2022 OFA Virtual Workshop

Omni-Path Express (OPX) Libfabric Provider

Overview & Status

Tim Thompson & Dennis Dalessandro

Cornelis Networks

Tuesday, April 26th 11:00 – 11:30am PST | 2:00 – 2:30pm EST

Notices and Disclaimers

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH CORNELIS NETWORKS PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN CORNELIS NETWORKS'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, CORNELIS NETWORKS ASSUMES NO LIABILITY WHATSOEVER, AND CORNELIS NETWORKS DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF CORNELIS NETWORKS PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT. CORNELIS NETWORKS PRODUCTS ARE NOT INTENDED FOR USE IN MEDICAL, LIFE SAVING, OR LIFE SUSTAINING APPLICATIONS.

Cornelis Networks may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined". Cornelis Networks reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.

All products, dates, and figures specified are preliminary based on current expectations, and are subject to change without notice. Roadmap not reflective of exact launch granularity and timing. The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

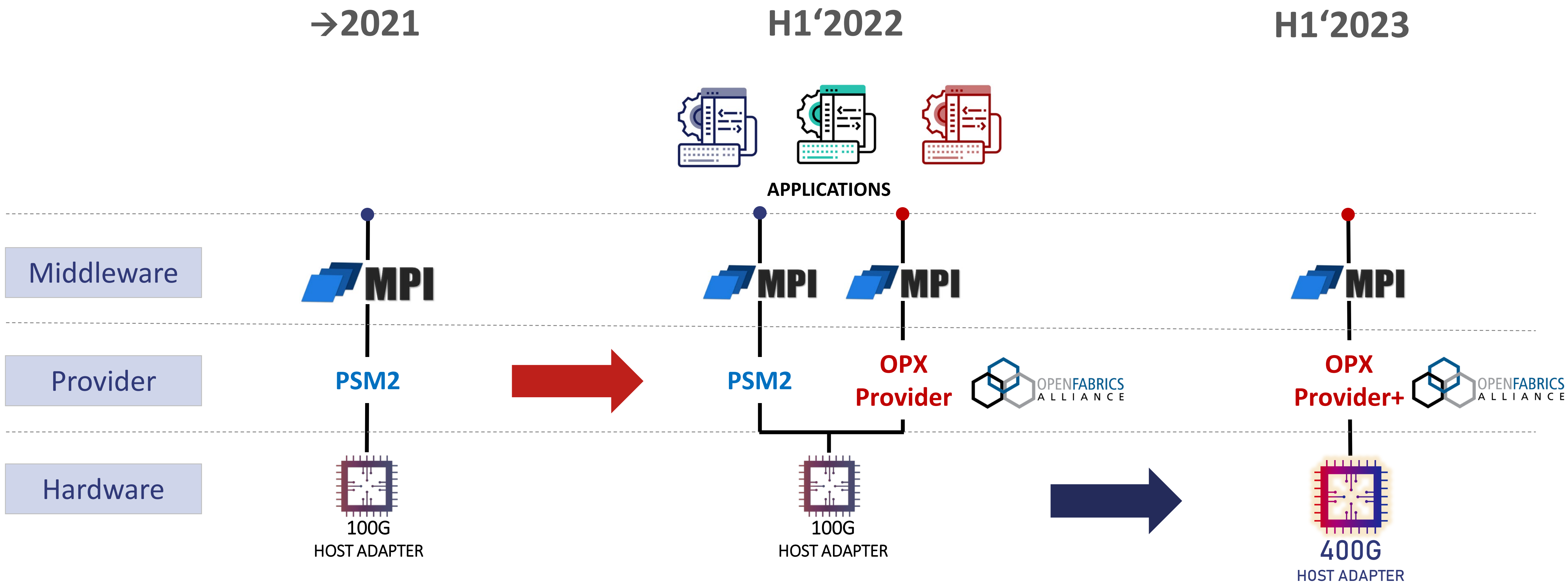
Any code names featured are used internally within Cornelis Networks to identify products that are in development and not yet publicly announced for release. Customers, licensees and other third parties are not authorized by Cornelis Networks to use code names in advertising, promotion or marketing of any product or services and any such use of Cornelis Networks' internal code names is at the sole risk of the user.

