HPC·AI
ADVISORY COUNCIL
NETWORK OF EXPERTISE

VASP
Performance Benchmark and Profiling

August 2020

b-initio
VASP
ackage
ienna
imulation

Note

- **The following research was performed under the HPC Advisory Council activities**
  - Systems
    - HPC Advisory Council Cluster Center Iris cluster
    - HPC Advisory Council Cluster Center Helios cluster
    - Texas Advanced Computing Center Frontera supercomputer

- **The following was done to provide best practices**
  - VASP performance overview over Intel based platforms
  - Understanding VASP communication patterns

- **More info on VASP**
  - https://www.vasp.at/

Vienna Ab initio Simulation Package (VASP)

- The Vienna Ab initio Simulation Package (VASP) is a computer program for atomic scale materials modelling, e.g. electronic structure calculations and quantum-mechanical molecular dynamics

- VASP computes an approximate solution to the many-body Schrödinger equation, either within density functional theory (DFT), solving the Kohn-Sham equations, or within the Hartree-Fock (HF) approximation, solving the Roothaan equations.

- Hybrid functionals that mix the Hartree-Fock approach with density functional theory are implemented as well. Furthermore, Green's functions methods (GW quasiparticles, and ACFDT-RPA) and many-body perturbation theory (2nd-order Møller-Plesset) are available in VASP

Helios Cluster Configuration

- **Helios cluster**
  - Supermicro SYS-6029U-TR4 / Foxconn Groot 1A42USF00-600-G 32-node cluster
  - Dual Socket Intel Xeon Gold 6138 CPU @ 2.00GHz
  - Mellanox ConnectX-6 HDR100 InfiniBand
  - Mellanox Quantum Switch HDR InfiniBand
  - Memory: 192GB DDR4 2677MHz RDIMMs per node
  - Lustre Storage

- **Software**
  - OS: RHEL 7.7, MLNX_OFED 4.7.3
  - MPI: HPC-X 2.6.0
  - Compiler: Intel 2020.1.217
  - VASP : 6.1
    - NSIM: 16
    - NCORE: 40
    - KPAR: 4

Iris Cluster Configuration

- **Iris cluster**
  - Dell C6400 32-node cluster
  - Dual Socket Intel(R) Xeon(R) Gold 6148 CPU @ 2.40GHz
  - Mellanox ConnectX-6 HDR100 100Gb/s InfiniBand adapters
  - Mellanox HDR Quantum Switch QM7800 40-Port 200Gb/s HDR InfiniBand
  - Memory: 192GB DDR4 2666MHz RDIMMs per node
  - Lustre Storage

- **Software**
  - OS: RHEL 7.7, MLNX_OFED 4.7.3
  - MPI: HPC-X 2.6.0
  - Compiler: Intel 2020.1.217
  - VASP : 6.1
    - NSIM:16,4
    - NCORE: 40
    - KPAR: 4

