

# **Optical interconnects for energy efficient HPC**

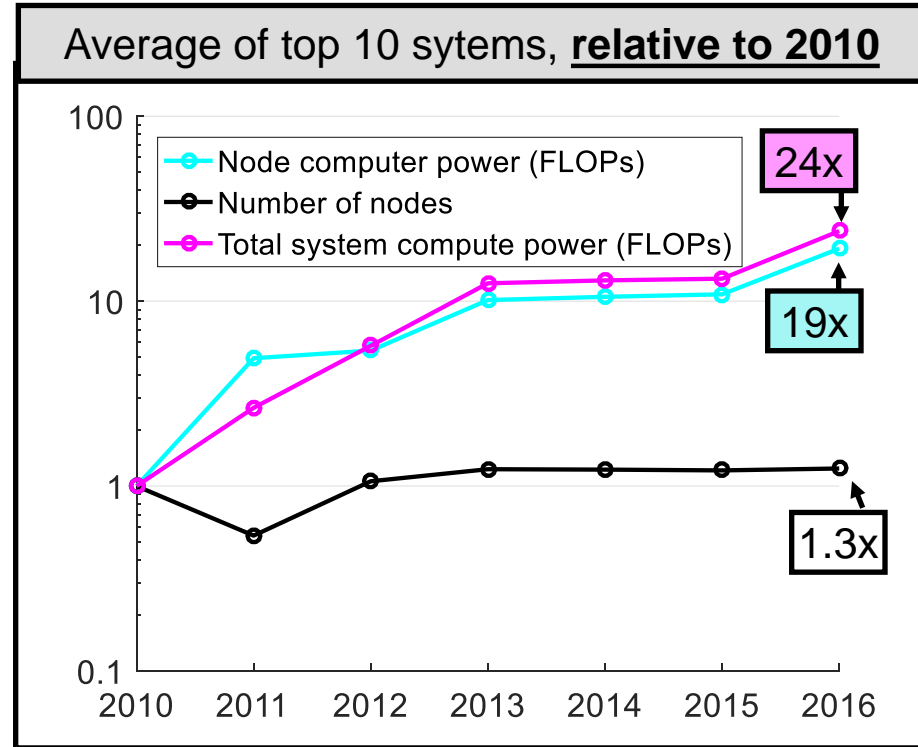
Keren Bergman, Sébastien Rumley

Columbia University  
Lightwave Research Lab

# Trends in ultra-scale HPC

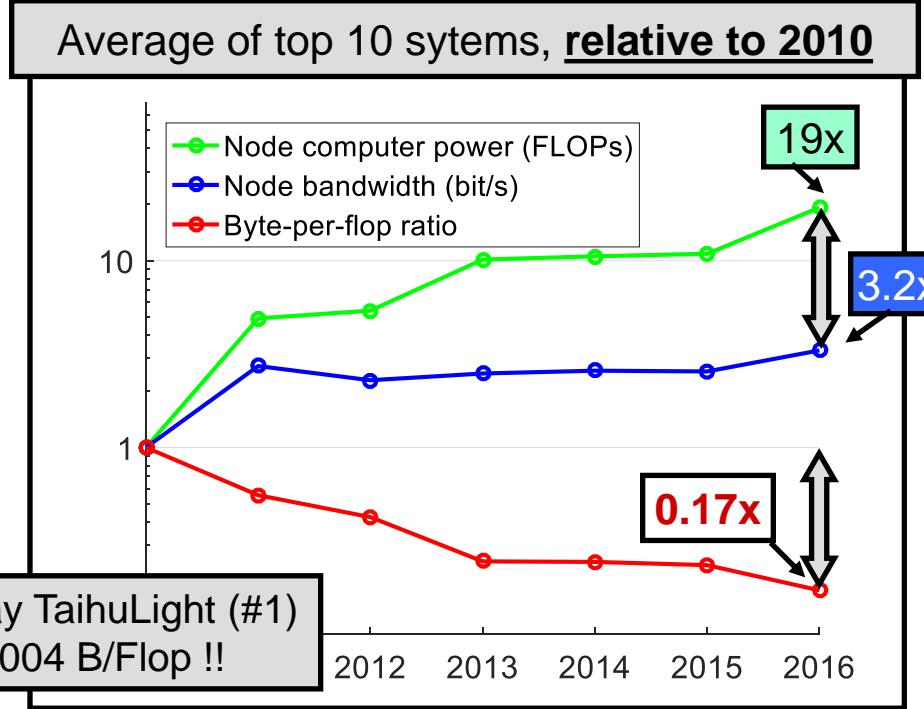
- Evolution in the last six years:
  - Average total compute power:
    - 0.86 PFlops  $\rightarrow$  21 PFlops
    - $\sim 24x$  increase
  - Average node compute power:
    - 31GFlops  $\rightarrow$  600GFlops
    - $\sim 19x$  increase
  - Average number of nodes
    - 28k  $\rightarrow$  35k
    - $\sim 1.3x$  increase

