



NETWORK OF EXPERTISE



# ECLIPSE Best Practices

## Performance, Productivity, Efficiency

March 2009



**Schlumberger**

# ECLIPSE Performance, Productivity, Efficiency



- **The following research was performed under the HPC Advisory Council activities**
  - HPC Advisory Council Cluster Center
- **Special thanks to AMD, Dell, Mellanox, Schlumberger**
  - In particular to
    - Joshua Mora (AMD)
    - Jacob Liberman (Dell)
    - Gilad Shainer and Tong Liu (Mellanox Technologies)
    - Owen Brazell (Schlumberger)
- **For more info please refer to**
  - <http://hpcadvisorycouncil.mellanox.com/>



- **Distinguished HPC alliance (OEMs, IHVs, ISVs, end-users)**
  - More than 60 members worldwide
  - 1st tier OEMs (AMD, Dell, HP, Intel, Sun) and systems integrators across the world
  - Leading OSVs and ISVs (LSTC, Microsoft, Schlumberger, Wolfram etc.)
  - Strategic end-users (Total, Fermi Lab, ORNL, OSU, VPAC etc.)
  - Storage vendors (DDN, IBRIX, LSI, Panasas etc)
- **The Council mission**
  - Bridge the gap between high-performance computing (HPC) use and its potential
  - Bring the beneficial capabilities of HPC to new users for better research, education, innovation and product manufacturing
  - Bring users the expertise needed to operate HPC systems
  - Provide application designers with the tools needed to enable parallel computing
  - Strengthen the qualification and integration of HPC system products

# HPC Advisory Council Activities



Home | Blog | Council Members | Cluster Center | Network of Experts | Technical Content | Contact



**Mellanox Cluster Center**  
Apply for cluster access

The HPC Advisory Council is also a community effort support center for HPC end-users, providing the following capabilities:

- Mellanox Cluster Center - the center provides a unique ability to access the latest Mellanox and the HPC Advisory Council member technology, even before it reaches the public availability. It provides the Council members and any HPC end user with a development, testing, benchmarking and optimization environment.
- HPC Advisory Council support group - provide a support center for consultations, operations, issues etc. for the HPC end-users

**JOIN TODAY!** To become an HPC Advisory Council member please refer to the HPC Advisory Council Application (PDF)

- READ THE HPC ADVISORY COUNCIL BLOG
- Current Member Roster



**Performance on Overdrive**


Mellanox is dedicated to building a distinguished HPC alliance by working closely with our chosen partners and customers to ensure the best total solution is available to end-customers. The HPC Advisory Council includes best-in-class original equipment manufacturers (OEMs), strategic technology suppliers, independent software vendors (ISVs) and selected end-users across the entire HPC market segments.

## Network of Expertise

## Best Practices



- Oil and Gas
- Automotive
- Bioscience
- Weather
- CFD
- Quantum Chemistry
- and more....

## HPC Outreach and Education



### Torrey Pines High School

A PROUD MEMBER OF THE SAN DIEGUITO UNION HIGH SCHOOL DISTRICT

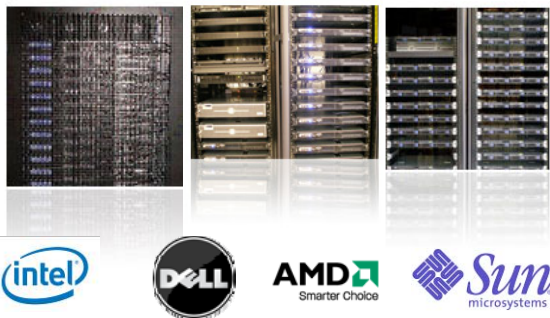




Mathematics Club Torrey Pines High School 2007-2008



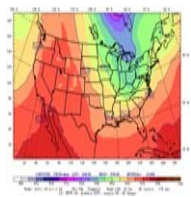

## Cluster Center

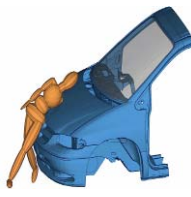
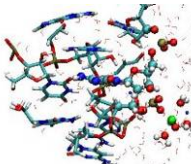
End-user applications benchmarking center

