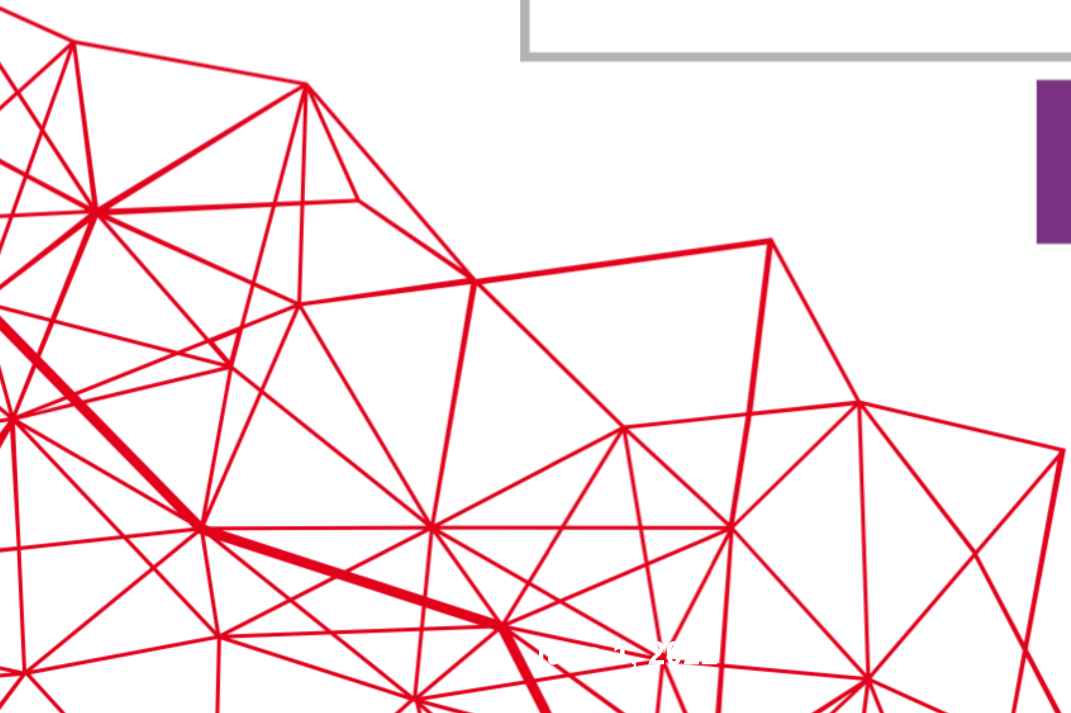
ISC High Performance

TRANSFORMING THE FUTURE

MAY 29 – JUNE 2, 2022 | HAMBURG, GERMANY



LLM: Realizing Low-Latency Memory by Exploiting Embedded Silicon Photonics for Irregular Workloads

Marjan Fariborz, Mahyar Samani, Pouya Fotouhi, Roberto Proietti, Il-Min Yi, Venkatesh Akella, Jason Lowe-Power, Samuel Palermo, and S. J. Ben Yoo

University of California Davis, Texas A&M University

Outline

- Motivation
- Background on Silicon Photonic
- LLM Architecture
- Evaluation methodology
- Evaluation results
- Conclusion

Large Scale Irregular Application

- Modern applications have irregular memory access pattern with low locality
- Memory system is the bottleneck
- Rethink the architecture of the memory systems for these applications.
  - Low latency
  - High bandwidth
  - Low memory access variation
- Main source of latency is the contention caused by sharing resources.

Recommendation System

Mining large graphs

Speech Recognition

Sources of Contention

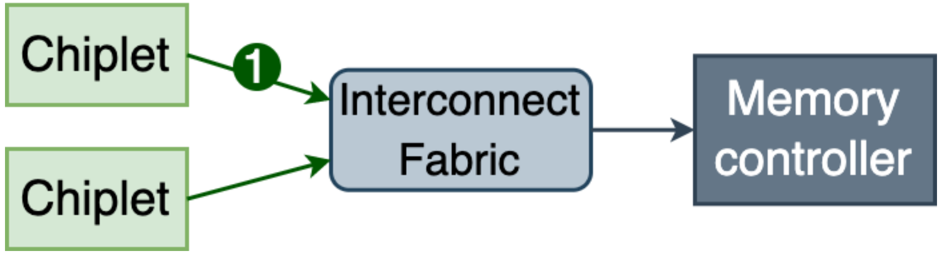
1. **Interconnect:**

Post-Moore’s law era: Replacing large monolithic dies into smaller “chiplets”.

Interconnection between chiplets have challenges.

Chiplets require to share interconnect resources.

❶ Interconnect fabric



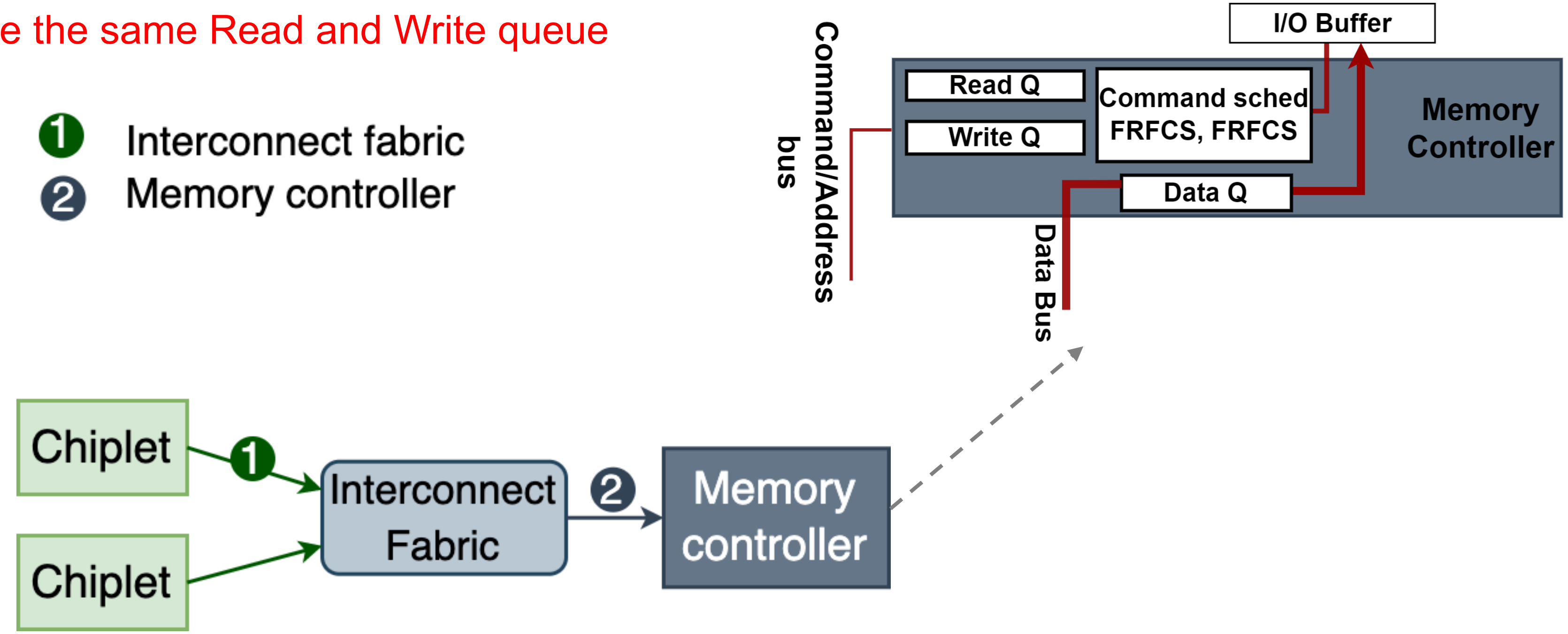
Sources of Contention

2. Memory Controller:

Single memory controller per channel

- Single command and data bus.
- Memory timing constraints.
- Maintain low latency and high throughput

