Optically Connected Memory for Disaggregated Data Centers

Jorge Gonzalez[1] Alexander Gazman[2] Maarten Hattink[2] Mauricio G.Palma[1]
Meisam Bahadori[3] Ruth Rubio-Noriega[4] Lois Orosa[5]
Madeleine Glick[2] Onur Mutlu[5] Keren Bergman[2] Rodolfo Azevedo[1]

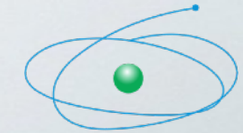
1

2

3

4

5



Executive Summary

- **Motivation:** Off-chip memory bandwidth is limited to short distances, being challenging to disaggregate the main memory.
- **Problem:** Current memory interconnects does not scale for disaggregation in datacenters.
- **Contributions:**
  - New optical point-to-point disaggregated main memory system for current DDR standards.
  - Study how a processor interacts with the disaggregated memory subsystem.
  - Evaluates a SiP link with state-of-the-art optical devices.
- **Results:**
  - OCM is 5.5x faster than 40G NIC-based disaggregated memory.
  - OCM has10.7% energy overhead compared to the DDR DRAM energy consumption for data movement.
- **Conclusion:** OCM is a promising step towards future data centers with disaggregated main memory.

Outline

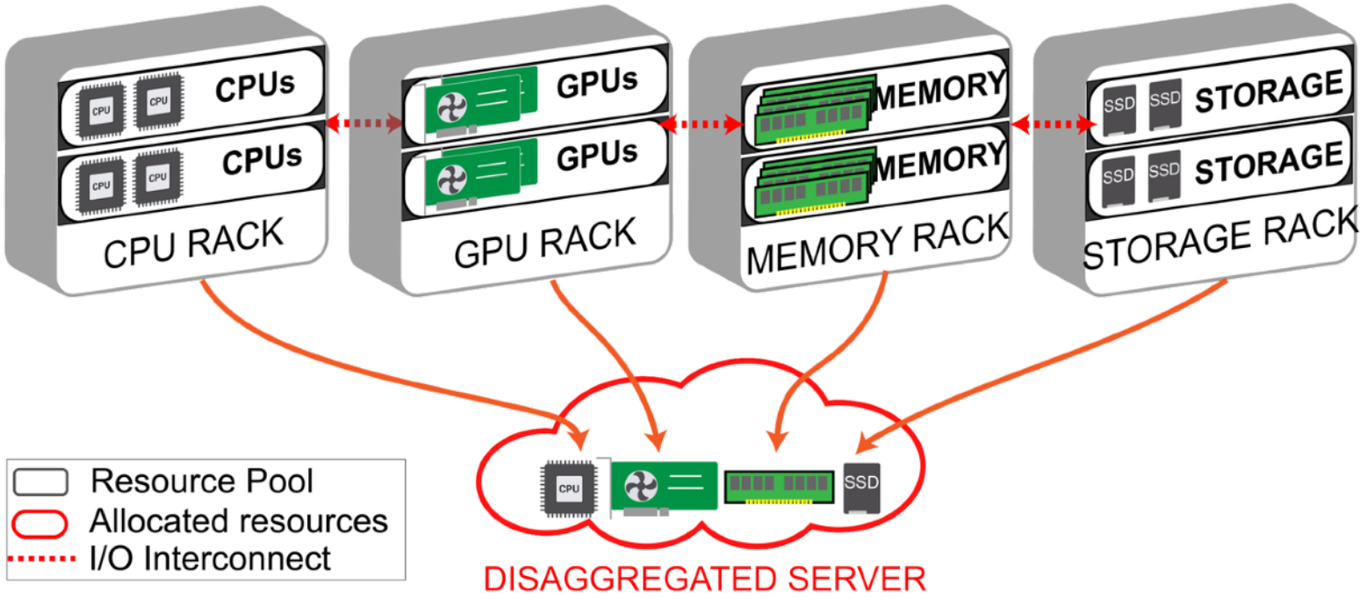
- Introduction
- Background
- Motivation and Goal
- Optically Connected Memory (OCM)
- Evaluation
- Conclusion

Outline

Disaggregated Data Centers

- **Growing gap** between computing power and communication.
  - #1 Top500 2019 (Summit) has ratio communication/computing (Bytes/FLOP) **8X lower than #1 Top500 2017**.
  - Data Centers **maintain most data inside the node.**
- **Improve performance** by increasing the available resources.
- **Underutilization** of resources still occurs.

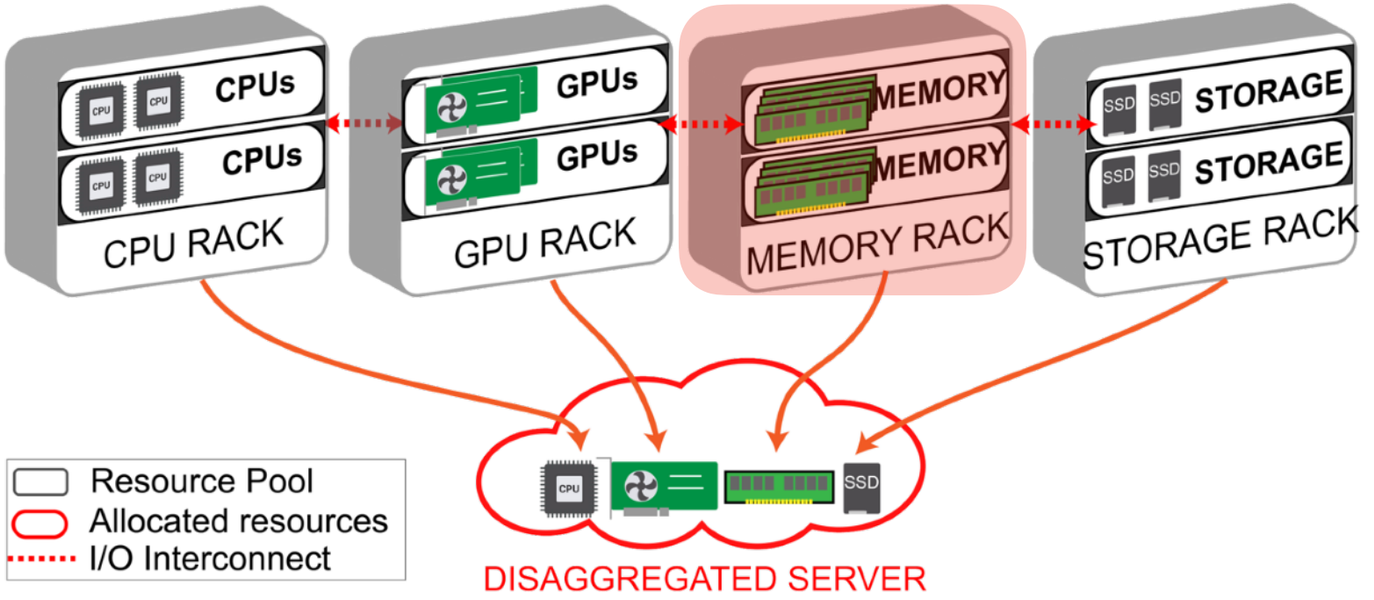
- **Disaggregated systems:**
  - **Approach**: a network of resources, rather than a network of servers.
  - **Efficient allocation** of resources
  - Memory is **a critical resource.**
  - **High-bandwidth** interconnection at **rack distance** (<10m).



Disaggregated Data Centers

- **Growing gap** between computing power and communication.
  - #1 Top500 2019 (Summit) has ratio communication/computing (Bytes/FLOP) **8X lower than #1 Top500 2017**.
  - Data Centers **maintain most data inside the node.**
- **Improve performance** by increasing the available resources.
- **Underutilization** of resources still occurs.

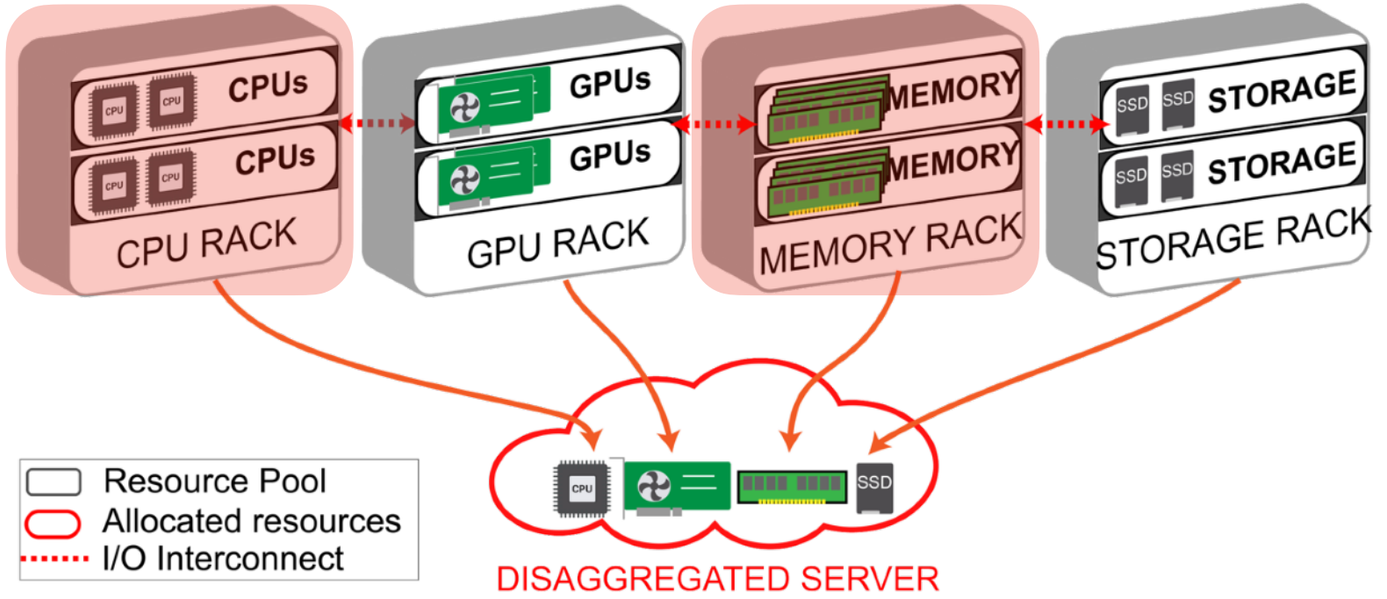
- **Disaggregated systems:**
  - **Approach**: a network of resources, rather than a network of servers.
  - **Efficient allocation** of resources
  - Memory is **a critical resource.**
  - **High-bandwidth** interconnection at **rack distance** (<10m).



Disaggregated Data Centers

- **Growing gap** between computing power and communication.
  - #1 Top500 2019 (Summit) has ratio communication/computing (Bytes/FLOP) **8X lower than #1 Top500 2017**.
  - Data Centers **maintain most data inside the node.**
- **Improve performance** by increasing the available resources.
- **Underutilization** of resources still occurs.

- **Disaggregated systems:**
  - **Approach**: a network of resources, rather than a network of servers.
  - **Efficient allocation** of resources
  - Memory is **a critical resource.**
  - **High-bandwidth** interconnection at **rack distance** (<10m).



Outline

Introduction

Background

Motivation and Goal

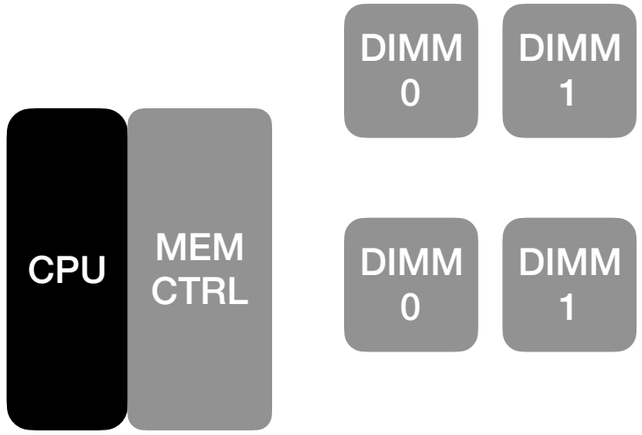
Optically Connected Memory (OCM)

Evaluation

Conclusion

Electrical Memory Interface

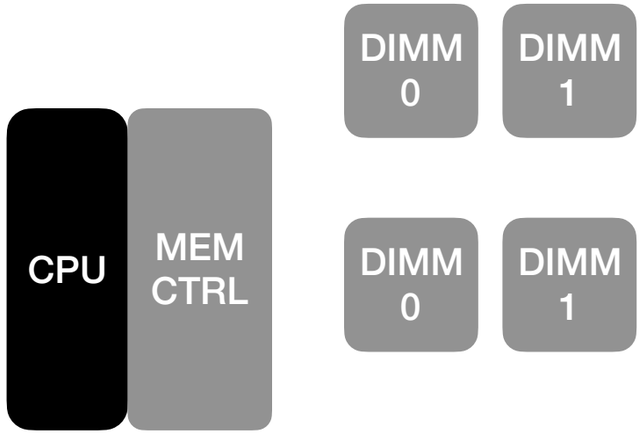
Single Channel



Dual Channel

Electrical Memory Interface

Single Channel

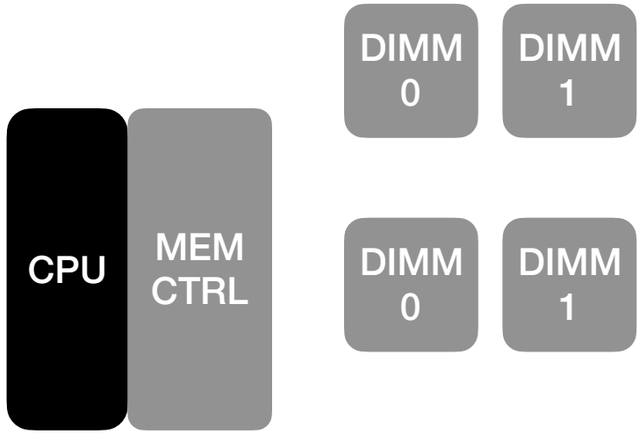


Dual Channel

- **Single channel:**

Electrical Memory Interface

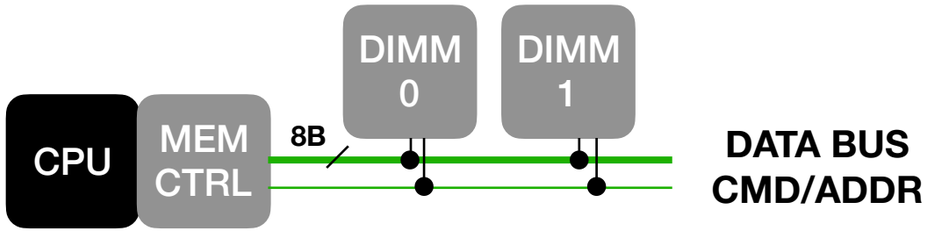
Single Channel



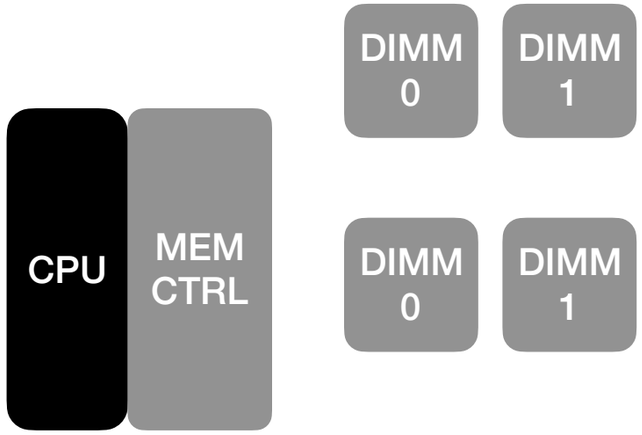
Dual Channel

- **Single channel:**
  - Shared data bus.

Electrical Memory Interface



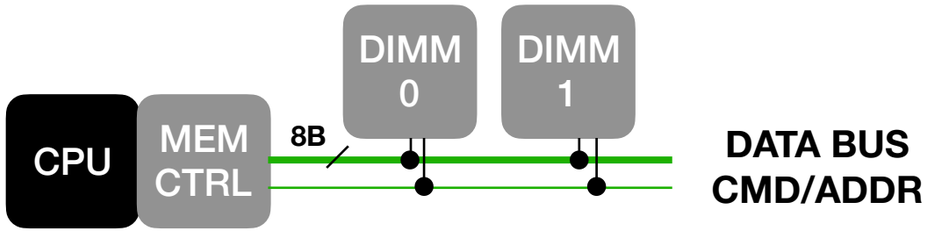
Single Channel



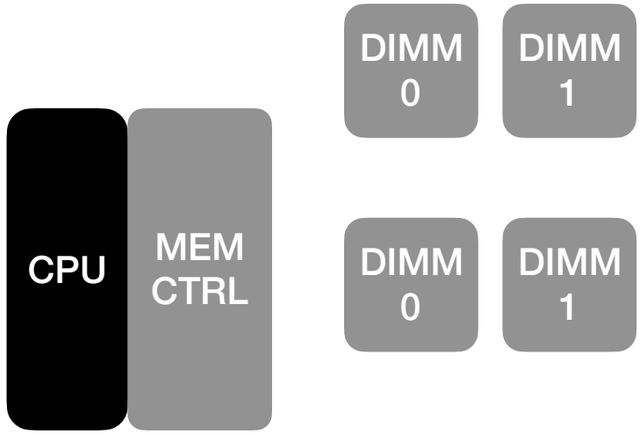
Dual Channel

- **Single channel:**
  - Shared data bus.

Electrical Memory Interface



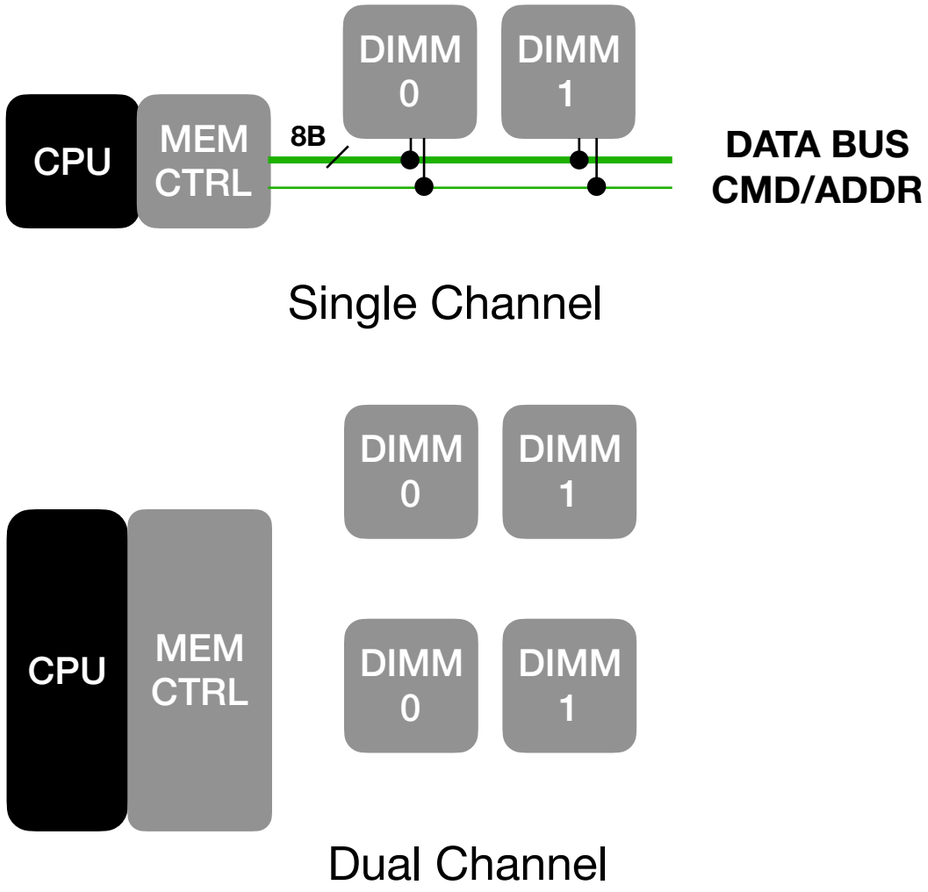
Single Channel



Dual Channel

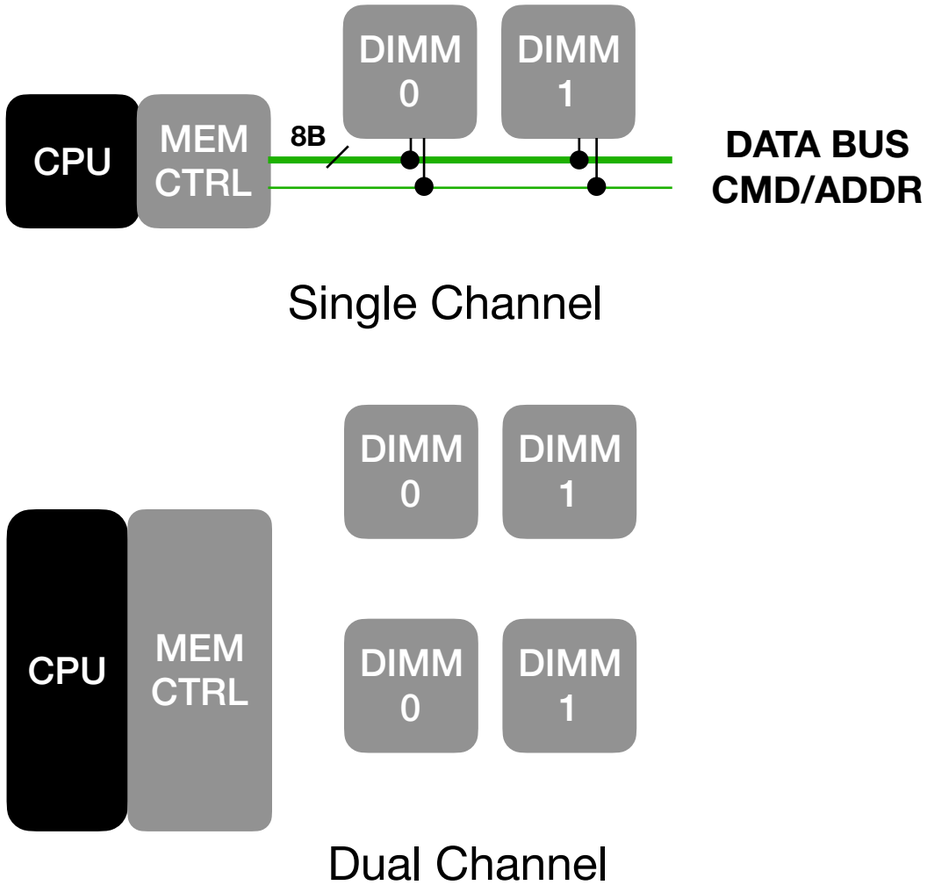
- **Single channel:**
  - Shared data bus.
  - Only 1 DRAM DIMM can be accessed.

