BERKELEY LAB
Bringing Science Solutions to the World

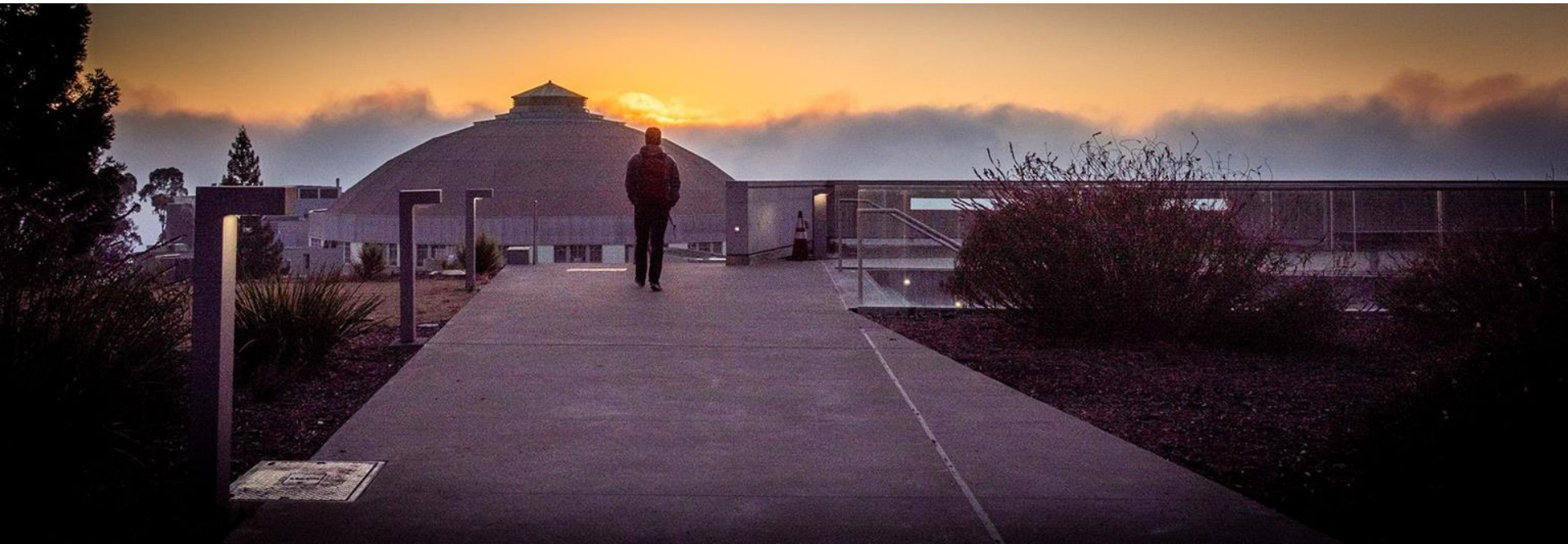
PINE

arpa·e
CHANGING WHAT'S POSSIBLE

ENLITENED

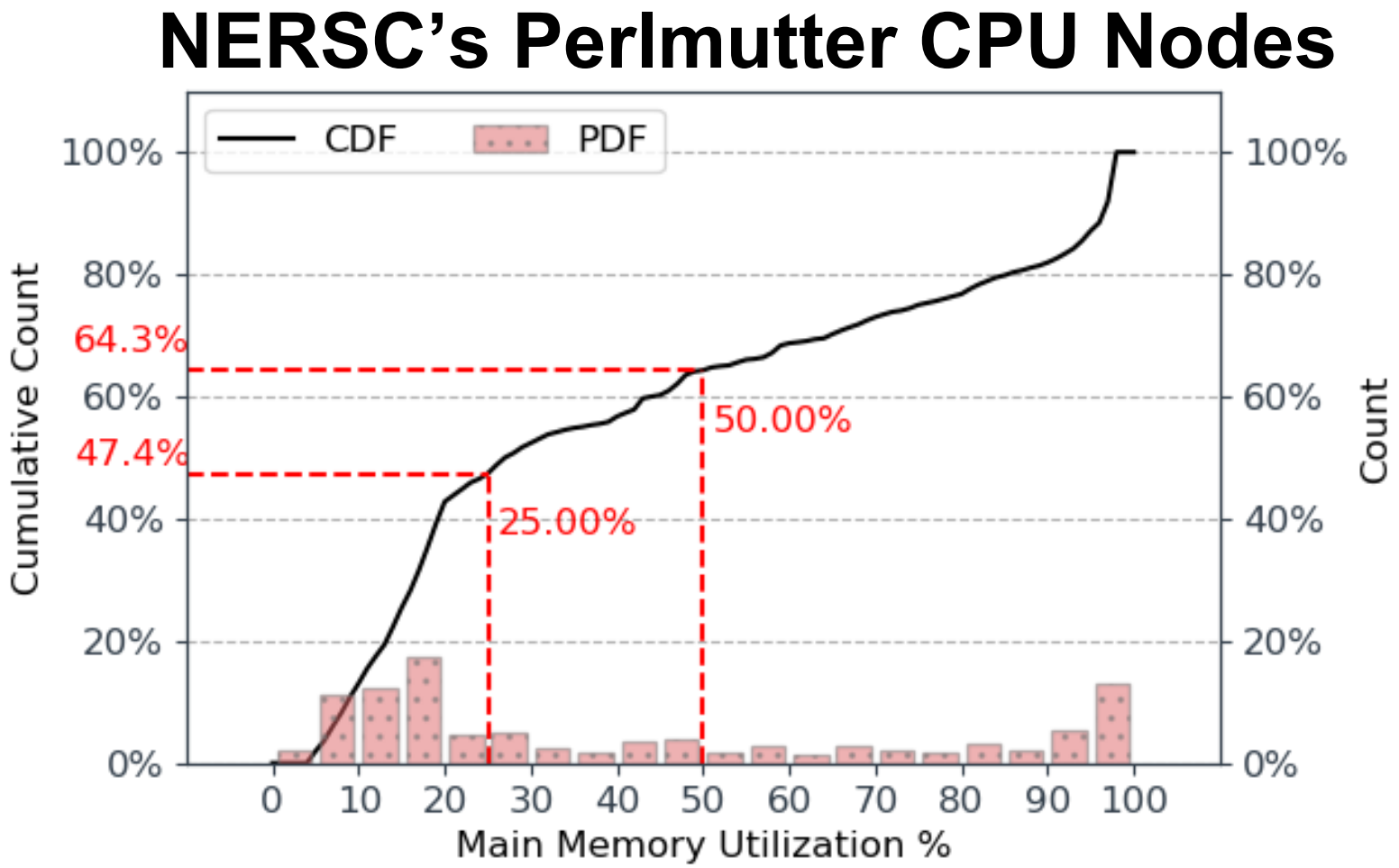
Efficient Intra-Rack Resource Disaggregation in HPC Using Co-Packaged DWDM Photonics

George Michelogiannakis, Yehia Arafa, Brandon Cook, Liang Yuan Dai, Abdel-Hameed Badawy, Madeleine Glick, Keren Bergman, John Shalf
Contact: mihelog@lbl.gov



