



Amber Performance Benchmark and Profiling

January 2013













• 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

- AMBER performance overview
- Understanding AMBER communication patterns
- Ways to increase AMBER productivity
- MPI libraries comparisons

For more info please refer to

- <u>http://www.dell.com</u>
- <u>http://www.intel.com</u>
- <u>http://www.mellanox.com</u>
- <u>http://ambermd.org</u>

AMBER Application



• AMBER

- Software for analyzing large-scale molecular dynamics (MD) simulation trajectory data
- Reads either CHARMM or AMBER style topology/trajectory files as input, and its analysis routines can scale up to thousands of compute cores or hundreds of GPU nodes with either parallel or UNIX file I/O
- AMBER has dynamic memory management, and each code execution can perform a variety of different structural, energetic, and file manipulation operations on a single MD trajectory at once
- The code is written in a combination of Fortan90 and C, and its GPU kernels are written with NVIDIA's CUDA API to achieve maximum GPU performance





Test Cluster Configuration

- Dell™ PowerEdge™ R720xd 16-node (256-core) "Jupiter" cluster
 - Dual-Socket Eight-Core Intel E5-2680 @ 2.70 GHz CPUs (Static max Perf in BIOS)
 - Memory: 64GB memory, DDR3 1600 MHz
 - OS: RHEL 6.2, OFED 1.5.3 InfiniBand SW stack
 - Hard Drives: 24x 250GB 7.2 RPM SATA 2.5" on RAID 0
- Intel Cluster Ready certified cluster
- Mellanox ConnectX-3 FDR InfiniBand VPI adapters
- Mellanox SwitchX SX6036 InfiniBand switch
- Compilers: Intel Composer XE 2011
- MPI: Intel MPI 4 U3, Open MPI 1.5.5, Platform MPI 8.2
- Application: AMBER 11, AmberTools 1.5
- External libraries: charm-6.4.0, fftw-2.1.3, TCL 8.3
- Benchmark workload: Cellulose_production_NVE_256_128_128 (408,609 atoms)





- Intel® Cluster Ready systems make it practical to use a cluster to increase your simulation and modeling productivity
 - Simplifies selection, deployment, and operation of a cluster
- A single architecture platform supported by many OEMs, ISVs, cluster provisioning vendors, and interconnect providers
 - Focus on your work productivity, spend less management time on the cluster

Select Intel Cluster Ready

- Where the cluster is delivered ready to run
- Hardware and software are integrated and configured together
- Applications are registered, validating execution on the Intel Cluster Ready architecture
- Includes Intel® Cluster Checker tool, to verify functionality and periodically check cluster health

About Dell PowerEdge R720xd



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

AMBER Performance – Processor Generations



- Intel E5-2600 Series (Sandy Bridge) outperforms prior generations
 Up to 76% higher performance than Intel Xeon X5670 (Westmere) at 16-node
- System components used:
 - Jupiter: 2-socket Intel E5-2680 @ 2.7GHz, 1600MHz DIMMs, FDR IB, 24 disks
 - Janus: 2-socket Intel x5670 @ 2.93GHz, 1333MHz DIMMs, QDR IB, 1 disk



AMBER Benchmark

(Cellulose)

AMBER Performance – Network Interconnect



- FDR InfiniBand delivers the highest performance
 - 30%+ higher performance versus 40GbE
 - 60%+ higher performance versus 10GbE

• FDR InfiniBand provides the best scalability performance on Ethernet

- 61% better performance than 10GbE at 16 nodes
- 35% better performance than 40GbE at 16 nodes
- 16% better performance than QDR InfiniBand at 16 nodes
- 1GbE does not scale at all

AMBER Benchmark

(Cellulose)



AMBER Performance – MPI



- Intel MPI allows AMBER to achieve the highest performance
 - Up to 13% performance gains by using Intel MPI versus Open MPI
 - The default DAPL provider is used for Intel MPI; openib btl is used for Open MPI
 - Processor binding is enabled for Open MPI; pinning is enabled by default in Intel MPI



AMBER Benchmark (Cellulose)

NETWORK OF EXPERTISE

AMBER Profiling – MPI/User Time Ratio



- Computation time dominates the overall run time
 - More time spent on computation than communications
- Compute time is reduced as the cluster scales
 - Job can complete faster by spreading the workload to more compute nodes



AMBER Profiling (Cellulose)

AMBER Profiling – MPI Message Size



• Majority of MPI messages are messages in the small message sizes

- In the range of 0 to 64B and 16KB to 64KB
- A spike in midrange messages around 16 to 64KB for larger node counts
- Significantly large volume of messages being sent between processes





AMBER Profiling – % MPI Time and Calls



- Large percentage of calls for non-blocking MPI communications
 - 32% in MPI_Irecv, MPI_Isend, MPI_Waitany
- Time spent in MPI is mixed between collectives and non-blocking communications
 - 31% in MPI_Allgatherv, 18% in MPI_Allreduce, 16% in MPI_Waitall and MPI_Waitany



AMBER Profiling

(Cellulose, 16-node) % Time Spent of MPI Calls



16 Nodes

AMBER Profiling – Time Spent on MPI Calls



The cellulose dataset shows spikes in data communications

- Shows spike in data communication for 1 in every 2-3 ranks
- The spikes are caused by time spent on MPI_Waitany (for pending MPI_Isend/Irecv)



AMBER Profiling – MPI Buffer Size Distributions



- The data transfers are taken place in the small and midrange messages
 - MPI_Recv are in the range of 64B to 256B
 - MPI_Isend/Irecv are in the range of 4KB to 64KB
 - MPI_Allgatherv occurs in the range around 64KB



AMBER Profiling – Distribution of MPI Data



Majority of communications takes place with the close neighboring ranks

- There are also some communications take place between far ranks
- Large transfers up to 4GB take place between close ranks at 16 nodes
- Data transfer drops from 5.2GB per rank (at 4 nodes) to 4GB per rank (at 16 nodes)



AMBER Profiling – Aggregated Transfer



- Aggregated data transfer refers to:
 - Total amount of data being transferred in the network between all MPI ranks collectively
- Huge data transfer takes place between processes distributed in the network
 - The total data transfer between processes is around 2TB at 16 nodes



AMBER – Summary



AMBER is a compute and network throughput intensive application

• Performance:

- Intel Xeon E5-2680 on the "Jupiter" cluster and FDR InfiniBand enable AMBER to scale
- "Jupiter", the E5-2680 cluster performs up to 76% over "Janus" the X5670 cluster
- Network:
 - FDR InfiniBand allows AMBER to run at the highest network throughput at 56Gbps
 - FDR InfiniBand delivers 16% better performance than QDR InfiniBand at 16 nodes
 - FDR InfiniBand outperforms Ethernet (10GbE and 40GbE) by +35-61% at 16 nodes
- MPI:
 - Intel MPI performs 13% better Open MPI at 16 nodes
- Profiling:
 - Heavy data communications in small and midrange message sizes 0-64B and 4KB-16KB



Thank You HPC Advisory Council



All trademarks are property of their respective owners. All information is provided "As-Is" without any kind of warranty. The HPC Advisory Council makes no representation to the accuracy and completeness of the information contained herein. HPC Advisory Council Mellanox undertakes no duty and assumes no obligation to update or correct any information presented herein