HPC ADVISORY COUNCIL

AMR (Adaptive Mesh Refinement) Performance Benchmark and Profiling

January 2011

DELL intel Mellanox TECHNOLOGIES

Acknowledgment:
- **The DoD High Performance Computing Modernization Program**
- **John Bell from Lawrence Berkeley Laboratory**

Note

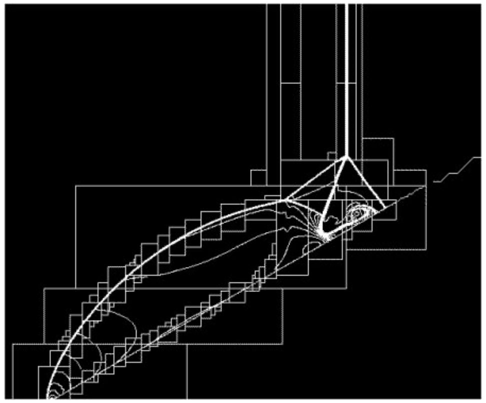
- **The following research was performed under the HPC Advisory Council activities**
  - Participating vendors: Intel, Dell, Mellanox
  - Compute resource - HPC Advisory Council Cluster Center

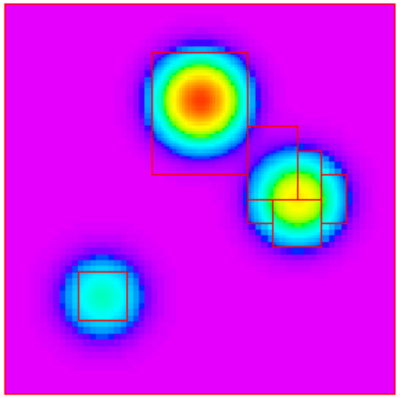
- **We would like to acknowledge**
  - The DoD High Performance Computing Modernization Program for providing access to the FY 2009 benchmark suite
  - John Bell from Lawrence Berkeley Laboratory for developing the application

- **For more info please refer to**
  - http://www.dell.com
  - http://www.intel.com
  - http://www.mellanox.com

AMR Application

- **AMR - Adaptive Mesh Refinement (AMR)**
  - A collection of 3 applications for solving a wide variety of problems that benefit from grids with adaptive, inhomogeneous spatial resolution
  - AMR is the product of the Center for Computational Sciences and Engineering at Lawrence Berkeley National Laboratory
  - This particular benchmark makes use of the HyperClaw application for solving a gasdynamic problem; it is written primarily in C++





Objectives

- **The following was done to provide best practices**
  - AMR performance benchmarking
  - Interconnect performance comparisons
  - Understanding AMR communication patterns
  - Ways to increase AMR productivity
  - Compilers and MPI libraries comparisons

- **The presented results will demonstrate**
  - The scalability of the compute environment to provide nearly linear application scalability
  - The capability of AMR to achieve scalable productivity
  - Considerations for power saving through balanced system configuration

Test Cluster Configuration

HPC ADVISORY COUNCIL

- **Dell™ PowerEdge™ M610 38-node (456-core) cluster**
  - Six-Core Intel X5670 @ 2.93 GHz CPUs
  - Memory: 24GB memory, DDR3 1333 MHz
  - OS: RHEL 5.5, OFED 1.5.2 InfiniBand SW stack
- **Intel Cluster Ready certified cluster**
- **Mellanox ConnectX-2 InfiniBand adapters and non-blocking switches**
- **MPI: Intel MPI 4.0, MVAPICH2 1.5.1p1, Open MPI 1.5.1, Platform MPI 8.0.1**
- **Compilers: GNU Compilers 4.1.2, Intel Compilers 11.1**
- **Storage: Lustre 1.8.5**
- **Application: AMR (2006 version of the code)**
- **Benchmark dataset: Standard**

