OPTICAL INTERCONNECTS FOR EXASCALE SYSTEMS

Moray McLaren
HP Labs
13th October 2010

TALK OUTLINE

- Characteristics of optical interconnects
- CMOS nanophotonics and high radix routers
- Exascale Systems

PHOTONICS AS A DISRUPTIVE TECHNOLOGY

OPTICAL INTERCONNECTS DISRUPTIVE TECHNOLOGY OR…

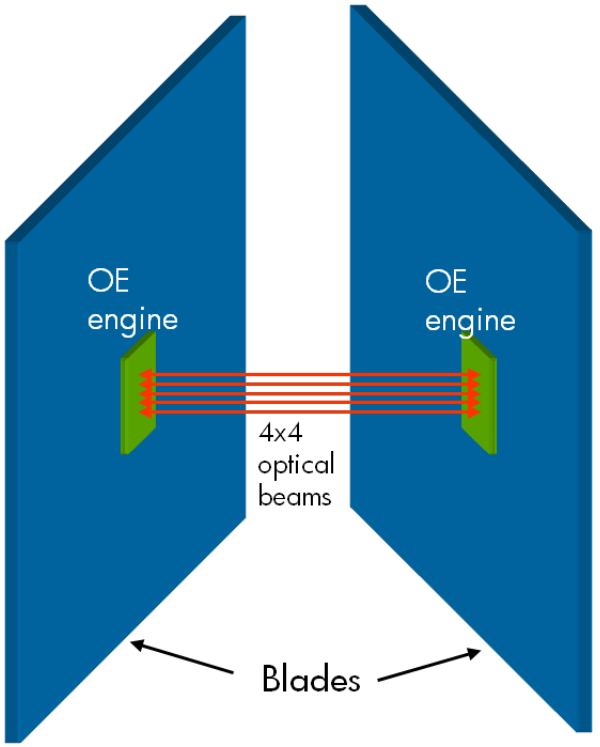
….GOLD PLATED PLUMBING?

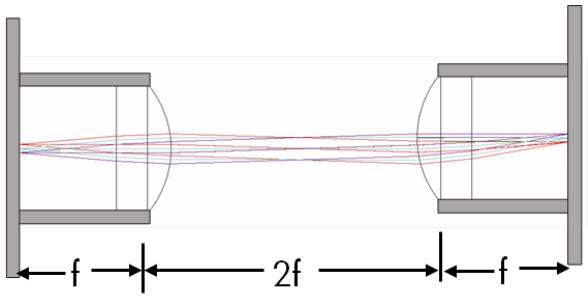
POTENTIALLY DISRUPTIVE CHARACTERISTICS

- Freespace capability
- Broadcast
- Circuit switching
- Distance independence
- Power efficiency
- Bandwidth density
- EMI immunity

FREESPACE

- Board to board interconnect using 4x4 VCSEL arrays
- Telecentric lenses allow for misalignment
- No fibers..
- Disadvantages
  - Range limited by divergence
  - Very specific to packaging solution
  - Hard to scale with WDM





BROADCAST

Broadcast based optical fabric

Optical backplane demonstrator

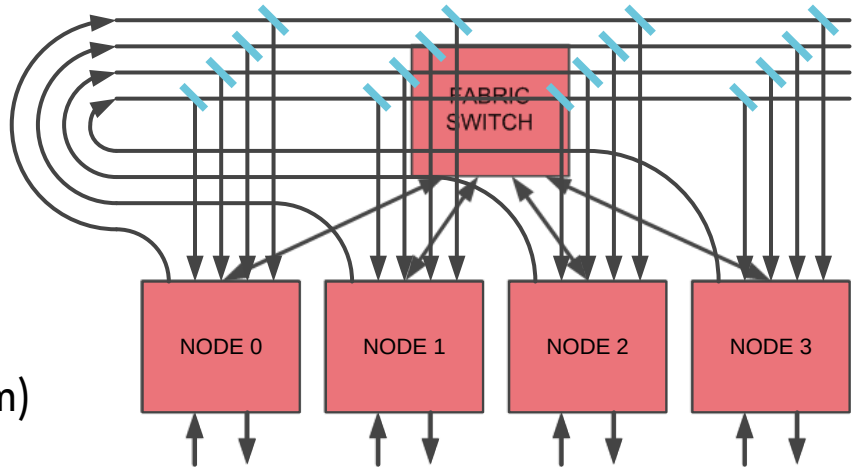
- Replace CMOS fabric ASICS with multiple broadcast buses
- Very low cost plastic waveguides
- Scalable to 16 line cards

Advantages

- Low power
- Simple passive backplane
- Upgradeable through CWDM

Broadcast for Exascale?

- Potential application to memory buses
- (but bandwidth, not capacity is the problem)



OPTICAL SWITCHING

- "I've got a 10ps optical switch, why can't you use it"
- "It's only a bit of logic to turn a circuit switch into a packet switch isn't it?"
- Many possible implementations
  - MEMS, ring resonators
- Advantages
  - Very low power operation
  - Very low through latency
- Disadvantages
  - It's not a packet switch…
- Successful Telco use model – resilience and provisioning

Glimmerglass MEMS optical switch

DISTANCE INDEPENDENCE

ELECTRONIC HIERARCHY

- Short on chip <3mm
- Global on hip <20mm
- Local off chip <200mm
- Chassis level <1000mm
- Cabled <6m
- Active cable <12m
- Active optical <100m

OPTICAL HIERARCHY

- Short on chip <3mm
- Global on chip <20mm
- Photonic <100m

BANDWIDTH DENSITY

CHOKE POINTS TODAY

At the enclosure boundary
- Screened cables

At the card edge
- Connector density
- Crosstalk limitations

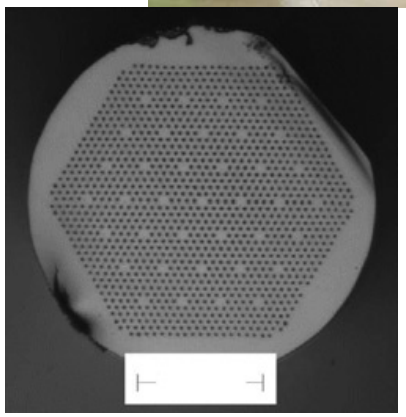
At chip edge
- Pincount limitation <6000
- Pin data rate limitation

OPTICAL BANDWIDTH SCALING

– Multiple wavelengths..

– Multiple fibers

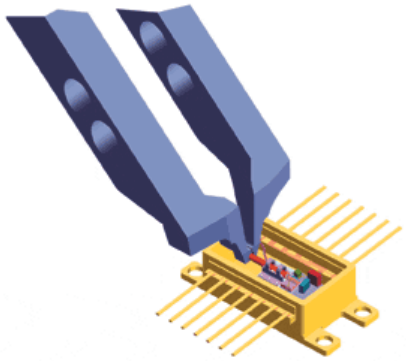
– Multiple cores..