Requests targeting the same channel share the same Read and Write queue

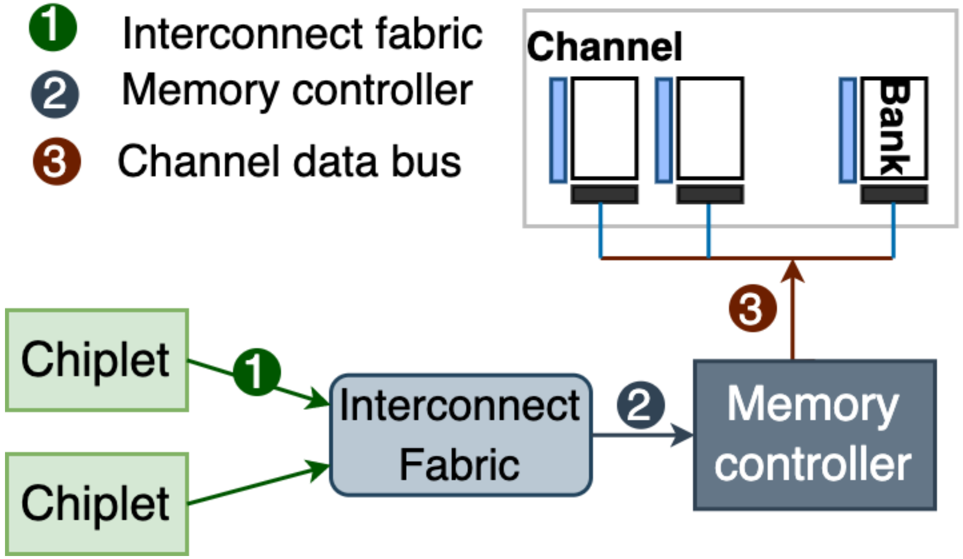


Sources of Contention

3. **Memory channel :**

DRAM systems are organized into a hierarchy of channels, banks, rows, and columns to exploit locality and parallelism.

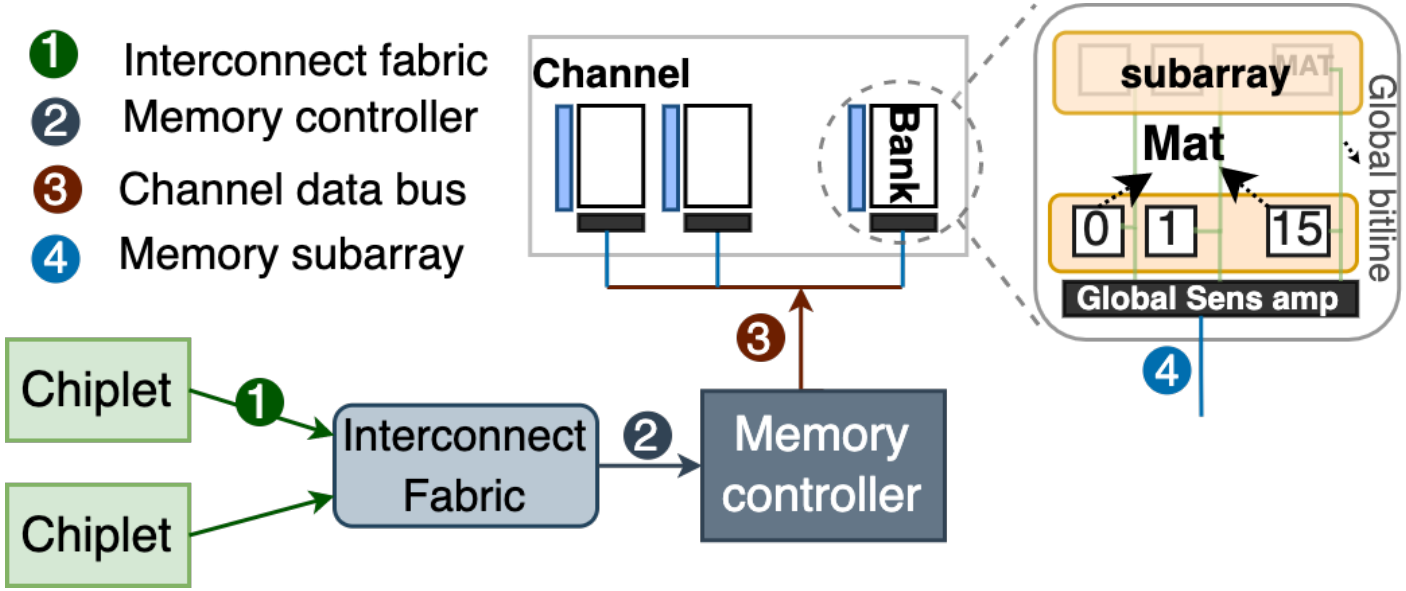
Memory banks inside of the single memory channel share the same data and command bus.



Sources of Contention

4. **Memory bank (bank conflict):**

   - Single sense amplifier
   - Global bitlines
   - Command decoder



Related Work

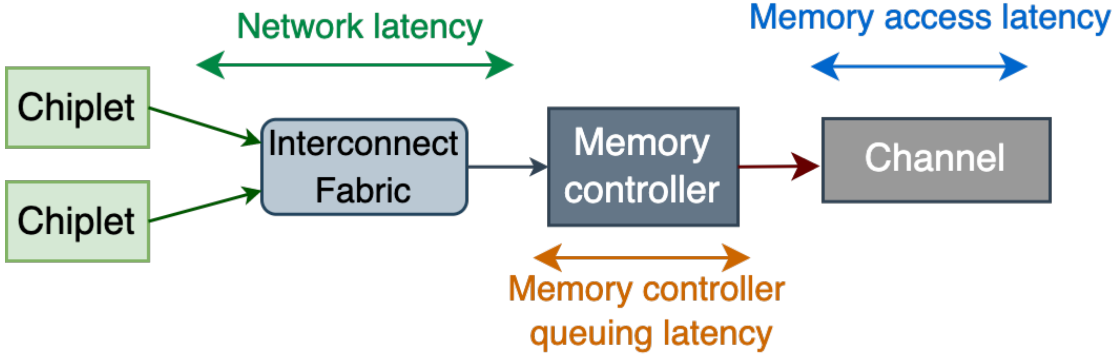
- **"Combining memory and a controller with photonics through 3d-stacking to enable scalable and energy-efficient system"**, Rohbani *et al.*, ISCA2021
- **"Reducing Memory Access Latency with Asymmetric DRAM Bank Organizations"**, son *et al.*, ISCA 2013
- **"A case for exploiting subarray-level parallelism (SALP) in DRAM"**, Kim *et al.*, ISCA.2012

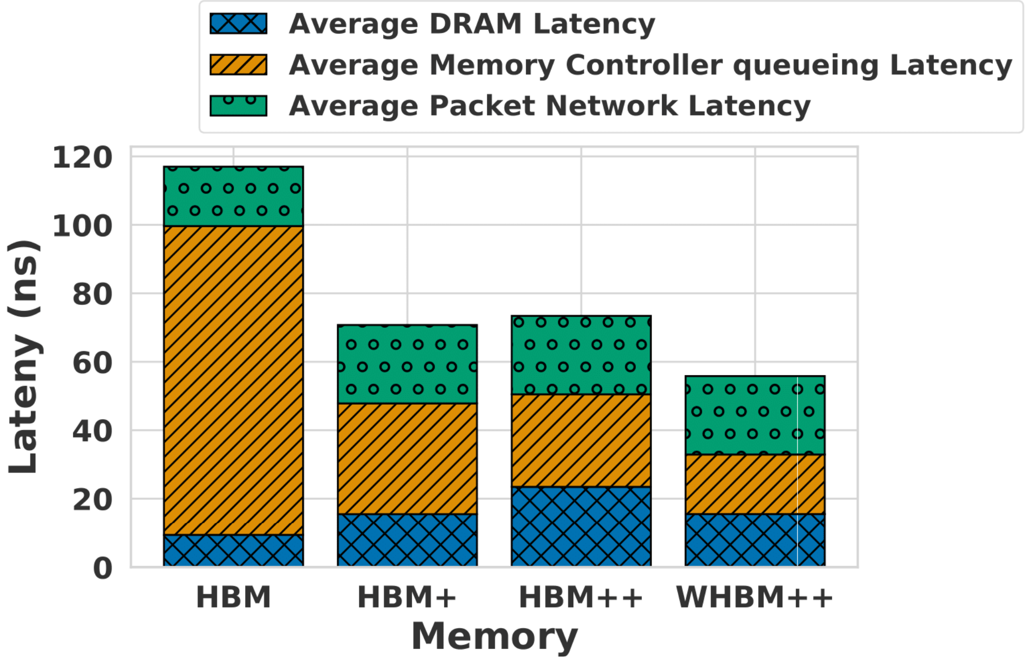
→ Low Latency

- **"Fine-Grained DRAM: Energy-Efficient DRAM for Extreme Bandwidth Systems"**, O'Connor *et al.*, Micro 2017
- **"Combining memory and a controller with photonics through 3D-stacking to enable scalable and energy-efficient systems"**, Udipi *et al.*, ISCA2011
- **"Re-architecting dram memory systems with monolithically integrated silicon photonics"**, Beamer *et al.*, ISCA 2010

