HPC·AI
ADVISORY COUNCIL
NETWORK OF EXPERTISE

BSMBench
Performance Benchmark and Profiling

August 2020

Note

- **The following research was performed under the HPC Advisory Council activities**
  - System: Texas Advanced Computing Center (TACC)

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

- **More info on BSMBench**
  - https://gitlab.com/edbennett/BSMBench

BSMBench

- **Open source HPC benchmarking tool**
- **Based on simulation code used for studying strong interactions in particle physics**
- **Includes the ability to tune the ratio of communication over computation**
- **Includes 3 examples that show the performance of the system**
  - Problem that is computationally dominated (marked as Communications)
  - Problem that is communication dominated (marked as Compute)
  - Problem in which communication and computational requirements are balanced (marked as Balance)
- **Used to simulate workload such as Lattice Quantum ChromoDynamics (QCD), and by extension its parent field, Lattice Gauge Theory (LGT), which make up a significant fraction of supercomputing cycles worldwide**

InfiniBand Adaptive Routing

- InfiniBand defines variety of routing algorithm that can be configured by the Subnet Manager
- Adaptive routing is a switch ability to select the best and least busiest route for each network packet, in order to spread the network traffic and avoid point to point congestions
- The following analysis review the BSMBench performance utilizing InfiniBand adaptive routing

