Performance and Configuration of Hierarchical Ring Networks for Multiprocessors

V. Carl Hamacher*
Department of Electrical and Computer Engineering
Queen's University
Kingston, Ontario, Canada K7L 3N6

Hong Jiang†
Department of Computer Science and Engineering
University of Nebraska-Lincoln
Lincoln, Nebraska 68588-0115

Abstract

Analytical queueing network models for expected message delay in 2-level and 3-level hierarchical-ring interconnection networks (INs) are developed. Such networks have recently been used in commercial and research prototype multiprocessors. A major class of traffic carried by these INs consists of cache line transfers, and associated coherency control messages, between processor caches and remote memory modules in shared-memory multiprocessors. Memory modules are assumed to be evenly distributed over the processor nodes. Such traffic consists of short, fixed-length messages. They can be conveniently transported using the slotted ring transmission technique, which is studied here. The message delay results derived from the models are shown to be quite accurate when checked against a simulation study. The comparisons to simulations include heavy traffic situations where queueing delays in ring crossover switches are significant for ring utilization levels of 80 *to* 90%. *As well as facilitating analysis, the analytical models can be used to determine optimal sizes for the rings at different levels in the hierarchy under specified traffic distributions in a system with a given total number of processor nodes. Optimality is in terms of minimizing average message delay. A specific example of such a design exercise is provided for the uniform traffic case.*

1 Introduction

A main hardware component in a multiprocessor system is the interconnection network (IN) that connects together processors and remote memory modules. One such IN structure, hierarchical slotted rings, is an interesting base on which to build large scale shared-memory multiprocessors. They have received a great deal of attention recently, both in academia [12, 17, 14, 5, 7, 10, 6] and in industry [16, 3, 4]. The salient features of this class of INs are: (1) the physical locality of hierarchical rings blends naturally with that of computational locality of shared-memory multiprocessing [12, 7], (2) the hierarchical ring structure provides natural and efficient broadcasting and multicasting capabilities that are crucial for process coordination and cache coherence protocols [5], and (3) hierarchical rings have an inherent and unique capability of "diluting" the impact of hot-spot traffic [17, 7]. Nevertheless, a more popular choice for INs seems to be meshes. This, as noted in [12], may stem from the fact that mesh-connected systems are relatively easy to build using off-the-shelf routers and processors and have good scalability characteristics. While meshes have superior scaling characteristics relative to hierarchical rings, both of the only two comparative studies of hierarchical rings and meshes in the literature, one based on an approximate modeling [6] and the other based on detailed execution-driven simulations [12], concluded that hierarchical rings outperform meshes under some practical workloads. More specifically, [12] found that hierarchical rings perform significantly better than meshes for system sizes up to 121 processors if the workload exhibits moderate to high memory access locality. Even if there is no memory locality, [12] observed that hierarchical ring systems perform better than meshes for systems with large cache lines either if the system is small, or if the global ring has double the normal bandwidth.

Exact analytical modeling of hierarchical slotted-ring networks is intractable because of the phenomenon of "clustering" of occupied slots in the ring as observed in [11, 1]. As a result, analytical studies of such networks have been based on approximation techniques [11, 1, 17]. With the exception of [17], which analyzed 2-level structures, hierarchical ring structures have not been studied analytically so far despite the existence of many analytical studies in the literature on single-level rings [11, 1]. Paper [17] evaluated the performance of two-level ring structures under cache-coherent traffic in the form of hot-spot patterns. It considered the source removal transmission protocol. Two other recent performance studies on hierarchical ring networks were based entirely on simulations [7, 12].

In this paper, we use approximate analytical techniques to model the message delay performance of 2-level and 3-level hierarchical ring networks that operate under a destination removal protocol, as opposed to source removal. The former is more efficient in terms of network channel utilization and has been em-

*Supported by an NSERC (Canada) Research G rant
†Supported in part by a Nebraska Research Initiative (NRI) Research Grant

0190-3918/97 $10.00 © 1997 IEEE

ployed in recent research prototypes [15, 14]. We consider both uniform and localized traffic patterns that are typical of shared-memory multiprocessing applications. A main objective of the paper is to gain important insights based on the performance measures obtained analytically, into the optimal design of hierarchical ring systems. That is, for a given total node size and traffic environment, how should one determine the size of rings on different levels to minimize the expected message delay?

The paper is organized as follows. Section 2 presents a description of the hierarchical interconnection network model, including enough structural and operational detail for performance evaluation purposes. Section 3 presents message delay models using queueing models to capture the effect of contention. The analytical models developed are validated through extensive simulations and accuracy of the analytical models is assessed in Section 4. Section 5 addresses the issue of optimal configuration using the analytical models developed in Section 3. Finally, some concluding remarks and prospects for future work are made in Section 6.

2 Hierarchical Ring Networks

The hierarchical slotted-ring IN studied here consists of unidirectional rings, as employed in [3, 4]. Processor node clusters are only connected to local rings, as shown in Figure 1. Each segment, called a station, connects one cluster into the ring. The station switch, S, removes an incoming ring message into its cluster interface if it is the destination, or sends the message on around the ring otherwise. This message-handling protocol is the same as that used in destination-remove, slotted, Local Area Networks (LANs) [11]. A switch introduces a pending transmit message from its cluster interface into the downstream station as soon as it observes its own ring input side to be empty. Ring traffic is thus never blocked. In the context of memory read/write messages in shared-memory multiprocessors, operations can be described briefly as follows. At the destination station, the message has priority on the cluster bus. If the target memory module is free to handle the request, it starts the operation (a read or a write), and immediately sends a positive acknowledgment message back to the source station, where the acknowledgment is removed by the source station switch. A negative acknowledgement is returned if the target memory module is busy, and the read/write request message will need to be tried again later by the source. If the destination memory module is free, a write operation requires a request and acknowledgement message. A read operation requires three messages: one to send the read request, an acknowledgement, and a later one from the destination station to return the requested data. These details are not actually needed for the network performance modeling done later, but they explain the use of the destination-remove protocol in the shared-memory application. Efficiency is enhanced if acknowledgements immediately use the slot vacated by the request. This operational possibility is modeled in the analytic and simulation study reported here.

