CacheCloud
Towards Speed of Light Datacenter Communication

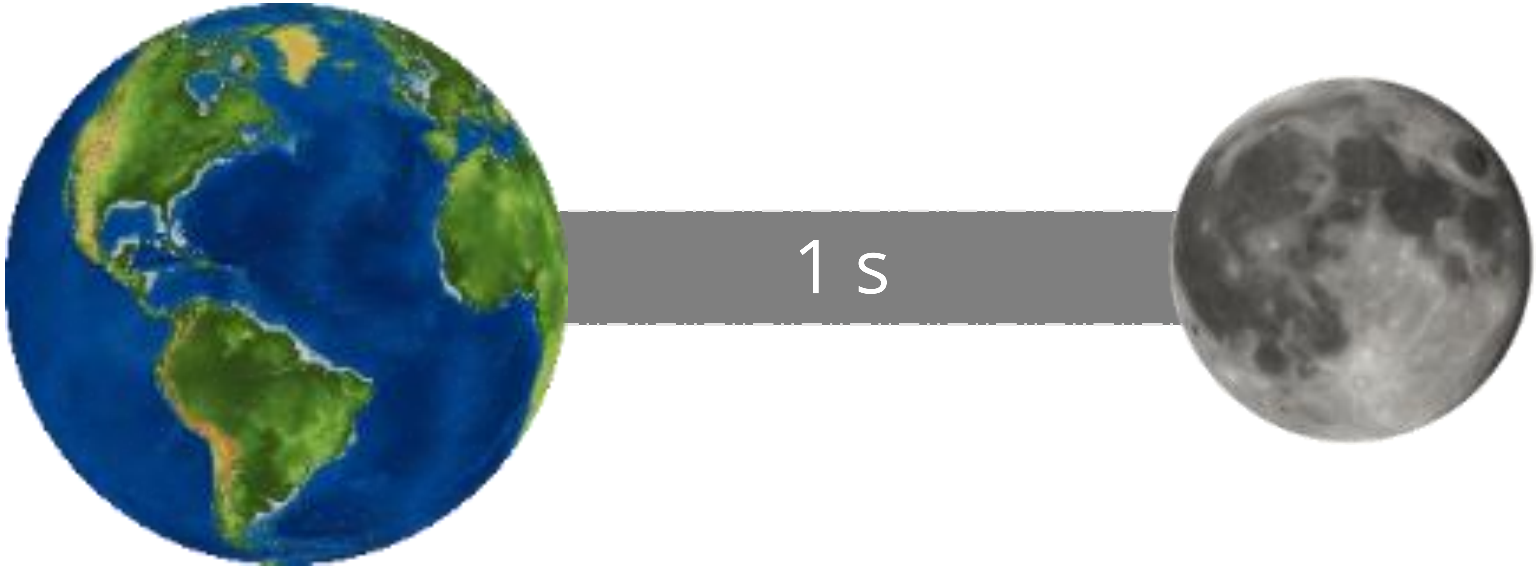
Shelby Thomas

Geoff Voelker

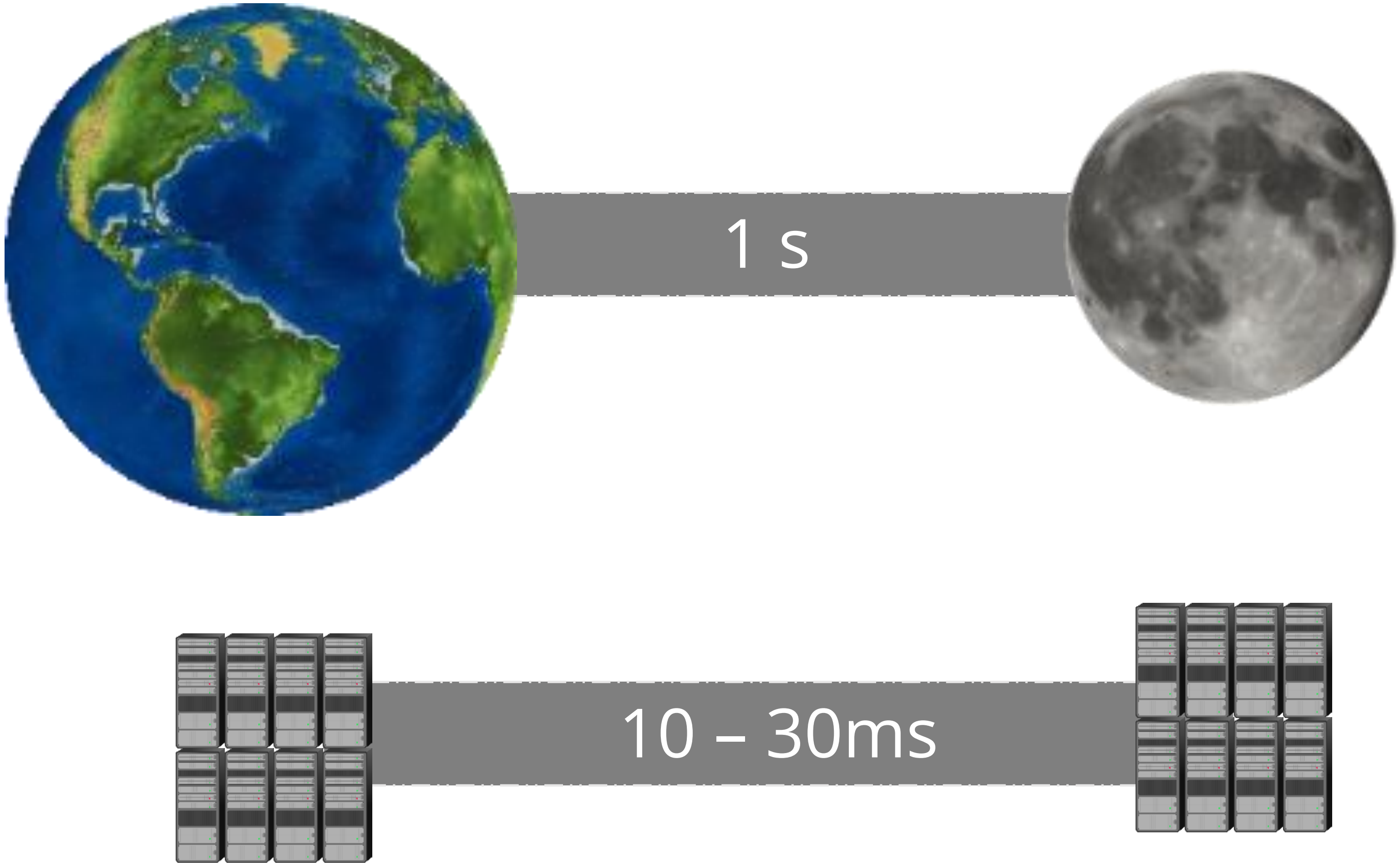
George Porter

University of California, San Diego

The Speed of Light Baseline



The Speed of Light Baseline



1 s

10 – 30ms

[Singla Et. Al.]

Speed of Light in the **Wide Area**

Bufferbloat and congestion

DNS resolution

Distance

10x – 1000x slower than speed of light propagation in WAN

[Singla Et. Al.]

Speed of Light in the **Wide Area**

Bufferbloat and congestion

What about the Local Area?

Distance

10x – 1000x slower than speed of light propagation in WAN

[Singla Et. Al.]

Speed of Light in the **Local** Area

~~Bufferbloat and congestion~~

~~DNS resolution~~

~~Distance~~

Expectation: Datacenter applications are closer to the speed of light than wide area networks.

Speed of Light in the **Local** Area

~~Bufferbloat and congestion~~

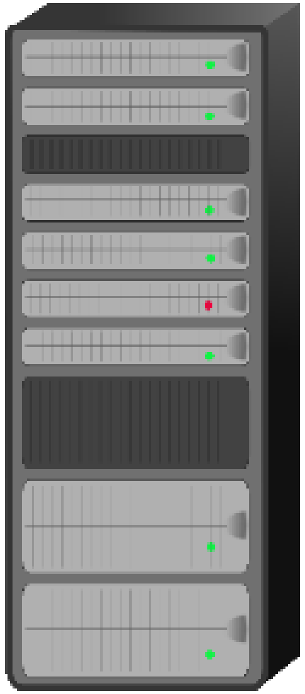
~~DNS resolution~~

~~Distance~~

Reality: Datacenter applications are **10x – 1000x slower** than speed of light propagation in the LAN

Different Distances Same Performance Gap

NORTH AMERICA
ASIA
EUROPE
Atlantic Ocean
Pacific Ocean
AFRICA
SOUTH AMERICA
AUSTRALIA
Indian Ocean
Pacific Ocean

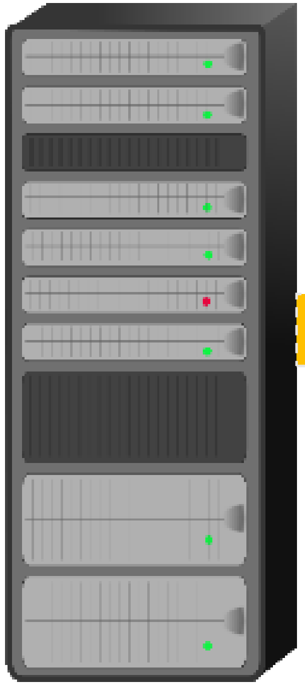


Datacenter Propagation Delay

100ns @ 20m

[Coleman Et. Al.]

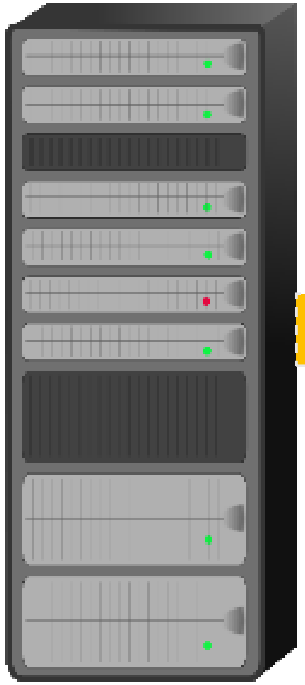
Datacenter Propagation Delay



Datacenter applications do not run at nanosecond scale

100ns @ 20m

Datacenter Propagation Delay



Datacenter applications do not run at nanosecond scale

100ns @ 20m

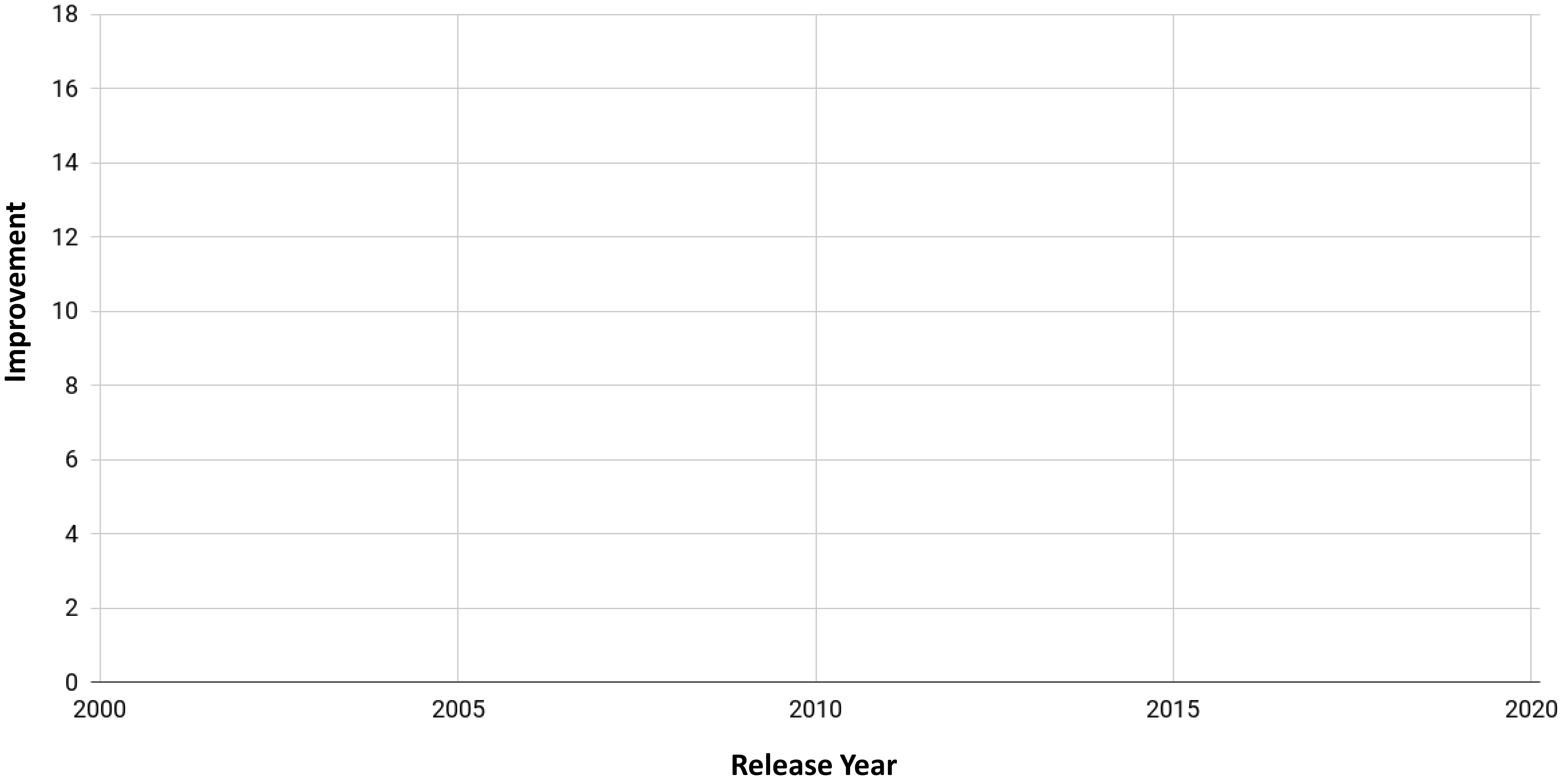
Application	Latency	Gap
Key Value Stores	4.3us – 100us	**43x**
Industry End-to-End RPC	75us	**750x**
Consensus	100us – 300us	**1000x**

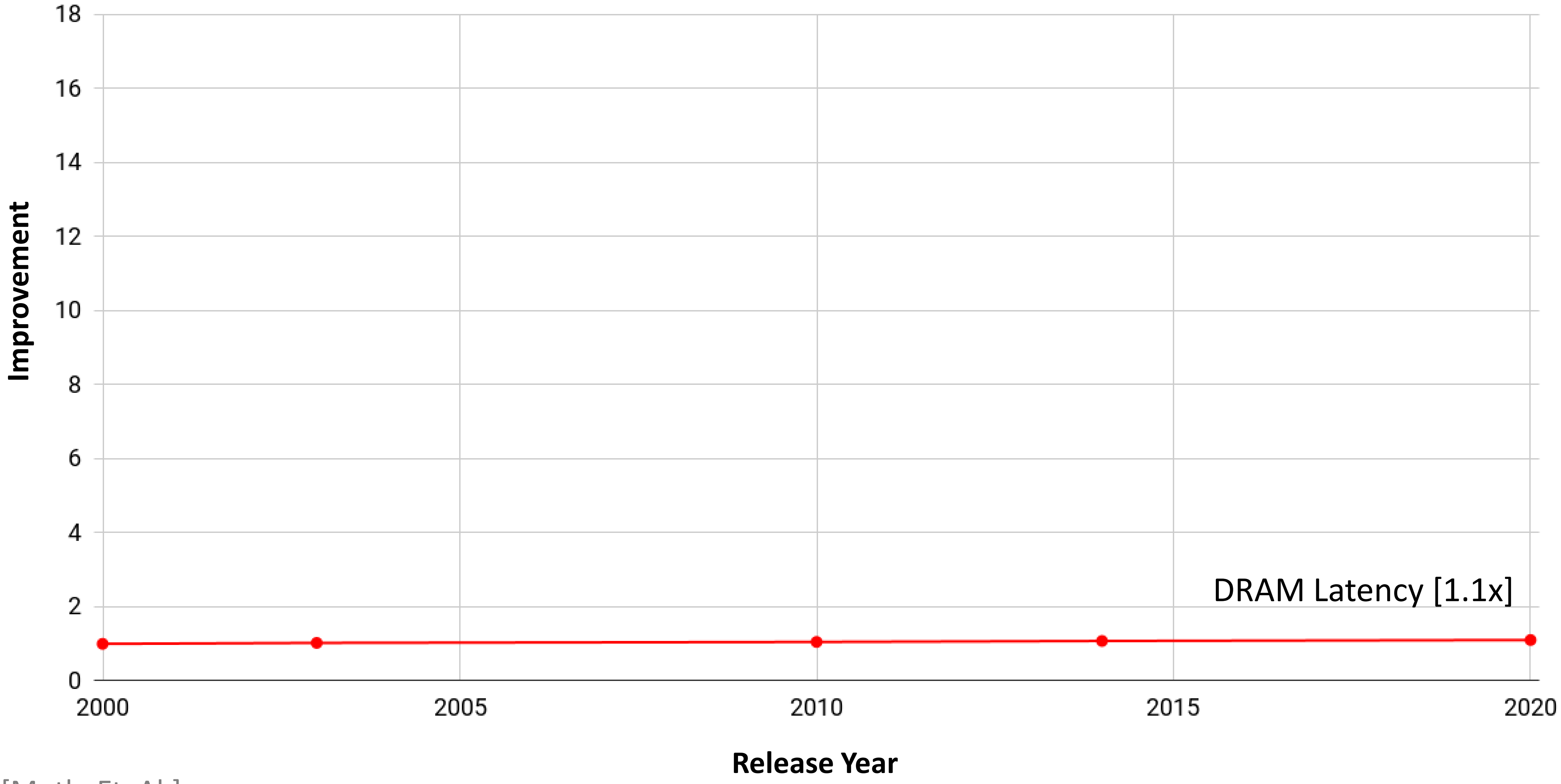
Datacenter Propagation Delay

Datacenter applications do not run at nanosecond scale

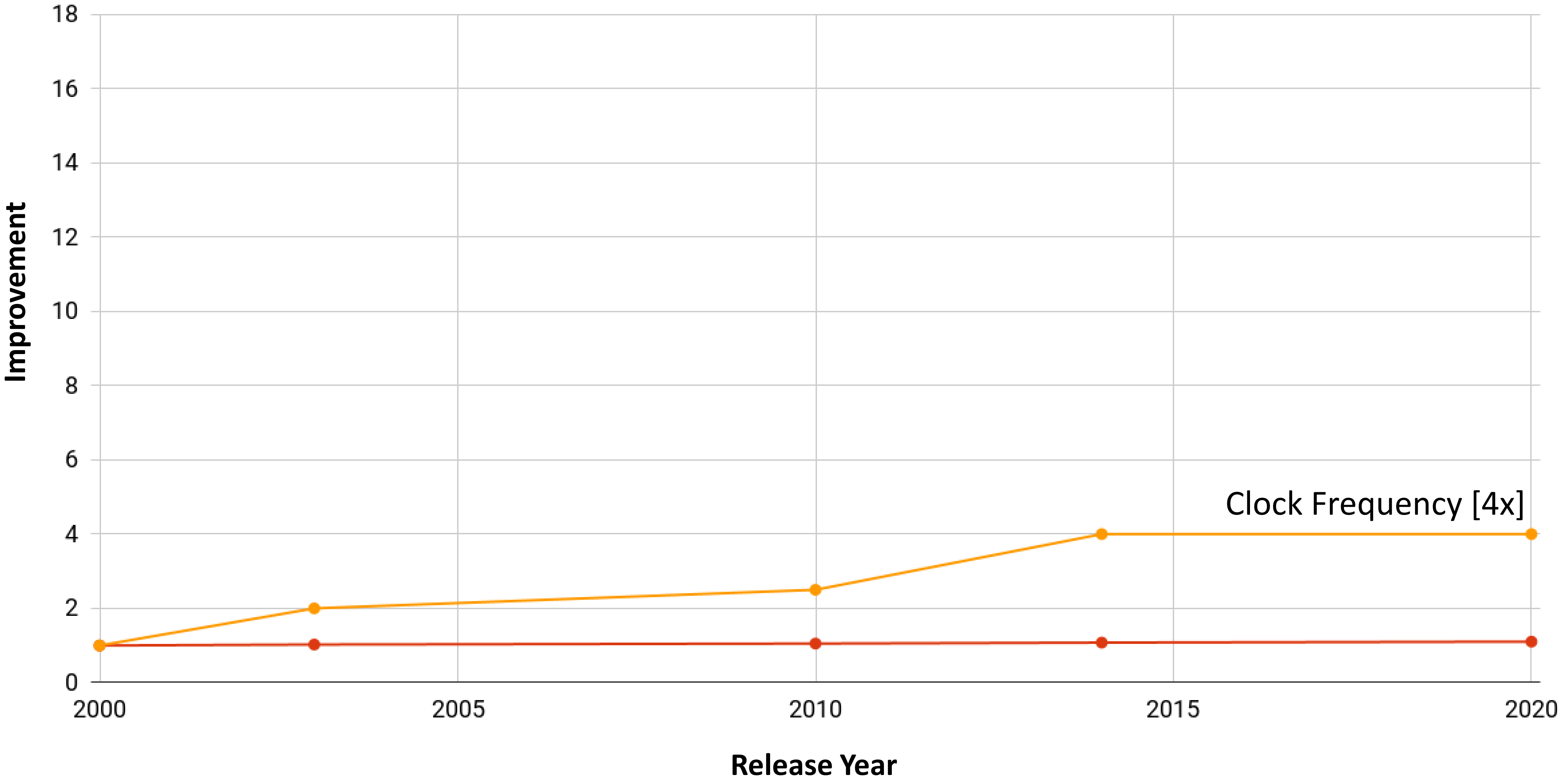
How has our infrastructure evolved?

