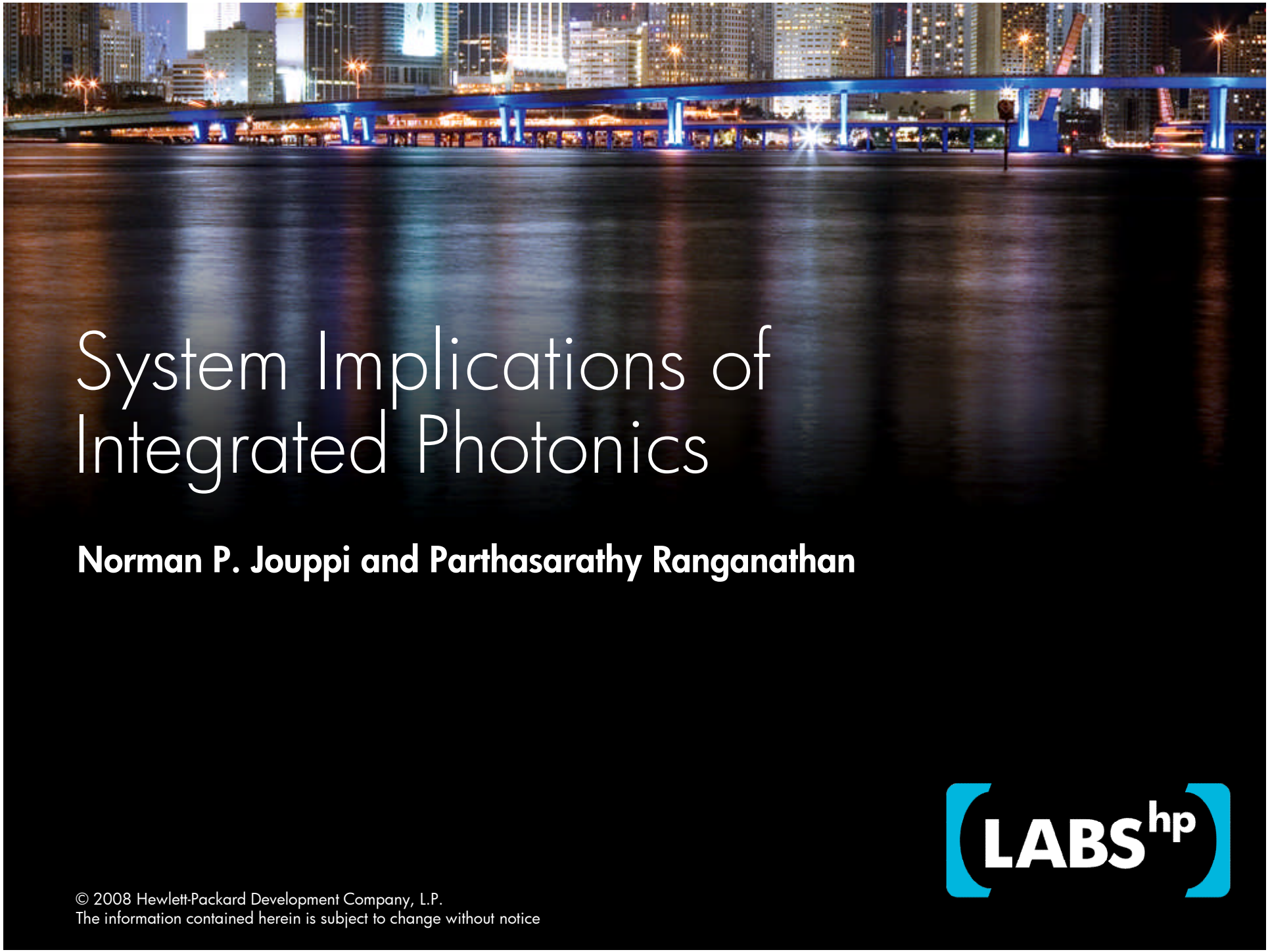


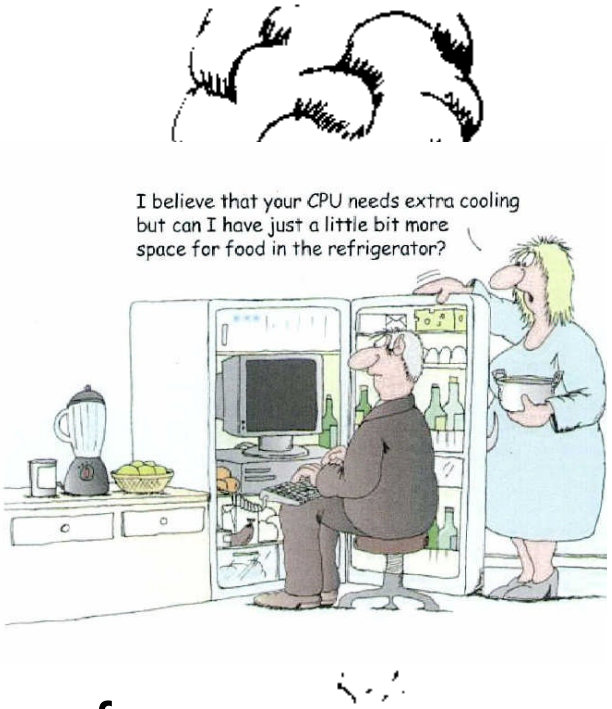
# System Implications of Integrated Photonics

**Norman P. Jouppi and Parthasarathy Ranganathan**

# Today's talk

- Introduction (Partha)
- Nanophotonics and Capabilities (Norm)
- Potential Impact and Early Results (Norm)
- Some System Implications (Partha)

# Low Power important in all markets





- from processors to data centers
- from handhelds to supercomputers



LABS hp

# Large body of prior work

Energy-efficient technology

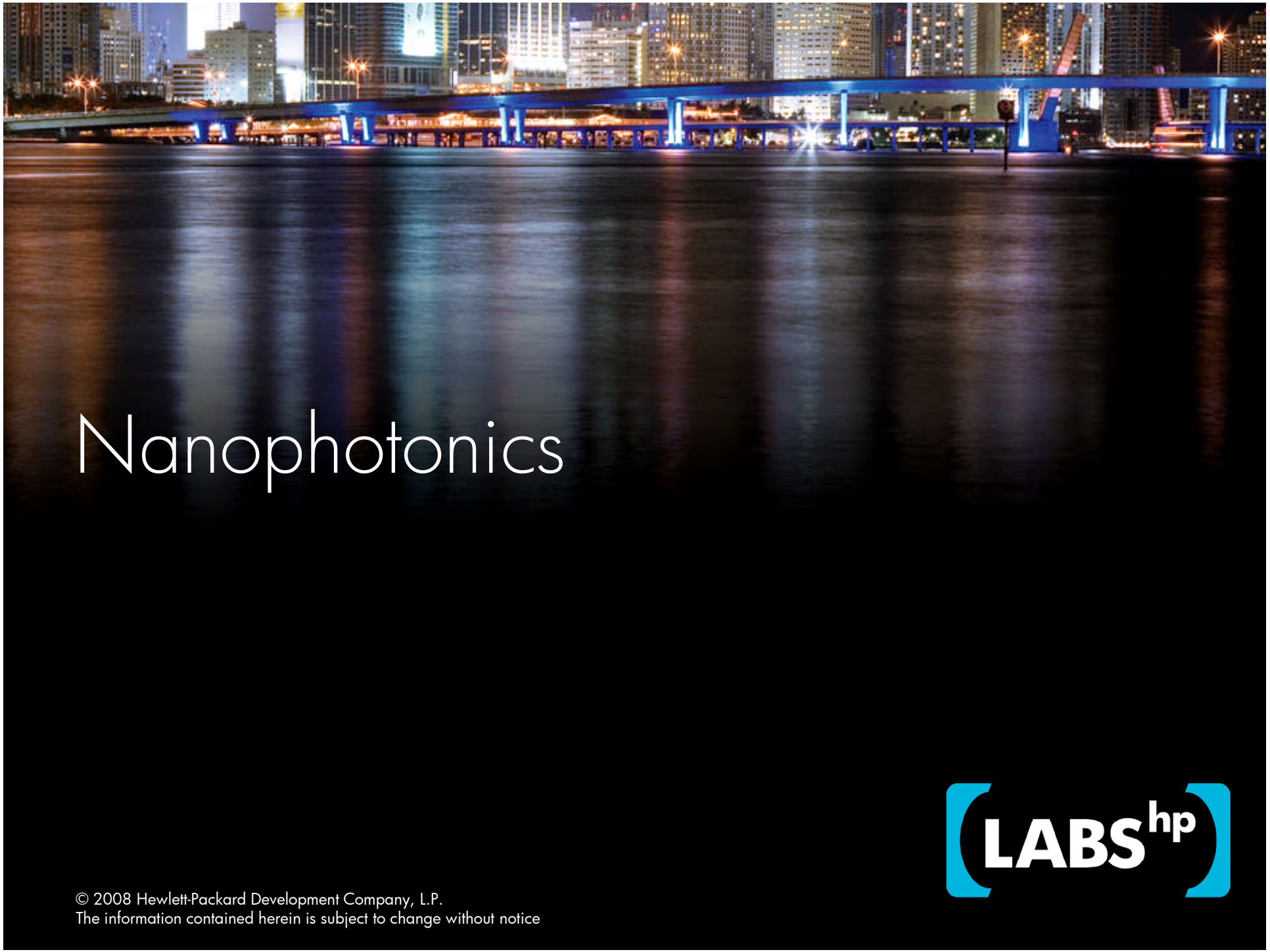
Energy-efficient resource mgmt

# Today's talk

Integrated photonics

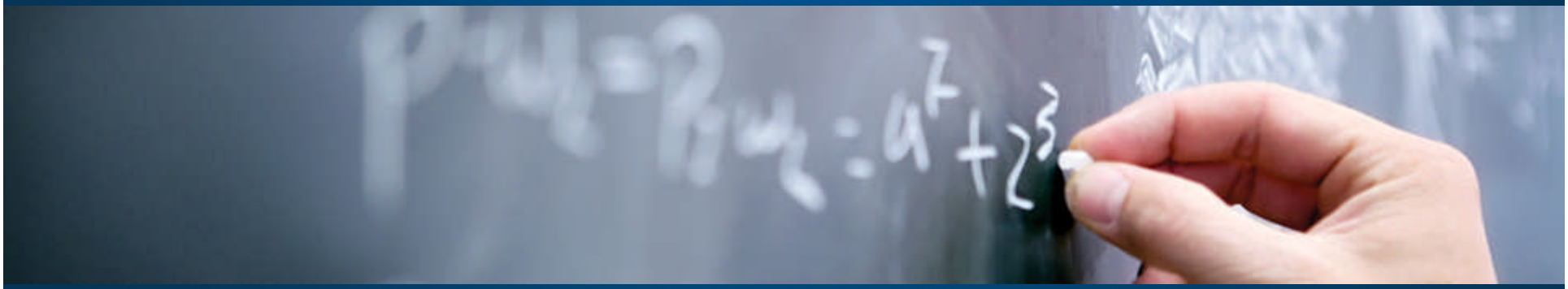
Disaggregated datacenters

LABS hp



# Nanophotonics

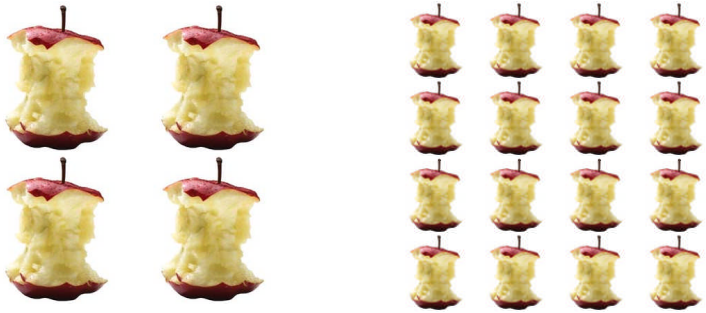
# Current Trends



# Cores Per Die

Past

Present



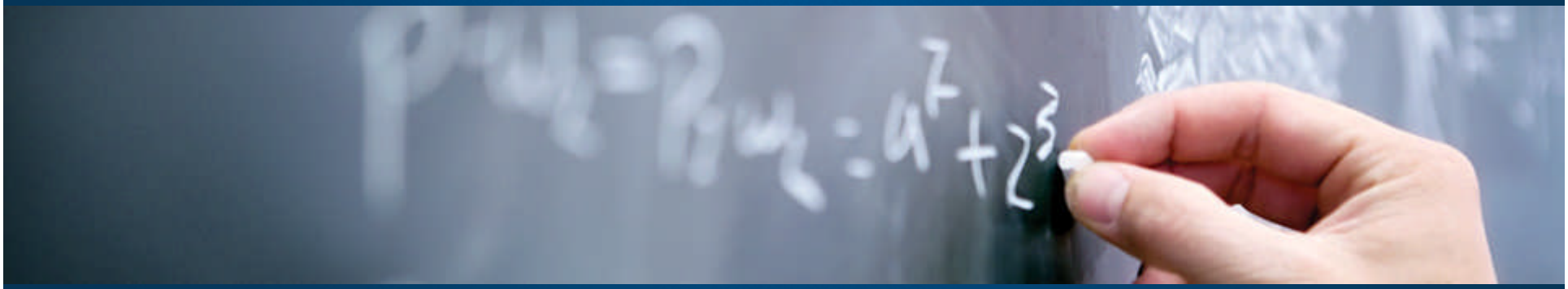
Future

# Speed Bumps on the Road to 2017

- 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

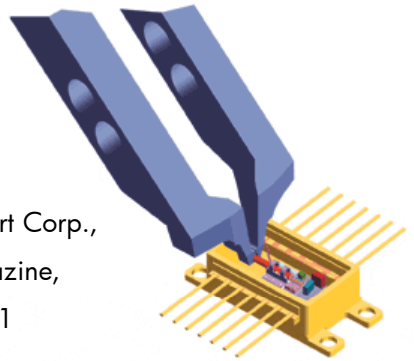
⇒We need a disruptive technology

# Capabilities of Emerging Integrated Photonics



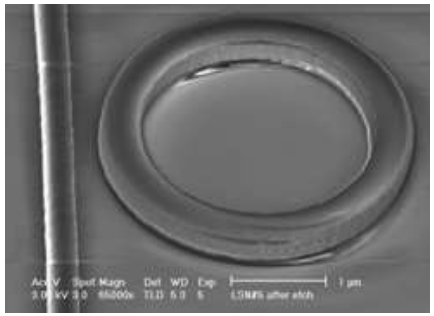
# What are Integrated Photonics?

