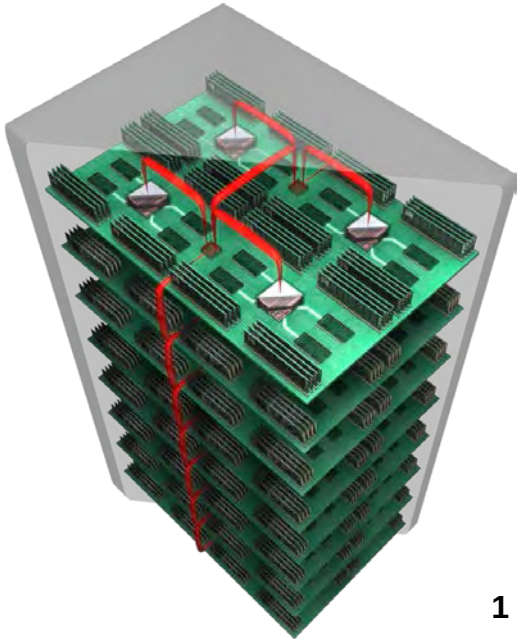
# Crosslayer Design and Modeling for Future Optical Interconnects

**MODSIM 2017, August 10th, Seattle, WA**

Keren Bergman

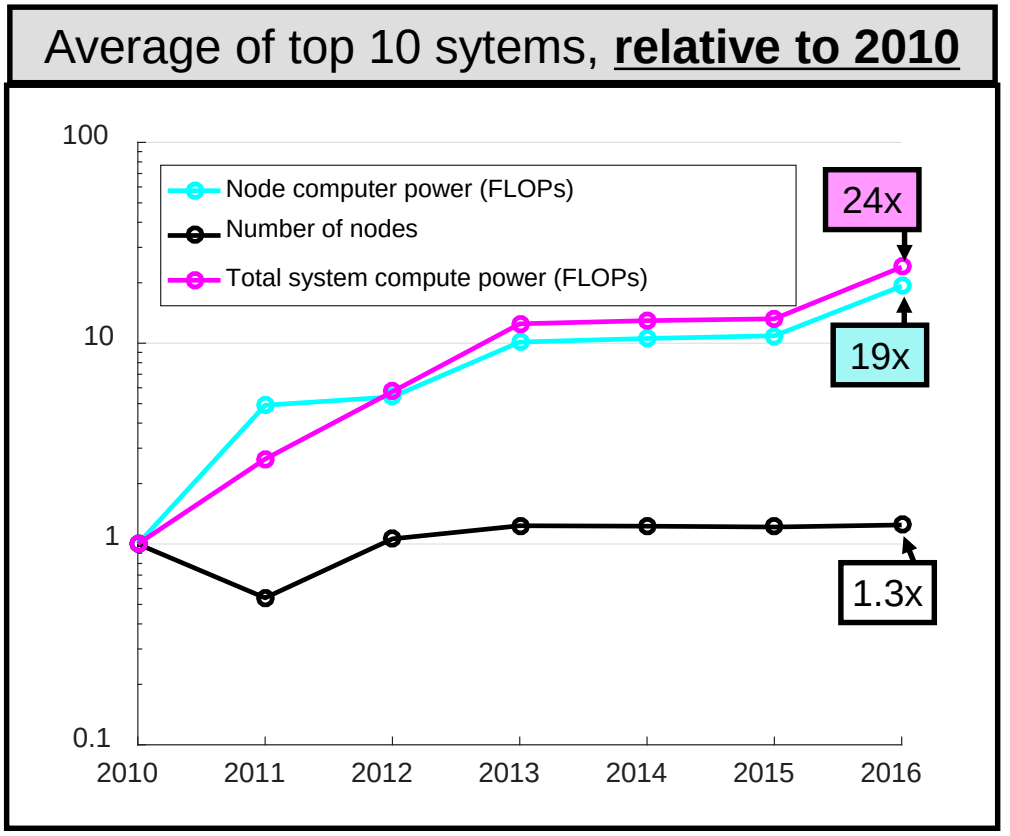
*Lightwave Research Lab, Columbia University*

*New York*



# 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

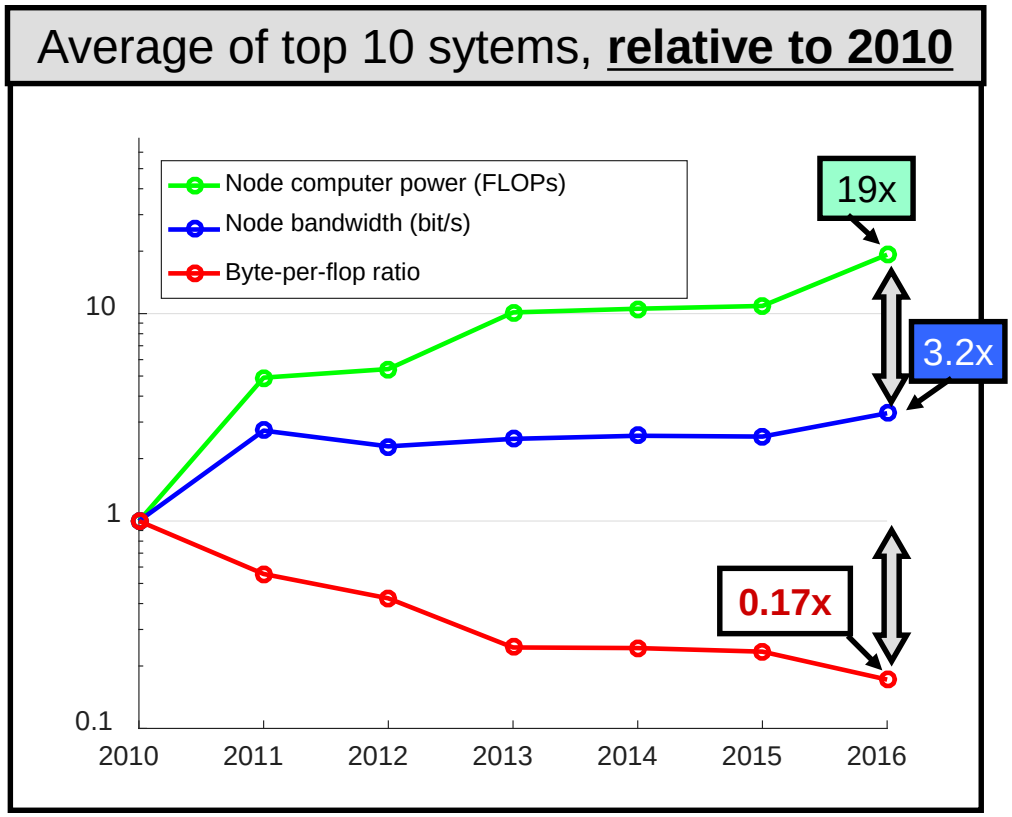



[top500.org, S. Rumley, et al. Optical Interconnects for Extreme Scale Computing Systems, Elsevier PARCO 64, 2017]

→ 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 performance**
  - **…while satisfying constraints (**20MW, $200M)
  - **…with reasonably useful applications**

**1.25 ExaFLOP**
**X 0.01 B/FLOP**

**= 125 Pb/s injection BW**
**X 4 hops**

**= 500 Pb/s installed BW**

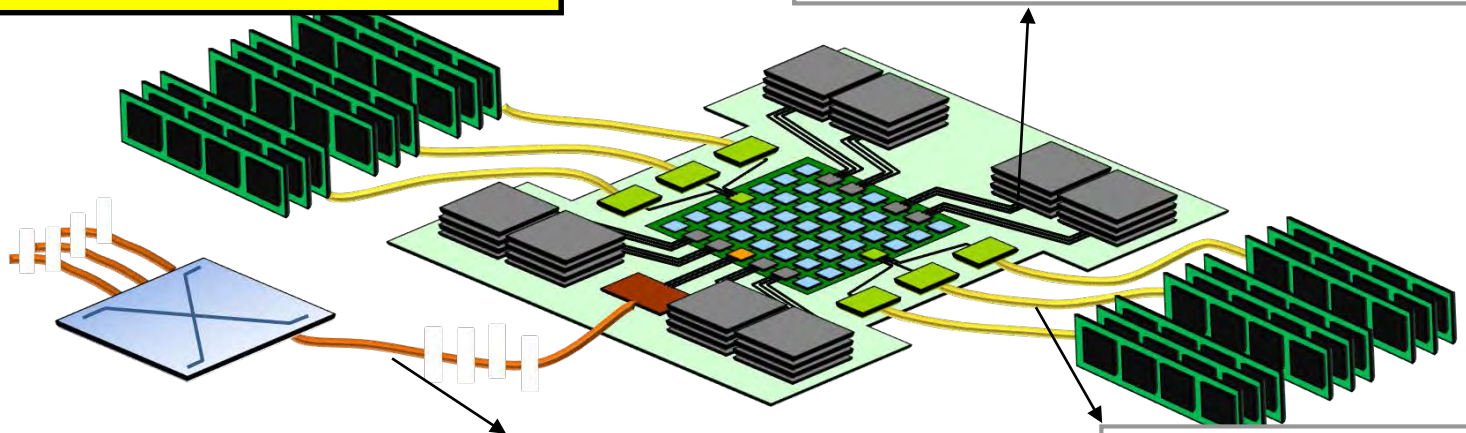
- Assume 15% of $ budget for interconnect:
  - 15% x $200M / **500 Pb/s** = 6 **¢/Gb/s**
  - Bi-directional links must thus be sold for ~10 ¢/Gb/s
    - Today: optical 10$/Gb/s
      electrical 0.1-1 $/Gb/s
- Assume 15% of power budget for interconnect:
  - 15% x 20MW / **125 Pb/s** = 24 **mW/Gb/s** = 24 **pJ/bit**
    = budget for communicating a bit end-to-end

→ 6 pJ/bit per hop
→ 4 pJ/bit for switching — today ~20 pJ/bit
→ 2 pJ/bit for transmission — today ~10 pJ/bit (elec)

# Exascale supercomputing node

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

Near memory bandwidth:
10-30 TF x 8bit x 0.5B/F = **40 - 120Tb/s**



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

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

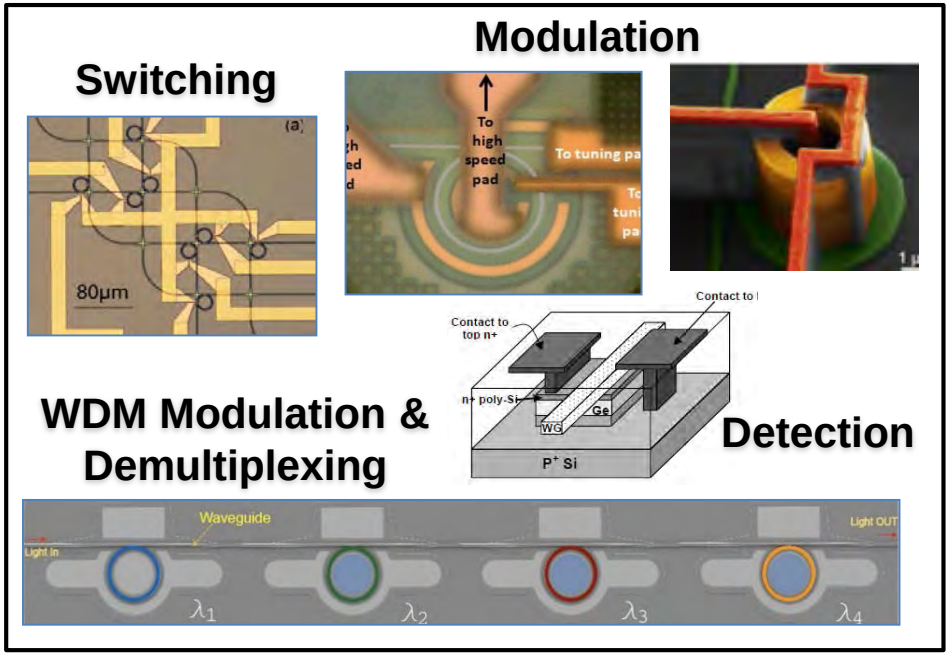
Interconnect bandwidth:
**0.01 B/F** → 0.8 – 2.4 Tb/s

- Consider 50k nodes
  - Total injection bandwidth ~ 100 Pb/s
    - 4 ZB per year at 1% utilization
  - Total cumulated unidirectional bandwidth: ~ 500 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



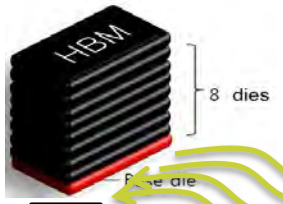

- Silicon wafer-scale CMOS
  - Integration, density scaling
  - CMOS fabrication tools
  - 2.5D and 3D platforms

S. Rumley et al. "Silicon Photonics for Exascale Systems", IEEE JLT 33 (4), 2015.

# Optically-Connected Memory Architecture

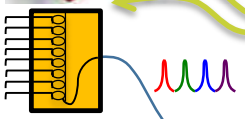
- Separate memory stacks & processors onto different SiPh interposers -- alleviate area limit
- Fabricate waveguides, switches & transceiver on the SiPh interposer -- active functionality

8GB/cube




25 wavelengths x
40 Gb/s/λ
= 1 Tb/s per fiber

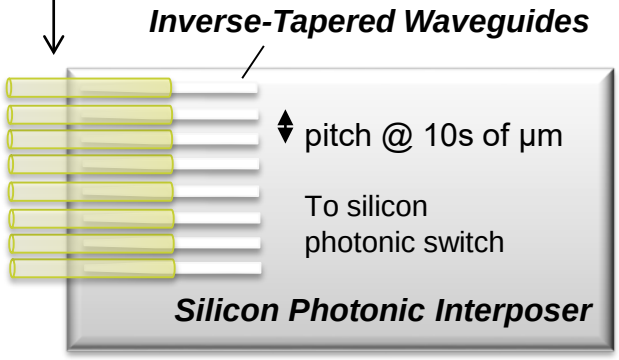
Fiber

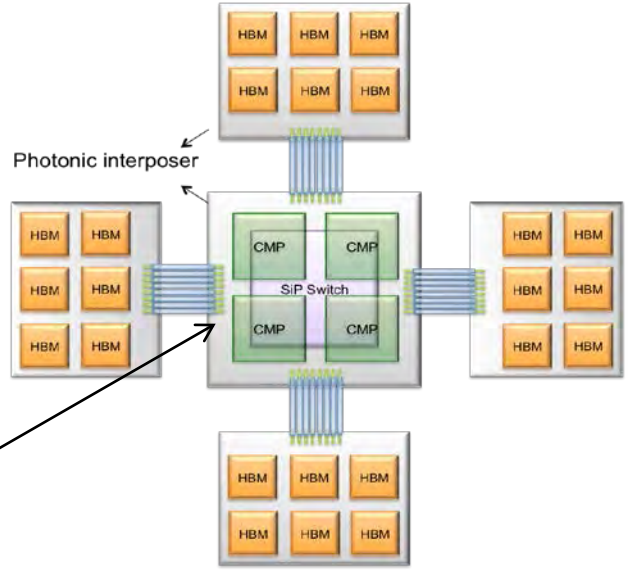


WDM transceiver &
wavelength selective switch

Each direction: 2 fibers x 1 Tb/s
≈ 256 GB/s (HBM Gen2)

