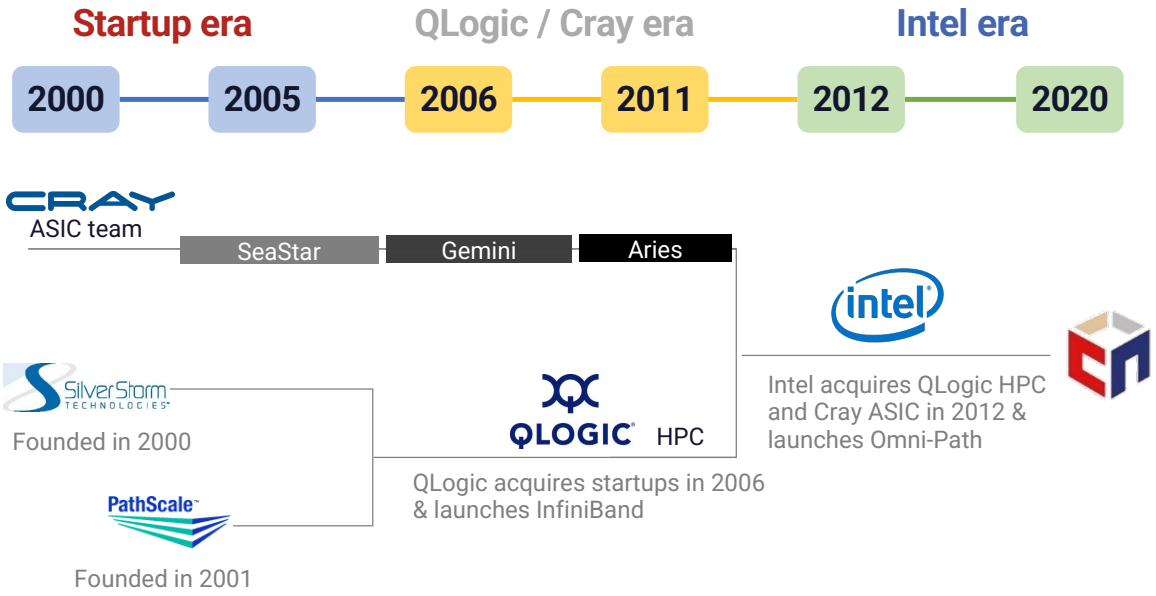
OPX – Informed Development for a Next-Gen libfabric Provider

Charles Shereda, Engineering Manager, OPX Development

Ben Lynam, Senior Software Engineer, OPX Development

MPICH BOF, SC22

Who we are



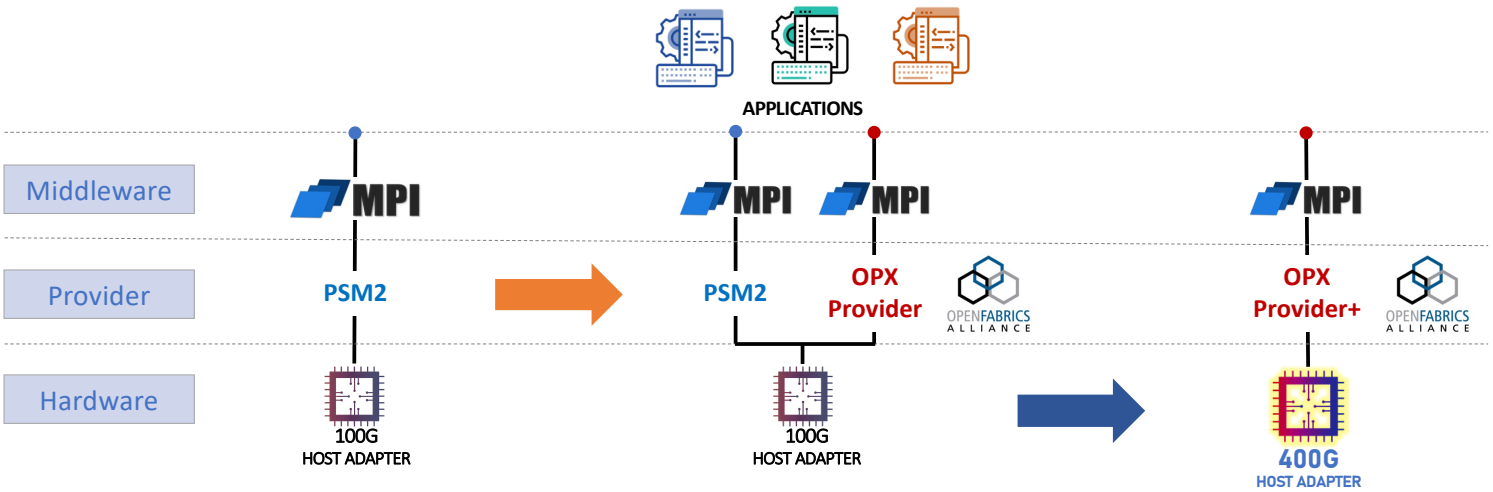
Technology built on ~$1B investment over 20 years

