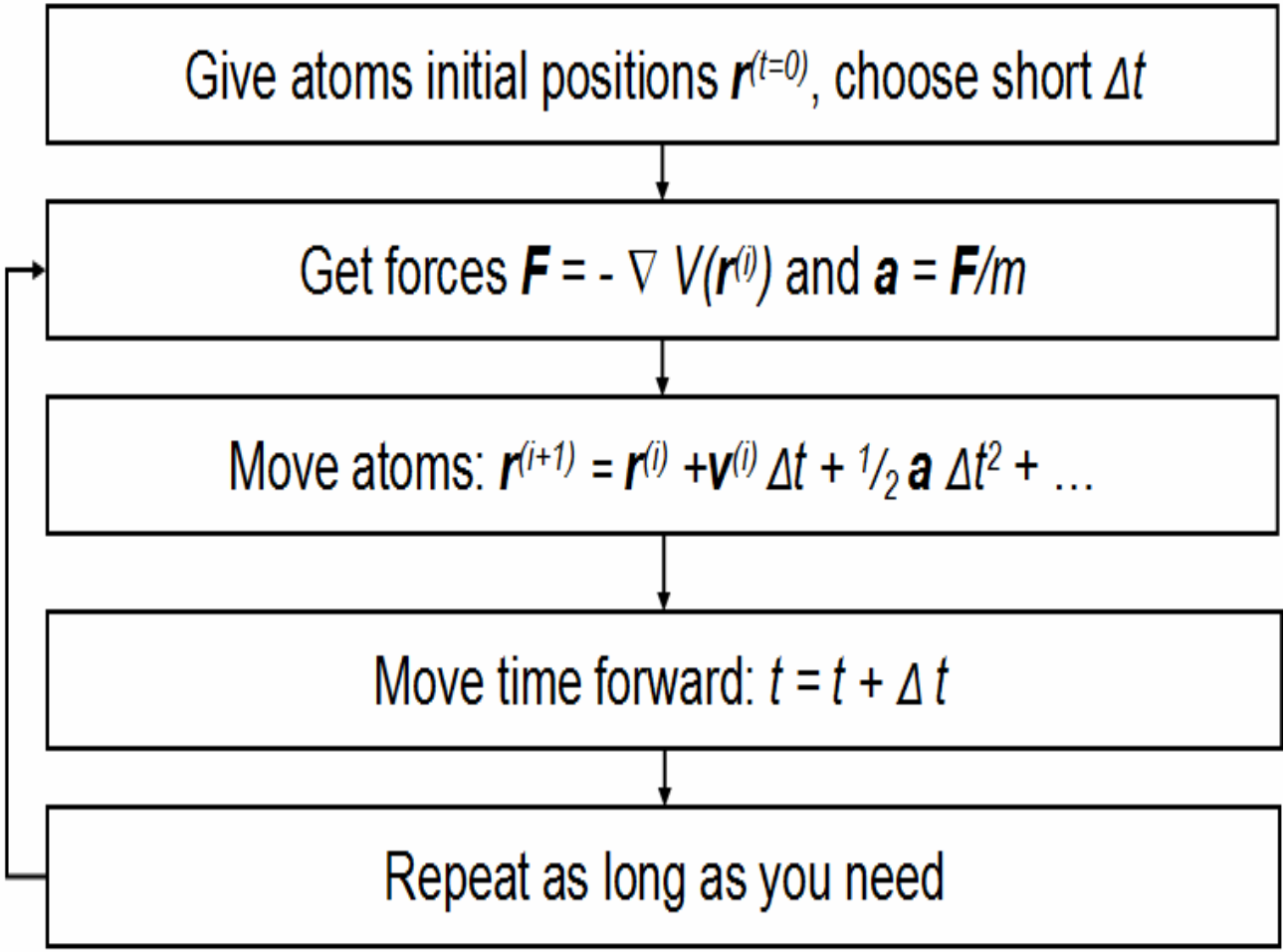
Molecular Dynamics - Gromacs

JURECA BOOSTER-CLUSTER / P. Petkov, S. Markov & V. Pavlov

Molecular dynamics

- Molecular Dynamics (MD) - solving Newtonian equations (F=m*a) of motion of atomic system to calculate its time evolution
- Forces are calculated from classical approximation of interatomic potential function called Force Field (FF). In general FF:
- Bonded terms (bonds, angles, dihedrals, …) – Real Space (Particle) Node – PP node
- Long-range interactions, Electrostatic and Van der Waals terms - Particle Mesh Ewald Algorithm (PME Node)



