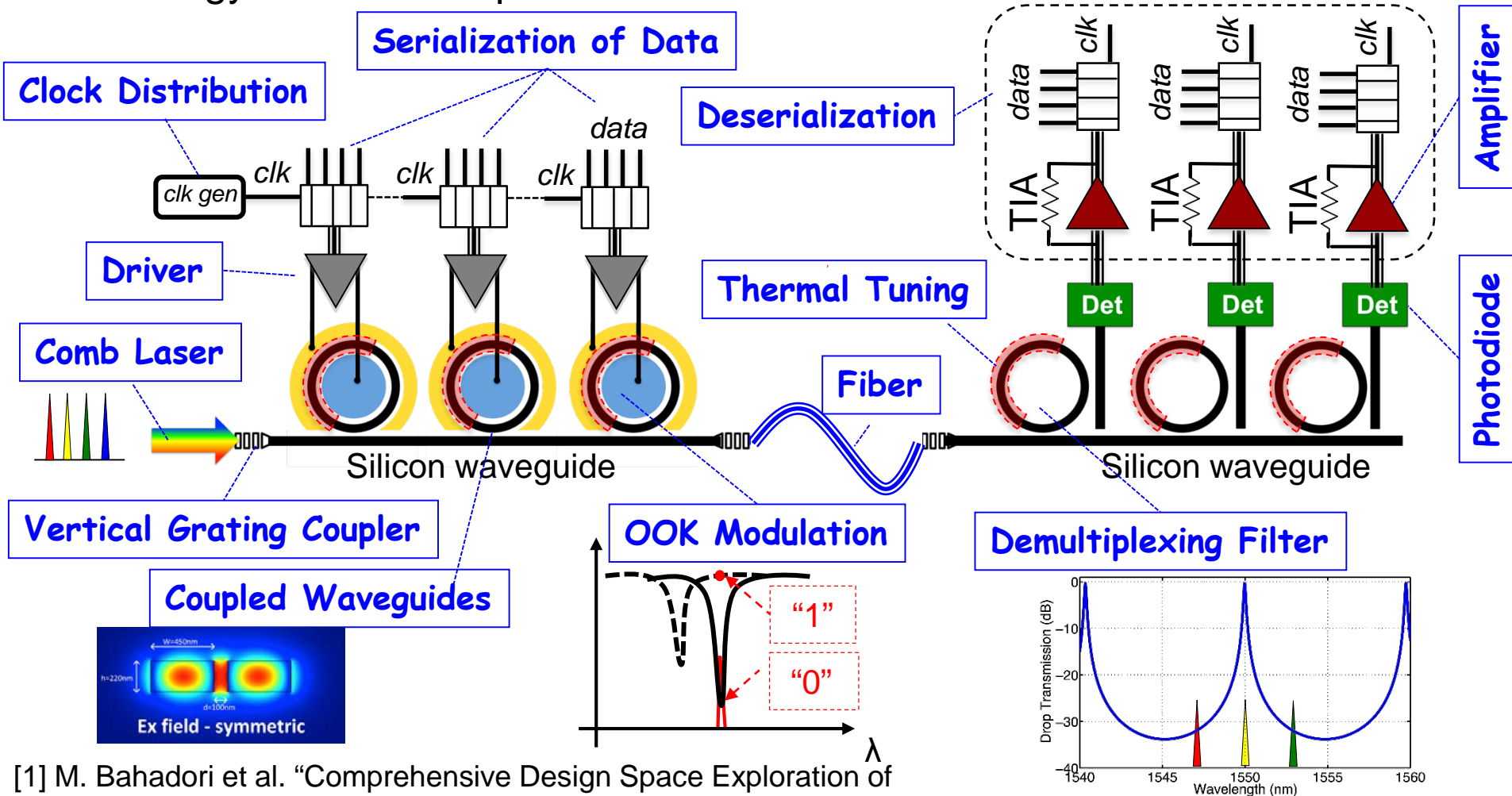
10

- Leverage dense WDM bandwidth density
- Photonic switching
- Distance-independent, cut-through, bufferless
- Bandwidth-energy optimized interconnects



