Graph500
Performance Benchmark and Profiling

March 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**
  - Graph500 performance overview
  - Understanding Graph500 communication patterns
  - Ways to increase Graph500 productivity
  - MPI libraries comparisons
- **For more info please refer to**
  - http://www.dell.com
  - http://www.intel.com
  - http://www.mellanox.com
  - http://www.graph500.org

Graph500

- The Graph500 is a rating of supercomputer systems, focused on Data intensive loads
  - The project was announced on International Supercomputing Conference (ISC) in June 2010
  - The first list was published at the ACM/IEEE Supercomputing Conference in November 2010
- Graph500 Benchmark
  - Stresses communication subsystem, instead of counting double precision floating-point
  - Based on a breadth-first search in a large undirected graph
    - Model of Kronecker graph with average degree of 16
  - Contains two computation kernels in the benchmark:
    - 1st kernel is to generate the graph and compress it into sparse structures CSR or CSC
    - 2nd kernel does a parallel BFS search of some random vertices (64 search iterations per run).
  - Performance metric used to rank the supercomputers is GTEPS (10^9 Traversed edges/second)
  - For more information: http://www.graph500.org/referencecode

Objectives

- **The presented research was done to provide best practices**
  - Graph500 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. Turbo Mode disabled unless otherwise stated
  - Memory: 64GB memory, DDR4 2133 MHz, Memory Snoop Mode: Cluster-on-Die
  - 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
- **Mellanox Connect-IB FDR InfiniBand adapters**
- **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**
- **Code Implementations:**
  - Graph500 Reference Implementation, version 2.1.4
  - Tuned MPI implementation with 2-D data distribution

PowerEdge R730

Massive flexibility for data intensive operations

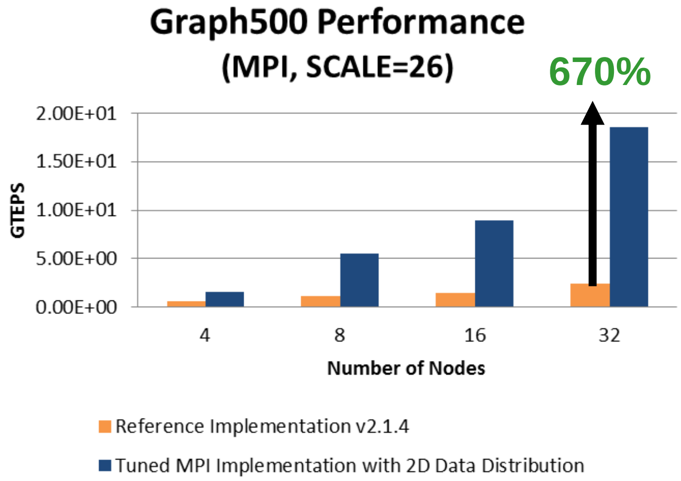
- **Performance and efficiency**
  - Intelligent hardware-driven systems management with extensive power management features
  - Innovative tools including automation for parts replacement and lifecycle manageability
  - Broad choice of networking technologies from GigE to IB
  - Built in redundancy 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

Graph500 Performance – Graph500 Implementations

- **There are 2 implementations that can be obtained from Graph500 web site**
- **Reference Implementation version 2.1.4**
  - Default version
  - http://www.graph500.org/sites/default/files/files/graph500-2.1.4.tar.bz2
  - Workaround is applied in the mpi_workarounds.h for problems with hanging processes
- **Tuned MPI implementation with 2-D data distribution**
  - Implementation by Indiana University
  - http://www.graph500.org/sites/default/files/files/mpi-tuned-2d.tar.bz2
  - This implementation has a dependency on libhugetlbfs

