AN APPROXIMATE QUEUEING MODEL FOR PACKET SWITCHED MULTISTAGE INTERCONNECTION NETWORKS[1]

by

T. N. Mudge and B. A. Makrucki
Department of Electrical and Computer Engineering
and the Program in Computer, Information and Control Engineering
University of Michigan
Ann Arbor, Michigan 48109
313/764-0203

Abstract--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 develops an approximate queueing model for packet switched multistage interconnection networks that have unique interconnection paths. The model assumes that packet buffering occurs at each stage of a network and that the buffers are of finite size. The major source of approximation is the assumption that packet interarrival times at interstage points in a network are exponentially distributed, i.e., packet arrivals at these points form a Poisson process. This assumption leads to simple expressions for the average time required for a packet to traverse a network and for the average rate of flow out of the network at the destination side. The accuracy of these expressions are compared to simulations and they are found to yield results that, for the most part, differ by less than 10%. It is concluded that the model offers a simple and economical way of deriving approximate performance figures for packet switched multistage interconnection networks, that could otherwise only be obtained by simulation.

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. Prominent among these are the following: the networks of Clos and Benes [Clo53,Ben65], the Banyan networks* of Goke and Lipovski [GoL73], the Data Manipulator network of Feng [Fen74], the Omega network* of Lawrie [Law75], the STARAN flip network* of Batcher [Bat76], the Indirect Binary n-cube* of Pease [Pea77], the Generalized Cube* of Siegel and Smith [SiS78, SiM81b], the Delta network* of Patel [Pat79], the Baseline network* of Wu and Feng [WuF80], and the Augmented Data Manipulator network of Siegel and McMillen [SiM81a]. Furthermore, as a consequence of the rapid development of VLSI circuit technology, the feasibility of crossbar networks* is being reexamined [Fra80,MaM82b]. For a good survey of the state-of-the-art in networks see Siegel [Sie80]. A unified theory developed by Lipovski and Malek that characterizes the inter-relationship of many of these networks will appear in [LiM82].

This paper describes an analysis of a class of packet switched multistage interconnection networks that have unique interconnection paths (e.g. those marked with an asterisk in the above list have the unique interconnection path property). An approximate model is developed that is appropriate for use in multiple instruction, multiple data stream (MIMD) multiprocessor systems. Earlier research which is relevant to the model and analysis presented here includes the work by Dias and Jump [DiJ81] which develops some interesting results about the throughput and bandwidth of buffered Delta networks by using a combination of discrete time analysis and simulation, and the work described in [MaM82a,MaM82c].

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 presents the simulation results and Section 5 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 together with the simplifying assumptions.

The multiprocessor architecture is shown in Figure 1. 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[2], or *packet sources*, that emit communication packets of constant length. A packet consists of a destination address field and a data field. The data field may consist of a fixed number of data words or other types of information such as a request to read and return data from the destination. When a source emits a packet it is destined for exactly one of the n *sink* devices (PE's here), i.e., packet broadcasting is not modeled.

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

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

[2] We will assume for convenience that $n = b^x$, that $b = 2^k$, with k an integer, and that x is the number of stages in the multistage interconnection network.

Preprint from the proceedings of
The 3rd International Conference
on Distributed Computing Systems
October,1982.

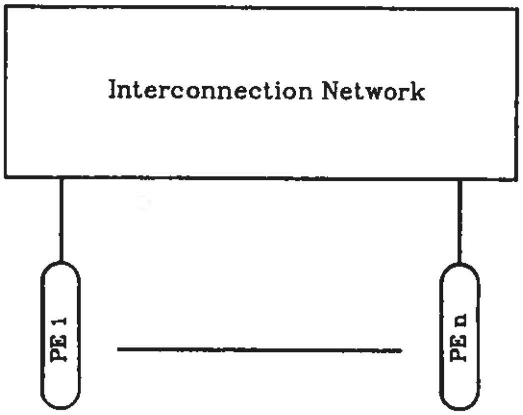

Figure 1. System Architecture.