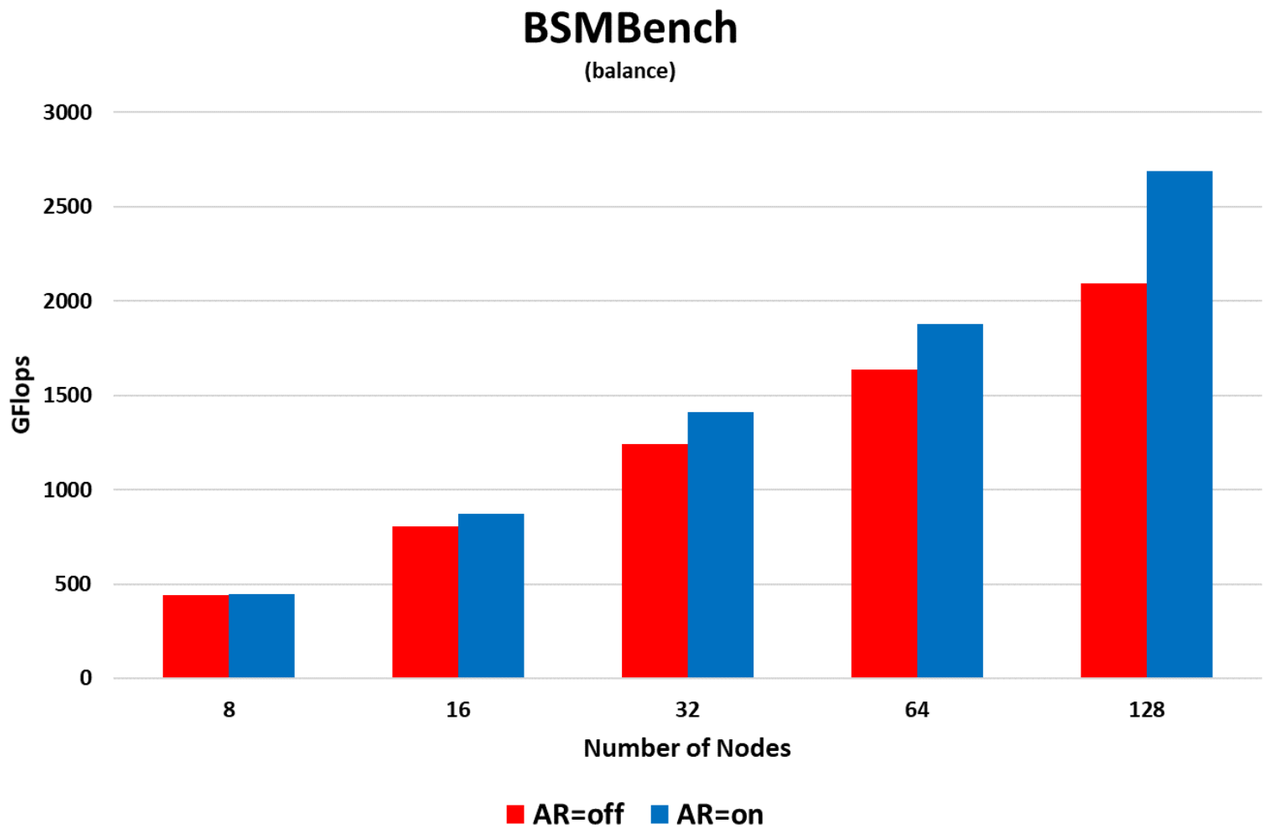
Cluster Configuration

- **Texas Advanced Computing Center (TACC) “Frontera” supercomputer**
  - Dual Socket Intel(R) Xeon(R) Platinum 8280 CPU @ 2.70GHz
  - Mellanox ConnectX-6 HDR100 InfiniBand
  - Mellanox Quantum Switch HDR InfiniBand
  - Memory: 192GB DDR4 2677MHz RDIMMs per node
  - Lustre Storage

- **Software**
  - OS: RHEL 7.8, MLNX_OFED 5.0.2
  - MPI: HPC-X 2.6.0
  - BSMBench : 1.1a
  - Compiler: Intel 2020.1.217

