HPC·AI
ADVISORY COUNCIL
NETWORK OF EXPERTISE

RELION
Performance Benchmark and Profiling

April 2021

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**
  - RELION performance overview over Intel based platforms
  - Understanding RELION communication patterns

- **More info on RELION**
  - https://github.com/3dem/relion
  - https://www3.mrc-lmb.cam.ac.uk/relion/index.php/Main_Page

RELION

- **RELION (REgularized LIkelihood OptimizatioN) is an open-source program for the refinement of macromolecular structures by single-particle analysis of electron cryo-microscopy (cryo-EM) data**
- **RELION (REgularized LIkelihood OptimizatioN) implements an empirical Bayesian approach for analysis of electron cryo-microscopy (Cryo-EM)**
- **RELION provides refinement methods of singular or multiple 3D reconstructions as well as 2D class averages**
- **RELION is an important tool in the study of living cells**

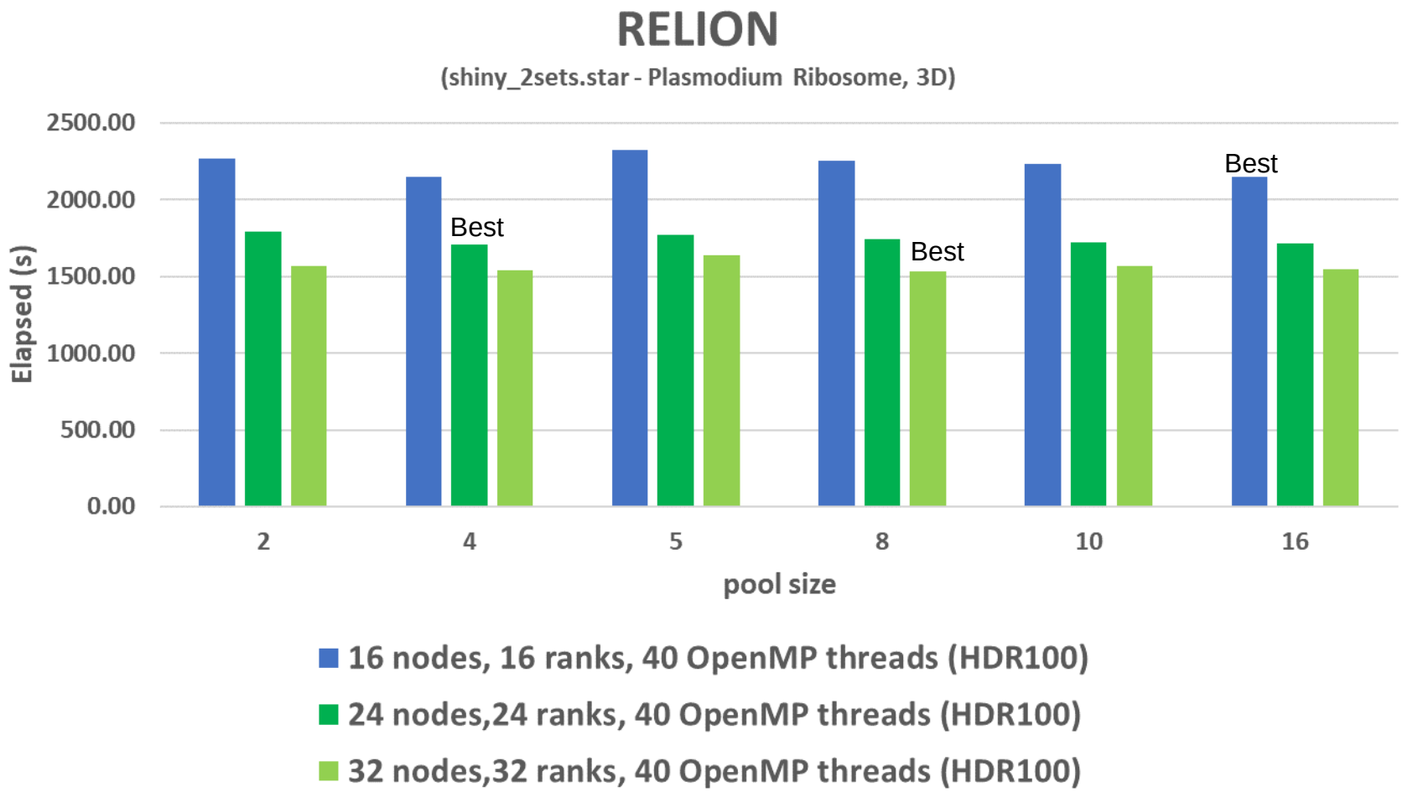
Cluster Configuration

- **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 8.3, MLNX_OFED 5.2.1
  - MPI: HPC-X 2.7.0
  - Relion 3.1.0
  - Input: shiny_2sets.star (Plasmodium Ribosome, 3D)

