Architectural Opportunities and Challenges from Emerging Photonics in Future Systems

George Michelogiannakis, John Shalf
Lawrence Berkeley National Laboratory

Benjamin Aivazi, Yiwen Shen, Keren Bergman, Madeleine Glick
Columbia University

Larry Dennison
NVIDIA

Overview

- Specialization in future HPC and datacenter systems will stress the network
- Optical advancements bring significant promise but not as simple drop-in replacements of existing ones
  - Node and system reconfigurability
- Challenges remain
  - Including simulating optics devices at a system scale

HPC System Trends

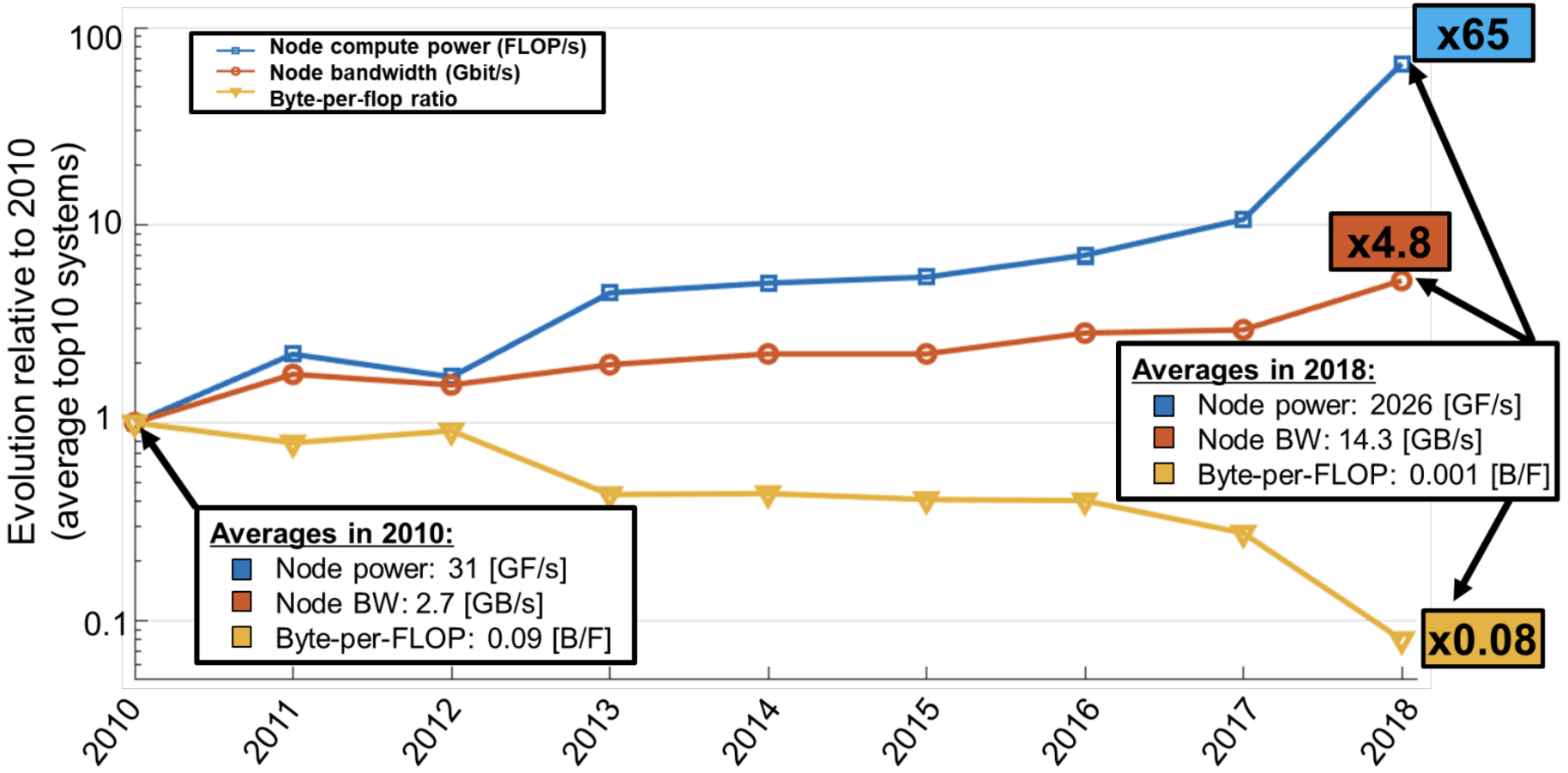
- Summit supercomputer at ORNL
  - Top performance in Linpack (top500.org results) with 122.3 PetaFLOPS
  - 13MW ⇨ **13.9 GFLOPs / Watt**
  - 6 GPUs per node. 2 CPUs

- Next challenge: Exascale computing within 20MW
  - 50 GLOPs / Watt

TOP 500
The List.

Top 10 System Trends

Performance/Communications Trends for Top 10 (2010-2018)



Sunway TaihuLight (Nov 2017) B/F = 0.004; Summit HPC (June 2018) B/F = 0.0005 → **8X decrease**

Keren Bergman, "Empowering Flexible and Scalable High Performance Architectures with Embedded Photonics", IPDPS 2018

Future Energy Goals

- 14 GFLOPs / Watt (Summit) ⇨ 72 pJ / FLOP
  - 0.36 pJ / bit
- Exascale target: 50 GLOPs / Watt ⇨ 20 pJ / FLOP
  - 0.1 pJ / bit

- **Total budget**
- The above assume 200 bits / FLOP

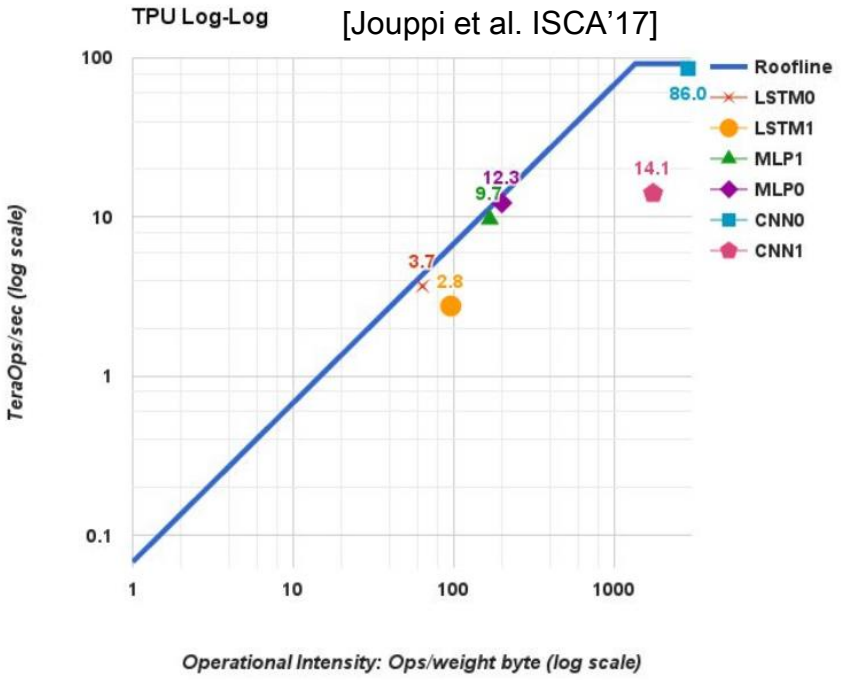
Data Movement Energy:	
– Access SRAM	O(10fJ/bit)
– Access DRAM cell	O(1 pJ/bit)
– Movement to HBM/MCDRAM (few mm)	O(10 pJ/bit)
– Movement to DDR3 off-chip (few cm)	O(100 pJ/bit)

Specialization May Be Limited By IO
Google's TPU as an Example

- Dedicated hardware for DNNs
  - Peak compute capacity: 92 TOPS/s (8-bit precision)
  - Peak bandwidth: 34 GB/s
- Must reuse a byte 2706 times to fully exploit compute capacity
  - Operational intensity: 2.7KOPs/byte, hit rate: 99.96%, 0.003 bit/OP
- Only **two** operations have high operational intensity: CNN0 and CNN1
- Operational intensity of others (e.g., translate and Rankbrain which are **90%** of the applications) are **1 – 1.5** orders of magnitude smaller
- LSTM0 would require **40x** more bandwidth to (theoretically) allow full TPU utilization