- The 2000 telecom bubble based on discrete optics
  - Think pre-Noyce/Kilby era in electronics
  - Components are measured in mm
  - Hand alignment
  - Expensive and not scalable



Source: Newport Corp., Assembly Magazine, September 2001

- Recent research is on integrated photonics
  - Think post-Noyce/Kilby era in electronics
  - Components are measured in a few μm
  - Manufacture many thousands per die
  - Advances in lithography yield better devices



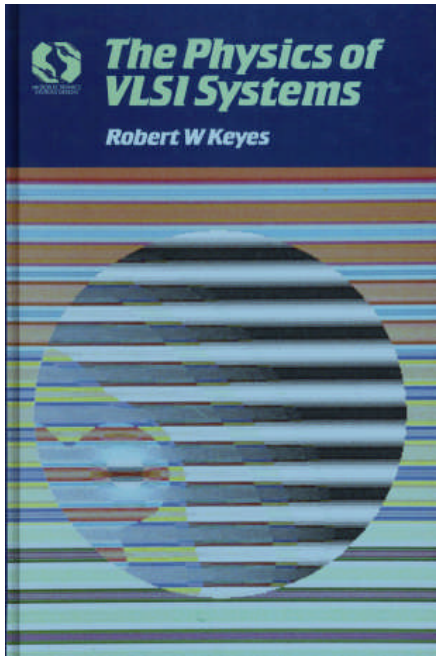


LABS hp

# Important Technology Characteristics

- Several things are important for a successful technology, including:
  - Gain (leading to fan-in and fan-out)
  - Power efficiency
- Reference:

  "The Physics of VLSI Systems" by Robert W. Keyes, 1987.






LABS hp

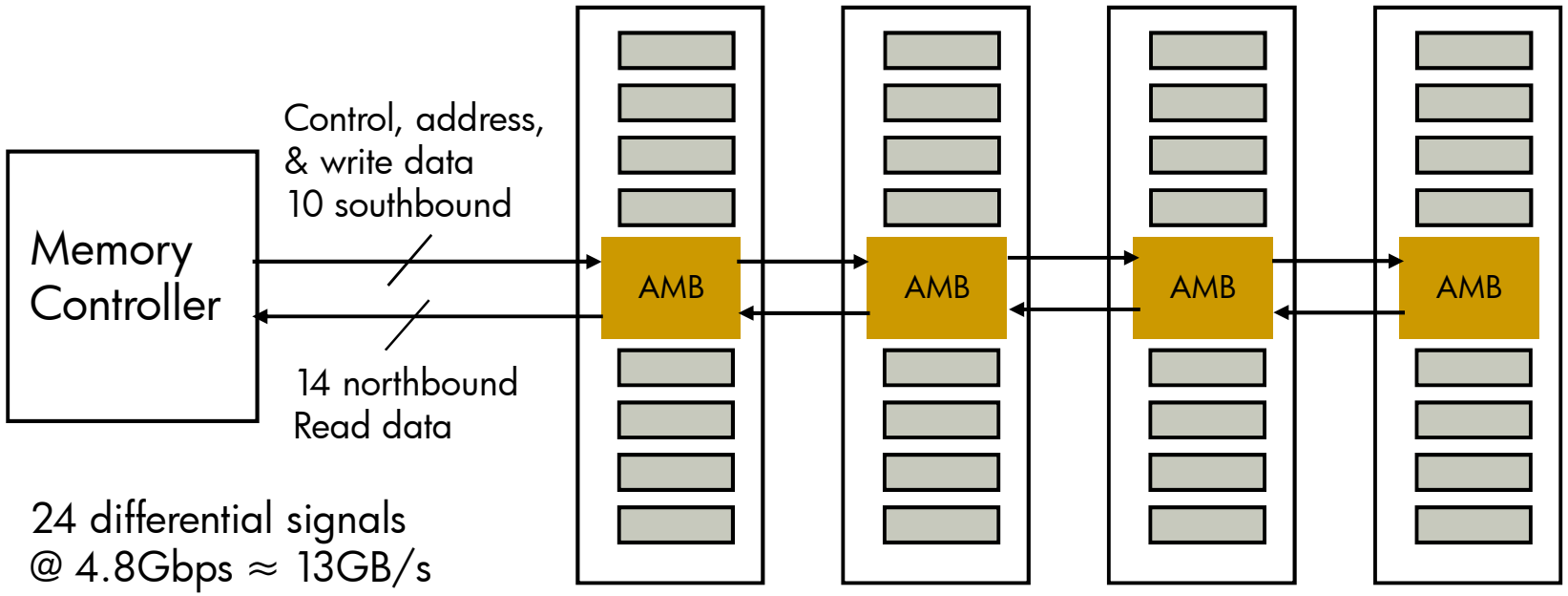
# Gain

- Transistors have good gain
  - Electronics is good for computation
- Photons don't like to interact
  - Photonics is good for long distance communication
  - How long is long?
    - Depends on size -> capacitance -> power of device
    - mm devices: ~30 meters
    - μm devices: ~30 millimeters

# Fan-in and Fan-out

- Important for efficient system design
  - Not economically feasible at signaling rates >2Gbs in electrical systems due to stub problems
  - Possible in optics by using splitters and combiners
- Electrical point-to-point links do not scale well
  - Adds to pin bandwidth limitations
  - Repeated buffering of signals adds delay & power
    - FBDIMM example

