UNIVERSITÀ DI TRENTO

Benchmarking Ethernet Interconnect for HPC/AI workloads

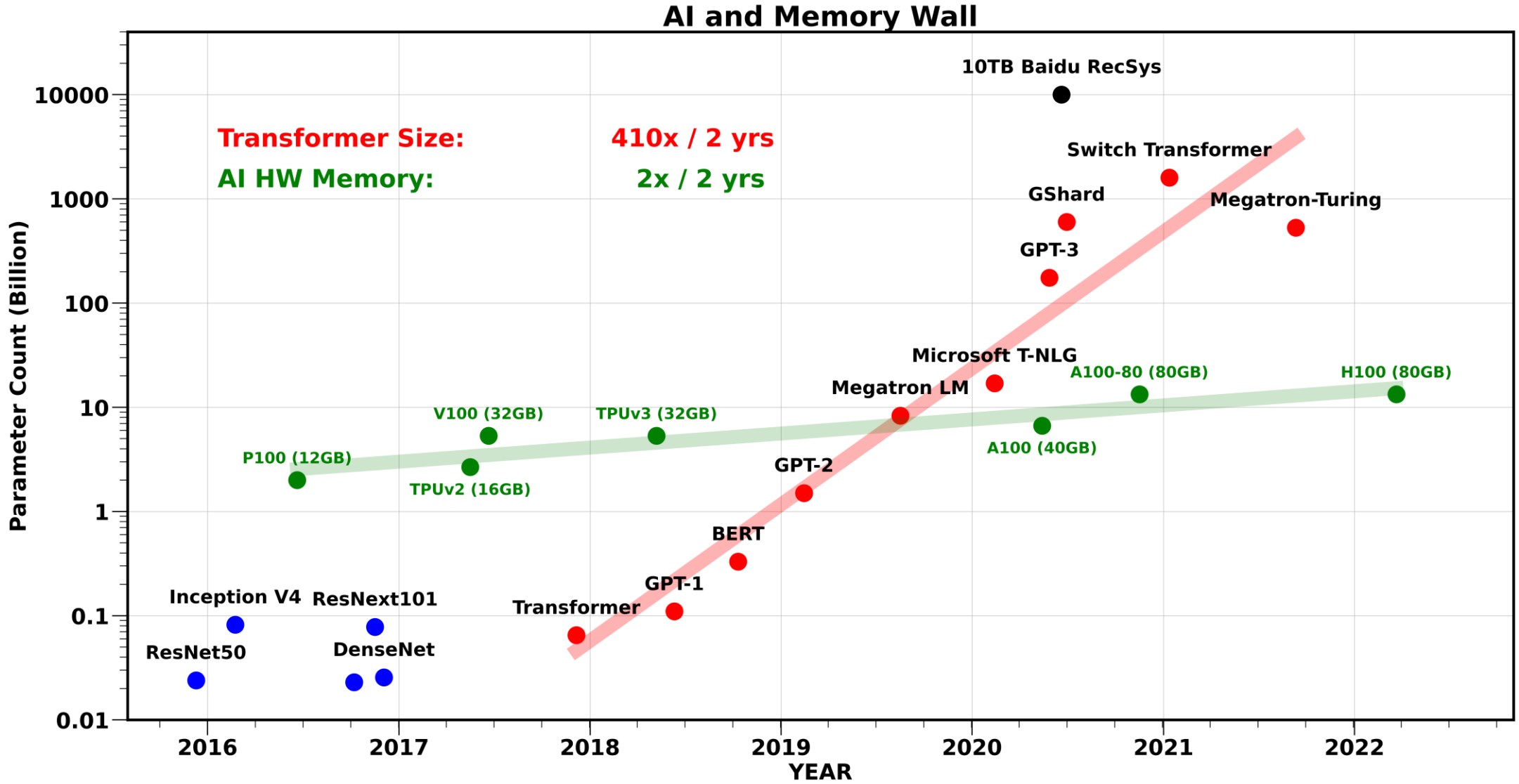
L. Pichetti *D. De Sensi* *K. Sivalingam* *S. Nassyr*
M. Turisini *D. Cesarini* *D. Pleiter* *A. Artigiani* *F. Vella*

SC24 Atlanta, GA hpc creates.

HiCREST

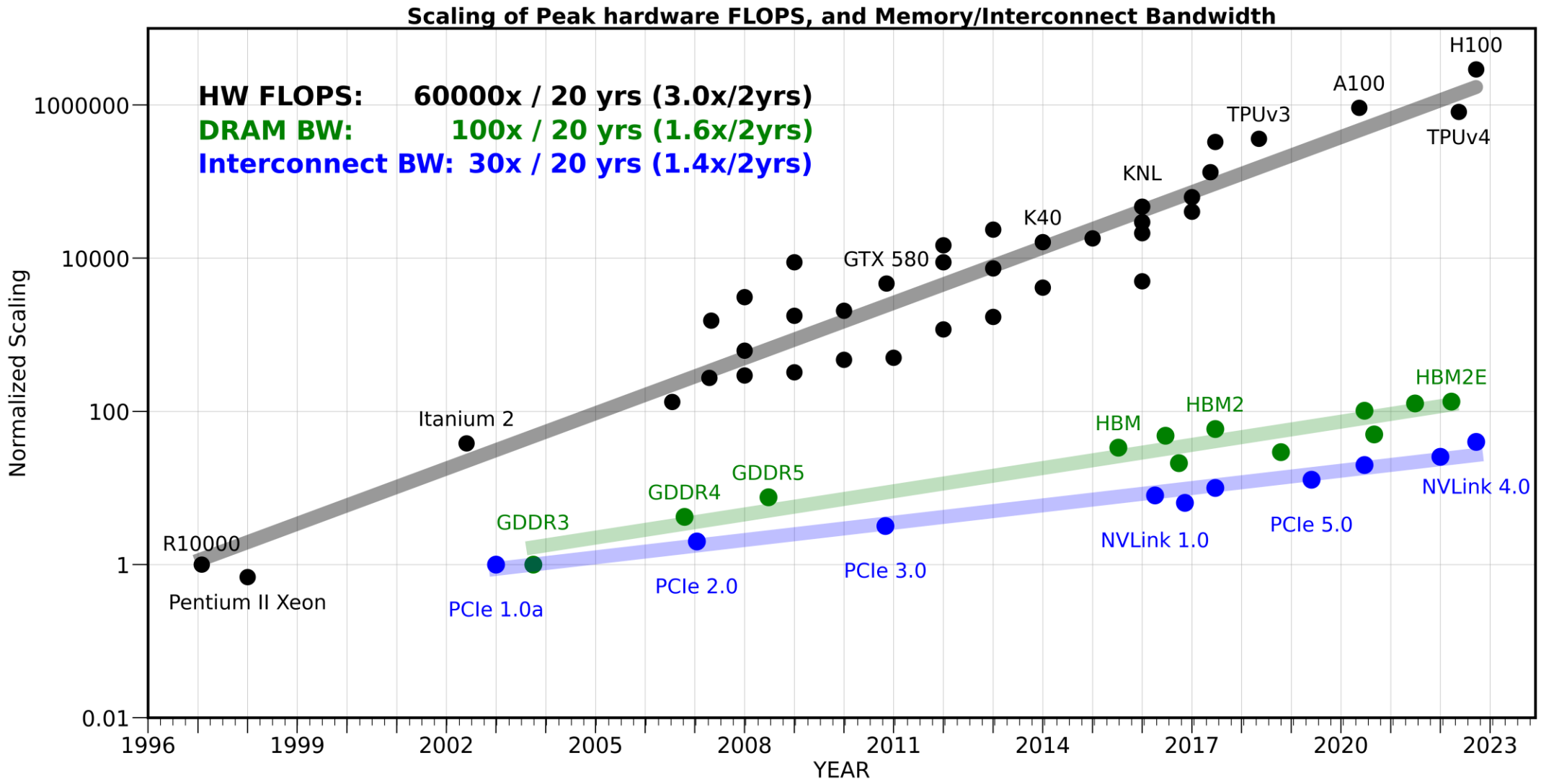
IXPUG INTEL EXTREME PERFORMANCE USERS GROUP

Introduction – The data-movement bottleneck



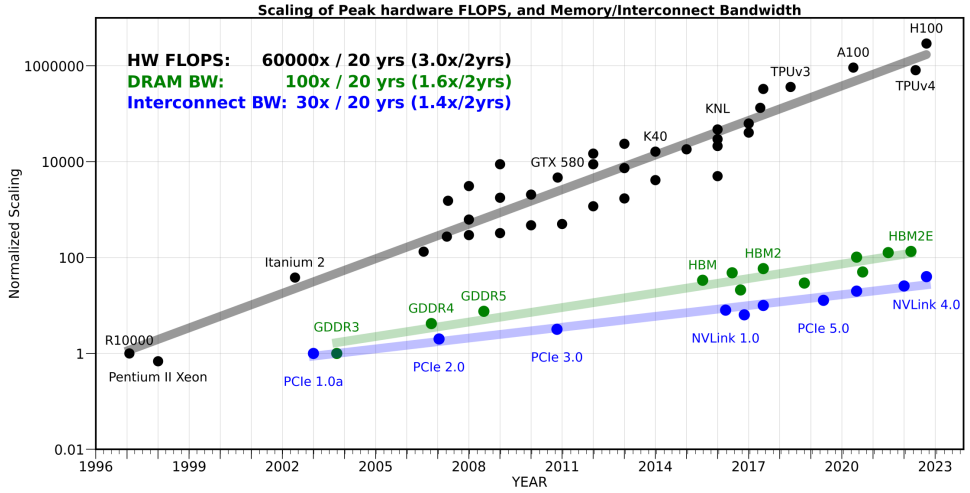
Gholami et al. "AI and Memory Wall." IEEE Micro 44 (2024).

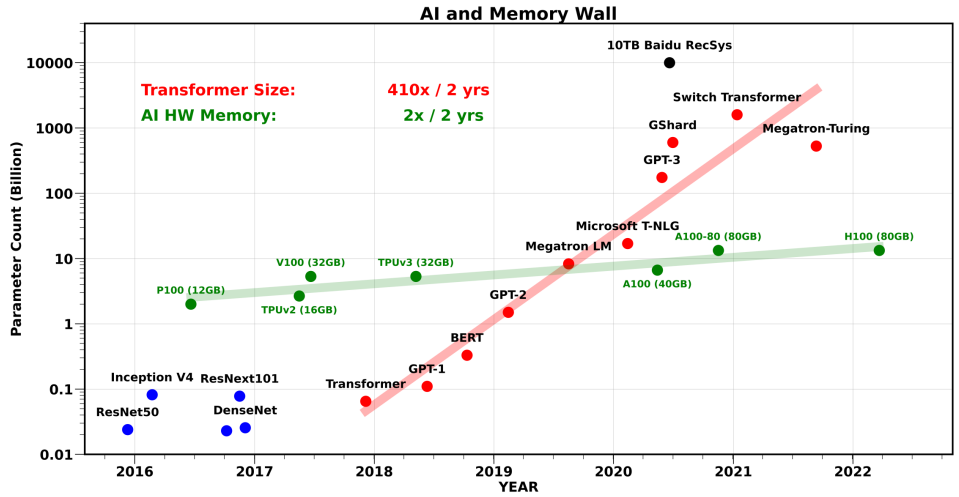
Introduction – The data-movement bottleneck

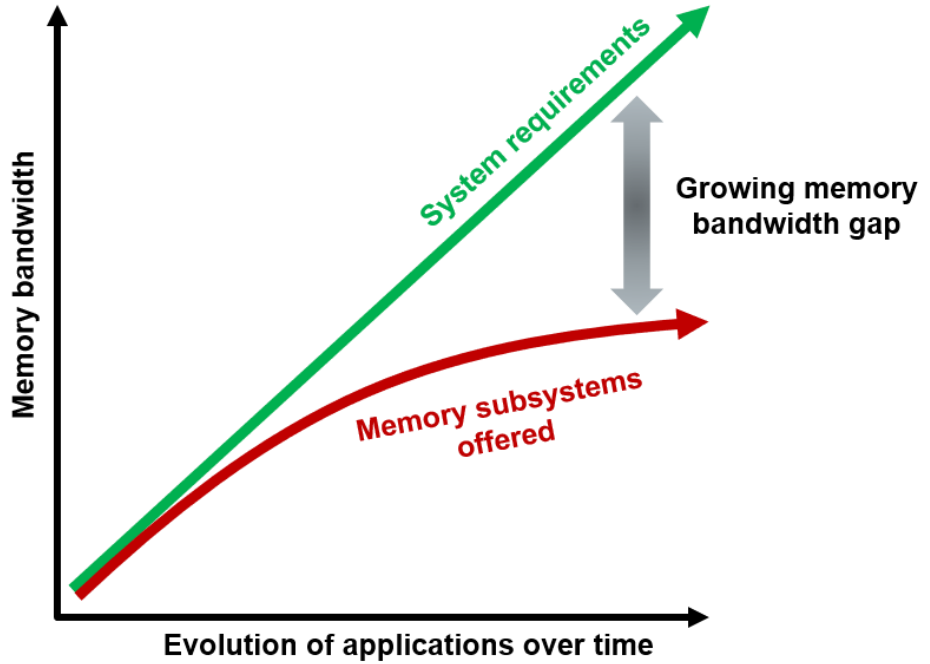


Gholami et al. "AI and Memory Wall." IEEE Micro 44 (2024).

Introduction – The data-movement bottleneck

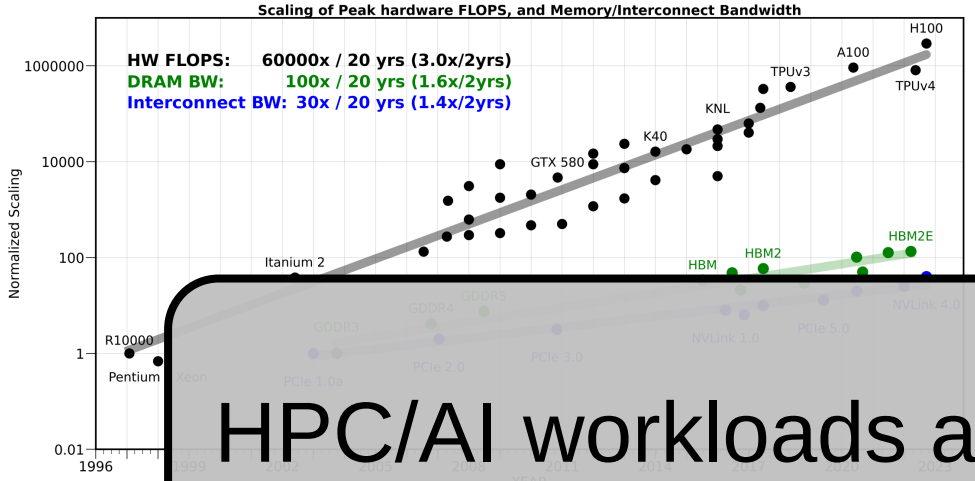


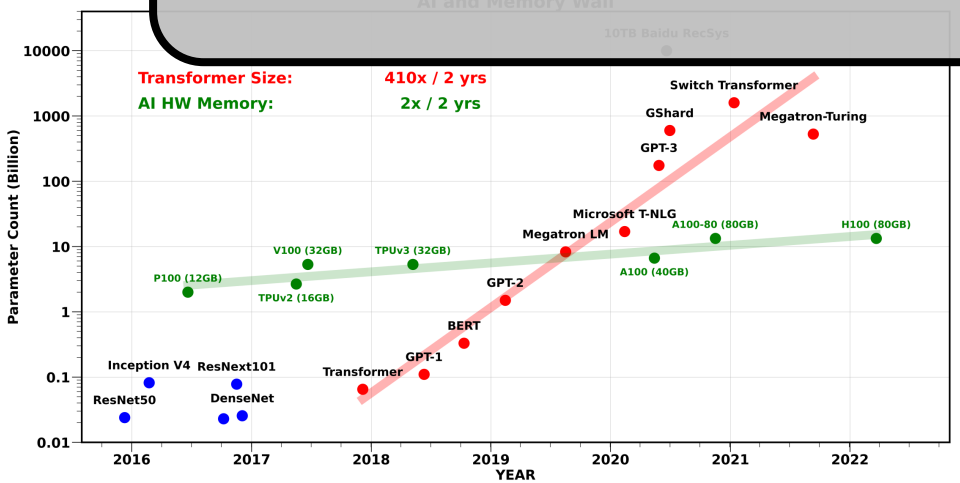




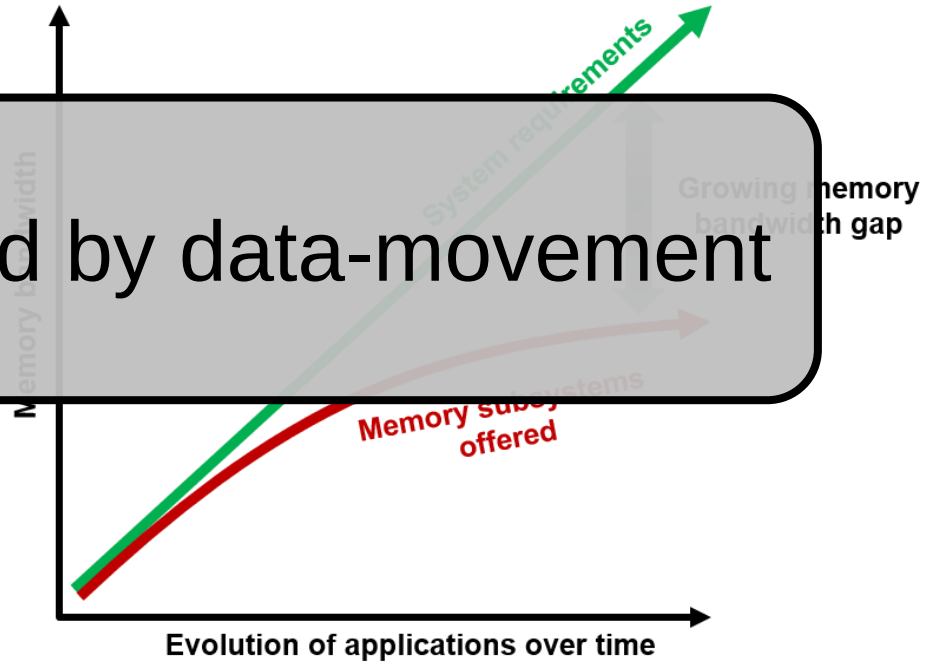
Gholami et al. “AI and Memory Wall.” IEEE Micro 44 (2024).

Introduction – The data-movement bottleneck





HPC/AI workloads are bottlenecked by data-movement



Gholami et al. "AI and Memory Wall." IEEE Micro 44 (2024).

Introduction – Ethernet and InfiniBand

InfiniBand vs **Gigabit Ethernet** Top500.

Introduction – Ethernet and InfiniBand

InfiniBand vs **Gigabit Ethernet** Top500.

! DISCLAIMER !

Top500 is **significative for the HPC** trends.

However, keep in mind that **does not represent many company and data-centric clusters**.

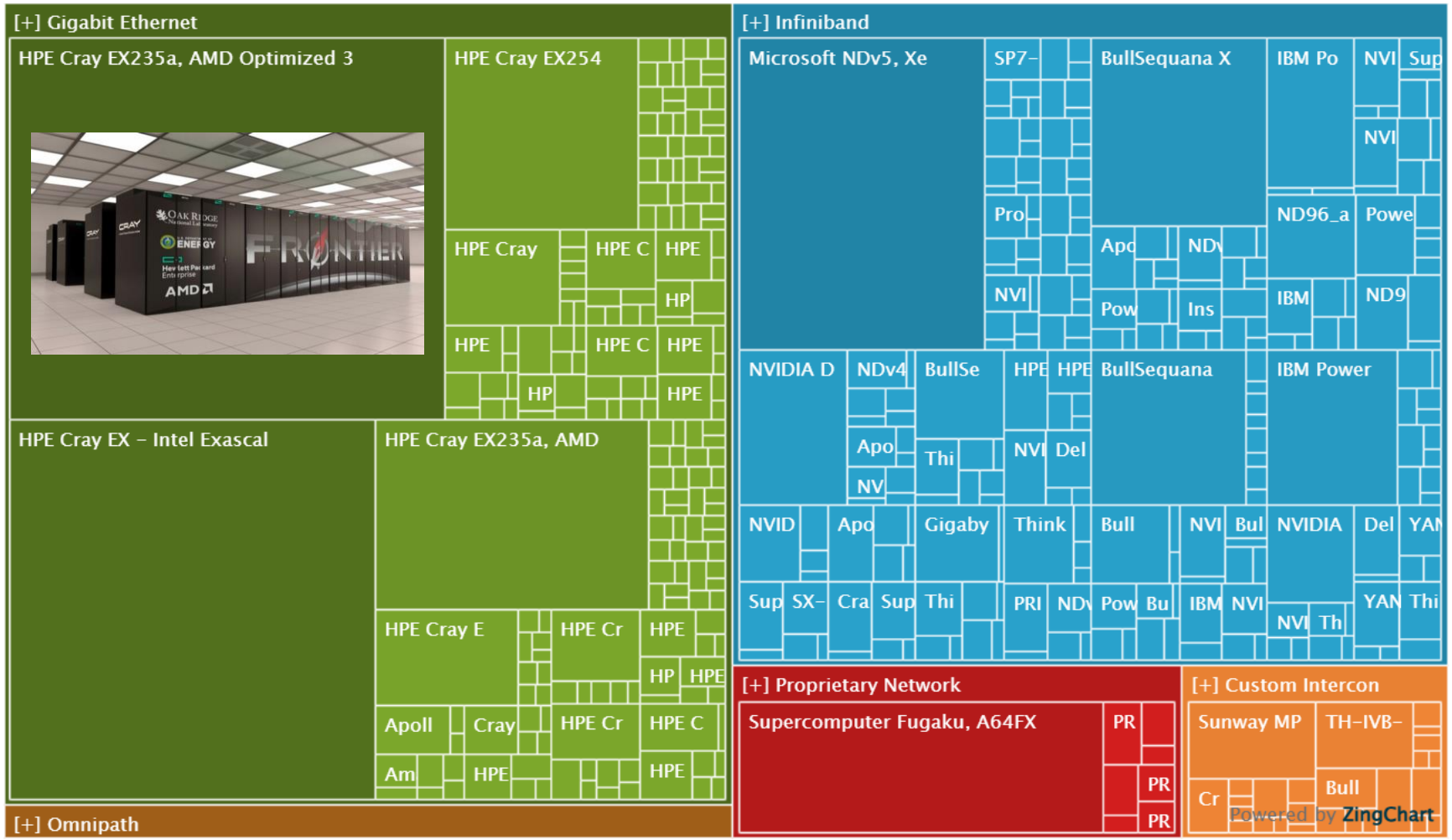
Introduction – Ethernet and InfiniBand

InfiniBand vs **Gigabit Ethernet** Top500.

Source: https://top500.org/

Introduction – Ethernet and InfiniBand

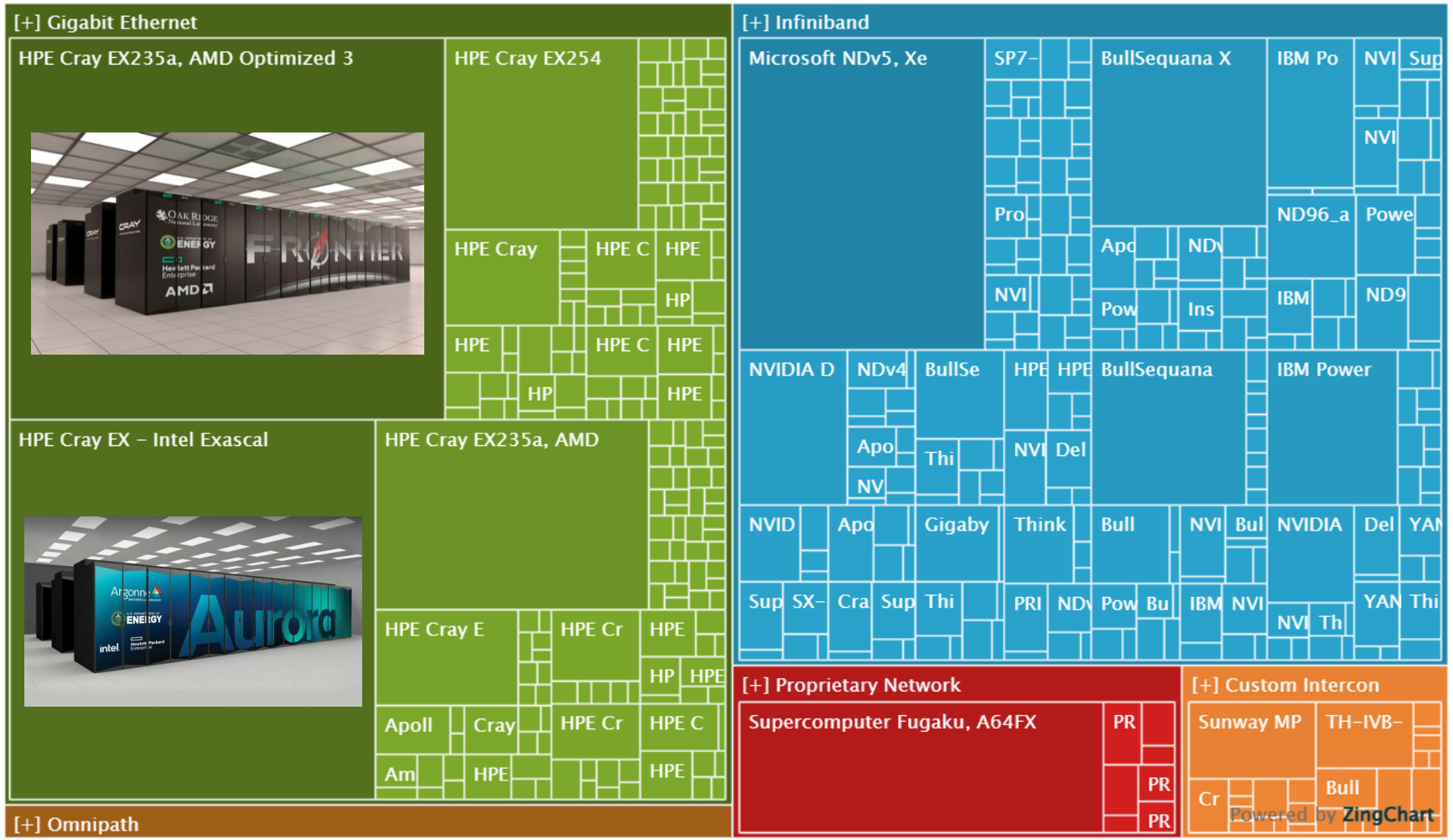
InfiniBand vs **Gigabit Ethernet** Top500.



Source: https://top500.org/

Introduction – Ethernet and InfiniBand

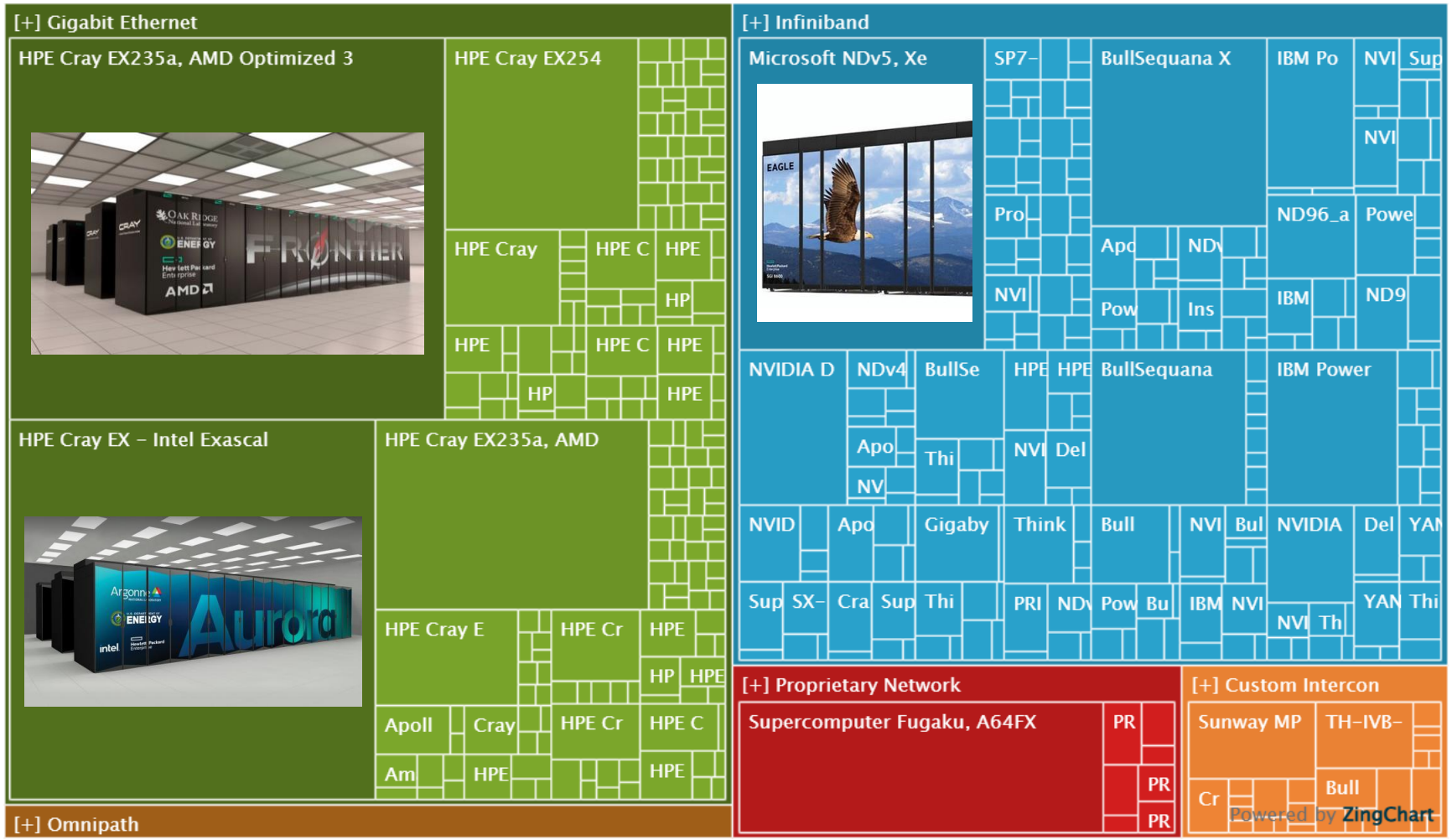
InfiniBand vs **Gigabit Ethernet** Top500.



Source: https://top500.org/

Introduction – Ethernet and InfiniBand

InfiniBand vs **Gigabit Ethernet** Top500.



Source: https://top500.org/

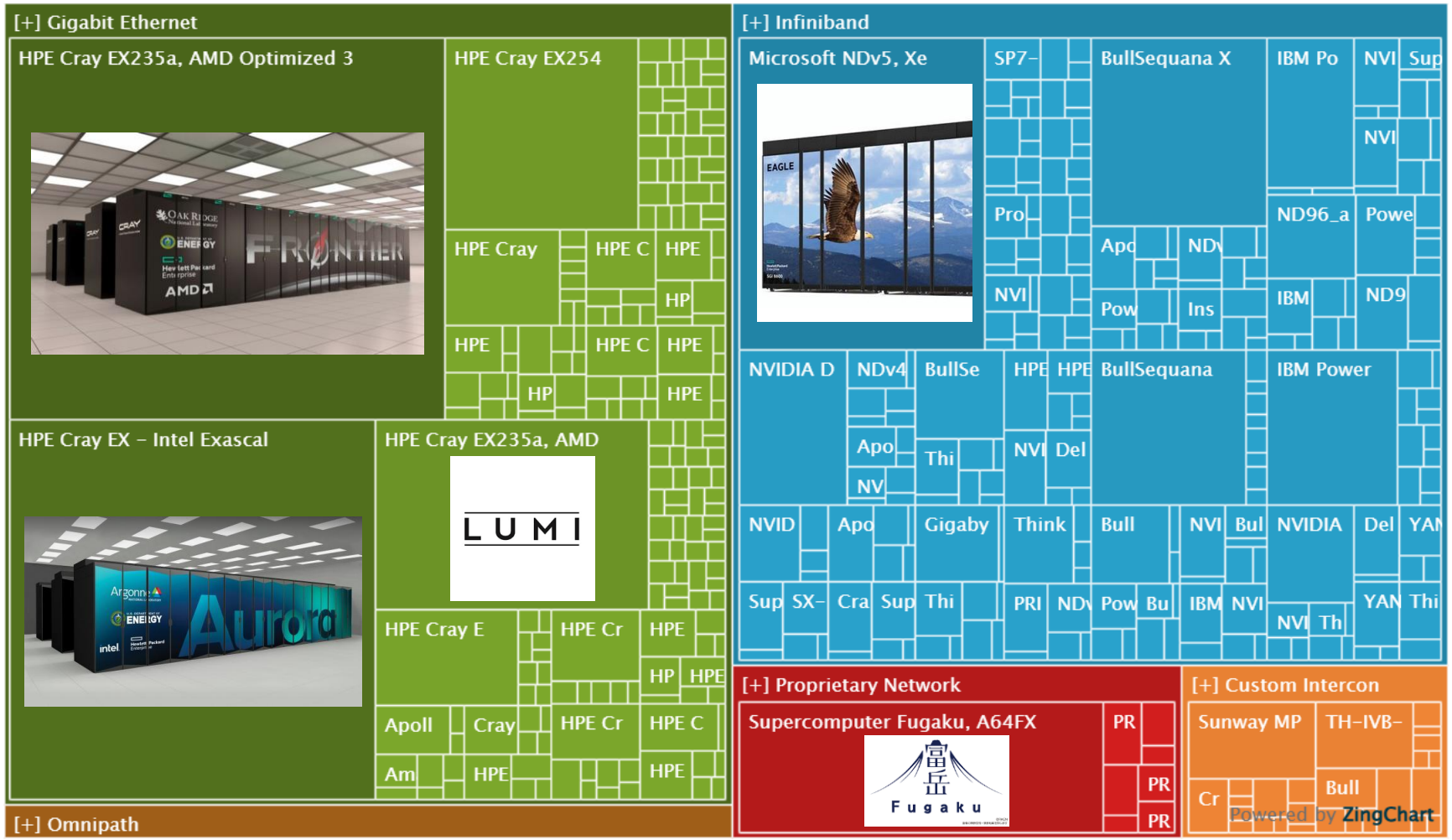
Introduction – Ethernet and InfiniBand

InfiniBand vs **Gigabit Ethernet** Top500.

Source: https://top500.org/

Introduction – Ethernet and InfiniBand

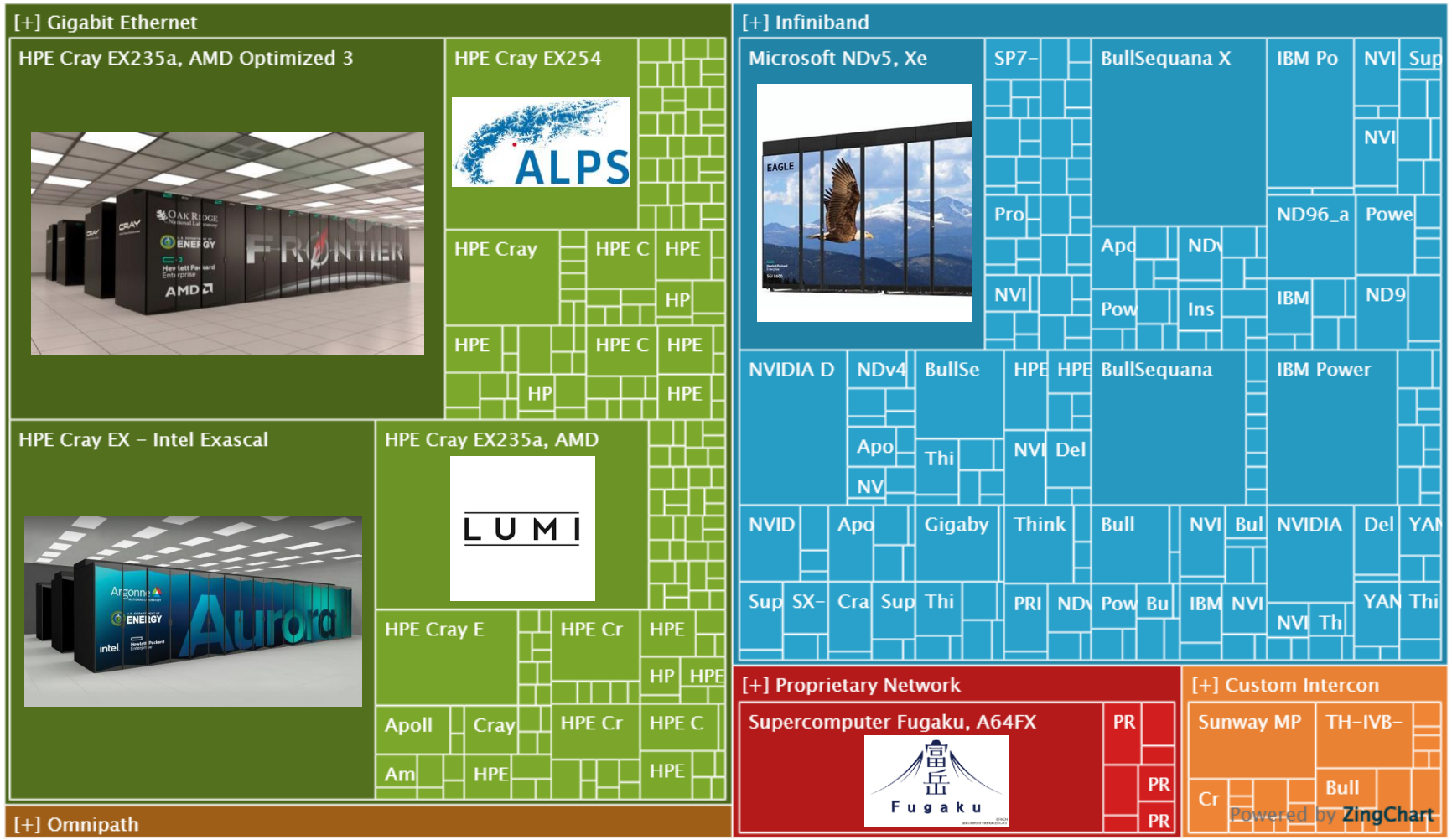
InfiniBand vs **Gigabit Ethernet** Top500.



