On Some Architectural Issues of Optical Hierarchical Ring Networks for Shared-Memory Multiprocessors

Hong Jiang

Dept. of Computer Sci. & Engr.
University of Nebraska-Lincoln
Lincoln, Nebraska 68588-0115

Clement Lam and V. Carl Hamacher *

Dept. of Elec. & Computer Engr.
Queen's University
Kingston, Ontario, Canada K7L 3N6

Abstract

Optical hierarchical ring networks with 2 and 3 levels for multiprocessors are studied through simple analytical modeling and extensive simulations. The performance of the four possible deflection routing schemes to resolve contentions is simulated and found to be relatively the same. Comparison of deflection routing and buffering, under the assumption that each slot contains one bit along the temporal dimension, shows that the transaction delays in systems using deflection routing increase faster than in systems with buffering with an increase in traffic intensity. However, the performance gain by reconfiguring from a 2-level deflection system to a 3-level system is significant, and the gain can outperform buffering in a 2-level system. It is postulated, nevertheless, that deflection routing should outperform the buffering scheme when each slot contains more bits along the temporal dimension, because the o-e and e-o cost of the latter is proportional to the number of bits whereas it is constant for the former. Non-contention optimal configurations are found by minimizing the maximum transaction delay and the average transaction delay. However, when contentions are considered, those configurations that minimize the average non-contention delay perform worse than those which minimize the maximum non-contention delay. The poor performance is the result of quick saturation at the global ring. However, configurations that result from minimizing the maximum or average non-contention delay may be far from the true optimal configuration specific to a particular workload, especially when the traffic load is high, and traffic is localized.

1 Introduction

In the research community, shared-memory multiprocessors, as well as the closely-related multicomputer structures, have provided a number of challenging problems and new areas of investigation. The structure of the interconnection network (IN) that allows processors to access remote memory modules is of critical importance in achieving high performance.

The use of optic fiber and, subsequently, of optical networking technology in LANs/WANs has made it possible to design and construct, albeit in a laboratory environment, optical INs for multiprocessors [6, 13].

In this paper we study the hierarchical ring structure, an IN form extended naturally from the single-ring based shared memory multiprocessor systems for improved system scalability. Our motivation in studying this type of IN structure stems from the observation that the hierarchical structure takes full advantage of the spatial locality of communication often exhibited in multiprocessors. Spatial locality, a key to size scalability according to Bell [1], measures the likelihood of a processor communicating with a physically near neighbor. This paper focuses on a class of optical hierarchical ring structures as INs for shared-memory multiprocessors. Examples of multiprocessor systems using such network structure as INs that operate electronically are The GigaMax [15], Paradigm [4], KSR-1 [5], and Hector [14]. More specifically, this paper investigates important architectural issues pertaining to hierarchical ring networks in general and optical ring networks in particular and studies the impact of certain design parameters on system performance under some practical loading conditions. The work presented here is an expansion on work presented in [8], with an explicit emphasis on the optical aspects.

The paper is organized as follows. Section 2 describes the hierarchical interconnection network model, including enough structural and operational detail for performance evaluation purposes. Section 3 introduces flow control mechanisms that resolve conflicts arising during packet transmissions. Section 4 presents optimal structure design subject to some constraints such as traffic conditions and the depth of the

*Supported by an NSERC (Canada) Research Grant

0-8186-7101-7/95 $4.00 © 1995 IEEE

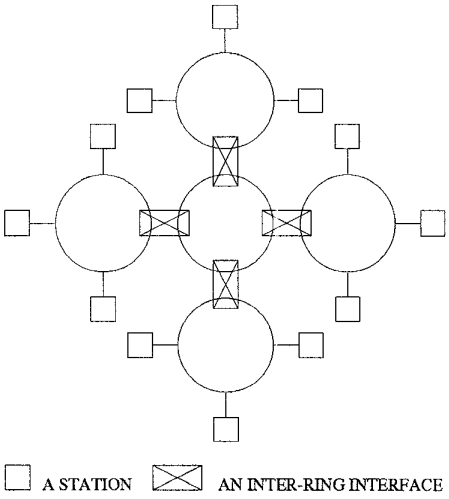

Figure 1: 2-level Hierarchical Ring Network with 12 Stations.

hierarchy. Two major flow control schemes, namely, the deflection routing approach (including all four possible schemes) and the buffering approach, are investigated for their effectiveness, with respect to features unique to fiber optic networks, through extensive simulations in Section 4. Also presented in this section are other simulation results comparing various design alternatives for the hierarchical ring structure. Finally, some concluding remarks on future work are made in Section 5.