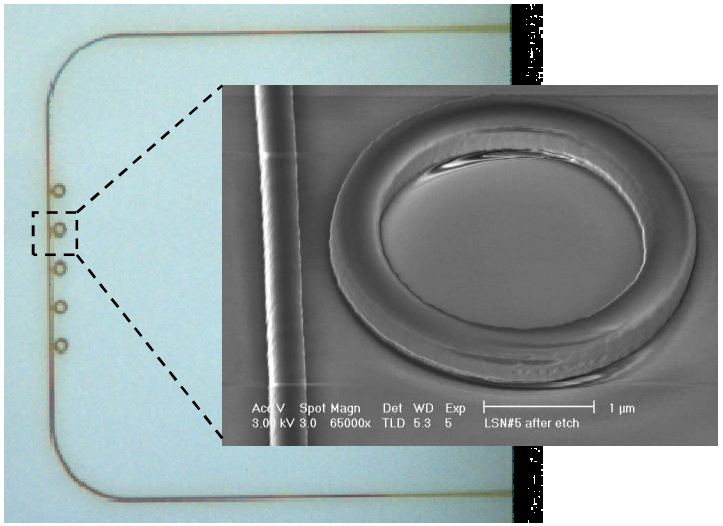
# FBDIMM Memory System

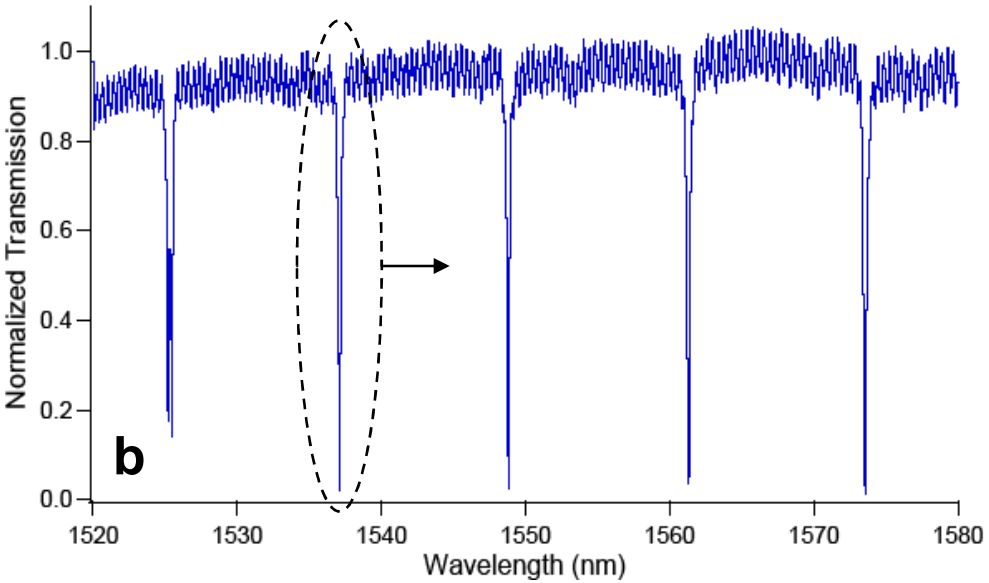


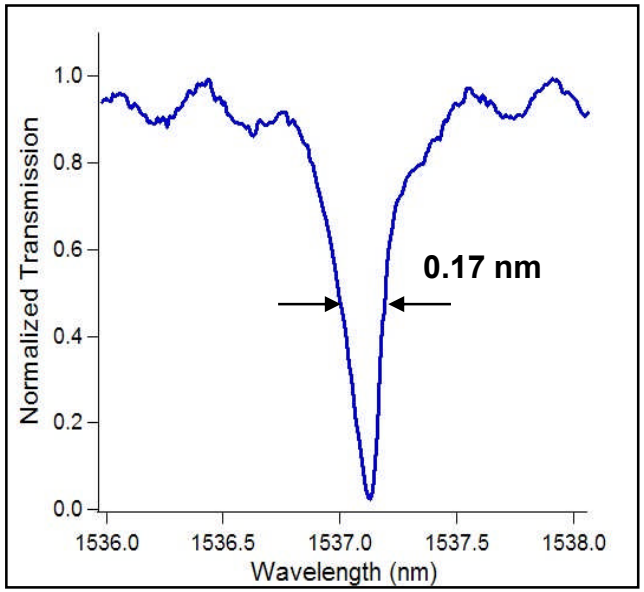

- Latency: Multi-hops
- Power: re-transmission
- Cost

# Si Microrings

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



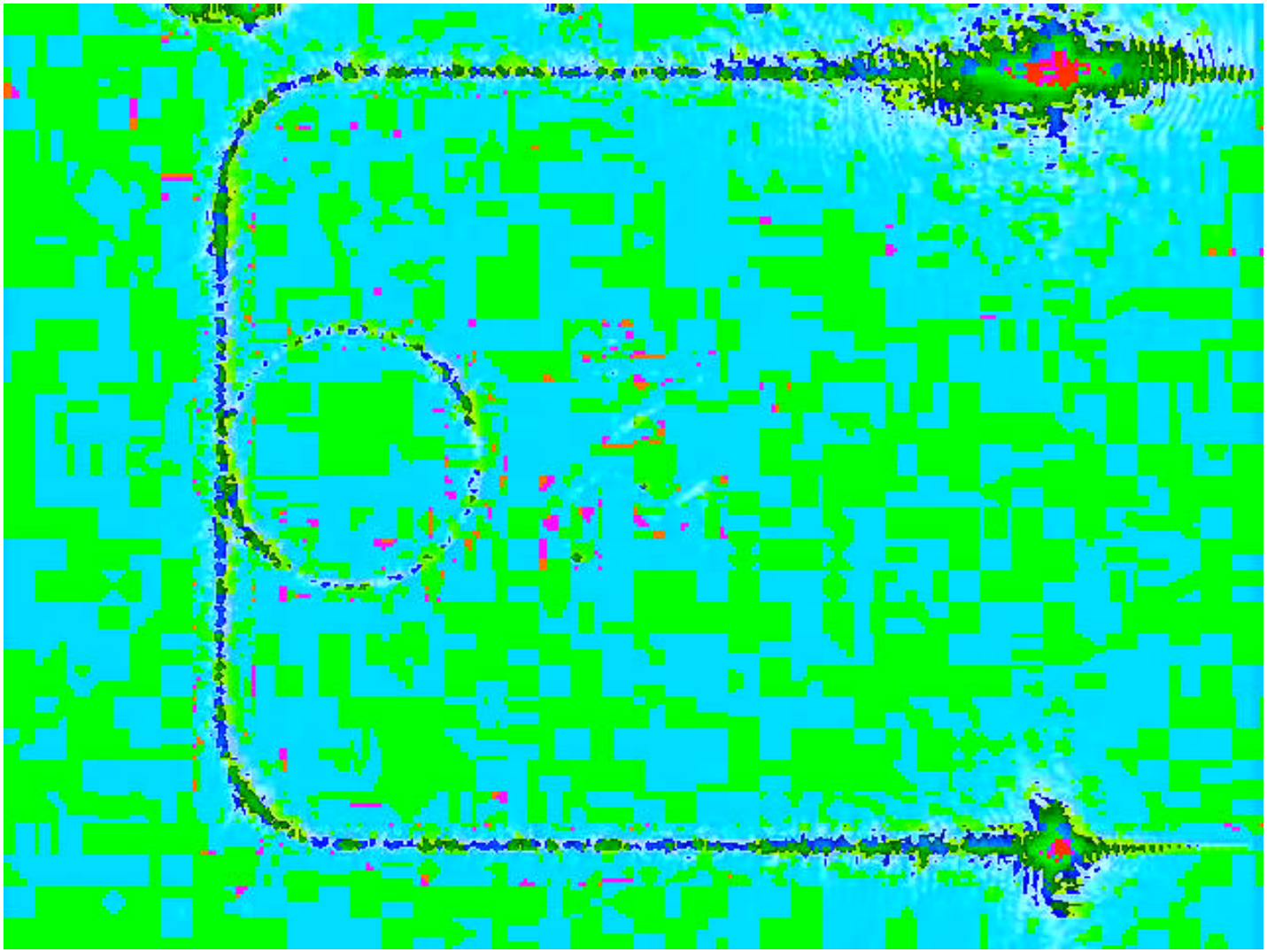




Q. Xu, D. Fattal, and RGB, Opt. Express 16, 4309-4315 (2008) — World Record!



LABS hp

# Si Ring Resonator in Action

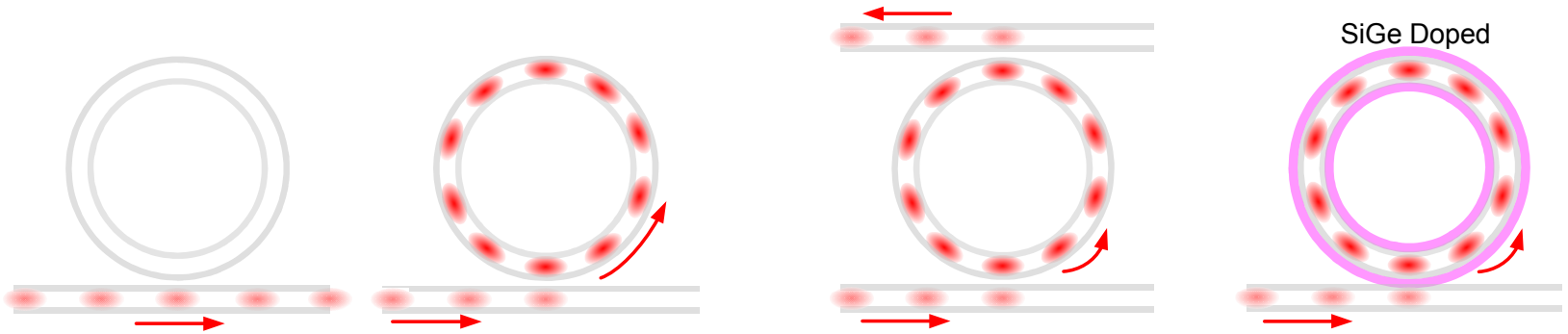




LABS hp

# 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



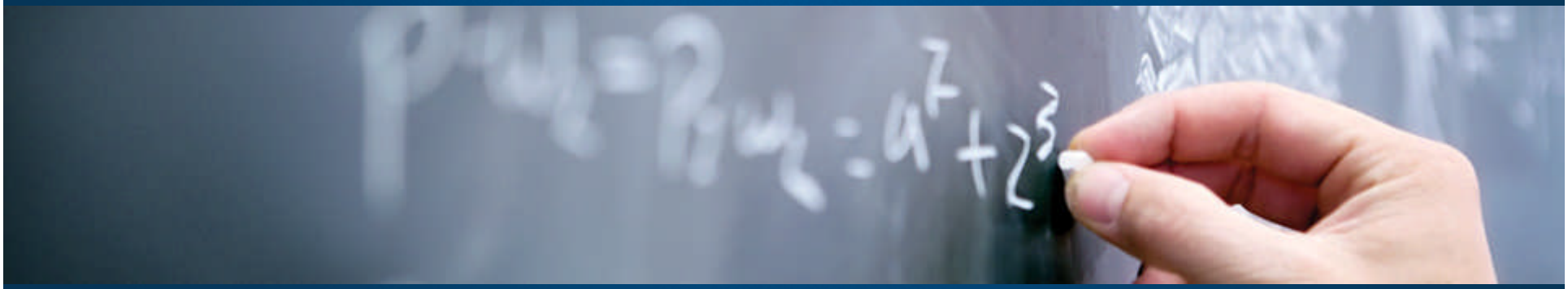
LABS hp

# Power Efficiency

- Hybrid actively mode-locked lasers or comb lasers
  - Produce all wavelengths from a single source
  - Track with temperature
- Si microring modulators
  - Parallel buses with clock forwarding (no SERDES)
  - DWDM: 256 waveguides × 64 wavelengths each = 256 × 64 Xbar
  - Analog drivers for both modulators and detectors (no A/D)
  - Femtofarad-class low-power receiverless detectors

=> Low power 10 Gb/s signaling

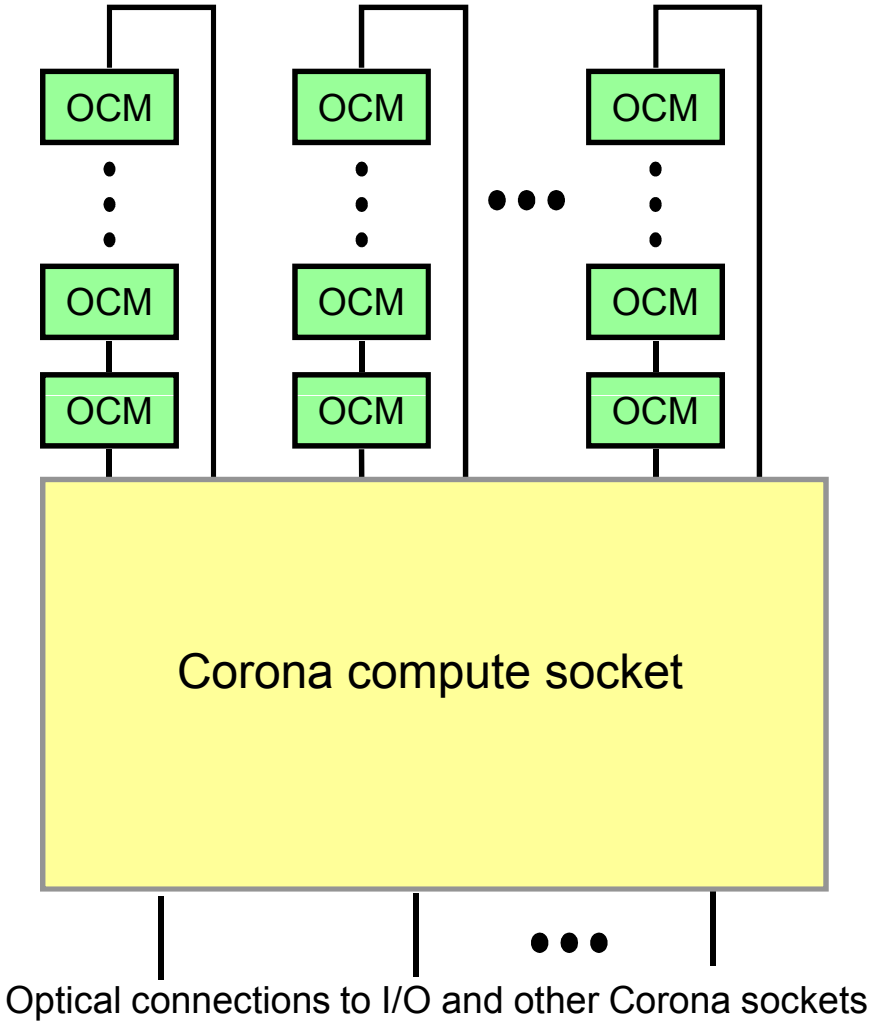
# Potential Impact in 2017



# The *Corona Manifesto*

- Take full advantage of nanophotonics
  - Don't just replace today's wires with optics
  - Redesign the multi-core processor from the ground up
  - No off-chip or cross-chip electrical wires
  - Restore balance: memory bandwidth scales with cores
  - All memory readily reachable from all cores