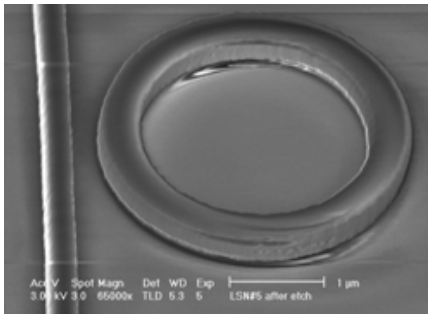
INTEGRATED PHOTONICS

Only way to fully exploit bandwidth density

- The 2000 telecom bubble based on discrete opt
  - Components are measured in mm
  - Hand alignment
  - Expensive and not scalable
- Integrated nanophotonics
  - Manufacture many thousands per die
  - Advances in lithography -> better devices
- Current generation CMOS photonics
  - Relatively large MZ modulators
  - No power advantage over VCSELs
  - Limited WDM capability
- Use resonant devices….

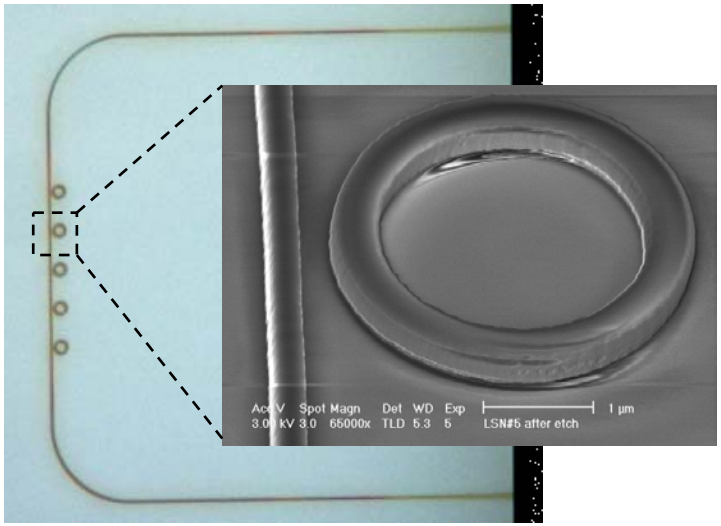


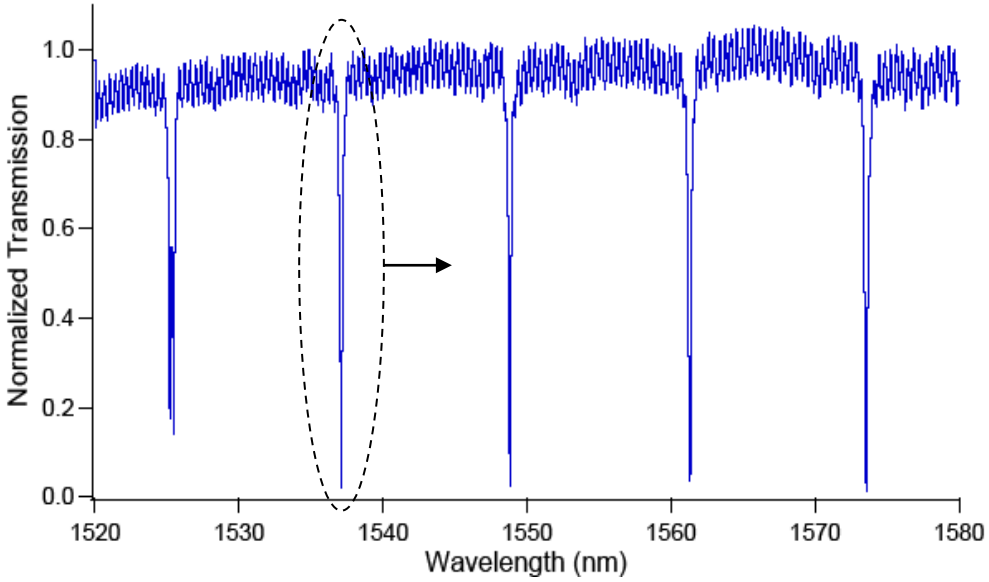
Source: Newport Corp., Assembly Magazine, September 2001

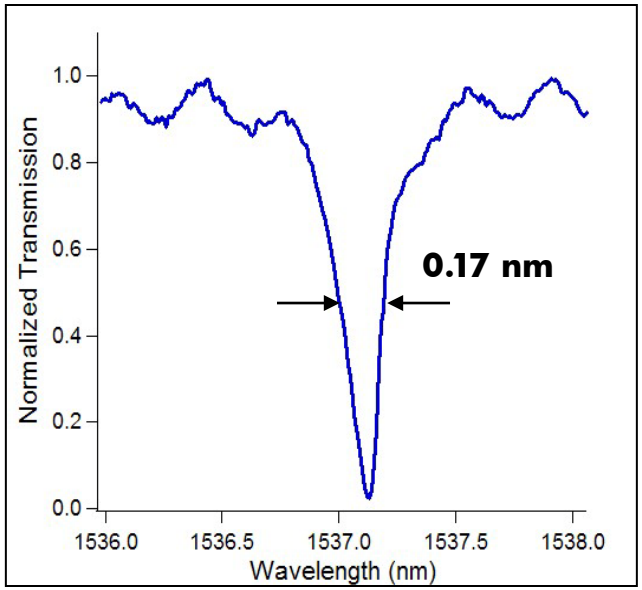


SI MICRORINGS

- Example: 5 cascaded microring resonators, slightly different radii ~ 1.5 mm.
- High Q of 9,000 (BW ~ 20 GHz) and high extinction ratio of 16 dB.



