Performance of Collective Communications on Interconnection Networks with Fat nodes and Edges

Milos Ohlidal, Jiri Jaros, Vaclav Dvorak
Brno University of Technology, Czech Republic
{ohlidal, jarosjir,dvorak}@fit.vutbr.cz

Abstract

The paper deals with scheduling collective communications in the minimum number of communication steps; it shows how to generalize the known results regarding time complexity of collective communications on common direct networks for the same networks with fat nodes and edges. Models of node architecture composed of several processor cores that share a router are discussed. Examples of communication algorithms on fat K-ring networks with 8 to 32 processors are summarized and given in detail. The results show that fat networks, depending on their configuration, can provide a range of communication performance at a lower cost.

1. Introduction

As microelectronic technology is turning to multi-core and multi-threaded processors for more performance and less power consumption, networks interconnecting such "fat" nodes are of interest. Pair-wise (point-to-point) as well as collective (group) communications involving all processors are frequently used in the solution of demanding problems in parallel and their timing complexity has a dramatic impact on performance. Since processors are connected only sparsely, the message can reach a destination processor directly, if source and destination processors are neighbors, or else through some intermediate routers. The communication time from issuing the send request by one CPU until receiving data by another CPU represents an overhead of parallel processing which has to be minimized. It does not occur in sequential processing and therefore reduces the potential speedup below P if P processors are engaged in processing.

In this paper we are going to analyze only the collective communications involving all processors: one-to-all broadcast (OAB), all-to-all broadcast (AAB), one-to-all scatter (OAS, a private message to each partner), all-to-one gather (AOG), all-to-all scatter (AAS). Given the amount of computation, the only thing that matters in obtaining the highest performance is collective communication times. Their values depend strongly, beside other parameters, on the message lengths (assumed uniform across processors in one collective communication), processor count, and topology of the network.

A class of interconnection networks of interest in this paper covers direct networks, which for performance-driven environments converge on the use of pipelined (wormhole, WH) message transmission and source-based routing algorithms. The use of more processors in one node is an interesting idea that can improve certain properties of interconnection networks. For example, expanding the hypercube vertices into cycles leads to cube-connected cycles [1] with constant node degree d=3. The diameter D=log$P$ is still logarithmic, but the scalability is limited to too few values of P (P = 24, 64, 160…). On the other hand, if the vertices of the hypercube are expanded, but now into sets of processors connected by the crossbar switch inside the router, scalability is improved. Nodes can contain any number of processors m = 1, 2, 3, etc., and $P = m2^d$. The node degree now grows more slowly than in the hypercube, d = log (P/m). Due to these favorable features has the fat cube topology been recently used in commercial DSM NUMA machine Origin 3000 (SGI). Also fat nodes with 4 Opteron processors have been used in 3D fat cube network [3] and nodes with 8 CPUs are connected into K-ring network in Swiss-T1 cluster [4].

In the rest of the paper, we want to analyze the complexity of collective communications if fat nodes with more than one processor and/or fat (multiple) edges are introduced into the network. The results concerning the special case of fat cubes are available [2], but here we will approach the problem in a general scope. Section 2 deals with collective communications

Proceedings of the International Conference on Networking, International Conference on Systems and International Conference on Mobile Communications and Learning Technologies (ICNICONSMCL'06)
0-7695-2552-0/06 $20.00 © 2006 IEEE

IEEE COMPUTER SOCIETY

on several base network topologies, which serve as candidates for getting fatter nodes. In Section 3 we discuss the router architecture and complexity of collective communications within fat nodes. Finally Section 4 gives upper complexity bounds for any topology; particular results for the K-ring network, that seems to perform best in collective communications among all other topologies, are presented as an illustration. Applications of the foregoing analysis and future possible extensions are given in Conclusions.