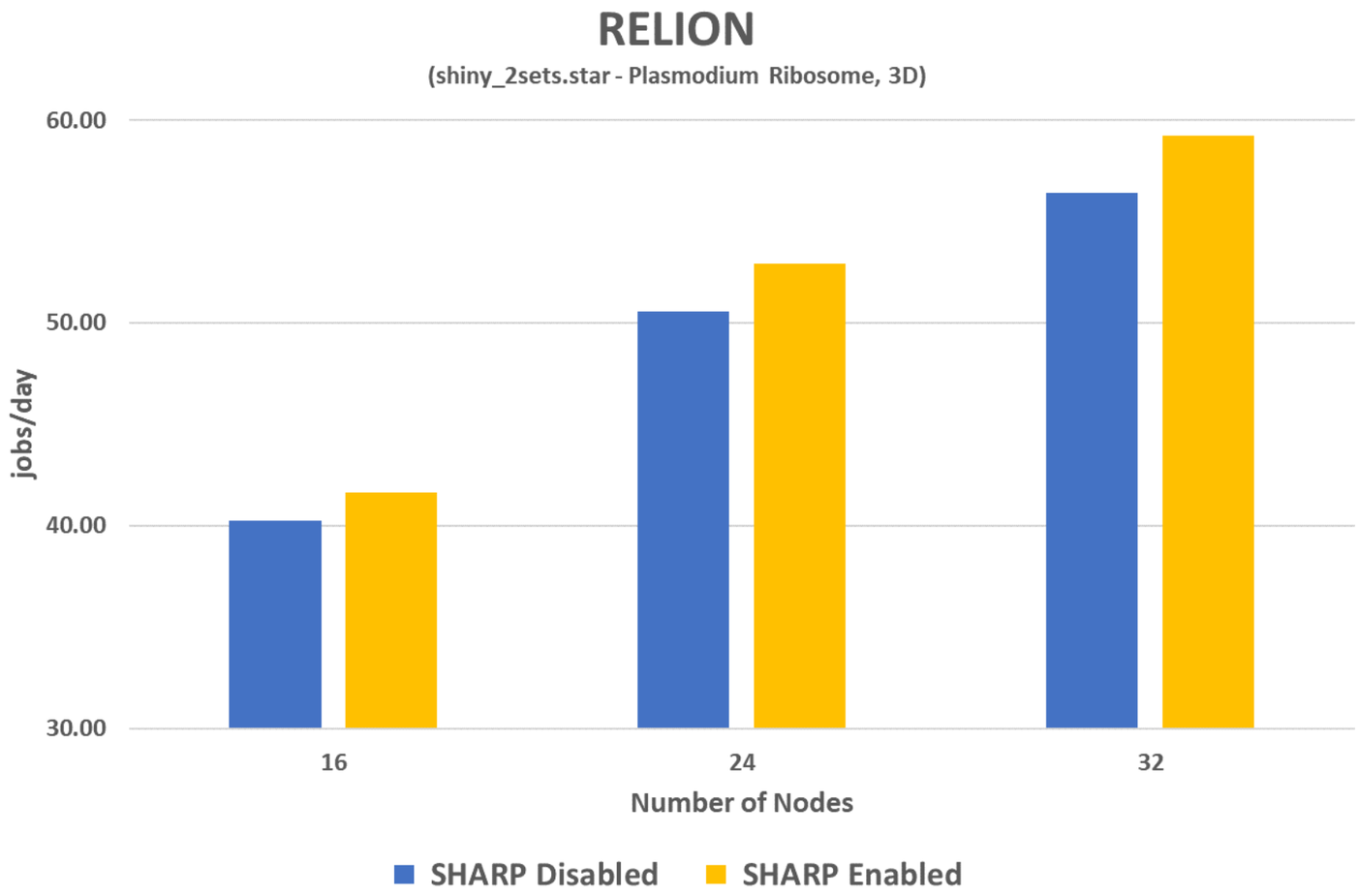
RELION Performance

- Different pool sizes per node count for highest performance



