HPC ADVISORY COUNCIL

AMG2013
Performance Benchmark and Profiling

July 2013

DELL intel Mellanox TECHNOLOGIES

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**
  - AMG performance overview
  - Understanding AMG communication patterns
  - Ways to increase AMG productivity
  - MPI libraries 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**
  - AMG performance benchmarking
  - Interconnect performance comparisons
  - MPI performance comparison
  - Understanding AMG communication patterns

- **The presented results will demonstrate**
  - The scalability of the compute environment to provide nearly linear application scalability
  - The capability of AMG to achieve scalable productivity

AMG

- **AMG**
  - A parallel algebraic multigrid solver for linear systems arising from problems on unstructured grids
- **AMG2013**
  - AMG2013 is a SPMD code, written in ISO-C, which uses MPI communications
  - Parallelism is achieved by data decomposition by subdividing the grid into logical P x Q x R (in 3D) chunks of equal size
- **AMG2013 is a highly synchronous code for scalability**
  - The communications and computations patterns exhibit the surface-to-volume relationship common to many parallel scientific codes
  - Parallel efficiency is largely determined by the size of the data "chunks" mentioned above, and the speed of communications and computations on the machine
  - AMG2013 is also memory-access bound, doing only about 1-2 computations per memory access, so memory-access speeds will also have a large impact on performance

Test Cluster Configuration

- **Dell™ PowerEdge™ R720xd 32-node (512-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 2.0 InfiniBand SW stack
  - Hard Drives: 24x 250GB 7.2 RPM SATA 2.5" on RAID 0
- **Mellanox ConnectX-3 FDR InfiniBand VPI adapters**
- **Mellanox SwitchX SX6036 InfiniBand VPI switch**
- **Intel Cluster Ready certified cluster**
- **MPI: Open MPI 1.6.4, Intel MPI 4.1.0, MVAPICH2 1.9**
- **Application: AMG2013May20**
- **Benchmark datasets:**
  - AMG2013May20 – 3D 7-point Laplace on cube (each rank solving 150x150x150)
  - http://www.nersc.gov/systems/trinity-nersc-8-rfp/draft-nersc-8-trinity-benchmarks/amg/

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

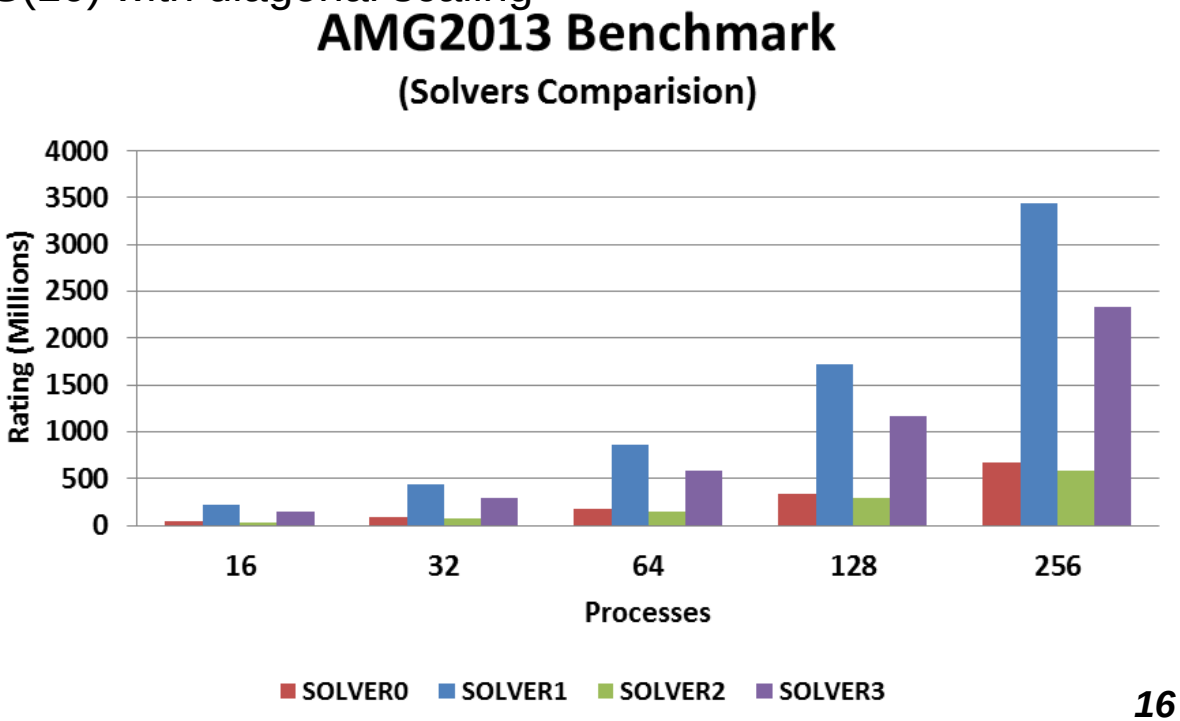
PowerEdge R720xd

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

AMG Performance – Solvers

- **Different linear solvers uses different interpolation algorithm**
  - Solver 1 provides the highest performance rating
  - All solvers shows very good scalabilty up to 256 processes
- **Solver ID and its corresponding descriptions**
  - 0 - PCG with AMG precondition
  - 1 - PCG with diagonal scaling
  - 2 - GMRES(10) with AMG precondition
  - 3 - GMRES(10) with diagonal scaling