High-Density Fiber
Coupling Array
(up to 72)




24 fibers per coupling assembly
= 6 cubes, 48 GB capacity, 1.5 TB/s

4 coupling assemblies
= 24 cubes, 192 GB capacity, 6 TB/s

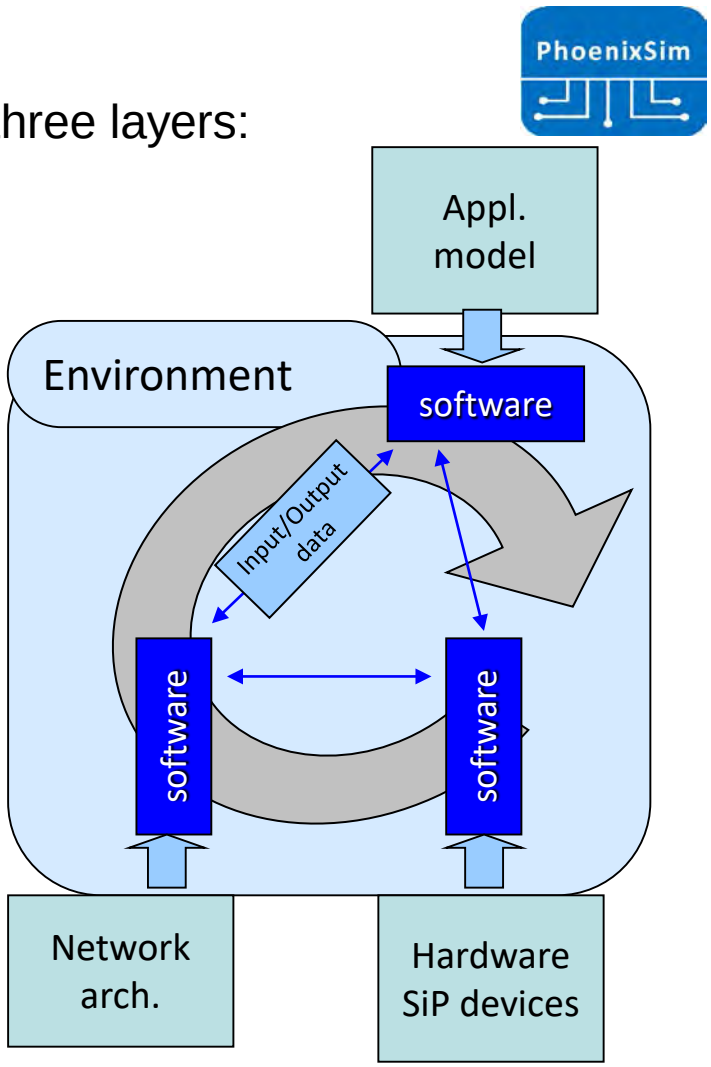
U.S. DEPARTMENT OF ENERGY

SBIR Collaborative Project:
Photonic Memory Controller Module

FREEDOM PHOTONICS
PLC CONNECTIONS
ANALOG PHOTONICS

# PhoenixSim: Integrated Multi-Level Modeling and Design Environment

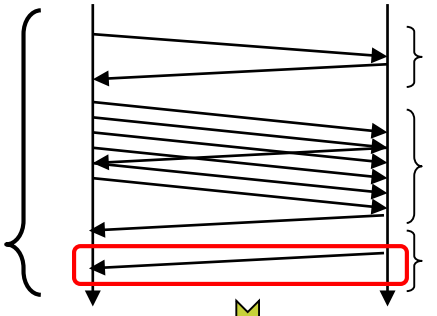
- Integrated design/modeling environment across three layers:
  - **Application** IO primitives
    - Copy memory array to remote location
    - Send, multicast, broadcast messages
    - Thread synchronization (e.g. barrier)
  - **Network** architecture and protocols
    - Link locking mechanisms (frame detection)
    - Network topology (routing)
    - Arbitration of shared buses, switches
  - **Si Photonic Hardware** implementations
    - Silicon photonics modulators, switches
- Complete “toolbox” of models at each layer
  - Ensure interoperability among models
  - **Cross-layer co-optimization is Key**



# Multi-layer environment

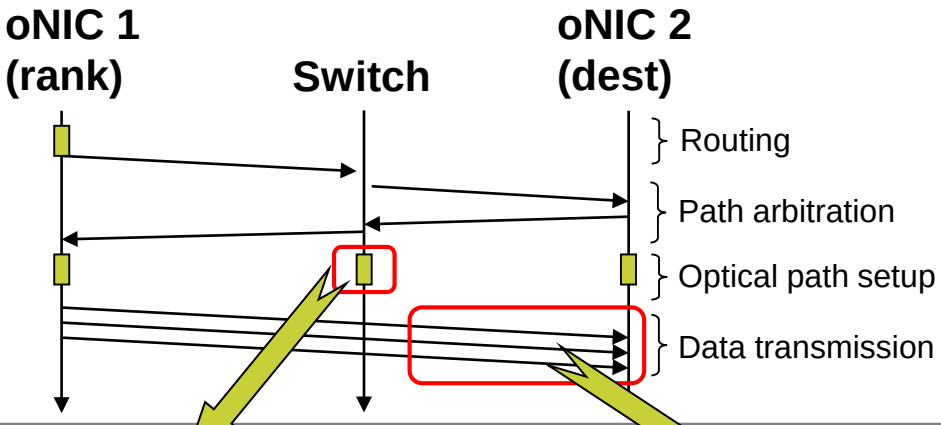
Thread ID

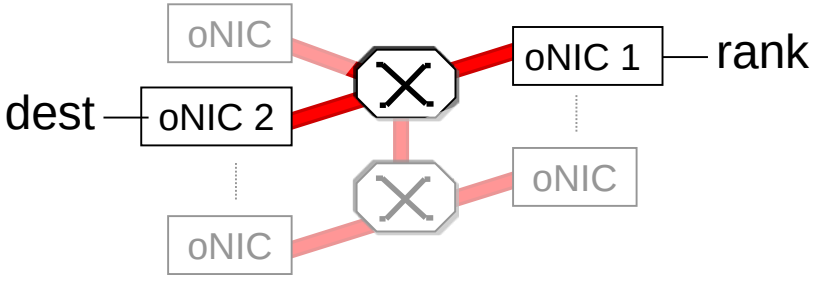
```
void work_in_parallel(int rank) {
  int[] array = calculate_local_array(rank);
  int dest = determine_next_dest(array);

  copy_array_remote(array, dest, address);
}
```



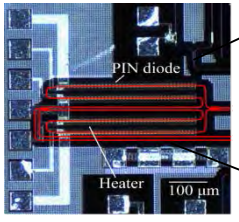

Application

Network

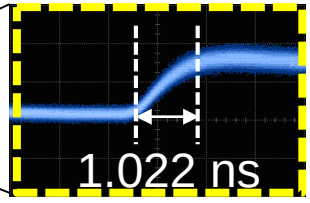


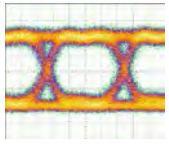

Hardware

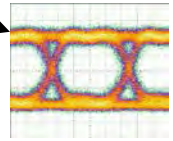



SiP Switch




SiP WDM
Demux



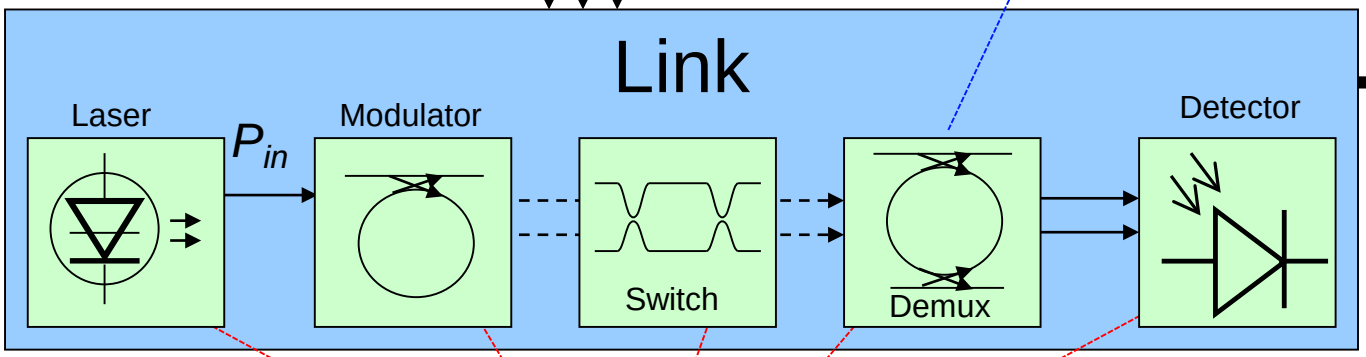


# Methodology - Abstraction of Physical Devices

Input laser power,
Number of wavelengths,
Modulation rate

Param

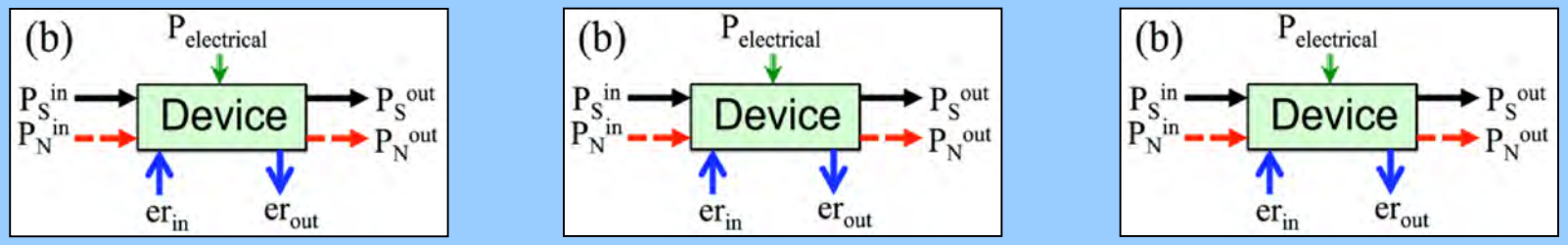
Explore both device and link parameters to ***optimize*** bandwidth or energy efficiency




