Demystifying the Characteristics of 3D-Stacked Memories: A Case Study for the Hybrid Memory Cube (HMC)

Ramyad Hadidi, Bahar Asgari , Burhan Ahmad Mudassar, Saibal Mukhopadhyay, Sudhakar Yalamanchili, and Hyesoon Kim

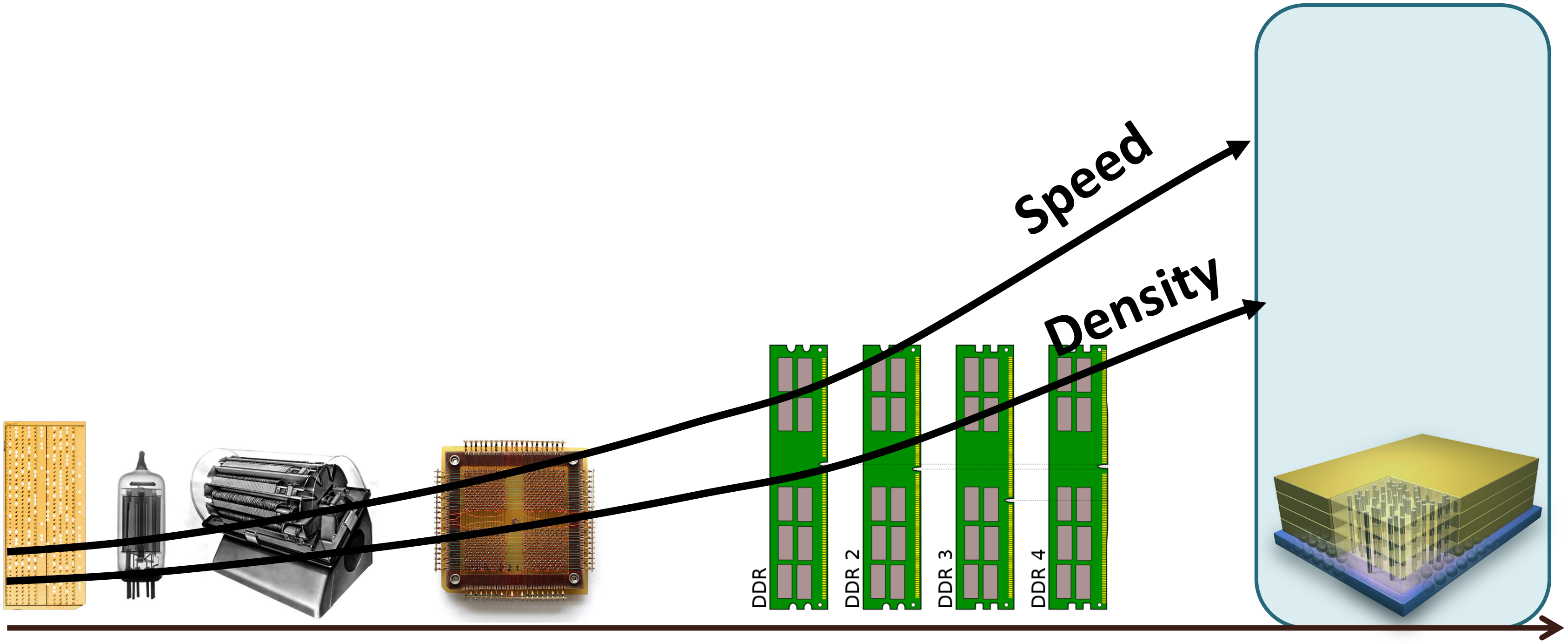
IISWC'17 Talk

Georgia Tech

comparch

CASL
Computer Architecture
and Systems Laboratory

Memory Evolution



3D-Stacking Technology

Provides opportunities & novel features

3D-DRAMs:

- Provide higher bandwidth and density
- Enable lower power consumption
- Motivate processing-in-memory

HMC is an example of such memories.

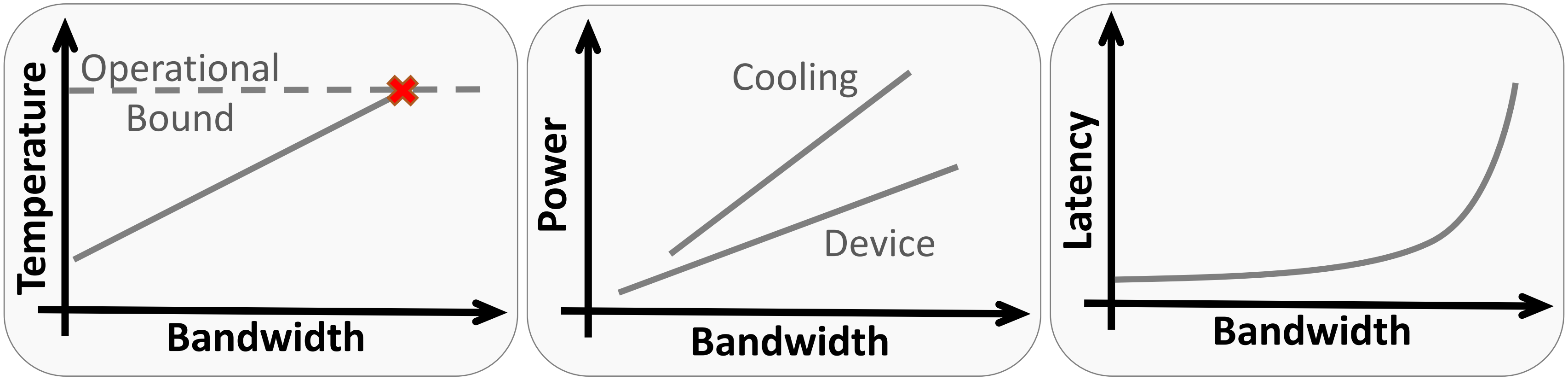
New Considerations

New **internal organization**

New **thermal** behavior

New **latency** and **bandwidth** hierarchy

New packet-switched **interface**



Contributions

We evaluate a real system with HMC 1.1 to:

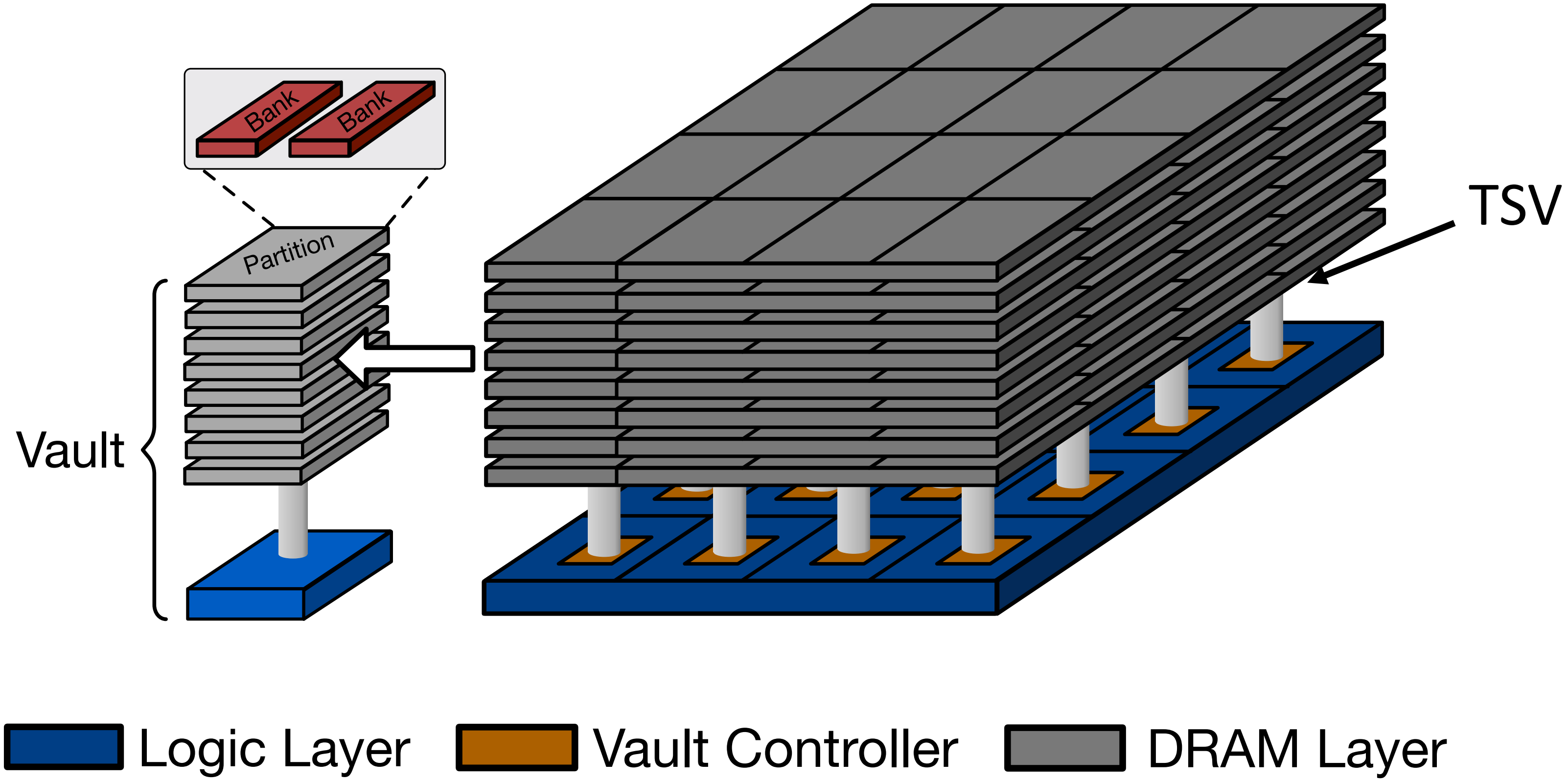
- Study new memory organization
- Present bandwidth, power, and temperature relationships
- Investigate required cooling power
- Explore contributing factors to latency

AC510

To realize the full-system impact of 3D-stacked memories and HMC in particular.

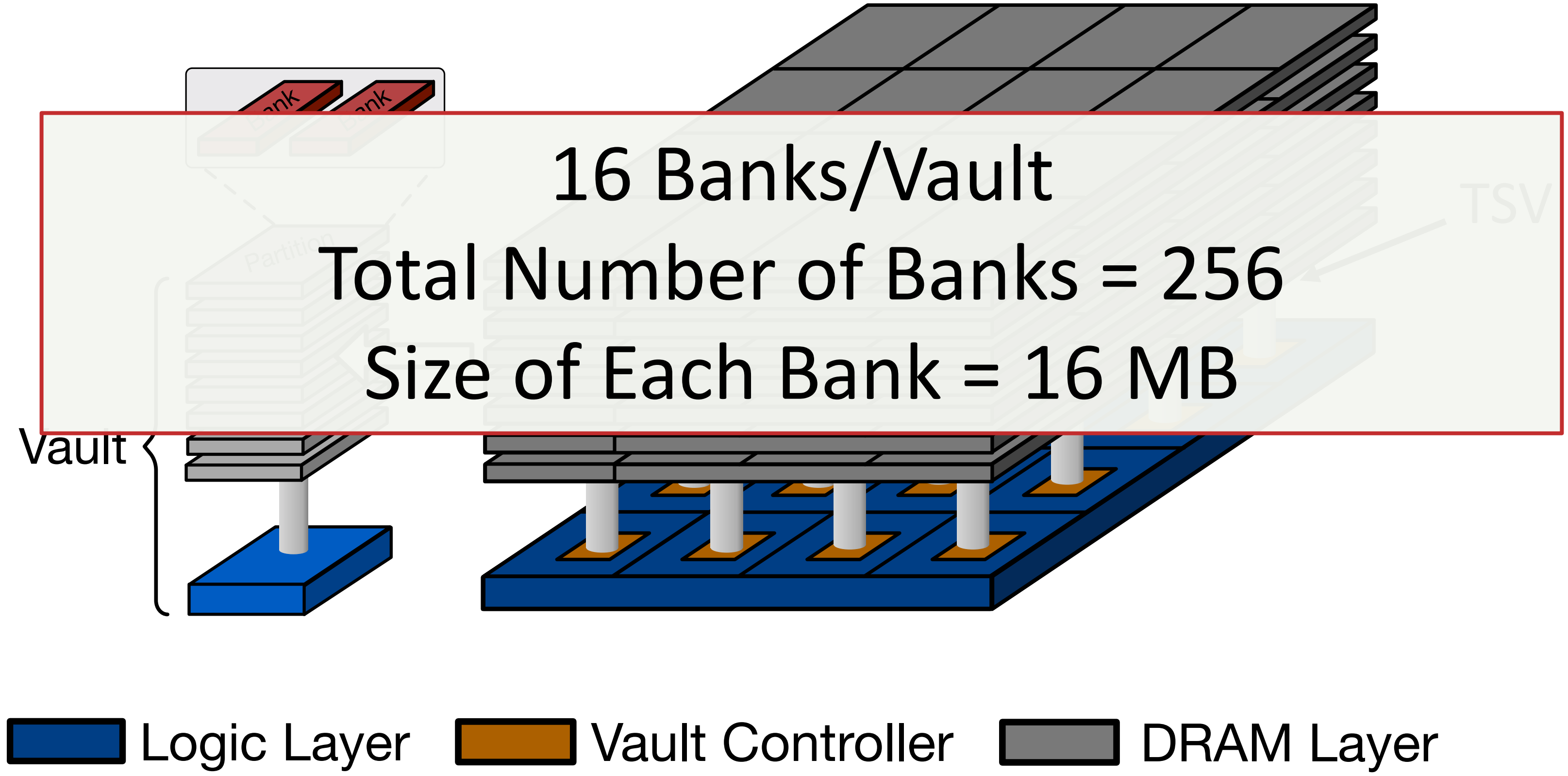
Hybrid Memory Cube (HMC)

HMC 1.1 (Gen2): 4GB size



Hybrid Memory Cube (HMC)

HMC 1.1 (Gen2): 4GB size



16 Banks/Vault

Total Number of Banks = 256

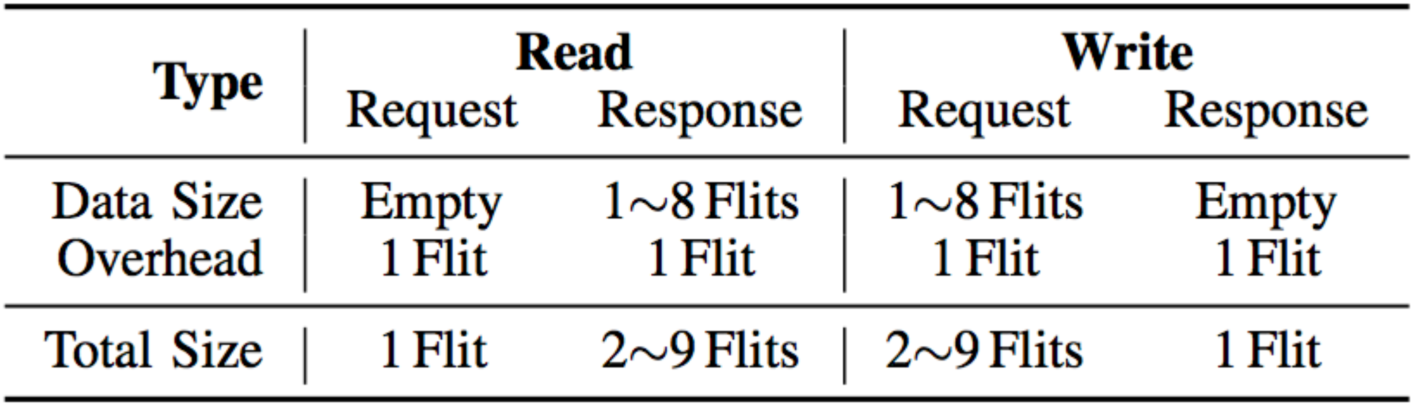
Size of Each Bank = 16 MB

HMC Communication I

Follows a serialized **packet-switched** protocol

Partitioned into 16-byte *flit*

Each transfer incurs 1 flit of overhead



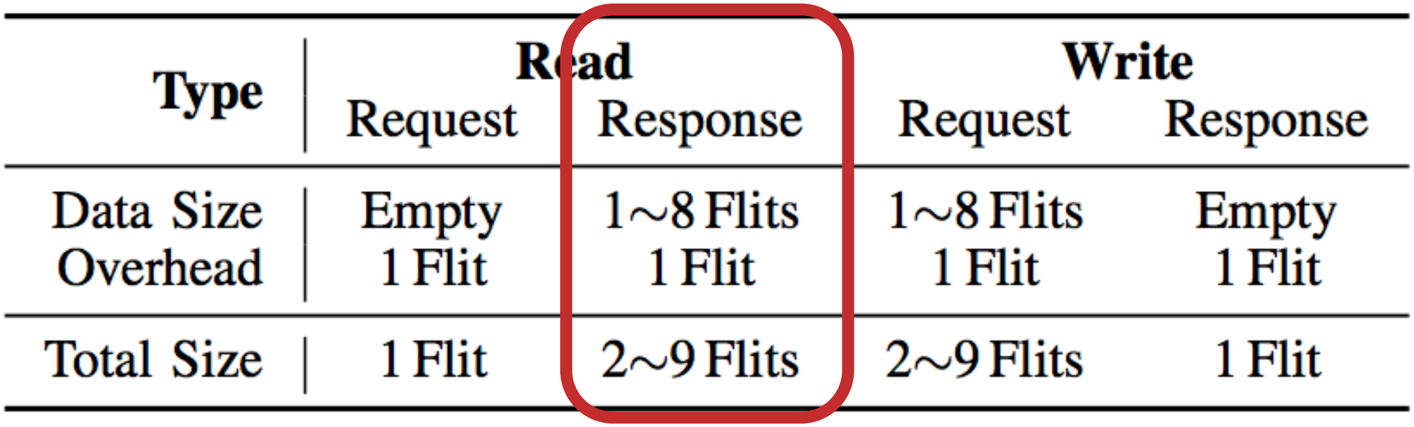
Type	Read		Write	
	Request	Response	Request	Response
Data Size	Empty	1∼8 Flits	1∼8 Flits	Empty
Overhead	1 Flit	1 Flit	1 Flit	1 Flit
Total Size	1 Flit	2∼9 Flits	2∼9 Flits	1 Flit

HMC Communication I

Follows a serialized **packet-switched** protocol

Partitioned into 16-byte *flit*

Each transfer incurs 1 flit of overhead

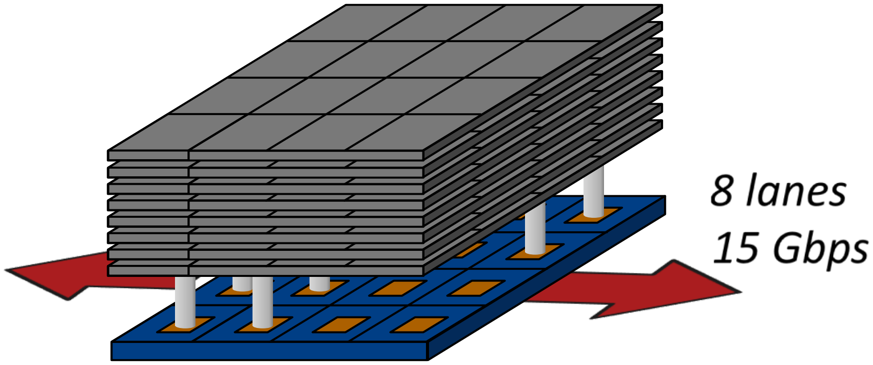


Type	Read		Write	
	Request	Response	Request	Response
Data Size	Empty	1∼8 Flits	1∼8 Flits	Empty
Overhead	1 Flit	1 Flit	1 Flit	1 Flit
Total Size	1 Flit	2∼9 Flits	2∼9 Flits	1 Flit

HMC Communication II

Two/Four full duplex external links:

