

# NEMO

## Performance Benchmark and Profiling

May 2011



- **The following research was performed under the HPC Advisory Council HPC|works working group activities**
  - Participating vendors: HP, Intel, Mellanox
  - Compute resource - HPC Advisory Council Cluster Center
  
- **For more info please refer to**
  - <http://www.hp.com/go/hpc>
  - [www.intel.com](http://www.intel.com)
  - [www.mellanox.com](http://www.mellanox.com)
  - <http://www.nemo-ocean.eu>

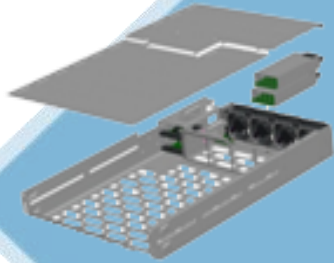
- **NEMO is a state-of-the-art modeling framework for**
  - Oceanographic research
  - Operational oceanography seasonal forecast
  - Climate studies
- **NEMO includes 4 major components**
  - The blue ocean (ocean dynamics, NEMO-OPA)
  - The white ocean (sea-ice, NEMO-LIM)
  - The green ocean (biogeochemistry, NEMO-TOP)
  - The adaptative mesh refinement software (AGRIF)
- **NEMO is used by a large community:** 240 projects in 27 countries
  - Under the CeCILL license (public license) controlled by a European Consortium between [CNRS](#), [Mercator-Ocean](#), [UKMO](#) and [NERC](#)
- **NEMO is part of DEISA benchmark suite** (<http://www.deisa.eu>)

- **The presented research was done to provide best practices**
  - File-system performance comparison
  - MPI libraries comparisons
  - Interconnect performance benchmarking
  - NEMO Application profiling
  - Understanding NEMO communication patterns
  
- **The presented results will demonstrate**
  - Balanced compute environment determines application performance

- **HP ProLiant SL2x170z G6 16-node cluster**
  - Six-Core Intel X5670 @ 2.93 GHz CPUs
  - Memory: 24GB per node
  - OS: CentOS5U5, OFED 1.5.3 InfiniBand SW stack
- **Mellanox ConnectX-2 InfiniBand QDR adapters and switches**
- **Fulcrum based 10Gb/s Ethernet switch**
- **MPI**
  - Intel MPI 4, Open MPI 1.7, Platform MPI 8.0.1
- **Compilers: Intel Compilers 11.1.064**
- **Application: NEMO 3.2**
- **Libraries: Intel MKL 2011.3.174, netCDF 2.122**
- **Benchmark workload**
  - OPA (the ocean engine), confcoef=25

# About HP ProLiant SL6000 Scalable System

- **Solution-optimized for extreme scale out**



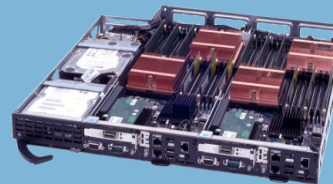
ProLiant z6000 chassis  
Shared infrastructure  
– fans, chassis, power



