HPC ADVISORY COUNCIL

Virtual Performance Solution (VPS) 2013.01 Performance Benchmark and Profiling

May 2014

esi get it right®

DELL

intel

Mellanox TECHNOLOGIES

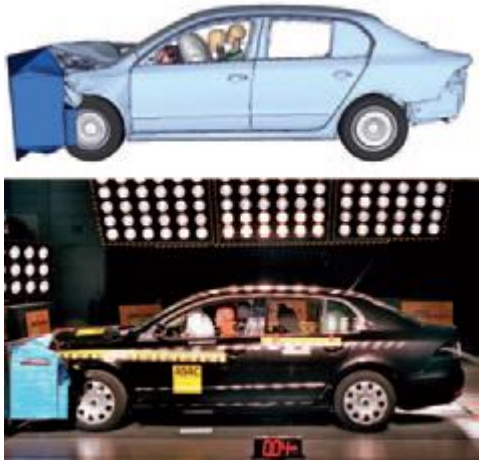
Note

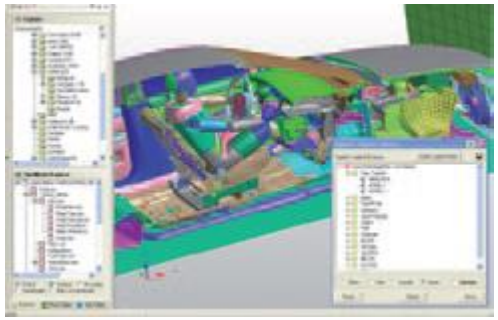
- **The following research was performed under the HPC Advisory Council activities**
  - Participating vendors: ESI Group, Intel, Dell, Mellanox
  - Compute resource - HPC Advisory Council Cluster Center

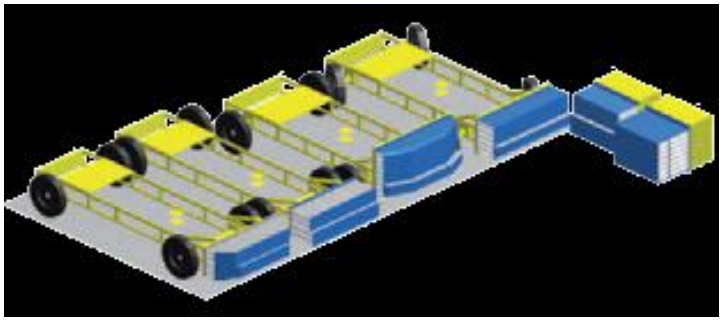
- **The following was done to provide best practices**
  - VPS performance overview
  - Understanding VPS communication patterns
  - Ways to increase VPS productivity
  - MPI libraries comparisons

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

- **Virtual Performance Solution (VPS)**
  - Originated from **PAM-CRASH**
  - Software package from ESI Group
  - Used for crash simulation
  - Design of occupant safety systems
  - Primarily used in the automotive industry
  - Simulate the performance of a proposed vehicle design
  - Evaluate the potential for injury to occupants in multiple crash scenarios







Objectives

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

- **The presented results will demonstrate**
  - The scalability of the compute environment
  - The scalability of the compute environment/application
  - Considerations for higher productivity and efficiency

Test Cluster Configuration

- **Dell™ PowerEdge™ R720/R720xd 32-node (640-core) "Jupiter" cluster**
  - Dual-Socket Hexa-Core Intel E5-2680 V2 @ 2.80 GHz CPUs
  - Memory: 64GB memory, DDR3 1600 MHz, Dual Rank
  - OS: RHEL 6.2, OFED 2.1-1.0.6 InfiniBand SW stack
  - Hard Drives: R720xd: 24x 250GB 7.2 RPM SATA 2.5" on RAID 0. R720: 16x250GB on RAID 0
- **Intel Cluster Ready certified cluster**
- **Mellanox Connect-IB FDR InfiniBand and ConnectX-3 Ethernet adapters**
- **Mellanox SwitchX 6036 VPI InfiniBand and Ethernet switches**
- **MPI executables provided: Platform MPI 8.3, Open MPI 1.4**
- **MPI used: Platform MPI 9.1, Open MPI 1.8 based on Mellanox HPC-X 1.0.0rc4**
- **Application: VPS 2013.01**
- **Benchmarks:**
  - Crash_NEON_FINE_CAR2CAR – Chrysler Neon CAR2CAR 56km/h, 120ms, Single Precision (unless otherwise stated)

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

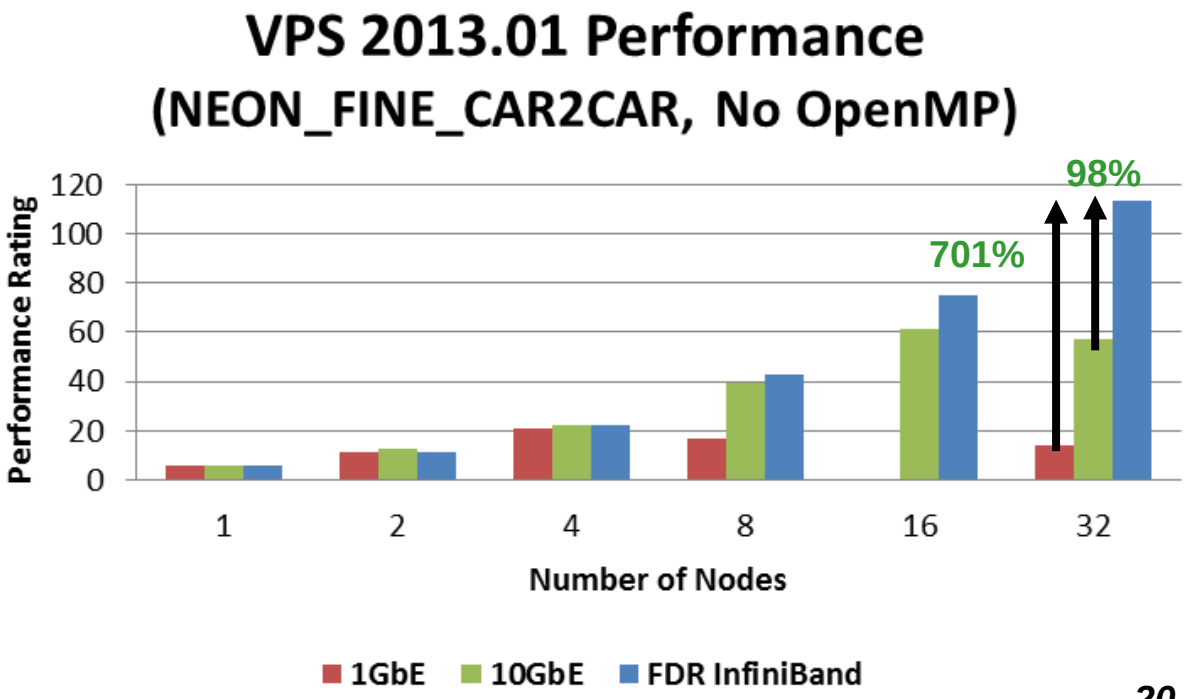
PowerEdge R720/R720xd

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 GigE 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

VPS Performance – Network (MPI)

- **FDR InfiniBand delivers the best network scalability performance**
  - Provides up to 701% higher performance than 1GbE at 32 nodes
  - Provides up to 98% higher performance than 10GbE at 32 nodes
  - FDR IB scales linearly while 10/40GbE has scalability limitation beyond 16 nodes
  - Result for 1GbE at 16 nodes was excluded due to error termination at runtime

