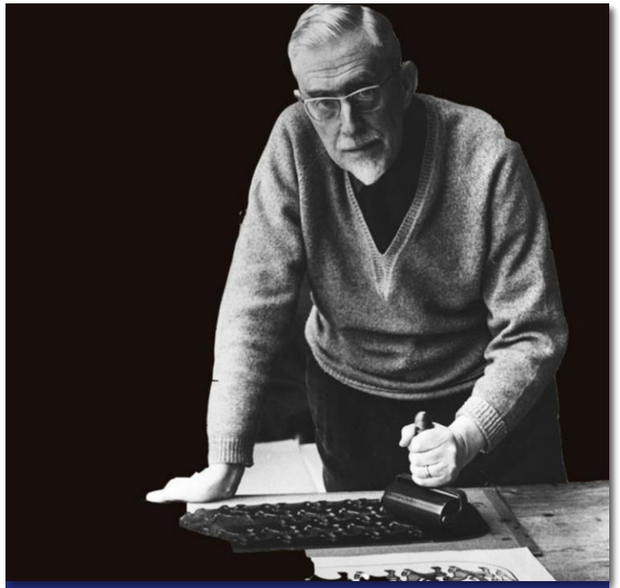
Purpose Built High-Performance Fabrics for HPC/HPDA/AI

ExaComm 2021
Philip Murphy
July 2, 2021

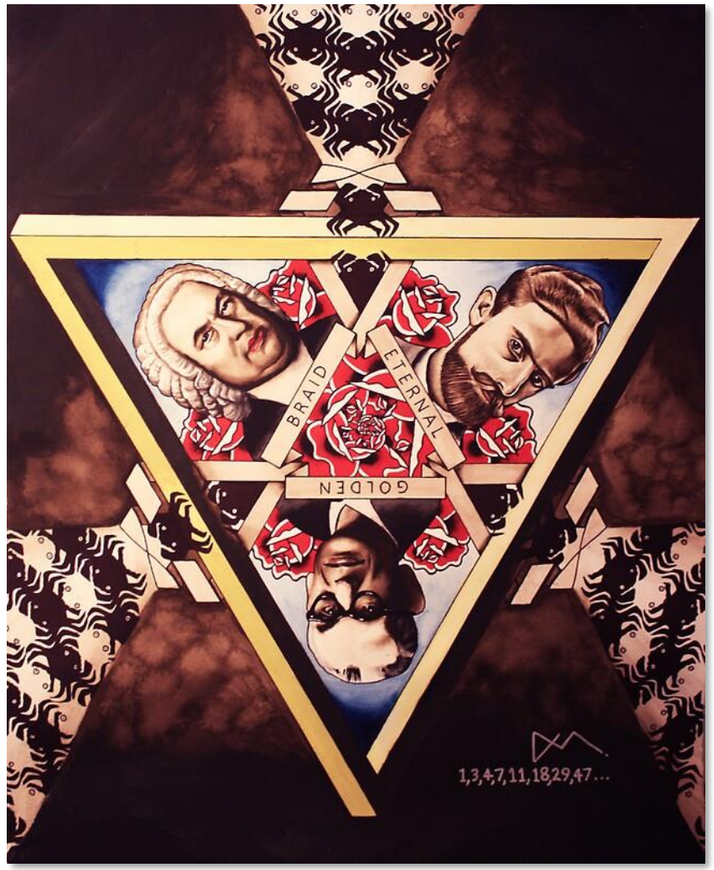
CORNELIS™ NETWORKS

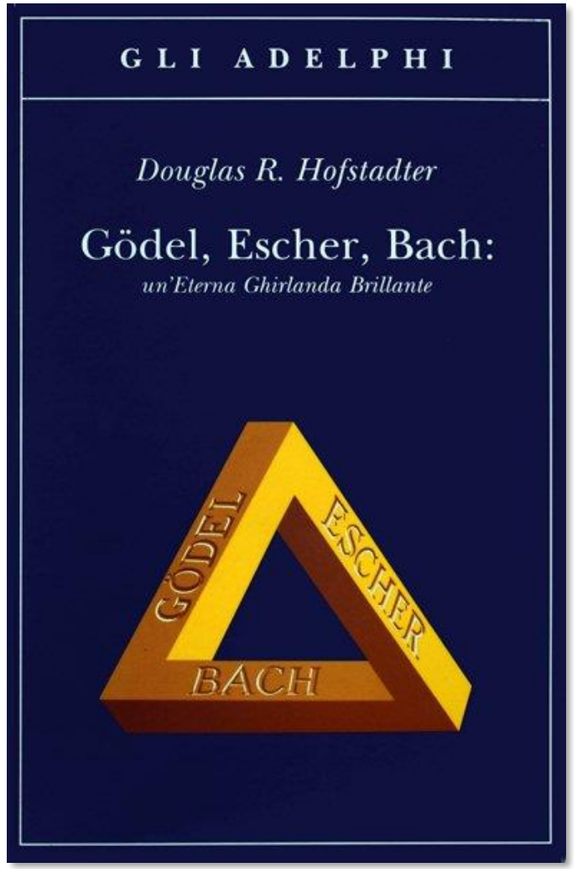
Cornelis Networks…

Name inspired by a book by Douglas R. Hofstadter: *Gödel, Escher, Bach: An Eternal Golden Braid (*1979), an early AI work illuminating the foundations of intelligence



Maurits Cornelis Escher
Dutch graphic artist
who made mathematically
inspired art
1898-1972



