HPC·AI
ADVISORY COUNCIL
NETWORK OF EXPERTISE

MILC
Performance Benchmark and Profiling

June 2020

Note

- **The following research was performed under the HPC Advisory Council activities**
  - Compute resource - HPC Advisory Council Cluster Center

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

- **More info on MILC**
  - https://github.com/milc-qcd/milc_qcd

MIMD Lattice Computation (MILC)

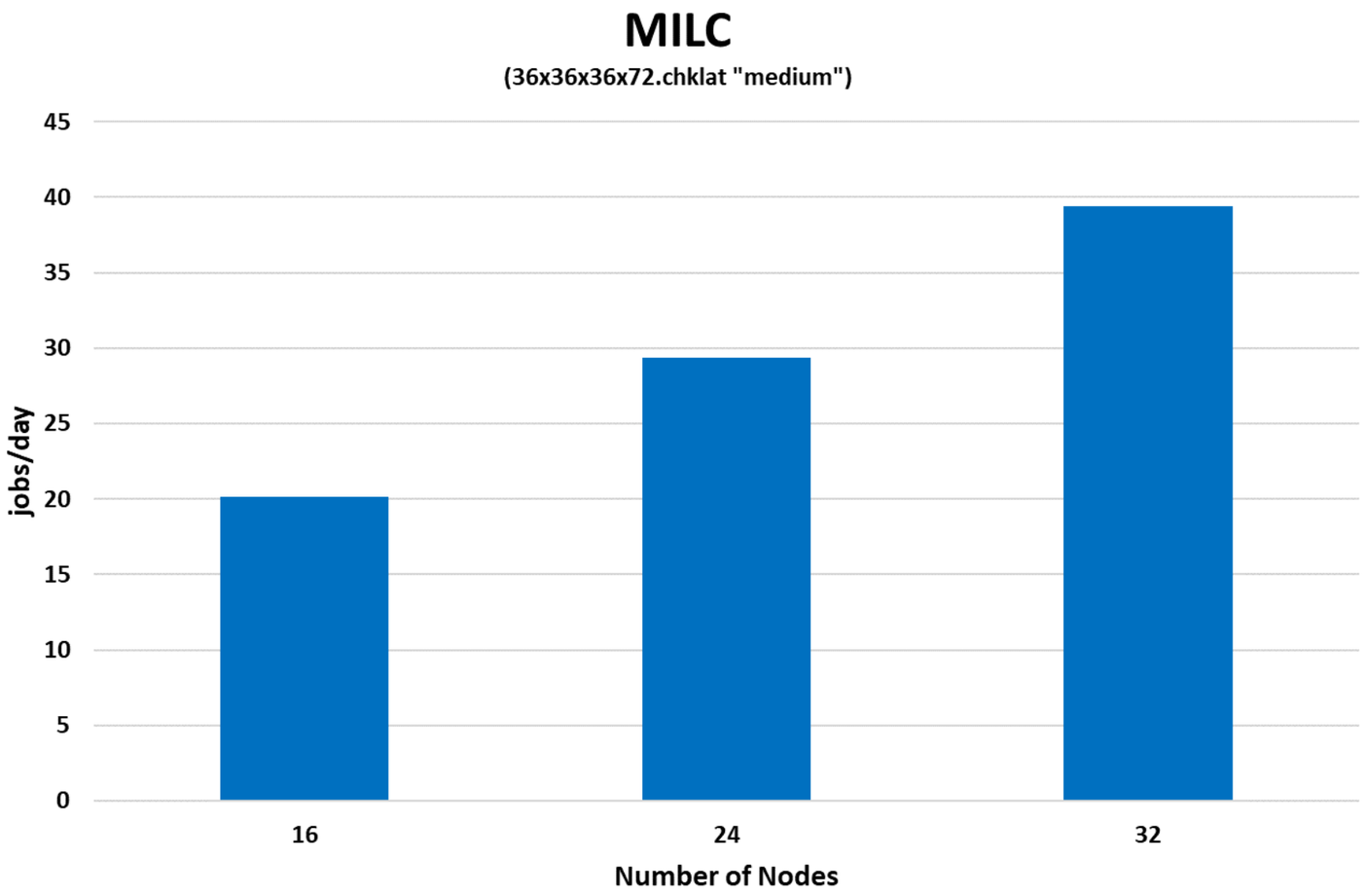
- The MIMD Lattice Computation (MILC) represents part of a set of codes used to study quantum chromodynamics (QCD), the theory of the strong interactions of subatomic physics
- It performs simulations of four dimensional SU(3) lattice gauge theory on MIMD parallel machines
- "Strong interactions" are responsible for binding quarks into protons and neutrons and holding them all together in the atomic nucleus
- The MILC collaboration has produced application codes to study several different QCD research areas

