Performance Analysis of Combining Multistage Interconnection Networks*

Prasant Mohapatra
Dept. of Electrical & Computer Engineering
Iowa State University
Ames, IA 50011

Sheldon Wong and Chita R. Das
Dept. of Computer Science & Engineering
The Pennsylvania State University
University Park, PA 16802

Abstract

Concurrent access to a shared variable may cause network saturation in parallel computers. This problem, commonly termed as hot spot contention, can be alliviated by combining requests destined to the hot memory module. In this paper, we propose an analytical model to predict performance of combining multistage interconnection networks. The model considers realistic assumptions like finite length buffers in the switches, deterministic service time, finite degree of combining. Simulation results are used to validate the analytical model.

1 INTRODUCTION

The multistage interconnection network (MIN) has been used as an effective interconnection medium in many shared memory parallel computers [1]. The network delay of MINs is low with moderate intensity uniform traffic. But in a multiprocessor environment the traffic pattern may not be uniform. Access to locks for process synchronization, and loop index variables may create non-uniformity in the access pattern.

Pfister and Norton [2] first investigated the effect of non-uniform traffic distribution which leads to the *hot spot* problem. A hot spot severely degrades the performance of the network. It is found that all traffic, not just those destined for the hot spot, are affected. This happens due to tree saturation. They suggested the idea of combining to alleviate the effects of non-uniform traffic distribution.

Lee [3] has provided a delay analysis on combining networks considering idealized combining where a hot request can possibly combine with an infinite number of other hot requests. Kang et al. [4] presented an analytical model of the NYU Ultracomputer combining scheme. It involved pairwise combining, finite size buffers, and synchronous operation. Merchant [5] provided an analysis of banyan combining networks assuming infinite degree of combining.

The proposed analytical model incorporates finite degree of combining as well as finite length buffers in the SEs. In addition, the model allows for asynchronous communication and incorporates blocking of requests (instead of discarding them). The proposed model is an extension of our work reported in [6], where it is shown that an n-stage MIN under uniform traffic distribution can be modeled by a network of n $M/D/1/L$ queueing centers. In this paper, the model will be expanded upon to include the element of combining which has been discussed earlier.

*This research was supported in part by the National Science Foundation under grant MIP-9104485.

2 MODEL PLATFORM

To connect N processing elements (PEs) to N memory modules (MMs), an $(N \times N)$ MIN constructed using $(s \times s)$ switching elements (SEs) has $n = log_s N$ stages with N/s SEs per stage. The SEs of the MIN have finite buffers of size L located at the input ports of the switches. Additionally, the following assumptions also apply to the model:

(1) Each processor generates requests independently at a rate λ, and the inter-request times are exponentially distributed.

(2) A fraction h of the requests (hot traffic) are directed to the hot MM. The remaining $(1 - h)$ requests (cold traffic) are uniformly distributed among all MMs.

(3) A hot request entering an SE may be combined with an enqueued hot request only if that enqueued request is not fully combined (i.e., it has not yet reached the maximum degree of combining, cd) and is not at the head of its queue.

(4) A request is blocked at a stage if its destination buffer at the next stage is full. However, a hot request may still enter the next stage through combining.

3 COMBINING ANALYSIS

Fig. 1. An (8x8) MIN with a hot spot (MM0).

The switching elements of a MIN can each be modelled as $M/D/1/L$ queueing centers as they have finite buffers and deterministic service time. It has been shown in [6] that for an n-stage MIN affected by a single hot spot, all possible PE-to-MM routes can be modeled by $(n+1)$ queueing network models. Figure 1 shows the fan-in tree (bold paths) from all PEs to the hot MM0. The total traffic rate at the input port of an SE at stage i in the fan-in tree of a hot MM is $(1-h)\lambda_i + hs^i\lambda_i$. The $(1-h)\lambda_i$ term represents the cold traffic entering that SE port, while the $hs^i\lambda_i$ term is the hot traffic rate.

3.1 Combining Probability

Let Z_i be the probability of combining "seen" by a hot request as it arrives at the ith stage. For each possible occupancy length of a buffer, where occupancy length is defined to be the number of requests present in a queue, it is necessary to find the corresponding probability of combining, $Z_i(k)$, where $0 \le k \le L$.

The probability that an occupied buffer space contains a cold request is just the ratio of cold requests in the incoming traffic rate to the entire traffic rate. Thus, if the incoming traffic rate is $\lambda = H + C$, where H and C are the hot and cold traffic rates, respectively, then the cold probability is just $P_{cold} = C/\lambda$. Similarly, the probability of an occupied buffer space containing a single hot request is just $P_{1hot} = H/\lambda$. Probabilities for multiple hot requests ($m = 2, 3, \cdots, cd$) are just $P_{mhot} = P_{1hot}^m$. These probabilities are then normalized to obtain $P_c, P_{1,h}, P_{2,h}, \cdots, P_{cd,h}$.