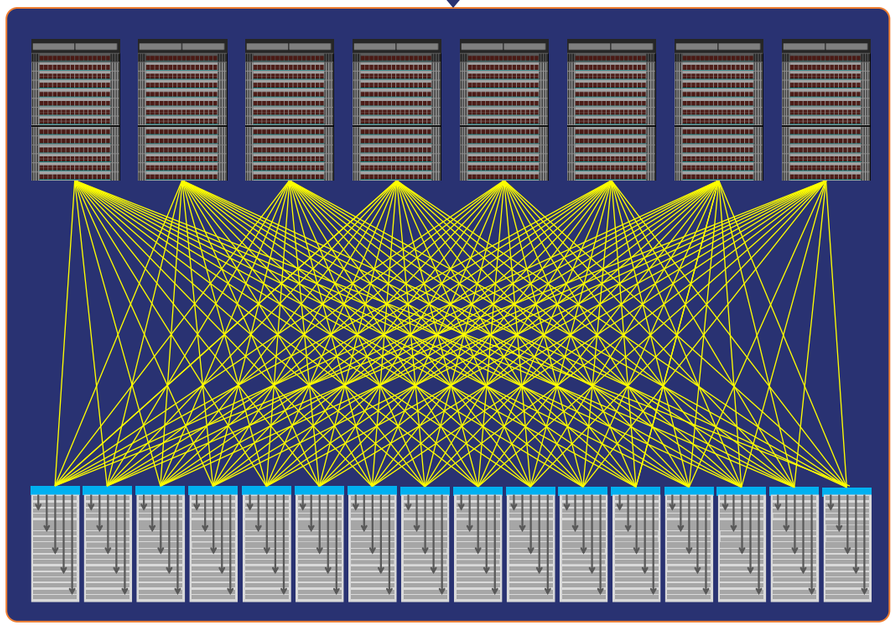


Interconnect Driven by Data-Centric Convergence

High Performance Computing

High Performance Data Analytics

Artificial Intelligence



Fabric essentials for application performance and scale

- Low scalable latency
- Extreme message rate
- Highly scalable bandwidth
- Outstanding collectives perf
- Optimized data flow

Cornelis™ Omni-Path™ enables convergence

- Exceptional latency and message rates
- Incrementally scalable bandwidth
- Balanced host / network functionality
- High performance collectives
- Optimized for HPC, analytics, and AI frameworks

Cornelis Omni-Path Fabrics Roadmap

- Continued commitment to customers and **HPC/HPDA/AI** ecosystem
- Delivering industry leading **application performance** and unmatched **price/performance**
- Supporting a wide **variety of industry CPUs and accelerators** to ensure solution choice

Today

CN-OPA100
PCIeV3.0/4.0 adapters
48p Edge Switches
288P & 1152p Directors

Q4'21
Omni-Path Express
OFA/OFI

Q4'22/Q1'23

CN-OPA400*
PCIeV5.0
High-density Edge Switches
High-density Directors

2025

CN-OPA800*
PCIeV6.0
High-density Edge Switches
High-density Directors

Next…

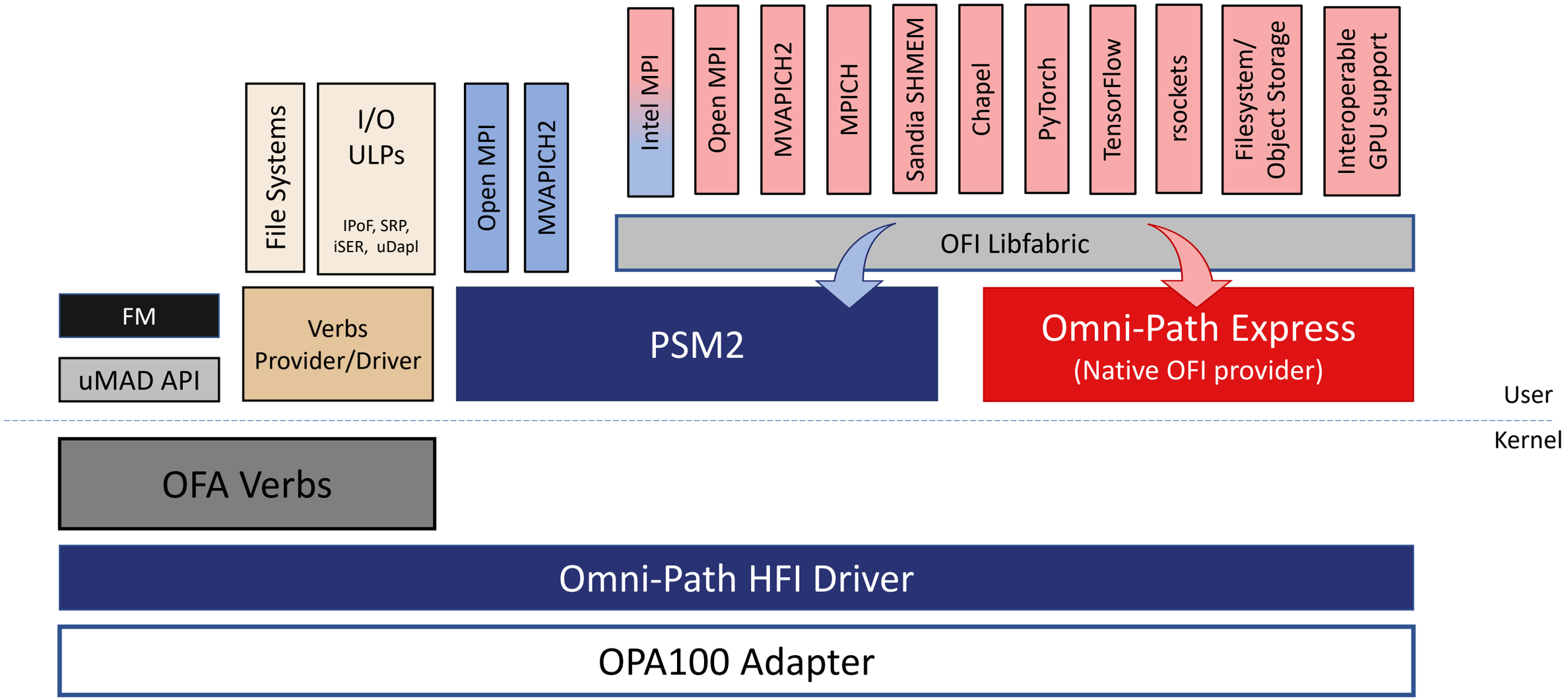
800 Gbps
400 Gbps
400 Gbps
200 Gbps
200 Gbps
100 Gbps

Increasing Bandwidth and Network Intelligence to target HPC, AI and HPDA workloads

*Future features/options are subject to change without notice.

