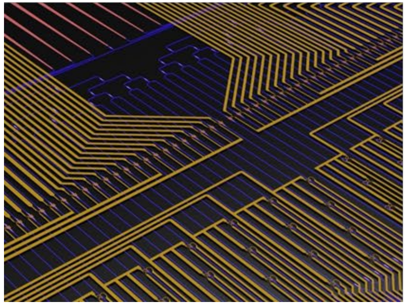
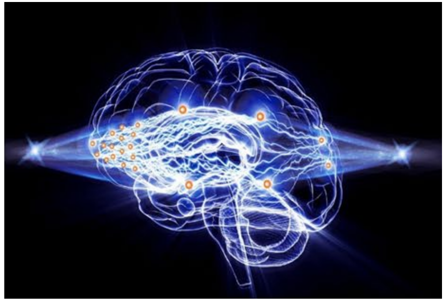
Processing in Memory for AI Acceleration with Silicon Photonics





Mini Keynote

MPSoC Forum 2024, July 7-12, Kanazawa, Japan

Sudeep Pasricha

Department of Electrical and Computer Engineering

Colorado State University, Fort Collins, CO, USA

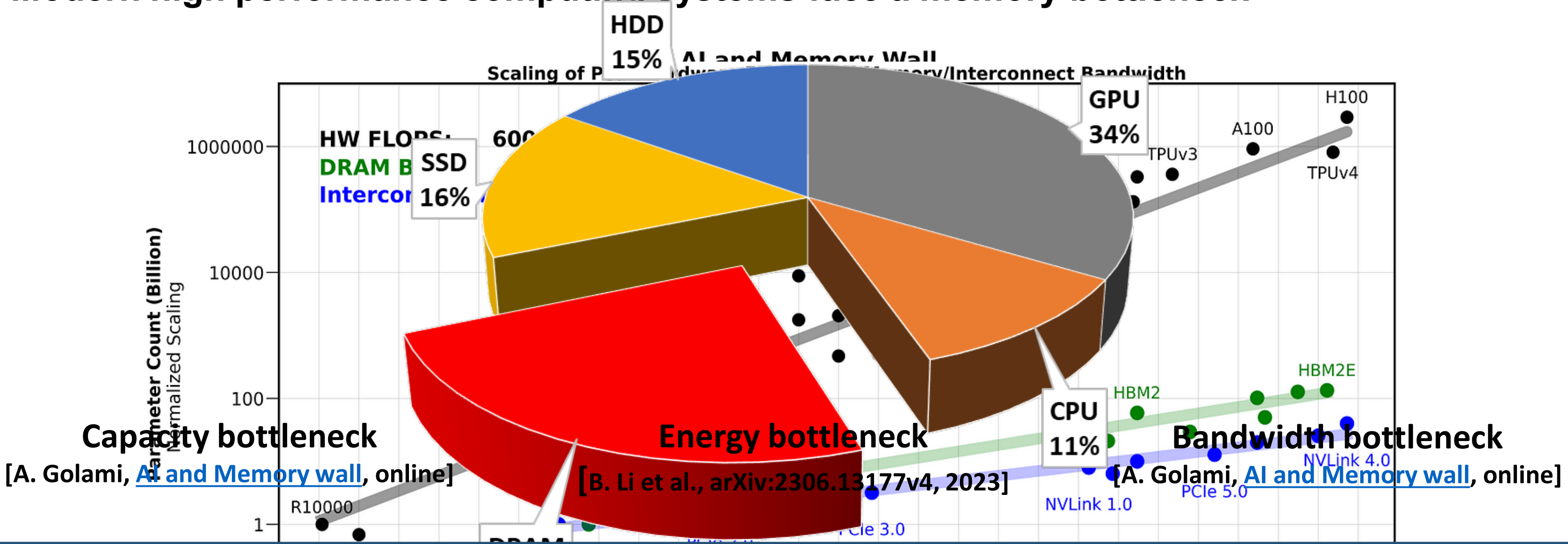
sudeep@colostate.edu

Colorado State University

EPIC Lab @ CSU

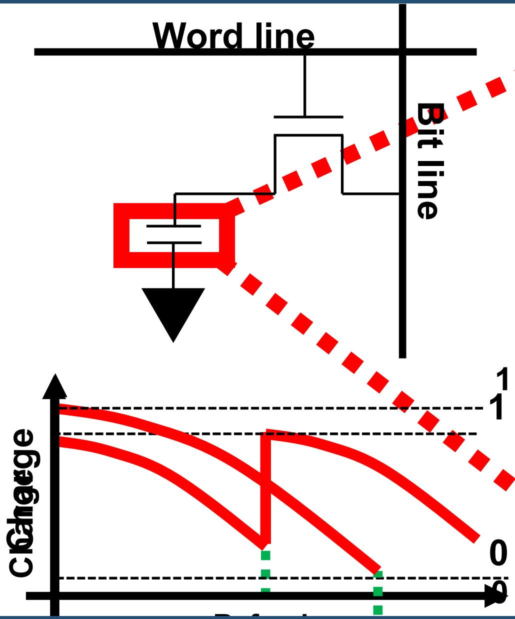
