Effect of Non-uniform Traffic on the Performance of Multistage Interconnection Networks

M. Atiquzzaman
Dept. of Computer Science
La Trobe University
Melbourne 3083, Australia.
atiq@latcs1.lat.oz.au

M.S. Akhtar
Dept. of Electrical Engg.
King Fahd Univ. of Pet. & Minerals
Dhahran 31261, Saudi Arabia

Abstract

Multistage interconnection networks are used to connect processors to memories in shared memory multiprocessor systems. Hot spots in shared memory multiprocessor systems result in a degradation the in network performance. An analytical model to evaluate the performance of an unbuffered MIN has been developed. A comparison of performance between unbuffered and buffered MINs under hot spot traffic pattern has been presented in this paper. It is found that an an unbuffered network may perform better than a buffered network in the presence of a non-uniform traffic pattern. Finally, a *hybrid mode network* is suggested for optimum performance under different traffic conditions.

1 Introduction

Tightly-coupled multiprocessor systems with a variety of interconnection networks have been proposed, analyzed, and built in the last two decades [1]. Multistage interconnection networks (MINs) have been found to be a very cost-effective solution to connect a large number of processors and memories. MINs consist of a number of stages of small switching elements (SEs), the successive stages being connected by a permutation function. The Omega network is a blocking type of MIN where contention arising at SEs results in performance degradation. Performance evaluation of buffered and unbuffered MINs in the presence of *uniform* traffic in the network has been presented in the literature [2, 3]. The assumption of uniform traffic in the network is rather restrictive and does not hold in many real-world applications [4, 5].

Hot spots arise in MINs due to the use of shared variables used for locking, global and barrier synchronization, pointers to shared queues, etc. In the case of a buffered MIN, a phenomenon called *tree saturation* severely degrades the performance of the network. In the case of an unbuffered network, the performance degradation is due to the high rate of contention at the SEs carrying the hot memory traffic.

The *first objective* of this paper is to develop an analytical model for the performance evaluation of an unbuffered Omega network in the presence of a *hot spot traffic*. A recursive technique is used to develop the model. It has previously been reported in the literature that buffered MINs have higher bandwidth than unbuffered networks. The above conclusion was based on the assumption of uniform traffic. A *second objective* of this paper is to compare the performance of an unbuffered network with that of a single-buffered MIN, when the networks are operating in the presence of a hot spot traffic. It will be shown that *an unbuffered network may perform better than a buffered network under non-uniform traffic pattern.* Consequently, a *hybrid mode* MIN is proposed to optimize the performance of the network for both uniform and non-uniform traffic patterns.

The modeling assumptions are given in section 2. The analytical model is described in section 3. Performance results for unbuffered and buffered networks are presented in section 4, followed by conclusions in section 5.

2 Modeling Assumptions

An Omega network [6] will be used as an example of a MIN to be analyzed. It connects N processors to N memories using $n = \log_2 N$ stages of $N/2$ crossbar SEs per stage, each SE having two input and two output lines. The $i-$th stage will be denoted by $S_i, 0 \leq i \leq n-1$. A perfect shuffle is used to connect adjacent stages. For modeling purposes, the following assumptions are made regarding the network and its operation.

1. There are $N = 2^n$ processors and N memory modules in the system, where n is a positive integer. The $i-$th processor and the $j-$th memory will be denoted by PE_i and MM_j respectively, where $0 \leq i, j \leq N - 1$.

2. A *packet* switched network operating in the *synchronous* mode is assumed. *Destination tag* routing in used to route the packets. In the case of a conflict at any SE, the SE randomly selects one input and the rejected one is ignored or blocked depending on the network being unbuffered or buffered respectively.

3. *Temporal* and *spatial* independence of requests are assumed. The probability that a processor generates a memory request at the beginning of a cycle is p_0.

4. Memory reference patterns are *not uniform*. MM_h is a *hot memory module* for all processors. If PE_i generates a request, the probability that it will request MM_h is q, whereas, the probability of requesting any module MM_j, $j \neq h$, is $q' = (1 - q)/(M - 1)$. Also $q > q'$.

5. Requests directed to the hot memory module are called *hot requests*. The links and SEs which carry hot requests are called *hot links* and *hot SEs* respectively.

Assumption 3 is unrealistic, but it makes the analysis simple without introducing significant errors [2].

3 Performance Analysis

The performance of the network will be measured by the average memory bandwidth (AMBW) which is defined to be the expected number of memory modules active during any memory cycle.

$$AMBW = \sum_{i=1}^{N} p(i) \tag{1}$$

where, $p(i)$ is the data rate at the $i-$th output link of the network. Starting with the request rate of the PEs, the data rates at the different stages of the network will be calculated recursively. The rest of this section presents a recursive method of calculating the data rates.