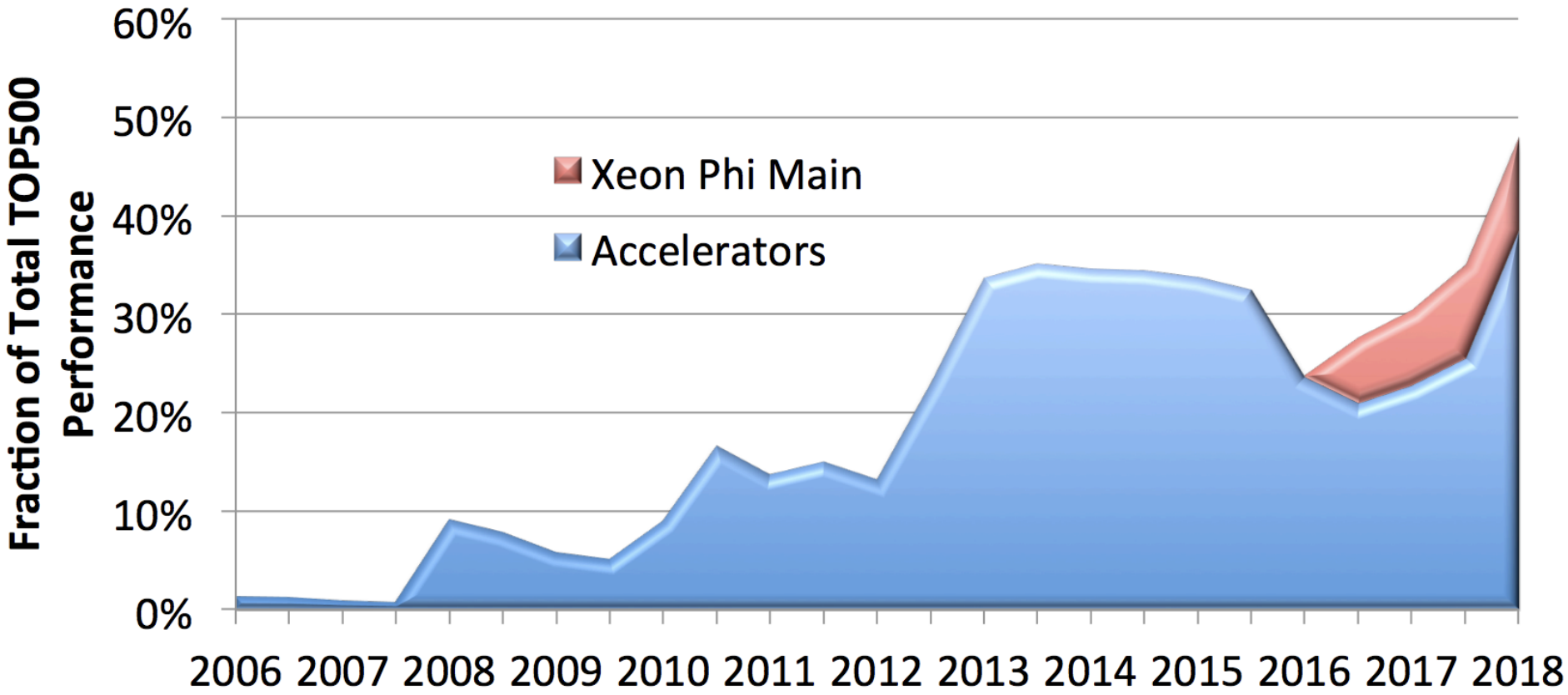
[Google cloud]

[Jouppi et al. ISCA'17]



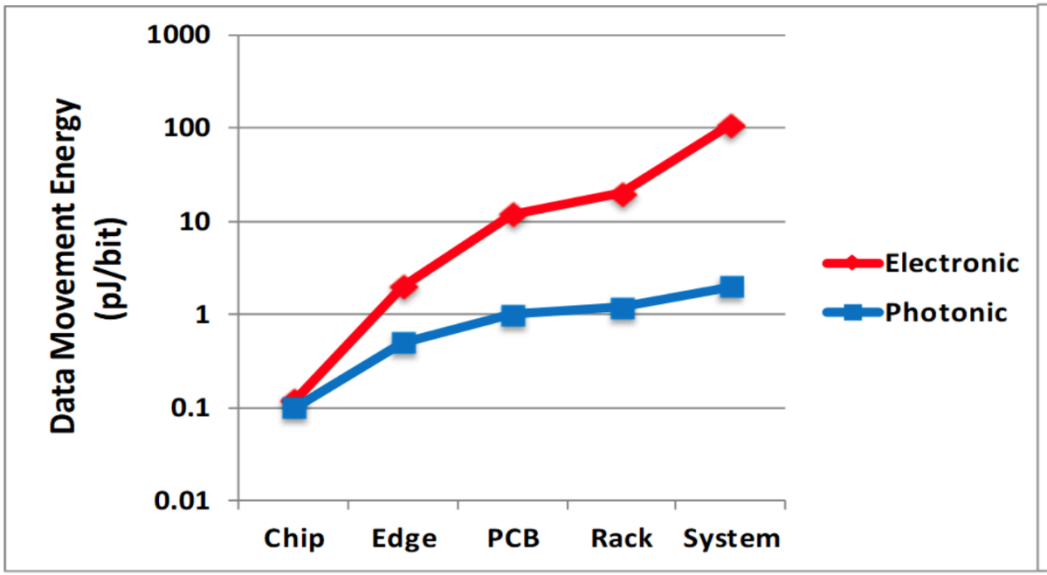
[Keren Bergman]