Omni-Path Express Software Stack – OFI Libfabrics

OFI Libfabric support provides access to significant industry contributions across all key areas.



Portfolio Overview

End-to-end fabric solution driving industry leading performance and TCO

Host and Management Software	Standard and Custom Adapters	Standard and Custom Edge Switches	Director Class Switches	Copper and Active Optical Cables	InfiniBand & Ethernet Gateways
					
					

Omni-Path (now): Complete product portfolio, in production, 500+ sites, single/multi 100Gbps rails

Omni Path Express (2H'21): Dramatic performance and functionality enhancements, seamless upgrade

Omni-Path Next Generation (4Q'22): Enhanced ASICs, acceleration features, single/multi 400Gbps rails

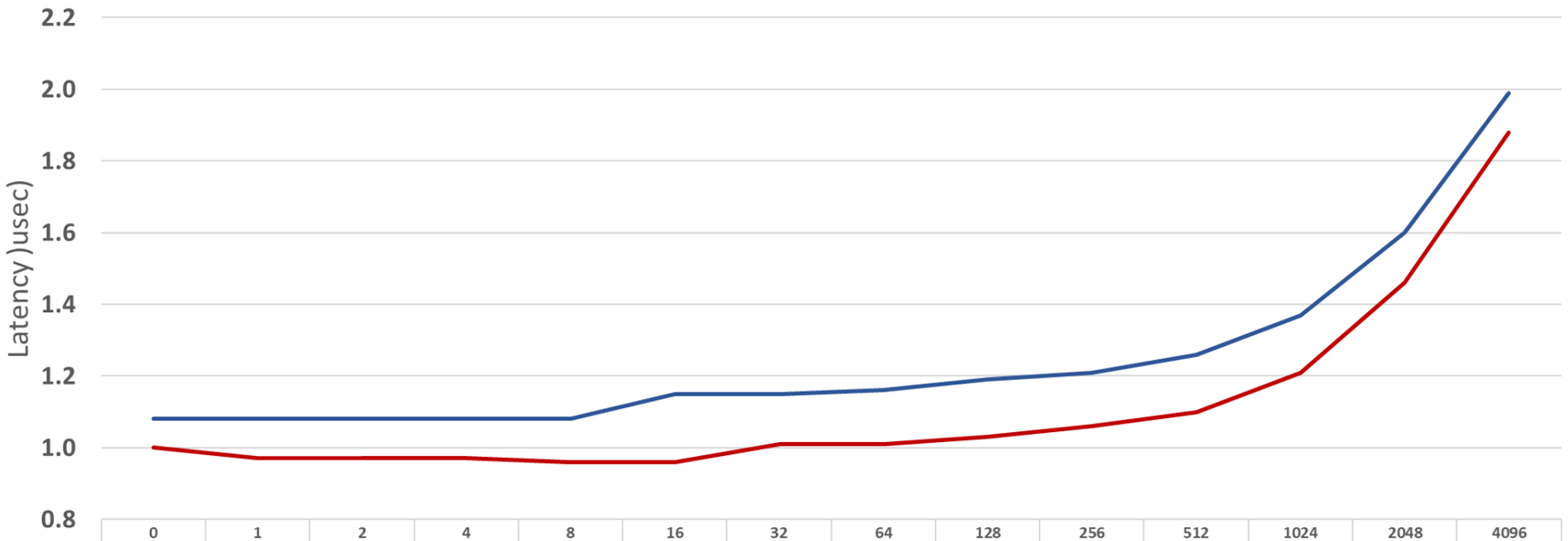
Omni-Path Express

- Next generation native OFI/libfabric provider
  - Architected for performance
  - Premier implementation of OFI
  - Initial development efforts focused on small message **latency**, **software footprint,** and **message rate**

Latency

- Average of 12% Latency Improvements
- MPI latency through a switch hop nearing 900 ns
- Continuing to drive down through leveraging OFI

IMB-MPI1 pingpong Latency
2 Socket Intel Xeon Scalable 8352Y (32c, 2.2Ghz)
IntelMPI 2020.4.304
PSM2 (IFS 10.11.0.0.576) vs OPX (DUAL RAIL)



	0	1	2	4	8	16	32	64	128	256	512	1024	2048	4096
1PPN PSM2	1.03	1.03	1.03	1.03	1.03	1.10	1.10	1.11	1.14	1.16	1.21	1.32	1.55	1.94
1PPN OPX	0.95	0.92	0.92	0.92	0.91	0.91	0.96	0.96	0.98	1.01	1.05	1.16	1.41	1.83

Message Size (bytes)