ProLiant SL160z G6    ProLiant SL165z G7  
Large memory  
-memory-cache apps



ProLiant SL170z G6  
Large storage  
-Web search and database apps




ProLiant SL2x170z G6  
Highly dense  
- HPC compute and  
web front-end apps

Save on cost and  
energy -- per node,  
rack and data  
center

Mix and match  
configurations

Deploy with  
confidence



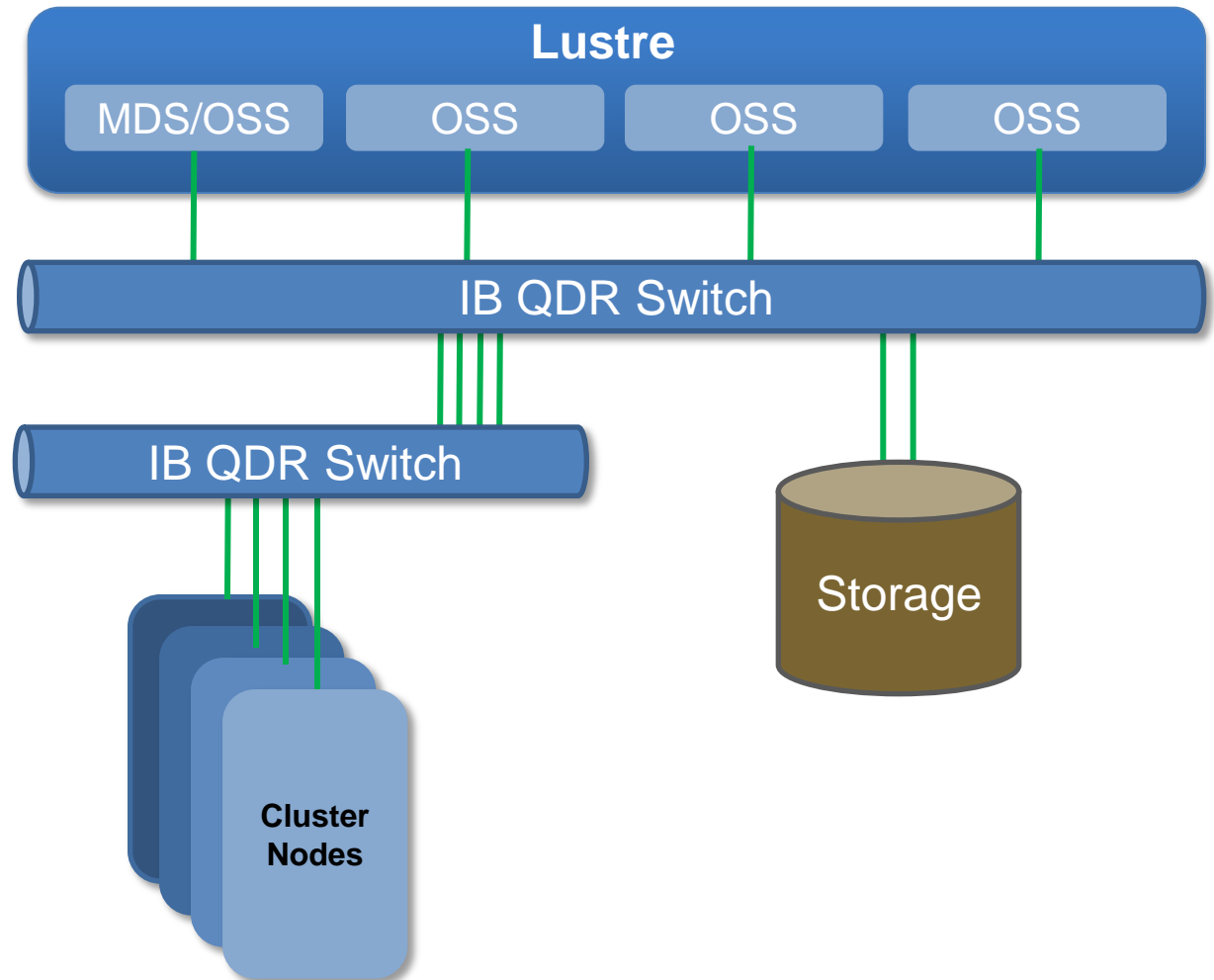
#1  
Power  
Efficiency\*

\* SPECpower\_ssj2008  
[www.spec.org](http://www.spec.org)  
17 June 2010, 13:28