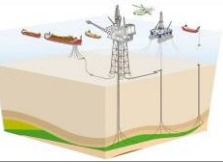




THE WEATHER RESEARCH & FORECASTING MODEL


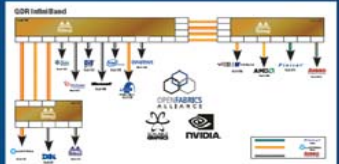





## HPC Technology Demonstration

### 40Gb/s InfiniBand Distributed Visualization over SCinet

40Gb/s InfiniBand SCinet Participants



- **Ways to improve performance, productivity, efficiency**
  - Knowledge, expertise, usage models
- **The following presentation reviews:**
  - ECLIPSE performance benchmarking
  - Interconnect performance comparisons
  - ECLIPSE productivity
  - Understanding ECLIPSE communication patterns
  - Power-aware simulations

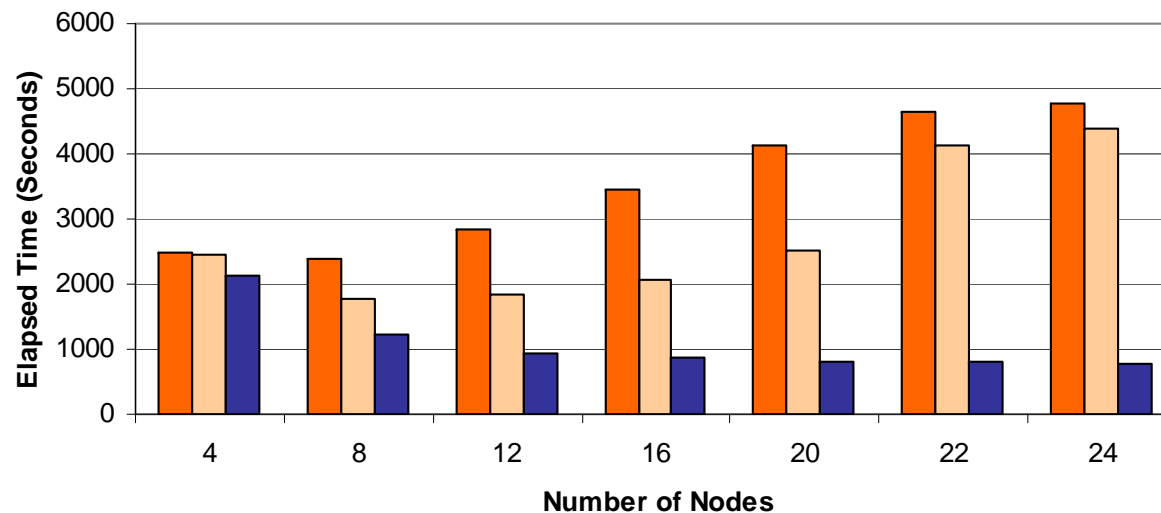
# Test Cluster Configuration

- **Dell™ PowerEdge™ SC 1435 24-node cluster**
  - Dual socket 1-U rack server supports 8x PCIe and 800MHz DDR2 DIMMs
  - Energy efficient building block for scalable and productive high performance computing
- **Quad-Core AMD Opteron™ Model 2382 processors (“Shanghai”)**
  - Industry leading technology that delivers performance and power efficiency
  - Opteron™ processors provide up to 21 GB/s peak bandwidth per processor
- **Mellanox® InfiniBand ConnectX® DDR HCAs and Mellanox InfiniBand DDR Switch**
  - High-performance I/O consolidation interconnect – transport offload, HW reliability, QoS
  - Supports up to 40Gb/s, 1usec latency, high message rate, low CPU overhead and zero scalable latency
- **Application: Schlumberger ECLIPSE Simulators 2008.2**
  - ECLIPSE is the market leading reservoir simulator
  - Has become the de-facto standard for reservoir simulation in the oil and gas industry
- **Benchmark Workload**
  - 4 million cell model ( 2048 200 10) Blackoil 3 phase model with ~ 800 wells