Specialization is Increasing

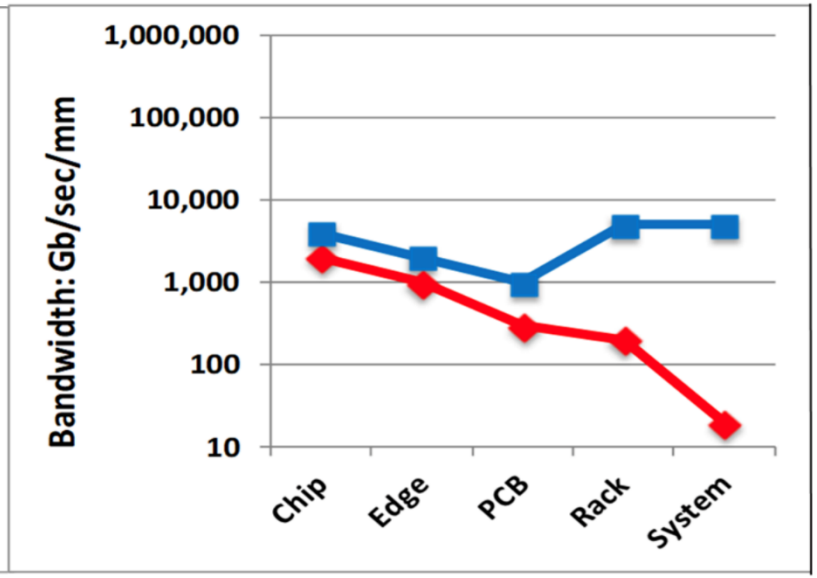


The Photonic Opportunity

The Photonic Opportunity for Data Movement



Reduce Energy Consumption



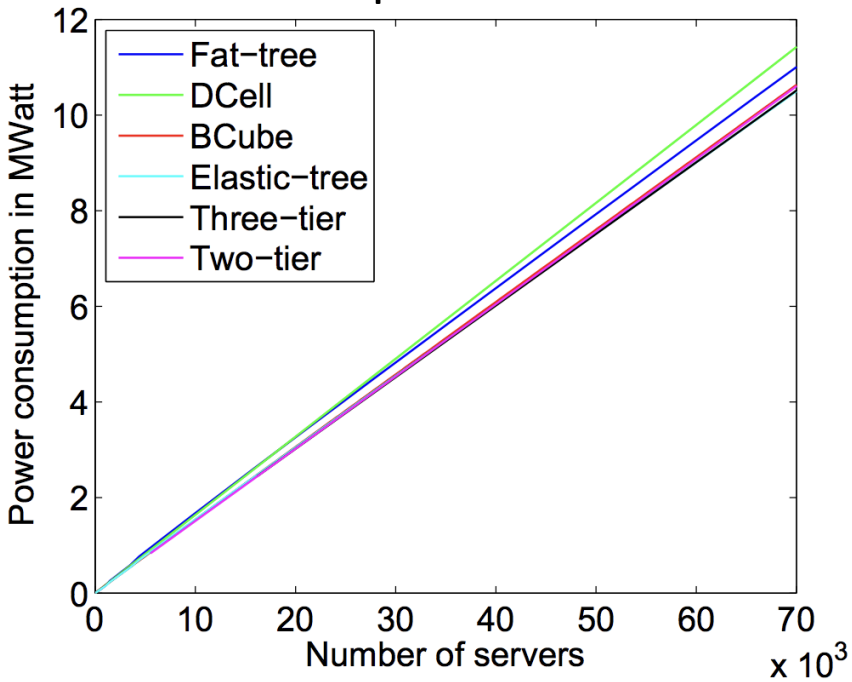
Eliminate Bandwidth Taper

R. Lucas et al., "Top ten exascale research challenges," DOE ASCAC subcommittee Report, 2014

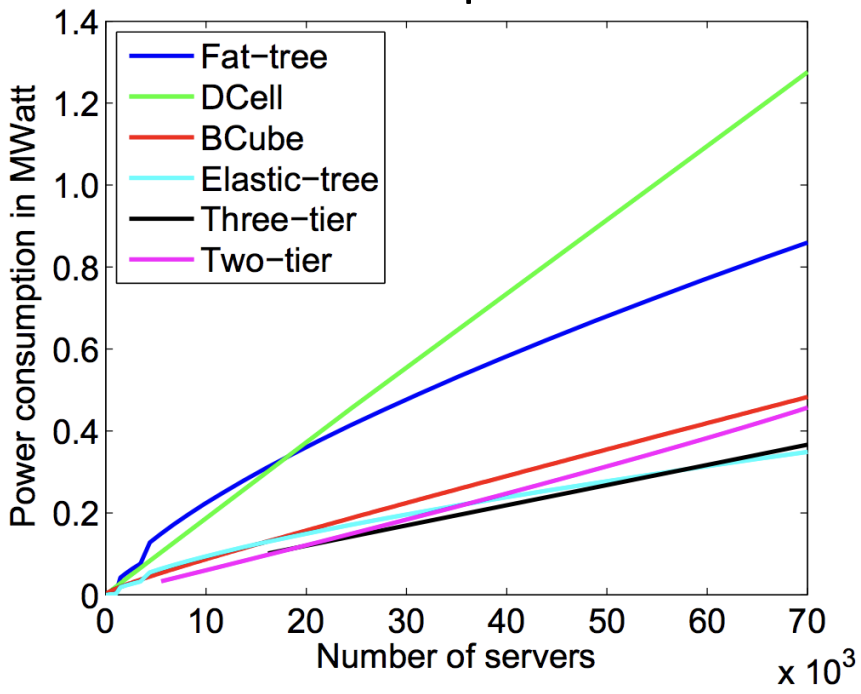
Drop-In Replacements Not Enough

- Even if we have a network that consumes no energy, we cannot reach a **2x** improvement
  - Only 4% to 12% of total power is in the network
- Key: use emerging photonic components to change the architecture

Total power



Network power

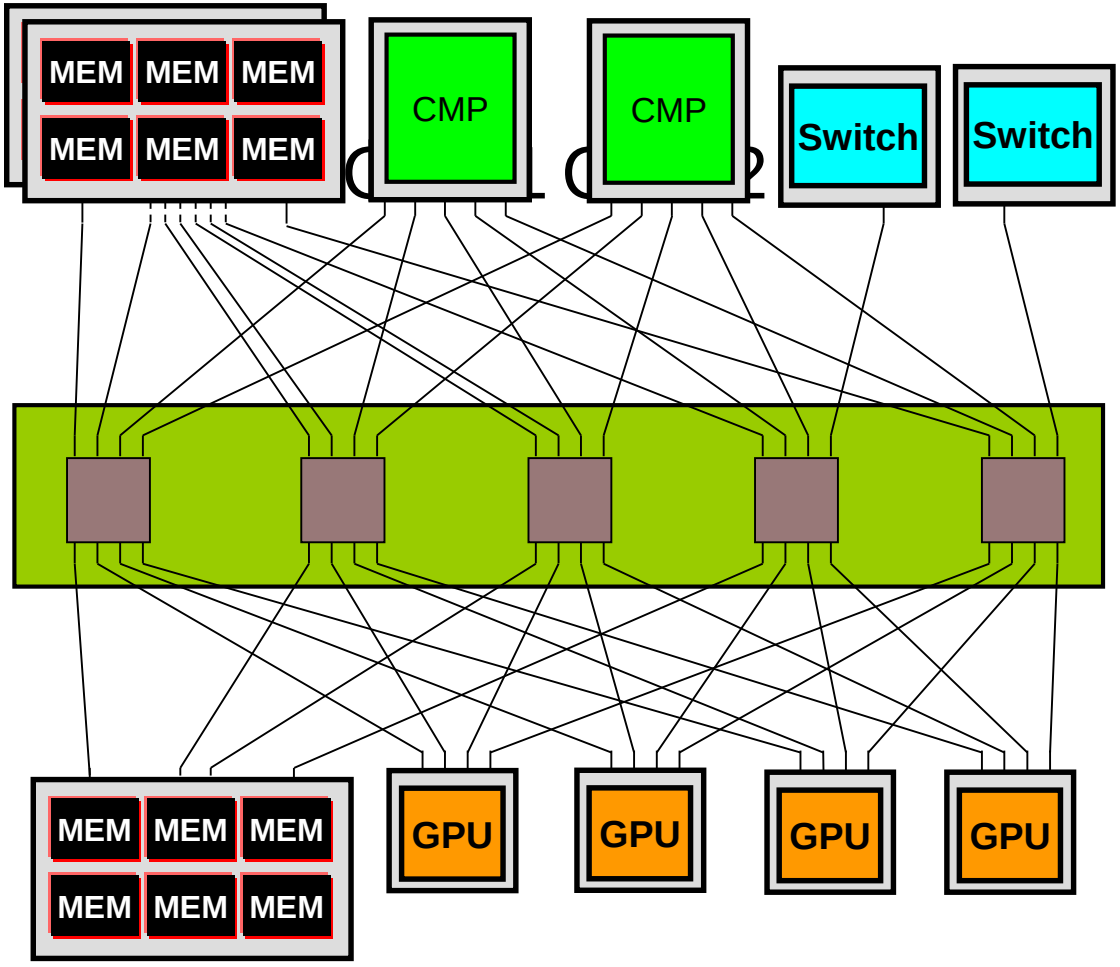


Rastin Pires et al, "Power Consumption Analysis of Data Center Architectures", GreeNets 2011