2 Hierarchical Ring Network Architecture

A general m-level hierarchical ring network logically resembles a tree in which every node represents a ring. The root of the tree, level m, is the *global ring*; and the nodes at level 1, the lowest level, to which the leaves are attached are *local rings*. These structural features put hierarchical ring networks into a class of Non-Uniform Memory Access (NUMA) architectures in which memory access delay depends on the memory location. In a shared memory multiprocessor system, the leaves of the tree are *processor clusters* or *stations* and they contain shared memory modules as well as processors. Figure 1 shows a 2-level hierarchical ring network with 12 processor clusters or stations.

Each station is a collection of processors (possibly one) and memory modules connected by a bus, called a *cluster bus*. All components within a station (or cluster) operate electronically while the hierarchical rings that connect the stations operate optically. Two types of interfaces which can be realized by simple logic are needed in the system. A *station interface*, which is basically an e-o-e transceiver, connects a processor cluster to a local ring. The rings operate in a similar manner as Qiao and Melhem's optical bus [11], except that each station taps (reads and writes) on the same spot of the ribbon of fibers. Messages are transmitted in slots, or packets, where each slot is of length τ, called a *slot cycle*, the time it takes an optical pulse to travel between two neighboring stations assuming that stations are equally spaced. Thus, if there are w fibers in the ribbon and l optical pulses can be accomodated in τ and enough time to process routing information, then each slot (packet) is of $s = w \times l$ bits. We assume that, throughout the paper, $l \geq 1$ and $w \geq 2 \log N$ where N is the total number of stations. That is, when $l = 1$, w should be at least enough to contain routing information (source and destination addresses).

Rings at different levels are interconnected by *inter-ring interfaces*, which are essentially 2×2 crossbar switches. Packet transfer in the network is synchronized at the slot (τ) boundaries. Each ring in the network is unidirectional and is divided into fixed-size slots or segments. The packets can be realized by associating a set of latches with each station interface and inter-ring interface on every ring. Therefore, within a slot cycle, a packet can be transferred between two adjacent station interfaces on a local ring or two adjacent inter-ring interfaces on a higher level ring; between two adjacent inter-ring interfaces on different rings. The type of packet transfer within a station is not considered in this paper since only traffic which involves the network is of interest.

The system provides a single address space in which all processors can access all memory locations transparently for memory read/write operations. Besides data, each packet contains the source and destination addresses for routing purposes. When a read or write request is introduced into the network from a station, the request packet will move around the rings until it reaches the destination station. At the destination station, if the target memory is free, the request is accepted and a positive acknowledgement is sent back in the same slot cycle if the request is a write. A write memory *transaction* is assumed to be completed when the acknowledgement packet is finally removed at the source station. For a read request that is accepted, the request packet is removed at the destination and a *response* packet will be sent back at a later time with the required data. A read transaction is completed when the response packet is removed by the source station. If the memory is not free to accept the read request, a negative acknowledgement is sent back to the source

immediately. The request will be resent by the source station at a later time.

3 Flow Control Strategies

Simultaneous requests for use of a slot or output channel can cause conflicts. For example, when a new packet from a station requests a transfer from the station into a slot on the local ring, a conflict occurs if another packet on the local ring is to be transferred to the same slot in the next clock cycle. Another example of conflict is one that occurs at inter-ring interface when two packets on different rings request the same output channel. However, if a station is receiving a packet while there is a pending transfer of a new packet to the local ring, both transfers can complete successfully in the same slot cycle.

To resolve these conflicts, some flow control strategies must be employed. Here we describe to two major schemes relevant to hierarchical ring networks, namely, the buffering scheme and the deflection routing scheme. In Section 5, the effectiveness and efficiency of these two schemes will be studied through extensive simulations.

3.1 Deflection Routing Scheme

This scheme resovles a conflict by granting one of the two contenting requests while deflecting the other, hence the name. More specifically, when two packets from two different ring levels arrive at an interface at the same clock cycle and request the same output link, one packet is allocated the requested link and the other is deflected to the other output. With a 2×2 crossbar switch, four deflection routing schemes are possible to resolve the conflict, depending on which type of traffic (i.e., same ring vs. cross ring) or which of the two competing rings is given priority. They are the HRP (*Higher Ring Priority*) scheme, the LRP (*Lower Ring Priority*) scheme, the CRP (*Current Ring Priority*) scheme, and the ORP (*Other Ring Priority*) scheme, respectively. Their operations are summarized in the following table.

scheme	*operations*
HRP	grants the requested link to the packet coming from the higher level in the hierarchy
LRP	grants the requested link to the packet coming from the lower level
CRP	allocates the requested link to the packet at the same level as the requested link
ORP	gives priority to the packet which is crossing from one level to another

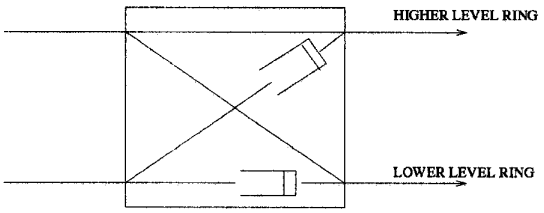

Figure 2: A 2x2 crossbar switch with output buffers arranged for the higher level ring priority (HRP) case.