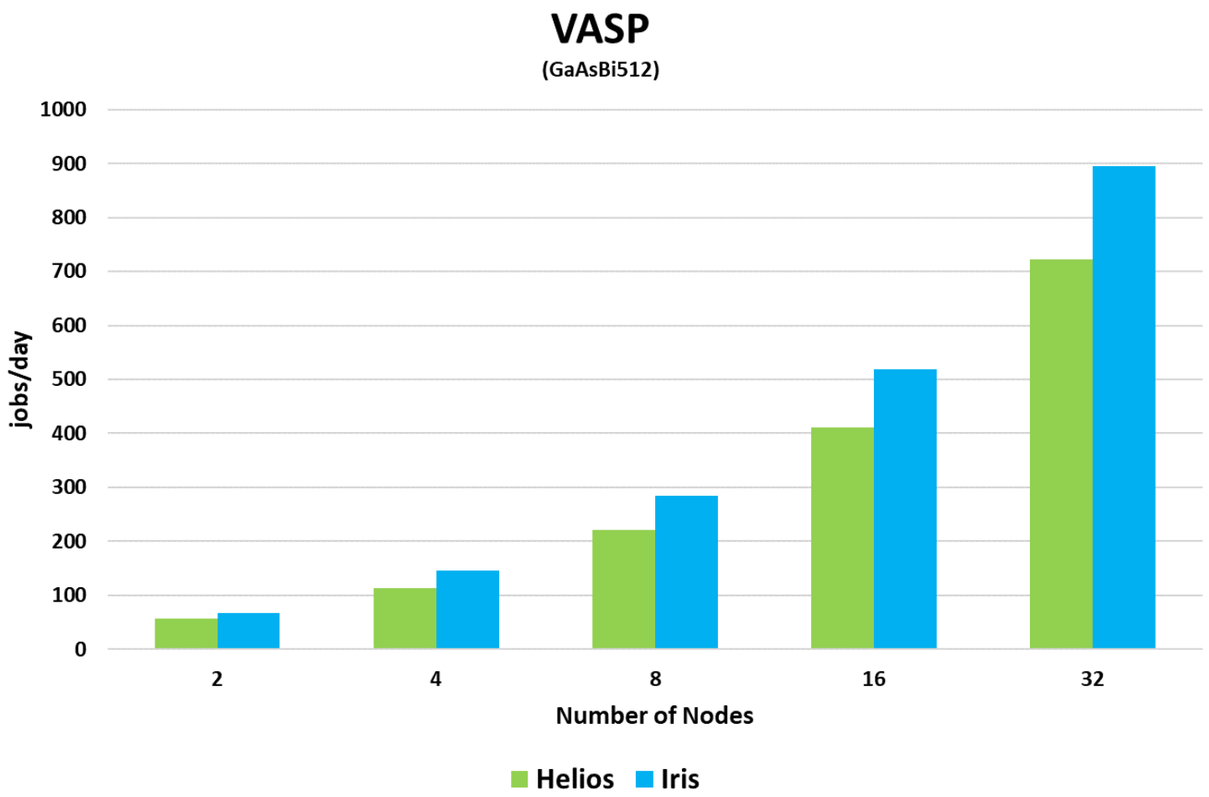
Frontera Cluster Configuration

- **Frontera cluster (TACC)**
  - Dual Socket Intel(R) Xeon(R) Gold 8280 CPU @ 2.70GHz
  - Mellanox ConnectX-6 HDR100 100Gb/s InfiniBand adapters
  - Mellanox HDR Quantum Switch QM7800 40-Port 200Gb/s HDR InfiniBand
  - Lustre Storage

- **Software**
  - OS: RHEL 7.8, MLNX_OFED 5.0.2
  - MPI: HPC-X 2.6.0
  - Compiler: Intel 2020.1.217
  - VASP : 6.1
    - NSIM:16
    - NCORE: 56
    - KPAR:4