Lower is better

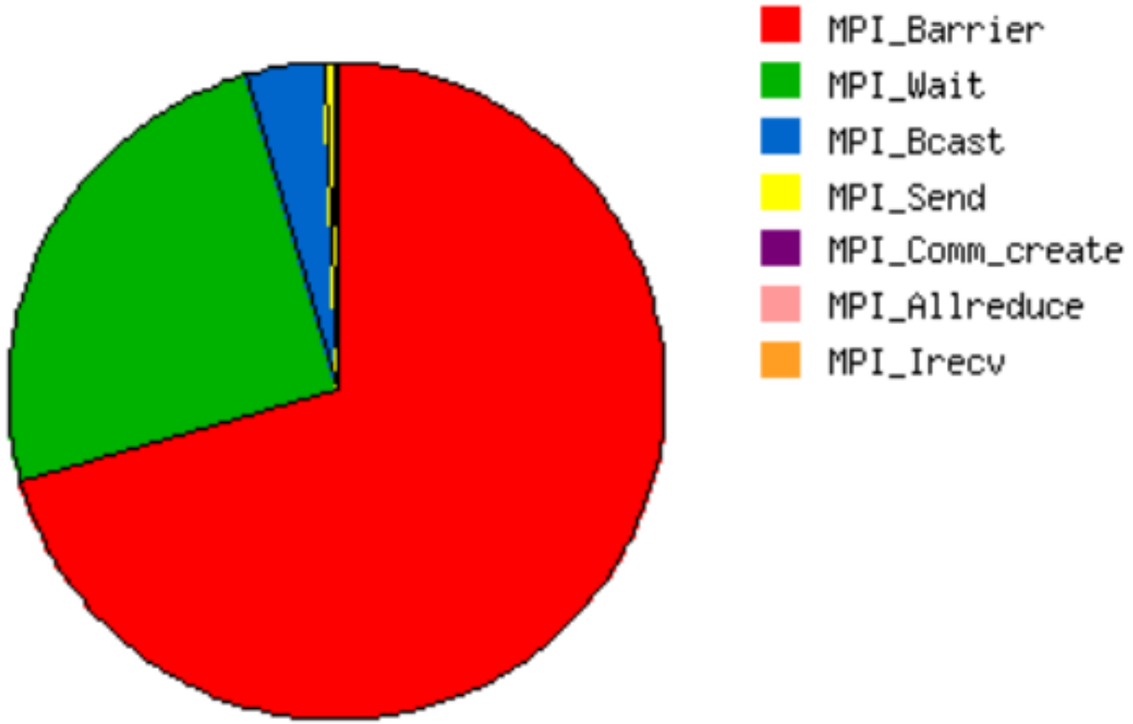
RELION Performance – SHARP In-Network Computing