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 HDR InfiniBand
  - Mellanox Quantum Switch HDR InfiniBand
  - Memory: 192GB DDR4 2677MHz RDIMMs per node
  - Lustre Storage

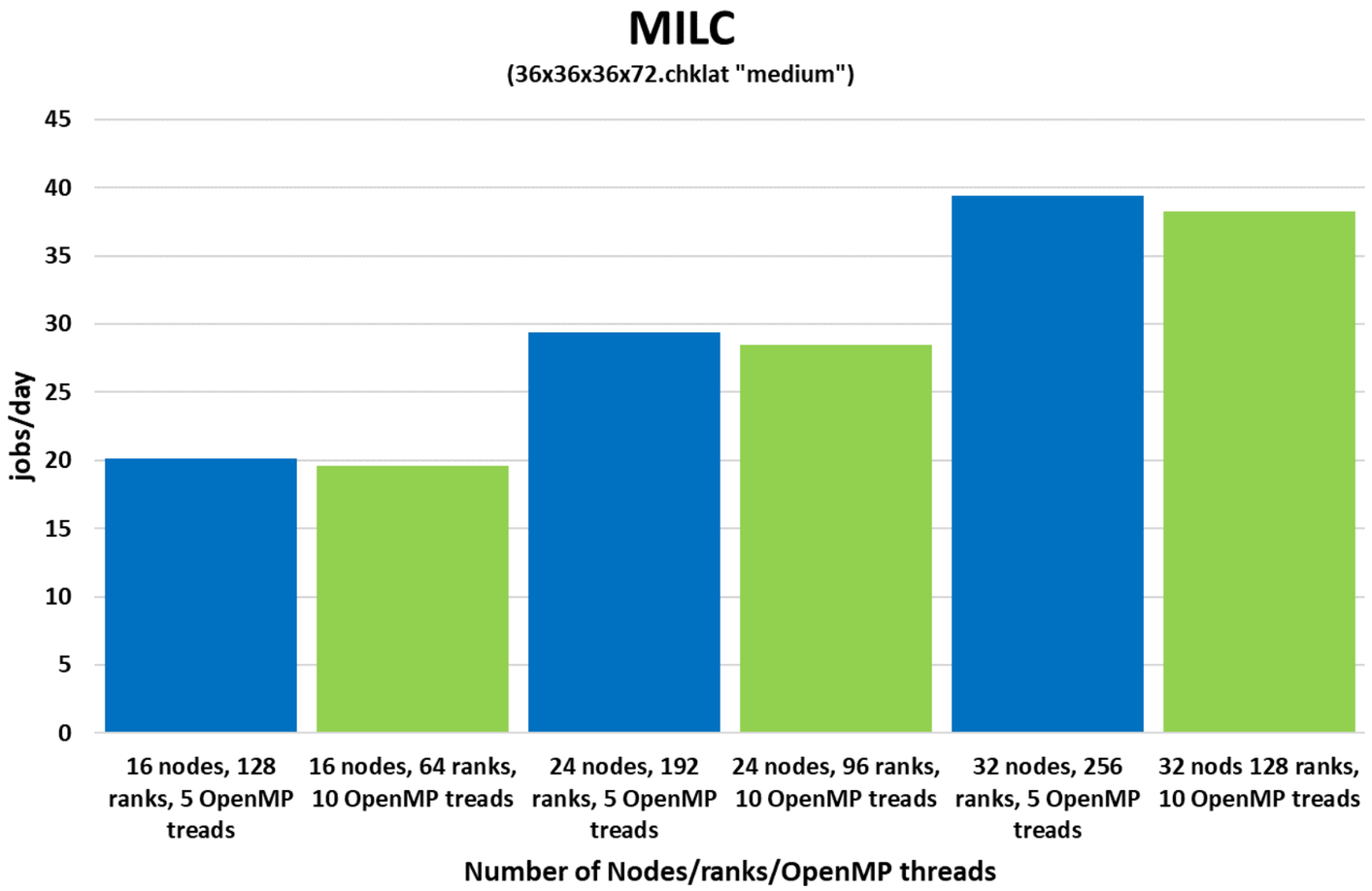
- **Software**
  - OS: RHEL 7.7, MLNX_OFED 4.7.3
  - MPI: HPC-X 2.6.0
  - MLIC: develop branch of https://github.com/milc-qcd/milc_qcd, commit 77d89f04bdc8fb55ebd40d555cb1f54c4b39d105 (May 29 1 2020)