Electrical Memory Interface



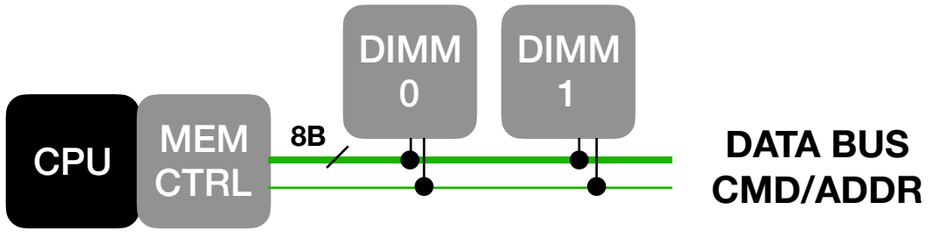
- **Single channel:**
  - Shared data bus.
  - Only 1 DRAM DIMM can be accessed.
  - 1 memory access = 8 transactions of 8B (e.g.: 64B cache line).

Electrical Memory Interface

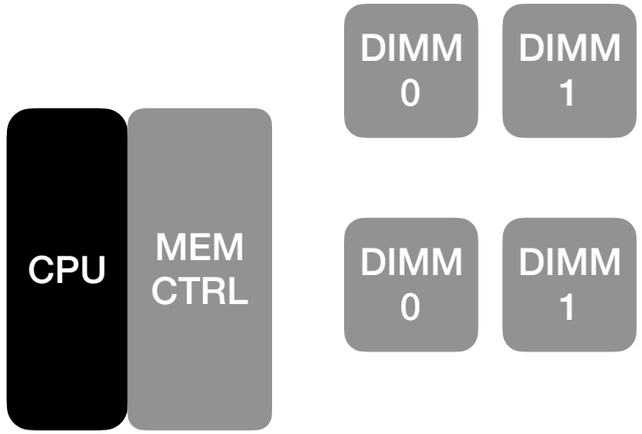


- **Single channel:**
  - Shared data bus.
  - Only 1 DRAM DIMM can be accessed.
  - 1 memory access = 8 transactions of 8B (e.g.: 64B cache line).
- **Multichannel:**

Electrical Memory Interface



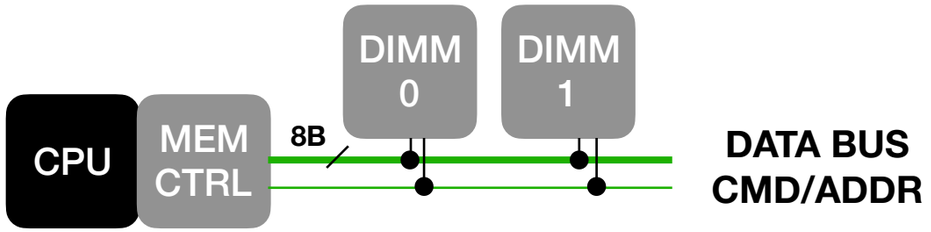
Single Channel



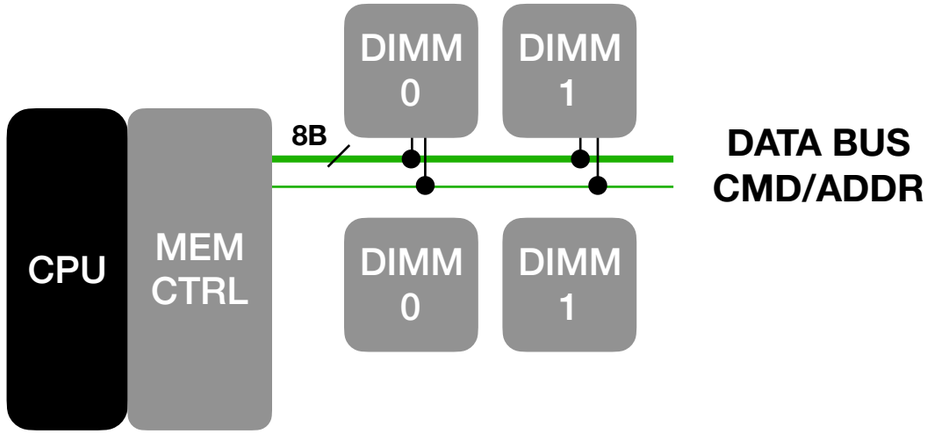
Dual Channel

- **Single channel:**
  - Shared data bus.
  - Only 1 DRAM DIMM can be accessed.
  - 1 memory access = 8 transactions of 8B (e.g.: 64B cache line).
- **Multichannel:**
  - Add physical channels to access data in parallel.

Electrical Memory Interface



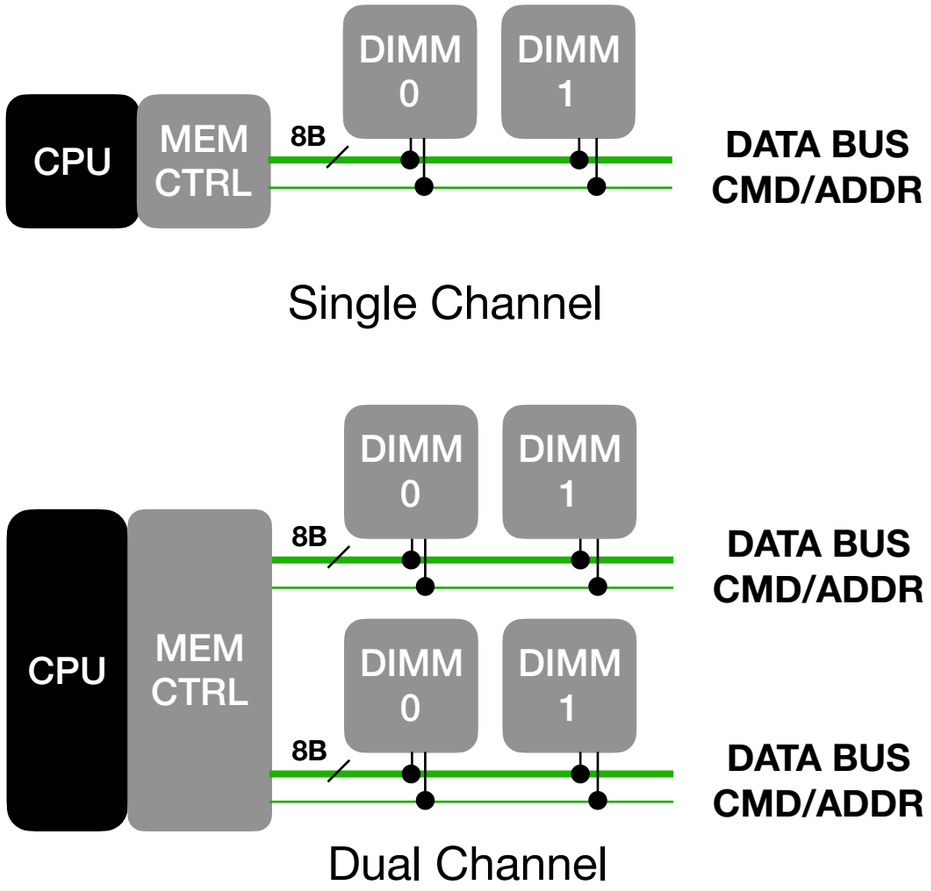
Single Channel



Dual Channel

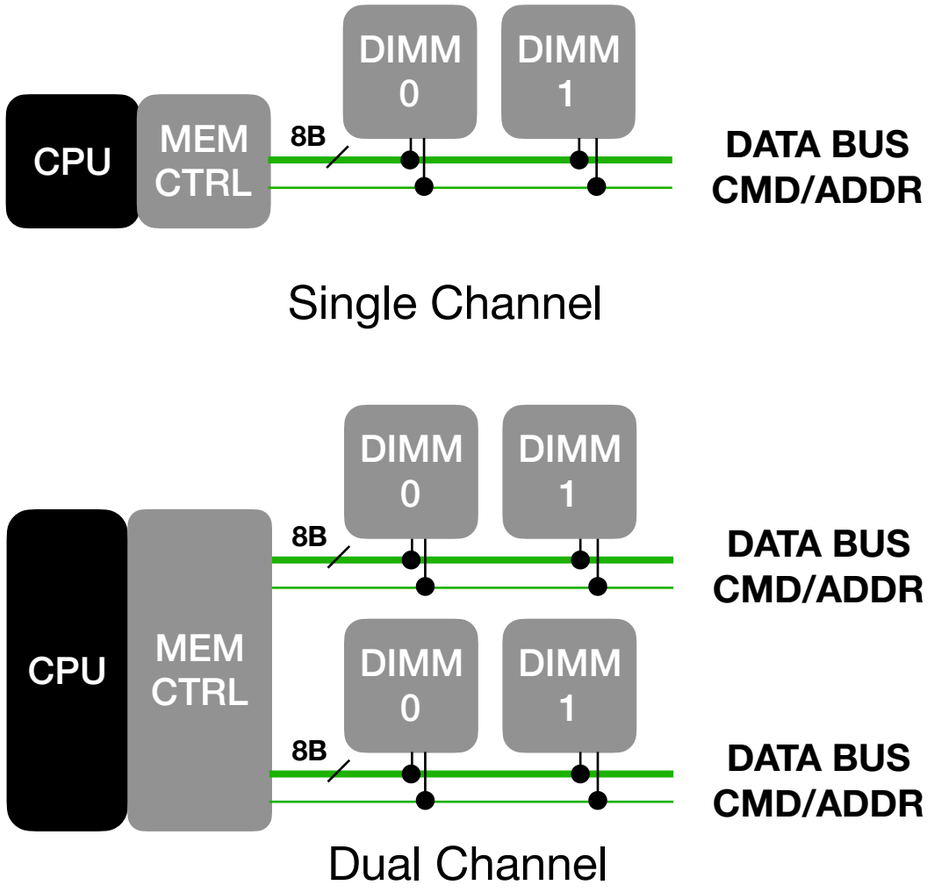
- **Single channel:**
  - Shared data bus.
  - Only 1 DRAM DIMM can be accessed.
  - 1 memory access = 8 transactions of 8B (e.g.: 64B cache line).
- **Multichannel:**
  - Add physical channels to access data in parallel.

Electrical Memory Interface



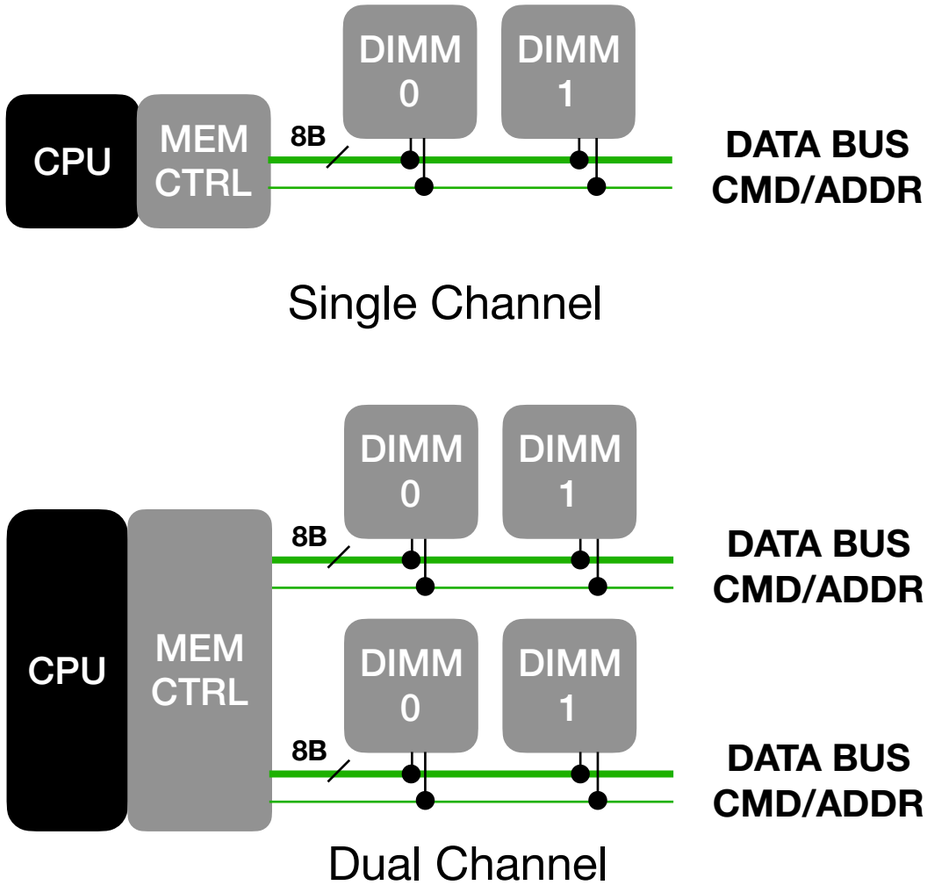
- **Single channel:**
  - Shared data bus.
  - Only 1 DRAM DIMM can be accessed.
  - 1 memory access = 8 transactions of 8B (e.g.: 64B cache line).
- **Multichannel:**
  - Add physical channels to access data in parallel.

Electrical Memory Interface



- **Single channel:**
  - Shared data bus.
  - Only 1 DRAM DIMM can be accessed.
  - 1 memory access = 8 transactions of 8B (e.g.: 64B cache line).
- **Multichannel:**
  - Add physical channels to access data in parallel.
- **Electrical constraints:** wiring, pins and short distances (few cm).

Electrical Memory Interface



- **Single channel:**
  - Shared data bus.
  - Only 1 DRAM DIMM can be accessed.
  - 1 memory access = 8 transactions of 8B (e.g.: 64B cache line).
- **Multichannel:**
  - Add physical channels to access data in parallel.
- **Electrical constraints:** wiring, pins and short distances (few cm).

How can we **disaggregate main memory**?

Outline

- Introduction
- Background
- **Motivation and Goal**
- Optically Connected Memory (OCM)
- Evaluation
- Conclusion

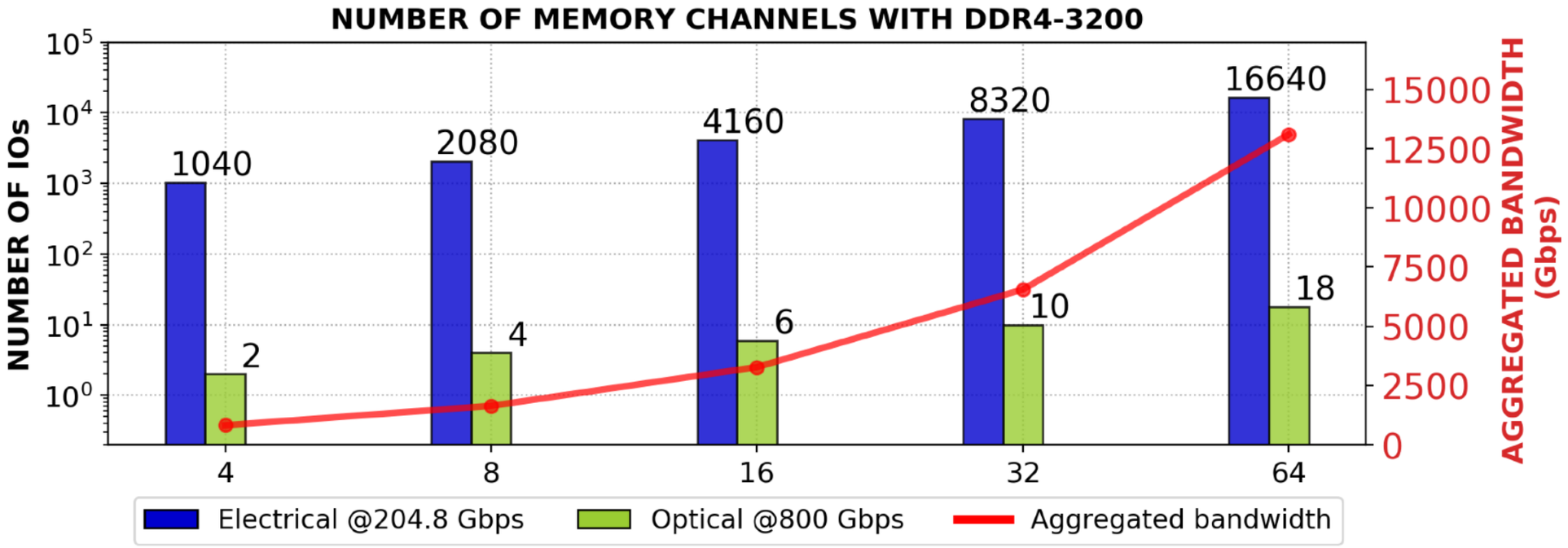
Motivation: Approaching the Memory Wall

Motivation: Approaching the Memory Wall

- Memory off-chip bandwidth **is limited.**

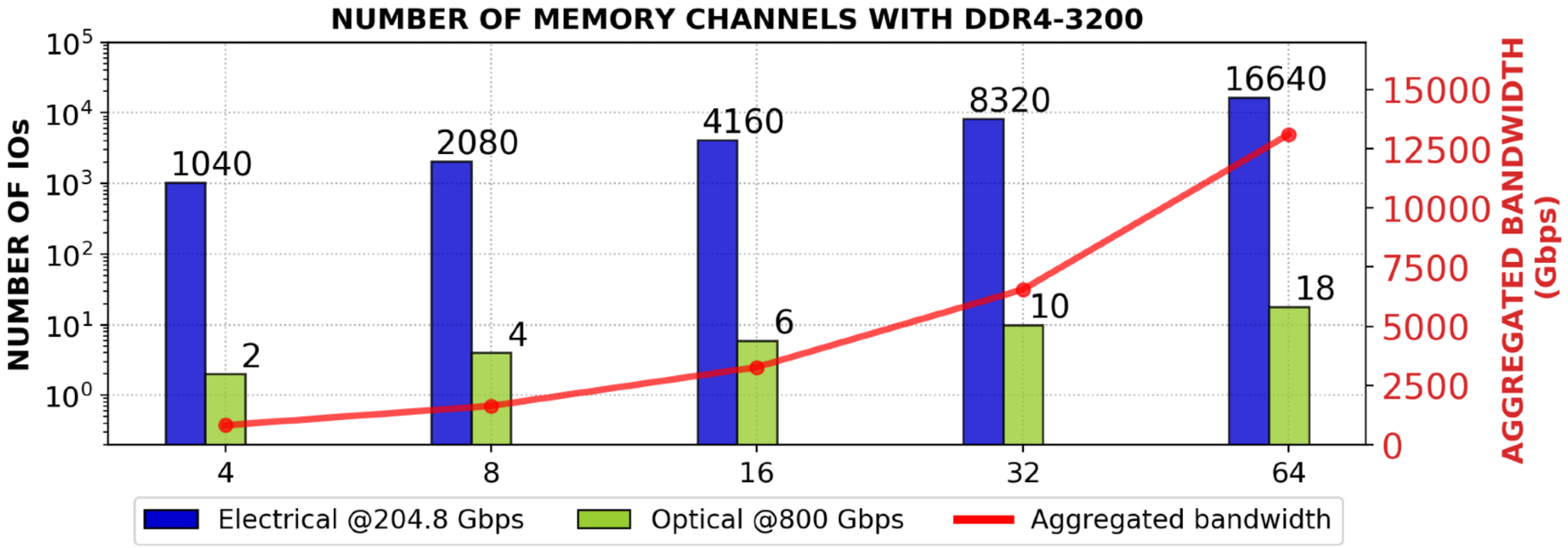
Motivation: Approaching the Memory Wall

- Memory off-chip bandwidth **is limited.**
- **Photonics characteristics:** a) high-bandwidth, b) low-energy, c) distance independent.



Motivation: Approaching the Memory Wall

- Memory off-chip bandwidth **is limited.**
- **Photonics characteristics:** a) high-bandwidth, b) low-energy, c) distance independent.



We can **scale** the number of **memory channels with photonics.**

Motivation: Photonics for Memory Disaggregation

- Prior works show that photonics is a promising solution for scalability **[Bahadori+, JLT'16], [Anderson+ OFC'18], [Brunina+, JSTQE'13]**

- **No comprehensive analysis** that evaluates **both**:

  - **How a processor interacts** with a disaggregated memory subsystem executing real applications.

  - SiP link design to **estimate the number of optical devices and energy consumption** for DDR standards.

Goal

Propose an **Optically Connected Memory (OCM) architecture** and study how photonics can enable disaggregation for memory systems.

1. Study **how a processor interacts** with a disaggregated memory subsystem while executing real applications.

2. Design and evaluate a SiP link to **estimate the number of optical devices and energy consumption** for current DDR standards.