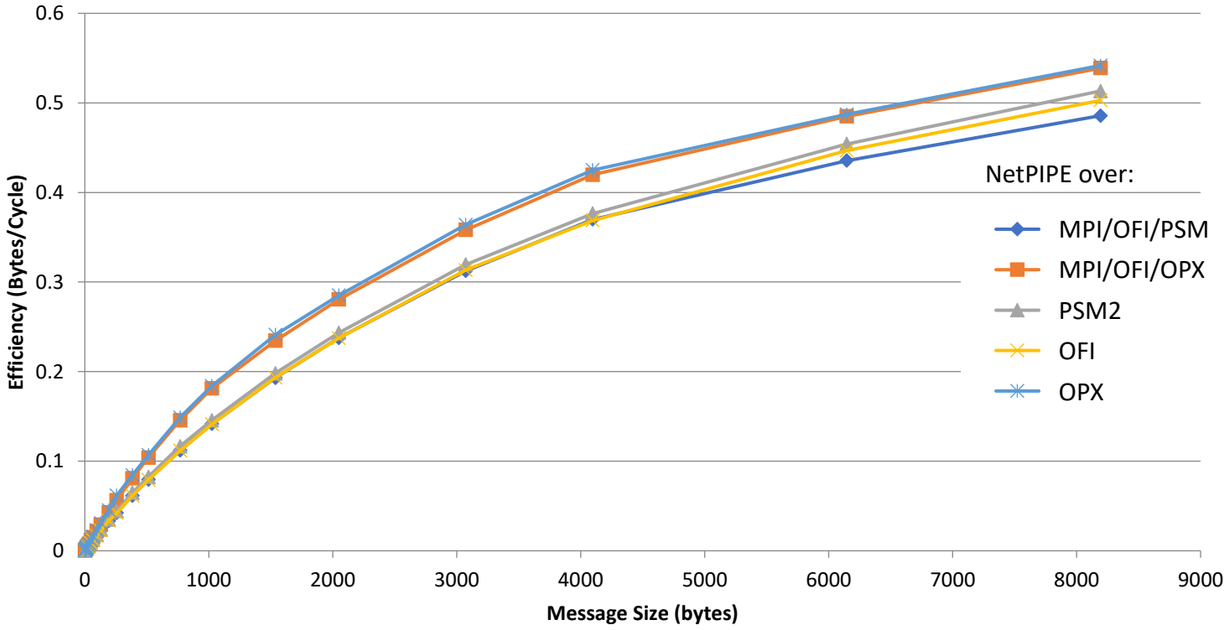
Libfabric Omni-Path Express (OPX) provider

Minimize Software Overhead

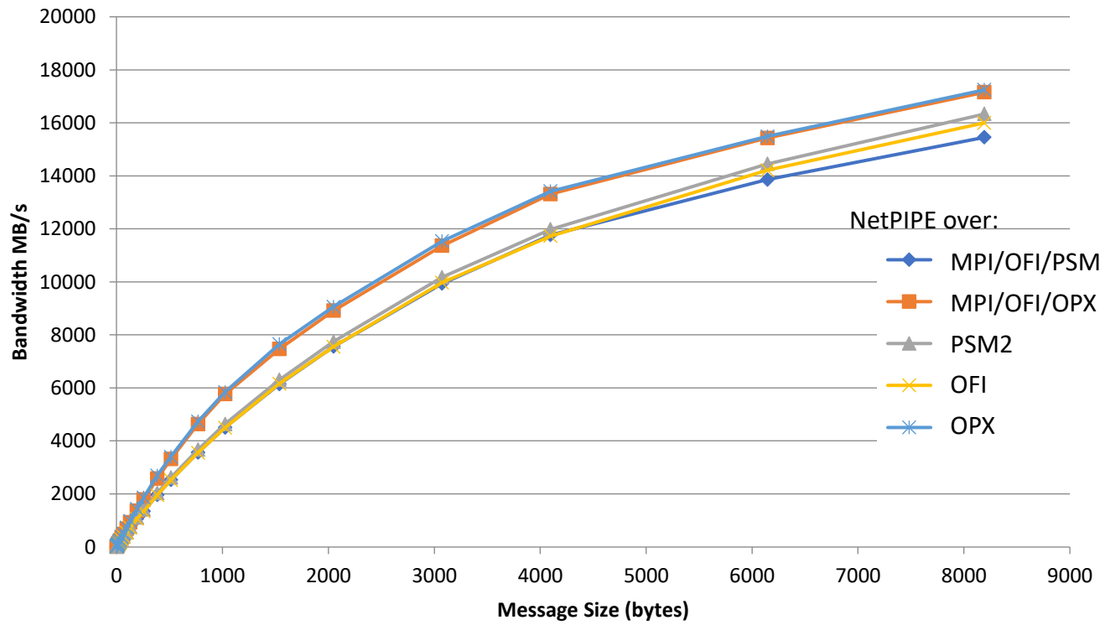
- Up to 45% realizable efficiency gains in MPI
- Efficiency as measured by Bytes/Cycle
- A good measure of overall footprint in instructions, cache use
- Realizable as latency and message rate improvements
- Real world improvements in MPI applications (MPI adds overhead in the noise with OFI)