Memory Bottlenecks in Computing

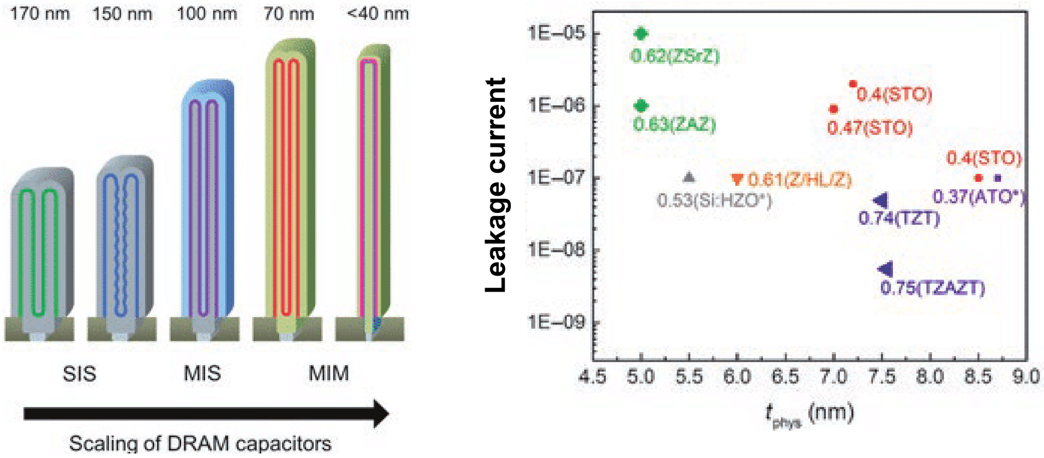
- **Modern high performance computing systems face a memory bottleneck**



For future computing systems, alternate memory technologies need to be considered urgently

DRAM Scaling Woes

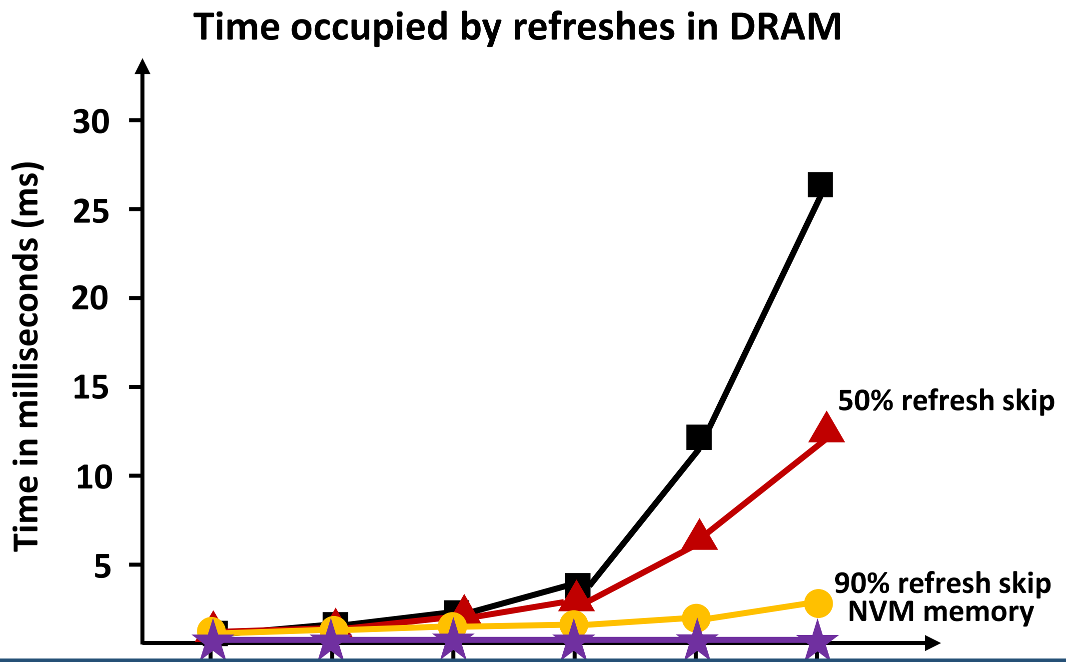




(*Left*) DRAM Capacitor structure scaling; *(Right)* Leakage current vs capacitor thickness [K. S. Kim and M. Popovici, *MRS Bulletin*, vol. 43, 2018]