About Intel® Cluster Ready

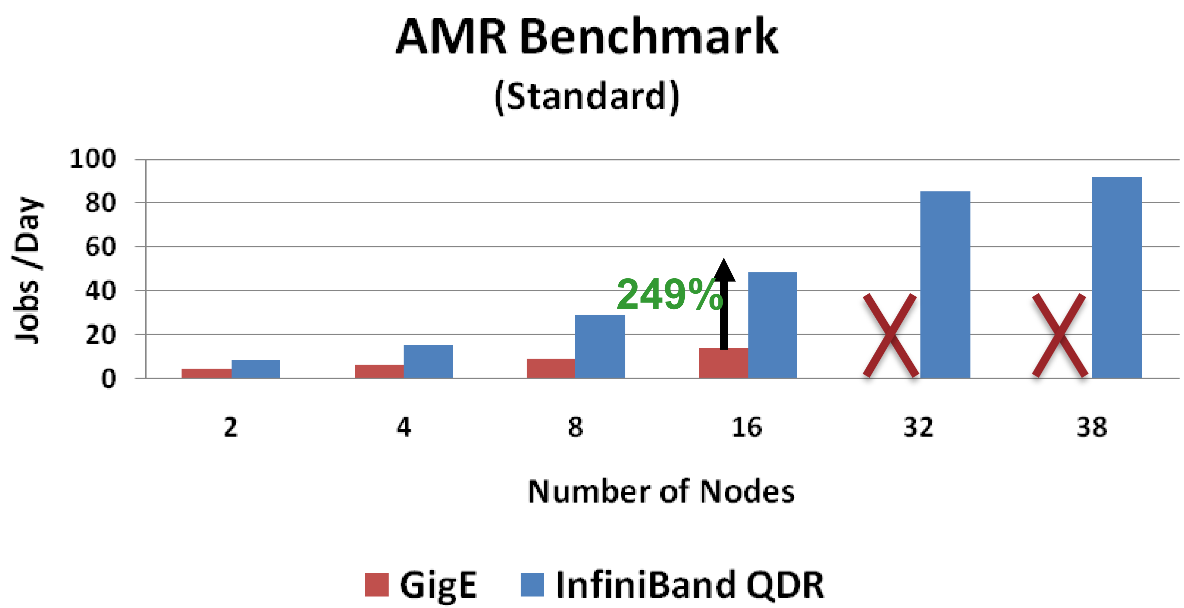
- **Intel® Cluster Ready systems make it practical to use a cluster to increase your simulation and modeling productivity**
  - Simplifies selection, deployment, and operation of a cluster
- **A single architecture platform supported by many OEMs, ISVs, cluster provisioning vendors, and interconnect 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

Dell PowerEdge Servers helping Simplify IT

- **System Structure and Sizing Guidelines**
  - 38-node cluster build with Dell PowerEdge™ M610 blade servers
  - Servers optimized for High Performance Computing environments
  - Building Block Foundations for best price/performance and performance/watt
- **Dell HPC Solutions**
  - Scalable Architectures for High Performance and Productivity
  - Dell's comprehensive HPC services help manage the lifecycle requirements.
  - Integrated, Tested and Validated Architectures
- **Workload Modeling**
  - Optimized System Size, Configuration and Workloads
  - Test-bed Benchmarks
  - ISV Applications Characterization
  - Best Practices & Usage Analysis

AMR Performance – Interconnects

- **InfiniBand enables higher throughput and cluster productivity**
  - Shows performance gain over GigE starting with 2-node
  - Up to 249% gain in productivity over GigE on a 16-node cluster
- **The performance gap widens as the node count increases**
  - 4 InfiniBand QDR nodes with outperforms 16 GigE nodes
- **GigE testing is limited to 16-node due to switch port availability**



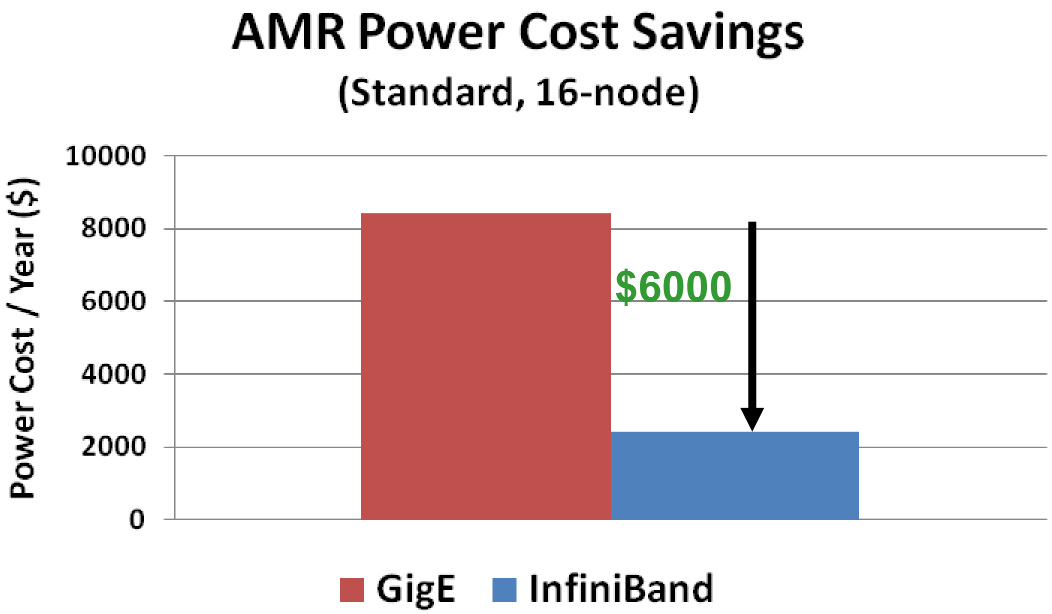
Higher is better

12 Cores/Node

Power Cost Savings with Different Interconnect

HPC ADVISORY COUNCIL

- **To finish the same number of AMR jobs with InfiniBand QDR or GigE**
  - Using InfiniBand QDR saves up to $6000 in electricity cost
  - Yearly based on a 16-node cluster
- **As cluster size increases, more power can be saved**



