HPC ADVISORY COUNCIL

VASP
Performance Benchmark and Profiling

January 2013

AMD · DELL · Mellanox TECHNOLOGIES · VASP (Vienna Ab-initio Simulation Package)

Note

- **The following research was performed under the HPC Advisory Council activities**
  – Participating vendors: AMD, Dell, Mellanox
  – Compute resource –
    - HPC Advisory Council Cluster Center
- **For more info please refer to**
  – http://www.amd.com
  – http://www.dell.com/hpc
  – http://www.mellanox.com
  – http://www.vasp.at

VASP

- **VASP**
  - Stands for “Vienna Ab-initio Simulation Package”
  - Performs ab-initio quantum-mechanical molecular dynamics (MD)
    - using pseudopotentials and a plane wave basis set
  - The code is written in FORTRAN 90 with MPI support
  - Access to the code may be given by a request via the VASP website
- **The approach that used in VASP is based on the techniques:**
  - A finite-temperature local-density approximation, and
  - An exact evaluation of the instantaneous electronic ground state at each MD-step using efficient matrix diagonalization schemes and an efficient Pulay mixing
- **These techniques avoid problems in original Car-Parrinello method**
  - which is based on the simultaneous integration of electronic and ionic equations of motion

Objectives

- **The following was done to provide best practices**
  - VASP performance benchmarking
  - Understanding VASP communication patterns
  - Ways to increase VASP productivity
  - Compilers and network interconnects comparisons

- **The presented results will demonstrate**
  - The scalability of the compute environment
  - The capability of VASP to achieve scalable productivity
  - Considerations for performance optimizations

Test Cluster Configuration

- Dell™ PowerEdge™ R815 11-node (704-core) cluster
- AMD™ Opteron™ 6276 (code name "Interlagos") 16-core @ 2.3 GHz CPUs
- 4 CPU sockets per server node
- Mellanox ConnectX®-3 FDR InfiniBand Adapters
- Mellanox SwitchX™ 6036 36-Port InfiniBand switch
- Memory: 128GB memory per node DDR3 1333MHz
- OS: RHEL 6.2, SLES 11.2 with MLNX-OFED 1.5.3 InfiniBand SW stack
- MPI: Intel MPI 4 Update 3, MVAPICH2 1.8.1, Open MPI 1.6.3 (w/ dell_affinity 0.85)
- Math Libraries: ACML 5.2.0, Intel MKL 11.0, SCALAPACK 2.0.2
- Compilers: Intel Compilers 13.0, Open64 4.5.2
- Application: VASP 5.2.7
- Benchmark workload:
  - Pure Hydrogen (MD simulation, 10 iconic steps, 60 electronic steps, 264 bands, IALGO=48)

Dell™ PowerEdge™ R815 11-node cluster

HPC ADVISORY COUNCIL

- **HPC Advisory Council Test-bed System**
- **New 11-node 704 core cluster - featuring Dell PowerEdge™ R815 servers**
  - Replacement system for Dell PowerEdge SC1435 (192 cores) cluster system following 2 years of rigorous benchmarking and product EOL
    - System to be redirected to explore HPC in the Cloud applications
- **Workload profiling and benchmarking**
  - Characterization for HPC and compute intense environments
  - Optimization for scale, sizing and configuration and workload performance
  - Test-bed Benchmarks
    - RFPs
    - Customers/Prospects, etc
  - ISV & Industry standard application characterization
  - Best practices & usage analysis

About Dell PowerEdge™ Platform Advantages

HPC ADVISORY COUNCIL

Best of breed technologies and partners

Combination of AMD Opteron™ 6200 series platform and Mellanox ConnectX®-3 InfiniBand on Dell HPC

Solutions provide the ultimate platform for speed and scale

- Dell PowerEdge R815 system delivers 4 socket performance in dense 2U form factor
- Up to 64 core/32DIMMs per server – 1344 core in 42U enclosure

Integrated stacks designed to deliver the best price/performance/watt

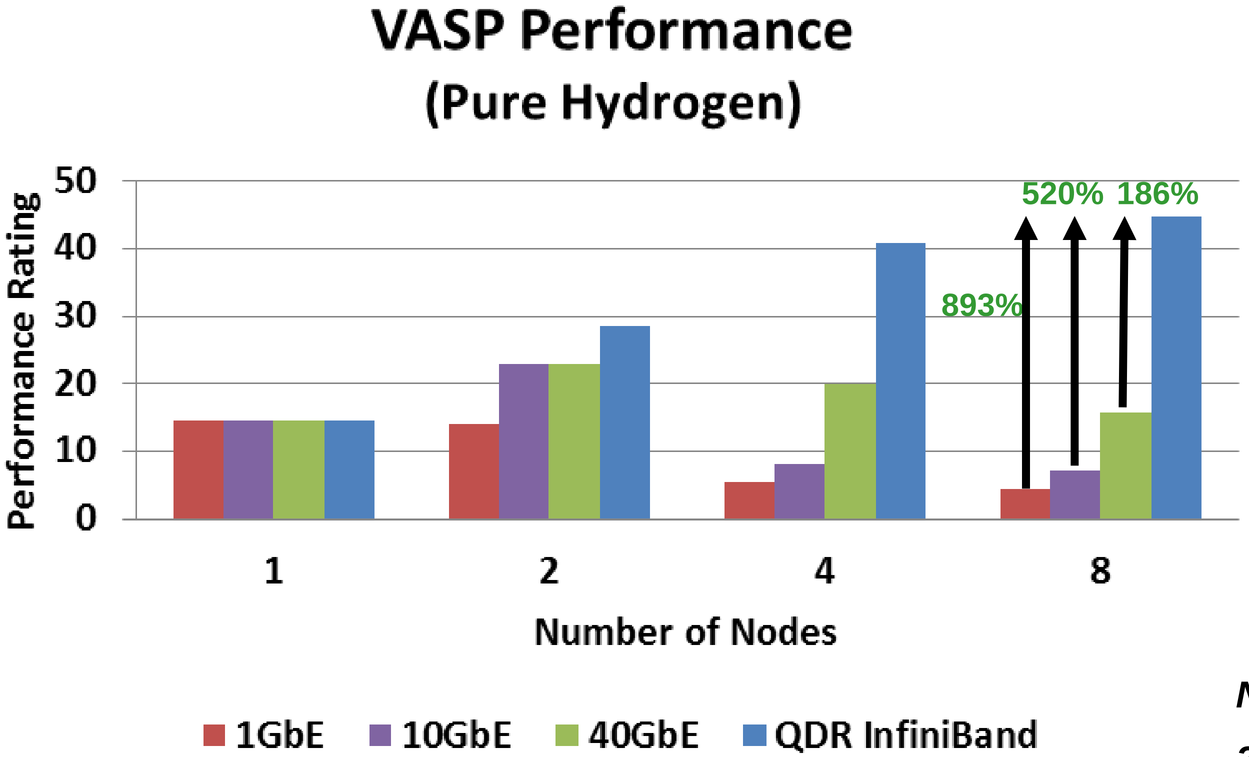
- 2x more memory and processing power in half of the space
- Energy optimized low flow fans, improved power supplies and dual SD modules

