Performance Evaluation of Next-generation Data Center and HPC Networks with Co-packaged Optics

Pavlos Maniotis

IBM Thomas J. Watson Research Center, Yorktown Heights, New York, USA

9th OMNeT++ Community Summit, November 2-3, 2022

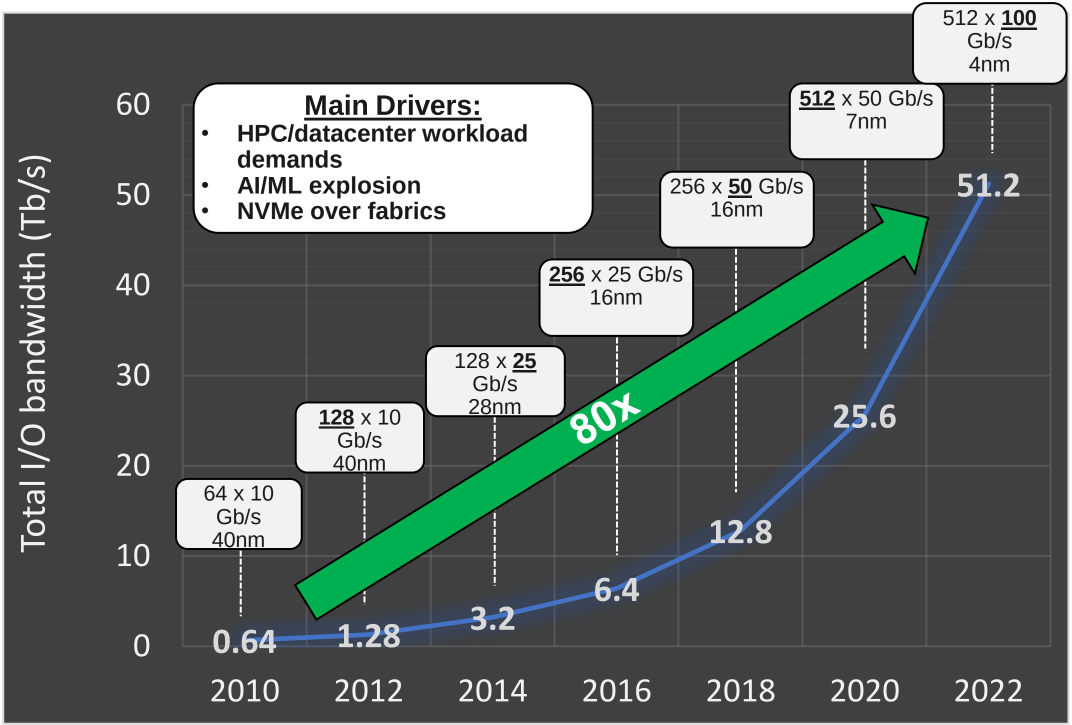
Outline

- Why are we interested in co-packaged optics?
- What are we doing in MOTION research project?
- How much additional bandwidth can we get with co-packaging?
- How is system architecture affected?
- What about performance?
  (a) Network performance analysis with synthetic traffic
  (b) Job placement analysis with VM traces
- Conclusion

Outline

- Why are we interested in co-packaged optics?
- What are we doing in MOTION research project?
- How much additional bandwidth can we get with co-packaging?
- How is system architecture affected?
- What about performance?
  (a) Network performance analysis with synthetic traffic
  (b) Job placement analysis with VM traces
- Conclusion

Switch evolution

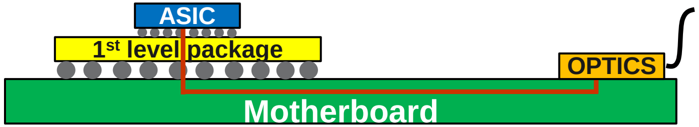


Source: https://www.nextplatform.com/
- /2019/12/12/broadcom-launches-another-tomahawk-into-the-datacenter/
- /2022/04/01/spectrum-4-ethernet-leaps-to-800-gb-sec-with-nvidia-circuits/

- Doubling alternatingly the # of SerDes lanes or the data rate per lane has led to an 80x increase in total switch I/O bandwidth
- Latest switch generation: 51.2 Tb/s 2x data rate per lane + 4nm process
- Demand for further bandwidth scaling is still here and has opened the way to new ideas and solutions (e.g., co-packaged optics for 102.4 Tb/s and beyond)

Why are we interested in co-packaged optics?

Today's approach: pluggable optics



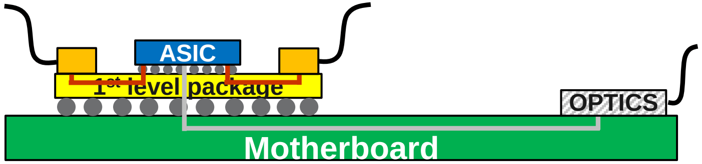
Limiting factors:

(a) Pin density – Larger ASICs are package pin constrained

(b) High power consumption – lengthy wires for driving optics

(c) High cost – optics account for 50% or more of the total cost*

w/ co-packaged optics:



The promise:

(a) An extra dimension for wiring chip pins

(b) Much shorter wires → Low-power SERDES → 25-50% reduction in power consumption over pluggable optics**

(c) Reduced cost through simpler ASICs + I/O modules → 50% reduction in cost per capacity compared to pluggable optics**

* A. Zilkie, High Density Silicon Photonics for Co-packaged Optics and Coherent Optical Engines, ARPA-E ENLITENED Phase 2 Kick-off Meeting, Jan 2021

** C. Minkenberg, et al., (2021), Co-packaged datacenter optics: Opportunities and challenges. IET Optoelectron, 15: 77-91.