VASP Performance - Clusters

- VASP demonstrated 24% higher performance on the Iris cluster



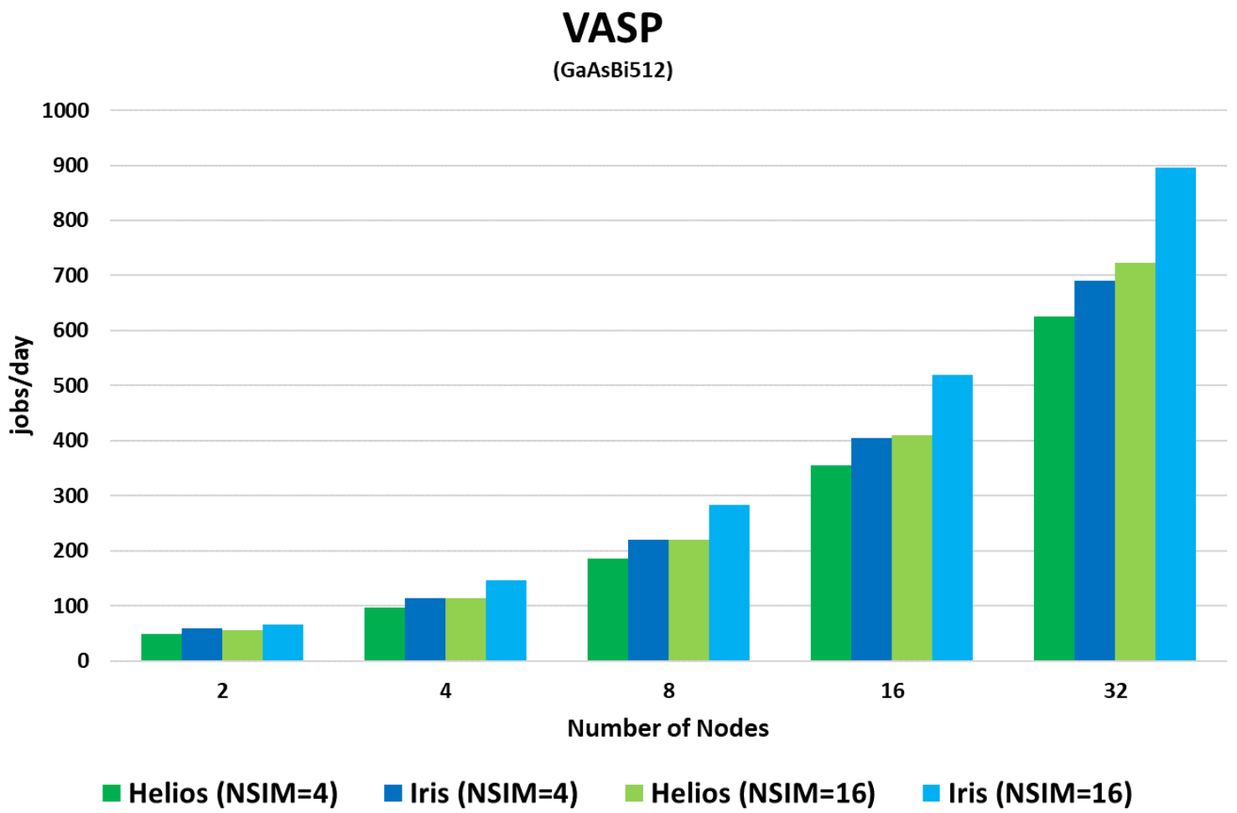
NSIM=16

Higher is Better

NSIM: Sets the number of bands that are optimized simultaneously by the RMM-DIIS algorithm

VASP Performance - NSIM

- With NSIM 16 VASP demonstrated 30% higher performance versus NSIM 4



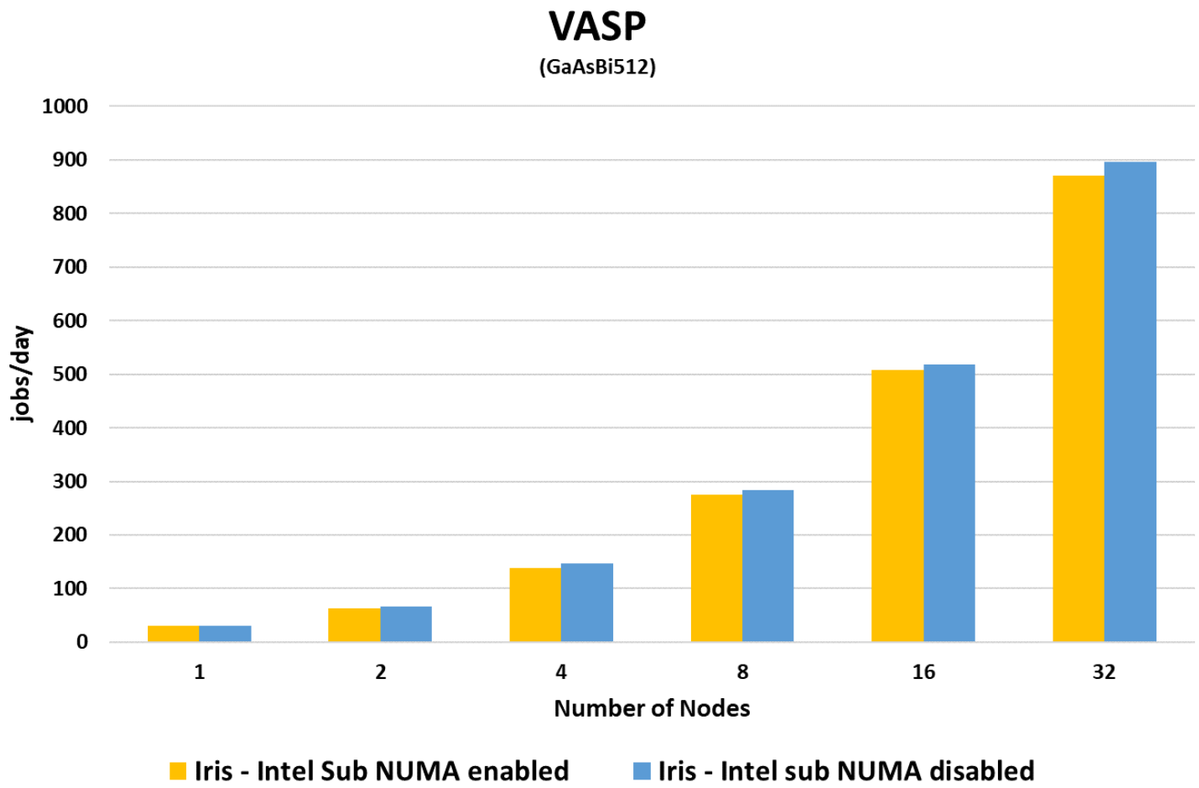
NSIM=4, 16

Higher is Better

NSIM: Sets the number of bands that are optimized simultaneously by the RMM-DIIS algorithm