All products, computer systems, dates and figures specified are preliminary based on current expectations and are subject to change without notice. Material in this presentation is intended as product positioning and not approved end user messaging.

Performance tests are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.

Cornelis Networks technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration.

Omni-Path Evolution



New Software

Significant functionality and performance enhancements via ***Libfabric over Omni-Path Express***

New Hardware

Optimized performance via ***Premier OFI Adapter***

PSM: Performance Scaled Messaging
OPX: Libfabric over Omni-Path Express

**Simplification for concept illustration. Future features/options are subject to change without notice.*

Introducing Omni-Path Express Libfabric Provider

Enabling Dramatic Performance Improvements

- Optimized for high-performance converged infrastructures
  - Host architecture based on OpenFabrics Interfaces (OFI)
  - Access to industry standard frameworks and ongoing open-source development
  - Significant application performance gains resulting from accelerated fabric performance
    - Improved time-to-solution and return on investment
  - Foundational for next generation Omni-Path fabric architecture
    - Seamless transition to future Omni-Path platforms
  - Broad support coming for application-critical technologies
    - All popular MPIs, AI frameworks, Object Storage file systems like DAOS, and all popular GPUs

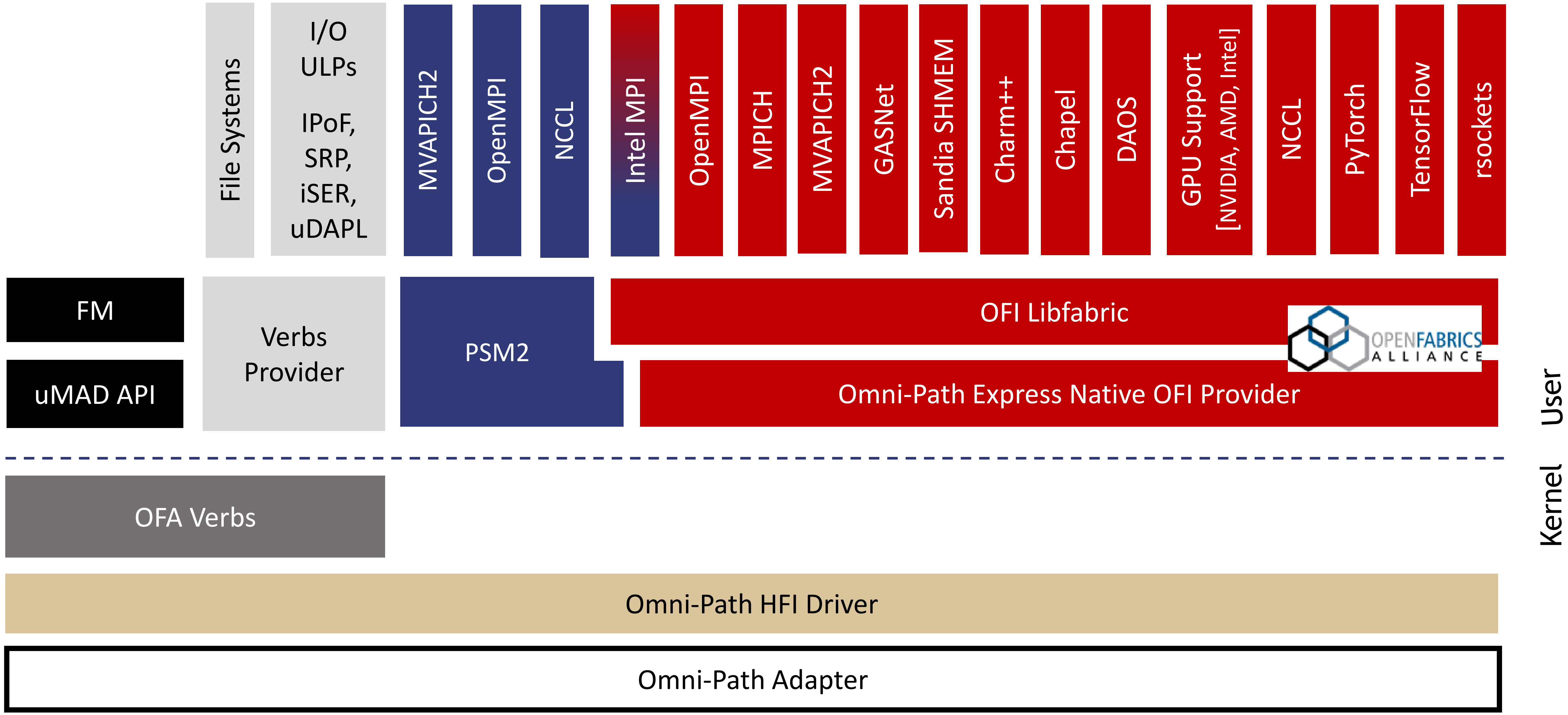
Clearing Up Library Confusion

- PSM2
  - Native OPA Provider, has a history
  - The original way of supporting MPIs on Omni-Path
  - Support continues in OPA100
- Libfabric Through PSM2
  - Libfabric uses PSM2
  - Two layers of APIs (not optimal)
- PSM3
  - Fork of PSM2 to support Ethernet by Intel
- Omni-Path Express often referred to as: OPX (this talk)
  - Native Omni-Path support for Libfabric
  - Replaces PSM2 eventually
  - All the benefits of Libfabric with optimized performance

Omni-Path Express Host Software Stack

Accelerating the Next Level of Application Performance

- Fully open-sourced messaging software stack
- Leveraging libfabric with lightweight OFI Provider
- Facilitating rapid adoption of optimized communication libraries
- Foundational hardware/software co-design driving innovation



**Simplification for concept illustration. Future features/options are subject to change without notice.*

Omni-Path Express Design Philosophy

- PSM2
  - Support applications with well-engineered, durable API
  - Handle HW access to achieve performance (provider)

- OPX
  - Leave application support/API to upper layer Libfabric (and the community)
  - Focus on the part that really matters for performance: The Provider

- OPX designed from ground up to be performance optimal
  - Must be as good as PSM2....turns out it gets even better!
  - Instruction count and cache line footprint are a major goal