Outline

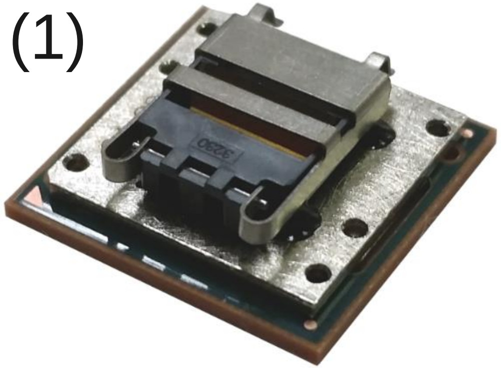
- Why are we interested in co-packaged optics?
- What are we doing in MOTION research project?
- How much additional bandwidth can we get with co-packaging?
- How is system architecture affected?
- What about performance?
  (a) Network performance analysis with synthetic traffic
  (b) Job placement analysis with VM traces
- Conclusion

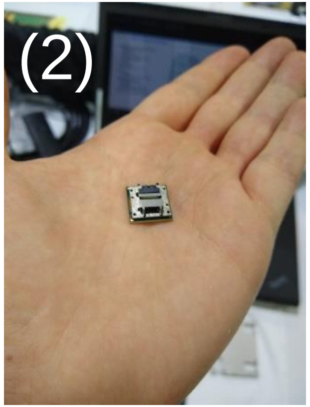
MOTION research project

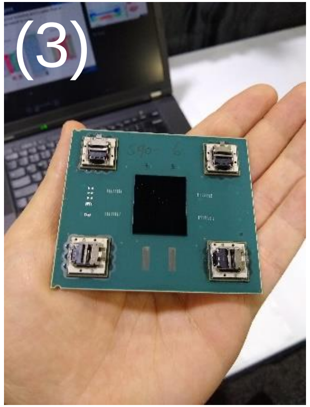
Multi-wavelength **O**ptical **T**ransceivers **I**ntegrated **O**n **N**ode

Target Specifications	
Phase 1	**Phase 2**
16 Channels	32 Channels
NRZ	PAM4
56 GBd / 56 Gb/s per channel	56GBd / 112 Gb/s per channel
0.9 Tb/s per module	3.58 Tb/s per module
BW density: 5.3 Gb/s/mm^2	BW density: 21.2 Gb/s/mm^2
<4 pj/bit (3.2W)	<2 pj/bit (7W)
2 dB Optical margin, >30m w/ connectors	
Temperature: 0-70°C	
WxDxH: 13 x 13 x 4 mm	

(1)





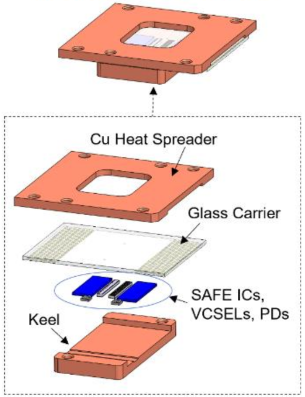


More hardware details:

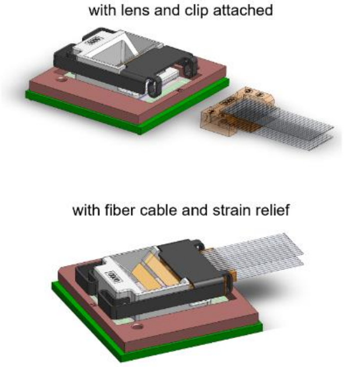
D. Kuchta et al., An 800 Gb/s, 16 Channel, VCSEL-Based, co-Packaged Transceiver With Fast Laser Sparing, Tu1F.1, ECOC 2022

Packaging details

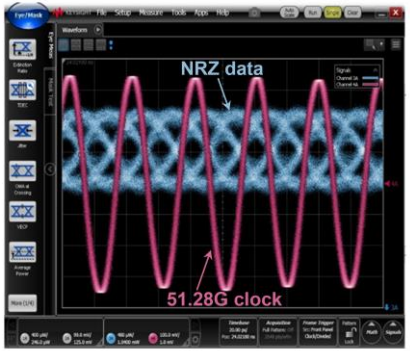
(a) Optical subassembly

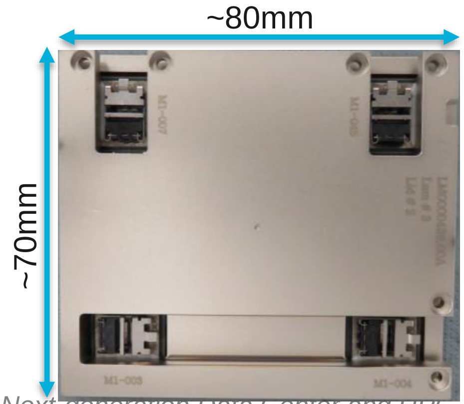


(b) Final assembly



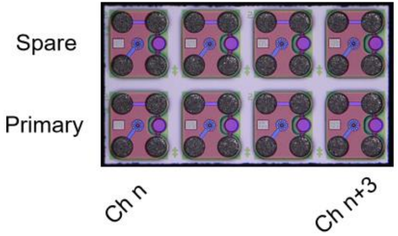
50 Gb/s NRZ data

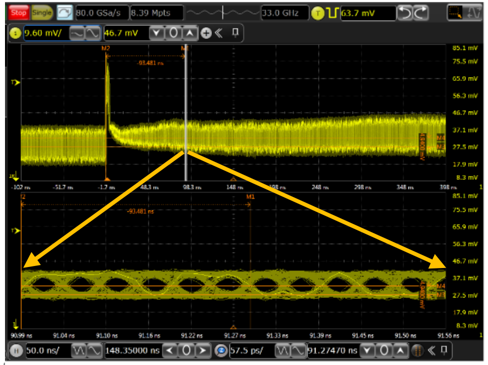




Increased reliability through fast laser sparing

- MOTION has 2:1 laser redundancy on every channel
- Simulation shows ~1000x improvement in reliability at the end of 10 years of service →