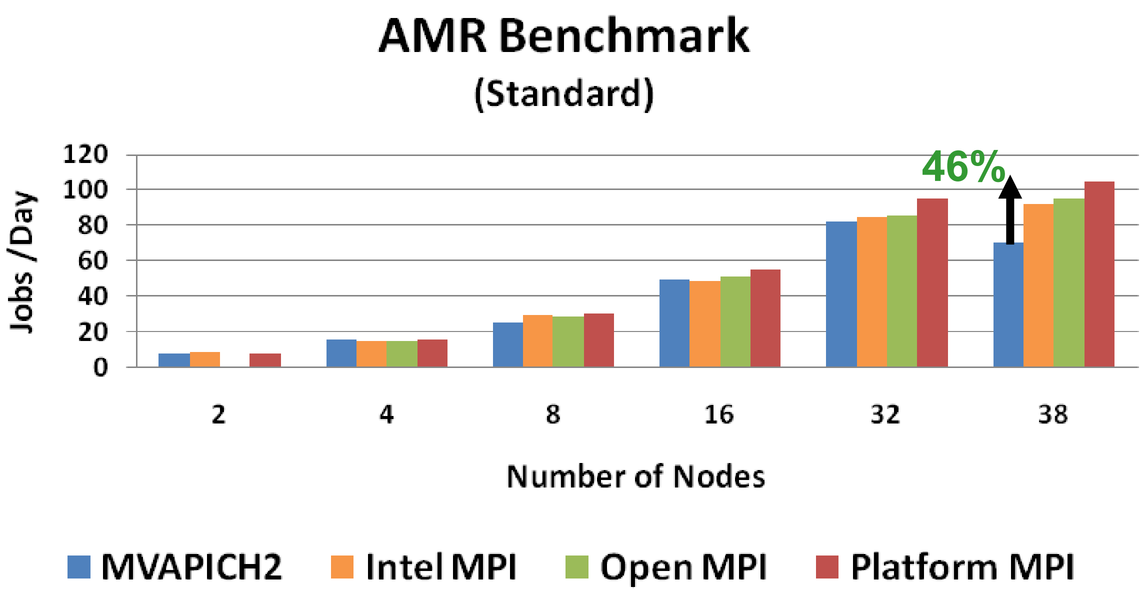
*$/KWh = KWh * $0.20*

For more information - http://hightech.lbl.gov/documents/DATA_CENTERS/svrpwrusecompletefinal.pdf

Lower is better

AMR Performance – MPI Implementations

- **Platform MPI shows the best scaling among all MPI implementations tested**
  - Shows 46% better compared to MVAPICH2
- **MVAPICH2 shows a sudden performance drop from 32-node to 38-node**
  - The exact cause is unknown but is reproducible only with MVAPICH2