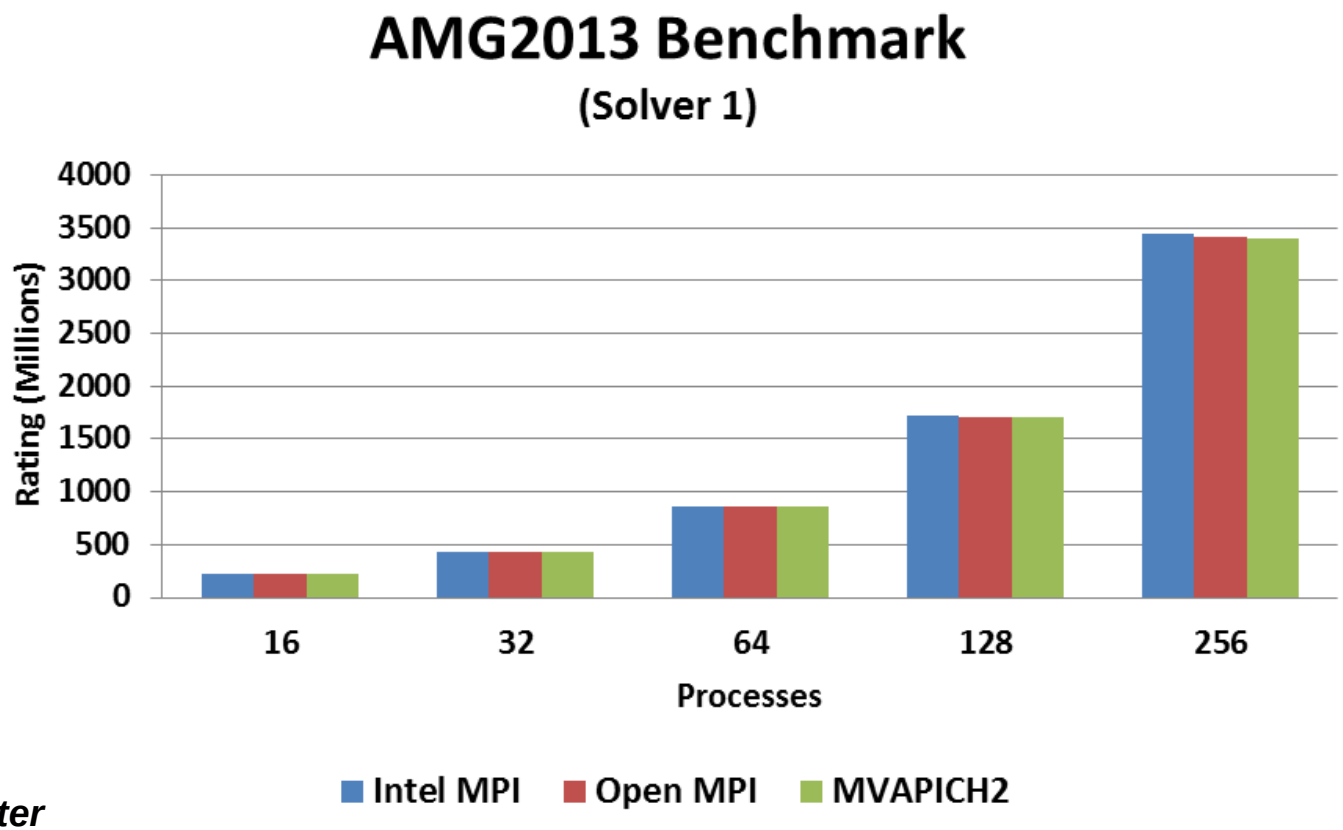


Higher is better

16 Processes/Node

AMG Performance – MPI

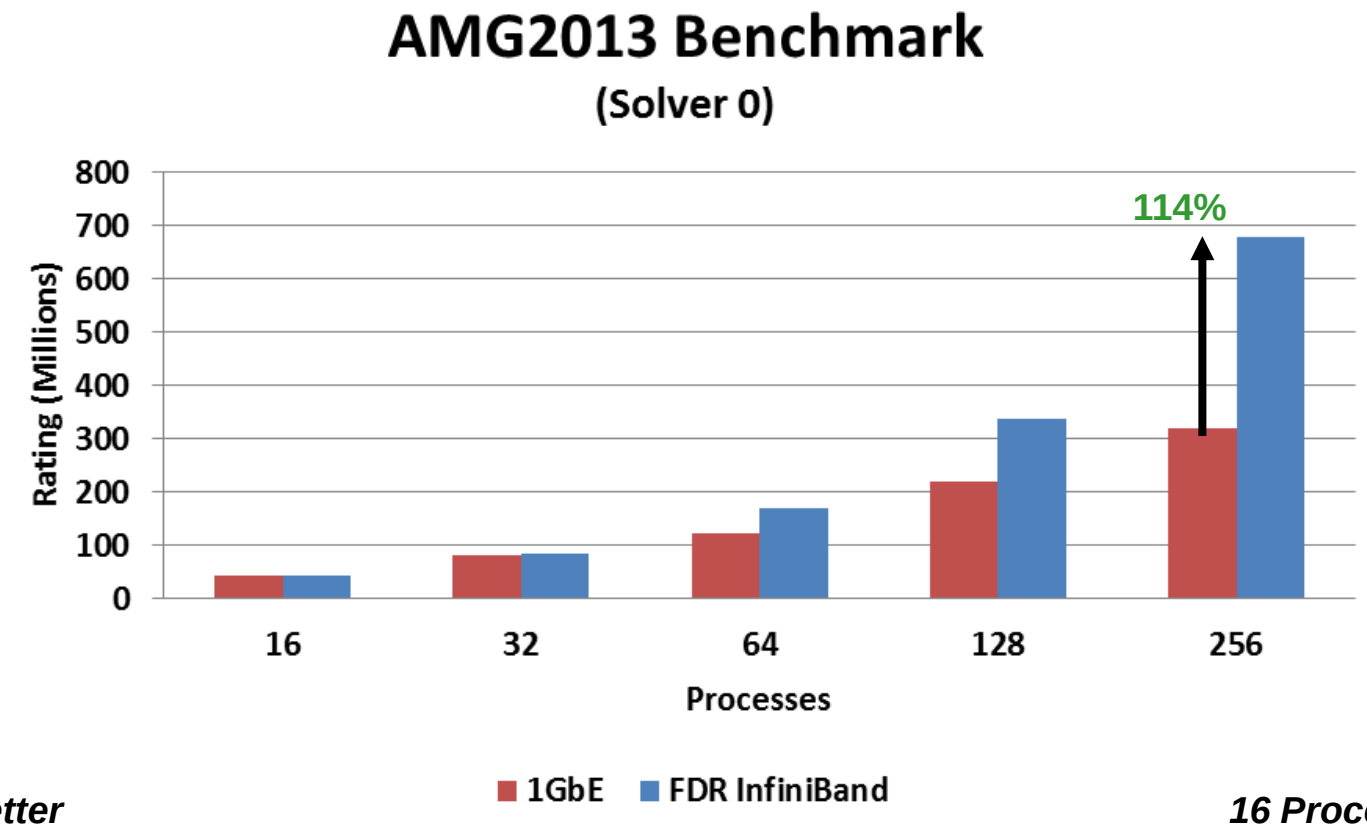
- **All MPI implementation tested performs roughly the same**
  - Reflects that input data and code spends small amount of time in communications



Higher is better

AMG Performance – Interconnect

- **FDR InfiniBand provides better scalability performance than Ethernet**
  - 114% better performance than 1GbE at 16 nodes