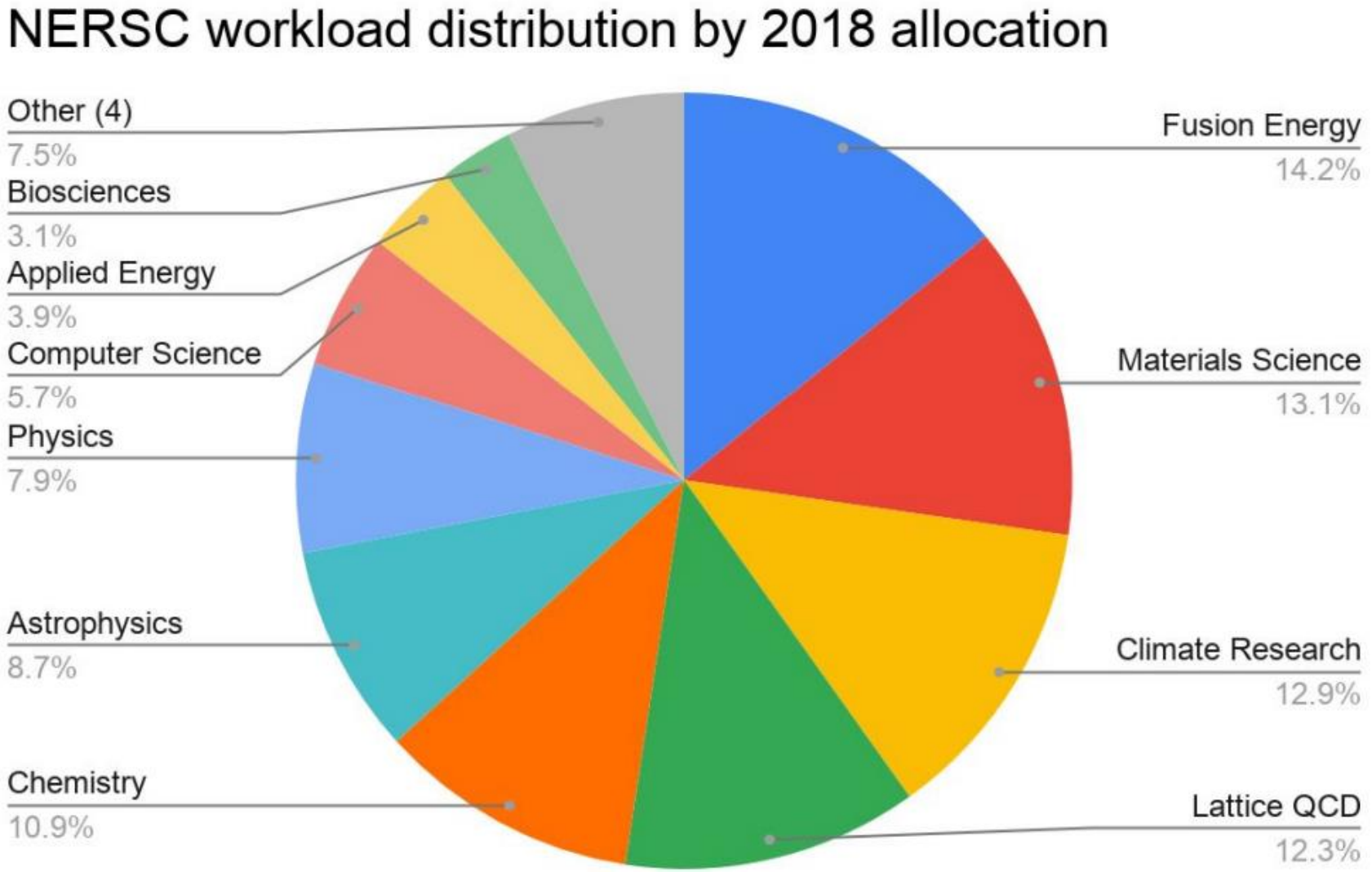
Why Resource Disaggregation

Expensive resources are consistently underutilized



J Li et al., "Analyzing Resource Utilization in an HPC System: A Case Study of NERSC's Perlmutter"

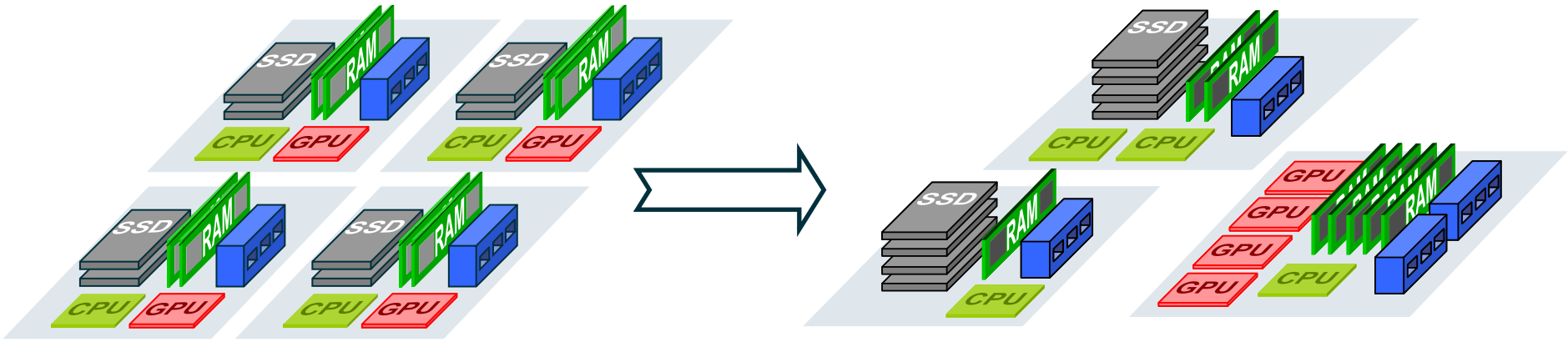
Reminder: NERSC's Workload



4300 active users. 850 projects. 10 codes make up 50% of the workload. 50 codes 84%.

https://portal.nersc.gov/project/m888/nersc10/workload/N10_Workload_Analysis.latest.pdf

Intra-rack resource disaggregation achieves most of the benefit with a fraction of the overhead



Michelogiannaks et al., “A Case for Intra-Rack Resource-Disaggregation in HPC”, ACM TACO 2022

J Li et al., “Analyzing Resource Utilization in an HPC System: A Case Study of NERSC’s Perlmutter”, ISC 2023

