SOME ISSUES IN HIERARCHICAL INTERCONNECTION NETWORK DESIGN

Sivarama P. Dandamudi and Derek L. Eager

SCR-TR-156, April 1989

School of Computer Science, Carleton University
Ottawa, Canada, KIS 5B6

Some Issues in Hierarchical Interconnection Network Design

Sivarama P. Dandamudi
Center for Parallel and Distributed Computing
School of Computer Science
Carleton University
Ottawa, Ontario, K1S 5B6
Canada

Derek L. Eager
Department of Computational Science
University of Saskatchewan
Saskatoon, Saskatchewan, S7N 0W0
Canada

ABSTRACT

Hierarchical interconnection networks (HINs) have been proposed as a cost-effective way to interconnect large numbers of processors in a multicomputer system. This paper considers two issues concerning the design of binary-hypercube based HINs -- the optimum cluster size and the optimum number of levels in the hierarchy. The optimum cluster size is dependent on the underlying degree of locality in communication. For all of the various locality characterizations that we consider, the optimum cluster size in a two-level binary-hypercube based HIN is shown to be either 4 or 8, over a wide range of network sizes. It is shown that increasing the number of levels to three yields at most a moderate improvement in the cost-benefit ratio, depending on the type of locality characterization considered. For some locality characterizations, the cost-benefit ratio deteriorates. Thus, an appropriate design for a binary-hypercube based HIN appears to be a two-level hierarchy, with a cluster size of either 4 or 8.

Key words: Binary hypercubes, Hierarchical networks, Interconnection networks, Multicomputers.

1. Introduction

Hierarchical interconnection networks (HINs) have been proposed as a cost-effective way to interconnect large numbers of processors in a distributed-memory MIMD (multicomputer) system [Dandamudi & Eager 1987]. HINs reduce the link cost substantially at the expense of moderately increasing message delivery times. The structure of HINs can be informally described as follows. The N nodes in the system are grouped into K_1 clusters of $n_1 = N/K_1$ nodes each. It is assumed here, for convenience, that K_1 evenly divides N, and that each cluster is of identical size, although in practice there would be no reason not to permit clusters of varying sizes. Each cluster of n_1 nodes is linked together internally by a level 1 interconnection network. One node from each cluster acts as an interface node to the rest of the system. These K_1 interface nodes may be linked by a single level 2 interconnection network, or they may be themselves grouped into K_2 clusters of $n_2 = K_1/K_2$ nodes each, with each cluster linked together internally by a separate level 2 network. In this later case, one node from each level 2 cluster is selected as a level 2 interface node, and linked with other interface nodes by a level 3 network, and so on. Figure 1 shows an example HIN with two levels. This HIN is denoted as a "BH/BH" HIN since both levels use a binary hypercube (BH) network. (For details on the structure of other hierarchical interconnection networks for shared-memory multiprocessor systems and distributed-memory multicomputer systems, see [Swan *et al.* 1977, Gajski *et al.* 1983, Agrawal & Mahgoub 1985, Carlson 1985, Hwang & Ghose 1987].)

An advantage of HINs is that they reduce the degree (i.e., the number of links connected to a node) of the majority of nodes. The degree of the remaining nodes is the same as that of the corresponding non-hierarchical network. For example, consider a binary hypercube network with $N = 2^D$ nodes. The degree of each node in this network is D (assuming full-duplex links). In contrast, in a two-level BH/BH HIN with a cluster size of $n = 2^d$, only $\frac{N}{2^d}$ nodes will have a degree of D (the interface nodes), and the remaining nodes will have only a degree of d. If $N = 1024$ (i.e., $D = 10$) and $n = 16$ (i.e., $d = 4$), then the BH/BH HIN will have 960 nodes with degree 4 and 64 nodes with degree 10. This is in contrast to a degree of 10 for all 1024 nodes in the BH network.

This aspect of HINs is important in system implementation. For example, consider the NCUBE/ten system [Hayes *et al.* 1986], which organizes 1024 nodes (each consisting of a processor and memory) as a binary hypercube. Each node in this system has a total of 20 half-duplex link connections. Using custom-made VLSI chips, the designers could pack as many as 64 nodes on a single 16" × 22" printed-circuit board (PCB). The total system consists of 16 such PCBs, and inter-PCB connections require as many as 512 wires from each PCB. (The NCUBE/ten actually uses 640 wires to allow connections to I/O devices.) The use of the BH/BH HIN (with $d = 6$ and $D = 10$) would greatly reduce the number of wires needed for inter-PCB connection (from 512 wires to 8 wires). This reduction in the number of link connections has two implications on system design. First, it reduces the demand for both chip area and PCB area. Thus, more nodes can be packed in a given PCB area. Second, since inter-PCB connection is greatly simplified, it is feasible to increase the system size N substantially. This second factor is important in implementing large parallel systems.

This paper considers two issues concerning the design of binary-hypercube based HIN -- the optimum cluster size and the optimum number of levels in the hierarchy. This requires measures of both network cost and network performance. The *link cost, L,* as given by the number of links, is used to measure the network cost. Since bus-based structures are not treated here, the number of link connections need not be considered. The performance of the network is measured by the *average internode distance, P*, as given by the average number of links that must be traversed by a message.

In general, trying to minimize one of the above two measures results in an increase in the other. A useful measure is the *LP product,* defined as $L \times P$. The *LP product* can be interpreted as a cost-benefit ratio (with L representing the cost of the network and $1/P$ representing the benefit). Thus, benefit per unit cost can be maximized by minimizing the *LP product.* Similar measures have been used by other researchers [Bhuyan & Agrawal 1982, Agrawal *et al.* 1986, McCrosky 1986].

Rather than working in terms of raw *LP product* values, here we use values relative to those of a "reference" network. The *LP ratio* measure is defined as follows:

$$LP \ \ ratio = \frac{L_H \times P_H}{L_R \times P_R}$$

where the subscript 'H' refers to the HIN under consideration and the subscript 'R' refers to the reference network. Smaller *LP ratios* are preferred. In this paper, the non-hierarchical binary hypercube is used as the reference network.

The *LP ratio* is dependent on the underlying degree of locality in communication. Locality can be characterized in several ways. Let $\phi(l)$ denote the probability that a message is destined to some specific node that is at a distance of l links from the source node in the reference network, l_{max} denote the maximum internode distance in the reference network, and *NumNodes* (l) denote the number of nodes at a distance of l links from a source node in the reference network. Then, the average internode distance, P, in a network of interest, is given by

$$P = \sum_{l=1}^{l_{max}} d(l\ )\ \phi(\ l)\ NumNodes\,(l\ ) \tag{1}$$

where $d(l)$ denotes the average distance from a source node to a node that would be of distance l from the source node in the reference network. Different choices for $\phi(l)$ reflect differing degrees and types of locality, and lead to different average internode distances and, in turn, to different values of the *LP ratio*. The following five types of distributions for $\phi(l)$ are considered here:

(*i*) Uniform (UNIF)

(*ii*) Decreasing probability (DPF)

(*iii*) Reed's sphere of locality (R-SOL)

(*iv*) Sphere of locality (SOL)

(*v*) Generalized sphere of locality (G-SOL)

The first four types of distributions have been used previously in the literature [Wu & Liu 1981, Reed 1983, Dandamudi & Eager 1987]. The fifth is proposed here as a generalization of SOL. For each type of locality distribution, expressions can be derived for the average internode distance, in terms of the parameters of the distribution parameters.

These expressions are complex, however, and thus optimum network designs are found numerically in the following sections.

The paper is organized as follows. Section 2 presents the optimum cluster size analysis for a two-level binary hypercube based HIN, the BH/BH network. Section 3 considers the impact of increasing the number of levels in the hierarchy. Section 4 concludes the paper by summarizing the results.

2. The Optimum Cluster Size for the BH/BH Network

Let the total number of nodes be $N = 2^D$, and the number of nodes in a level 1 cluster of the BH/BH network be $n = 2^d$. Then L_{BH} and $L_{BH/BH}$ are given by:

$$L_{BH} = D\, 2^{D-1} \tag{2}$$

$$L_{BH/BH} = d\, 2^{D-1} + (D-d)\, 2^{D-d-1} \tag{3}$$

To derive the desired LP ratios, it remains to find expressions for the average internode distances in the BH and BH/BH networks. Let P_Y^X denote the average internode distance in network Y with a communication distribution of type X. For example, $P_{BH/BH}^{UNIF}$ represents the average internode distance in a BH/BH network with uniform communication. The following five subsections derive the average internode distances for the BH and BH/BH networks with each of the five communication distributions that are considered here. Finally, in Section 2.6, these results are used to derive optimum cluster sizes.

2.1. Uniform

In uniform communication, each pair of distinct nodes i and j exchange messages at an identical rate. The assumption of uniform communication is appealing because one would expect real computations to be characterized by some degree of locality, implying that this assumption should result in an upper bound on path lengths. For the BH network with $N = 2^D$ nodes and uniform communication, $\phi(l)$ is given by

$$\phi(l) = \frac{1}{2^D - 1} \qquad 1 \leq l \leq D \tag{4}$$

Then,

$$P_{BH}^{UNIF} = \frac{1}{2^D - 1}\sum_{l=1}^{D}\binom{D}{l} l = \frac{D2^{D-1}}{2^D - 1} \tag{5}$$

For the BH/BH network, the average internode distance is given by

$$P_{BH/BH}^{UNIF} = P'_{BH/BH} + P''_{BH/BH} \tag{6}$$

where

$P'_{BH/BH}$ = that portion of the average internode distance contributed by messages whose source and destination nodes are within the source cluster, and

$P''_{BH/BH}$ = that portion of the average internode distance contributed by messages whose source and destination nodes are in two different clusters.

It is straightforward to derive $P'_{BH/BH}$ as

$$P'_{BH/BH} = \left(\frac{2^d - 1}{2^D - 1}\right)\left[\frac{d2^{d-1}}{2^d - 1}\right]$$

where the first factor gives the fraction of messages that travel within a single cluster and the second factor gives the average internode distance travelled by these messages. The fraction of messages whose source and destination nodes are in two different clusters is $\left(1 - \frac{2^d - 1}{2^D - 1}\right)$. Since each such message travels on average $\frac{d}{2}$ links to reach the source cluster interface node, $\frac{d}{2}$ links from the destination cluster interface node to the final destination node, and $\frac{(D-d)2^{D-d-1}}{2^{D-d} - 1}$ links from the source cluster interface node to the destination cluster interface node, we obtain

$$P_{BH/BH}^{UNIF} = \left(\frac{2^d - 1}{2^D - 1}\right)\left[\frac{d2^{d-1}}{2^d - 1}\right] + \left(1 - \frac{2^d - 1}{2^D - 1}\right)\left[d + \frac{(D-d)2^{D-d-1}}{2^{D-d} - 1}\right]$$

2.2. Decreasing Probability

With this type of communication locality, the probability that a given message is destined for a node at a distance of l links from the source node decreases as a smooth function of l. A wide variety of potentially suitable distribution functions exhibiting such behaviour exist. Reed has proposed [Reed 1983]

$$\phi(l) = \frac{Decay\ (a,\ l_{max})}{NumNodes\ (l\ )} a^l \qquad 0 < a < 1, \quad 1 \leq l \leq D$$

where $Decay\ (a,\ l_{max})$ is a normalizing constant chosen to ensure that the $\phi(l)$ sum to one. As a approaches one, communication becomes uniform, as in the previous section. Conversely, as a approaches zero, communication becomes nearest neighbour only (communication will occur only between directly connected nodes). Choices of a between these two extremes lead to varying degrees of communication locality.

