Scaling OFA: Beyond RDMA?

William Magro
Director & Chief Technologist
Technical Computing Software
Intel Software and Services Group

#OFADevWorkshop
HPC – Powering Breakthroughs

- Energy Exploration
- Computational Race
- Medical Imaging
- Climate Weather Modeling
- Scientific Research
- Security
- CAD Manufacturing
- Financial Analyses
- Digital Content Creation
OFA: Powering HPC

OFED now powers 42% of Top500

Image source: openfabrics.org, top500.org
OFA: Making Fast Fabrics Accessible

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>Alliance is formed. Initial support on InfiniBand.</td>
</tr>
<tr>
<td>2005</td>
<td>OFA software in Linux kernel. Major foundation for common Linux stack, elimination of proprietary stacks</td>
</tr>
<tr>
<td>2006</td>
<td>OFED develop starts. 10GbE iWARP support added.</td>
</tr>
<tr>
<td>2007</td>
<td>First OFED version released. OFED available in major Linux OS distributions</td>
</tr>
<tr>
<td>2008</td>
<td>OFA and UNH-IOL interoperability test events held. Expansion to commercial and datacenter apps.</td>
</tr>
<tr>
<td>2009</td>
<td>Fourth OFED version released. Major enhancements to Linux and Windows OFED.</td>
</tr>
<tr>
<td>2010</td>
<td>Fifth OFED version released. Support for 10GbE RoCE added.</td>
</tr>
</tbody>
</table>

- Fast fabrics offer distinct value vs. traditional networks
- Linux community accepts the notion of fast fabrics
Is it Really About RDMA?

What makes a fast fabric?

– High peak bandwidth
– Low small-message latency
– Direct application access with kernel bypass
– Multiple, protected inter-application “channels”
– Data movement engine

RDMA is *one* capability of *some* fast fabrics
MPI on RDMA

What MPI wants to do:

<table>
<thead>
<tr>
<th>Sender</th>
<th>Receiver</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

MPI Semantic

What MPI does on RDMA:

Simple Eager Message

Sender | Receiver
---|---

Rendezvous

Sender | Receiver
---|---

Request

READ

Ack

Software overheads a growing concern
And What About…

- Unexpected message handling
- Tag matching
- Connection management
- Memory footprint
- Memory registration

- Collectives
- Non-contiguous message
- Unnecessary software complexity for consumers
- MPI3
- PGAS

Reaching Exascale demands extreme efficiency: in performance and in energy

A semantic mismatch between the most common fast fabric (RDMA) and is most common use (MPI)
What if…

• We had application centric APIs?
• Software led, rather than lagged, hardware?
• OpenFabrics led the way?

Would hardware follow the software?
What if…OpenFabrics Led?

– OFA is the framework for fast fabrics in Linux
– It’s the framework that is valuable
  • RDMA is one key capability
– OFA is the focal point for today’s fast fabric producers and consumers
– OFA has addressed and can prevent a fragmented fabric ecosystem
– A “software first” approach helps solidify requirements
– Extant software interfaces can lead hardware
– Standards could target known, valuable capabilities
– Hardware vendors gain flexibility in implementation
Can OFA Get Us There?

- Existing framework provides the infrastructure
Can OFA Get Us There?

- Existing framework provides the infrastructure
- Prior work shows value of app-centric APIs
  - MX, MXM, PSM, TSM, Portals, etc.
- Prior work shows app-centric APIs can coexist with RDMA
  - DAPL extension framework
- Discoverability of fabric capabilities is key
- Long-lived interfaces are critical
Where to Start?

- **Message Passing Interface**
  - MPI is the most prevalent usage of RDMA today
  - MPI3 standard just released adds new semantics
- **Partitioned Global Address Space (PGAS)**
  - Growing relevance and interest for HPC
- **Parallel file systems**
  - Use RDMA for block storage
  - Could benefit from tailored APIs for file/object access
- **Use cases beyond HPC**
  - Database, Hadoop, Big Data, Financial Services, O&G

Identify common “Communication Building Blocks”
Let’s Get Started!

– Build on OFA’s established industry position

– Embrace software that leads versus lags hardware

– Start developing application centric APIs

– Start developing software implementations of new fabric capabilities

Keep OFA the “Home for Fast Fabrics”