In the HRP scheme, if the requested output link is at the higher level, the packet at the lower level is deflected into the same ring on which it arrived. When the requested link is at the lower level, the lower-level packet is deflected to the higher level ring. This scheme is the deflection routing version of the Hector [14] buffering scheme. Whereas in LRP, a packet at a higher level is deflected to the same ring if the requested link is at the lower level. Otherwise, the higher-level packet is deflected to the lower level ring if the output at the higher level is requested.

In the ORP scheme, every packet which tries to stay on the same ring may potentially be deflected to another ring at each inter-ring interface the packet passes. Yet, unlike the other schemes, no packet which wishes to cross to a different ring level is obstructed. Finally, in the CRP scheme, the packet wishing to stay on the ring it is currently on is granted its request.

3.2 Buffering Scheme

In this scheme, FIFO buffers are associated with the outputs of the crossbar switches at the inter-ring interfaces. While requests from the same input are serviced on a FIFO basis, those coming from different inputs may be assigned different priorities. In Hector [14], priority is given to the packet on a higher level ring while the other packet is buffered until the slot is free. The reason behind this priority scheme is to minimize the delays of the packets that are descending the hierarchy [7]. An inter-ring interface with output buffers is depicted in Figure 2. The location of the buffers in the diagram reflects the fact that priority is given to the packet on the higher level ring.

However, with the buffering scheme, a packet is dropped when it finds the buffer full. To deal with dropped packets, a time-out mechanism can be used in which a source retransmits a request when no response is received within a time-out period. However, it is pointed out in retrospect by the Hector designers [7] that the design of allowing packets to be lost is unfortunate. The mechanism not only contributes to higher latencies in memory access, but also complicates software. Most importantly, a new cache consistency mechanism [12] designed for the system does

not tolerate lost packets.

Although one solution to prevent packet loss is to increase the size of the buffers as suggested by Stumm et al [7], it leads to higher cost. While all these may be tolerable in the electronically operated networks, they present a severe drawback in optical networks where buffering can be unacceptably expensive. This is because pure optical buffers are somewhat impractical while electrical buffers coupled with optical links incur *o-e-o* (optical to electrical or electrical to optical conversion) delays. These problems motivate the design of cost-effective schemes that eliminate packet loss. The deflection strategy described earlier is one such solution.

4 Optimal Structure Design

The performance of a hierarchical ring network depends on how the hierarchy is structured or organized. For a given structure or configuration, on the other hand, a hierarchical ring network performs variably depending on the traffic (or workload) conditions applied. It is clearly nontrivial, if not impossible, to find an optimal structure for such a network, especially one which optimizes performances under most conditions. Further, for the word "optimal" to be meaningful in our context, it must be subject to some objective (or optimizing) measures, such as delay or throughput. For the sake of simplicity and clarity, this paper will focus on optimal-structure designs based on minimizing transaction delays under no contention.

To obtain useful insight into the relationship between the system configuration and its performance, we first derive some simple performance expressions with practical traffic conditions, or workload. Then optimal design parameters for the system subject to the workload is derived. Due to the lack of space and high complexity, in this paper we only consider hierarchical ring networks with 1, 2, and 3 levels, denoted H1, H2, and H3, respectively.

Communication locality is a common property shared by most application programs. Thus, delay measures for localized traffic become important. In this paper, we adopt the *clusters of locality* model of Holliday and Stumn [7], in the context of clustered communication patterns. In their model [7], an m-tuple, $P = (P_1, P_2, ..., P_m)$, is used to describe the behavior of locality in an m-level clustered application, $S = (S_1, S_2, ..., S_m)$ ($\sum_{i=1}^{m} S_i = N$, the number of processors in the system). That is, P_1 is the probability that a processor communicates with other processors of the same cluster 1 (of size S_1), P_2 the probability that a processor communicates with processors in cluster 2 given that it does not communicate within its own cluster and, in general, P_i the probability that a processor communicates with processors in cluster i given that it does not communicate with any processor in clusters 1 through $i-1$.