2. Group communications: models, lower complexity bounds and real complexity

The simplest linear time model of communication in distributed memory systems uses a number of communication steps (rounds). In distance-sensitive store-and-forward (SF) networks, one step is a set of parallel (simultaneous) hops of packets between adjacent nodes. The start-up latency t_0 is the software and hardware latency in the source and destination nodes for initializing the DMA transfer between memory and the processor. Additional serialization latency is proportional to the message length m (in bytes) and to per byte transfer time t_1. The serialization latency is in SF networks incurred in every hop between neighbor nodes, $h \times (m\ t_1)$ in total, where h is the distance (in hops) of source and destination nodes. In wormhole (WH) networks each step includes start-up latency t_0, the serialization latency mt_1 and a small router delay t_d h-times along a traversed path; the serialization latency is incurred only once per step. Several messages between source-destination pairs, not necessarily the neighbors, can proceed concurrently and can be combined into a single step if their paths are disjoint. Of course, for simplicity, we assume no contention for channels and no due delays.

Further, we will consider only bi-directional full-duplex links. The number k of bi-directional channels between the CPU and a router (ports), that can be engaged in communication simultaneously, has a decisive impact on the number of communication steps; 1-port (k=1) or all-port (k=d) models will be considered, as they are most common. We will not assume combining nodes with facility to combine/extract partial messages but exclusively the non-combining nodes that can only retransmit /consume original messages.

The topology is one of the key design factors of an interconnection network. There is a large body of theoretical research on optimal topologies, based on graph theory metrics such as average distance d_a, network diameter D, and bisection width B_C, among others. These parameters have a direct impact on network performance, Tab.1. As far as the broadcast communication (OAB) in SF network is concerned, the number of steps cannot be less than network diameter D, because this is the worst case even for point-to-point communication. For WH switching the distance between nodes is not that important and the lower bound on the number of steps $s = \lceil \log_{k+1} P \rceil$ is given by the number of nodes informed in each step, that is initially 1, $1+1\times k$ after the first step, $(k+1)+(k+1)\times k = (k+1)^2$ after the second step, etc.,…, and $(k+1)^s \geq P$ nodes after step s.

Table 1. Lower bounds on complexity of collective communications

CC	SF	WH
OAB	D	$\lceil \log_{k+1} P \rceil = \lceil (\log P)/\log(k+1) \rceil$
AAB	$\lceil (P-1)/k \rceil$	$\lceil (P-1)/k \rceil$
OAS	$\lceil (P-1)/k \rceil$	$\lceil (P-1)/k \rceil$
AAS	$\lceil S/(Pk) \rceil$	$\lceil (d/k)P^2/(2\ B_C) \rceil$

In case of AAB communication (SF or WH), since each node has to accept P–1 distinct messages, the lower bound is $\lceil (P-1)/k \rceil$ steps. A similar bound applies to OAS communication, because each node can inject into the network not more than k messages in one step; for irregular networks with non-constant node degree d we should use the lowest value of k for AAB and the source node port model for OAS. The common strategy with SF OAS is to send messages to the farthest nodes first and then pipeline them with messages to the nodes less and less remote. The optimum broadcast tree is therefore different from that for OAB. In WH OAS we use different strategy: P–1 pair-wise communications, k of them per step, must be packed into the lowest number of steps in such a way that paths traversed in the optimum broadcast tree are edge-disjoint in each step. Note also that for 1-port models the bound P-1 is valid only for networks with Hamiltonian cycle.

For AAS communication pattern each of P processor sends an individual message to each of P-1 partners. If S is the sum of the shortest distances of all node pairs, then the average distance of nodes $d_a = S/P^2$, [3]. With concurrent communication on k=d ports (all-port model), the number of communication steps for SF switching cannot be less than $S/(Pd) = P\ d_a/d$.

0-7695-2552-0/06 $20.00 © 2006 IEEE

IEEE COMPUTER SOCIETY

Another lower bound for AAS can be obtained considering that one half of messages from each processor cross the bisection and the other half do not. There will be altogether 2 $(P/2)(P/2)$ of such messages in both ways and up to $(k/d)B_C$ messages in one step, where B_C is the network bisection width [3]. This gives $(d/k)P^2/(2B_C)$ steps. This second lower bound applies to WH as well as SF switching, but is more appropriate to WH switching, since point-to-point messages (and not neighbor-to-neighbor messages as in SF switching) are considered.

