Mellanox TECHNOLOGIES
HPC ADVISORY COUNCIL

NETWORK OF EXPERTISE

ECLIPSE Performance Benchmarks and Profiling

January 2009

AMD Smarter Choice

DELL

Mellanox TECHNOLOGIES

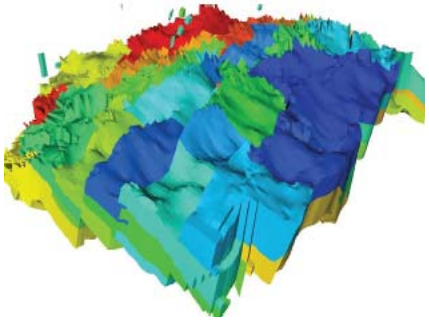
Schlumberger

Note

- **The following research was performed under the HPC Advisory Council activities**
  - AMD, Dell, Mellanox, Schlumberger
  - HPC Advisory Council Cluster Center
- **For more info please refer to**
  - www.mellanox.com, www.dell.com/hpc, www.amd.com, http://www.slb.com/

Schlumberger ECLIPSE

- **Oil and gas reservoir simulation software**
  - Developed by Schlumberger
- **Offers multiple choices of numerical simulation techniques for accurate and fast simulation for**
  - Black-oil
  - Compositional
  - Thermal
  - Streamline
  - Others



- **ECLIPSE support MPI to achieve high performance and scalability**

Objectives

- **The presented research was done to provide best practices**
  - ECLIPSE performance benchmarking
  - Interconnect performance comparisons
  - Ways to increase ECLIPSE productivity
  - Understanding ECLIPSE communication patterns
  - Power-efficient simulations

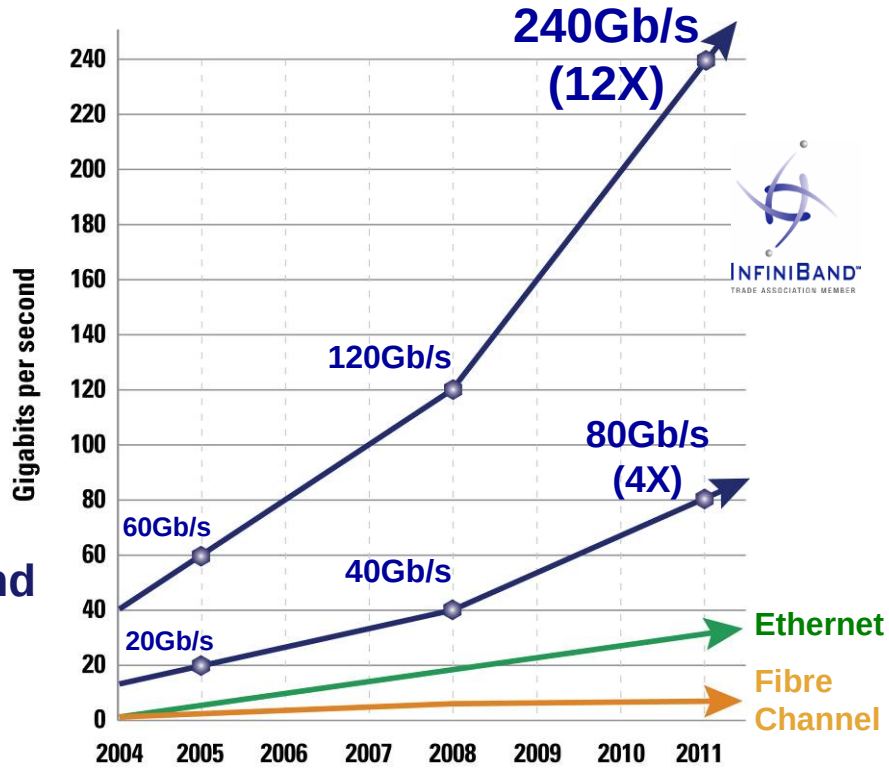
Test Cluster Configuration – System Upgrade

- **Dell™ PowerEdge™ SC 1435 24-node cluster**
- **Quad-Core AMD Opteron™ Model 2382 processors (“Shanghai”)**
- **Mellanox® InfiniBand ConnectX® DDR HCAs**
- **Mellanox® InfiniBand DDR Switch**
- **Memory: 16GB memory, DDR2 800MHz per node**
- **OS: RHEL5U2, OFED 1.3 InfiniBand SW stack**
- **MPI: Platform MPI 5.6.5**
- **Application: Schlumberger ECLIPSE Simulators 2008.2**
- **Benchmark Workload**
  - 4 million cell model (2048 200 10) Blackoil 3 phase model with ~ 800 wells

Mellanox InfiniBand Solutions

- **Industry Standard**
  - Hardware, software, cabling, management
  - Design for clustering and storage interconnect
- **Price and Performance**
  - 40Gb/s node-to-node
  - 120Gb/s switch-to-switch
  - 1us application latency
  - Most aggressive roadmap in the industry
- **Reliable with congestion management**
- **Efficient**
  - RDMA and Transport Offload
  - Kernel bypass
  - CPU focuses on application processing
- **Scalable for Petascale computing & beyond**
- **End-to-end quality of service**
- **Virtualization acceleration**
- **I/O consolidation Including storage**