Optimized for long-term capital and operating investment protection

- System expansion
- Component upgrades and feature releases

VASP Performance – Interconnect

- **QDR InfiniBand delivers the best performance for VASP**
  - Up to 186% better performance than 40GbE on 8 nodes
  - Over 5 times better performance than 10GbE on 8 nodes
  - Over 8 times better performance than 1GbE on 8 nodes
- **Scalability limitation seen with Ethernet networks**
  - 1GbE, 10GbE and 40GbE performance starts to decline after 2 nodes



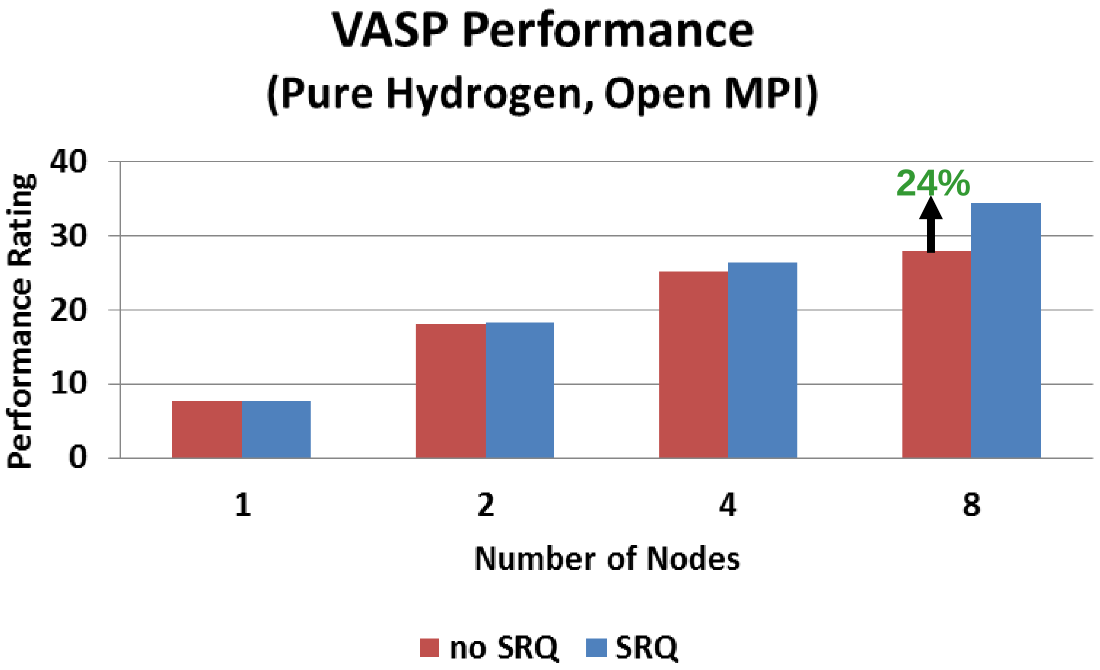
NPAR=8

32 Cores/Node

Higher is better

VASP Performance – Open MPI w/ SRQ

- **Using SRQ enables better performance for VASP at high core counts**
  - 24% higher performance than Open MPI at 8 nodes
- **Flags used for enabling SRQ in Open MPI:**
  - -mca btl_openib_receive_queues S,9216,256,128,32:S,65536,256,128,32
  - Processor binding is enabled for both cases