CORNELIS NETWORKS

Cornelis Networks era

- Acquired Intel Omni-Path business
- Delivering complete networking solution
- Supporting 500+ global deployments
- Developing strong ecosystem support
- Serving Government, Academic, Scientific, & Commercial segments
- Enhancing Omni-Path solutions with next generation development

Omni-Path Evolution



New Software

Significant functionality and performance enhancements via

Libfabric over Omni-Path Express

New Hardware

Optimized performance via

Premier OFI Adapter

Broad Ecosystem Support

Broad ecosystem support targeted across…

- All popular CPUs and accelerators, ensuring solution choice
- Key application-critical technologies, libraries, and frameworks
- Leading OEMs, ISVs, and reseller partners

Communication API	CPU	Storage	AI/ML	GPU
Open MPI, Intel MPI, MVAPICH, MPICH, Sandia OpenSHMEM, Charm++, Chapel, GasNet	Intel, AMD	Lustre, IBM Spectrum Scale, BeeGFS, DAOS (Intel)	TensorFlow, PyTorch	NVIDIA, Intel*, AMD*

All other names, logos, and brands maybe claimed as the property of others

* Support planned

Engineering OPX

- Upstream-first, open source
- OFIWG participant and libfabric provider
- Use PSM2 as performance baseline
- Compare build-to-build performance as well
- Utilize multiple test systems with different core counts
- Optimal protocol and HW paths are selected at runtime
  - Each protocol exploits its own sw/hw path
  - Eager
  - Multipacket Eager
  - Rendezvous