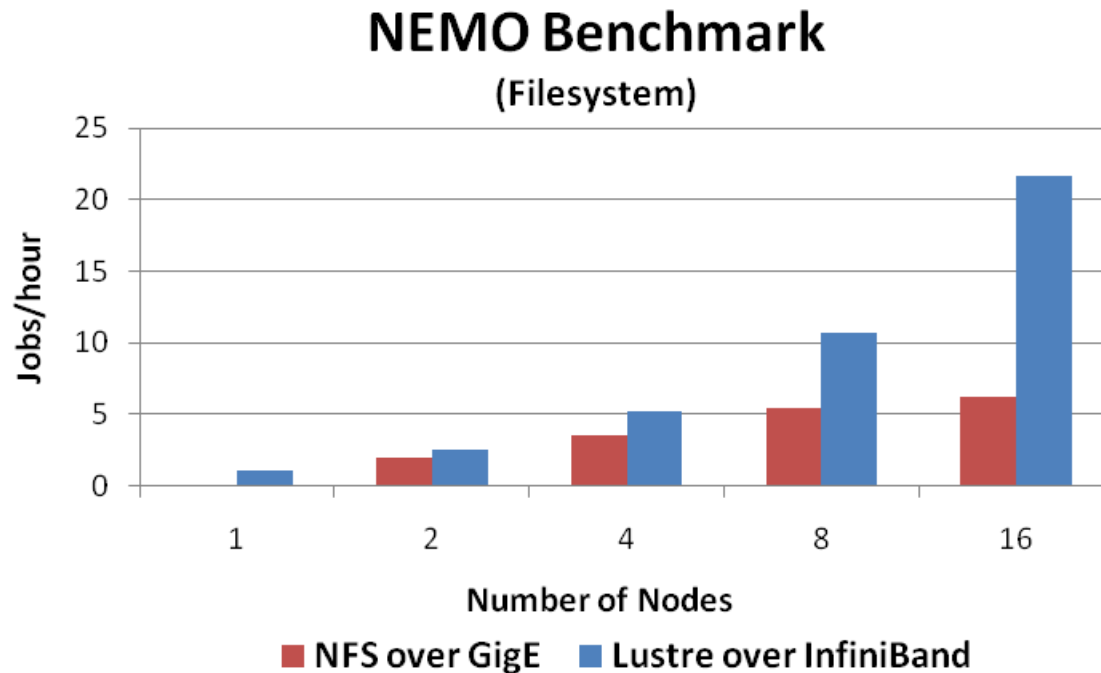
# Lustre File System Configuration

- **Lustre Configuration**

- 1 MDS
- 4 OSS (Each has 2 OST)
- InfiniBand based Backend storage
- All components are connected through InfiniBand QDR interconnect



- **File I/O performance is important to NEMO performance**
  - InfiniBand powered Lustre file system enables application scalability
  - NFS over GigE doesn't meet application file I/O requirement