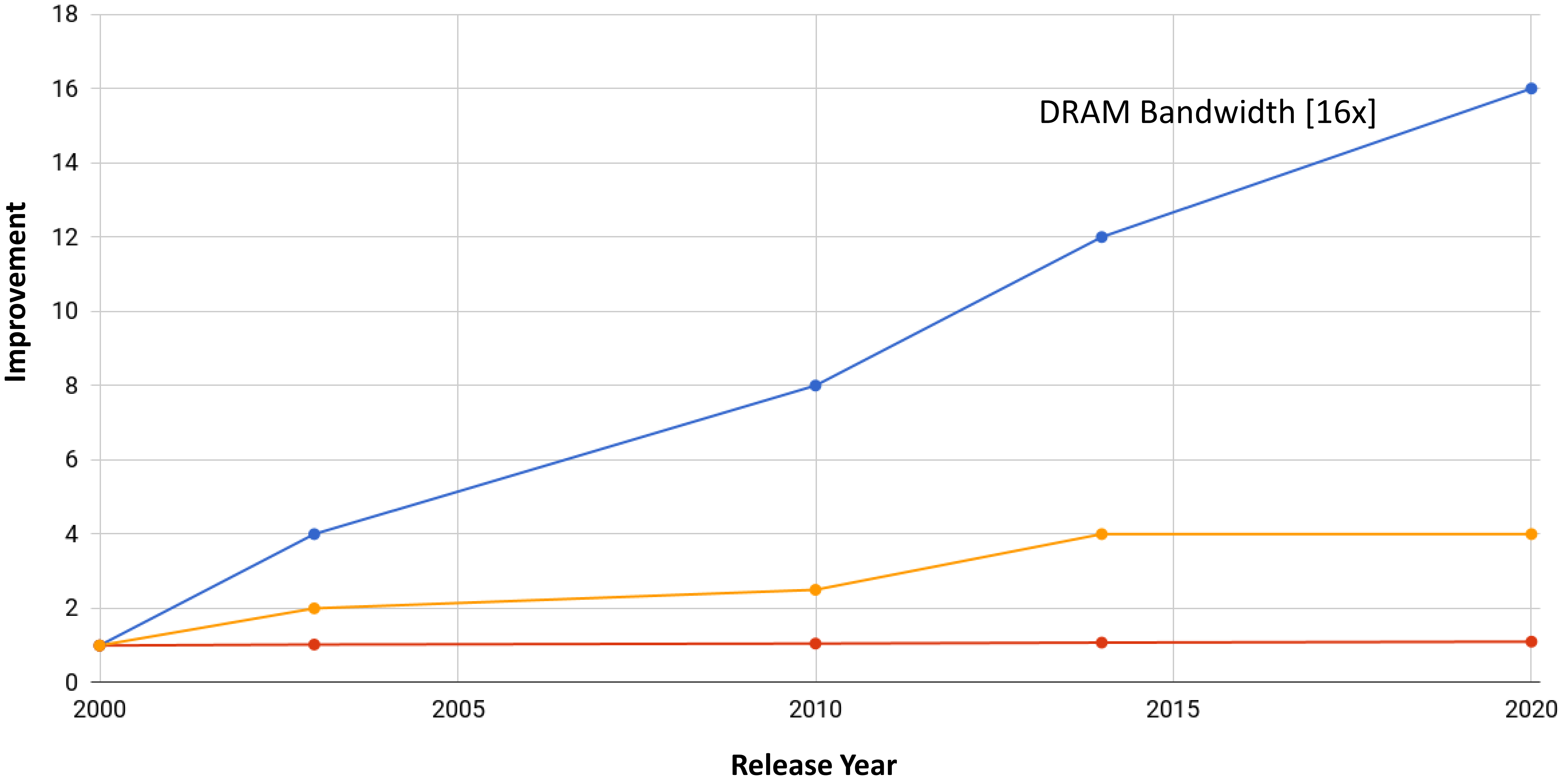
Application	Latency	Gap
Key Value Stores	4.3us – 100us	**43x**
Industry End-to-End RPC	75us	**750x**
Consensus	100us – 300us	**1000x**


18
16
14
12
10
8
6
4
2
0
Improvement
2000
2005
2010
2015
2020
Release Year

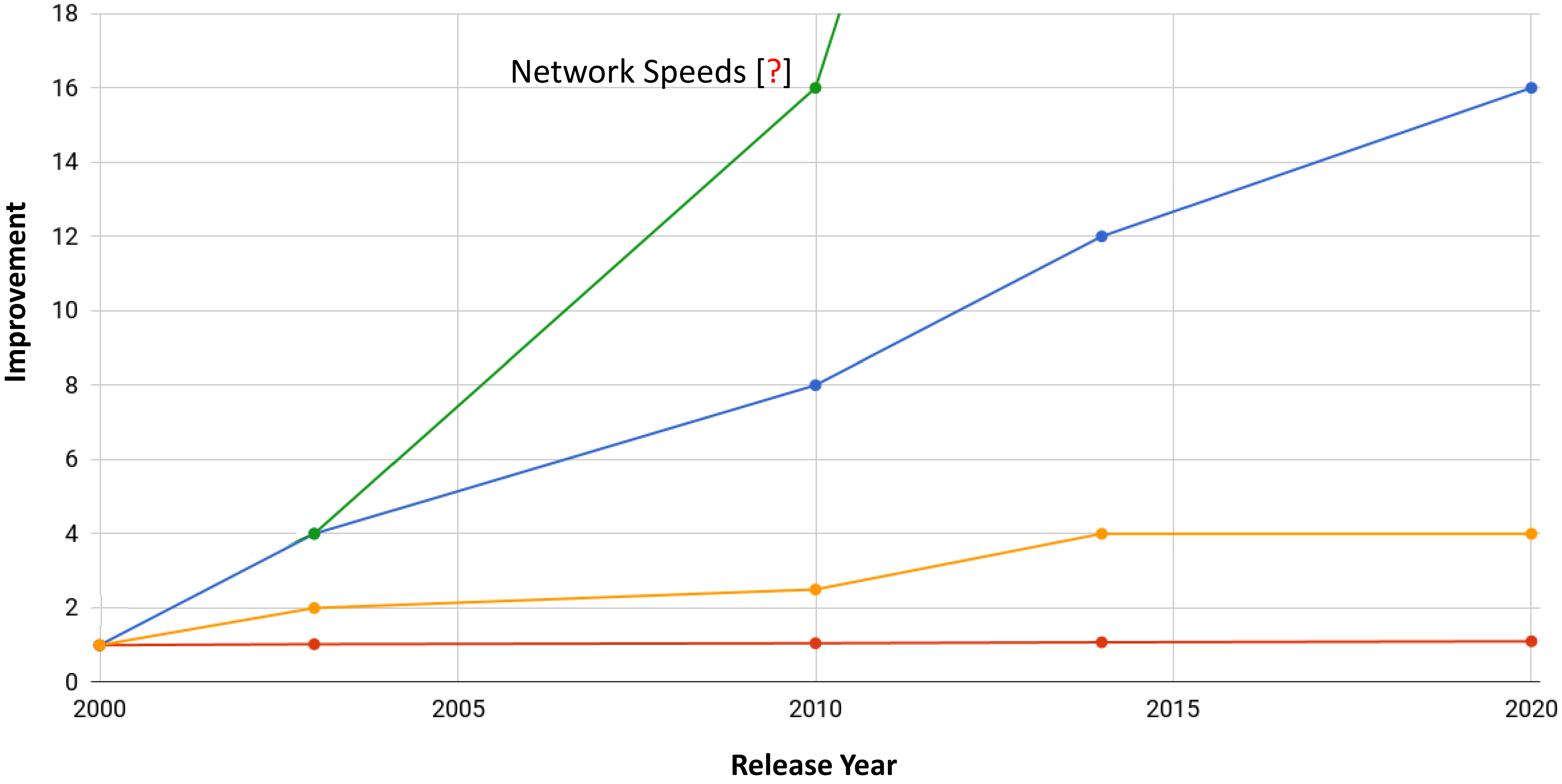

Improvement
18
16
14
12
10
8
6
4
2
0
2000
2005
2010
2015
2020
Release Year
DRAM Latency [1.1x]

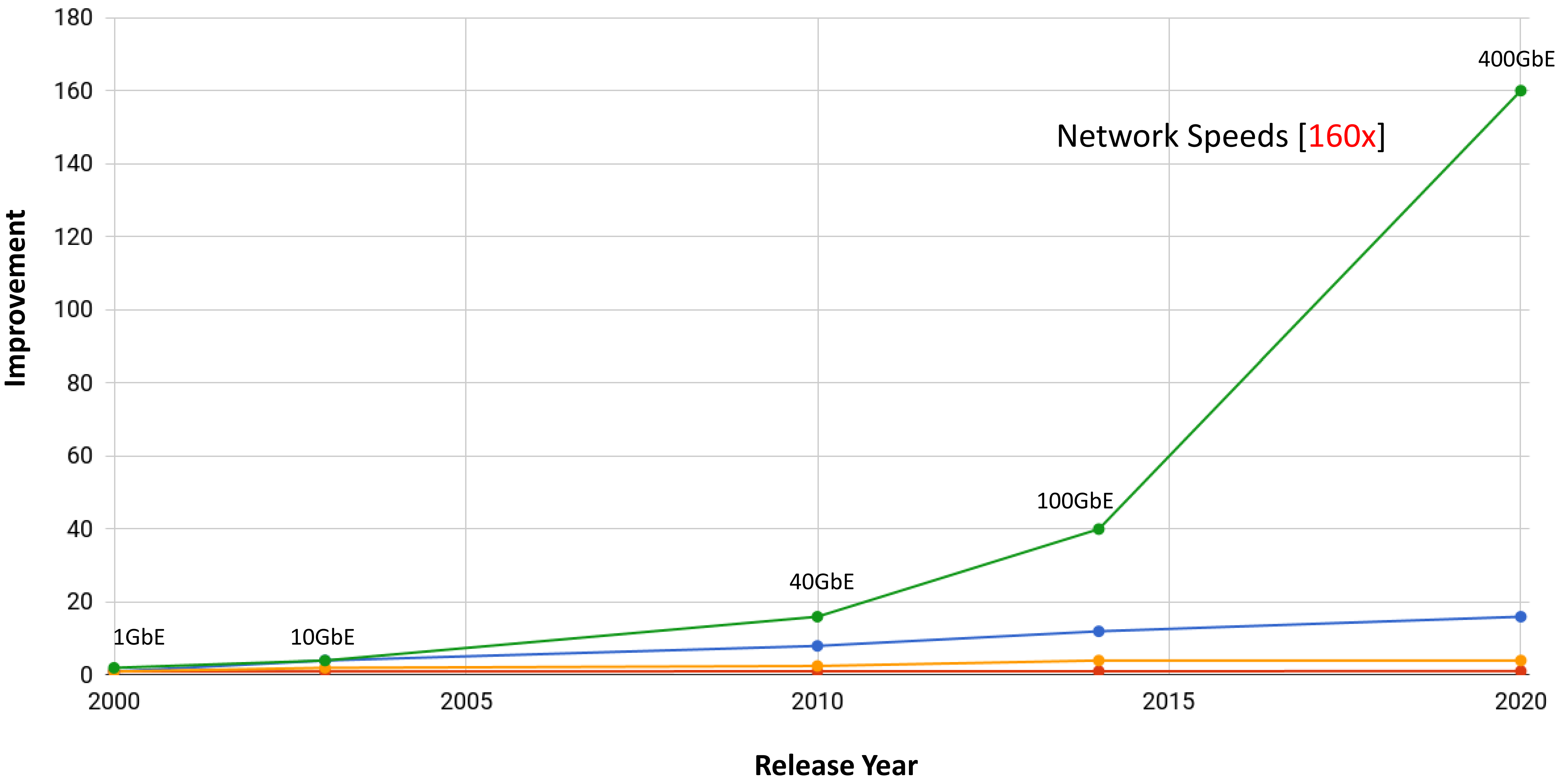
[Mutlu Et. Al.]

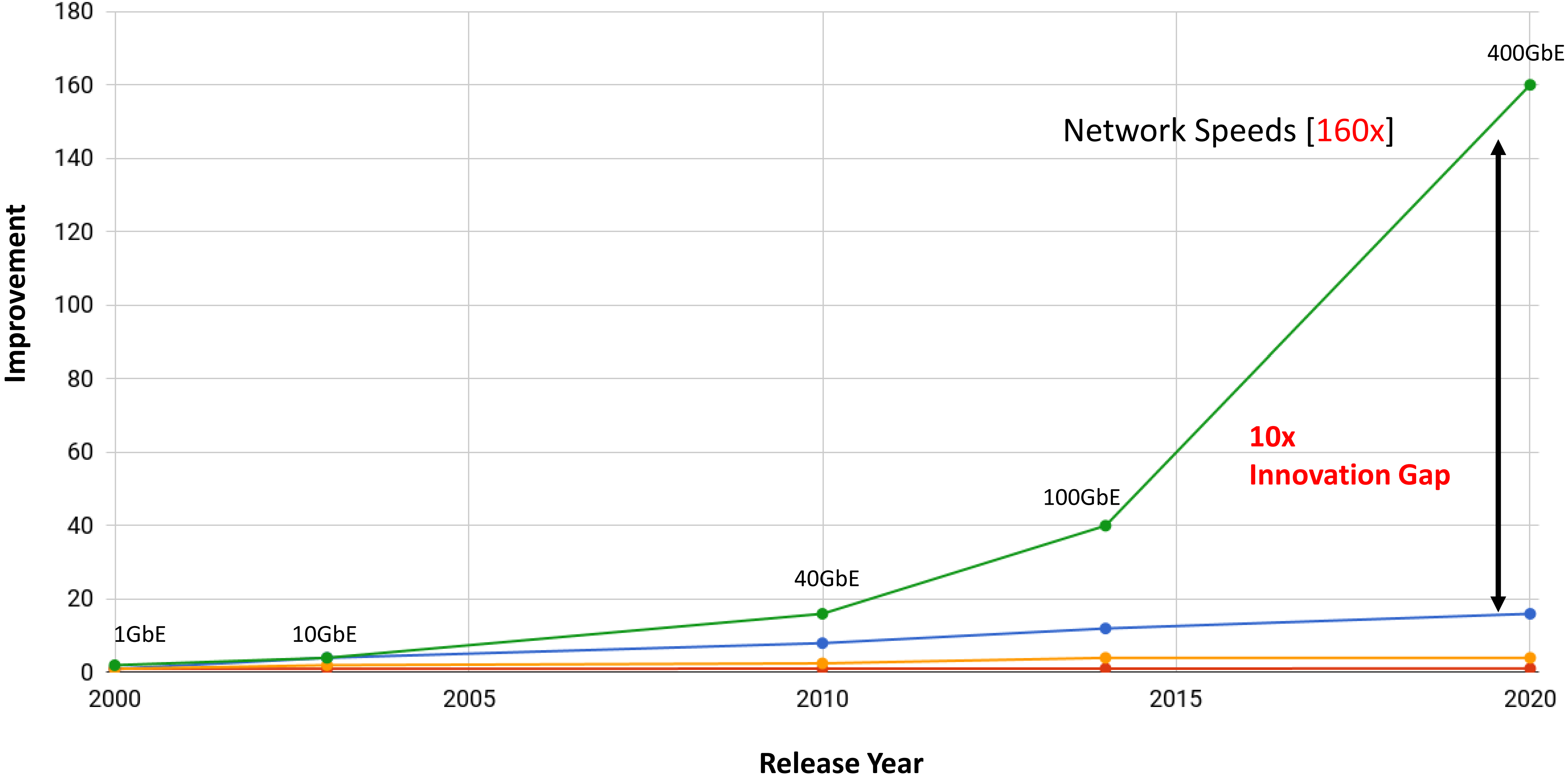

Improvement
18
16
14
12
10
8
6
4
2
0
2000
2005
2010
2015
2020
Release Year
Clock Frequency [4x]