VASP Performance – sub-NUMA

- 3% higher performance with sub-NUMA disable mode



NSIM=16

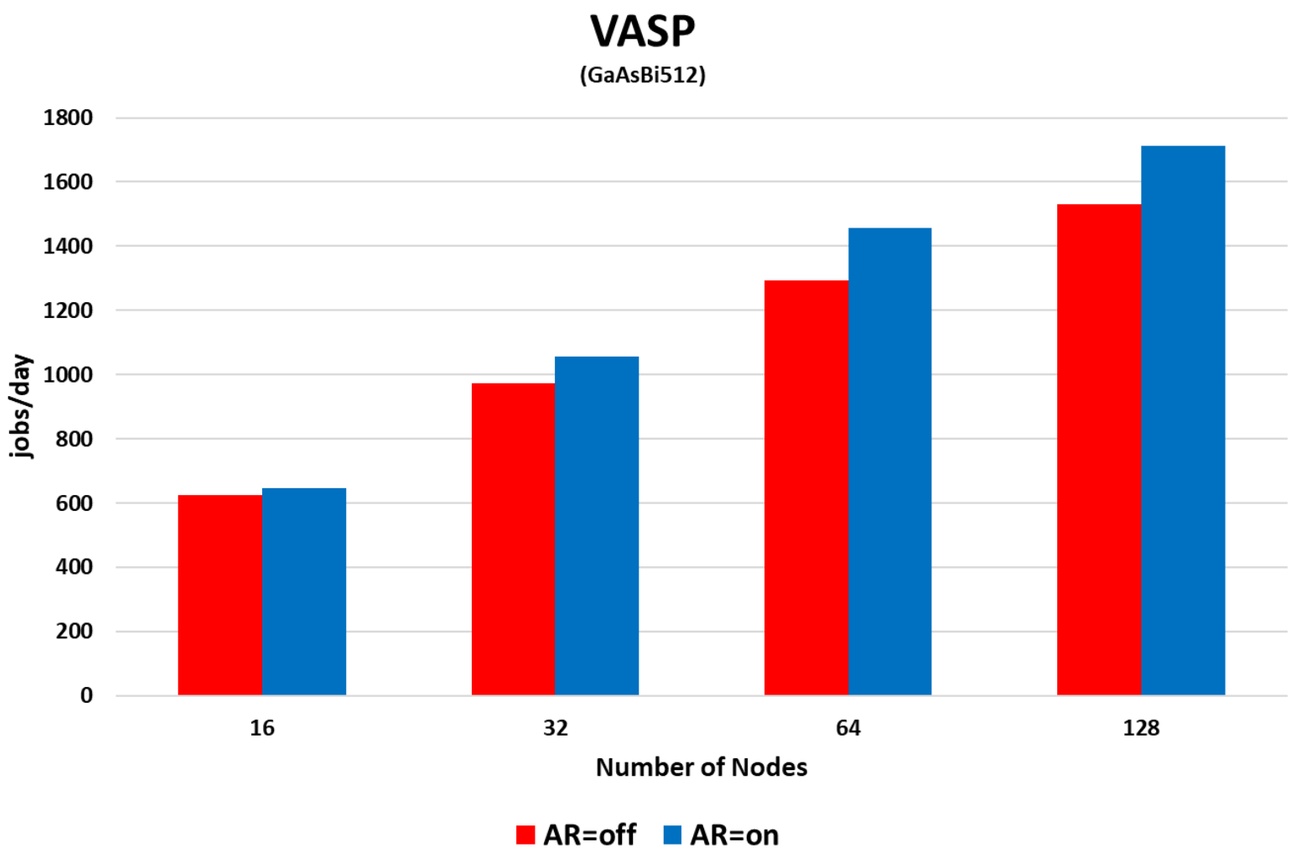
BIOS Sub-NUMA – Enabled/ Disabled

Higher is Better

Sub-NUMA Clustering divides the cores, cache, and memory of the processor into multiple NUMA domains
NSIM: Sets the number of bands that are optimized simultaneously by the RMM-DIIS algorithm