JURECA

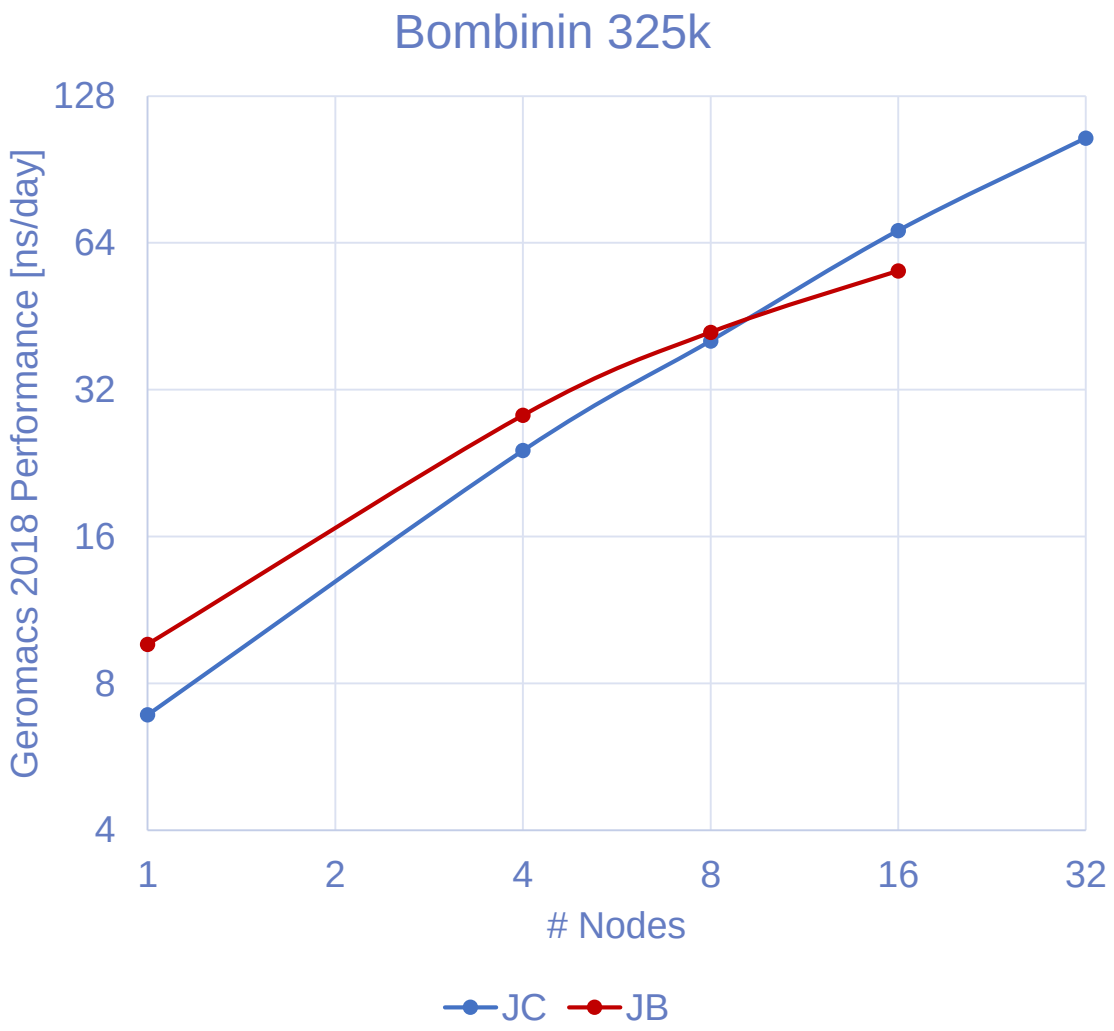
JURECA Cluster

- 1882 compute nodes based on dual-Socket Intel Xeon Haswell - 2680v3, 24 cores, 2.5 GHz
- Mellanox InfiniBand EDR100 Gb/s network !
- Full fat-tree topology
- 2.2 PF/s
- Standard configuration: 128 GB main memory
  - 128 nodes with 256 GB
  - 64 nodes with 512 GB
- 75 nodes with 2× Nvidia K80 GPUs

JURECA Booster

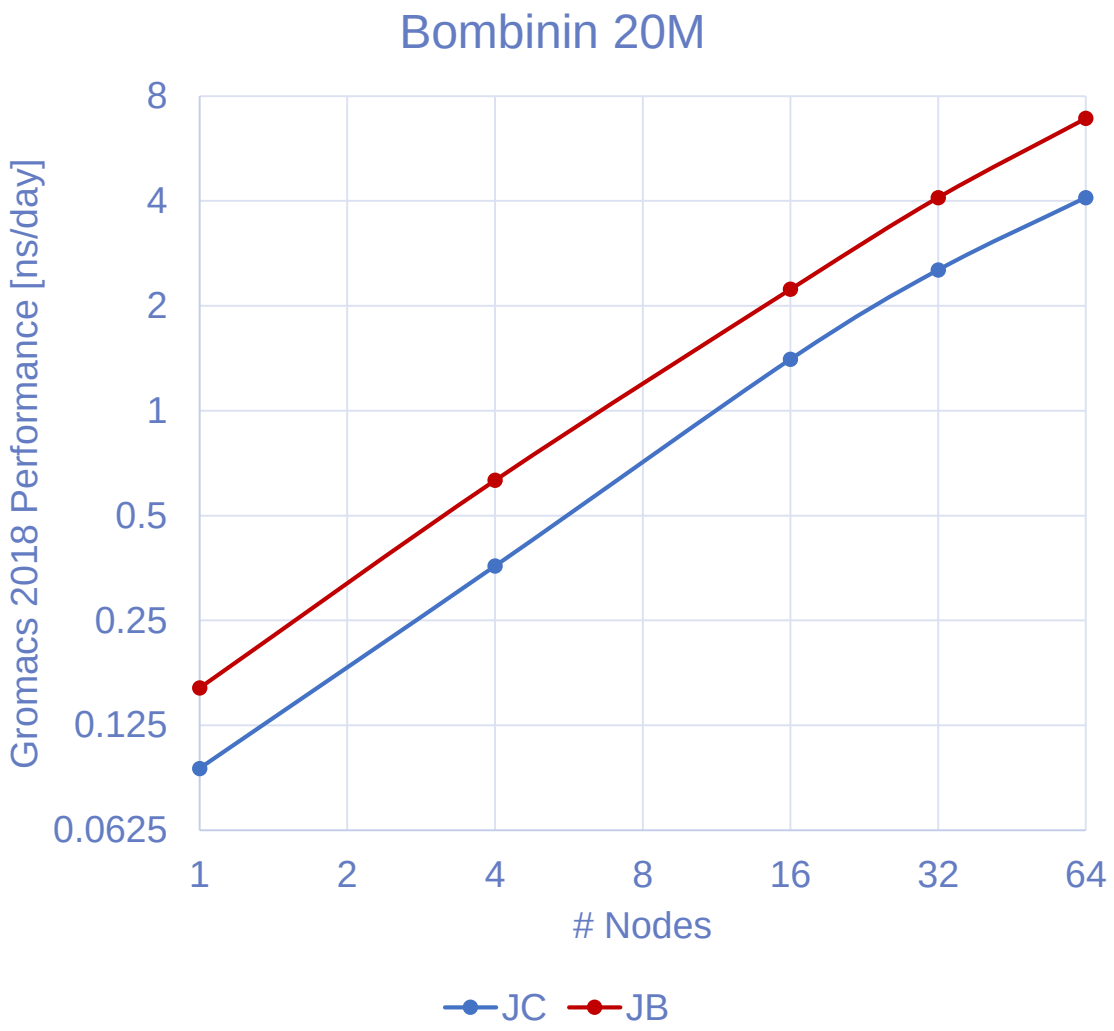
- 1640 compute nodes based on Intel Xeon Phi 7250-F - 7250-F, 68 cores, 1.7 GHz, **quad** mode, 96 GB main memory, + 16 GB MCDRAM, Currently: **hybrid50** mode
- Intel Omni-Path Architecture 100 Gb/s network !
- Full fat-tree topology
- 5 PF/s

Performance test – Bombinin 325k atoms Rc=1.2nm



- JURECA (JC):
  - 24 MPIs per Node
  - 2 OpenMP Threads per MPI
  - NPME ~0.125*(Total #MPIs)
- JURECA BOOSTER (JB):
  - 64 MPIs per Node
  - 2 OpenMP Threads per MPI
  - NPME ~(0.3 – 0.4)*(Total #MPIs)

Performance test – Bombinin 20M atoms Rc=1.2nm



- JURECA (JC):
  - 24 MPIs per Node
  - 2 OpenMP Threads per MPI
  - NPME ~0.25*(Total #MPIs)
- JURECA BOOSTER (JB):
  - 64 MPIs per Node
  - 2 OpenMP Threads per MPI
  - NPME ~(0.3 – 0.4)*(Total #MPIs)