For the BH network,

$$Decay\ (a,\ D) = \frac{1 - a}{a\,(1 - a^D)}$$

and

$$\phi(l) = \left[\frac{1 - a}{a(1 - a^D)}\right] \frac{a^l}{\binom{D}{l}} \tag{7}$$

Therefore,

$$\begin{aligned} P_{BH}^{DP} &= \left[\frac{1 - a}{a\,(1 - a^D)}\right] \sum_{l=1}^{D} l\ a^l \\ &= \frac{Da^{D+1} - (D\ + 1)\ a^D + 1}{(1 - a)\left[1 - a^D\right]} \end{aligned} \tag{8}$$

For the BH/BH network, the average internode distance is given by

$$P_{BH\ /\ BH}^{DP} = P'_{\ BH\ /\ BH} + P''_{\ BH\ /\ BH} \tag{9}$$

where $P'_{\ BH\ /\ BH}$ and $P''_{\ BH\ /\ BH}$ are defined as in Eq. (6).

Considering first $P'_{\ BH\ /\ BH}$, it is easy to derive that

$$P'_{BH/BH} = \sum_{l=1}^{d} \binom{d}{l} l \; \phi(l) \tag{10}$$

In computing $P''_{BH/BH}$, the address of a node is divided into two parts; the most significant (D-d) bits giving the cluster identity, and the least significant d bits identifying a node within a cluster. The distance between a source node and a destination node is broken up into four terms: i+j+k+m. The term i represents the distance from the source node to the source cluster interface node, and j represents the distance in the level 2 network. The least significant d bits of the destination node address are further divided into two groups: one group consists of those bits of the destination node address that are identical to the corresponding bits of the source node address, and yet that differ from those of the source node cluster interface node, and the other group consists of the rest of the least significant bits. The term, k represents the distance contributed by the former group of address bits, and m represents the distance contributed by the later group. Recall that $\phi(l)$ is interpreted in terms of the distances within the reference network (BH). A distance $i + j + k + m$ in the BH/BH network thus corresponds to a distance of $j + k + m$ in the BH network. $P''_{BH/BH}$ can then be derived as

$$P''_{BH/BH} = \frac{1}{2^d} \sum_{i=0}^{d} \sum_{j=1}^{D-d} \sum_{k=0}^{i} \sum_{m=0}^{d-i} (i + j + k + m) \binom{d}{i} \binom{D-d}{j} \binom{i}{k} \binom{d-i}{m} \phi(j + k + m) \tag{11}$$

2.3. Reed's Sphere of Locality

A simpler model of locality places each node at the centre of a "sphere of locality" [Reed 1983, Reed & Schwetman 1983, Reed & Grunwald 1987]. A given message is destined for a node within the source node's sphere of locality with some high probability α, and to a node outside the sphere of locality with the low probability (1- α). A message is destined for each node within the same category (either within or exterior to the sphere of locality) with identical probability. Let L denote the maximum number of links a message may cross and remain in the sphere of locality centered at its source (i.e., L is the radius of the sphere). The number of nodes contained in a sphere of locality for the BH network is $\sum_{l=1}^{L}\binom{D}{l}$.

Then,

$$\phi(l)=\begin{cases}\dfrac{\alpha}{\sum_{l=1}^{L}\binom{D}{l}} & 1\leq l\leq L\\[2ex] \dfrac{1-\alpha}{2^{D}-\sum_{l=0}^{L}\binom{D}{l}} & L< l\leq D\end{cases}\tag{12}$$

P_{BH}^{R-SOL} is given by

$$P_{BH}^{R-SOL}=\frac{\alpha\sum_{l=1}^{L}\binom{D}{l}l}{\sum_{l=1}^{L}\binom{D}{l}}+\frac{(1-\alpha)\sum_{l=L+1}^{D}\binom{D}{l}l}{\sum_{l=L+1}^{D}\binom{D}{l}}\tag{13}$$

$P_{BH/BH}^{R-SOL}$ is given by Eqs. (9), (10) and (11) by using Eq. (12) for $\phi(l)$.

2.4. Sphere of Locality

Another similar but distinct type of locality characterization (SOL) divides the total number of nodes N into disjoint groups of 2^{s} nodes each [Wu & Liu 1981, Dandamudi & Eager 1987]. All nodes within a group share the same sphere of locality, as comprised of the nodes within that group. (In contrast, in R-SOL each node has its own distinct sphere of locality, and this sphere overlaps with other spheres). A message is destined for a node within the same sphere of locality as the source node with some high probability α, and to a node in a different sphere of locality with probability $(1-\alpha)$. A message is destined for each node within the same category (either within or exterior to the sphere of locality) with identical probability.

P_{BH}^{SOL} is given by (similar to Eq. (6)),

$$P_{BH}^{SOL} = \alpha\left[\frac{s\,2^{s-1}}{2^{s}-1}\right] + (1-\alpha)\left[\frac{D\,2^{D-1} - s\,2^{s-1}}{2^{D}-2^{s}}\right] \qquad (14)$$

For the BH/BH network, the average internode distance is given by

$$P_{BH/BH}^{SOL} = \alpha\,P_{iSOL} + (1-\alpha)\,P_{oSOL} \qquad (15)$$

where

P_{iSOL} = average distance between two nodes in the same sphere of locality, and

P_{oSOL} = average distance between two nodes in different spheres of locality.

Consider the following two cases: $(i)\ d \le s$, and $(ii)\ d \ge s$.

Case (i): $d \le s$

Since $d \le s$, the number of clusters within a sphere of locality is 2^{s-d}. In this case, P_{iSOL} can be derived as

$$P_{iSOL} = \beta\,P_{CL\text{-}avg} + (1-\beta)\left[2P'_{CL\text{-}avg} + P_{NCL\text{-}avg}\right]$$

where

β = fraction of messages between two nodes in the same sphere of locality that also lie within the same cluster;

$P_{CL\text{-}avg}$ = avearge internode distance in a cluster;

$P'_{CL\text{-}avg}$ = avearge number of links traversed in the source (or destination) cluster by messages whose source and destination nodes lie within the same sphere of locality, but in different clusters;

$P_{NCL\text{-}avg}$ = avearge number of links traversed by messages whose source and destination nodes lie within the same sphere of locality, but in different clusters.

It is straightforward to derive that

$$\beta = \frac{2^{d}-1}{2^{s}-1},\quad P_{CL\text{-}avg} = \frac{d\,2^{d-1}}{2^{d}-1},\quad P'_{CL\text{-}avg} = \frac{d}{2},\ \text{and}\ P_{NCL\text{-}avg} = \frac{(s-d)2^{s-d-1}}{2^{s-d}-1}.$$

Therefore,

$$P_{iSOL}=\left[\frac{d2^{d-1}}{2^s-1}\right]+\left[\frac{2^s-2^d}{2^s-1}\right]\left\{d+\frac{(s-d)2^{s-d-1}}{2^{s-d}-1}\right\} \tag{16}$$

P_{oSOL} can be derived similarly and is given by

$$P_{oSOL}=d+\frac{(D-s)2^{D-s-1}}{2^{D-s}-1} \tag{17}$$

Case (ii): $d \geq s$

For this case,

$$P_{iSOL}=\frac{s2^{s-1}}{2^s-1} \tag{18}$$

and

$$\begin{aligned} P_{oSOL} &= \left[\frac{2^d-2^s}{2^D-2^s}\right]\left\{\frac{d2^{d-1}-s2^{s-1}}{2^d-2^s}\right\}+\left[1-\frac{2^d-2^s}{2^D-2^s}\right]\left\{2\frac{d}{2}+\frac{(D-d)2^{D-d-1}}{2^{D-d}-1}\right\} \\ &= \left[\frac{d2^{d-1}-s2^{s-1}}{2^D-2^s}\right]+\left[\frac{2^D-2^d}{2^D-2^s}\right]\left[d+\frac{(D-d)2^{D-d-1}}{2^{D-d}-1}\right] \end{aligned} \tag{19}$$

2.5. Generalized Sphere of Locality

The generalized sphere of locality (G-SOL) type of distribution provides a characterization of locality that is intermediate in detail between that of SOL and R-SOL and that of DP. The G-SOL type of distribution generalizes SOL by allowing more than two "layers" of locality. The total number of nodes is divided into disjoint spheres of locality, which are then each divided into subspheres of locality, and so on. Each node is in a locality containing only itself, said to be a layer 0 locality. A message from a node is destined to some node within its containing layer i locality, but not within its containing layer i-1 locality, with probability α_i (and to each node within this layer with equal probability); locality implies that $\alpha_j \geq \alpha_k$ for $0 < j < k$. Since a node may not send messages to itself in this model, $\alpha_0 = 0$.

For the BH based networks under consideration here, it is convenient to assign all nodes with matching most significant (D-i) address bits to the same layer i locality. $F(i)$ will be used to denote the probability that a message is destined for some node in the source node's containing level i locality. From the definition of α_k, $F(i) = \sum_{k=0}^{i} \alpha_k$.

As an example of how this characterization of communication locality could be applied, consider nearest-neighbour communication in which each node communicates with its four neighboring nodes in a two-dimensional torus with $N = 2^D$ nodes. If D is even, then the torus is $2^{D/2} \times 2^{D/2}$; if D is odd, it is $2^{(D+1)/2} \times 2^{(D-1)/2}$. In such a structure, $F(i)$ can be computed as follows, assuming again that layer i localities are to contain 2^i nodes. If i is even, a layer i locality represents a square mesh of $2^{i/2} \times 2^{i/2}$; if i is odd, a layer i locality represents a mesh of $2^{(i+1)/2} \times 2^{(i-1)/2}$. Then, it is easy to derive $F(i)$, $0 < i < D$, as

$$F(i) = \begin{cases} 1 - \dfrac{1}{\sqrt{2^i}} & \text{if } i \text{ is even} \\ 1 - \dfrac{1.5}{\sqrt{2^{i+1}}} & \text{if } i \text{ is odd} \end{cases} \qquad (20)$$

Some sample $F(i)$ values are shown in Table 1 for a 1024-node torus. It should be noted that this characterization of communication locality only approximately represents the true locality (nearest-neighbour communication) that is present.