Targets of Memory Disaggregation Hardware

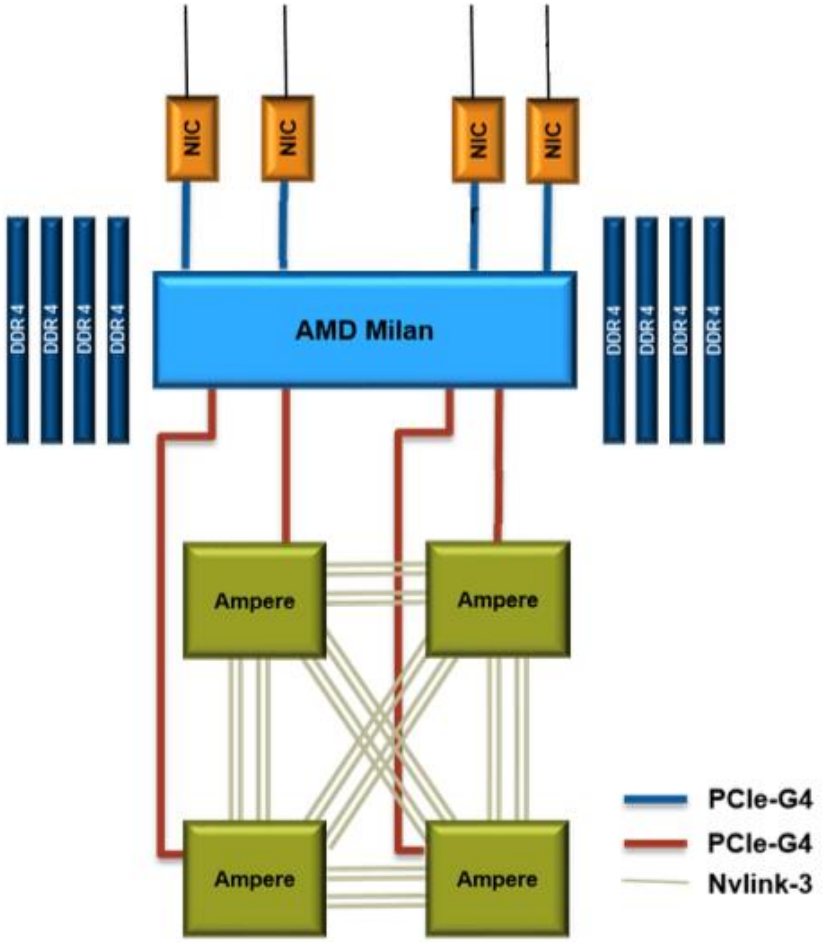
Our design goals

- Satisfy each chip's maximum escape bandwidth
- Achieve comparable BER with today's HPC systems
  - Less than 10^{-18}
  - We use forward error correction (FEC) and take into account its latency
- Minimize energy overhead
- Minimize latency overhead
  - Will be imposed to latency-sensitive communication such as to and from memory

NERSC's Perlmutter GPU Rack

Study based on NERSC's Perlmutter rack:

- 128 GPU accelerated nodes
- Each node has one AMD Milan CPU
- Eight 3200 MHz DDR4 modules per CPU
- Four NVIDIA Ampere A100 GPUs
- Each GPU has 40 GB of co-located HBM



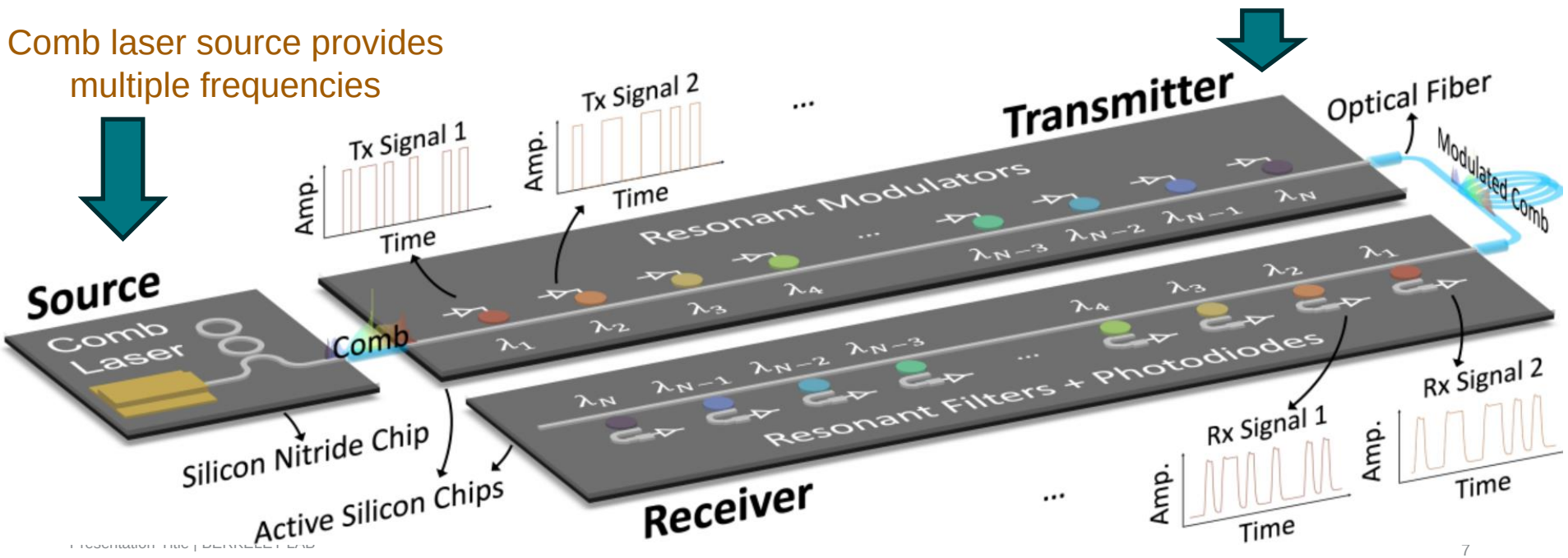
With Emerging Co-Packaged Photonics

To maximize bandwidth density and meet goals

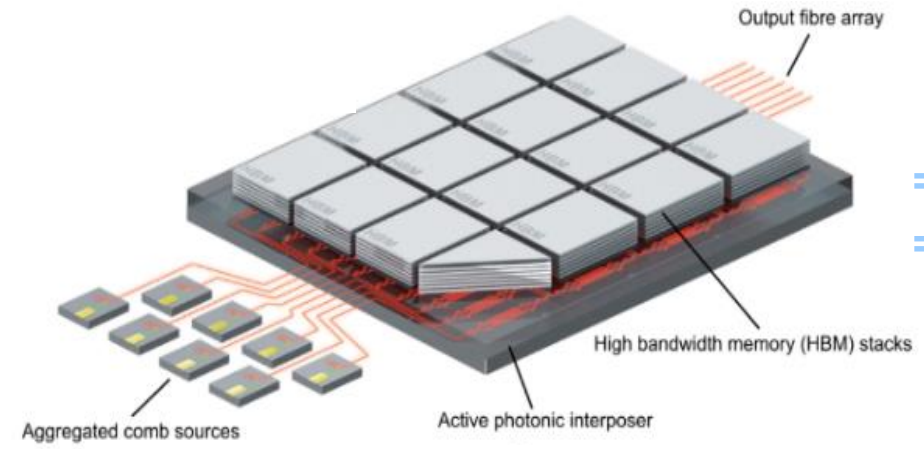
- To maximize bandwidth density, we use prototypes of dense wavelength division multiplexed links that are co-packaged
  - Bandwidths range from 100 Gbps to 2 Tbps
  - Rely on silicon comb laser sources

