Does Time-Division Multiplexing Close the Gap Between Memory and Optical Communication Speeds?

X. Yuan, R. Gupta and R. Melhem

The University of Pittsburgh, Pittsburgh, PA 15260

Abstract. Optical interconnection networks have the potential of transmitting data much faster than the speed at which this data can be generated at the transmitting nodes or consumed at the receiving nodes. Even when the communication protocols are simplified and most of the protocol signaling at the end nodes is done in hardware, the data cannot be processed at the end nodes at a rate faster than the rate of memory operations. We study the effect of multiplexing the network on both its throughput and latency for different ratios of memory to link transmission speeds. We found that multiplexing reduces the impact of the gap between data transmission and data generation/consumption speeds.

1 Introduction

Optical interconnection networks have the potential of offering bandwidths that are orders of magnitude larger than those of electronic interconnection networks. However, packet switching techniques, which are usually used in electronic interconnection networks are not quite suitable when optical transmission is used. The absence of appropriate photonic logic devices makes it extremely impractical to process packet routing information in the photonic domain. Moreover, conversion of this information into the electronic domain increases the latency at intermediate nodes relative to the internode propagation delay, especially in multiprocessor networks where the internode propagation delays are very small. Hence, to take full advantage of optical transmission, all–optical paths should be established between the communicating nodes before data is transmitted.

In Order to exploit the large bandwidth of all–optical networks in massively parallel systems, the speed mismatch between the components of the communication system must be resolved. An imminent speed mismatch in systems involving all–optical networks is the gap between the fast transmission of signals in the optical domain and the slow network control, which is usually performed in the electronic domain. Another gap is between the high optical data transmission rate and the relatively slow data generation/consumption rate at the source and destination nodes.

Time–Division–Multiplexing (TDM) techniques [3, 4] have been proposed to resolve the speed mismatch between the fast optical data network and slow

* This work was supported in part by NSF awards MIP-9633729 and CCR-9157371.

network control. However, to our knowledge, no previous work has considered the impact of the speed mismatch between the fast data transmission and the slow data generation/consumption, which is bounded by the electronic speed. Specifically, the bandwidth of optical links may not be fully utilized if the data generation/consumption rate at the end points of a connection cannot match the transmission speed. In this paper, we present a study of the effect of this mismatch. First, however, we briefly describe *Time Division Multiplexing* (TDM) as applied to interconnection networks.

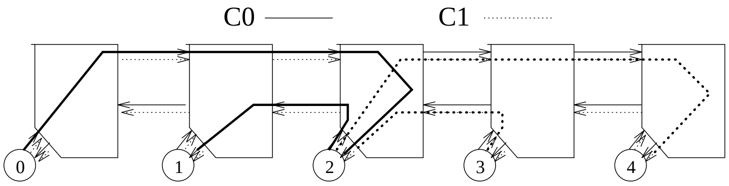
2 Time division multiplexing (TDM)

We consider switching networks in which the set of connections that can be established simultaneously may be changed by changing the state of the network. A set of connections that can be supported simultaneously will be called a *configuration*. In a TDM system, each link is multiplexed in the time domain to support multiple virtual channels. Hence, multiple connections can co-exist on a link, and thus, multiple configurations can co-exist in the network.

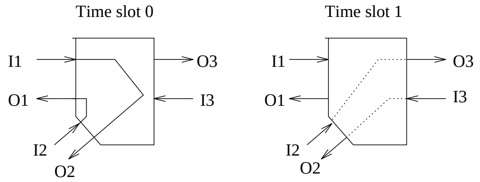
Figure 1 (a) shows an example of such a system. In this example, each processor has an input and an output connection to a 3×3 switch. Each link is multiplexed with degree 2 by dividing the time domain into 2 time slots, and using alternating time slots for supporting two channels, $c0$ and $c1$. Four connections, $(0, 2)$, $(2, 1)$, $(2, 4)$ and $(3, 2)$, are established with connections $(0, 2)$ and $(2, 1)$ using channel $c0$, and connections $(2, 4)$ and $(3, 2)$ using channel $c1$. The switches are globally synchronized at timeslot boundaries, and each switch is set to alternate between the two states that are needed to realize the connections. For example, Figure 1 (b) shows the two states that the 3×3 switch attached to processor 2 must realize for the establishment of the connections shown in Figure 1 (a). Note that each switch is an electro–optical switch ($Ti : LiNbO_3$ switch, for example [1]) which connects optical inputs to optical outputs without optical/electronic conversion. The state of the switch is controlled by an electronic signal.