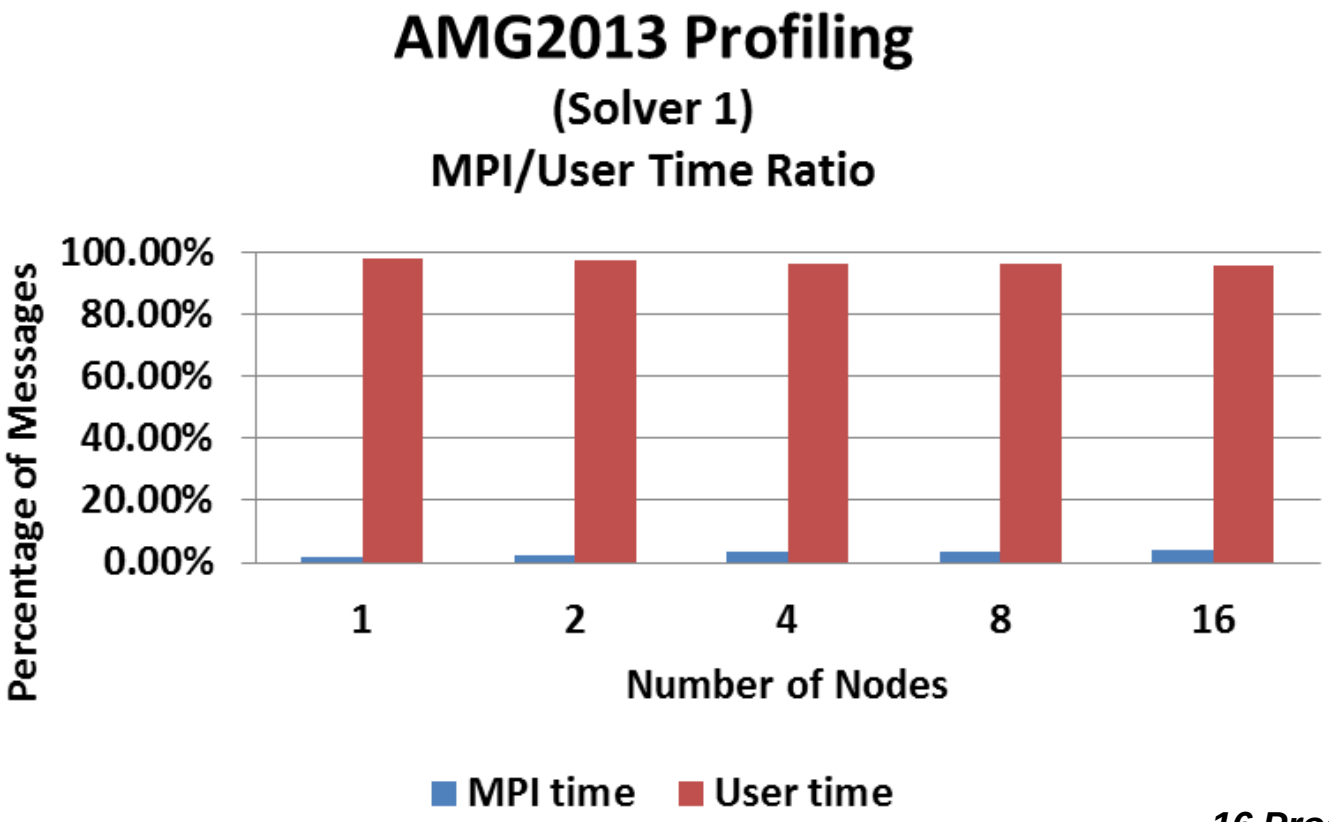
Higher is better

16 Processes/Node

*Rating = System Size * Iterations / Solve Phase Time*

AMG Profiling – CPU / Network Ratio

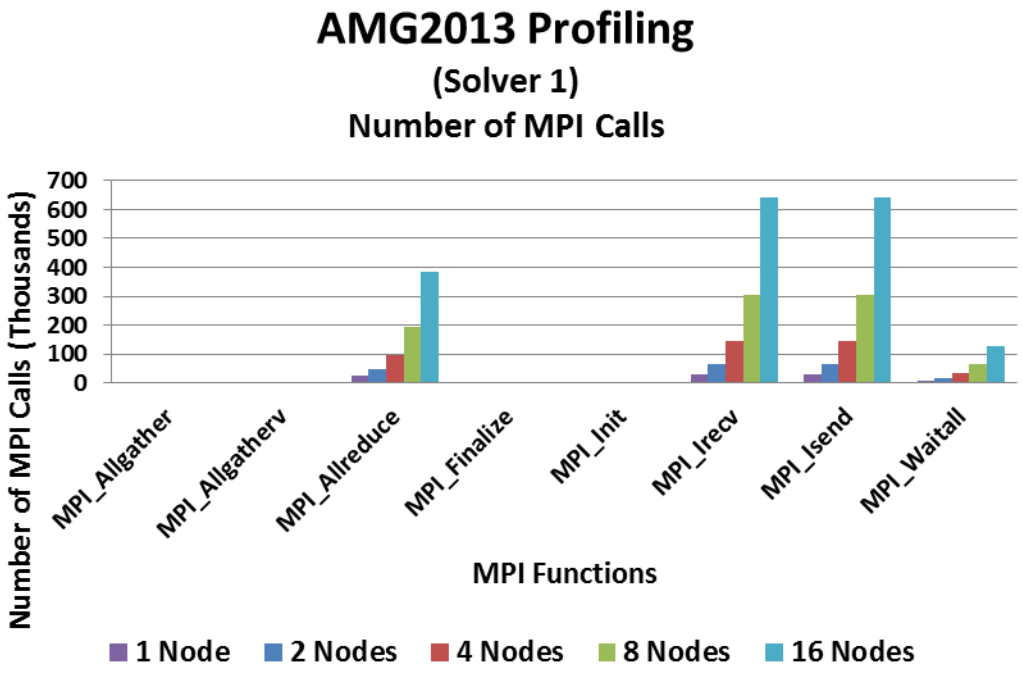
- **Very small percentage of time is spent on MPI**
  - For this input dataset at this scale up to 16 nodes (or 256 processes)