Table 1 Example *F*(*i*) values for a 1024-node torus

i	number of nodes in a layer i locality $(= 2^i)$	$F(i)$
1	2	0.25
2	4	0.5
3	8	0.625
4	16	0.75
5	32	0.8125
6	64	0.875
7	128	0.90625
8	256	0.9375
9	512	0.953125
10	1024	1

With the G-SOL type of distribution, P_{BH}^{G-SOL} can be derived as

$$P_{BH}^{G-SOL} = \sum_{i=1}^{D} [F(i) - F(i-1)]\, P_{avg}(i) \tag{21}$$

where $P_{avg}(i)$ is the average internode distance between a source node and the nodes that are in the containing layer i locality but not in the containing layer i-1 locality. Note that such nodes must have addresses that differ in the i-1st least significant bit in comparison to the source node address, but whose (D-i) most significant bits match those of the source node address. Since the number of potential destination nodes at a distance (j+1) from the source node, where j ($j \geq 0$) is contributed by the least significant (i-1) bits and the '1' is contributed by the (i-1)th least significant bit, is equal to $\binom{i-1}{j}$, $P_{avg}(i)$ is given by

$$P_{avg}(i) = \frac{\sum_{j=0}^{i-1}\left\{\binom{i-1}{j}(j+1)\right\}}{2^{i} - 2^{i-1}} = \frac{i+1}{2} \tag{22}$$

For the BH/BH network, the average internode distance is given by

$$P_{BH/BH}^{G-SOL} = P'_{BH/BH} + P''_{BH/BH} \tag{23}$$

where

$P'_{BH/BH}$ = that portion of the average internode distance contributed by messages whose source and destination nodes are within the source cluster, and

$P''_{BH/BH}$ = that portion of the average internode distance contributed by messages whose source and destination nodes are in two different clusters.

$P'_{BH/BH}$ and $P''_{BH/BH}$ can be derived as

$$P'_{BH/BH} = \sum_{i=1}^{d}[F(i) - F(i-1)]\, P_{avg}(i) \tag{24}$$

and,

$$P''_{BH/BH} = \sum_{i=d+1}^{D}[F(i) - F(i-1)]\, P''_{avg}(i) \tag{25}$$

where $P''_{avg}(i)$ is defined in a similar manner as $P_{avg}(i)$, but is conditioned on source and destination nodes being in distinct clusters. Since all destination nodes have to be nodes that are in the layer i locality containing the source node, but not in the layer i-1 locality, the ith address bit of all these destination nodes should differ from the corresponding bit in the address of the message source node. The number of clusters composed of such nodes is 2^{i-d-1}. Thus, $P''_{avg}(i)$ is given by

$$P''_{avg}(i) = \frac{1}{2^{d}2^{i-d-1}2^{d}} \sum_{j=0}^{d} \sum_{k=0}^{i-d-1} \sum_{m=0}^{d} \left[\binom{d}{j}\binom{i-d-1}{k}\binom{d}{m}(j+k+m+1)\right]$$

where j represents the number of links travelled by a message to reach the interface node in the source cluster, k represents the number of links travelled in the level 2 network and m represents the number of links travelled from the interface node in the destination cluster to the destination node. The '1' represents the distance contributed by the ith address bit. This expression can be simplified to yield

$$P''_{avg}(i) = \frac{d+i+1}{2} \tag{26}$$

Combining Eqs. (22) through (26) yields

$$P^{G-SOL}_{BH/BH} = \sum_{i=1}^{d}[F(i) - F(i-1)](i+1)/2 + \sum_{i=d+1}^{D}[F(i) - F(i-1)](d+i+1)/2 \tag{27}$$

2.6. Discussion of Results

Figures 2-6 show results derived from the preceding analyses concerning optimum cluster size in the BH/BH network, for the five types of communication distributions studied. For each type of distribution, the system size is varied from 2^7 to 2^{16}. Each figure consists of two graphs, one gives optimum cluster sizes for varying system sizes, and the other gives the *LP ratios* when those optimum cluster sizes are utilized. (Note that clusters were constrained to be of size 2^d for some *d*.)

These results show that, in all cases, the optimum cluster size varies only between 4 and 8 over the range network sizes considered. This small variation can be explained by the fact that the number of links, $L_{BH/BH}$, is minimized for these cluster sizes over a wide range of system sizes, as is now shown. Let $2^{d_{opt}}$ be the cluster size that minimizes $L_{BH/BH}$. By differentiating $L_{BH/BH}$ as given by Eq. (3) and setting the resulting expression to zero, one obtains

$$\begin{aligned} d_{opt} &= D + \frac{1 - 2^{d_{opt}}}{\ln 2} \\ &\approx D + 1.4423 - 1.4423\ 2^{d_{opt}} \end{aligned}$$

For values of *D* ranging from 7 to 9, d_{opt} (when constrained to be integral) is 2; for values ranging from 9 to 16, d_{opt} is 3. (When *D* is 9, d_{opt} has two possible values.) As the figures show, the impact of various types of locality characterizations is relatively minor in comparison to the impact of the number of links. These plots also show that the *LP ratio* obtained with uniform message routing gives a reasonable estimate of the *LP ratio* obtained with the other locality characterizations considered in this section.

The queueing analysis in [Dandamudi & Eager 1987] has shown that the larger the cluster size in the BH/BH network, the larger the message transmission capacity that is required for the level 2 links. This would eliminate large cluster sizes from the set of feasible designs. Fortunately, the analysis done in this section suggests that the optimum cluster size in this network is 4 or 8, which is well within the range of feasible cluster sizes.

Abraham and Davidson [1986] consider optimum cluster size issue in two-level hierarchical interconnection networks for shared-memory multiprocessor systems. Their conclusions are somewhat different from the results presented here mainly because they obtain optimum cluster sizes that minimize the communication delay whereas here we are minimizing the *LP ratio* which represents a cost-benefit ratio.

3. Increasing the Number of Hierarchy Levels

This section concerns the issue of the optimum number of levels in a binary-hypercube based HIN. Specifically, the impact of increasing the number of levels from two to three is investigated. Let 2^{d1} be the number of nodes in a level 1 cluster and 2^{d2} be the number of nodes in a level 2 cluster. In this case, the level 3 network contains 2^{d3} nodes, where $d3 = D - d1 - d2$. Then, the number of links in the BH/BH/BH (or 3BH for short) network is given by

$$L_{3BH} = d1 2^{D-1} + d2 2^{D-d1-1} + d3 2^{d3-1} \tag{28}$$

P_{3BH} is given by:

$$P_{3BH} = P'_{3BH} + P''_{3BH} + P'''_{3BH} \tag{29}$$

where

P'_{3BH} = that portion of the average internode distance contributed by messages whose source and destination nodes are within the same level 1 cluster;

P''_{3BH} = that portion of the average internode distance contributed by messages whose source and destination nodes are in two different level 1 clusters that belong to the same level 2 cluster;

P'''_{3BH} = that portion of the average internode distance contributed by messages whose source and destination nodes belong to different level 2 clusters.

The remainder of this section derives expressions for P_{3BH}, for each locality characterization.

3.1. Uniform

With uniform communication, the probability that the source and destination nodes of a message are within the same level 1 cluster is $\frac{2^{d1}-1}{2^{D}-1}$, and the probability that the source and destination nodes of a message are in two different level 1 clusters that belong to the same level 2 cluster is $\frac{[2^{d2}-1]2^{d1}}{2^{D}-1}$.

It is straightforward then to derive the following:

$$P^{UNIF}_{3BH} = \left\{\frac{2^{d1}-1}{2^{D}-1}\right\}\left[\frac{d1\,2^{d1-1}}{2^{d1}-1}\right] + \left\{\frac{[2^{d2}-1]2^{d1}}{2^{D}-1}\right\}\left[d1 + \frac{d2\,2^{d2-1}}{2^{d2}-1}\right] + \left\{1 - \left\{\frac{2^{d1}-1}{2^{D}-1}\right\} - \left\{\frac{[2^{d2}-1]2^{d1}}{2^{D}-1}\right\}\right\}\left[d1 + d2 + \frac{d3\,2^{d3-1}}{2^{d3}-1}\right] \quad (28)$$

3.2. Decreasing probability and Reed's sphere of locality

P'_{3BH} and P''_{3BH}, as defined above, are given by (from Eqs. (10) and (11))

$$P'_{3BH} = \sum_{l=1}^{d1}\binom{d1}{l}\, l\ \phi(l)$$

and

$$P''_{3BH} = \frac{1}{2^{d1}}\sum_{i=0}^{d1}\sum_{j=1}^{d2}\sum_{k=0}^{i}\sum_{m=0}^{d1-i}(i+j+k+m)\binom{d1}{i}\binom{d2}{j}\binom{i}{k}\binom{d1-i}{m}\phi(j+k+m)$$

where $\phi(l)$ is given by Eq. (7) for DP, and by Eq. (12) for R-SOL. P'''_{3BH} can be derived through a similar argument as that used to derive $P''_{BH/BH}$, as

$$P'''_{3BH} = \frac{1}{2^{d1+d2}}\sum_{i=0}^{d1}\sum_{j=0}^{d2}\sum_{k=1}^{d3}\sum_{m=0}^{j}\sum_{n=0}^{d2-j}\sum_{x=0}^{i}\sum_{y=0}^{d1-i}(i+j+k+m+n+x+y) \binom{d1}{i}\binom{d2}{j}\binom{d3}{k}\binom{j}{m}\binom{d2-j}{n}\binom{i}{x}\binom{d1-i}{y}\phi(k+m+n+x+y)$$

These component internode distances can be used in Eq. (29) to compute P^{DP}_{3BH} and P^{R-SOL}_{3BH}.

3.3. Sphere of locality

For the 3BH network, the average internode distance P^{SOL}_{3BH} is given by Eq. (15). Consider the following three cases: (*i*) $s \le d1$ (*ii*) $d1 \le s \le d1 + d2$, and (*iii*) $d1 + d2 \le s \le D$. For each case, P_{iSOL} and P_{oSOL} are obtained as in Section 2.4.