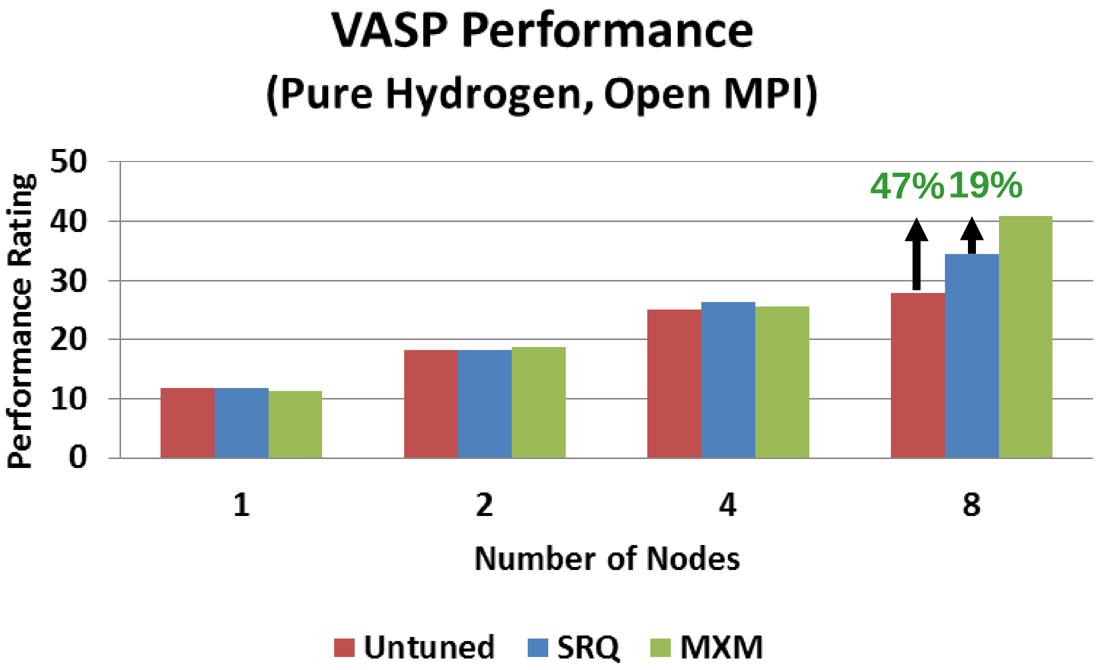
NPAR=8

32 Cores/Node

Higher is better

VASP Performance – Open MPI w/ MXM

- **Enabling MXM enables better performance for VASP at high core counts**
  - 47% higher job productivity than the untuned Open MPI run at 8 nodes
  - 19% higher job productivity than the SRQ-enabled Open MPI at 8 nodes
- **Flags used for enabling MXM in Open MPI:**
  - -mca mtl mxm -mca btl_openib_free_list_num 8192 -mca btl_openib_free_list_inc 1024 -mca mpi_preconnect_mpi 1 -mca btl_openib_flags 9
  - Processor binding using dell_affinity.exe for all 3 cases



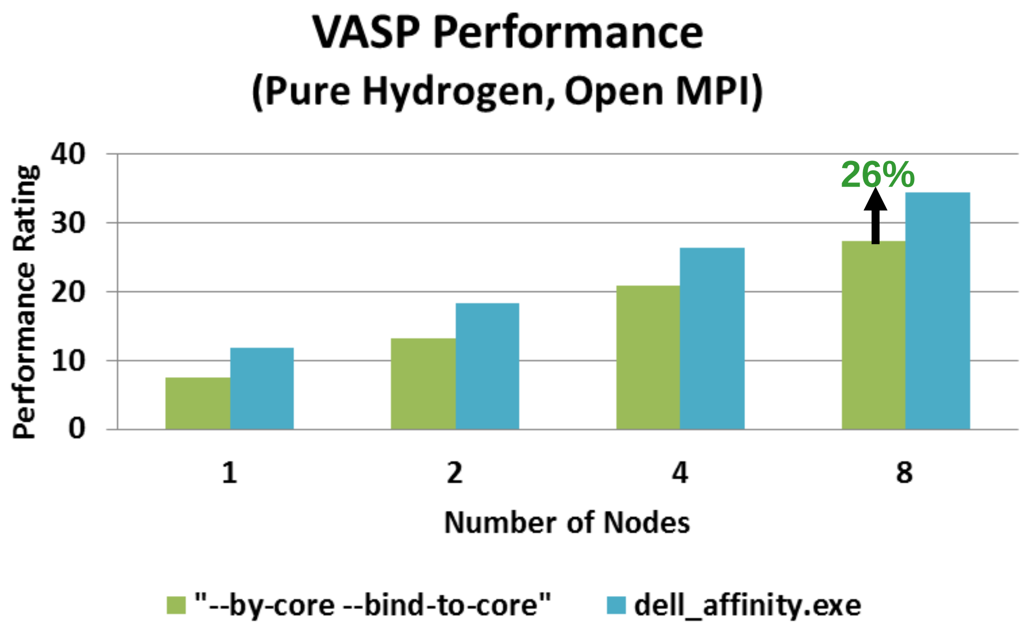
NPAR=8

Higher is better

32 Cores/Node

VASP Performance – Processor Binding

- Processor binding is crucial for achieving he best performance on AMD Interlagos
  - Allocating MPI processes on the most optimal cores allows Open MPI to perform
- Options that are used between the 2 cases:
  - bind-to-core: OMPI param: **--bind-to-core** (OMPI compiled with hwloc support)
  - dell_affinity.exe: **dell_affinity.exe -v -n 32 -t 1**
- dell_affinity is described at the HPC Advisory Council Spain Conference 2012:
  - http://www.hpcadvisorycouncil.com/events/2012/Spain-Workshop/pres/7_Dell.pdf
  - Works with all open source and commercial MPI libraries on Dell platforms



