Available at
www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT®

ELSEVIER

Information Processing Letters 86 (2003) 33–38

Information Processing Letters

www.elsevier.com/locate/ipl

A bandwidth latency tradeoff for broadcast and reduction

Peter Sanders [a,*], Jop F. Sibeyn [b]

[a] *Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany*
[b] *Department of Computing Science, Umeå University, 901 87 Umeå, Sweden*

Received 9 August 2000; received in revised form 16 January 2002

Communicated by F. Dehne

Abstract

The "fractional tree" algorithm for broadcasting and reduction is introduced. Its communication pattern interpolates between two well known patterns—sequential pipeline and pipelined binary tree. The speedup over the best of these simple methods can approach two for large systems and messages of intermediate size. For networks which are not very densely connected the new algorithm seems to be the best known method for the important case that each processor has only a single (possibly bidirectional) channel into the communication network.

Keywords: Collective communication; Broadcast; Reduction; Tree; Single ported; Half-duplex; Full-duplex; Parallel algorithms; Mesh; Hierarchical Crossbar

1. Introduction

Consider P processing units, PUs, of a parallel machine. *Broadcasting*, the operation in which one processor has to send a message M to all other PUs, is a crucial building block for many parallel algorithms. Since it can be implemented once and for all in communication libraries such as MPI [7], it makes sense to invest into algorithms which are close to optimal for all P and all message lengths k. Since broadcasting is sometimes a bottleneck operation, even constant factors should be considered. In addition, by reversing the direction of communication, broadcasting algorithms can usually be turned into reduction algorithms. *Reduction* is the task to compute a generalized sum $\bigoplus_{i<P} M_i$, where initially message M_i is stored on PU i and where "$\oplus$" can be any associative operator. Broadcasting and reduction are among the most important communication primitives. For example, some of the best algorithms for matrix multiplication or dense matrix–vector multiplication have these two functions as their sole communication routines [5].

We study broadcasting long messages for a simple synchronous, symmetric communication model which is intended as a least common denominator of practical protocols able to support high bandwidth for long messages: It takes time $t + k$ to transfer a message of size k regardless which PUs are involved. This is realistic on many modern machines where network la-

* Corresponding author.
E-mail addresses: sanders@mpi-sb.mpg.de (P. Sanders), jopsi@mpi-sb.mpg.de (J.F. Sibeyn).
URLs: http://www.mpi-sb.mpg.de/~sanders, http://www.cs.umu.se/~jopsi.

0020-0190/02/$ – see front matter
doi:10.1016/S0020-0190(02)00473-8