OPX Transfer Methods and Thresholds

PIO Send -> Programmed IO Send –> 0 – 16K

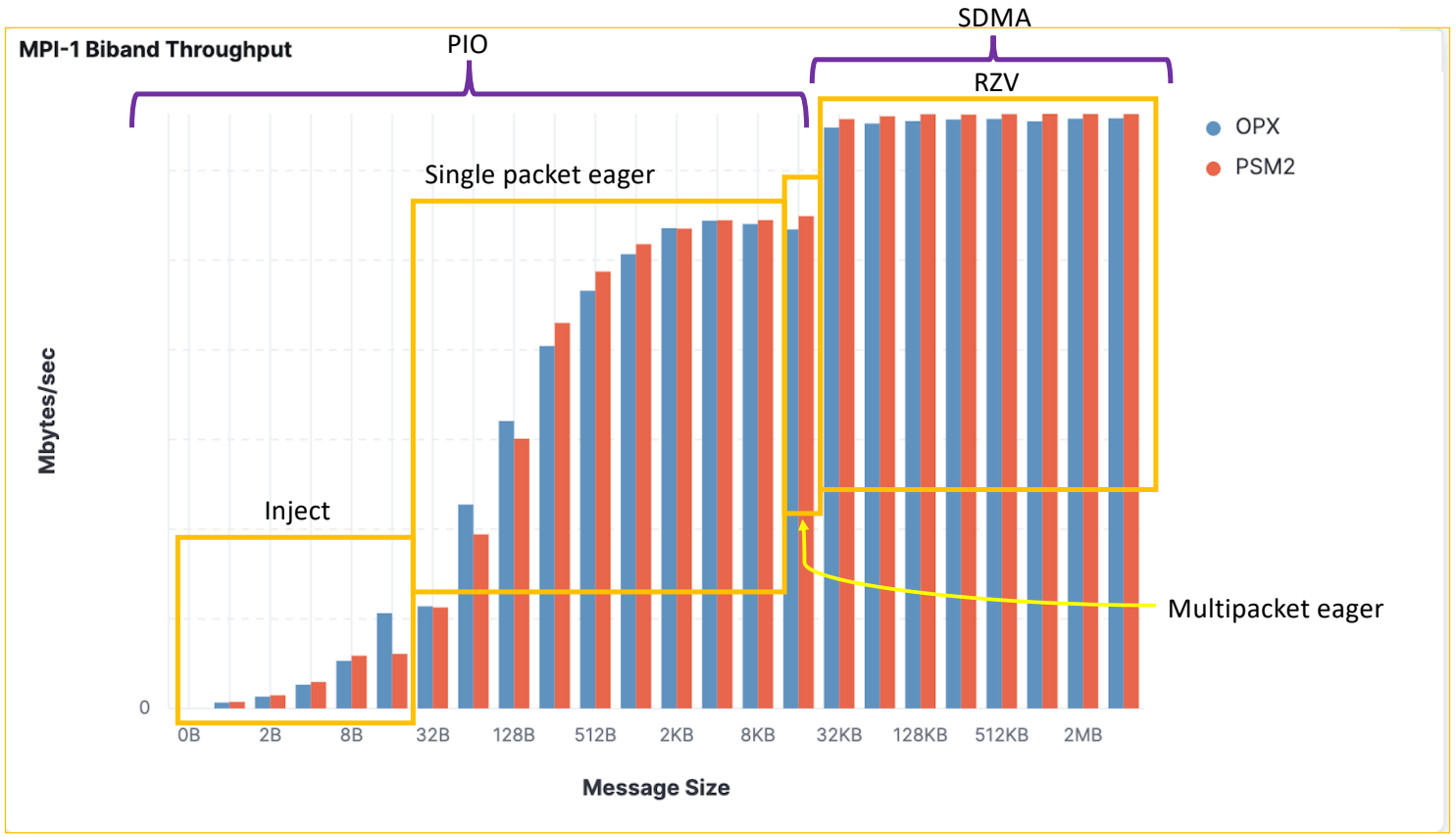
- Data transfer using CPU via memory mapped IO
- 0-16B - Inject (packet w/o payload – header only)
- 17B – 8K – Single packet eager*
- 8K+1 – 16K – Multipacket eager
- 16K+1 – RZV that may utilize either PIO or SDMA
- Faster than SDMA, but resource limited (limited pool of credits per endpoint)

SDMA -> Send DMA -> Minimum of (16K + 1)