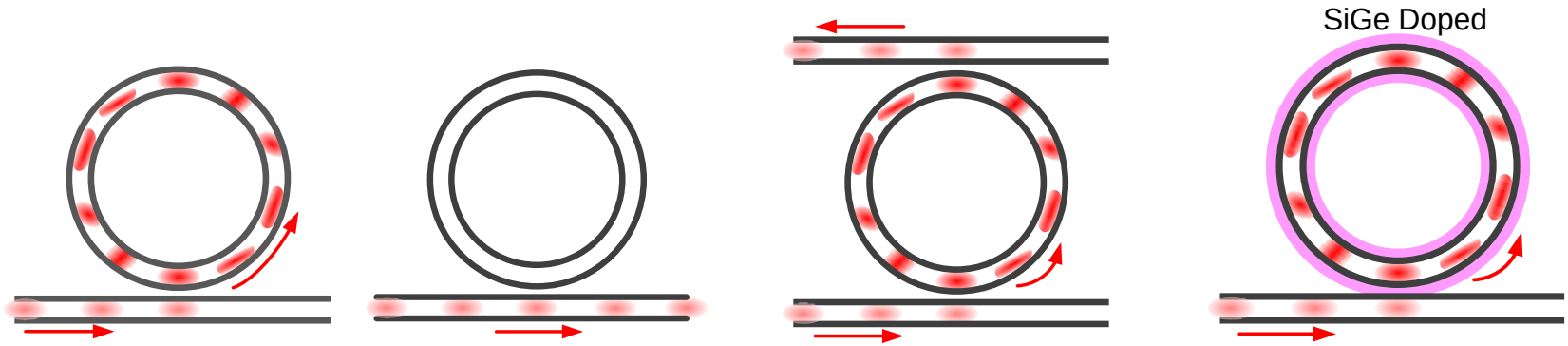




Q. Xu, D. Fattal, and RGB, Opt. Express 16, 4309-4315 (2008) — **World Record!**

RING RESONATORS

One basic structure, 3 applications

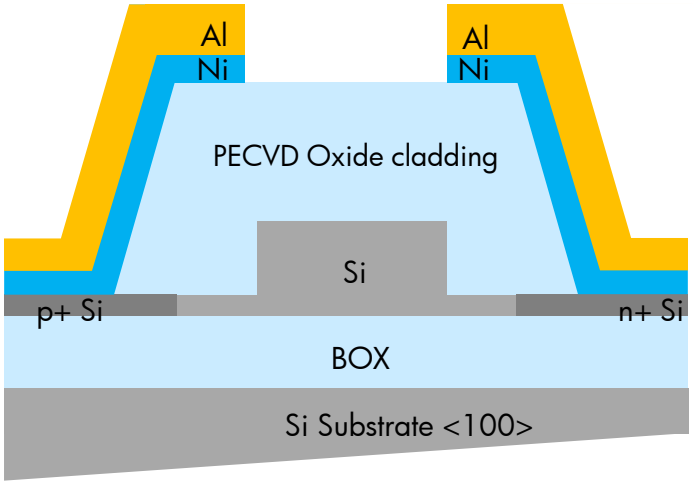


- A modulator – move in and out of resonance to modulate light on adjacent waveguide
- A switch – transfers light between waveguides only when the resonator is tuned
- A wavelength specific detector - add a doped junction to perform the receive function

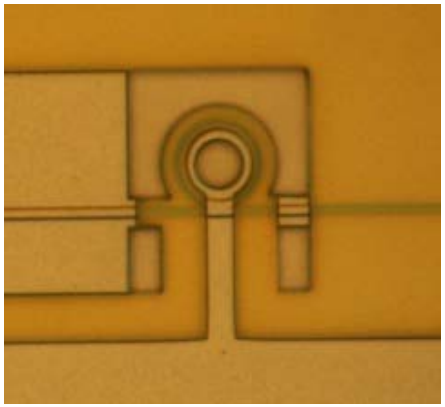
SILICON INTEGRATED CIRCUITS

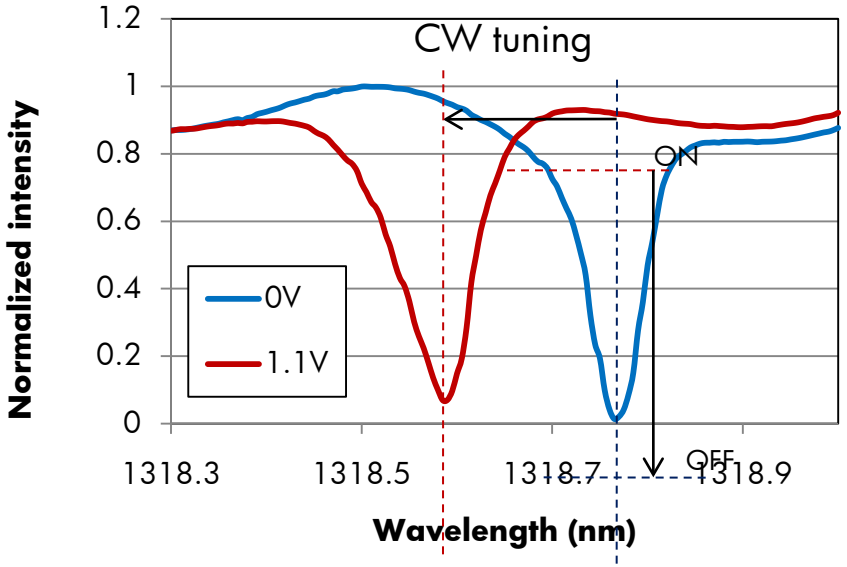
- 10 μm silicon ring resonators
  - Charge injection
  - 1310 nm (compatible with Ge detectors)
- Experimental Results
  - Q ~ 10,000
  - 0.18 nm shift
  - 18 dB extinction
  - 3 Gbps modulation
  - 54 fJ/bit

