The Effect of In-Network Computing-Capable Interconnects on the Scalability of CAE Simulations

Ophir Maor, Gerardo Cisneros, David Cho, Yong Qin, Gilad Shainer
HPC Advisory Council
The HPC-AI Advisory Council

- World-wide HPC non-profit organization
- More than 400 member companies / universities / organizations
- Bridges the gap between HPC-AI usage and its potential
- Provides best practices and a support/development center
- Explores future technologies and future developments
- Leading edge solutions and technology demonstrations
HPC-AI Advisory Council Cluster Center (Examples)

- Supermicro / Foxconn 32-node cluster
  - Dual Socket Intel(R) Xeon(R) Gold 6138 CPU @ 2.00GHz

- Dell™ PowerEdge™ R730/R630 36-node cluster
  - Dual Socket Intel® Xeon® 16-core CPUs E5-2697A V4 @ 2.60 GHz

- IBM S822LC POWER8 8-node cluster
  - Dual Socket IBM POWER8 10-core CPUs @ 2.86 GHz
  - GPU: NVIDIA Kepler K80 GPUs
Multiple Applications Best Practices Published

- Abaqus
- ABySS
- AcuSolve
- Amber
- AMG
- AMR
- ANSYS CFX
- ANSYS FLUENT
- ANSYS Mechanical
- BQCD
- BSMBench
- CAM-SE
- CCSM
- CESM
- COSMO
- CP2K
- CPMD
- Dacapo
- Desmond
- DL-POLY
- Eclipse
- FLOW-3D
- GADGET-2
- Graph500
- GROMACS
- Himeno
- HIT3D
- HOOMD-blue
- HPCC
- HPCG
- HYCOM
- ICON
- Lattice QCD
- LAMMPS
- LS-DYNA
- miniFE
- MILC
- MSC Nastran
- MR Bayes
- MMS
- MPQC
- NAMD
- Nekbone
- NEMO
- Nekbone
- OpenFOAM
- OpenMX
- OptiStruct
- PARATEC
- PFA
- PFLOTRAN
- Quantum ESPRESSO
- RADIOSS
- SNAP
- SPECFEM3D
- STAR-CCM+
- STAR-CD
- VASP
- VSP
- WRF
Exponential Data Growth Everywhere

source: IDC
Breaking the Application Latency Wall

10 years ago
- Network: ~10 microsecond
- Communication Framework: ~100 microsecond

Today
- Network: ~0.1 microsecond
- Communication Framework: ~10 microsecond

Future
- Network: ~0.05 microsecond
- Communication Framework: ~1 microsecond
The Ever Growing Demand for Higher Performance

Terascale

Petascale

Exascale

SMP to Clusters

Single-Core to Many-Core

Co-Design

Application
Software
Hardware
Data Centric Data Center
SHARP - Scalable Aggregation and Reduction Technology

• Reliable Scalable General Purpose Primitive
  – In-network Tree based aggregation mechanism
  – Large number of groups
  – Multiple simultaneous outstanding operations

• Applicable to Multiple Use-cases
  – HPC Applications using MPI / SHMEM
  – Distributed Machine Learning applications

• Scalable High Performance Collective Offload
  – Barrier, Reduce, All-Reduce, Broadcast and more
  – Sum, Min, Max, min-loc, max-loc, OR, XOR, AND
  – Integer and Floating-Point, 16/32/64 bits
SHARP Allreduce Performance

Allreduce Latency

Latency (usec)

Cluster Size (Nodes)

SHARP - 8B
SHARP - 128B
Software - 8B
Software - 128B
SHARP Allreduce Performance

Allreduce Latency

Latency (usec)

Cluster Size (Nodes)

SHARP - 1024B
SHARP - 2048B
Software - 1024B
Software - 2048B
The Niagara Supercomputer – University of Toronto
SHARP AllReduce Performance
1500 Nodes, 60K MPI Ranks, Dragonfly+ Topology

MPI AllReduce Latency
1500 Nodes, 1PPN

- HPC-X SHARP
- Software MPI

MPI AllReduce Latency
1500 Nodes, 40PPN, 60K MPI Ranks

- HPC-X SHARP
- Software MPI
OpenFOAM

• OpenFOAM® (Open Field Operation and Manipulation) CFD

• Toolbox in an open source CFD applications that can simulate
  – Complex fluid flows involving
  – Chemical reactions
  – Turbulence
  – Heat transfer
  – Solid dynamics
  – Electromagnetics
  – The pricing of financial options

• OpenFOAM support can be obtained from OpenCFD Ltd
OpenFOAM Scalability per Interconnect Technology

OpenFOAM Scalability
(motorBike_160)

Number of Nodes

Jobs/Day

0% 20% 40% 60% 80% 100%

4 8 16 32

OmniPath  EDR InfiniBand
OpenFOAM MPI Performance

OpenFOAM
(motorBike_160)

Jobs/Day

Number of Nodes

4
8
16
32

EDR Intel MPI
EDR HPC-X
Summary

- HPC cluster environments impose high demands on connectivity throughput and low latency with low CPU overhead, network flexibility, and high efficiency.
- Fulfilling these demands enables the maintenance of a balanced system that can achieve high application performance and high scaling.
- With the increase in number of CPU cores and application threads, there is a need to develop a new HPC cluster architecture - a data-focused architecture.
- The Co-Design collaboration enables the development of In-Network Computing technology that breaks the performance and scalability barriers.
- The OpenFoam application was benchmarked for this study to demonstrate the advantages of In-Network Computing technology.
- We have witnessed nearly 30% performance advantage and linear scalability with InfiniBand In-Network Computing technology.
2018 HPC-AI Advisory Council Activities

• HPC-AI Advisory Council
  – Application best practices, case studies
  – Benchmarking center with remote access for users
  – World-wide conferences

• 2018 Conferences
  – USA (Stanford University) – February
  – Switzerland (CSCS) – April
  – Australia - August
  – Spain (BSC) – Sep
  – China (HPC China) – October

• 2018 Competitions
  – APAC HPC-AI Competition - March
  – China - 6th Annual RDMA Competition - May
  – ISC Germany - 7th Annual Student Cluster Competition - June

• For more information
  – [www.hpcadvisorycouncil.com](http://www.hpcadvisorycouncil.com)
  – info@hpcadvisorycouncil.com
Thank You