Case (i): $s \le d1$

$$P_{iSOL} = \frac{s\, 2^{s-1}}{2^{s} - 1}$$

$$P_{oSOL} = \left[\frac{d1 2^{d1-1} - s\, 2^{s-1}}{2^{D} - 2^{s}}\right] + \left[\frac{2^{d1+d2} - 2^{d1}}{2^{D} - 2^{s}}\right]\left\{d1 + \frac{d2 2^{d2-1}}{2^{d2} - 1}\right\} + \left[\frac{2^{D} - 2^{d1+d2}}{2^{D} - 2^{s}}\right]\left\{d1 + d2 + \frac{d3\, 2^{d3-1}}{2^{d3} - 1}\right\}$$

Case (ii): $d1 \le s \le d1 + d2$

$$P_{iSOL} = \left[\frac{d1\, 2^{d1-1}}{2^{s} - 1}\right] + \left[\frac{2^{s} - 2^{d1}}{2^{s} - 1}\right]\left\{d1 + \frac{(s - d1)\, 2^{s-d1-1}}{2^{s-d1} - 1}\right\}$$

$$P_{oSOL} = \left[\frac{2^{d1+d2} - 2^{s}}{2^{D} - 2^{s}}\right]\left\{d1 + \frac{(d1 + d2 - s)\, 2^{d1+d2-s-1}}{2^{d1+d2-s} - 1}\right\} + \left[\frac{2^{D} - 2^{d1+d2}}{2^{D} - 2^{s}}\right]\left\{d1 + d2 + \frac{d3\, 2^{d3-1}}{2^{d3} - 1}\right\}$$

Case (iii): $d1 + d2 \le s \le D$

$$P_{iSOL} = \left[\frac{d1\, 2^{d1-1}}{2^{s} - 1}\right] + \left[\frac{2^{d1+d2} - 2^{d1}}{2^{s} - 1}\right]\left\{d1 + \frac{d2\, 2^{d2-1}}{2^{d2} - 1}\right\} + \left[\frac{2^{s} - 2^{d1+d2}}{2^{s} - 1}\right]\left\{d1 + d2 + \frac{(s - d1 - d2)\, 2^{s-d1-d2-1}}{2^{s-d1-d2} - 1}\right\}$$

$$P_{oSOL} = d1 + d2 + \frac{(D - s)\, 2^{D-s-1}}{2^{D-s} - 1}$$

3.4. Generalized sphere of locality

In this case, P'_{3BH} and P''_{3BH} are obtained (as in Section 2.5) as

$$P'_{3BH} = \sum_{i=1}^{d1} [F(i) - F(i-1)]\left[\frac{i+1}{2}\right]$$

and

$$P''_{3BH} = \sum_{i=d1+1}^{d1+d2} [F(i) - F(i-1)]\left[\frac{d1+i+1}{2}\right]$$

P'''_{3BH} is given by

$$P'''_{BH} = \sum_{i=d1+d2+1}^{D} [F(i) - F(i-1)]\; P'''_{avg}(i)$$

where $P'''_{avg}(i)$ is defined in a similar manner as $P_{avg}(i)$ was defined in Section 2.5, and can be derived as

$$P'''_{avg}(i) = \frac{\sum_{j=0}^{d1}\sum_{k=0}^{d2}\sum_{x=0}^{i-d1-d2-1}\sum_{y=0}^{d2}\sum_{z=0}^{d1}\left[\binom{d1}{j}\binom{d2}{k}\binom{i-d1-d2-1}{x}\binom{d2}{y}\binom{d1}{z}(j+k+x+y+z+1)\right]}{2^{d1+d2}\,2^{i-d1-d2-1}\,2^{d1+d2}}$$

$$= \frac{d1+d2+i+1}{2}$$

These component internode distances can be used in Eq. (29) to compute P^{G-SOL}_{3BH}.

3.5. Discussion of Results

Figures 7-11 show results concerning optimum cluster sizes and corresponding *LP ratios* for the 3BH network. The network size is varied from 2^7 to 2^{16} as in Section 2.6. Cluster sizes of 1 were not permitted in this study, since with a cluster size of 1 the resulting network is no longer three-level. The plots for UNIF, SOL and G-SOL suggest that using a three level HIN may improve the *LP ratio,* but only moderately. For example, with UNIF distribution, when the network size is 2^7, the *LP ratio* improves from 0.6 to 0.5 when the optimum 3BH network is used. When the network size is 2^{16} the corresponding values are 0.34 and 0.25. In both the cases, the additional improvement in *LP ratio* is less than or equal to 10% (e.g., the *LP ratio* reduces from 1 associated with the BH network to 0.6 when the BH/BH network is used and from 1 to 0.5 with the 3BH network). For DPF and R-SOL, using a three-level network actually increases the *LP ratio*. (However, for smaller values of *a* in the DP distribution, and large networks, a three level network does offer small improvements in the *LP ratio*.)

It appears from this data that a two-level HIN provides a reasonable compromise between the savings achieved in link cost with a hierarchical design and the penalty paid in terms of increased average internode distance.

4. Summary

This paper considered two issues concerning the design of binary-hypercube based HINs -- the optimum cluster size and the optimum number of levels in the hierarchy. The optimum cluster size is dependent on the underlying degree of locality in communication. Five types of locality characterizations were considered here. For all the locality characterizations, the optimum cluster size in a two-level binary-hypercube based HIN is shown to be either 4 or 8, over a wide range of network sizes. Further, increasing the number of levels to three yields at most a moderate improvement in the cost-benefit ratio, depending on the type of locality characterization considered. For some locality characterizations, the cost-benefit ratio deteriorates. Thus, an appropriate design for a binary-hypercube based HIN appears to be a two-level hierarchy, with a cluster size of either 4 or 8.

REFERENCES

[Abraham & Davidson 1986]
S. G. Abraham and E. S. Davidson, "A Communication Model for Optimizing Hierarchical Multiprocessor Systems," *Proc. 1986 Int. Conf. Parallel Processing,* 1986, pp. 467-474.

[Agrawal & Mahgoub 1985]
D. P. Agrawal and I. E. O. Mahgoub, "Performance Analysis of Cluster-Based Supersystems," *Proc. 1st Int. Conf. on Supercomputing Systems,* St. Petersburg, Florida, IEEE Computer Society, 1985, pp. 593-602.

[Agrawal *et al.* 1986]

D. P. Agrawal, V. K. Janakiram, and G. C. Pathak, "Evaluating the Performance of Multicomputer Configurations," *Computer,* Vol. 19, 1986, pp. 23-37.

[Bhuyan & Agrawal 1982]

L. N. Bhuyan and D. P. Agrawal, "A General Class of Processor Interconnection Strategies," *Proc. 9th Symp. on Comp. Arch.,* 1982, pp. 26-29.

[Carlson 1985]

D. Carlson, "The Mesh With A Global Mesh: A Flexible, High-Speed Organization for Parallel Computation," *Proc. 1st Int. Conf. on Supercomputing Systems,* St. Petersburg, Florida, IEEE Computer Society, 1985, pp. 618-627.

[Dandamudi & Eager 1987]

S. P. Dandamudi and D. L. Eager, "Hierarchical Interconnection Networks for Multicomputer Systems," to appear in *IEEE Trans. Computers* (Also available as Tech. Rep. 87-11, Department of Computational Science, University of Saskatchewan, Saskatoon, 1987).

[Gajski *et al.* 1986]

D. Gajski, D. Kuck, D. Lawrie, and A. Sameh, "CEDAR," Department of Computer Science, University of Illinois, Urbana, 1983. (Reprinted in *Tutorial on Supercomputer: Design and Applications,* (ed.) K. Hwang, IEEE Computer Science Press, 1983, pp. 251-275.)

[Hayes *et al.* 1986]

J. P. Hayes, T. N. Mudge, Q. F. Stout, S. Colley, and J. Palmer, "Architecture of a Hypercube Supercomputer," *Proc. 1986 Int. Conf. Parallel Processing,* 1986, pp. 653-660.

[Hwang & Ghose 1987]

K. Hwang and J. Ghose, "Hypernet: A Communication-Efficient Architecture for Constructing Massively Parallel Computers," *IEEE Trans. Computers,* Vol. C-36, No. 12, December 1987, pp. 1450-1466.

[McCrosky 1986]

C. D. McCrosky, "Message Passing in Synchronous Hypercubes," to appear in *Computer System Science and Engineering.*(Also available as Tech. Rep. 86-4, Department of Computational Science, University of Saskatchewan, Saskatoon, 1986).

[Reed 1983]

D. A. Reed, "Queueing Network Models of Multimicrocomputer Networks," *Proc. 1983 ACM SIGMETRICS Conf. on Measurement and Modeling of Computer Systems,* Minneapolis, Minnesota, 1983, pp. 190-197.

[Reed & Schwetman 1983]

D. A. Reed and H. D. Schwetman, "Cost-Performance Bounds for Multimicrocomputer Networks," *IEEE Trans. Computers,* Vol. C-32, No. 1, January

1983, pp. 83-95.

[Reed & Grunwald 1987]

D. A. Reed and D. C. Grunwald, "The Performance of Multicomputer Interconnection Networks," *Computer,* Vol. 20, No. 6, June 1987, pp. 63-73.

[Swan et al. 1977]

R. J. Swan, A. Bechtolsheim, K. -W. Lai, and J. K. Ousterhout, "The Implementation of the Cm* Multi-Microprocessor," *Proc. of the National Computer Conference,* 1977, pp. 645-655. (Reprinted in *Tutorial on Parallel Processing,* (eds.) R. H. Kuhn and D. A. Padua, IEEE Computer Society Press, 1981, pp. 154-164.)

[Wu & Liu 1981]

S. W. Wu and M. T. Liu, "A Cluster Structure as an Interconnection Network for Large Multimicrocomputer Systems," *IEEE Trans. Computers,* Vol. C-30, No. 4, April 1981, pp. 254-264.