JURECA Cluster+Booster architecture

Cluster

1884 nodes
Mellanox EDR IB
1.8 + 0.4 PF/s

Bridging: 20 Tb/s

Booster

1640 nodes
Xeon Phi 7250-F
Intel OPA
5 PF/s peak

198 MPI bridge nodes

Routing: 1.4 Tb/s

IP routing: 2 Tb/s

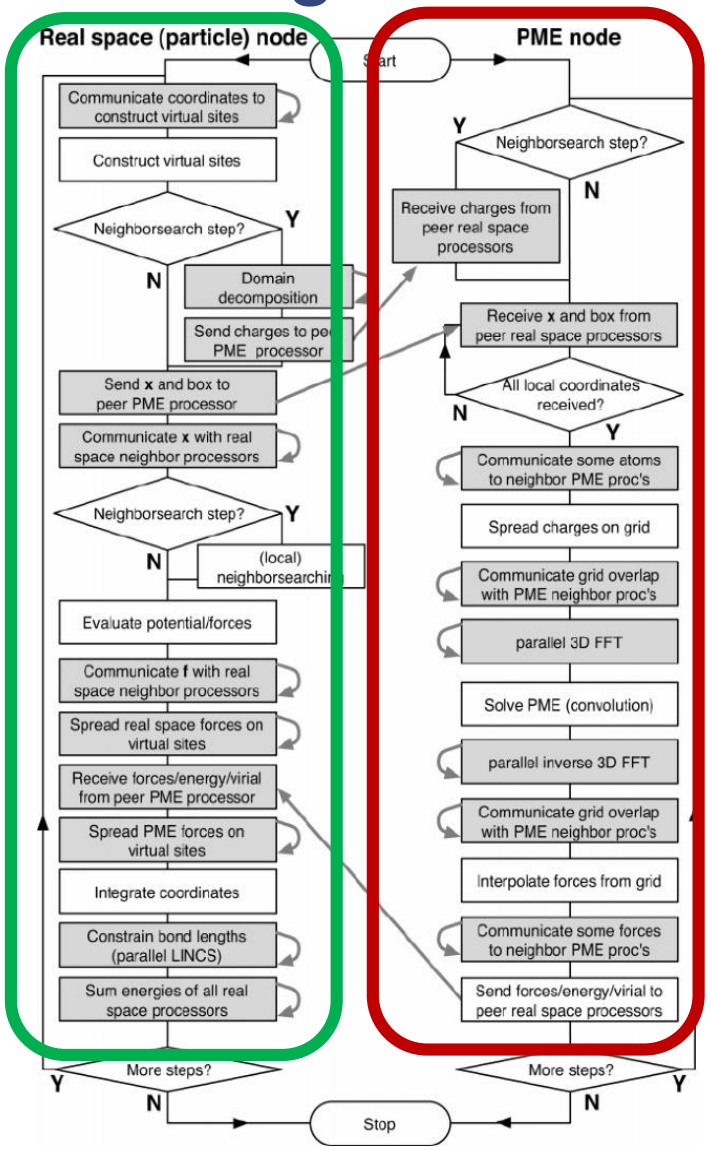
JUST Storage cluster

JURECA Cluster-Booster communication

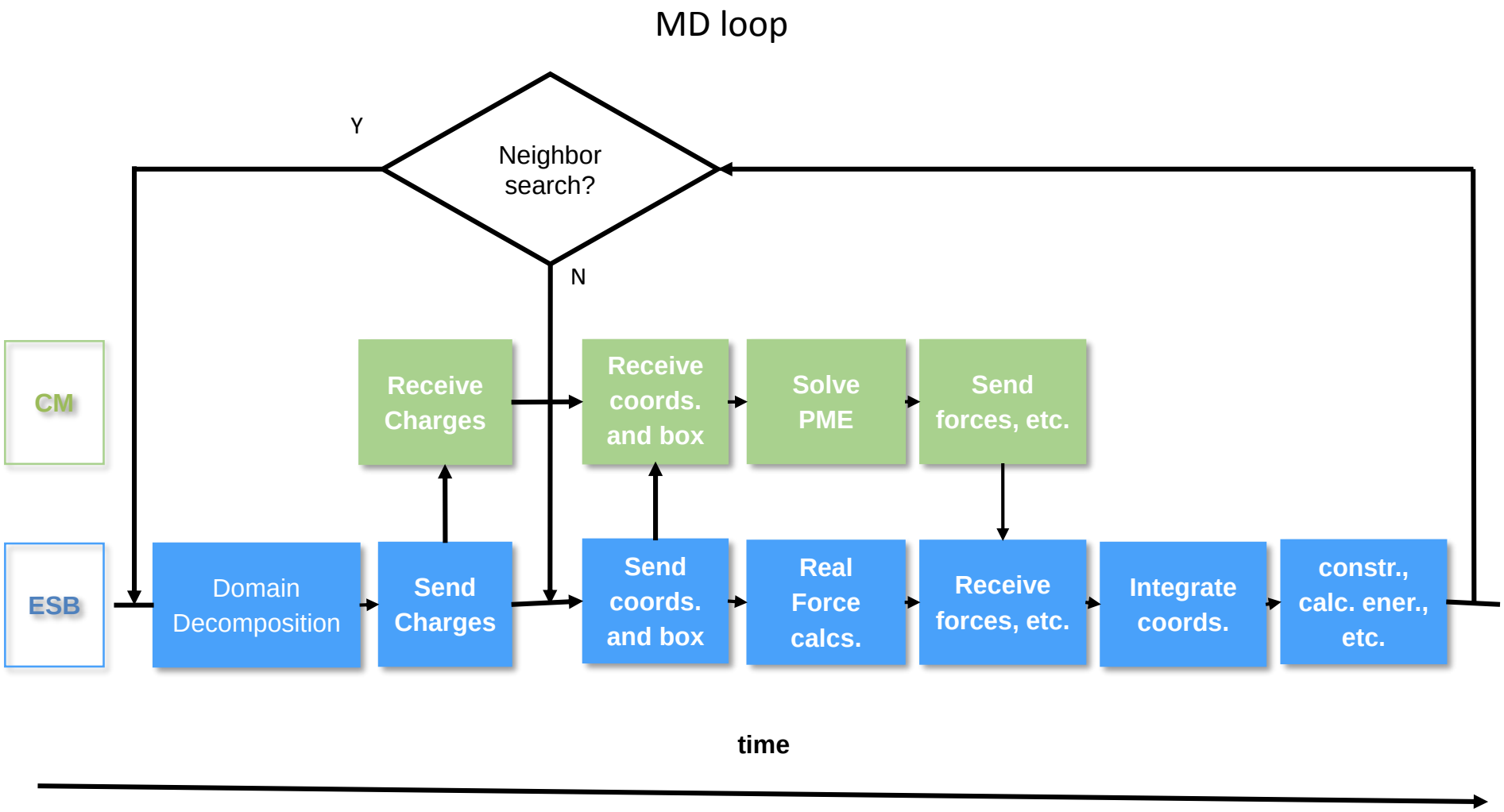
- **Hardware**
  - 198 bridge nodes: each node with 1× EDR InfiniBand HCA and 1× Omni-Path Architecture HFI
    - *Ca. 20 TB/s line-speed bridging capabilities*
- **Software**
  - Slurm 17.11 + ParaStation Modulo: psslurm plugin
    - *Unique feature: Support for heterogeneous jobs with common* ***MPI_COMM_WORLD***
    - *Note: Official Slurm support for heterogeneous MPI jobs started in version 18.08*
  - ParaStation Modulo: Gateway (GW) protocol
    - *ParaStation psgwd gateway daemons launched on bridge nodes*
    - *GW protocol integrated in ParaStation pscom layer ⇨ Transparent for MPI application using ParaStation MPI (except routing choices)*