- NetPIPE Benchmark with Cold Caches
- OPX Provider
- PSM

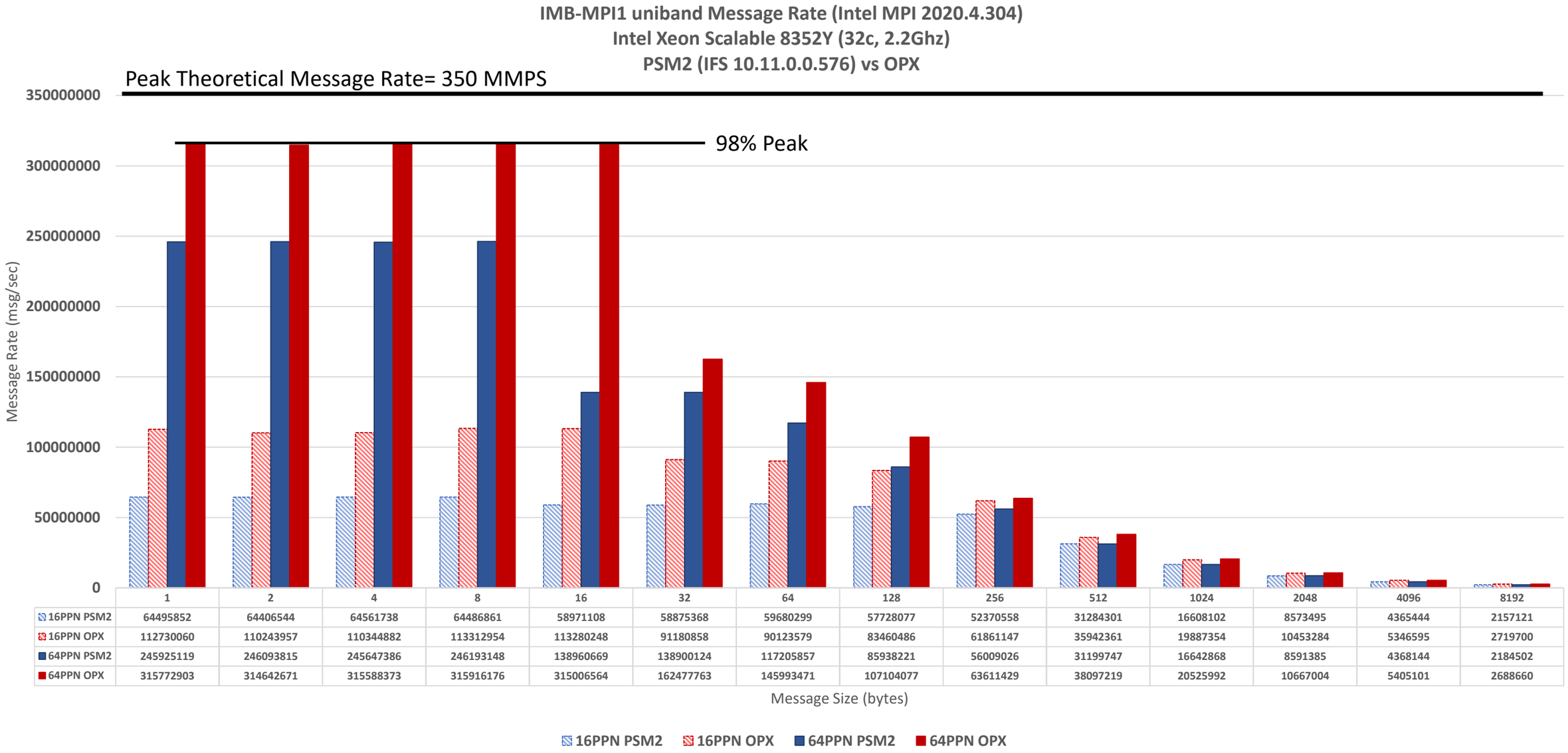
NetPIPE Efficiency: Bytes/Cycle



NetPIPE Bandwidth: MB/s



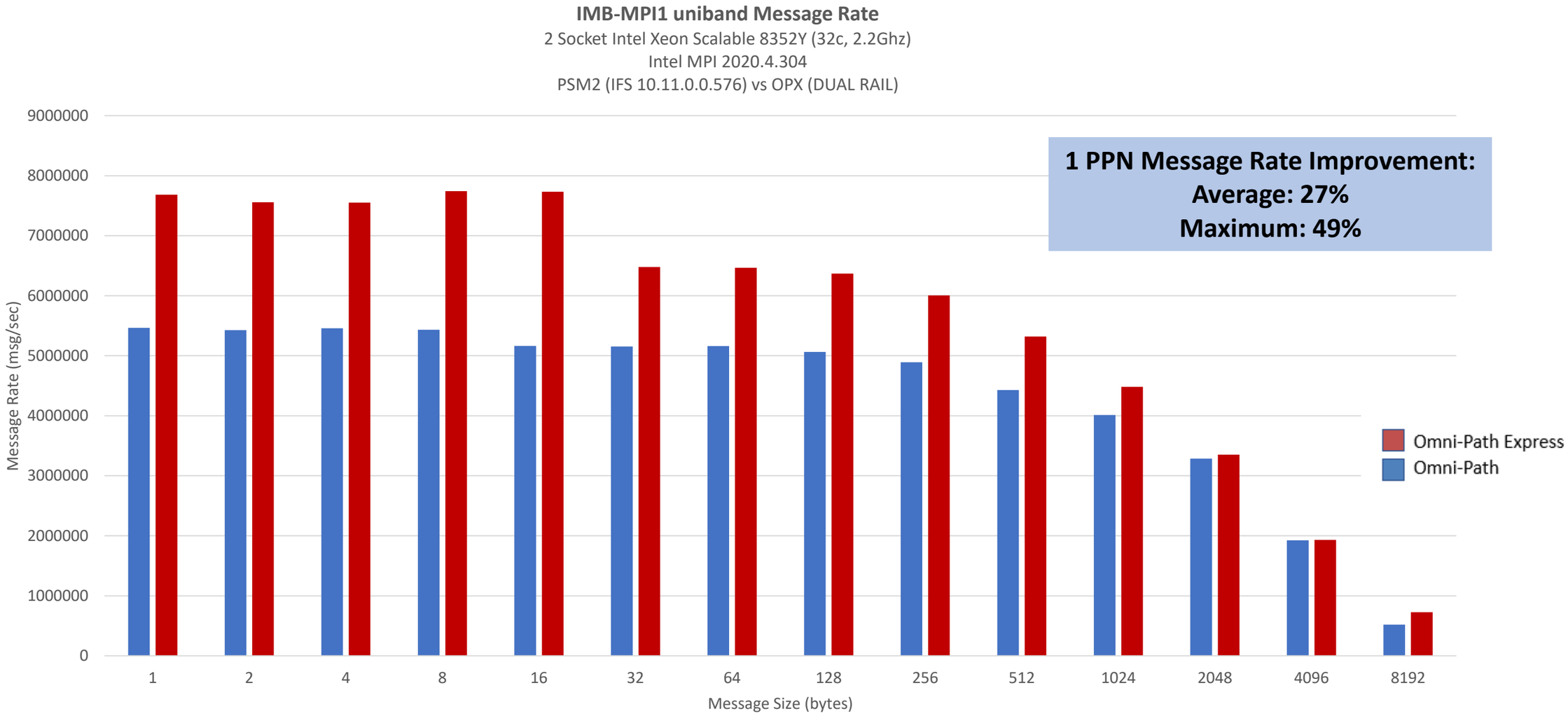
OPX Provider Message Rate Improvements



	1	2	4	8	16	32	64	128	256	512	1024	2048	4096	8192
16PPN PSM2	64495852	64406544	64561738	64486861	58971108	58875368	59680299	57728077	52370558	31284301	16608102	8573495	4365444	2157121
16PPN OPX	112730060	110243957	110344882	113312954	113280248	91180858	90123579	83460486	61861147	35942361	19887354	10453284	5346595	2719700
64PPN PSM2	245925119	246093815	245647386	246193148	138960669	138900124	117205857	85938221	56009026	31199747	16642868	8591385	4368144	2184502
64PPN OPX	315772903	314642671	315588373	315916176	315006564	162477763	145993471	107104077	63611429	38097219	20525992	10667004	5405101	2688660