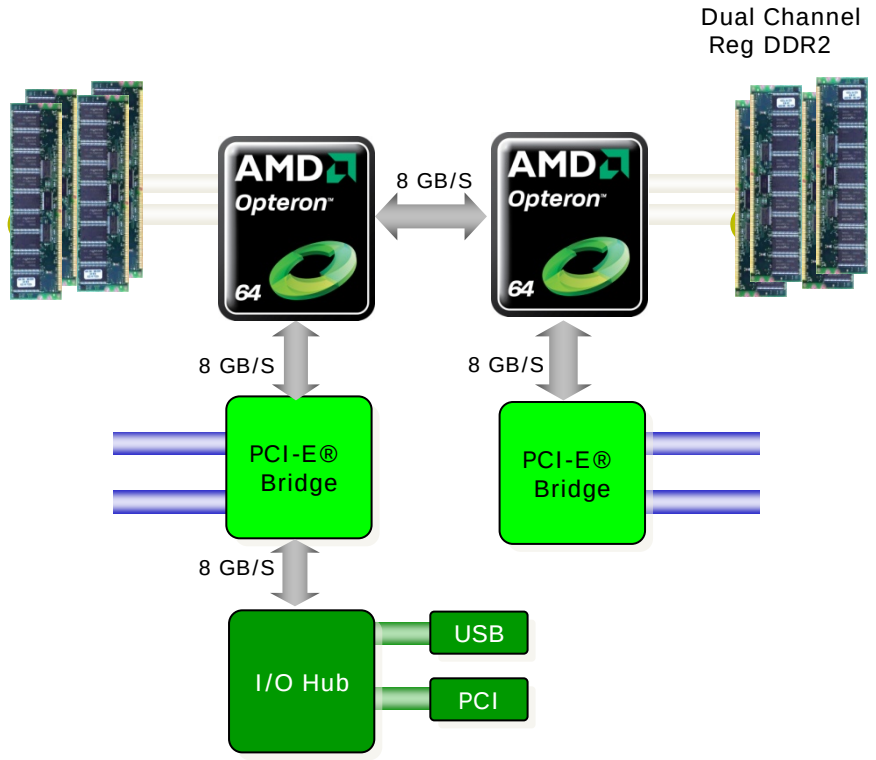
The InfiniBand Performance Gap is Increasing



InfiniBand Delivers the Lowest Latency

Quad-Core AMD Opteron™ Processor

- **Performance**
  - Quad-Core
    - Enhanced CPU IPC
    - 4x 512K L2 cache
    - 6MB L3 Cache
  - Direct Connect Architecture
    - HyperTransport™ technology
    - Up to 24 GB/s peak per processor
  - Floating Point
    - 128-bit FPU per core
    - 4 FLOPS/clk peak per core
  - Integrated Memory Controller
    - Up to 12.8 GB/s
    - DDR2-800 MHz or DDR2-667 MHz
- **Scalability**
  - 48-bit Physical Addressing
- **Compatibility**
  - Same power/thermal envelopes as 2nd / 3rd generation AMD Opteron™ processor

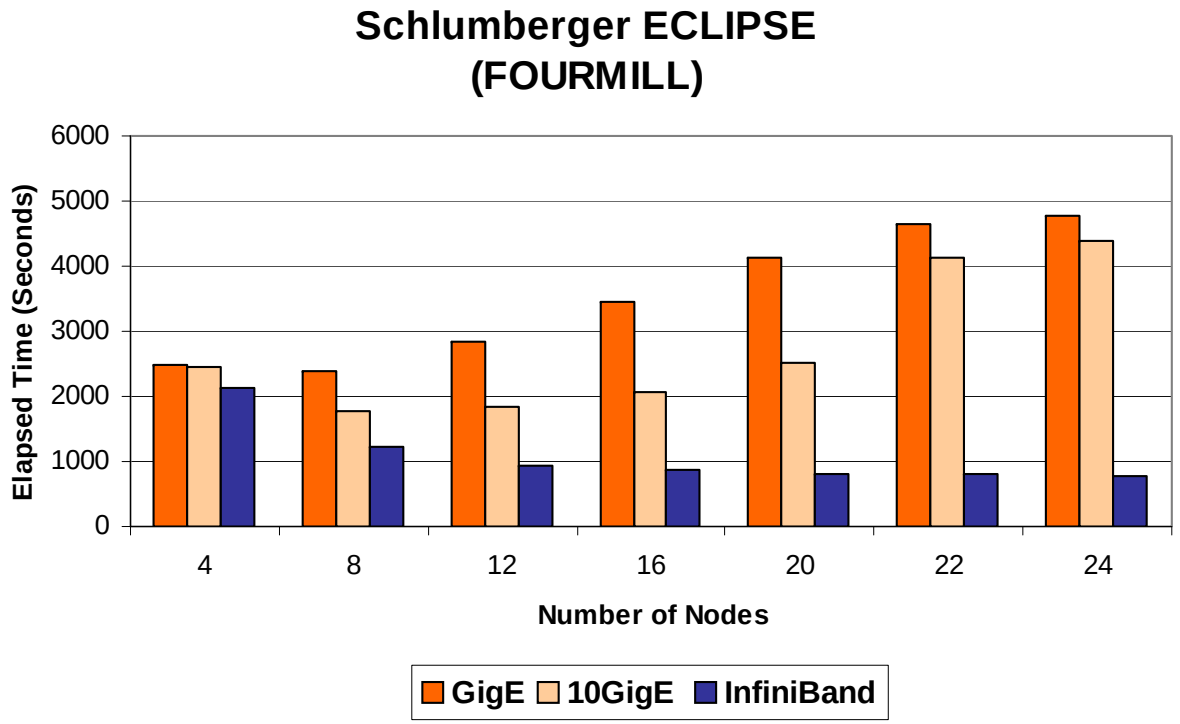


Dell PowerEdge Servers helping Simplify IT

- **System Structure and Sizing Guidelines**
  - 24-node cluster build with Dell PowerEdge™ SC 1435 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