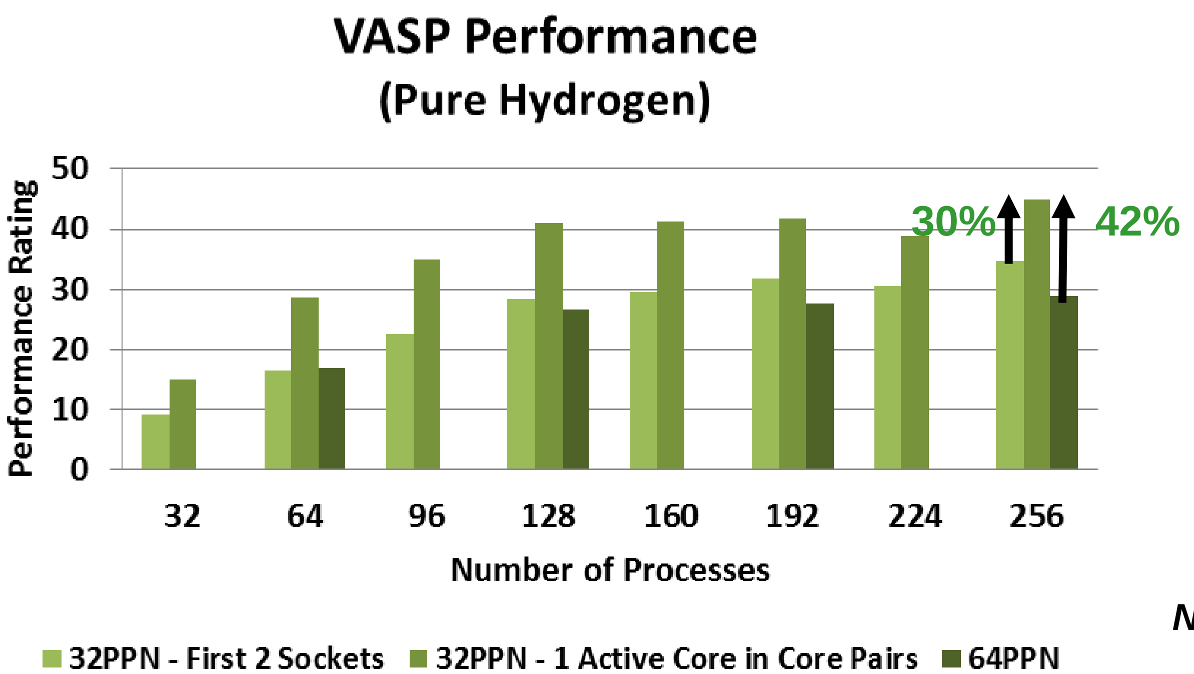
NPAR=8

32 Cores/Node

Higher is better

VASP Performance – Processes Per Node

- **Running 1 active core in core pairs yield higher system utilization**
  - 42% gain in performance with 64 PPN versus 32 PPN (with 1 active core) for 8 nodes
  - 1 floating point unit (FPU) is shared between 2 CPU cores in a package
- **Using 4P servers deliver higher performance than 2P servers**
  - 30% gain with 4P server (32 PPN with 1 active core/package) than 32PPN in a 2P

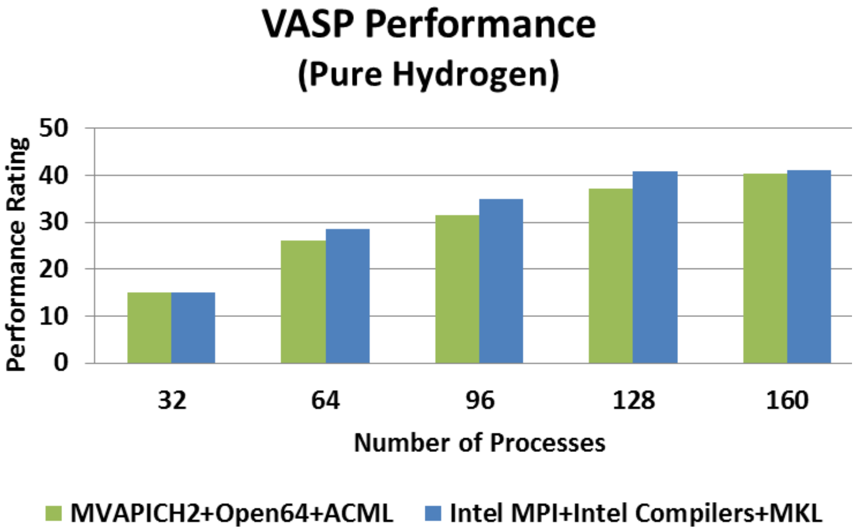


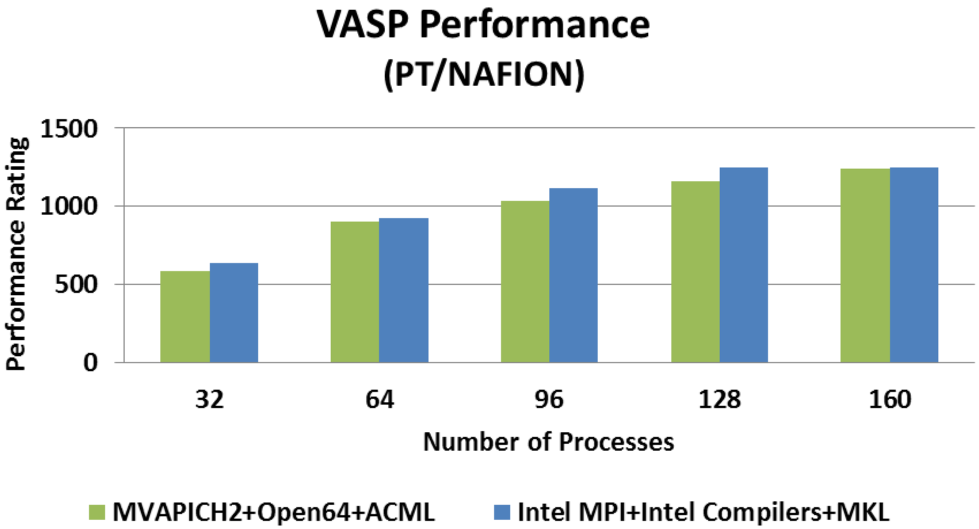
NPAR=8

Higher is better

VASP Performance – Software Stack

- **Free software stack performs comparably to Intel software stack**
  - Free software stack: MVAPICH2, Open64 compilers, ACML and ScaLAPACK
- **Specifications regarding the runs:**
  - Intel Compiler/MKL: Default optimization in Makefile with minor modification to loc
  - MVAPICH2 runs with processor affinity "**dell_affinity.exe -v -n32 -t 1**"
  - Open64: Compiler flags: **"-march=bdver1 -mavx –mfma**"
  - Changes to interface blocks of certain modules to support Open64 can be accessed and acquired from the University of Vienna