Intel MPI + IMB uniband (Small Message Rate)

Intel Xeon, 1ppn, peak per process

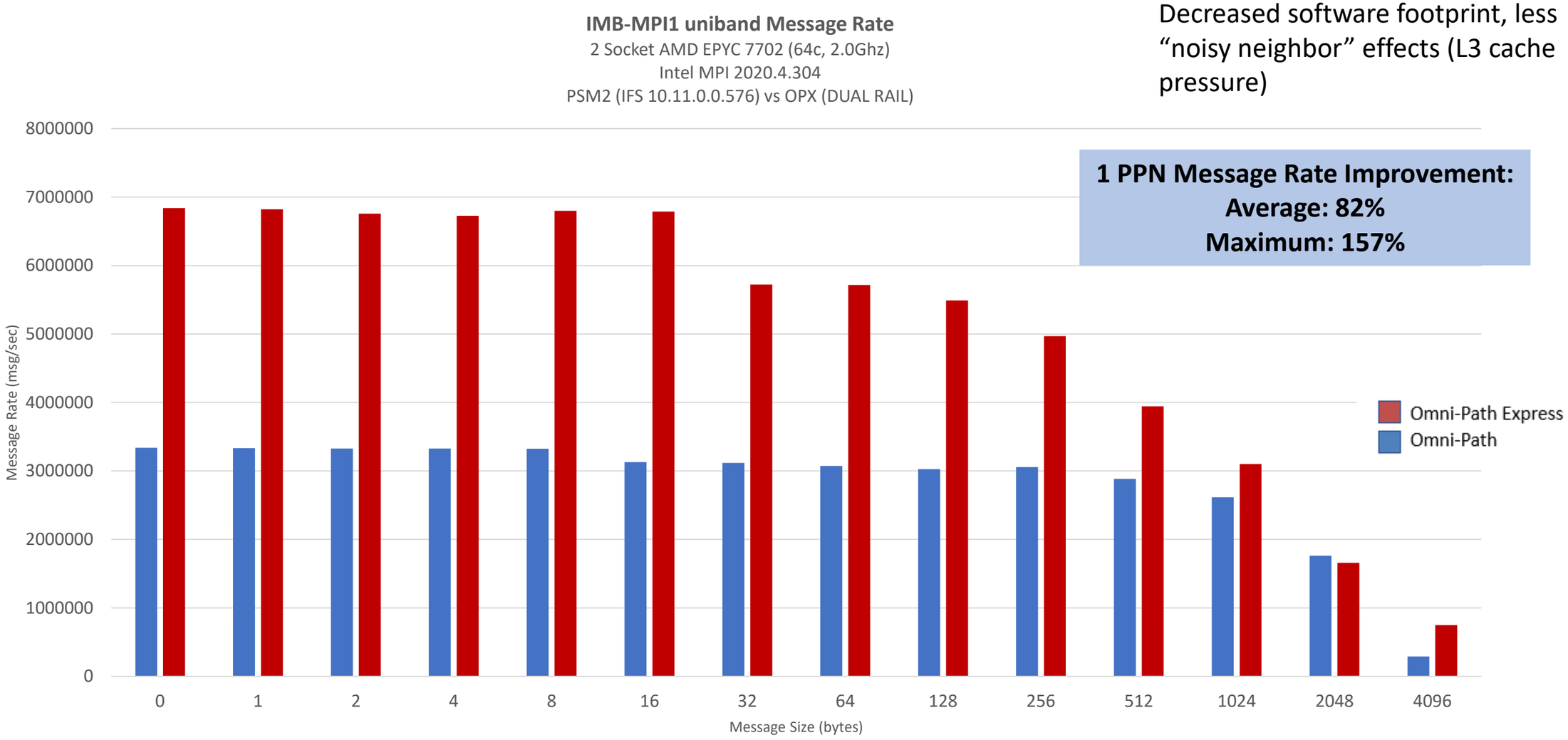


Where applicable: Red Hat Enterprise Linux 8.3 (Ootpa). 4.18.0-240.el8.x86_64 kernel16xDDR4, 512 GB, 3200 MT/s. itlb_multihit:Not affected. l1tf:Not affected. mds:Not affected. meltdown:Not affected. spec_store_bypass:Mitigation: Speculative Store Bypass disabled via prctl and seccomp. spectre_v1:Mitigation: usercopy/swapgs barriers and __user pointer sanitization. spectre_v2:Mitigation: Enhanced IBRS, IBPB: conditional, RSB filling. srbds:Not affected. tsx_async_abort:Not affected. Intel Fabric Suite (IFS) 10.11.0.0.576. HFI driver parameters: pcie_caps=0x51 krcvqs=3 ipoib_accel=0.

