12th ANNUAL WORKSHOP 2016

OPENFABRICS INTERFACES: PAST, PRESENT, AND FUTURE

Sean Hefty
OFIWG Co-Chair
[ April 5th, 2016 ]
**Optimized SW path to HW**
- Minimize cache and memory footprint
- Reduce instruction count
- Minimize memory accesses

**Expand open source community**
- Inclusive development effort
- App and HW developers

**Software interfaces aligned with application requirements**
- Careful analysis of requirement

**Implementation Agnostic**
- Good impedance match with multiple fabric hardware
  - InfiniBand*, iWarp, RoCE, Ethernet, UDP offload, Intel®, Cray*, IBM*, others

**Open Source**

**Application-Centric**

**Scalable**
Give us a high-level interface!

Give us a low-level interface!

MPI developers

OFI strives to meet both requirements
OFI SOFTWARE DEVELOPMENT STRATEGIES
One Size Does Not Fit All

Application

App uses OFI features

Provider

Provider optimizes for OFI features

Fabric Services

Application

Common optimization for all apps/providers

Provider

Provider supports low-level features only

Application

App optimizes based on supported features

Provider

OFI

Provider supports low-level features only
OFI DEVELOPMENT STATUS

Application

App uses OFI features

App optimizes based on supported features

libfabric

Common optimization for all apps/providers

Provider

Provider optimizes for OFI features

Provider supports low-level features only

Many apps

Few apps

Provider’s choice

Fabric Services

Many apps

Few apps

App uses OFI features

App optimizes based on supported features

libfabric

Common optimization for all apps/providers

Provider

Provider optimizes for OFI features

Provider supports low-level features only

Fabric Services
Because of the OFI-provider gap, not all apps work with all providers.

* Other names and brands may be claimed as the property of others
LIBFABRIC SCALABILITY

Developed to evaluate the Aurora software stack at scale and assist applications in the transition from Mira to Aurora

Native provider implementation that directly uses the Blue Gene/Q hardware and network interfaces for communication

By Courtesy Argonne* National Laboratory, CC BY 2.0, https://commons.wikimedia.org/w/index.php?curid=24653857

* Other names and brands may be claimed as the property of others
LIBFABRIC SCALABILITY

IBM MPICH / PAMI
- IBM XL C compiler for BG, v12.1
- Optimized for single-threaded latency
- ./comm/xl.legacy.ndebug/bin/mpicc
- v1r2m2

MPICH / CH4 / libfabric
- gcc 4.4.7
- global locks, inline, direct, etc.
- Provider not optimized for performance

32 nodes on ALCF Vesta machine

Completely subjective software stack comparison

- PAMI
- PAMID
- MPICH
- CH4 OFI
- libfabric
- BG/Q Provider
- hardware
**LIBFABRIC SCALABILITY**

**OSU* MPI Performance Tests v5.0**

### Latency (us)

<table>
<thead>
<tr>
<th>Bytes</th>
<th>IBM</th>
<th>OFI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>1.5</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>64</td>
<td>8</td>
<td>6.5</td>
</tr>
<tr>
<td>512</td>
<td>16</td>
<td>12</td>
</tr>
<tr>
<td>4096</td>
<td>32</td>
<td>24</td>
</tr>
</tbody>
</table>

### Bandwidth (MB/s)

<table>
<thead>
<tr>
<th>Bytes</th>
<th>IBM</th>
<th>OFI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>64</td>
<td>100</td>
<td>50</td>
</tr>
<tr>
<td>512</td>
<td>1000</td>
<td>500</td>
</tr>
<tr>
<td>4096</td>
<td>32768</td>
<td>16384</td>
</tr>
</tbody>
</table>

### Messages/s

<table>
<thead>
<tr>
<th>Bytes</th>
<th>IBM</th>
<th>OFI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10000000</td>
<td>10000000</td>
</tr>
<tr>
<td>8</td>
<td>1000000</td>
<td>1000000</td>
</tr>
<tr>
<td>64</td>
<td>100000</td>
<td>100000</td>
</tr>
<tr>
<td>512</td>
<td>10000</td>
<td>10000</td>
</tr>
<tr>
<td>4096</td>
<td>1000</td>
<td>1000</td>
</tr>
<tr>
<td>32768</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

**MPI scale out testing:**
- cpi – 1M ranks,
- ISx benchmark – 0.5M ranks

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit [http://www.intel.com/performance](http://www.intel.com/performance).