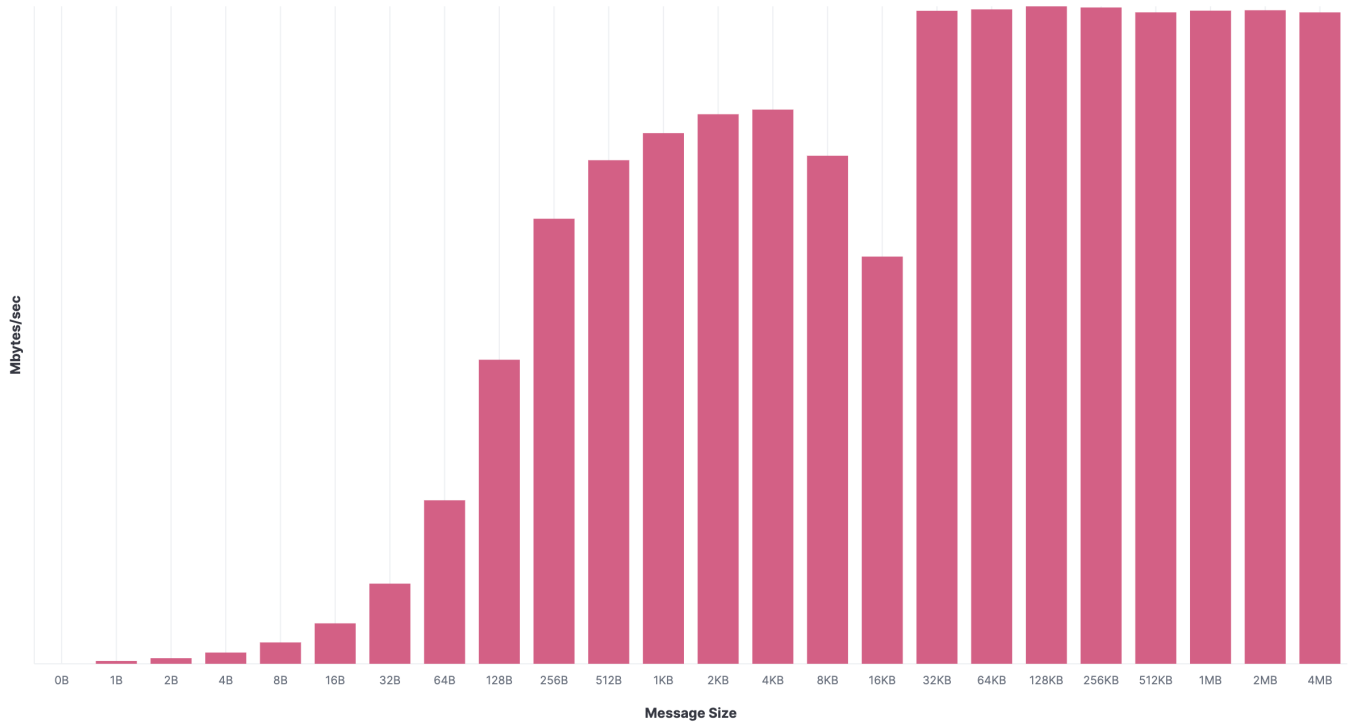
- Data transfer using SDMA engines on HFIs
- Syscall required
- Either one-sided or RZV
- Expected TID would be RZV only

* May fall back to multipacket eager depending on credit availability

Thresholds Graphed (IMB tests)

