1-2001

Hierarchical Ring Network Configuration and Performance Modeling

V. Carl Hamacher
hamacher@eleceng.ee.queensu.ca

Hong Jiang
University of Nebraska - Lincoln, jiang@cse.unl.edu

Hierarchical Ring Network Configuration and Performance Modeling

V. Carl Hamacher, *Senior Member*, *IEEE*, and Hong Jiang, *Member*, *IEEE*

Abstract—Approximate analytical queuing network models for expected message packet delay in 2-level and 3-level hierarchical ring interconnection networks (INs) are developed. A major class of traffic carried by these INs consists of cache line transfers between processor caches and remote memory modules in shared-memory multiprocessors. Such traffic consists of short, fixed-length messages; they can be conveniently transported by the slotted-ring transmission technique which is studied here. The packet delay results derived from the models are shown to be quite accurate when checked against a simulation study. As well as facilitating analysis, the analytical models can be used to determine optimal sizes for the rings at different levels in the hierarchy, where optimality is in terms of minimizing average packet delay.

Index Terms—Interconnection networks, hierarchical rings, slotted rings, shared-memory multiprocessors, queuing models, message-passing performance.

1 Introduction

A main hardware component in a multiprocessor system is the interconnection network (IN) that connects together processors and 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 [14], [18], [15], [5], [9], [10], [12], [6] and in industry [17], [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 [14], [9], 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 [18], [9]. Nevertheless, a more popular choice for INs seems to be meshes. This, as noted in [14], 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, two comparative studies of hierarchical rings and meshes in the literature, one based on approximate modeling [6] and the other based on detailed execution-driven simulations [14], concluded that hierarchical rings can outperform meshes under some practical workloads. More specifically, [14] found that hierarchical rings perform significantly better than meshes for system sizes up to about 120 processor nodes if the workload exhibits moderate to high memory access locality. Even if there is no memory locality, [14] 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 [13], [1]. As a result, analytical studies of such networks have been based on approximation techniques [13], [1], [18]. With the exception of [18], which analyzed 2-level structures, hierarchical ring structures have not been studied analytically so far despite the existence of analytical studies in the literature on single-level rings [13], [1]. In [13], buffering and queuing effects were not included at the input ports and contention for slot access was modeled only in the single-level ring case. Bhuyan et al. [1] extended the model in [13] to incorporate buffers at input ports and to consider a double-ring system where two unidirectional slotted rings were put in parallel. Zhang and Yan [18] analyzed a 2-level hierarchical ring system with emphasis on finding relative performances of a few cache coherence protocols and the impact of hot spot traffic. Thus, the models developed in [18] were geared toward specific coherence protocols under the hot spot traffic condition. Further, all models in [13], [1], [18] assumed a source removal packet transfer protocol. Two other recent performance studies on hierarchical ring networks were based entirely on simulations [9], [14].

In this paper, we use approximate analytical techniques to model the packet delay performance of 2-level and 3-level hierarchical ring networks that operate under a full range of applied load conditions and a destination removal protocol, as opposed to source removal. The destination removal protocol is more efficient in terms of network channel utilization and has been employed in recent research prototypes [16], [15]. The model is used to gain

- *V.C. Hamacher is with the Department of Electrical and Computer Engineering, Queen's University, Kingston, Ontario, Canada K7L 3N6. E-mail: hamacher@eleceng.ee.queensu.ca.*
- *H. Jiang is with the Department of Computer Science and Engineering, University of Nebraska-Lincoln, Lincoln, NE 68588-0115. E-mail: jiang@cse.unl.edu.*

Manuscript received 24 Mar. 1997; revised 18 Aug. 1998; accepted 15 May 2000.

For information on obtaining reprints of this article, please send e-mail to: tc@computer.org, and reference IEEECS Log Number 104744.

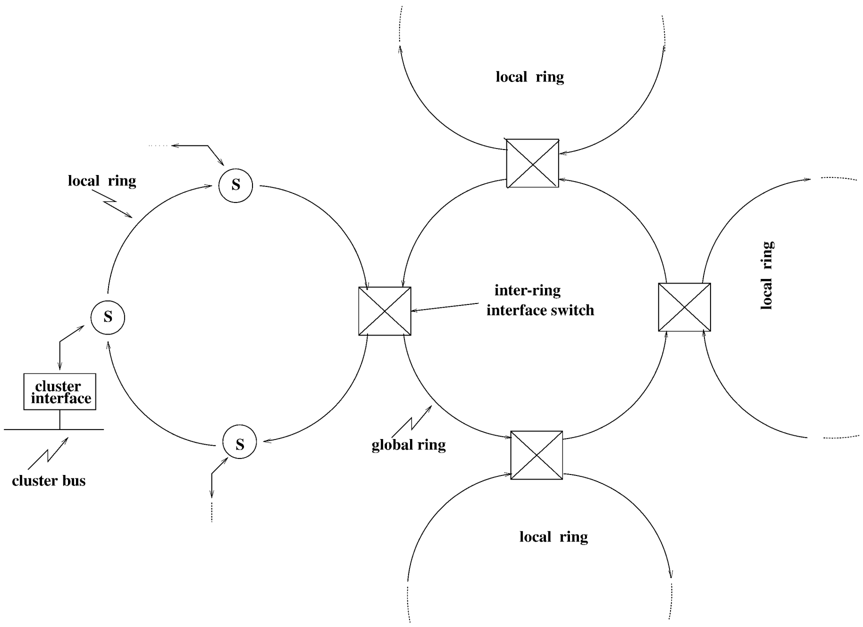


Fig. 1. A 2-level hierarchically structured multiprocessor.

important insights 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 packet delay? The effect of doubling the bandwidth of the global ring, after that ring is shown to be a traffic bottleneck, is also determined.

