NSF CNS #1318981: Power-Efficient 3D Reconfigurable Photonic Interconnect for Multicores
In collaboration with George Washington University (GWU) with PI, Ahmed Louri

SHARP: Shared Heterogeneous Architecture with Reconfigurable Photonic Network-on-Chip

Scott VanWinkle and Avinash Kodi

Technologies for Emerging Computer Architecture Laboratory (TEAL)
School of Electrical Engineering and Computer Science
Ohio University, Athens OH, USA

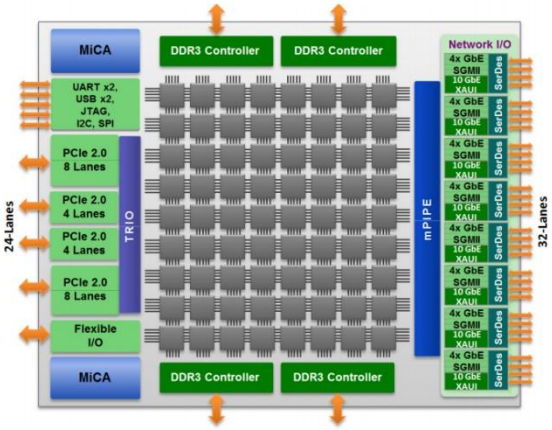
CSR PI Meeting, Orlando, Florida, June 2, 2017

Contact Website: http://oucsace.cs.ohiou.edu/~avinashk/

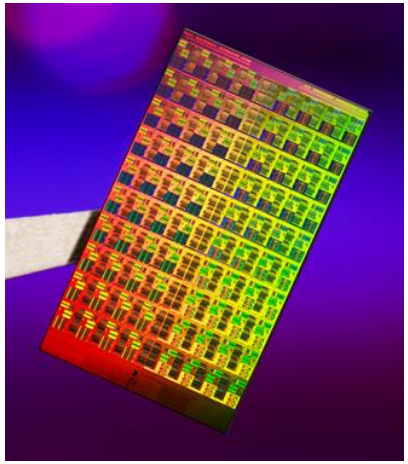
Outline

- Introduction & Motivation
- SHARP
  - Architecture & Implementation
  - Dynamic Bandwidth & Power Scaling
  - Machine Learning
- Performance Analysis
- Other Research Accomplishments

Multicores & Networks-on-Chip



TILE-Gx72[1]



80-core Intel TeraFlops[2]

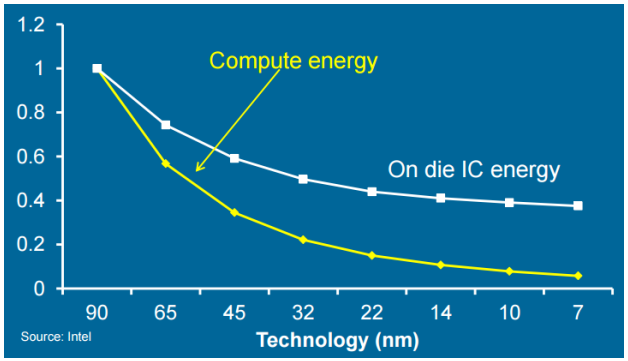
2880-core KEPLER (Nvidia)[3]

- With increasing number of heterogeneous cores, communication-centric design paradigm is becoming critical (Networks-on-Chip)
  - Energy consumed for communication is increasing
  - Traffic patterns exhibit temporal and spatial fluctuations

[1] http://www.tilera.com/products/processors/TILE-Gx_Family [2] http://www.intel.com/pressroom/kits/teraflops/ [3] http://www.nvidia.com/object/nvidia-kepler.html

Energy & Bandwidth Limitations

- Interconnect energy is scaling slower than compute energy as technology and number of cores continue to scale



Source: S. Borkar, Exascale Computing- a fact or fiction?, Intel, 2013
http://www.ipdps.org/ipdps2013/SBorkar_IPDPS_May_2013.pdf

- Reduced bandwidth/throughput due to voltage/frequency scaling

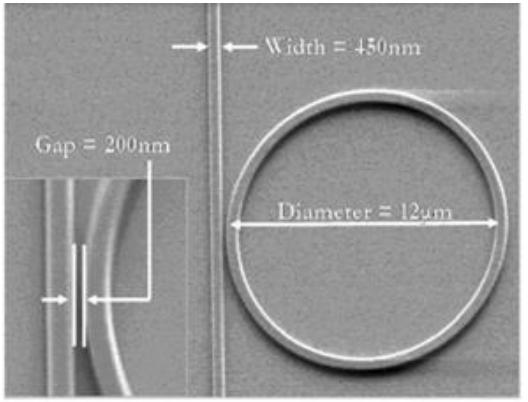
➢ Potential Solutions:
Photonics, 3D Stacking, Wireless

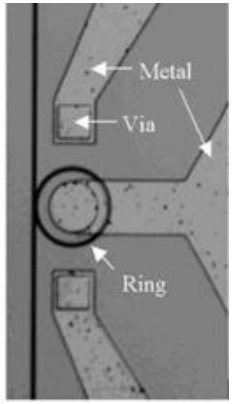
- Traffic patterns exhibit temporal and spatial fluctuations

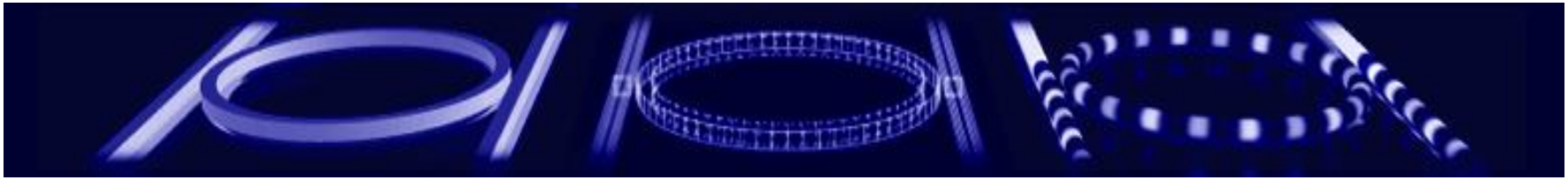
➢ Potential Solution:
Dynamic Bandwidth Reconfiguration

Why Photonics?

- Photonics provides
  - ✓ Low energy (7.9 fJ/bit)
  - ✓ Small footprint (~2.5 μm)
  - ✓ High bandwidth (~40 Gbps)
  - ✓ Low latency (10.45 ps/mm)
  - ✓ CMOS compatible