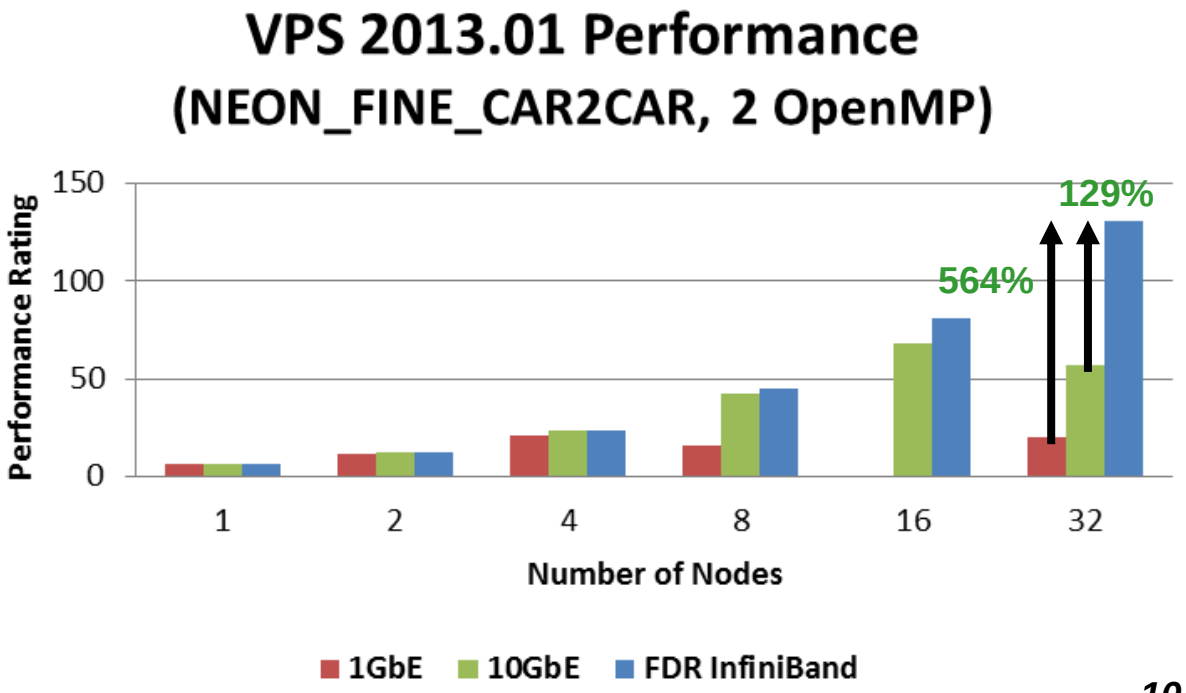


Higher is better

20 MPI proc/node

VPS Performance – Network (Hybrid)

- **Similar scalability seen with 2 OpenMP thread spawn per process**
  - FDR IB provides up to 564% higher performance vs 1GbE, and 129% vs 10GbE
  - FDR IB scales linearly while 10GbE has scalability limitation beyond 16 nodes
  - Scalability of 1GbE drops after 4 nodes

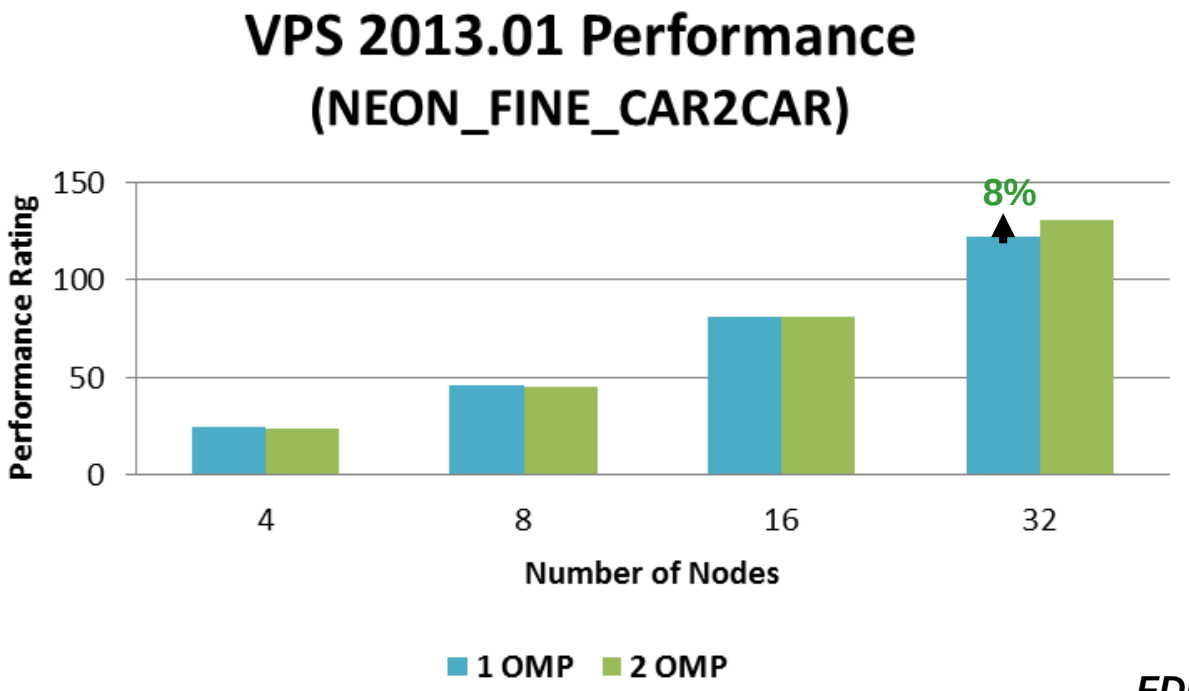


Higher is better

10 MPI / 2 OMP

VPS Performance – MPI-OpenMP Hybrid

- **Hybrid mode allows higher performance at scale**
  - 2 OpenMP threads per process provides 8% higher at 32 nodes
  - Slightly better performance if OpenMP is not used on smaller node counts
  - Hybrid mode expect to provide higher performance at larger scale