VASP Performance - Adaptive Routing

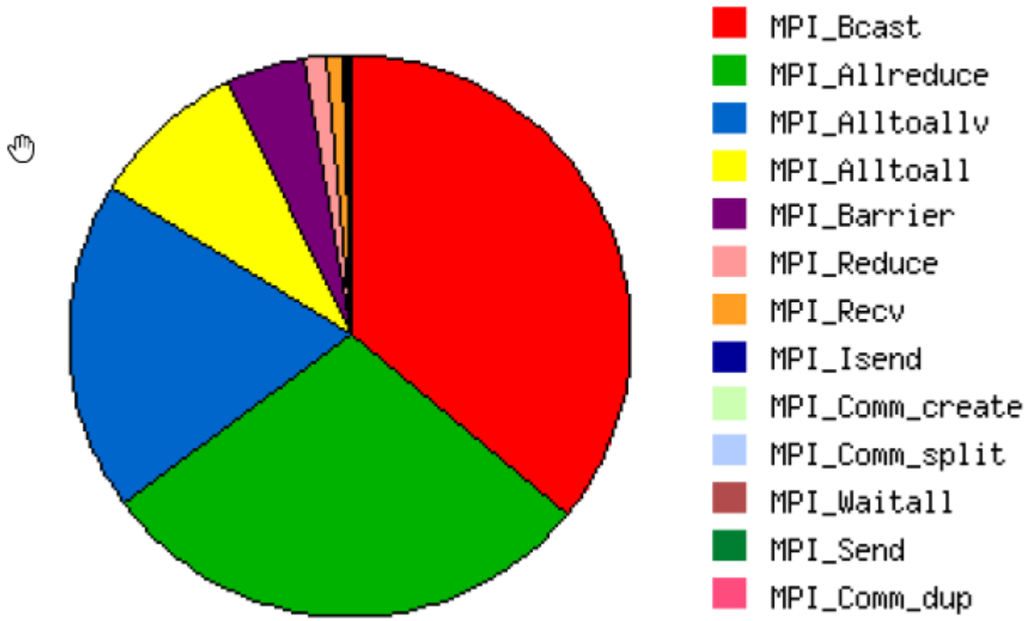
- **InfiniBand adaptive routing enable 13% higher performance**