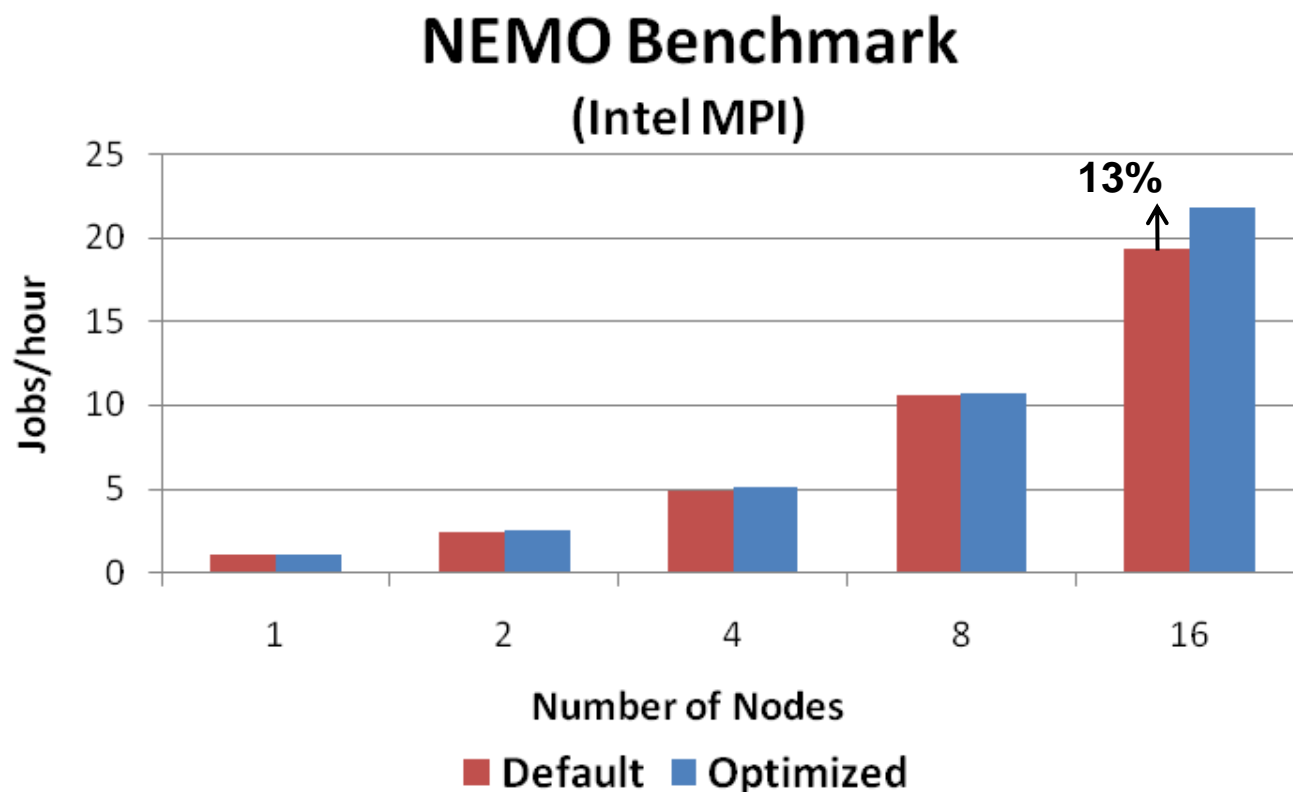


*Higher is better*

*Open MPI over InfiniBand QDR  
12-cores per node*

- **Intel MPI with tuning runs 13% faster than default mode at 16 nodes**

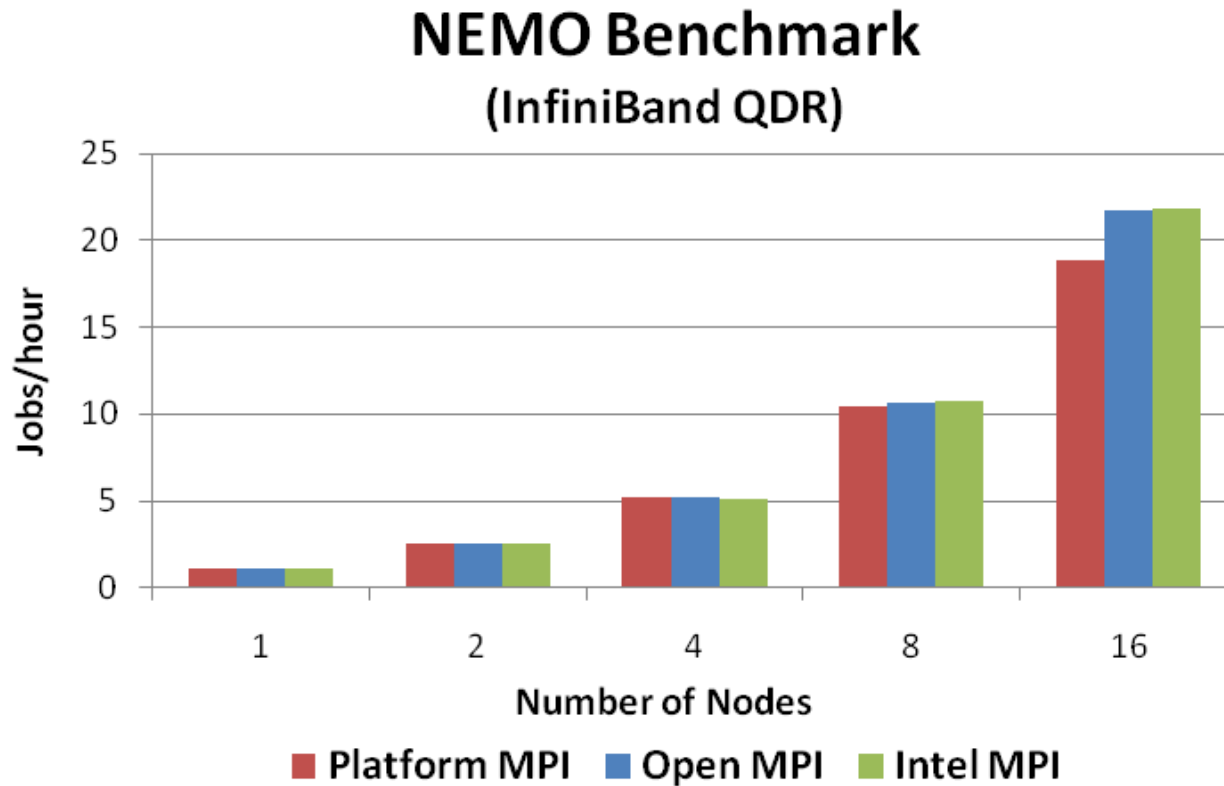
- `-genv I_MPI_RDMA_TRANSLATION_CACHE 1 -genv I_MPI_RDMA_RNDV_BUF_ALIGN 65536 -genv I_MPI_DAPL_DIRECT_COPY_THRESHOLD 65536`