Source: https://top500.org/

Introduction – Ethernet and InfiniBand

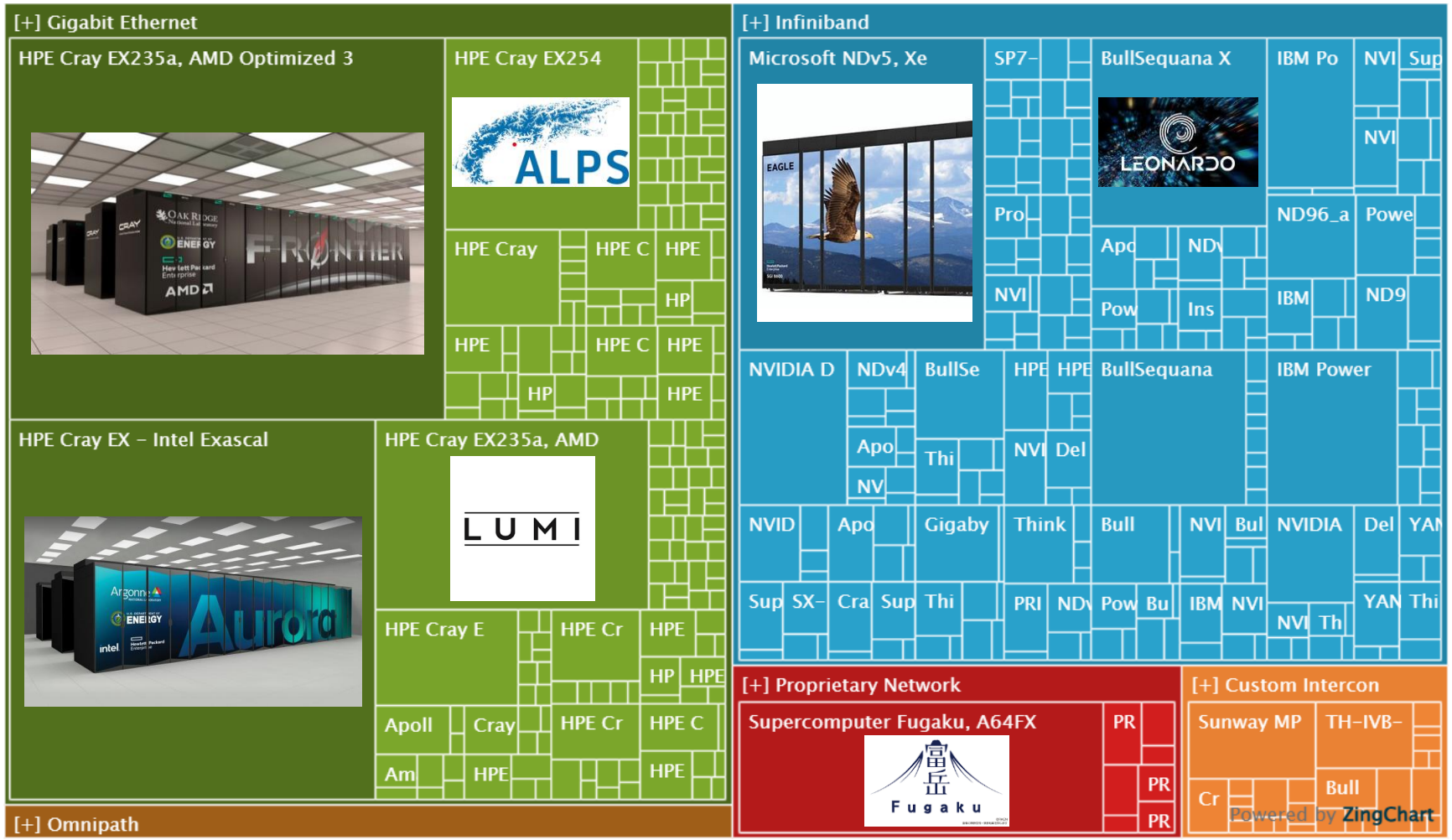
InfiniBand vs **Gigabit Ethernet** Top500.



Source: https://top500.org/

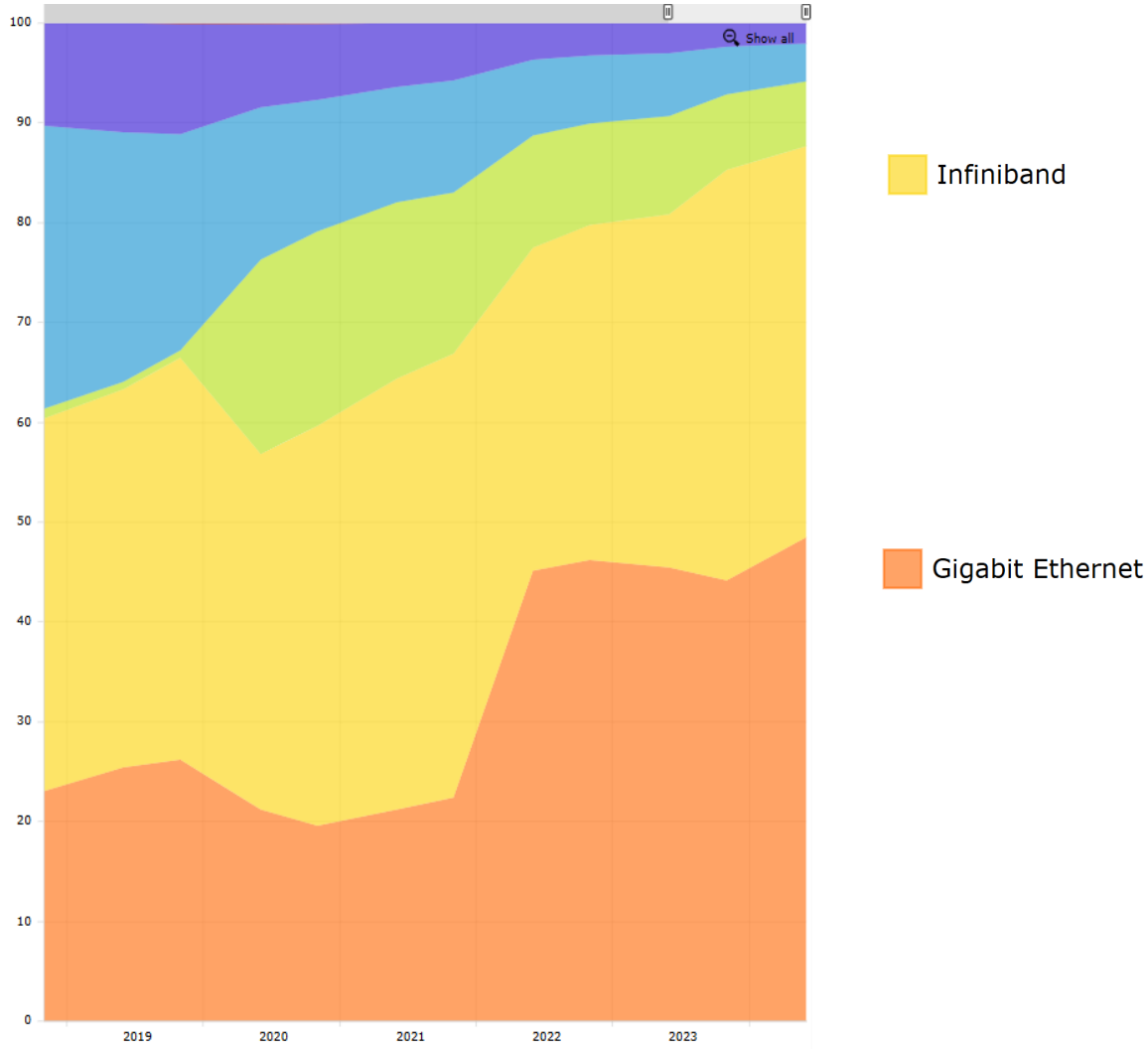
Introduction – Ethernet and InfiniBand

InfiniBand vs **Gigabit Ethernet** Top500.



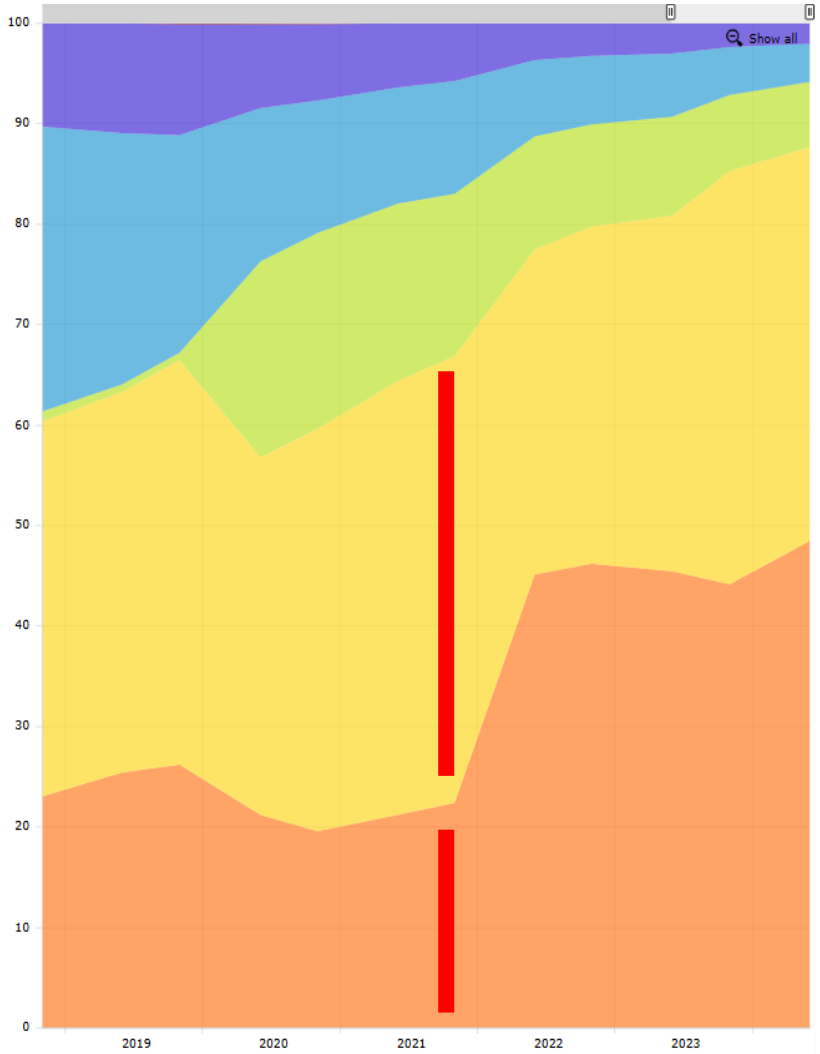
Source: https://top500.org/

Introduction – Trend over years



Source: https://top500.org/

Introduction – Trend over years



Infiniband

44%

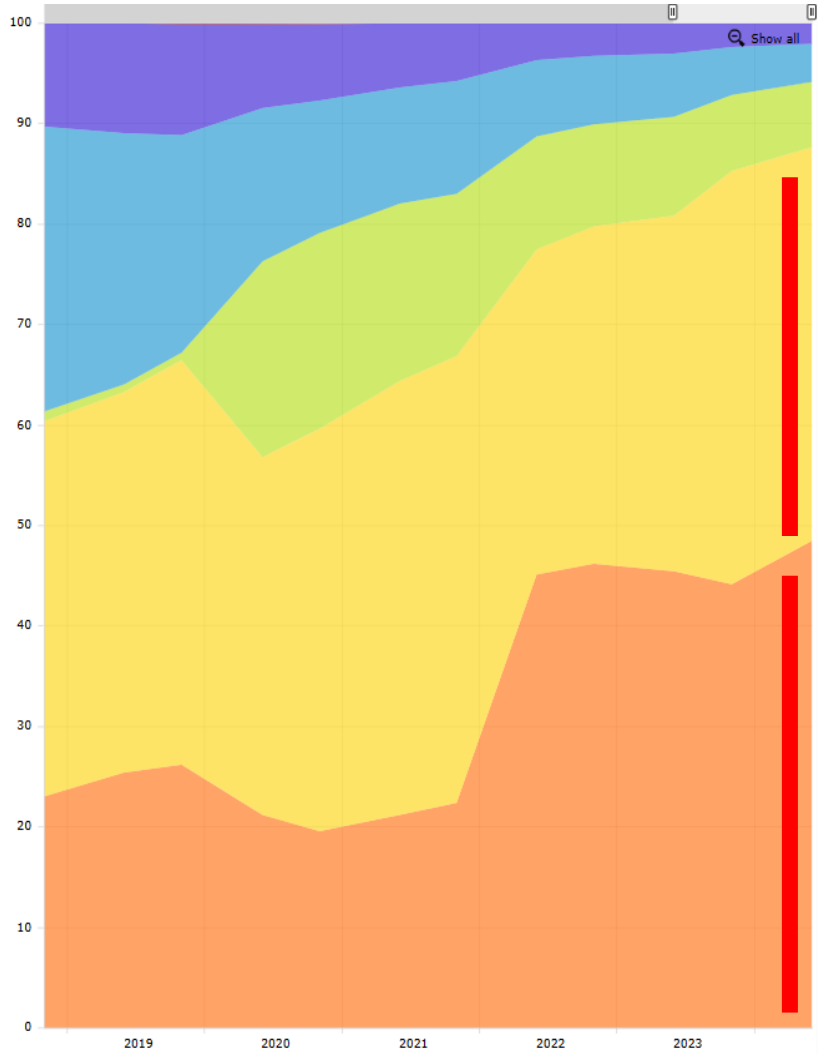
66%

Gigabit Ethernet

22%

Source: https://top500.org/

Introduction – Trend over years



Infiniband

44%

39%

66%

87%

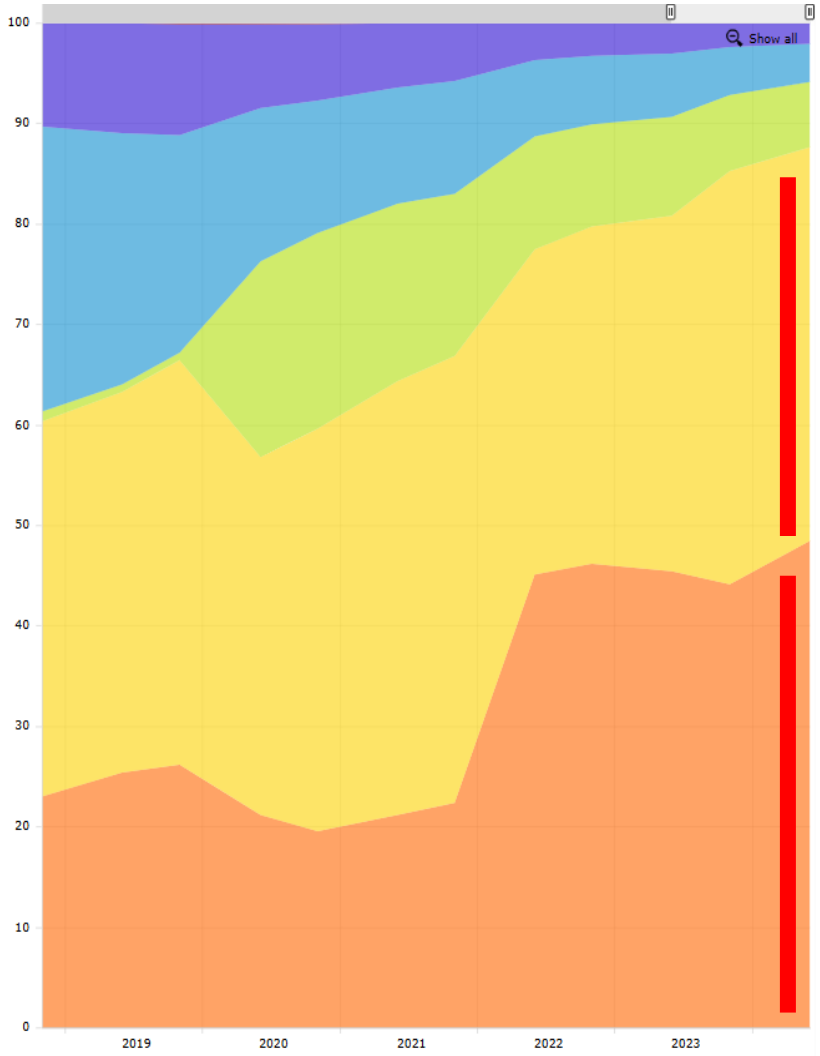
Gigabit Ethernet

22%

48%

Source: https://top500.org/

Introduction – Trend over years



Source: https://top500.org/

Infiniband

44%

39%

66%

87%

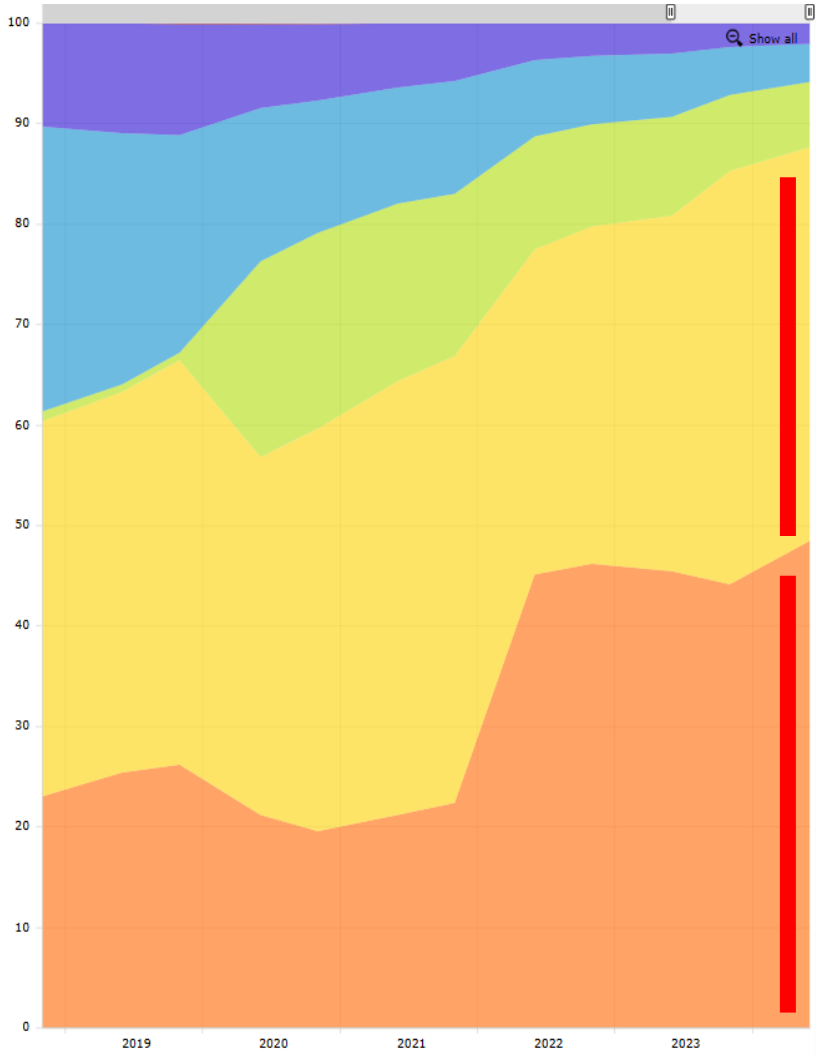
Gigabit Ethernet

22%

48%

Why Ethernet is currently preferred?

Introduction – Trend over years



Infiniband

44%

39%

66%

87%

Gigabit Ethernet

22%

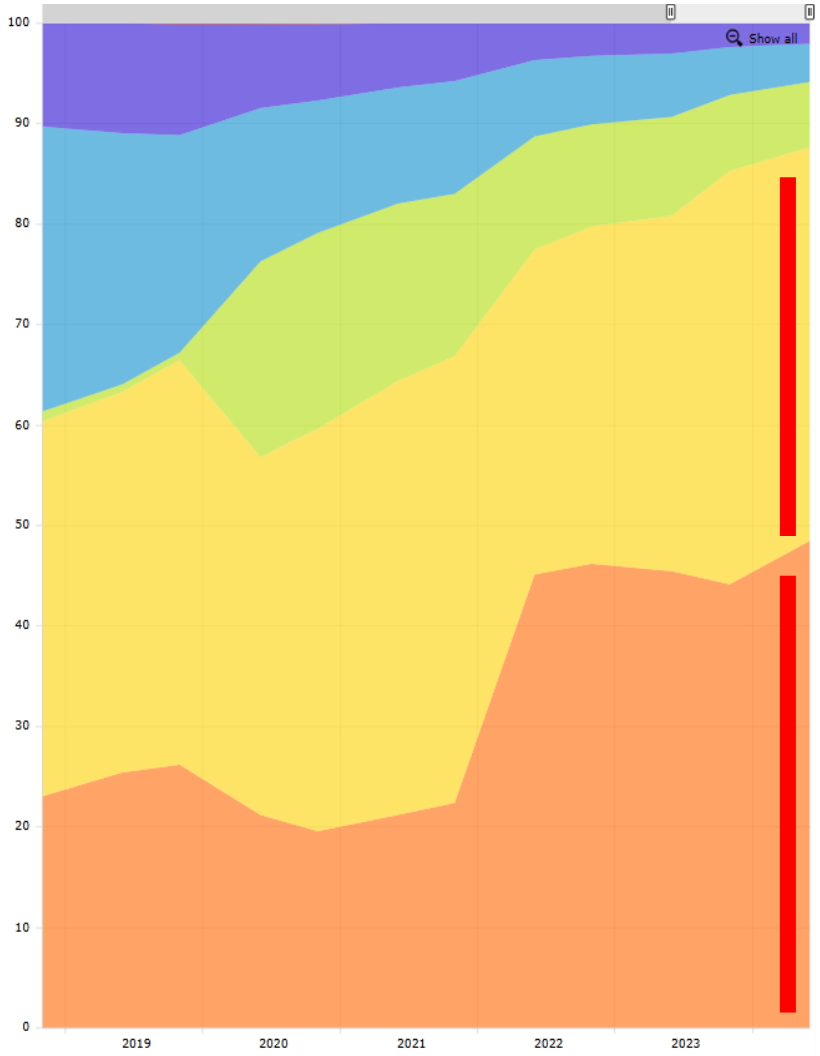
48%

Source: https://top500.org/

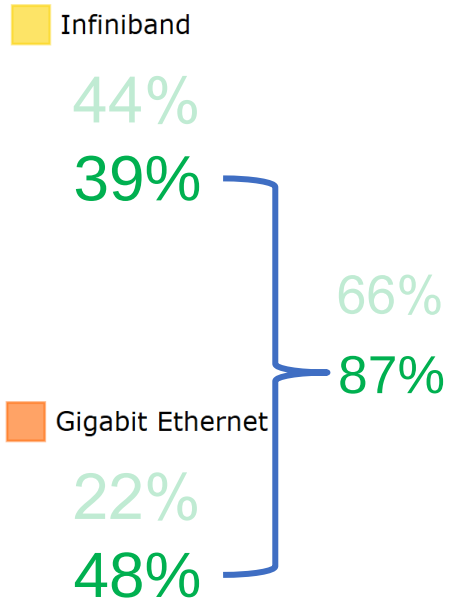
Why Ethernet is currently preferred?

Only a **few companies provide InfiniBand** solutions → risk of vendor **lock-in**

Introduction – Trend over years



Source: https://top500.org/

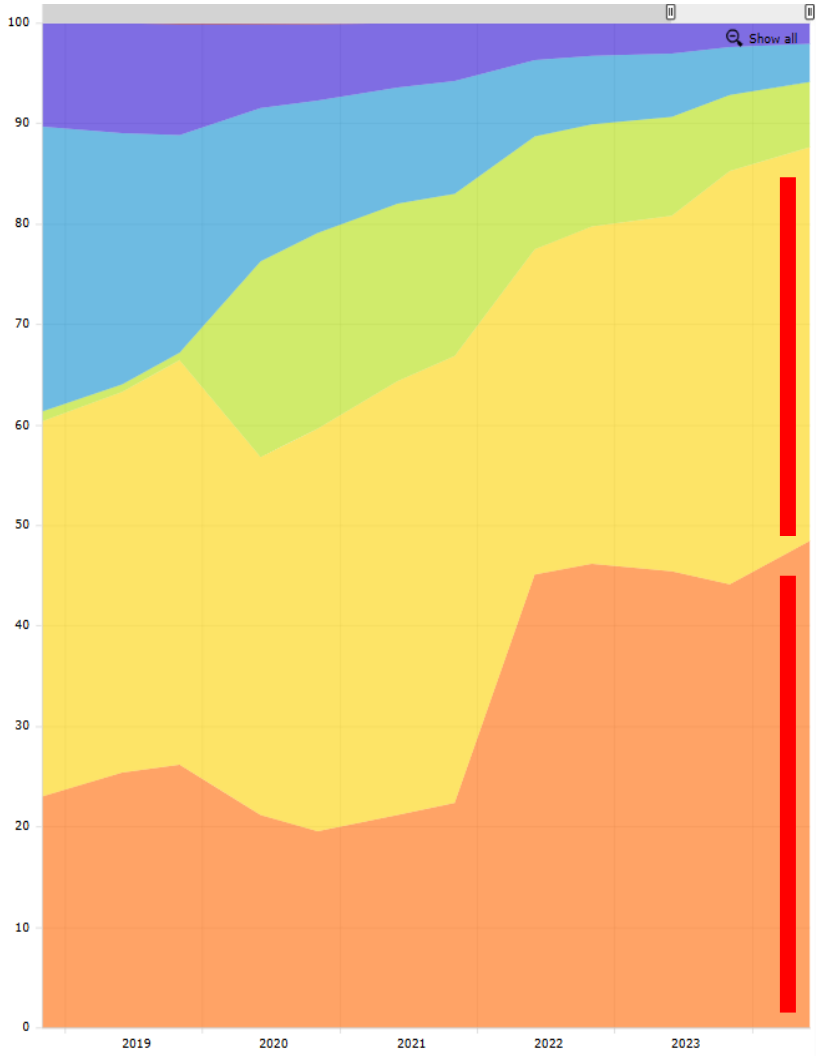


Why Ethernet is currently preferred?

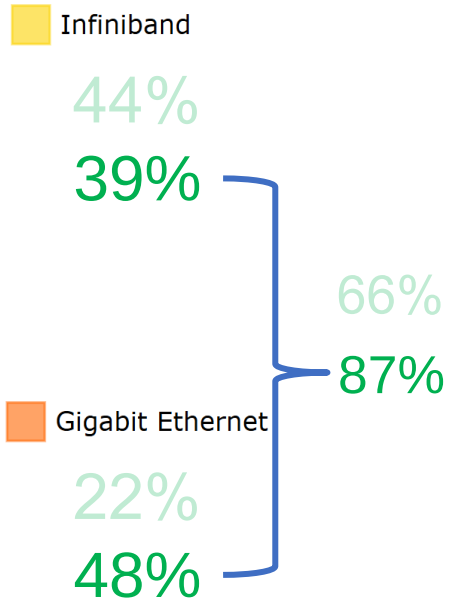
Only a **few companies provide InfiniBand** solutions → risk of vendor **lock-in**

Increasing interest in Ethernet solutions.

Introduction – Trend over years



Source: https://top500.org/



Why Ethernet is currently preferred?

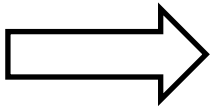
Only a **few companies provide InfiniBand** solutions → risk of vendor **lock-in**

Increasing interest in Ethernet solutions.

However, Ethernet protocols have **limitations**.

Introduction – Remote Direct Memory Access

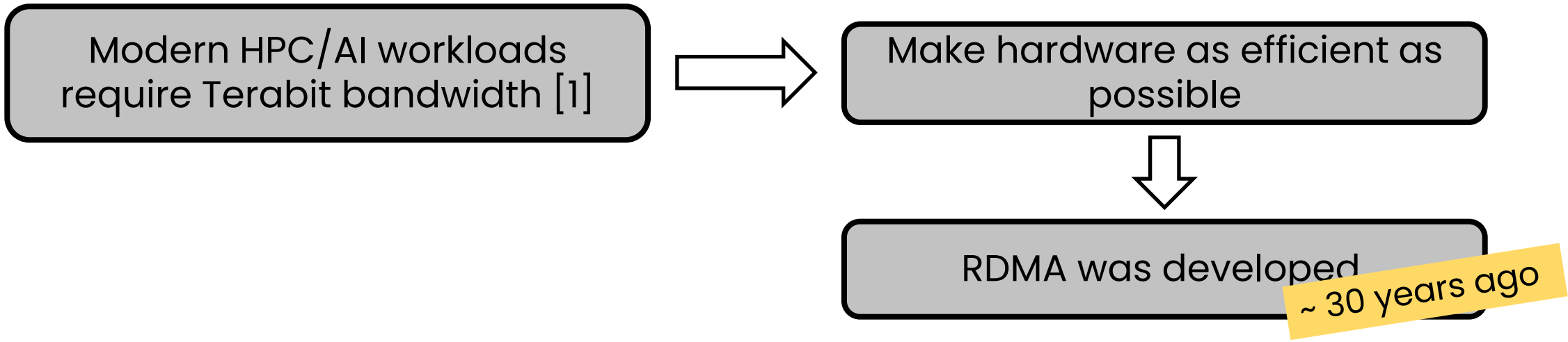
Modern HPC/AI workloads require Terabit bandwidth [1]



Make hardware as efficient as possible

[1] Hoefler et al. “HammingMesh: A Network Topology for Large-Scale Deep Learning.” SC22.
[2] T. Hoefler et al., "Data Center Ethernet and Remote Direct Memory Access: Issues at Hyperscale," in Computer.

Introduction – Remote Direct Memory Access



[1] Hoefler et al. “HammingMesh: A Network Topology for Large-Scale Deep Learning.” SC22.
[2] T. Hoefler et al., "Data Center Ethernet and Remote Direct Memory Access: Issues at Hyperscale," in Computer.

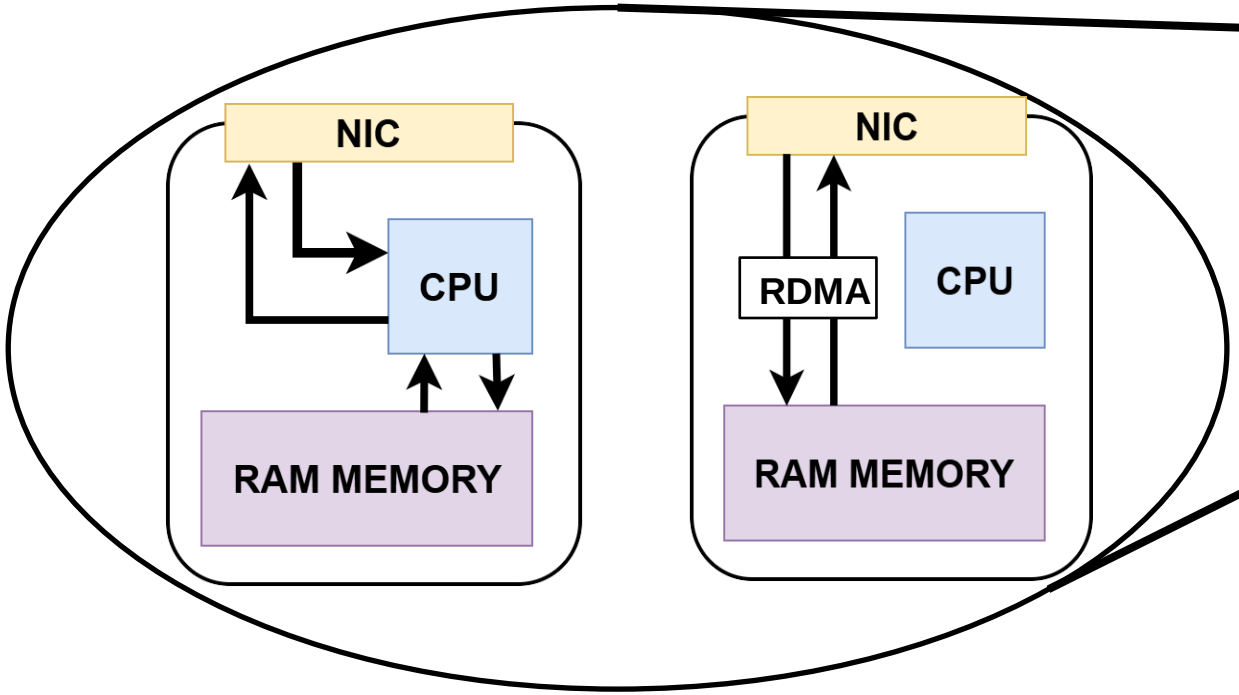
Introduction – Remote Direct Memory Access

Modern HPC/AI workloads require Terabit bandwidth [1]

Make hardware as efficient as possible

RDMA was developed

~ 30 years ago



[1] Hoefler et al. "HammingMesh: A Network Topology for Large-Scale Deep Learning." SC22.
[2] T. Hoefler et al., "Data Center Ethernet and Remote Direct Memory Access: Issues at Hyperscale," in Computer.

Introduction – Remote Direct Memory Access

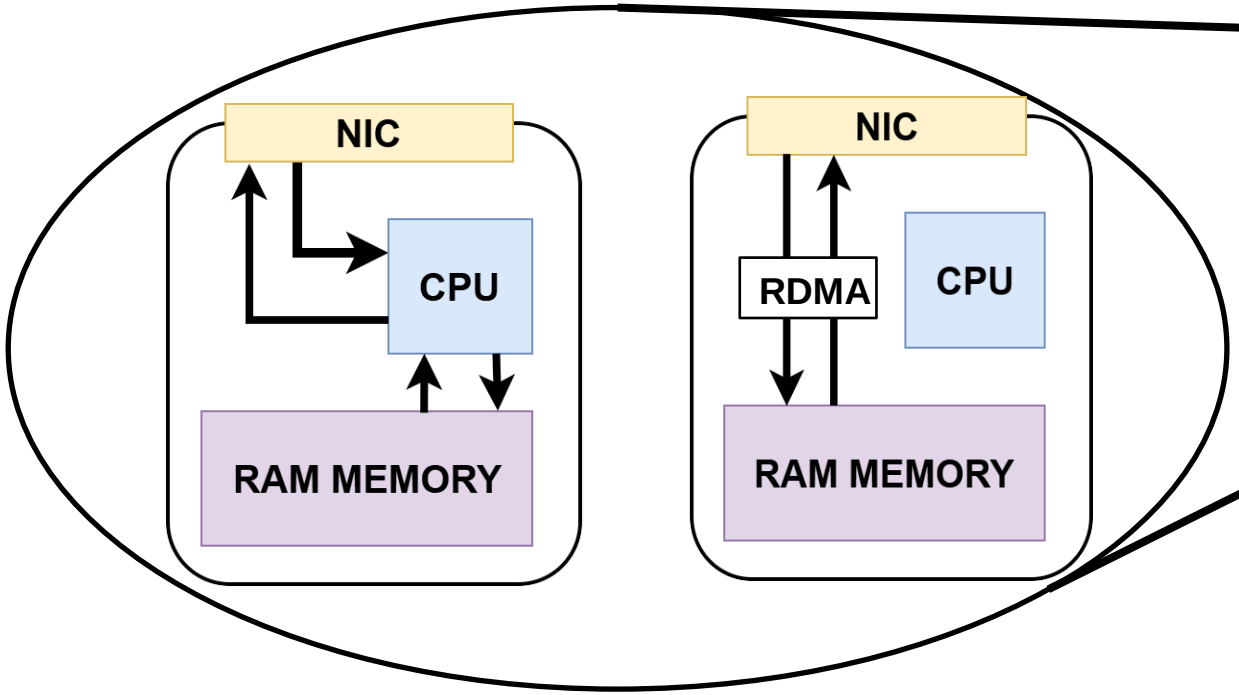
Modern HPC/AI workloads require Terabit bandwidth [1]

Make hardware as efficient as possible

RDMA was developed

~ 30 years ago

IB-based RDMA becomes a standard de facto.



[1] Hoefler et al. "HammingMesh: A Network Topology for Large-Scale Deep Learning." SC22.
[2] T. Hoefler et al., "Data Center Ethernet and Remote Direct Memory Access: Issues at Hyperscale," in Computer.

Introduction – Remote Direct Memory Access

Modern HPC/AI workloads require Terabit bandwidth [1]

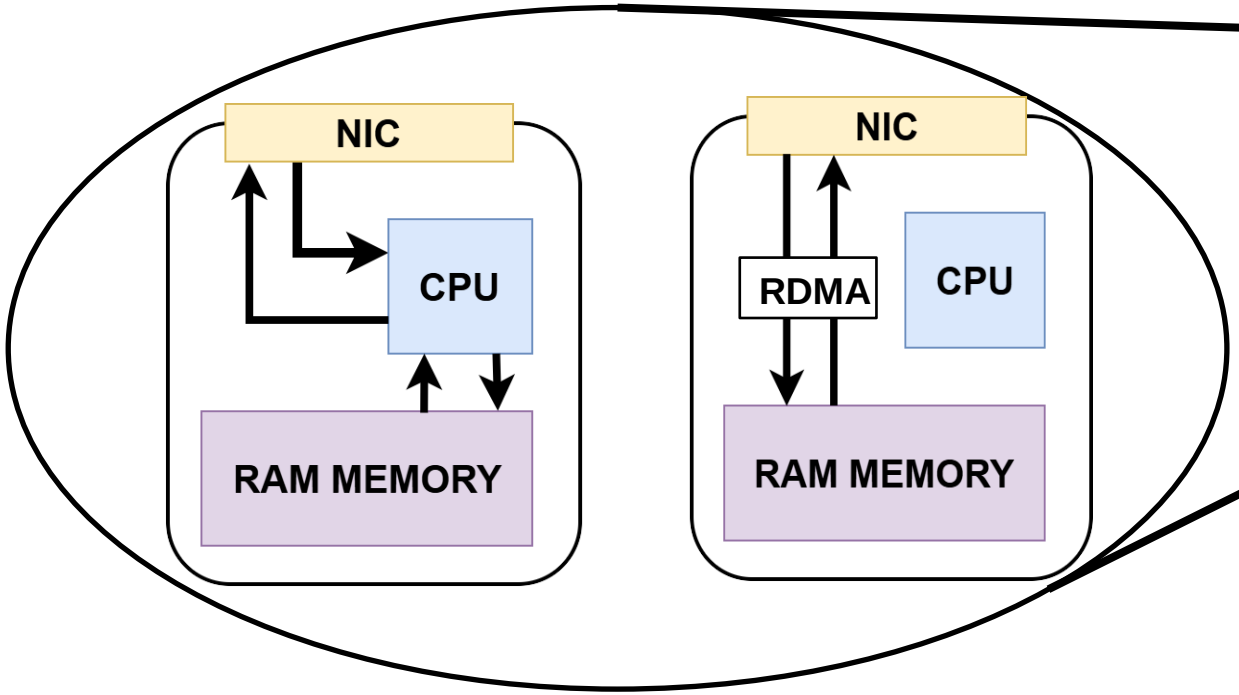
Make hardware as efficient as possible

RDMA was developed

~ 30 years ago

IB-based RDMA becomes a standard de facto.