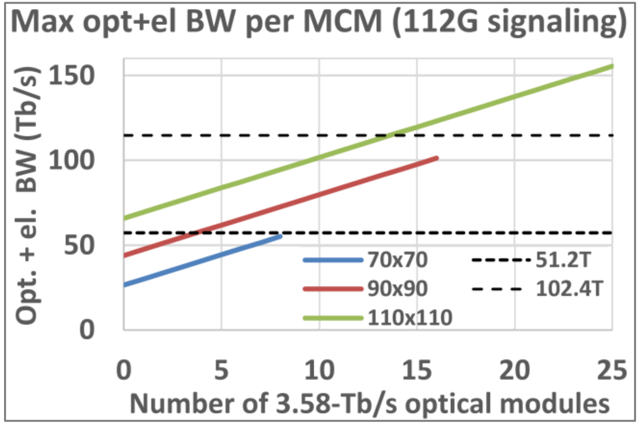


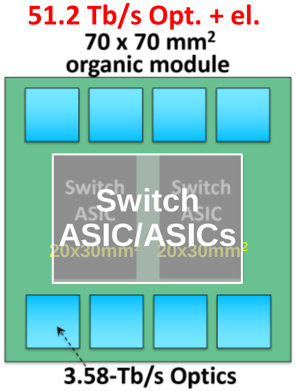
Outline

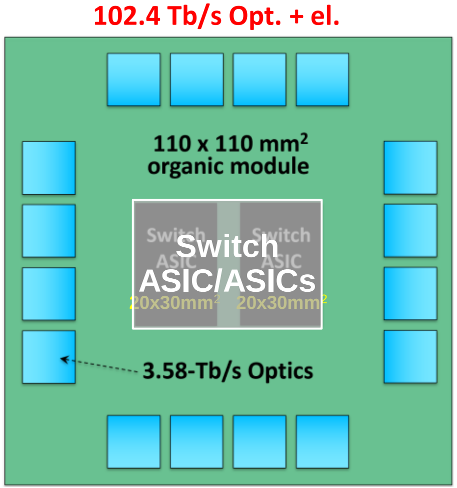
- Why are we interested in co-packaged optics?
- What are we doing in MOTION research project?
- How much additional bandwidth can we get with co-packaging?
- How is system architecture affected?
- What about performance?
  (a) Network performance analysis with synthetic traffic
  (b) Job placement analysis with VM traces
- Conclusion

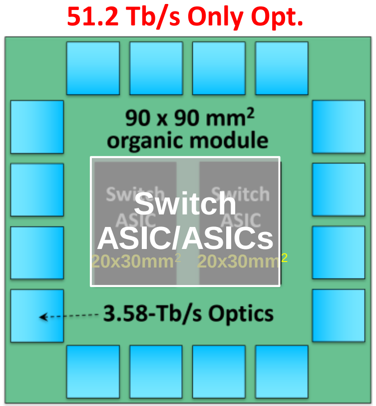
How much additional bandwidth?

	70x70 mm^2	90x90 mm^2	110x110 mm^2
ASICs	2x 20x30 mm^2 Up to 1024 SerDes @ 112 Gb/s signaling		
Pins for high-speed I/O	25%		
Fill factor	40%		
BW density	21.2 Gb/s/mm^2		
Opt. BW	29 Tb/s	57 Tb/s	90 Tb/s
Power Cons.	56 W	112 W	175 W









P. Maniotis, Performance Evaluation of Next-generation Data Center and HPC Networks with Co-packaged Optics, OMNeT++ Summit 2022

Outline

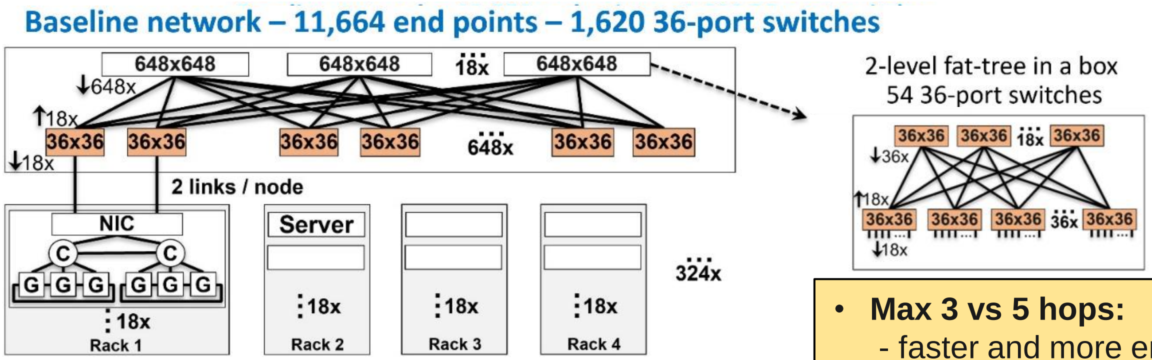
- Why are we interested in co-packaged optics?
- What are we doing in MOTION research project?
- How much additional bandwidth can we get with co-packaging?
- How is system architecture affected?
- What about performance?
  (a) Network performance analysis with synthetic traffic
  (b) Job placement analysis with VM traces
- Conclusion