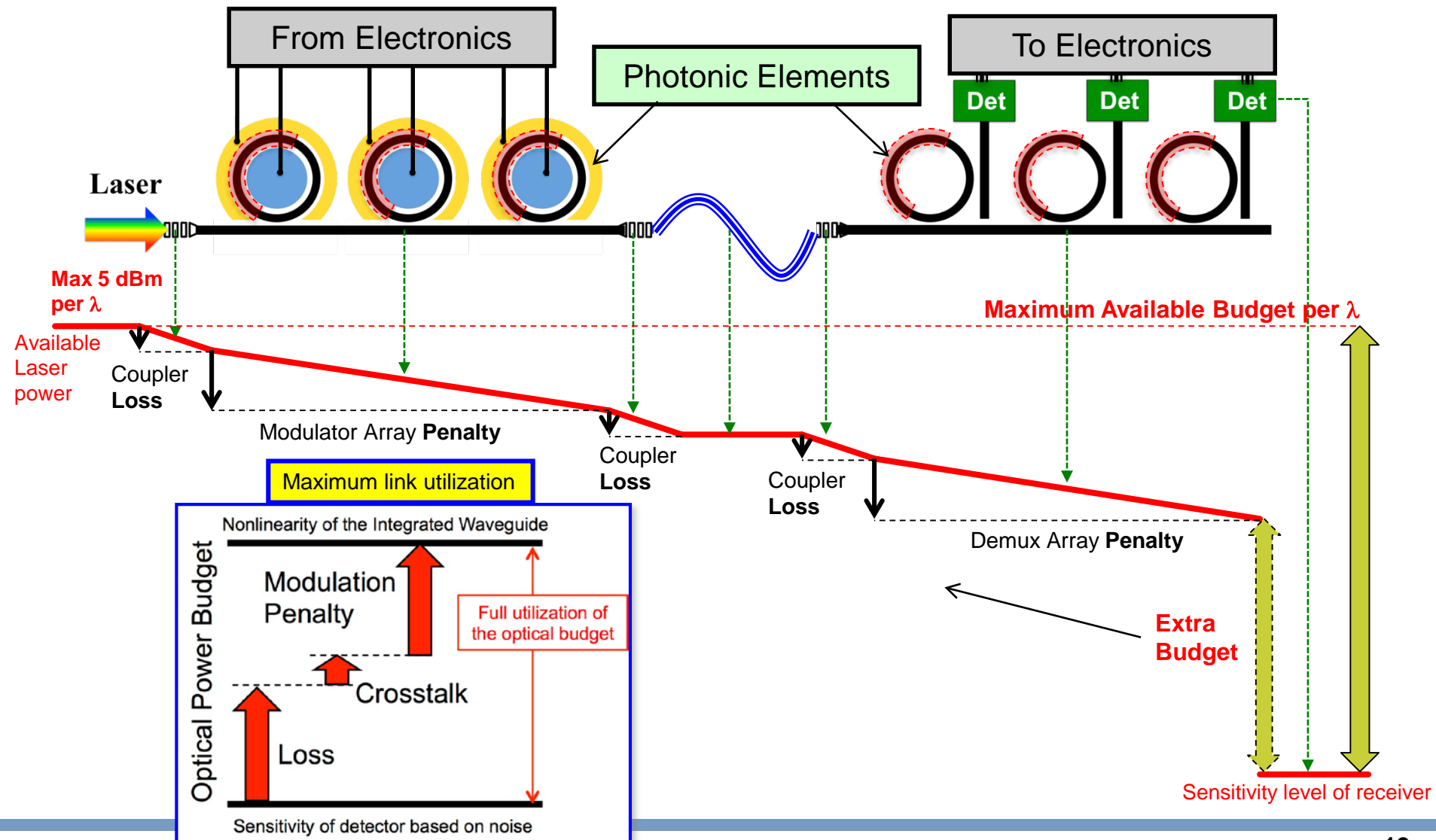
Silicon Photonic Link Design

- Co-existence of Electronics and Photonics
- Energy-Bandwidth optimization

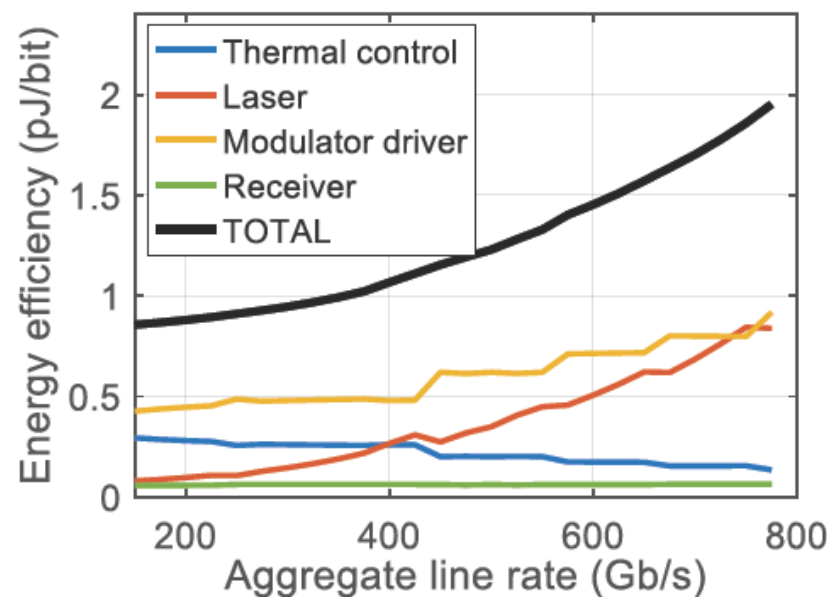
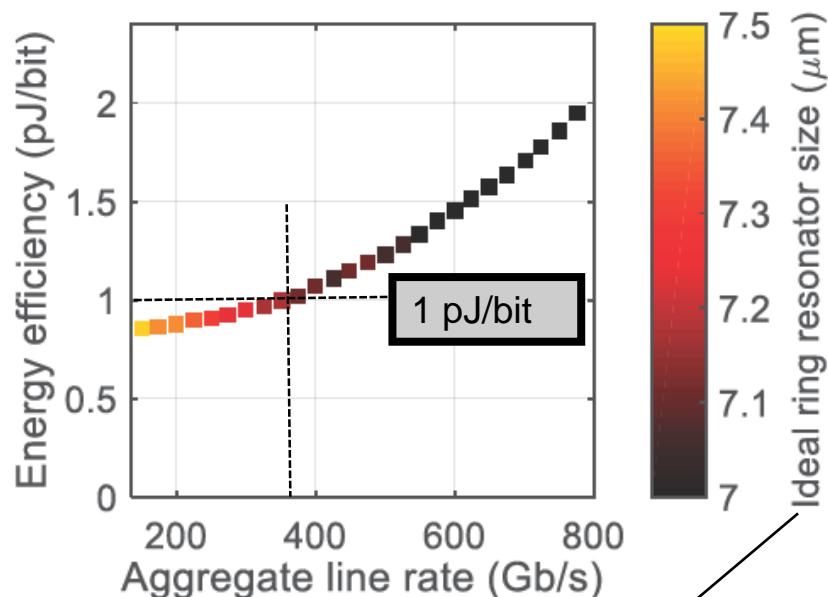
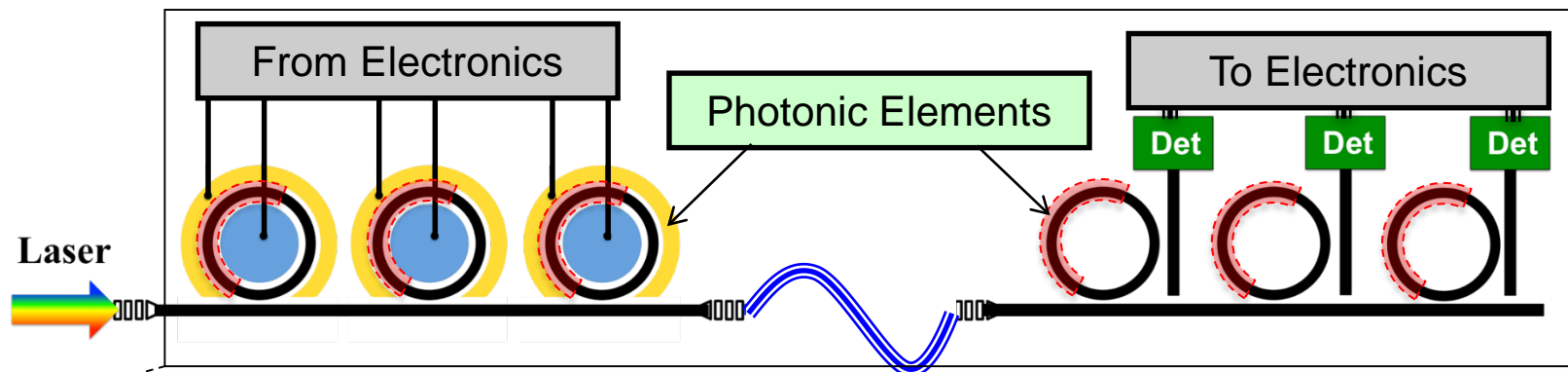


[1] M. Bahadori et al. "Comprehensive Design Space Exploration of Silicon Photonic Interconnects," IEEE JLT 34 (12), 2015.

Utilization of Optical Power Budget



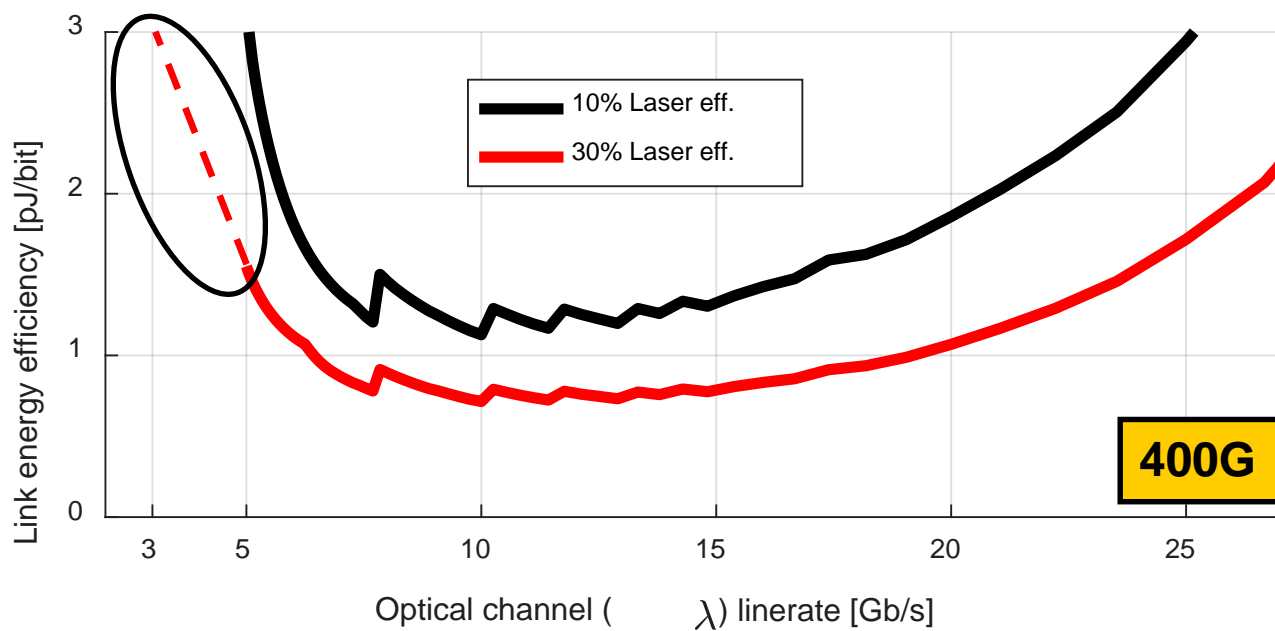
All-Parameter Optimization: Min Energy Design