RDMA was **further adapted** to Ethernet (RoCE)



[1] Hoefler et al. "HammingMesh: A Network Topology for Large-Scale Deep Learning." SC22.
[2] T. Hoefler et al., "Data Center Ethernet and Remote Direct Memory Access: Issues at Hyperscale," in Computer.

Introduction – Remote Direct Memory Access

Modern HPC/AI workloads require Terabit bandwidth [1]

Make hardware as efficient as possible

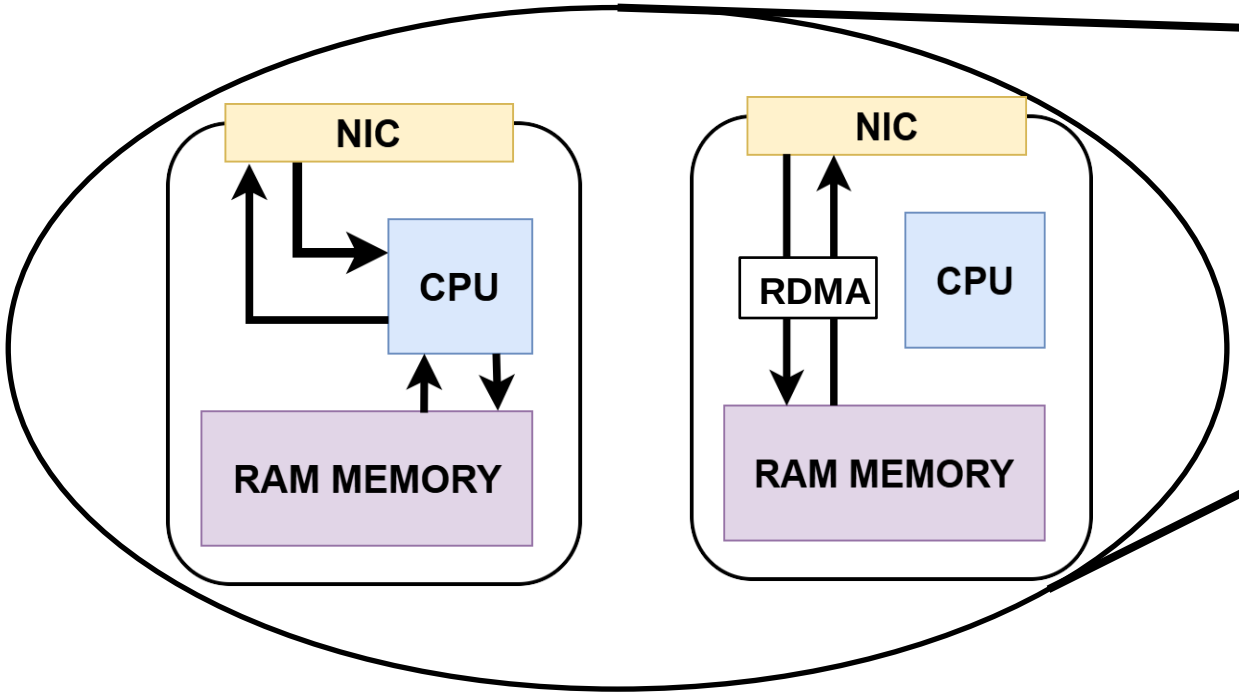
RDMA was developed

~ 30 years ago

IB-based RDMA becomes a standard de facto.

RDMA wa[illegible]

RoCE assumptions no longer fit network requirements [2].



[1] Hoefler et al. "HammingMesh: A Network Topology for Large-Scale Deep Learning." SC22.
[2] T. Hoefler et al., "Data Center Ethernet and Remote Direct Memory Access: Issues at Hyperscale," in Computer.

Introduction – RoCE limitations

To maximize performance,
RDMA require a
lossless networks [2]

[2] Hoefler, T. et al., "Data Center Ethernet and Remote Direct Memory Access: Issues at Hyperscale," in Computer.
[3] De Sensi, D. et al. “An In-Depth Analysis of the Slingshot Interconnect.” SC20.

Introduction – RoCE limitations

To maximize performance,
RDMA require a
lossless netwo

This is the case of IB, but not of ETH

[2] Hoefler, T. et al., "Data Center Ethernet and Remote Direct Memory Access: Issues at Hyperscale," in Computer.
[3] De Sensi, D. et al. “An In-Depth Analysis of the Slingshot Interconnect.” SC20.

Introduction – RoCE limitations

To maximize performance, RDMA require a **lossless networks** [2]

PFC protocol and larger header to address the deriving issues.

[2] Hoefler, T. et al., "Data Center Ethernet and Remote Direct Memory Access: Issues at Hyperscale," in Computer.
[3] De Sensi, D. et al. “An In-Depth Analysis of the Slingshot Interconnect.” SC20.

Introduction – RoCE limitations

To maximize performance, RDMA require a **lossless networks** [2]

PFC protocol and larger header to address the derivi[illegible]sues.

They **reduce efficency** [2].

[2] Hoefler, T. et al., "Data Center Ethernet and Remote Direct Memory Access: Issues at Hyperscale," in Computer.

[3] De Sensi, D. et al. “An In-Depth Analysis of the Slingshot Interconnect.” SC20.

Introduction – RoCE limitations

To maximize performance, RDMA require a **lossless networks** [2]

PFC protocol and larger header to address the derivi[illegible]sues

They **reduce efficency** [2]

Require strong **tuning** [2].

[2] Hoefler, T. et al., "Data Center Ethernet and Remote Direct Memory Access: Issues at Hyperscale," in Computer.

[3] De Sensi, D. et al. “An In-Depth Analysis of the Slingshot Interconnect.” SC20.

Introduction – RoCE limitations

To maximize performance, RDMA require a **lossless networks** [2]

PFC protocol and larger header to address the derivi[illegible]sues

They **reduce efficency** [2]

Require strong **tuning** [2].

Updated standard protocols → Ultra Ethernet Consortium

[2] Hoefler, T. et al., "Data Center Ethernet and Remote Direct Memory Access: Issues at Hyperscale," in Computer.

[3] De Sensi, D. et al. “An In-Depth Analysis of the Slingshot Interconnect.” SC20.

Introduction – RoCE limitations

To maximize performance, RDMA require a **lossless networks** [2]

PFC protocol and larger header to address the derivi[illegible]sues

They **reduce efficency** [2]

Require strong **tuning** [2].

Updated standard protocols → Ultra Ethernet

Not released yet

[2] Hoefler, T. et al., "Data Center Ethernet and Remote Direct Memory Access: Issues at Hyperscale," in Computer.
[3] De Sensi, D. et al. “An In-Depth Analysis of the Slingshot Interconnect.” SC20.

Introduction – RoCE limitations

To maximize performance, RDMA require a **lossless networks** [2]

PFC protocol and larger header to address the derivi[illegible]sues

They **reduce efficency** [2]

Require strong **tuning** [2].

Updated standard protocols → Ultra Ethernet

Not released yet

Standard protocols with smart networks → NVIDIA - SpectrumX

[2] Hoefler, T. et al., "Data Center Ethernet and Remote Direct Memory Access: Issues at Hyperscale," in Computer.
[3] De Sensi, D. et al. “An In-Depth Analysis of the Slingshot Interconnect.” SC20.

Introduction – RoCE limitations

To maximize performance, RDMA require a **lossless networks** [2]

PFC protocol and larger header to address the derivi[illegible]sues

They **reduce efficency** [2]

Require strong **tuning** [2].

- Updated standard protocols → Ultra Ethernet
  Not released yet
- Standard protocols with smart networks → NVIDIA - SpectrumX
- Vendors customized versions
  - HPE-Cray - Slingshot
  - Huawei Lossless ETH

[2] Hoefler, T. et al., "Data Center Ethernet and Remote Direct Memory Access: Issues at Hyperscale," in Computer.

[3] De Sensi, D. et al. “An In-Depth Analysis of the Slingshot Interconnect.” SC20.

Introduction – RoCE limitations

To maximize performance, RDMA require a **lossless networks** [2]

PFC protocol and larger header to address the derivi[illegible]sues

They **reduce efficency** [2]

Require strong **tuning** [2].

- Updated standard protocols → Ultra Ethernet
  - **Not released** yet
- Standard protocols with smart networks → NVIDIA - SpectrumX
- Vendors customized versions
  - HPE-Cray - Slingshot
    - Benchmarked in [3]
  - Huawei Lossless ETH

[2] Hoefler, T. et al., "Data Center Ethernet and Remote Direct Memory Access: Issues at Hyperscale," in Computer.

[3] De Sensi, D. et al. "An In-Depth Analysis of the Slingshot Interconnect." SC20.

Introduction – RoCE limitations

To maximize performance, RDMA require a **lossless networks** [2]

PFC protocol and larger header to address the derivi[illegible]sues

They **reduce efficency** [2]

Require strong **tuning** [2].

- Updated standard protocols → Ultra Ethernet
  - **Not released** yet
- Standard protocols with smart networks → NVIDIA - SpectrumX
- Vendors customized versions
  - HPE-Cray - Slingshot
    - Benchmarked in [3]
  - Huawei Lossless ETH
    - What we benchmark

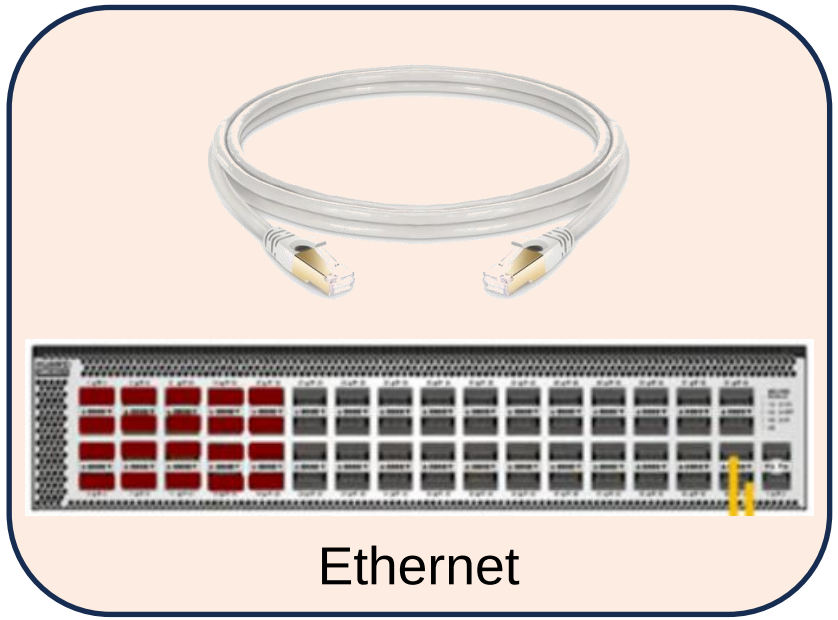
[2] Hoefler, T. et al., "Data Center Ethernet and Remote Direct Memory Access: Issues at Hyperscale," in Computer.

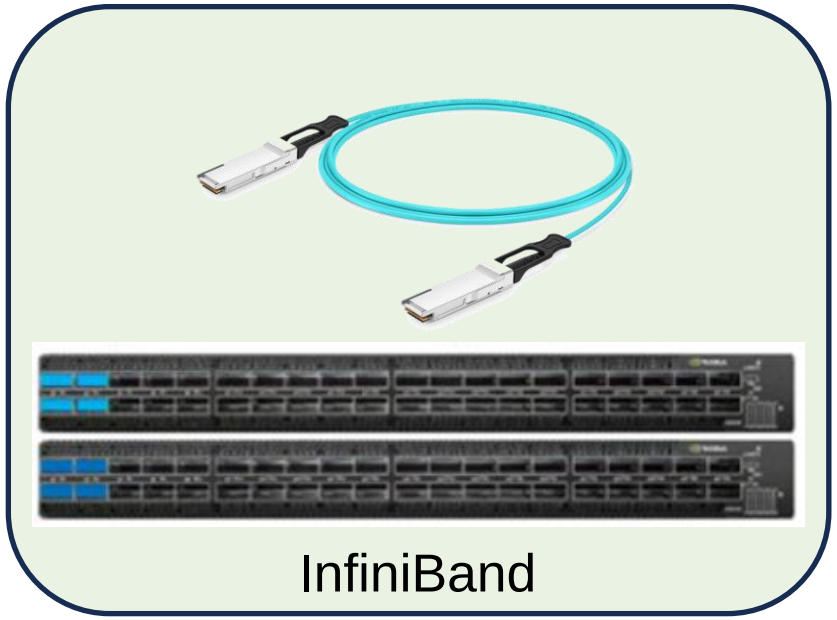
[3] De Sensi, D. et al. "An In-Depth Analysis of the Slingshot Interconnect." SC20.

Method – Our goals

Our goal is to **benchmark lossless ethernet** performance in an HPC scenario.

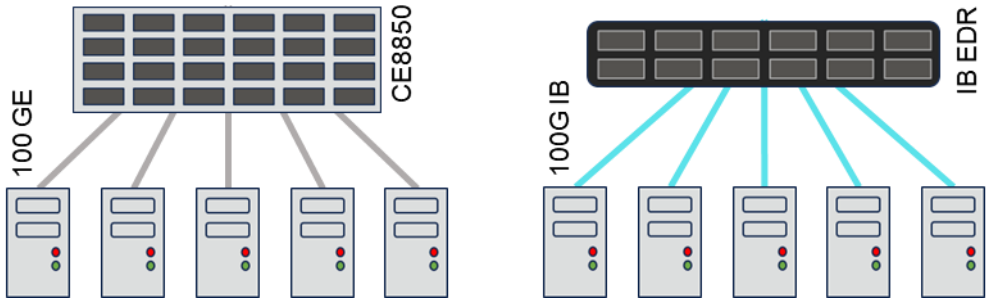
1. Lossless ETH *vs* IB comparison
2. System benchmark

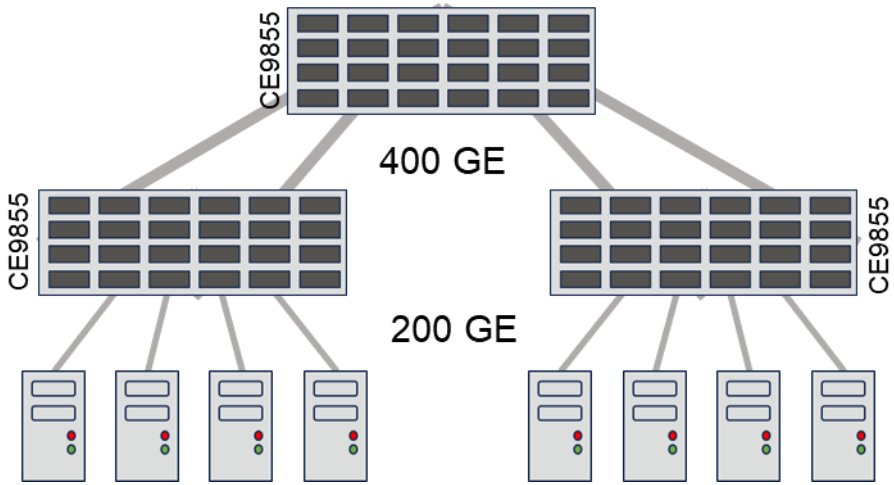




Method – Our resources

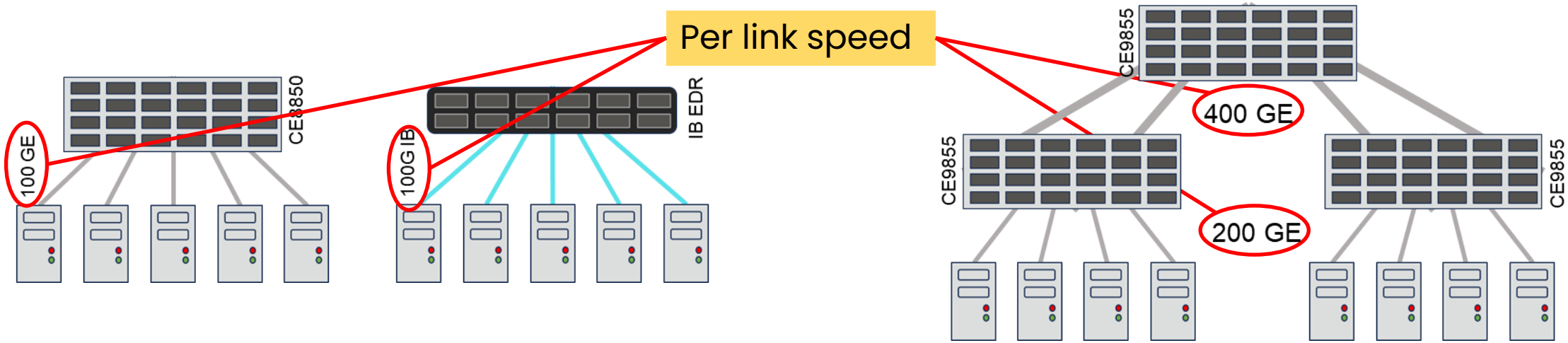
Testsbed: **HAICGU system** and **Nanjing cluster**, respectively provided by OEHI and Huawei.





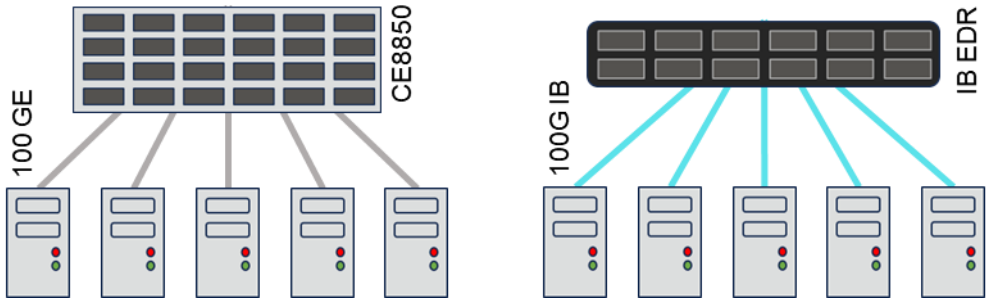
Method – Our resources

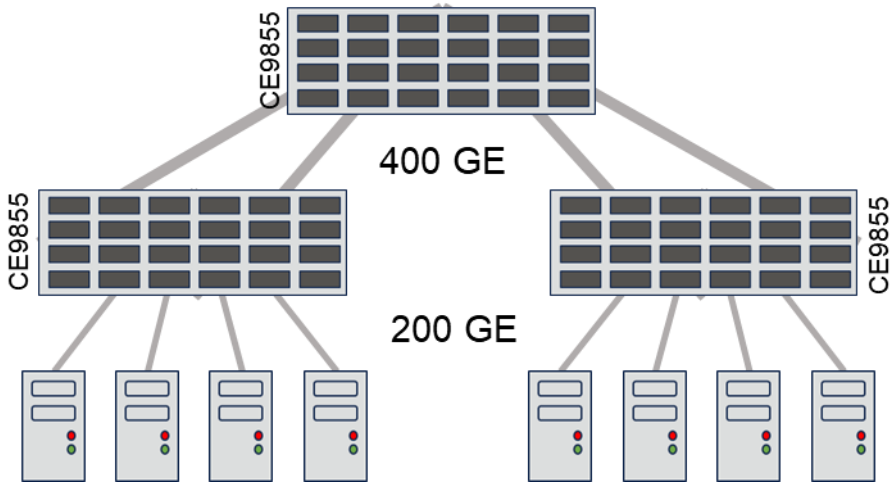
Testsbed: **HAICGU system** and **Nanjing cluster**, respectively provided by OEHI and Huawei.



Method – Our resources

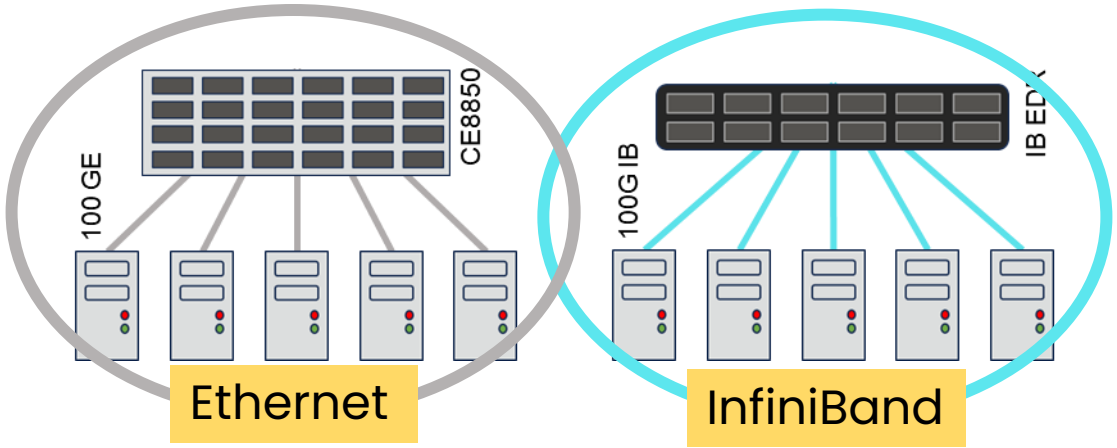
Testsbed: **HAICGU system** and **Nanjing cluster**, respectively provided by OEHI and Huawei.

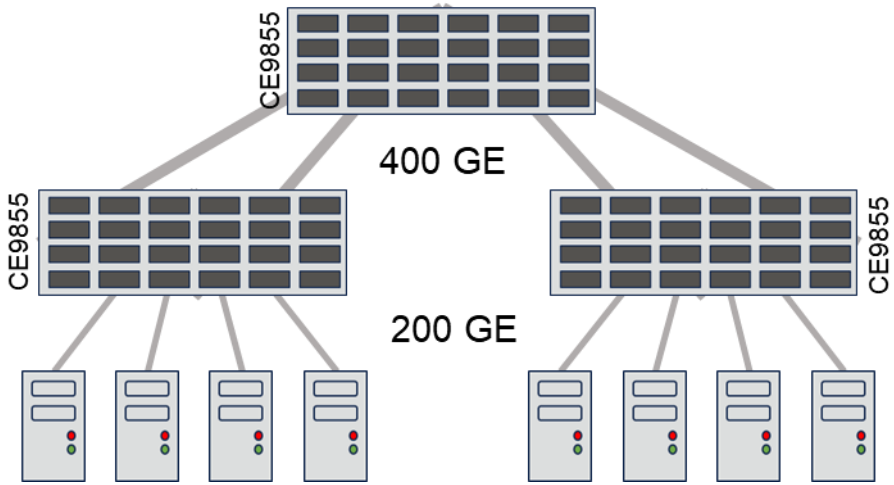




Method – Our resources

Testsbed: **HAICGU system** and **Nanjing cluster**, respectively provided by OEHI and Huawei.

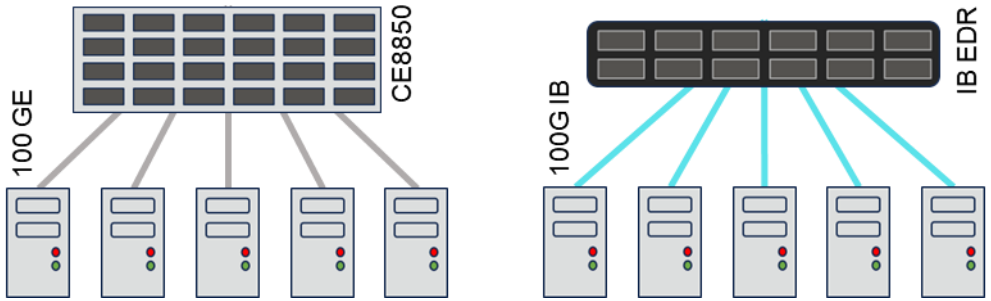


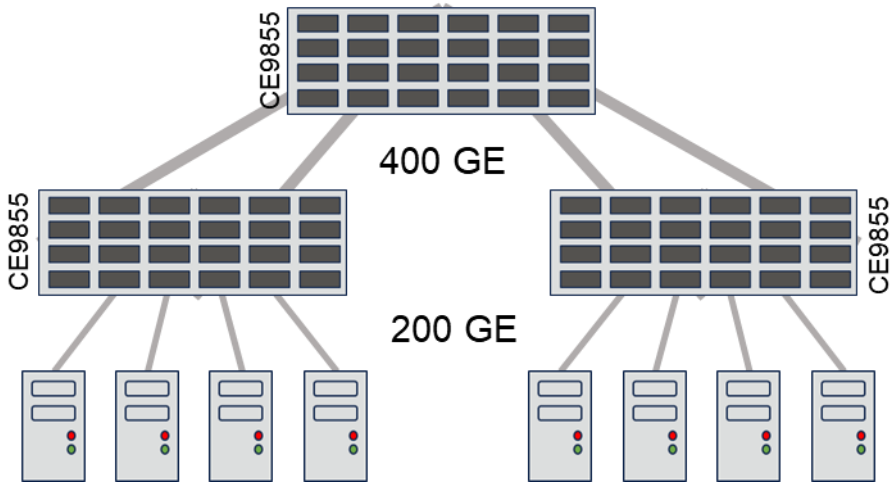


- **IB vs ETH** comparable partitions.
- Identical host, same topology.
- 100 Gbit/s peak connection.

Method – Our resources

Testsbed: **HAICGU system** and **Nanjing cluster**, respectively provided by OEHI and Huawei.

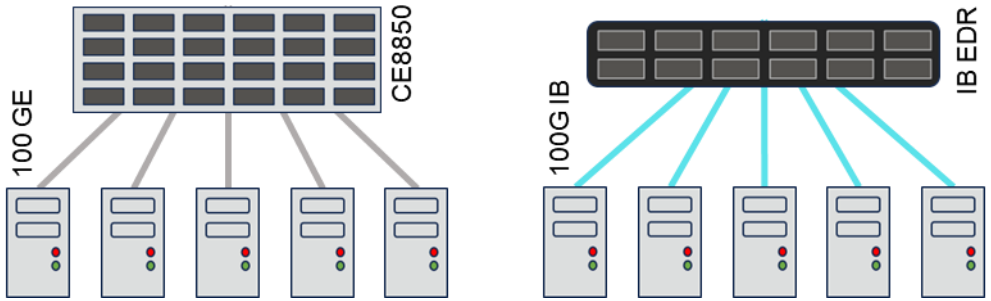




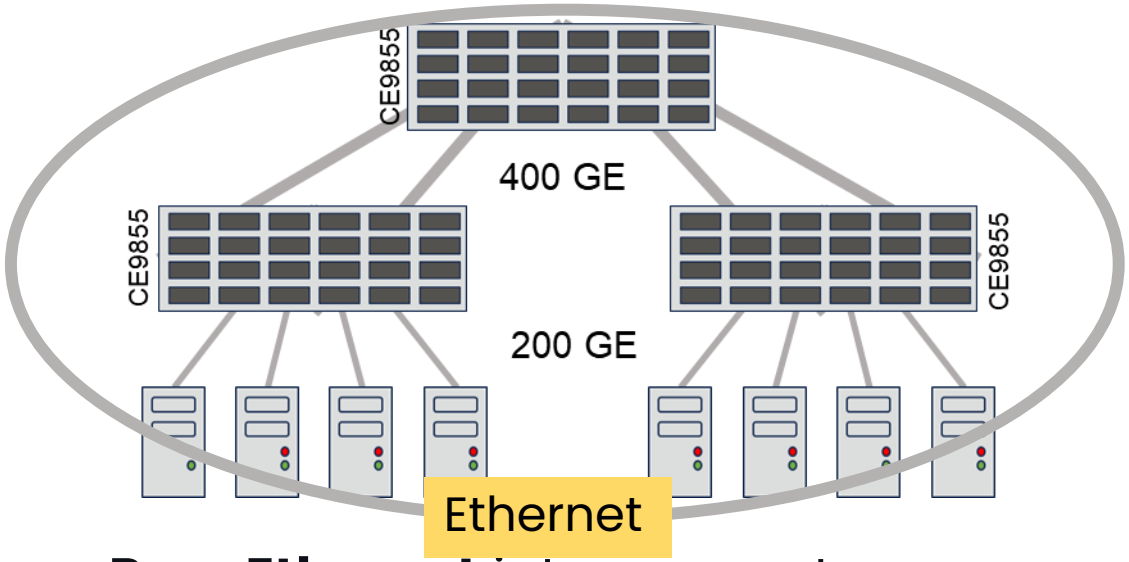
- **IB vs ETH** comparable partitions.
- Identical host, same topology.
- 100 Gbit/s peak connection.

Method – Our resources

Testsbed: **HAICGU system** and **Nanjing cluster**, respectively provided by OEHI and Huawei.



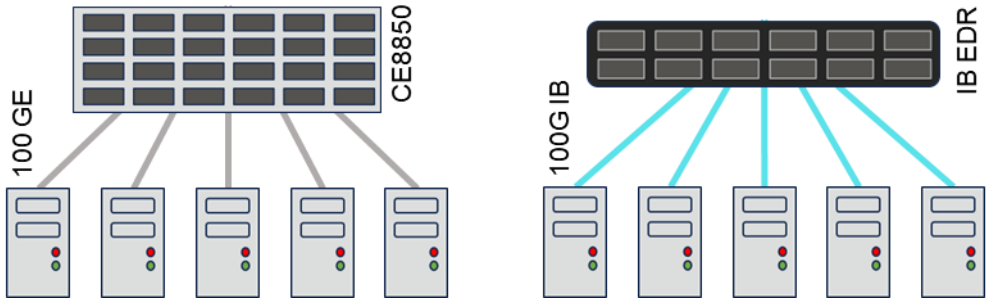
- **IB vs ETH** comparable partitions.
- Identical host, same topology.
- 100 Gbit/s peak connection.

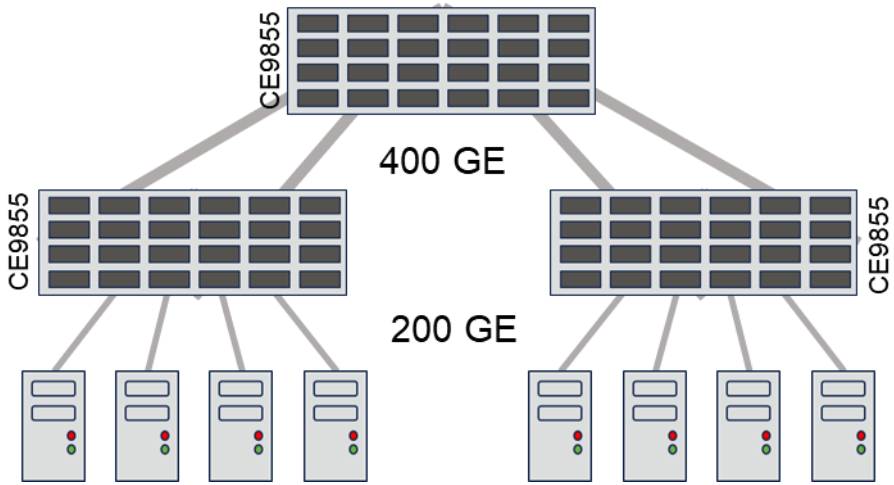


- **Pure Ethernet** interconnect.
- Two-level fat-tree topology.
- Cutting-edge switches.
- 200 Gbit/s peak connection.

Method – Our resources

Testsbed: **HAICGU system** and **Nanjing cluster**, respectively provided by OEHI and Huawei.



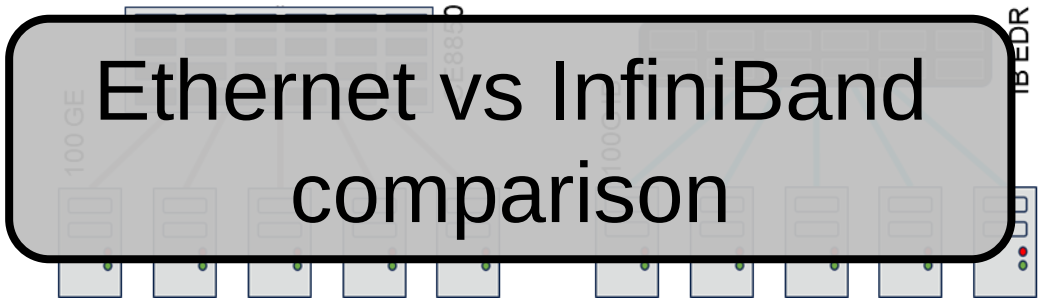


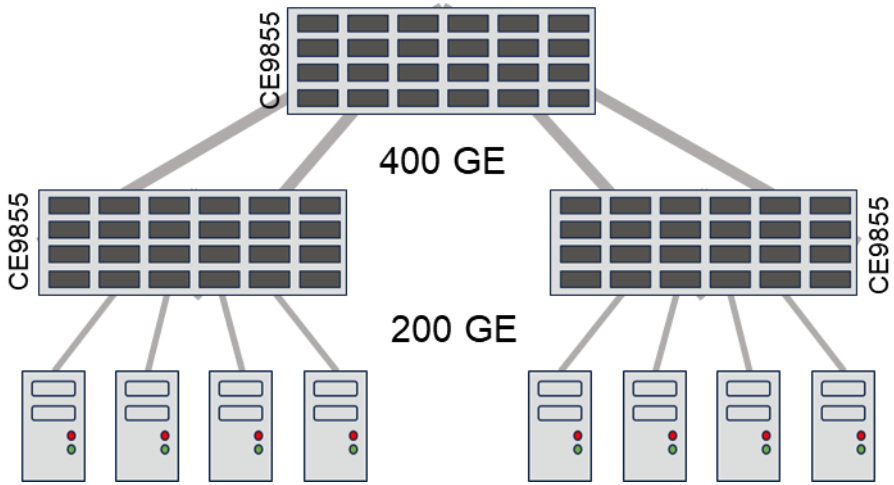
- **IB vs ETH** comparable partitions.
- Identical host, same topology.
- 100 Gbit/s peak connection.

- **Pure Ethernet** interconnect.
- Two-level fat-tree topology.
- Cutting-edge switches.
- 200 Gbit/s peak connection.

Method – Our resources

Testsbed: **HAICGU system** and **Nanjing cluster**, respectively provided by OEHI and Huawei.



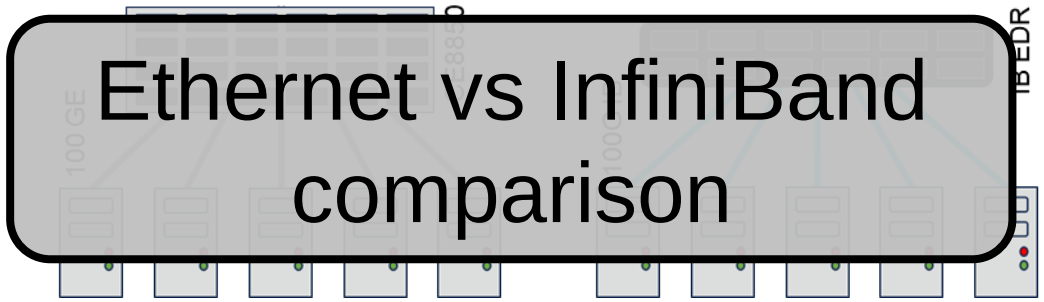


- **IB vs ETH** comparable partitions.
- Identical host, same topology.
- 100 Gbit/s peak connection.

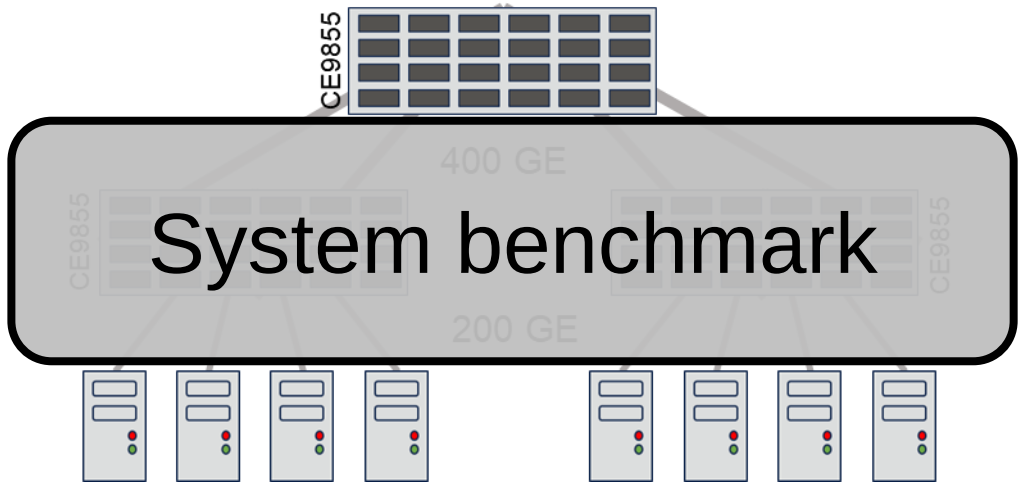
- **Pure Ethernet** interconnect.
- Two-level fat-tree topology.
- Cutting-edge switches.
- 200 Gbit/s peak connection.

Method – Our resources

Testsbed: **HAICGU system** and **Nanjing cluster**, respectively provided by OEHI and Huawei.



- **IB vs ETH** comparable partitions.
- Identical host, same topology.
- 100 Gbit/s peak connection.

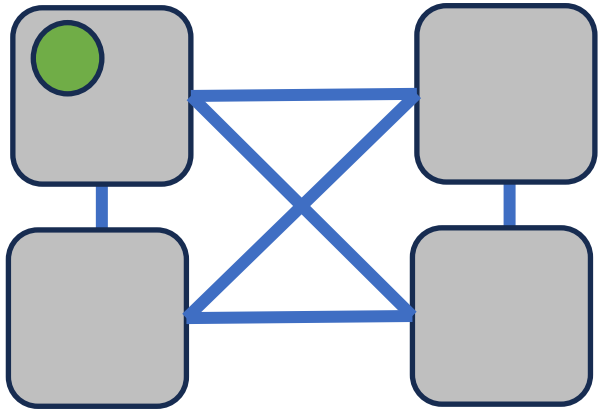


- **Pure Ethernet** interconnect.
- Two-level fat-tree topology.
- Cutting-edge switches.
- 200 Gbit/s peak connection.