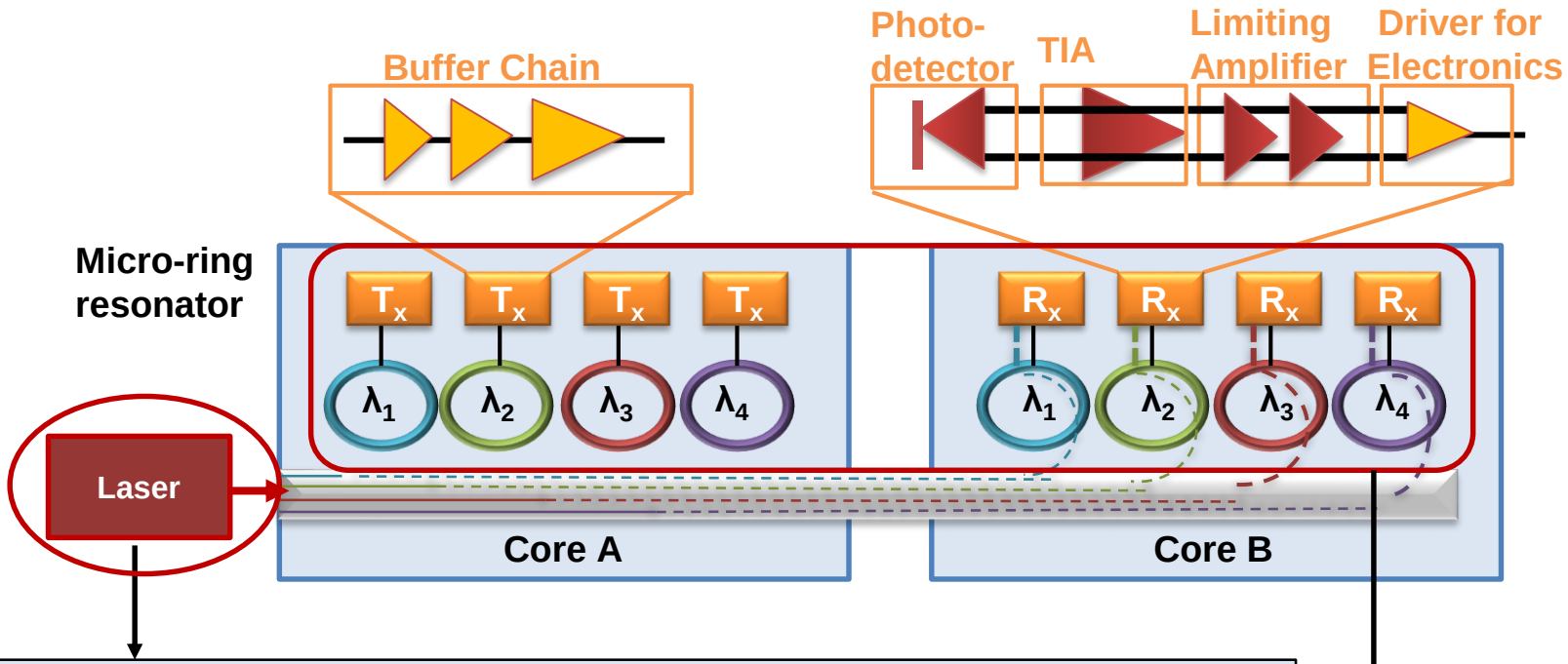






1. L. Xu, W. Zhang, Q. Li, J. Chan, H. L. R. Lira, M. Lipson, K. Bergman, “40-Gb/s DPSK Data Transmission Through a Silicon Microring Switch," *IEEE Photonics Technology Letters* 24.
2. S. Manipatruni, K. Preston, L. Chen, and M. Lipson, “Ultra-low voltage, ultra-small mode volume silicon microring modulator,” Opt. Express 18, 18235-18242 (2010)

Photonic Link



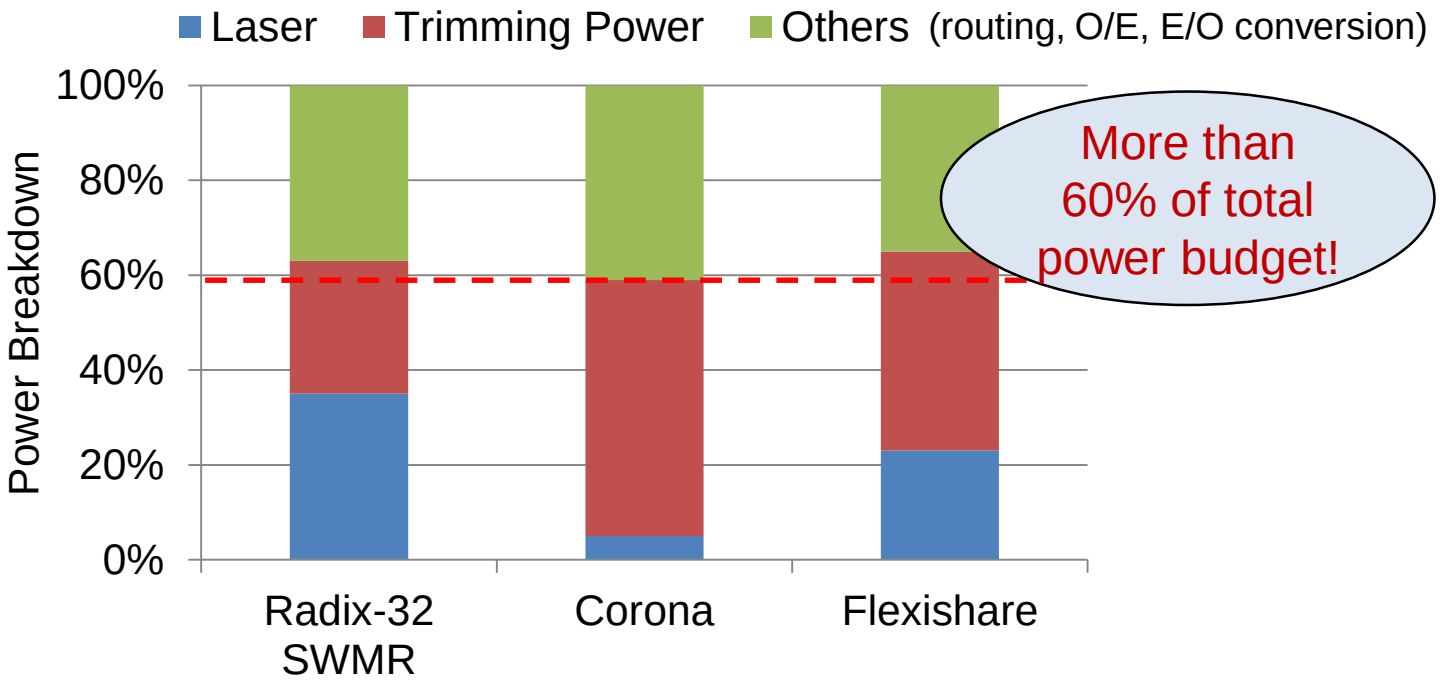
Laser power

- Compensates for a variety of light losses along its path

Trimming power

- Microring resonators are sensitive to temperature variations.
- They require additional trimming power to maintain their resonant wavelength

Static Power Challenge



- The laser source and on-chip micro-ring resonators trimming power represent the majority of network power