Reconfigurability

- Intra node
  - Resource disaggregation
- System-wide
  - Bandwidth steering

Optical Switches on Nodes

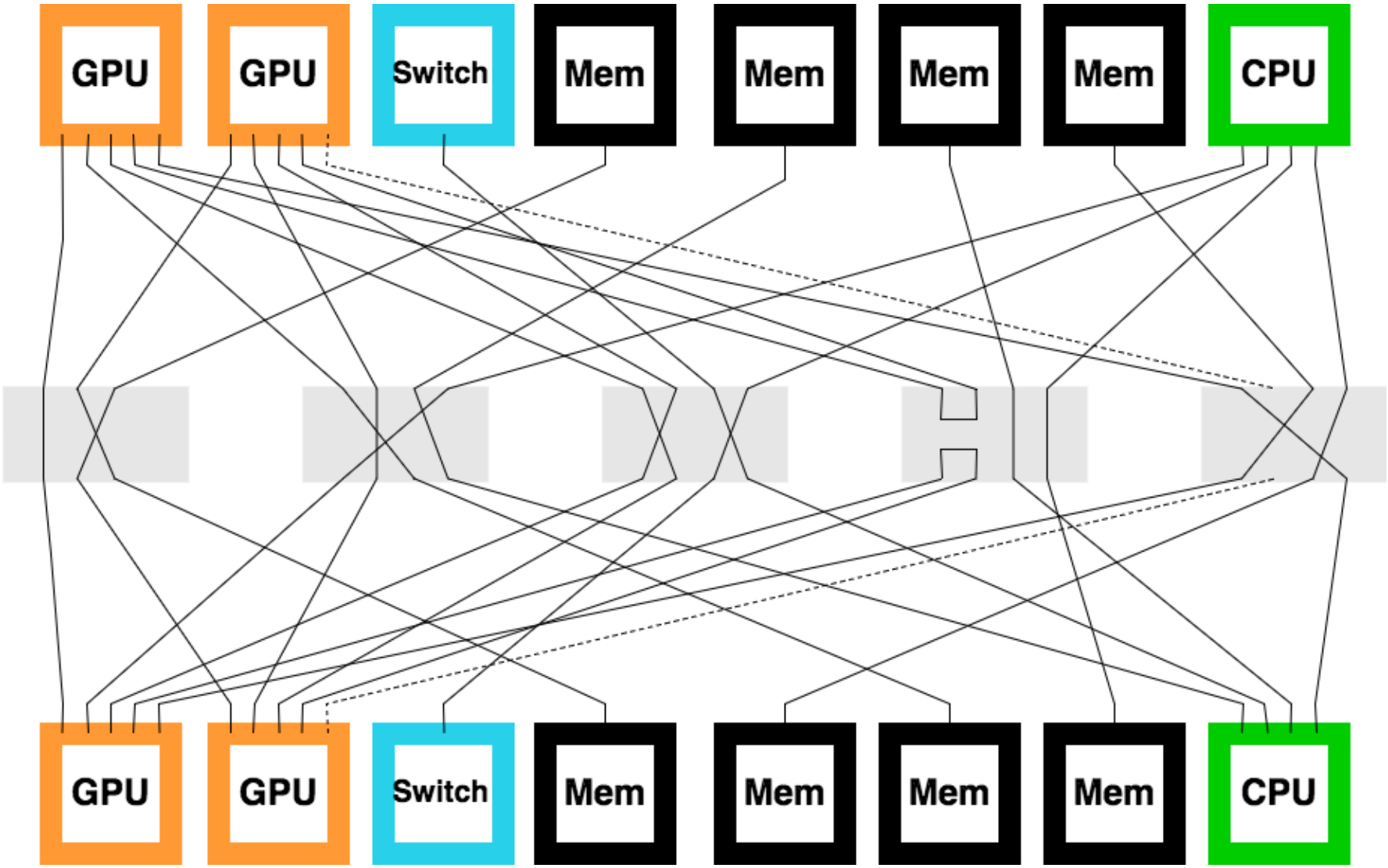


Keren Bergman, “PINE: An Energy Efficient Flexibly Interconnected Photonic Data Center Architecture for Extreme Scalability”, OI 2018