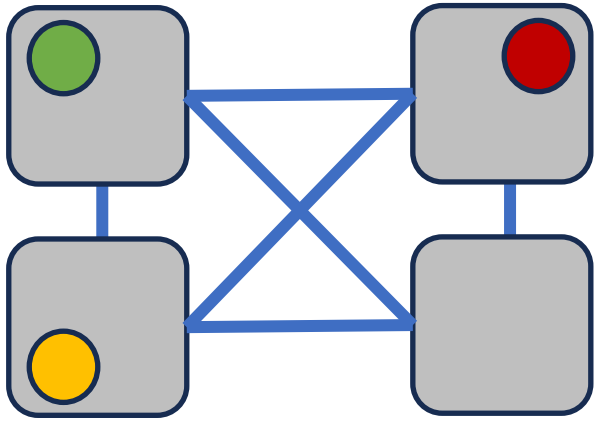
Method – Our tests

AI and HPC workloads rely on primitives of simple communication.

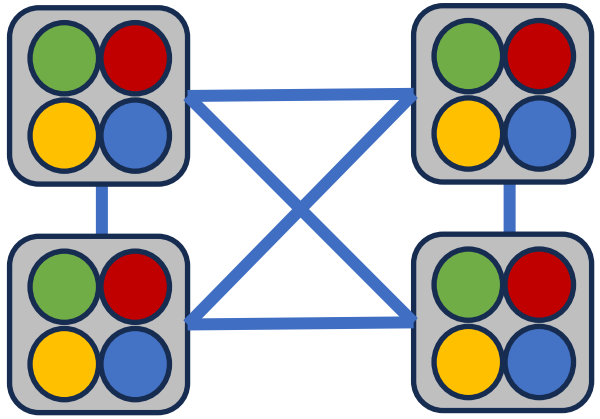
Benchmark suite used: **BLink** [4].



Peer-To-Peer communication



incast communication



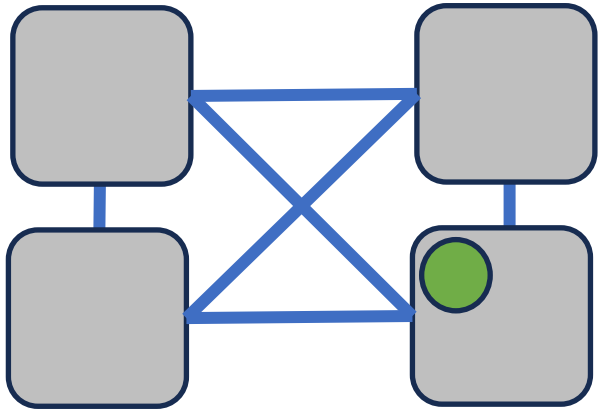
Collective communication

[4] GitHub source: https://github.com/DanieleDeSensi/blink

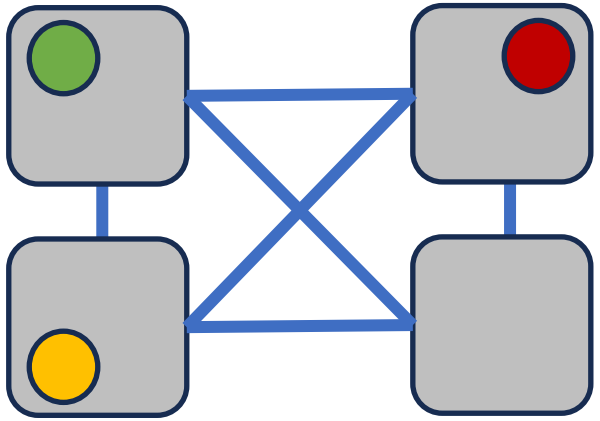
Method – Our tests

AI and HPC workloads rely on primitives of simple communication.

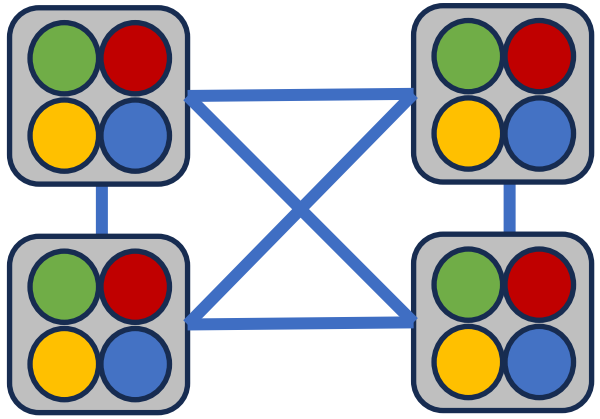
Benchmark suite used: **BLink** [4].



Peer-To-Peer communication



incast communication



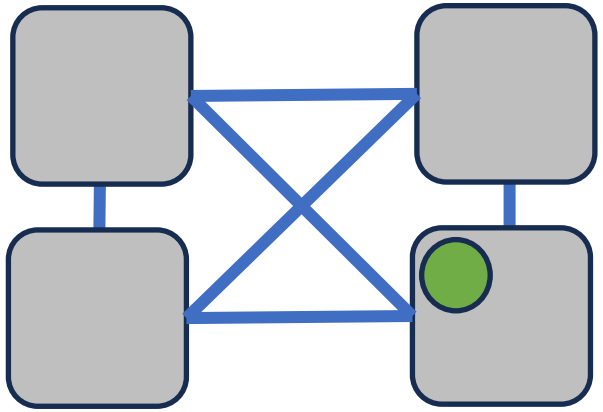
Collective communication

[4] GitHub source: https://github.com/DanieleDeSensi/blink

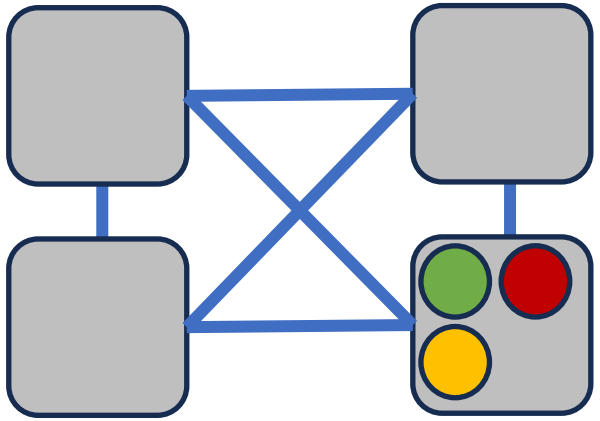
Method – Our tests

AI and HPC workloads rely on primitives of simple communication.

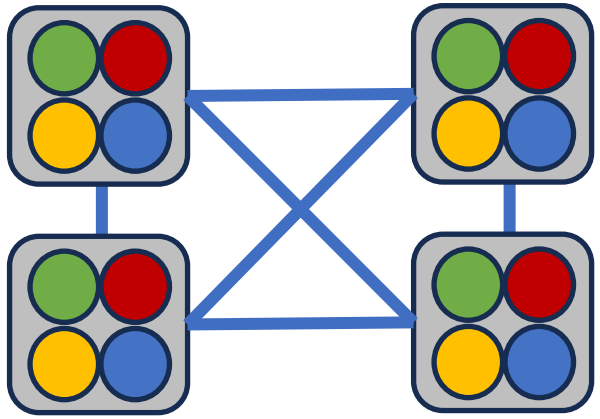
Benchmark suite used: **BLink** [4].



Peer-To-Peer communication



incast communication



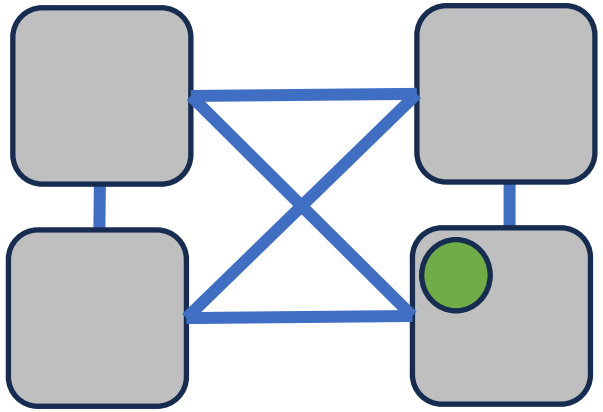
Collective communication

[4] GitHub source: https://github.com/DanieleDeSensi/blink

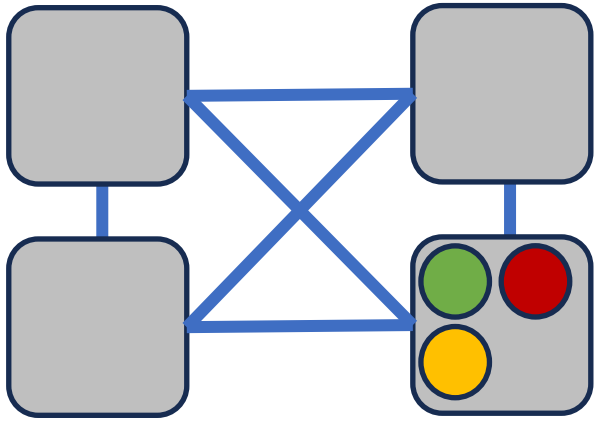
Method – Our tests

AI and HPC workloads rely on primitives of simple communication.

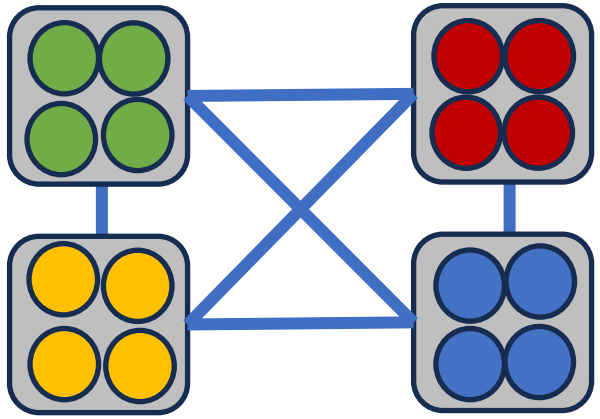
Benchmark suite used: **BLink** [4].



Peer-To-Peer communication



incast communication



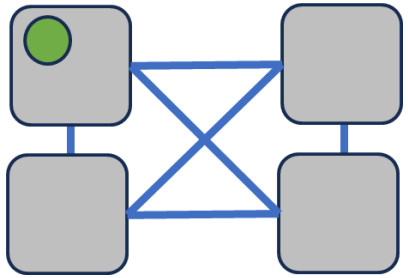
Collective communication

[4] GitHub source: https://github.com/DanieleDeSensi/blink

Method – Our expectations

Method – Our expectations

Peer-To-Peer

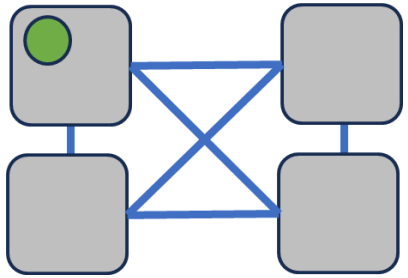


- Benchmarks a **single connection**.
- We expect to **saturate** theoretical **bandwidth**.

[2] Hoefler et al., "Data Center Ethernet and Remote Direct Memory Access: Issues at Hyperscale," in Computer.

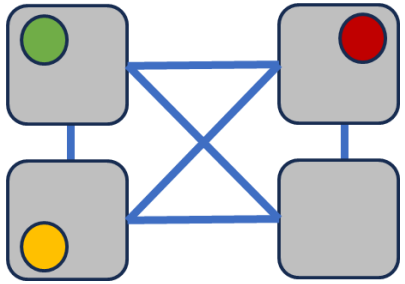
Method – Our expectations

Peer-To-Peer



- Benchmarks a **single connection**.
- We expect to **saturate** theoretical **bandwidth**.

Incast

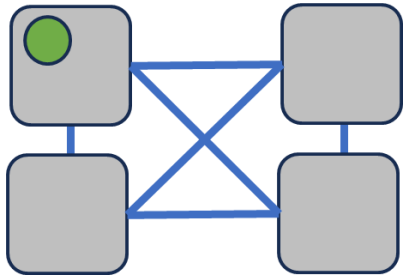


- Real-world **unbalanced situation**.
- **Congestion control** must manage traffic and achieve the theoretical peak.

[2] Hoefler et al., "Data Center Ethernet and Remote Direct Memory Access: Issues at Hyperscale," in Computer.

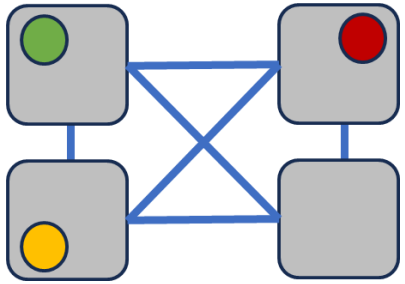
Method – Our expectations

Peer-To-Peer



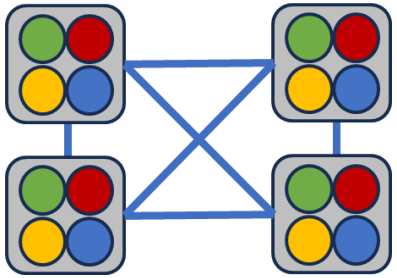
- Benchmarks a **single connection**.
- We expect to **saturate** theoretical **bandwidth**.

Incast



- Real-world **unbalanced situation**.
- **Congestion control** must manage traffic and achieve the theoretical peak.

Collectives

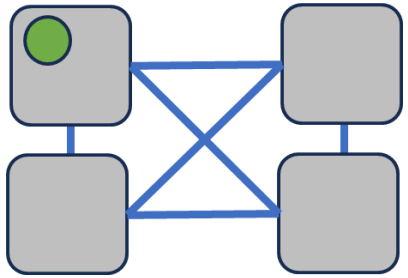


- Real-world operations that **stress the whole interconnect**.
- **The network** must manage the traffic.

[2] Hoefler et al., "Data Center Ethernet and Remote Direct Memory Access: Issues at Hyperscale," in Computer.

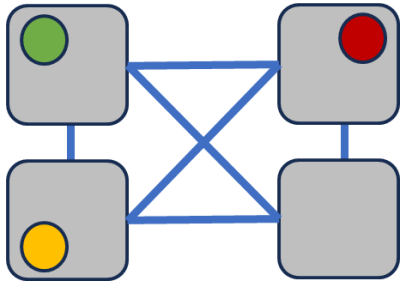
Method – Our expectations

Peer-To-Peer



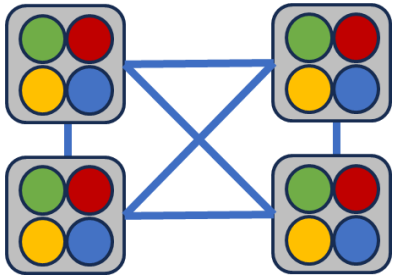
- Benchmarks a **single connection**.
- We expect to **saturate** theoretical **bandwidth**.

Incast



- Real-world **unbalanced situation**.
- **Congestion control** must manage traffic and achieve the theoretical peak.

Collectives



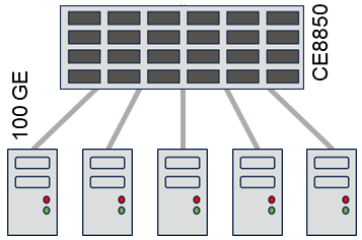
- Real-world operations that **stress the whole interconnect**.
- **The network** must manage the traffic.

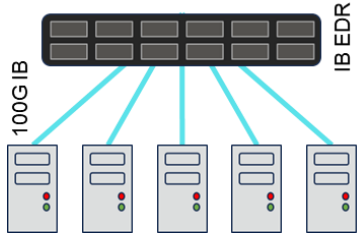
General for all tests

ETH headers are **~3x larger** than IB headers (RoCEv2) [2]. On all tests, we expect better IB performance **on small messages**.

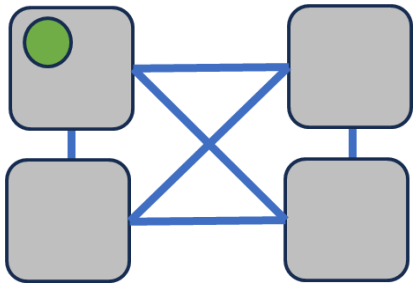
[2] Hoefler et al., "Data Center Ethernet and Remote Direct Memory Access: Issues at Hyperscale," in Computer.

Experimental results: ETH vs IB





	Peer-To-Peer	Incast	All-To-All
ETH vs IB comparison			
System benchmark			

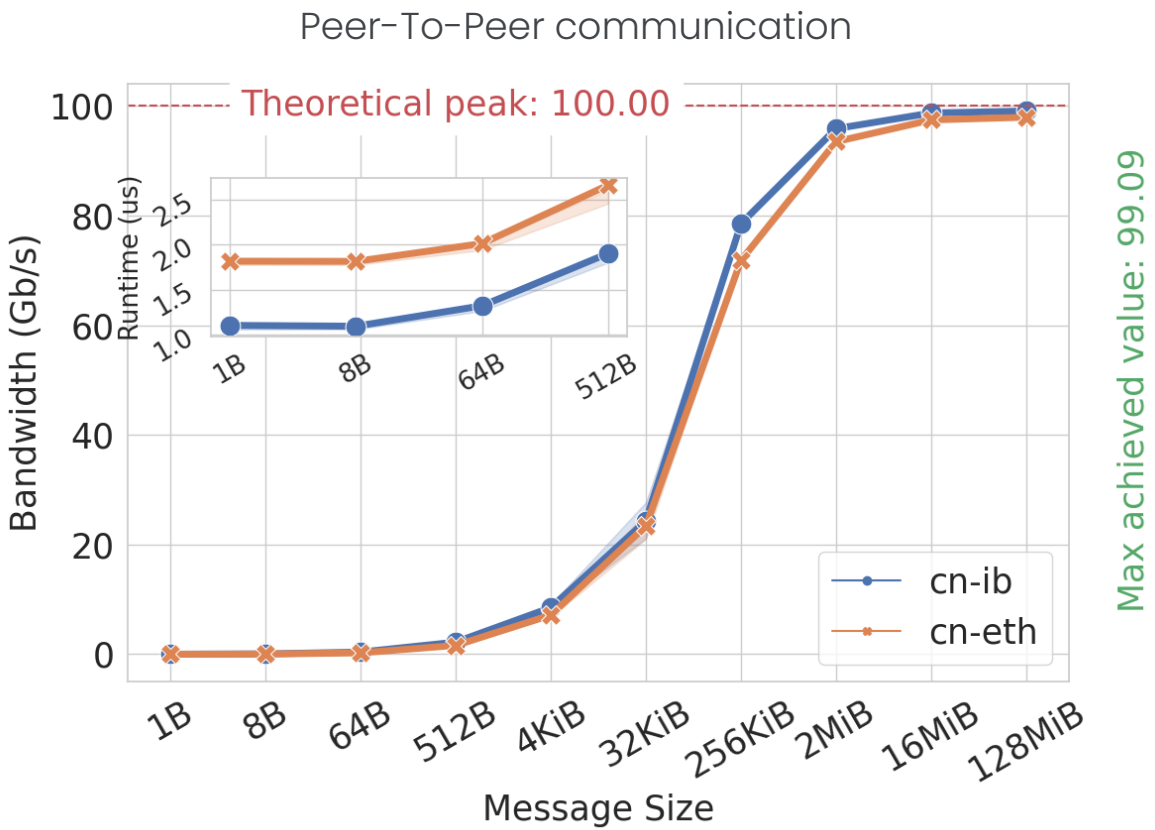


Results – ETH vs IB

Peer-To-Peer comparison:

- The test saturates the nominal bandwidth;
- Approximatively same large message performance;
- IB outperforms ETH over small messages.

Size	cn-ETH (peak%)	cn-IB (peak%)	ETH/IB ratio
1B	0.0045 %	0.0074 %	60.18 %
8B	0.04 %	0.06 %	60.03 %
512B	01.61 %	02.20 %	73.10 %
32KiB	23.40 %	24.32 %	96.21 %
2MiB	93.49 %	95.86 %	97.53 %
16MiB	97.45 %	98.71 %	98.72 %
128MiB	97.94 %	99.07 %	98.86 %

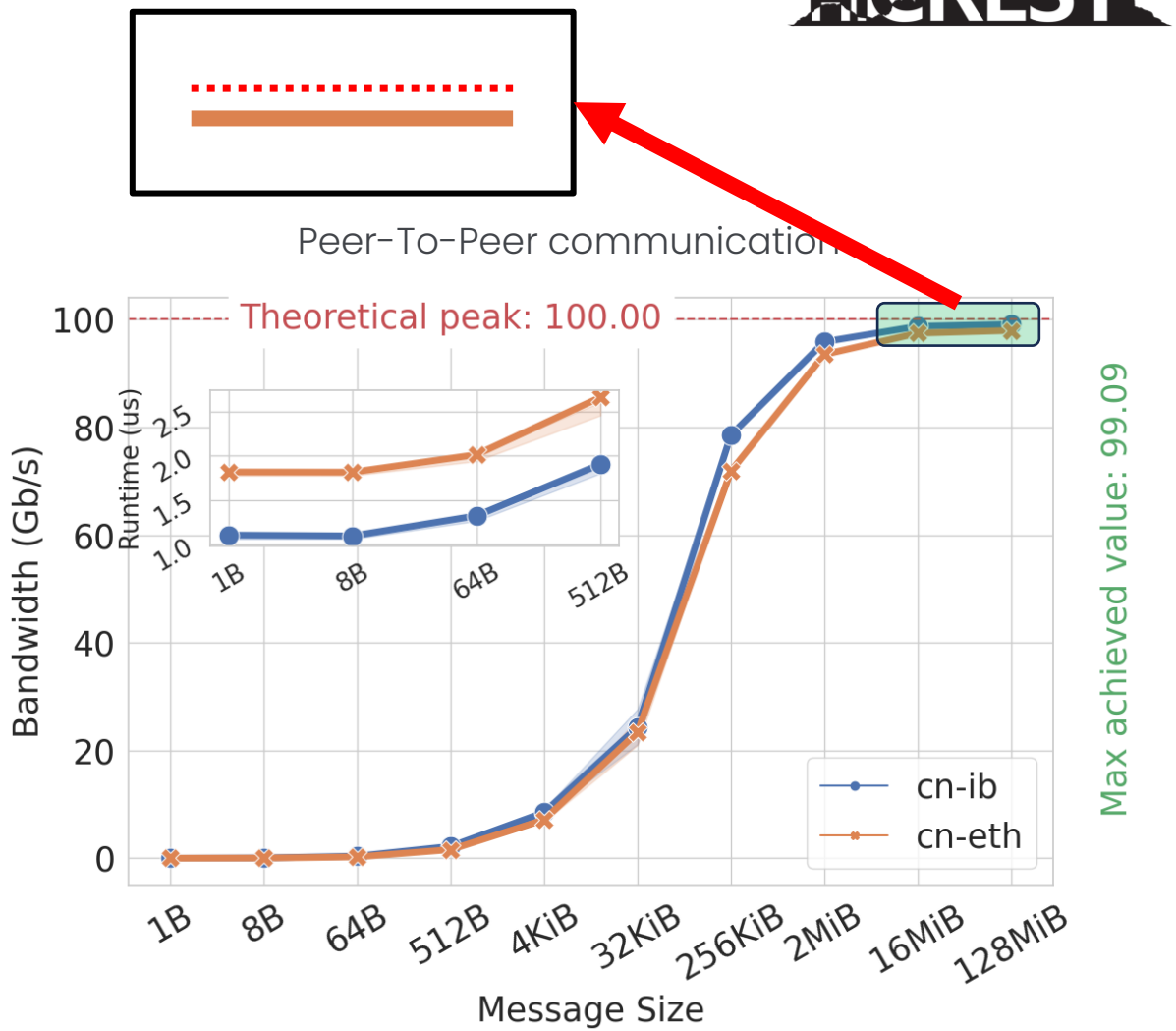


Results – ETH vs IB

Peer-To-Peer comparison:

- The test saturates the nominal bandwidth;
- Approximatively same large message performance;
- IB outperforms ETH over small messages.

Size	cn-ETH (peak%)	cn-IB (peak%)	ETH/IB ratio
1B	0.0045 %	0.0074 %	60.18 %
8B	0.04 %	0.06 %	60.03 %
512B	01.61 %	02.20 %	73.10 %
32KiB	23.40 %	24.32 %	96.21 %
2MiB	93.49 %	95.86 %	97.53 %
16MiB	97.45 %	98.71 %	98.72 %
128MiB	97.94 %	99.07 %	98.86 %

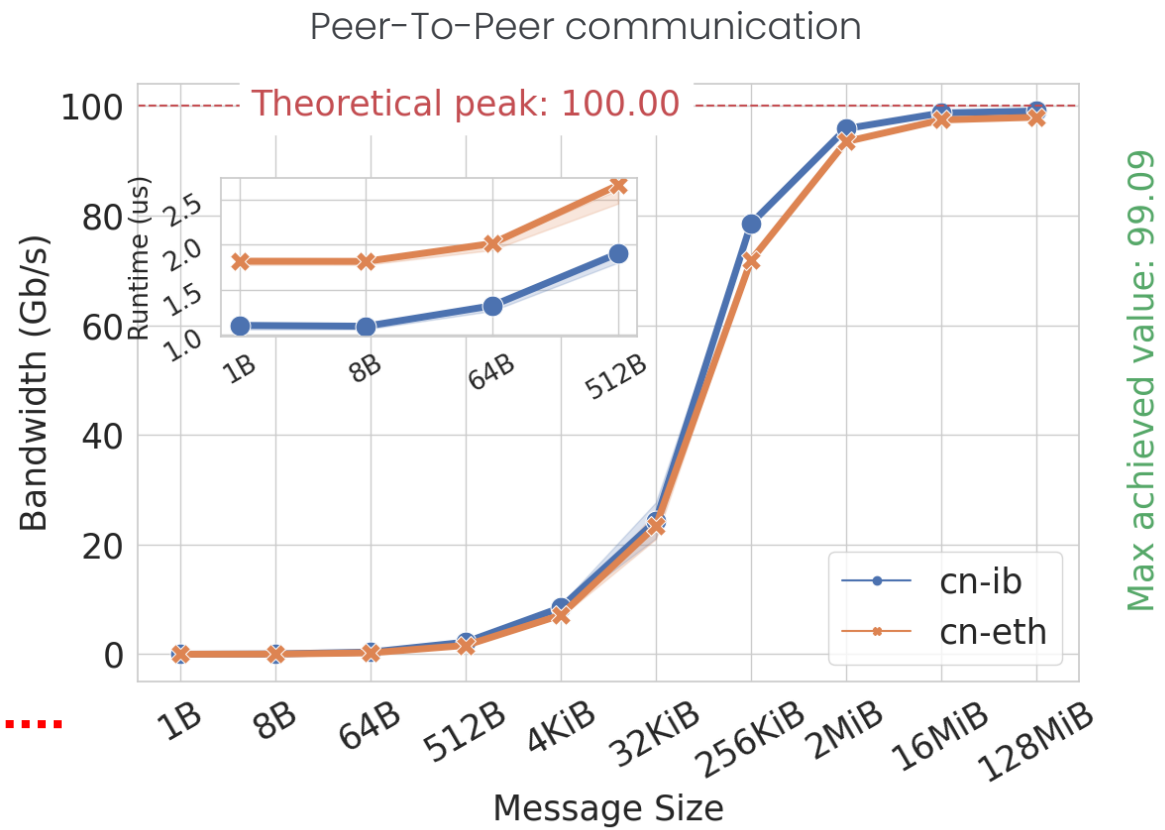


Results – ETH vs IB

Peer-To-Peer comparison:

- The test saturates the nominal bandwidth;
- Approximatively same large message performance;
- IB outperforms ETH over small messages.

Size	cn-ETH (peak%)	cn-IB (peak%)	ETH/IB ratio
1B	0.0045 %	0.0074 %	60.18 %
8B	0.04 %	0.06 %	60.03 %
512B	01.61 %	02.20 %	73.10 %
32KiB	23.40 %	24.32 %	96.21 %
2MiB	93.49 %	95.86 %	97.53 %
16MiB	97.45 %	98.71 %	98.72 %
128MiB	97.94 %	99.07 %	98.86 %

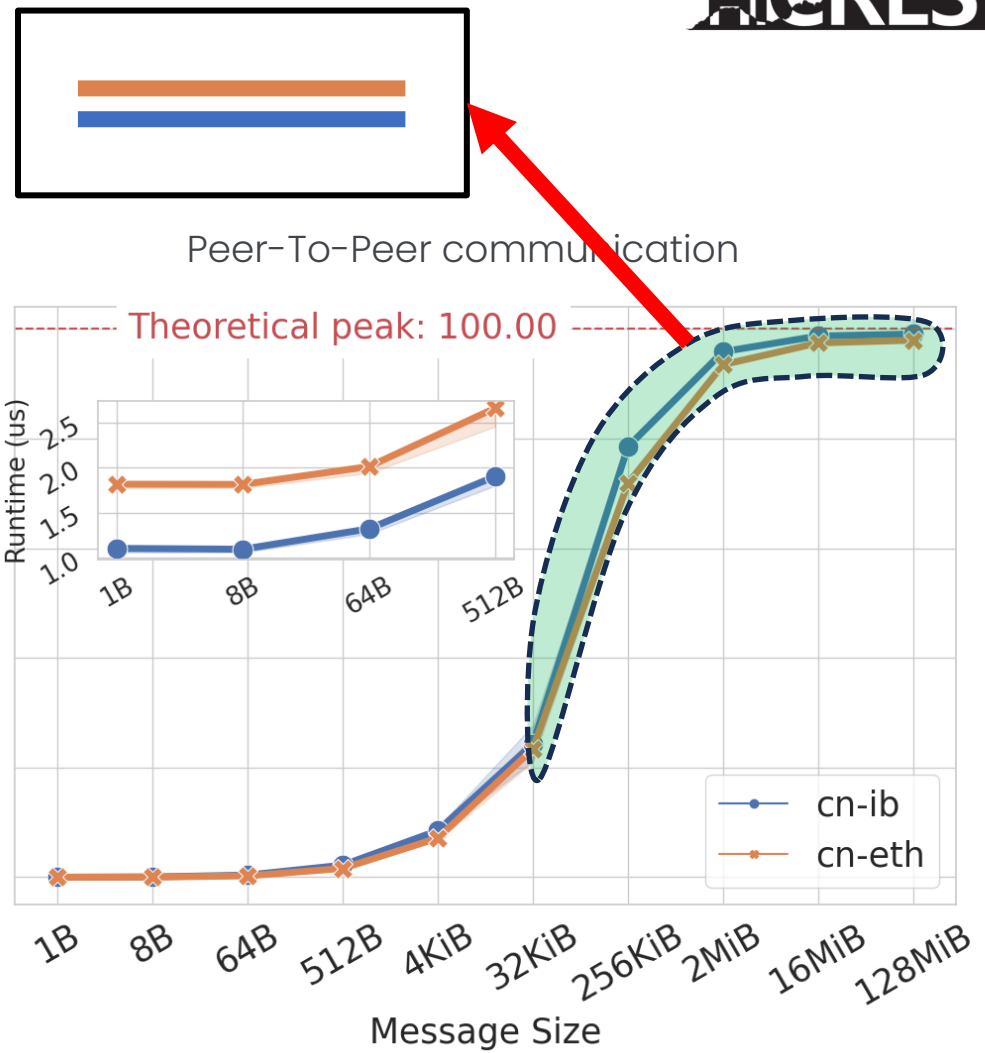


Results – ETH vs IB

Peer-To-Peer comparison:

- The test saturates the nominal bandwidth;
- Approximatively same large message performance;
- IB outperforms ETH over small messages.

Size	cn-ETH (peak%)	cn-IB (peak%)	ETH/IB ratio
1B	0.0045 %	0.0074 %	60.18 %
8B	0.04 %	0.06 %	60.03 %
512B	01.61 %	02.20 %	73.10 %
32KiB	23.40 %	24.32 %	96.21 %
2MiB	93.49 %	95.86 %	97.53 %
16MiB	97.45 %	98.71 %	98.72 %
128MiB	97.94 %	99.07 %	98.86 %
			< 4%



Results – ETH vs IB

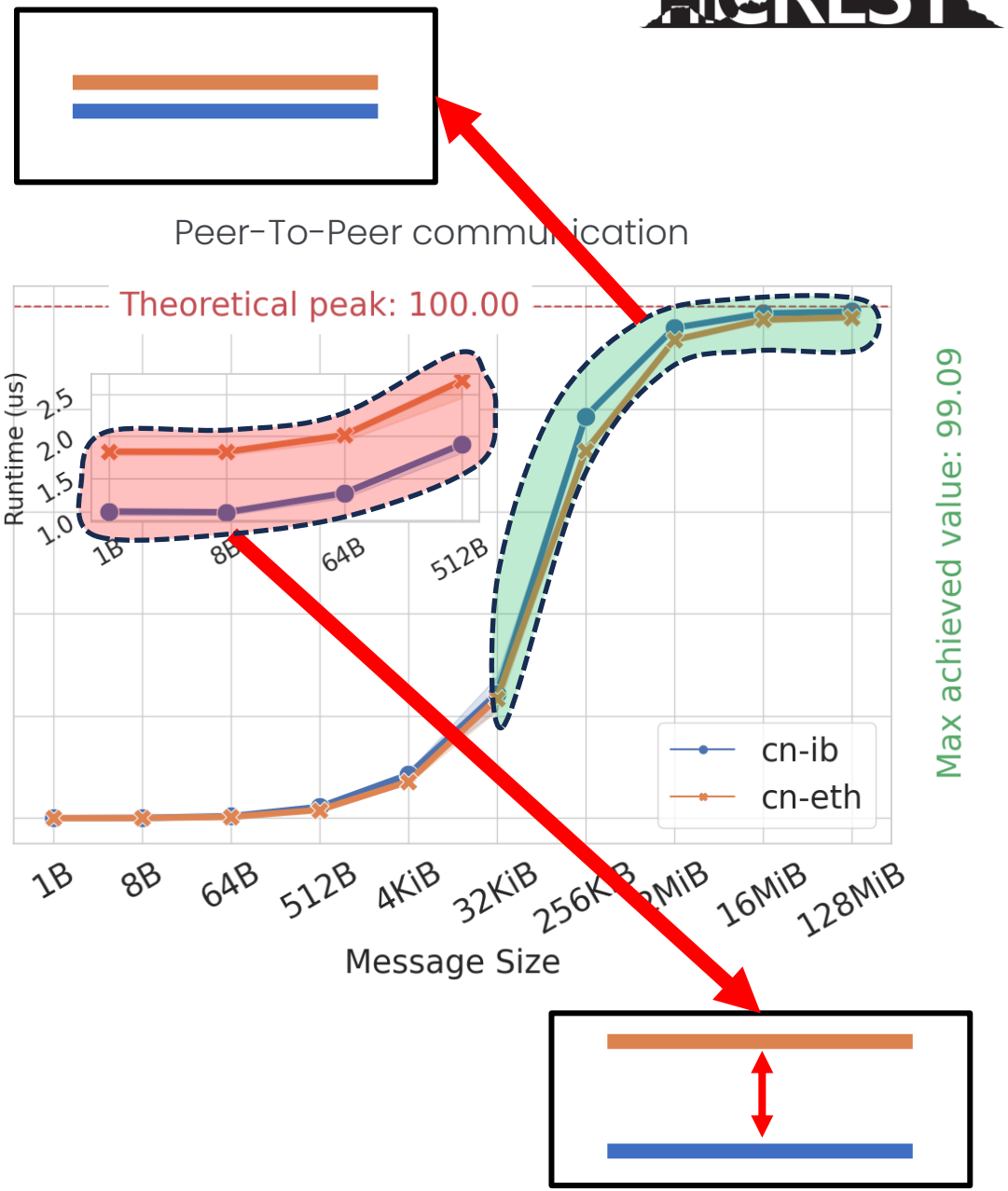
Peer-To-Peer comparison:

- The test saturates the nominal bandwidth;
- Approximatively same large message performance;
- IB outperforms ETH over small messages.

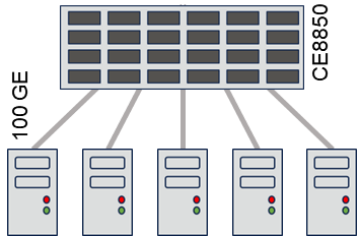
Size	cn-ETH (peak%)	cn-IB (peak%)	ETH/IB ratio
1B	0.0045 %	0.0074 %	60.18 %
8B	0.04 %	0.06 %	60.03 %
512B	01.61 %	02.20 %	73.10 %
32KiB	23.40 %	24.32 %	96.21 %
2MiB	93.49 %	95.86 %	97.53 %
16MiB	97.45 %	98.71 %	98.72 %
128MiB	97.94 %	99.07 %	98.86 %

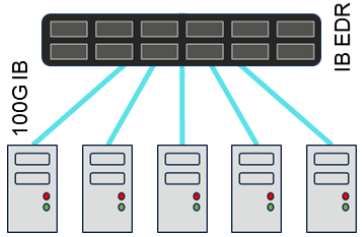
> 30%

< 4%

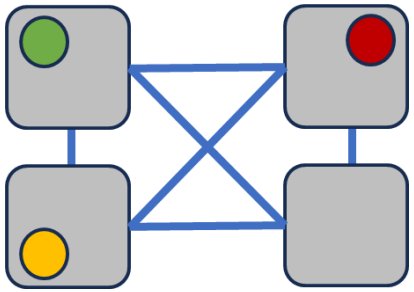


Experimental results: ETH vs IB





	Peer-To-Peer	Incast	All-To-All
ETH vs IB comparison			
System benchmark			

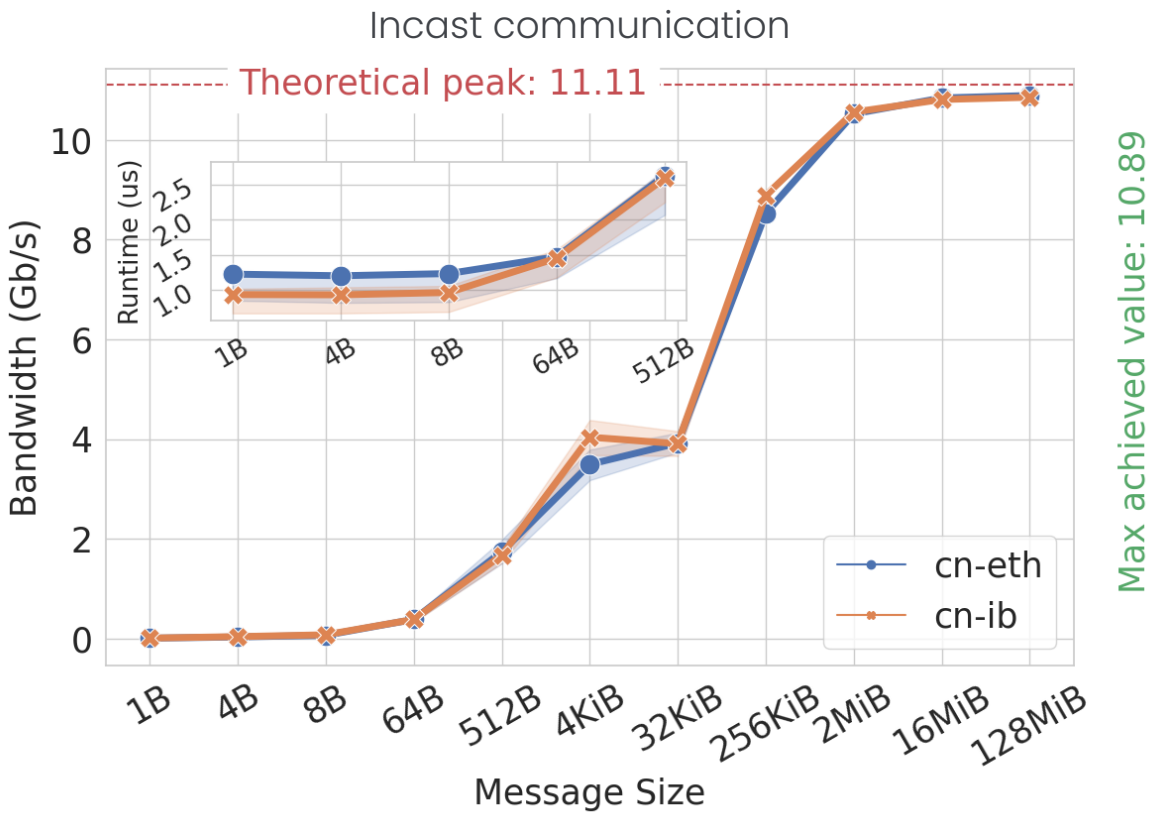


Results – ETH vs IB

Incast comparison:

- Also Incast exposes the expected results.
- *Wrt* Peer-To-Peer, Incast communication:
  - Performed slightly better.
  - Is more sensitive to algorithm change.