MILC Performance - Scalability



Higher is better

MILC Performance – OpenMP Threads



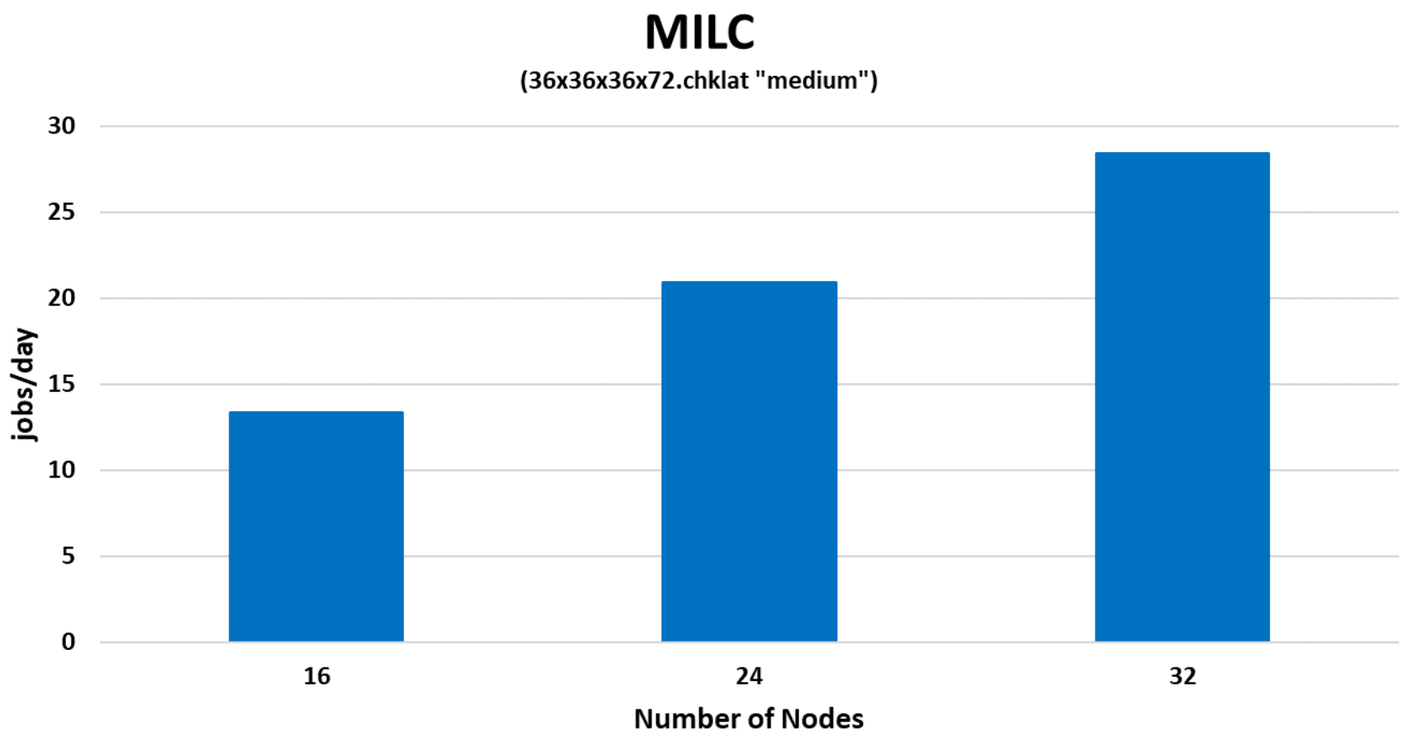
Higher is better

Cluster Configuration

- **Thor cluster**
  - Dell™ PowerEdge™ R730/R630 36-node cluster
  - Dual Socket Intel Xeon CPU E5-2697A v4 @ 2.60GHz
  - Mellanox ConnectX-6 HDR100 InfiniBand
  - Mellanox Quantum Switch HDR InfiniBand
  - Memory: 256GB DDR4 2400MHz RDIMMs per node
  - Lustre Storage