There are two main advantages for optical TDM networks. First, multiplexing increases the number of connections that can co-exist in the network, thus, increasing the chance of successfully establishing a connection. The cost of control is amortized over the number of co-existing connections, thus reducing the control overhead. This effect was studied in [4], where TDM was shown to reduce the impact of the gap between data transmission and control network speeds.

The second advantage of TDM is that it can reduce the impact of the speed mismatch between the large bandwidth of optical transmission rate and the low data generation rate at each node, especially when transmitted data is to be fetched from (or stored into) memory. In other words, if data cannot be fetched from (or stored into) memory fast enough to match the optical transmission bandwidth, the extra bandwidth in the path will be wasted. In such cases, multiplexing allows the large optical bandwidth to be shared among multiple connections. Notice that under current technology, assuming that 64 bits



(a) A TDM linear array



(b) Switch states

Fig. 1. Time division multiplexing

can be fetched from (or stored into) memory in 40ns, the end nodes can have 1.6Gb/s data generation/consumption rate, while optical network with bandwidth of 250Gb/s have been demonstrated [2]. We focus on the impact of this gap on TDM systems.

3 Effect of memory speed on network performance

Let's define the *link/memory bandwidth ratio* to be the ratio between the time it takes to fetch/store a data packet from memory to the time it takes to transmit that data packet between two nodes on an already established all-optical connection. Time will be measured in terms of time slots, where a time slot is large enough to transmit a packet between two nodes on an optical connection. In a multiprocessing environment, where nodes are at most a few feet apart, it is reasonable to assume that the time for packet transmission is independent of the length of the optical connection (light signals travel at a speed of about one foot per nano-second).

We conducted extensive simulations to study the effect of the memory speed on the communication performance of TDM systems. This section presents some sample results of our study. We simulated random communications on a 16×16 torus network which uses the *conservative backward reservation protocol* [4] for the dynamic reservation of all-optical paths. To support this protocol, a shadow control network, which has the same topology as the optical data network, is needed. The control network, is an electronic network which operates in a packet switching fashion, and in which control packets are processed at each intermediate node to execute a distributed protocol for the establishment of optical data paths. The speed of the control network is characterized by the control packet processing time, which is the time required at each intermediate node to process a control packet, and the control packet propagation time, which is the time to transmit the packet to between two nodes on the control network.

We modified the simulation package described in details in [4] to account for memory speed, and we studied the effect of the link to memory speed ratio on the efficiency of the communication. In the simulation, during each time slot, a message is generated at each node with a fixed probability, r. Each message has a fixed size measured in terms of the number of data packets it contains. When a message is generated, a distributed control protocol is executed by exchanging control packets on the control network to establish a connection for the message on the multiplexed data network. When the connection is established using a given TDM channel, transmission proceeds on the data network at a rate of one data packet every K time slots, where K is the multiplexing degree of the data network.

The performance metrics used to estimate the efficiency of the communication network are the *maximum throughput*, and the *average message delay*. The maximum throughput is defined as the number of packets delivered to their destinations per time slot when the network is saturated. The message delay is the length of the period, in time slots, between the time a message is submitted to the network to the time at which the first packet in the message is received at the destination. Message delay is measured at low traffic conditions. Specifically, we will measure the delay when, r, the messages generation rate at each node, is much smaller than the rate at which the network saturates.

