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
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

- Doubling alternatingly the number 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 Tbps; 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-packed optics for 102.4 Tbps and beyond).

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/
Why are we interested in co-packaged 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*

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
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
## Target Specifications

<table>
<thead>
<tr>
<th>Phase 1</th>
<th>Phase 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>16 Channels</td>
<td>32 Channels</td>
</tr>
<tr>
<td>NRZ</td>
<td>PAM4</td>
</tr>
<tr>
<td>56 GBd / 56 Gb/s per channel</td>
<td>56GBd / 112 Gb/s per channel</td>
</tr>
<tr>
<td>0.9 Tb/s per module</td>
<td>3.58 Tb/s per module</td>
</tr>
<tr>
<td>BW density: 5.3 Gb/s/mm²</td>
<td>BW density: 21.2 Gb/s/mm²</td>
</tr>
<tr>
<td>&lt;4 pj/bit (3.2W)</td>
<td>&lt;2 pj/bit (7W)</td>
</tr>
<tr>
<td>2 dB Optical margin, &gt;30m w/ connectors</td>
<td></td>
</tr>
<tr>
<td>Temperature: 0-70°C</td>
<td></td>
</tr>
<tr>
<td>WxDxH: 13 x 13 x 4 mm</td>
<td></td>
</tr>
</tbody>
</table>
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
Cu Heat Spreader
Glass Carrier
SAFE ICs, VCSELs, PDs
Keel

(b) Final assembly
with lens and clip attached
with fiber cable and strain relief

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

50 Gb/s NRZ data

~80mm

~70mm
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?

<table>
<thead>
<tr>
<th></th>
<th>70x70 mm²</th>
<th>90x90 mm²</th>
<th>110x110 mm²</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASICs</td>
<td>2x 20x30 mm²</td>
<td>Up to 1024 SerDes @ 112 Gb/s signaling</td>
<td></td>
</tr>
<tr>
<td>Pins for high-speed I/O</td>
<td>25%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fill factor</td>
<td>40%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BW density</td>
<td>21.2 Gb/s/mm²</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Opt. BW</td>
<td>29 Tb/s</td>
<td>57 Tb/s</td>
<td>90 Tb/s</td>
</tr>
<tr>
<td>Power Cons.</td>
<td>56 W</td>
<td>112 W</td>
<td>175 W</td>
</tr>
</tbody>
</table>
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

**Baseline network – 11,664 end points – 1,620 36-port switches**

- Link rate: 100 Gb/s
  - Bisection BW: 1,166.4 Tb/s

**MOTION network – 11,552 end points – 228 152-port switches**

- 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

<table>
<thead>
<tr>
<th>Link rate: 100 Gb/s</th>
<th>Bisection BW: 409.6 Tb/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Link rate: 400 Gb/s</td>
<td>Bisection BW: 1,638.4 Tb/s</td>
</tr>
</tbody>
</table>

(a) Baseline network – 12,288 end points
272 switches – 3:1 oversubscription

(b) MOTION-enabled network – 12,288 end points
160 switches – 3:1 oversubscription

What’s the catch? Any drawbacks?
(a) system availability, and
(b) system security become more important

More servers are affected if a switch goes down or gets compromised

- 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

What’s the catch? Any drawbacks?
(a) system availability, and
(b) system security become more important

More servers are affected if a switch goes down or gets compromised

- 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)

**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

---


Simulation setup

<table>
<thead>
<tr>
<th>Traffic Generators</th>
<th>BitButterfly</th>
<th>BitShuffle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Packet size</td>
<td>1-to-1: 71%</td>
<td>1-to-1: 59%</td>
</tr>
<tr>
<td>Arrivals distribution</td>
<td>2-to-1: 29%</td>
<td>2-to-1: 41%</td>
</tr>
<tr>
<td>Rate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Load</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[0.1-1]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Network Interface Cards</th>
<th>BitRotation</th>
<th>BitTranspose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Buffer size</td>
<td>1-to-1: 79%</td>
<td>1-to-1: 79%</td>
</tr>
<tr>
<td>Delay</td>
<td>3-to-1: 21%</td>
<td>3-to-1: 21%</td>
</tr>
<tr>
<td>BW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100, 400 Gb/s</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

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

Early throughput saturation for baseline without congestion management

Max theoretical throughput
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

VM placement example

Job A: 1 2 3 4 5
Job B: 1 2 3 4 5

- 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

Distribution of requests according to their size

- Cost is 1 hop max
- No spine crossing

- Cost is up to 3 hops
- Need to cross the spine
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.
Acknowledgements: MOTION 1 & 2 team members & sponsor

- IBM Research

- IBM Bromont

- IBM Server Packaging

- IBM Server Development

- IBM Supply Chain Engineering

- II-VI Finisar Corporation

- 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.”