Optimal design based on physical dimensions

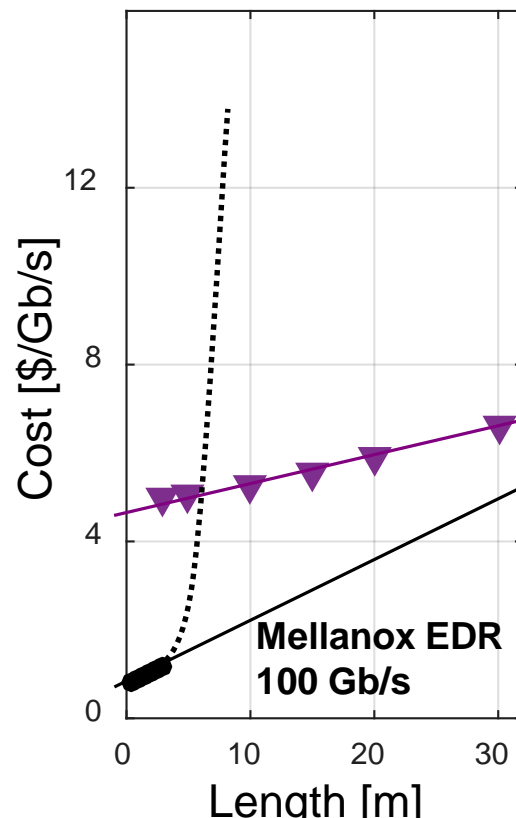
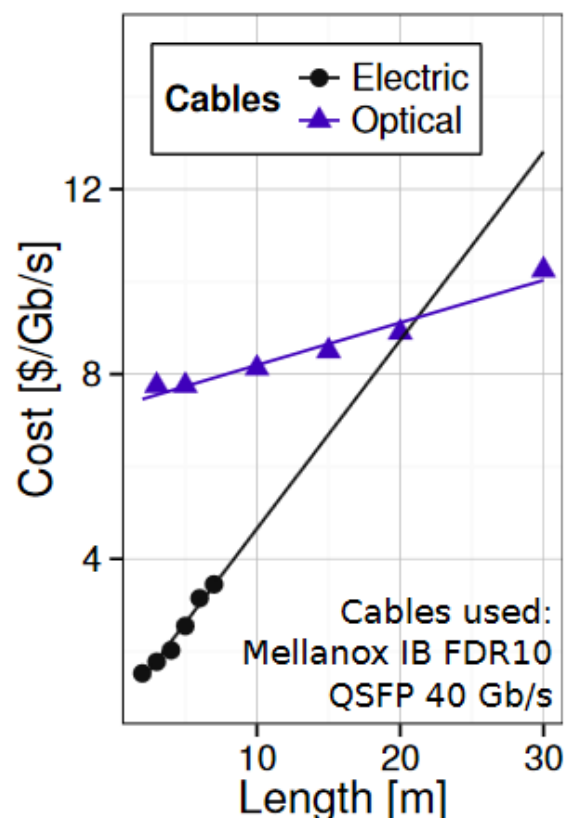
Breakdown of consumption for given bandwidth

Highly parallel, “SERDES-less” links



- Investigate “many-channel” architectures with low bitrate wavelengths
 - Leads to poor laser utilization
 - Only possible with high laser efficiency
 - But may allow drastic simplification of drivers and SERDES blocks

Cost per bandwidth – declining but slowly

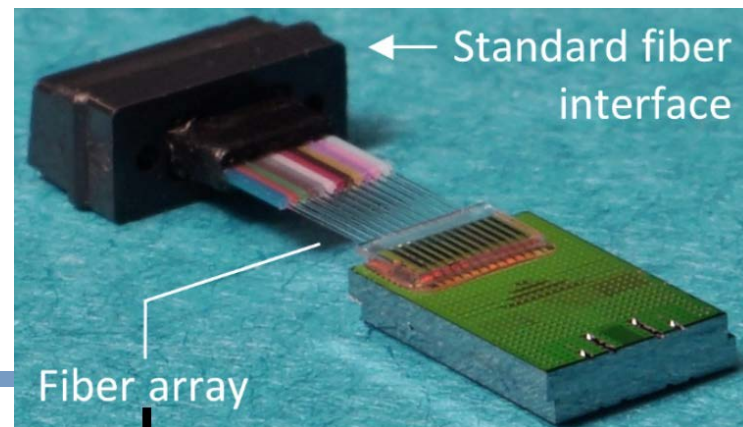
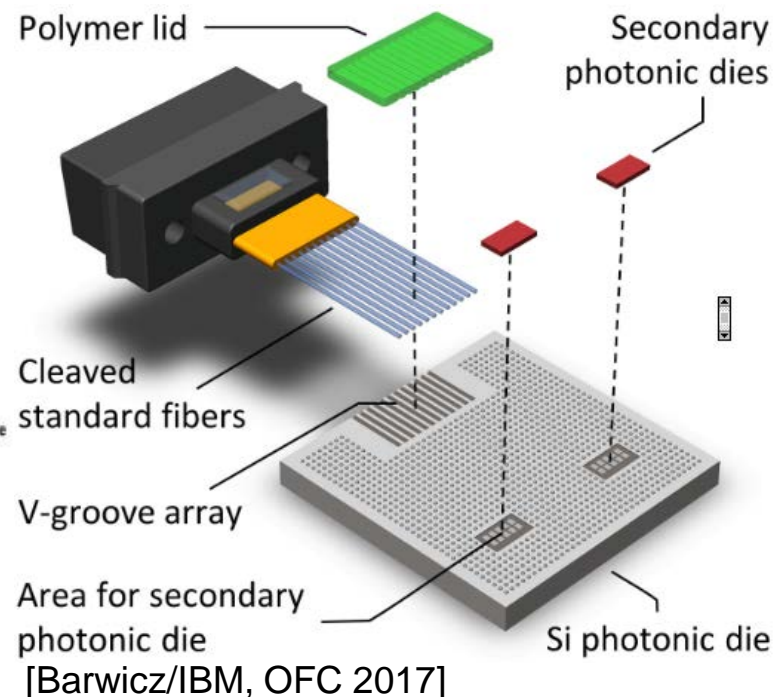
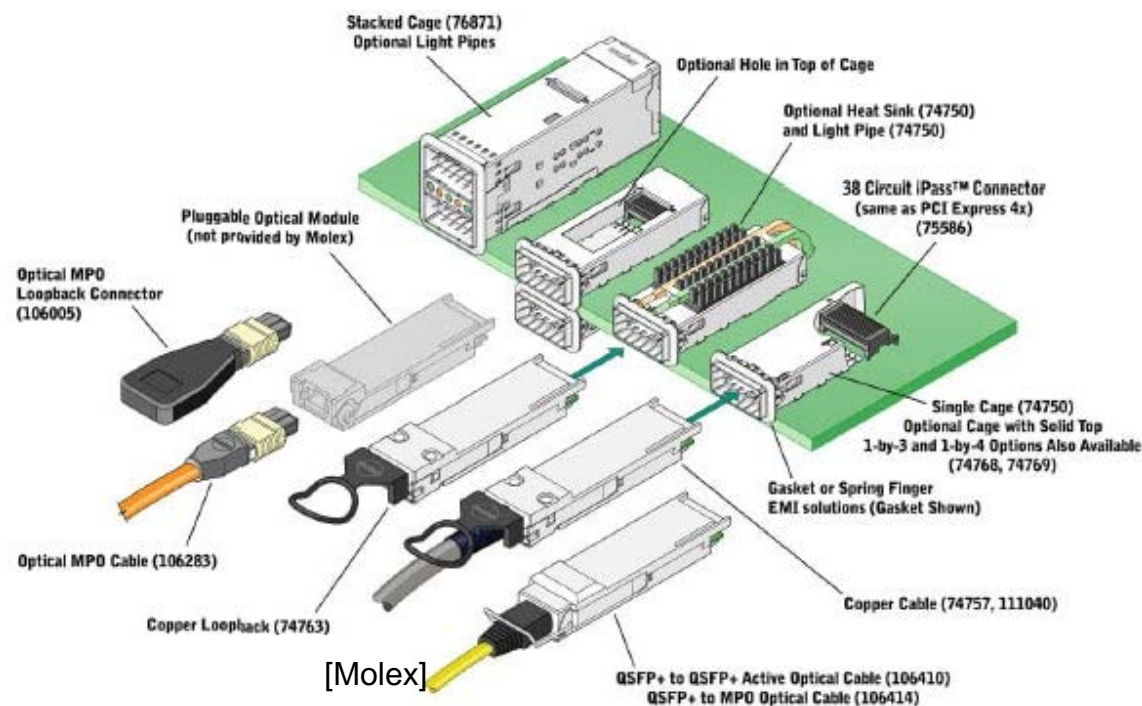


[may 2017]

- Today (2017):
 - 100G (EDR) best \$/Gb/s figure
 - Copper cable have shorter reaches due to higher bit-rate
 - Optics: Not even ½ order of magnitude price drop over 4 years
 - But electrical-optical gap is shrinking