Cross section

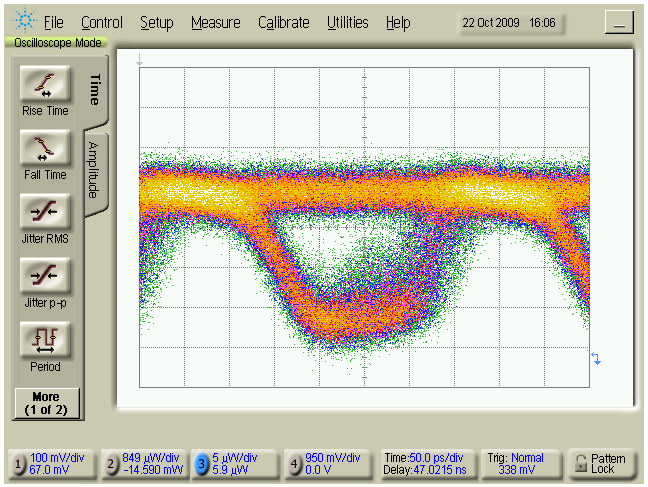


Top view



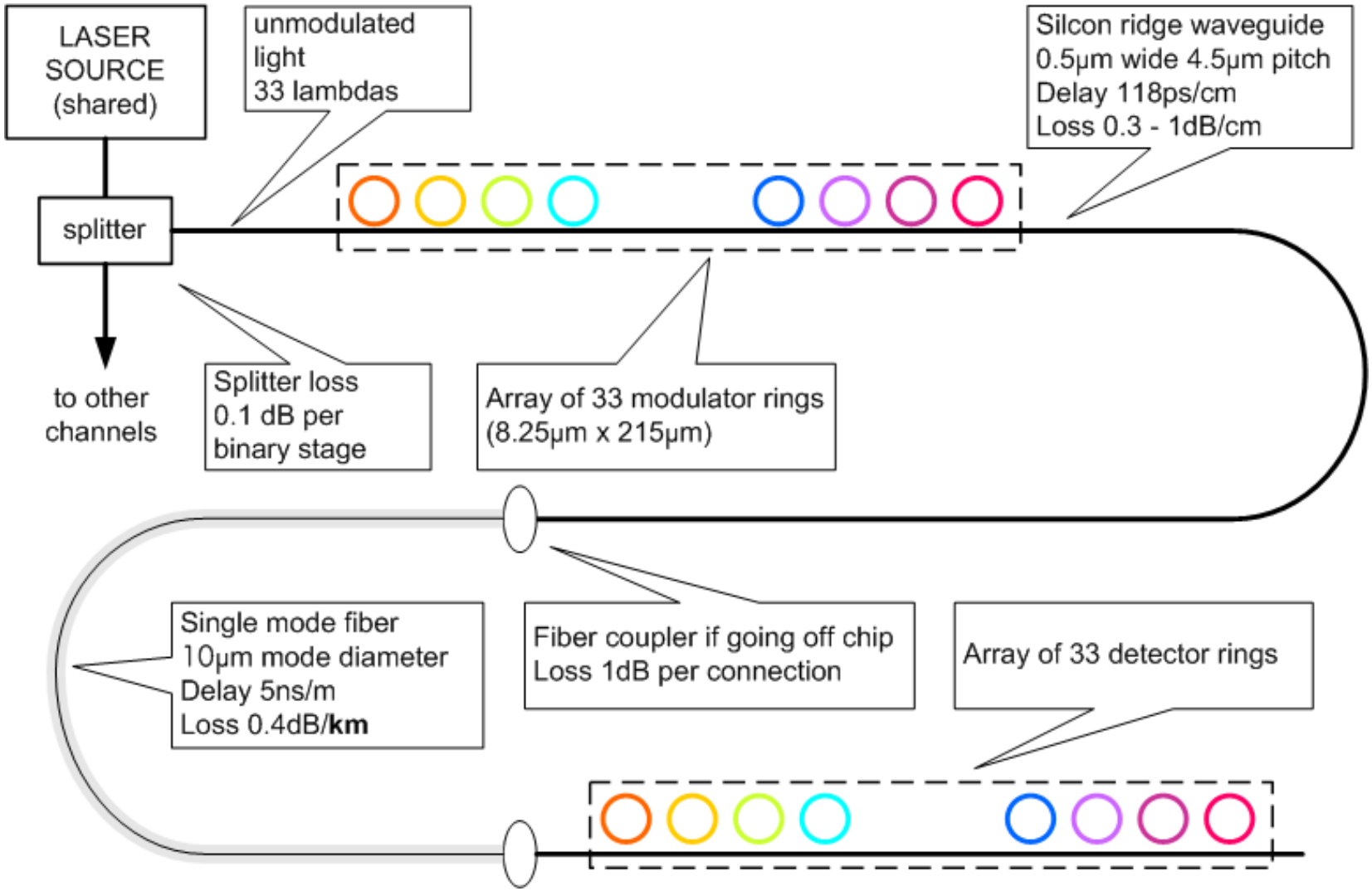


Eye diagram RZ 3 Gbps



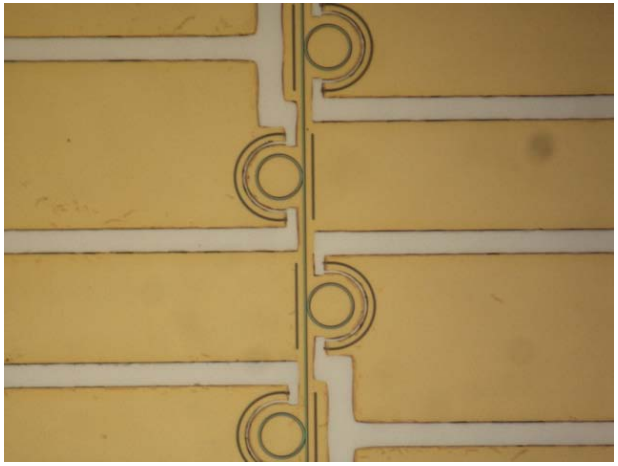
Zhen Peng – HP labs

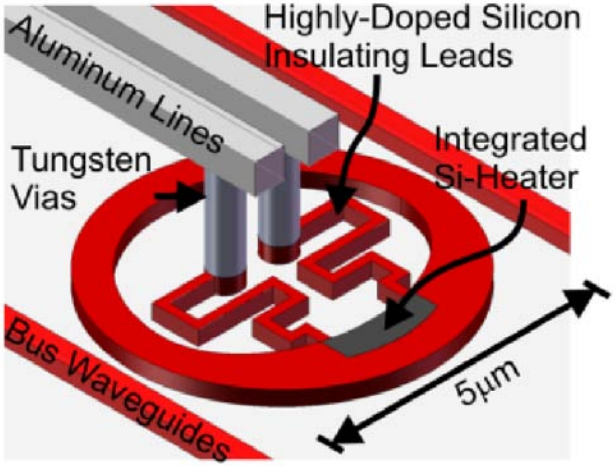
DWDM POINT TO POINT LINK



TECHNICAL CHALLENGES - TUNING

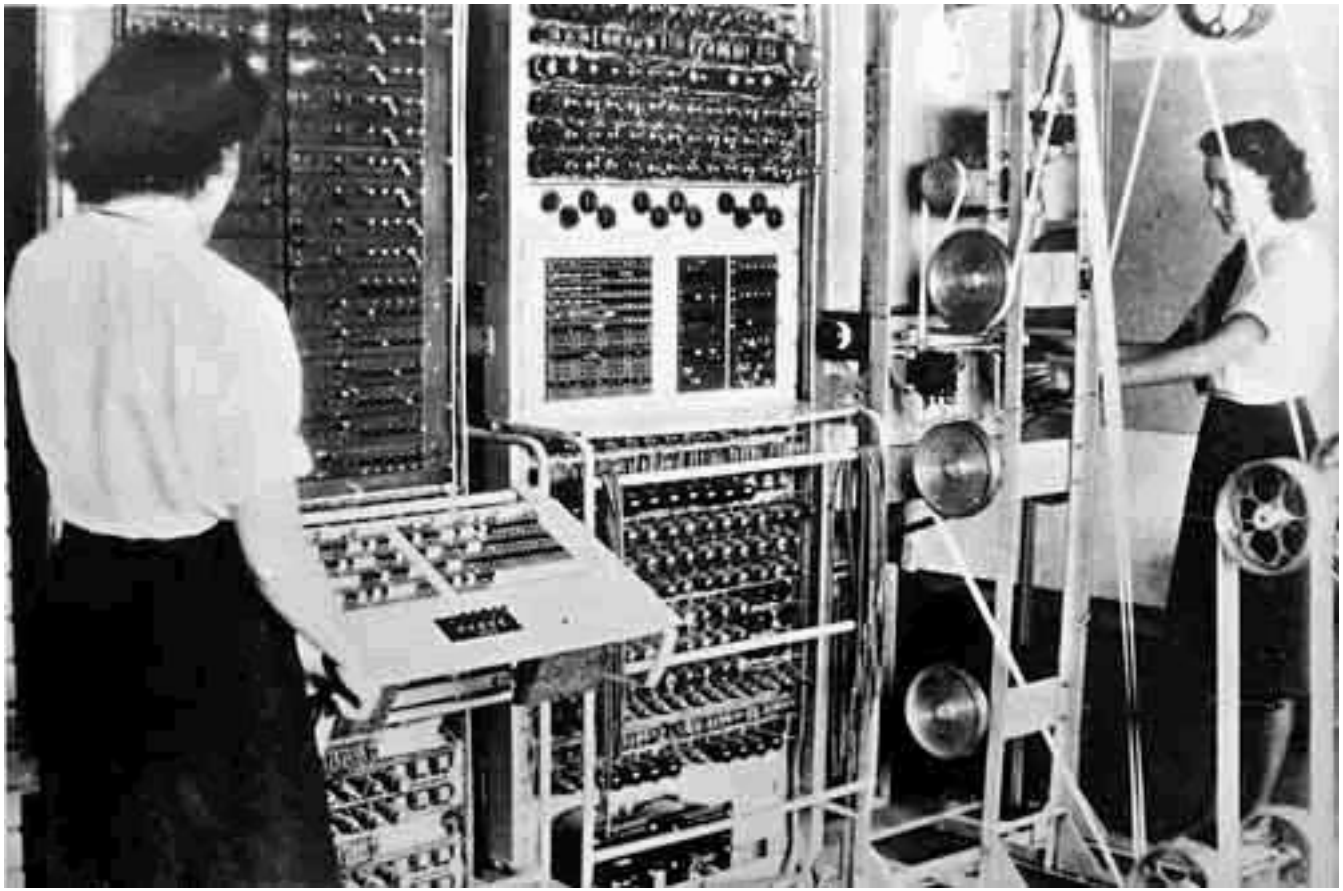
- Tuning is necessary to:
  - Compensate for fabrication variations
  - Correct for temperature variations
- Thermal tuning
  - Simple to implement
  - Minimize tuned thermal mass
- Alternate approaches
  - Self compensated rings
  - Still need to tune for process variation