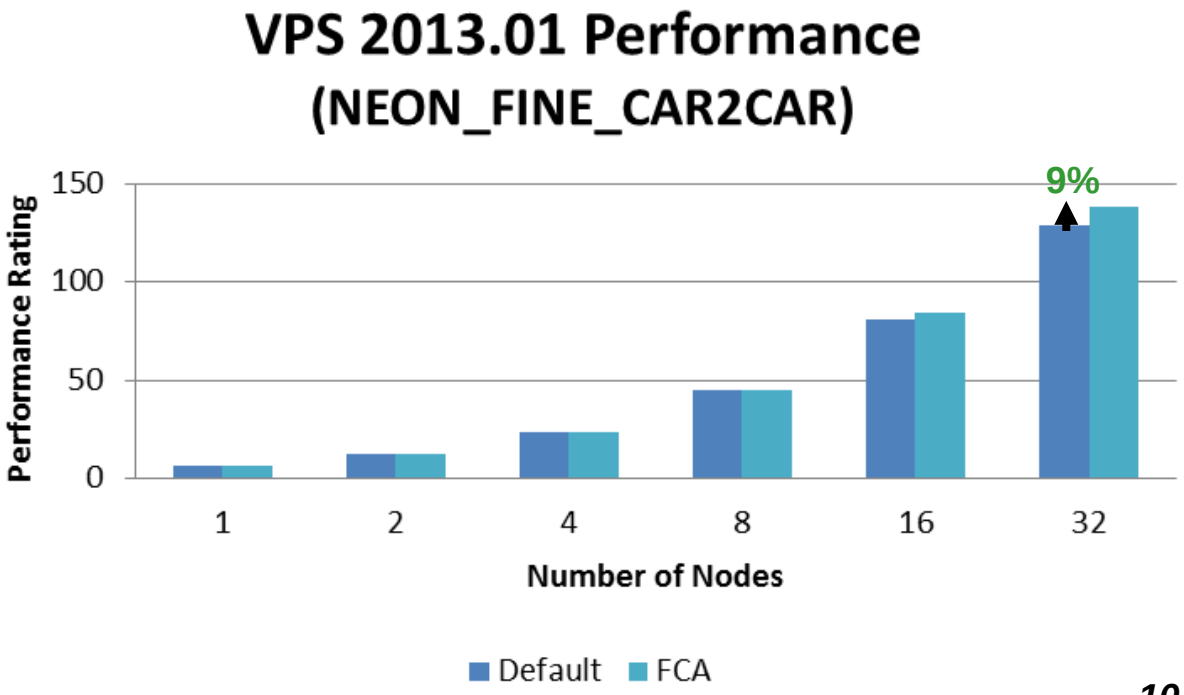


Higher is better

FDR InfiniBand

VPS Performance – FCA

- **Enabling FCA provides additional speedup for Open MPI**
  - MPI collective accelerations provide ~9% speedup at 32 nodes
- **Runtime flags used:**
  - Enabling FCA: -mca coll_fca_enable 1 -mca coll_fca_np 0
  - Other tuned flags used for FCA, to mitigate node imbalances effect in MPI_Allreduce:
  - –map-by slot -x fca_mpi_slow_sleep=0 -x fca_mpi_slow_num_polls=100000000"

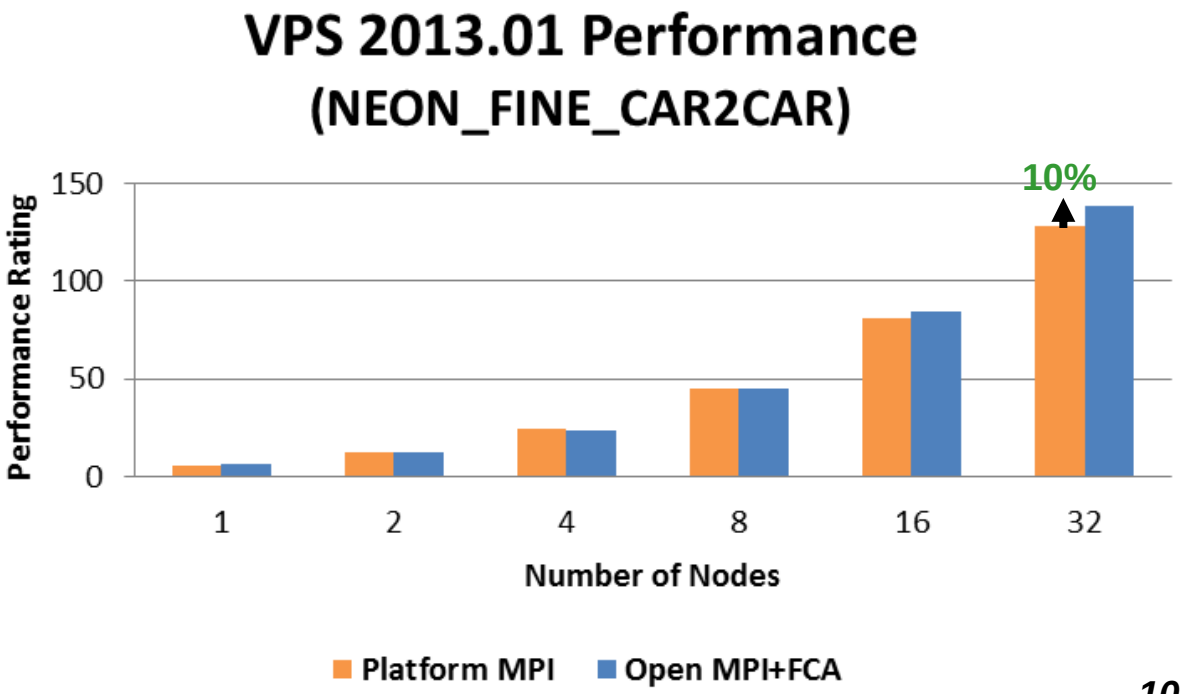


Higher is better

10 MPI/2 OMP/node

VPS Performance – MPI Implementations

- **Tuned Open MPI delivers higher performance for VPS**
  - Open MPI with FCA runs 10% faster than Platform MPI
  - Default MPI implementation used in VPS is Platform MPI
  - VPS supports OMPI 1.4 but need more recent version to work for network
  - Modifications (on pamworld) and run script to make Open MPI 1.8 to work

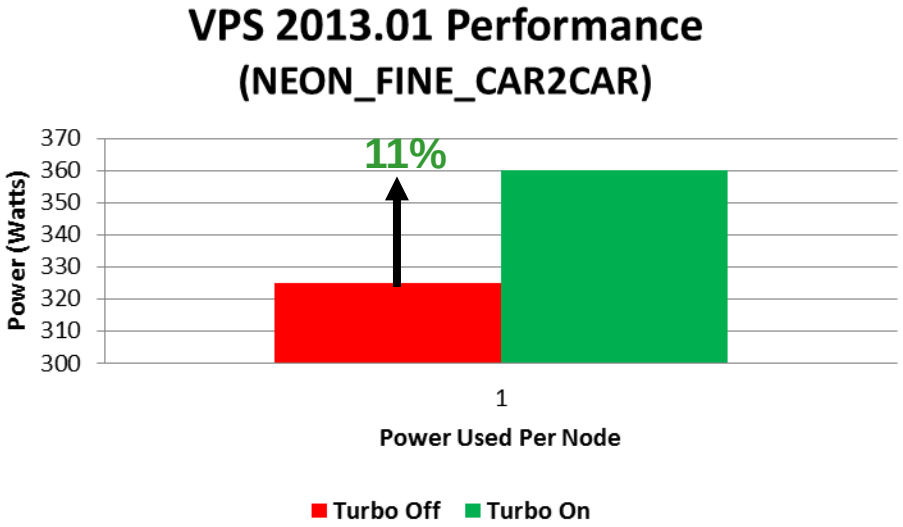


Higher is better

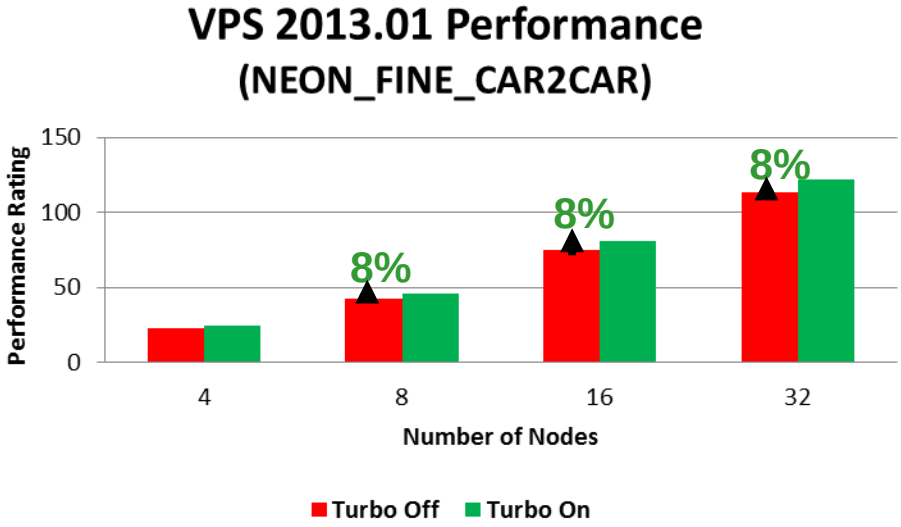
10 MPI/2 OMP/node

VPS Performance – Turbo Mode

- **Enabling Turbo mode results in higher application performance**
  - Up to 8% of the improvement seen by enabling Turbo mode
  - At a cost of ~11% of higher power utilization per node
  - Boosting base frequency; consequently resulted in higher power consumption
  - Power measurement is gathered from the iDRAC management interface on R720
- **Using kernel tools called "msr-tools" to adjust Turbo Mode dynamically**
  - Allows dynamically turn off/on Turbo mode in the OS level