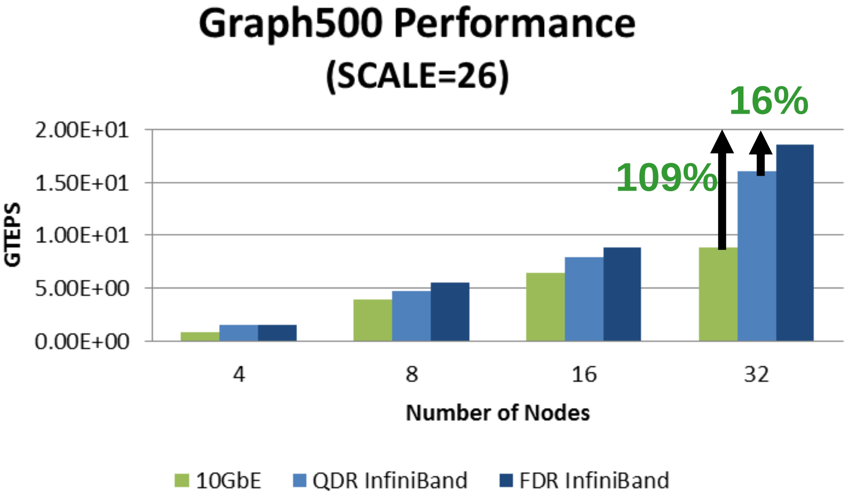
Graph500 Performance – Graph500 Implementations

- **Tuned Impl. with 2-D data distribution outperforms Reference Implementation**
  - Outperforms the reference implementation by over 6.7 times
- **Most implementations in listings are defined as “Custom” implementation**
  - There are many custom (code) implementations, but they are not publically obtainable
  - The algorithm used in code can vary the performance, and can dramatically improve performance



Graph500 Performance – Network Interconnect

- **FDR InfiniBand delivers higher performance than other network interconnects**
  - FDR IB outperforms QDR IB by 16% at 32 nodes, 10GbE by 109% at 32 nodes
  - InfiniBand outperforms Ethernet in MPI_Alltoallv performance
  - The hybrid MPI-OpenMP case is used