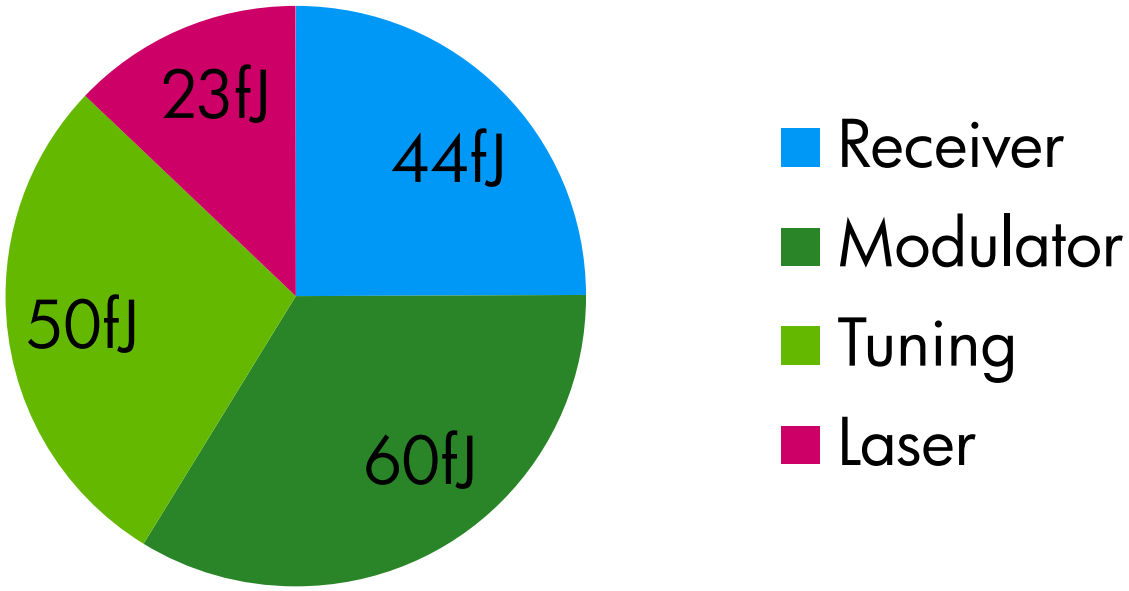
M. Watts et. al. Sandia National Lab

COMPUTERS WITH LARGE NUMBERS OF INTEGRATED HEATERS HAVE BEEN BUILT....



Colossus Mark 2 computer

POINT-TO-POINT POWER BUDGET



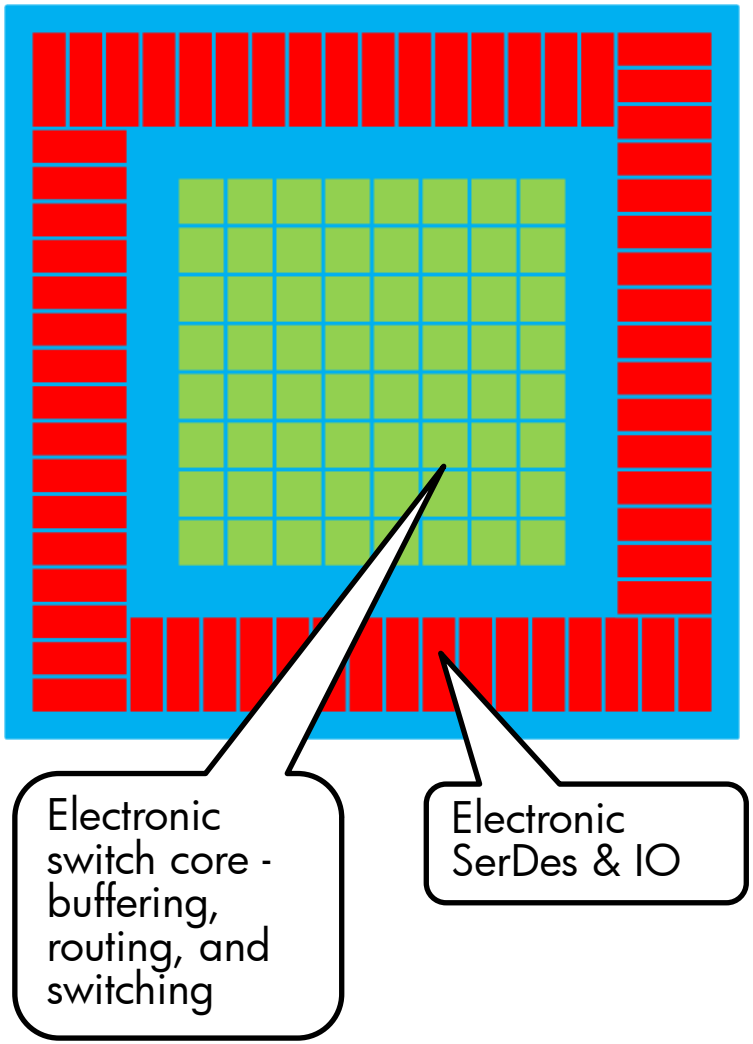
- 10Gbit/s per wavelength
- 177fJ/bit assuming 32nm process
- No clock recovery and latching - not directly comparable to electronic numbers
- Idle link still needs laser and tuning power