$\rightarrow$  Node compute power main contributor to performance growth



# Interconnect trends in ultra-scale HPC

- Node compute power growth:
  - Average node compute power:
    - 31GFlops → 600GFlops
    - ~19x increase
  - Average node bandwidth
    - 2.7GB/s → 7.8GB/s
    - ~3.2x increase
  - Average byte-per-flop ratio
    - 0.06 B/Flop → 0.01 B/Flop
    - ~6x **decrease**



→ **Growing gap for Interconnect scaling!**

→ **Can this trend continue?**

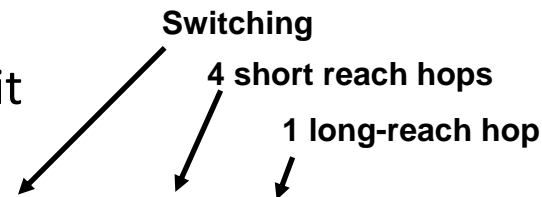
[S. Rumley, et al. Optical Interconnects for Extreme Scale Computing Systems, accepted for publication, Elsevier PARCO]

# Interconnect energy efficiency

- Today's interconnect energy consumption:

- Switching a bit:  $\sim 30$  pJ/bit
- Sending a bit, short-reach electrical link:  $\sim 10$  pJ/bit
- Sending a bit, long-reach optical link:  $\sim 30$  pJ/bit

→ Total budget (diameter 3, one optical link):  $(4 \times 30) + (4 \times 10) + 30 = 170$  pJ/bit

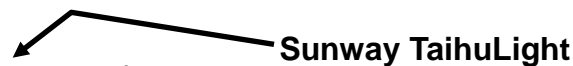


- Assume:

- Exascale at 75% efficiency = 1.3 EFlop peak, 0.004 Byte/Flop

→ 40 Pb/s total interconnect bandwidth

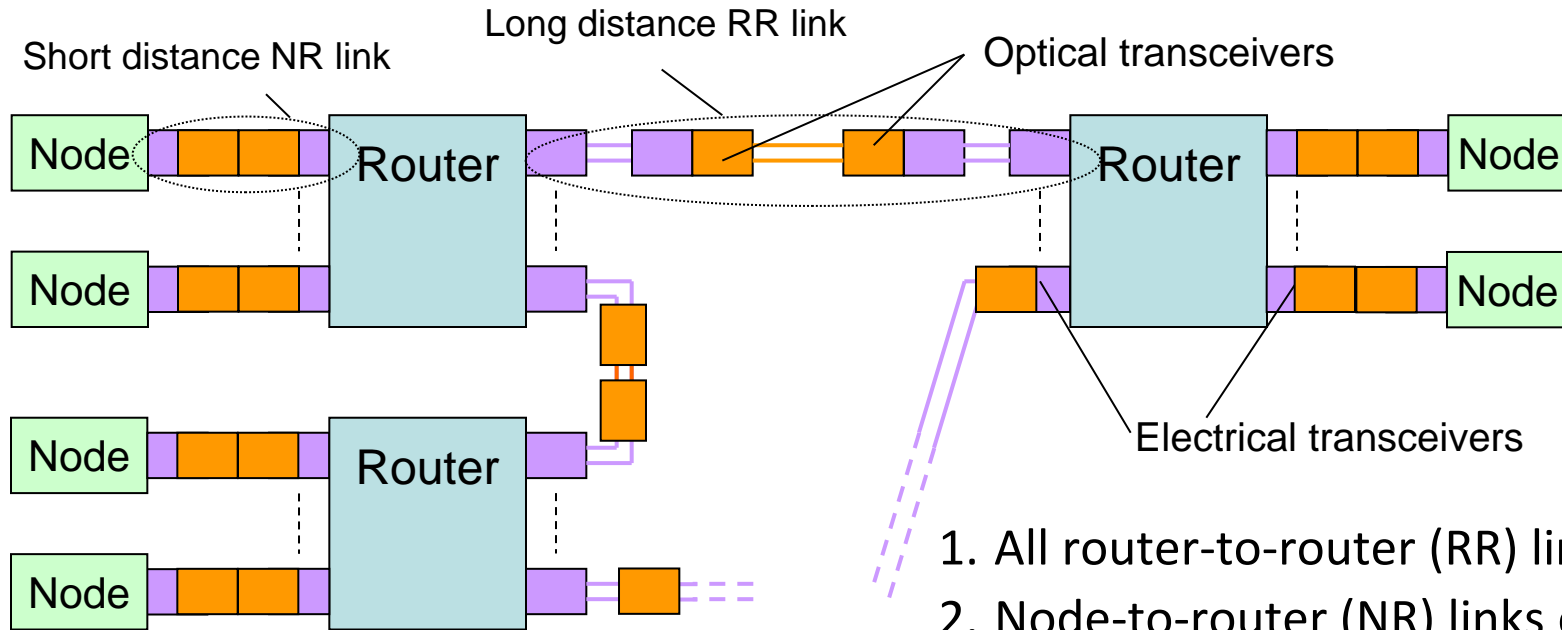
→ **Total interconnect consumption:  $(170 \cdot 10^{-12}) \times (40 \cdot 10^{-15}) = 6.8$  MW**



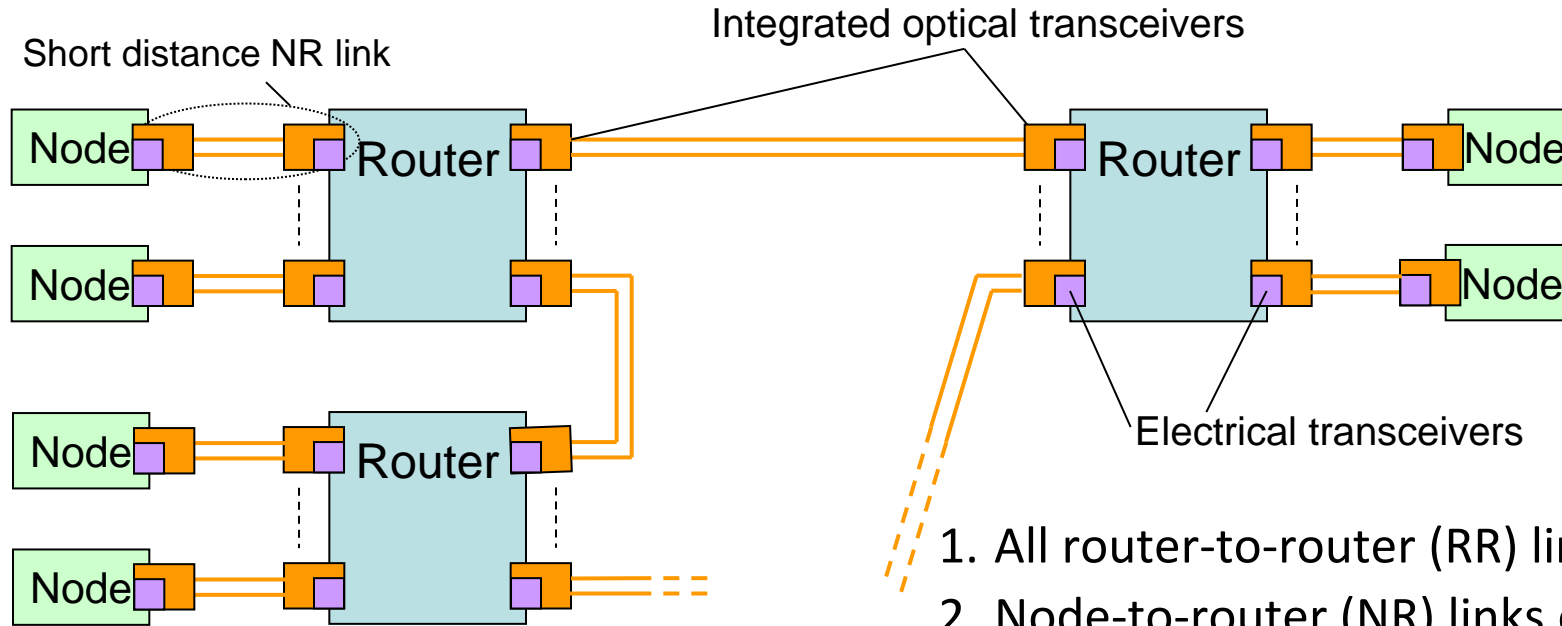
- 6.8 MW: More than a third of target 20MW Exascale power budget!