A case study from the HPC area

Link rate: 100 Gb/s

↓

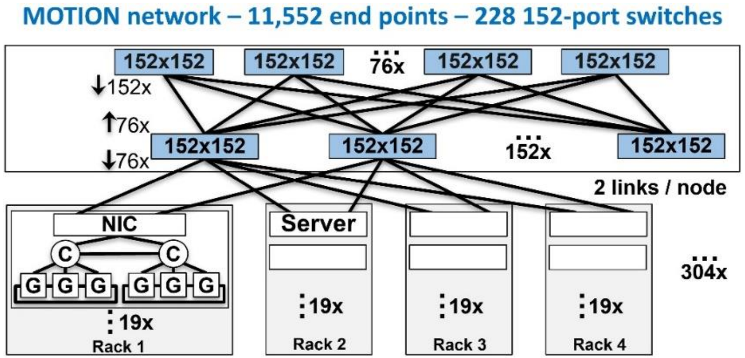
Bisection BW:
1,166.4 Tb/s



Link rate: 400 Gb/s

↓

Bisection BW:
4,620.8 Tb/s



- **Max 3 vs 5 hops:**
  - faster and more energy efficient packet deliveries
  - less network contention
- **86% fewer switch modules**
  - reduced cost / power consumption
  - less administration / management overhead
- **4.2x more servers per 1st-level switch**
  - improved network locality
  - e.g., 1 hop max for >3.5K cores (assuming 48-core servers)
- **4x higher bisection bandwidth**

A case study from the Cloud area

Link rate: 100 Gb/s

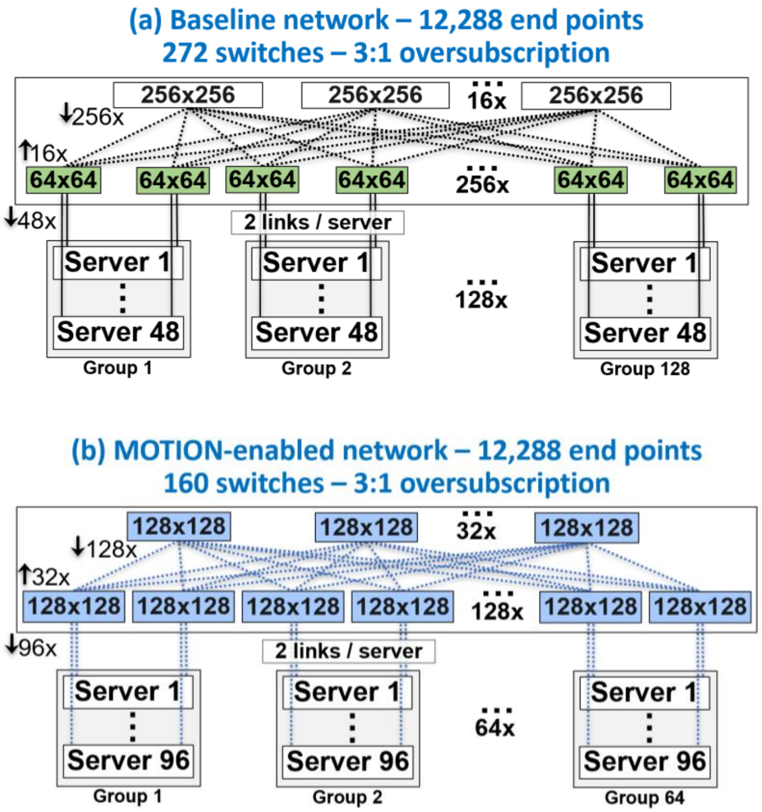
↓

Bisection BW: 409.6 Tb/s

Link rate: 400 Gb/s

↓

Bisection BW: 1,638.4 Tb/s



<u>What's the catch? Any drawbacks?</u>

(a) system availability, and

(b) system security become more important

<u>More servers are affected if a switch goes down or gets compromised</u>

- **Same number of hops**
- **41% fewer switches**
  - reduced cost / power consumption
  - less administration / management overhead
- **2x more servers per 1st-level switch**
  - improved network locality
  - e.g., 1 hop max for >4.5K cores (assuming 48-core servers)
- **4x higher bisection bandwidth**

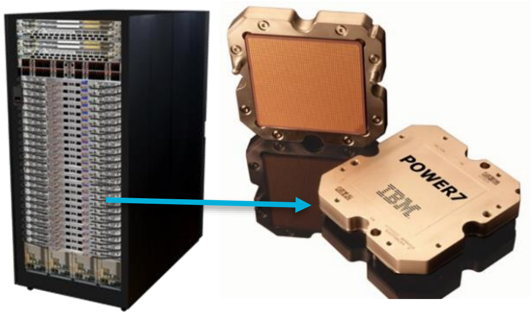
P. Maniotis, Performance Evaluation of Next-generation Data Center and HPC Networks with Co-packaged Optics, OMNeT++ Summit 2022

Outline

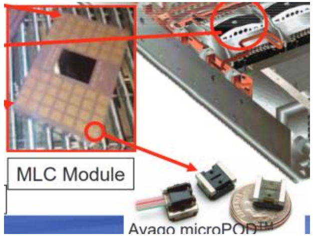
- Why are we interested in co-packaged optics?
- What are we doing in MOTION research project?
- How much additional bandwidth can we get with co-packaging?
- How is system architecture affected?
- What about performance?
  (a) Network performance analysis with synthetic traffic
  (b) Job placement analysis with VM traces
- Conclusion