ECLIPSE Performance Results - Interconnect

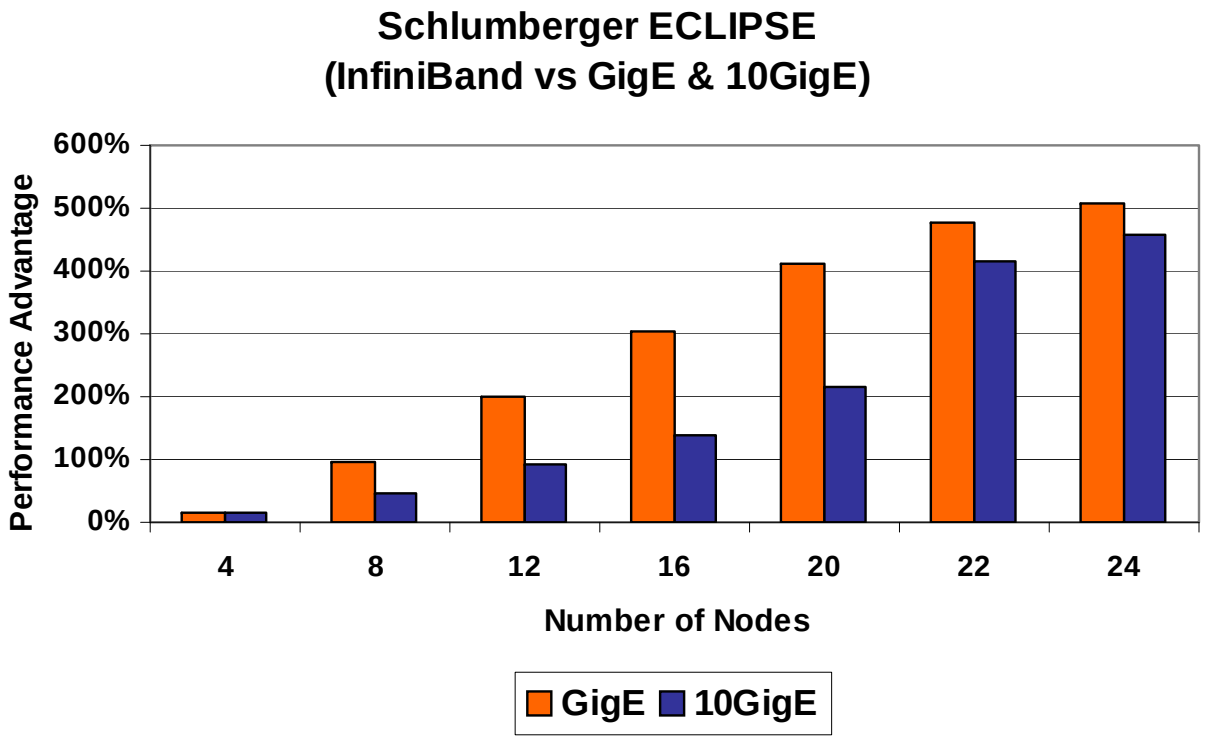
- **InfiniBand enables highest scalability**
  - Performance accelerates with cluster size
- **Performance over GigE and 10GigE is not scaling**
  - Slowdown occurs beyond 8 nodes



Lower is better

Single job per cluster size

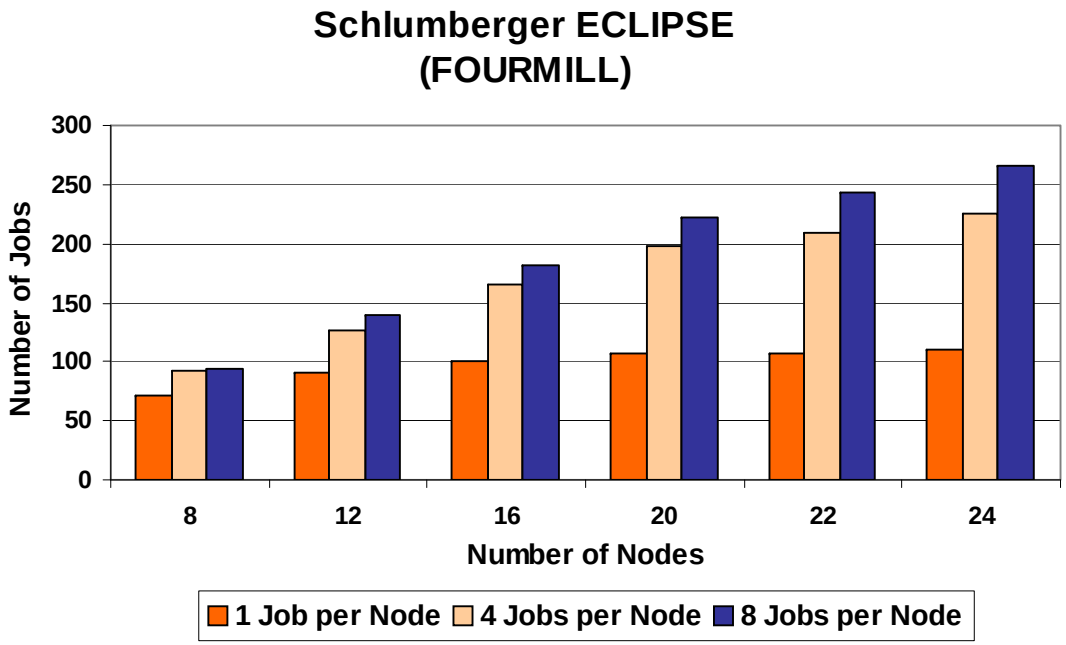
ECLIPSE Performance Results - Interconnect



