UCX: An Open Source Framework for HPC Network APIs and Beyond

Presented by: Pavel Shamis / Pasha

ORNL is managed by UT-Battelle for the US Department of Energy
Co-Design Collaboration

The Next Generation HPC Communication Framework

Collaborative Effort

Industry, National Laboratories and Academia
Challenges

• Performance Portability (across various interconnects)
  – Collaboration between industry and research institutions
    • …but mostly industry (because they built the hardware)

• Maintenance
  – Maintaining a network stack is time consuming and expensive
  – Industry have resources and strategic interest for this

• Extendibility
  – MPI+X+Y ?
  – Exascale programming environment is an ongoing debate
## Challenges (CORAL)

<table>
<thead>
<tr>
<th>Feature</th>
<th>Summit</th>
<th>Titan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Application Performance</td>
<td>5-10x Titan</td>
<td>Baseline</td>
</tr>
<tr>
<td>Number of Nodes</td>
<td>~3,400</td>
<td>18,688</td>
</tr>
<tr>
<td>Node performance</td>
<td>&gt; 40 TF</td>
<td>1.4 TF</td>
</tr>
<tr>
<td>Memory per Node</td>
<td>&gt;512 GB (HBM + DDR4)</td>
<td>38GB (GDDR5+DDR3)</td>
</tr>
<tr>
<td>NVRAM per Node</td>
<td>800 GB</td>
<td>0</td>
</tr>
<tr>
<td>Node Interconnect</td>
<td>NVLink (5-12x PCIe 3)</td>
<td>PCIe 2</td>
</tr>
<tr>
<td>System Interconnect (node injection bandwidth)</td>
<td>Dual Rail EDR-IB (23 GB/s)</td>
<td>Gemini (6.4 GB/s)</td>
</tr>
<tr>
<td>Interconnect Topology</td>
<td>Non-blocking Fat Tree</td>
<td>3D Torus</td>
</tr>
<tr>
<td>Processors</td>
<td>IBM POWER9, NVIDIA Volta™</td>
<td>AMD Opteron™, NVIDIA Kepler™</td>
</tr>
<tr>
<td>File System</td>
<td>120 PB, 1 TB/s, GPFS™</td>
<td>32 PB, 1 TB/s, Lustre®</td>
</tr>
<tr>
<td>Peak power consumption</td>
<td>10 MW</td>
<td>9 MW</td>
</tr>
</tbody>
</table>
UCX – Unified Communication X Framework

• Unified
  – Network API for multiple network architectures that target HPC programing models and libraries

• Communication
  – How to move data from location in memory A to location in memory B considering multiple types of memories

• Framework
  – A collection of libraries and utilities for HPC network programmers
History

MXM
- Developed by Mellanox Technologies
- HPC communication library for InfiniBand devices and shared memory
- Primary focus: MPI, PGAS

UCCS
- Developed by ORNL, UH, UTK
- Originally based on Open MPI BTL and OPAL layers
- HPC communication library for InfiniBand, Cray Gemini/Aries, and shared memory
- Primary focus: OpenSHMEM, PGAS
- Also supports: MPI

PAMI
- Developed by IBM on BG/Q, PERCS, IB VERBS
- Network devices and shared memory
- MPI, OpenSHMEM, PGAS, CHARM++, X10
- C++ components
- Aggressive multi-threading with contexts
- Active Messages
- Non-blocking collectives with hw acceleration support

Decades of community and industry experience in development of HPC software
What we are doing differently...

• UCX **consolidates** multiple industry and academic efforts
  – Mellanox MXM, IBM PAMI, ORNL/UTK/UH UCCS, etc.

• Supported and maintained by industry
  – IBM, Mellanox, NVIDIA, Pathscale
What we are doing differently...

• Co-design effort between national laboratories, academia, and industry

    Co-design

    Applications: LAMMPS, NWChem, etc.

    Programming models: MPI, PGAS/Gasnet, etc.

    Middleware:

    Driver and Hardware
UCX: An Open Source Framework for HPC Network APIs and Beyond
A Collaboration Efforts

- Mellanox co-designs network interface and contributes MXM technology
  - Infrastructure, transport, shared memory, protocols, integration with OpenMPI/SHMEM, MPICH

- ORNL co-designs network interface and contributes UCCS project
  - InfiniBand optimizations, Cray devices, shared memory

- NVIDIA co-designs high-quality support for GPU devices
  - GPUDirect, GDR copy, etc.