Higher is better

RELION MPI Profiling

- **41% of the overtime is MPI Communication**



32 Nodes, 1 rank per node

RELION MPI Profiling

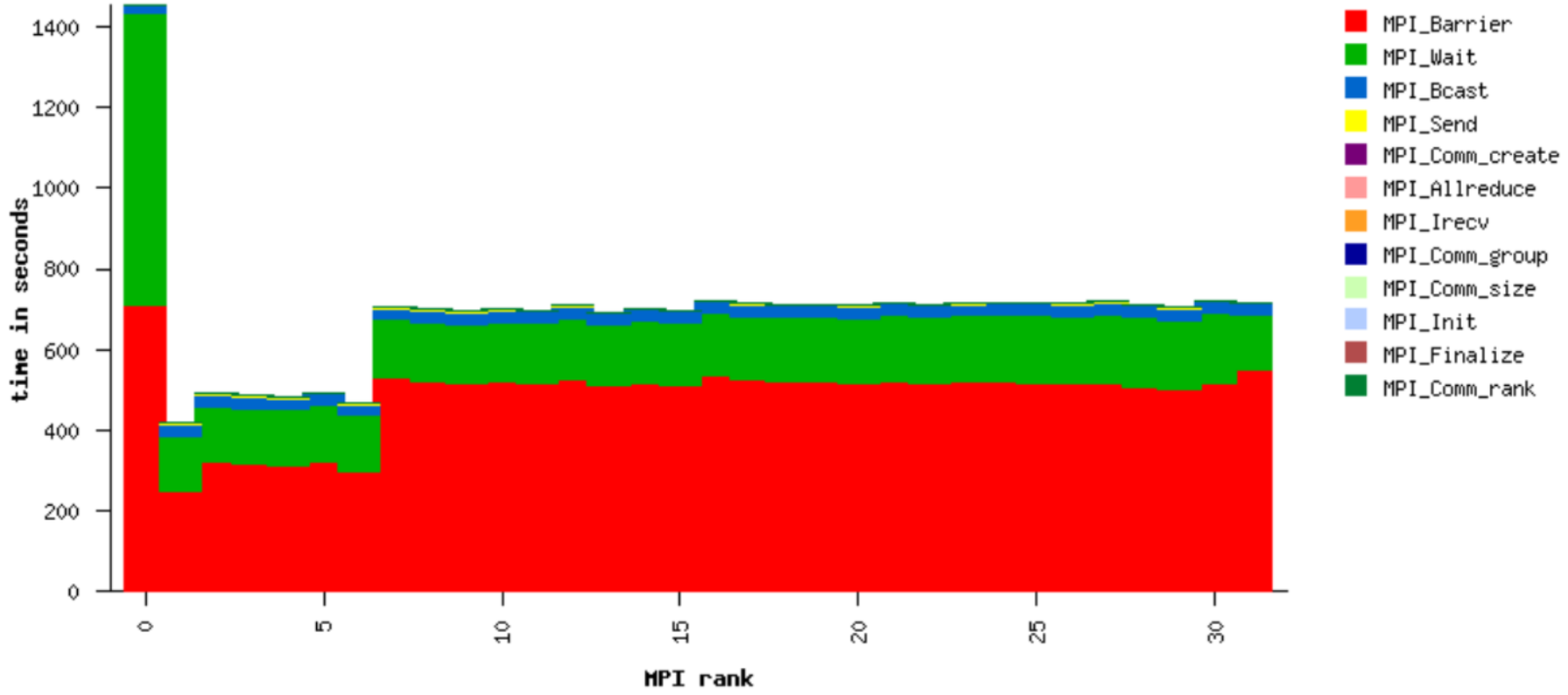
- **70% of MPI Communication spent on MPI_Barrier**
- **25% MPI_Wait**

Communication Event Statistics (% detail, --- error)									
	Comm Size	Buffer Size	Ncalls	Total Time	Avg Time	Min Time	Max Time	%MPI	%Wall
MPI_Barrier	32	0	10816	1.556102e+04	1.438704e+00	9.536700e-07	1.518000e+01	70.59	29.46
MPI_Wait	0	0	69957	5.494780e+03	7.854511e-02	0.000000e+00	9.962800e+00	24.93	10.40
MPI_Bcast	32	335544320	4800	7.583216e+02	1.579837e-01	8.930000e-02	3.533100e-01	3.44	1.44
MPI_Send	0	536870912	900	6.514308e+01	7.238120e-02	4.050200e-02	3.048700e-01	0.30	0.12
MPI_Bcast	31	25165824	930	2.964804e+01	3.187961e-02	2.493300e-02	4.634200e-02	0.13	0.06
MPI_Bcast	31	16777216	558	2.097387e+01	3.758758e-02	2.145100e-02	7.531400e-02	0.10	0.04
MPI_Bcast	31	33554432	558	1.867219e+01	3.346271e-02	9.944000e-03	9.840500e-02	0.08	0.04
MPI_Send	0	402653184	240	1.120947e+01	4.670613e-02	3.101500e-02	1.367500e-01	0.05	0.02
MPI_Bcast	31	8388608	372	1.113004e+01	2.991945e-02	1.970200e-02	4.617700e-02	0.05	0.02
MPI_Bcast	31	58720256	372	1.004068e+01	2.699108e-02	1.626000e-02	6.067600e-02	0.05	0.02