Size	cn-ETH (peak%)	cn-IB (peak%)	ETH/IB ratio
1B	0.07 %	0.09 %	81.20 %
8B	0.59 %	0.69 %	85.89 %
512B	15.66 %	15.06 %	103.92 %
32KiB	35.27 %	35.15 %	100.33 %
2MiB	94.66 %	94.98 %	99.67 %
16MiB	97.53 %	97.25 %	100.29 %
128MiB	97.97 %	97.65 %	100.33 %



Results – ETH vs IB

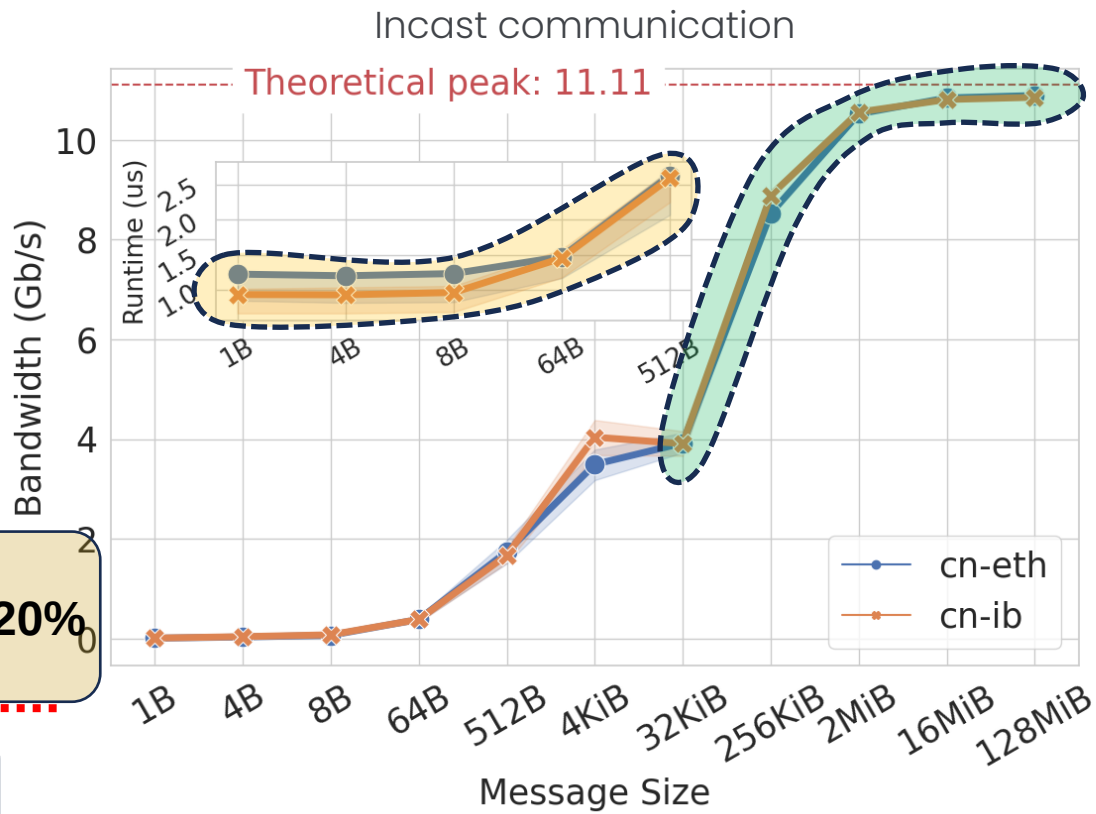
Incast comparison:

- Also Incast exposes the expected results.
- *Wrt* Peer-To-Peer, Incast communication:
  - Performed slightly better.
  - Is more sensitive to algorithm change.

Size	cn-ETH (peak%)	cn-IB (peak%)	ETH/IB ratio
1B	0.07 %	0.09 %	81.20 %
8B	0.59 %	0.69 %	85.89 %
512B	15.66 %	15.06 %	103.92 %
32KiB	35.27 %	35.15 %	100.33 %
2MiB	94.66 %	94.98 %	99.67 %
16MiB	97.53 %	97.25 %	100.29 %
128MiB	97.97 %	97.65 %	100.33 %

< 20%

~ same



Results – ETH vs IB

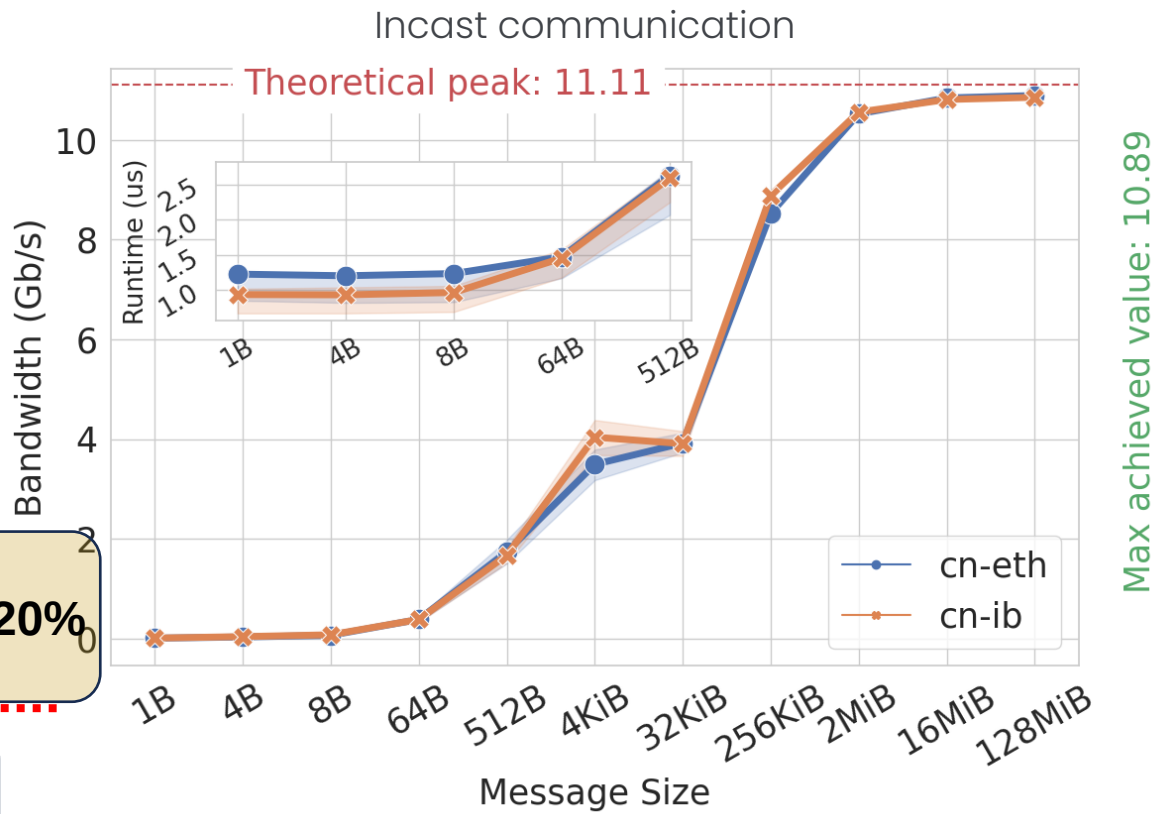
Incast comparison:

- Also Incast exposes the expected results.
- *Wrt* Peer-To-Peer, Incast communication:
  - Performed slightly better.
  - Is more sensitive to algorithm change.

Size	cn-ETH (peak%)	cn-IB (peak%)	ETH/IB ratio
1B	0.07 %	0.09 %	81.20 %
8B	0.59 %	0.69 %	85.89 %
512B	15.66 %	15.06 %	103.92 %
32KiB	35.27 %	35.15 %	100.33 %
2MiB	94.66 %	94.98 %	99.67 %
16MiB	97.53 %	97.25 %	100.29 %
128MiB	97.97 %	97.65 %	100.33 %

< 20%

~ same



Results – ETH vs IB

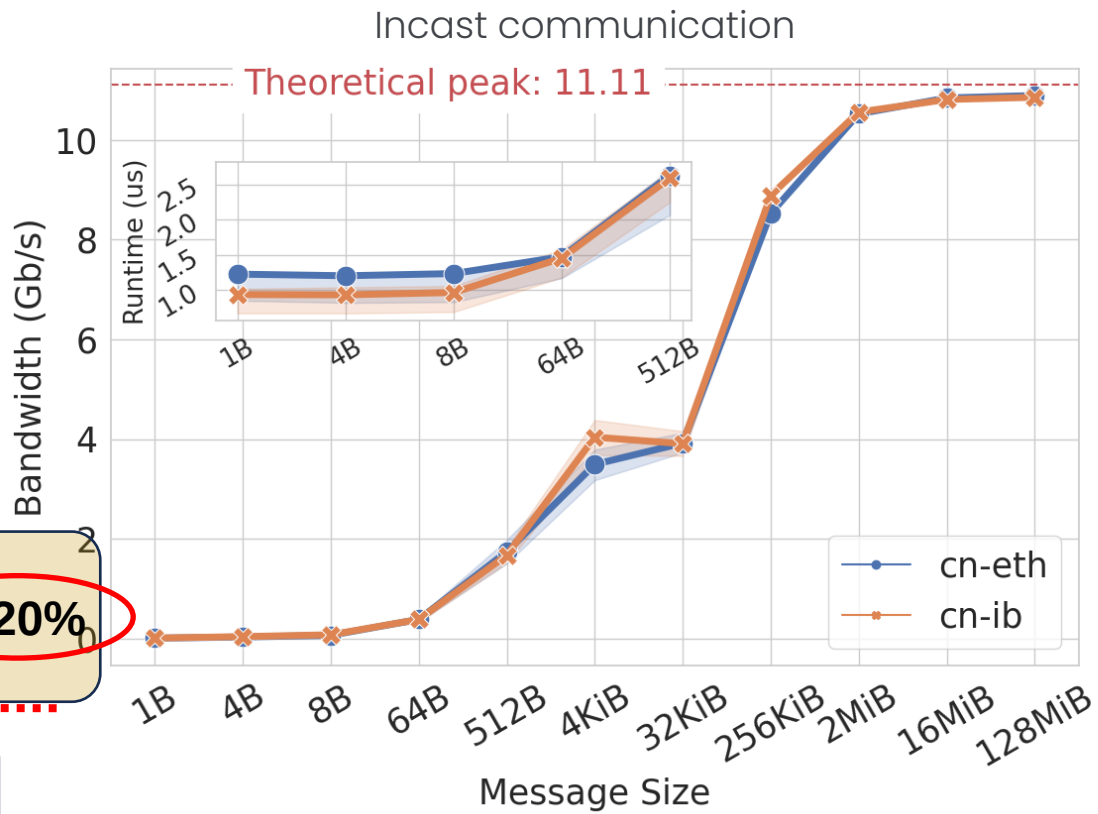
Incast comparison:

- Also Incast exposes the expected results.
- *Wrt* Peer-To-Peer, Incast communication:
  - Performed slightly better.
  - Is more sensitive to algorithm change.

Size	cn-ETH (peak%)	cn-IB (peak%)	ETH/IB ratio
1B	0.07 %	0.09 %	81.20 %
8B	0.59 %	0.69 %	85.89 %
512B	15.66 %	15.06 %	103.92 %
32KiB	35.27 %	35.15 %	100.33 %
2MiB	94.66 %	94.98 %	99.67 %
16MiB	97.53 %	97.25 %	100.29 %
128MiB	97.97 %	97.65 %	100.33 %

< 20%

~ same



Results – ETH vs IB

Incast comparison:

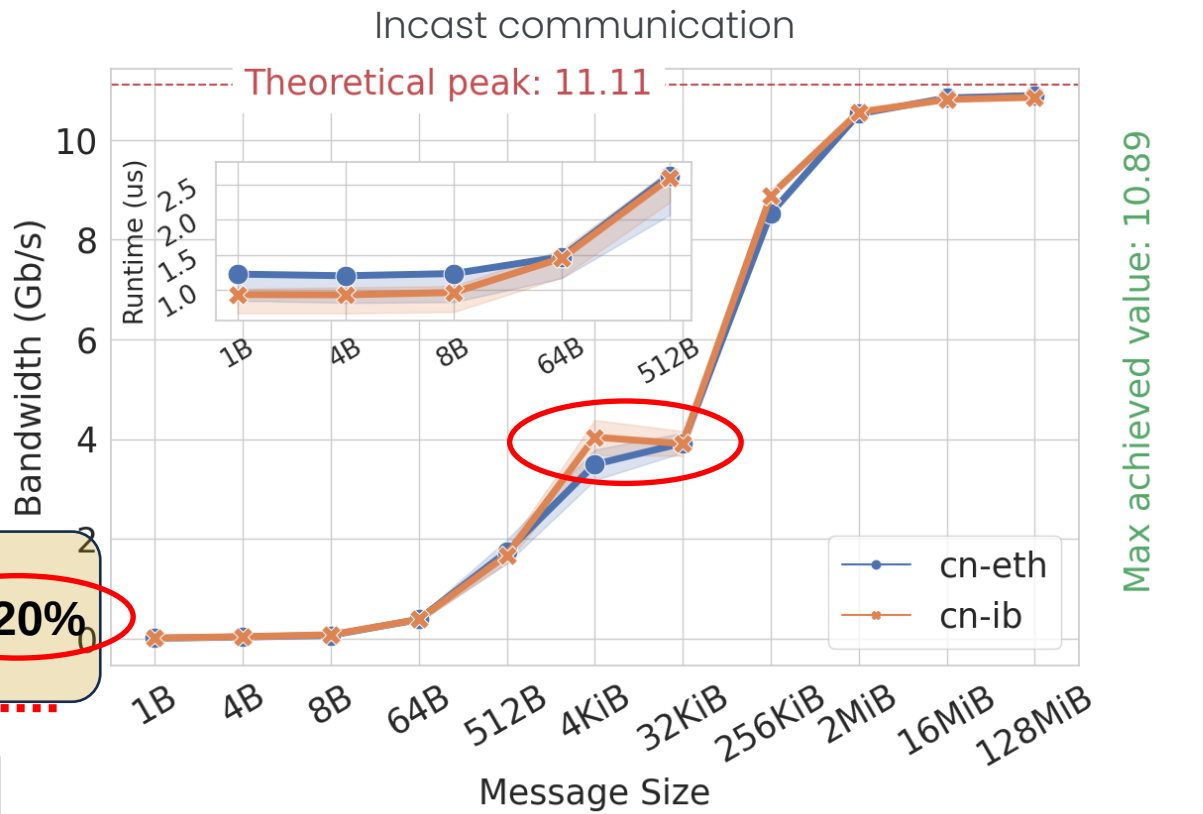
- Also Incast exposes the expected results.
- *Wrt* Peer-To-Peer, Incast communication:
  - Performed slightly better.
  - Is more sensitive to algorithm change.

From an **Eager** to a **Rendezvous** protocol.

	cn-ETH	cn-IB (peak%)	ETH/IB ratio
		0.09 %	81.20 %
8B	0.59 %	0.69 %	85.89 %
512B	15.66 %	15.06 %	103.92 %
32KiB	35.27 %	35.15 %	100.33 %
2MiB	94.66 %	94.98 %	99.67 %
16MiB	97.53 %	97.25 %	100.29 %
128MiB	97.97 %	97.65 %	100.33 %

< 20%

~ same



Results – ETH vs IB

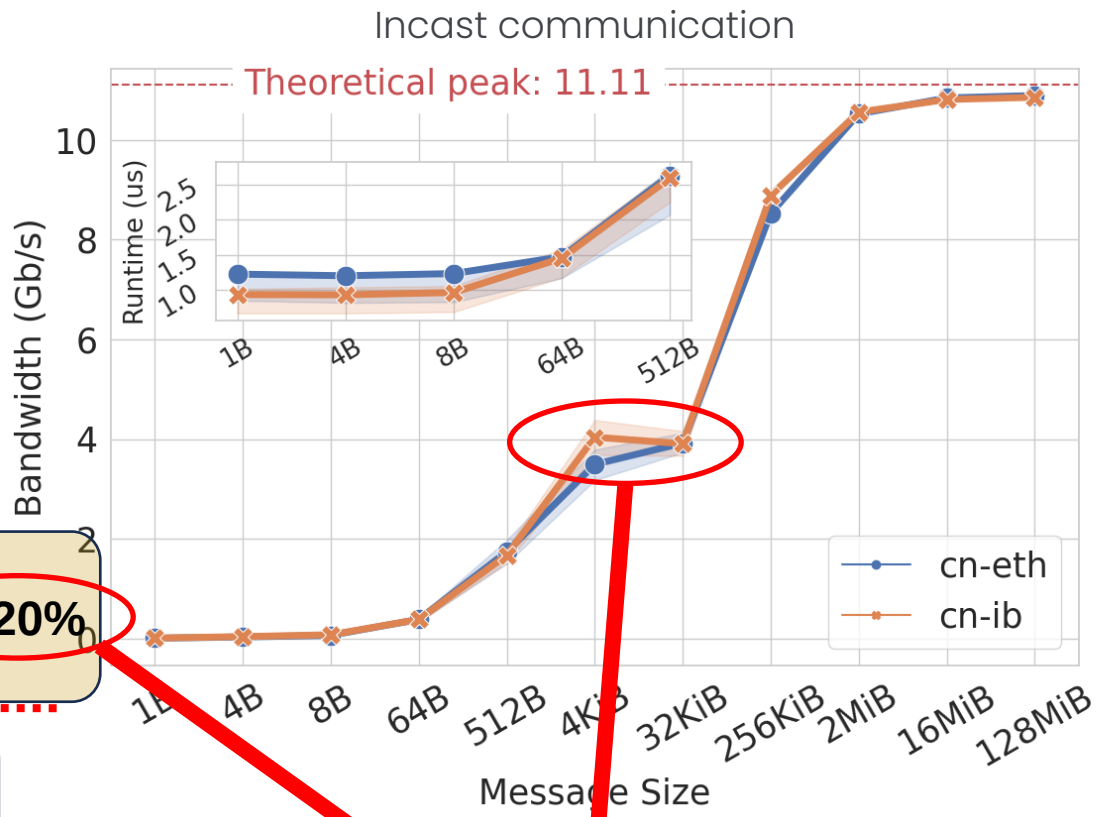
Incast comparison:

- Also Incast exposes the expected results.
- *Wrt* Peer-To-Peer, Incast communication:
  - Performed slightly better.
  - Is more sensitive to algorithm change.

Size	cn-ETH (peak%)	cn-IB (peak%)	ETH/IB ratio
1B	0.07 %	0.09 %	81.20 %
8B	0.59 %	0.69 %	85.89 %
512B	15.66 %	15.06 %	103.92 %
32KiB	35.27 %	35.15 %	100.33 %
2MiB	94.66 %	94.98 %	99.67 %
16MiB	97.53 %	97.25 %	100.29 %
128MiB	97.97 %	97.65 %	100.33 %

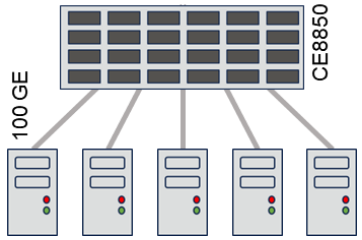
< 20%

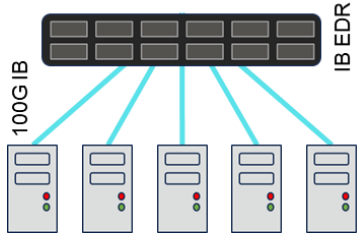
~ same



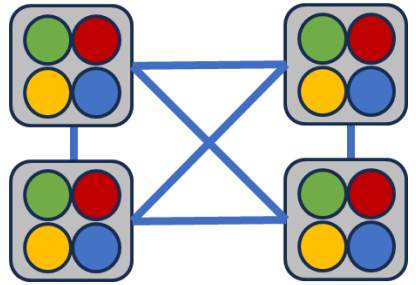
Theoretically hard to predictable results

Experimental results: ETH vs IB





	Peer-To-Peer	Incast	All-To-All
ETH vs IB comparison			
System benchmark			

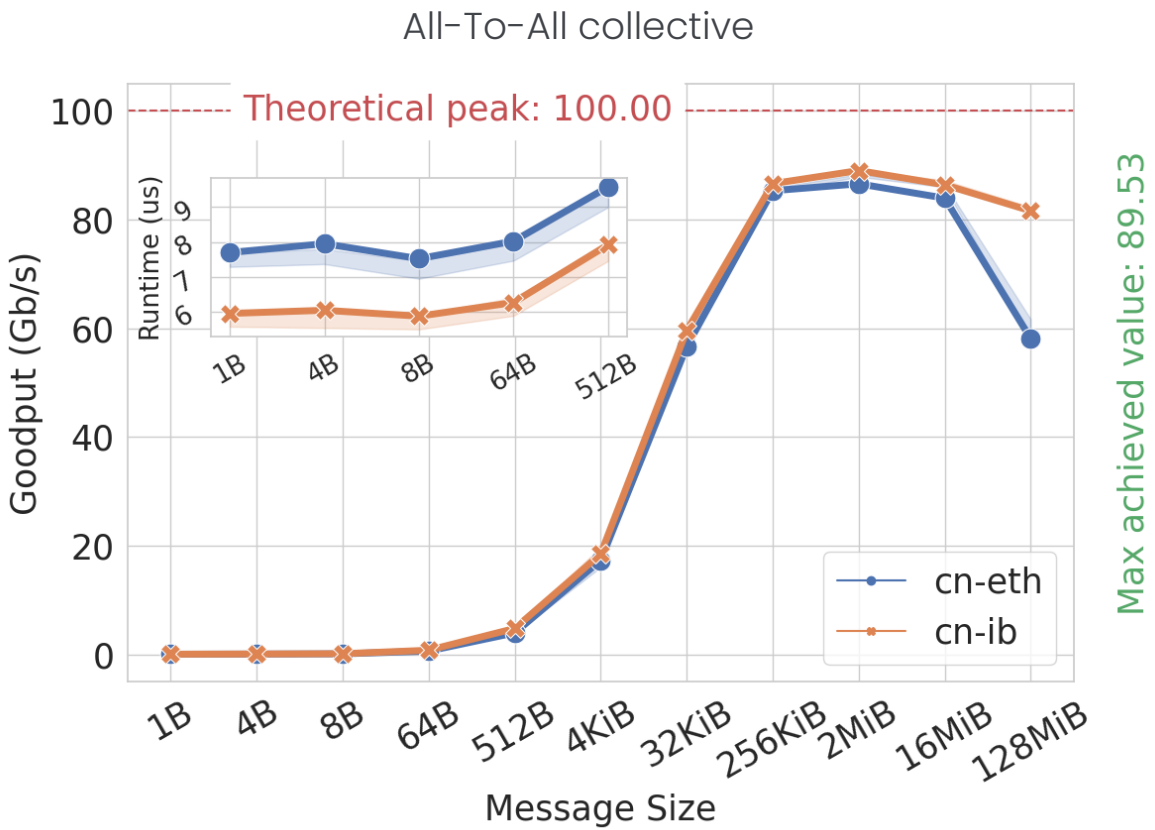


Results – ETH vs IB

All-To-All comparison:

- The performance drop on 128 MiB buffer size still under investigation.
- On the other large buffer cases the performance gap is tight.

Size	cn-ETH (peak%)	cn-IB (peak%)	ETH/IB ratio
1B	0.0095 %	0.012 %	77.3 %
8B	0.08 %	0.1 %	79.1 %
512B	3.9 %	4.7 %	82.7 %
32KiB	56.6 %	59.5 %	95.1 %
2MiB	86.6 %	89 %	97.3 %
16MiB	84 %	86.4 %	97.7 %
128MiB	58.1 %	81.6 %	71.2 %

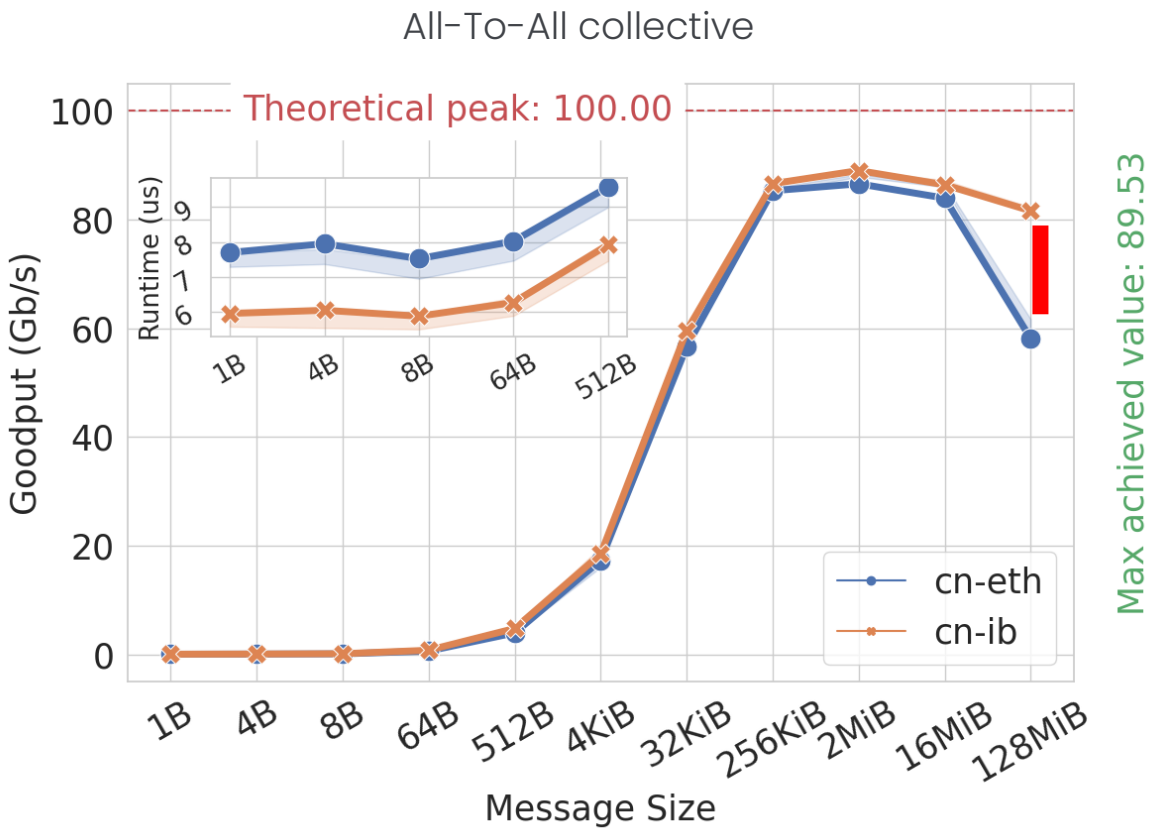


Results – ETH vs IB

All-To-All comparison:

- The performance drop on 128 MiB buffer size still under investigation.
- On the other large buffer cases the performance gap is tight.

Size	cn-ETH (peak%)	cn-IB (peak%)	ETH/IB ratio
1B	0.0095 %	0.012 %	77.3 %
8B	0.08 %	0.1 %	79.1 %
512B	3.9 %	4.7 %	82.7 %
32KiB	56.6 %	59.5 %	95.1 %
2MiB	86.6 %	89 %	97.3 %
16MiB	84 %	86.4 %	97.7 %
128MiB	58.1 %	81.6 %	71.2 %



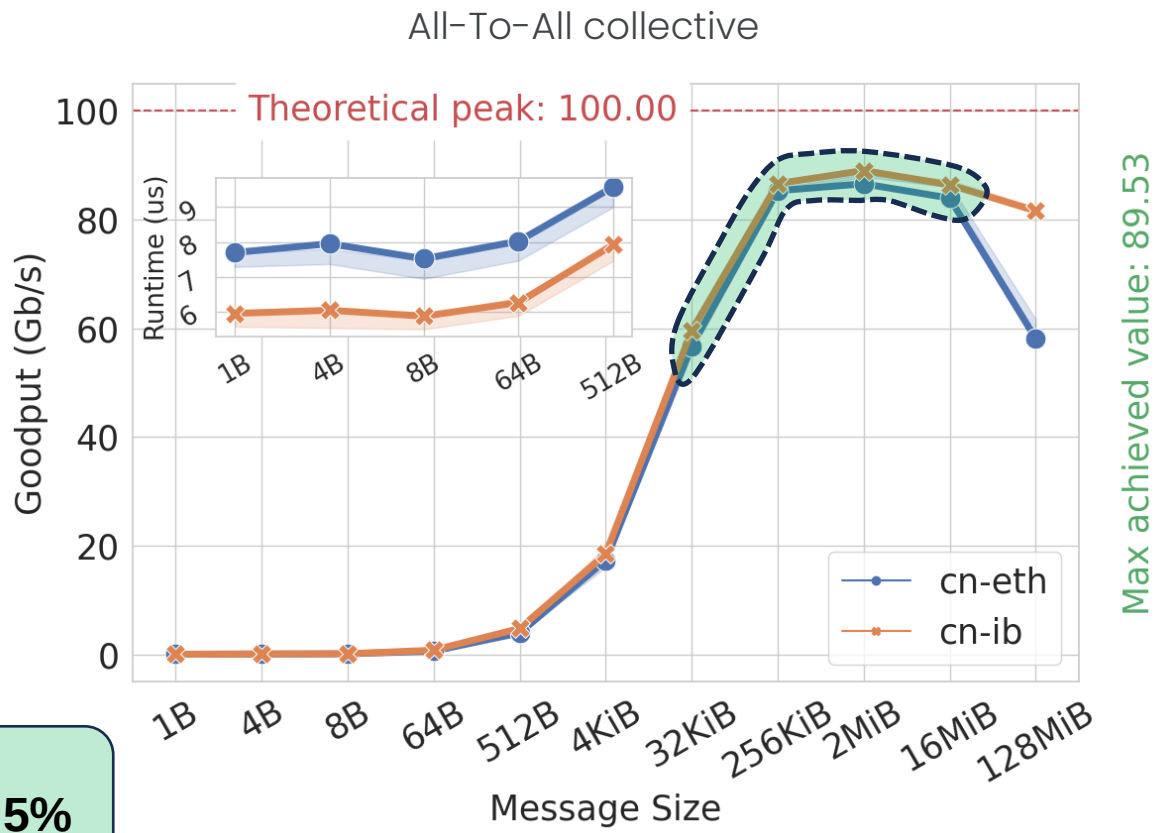
Results – ETH vs IB

All-To-All comparison:

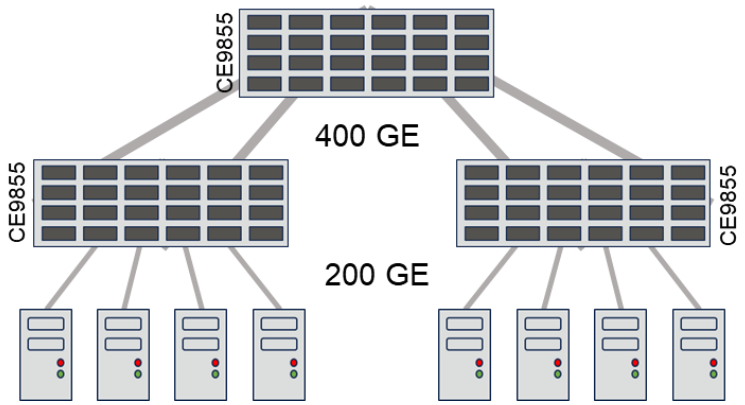
- The performance drop on 128 MiB buffer size still under investigation.
- On the other large buffer cases the performance gap is tight.

Size	cn-ETH (peak%)	cn-IB (peak%)	ETH/IB ratio
1B	0.0095 %	0.012 %	77.3 %
8B	0.08 %	0.1 %	79.1 %
512B	3.9 %	4.7 %	82.7 %
32KiB	56.6 %	59.5 %	95.1 %
2MiB	86.6 %	89 %	97.3 %
16MiB	84 %	86.4 %	97.7 %
128MiB	58.1 %	81.6 %	71.2 %

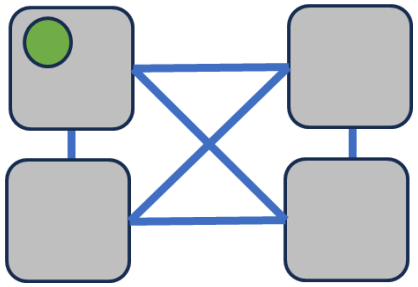
< 5%

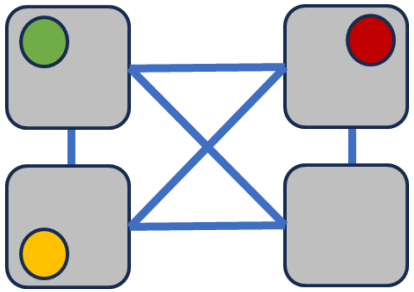


Experimental results: switch crossing



	Peer-To-Peer	Incast	All-To-All
ETH vs IB comparison			
System benchmark			



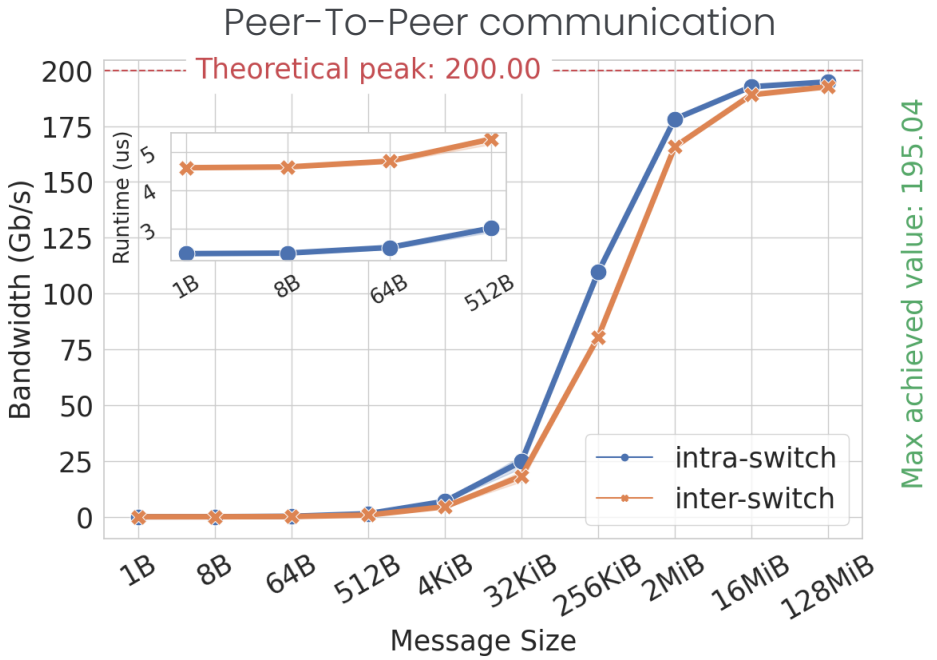


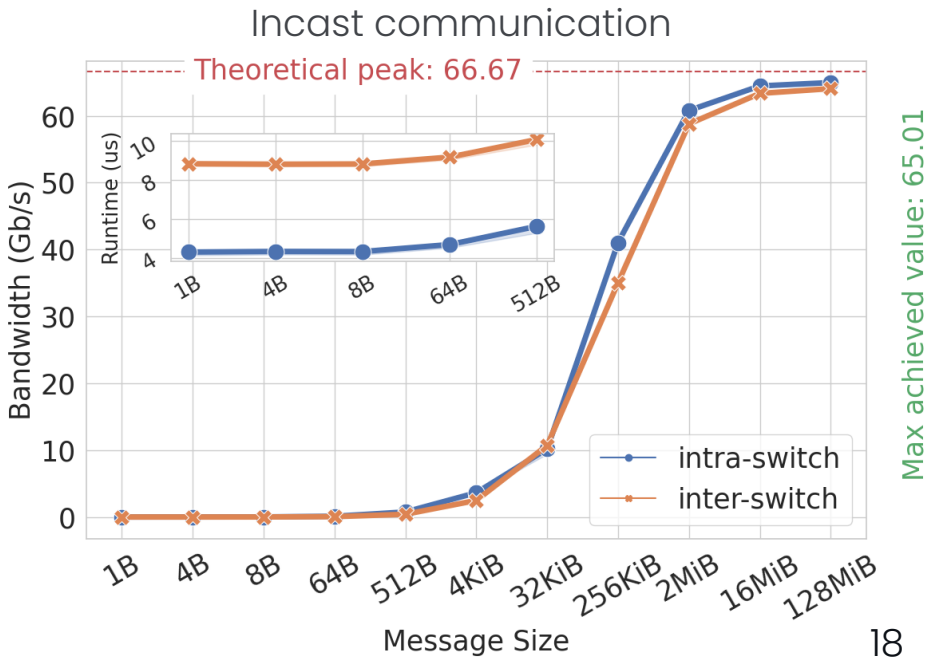
Results – System benchmark

Spine-switch overhead:

- We estimate the latency for crossing one link and one switch as 1.11 µs;
- Buffer larger than 32KiB amortize the spine-switch overhead.