Intel MPI + IMB uniband (Small Message Rate)

AMD EPYC, 1ppn, peak per process

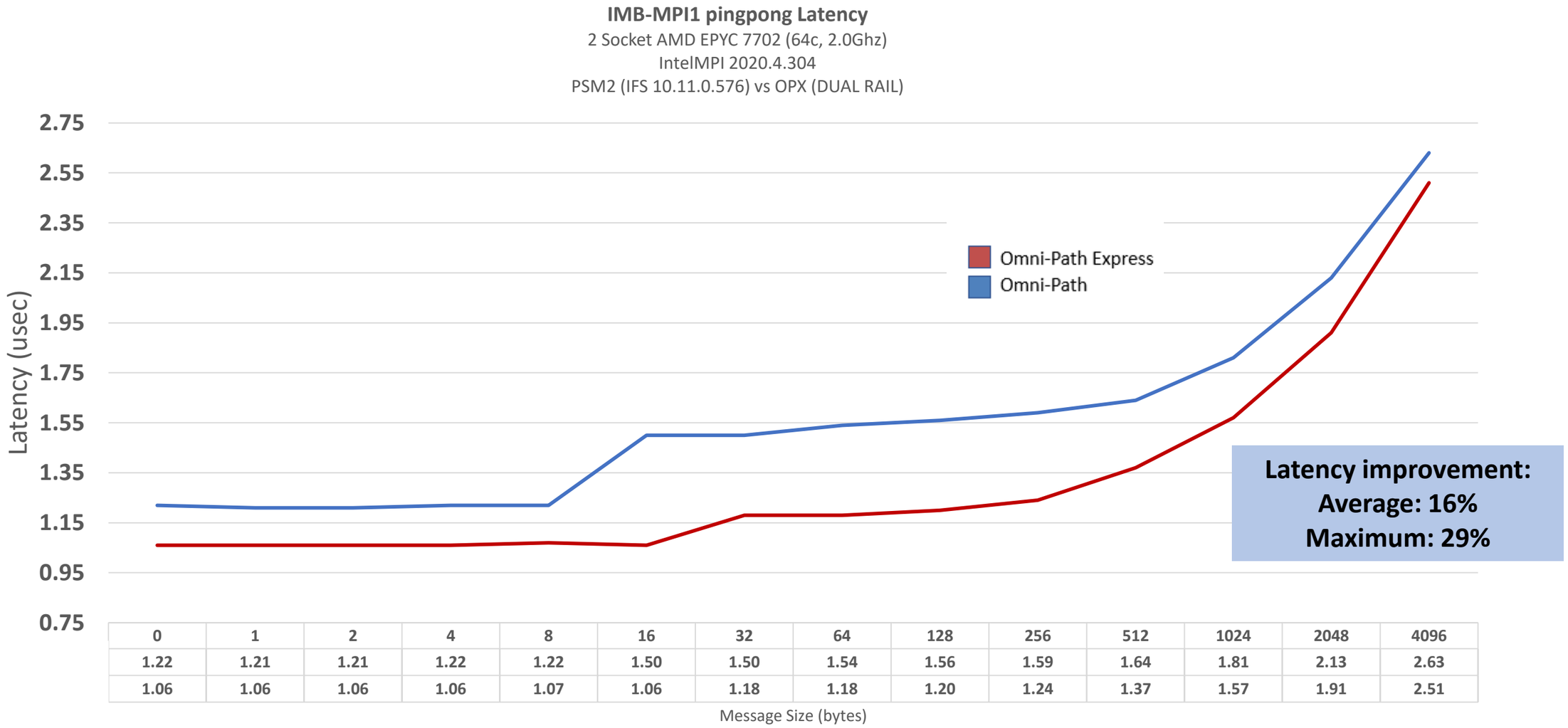
Decreased software footprint, less “noisy neighbor” effects (L3 cache pressure)



Where applicable: Red Hat Enterprise Linux 8.3 (Ootpa). 4.18.0-240.el8.x86_64 kernel16xDDR4, 512 GB, 3200 MT/s. itlb_multihit:Not affected. l1tf:Not affected. mds:Not affected. meltdown:Not affected. spec_store_bypass:Mitigation: Speculative Store Bypass disabled via prctl and seccomp. spectre_v1:Mitigation: usercopy/swapgs barriers and __user pointer sanitization. spectre_v2:Mitigation: Enhanced IBRS, IBPB: conditional, RSB filling. srbds:Not affected. tsx_async_abort:Not affected. Intel Fabric Suite (IFS) 10.11.0.0.576. HFI driver parameters: pcie_caps=0x51 krcvqs=3 ipoib_accel=0.

Intel MPI + IMB PingPong (Small Message Latency)

