BNL-22531
AMD-758

Conf-770812--1

# THROUGHPUT IN LOCALLY BALANCED COMPUTER SYSTEM MODELS*

Andrew S. Noetzel

Applied Mathematics Department
Brookhaven National Laboratory
Upton, NY 11973

March, 1977

*Work performed under the auspices of the ERDA.

# ABSTRACT

The optimization of throughput in locally balanced queueing network models is investigated. A general result, useful in the design of computer system models, shows that throughput is a non-decreasing function of the number of customers contained in any subnetwork. The throughput of the two queue network, which duplicates (through Norton's theorem) that of more complex networks, is characterized in terms of the processing rates at each queue. Then processor allocation algorithms that maximize throughput are shown for the case where processing power can be switched between queues. The maximization of throughput is shown first in the case that processing power allocations to a queue depend on the queue state only, and then, in an extension of known locally balanced queue, the case in which processing power is allocated on the basis of an entire subnetwork state. The latter case provides a simple and optimum rule for processor allocations that maximize throughput in networks containing multiprocessor systems.

## 1. Introduction

The complete solution of queueing networks of arbitrary configuration is possible only for the class of networks that have been variously described as separable, or having the local balance property or the product form solution [13,3,8]. In its recent development, this class of networks has been made to reflect features such as general service time distributions, various scheduling disciplines and multiple classes of customers. As a result the local balance model has found wide utility in the analysis of computer systems and networks [1,3]. It has been implemented as the basis of several interactive systems that yield quick analyses of computer systems and configurations [7,12]. Because of its unique tractability in general network configurations, the network model has also been studied as an approximation to networks whose queues do not meet the local balance requirement [5].

A central consideration in the use of network models for the development of computer systems, is that of throughput. In open network models, throughput is determined by the rate of the assumed source. In closed networks, throughput is determined by the combined effects of congestion at the queues of the network, usually with no one queue being identifiable as the only bottleneck.

In this paper, we consider the optimization of throughput in closed, locally balanced computer system models. First, the relationship of the processing rate of a subnetwork, or the throughput of a closed network, to the number of customers contained therein, is established. Then throughput is related to the processing rates of the queues in a closed two queue network. The two queue network is important because of its role as a model of larger networks, through the construction of Norton's equivalent queues.

These results are then used to show the optimization of throughput in locally balanced networks in which processing power tradeoffs between queues are possible. This model will reflect the cases in which several independent queues or logical processes are located at a multiprocessor system whose processors can be switched between the functions. A processor allocation rule for subnetworks, which extends the known queueing models with product form solutions, is demonstrated. The processor allocation rule can be shown to yield maximum throughput for the subnetwork.

## 2. Locally Balanced Networks and Norton's Theorem

The separable or locally balanced networks to be considered have fixed topology. After leaving queue i, a customer will go to queue j with fixed probability $p_{ij}$. Let P be the matrix of transition probabilities $p_{ij}$, $1 \le i,j \le M$, for a network of M queues. Let $\lambda_i$ be the mean flow rate into queue i. If $L = \left[\lambda_1\ \lambda_2\ \dots\ \lambda_M\right]$ is a vector of relative flow rates for the network, then $LP=L$. If the network is open, L is determined by the absolute input rate to the network. For closed networks, the above relation determines L to within a constant. Then the actual throughput at queue i, when the network contains N customers, is given by $\tau_i(N) = \lambda_i \frac{G(N-1)}{G(N)}$, where G(n) is the normalization factor for some L when the network contains n customers. G(n) is computed by the convolution algorithm [2,11,13]. Since the throughputs at the various branches of the network remain fixed as N and the processing rates at the queues are varied, the network throughput, or total processing rate, is optimized as throughput in any branch of the network is optimized.