3.1 Effect of memory speed on throughput

Figure 2 shows the maximum throughput that the network can achieve with different multiplexing degrees and different link/memory bandwidth ratios. In this experiment, we set the message size to be 8 packets, the control packet processing time to be 1 time slot and the control packet propagation time to be 1 time slot. It is clear from Figure 2 that for any link/memory bandwidth ratio, the throughput increases when the multiplexing degree increases. This is due to the decrease in the connection establishment time.

For a given multiplexing degree, K, the data generation/consumption at the end nodes is not a bottleneck when the memory speed is the same as the network speed (link/memory bandwidth ratio equal to 1). For slower memory speeds, the network throughput does not decrease as long as data can be fetched from (or stored into) memory faster that the speed at which it can be transmitted on the network. That is, as long as one packet is ready for transmission every K time slots. However, for each multiplexing degree, there exists a link/memory bandwidth ratio, called *the throughput critical ratio*, CR_t, after which data are not available when it can be transmitted. In other words, when the link/memory bandwidth ratio is larger than CR_t, the memory speed does not affect the network throughput and when the link/memory bandwidth ratio is smaller than CR_t, the throughput decreases (almost linearly) with the decrease in memory speed.

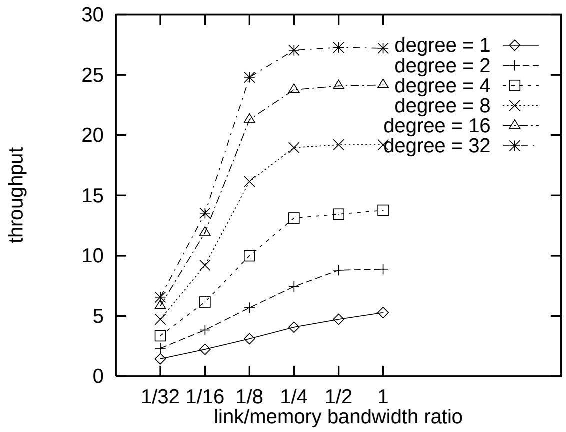


Fig. 2. Effect of memory speed on the maximum throughput (message size = 8)

As shown in Figure 2, CR_t in a multiplexed network is smaller than CR_t in a non-multiplexed network (degree = 1). For example, CR_t for non-multiplexed network is 1, while CR_t for networks with multiplexing degree 16 is between $\frac{1}{4}$ and $\frac{1}{8}$ in this experiment. In other words, for non–multiplexed networks, any mismatch between the memory and the transmission speed will reduce the network throughput, while when the multiplexing degree is 16, slowing down the memory speed up to $\frac{1}{8}$ of the optical transmission speed does not degrade the throughput. This indicates that TDM results in less memory pressure on the source nodes to achieve its maximum throughput. Hence TDM bridges the gap between the data transmission speed and the memory speed.

It is possible to approximate CR_t for a given multiplexing degree, K. Specifically, if C is the average number of connections that originate or terminate at a node at any given time, then in order not to reduce the throughput, the memory should be fast enough to fetch (or store) C packets every K time slots. This will allow one packet to be sent over each connection every K time slots, thus fully utilizing the bandwidth available on the K-way multiplexed connections. This gives:

.

$$CR_t = \frac{C}{K}.$$

Clearly, the value of C is between 0 and K since at most K channels can be established at any source or destination. Moreover, C should increase when K increases since multiplexing increases the number of connections that can co-exist at the source or destination nodes. In fact, in an ideal network in which network control is very fast and efficient, it is reasonable to assume that doubling K will double C, thus making CR_t unaffected by the multiplexing degree. However, as can be seen from Figure 2, CR_t decreases with K, which means that C is a sublinear function of K. The decrease in CR_t is not linearly proportional to K

since C increases with K.