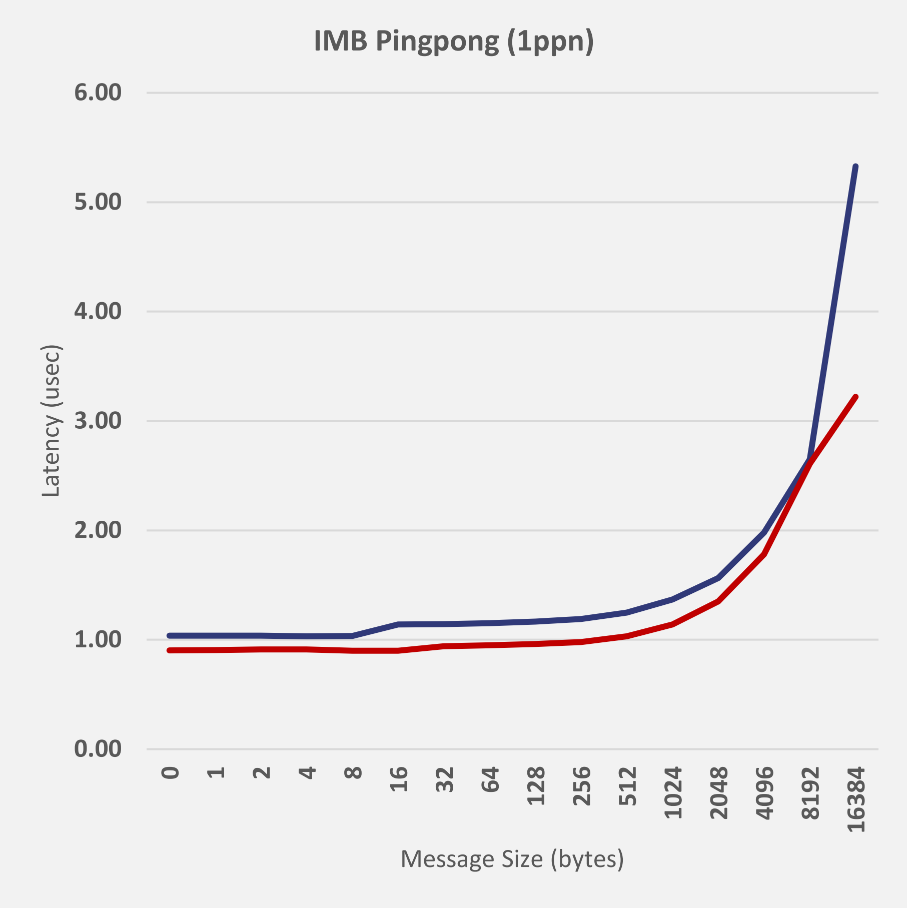
- Bottom line
  - Performance rules

Significant Performance Improvements

Intel Xeon Icelake Platform

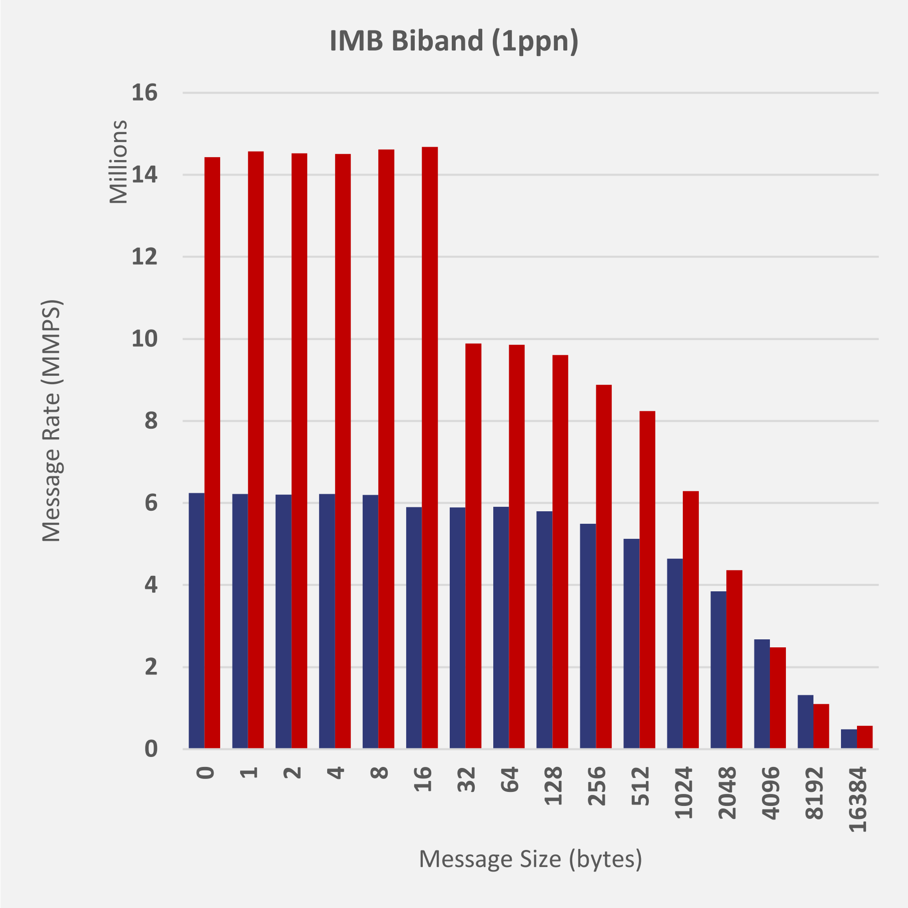
✓ Latency

Up to 20% latency improvement



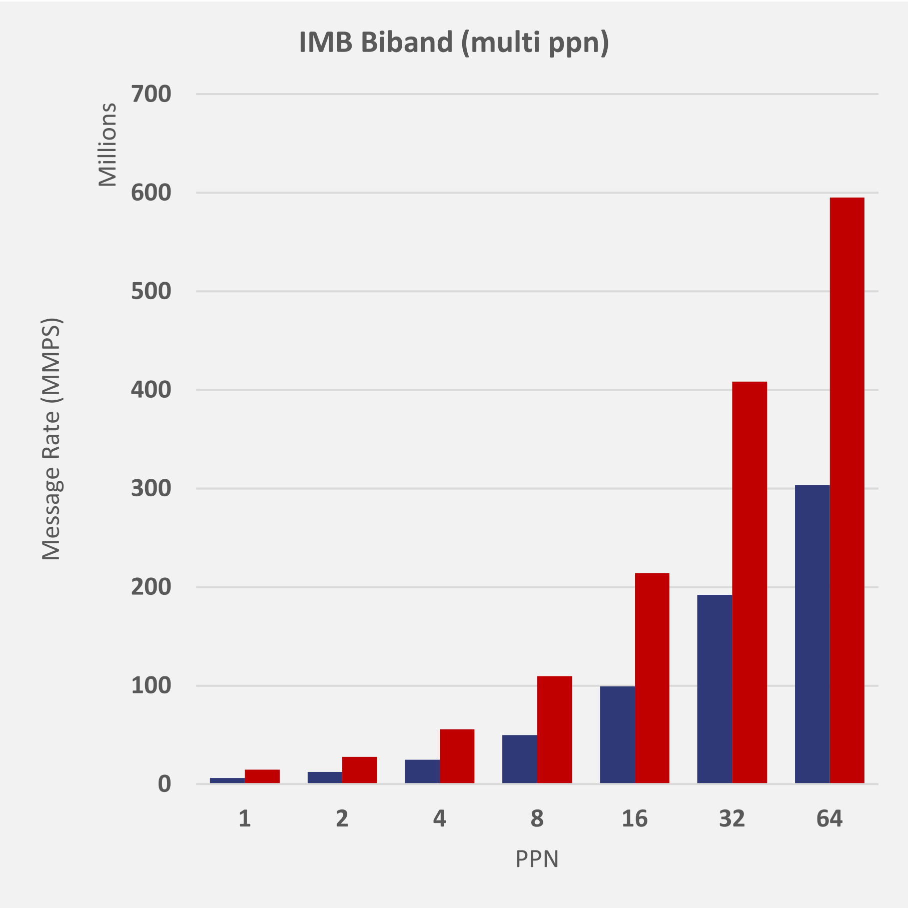
✓ Message Rate

Up to 2.4X messages/sec per core



✓ Scalability

Linear Scalability at double the throughput



🟦 PSM2 Provider 🟥 OPX Provider

Test Configuration:
2-socket Intel® 3rd Generation Xeon® Scalable (Icelake) Platinum 8358, Dual Rail OPA100, BIOS: Snoop Hold-off Response Timer=11, Energy Efficient Turbo=DISABLED, C-States=DISABLED
Rocky Linux 8.4 (Green Obsidian), Kernel 4.18.0-305.19.1.el8_4.x86_64, IntelMPI 2019.6, IMB 2019.6, IFS 10.11.1.1.1, OPX Build 225