- Width of 8 or 16 lanes
- Configurable speeds of 10, 12.5, and 15 Gbps

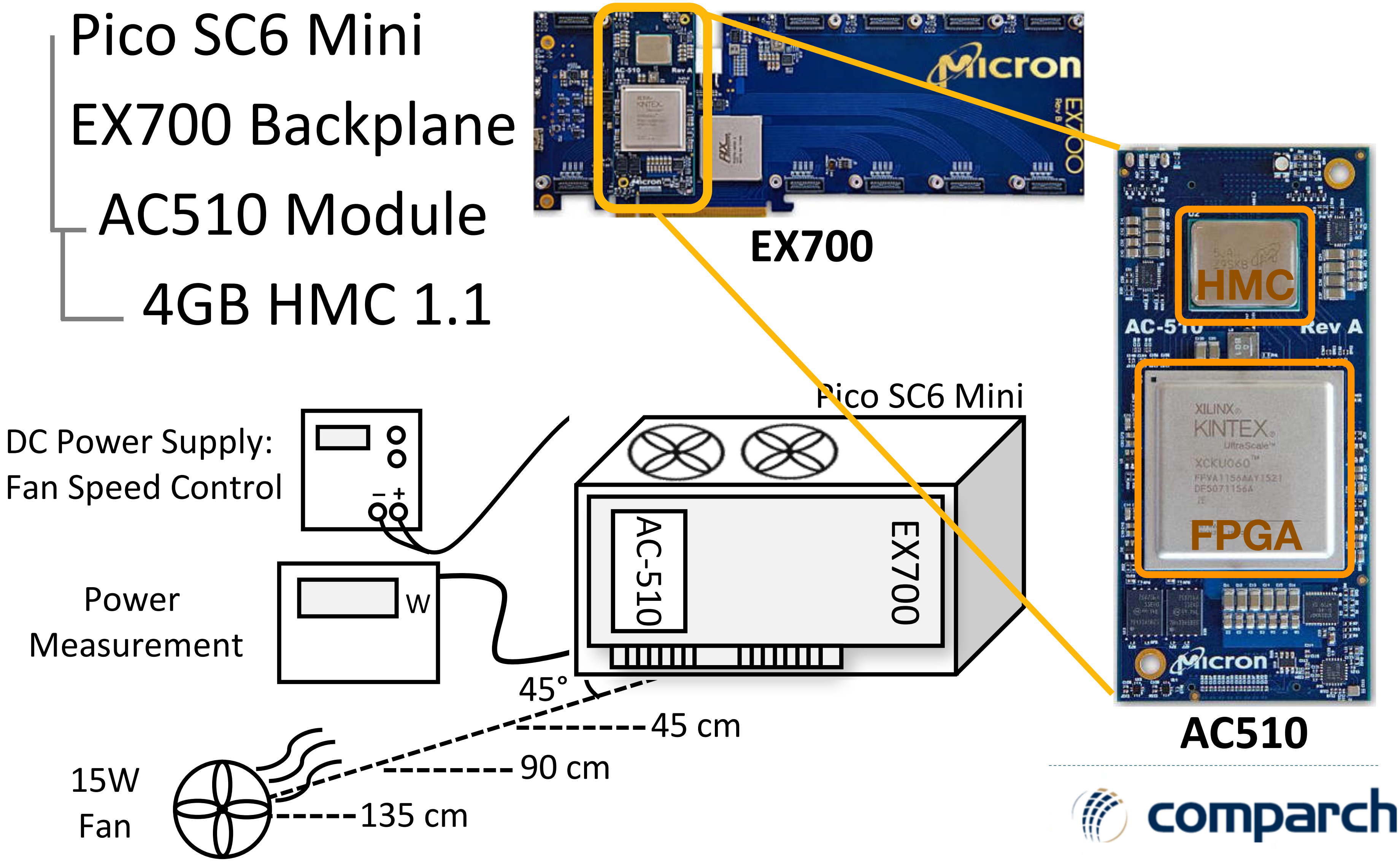


Our evaluated system

2 external links – 8 lanes each

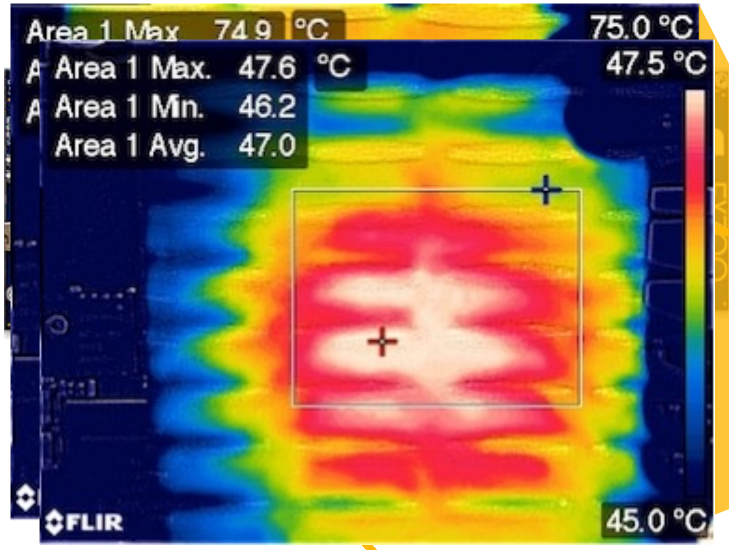
Experimental Setup I

- Pico SC6 Mini
  - EX700 Backplane
    - AC510 Module
      - 4GB HMC 1.1

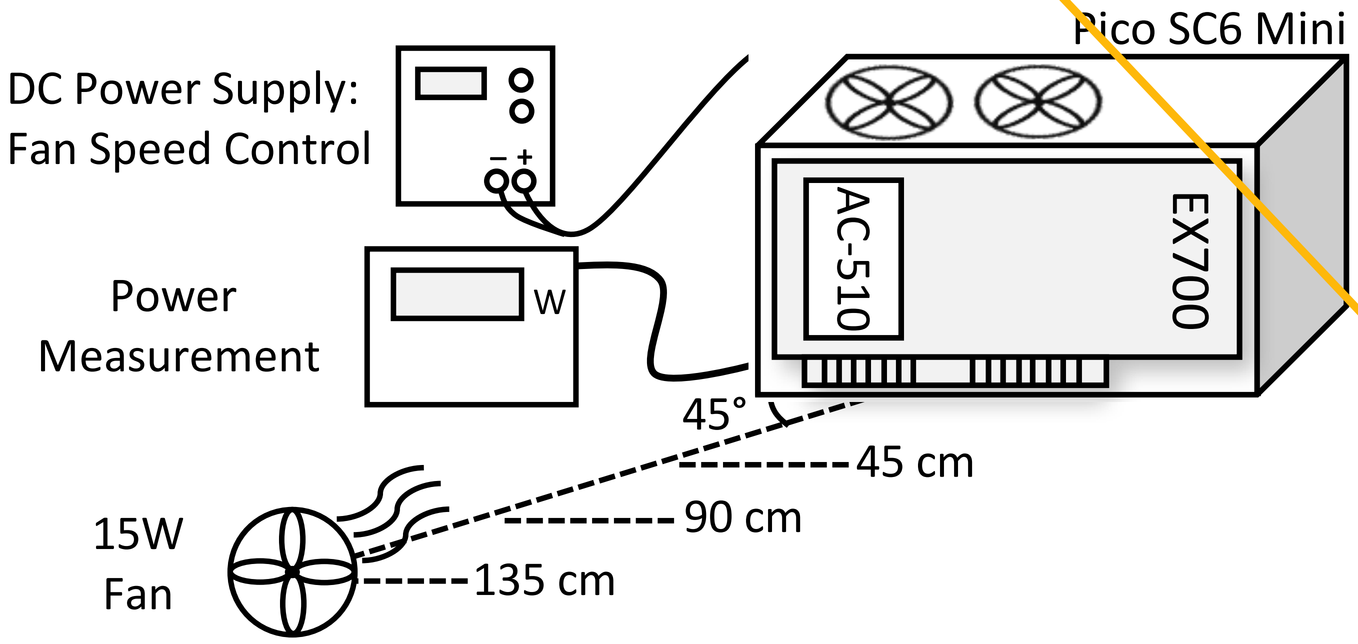


Experimental Setup I

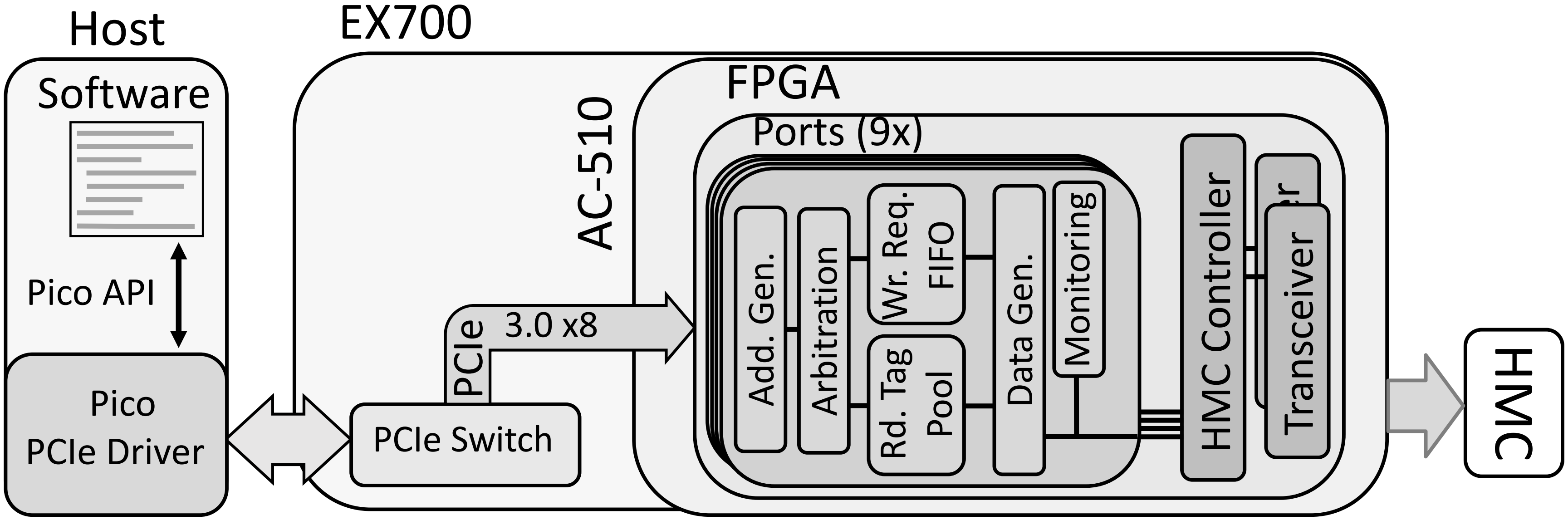
- Pico SC6 Mini
  - EX700 Backplane
    - AC510 Module
      - 4GB HMC 1.1



AC510



Experimental Setup II

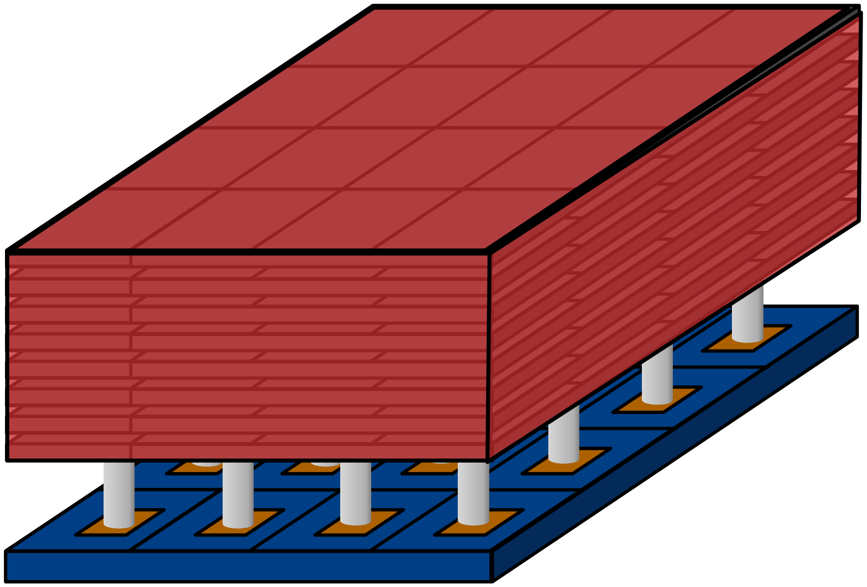


FPGA frequency: 187.5 MHz

Modified GUPS (giga updates per second) benchmark

Apply different masks to addresses

Access Patterns



Accessing Less Banks

16 vaults

8 vaults

4 vaults

2 vaults

1 vault (16 banks)

8 banks

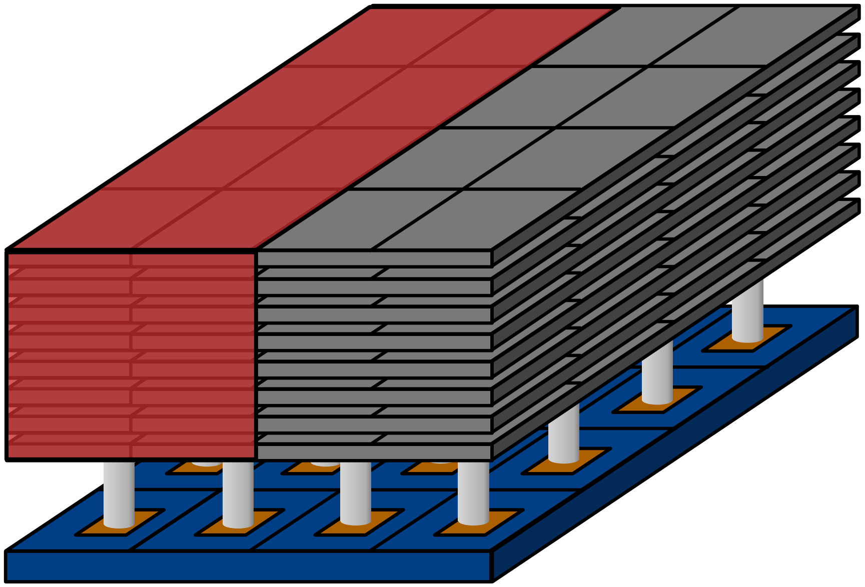
4 banks

2 banks

1 bank

Access Patterns

Access Patterns



Accessing Less Banks

16 vaults

8 vaults

4 vaults

2 vaults

1 vault
(16 banks)

8 banks

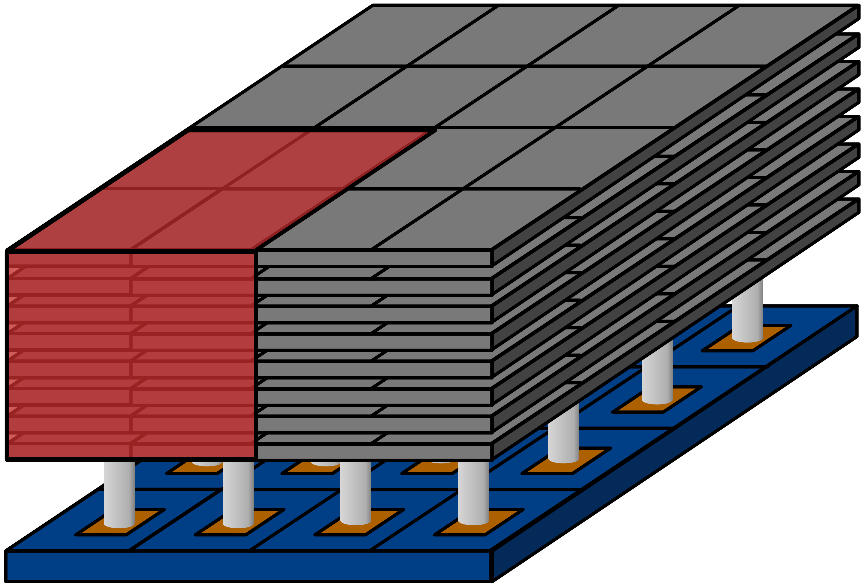
4 banks

2 banks

1 bank

Access Patterns

Access Patterns



Accessing Less Banks

16 vaults

8 vaults

4 vaults

2 vaults

1 vault (16 banks)

8 banks

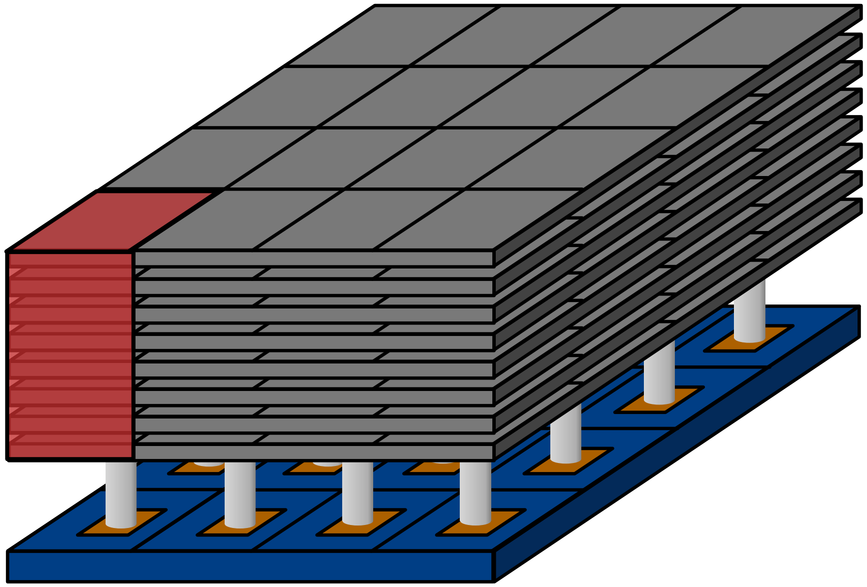
4 banks

2 banks

1 bank

Access Patterns

Access Patterns



Accessing Less Banks

16 vaults

8 vaults

4 vaults

2 vaults

1 vault (16 banks)

