DATE 18
DESIGN, AUTOMATION & TEST IN EUROPE
19 – 23 March, 2018 · ICC · Dresden · Germany
The European Event for Electronic System Design & Test

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

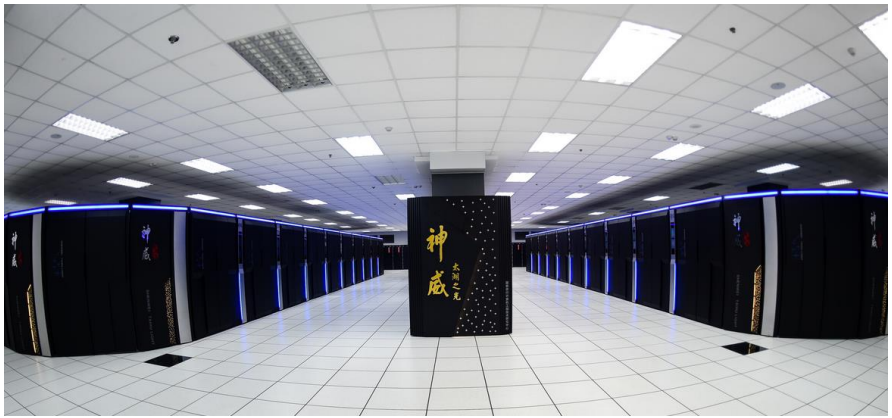
香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Outline

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

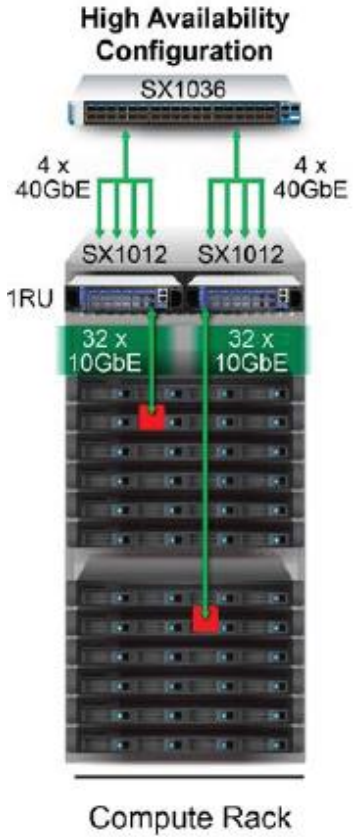
Large scale high performance computing systems