# ECLIPSE Performance - Interconnect

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

**Schlumberger ECLIPSE  
(FOURMILL)**



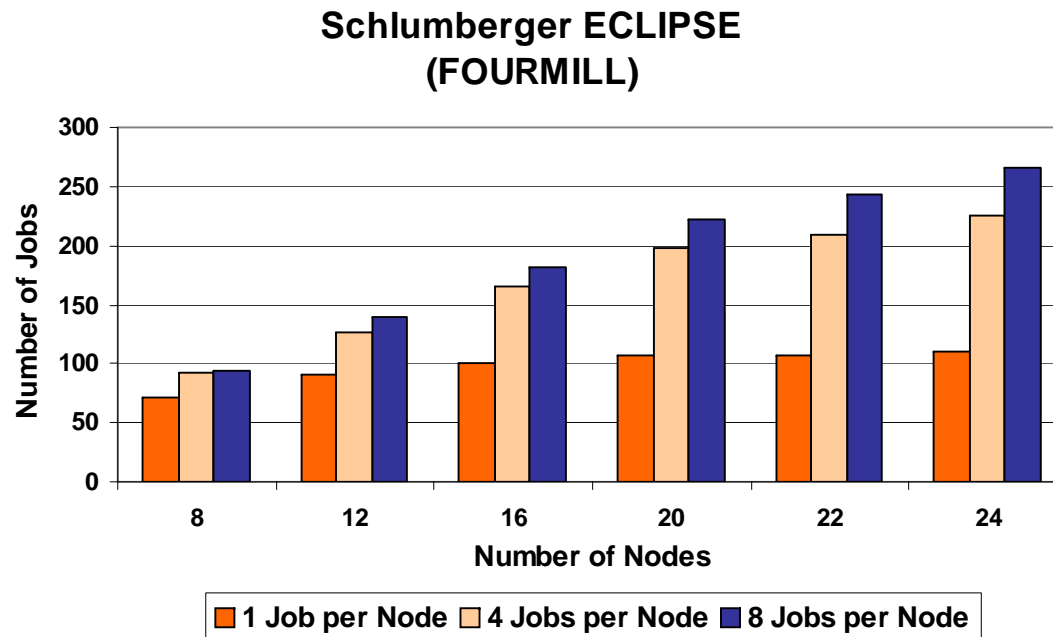
*Lower is better*

■ GigE ■ 10GigE ■ InfiniBand

*Single job per cluster size*

# ECLIPSE Productivity

- **By allowing multiple jobs to run simultaneously, ECLIPSE productivity can be increased**
  - Maximum gain with one job per core, maximum % improve with 1 job per 2 cores
- **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%**