Gromacs Flow Chart and Application partitioning

- In GROMACS one can divide MPI ranks into two groups one for real-space calculations (PP nodes) and the rest dedicated to PME calculations (PME nodes).
- At the beginning of the time step, each PP node **sends coordinates and charges** to its corresponding PME node and ones the **PME calculations are done** each PME node **sends the resulting forces back** to the corresponding PP node.
- Meanwhile, **all collective communications** go **only between PME nodes** and **only between PP nodes** which **overlaps FFT communications (exchanged between PME nodes only) with real-space calculations.**



Application mapping

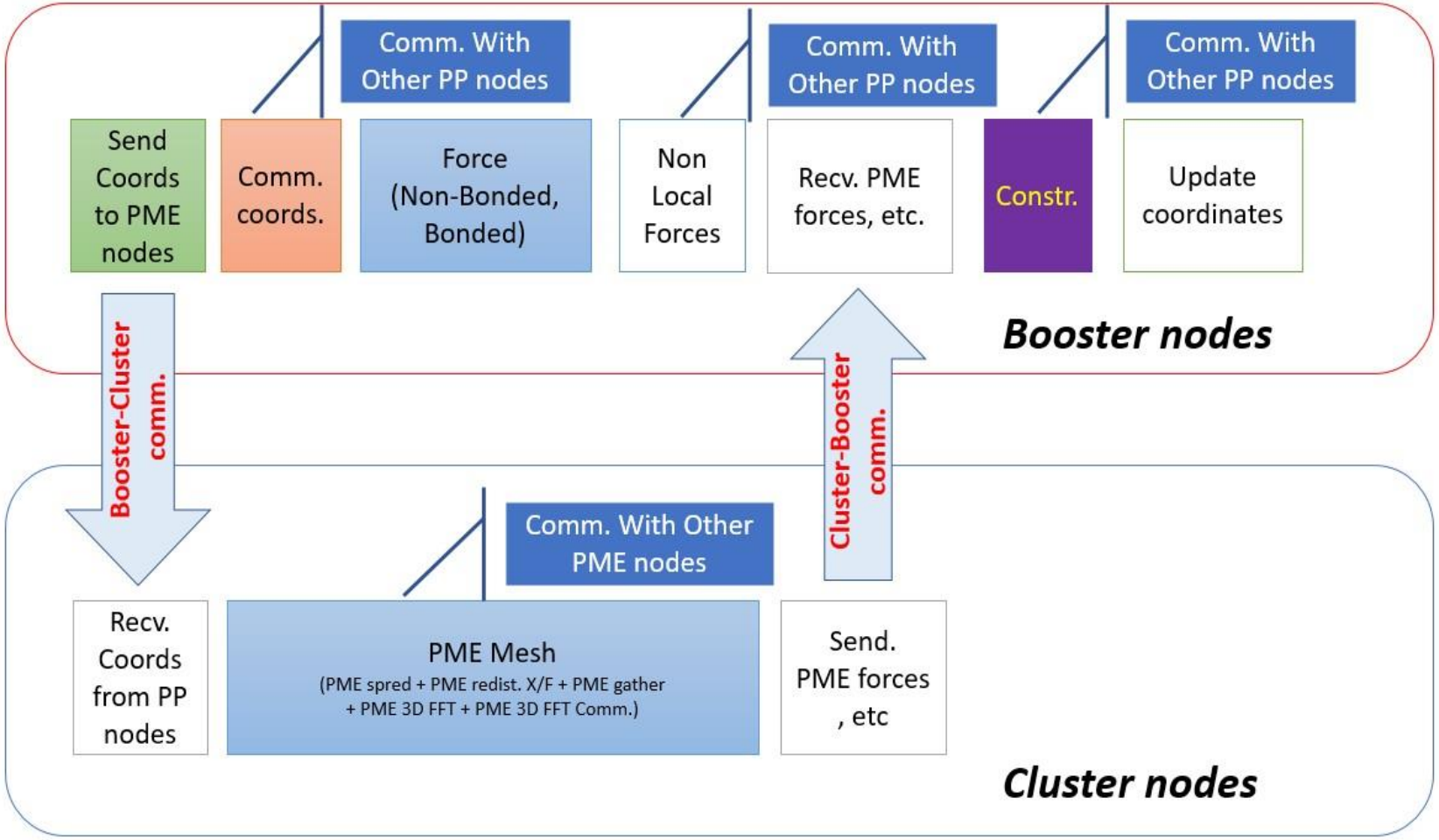


Gromacs performance counters

Gromacs performance counters description (herein we will mention only a subset of them, mainly those used in the analysis due to their non-negligible influence on the performance scalability judgement):

- PP MPI processes
  - Send X to PME – time spent by PP ranks for sending atoms' coordinates to PME ranks
  - Comm. coord. – time spent to exchange atoms' data used for calculations wih other domains
  - Force – pair interactions force calculations
  - Wait + Comm. F (named "non-local forces" in the next slide) – time needed to communicate forces contributions calculated on other domains
  - Update (named "Update coordinates" in the next slide) – set new coordinates and velocities at the end of the timestep calculations
  - Constraints – in the particular case, calculations and communication time need for keeping bonds length fixed and water molecule rigid.
- PME PMI processes
  - PME mesh – the time needed for PME forces calculations
  - PME redist. X/F – redistribute atoms parameters and coordinates before each FFT calculation
  - PME spread and PME gather – spreading and gathering the data across PME ranks
  - PME 3D-FFT and PME 3D-FFT Comm. – time spent in FFT itself and needed communications, respectively.