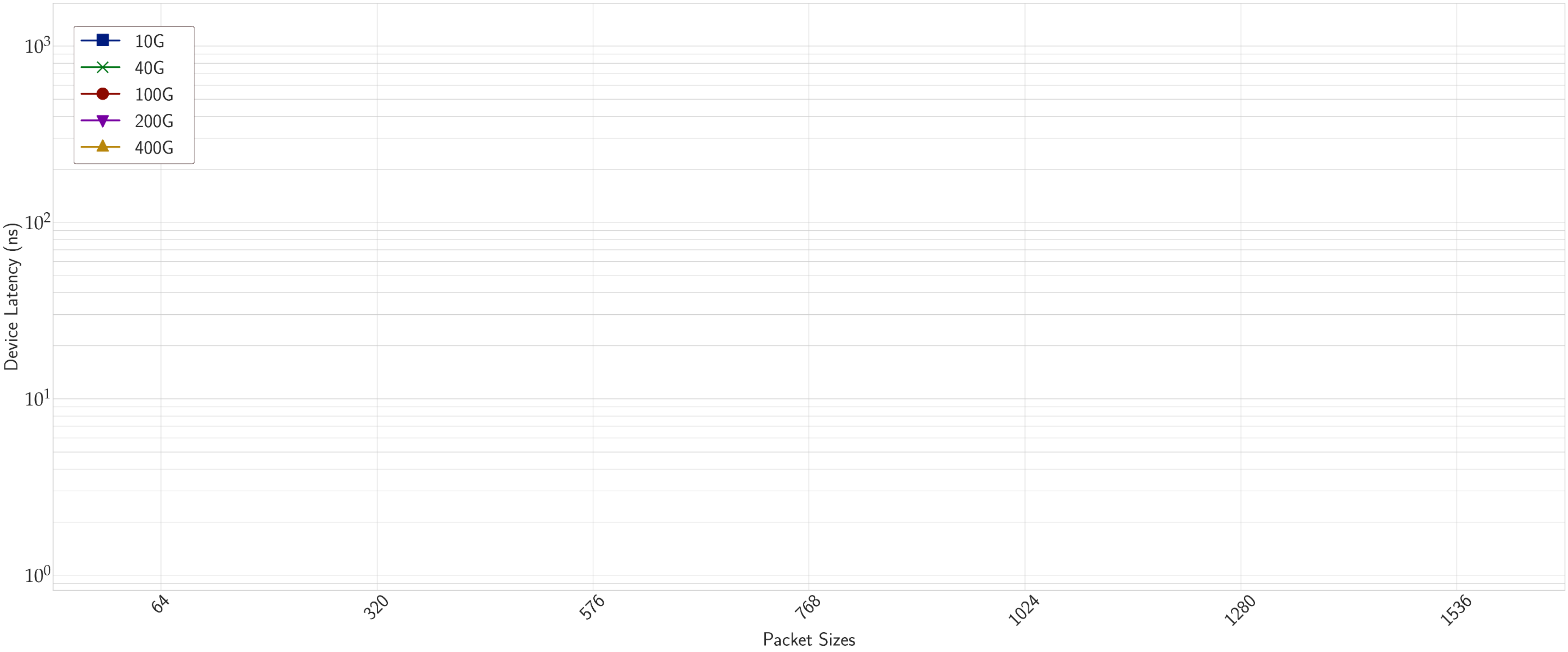


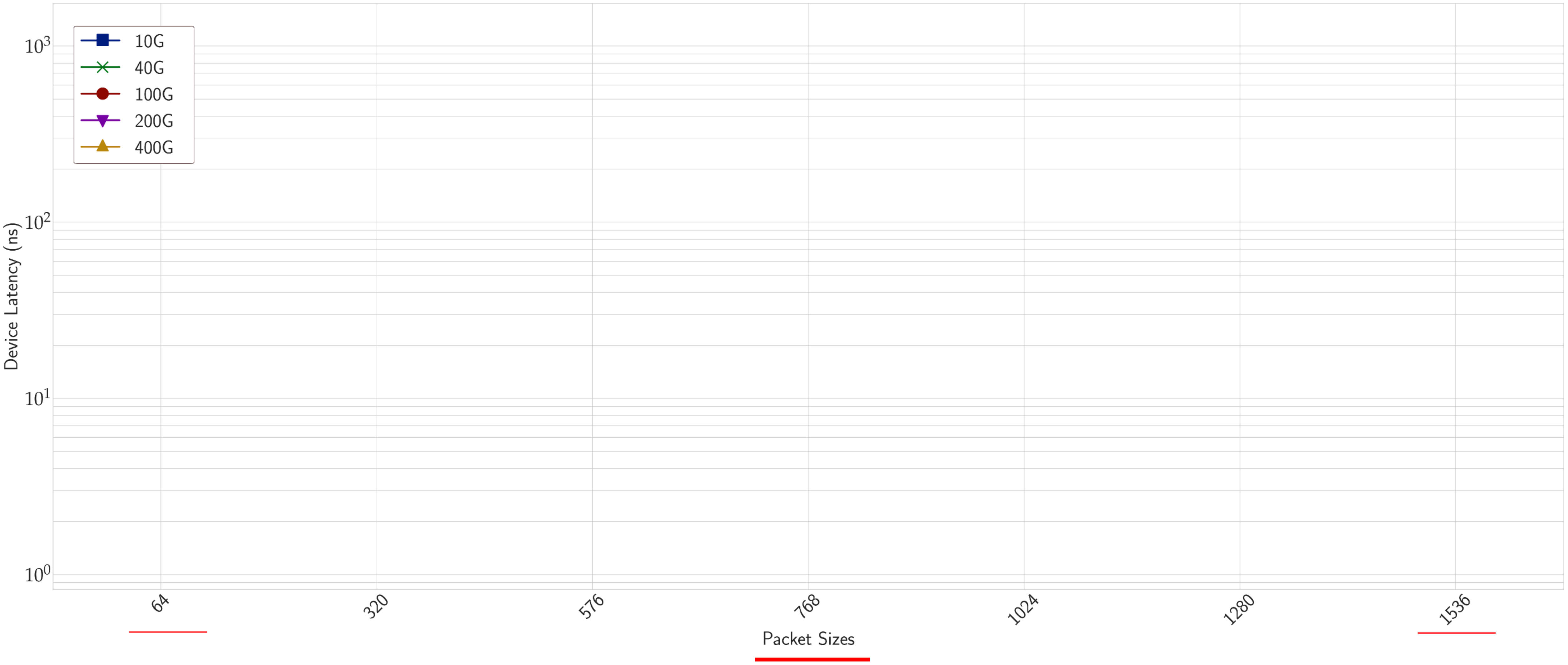
DRAM Bandwidth [16x]
Improvement
18
16
14
12
10
8
6
4
2
0
2000
2005
2010
2015
2020
Release Year

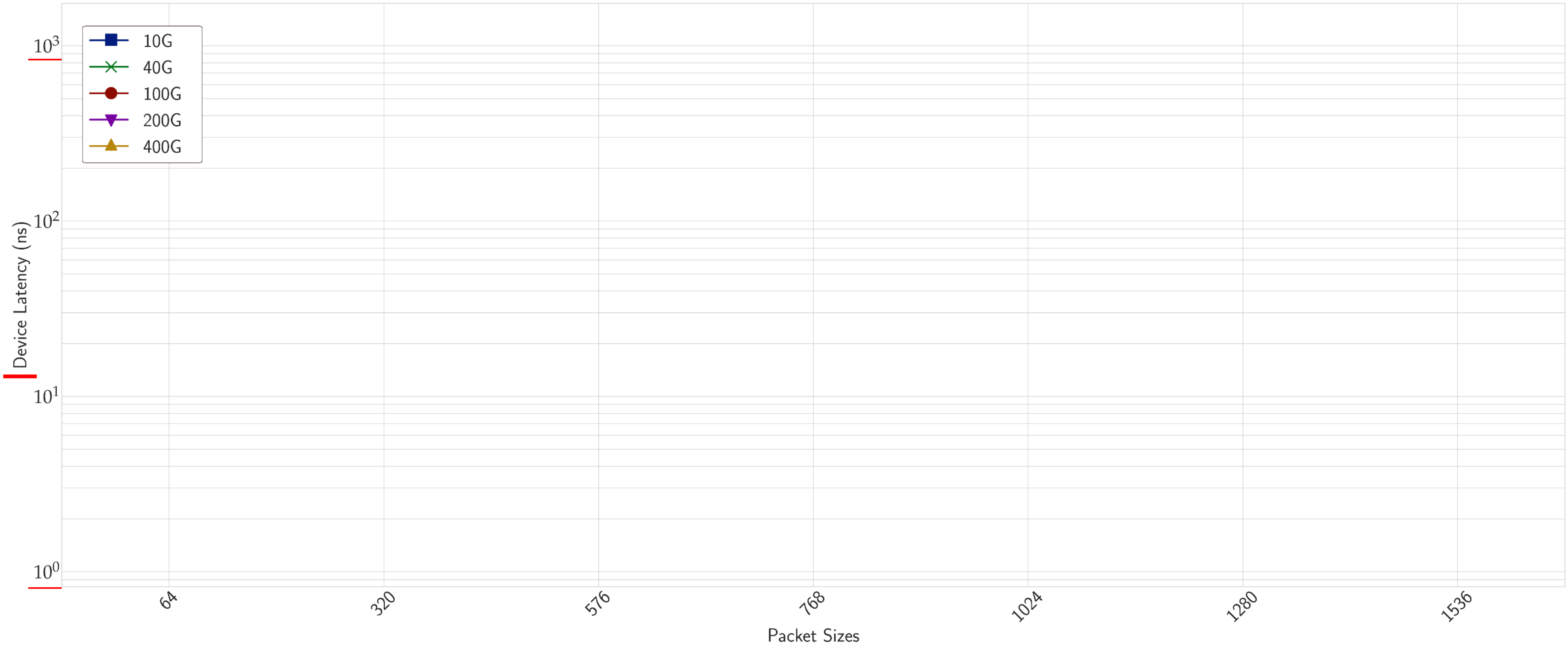

Network Speeds [?]
Improvement
0
2
4
6
8
10
12
14
16
18
2000
2005
2010
2015
2020
Release Year

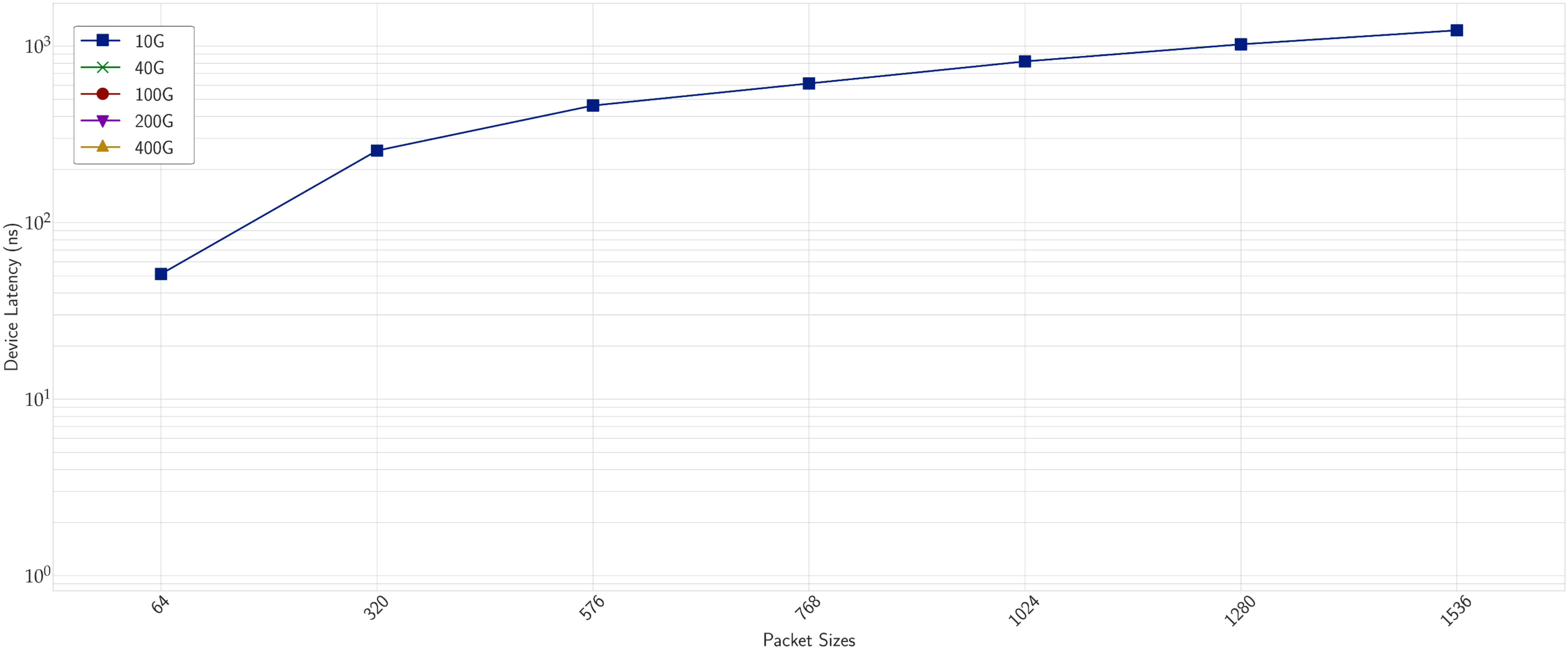

Network Speeds [160x]
400GbE
100GbE
40GbE
10GbE
1GbE
Improvement
180
160
140
120
100
80
60
40
20
0
2000
2005
2010
2015
2020
Release Year

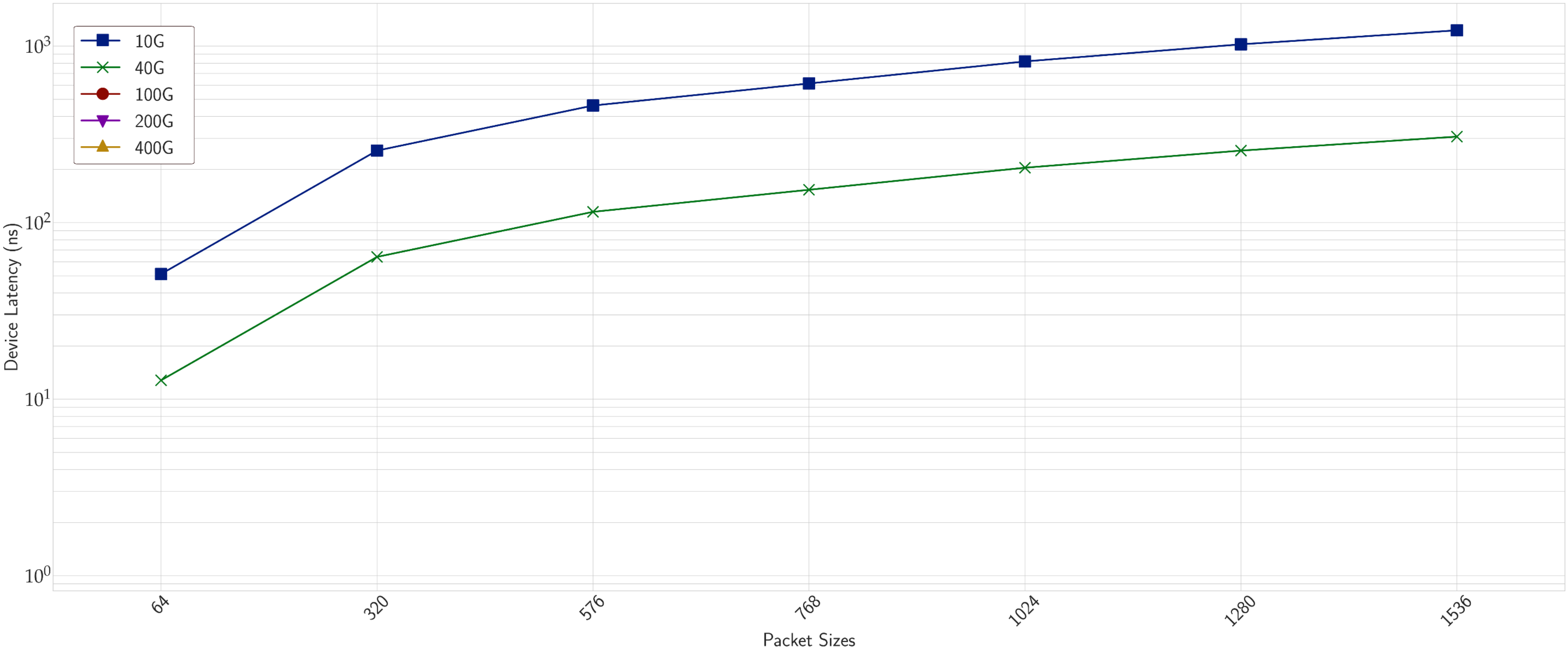

400GbE
Network Speeds [160x]
10x
Innovation Gap
100GbE
40GbE
1GbE
10GbE
180
160
140
120
100
80
60
40
20
0
Improvement
2000
2005
2010
2015
2020
Release Year

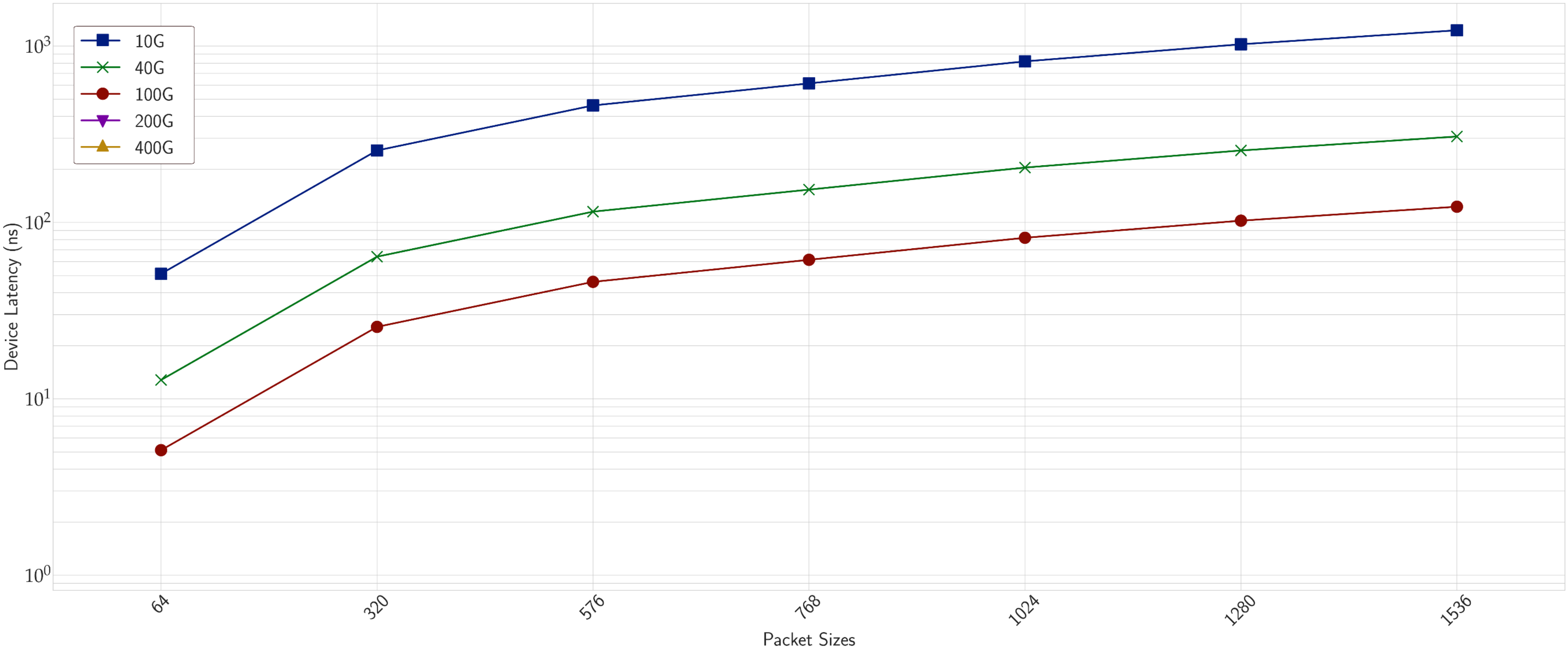

10G
40G
100G
200G
400G
Device Latency (ns)
10^3
10^2
10^1
10^0
64
320
576
768
1024
1280
1536
Packet Sizes