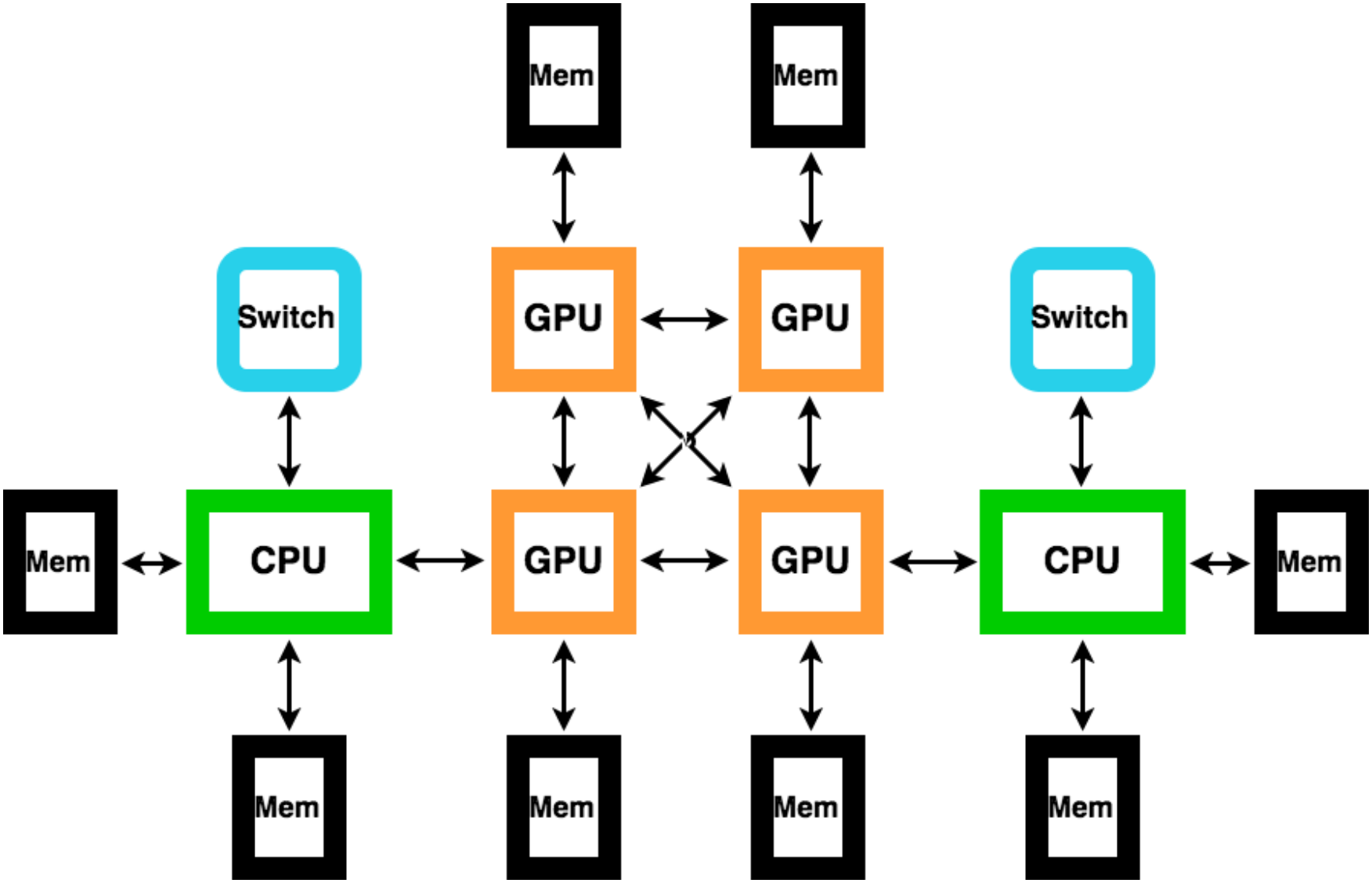
COLUMBIA UNIVERSITY

Intra-Node Reconfigurability

BERKELEY LAB



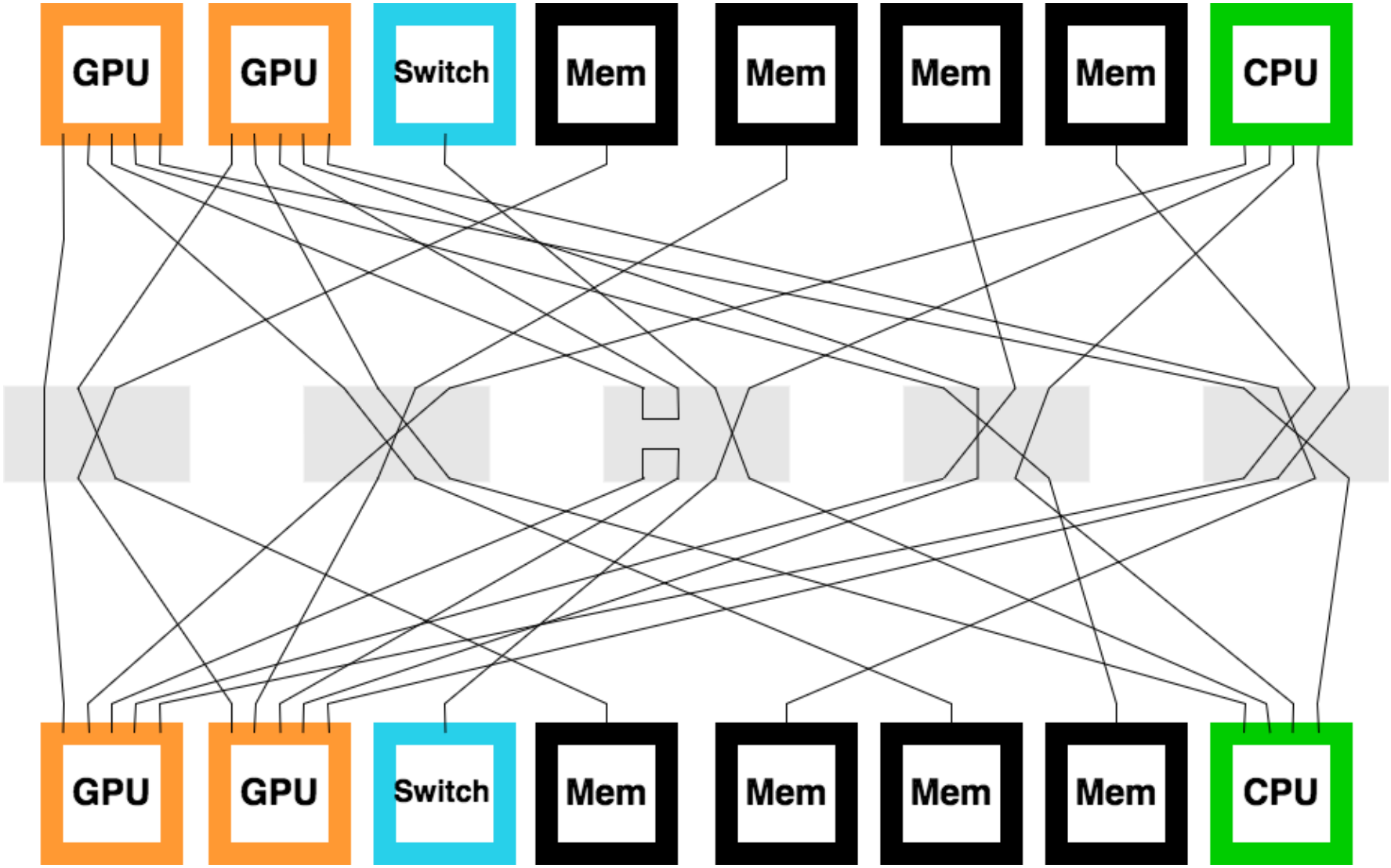
Intra-Node Reconfigurability



COLUMBIA UNIVERSITY

Intra-Node Reconfigurability

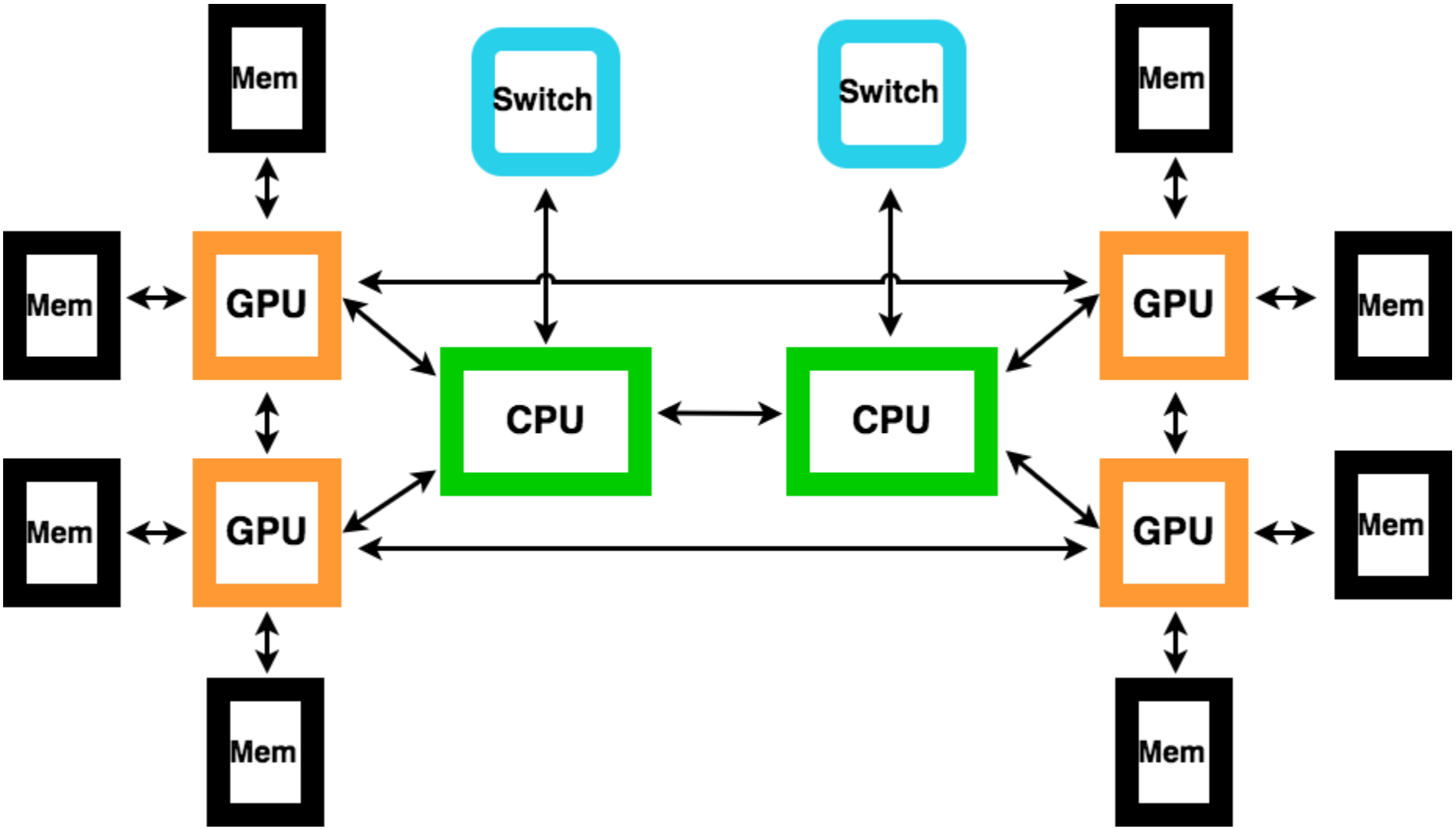
BERKELEY LAB



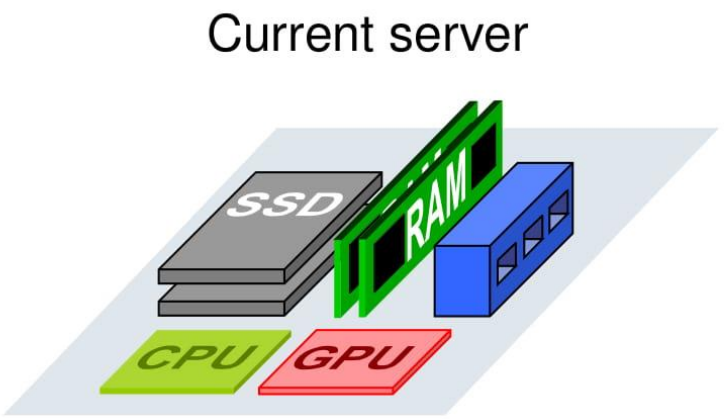
COLUMBIA UNIVERSITY

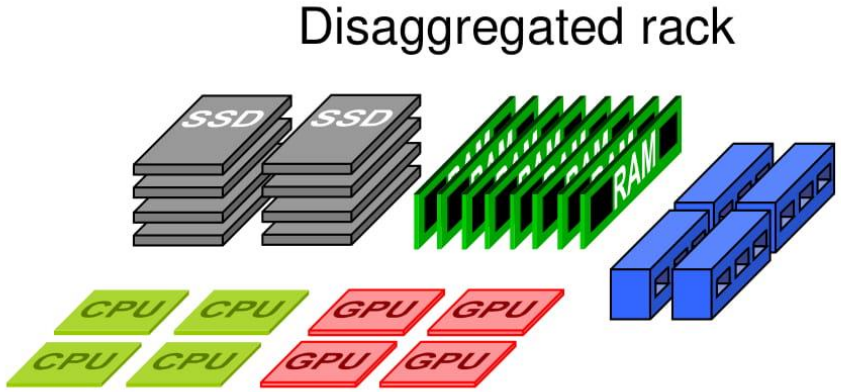
Intra-Node Reconfigurability

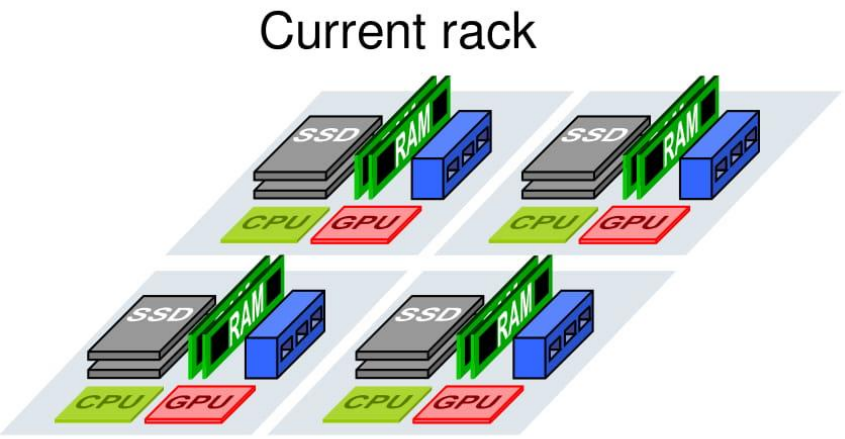
BERKELEY LAB

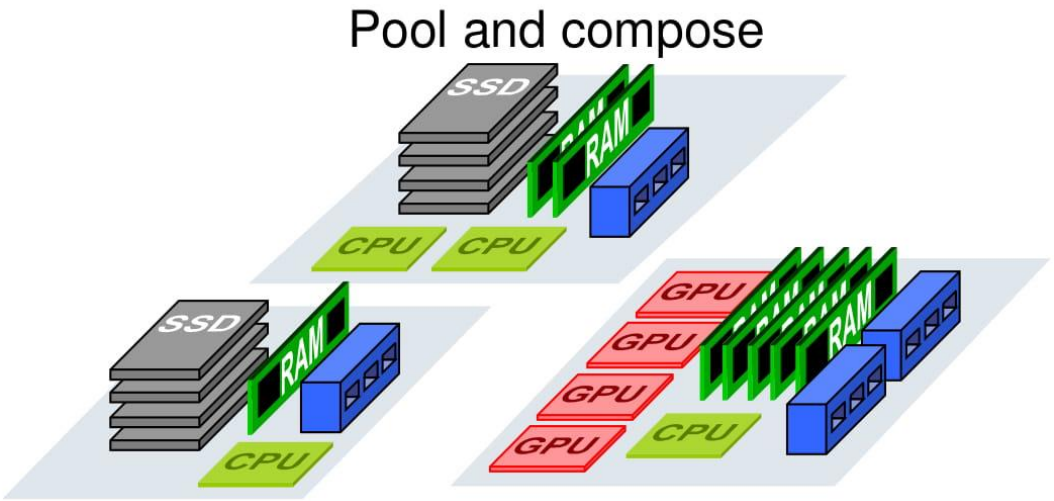


Aggregate Remote Resources









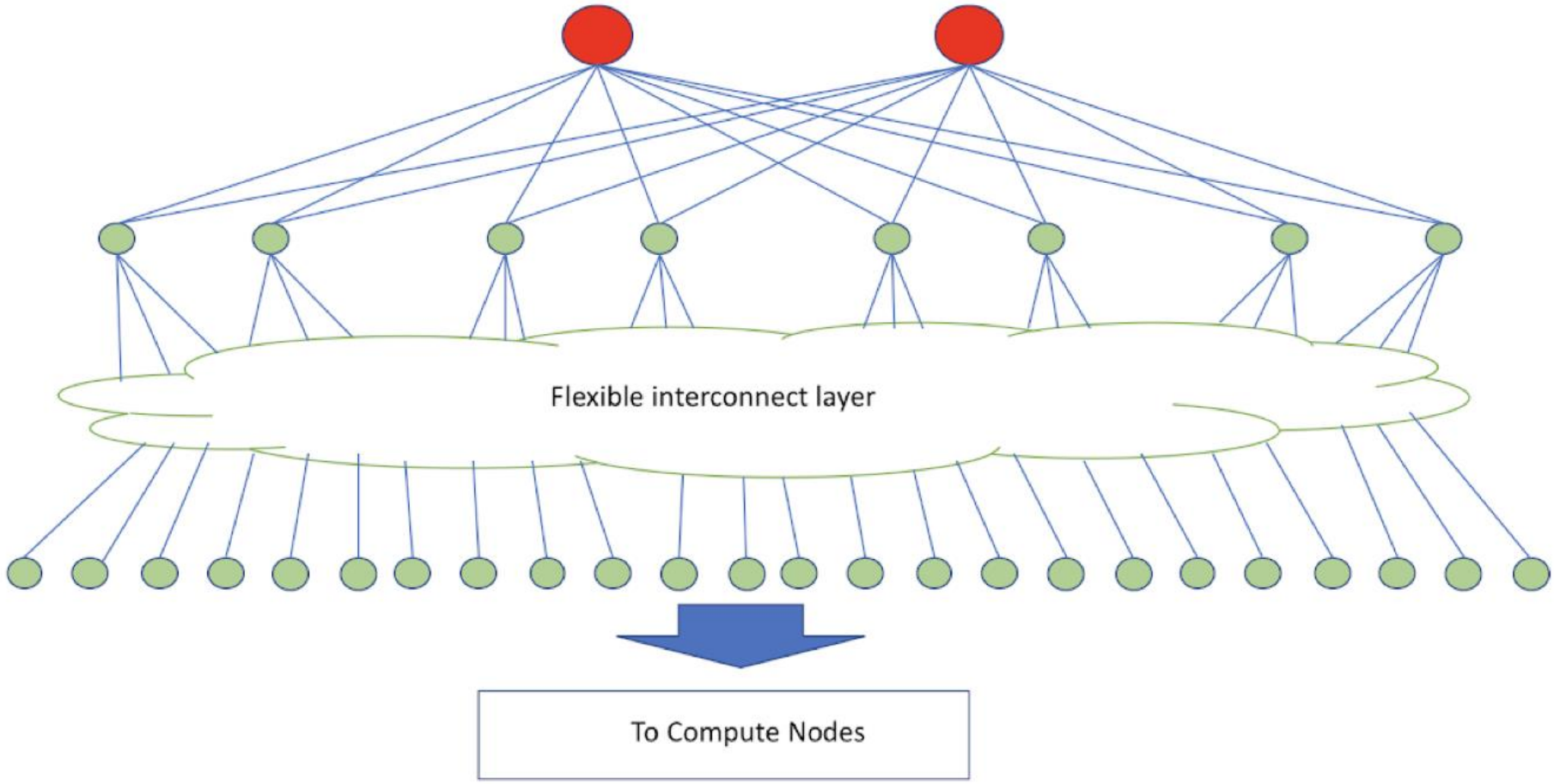