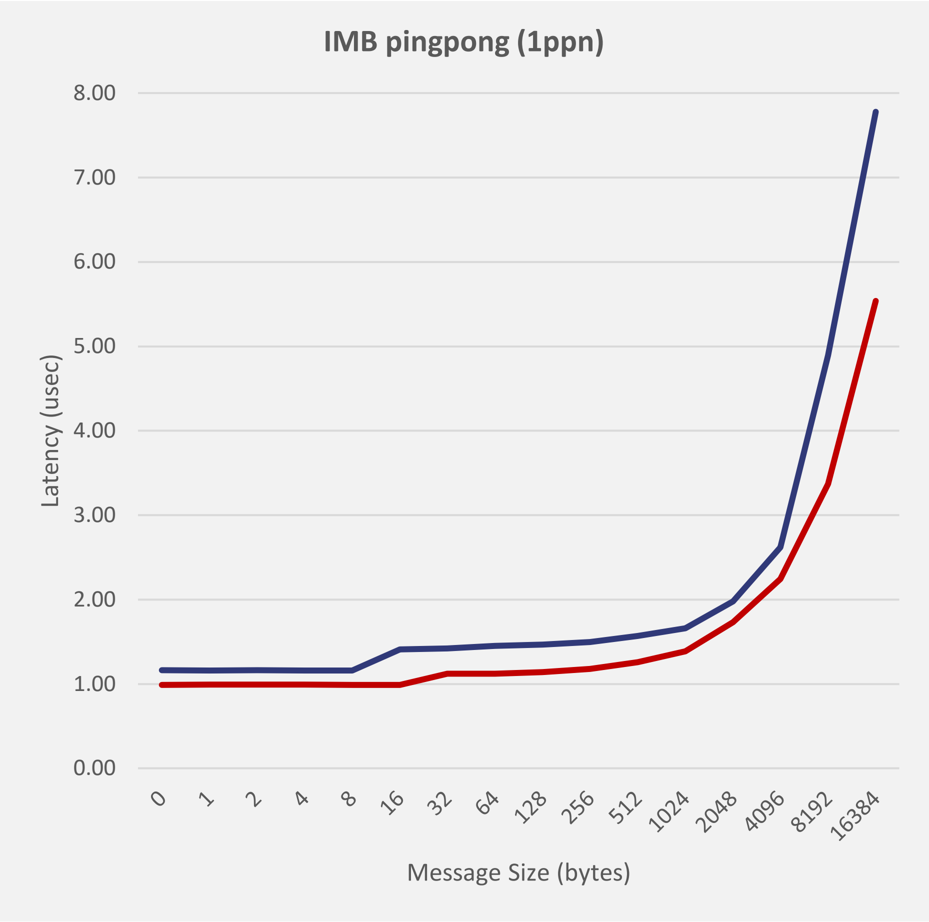
Significant Performance Improvements

AMD EPYC Milan Platform

CORNELIS NETWORKS

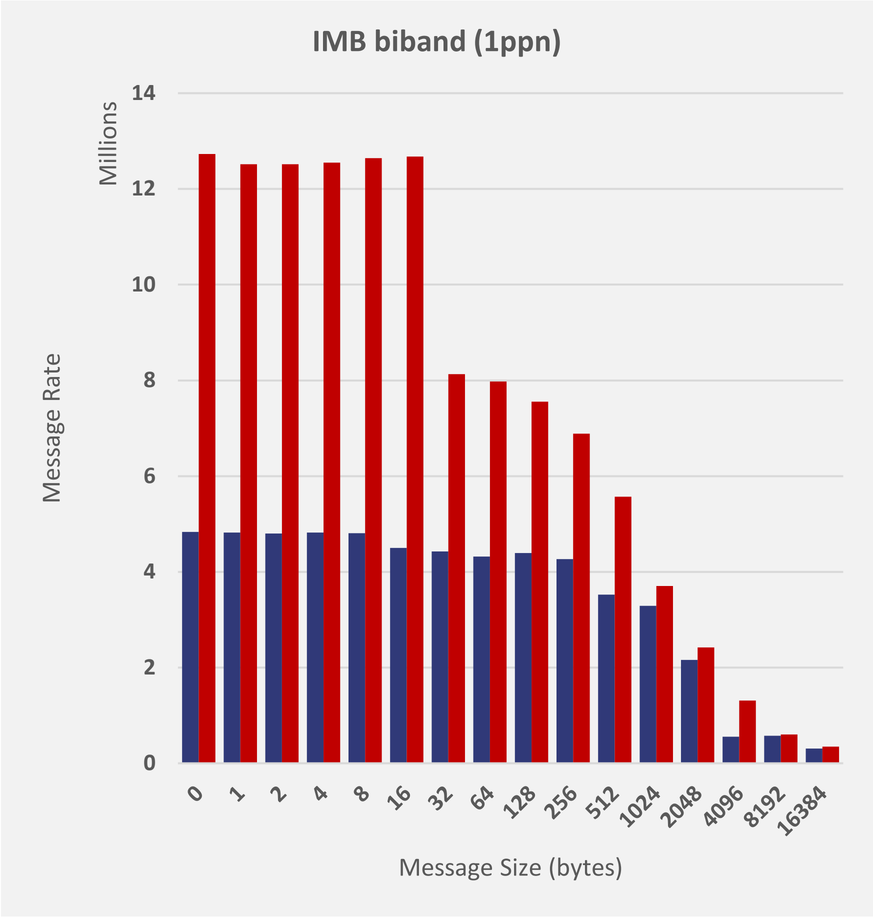
✓ Latency

Up to 25% latency improvement



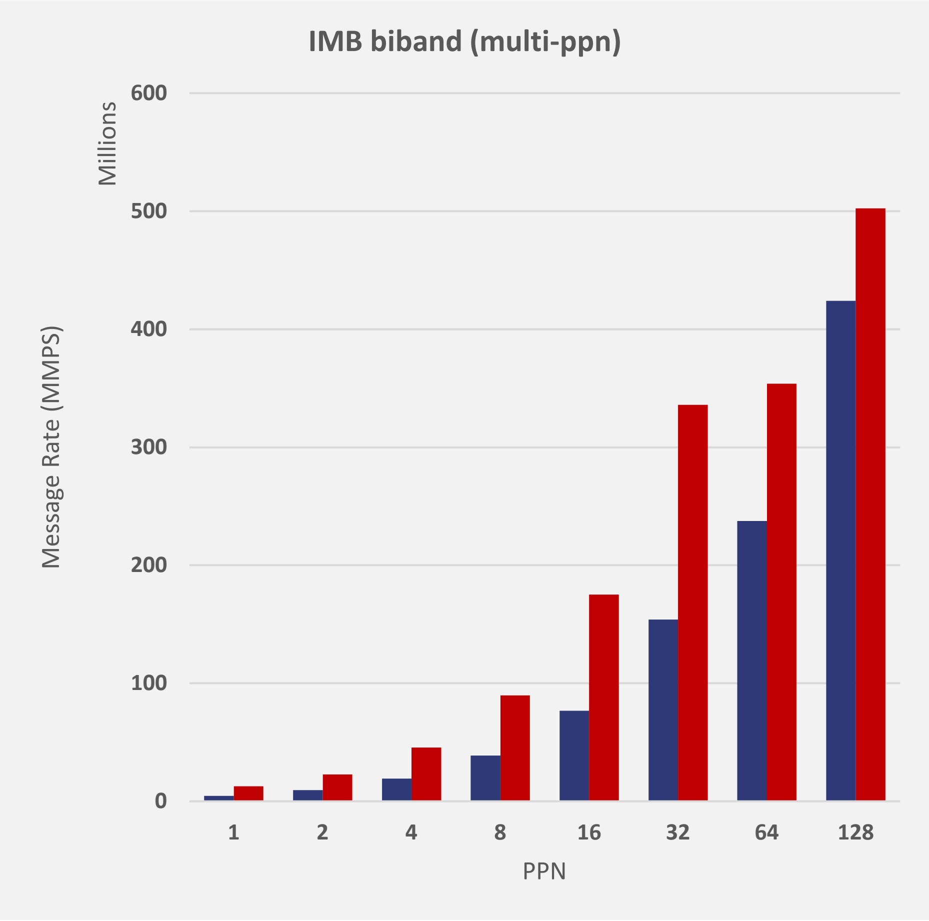
✓ Message Rate

Up to 2.6X messages/sec per core



✓ Scalability

Linear Scalability at double the throughput



🟦 PSM2 Provider 🟥 OPX Provider

Test Configuration:
2-socket AMD EPYC (Milan) 7713, Dual Rail OPA100, xGMI Frequency Locked, xGMI Link Width Locked, P-State Disabled, PCIe Slot Frequency Locked
CentOS Linux 8.3, IntelMPI 2019.6, IMB 2019.6, IFS 10.11.1.1.1, OPX Build 223

