ANALYSIS OF MULTISTAGE NETWORKS WITH UNIQUE INTERCONNECTION PATHS [1]

by

T. N. Mudge and B. A. Makrucki
Department of Electrical and Computer Engineering
University of Michigan
Ann Arbor, MI48109
313/764-0203

Abstract

This paper describes an analysis of multistage interconnection networks with unique interconnection paths where queues are placed in the $b \times b$ crossbar switches from which the networks are constructed. An approximate model of the network's behavior is developed. From this model, communication delay time and network throughput are derived. In addition, using the model, queue lengths may be chosen so that the network satisfies certain performance requirements.

1. Introduction.

In high performance multiprocessor computer systems, a high capacity communication channel is typically required if processor communication/data movement is not to be a system bottleneck. A wide variety of interconnection networks have been developed to satisfy this need. This paper describes an analysis of a class of packet switched multistage interconnection networks that have unique interconnection paths. An approximate model is developed that is appropriate for use in multiple instruction, multiple data stream (MIMD) multiprocessor systems. Due to space limitations, this paper has been abstracted from [MaM82] where complete derivations and references are given.

Section 2 describes the multiprocessor model assumptions and the performance measures to be derived (several more are implicit in the analysis). Section 3 describes the queueing analysis (with approximations) of the model and derives expressions for the performance measures developed in section 2. Section 4 is the conclusion.

2. System Model, Operation, and Assumptions.

This section describes the multiprocessor system to be modeled. System operation in the context of the model is described and simplifying system operation assumptions are described.

The multiprocessor architecture is shown in Figure 1 (all figures are at the end of the text). The system consists of processing elements (PE's, processors with local memory) connected by an interconnection network.

The system to be modeled is a packet switched system with n PE's ($n = b^z$, $b = 2^k$, k an integer, and z is the number of stages in the multistage interconnection network), or *packet sources*, that emit communication packets of constant length. A packet consists of a destination address field and a data field. For example, a packet may be a memory word or several memory words. When a source emits a packet, the packet is destined for one of the n *sink* devices (PE's here), i.e., packet broadcasting is not modeled.

Two network performance measures are derived, they are measures of communication delay time and network throughput. They are:

(1) *Packet delay time* (*PDT*) - this is the *average* time required for a packet to reach its destination. *PDT* is the average delay encountered by a packet from the time of emission from its source to the time of arrival at its destination.

(2) *Network throughput* (*NTP*) - this is the *average* rate of packet flow out of the network at the destination side.

The interconnection network for this analysis is assumed to have the unique transmission path property. That is, for any transmission path required there exists only a unique set of switch settings that will yield the transmission path, or put another way there is only one choice of a transmission path for every route between PEs. Multistage networks which satisfy this requirement are those that are bit controlled, i.e., where each stage in the network determines its switch settings by using a unique field from the destination address tag. Furthermore, these fields are mutually disjoint. Multistage networks which exhibit this property include delta, generalized cube, and omega networks. The requirement of unique transmission paths will be seen later in the network analysis.

2.1. Assumptions.

(1) All processors behave independently.

(2) Each processor emits packets as a Poisson process with rate λ. Thus program behavior is modeled as follows: a program executing on processor i emits packets at any time, the average time between packet emissions is $\frac{1}{\lambda}$.

(3) When a processor emits a packet, the selection of a destination sink is assumed to be uniformly distributed over all sink devices. This is an approximation, since a processor will not emit packets

[1] This research was supported in part by NSF grant MCS-8009315.

Preprint from the proceedings of
14th Southeastern Symposium on
System Theory
April 15-16, 1982

destined for itself.

From the bit controlled property, and the uniform distribution assumption, it may be seen that for every packet traversing the network at stage i, its position at the next stage is distributed uniformly over all outputs of the next stage switch to which it is routed (each bit of the address field is either 0 or 1 with equal probability). Thus packet routing at any stage is *independent* of all other stage routings. This allows a stage-by-stage queueing analysis to be done. Figure 2 shows the queueing network configuration of a $b \times b$ crossbar switch on which delta networks are based.