Keren Bergman, "PINE: An Energy Efficient Flexibly Interconnected Photonic Data Center Architecture for Extreme Scalability", OI 2018

Node Reconfigurability Challenges

- Photonic switches with sufficient radix
- Efficient conversion to optics
  - In package?
- How changing node configuration affects network traffic, scheduling, and system management [1]

[1] D. Z. Tootaghaj et al., “Evaluating the combined impact of node architecture and cloud workload characteristics on network traffic and performance/cost,”,2015 IEEE International Symposium on Workload Characterization.

Use Optics for Efficient B/W Steering



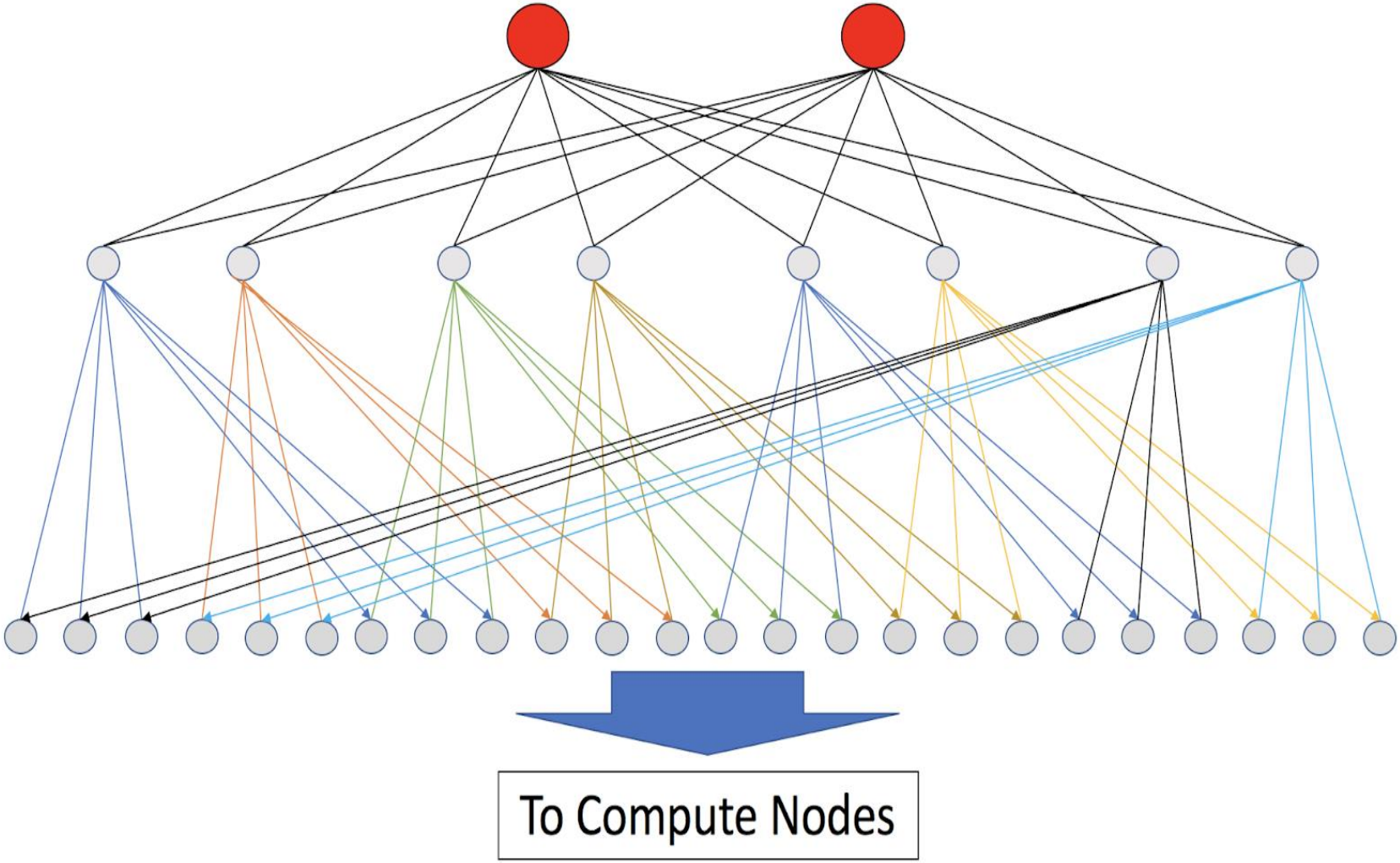
[Min Yee (Jason) The]