This model of locality characterizes the behaviors of many parallel applications where a process (or task) communicates mostly (or only) with its nearest neighbors or neighbors within a certain radius [2, 10]. Further, this characterization is independent of the architecture of the underlying multiprocessor on which the application may run. In other words, the communication locality of an application, as opposed to physical locality of a parallel machine, is captured in this model. This allows us to compare systems of different architectures under the same loading conditions. In the following analysis, we assume 3-level clustered traffic, namely, $S = (1, x, N-x-1)$ and $P = (P_1, P_2, 1.0)$, where N is the number of processes (or tasks), $x \in \{2, 3, ..., N-2\}$. That is, each process communicates with itself (i.e., cluster 1) with probability P_1 and communicates with processes of its local cluster (cluster 2) of size x with probability P_2 given that it does not communicate with itself.

However, some simple mapping of the communication locality model, which is network topology independent, onto a given network topology is needed before the analysis of average delay can be done. For simplicity, clusters are evenly mapped with respect to a given processor such that it is the geometric center of its corresponding cluster. For hierarchical rings the locality is defined on a one-dimensional grid (or linear array) [7].Thus, processors are numbered from left to right in a logically formed tree from the ring hierarchy and the cluster set 2 for processor i would consist of processors $i-2$, $i-1$, $i+1$, and $i+2$ (modulo N). Clearly, the mapping is also direct.The analysis is carried out for each of the networks under study. Let D_t denote the average delay under localized traffic for network topology t, where $t \in \{H1, H2, H3\}$. In the following analysis, N is the number of stations in a system and l (or l_1 and l_2) is the number of nodes on the global ring (or level-1 and level-2 ring).

H1: Since every processor connects to the same ring, message delay is independent of traffic locality. Thus,

$$D_{H1} = n^2$$

H2: A message can experience two distinctive delays, N/l when travelling in local ring and $2N/l + l$ when travelling globally. Further, depending on the logical number of the source processor and the value of x relative to N/l, a message's target processor in a local cluster may or may not reside inside the same local ring. Conditioning on a message not being homebound, we have,

$D_{H2} =$

$$\begin{cases} P_2[\frac{(x+2)(N+l^2)}{4N} + \frac{N}{l}] + \frac{(1-P_2)}{(N-x-1)}\{\frac{N(2N-x-1)}{l} - (\frac{N}{l})^2 \\ +N(l-1) - \frac{x(x+2)(N+l^2)}{4N}\} \ldots\ldots\ldots \text{ for } x \le \frac{N}{l} - 1 \\ \frac{P_2}{x}[(x+1)l + \frac{N(2x+1)}{l} - \frac{N^2}{l^2} - N] \\ +(1-P_2)(\frac{2N}{l} + l) \ldots\ldots\ldots\ldots\ldots \text{ for } x \ge \frac{2N}{l} - 1 \\ P_2[\frac{N(1+x)}{lx} + \frac{(x-2)(N+l^2)}{4N} + \frac{l}{x}] + (1-P_2)\{\frac{N}{l} \\ +\frac{N+l^2}{N-x-1}[\frac{N(l-1)}{l^2} - \frac{x(x+2}{4N}]\} \ldots\ldots\ldots\ldots \text{ otherwise} \end{cases}$$

In the above analysis, we have taken into consideration the edge effects which occur when a processor has part of its cluster set located out side its home local ring. That is, even for messages destined for nodes of the same cluster, the latency may involve inter-ring (or global ring) communication. This edge effect diminishes as the value x, in comparison with N/l, decreases. In more general situation, such edge effects can have significant impact on performance.

H3: In a three-level network, a message may experience one of three delays, $\frac{N}{l_1 l_2}$, $\frac{2N}{l_1 l_2} + l_1$, or $\frac{2N}{l_1 l_2} + 2l_1 + l_2$, depending on the destination, similar to the case of H2.

To derive expressions that consider edge effects as in the case of H2, we must now take into account the added level of physical locality, namely, the level-1 and level-2 local rings. This results in a considerably more complicated situation where the communication locality of a station can be completely contained in, be partially overlapped with, or completely contain the level-1 local ring and the level-2 local ring respectively. Nevertheless, the derivation for H3 is similar to that for H2 in principles. The following expressions give the average delays for a H3 network where $x \le \frac{N}{2l_2} - 1$, that is, the communication locality is completely contained in the level-2 local ring for any station. This situation emcompasses arguably the communication patterns of most practical applications. Expressions for other values of x relative to local ring size can be derived similarly.