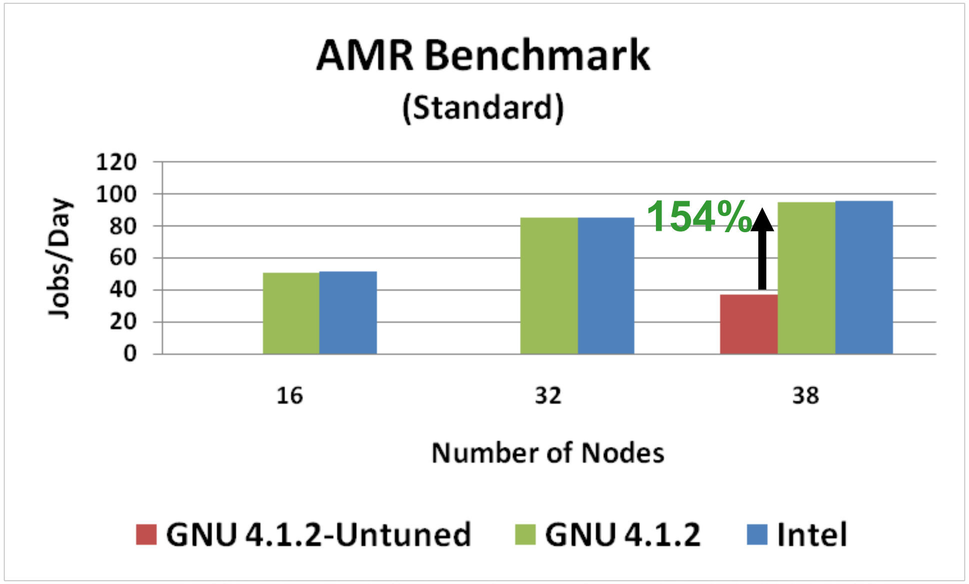
Higher is better

Intel Compilers

12 Cores/Node

AMR Performance – Compilers

- **Intel and tuned GNU compilers provide similar CPU utilization**
- **Tuned GNU compilers show better CPU utilization versus non-tuned GNU**
  – Up to 154% better performance than without using optimized flags
- **Compiler optimization flags used:**
  – Intel: " -O3 -ip -xSSE4.2 -w -ftz -align all -fno-alias -fp-model fast=1 -convert big_endian”
  – GNU: “-O3 -ffast-math -ftree-vectorize -ftree-loop-linear -funroll-loops”