→ High Bandwidth

End-to-End Latency Analysis





	HBM (HBM2.0)	HBM+	HBM++	WHBM++
Channel/stack	8	8	8	8
Pseudo-channel/channel	2	4	16	8
Banks/channel	16	16	32	32
Pins/pseudo-channel	64	32	8	64

Low data bus contention

Low data bus contention
Lower bank conflict

Higher peak bandwidth

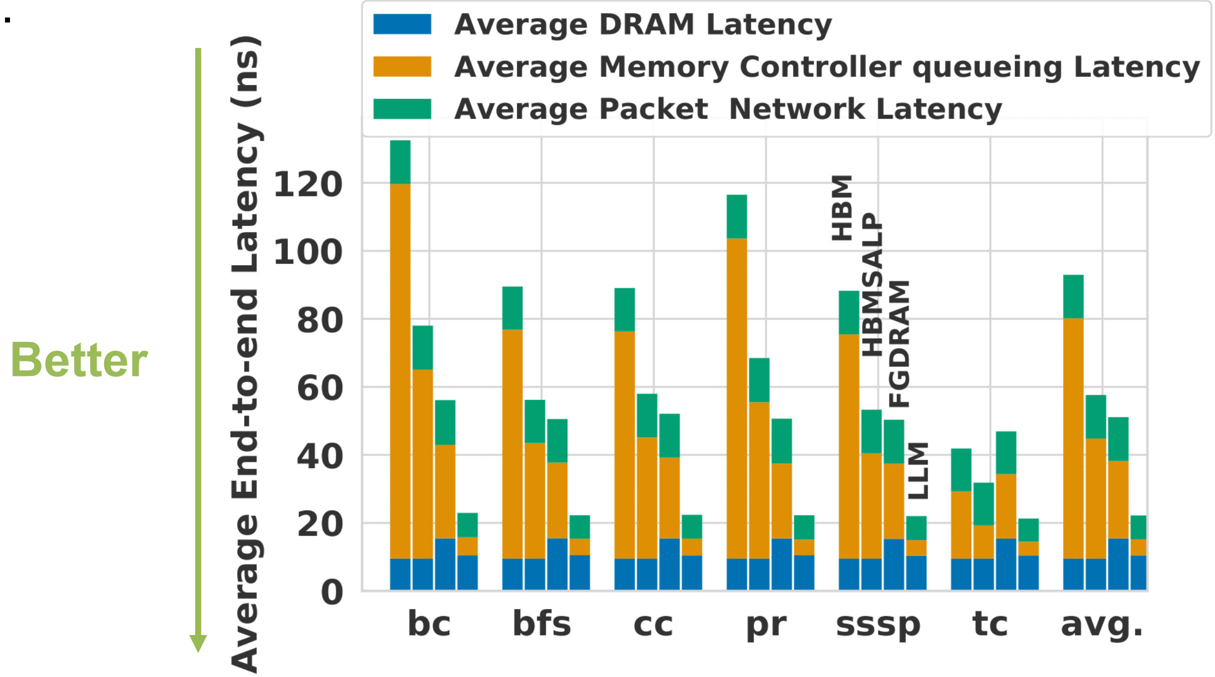
Low Latency Memory (LLM)

Design:

- **Remove the Contention**
- Co-design the components on the data path.
- All optical data plane with **Silicon Photonic** (SiPh).

Benefits:

- Lower queuing latency
- Lower latency variation
- Low energy per bit using SiPh
  - More parallelism → High bandwidth



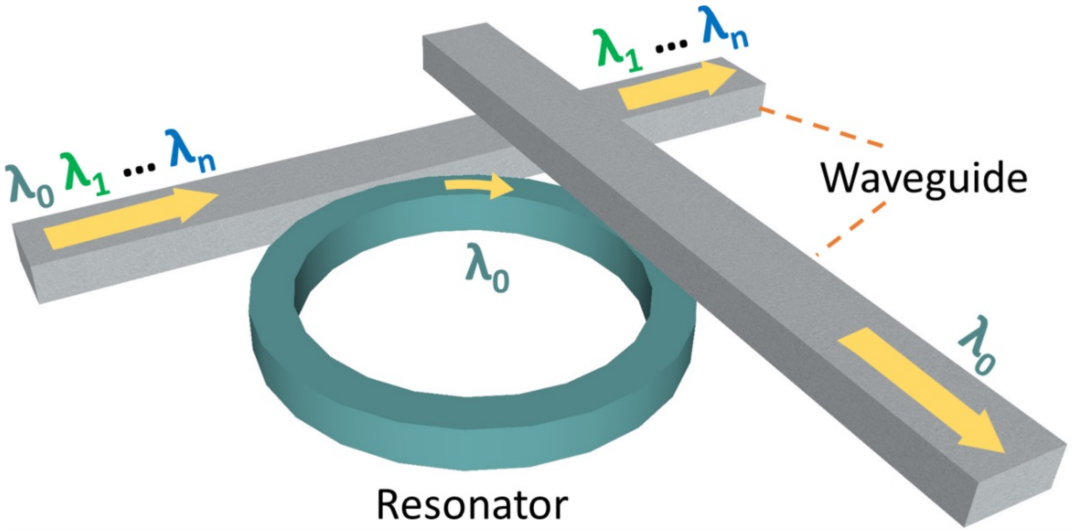
Outline

- Motivation
- Background on Silicon Photonic
    - Microring
    - Silicon-Photonic Link
    - Array Waveguide Grating Router
- LLM Architecture
- Evaluation methodology
- Evaluation results
- Conclusion

Background on Silicon Photonic (SiPh)

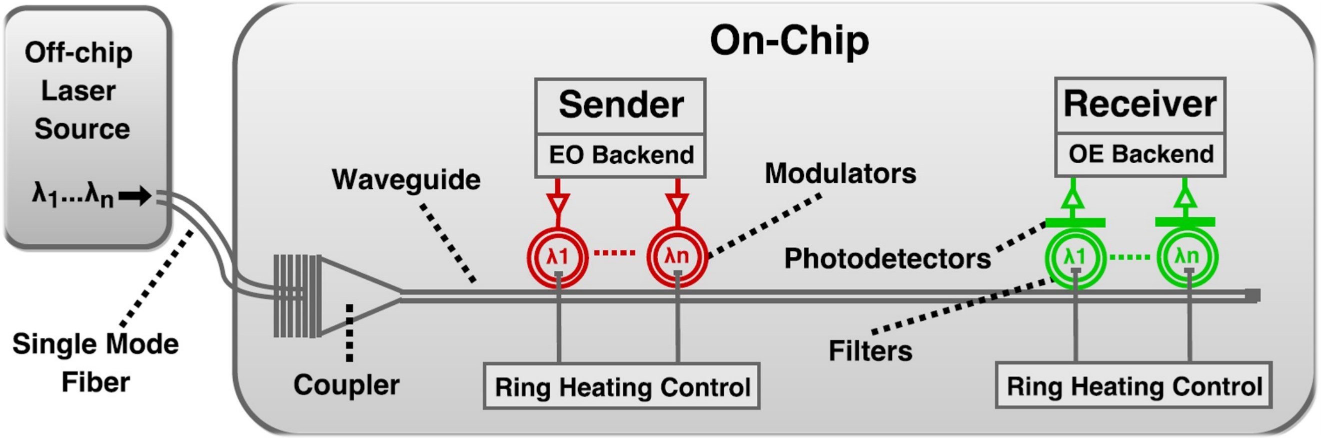
- Microring Resonators (MRs)
  - Resonates at a particular wavelength (**λ**)
  - Filtering
  - Modulation

