The Race Towards
Co-Design Architecture For Next Generation HPC Systems

Gilad Shainer
HPC Advisory Council
Spain, September 2015
Global Perspective
The Race Towards Exascale

Performance

- Terascale
- Petascale
- Exascale

Technology Development

- SMP to Clusters
- Single-Core to Multi-Core

Image: "Roadrunner"
The Road to Exascale Computing
Co-Design Architecture – From Discrete to System Focused
Exascale will be Enabled via Co-Design Architecture

- Software – Hardware
- Hardware – Hardware (e.g. GPU-Direct)
- Software – Software (e.g. OpenUCX)
- Industry – Users – Academia

Standard, Open Source, Eco-System
Programmable, Configurable, Innovative
Software-Hardware Co-Design? Example: Breaking the Latency Wall

- Today: Network devices are in 100ns latency today
- Challenge: How to enable the next order of magnitude performance improvement?
- Solution: Co-Design - mapping the communication frameworks on all active devices
- Result: reduce HPC communication frameworks latency by an order of magnitude

Co-Design Architecture Paves the Road to Exascale Performance
The Future of Latency – Application Level

- Migrating complete operations from the software / CPU to the Network
- Hardware – Software co-design
The Road to Exascale – Co-Design System Architecture

- The road to Exascale requires order of magnitude performance improvements
- Co-Design architecture enables all active devices to become co-processors

Mapping Communication Frameworks on All Active Devices
The Elements of the Co-Design Architecture

Offloading

- The Elements of the Co-Design Architecture

Applications (Innovations, Scalability, Performance)

Communication Frameworks (MPI, SHMEM/PGAS)

Offloading Technologies: In-Network Computing

- Software-Defined X
- Direct Communications
- Flexibility
- RDMA
- GPUDirect
- Programmability
- Heterogeneous System
- Virtualization
- Backward and future Compatibility

Co-Design Implementation Via Offloading Technologies
The Next Generation
HPC Software Framework
Collaborative Effort
Industry, National Laboratories and Academia
What We Don’t Want to Achieve

How Standards Proliferate:

Situation: There are 14 competing standards.

14?! Ridiculous! We need to develop one universal standard that covers everyone’s use cases. Yeah!

Soon:

Situation: There are 15 competing standards.
Different Model - Co-Design Effort

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

- Applications: LAMMPS, NWChem, etc.
- Programming models: MPI, PGAS/Gasnet, etc.
- Middleware: UCX
- Driver and Hardware
UCX Framework Mission

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

API
Exposes broad semantics that target data centric and HPC programming models and applications

Performance oriented
Optimization for low-software overheads in communication path allows near native-level performance

Production quality
Developed, maintained, tested, and used by industry and researcher community

Community driven
Collaboration between industry, laboratories, and academia

Research
The framework concepts and ideas are driven by research in academia, laboratories, and industry

Cross platform
Support for Infiniband, Cray, various shared memory (x86-64 and Power), GPUs

Co-design of Exascale Network APIs
The UCX Framework

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

Functionality: Platform abstractions and data structures

**UC-T for Transport**
Low-level API that expose basic network operations supported by underlying hardware

Functionality: work request setup and instantiation of operations

**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.
UCX High-level Overview

Applications

- MPI
- OpenSHMEM, UPC, CAF, X10, Chapel, etc.
- Parsec, OCR, Legions, etc.
- Burst buffer, ADIOS, etc.

UC-P (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:
RMAs, 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 driver
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

OFA Verbs Driver
Cray Driver
OS Kernel
Cuda

Hardware
Collaboration

- **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**
  - GPU-Direct, GDR copy, etc.

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

- **UH/UTK focus on integration with their research platforms**
UCX Information

Unified Communication - X Framework

WEB:
www.openucx.org

Mailing List:
https://elist.orl.gov/mailman/listinfo/ucx-group
ucx-group@elist.orl.gov
GPUDirect RDMA

GPU-GPU Internode MPI Latency

- Lower is Better
- 8.6X
- 2.37 usec

GPU-GPU Internode MPI Bandwidth

- Higher is Better
- 10X Increase in Throughput

88% Lower Latency

Source: Prof. DK Panda
GPUDirect RDMA and Sync (3.0 and 4.0)

- **Hardware – Hardware co-design**

![Diagram showing GPUDirect RDMA and Sync](image)

**HOOMD-blue Performance**
*(LJ Liquid Benchmark, 512K Particles)*

- **Network Performance Chart**
  - **Without GPUDirect RDMA**
  - **GPUDirect RDMA**

- **Universities and Organizations**
  - University of Cambridge
  - Lockheed Martin
  - NVIDIA
  - IBM
  - UCSD
  - CSCS
  - Boehringer Ingelheim
  - Fraunhofer IWM
  - UNIST
  - CERN
  - TELUS
  - NASA
  - LSI Logic
  - Northrop Grumman
  - HP
  - SpaceX
  - Dell
  - General Mills
  - U.S. Air Force
  - GE
  - Telstra
  - UC Berkeley
HPC music is an advanced research project about High Performance Computing and Music Production dedicated to enable HPC in music creation. Its goal is to develop HPC cluster and cloud solutions that further enable the future of music production and reproduction.
Thank You!

Web: www.hpcadvisorycouncil.com
Email: info@hpcadvisorycouncil.com
Facebook: http://www.facebook.com/HPCAdvisoryCouncil
Twitter: www.twitter.com/hpccouncil
YouTube: www.youtube.com/user/hpcadvisorycouncil