- 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



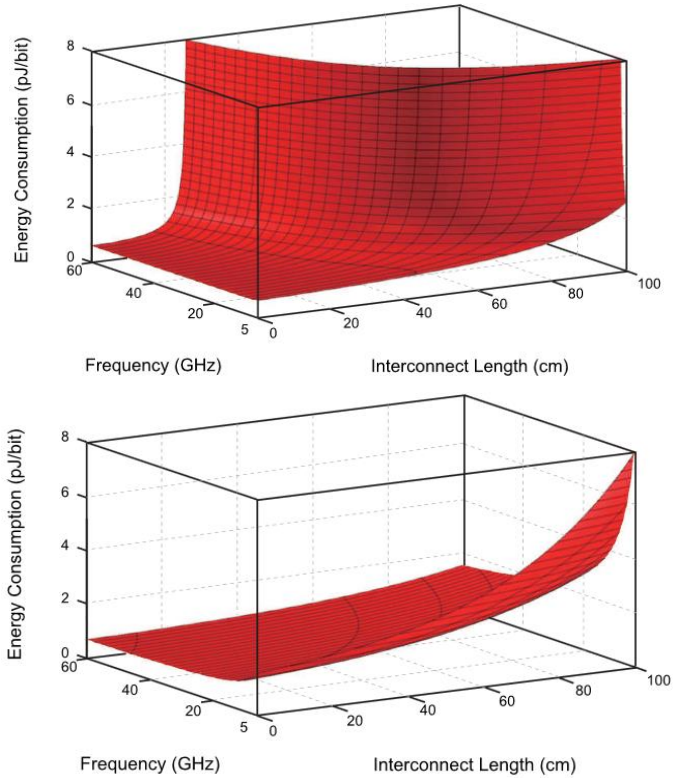
Rack-scale computing system

- **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**



[Z. Wang et al, 2015

Related works on optical interconnected rack

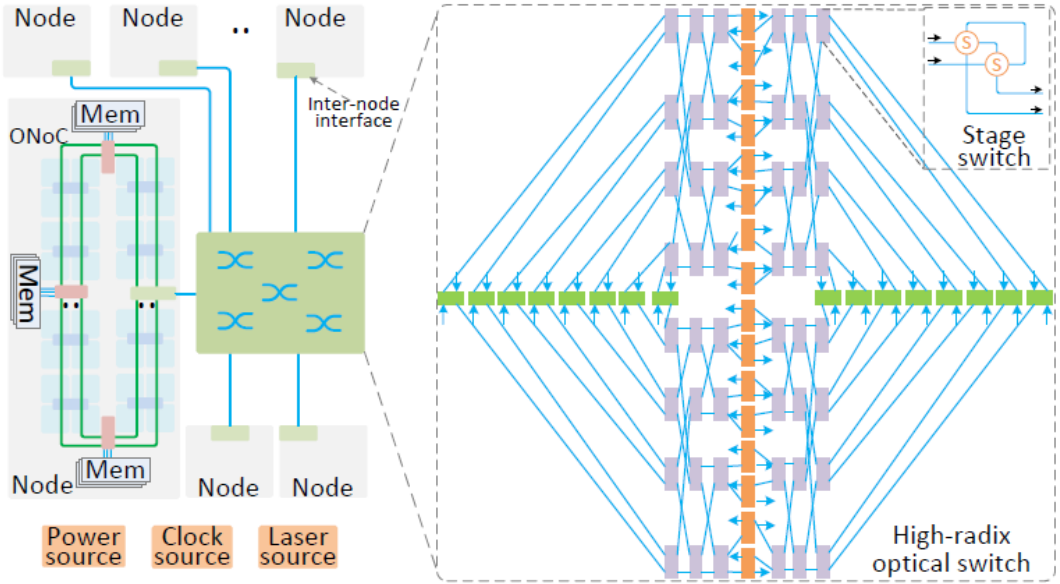
- **Inter/intra-chip optical network:**
  - **Hybrid E/O NoC** [N. Kirman et al 2006; Y. Pan et al, 2010; Y. Ye et al, 2012; X. Wu et al, 2014]
  - **All-optical NoC** [S. Koohi et al, 2012;Y. Ye et al, 2013; S. Bartolini et al, 2013; M. Browning et al, 2013]
  - **Inter and intra-chip optical network**[X. Wu et al, 2013; X. Wu et al, 2015]
- **Rack-scale optical network**[J. Kim et al, 2009; S. Liu et al, 2014; J. Chen et al, 2015, V. Shrivastav et al, 2017]
  - **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**

Outline

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

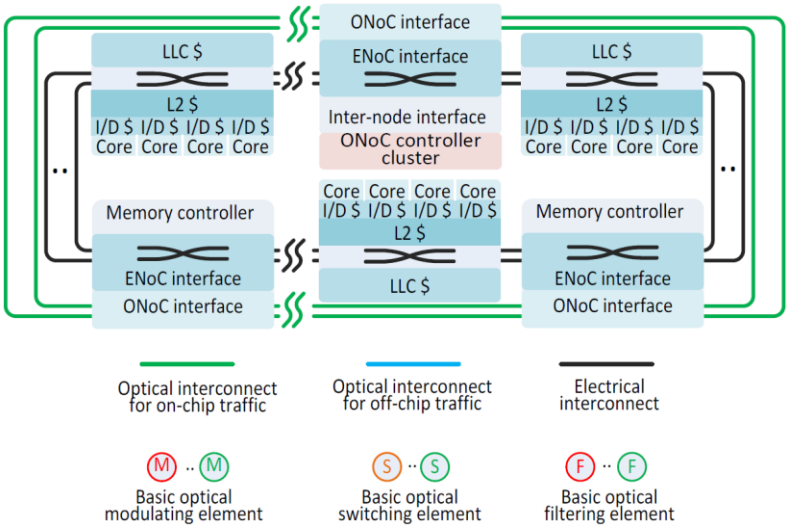
Rack-scale optical network architecture overview