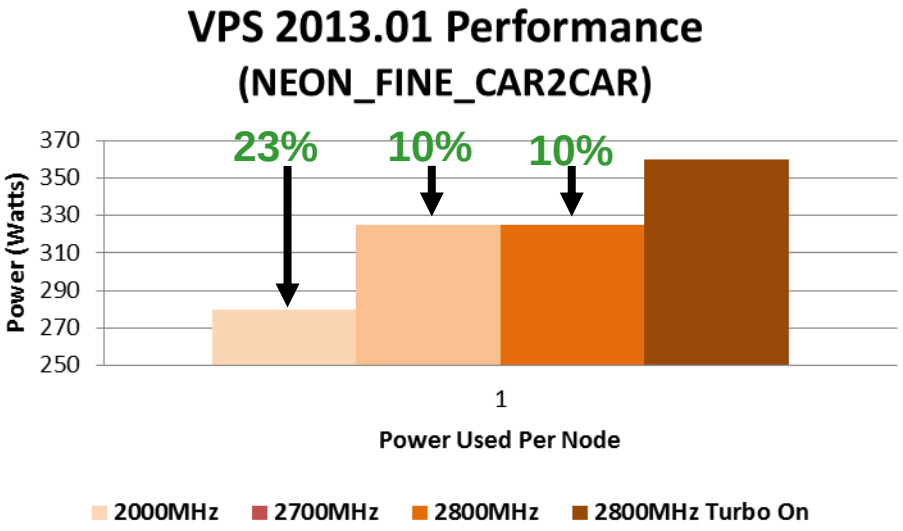
Lower is better



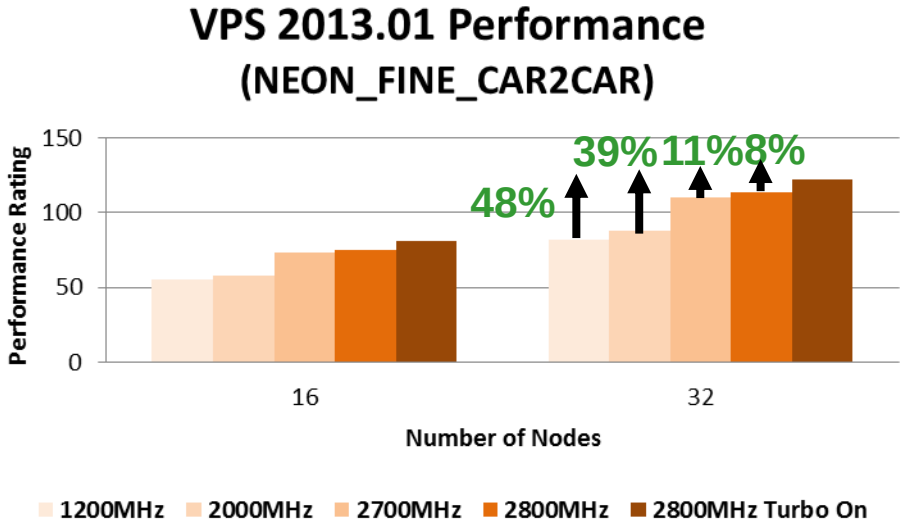
Higher is better

VPS Performance – CPU Frequencies

- **Running at higher CPU clock improves VPS performance**
  - For example, Running CPU at 2000MHz on all nodes saves 23% of system power
  - While performance is improved by 49% when using 2800MHz (Turbo) vs 2000MHz
- **Better Power/Performance efficiency is observed**
  - When clock speed around 2700MHz or 2800MHz with Turbo off



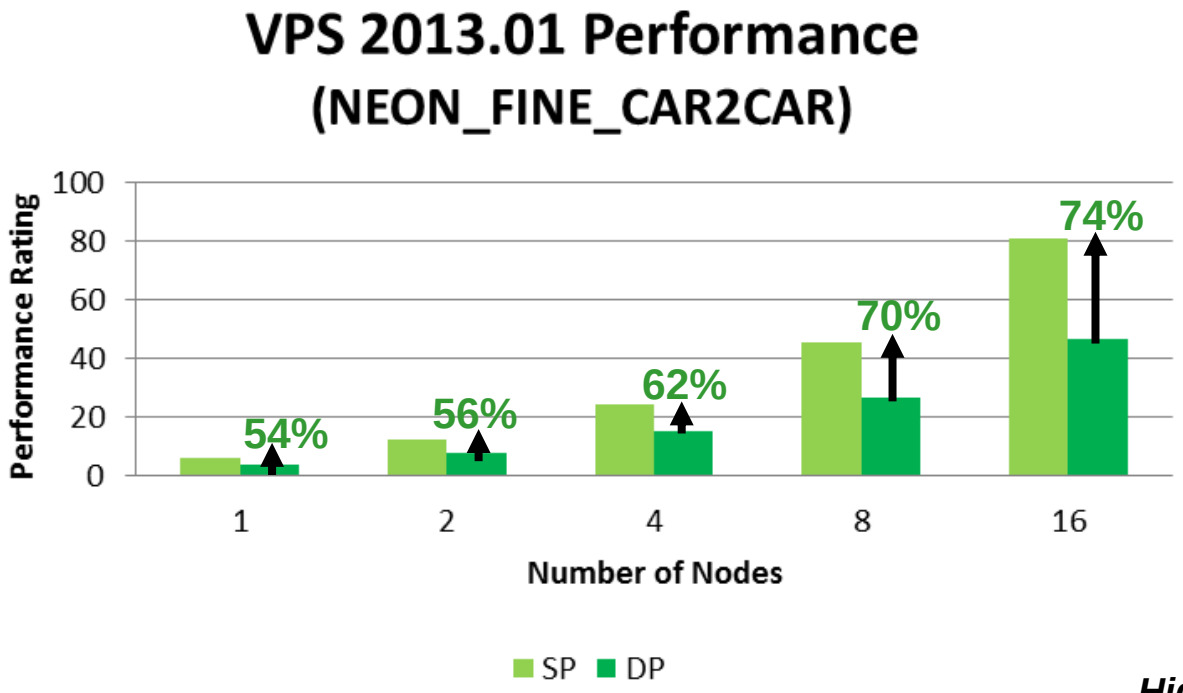
Open MPI



20 MPI proc/node

VPS Performance – Floating Point Precisions

- **Running Double Precision takes long than running at Single Precision**
  - DP takes more time than SP, by 54% on a single node
  - Some models require to run in DP to reach convergence, or crash when using SP
- **The difference in ratio between DP and SP increases as it scales**
  - Since DP provides higher precision in calculation, thus requires more data transferred
  - With data grows faster for DP as it scales, thus explains DP is slower than SP