In what follows we will consider only the distance-insensitive wormhole (WH) networks. Selected network topologies are depicted in Fig.1. Some of them deserve few comments. The mathematical problem of maximizing the size of a graph P for the given node degree d and the diameter D led to a so called Moore bound [5] for non-oriented graphs (bidirectional links)

$$P \le 1+d+d(d-1)+d(d-1)^2+\ldots+d(d-1)^{D-1}, \qquad (1)$$

that can be reached in only a few cases (e.g. $d = 3$, $D = 2$, $P = 10$ nodes, Fig.1a). Note that this is the only graph without a Hamiltonian path. Graphs close in size to (1) can be obtained by discrete optimization. For example AMP (A Minimum Path) topology [5] is a result of genetic graph optimization. Optimum graphs with $d = 4$ and one extra node (a system controller SC) were found for $P = 5, 8, 12, 13, 14, 32, 36, 40, 53, 64, 128, 256$. Here we will consider a modified version without the system controller, where one out of P compute nodes will act as a SC. AMP topology with 8 nodes is depicted in Fig.1g. Note that this topology is not symmetric, so that table routing has to be used. Octagon network at Fig.1 was described in ref. [6].

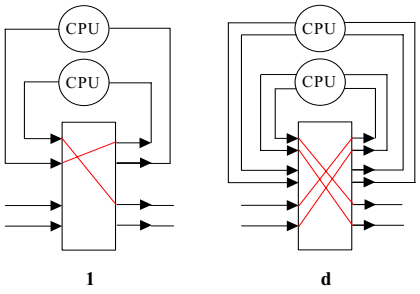


Figure 1.
Router models for fat nodes (*m*=2, *d*=2, *f* =1)
1) one-port model d) d-port (all-port) model

Among *symmetric* networks of degree four, the Midimew (Minimal Distance Mesh with Wrap-around links) has the lowest average distance and diameter, [7]. The graph at Fig. 1e can be easily redrawn in 2×4 rectangular form. Finally the K-ring [4] or a folded hypercube [1] can be constructed from a standard hypercube by connecting each node to the unique node that is most remote to it.

Table 2 gives real complexities of collective communications. Some of them are well known (hypercube), other were obtained numerically with the use of Bayesian Optimization Algorithm (BOA) and a combination of BOA and HSGA (Hybrid parallel Genetic Simulated Annealing), [8]. Fat digits denote that the theoretical lower bound has been reached.

3. Router architecture and local communication

The router model for fat nodes deserves some explanation, because it is a certain generalization of the router model used in connection with thin nodes. In the simplest case, processors are connected to the router by a single link as in Fig. 1. This so-called one-port model ("1") allows each of m or less processors to send a message either outside to a remote processor or to the local processor inside the same node. In d-port model each processor can send up to d distinct messages simultaneously, either outside or locally. In

Table 2.
Parameters and complexity of collective communications on base topologies

	P	d	D	e	S	B_C
Ladder	8	2-3	4	10	112	4
Twisted Ladder	8	2-3	3	10	100	8
Hypercube	8	3	3	12	96	8
Moore	10	3	2	15	110	10
Octagon	8	3	2	12	88	10
Midimew	8	4	2	16	80	12
AMP	8	4	2	16	80	16
K-ring	8	4	2	16	80	16
Fully connected	8	7	1	28	56	32

	OAB	AAB	OAS	AAS
Ladder	**4**	**4**	4	8
Twisted Ladder	**3**	**4**	**4**	**6**
Hypercube	2	3	3	4
Moore	2	3	3	5
Octagon	2	3	3	**4**
Midimew	2	2	2	3
AMP	2	2	2	**3**
K-ring	2	2	2	**3**
Fully connected	1	1	1	1

0-7695-2552-0/06 $20.00 © 2006 IEEE
IEEE COMPUTER SOCIETY

fact, both the models are special cases of the "*k*-port" model, where k =1 or d. In the context of thin networks (m = 1 and f = 1) these models are known as one-port and all-port models.