Assume that each processor has a request rate of $p_{-1,0}$. The requests are fed to the SEs at S_0. A request arriving at a SE at S_i is routed to the upper or lower output link depending on whether the $i-$th bit of the destination address is 0 or 1 respectively. Let's denote the data rates at the hot and non-hot output links of a hot SE at S_i by $p_{i,0}$ and $p_{i,1}$ respectively. The inputs to the hot SEs at stage S_i have data rate $p_{i-1,0}$. Let the other non-hot data rates at the outputs of S_i be $p_{i,2}, p_{i,3}, \cdots, p_{i,i+1}$, which are dependent on the data rates $p_{i-1,1}, p_{i-1,2}, \cdots, p_{i-1,i}$ respectively at the non-hot input links of the non-hot SEs. The input-output data rate dependency for the different stages can be expressed by a *data rate dependency tree* as shown in figure 1. The tree shows the dependency of the output data rates of a stage on the input data rates of that stage. The boxed data rates are those at the hot links. The data rates in bold typeface are those which depend not only on the data rate indicated in the tree, but also on the ratio of q/q' as will be explained below. The superscript associated with a data rate shows the number of links at that stage having the data rate.

Lemma 1 *There will be a total of $i + 2$ different data rates from $p_{i,0}$ to $p_{i,i+1}$ at the outputs of S_i. The number of output links having non-hot data rate $p_{i,j}$, $0 \leq i \leq n - 1, 1 \leq j \leq i + 1$, is $N/2^{i-j+2}$. The number of links having the hot data rate $p_{i,0}, 0 \leq i \leq n - 1$ is $N/2^{i+1}$.*

The data rates at the output links of S_1 will first be determined, which will then be used to generalize the data rates at the output link of any stage. Let's consider a switch at S_1, as shown in figure 2, with the input and output data rates as marked on the links. For the purpose of illustration, assume that $h \geq N/2$. y_1 will then be the hot output link and will be labeled with $p_{1,0}$. The data rate at y_1 is equal to the probability that there will be a request routed at the y_1 output, and is given by

$$p_{1,0} = \Pr[x_0 \to y_1] \Pr[x_1 \to y_1] + \Pr[x_0 \to y_1] \Pr[x_1 \not\to y_1] + \Pr[x_1 \to y_1] \Pr[x_0 \not\to y_1] \tag{2}$$

where, $\Pr[x_i \rightarrow y_j]$ denotes the probability that a memory request is routed to y_j from x_i, and $\Pr[x_i \not\rightarrow y_j]$ denotes the probability that no request is routed to y_j from x_i.
$\Pr[x_0 \rightarrow y_1] = \Pr[$a message is present at $x_0]$. $\Pr[$the message at x_0 is routed to $y_1]$
The three different data rates at the outputs of S_1 can be shown [7] to be as follows.

$$p_{1,0} = 2(p_{0,0})\left(\frac{(N/4)q' + q - q'}{(N/2-1)q' + q}\right) - (p_{0,0})^2\left(\frac{(N/4)q' + q - q'}{(N/2-1)q' + q}\right)^2 \tag{3}$$

$$p_{1,1} = 2(p_{0,0})\left(\frac{(N/4)q'}{(N/2-1)q' + q}\right) - (p_{0,0})^2\left(\frac{(N/4)q'}{(N/2-1)q' + q}\right)^2 \tag{4}$$

$$p_{1,2} = p_{0,1} - (p_{0,1})^2\left(\frac{1}{2}\right)^2 \tag{5}$$

Similarly, the general expressions for $p_{i,j}$ can be expressed as follows.

$$p_{i,j} = p_{i-1,j-1} - (p_{i-1,j-1})^2\left(\frac{1}{2}\right)^2 \quad \text{for } 1 \le i \le n-1, 2 \le j \le i+1 \tag{6}$$

$$\begin{aligned} p_{i,j} \quad = \quad & 2(p_{i-1,0})\left(\frac{(N/2^{i+1})q' + (q-q')(1-j)}{(N/2^i - 1)q' + q}\right) - \\ & (p_{i-1,0})^2\left(\frac{(N/2^{i+1})q' + (q-q')(1-j)}{(N/2^i - 1)q' + q}\right)^2 \quad \text{for } 0 \le i \le n-1, 0 \le j \le 1 \end{aligned} \tag{7}$$

Equations (6) and (7) recursively define the data rates at the output links of the non-hot and hot SEs respectively.