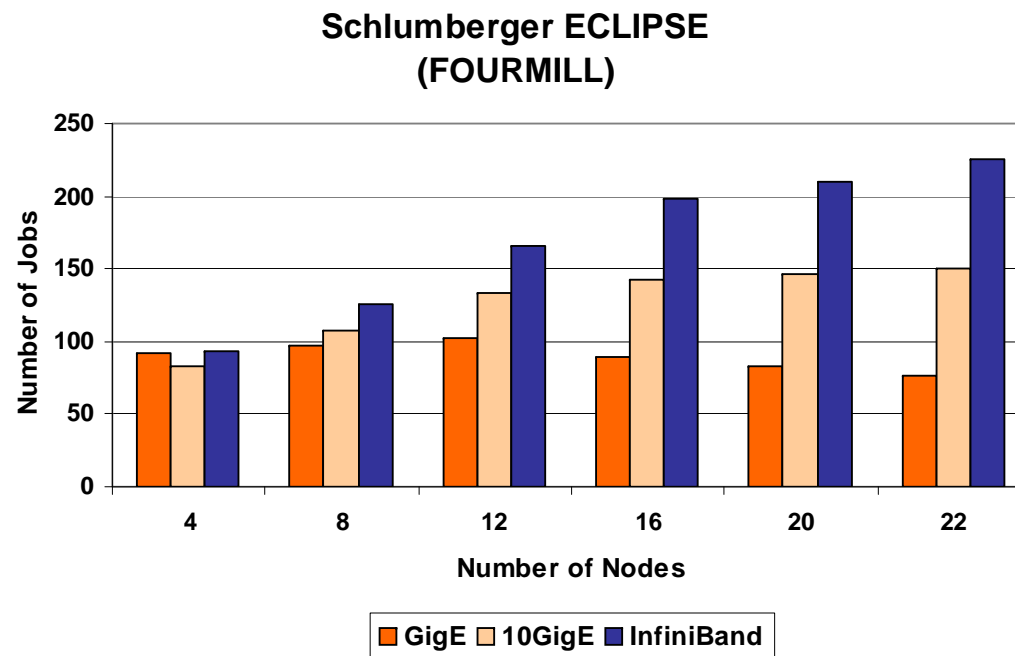
*Higher is better*

*InfiniBand*



# ECLIPSE Productivity – Interconnect Comparison

- **Productivity improvement depends on the cluster interconnect**
  - As number of job increases, I/O becomes the system bottleneck
- **InfiniBand demonstrates scalable productivity compared to Ethernet**
  - GigE shows performance decrease beyond 8 nodes
  - 10GigE demonstrates no scaling beyond 16 nodes