Size	p2p			ic		
	Runtime (µs)		Slow-dw	Runtime (µs)		Slow-dw
	Intra-	Inter-		Intra-	Inter-	
1B	2.35	4.58	1.96 x	1.21	2.33	2.07 x
8B	2.36	4.60	1.95 x	1.22	2.33	2.04 x
512B	3.01	5.33	1.79 x	1.67	2.77	1.83 x
32KiB	10.7	14.7	1.36 x	8.79	8.44	0.95 x
2MiB	94.1	101	1.07 x	209	217	1.03 x
16MiB	696	709	1.02 x	1563	1592	1.01 x
128MiB	550	556	1.01 x	12397	12573	1.01 x



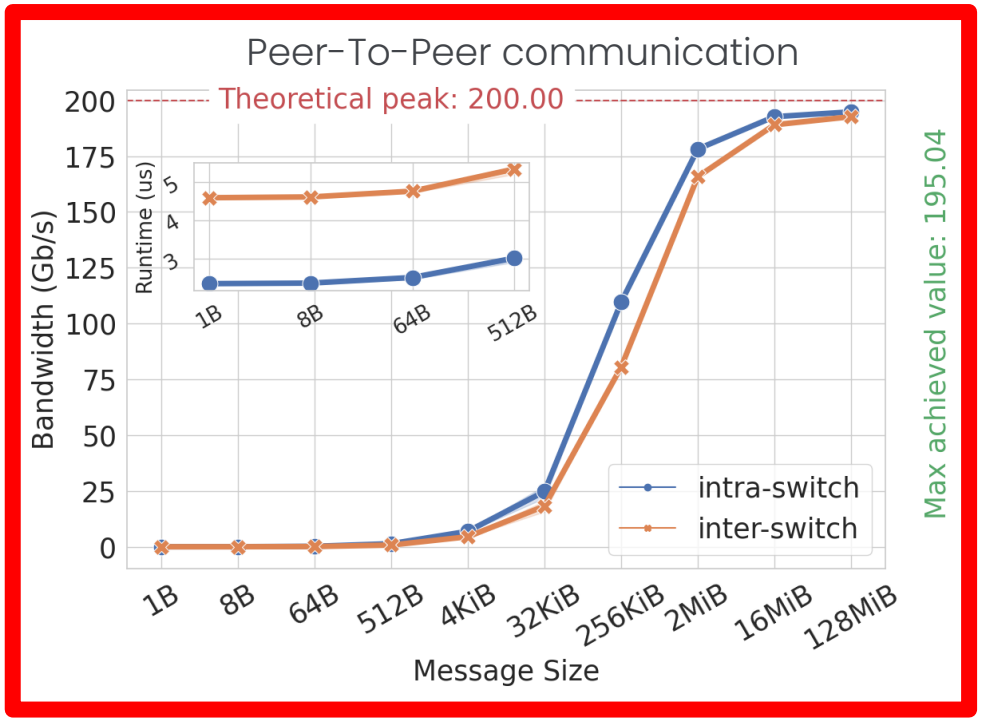


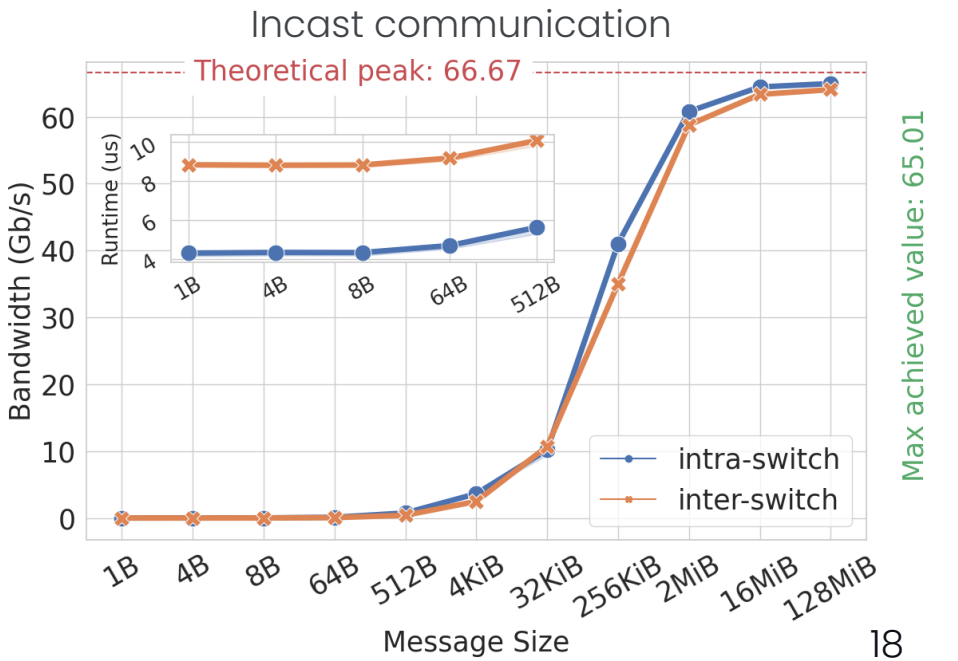
Results – System benchmark

Spine-switch overhead:

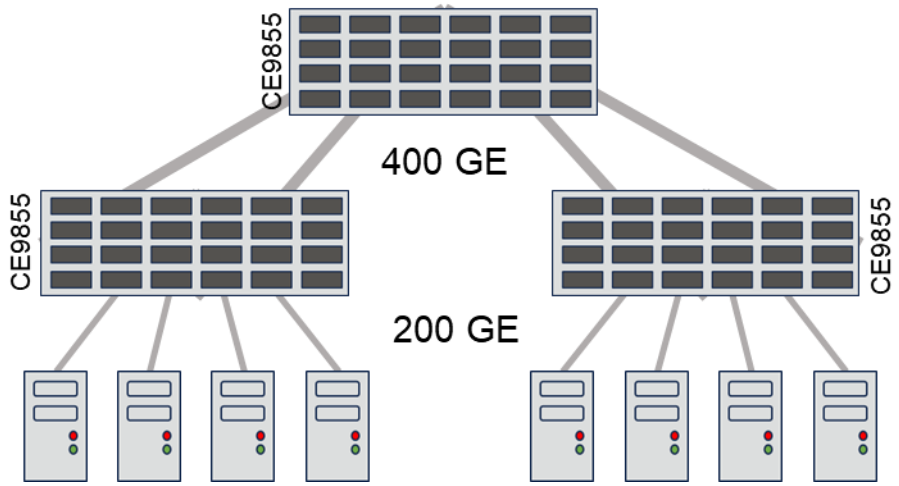
- We estimate the latency for crossing one link and one switch as 1.11 µs;
- Buffer larger than 32KiB amortize the spine-switch overhead.

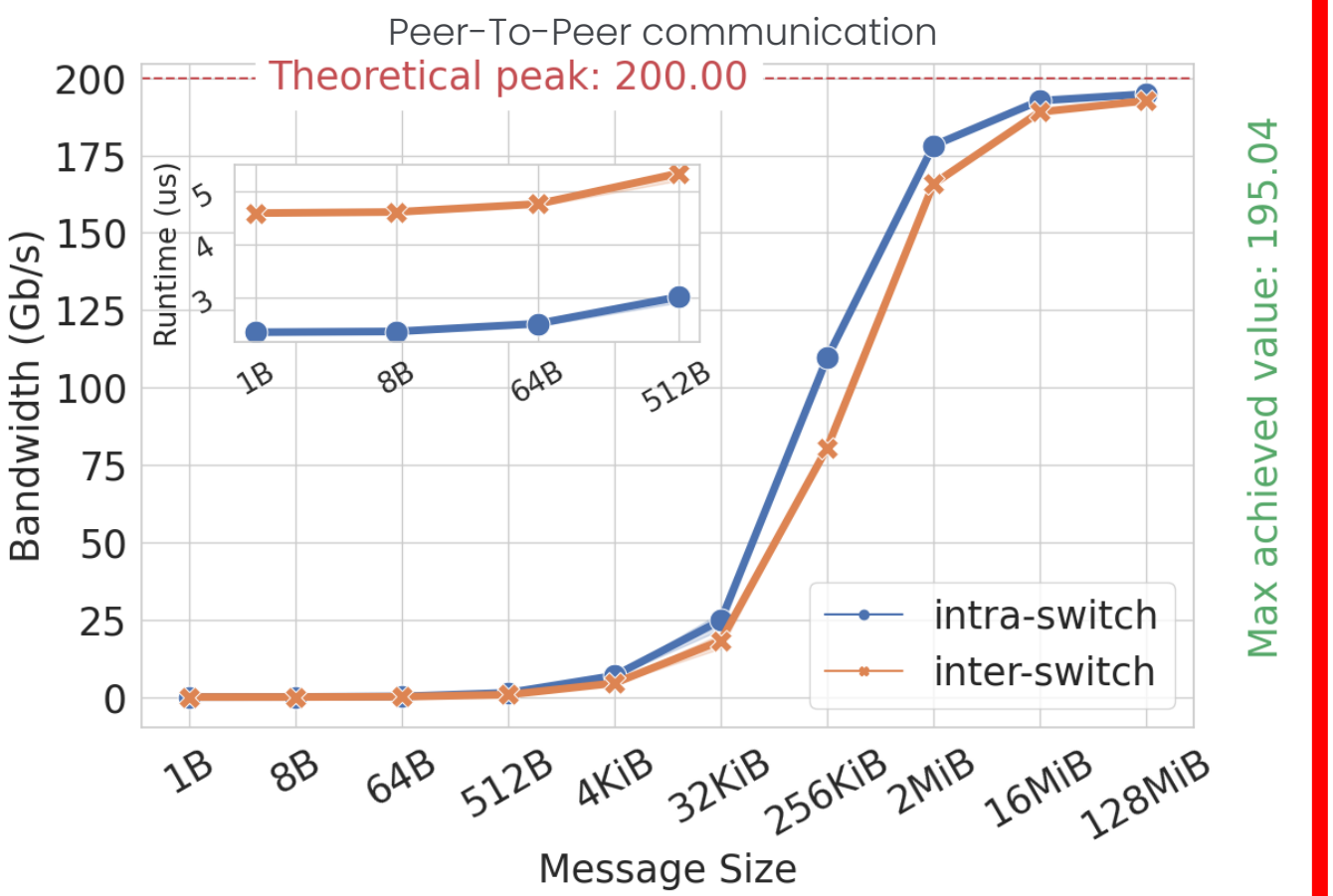
Size	p2p			ic		
	Runtime (µs)		Slow-dw	Runtime (µs)		Slow-dw
	Intra-	Inter-		Intra-	Inter-	
1B	2.35	4.58	1.96 x	1.21	2.33	2.07 x
8B	2.36	4.60	1.95 x	1.22	2.33	2.04 x
512B	3.01	5.33	1.79 x	1.67	2.77	1.83 x
32KiB	10.7	14.7	1.36 x	8.79	8.44	0.95 x
2MiB	94.1	101	1.07 x	209	217	1.03 x
16MiB	696	709	1.02 x	1563	1592	1.01 x
128MiB	550	556	1.01 x	12397	12573	1.01 x



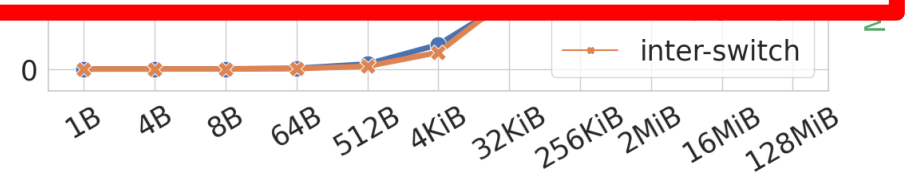


Results – System benchmark

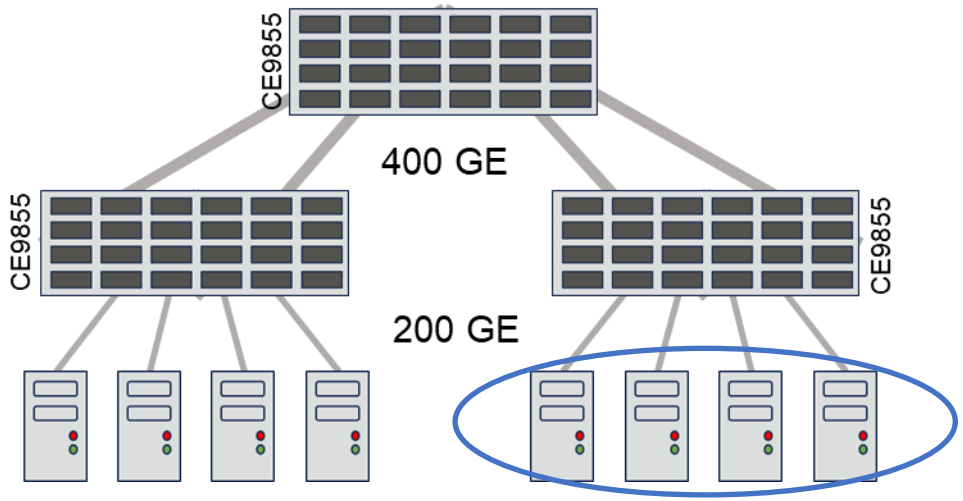


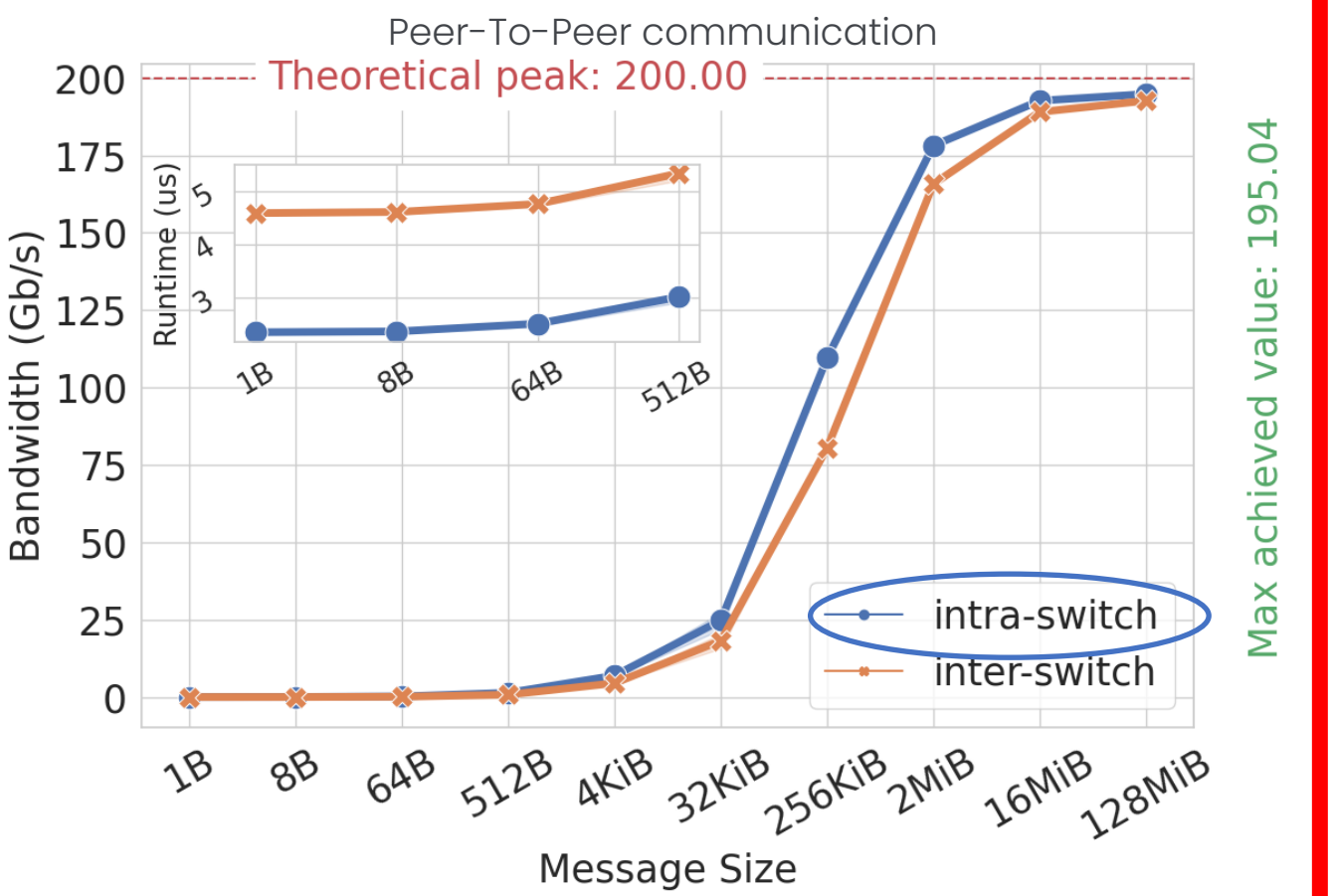


128MiB	550	556	1.01 x	12397	12573	1.01 x



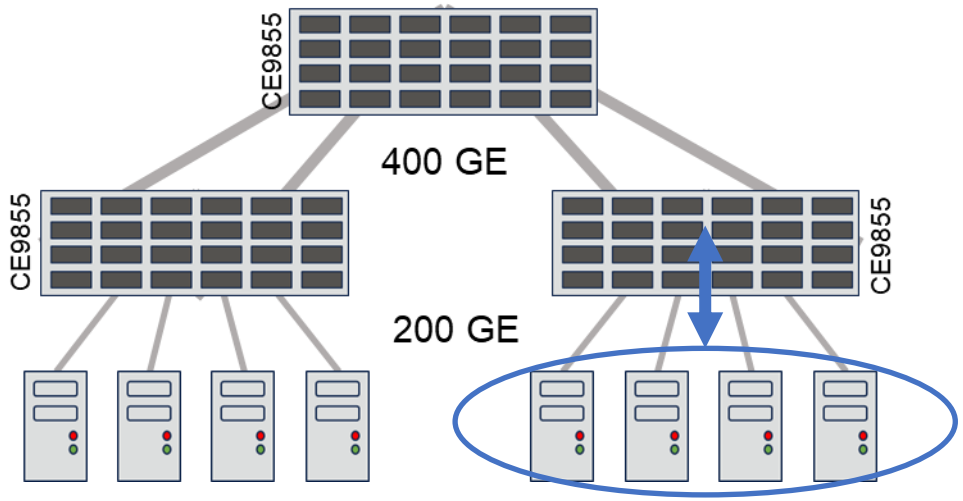
Results – System benchmark

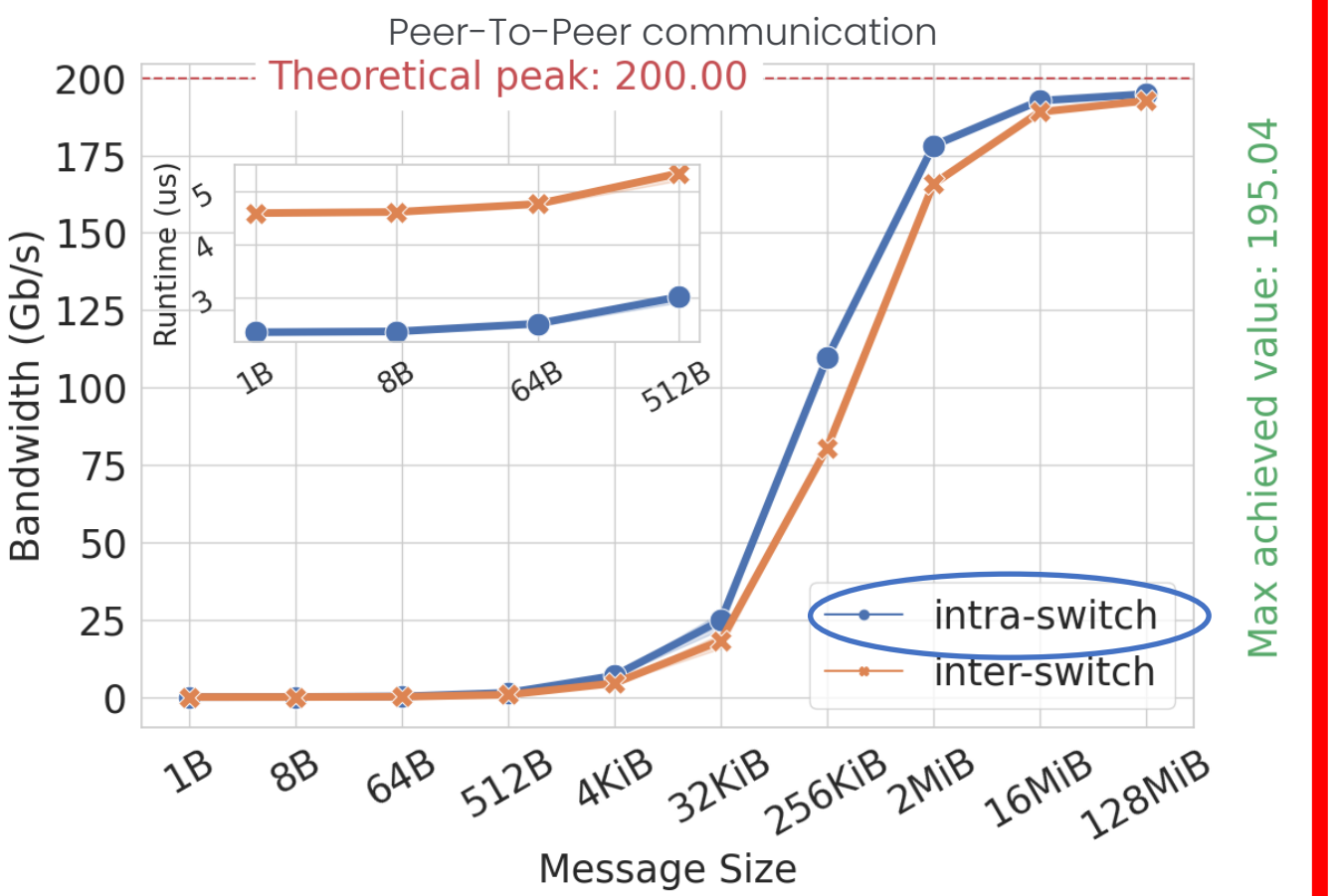




128MiB	550	556	1.01 x	12397	12573	1.01 x

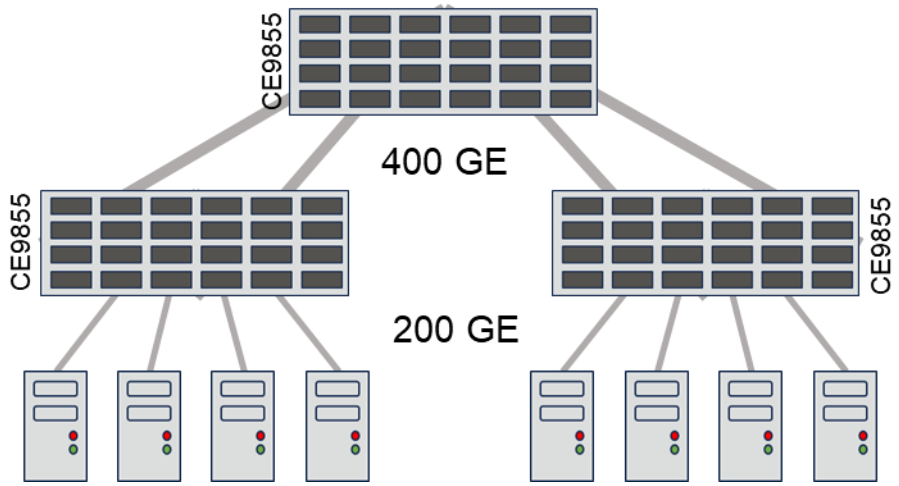
Results – System benchmark

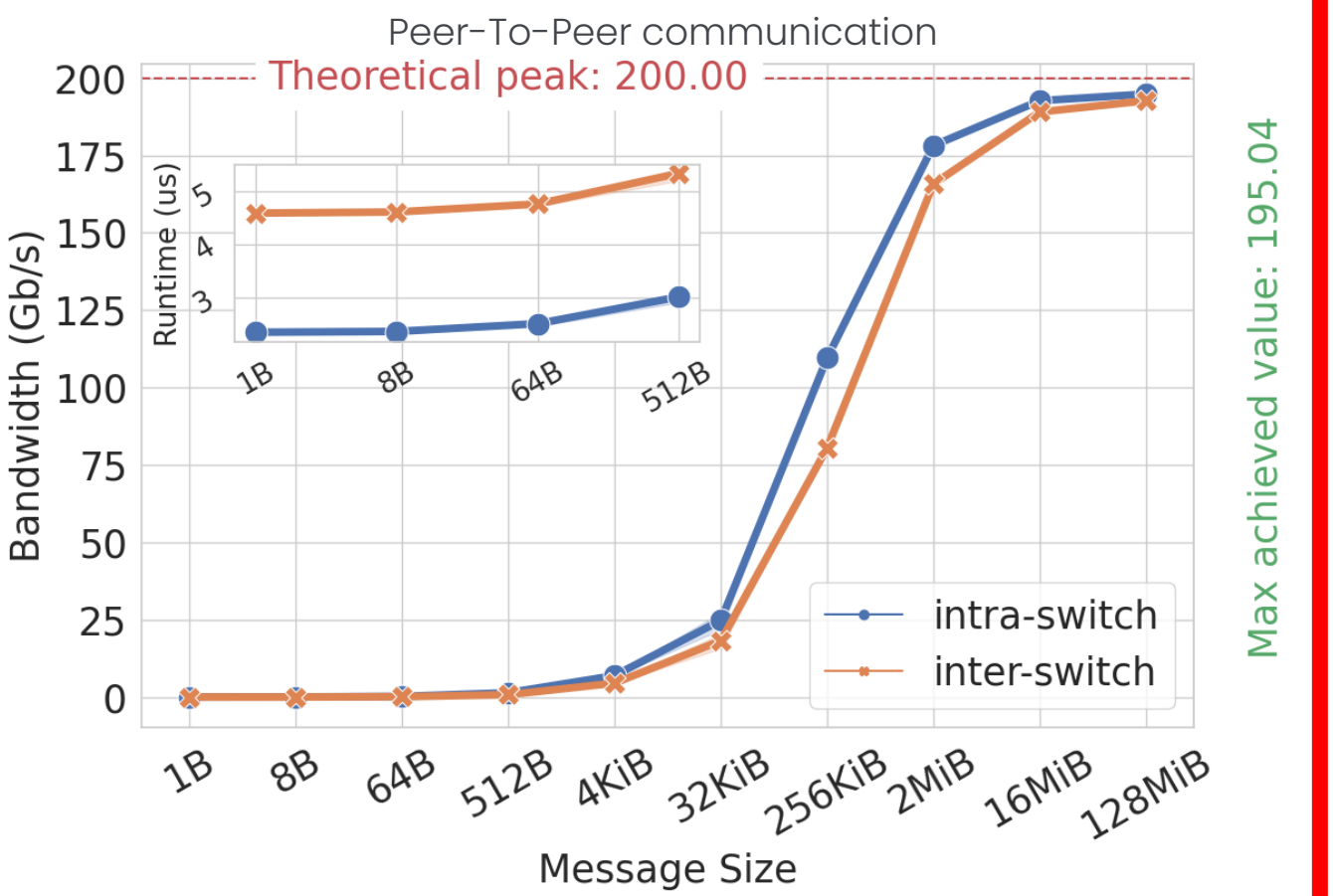




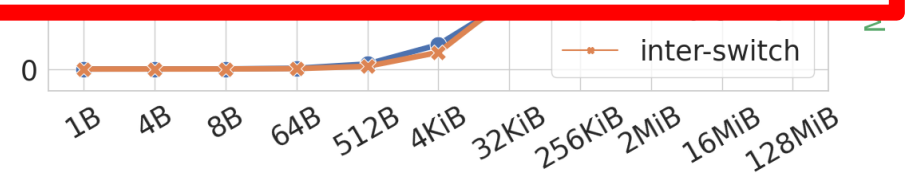
128MiB	550	556	1.01 x	12397	12573	1.01 x

Results – System benchmark

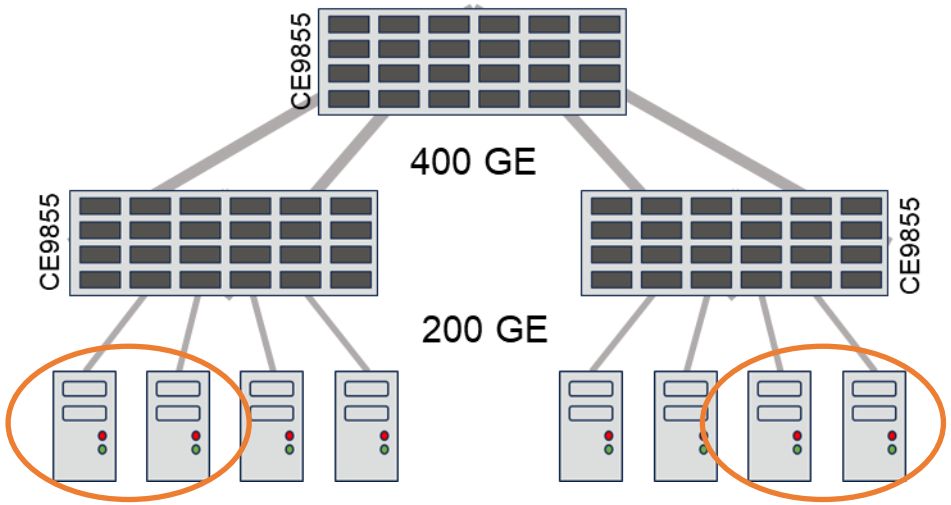


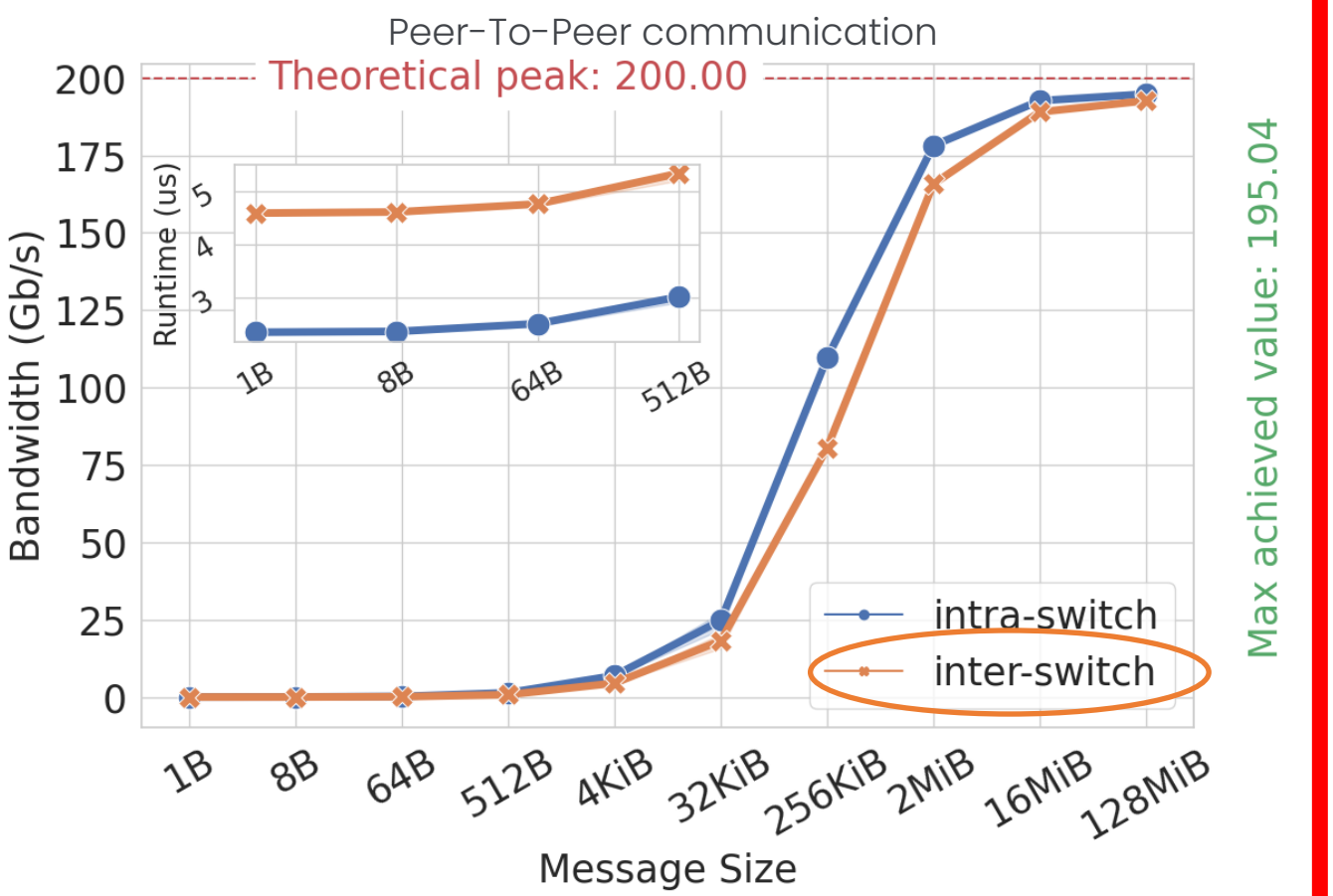


128MiB	550	556	1.01 x	12397	12573	1.01 x



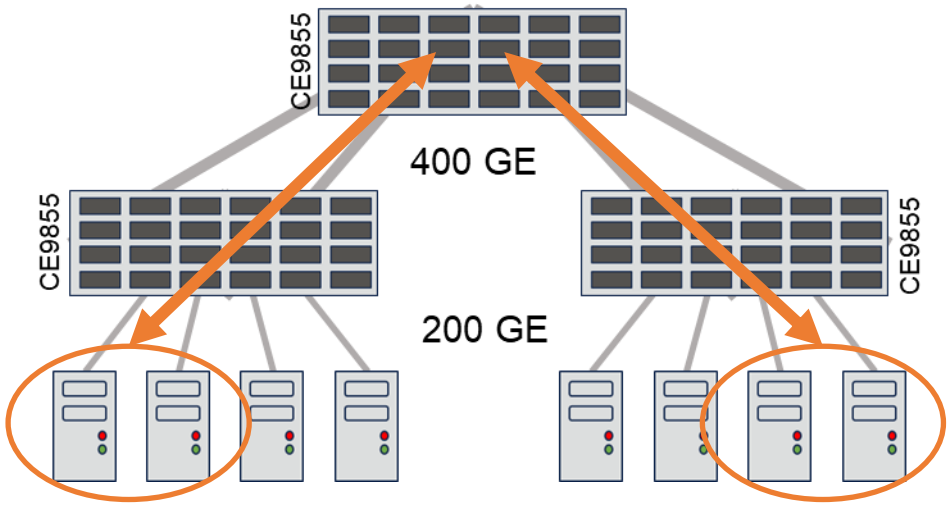
Results – System benchmark

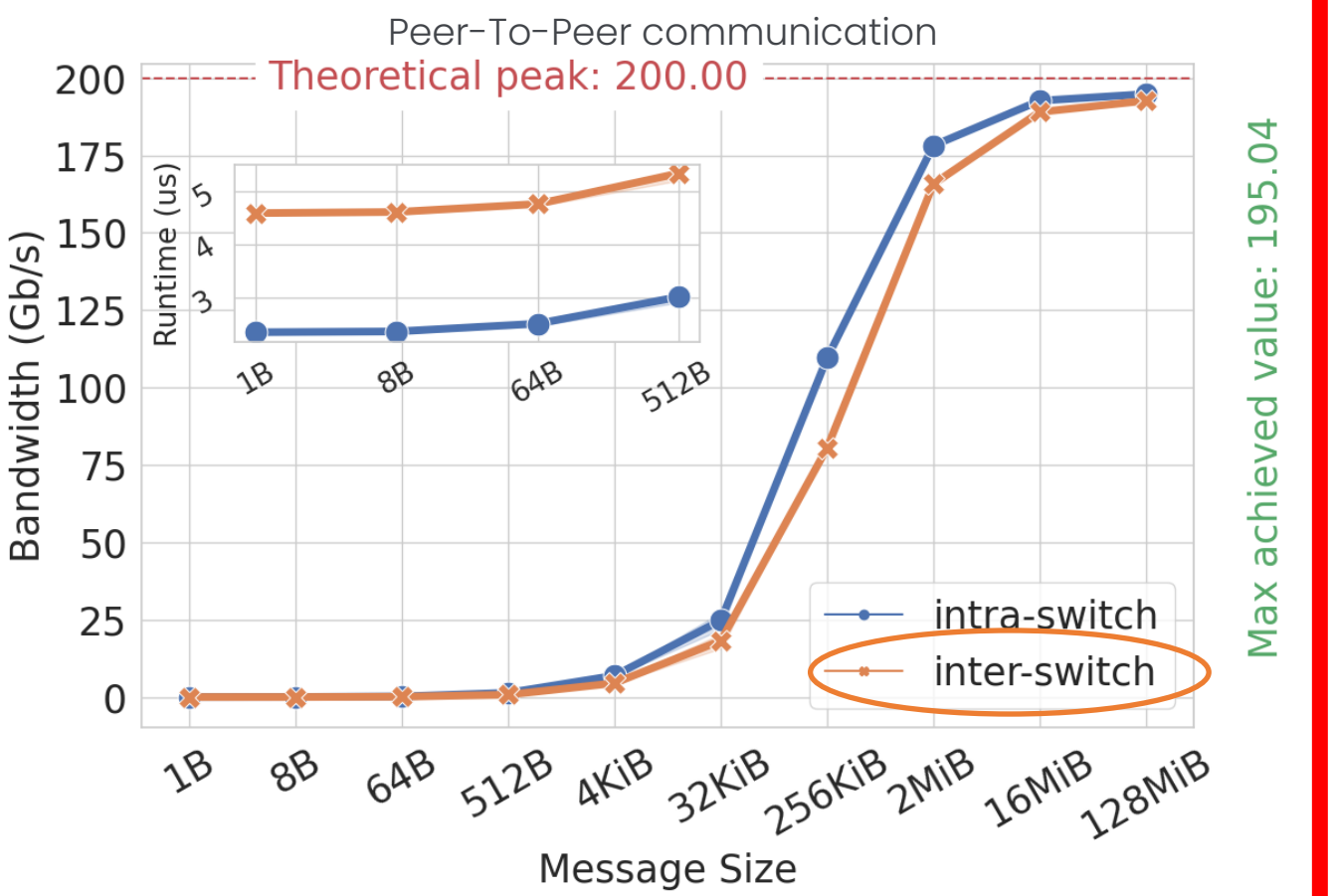




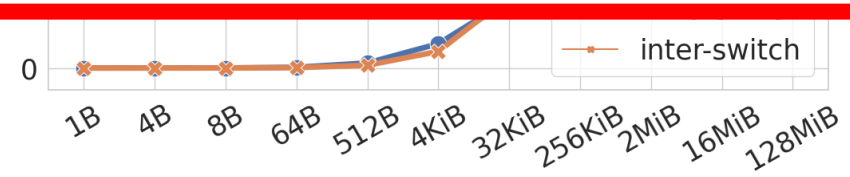
128MiB	550	556	1.01 x	12397	12573	1.01 x