HIGH RADIX ROUTERS

Subtitle Placeholder

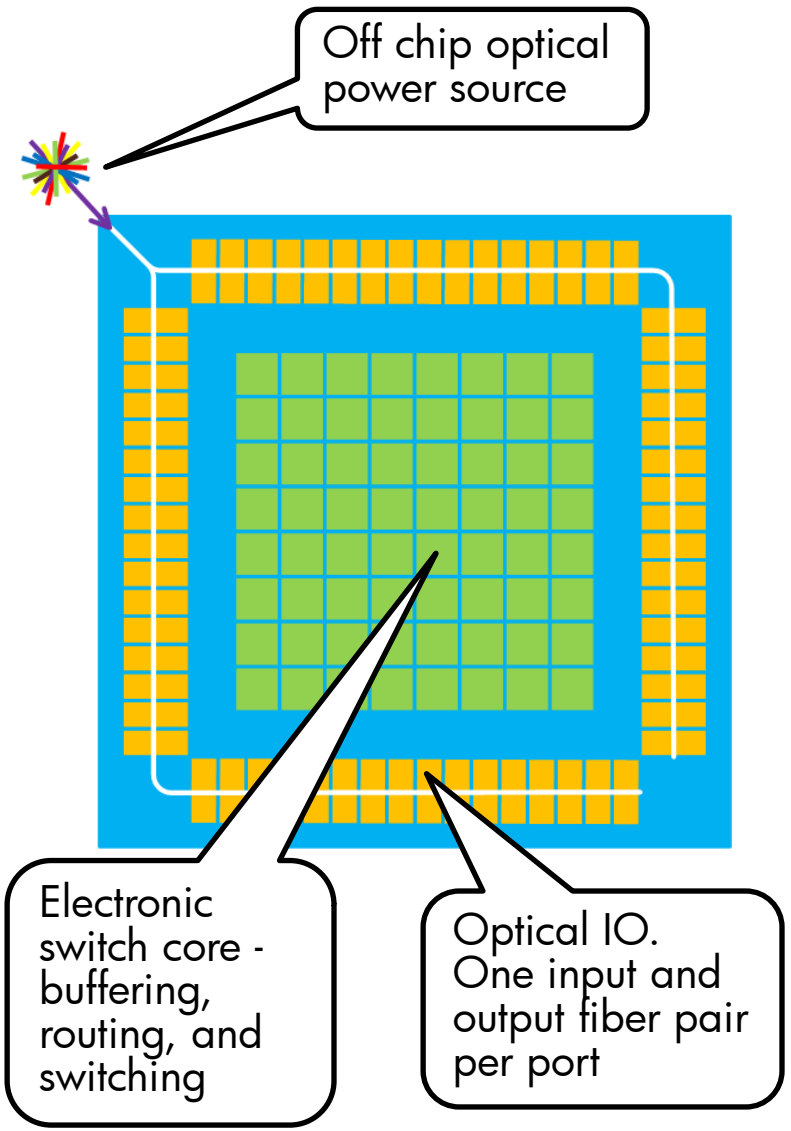
STATE OF THE ART ELECTRONIC ROUTER

- Total chip IO limited by package constraints
- Increasing port data rates only possible by reducing port count
- Overall chip power dominated by SerDes and IO power
- Pin data rates limited by power and signal integrity
- Limited cable length, (<6m today, decreasing with data rate)
-



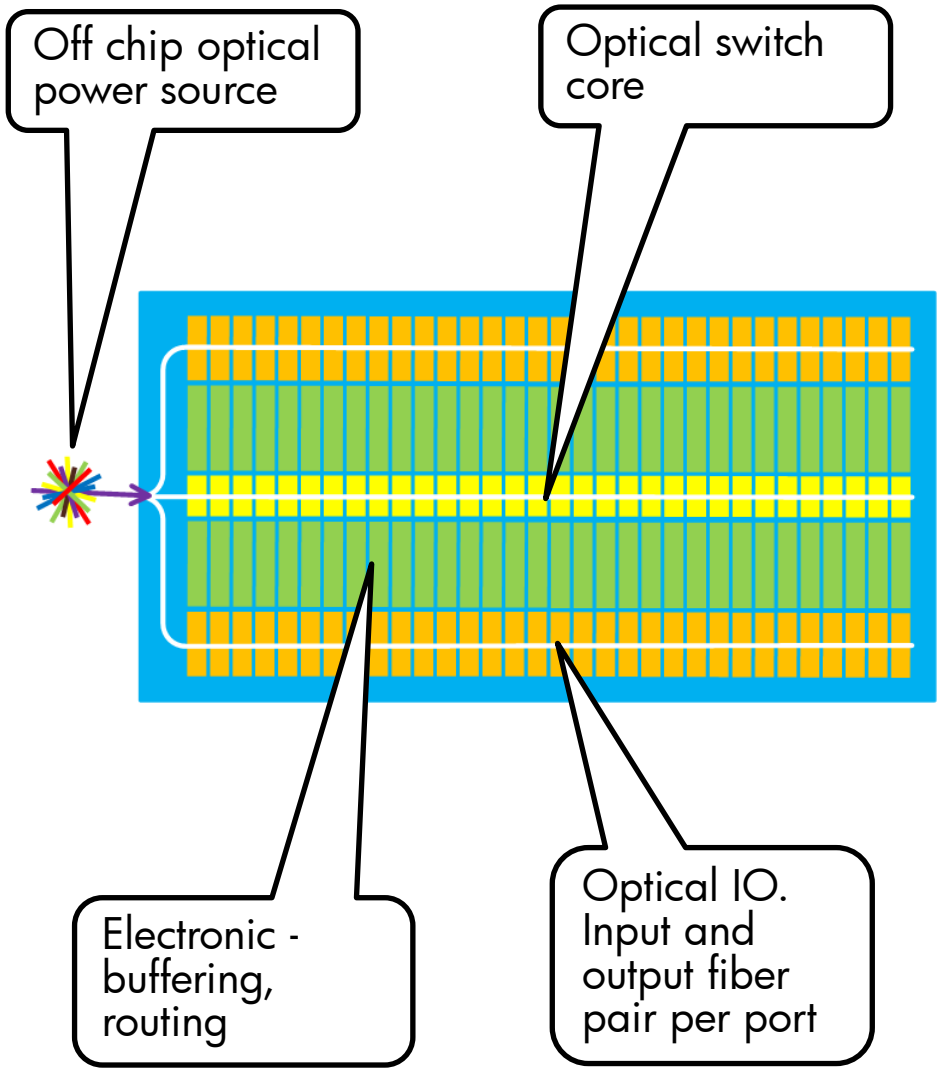
INTEGRATED PHOTONIC IO ROUTER

- High bandwidth density at chip edge due to WDM fiber interconnect
- Greatly reduced IO power
- Electronic switch core unchanged. Buffering, routing and switch functions still implemented electronically
- Unlimited cable length
- Scale port bandwidth by adding wavelengths



INTERGRATED PHOTONIC IO & CORE

- Optical switch core eliminates long electronic wires, enables full crossbar, reducing power
- Chip area dominated by electronic input buffering
- Area (&cost) scales ~linearly with port count
- Folding possible for large switches
- Switch arbitration may be optical or electronic



ROUTER IO POWER SCALING

IO power in watts for 64, 100 and 144 port switches

Generation	Port BW(Gbps)	IO type	fJ/bit	64	100	144
45nm	80	Electronic	7000	35.8W	56.0W	80.6W
	80	Optical	451	2.3W	3.6W	5.2W
35nm	160	Electronic	5048	51.7W	80.8W	116.3W
	160	Optical	284	2.9W	4.5W	6.5W
22nm	320	Electronic	4049	82.9W	129.6W	186.6W
	320	Optical	191	3.9W	6.1W	8.8W

- Assumes each generation will require a doubling of port speed to match improvements in processor performance
- Total device power must be less than 130W for normal forced air cooling with heatsink, 200W possible with special cooling. Assume max power budget of 50% (65W) for IO.
- Conservative modulation rate enables power efficient DDR clocking (data rates is 2x system clock rate)

ROUTER CORE POWER SCALING

CORE POWER = CHIP POWER – IO POWER

Generation	Port BW	core type	64	100	144
45nm	80Gbps	Electronic	41.8W	72.7W	120.7W
		Optical	13.2W	17.4W	31.9W
35nm	160Gbps	Electronic	38.0W	65.9W	109.0W
		Optical	22.9W	27.7W	50.9W
22nm	320Gbps	Electronic	52.4W	91.9W	153.8W
		Optical	34.2W	41.3W	76.3W

- Core speed doubles with each generation to match improvements in processor performance
- Total device power must be less than 130W for normal forced air cooling with heatsink. IO power negligible with photonic IO.
- Photonic switch core allows low power operation at high port counts and high bandwidth.
- Router chip area (&cost) scales linearly with port count for optical switch core case.