32 Nodes, 1 rank per node

RELION MPI Profiling

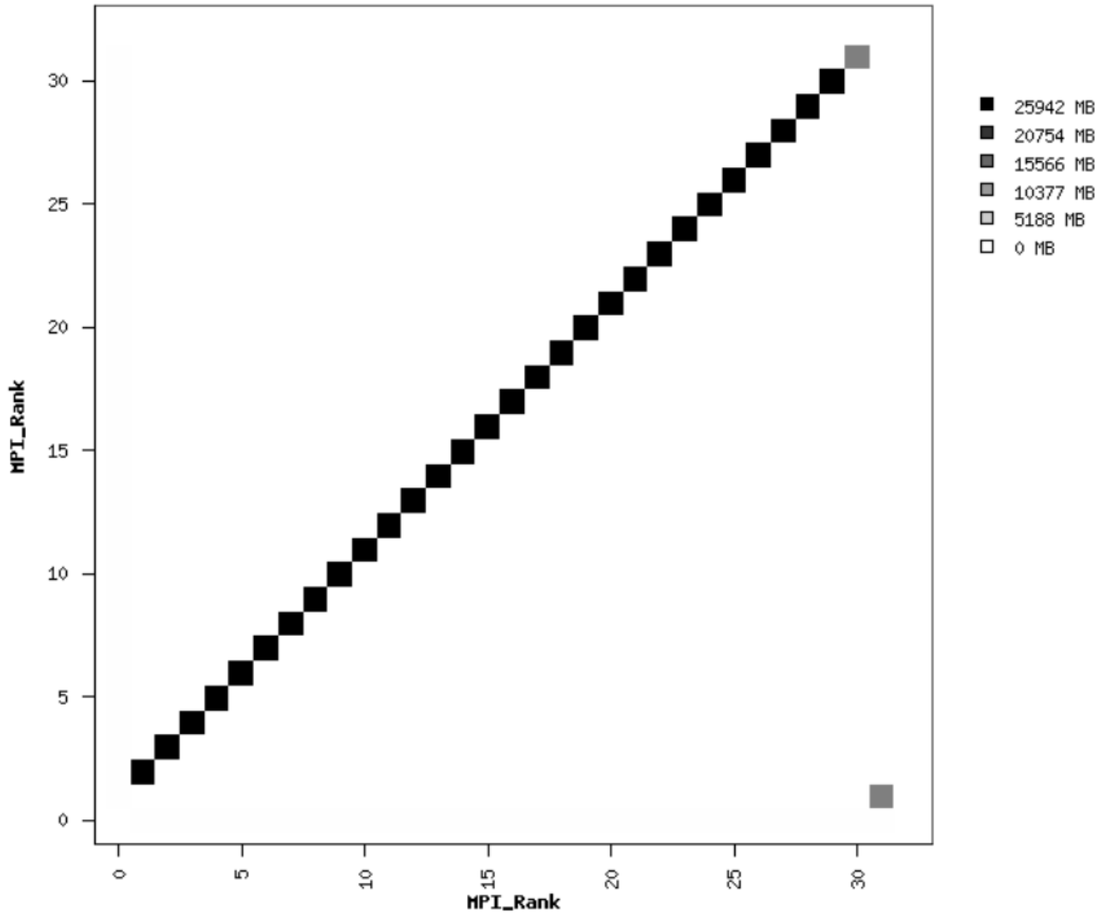
- **Rank 0 – is not a compute rank, use to distribute the job**