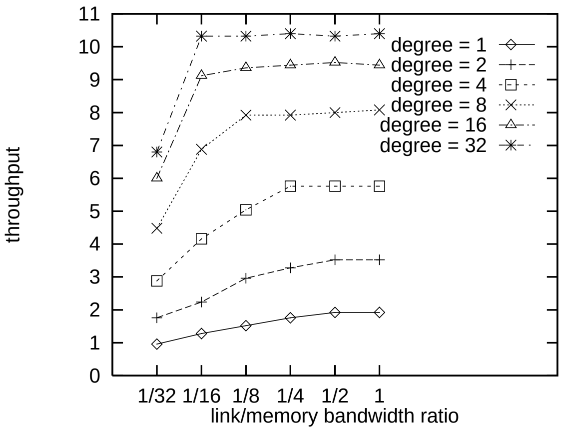


Fig. 3. Throughput when control network is slow (message size = 8)

The above argument can be further validated by considering the effect of the control network speed. Figure 3 shows the communication performance with a slow control network speed. This experiment has the same setting as in Figure 2 except that the control network speed is 4 times slower. We observe that, for a given multiplexing degree, slowing down the control network results in a smaller CR_t (in addition to reducing the network's throughput). This can be explained by noting that a slower control network reduces the number of established connections, and thus reduces C. Moreover, notice that the decrease in CR_t with K is closer to linear than the case of the faster control network. Again, this is because, with a slow control network, the communication bottleneck is shifted from the data network to the control network, and thus C does not increase as much with K.

The number of connections C is also affected by the message size. Specifically, long messages will result in long connections, which will tend to increase the number of connections established in the network, thus increasing C. Figure 4 shows the effect of the message size, measured in terms of the number of data packets, on the maximum throughput. In this experiment, we set the multiplexing degree to be 16 and set all other parameters to be the same as in Figure 2. We can see from this figure that, for message size 2, CR_t is $\frac{1}{16}$, while for message size 32, CR_t is between $\frac{1}{2}$ and $\frac{1}{4}$. Hence, TDM is more effective in bridging the gap between memory and data transmission speeds when the message size is small.

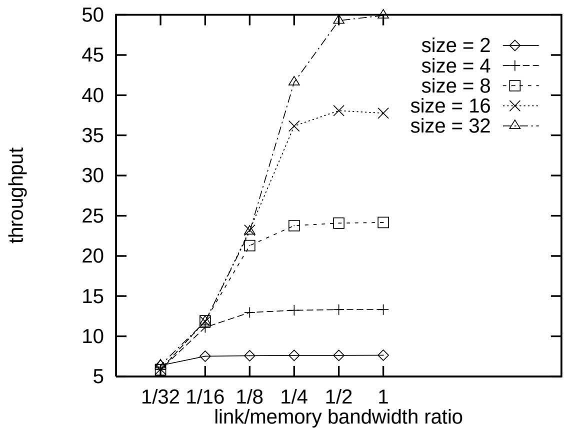


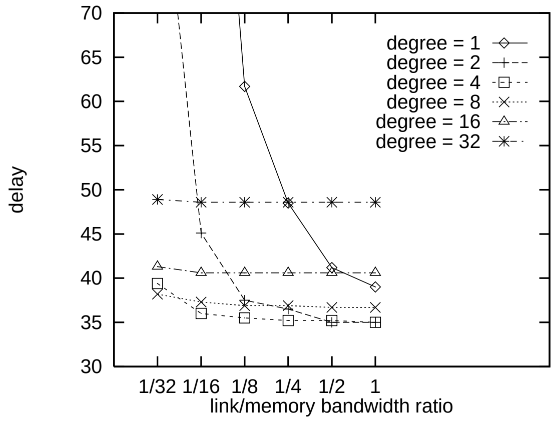
Fig. 4. Effect of message size on CR_t (degree = 16)