- **High-radix integrated optical switch fabric**
- **Unified inter/intra-chip optical interconnect**
- **Hybrid electrical/optical NoC**
- **Multi-domain circuit switching**



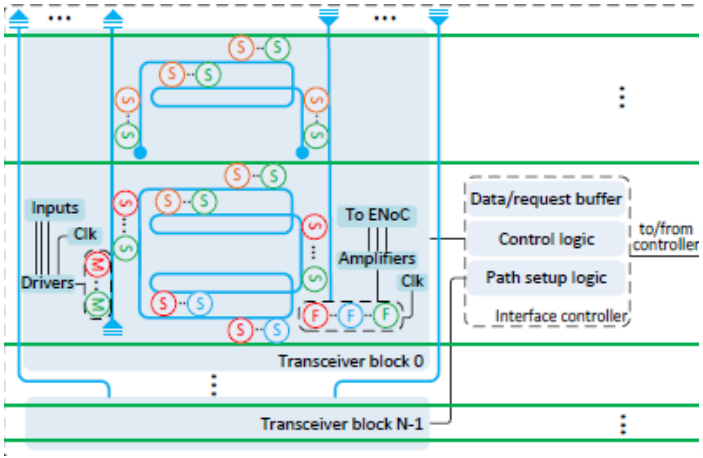
Hybrid E/O server node network architecture