➢ Potential Solution:
Dynamic Power Scaling

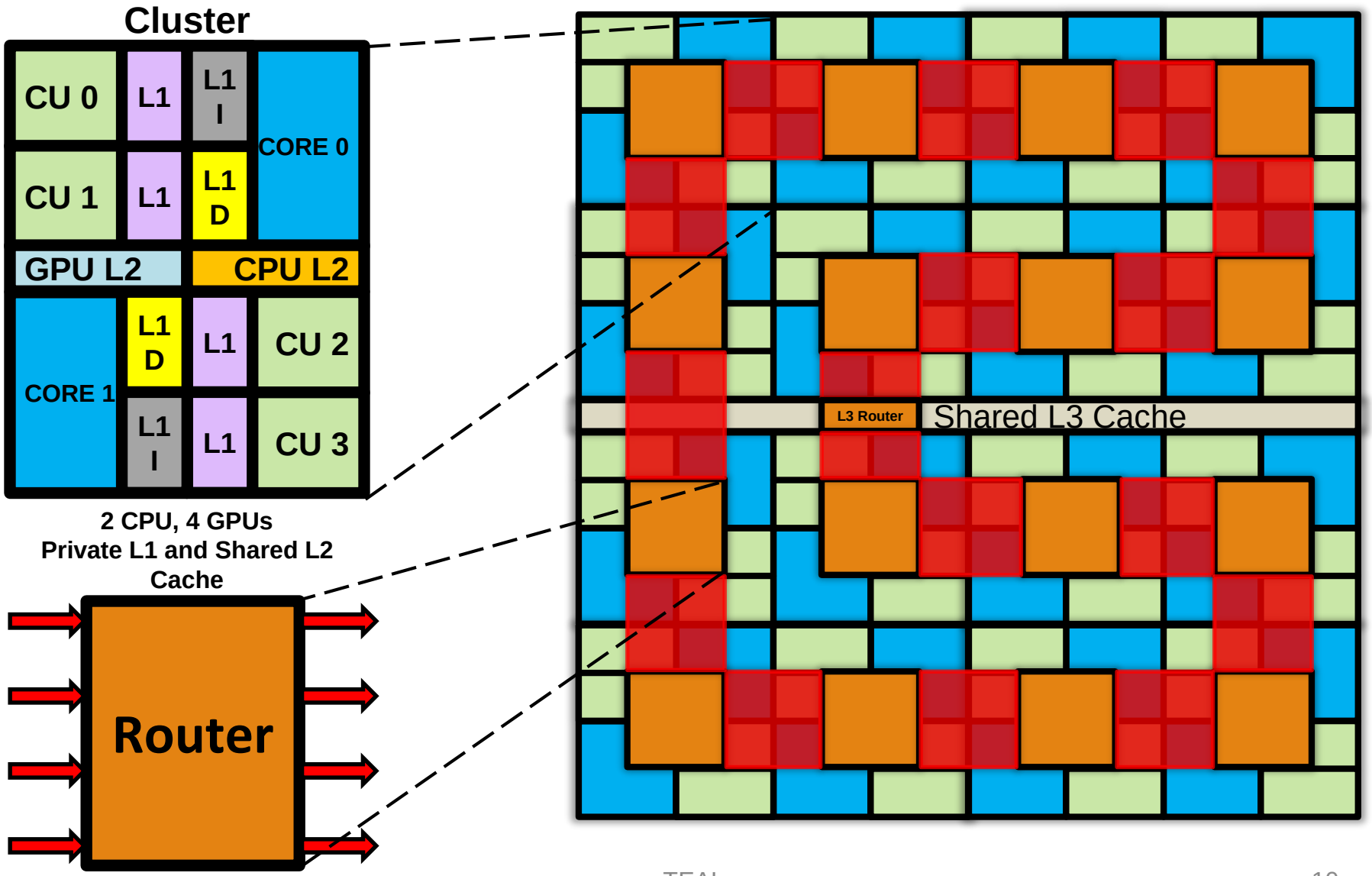
Outline

- Introduction & Motivation
- SHARP
  - Architecture & Implementation
  - Dynamic Bandwidth & Power Scaling
  - Machine Learning
- Performance Analysis
- Other Research Accomplishments

SHARP: Shared Heterogeneous Architecture for Reconfigurable Photonic Network-on-Chip

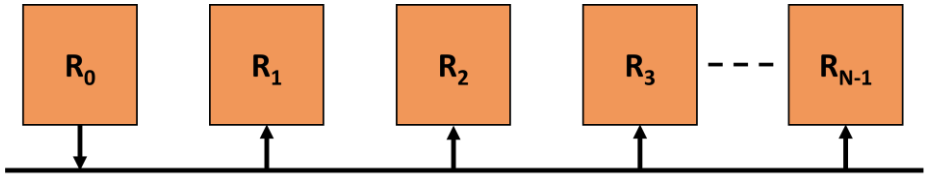
- Key goal
  - Provide scalable bandwidth and save static optical power while meeting performance constraints
- Hardware Design
  - Ring-based photonic crossbar that combines both CPU and GPU cores together into a cluster
  - Propose R-SWMR to reduce power consumption
- Novelty
  - Fine-grain dynamic bandwidth allocation without global coordination
  - Refine the bandwidth and power scaling using machine learning algorithms
- Main result
  - Static power savings more than 45% - 65%, with 0.3% -14% penalty on throughput

SHARP Architecture: Checkerboard Pattern



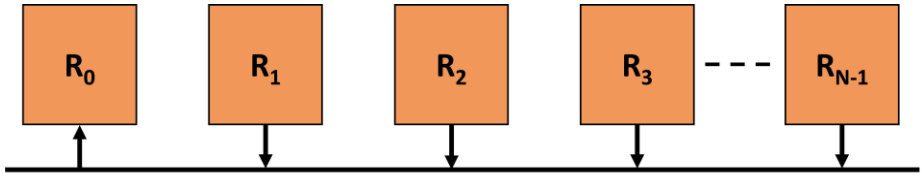
Inter-Router Communication

Single Write Multiple Reader (SWMR)

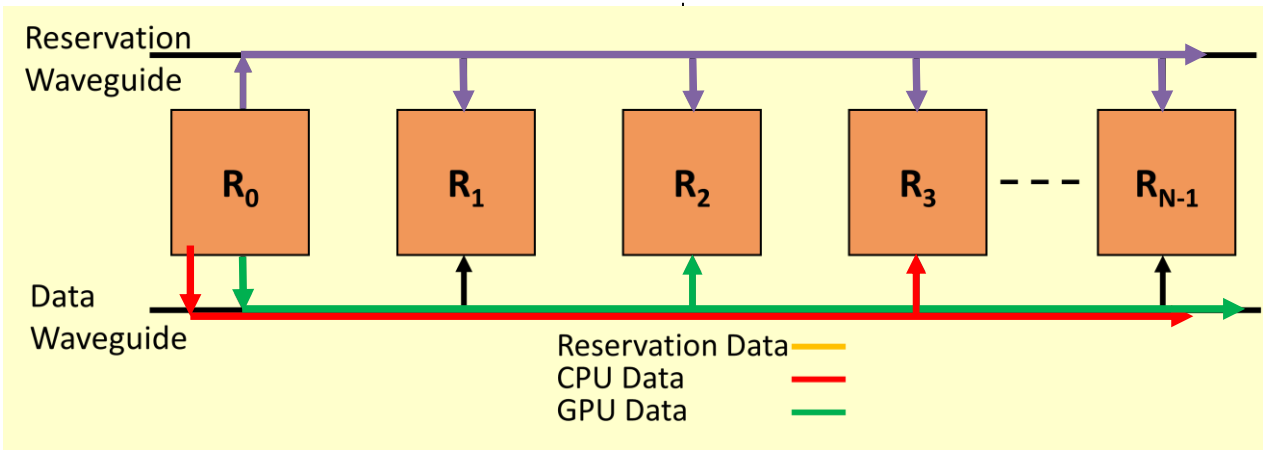


- Dedicated crossbar channel – local control & coordination
- Broadcast optical signal – consumes more power

Multiple Write Single Reader (MWSR)

