HPC·AI
ADVISORY COUNCIL
NETWORK OF EXPERTISE

GROMACS
Performance Benchmark and Profiling

Oct 2018

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**
  - GROMACS performance overview over AMD EPYC based platforms
  - Understanding GROMACS communication patterns

- **More info on GROMACS**
  - http://www.gromacs.org/

GROMACS (GROningen MAchine for Chemical Simulation)

- **A molecular dynamics simulation package**
- **Primarily designed for biochemical molecules like proteins, lipids and nucleic acids**
  - A lot of algorithmic optimizations have been introduced in the code
  - Extremely fast at calculating the nonbonded interactions
- **Ongoing development to extend GROMACS with interfaces both to Quantum Chemistry and Bioinformatics/databases**
- **An open source software released under the GPL**

GROMACS
FAST. FLEXIBLE. FREE.

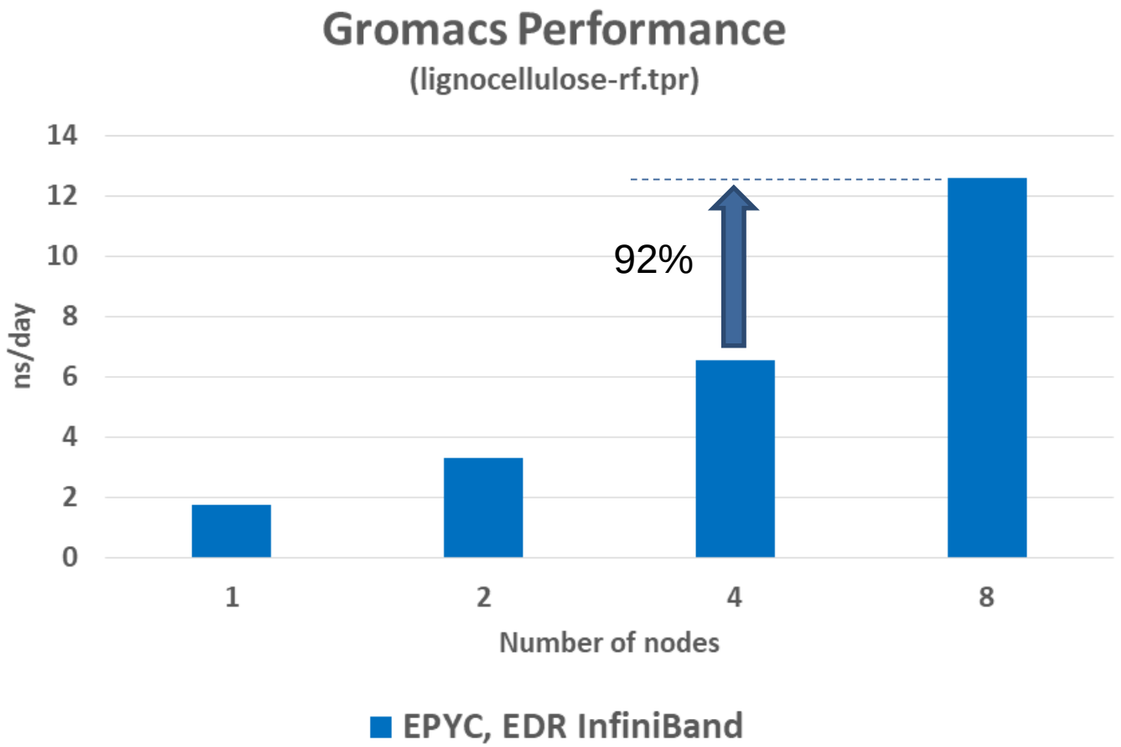
Cluster Configuration

- **Venus cluster**
  - Supermicro AS -2023US-TR4 8-node cluster
  - Dual Socket AMD EPYC 7551 32-Core Processor @ 2.00GHz
  - Mellanox ConnectX-5 EDR 100Gb/s InfiniBand
  - Mellanox Switch-IB 2 SB7800 36-Port 100Gb/s EDR InfiniBand switch
  - Memory: 256GB DDR4 2677MHz RDIMMs per node
  - 240GB 7.2K RPM SSD 2.5" hard drive per node

- **Software**
  - OS: RHEL 7.5, MLNX_OFED 4.4
  - MPI: HPC-X 2.2

GROMACS Performance

- Version 2018.2
- Input “lignocellulose-rf.tpr”