*Bandwidth*

*Energy per bit*

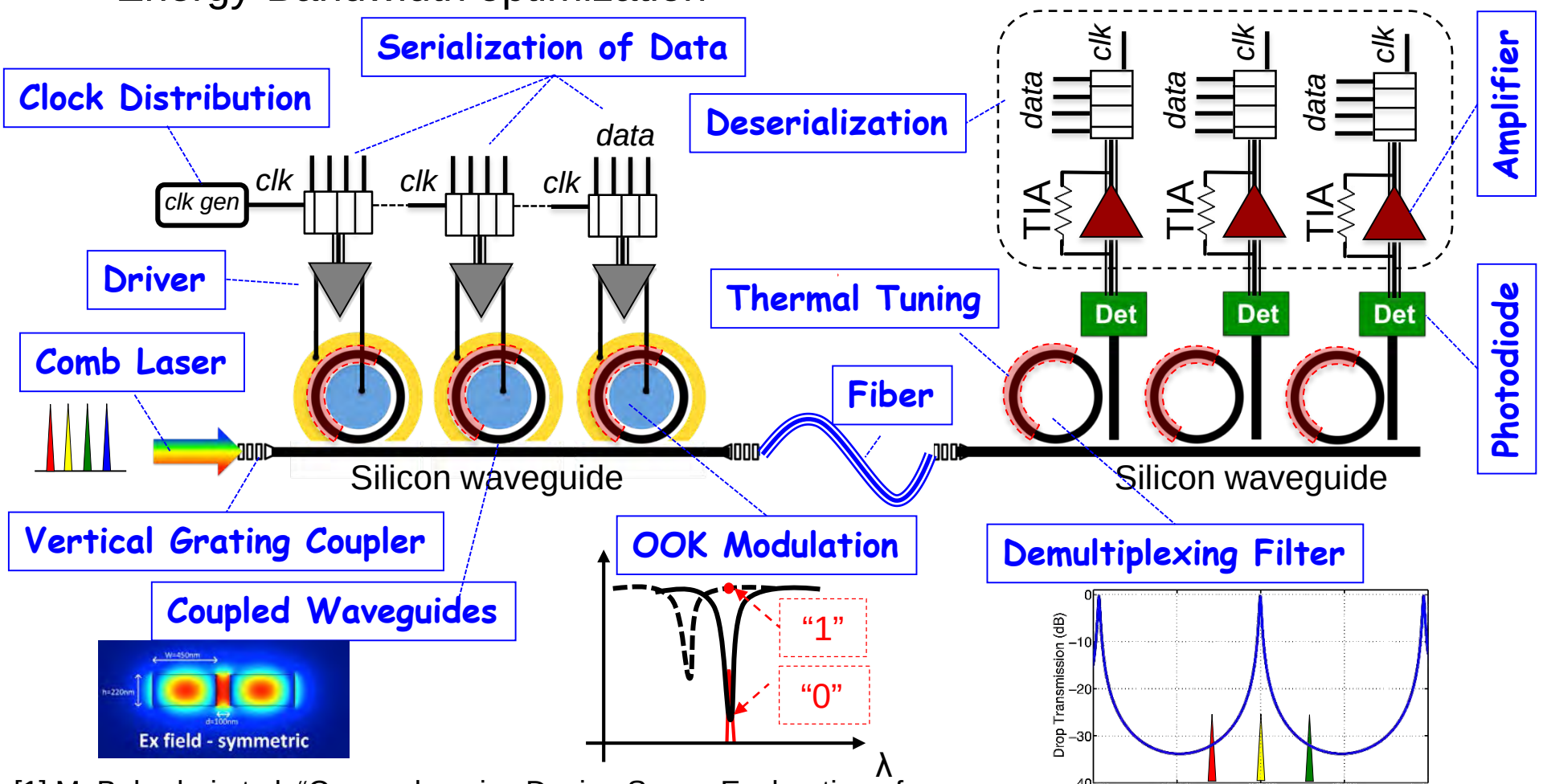
Abstract Physical Models




*Model the flow and characteristics of optical signal along the link*

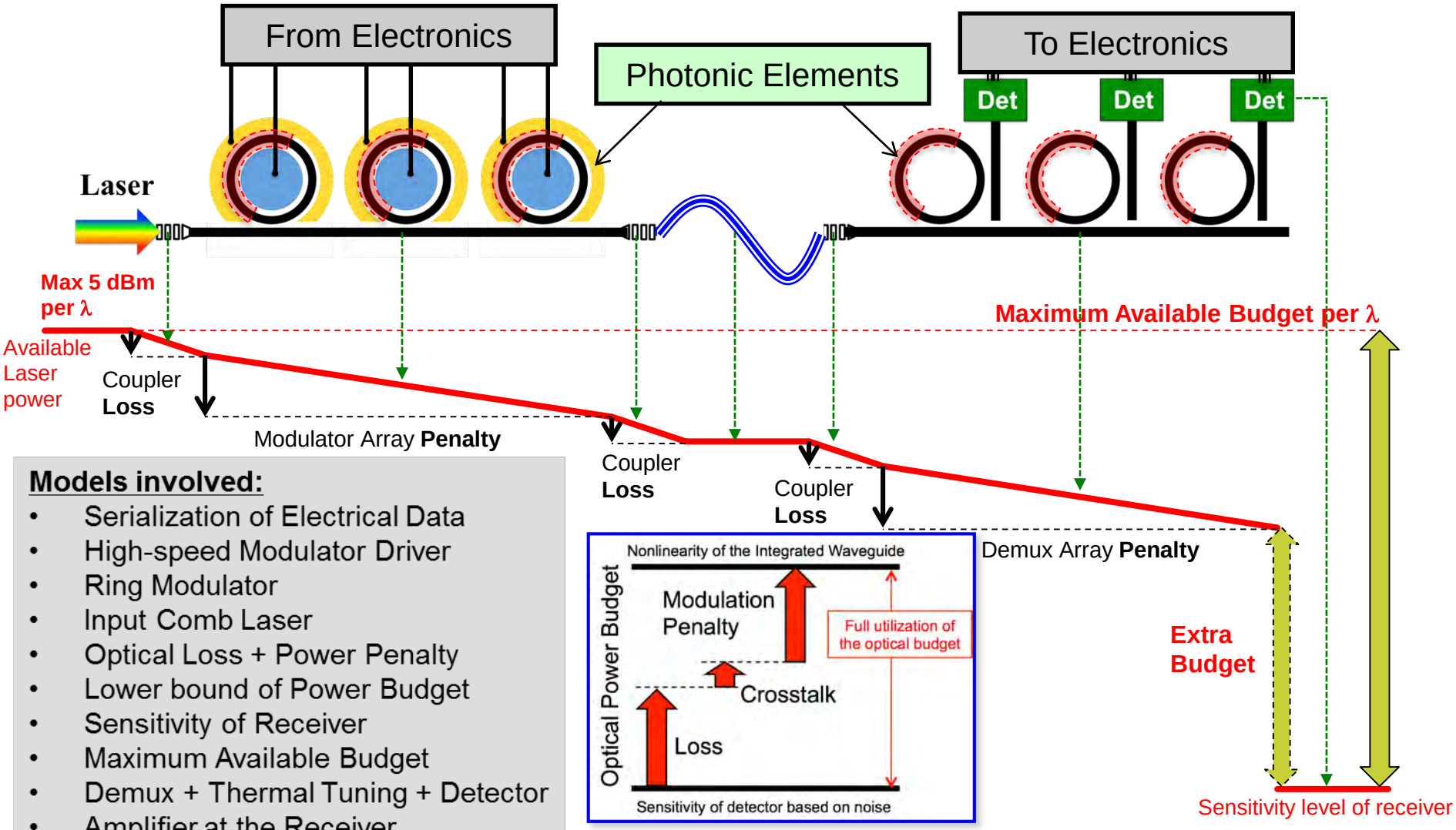
# Physical - Silicon Photonic Link Design

- 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.

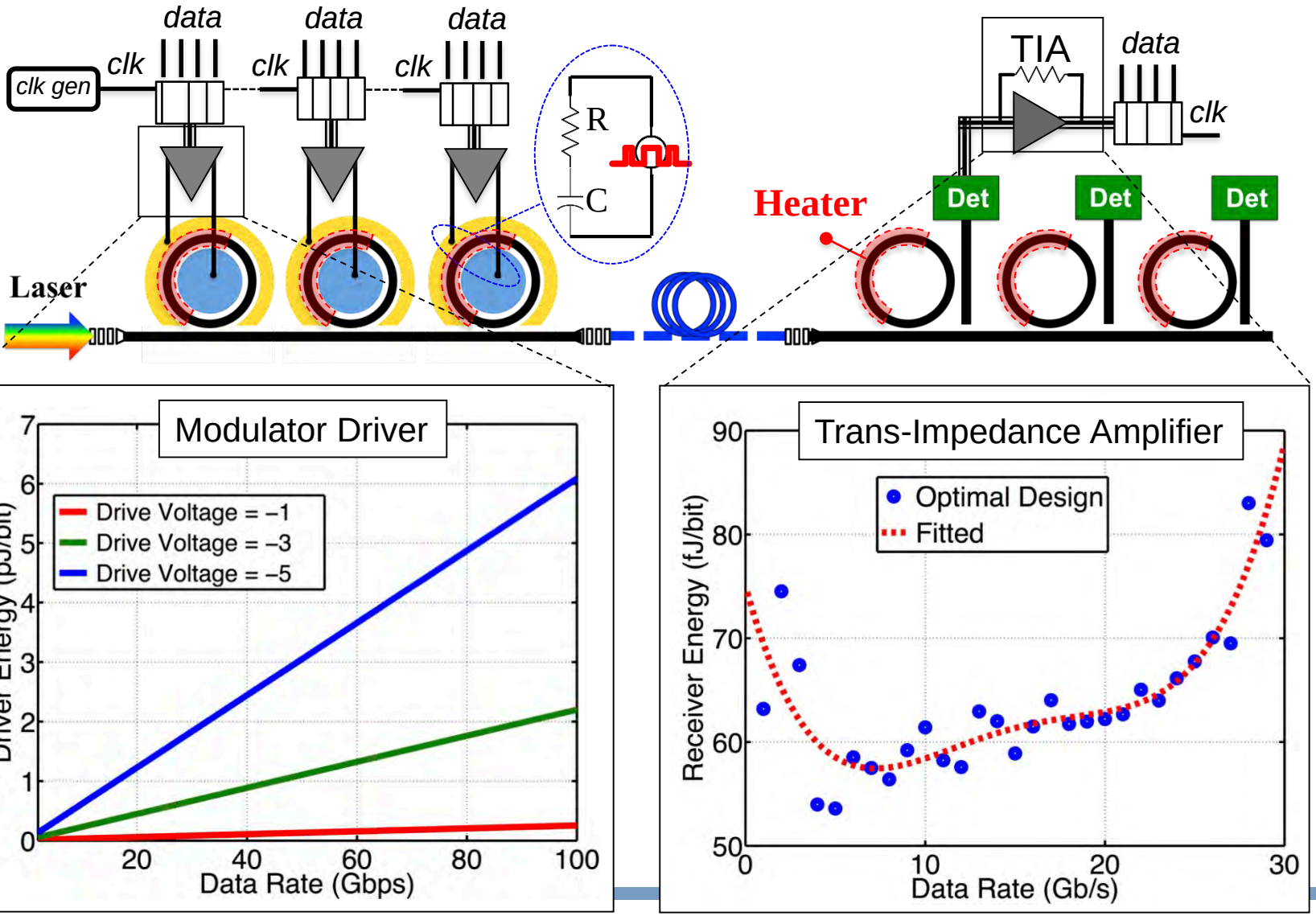
# Utilization of Optical Power Budget



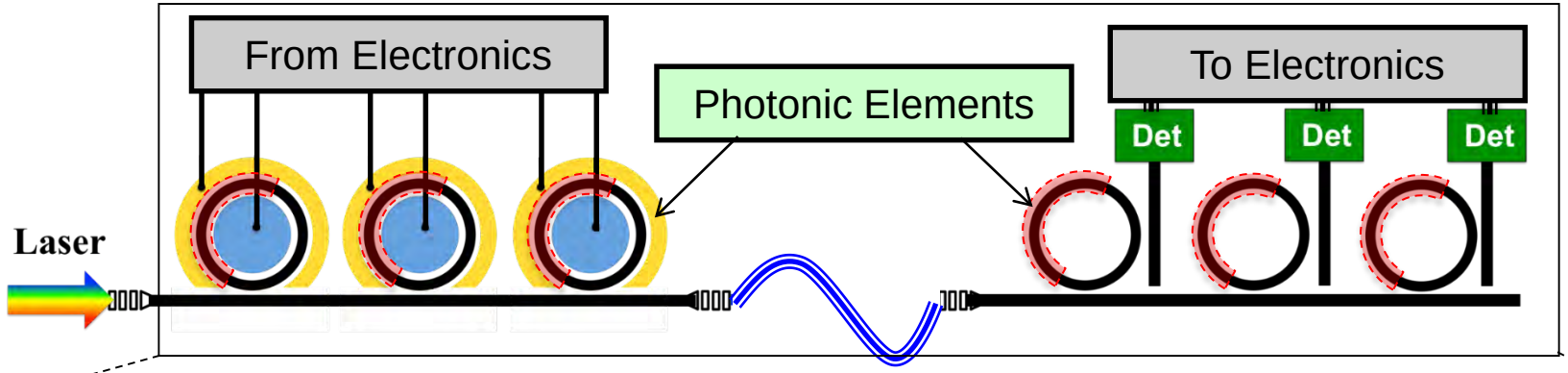

**Models involved:**

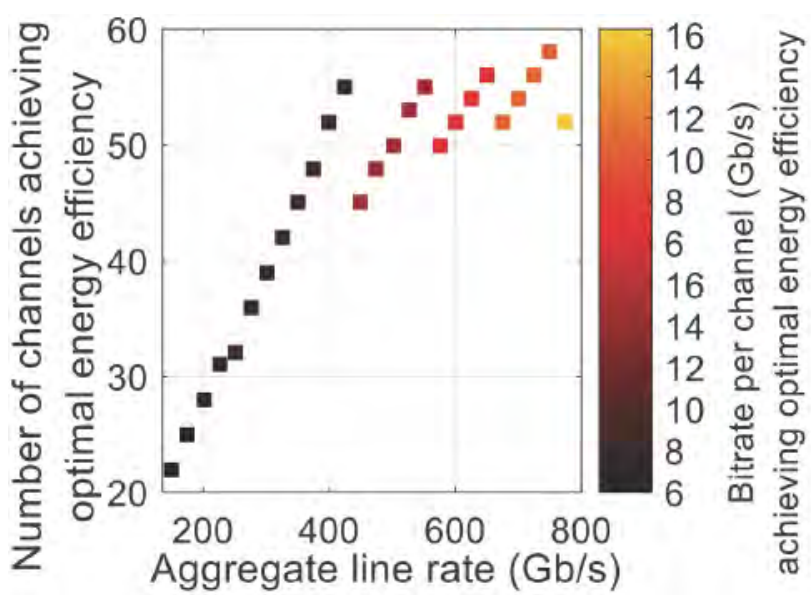
- Serialization of Electrical Data
- High-speed Modulator Driver
- Ring Modulator
- Input Comb Laser
- Optical Loss + Power Penalty
- Lower bound of Power Budget
- Sensitivity of Receiver
- Maximum Available Budget
- Demux + Thermal Tuning + Detector
- Amplifier at the Receiver
- Deserialization