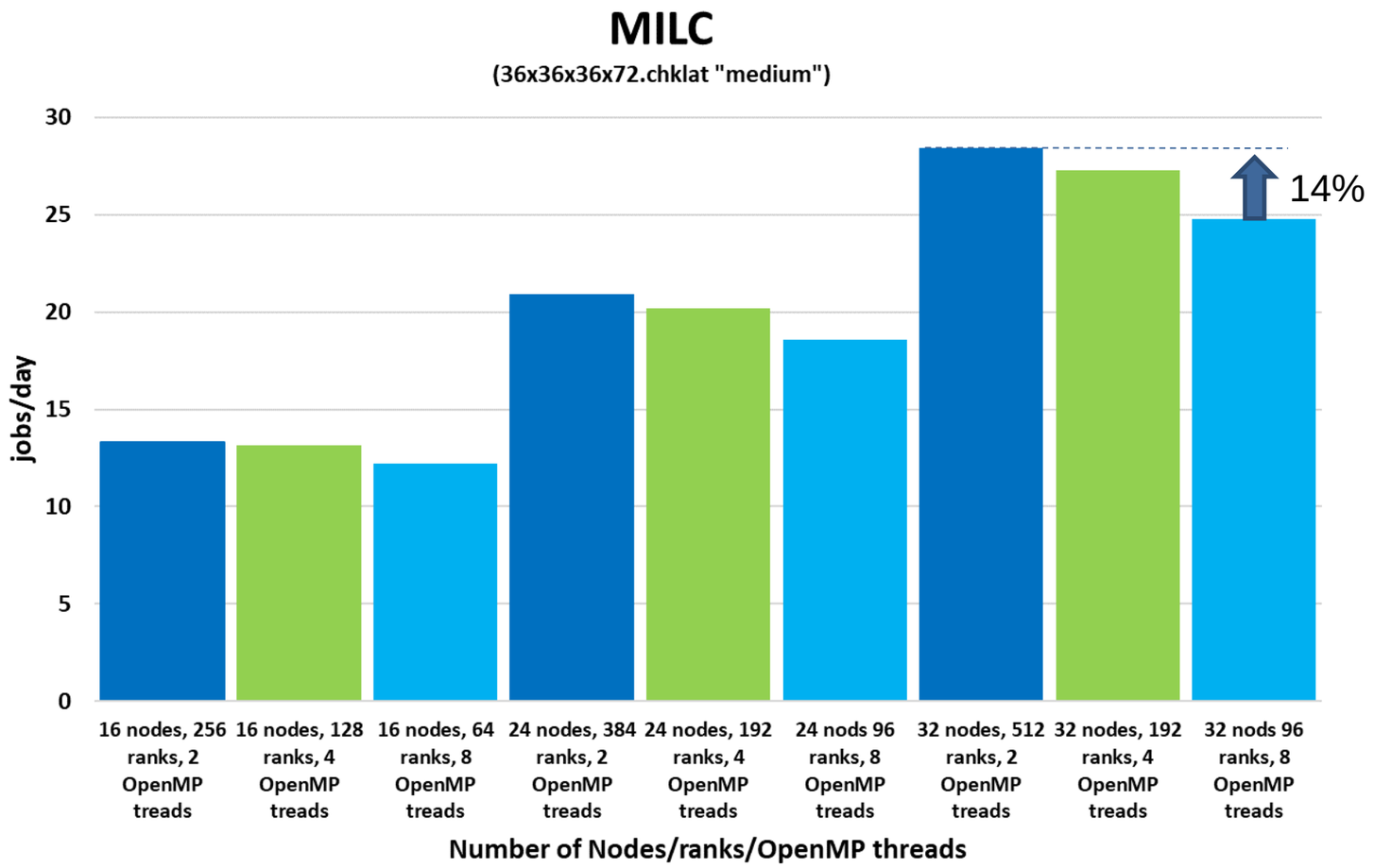
- **Software**
  - OS: RHEL 7.7, MLNX_OFED 4.7.3
  - MPI: HPC-X 2.6.0
  - MLIC: develop branch of https://github.com/milc-qcd/milc_qcd, commit 77d89f04bdc8fb55ebd40d555cb1f54c4b39d105 (May 29 1 2020)