Charge-based, volatile storage mechanism of DRAM limits main memory scaling

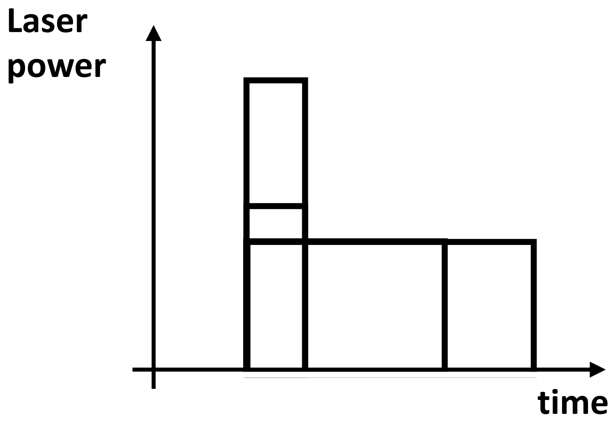
Reclaiming DRAM Refresh Overheads



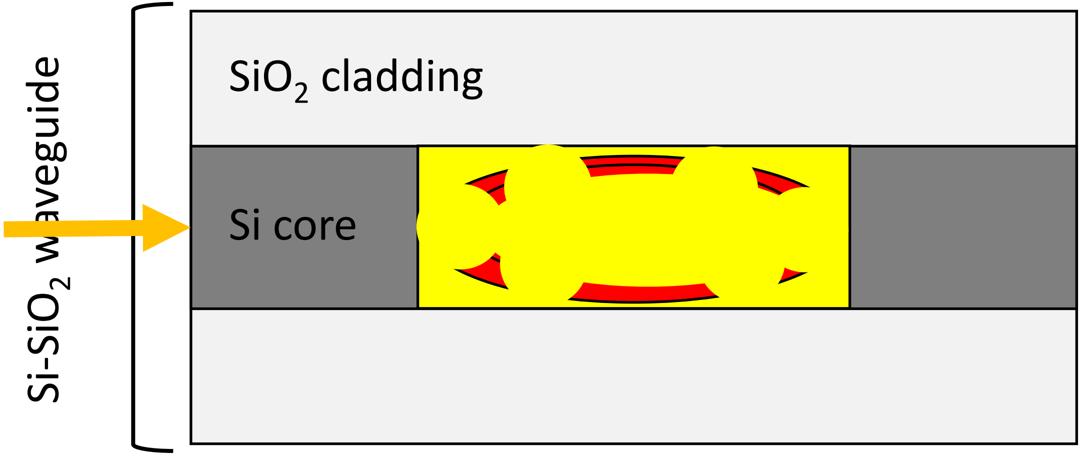
Non-volatile memory cells have promising potential over 1T-1C cells in a main memory architecture

Optically Programmed PCM Cells

Crystalline | Amorphous

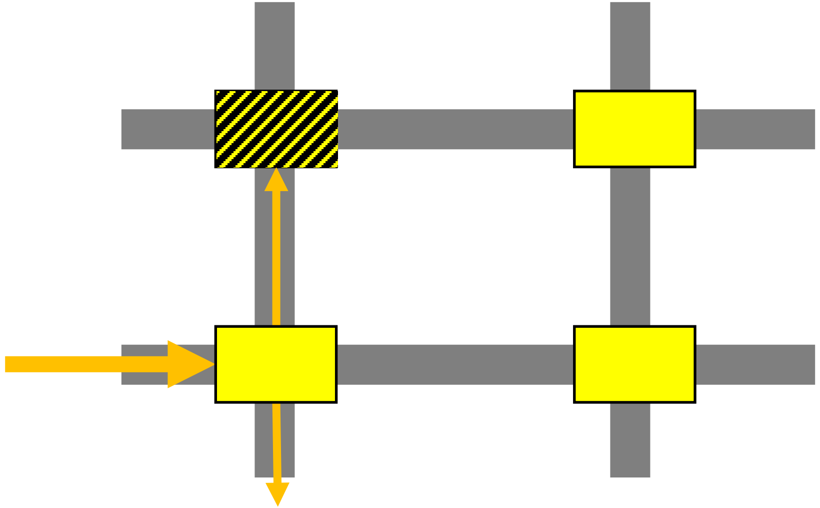


Optical PCM (OPCM)



OPCM cells can be used as building blocks for innovative optical main memory architectures

New Challenges: Thermal Crosstalk



Data corruption in crossbar-based OPCM memory after 4 writes to adjoining rows.

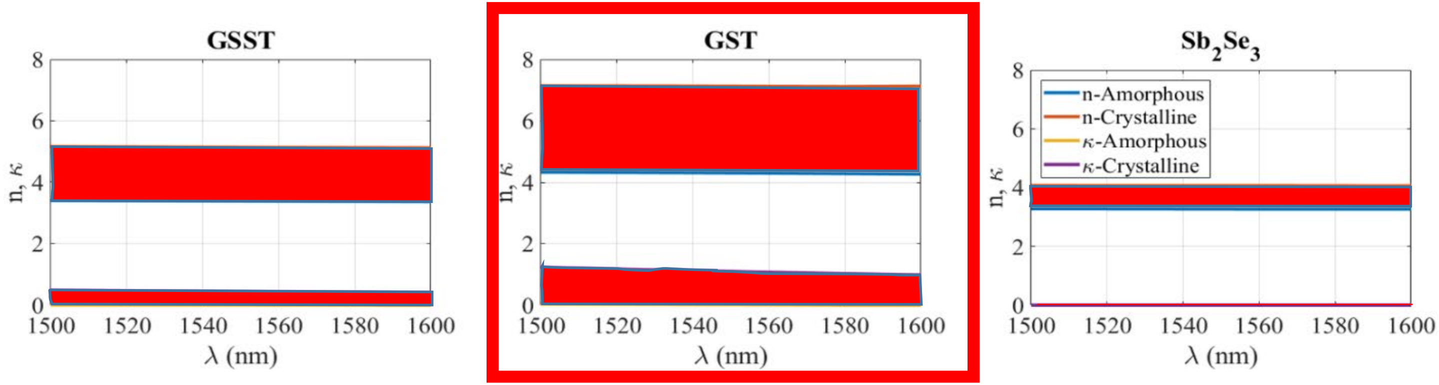
To preserve data integrity, OPCM cells must be isolated

