Intra-Rack Resource Disaggregation Using Emerging Photonics
Intra-Rack Resource Disaggregation Using Emerging Photonics

George Michelogiannakis
Research Scientist
Lawrence Berkeley National Laboratory
Resources in HPC are Underutilized

Resources in HPC are Underutilized

Intra-Rack Disaggregation Suffices

Towards Racks That Pool and Compose

Current server

Current rack

Disaggregated rack

Pool and compose
Photonic-Based Hardware Implementation
Photonic Switch at the Center of a Blade

Multi-chip modules (MCMs) of homogeneous resources and comb laser sources
Scaling Up To a Rack
All-to-All Optical Switch

• Popular example: arrayed waveguide grating routers (AWGRs). All-to-all connectivity. Equal number of wavelengths from every source to every destination

• State of the art example: 400 x 400, 400 wavelengths per port, 25 GHz bandwidth per wavelength

• Y. Hida et al., “400-channel 25-ghz spacing arrayed-waveguide grating covering a full range of c- and l-bands,” in OFC 2001

• Challenge: Single wavelength’s bandwidth may not be adequate. Requires indirect routing

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

- Popular examples: microelectromechanical systems (MEMS), microring resonators (MRRs)
- All wavelengths are connected from one port to another port
- State of the art example (MEMS): 240x240, one wavelength per port
- Challenge: Quantization of bandwidth. For example, 21 wavelengths escape an MCM

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

- “Intermediate” optical switches exist: Wave-selective
- Wave-selective and spatial switches reconfigure in micro to milli seconds
- CXL can be the link protocol
- Forward error correction can achieve necessary bit error rates with only an additional 5ns latency
- With modern photonic links, we meet all design goals. But what about the additional latency?
Speedup Over Electronic Switches

[Bar chart showing speedup over electronic for different benchmarks and architectures: CPU timing, CPU OOO, GPU]
[Call to Action]

• Lets build efforts on the usability and programmability of resource-disaggregated systems: resource allocation algorithms, usage monitoring, programming models, runtime management, etc.

• Contact the speaker: mihelog@lbl.gov

• More information on the project ecosystem:
  
  • https://arpa-e.energy.gov/sites/default/files/Columbia_Bergman_ENLITENED_Annual_Meeting.pdf
Thank you!