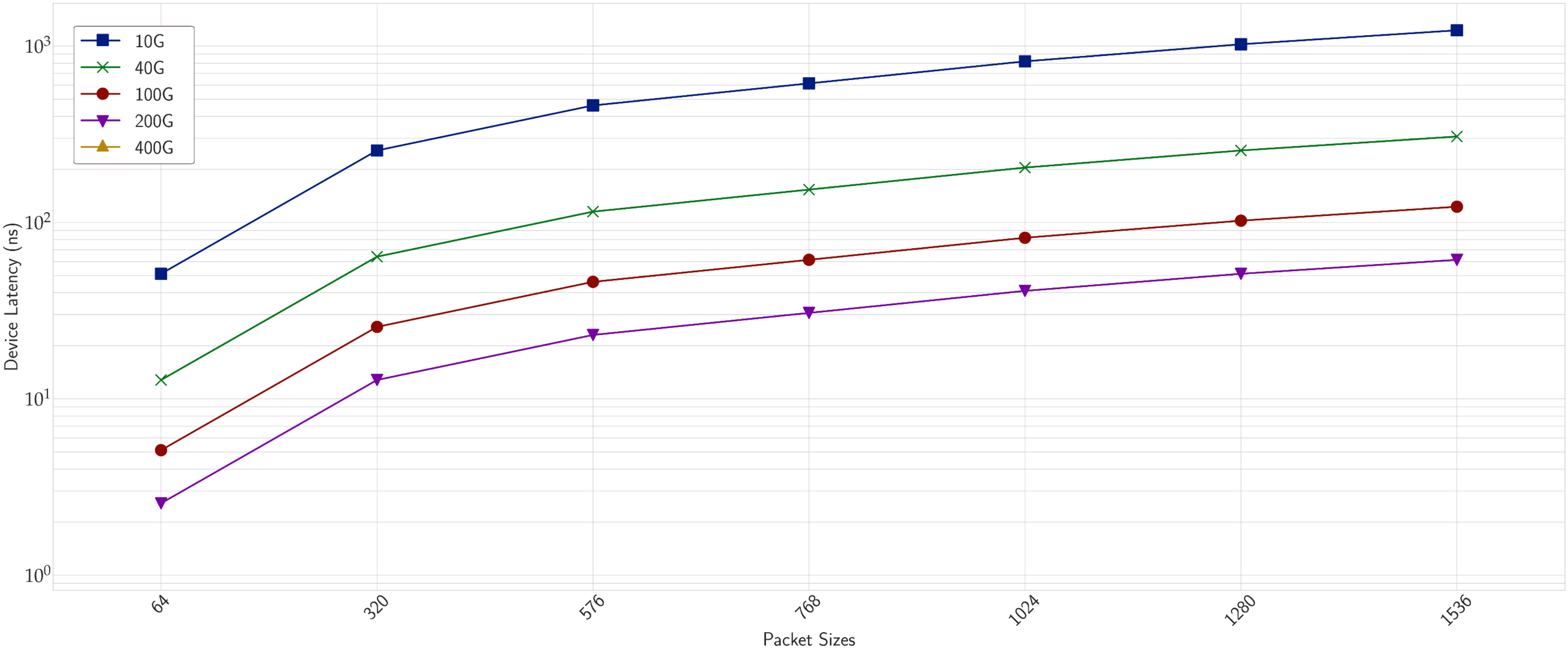

10G
40G
100G
200G
400G
10^3
10^2
10^1
10^0
Device Latency (ns)
64
320
576
768
1024
1280
1536
Packet Sizes

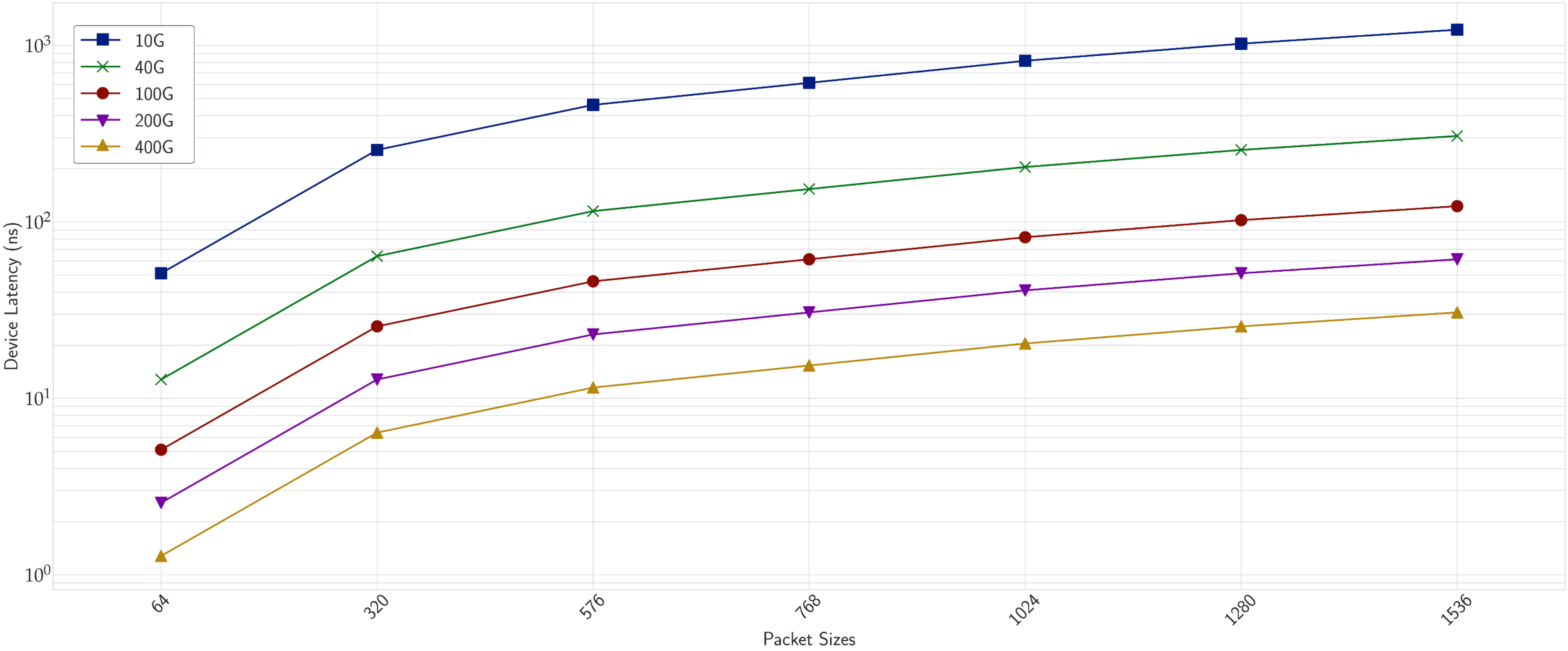

10G
40G
100G
200G
400G
Device Latency (ns)
10^3
10^2
10^1
10^0
64
320
576
768
1024
1280
1536
Packet Sizes


10G
40G
100G
200G
400G
10^3
10^2
10^1
10^0
Device Latency (ns)
64
320
576
768
1024
1280
1536
Packet Sizes

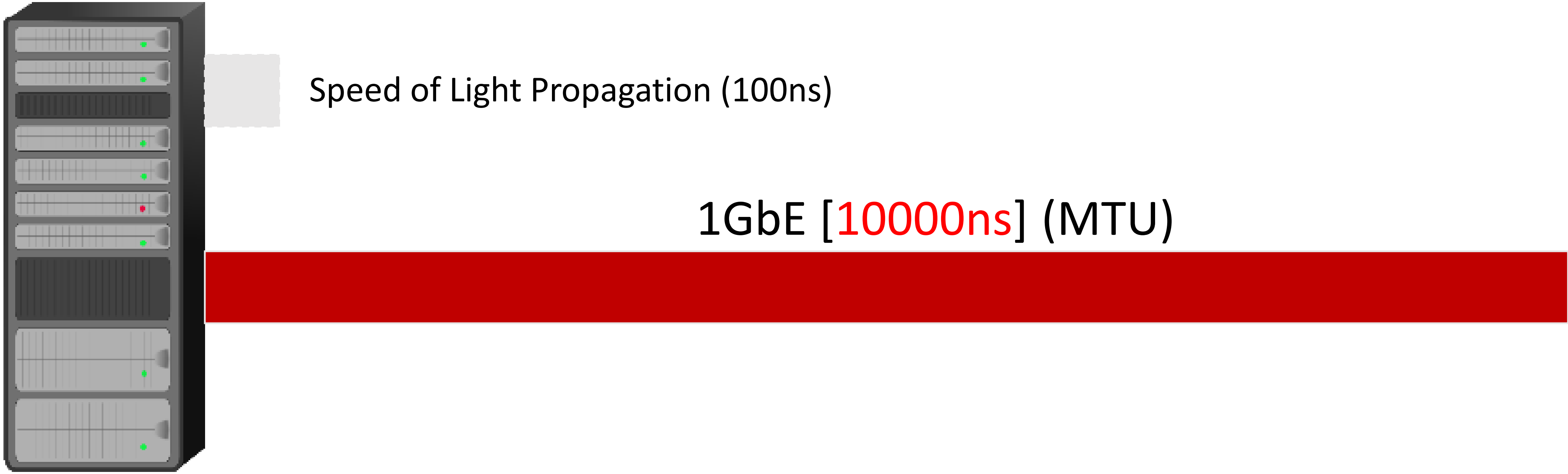

10G
40G
100G
200G
400G
Device Latency (ns)
10^3
10^2
10^1
10^0
64
320
576
768
1024
1280
1536
Packet Sizes


10G
40G
100G
200G
400G
Device Latency (ns)
10^3
10^2
10^1
10^0
64
320
576
768
1024
1280
1536
Packet Sizes


10G
40G
100G
200G
400G
Device Latency (ns)
10^3
10^2
10^1
10^0
64
320
576
768
1024
1280
1536
Packet Sizes


10G
40G
100G
200G
400G
Device Latency (ns)
10^3
10^2
10^1
10^0
64
320
576
768
1024
1280
1536
Packet Sizes

Interpacket Gap Decreased Drastically

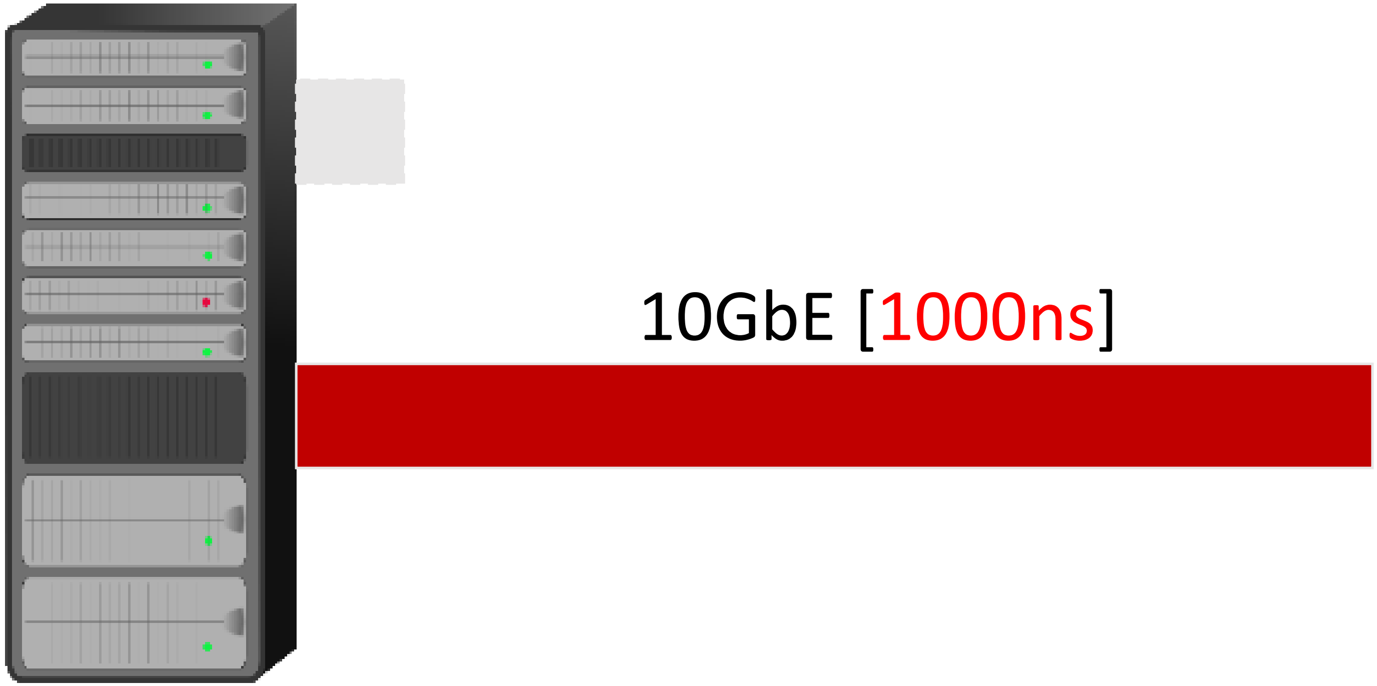


Speed of Light Propagation (100ns)

1GbE [10000ns] (MTU)

Interpacket Gap Decreased Drastically

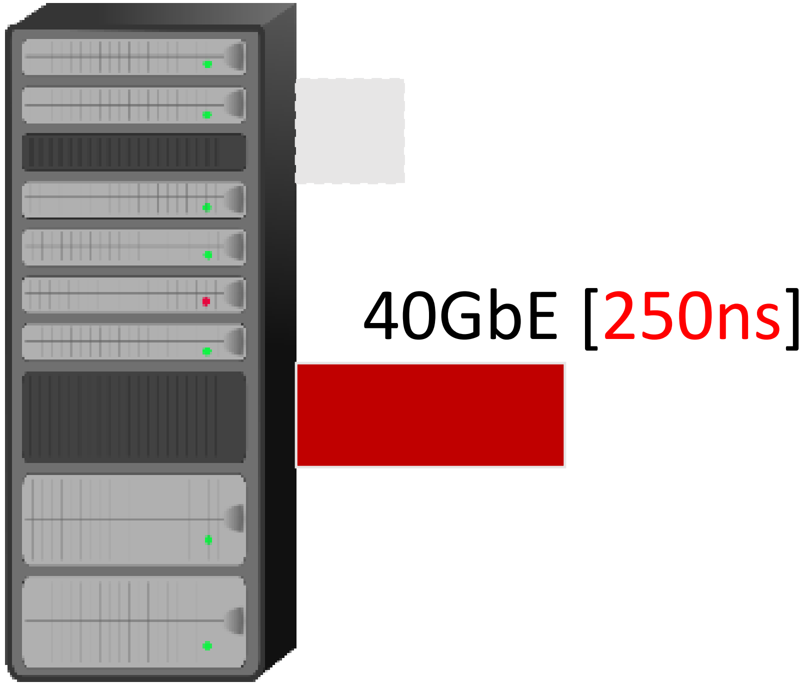
Interpacket Gap Decreased Drastically



10GbE [1000ns]

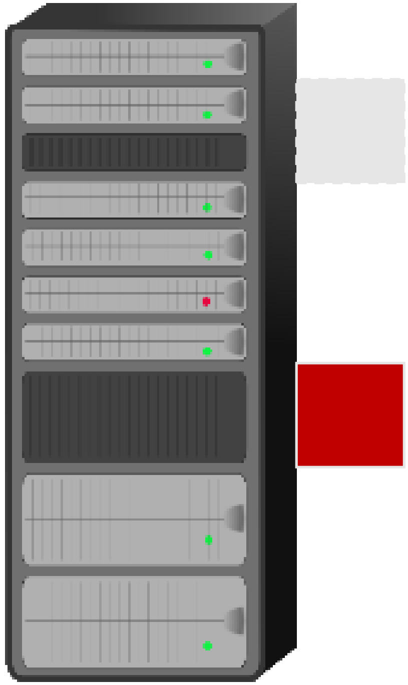
Interpacket Gap Decreased Drastically

Interpacket Gap Decreased Drastically

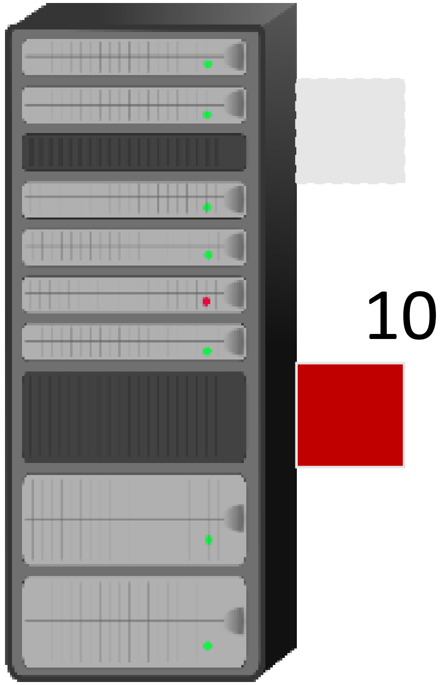


40GbE [250ns]

