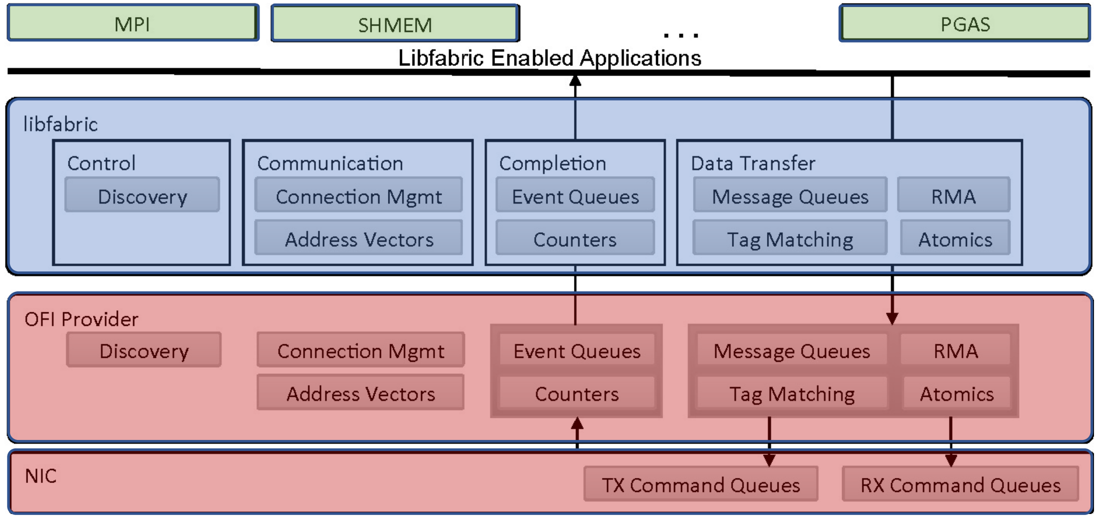
Cornelis Networks OPX Libfabric provider

- OPX is a new Libfabric provider for Omni-Path 100 HPC fabrics
- OPX was started as a fork of the BGQ Libfabric provider and code from PSM2 (register defs, driver calls, etc). The core hardware logic is largely written from scratch.
- OPX is a pure implementation of a Libfabric provider, and has no external dependencies
- OPX makes no assumptions when provider Caps are exchanged, some middlewares had bugs…MPICH did not.
- Not yet up-streamed to Libfabric. Pull request coming soon. Target release is Libfabric 1.15 (March 2022)



MPI | SHMEM | . . . | PGAS

Libfabric Enabled Applications

libfabric

Control: Discovery

Communication: Connection Mgmt, Address Vectors

Completion: Event Queues, Counters

Data Transfer: Message Queues, RMA, Tag Matching, Atomics

OFI Provider: Discovery, Connection Mgmt, Address Vectors, Event Queues, Counters, Message Queues, RMA, Tag Matching, Atomics

NIC: TX Command Queues, RX Command Queues

OPX Features and Design Considerations

- PSM2 is the original user space Api for Omni-Path. Lineage goes back to previous InfiniBand products and has a long history.
- OPX on Libfabric is a drop-in replacement for PSM2, is easy to use for any Omni-Path 100 fabric, uses an unmodified kernel module (hfi1), and requires no driver tuning (may require a few app-level ENVs)
- Reduced message latency and improved bandwidth compared to PSM2 for messages under 8K
- 0-16 byte messages have data payload delivered in the packet header instead of an additional EAGER payload packet
- Hybrid Solicitation-based reliability protocol with some pre-emptive ACKs
- Onload (rank-level) and Offload (node-level) reliability models. Onload is default, Offload is incomplete
- Less intrusive to the hosted HPC application in terms of instruction count and cache-line footprint
- Supports RMA/One-sided messages, works with OpenSHMEM (passes test bucket)
- Supports FabricDirect (Compiled MPICH with OPX and FabricDirect, took over 70 Gigs of memory)
- Performant implementation of intra-node comms via Lock-free queuing

OPX Current Status and Limitations

- No Bulk-transfer (SDMA) support. Non-performant for message lengths over 16K
- No multi-packet Eager. Latency spikes at just under 8k for high-core count systems
- In-node scaling issues related to use of the PCIe bus. OPX has not yet been tuned with respect to PCIe accesses. Especially apparent on older hardware like Broadwell.
- Reliability protocol needs to be tuned; we can still get more performance!
- GPU support is in development.
- Supports only the leanest, fastest Libfabric caps (like FI_MR_SCALABLE, FI_PROGRESS_MANUAL). Support for rich user features is still in development. FI_THREADSAFE support.
- Stable on 4 nodes. Initial testing on larger fabrics has OPX looking decent, but still much testing to do.
- Passes most of the MPICH test bucket with IMPI, MPICH, and OpenMPI.
- Alpha-level provider, still has issues and is under active development. **PSM2 should be used by all users who want support and stability**.