- Live with that power consumption?
- Tolerate further decrease in byte/flop?
- **Improve interconnect energy efficiency!**

# Improving energy efficiency



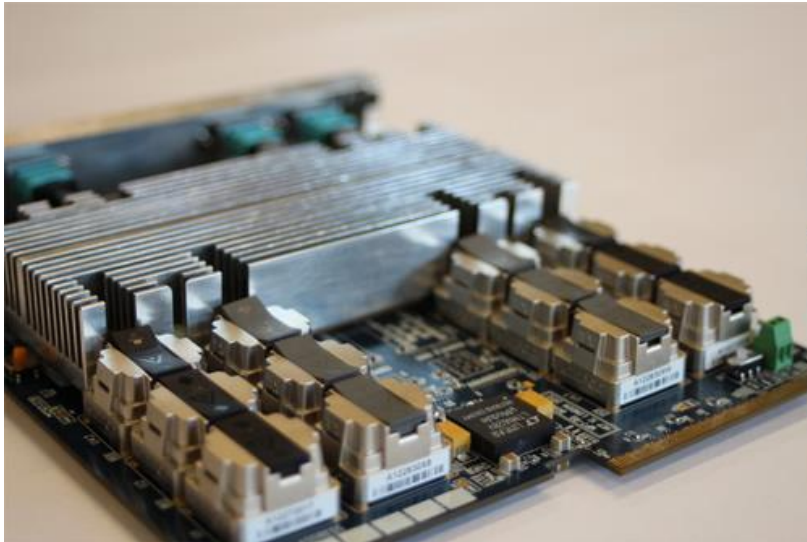
# Improving energy efficiency



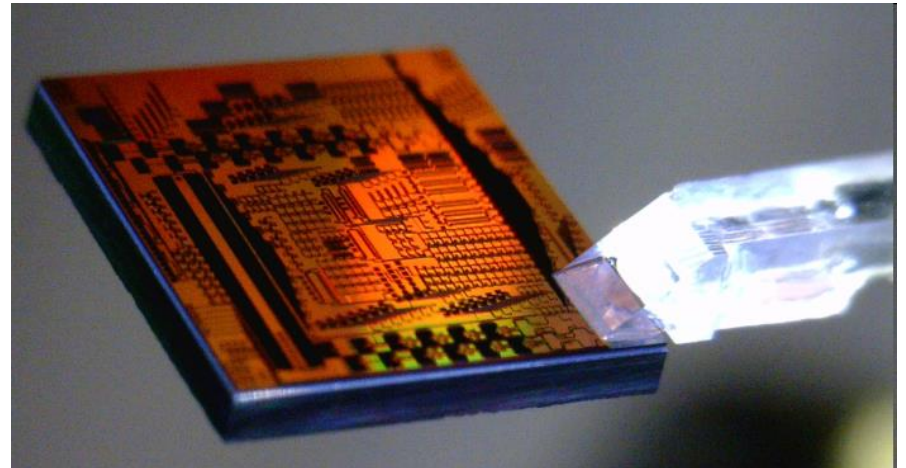
1. All router-to-router (RR) links optical
2. Node-to-router (NR) links optical
3. Co-packaged optical transceivers
- (4. Improve routers)