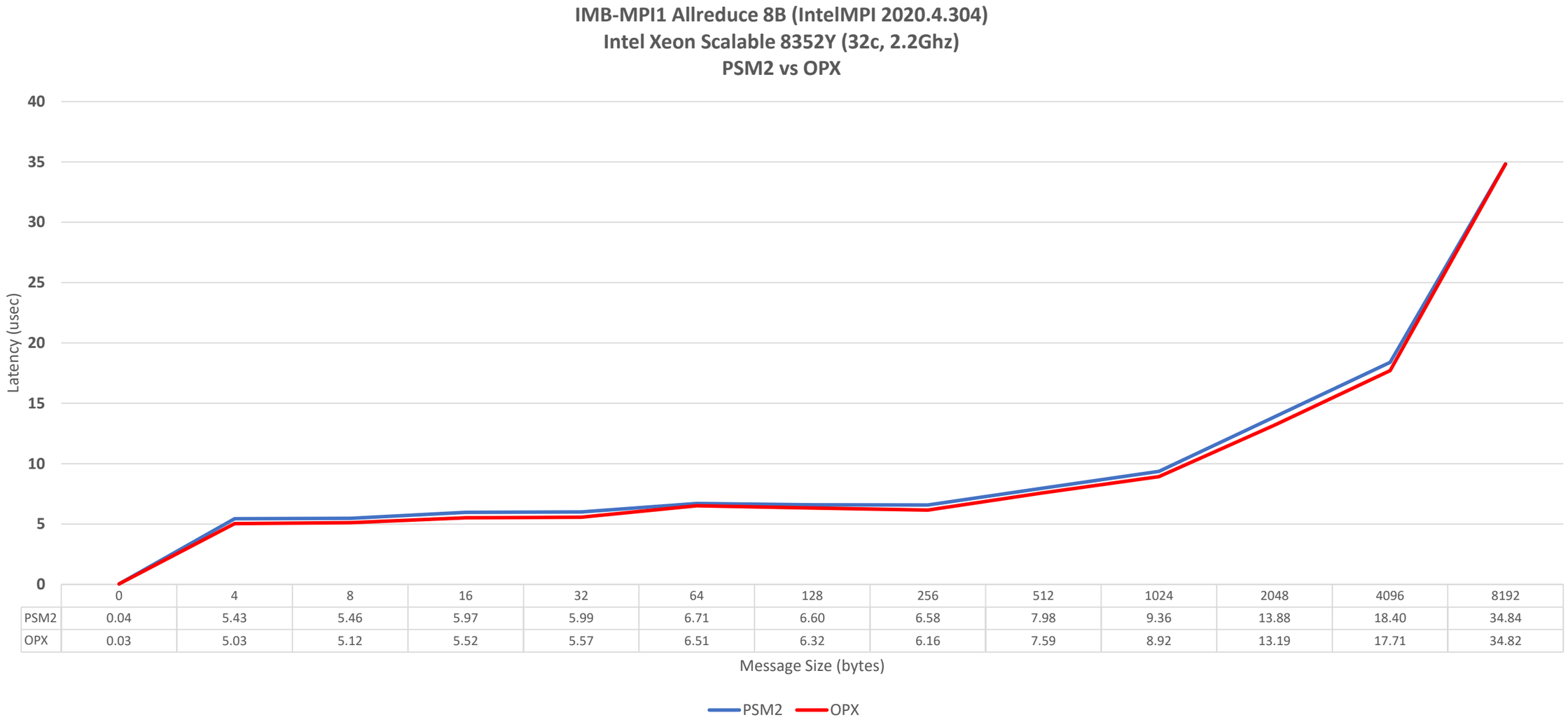
AMD EPYC, 1ppn



0	1	2	4	8	16	32	64	128	256	512	1024	2048	4096
1.22	1.21	1.21	1.22	1.22	1.50	1.50	1.54	1.56	1.59	1.64	1.81	2.13	2.63
1.06	1.06	1.06	1.06	1.07	1.06	1.18	1.18	1.20	1.24	1.37	1.57	1.91	2.51

Where applicable: Red Hat Enterprise Linux 8.3 (Ootpa). 4.18.0-240.el8.x86_64 kernel16xDDR4, 512 GB, 3200 MT/s. itlb_multihit:Not affected. l1tf:Not affected. mds:Not affected. meltdown:Not affected. spec_store_bypass:Mitigation: Speculative Store Bypass disabled via prctl and seccomp. spectre_v1:Mitigation: usercopy/swapgs barriers and __user pointer sanitization. spectre_v2:Mitigation: Enhanced IBRS, IBPB: conditional, RSB filling. srbds:Not affected. tsx_async_abort:Not affected. Intel Fabric Suite (IFS) 10.11.0.0.576. HFI driver parameters: pcie_caps=0x51 krcvqs=3 ipoib_accel=0.

Allreduce (16 Nodes, 64 PPN)



	0	4	8	16	32	64	128	256	512	1024	2048	4096	8192
PSM2	0.04	5.43	5.46	5.97	5.99	6.71	6.60	6.58	7.98	9.36	13.88	18.40	34.84
OPX	0.03	5.03	5.12	5.52	5.57	6.51	6.32	6.16	7.59	8.92	13.19	17.71	34.82

Cornelis Networks Omni-Path 400 Overview

Extending decades of investments:

- OpenFabrics Alliance
- QLogic InfiniBand
- Cray Aries
- Intel® Omni-Path Architecture

Delivering industry-leading fabric:

- Artificial intelligence
- Data analytics
- Traditional modeling & simulation

Best-in-Class OFI-Focused AI / HPC Adapter

High-performance data pipeline complemented with scalable OFI accelerators

High-Bandwidth Low-Diameter Topologies

High-radix 400Gbps switch building blocks supporting tree and advanced topologies

Low Latency at Any Utilization and Scale

Minimum fall-through latency coupled with advanced traffic management

Advanced Adaptive Routing and QoS

Multi-tenant security and QoS via dynamic virtual fabrics and fine-grained adaptive routing

State-of-the-Art Congestion Control

Comprehensive network telemetry synthesized to pace senders and modulate path selection

Summary and Future Trends

- Independent company focused on high-performance scale-out interconnects for modeling/simulation, data analytics, and AI
- Co-optimizing for a wide variety of processors and accelerators
- Committed to open standards and delivering best-in-class solutions based on OpenFabrics Interfaces
- Delivering best performance and price/performance today for MVAPICH2, other MPIs, PGAS, AI frameworks, DAOS, and accelerators
- Investing in next-generation for increased bandwidth and application-focused acceleration technologies
- Highest performance and best price/performance for converged infrastructure

Thank You