Simulation setup

PERCS HPC Cluster

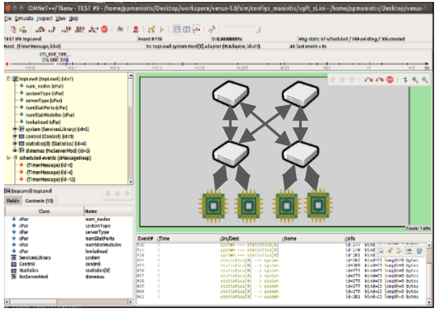


- World's first HPC with co-packaged optics
- Developed as part of the PERCS program and released in 2011 (Productive, Easy-to-use, Reliable Computing System)
- 96 computing nodes organized in 12 drawers
- 12 TB RAM (128 GB / computing node)



A. Benner, "Optical interconnect opportunities in supercomputers and high end computing," OFC/NFOEC, 2012, pp. 1-60.

Venus network simulator



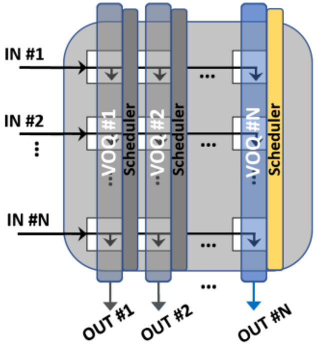
- Discrete event simulator built on top of OMNEST (140K lines of C/C++ code)
- Developed at IBM Research Zurich Labs. Has been used during the development of multiple HPC systems
- Fat tree, XGFT, Mesh, Multi-dimensional mesh, Hypercube, Torus, Dragonfly(+), Flattened butterfly
- Ethernet, Infiniband, Myrinet, Optically interfaced switches, Optical switches

*R. Birke, et.al., "Towards massively parallel simulations of massively parallel high-performance computing systems," *SIMUTOOLS '12, ICST*, 2012, Brussels, BEL, 291–298.

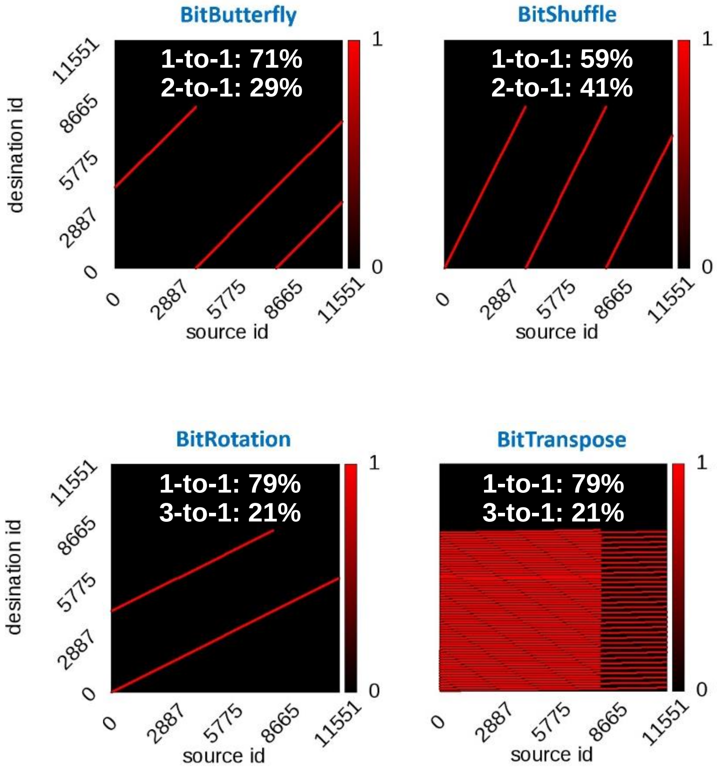
Simulation setup

Traffic Generators	
Packet size	1,500 bytes
Arrivals distribution	Geometric
Rate	100, 400 Gb/s
Load	[0.1-1]

Network Interface Cards	
Buffer size	512 KB
Delay	100 ns
BW	100, 400 Gb/s

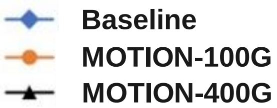


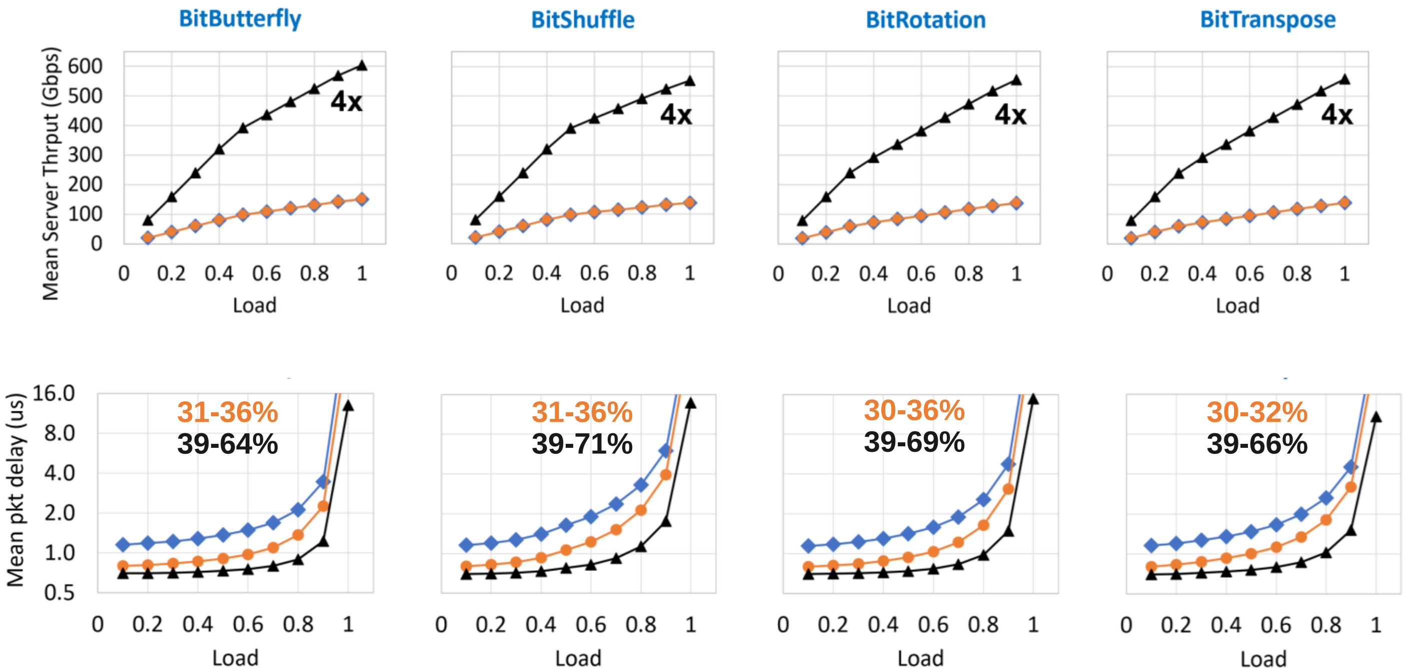
Switches	
Radix	36x36, 152x152
Architecture	Input buffers with VOQs
Flow control	Credit-based
Routing	Random
Buffer size	128 KB / port
Delay	100 ns
BW per port	100, 400 Gb/s