BSMBench Performance – Balance Benchmark

- **InfiniBand adaptive routing enables 28% higher performance at 128 nodes**

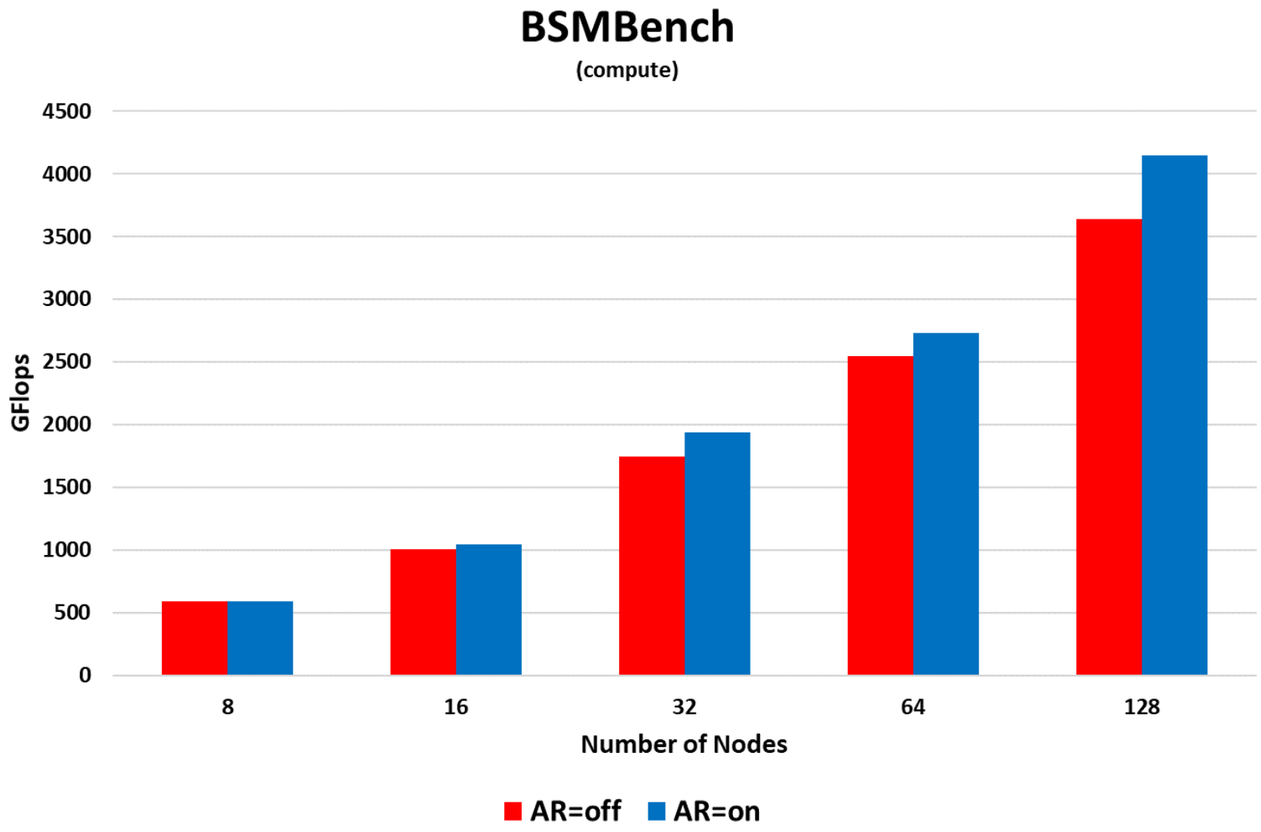


AR – Adaptive Routing

Higher is better

BSMBench Performance – Compute Benchmark

- **InfiniBand adaptive routing enables 14% higher performance at 128 nodes**

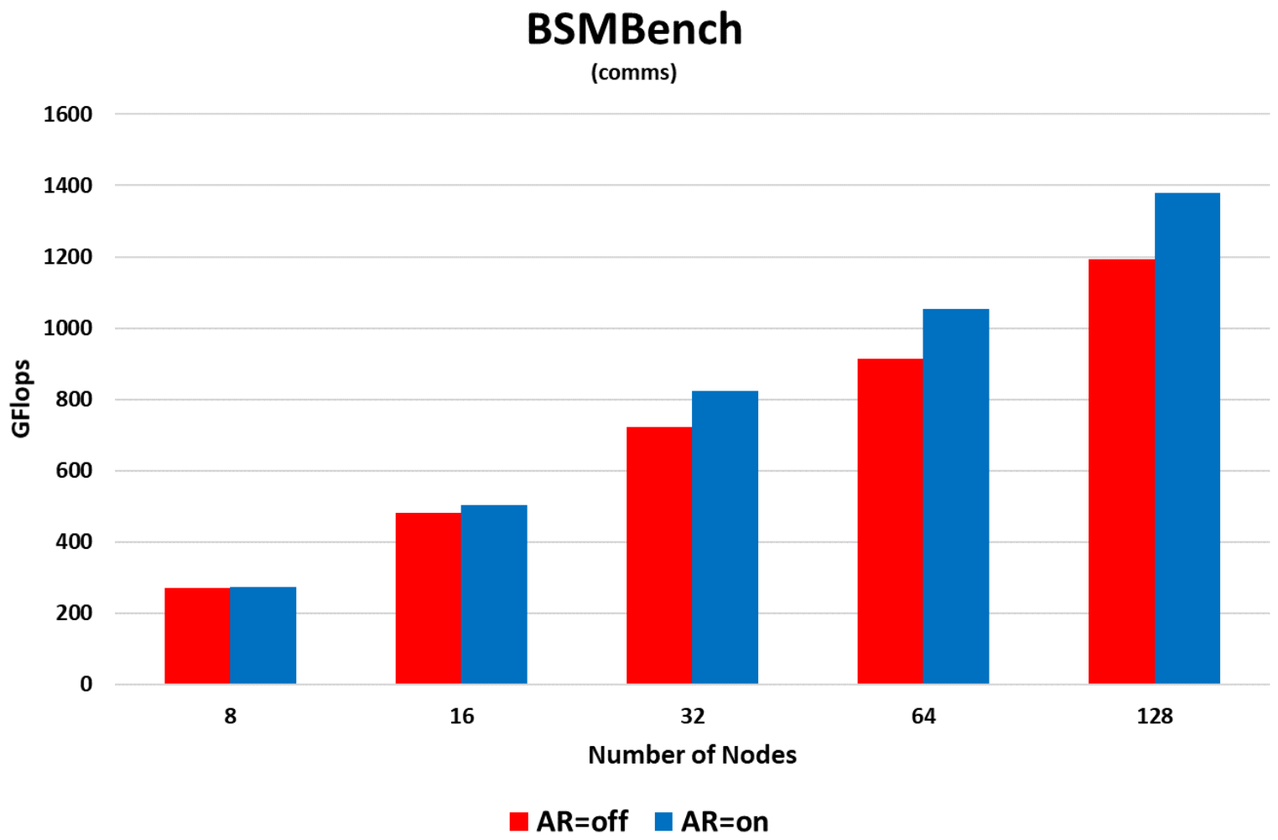


AR – Adaptive Routing

Higher is better

BSMBench Performance – Comms Benchmark

- **InfiniBand adaptive routing enables 16% higher performance at 128 nodes**



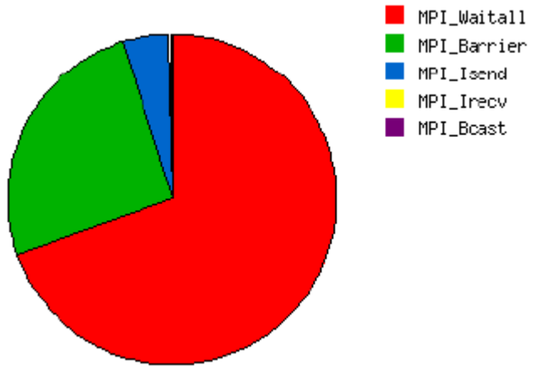
AR – Adaptive Routing

Higher is better

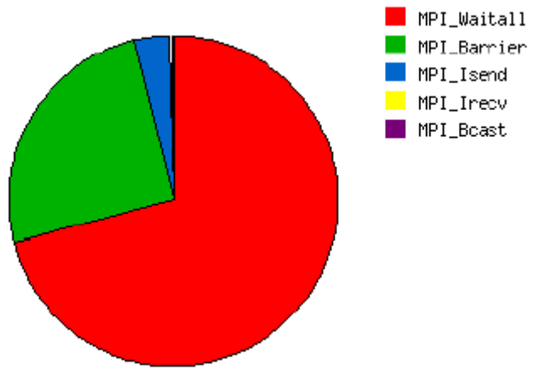
BSMBench MPI Profile

- 40-60% of the application run time is spend in MPI Communication time
- From the MPI communication time, ~70% is MPI WaitAll
- From the MPI communication time, ~25% of MPI Barrier

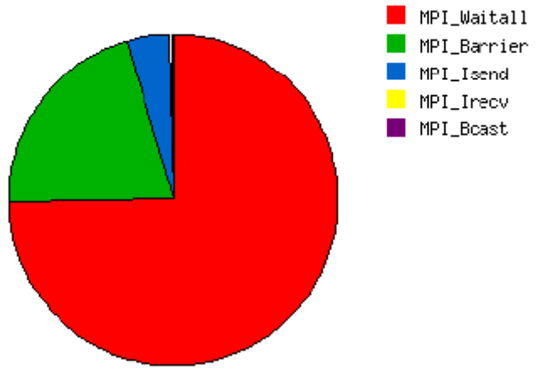
balance



compute



comms



32 nodes, with adaptive routing

BSMBench MPI Profile Details

Balance

Communication Event Statistics (% detail, --- error)									