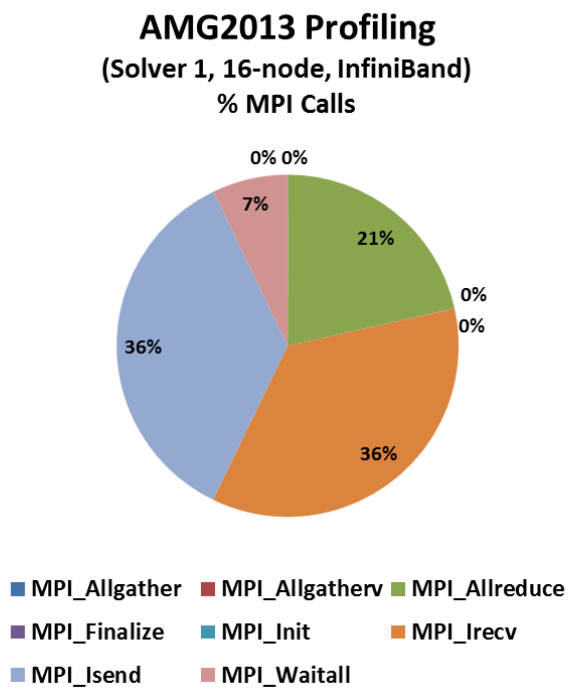


16 Processes/Node

AMG Profiling – Number of MPI Calls

- **AMG utilizes non-blocking communications in most data transfers**
  - MPI_Irecv(36%), MPI_Isend(36%) and MPI_Allreduce(21%) at 16 nodes

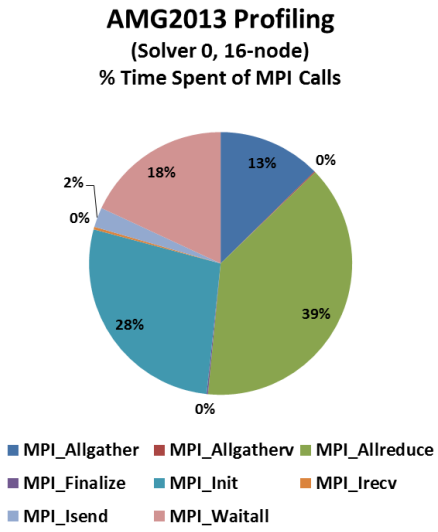


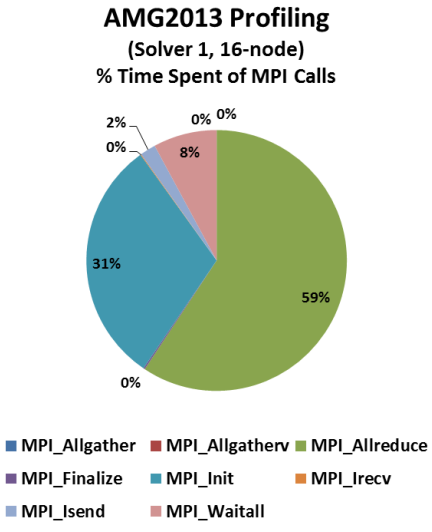


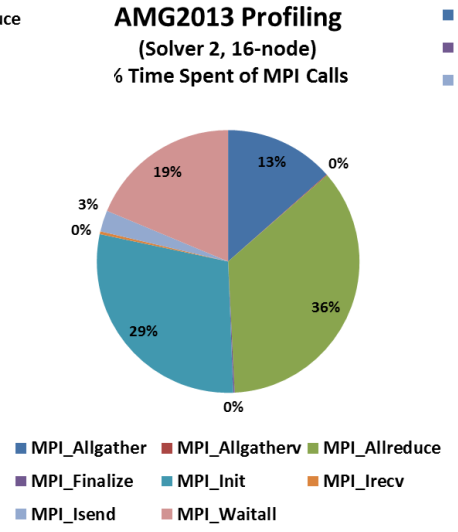
16 Processes/Node

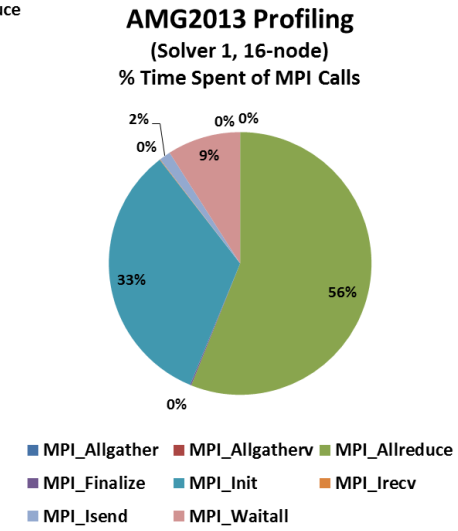
AMG Profiling – % Time Spent on MPI

- **The most time consuming MPI calls are for collective communications**
  - MPI_Allreduce(59%) for Solver 1
  - Roughly the same communicaiton patterns seen for other solvers