(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 network for this analysis is assumed to have the unique interconnection path property. That is, for any interconnection path required there is a unique set of switch settings that will yield that path, or put another way, there is only one choice of a interconnection path for every route between PE's. Multistage networks which satisfy this requirement are those that are *bit-controlled*—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.

The assumptions are as follows:

(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 random points in time, with the only constraint being that the average time between packet emissions is $1/\lambda$. The validity of approximating the above type of processor behavior with Poisson processes or their counterpart in discrete time, Bernoulli processes, has been discussed in numerous places (see for example [BaS76,Bha75]).

(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 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 the interconnection networks are based.

(4) The queues in the basic crossbar (see Figure 2) used for buffering packets are finite in size—they can hold L packets. The analysis for multistage packet switched interconnection networks in which L is allowed to be unbounded is given in [MaM82a,MaM82c]—see the "light loading" case.

(5) The routing mechanisms in the crossbars at each stage that dispatch packets forward to the next stage are assumed to be exponential servers. Deterministic servers would be a more accurate model, however, the resulting analysis becomes intractable.

3. Queueing Network Analysis.

This section analyzes the queueing network representation of the class of networks described in Section 2. The analysis will be performed in a stage-wise fashion with the output of stage i being the input to stage $i+1$.

Due to the symmetry in all processor source emission rates and destination distributions--all network output ports are equally likely to be an emission destination--all queues in stage i $(0 \le i \le z-1)$ have the same steady-state solution for such quantities as the average number in the queue and the average time

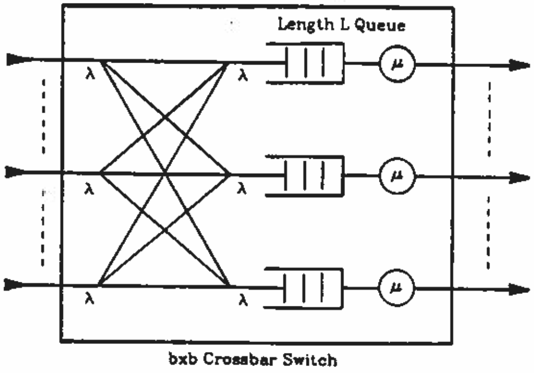

Figure 2. A $b \times b$ Crossbar With Queues.

spent in the queue. A solution for a single queue in stage i suffices for stage i analysis.

Consider the first stage. Each queue in the first stage sees a Poisson process with rate λ at its input. (A packet emitted from a processor enters a first stage $b \times b$ crossbar switch, there it is routed to one of the b output queues as seen in Figure 2. Thus the processor emission process branches into b Poisson processes each with rate λ / b. Each queue in a first stage $b \times b$ crossbar sees b contributions and thus sees a rate $\sum_{i=1}^{b} \frac{\lambda}{b}$ at its input).

In general, results for $M/M/1/L$ queues are available (see for example [GrH74]). Let:

$$p_k = Pr\{\, k \textit{ packets in the queueing station } \} \qquad 0 \le k \le L$$

be the steady-state, general time, queue occupancy probabilities for an $M/M/1/L$ queueing station (including the server). Then:

$$p_k = \frac{1-\rho}{1-\rho^{L+1}}\,\rho^k \qquad 0 \le k \le L, \quad \rho \ne 1 \tag{1}$$

Where $\rho = \lambda / \mu$. The $\rho = 1$ case will be ignored; it can simply be treated as a separate case of the following analysis.

Let N be the random variable representing the steady-state number of packets in an $M/M/1/L$ queue (including the server position). Then:

$$\begin{aligned} E[N] &= \sum_{k=1}^{L} k p_k \\ &= \frac{\rho[1-(L+1)\rho^L + L\rho^{L+1}]}{(1-\rho)(1-\rho^{L+1})} \end{aligned} \tag{2}$$

Using Little's equation then:

$$E[T] = \frac{E[N]}{\lambda(1-p_L)} \tag{3}$$

where T represents the time spent in the queueing station.