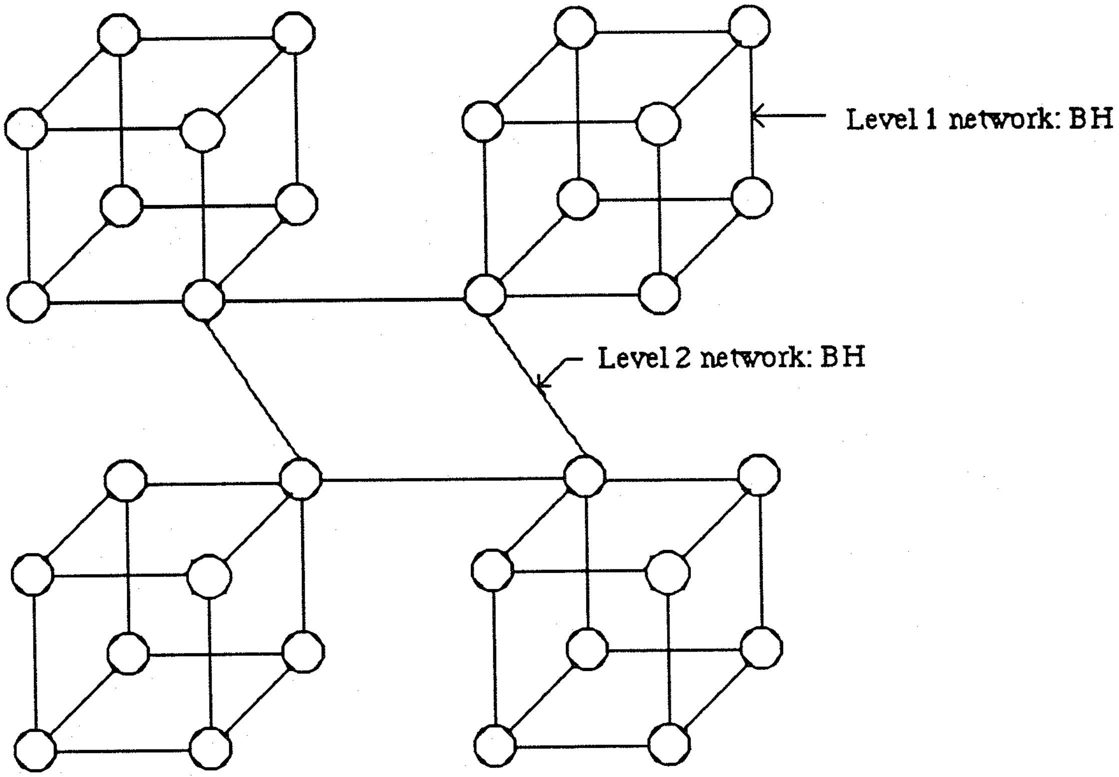
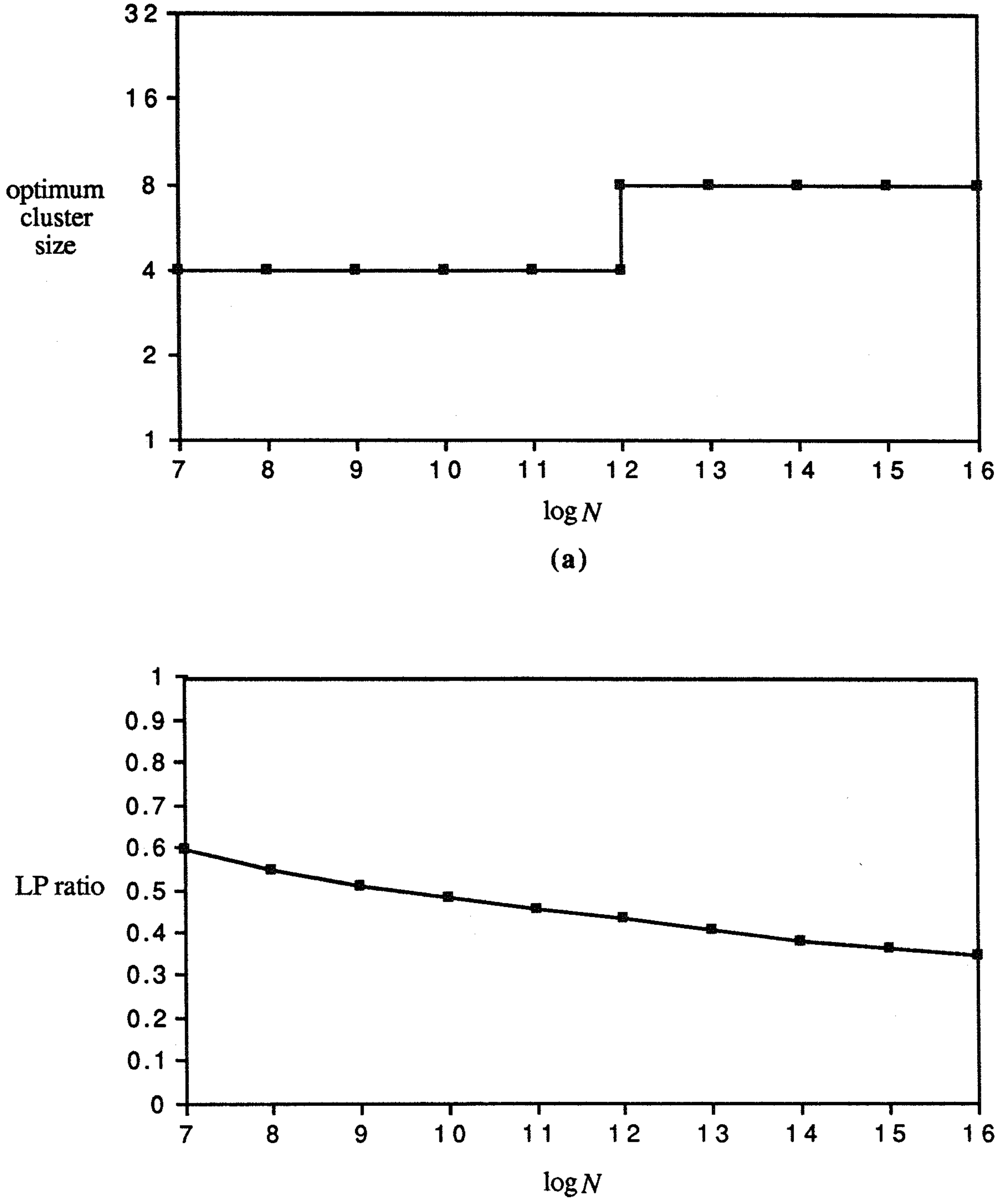


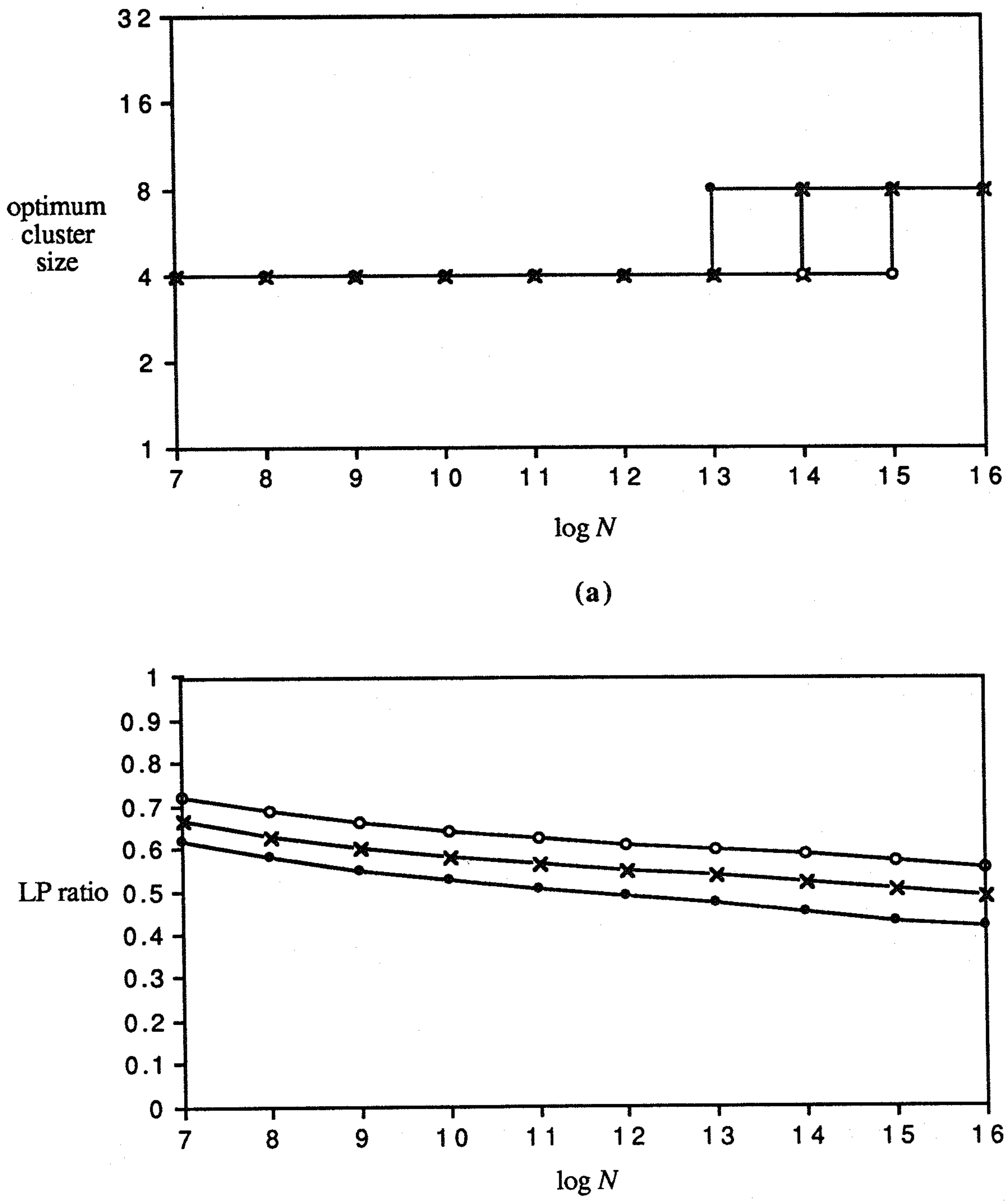
Figure 1 An example hierarchical interconnection network, BH/BH



(a)

(b)

Figure 2 Optimum cluster size and the corresponding LP ratio for the BH/BH network (UNIF)