* Other names and brands may be claimed as the property of others
ADDRESSING THE OFI-PROVIDER GAP

**Libfabric Framework**

- **Provider Services**
  - Logging
  - Environment variables

- **Components**
  - templates, lists, rbtree, hash table, free pool, ring buffer, stack, ...

- **Base Class Implementations**
  - fabric, domain, EQ, wait sets, AV, CQ, ... SHM primitives

**Utility Provider**

- Interface ‘extensions’ – for consistency
- Assist in provider development
- Enhance core provider

**Core Provider**

libfabric API
UTILITY PROVIDER

Performance is a primary objective

Layered over simpler core endpoints

E.g. RDM over DGRAM

Minimal functionality

Optional functionality

Requested application modes

Optional application modes

Core Provider

Utility Provider

Interface

Msg
Tagged
RMA
Atomic
Scalable EP
Shared RX
Shared TX
Multicast
FI_MULTI_RECV
FI_REMOTE_CQ_DATA
FI_INJECT
FI_RMA_EVENT
FI_SOURCE
FI_DIRECTED_RECV

Caps

Msg
Tagged
RMA
Atomic
Scalable EP
Shared RX
Shared TX
Multicast
FI_MULTI_RECV
FI_REMOTE_CQ_DATA
FI_INJECT
FI_RMA_EVENT
FI_SOURCE
FI_DIRECTED_RECV

Attrs

FI_MR_BASIC
FI_MR_SCALABLE
FI_CONTEXT
FI_LOCAL_MR
FI_MSG_PREFIX
FI_ASYNC_IOV
FI_RX_CQ_DATA

Modes

Performance is a primary objective

Layered over simpler core endpoints

E.g. RDM over DGRAM

Minimal functionality

Optional functionality

Requested application modes

Optional application modes

Core Provider

Utility Provider
Beyond HPC

Sockets – TCP/UDP

NetworkDirect

Beyond Linux*

Beyond Linux*

Enterprise, Cloud, Storage (NVM)

Stronger engagement with these communities

* Other names and brands may be claimed as the property of others
TARGET SCHEDULE

- Driven by implementation feedback
- Improve error handling, flow control
- Better support for non-traditional fabrics
- Optimize completion handling
- Address deferred features

- RDM over DGRAM Util
- RDM over MSG Util
- Shared Memory

- Utility provider is ongoing

- New Core Providers
  - Traditional and non-traditional RDMA providers
  - ABI 1.1

<table>
<thead>
<tr>
<th>2016</th>
<th>2017</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

OpenFabrics Alliance Workshop 2016
SUMMARY

- OFIWG development model working well
- Interest in OFI and libfabric is high
- Growing community
- Significant effort being made to simplify the lives of developers
  - Applications and providers

OFI is so good
LEGAL DISCLAIMER & OPTIMIZATION NOTICE

- No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document. Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade. This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps. The products and services described may contain defects or errors known as errata which may cause deviations from published specifications. Current characterized errata are available on request.

- Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, 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.

- Copyright © 2016, Intel Corporation. All rights reserved. Intel, Pentium, Xeon, Xeon Phi, Core, VTune, Cilk, and the Intel logo are trademarks of Intel Corporation in the U.S. and other countries.

- *Other names and brands may be claimed as the property of others

---

Optimization Notice

Intel’s compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Notice revision #20110804
THANK YOU

Sean Hefty

OFIWG Co-Chair