3.2 Effect of memory speed on communication delay

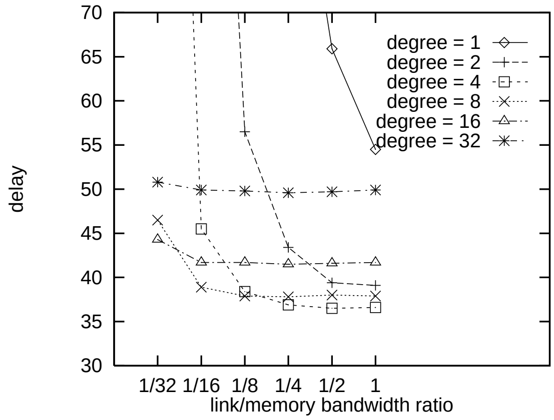
The communication delay is defined in this section as the average delay between the time a connection request is generated to the time at which the first packet of the data is received at the destination node. The delay in establishing a data communication path can be due to conflict induced by the control network or due to conflict induced by the data network. The former conflict results from control packets contending for the same buffers in the packet-switched control network, while the latter conflict results from the unavailability of data channels. Either conflicts causes the control packets to be blocked waiting for resources to be released (buffers in the former case and data channels in the latter case). In order to avoid deadlock, we use a timeout strategy in our simulation; a control packet is dropped after a certain timeout period and a "fail" message is sent to the sending node, which re-starts the path reservation phase after a random waiting period.

In this section, the communication delay will be estimated at low traffic compared to the saturation traffic considered in the last section. When the network is saturated, the throughput is equal to the generation rate. For instance a throughput of 24 packets per time slot in Figure 2 corresponds to a generation rate $r = \frac{24}{8*256} = 0.0117$ messages per node per time slot. In Figure 5, we show the delay for $r = 0.0005$ and 0.0015 in a 256 node network as a function of the link/memory bandwidth ratio, assuming 8-packet messages and different multiplexing degrees.

The multiplexing degree affects the delay in establishing a connection in two opposing manners. On the one hand, a larger multiplexing degree means that there is more data channels available for establishing the connection, thus reducing the delay. We will refer to this effect by $E_{channel}$. On the other hand, with multiplexing degree of K, one packet of a given message is sent on an established connection every K time slots, which means that connections are held longer when the multiplexing degree is larger. This in turns increases the probability of blocking and thus increases the delay. We will refer to this effect by $E_{duration}$. That is, when K increases, $E_{channel}$ causes the delay to decrease while $E_{duration}$ causes the delay to increase. Because of the two opposing effects, the minimum delay is obtained at a certain multiplexing degree, K_{opt}, which depends on many factors (see [4]).



(a) generation rate = 0.0005



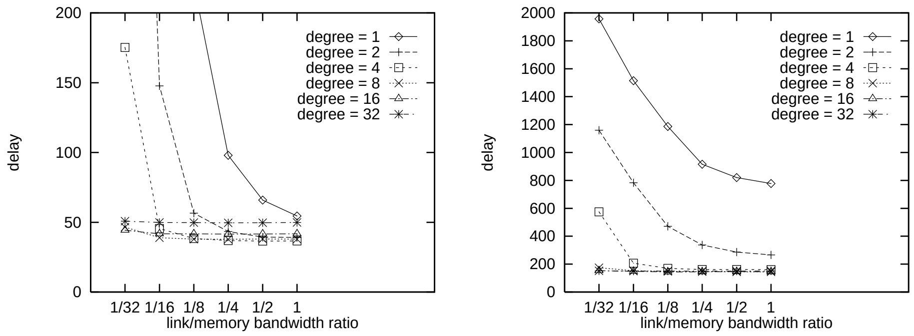
(b) generation rate = 0.0015

Fig. 5. Effect of memory speed on the delay (message size = 8)

Although we will not study the effect of $E_{channel}$ and $E_{duration}$ on the optimum multiplexing degree, we will use $E_{duration}$ to clarify the effect of the link/memory bandwidth ratio on the delay. Specifically, as we did in the last section for the throughput, we define *the delay critical ratio*, CR_d, as the link/memory bandwidth ratio below which the delay increases. That is, slowing down the memory up to CR_d does not affect the delay. For very low generation rates (as is the case for Figure 2(a)), the probability of having more than one connection established per node is very low, and thus slowing down the memory speed by a factor of up to the multiplexing degree K does not affect the delay. However, when the memory speed is slowed down by a factor larger than K, the duration of each connection is increased since data will not be available for transmission when needed. This will cause the delay to increase due to the $E_{duration}$ effect. Hence, $CR_d = \frac{1}{K}$. When the generation rate, r, increases (as is the case for Figure 2(b)), the number of connections established per node increases, thus causing CR_d to be larger than $\frac{1}{K}$. Hence, in general,

$$CR_d = \frac{C}{K}.$$

where C is the average number of connections established per node.