Results – System benchmark





128MiB	550	556	1.01 x	12397	12573	1.01 x

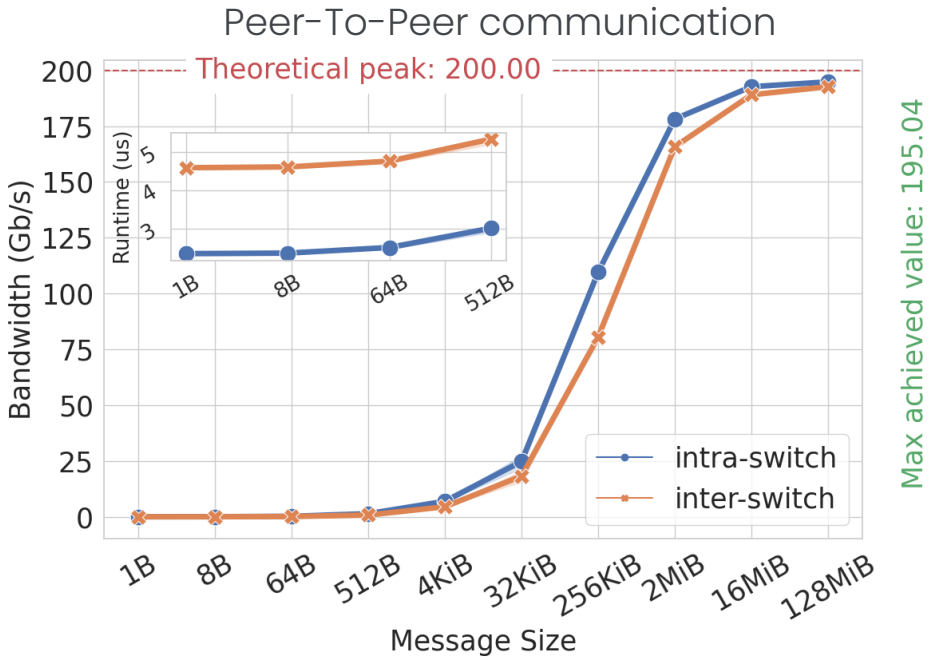


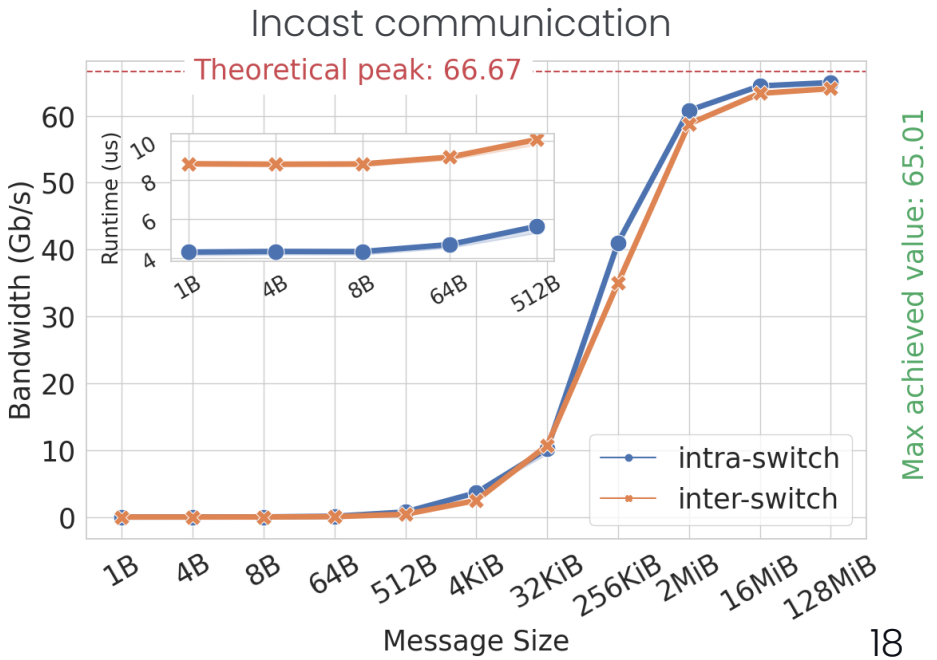
Results – System benchmark

Spine-switch overhead:

- We estimate the latency for crossing one link and one switch as 1.11 µs;
- Buffer larger than 32KiB amortize the spine-switch overhead.

Size	p2p			ic		
	Runtime (µs)		Slow-dw	Runtime (µs)		Slow-dw
	Intra-	Inter-		Intra-	Inter-	
1B	2.35	4.58	1.96 x	1.21	2.33	2.07 x
8B	2.36	4.60	1.95 x	1.22	2.33	2.04 x
512B	3.01	5.33	1.79 x	1.67	2.77	1.83 x
32KiB	10.7	14.7	1.36 x	8.79	8.44	0.95 x
2MiB	94.1	101	1.07 x	209	217	1.03 x
16MiB	696	709	1.02 x	1563	1592	1.01 x
128MiB	550	556	1.01 x	12397	12573	1.01 x



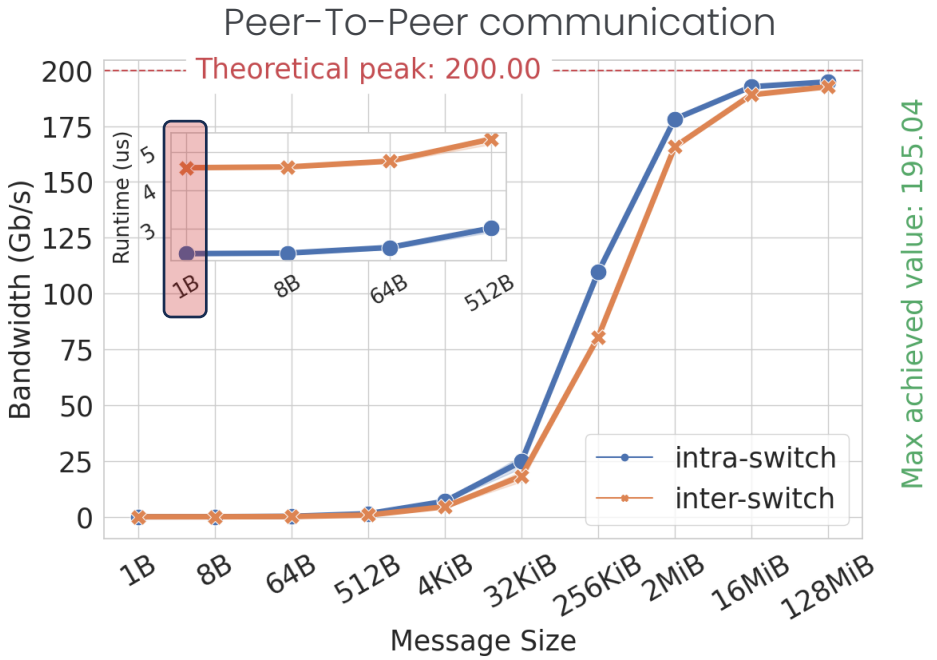


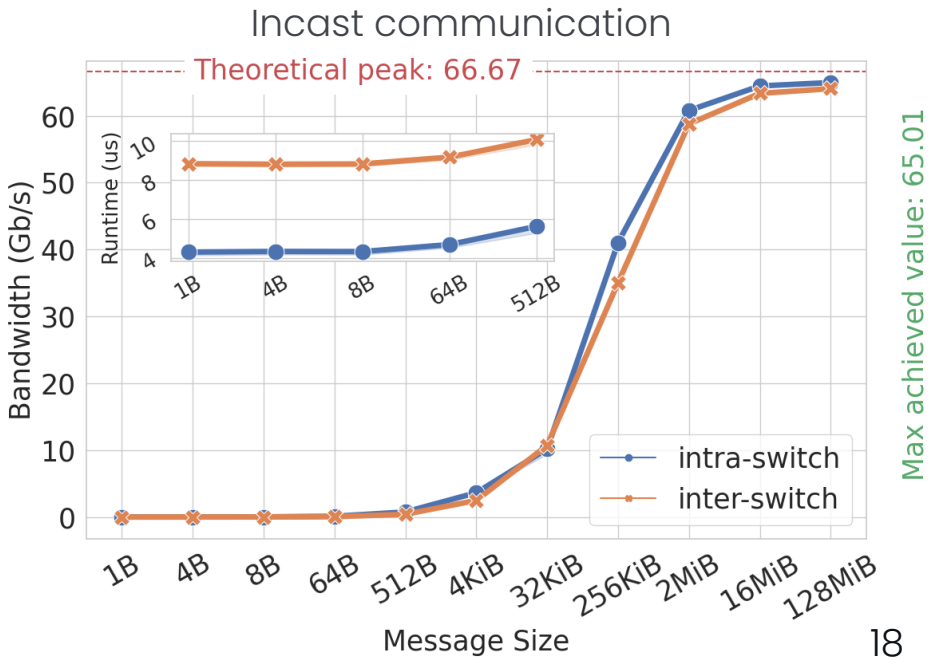
Results – System benchmark

Spine-switch overhead:

- We estimate the latency for crossing one link and one switch as 1.11 µs;
- Buffer larger than 32KiB amortize the spine-switch overhead.

Size	p2p			ic		
	Runtime (µs)		Slow-dw	Runtime (µs)		Slow-dw
	Intra-	Inter-		Intra-	Inter-	
1B	2.35	4.58	1.96 x	1.21	2.33	2.07 x
8B	2.36	4.60	1.95 x	1.22	2.33	2.04 x
512B	3.01	5.33	1.79 x	1.67	2.77	1.83 x
32KiB	10.7	14.7	1.36 x	8.79	8.44	0.95 x
2MiB	94.1	101	1.07 x	209	217	1.03 x
16MiB	696	709	1.02 x	1563	1592	1.01 x
128MiB	550	556	1.01 x	12397	12573	1.01 x



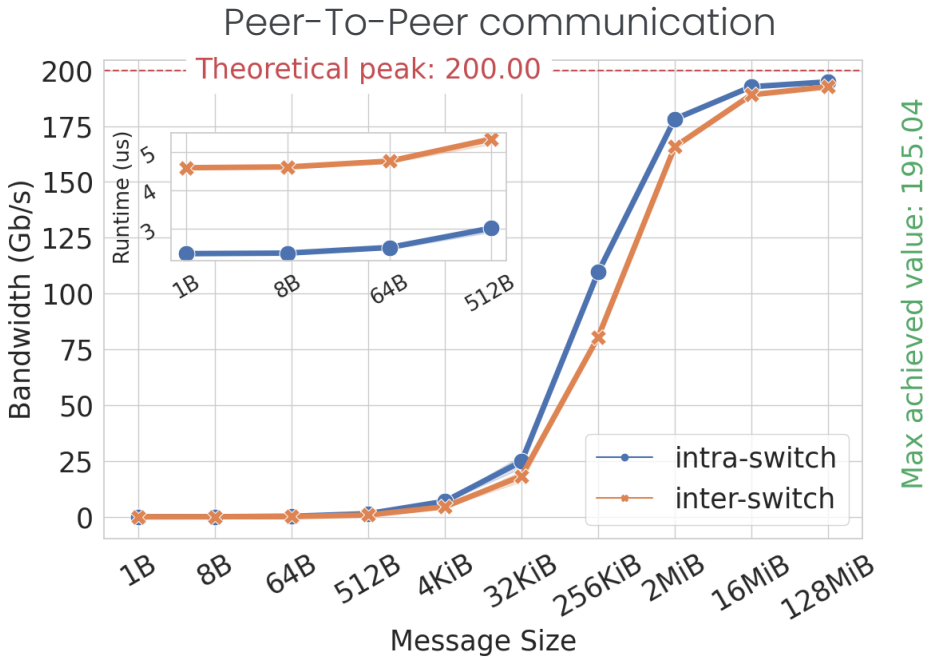


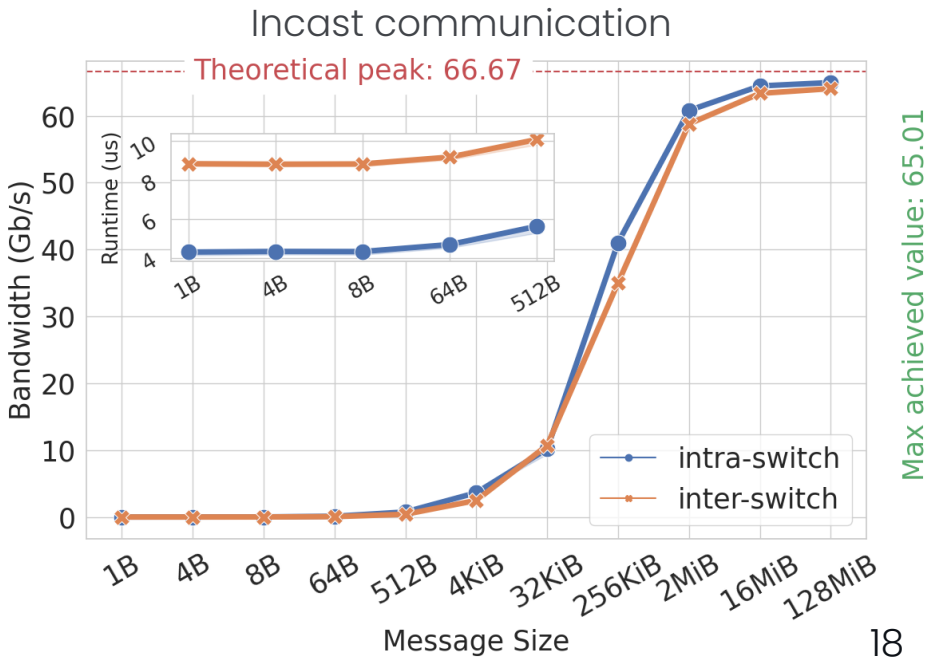
Results – System benchmark

Spine-switch overhead:

- We estimate the latency for crossing one link and one switch as 1.11 µs;
- Buffer larger than 32KiB amortize the spine-switch overhead.

Size	p2p			ic		
	Runtime (µs)		Slow-dw	Runtime (µs)		Slow-dw
	Intra-	Inter-		Intra-	Inter-	
1B	2.35	4.58	1.96 x	1.21	2.33	2.07 x
8B	2.36	4.60	1.95 x	1.22	2.33	2.04 x
512B	3.01	5.33	1.79 x	1.67	2.77	1.83 x
32KiB	10.7	14.7	1.36 x	8.79	8.44	0.95 x
2MiB	94.1	101	1.07 x	209	217	1.03 x
16MiB	696	709	1.02 x	1563	1592	1.01 x
128MiB	550	556	1.01 x	12397	12573	1.01 x



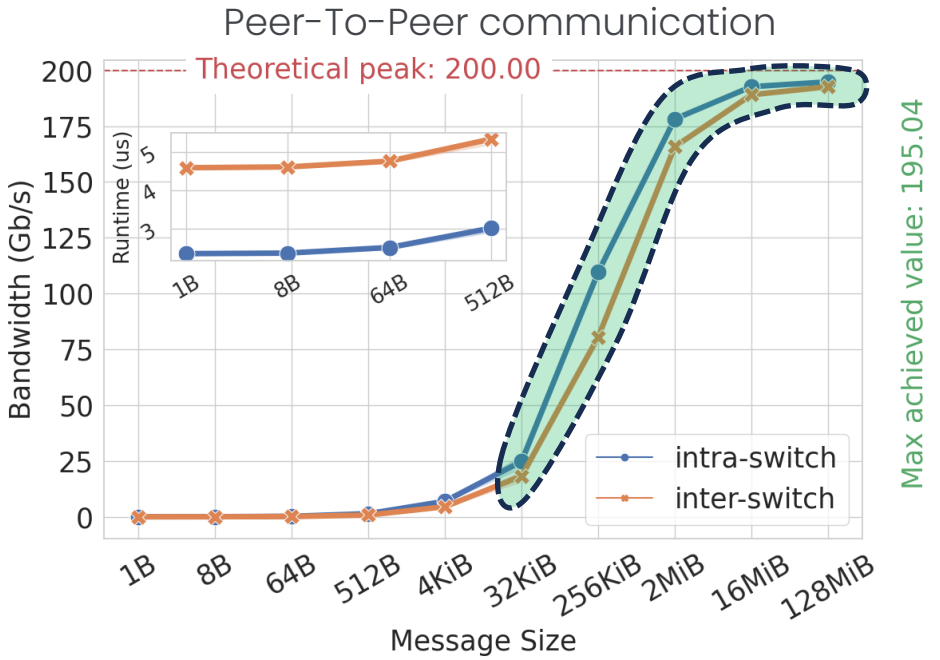


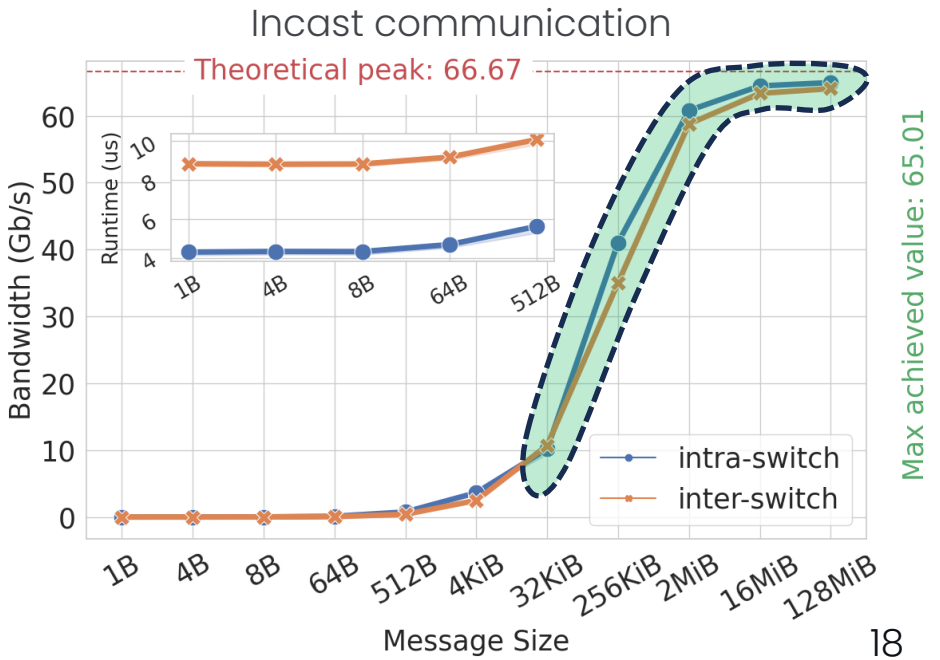
Results – System benchmark

Spine-switch overhead:

- We estimate the latency for crossing one link and one switch as 1.11 µs;
- Buffer larger than 32KiB amortize the spine-switch overhead.

Size	p2p			ic		
	Runtime (µs)		Slow-dw	Runtime (µs)		Slow-dw
	Intra-	Inter-		Intra-	Inter-	
1B	2.35	4.58	1.96 x	1.21	2.33	2.07 x
8B	2.36	4.60	1.95 x	1.22	2.33	2.04 x
512B	3.01	5.33	1.79 x	1.67	2.77	1.83 x
32KiB	10.7	14.7	1.36 x	8.79	8.44	0.95 x
2MiB	94.1	101	1.07 x	209	217	1.03 x
16MiB	696	709	1.02 x	1563	1592	1.01 x
128MiB	550	556	1.01 x	12397	12573	1.01 x



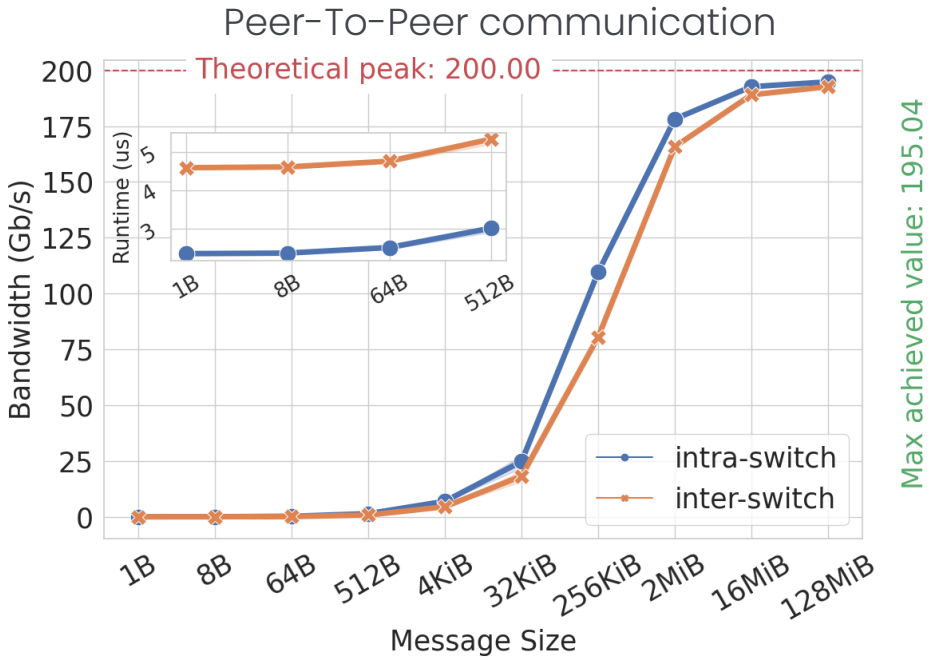


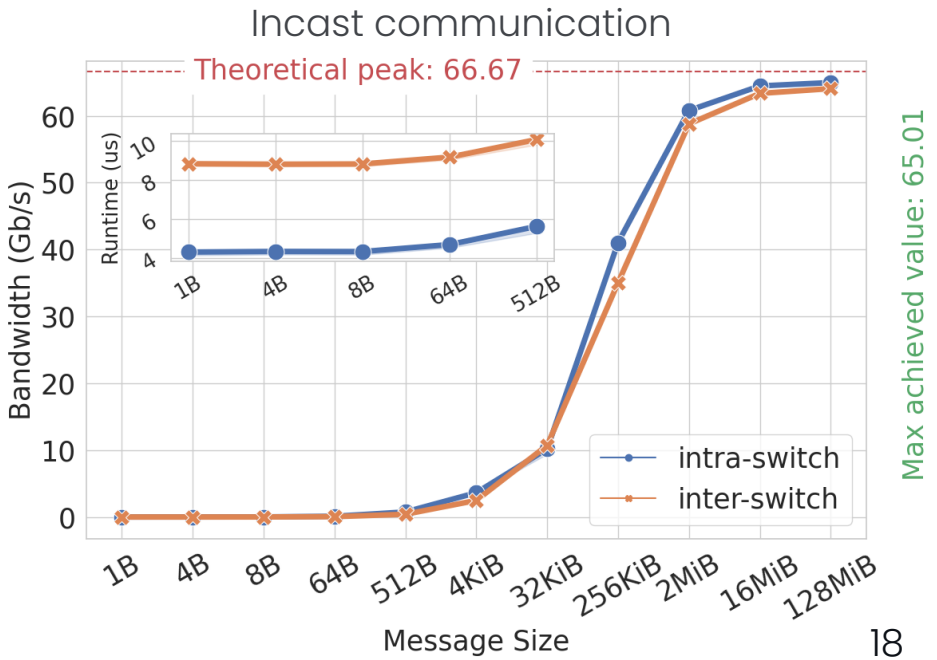
Results – System benchmark

Spine-switch overhead:

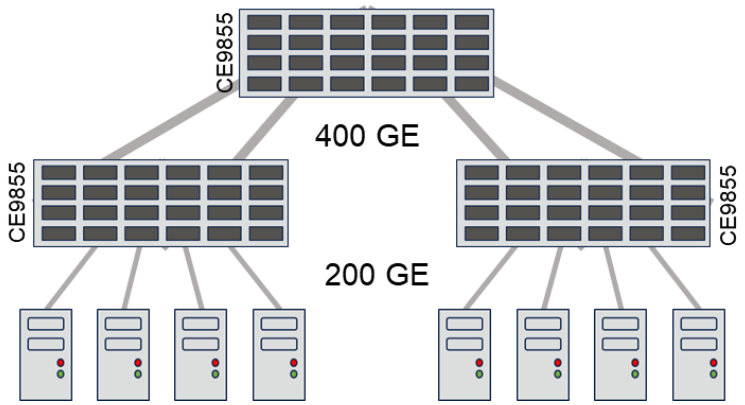
- We estimate the latency for crossing one link and one switch as 1.11 µs;
- Buffer larger than 32KiB amortize the spine-switch overhead.

Size	p2p			ic		
	Runtime (µs)		Slow-dw	Runtime (µs)		Slow-dw
	Intra-	Inter-		Intra-	Inter-	
1B	2.35	4.58	1.96 x	1.21	2.33	2.07 x
8B	2.36	4.60	1.95 x	1.22	2.33	2.04 x
512B	3.01	5.33	1.79 x	1.67	2.77	1.83 x
32KiB	10.7	14.7	1.36 x	8.79	8.44	0.95 x
2MiB	94.1	101	1.07 x	209	217	1.03 x
16MiB	696	709	1.02 x	1563	1592	1.01 x
128MiB	550	556	1.01 x	12397	12573	1.01 x

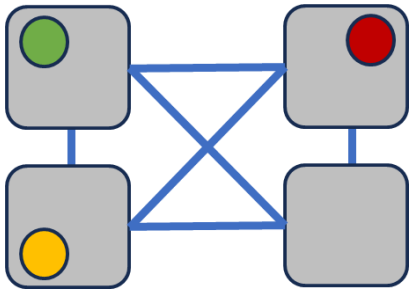


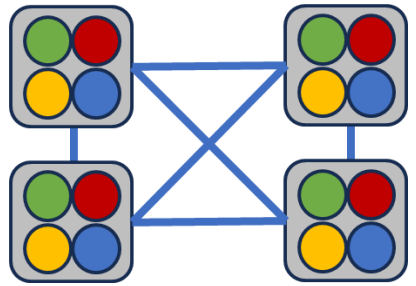


Experimental results: full system



	Peer-To-Peer	Incast	All-To-All
ETH vs IB comparison			
System benchmark			



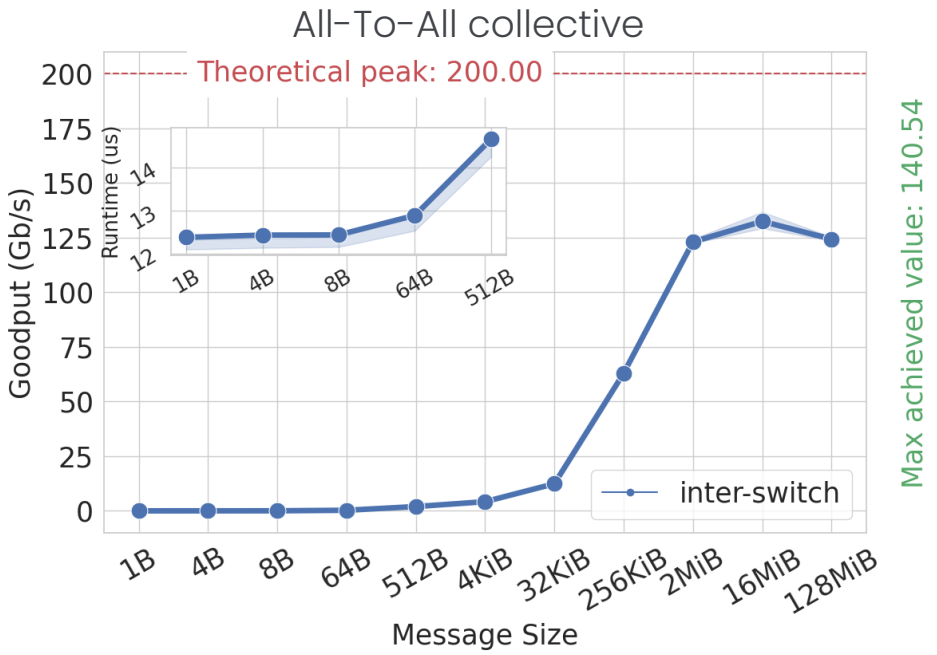


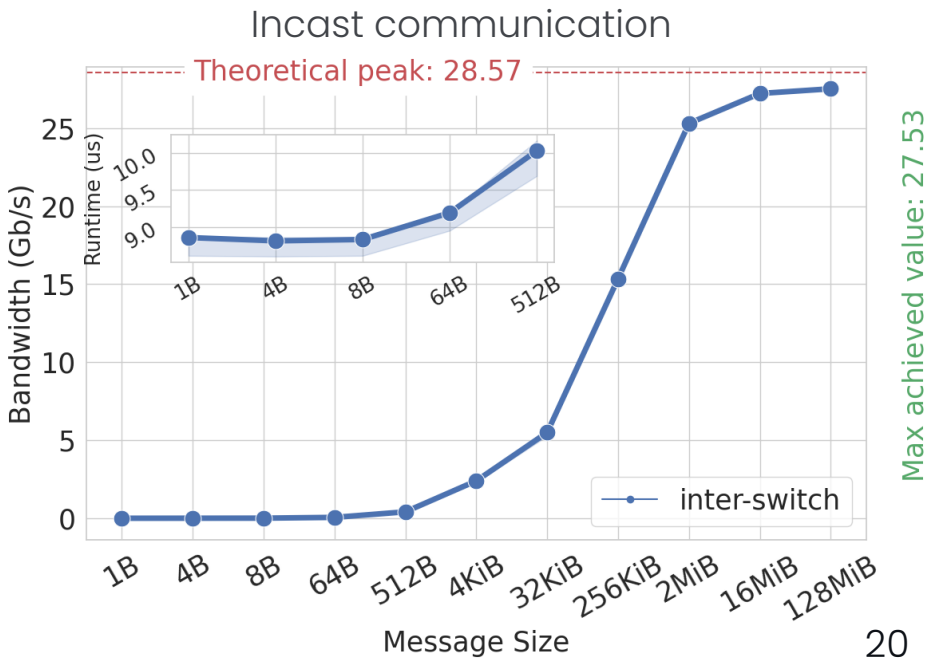
Results – System benchmark

Whole cluster performances:

- All-To-All grows until 66% of the peak.
- Incast saturate the bandwidth also on the whole cluster.

Size	ic		a2a	
	Bandwidth (Gb/s)	Peak %	Goodput (Gb/s)	Peak %
1B	0.0009	0.003 %	0.004	0.002 %
8B	0.007	0.025 %	0.036	0.018 %
512B	0.411	1.44 %	1.96	0.98 %
32KiB	5.49	19.22 %	12.4	6.20 %
2MiB	25.3	88.55 %	123.1	61.55 %
16MiB	27.21	95.24 %	132.46	66.23 %
128MiB	27.51	96.30 %	124.36	62.18 %



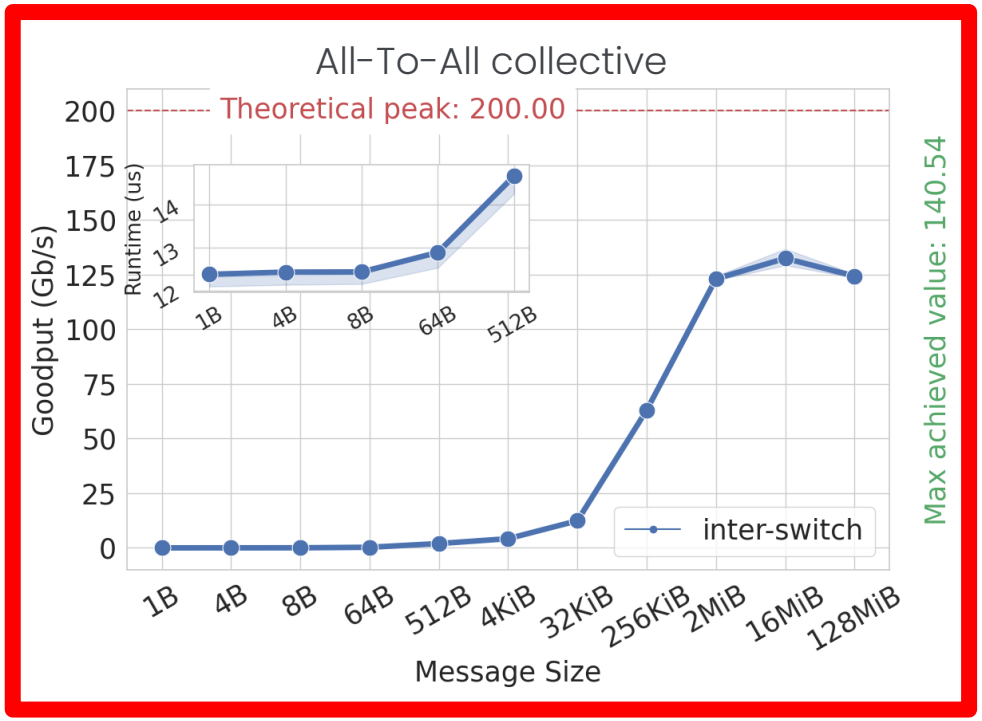


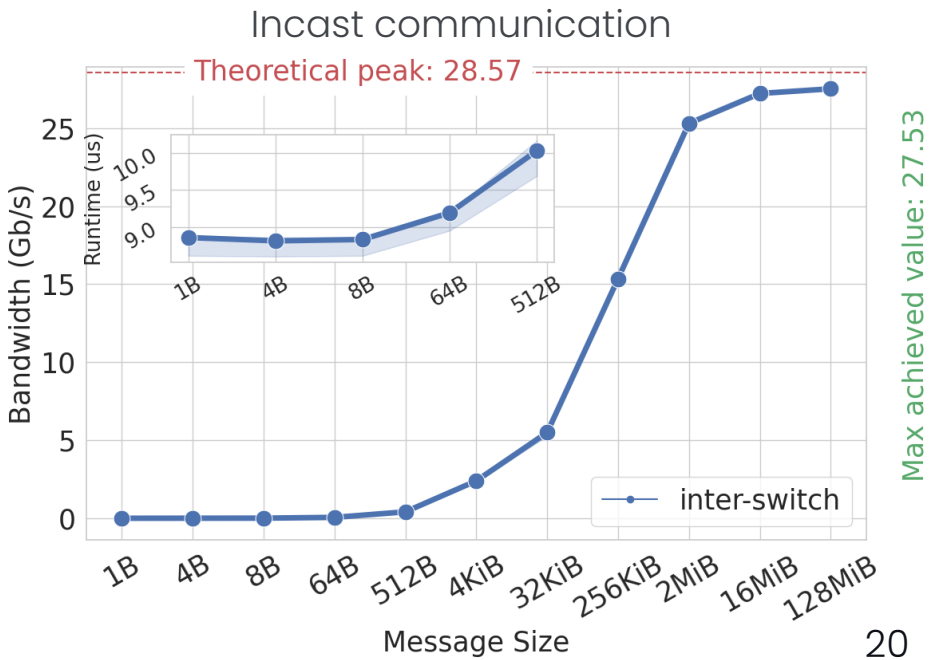
Results – System benchmark

Whole cluster performances:

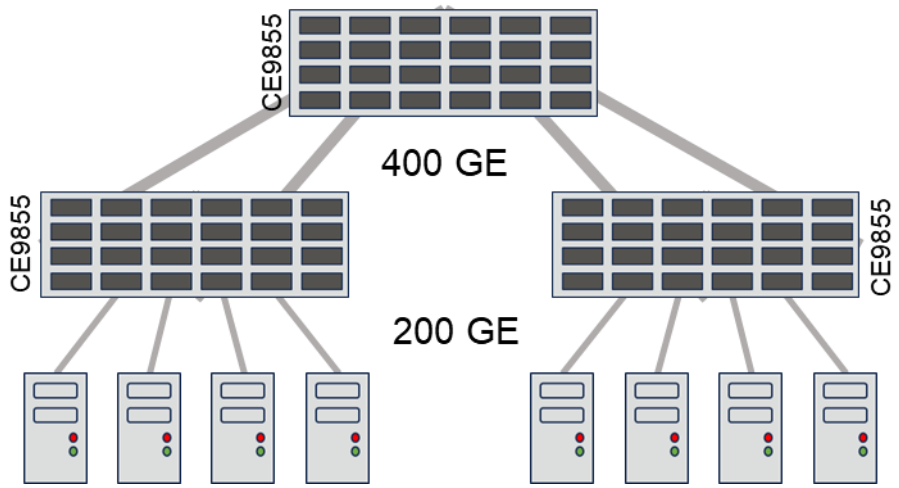
- All-To-All grows until 66% of the peak.
- Incast saturate the bandwidth also on the whole cluster.