In general, for a certain buffer occupancy l and maximum combining degree cd, the probability of a valid, but non-combinable, configuration is

$$P_{valid,non}(l) = \sum_{t=0}^{l-1} \binom{l-1}{t} P_c^t P_{cd,h}^{l-1-t}, \qquad (1)$$

where P_c^t is the probability that there are t cold requests in a queue of occupancy l. Since the head queue space is not considered due to the enqueueing/dequeueing assumption, $P_{cd,h}^{l-1-t}$ is the probability that the remaining $(l-1-t)$ buffer spaces each contain a hot request which has reached the maximum combining degree cd.

The general equation for the probability of a valid and combinable configuration with occupancy l and maximum combining degree cd is

$$P_{valid,comb}(l) = \sum_{u=1}^{l-1} \sum_{v=0}^{u-1} P_c^{l-1-u} \left(\sum_{w=1}^{cd-1} P_{w,h} \right) \left(\binom{u-1}{v} P_c^v P_{cd,h}^{u-1-v} \right). \qquad (2)$$

A valid and combinable configuration will have exactly one buffer space which has a combinable hot request (i.e., the number of hot requests in that space is less than cd). The P_c^{l-1-u} term of equation (2) is the probability that there are $l-1-u$ cold requests behind the combinable hot request in the queue. The summation of $P_{w,h}$ is the probability of a combinable request. The remaining term represents the probability that the remaining spaces which are ahead of the combinable request (excluding the head buffer space) are combinations which have v cold requests and $u-1-v$ fully combined hot requests. Thus, the probability of combining at stage i for a buffer occupancy l and a maximum combining degree cd, is simply

$$Z_i(l) = \frac{P_{valid,comb}(l)}{P_{valid,non}(l) + P_{valid,comb}(l)}. \qquad (3)$$

The overall probability of combining that a hot request encounters upon its approach into a queue of size L at stage i is then

$$Z_i = \sum_{l=2}^{L} p_l^{(L)} Z_i(l), \qquad (4)$$

where $p_l^{(L)}$ represents the probability of the buffer occupancy being l in a queue of size L [7].

3.2 Arrival and Departure Rates

The hot MM queueing network model is shown in Figure 2. After every queueing stage i, there is an input arc, b_{i+1}, which represents the addition of hot traffic generated by the other processors of the system. Additionally, immediately before the queue of each stage i, there is an output arc, a_i, which represents the effective reduction in the hot request rate due to combining at stage i. Thus, it can be seen that the departure rate from a stage $(i-1)$ will be different from the arrival rate into stage i. These input/output arcs affect only the hot traffic. The amount of cold traffic that departs from stage $(i-1)$ is the same as that which enters stage i.

Fig. 2. Hot MM queuing network.

Considering the hot spot saturation tree in Figure 1, it can be seen that at the junction where arc b_i joins the hot spot path model, both have subtrees which lead back to the same number of processors. Since each processor generates the same amount of hot requests, b_i is equal to H_i. Hence, $H_i' = H_i + b_i = 2H_i$. After merging in the additional hot requests from arc b_i, the revised traffic rate is given by $\lambda_i' = H_i' + C_i$.

With an entering hot request rate of H_i' into the stage i queue, there is a probability Z_i that combining will occur. Arc a_i represents those hot requests which are able to combine, and thus, are no longer part of the effective traffic rate. The value of a_i is determined to be $(H_i' \times Z_i)$. Since a hot request rate of a_i is removed, the new and correct effective hot request rate entering a stage i queue is then

$$H_i'' = H_i' - H_i' Z_i = H_i'(1 - Z_i). \qquad (5)$$

The overall effective traffic rate entering the stage i queue is

$$\lambda_i'' = H_i'' + C_i. \quad (6)$$

The departure rate λ_{i+1} from a stage i can be derived from the $M/D/1/L$ queue analysis [6] and is expressed as

$$\lambda_{i+1} = \frac{\lambda_i''(1-x_i)}{p_0^{(L)}(i) + \lambda_i'' d(1-x_i)}, \ 0 \le i \le (n-1) \quad (7)$$

where λ_i'' is the revised arrival rate into stage i. x_i is the probability of blocking at stage i, and d is the deterministic service time of the SE.

3.3 Delay Analysis

The queueing network is evaluated from stage 0 thru stage $(n-1)$ using the equations (1-7). Once the departure rate, λ_n, is obtained, the hot content of the traffic, H_n, is doubled to take into account the input arc b_n. The new traffic rate, $\lambda_n' = 2H_n + C_n = H_n' + C_n$, is then the actual effective traffic rate seen entering the hot MM.

Throughout the network evaluation, a running sum is kept of all the departing arcs a_i's. This is because each hot request entering the hot MM might actually contain several hot requests due to combining. Adding this sum to H_n' will give the overall hot throughput. Since cold traffic is never dismissed or added to the queueing network, C_n itself is the overall cold throughput at the hot MM.

Using Little's law, the delay at stage i is determined to be

$$E[T_i] = \frac{\sum_{k=1}^{L+1} k p_k^{(L)}}{\lambda_i}, \qquad for\ 0 \le i \le (n-1). \quad (8)$$

The total hot spot delay is the sum of delay at all n stages.