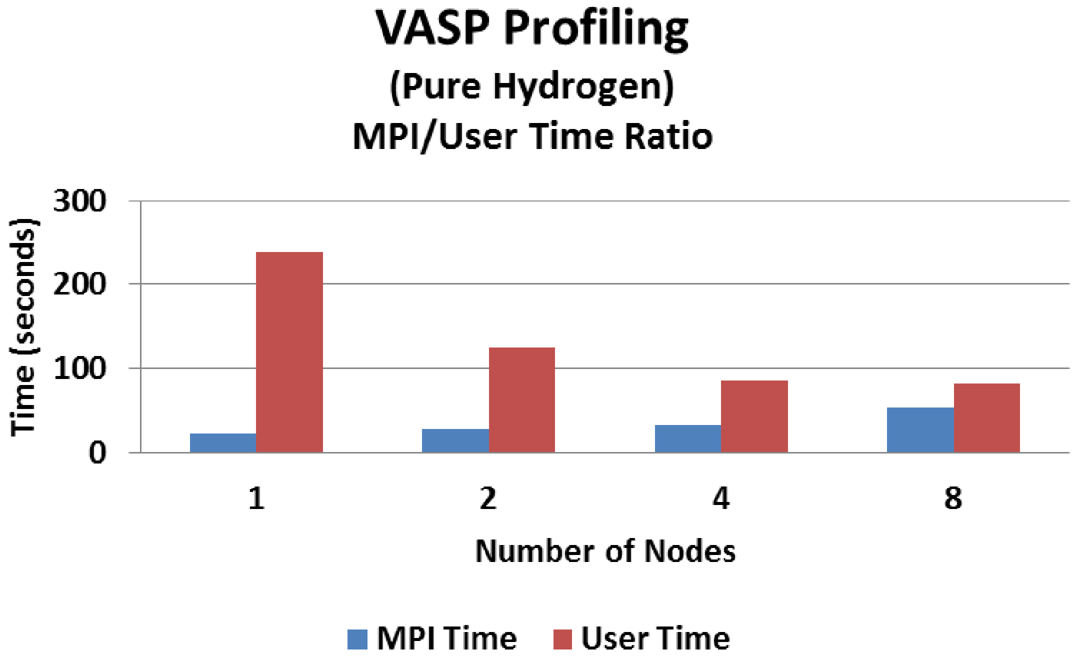


Higher is better

32 Cores/Node

VASP Profiling – MPI/User Time Ratio

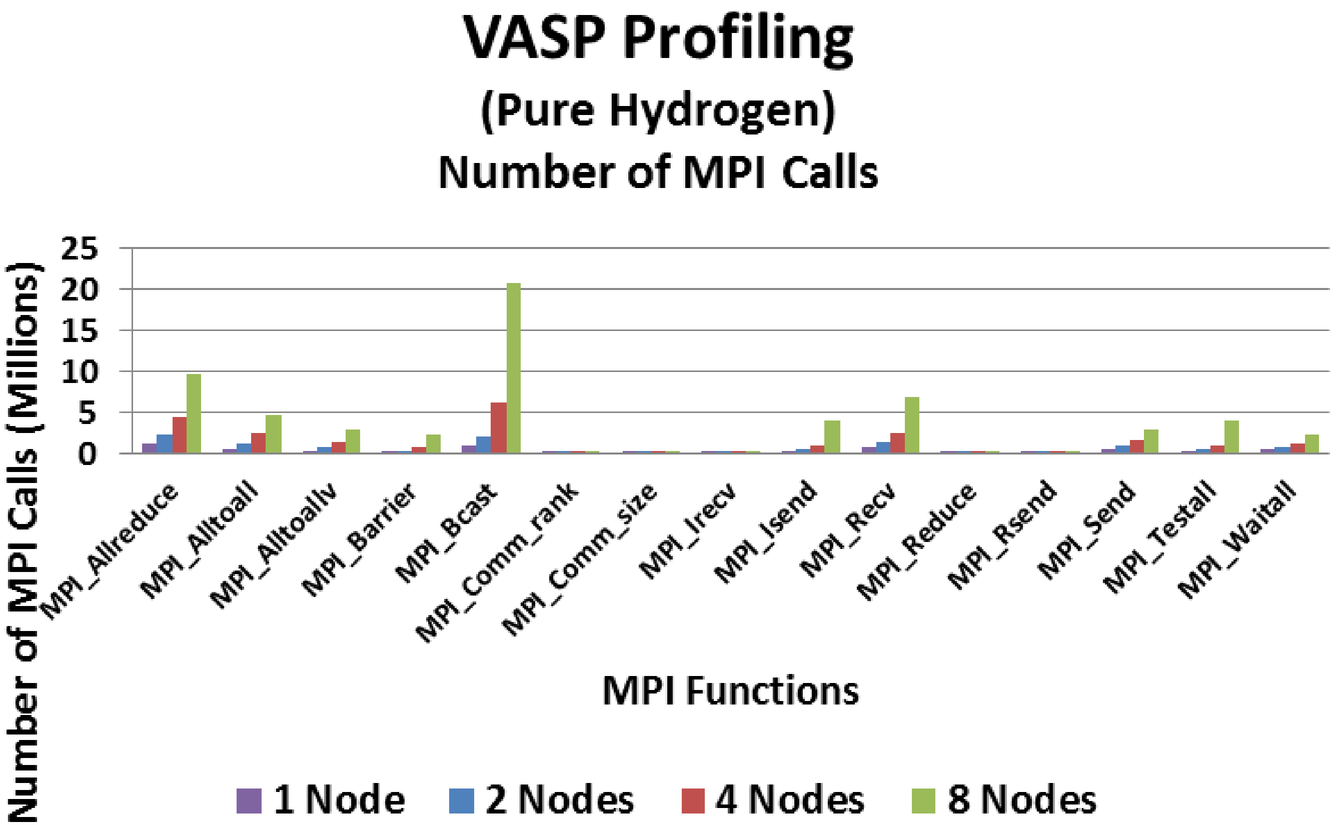
- **QDR InfiniBand reduces the amount of time for MPI communications**
  - MPI Communication time increase gradually as the compute time reduces