3. Network Queueing Analysis.

This section analyzes interconnection network behavior to find *PDT* and *NTP*.

The analysis will be performed in a stage-wise fashion. Initially the first stage will be analyzed (note that *all* queues in a given stage behave similarly due to symmetry in processor emission rates and uniform destination distributions) to find queue behavior. The results of the first stage may then be used to find second stage behavior. Likewise for all successive stages. Approximations will be made in order to make the analysis tractable.

Again, each packet entering a $b \times b$ switch is randomly destined (with uniform distribution) for one of the b output queues. All inputs to all $b \times b$ switches in the first stage are Poisson processes with rate λ. Thus, by decomposition and superposition of Poisson processes, each queue in the first stage sees a Poisson process with rate $\sum_{i=1}^{b} \frac{\lambda}{b} = \lambda$ at its input. Figure 2 shows the situation.

The exponential servers model randomness associated with the time required to move a packet from one stage to the next. Synchronous queue servers take a non-zero amount of time to move a packet to the next stage. Thus, multiple packets may try to enter a queue simultaneously (in a synchronous design), some of which will be delayed. The use of an exponential server is an attempt to models this interstage data movement delay, without unduly complicating the analysis.

All queues in the first stage behave identically and independently. Therefore, it suffices to analyze a single queue of the first stage.

3.1. First Stage, Single Queue Analysis.

For an $M/M/1/L$ queue, results from queueing theory are available [GrH74]. Let,

$$p_k = Pr\{k \text{ packets in the queueing station}\} \quad 0 \le k \le L$$

be the steady-state, general-time, occupancy probabilities for a queueing station. Then,

$$p_k = \begin{cases} \frac{1-\rho}{1-\rho^{L+1}}\,\rho^k & 0 \le k \le L,\ \rho \ne 1,\ M/M/1/L \quad (1.1) \\ \frac{1}{L+1} & 0 \le k \le L,\ \rho = 1,\ M/M/1/L \quad (1.2) \end{cases}$$

Where $\rho = \frac{\lambda}{\mu}$.

Let N be the random variable representing the steady-state number of packets in a queueing station (N will be subscripted with a queue number later). Then:

$$E[N] = \sum_{k=1}^{L} k p_k = \frac{\rho[1-(L+1)\rho^L + L\rho^{L+1}]}{(1-\rho)(1-\rho^{L+1})} \qquad M/M/1/L \quad (2)$$

Thus the first stage queue is easily analyzed. For successive stages the situation is generally complicated.

Consider the interdeparture process at the output of an $M/M/1/*$ (* denotes any queue length including infinite) queueing station with Poisson input rate λ and Poisson service rate μ, its probability density may be shown to be [MaM82]:

$$f_T(t) = \begin{cases} \frac{1-\rho-p_0}{1-\rho}\,\mu e^{-\mu t} + \frac{p_0}{1-\rho}\,\lambda e^{-\lambda t} & \rho \ne 1 \quad (3.1) \\ \frac{1}{L+1}\,\mu^2 t e^{-\mu t} + \frac{L}{L+1}\,\mu e^{-\mu t} & \rho = 1 \quad (3.2) \end{cases}$$

(3.1) is equivalent to a 2-hyperexponential server as shown in Figure 3.a provided $0 \le \frac{p_0}{1-\rho} \le 1$. (In actuality, p_0 for $M/M/1/L$ queues is greater than $1-\rho$ so Figure 3.a is not strictly valid for $M/M/1/L$ queues.) (3.2) is equivalent to Figure 3.b. Note that the interdeparture process is a renewal process with a *non-memoryless* interdeparture time distribution. The non-memorylessness of the interdeparture process implies that processes input to second stage queues are not renewal processes unless, to an approximation, either $p_0 \approx 1-\rho$, $p_0 \approx 0$, or $\rho = 1$ with $\frac{L}{L+1} \approx 1$. [If processes input to queueing stations are not renewal processes, analysis is very complex and simulation is a simpler technique for obtaining accurate results.]