# Considering the electronics

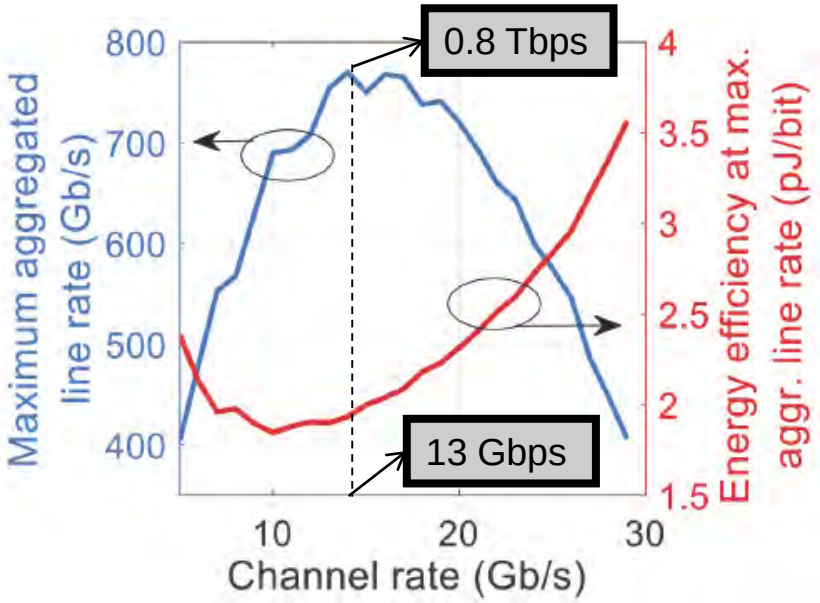


# All-Parameter Optimization: Max Bandwidth Design



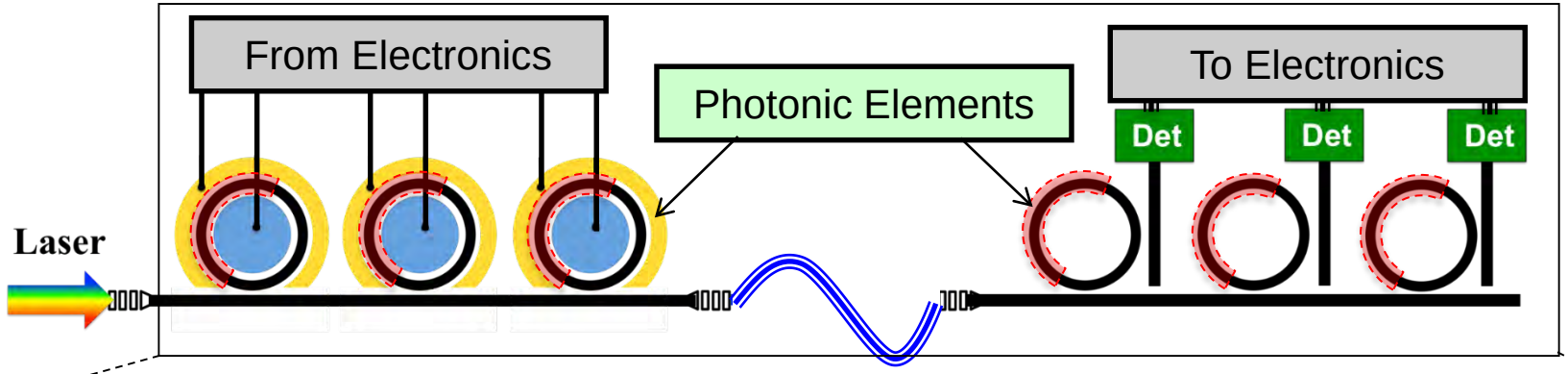



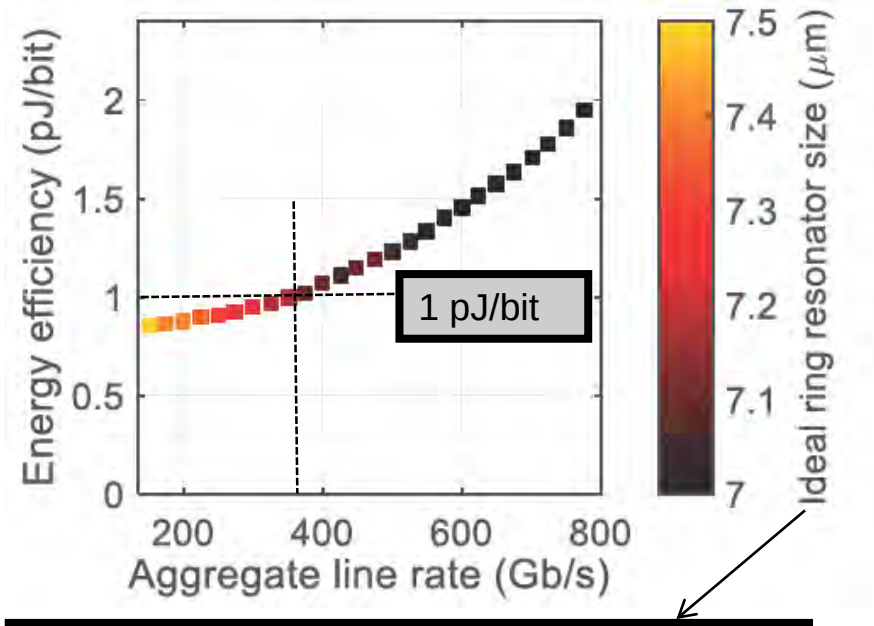

Optimal choice of Number of channels and Bitrate of each channel for Max Bandwidth design



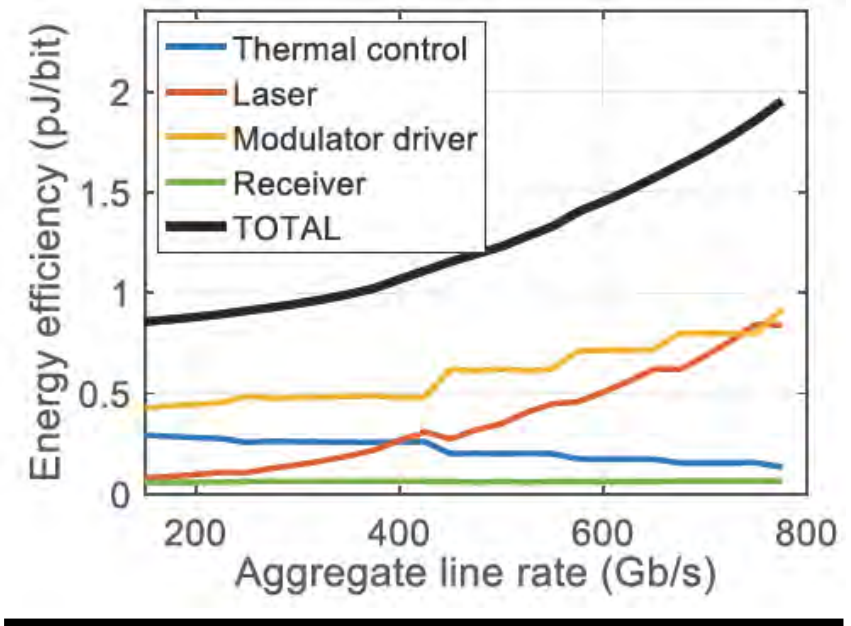

Full utilization of link for max bandwidth density

# All-Parameter Optimization: Min Energy Design



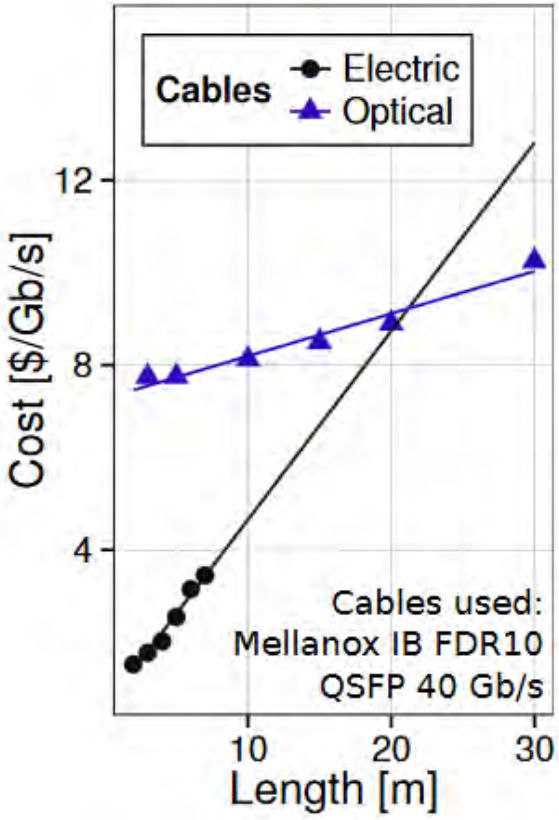




Optimal design based on physical dimensions

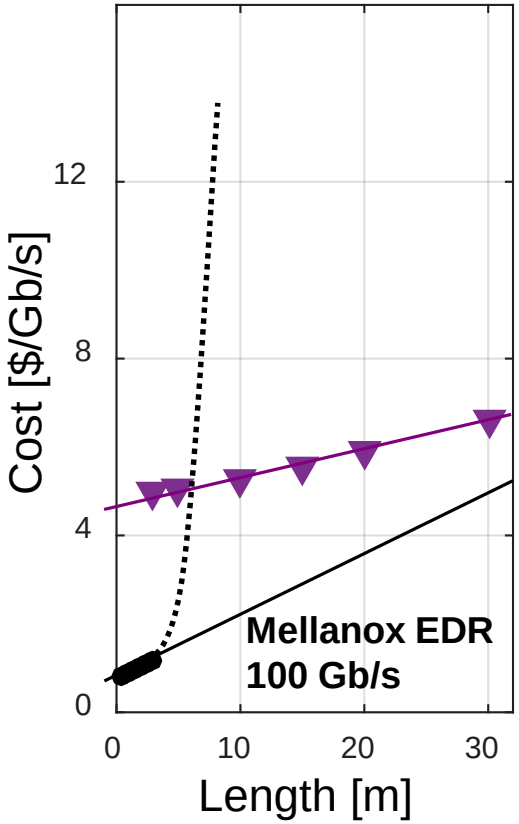



Breakdown of consumption for given bandwidth

# Cost per bandwidth – declining but slowly




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

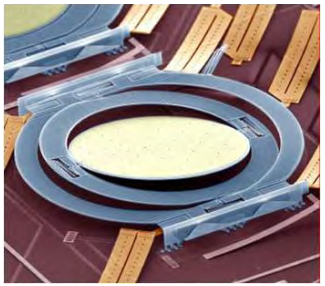



[may 2017]

- 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

# Beyond the Link: Photonic Switching

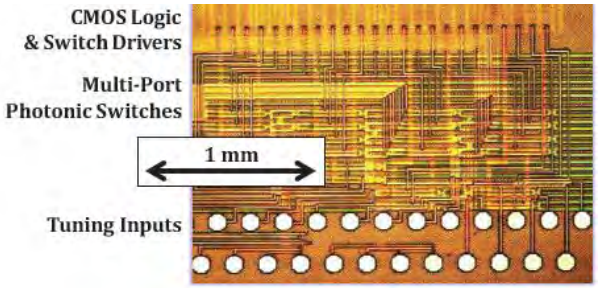
## ➢ MEMS-based Switches



[Lucent Technologies' Lambda Router]

- Free-space propagation
- High actuating voltage
- Broadband
- Low loss/low crosstalk
- Bulky
- Slow
- Scalable
- Cost ultimately limit by installation & calibration

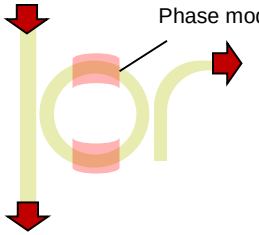
## ➢ SiP-based PIC Switches



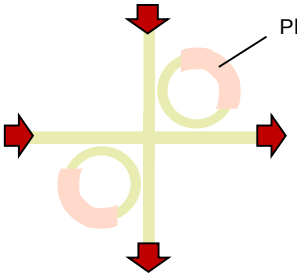

[Benjamin Lee, OFC 2013, PDP paper]

- Planar lightwave circuits
- Broadband/Wavelength Selective
- High integration (small footprint)
- Fast (E-O effect)
- Lossy/relatively high crosstalk
- Rather Scalable
- CMOS/PIC monolithic integration
- Cost can be low benefiting from mature CMOS industry

