AN OVERVIEW OF OFI UTILITY PROVIDERS: IMPLEMENTATION AND USES

Alexia Ingerson
Intel® Corporation
March 2019
STRUCTURE OVERVIEW

OpenFabrics Alliance Workshop 2019
RXM PROVIDER
RXM PROVIDER

RDM EP with full msg, tagged, RMA, and atomic functionality

Targets verbs, tcp, and propriety hardware with reliable-connected semantics

Uses core provider messaging and RMA

MR registration considerations

Handles connection management

Targets a few hundred nodes
On-demand connections:

**RXM EP**
- MSG EP 0
  - fi_send()
  - fi_connect()
  - fi_accept()
- MSG EP 1
  - fi_send()
  - fi_connect()
  - fi_accept()
- RDM EP
  - fi_listen()

**RXM EP**
- MSG EP 0
  - fi_listen()
  - fi_connect()
  - fi_accept()
- PEP
  - fi_send()
- RDM EP
  - fi_addr 0
  - fi_addr 1
  - fi_ep_bind()
Simultaneous connections:

```
fi_sockaddr_in://192.168.0.111:8090
fi_send()
fi_connect()
fi_sockaddr_in://192.168.0.112:8090
fi_send()
fi_reject()
```

```
fi_accept()
fi_connect()
fi_sockaddr_in://192.168.0.111:8090
fi_sockaddr_in://192.168.0.112:8090
```
RXM PROTOCOLS

Eager

TX BUFFER

RX BUFFER

Pre-registered tx/rx MRs

Rendezvous

Register MR

TX BUFFER

RX BUFFER

Close MR

fi_inject

fi_send

TX

RX

Register MR

Register MR

Close MR

REQ

READ

ACK

fi_send

SAR
RXD PROVIDER

RDM EP with full msg, tagged, RMA, and atomic functionality

Segmentation + core provider messaging

Handles per-peer flow control (packet tracking + retrying)

RXD EP

Targets verbs UD, udp, and propriety hardware with datagram semantics

MR registration considerations

Targets +1000 nodes
RXD PROTOCOLS

Eager

TX BUFFER

Pre-registered tx/rx MRs

Time and retry packets if unacked

RX BUFFER

SAR

TX BUFFER

TX BUFFER

RX BUFFER

RX BUFFER

TX BUFFER

TX BUFFER

RX BUFFER

RX BUFFER

TX BUFFER

TX BUFFER

RX BUFFER

RX BUFFER

TX BUFFER

TX BUFFER

RX BUFFER

RX BUFFER

TX BUFFER

TX BUFFER

RX BUFFER

RX BUFFER
OUT OF ORDER REASSEMBLY

TX BUFFER

0 1 2 3 4 5 6 7 8

RX BUFFER

0 1 3 2 4 7 6 5 8

ACK 3

ACK 7

OpenFabrics Alliance Workshop 2019
RDM EP with full msg, tagged, RMA, and atomic functionality

Stand alone provider for local communication

Per-endpoint shared memory region

Disabled on non-Linux platforms (CMA)

RDM EP with full msg, tagged, RMA, and atomic functionality

Disabled on non-Linux platforms (CMA)

Stand alone provider for local communication

Per-endpoint shared memory region
SHARED MEMORY STRUCTURE

**Per-endpoint shared memory region**

- Single command queue for incoming messages
- Response queue for messages requiring an ACK
- Pool of bounce buffers for medium-sized messages
- Peer map for address exchange

**smr_region**
EP initialized info / resources

- Command Queue
- Response Queue
- Inject Pool
- Peer Map
### SHM ADDRESSING

#### Endpoint name: 11111:0:0

<table>
<thead>
<tr>
<th>0</th>
<th>22222</th>
<th>UNSPEC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>33333</td>
<td>UNSPEC</td>
</tr>
<tr>
<td>2</td>
<td>44444</td>
<td>UNSPEC</td>
</tr>
</tbody>
</table>

#### Endpoint name: 22222:0:0

<table>
<thead>
<tr>
<th>0</th>
<th>11111</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>33333</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>44444</td>
<td>1</td>
</tr>
</tbody>
</table>

#### Endpoint name: 33333:0:0

<table>
<thead>
<tr>
<th>0</th>
<th>11111</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>22222</td>
<td>UNSPEC</td>
</tr>
<tr>
<td>2</td>
<td>44444</td>
<td>2</td>
</tr>
</tbody>
</table>

#### Endpoint name: 44444:0:0

<table>
<thead>
<tr>
<th>0</th>
<th>11111</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>22222</td>
<td>UNSPEC</td>
</tr>
<tr>
<td>2</td>
<td>33333</td>
<td>UNSPEC</td>
</tr>
</tbody>
</table>
SHM PROTOCOLS

Inline

Rx Command Queue

CMA BUFF

RESP

Inject

Rx Command Queue

IOV

Rx Command Queue

BUFF
SHM RMA FAST PATH

**ssize_t process_vm_writev(pid_t pid,**

```c
const struct iovec *local_iov,
unsigned long liovcnt,
const struct iovec *remote_iov,
unsigned long riocnt,
unsigned long flags);
```

**Requirements for fast path:**

- ✓ mr_mode: FI_MR_VIRT_ADDR
- ✓ no RMA ordering requested (RAR, RAW, RAS, WAR, WAW, WAS, SAR, SAW)
- ✓ message is not sending remote CQ data
FUTURE ENHANCEMENTS

RXM
- Reduce header overhead
- MR cache improvements
- Improve start-up times
- Add HW-specific optimizations
- Rendezvous read -> write

MR
- Threading, progress
- Performance
- Non-host memory transfers
- xpmem

Mrail
- Under development

SHM

RXD
- Threading, progress
- Performance
- Packet buffering
- Scaling, flow control
- Reliable reassembly
THANK YOU

Alexia Ingerson
Intel Corporation
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. For more complete information visit www.intel.com/benchmarks.

INFORMATION IN THIS DOCUMENT IS PROVIDED “AS IS”. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO THIS INFORMATION INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

Copyright © 2018, 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.

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