* For patterns see: *Principles and Practices of Interconnection Networks* from W. J. Dally and B. P. Towles

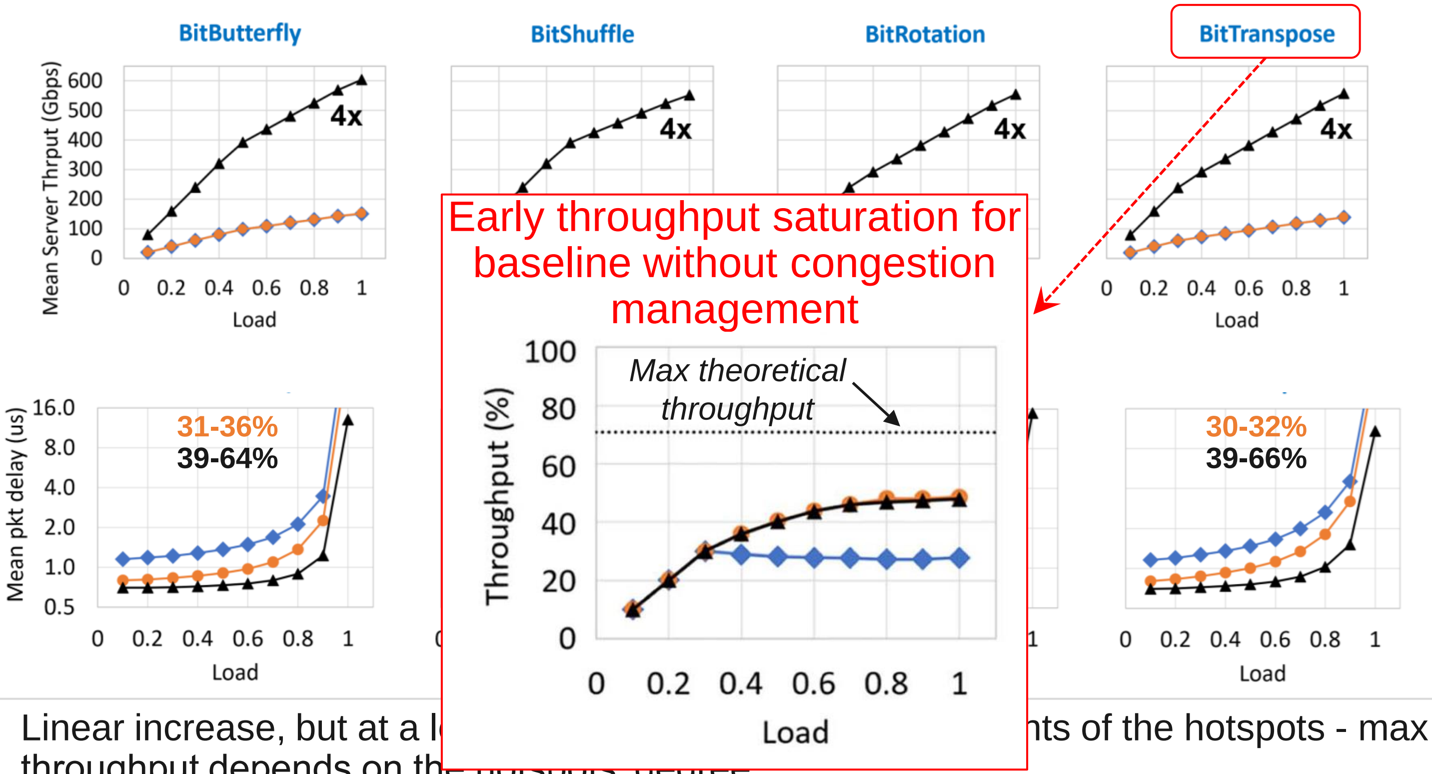
Simulation results





- Linear increase, but at a lower rate beyond the saturation points of the hotspots - max throughput depends on the hotspots' degree
- 4x better higher throughput performance in terms of absolute throughput
- Significant improvements of up to 71% for mean packet delay

