F2: Communication

Lars Karlsson

2009-03-27

Outline

- Point-to-point
- Broadcast
  - One-to-all (broadcast)
  - All-to-all
- Reduction
  - All-to-one
  - All-to-all
- Prefix sum (scan)
- Personal communication
  - One-to-all (scatter)
  - All-to-all
- Circular shift

Point-to-point (`MPI_Send` / `MPI_Recv`)

- An m-word message from one process to another takes time

$$t_s + t_w m$$

  according to our basic communication cost model.
- t_s is the startup cost.
- t_w is the *word transfer time* or the *inverse bandwidth*.
- *We will assume cut-through routing and ignore the hop delay in the rest of this lecture. See the course literature for the details.*
- Typically, t_s is in the microsecond range whereas t_w is in the nanosecond range.
- Beware that in all analyses the word size is implicit which means that you must be careful when you calculate t_w from a bandwidth given in MB/s.

One-to-all Broadcast (`MPI_Bcast`)

One process sends an m-word message to all other processes.
Cost:

- Ring algorithm

$$\left\lceil \frac{p}{2} \right\rceil (t_s + t_w m).$$

- Recursive doubling

$$(t_s + t_w m) \log_2 p.$$

All-to-all Broadcast (`MPI_Allgather`)

All processes have their own m-word message that they broadcast to all other processes.
Cost:

- Ring algorithm
$$(p-1)(t_s + t_w m).$$
- Mesh algorithm
$$(\sqrt{p}-1)(2t_s + t_w m).$$
- Hypercube algorithm
$$t_s \log_2 p + t_w m(p-1).$$
- **Note:** all algorithms have the same transfer times but different startup costs.

Reduction (`MPI_Reduce` / `MPI_Reduce_scatter`)

- All-to-one reduction (`MPI_Reduce`) is the dual to one-to-all broadcast.
- All-to-all reduction (`MPI_Reduce_scatter`) is the dual to all-to-all broadcast.

All-reduce (`MPI_Allreduce`)

Each process has an m-word message that is to be reduced and a copy of the result is left on each process.
Cost:

- Hypercube algorithm

$$(t_s + t_w m) \log_2 p.$$

- **Note:** prefix sums can be computed with the same communication pattern and cost.

Scatter and Gather (`MPI_Scatter` / `MPI_Gather`)

(Scatter:) one process sends a unique m-word message to every other process.
The dual operation is gather.
Cost:

- Hypercube algorithm

$$(t_s + t_w m) \log_2 p.$$

All-to-all personalized (`MPI_Alltoall`)

Each process has a unique m-word message for each of the other processes.
Cost:

- Ring algorithm

$$(p-1)(t_s + t_w m\frac{p}{2}).$$

- Mesh algorithm

$$(\sqrt{p}-1)(2t_s + t_w mp).$$

- Hypercube algorithm

$$(p-1)(t_s + t_w m).$$

Circular shift (not in MPI)

Shift by q steps so that the message initially at process i ends up at process $(i + q) \mod p$.
Cost:

- Ring algorithm

$$\min\{q, p - q\}(t_s + t_w m).$$

- Mesh algorithm (upper bound)

$$(\sqrt{p} + 1)(t_s + t_w m).$$

- Hypercube algorithm (upper bound)

$$(2 \log_2 p - 1)(t_s + t_w m).$$