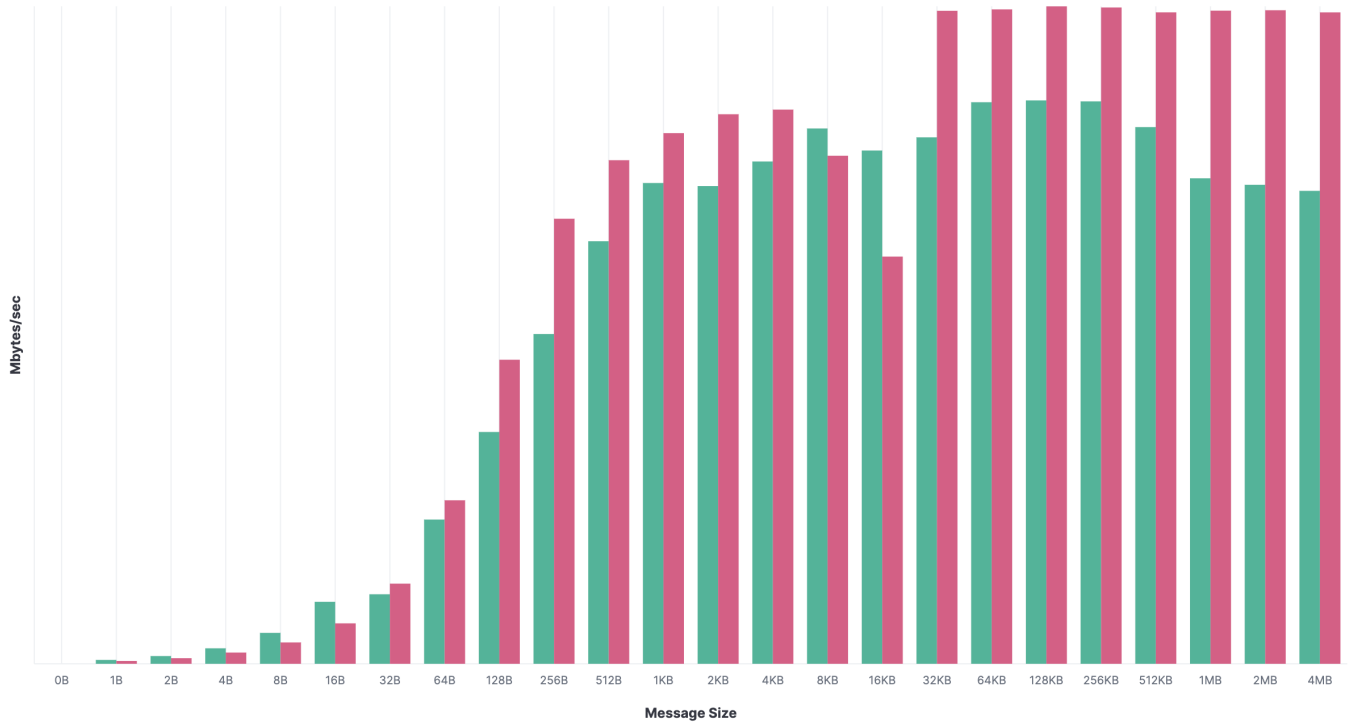


48 PPN runs on Skylake



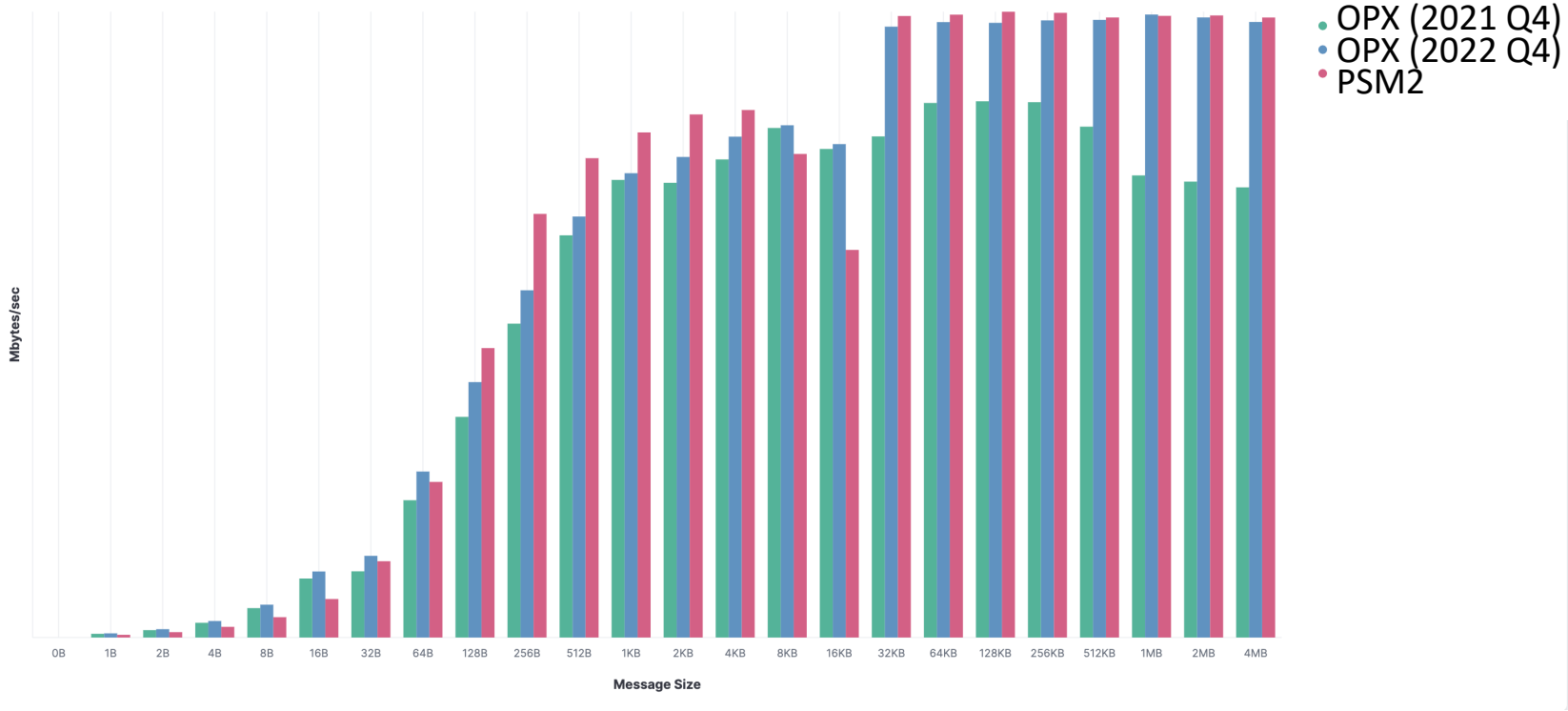
- PSM2

48 PPN runs on Skylake

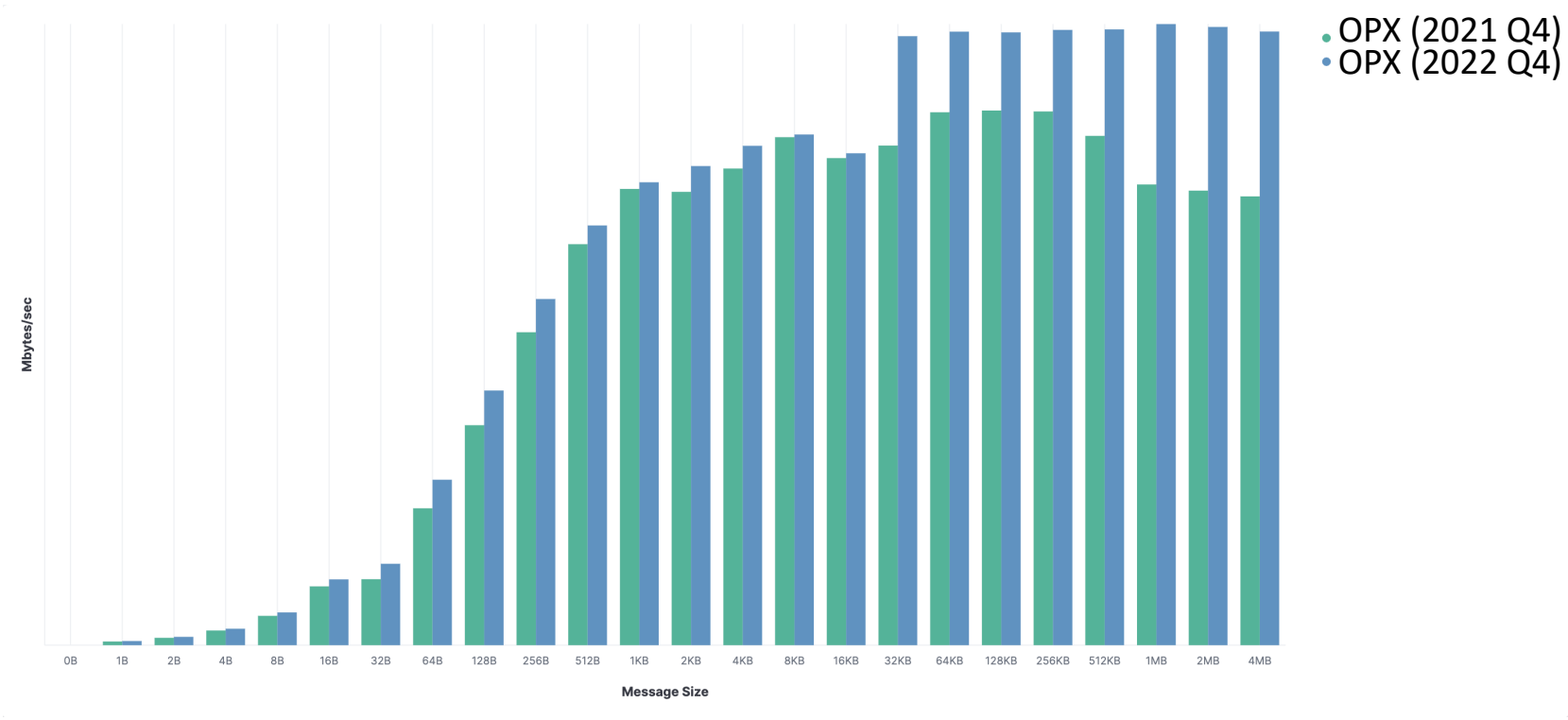


- OPX (2021 Q4)
- PSM2

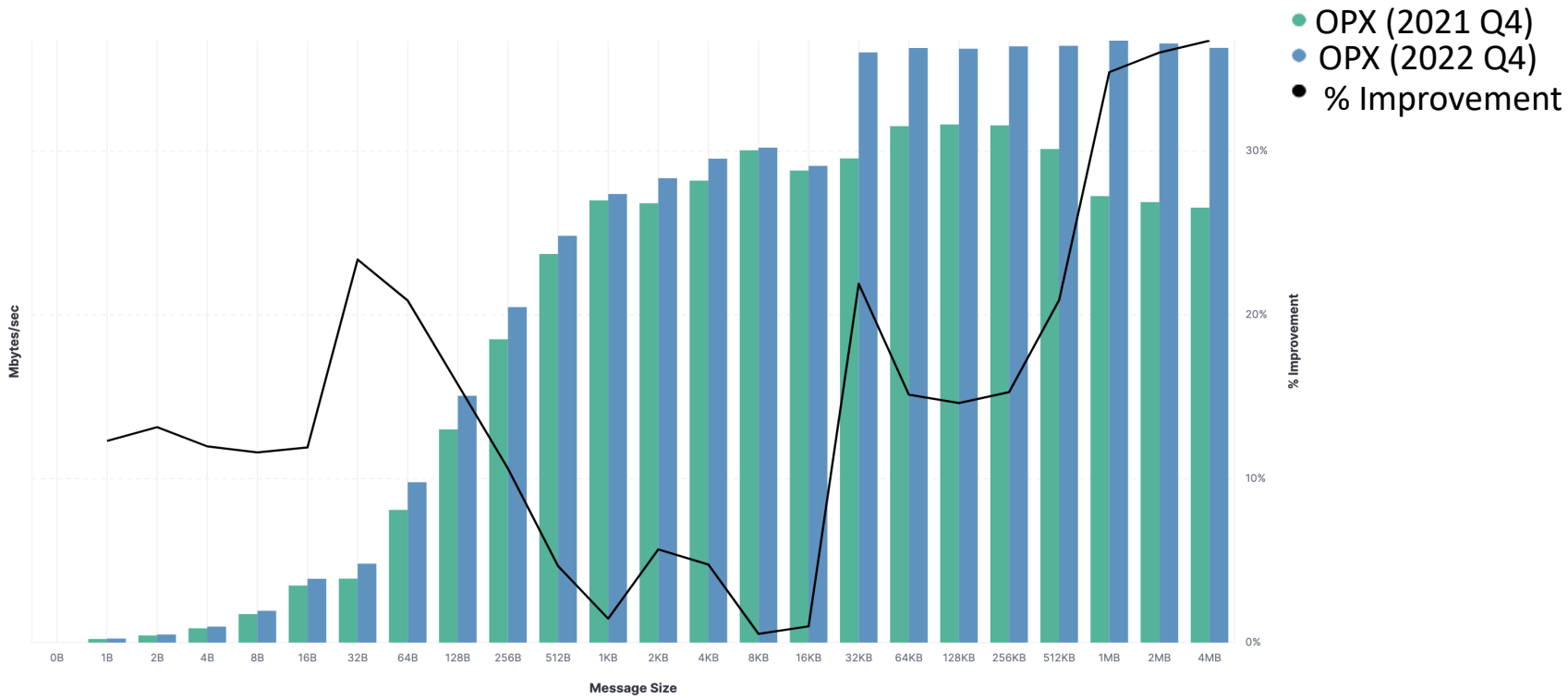
48 PPN runs on Skylake



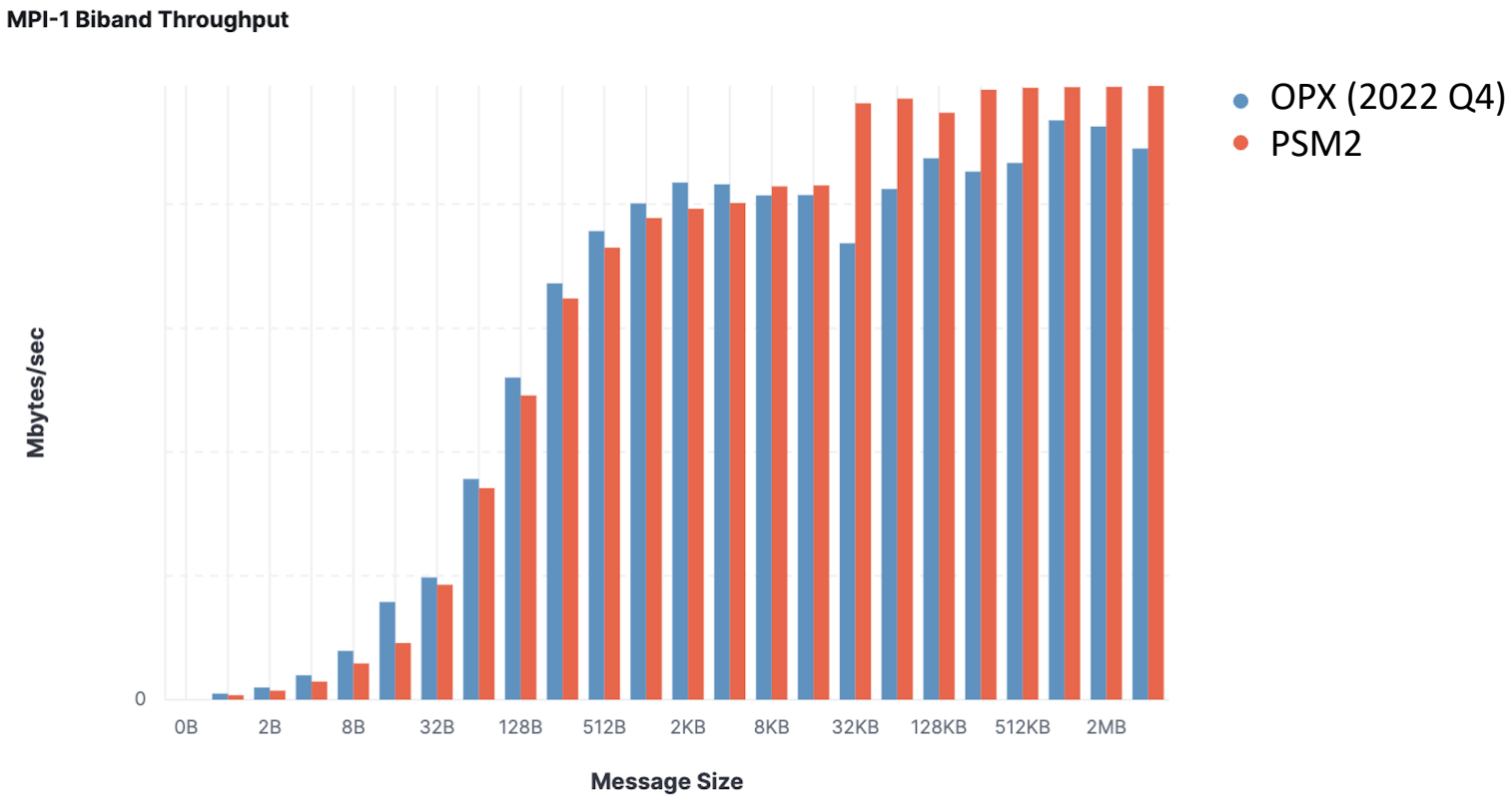
48 PPN runs on Skylake



48 PPN runs on Skylake



128 PPN run on AMD Milan



Omni-Path Express vs PSM2

Processing a Packet

- Optimized incoming packet processing (Do a single MPI_Recv(...))
  - Intel SDE testing shows tremendous improvement in instruction count
  - Significant improvements in cache line footprint

	PSM2	OPX	Improvement
Instruction count	3064	1170	62%
Cache lines for code	205	124	40%
Cache line loads	93	55	41%
New cache line access	354	209	41%

- Every commit is checked to ensure no regressions

MPICH test bucket

- MPICH tests run in addition to performance tests

- Our MPICH test focus is more on correctness and coverage

OPX Status

- OPX is part of libfabric 1.16.1
  - https://github.com/ofiwg/libfabric
  - Peacefully coexists with PSM2

- Omni-Path Express Suite (OPXS)
  - Cornelis's supported release vehicle
  - OPXS release containing OPX : **10.12.0.0.22**, just released

Thank You