Interpacket Gap Decreased Drastically

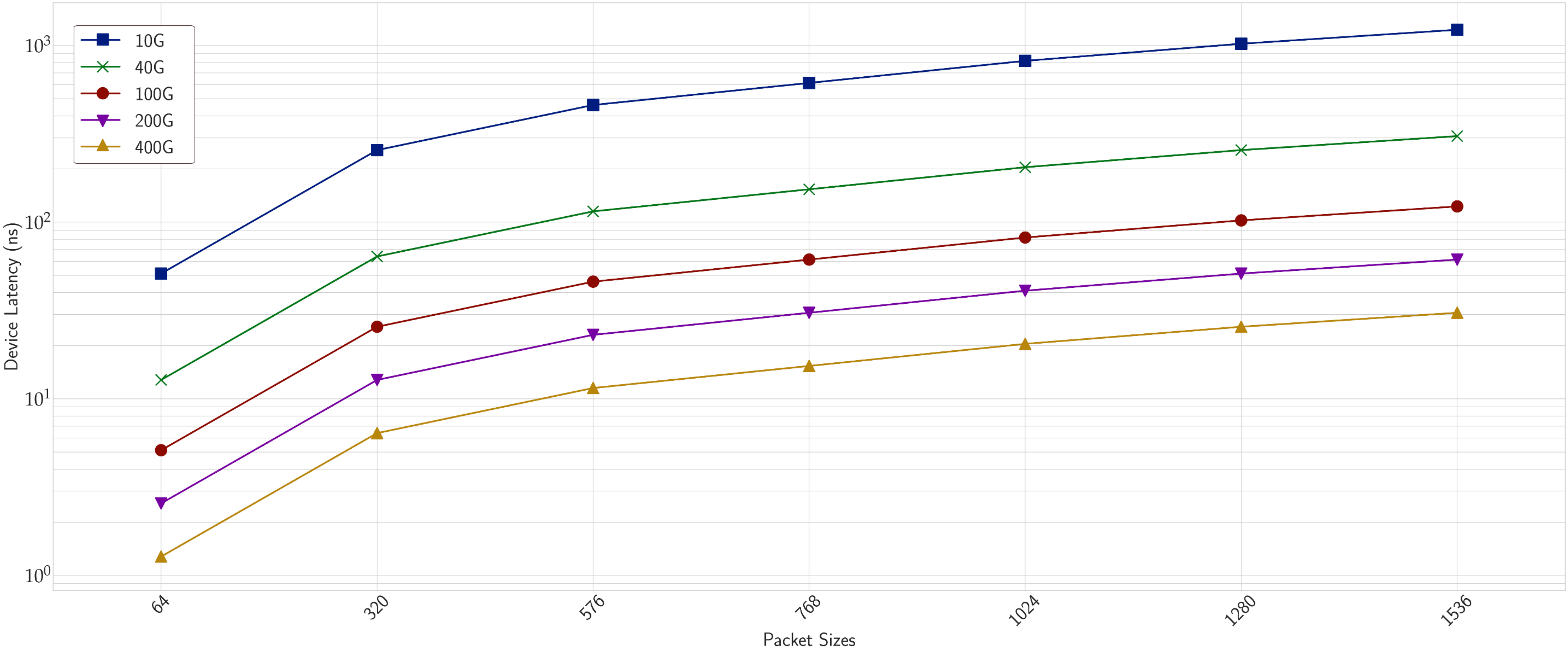


Interpacket Gap Decreased Drastically

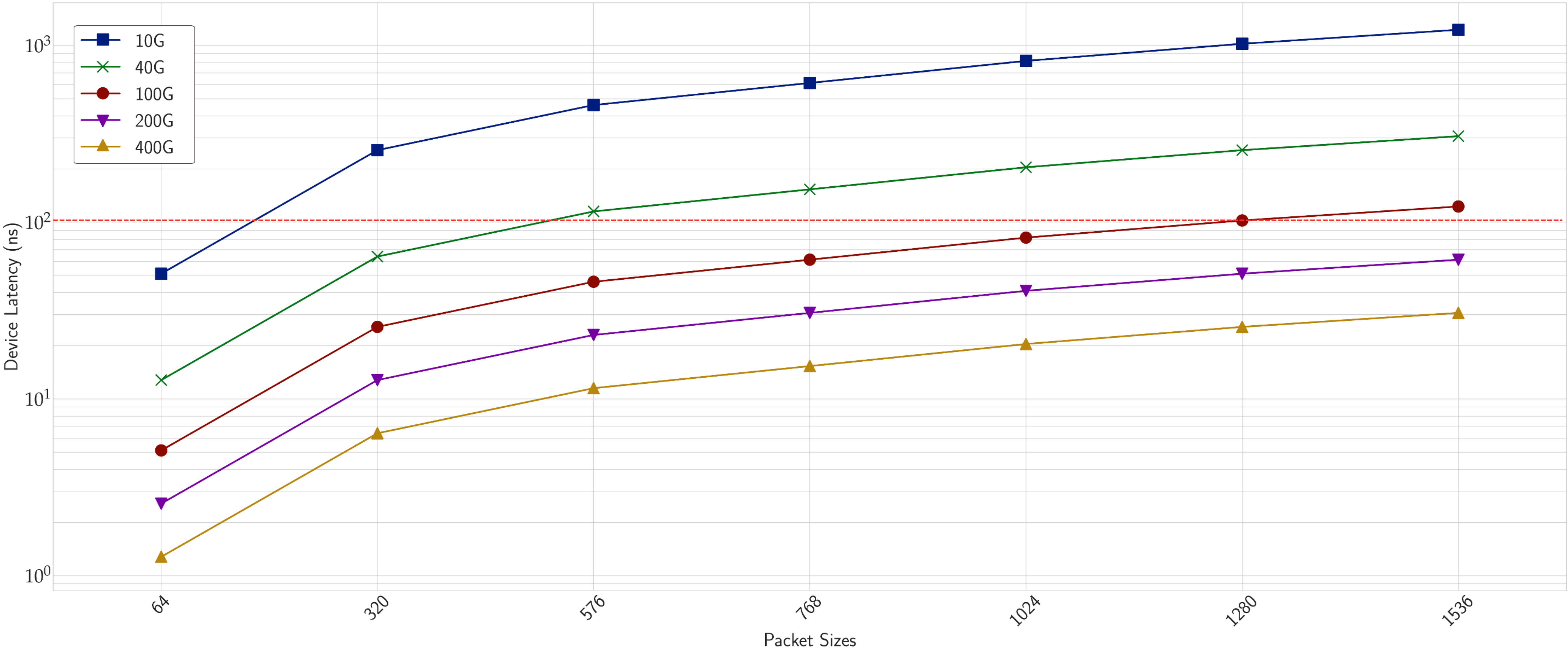


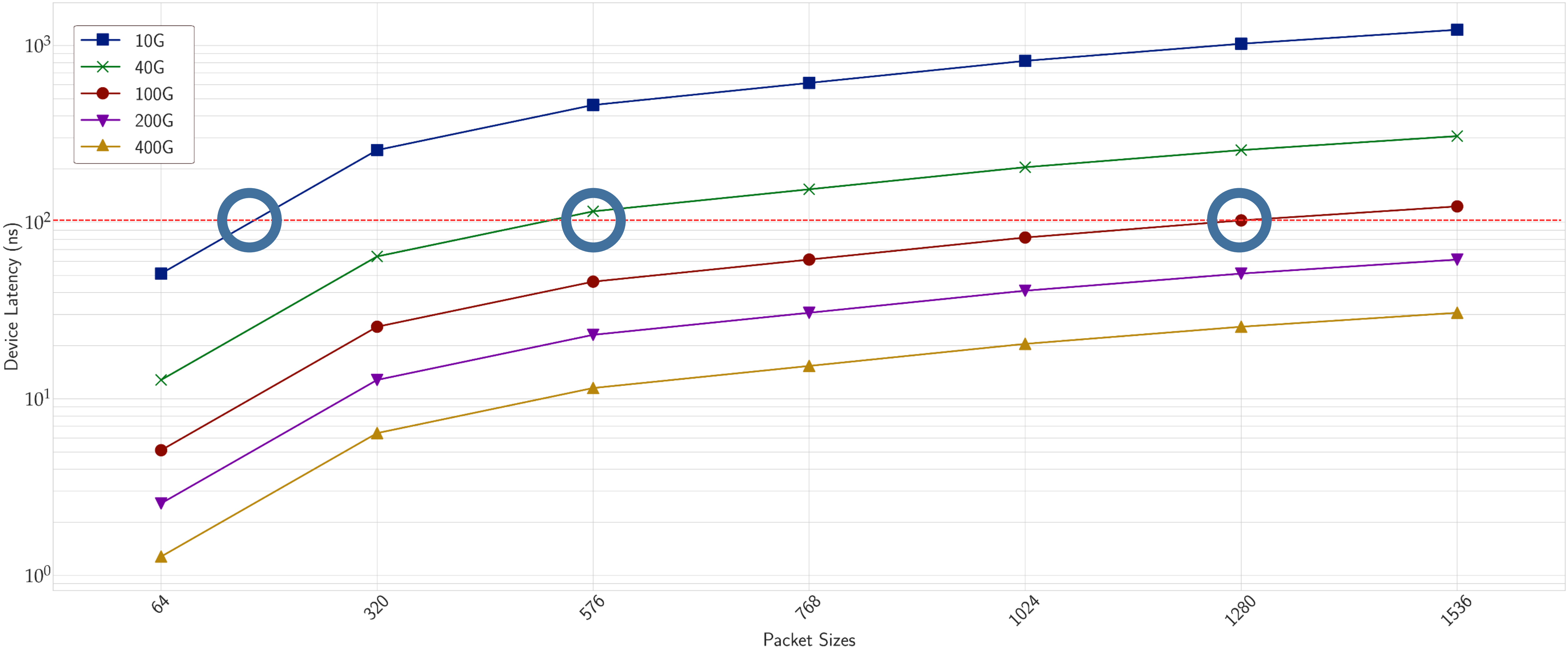
100GbE [100ns]

DRAM Latency?



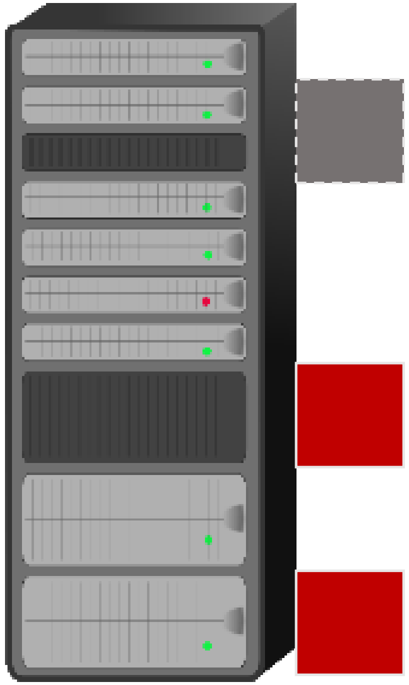
DRAM Latency ~ 100ns





Interpacket Gap < DRAM Latency

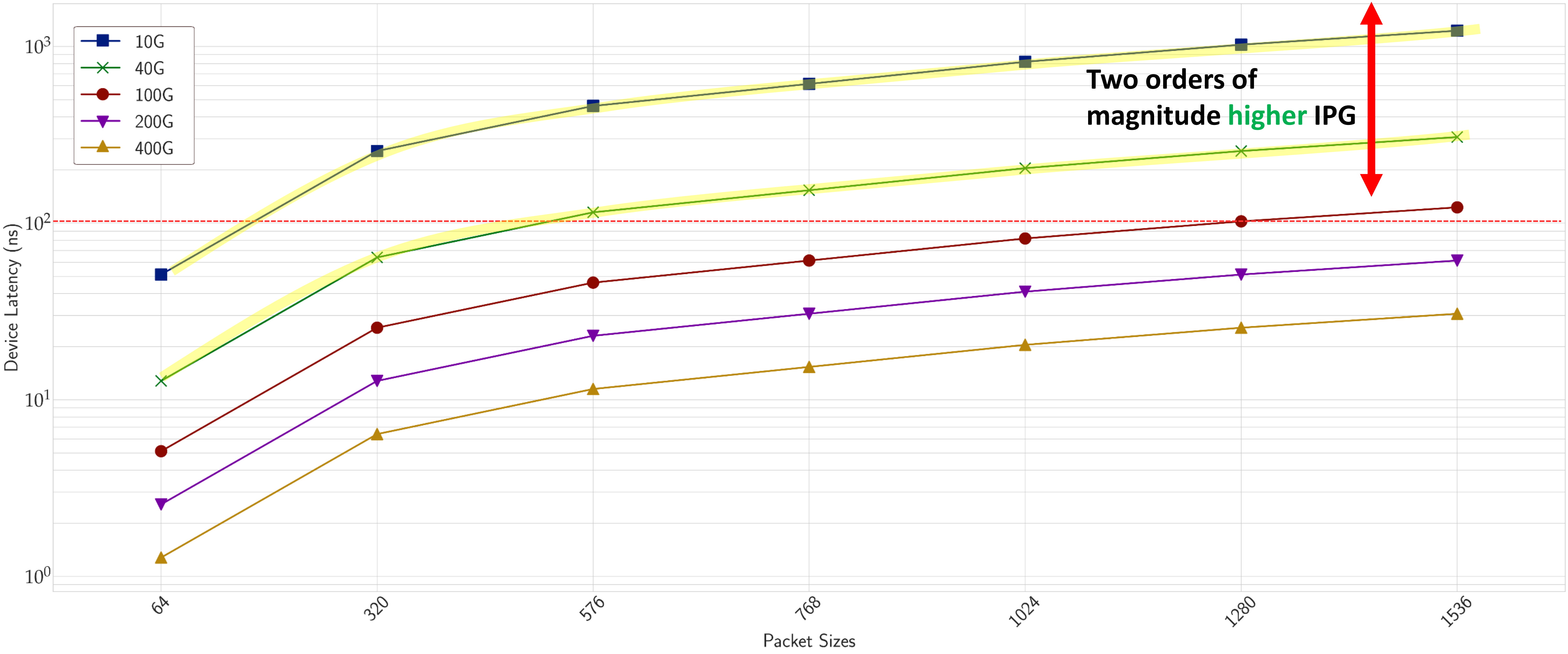
Latency Convergence



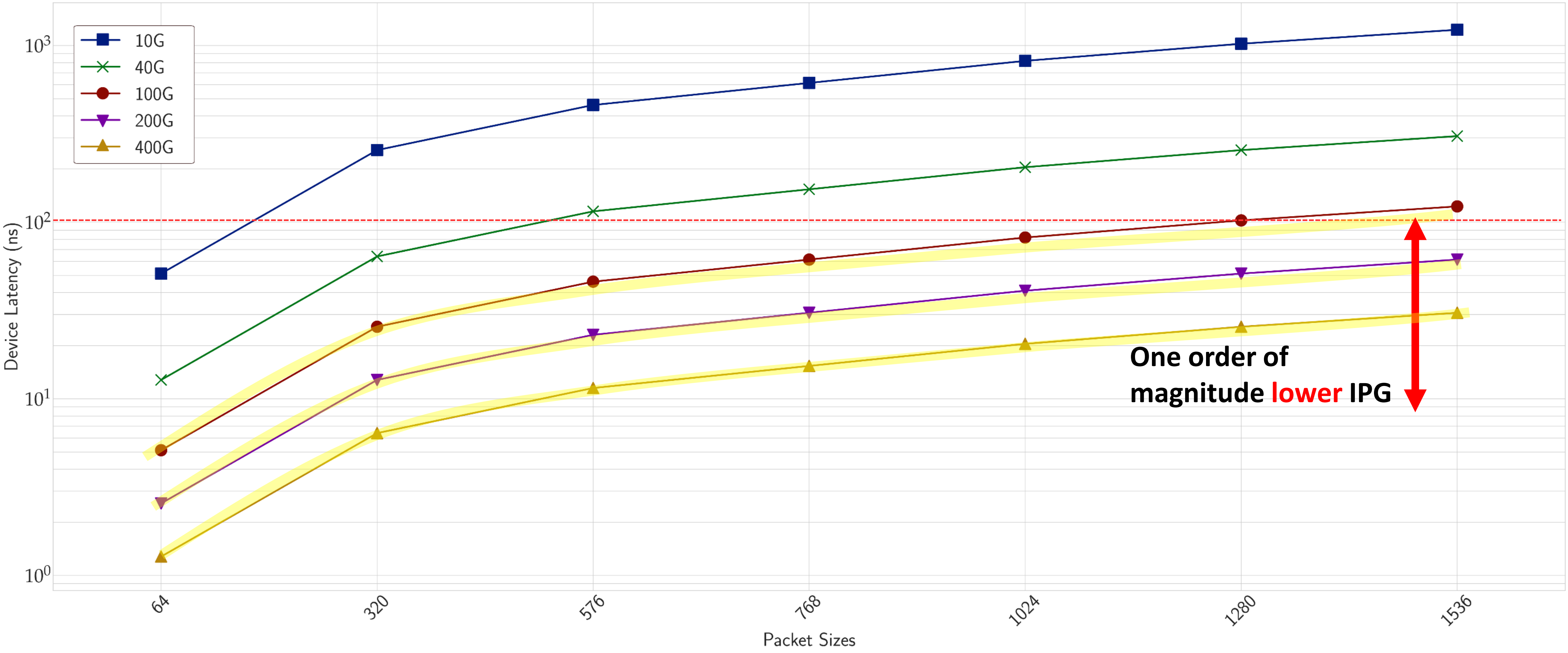
Speed of Light Propagation [100ns]

40GbE/100GbE Interpacket gap [100ns]

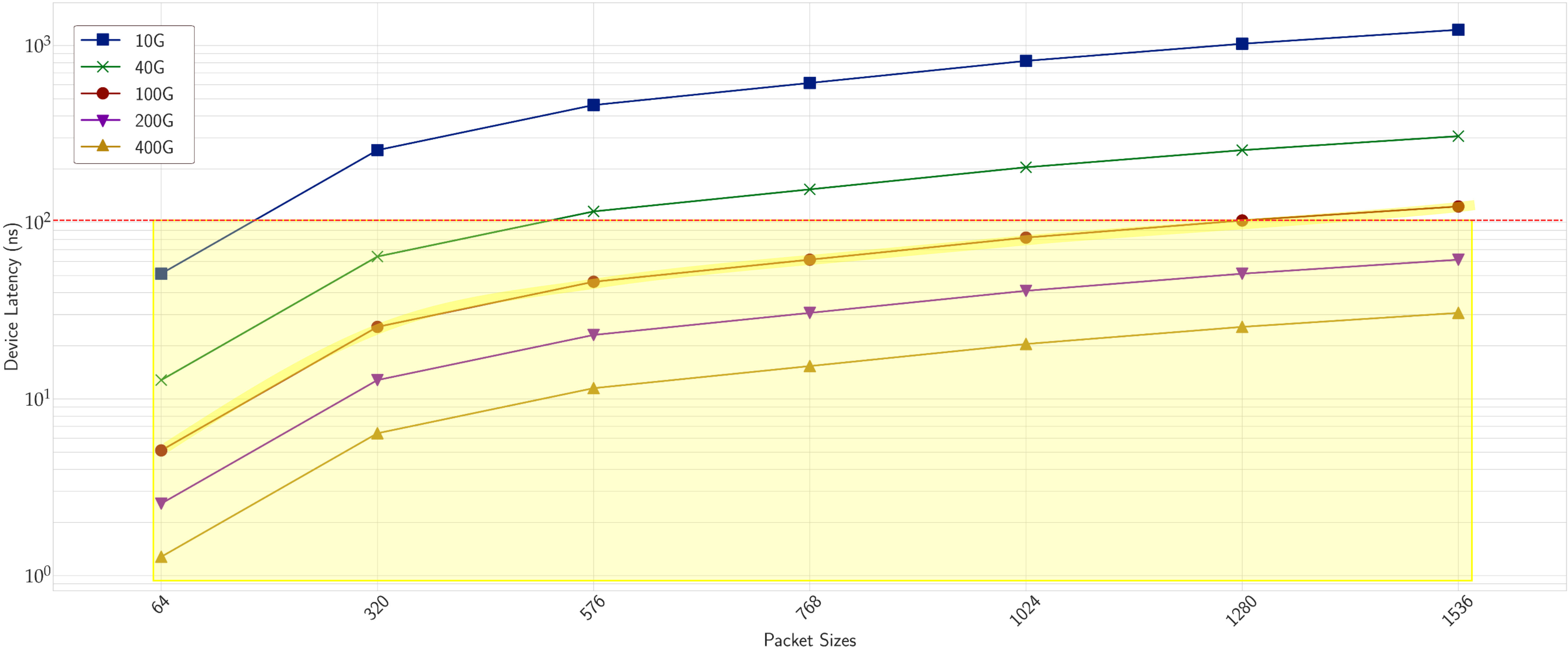
DRAM Latency [100ns+]



