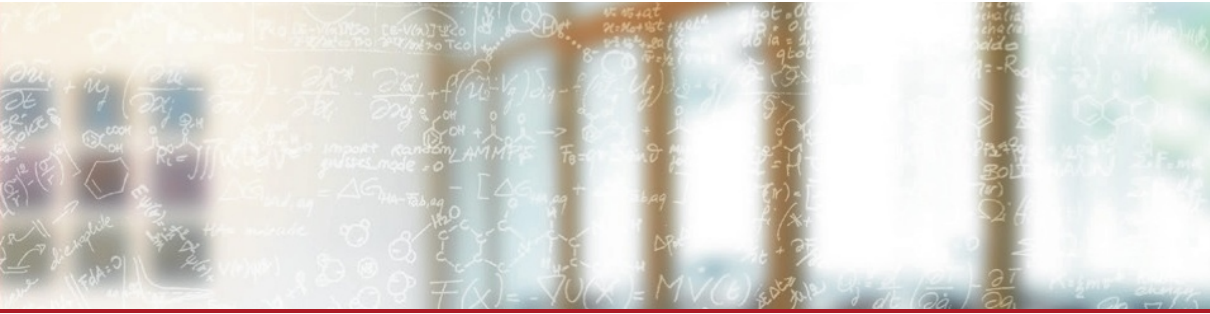
CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

ETH zürich



Towards an optimal allreduce communication in message-passing systems

EuroMPI/USA'20

Andreas Jocksch (CSCS), Noe Ohana (SPC EPFL), Emmanuel Lanti (SPC EPFL), Vasileios Karakasis (CSCS) and Laurent Villard (SPC EPFL)

23 September, 2020

Motivation

- Goal: Fast allreduce operation as MPI persistent collective communication
- Collective communication provides complex communication patterns which can be highly optimised
- Allreduce
  - Frequently used on HPC systems
  - Applications in data science but also in other fields
- Related work for algorithmic optimisations and for other collectives (all-to-all)

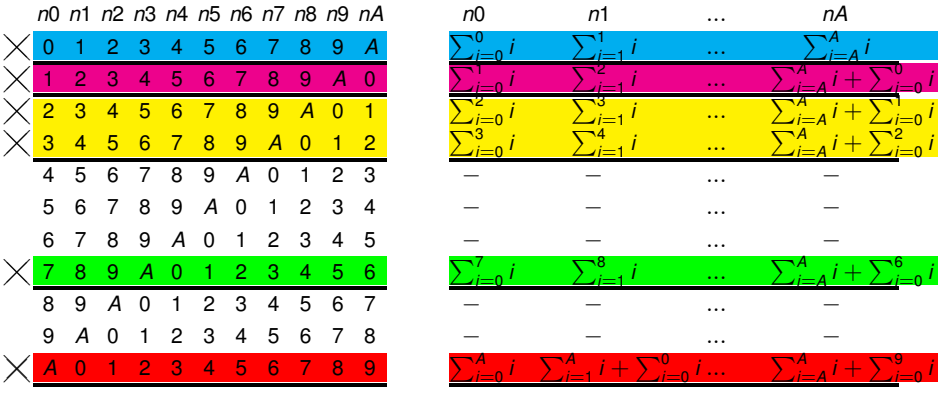
Algorithms

- All algorithms exploit shared memory on the node
- One or more (max all) cores used for communication between nodes
- Generation of bytecode in the setup phase
- Repeated execution of the bytecode

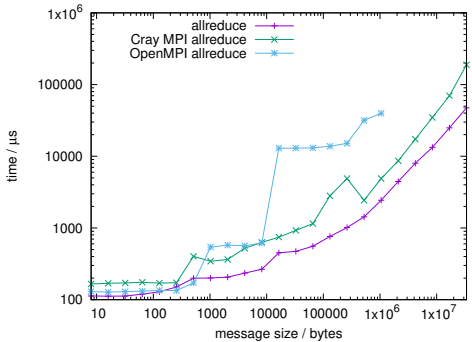
Algorithms contd.

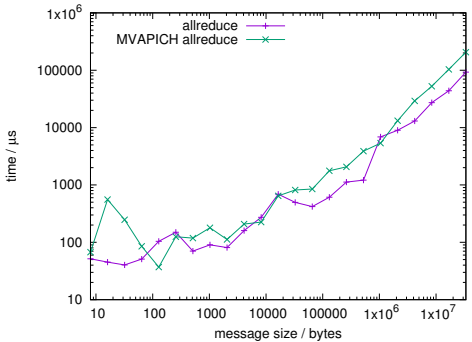
- Generalisation of cyclic shift with flexible factors between steps
- Long messages
  - Reduce_scatter followed by allgatherv
  - Variable message size
- Short messages
  - Allgather with prefix operation

Cyclic shift with prefix operation



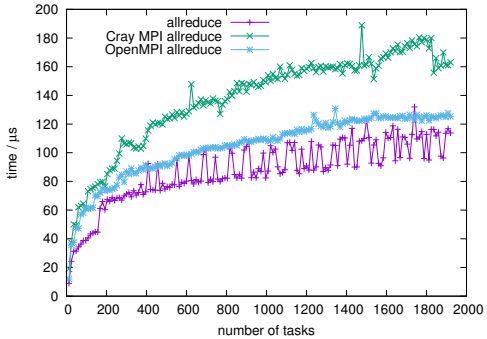
Benchmarks

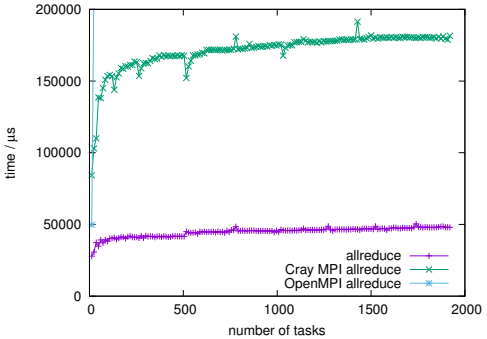




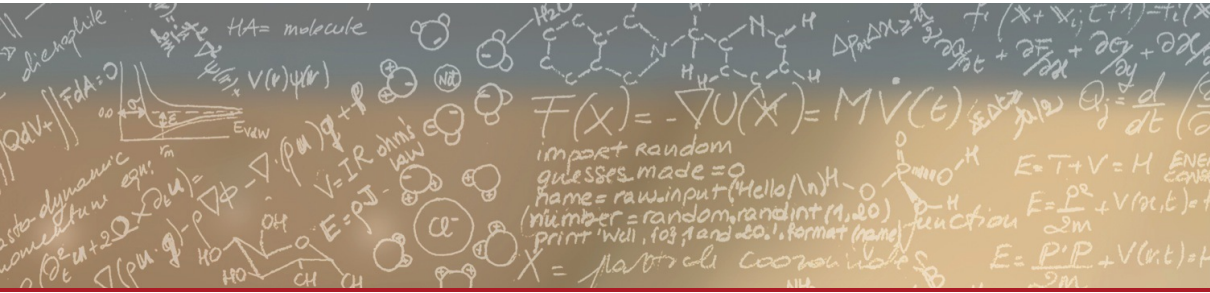
Allreduce on 160 nodes with 9600 MPI processes Cray (left) and on 17 nodes with 408 MPI processes Infiniband (right)

Benchmarks





Allreduce with 8 bytes (left) and 33554432 bytes (right), Cray



Thank you for your attention.

https://github.com/eth-cscs/ext_mpi_collectives