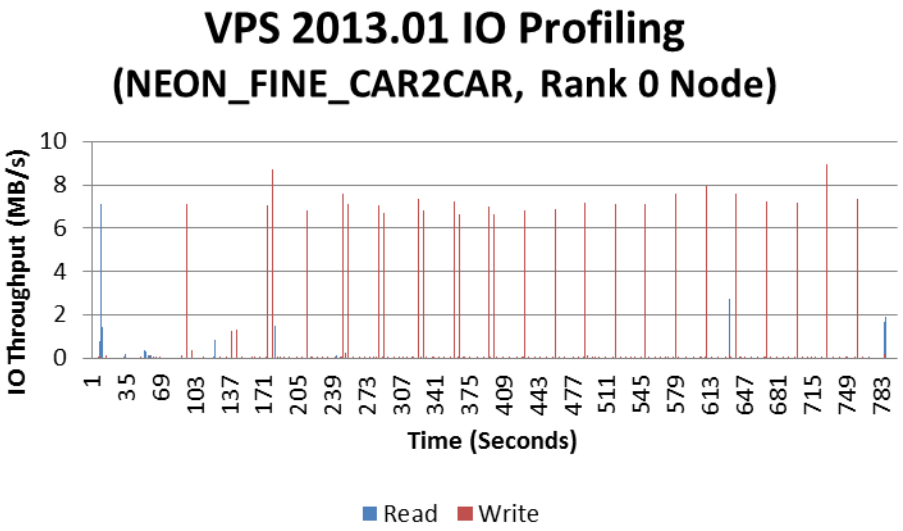


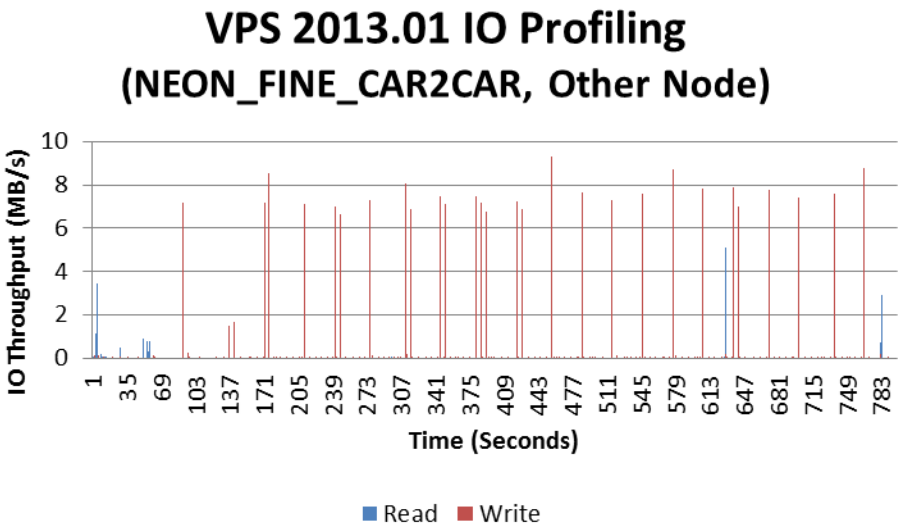
Lower is better

Higher is better

VPS Profiling – IO Profiling

- **Both rank 0 node and other nodes perform similar disk operations**
  - Disk read occurs mostly at the beginning of a run
  - Recurring disk writes takes place throughout the job run
  - Could potentially benefit by using parallel file system



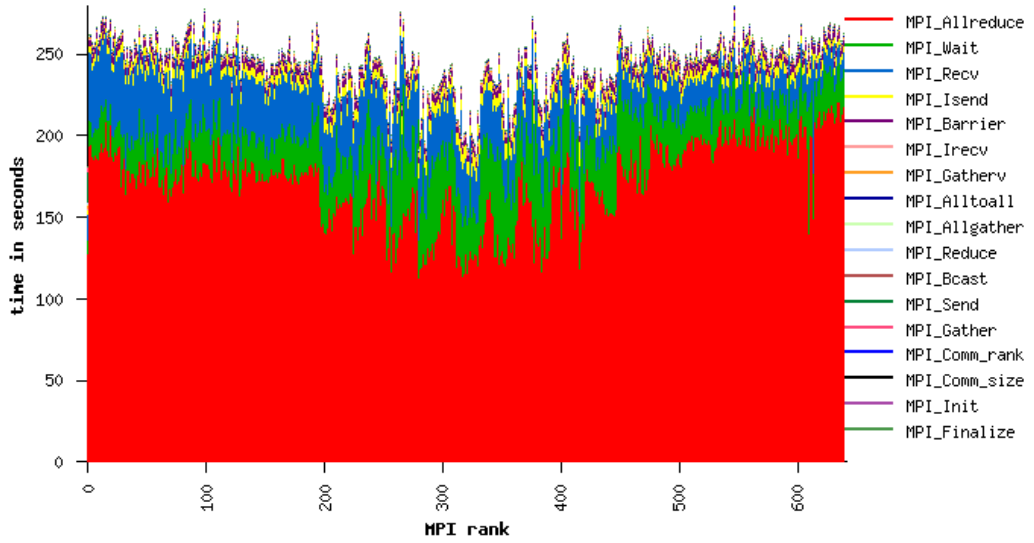


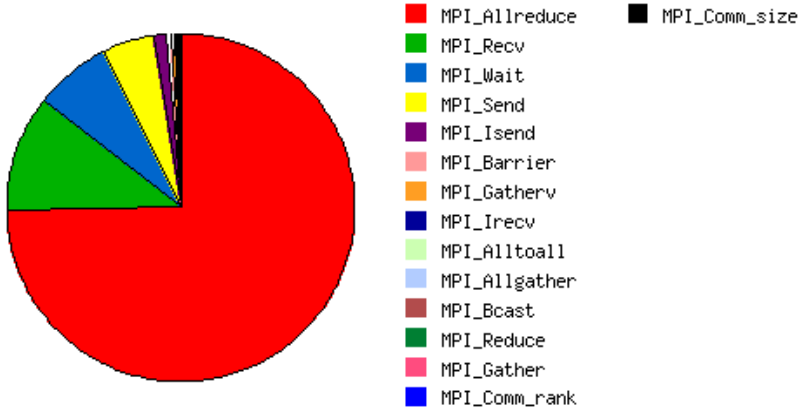
FDR InfiniBand

VPS Profiling – MPI Communication Time

- **MPI communication time consumption at 32 nodes**
  - MPI Time: MPI_Allreduce(71%), MPI_Wait(13%), MPI_Recv(12%), MPI_Isend(3%)
  - Wall Time: MPI_Allreduce(24%), MPI_Wait(4%), MPI_Recv(4%), MPI_Isend(1%)
  - FDR InfiniBand is used