32 Cores/Node

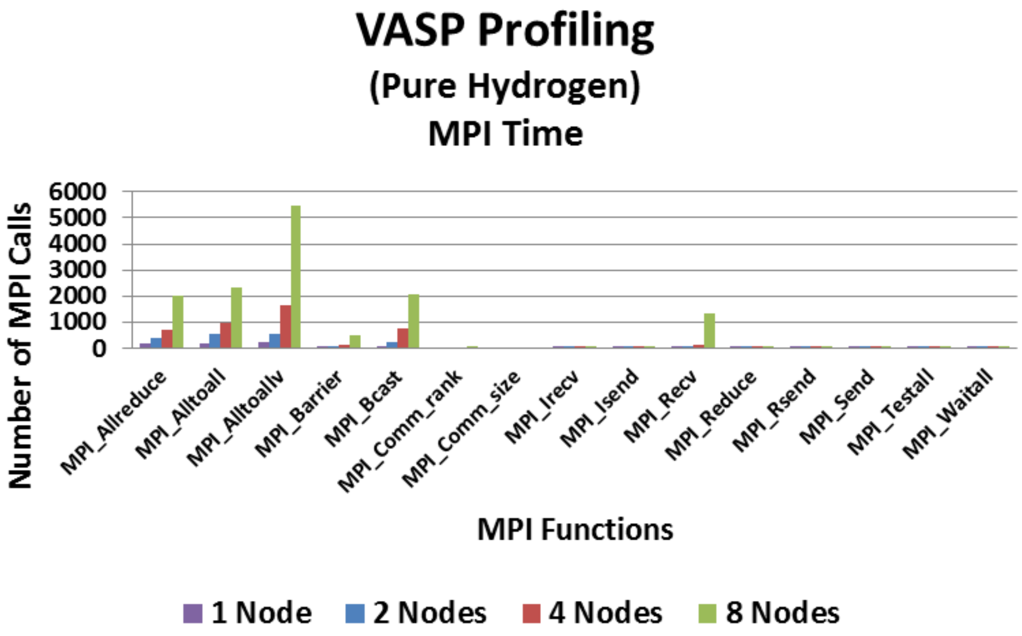
VASP Profiling – Number of MPI Calls

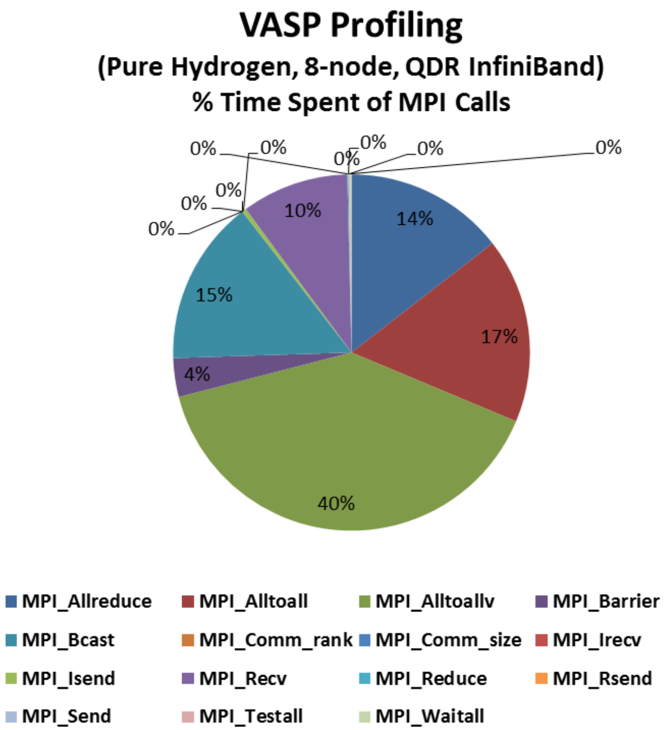
- **The most used MPI functions are for MPI collective operations**
  – MPI_Bcast(34%), MPI_Allreduce(16%), MPI_Recv(11%), MPI_Alltoall(8%) at 8 nodes
  – Collective operations cause communication time to grow at larger node counts



VASP Profiling – Time Spent of MPI Calls

- **The time in communications is taken place in the following MPI functions:**
  – MPI_Alltoallv(40%) MPI_Alltoall(17%), MPI_Bcast (15%) at 8 nodes