The bit width of the local ring is assumed to be enough to carry full information for a memory word write message or a two-word reply message to a read request. This wide-slot format is used in both [3] and [4].

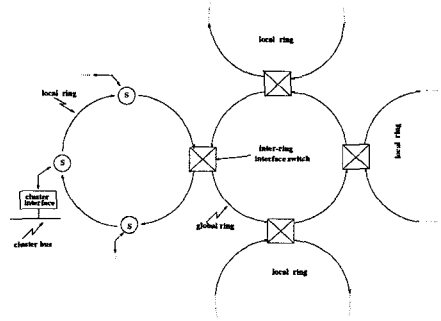

Figure 1: A 2-Level Hierarchically Structured Multiprocessor

A local ring can be expanded to any desired number of segments because each station is a regenerative repeater in the electrical sense. However, from a performance standpoint, message transfer delay will increase linearly, degrading performance. To alleviate the performance problem, a higher level ring can be added in the form of a global segmented ring that is used to interconnect local rings, as shown in Figure 1. It operates much like a local ring, with its source and destination stations being local ring interfaces instead of cluster interfaces. This structure can be extended to even higher levels. Message blocking can occur at the crossover switch between two rings. For example, in a 2-level system, if a message from a local ring needs to move up to the global ring at the same time that a continuing message on the global ring arrives at the crossover switch, there is contention for the downstream link on the global ring, and only one message can proceed. The other message must be temporarily buffered in the crossover switch to insure that messages are never lost in the network. Details will be given in Section 3.2.

3 Contention (Queueing) Model for Message Delay

In [6, 10] we developed message delay and throughput performance measures for hierarchical rings in the light traffic (no contention) situation. While contention-free models are easy to develop and useful for rough network comparison purposes, any detailed evaluation of a network must consider contentions that occur. Further, only contention models can identify potential system performance bottlenecks. In this section, analytical models will be developed to capture the effect of contention under the full range of the applied loads.

3.1 Message Destination Distribution

Applications that run on shared-memory multiprocessors will have different patterns of message destination locality as the processor clusters (containing one or more processors) make memory read/write requests to remote memory modules. These patterns

may range from situations where a cluster references mainly only a small number of other cluster memories (high locality) to situations where references are are uniformly distributed over all other clusters (low/no locality). In the first case, clusters that reference each other often should be located on the same local ring. Conversely, if such situations dominate, the size of the local ring in a hierarchical ring network can be chosen to best match the size of the typical locality sets. If applications tend to have uniform destination distributions, then for a fixed total number of clusters, the various ring sizes can be chosen to minimize average message delay. An example of this network design optimization is given in Section 5.

In the models to be developed, the following parameters reflect message destination locality. In H2 (2-level systems), P is the probability that a message is destined for a cluster on the same local ring, with $1 - P$ being the probability that it will need to move over the global ring to a different local ring. In H3 (3-level systems), P_L is the probability of a "same local ring" destination. P_M is the probability that the message is destined for another local ring attached to the same intermediate ring; while $P_G = 1 - (P_L + P_M)$ is the probability that the message must move all the way up through the global ring, eventually moving down through the hierarchy to a local ring on a different intermediate ring.

3.2 Queues in the Network

FIFO queues are associated with each local ring station interface and inter-ring interface, as shown in Figure 2 and Figure 3, respectively. At a station interface, shown in Figure 2, the message packet at the head of the queue waits until an empty slot passes by, or a full slot destined to the local station arrives and the packet is removed from the slot by the station, at which time the head packet is transmitted onto the slot. Thus, a slot is deemed *empty* if it (1) contains no valid packet, or (2) contains a packet destined to the local station and will be removed by it. The transmitted packet will then travel to its destination station unblocked if the destination is on the local ring, or to the inter-ring interface otherwise. At the inter-ring interface, shown in Figure 3, the packet joins the FIFO queue for the higher level ring. Once at the head of the queue, the packet follows similar steps as in the case of a local station interface; that is, the packet rides on the first empty slot to join the FIFO queue at another inter-ring interface connecting down to the destination ring, or up to a higher-level ring, depending on the destination. Ultimately, the packet is removed from the ring by the destination station. Thus, the message delay, d (see Figure 2), of a packet is the sum of (1) *queueing delays* at all FIFO queues on its entire path from source station to destination station, (2) *slot access time* at all interfaces on its path, that is, the time between when the packet reaches the head of a FIFO queue and when it gets an empty slot, (3) *slot traverse time*, the total time the packet spends moving through ring segment slots on its entire path, and (4) a final time step into the destination station bus buffer.

Part (3) of the message delay is uniquely determined by the source and destination addresses and the network configuration, independent of traffic density and contention. Clearly parts (1) and (2) of the message delay capture the effect of contention, and hence are traffic density dependent. Unfortunately, it is extremely difficult to model the contention exactly, due to the dependence among *full slots*. This dependence, also known as "clustering of full slots", has been observed in [11, 1, 17], where, as traffic intensifies, full slots tend to cluster together to form "trains" of slots, as opposed to full slots being uniformly distributed on the rings. This dependence makes an exact analysis intractable [9]. A second factor that complicates the exact analysis is the issue of finite buffers. To make the analysis tractable and simple, we circumvent the problems by making two main simplifying assumptions. First, we assume that the event of a slot being full is independent of that of other slots. Second, we assume the FIFO buffers at all interfaces are infinite in size. Fortunately, these assumptions have been shown to be not problematic as shown in [11, 1, 17] and by our own simulation validation studies.