$D_{H3} =$

$$\begin{cases} P_2[\frac{(x+2)(N+l_1^2 l_2)}{4N} + \frac{N}{l_1 l_2}] + \frac{1-P_2}{N-x-1}[\frac{N(2N-x-1)}{l_1 l_2} \\ -\frac{N^2}{l_1^2 l_2^2} - \frac{N(l_1+1)}{l_2} + N(2l_1 + l_2 - 1) \\ -\frac{(x^2+2x)(N+l_1^2 l_2)}{4N}] \ldots\ldots\ldots\ldots \text{ for } x \le \frac{N}{l_1 l_2} - 1 \\ P_2[\frac{xN+N+l_1^2 l_2}{l_1 l_2 x} + \frac{(x-2)(N+l_1^2 l_2)}{4N}] + \frac{1-P_2}{N-x-1}[N(2l_1 \\ +l_2 - 1) - \frac{N^2}{l_1^2 l_2^2} + \frac{N(2N-x-1)}{l_1 l_2} - \frac{N(l_1+1)}{l_2} - \frac{x l_1^2 l_2)}{2N} \\ -\frac{x(x+2+2l_1)}{4}] \ldots\ldots\ldots \text{ for } \frac{N}{l_1 l_2} \le x < \frac{2N}{l_1 l_2} - 1 \\ \frac{P_2}{x}[\frac{N(2x+1)}{l_1 l_2} - \frac{N^2}{l_1^2 l_2^2} - \frac{N}{l_2} + l_1(x+1)] + \frac{1-P_2}{N-x-1} \\ (\frac{2N(N-x-1)}{l_1 l_2} - \frac{N l_1}{l_2} + N(2l_1 + l_2 - 1) \\ -l_1(x+1)) \ldots\ldots\ldots \text{ for } \frac{2N}{l_1 l_2} - 1 \le x \le \frac{N}{2l_2} - 1 \end{cases}$$

With the estimates of the average delay in localized traffic for a given network configuration described above, the next logical question to ask is what would be the best configuration for the hierarchical networks that minimizes the average delay, given a localized traffic pattern. The answer to the question may be found by deriving optimal values for l in H2 and l_1 and l_2 in H3 that minimize D_{H2} and D_{H3}, respectively, while keeping values P_1, P_2, and x constant. Unfortunately, however, the optimal values for H2 and H3 don't seem to exist. This is provably true for H2 when the edge effects are ignored. To obtain useful insight into the relationship among various parameters, with regard to best performance, we plotted several 3-D surfaces that show the average delays of H2 and H3, respectively, as functions of two variables (e.g., x and l for H2, and l_1 and l_2 for H3) while fixing other parameters. The Maple-V software [3] was used for plotting the surfaces and computing numerical solutions reported in this paper.

In Figure 3, the average delays under localized traffic for the two-level hierarchical rings are plotted against x, the communication locality size, and l, the global ring size, using the expressions for D_{H2} developed in the previous subsection. The three surfaces shown in the figure correspond to three different values of P_2 respectively, with $P_2 = 0.2$ being on the top and $P_2 = 0.8$ being on the bottom. With this visualization of the expression and the help of a widely available numerical package such as Maple-V, one can easily find the optimal value for l, given N, P_2 and x. For example, when $P_2 = 0.8$ and $N = 400$, $l = 34$ for $x = 4$, $l = 26$ for $x = 20$, and $l = 23$ for $x = 40$.

Similarly, Figure 4 shows the situation for the three-

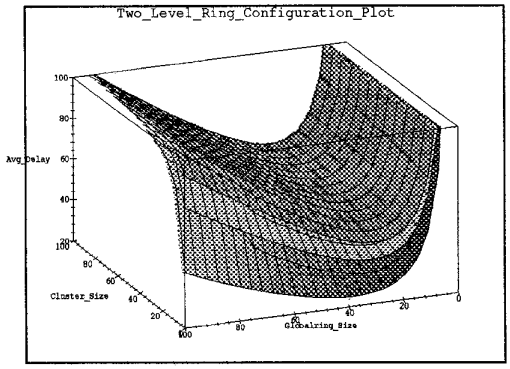

Figure 3: 3-D Plots for Two-Level Ring for $N = 400$, $P_2 = \{0.2, 0.4, 0.8\}$

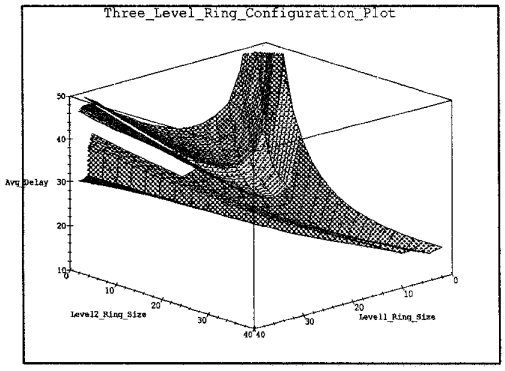

Figure 4: 3-D Plots for Three-Level Ring for $N = 400$, $P_2 = 0.8$ and $x = \{4, 20, 40\}$

level hierarchical rings. In this case, however, the delays are plotted against l_1, the size of the first level global ring, and l_2, the size of the second level global ring, while P_2 and x are kept constant. Again, the expressions for D_{H3} developed in the previous subsection were used for the plotting. The three surfaces correspond to three different values of x ($x = \{4, 20, 40\}$), with $x = 40$ on the top and $x = 4$ at the bottom. Examples of optimal configurations, denoted in a two-tuple (l_1,l_2), are: (6,20), (7,10), and (11,5), for $N = 400$ and $P_2 = 0.8$.