- Silicon-Photonic Link
  - Off/On-Chip laser
  - MRs as modulator at sender
  - MRs as filters at receiver
  - Heat-control to tune MRs



λ_1 ... λ_n
λ_0 λ_1 ... λ_n
Waveguide
λ_0
λ_0
Resonator

Filtering out using MRs

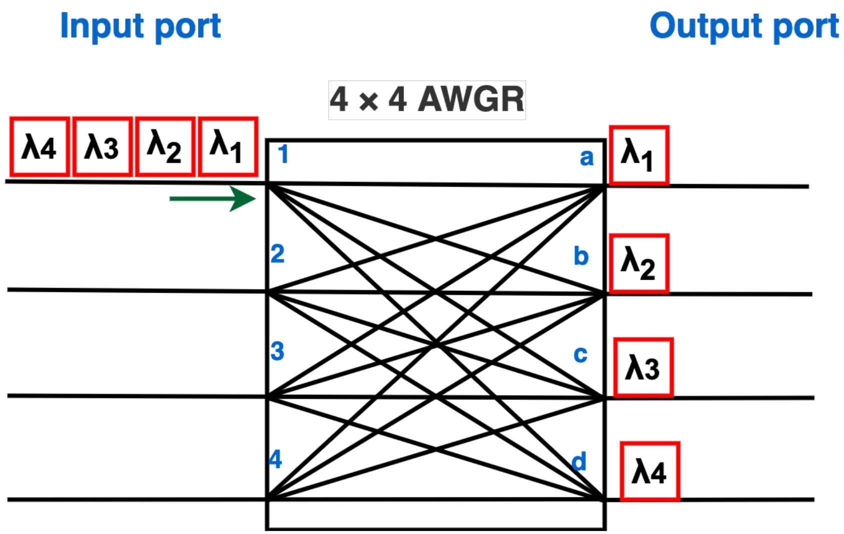


Reference SiP Link

Arrayed Waveguide Grating Router (AWGR)

- AWGR
  - Wavelength multiplexer
  - Passive device
  - Bidirectional
  - Compact layout (<1mm^2)

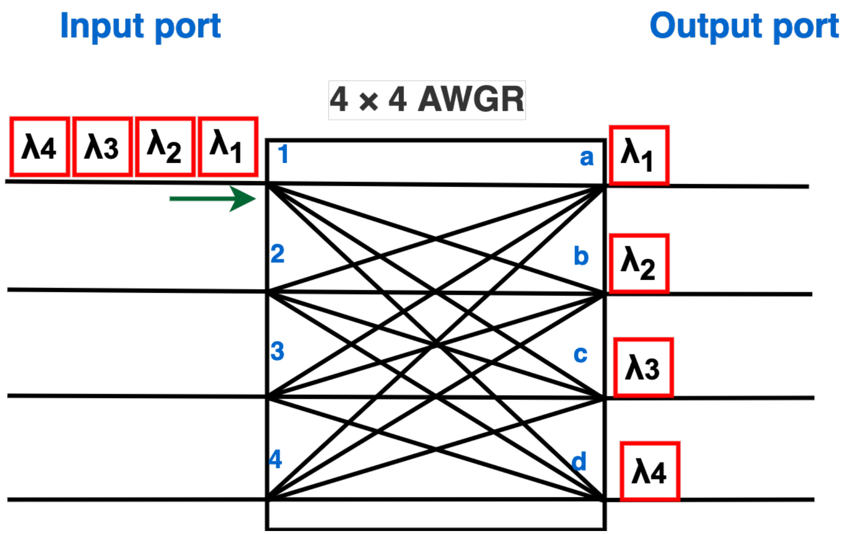
Contention-free one-to-all



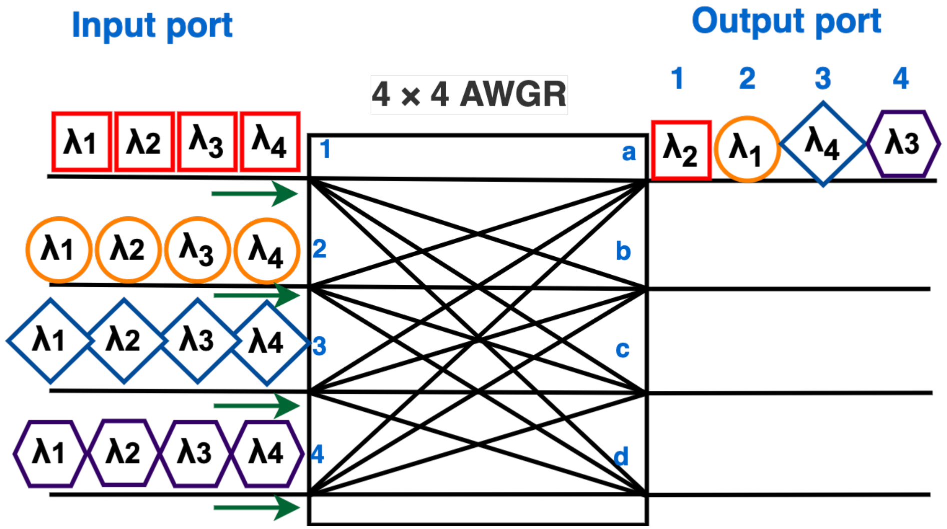
Arrayed Waveguide Grating Router (AWGR)

- AWGR
  - Wavelength multiplexer
  - Passive device
  - Bidirectional
  - Compact layout (<1mm^2)

Contention-free one-to-all



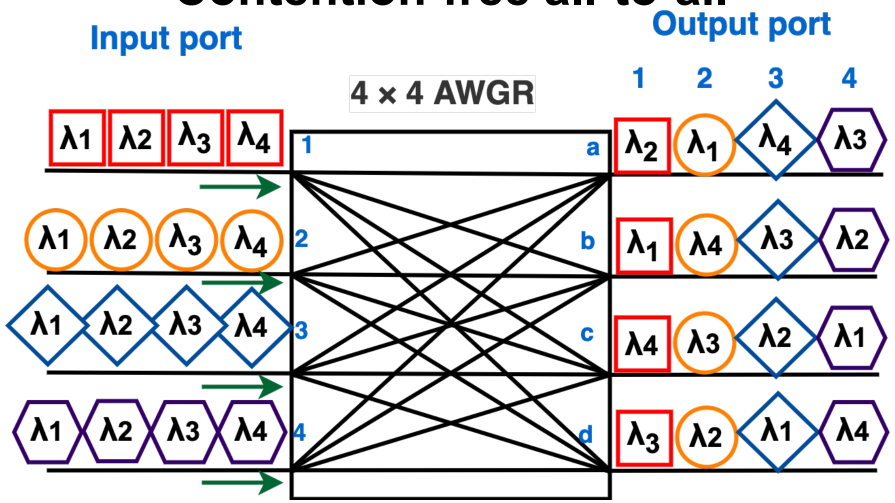
Contention-free all-to-one



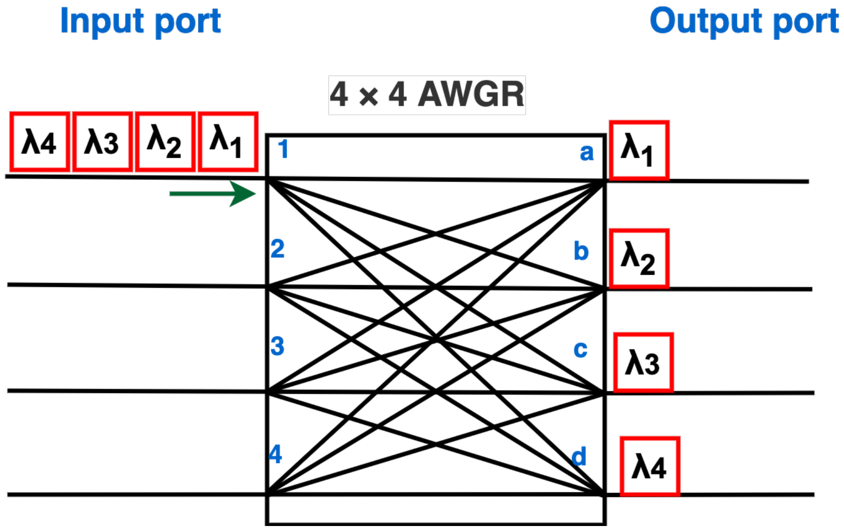
Arrayed Waveguide Grating Router (AWGR)