- **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**



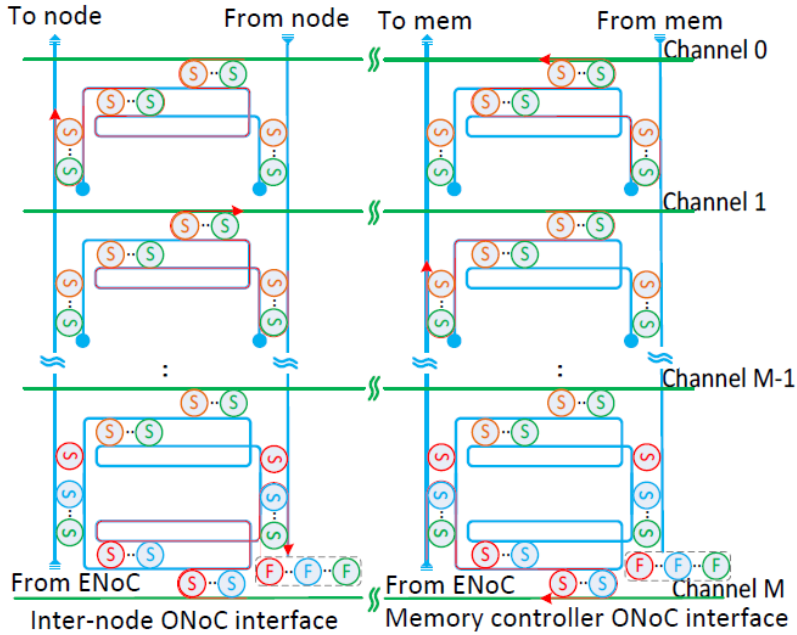
Optical inter-node interface

- 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



Optical transceiver block

- **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**

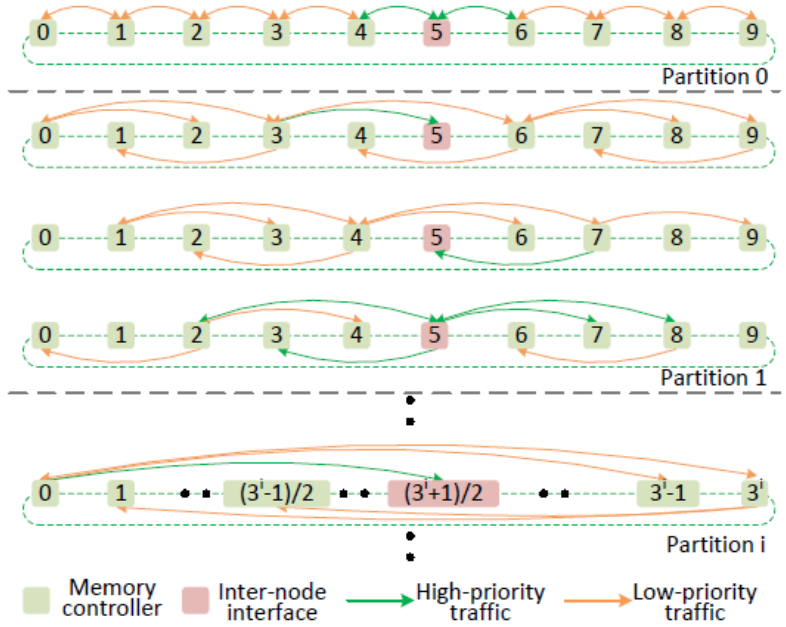


Optical switch control

- **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**

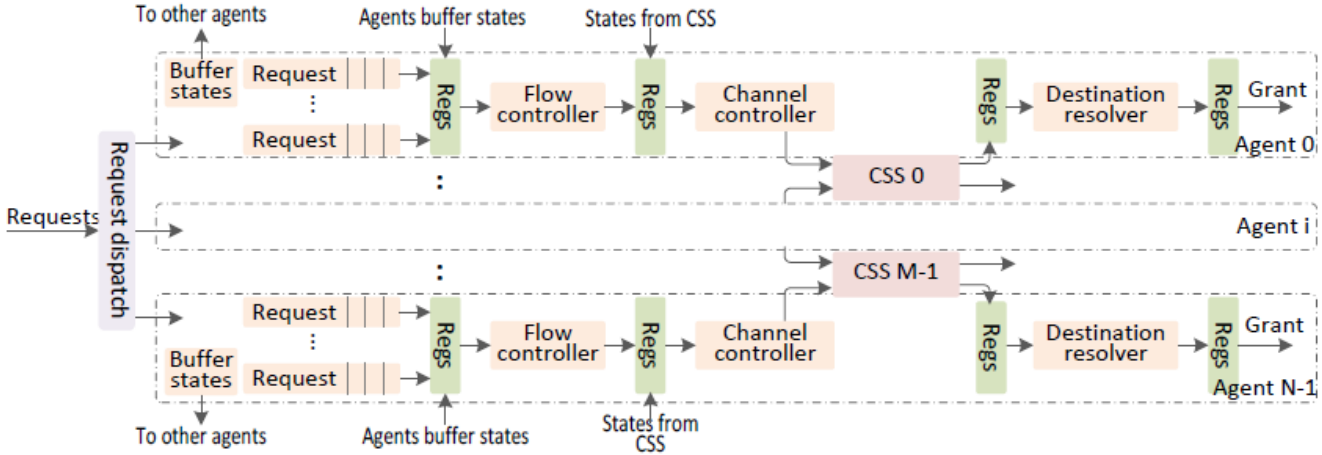
ONoC channel partition