# Corona System Overview




10 teraflops compute performance

10 terabytes/s memory bandwidth

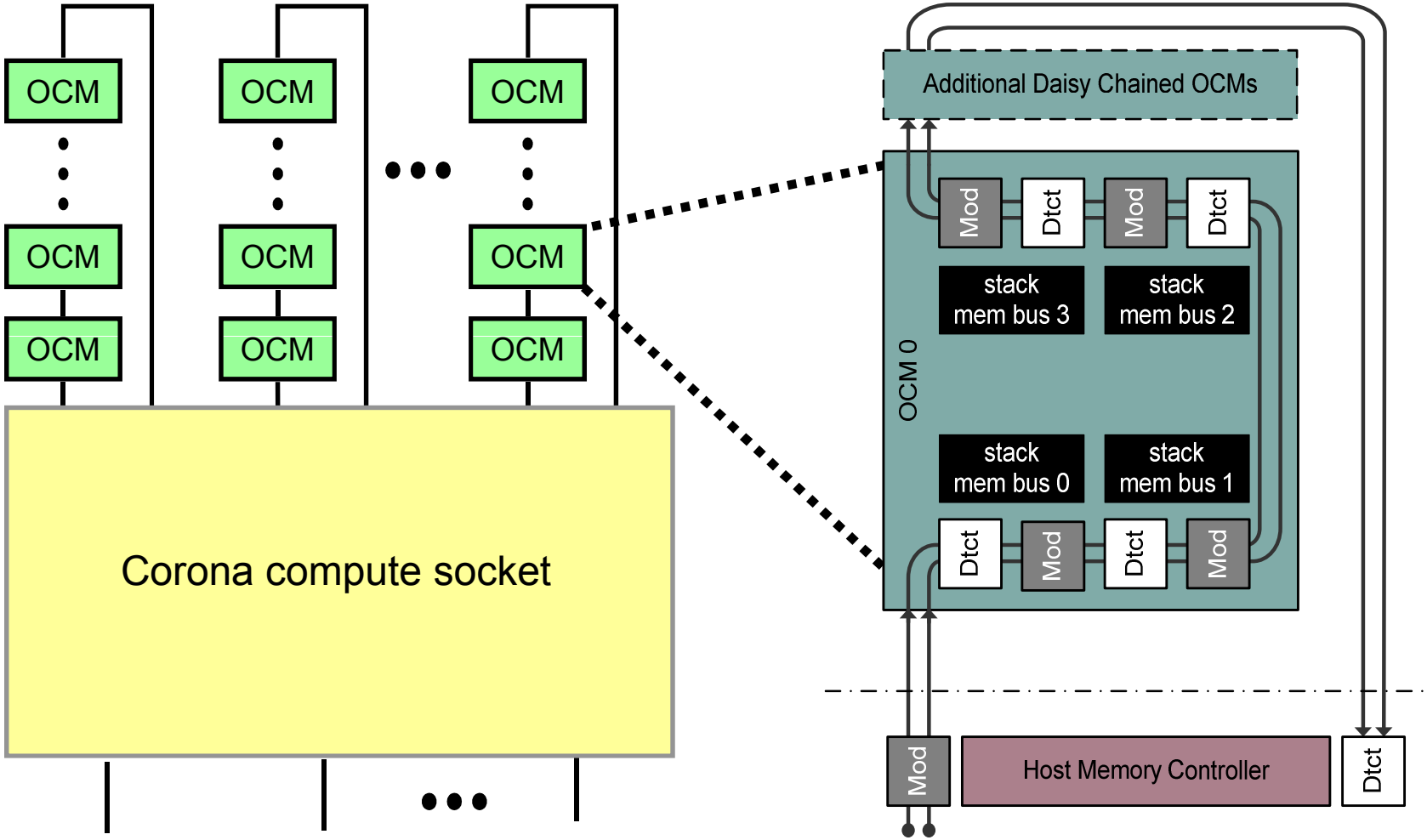
20 terabytes/s on-chip interconnect

All off-socket and cross-socket communication is optical

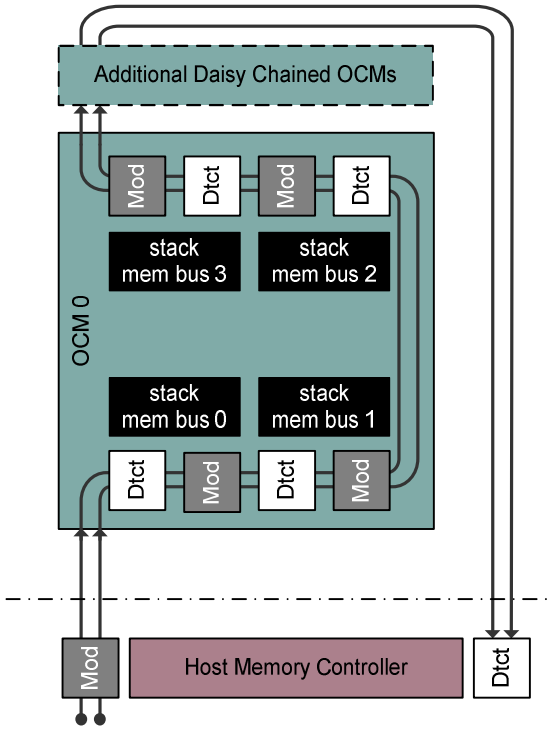


LABS hp

# Optically Connected Memory (OCM)

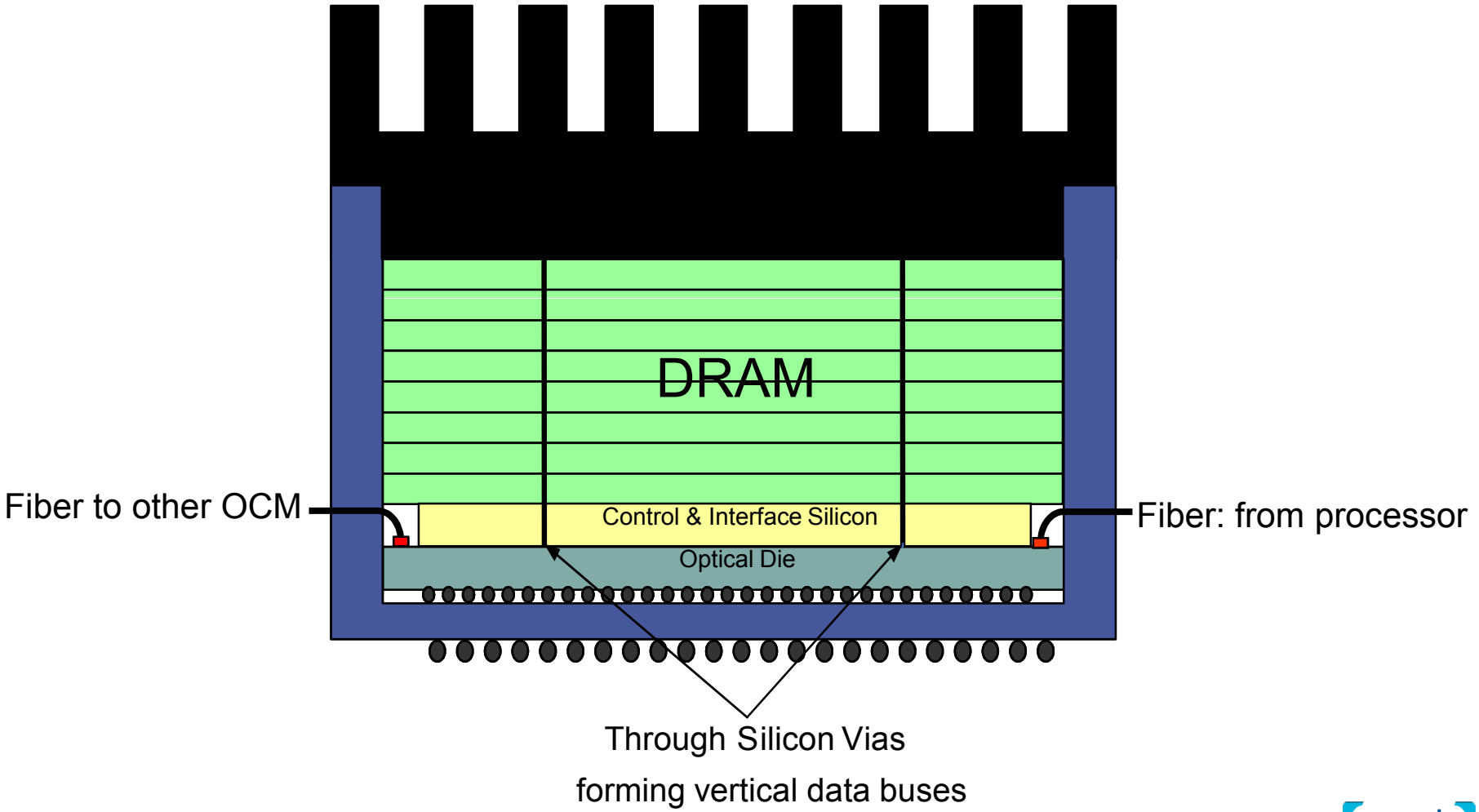


# Optically Connected Memory

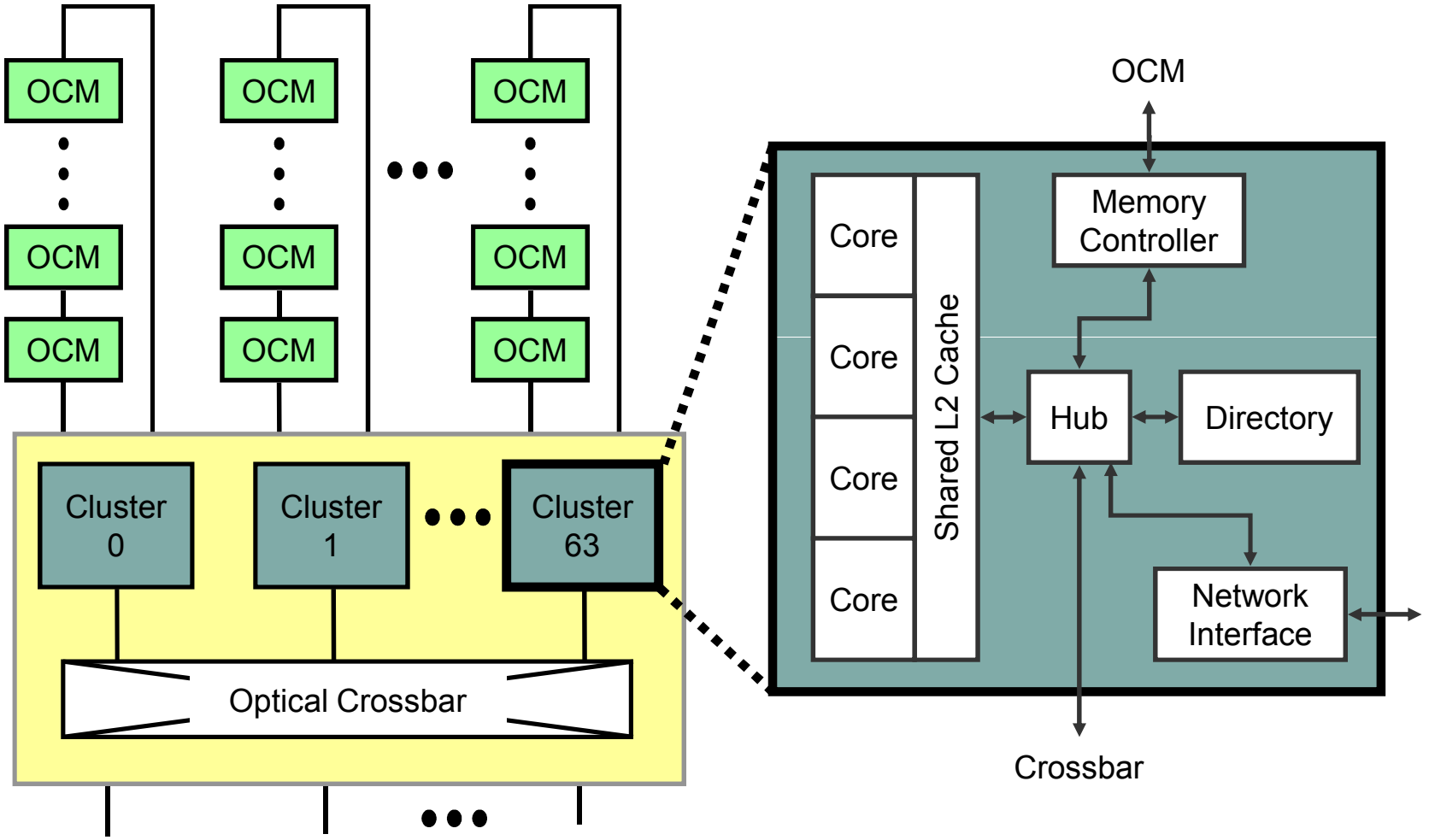



- Master/slave bus on waveguide loop
  - Optical power from processor
  - Processor modulates for data out
  - OCM modulates for return data
- Multiple optical interfaces per chip stack
  - Eliminates electronic global wiring
- OCMs communicate via DWDM
  - High bandwidth
- Accessed in parallel, no receive 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
- High bandwidth at low power