Outline

Introduction

Background

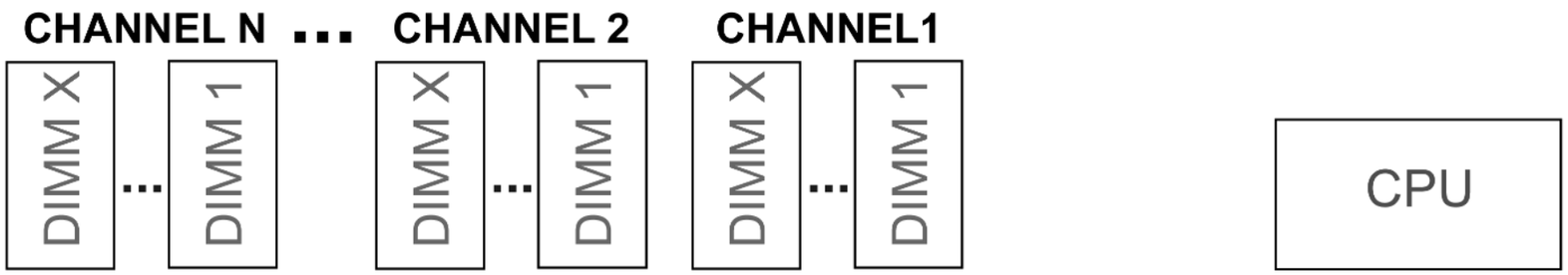
Motivation and Goal

Optically Connected Memory (OCM)

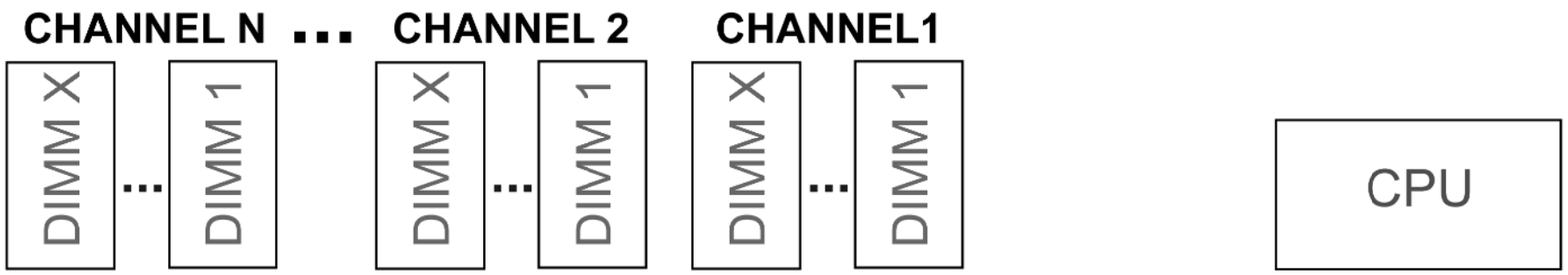
Evaluation

Conclusion

Optically Connected Memory (OCM)

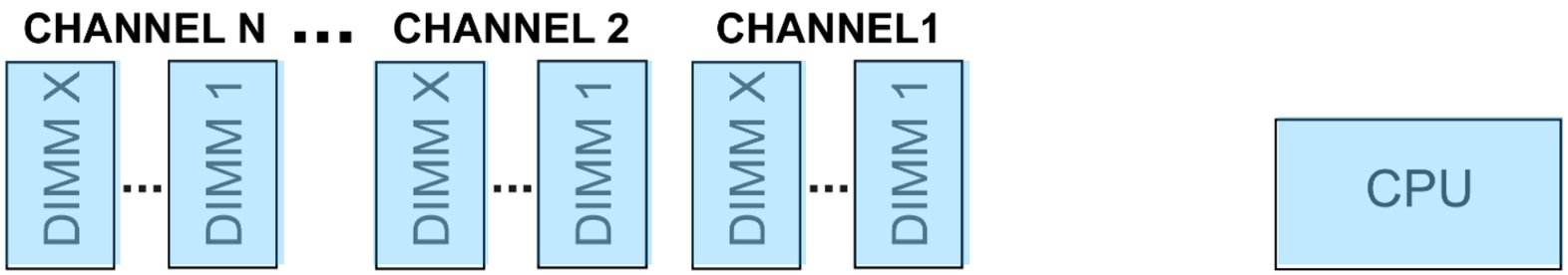


Optically Connected Memory (OCM)



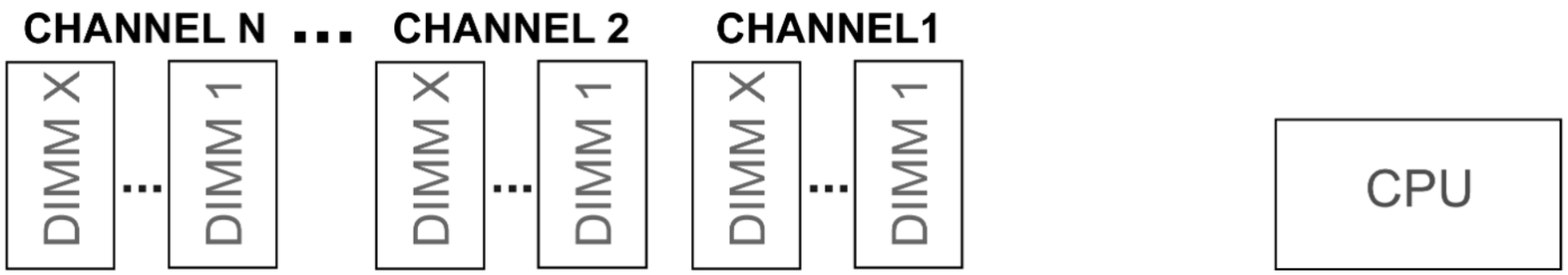
- OCM connects a **processor to N-memory channels**.
- Each channel has a set of DRAM DIMMs, e.g.: **two DRAM DIMMs per channel.**

Optically Connected Memory (OCM)



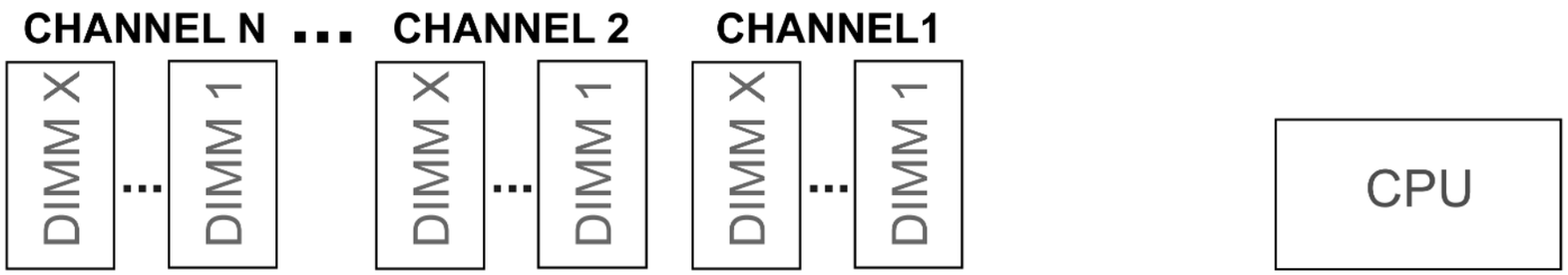
- OCM connects a **processor to N-memory channels**.
- Each channel has a set of DRAM DIMMs, e.g.: **two DRAM DIMMs per channel.**

Optically Connected Memory (OCM)

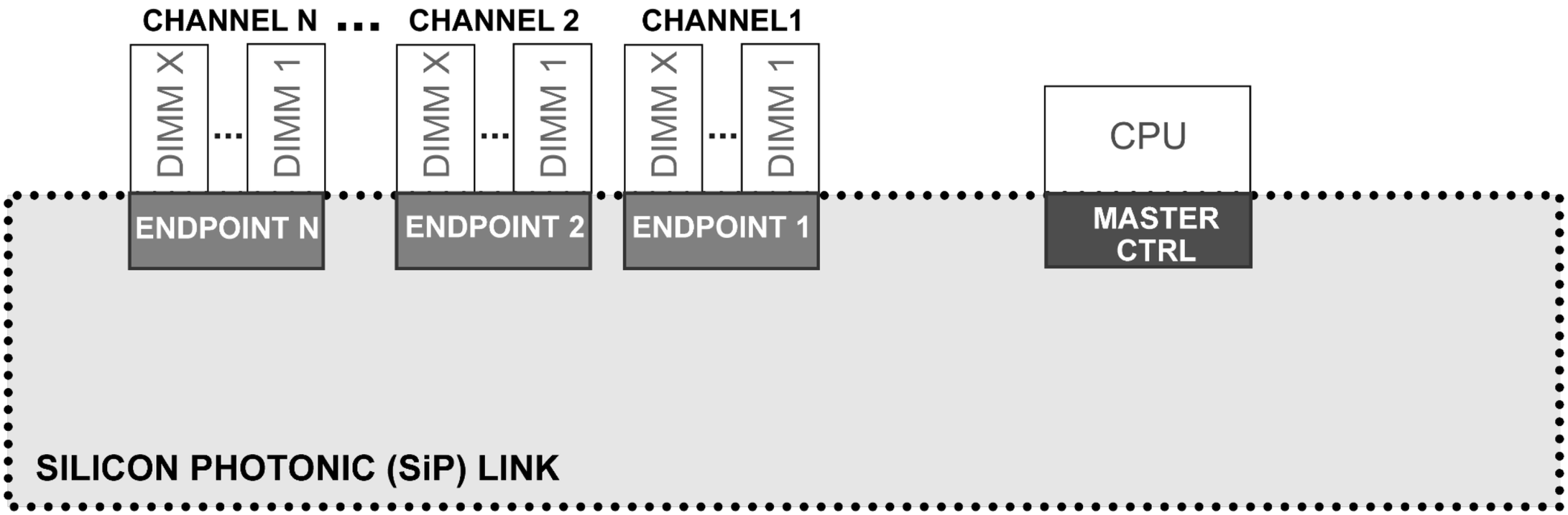


- OCM connects a **processor to N-memory channels**.
- Each channel has a set of DRAM DIMMs, e.g.: **two DRAM DIMMs per channel.**

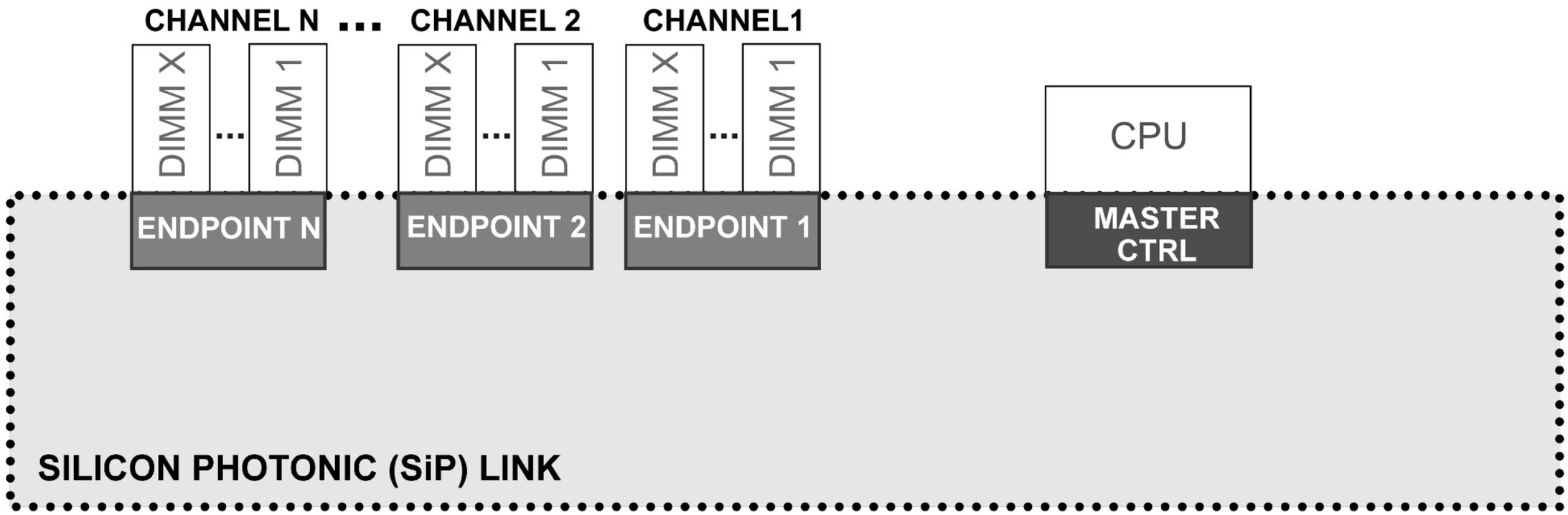
Optically Connected Memory (OCM)



Optically Connected Memory (OCM)

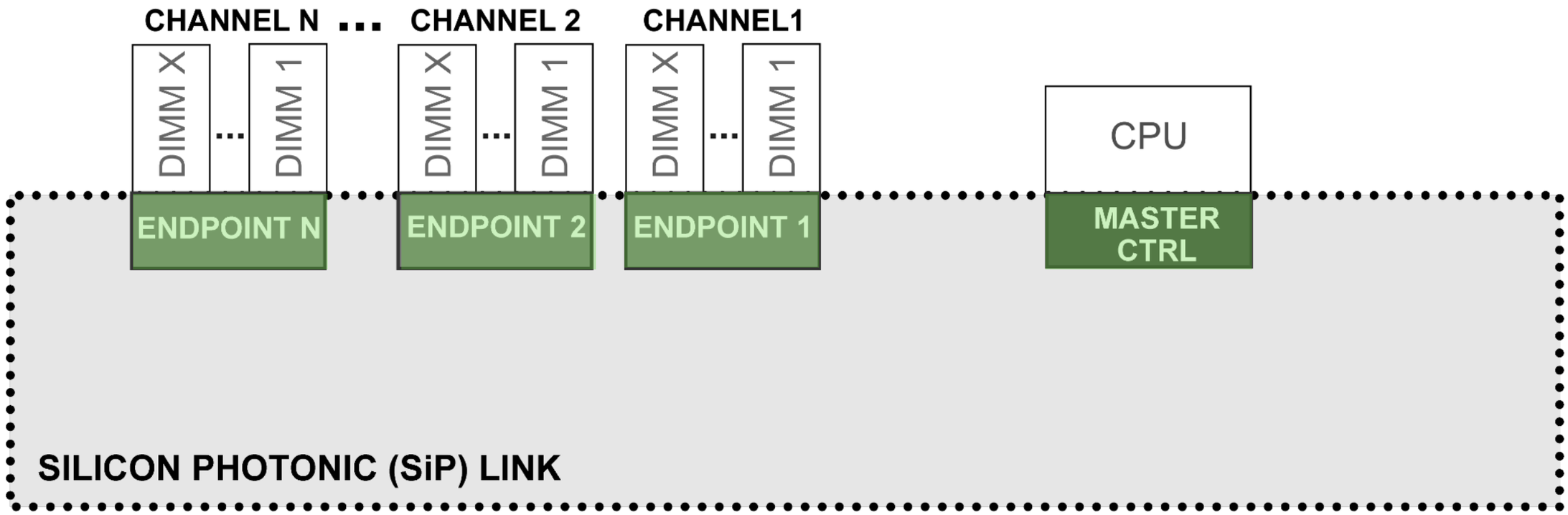


Optically Connected Memory (OCM)



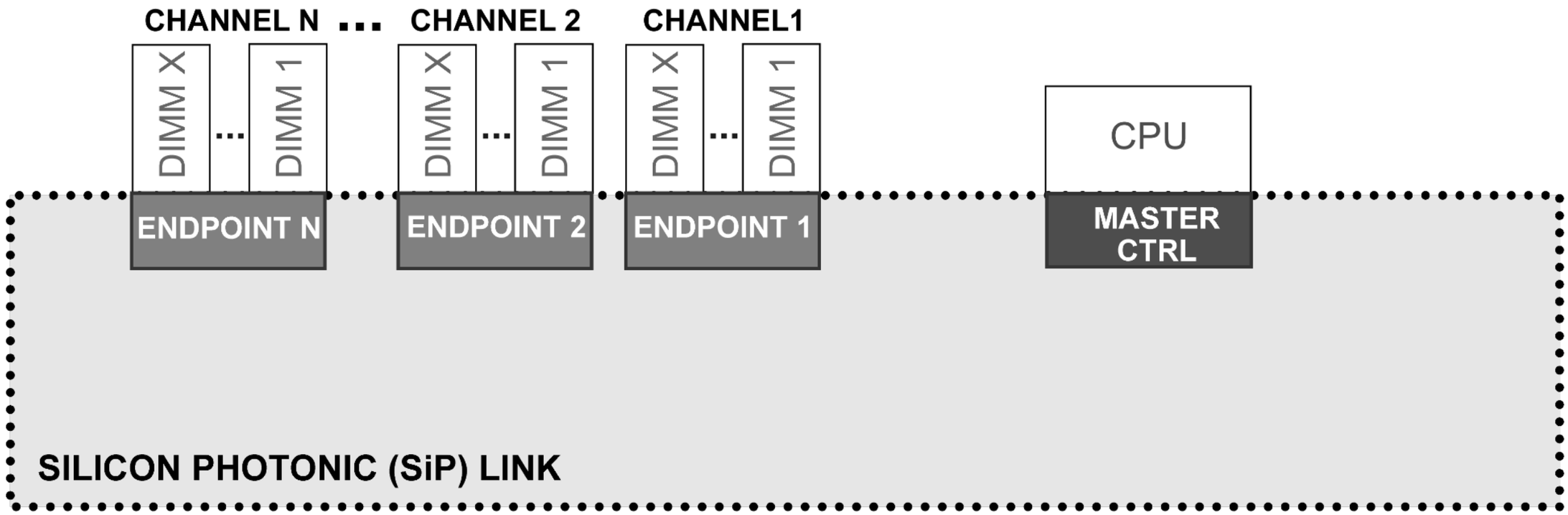
- OCM has a **SiP link** based on **state-of-the-art photonic devices.**
- It has a **master controller** on the **processor side,** and **enpoint controllers** on the **DRAM DIMM side**.
- Each controller **is based on Microring Resonators (MRRs)**

Optically Connected Memory (OCM)



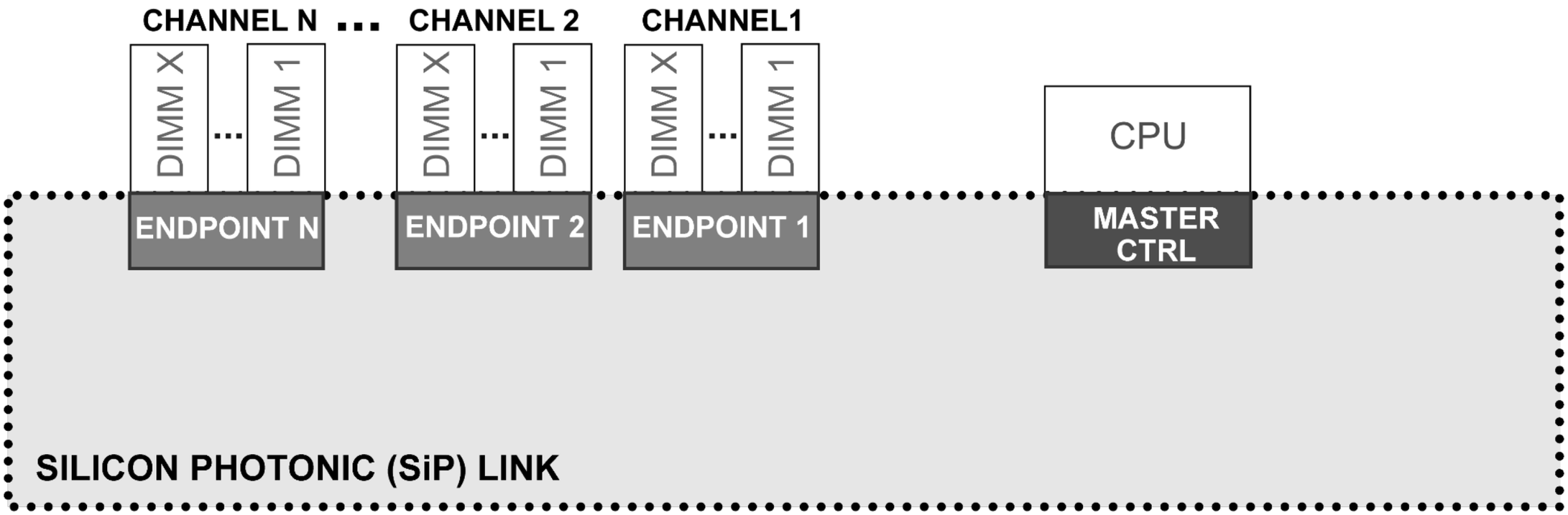
- OCM has a **SiP link** based on **state-of-the-art photonic devices.**
- It has a **master controller** on the **processor side,** and **enpoint controllers** on the **DRAM DIMM side**.
- Each controller **is based on Microring Resonators (MRRs)**

Optically Connected Memory (OCM)

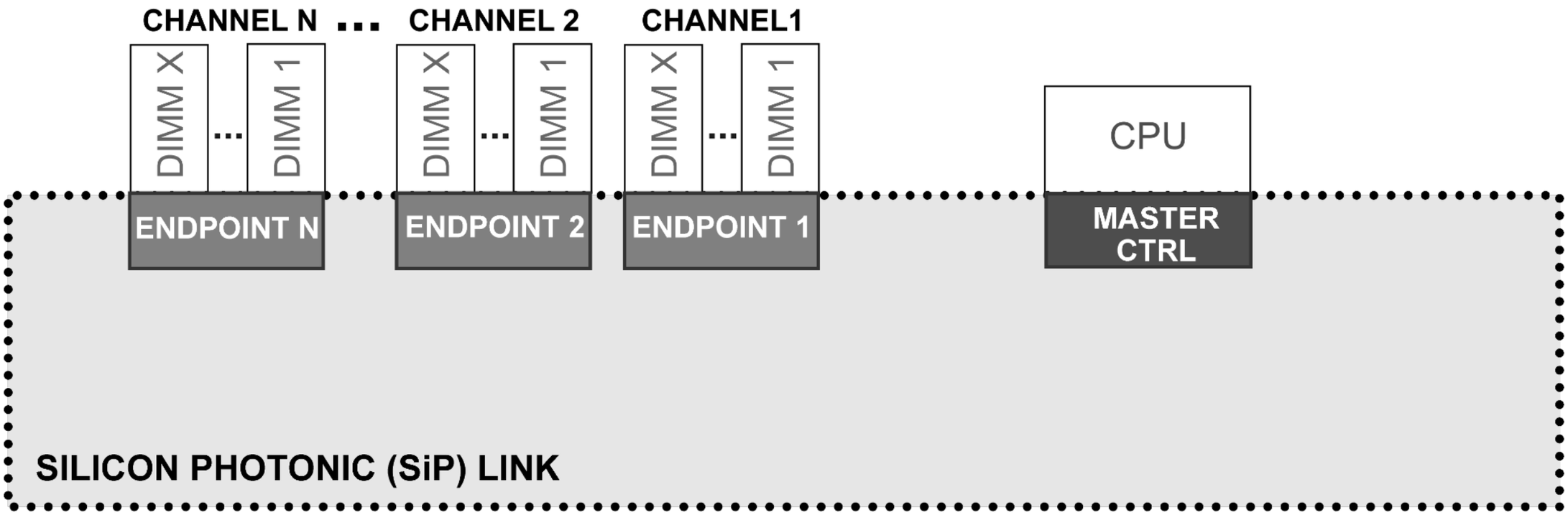


- OCM has a **SiP link** based on **state-of-the-art photonic devices.**
- It has a **master controller** on the **processor side,** and **enpoint controllers** on the **DRAM DIMM side**.
- Each controller **is based on Microring Resonators (MRRs)**

