

Exascale 2020

2010 IN2P3

patrick.demichel@hp.com HPC EMEA



HP Labs around the world

Beijing

Tokyo

Palo Alto

Bristol

St. Petersburg

Haifa

Bangalore

7 locations

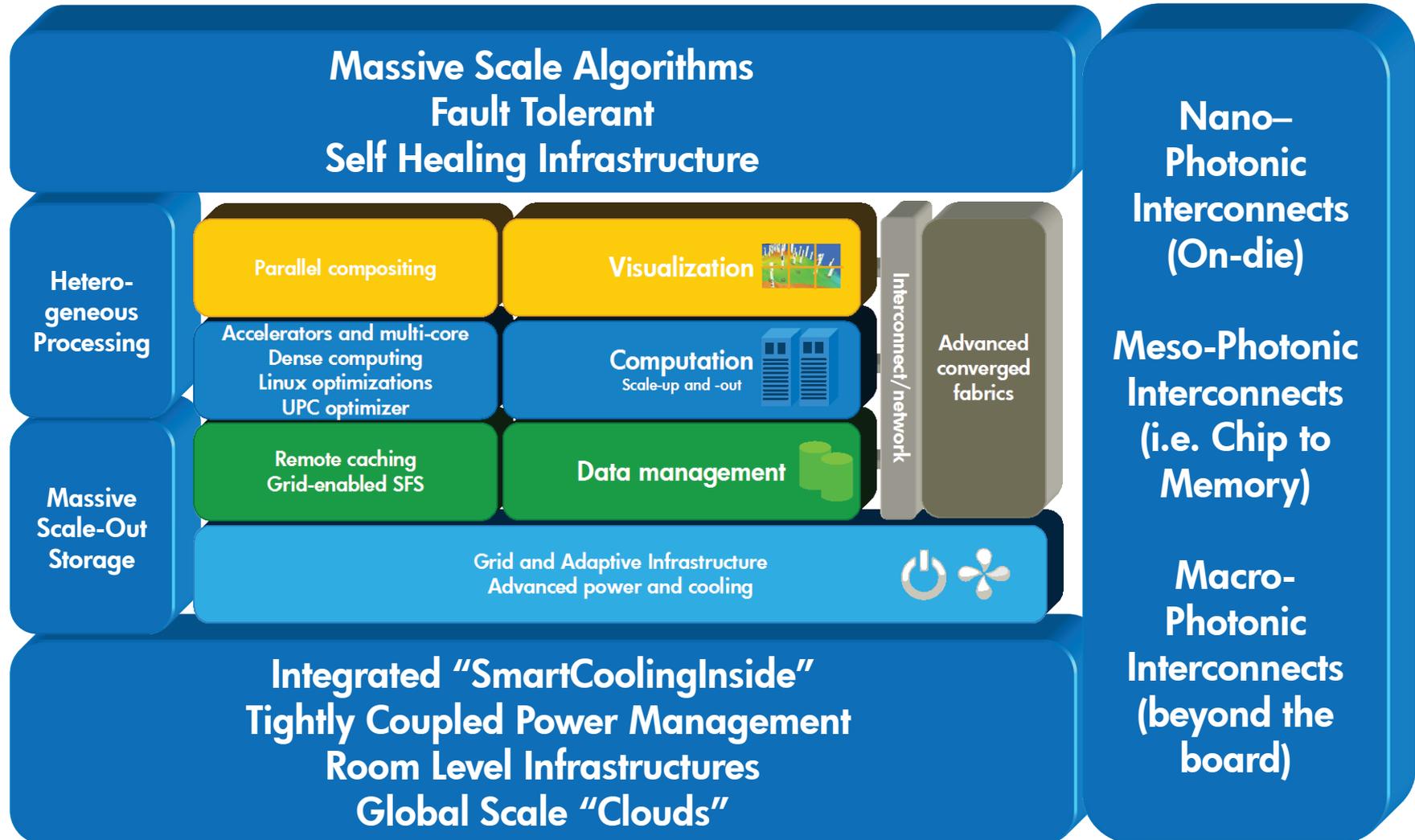
600 researchers in 23 newly formed labs

5 research themes with 20-30 projects at a time

Exascale Computing laboratory directed by Norm Jouppi



Extending the vision: ExaScale Program



The Power Wall

– At current trends, exascale systems are expected to consume > **100MW** in 2018

– Energy today:

- DP FP Op 50 pJ/op
- Access L1D Cache 33pJ
- Access L2D Cache **150pJ**
- Access DRAM **2000pJ**



– How can we reduce data movement & make it use less energy?

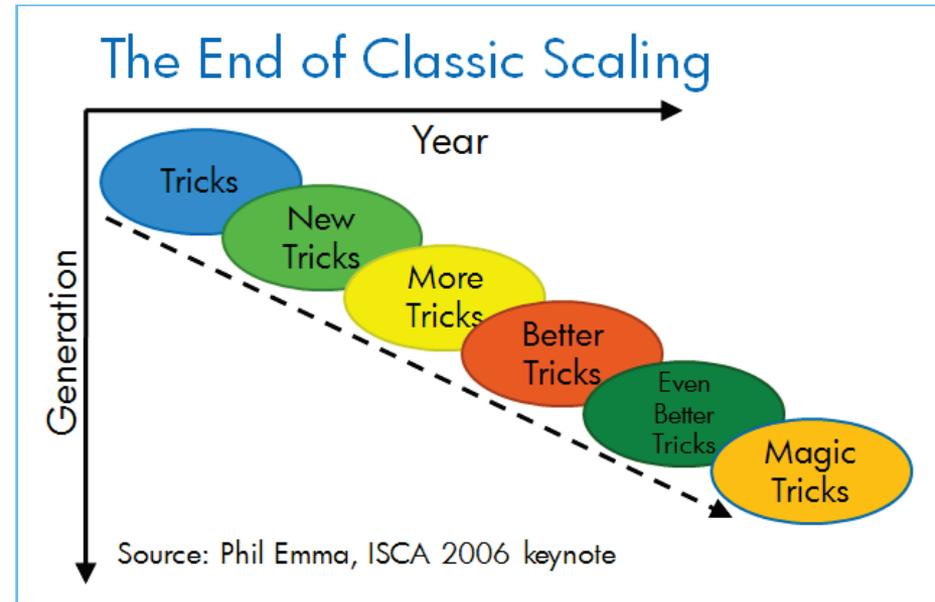
– DARPA UHPC studies:

http://www.darpa.mil/IPTO/solicit/solicit_open.asp

Speed Bumps on the Road to 2018

- Off-chip bandwidth requirements will scale geometrically
 - (Up to 10 TB/s)
- ITRS pin counts increase from a max of 3072 pins today to:
 - **3072 pins in 2017!**
- On-chip bandwidths scale geometrically too
 - **Interconnect power is a tougher constraint at each generation**
 - Mesh and ring bandwidth and latency vary based on data placement
- Non-uniform latencies & bandwidth complicate programming
 - **Programmer has to worry about placement of data & threads**
 - Placement needs to change with each new chip

⇒ We need disruptive technologies



Boost Performance/TCO 10X

– Energy Efficiency:

- *Interconnects* through photonics
 - 5x (vs. copper) in near term (5y) with chip-to-chip optics
 - 10x (long-term) with nanophotonics (and 10x bandwidth)
- *Computing nodes* through low power CPUs
- *Memory hierarchy* through NV memory (memristors)

– Manageability: 1 operator / petascale (100K servers)

>100x improvement vs. today's admin/server ratio – ranging from 1:1000 (yahoo) to 1:100 (HP-IT)

– Scalability: to 1M nodes

– Programming:

- Reliability, programmability, efficiency



Optical interconnect



1. Optical Interconnect at All Scales

The Optically Interconnected Datacenter

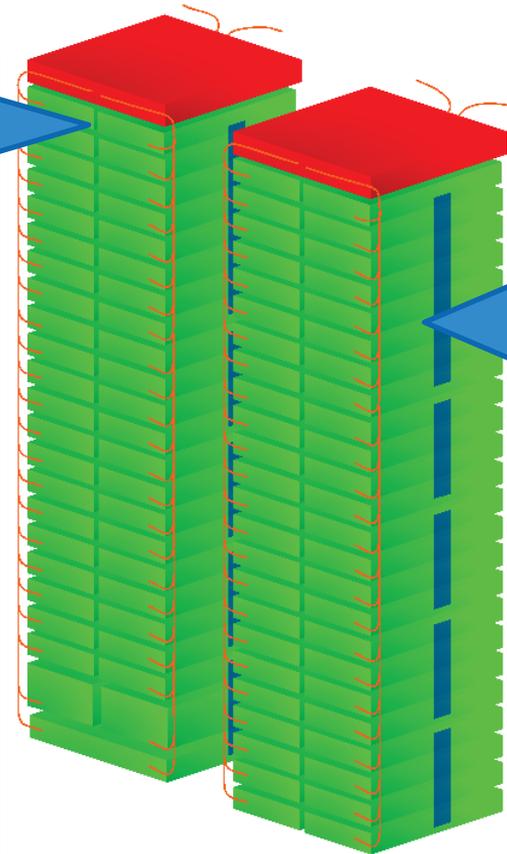
Networking

Research Challenges

- High radix switches
- Optical fabrics
- High radix routers
- Connectors, engines, media

Opportunities

- Uniform bandwidth and latency → high flexibility, new programming models
- Lower TCO through power saving, ease of installation, flexibility



Memory/CPU Interconnect

Research Challenges

- Low cost optical bus structures
- New memory architectures
- Large Scale Integrated Nanophotonics

Opportunities

- New architectures
- Flexibility, re-configurability
- New packaging options
- Lower power