These equations (1,2,3) may be applied directly to determine first stage queue behavior. To determine successive stage behaviors, consider the output process of a first stage $M/M/1/L$ queue.

Let T_I be the random variable representing the time between departures of packets from an $M/M/1/*$ queue ($*$ denotes any queue length). Also, approximate departure point probabilities, π_k, with arrival point probabilities, p_k. Then the probability density of T_I is approximately [MaM82a]:

$$f_{T_I}(t) = \frac{1-\rho-p_0}{1-\rho}\,\mu e^{-\mu t} + \frac{p_0}{1-\rho}\lambda e^{-\lambda t} \qquad \rho \ne 1 \tag{4}$$

Since this density is *not* Poisson (i.e., it is not memoryless), input processes to second stage queues are *not* Poisson, they are not even renewal processes [Cin75]. The characterization of input processes to second (and succeeding) stage queues is very difficult. However, particularly if the $b \times b$ switches are large, the inputs to second stage queues are *approximately* Poisson. The approximate analysis used for stages 1 through $z-1$ will assume Poisson input processes. Simulation results are then used to examine the validity of the approximation.

To find the equivalent Poisson rate of second stage queue input processes, consider the average output rate of first stage queues. Call this λ_1, then:

$$\lambda_1 = \frac{1}{E[T_{I_0}]}$$

Where T_{I_0} is T_I for stage 0.

The approximation in general is to take λ_i to be a Poisson input rate for stage i queues, $1 \le i \le z-1$. That is:

$$\begin{aligned} \lambda_i &= \frac{1}{E[T_{I_{i-1}}]} \\ &= \frac{\lambda_{i-1}\mu(1-\rho_{i-1}^{L+1})}{\mu - \lambda_{i-1}\rho_{i-1}^{L+1}} \qquad 1 \le i \le z-1 \end{aligned} \tag{5}$$

Rewriting (2) with subscripts:

$$E[N_i] = \frac{\rho_i[1-(L+1)\rho_i^L + L\rho_i^{L+1}]}{(1-\rho_i)(1-\rho_i^{L+1})} \qquad 0 \le i \le z-1$$

and

$$E[T_i] = \frac{E[N_i]}{\lambda_i(1-p_{Li})} \qquad 1 \le i \le z-1 \tag{6}$$

where p_{Li} is p_L for stage i queues, $\rho_0 = \lambda / \mu$.

Using the approximate results given in equations 5 and 6, network performance measures may be determined.

PDT is simply the sum of $E[T_i]$ for $i = 0, 1, 2, \cdots, z-1$, if the packet does not get rejected. Thus:

$$PDT = \sum_{i=0}^{z-1} E[T_i] \tag{7}$$

and *NTP* is simply the rate of packet flow out of the n queues in stage $z-1$. Each server is busy $1 - p_{0z-1}$ (p_{0z-1} is p_0 for stage $z-1$) fraction of the time. Hence:

$$NTP = n\mu(1 - p_{0z-1}) \tag{8}$$

To account for packet loss and improve the accuracy of this analysis the rejection/resubmission process at any stage in the network may be modeled as a Bernoulli process (see [MaM82a,MaM82c] for a discussion of this technique).

4. Simulation Results.

Since the analysis relies on the approximations assumed in Section 2 and the approximation that interstage packet arrivals form a Poisson process assumed in Section 3, simulations were used to evaluate the effectiveness of the approximations. This section describes the results obtained.

Simulations of the queueing network of Figure 3 were run. This network of queues is equivalent in behavior to a path through the network of queues described in Section 2. That is, a packet entering the interconnection network with queues experiences the same average delay as one entering the equivalent network of Figure 3. Server utilizations are also the same in the equivalent network. Thus it suffices to simulate this equivalent network in order to estimate the effectiveness of the approximations proposed.