COMET Photonic Main Memory

3. *COMET* Architecture	• Optical loss-aware architecture design • Optimizations for energy efficiency and reliability
2. Memory Cell Design	• Memory cell design for thermal crosstalk mitigation and efficient phase transition
1. PCM Selection	• Explored PCMs to determine the best material based on efficiency for optical memory use case

F. Sunny, A. Shafiee, B. Charbonnier, M. Nikdast, S. Pasricha, "COMET: A Cross-Layer Optimized Optical Phase Change Main Memory Architecture", ***IEEE/ACM DATE, Mar 2024.***

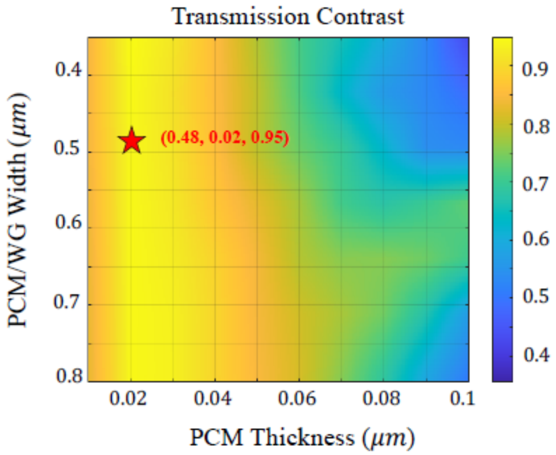
1. Phase Change Material Selection



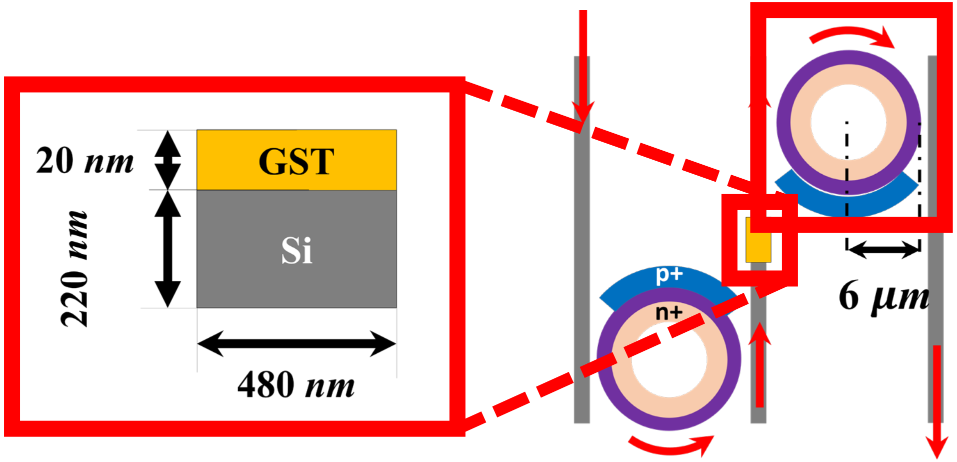
- **High transmittance contrast = High n (refractive index) contrast**
- **High κ (extinction coefficient) contrast = Energy efficient transition between states**
  - As κ relates to the amount of energy transferred to the bulk

PCM: Phase Change Material

2. OPCM Memory Cell Design



geometric configuration with values for (width, thickness, transmission contrast ratio)