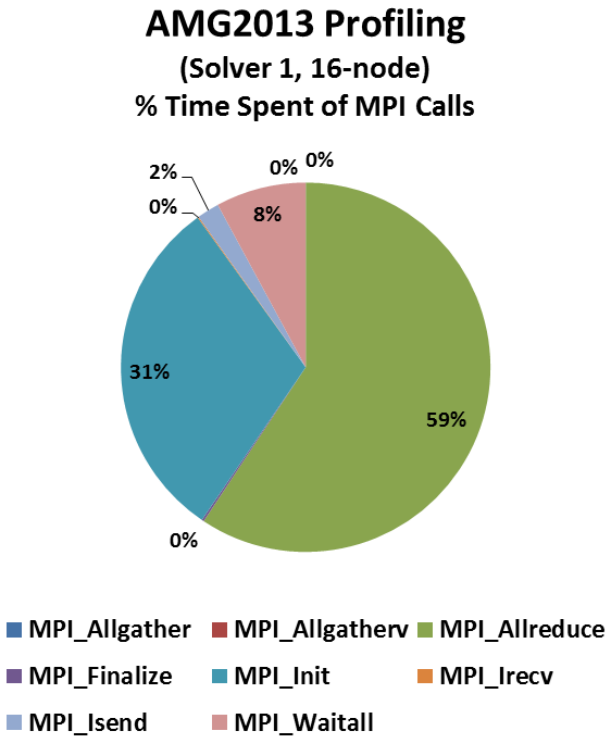


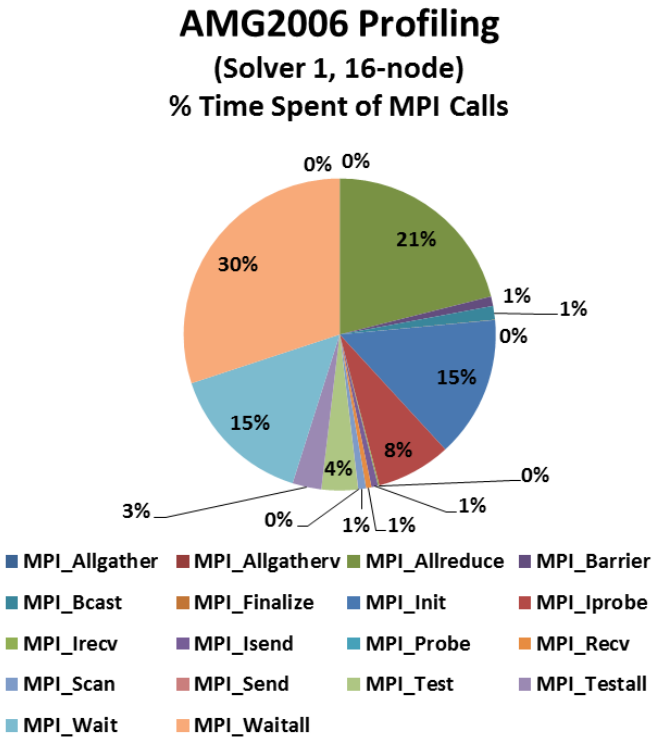




AMG Profiling – Communication Patterns

- **The MPI communications appear to be different from prior version**
  - AMG2013: MPI_Allreduce(59%)
  - AMG2006: MPI_Waitall (30%), MPI_Allreduce (21%)