*Higher is better*

*12-cores per node*

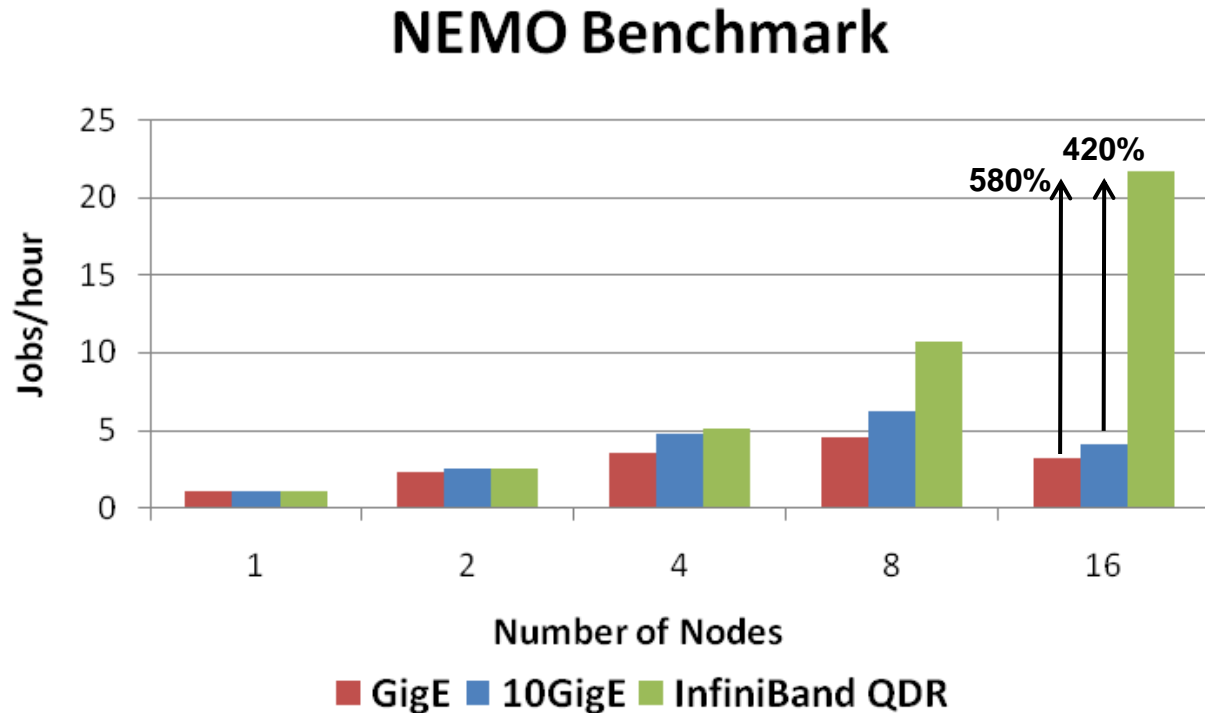
- Intel MPI and Open MPI are faster at 16 nodes



