Rack-Scale Optical Network for High Performance Computing Systems

Peng Yang, Zhengbin Pang, Zhifei Wang
Zhehui Wang, Min Xie, Xuanqi Chen, Luan H. K. Duong, Jiang Xu
Outline

• Introduction
• Rack-scale inter/intra-chip optical network (RSON)
• Evaluation and analysis
• Conclusion
Scientific computing, big data, and AI applications
  - Weather prediction, SNS, deep neural networks...

HPC and cloud computing
  - Sunway TianhuLight with ~10M cores (1ST, Nov 2017)

Exascale system requirements:
  - Total computing power: ExaFlop/s
  - Energy efficiency: 50 GFLOPS/W
  - Interconnection side: ~20 pJ/bit
Rack-scale computing system

• Rack-scale computing system:
  • Merged scale-up and scale-out model
  • Close-coupled computing and communication
  • Widely employed standardized structure: HP Moonshot, Intel RSA

• Intra-rack communication is becoming the bottleneck
  • ~80% traffic stays within the rack in cloud computing
  • Efficient and low latency access to intra-rack resource
  • Dominant power in system interconnection
Optical interconnection for rack/large scale systems

• Available solutions
  • Ethernet, Fibre Channel, InfiniBand...
  • Electrical Top-of-Rack (ToR) switch
  • ENoC for multicore chips

• Optical interconnection to address the electrical link challenges
  • High energy consumption
  • Limited data rate
  • Chip pin counts
  • High latency

1 April 2018
Peng Yang

[Z. Wang et al, 2015]
Related works on optical interconnected rack

- Inter/intra-chip optical network:
  - Inter and intra-chip optical network [X. Wu et al, 2013; X. Wu et al, 2015]

  - Optical switch and path control is oversimplified
  - Electrical intra-chip network

- The unified inter/intra-chip rack-scale optical network is not fully investigated!
  - Different traffic for on-chip and off-chip communication
  - Co-consider inter-chip and intra-chip network architectures
  - Detailed control for active switching
Introduction

Rack-scale inter/intra-chip optical network (RSON)

Evaluation and analysis

Conclusion
Rack-scale optical network architecture overview

- High-radix integrated optical switch fabric
- Unified inter/intra-chip optical interconnect
- Hybrid electrical/optical NoC
- Multi-domain circuit switching
• Sufficient computing power and memory space
  • Multiple core clusters: four cores with I/D $\$
  • Multiple memory controllers

• Hybrid and complementary ENoC and ONoC
  • ENoC for legacy connection of on-chip resources
  • ONoC connecting memory controllers and inter-node interface
  • High-bandwidth bypass for memory access between on-chip and off-chip domains

Hybrid E/O server node network architecture

1 April 2018 Peng Yang
Optical inter-node interface

- Efficient interaction between on-chip and off-chip networks
  - Similar architecture as ONoC interface
  - Hold ONoC controller to control channels
  - Handle all path requests to the optical switch

- Essential to break the performance gap between intra-chip and inter-chip domain

- Avoiding the power hungry E/O/E conversion
• Power efficient E/O/E conversion
  • Shared Tx & Rx among different data channels
  • Optical weaving serdes
• Merged on-chip and off-chip domain
  • Bidirectional switching between horizontal and vertical links
• Efficient control signals exchanging
  • Unique wavelength for each ONoC interface via channel M
Optical switch control

- Requiring fast control signaling between server node and switch
  - Separate control and data links
  - Internal forwarding request/grant

- Fully desynchronized round-robin arbitration
  - Reduce the synchronized contention for the same output port/path
  - Increase the path reservation success ratio

- Efficient implementation of the 64x64 arbiter
  - Power: ~250 mW
  - Area: ~1.07 mm$^2$
ONoC channel partition

- Reduce the arbitration overhead
  - Divide channels into partitions
  - Scalable and low arbitration complexity, $O(1)$
- Fulfill different traffic demand
  - Dynamically allocate data channels to each partitions
- Different priority for specific traffic
Control implementation

- Reduce control system latency and increase processing efficiency
  - Pipelined structure
  - Parallel for each interface agent and channel section solver
- Low power overhead: ~61 mW for 16 interfaces
Outline

• Introduction
• Rack-scale inter/intra-chip optical network (RSON)
• Evaluation and analysis
• Conclusion
Simulation environment and setup

- JADE full-system simulator
- InfiniBand and optical switch for comparison
- APEX benchmark
- McPAT and CACTI for electrical power evaluation
- 64-node, bandwidth from 100 to 800Gbps: 25 Gbps/wavelength

<table>
<thead>
<tr>
<th>Item</th>
<th>Configuration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core cluster</td>
<td>8, 4 cores/cluster</td>
</tr>
<tr>
<td>L1 I/D $</td>
<td>64KB/core, private</td>
</tr>
<tr>
<td>L2 $</td>
<td>512KB/cluster, shared</td>
</tr>
<tr>
<td>LLC</td>
<td>2.5MB/slice</td>
</tr>
<tr>
<td>Coherence protocol</td>
<td>Directory based MOSI</td>
</tr>
<tr>
<td>Electrical link</td>
<td>128-bit width/direction</td>
</tr>
<tr>
<td>Memory bandwidth</td>
<td>480GB/s/port</td>
</tr>
<tr>
<td>Memory capacity</td>
<td>16GB/port</td>
</tr>
</tbody>
</table>
• RSON achieves most performance, 6.8X at best
• High bandwidth interconnect benefits RSON more
Interconnection energy efficiency

<table>
<thead>
<tr>
<th>Item</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laser efficiency</td>
<td>0.33 [48,49]</td>
</tr>
<tr>
<td>MR insertion loss</td>
<td>1~1.8 dB</td>
</tr>
<tr>
<td>MR passing loss</td>
<td>0.06~0.3 dB</td>
</tr>
<tr>
<td>Edge coupling loss</td>
<td>1~1.5 dB</td>
</tr>
<tr>
<td>Waveguide crossing loss</td>
<td>0.3~0.8 dB</td>
</tr>
<tr>
<td>Waveguide propagation loss</td>
<td>0.8~1.3 dB/cm</td>
</tr>
<tr>
<td>Photodetector sensitivity</td>
<td>-15 ~ -20 dBm</td>
</tr>
</tbody>
</table>
• Goes down for all three designs
• Better improvement for Trad. IB
  • Still higher than RSON
  • Beyond the power budget on exascale
Outline

• Introduction
• Rack-scale inter/intra-chip optical network (RSON)
• Evaluation and analysis
• Conclusion
Conclusion

• Propose the inter/intra-chip rack-scale optical network
  • The architectural design of inter-chip and intra-chip optical network
• Efficient channel control for optical switch and ONoC
• Systematic evaluation on RSON
  • Promising solution for energy efficient rack-scale high performance computing system
Reference

Q&A