Metrics

Discrete optics

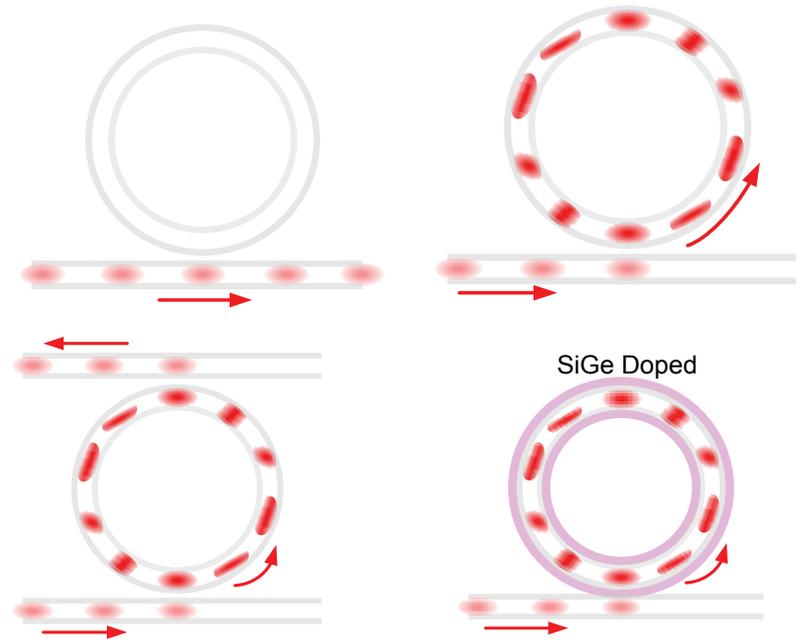
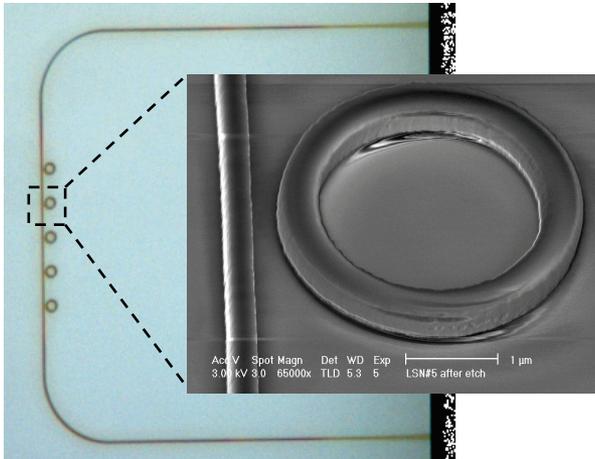
- 5x higher BW density
- 5x lower power

Integrated photonics

- 20x higher BW/pin
- 5x further power reduction

Ring Resonator : one basic structure, 3 applications

- CMOS nanophotonics
- Wavelength Division Multiplexing
- **Radically lower power** : <200fJ/bit
- Chip to chip anywhere in the data center
 - Single mode fiber 0.4 dB/Km
- **"Cost" independent of length**



- Example: 5 cascaded microring resonators, slightly different radii ~ **1500 nm**.
- High **Q** of **9,000** (BW ~ 20 GHz) and high extinction ratio of 16 dB.

- **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

Optical Waveguide

– Hollow Metal Waveguides⁽¹⁾ (HMWG)

- Low propagation loss – light rays travel at near grazing angle to metal walls
- Low numerical aperture
- Prop delay 33psec/cm

Air core

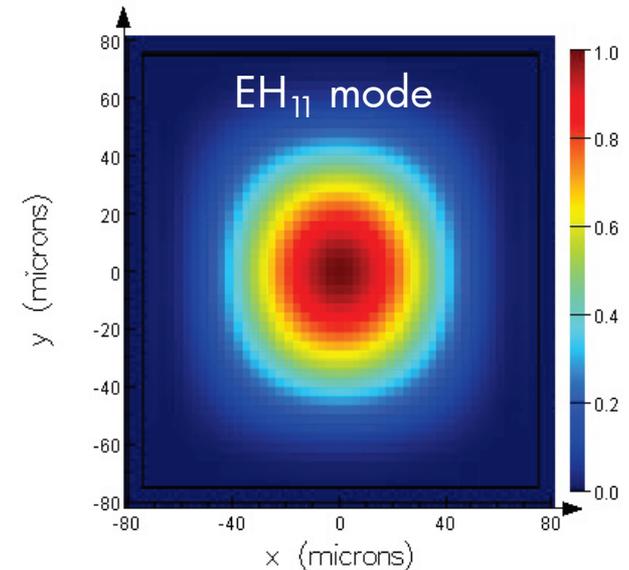
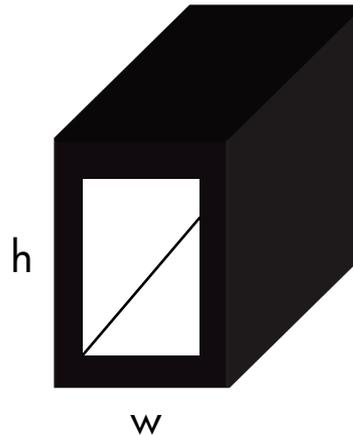
Ag clad (n, k) = (0.15+i 5.68)

$w = 150\mu\text{m}, h = 150\mu\text{m}$

$\alpha = 0.0015 \text{ dB/cm}$

$n_{\text{eff}} \sim 1$

NA ~ 0.01



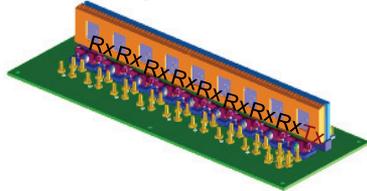
(1) E. Marcatili *et al.*, *Bell Syst. Tech. J.* 43, 1783 (1964).

HP photonics technologies roadmap

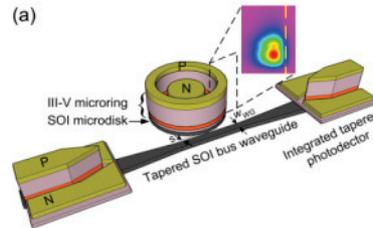
Active cable



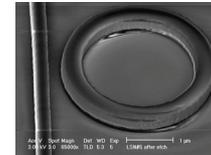
Optical Bus



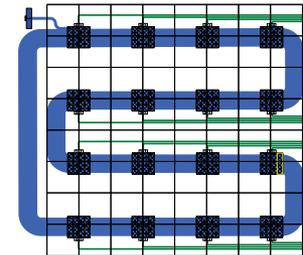
Hybrid laser cable



Silicon PIC



On-chip interconnect



Optical switch demonstrator

- Objectives

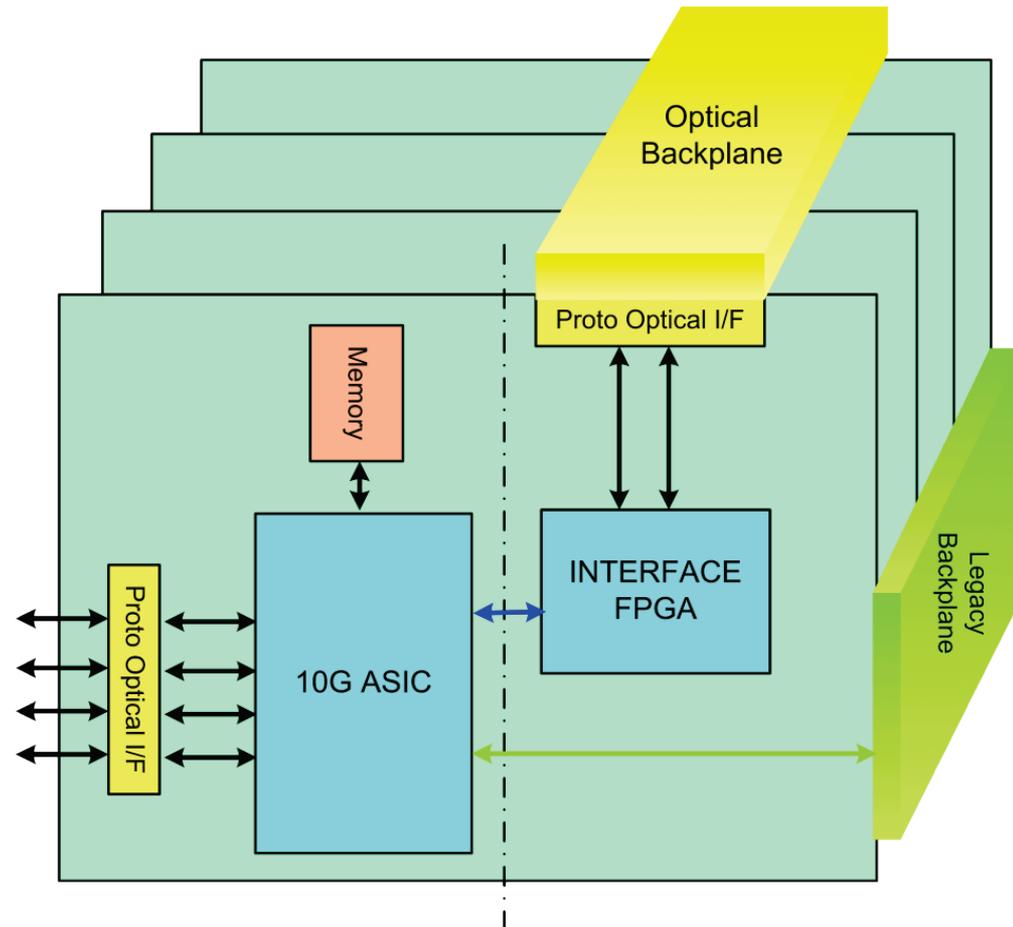
- Enable >10Gb/s signaling
- Demo “switchless” fabrics
- Lower power
- Understand implementation issues for new technologies

- Key technologies

- Optical broadcast buses
- Alignment tolerant connectors
- Low cost optical engines
- Multi bus based fabrics

- Implementation

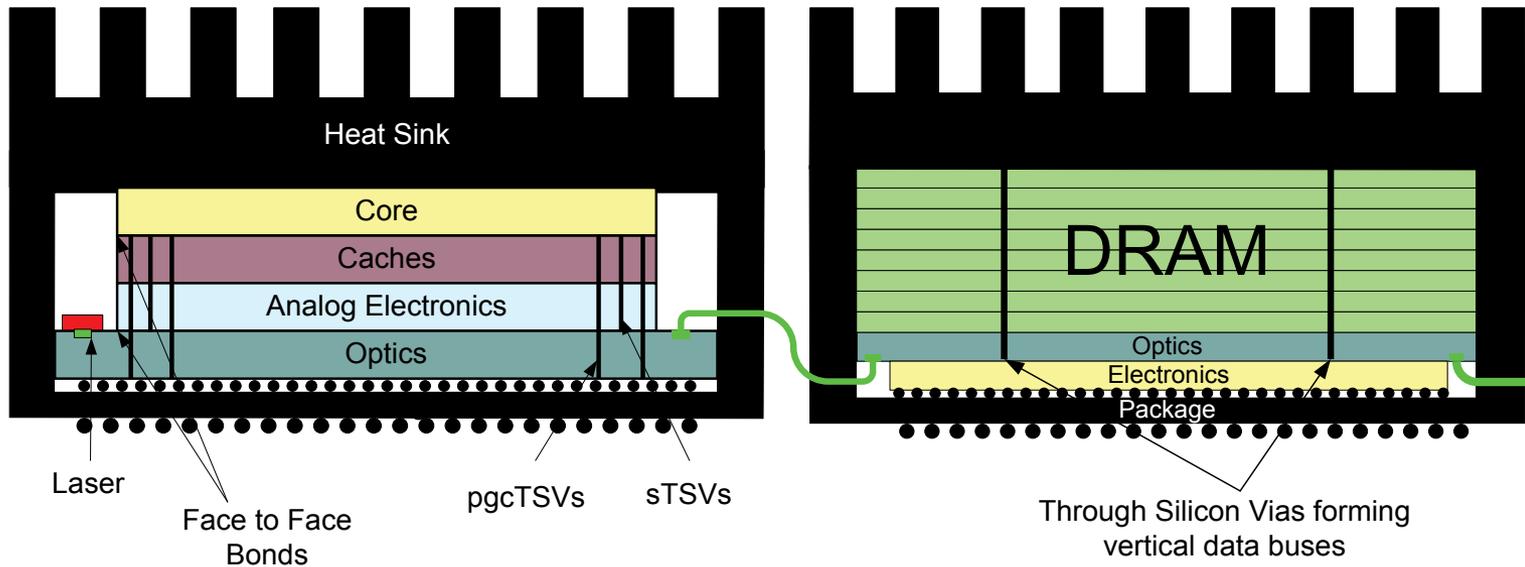
- Leverage existing switch design
- FPGAs for flexible prototyping
- Joint project with ProCurve