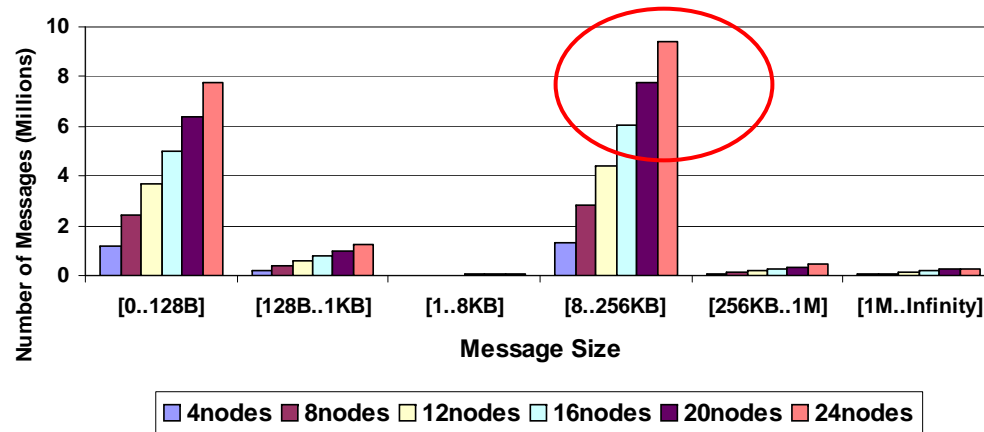


*Higher is better*

*4 Jobs on each node*

# ECLIPSE Profiling – Data Transferred

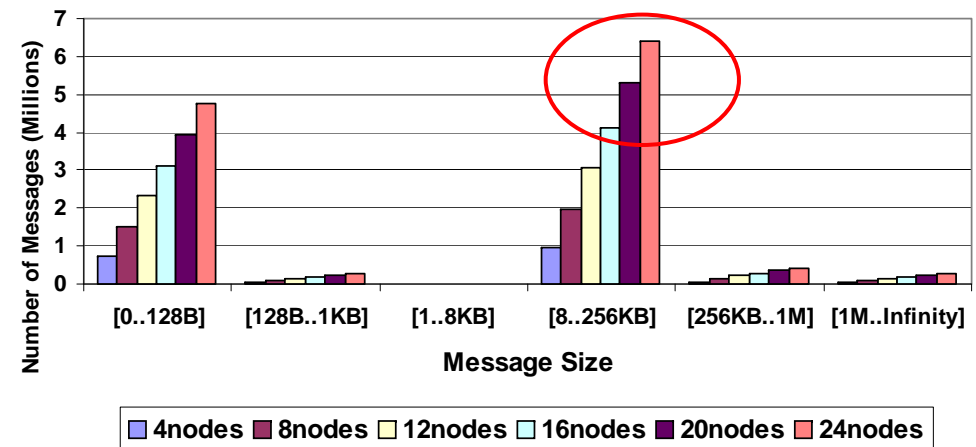
ECLIPSE MPI Profiling  
MPI\_Isend



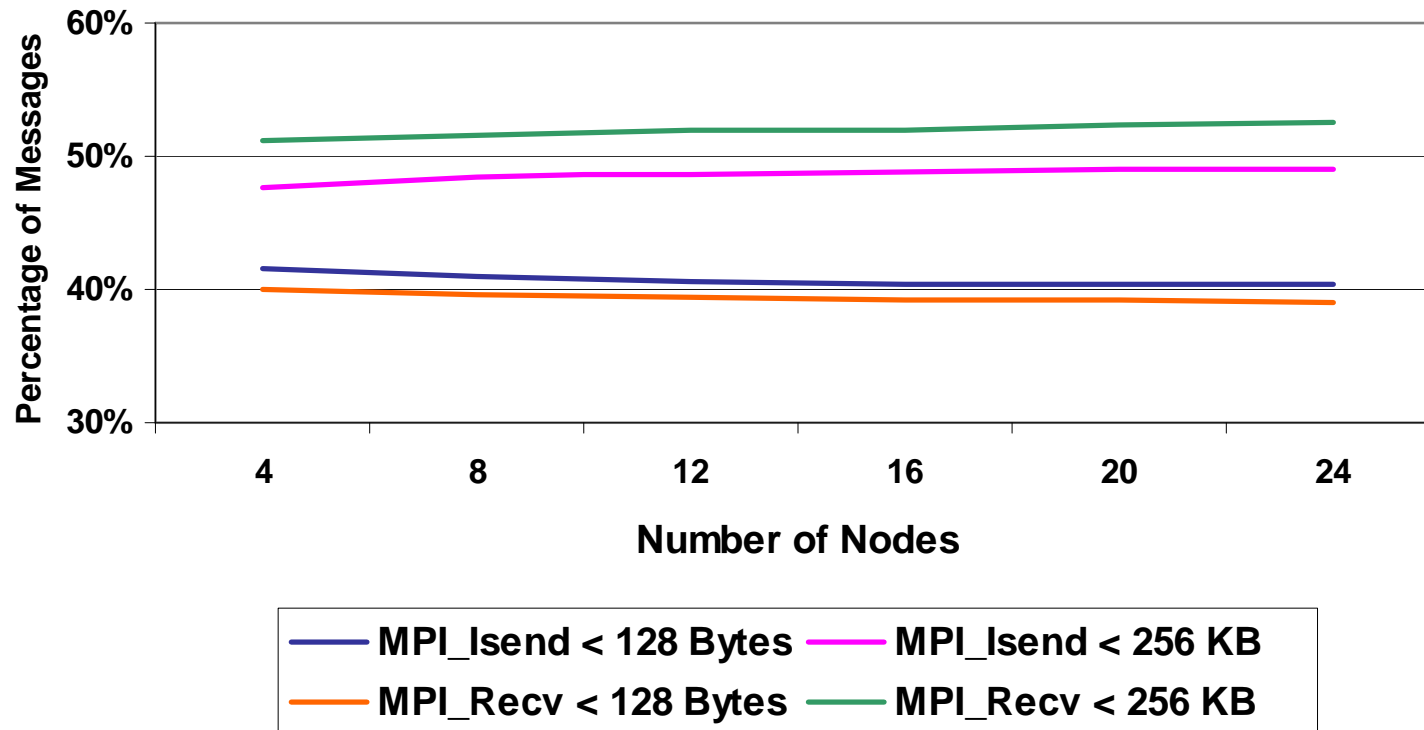
- MPI message profiling
- Determine Application sensitivity point
  - Both Send and Receive