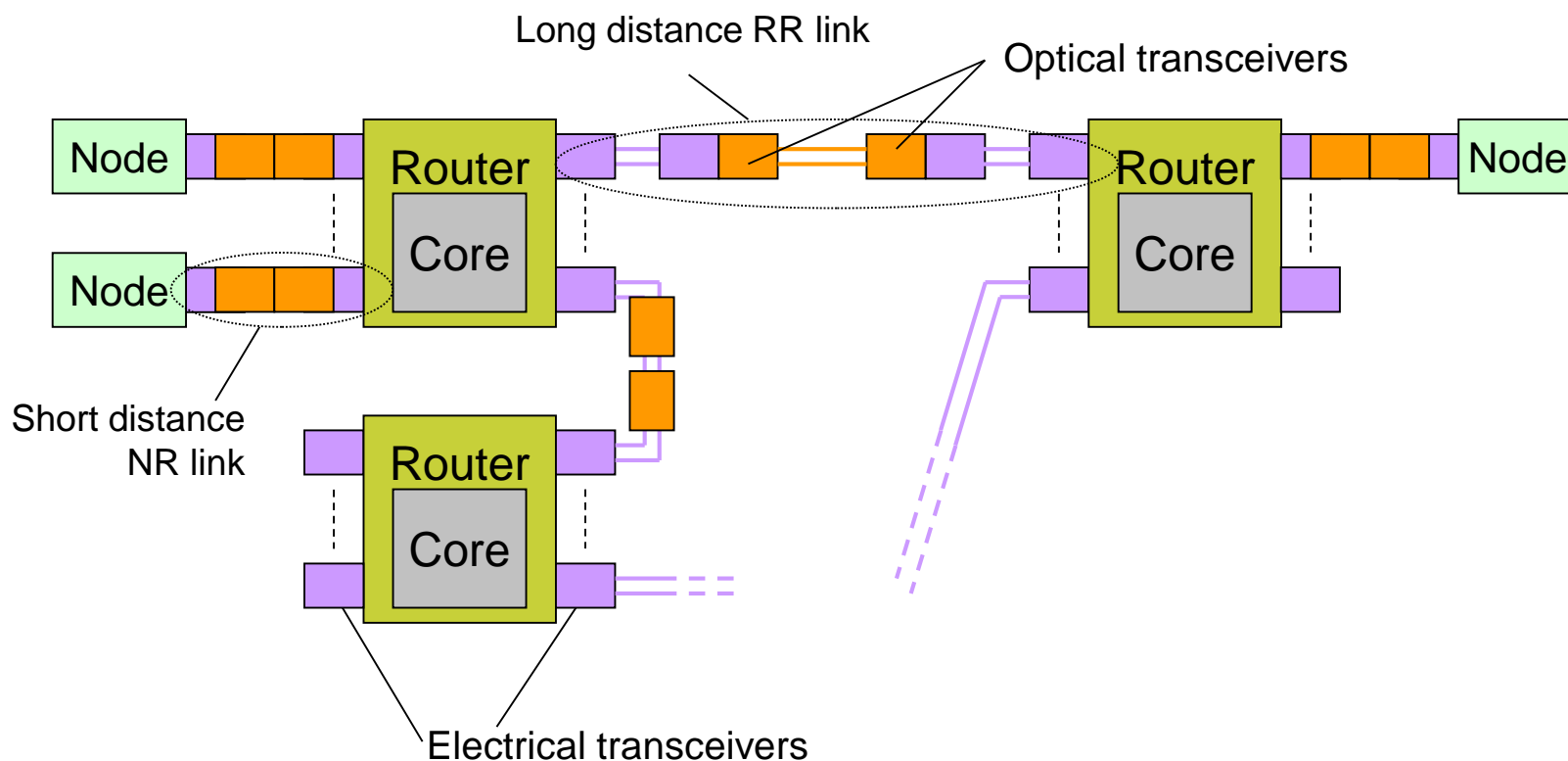
[Besta et al. "Slim Fly: A Cost Effective Low-Diameter Network Topology", SuperComputing 2014]

Packaging and connector drive costs



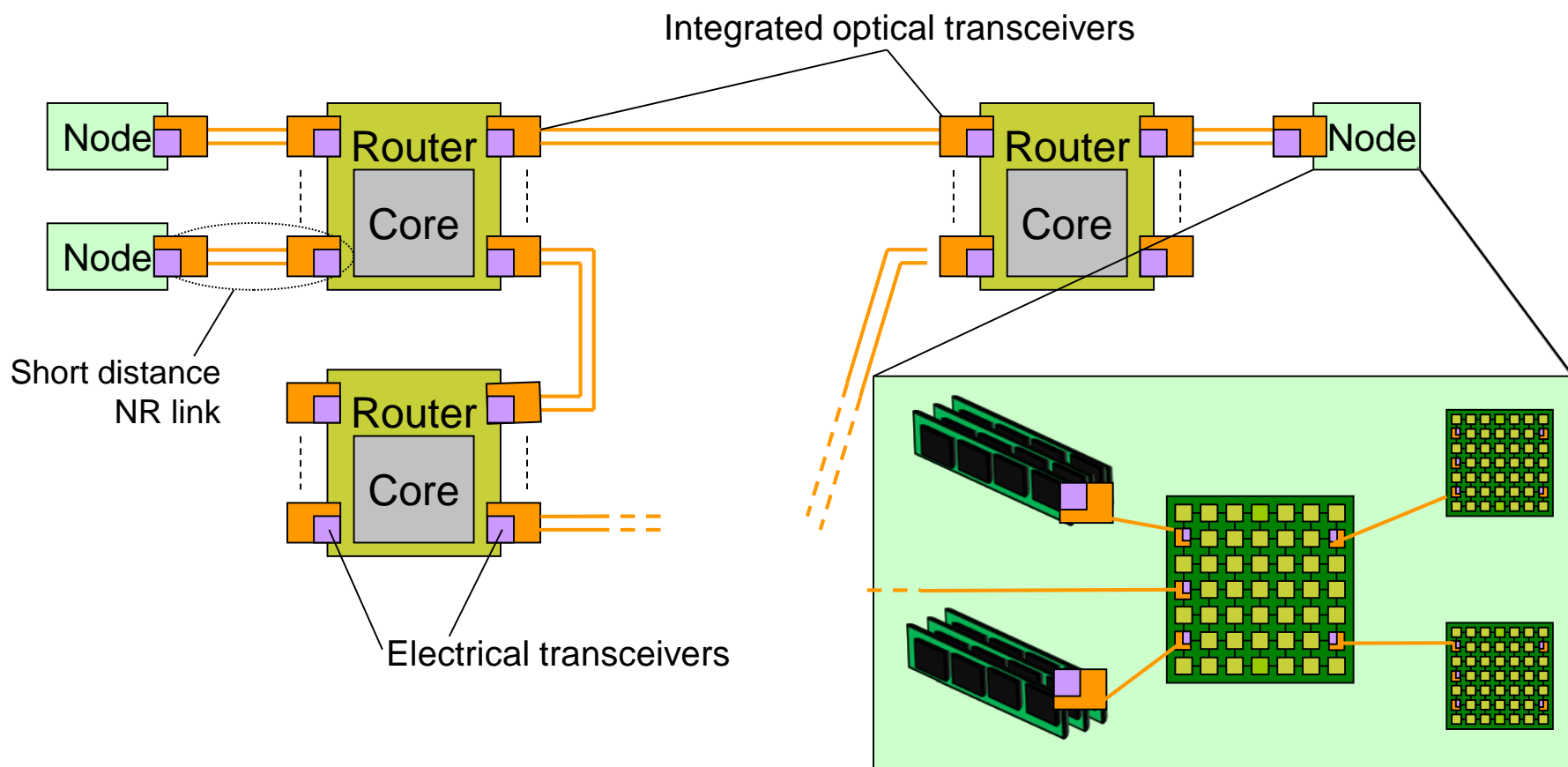
Toward integrated optical transceivers

- Fundamental step to reduce cost and power of optics: co-integration



Towards integrated, general-purpose optical transceivers

- Fundamental step to reduce cost and power of optics





Cost vs. energy

■ Cost of energy over lifetime (\$100/MWh)
■ Procurement cost

Desktop PC (5 years, 100W, \$2000)

Roadrunner (5 years, 2.35 MW, M\$100)

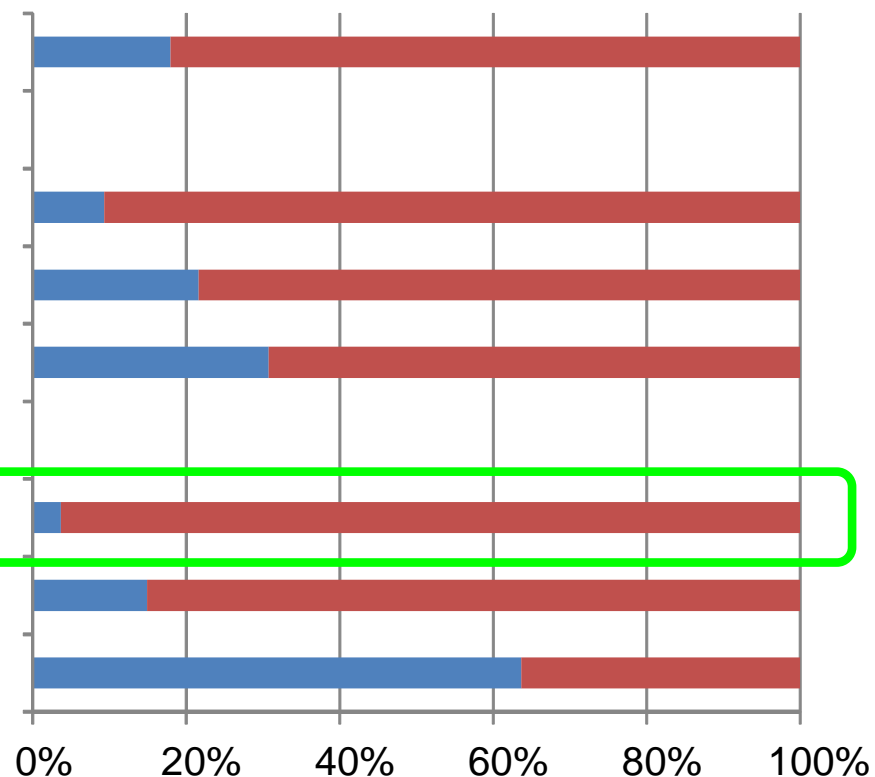
Cori (8 years, 3.9 MW, M\$100)