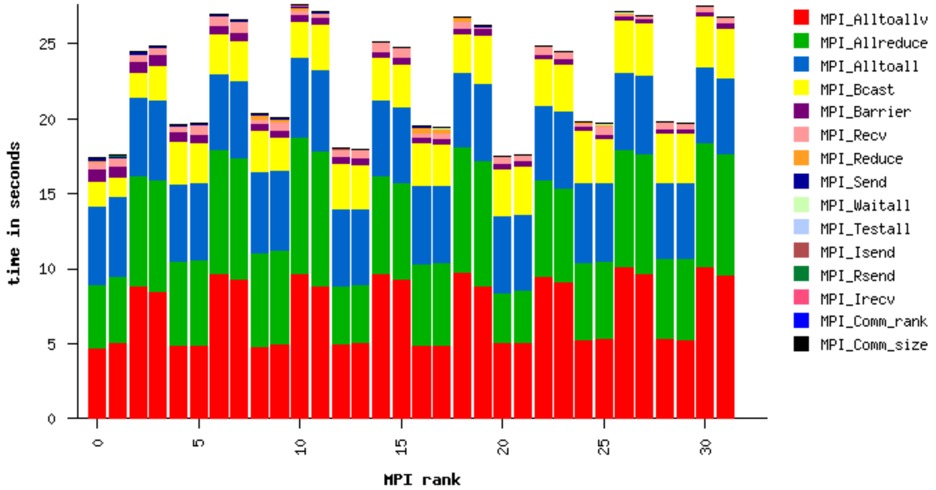




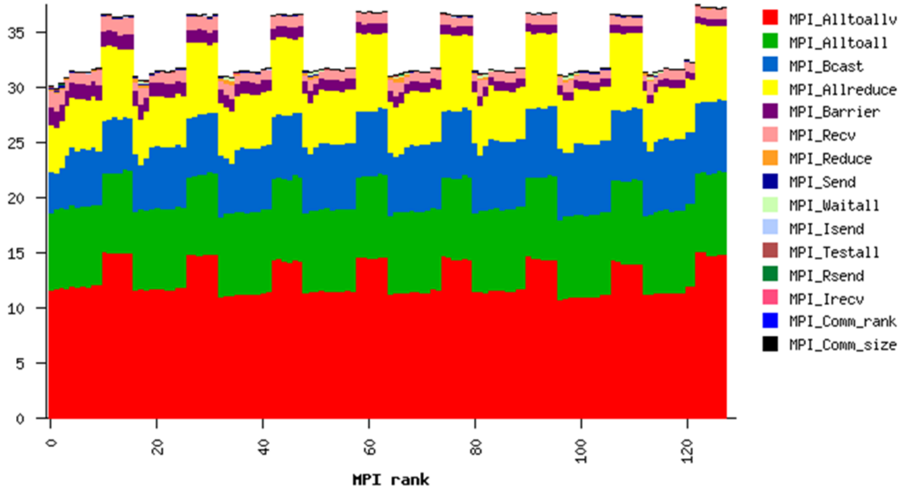
VASP Profiling – MPI Message Sizes

- **Uneven communication seen for MPI messages**
  - More MPI collective communication takes place on certain MPI ranks

1 Nodes – 32 Processes



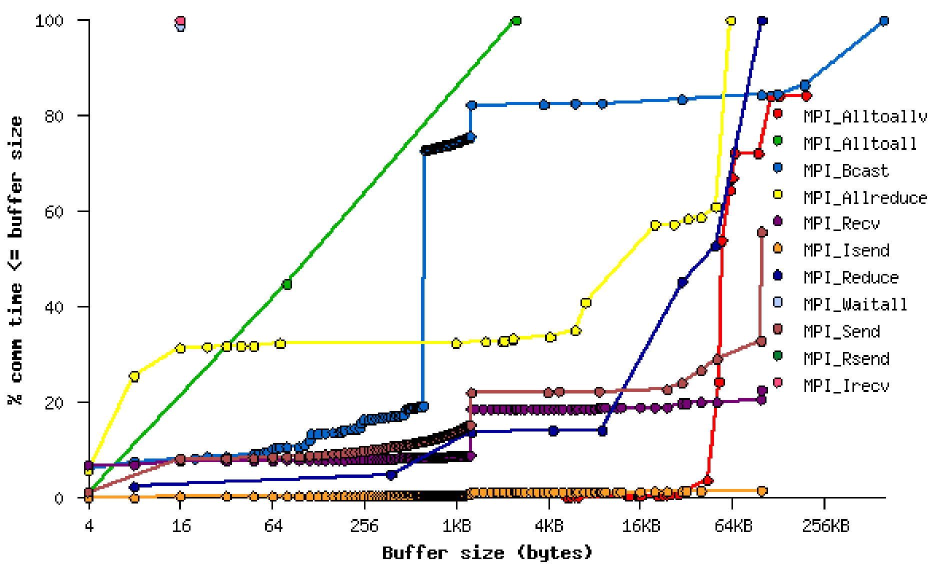
4 Nodes – 128 Processes



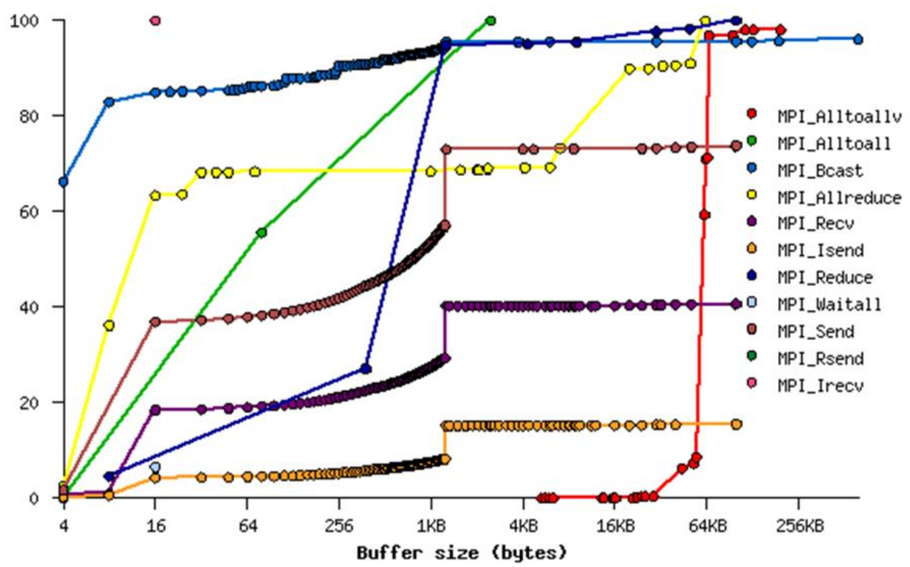
VASP Profiling – MPI Message Sizes

- **Communication pattern changes as more processes involved**
  - 4 nodes: Majority of messages are concentrated at 64KB
  - 8 nodes: MPI_Alltoallv is the largest MPI time consumer, is largely concentrated at 64KB
  - 8 nodes: MPI_Bcast is the most frequently called MPI API, is largely concentrated at 4B