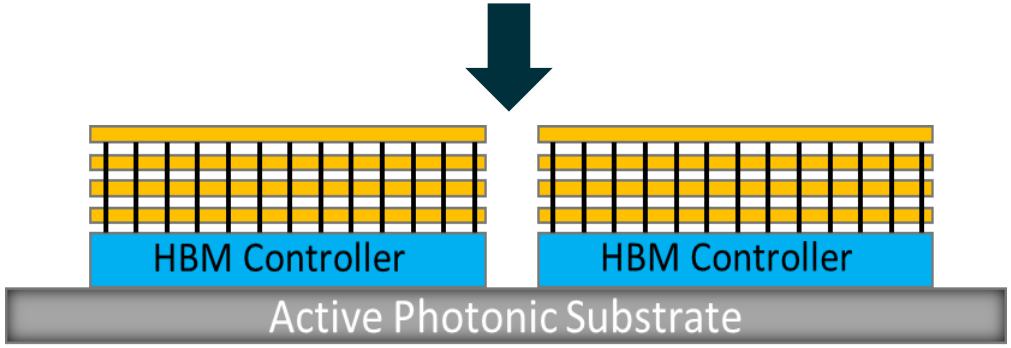
Rings modulate different frequencies

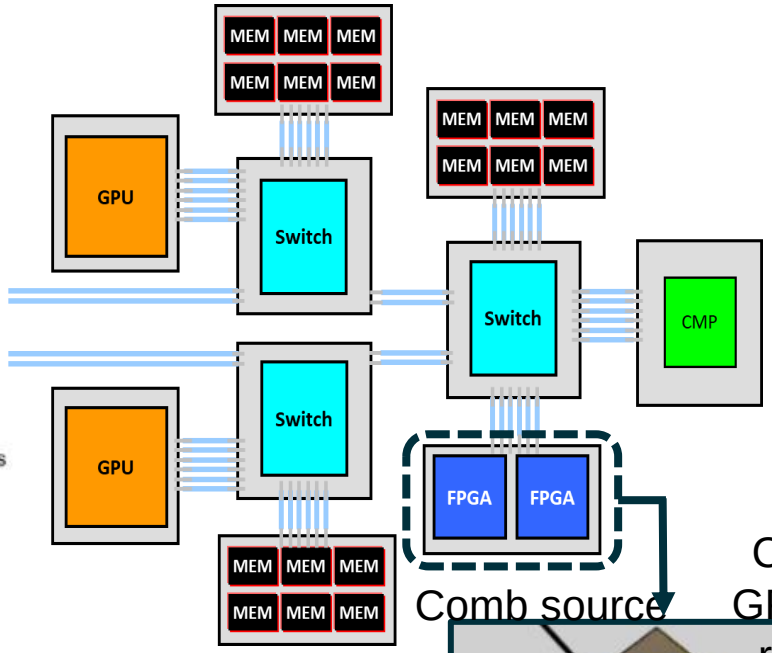
Comb laser source provides multiple frequencies