Next, we compare the performance of the three-level hierarchical rings with that of the two-level hierarchical rings in Figures 5, 6 and 7. The narrow stripes in these figures correspond to the delays of the two-level rings at respective communication localities. The difference among these three figures lies in their corresponding value of x, the size of communication locality. In all the cases, the three-level rings consistantly outperform their two-level counterparts by a big margin. This performance advantage of the former, however, comes at its relatively high cost.

5 Comparisons and Discussions

In this section an event-driven simulator is used to study 2-level and 3-level hierarchical ring networks. All the simulation results presented here have a 95% confidence interval. The results are organized into three parts. The first part focuses on the comparative performance of the different flow control strategies and the tradeoffs between the deflection routing scheme and the buffering scheme. The other two parts concentrate on the performance of different network topologies. First, results of 2-level hierarchies are compared to 3-level ones to show the performance gain in 3-level structures. Then, simulation results which relate to the two design methods for an optimal topology discussed in the previous section are presented.

The performances of the four possible deflection routing schemes and the buffering scheme are studied via simulation. The experiments assume the optimal topology designed by minimizing the maximum transaction delay in the system [8]. Due to time constraint of our study, it is assumed that each slot cycle contains only one bit along the temporal dimension (i.e., one optical pulse is accomodated in τ, $l = 1$) while a large number of bits are associated along the spatial dimension (i.e., the ribbon contains many fibers). This assumption is clearly in favor of the buffering scheme because the time it takes to buffer is the same as the time it takes to process the routing information in the deflection scheme, hiding the severe drawback of the former. Therefore, it is postulated that, when $l \gg 1$, the deflection schemes should outperform the buffering scheme. This is because all the routing information is contained in the first bit of the slot (along the spatial dimension or parallel) and thus the former involves only a time of one bit in processing routing information whereas the latter spends a time proportional to l in buffering. Performance study is under way for comparing the two schemes for $l \gg 1$.

The physical locality-of-communication model, instead of the computational locality model, is used because the model has direct control of the percentages of transactions that are local, second level, or global. That is, in the case of 2-level ring network, a locality

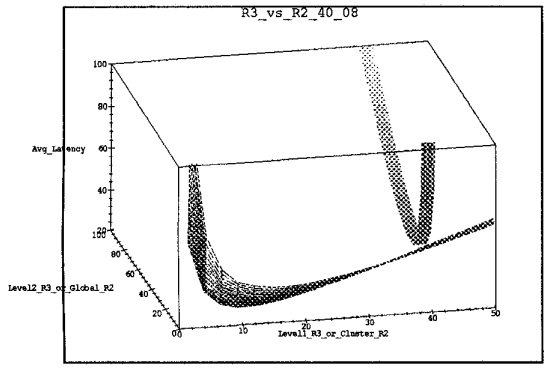

Figure 5: 3-D Plots Comparing Three-Level Ring with Two-Level Ring for $N = 400$, $P_2 = 0.8$ and $x = 40$

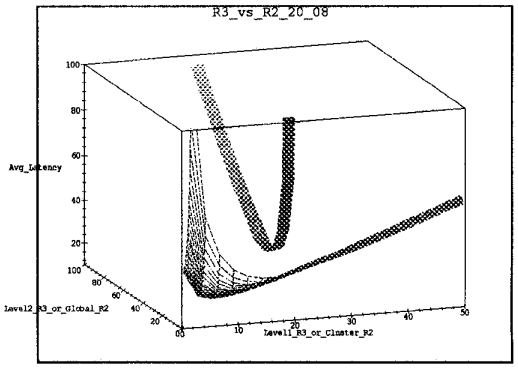

Figure 6: 3-D Plots Comparing Three-Level Ring with Two-Level Ring for $N = 400$, $P_2 = 0.8$ and $x = 20$

of, say, 80% implies that 80% of all transactions take place inside local rings. Similarly, in a 3-level ring network, a locality of $P = (0.9, 0.5, 1.0)$ means that 90% of the transactions are local, 50% of the rest are second level transactions, and the remainder are global with probability 1.0.