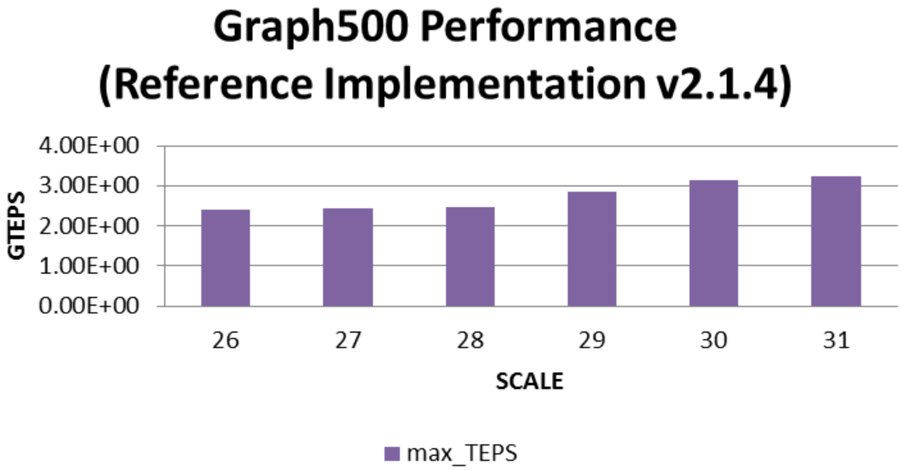


Tuned MPI implementation with 2-D data distribution

4 MPI tasks Per Node

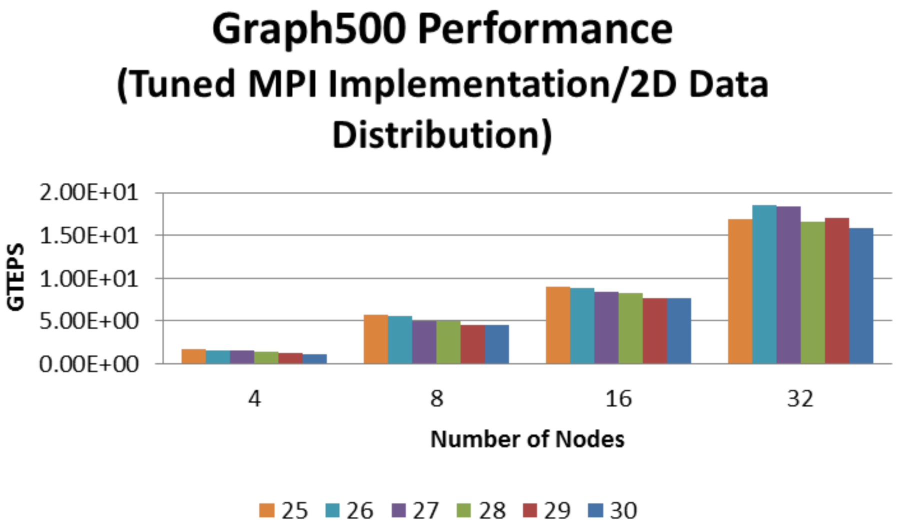
Graph500 Performance – Scale Sizes

- **For the referenced implementation:**
  - Increasing the scale parameter would increase memory consumption
  - Higher Scale value would gradually increase the GTEPS results



Graph500 Performance – Scale Sizes

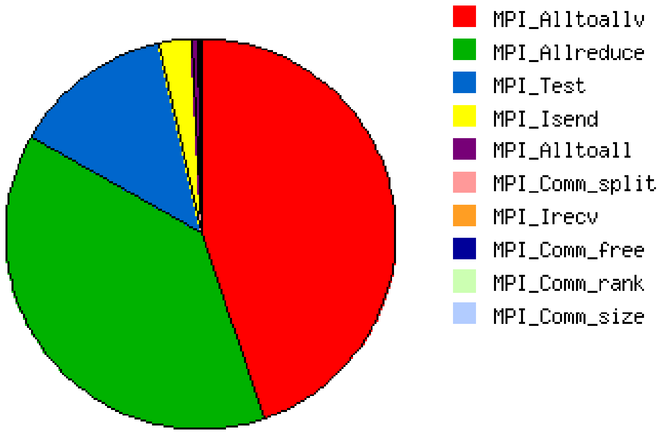
- **For the tuned MPI implementation with 2-D data distribution:**
  - Increasing the scale parameter would also increase memory consumption
  - Higher scale sizes does not seem to increase the GTEPS results



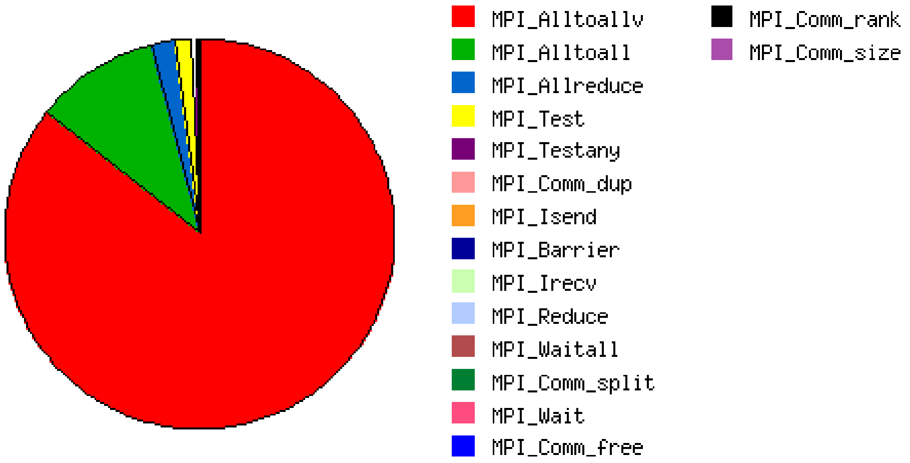
Graph500 Performance – MPI Profiles

- **Both Graph500 implementation shows high usage for MPI Collective Operations:**
  - MPI_Alltoallv affects Graph500 performance for both of the implementations
  - Ref Imp: MPI_Alltoallv (97%), MPI_Allreduce (1%), MPI_Test (0.8%)
  - Tuned MPI: MPI_Alltoallv (44%), MPI_Alltoall (38%) MPI_Allreduce (13%)
  - This shows MPI collective communications has a direct impact on Graph500 performance