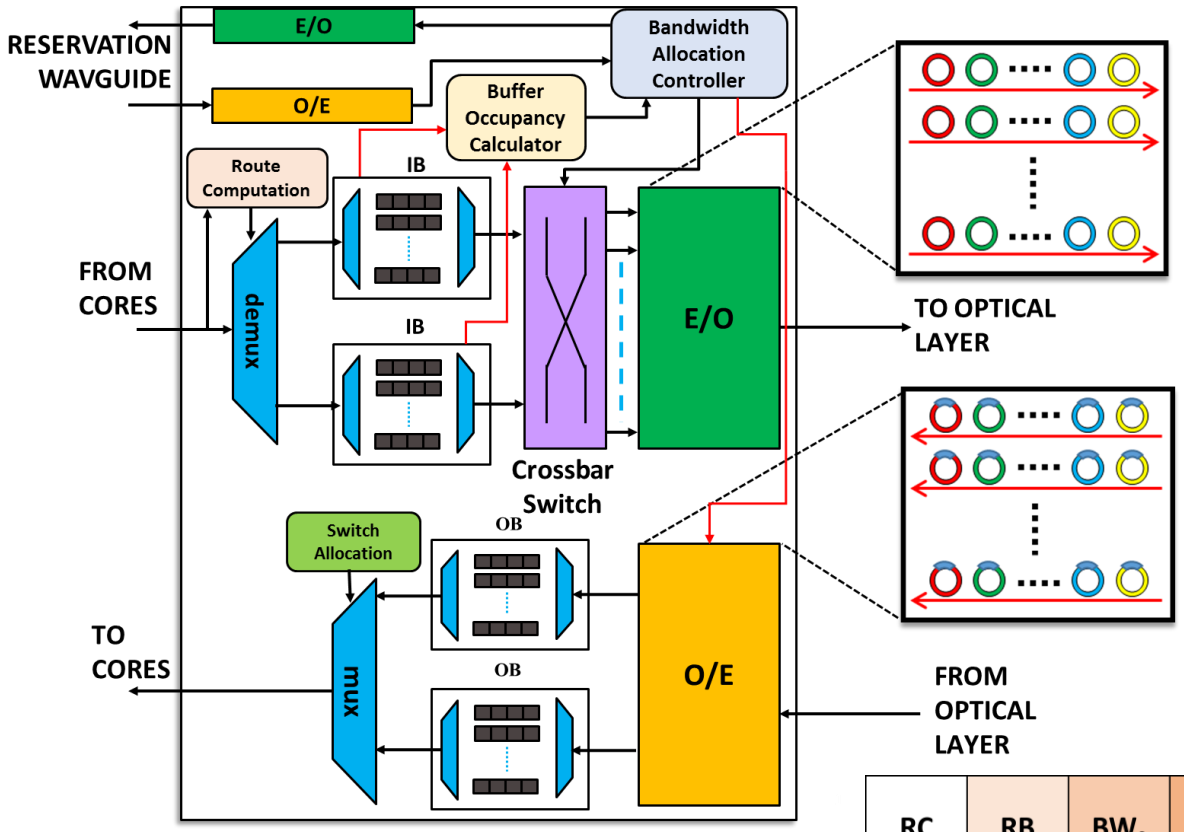


- Shared crossbar channel – global control & coordination
- Point-to-point optical signal – consumes less power



- Y. Pan, P. Kumar, J. Kim, G. Memik, Y. Zhang and A. Choudhary, "Firefly: Illuminating Future Network-on-Chip with Nanophotonics", Proc. International Symposium on Computer Architecture, 2009, pp. 429- 440

Router Microarchitecture



Router Pipeline

RC	RB	BW_S	E/O	OL	O/E	BW_D	SA	
	RC	RB	BW_S	E/O	OL	O/E	BW_D	SA
		RC	RB	BW_S	E/O	OL	O/E	BW_D	SA

Dynamic Bandwidth Scaling

- Predict buffer occupancy

$$\beta_{Ocup(CPU)} = \frac{\sum_{i=0}^{j-1} Buf_i \times a_i}{j}$$

$$\beta_{Ocup(GPU)} = \frac{\sum_{i=0}^{k-1} Buf_i \times a_i}{k}$$

- Re-allocate bandwidth between CPU and GPU cores
- Prioritize CPU requests over GPU requests
- Make decisions locally using R-SWMR

Step 0: For each individual routers R_0 through R_{N-1} complete steps 1 through 7

Step 1: Calculate the buffer occupancy β_{Ocup} for each input buffer $\beta_{Ocup\text{-}0}$ through $\beta_{Ocup\text{-}(j-1)}$ in router R_ω

Step 2: β_{Ocup} for routers $\beta_{Ocup\text{-}0}$ through $\beta_{Ocup\text{-}(j-1)}$ are sent to Buffer Occupancy Calculator

Step 3: Calculate β_{CPU} using $\beta_{Ocup\text{-}0}$ through $\beta_{Ocup\text{-}(k-1)}$

Step 4: Calculate β_{GPU} using $\beta_{Ocup\text{-}k}$ through $\beta_{Ocup\text{-}(j-1)}$

Step 5: Determine the amount of bandwidth to be allocated to the CPU and GPU core types:

- If $\beta_{GPU} = 0$ and $\beta_{CPU} > 0$
  - $GPU_{Bandwidth} = 0\%\ Bandwidth$
  - $CPU_{Bandwidth} = 100\%\ Bandwidth$
- Else if $\beta_{CPU} = 0$ and $\beta_{GPU} > 0$
  - $GPU_{Bandwidth} = 100\%\ Bandwidth$
  - $CPU_{Bandwidth} = 0\%\ Bandwidth$
- Else if $\beta_{GPU} < \beta_{GPU\text{-}UpperBound}$
  - $GPU_{Bandwidth} = 25\%\ Bandwidth$
  - $CPU_{Bandwidth} = 75\%\ Bandwidth$
- Else if $\beta_{CPU} < \beta_{CPU\text{-}UpperBound}$
  - $GPU_{Bandwidth} = 75\%\ Bandwidth$
  - $CPU_{Bandwidth} = 25\%\ Bandwidth$
- Else
  - $GPU_{Bandwidth} = 50\%\ Bandwidth$
  - $CPU_{Bandwidth} = 50\%\ Bandwidth$

Step 6: Send reservation packet via reservation-assisted SWMR link

Step 7: Transmit Data Using specified wavelengths on the first come first serve basis

Dynamic Power Scaling

- Power scaling by turning off select wavelengths (from 64-48-32-16-8)
- On-chip laser with a 2ns turn-on delay has been proposed [1,2,3]
- Reservation windows set to 500 and 2000 cycles

Step 0: For each individual routers R_0 through R_{N-1}
complete steps 1 through 9
Step 1: Calculate the buffer occupancy β_{Ocup} for each input buffer β_{Ocup-0} through $\beta_{Ocup-(j-1)}$ in router R_ω
Step 2: β_{Ocup} for routers β_{Ocup-0} through $\beta_{Ocup-(j-1)}$ are sent to Buffer Occupancy Calculator
Step 3: Calculate β_{CPU} using β_{Ocup-0} through $\beta_{Ocup-(k-1)}$
Step 4: Calculate β_{GPU} using β_{Ocup-k} through $\beta_{Ocup-(j-1)}$

⋮

Step 7: Transmit Data Using specified wavelengths on the first come first serve basis
Step 8: For each reservation window RW, sum the total buffer occupancy β_{Total} for each cycle
Step 9: At the end of RW, determine the number of wavelengths WL for the outgoing waveguide at Router R_ω:
If $\beta_{Total} > Threshold_{upper}$
WL= 64 *Wavelengths*
Else If $\beta_{Total} > Threshold_{mid\text{-}upper}$
WL= 48 *Wavelengths*
Else If $\beta_{Total} > Threshold_{mid\text{-}lower}$
WL= 32 *Wavelengths*
Else If $\beta_{Total} > Threshold_{lower}$
WL= 16 *Wavelengths*
Else
WL= 8 *Wavelengths*

[1] K. P. E. Kotelnikov, A. Katsnelson and I. Kudryashov, “Highpower single-mode ingaasp/inp laser diodes for pulsed operation,” Proceedings of SPIE, vol. 8277 827715, pp. 1–6, 2012.
[2] M. Heck and J. Bowers, “Energy Efficient and Energy Proportional Optical Interconnects for Multi-core Processors: Driving the Need for On-chip Sources”, IEEE Journal of Selected Topics in Quantum Electronics 20(4)(2014), pp. 1-12.
[3] T. Wang, H. Liu, A. Lee, F. Pozzi and A. Seeds, “1.3-μm InAs/GaAs Quantum-dot Lasers Monolithically Grown on Si Substrates”, Optics Express 19(12)(2011), pp. 11381-11386

Machine Learning for Power Scaling

- Why Machine learning?
  - Machine learning uses a proactive technique instead of reactive

- The machine learning uses linear ridge regression with the following error function:

$$\tilde{E}(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\{\boldsymbol{w}^T\phi(x_n) - t_n\}^2 + \frac{\lambda}{2}\|\boldsymbol{w}\|^2$$

- β_{Total} gets replaced with the number of predicted packets injected into each router for the next reservation window

[1] C. M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics). Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