The paper is organized as follows: Section 2 presents a description of the hierarchical ring interconnection network model, including enough structural and operational detail for performance evaluation purposes. Section 3 develops packet delay models using queuing models to capture the effect of contention. The analytical models developed are validated through extensive simulations in Section 4. Section 5 addresses the issue of optimal configuration using the analytical models developed in Section 3. We also include the effect of doubling the bandwidth of the global ring. Finally, some concluding remarks and prospects for future work are made in Section 6. An earlier version of this paper, with a slightly different analytical model, appeared in [7].

2 Hierarchical Ring Networks

The hierarchical slotted-ring IN studied here consists of unidirectional rings, as employed in [3], [4], [15], [16]. Processor node clusters are only connected to local rings, as shown in Fig. 1. Each segment, called a station, connects one cluster into the ring. The station switch, S, removes an incoming ring packet into its cluster interface if it is the destination or sends the packet on around the ring otherwise. This packet-handling protocol is the same as that used in destination-remove, slotted, Local Area Networks (LANs) [13]. The switch also introduces a pending transmit packet 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, packets have 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 acknowledgment 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 acknowledgment message. A Read operation requires three messages: one to send the Read request, an acknowledgment, 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.

The bit width of a slot in 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] and we will refer to this slot quantity of information as a packet. Current cache line sizes in multiprocessor systems consist of multiple words [8], which will not fit into one slot. Therefore, cache line transfers would need to consist of multiple-packet messages. This presents no difficulty for the IN described here because a wide-slot packet is large enough to contain source and destination node addresses and can therefore move through the IN as an independently routed unit. Also, the order of packets from any one source is maintained as they reach their destination.

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

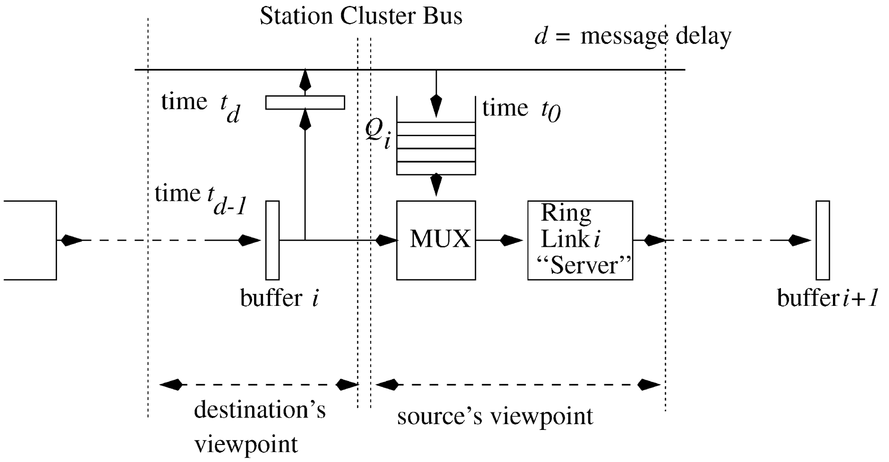


Fig. 2. Structure of local station interface.

standpoint, packet 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 Fig. 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. Packet blocking can occur at the crossover switch between two rings. For example, in a 2-level system, if a packet from a local ring needs to move up to the global ring at the same time that a continuing packet on the global ring arrives at the crossover switch, there is contention for the downstream link on the global ring, and only one packet can proceed. The other packet must be temporarily buffered in the crossover switch to insure that packets are never lost in the network. Details will be given in Section 3.2.

3 Contention (Queuing) Model for Packet Delay

In [6], [12], we developed packet 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 a full range of applied loads.

3.1 Packet 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 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 packet delay. An example of this network design optimization is given in Section 5.

In the models to be developed, the following parameters reflect packet destination locality. In H2 (2-level systems), P is the probability that a packet 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 packet 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 packet 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 interring interface, as shown in Fig. 2 and Fig. 3, respectively. At a station interface, shown in Fig. 2, the 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 interring interface otherwise. At the interring interface, shown in Fig. 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 accesses the first empty slot and moves around the ring to join the FIFO queue at another interring 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 packet delay, d (see Fig. 2), of a packet is the sum of

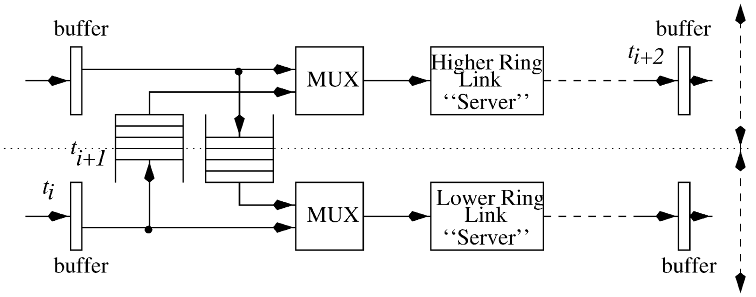


Fig. 3. Structure of interring interface.

1. *queuing 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 packet 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 packet 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 [13], [1], where, as traffic intensifies, full slots tend to cluster together to form "trains" of slots, as opposed to full slots being uniformly and independently distributed on the rings. This dependence makes an exact analysis intractable [11]. 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 in [13], [1], [18] and by our own simulation validation studies.

With the above assumptions, we model the contention in parts 1 and 2 of packet delay using the M/G/1 queuing center model, similar to the approach in [1] and [18] 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 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 Figs. 2 and 3 in narrow bars. It only needs to have capacity for one packet because:

1. Packet arrivals occur only at discrete time points and the associated ring link "server" has a constant service time of 1 discrete time step; and
2. This ring link buffer has priority over station FIFO queues and interring crossover queues in competing for access to the ring link "server." This priority policy is consistent with the implementations of the NUMAchine [15], [14] 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 packet traffic and packet travel patterns. This will become clear later.