# Integrated optics – possible solutions

- Vertical Cavity Surface Emitting Laser (**VCSEL**) links
- Silicon Photonics



[Imperial College – MP7 board]

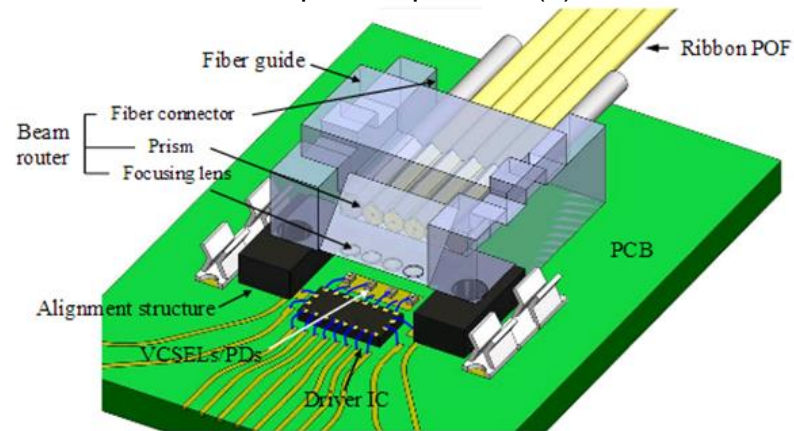


[PLCconnections – OpSIS]

# VCSELS

- Vertical emission by definition
  - If directly mounted on top of CMPs or ASICs, problem with heat sink
  - If “2.5D” integrated, need to electrically “escape” the ASIC/CMP
    - Subject to pin, substrate limitations
    - Less power efficient due to increased capacitances
- Bandwidth density: around 1-5 Gb/s/mm<sup>2</sup> (now)
  - 5 Gb/s/mm<sup>2</sup> → 20 cm<sup>2</sup> footprint for 10 Tb/s
  - Paths for scaling to ~50 Gb/s/mm<sup>2</sup>
    - Shortwave wavelength division multiplexing (SWDM)
    - Higher data-rates, modulation formats (PAM4) (increased latency)

H.-S. Lee, et al., Optics Express 19(5) 2011

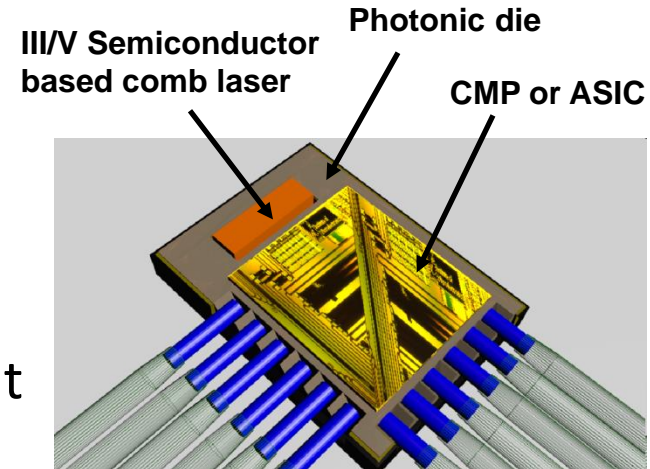
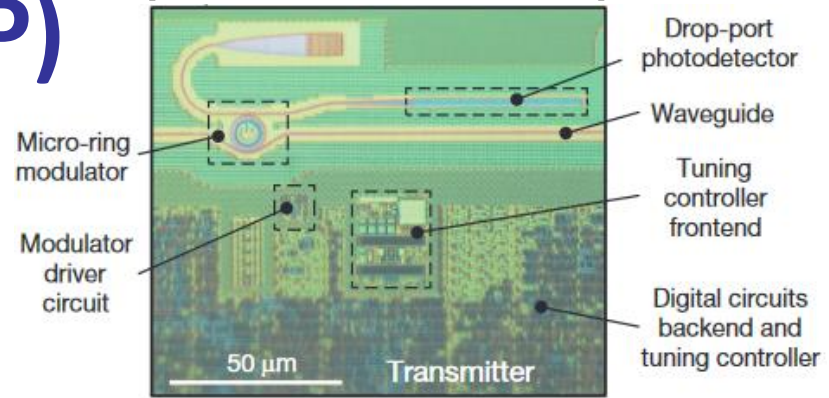