Outline

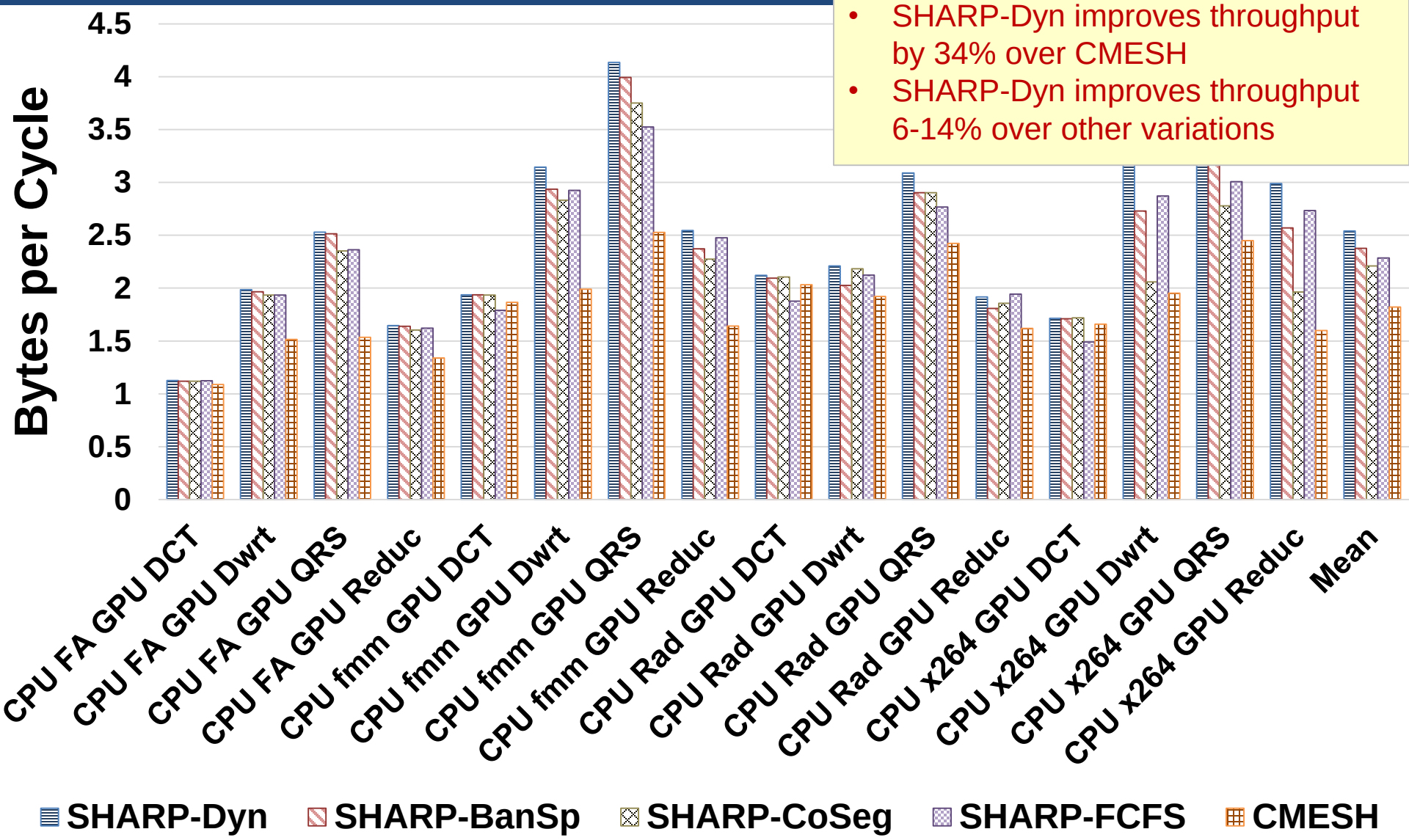
- Introduction & Motivation
- SHARP
  - Architecture & Implementation
  - Dynamic Bandwidth & Power Scaling
  - Machine Learning
- Performance Analysis
- Other Research Accomplishments

Simulation Methodology

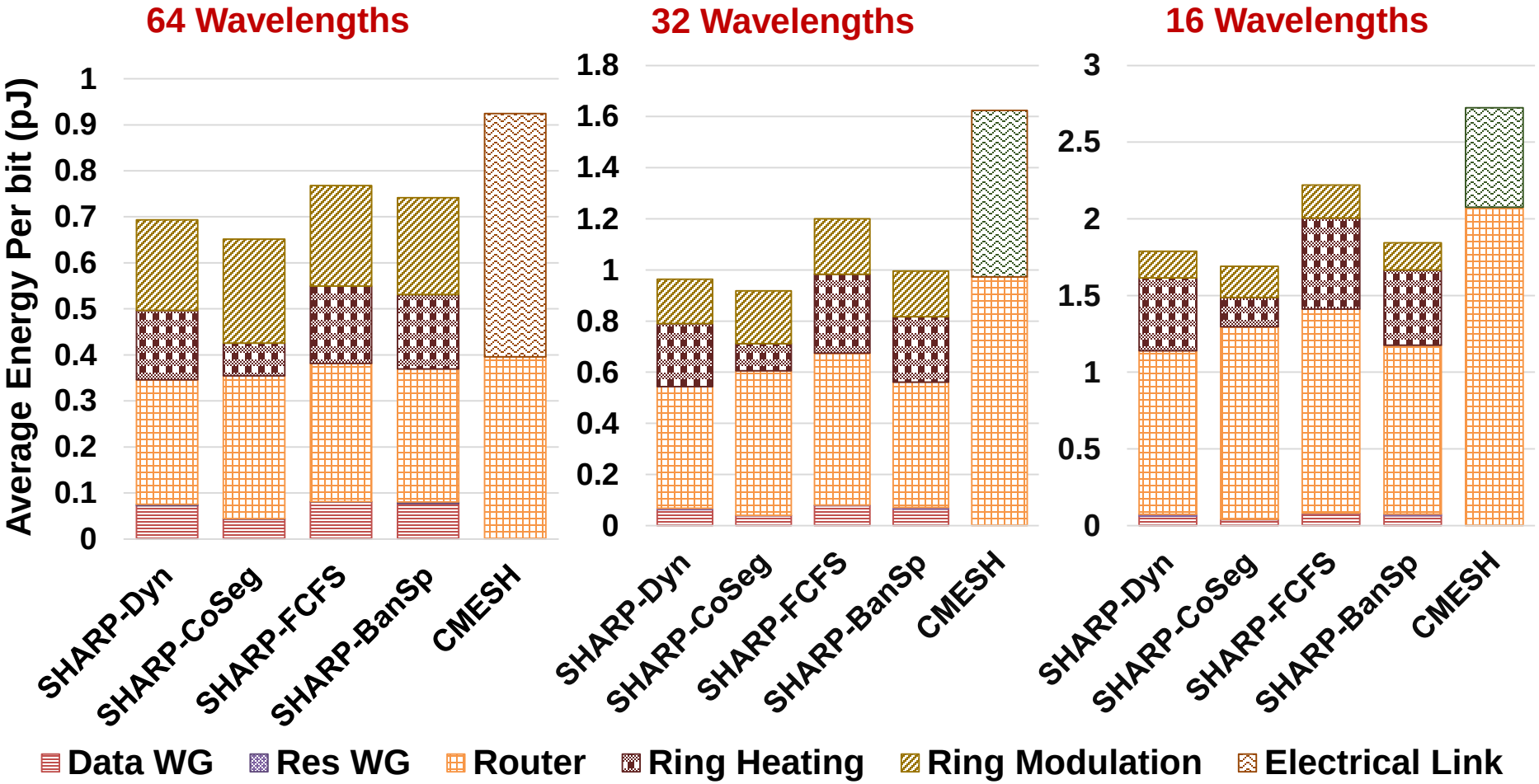
- Architecture Specifications
  - 32-CPU (32KB L1 Instruction, 64KB L1 Data, 256KB L2 Cache)
  - 64-GPU (64KB L1 Instruction, 512KB L2 Cache)
  - Traces collected using Multi2Sim on CPU Benchmarks (PARSEC 2.1 and Splash-2) and GPU Benchmarks (OpenCL SDK)
- Networks-on-Chip Model
  - Cycle-accurate simulator based on Netsim + Dsent for power analysis
  - Compared against SHARP-Dyn variations and CMESH
    - FCFS – First Come First Serve
    - CoSeg – Core Segregation
    - BanSp – Bandwidth Split
- Performance Analysis
  - Analyzed throughput, energy/bit
  - Sensitivity to different wavelengths (16, 32, 64)
  - Reconfiguration window sizing (500 and 2000 cycles)

