NAMD
Performance Benchmark and Profiling

January 2015
Note

The following research was performed under the HPC Advisory Council activities
- Participating vendors: Intel, Dell, Mellanox
- Compute resource - HPC Advisory Council Cluster Center

The following was done to provide best practices
- NAMD performance overview
- Understanding NAMD communication patterns
- Ways to increase NAMD productivity
- MPI libraries comparisons

For more info please refer to
- http://www.dell.com
- http://www.intel.com
- http://www.mellanox.com
- http://www.ks.uiuc.edu/Research/namd/
NAMD

- A parallel molecular dynamics code that received the 2002 Gordon Bell Award
- Designed for high-performance simulation of large biomolecular systems
  - **Scales to hundreds of processors and millions of atoms**
- Developed by the joint collaboration of the Theoretical and Computational Biophysics Group (TCB) and the Parallel Programming Laboratory (PPL) at the University of Illinois at Urbana-Champaign
- NAMD is distributed free of charge with source code
Objectives

• The presented research was done to provide best practices
  – NAMD performance benchmarking
    • MPI Library performance comparison
    • Interconnect performance comparison
    • CPUs comparison
    • Compilers comparison

• The presented results will demonstrate
  – The scalability of the compute environment/application
  – Considerations for higher productivity and efficiency
Test Cluster Configuration

- **Dell PowerEdge R730 32-node (896-core) “Thor” cluster**
  - Dual-Socket 14-Core Intel E5-2697v3 @ 2.60 GHz CPUs
  - Memory: 64GB memory, DDR4 2133 MHz
  - OS: RHEL 6.5, OFED 2.3-2.0.5 InfiniBand SW stack
  - Hard Drives: 2x 1TB 7.2 RPM SATA 2.5” on RAID 1
  - Memory Snoop Mode: Cluster-on-Die
  - Turbo Mode disabled unless otherwise stated

- **Dell PowerEdge R720xd 32-node (640-core) “Jupiter” cluster**
  - Dual-Socket 10-Core Intel E5-2680v2 @ 2.80 GHz CPUs
  - Memory: 64GB memory, DDR3 1600 MHz
  - OS: RHEL 6.2, OFED 2.3-2.0.5 InfiniBand SW stack
  - Hard Drives: 24x 250GB 7.2 RPM SATA 2.5” on RAID 0

- **Mellanox ConnectX-3 QDR InfiniBand and 40GbE VPI adapters**
- **Mellanox SwitchX SX6036 VPI InfiniBand and Ethernet switches**
- **MPI: Mellanox HPC-X v1.2.0-292, Intel MPI 5.0.2.044**
- **Compilers: Intel Composer XE 2015.1.133, GNU Compilers 4.9.1**
- **Application: NAMD 2.10**
- **Benchmarks:**
  - ApoA1 benchmark (92,204 atoms, 12A cutoff)
  - Apolipoprotein A1: Models bloodstream lipoprotein particle
PowerEdge R730
Massive flexibility for data intensive operations

• **Performance and efficiency**
  – Intelligent hardware-driven systems management with extensive power management features
  – Innovative tools including automation for parts replacement and lifecycle manageability
  – Broad choice of networking technologies from GigE to IB
  – Built in redundancy with hot plug and swappable PSU, HDDs and fans

• **Benefits**
  – Designed for performance workloads
    • from big data analytics, distributed storage or distributed computing where local storage is key to classic HPC and large scale hosting environments
    • High performance scale-out compute and low cost dense storage in one package

• **Hardware Capabilities**
  – Flexible compute platform with dense storage capacity
    • 2S/2U server, 6 PCIe slots
  – Large memory footprint (Up to 768GB / 24 DIMMs)
  – High I/O performance and optional storage configurations
    • HDD options: 12 x 3.5" - or - 24 x 2.5 + 2x 2.5 HDDs in rear of server
    • Up to 26 HDDs with 2 hot plug drives in rear of server for boot or scratch
NAMD Performance – Network Interconnect

- **FDR InfiniBand outperforms 1GbE and 10GbE on every node size**
  - InfiniBand runs faster than 1GbE by 5x, 10GbE by 7x at 4 nodes / 112 MPI processes
  - Performance differences widen as the cluster scales to 32 nodes / 896 NP
  - High core count per CPU generates more network communications per node
    - Scalability issue for Ethernet beyond 2 nodes

![NAMD Performance](image)

**Thor Cluster**
28 Cores Per Node
NAMD Profiling – User/MPI calls

- **NAMD shows high usage for MPI communications**
  - With RDMA, FDR IB reduces network overhead; allows CPU to focus on computation
  - Ethernet consumes about 87-89% on computation, while FDR IB consumes 41%

### NAMD Profiling

- **(32-node, 1Gbe)**
  - **MPI/User Time Ratio**
  - 89% User Time, 11% MPI Time