- AWGR
  - Wavelength multiplexer
  - Passive device
  - Bidirectional
  - Compact layout (<1mm^2)

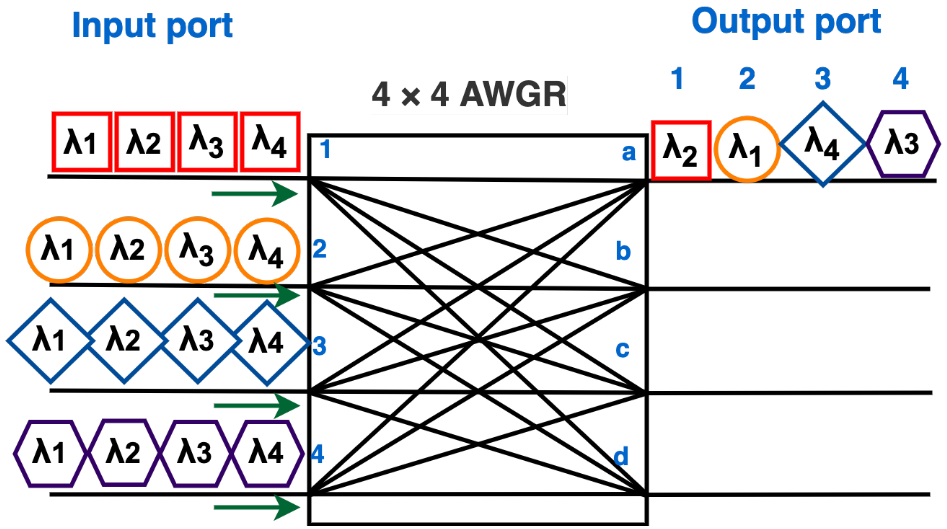
Contention-free all-to-all



Contention-free one-to-all



Contention-free all-to-one



Outline

- Motivation
- Background on Silicon Photonic
- LLM Architecture
  - Processor memory interconnect
  - Memory controller
  - Memory microarchitecture
  - Organization
- Evaluation methodology
- Evaluation results
- Conclusion

Low Latency Memory (LLM)

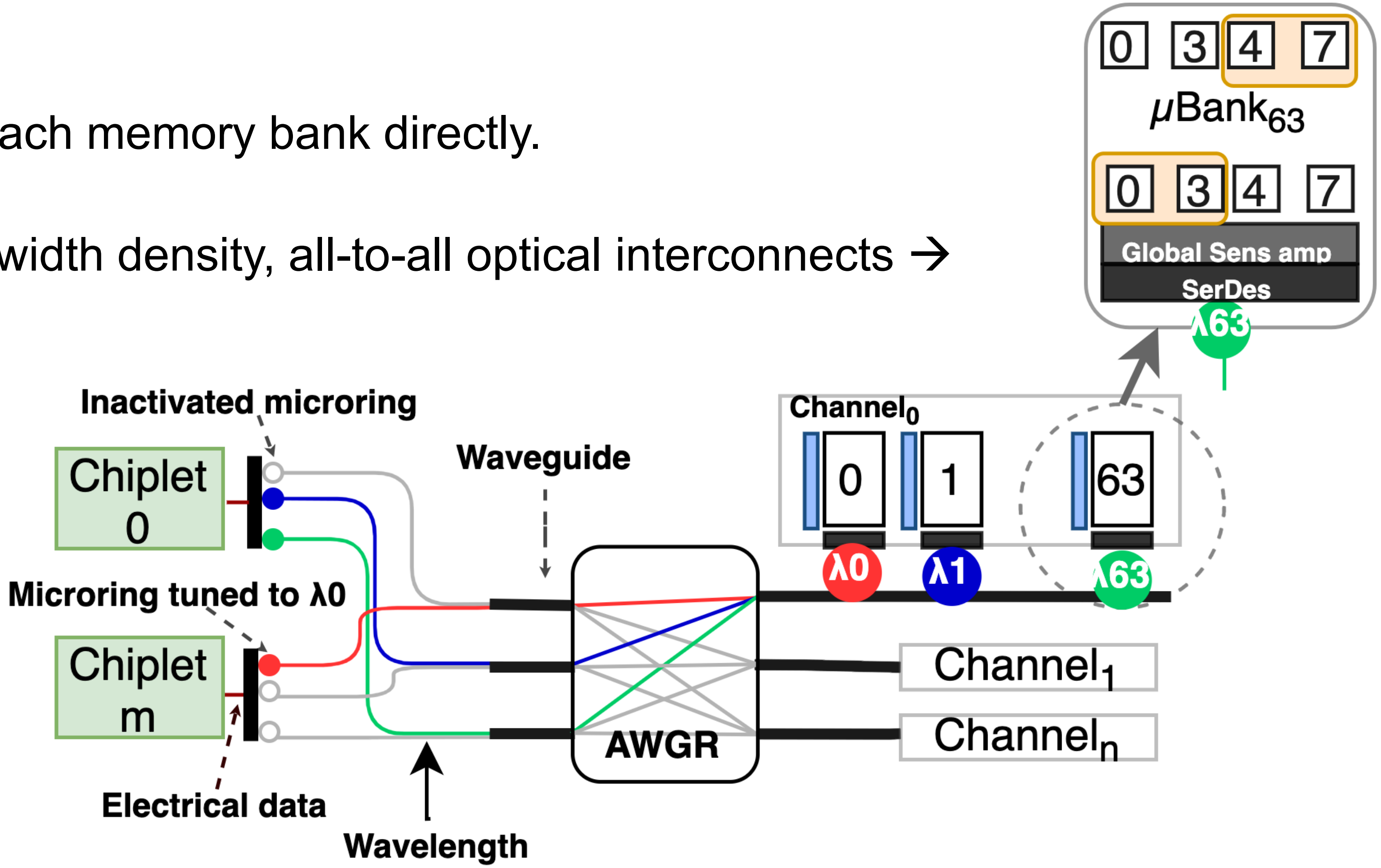
Removing end-to-end contention:

- Ground up co-design of the entire path
  - Interconnect
  - Memory controller
  - Memory microarchitecture
- Separating data and control plane:
  - Optical data plane
  - Electrical control plane

LLM: Processor Memory Interconnect

Data plane

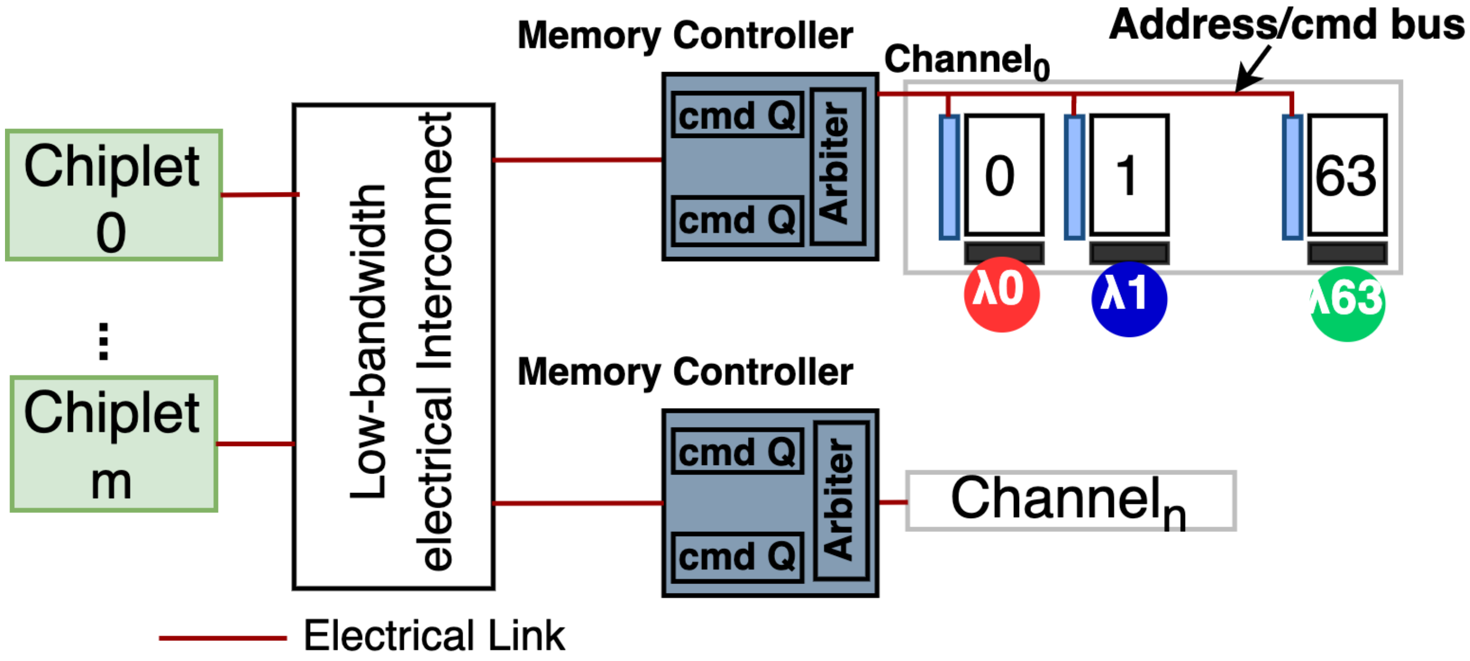
- Connecting each chiplet to each memory bank directly.
- Using low energy, high bandwidth density, all-to-all optical interconnects → AWGR.