- **InfiniBand outperforms GigE by up to 500% and 10GigE by up to 457%**
  - As node number increases, bigger advantage is gained

ECLIPSE Performance Results - Productivity

- **InfiniBand increases productivity by allowing multiple jobs to run simultaneously**
  - Providing required productivity for reservoir simulations
- **Three cases are presented**
  - Single job over the entire systems
  - Four jobs, each on two cores per CPU per server
  - Eight jobs, each on one CPU core per server
- **Eight jobs per node increases productivity by up to 142%**

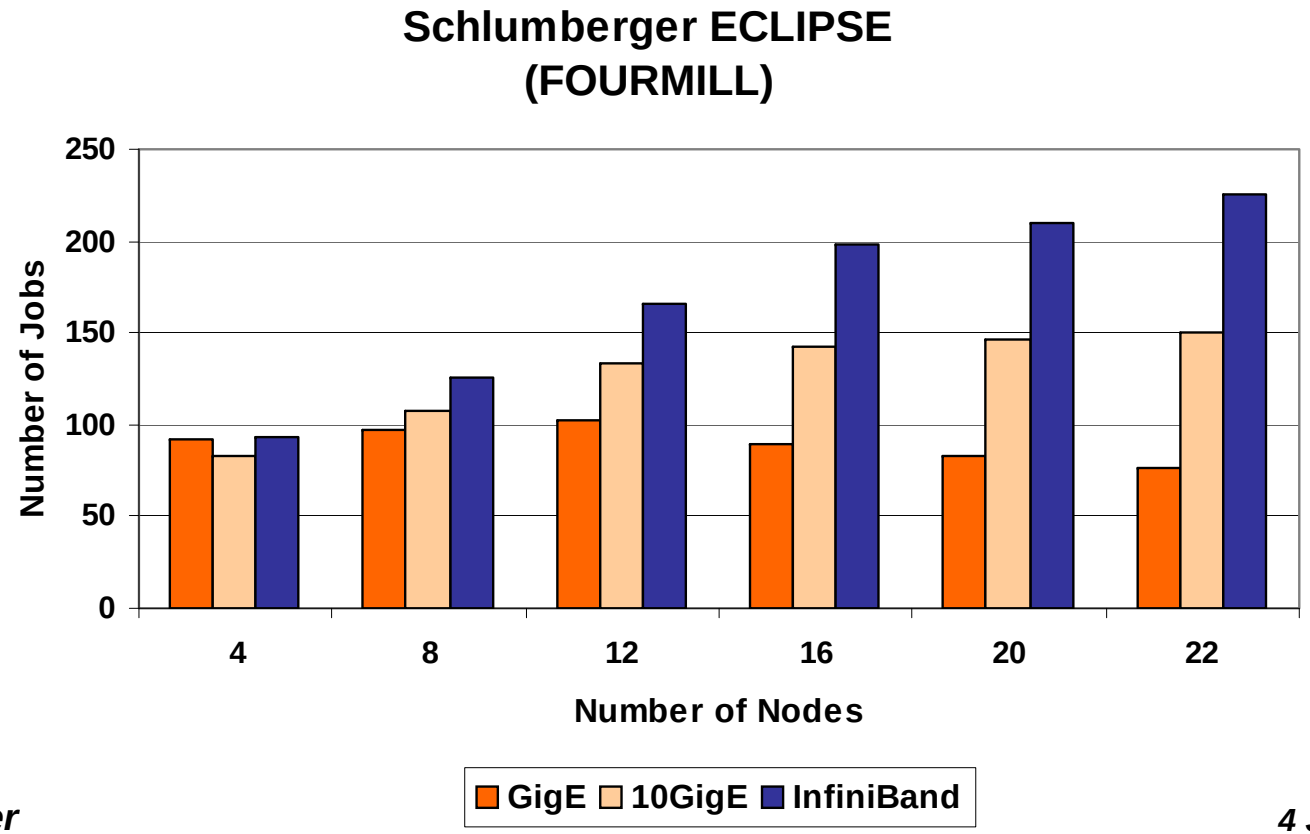


Higher is better

InfiniBand

ECLIPSE Performance Results - Productivity

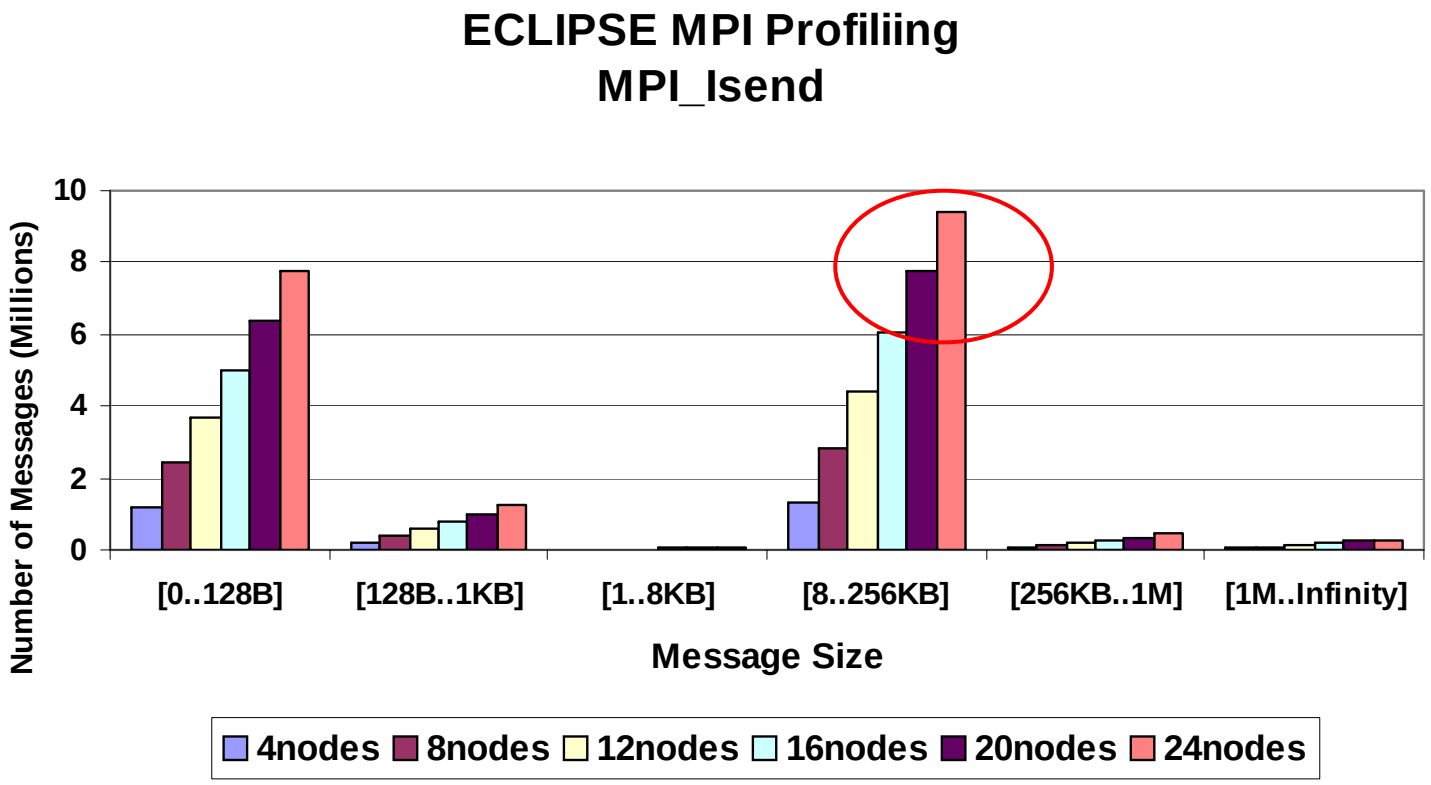
- **InfiniBand offers unparalleled productivity compared to Ethernet**
  - GigE shows performance decrease beyond 8 nodes
  - 10GigE demonstrates no scaling beyond 16 nodes