Titan (6 years, 8.2 MW, M\$97)

Typical 100G transceivers (8 years, 2.5W, \$500)

Hypothetical 1\$/Gb/s transceivers (8 years)



Hypothetical 0.1\$/Gb/s transceivers (8 years)



- Transceiver procurement cost dominates TCO
 - Same energy/procurement ratio as Cori (30% / 70%) with 0.4 \$/Gb/s
 - This is a factor of 10 from the current cost figure

The bright side of optical switching

- Optical switches can have astonishing aggregate bandwidths
 - Absence of signal introspection
 - Possibility to receive dense WDM signals (> 1 Tb/s) on each port
- Translates into alluring \$/Gb/s figures:

Type	# ports	Bandwidth per port	Total bandwidth	Price	Price per Gb/s	
Calient S320 Mems switch	320 ports	400Gbps (with 16 wavelengths at 25G – CDAUI-16 signaling)	128 Tb/s	\$40k	0.3 \$/Gb/s	
Mellanox SB7790	36 ports	100 Gbps (4xEDR Infiniband)	3.6 Tb/s	\$12k	3.3 \$/Gb/s	

→ Optical switching >10 cheaper than electrical packet routing...

Optical switch vs. Electrical packet router

Optical switch: Bufferless

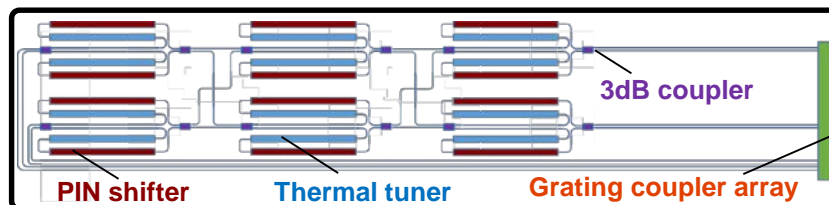
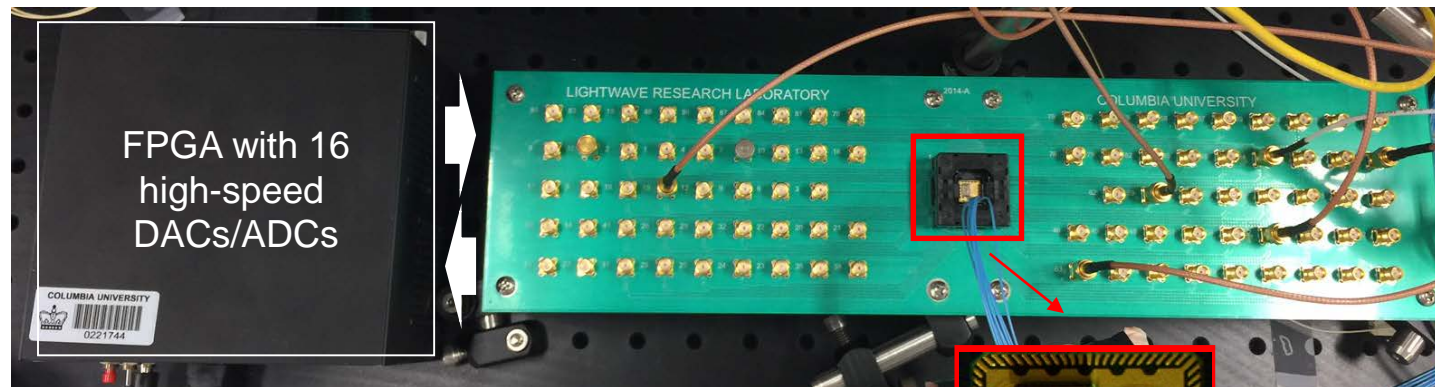


Electrical packet router: Buffered

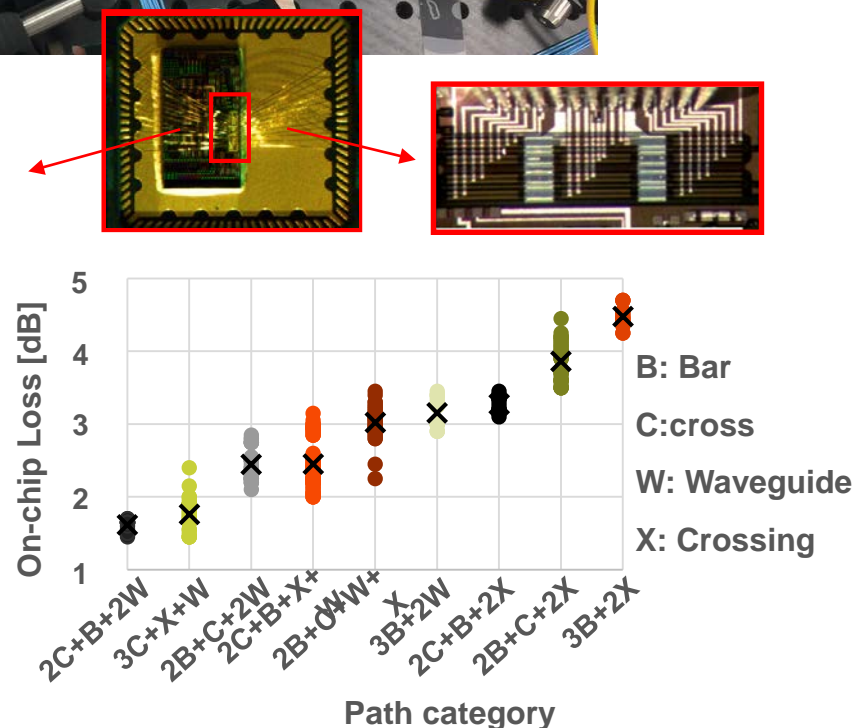


- (Random access) buffering plays a crucial role.
 - Without buffering, end-to-end scheduling is required
 - With buffering, scheduling made link-after-link
 - Without buffering, no back-to-back transmissions
- Packet routers also allow differentiated QoS, error correction, etc.
- **Packet router offers much higher “value” than optical switches**
- My personal guess:
 - At least 30x more value (for 10ns optical switching time)
 - Optical switches not competing with packet routers for “daily switching”

Our FPGA-Controlled Switch Test-Bed



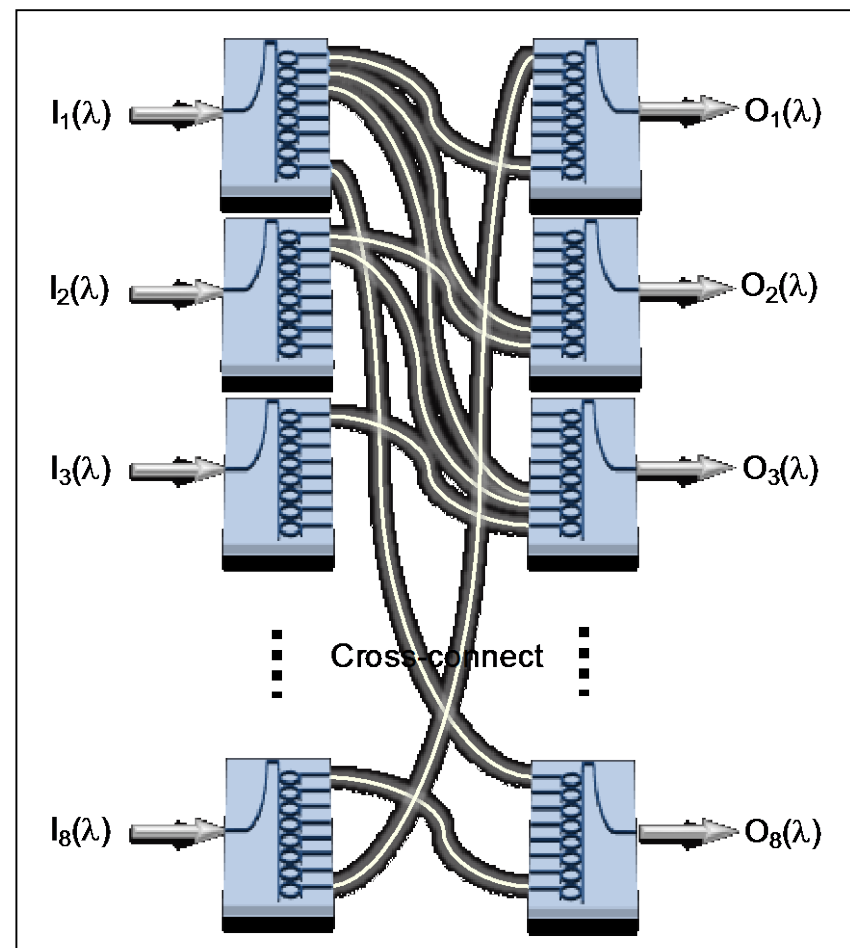
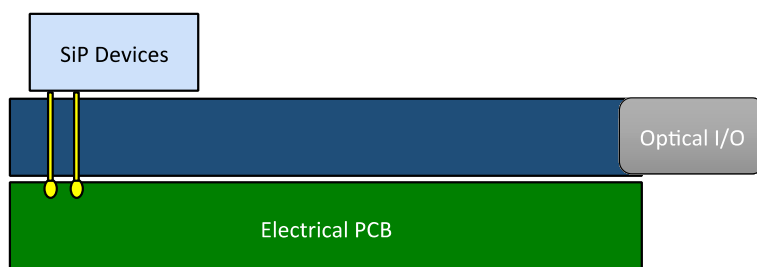
- 16 high-speed DACs enable test and control of integrated photonic switch circuit



Transitioning to Novel Modular Architectures...