	Comm Size	Buffer Size	Ncalls	Total Time	Avg Time	Min Time	Max Time	%MPI	%Wall
MPI_Waitall	0	0	1174369280	5.335985e+05	4.543703e-04	0.000000e+00	1.543000e-01	69.64	40.10
MPI_Barrier	1792	0	587188224	1.953096e+05	3.326184e-04	4.053100e-06	2.193200e-01	25.49	14.68
MPI_Isend	0	6144	6710681600	1.325725e+04	1.975544e-06	0.000000e+00	3.861000e-03	1.73	1.00
MPI_Isend	0	28672	1342136320	6.485218e+03	4.832011e-06	0.000000e+00	2.105300e-02	0.85	0.49
MPI_Isend	0	7168	2684272640	4.644740e+03	1.730353e-06	0.000000e+00	3.454000e-03	0.61	0.35

Comms

Communication Event Statistics (% detail, --- error)									
	Comm Size	Buffer Size	Ncalls	Total Time	Avg Time	Min Time	Max Time	%MPI	%Wall
MPI_Waitall	0	0	1174369280	5.579825e+05	4.751337e-04	0.000000e+00	1.087600e-01	74.76	47.10
MPI_Barrier	1792	0	587188224	1.545311e+05	2.631713e-04	4.768400e-06	1.866600e-01	20.70	13.04
MPI_Isend	0	6144	6710681600	1.233387e+04	1.837946e-06	0.000000e+00	2.077500e-02	1.65	1.04
MPI_Isend	0	28672	1342136320	5.617285e+03	4.185331e-06	0.000000e+00	3.411100e-03	0.75	0.47
MPI_Isend	0	7168	2684272640	4.219894e+03	1.572081e-06	0.000000e+00	4.331100e-03	0.57	0.36