Socket



Corona Chip Stack

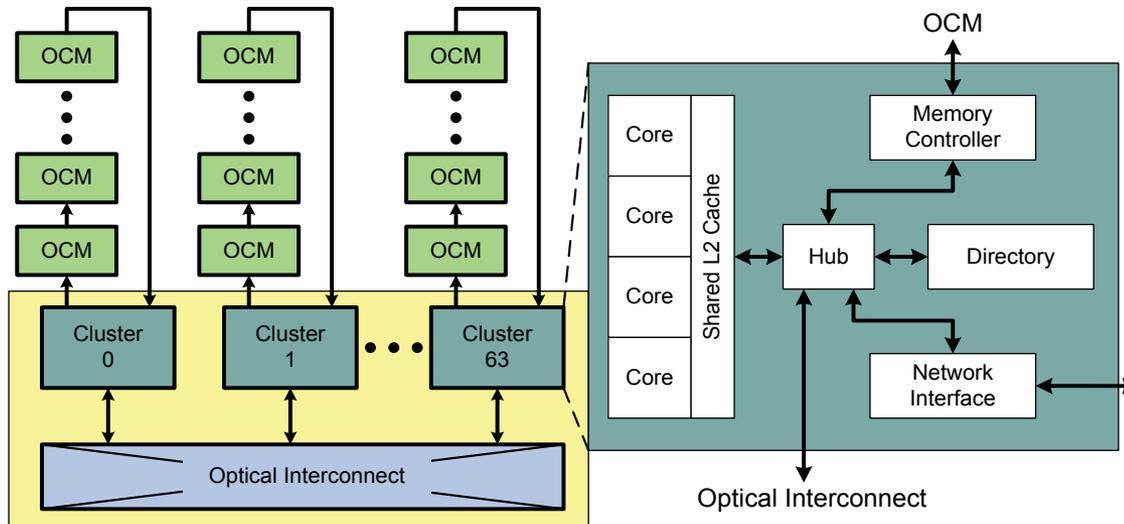


- Stacking technology minimizes electrical path lengths
- Compatible layers, each tailored to their function
- Chip to chip and intrachip communications > 5mm are optical
- Combine photonic interconnect with stacked DRAM
- Amortizes optics cost across multiple devices
- Allows high bandwidth access to DRAM array
- Multiple vertical data buses using TSVs (Through Silicon Vias)
- Could mix memory types in stack (NVRAM?)

Corona* many core architecture



Optically enabled 256 core processor



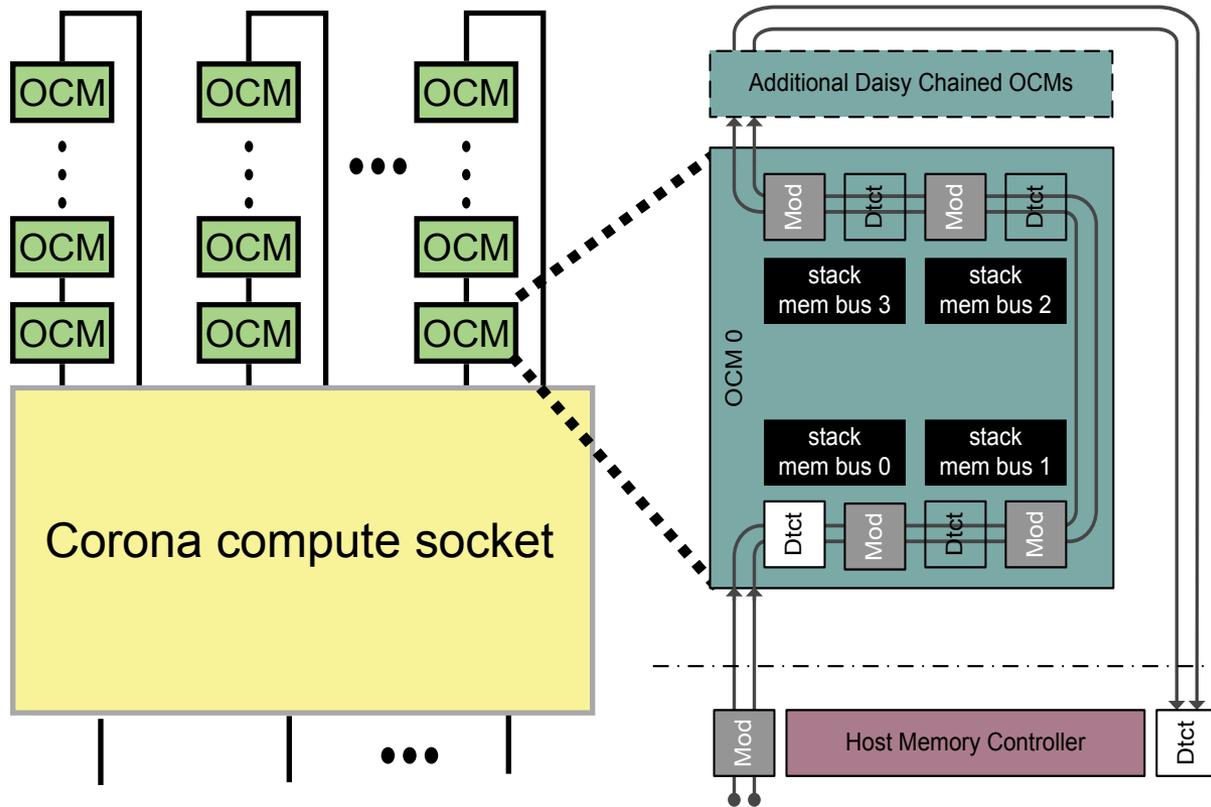
- Per each of 64 clusters:
 - Cores: 4
 - Memory controllers: 1
 - L2 cache: 4 MB, 16-way
- Per-core:
 - Frequency: 5 GHz in-order
 - Threads: 4
 - L1 I-Cache: 16 KB, 4-way
 - L1 D-Cache: 32 KB, 4-way
 - Issue: 2-wide in-order
 - 64 b SIMD FP width 4 + Fused FP operations

- High off socket bandwidth requires high intrachip bandwidth
- 20 Terabyte/s optical processor intrachip interconnect (1 byte/flop)
- All optical arbitration

vs. 2D electronic mesh

- 16x higher BW
- ~4x lower interconnect power
- Uniform 8 cycle latency vs. 5 cycles per hop mesh latency