Gromacs performance counters (II)

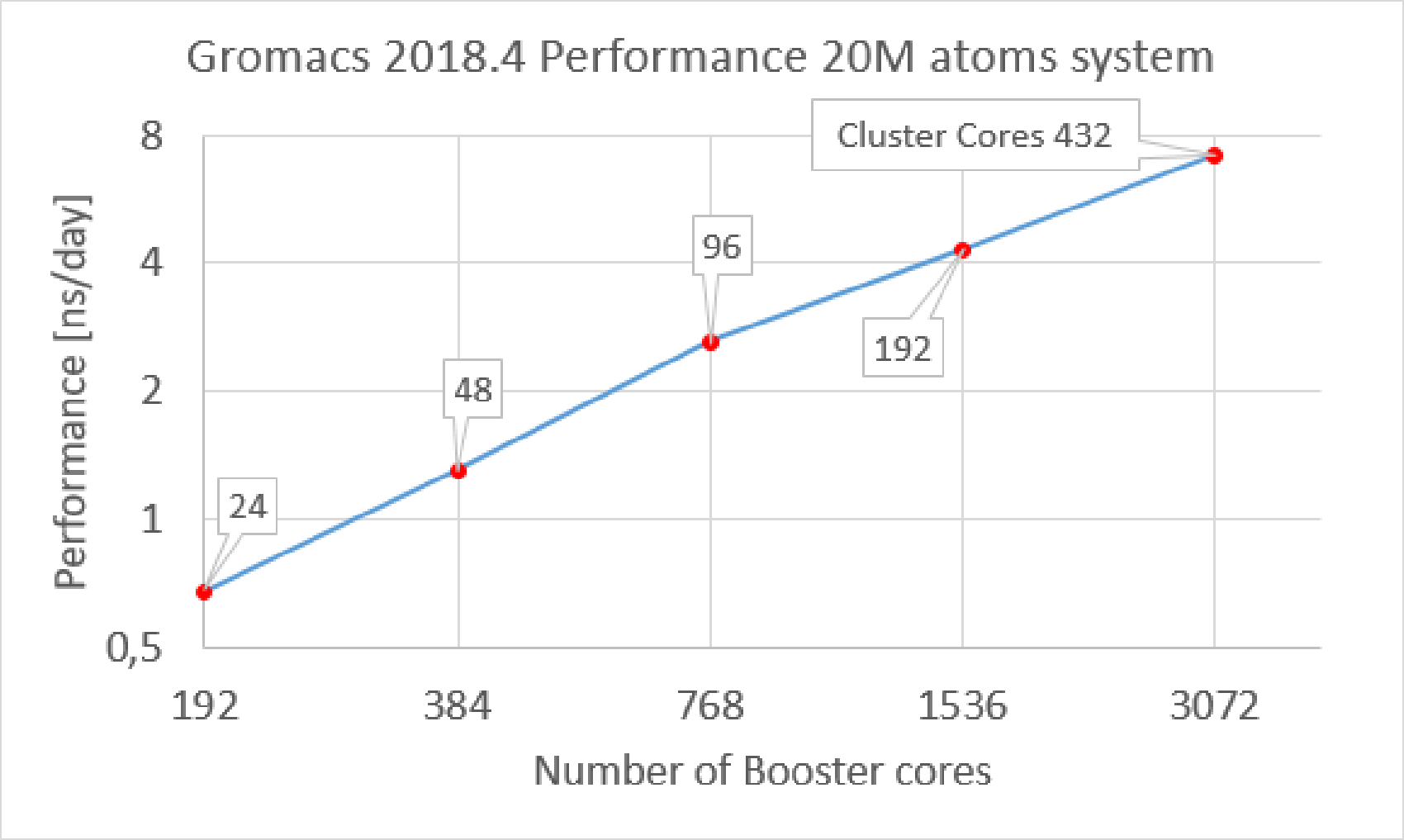


Test simulations

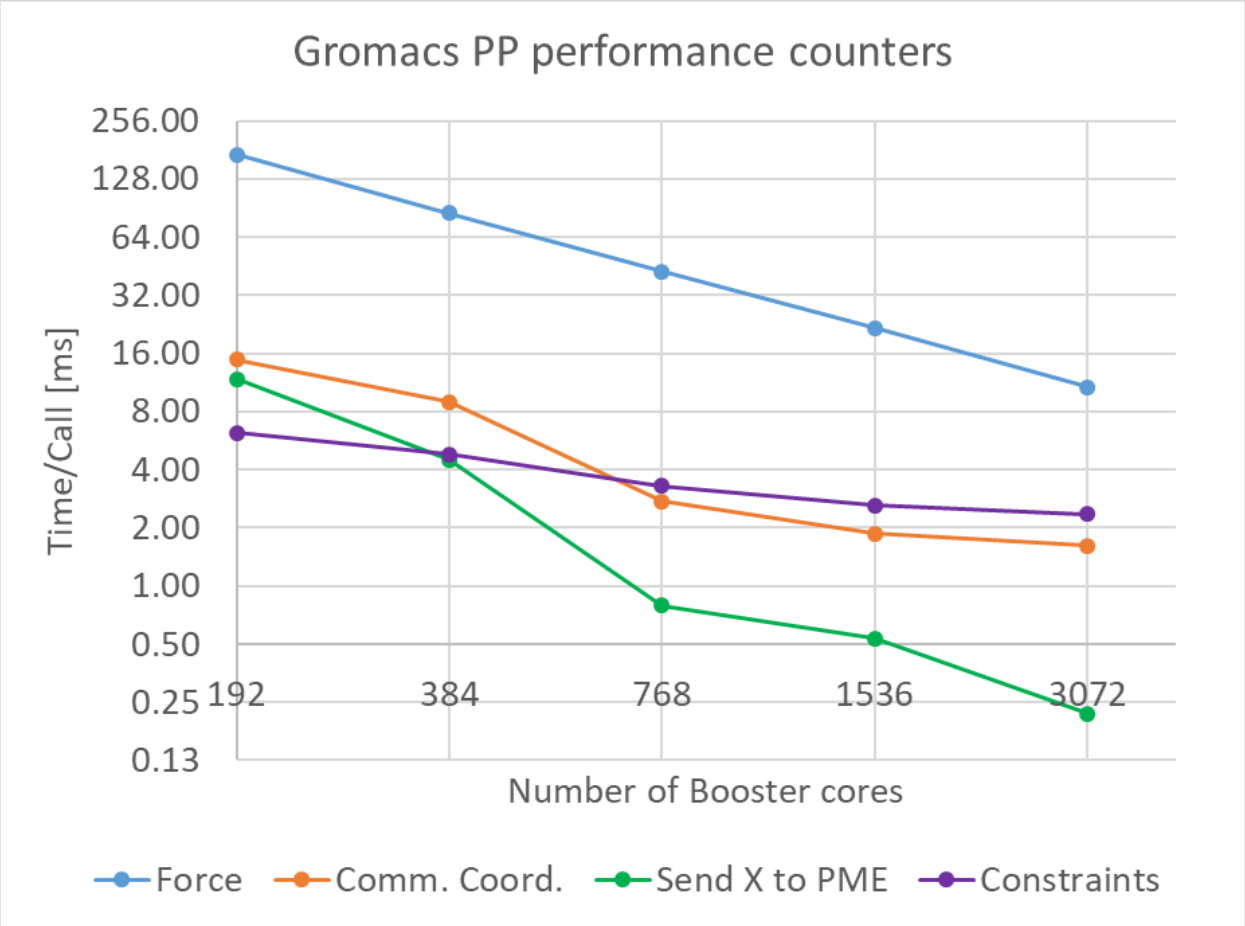
BOOSTER NODES	TOTAL NUMBER OF BOOSTER CORES	BOOSTER MPIS / THREADS PER MPI	CLUSTER NODES	TOTAL NUMBER OF CLUSTER CORES	CLUSTER MPIS / THREADS PER MPI	NUMBER OF GATEWAYS
3	192	192/2	1	24	24/2	8
6	384	384/2	2	48	48/2	8
12	768	768/2	4	96	96/2	32
24	1536	1536/2	8	192	192/2	64
48	3072	3072/2	18	432	432/2	128