- Sensitivity points:
  - 8K-256KB messages
    - Relate to data communications
  - 0-128B messages
    - Relate to data synchronizations

ECLIPSE MPI Profiling  
MPI\_Recv



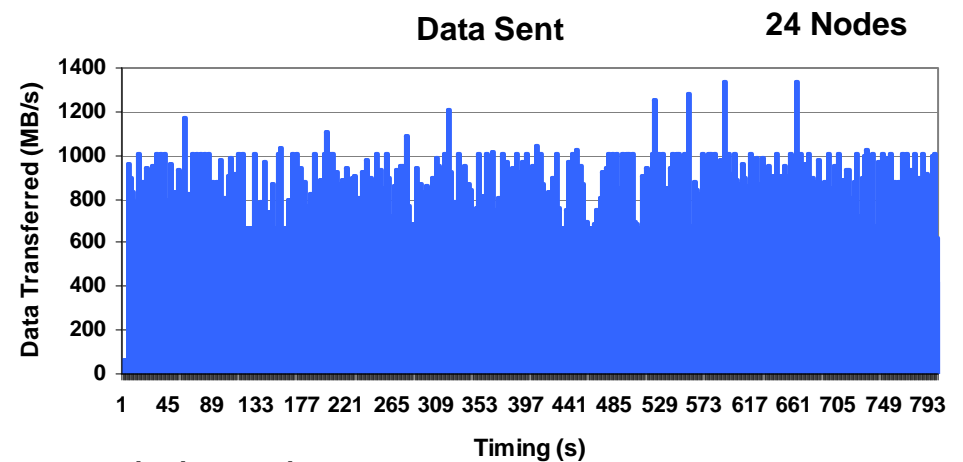
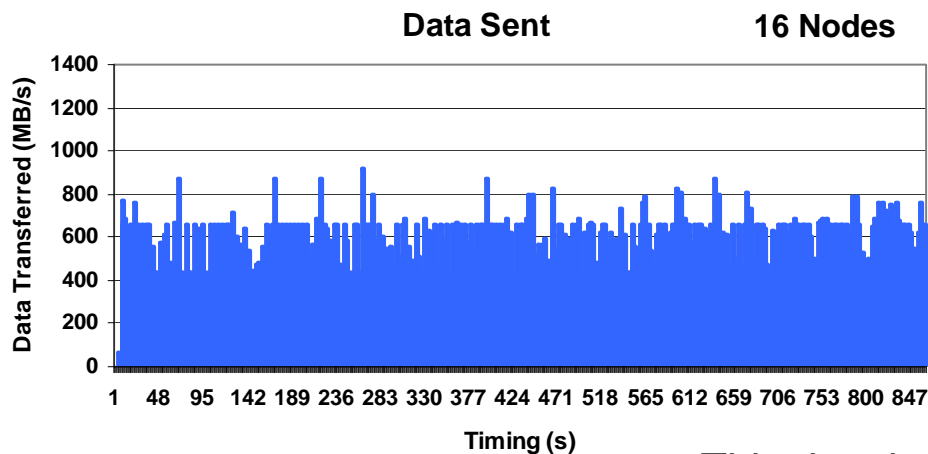
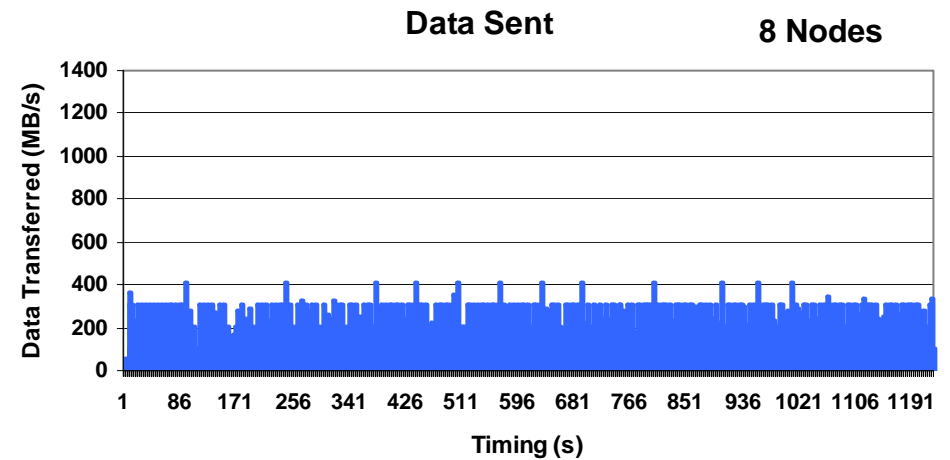
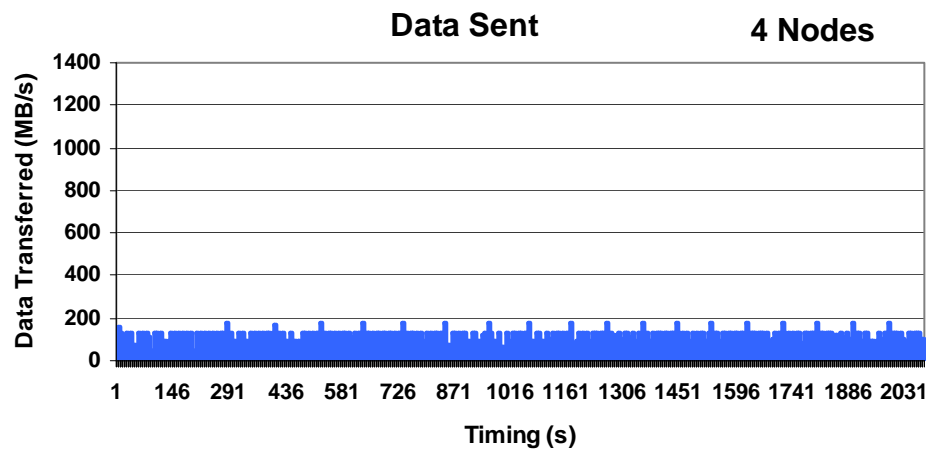
## Eclipse MPI Profiling