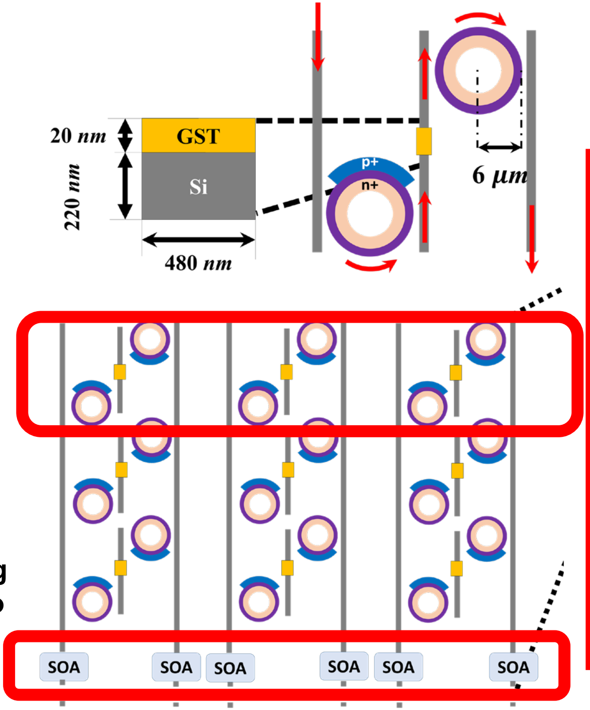


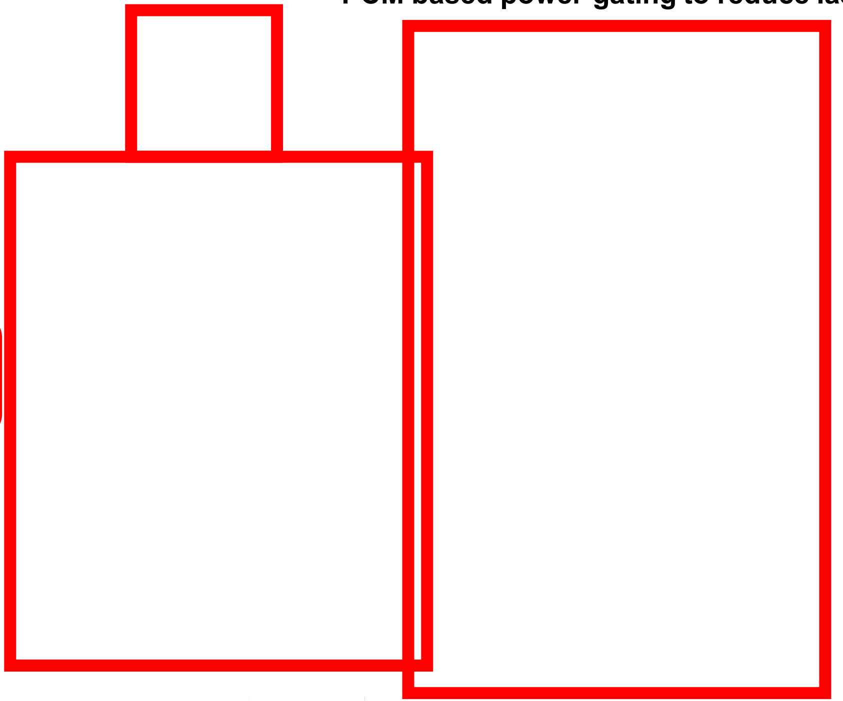
Microring resonators (MRs): Wavelength sensitive photonic switches

- **The bulk and dimensions of the OPCM cell impact n and κ values**
- **High extinction coefficient (κ) contrast between amorphous and crystalline states needed**
  - So that data readout is reliable
  - To accommodate additional transmission levels for MLC operation

3. *COMET* Architecture Overview

- **PCM based power-gating to reduce laser power**





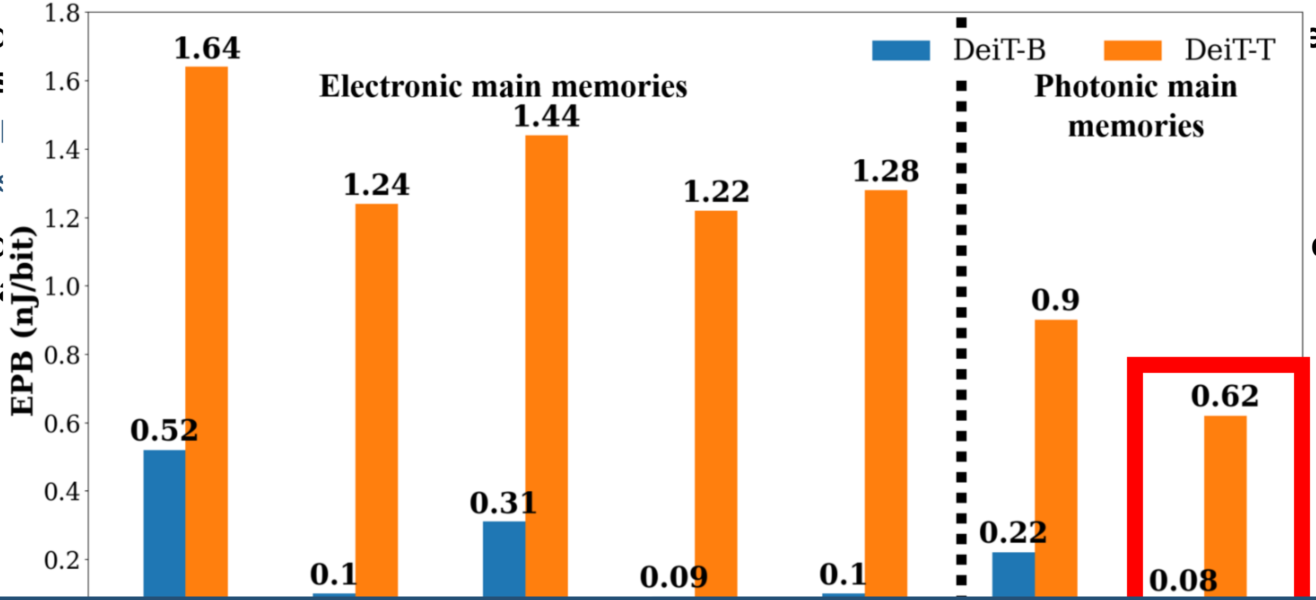
SOA-based intermediate gain tuning for improving signal-to-noise ratio (SNR)

- **Subarrays remain completely disaggregated**
- **Reduced crosstalk and thermal noise**
- **Reduction in input power requirement**