Embedded Photonic Connectivity

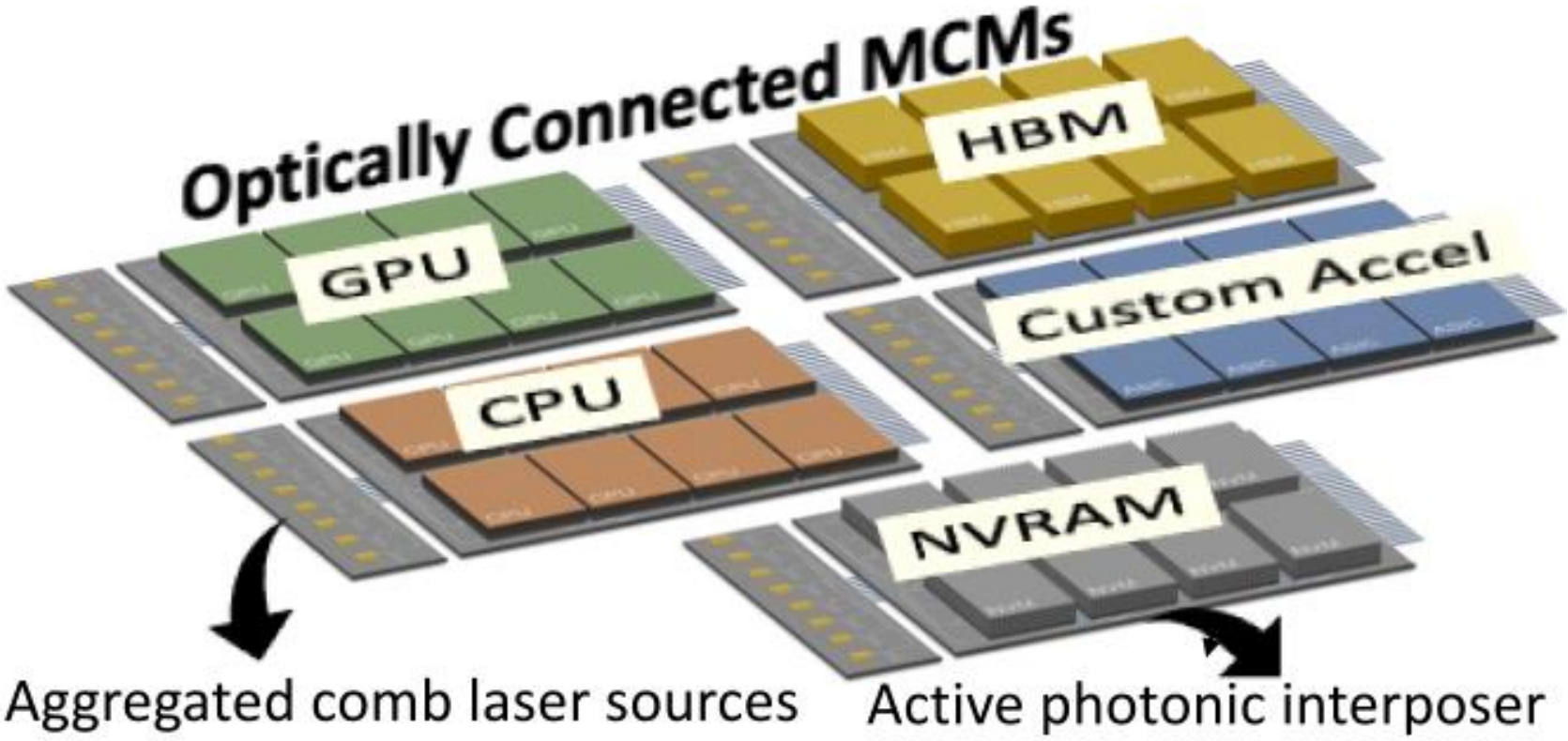






Building up a Rack Using MCMs

Demonstrated in 2.5D and 3D. Can use UCIe or CXL

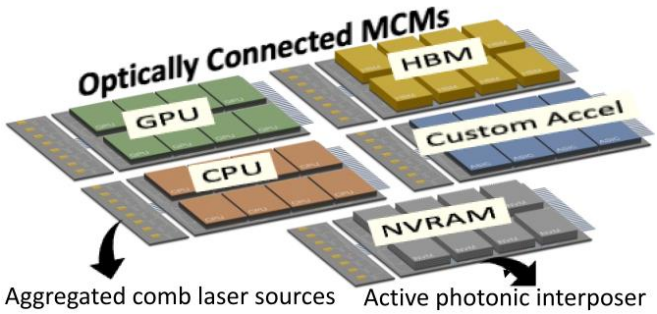


Building up a Rack Using MCMs With Integrated Photonics

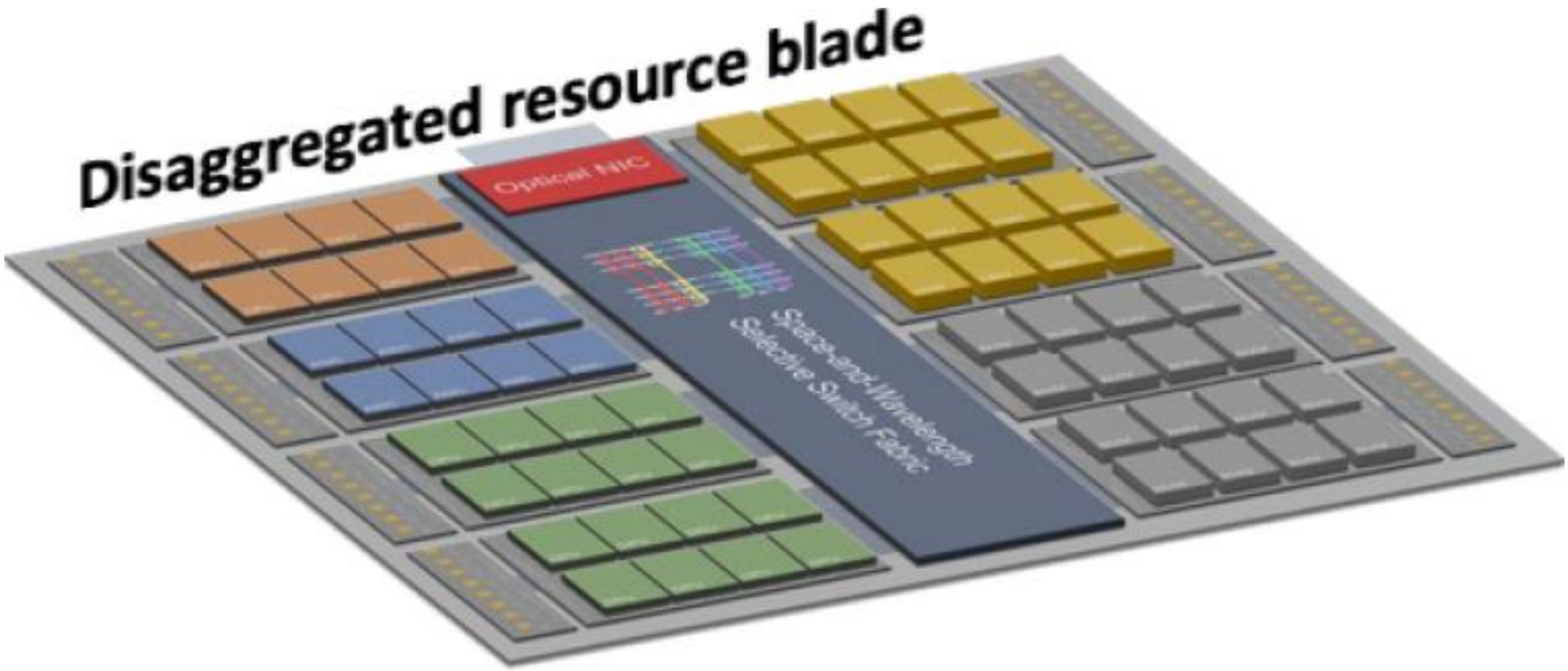
Demonstrated in 2.5D and 3D. Can use UCIe or CXL

Per MCM: 32 fibers, 64 wavelengths, 25 Gbps per

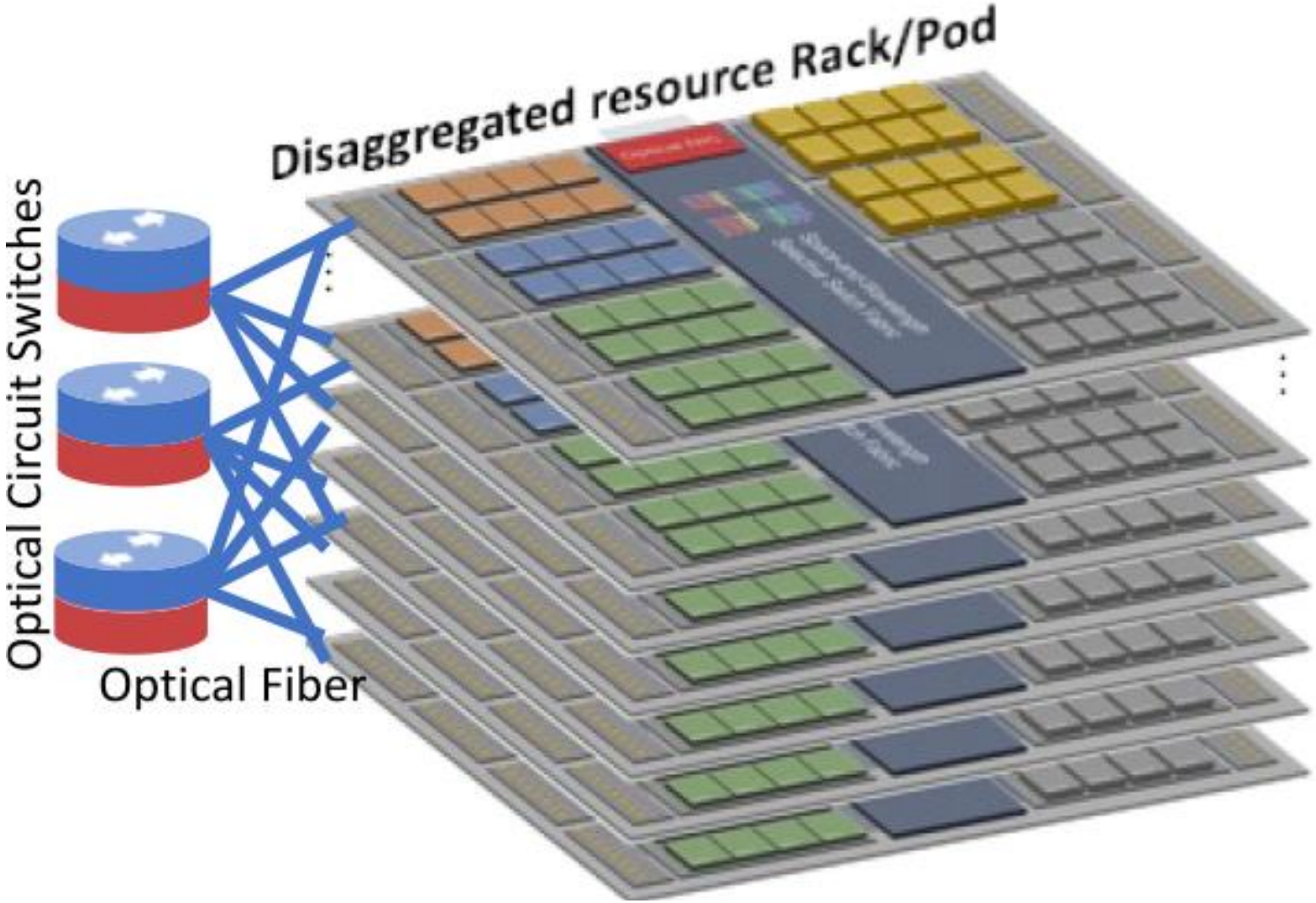
Chip type	Chips per MCM	# MCMs per rack
CPU	14	10
GPU	3	171
NIC	203	3
HBM	4	128
DDR4 (module)	27	38
Total		350