# OCM Chip Stack

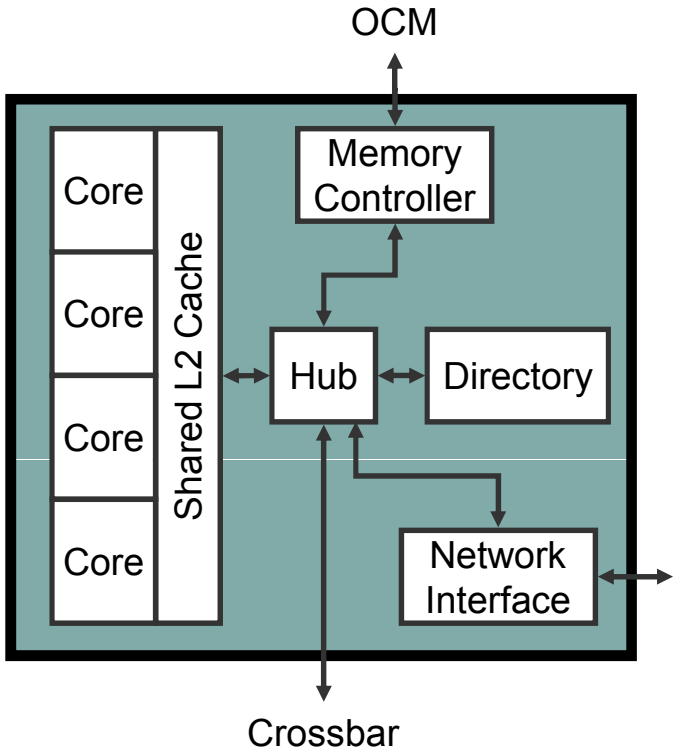


# Corona Compute Socket

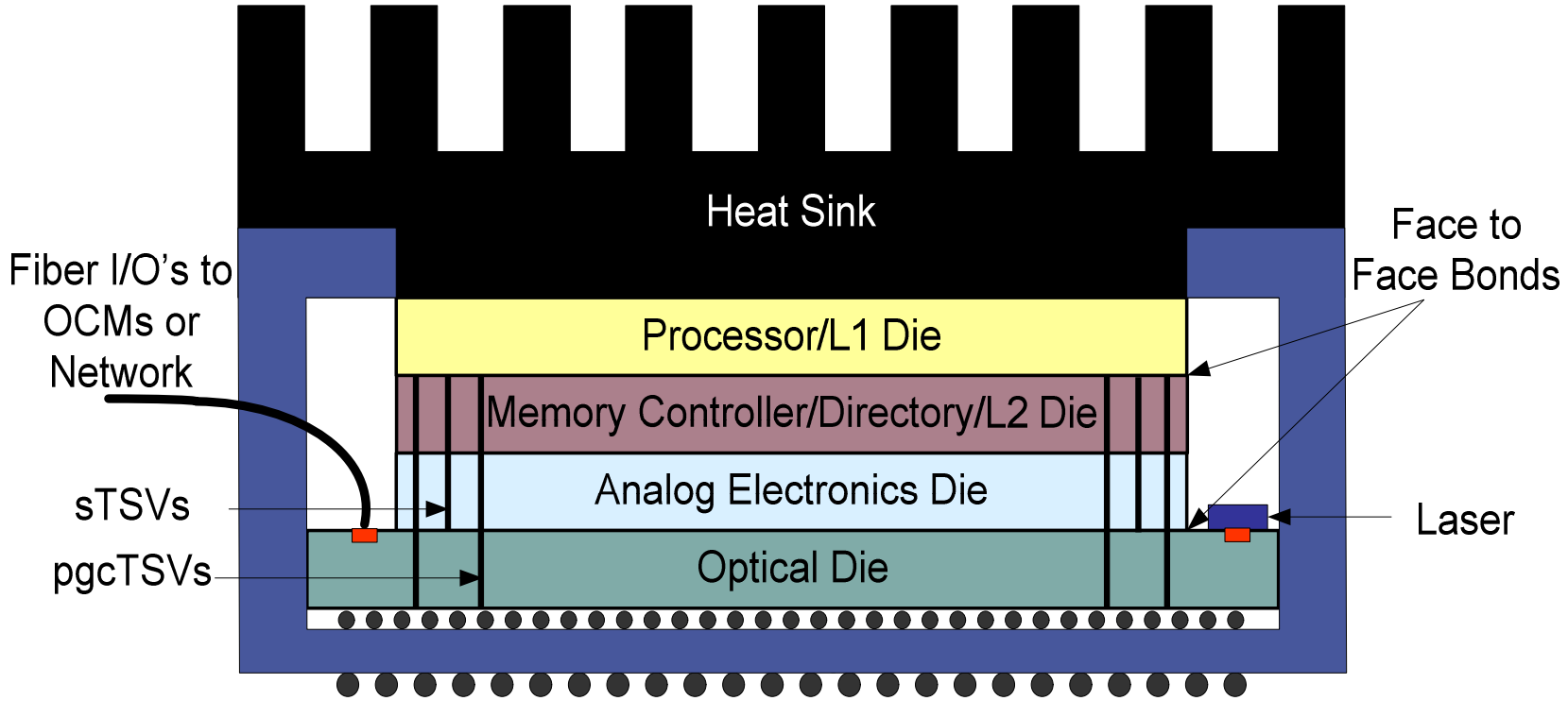


# Corona Cluster Parameters

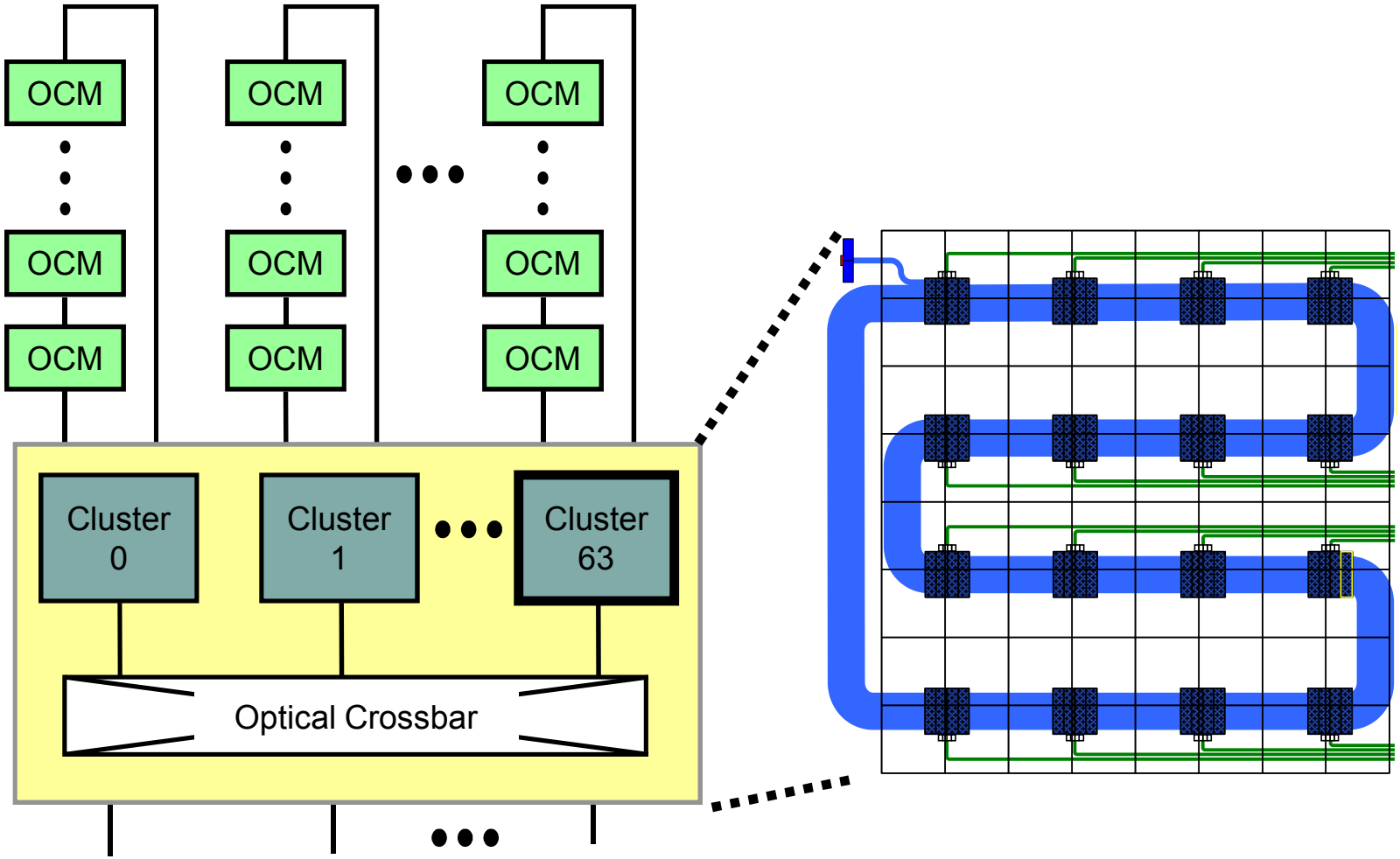
- Per each of 64 clusters:
  - Cores: 4
  - Memory controllers: 1
  - L2 cache: 4 MB, 16-way, 64B lines
- Per-core:
  - Frequency: 5 GHz
  - Threads: 4
  - L1 I-Cache: 16 KB, 4-way, 64B lines
  - L1 D-Cache: 32 KB, 4-way, 64B lines
  - Issue: 2-wide in-order
  - 64 b SIMD FP width 4 + Fused FP operations



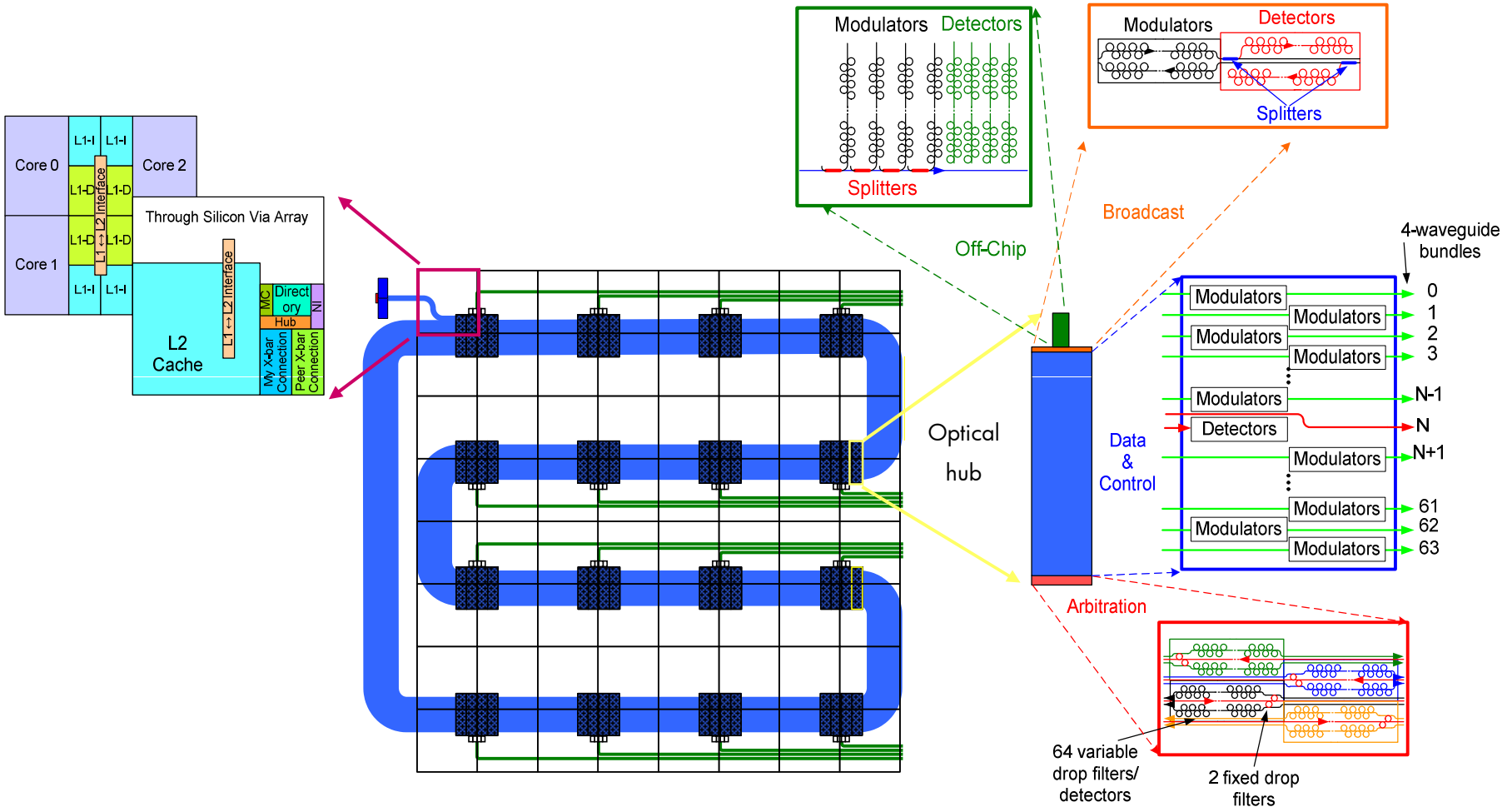
# Corona Chip Stack



# On-chip Interconnect

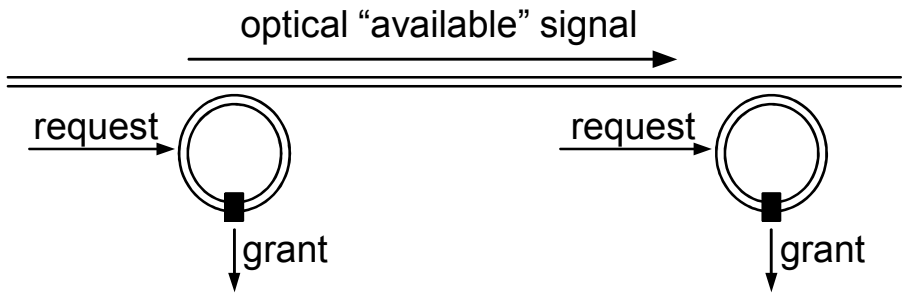


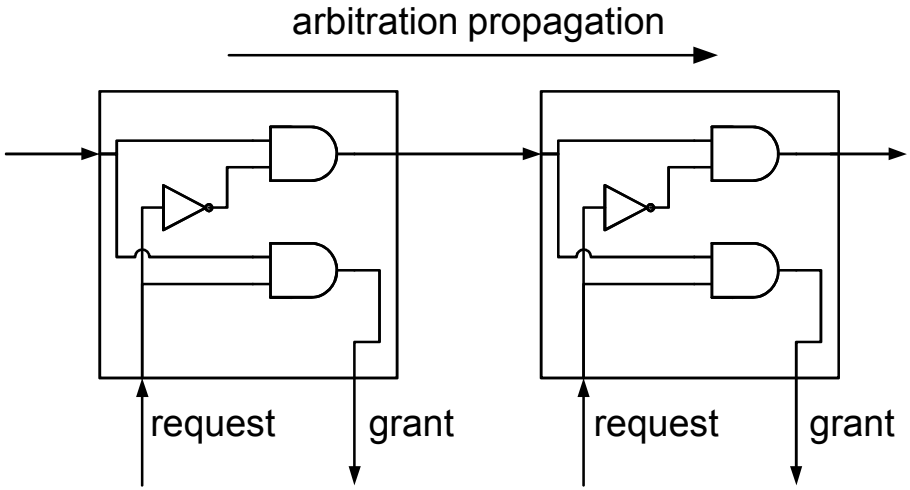
# The Optical Crossbar



# All-optical Arbitration

- A single micro-ring both asserts request and detects success or failure
- Requester tries to divert one wavelength
  - Detected power: success/failure
- Off resonance micro-rings add no delay and negligible loss –> highly scalable
- Arbitration time is light propagation time
- DWDM –> many concurrent arbitrations