8 banks

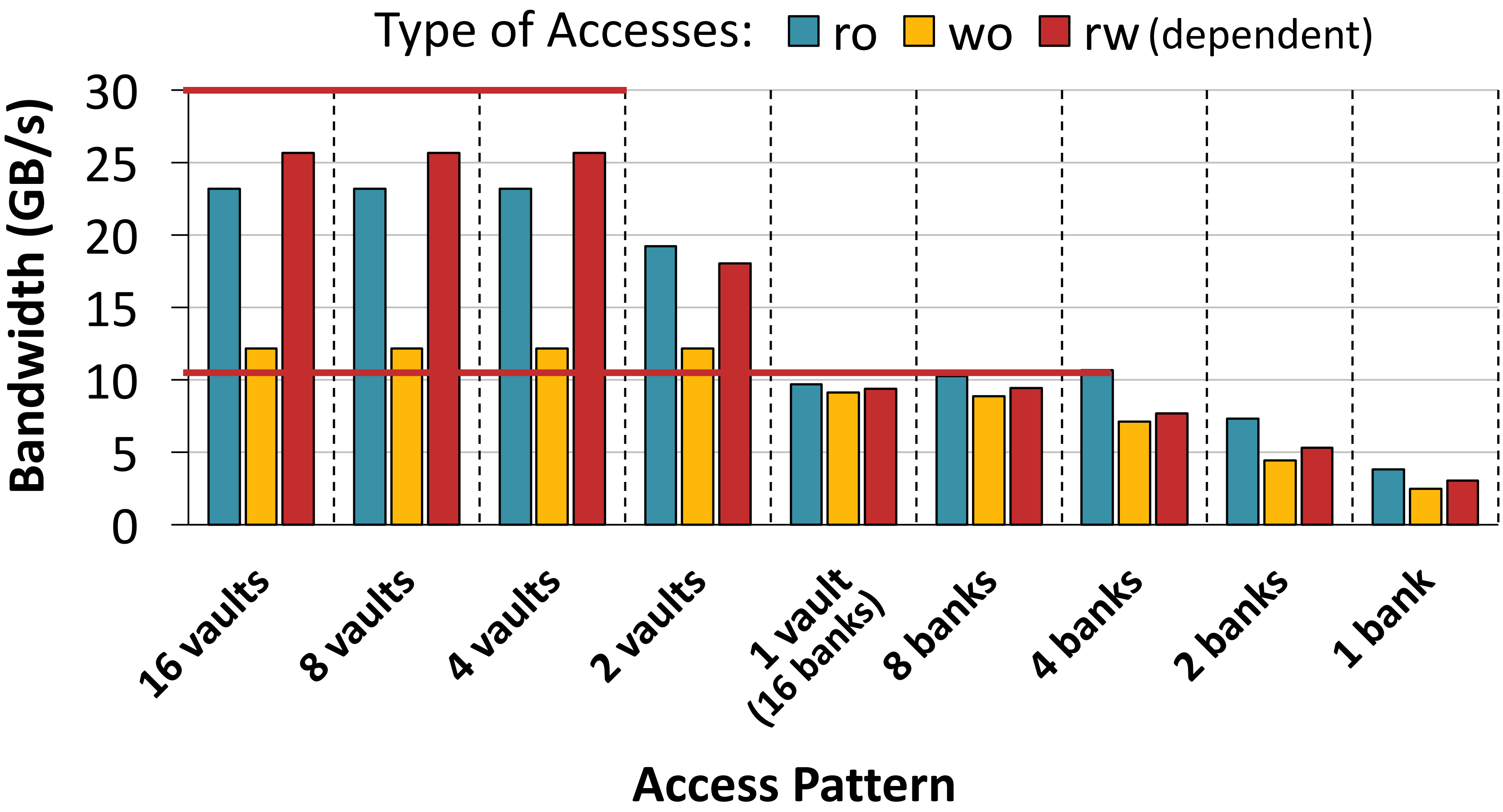
4 banks

2 banks

1 bank

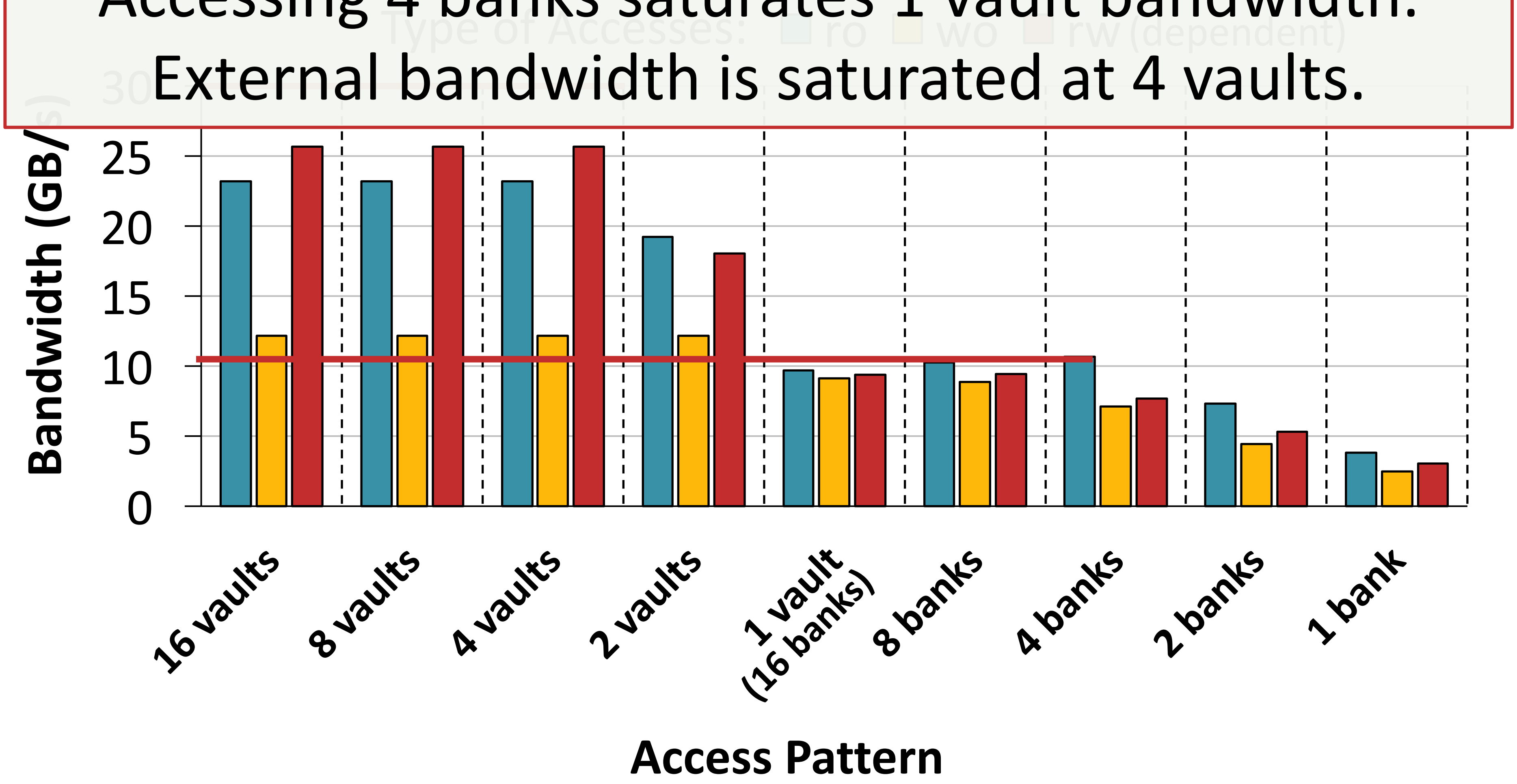
Access Patterns

Bandwidth

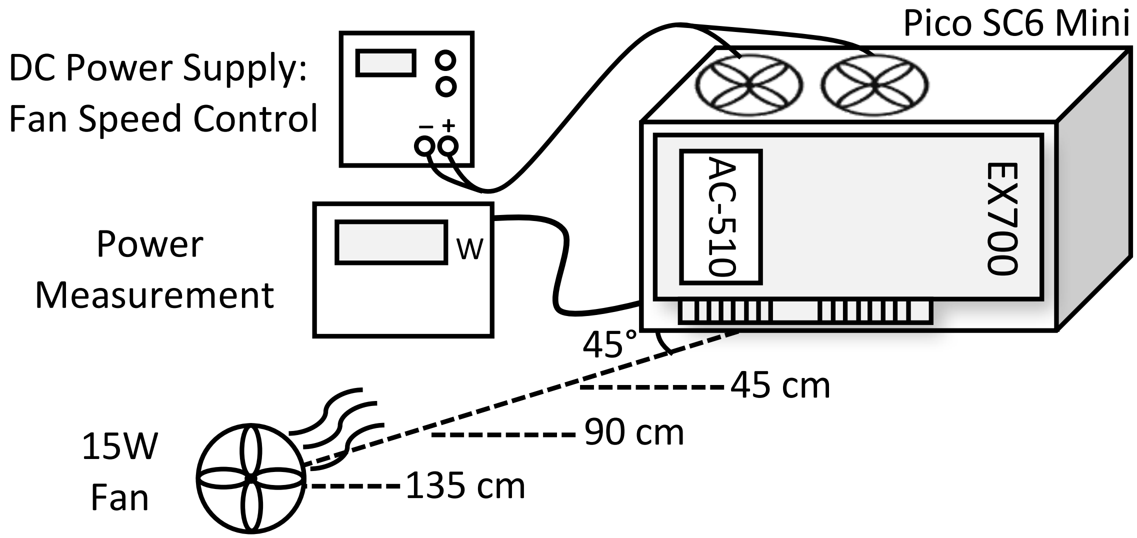

Type of Accesses:
ro
wo
rw (dependent)
Bandwidth (GB/s)
0
5
10
15
20
25
30
16 vaults
8 vaults
4 vaults
2 vaults
1 vault (16 banks)
8 banks
4 banks
2 banks
1 bank
Access Pattern

Bandwidth

Accessing 4 banks saturates 1 vault bandwidth.
External bandwidth is saturated at 4 vaults.

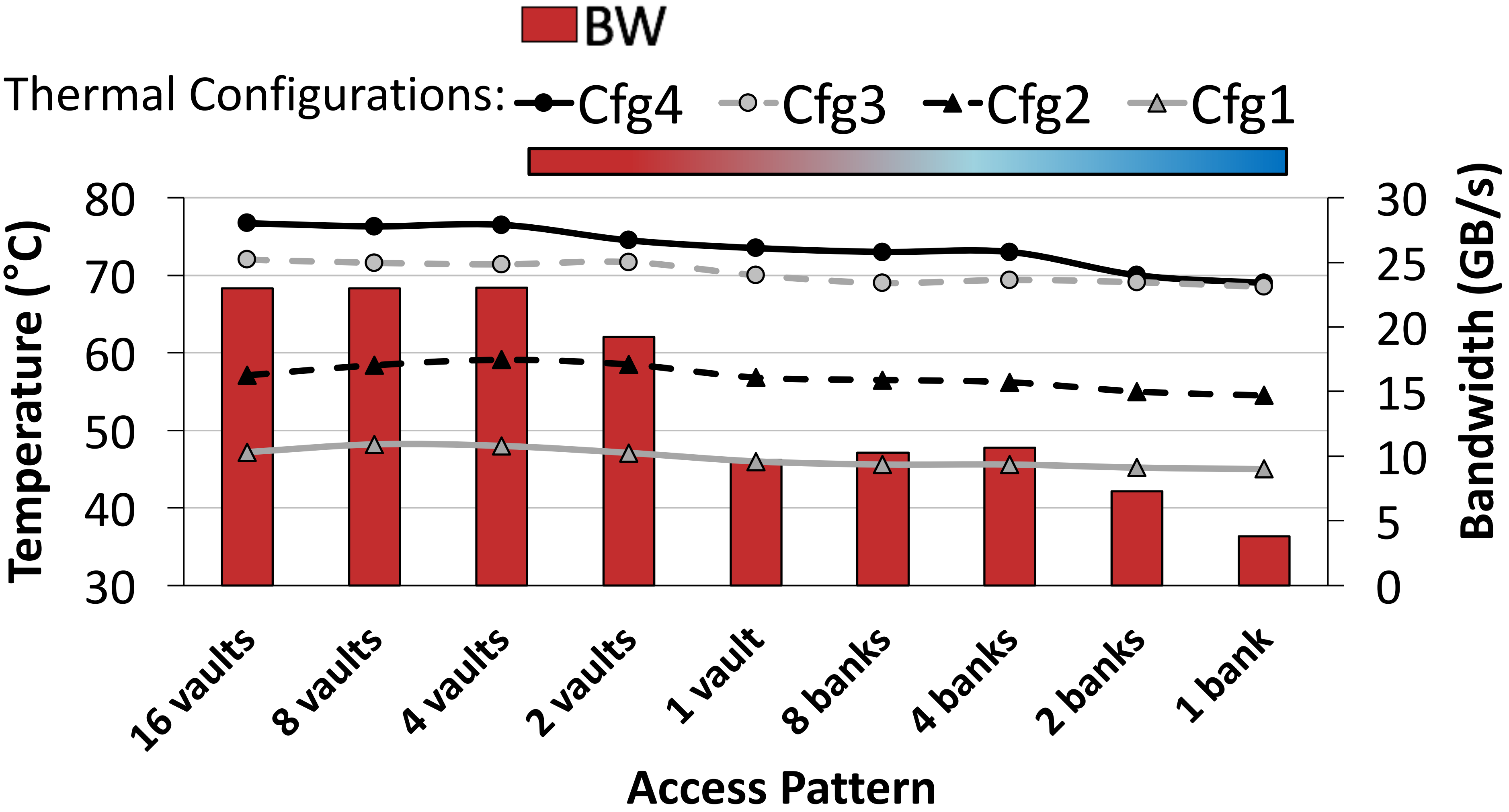


Thermal/Power Experiments

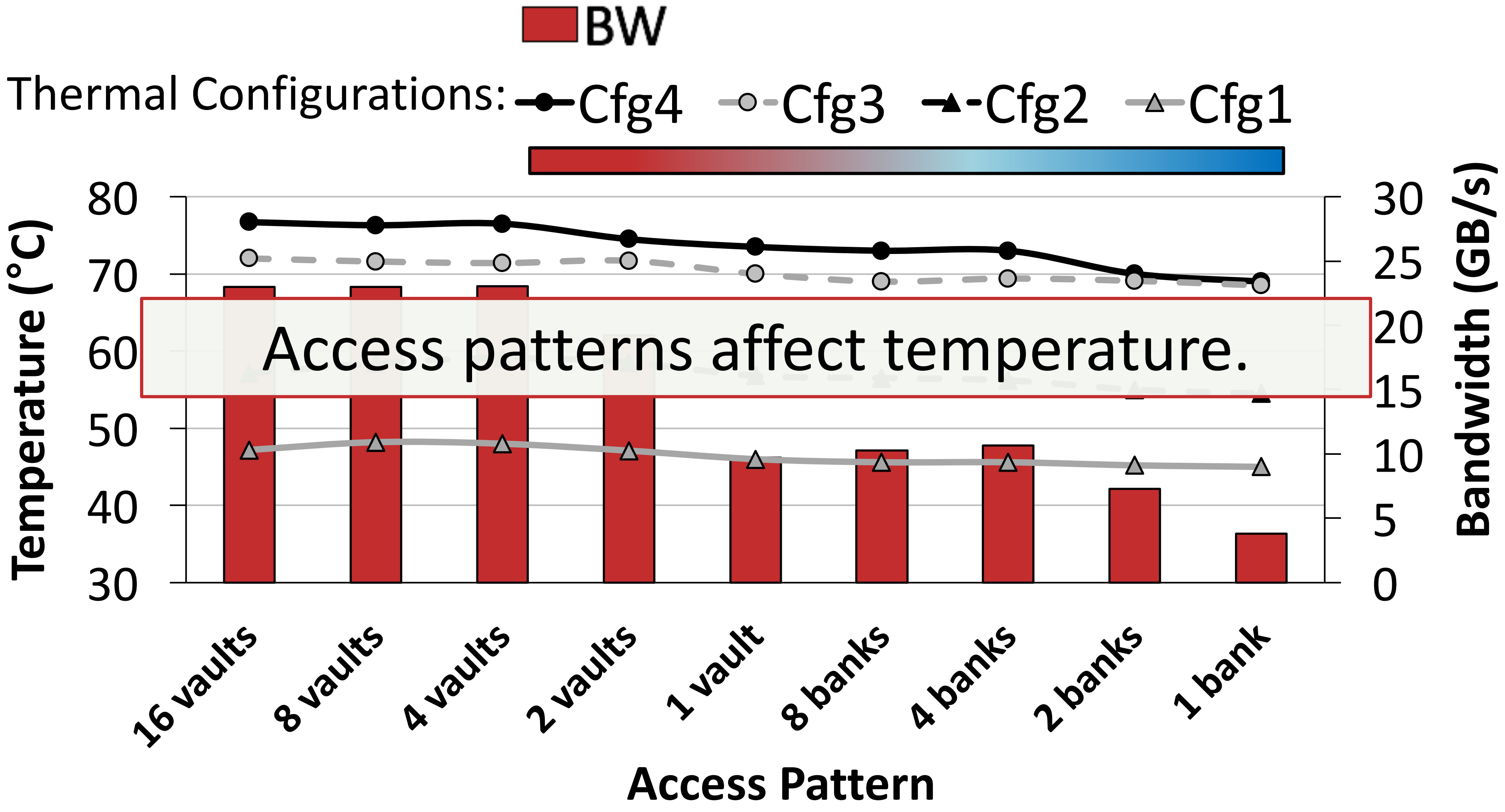


Configuration Name	DC Power Supply: Voltage	DC Power Supply: Current	15 W Fan Distance	Average HMC Idle Temperature
Cfg1	12 V	0.36 A	45 cm	43.1° C
Cfg2	10 V	0.29 A	90 cm	51.7° C
Cfg3	6.5 V	0.14 A	90 cm	62.3° C
Cfg4	6.0 V	0.13 A	135 cm	71.6° C

Temperature (read only)

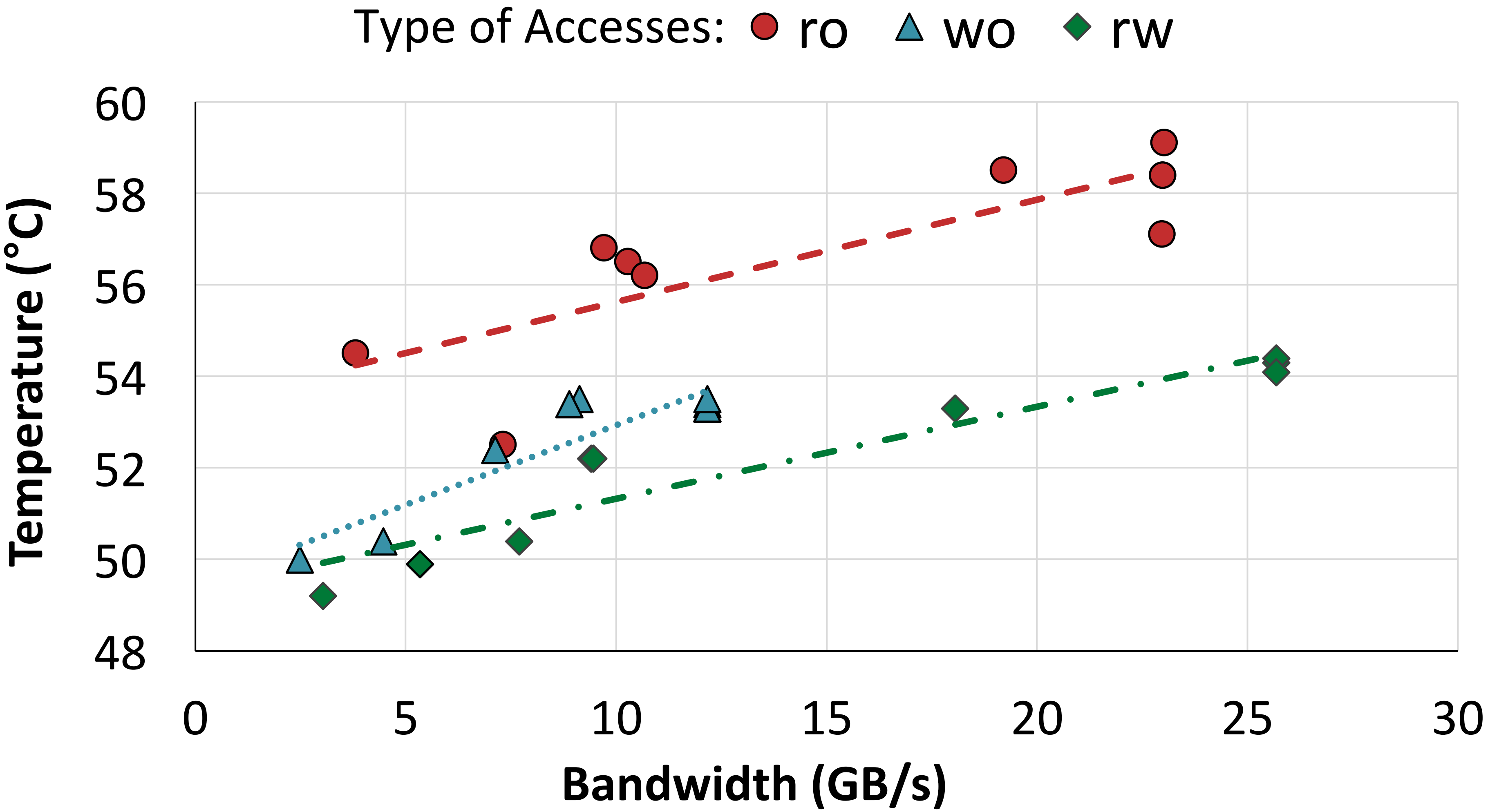


Temperature (read only)