# Silicon Photonics (SiP)

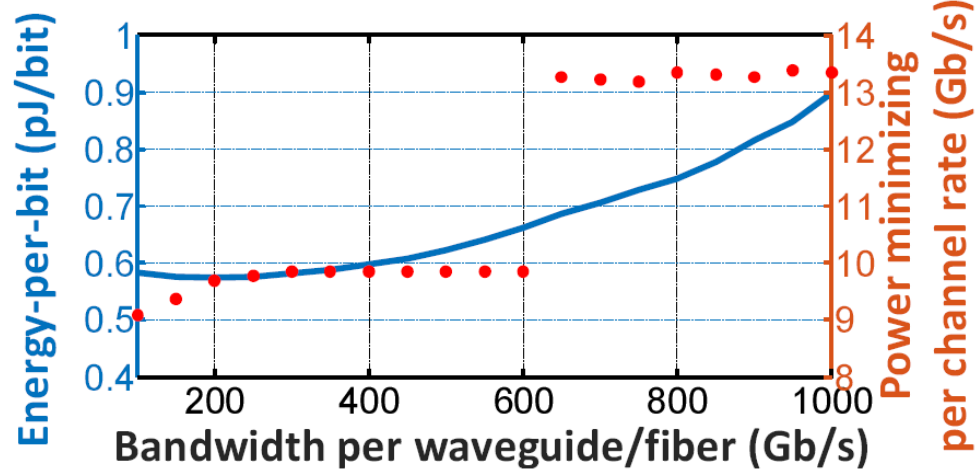
- All optical operations in silicon (except light generation)
- Single-die approach
  - Zero-Change Photonics
    - Photonic and electrical structures in the same die
- Two-dies approach (2.5D and 3D)
  - Photonic die made in SiP optimized process
    - Permits to use of Germanium for detectors
  - Electrical die for digital logic and photonic drivers
  - Two dies combined with flip-chip attachment
- $\sim 100 \text{ Gb/s/mm}^2$  - can scale to  $1 \text{ Tb/s/mm}^2$

[C. Sun *et al.*, Nature 528, 2015]



# Silicon photonics

- Chip-edge coupling possible
  - Pitch between couplers can be as low as  $20\mu\text{m}$  [1]
  - Support for 100+ fibers around the optical chip
- High bandwidth density
  - 320 Gb/s per waveguide/fiber demonstrated [2]
    - Higher densities possible up to 1 Tb/s
- Energy efficiency:
  - Below 1 pJ/bit [3] (excluding SERDES)



[1] F. E. Doany, et al. Journal of Lightwave Technology 29(4), 2011  
 [2] R. Ding, et al. IEEE Photonics Journal 6(3), 2014  
 [3] R. Hendry, et al. Hot Interconnects, 2014.

# Conclusions

- A power wall is clearly approaching for interconnects
  - So far, avoided by means of bandwidth tapering
  - But how far can we further go? 0.001 B/Flop ? 0.0001 B/Flop?
- Energy-efficiency improvements are possible
  - In particular, more integrated, energy optimized optics
- Future integrated photonic solutions:
  - MCM-to-MCM (Multi-chip module) links with VCSELs
    - More mature, but bandwidth density limited
  - Chip-to-chip optical links with Silicon Photonics
    - Long-term solution, manufacturing ecosystem challenges