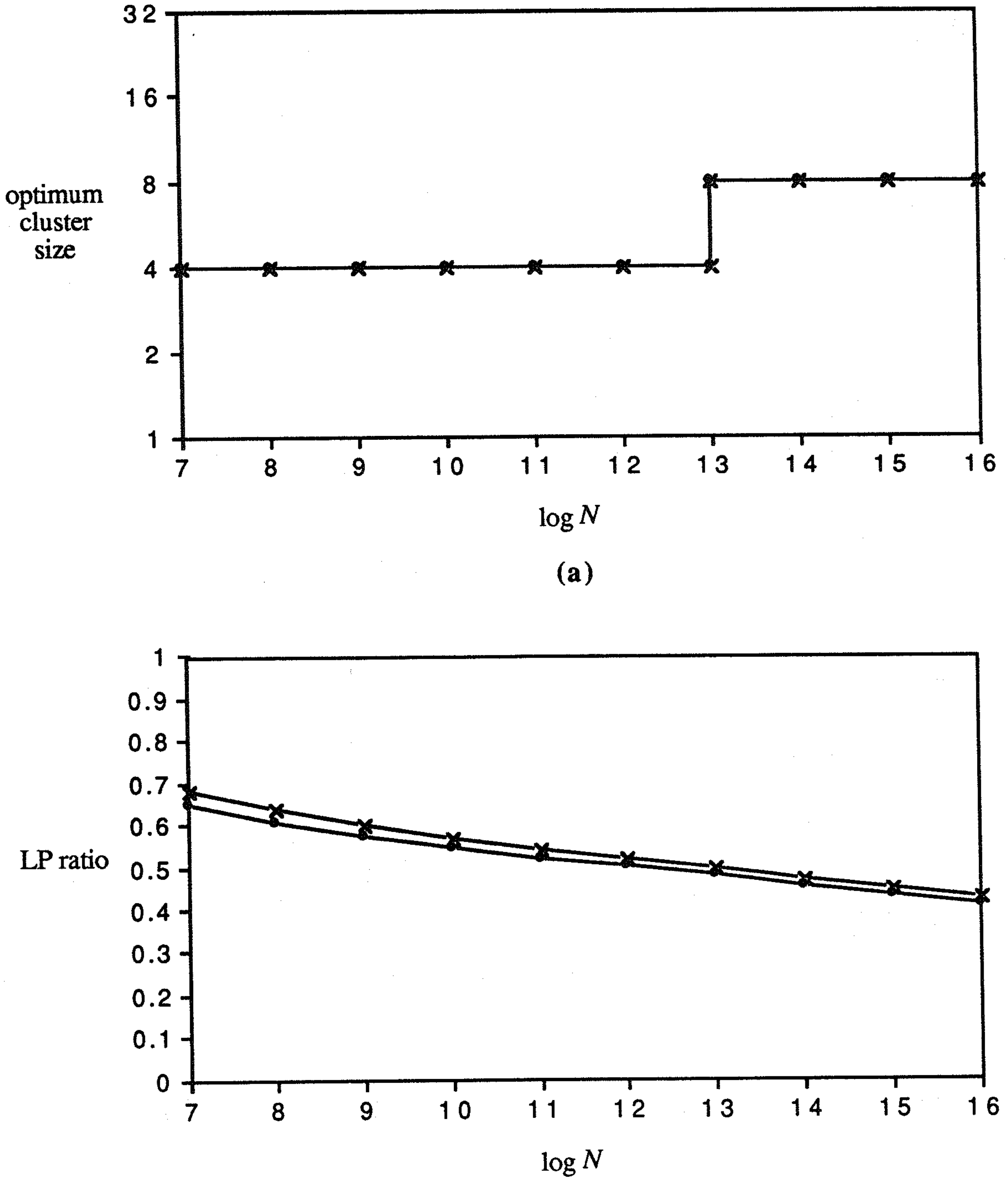


(a)

(b)

Figure 3 Optimum cluster size and the corresponding LP ratio for the BH/BH network (DPF)

○ : $a = 0.3$ × : $a = 0.5$ ● : $a = 0.7$

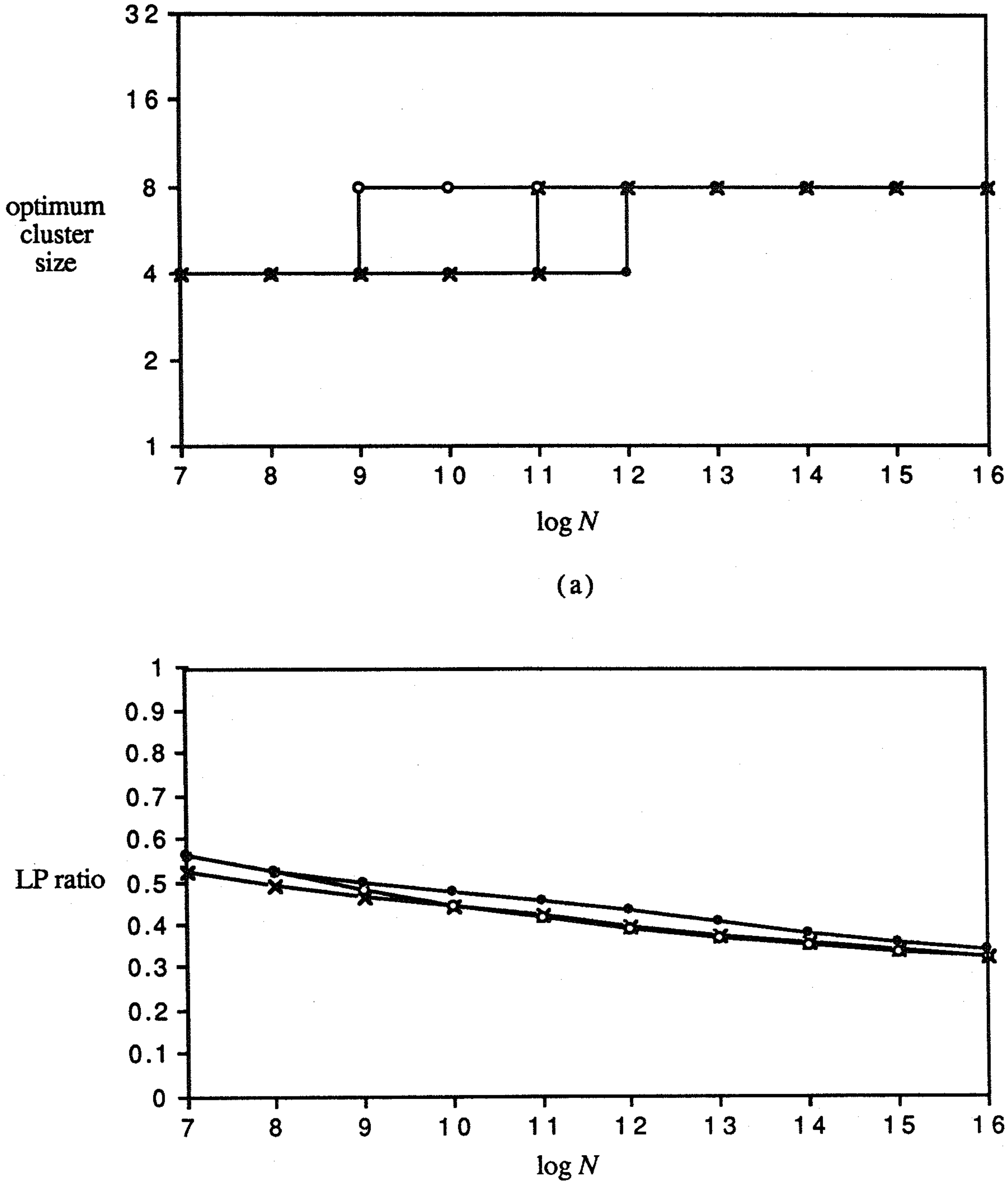


(a)

(b)

Figure 4 Optimum cluster size and the corresponding LP ratio for the BH/BH network (R-SOL with $\alpha = 0.75$)

× : $L = 1$ ● : $L = 2$



(a)

(b)

Figure 5 Optimum cluster size and the corresponding LP ratio for the BH/BH network (SOL with $\alpha = 0.75$)

× : $n = 2^2$　　○ : $n = 2^3$　　● : $n = 2^4$

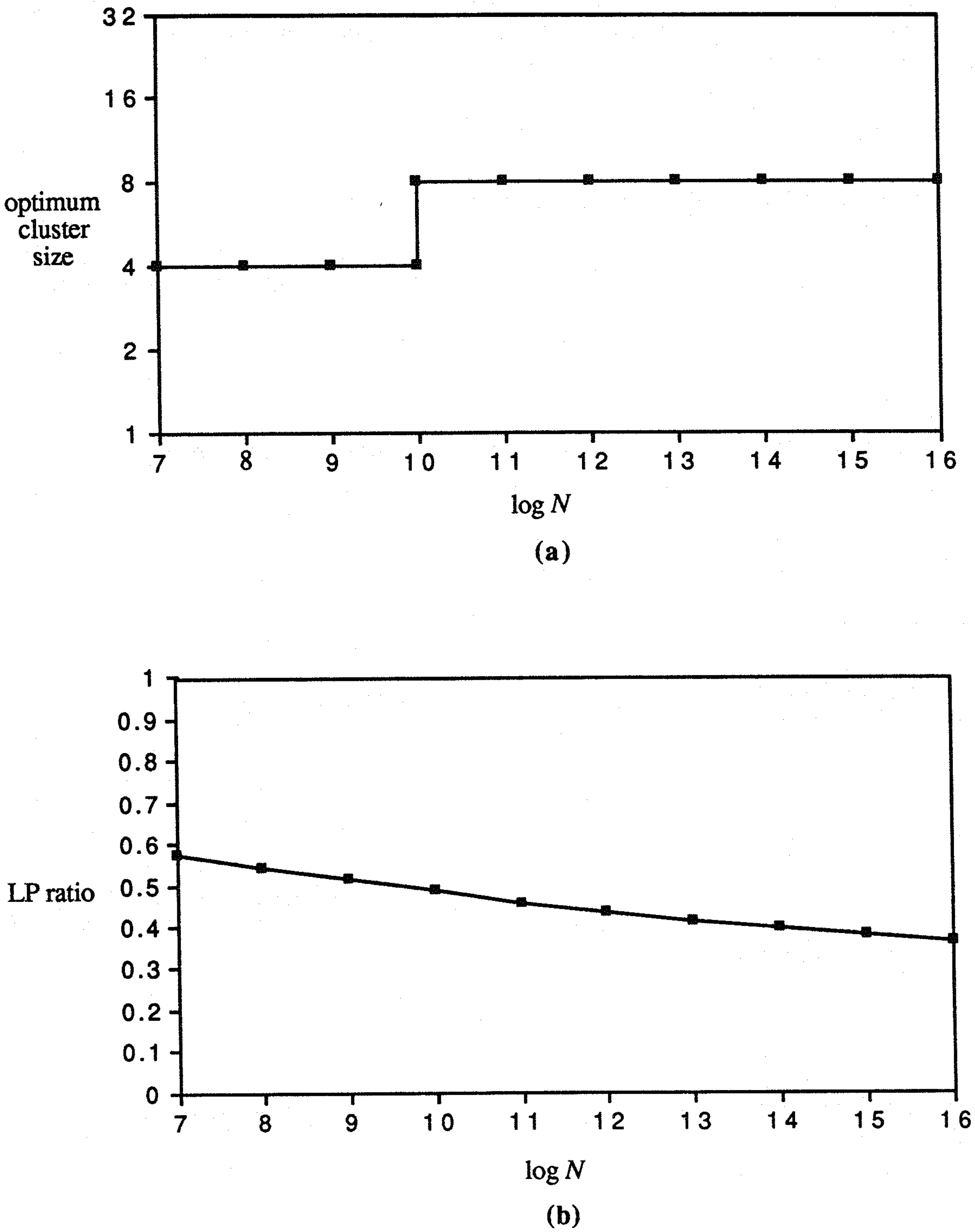
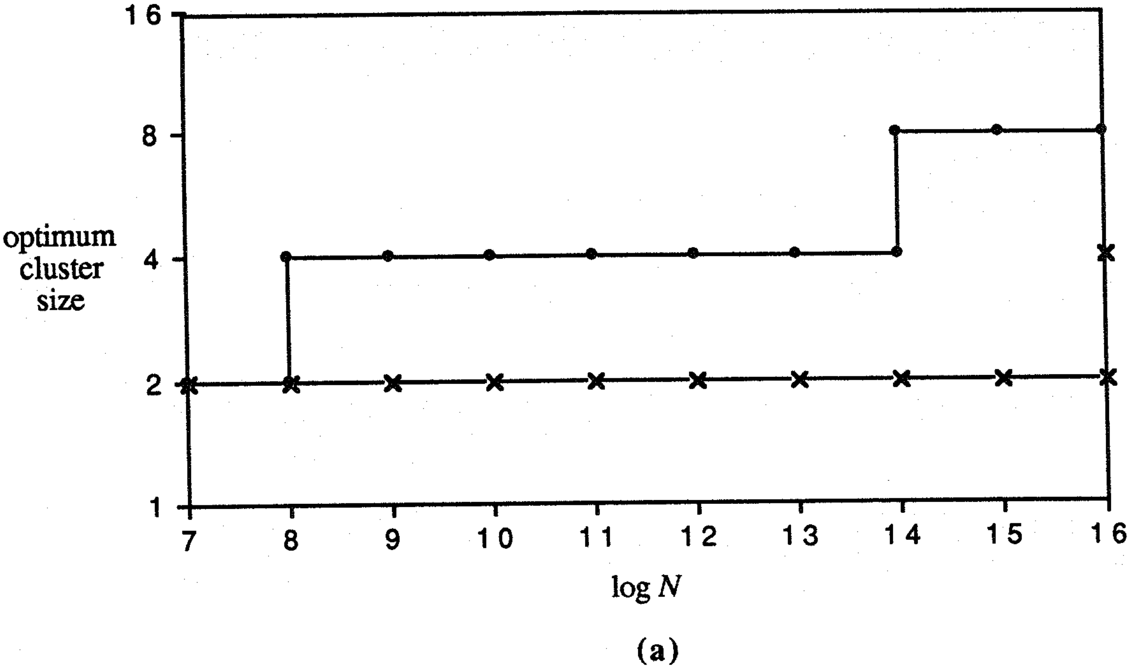