Three cases arise:

Case I

$p_0 \approx 1-\rho$ corresponds to a *lightly loaded* $M/M/1/L$ queueing station with an output process that is close to a Poisson process with rate λ. Note that $p_0 \approx 1-\rho$ is also an approximation to an $M/M/1/\infty$ queue, which has $p_0 = 1-\rho$. Since $\frac{1}{\mu}$ is the average time required for a packet transfer between stages and $\frac{1}{\lambda}$ is the average time between packet emissions (i.e., message emissions) from processors, it is our contention that $\mu \gg \lambda$ in a typical multiprocessor system running in an independent processor, MIMD mode. The condition $\mu \gg \lambda$ ensures case I.

Cases II and III (which correspond to a *saturated* $M/M/1/L$ queueing station and an $M/M/1/L$ queueing station with $\rho \approx 1$) are described in [MaM82]. Due to space limitations and our contention that case I is most applicable in the majority of situations, only its analysis is presented here.

In case I, first stage output processes are approximately Poisson so the second stage has approximately Poisson inputs. Obviously, there is some inaccuracy involved unless $p_0 = 1-\rho$; however, the region of validity for this approximation may be characterized in terms of

allowable values for p_0, ρ, and L.

Consider the region for which the lightly loaded $M/M/1/L$ approximation applies. Since an $M/M/1/\infty$ queue has a Poisson interdeparture process with rate λ, a simple bound on the difference between $p_{0_{M/M/1/\infty}}$ and $p_{0_{M/M/1/L}}$ will suffice (only p_0 affects $f_T(t)$ for both $M/M/1/L$ and $M/M/1/\infty$ queues when $\rho \neq 1$) for a bound on error. Define D to be the maximum allowable difference between $p_{0_{M/M/1/\infty}}$ and $p_{0_{M/M/1/L}}$ in relative error with respect to $p_{0_{M/M/1/\infty}}$. Then:

$$D \geq \frac{p_{0_{M/M/1/L}} - p_{0_{M/M/1/\infty}}}{p_{0_{M/M/1/\infty}}} \tag{4}$$

And

$$D \geq \frac{\frac{1-\rho}{1-\rho^{L+1}} - (1-\rho)}{1-\rho}$$

That is, $\rho \leq \left(\frac{D}{D+1}\right)^{\frac{1}{L+1}}$ will satisfy (4). Thus selecting D places a bound on the region of validity (with respect to D) for the light loading approximation.

Case II and III validity regions are described fully in [MaV82].

Figure 4 shows the regions of analysis validity for $p_{0_{max}} = .05$. As can be seen from the graph, $L \geq 30$ implies the analysis is relatively accurate for almost all ρ.

3.2. Network Analysis.

Using the approximations from section 3.1, the network will now be analyzed to find *PDT* and *NTP* from section 2.

Since queues are finite in length, packets will be rejected when they attempt to enter full queues. Due to this effect, packets may be lost at any stage of the network. That is, the network does not exhibit the blocking property [DiJ80] (which is an intractable analysis except for the simplest networks of queues). When a packet is lost, it is routed back to its source to be resubmitted (alternatively, buffers could be placed between stages for lost packets, but this amounts to lengthening queues). Figure 5 shows the packet return path configuration. The submission/resubmission process resembles a Bernoulli process where the probability of event occurrence is:

$$p = Pr\{\ \textit{packet is not rejected at any queue}\ \}.$$

This approximation is supported by discrete time analysis and simulation.

The average number of trials before the first event occurrence in a Bernoulli process is:

$$E[\textit{number of trials before event occurrence}] = \frac{1-p}{p}$$

Define the following:

$$p_{ki} = Pr\{\ k \textit{ packets in queue } i \textit{ at an arrival}\ \}$$

$$T_i = \begin{cases} \textit{random variable representing the time spent} \\ \textit{by a packet in queue } i\textit{, service time} \\ \textit{is included in this measurement.} \end{cases}$$

$$E[T_{reject}] = \begin{cases} \textit{the average time from packet emission} \\ \textit{to return when the packet is lost (rejected)} \\ \textit{somewhere in the network.} \end{cases}$$