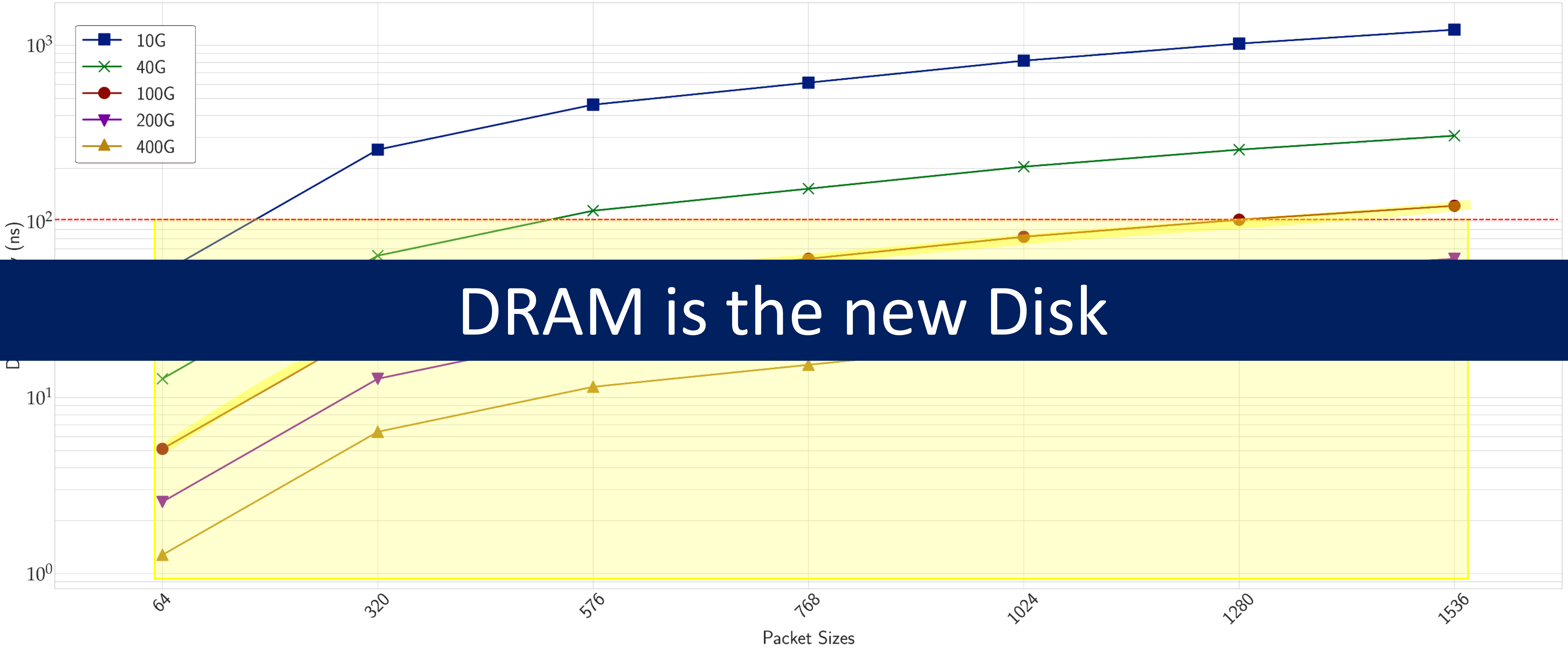
Two orders of magnitude higher IPG



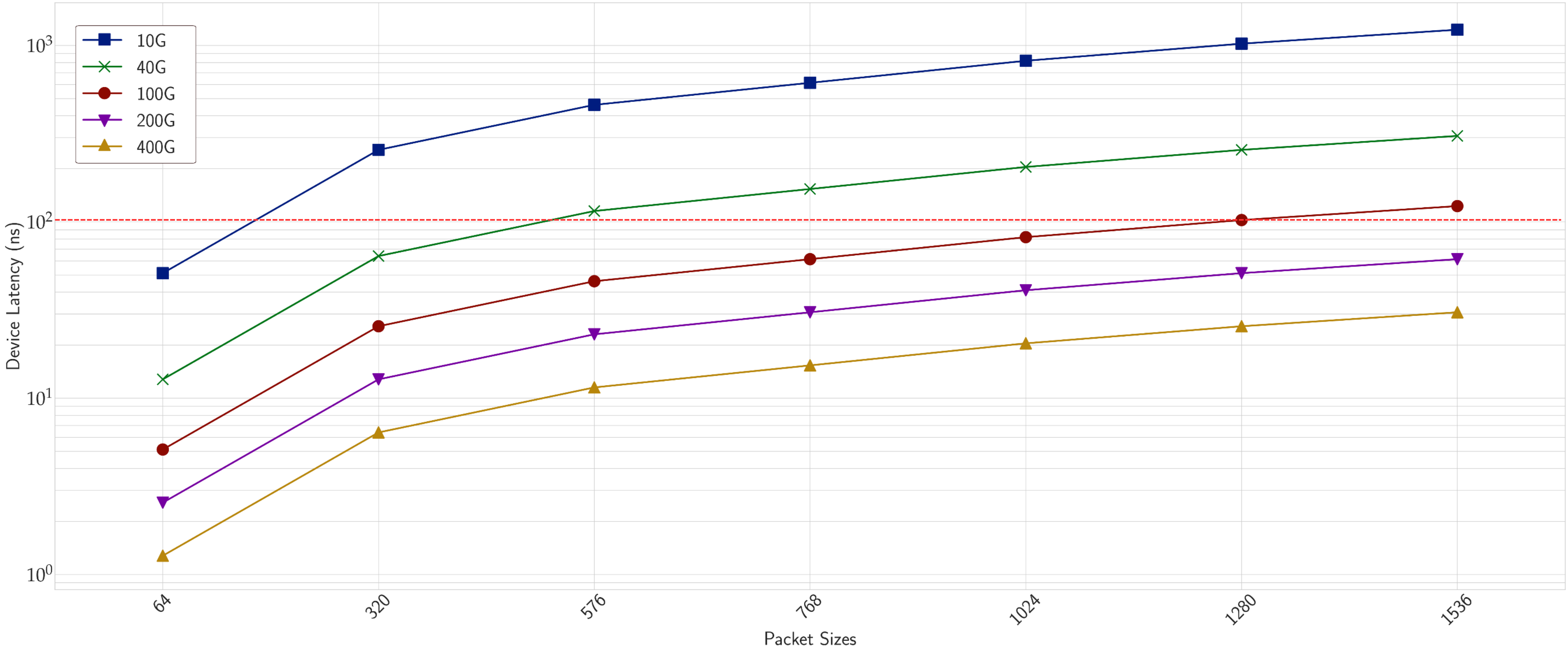
One order of magnitude lower IPG



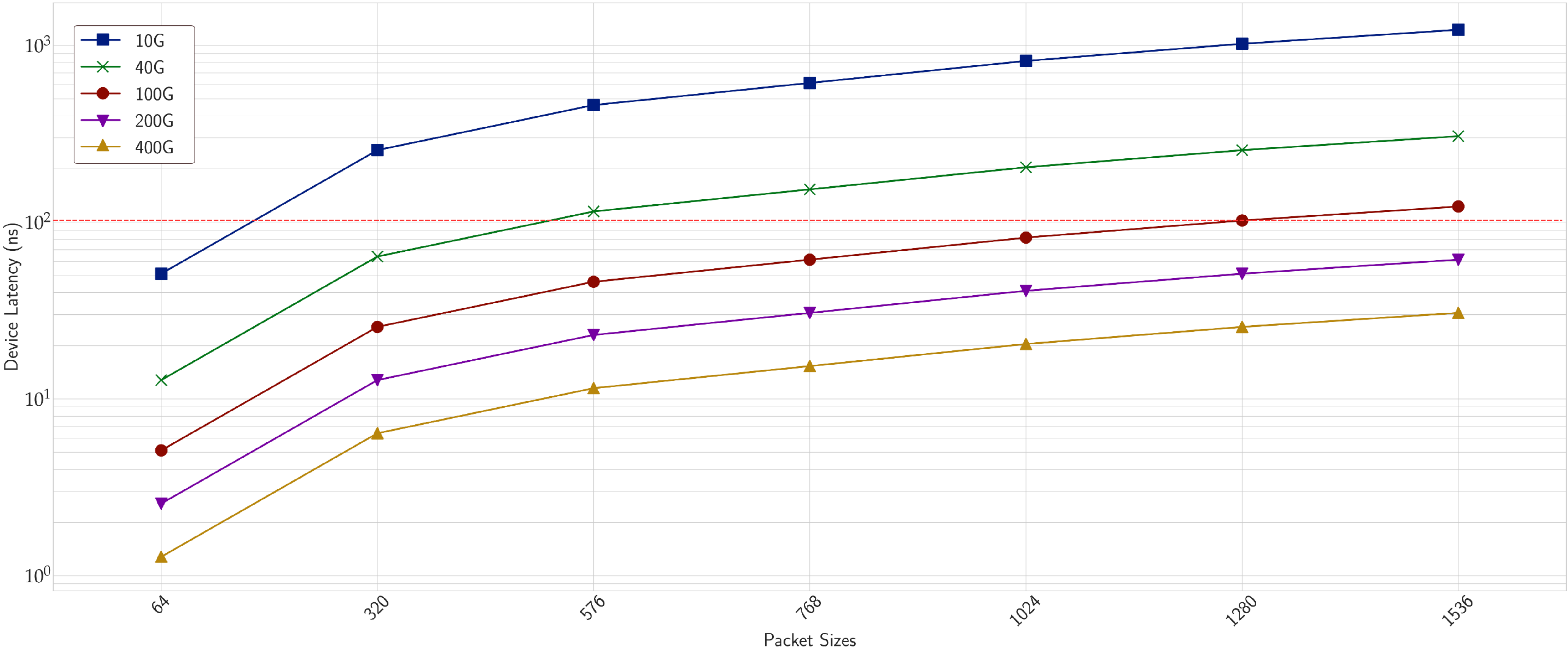
DRAM accesses need to be minimized or eliminated



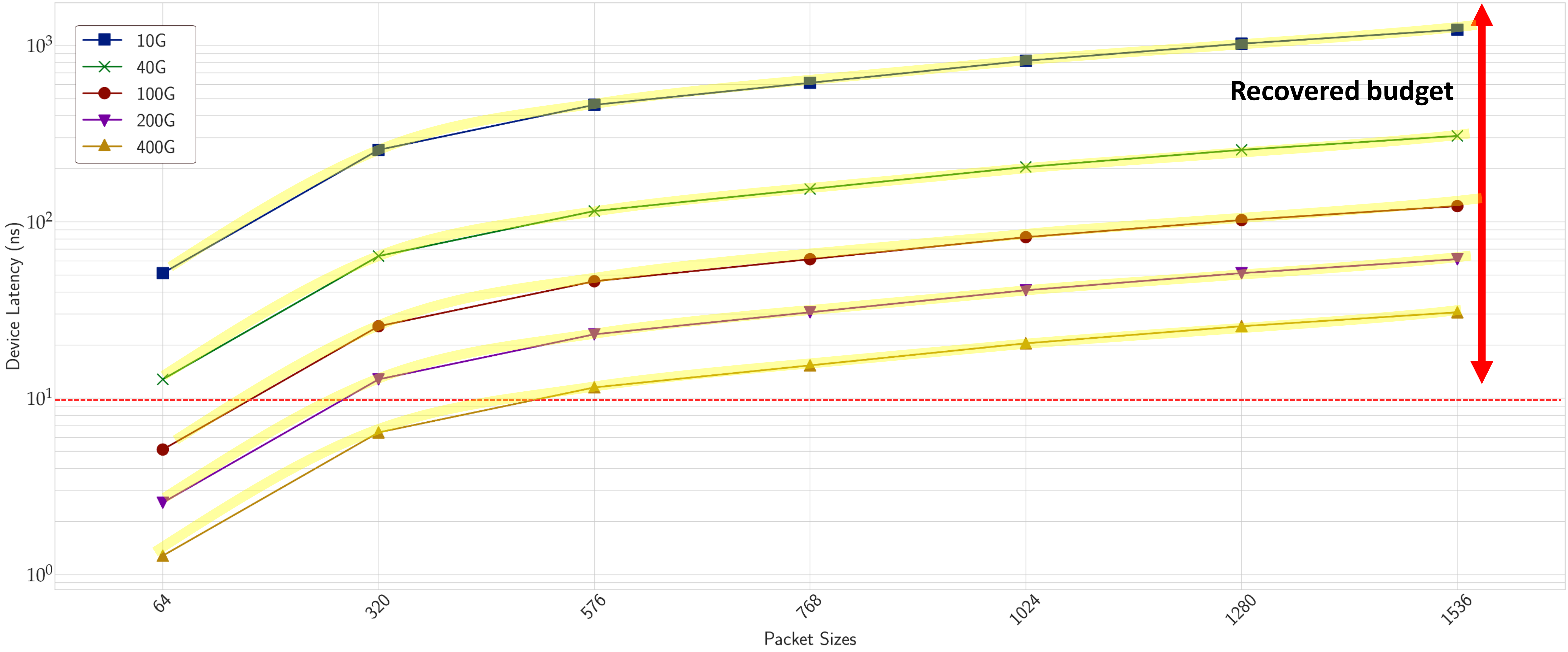
DRAM is the new Disk



How do we overcome DRAM limitation?

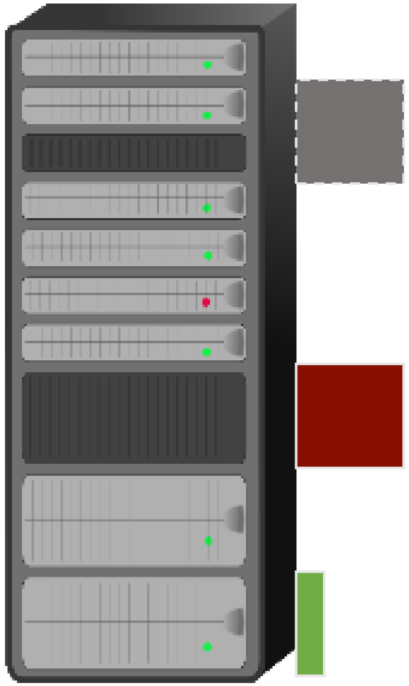


How do we overcome DRAM limitation? → **Pivot to SRAM**



How do we overcome DRAM limitation? → **Pivot to SRAM**

Faster than Light SRAM

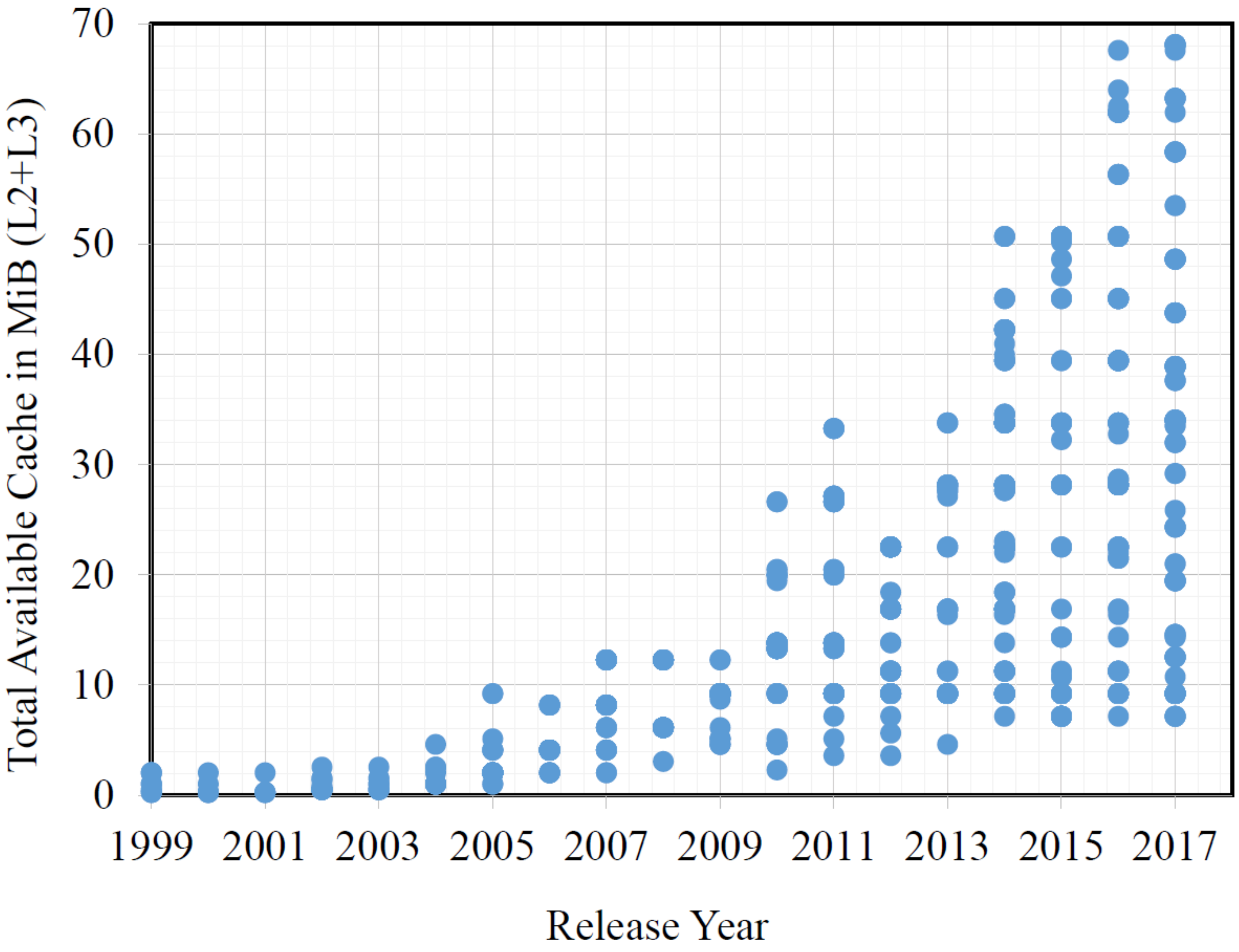


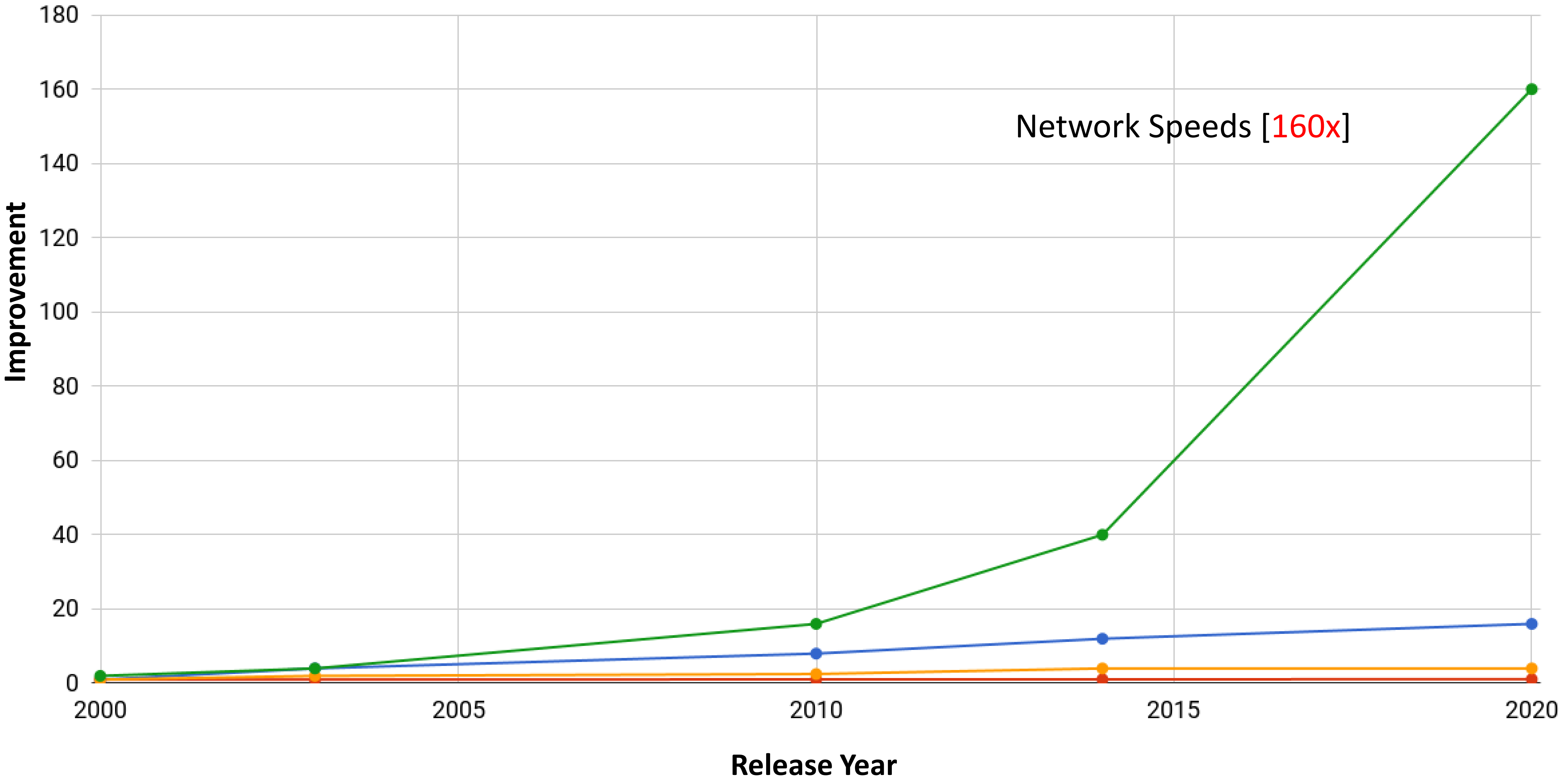
Speed of Light Propagation [100ns]