3.3 Definitions and Assumptions

Time is discretized into clock ticks. One tick is the time needed for a packet to move between adjacent slot segments in any ring or from a ring link buffer to a FIFO queue in an interring interface (see Fig. 3) or from a ring link buffer to a station cluster bus buffer at a destination (see Fig. 2). 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 packets per clock tick arriving at a local ring station FIFO queue,
2. packet destination locality in H2 is determined by probability P as defined in Section 3.1,
3. packet 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 3-level networks,
7. G: number of lower-level rings connected directly 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 interring interface FIFO follows a Poisson process.
2. One 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 Fig. 2).

3.4 General Model

The basic idea of this analysis is to solve the M/G/1 queuing model for all FIFO queues (local stations and interring interfaces), which will give rise to expected queue lengths at all FIFO queues. In order to do this, we need ring utilizations. Using Little's result, these results can then be used to derive expected packet delays as follows.

Let Q_i, $1 \leq i \leq N$, denote the queue length at 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 interring interface i, $1 \leq i \leq \frac{N}{L}$ 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, $1 \leq i \leq \frac{N}{L}$ and $1 \leq j \leq \frac{N}{LM}$. Further, let U_L, U_M, and U_G represent the ring link utilizations in local, intermediate, and global rings, respectively. In steady state, Little's result applies and the expected packet delays for H2 and H3, T_{H2} and T_{H3}, are:

$$T_{H2} = \frac{\text{Average Number of Packets in System}}{\text{System Throughput}} =$$

$$\frac{\sum_{i=1}^{N}\overline{Q}_i + \sum_{i=1}^{\frac{N}{L}}(\overline{Q}_{G-L(i)} + \overline{Q}_{L-G(i)} + 2L\lambda(1-P)) + (N+\frac{N}{L})U_L}{N\lambda}$$

$$\frac{+\frac{N}{L}U_G + N\lambda}{N\lambda}. \tag{3.1}$$

$$\begin{aligned} T_{H3} &= \frac{\text{Average Number of Packets in System}}{\text{System Throughput}} \\ &= \Bigg[\sum_{i=1}^{N}\overline{Q}_i + \sum_{i=1}^{\frac{N}{L}}(\overline{Q}_{M-L(i)} + \overline{Q}_{L-M(i)} + 2L\lambda(1-P_L)) \\ &\quad + \sum_{i=1}^{\frac{N}{LM}}(\overline{Q}_{M-G(i)} + \overline{Q}_{G-M(i)} \\ &\quad + 2LM\lambda P_G) + \left(N + \frac{N}{L}\right)U_L + \left(\frac{N}{L} + \frac{N}{LM}\right)U_M \\ &\quad + \frac{N}{LM}U_G + N\lambda\Bigg] \times \frac{1}{N\lambda}. \end{aligned} \tag{3.2}$$

In each equation, $\overline{Y}$ denotes the expected value of the variable Y. The numerator in (3.1) represents the total population (of packets) in FIFO queues, interring interfaces, and rings. Aside from average queue lengths, $\overline{Q}$, interface and ring packet occupancies are accounted for as follows: The term $2L\lambda(1-P)$ accounts for packets in the two links leading from ring buffers to FIFO queues, as shown in Fig. 3 (the step from t_i to t_{i+1}). The terms $(N+\frac{N}{L})U_L$ and $\frac{N}{L}U_G$ account for packets in all local rings and the global ring, respectively, and the term $N\lambda$ accounts for packets in all links leading into station cluster bus buffers, as shown in Fig. 2 (the step from t_{d-1} to t_d). The denominator represents the system throughput. An implicit assumption here is that the system is nonsaturated and in steady state, making the system throughput equal to the total packet arrival rate. Similar comments apply to the terms in T_{H3}.

3.5 Ring Utilizations

In H2 and H3, all local rings have $L+1$ links, with the extra link being needed to incorporate the interring 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 packet 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 packets.

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 packet 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

$$U_L = \frac{L\lambda}{2} + \frac{L\lambda(1-P)}{2} = \frac{L\lambda(2-P)}{2}. \tag{3.3}$$

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

$$U_G = N\lambda(1-P)T\frac{G}{2} \times \frac{1}{GT} = \frac{N\lambda(1-P)}{2}. \tag{3.4}$$

Also note that

$$\lambda_{L-G} = \lambda_{G-L} = L\lambda(1-P). \tag{3.5}$$

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 → M → L with probability P_M; and

"Global": all source traffic going L → M → G → M → L with probability $1 - P_L - P_M$.

U_L: Over T time steps, there are two sources of traffic going onto each local ring: Q_i and Q_{M-L}. All packets from Q_i, whether L, L → M → L, or L → M → G → M → 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.

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

1. L → M → L packets 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 → M → L traffic to any particular local ring and each such packet uses $(L+1)/2$ links. Hence, over T, the number of links used by these packets is:

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

2. L → M → G → M → L packets from all $(N/L)-1$ other local rings and, arguing as in 1, over T the number of links used by these packets is:

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

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

$$\begin{aligned} U_L &= L\lambda T\frac{L+1}{2} + \frac{A1+B1}{(L+1)T} \\ &= \frac{L\lambda}{2} + \frac{L\lambda P_M}{2} + \frac{L\lambda(1-P_M-P_L)}{2} \\ &= \frac{L\lambda}{2} + \frac{L\lambda(1-P_L)}{2} = \frac{L\lambda(2-P_L)}{2}. \end{aligned} \tag{3.6}$$

Note that

$$\lambda_{M-L} = \lambda_{L-M} = L\lambda P_M + L\lambda(1-P_M-P_L) = L\lambda(1-P_L). \tag{3.7}$$

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 → M → L and L → M → G → M → L traffic classes use $(M+1)/2$ links, the number of links used by the first traffic source 1) over T is:

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

The second traffic source (2) is the L → M → G → M → 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 packet uses $(M-1)/2$ links. Hence, over T, the number of links used by the second traffic source (2) is:

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

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

$$\begin{aligned} U_M &= \frac{A2+B2}{(M+1)T} = \frac{LM\lambda(1-P_L)}{2} + \frac{LM\lambda(1-P_M-P_L)}{2} \\ &= \frac{LM\lambda}{2}(2-2P_L-P_M). \end{aligned} \tag{3.8}$$