MILC Performance - Scalability



Higher is better

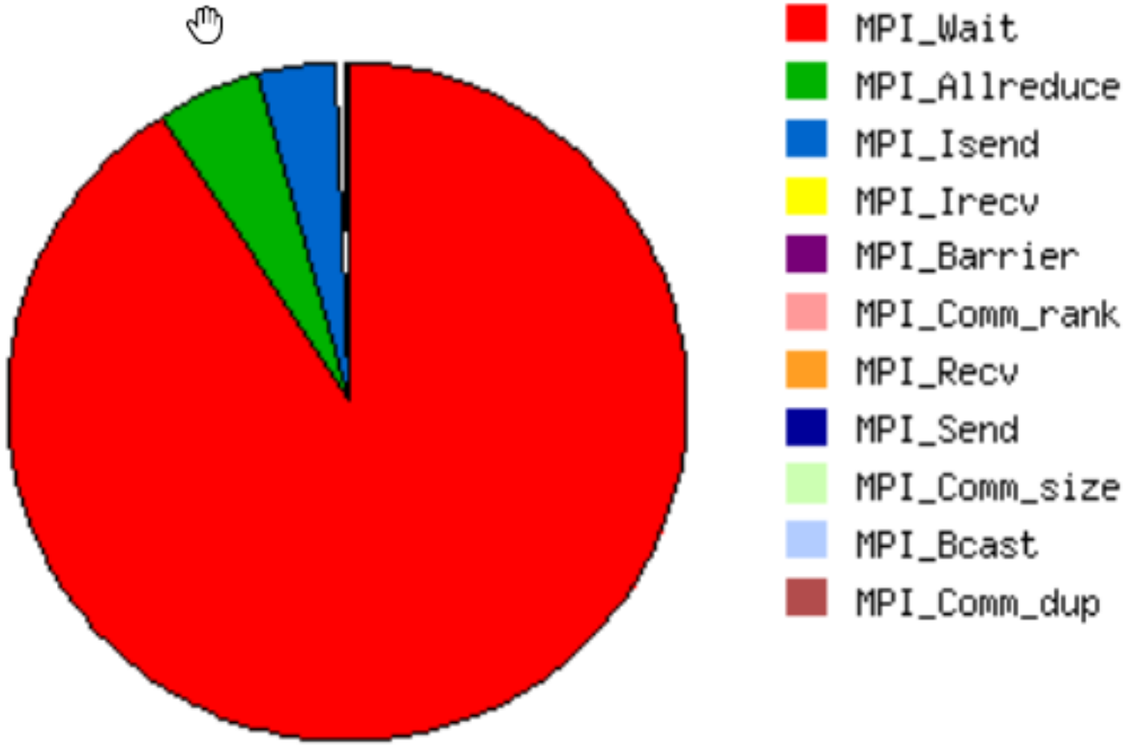
MILC Performance - OpenMP Threads



Higher is better

MILC Profile on 32 Nodes Helios Cluster

- **18% MPI Communication**



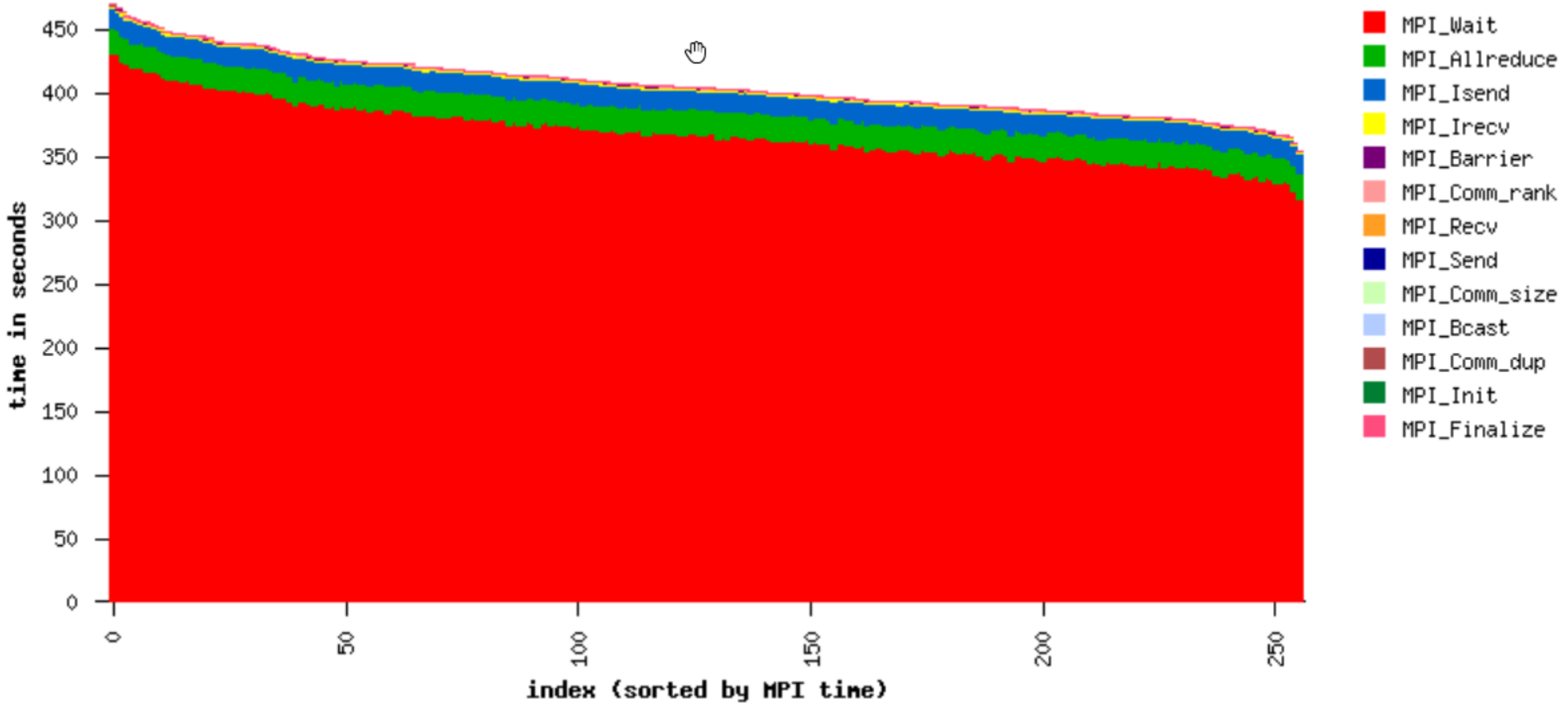
MILC Profile on 32 Nodes Helios Cluster

- 91% of MPI Communication spent on MPI_Wait
- 5% MPI Allreduce 8 bytes
- Async send and receive communication