32 Nodes, 1 rank per node

RELION MPI Profiling

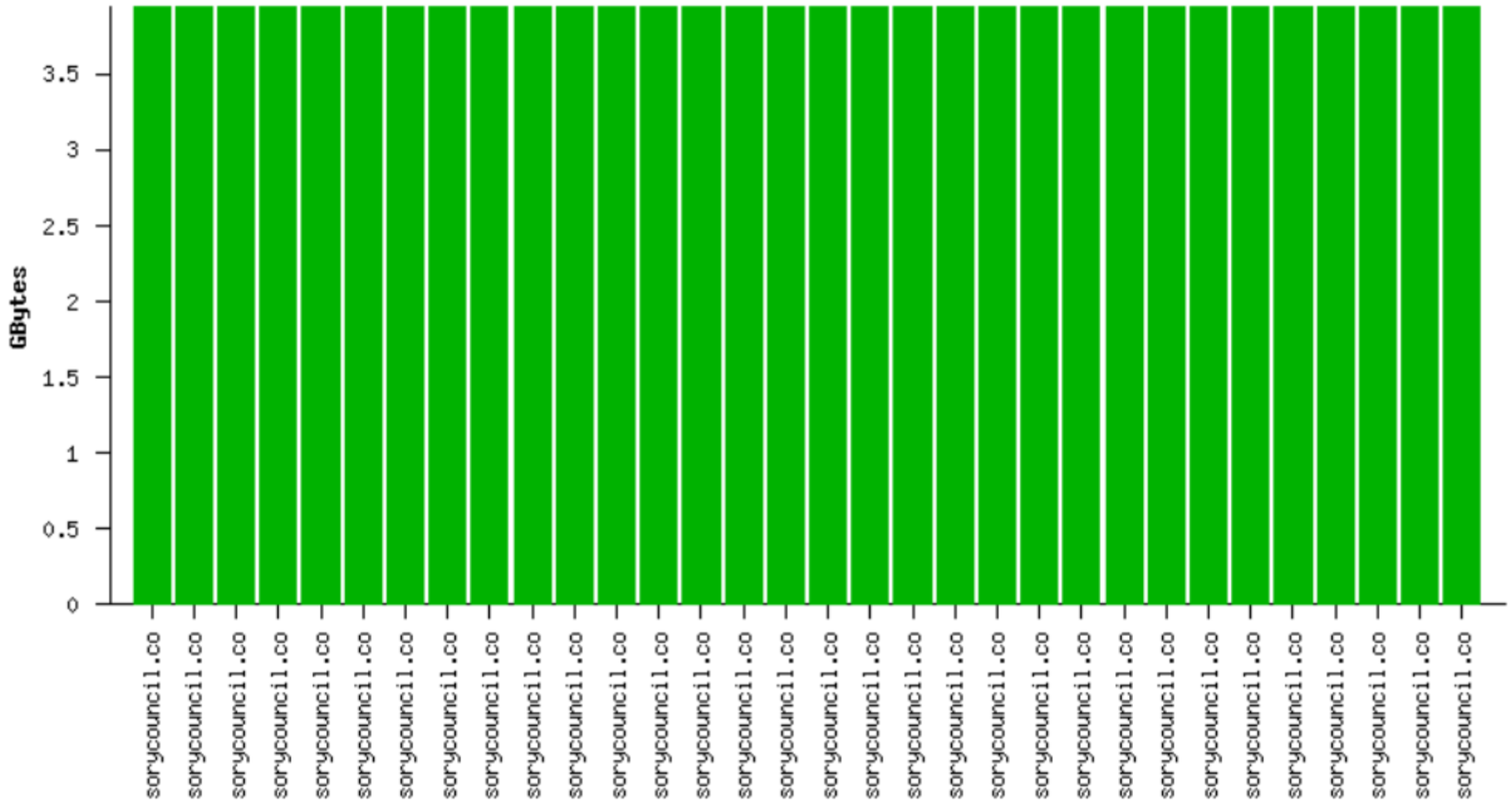
- **Ring communication**



32 Nodes

RELION Profiling

- **Memory footprint**



32 Nodes

Performance Analysis Summary

- **RELION performance testing**
  - Pool size 4,8,16 gave best performance on 16,24,32 nodes
  - SHARP In-Network Computing reduces MPI time by 13% and increase overall application performance by 5%
    - Performance advantages increases with system size, up to 32 nodes were tested

- **RELION Profile**
  - Rank #0 does not perform computation
  - Mostly MPI_Barrier (70%)
  - Ring communication matrix

HPC·AI ADVISORY COUNCIL
NETWORK OF EXPERTISE

Thank You