- **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

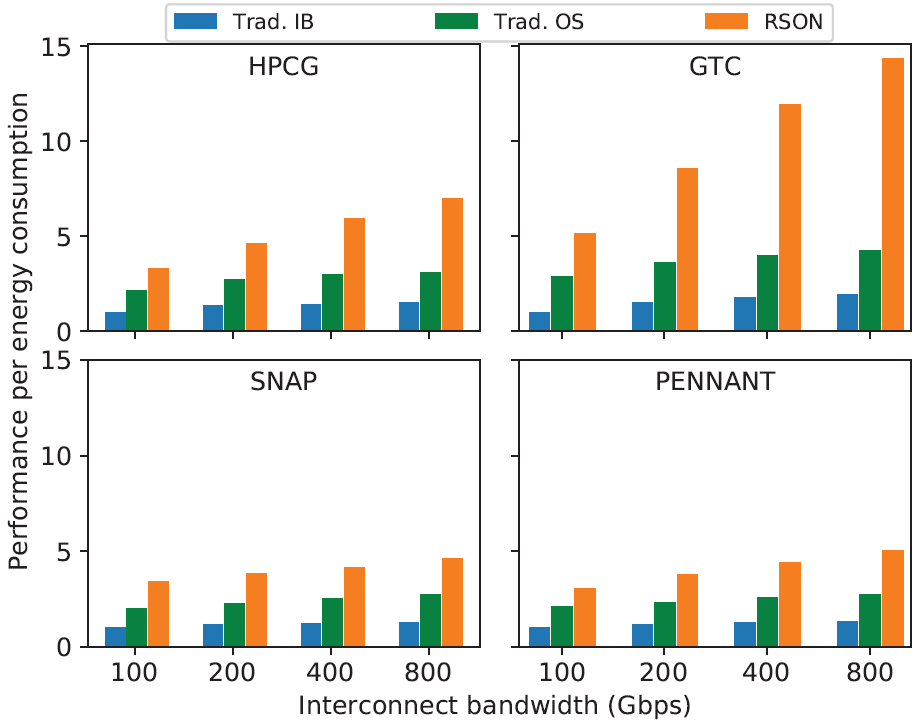
Simulation environment and setup

- 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

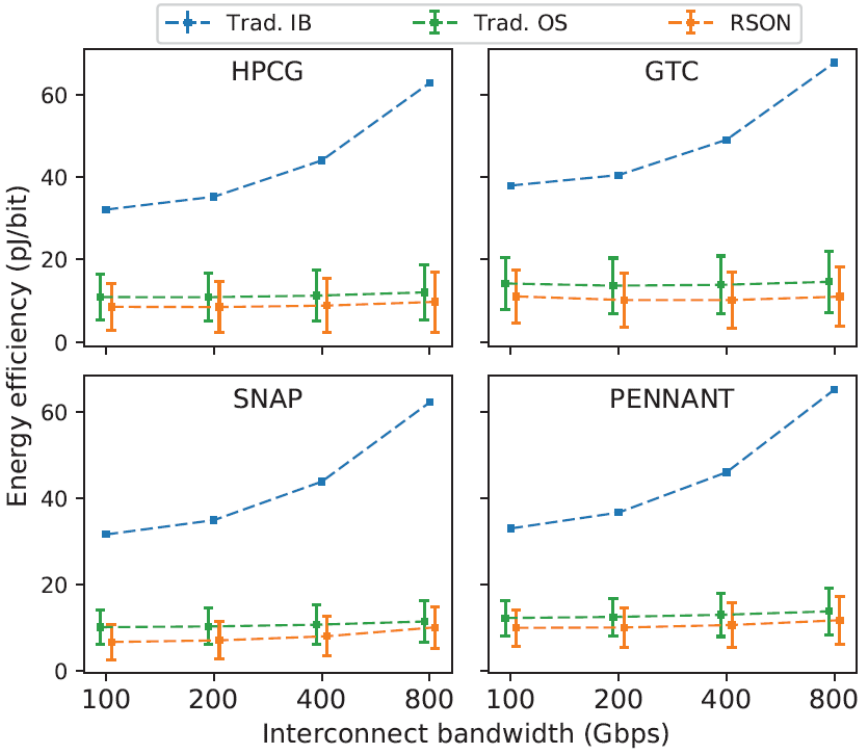
Item	Configuration
Core cluster	8, 4 cores/cluster
L1 I/D $	64KB/core, private
L2 $	512KB/cluster, shared
LLC	2.5MB/slice
Coherence protocol	Directory based MOSI
Electrical link	128-bit width/direction
Memory bandwidth	480GB/s/port
Memory capacity	16GB/port