Note that the approximation is expected to become more accurate as L and b are increased (increasing L is discussed in [MaM82a]) because the superposition of a large number of independent processes tends to an approximate Poisson process (here it will be seen that for $b \ge 4$ the approximation is relatively accurate in predicting *PDT* and *NTP*). Strictly speaking the

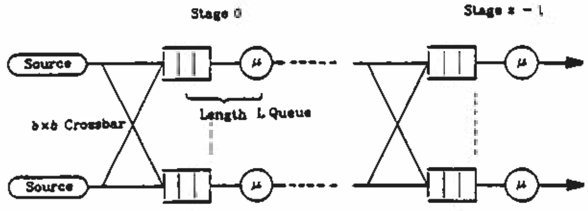

Figure 3. The Simulation Network.

superimposed independent processes should be "sparse" for their combination to exhibit Poisson behavior. In our terms this means the packet rates (λ's) should be very low. Inasmuch as this is not the case the result is an approximate model. We have discussed the low packet rate situation in [MaM82a,MaM82c]--see "light loading" case.

Table I shows simulated and calculated results for *PDT* and *NTP*. (All tables are after the references.) Table I is partially derived from Tables II and III. Tables II and III show results for average behavior within the network. Note that in tables II and III, $E[T_{I_0}]_s$ is close to $E[T_{I_0}]_c$ so the % error is written as 0. This measure has no significance in that input processes to the first stage *are* Poisson by assumption. Note also, that T_i' is the time spent in queue, not including service, and that *PDT'* is *PDT* not including the z service times. We have made it a point to show *PDT'* rather than *PDT* so that the % error is not reduced by the addition of the z service times. In spite of this the simple expressions given in equations 7 and 8 yield results that are fairly accurate--most results have less than a 10% error although there are occasionally errors up to 19%.

5. Conclusions.

A more accurate analysis would entail computing a probability distribution for the time spent by packets in the network. Even with the assumption of Poisson arrivals at the input to the network and exponential servers at each stage, deriving this distribution is essentially impossible. Thus simulation appears to be the only alternative to developing a rough model. In this paper we have developed such a model; through equations 7 and 8 it provides a simple way of calculating approximate performance figures for packet switched multistage interconnection networks. These figures are usually no more than 10% in error.

6. References

[BaS76]
F. Baskett, and A. J. Smith, "Interference in Multiprocessor Computer Systems with Interleaved Memory," *CACM*, Vol. 19, No. 6, June 1976, pp. 327-334.

[Bat76]
K. E. Batcher, "The flip network in STARAN," *Proc. of the 1976 Int'l Conf. on Parallel Processing*, August 1976, pp. 65-71.

[Ben65]
V. E. Benes, *Mathematical Theory of Connecting Networks and Telephone Traffic*, Academic Press, New York 1965.

[Bha75]
D. P. Bhandarkar, "Analysis of Memory Interference in Multiprocessors," *IEEE Trans. Computers*, Vol. C-24, No. 9, September 1975, pp. 897-908.

[Cin75]
E. Cinlar, *Introduction to Stochastic Processes*, Prentice-Hall, 1975.

[Clo53]
C. Clos, "A Study of Non-blocking Switching Networks," *The Bell System Technical Journal*, Vol. 32, March 1953, pp. 406-424.

[DiJ81]
D. M. Dias and J. R. Jump, "Analysis and Simulation of Buffered Delta Networks," *IEEE Trans. Computers*, Vol. C-30, No. 4, April 1981, pp. 273-282.

[Fen74]
T. Feng, "Data Manipulating Functions in Parallel Processes and their Implementations," *IEEE Trans. Computers*, Vol. C-23, No. 3, March 1974, pp. 309-318.

[Fra80]
M. A. Franklin, "VLSI Performance Comparison of Banyan and Crossbar Connection Networks," in [Sie80], pp. 20-28.

[GoL73]
R. Goke, G. J. Lipovski, "Banyan Networks for Partitioning on Multiprocessor Systems," *Proc. of the 1-st Annual Symposium on Computer Architecture*, 1973, pp.21-30.

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

[Law75]
D. H. Lawrie, "Access and Alignment of Data in an Array Processor," *IEEE Trans. Computers*, Vol. C-24, No. 18, December 1975, pp. 1145-1155.