AR – Adaptive Routing

VASP MPI Profile, 32 Nodes Iris Cluster

- **35% of VASP time is spent in MPI**



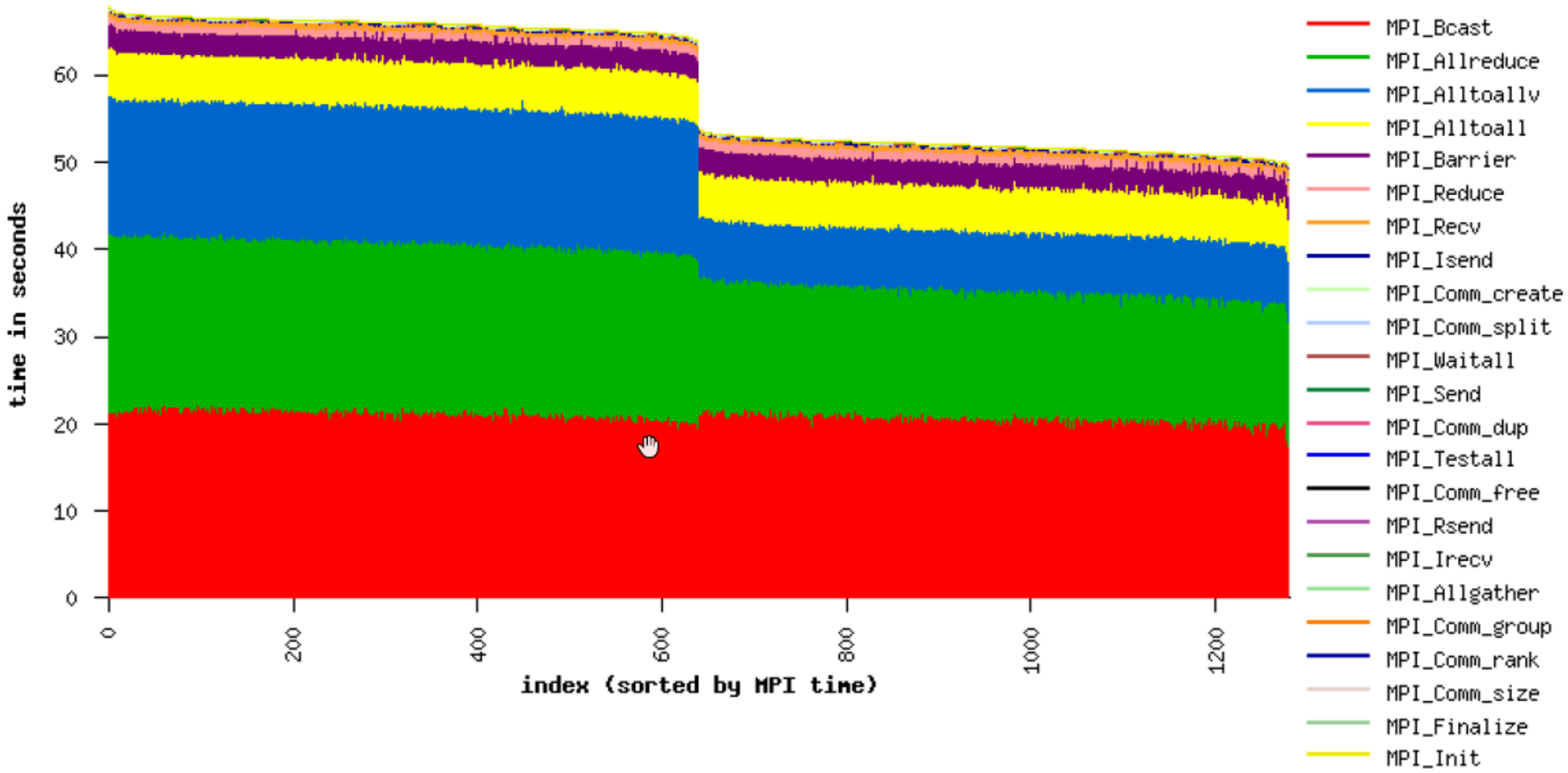
VASP MPI Profile, 32 Nodes Iris Cluster

- ~35% of MPI Communication spent on MPI_Bcast
- ~30% MPI Allreduce on large messages
- Alltoallv, and Alltoall are used in medium/large message size

Communication Event Statistics (% detail, --- error)									
	Comm Size	Buffer Size	Ncalls	Total Time	Avg Time	Min Time	Max Time	%MPI	%Wall
MPI_Bcast	0	2097152	8847360	1.885305e+04	2.130924e-03	4.200900e-04	4.752300e-02	25.05	8.93
MPI_Allreduce	0	57344	69852080	1.430066e+04	2.047278e-04	2.813300e-05	4.530700e-01	19.00	6.77
MPI_Alltoallv	0	3584	13634680	8.035205e+03	5.893211e-04	1.330400e-04	1.417200e-02	10.68	3.81
MPI_Alltoall	0	7168	24184320	4.219188e+03	1.744597e-04	7.867800e-06	7.304900e-03	5.61	2.00
MPI_Alltoallv	0	5120	16040800	3.862791e+03	2.408104e-04	1.289800e-04	7.963200e-03	5.13	1.83
MPI_Barrier	0	0	28041560	3.386375e+03	1.207627e-04	0.000000e+00	1.754100e-01	4.50	1.60
MPI_Allreduce	0	8	16107480	2.631645e+03	1.633803e-04	0.000000e+00	4.276000e-01	3.50	1.25
MPI_Bcast	0	4	17267840	2.325260e+03	1.346584e-04	0.000000e+00	1.777600e-01	3.09	1.10
MPI_Allreduce	0	16	27645280	2.080444e+03	7.525495e-05	9.536700e-07	2.568000e-02	2.76	0.99
MPI_Alltoall	0	320	29378560	1.519206e+03	5.171137e-05	1.907300e-06	1.935000e-02	2.02	0.72
MPI_Alltoallv	0	4096	2406120	1.426994e+03	5.930685e-04	1.339900e-04	9.658100e-03	1.90	0.68
MPI_Allreduce	0	40960	20533680	1.420516e+03	6.917982e-05	3.385500e-05	5.370900e-03	1.89	0.67
MPI_Alltoall	0	4	30720	9.946785e+02	3.237886e-02	1.096700e-05	7.618100e-01	1.32	0.47
MPI_Reduce	0	57344	20480	9.457655e+02	4.617995e-02	3.600100e-05	5.006700e-01	1.26	0.45
MPI_Bcast	0	1024	10983076	8.542726e+02	7.778081e-05	0.000000e+00	4.996100e-03	1.14	0.40
MPI_Bcast	0	1280	5890560	8.163482e+02	1.385858e-04	9.536700e-07	7.362100e-03	1.08	0.39
MPI_Bcast	0	256	12926088	7.847797e+02	6.071285e-05	0.000000e+00	4.300100e-03	1.04	0.37