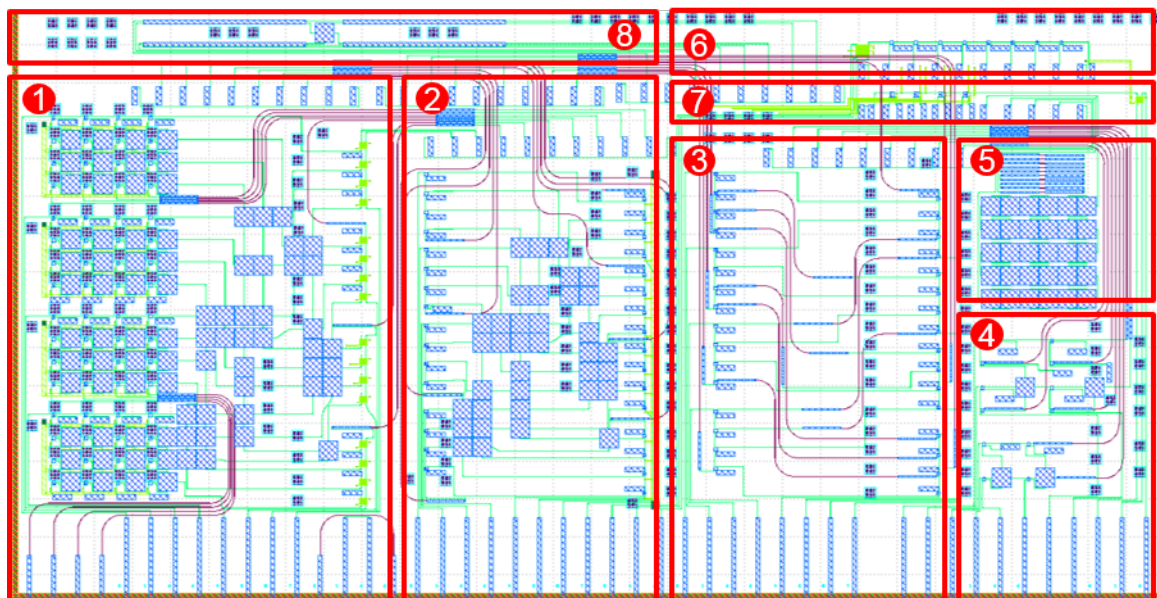
- Modular architecture and control plane
- Avoids on chip crossings
- Fully non-blocking
- Path independent insertion loss
- Low crosstalk

ed integration method



[Dessislava Nikolova*, David M. Calhoun*, Yang Liu, Sébastien Rumley, Ari Novack, Tom Baehr-Jones, Michael Hochberg, Keren Bergman, Modular architecture for fully non-blocking silicon photonic switch fabric, *Nature Microsystems & Nanoengineering* 3 (1607) (Jan 2017).]

AIM Datacom 2nd Tapeout Run

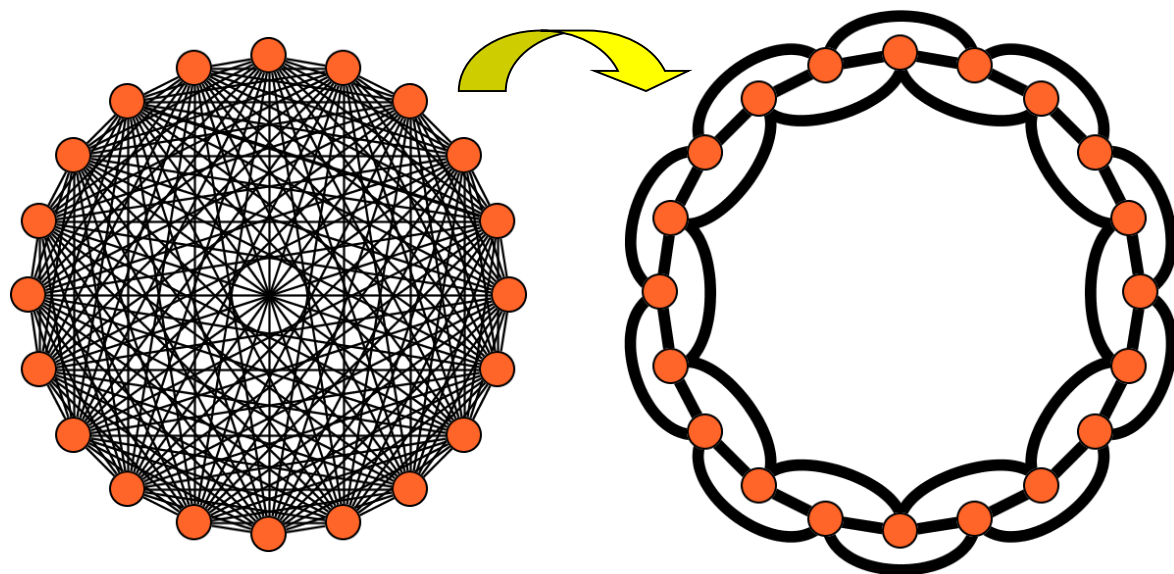
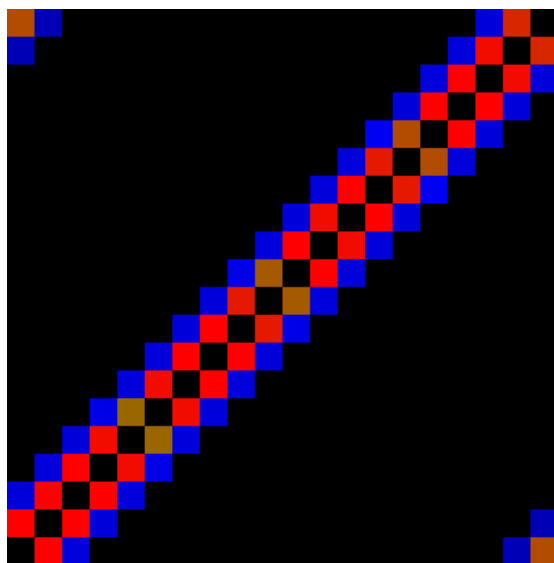


- More efforts poured into monolithically-integrated photonic devices
- Novel switch devices and test structures submitted to AIM platform

	Device	Area
1	4x4x4 λ Space-and-wavelength switch	1.9mm x 2.6mm
2	4x4 Si space switch	1.4mm x 2.3mm
3	4x4 Si/SiN two-layered space switch	1.5mm x 2.3mm
4	2x2 double-gated/single-gated ring switch	0.8mm x 1.4mm
5	Crossing and escalator test structure	0.6mm x 1mm
6	1x2x8 λ MUX with rings	1.2mm x 0.2mm
7	1x2x4 λ MUX with micro-disks	0.6mm x 0.2mm
8	2x2 double-gated MZM switch	3mm x 0.4mm

What optical networks are good at: bandwidth steering

- Put your fibers where traffic is
- Example: 3D stencil traffic over Dragonfly
 - 20 groups, 462 nodes per group, 24x24x16 stencil



- Per workload reconfiguration – avoids switching time overhead
- No need for ultra-large radixes – 8x8 is sufficient