# MRR Element Model

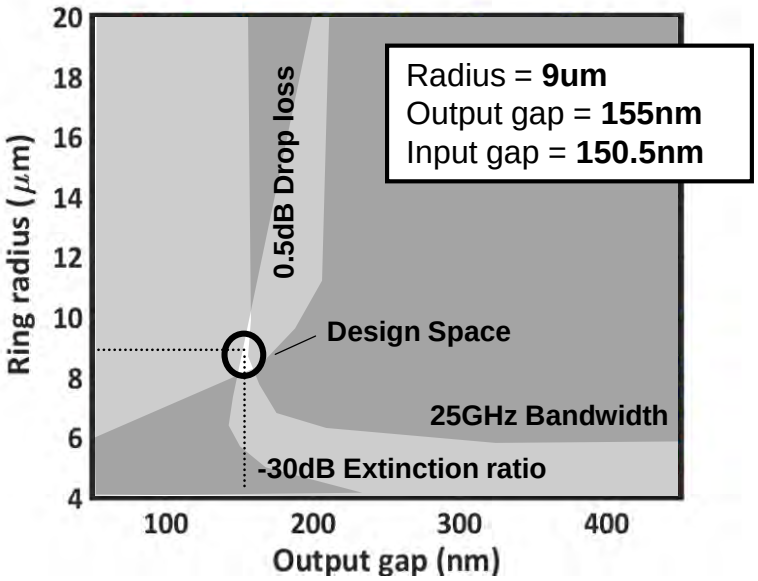



**: Critical coupling**



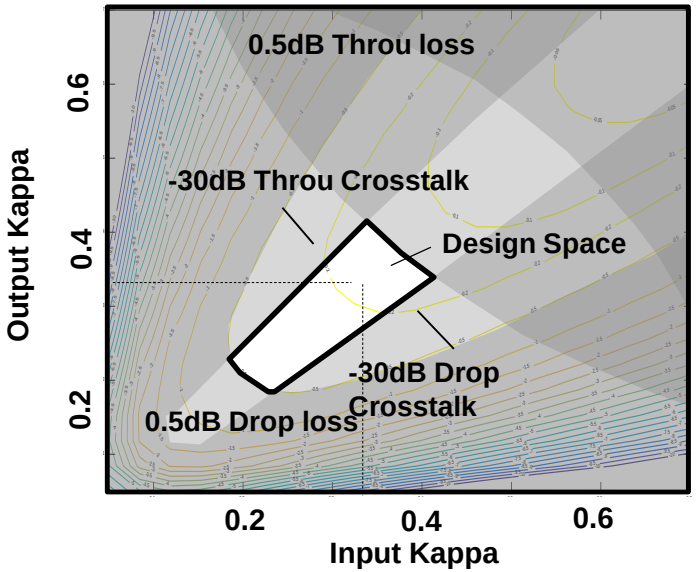

**: Balance between Cross and Bar**

**1x2 Add-drop MRR**




- Drop loss: 0.35 dB
- Thru loss: 0.1 dB; Xtalk : -29.3 dB

**2x2 MRR SE**



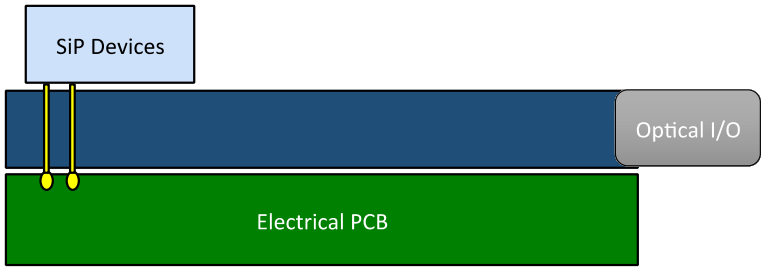

- Drop loss: 0.17 dB; Xtalk: -32.4 dB
- Thru loss: 0.19 dB; Xtalk : -31.7 dB

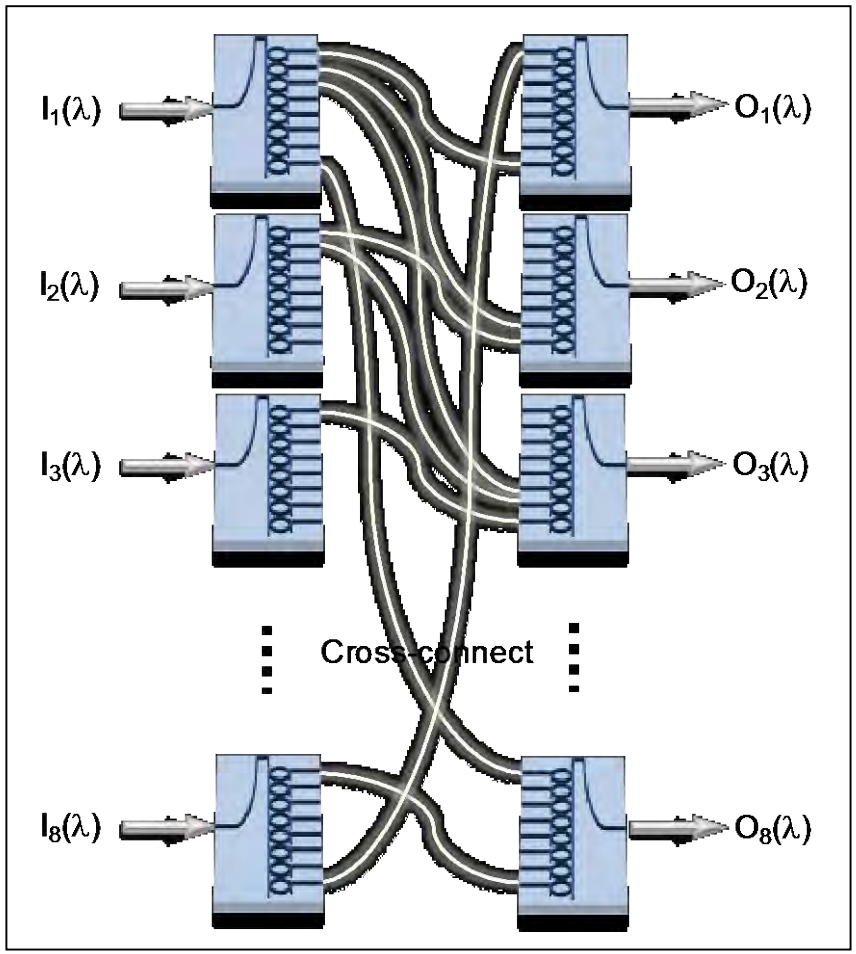
*S. Rumley, et al., "PhoenixSim: Crosslayer design and modeling of silicon photonic interconnects," AISTECS workshop, ACM, 2016.*

# Transitioning to Novel Modular Architectures…

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

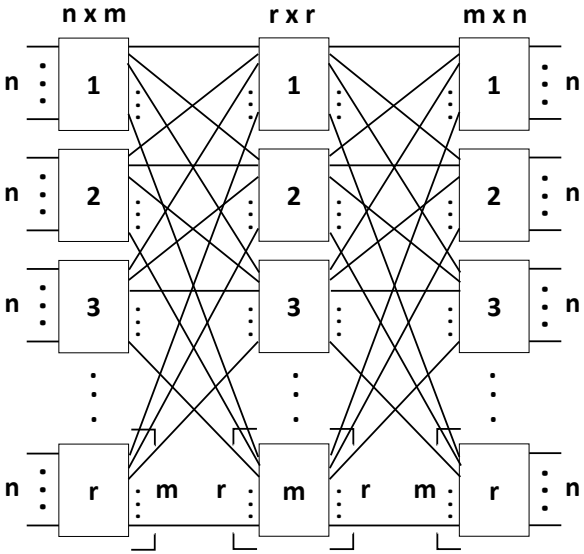
ed integration method

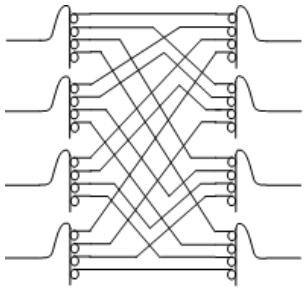



*[Dessislava Nikolova*, David M. Calhoun*, Yang Liu, Sébastien Rumley, Ari Novack, Tom Baehr-Jones, Michael Hochberg, Keren Bergman , Modular architecture for fully non-blocking silicon photonic switch fabric , Nature Microsystems & Nanoengineering **3** (1607) (Jan 2017).]*

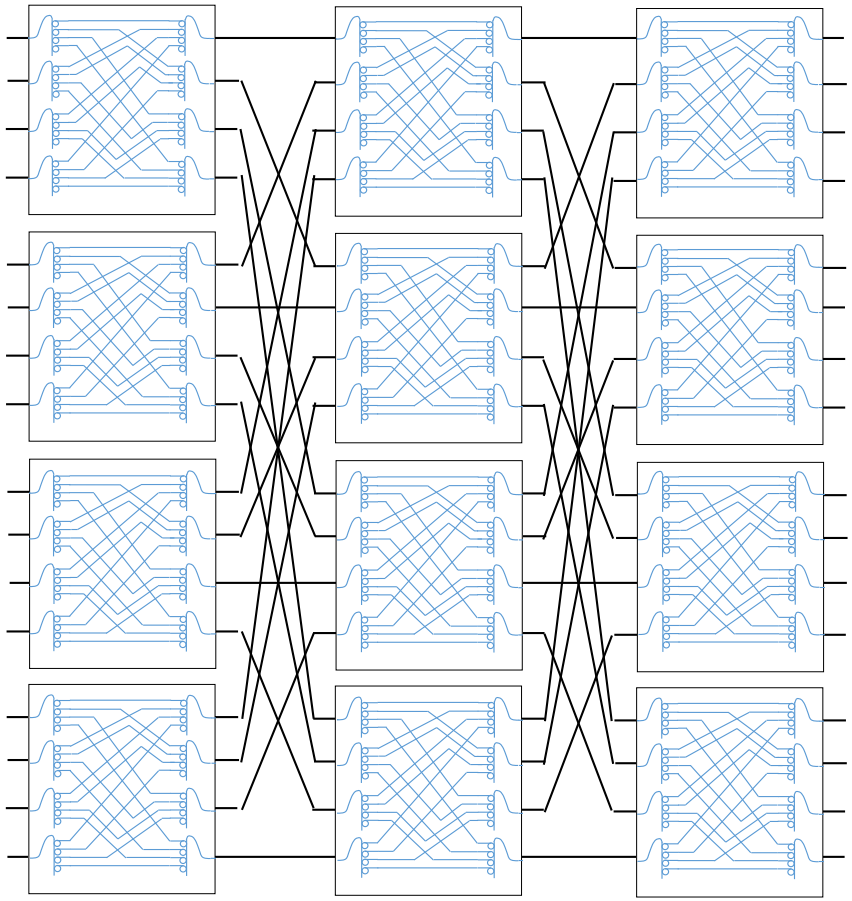
# Clos-of-Switch-and-Select Architecture



4 x 4

16 x 16

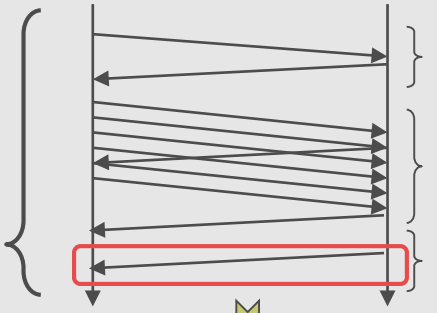


- Offering a suitable balance that keeps the number of stages to the modest value of three while largely reducing the required number of MRRs

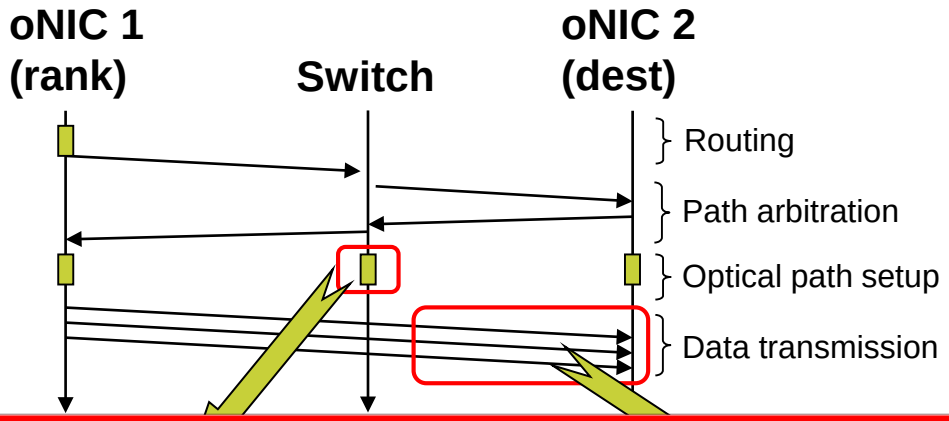
# Network layer

Thread ID

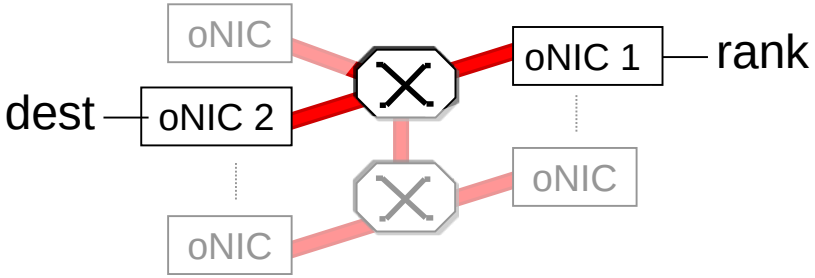
```
void work_in_parallel(int rank) {
  int[] array = calculate_local_array(rank);
  int dest = determine_next_dest(array);

  copy_array_remote(array, dest, address);
}
```