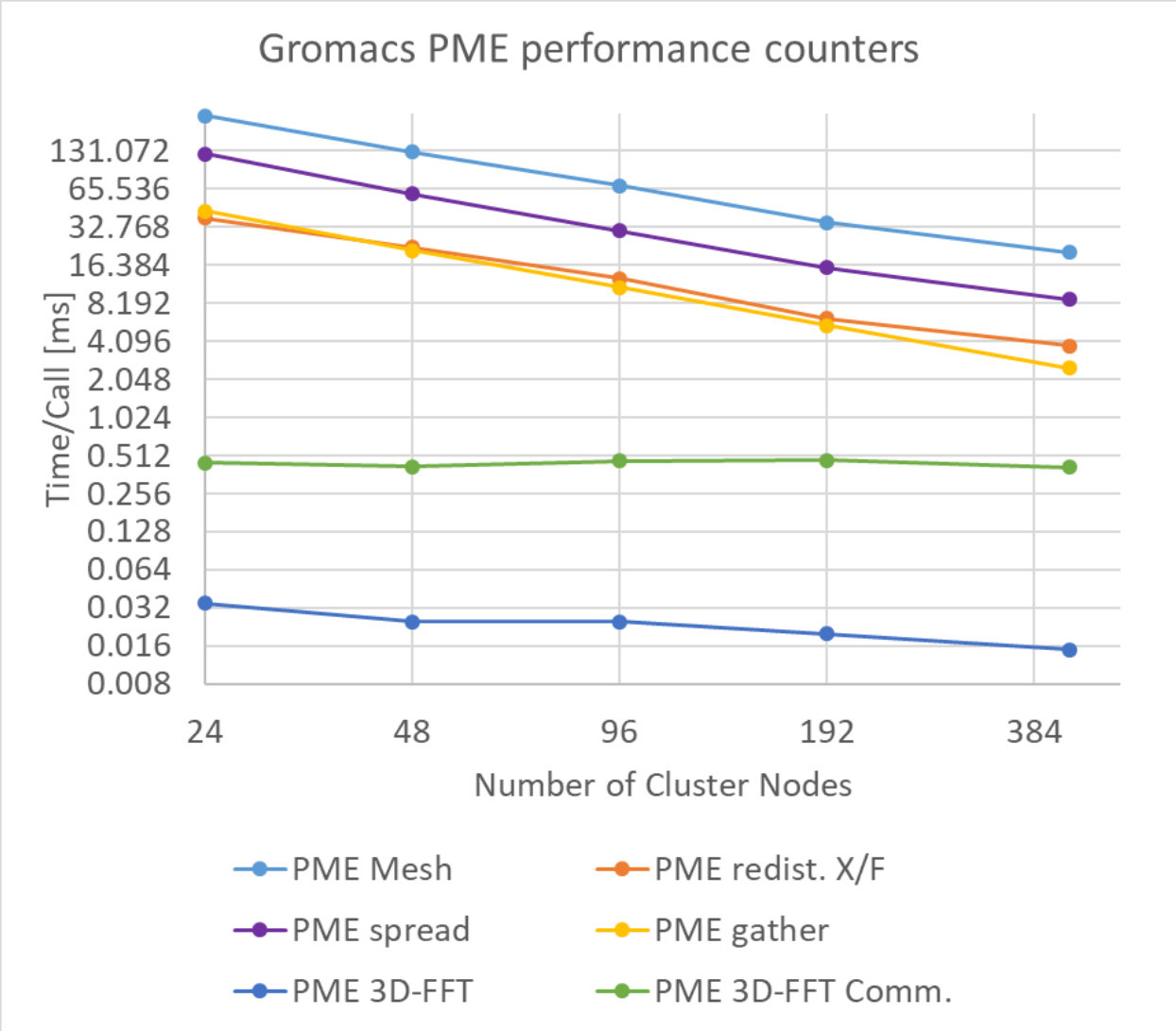
N.B.: The job parameters like PP/PME MPI ranks ratio and number of gateways were optimized for maximal performance for given number of BOOSTER nodes.