Intra-node bandwidth steering

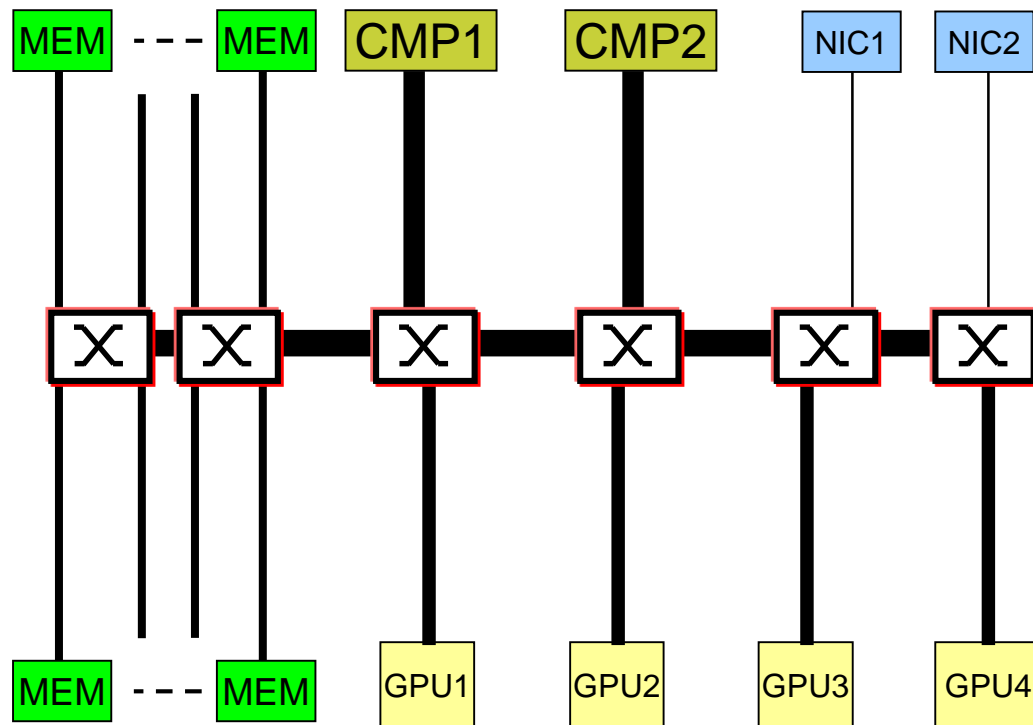
- Emerging architectural concept:
unified interconnect fabric

- AMD's Infinity fabric
- HPE's "The Machine"
- Gen-Z

- Highly flexible!

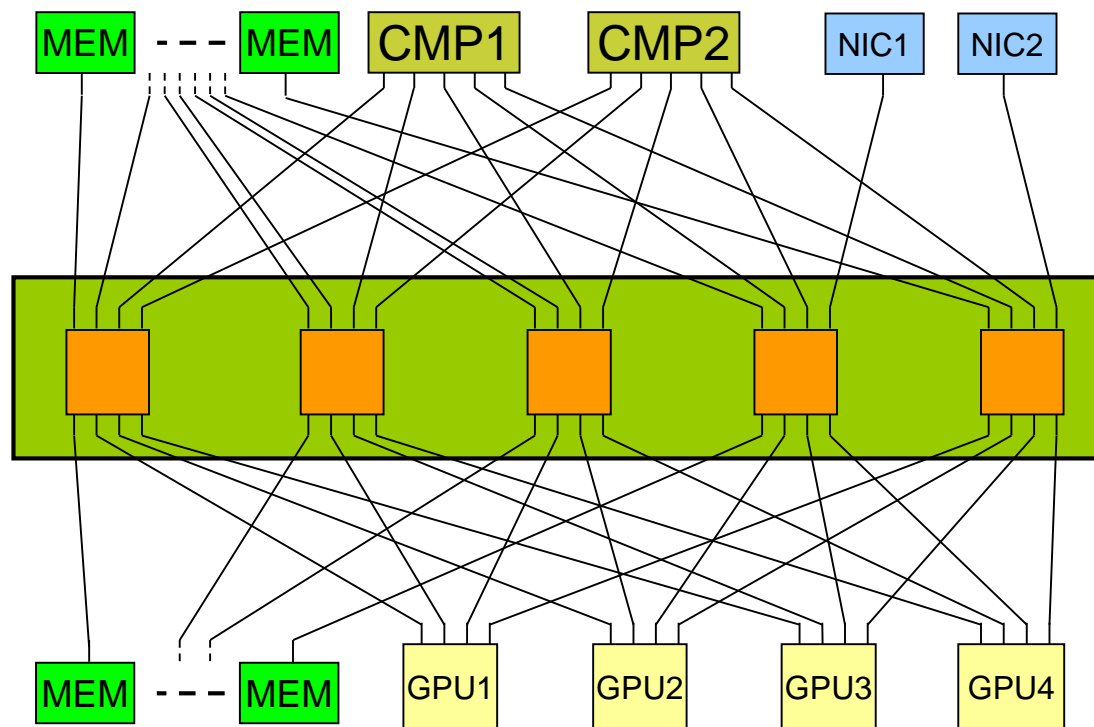
- But involves many hops

- Latency
- Cost/power of routers
- Many chip-to-chip PHYs

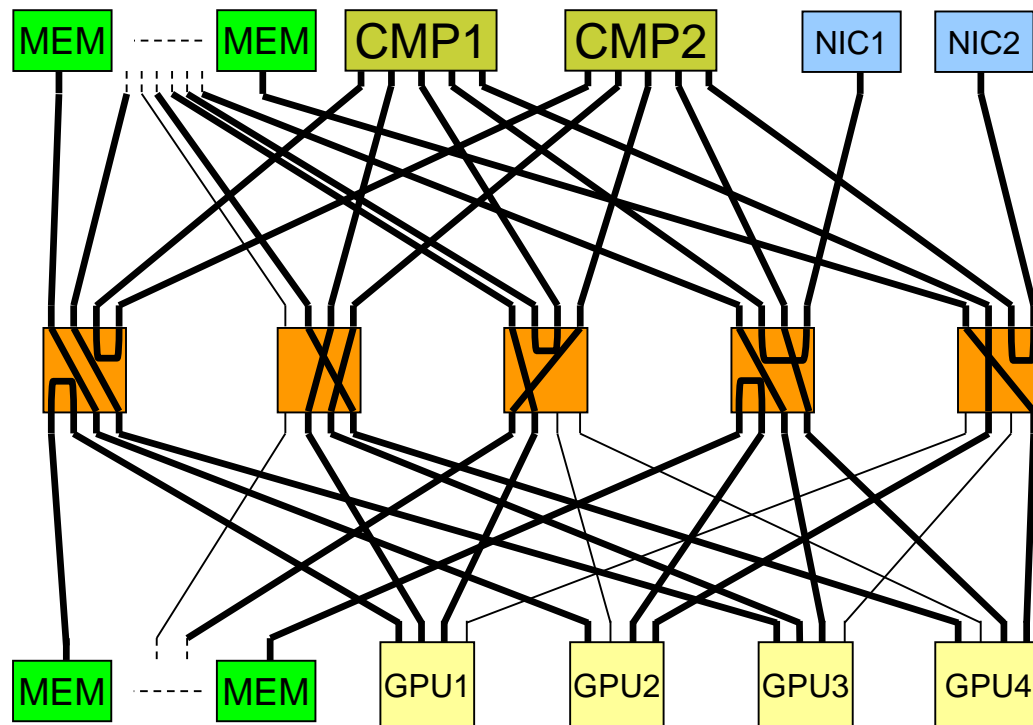
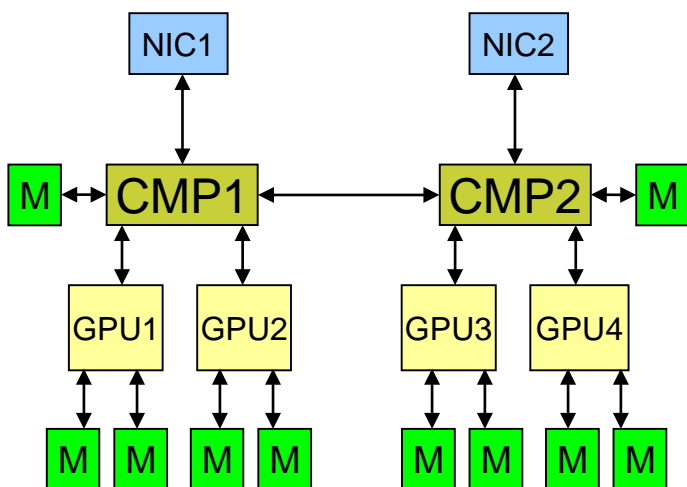