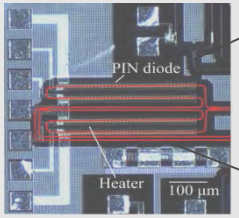

Application



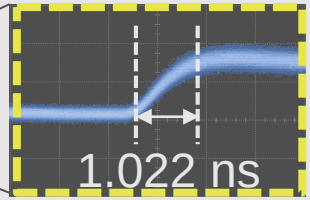

Network



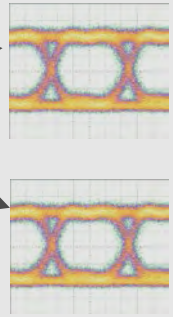

Hardware




SiP Switch

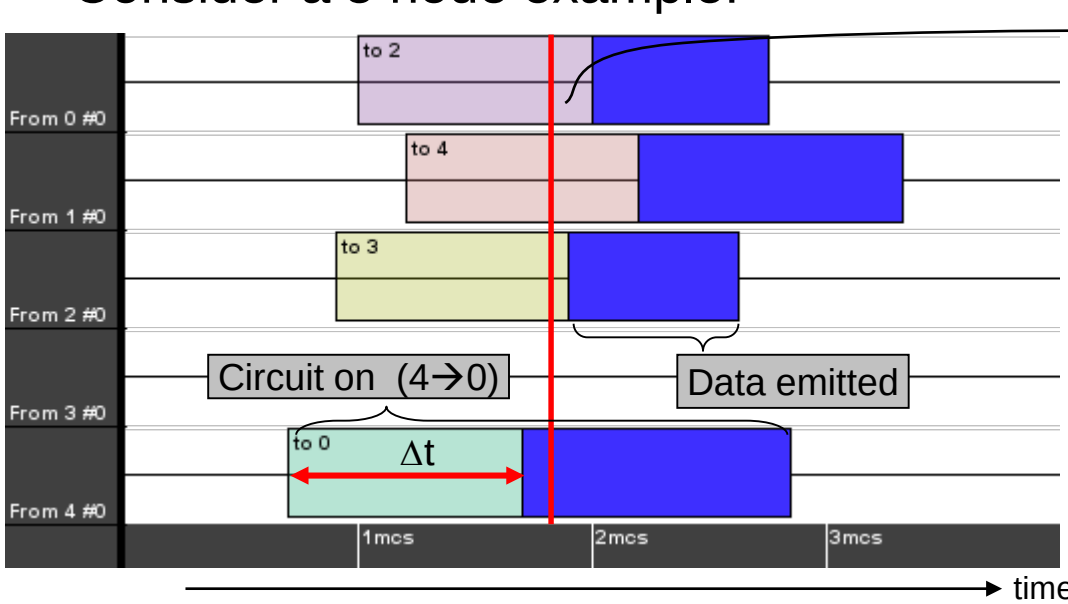


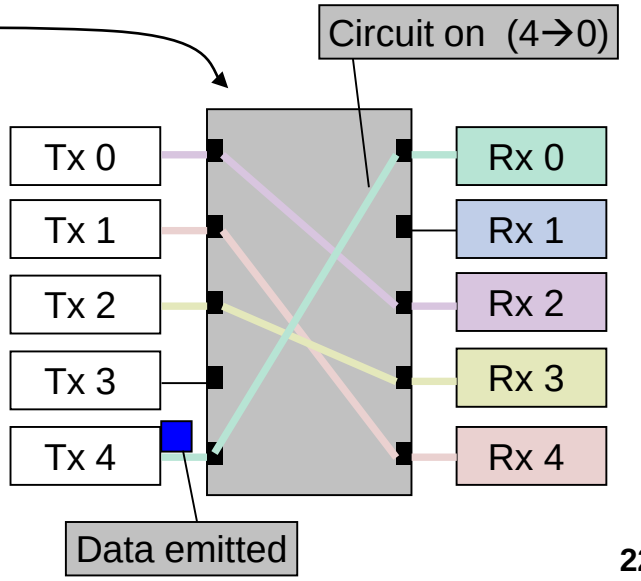

SiP WDM
Demux



# Network layer

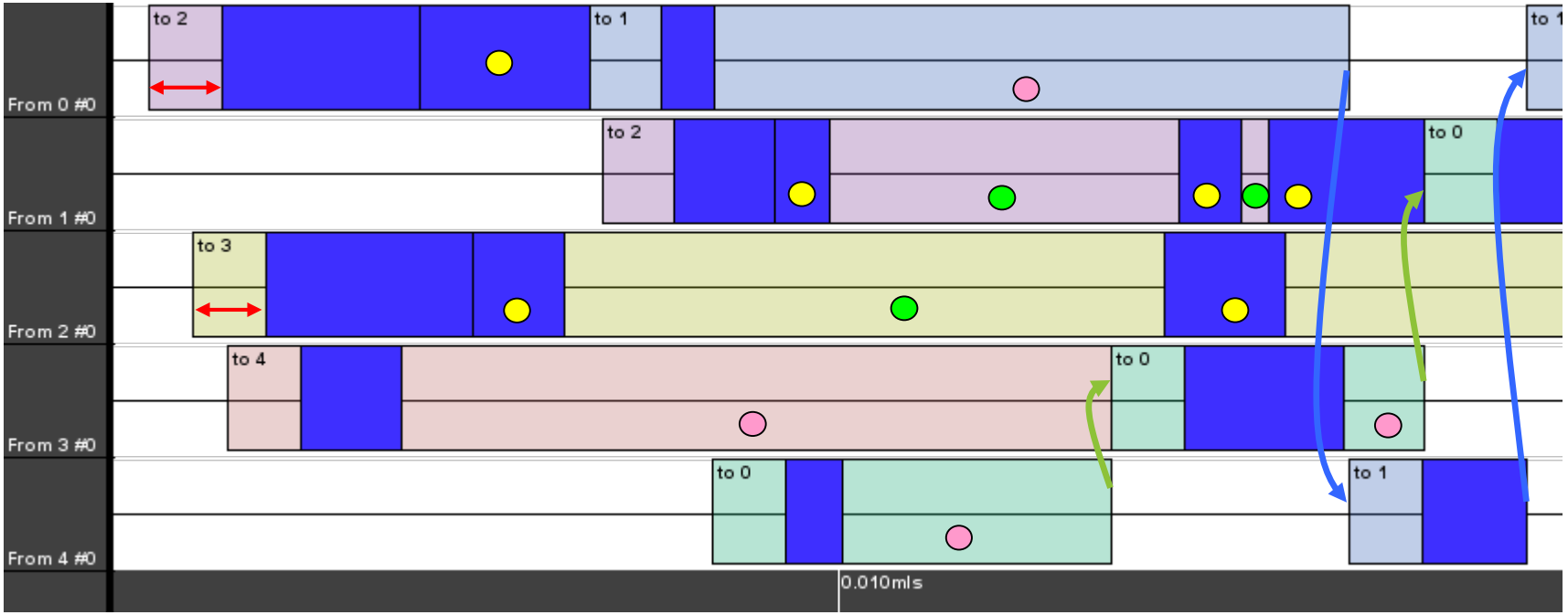
- Implemented circuit level arbitration
  - Data or packets emitted by application layer delayed while circuit is set up
    - Circuit setup is assumed to take a predefined time $\Delta t$
  - Includes prediction mechanism:
    - Keep circuit on if high probability of being reused
    - Prefetch next circuit if next destination highly probable
  - Supposes a fully non-blocking physical layer
    - A circuit can always be established as long as input and output ports are free
- Consider a 5 node example:



# Circuit arbitration - visualization

- Arbitration at play under random packet arrivals (30% load)
  - Correlated destinations: packet goes to next index with prob 50%




Circuit setup time

Dependency on dest 0

Dependency on dest 1

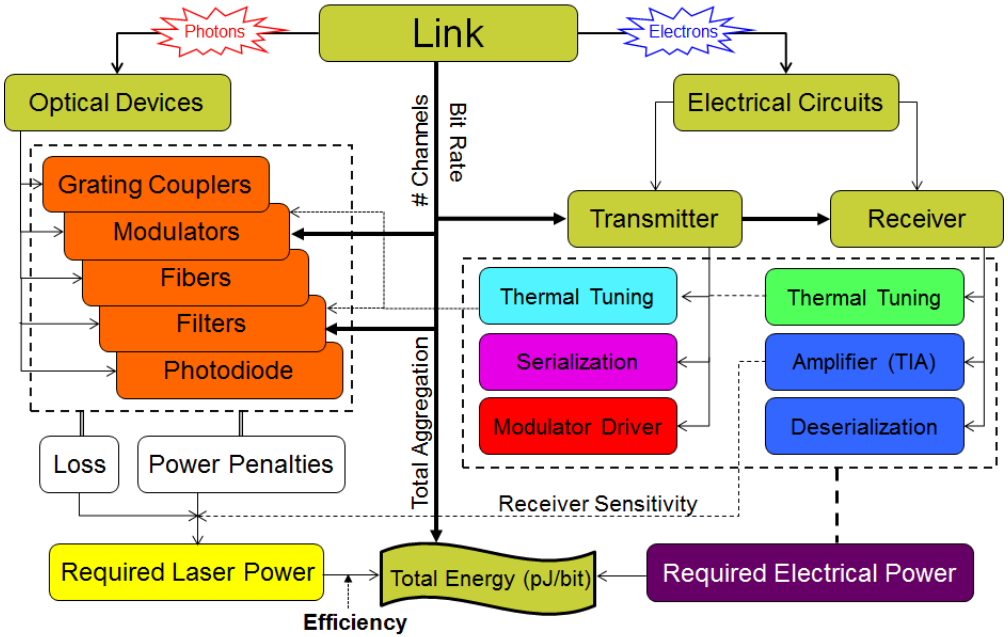
Instances of successful circuit reuse

Instances of circuit hold “for nothing”

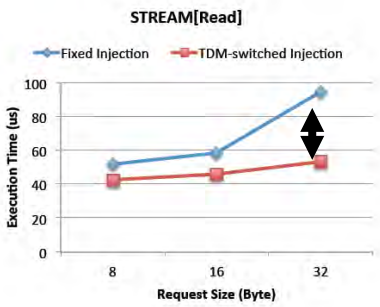
Instance of circuit rightfully hold

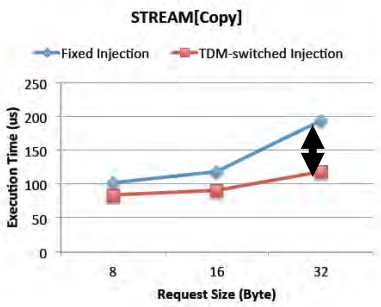
# Photonic Interconnected Memory - ModSim

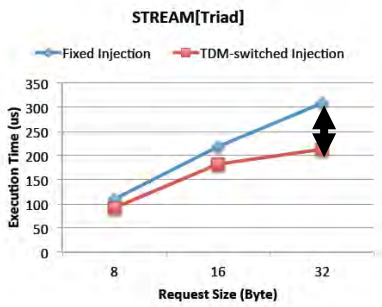
- SST's event-based simulation allows accurate/tractable system performance.
- Efficacy of different interconnect topologies evaluated, user-defined system components
- Performance of memory reads, writes, etc. simulated for performance evaluation
- Simulation results from optically-connected memories evaluated against conventional busses and electrical networks.







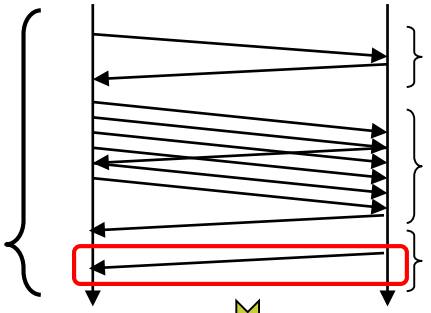
X 1.5-2 Execution time for low radix nanosecond switching

# Application layer

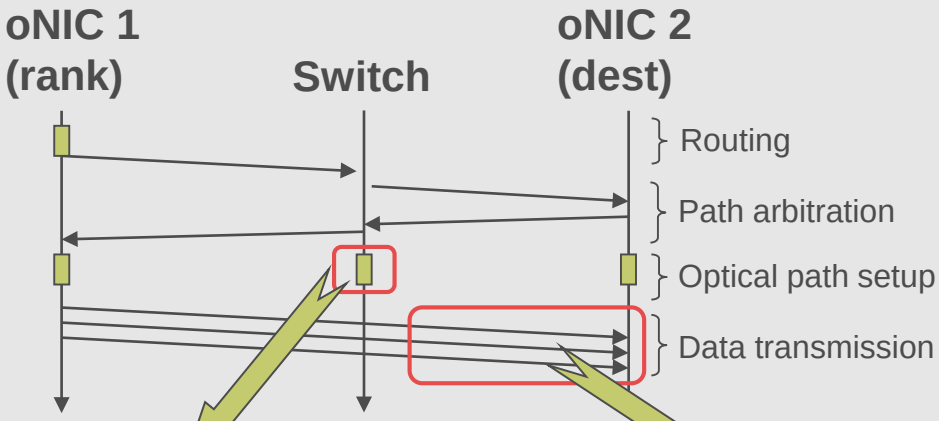
Application

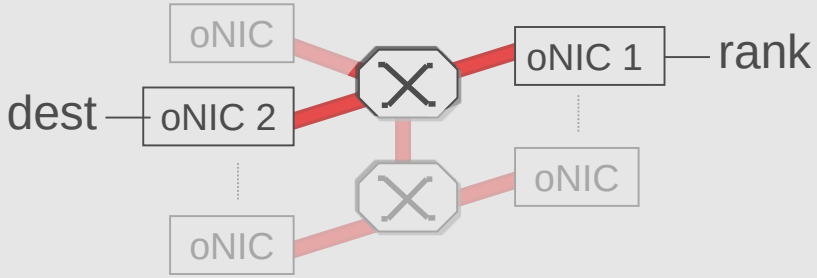
Thread ID

```
void work_in_parallel(int rank) {
  int[] array = calculate_local_array(rank);
  int dest = determine_next_dest(array);

  copy_array_remote(array, dest, address);
}
```

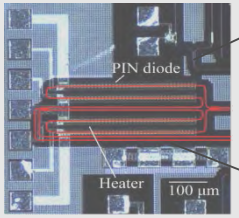



Network

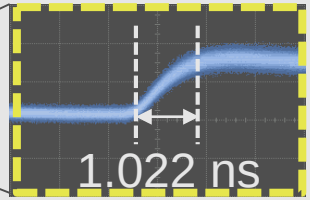


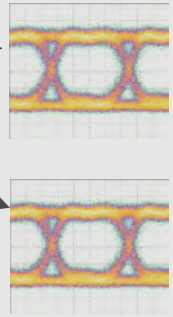

Hardware




SiP Switch




SiP WDM
Demux



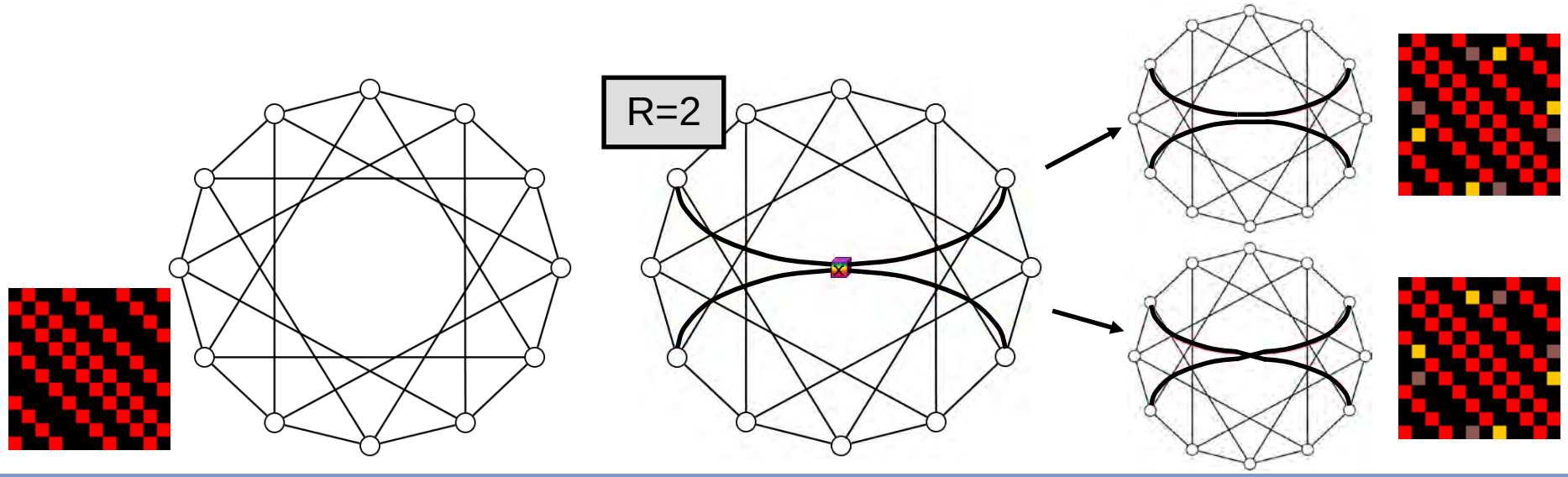
# Hybrid switching interconnects

- Do NOT use optical switches ***in place*** of packet routers!