Compute

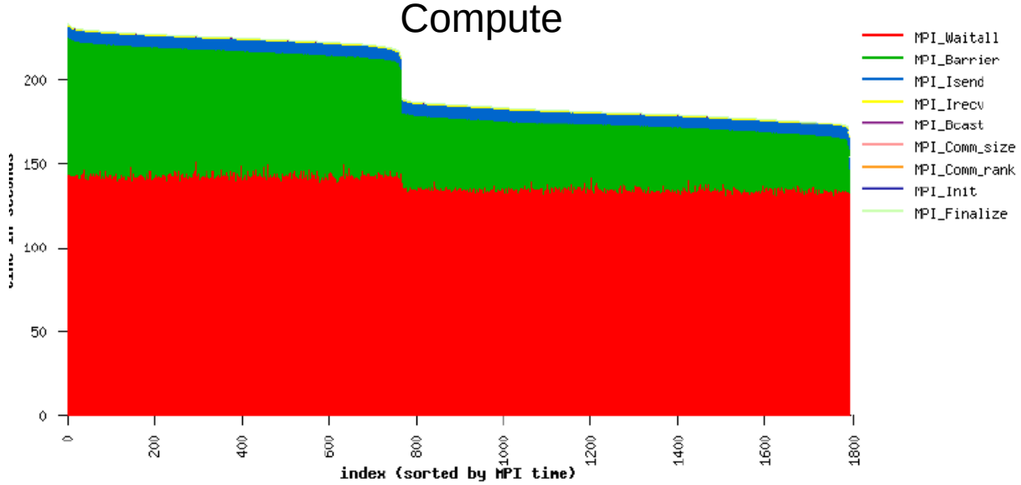
Communication Event Statistics (% detail, --- error)									
	Comm Size	Buffer Size	Ncalls	Total Time	Avg Time	Min Time	Max Time	%MPI	%Wall
MPI_Waitall	0	0	293565440	2.532195e+05	8.625658e-04	0.000000e+00	1.706200e-01	70.82	31.14
MPI_Barrier	1792	0	146786304	9.104175e+04	6.202332e-04	6.914100e-06	2.578100e-01	25.46	11.20
MPI_Isend	0	57344	335503360	3.285262e+03	9.792038e-06	0.000000e+00	3.661200e-03	0.92	0.40
MPI_Isend	0	12288	1677516800	2.910765e+03	1.735163e-06	0.000000e+00	5.701100e-03	0.81	0.36
MPI_Isend	0	49152	251627520	2.278789e+03	9.056201e-06	0.000000e+00	3.616100e-03	0.64	0.28