(a) fast control network (b) slow control network

Fig. 6. Effect of control network speed on the delay (message size = 8, $r = 0.0015$)

In Figures 6, we show the effect of the control network speed where Figures 6(a) is a re-drawing of Figure 5(b) using a different scale, and Figure 6(b) is the delay for the case when the control network is slowed down by a factor of four. Clearly, slowing down the control network increases the delay, but the value of CR_d seems to be relatively unaffected since the value of C is more affected by the generation rate, r, rather than by the control speed.

Finally, we show in Figure 7 the effect of the message size on the delay. For a given link/memory bandwidth ratio, longer messages lead to increased delay due

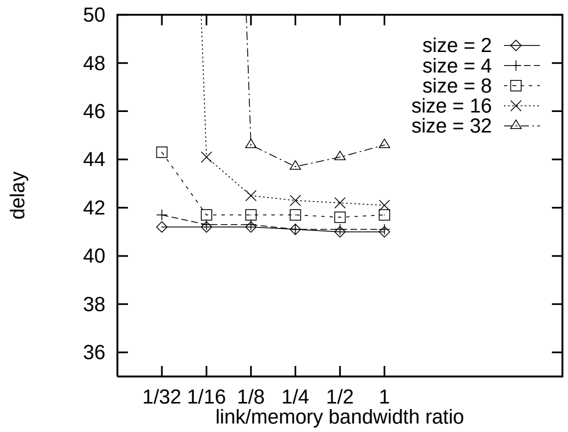


Fig. 7. Effect of message size on the delay (degree = 16, r = 0.0015)

to the $E_{duration}$ effect. Also, longer messages means larger traffic on the network (message generation rate is fixed at $r = 0.0015$), which in turns increases CR_d as discussed earlier.

4 Conclusion

In this work, we study the impact of the gap between memory and optical data transmission speeds on 2D torus networks. We define throughput and delay critical ratios as the link/memory speed ratios after which slower memory will decrease the throughput and increase the delay, respectively. We found that time-division multiplexed networks always result in smaller critical ratios than non-multiplexed networks. Large message sizes and fast control networks increase the number of connections at any give node, thus increasing the critical ratio due to the increase pressure on the memory. We conclude that, in addition to reducing the impact of the gap between data transmission speed and network control speed, TDM also absorbs the effect of mismatch between the memory speed and the optical communication speed, especially for small message sizes or slow network control. That is, TDM leads to the efficient utilization of the large optical bandwidth.

References

1. H. Scott Hinton, "Photonic Switching Using Directional Couplers," *IEEE Communication Magazine*, Vol. 25, No. 5, pp 16-26, 1987.
2. P.R. Prucnal, I. Glesk and J.P. Sokoloff, "Demonstration of All–Optical Self–Clocked Demultiplexing of TDM Data at 250Gb/s", In *Proceedings of the first international workshop on massively parallel processing using optical interconnections.* Pages 106–117, Cancun, Mexico, April 1994.

3. C. Qiao and R. Melhem, “Reconfiguration with Time Division Multiplexed MIN's for Multiprocessor Communication.” *IEEE Trans. on Parallel and Distributed Systems*, Vol. 5, N0. 4, April 1994.
4. X. Yuan, R. Melhem and R. Gupta, “Distributed path reservation algorithms for multiplexed all–optical interconnection networks.” *Int. Symp. on High Performance Computer Architecture - HPCA-3*, 1997.