To calculate the AMBW, the set of data rates $P_{n-1} = \{p_{n-1,0}, p_{n-1,1}, \cdots, p_{n-1,n}\}$ are calculated recursively using equations (6) and (7). The bandwidth is then given (by equation (1)) as

$$AMBW = p_{n-1,0} + \sum_{i=1}^{n} 2^{i-1} p_{n-1,i} \tag{8}$$

where, 2^{i-1} is the number of output links at S_{n-1} having the data rate $p_{n-1,i}$. The above model is used to calculate the performance of an unbuffered network in the next section.

4 Results

In this section, the performance figures for unbuffered and buffered MINs in the presence of hot spot traffic pattern are presented.

Unbuffered network: Figure 3 shows the AMBW vs. the hot spot probability for various values of processor request rates and a network size of 256. The bandwidth decreases because of increased contention at the hot SEs with increasing q. However, the bandwidth increases with increased processor request rates.

The bandwidth as a function of the hot spot probability for different network sizes and for processor request rate of 0.8 is shown in figure 4 . The degradation in performance with increasing hot spot probability is much more pronounced for a large network size.

Figure 5 shows the bandwidth as a function of the network size for different q. For each curve, a constant ratio of q/q' has been maintained. If the ratio q/q' can be maintained constant when the network size is increased, a linear increase in bandwidth can be expected. $q = q'$ corresponds to the URM case. It is observed that even with high hot spot probability ($q = 40q'$), the degradation as compared to the uniform traffic is not significant.

Comparison between unbuffered and buffered network: In figure 6, we present a performance comparison between unbuffered (solid line) and single-buffered (dashed line) Omega networks for uniform traffic and for different degrees of hot spot traffic. Results for the buffered network have been obtained from the analytical models described in [8, 9]. As has been proved earlier in the literature, the buffered network has a higher bandwidth than the unbuffered one for uniform traffic. The above fact is not always true in the case of non-uniform traffic. As an example, for $q = 0.1$ and 0.2, the unbuffered network performs better than the buffered network for processor request rates greater than 0.5 and 0.2 respectively. For higher hot spot probabilities, the unbuffered network performs better than the buffered network for even lower processor request rates. This is due to tree saturation in buffered networks operating under hot spot traffic. As hot spot probability increases, tree saturation occurs at lower processor request rates. If multiple buffers are used at the SEs, it simply takes more time for the onset of tree saturation but the performance is the same once tree saturation occurs. This leads us to suggest a *hybrid network* which can switch between buffered and unbuffered modes. Under normal traffic conditions, the network will operate in a buffered mode, but will switch itself to the unbuffered mode at the onset of tree saturation. Such a hybrid network will provide optimum performance in the presence of both uniform and non-uniform traffic patterns. The results obtained from the analytical models have been validated by simulation results.

5 Conclusions

In this paper, an analytical model to study the bandwidth of unbuffered multistage interconnection networks in the presence of a hot spot traffic has been developed. It has been found that the degradation in the bandwidth with an increase in the hot spot traffic is more pronounced for larger networks than the smaller ones. It has been verified that buffered networks perform better than unbuffered networks for uniform traffic pattern. On the contrary, *unbuffered networks perform better than buffered networks in the presence of a hot spot traffic.* A *hybrid network* has been proposed to optimize the performance of a MIN for both uniform and non-uniform traffic conditions. The hybrid network will normally work in the buffered mode, but will switch to an unbuffered mode whenever congestion in the network is detected. The method of analysis used in this paper can also be applied for the analysis of other Banyan-type networks in the presence of a non-uniform traffic.

References

[1] T.Y. Feng, "A survey of interconnection networks," *Computer*, vol. 14, pp. 12–27, December 1981.

[2] J.H. Patel, "Performance of processor-memory interconnections for multiprocessors," *IEEE Transactions on Computers*, vol. C-30, no. 10, pp. 771–780, October 1981.

[3] D.M. Dias and J.R. Jump, "Analysis and simulation of buffered Delta networks," *IEEE Transactions on Computers*, vol. C-30, no. 4, pp. 271–282, April 1981.

[4] L.N. Bhuyan, "An analysis of processor-memory interconnection networks," *IEEE Transactions on Computers*, vol. C-34, no. 3, pp. 279–283, March 1985.

[5] W.T. Chen and J.P. Sheu, "Performance analysis of multistage interconnection networks with hierarchical requesting model," *IEEE Transactions on Computers*, vol. 37, no. 11, pp. 1438–1442, November 1988.