Performance per energy consumption

- **RSON achieves most performance, 6.8X at best**
- **High bandwidth interconnect benefits RSON more**



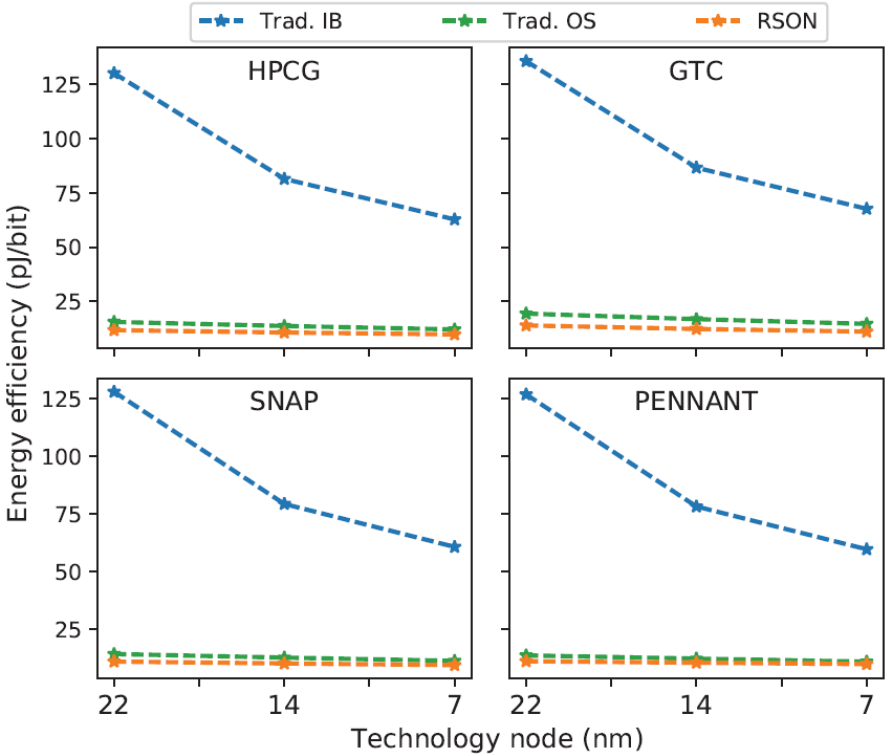
Interconnection energy efficiency



Item	Value
Laser efficiency	0.33 [48, 49]
MR insertion loss	1∼1.8 dB
MR passing loss	0.06∼0.3 dB
Edge coupling loss	1∼1.5 dB
Waveguide crossing loss	0.3∼0.8 dB
Waveguide propagation loss	0.8∼1.3 dB/cm
Photodetector sensitivity	-15 ∼ -20 dBm