*Higher is better*

*12-cores per node*

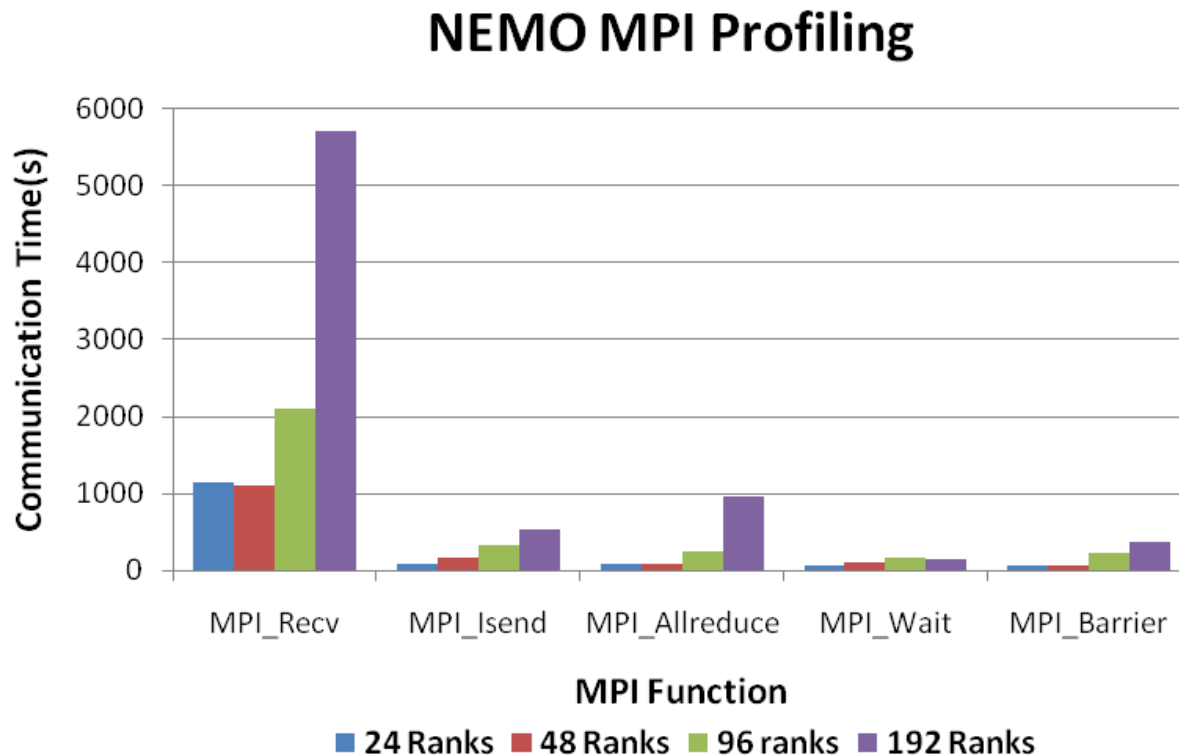
- **InfiniBand enables highest performance and linear scalability for NEMO**
  - 420% faster than 10GigE and 580% faster than GigE at 16 nodes