Optically Connected Memory (OCM)

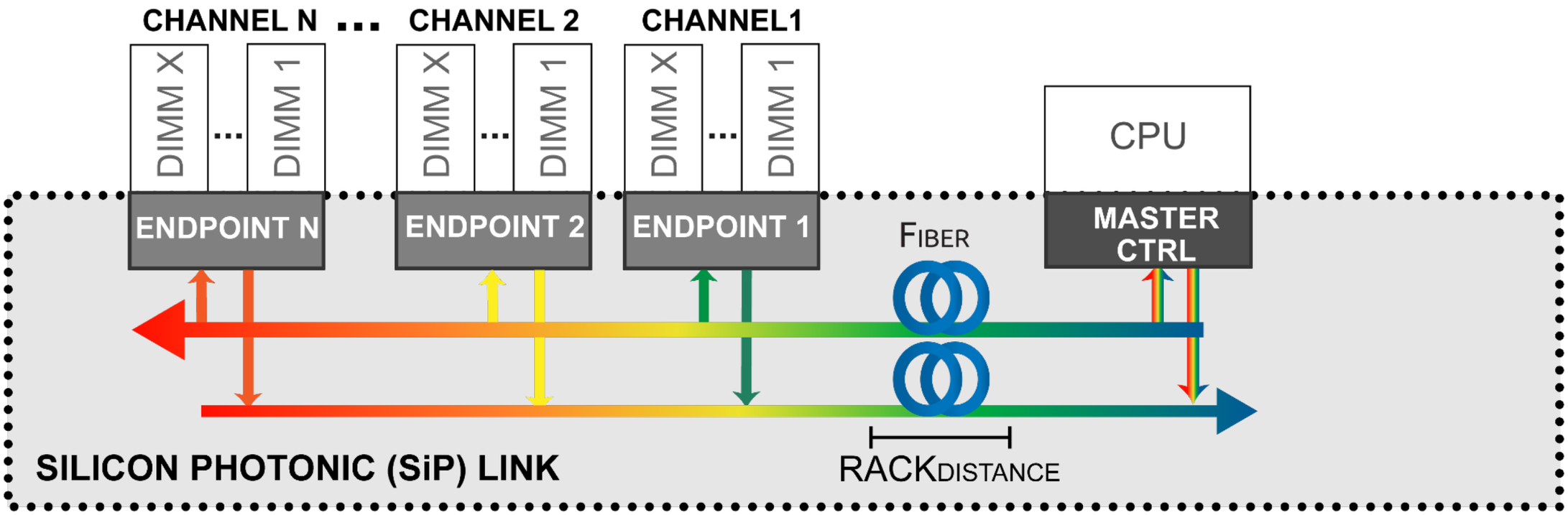


Optically Connected Memory (OCM)



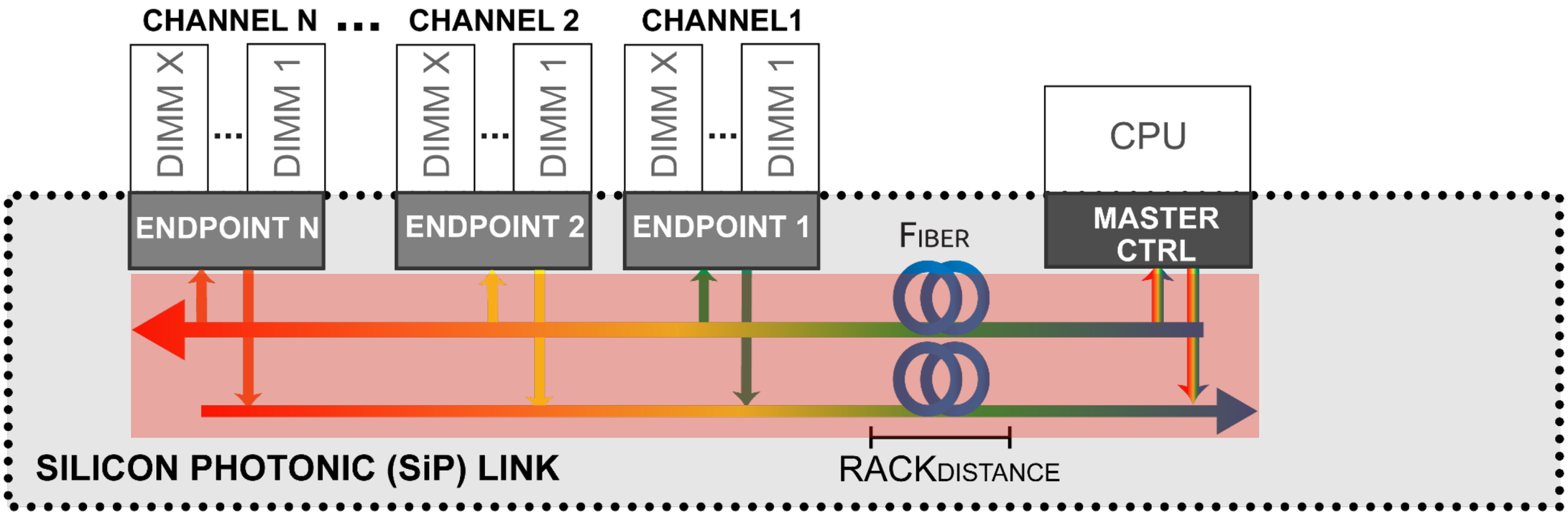
- **SiP link** is **bidirectional and composed of sublinks**, e.g.: two unidirectional sublinks.
- OCM uses a **fiber (waveguide) to connect processor and memory**, because they are placed at rack distance.

Optically Connected Memory (OCM)



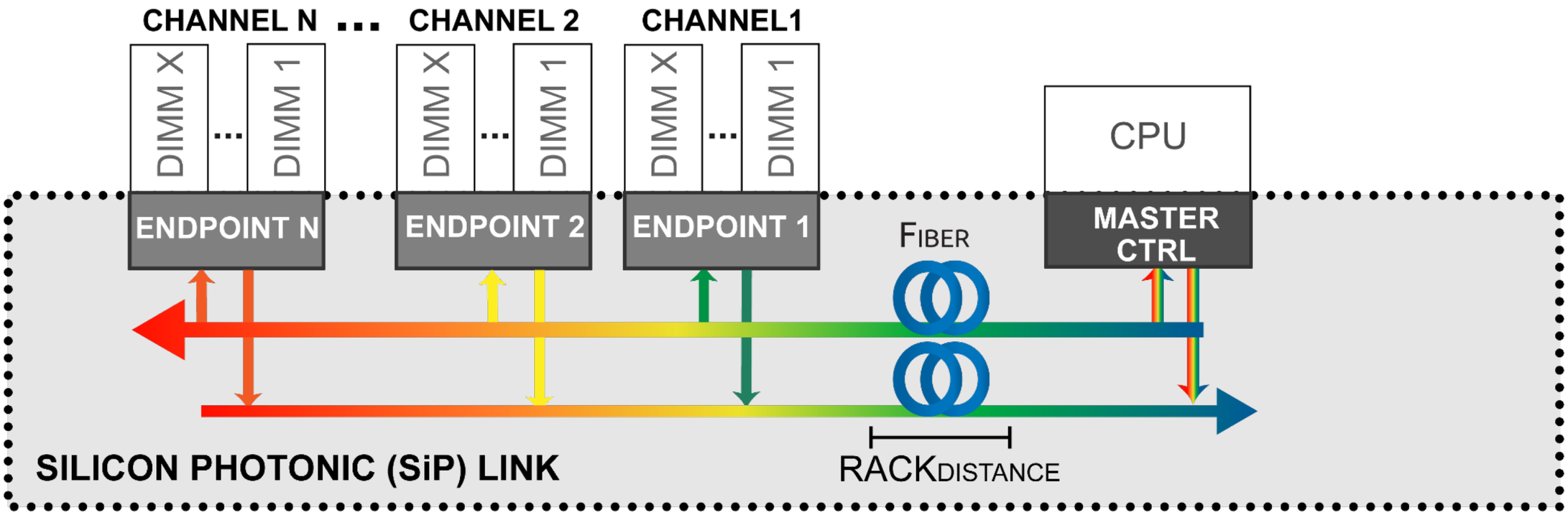
- **SiP link** is **bidirectional and composed of sublinks**, e.g.: two unidirectional sublinks.
- OCM uses a **fiber (waveguide) to connect processor and memory**, because they are placed at rack distance.

Optically Connected Memory (OCM)



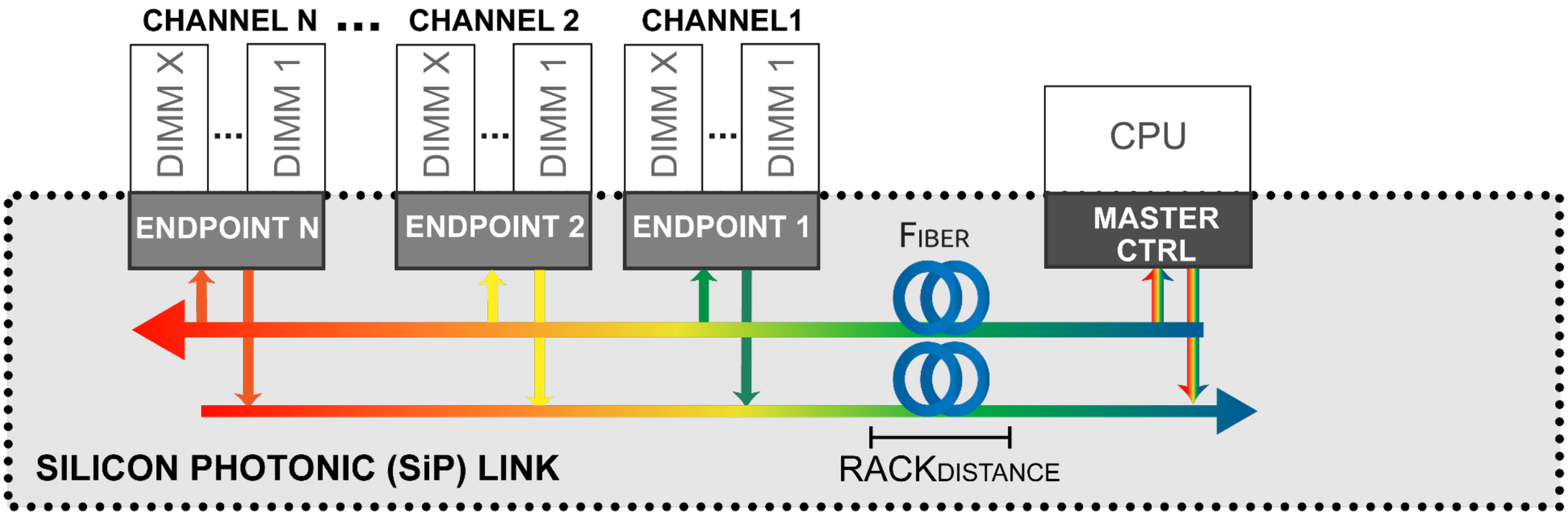
- **SiP link** is **bidirectional and composed of sublinks**, e.g.: two unidirectional sublinks.
- OCM uses a **fiber (waveguide) to connect processor and memory**, because they are placed at rack distance.

Optically Connected Memory (OCM)

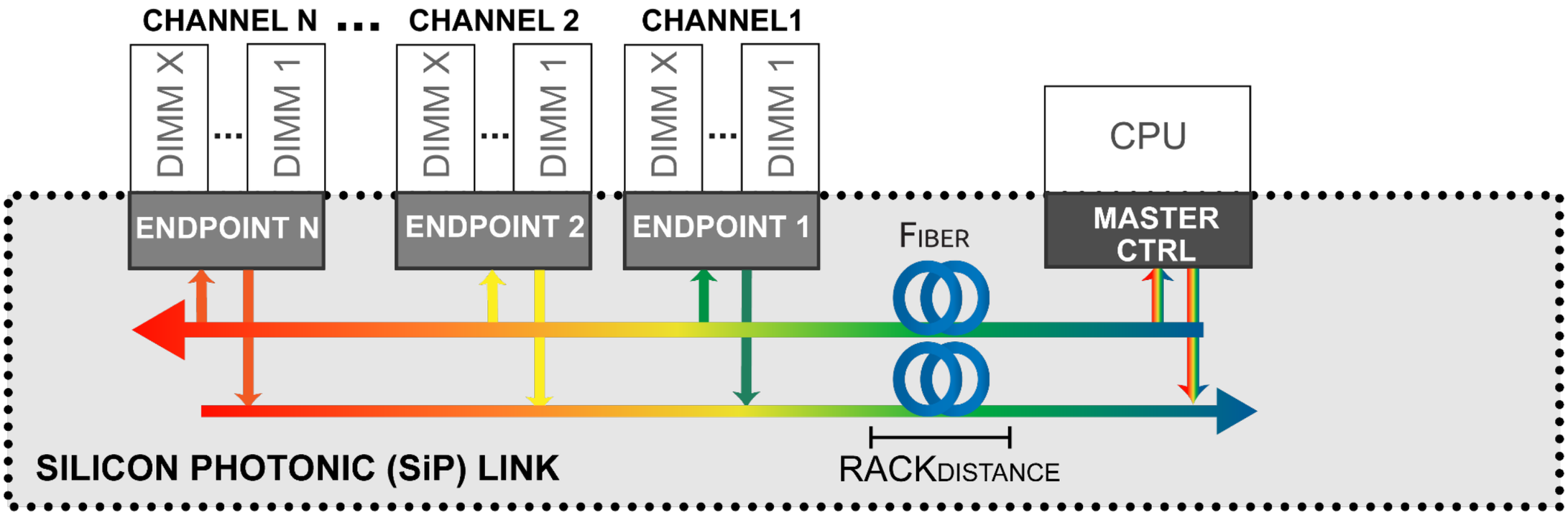


- **SiP link** is **bidirectional and composed of sublinks**, e.g.: two unidirectional sublinks.
- OCM uses a **fiber (waveguide) to connect processor and memory**, because they are placed at rack distance.

Optically Connected Memory (OCM)

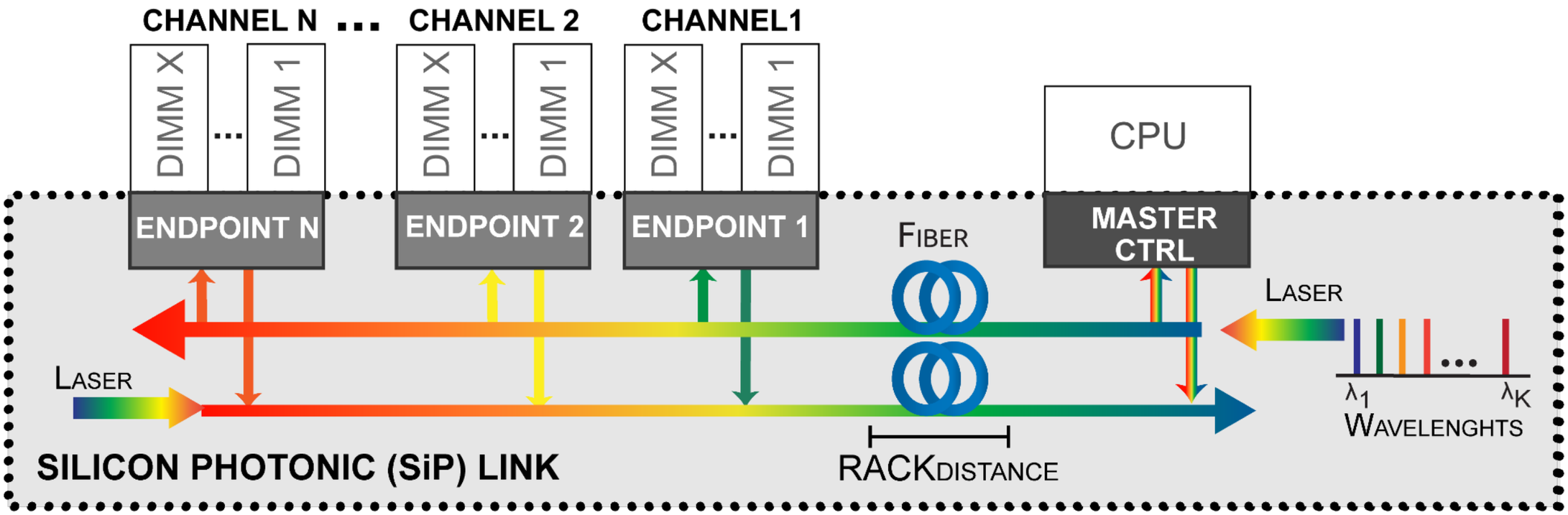


Optically Connected Memory (OCM)



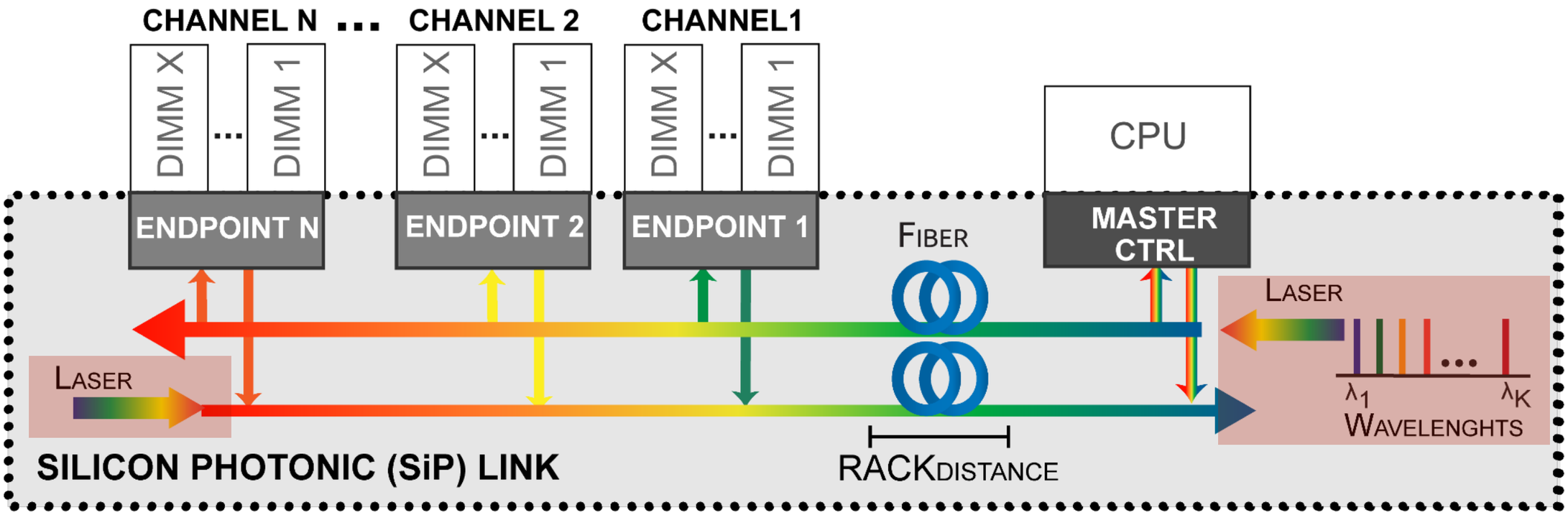
- **Laser** source **generates light in multiple wavelengths (λs)** simultaneously.

Optically Connected Memory (OCM)



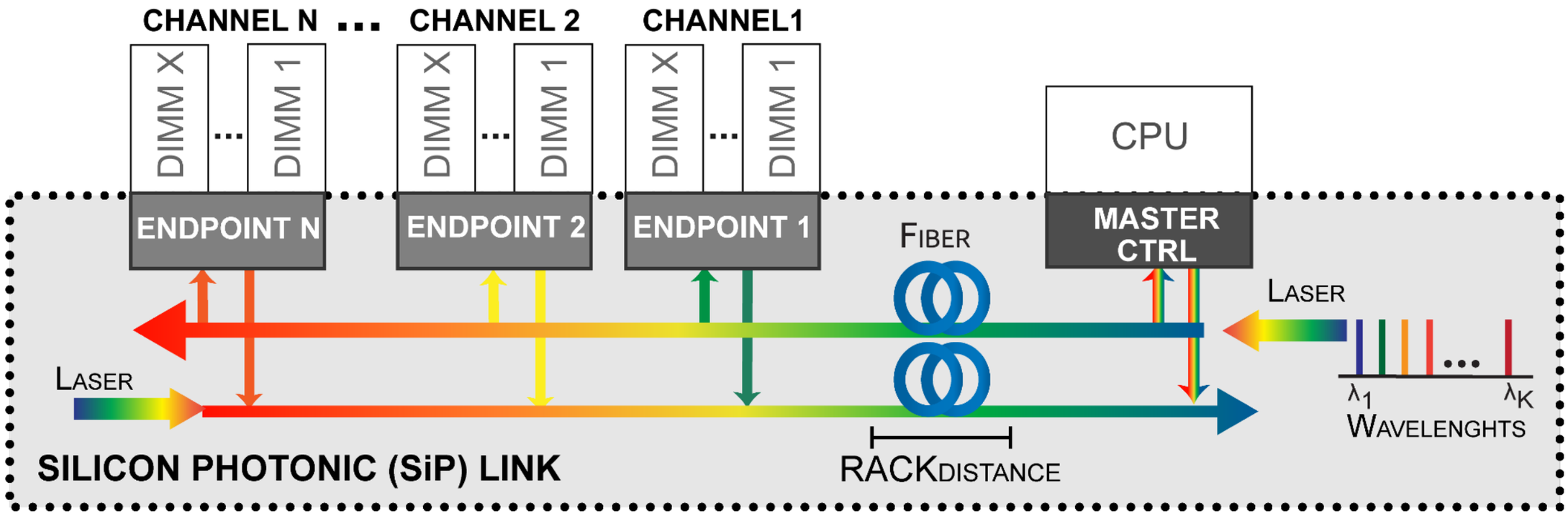
- **Laser** source **generates light in multiple wavelengths (λs)** simultaneously.