Because collective communication on fat networks can be visualized as composed of two parts, communication between fat nodes and inside them, we will now discuss the complexity of collective communications among crossbar-connected processors. We assume the *k*-port router model.

(OAB) Unless the router has a special broadcasting hardware, broadcast proceeds by multiplying the number of informed nodes recursively. At the beginning source processor has a message. In step 1, k new processors will be informed from the source processor, in step 2 it will be $k(k+1)$ newly informed processors plus k+1 processors informed earlier, together $(k+1)^2$, then $(k+1)^3$, etc. All m processors will thus be informed in s steps where

$$(k+1)^s \geq m \quad \text{or} \quad s = \lceil \log_{k+1} m \rceil. \qquad (2)$$

(AAB) Let us assume that each processor has a single message to share with other local processors. The fastest AAB among m nodes can be done on the router crossbar as m-1 permutations, k permutations at a time, in $\lceil (m-1)/k \rceil$ steps.

(OAS) The local OAS requires at least $\lceil (m-1)/k \rceil$ steps, because the source processor can emit not more than k messages at a time.

(AAS) The local AAS among m processors can be implemented on the router crossbar as $(m-1)$ permutations at a rate k permutations in one step, i.e. in $\lceil (m-1)/k \rceil$ steps.

4. Complexity of collective communications on fat networks

Theorem 1**.** Complexity of non-combining collective communications on a wormhole fat network with m k-port processors per node, measured by the number of communication steps, is

OAB) $\boldsymbol{\tau^{OAB}} = \tau^{OAB}(k) + \lceil \log_{k+1} m \rceil$

AAB) $\boldsymbol{\tau^{AAB}} = \tau^{AAB}(k)\lceil m/f \rceil + \lceil (P/m)(m-1)/k \rceil$

OAS) $\boldsymbol{\tau^{OAS}} = \tau^{OAS}(k)\lceil m/f \rceil + \lceil (m-1)/k \rceil$

AAS) $\boldsymbol{\tau^{AAS}} = \tau^{AAS}(k)\lceil m^2/f \rceil + \lceil (m-1)/k \rceil$,

where $\tau^{CC}(k)$ is the complexity of collective communication CC on the k-port generic thin network with P/m processors and simple edges, and f is multiplicity of external links in the fat network with m processors per node.

Proof. Collective communications CC can be thought of as composed of two parts, inter-node (global) CC among all fat nodes and then local intra-node CC within nodes. Global CC among P/m fat nodes proceeds similarly as in thin networks, except that number of messages between source and destination nodes differs according to the CC type:

OAB: 1 message only is propagated to remote nodes. In the most remote node, $m-1$ local processors remain to be informed. By making use of result (2), it will take $\lceil \log_{k+1} m \rceil$ steps.

AAB: m messages stored in m *source* node processors are transferred through intermediate nodes to m processors in a destination node during a „super-step". There are many such source-destination node pairs communicating in parallel. If we had only simple links, this giant „super-step" would take m steps, because messages from m processors would be serialized on a single link; otherwise with multiple links it will take $\lceil m/f \rceil$ steps. When the inter-node communication is over, we are left with P/m distinct messages in each node processor that are to be distributed locally. As shown in (AAB), the local AAB with 1 message per CPU will require $\lceil (m-1)/k \rceil$ steps and therefore P/m messages per CPU can be distributed in not more than $\lceil (P/m)(m-1)/k \rceil$ steps.

OAS: a single super-message from the source node to destination nodes will consist again of m messages and the number of steps to get them from one node to another can be reduced by parallel transmission from several processors if fat links (f >1) are available. The number of steps needed is given as above in the case of AAB. The local communication will require $\lceil (m-1)/k \rceil$ steps as shown in (OAS).

AAS: the block of m^2 messages (a super-message) from the source node (m source CPUs in one node, each of them sending m messages to a destination node) goes through intermediate nodes to the destination node. For k-port processors, up to k super-messages will go from one source node to k destination nodes in parallel. With multiple links it will require $\lceil m^2/f \rceil$ steps. Local AAS will be completed in $\lceil (m-1)/k \rceil$ steps as shown in (AAS), q.e.d.