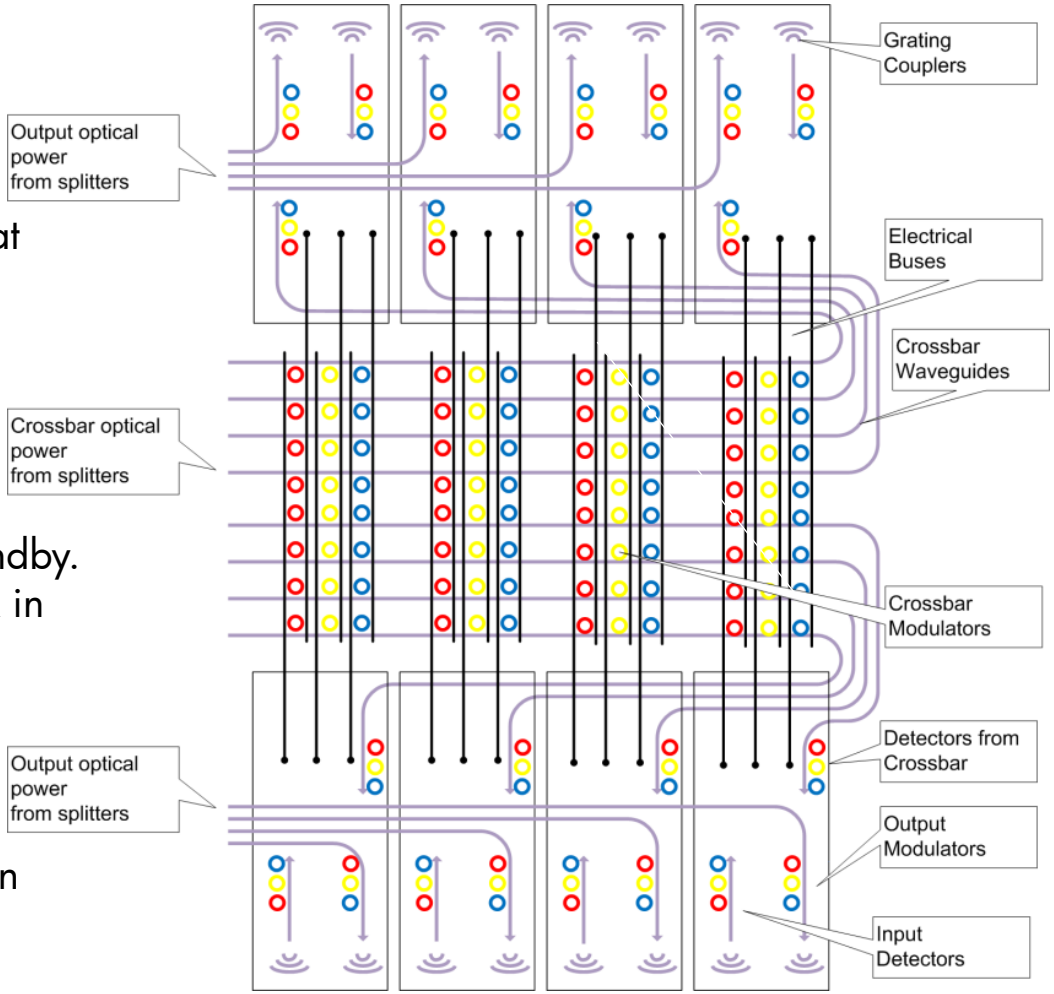
PHOTONIC ROUTER

- Pair of fibers per port
- Buffering and routing logic in CMOS
- OEOEO structure with electronic buffering at inputs and outputs shown
- OEO switches possible with no buffering at outputs.
- Fully non-blocking crossbar
- Clustering to reduce number of rings in standby. Multiple ports share single modulator block in optical core (2-way shown)
- Optical or electronic arbitration

  Optical simple, fast, low-power

  Electronic allows more complex arbitration algorithms, age-based, QoS etc.

  Need to evaluate trade-off at system level

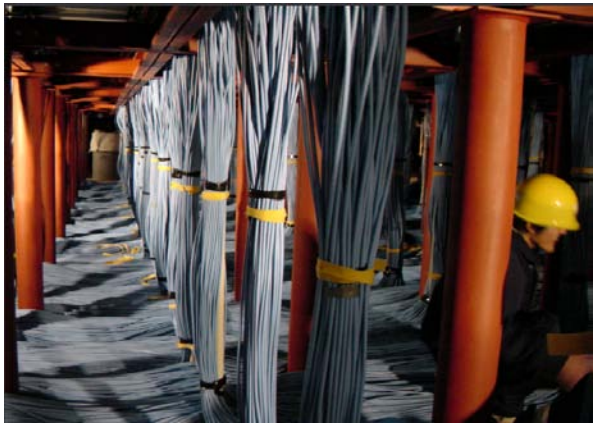


EXASCALE SYSTEMS

EXASCALE INTERCONNECT CHALLENGES

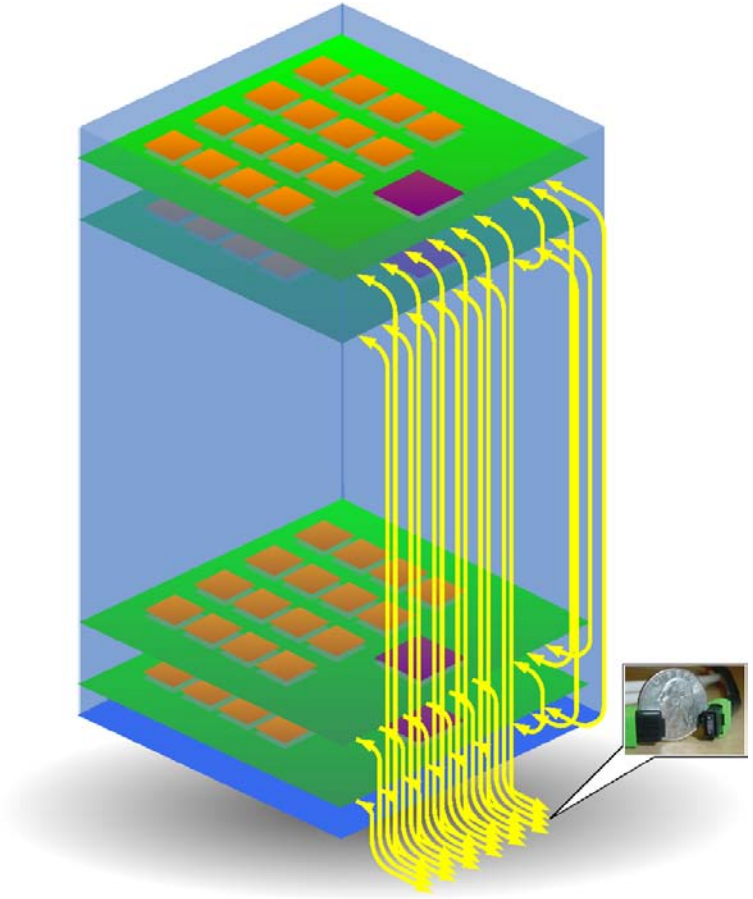
Subtitle Placeholder Goes Here

- Topology choices
  - Cabling complexity
  - Network diameter
- Enclosure level deployment
- Cost trade-off in global bisection bandwidth
  - What size of domain is it desirable to maintain full bisection bandwidth?
- Central switches
  - Impractical at Exascale level
- Power
  - Proportionality
  - Ability to power down unused resources