Photonic Switch At the Center of a Blade

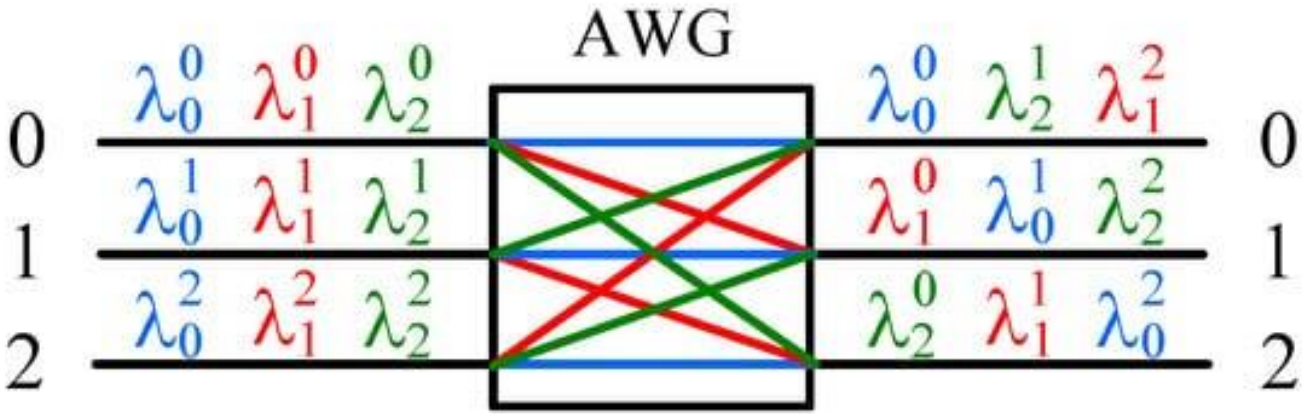


Building Up a Rack

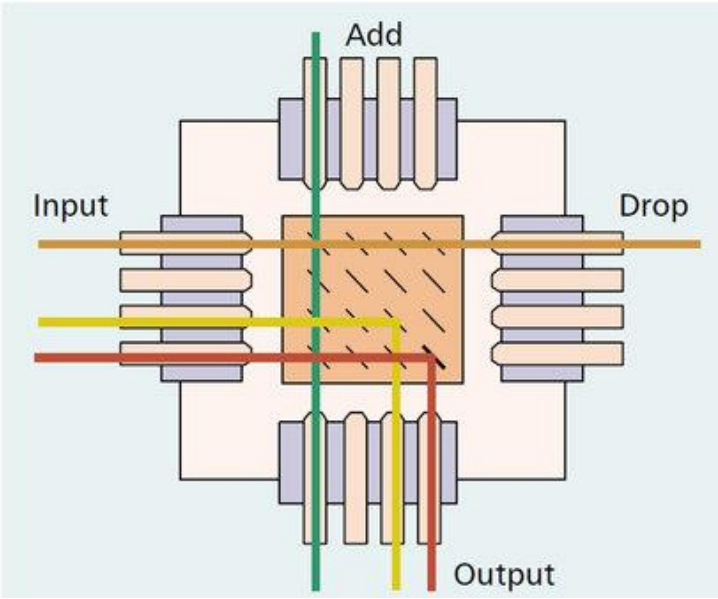


Types of Photonic Switches

Reconfigurable spatial or all-to-all arrayed waveguide grating routers (AWGRs)



B Lin., “Generalization of an Optical ASA Switch”, 2019



Q Cheng et al., “Photonic Switching in High Performance Datacenters”, 2018

Optical Switch State of the Art

- Some switches can achieve more than 25 Gbps per wavelength. We pessimistically use 25 Gbps for all
- We also included wave-selective (few-to-few)

	Switch Type	State of the art
Switch Radix	Cascaded AWGRs Spatial / Wave selective	370 256
Gbps per wavelength	All switches	25
Wavelengths per port	Cascaded AWGRs Spatial / Wave Selective	370 256

Challenges With Optical Switches

Neither spatial nor AWGR are a perfect fit

Spatial Switches

- Quantization of bandwidth
  - I.e., too much bandwidth between some sources, and none between others
- Have to be reconfigured
  - Requires control plane and downtime

AWGRs

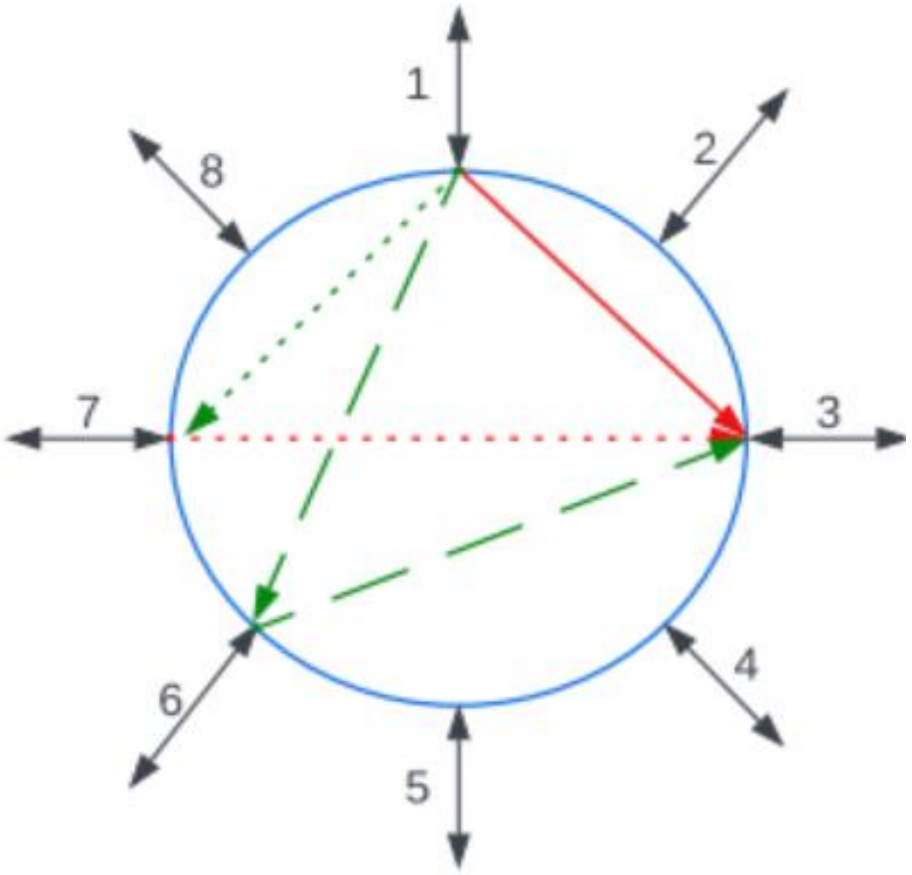
- Low point-to-point bandwidth
  - Poor bandwidth utilization
- Expected to be smaller-radix
  - Spatial switches likely to grow in radix faster

Can we design our rack with AWGRs to avoid reconfiguring spatial switches?

Indirect Routing to Increase AWGR Point-to-Point BW

Local decisions. Have to broadcast or piggyback congestion information

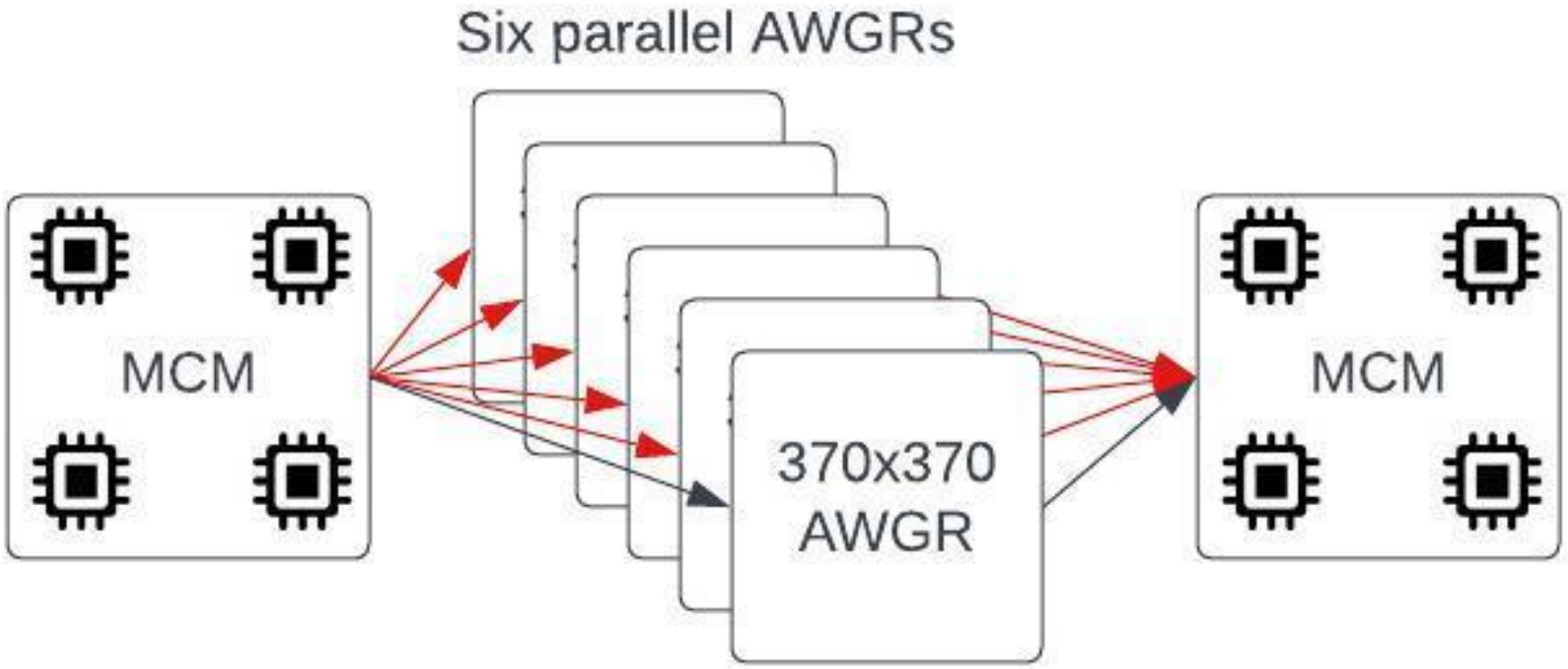
If an endpoint's wavelength to the desired destination is already in use, the endpoint picks an intermediate source whose wavelength to the destination is available