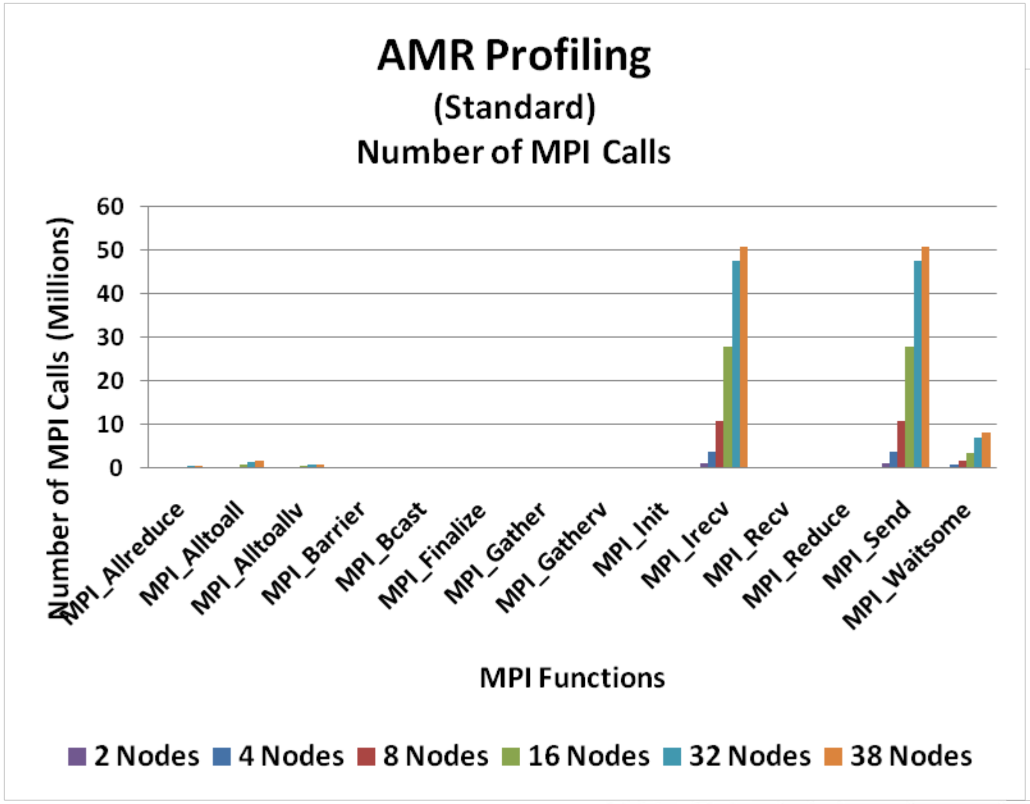


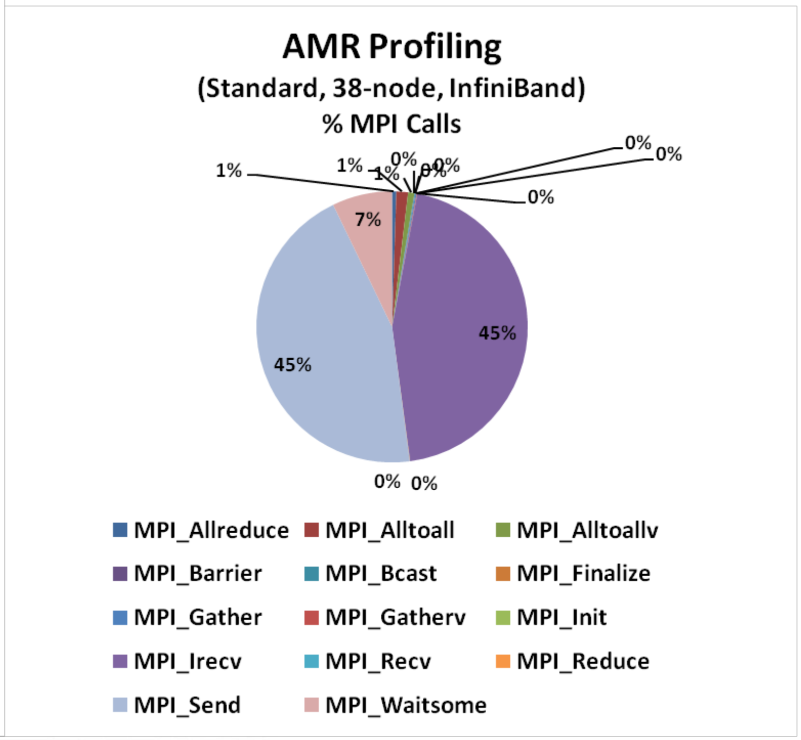
Higher is better

Open MPI 1.5
12 Cores/Node

AMR Profiling – Number of MPI Calls

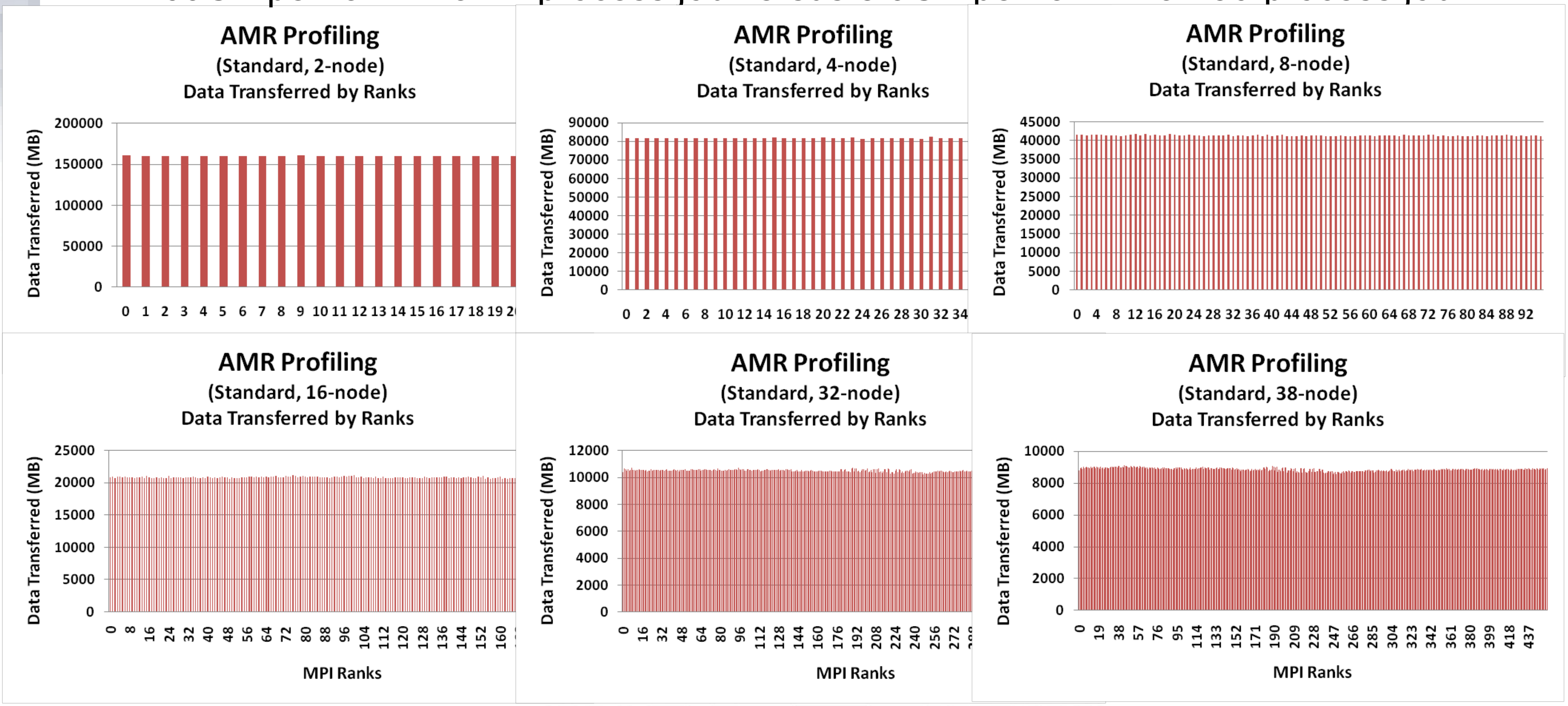
- **MPI_Irecv and MPI_Send dominates 90% of all MPI calls**
  - Each MPI call is accounted for about 45% of all MPI functions on a 38-node job
- **Non-blocking receives (MPI_Irecv) enable maximum efficiency**
  - Allow processes to compute while receiving in background
- **MPI calls increase proportionally with the node count**