Communication Event Statistics (% detail, --- error)									
	Comm Size	Buffer Size	Ncalls	Total Time	Avg Time	Min Time	Max Time	%MPI	%Wall
MPI_Wait	0	0	5577428992	9.441883e+04	1.692874e-05	0.000000e+00	2.310400e-02	90.83	16.59
MPI_Allreduce	256	8	317123072	5.073033e+03	1.599705e-05	2.861000e-06	1.188200e-02	4.88	0.89
MPI_Isend	0	98304	1965566208	2.652641e+03	1.349556e-06	0.000000e+00	7.525000e-03	2.55	0.47
MPI_Isend	0	49152	636610304	7.591547e+02	1.192495e-06	0.000000e+00	7.477000e-03	0.73	0.13
MPI_Isend	0	196608	167205120	4.598280e+02	2.750083e-06	0.000000e+00	1.429100e-03	0.44	0.08
MPI_Irecv	0	98304	1965566208	2.592905e+02	1.319165e-07	0.000000e+00	7.382900e-03	0.25	0.05
MPI_Irecv	0	196608	167205120	1.081722e+02	6.469433e-07	0.000000e+00	4.210900e-03	0.10	0.02
MPI_Barrier	256	0	840960	6.590245e+01	7.836574e-05	9.536700e-07	1.081100e-01	0.06	0.01
MPI_Irecv	0	49152	636610304	4.963680e+01	7.797046e-08	0.000000e+00	1.311800e-03	0.05	0.01