Throughput Achieved (Dynamic Bandwidth Scaling)

- SHARP-Dyn improves throughput by 34% over CMESH
- SHARP-Dyn improves throughput 6-14% over other variations

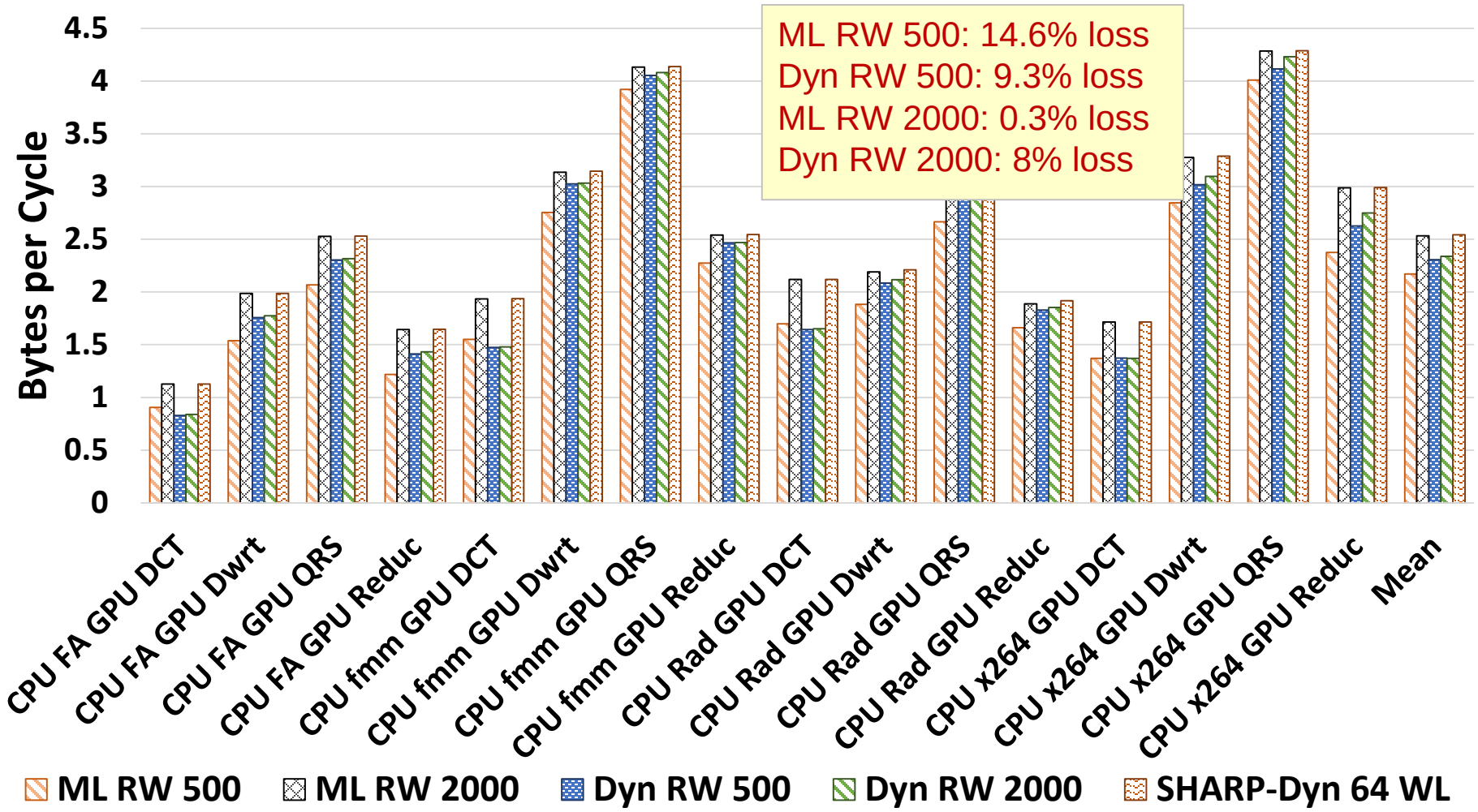


Energy per bit (Dynamic Bandwidth Scaling)