The throughput characteristics of two queue networks are of considerable importance because of the reduction of arbitrary closed networks to an equivalent two queue network, made possible by Norton's theorem for computer networks [4]. The behavior of a particular queue q within a closed network $\eta$ is the same as that of q in a network with one other queue, which serves as the equivalent of all the queues of $\eta$ except for q (see figure 1). The processing rate U(n) of the equivalent queue, when it contains n customers, is determined by constructing network $\eta'$, which is the same as $\eta$ except for a short circuit in place of q. U(n) is simply the mean throughput in the link replacing q, when $\eta'$ contains n customers. To study via the equivalent network the effects of q in $\eta$ when $\eta$ contains N customers, the rates U(1),...,U(N) must be measured in the link in $\eta'$, when $\eta'$ contains 1,...,N customers.

It is possible to analyze in the same manner as q the behavior of any subnetwork $\sigma$ that can be isolated by a single pair of input and output terminals. Network $\eta$ may be therefore reduced to a two queue network; an equivalent queue for $\sigma$, and an equivalent for the complement of $\sigma$ in $\eta$. Study of throughput in the two queue network is straightforward, since the normalization constant G(N) is expressed as a single convolution.

## 3. Throughput as a Function of Load

An important characteristic of separable queueing networks that can be determined with the aid of Norton's Theorem is that the throughput, or output rate, of any subnetwork is a nondecreasing function of the load, or the number of customers in the subnetwork. It is first shown that this holds for the equivalent two-queue network.

Theorem 1. Let $\mu(m)$ and $U(n)$ be the processing rates of two locally balanced queues in a closed two queue network, when they contain m and n customers, respectively. Let $\tau(N)$ be the mean throughput of the closed network when it contains N customers. Then if $\mu(n+1)\geq\mu(n)$ and $U(n+1)\geq U(n)$ for all $0<n\leq N$, then $\tau(N+1)\geq\tau(N)$.

The proof follows from the form of $\tau(N)=\frac{G(N-1)}{G(N)}$ . The details are in Appendix A.

The characteristic of nondecreasing throughput with increasing load may now be proven for all locally balanced networks, through the use of Norton's Theorem.

Theorem 2. Let $\tau(n)$ be the mean throughput in any branch of a closed network of locally balanced queues that contains n customers. If the processing rate of each queue is a nondecreasing function of the number of customers at the queue, then for all $n>0$, $\tau(n+1)\geq\tau(n)$.

Proof. The proof is by induction on M, the number of queues in the network. Let $\tau_M(n)$ be the mean throughput in some branch b of a network of M queues that contains n customers. Let $\mu(n)$ be the processing rate of an arbitrary queue when it contains n customers, and let $\mu(n+1)\geq\mu(n)$. For M=1, then $\tau_1(n+1)=p\mu(n+1)\geq p\mu(n)=\tau_1(n)$, where p is the probability that a customer takes branch b in returning to the queue.

Let U(n) be the processing rate of the equivalent queue for an M-1 queue network with respect to branch b. Then by Norton's Theorem and the inductive hypothesis, $U(n+1)=\tau_{M-1}(n+1)\geq\tau_{M-1}(n)=U(n)$. And if the equivalent queue is placed in series with a queue with processing rate $\mu(n+1)\geq\mu(n)$, $1\leq n\leq N$, and there are N customers in this network, then the throughput of the closed two queue network is $\tau_M(N)$. But then $\tau_M(N+1)\geq\tau_M(N)$ by Theorem 1. This completes the proof.

The usual interpretation of a queue meets the conditions of this theorem. For the queue considered by Gordon and Newell [6], a fixed number r of processors are available. If a single customer can be serviced on a processor at rate $\mu$, then $\mu(n)=n\mu$ for $n\leq r$, and $\mu(n)=r\mu$ for $n>r$.

## 4. Throughput with Queue-State Dependent Processing Rates

A queue may have the local balance property if each customer has an exponential processing time, or else if the queueing discipline is processor sharing or preemptive last-come-first-served. The total processing rate of the queue may be any function of the number of customers in the queue. If $\mu^{-1}$ is the mean processing time of one customer, and r(n) is the number of available processors when the queue contains n customers, then $\mu(n)=r(n)\mu$. These processing rates will be called queue-state dependent.