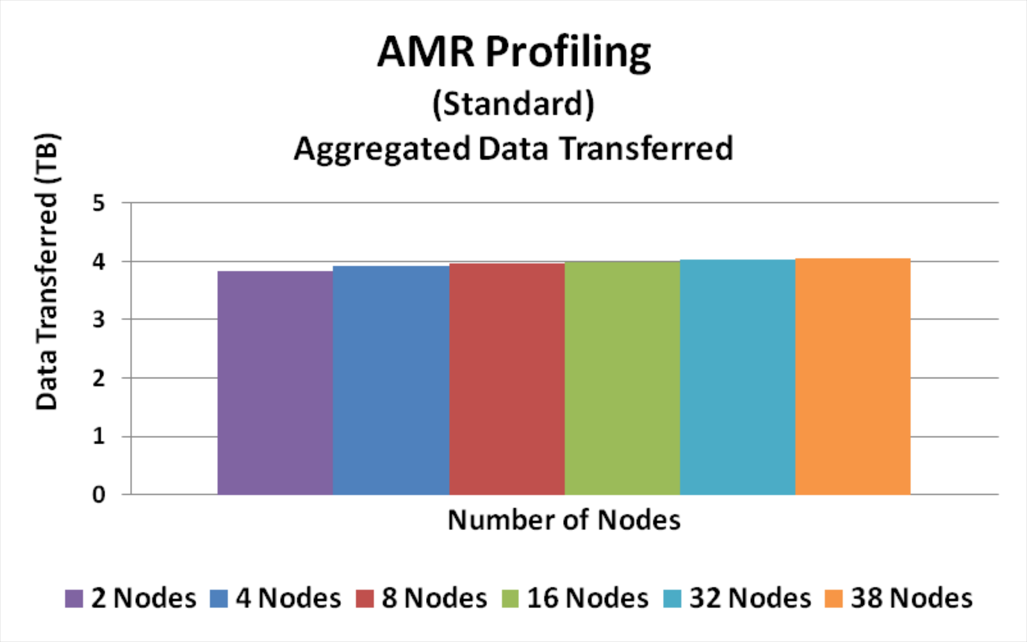
AMR Profiling – Data Transfer Per Process

- **Data transferred to each process is roughly the same**
  – Shows good balance in data distributions and job separation for computation
- **As the cluster scales, less data is driven per rank and per node**
  – 160GB per rank in a 24-process job versus 8.9GB per rank in a 456-process job



AMR Profiling – Aggregated Data Transfer

- **Aggregated data transfer refers to:**
  - Total amount of data being transferred in the network between all MPI ranks collectively
- **The total data transfer remains roughly the same as the cluster scales**
  - AMR can efficiently distributes data without generating extra data overheads on network
- **Demonstrates the advantage and importance of scalable network interconnect**
  - InfiniBand QDR can deliver bandwidth needed to push 4TB of data across the network