Optically Connected Memory (OCM)



- **Laser** source **generates light in multiple wavelengths (λs)** simultaneously.

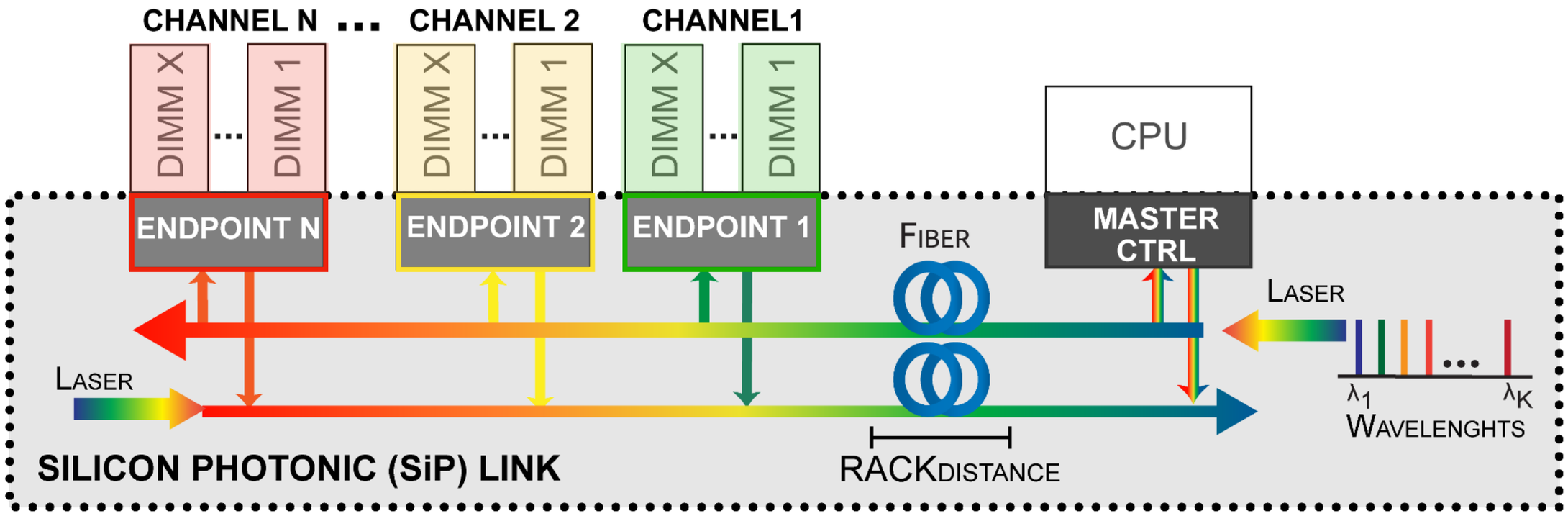
Optically Connected Memory (OCM)



OCM uses **DWDM** to achieve **high aggregated bandwidth** using **multiple wavelength (λs)**:

- Same **λs for all** DRAM DIMMs in a memory channel.
- A **single fiber can carry multiple λs** (requires less wires than an electrical bus).

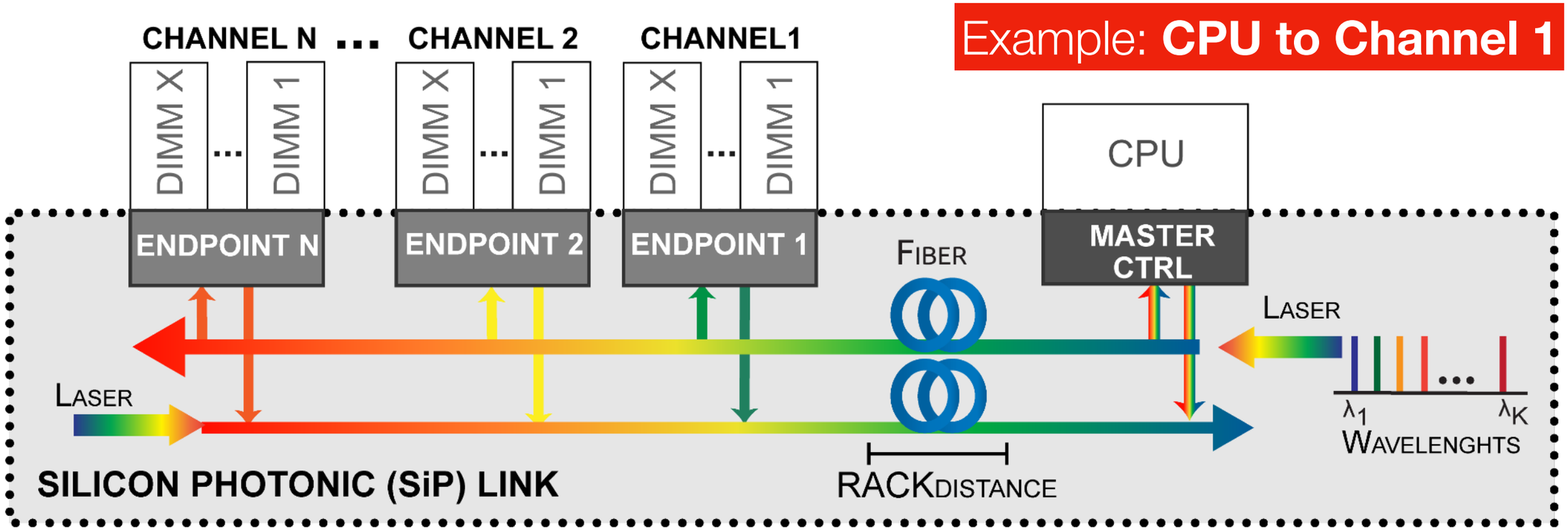
Optically Connected Memory (OCM)



OCM uses **DWDM** to achieve **high aggregated bandwidth** using **multiple wavelength (λs)**:

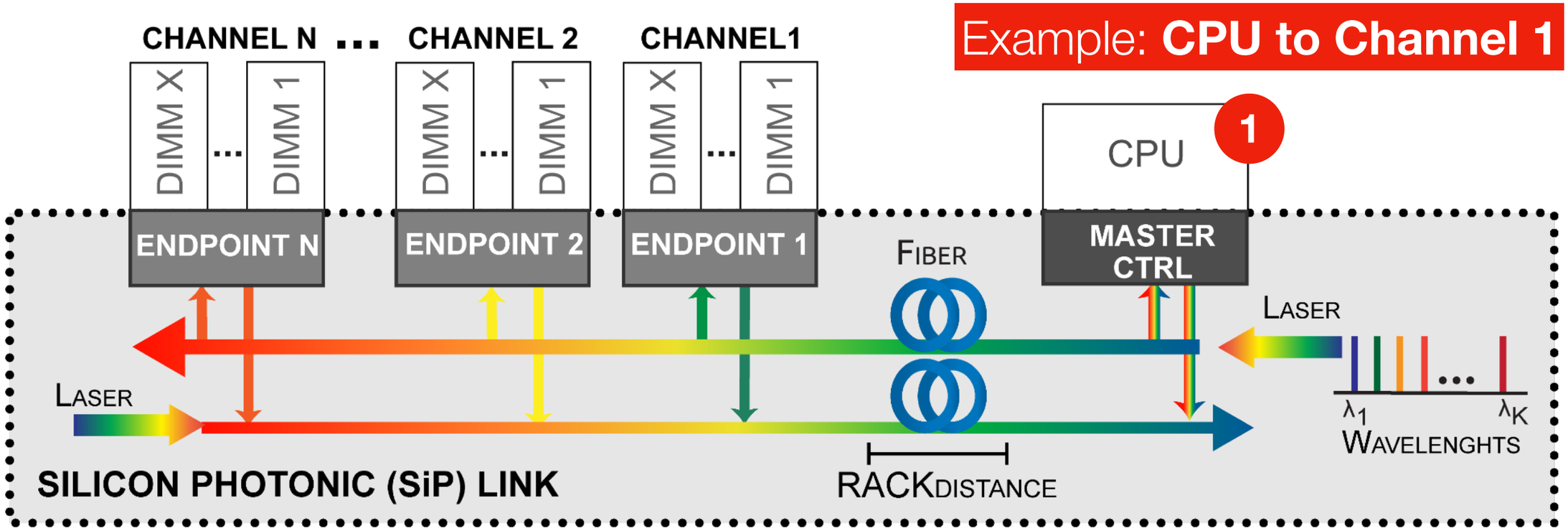
- Same **λs for all** DRAM DIMMs in a memory channel.
- A **single fiber can carry multiple λs** (requires less wires than an electrical bus).

OCM Read/Write Operation



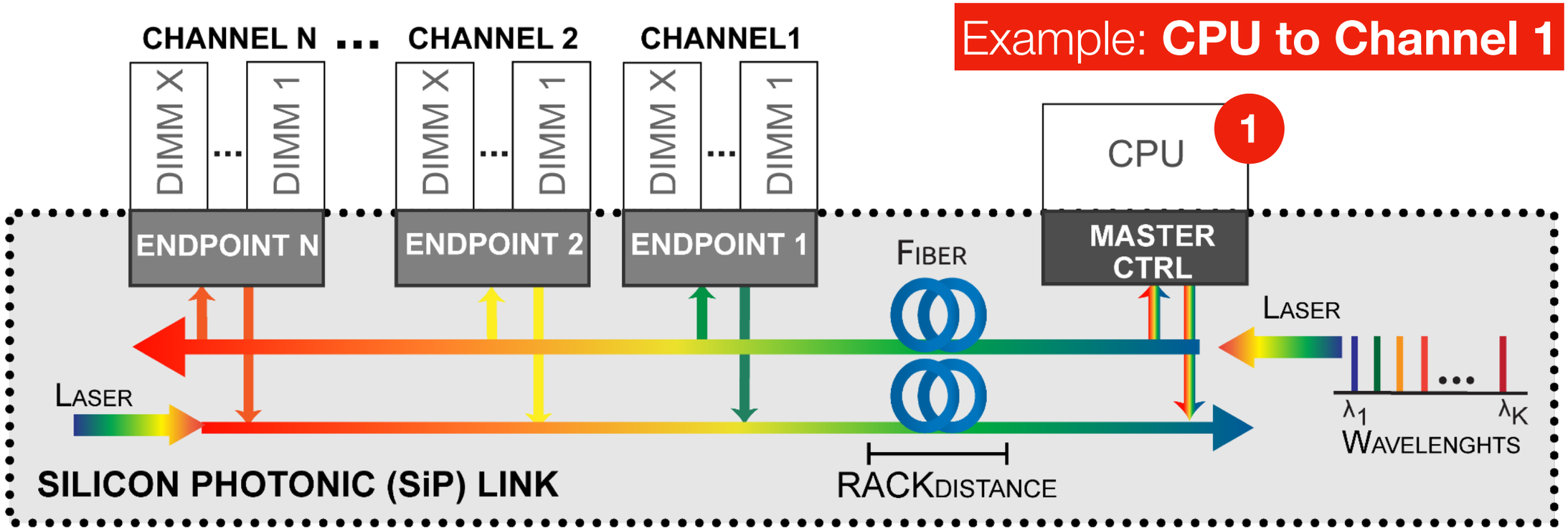
OCM **operation steps:**

OCM Read/Write Operation



OCM **operation steps:**

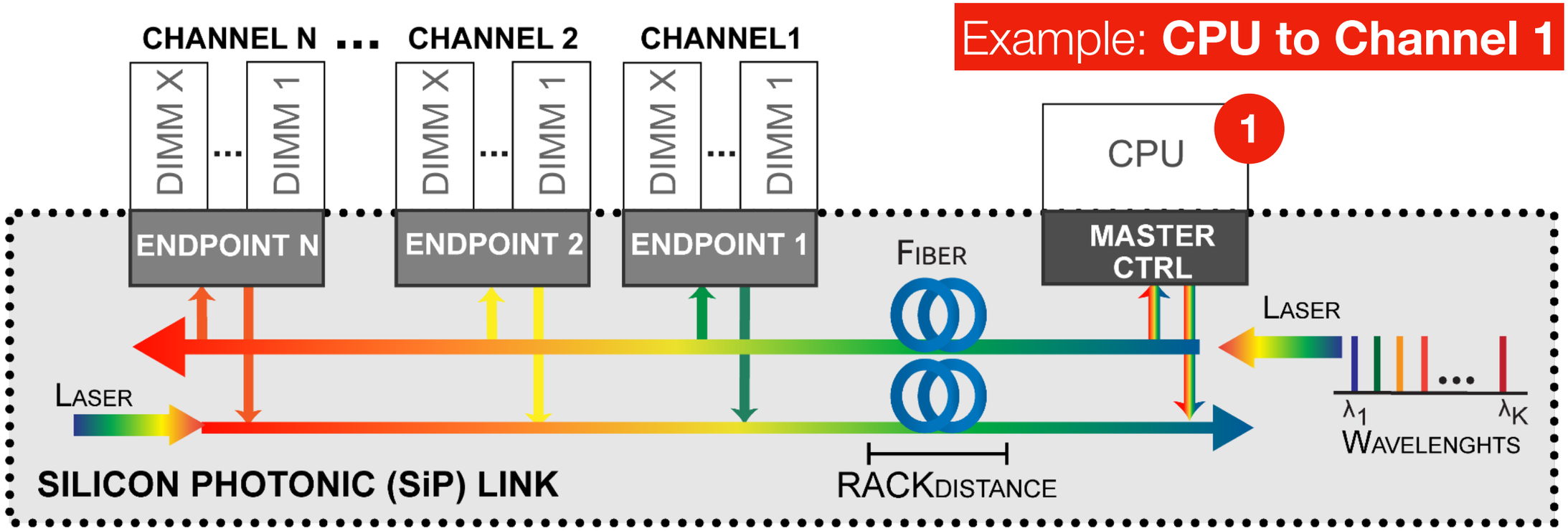
OCM Read/Write Operation



OCM **operation steps:**

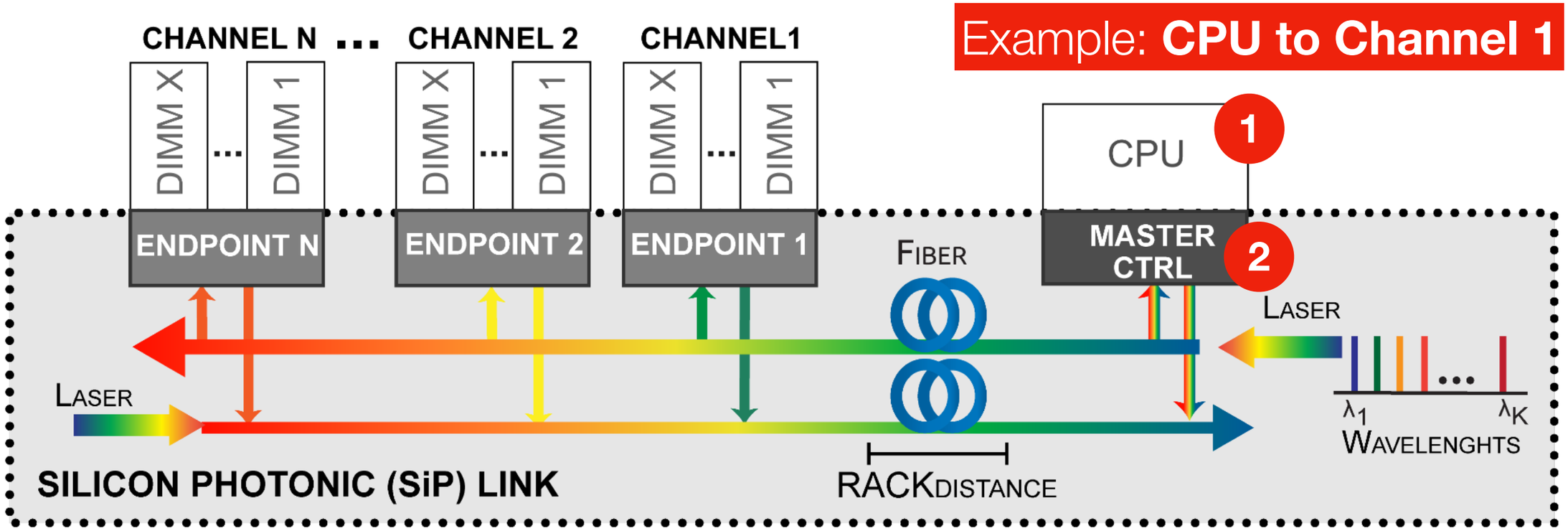
1. Processor **request of Read or Write** operation.