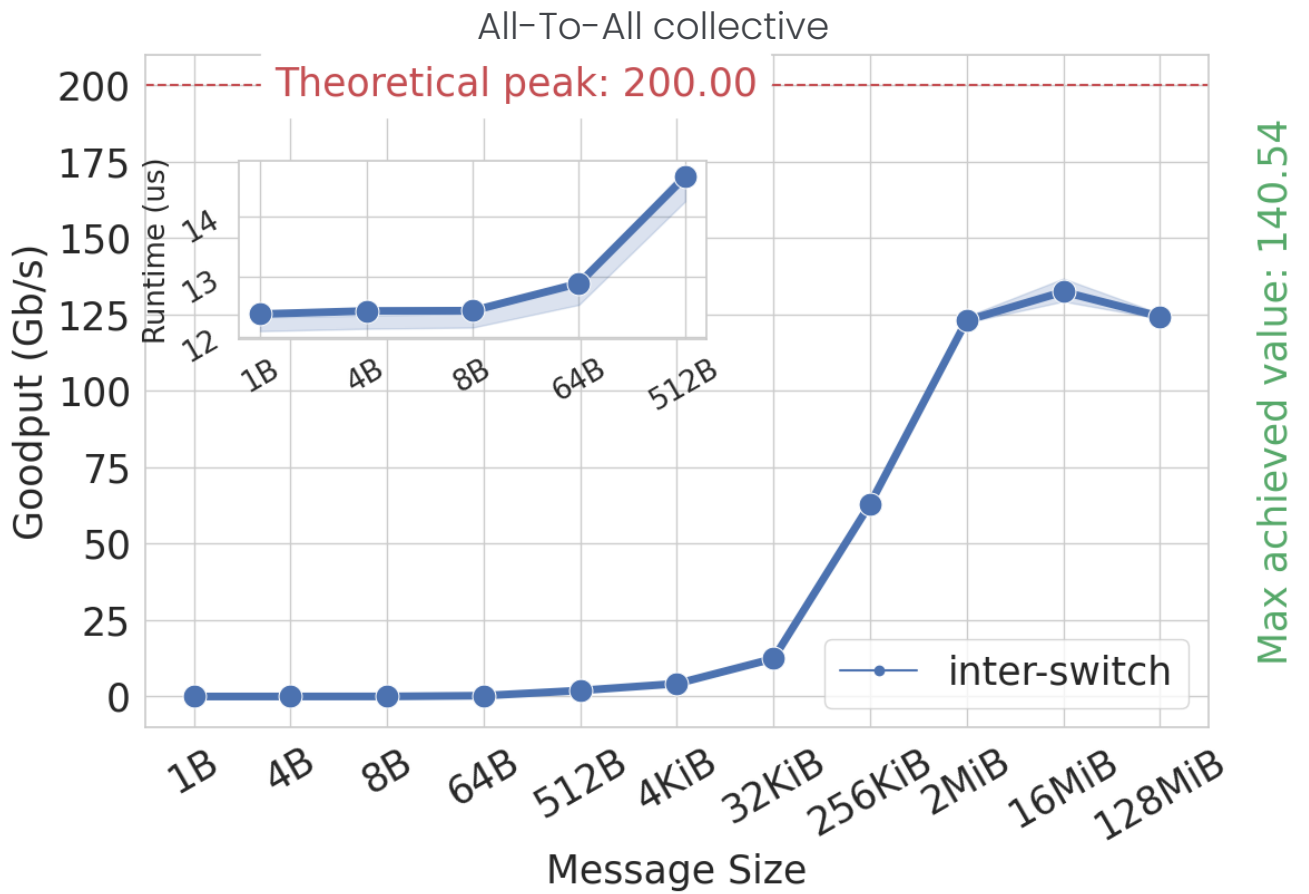
Size	ic		a2a	
	Bandwidth (Gb/s)	Peak %	Goodput (Gb/s)	Peak %
1B	0.0009	0.003 %	0.004	0.002 %
8B	0.007	0.025 %	0.036	0.018 %
512B	0.411	1.44 %	1.96	0.98 %
32KiB	5.49	19.22 %	12.4	6.20 %
2MiB	25.3	88.55 %	123.1	61.55 %
16MiB	27.21	95.24 %	132.46	66.23 %
128MiB	27.51	96.30 %	124.36	62.18 %





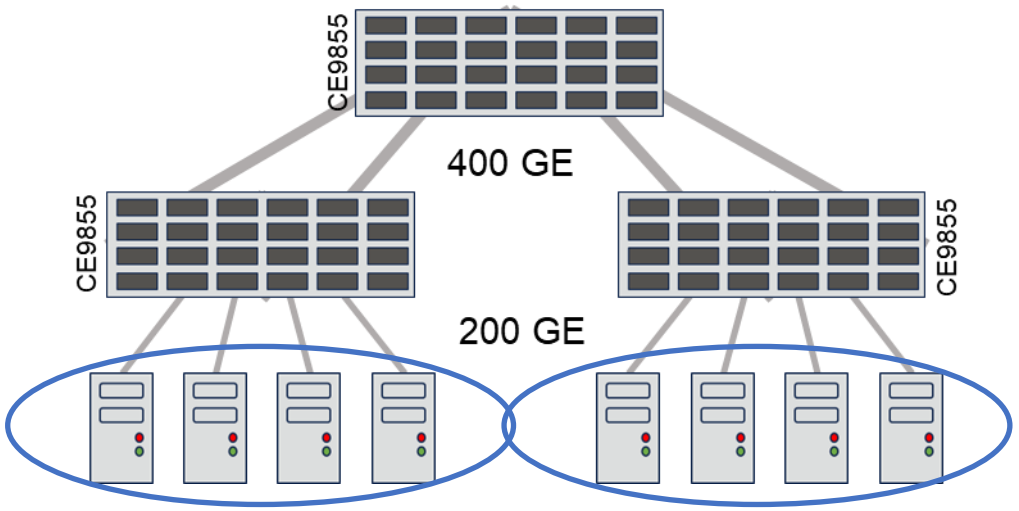
Results – System benchmark

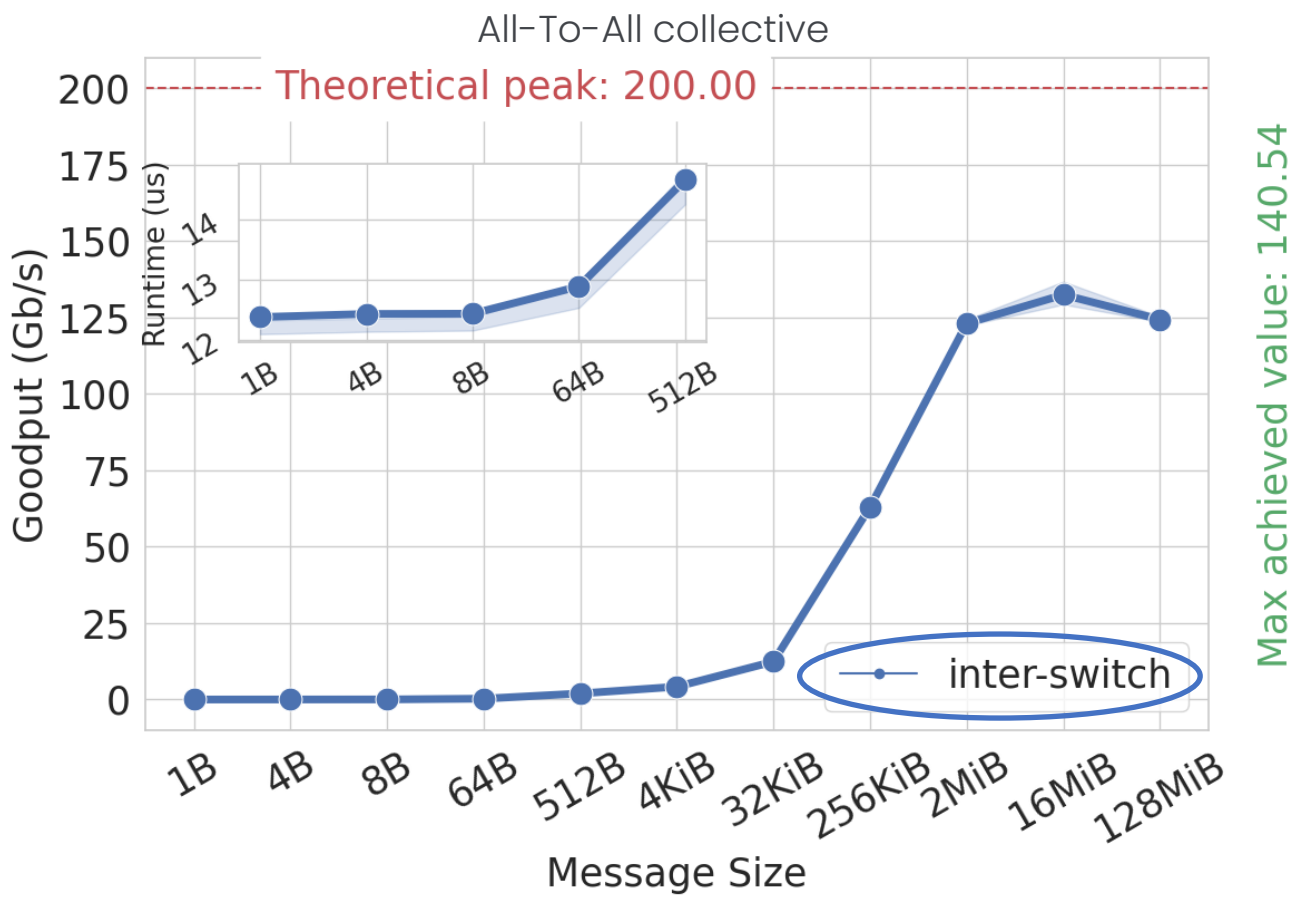




128MiB	27.51	96.30 %	124.36	62.18 %

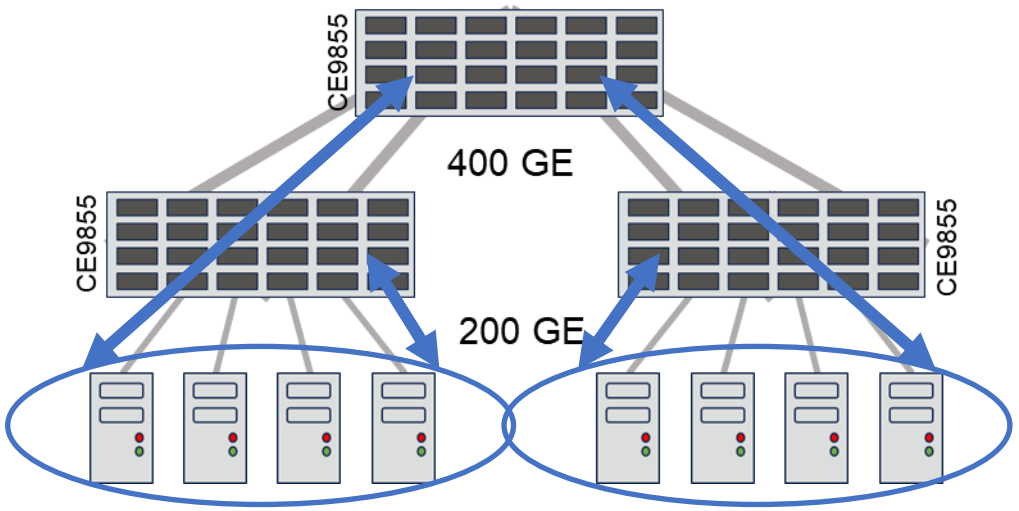
Results – System benchmark

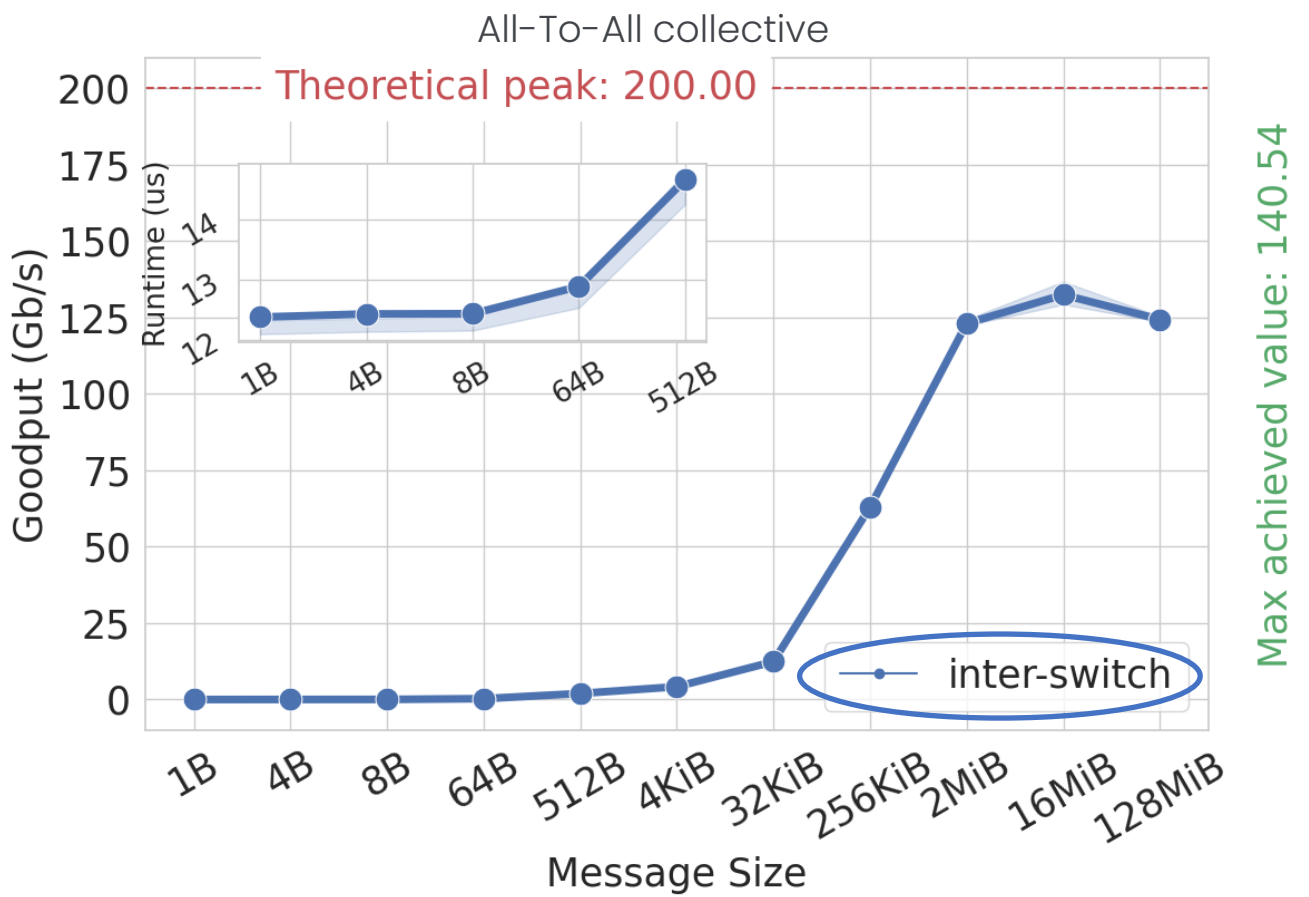




128MiB	27.51	96.30 %	124.36	62.18 %

Results – System benchmark





128MiB	27.51	96.30 %	124.36	62.18 %

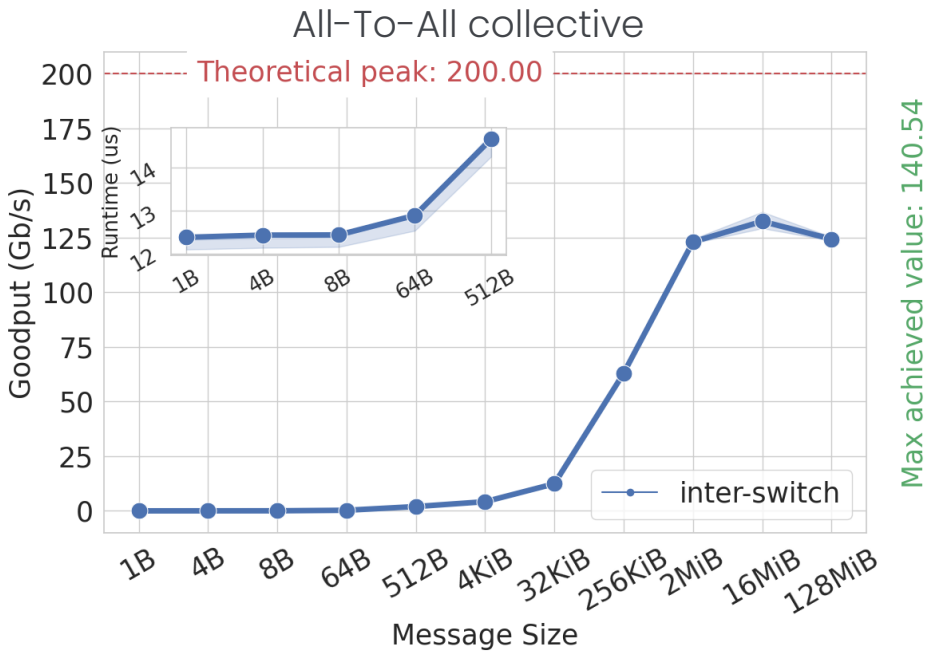
All-To-All collective
Theoretical peak: 200.00
inter-switch
Message Size

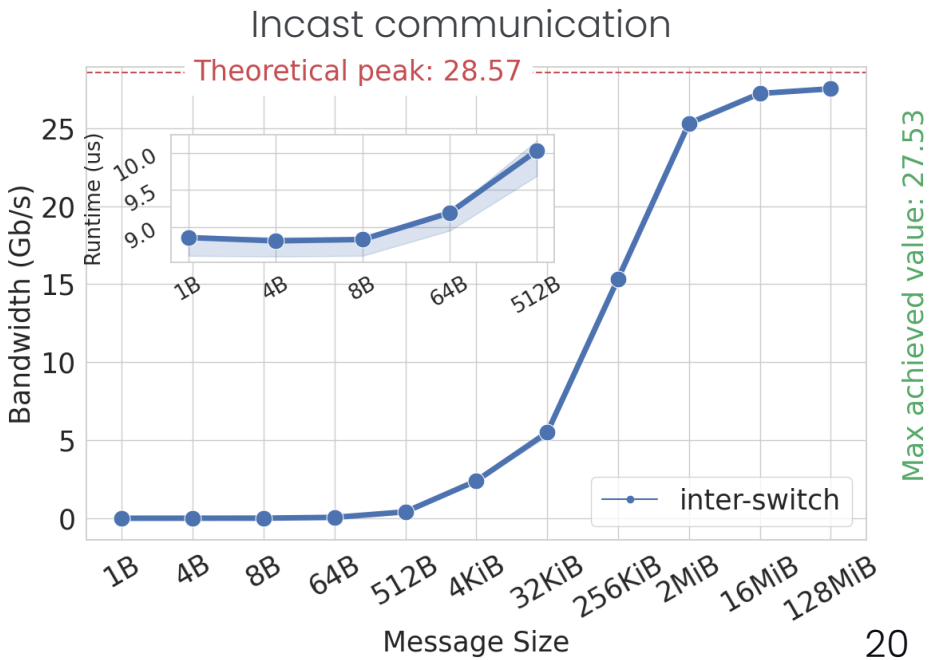
Results – System benchmark

Whole cluster performances:

- All-To-All grows until 66% of the peak.
- Incast saturate the bandwidth also on the whole cluster.

Size	ic		a2a	
	Bandwidth (Gb/s)	Peak %	Goodput (Gb/s)	Peak %
1B	0.0009	0.003 %	0.004	0.002 %
8B	0.007	0.025 %	0.036	0.018 %
512B	0.411	1.44 %	1.96	0.98 %
32KiB	5.49	19.22 %	12.4	6.20 %
2MiB	25.3	88.55 %	123.1	61.55 %
16MiB	27.21	95.24 %	132.46	66.23 %
128MiB	27.51	96.30 %	124.36	62.18 %



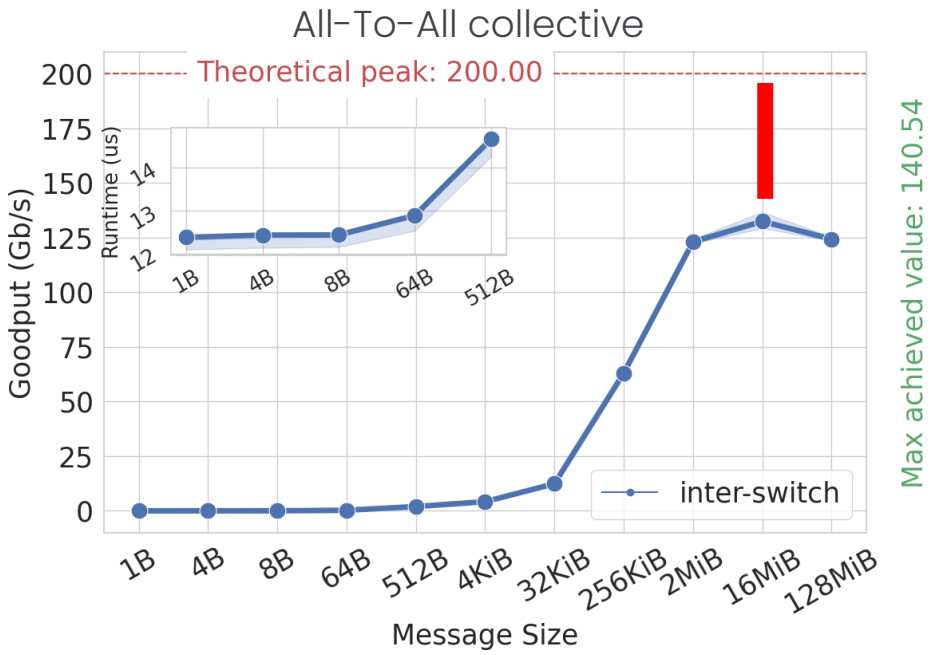


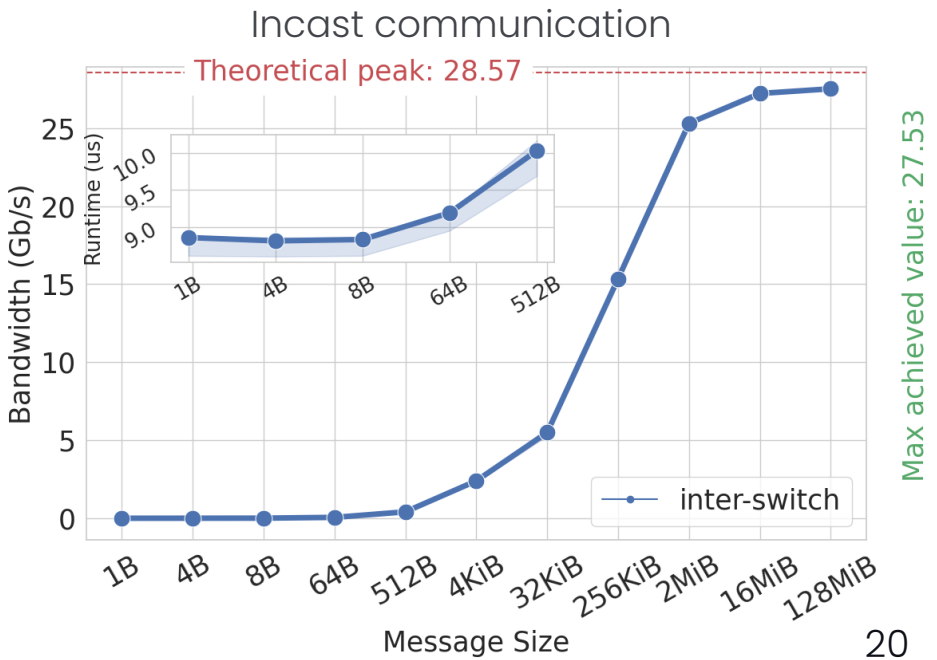
Results – System benchmark

Whole cluster performances:

- All-To-All grows until 66% of the peak.
- Incast saturate the bandwidth also on the whole cluster.

Size	ic		a2a	
	Bandwidth (Gb/s)	Peak %	Goodput (Gb/s)	Peak %
1B	0.0009	0.003 %	0.004	0.002 %
8B	0.007	0.025 %	0.036	0.018 %
512B	0.411	1.44 %	1.96	0.98 %
32KiB	5.49	19.22 %	12.4	6.20 %
2MiB	25.3	88.55 %	123.1	61.55 %
16MiB	27.21	95.24 %	132.46	66.23 %
128MiB	27.51	96.30 %	124.36	62.18 %



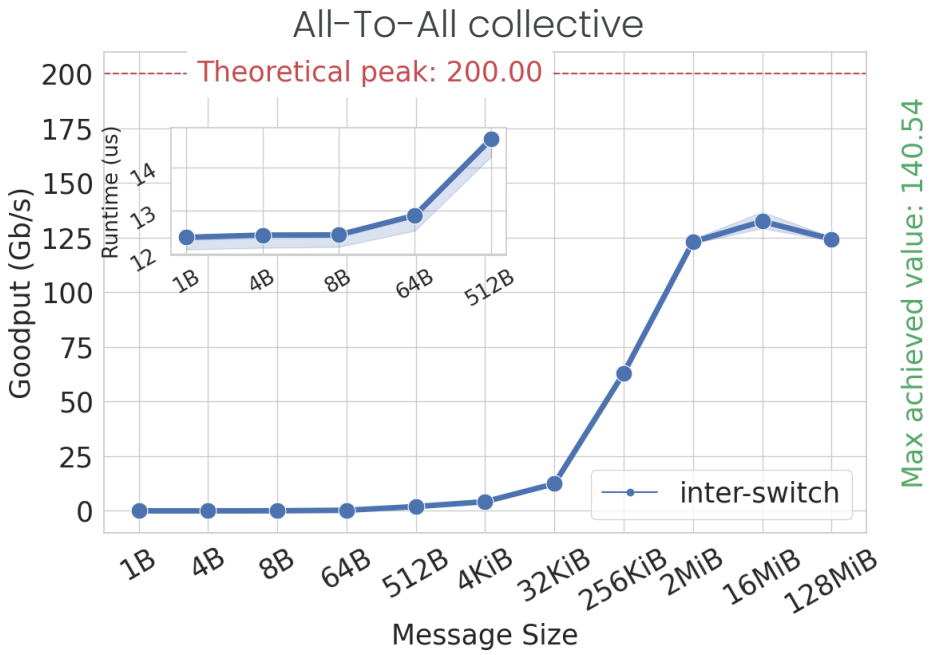


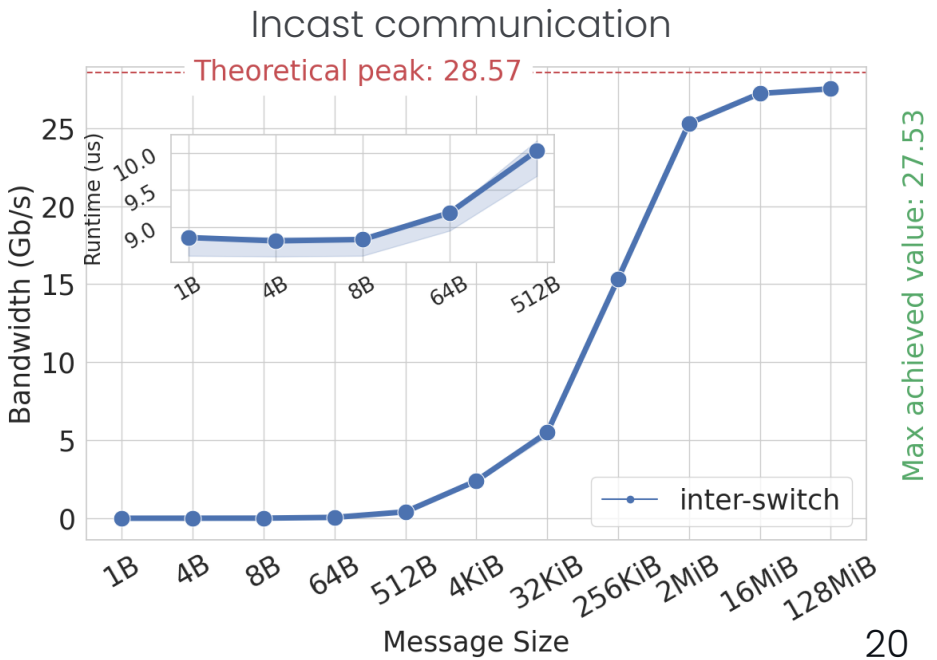
Results – System benchmark

Whole cluster performances:

- All-To-All grows until 66% of the peak.
- Incast saturate the bandwidth also on the whole cluster.

Size	ic		a2a	
	Bandwidth (Gb/s)	Peak %	Goodput (Gb/s)	Peak %
1B	0.0009	0.003 %	0.004	0.002 %
8B	0.007	0.025 %	0.036	0.018 %
512B	0.411	1.44 %	1.96	0.98 %
32KiB	5.49	19.22 %	12.4	6.20 %
2MiB	25.3	88.55 %	123.1	61.55 %
16MiB	27.21	95.24 %	132.46	66.23 %
128MiB	27.51	96.30 %	124.36	62.18 %



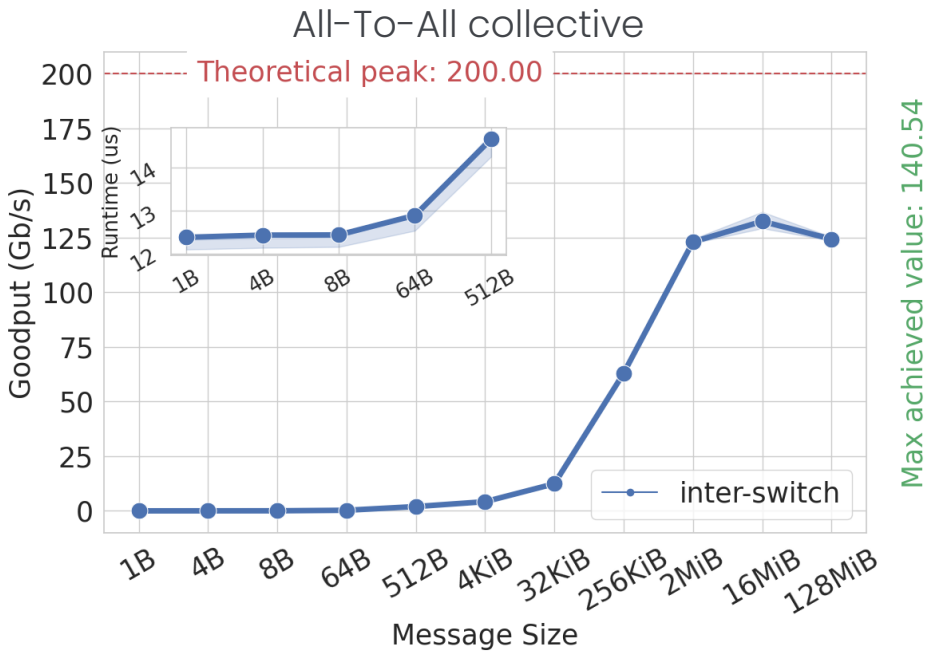


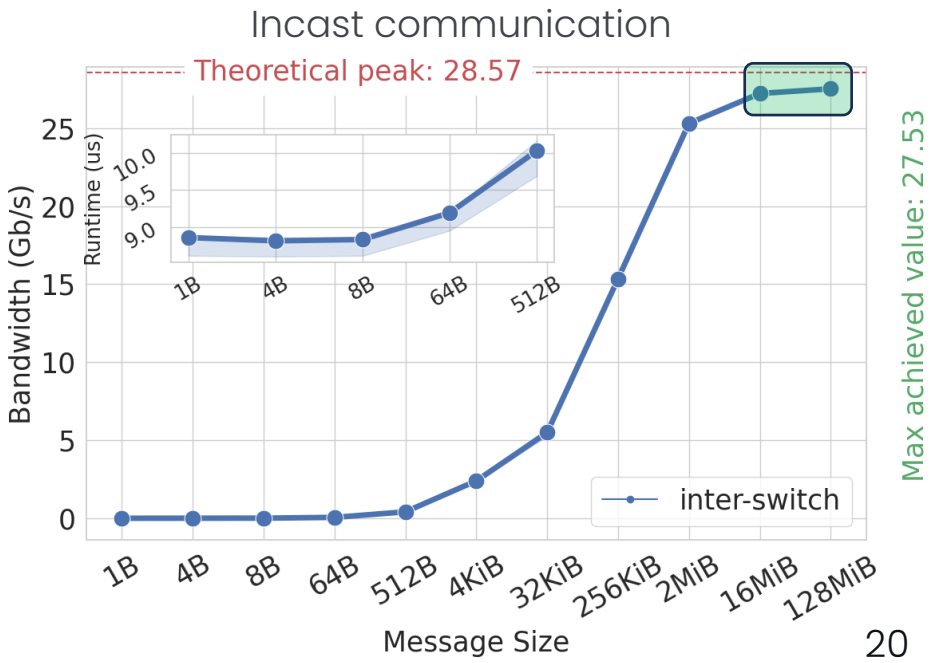
Results – System benchmark

Whole cluster performances:

- All-To-All grows until 66% of the peak.
- Incast saturate the bandwidth also on the whole cluster.

Size	ic		a2a	
	Bandwidth (Gb/s)	Peak %	Goodput (Gb/s)	Peak %
1B	0.0009	0.003 %	0.004	0.002 %
8B	0.007	0.025 %	0.036	0.018 %
512B	0.411	1.44 %	1.96	0.98 %
32KiB	5.49	19.22 %	12.4	6.20 %
2MiB	25.3	88.55 %	123.1	61.55 %
16MiB	27.21	95.24 %	132.46	66.23 %
128MiB	27.51	96.30 %	124.36	62.18 %



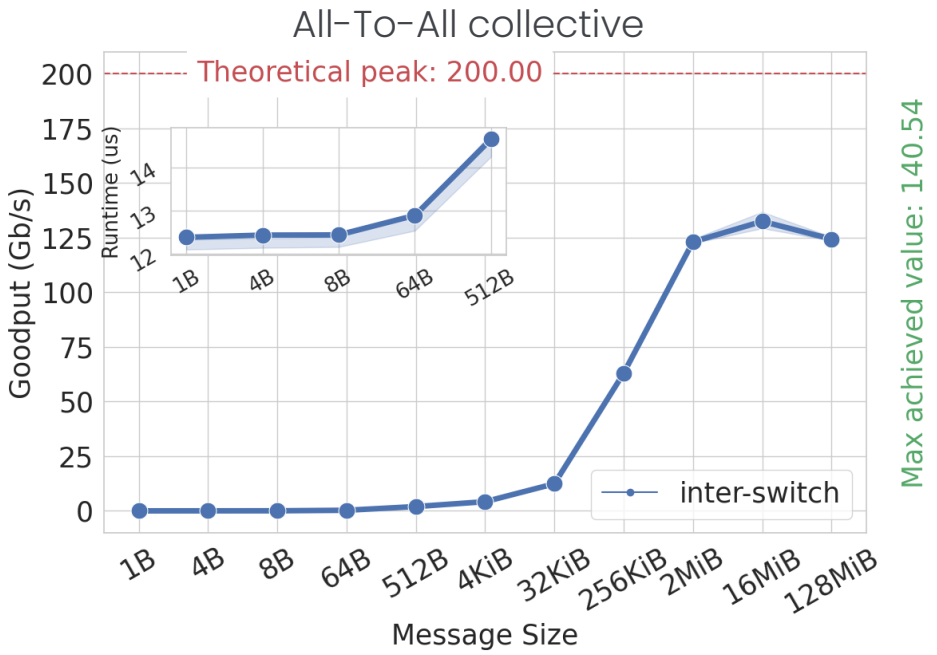


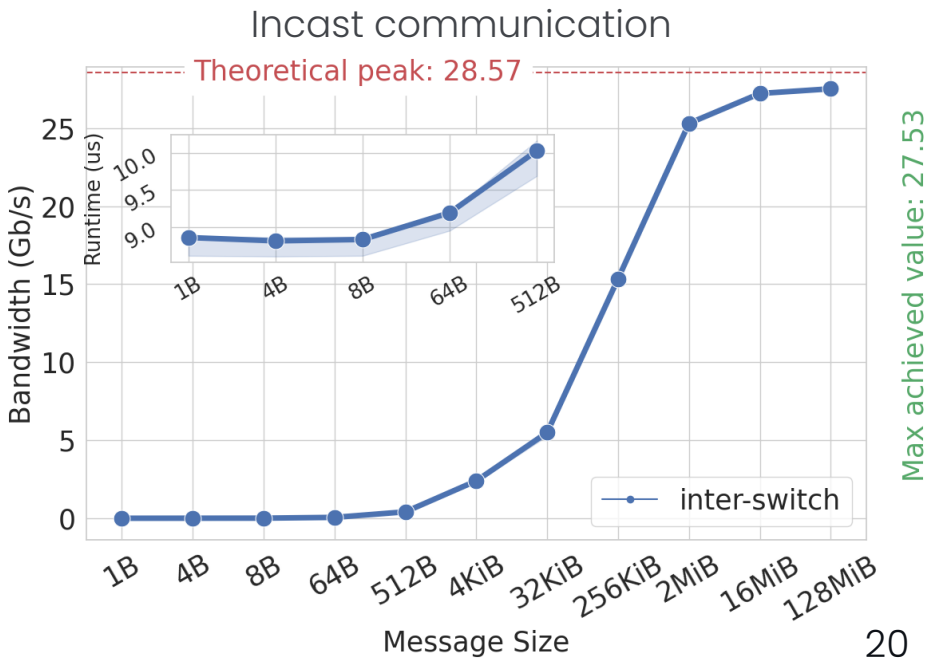
Results – System benchmark

Whole cluster performances:

- All-To-All grows until 66% of the peak.
- Incast saturate the bandwidth also on the whole cluster.

Size	ic		a2a	
	Bandwidth (Gb/s)	Peak %	Goodput (Gb/s)	Peak %
1B	0.0009	0.003 %	0.004	0.002 %
8B	0.007	0.025 %	0.036	0.018 %
512B	0.411	1.44 %	1.96	0.98 %
32KiB	5.49	19.22 %	12.4	6.20 %
2MiB	25.3	88.55 %	123.1	61.55 %
16MiB	27.21	95.24 %	132.46	66.23 %
128MiB	27.51	96.30 %	124.36	62.18 %





Conclusions

Considering the **upcoming UltraEthernet standards**, **we expect growth in Ethernet solutions** like the benchmarked Lossless Ethernet.

Our main findings are:

- From a **bandwidth perspective** Lossless Ethernet has approximately the same performance as InfiniBand.
- From a **latency perspective** InfiniBand still outperformed Ethernet.

Given the promising results:

- We plan **larger-scale benchmarks**;
- Characterization of **specific workloads**.

Thank you
for your attention

Questions are welcome

UNIVERSITÀ
DI TRENTO