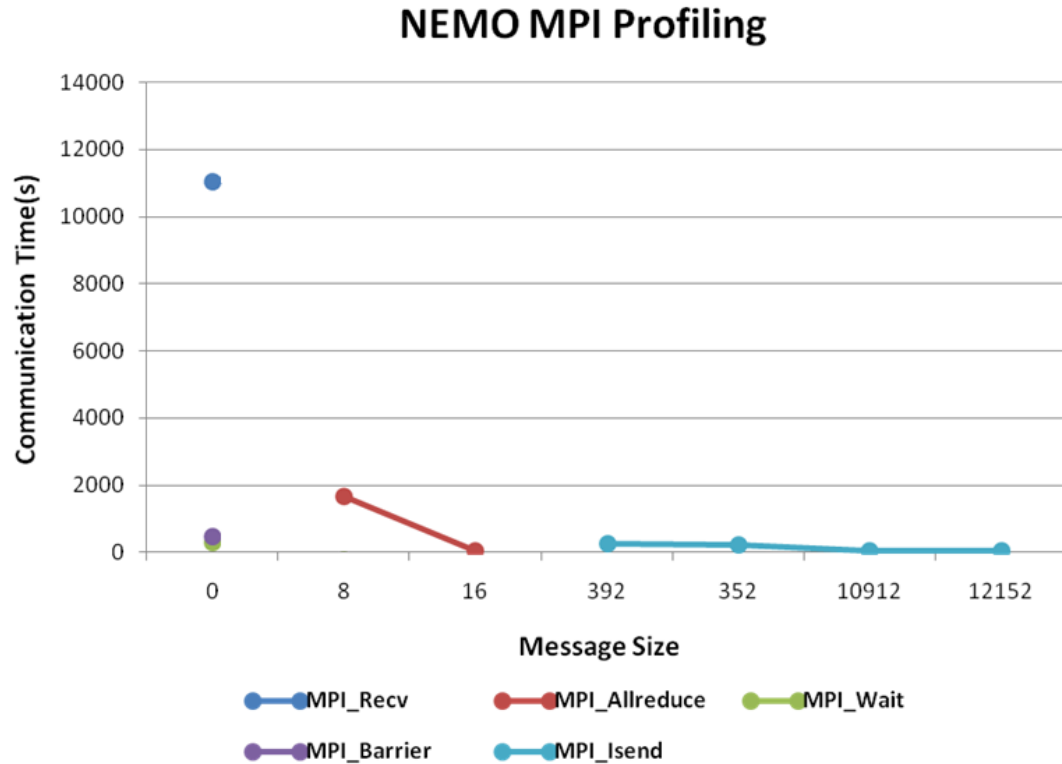
*Higher is better*

*12-cores per node*

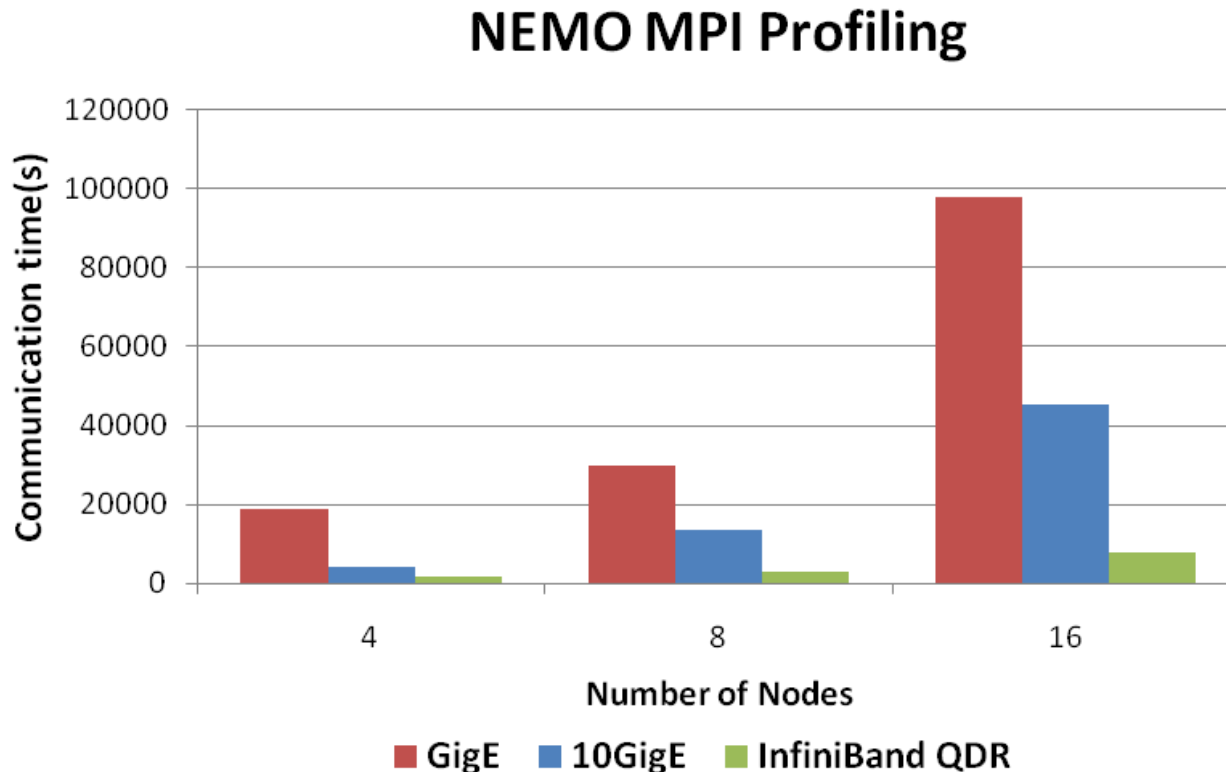
- **MPI point-to-point communication overhead is dominated**
  - Point-to-point: MPI\_Isend/recv
  - Collectives: MPI\_Allreduce overhead increases faster after 8 nodes



- Most messages are small messages: <12KB



- **InfiniBand QDR has least communication overhead**
  - 13% of total MPI time over GigE
  - 17% of total MPI time over 10GigE



- **NEMO performance benchmark demonstrates**
  - InfiniBand QDR delivers higher application performance and linear scalability
    - 420% higher performance than 10GigE and 580% higher than GigE
  - Intel MPI tuning can boost application performance by 13%
  - Application has intensive file I/O operations
    - Lustre over InfiniBand eliminates NFS bottleneck and enables application performance
- **NEMO MPI profiling**
  - Message send/recv creates big communication overhead
  - Most are small message used by NEMO
  - Collectives overhead increases as cluster size scales up

# 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