32 Nodes/640 MPI



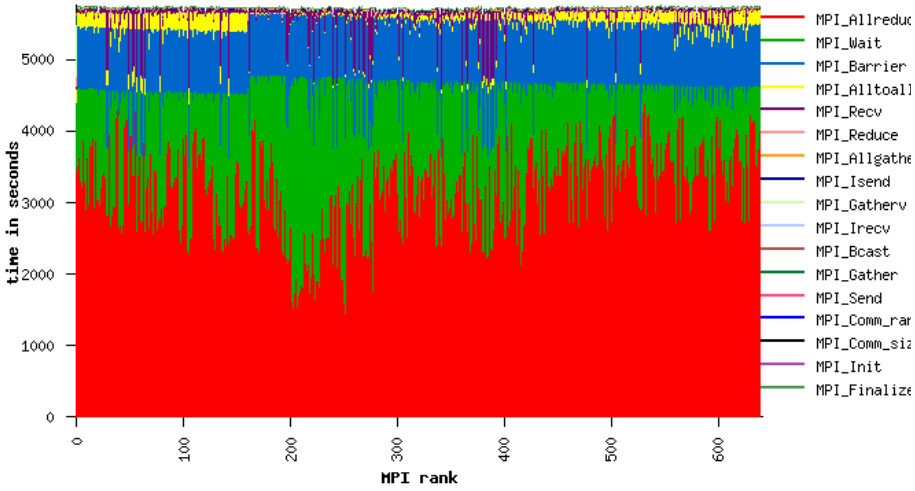


FDR InfiniBand

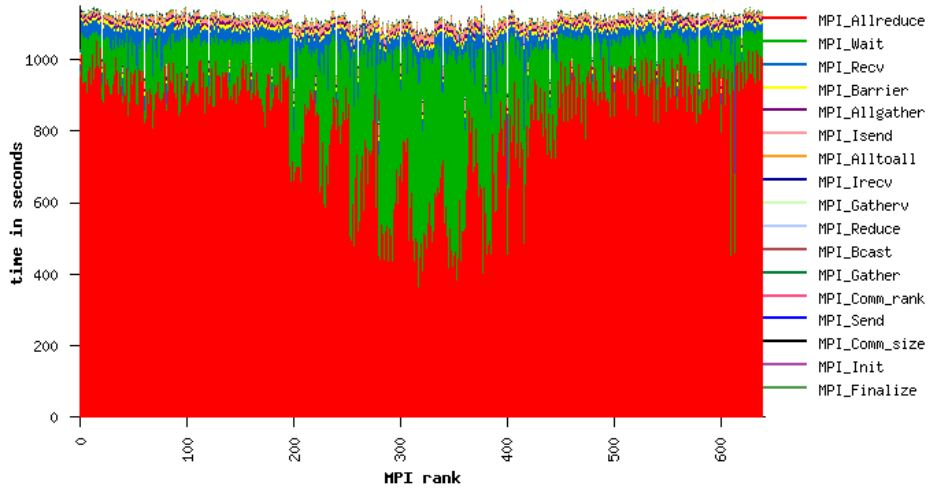
VPS Profiling – MPI Communication Time

- **Identified MPI overheads by profiling communication time**
  - VPS uses different MPI communication method extensively
    - collective, point-to-point and non-blocking operations
- **Ethernet spends more in collective operations**
  - 10GbE vs FDR IB: Spent longer time in MPI_Allreduce
  - 1GbE vs FDR IB: Spent way longer time in MPI_Allreduce, MPI_Barrier

1GbE



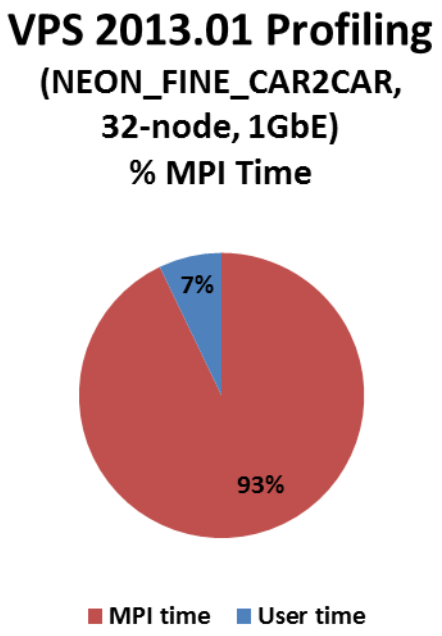
10GbE

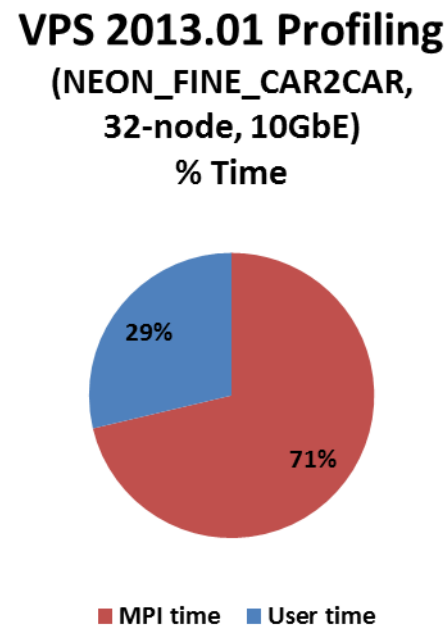


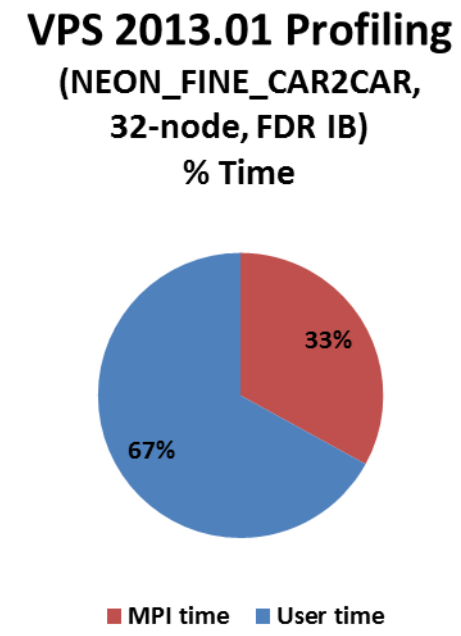
20 Processes/Node

VPS Profiling – User/MPI Time Ratio

- **VPS spent more time in computation than communication for FDR IB**
  - Other network spent more time in communication at 32 nodes
  - FDR IB consumes 33% of runtime in comm, vs 10GbE: 71% and 1GbE: 93%
  - FDR InfiniBand provides more time for computation, thus the most efficient network

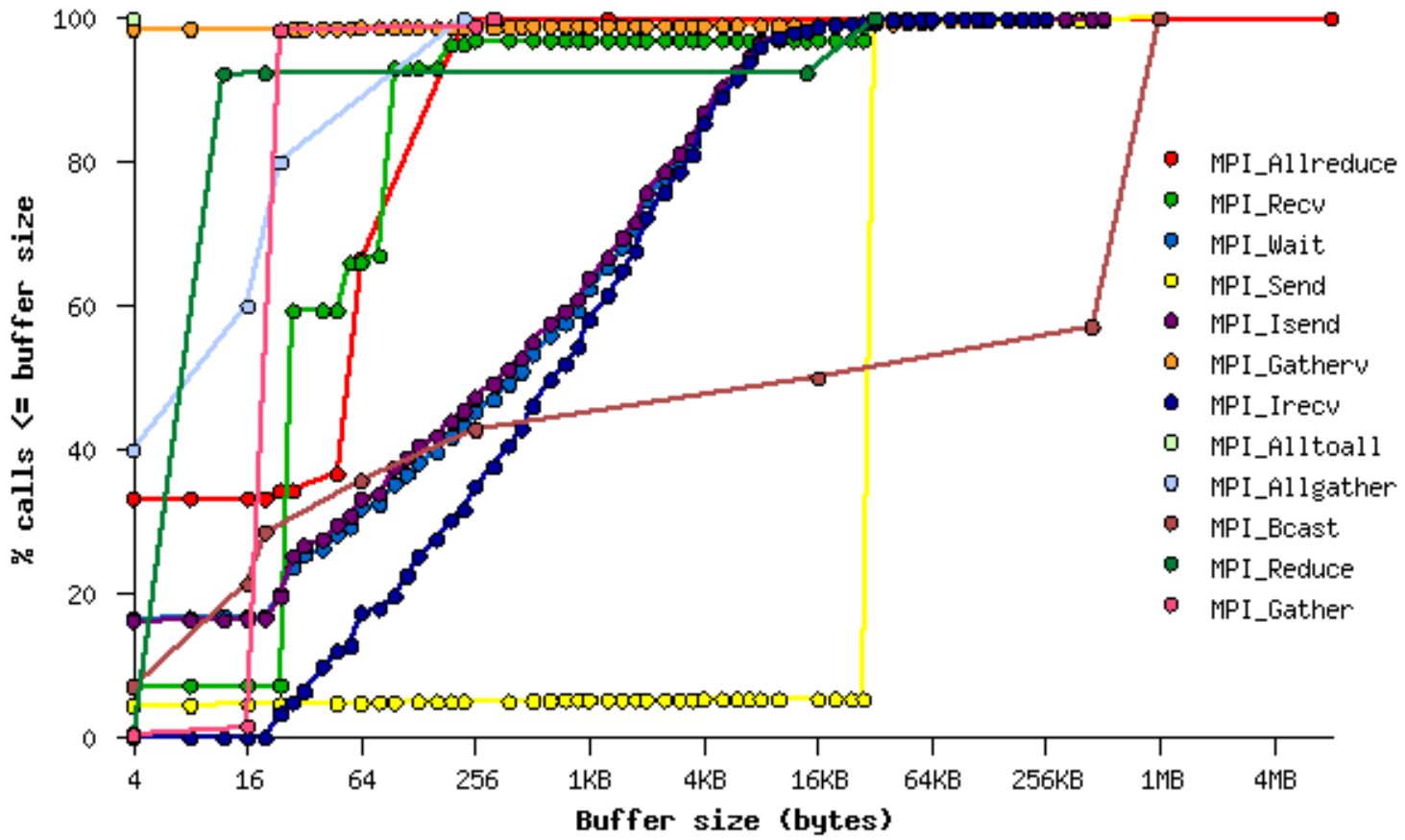






VPS Profiling – MPI Time Spent

- **The most time consuming MPI for VPS is MPI_Allreduce**
  - MPI_Allreduce consumes 60% of all MPI time
  - Majority of MPI_Allreduce takes place at 4B and 224B