Simulation results



- Linear increase, but at a l... nts of the hotspots - max throughput depends on the hotspots' degree
- 4x better higher throughput performance in terms of absolute throughput
- Significant improvements of up to 71% for mean packet delay

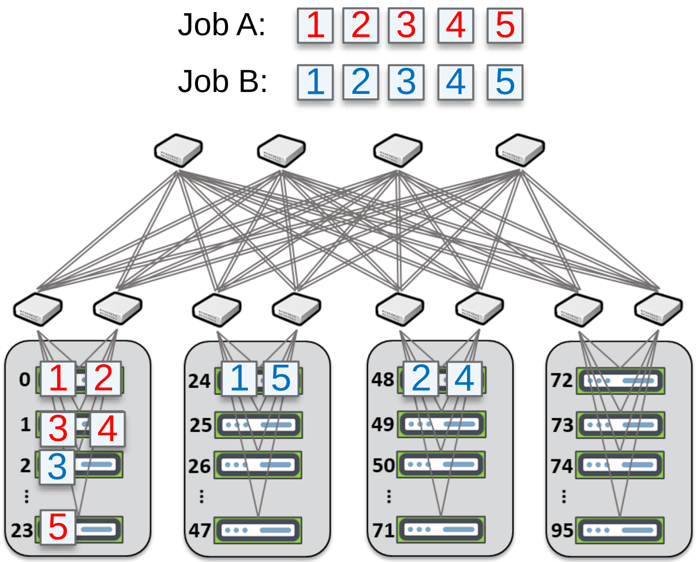
Outline

- Why are we interested in co-packaged optics?
- What are we doing in MOTION research project?
- How much additional bandwidth can we get with co-packaging?
- How is system architecture affected?
- What about performance?
  (a) Network performance analysis with synthetic traffic
  (b) Job placement analysis with VM traces
- Conclusion

Simulation setup

M. C. Silva Filho, et al., "**CloudSim Plus**: A cloud computing simulation framework pursuing software engineering principles for improved modularity, extensibility and correctness," 2017 IFIP/IEEE Symposium on Integrated Network and Service Management (IM), 2017, pp. 400-406

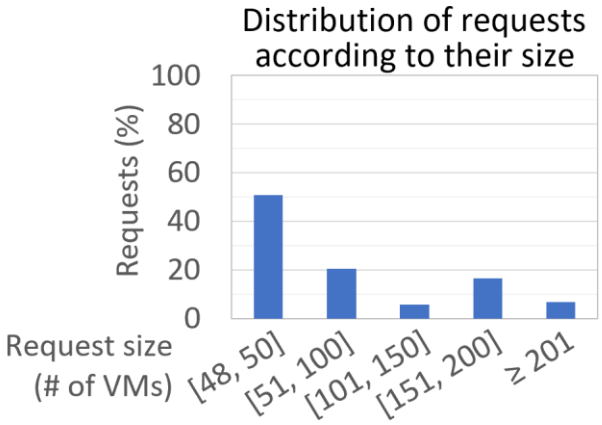
<u>VM placement example</u>



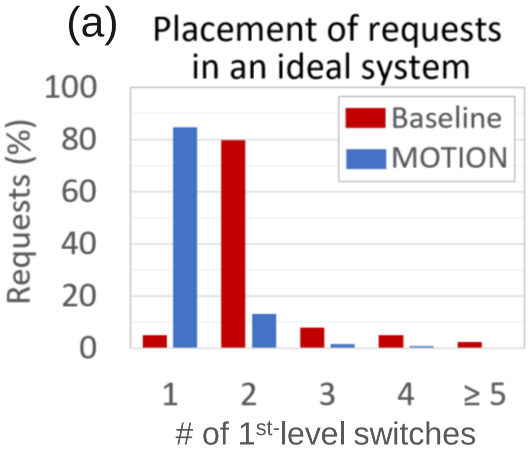
- Cost is 1 hop max
- No spine crossing

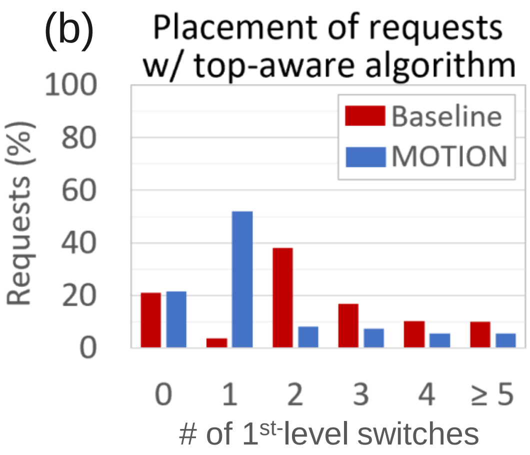
- Cost is up to 3 hops
- Need to cross the spine

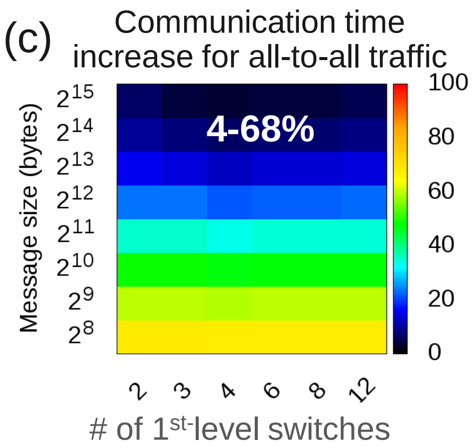
- VM trace from the publicly available *AzureTracesForPacking2020* dataset https://github.com/Azure/AzurePublicDataset/blob/master/AzureTracesForPacking2020.md
- 630 VM group requests from a 7-day period (>62.5K VMs)
- Interarrival times: [min / avg / max / stddev] → [0s / 14.1m / 1.2d / 1.27h]
- Lifetimes: [min / avg / max / stddev] → [3s / 1.49d / 89.9d / 8.17d]
- Server configuration: 48 cores, 384 GiB RAM, 100 or 400 Gb/s / NIC