32 nodes, with adaptive routing

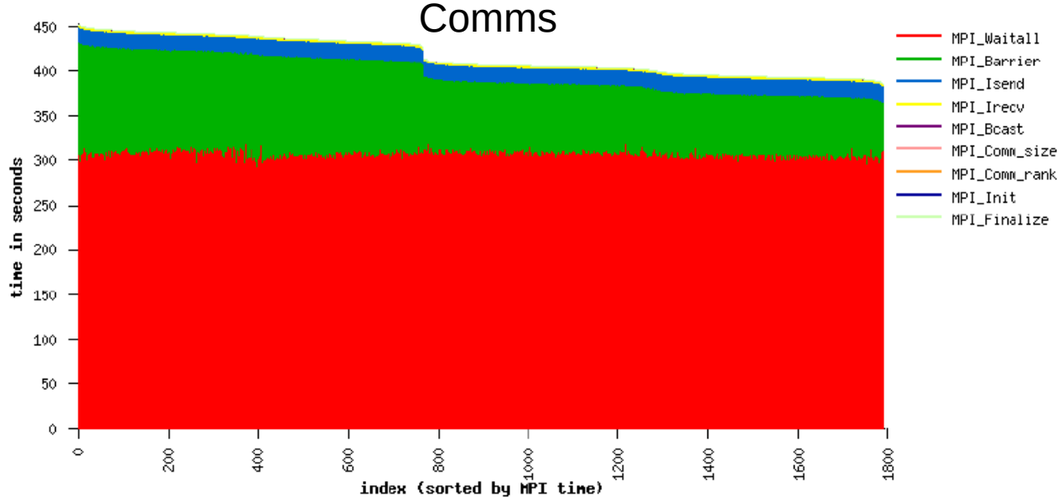
BSMBench MPI Profile Time Based

- **~20% is imbalance**

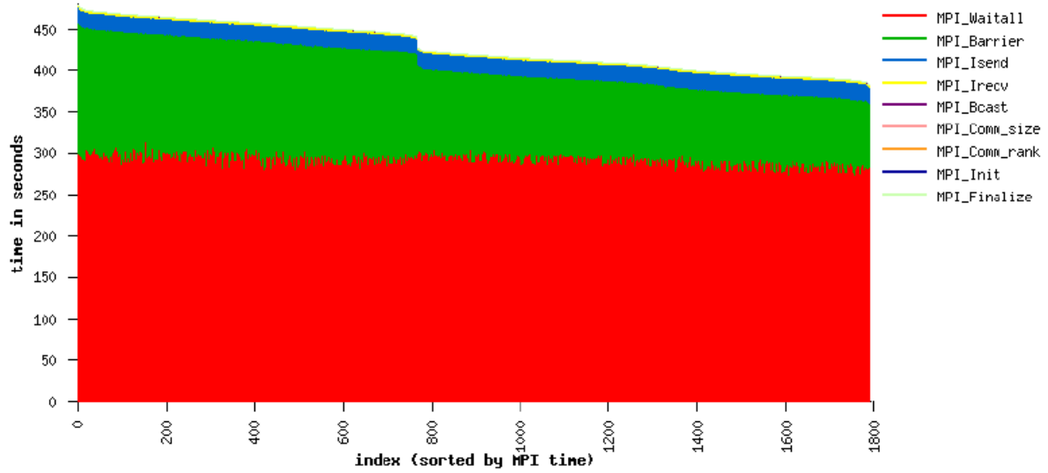
Compute



Comms



Balance

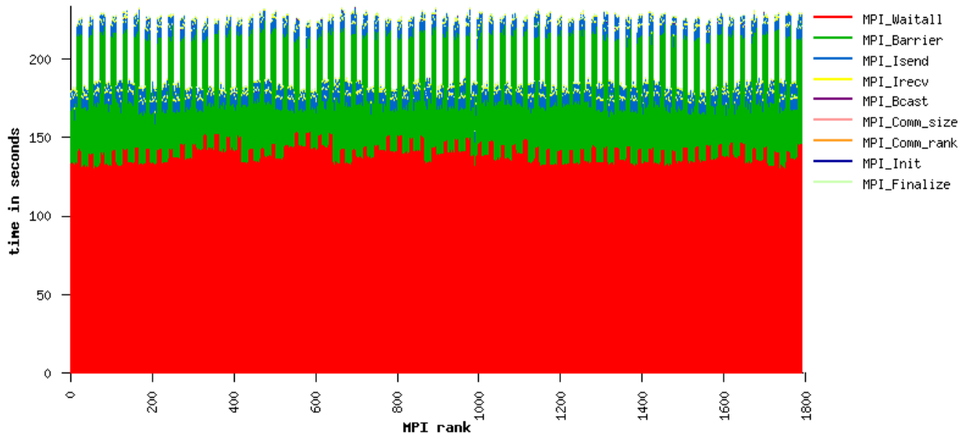


32 nodes, with adaptive routing

BSMBench Profile Rank Based

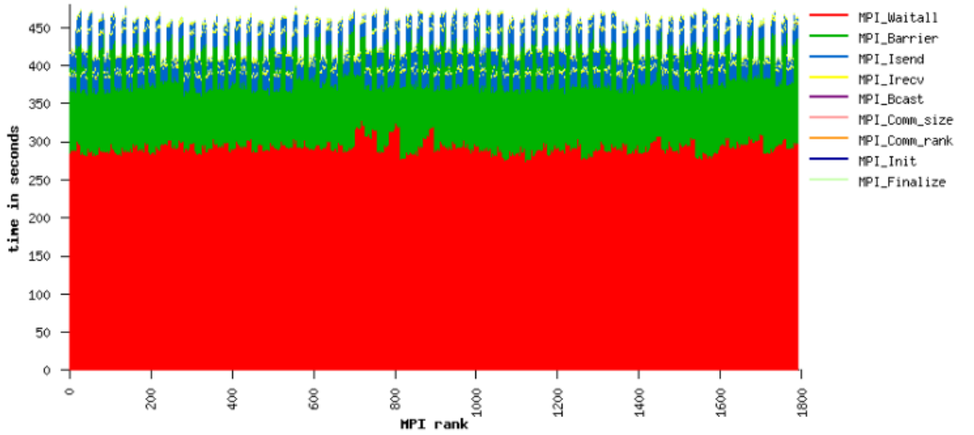
- **MPI Barrier imbalance between the sockets**