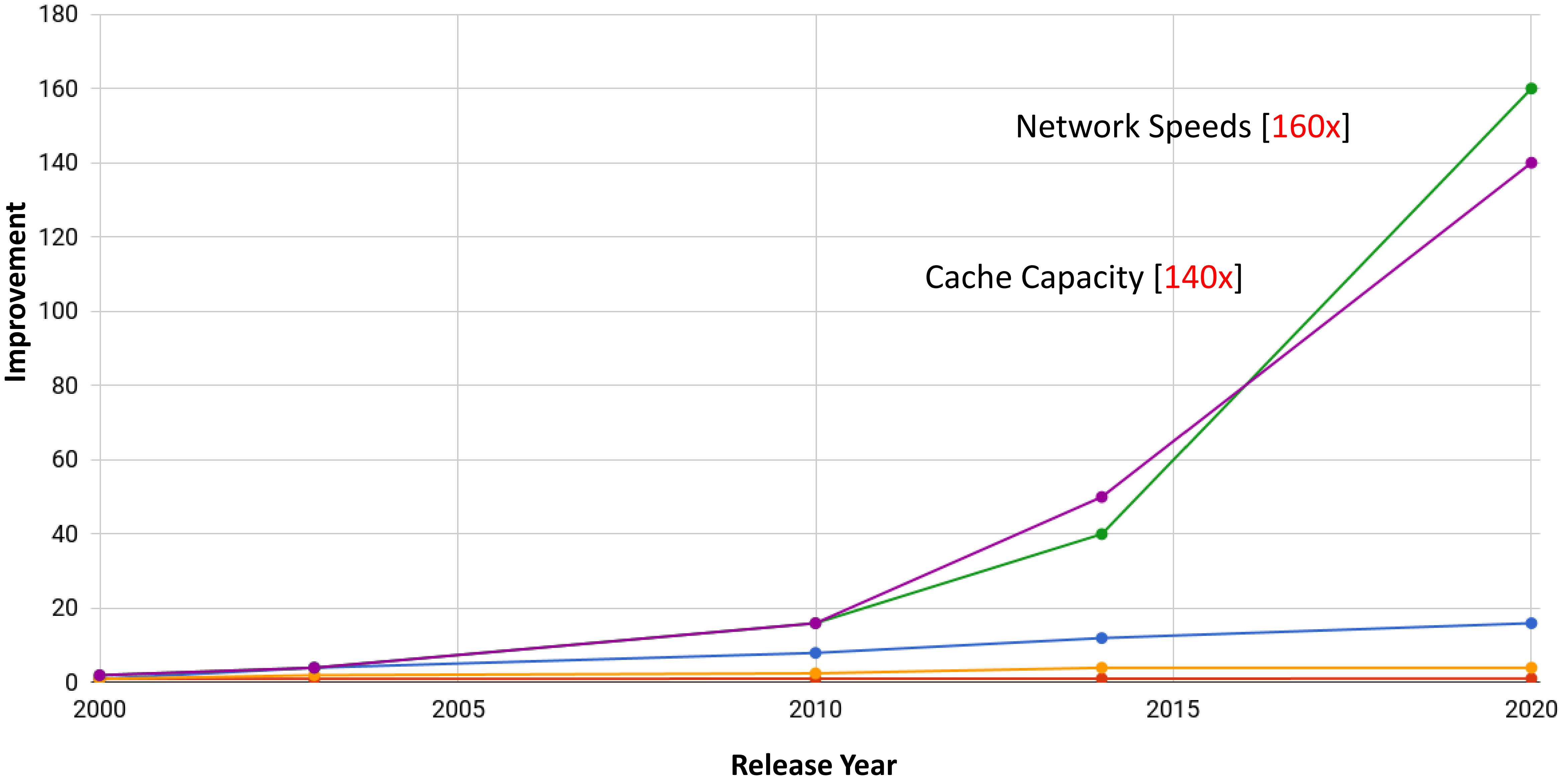
100GbE Interpacket gap [100ns]

SRAM Latency [10ns]

SRAM Capacity Continues to Scale



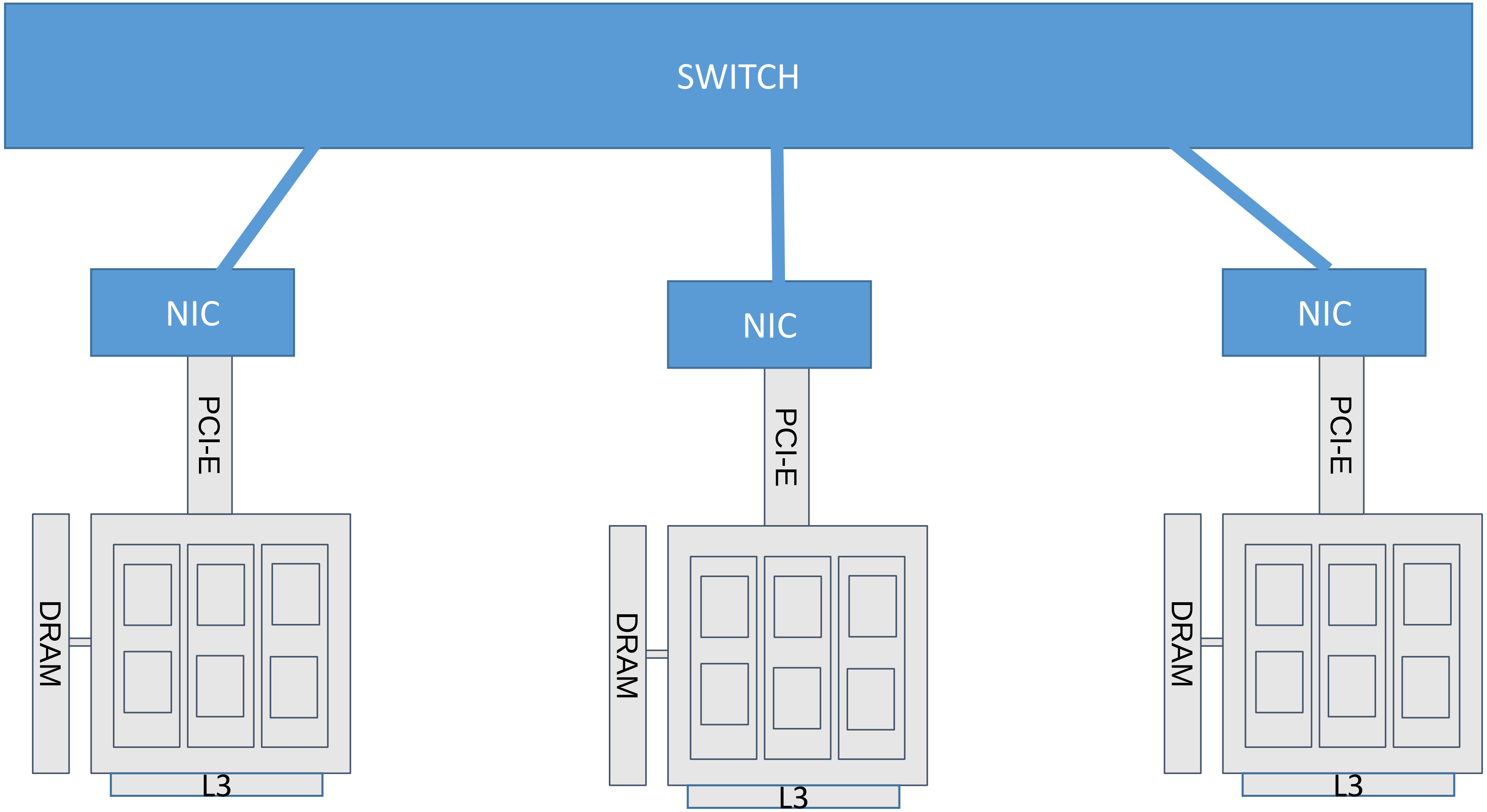

Network Speeds [160x]
Improvement
180
160
140
120
100
80
60
40
20
0
2000
2005
2010
2015
2020
Release Year



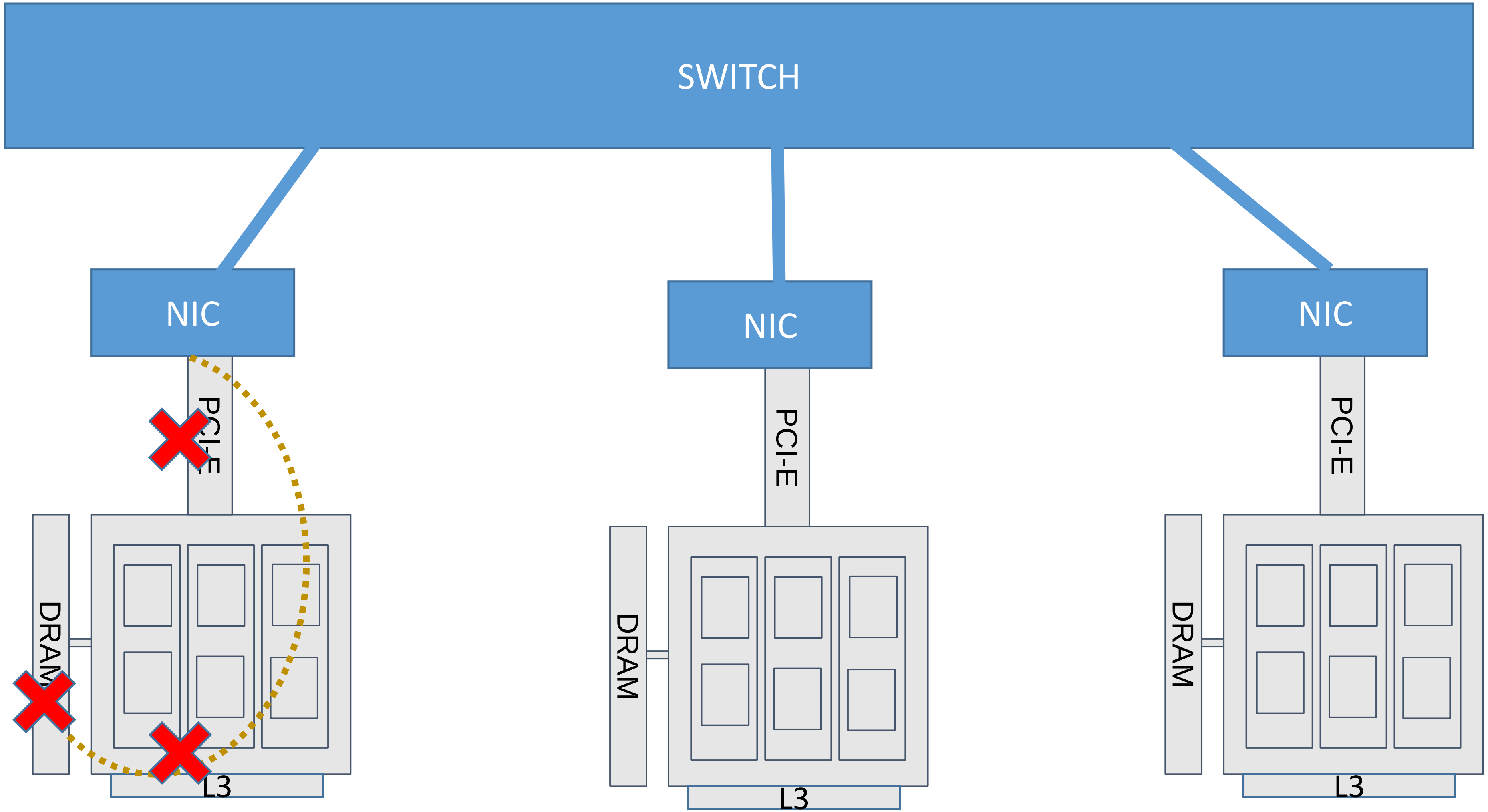
Network Speeds [160x]

Cache Capacity [140x]