Then,

$$p = \prod_{i=1}^{z}(1 - p_{Li})$$

With the Bernoulli submission/resubmission approximation:

$$PDT \approx \sum_{i=1}^{z} E[T_i] + \frac{1-p}{p} E[T_{reject}] \tag{5}$$

And

$$\begin{aligned} E[T_{reject}] = {} & (1-p_{L1})p_{L2}E[T_1] \\ & + (1-p_{L1})(1-p_{L2})p_{L3}(E[T_1] + E[T_2]) \\ & + \cdots \\ & + (1-p_{L1})(1-p_{L2})\cdots(1-p_{L\,z-1})p_{Lz}(E[T_1] \\ & \qquad + \cdots + E[T_{z-1}]) \\ = {} & \sum_{i=2}^{z}\left[\sum_{j=1}^{i-1} E[T_j]\right] p_{Li}\prod_{j=1}^{i-1}(1-p_{Lj}) \qquad z \geq 2 \end{aligned} \tag{6}$$

Resubmission processing by sources is assumed to be instantaneous.

Case I Analysis.

For the light loading approximation, Figure 6 shows a series of z queues that represents a transmission path through the network. The equivalence of each of the queues in the transmission path relies on the equal rate, uniform destination distribution, and unique transmission path assumptions. The analysis may be extended (by considering Poisson flow rates) to accommodate unequal rates with general destination distributions.

From Little's formula

$$E[T_i] = \frac{E[N_i]}{\lambda(1-p_{Li})} \qquad i = 1, 2, \cdots, z.$$

From (1.1):

$$p_L \equiv p_{Li} = \frac{1-\rho}{1-\rho^{L+1}}\rho^L \qquad i = 1, 2, \cdots, z.$$

$$1 - p_L \equiv 1 - p_{Li} \doteq \frac{1-\rho^L}{1-\rho^{L+1}} \qquad i = 1, 2, \cdots, z.$$

So from (2):

$$E[T] \equiv E[T_i] = \frac{1-(L+1)\rho^L + L\rho^{L+1}}{\mu(1-\rho)(1-\rho^L)} \qquad i = 1, 2, \cdots, z.$$

And,

$$p = \prod_{i=1}^{z}(1-p_{Li}) = \left(\frac{1-\rho^L}{1-\rho^{L+1}}\right)^z$$

From (6):

$$E[T_{reject}] = p_L \sum_{i=2}^{z} (i-1)E[T](1-p_L)^{i-1}$$

$$= p_L E[T] \sum_{k=1}^{z-1} k(1-p_L)^k$$

$$= \frac{\beta}{1-\beta}\left(1 - z\beta^{z-1} + (z-1)\beta^z\right)E[T] \qquad z \geq 2$$

$$\beta = 1 - p_L = \frac{1-\rho^L}{1-\rho^{L+1}}$$

$$p = \beta^z$$

Therefore,

$$PDT = \left\{ z + \frac{1-\beta^z}{(1-\beta)\beta^{z-1}}\left(1 - z\beta^{z-1} + (z-1)\beta^z\right)\right\} E[T]$$

To find *NTP*, simply subtract the rate of network packet loss from $n\lambda$, the packet input rate.

$$NTP = n\lambda - \sum_{i=1}^{z} n\lambda p_{Li}$$

$$= n\lambda \left[\frac{1 + (z-1)\rho^{L+1} - z\rho^L}{1-\rho^{L+1}} \right]$$

PDT and *NTP* for $n = 64$, $b = 4$, verses ρ for several L are shown in Figure 7. Note that as $L \to \infty$, $NTP \to n\lambda$ as would be expected because no packets are lost when L is very large (as is also the case for networks with blocking, but in that case $NTP = n\mu(1 - p_{0z})$). Note also that as $L \to \infty$ (or L gets large) the cost of the network becomes large for both delta networks and crossbars (i.e., the largest cost factor of the network becomes the queues) so it is actually less costly to use a crossbar because only n queues are required whereas for delta networks $n \log_b n$ queues are needed. With a single stage network such as a crossbar, *PDT* is considerably better also.

Case II and III results are derived in [MaM82].