OCM Read/Write Operation



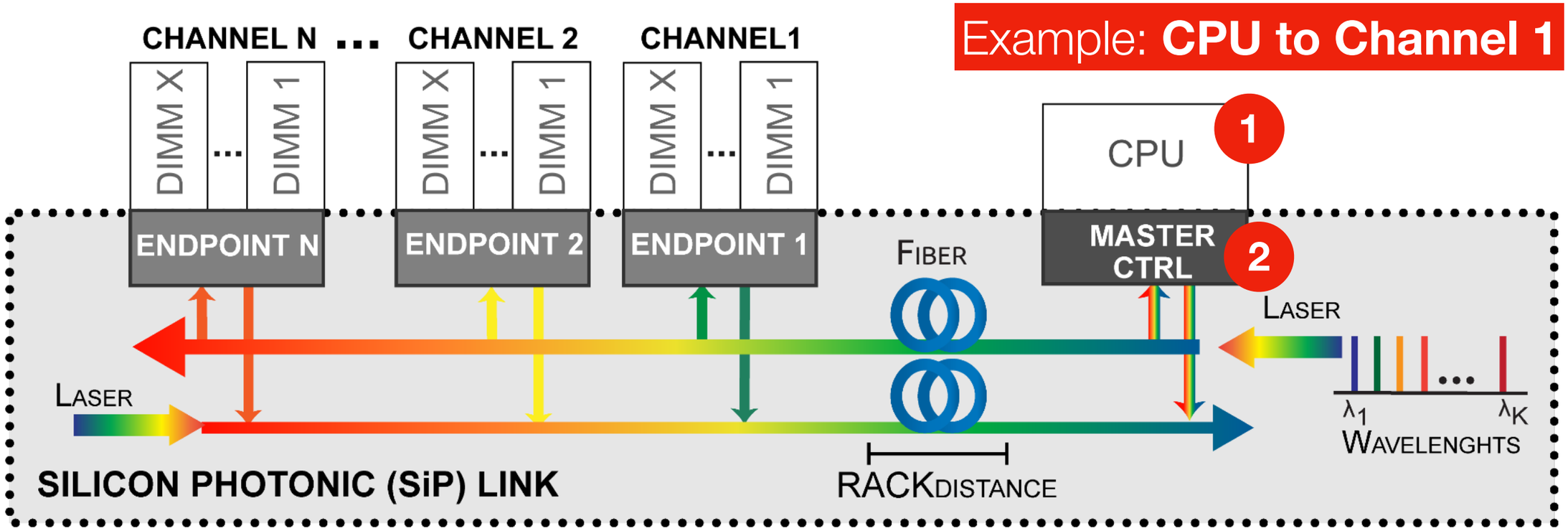
OCM **operation steps:**

OCM Read/Write Operation



OCM **operation steps:**

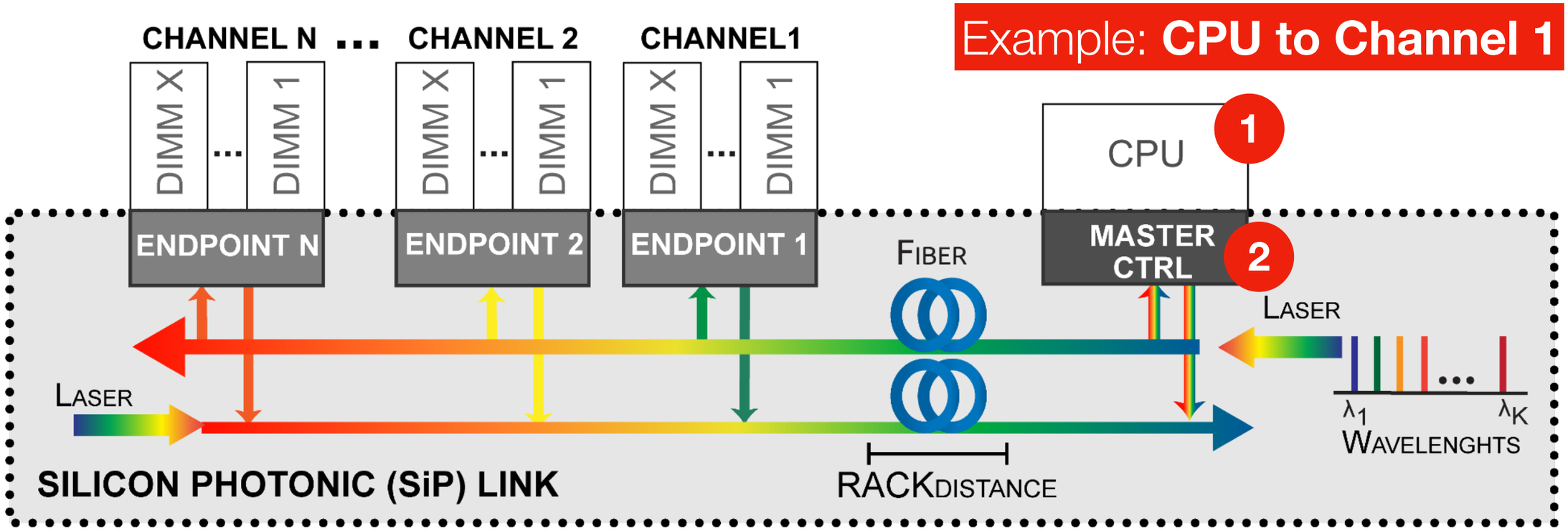
OCM Read/Write Operation



OCM **operation steps:**

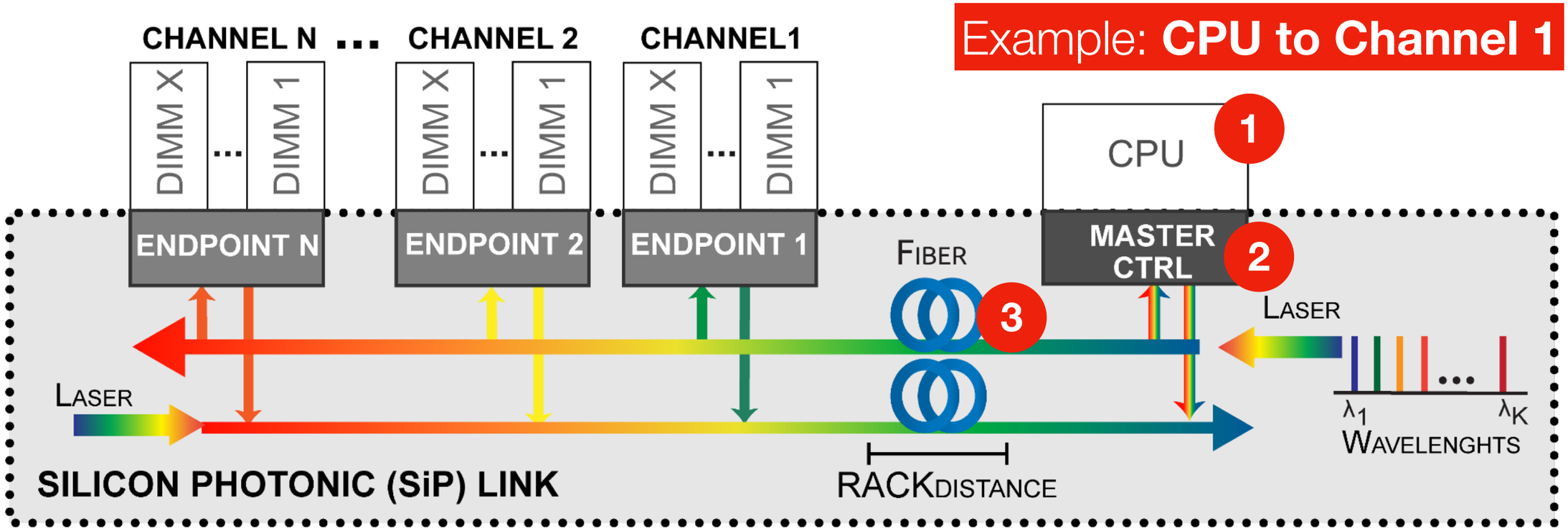
2 **Electrical** signals **to optical domain conversion** for TX **(Modulation)**.

OCM Read/Write Operation



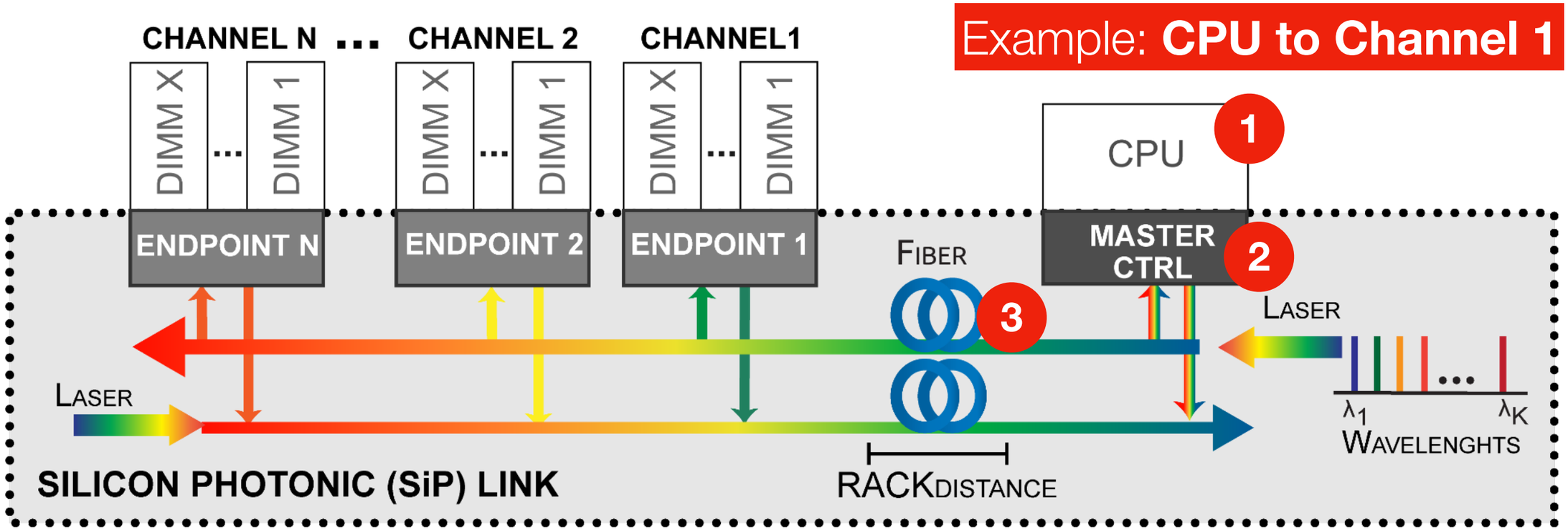
OCM **operation steps:**

OCM Read/Write Operation



OCM **operation steps:**

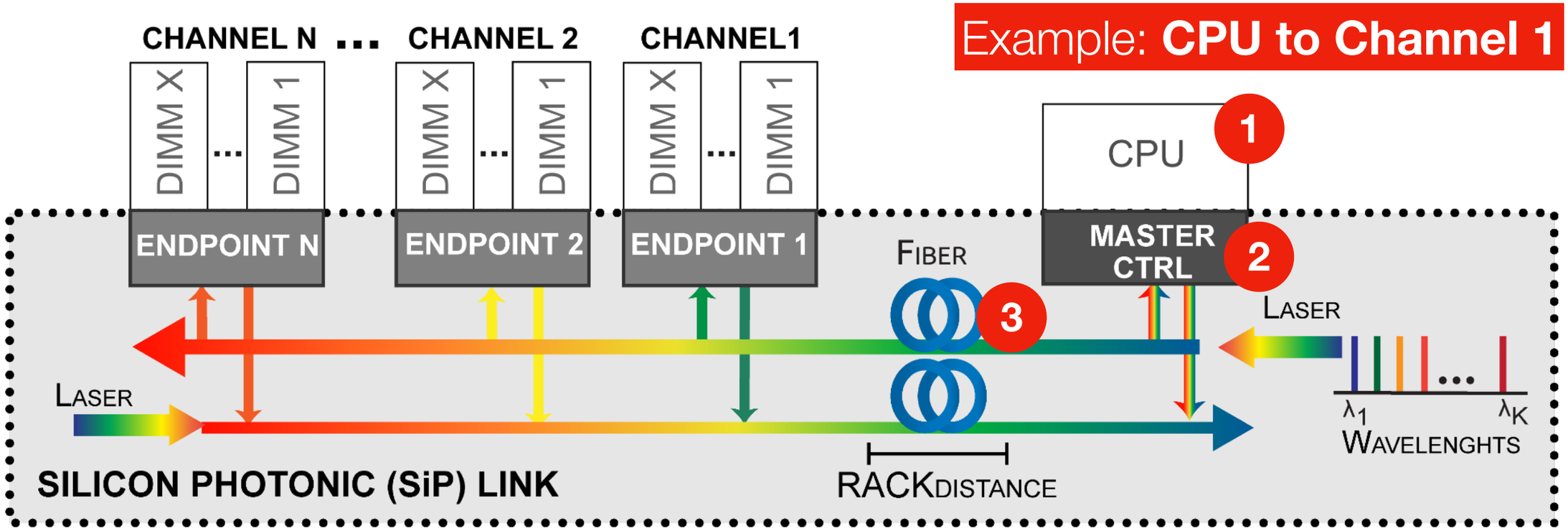
OCM Read/Write Operation



OCM **operation steps:**

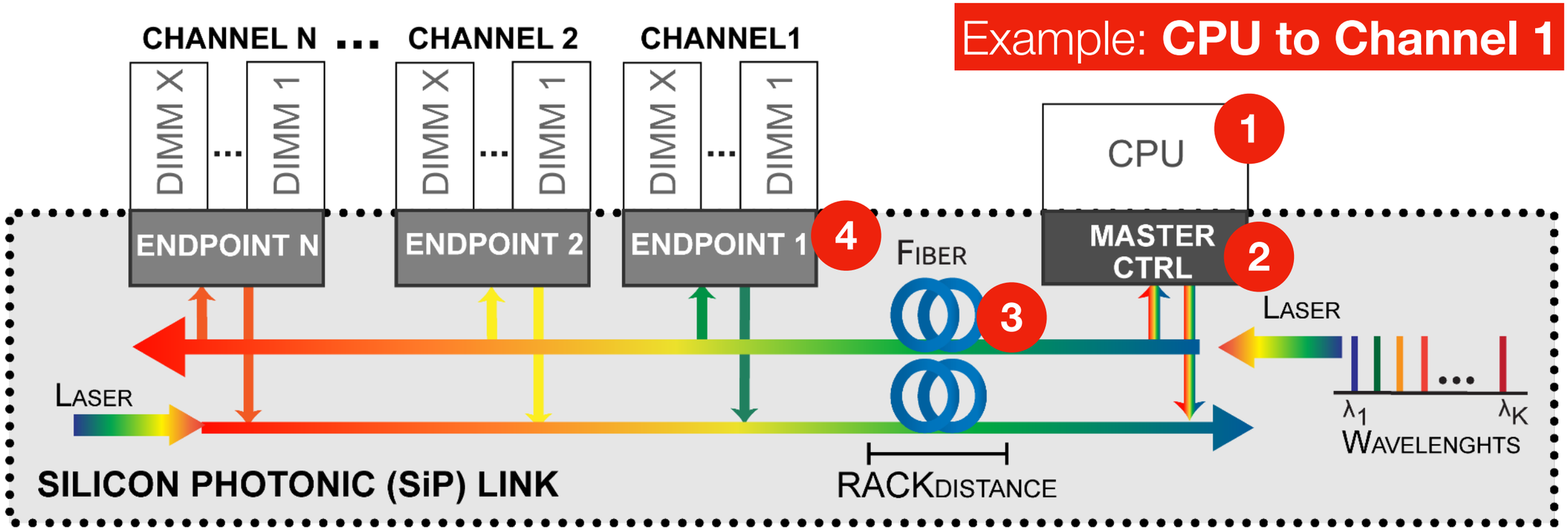
3 Signal **propagation through the optical fiber.**

OCM Read/Write Operation



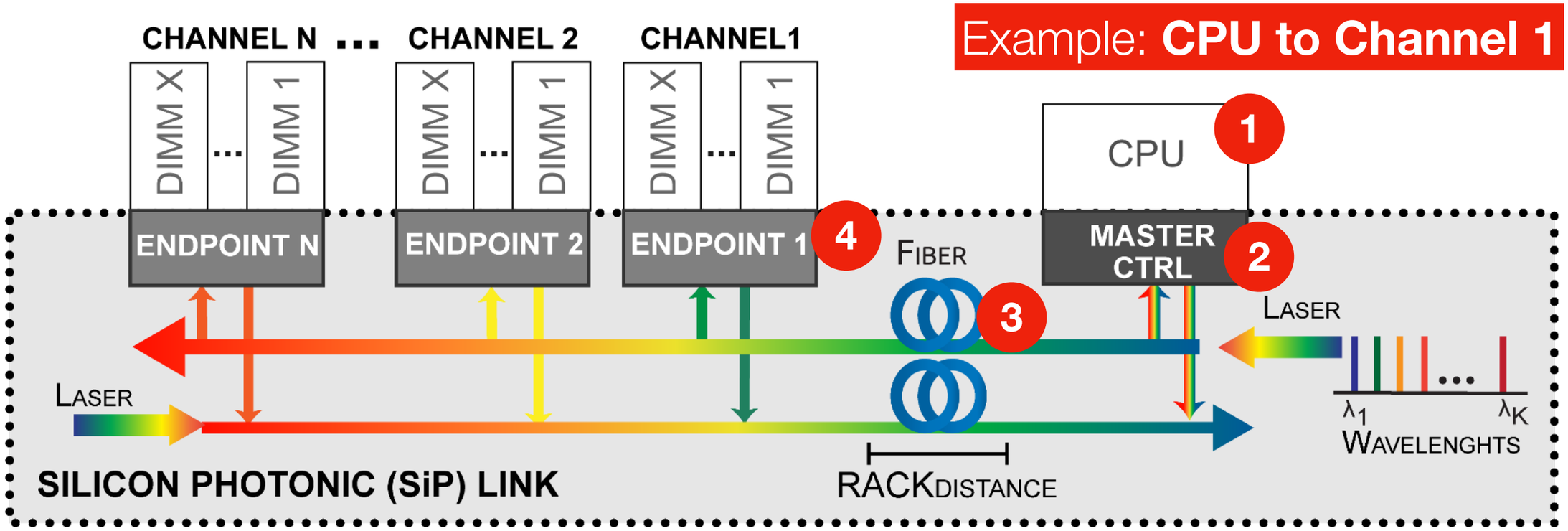
OCM **operation steps:**

OCM Read/Write Operation



OCM **operation steps:**

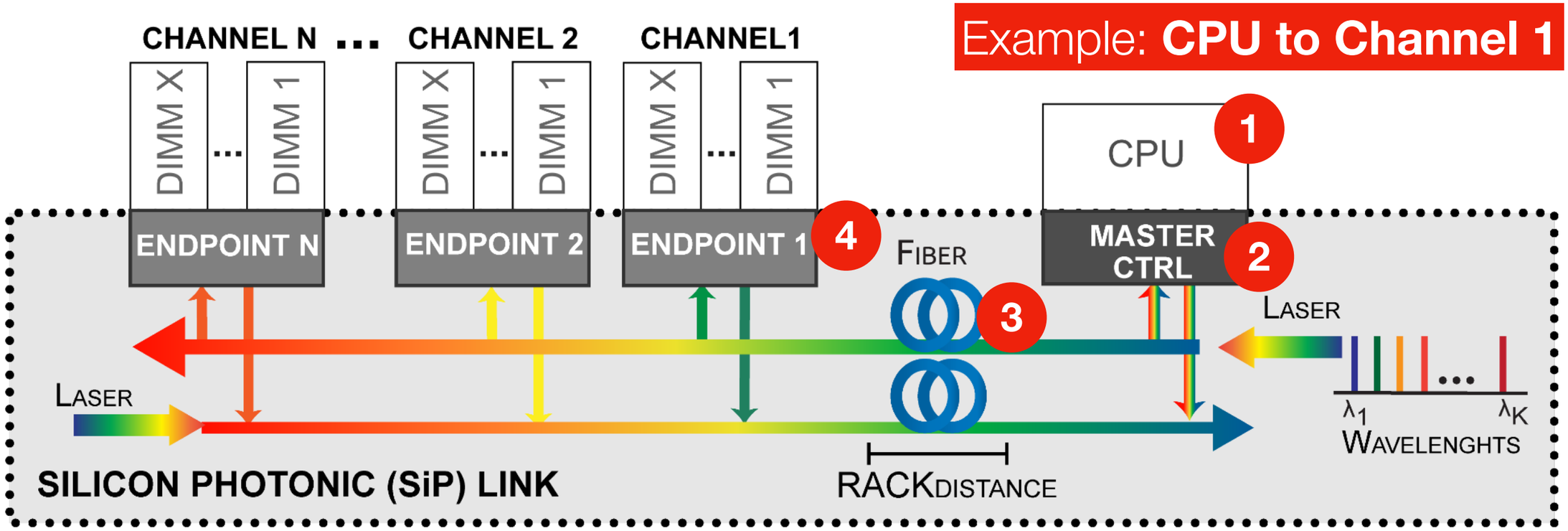
OCM Read/Write Operation



OCM **operation steps:**

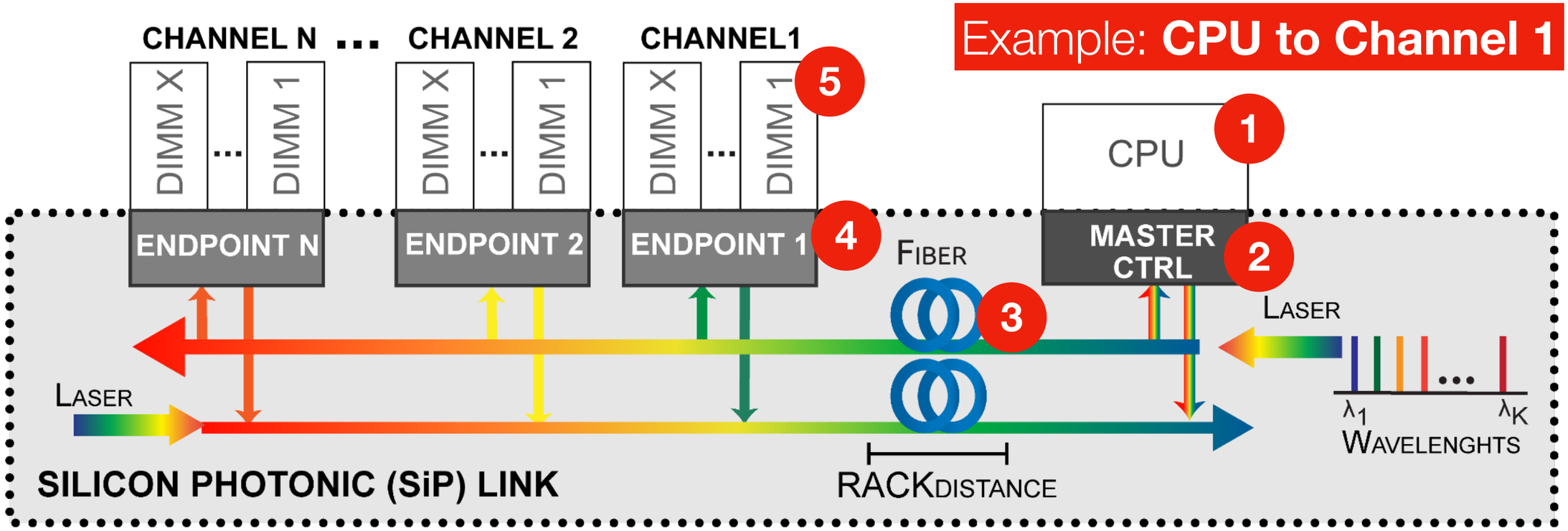
4 **Optical to electrical conversion** at **RX (Demodulation).**

OCM Read/Write Operation



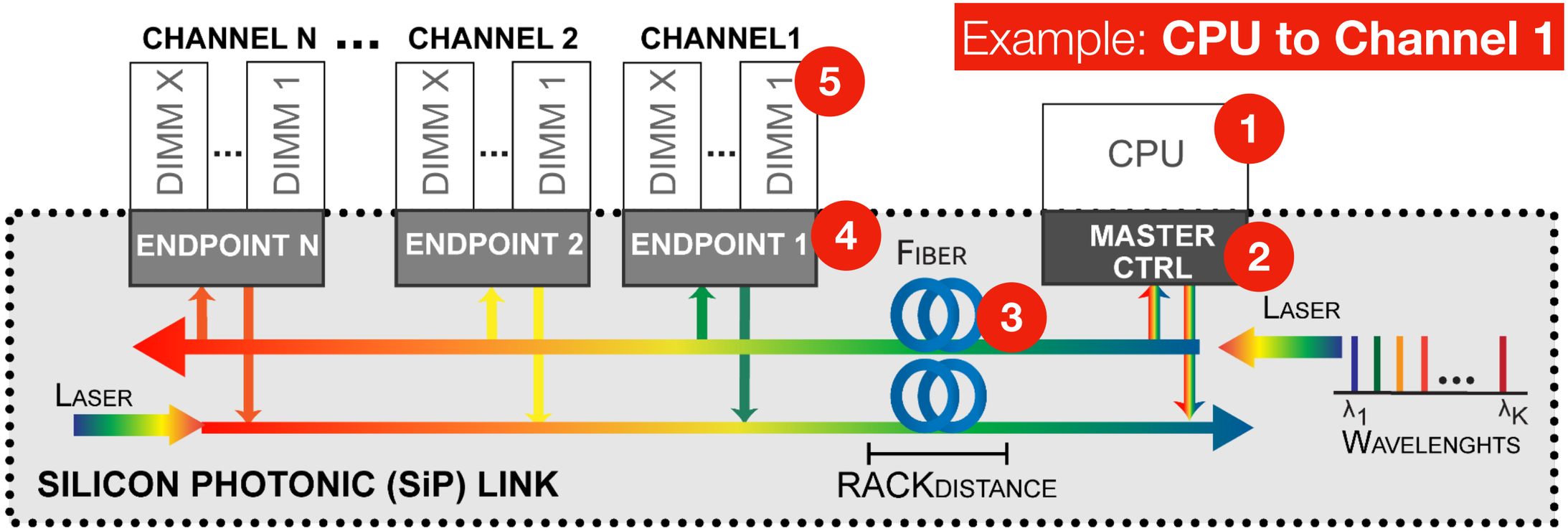
OCM **operation steps:**

OCM Read/Write Operation



OCM **operation steps:**

OCM Read/Write Operation



OCM **operation steps:**

5 **WR: Store data** in DRAM DIMM. **RD: Load data** and **repeat steps 1 to 4 (from DIMM to CPU)**

OCM Timing

- OCM **DDR latency** and **memory controller latency** is the same as conventional DDR.
- **OCM latency overhead**: **SERDES latency + Distance propagation**
- OCM incurs in **latency overhead** because of:
  - **Signal conversion** electrical to/from optical.
  - **Signal propagation** latency at rack distance.