Performance scalability results – 20M atoms MD simulation

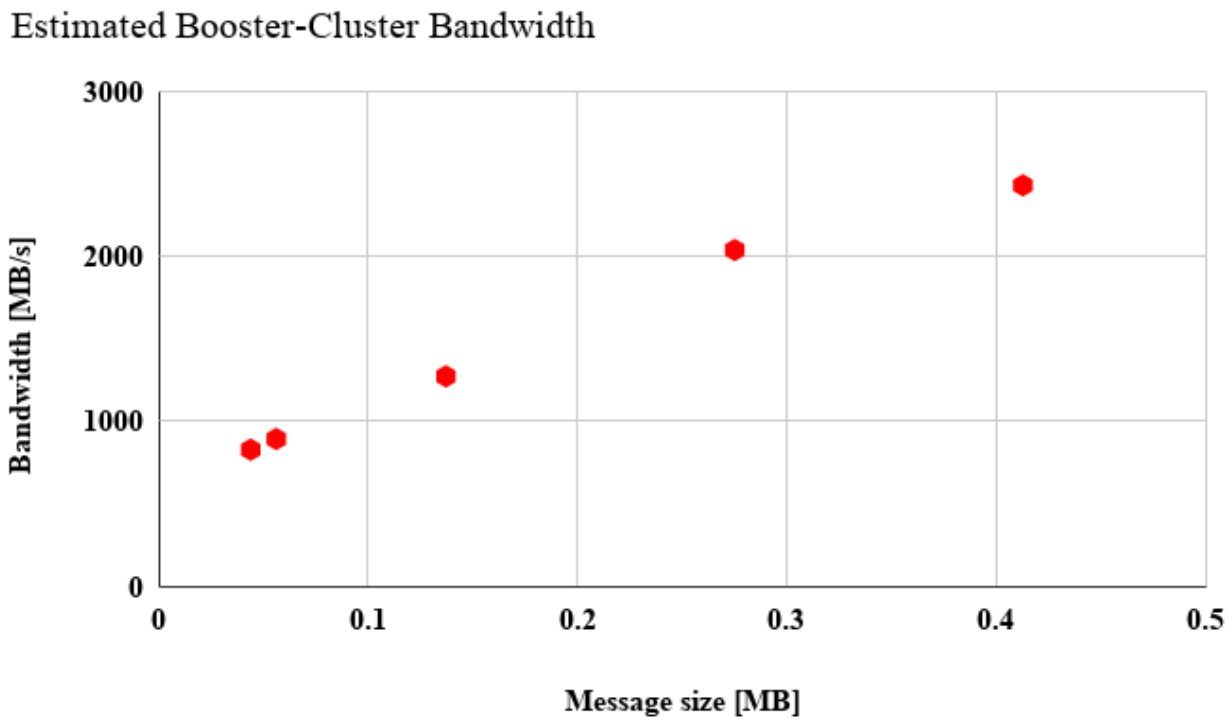


Gromacs performance counters analysis





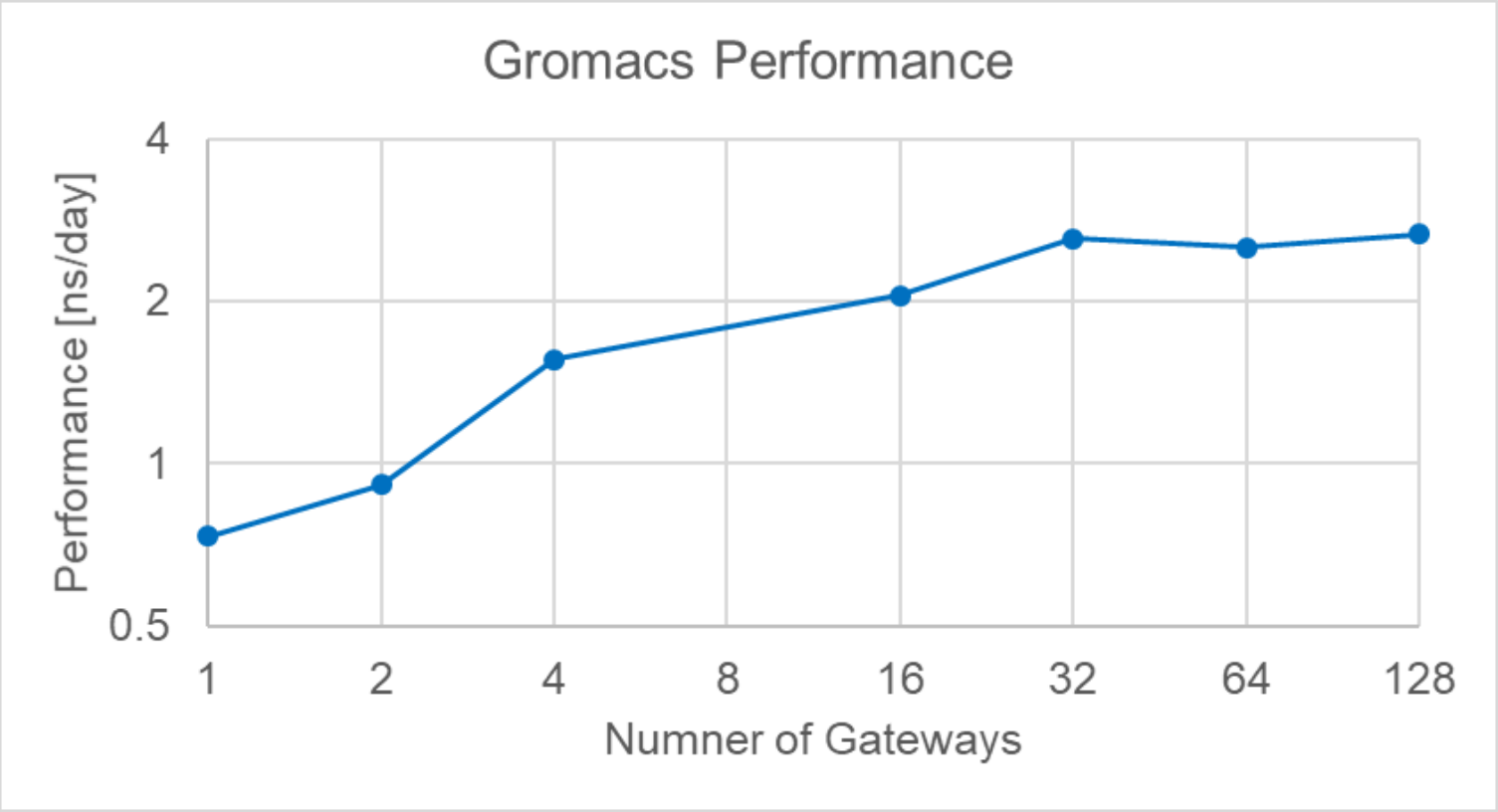
Estimated Cluster-Booster bandwidth



- Test simulations with 2.2 M and 20 M atoms on a various number of Booster and Cluster nodes keeping nPP/nPME ration about 3 to have the same domain decomposition as in performance scalability test, but the different length of the messages.
- The number of gateways used was always 8.