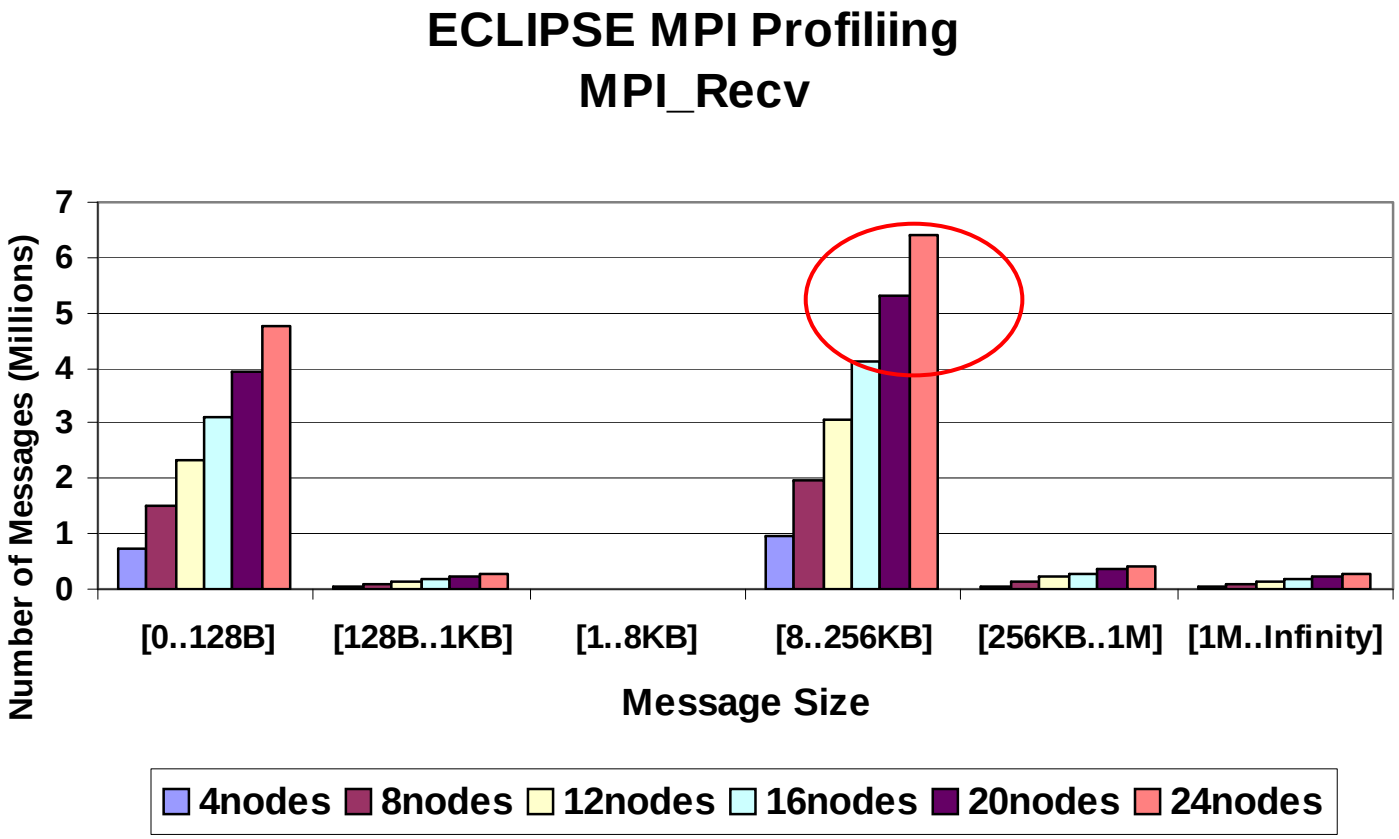
Higher is better

4 Jobs on each node

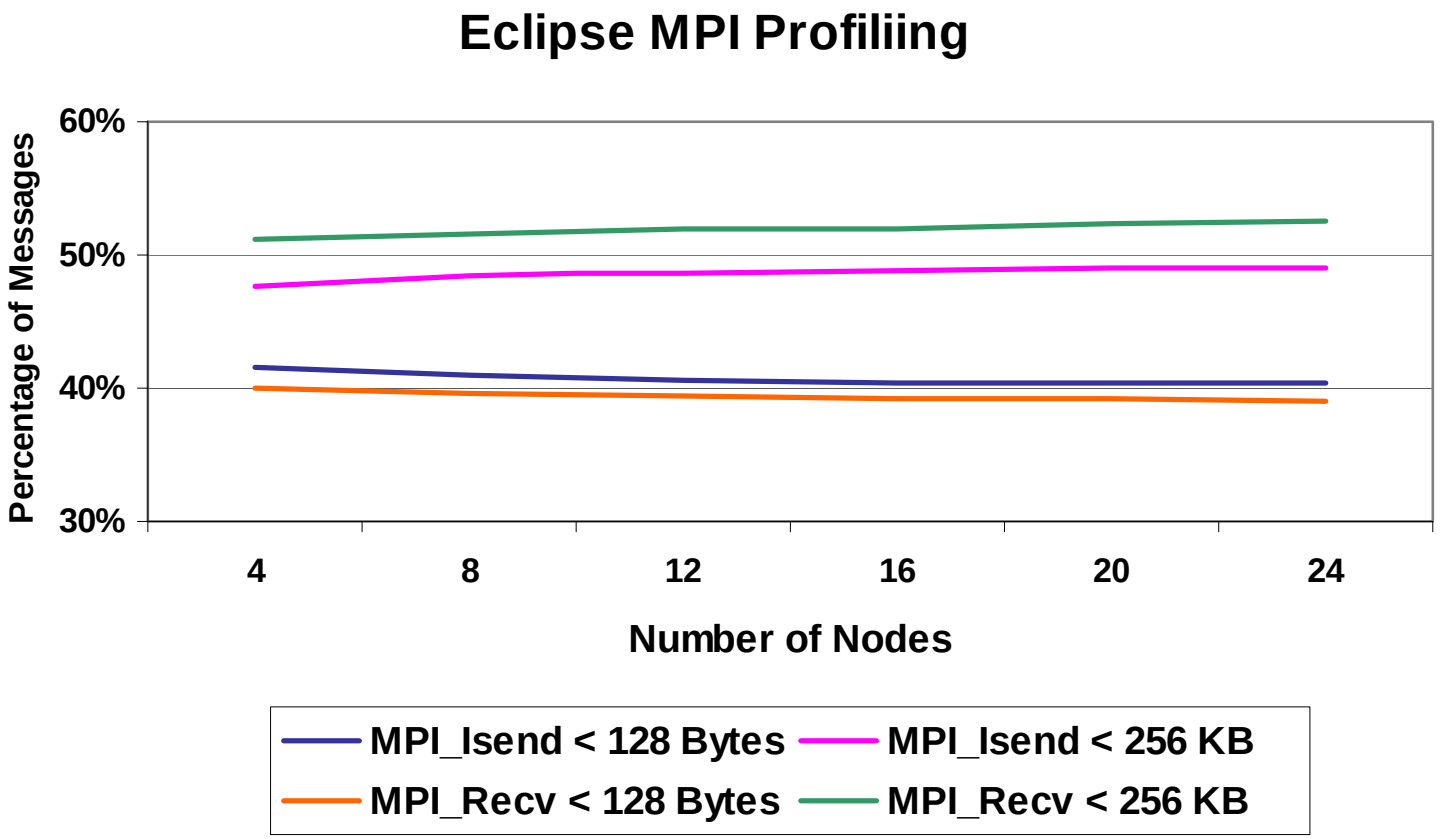
ECLIPSE Profiling – Data Transferred



ECLIPSE Profiling – Data Transferred



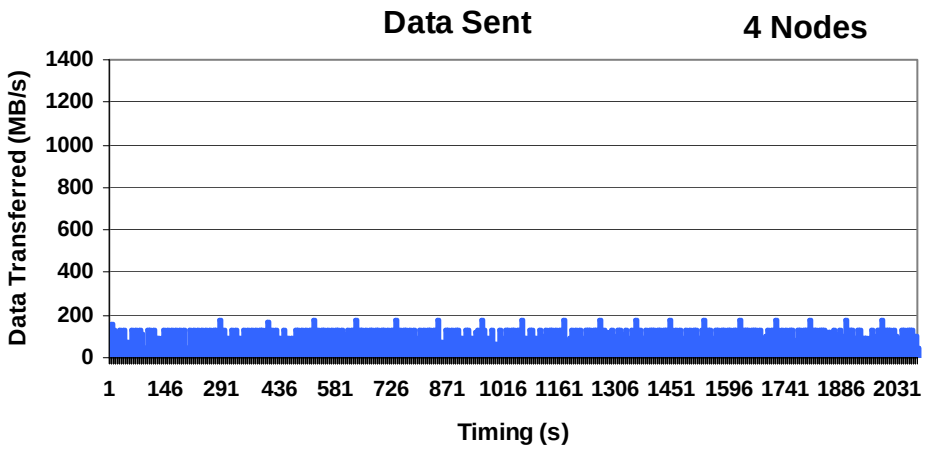