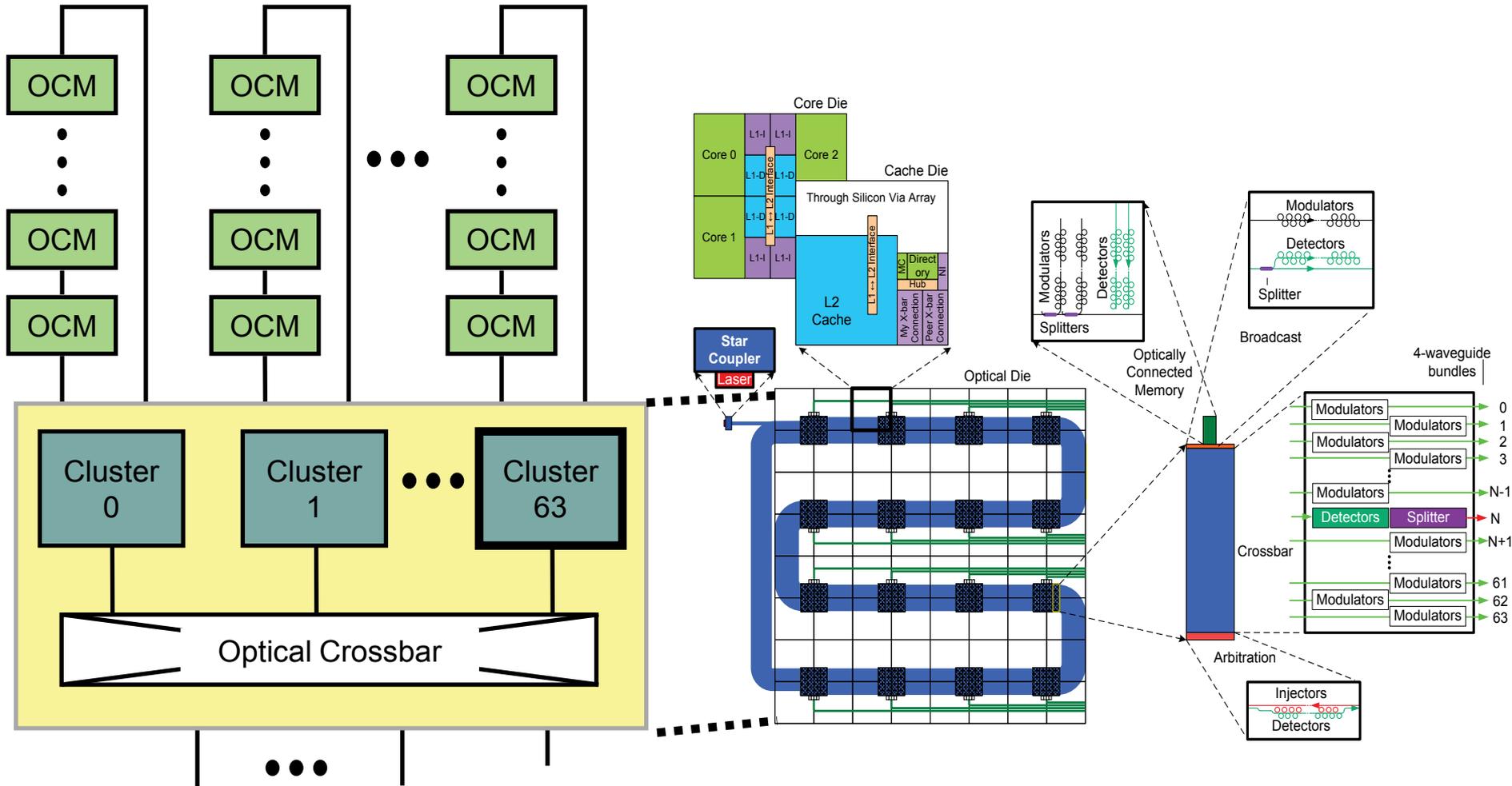
Optically Connected Memory



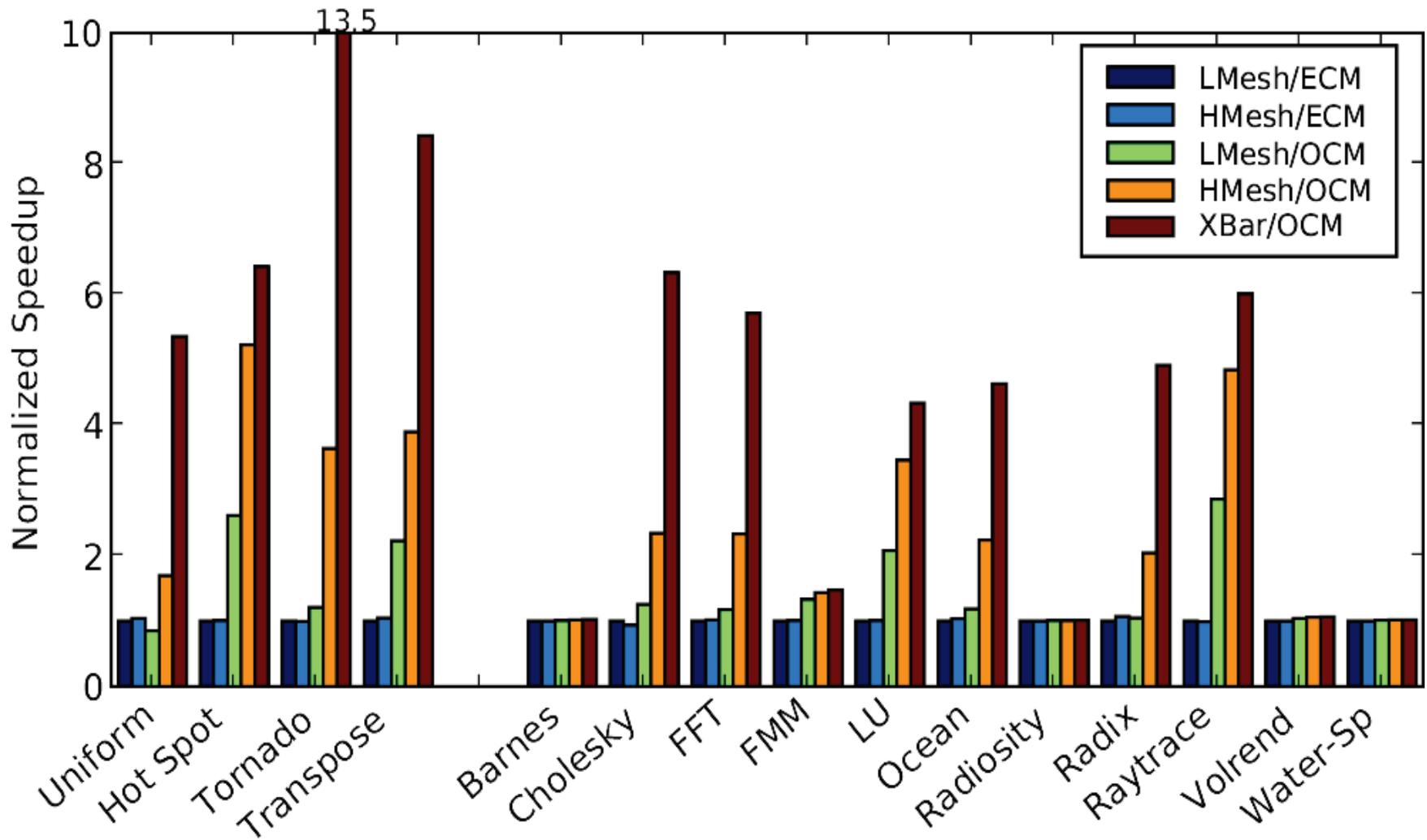
- Master/slave bus on waveguide loop
 - Optical power from processor
 - Processor modulates for data out
 - OCM modulates for return data
- Multiple optical interfaces per chip stack
 - Eliminates electronic global wiring
- OCMs communicate via DWDM
 - High bandwidth
- Accessed in parallel, no receive and retransmit like FBDIMM
 - Large capacities with low latency and power
- OCM only activates one DRAM mat per cache line fill/write
 - Less overfetching (in conventional DIMM 128X) → much lower power

Much higher bandwidth at very low power

The Optical Crossbar

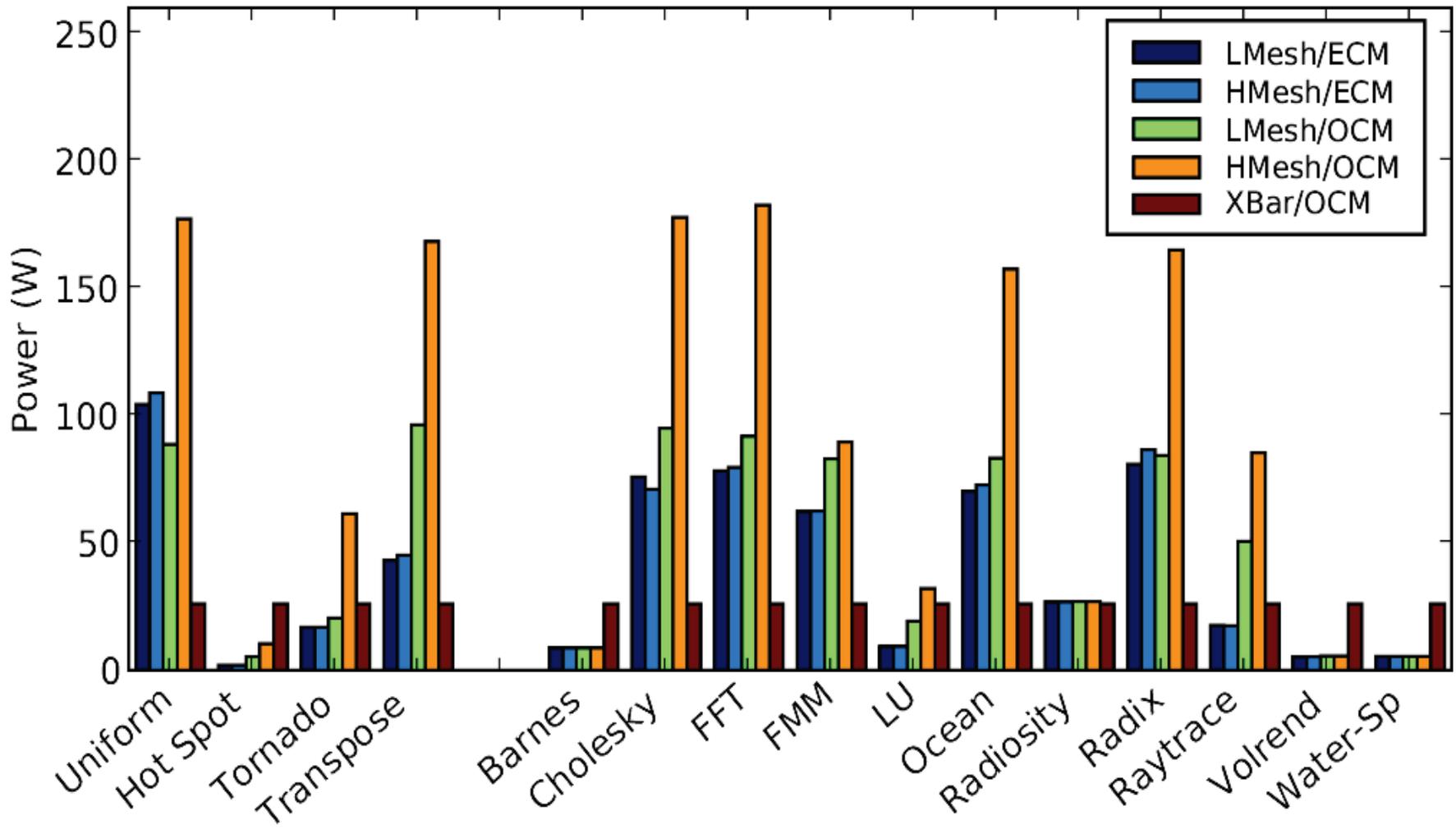


Performance (LMesh/ECM = 1)



Applications that don't fit in cache show 4-6X improvements with Xbar

On-chip Network Power



Optics can reduce network power of aps that don't fit in cache by 6X

Corona Benefits from Optics



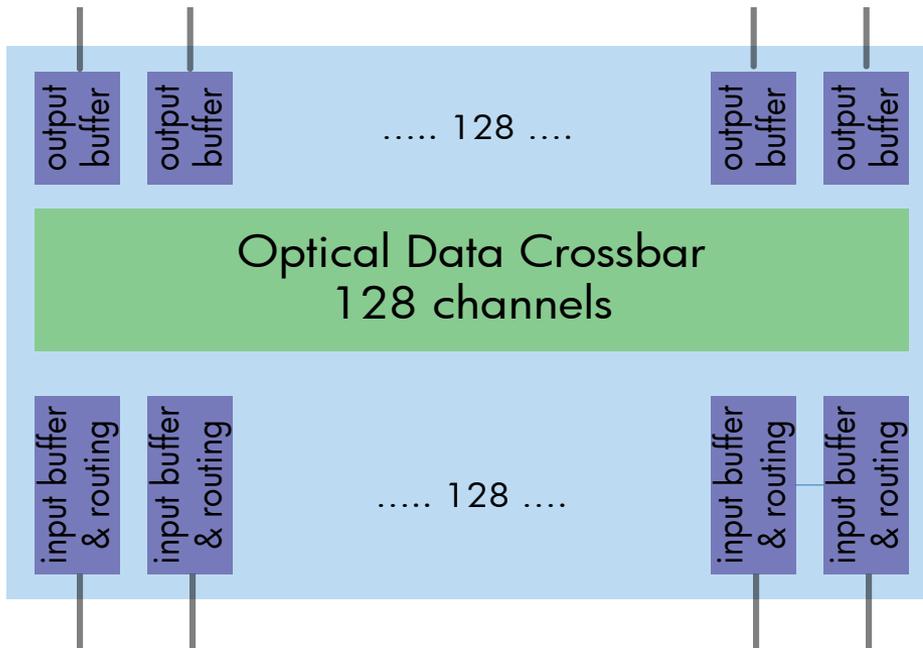
- Bandwidth scales to 1,000 threads
 - 10 Tflops
 - 10 TB/s off-chip bandwidth
 - 20 TB/s bandwidth between cores
 - Modest power requirements 200Watts
- Low, uniform latencies between cores & memory
 - Optical crossbar
- Coherent shared memory

system



High Radix CMOS nanophotonic switch

- 64 to 128 DWDM ports
- <math><200\text{fJ/bit}</math> IO power
- 160 – 640 Gbps/s per port
- Electronic or nanophotonic switch core