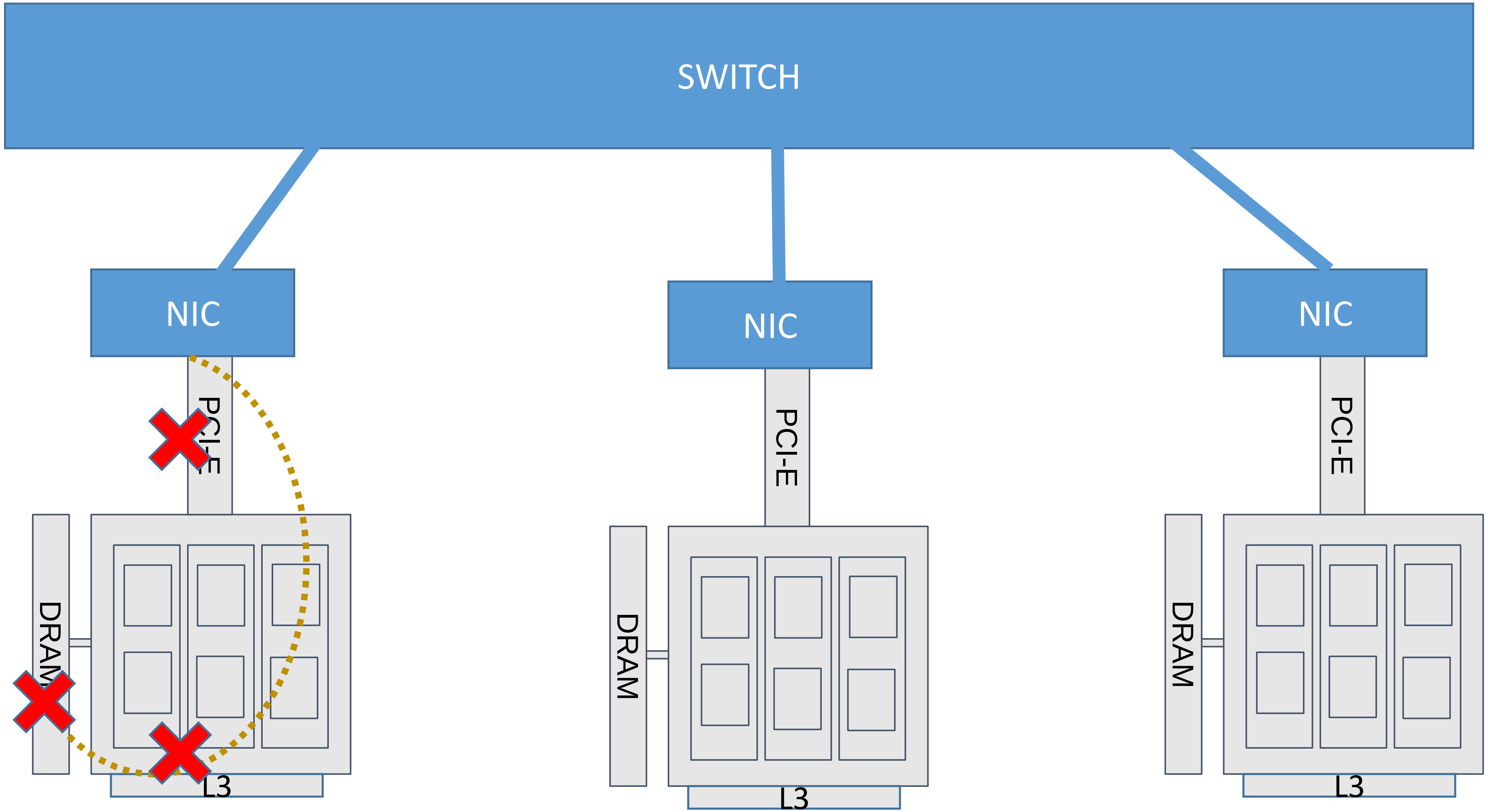
Conventional System Model



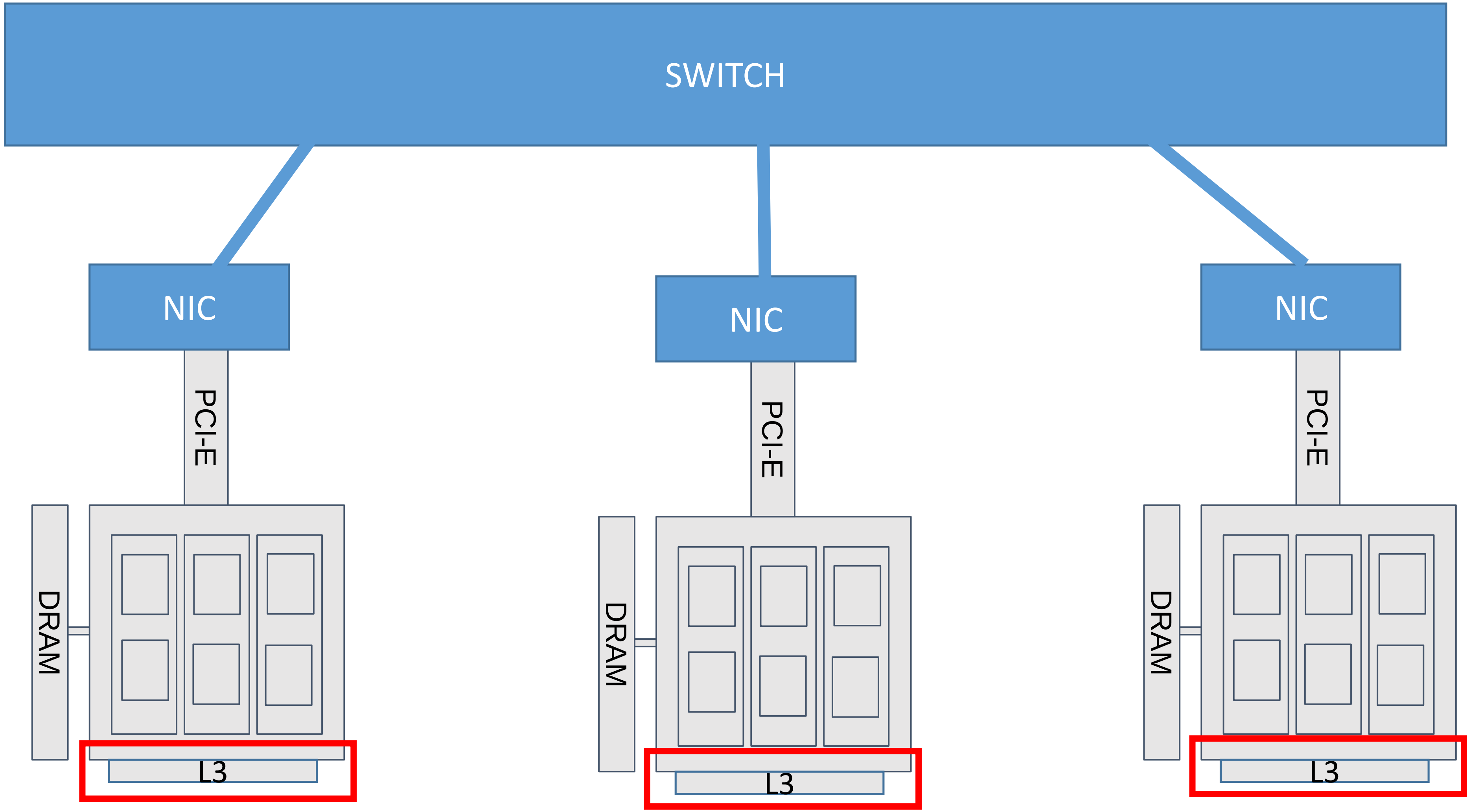
Missing is Expensive



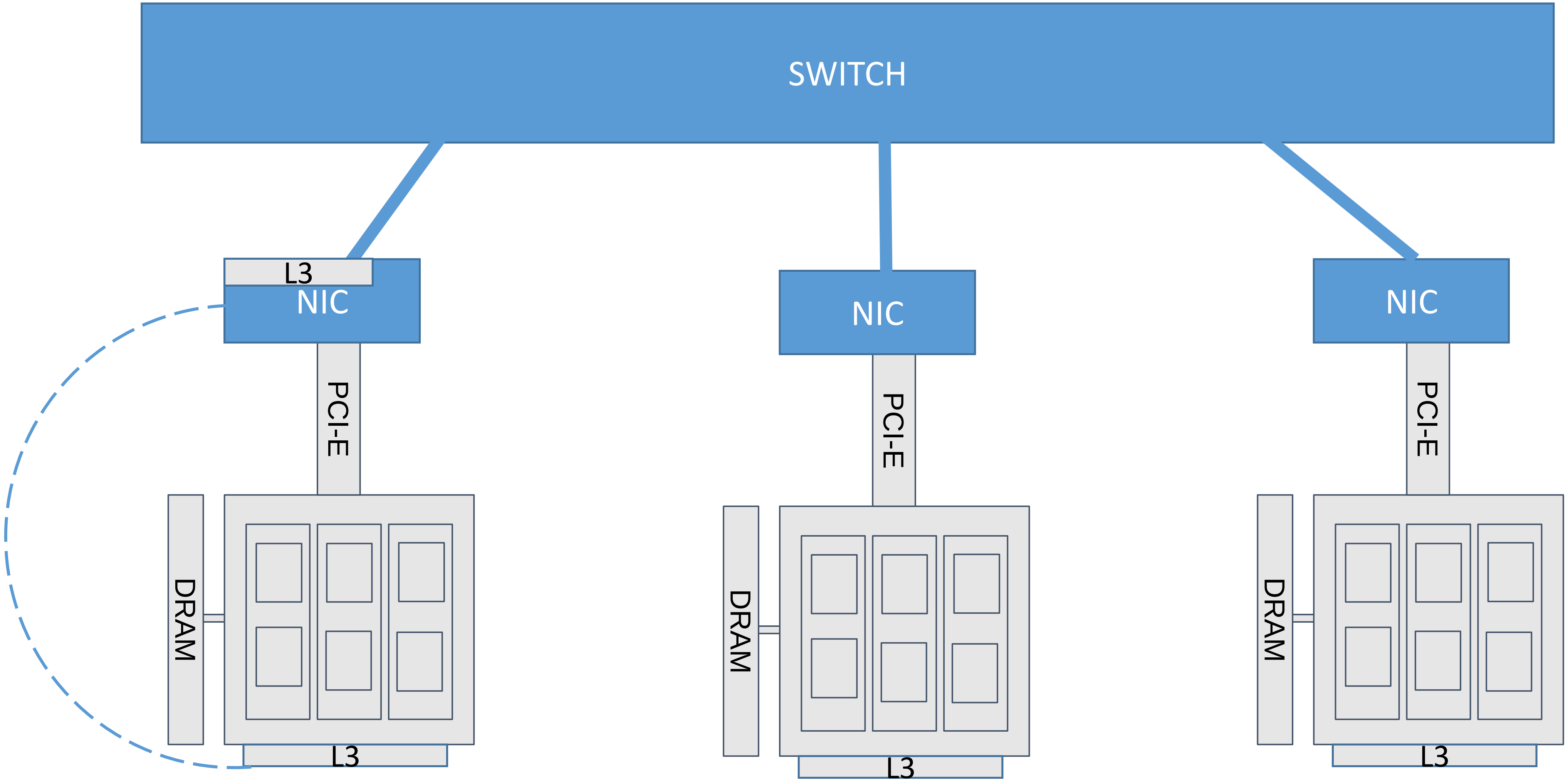
Can we Predict L3 Misses *Before They Happen*?



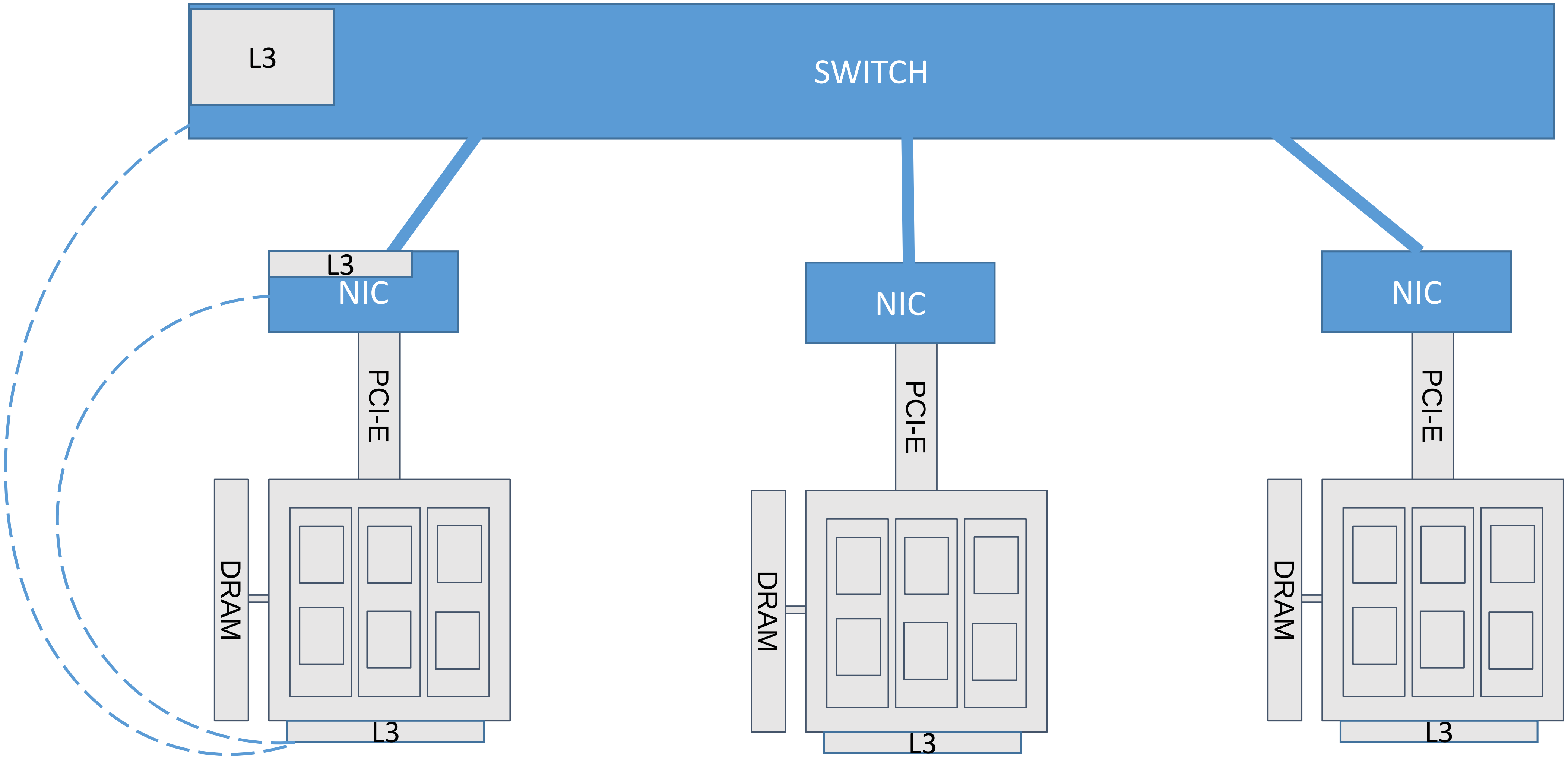
Idea: Global View of L3 Caches Across Cluster



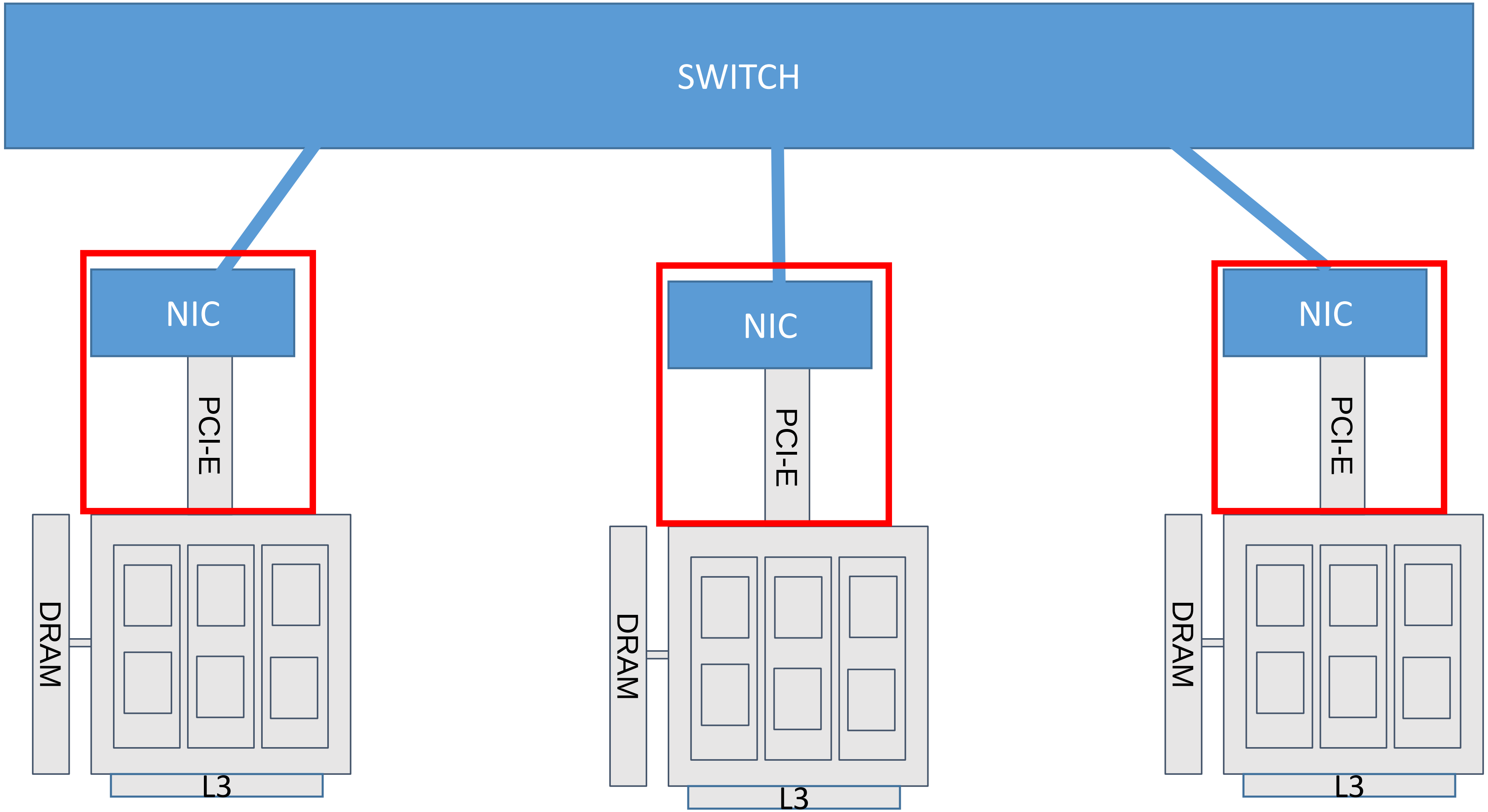
Closer Integration with Networks and Caches



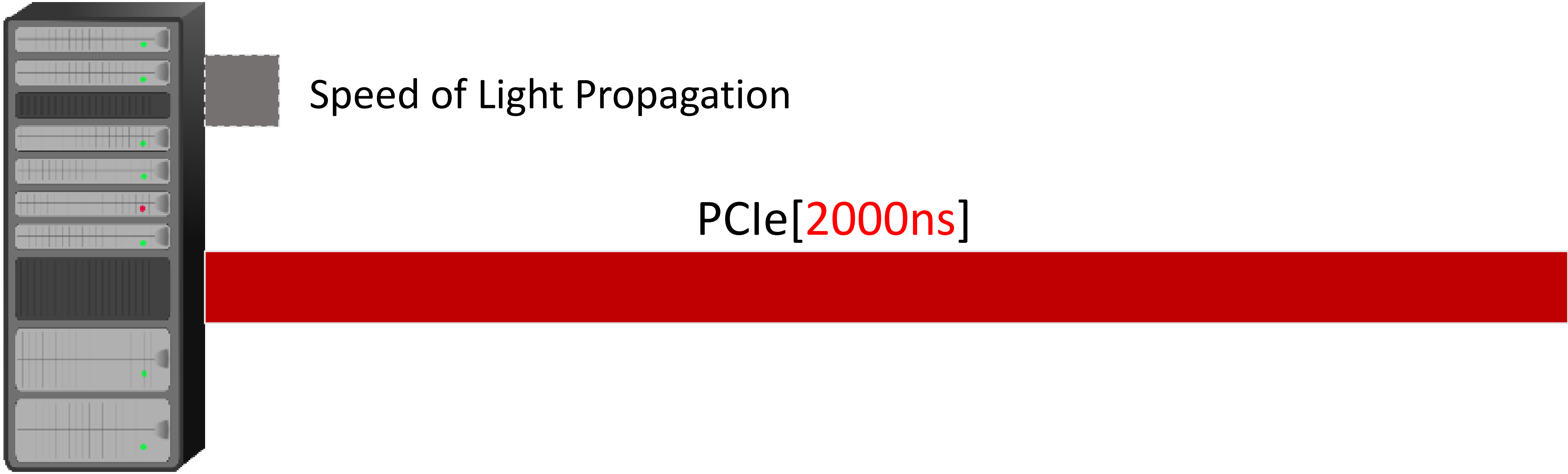
Reroute Packets Based on In-Network Data



Improving Interconnects



Interconnects are Significant Bottlenecks

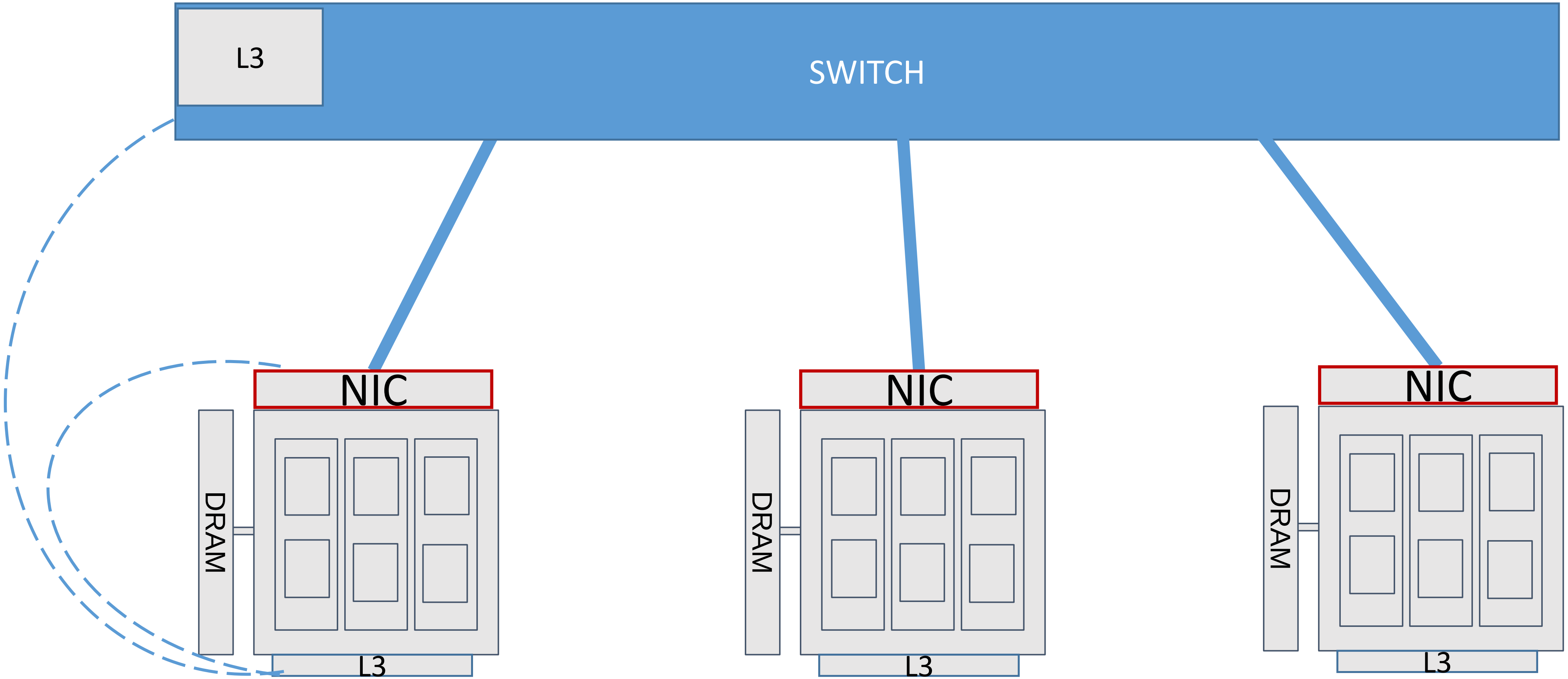


Speed of Light Propagation

PCIe[2000ns]

CacheCloud

Co-Packaging Network Interfaces and Cache



CacheCloud: Takeaways

We are more than an **order of magnitude away** from speed of light propagation

Most hardware components **have stopped scaling** while networks **scale up**

Cache is the new DRAM, DRAM is the new Disk

Architecture, Network and OS **integration** is crucial

Issues and Feedback for CacheCloud

Q: Won’t this be solved with new hardware?

- SmartNICs, Accelerators, FPGAs, etc.
- Coherency, programmability, size, and cost still a problem

Q: Datasets are too large….but do all applications have large dataset?

- Network function virtualization (NFV)
- Coordination services

Feedback:

- How do hardware accelerators (TPU) change network and end-host design?
- Starting from scratch what architecture should we build?
- Starting from scratch what network protocols should we build?

Thank You

UC San Diego

Computer Science and Engineering

NSF