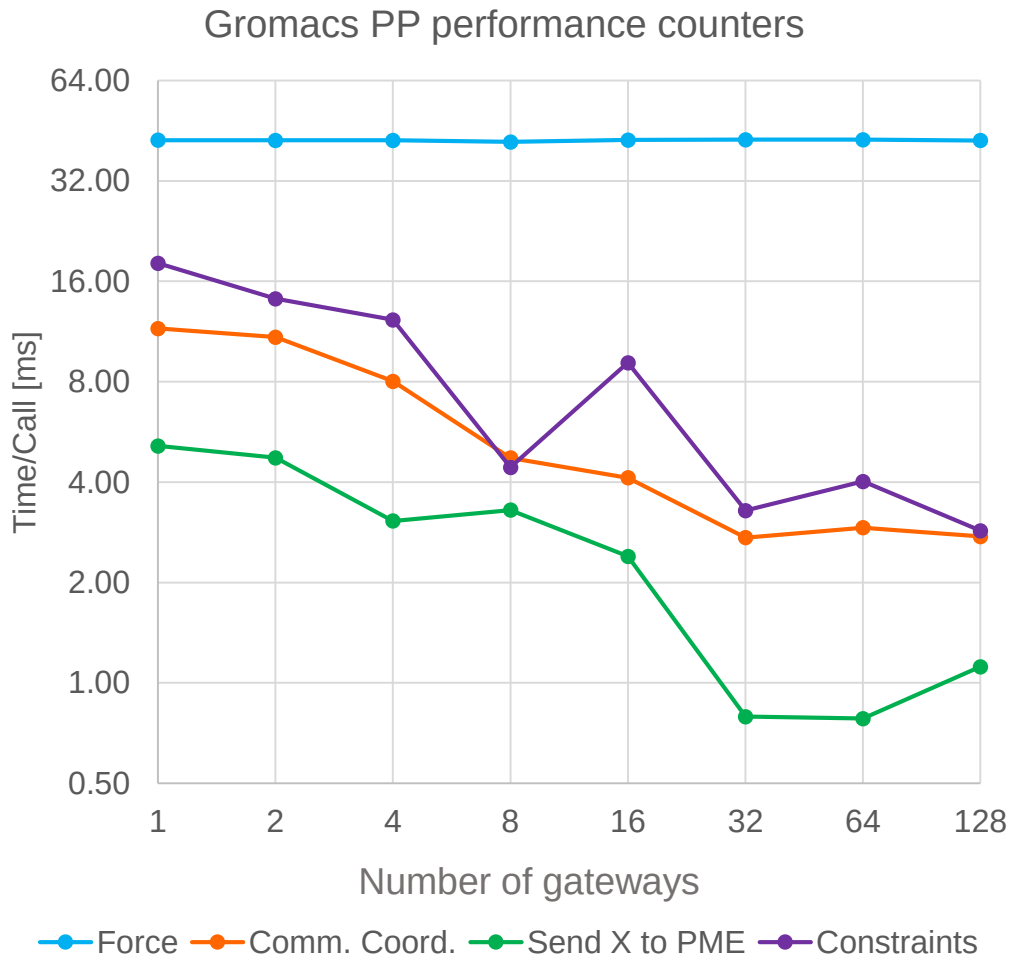
in a very good agreement with the result from the synthetic tests reported in https://bscw.zam.kfa-juelich.de/bscw/bscw.cgi/2712620

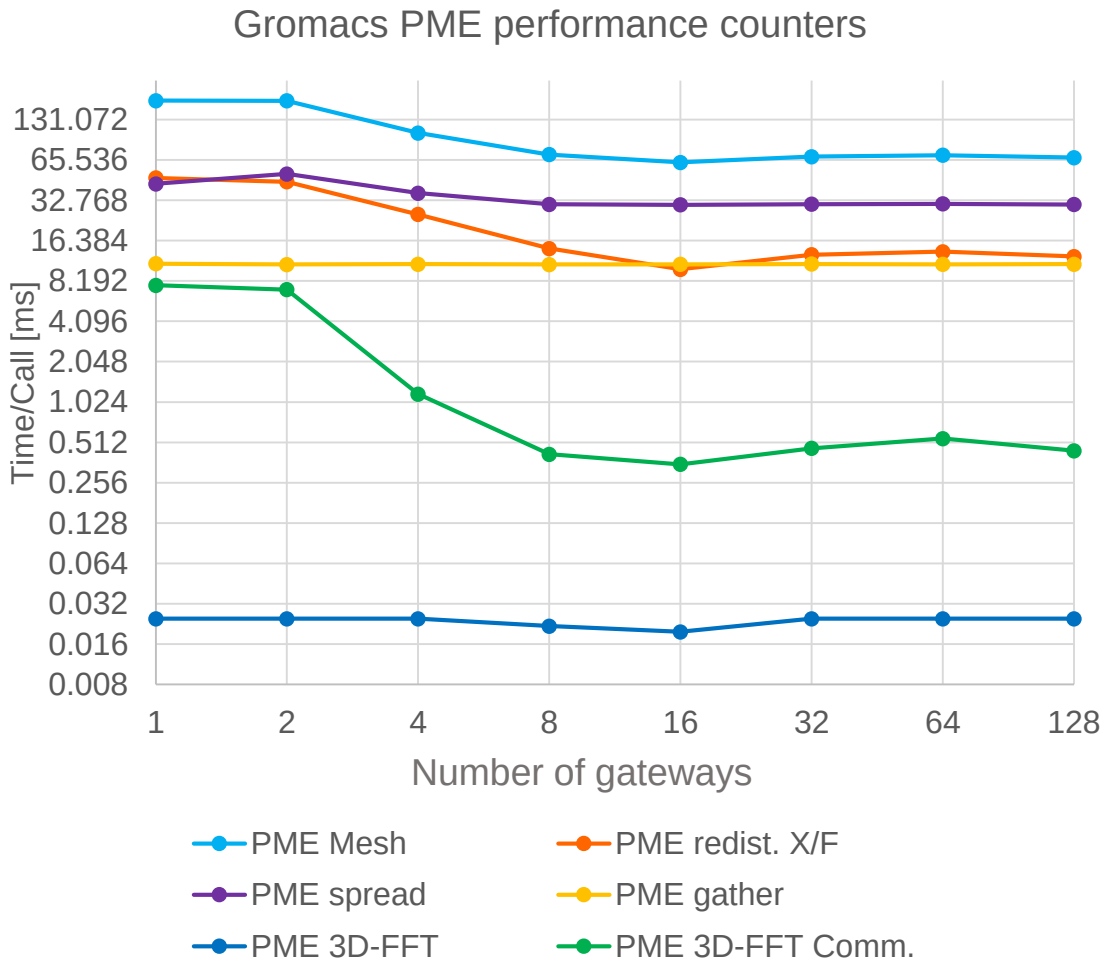
Performance vs. Number of Gateways



- 12 Booster nodes and 4 Cluster nodes with a different number of gateways

Gromacs performance counters vs. number of gateways





Conclusions

- The simulations showed good enough performance scalability and proper utilizations of the computational resources due to a good module interconnecting system.
- inter-modular influence on intra-modular communications bandwidth?
- In general our results show that JURECA system could be used for large-scale MD simulations with Gromacs in Cluster-Booster configuration.

Acknowledgements

- Special thanks to Dorian Krause, Brian Wylie, Ilya Zhukov, Jacopo De Amicis and Estela Suarez.

DEEP Projects

The DEEP projects have received funding from the European Union's Seventh Framework Programme (FP7) for research, technological development and demonstration and the Horion2020 (H2020) funding framework under grant agreement no. FP7-ICT-287530 (DEEP), FP7-ICT-610476 (DEEP-ER) and H2020-FETHPC-754304 (DEEP-EST).