- **Lockstep operation:**
  - **Parallel access** to DRAM DIMMs in the same memory channel.
  - Larger **cache line size, e.g.: 128B, split** into the DRAM DIMMs on the same channel.
  - Allow to **reduce the latency overhead.**

Outline

Introduction

Background

Motivation and Goal

Optically Connected Memory (OCM)

Evaluation

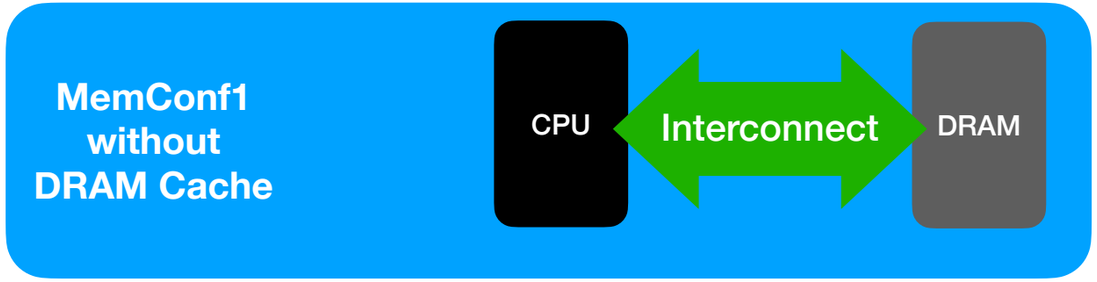
Conclusion

Experimental Setup

- **System-level performance**: modified version of ZSIM simulator.
  - **Benchmarks**: SPEC06, SPEC17, Parsec, Splash, GAPS

Baseline	Processor	3 GHz, 8 cores, 128B cache lines
	Cache	32KB L1(D+I), 256KB L2, 8MB L3
MemConf1	Mem	4 channels, 2 DIMMs/channel, DDR4-2400
MemConf2	Mem	1 channel, 2 DIMMs/channel, DDR4-2400
	DRAM cache	4GB stacked, 4-way, 4K pages, FBR, DDR4-2400
OCM	SERDES	latency: 10 - 340 cycles
	Fiber	latency: 30/60/90 cycles (2/4/6 meters roundtrip)
NIC	40G PCIe	latency: 1050 cycles

Setup Parameters

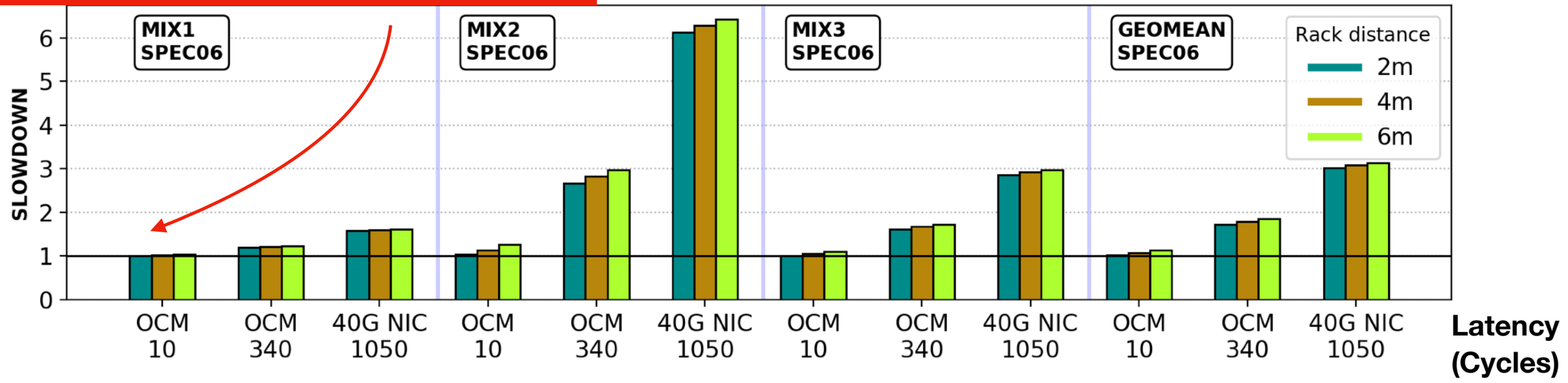


- **SiP link evaluation**: custom model in our PhoenixSim simulator **[Rumley+, AISTECS'16].**
  - Modified to **sustain current DDR4-2400** memory systems.

Results MemConf1 w/out DRAM Cache

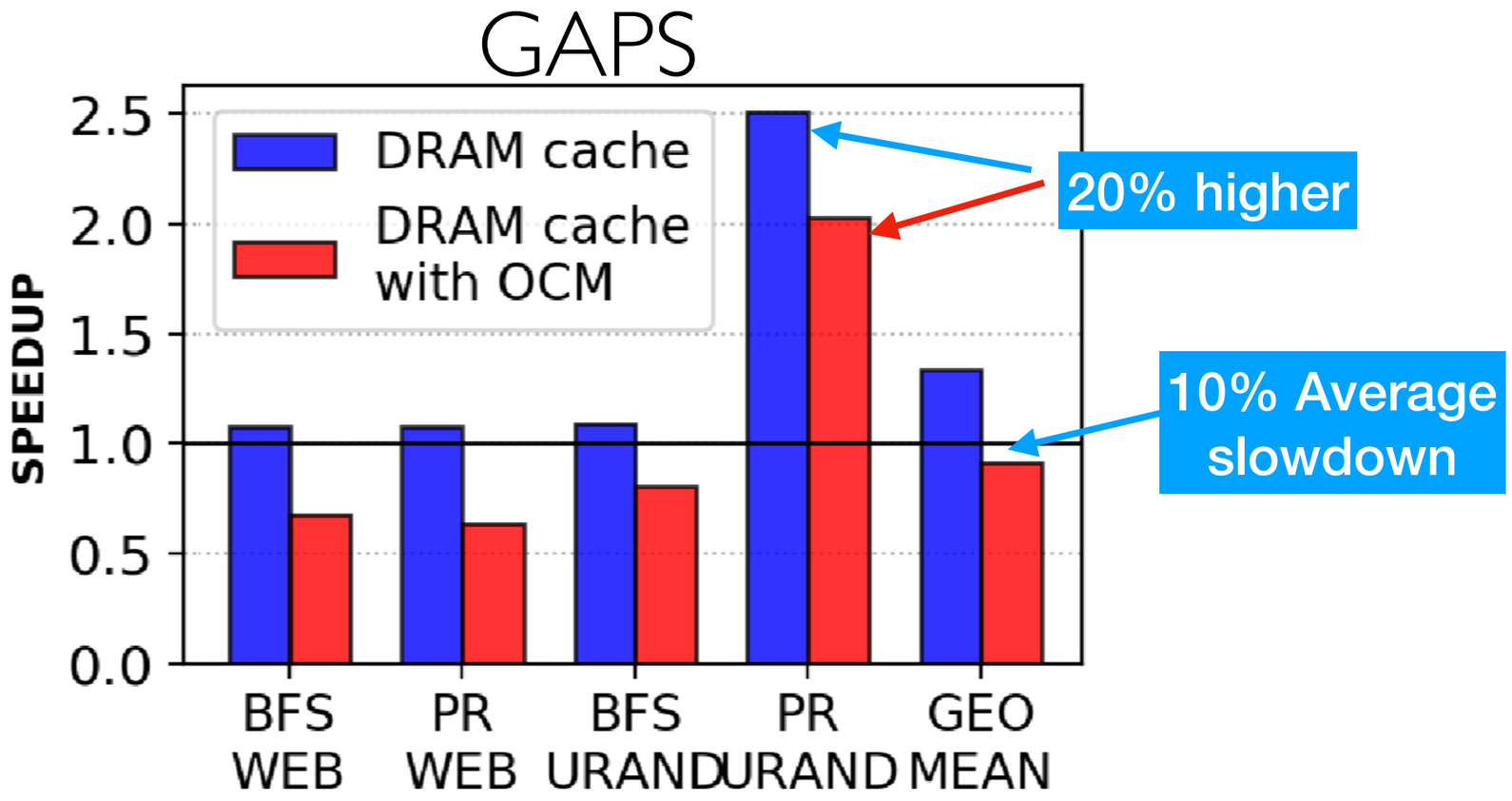
SPEC06

Lower delay is close to baseline performance



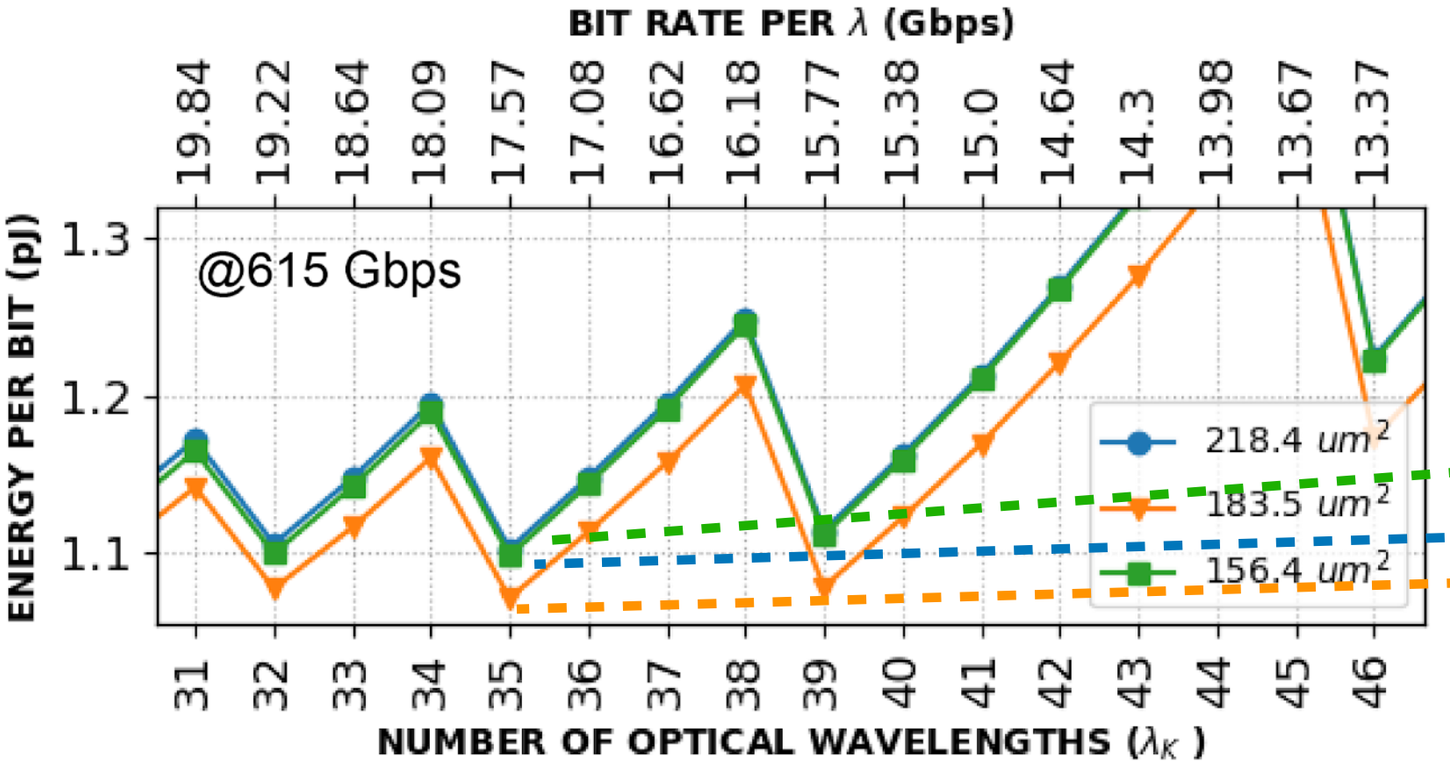
- As expected higher OCM latency degrades performance.
  - Average **slowdown is 1.07x** with the **optimistic scenario.**
  - Average **slowdown is 1.78x** with the **worst-case scenario.**
- OCM **is up to 5.5x faster** than 40G NIC.

Results MemConf2 with DRAM Cache



- For most of the benchmarks, OCM has less than **38% slowdown.**
- PR with Urand input graph **exhibits > x2 speedup** on both OCM and non-disaggregated scenario.

SiP Link Results



Lowest energy consumption

1.07 pj/bit @ 615 Gbps

- For both Gbps rates, 183.5 um² rings have the **lowest energy consumption.**
- **DDR4-2400 bandwidth:** 2x615 Gbps link: **35** wavelengths **(MRRs)** @ 17.57 Gbps.

Additional **details in the paper:**

- More **OCM architecture** details in the paper:
  - Timing diagram and operation.
- More **results** in the paper:
  - Measured memory footprint.
  - Multithreaded results.
  - Multiprogrammed results.
  - Additional SiP link configuration.

Outline

Introduction

Background

Motivation and Goal

Optically Connected Memory (OCM)

Evaluation

Conclusion

Conclusion

Conclusion

- We proposed and evaluated Optically Connected Memory (OCM), a new optical architecture for disaggregated main memory systems, compatible with current DDR DRAM technology.
- We made **three key observations:**
  1. OCM has **low energy overhead of only 10.7%** compared to DDR DRAM data movement energy consumption.
  2. OCM **performs x5.5 faster** than a 40G NIC- based disaggregated memory.
  3. OCM **can fit the bandwidth requirements** of commodity DDR4 DRAM modules.
- We conclude that OCM is a promising step towards future data centers with disaggregated main memory.

Optically Connected Memory for Disaggregated Data Centers

Jorge Gonzalez[1] Alexander Gazman[2] Maarten Hattink[2] Mauricio G.Palma[1]
Meisam Bahadori[3] Ruth Rubio-Noriega[4] Lois Orosa[5]
Madeleine Glick[2] Onur Mutlu[5] Keren Bergman[2] Rodolfo Azevedo[1]

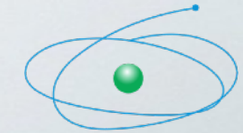
1

2

3

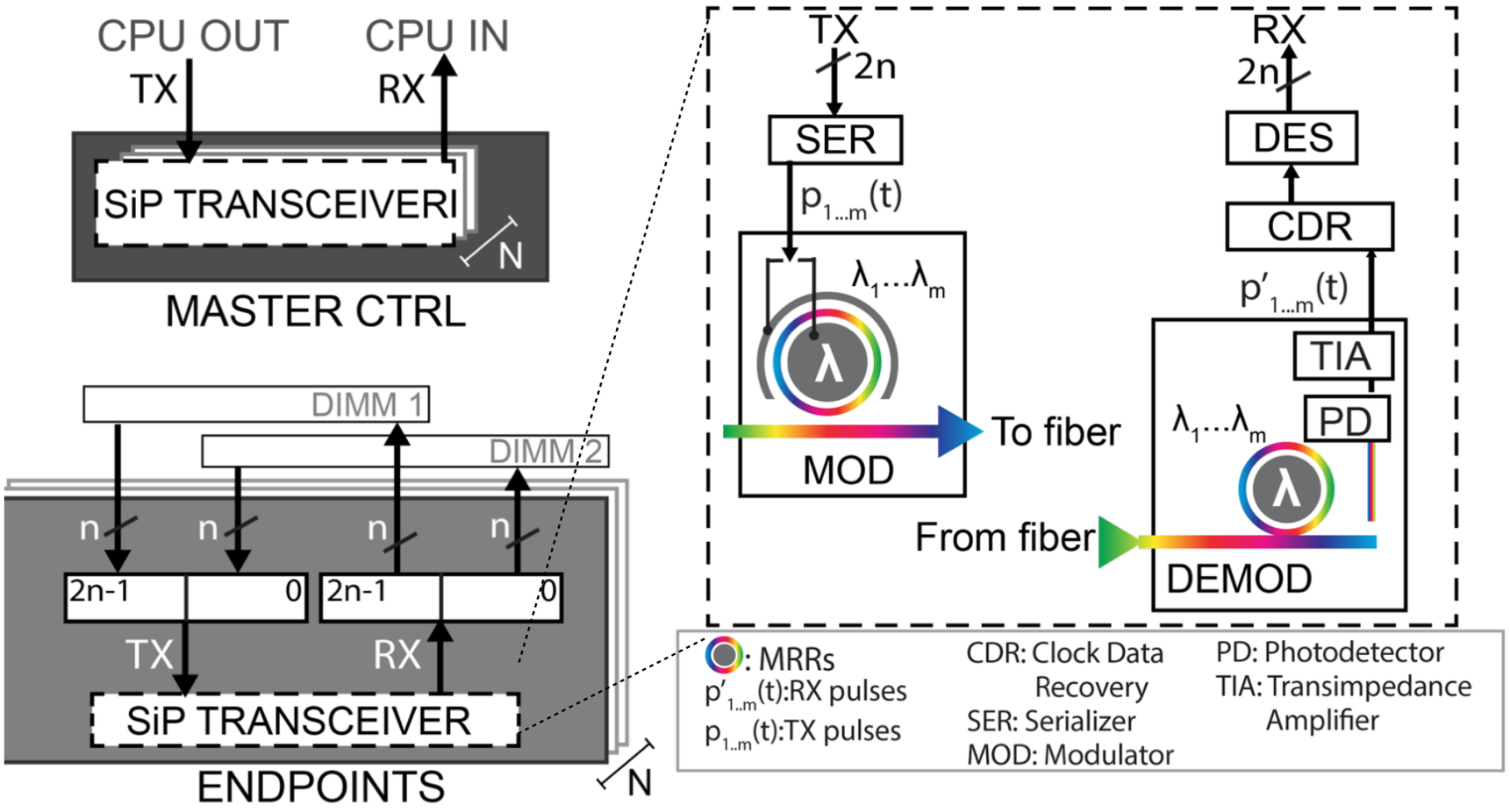
4

5



SiP Link Controllers and Transceivers

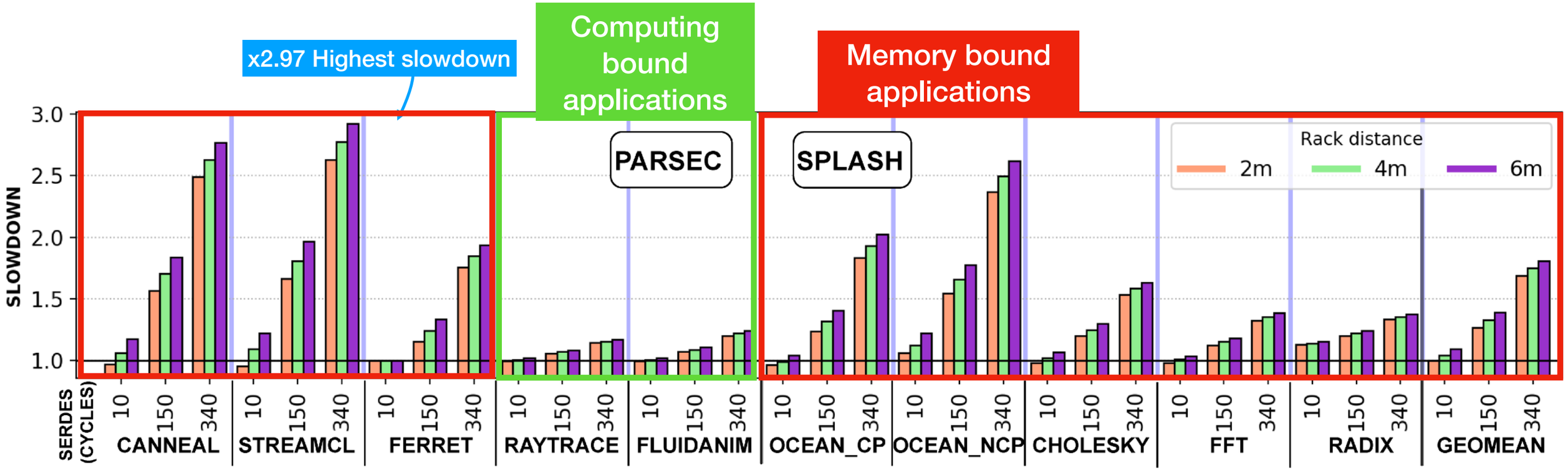
SiP link design based on state-of-the art devices **[Bahadori+, JLT'16], [Bahadori+, JLT'18], [Bahadori+, OI'16] , [Polster+, TVLSI'16]**



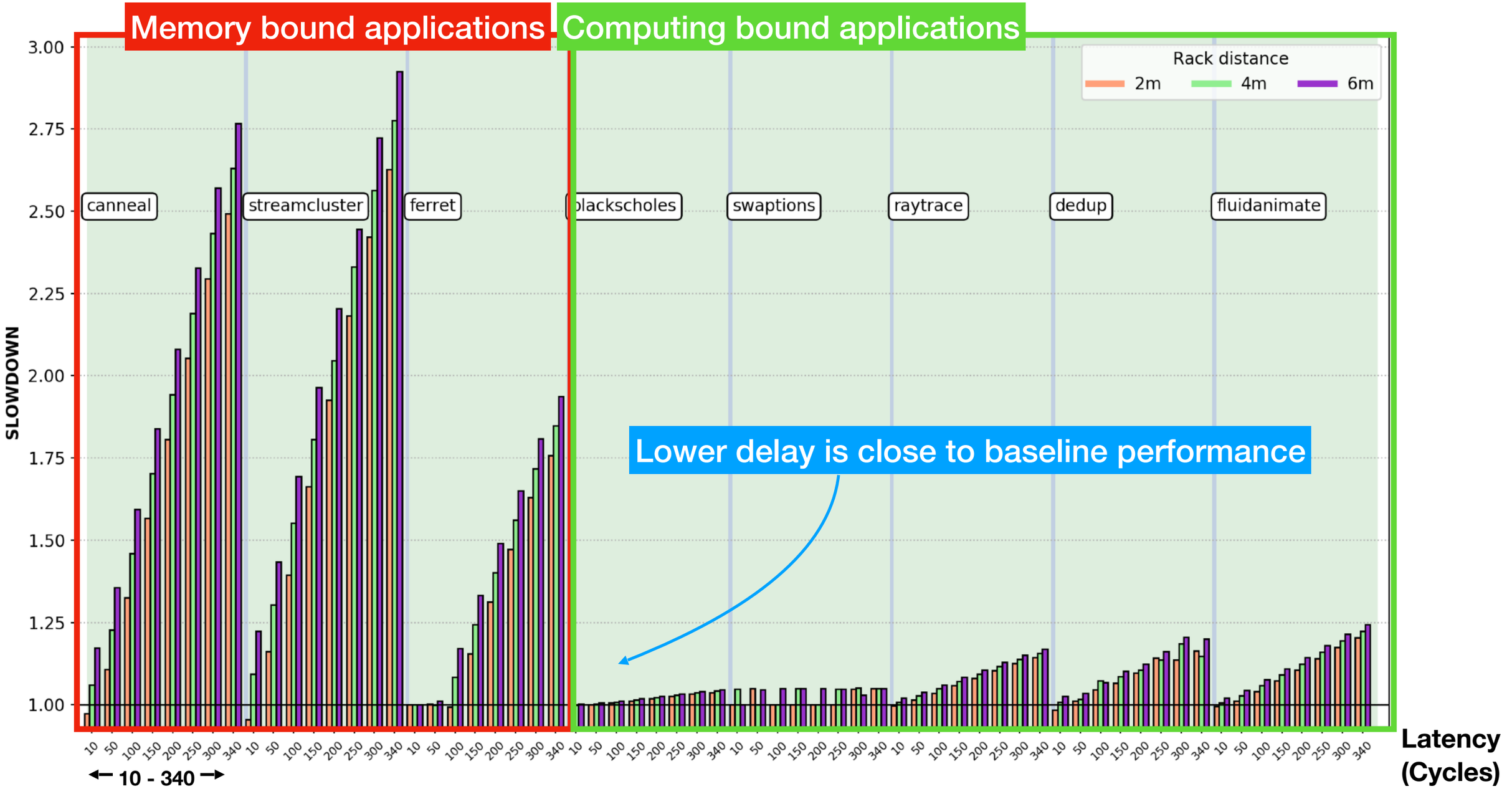
- **Controllers** have transceivers:
  - Master, on the processor side.
  - Endpoint, on the DRAM DIMM
- A **transceiver** is a microring resonator (MRR) array for TX (modulators) and RX (demodulators).