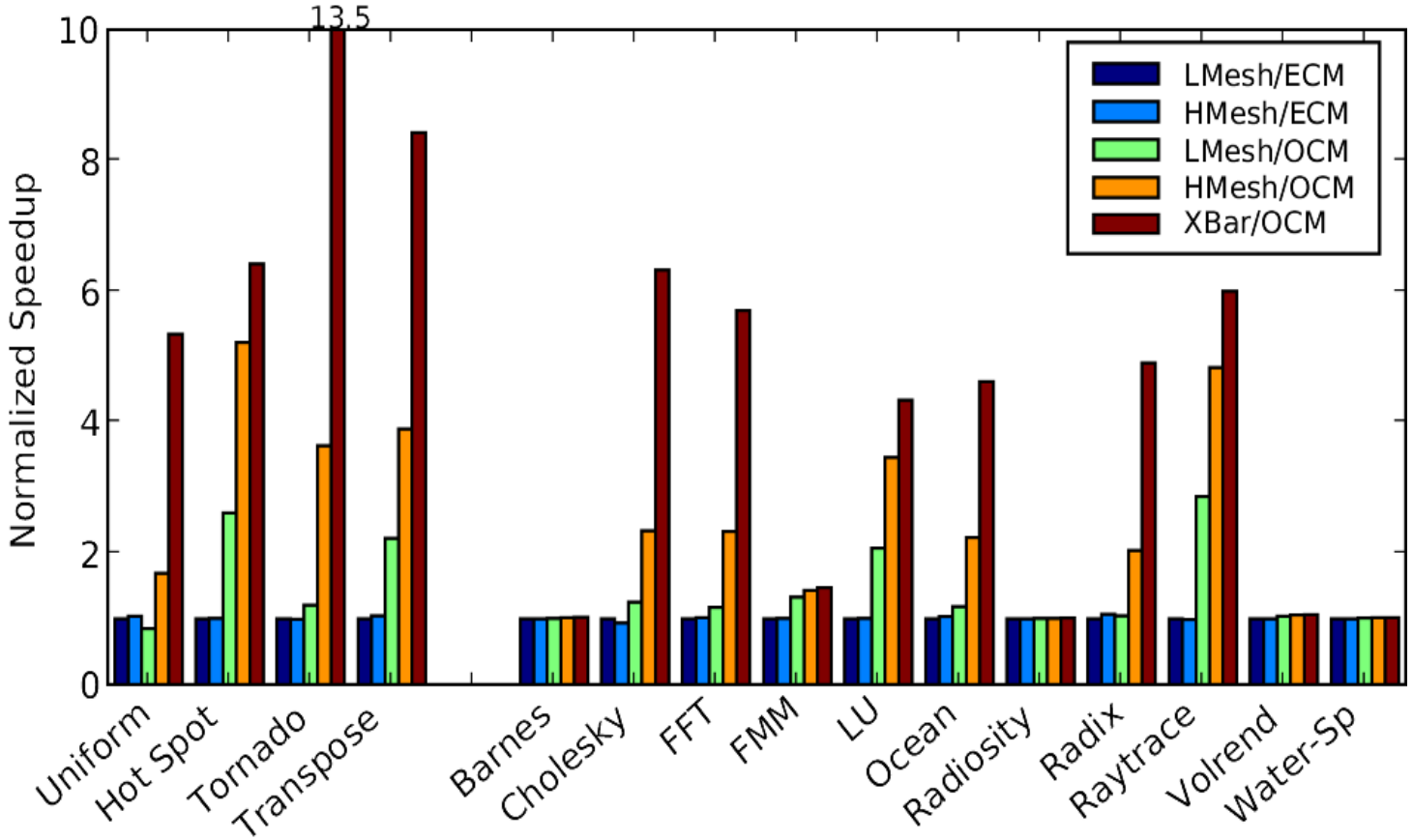




equivalent electronic circuit

# Performance

- Compare 5 systems using:
  - Three different on-chip interconnects
    - Electrical 2D on-chip mesh, 0.64 TB/s and 5 cycle hops (LMesh)
    - Electrical 2D on-chip mesh, 1.28 TB/s and 5 cycle hops (HMesh)
    - Optical crossbar, 20.48 TB/s and 8 cycles total
  - Two different memory subsystems
    - Electrical 0.96 TB/s, 1536 signal pins, memory latency is 20 ns
    - Optical 10.24 TB/s, 256 fibers, memory latency is 20 ns
- Simulate using COTSon + M5
  - 4 synthetic benchmarks
  - SPLASH-2

# Performance (LMesh/ECM = 1)

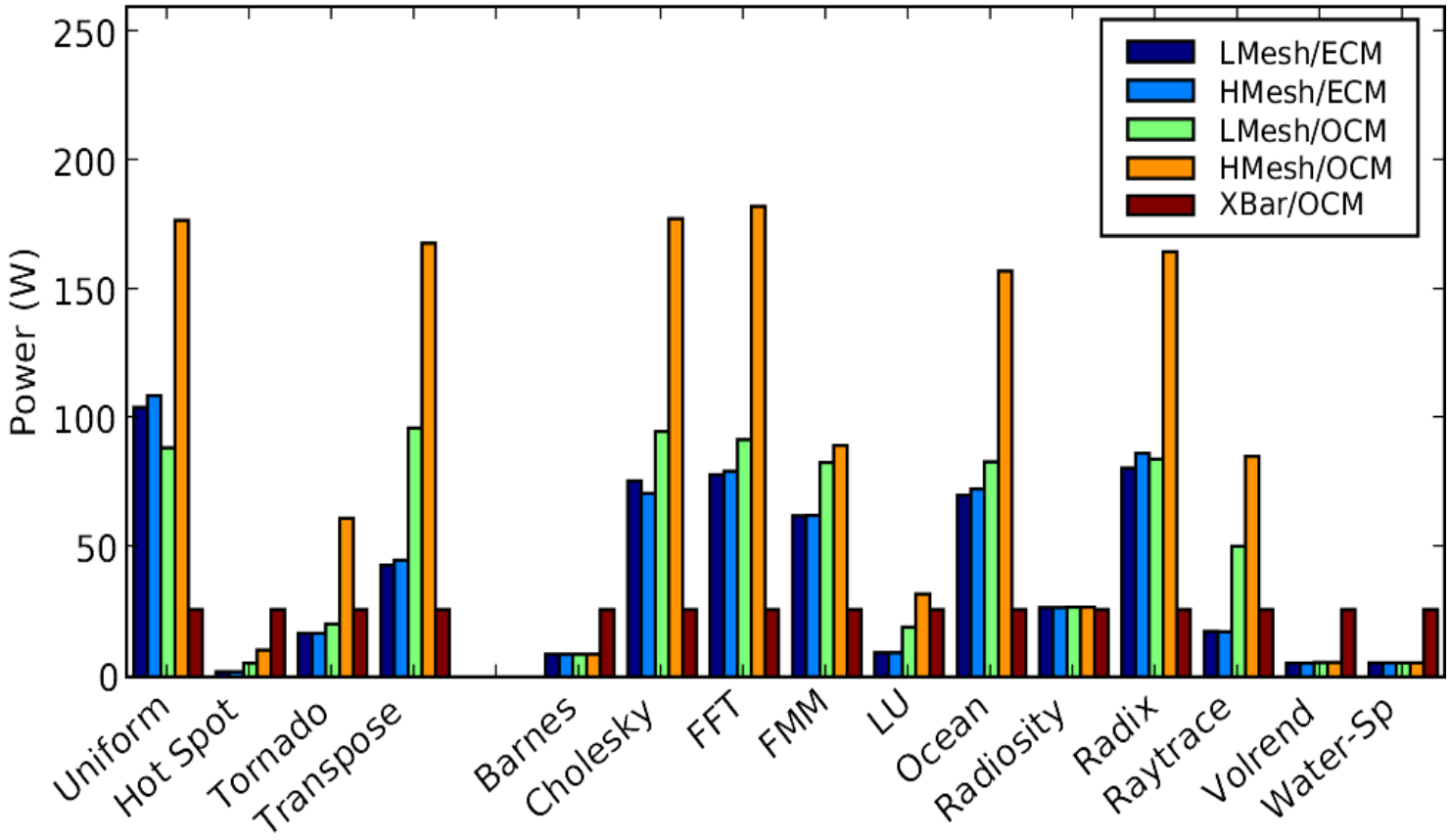



Applications that don't fit in cache show 4-6X improvements with Xbar



LABS hp

# On-chip Network Power




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

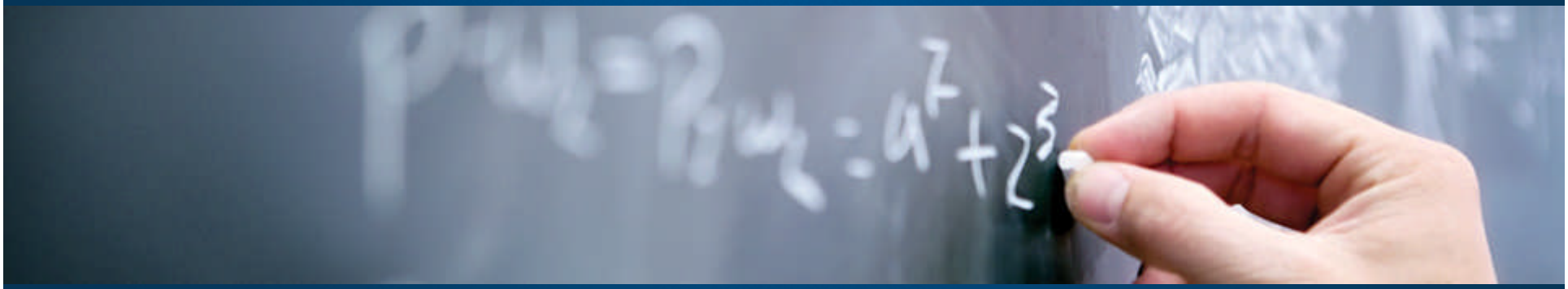


LABS hp

# Optics Can Remove the Bottlenecks

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

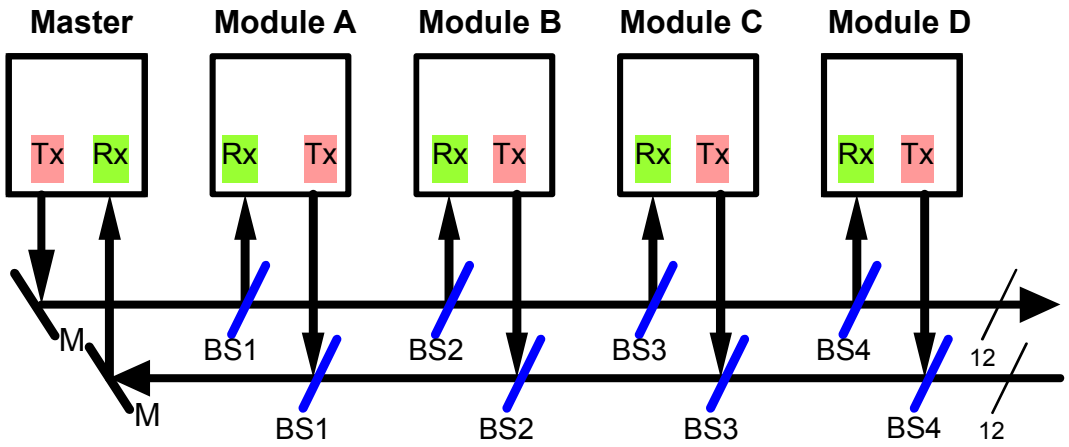
# Near-term Technologies



# Optical Buses

- Preview of upcoming Hot Interconnects presentation
  - "A High-Speed Optical Multi-drop Bus for Computer Interconnections," Mike Tan et. al.

# Optical Multidrop Bus – A Master Slave Bus

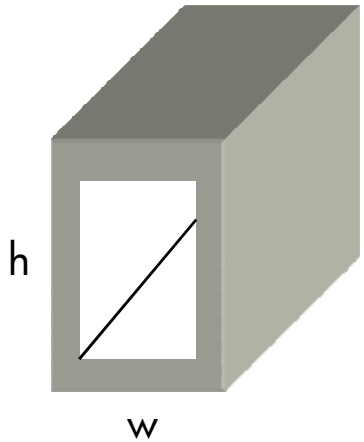


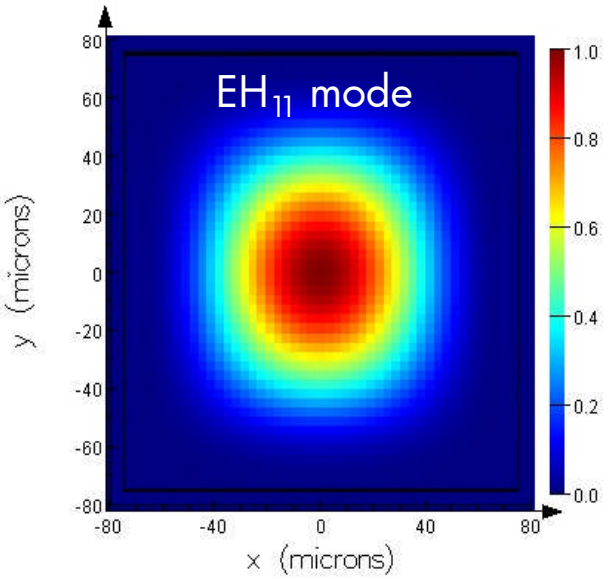

- Replace electrical transmission line with optical waveguides
- Replace electrical stubs with optical taps
- Two Unidirectional buses: 12 bit wide @ 10Gb/s = 30GB/s
  - Master broadcasts to each module on the bus;
  - Distribute optical power equally among modules
  - Each module sends data back to the master at full bus bandwidth
- Lower latency with reduced power

# 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$\mu$m, h =150$\mu$m
$\alpha$ = 0.0015 dB/cm
$n_{eff}$ ~ 1
NA ~ 0.01