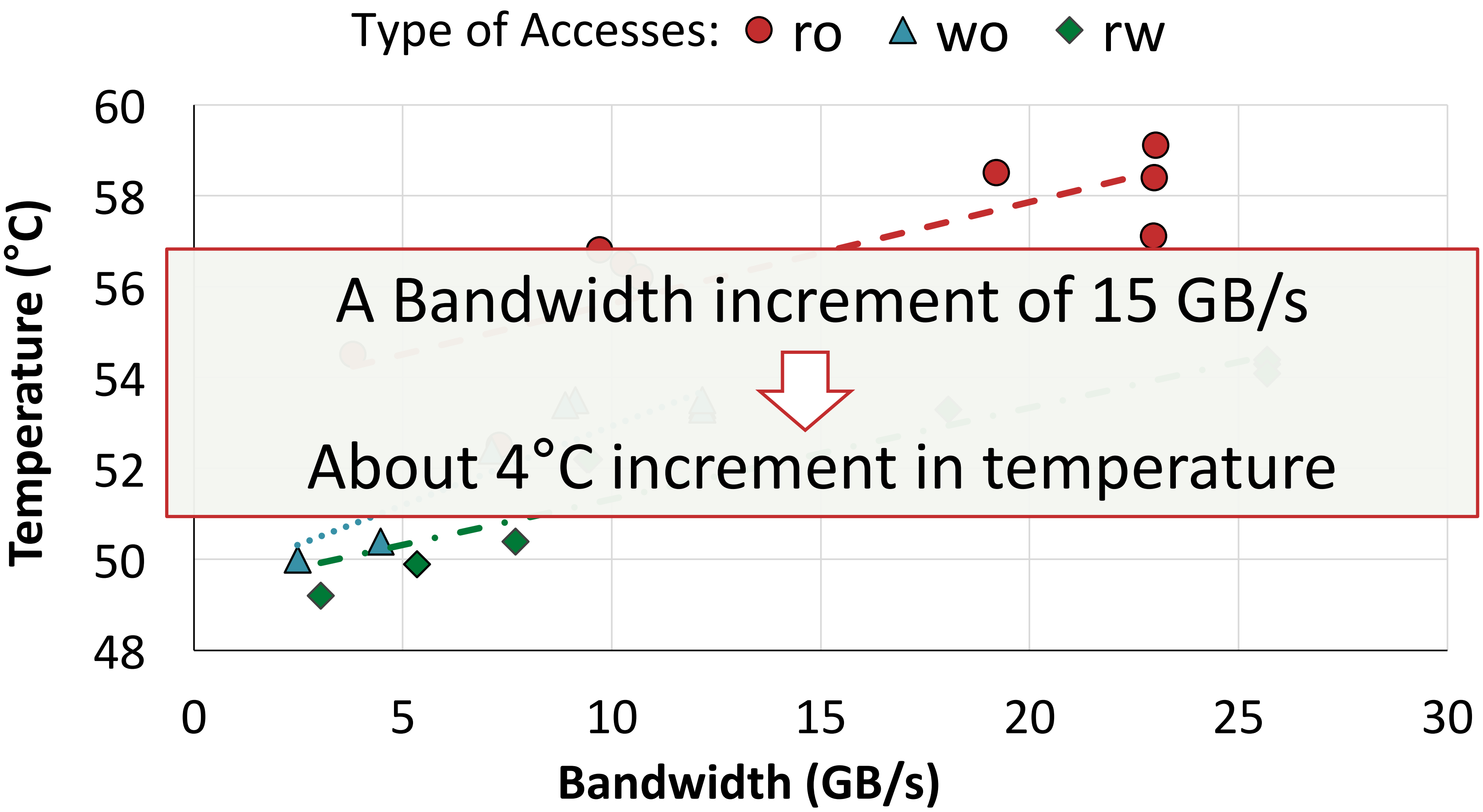


Access patterns affect temperature.

Temperature & Bandwidth



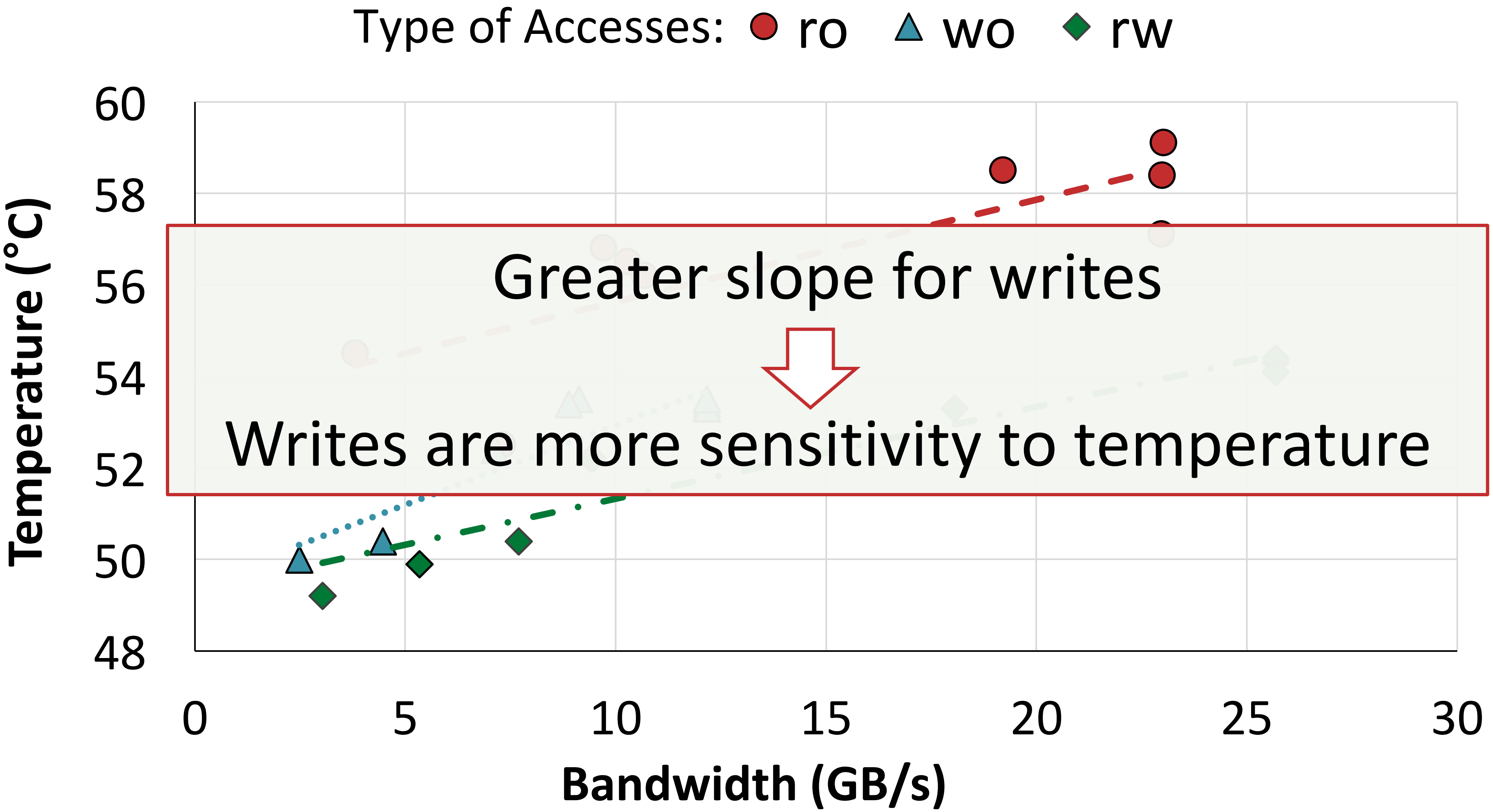
Temperature & Bandwidth



A Bandwidth increment of 15 GB/s

About 4°C increment in temperature

Temperature & Bandwidth

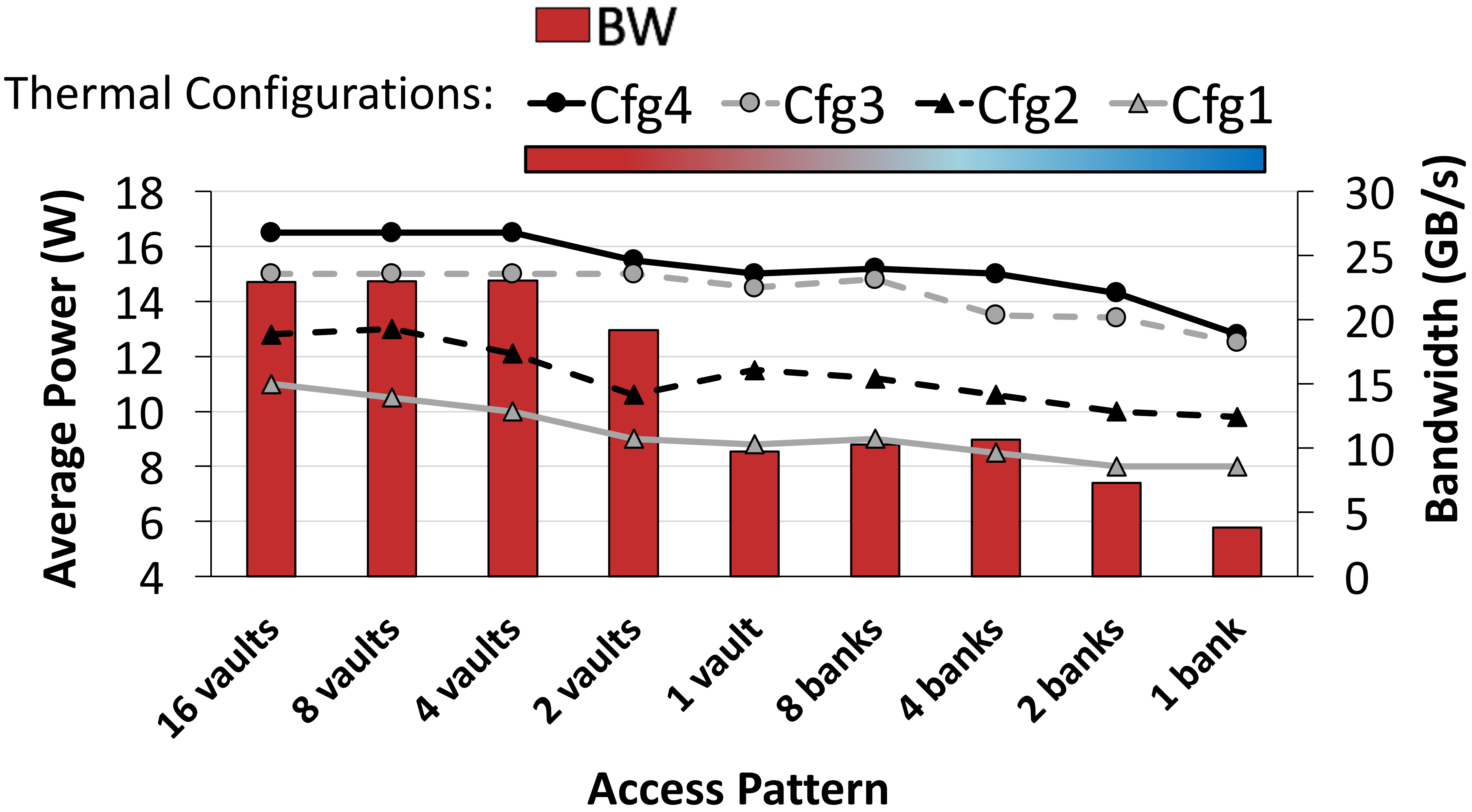


Greater slope for writes

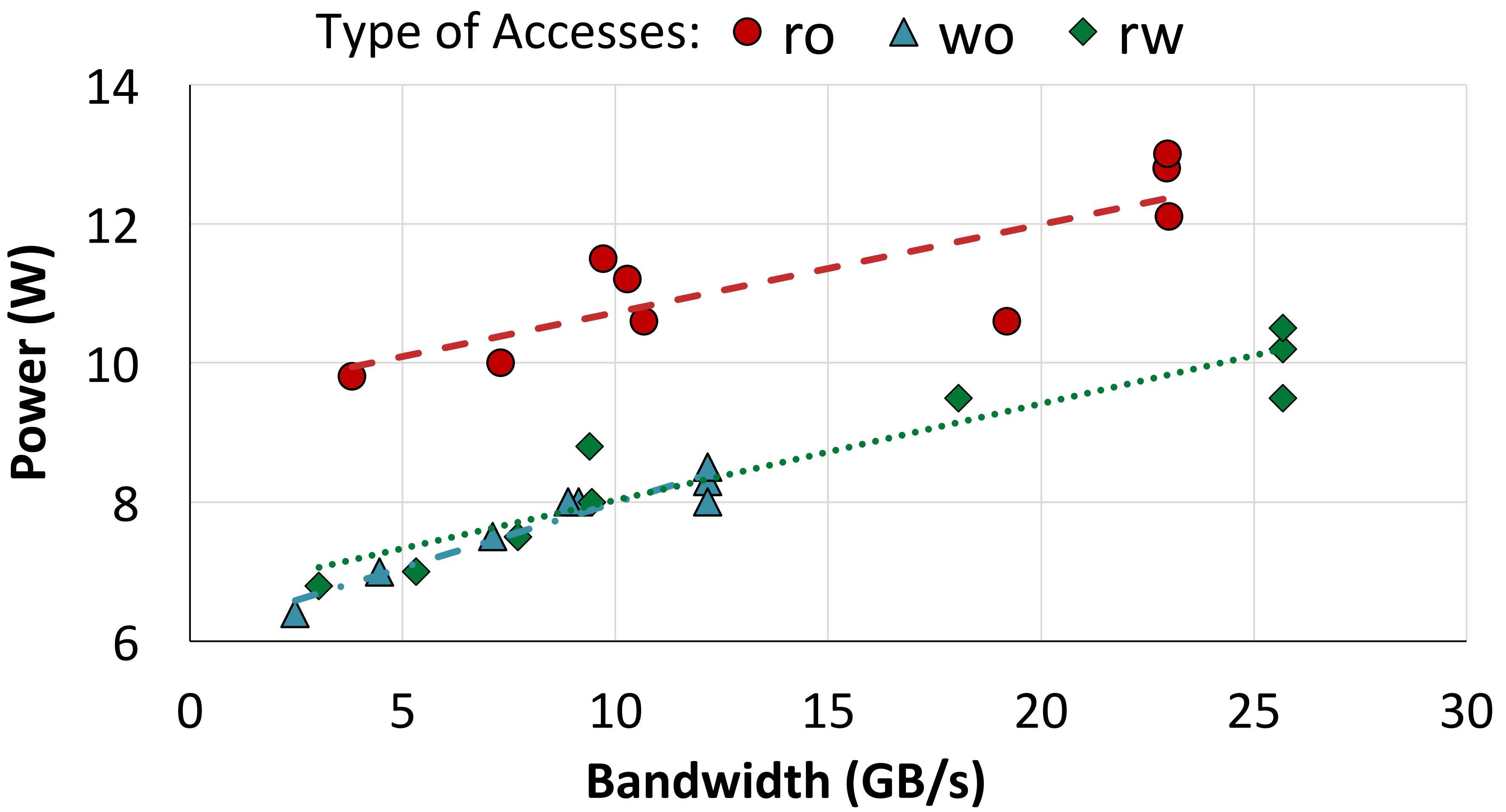
⇩

Writes are more sensitivity to temperature

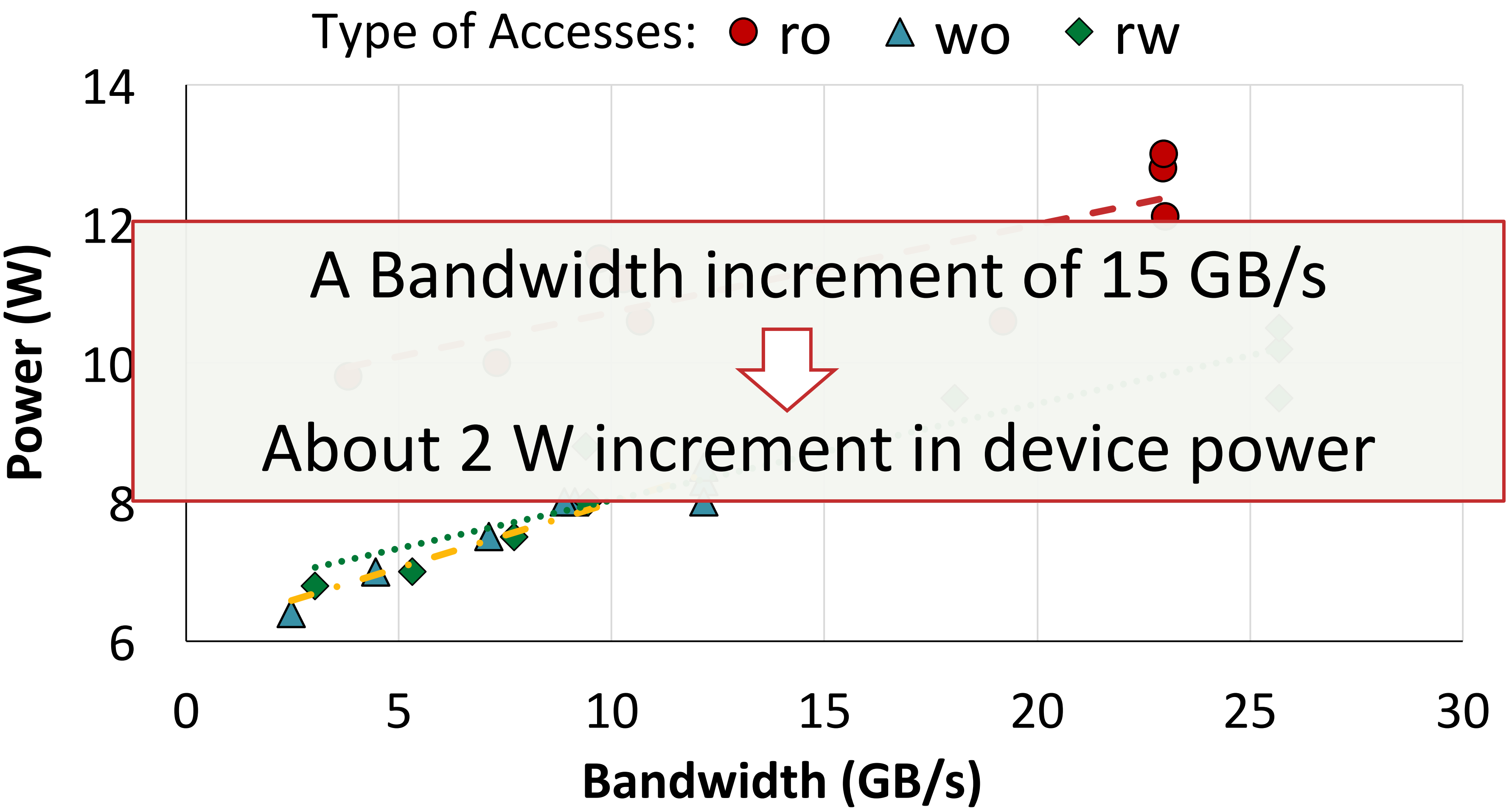
Device Power Consumption (read only)