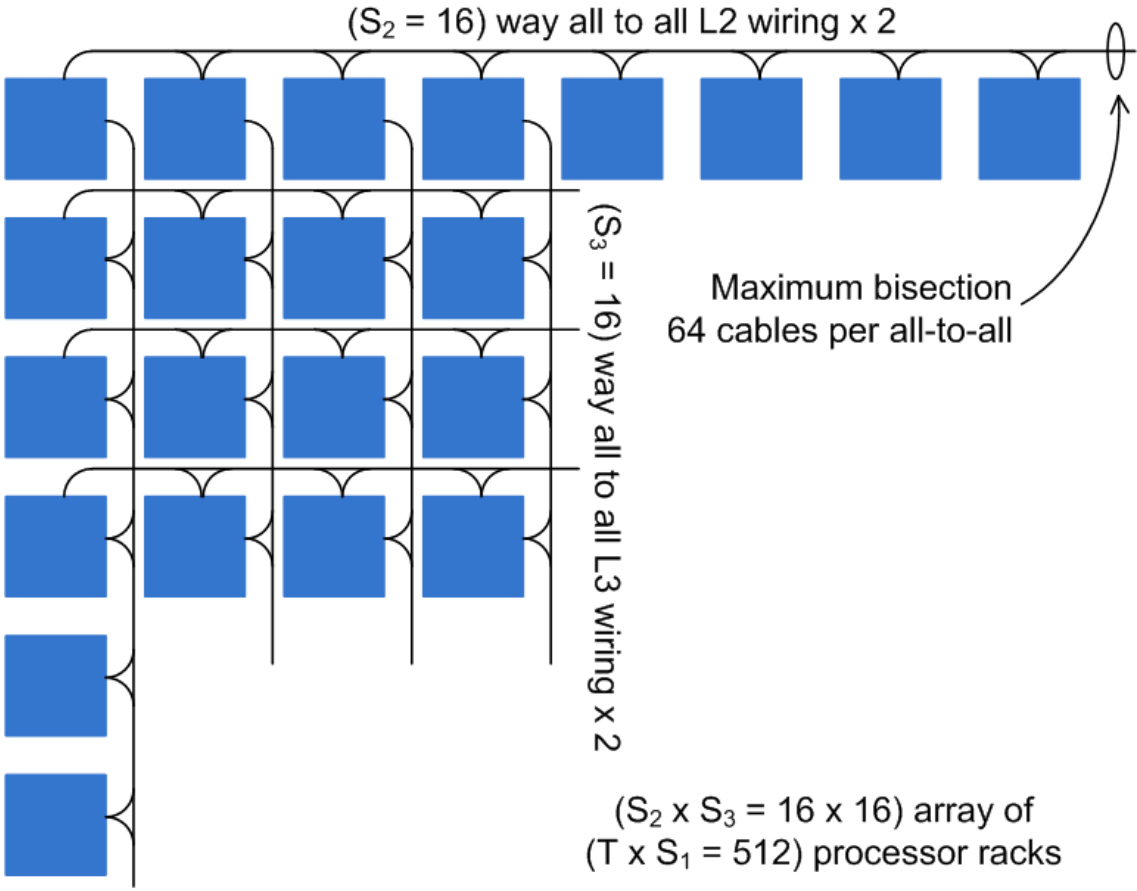
ENCLOSURE LEVEL CONNECTIVITY

- Blade card
  - 16, ~10Tflops processors
  - 96 port packet switch
  - 2, 160Gbyte/s links per processor
- 10TB/s blade to backplane bandwidth
  - 128 fibers, 32mm at 250um pitch
  - Bandwidth constrained by power (&cost) not card-edge
- 32 blade backplane
  - 32 way all to all wiring between blades within enclosure
  - Consolidate links to multi-fiber connectors
- External connectivity
  - 160TB/s external bandwidth
  - 32, 64 fiber connections



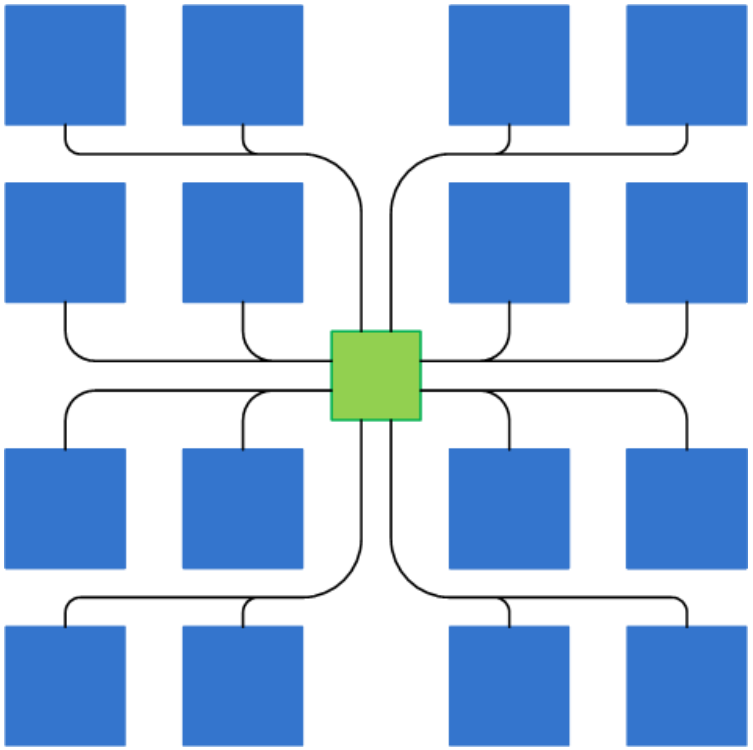
OVERALL SYSTEM

- 256 enclosures
- No central switches
- Flattened butterfly/HyperX connectivity
- 5 /10 Petabytes/s bisection bandwidth
- Replicated fiber wiring
- Wide range of link lengths



MIXED PACKET & CIRCUIT SWITCHING?

- Use circuit switches to configure connectivity between enclosures
- Modify circuit switches for changes in usage patterns and response to hardware failures (Telco model)
- Reduces power by minimizing number of times packets are actively switched
- Increase application bandwidth by minimizing deroutes



Q&A