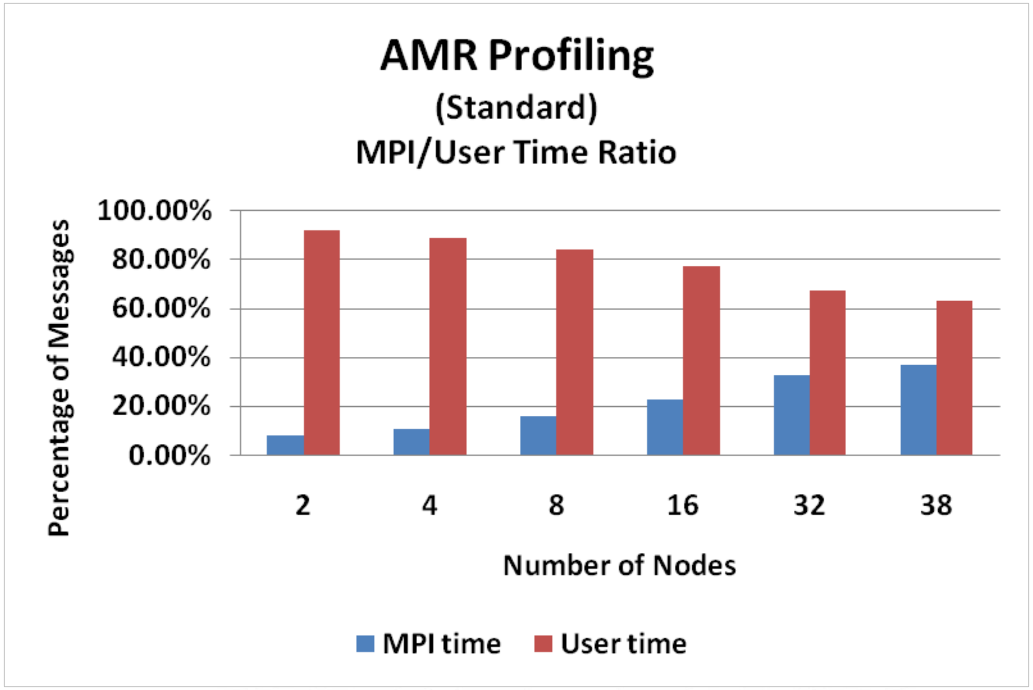


InfiniBand QDR

AMR Profiling – MPI/User Time Ratio

HPC ADVISORY COUNCIL

- **The MPI/User time ratio shows AMR is a compute-bound application**
  - More than 80% of the time spent on user code with the standard dataset
  - A small time percentage is spent for communications between the MPI ranks
- **Computational work is reduced per node as the cluster size increases**
  - More nodes take on computation, thus reduces percentage in user time