Results for 2-level and 3-level systems show that the differences among deflection routing schemes are insignificant when communications are localized or when the rate of requests is low. LRP and ORP perform slightly better than HRP and CRP when the system becomes more heavily utilized. The performance between LRP and ORP or the performance between HRP and CRP cannot be differentiated. When the performance of the buffering scheme is compared to the deflection schemes, the simulation results show average transaction delays of deflection routing schemes are comparable to the buffering scheme only when the traffic is highly localized or when the traffic intensity is low. The performance of deflection routing degrades at a faster rate than buffering with a decrease in locality of communication or an increase in traffic intensity. Figures 8 and 9 illustrate some of the simulation results. More detailed results can be found in [8] and [9].

Next we compare the performances of 2-level and 3-level systems. In order to compare hierarchies with different numbers of levels, the computational locality-of-communication model is used because it is independent of the network configuration. The experiments also assume the optimal configuration designed by minimizing the maximum transaction delay.

The 2-level topology with $N = 450$ and $L = (15, 30)$, and the 3-level topology with $N = 468$ and $L = (6, 6, 13)$, are used in the first set of experiments. Communication locality with $x = 4$ and P_2 varied as 0.9, 0.6, and 0.3 is used. A request rate of $\lambda = 0.0025$ is chosen so that both configurations are below saturation. The topologies are chosen so that the number of stations in the 2-level and 3-level hierarchy are similar. Moreover, because results from the last subsection show no significant difference among the deflection schemes, only HRP and buffering scheme results are presented. Figure 10 shows the plot of the average transaction delays versus the fraction of transactions to a station outside the locality set.

The second set of results studies a larger system with topologies $L = (22, 46)$ and $L = (8, 8, 16)$. The communication pattern is specified by $x = 4$, $P_2 = 0.9$, and $\lambda = 0.0012$. Similar to before, the locality of communication is varied. The simulation results are shown in Figure 11, where the average transaction delays are plotted against the fraction of transactions not within a locality set.

Simulation results show that the performance gains in 3-level structures are quite significant. In fact, in cases where the traffic exhibits localization, the performance of deflection routing in a 3-level hierarchy is better than in a 2-level hierarchy with buffers. Moreover, the degradation in performance from the buffering to deflection routing is less significant in a 3-level system than a 2-level system. In fact, Figure 10 and Figure 11 show that the average transaction delay of a 3-level system with deflection routing is comparable to the average transaction delay of a 2-level system with buffering. Therefore, an alternative to replacing deflection routing with buffering in a 2-level system in

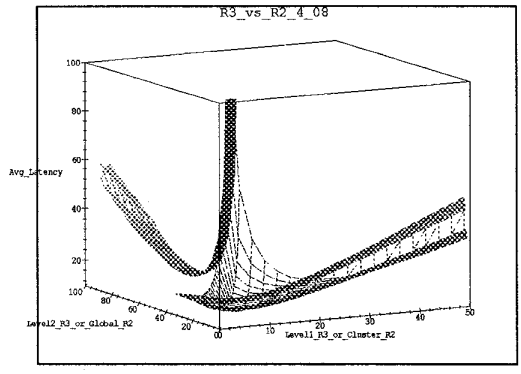

Figure 7: 3-D Plots Comparing Three-Level Ring with Two-Level Ring for $N = 400$, $P_2 = 0.8$ and $x = 4$

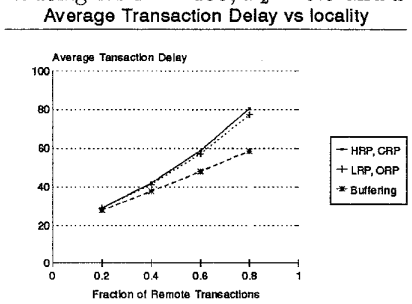

Figure 8: Avg. Transaction Delay in a 2-level System with $N = 512$, $L = (16, 32)$, a nd $\lambda = 0.001$.

order to improve performance is to reconfigure from 2 levels to 3 levels.From a hardware complexity standpoint, such a reconfiguration might be attractive, because buffer management at high clock rates may be more difficult to implement than deflection routing.

6 Concluding Remarks and Future Work

This paper has studied some important architectural issues, namely, optimal structures and flow control strategies, in optical hierarchical ring networks for multiprocessors. While finding an optimal topology is critical for guaranteed performance of a hierarchical ring network, it is an extremely difficult task because it is affected by many factors that are either unknown or dynamic. Thus, in this paper, we are interested in obtaining useful insights into such a complex issue. As a result, optimal configurations that minimize transaction latency and ignore contentions, but with practical loading conditions, are derived, and their perfor-

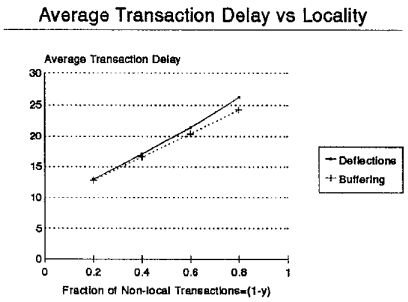

Figure 9: Avg. Transaction Delay in a 3-level System with $N = 504$, $L = (7, 6, 12)$, and $P_2 = (y, 0.8, 1.0)$.