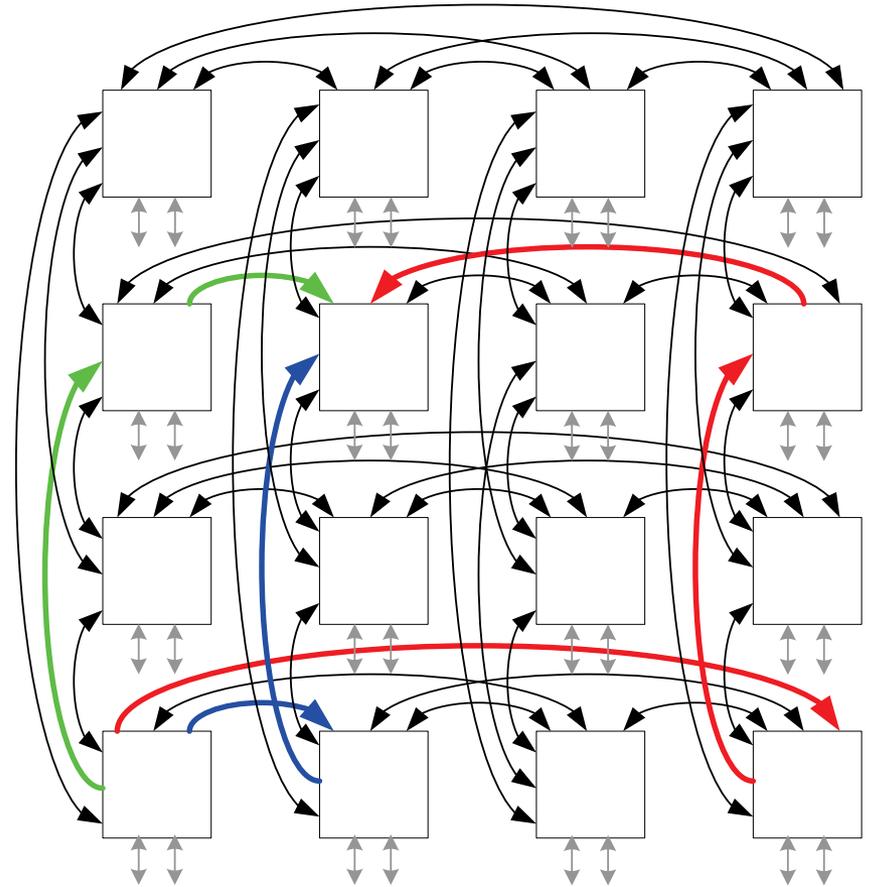
Advantages

- Switch size unconstrained by device IO limits
- Port bandwidth scalable by increasing number of wavelengths
- Optical link ports can directly connect to anywhere within the data centre
- Greatly increased connector density, reduced cable bulk

HyperX networks

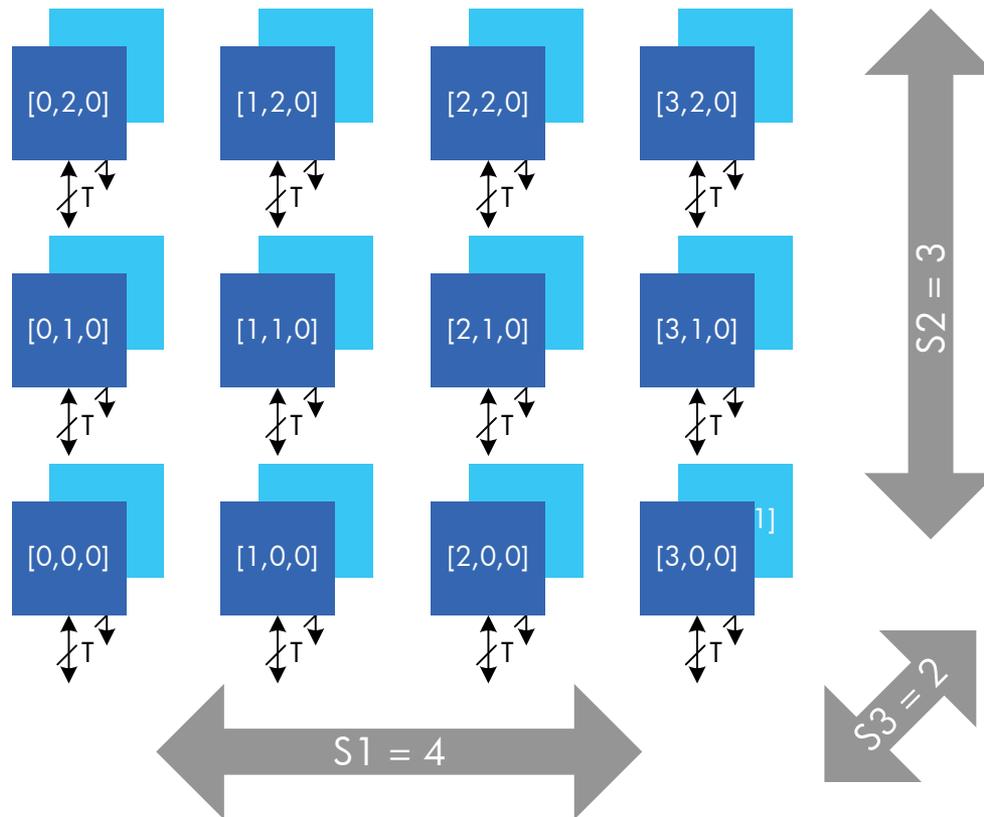
Fully connected sub-networks in multiple dimensions

- Superset of “flattened butterfly” networks¹ and hypercubes
- Fully connected networks offer lowest hop count but limited scalability
- Multiple dimensions increase scalability at the expense of hop count
- Many alternate paths with one or more additional hop
- Non-minimal routes required for full bisection bandwidth



1. Kim, Dally, Abts ISCA2007

General HyperX Structure (L,S,K,T)



- L is an integer number of dimensions
- S is an L-length integer vector specifying how many switches there are in each dimension
- K is an L-length rational vector specifying the relative bandwidths of links in each dimension
- T is the number of terminals attached to each switch

Switch connectivity

- Each switch is fully connected to its peers in each dimension

128K network using 128 port switches

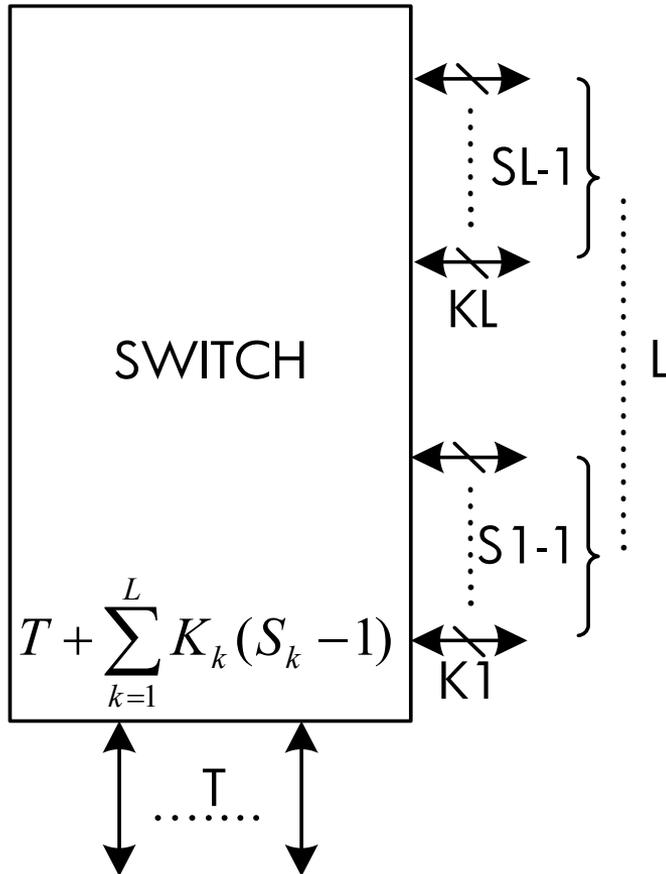
Regular HyperX

B	L	S	K	T	Switches
0.125	4	7	2	76	2401
0.25	3	14	2	54	2744
0.5	3	16	2	35	4096
1	4	10	3	19	10000

General HyperX

B	L	S	K	T	Switches
0.125	3	(5, 19, 19)	(4, 1, 1)	76	1805
0.25	3	(3, 27, 30)	(9, 1, 1)	54	2430
0.5	3	(3, 35, 36)	(12, 1, 1)	35	3780
1	3	(5, 38, 38)	(8, 1, 1)	19	7220

Switch port usage



- Each dimension requires switch ports $K_k (S_k - 1)$
- The total number of ports per switch:
- For non blocking behaviour, in every dimension $K_k S_k \geq 2T$
- The relative bandwidth of a network is:

$$\beta \equiv \frac{\min(K_k S_k)}{2T}$$

- In a regular HyperX:
 - K_k is the same for all k ;
 - S_k is the same for all k

Finding an optimal HyperX configuration

Need at least N ports:

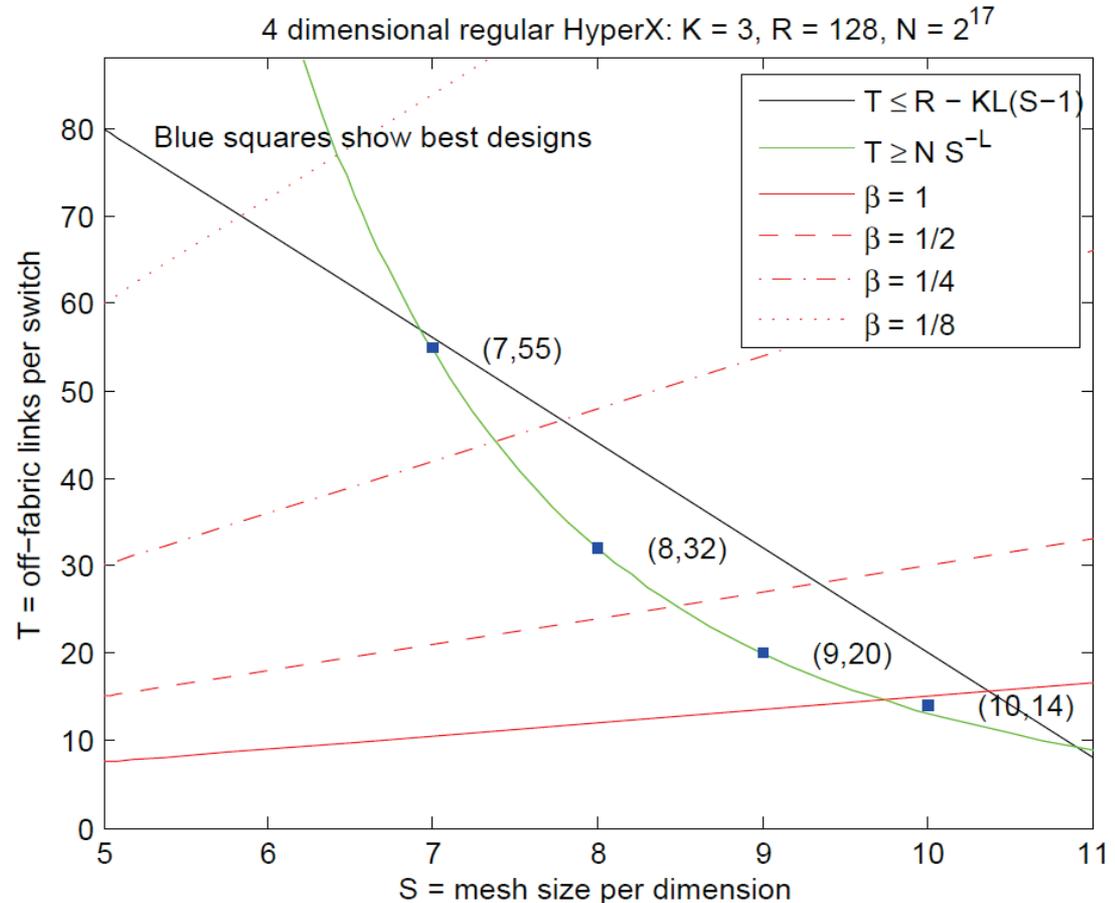
$$T \prod_{k=1}^L S_k \geq N$$

Size of switch R is fixed:

$$T + \sum_{k=1}^L K_k (S_k - 1) \leq R$$

Target bandwidth is β :

$$K_k S_k \geq 2\beta T$$



Now search the design space in L, S & K

Other constraints can be added e.g. $K \geq 2$ (redundant paths)

Dimension, Adaptive, Load balanced routing

- Initially mark offset dimensions as available for “deroute”
- Find an offset dimension that is unblocked
 - If available go there
- If none available: find an unblock switch in an offset dimension available for “deroute”
 - If available take that path
 - Mark this dimension as NOT deroutable
- If none available
 - push the packet into a minimal, dimension order, deterministic VC

128K ports

$L=3$

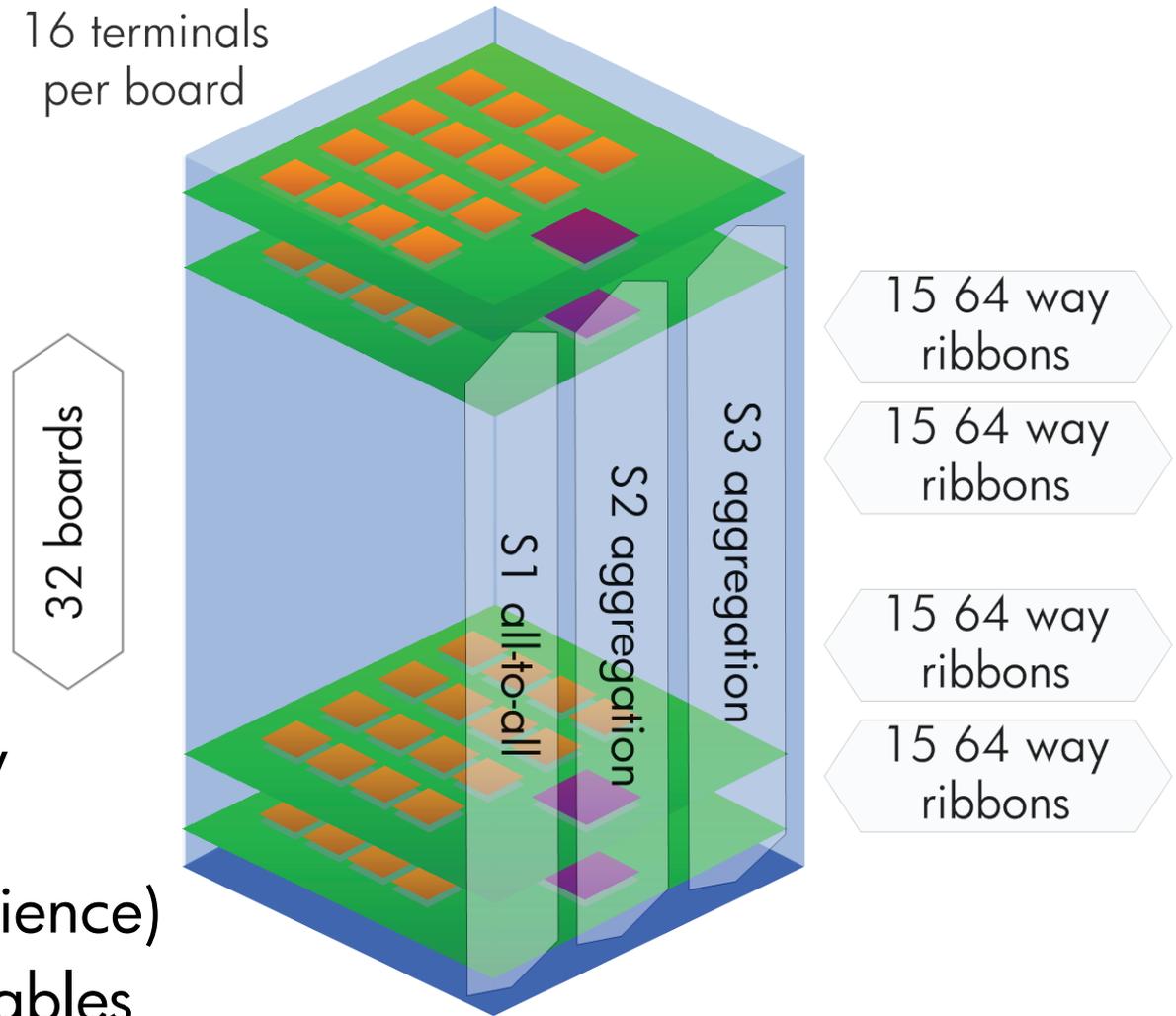
$S=(32,16,16)$

$K=(1,2,2)$

$T=16$

- $2 \times 32 \times 2 = 128$ way parallel ribbons
- (or 2×64 for resilience)
- $2 \times 15 \times 2 = 60$ cables per enclosure

16 terminals
per board

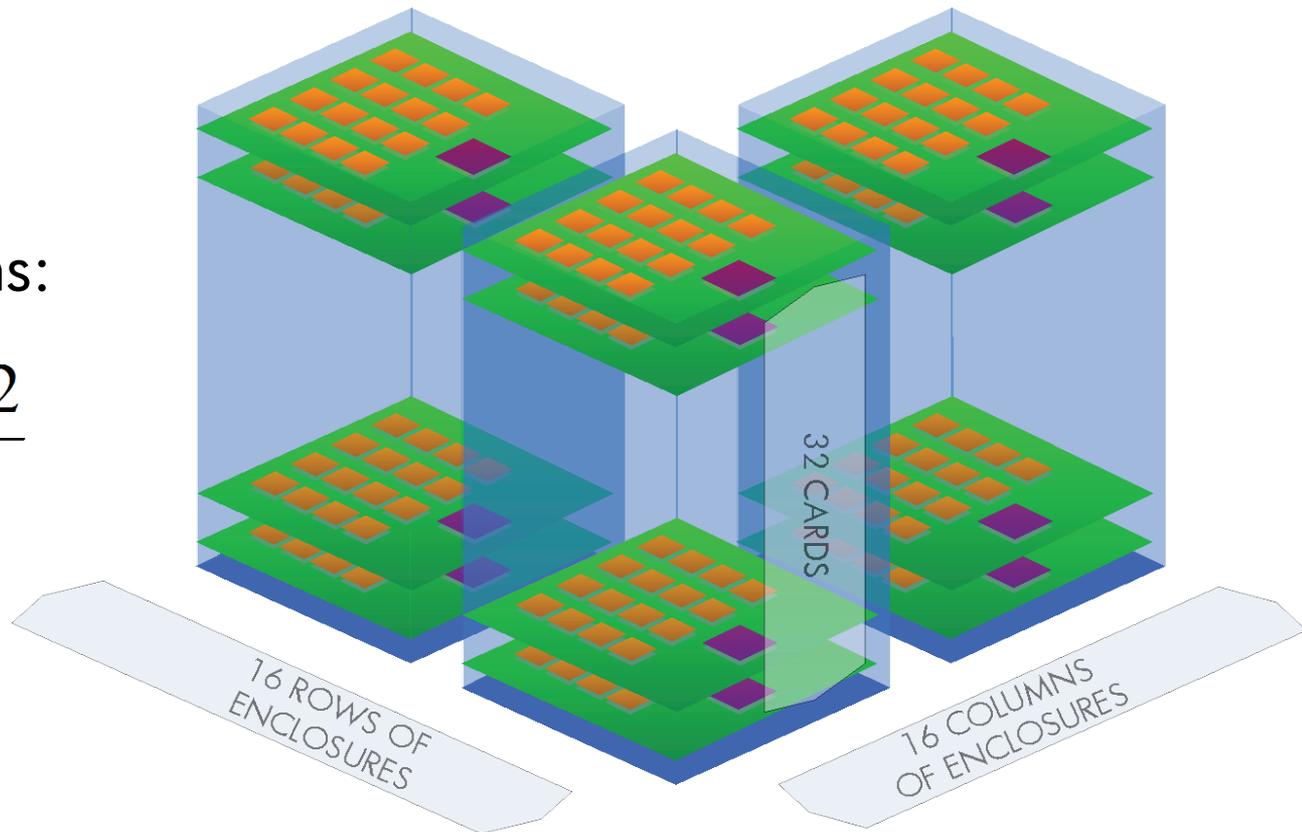


128K node system...