Energy efficiency vs. technology node

- 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

[1] P. Kapur and K. Saraswat, “Comparisons between electrical and optical interconnects for on-chip signaling,” in Interconnect Technology Conference, 2002.roceedings of the IEEE 2002 International, 2002, pp. 89 – 91.
[2] Y. Pan, P. Kumar, J. Kim, G. Memik, Y. Zhang, and A. Choudhary, “Firefly: illuminating future network-on-chip with nanophotonics,” In Proc. of the International Symposium on Computer Architecture, vol. 37, no. 3, pp. 429–440, 2009.
[3] M. J. Cianchetti, J. C. Kerekes, and D. H. Albonesi, “Phastlane: a rapid transit optical routing network,” in Proceedings of the 36th annual international symposium on Computer architecture. New York, USA: ACM, 2009, pp. 441–450.
[4] Y. Pan, J. Kim, and G. Memik, “FlexiShare: Channel sharing for an energy-efficient nanophotonic crossbar,” in High Performance Computer Architecture, IEEE 16th International Symposium on, 2010, pp. 1–12.
[5] N. Kirman, M. Kirman, R. K. Dokania, J. F. Martinez, A. B. Apsel, M. A. Watkins, and D. H. Albonesi, “Leveraging Optical Technology in Future Busbased Chip Multiprocessors,” in Proceedings of the 39th Annual IEEE/ACM International
Symposium on Microarchitecture. Washington, DC, USA: IEEE Computer Society, 2006, pp. 492–503.
[6] S. Pasricha and N. Dutt, “ORB: An on-chip optical ring bus communication architecture for multi-processor systems-on-chip,” in Design Automation Conference, Asia and South Pacific, 2008, pp. 789–794.
[7] S. Bahirat and S. Pasricha, “UC-PHOTON: A novel hybrid photonic network-on-chip for multiple use-case applications,” in Quality Electronic Design (ISQED), 2010 11th International Symposium on, march 2010, pp. 721 –729.
[8] X. Wu, J. Xu, Y. Ye, Z. Wang, M. Nikdast, and X. Wang, “Suor: Sectioned undirectional optical ring for chip multiprocessor,” ACM Journal on Emerging Technologies in Computing Systems (JETC), vol. 10, no. 4, p. 29, 2014.
[9] S. Le Beux, J. Trajkovic, I. O’Connor, G. Nicolescu, G. Bois, and P. Paulin, “Optical Ring Network-on-Chip (ORNoC): Architecture and design methodology,” in Design, Automation Test in Europe Conference Exhibition, 2011, pp. 1–6.
[10] S. Koohi and S. Hessabi, “All-Optical Wavelength-Routed Architecture for a Power-Efficient Network on Chip,” Computers, IEEE Transactions on, vol. PP, no. 99, pp. 1–1, 2012.
[11] S. Bartolini, L. Lusnig, and E. Martinelli, “Olympic: A hierarchical all-optical photonic network for low-power chip multiprocessors,” in Digital System Design (DSD), 2013 Euromicro Conference on, Sept 2013, pp. 56–59.
[12] Z. Wang, H. Gu, Y. Yang, and Y. Li, “Ring based Optical Network-on-Chip,” Optics Communications, vol. 285, no. 6, pp. 1010 – 1016, 2012.
[13] M. Browning, C. Li, P. Gratz, and S. Palermo, “Luminoc: A low-latency, high-bandwidth per watt, photonic network-on-chip,” in System Level Interconnect Prediction (SLIP), 2013 ACM/IEEE International Workshop on, June 2013.
[14] H. Li, H. Gu, and Y. Yang, “A hierarchical cluster-based optical network-onchip,” in Future Computer and Communication (ICFCC), 2010.
[15] J. Kim, W. Dally, S. Scott, and D. Abts, “Cost-efficient dragonfly topology for large-scale systems,” in Optical Fiber Communication, 2009. OFC 2009. Conference on, March 2009.
[16] D. Vantrease, R. Schreiber, M. Monchiero, M. McLaren, N. P. Jouppi, M. Fiorentino, A. Davis, N. Binkert, R. G. Beausoleil, and J. H. Ahn, “Corona: System Implications of Emerging Nanophotonic Technology,” in Computer Architecture, 35th International Symposium on, 2008, pp. 153–164.
[17] P. Costa, H. Ballani, K. Razavi, and I. Kash, “R2c2: A network stack for rack-scale computers,” in ACM SIGCOMM Computer Communication Review, 2015.
[18] T. Benson, A. Akella, and D. A. Maltz, “Network traffic characteristics of data centers in the wild,” in ACM SIGCOMM conference on Internet measurement. ACM, 2010.


Q&A