Let us note that global and local CC do not have to follow one another, but can be overlapped, because only f node processors communicate simultaneously in one step, each on k internal links to the router and then on $k \leq d$ external links to the neighbor node. Remaining $m - f$ processors, different in each step,

IEEE COMPUTER SOCIETY

can engage in a local CC going on simultaneously with the global CC. From this point of view the upper bounds declared in Theorem 1 are a bit pessimistic, because they do not count on any overlap.

Efficiency of AAB (AAS) can alternatively be improved with combining messages targeted for the same node. Unpacking and distributing messages inside a node by one processor can often be faster than sending separate messages, each with a startup delay. Also, if node processors share an L2 cache, each processor can pick up its part from this cache. The whole communication consists then of two distinct parts, message passing among nodes and shared memory communication within the nodes.

As an illustration, performance of collective communication for various alternative networks derived from a single generic architecture (K-ring network) is compared in Table 3. The upper bounds of the number of communication steps according to Theorem 1 are given, based on the minimum obtained values (highlighted) for the generic thin K-ring network with 8, 16, and 32 processors. These base values were again obtained by means of evolutionary algorithms [8].

Table 3. Upper bounds on complexity of CC on the all-port WH K-ring network

P	m	f	d	D	OAB	AAB	OAS	AAS
8	1	1	4	2	2	2	2	3
8	2	1	3	1	2	4	3	5
8	2	2	3	1	2	3	2	3
16	1	1	5	3	2	3	3	5
16	2	1	4	2	3	6	5	13
16	2	2	4	2	3	4	3	7
16	4	1	3	1	2	7	5	17
16	4	2	3	1	2	5	3	9
32	1	1	6	4	2	6	6	9
32	2	1	5	3	3	10	7	21
32	2	2	5	3	3	7	4	11
32	2	4	5	3	3	7	4	6
32	4	1	4	2	3	14	9	49
32	4	2	4	2	3	10	5	25
32	4	4	4	2	3	8	3	13

5. Conclusions

Technology of fat interconnection networks has several advantages over traditional networks:
- makes some small networks more scalable, even though the interconnection graph of a network is not scalable at all (Moore, twisted ladder) or only partially scalable (Octagon, AMP);
- it provides in many cases cheaper network implementation in terms of hardware cost and often more suitable for networking systems on chip;
- performance in one-to-all collective communications OAB and OAS is comparable to generic base networks, sometimes even better;
- performance in all-to-all collective communications AAB and AAS is worse than that of base networks, but it can be controlled by multiplicity of links and by overlapping local and global communications.

Future research should select topologies with the best potential for specific (real-time) applications and/or communication patterns that could be used for network-on-chips (NoC).

6. References

[1] Zomaya, A.: Parallel and Distributed Computing Handbook. McGraw Hill, 1996.

[2] V. Dvorak : Scheduling Collective Communications on Wormhole Fat Cubes. To appear in Proc. of the 17th SBAC-PAD Symposium on Computer Architecture and High Performance Computing, IEEE Press, 2005.

[3] C.N. Keltcher et al.: The AMD Opteron Processor for Multiprocessor Servers. IEEE Micro, March/April 2003, pp.66 – 76.

[4] Gabrielyan, E., Hersch, R.D.: Efficient Liquid Schedule Search Strategies for Collective Communications. Proc. of ICON 2004 - 12th IEEE International Conference on Networks, Singapore, Vol. 2, November 16-19, 2004, pp 760-766.

[5] Chalmers, A.-Tidmus, J.: Practical Parallel Processing. International Thomson Computer Press, 1996.

[6] Karim, F. – Nguyen, A.: An Interconnect Architecture for Networking Systems on Chips. IEEE Micro, Sept. – Oct. 2002, pp. 36-45.

[7] Puente, V. et al.: Improving Parallel System Performance by Changing the Arrangement of the Network Links. Proc. of the International Conference on Super-computing, May 2000, p.44-53.