Also note that

$$\lambda_{G-M} = \lambda_{M-G} = LM\lambda(1-P_M-P_L). \tag{3.9}$$

U_G: Over T there are $N\lambda(1-P_M-P_L)T$ packets that follow the L → M → G → M → L path, each of which uses $G/2$ links; but GT links are available, thus

$$U_G = \frac{N\lambda(1-P_L-P_M)}{2}. \tag{3.10}$$

3.6 Derivation of Average Queue Lengths

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

H2:

$\overline{Q}_i$: Slot access time at a local station will be 0 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 component of (3.1) 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 slot access time for the HOL message packet 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 the queue. When a new packet arrives, it must wait s time units for each item ahead of it and then wait s more units for its own service. Because of the memoryless property of the stochastic process, we have $W = s + s\overline{Q}_i$. Therefore,

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

$$\text{so} \quad \overline{Q}_i = \frac{s\lambda}{1-s\lambda}, \quad \text{for} \quad s = \frac{p}{1-p}. \tag{3.11}$$

The probability that a slot is full is U_L. The probability that it is continuing past the current point can be shown to be $\frac{L-(1+P)}{L}$ by a detailed consideration of the possible source and destination of each packet that appears in the input side link buffer of a local station. Therefore,

$$p = U_L \frac{L - (1 + P)}{L}. \tag{3.12}$$

$\overline{Q}_{L-G(i)}$: Similar to the local station queue Q_i, the average queue length of the upper-going FIFO in an interring switch is:

$$\overline{Q}_{L-G(i)} = \frac{s\lambda_{L-G}}{1 - s\lambda_{L-G}} \quad \text{for} \quad s = \frac{p_{L-G}}{1 - p_{L-G}}, \tag{3.13}$$

where

$$p_{L-G} = U_G \frac{G - 2}{G}. \tag{3.14}$$

$\overline{Q}_{G-L(i)}$: The average queue length of the downward-going FIFO in an interring switch is

$$Q_{G-L(i)} = \frac{s\lambda_{G-L}}{1 - s\lambda_{G-L}}, \quad \text{for} \quad s = \frac{p_{G-L}}{1 - p_{G-L}}, \tag{3.15}$$

where

$$p_{G-L} = \frac{PL\lambda}{2} \tag{3.16}$$

because the only traffic continuing on the local ring through the interface switch is local traffic.

H3: Packet destination localities are given in terms of the probabilities P_L, P_M, and P_G, where $P_G = 1 - P_L - P_M$.

The average length of the input queue, $\overline{Q}_i$, is the same as in (3.11), with $p = U_L \frac{L-(1+P_L)}{L}$.

For the other four queues, the average lengths $\overline{Q}_{L-M(i)}$, $\overline{Q}_{M-L(i)}$, $\overline{Q}_{M-G(i)}$, and $\overline{Q}_{G-M(i)}$, have expressions similar to (3.13), with traffic rate factors λ_{L-M}, λ_{M-L}, λ_{M-G}, and λ_{G-M}, respectively. The p factors are

$$\begin{aligned} p_{L-M} &= U_M \left[\frac{M - (1 + \frac{P_M}{P_M + P_G})}{M} \right] \quad \text{for} \quad \overline{Q}_{L-M(i)} \\ p_{M-L} &= \frac{L\lambda P_L}{2} \quad \text{for} \quad \overline{Q}_{M-L(i)} \\ p_{M-G} &= U_G \frac{G-2}{G} \quad \text{for} \quad \overline{Q}_{M-G(i)} \\ \text{and} \quad p_{G-M} &= \frac{LM\lambda P_M}{2} \quad \text{for} \quad \overline{Q}_{G-M(i)}. \end{aligned}$$

3.7 Expected Packet Delay

The expressions for ring utilizations, traffic rates, 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.

The required sequence of substitutions is as follows in converting the global expression (3.1) for T_{H2} into an explicit expression involving only the structural parameters N, L and $G = N/L$ and the traffic parameters λ and P: First, substitute from (3.3) for U_L into (3.12) for p and then substitute this explicit expression for p into (3.11) to obtain an explicit expression for $\overline{Q}_i$. Similarly, use (3.4), (3.14), (3.5), and (3.13) to obtain an explicit expression for $\overline{Q}_{L-G(i)}$ and use (3.16), (3.5), and (3.15) to obtain an explicit expression for $\overline{Q}_{G-L(i)}$. Then, use these three average queue length expressions, along with (3.3) and (3.4) for U_L and U_G in (3.1) to derive an explicit expression for T_{H2}.

After performing a number of algebraic simplifications, we have

$$T_{H2} = T_1 + PT_2 + (1 - P)(T_3 + T_4 + T_5) + 1, \tag{3.17}$$

where

$$\begin{aligned} T_1 &= \frac{X}{1 - X(1 + \lambda)} \quad \text{for} \quad X = \frac{\lambda}{2}(2 - P)[L - (1 + P)] \\ T_2 &= \frac{L+1}{2} \\ T_3 &= \frac{Y(\frac{N}{L} - 2)}{2 - (1 + Y)Y(\frac{N}{L} - 2)} \quad \text{for} \quad Y = L\lambda(1 - P) \\ T_4 &= \frac{PL\lambda}{2 - PL\lambda[1 + L\lambda(1 - P)]} \\ T_5 &= 2 + (L + 1) + \frac{N}{2L}. \end{aligned}$$

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 packet; T_3 represents average waiting time in $Q_{L-G(i)}$ for a remote packet 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 packet moving down from the global ring to a (destination) local ring; and T_5 represents average path length for a remote packet. The final "1" term in the T_{H2} expression represents the time step needed to move a packet from the ring buffer at the destination station into the station interface, as indicated in Fig. 2.