Compute



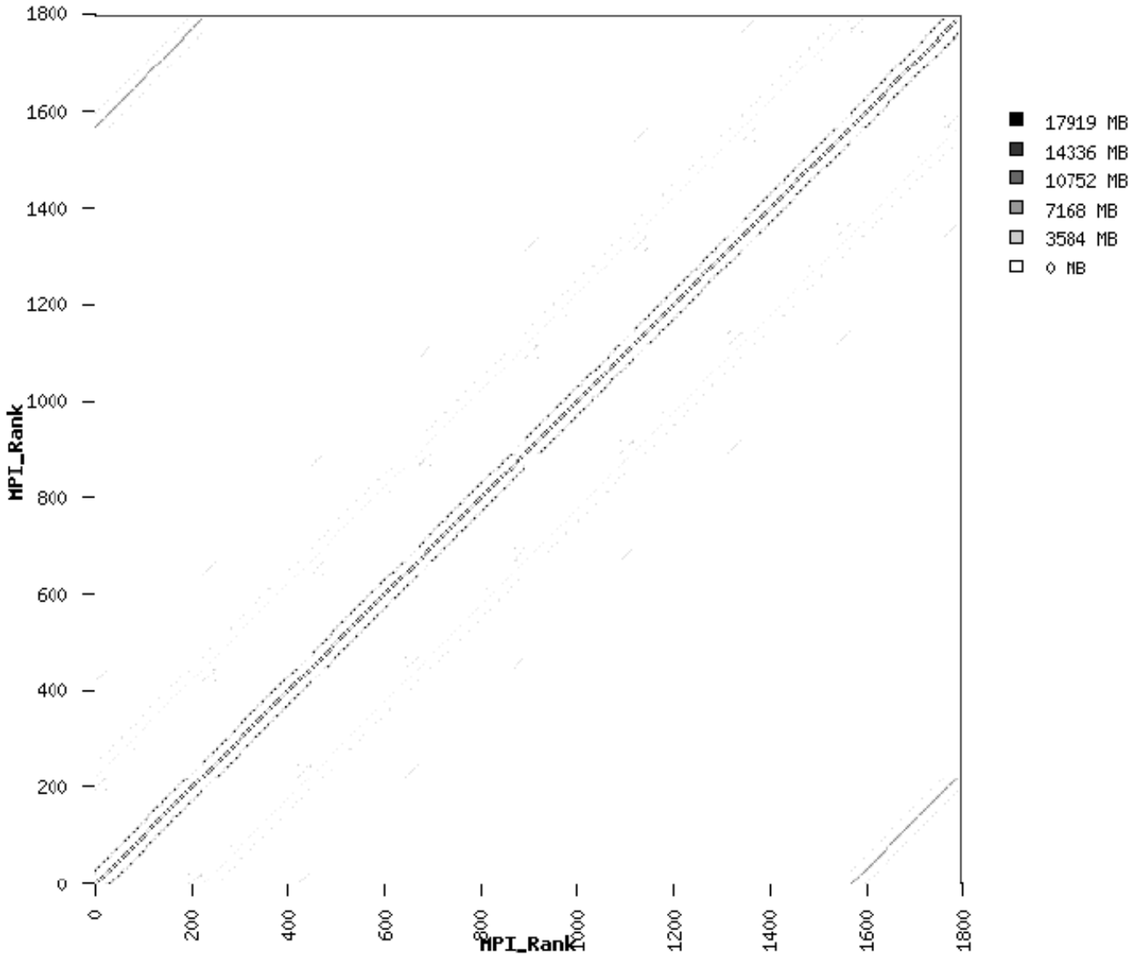
Comms

Balance



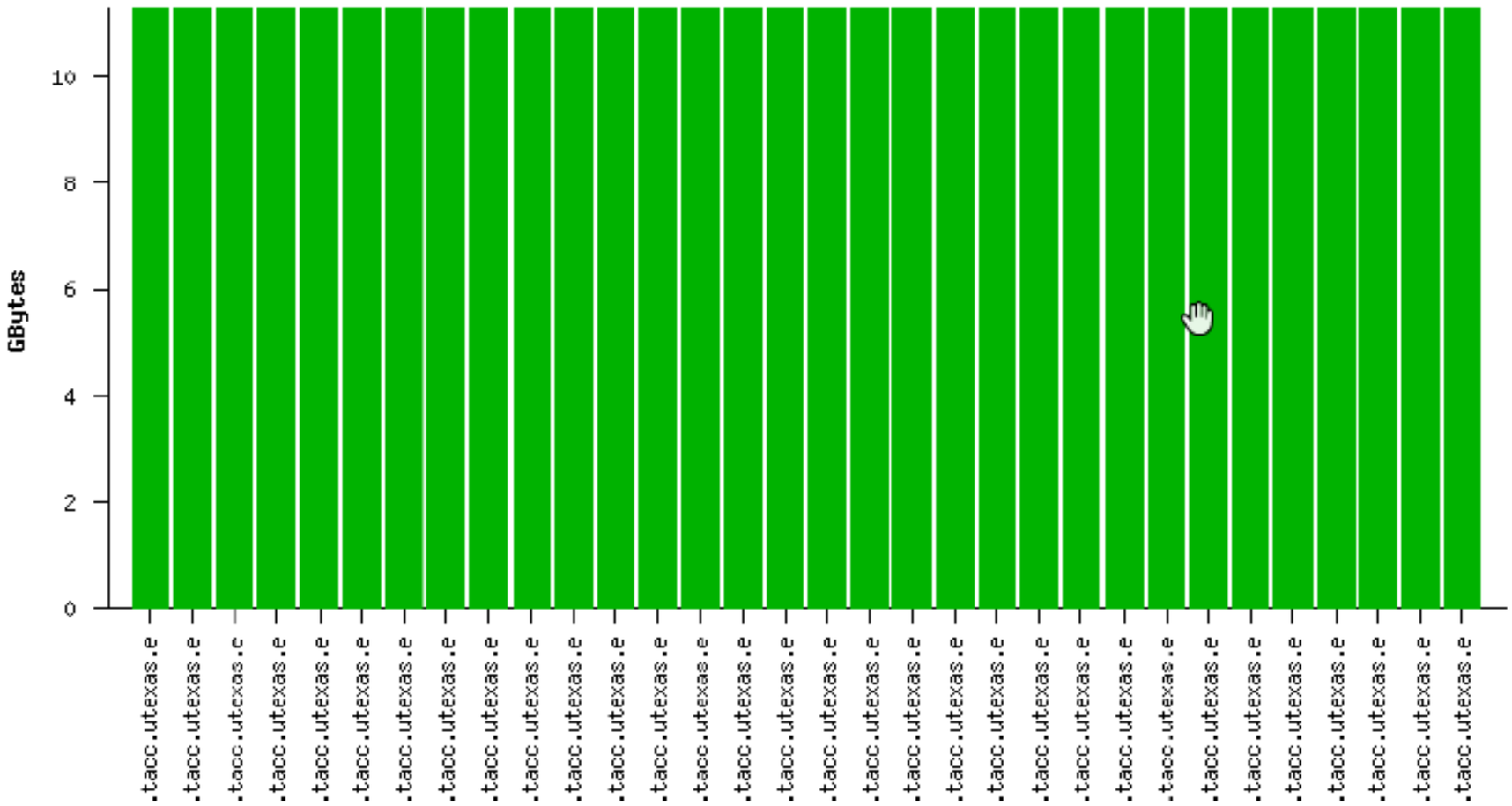
32 nodes, with adaptive routing

BSMBench Profile – Communication Matrix



32 nodes, with adaptive routing

BSMBench Profile – Memory Footprint



32 nodes, with adaptive routing

Summary

- **By using the InfiniBand adaptive routing, up to 28% higher performance was achieved**
  - Testing with 128 nodes

- **BSMBench MPI Profiling presented**
  - Collective operations
  - Barrier, WaitAll and Isend were the main MPI calls
  - MPI Barrier imbalance between the sockets
  - Communication pattern mostly to neighbor ranks
  - 12GB of memory usage over 32 nodes

Thank You