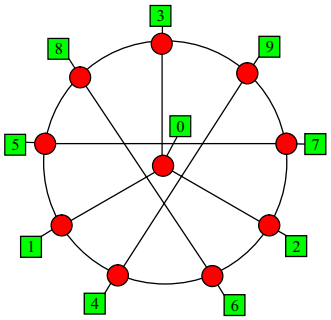
[8] Jaroš, J. - Ohlídal, M. - Dvořák, V.: Evolutionary Design of Group Communication Schedules for Interconnection Networks. Lecture Notes in Computer Sciences 3733, Berlin, DE, Springer, 2005, s. 472-481, ISBN 1122-3344.

Acknowledgement

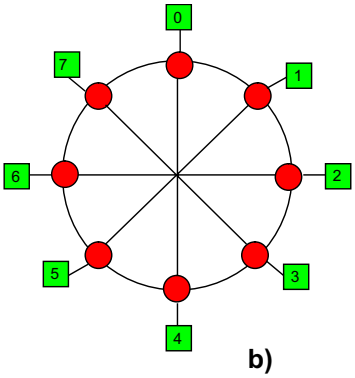
This research has been carried out under the financial support of the research grant "Network Architectures of Embedded Systems Networks", GA102/05/0467, Grant Agency of Czech Republic, 2005-2007.

0-7695-2552-0/06 $20.00 © 2006 IEEE

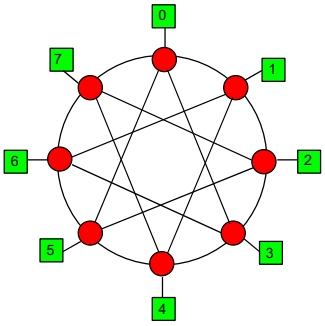
IEEE COMPUTER SOCIETY



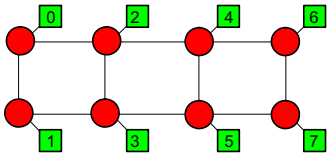
a)



b)

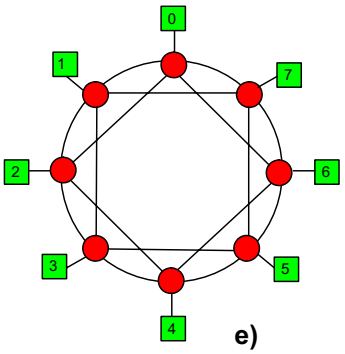


c)

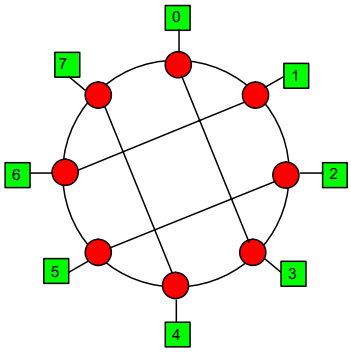


d)

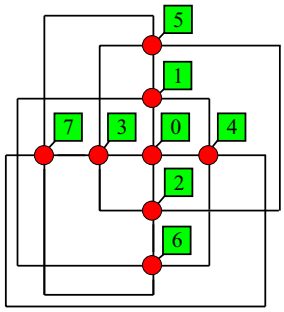
Fig.1. Generic networks with 8 (10) nodes. a) Moore graph, b) Octagon, c) K-ring, d) Ladder



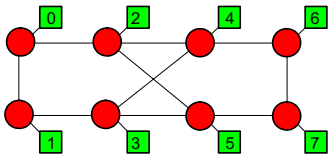
e)



f)



g)



h)

Fig.1. Generic networks with 8 nodes. e) Midimew, f) HyperCube, g) AMP, h) Twisted Ladder

Proceedings of the International Conference on Networking, International Conference on Systems and International Conference on Mobile Communications and Learning Technologies (ICNICONSMCL'06)
0-7695-2552-0/06 $20.00 © 2006 IEEE

IEEE COMPUTER SOCIETY