A similar sequence of substitutions (using (3.6), (3.8), and (3.10) for ring utilizations, expressions similar to (3.11) and (3.13) for average queue lengths along with corresponding p factors, and (3.7) and (3.13) for packet rates at crossovers) and algebraic rearrangements and simplifications can be used to derive the following expression for expected packet delay in 3-level ring structures. The final result is:

$$\begin{aligned} T_{H3} = T_6 + P_L T_7 + P_M(T_8 + T_9 + T_{10}) \\ + P_G(T_8 + T_9 + T_{11} + T_{12} + T_{13}) + 1, \end{aligned} \tag{3.18}$$

where

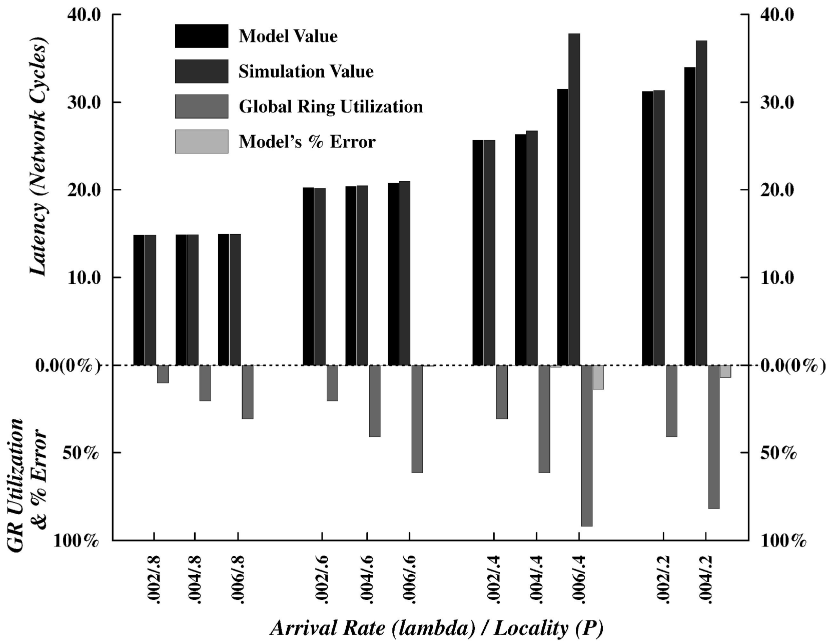


Fig. 4. Comparison between the model and simulation for an H2 system where $N = 512$ and $L = 16$; that is, 32 local rings with 16 stations each.

$$T_6 = \frac{PZ}{1 - Z(1+\lambda)} \quad \text{for} \quad Z = \frac{\lambda}{2}(2 - P_L)[L - (1 + P_L)]$$

$$T_7 = \frac{L+1}{2}$$

$$T_8 = \frac{1}{(1/p_{L-M}) - (1 + \lambda_{L-M})}$$

$$\text{with} \quad p_{L-M} = U_M \left[\frac{M - (1 + \frac{P_M}{P_M + P_G})}{M} \right],$$

$$U_M = \frac{LM\lambda(2P_G + P_M)}{2}, \quad \text{and}$$

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

$$T_9 = \frac{L\lambda P_L}{2 - L\lambda P_L[1 + L\lambda(1 - P_L)]}$$

$$T_{10} = (L+1) + \frac{M+1}{2} + 2$$

$$T_{11} = \frac{\lambda P_G(N - 2LM)}{2 - (1 + LM\lambda P_G)\lambda P_G(N - 2LM)}$$

$$T_{12} = \frac{LM\lambda P_M}{2 - LM\lambda P_M(1 + LM\lambda P_G)}$$

$$T_{13} = (L+1) + (M+1) + \frac{N}{2LM} + 4.$$

As with the T_{H2} expression (3.17), each of the terms in (3.18) for T_{H3} has an interpretation that is directly related to the network. Briefly, T_6 represents local station queuing delay; T_7, T_{10}, and T_{13} represent path lengths for local, intermediate, and global packets respectively; T_8 and T_9 represent the up-queue and down-queue delays in switches between local rings and intermediate rings; and T_{11} and T_{12} represent up and down queuing 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 [12], 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 percent confidence intervals and, so, these intervals are not shown.

In Fig. 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 8.3 percent and 16.7 percent occur at global ring utilizations of 82 percent and 92 percent, 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.

Fig. 5 shows a comparison between our model and the simulations for an H3 system. As with the case of H2, our model agrees very well with the simulation, with the worst error being 7.7 percent at a global ring utilization of 81 percent.

Our final comparison between model and simulation is shown in Fig. 6 for three H3 configurations, again revealing very good agreement except at high global ring utilization levels.

The more important point brought out by Fig. 6, however, relates to the relationship between average packet delay performance and network configuration at different traffic levels. Consider the following: Assume a distribution of message packet destinations that is characterized by the

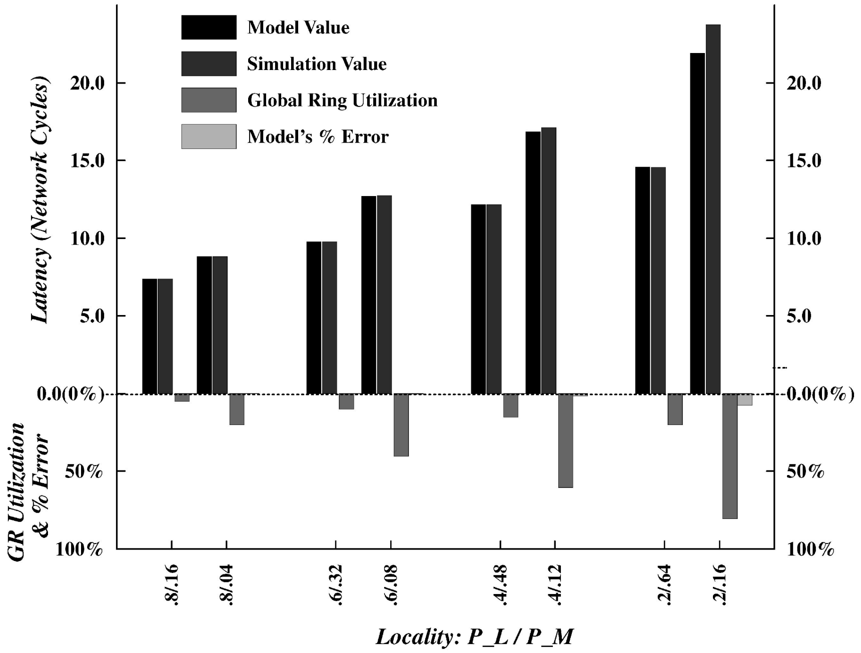


Fig. 5. Comparison between the model and simulation for an H3 system where $N = 504$, $L = 7$, $M = 6$, $G = 12$, and $\lambda = 0.005$.

