PARALLEL PROGRAMMING

Assignment 1

MPI – Getting Started

Tammo Johannes Herbert, Rico Jasper (319396), Erik Rudisch

24.04.2013

Exercise 3 – Collective Communications

In Exercise 2b you have implemented a variant of a broadcast operation: before the (series of) communication only one node knows the data, after the communication all nodes are aware of it.

(a) Influence of the interconnection network

Assume that each MPI process is executed on its own node and that the interconnection network has a certain node to node latency and a certain bandwidth between a node and the interconnection network. How long does it take (ignoring computation) until all processes of Exercise 2b are aware of the message?

- p number of processes
- n size of message in bytes
- α latency in seconds
- β bandwidth/data rate in bytes per second

Solution

Usually network delays are separated into four delay components: Queuing, processing, transmission and propagation delay. Here, we only consider the transmission delay $n\beta$ and the propagation delay α. Since there are $p-1$ consecutive transmission, the resulting delay is:

$$(p-1)\cdot(\alpha+n\beta)$$

(b) Other ways to broadcast

The broadcast in Exercise 2b is rather inefficient as most nodes do nothing most of the time.

Describe multiple ideas to realize the broadcast more efficiently (using only send/receive). Analyze each idea as in Exercise 3a and compare them. Is there a “best” variant to broadcast, or do we need different algorithms for, e. g., different message sizes?

Solution

A simple approach which relies on the known MPI_Send and MPI_Recv functions is broadcast-via-unicast. The original source simply sends the message to each other recipient. Therefore the formula changes to $\alpha+(p-1)\cdot n\beta$, which is a difference of $(p-2)\cdot\alpha$. This assumes, that the transmission to the next recipients starts right away without waiting for the previous message to be off the line. Such an implementation is applicable for small messages and few recipients.

A more advanced approach is based on multicast or actual broadcast. In this case the message is duplicated by the network for each recipient. The formula would be $\alpha+n\beta$, which doesn’t consider any processing or queue delays caused by the network. Also multicast could be fairly complicated to set up and therefore to much effort for small messages. Broadcast on the other hand could flood the network and also reach hosts which are not interested in the delivered messages.