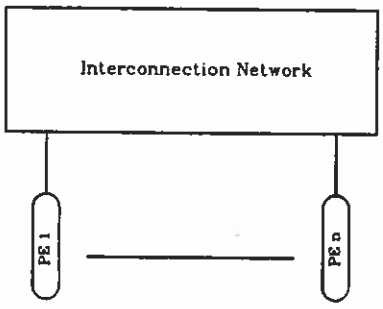

Figure 1. System architecture.

4. Conclusion.

An analysis of a class of packet switched, multistage interconnection networks (that exhibit the bit controlled property) was presented. The analysis (with approximations) allows network communication delay and network throughput to be evaluated for certain combinations of queue lengths, interstage transfer rates, and processor packet emissions rates. From the analysis, queue lengths that satisfy performance requirements may be chosen.

5. References.

[DiJ80]
D. M. Dias, and J. R. Jump, "Analysis and Simulation of Buffered Delta Networks," *Proc. Workshop on Interconnection Networks*, Purdue University, April 21-22, 1980.

[GrH74]
D. Gross, and C. M. Harris, *Fundamentals of Queueing Theory*, John Wiley and Sons Inc., New York, 1974.

[MaM82]
B. A. Makrucki, and T. N. Mudge, *A Queueing Model of Delta Networks*, SEL Report No. 159, Department of Electrical and Computer Engineering, University of Michigan, January 1982.

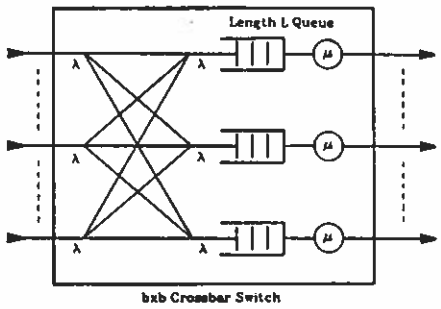

Figure 2. b x b crossbar switch.

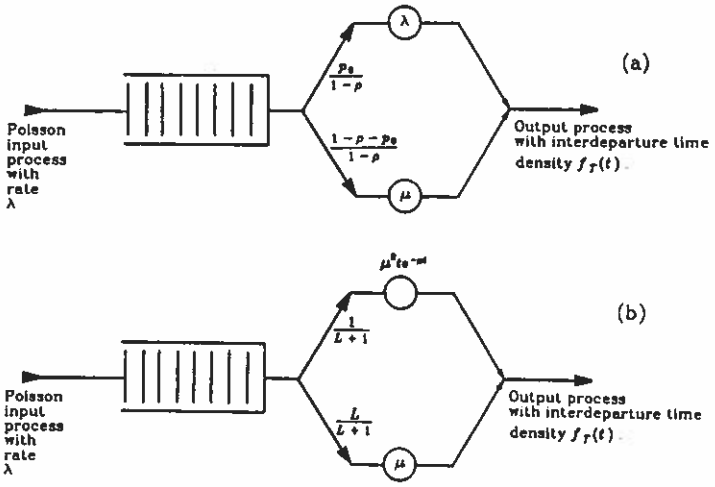

Figure 3. Equivalent first stage queue

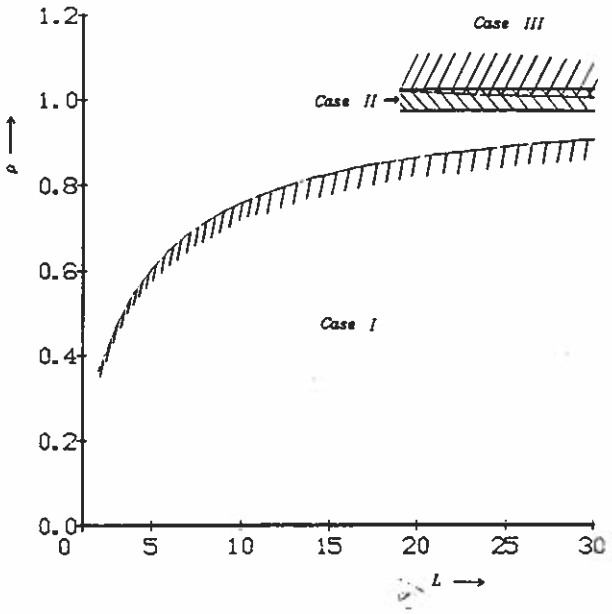

Figure 4. Approximation validity regions.