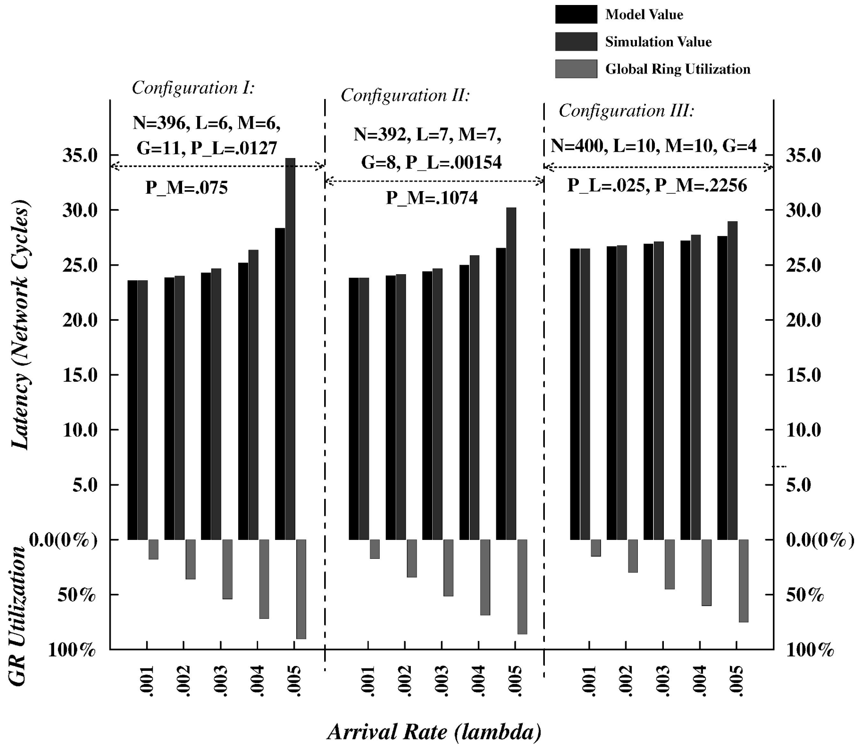


Fig. 6. Comparison between the model and simulation for three H3 configurations with a uniform distribution of message packet destinations.

application, not related to network configuration. For example, in the uniform distribution, all processor cluster nodes are equally likely as the destination of a message packet. This presents the most demanding case for any multiprocessor network. There is no locality that can be exploited.

Fig. 6 shows such a case. System size 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 packet destination distribution.

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

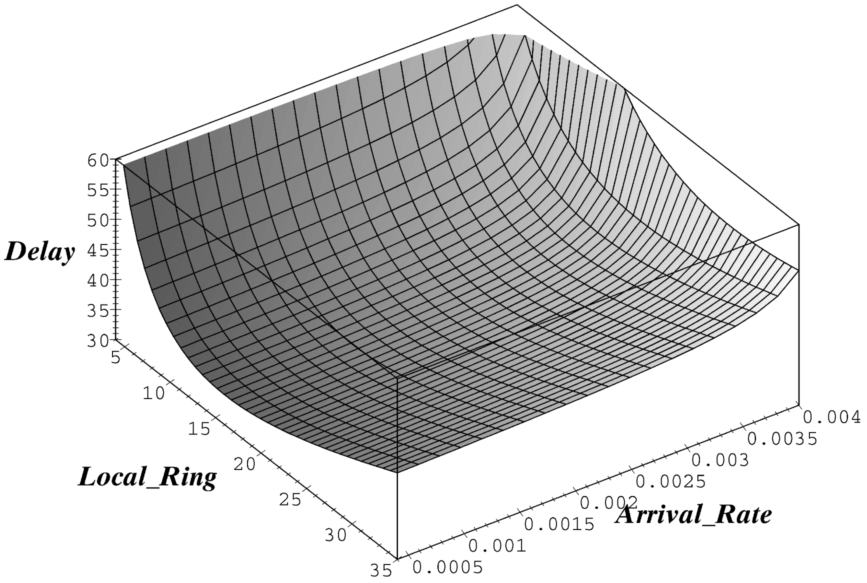


Fig. 7. 3D plot for H2 delay with $N = 500$ and uniform message packet destination distributions.

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 [14], an independent detailed simulation study of H3 systems showed that good configurations for the heavy uniform traffic case 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 and Bottlenecks

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.

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 (3.17) and (3.18) can be used to find optimal values of L and M. All 3D 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 packet destination case here. This is simply the case in which all other $N - 1$ nodes are equally likely as 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.

Fig. 7 shows a 3D 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 ($\lambda = 0.0005$), L is optimal near 16, shifting to larger values as λ increases, with L being optimal near 28 for $\lambda = 0.004$.

In Figs. 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}.

A general rule-of-thumb can be concluded from the results of Sections 4 and 5 for the uniform traffic case: As the traffic intensity rate λ moves from light to heavy, the proportional ring sizes for optimal network configurations shift from 1:2 in H2 and 1:1:2 in H3 to 2:1 in H2 and 2:2:1 in H3.