[LiM82]
G. J. Lipovski, M. Malek, "A Theory for Multicomputer Interconnection Networks" *IEEE Trans. Computers*, (to appear).

[MaM82a]
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, pp.

[MaM82b]
T. N. Mudge, B. A. Makrucki, "Probabilistic Analysis of a Crossbar Switch," *Proceedings of the 9-th Annual International Symposium on Computer Architecture*, April 1982, to appear.

[MaM82c]
T. N. Mudge and B. A. Makrucki, "Analysis of Multistage Networks with Unique Interconnection Paths," *Proceedings of the 14-th Southeastern Symposium on System* Theory, April 1982, to appear.

[Pat79]
J. H. Patel, "Processor-Memory Interconnections for Multiprocessors," *Proc. 6-th Annual Symposium on Computer Architecture*, IEEE, April 1979, pp. 168-177.

[Pea77]
M. C. Pease, "The Indirect Binary n-Cube Microprocessor Array," *IEEE Trans. Computers*, Vol. C-26, No. 5, May 1977, pp. 458-473.

[Sie80]
H. J. Siegel, (Ed.), *Proceedings of the Workshop on Interconnection Networks*, Purdue University, April 21-22, 1980.

[SiM81a]
H. J. Siegel, R. J. McMillen, "Using the Augmented Data Manipulator Network in PASM," *Computer*, Vol. 14, No. 2, February 1981, pp. 23-33.

[SiM81b]
H. J. Siegel, R. J. McMillen, "The Multistage Cube: A Versatile Interconnection Network," *Computer*, Vol. 14, No. 12, December 1981 (to appear).

[SiS78]
H. J. Siegel, S. D. Smith, "Study of Multistage SIMD Interconnection Networks" *Proc. of the 5-th Annual Symposium on Computer Architecture*, April 1978, pp. 223-229.

[WuF80]
C-L. Wu, T-Y. Feng, "On a Class of Multistage Interconnection Networks," *IEEE Trans. Computers*, Vol. C-29, No. 8, August 1980, pp. 694-702.

Results for *PDT* and *NTP*									
b	L	ρ	n	PDT'_s	PDT'_c	% error	$NTP_s/n\mu$	$NTP_c/n\mu$	% error
4	4	.749	1024	932	1013	8.7	.532	.568	6.8
4	4	.901	"	1056	1136	7.6	.563	.598	10.5
4	4	1.25	"	1233	1304	5.8	.603	.624	3.0
4	4	2.5	"	1411	1481	5.0	.612	.636	3.9
4	8	.749	"	1895	1967	3.8	.676	.688	1.8
4	8	.901	"	2448	2510	2.5	.735	.741	.8
4	8	1.25	"	3227	3168	-1.8	.804	.772	-.98
4	8	2.5	"	3597	3642	1.3	.739	.777	5.1
8	4	.749	512	654	641	-2.0	.657	.609	-7.3
8	4	.901	"	719	724	.7	.688	.653	-5.1
8	4	1.25	"	915	875	-4.4	.716	.698	-2.5
8	4	2.5	"	1047	1040	-.7	.789	.720	-8.7
8	8	.749	"	1291	1217	-5.7	.786	.706	-10.2
8	8	.901	"	1720	1613	-6.2	.847	.777	-8.3
8	8	1.25	"	2295	2166	-5.6	.877	.828	-5.6
8	8	2.5	"	2360	2623	-.3	.871	.837	-3.9

Where,

$$PDT' = PDT - \frac{z}{\mu}$$

Values subscripted with "s" are simulation values, those subscripted with "c" are calculated values.

Table 1. Results for *PDT* and *NTP*.