Results MemConf1 w/out DRAM Cache

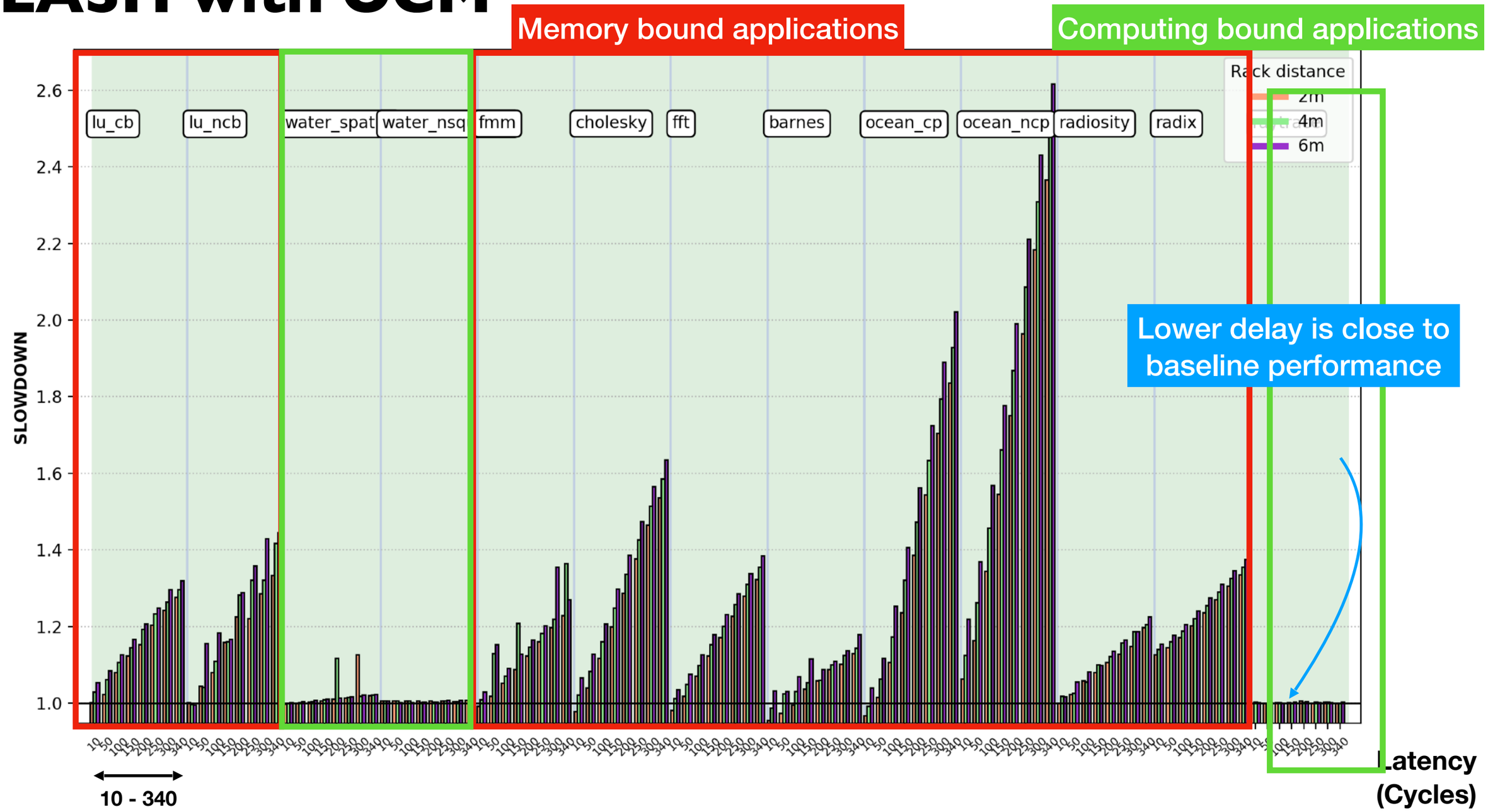
- **Average slowdown** is 1.3x for Splash2x and 1.4x for Parsec.



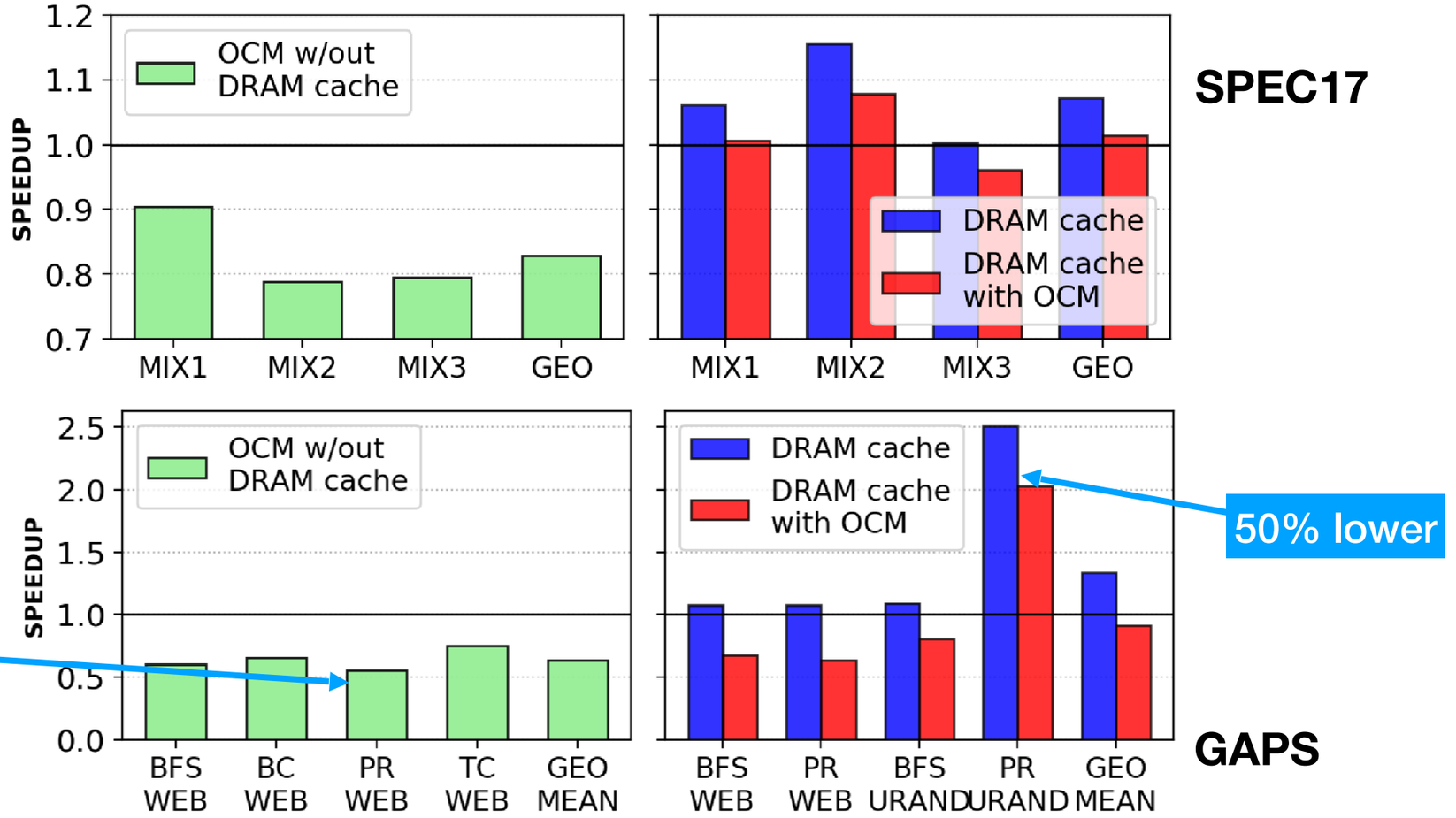
PARSEC with OCM



SPLASH with OCM

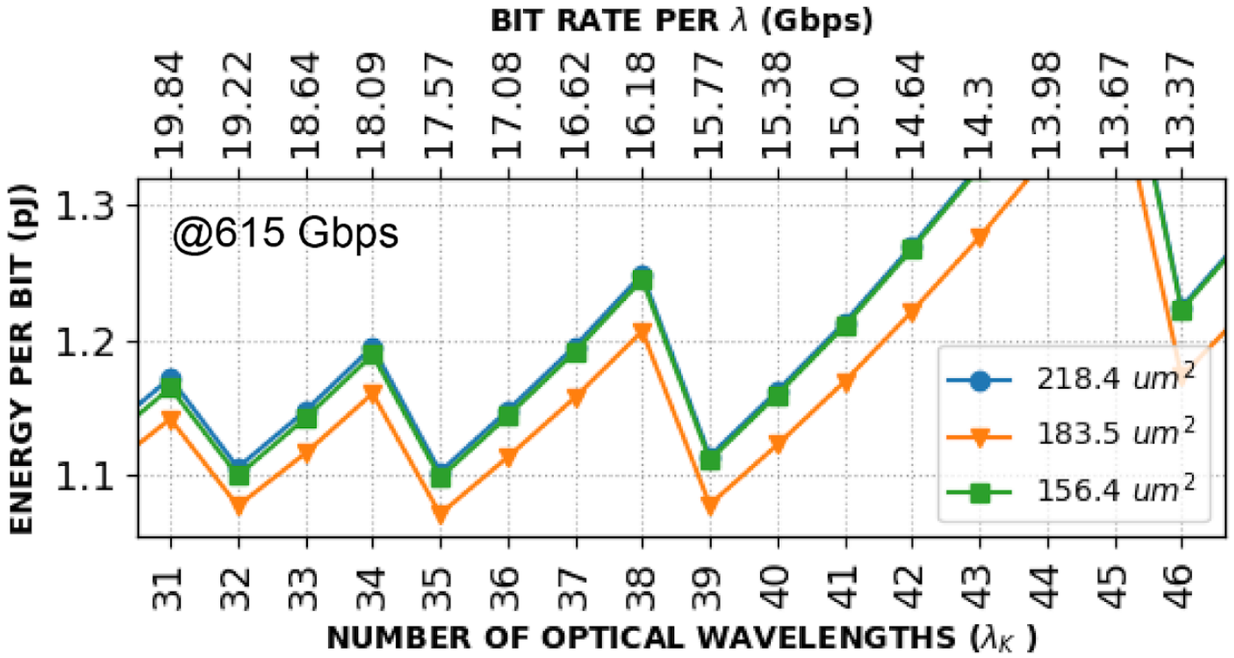


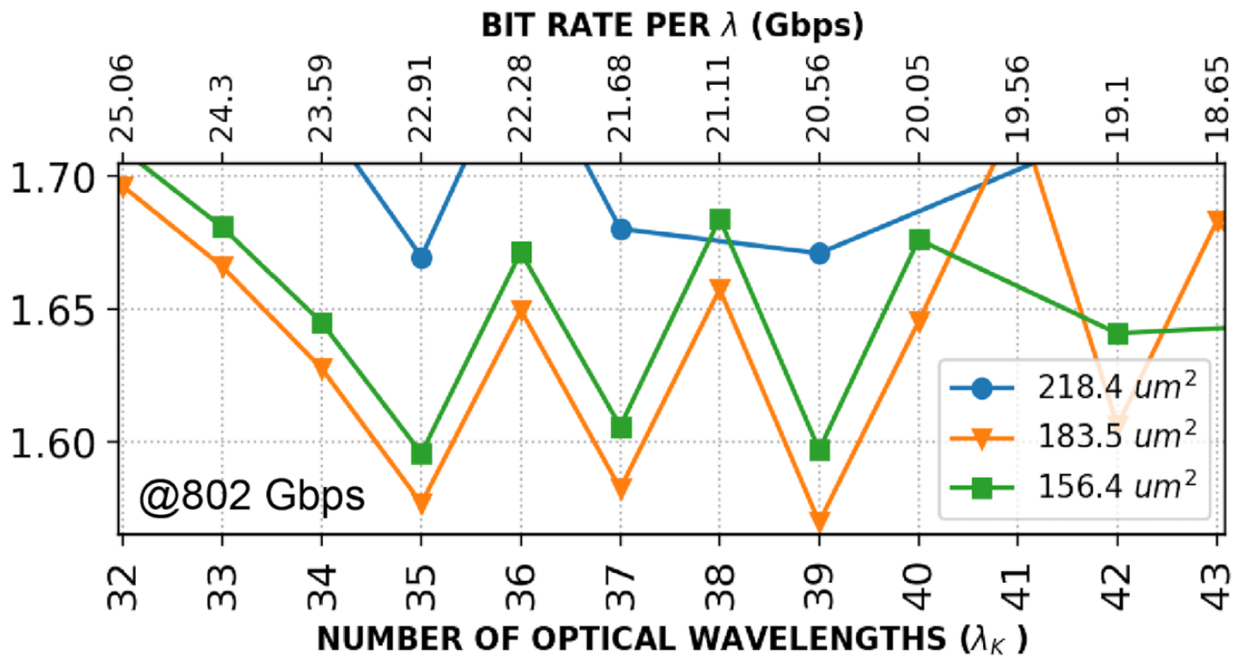
Results MemConf2 with DRAM Cache



- For most of the benchmarks, OCM has less than **38% slowdown.**
- PR with Urand input graph **exhibits > x2 speedup** on both OCM and non-disaggregated scenario.

SiP Link Results





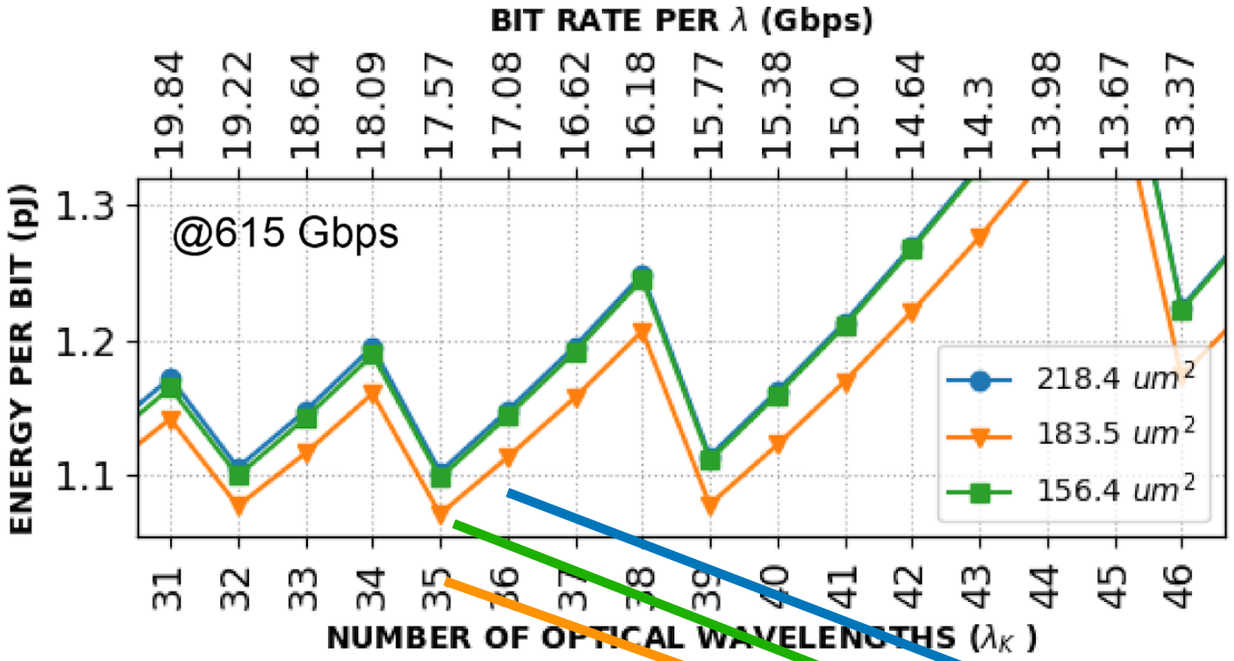
1.07 pj/bit @ 615 Gbps

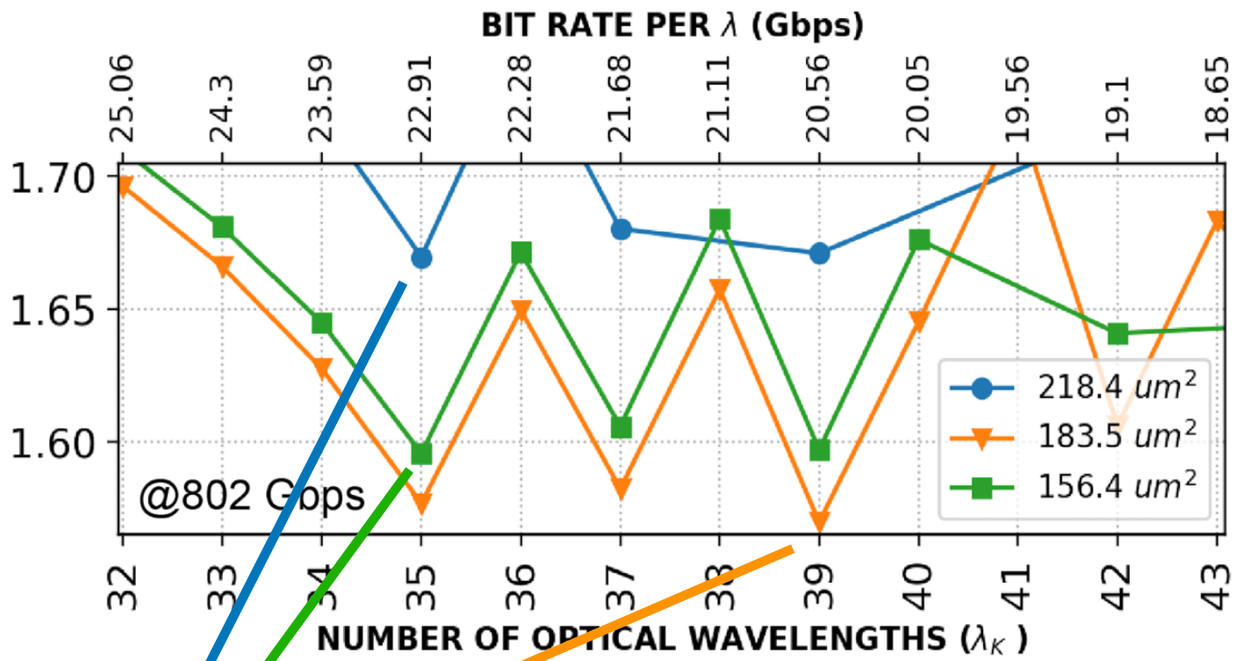
Lowest energy consumption

1.57 pj/bit @ 802 Gbps

- For both Gbps rates, 183.5 um² rings consume the **lowest energy.**
- 615 Gbps link: 35 wavelengths **(MRRs)** @ 17.57 Gbps. **Area** overhead: 51.4E-3 mm²
- 802 Gbps link: 39 wavelengths **(MRRs)** @ 20.56 Gbps. **Area** overhead: 57.3E-3 mm²

SiP Link Results





1.07 pj/bit @ 615 Gbps

Lowest energy consumption

1.57 pj/bit @ 802 Gbps

- For both Gbps rates, 183.5 um² rings consume the **lowest energy.**
- 615 Gbps link: 35 wavelengths **(MRRs)** @ 17.57 Gbps. **Area** overhead: 51.4E-3 mm²
- 802 Gbps link: 39 wavelengths **(MRRs)** @ 20.56 Gbps. **Area** overhead: 57.3E-3 mm²