Adversarial Traffic Pattern

0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

[Min Yee (Jason) The]

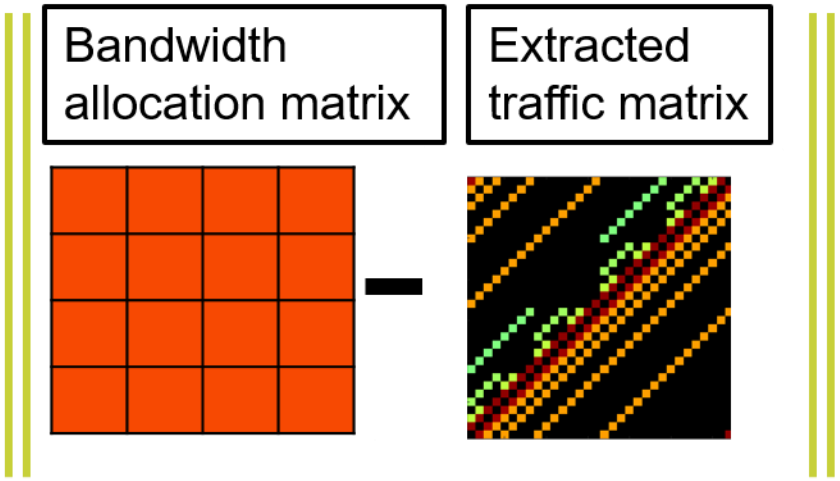
Bandwidth Steered



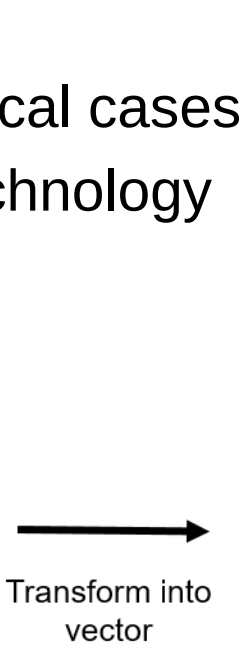
[Min Yee (Jason) The]

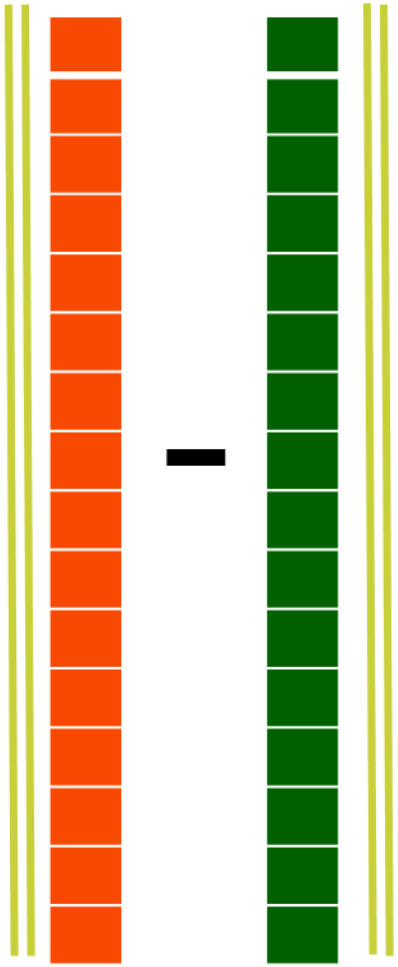
Algorithmically Challenging

- NP-hard optimally
- Respect physical limitations
- Understand implications in pathological cases
- Solid models of underlying optics technology
  - Cost of reconfiguration



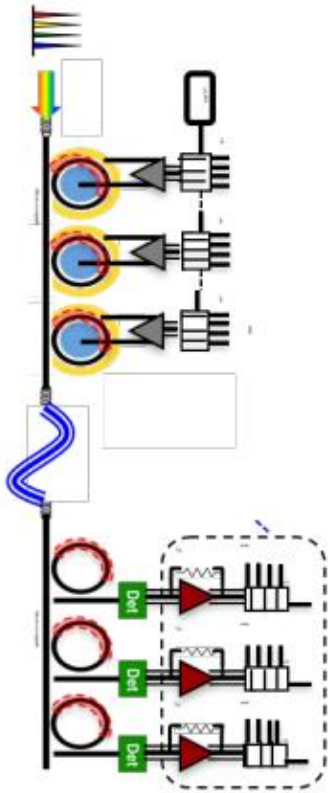
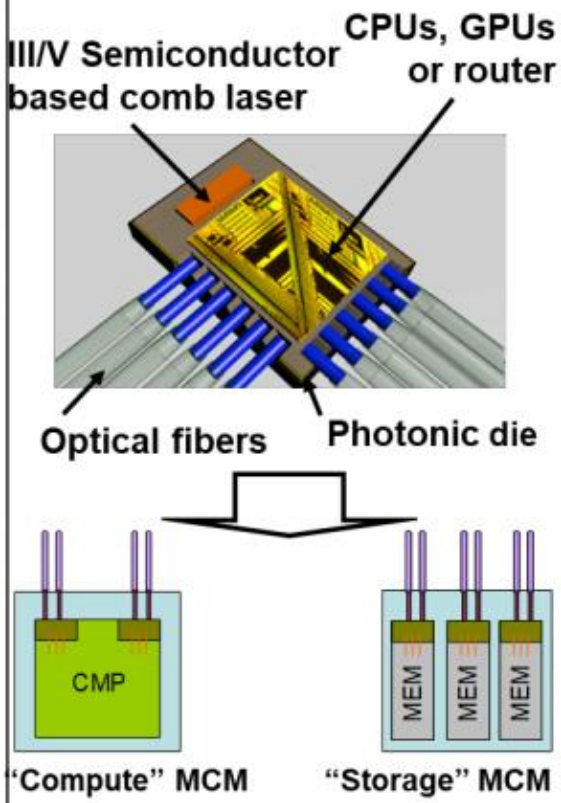
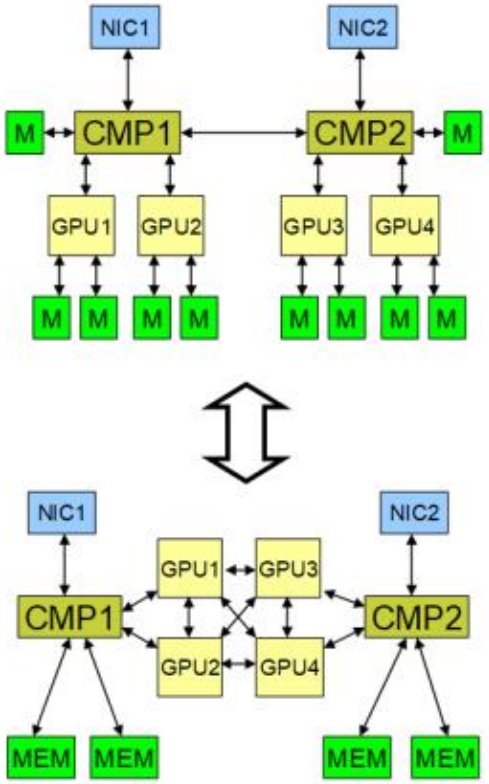
Transform matrix form into vector form





[Min Yee (Jason) The]