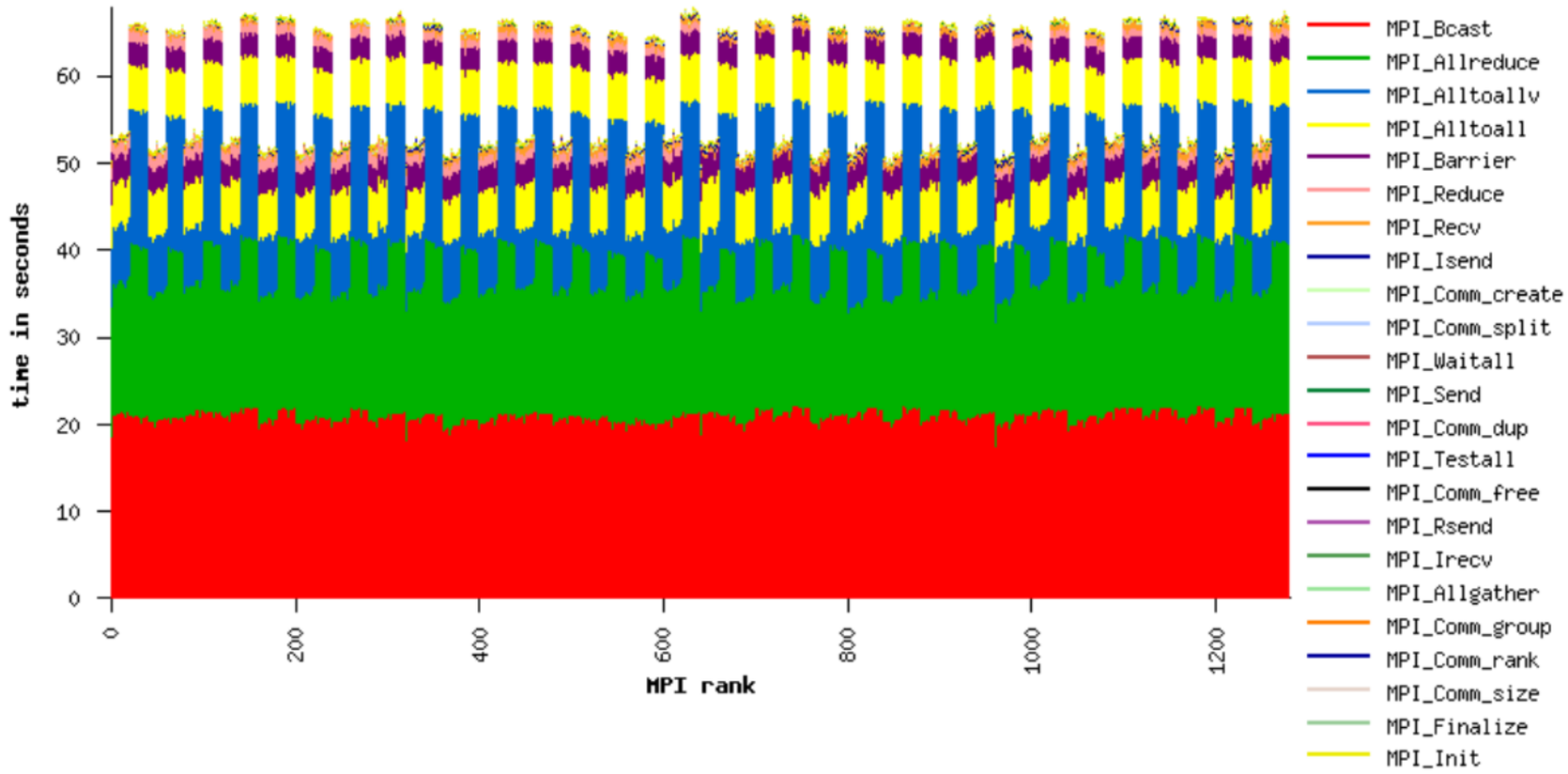
VASP MPI Profile, 32 Nodes Iris Cluster

- **Socket Imbalance**



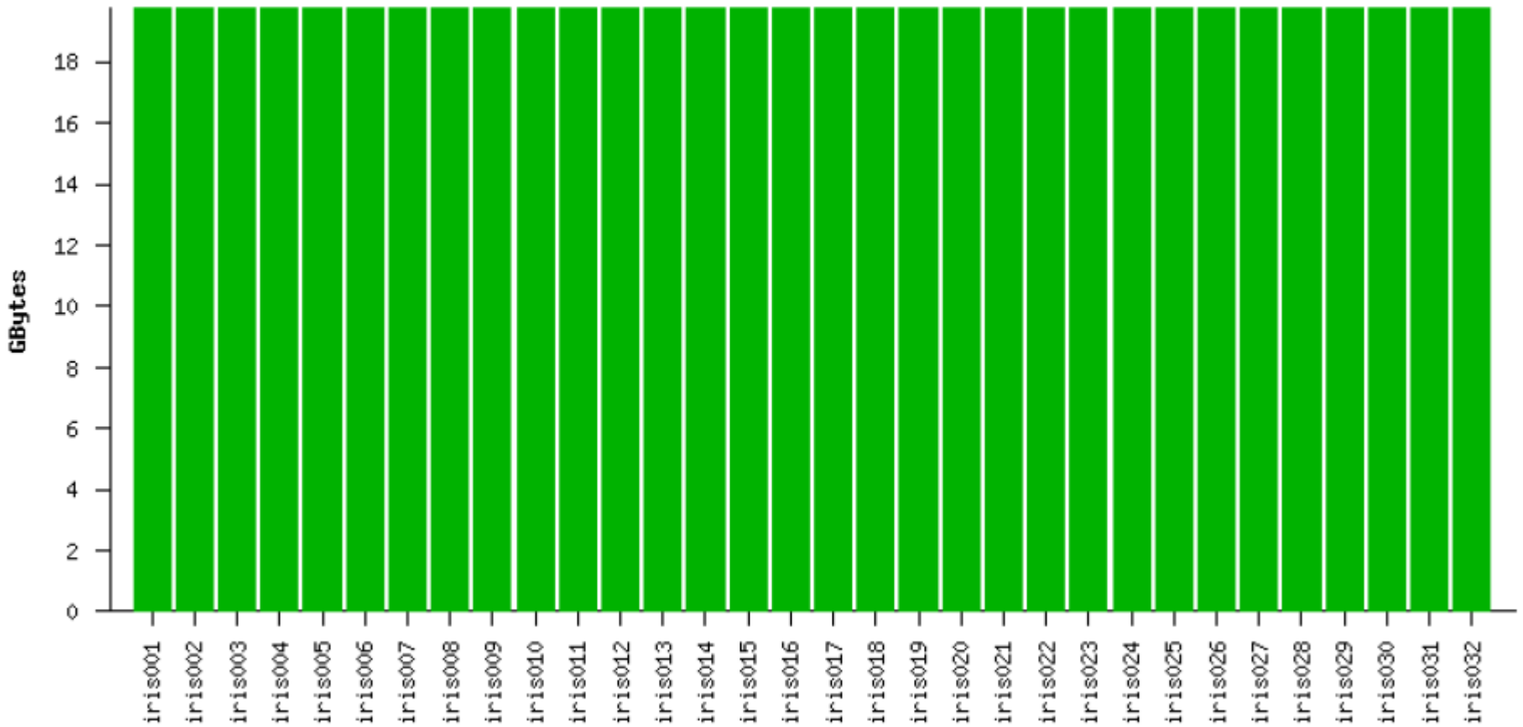
VASP MPI Profile, 32 Nodes Iris Cluster

- **Socket Imbalance**



VASP MPI Profile, 32 Nodes Iris Cluster

- **Memory footprint**



Summary

- **VASP performance**
  - 24% difference between Helios (Intel Gold 6138 2GHz) and Iris (intel 6148 Gold 2.4Ghz) clusters, 32 nodes
  - 30% higher performance when with NSIM=16 versus NSIM=4 on Iris cluster
  - BIOS sub-NUMA disabled mode increases the performance by 3% on 32 nodes Iris cluster
  - InfiniBand adaptive routing increases VASP performance by 13% on 128 nodes on Frontera cluster (TACC)

- **VASP MPI Profile**
  - 35% of communication, mostly collective operations
  - Broadcast, Allreduce and Alltoall/Alltoallv are the major calls
  - 18GB of memory usage over 32 nodes
  - Allreduce and Alltoall imbalance between the sockets

HPC-AI ADVISORY COUNCIL
NETWORK OF EXPERTISE

Thank You