SOA: Semiconductor Optical Amplifier; MR: Microring Resonator

Optical ML Accelerator Case Study

- **We consic** ... **e as photonic s**
  - DOTA [I
  - DOTA is
- **We consic** ... **earch/deit] for our ana**



co

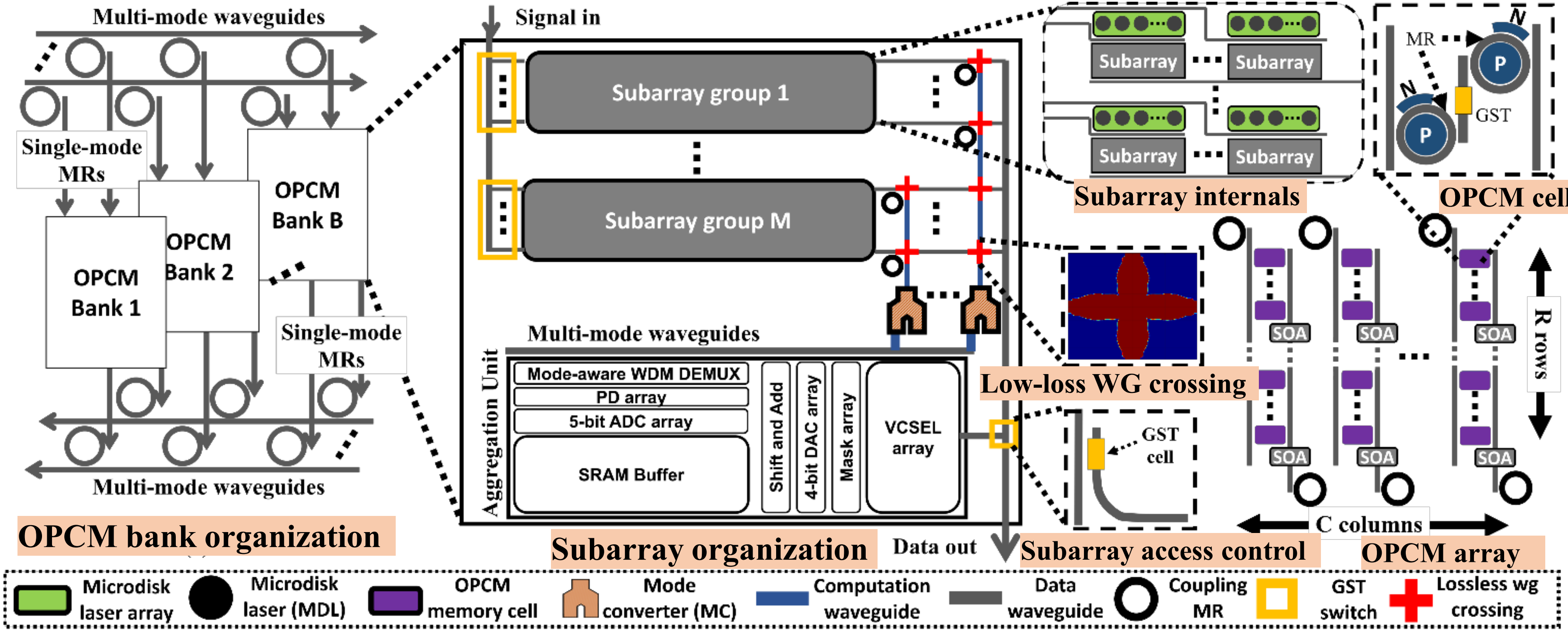
EP

COMET+DOTA achieves at best 2.06× lower EPB against 3D_DDR4+DOTA and 2.7× better EPB against COSMOS+DOTA