- **Majority of MPI messages are large size**
  - Percentage slightly increase with cluster size
- **Demonstrating the need for highest throughput**

# Interconnect Usage by ECLIPSE

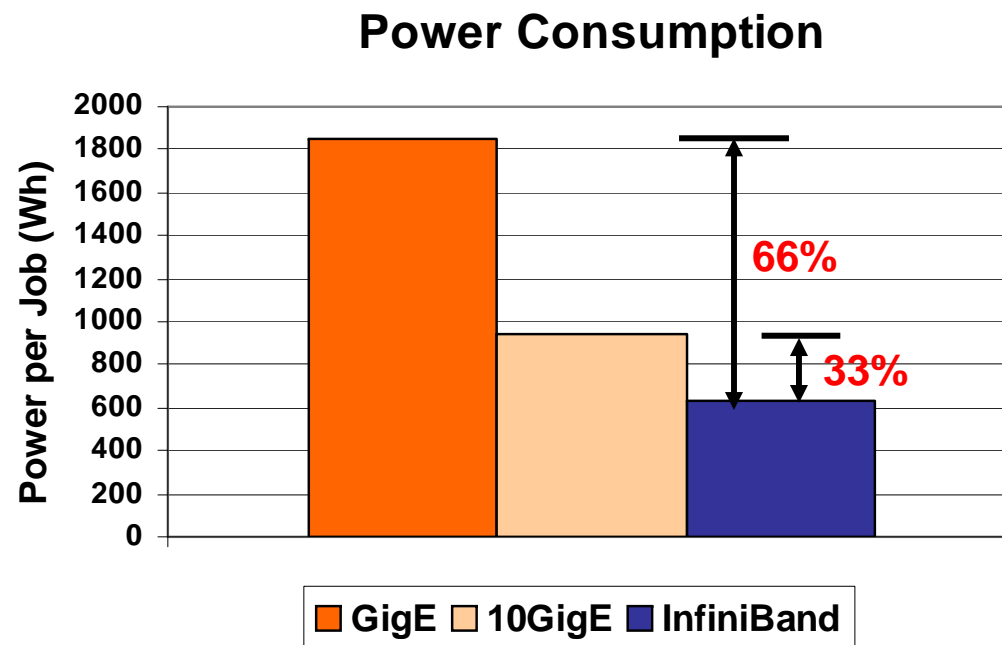
- Total server throughput increases rapidly with cluster size