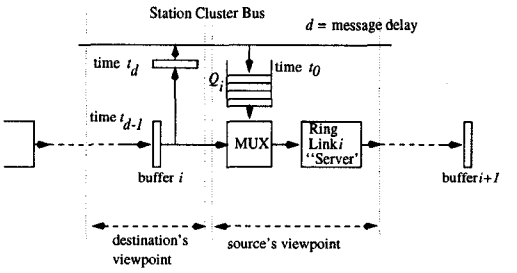

Figure 2: Structure of Local Station Interface

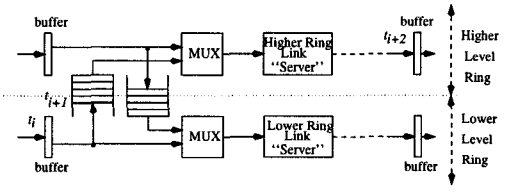

Figure 3: Structure of Inter-Ring Interface

With the above assumptions, we model the contention in parts (1) and (2) of message delay using the M/G/1 queueing center model, similar to the approach in [1] and [17] where source-remove one-level and two-level rings, respectively, are analyzed. The key in this method lies in finding the expected service time of the M/G/1 service center which models a particular interface FIFO queue. This expected service time is effectively the expected time that a packet at the head of the queue waits before it gets an empty slot. In what follows we first define the necessary parameters and list assumptions for the analysis and then give a detailed description of the analytical model.

It should also be noted that, from the modeling viewpoint, there is also a buffer, called a *ring link buffer*, associated with each ring link in the system, as shown in Figures 2 and 3 in narrow bars. It only needs to have capacity of 1 because:

i) arrivals occur only at discrete time points, and

the associated ring link "server" has a constant service time of 1 discrete time step; and

ii) this ring link buffer has priority over station FIFO queues and inter-ring crossover queues in competing for access to the ring link "server". This priority policy is consistent with the implementations of the NUMAchine [14, 12] and KSR [4].

We will not need a specific notation to identify these buffers because their total occupancies can be derived from ring utilization, which can be calculated directly from input message traffic and message travel patterns. This will become clear later.

3.3 Definitions and Assumptions

Time is discretized into clock ticks, where one tick is the time needed for a packet to move between adjacent slot segments in any ring. The models to be developed are based on the following system parameters:

1. λ: identical traffic arrival rate at each local station, i.e., number of independent message packets per clock tick arriving at a local ring station FIFO queue.
2. message destination locality in H2 is determined by probability P as defined in Section 3.1.
3. message destination locality in H3 is determined by probabilities P_L and P_M as defined in Section 3.1.
4. N: total number of local stations in the network.
5. L: number of stations on a local ring.
6. M: number of local rings on an intermediate ring in the case of 3-level ring network.
7. G: number of lower-level rings connected to the global ring. Note that $G = \frac{N}{L}$ in 2-level ring networks and $G = \frac{N}{LM}$ in 3-level ring networks.

Furthermore, we make the following assumptions:

1. The traffic arrival rate at each station and inter-ring interface FIFO follows a Poisson process.
2. One message packet can be completely carried by one slot.
3. A packet is removed from the network by the destination immediately after it reaches the destination station cluster bus buffer (see Figure 2).

3.4 General Model

The basic idea of this analysis is to solve the M/G/1 queueing model for all FIFO queues (local stations and inter-ring interfaces), which will give rise to expected queue lengths at all FIFO queues. We also need ring utilizations. Using Little's result [13], these results can then be used to derive expected message delays as follows.

Let Q_i, $1 \le i \le N$, denote the queue length of local station $S(i)$, and let $Q_{L-G(i)}$ and $Q_{G-L(i)}$ denote, respectively, the local-ring to global-ring FIFO queue length and the global-ring to local-ring FIFO queue length of the inter-ring interface i, $1 \le i \le \frac{N}{L}$ for the 2-level ring. Similarly, for the 3-level ring, let $Q_{M-L(i)}$, $Q_{L-M(i)}$, $Q_{G-M(j)}$, and $Q_{M-G(j)}$ denote, respectively, the middle-ring to local-ring, local-ring to middle-ring, global-ring to middle-ring, and middle-ring to global-ring FIFO queue lengths. Here, $1 \le i \le \frac{N}{L}$ and $1 \le j \le \frac{N}{LM}$. Further, let U_L, U_M, and U_G represent the ring utilizations at local, intermediate, and global rings, respectively. In steady state, Little's result applies and the expected message delays for H2 and H3, T_{H2} and T_{H3}, are:

$$\begin{aligned} T_{H2} &= \frac{\text{Average Number of Packets in System}}{\text{System Throughput}} = \frac{\overline{n}}{\overline{X}} \\ &= \frac{\sum_{i=1}^{N} \overline{Q}_i + \sum_{i=1}^{\frac{N}{L}} (\overline{Q}_{G-L(i)} + \overline{Q}_{L-G(i)})}{N\lambda} \\ &\quad + \frac{(N + \frac{N}{L})U_L + \frac{N}{L}U_G}{N\lambda} \end{aligned} \tag{1}$$

$$\begin{aligned} T_{H3} &= \frac{\text{Average Number of Packets in System}}{\text{System Throughput}} = \frac{\overline{n}}{\overline{X}} \\ &= [\sum_{i=1}^{N} \overline{Q}_i + \sum_{i=1}^{\frac{N}{L}} (\overline{Q}_{M-L(i)} + \overline{Q}_{L-M(i)}) \\ &\quad + \sum_{i=1}^{\frac{N}{LM}} (\overline{Q}_{M-G(i)} + \overline{Q}_{G-M(i)}) + (N + \frac{N}{L})U_L \\ &\quad + (\frac{N}{L} + \frac{N}{LM})U_M + \frac{N}{LM}U_G] \times \frac{1}{N\lambda} \end{aligned} \tag{2}$$

In each equation, $\overline{Y}$ denotes the expected value of the variable Y, the numerator represents the total population (number of packets) in the network, including all FIFO queues and those in the rings. The latter quantity, packets in all rings, is derived from the ring utilizations. The denominator represents the system throughput. An implicit assumption here is that the system is non-saturated and in steady state, making the system throughput equal to the total packet arrival rate.