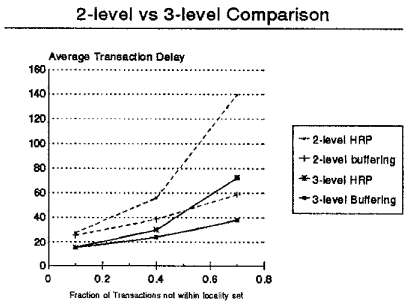

Figure 10: 2-level vs 3-level system avg. transaction delay with $N \approx 450$.

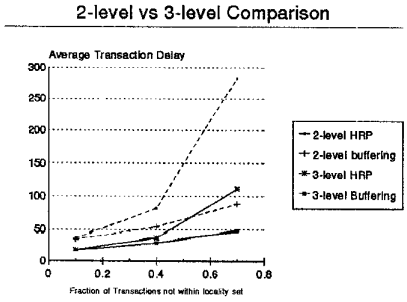

Figure 11: 2-level vs 3-level system avg. transaction delay with $N \approx 1024$.

mances are studied via simulations. The simulation study indicates that optimal network configuration is both communication-pattern and traffic-intensity dependent.

Extensive simulation studies are carried out to investigate the performances of various flow control strategies. The performance of the four possible deflection routing schemes, ones suitable for optical hierarchical networks, are basically the same using network configurations which minimize the maximum transaction delay under no contention. The delay is comparable to buffering systems under low load. The performance of deflection routing degrades significantly with either an increase in traffic intensity or a decrease in the locality of communications because of heavy contention at the global ring in 2-level systems. Although the transaction delays are less sensitive to traffic intensity in 3-level systems, the performance is still poor under high load. Under the current assumption that $l = 1$, the performance of buffering is always better than deflection routing. This, however, can be misleading because delay incurred on deflection schemes is constant (one bit of time for processing routing information), independent of l, while the delay due to buffering is proportional to l. Thus, it is postulated that as l increases deflection schemes outperform the buffering scheme. Moreover, the advantage of buffering in 3-level systems is not as significant as in 2-level systems. Since the performance of 3-level systems is significantly better than 2-level systems with deflection routing, reconfiguring from 2 levels to 3 levels is a good alternative to gain similar or better performance than by adding buffers to the 2-level system. Deflection routing is not suitable for 2-level or 3-level systems under high load or highly non-localized traffic because the global ring saturates quickly which causes a packet to be deflected a number of times before it can reach its destination. In addition, deflection routing is found to be more sensitive to configuration changes than buffering.

References

[1] Gordon Bell. Ultracomputers: A teraflop before its time. *Commun. of the ACM*, 35(8):27–47, August 1992.

[2] D. P. Bertsekas and J. N. Tsitsiklis. *Parallel and Distributed Computation–Numerical Methods.* Prentice-Hall, 1989.

[3] 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.

[4] 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.

[5] Thomas H. Dunigan. Kendall square multiprocessor: Early experiences and performance. *Kendall Square Research Tech. Note*, August 1992.

[6] E.E.E. Frietman, L. Dekker, and W. Smit. Massively parallel processing: Optical interconnects according to a system to device approach. *Proceedings of MPPOI94 (1st Intl. Workshop on Massively Parallel Processing Using Optical Interconnections)*, pages 94–105, April 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 1989.

[8] 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, August 14-19 1995.

[9] Clement Lam, Hong Jiang, and V. Carl Hamacher. Design and analysis of hierarchical ring networks for shared-memory multiprocessors. *TR-95-08, Dept. of CSE, University of Nebraska-Lincoln*, April 1995.

[10] F. Thomson Leighton. Introduction to parallel algorithms and architectures: Arrays, trees, hypercubes. *Morgan Kaufmann Publishers, San Mateo, CA*, 1992.

[11] C. Qiao and R. G. Melhem. Time-division optical communications in multiprocessor arrays. *IEEE Transactions on Computers*, 42(5):577–590, May 1993.

[12] M. Stumm, Z. Vranesic, R. White, R. Unrau, and K. Farkas. Experiences with the hector multiprocessor. *U. of Toronto Tech. Report CSRI-276*, October 1992.

[13] R. J. Vetter and H. C. Du. Distributed computing with high-speed optical networks. *IEEE Computer*, 26(2):8–18, February 1993.

[14] 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.

[15] 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.