Simulation Results for b = 4

L	ρ	Stage	$E[T_K]_s$	$E[T_K]_a$	% error	$E[T'_i]_s$	$E[T'_i]_a$	% error
4	.749	0	267	267	0	219	229	4.9
"	"	1	298	288	-3.4	200	212	6.4
"	"	2	334	305	-8.7	189	200	5.6
"	"	3	328	319	-2.7	169	190	12.5
"	"	4	362	332	-8.3	155	182	17.2
"	.901	0	222	222	0	275	274	.3
"	"	1	268	254	-5.2	223	241	8.2
"	"	2	303	278	-8.3	197	220	12.1
"	"	3	324	296	-8.6	191	206	8.0
"	"	4	331	312	-5.7	170	196	14.5
"	1.25	0	160	160	0	350	355	1.4
"	"	1	231	219	-5.2	256	277	8.2
"	"	2	273	252	-7.7	231	243	5.2
"	"	3	298	276	-7.4	197	222	13.3
"	"	4	322	295	-8.4	199	207	4.0
"	2.5	0	80	80	0	491	488	-.6
"	"	1	208	201	-3.4	280	298	6.3
"	"	2	256	241	-5.9	229	254	11.2
"	"	3	283	267	-5.7	212	229	7.9
"	"	4	310	288	-7.1	199	212	6.4
8	.749	0	267	267	0	437	421	-3.7
"	"	1	276	272	-1.4	389	405	4.1
"	"	2	278	277	-.4	374	391	4.5
"	"	3	291	281	-3.4	354	380	7.3
"	"	4	287	285	-.7	341	370	8.5
"	.901	0	222	222	0	583	592	1.5
"	"	1	242	236	-2.5	502	531	5.8
"	"	2	252	247	-2.0	494	490	-.8
"	"	3	260	255	-1.9	450	460	2.2
"	"	4	285	282	-1.1	419	437	4.3
"	1.25	0	160	160	0	956	923	-3.5
"	"	1	210	206	-1.9	625	668	6.9
"	"	2	231	226	-2.2	593	575	-3.1
"	"	3	244	239	-2.0	465	520	11.7
"	"	4	245	249	1.6	586	482	-17.9
"	2.5	0	80	80	0	1281	1268	-1.0
"	"	1	192	200	4.0	812	765	-5.8
"	"	2	235	222	-5.6	518	590	14.1
"	"	3	236	236	0	545	530	-2.7
"	"	4	264	246	-6.7	441	489	10.9

Table II. Network behavior for *b* = 4.

Table of Simulation Results for b = 8								
L	ρ	Stage	$E[T_R]_s$	$E[T_R]_c$	% error	$E[T'_i]_s$	$E[T'_i]_c$	% error
4	.749	0	267	267	0	230	229	-.4
"	"	1	296	288	-2.7	215	212	-1.4
"	"	2	283	305	7.8	209	200	-4.3
"	.901	0	222	222	0	265	274	3.4
"	"	1	267	254	-4.9	216	241	11.6
"	"	2	248	278	12.1	238	220	-7.6
"	1.25	0	160	160	0	359	355	-1.1
"	"	1	228	219	-3.9	295	277	-6.1
"	"	2	237	252	6.3	261	243	-6.9
"	2.5	0	80	80	0	476	488	2.5
"	"	1	203	201	-1.0	274	298	8.8
"	"	2	220	241	9.5	297	254	-14.5
8	.749	0	267	267	0	442	421	-4.8
"	"	1	273	272	-.4	390	405	3.8
"	"	2	245	277	13.1	459	391	-14.8
"	.901	0	222	222	0	637	592	-7.1
"	"	1	241	236	-2.1	479	531	10.9
"	"	2	219	247	12.8	604	490	-18.9
"	1.25	0	160	160	0	922	923	.1
"	"	1	208	206	-1.0	744	688	-7.5
"	"	2	201	226	12.4	629	575	-8.6
"	2.5	0	80	80	0	1271	1268	-.2
"	"	1	203	200	-1.5	706	765	8.4
"	"	2	207	222	7.2	653	590	-9.6

Where,

$$\%\ error = \frac{calculated\ value - simulated\ value}{simulated\ value} \times 100$$

$$E[T'_i] = average\ time\ spent\ in\ queue\ i.$$

Table III. Network behavior for b = 8.