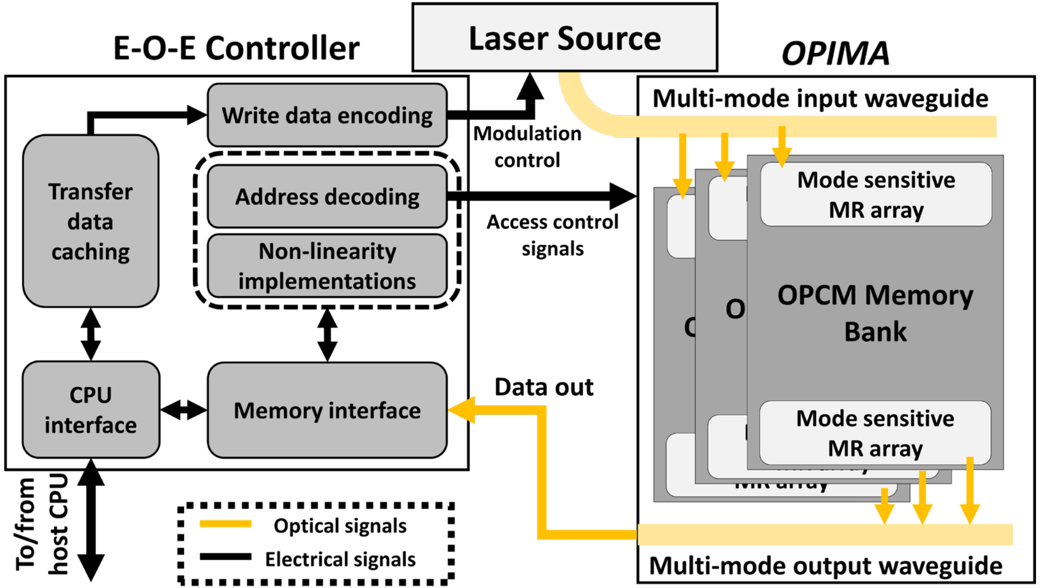
Processing in Optical Memory

- **Can we repurpose *COMET* for PIM?**
  - **Challenge 1: Supporting higher levels of parallelism**
    - Need to leverage additional mechanisms to increase memory access and computation parallelism beyond those offered by WDM
      - Leverage WDM+MDM for greater parallelism
  - **Challenge 2: Concurrent memory and computation operation**
    - Reads should be supported from a selected subarray or a group of subarrays as needed, without interrupting the main memory operation
      - Redesign bank and subarray architectures
  - **Challenge 3: Interference-free accesses**
    - When simultaneously read out, data from computation outputs and main memory accesses must not interfere with each other in an undesirable manner
      - Optimize waveguide topology and waveguide crossing design
  - **Challenge 4: Variable precision support**
    - Architecture should support PIM operations between parameters (e.g., CNN weights/activations) of any size, irrespective of bit density used in OPCM cells
      - Leverage TDM and optimize aggregation unit design

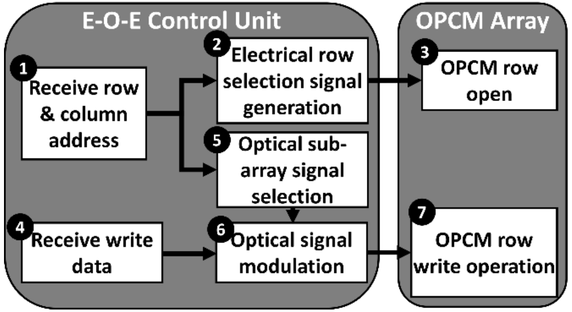
OPIMA Architecture Overview



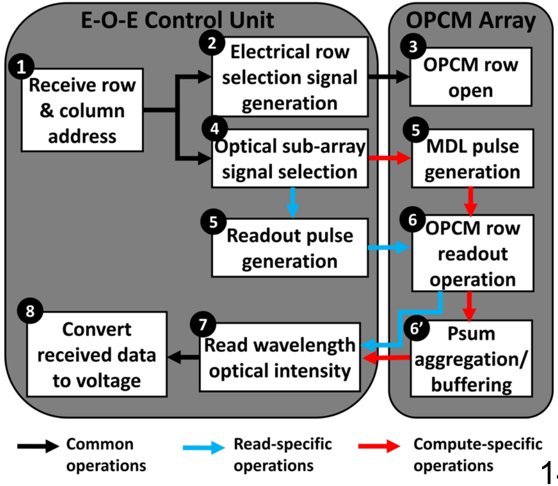
OPIMA Controller Design



Memory Write Control Flow

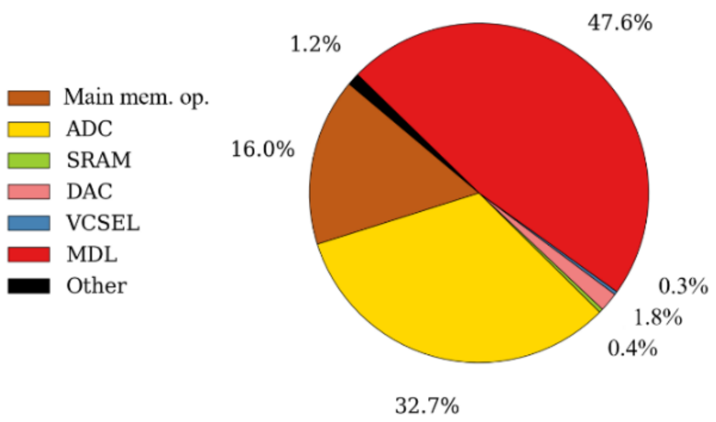


Memory Read + PIM Control Flow

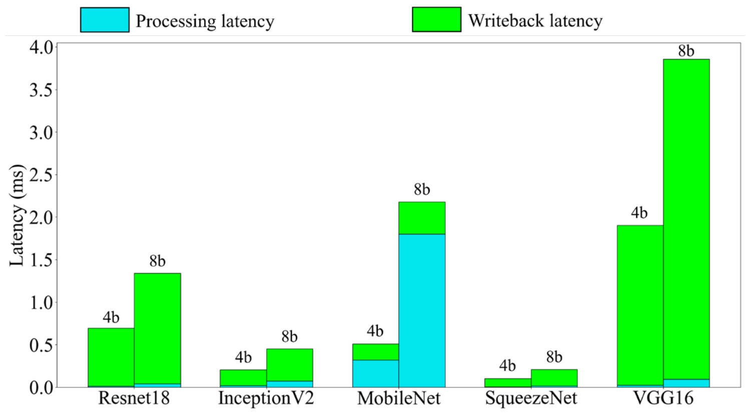


Experimental Analysis

Model	Dataset	Accuracy (fp32)	Accuracy (int8)	Accuracy (int4)	Parameter count
Resnet18	CIFAR100	75.3%	74.2%	72.6%	11584865 (11.6 M)
InceptionV2	SVHN	81.5%	80.8%	75.9%	2661960 (2.6 M)
MobileNet	CIFAR10	88.2%	87.5%	83.5%	4209088 (4.2 M)
SqueezeNet	STL-10	92.5%	90.3%	86.5%	1159848 (1.1 M)
VGG16	Imagenette	98.96%	96.25%	93.7%	134268738 (134.3 M)