Reference Implementation – 32 nodes



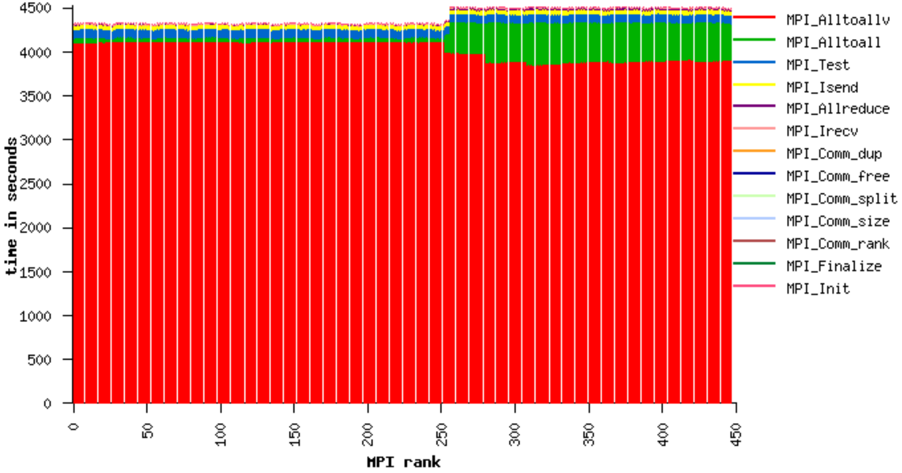
Tuned MPI implementation with 2-D data distribution – 32 nodes



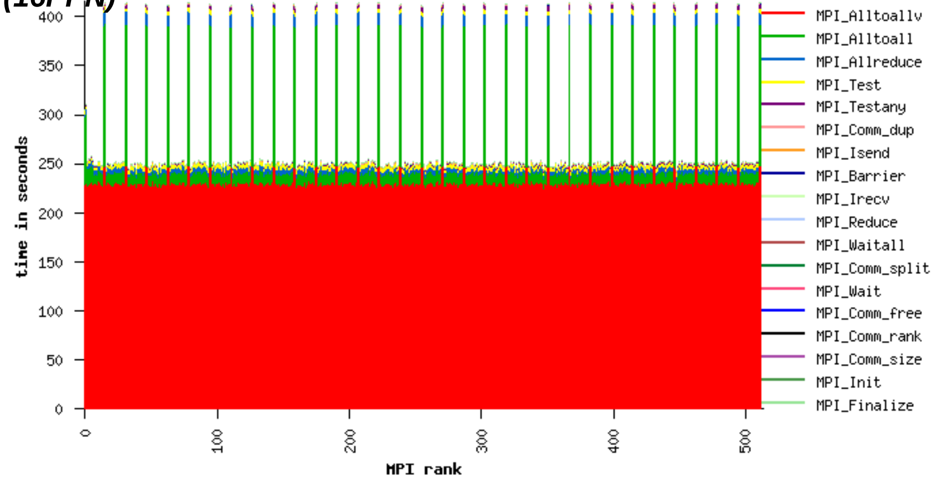
Graph500 Profiling – MPI calls

- **Both implementations appeared to have a substantial use in MPI_Alltoallv**
  - MPI_Alltoallv is a collective op for all processes send data to and receive data from all other processes
  - MPI Collective operations appears to have a direct impact on Graph500 performance
  - FDR InfiniBand outperforms other network interconnect on the performance of collective operations