ECLIPSE Profiling – Message Distribution

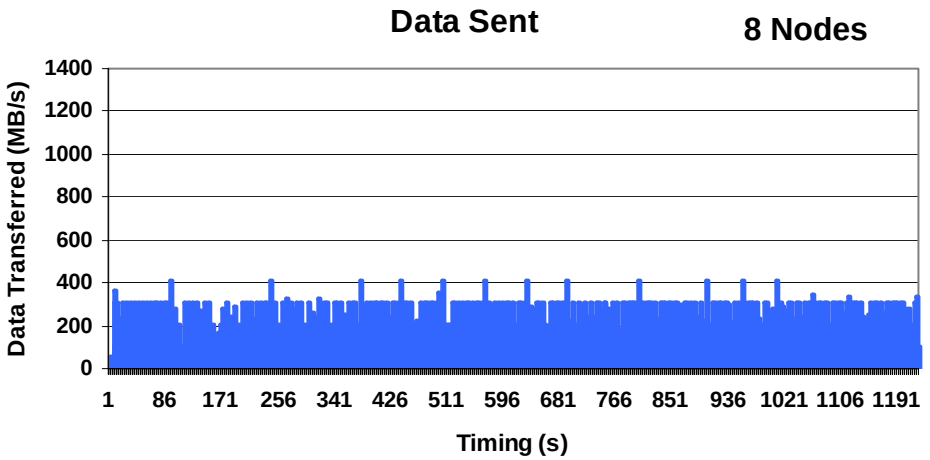


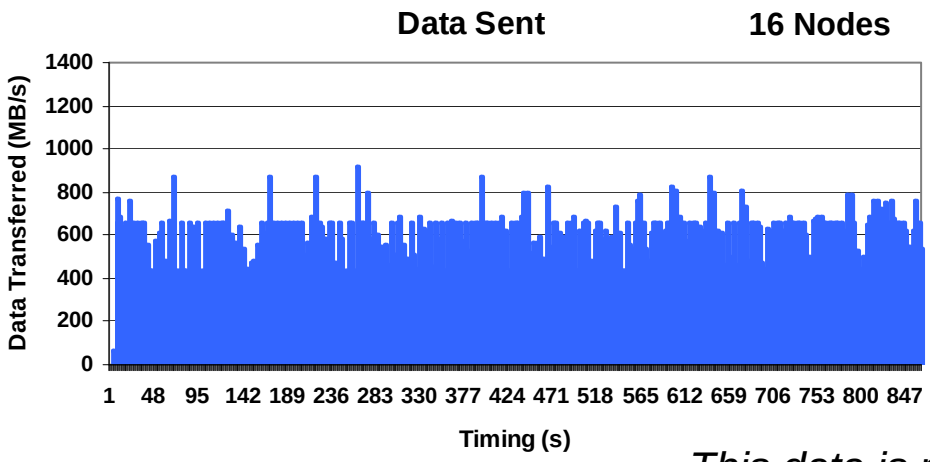
- Majority of MPI messages are large size
- Demonstrating the need for highest throughput