Fig. 3. Queuing models with combining effect.

The queueing network model including the combining effect is shown in Figure 3. For the group 1 queueing network, the hot traffic remains in the network until after stage 2. This can be seen in the fan-in tree of Figure 1. To access the group 1 MM (i.e., MM1) a cold request "*shares*" a route with other hot and cold traffic until after the stage 2 queue, at which time the hot traffic departs through the upper port of the SE and this group 1 request departs through the lower port. Since hot traffic is present in this group 1 system until after stage 2, combining and input hot traffic from other processors must still be accounted for in stages 0 and 1. The entire hot traffic rate is removed after stage 2, and only cold traffic reaches MM1. In the group 2 model, the hot traffic departs after stage 1, and in group 3's queueing network, the hot traffic is removed after stage 0. Up to those respective removal stages, combining and input hot traffic rates must be taken into account. In general, for a group i queueing network $(0 \le i \le n)$, the hot traffic is removed after stage $(n-i)$.

Analysis of a cold traffic queueing network is similar to that performed on the hot MM queueing network. The only exception is that after the removal of the hot traffic, only the cold traffic traverses toward the MM. Since the cold traffic does not participate in combining, for stages after the hot spot departure, the departure rate calculated at a stage i is the arrival rate into stage $(i+1)$. The departure rate after stage $(n-1)$ of a group i combining network, $\lambda_n(group\_i)$, is the actual throughput into a group i MM. Since all possible cold paths are represented by the $(n+1)$ different queueing networks, one of which is the hot spot queueing network, the throughput of each is used to calculate the average cold throughput of the system.

The delay contribution of each group must be weighed by the number of cold MMs in that group. If D_i is the delay calculated for a group i queueing network, then the average cold traffic delay is given by

$$D_{cold} = \frac{D_{hotspot} + g_1 D_1 + g_2 D_2 + \cdots + g_n D_n}{N}, \quad (9)$$

where $D_{hotspot}$ is the delay for the cold throughput of the hot spot combining network.

The overall average delay is determined by adding the average hot delay with the average cold delay, while weighing the hot and cold delays by their respective throughput contributions.

4 MODEL VALIDATION

In order to validate the analytical model that has been developed, we have compared the results with those obtained through simulation. An $(N \times N)$ baseline network with finite input-port buffers is used as the simulation platform. Requests at each PE are randomly generated with an exponential distribution of interarrival time with an average rate of λ requests per cycle. A random value is used in conjunction with the hot spot percentage, h, to determine whether a generated request is hot or cold.

Simulations were run for both (64×64) and (256×256) systems with hot spot percentages of 8%, and 16%. Comparisons of the normalized overall throughput versus average delay curves for the (64×64) are shown in Figures 4(a), and 4(b) for hot spot values of 8% and 16%, respectively. The difference between

the analysis and the simulation results is within 15%. A possible reason for the difference seen might be because at moderate to high request rates, the departure rate from a queueing center is almost deterministic, whereas, in the analysis, the departure rate is considered to be exponentially distributed.

Fig. 4. Model Validation for a (64 × 64) MIN.

The (256 × 256) system validation curves for hot spot percentages of 8% and 16% are shown in Figures 5(a) and 5(b), respectively.

5 CONCLUSIONS

A queuing model is proposed for analyzing performance of combining multistage interconnection networks. The model is general in the sense that it can be used for any degree of combining, and is based on realistic assumptions. The queuing analysis derived in [6] is used for solving the model. The proposed analytical model is shown to obtain results which compare favorably with those obtained through simulation, thus validating the developed model. Such a model would allow a designer not only to predict performance of a specific system under different traffic scenarios, but also analyze a variety of design implementations. A number of interesting issues such as, effects of combining degree, buffer length, are currently being studied.

References

[1] C. Wu and T. Feng, "On a Class of Multistage Interconnection Networks," *IEEE Trans. on Computers,* pp.694-702, Aug. 1980.

[2] G. Pfister and V. Norton, "Hot Spot Contention and Combining in Multistage Interconnection Networks," *Int. Conf. on Parallel Processing,* pp.790-797, 1985.

[3] G. Lee, "A Performance Bound of Multistage Combining Networks," *IEEE Trans. on Computers,* pp.1387-1395, Oct. 1989.

[4] B. Kang, G. Lee, and R. Kain, "Performance of Multistage Combining Networks,"*Int. Conf. on Parallel Processing,* pp.550-553, 1991.

[5] A. Merchant, "Analytical Models of Combining Banyan Networks," *Performance Evaluation Review,* pp.205-211, June 1992.

[6] P. Mohapatra and C. R. Das, "A Queuing Model for Finite-Buffered Multistage Interconnect Networks," *Int. Conf. on Parallel Processing*, pp. 210-213, Aug. 1993.

[7] P. Mohapatra, "Analytical Modeling of Combining Multistage Interconnection Networks," Technical Report, Department of Electrical and Computer Engineering, Iowa State University, 1994.

Fig. 5. Model validation for a (256 × 256) MIN.