[6] D.H. Lawrie, "Access and alignment of data in a array processor," *IEEE Transactions on Computers*, vol. C-24, no. 12, pp. 1145–1155, December 1975.

[7] M. Atiquzzaman and M.S. Akhtar, "Effect of non-uniform traffic on the performance of multistage interconnection networks," Tech. Rep., Computer Science Department, La Trobe University, Melbourne, Australia, 1993.

[8] M. Atiquzzaman and M.S. Akhtar, "Performance modeling of switching fabric for ATM switching node," *Australian Broadband Switching and Services Symposium*, pp. 581–588, July 15-17, 1992.

[9] M. Atiquzzaman and M.S. Akhtar, "Performance of buffered multistage interconnection networks in non uniform traffic environment," *7th International Parallel Processing Symposium*, April 13-16, 1993.

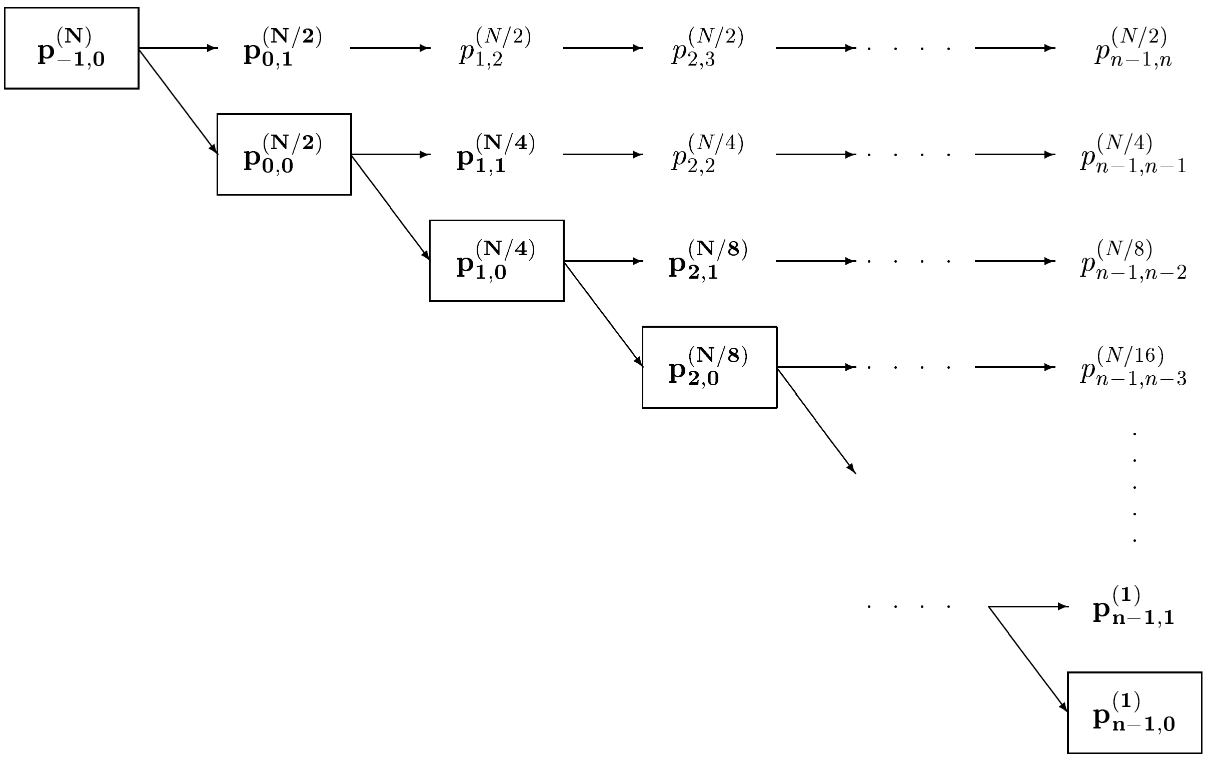

Figure 1: Data rate dependency tree for an $N \times N$ Omega network.

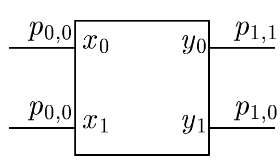

Figure 2: A typical 2×2 crossbar switch used in Omega networks

Figure 3: Bandwidth vs. probability of hot spot for different request rates

Figure 4: Bandwidth vs. probability of hot spot for different network sizes.

Figure 5: Bandwidth vs. network size for different q

Figure 6: Bandwidth comparison between unbuffered and buffered networks