3.5 Ring Utilizations

In H2 and H3, all local rings have $L + 1$ links, with the extra link being needed to incorporate the inter-ring interface to the intermediate level ring. All global rings have G links; while in H3, intermediate rings have $M + 1$ links, with the extra link incorporating the interface to the global ring.

Because of the destination-remove protocol, it is easy to see that, on average, a message traverses half of the links on any ring it moves over to reach its destination. This assumes that destinations are uniformly distributed inside the local, intermediate, and global sets of messages.

H2: Assuming symmetry over all stations, there are two types of utilizations: U_L for all local rings, and U_G for the global ring.

U_L: To derive U_L, consider a period of T time steps. During this time, there are two sources of traffic onto each local ring: one from local stations Q_i and the other from the global ring through Q_{G-L}. Traffic from

Q_i can be further divided into two parts, namely, those packets staying in the same local ring with probability P, and those going up to the global ring with probability $1-P$. They all use $(L+1)/2$ links on average. Thus traffic from Q_i uses $L\lambda T(L+1)/2$ links over time T.

The total traffic from global ring Q_{G-L} can be calculated as:

$$\lambda_{G-L} = \sum_{1}^{G-1} \frac{L\lambda(1-P)}{(G-1)} = L\lambda(1-P),$$

because $1/(G-1)$ of the global packets from each of the $G-1$ other local rings will be destined for any local ring. Of this traffic, each message uses $(L+1)/2$ links on average. Total number of links used by this traffic over T is $L\lambda(1-P)T(L+1)/2$. Since there are $(L+1)T$ links available over T, we have

$$U_L = \frac{L\lambda}{2} + \frac{L\lambda(1-P)}{2} = \frac{L\lambda(2-P)}{2} \quad (3)$$

$\underline{U_G}$: Each global message uses $G/2$ links on average, and there are GT links available over T. There are a total of $N\lambda(1-P)T$ messages over T, thus

$$U_G = N\lambda(1-P)T\frac{G}{2} \times \frac{1}{GT} = \frac{N\lambda(1-P)}{2} \quad (4)$$

Also note that

$$\lambda_{L-G} = \lambda_{G-L} = L\lambda(1-P) \quad (5)$$

H3: As in H2, consider a period of time T. We define the following locality terms:

"Local": all source traffic staying on the local ring with probability P_L;
"Middle": all source traffic going L $\rightarrow$ M $\rightarrow$ L with probability P_M; and
"Global": all source traffic going L $\rightarrow$ M $\rightarrow$ G $\rightarrow$ M $\rightarrow$ L with probability $1-P_L-P_M$.

$\underline{U_L}$: Over T time steps, there are two sources of traffic going onto each local ring: Q_i and Q_{M-L}. All messages from Q_i, whether L, L $\rightarrow$ M $\rightarrow$ L, or L $\rightarrow$ M $\rightarrow$ G $\rightarrow$ M $\rightarrow$ L bound, use $(L+1)/2$ links on average. Thus traffic from Q_i uses a total of $L\lambda T(L+1)/2$ links over T.

Messages coming down from Q_{M-L} can be divided into two groups:

i) L $\rightarrow$ M $\rightarrow$ L messages from other local rings attached to the same intermediate ring. There are $M-1$ such local rings; and each of them sends $1/(M-1)$ of their L $\rightarrow$ M $\rightarrow$ L traffic to any particular local ring; and each such message uses $(L+1)/2$ links. Hence, over T the number of links used by these messages are:

$$A1 = \sum^{M-1} L\lambda P_M \frac{(L+1)}{2(M-1)}T = L\lambda P_M T\frac{(L+1)}{2}$$

ii) L $\rightarrow$ M $\rightarrow$ G $\rightarrow$ M $\rightarrow$ L messages from all $(N/L)-1$ other local rings; and, arguing as in i), over T the number of links used by these messages are:

$$B1 = \sum^{N/L-1} L\lambda(1-P_M-P_L)\frac{(L+1)}{2(N/L-1)}T$$

$$= L\lambda(1-P_M-P_L)T\frac{(L+1)}{2}$$

But there are $(L+1)T$ links available over T. Therefore, combining link usage from Q_i traffic with $A1$ and $B1$, we have

$$U_L = L\lambda T\frac{L+1}{2} + \frac{A1+B1}{(L+1)T} = \frac{L\lambda(2-P_L)}{2} \quad (6)$$