Device Power & Bandwidth



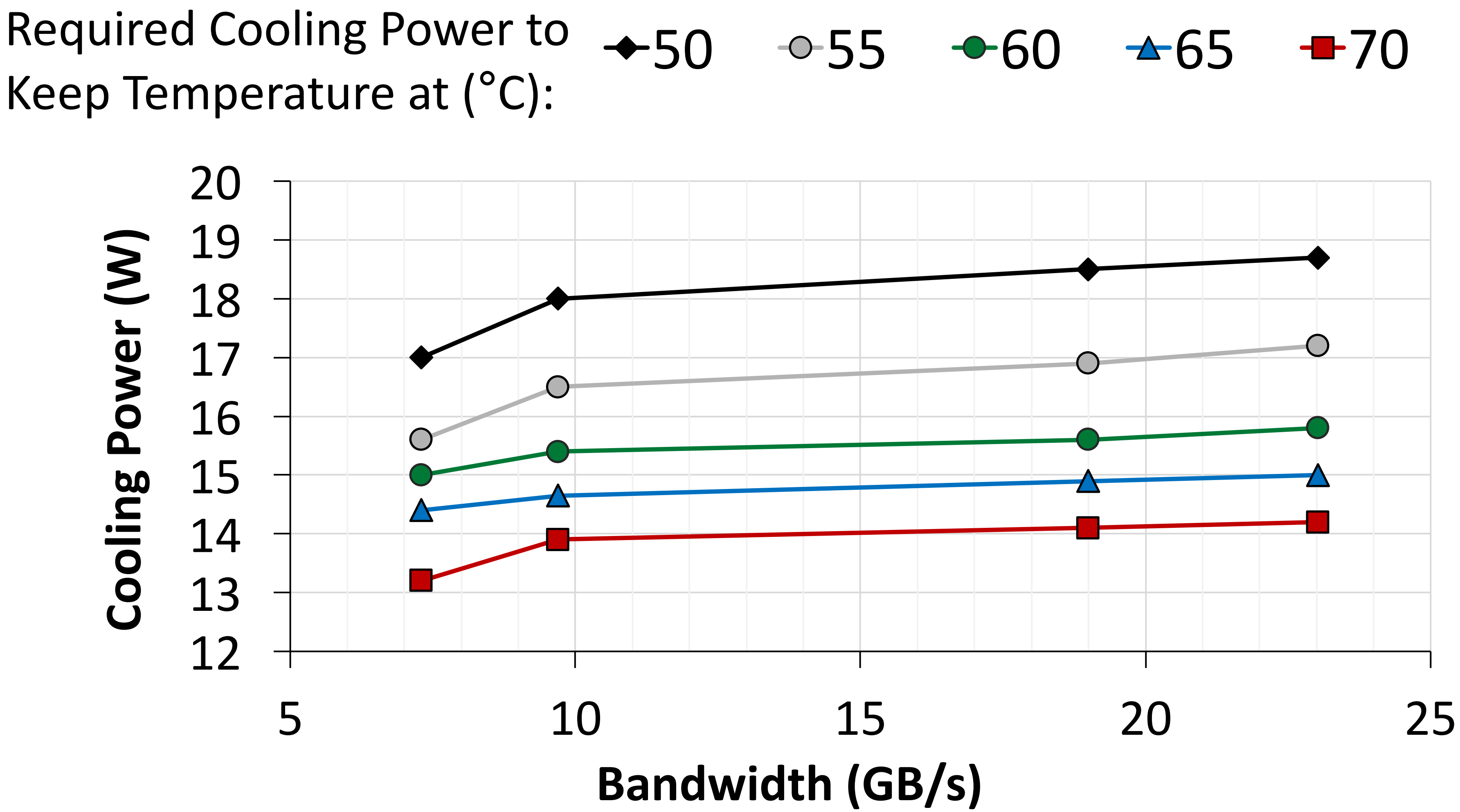
Device Power & Bandwidth



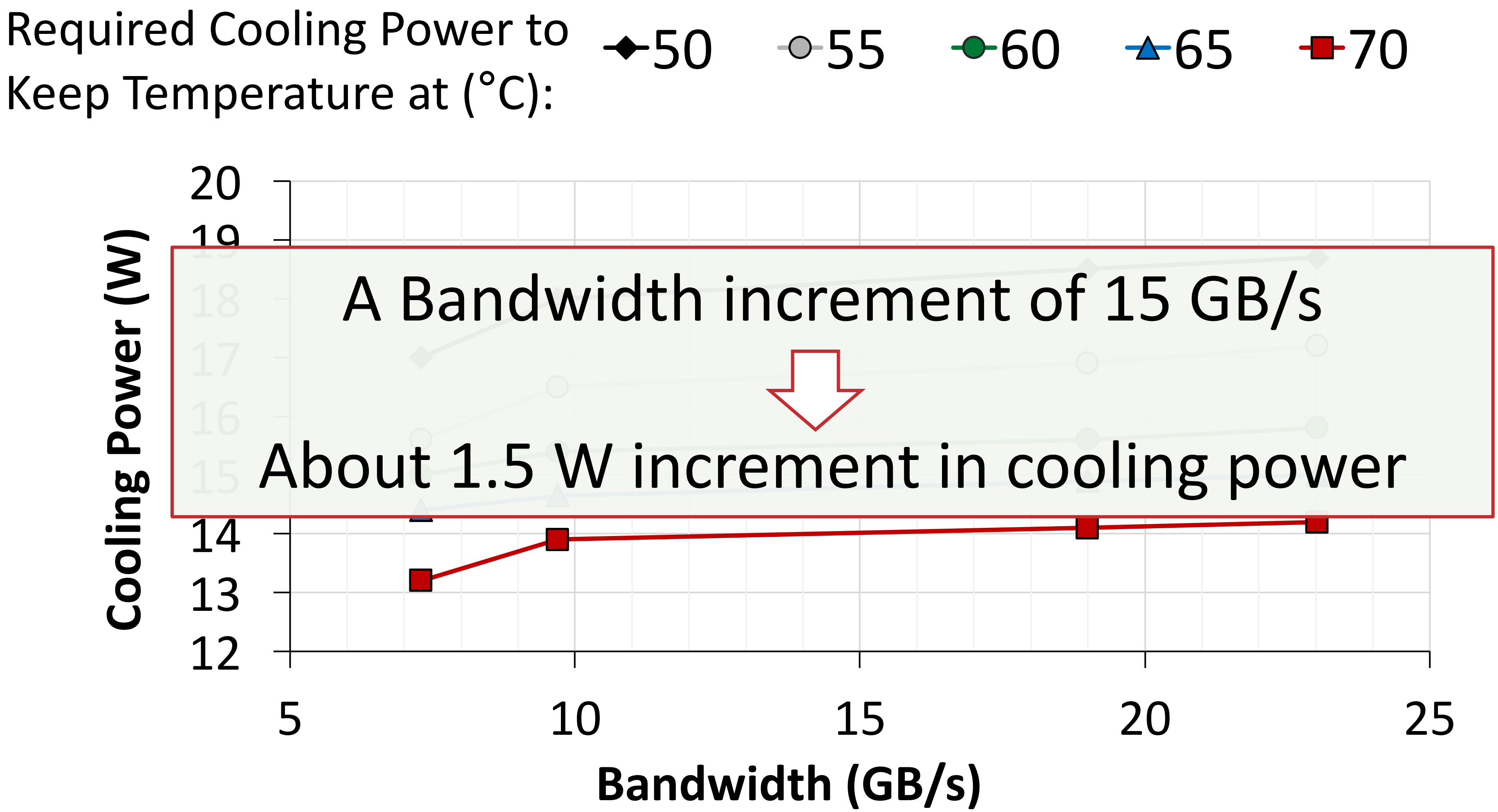
A Bandwidth increment of 15 GB/s

About 2 W increment in device power

Cooling Power Consumption (read only)



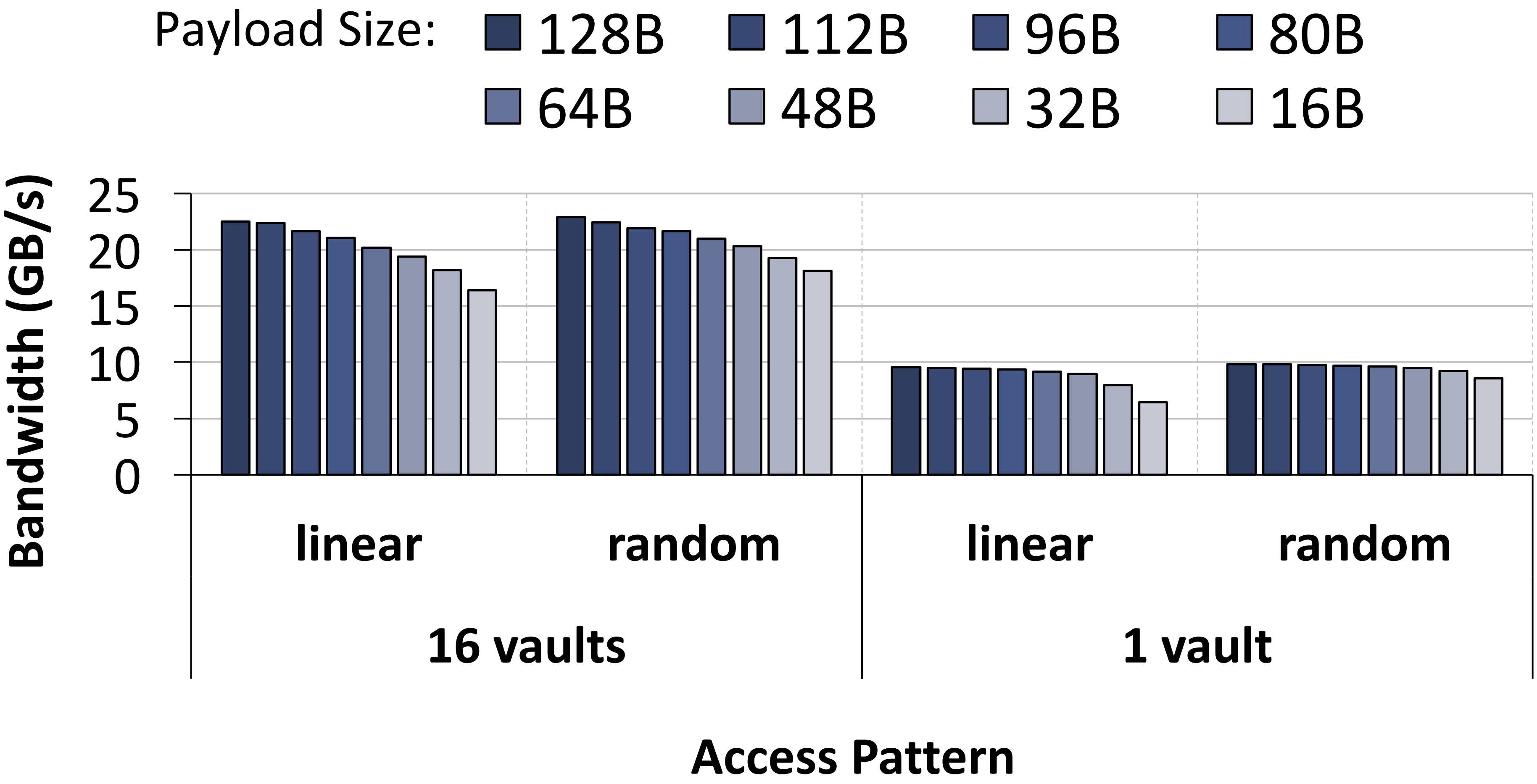
Cooling Power Consumption (read only)



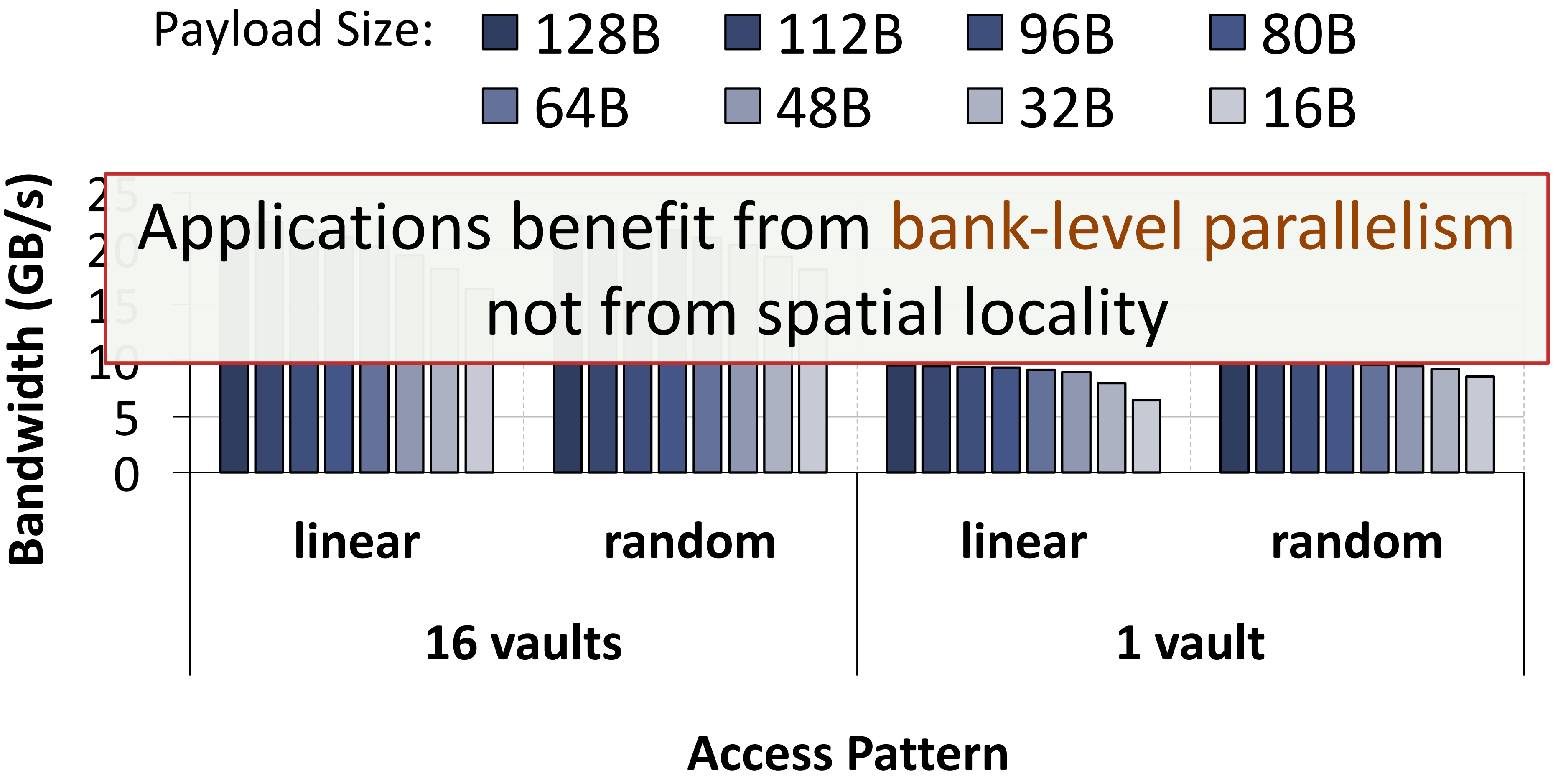
A Bandwidth increment of 15 GB/s

About 1.5 W increment in cooling power

Closed-Page Policy



Closed-Page Policy

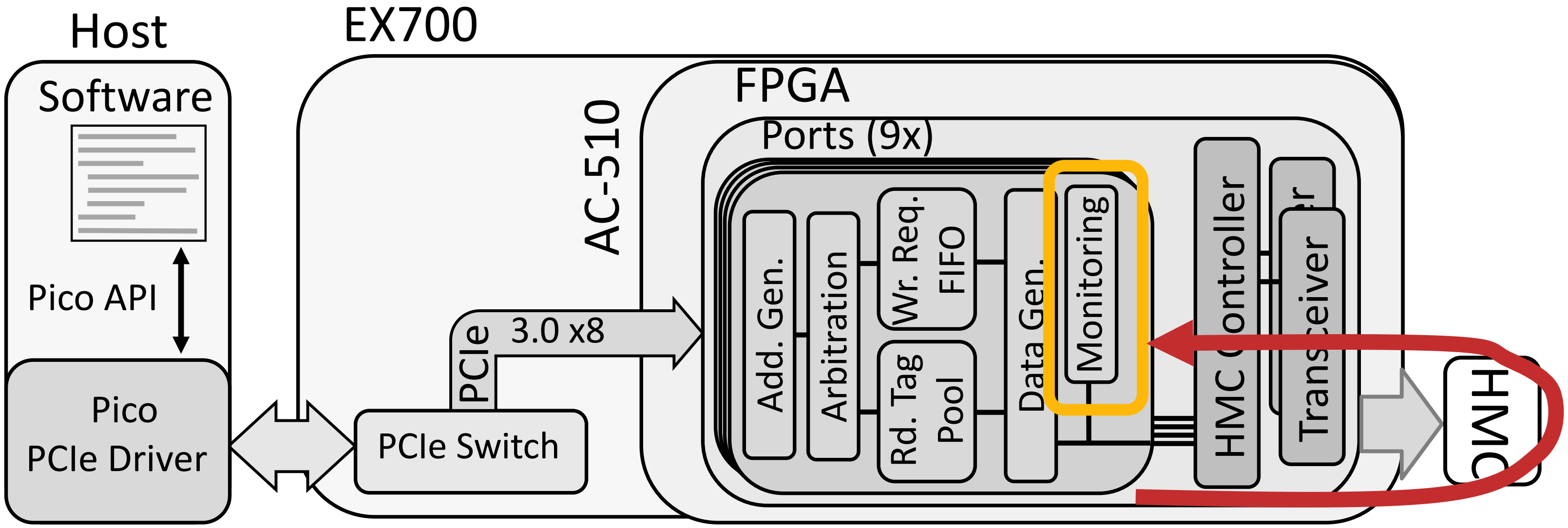


Applications benefit from bank-level parallelism not from spatial locality

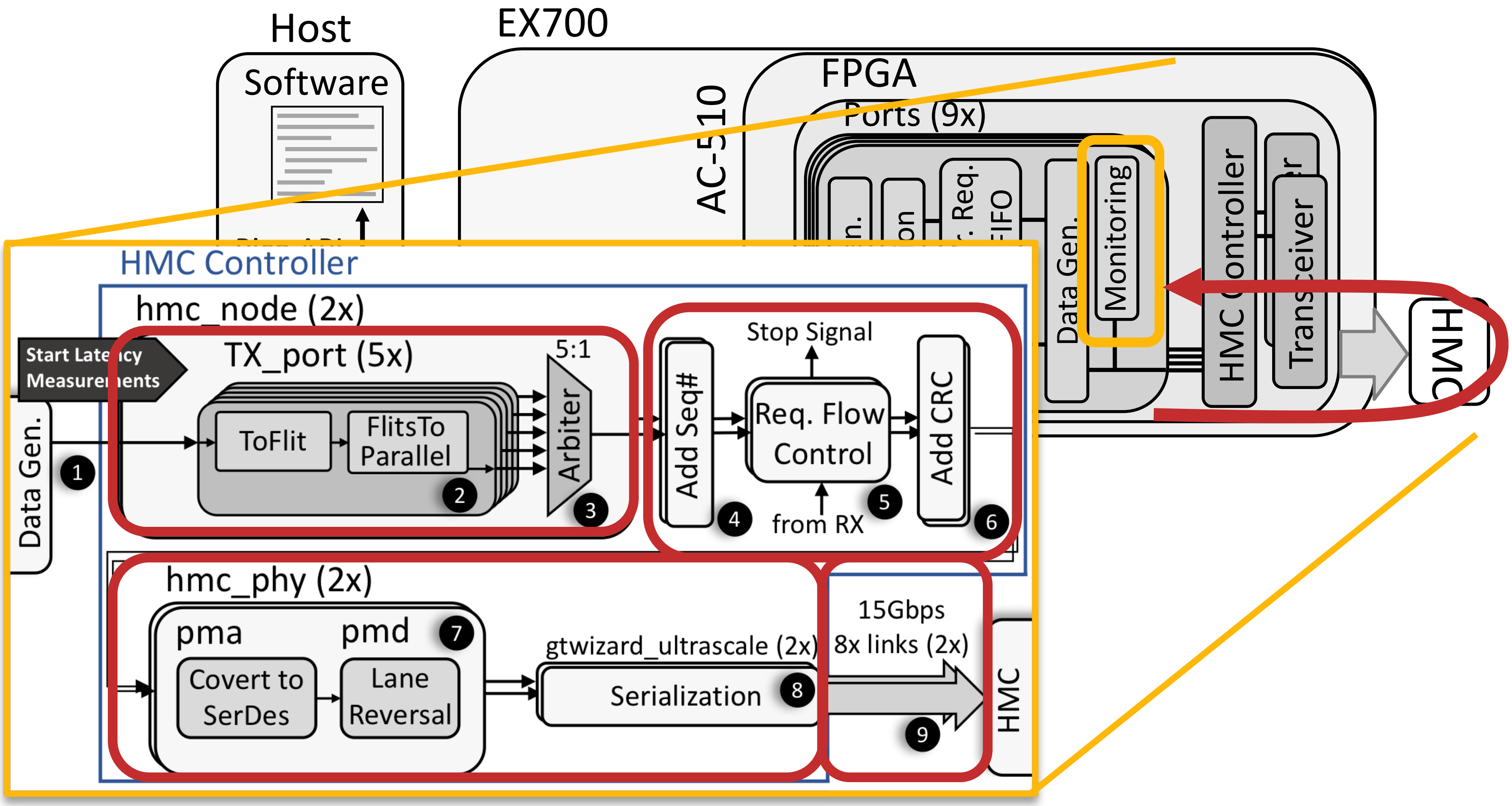
Achieving High Bandwidth

- Promote bank-level parallelism
- Remap data to avoid internal organization bottlenecks
- Concatenate requests to use bandwidth effectively