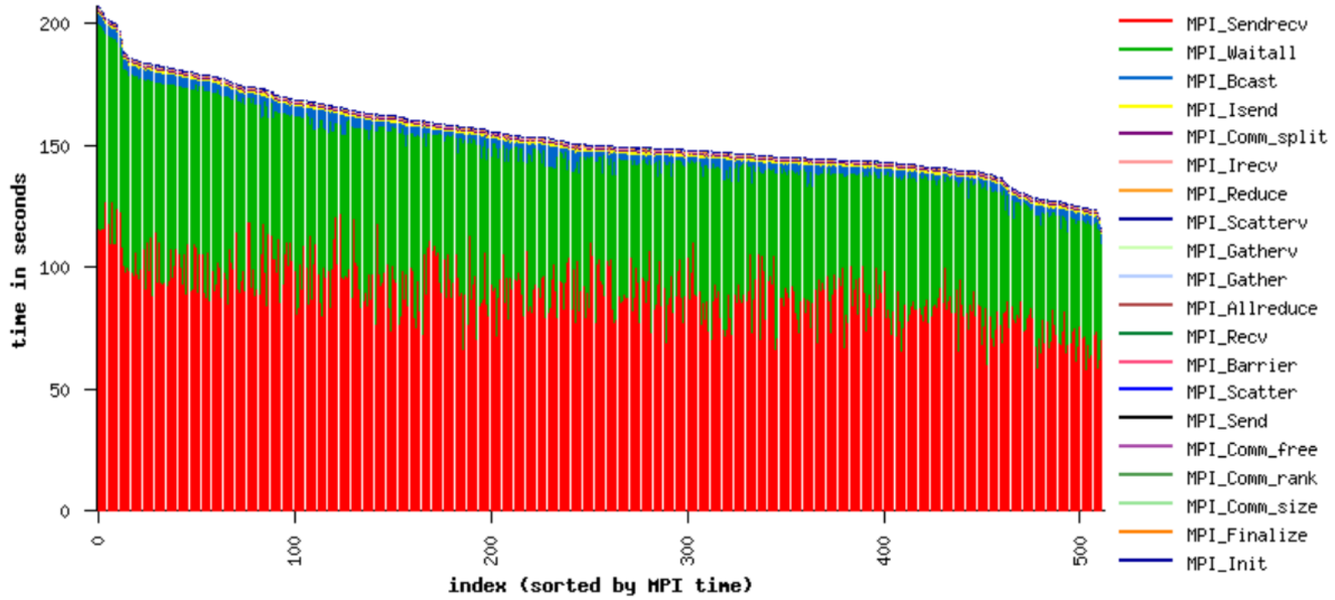


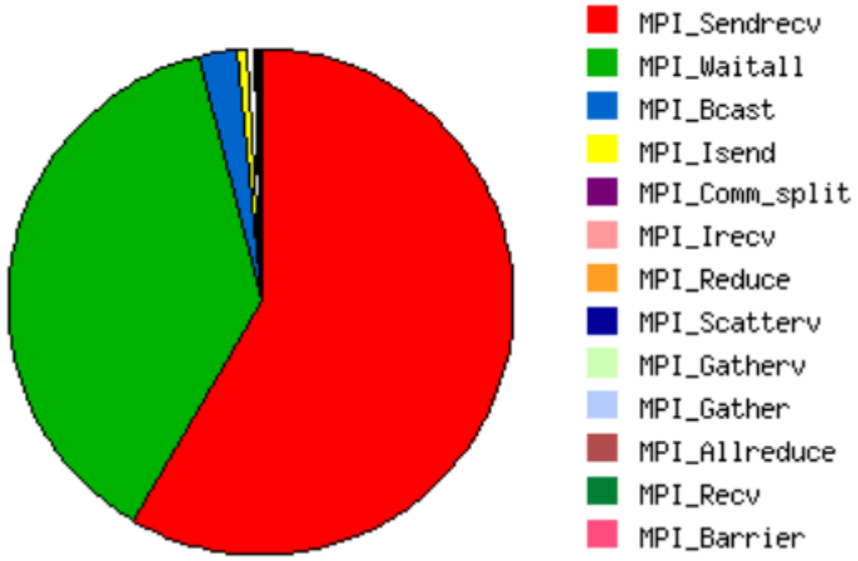
Higher is better

MPI Command: mpirun -np $nproc --map-by node --rank-by core --bind-to core -report-bindings --display-map -mca coll_hcoll_enable 0 -mca pml ucx -x UCX_NET_DEVICES=mlx5_0:1 cp2k.popt -i h2o-dft-ls-4.inp