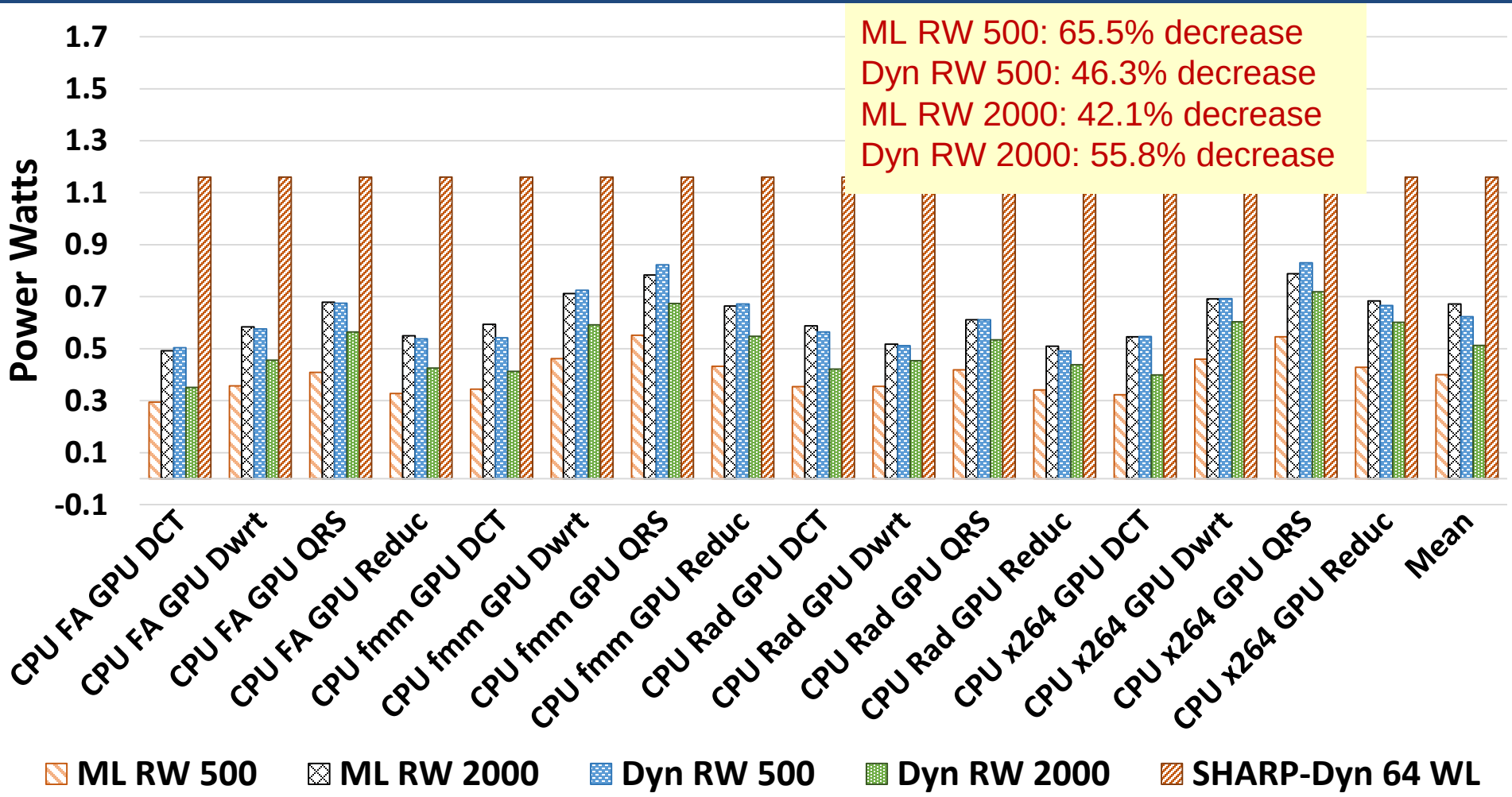


- SHARP-Dyn saves 24% more power than CMESH architecture

Machine Learning - Throughput



Machine Learning – Power Analysis

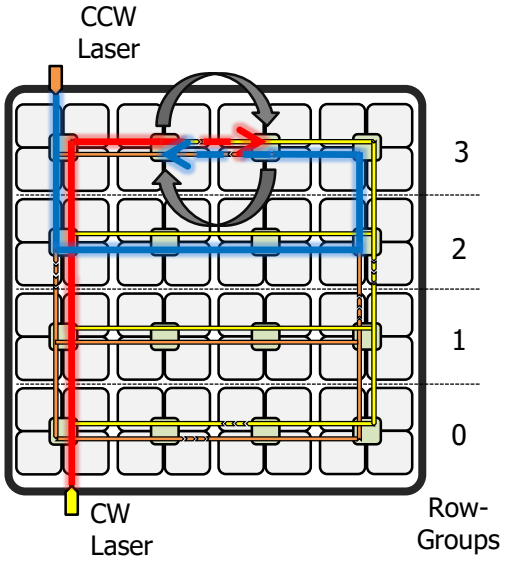


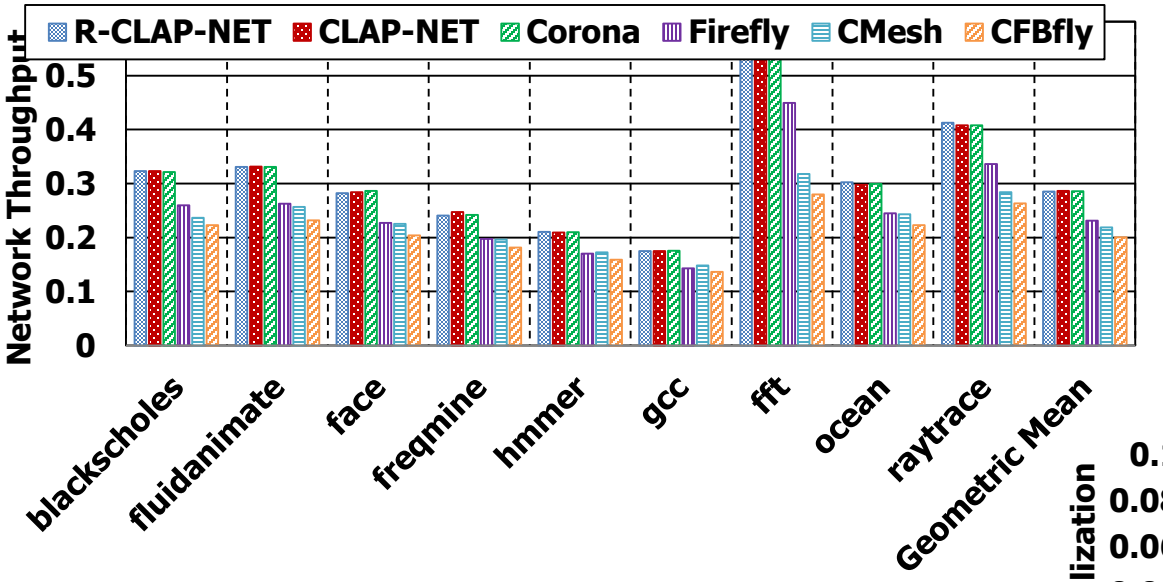
Outline

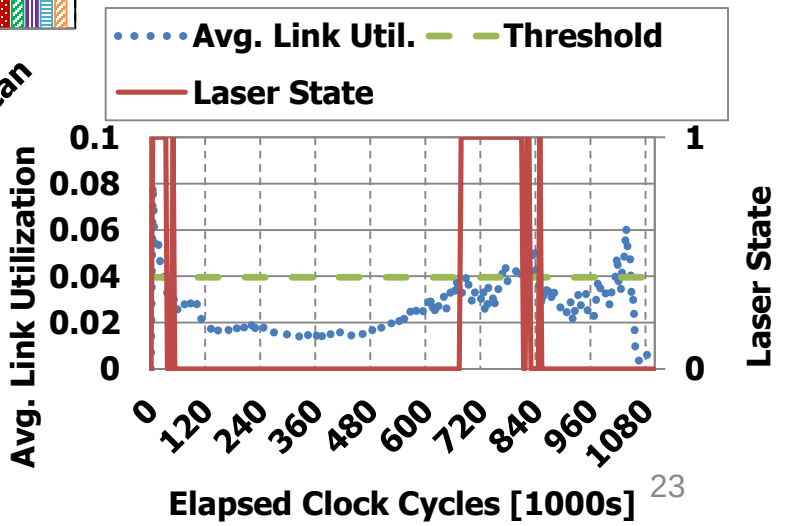
- Introduction & Motivation
- SHARP
  - Architecture & Implementation
  - Dynamic Bandwidth & Power Scaling
  - Machine Learning
- Performance Analysis
- Other Research Accomplishments

CLAPNET – Clockwise Counter-Clockwise Architecture

- Reduces signal propagation distance
- Split crossbar architecture to reduce router complexity (dual lasers)
- Bandwidth reconfiguration and power regulation