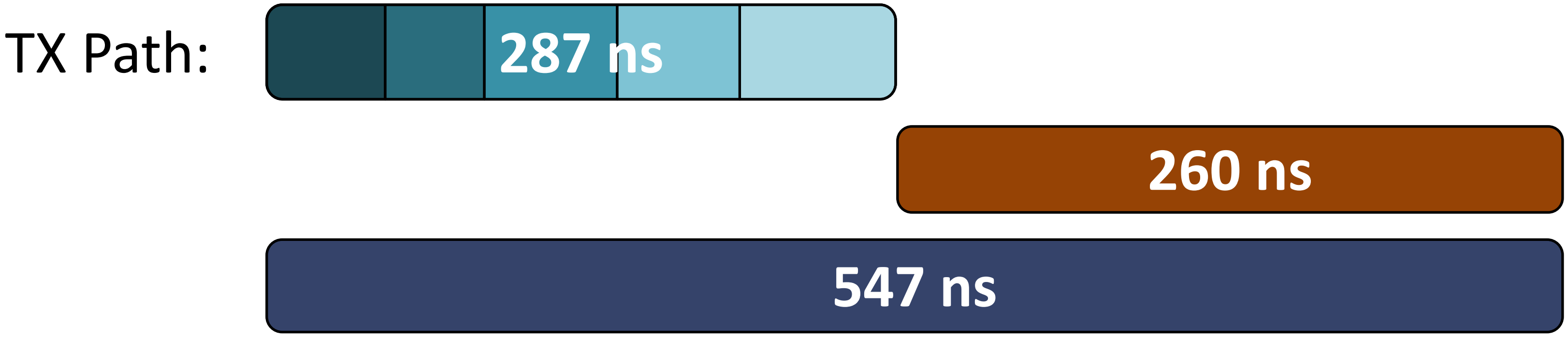
Latency Deconstruction



Latency Deconstruction



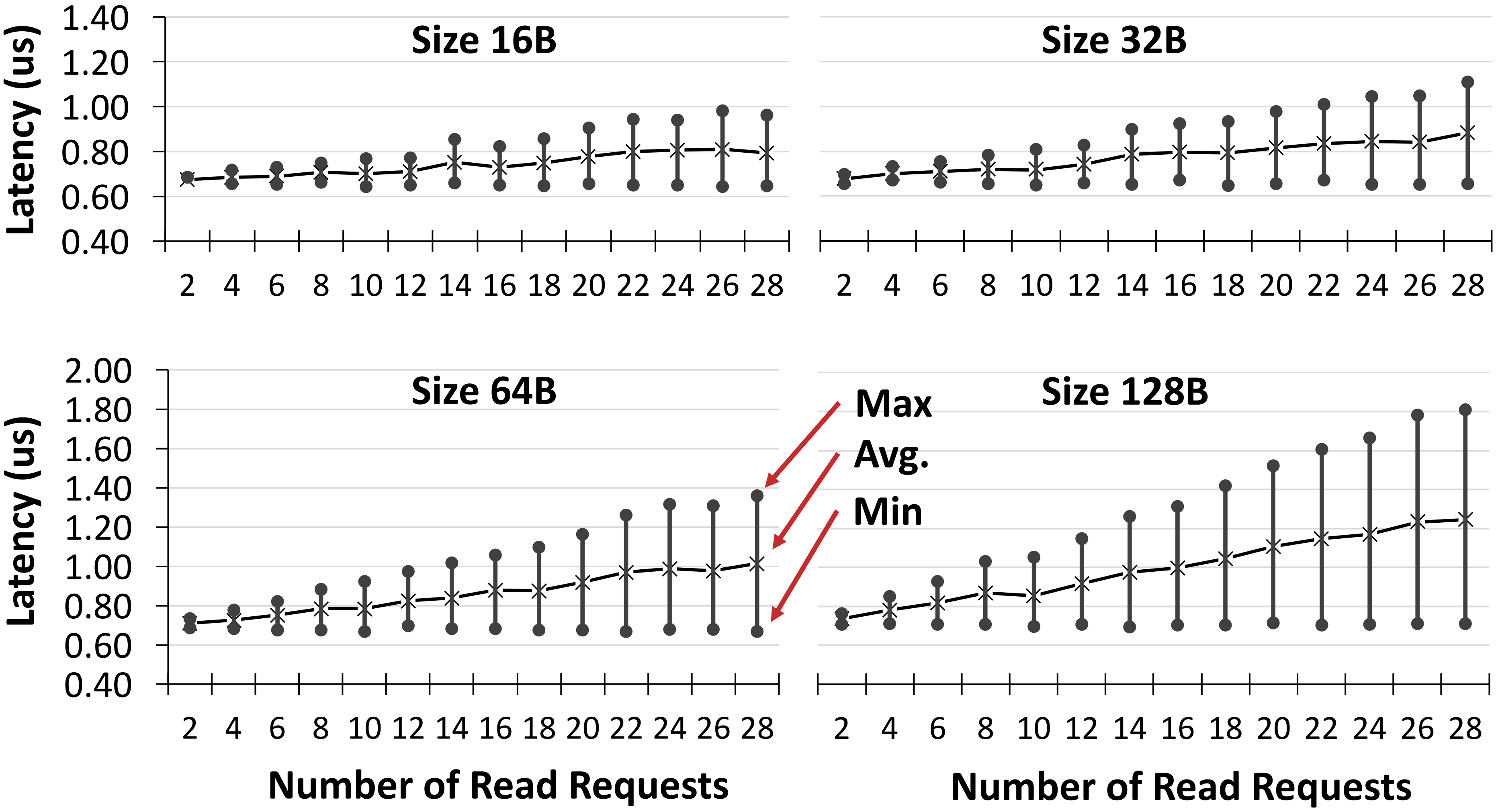
Latency Deconstruction Summary



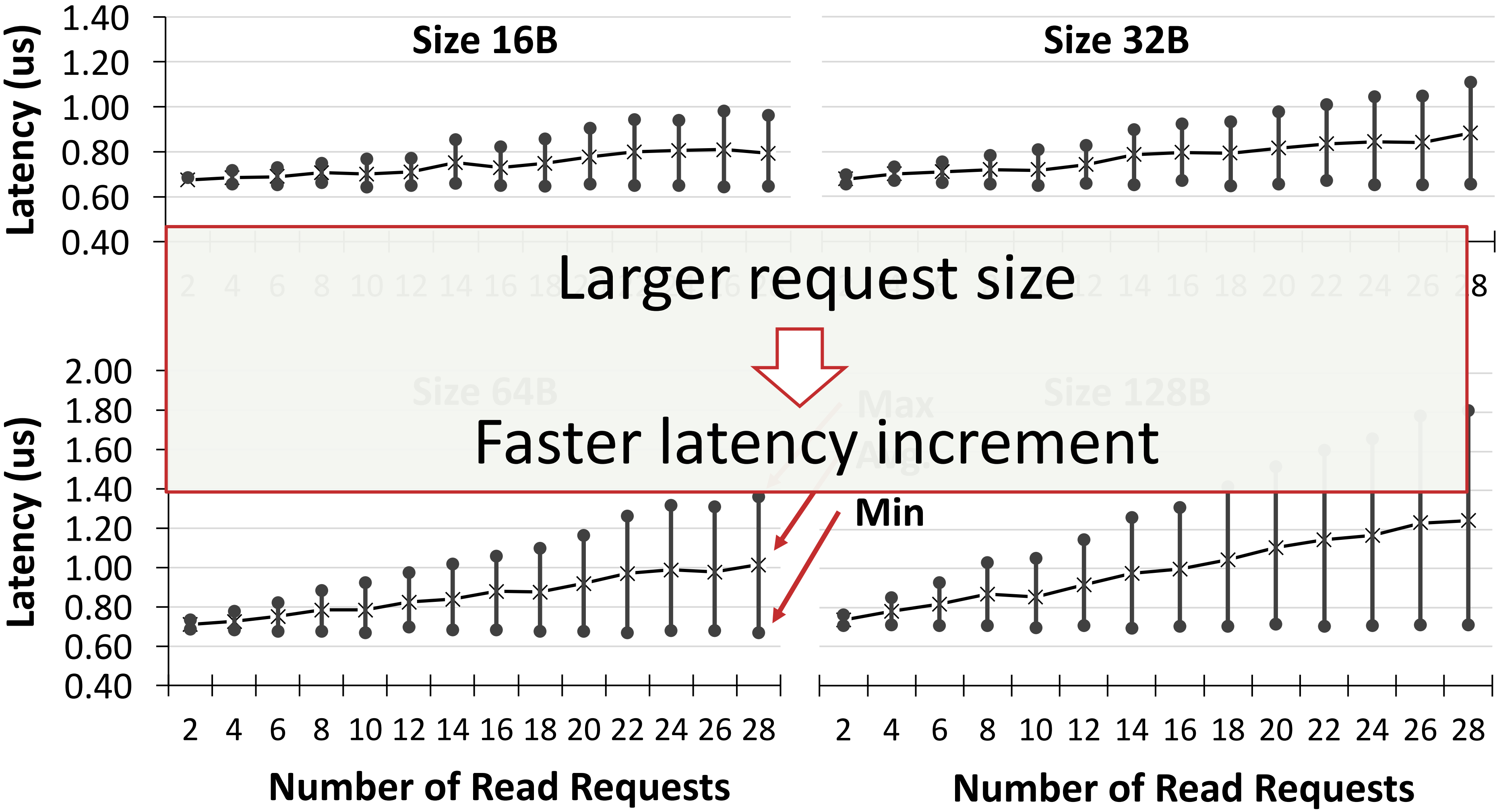
Conversion to flits & buffering	10 cycles
Round-robin arbitration among ports	2-9 cycles
Add packet fields & flow control	10 cycles
Serialization	10 cycles
Transmission (128B)	15 cycles