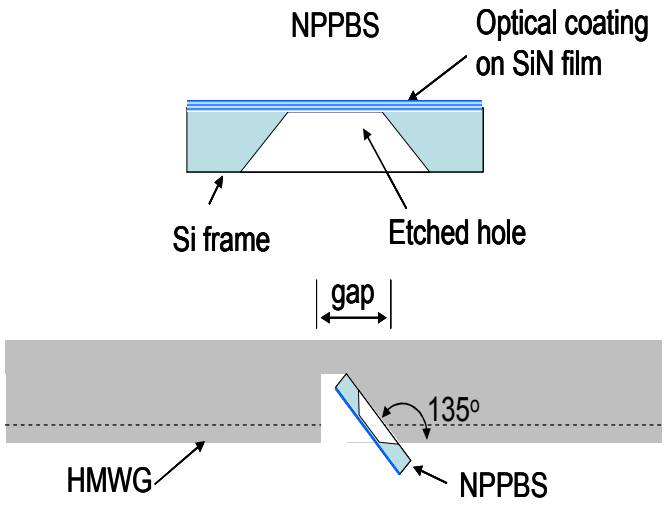


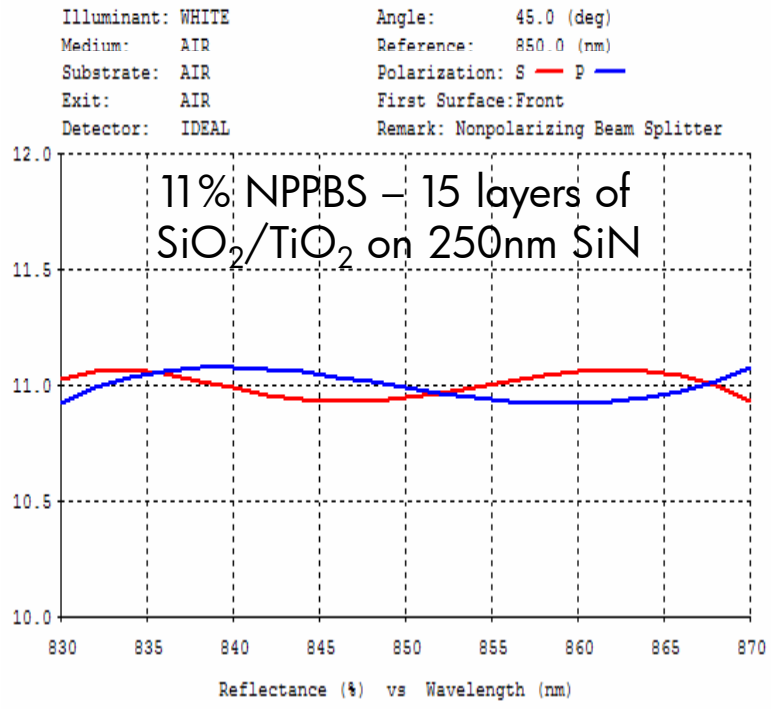

(1) E. Marcatili *et al., Bell Syst. Tech. J.* 43, 1783 (1964).

# Optical Taps

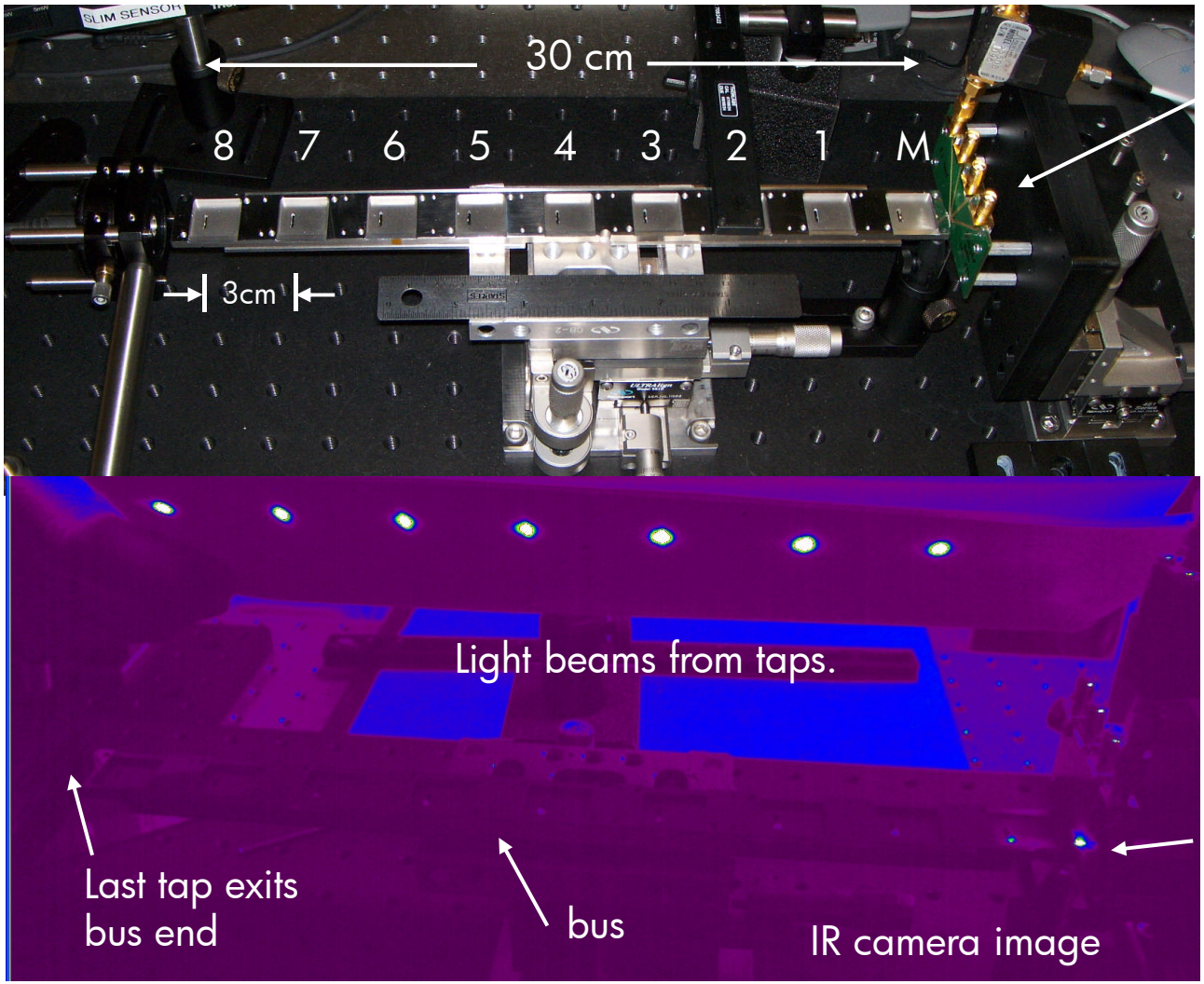
- Non-Polarizing Pellicle Beam Splitters
  - Low cost VCSELs randomly polarized
  - Negligible  beam-walk off




Insertion into HWMG @ 45°

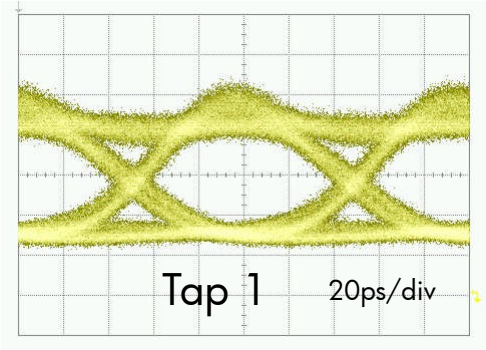


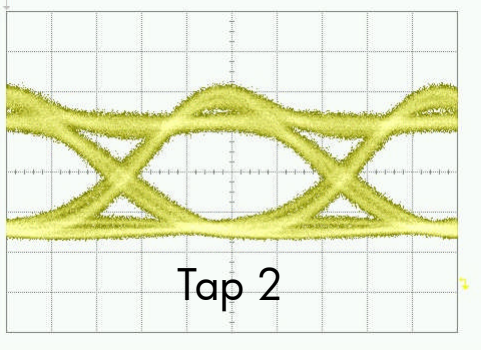
# 1x8 Fanout

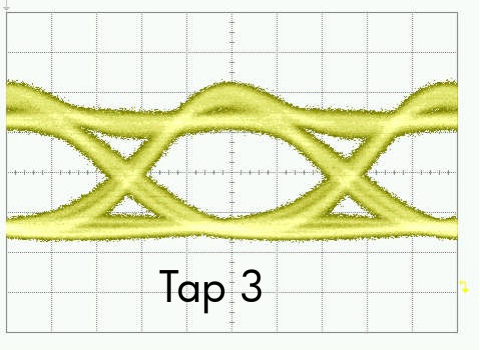


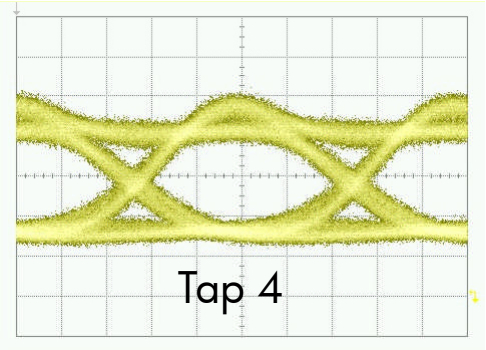

VCSEL driven from BERT thru bias-tee

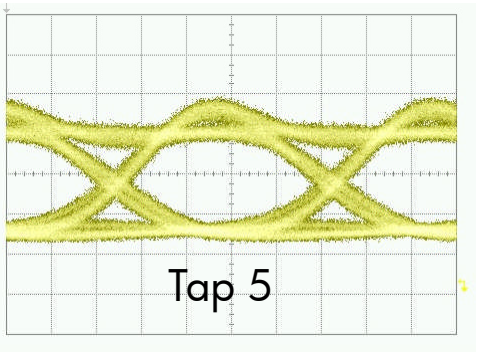
# 1x8 Fanout @ 10.3125Gbps (L=30cm)

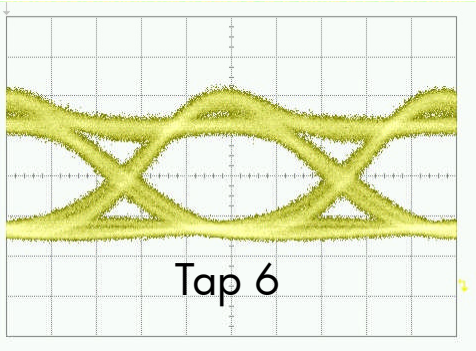


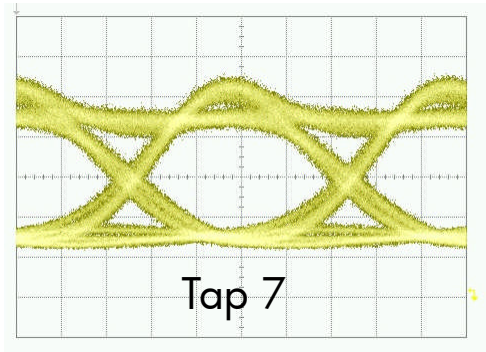



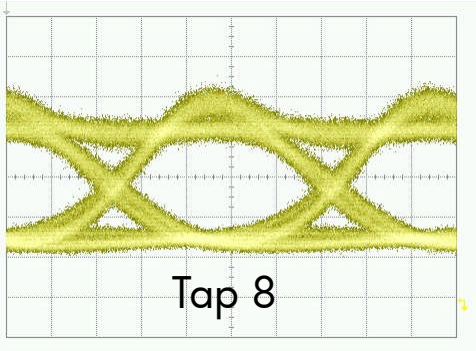



# Optical Bus Summary

- Can build today
- Provides good fan-in and fan-out (>8)
- Distance not an issue
- Composite structures (e.g., crossbars) possible

# Conclusions From Norm's Section

- Integrated photonics has the potential to:
  - Dramatically improve memory bandwidth
  - Significantly improve many-core performance
  - Reduce power
  - Simplify programming
  - All at the same time!
- Near term applications such as optical buses
  - Add significant system flexibility
  - Save latency and power

# Acknowledgements


- This includes contributions from many people:
  - All my ISCA 2008 coauthors
  - All my 2008 Hot Interconnects coauthors
- Special thanks for slide materials:
  - Mike Tan, Ray Beausoleil, Moray McLaren, Nathan Binkert, Jung Ho Ahn, Qianfan Xu

# System Implications

# Confluence of Optics and Other Systems Trends

multicores, virtualization, fabric convergence, non-volatile storage, manageability, power, resilience trends, volume/value blurring, web2.0 datacenters,SMB/BRIC, costs, flexibility, commodity…