→ Use optical switches ***in addition*** to packet routers

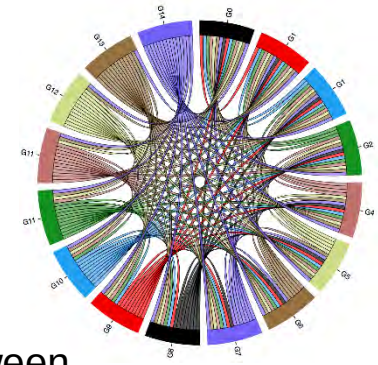
  - Coarse bandwidth steering across network clients: **optical switching**
  - Fine (per packet) bandwidth allocation: **packet routing**

- Bandwidth steering: equivalent to connectivity "re-wiring"
  - No need for large number of ports R
    - Allow for cheap, (soon) easy to fabricate silicon photonics switches



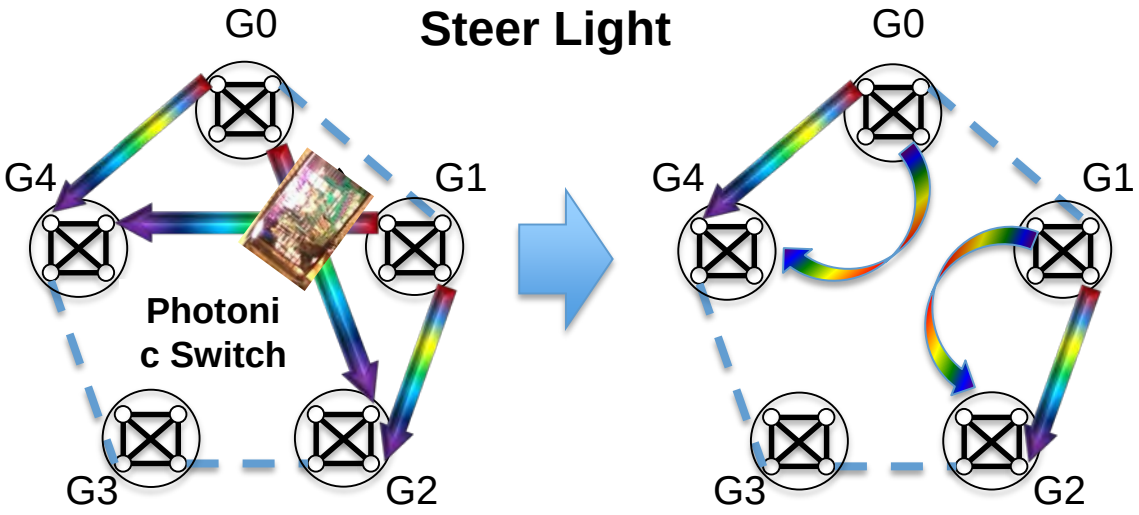
# Flexfly: A Reconfigurable Dragonfly

- **Incorporates photonic switching at inter-group level**
- **Reconfigure topology towards application traffic**

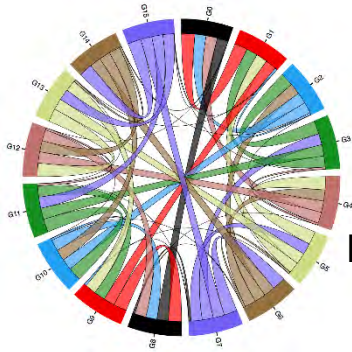


**Example:**
**Assume ±1 traffic pattern**

**Trading:**
**G0 adds a -1 link**
**G1 adds a +1 link**

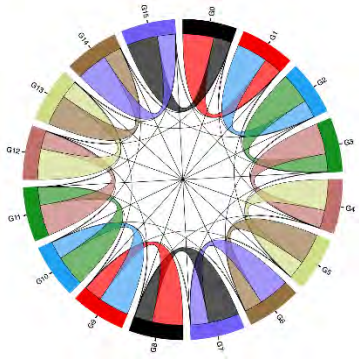



Anything in between



All to all

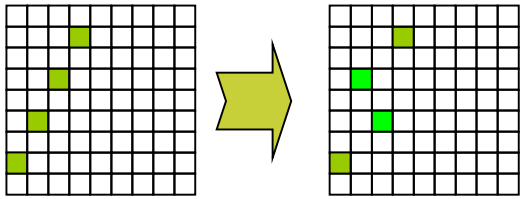
Neighbor intensive



# Flexfly construction

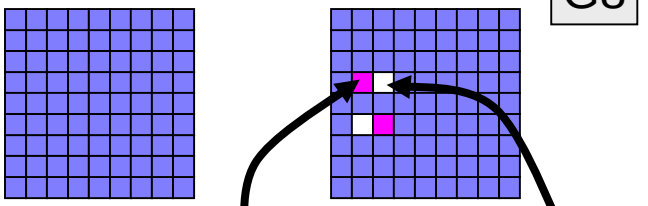
Before running parallel application:

4) Configure each switch to steer bandwidth



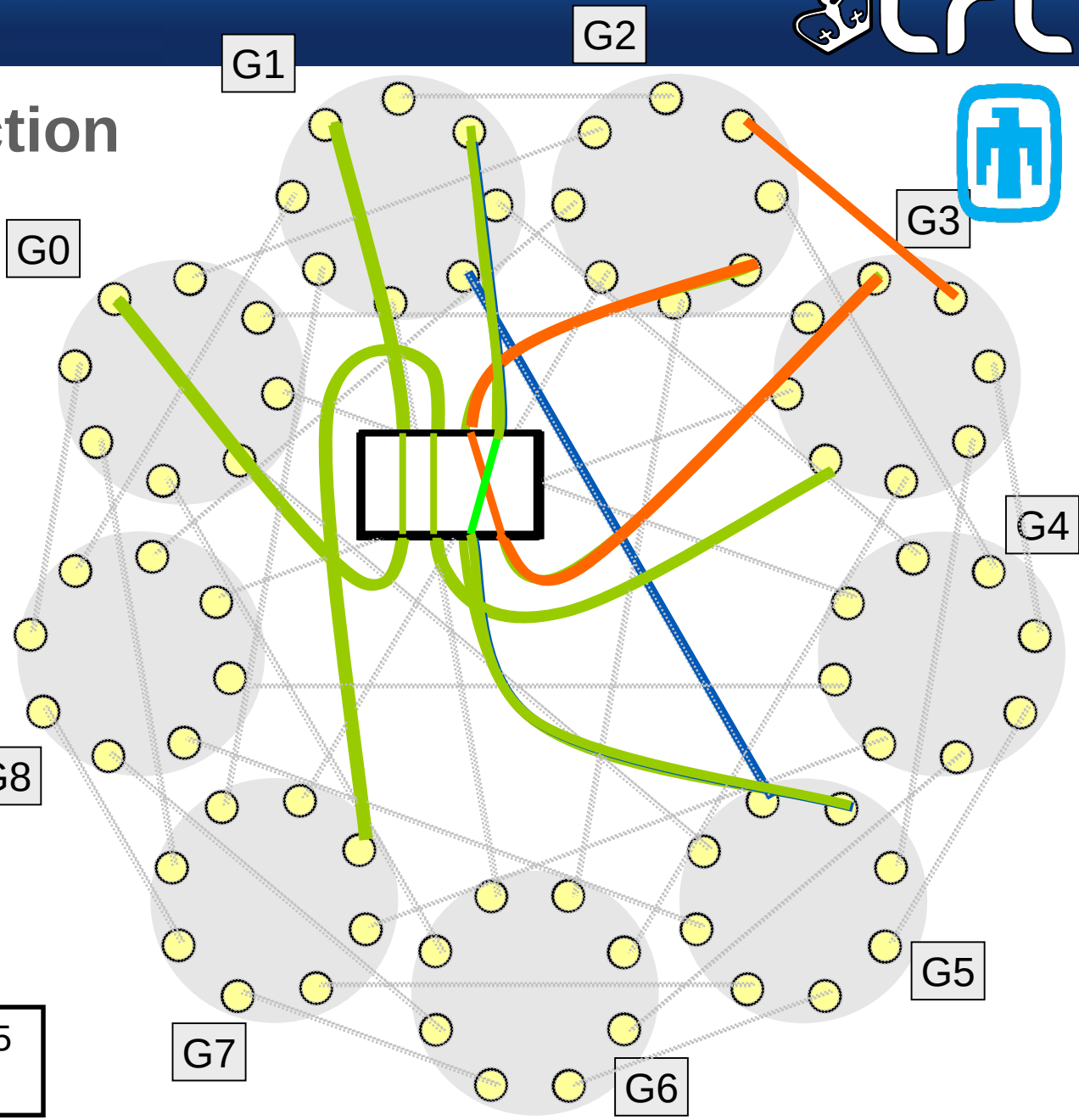
Original group-to group bandwidth distribution

New bandwidth distribution

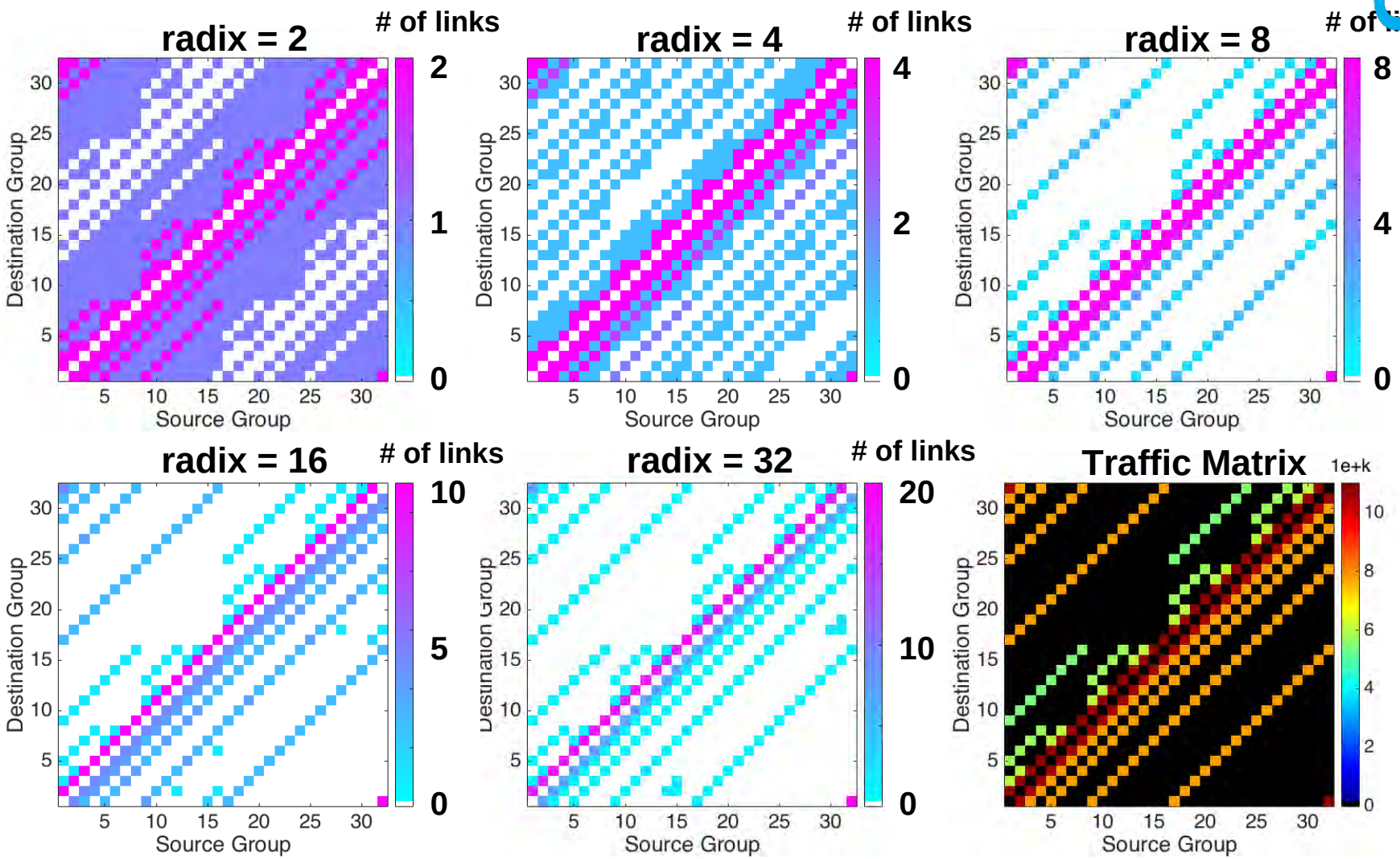


1 – 5 enjoys doubled BW

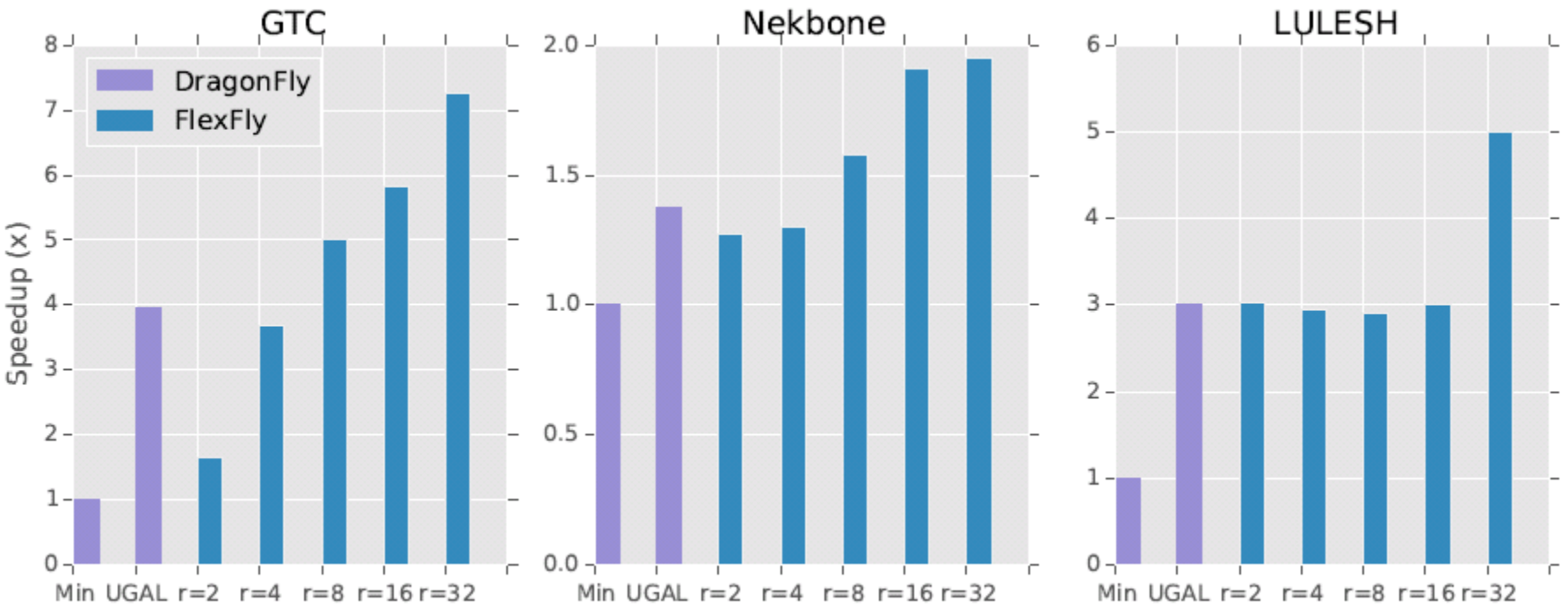
BW of 2 – 5 “stolen”