Note that

$$\lambda_{Q_{M-L}} = \lambda_{Q_{L-M}} = L\lambda(1-P_L) \quad (7)$$

$\underline{U_M}$: There are two sources of traffic going onto each intermediate ring: (1) Up from all M local rings attached to it, through each Q_{L-M}, and (2) Down from the global ring, through Q_{G-M}. Since both L $\rightarrow$ M $\rightarrow$ L and L $\rightarrow$ M $\rightarrow$ G $\rightarrow$ M $\rightarrow$ L traffic classes use $(M+1)/2$ links, the number of links used by the first traffic source (1) over T is:

$$A2 = LM\lambda[P_M + (1-P_M-P_L)]T\frac{(M+1)}{2}$$

$$= LM\lambda(1-P_L)T\frac{(M+1)}{2}$$

The second traffic source is the L $\rightarrow$ M $\rightarrow$ G $\rightarrow$ M $\rightarrow$ L traffic from other intermediate rings; there are $G-1$ of them, and each one sends $1/(G-1)$ of its global traffic to each other intermediate ring. Each such message uses $(M-1)/2$ links. Hence, over T the number of links used by the second traffic source (2) is:

$$B2 = \sum^{G-1} LM\lambda(1-P_M-P_L)\frac{(M+1)}{2}\frac{1}{(G-1)}T$$

$$= LM\lambda T(1-P_M-P_L)\frac{(M+1)}{2}$$

But there are $(M+1)T$ links available over T. Therefore, combining $A2$ and $B2$ we have:

$$U_M = \frac{A2+B2}{(M+1)T} = \frac{LM\lambda}{2}(2-2P_L-P_M) \quad (8)$$

Also note that

$$\lambda_{Q_{L-M}} = \lambda_{Q_{M-L}} = L\lambda(1-P_L) \text{ and} \quad (9)$$

$$\lambda_{Q_{G-M}} = \lambda_{Q_{M-G}} = LM\lambda(1-P_M-P_L) \quad (10)$$

$\underline{U_G}$: Over T there are $N\lambda(1-P_M-P_L)T$ L $\rightarrow$ M $\rightarrow$ G $\rightarrow$ M $\rightarrow$ L messages, each of which uses $G/2$ links; but GT links are available, thus

$$U_G = \frac{N\lambda(1-P_L-P_M)}{2} \quad (11)$$

3.6 Derivation of Average Queue Lengths

Now, we need average queue lengths, $\overline{Q}$, everywhere, for both H2 and H3 systems.

H2:

$\overline{Q_i}$: Waiting (queueing) time at a local station, before getting into the ring link "server" (see Figure 2) will be zero if the upstream link buffer is empty at the time the packet arrives at the head of the line (HOL) position. Service in the first link traversed is counted in the U_L part of the $\overline{n}$ expression in 1, because technically, as soon as the HOL entry starts to get service in the first link, it can be considered that it has been dropped into the empty upstream link buffer.

If p is the probability that a slot is full AND continuing past the current point, then waiting time for the HOL message is:

$$s \triangleq \sum_{j=1}^{\infty} p^j(1-p)j = \frac{p}{1-p}$$

Now, applying Little's Law we get $\overline{Q}_i = W\lambda$, where W is the average waiting time in queue. When a new message arrives, it must wait s time units for each item ahead of it, and then wait s more units. Because of the memoryless property of the stochastic process, we have $W = s + s\overline{Q}_i$. Therefore

$$\overline{Q}_i = (s + s\overline{Q}_i)\lambda$$

$$\overline{Q}_i = \frac{s\lambda}{1-s\lambda}, \text{ for } s = \frac{p}{1-p} \qquad (12)$$

Now, $p = U_L\frac{L-(1+P)}{L}$, where P is locality. This expression for p takes into account the fact that a released slot at either a local station or a inter-ring switch is allowed to be used immediately. Substituting this expression of p into 12, we have

$$\overline{Q}_i = \frac{U_L(L-1-P)\lambda}{L - U_L(L-1-P)(1+\lambda)} \qquad (13)$$

$\overline{Q}_{L-G(i)}$: Similar to local station queue Q_i, the average queue length of the inter-ring switch is:

$$\overline{Q}_{L-G(i)} = \frac{s_x\lambda_{L-G}}{1 - s_x\lambda_{L-G}} \qquad (14)$$

Where $s_x = 1 + \frac{p}{1-p} = \frac{1}{1-p}$. This modification to s is needed for the following reason. One extra unit must be added to the waiting time of every message crossing between rings to denote the step into $Q_{L-G(i)}$ from the adjacent ring link buffer (because it is not accounted for in either U_L or U_G).

Similar to the Q_i discussion, $p = U_G\frac{N/L-2}{N/L}$ and substituting in 14, we have

$$\overline{Q}_{L-G(i)} = \frac{N\lambda_{L-G}}{N(1-\lambda_{L-G}) - U_G(N-2L)} \qquad (15)$$

$\overline{Q}_{G-L(i)}$: We have $p = PL\lambda/2$ because the only traffic continuing on the local ring through the interface switch is local traffic, leading to

$$\overline{Q}_{G-L(i)} = \frac{2\lambda_{G-L}}{2 - PL\lambda - 2\lambda_{G-L}} \qquad (16)$$

H3: Message destination localities are given in terms of the probabilities P_L, P_M, and P_G, where $P_G = 1 - P_L - P_M$. In terms of these probabilities, and using reasoning similar to that used for H2 systems, the resulting average queue lengths in H3 systems are:

$$\overline{Q}_i = \frac{U_L(L-1-P_L)\lambda}{L - U_L(L-1-P_L)(1+\lambda)} \qquad (17)$$

$$\overline{Q}_{L-M(i)} = \frac{\lambda_{Q_{L-M}}}{1 - U_M\frac{M-1-\frac{P_M}{P_M+P_G}}{M} - \lambda_{Q_{L-M}}} \qquad (18)$$

$$\overline{Q}_{M-L(i)} = \frac{2\lambda_{Q_{M-L}}}{2 - P_LL\lambda - 2\lambda_{Q_{M-L}}} \qquad (19)$$

$$\overline{Q}_{M-G(i)} = \frac{\lambda_{Q_{M-G}}}{1 - U_G(G-2)/G - \lambda_{Q_{M-G}}} \qquad (20)$$

$$\overline{Q}_{G-M(i)} = \frac{\lambda_{Q_{G-M}}}{1 - LM\lambda P_M/2 - \lambda_{Q_{G-M}}} \qquad (21)$$

3.7 Expected Message Delay

The expressions for ring utilizations and average queue lengths, developed in Sections 3.5 and 3.6, can now be used in the general model, described in Section 3.4, to derive expressions for the expected message delay in both the 2-level and 3-level ring structures.