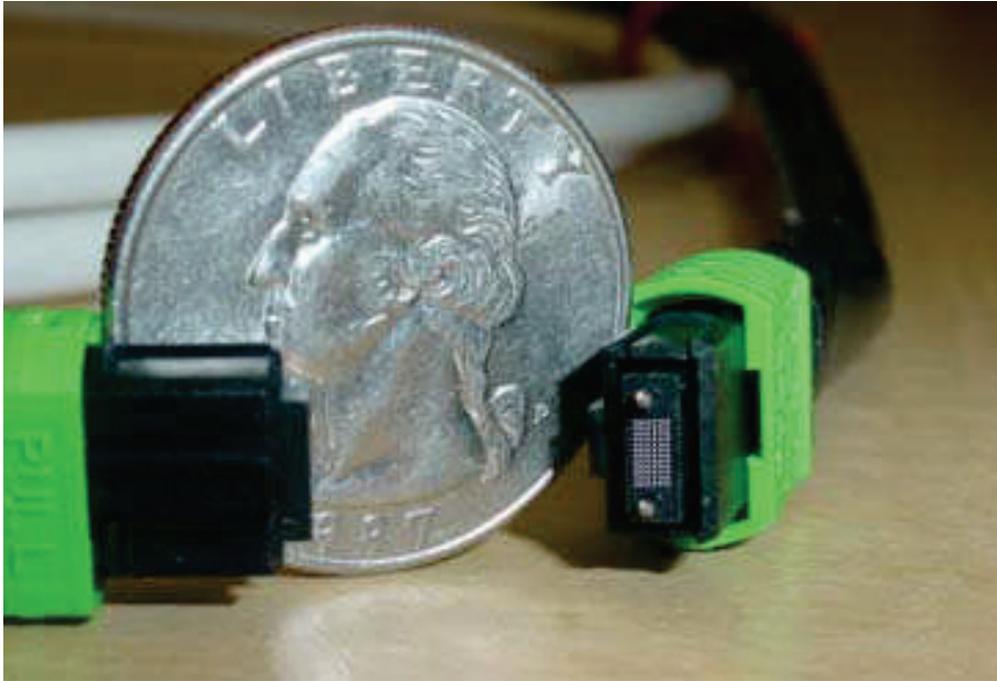
16 x 16 array of enclosures

Total fiber ribbons:

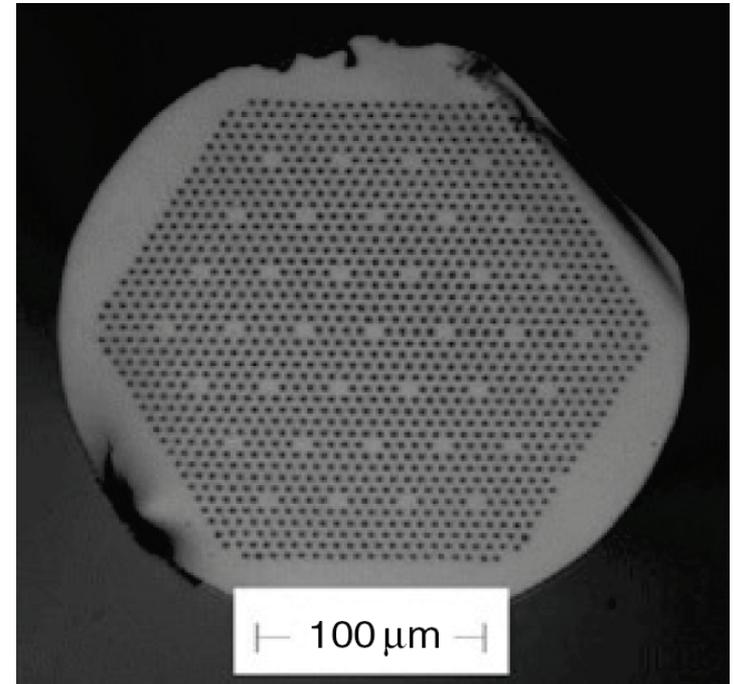
$$\frac{16^2 \times (15 + 15) \times 2}{2} = 7680$$



2 interesting fiber technologies...



72 way parallel fiber ribbon



37 core photonic crystal fiber (PCF)

Conclusions

- General HyperX combines the implementation advantages of a direct network with the network properties of logarithmic networks
- DAL routing provides highest throughput on HyperX networks under all traffic loads
- Characteristics of HyperX are well suited to emerging photonic technologies

Fault tolerance



Exascale MPP Fault Resiliency

- Exascale systems will have hundreds of thousands of CPUs, with reliability a key metric.
- HDD-based checkpoint/rollback scheme - the most widely-used method for MPP fault tolerance – is not scalable
- Example:
 - BlueGene/L – 500 teraflop
 - 12 minutes to take a checkpoint
 - Application execution must be suspended during checkpointing
 - 8% performance overhead with checkpointing interval of 2.5 hours

Local/Global Hybrid Checkpoint

- Two-level checkpoints: local and global
- Global – System-wide, for real “hard error”
- Local – Node-wide, for SEU, “soft error”

	Hardware	Software	Network	Environment	Human	Unknown
% Breakdowns	62%	18%	2%	1%	1%	16%

(Root Causes of failures in Terascale system)

Protected by local checkpoints

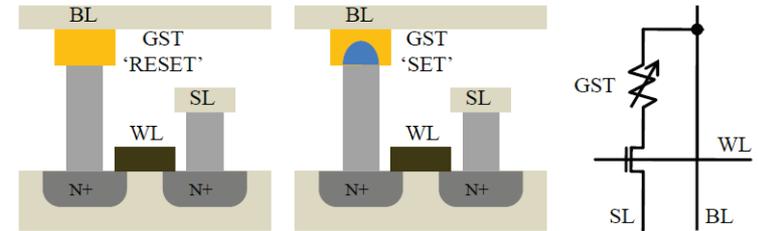
Protected by global checkpoints

Most of hardware errors are “soft error”

- Scaled to Exascale, the percentage of SEU becomes even larger.
- Most failures can be recovered by local checkpoints.

PCRAM

- PCRAM is much faster than Flash and HDD.
- PCRAM has 1000X longer life time than Flash.
- Use of PCRAM for checkpoints: Collaboration with Yuan Xie and Xiangyu Dong of PSE

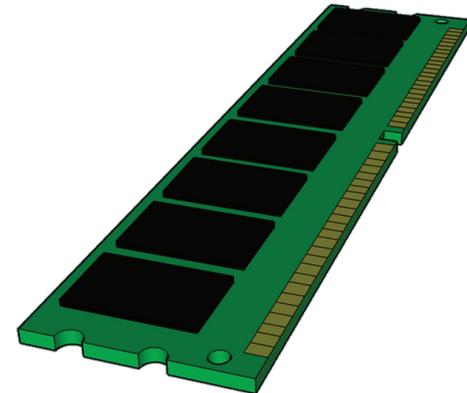


The schematic view of a PCRAM cell with NMOS access transistor (BL=Bitline, WL=Wordline, SL=Source/Drain line)

	HDD	NAND Flash	PCRAM
Cell size	-	4-6F ²	4-6F ²
Read time	~4ms	5us-50us	10ns-100ns
Write time	~4ms	2ms-3ms	100-1000ns
St'by power	~1W	~0W	~0W
Endurance	10 ¹⁵	10 ⁵	10 ⁸

Option 1: PCRAM DIMM

- Deploy PCRAM checkpoints in the form of memory DIMM.
- Connect to DDR memory bus.
- High bandwidth:
 - 10GB/s (DDR3-1333)
 - 32GB/s (projected bandwidth in Exascale system)



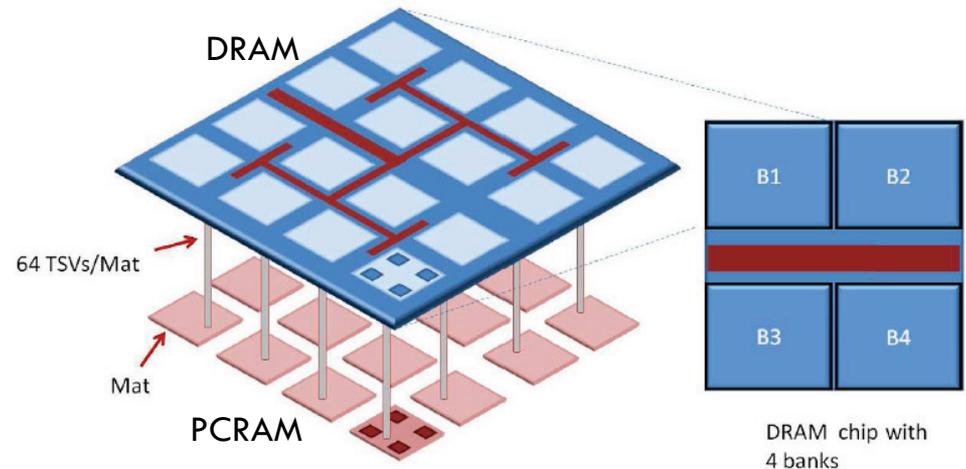
Option2: 3D PCRAM

– Stack PCRAM on top of DRAM

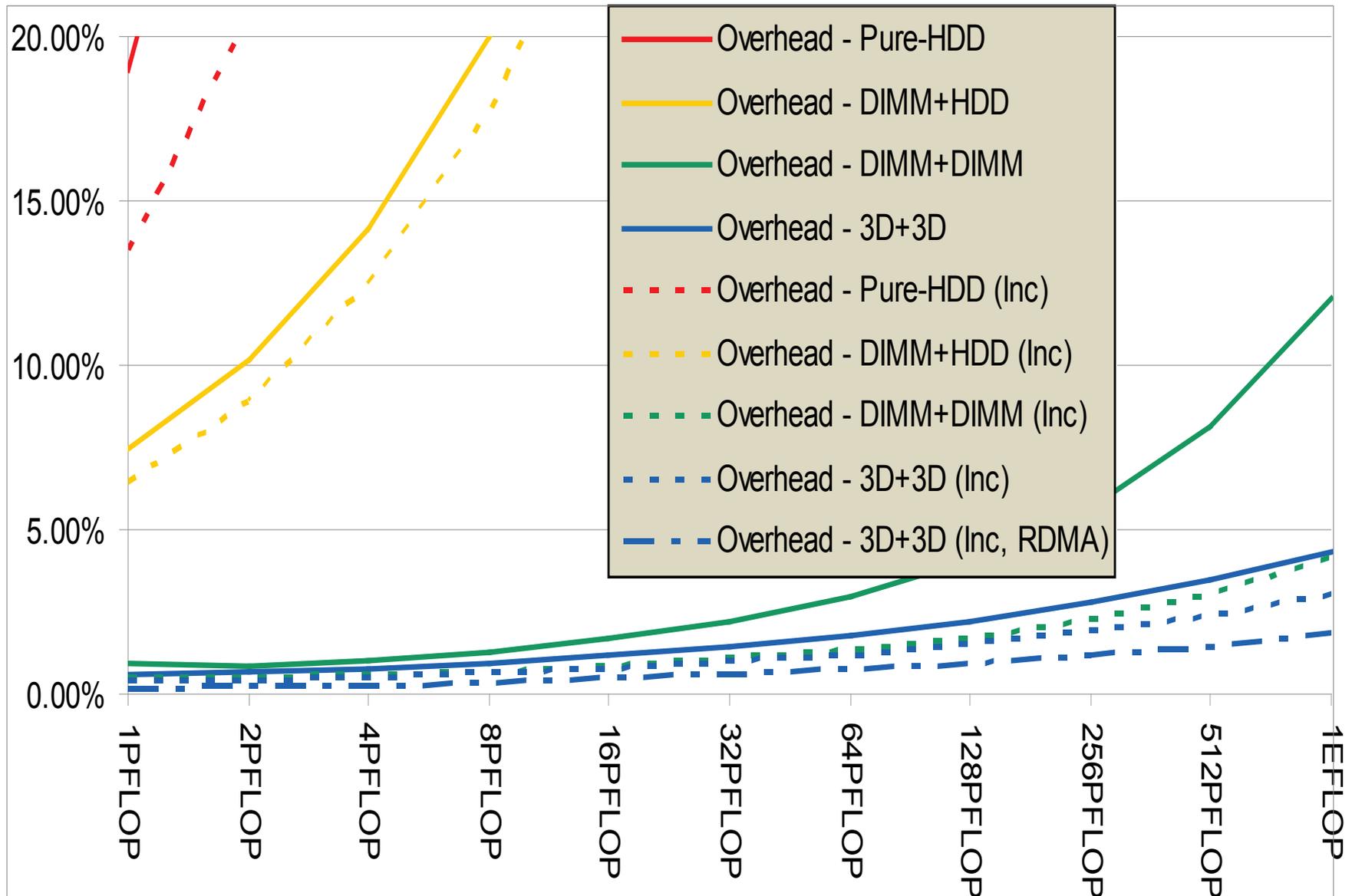
- Connecting DRAM and PCRAM mats using TSV
- Extra-high bandwidth
- Instant checkpoint (~0.8ms)