4 Nodes – 128 Processes



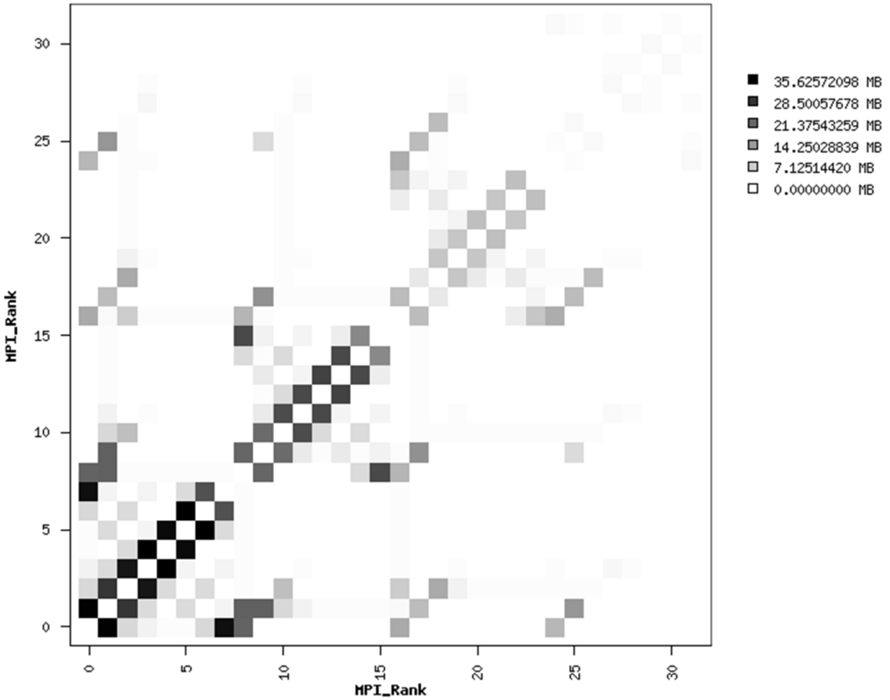
8 Nodes – 256 Processes



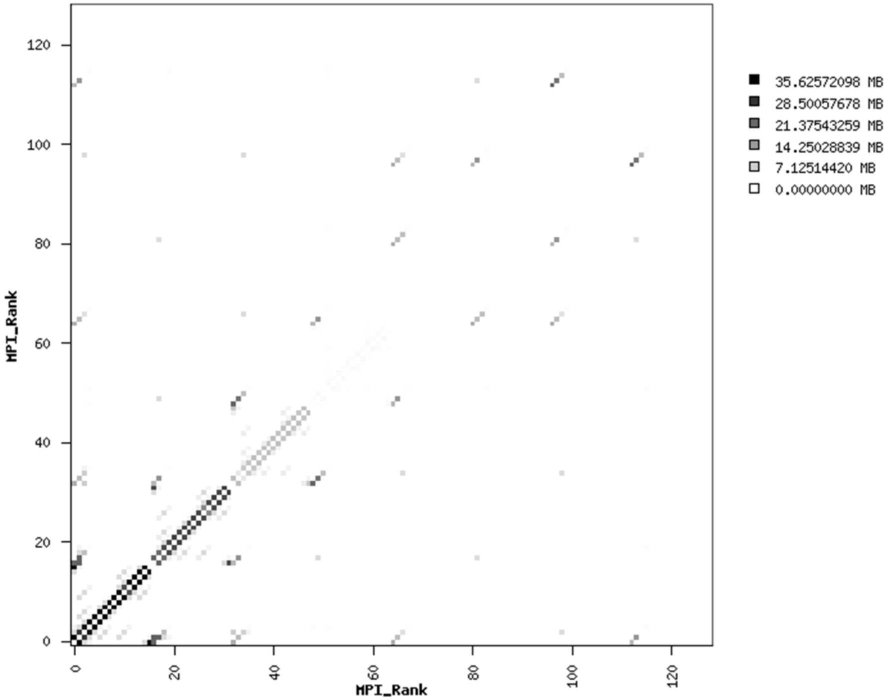
VASP Profiling – Point To Point Data Flow

- **The point to point data flow shows the communication pattern of VASP**
  - VASP mainly communicates on lower ranks
  - The pattern stays the same as the cluster scales

1 Nodes – 32 Processes



4 Nodes – 128 Processes



Summary

- **Low latency in network communication is required to make VASP scalable**
  - QDR InfiniBand delivers good scalability and provides lowest latency among the tested:
    - 186% versus 40GbE, over 5 times better than 10GbE and over 8 times than 1GbE on 8 nodes
  - Ethernet would not scale and become inefficient to run beyond 2 nodes
  - Mellanox messaging accelerations (SRQ and MXM) can provide benefit for VASP to run at scale
  - Heavy MPI collective communication occurred in VASP
- **CPU:**
  - Running single core in core pairs performs 42% faster than running with both cores
  - “dell_affinity.exe” ensures proper process allocation support in Open MPI and MVAPICH2
- **Software stack:**
  - Free stack (Open64/MVAPICH2/ACML/ScaLAPACK) performs comparably to Intel stack
  - Additional performance is expected with source code optimizations and tuning using the latest development tools (such as Open64, ACML) that support AMD “Interlagos” architecture

Thank You

HPC Advisory Council