LLM: Processor Memory Interconnect

Electrical Control Plane

- Low bandwidth electrical interconnect
  - Requests
  - Handshaking signals



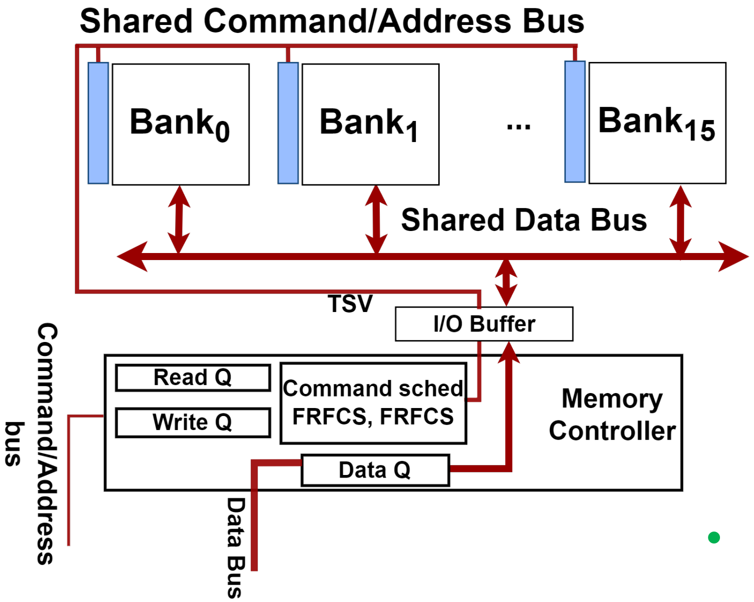
LLM: Memory Controller Architecture

HBM memory controller

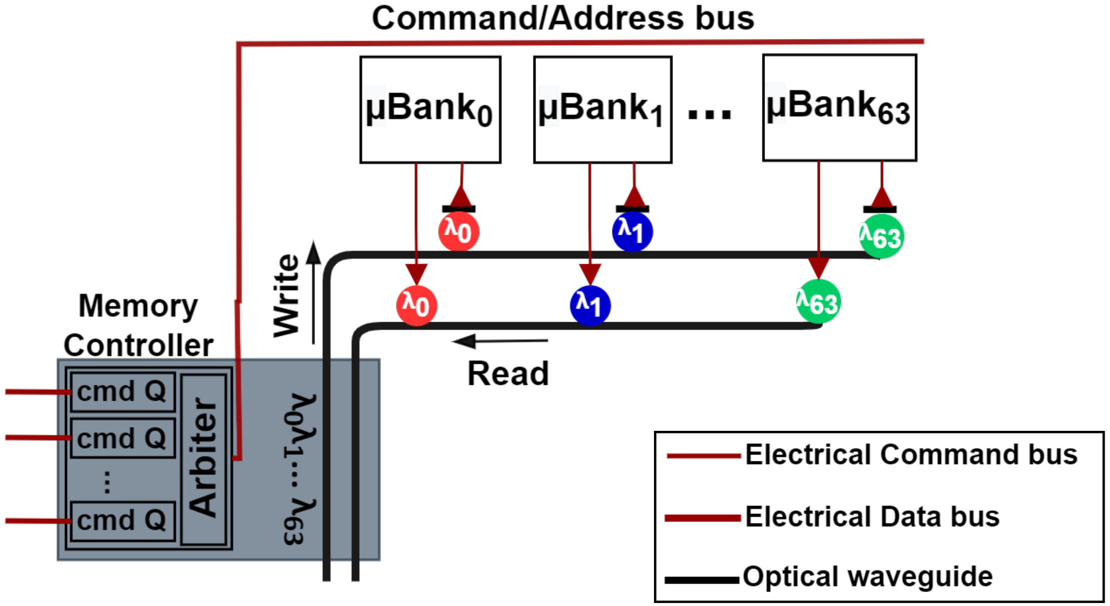
- Shared electrical data bus
- Long shared data and command queue

LLM memory controller

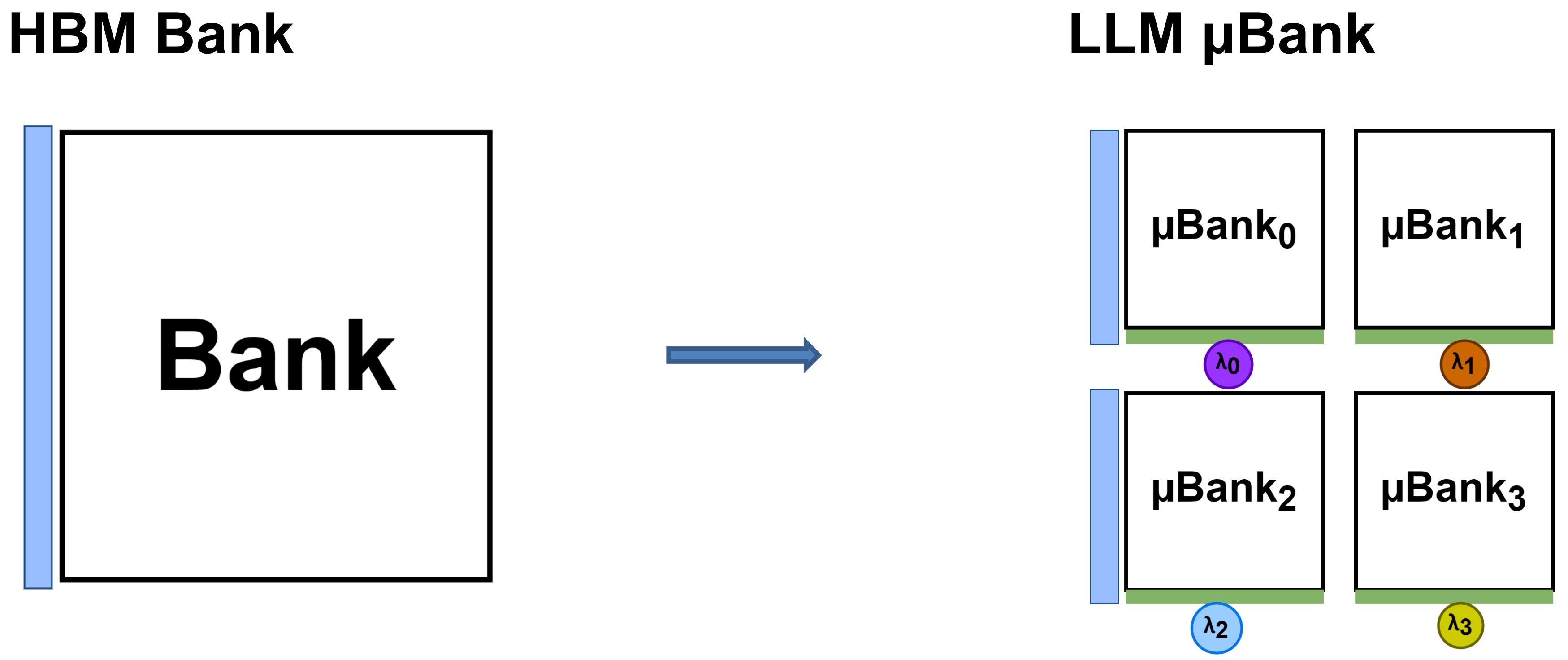
- Dedicated optical data link
- No data queue
- Dedicated single entry command queue



