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

Ophir Maor
HPC Advisory Council
ophir@hpcadvisorycouncil.com
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 Advisory Council Members
HPC-AI Advisory Council Cluster Center (Examples)

• **Supermicro / Foxconn 32-node cluster**
  • Dual Socket Intel Gold 6138 Skylake CPUs @ 2GHz

• **Dell™ PowerEdge™ R730/R630 36-node cluster**
  • Dual Socket Intel E5-2697A V4 Broadwell CPUs @ 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
- NwChem
- Octopus
- OpenAtom
- 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

Communication Framework

~10 microsecond

~100 microsecond

Today

Network

Communication Framework

~0.1 microsecond

~10 microsecond

Future

Network

Communication Framework

~0.05 microsecond

~1 microsecond
The Ever Growing Demand for Higher Performance

Terascale

2000

SMP to Clusters

Petascale

2005

“Roadrunner”

Single-Core to Many-Core

2010

Exascale

2015

2020

Co-Design

Application
Software
Hardware

“Summit” System

“Sierra” System

HW

SW

APP

Co-Design

June 5th-7th | Cleveland, OH

nafems.org/caase18

The Conference on Advancing Analysis & Simulation in Engineering | CAASE18
Need for Intelligent Faster Interconnect

CPU-Centric (Onload)

Data-Centric (Offload)

Must Wait for the Data
Creates Performance Bottlenecks

Faster Data Speeds and In-Network Computing Enable Higher Performance and Scale

Analyze Data as it Moves!
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

Cluster Size (Nodes)

Latency (usec.)

2 4 8 16 32 64 128

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

Allreduce Latency

Cluster Size (Nodes)

Latency (usec)

SHARP - 1024B
SHARP - 2048B
Software - 1024B
Software - 2048B
The Niagara Supercomputer – University of Toronto

Dragonfly+
SHARP AllReduce Performance
1500 Nodes, 60K MPI Ranks, Dragonfly+ Topology
ANSYS Fluent

- ANSYS Fluent software is a well known CFD tool available

- Fluent includes well-validated physical modeling capabilities to deliver fast, accurate results across the widest range of CFD and multiphysics applications.
Test Setup

- Dell PowerEdge R730 servers
  - Dual socket Intel(R) E5-2697A V4 CPUs at 2.60GHz
  - Mellanox ConnectX-5 EDR 100Gb/s InfiniBand adapters
  - Memory of 256GB DDR4 2400MHz RDIMMs per node
  - 1TB 7.2K RPM SSD 2.5" hard drive per node.
- Mellanox Switch-IB 2 SB7800 36-Port 100Gb/s EDR InfiniBand
- ANSYS Fluent v19
- Red Hat Enterprise Linux 7.4
- Fluent benchmarks: ice_2m, oil_rig_7m
ANSYS Fluent - MPI Benchmark (InfiniBand EDR)

![Graph showing solver rating vs. number of nodes for different MPI implementations, indicating non-scaling behavior at 35%.]
ANSYS Fluent - MPI Benchmark (InfiniBand EDR)
ANSYS Fluent – Adapter Comparison

19

![Graph showing performance comparison between Omni-Path and EDR InfiniBand for Ansys Fluent (oil_rig_7m) and Ansys Fluent (ice_2m) for different numbers of nodes. The bars indicate performance ratings, with Omni-Path and EDR InfiniBand results compared. The graphs show improvements with increasing node counts.]

- Ansys Fluent (oil_rig_7m) with 41% improvement.
- Ansys Fluent (ice_2m) with 29% improvement.
Data Centric Data Center
Summary

- HPC environments impose high performance demands
- The Co-Design collaboration enables the development of In-Network Computing that breaks the performance and scalability barriers
- The ANSYS Fluent application was benchmarked for this study to demonstrate the advantages of In-Network Computing technology
- We have witnessed nearly 41% performance advantage and linear scalability with InfiniBand In-Network Computing technology
2018 HPC-AI Advisory Council Activities

• HPC-AI Advisory Council
  – More than 400 members
  – Application best practices, case studies
  – Benchmarking center with remote access for users

• 2018 Conferences
  – Australia (Pawsey Supercomputing Centre) - August
  – Spain (Barcelona Supercomputing Center) - September
  – China (HPC China) – October

• 2018 Competitions
  – HPC-AI APAC Competition: March - August
  – 6th Annual RDMA Competition (China): May - October
  – 7th Annual Student Cluster Competition (ISC, Germany): June

• For more information
  – www.hpcadvisorycouncil.com, info@hpcadvisorycouncil.com
Thank You!

Ophir Maor
Application Performance
HPC Advisory Council
ophir@hpcadvisorycouncil.com