Substituting from expressions 3 and 4 for U_L and U_G, from expressions 13, 15, and 16 for average queue lengths, and from the expression 5 for message rates λ_{G-L} and λ_{L-G}, into expression 1 for expected message delay T_{H2}, and performing a number of algebraic rearrangements and simplifications of terms, leads to:

$$T_{H2} = T_1 + PT_2 + (1-P)(T_3 + T_4 + T_5) + 1 \qquad (22)$$

where $T_1 = \frac{X}{1-X(1+\lambda)}$, for $X = (\lambda/2)(2-P)(L-1-P)$, $T_2 = \frac{L+1}{2}$, $T_3 = \frac{1}{1-N\lambda(1-P)/2}$, $T_4 = \frac{1}{1-L\lambda(2-P)/2}$, and $T_5 = (L+1) + \frac{G}{2}$.

In this form, T_1 represents average waiting time in the local (source) station interface queue, Q_i; T_2 represents average path length for a local message; T_3 represents average waiting time in $Q_{L-G(i)}$ for a remote message moving up from a (source) local ring to the global ring; T_4 represents average waiting time in $Q_{G-L(i)}$ for a remote message moving down from the global ring to a (destination) local ring; and T_5 represents average path length for a remote message. The final "1" term in the T_{H2} expression 22 represents the time step needed to move a message from the ring

buffer at the destination station into the station interface, as indicated in Figure 2. (This term was not accounted for in the earlier expression 1 for T_{H2}.)

A similar sequence of substitutions (using expressions 6, 8, and 11 for ring utilizations, expressions 17, 18, 19, 20, and 21 for average queue lengths, and expressions 7, 9, and 10 for message rates at crossovers) and algebraic rearrangements and simplifications can be used to derive the following expression for expected message delay in 3-level ring structures. The final result is:

$$\begin{aligned} T_{H3} &= T_6 + P_L T_7 + P_M(T_8 + T_9 + T_{10}) \\ &\quad + P_G(T_8 + T_9 + T_{11} + T_{12} + T_{13}) + 1 \end{aligned} \quad (23)$$

where $T_6 = \frac{Y}{1-Y(1+\lambda)}$, for $Y = \frac{\lambda}{2}(2-P_L)(L-1-P_L)$, $T_7 = \frac{L+1}{2}$,

$$T_8 = \frac{1}{1-[\frac{L\lambda}{2}(2P_G+P_M)(M-1-\frac{P_M}{P_M+P_G})+L\lambda(P_M+P_G)]},$$

$T_9 = \frac{1}{1-L\lambda(2-P_L)/2}$, $T_{10} = (L+1) + \frac{M+1}{2}$, $T_{11} = \frac{1}{1-N\lambda P_G/2}$, $T_{12} = \frac{1}{1-LM\lambda(2P_G+P_M)/2}$, and $T_{13} = (L+1)+(M+1)+G/2$.

As with the T_{H2} expression 22, each of the terms in 23 for T_{H3} has an interpretation that is directly related to the network. Briefly, T_6 represents local station queueing delay; T_7, T_{10}, and T_{13} represent path lengths for local, intermediate, and global messages, respectively; T_8 and T_9 represent the up-queue and down-queue delays in switches between local ring and intermediate rings; and T_{11} and T_{12} represent up and down queueing delays between intermediate rings and the global ring.

4 Validation of the Analytical Models via Simulations

In this section we validate our analytical model through extensive simulations. In the simulation study, reported in [8], an event-driven simulator was used to study 2-level and 3-level hierarchical ring systems. All the simulation results presented here have very small 95% confidence intervals and so these intervals are not shown.

In Figure 4, results for an H2 system are plotted to show expected packet delay as a function of λ and locality. Since the global ring saturates faster than any other ring in the system, we also included its utilization. We were not able to compare the case of $P = 0.2$ and $\lambda > 0.004$ because the system entered saturation soon after that point. Nevertheless, it is clear from the figure that our model is very accurate with the exception of two points where errors of 7% and 14% occur at global utilizations of 80% and 90%, respectively. This discrepancy can be explained as a result of our model's inability to capture the "train effects" (see Section 3.2) at the near-saturated global ring conditions.

Figure 5 shows a comparison between our model and the simulations for an H3 system. Consistent with the case of H2, our model agrees very well with the simulation. In fact, the agreement in this case is better

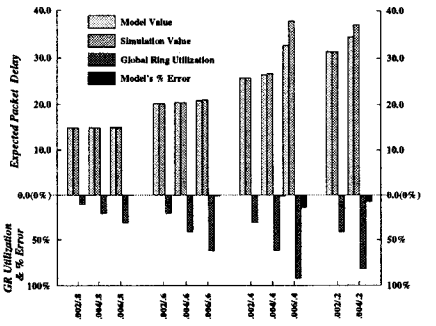

Figure 4: Comparison between the Model and Simulation for an H2 system where $N = 512$ and $L = 16$; that is, 32 local rings with 16 stations each.

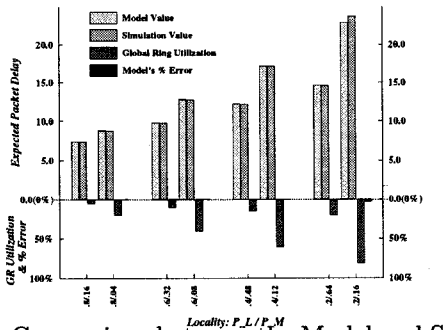

Figure 5: Comparison between the Model and Simulation for an H3 system where $N = 504$, $L = 7$, $M = 6$, $G = 12$ and $\lambda = 0.005$.

than H2. The improved accuracy may be viewed as a result of the "diluting" effect of the 3-level rings that alleviates the "train effects", thus making our model more accurate.