Intra-node bandwidth steering

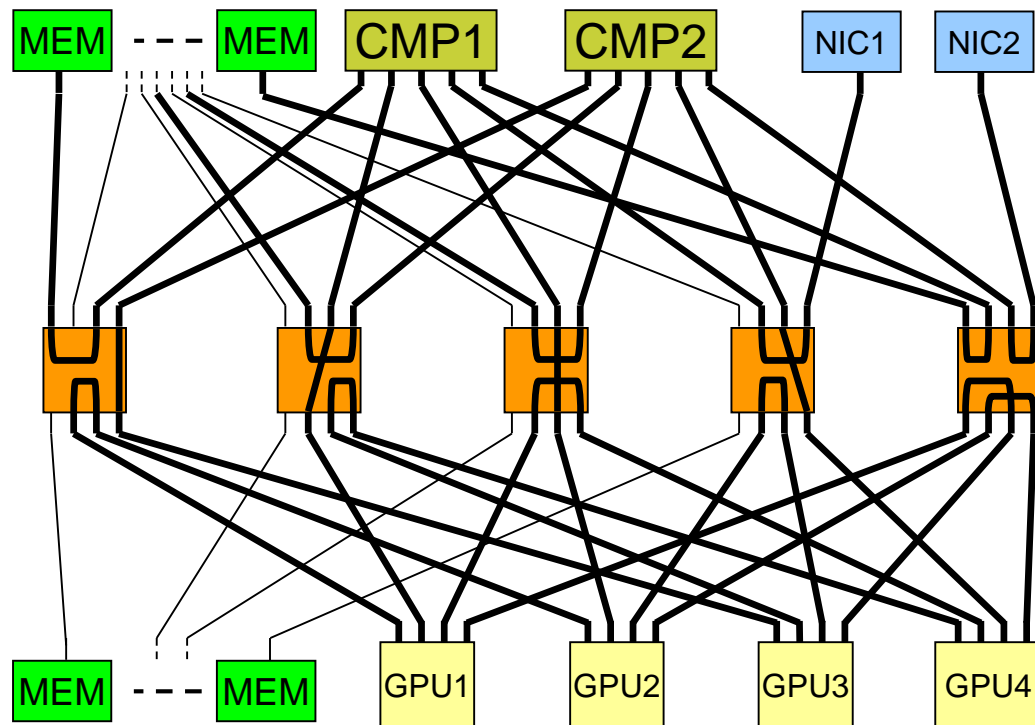
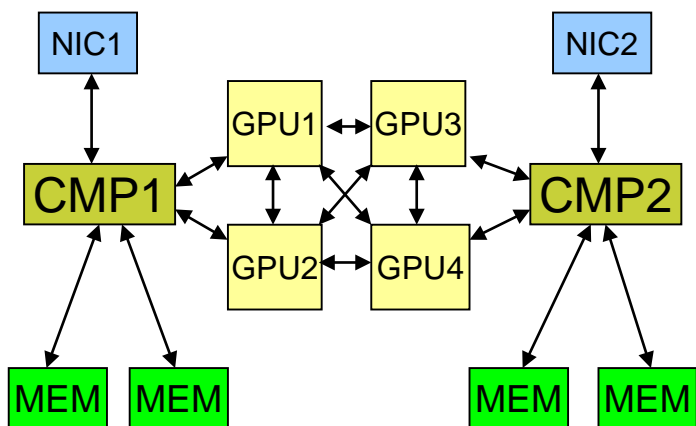
- Alternative concept: use many low-radix optical switches
 - 8x8 realizable with today's technology
 - Tens of switches can be collocated on a single chip
- Somehow less flexible than the packet routed counterpart
 - Not all-to-all
 - Reconfiguration takes microseconds
- But transparent for packets
 - Latency of point-to-point
 - Energy efficient



Conventional architecture



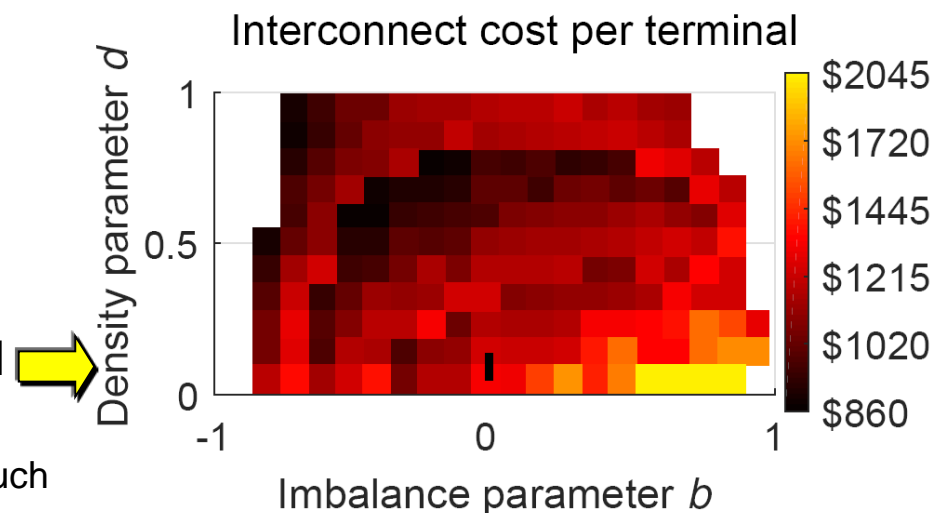
GPU centric / CMPs as data-accelerators



Conclusions

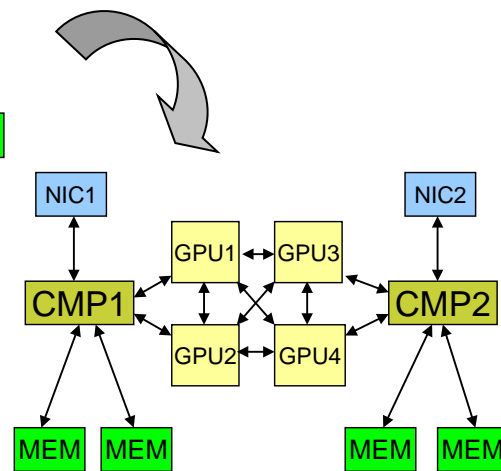
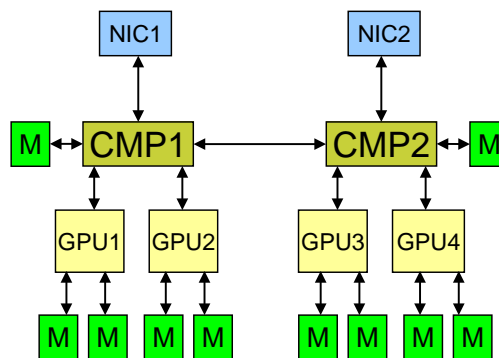
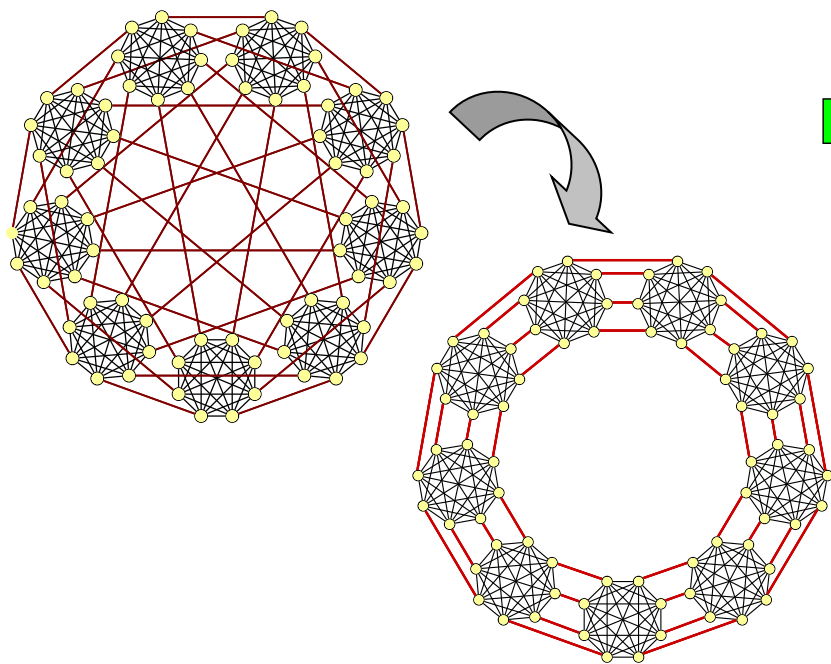
- Lack of bandwidth is threatening scalability
- For Exascale, need to work on (priority-sorted)

- Costs of the optical part
 - Automated packaging and testing
 - Increased integration
 - Larger market
- Power of electrical part
 - Packet routers
- Finely cost/performance-optimized topologies [1]
 - Taper optical bandwidth, but not too much
 - Get as much as we can from optical cables
- Power of optical part
 - Energy-wise optimized designs
 - Improved laser efficiency (technology or “tricks”)
 - Technological advances
- Costs of electrical part



Conclusions

- All-optical interconnects
- Optical switches and packets routers not directly comparable
- Bandwidth-steering based architectures should be explored
 - Optical switches used **in addition to** regular routers



Thank you!