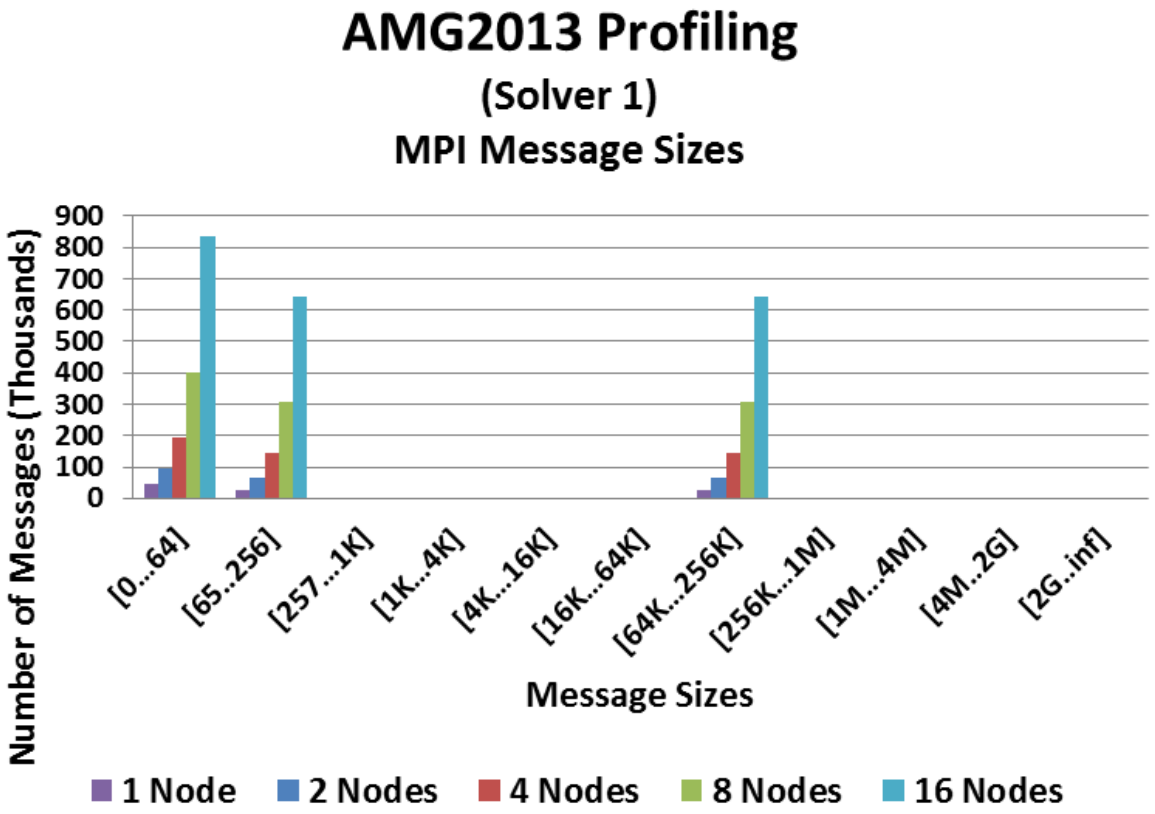




16 Processes/Node

AMG Profiling – MPI Message Sizes

- **AMG uses small and medium MPI message sizes**
  - Most message sizes are between 0B to 256B, and 64KB to 256KB

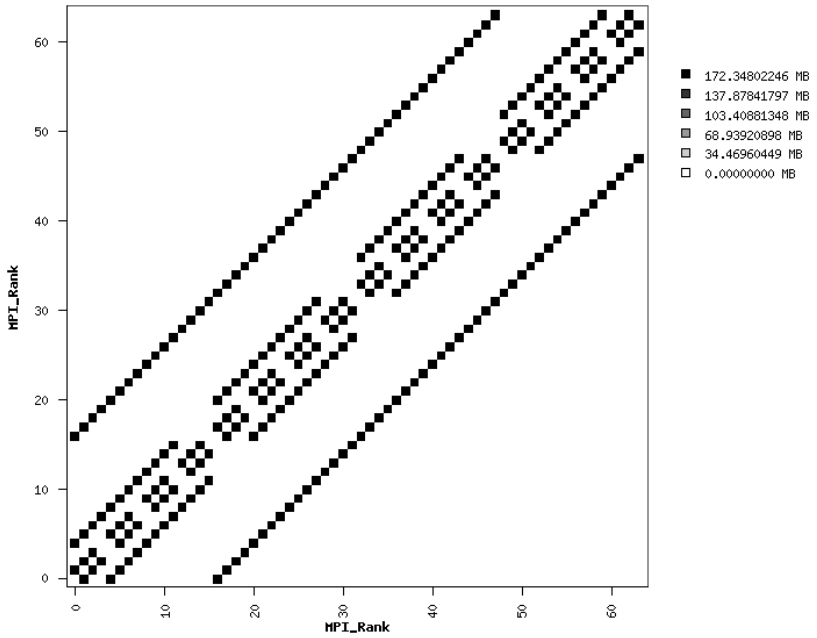


16 Processes/Node

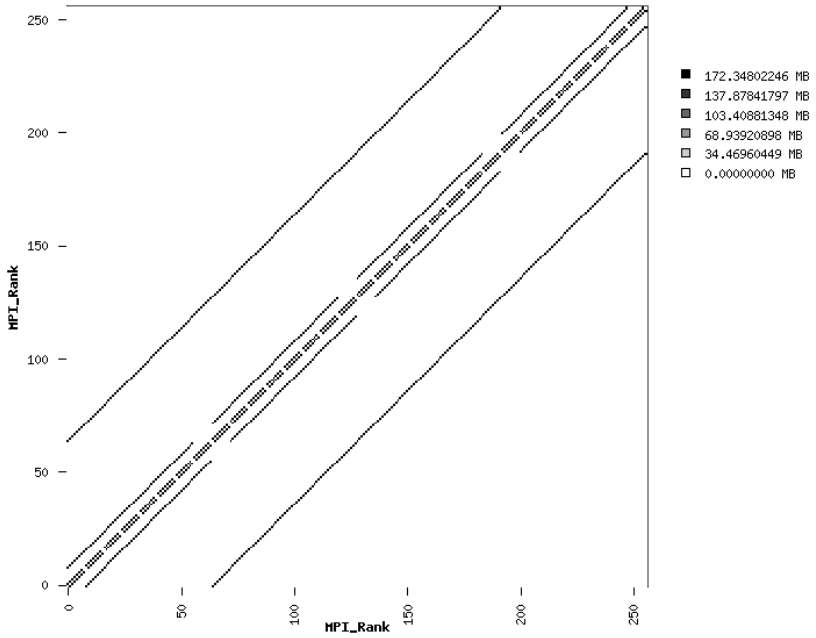
AMG Profiling – MPI Data Transfer

- **Distribution of data transfers between the MPI processes**
  - Non-blocking point-to-point data communications between processes are involved
  - Similar distribution is seen at larger node count