Freq.: 187.5 MHz
Cycle: 5.3 ns

Low-Load Latency



Low-Load Latency

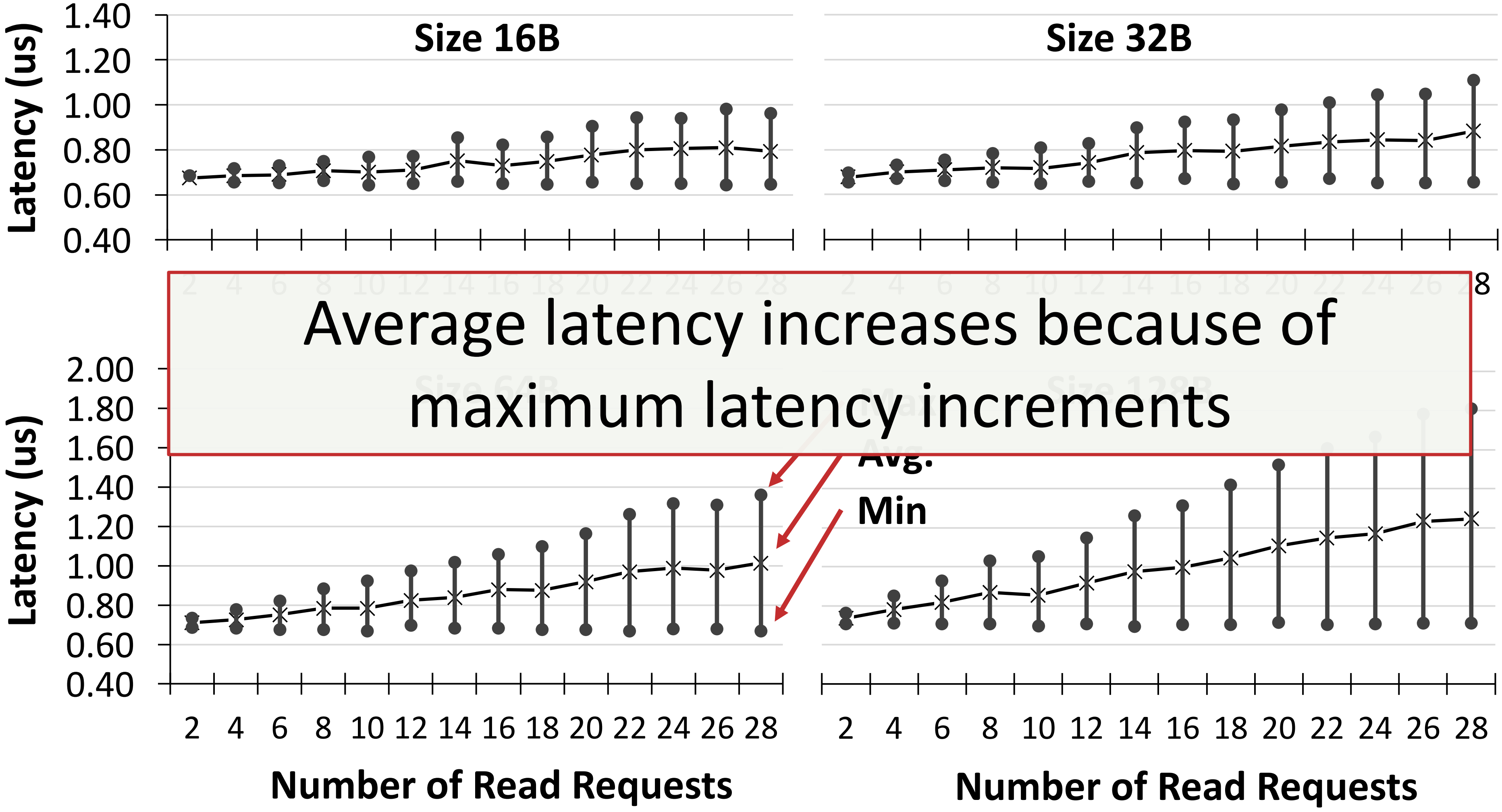


Larger request size

⇩

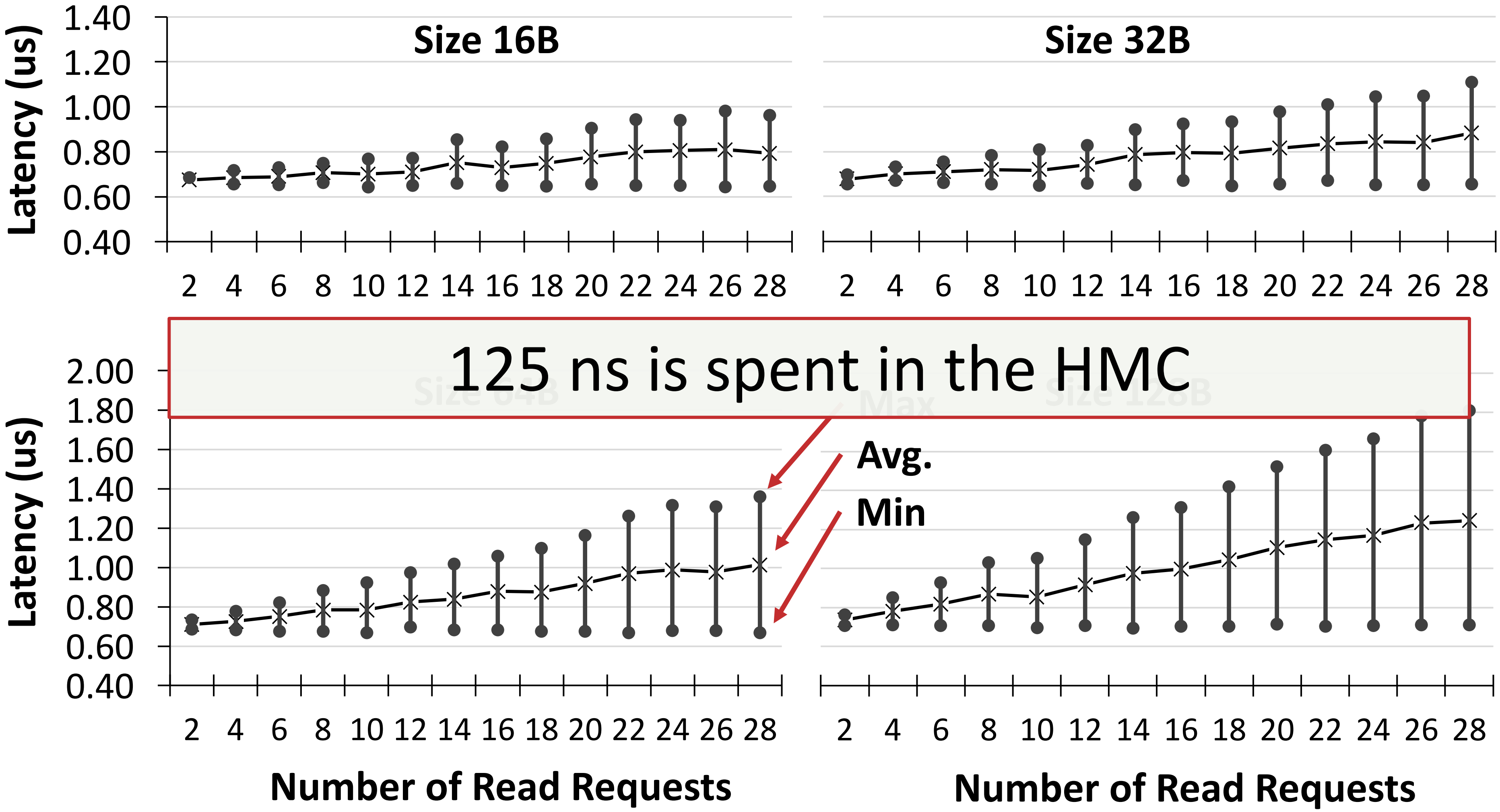
Faster latency increment

Low-Load Latency



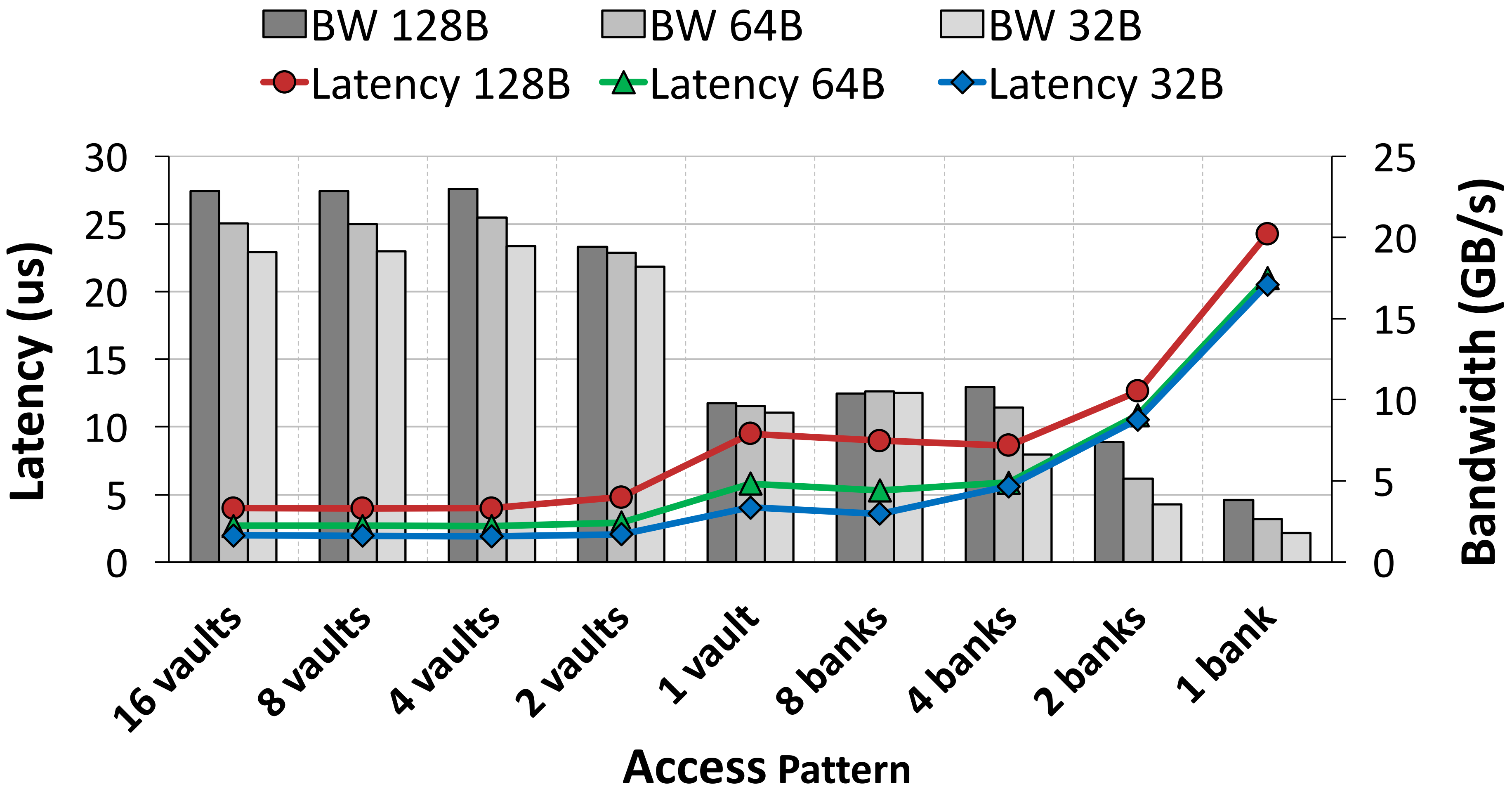
Average latency increases because of maximum latency increments

Low-Load Latency

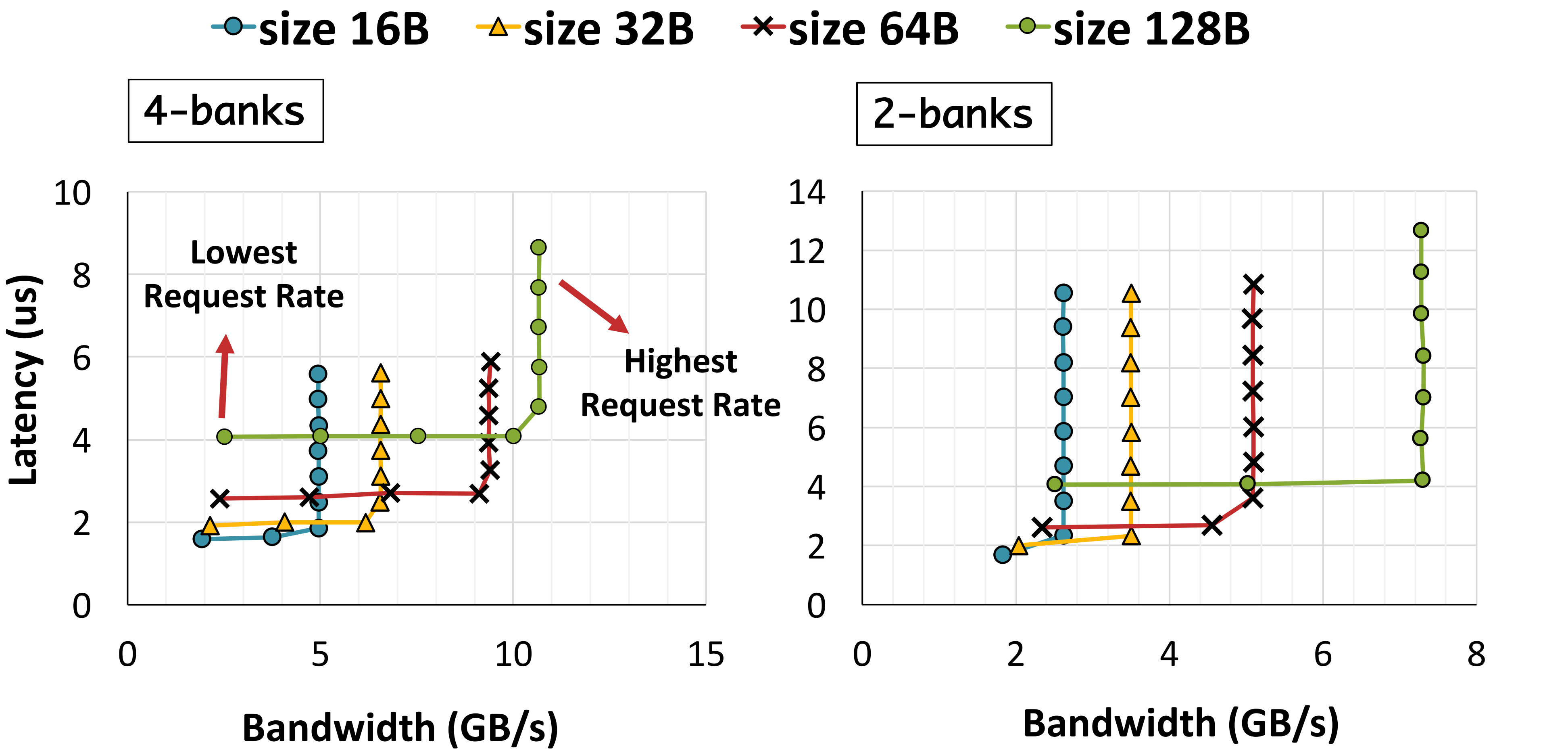


125 ns is spent in the HMC

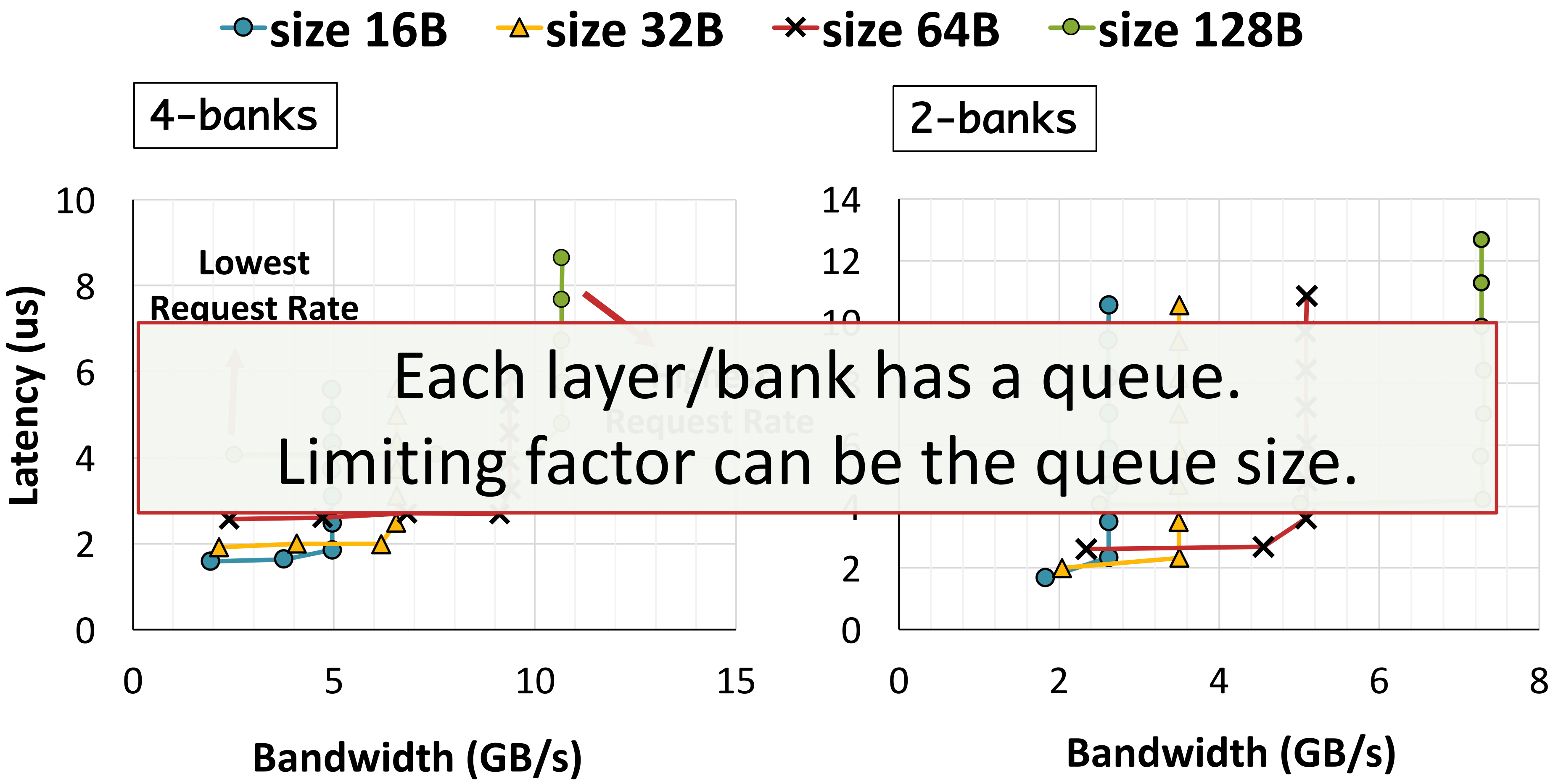
High-Load Latency



Latency-Bandwidth



Latency-Bandwidth



Each layer/bank has a queue.
Limiting factor can be the queue size.

Conclusions

- Mixing read and write requests and using large request sizes lead to effective use of bi-directional bandwidth.
- Distributing accesses prevents internal bottlenecks and exploits bank-level parallelism.
- Controlling the request rate to avoid high latency.
- Employing fault-tolerant mechanisms and using proper cooling solutions enables temperature-sensitive operations to reach a higher bandwidth.
- Reducing latency overhead of the infrastructure will greatly benefit latency.

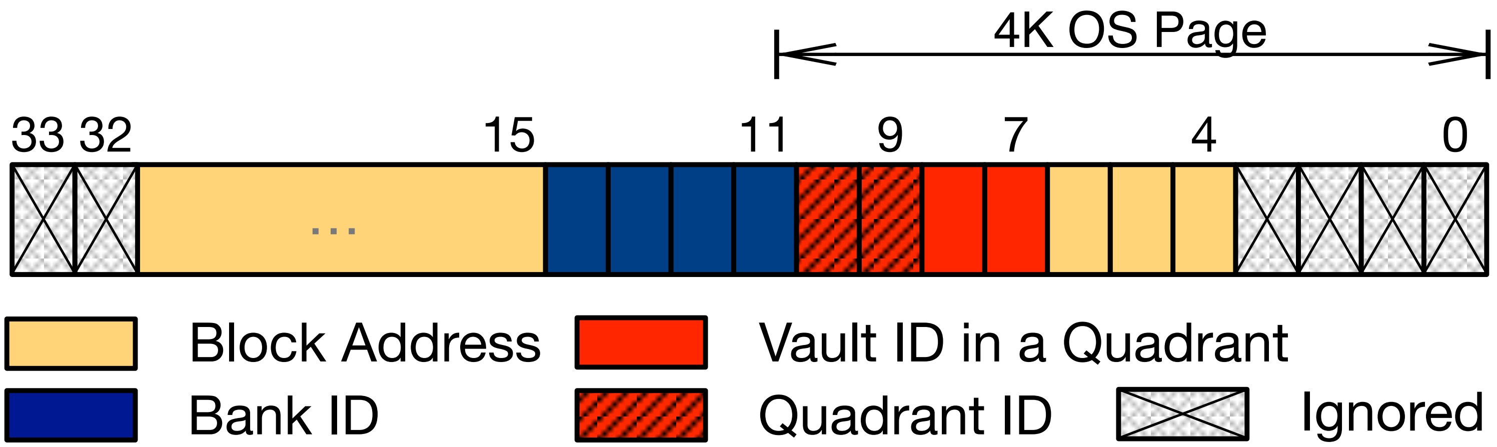
Backup Slides

HMC Memory Addressing

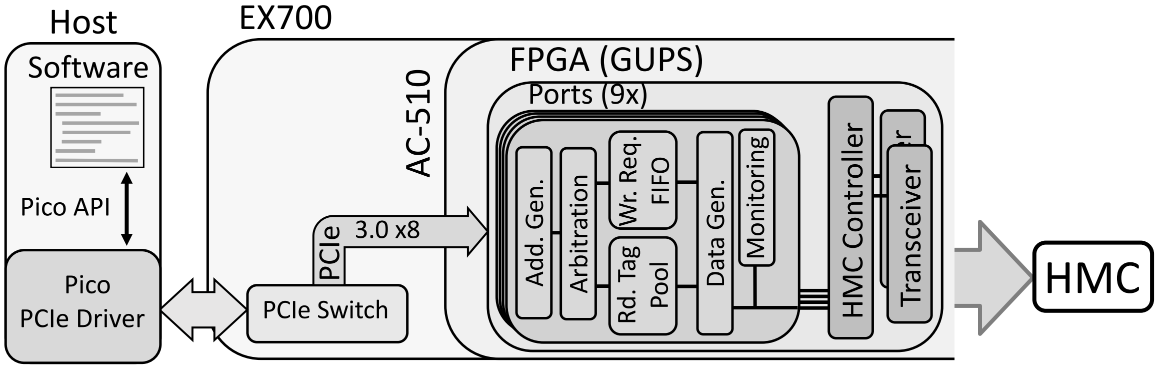
Closed-page policy Page Size = 256 B

Low-order-interleaving address mapping policy

34-bit address field:

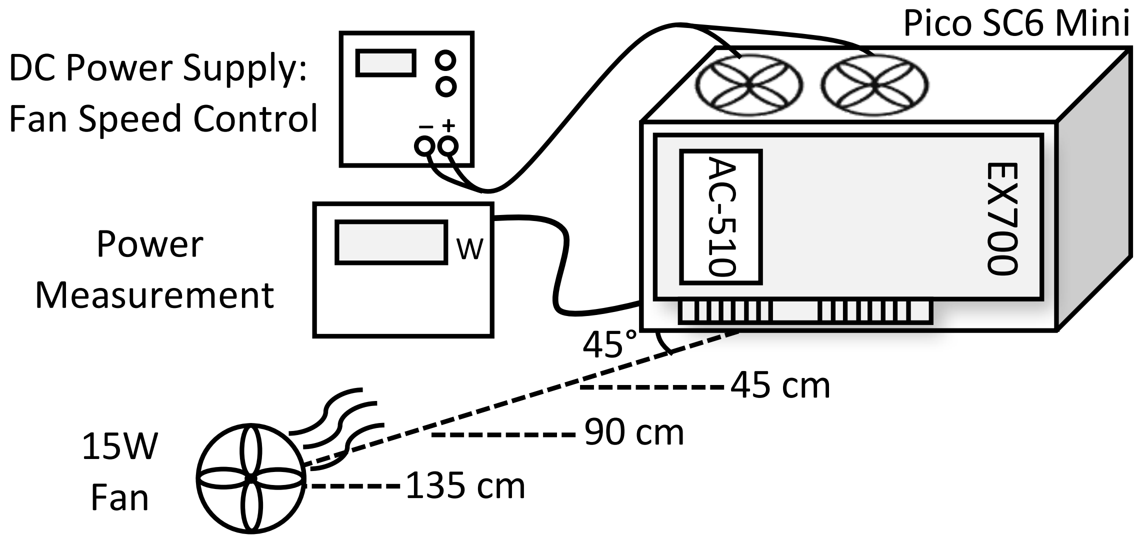


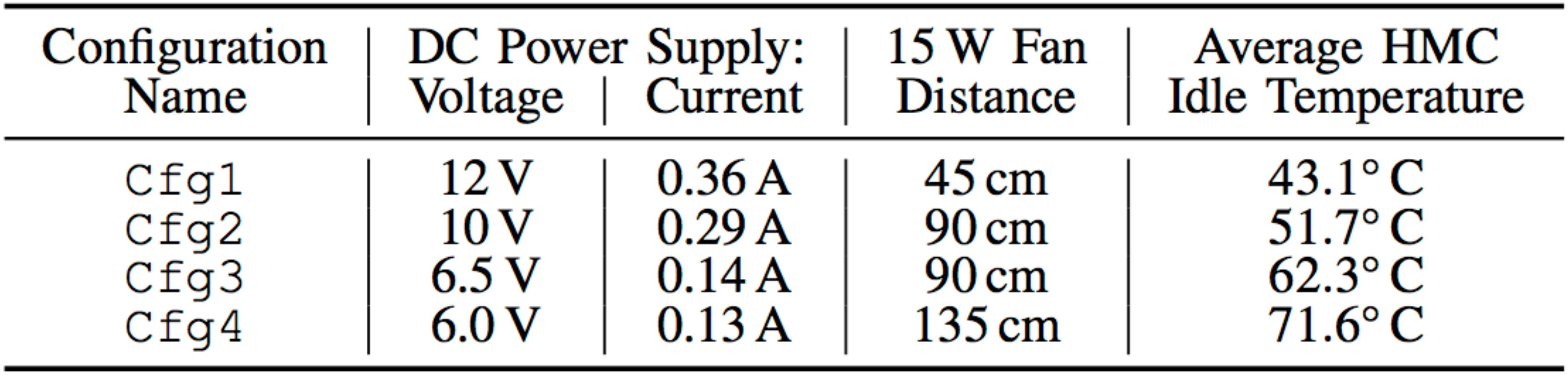
Experimental Setup III



	Full-scale GUPS	Small-scale GUPS	Stream GUPS
Addresses	Random Configurable Mask	Random Configurable Mask	Defined by User
Request Rate	Maximum	Configurable	Minimum
Experiment	Bandwidth Power Temperature High-Load Latency	Latency-Bandwidth	Integrity Check Low-Load Latency