GROMACS MPI Profiling

- **10.5% MPI and WallClock of 1471 seconds**





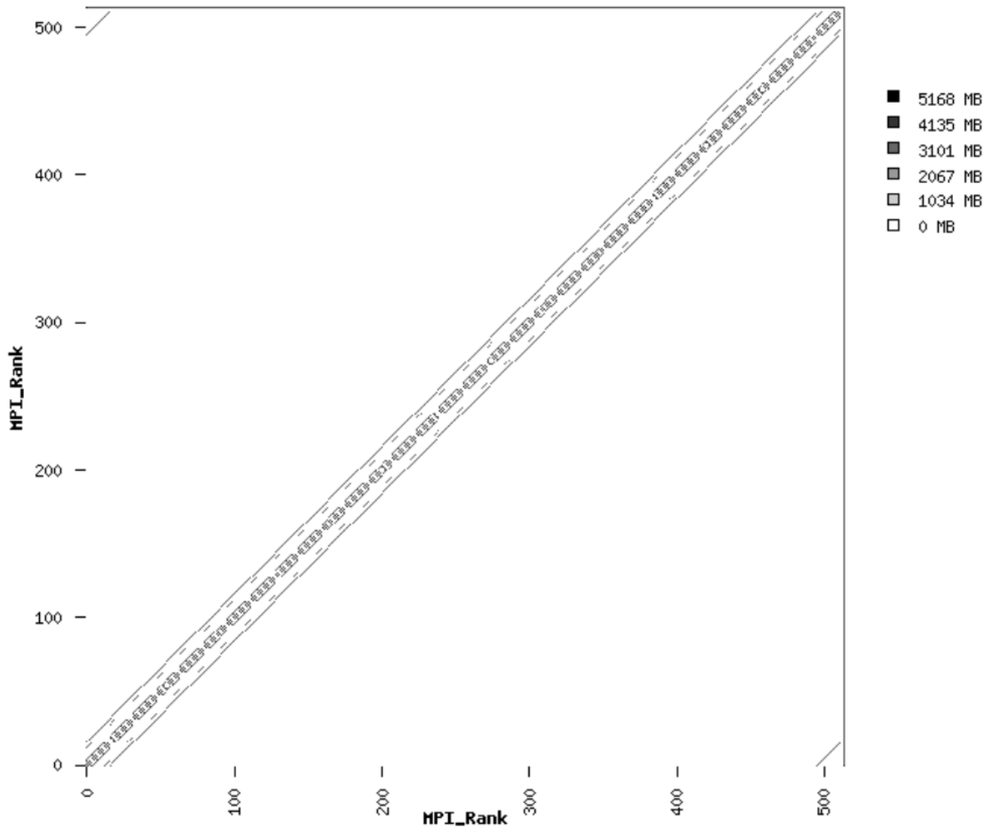
GROMACS MPI Profiling

- **Top MPI communication statistics**

Communication Event Statistics (% detail, --- error)									
	Comm Size	Buffer Size	Ncalls	Total Time	Avg Time	Min Time	Max Time	%MPI	%Wall
MPI_Sendrecv	0	32768	206823545	2.984589e+04	1.443061e-04	5.960500e-06	5.098100e-02	37.75	3.96
MPI_Waitall	0	0	153601536	2.978316e+04	1.938988e-04	0.000000e+00	1.066000e+00	37.67	3.95
MPI_Sendrecv	0	40960	71172429	5.556709e+03	7.807389e-05	6.914100e-06	5.057400e-02	7.03	0.74
MPI_Sendrecv	0	8	37025026	4.506340e+03	1.217106e-04	0.000000e+00	1.526000e-02	5.70	0.60
MPI_Sendrecv	0	5120	2384478	1.343952e+03	5.636251e-04	9.536700e-07	1.626100e-02	1.70	0.18
MPI_Sendrecv	0	49152	13987661	1.211238e+03	8.659335e-05	8.821500e-06	1.677700e-02	1.53	0.16
MPI_Sendrecv	0	28672	14742736	1.175827e+03	7.975638e-05	5.960500e-06	1.569300e-02	1.49	0.16
MPI_Sendrecv	0	14336	4723317	5.697625e+02	1.206276e-04	2.861000e-06	1.491400e-02	0.72	0.08
MPI_Bcast	512	3670016	512	5.211718e+02	1.012478e+00	5.232600e-02	1.070100e+00	0.66	0.07
MPI_Sendrecv	0	1536	2418173	3.829993e+02	1.583838e-04	0.000000e+00	1.429800e-02	0.48	0.05

GROMACS MPI Profiling

- **Most communication happen within the nearby ranks**



GROMACS Summary

- **GROMACS performance testing over AMD EYPC based platform**
  - 92% scaling was achieved from 4 to 8 nodes

- **GROMACS MPI profiling**
  - MPI communication accounts for 10.5% of overall wall clock time at 8 nodes
  - MPI_Sendrecv is 58% of MPI, MPI_Waitall is 37% of MPI and MPI_Bcast is 1% of MPI
  - Most communication is between nearby ranks

HPC·AI ADVISORY COUNCIL
NETWORK OF EXPERTISE

Thank You