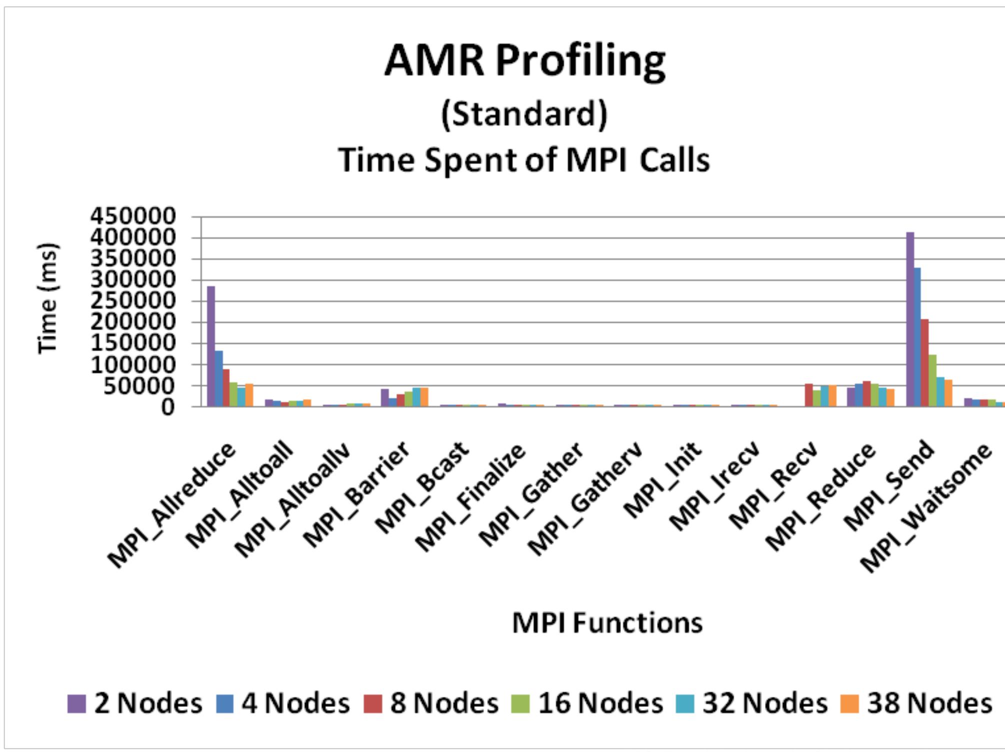


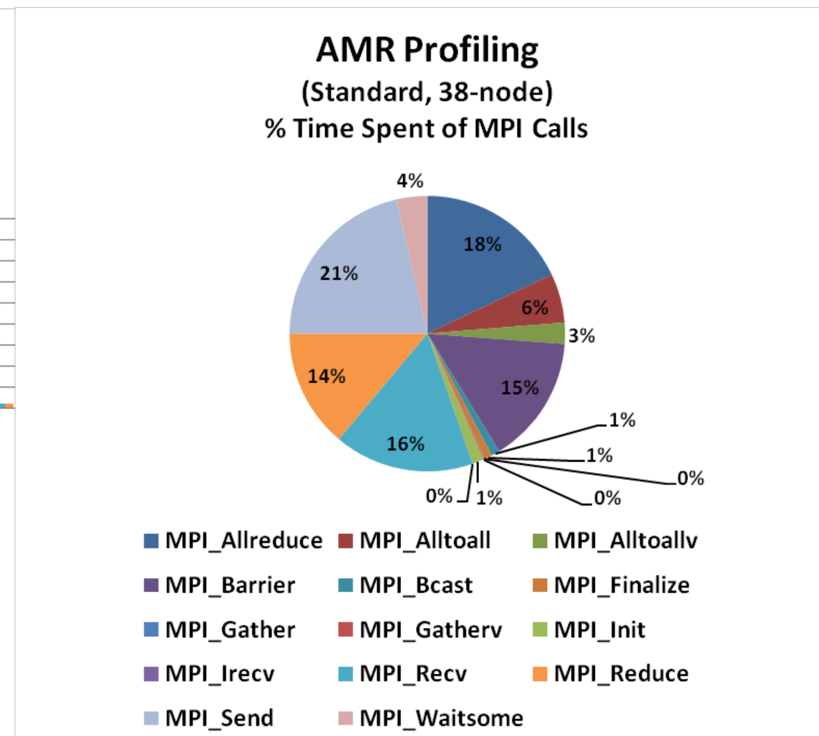
Higher is better

12 Cores/Node

AMR Profiling – Time Spent of by MPI Calls

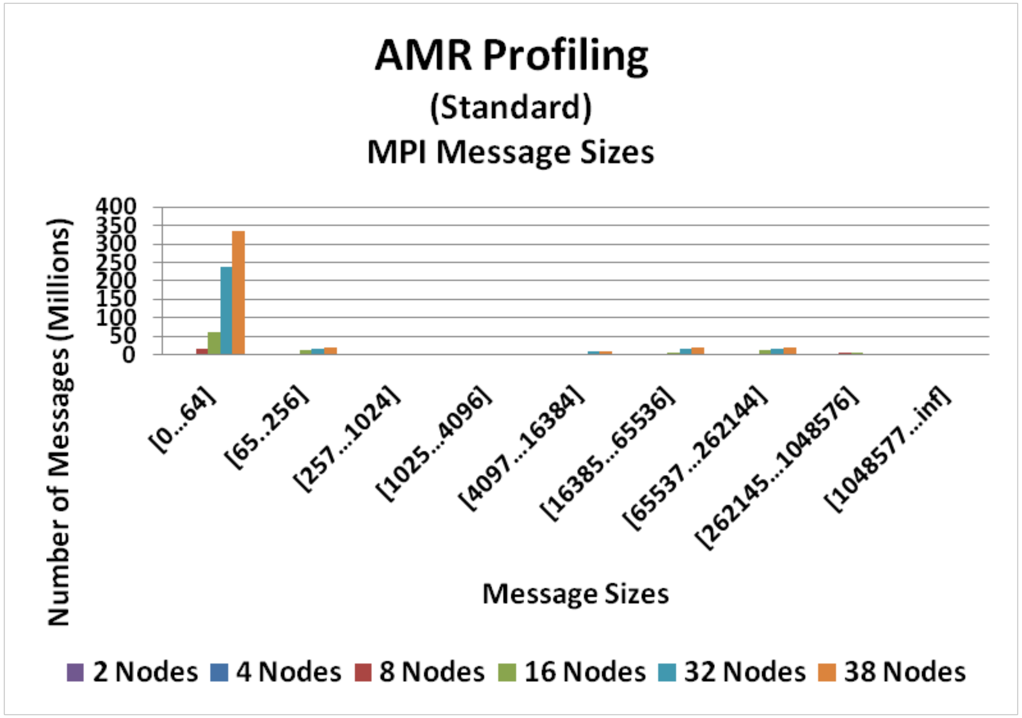
- **MPI_Send and MPI_Allreduce consume the most time on smaller node count**
  - Data transfer time (as in send and allreduce) is lowered dramatically
- **Communication time is reduced dramatically on larger node count**
  - As the data load is being spread across to more nodes on the network





AMR Profiling – MPI Message Sizes

- **Data transferred are concentrated in the small messages**
  - In the range between 0-byte to 64-byte
  - Small messages are generally for data synchronizations
- **Messages remains at the same sizes as the node count increases**



Conclusions

HPC ADVISORY COUNCIL

- **AMR with the standard dataset is mainly a compute-bound application**
  - Spends majority of the time in user time computation
  - Using optimized flags help to speed up computation on a per-node basis
- **AMR is sensitive to network interconnect performance**
  - Requires solid network interconnect for good data exchanges
  - InfiniBand outperforms GigE by providing network throughput needed for computation
- **Network interconnect performance becomes more important as cluster scales**
  - Shows roughly 40% of the time is spent on communications at 38-node
  - Computational work is spread across more nodes to reduce overall job run time
- **AMR allows efficient data transfer**
  - Use of non-blocking receives (MPI_Irecv) to allow computation while in data transfers
  - No extra overhead to the network as the cluster scales

Thank You

HPC Advisory Council