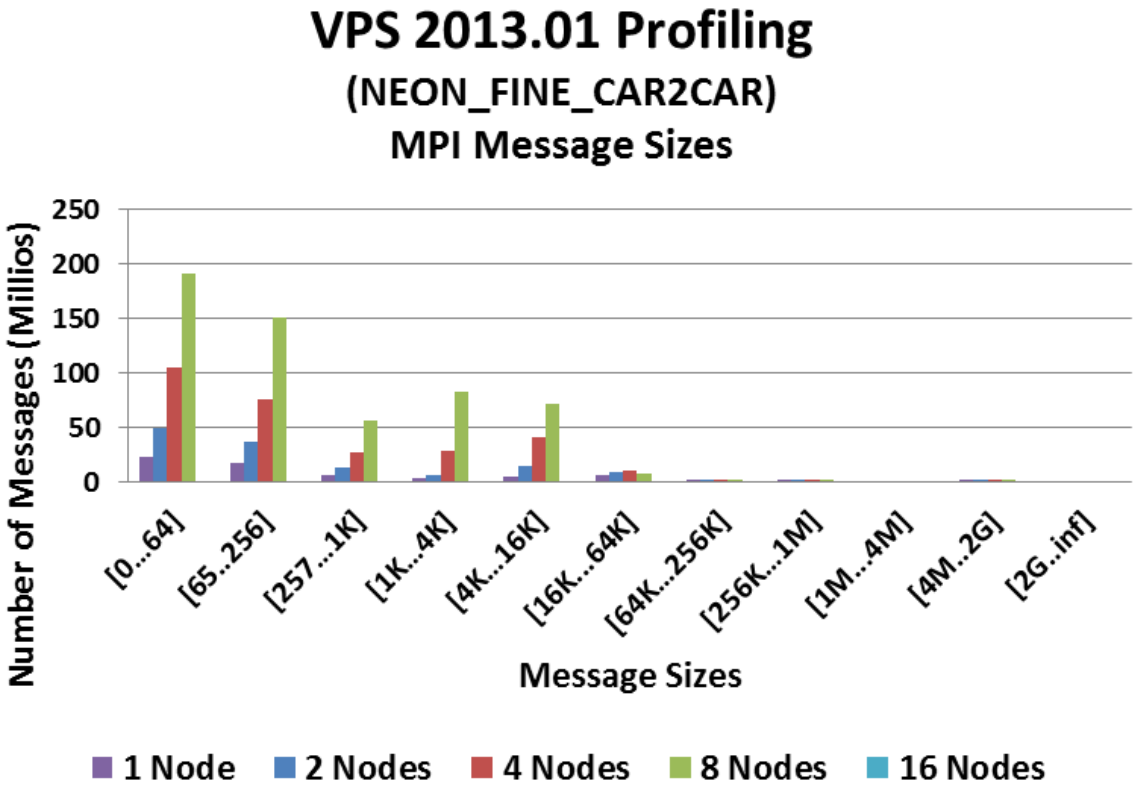


32 Nodes

FDR InfiniBand

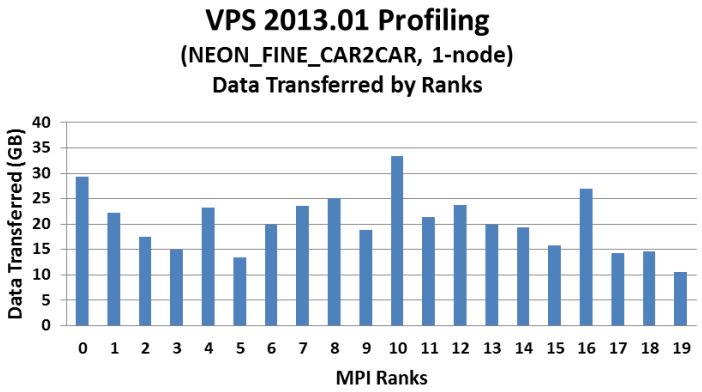
VPS Profiling – Message Sizes

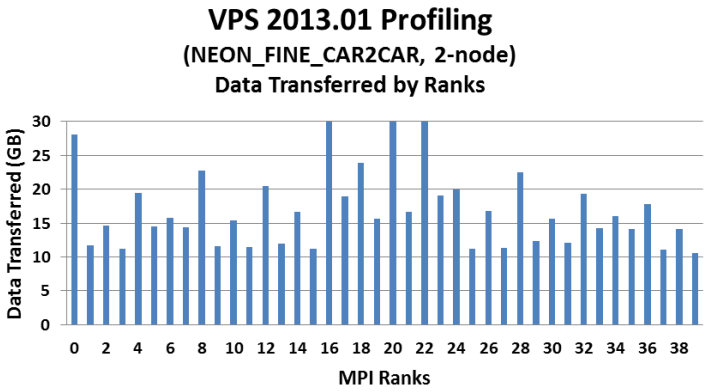
- **Majority of messages are small messages**
  - Messages are concentrated below 64KB
- **Number of messages increases with the number of nodes**

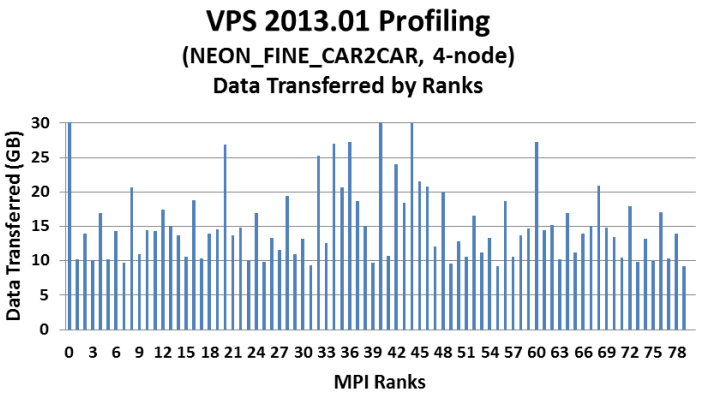


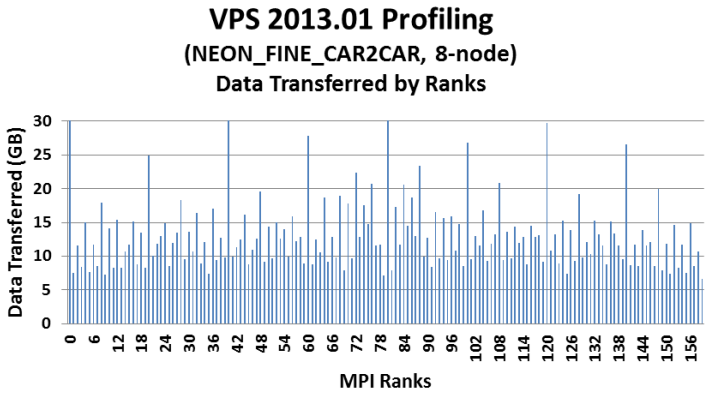
VPS Profiling – MPI Data Transfer

- **As the cluster grows, same amount of data transfers takes place**
  - From ~15-30GB per rank at 1 node vs 7-30GB at 8 nodes
  - Some node imbalances are seen through the amount of data transfers