Omni-Path Express vs PSM2

Processing a Packet

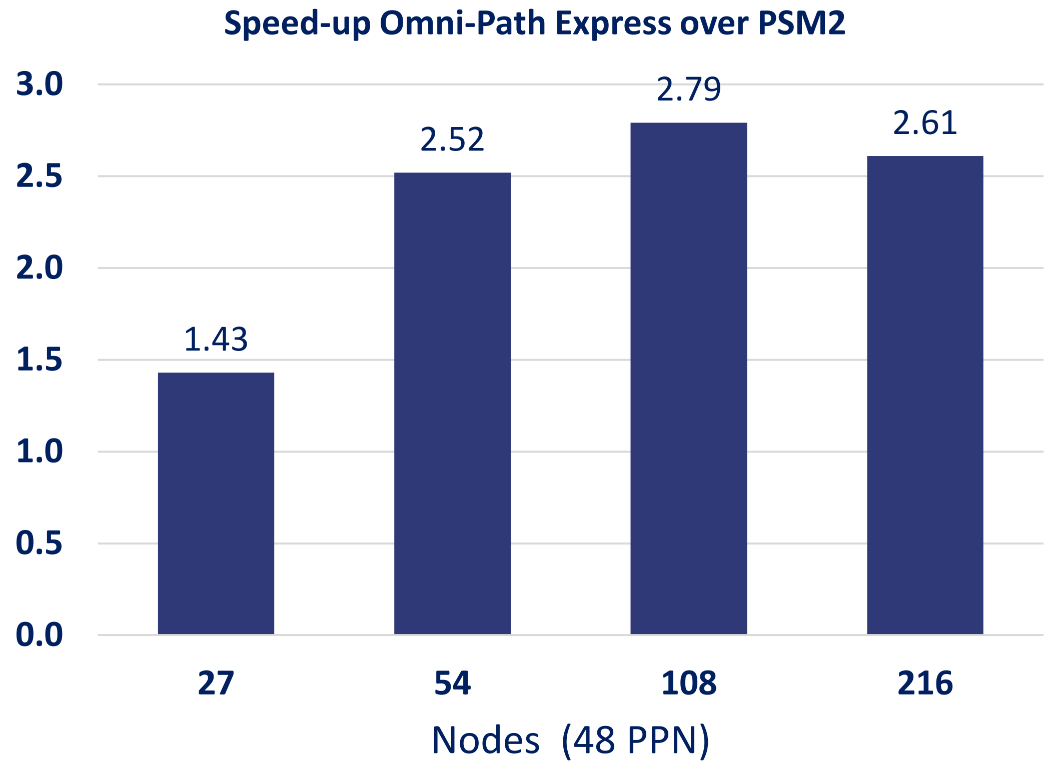
- Optimized incoming packet processing (Do a single MPI_Recv(...))
  - Intel SDE testing shows tremendous improvement in instruction count
  - Significant improvements in cache line footprint

	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

Early Customer Adoption

- Zuse Institute Berlin
- "Lise" System
  - 1270 Nodes with Omni-Path Interconnect
  - Test runs with OPX up to 100+ nodes
    - 6 real-world applications
    - 2 synthetic benchmarks
    - 10-20% improvements!



OpenFOAM, potentialFOAM solver:
Speed-up Omni-Path Express over PSM2
(weak scaling experiment)

See 2022 Hyperion HPC Forum for more details from NHR@ZIB!

Omni-Path Express: Current Status

- Beta: Upstream and accepted into Libfabric:main,
  - Will be in Libfabric release v1.15
  - Focused on small messages and latency improvements first
  - More updates coming to main, OPX under active development
- Breakthrough performance characteristics on current generation platforms
  - AMD Milan and Intel Ice Lake
- Working to upstream provider defaults for MPICH and Open MPI
- DAOS support under development and in testing
- Full GA Coming Soon!
  - Included in Cornelis OPXS software suite (formerly IFS)
  - Dedicated to upstream first development to Libfabric
    - Active and engaged in community
    - Get Involved: Happy to take patches via GitHub!

Support Targeted by Omni-Path Libfabric Provider

OPENFABRICS ALLIANCE

Communication APIs	Storage	AI/ML	GPU
Open MPI intel Intel MPI MVAPICH MPICH Sandia OpenSHMEM Charm++ CHAPEL GASNet	daos Intel	TensorFlow PyTorch	NVIDIA intel AMD

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

OPENFABRICS ALLIANCE

CORNELIS™ NETWORKS

2022 OFA Virtual Workshop

THANK YOU

Tim Thompson & Dennis Dalessandro

Cornelis Networks