About me

- Born in The Netherlands
- MSc Physics, PhD Physical Oceanography
- Climate Science for several years
- Worked 10 years as consultant for SURFsara in Amsterdam (PRACE, etc.)

- Since April 2018 working for Atos, based in Munich

- john.donners@atos.net

OpenMPI Tuning

2020

OpenMPI

- An Open-source MPI library
- Full MPI-3.1 standards conformance
- Thread-safe
- Tunable
- Many job schedulers supported
- High performance on all platforms, from laptop to cluster

- Use OpenMPI as a secondary option
  - Intel MPI is superior for the OPA network

Using OpenMPI

- module load openmpi
  - openmpi/gcc.9/3.1.5
  - openmpi/intel/3.1.6
- Compiler wrappers
  - mpicc
  - mpicxx
  - mpifort
- Underlying compiler depends on the module loaded

Tuned collectives

- Collectives are often an important part of MPI applications.
- Most collectives can be completed using multiple algorithms.
- It can be worthwhile to tune the algorithms for the MPI collectives that take up most of the time.
  - Compare this to the I_MPI_ADJUST_ family of environment variables
- This must be done for your particular application and particular setup.

- mpirun --mca coll_tuned_use_dynamic_rules 1 …

Tuned MPI collective algorithms

MPI collective	Algorithm count
Allreduce	6
Alltoall	6
Allgather	7
Allgatherv	6
Alltoallv	3
Barrier	7
Bcast	7
Reduce	7
Reduce_scatter	4
Gather	4
Scatter	3

How to use a different algorithm

- Runtime settings:
  - export OMPI_MCA_coll_tuned_use_dynamic_rules=1
  - export OMPI_MCA_coll_tuned_alltoall_algorithm=5

- Relatively easy to tune: Rerun a short job with all possible algorithms

- ompi_info -a:
  - *_algorithm_count
  - *_algorithm

Impact of different algorithms

- Paper „ Improvement of parallelization efficiency of batch pattern BP training algorithm using Open MPI” shows an improvement of 13.1%
- Own tests changing the collectives algorithm show similar improvements for multiple applications.

- See also documentation for I_MPI_ADJUST family

An unrelated remark about jobs

- Sometimes a problem occurs during a job
- Edit the job script to fix the problem
- Several iterations later it is not clear anymore what job script was used for which job. (based on my own experience)

- Print the job script in your job, to be certain of what job script was used.

```
cat $0
```

Thanks for your attention

john.donners@atos.net

Atos