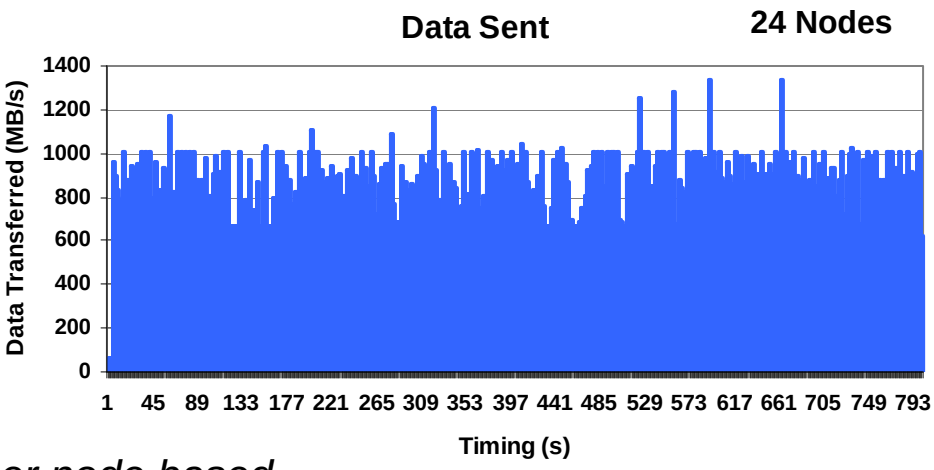
Interconnect Usage by ECLIPSE

- **Total server throughput increases rapidly with cluster size**









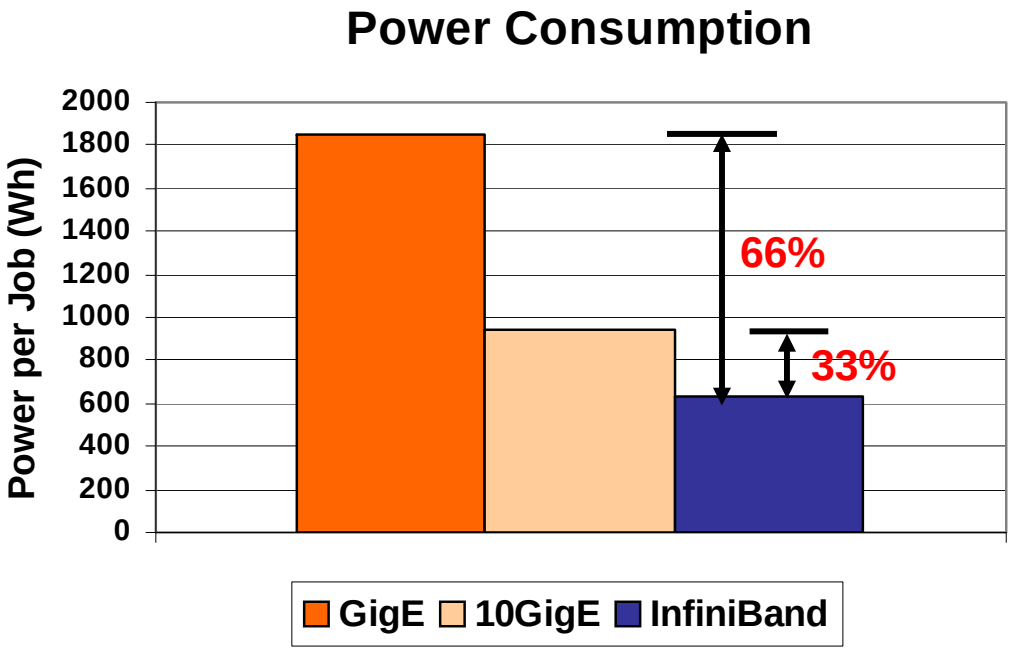
This data is per node based

ECLIPSE Profiling Summary - Interconnect

- **ECLIPSE was profiled to determine networking dependency**
- **Majority of data transferred between compute nodes**
  - Done with 8KB-256KB message size
  - Data transferred increases with cluster size
- **Most used message sizes**
  - <128B messages – mainly synchronizations
  - 8KB-256KB – data transferring
- **Message size distribution**
  - Percentage of smaller messages (<128B) slightly decreases with cluster size
  - Percentage of mid-size messages (8KB-256KB) increases with cluster size
- **ECLIPSE interconnects sensitivity points**
  - Interconnect latency and throughput for <256KB message range
  - As node number increases, interconnect throughput becomes more critical

Power Consumption/Productivity Comparison

- **InfiniBand enables power efficient simulations**
- **Reducing system power/job consumption up to 66% vs GigE and 33% vs 10GigE**
  - For productivity case – 4 jobs per node
  - When using single job approach, InfiniBand reduces power/job consumption by more than 82% compared to 10GigE



4 Jobs on each node

Conclusions

- **Eclipse is widely used to perform reservoir simulation**
  - Developed by Schulmberger
- **ECLIPSE performance and productivity relies on**
  - Scalable HPC systems and interconnect solutions
  - Low latency and high throughout interconnect technology
  - NUMA aware application for fast access to memory
  - Reasonable job distribution can dramatically improve productivity
    - Increasing number of jobs per day while maintaining fast run time
- **Interconnect comparison shows**
  - InfiniBand delivers superior performance and productivity in every cluster size
  - Scalability requires low latency and “zero” scalable latency
- **InfiniBand enables lowest power consumption per job**
  - Optimizing power/job ratio

Thank You

HPC Advisory Council

HPC@mellanox.com