Figure 6 Optimum cluster size and the corresponding LP ratio for the BH/BH network (G-SOL for F(*i*) values in Table 1)



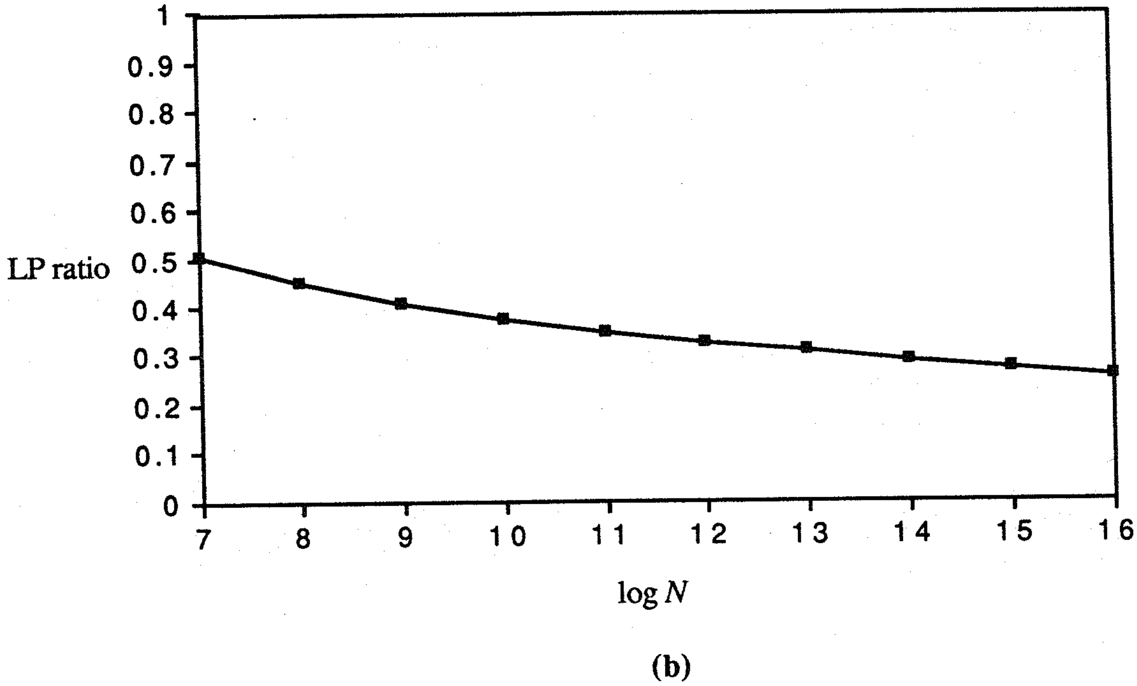


Figure 7 Optimum cluster size and the corresponding LP ratio for the BH/BH/BH network (UNIF)

× : 2^{d1} ● : 2^{d2}

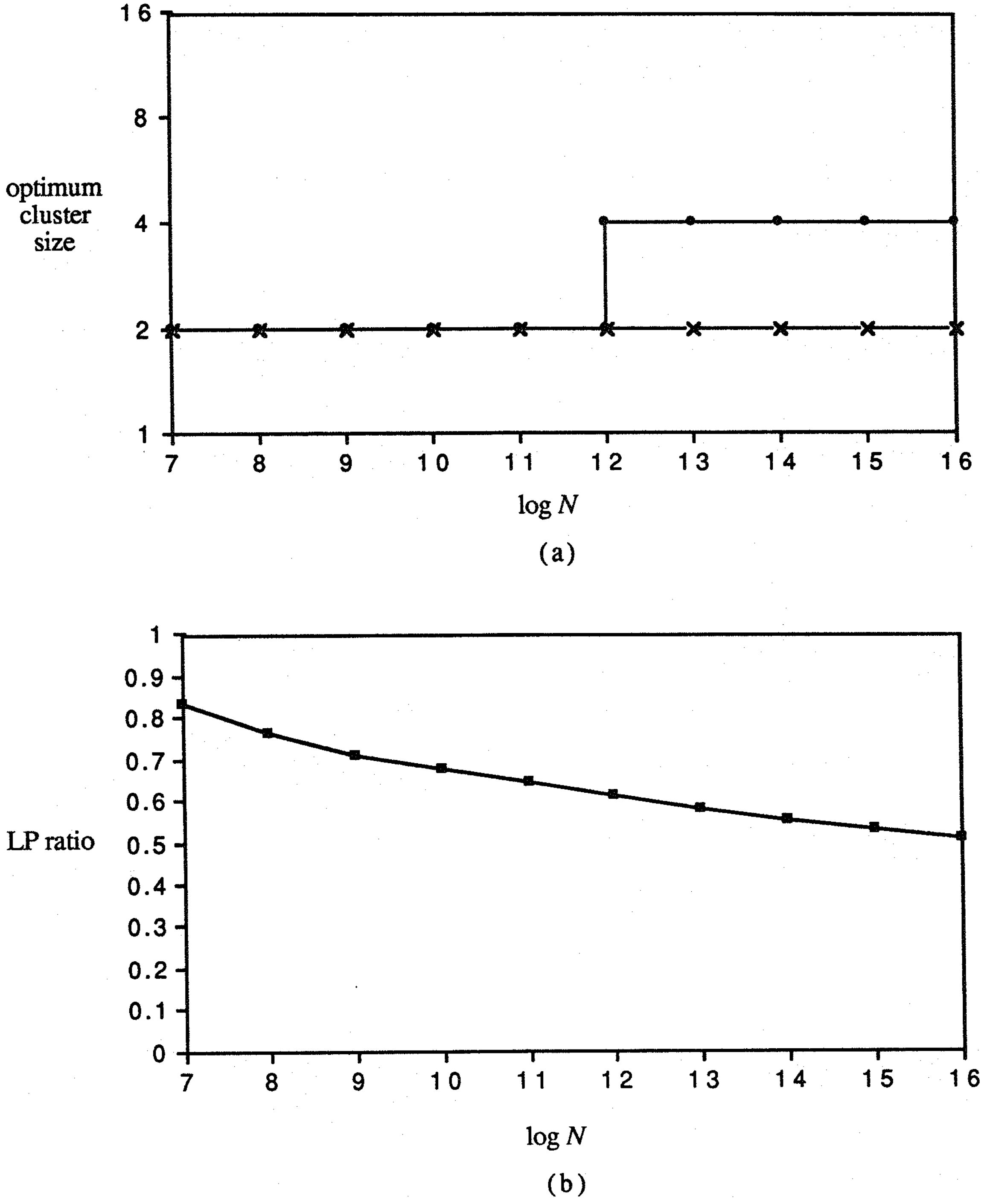
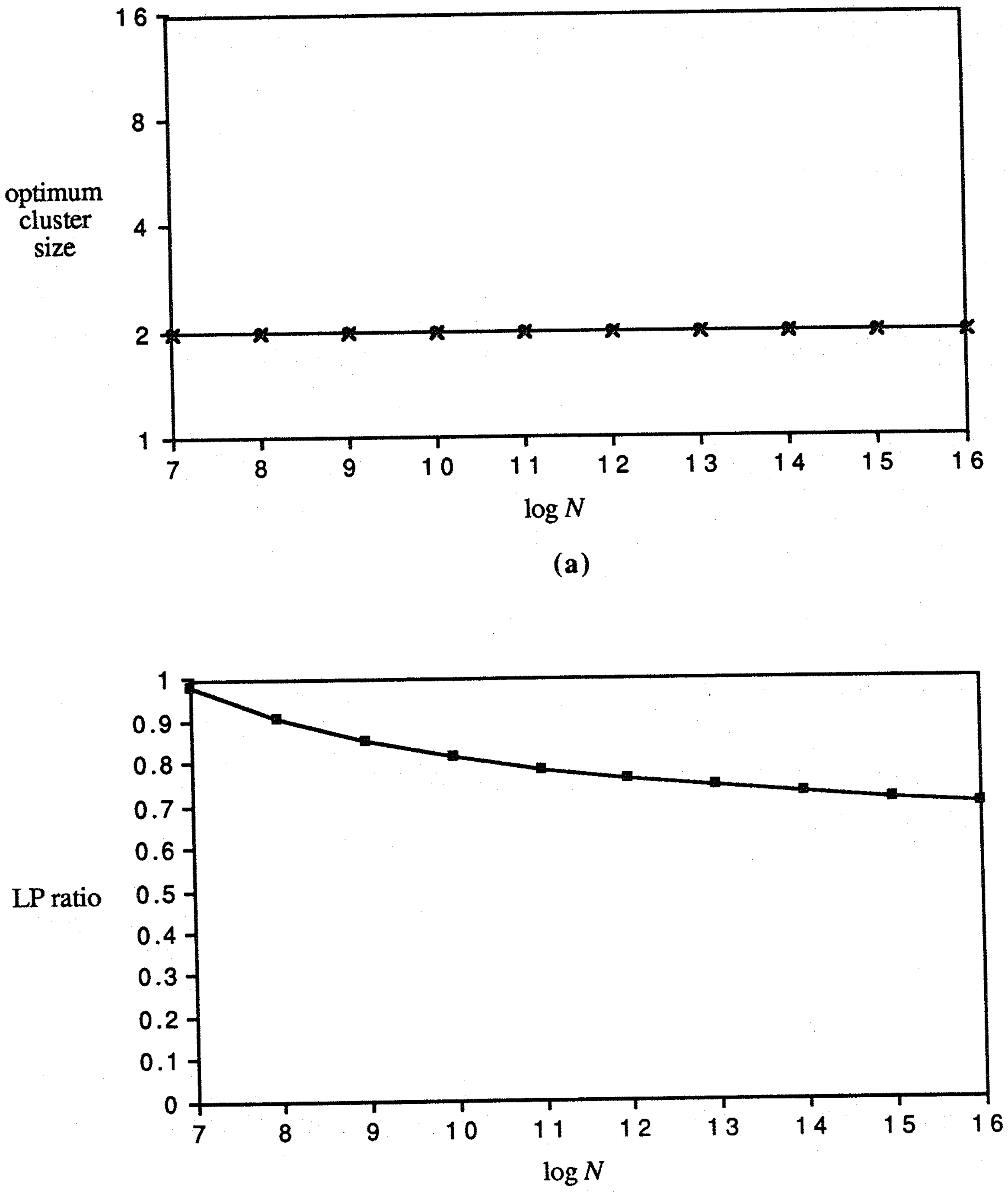


