Cost of some MPI routines

Laura Grigori

INRIA and LJLL, Sorbonne Université

October 2020
Plan

Cost of MPI routines
Cost of MPI routines, based on [Thakur et al., 2005]

Point-to-point communication (Send, Recv of a message of $n$ words)

$$\alpha + n\beta$$

MPI_Allgather: $n/P$ data from each process gathered on all processes cost based on recursive doubling algorithm (exchange $n/P$, $2n/P$, up to $2^{\log P - 1} n/P$ data in the last step):

$$\log P\alpha + \frac{P - 1}{P} n\beta$$
MPI_Broadcast: broadcasts \( n \) words from root to all processes, based on the binomial tree algorithm: First step, root sends data to (root + P/2); continue recursively with root and (root + P/2) as new roots.

\[
\log P(\alpha + n\beta)
\]
Cost of MPI routines (contd)

MPI_Alltoall: each process sends unique $n/P$ data to every other process. For long messages, pairwise exchange algorithm with $(P-1)$ steps:

$$(P - 1)\alpha + n\beta$$

MPI_Reduce: a global reduction operation on $n$ words of data, returns the result on the root. For short messages, reduction based on a binomial tree:

$$\log P(\alpha + n\beta + n\gamma)$$

MPI_Allreduce: a global reduction operation on $n$ words of data, returns the result on all processors. For short messages, similar to the recursive doubling algorithm used in MPI_Allgather.