Reference Implementation – 16 nodes



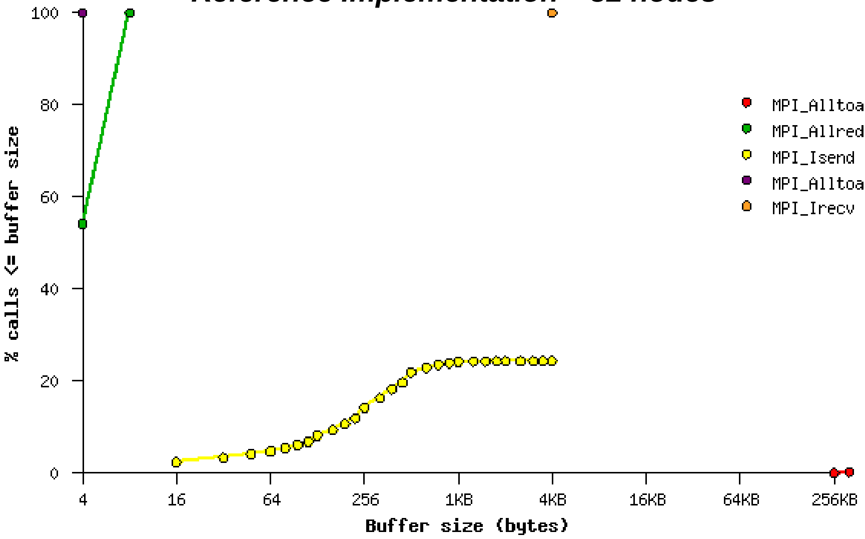
Tuned MPI implementation with 2-D data distribution – 32 nodes (16PPN)



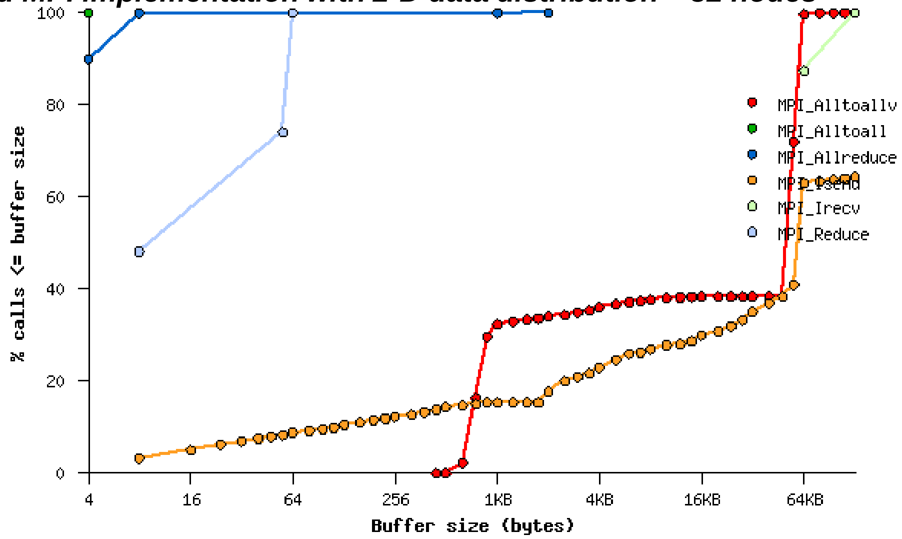
Graph500 Profiling – MPI Message Distribution

- **MPI_Alltoallv communications is the most time consuming MPI communication**
  - In Reference Implementation: Alltoallv seems to concentrated at 256KB and beyond
  - In Tuned Implementation: Alltoallv appears to be between 512B and 1KB, and 57KB and 64KB

Reference Implementation – 32 nodes



Tuned MPI implementation with 2-D data distribution – 32 nodes



Graph500 Summary

- **Performance of Graph500 depends on the implementation used**
  - Tuned MPI implementation with 2D distribution outperforms Reference Implementation by over 6 times
  - Other implementations of Graph500 exist and likely to improve performance, however not freely obtainable
- **InfiniBand FDR is the most efficient cluster interconnect for Graph500**
  - FDR InfiniBand outperforms QDR IB by 16%, and 10GbE by 109% at 32 nodes
  - InfiniBand outperforms Ethernet in MPI_Alltoallv performance
- **Graph500 Profiling**
  - Both Graph500 implementation shows high usage for MPI Collective Operations:
  - MPI_Alltoallv affects Graph500 performance for both of the implementations
  - Ref Imp: MPI_Alltoallv (97%), MPI_Allreduce (1%), MPI_Test (0.8%)
  - Tuned MPI: MPI_Alltoallv (44%), MPI_Alltoall (38%) MPI_Allreduce (13%)
  - This shows MPI collective communications has a direct impact on Graph500 performance

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 undertakes no duty and assumes no obligation to update or correct any information presented herein