Figure 8 Optimum cluster size and the corresponding LP ratio for the BH/BH/BH network
(DPF with $a = 0.3$)
× : 2^{d1} ● : 2^{d2}



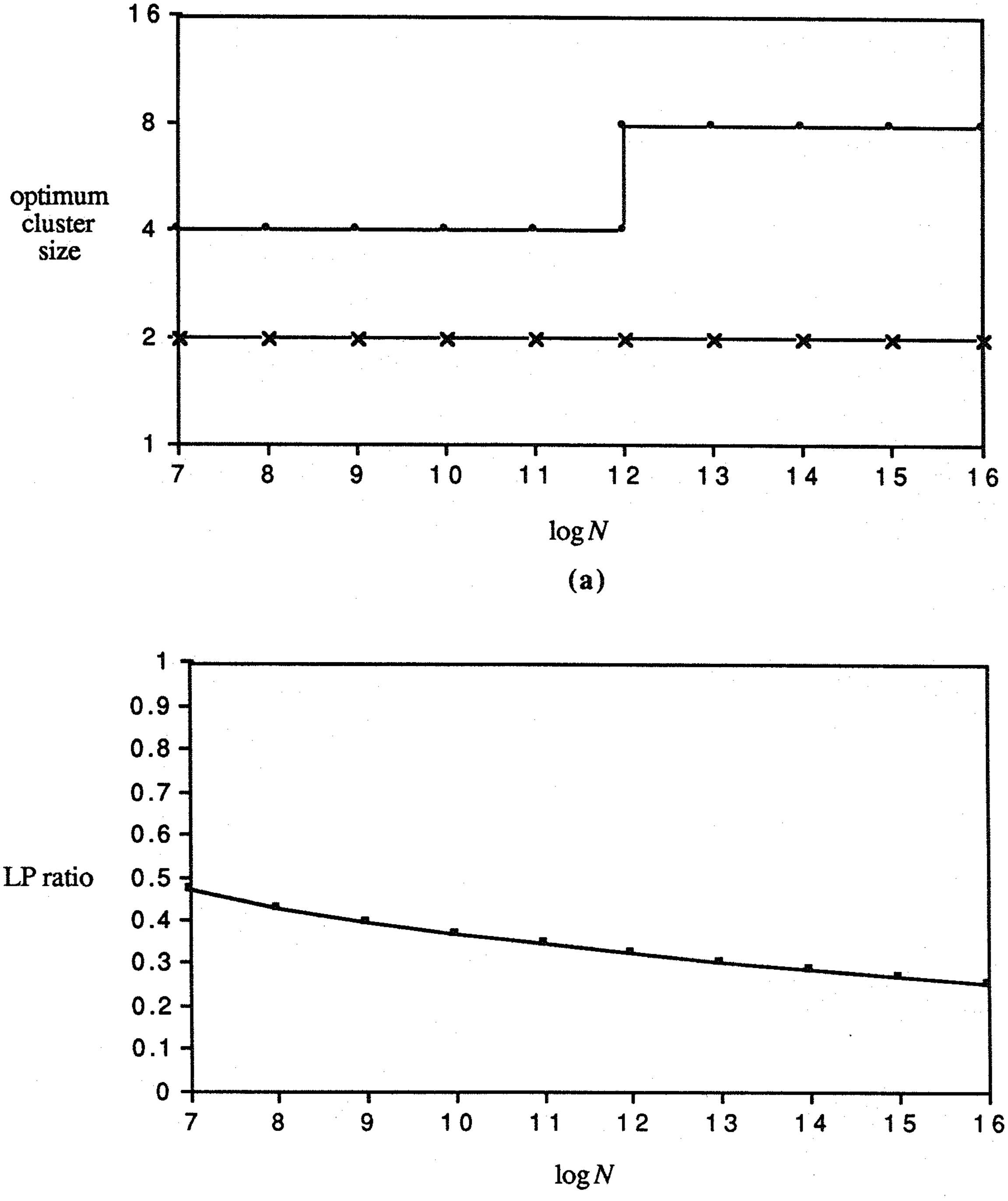
(a)

(b)

Figure 9 Optimum cluster size and the corresponding LP ratio for the BH/BH/BH network

(R-SOL with α =0.75 and L = 2)

× : $2^{d\,1}$ ● : $2^{d\,2}$



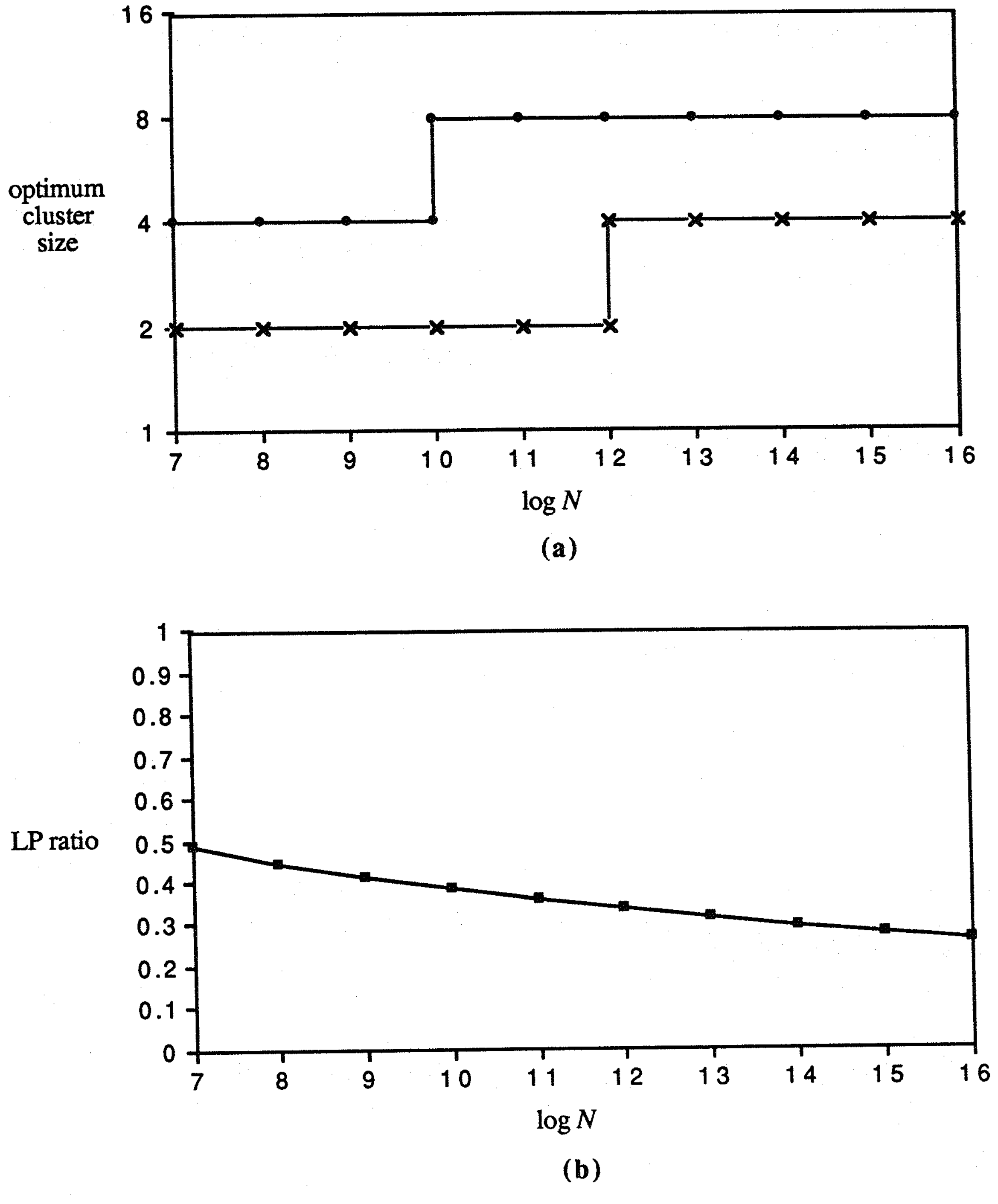
(a)

(b)

Figure 10 Optimum cluster size and the corresponding LP ratio for the BH/BH/BH network

(SOL with α =0.75 and $n = 2^3$)

× : 2^{d1} ● : 2^{d2}



(a)

(b)

Figure 11 Optimum cluster size and the corresponding LP ratio for the BH/BH/BH network

(G-SOL for F(i) values in Table 1)

× : 2^{d1} ● : 2^{d2}