– Some design issues

- Power spike during PCRAM writing
- Temperature
- Design complexity

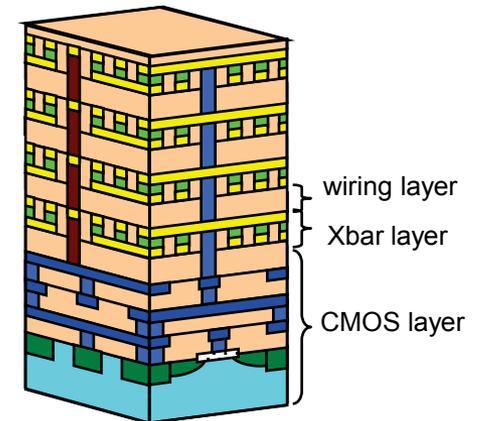


Scalability Trends



Memristors

- Similar to RRAM, but includes diode
 - Allows $4F^2$ cell
- 5nm junctions demonstrated
- Can fab with multiple layers
- Die can be stacked too
- → Very high storage density
- TiO_2 junction material
 - White paint, sunblock



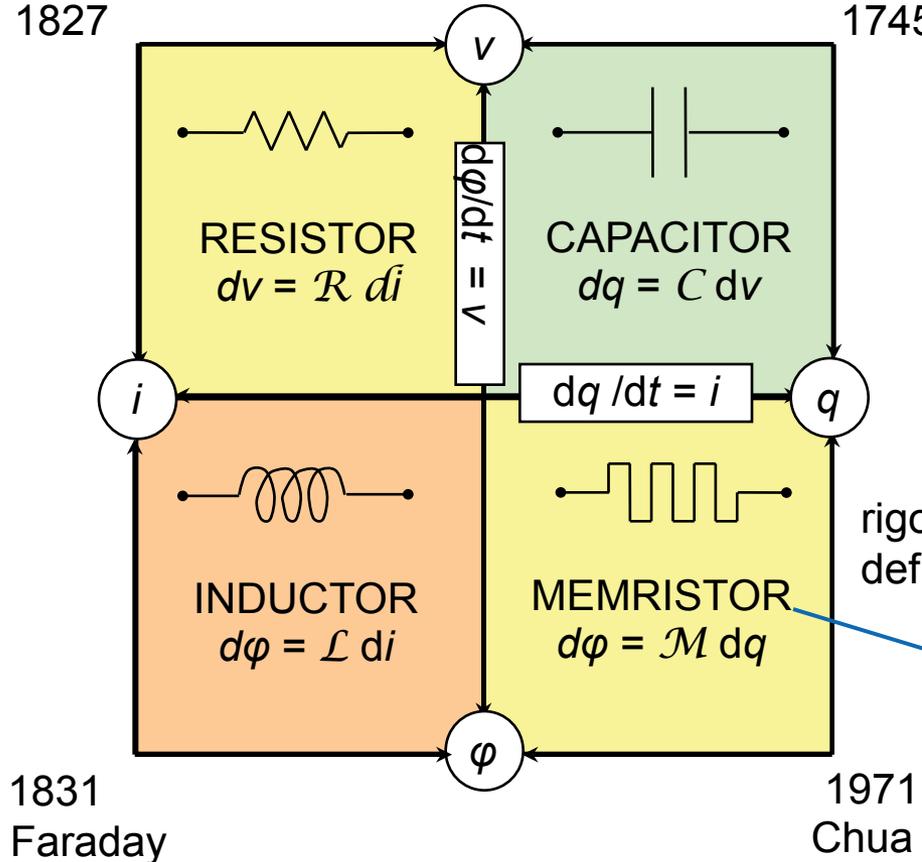
The Prediction of a New Circuit Element: the Memristor

Ohm
1827

Von Kleist
1745



L. O. Chua, *IEEE Trans. Circuit Theory*
18, 507 (1971)



rigorous
definition

$$v(t) = R[w, i(t)]i(t)$$

Quasi-static conduction eq.-
 R depends on state variable w

$$\frac{dw(t)}{dt} = f[w, i(t)]$$

Dynamical equation –
Evolution of state in time

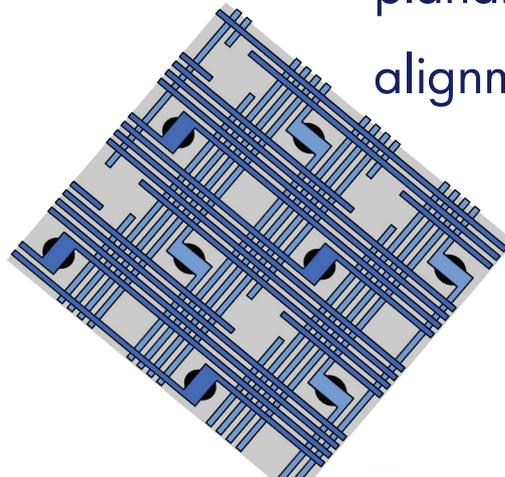
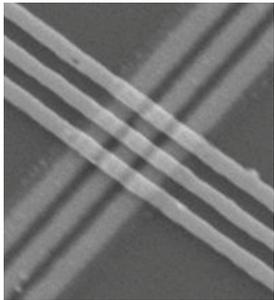
First Hybrid CMOS-Memristor Chip

Issues that had to be overcome:

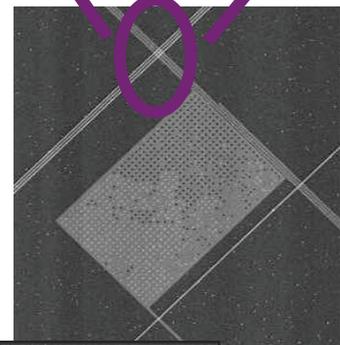
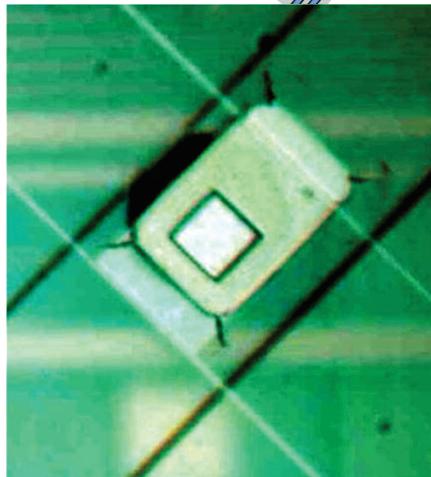
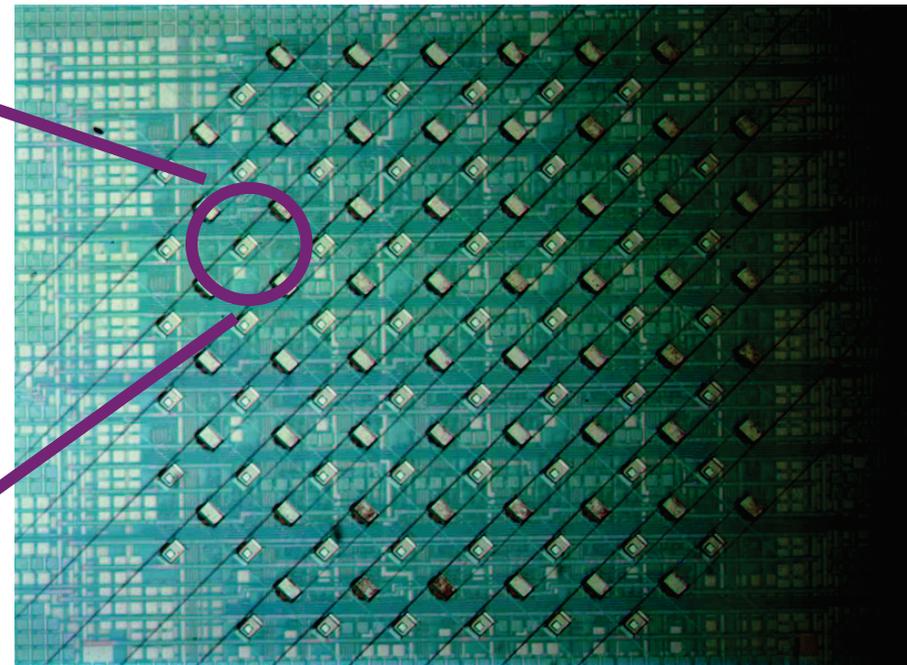
planarity

alignment of fine features

3x3 100nm nanowire
Crossbar junctions



CMOS chip with memristive devices



10 μ m
5E overview

Connecting the CMOS layer with the nanowire crossbar junctions

Questions People Ask about Memristors

- Switching voltages/currents (1- 2 Volts/ $<20\mu\text{A}$)
- Write/Erase/Read speed and energy ($<10\text{ns}$, $\sim\text{pJ}$)
- Data retention time (years – even millenia?)
- ON/OFF ratio ($>1000:1$); allows multiple bits
- Scaling limits ($<5\text{nm}$ - >1 terabit/sq cm/layer)
- Endurance $>10^5$ cycles
- Failure mechanism (wire electromigration)
- Nature of ON and OFF states (metal/insulator)
- Devices are evolving rapidly

Questions?