=

# Interesting opportunity to rethink system arch & mgmt

# Designing Future Servers & Datacenters

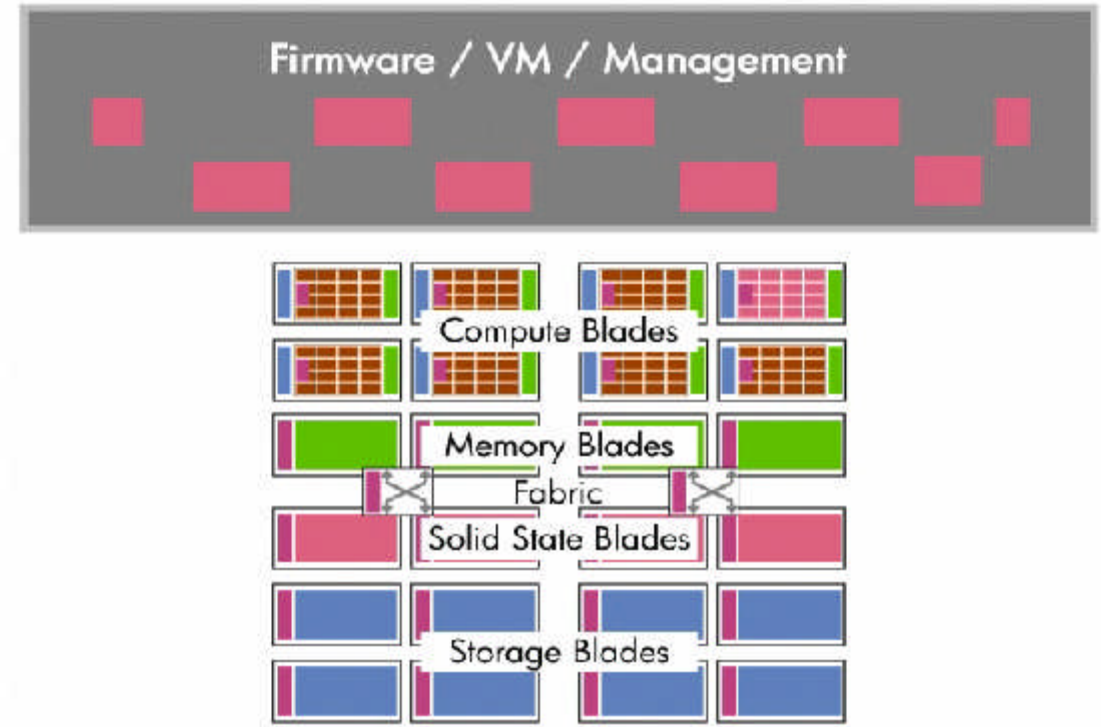
Proposal:

- power-efficient building blocks
- co-designed across hardware/software
- dynamically shared & configured as ensembles
- as needed, when needed

Why?

- One design: address power, m'gbility, scale, costs

# Designing Future Servers & Datacenters




cost-efficient building blocks across hardware/software,

dynamically shared and configured at datacenter level



LABS hp

# Several Interesting Research Directions

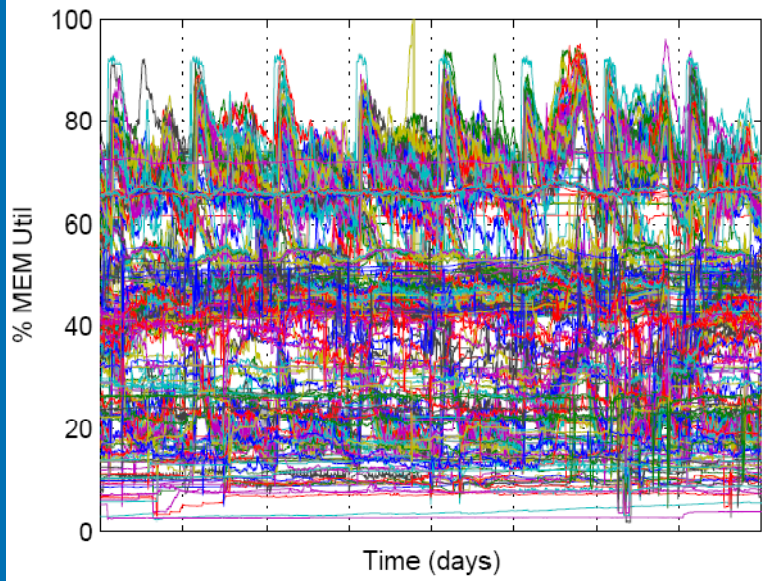
- Rethink architecture

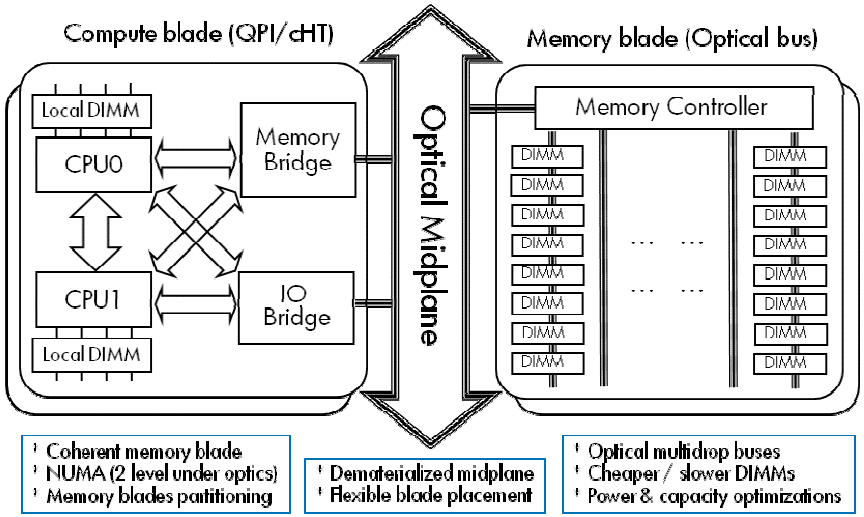
  [Beyond the "box" to the datacenter]

- Rethink management

  [Beyond the "platform" to the solution]

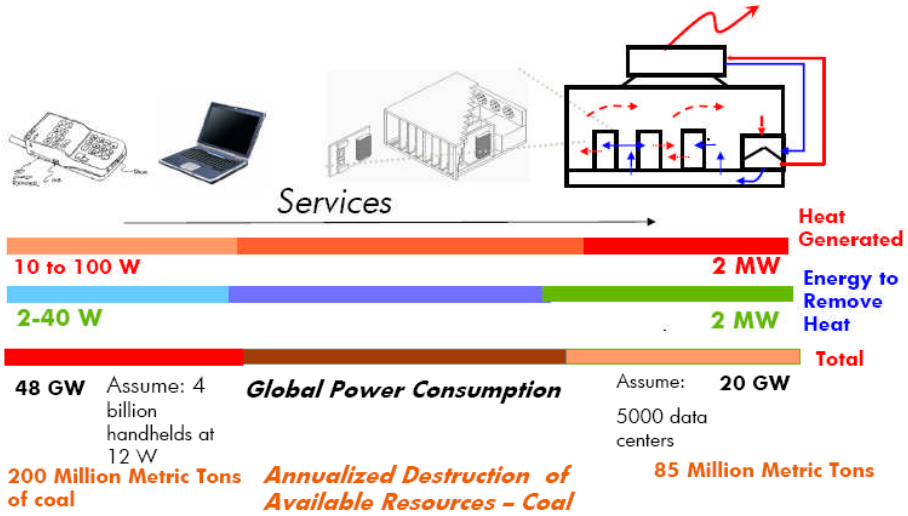
# E.g., Disaggregated systems



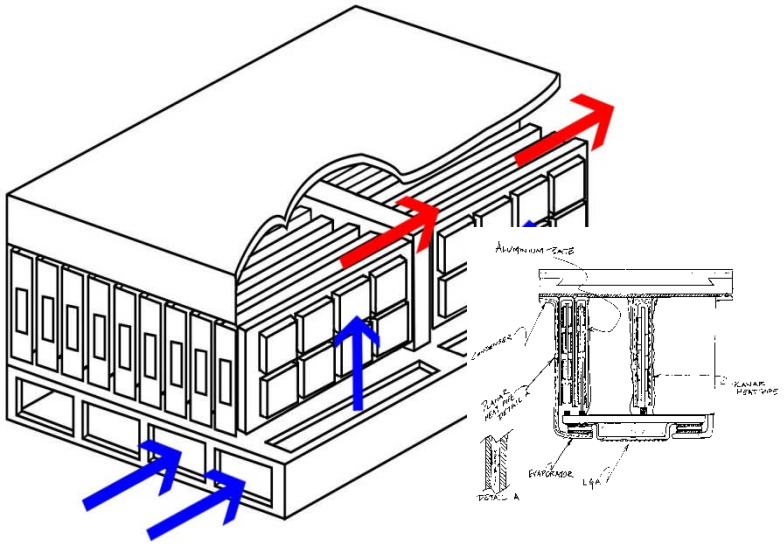




- Reduce memory power
- Enable non-volatile storage

# E.g., "Dematerialized" systems

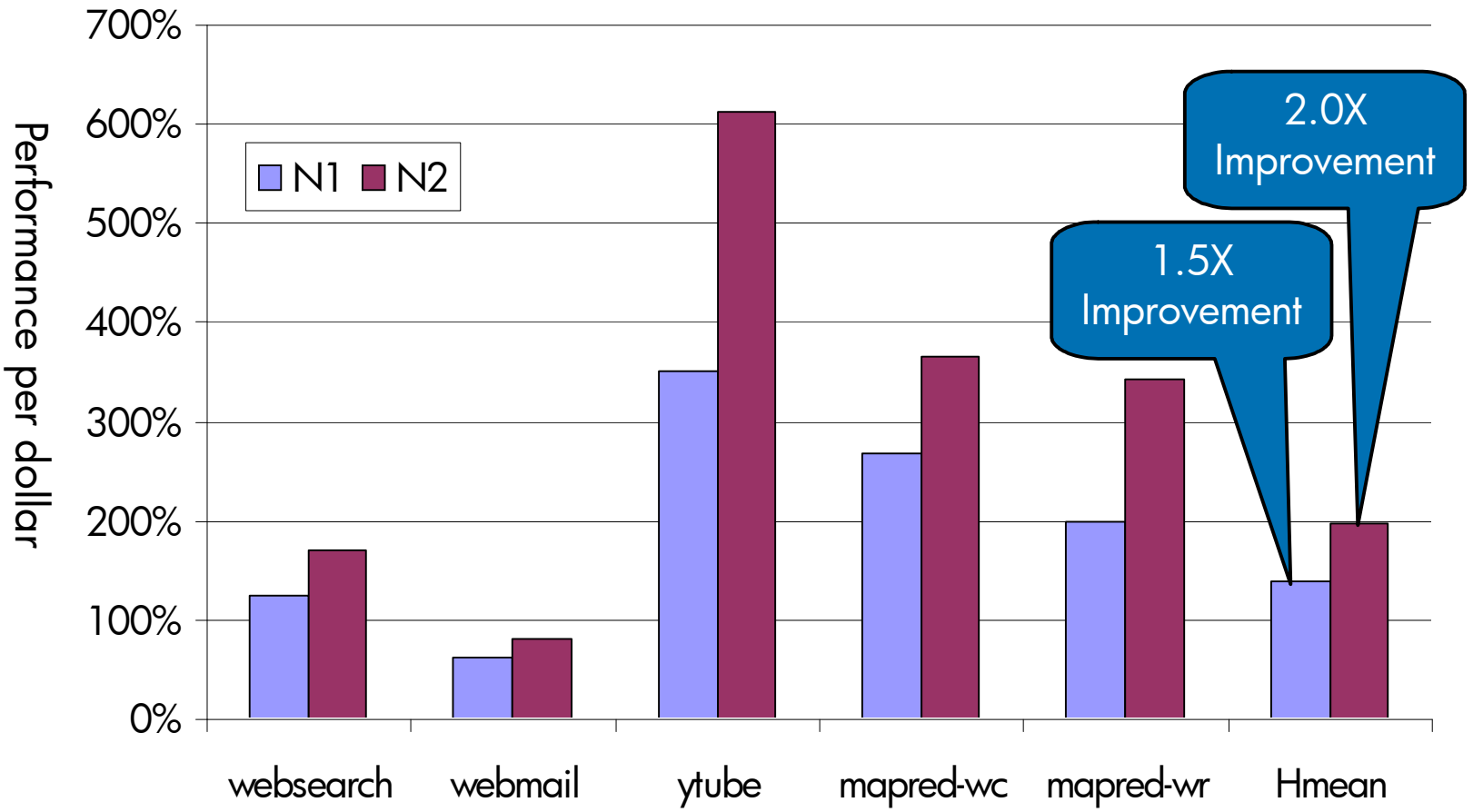



[Chandrakant Patel, Dematerializing the Ecosystem, Usenix08]



- Improved cable management
- Improved packaging efficiencies

# Early results (for web2.0) [isca2008]




*Even higher results possible with photonics…*



LABS hp

# Closing Remarks

- Integrated photonics had disruptive potential
  - Energy efficiency
  - Improved bandwidth
  - Simpler programming

- Future systems implications
  - New architectures & flexibility (e.g., optical buses)
  - Disaggregation and dematerialization enablement

# Closing Remarks

Integrated photonics

Disaggregated datacenters

LABS hp


hp