- Reducing bus conflict



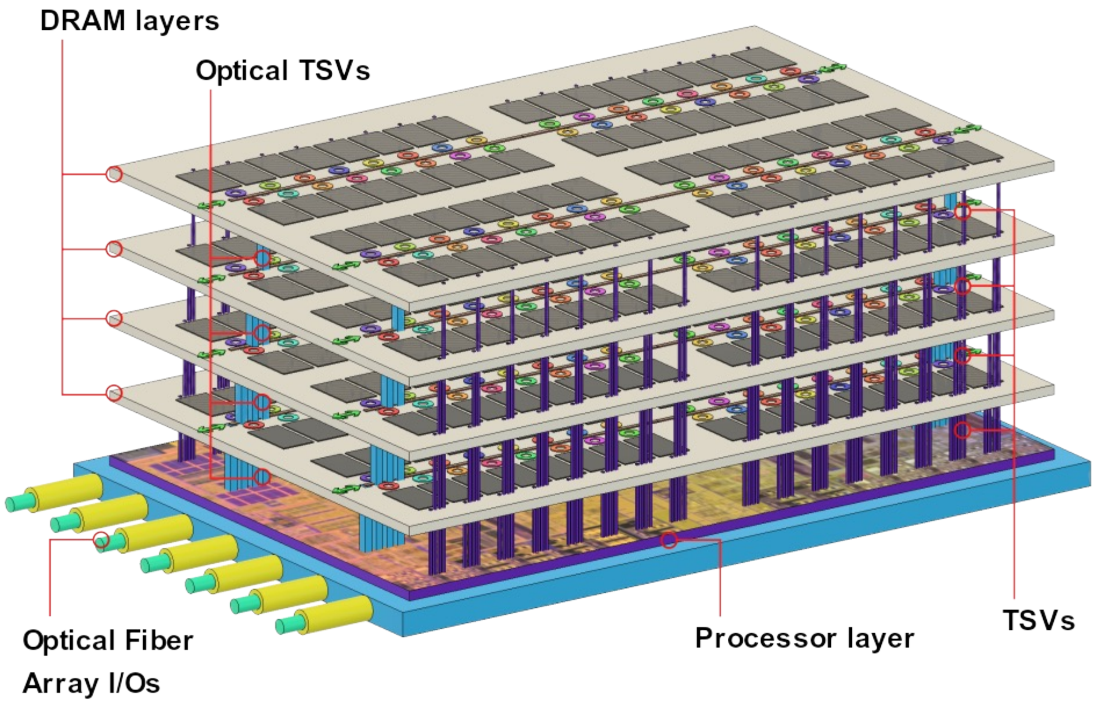
LLM: Bank Architecture



- 2x lower in-bank data movement latency
- 2x lower data movement energy
- 4x lower activation energy

LLM organization

- **3D stacks**
  - **High bandwidth**
  - **High capacity**
  - **Replace data TSVs with Vertical Optical Interconnect (VOI)**
- Non-stacked



Outline

- Motivation
- Background on Silicon Photonic
- LLM Architecture
- Evaluation methodology
- Evaluation results
- Conclusion

Evaluation methodology

gem5 simulator version 21.0

- Baseline memory systems:
  - HBM2.0
  - HBM + SALP
  - FGDRAM

- Synthetic test:
  - 32 traffic generators
  - Iso-Bandwidth test
    - Memories have the same peak bandwidth

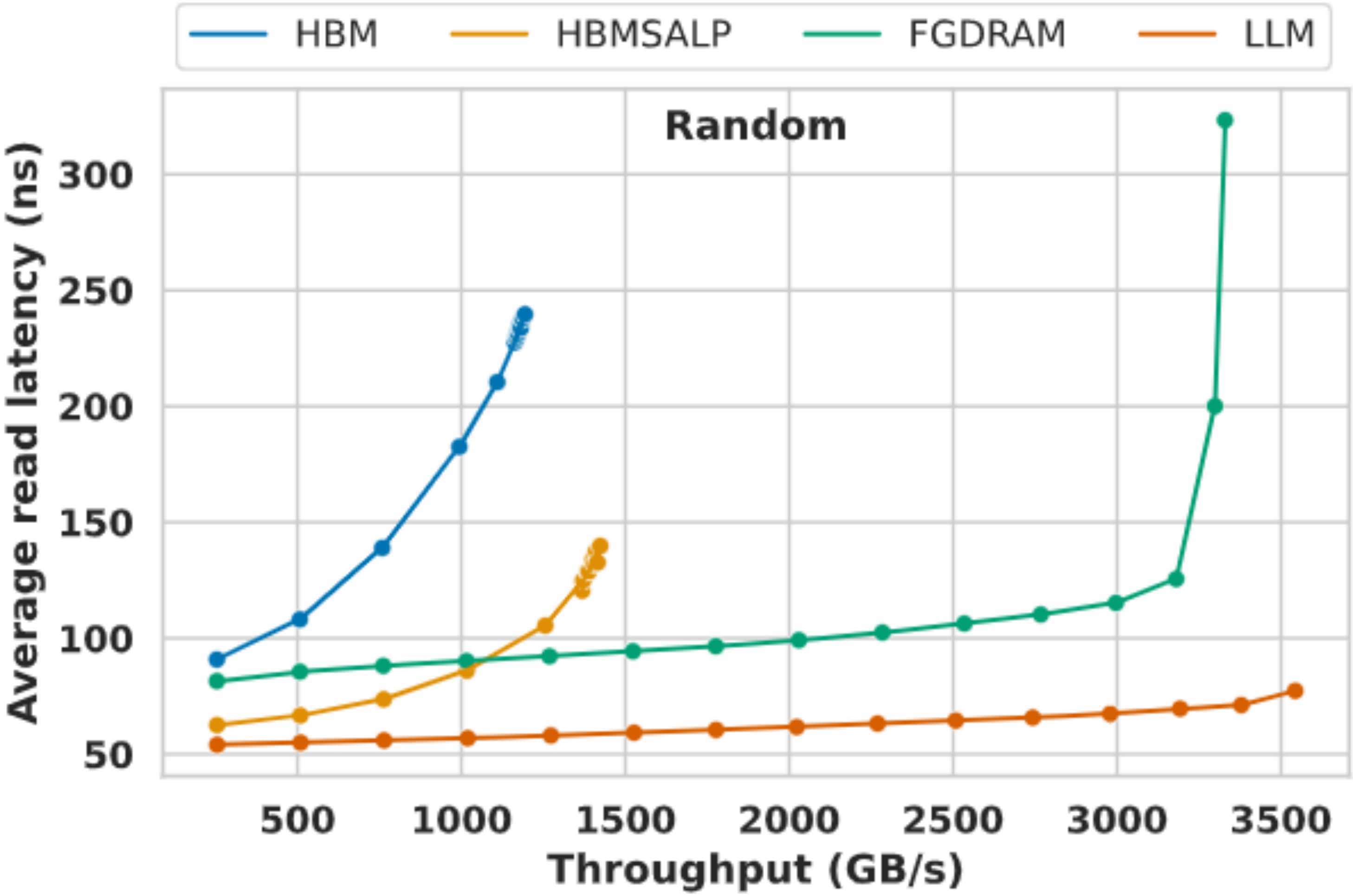
- Irregular Workloads:
  - 16 CPU cores
  - GAP Benchmark Suite (GAPBS)
  - Iso-capacity test
    - Memories have same capacity (8 channels)

Outline

- Motivation
- Background on Silicon Photonic
- LLM Architecture
- Evaluation methodology
- Evaluation results
  - Synthetic workload
  - Irregular workload
- Conclusion