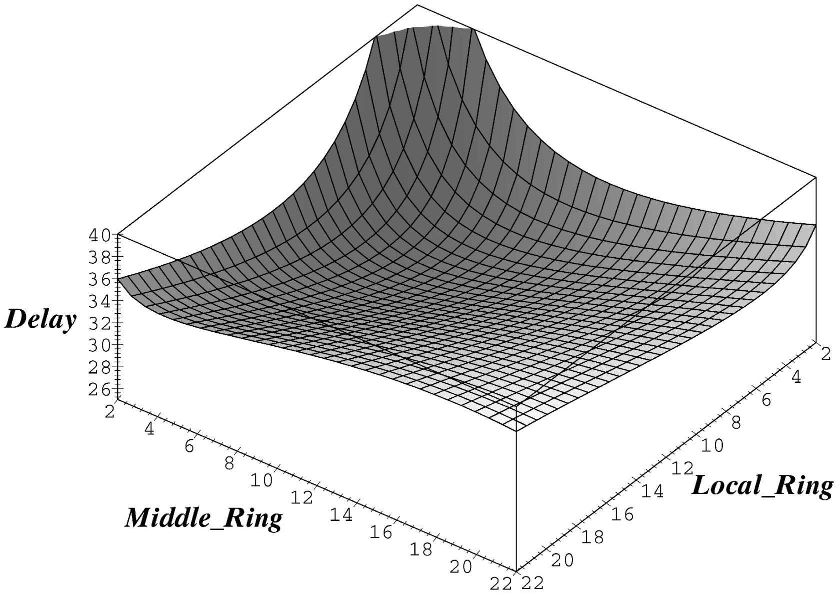


Fig. 8. 3D plot for H3 delay with $N = 500$, $\lambda = 0.002$, and uniform message packet destination distributions.

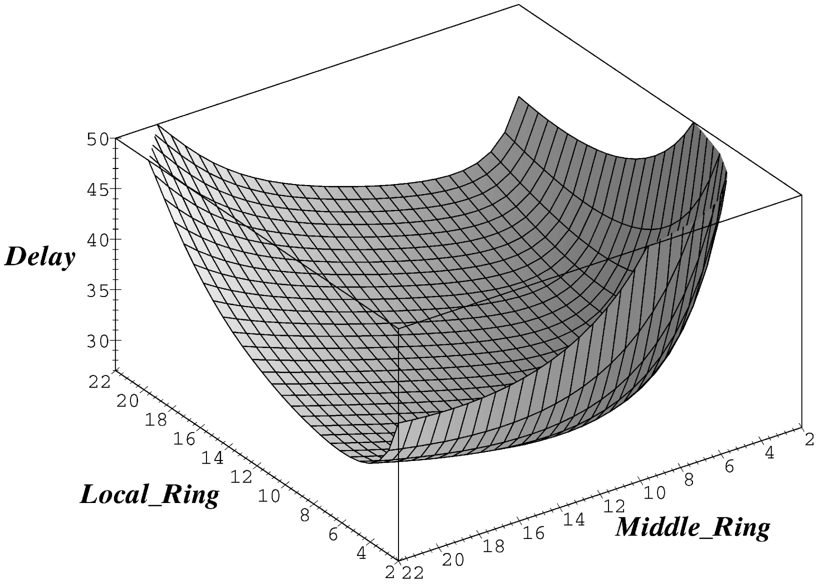


Fig. 9. 3D plot for H3 delay with $N = 500$, $\lambda = 0.004$, and uniform message packet destination distributions.

5.1 Global Ring Bottleneck

It is clear from numerical examples derived from either the analytical models or simulations that the global ring saturates first when there is low locality in the message traffic. For a uniform message packet distribution, Fig. 6 shows that choosing a ring size configuration with the global ring relatively smaller than the local and intermediate rings leads to lower utilization of the global ring and lower overall average packet delay. Put another way, the optimal configuration allows a higher traffic rate (more throughput) before saturation occurs.

Since a relatively small global ring represents a proportionally very small component of the hardware implementation cost of a full network, it is feasible to consider increasing its bandwidth. In [14], the authors propose doubling the global ring bandwidth. This can be achieved in one of two ways—doubling the physical width of the links or doubling the clock rate and adding a (pipeline) buffer in each link—as discussed in [14].

It is very easy to change the analytic queuing model to account for a double-bandwidth global ring. We will not give details here, but will state some numerical results.

For the uniform traffic case, an $N = 90$ H3 system with the configuration $(L, M, G) = (6, 5, 3)$ and a double-bandwidth global ring has a packet delay versus λ performance that is very close to that of an $N = 75$ system with the configuration $(L, M, G) = (5, 5, 3)$ and a regular global ring. As another example, an $N = 108$ H3 system with an $(L, M, G) = (6, 6, 3)$ configuration and a double-bandwidth global ring has a performance comparable to a regular $N = 90$ system with an $(L, M, G) = (6, 5, 3)$ configuration.

One way to view these results is that doubling the global ring bandwidth in these two cases allows an increase of 20 percent in system size, N, and total packet throughput,

$N\lambda$, for the same packet delay versus λ performance, as λ is varied over a wide operational range.

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 queuing 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 Section 5. 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 [9], we have shown in [10] 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 studying how to incorporate more general distribution specifications into the queuing models, enabling wider use of the models in network design and optimization.

Acknowledgments

The research of V. Carl Hamacher was supported by an NSERC (Canada) Research Grant; this research was done while he was a visiting professor in the Computer Science Department at the University of California at Riverside during 1996-1997. Facilities support there is also gratefully acknowledged. The research of Hong Jiang was supported in part by a Nebraska Research Initiative (NRI) Research Grant.

References

[1] L.N. Bhuyan, D. Ghosal, and Q. Yang, "Approximate Analysis of Single and Multiple Ring Networks," *IEEE Trans. Computers,* vol. 38, no. 7, pp. 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,* vol. 24, no. 2, pp. 33-46, Feb. 1991.