64 Processes



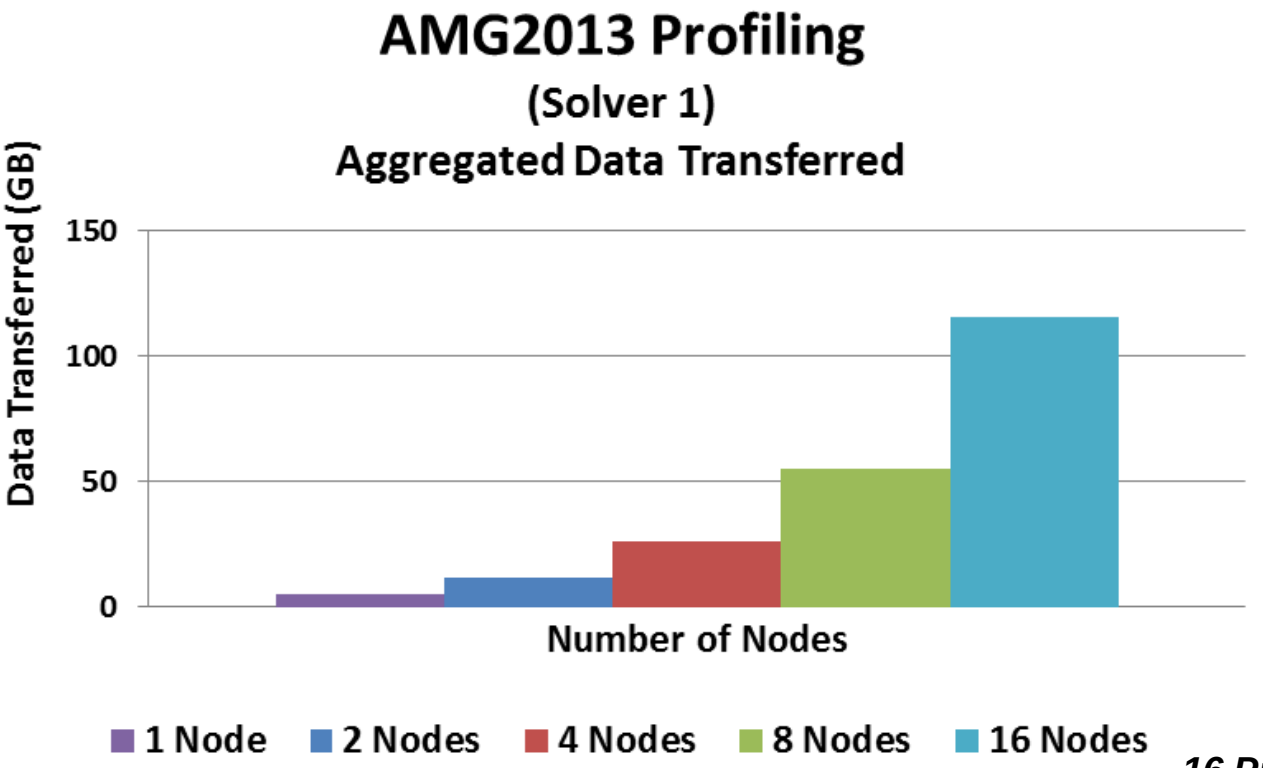
256 Processes



16 Processes/Node

AMG Profiling – Aggregated Transfer

- **Aggregated data transfer refers to:**
  - Total amount of data being transferred in the network between all MPI ranks collectively
- **Substantially larger data transfer takes place in AMG**
  - As node count doubles, amount of data transferred is more than double



Pure MPP

16 Processes/Node

AMG – Summary

- **AMG shows very good scalability to run at large scale**
  - Shows excellent scalability up to 256 cores (16 nodes) using pure MPI
  - Solver 1 shows the best performance rating amongst all of the solvers tested
- **Network**
  - FDR InfiniBand provides better scalability performance than Ethernet
    - Over 113% better performance than 1GbE at 16 nodes
  - All MPI implementation tested delivers roughly the same in performance
- **Profiling**
  - The most time consuming MPI calls is MPI_Allreduce
  - Most message sizes are between 0B to 256B, and 64KB to 256KB
  - AMG2013 changes the communication pattern significant from previous version

Thank You

HPC Advisory Council