DL-POLY
Performance Benchmark and Profiling

August 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
  – DL-POLY performance overview
  – Understanding DL-POLY communication patterns
  – Ways to increase DL-POLY productivity
  – Network Interconnect comparisons

• For more info please refer to
  – http://www.dell.com
  – http://www.intel.com
  – http://www.mellanox.com
Objectives

• The following was done to provide best practices
  – DL-POLY performance benchmarking
  – Interconnect performance comparisons
  – Processor generation performance comparison
  – Understanding DL-POLY communication patterns

• The presented results will demonstrate
  – The scalability of the compute environment to provide nearly linear application scalability
  – The capability of DL-POLY to achieve scalable productivity
• **DL-POLY**
  – Is a general purpose classical molecular dynamics simulation software
  – Developed at Daresbury Laboratory by I.T. Todorov and W. Smith.

• **DL_POLY_4**
  – General design provides scalable performance from a single processor workstation to a high performance parallel computer.
  – Can be compiled a parallel application code, provided an MPI2 instrumentation is available on the parallel machine
  – DL_POLY_4 offers fully parallel I/O as well as a netCDF alternative (HDF5 library dependence) to the default ASCII trajectory file
  – It is supplied in source form under license
Test Cluster Configuration

- **Dell™ PowerEdge™ R720xd 32-node “Jupiter” cluster**
  - 16-node Dual-Socket Ten-Core Intel E5-2680 V2 @ 2.80 GHz CPUs
  - 16-node Dual-Socket Eight-Core Intel E5-2680 @ 2.70 GHz CPUs
  - Memory: 64GB memory, DDR3 1600 MHz
  - OS: RHEL 6.2, OFED 2.0 InfiniBand SW stack
  - Hard Drives: 24x 250GB 7.2 RPM SATA 2.5” on RAID 0

- **Mellanox Connect-IB FDR InfiniBand adapters and ConnectX-3 Ethernet adapters**

- **Mellanox SwitchX SX6036 InfiniBand VPI switch**

- **Intel Cluster Ready certified cluster**

- **Compilers and Libraries: Intel Composer XE 2013.0.079**

- **Application: DL-POLY 4.04**

- **Benchmark Datasets:**
  - Sodium Chloride, 27K and 429K ions
About Intel® Cluster Ready

• **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
PowerEdge R720xd
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 GbE 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
DL-POLY Performance – Processors

- **E5-2680 V2 (Ivy Bridge) cluster outperforms prior generation**
  - Performs up to 19% better than E5-2680 cluster (Sandy Bridge) at 16 nodes
  - Performs up to 8% better than E5-2680 cluster (Sandy Bridge) at 16 nodes

- **System components used:**
  - IVB: 2-socket 10-core E5-2680 V2 @2.8GHz, 1600MHz DIMMs, FDR IB, 24 HDDs
  - SNB: 2-socket 8-core E5-2680 @ 2.7GHz, 1600MHz DIMMs, FDR IB, 24 HDDs

**Higher is better**

**FDR InfiniBand**

**DL-POLY Benchmark**

<table>
<thead>
<tr>
<th>NaCl 27K</th>
<th>NaCl 729K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance (Jobs/Hour)</td>
<td>Performance (Jobs/Day)</td>
</tr>
<tr>
<td>Number of Nodes</td>
<td>1</td>
</tr>
<tr>
<td>SNB</td>
<td>19%</td>
</tr>
<tr>
<td>IVB</td>
<td>19%</td>
</tr>
</tbody>
</table>

FDR InfiniBand
• DL-POLY demonstrates superior scalability using FDR InfiniBand
  – Performs closer to linear-scale as more nodes join the cluster
  – While Ethernet performance is limited after 4-node due to network traffic congestions

• FDR InfiniBand enables higher cluster productivity
  – Over 6 times versus 1GbE, 5 times versus 10GbE and 3 times versus 40GbE

DL-POLY Benchmark
(NaCl 729K)
**DL-POLY Performance – MXM for Open MPI**

- **MXM enables higher scalability performance for DL-POLY**
  - Provides 213% higher productivity than default Open MPI at 4 nodes for NaCl 27K
  - Provides 56% higher productivity than default Open MPI at 16 nodes for NaCl 729K

- **Flags used for enabling MXM in Open MPI:**
  - `-mca mtl mxm -mca pml cm -mca mtl mxm -mca mtl_mxm_np 0`
  - `-mca btl_openib_if_include mlx5_0:1 -x MXM_RDMA_PORTS=mlx5_0:1`
  - `-mca rmaps_base_dist_hca mlx5_0:1`

---

**Higher is better**

**20 Processes/Node**
DL-POLY Performance – MPI

- **Platform and Intel MPI perform the best at scale**
  - Platform MPI provides 13% higher performance versus Open MPI and 7% over Intel MPI

**DL-POLY Benchmark**
(NaCl 729K)

*Higher is better*
For DL-POLY, MPI communication time is reduced as the cluster scales
- As more nodes take on the computational work, the job completes faster
- Which reduces the communication time for each MPI call

The amount of computation time reduced by half as node count doubles
Majority of the MPI time is spent on MPI_Scatter
- MPI_Allreduce(26%), MPI_Scatter(25%), MPI_Recv(16%), MPI_Send(15%)
- Group communication is the majority of communication type for DL-POLY
In a 16-node job, MPI messages buffer sizes are concentrated in the midrange:

- MPI_Allreduce: Concentrated between 16KB and 256KB
- MPI_Scatter: Around 64KB
• The amount of data transferred to a process is reduced as cluster scales
  – About 21GB per rank is transferred on a single node job
  – About 2GB per rank is transferred on a 16-node job
• **Aggregated data transfer refers to:**
  - Total amount of data being transferred in the network between all MPI ranks collectively

• **Substantial increase in amount of data transfer between the models**
  - Around 50GB of data being exchanged for NaCl 27K model at 16 nodes
  - Around 600GB of data being exchanged for NaCl 729K model at 16 nodes

---

**DL-POLY Profiling – Aggregated Transfer**

---

**InfiniBand FDR**
• The point to point data flow shows the communication pattern
  – DL-POLY mainly communicates mainly its neighbors and close ranks
  – The localized communication pattern stays the same as the cluster scales
  – Data communication increases with the increase of interaction between atoms
DL-POLY – Summary

- **DL-POLY delivers superior linear scalability and performance**
  - DL-POLY can take advantage of additional compute power by using FDR InfiniBand

- **Superior network productivity needed for DL-POLY to run efficiently**
  - FDR InfiniBand performs 6 times faster vs 1GbE, 5 times vs 10GbE, 3 times vs 40GbE
  - Ethernet performance hinders the scalability of DL-POLY starting at 4-node

- **MXM enables higher scalability performance for Open MPI**
  - Provides higher productivity over default Open MPI by 56% at 16 nodes

- **Intel Ivy Bridge-EP series and FDR InfiniBand enable DL-POLY to scale**
  - The E5-2680 V2 (IVB) cluster outperforms E5-2680 (SNB) cluster by 19% at 16 nodes

- **MPI Performance and MPI Profiling**
  - Platform MPI provides 13% higher performance versus Open MPI and 7% over Intel MPI
  - Most time consuming MPI function: MPI_Allreduce(26%), MPI_Scatter(25%)
  - Significant data transfers in “midrange” data buffers are taken place for DL-POLY
Thank You
HPC Advisory Council