[4] T.H. Dunigan, "Multi-Ring Performance of the Kendall Square Multiprocessor," Oak Ridge Nat'l Laboratory Report TM-12331, Oct. 1994.

[5] K. Farkas, Z. Vranesic, and M. Stumm, "Scalable Cache Consistency for Hierarchically Structured Multiprocessors," *J. Supercomputing,* vol. 8, pp. 345-369, June 1995.

[6] V.C. Hamacher and H. Jiang, "Comparison of Mesh and Hierarchical Networks for Multiprocessors," *Proc. 1994 Int'l Conf. Parallel Processing,* vol. I, pp. 67-71, Aug. 1994.

[7] V.C. Hamacher and H. Jiang, "Performance and Configuration of Hierarchical Ring Networks for Multiprocessors," *Proc. 1997 Int'l Conf. Parallel Processing,* pp. 257-265, Aug. 1997.

[8] J.L. Hennessy and D.A. Patterson, *Computer Architecture: A Quantitative Approach,* second ed., San Francisco: Morgan Kaufmann, 1996.

[9] M. Holliday and M. Stumm, "Performance Evaluation of Hierarchical Ring-Based Shared Memory Multiprocessors," *IEEE Trans. Computers,* vol. 43, no. 1, pp. 52-67, Jan. 1994.

[10] H. Jiang, C. Lam, and V.C. Hamacher, "On Some Architectural Issues of Optical Hierarchical Ring Networks for Shared-Memory Multiprocessors," *Proc. Second Int'l Conf. Massively Parallel Processing Using Optical Interconnections (MPPOI),* pp. 345-353, Oct. 1995

[11] P.J.B. King and I. Mitrani, "Modeling a Slotted Ring Local Area Network," *IEEE Trans. Computers,* vol. 36, no. 5, pp. 554-561, May 1987.

[12] C. Lam, H. Jiang, and V.C. Hamacher, "Design and Analysis of Hierarchical Ring Networks for Multiprocessors," *Proc. 1995 Int'l Conf. Parallel Processing,* vol. I, pp. 46-50, Aug. 1995.

[13] W.M. Loucks, V.C. Hamacher, B.R. Preiss, and L. Wong, "Short-Packet Transfer Performance on Local Area Ring Networks," *IEEE Trans. Computers,* vol. 34, no. 11, pp. 1006-1014, Nov. 1985.

[14] G. Ravindran and M. Stumm, "A Performance Comparison of Hierarchical Ring- and Mesh-Connected Multiprocessor Networks," *Proc. 1997 Int'l Symp. High Performance Computer Architecture,* pp. 58-69, Feb. 1997.

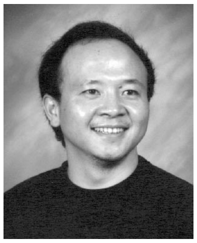
[15] Z. Vranesic, S. Brown, and M. Stumm, "The NUMAchine Multiprocessor," technical report, Dept. of Electrical and Computer Eng., Univ. of Toronto, June 1995.

[16] Z.G. Vranesic, M. Stumm, D.M. Lewis, and R. White, "Hector: A Hierarchically Structured Shared-Memory Multiprocessor," *Computer,* vol. 24, no. 1, pp. 72-79, Jan. 1991.

[17] A.W. Wilson, "Hierarchical Cache/Bus Architecture for Shared Memory Multiprocessors," *Proc. 14th Ann. Int'l Symp. Computer Architecture,* pp. 244-252, 1987.

[18] X. Zhang and Y. Yan, "Comparative Modeling and Evaluation of CC-NUMA and COMA on Hierarchical Ring Architectures," *IEEE Trans. Parallel and Distributed Systems,* vol. 6, no. 12, pp. 1316-1331, Dec. 1995.

V. Carl Hamacher received the BASc degree in engineering physics in 1963 from the University of Waterloo, Waterloo, Ontario, Canada, the MSc degree in electrical engineering in 1965 from Queen's University, Kingston, Ontario, Canada, and the PhD degree in electrical engineering in 1968 from Syracuse University, Syracuse, New York. From 1968 to the end of 1990, he was at the University of Toronto, where he was a professor in the Departments of Electrical Engineering and Computer Science. At Toronto, he held the position of director of the Computer Systems Research Institute from 1984-1988 and was chairman of the Division of Engineering Science from 1988-1990. Since January 1991, he has been a professor of electrical and computer engineering at Queen's University, where he was also Dean of the Faculty of Applied Science from January 1991 to June 1996. His current research interests are multiprocessors, multicomputers, and their interconnection networks, including local area networks. He is a coauthor of the text *Computer Organization* (McGraw-Hill, fourth edition, 1996). During 1978-1979, he was a visiting scientist at the IBM Research Laboratory in San Jose, California. For part of 1986, he was a research visitor at the Laboratory for Circuits and Systems associated with the University of Grenoble, France. During 1996-1997, he was a visiting professor in the Computer Science Department at the University of California at Riverside and in the LIP6 Laboratory of the University of Paris VI. Dr. Hamacher is a senior member of the IEEE, and a member of the ACM, a member of Sigma Xi, and a professional engineer in the Province of Ontario, Canada.



Hong Jiang received the BSc degree in computer engineering in 1982 from Huazhong University of Science and Technology, Wuhan, China, the MASc degree in computer engineering in 1987 from the University of Toronto, Canada, and the PhD degree in computer science in 1991 from Texas A&M University, College Station. Since August 1991, he has been at the University of Nebraska-Lincoln, where he is an associate professor in the Department of Computer Science and Engineering. His current research interests are computer architecture, parallel/distributed computing, performance evaluation, supercomputing, networking, computer storage systems, and computational engineering. He published numerous papers in major journals and international conferences in these areas and his research has been supported by the US Department of Defense, US National Science Foundation, and the State of Nebraska. Dr. Jiang is a member of the IEEE, the IEEE Computer Society, the ACM, and the ACM SIGARCH and ACM SIGCOMM.