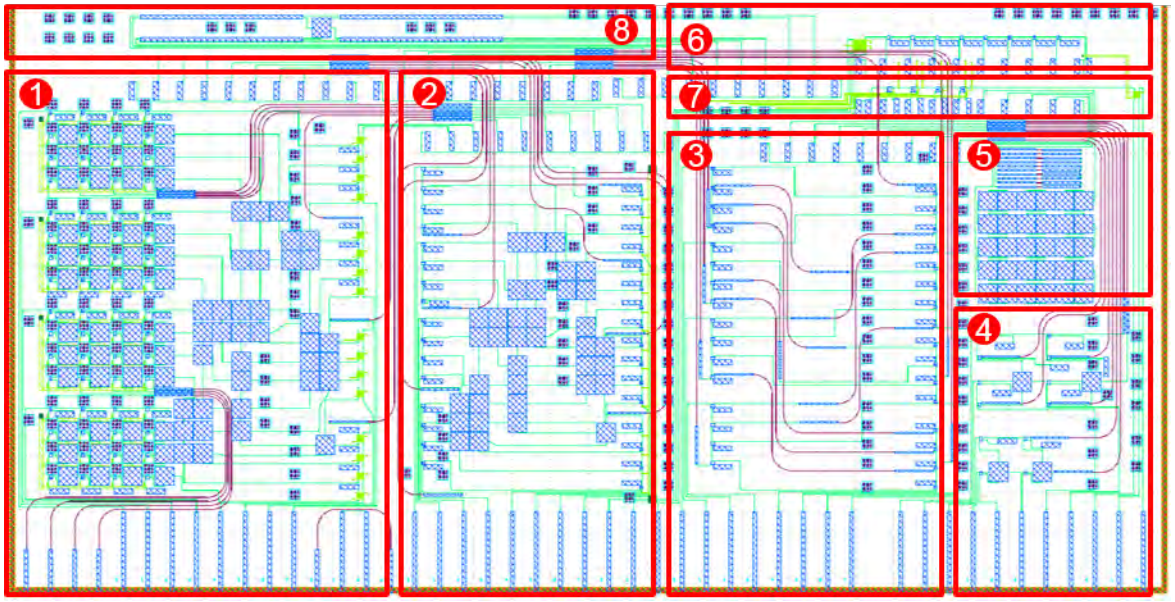
# Adapting topology for GTC application

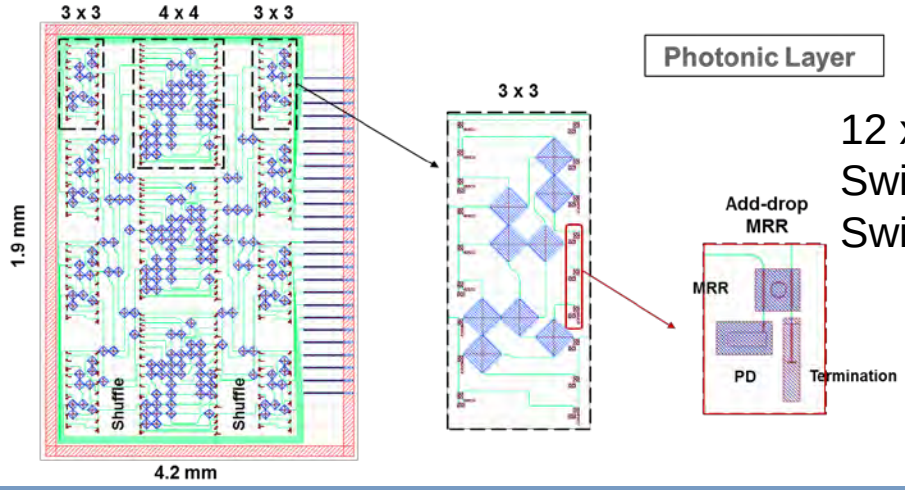


# Flexfly – simulated performance



# Implementation: AIM SiP Tapeout Run

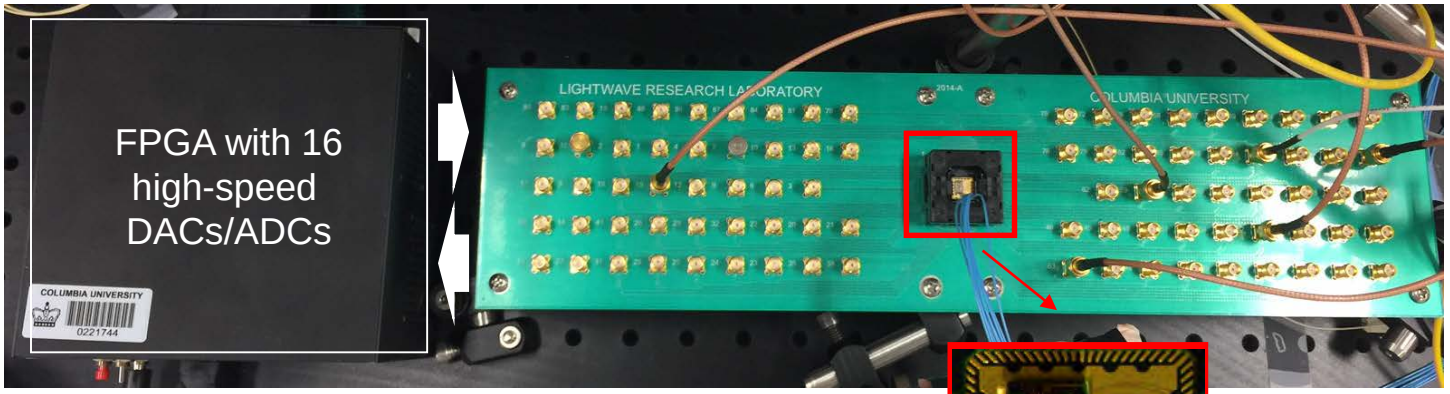


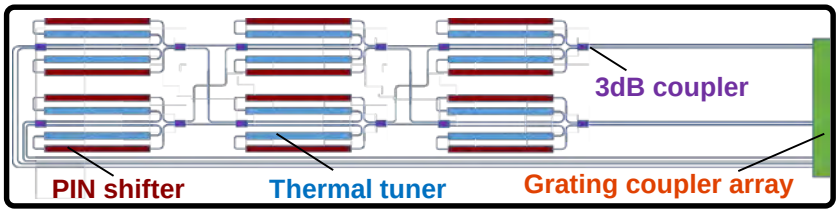



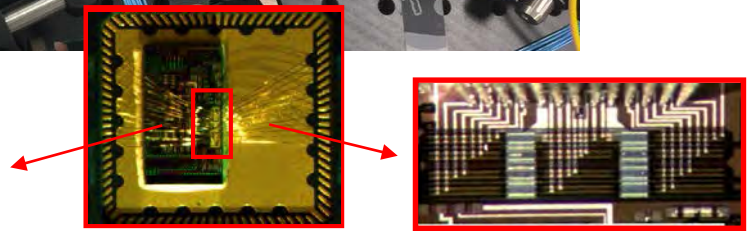

12 x12 Clos-of-Switch-and-Select Switch Layout

| | 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 |

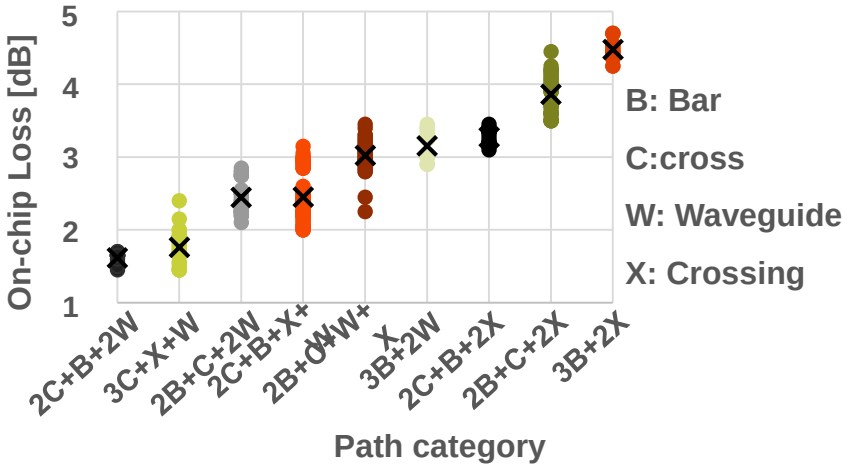
# Our FPGA-Controlled Switch Test-Bed



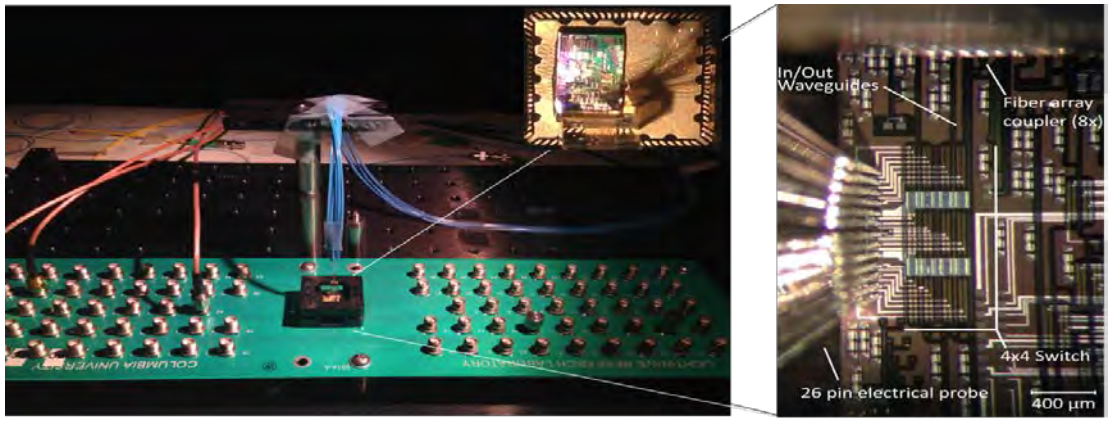



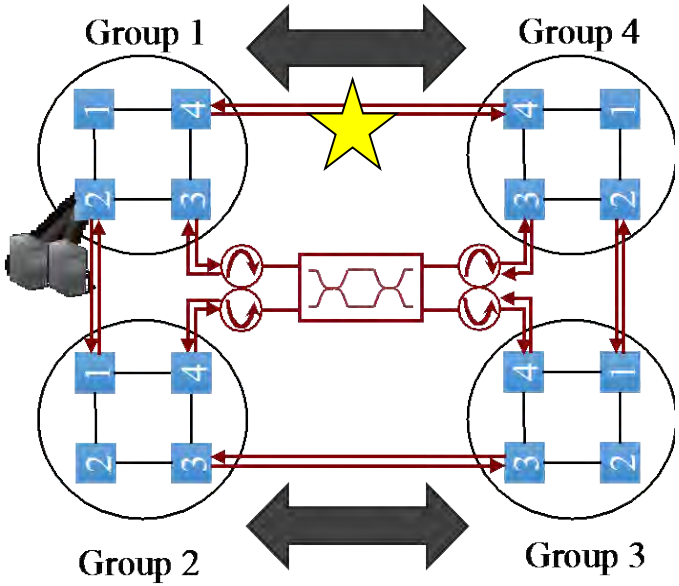



- 16 high-speed DACs enable test and control of integrated photonic switch circuit

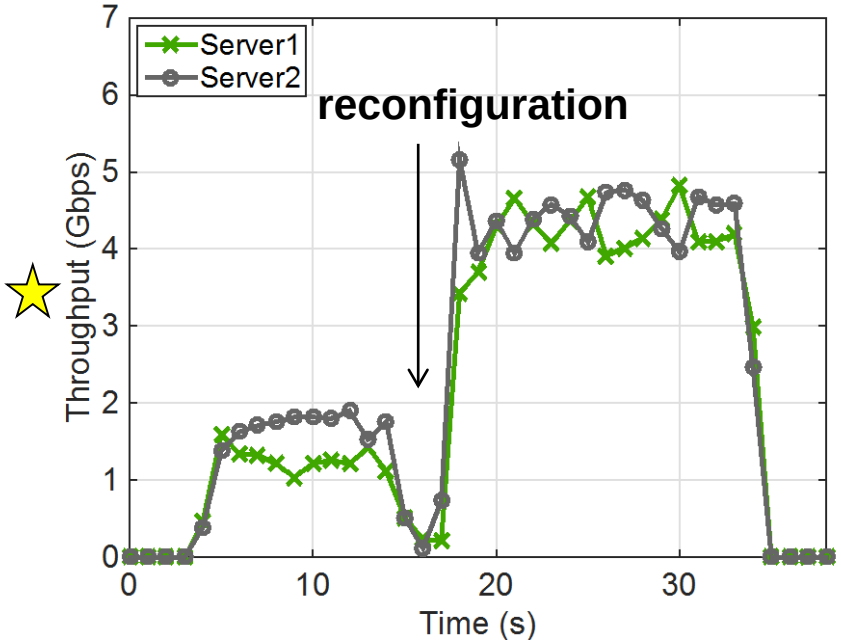


# Flexfly – testbed implementation




Group 1, Router 3
(owns reconfigured link)



# Conclusions

- Ultra-large scale interconnects are in high need for bandwidth
  - Interconnect bandwidth limitations among main HPC scalability threats
- Optics is playing a role and will continue to
  - But beware of costs and power consumption
  - Packaging is particularly important
  - Cost…
- Modeling/design must be cross-layer
- Optical switching in HPC:
  - Photonics switching for bandwidth steering
  - Flexfly: low port-count and cheap silicon photonics switches in HPC interconnects