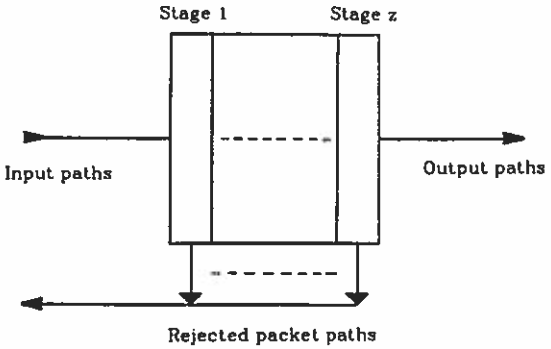

Figure 5. Resubmission path.

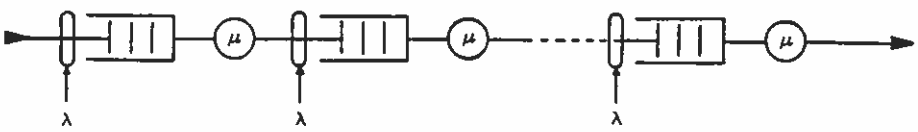

Figure 6. Case I equivalent series of queues.

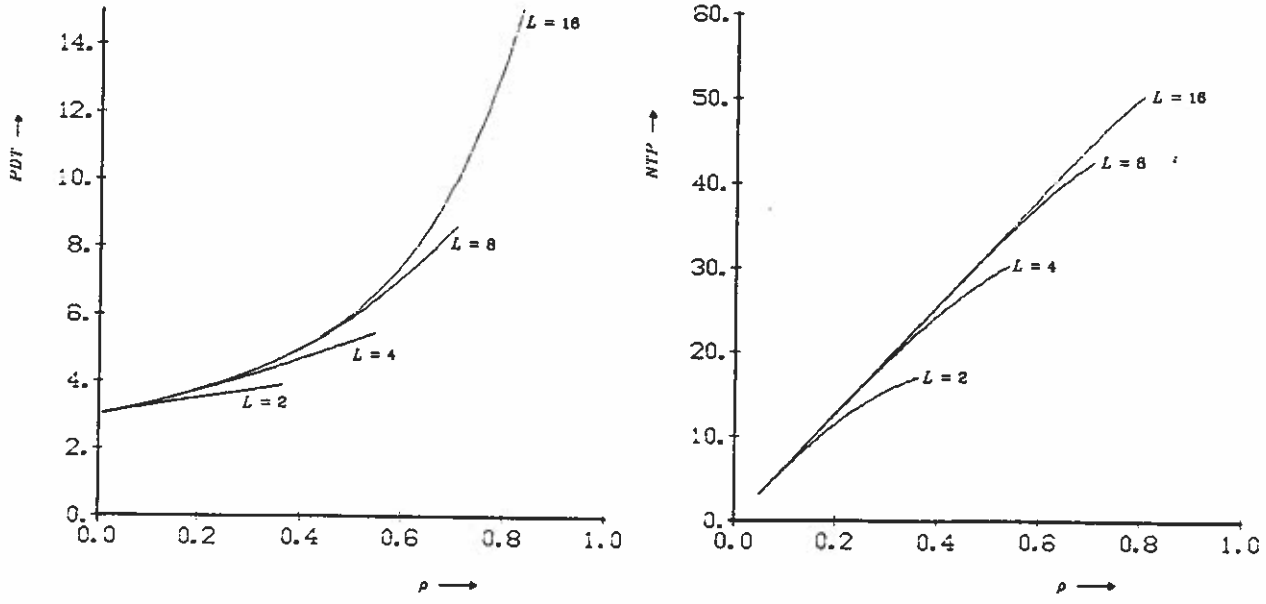

Figure 7. *PDT* and *NTP* for $n = 64$, $b = 4$.