Our final comparison between model and simulation is shown in Figure 6, again revealing very good agreement except at high global ring utilization levels.

The more important point brought out by Figure 6, however, relates to the relationship between average message delay performance and network configuration at different traffic levels. Consider the following. Assume a distribution of message packet destinations that is characterized by the application, not related to network configuration. For example, in the uniform distribution, all processor nodes are equally likely as destination of a message packet. This presents the most demanding case for any multiprocessor network. There is no locality that can be exploited.

Figure 6 shows such a case. N is close to 400 for all three configurations. As the configurations (L, M, G) vary, P_L, P_M, and P_G, must also vary to properly reflect a uniform message destination distribution.

The figure reveals that for light traffic ($\lambda = 0.001$), the $(L, M, G) = (6, 6, 11)$ configuration provides a lower average message delay than the $(10, 10, 4)$ configuration; while for heavy traffic ($\lambda = 0.005$, and global ring utilizations upwards of 75%), the opposite is true. In general, we have shown earlier [6] that the configuration leading to the lowest maximum distance be-

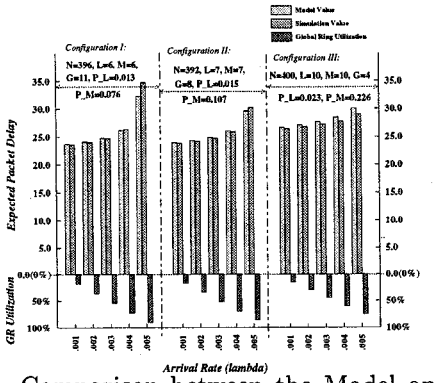

Figure 6: Comparison between the Model and Simulation for three H3 configurations, with a uniform distribution of message destinations.

tween any pair of nodes (the minimum diameter network) has L, M, and G sizes in proportions $1:1:2$. This is consistent with the $(6,6,11)$ configuration having the lowest average delay in the light traffic (and thus low contention) case. Correspondingly, in [12] an independent detailed simulation study of H3 systems showed that the best configurations for the heavy uniform traffic case (under their method for developing feasible configurations under heavy traffic) all had relatively small global rings. In particular, they derived $(L,M,G)=(6,3,3)$ for a particular $N=54$ network, and $(12,3,3)$ for an $N=108$ network. This tendency is qualitatively similar to our result that $(10,10,4)$ is better than $(6,6,11)$ for the heavy traffic case.

We will expand on this use of the model in configuration design in the next section.

5 Optimal Configurations

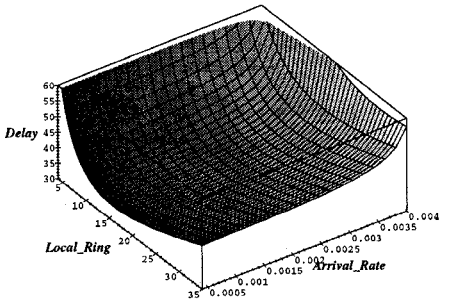

Figure 7: 3-D Plot for H2 Delay with $N=500$ and Uniform Message Destination Distributions

One very important issue in the design of hierarchical-ring systems is that of configuration. Our analytical model can predict expected packet delay accurately. It can now be used to answer the logical question: What is the best configuration for the hierarchical-ring network to minimize the average delay, given a particular application-based traffic pattern and system size? A quick answer to this question can be very helpful in enabling the system architect/designer to make sensible design decisions. The answer to the question may be found by deriving optimal values for L in H2, and L and M in H3, that minimize T_{H2} and T_{H3}, respectively.

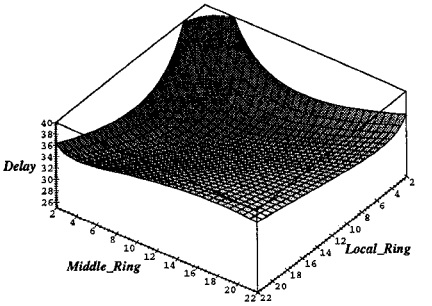

Figure 8: 3-D Plot for H3 Delay with $N=500$, $\lambda=0.002$ and Uniform Message Destination Distributions

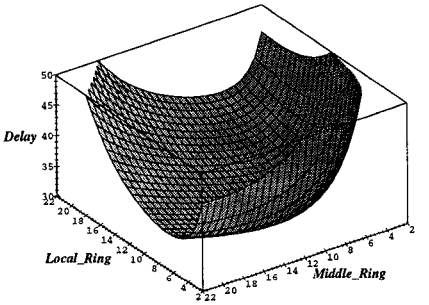

Figure 9: 3-D Plot for H3 Delay with $N=500$, $\lambda=0.004$ and Uniform Message Destination Distributions

The expressions for T_{H2} and T_{H3} are closed form functions of N, L, M, and traffic, which is uniquely defined by values of λ and locality (P, P_L, and P_M). Therefore, if one has some knowledge of the density (λ) and pattern (locality) of the traffic which the future system will likely be subject to, then for a given system size (N) it is possible to find values of L (for H2) and L and M (for H3) that minimize T_{H2} and T_{H3}, respectively, for given values of λ and application-based traffic locality. In this section, we show how expressions 22 and 23 can be used to find optimal values of L and M. All 3-D plots in this section were generated using the Maple-V software [2]. The design optimization question, as we have posed it, only makes sense if we are able to show how the physical network locality parameters P_L, P_M, and P_G $(=1-P_L-P_M)$, are functionally related to N, L, M, and $G=N/LM$, for a given application-based locality specification. As an example, we will deal with the uniform message destination case here. This is simply the case in which all other $N-1$ nodes are equally likely as message destinations from any particular source node. This traffic distribution is reflected in the following functional relationships: In H2, $P=(L-1)/(N-1)$; and in H3, $P_L=(L-1)/(N-1)$, $P_M=(M-1)L/(N-1)$, and $P_G=1-P_L-P_M=(G-1)LM/(N-1)$. These substitutions are made in T_{H2} and T_{H3} before plotting the Maple-V surfaces.