Simulation results







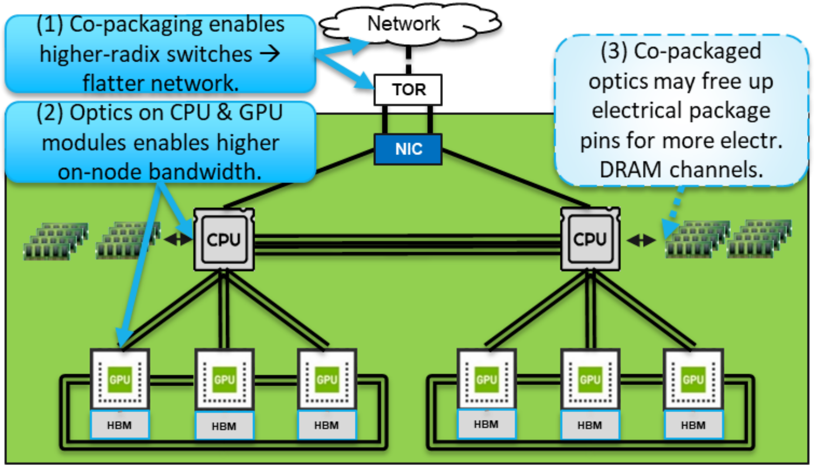
- Placing the VMs under the same 1st-level switch has 2 key advantages:
  - Cost is 1 hop max
  - No spine crossing
- High-radix switches can become a game changer in terms of network locality

Outline

- Why are we interested in co-packaged optics?
- What are we doing in MOTION research project?
- How much additional bandwidth can we get with co-packaging?
- How is system architecture affected?
- What about performance?
  (a) Network performance analysis with synthetic traffic
  (b) Job placement analysis with VM traces
- Conclusion

Conclusion

- Co-packaged optics can help in continuing bandwidth scaling in future HPC and data center networks
- Advantages in network architecture
  - (a) Simpler networks w/ fewer switch layers
  - (b) Higher bisection bandwidth
  - (c) Reduced switch count
  - (d) Improved network locality properties
- Advantages can be transferred to: NICs, CPUs, GPUs or other accelerators. More research needed in these areas.



Possible insertion points of co-packaged optics in HPC nodes

Acknowledgements: MOTION 1 & 2 team members & sponsor

- IBM Research
  - C. Baks, A. Benner, R. Budd, T. Dickson, F. Doany, W. Lee, M. Meghelli, P. Pepeljugoski, J. Proesel, M. Taubenblatt, L. Schares, M. Schultz, P. Maniotis, P. Stark, H. Ainspan, Z. Toprak Deniz, S. Dhawan, T. Dickson, N. Dupuis, P. Francese, B. Sadhu, M. Kossel, T. Morf, M. Brändli, S. Rylov, M. Cochet, C. Ozdag, A. Watanabe, H. Rahmani, D. Kuchta,

- IBM Bromont
  - L-M. Achard, P. Fortier, C. Dufort, E. Tucotte, C. Bureau, M. Pion, Y. Cossette, P. Ducharme, S. Desputeau, A. Janta-Polczynski, P. Minier, G. Jutras, P. McInnes, S. Whitehead, B. Sow

- IBM Server Packaging

  B. Parikh, S. Ostrander, S. Li, C. Setzer, H. Toy, J. Ross, K. Lange, M. Kapfhammer, B. Meiswinkel, C. Muzzy, E. Steiner, D. Smith, T. Saunders, G. Pomerantz, J. Coffin, K. Marston, K. Smith, T. Ahmed, L. Rapp, Y.Yao, T. Wassick, M. Warbrick, T. Lombardi, B. Singh, C. Walker, S. Iruvanti, D. Yannitty, P. Ramaglia, T. Weiss, M. Interrante, J. Sorbello, C. Arvin, M. Stalter, A. Perez, P. Torbet, T. Olowofela, J. Bunt, M. Fisher, T. Childress, J. Mingo, C. Savoy, S. Ruiz, D.Kohler, R. Seifts, R. Olson, H. Polgrean, J. Rowland, C. Thomas, E. Kastberg, A. Schetter, D. Babcock, A. Greenberg, D. Lord, R. Rodriguez, C. Taylor

- IBM Server Development
  - D. Becker, R. Laning, D. Dreps, M. Hoffmeyer, J. Eagle, F. Gholami, K. O'Connell, S. Canfield, S. Chun, R. Frota

- IBM Supply Chain Engineering
  - H Bagheri, K. Akasofu, C. Grosskopf, A. Tiano, T. Sass, E. Mallery

- II-VI Finisar Corporation

  F. Flens, D. Case, P. Chen, J. Glover, C. Kocot, K. Koski, G. Light, T. Nguyen, S. Pandy, S Quadery, K. Szczerba, B. Wang, P Westbergh, S. Pandey, H. Hayashigatani (131)

- Texas A&M University
  - N. Kim, A. Kumar, T. Liu, I. Yi, S. Palermo

Acknowledgment: "The information, data, or work presented herein was funded in part by the Advanced Research Projects Agency-Energy (ARPA-E), U.S. Department of Energy, under Award Number DE-AR0000846. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof."

CHANGING WHAT'S POSSIBLE