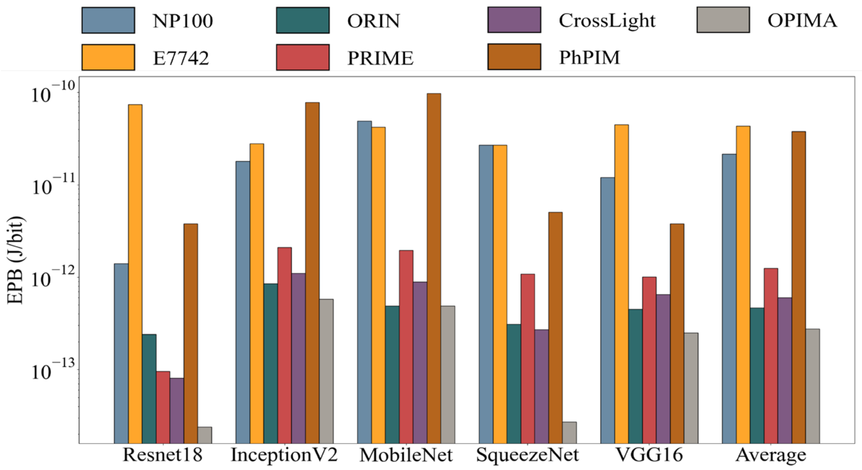


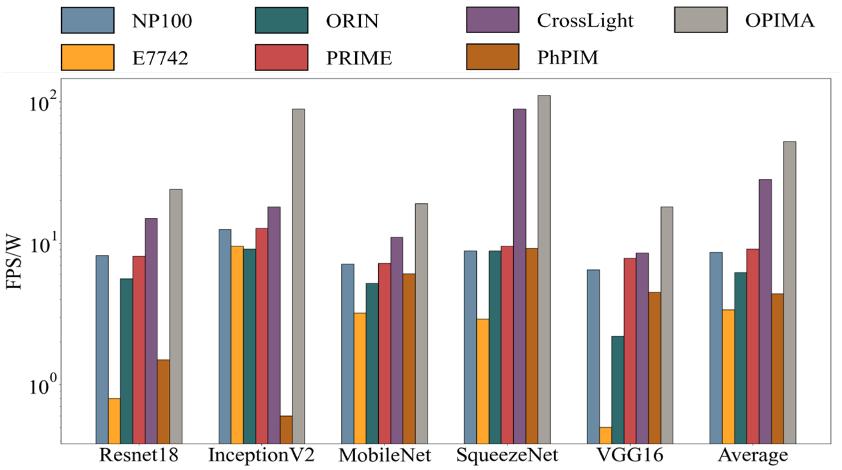
OPIMA power breakdown



OPIMA latency breakdown

EPB and FPS/W Comparisons





- ***OPIMA* outperforms CPU/GPU/NPU/PIM architectures by 83.1x (EPB) and 27.5x (FPS/W)**
  - Also outperforms SOTA photonic PIM architecture PhPIM by 186x and 55.3x

OPIMA is a promising PIM architecture for AI acceleration

Conclusions

- **Presented *COMET*, a low-loss, low-latency, and high throughput OPCM-based main memory architecture**

- ***COMET* consumes only 26% of power and has 2.7x lower EPB when paired with optical AI accelerator, vs. the only other OPCM-based main memory design, *COSMOS***

- **Extended *COMET* to design *OPIMA*, an innovative optical PIM architecture with high throughput, low latency, and high energy efficiency**

- ***OPIMA* outperforms various CPU/GPU/NPU/PIM architectures, and has 186x lower EPB and 55.3x higher FPS/W than the only other photonic PIM architecture, *PhPIM***

- **Optical PIM has excellent potential to be a very competitive solution for AI acceleration in emerging applications**

Acknowledgements

MONFORT FAMILY FOUNDATION

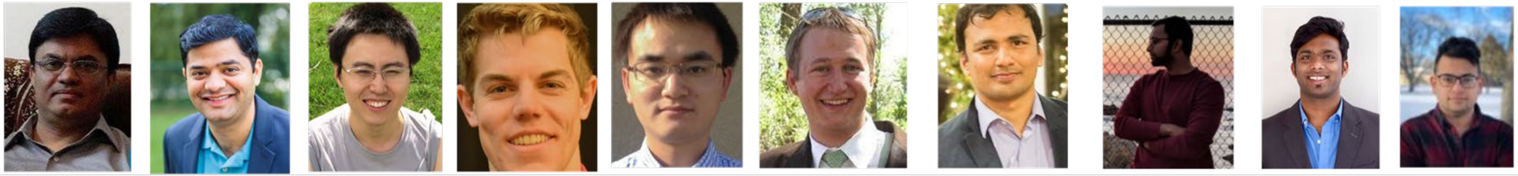
Hewlett Packard Enterprise

NSF

SRC

Ph.D. students

20+ (former, current)



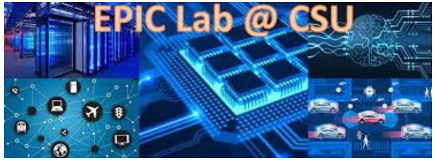
M.S. students

40+ researchers

Undergraduate students

100+ researchers

Faculty and industry collaborators:



Thank You!

Sudeep Pasricha
sudeep@colostate.edu