- IBM co-designs network interface and contributes ideas and concepts from PAMI

- UH/UTK focus on integration with their research platforms
Licensing

• Open Source
  – BSD 3 Clause license
  – Contributor License Agreement – BSD 3 based
## UCX Framework Mission

- Collaboration between industry, laboratories, and academia
- Create open-source production grade communication framework for HPC applications
- Enable the highest performance through co-design of software-hardware interfaces
- Unify industry - national laboratories - academia efforts

<table>
<thead>
<tr>
<th>API</th>
<th>Performance oriented</th>
<th>Production quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposes broad semantics that target data centric and HPC programming models and applications</td>
<td>Optimization for low-software overheads in communication path allows near native-level performance</td>
<td>Developed, maintained, tested, and used by industry and researcher community</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Community driven</th>
<th>Research</th>
<th>Cross platform</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collaboration between industry, laboratories, and academia</td>
<td>The framework concepts and ideas are driven by research in academia, laboratories, and industry</td>
<td>Support for Infiniband, Cray, various shared memory (x86-64 and Power), GPUs</td>
</tr>
</tbody>
</table>

Co-design of Exascale Network APIs
Architecture
UCX Framework

**UC-P for Protocols**
High-level API uses UCT framework to construct protocols commonly found in applications

**Functionality:**
Multi-rail, device selection, pending queue, rendezvous, tag-matching, software-atomics, etc.

---

**UC-T for Transport**
Low-level API that expose basic network operations supported by underlying hardware. Reliable, out-of-order delivery.

**Functionality:**
Setup and instantiation of communication operations.

---

**UC-S for Services**
This framework provides basic infrastructure for component based programming, memory management, and useful system utilities

**Functionality:**
Platform abstractions, data structures, debug facilities.
A High-level Overview

UCX: An Open Source Framework for HPC Network APIs and Beyond

UCX

Applications

UCP (Protocols) - High Level API
Transport selection, cross-transport multi-rail, fragmentation, operations not supported by hardware

- Message Passing API Domain: tag matching, rendezvous
- PGAS API Domain: RMA, Atomics
- Task Based API Domain: Active Messages
- I/O API Domain: Stream

UC-T (Hardware Transports) - Low Level API
RMA, Atomic, Tag-matching, Send/Recv, Active Message

- Transport for InfiniBand VERBs drivers: RC, UD, XRC, DCT
- Transport for Gemini/Aries drivers: GNI
- Transport for intra-node host memory communication: SYSV, POSIX, KNEM, CMA, XPMEM
- Transport for Accelerator Memory communication: GPU

UC-S (Services)
Common utilities
- Utilities
- Data structures
- Memory Management

Applications

MPICH, Open-MPI, etc.
OpenSHMEM, UPC, CAF, X10, Chapel, etc.
Parsec, ORC, Legions, etc.
Burst buffer, ADIOS, etc.

Hardware

OFA Verbs Driver
Cray Driver
OS Kernel
Cuda

UCX: An Open Source Framework for HPC Network APIs and Beyond
Preliminary Evaluation (UCT)

- Two HP ProLiant DL380p Gen8 servers
- Intel Xeon E5-2697 2.7GHz CPUs
- Mellanox SX6036 switch
- Single-port Mellanox Connect-IB FDR (10.10.5056)
- Mellanox OFED 2.4-1.0.4. (VERBS)
- Prototype implementation of Accelerated VERBS (AVERBS)
OpenSHMEM and OSHMEM (OpenMPI) Put Latency (shared memory)

Lower is better
OpenSHMEM and OSHMEM (OpenMPI) Put Injection Rate

Higher is better

Message Rate (put operations / second)

Higher is better

Connect-IB

Message Size

UCX: An Open Source Framework for HPC Network APIs and Beyond
OpenSHMEM and OSHMEM (OpenMPI) GUPs Benchmark

Higher is better

Connect-IB

UCX (mlx5)
OSHMEM (mlx5)
MPICH - Message rate – Preliminary Results

Connect-IB

“non-blocking tag-send”

Slide courtesy of Pavan Balaji, ANL - sent to the ucx mailing list
Where is UCX being used?

- Upcoming release of Open MPI 2.0
- Upcoming release of MPICH
- OpenSHMEM reference implementation
- PARSEC – runtime used on Scientific Linear Libraries
What Next?

• UCX Consortium!

• UCX Specification
  – Early draft is available online: [url: http://www.openucx.org/early-draft-of-ucx-specification-is-here/]

• Production releases
  – MPICH, Open MPI, Open SHMEM(s), Gasnet, and more…

• Support for more networks and applications and libraries

• UCX Hackathon 2016!
  – Will be announced on the mailing list and website
https://github.com/orgs/openucx

WEB: www.openucx.org
Contact: info@openucx.org

Mailing List:
https://elist.ornl.gov/mailman/listinfo/ucx-group
ucx-group@elist.ornl.gov
Acknowledgments
Acknowledgments

• Thanks to all our partners!
Questions?

Unified Communication - X Framework

WEB: www.openucx.org
Contact: info@openucx.org
    : https://github.com/orgs/openucx

Mailing List:
https://elist.ornl.gov/mailman/listinfo/ucx-group
ucx-group@elist.ornl.gov