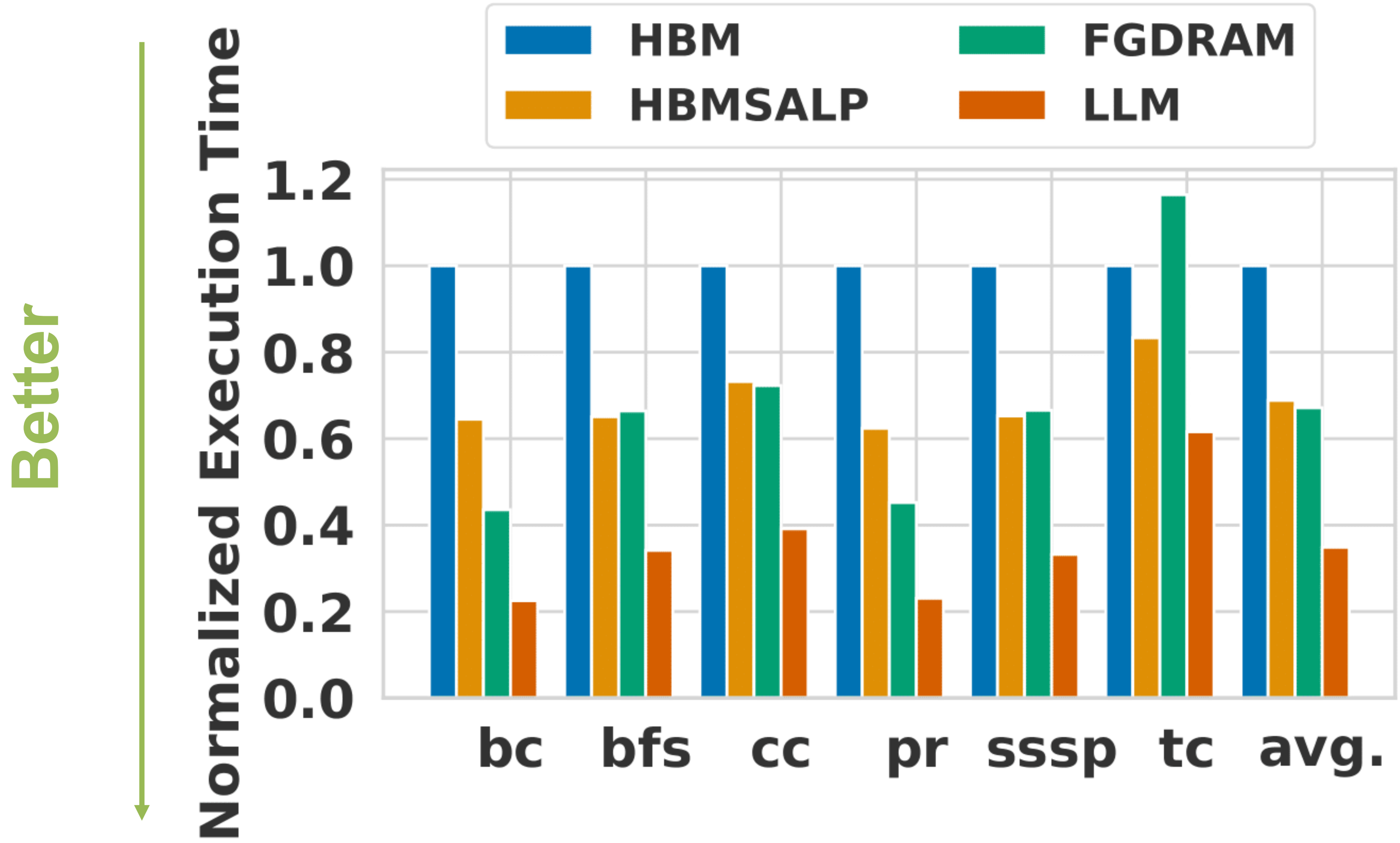
Evaluation: Synthetic Workload

- Number of channels:
  - HBM(SALP) > FGDRAM > LLM

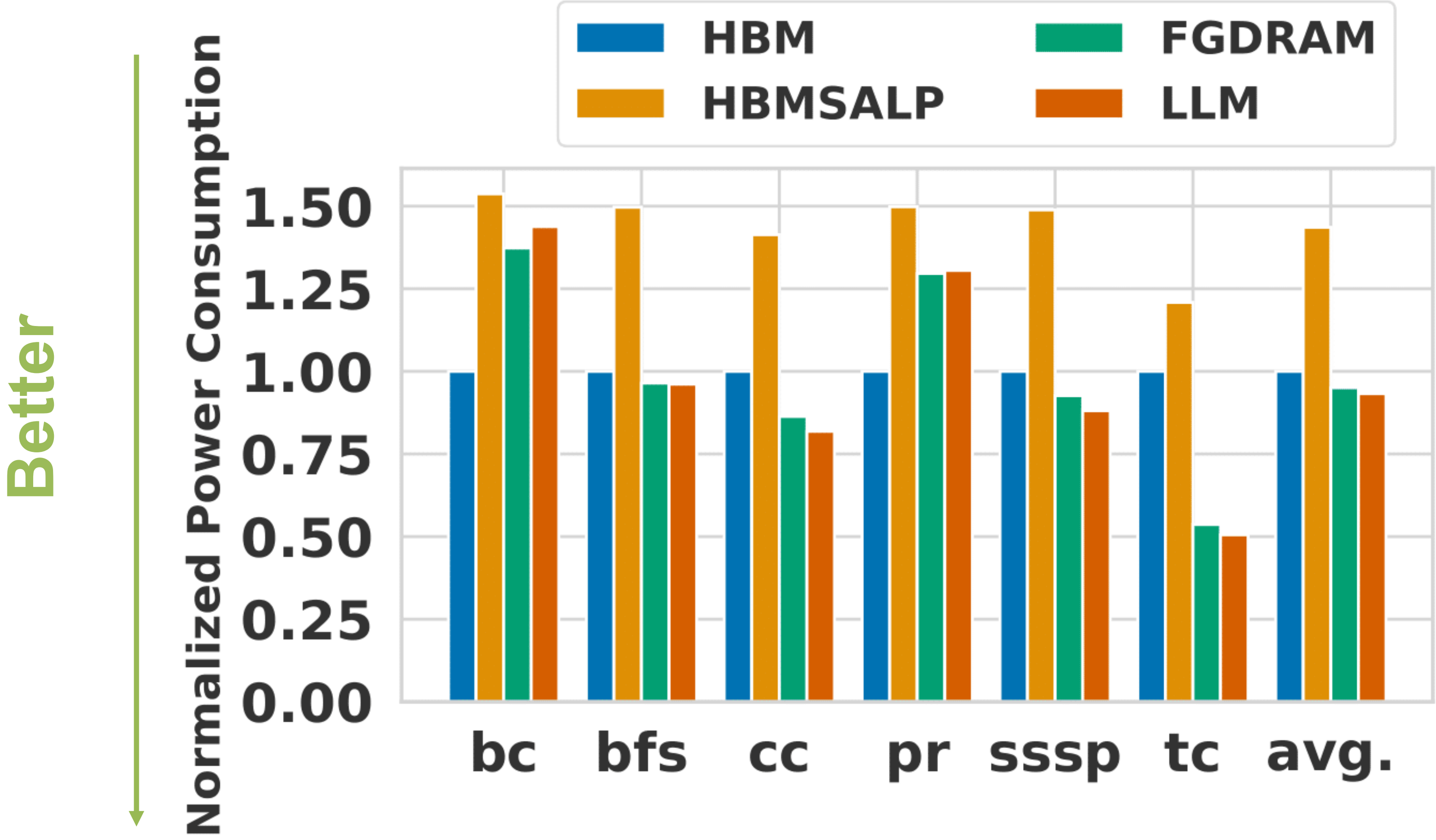


Check out our paper for more results!

Evaluation: Irregular Workloads - Execution Time



Evaluation: Irregular Workloads – Power Consumption



Outline

- Motivation
- Background on Silicon Photonic
- LLM Architecture
- Evaluation methodology
- Evaluation results
- Conclusion

Key Takeaways

- LLM proposes an end-to-end co-design that removes the contention on the data path.
- It proposes a new memory system optimized for applications with irregular access patterns.
- The use of optical links provide better data movement energy and higher bandwidth/mm^2.
- LLM achieves around 3× better execution time while maintaining the same power consumption as HBM2.0.
- Future Work:
  - Exploring the benefits of using LLM in graph accelerators
  - Evaluate the performance for other irregular/regular workloads

Thank You