*This data is per node based*

# Power Consumption/Productivity Comparison

- By enabling higher productivity on a given system, power/job decreases
- By using InfiniBand power/job consumption decreases by
  - Up to 66% vs GigE and 33% vs 10GigE
  - For productivity case – 4 jobs per node
- With a single job approach, InfiniBand reduces power/job consumption by more than 82% compared to 10GigE



*4 Jobs on each node*

- **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

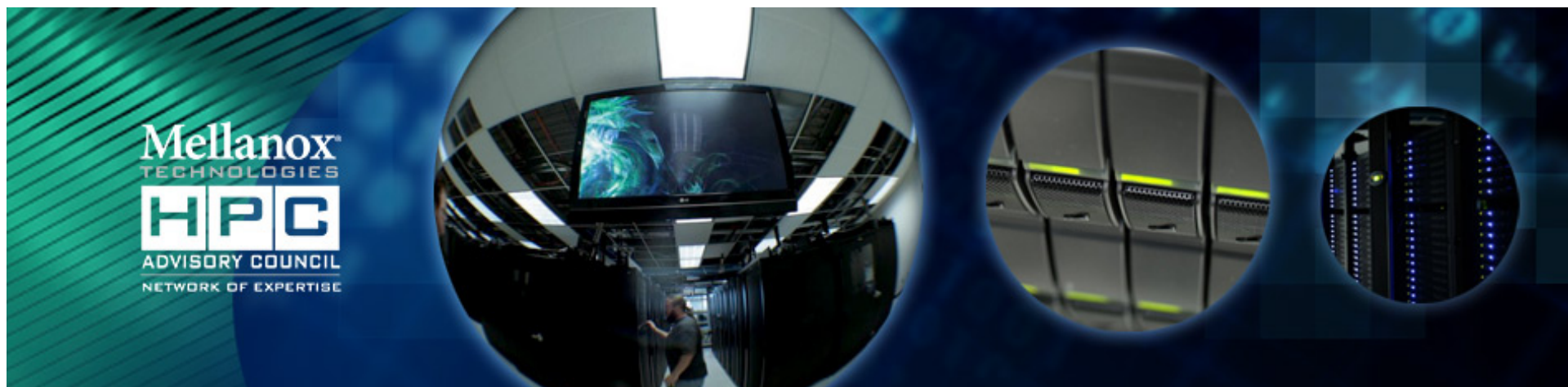
- **AMD Opteron “Shanghai” versus Opteron “Barcelona”**
  - Overall 20% performance increase
    - Both GFLOP/s and GB/s
  - While sustaining similar power consumption.
- **Memory throughput intensive applications (such as ECLIPSE)**
  - Performance is driven by north-bridge and DDR2 memory frequencies
  - Consume less energy than cache friendly compute intensive applications
- **Power management schemes**
  - Delivers the highest application performance per Watt ratio
  - Critical for low power consumption of the platform while in idle
  - Core frequency drop to as low as 800 MHz

- **Eclipse is widely used to perform reservoir simulation**
  - Developed by Schlumberger
- **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



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