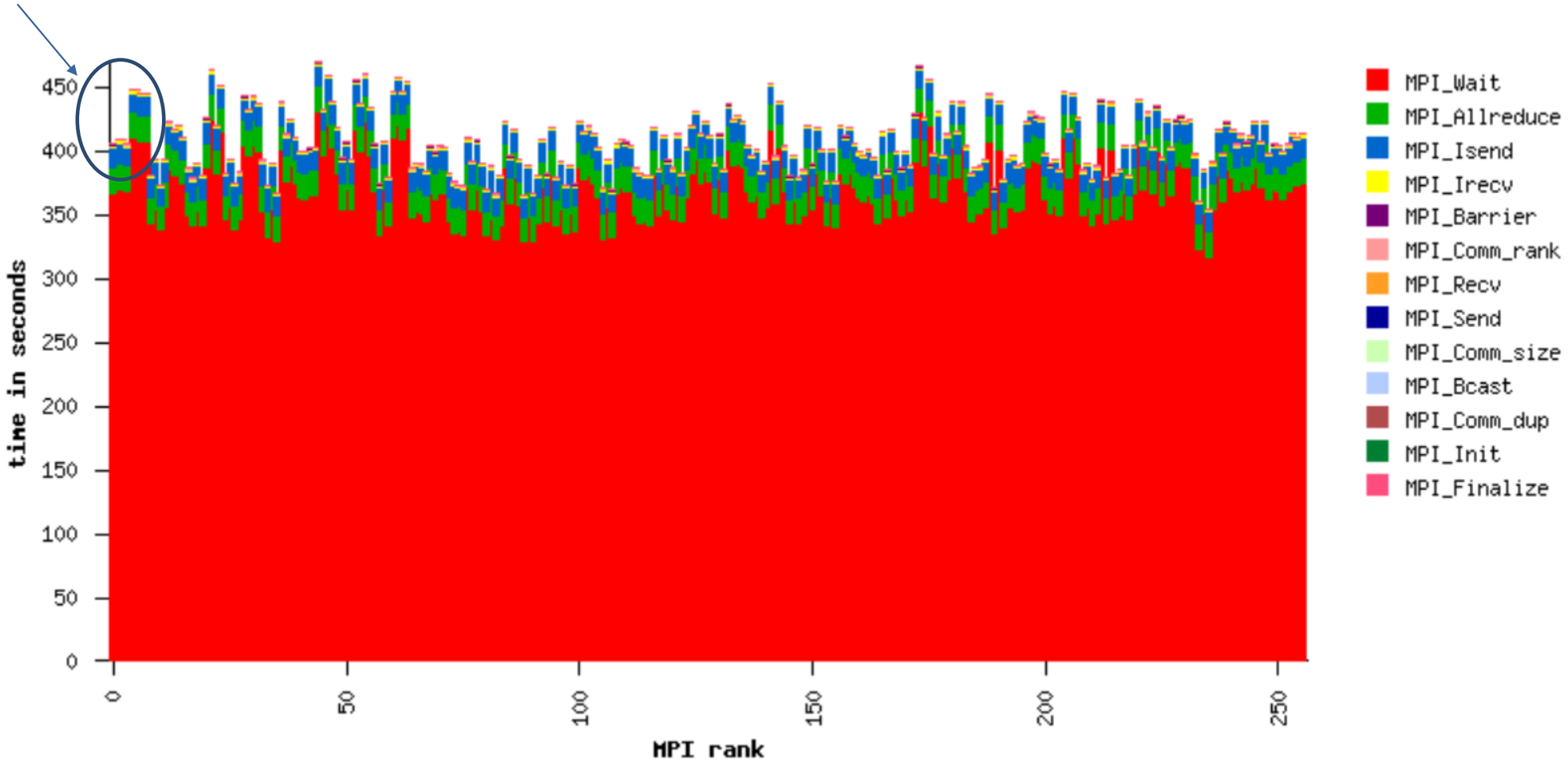
MILC Profile on 32 Nodes Helios Cluster

- **20% imbalance**



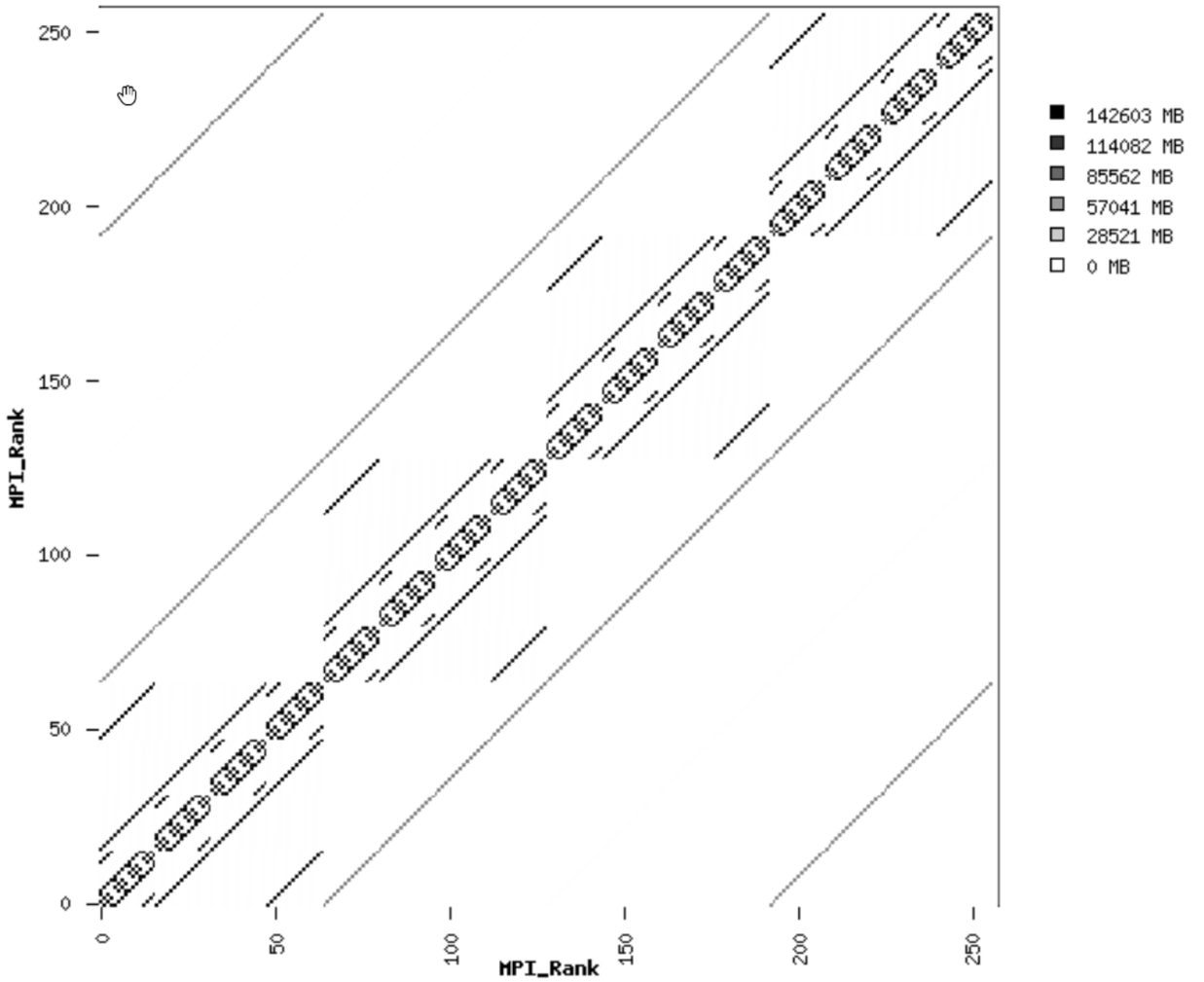
MILC Profile on 32 Nodes Helios Cluster

- **Imbalance between two sockets (4 processes per socket, 5 OpenMP threads)**



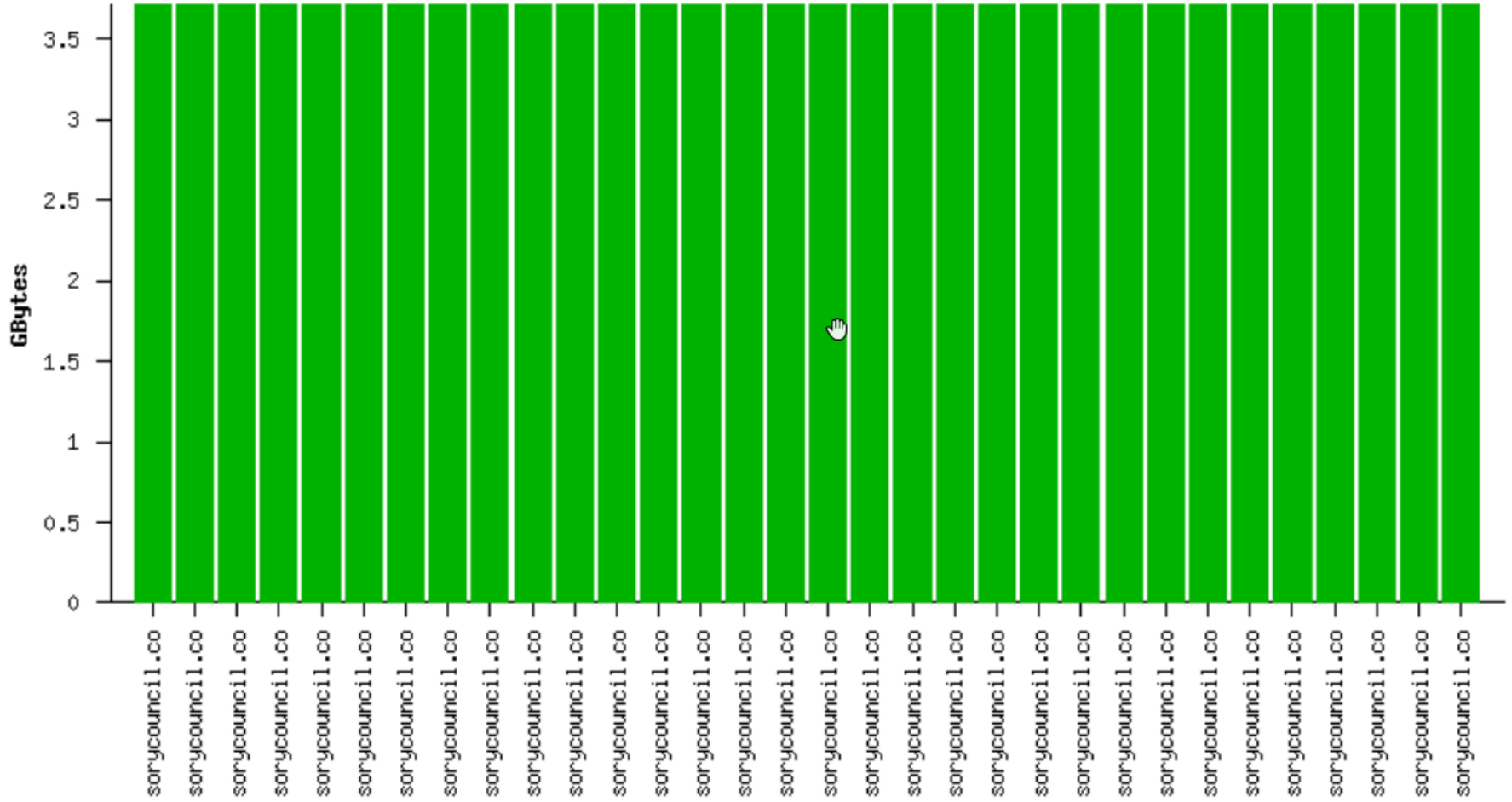
MILC Profile on 32 Nodes Helios Cluster

- **4D Communication Matrix**



MILC Profile on 32 Nodes Helios Cluster

- **Memory footprint**



Performance Analysis Summary

- **MILC performance testing over Intel based platform**
  - 95% scaling was achieved from 16 to 32 nodes for the medium benchmark on Helios Cluster (Skylake)
  - 3% difference when using 5 OpenMP threads comparing to 10 OpenMP threads on Helios Cluster (Skylake)
  - Super linear scaling was achieved from 16 to 32 nodes for the medium benchmark on Thor cluster (Broadwell)
  - 15% difference when using 2 OpenMP threads comparing to 8 OpenMP threads on Thor cluster (Broadwell)
  - Low number of OpenMP threads did better than high number of OpenMP threads

- **MILC Profile**
  - Async P2P communication with 8 byte MPI Allreduce collective
  - High percentage of MPI wait due to imbalance of the application
  - 4D communication matrix

HPC·AI ADVISORY COUNCIL
NETWORK OF EXPERTISE

Thank You