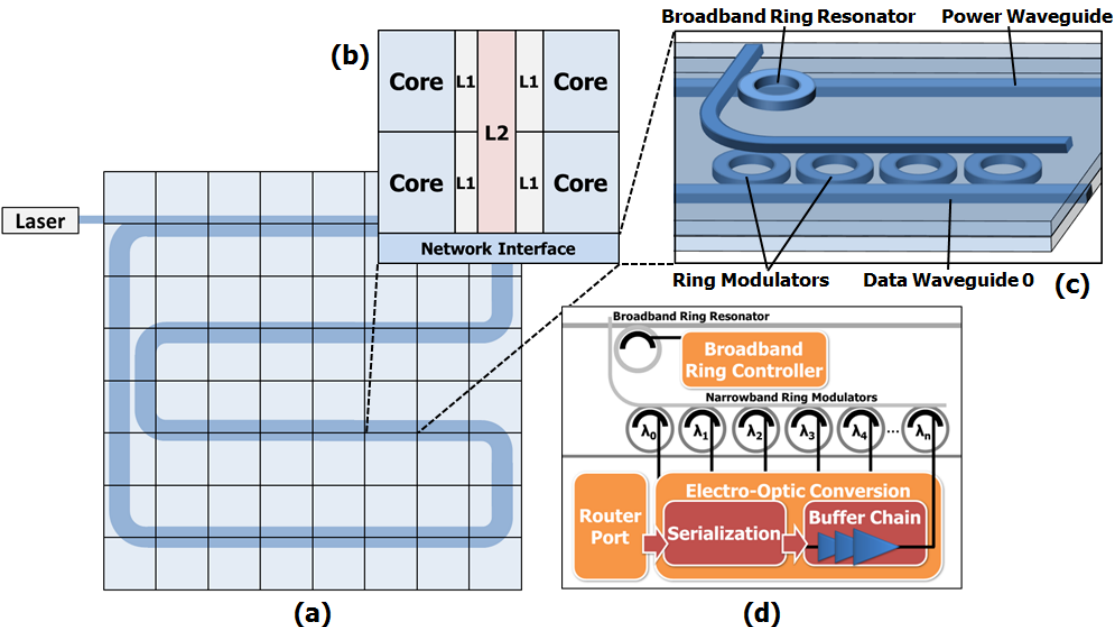




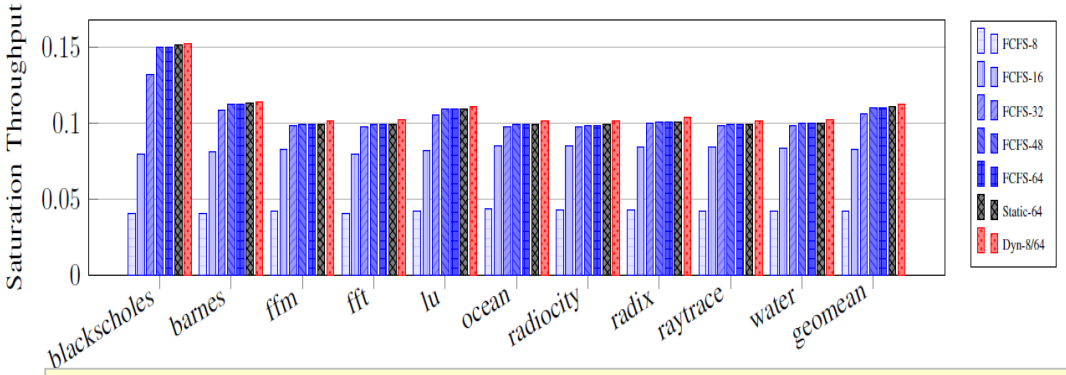


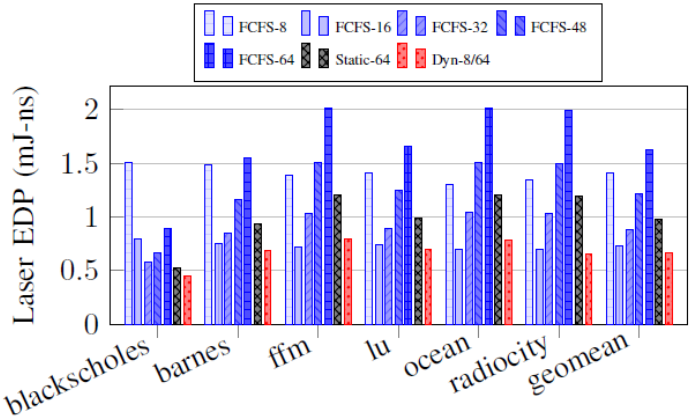
- Results indicated 13% improvement in throughput and 48% reduction in power consumption

Laser Pooling



- Share a pool of laser power between participating nodes that can be claimed on-demand
- Dynamically scaling laser power and the pool of optical channels
- FCFS allocation policy





- Improve energy-delay product by 31%

Conclusions

- Photonic interconnects can improve the performance/Watt when compared to traditional electrical interconnects
  - Wall-plug (coupling) efficiency of lasers
  - Thermal stability and sensitivity
- As traffic exhibits temporal and spatial fluctuations, bandwidth reconfiguration can improve throughput
- Static optical power (laser, trimming power) is a significant portion of the total power consumption and power reduction techniques are essential
- Machine learning algorithms could potentially provide higher throughput and power savings

Journal and Conference Publications & Students Graduated

JOURNALS

- Matthew Kennedy and Avinash Kodi, “Laser Pooling: Static and Dynamic Laser Power Allocation for On-Chip Optical Interconnects,” Accepted to appear in **IEEE/OSA Journal of Lightwave Technology (JLT)**, *Special Issue on Optical Interconnects Conference*, Sept/Oct 2017.
- Matthew Kennedy and Avinash Kodi, “CLAP-NET: Bandwidth Adaptive and Power Regulated Optical Crossbar Architecture,” **Elsevier Journal of Parallel and Distributed Systems (JPDC),** vol. 100, pp. 130-139, February 2017.
- Randy Morris, Evan Jolley and Avinash Kodi, “Extending the Performance and Energy-Efficiency of Nanophotonic Interconnects for Shared Memory Multicores,” **IEEE Transactions on Parallel and Distributed Systems (TPDS)**, vol. 25, no. 1, pp. 83-93, January 2014.
- Randy Morris, Avinash Kodi, Ahmed Louri and Ralph Whaley, “3D Stacked Nanophotonic Architecture with Minimal Reconfiguration,” **IEEE Transactions on Computers (TC)**, vol. 63, no. 1, pp. 243-255, January 2014.

CONFERENCES

- Ashif Sikder, Avinash Kodi and Ahmed Louri, “R-OWN: Reconfigurable Optical Wireless NoC Architectures,” *3rd ACM International Conference on Nanoscale Computing and Communication (NanoCom),* New York, NY, September 28-29, 2016**.**
- Matthew Kennedy and Avinash Kodi, “On Demand Laser Power Allocation for On-Chip Optical Interconnects” *Optical Interconnects Conference (OIC),* San Diego, CA, May 9-11, 2016.
- Scott VanWinkle, Matthew Kennedy, Dominic DiTomaso and Avinash Kodi, “Energy Efficient Optical Network-on-Chip Architecture for Heterogeneous Multicores,” *Optical Interconnects Conference (OIC),* San Diego, CA, May 9-11, 2016.
- Matthew Kennedy and Avinash Kodi, “Cross-Chip: Low Power Processor-to-Memory Nanophotonic Interconnect Architecture,” *Workshop on Energy-Efficient Networks of Computers (E2NC)* held in conjunction with *(IGSC’15)*, Las Vegas, NV, Dec 14-16, 2015.
- Ashif Sikdar, Matthew Kennedy, Avinash Kodi, Savas Kaya and Ahmed Louri, “OWN: Optical Wireless Network-on-Chips (NoCs) for Kilo-Core Architectures,” *23rd Annual Symposium on High-Performance Interconnects (Hot Interconnects),* Santa Clara, CA, August 26-28, 2015.
- Matthew Kennedy, Brian Neel and Avinash Kodi, “Runtime Power Reduction Techniques in On-Chip Photonic Interconnects,” *25th ACM’s Great Lakes VLSI Symposium (GLSVLSI)*, Pittsburgh, Pennsylvania, May 20-22, 2015.
- Matthew Kennedy and Avinash Kodi, “Design of Bandwidth Adaptive Nanophotonic Crossbars with Clockwise/Counter-Clockwise Optical Routing,” *28th International Conference on VLSI Design,* Bangalore, India, January 3-7, 2015.

STUDENTS GRADUATED

- Scott VanWinkle, “Shared Heterogeneous Architecture with Reconfigurable Photonic Network-on-Chips,” M.S. thesis, Ohio University, August 2017.
- Ashif Sikdar, “Emerging Technologies in On-Chip and Off-Chip Interconnection Network," M.S. thesis, Ohio University, August 2016.
- Matthew Kennedy, “Power-Efficient Nanophotonic Architectures for Intra- and Inter-Chip Communication," M.S. thesis, Ohio University, April 2016.
- Dominic DiTomaso, “Proactive and Reactive Fault Tolerant Network-on-Chips Architectures using Machine Learning,” Ph.D. Dissertation, August 2015.
- Brian Neel, “High-Performance Shared Memory Networking in Future Many-core Architectures Using Optical Interconnects," M.S. thesis, Ohio University, May 2014.

Questions?

THANK YOU!