Multiple Parallel AWGRs Satisfy Full Escape BW

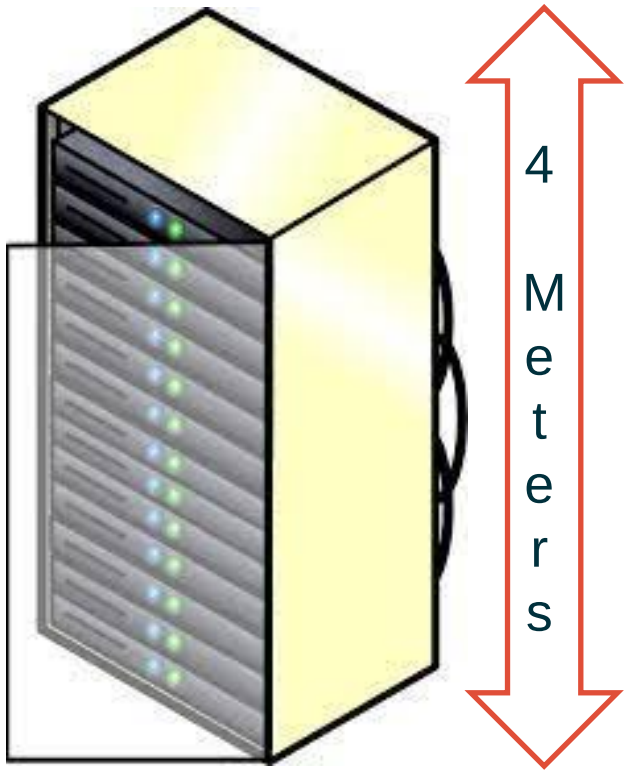
At least five wavelengths between any MCM pair. Thus, 125 Gbps

- 125 Gbps suffices 99.5% of the time between CPU and memory
  - That means, no indirect routing 99.5% of the time
- GPUs use more bandwidth but many pairs (i.e., HBM to HBM) use no bandwidth. After analysis, the probability that no indirect path exists is negligible



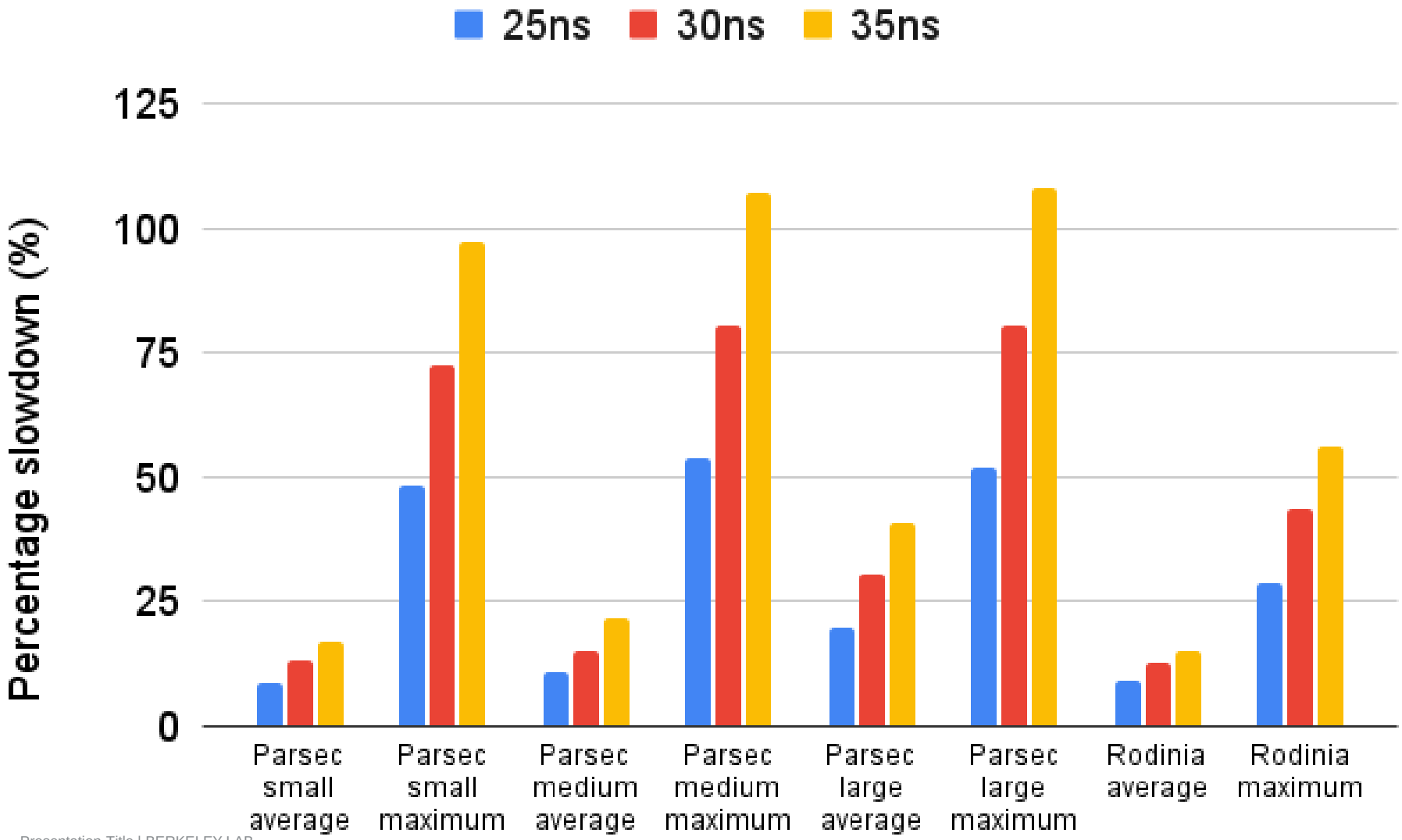
Additional Latency is 35ns

- Target reach for a rack is no more than 4 meters
  - Therefore, 20ns of propagation delay max
- 10ns for serialization delay for 200 Gbps
- FEC (lightweight) latency 2-3ns
- Total: We assume 35ns of end-to-end photonic latency