Shameless Plug: PINE Project

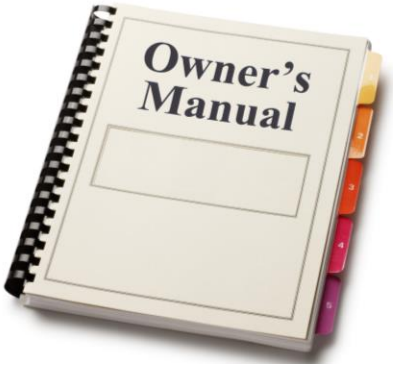
Energy optimized high bandwidth density links #1	Optically Connect MCMs – CPU/GPU/Memory #2	Adaptive connectivity bandwidth steering #3
		

Keren Bergman, "Empowering Flexible and Scalable High Performance Architectures with Embedded Photonics", IPDPS 2018

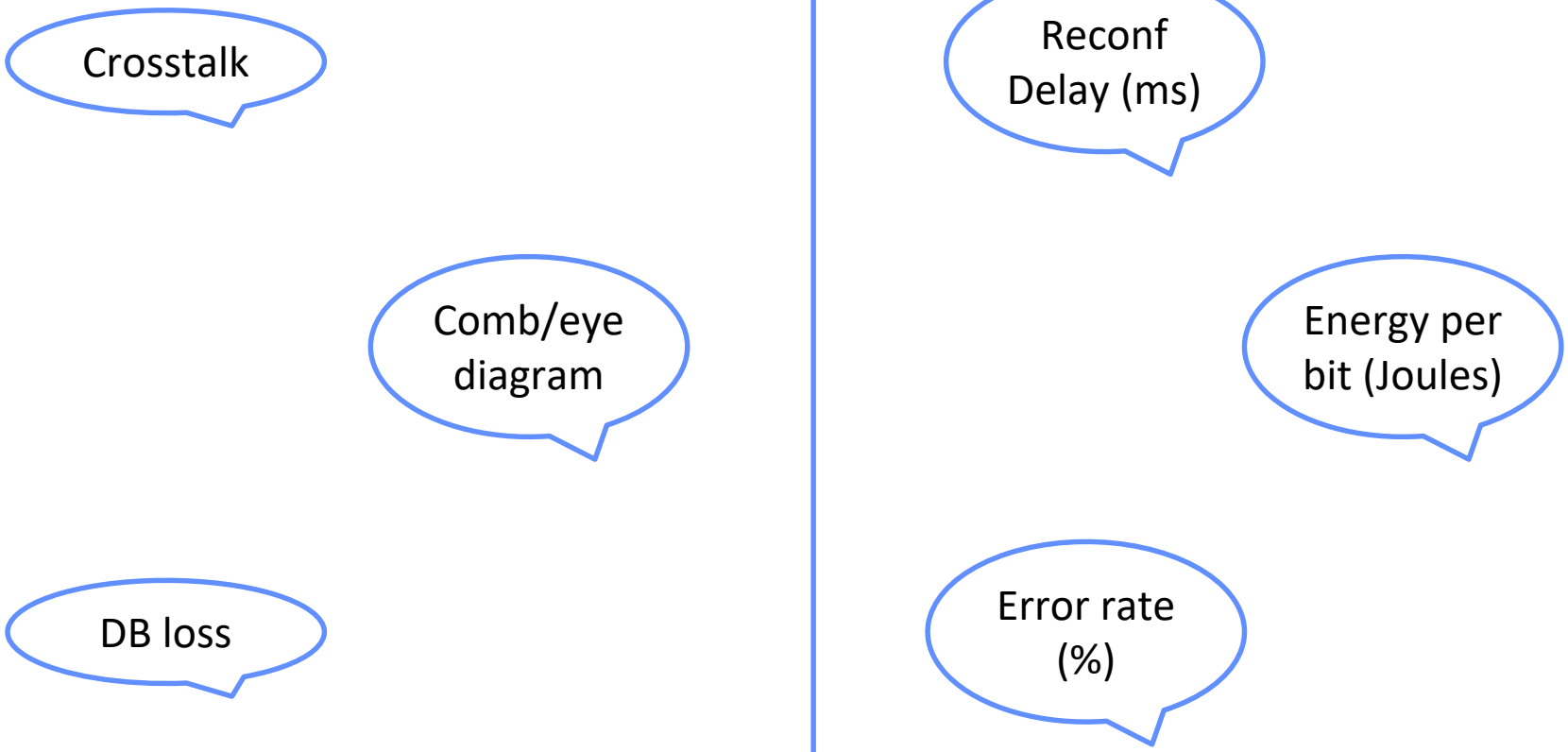
Making Use of New Components

- As an architect, new optical components are new toys that I add to my collection

- But in order for me to use this cool new toy, I need a user manual in a language I can understand

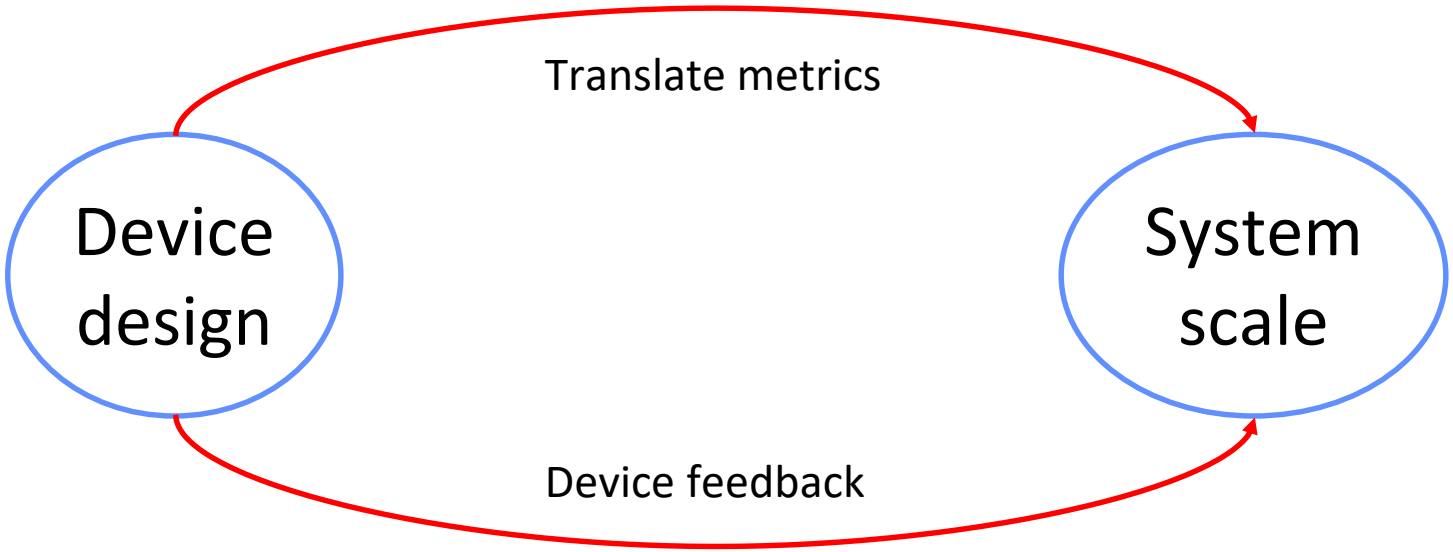


We Need to Speak the Same Language



Design Feedback

- Translate low-level metrics to architectural-level metrics
- Provide feedback to optimize devices for high-level impact
  - Prioritize reconfiguration time or energy?
  - Realize overhead of some choices, e.g., error rate
  - “Knobs” in the models are encouraged



Other Challenges/Opportunities

- Reconfigurability relies on predicting, monitoring or exposing application demands
  - Weigh the cost of reconfiguration
- How to use consecutive switch hops in the optical domain
  - Without conversion to electrical
- Faster reconfiguration and fast turn off/on lasers may change network design significantly

Optics May Not Replace Electronics

- "Electronics are approaching their limit"
  - "Optics will replace electronics"

- Electronics are fundamentally good at some aspects
  - E.g., computing such as for routing and reconfigurability
  - Packet switching -> higher utilization (dynamic traffic)

- Two options:
  - Give up electronics entirely and drastically re-design our networks with possible important drawbacks
    - With no overdesigning
  - Networks with both electronics and photonics
    - We just have to figure out exactly how much of each

Questions?

- Specialization in future HPC and datacenter systems will stress the network
- Optical advancements bring significant promise but not as simple drop-in replacements of existing ones
  - Node and system reconfigurability
- Challenges remain
  - Including simulating optics devices at a system scale