Figure 7 shows a 3-D plot of T_{H2} as a function of L and λ while the traffic pattern is uniform and $N=500$. In this figure, traffic density λ ranges from 0.0005, representing light traffic, to 0.004, representing the heavier traffic. As can be seen in the figure, there is an optimum of L for each λ value. For light traffic, L

is optimal near 16, shifting to larger values as λ increases.

In Figures 8 and 9 we plot T_{H3} as a function of L and M for $\lambda = 0.002$ and $\lambda = 0.004$, respectively, while keeping the traffic pattern uniform and $N = 500$. As expected, for each λ value there is a pair of optimal L and M values. In fact, for $\lambda = 0.002$ the optimal values for L and M are 6 and 7, respectively; whereas for $\lambda = 0.004$ values of 9 and 10 for L and M, respectively, minimize T_{H3}.

6 Concluding Remarks

Network configuration, that is, appropriate choices for the size of local, intermediate, and global rings, can be quickly and easily estimated by using the queueing models developed here, without resorting to time-consuming simulations, assuming that minimizing average message delay is the important criterion. We gave an example of such a design study in the previous section. As we noted, network optimization is only meaningful relative to a specified traffic intensity and message destination distribution that is determined by the application. In Section 5 we used a uniform distribution, which is easy to incorporate into the model. For more general application-based distributions, such as those described in [7], we have shown in [8] how to incorporate them into a simple model that is, however, only valid for very light traffic (no significant contention at crossover switches). We are currently incorporating the general distribution specifications into the queueing models, enabling wider use of the models in design evaluations.

References

[1] L. N. Bhuyan, D. Ghosal, and Q. Yang. Approximate analysis of single and multiple ring networks. *IEEE Transactions on Computers*, C-38(7):1022–1040, July 1989.

[2] B.W. Char, K.O. Geddes, G.H. Gonnet, B.L. Leong, M.B. Monagan, and S.M. Watt. *Maple V Language Reference Manual.* Spring-Verlag and Waterloo Maple Publishing, 1991.

[3] D. R. Cheriton, H. A. Goosen, and P. D. Boyle. Paradigm: A highly scalable shared- memory multicomputer architecture. *Computer*, 24(2):33–46, February 1991.

[4] T. H. Dunigan. Multi-ring performance of the kendall square multiprocessor. *Oak Ridge National Laboratory Report TM-12331*, October 1994.

[5] K. Farkas, Z. Vranesic, and M. Stumm. Scalable cache consistency for hierarchically structured multiprocessors. *The Journal of Supercomputing*, 8:345–369, June 1995.

[6] V. C. Hamacher and H. Jiang. Comparison of mesh and hierarchical networks for multiprocessors. *Proceedings of 1994 International Conference on Parallel Processing*, I:67–71, August 1994.

[7] M. Holliday and M. Stumm. Performance evaluation of hierarchical ring-based shared memory multiprocessors. *IEEE Transactions on Computers*, C-43(1):52–67, January 1994.

[8] H. Jiang, C. Lam, and V.C. Hamacher. On some architectural issues of optical hierarchical ring networks for shared-memory multiprocessors. *Proceedings of The Second International Conference on Massively Parallel Processing Using Optical Interconnections (MPPOI)*, pages 345–353, October 23-24 1995.

[9] P. J. B. King and I. Mitrani. Modeling a slotted ring local area network. *IEEE Transactions on Computers*, C-36(5):554–561, May 1987.

[10] C. Lam, H. Jiang, and V. C. Hamacher. Design and analysis of hierarchical ring networks for multiprocessors. *Proceedings of 1995 International Conference on Parallel Processing*, I:46–50, August 14-19 1995.

[11] W. M. Loucks, V. C. Hamacher, B. R. Preiss, and L. Wong. Short-packet transfer performance on local area ring networks. *IEEE Transactions on Computers*, C-34(11):1006–1014, November 1985.

[12] G. Ravindran and M. Stumm. A performance comparison of hierarchical ring- and mesh-connected multiprocessor networks. *Proceedings of 1997 International Symposium on High Performance Computer Architecture*, pages 58–69, February 1997.

[13] T.D. Todd and A.M. Bignell. Performance modeling of the SIGnet MAN backbone. *Proc. IEEE INFOCOM'90*, 1:192–199, June 1990.

[14] Z. Vranesic, S. Brown, and M. Stumm. The NUMAchine Multiprocessor. *Technical Report, Department of Electrical and Computer Engineering, University of Toronto*, June 1995.

[15] Z. G. Vranesic, M. Stumm, D. M. Lewis, and R. White. Hector: A hierarchically structured shared-memory multiprocessor. *IEEE Computer*, 24(1):72–79, January 1991.

[16] A. W. Wilson. Hierarchical cache/bus architecture for shared memory multiprocessors. In *Proceedings of the 14th Annual International Symposium on Computer Architecture*, page 1987, 1987.

[17] X. Zhang and Y. Yan. Comparative modeling and evaluation of CC-NUMA and COMA on hierarchical ring architectures. *IEEE Transactions on Parallel and Distributed Systems*, 6:1316–1331, December 1995.