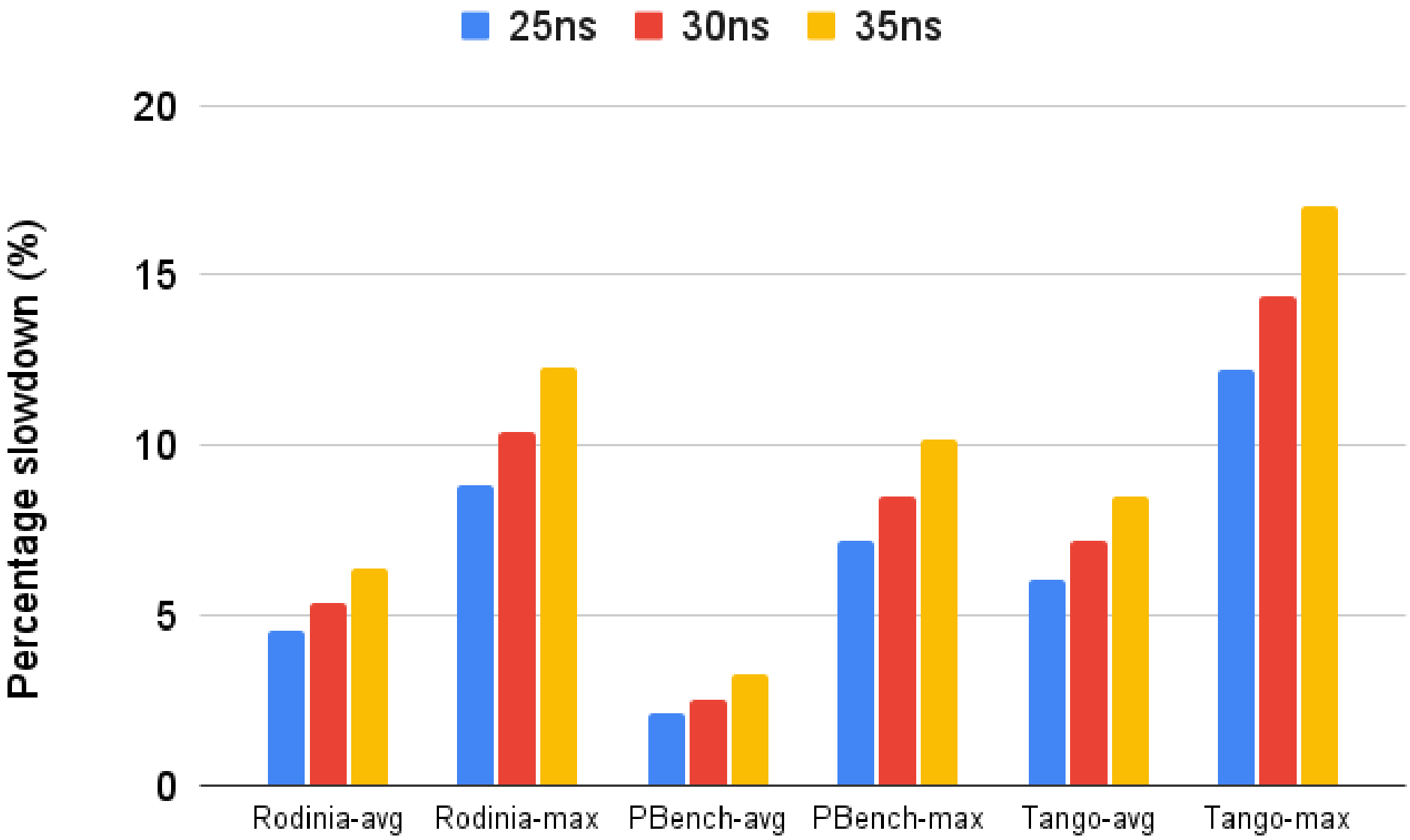


Slowdown for Out of Order (OOO) CPUs

Improving latency with better hardware is important

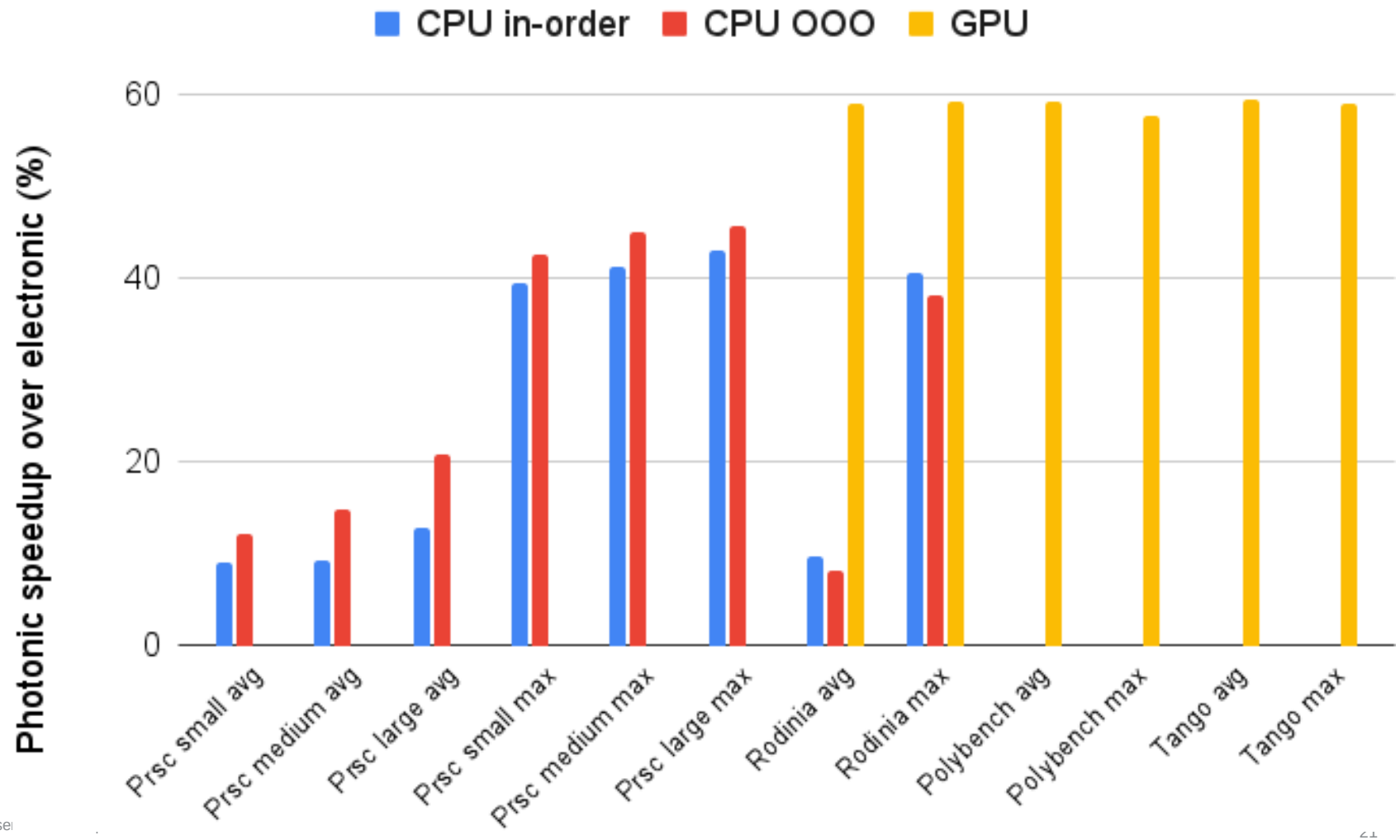


For NVIDIA A100 GPUs



Why Photonics: Compared to Electronic Switches

Electronic switches add 85ns latency (latest PCIe and Anton 3 networks)



Other Results

- Rack power overhead approximately 5%
- Intra-rack disaggregation preserves system throughput and reduces memory modules by 4x and NICs by 2x

Questions?