- **(32-node, 10Gbe)**
  - **MPI/User Time Ratio**
  - 87% User Time, 13% MPI Time

- **(32-node, FDR IB)**
  - **MPI/User Time Ratio**
  - 59% User Time, 41% MPI Time
NAMD Profiling – MPI calls

- **Time difference among interconnects appears in MPI_Iprobe**
  - MPI_Iprobe is a non-blocking test for data exchanges among the MPI processes
  - Network throughput appears to have a direct impact on NAMD performance
  - Time spent in MPI_Iprobe reduced from 1GbE to 10GbE, and to FDR InfiniBand
GNU 4.9.1 and Intel Compilers perform better than default GNU compilers

- GNU 4.4.6 is the default compilers available in the OS
  - GCC flags: "-m64 -O3 -fexpensive-optimizations -ffast-math"
  - ICC flags: "-O3 --enable-shared --enable-threads" --enable-float --enable-type-prefix"

Higher is better

Jupiter Cluster
20 Cores Per Node
• Intel MPI and HPC-X performs roughly the same when running at scale
  – Majority of the communications involve non-blocking communications

**NAMD Performance – MPI Libraries**

**NAMD Performance**
(ApoA1)

*FDR InfiniBand*
*Higher is better*

*Thor Cluster*
*28 Cores Per Node*
NAMD Performance – MPI Libraries and Compilers

- On par performance is seen with different MPI and compilers
  - With FDR IB, both MPI libraries able to scale NAMD to ~1000 CPU-core range
• Intel E5-2697v3 (Haswell) cluster outperforms prior CPU generation
  – Performs 24% higher than E5-2680v2 (Ivy Bridge) Jupiter cluster
    • Mostly due to the additional cores and difference in CPU speed

• System components used:
  – Jupiter: Dell PowerEdge R720: 2-socket 10c E5-2680v2 @ 2.8GHz, 1600MHz DIMMs, FDR IB
  – Thor: Dell PowerEdge R730: 2-socket 14c E5-2697v3 @ 2.6GHz, 2133MHz DIMMs, FDR IB

NAMD Performance
(apoa1)

FDR InfiniBand
Higher is better
NAMD Performance – CPU Core Frequency

- Running at higher clock rate allows greater performance improvement
  - Up to 52% higher performance from 2 GHz to 2.6 GHz, at 6-11% of gain in power
  - Up to 23% higher performance from 2.3 GHz to 2.6 GHz, at 6-11% of gain in power
  - Turbo clock turned off throughout these tests

**NAMD Performance (ApoA1)**

![Chart showing performance rating vs. number of nodes and CPU frequency]

- Higher is better
- 28 Cores Per Node
NAMD Performance – Cores Per Node

- Running more CPU cores provides more performance at some power
  - ~36% higher performance from 20 to 28 cores, at 9-13% of gain in power
  - ~10% higher performance from 24 to 28 cores, at 4-6% of gain in power
  - Turbo clock turned off throughout these tests
• Running more CPU cores provides more performance at some power
  - Up to 9-29% higher performance by enabling Turbo Mode, at 13-17% of gain in power
  - The Turbo gain diminishes as cluster scales

**NAMD Performance**
(ApoA1)

Higher is better

**Thor Cluster**

**NAMD Performance**
(ApoA1)
NAMD Profiling – MPI calls

- **NAMD shows high usage for MPI non-blocking communications:**
  - The performance of MPI_Iprobe affects NAMD performance
  - MPI Time: MPI_Iprobe (97%), MPI_Barrier (1%), MPI_Comm_dup (0.8%)
  - Wall Time: MPI_Iprobe (79%), MPI_Barrier (1%), MPI_Comm_dup (0.7%)
Communications for NAMD mostly concentrated in the midsize messages
- The point to point communications appear to be around 4KB to 10KB
NAMD Summary

• Scalability of NAMD can reach thousand of CPU cores and beyond
  – NAMD replies on the low latency of interconnect and high throughput
  – Intel MPI and HPC-X performs on par; Intel and the latest GNU 4.9.1 compilers outperforms default GNU 4.4.6 by ~6%
  – Running NAMD with higher CPU clock rate and cores per node provides better performance at lower additional power

• Good improvement seen from previous generation of servers
  – Provided up to 23% higher performance on a single node basis

• InfiniBand FDR is the most efficient cluster interconnect for NAMD
  – With RDMA, FDR IB reduces network overhead; allows CPU to focus on computation
  – InfiniBand runs faster than 1GbE by 7x, 10GbE by 5x at 4 nodes / 112 MPI processes; scalability grows as cluster scales

• NAMD Profiling
  – MPI_Iprobe consumes about 97% of MPI time or 79% of Wall time for non-blocking communications
  – The point-to-point message sizes appeared to be around 4-10KB
Thank You

HPC Advisory Council