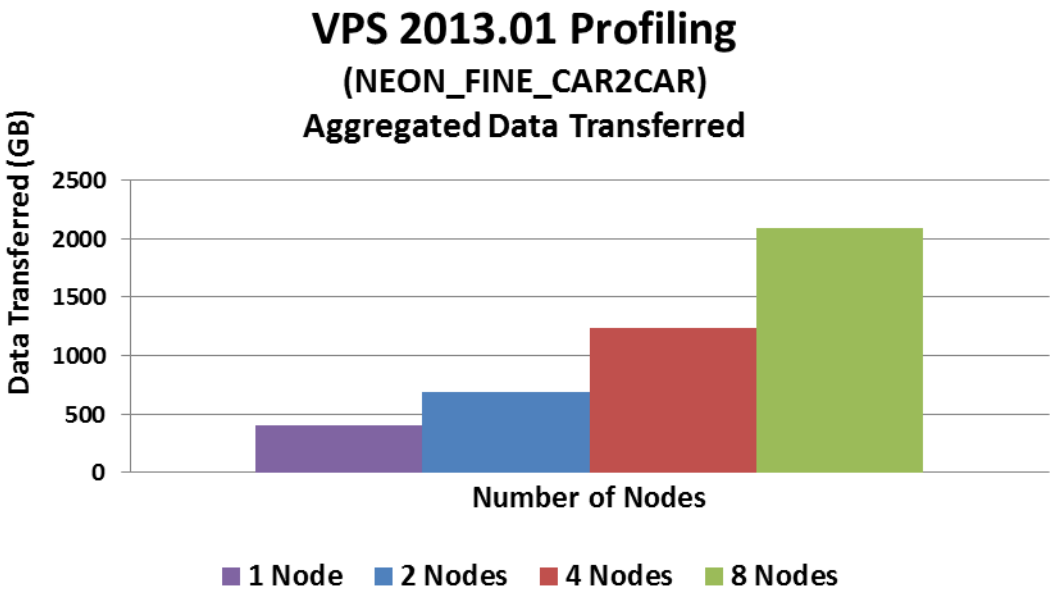






VPS Profiling – Aggregated Transfer

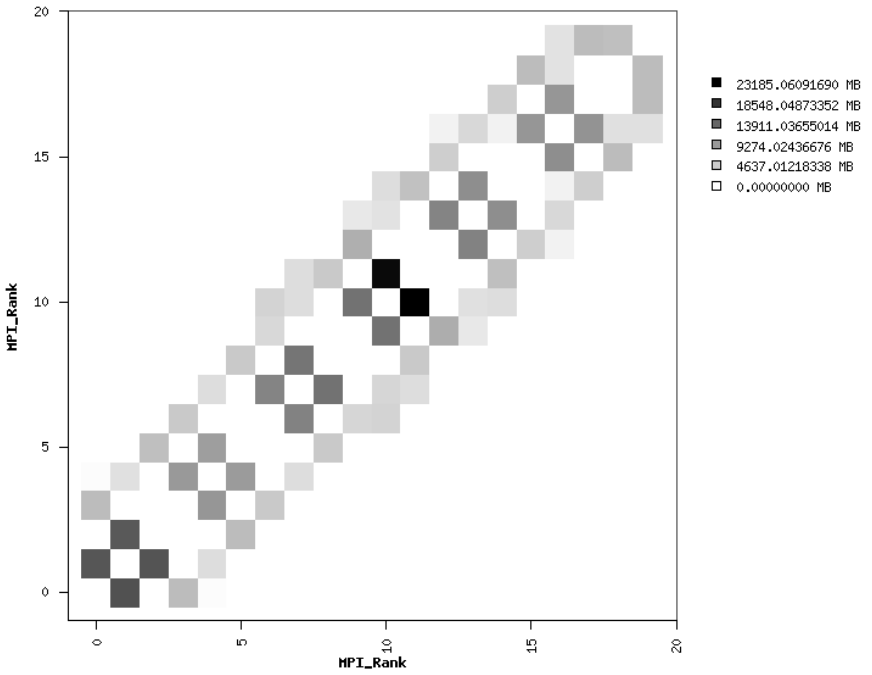
- **Aggregated data transfer refers to:**
  - Total amount of data being transferred in the network between all MPI ranks collectively
- **Very large data transfer takes place in VPS**
  - High network throughput is required for delivering the network bandwidth
  - 2TB of data transfer takes place between the MPI processes at 8 nodes



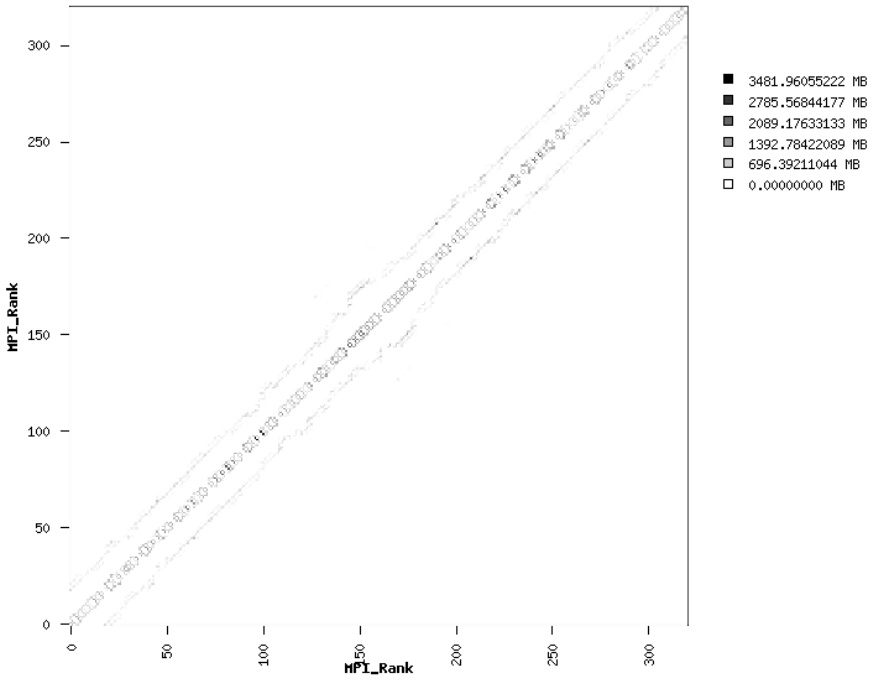
VPS Profiling – Point To Point Data Flow

- **The point to point data flow shows the communication pattern of VPS**
  - VPS mainly communicates mainly its neighbors and close ranks
  - The pattern stays the same as the cluster scales

4 Nodes – 40 Processes



32 Nodes – 320 Processes



VPS – Summary

- **Performance**
  - FDR InfiniBand delivers the highest network performance for VPS to scale
  - FDR IB provides higher performance against other networks
    - FDR IB delivers ~162% higher compared to 40GbE, ~178% vs 10GbE on a 32 node run
  - MPI-OpenMP Hybrid mode can provide better performance at scale
    - About 8% performance increase at 32 nodes with hybrid
  - Enabling Turbo mode results in higher application performance
    - Up to 8% of the improvement seen by enabling Turbo mode
    - At the expense of ~11% in higher power utilization
  - The default MPI implementation provides similarly as Open MPI 1.8 in HPC-X
    - With FCA enabled, Open MPI runs about 10% faster than Platform MPI at 32 nodes
- **MPI Profiling**
  - Majority of MPI communication time comes from MPI_Allreduce
    - About 71% of the time spent in MPI_Allreduce
  - Ethernet solutions consumes more time in communications
    - Spent 71%-93% of overall time in network due to congestion in Ethernet, while IB spent ~33%

Thank You

HPC Advisory Council