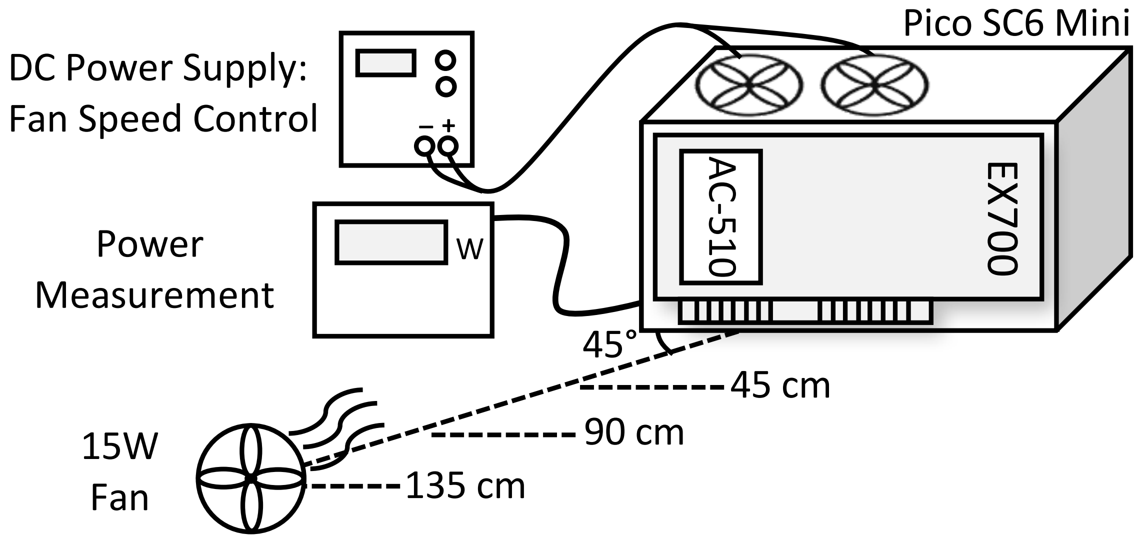
Thermal Configurations





Configuration Name	DC Power Supply: Voltage	Current	15 W Fan Distance	Average HMC Idle Temperature
Cfg1	12 V	0.36 A	45 cm	43.1° C
Cfg2	10 V	0.29 A	90 cm	51.7° C
Cfg3	6.5 V	0.14 A	90 cm	62.3° C
Cfg4	6.0 V	0.13 A	135 cm	71.6° C

Cooling Power

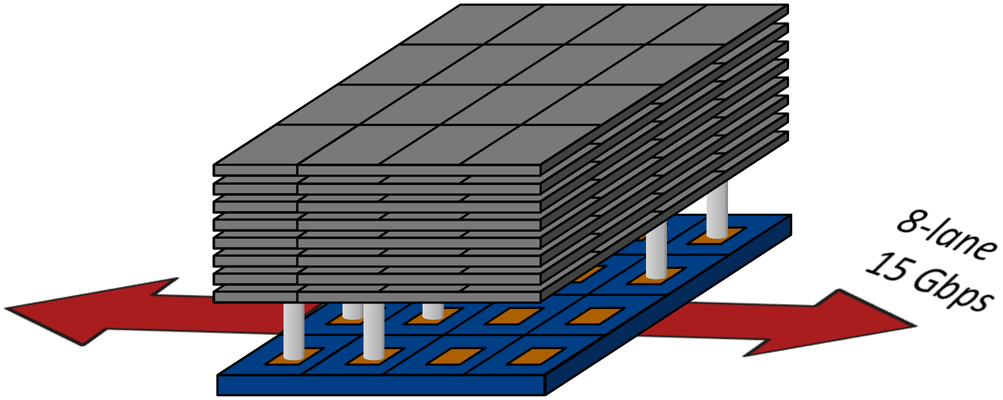


Configuration	Cooling Power
cfg1	19.32 W
cfg2	15.90 W
cfg3	13.90 W
cfg4	10.78 W

HMC Communication II

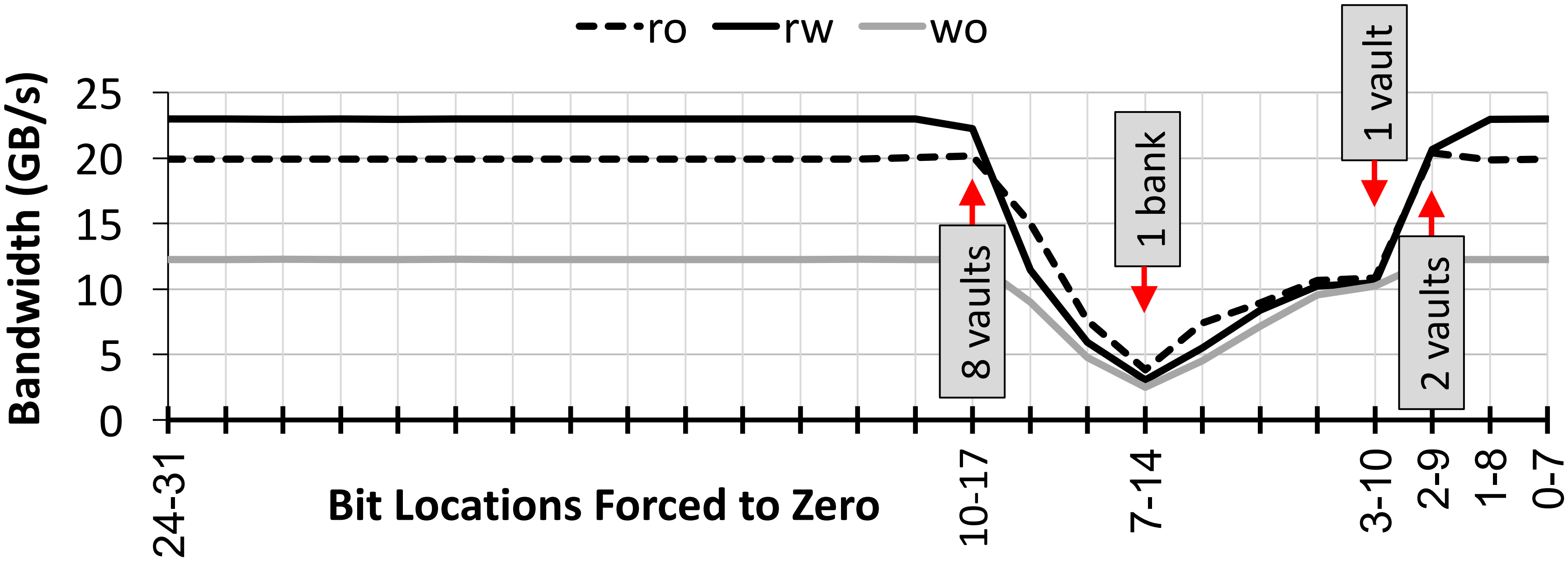
Two/Four full duplex external links:

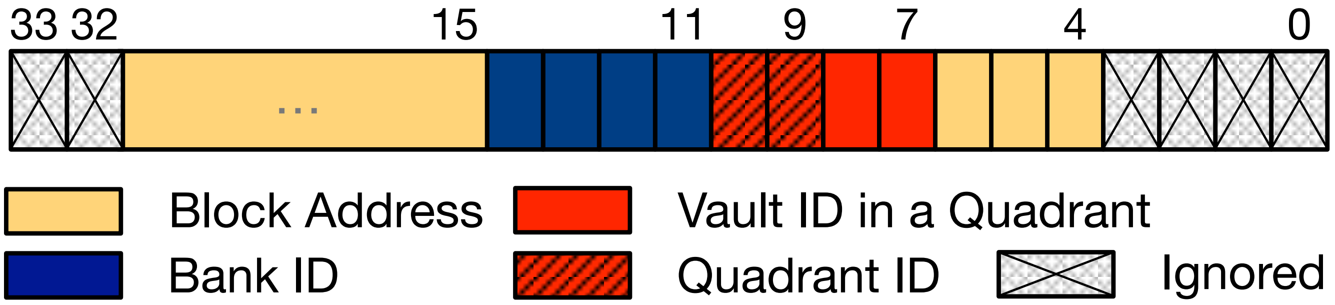
- Width of 16 or 8 lanes
- Configurable speeds of 10, 12.5, and 15 Gbps



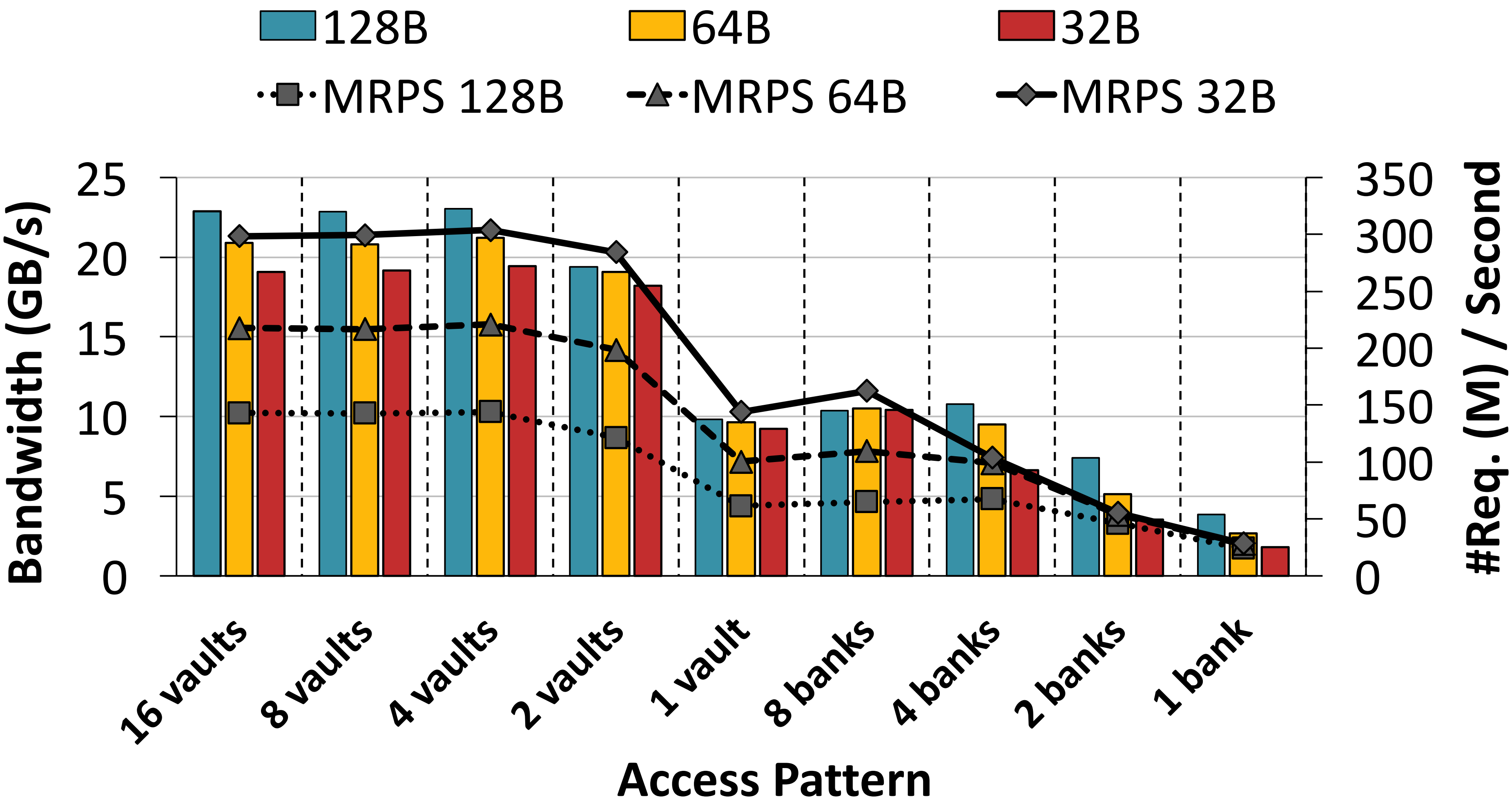
$$\begin{aligned} BW_{peak} &= 2\,\text{link} \times 8\,{}^{\text{lanes}}/_{\text{link}} \times 15\,\text{Gbps} \times 2\,\text{full duplex} \\ &= 480\,\text{Gbps} = 60\,\text{GB/s}. \end{aligned}$$

Address Mapping

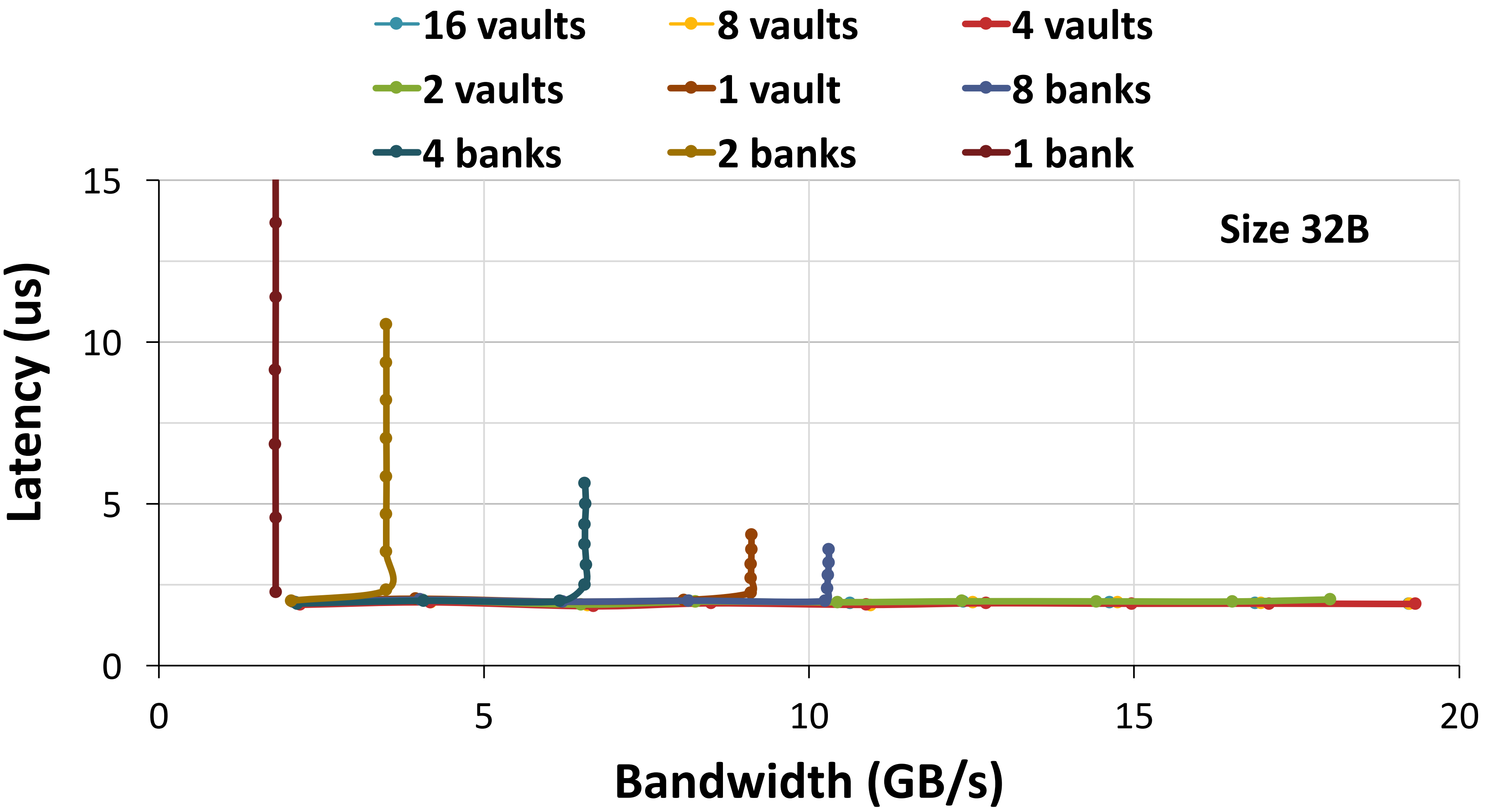




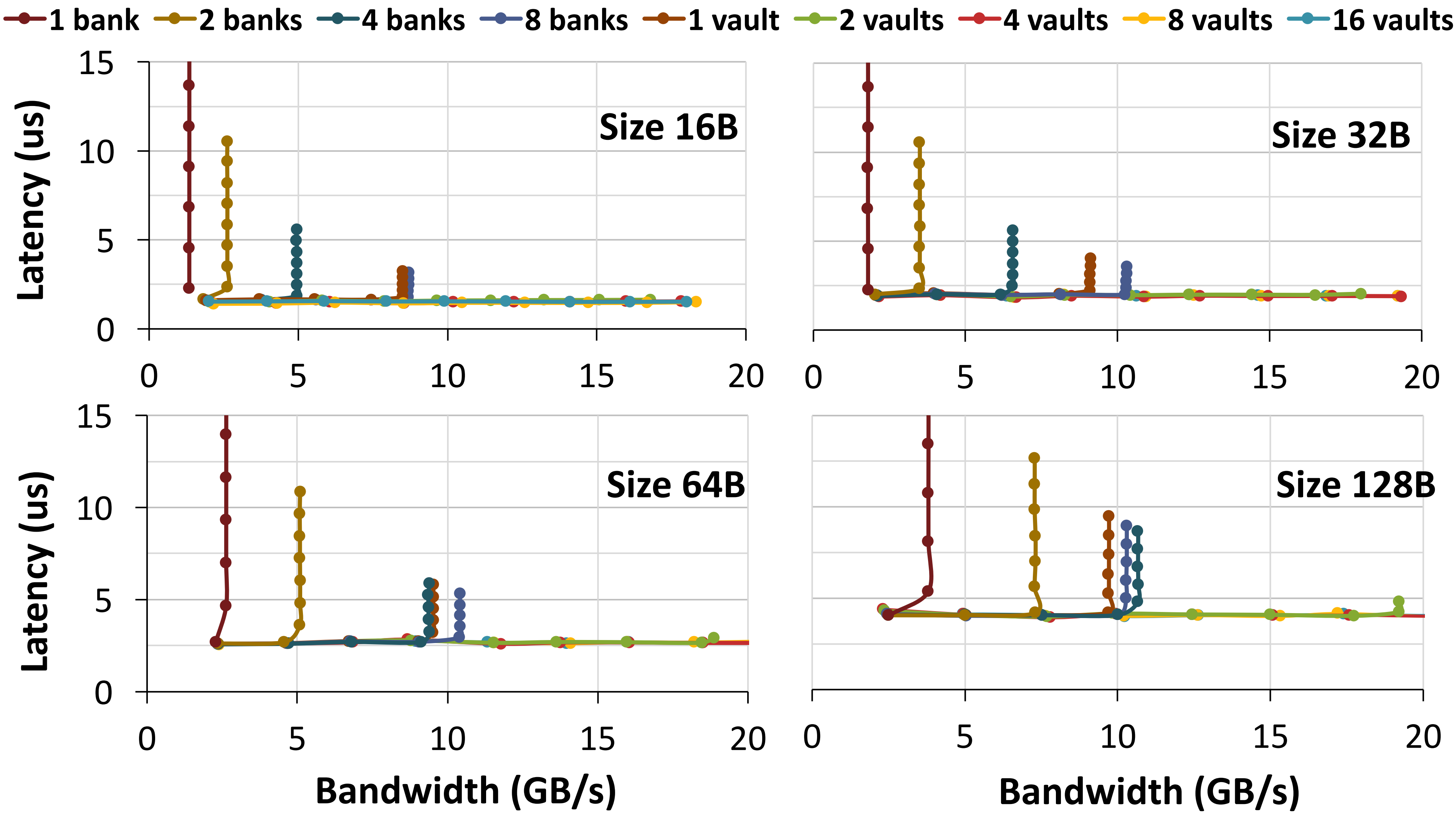
Bandwidth II



Latency-Bandwidth II



Latency-Bandwidth III


1 bank
2 banks
4 banks
8 banks
1 vault
2 vaults
4 vaults
8 vaults
16 vaults
Size 16B
Size 32B
Size 64B
Size 128B
Latency (us)
Latency (us)
Bandwidth (GB/s)
Bandwidth (GB/s)