There are several instances in which increased throughput can be obtained by dynamically switching processors from one queue to another. For example, two queues may represent two different processes or programs to be executed in one multiprocessor system. Then any or all of the processors may be available for either queue when they are not required for the other. However, with the developed model of a locally balanced queue--that is, a queue with queue-state dependent processing rates, the analysis of processing power tradeoffs is severely constrained. When the network state changes by means of a customer moving from one queue to another, processing power exchanges may take place only between the two queues involved in the transition.

Throughput with processing power tradeoffs between a single queue and the remainder of the network can be analyzed in the network consisting of the selected queue and the Norton's equivalent queue. But first, the relationship of throughput to processing rates in a two queue network will be established when the rates at each of the queues are independent of each other. This is expressed in the following theorem.

Theorem 3. Let $\mu(m)$ and $U(n)$ be the processing rates at the queues of a locally balanced two queue network, when the queues contain m and n customers, respectively, and suppose all the $U(n)$ and the $\mu(m)$ are mutually independent for $0<m,n\leq N$. Let $\tau(N)$ be the network throughput when it contains N customers. Then

a) $\tau(N)$, as a function of $U(n)$, $0<n\leq N$, has no extrema.

b) $\tau(N)$ is a nondecreasing function of $U(n)$ if $\mu(i)\geq\mu(j)$ for all $N\geq i>N-n$ and $j\leq N-n$.

c) If $\tau(N)$ is a nonincreasing function of $U(n)$, then it is a strictly increasing function of $\mu(N-n)$.

The theorem is proved by differentiating $\tau(N)=\frac{G(N-1)}{G(N)}$ with respect to $U(n)$, for $1\leq n\leq N$. The details are in Appendix B.

The optimum allocation of $R(N)$ available processors to a network that contains N customers, will now be shown. Suppose $r_i(n_i)$ processors are allocated to queue i, when it contains $n_i$ customers. A possible allocation strategy may be to hold some processing power in reserve, that is, $\Sigma r_i(n_i)<R(N)$ for some network state $(n_1 \ldots n_n)$; the reason for doing this may be to ensure that more processing power is available for $r_i(n_i+1)$, or even $r_i(n_i-1)$. But throughput cannot be optimized by such strategies.

Theorem 4. Let $R(N)$ be the processing power available to be allocated to the queues of a closed network of M locally balanced queues when the network contains $N>0$ customers. Let $r_i(n)$ be the processing power allocated queue i, when there are n customers at queue i, for $1 \le i \le M$. Let $\tau_M(N)$ be the mean throughput at some branch of the network.

a) Then for maximum $\tau_M(N)$, the available processing power must always be fully utilized; that is $\sum_{i=1}^{M} r_i(n_i)=R(N)$ for all $\sum_{i=1}^{M} n_i=N$, and $r_i(0)=0$ for $1 \le i \le M$.

b) And any processing power distribution meeting the above constraints provides the maximum throughput, and when $\tau_M(N)$ is maximum, $\tau_M(N)=kR(N)$, where k is a constant.

The proof is again by induction on the number of queues in the network. The M queue network is reduced to a two queue network by Norton's Theorem. Then Theorem 3 is used to show that for maximum throughput, each processing rate of the two queue network must be maximized. The details of the proof are in Appendix C.

From this result, it is seem that for maximum throughput with queue-state dependent rate assignments, the processing power added to one queue must be exactly that taken from another, whenever there is a transition between the queues. But for closed networks with more than two queues, this determines the processing power uniquely.

<u>Theorem 5</u>. Consider any network of $M>2$ locally balanced queues, containing N customers, in which any fraction of the processing power of R(N) processors may be assigned to any queue. If $r_i(n)$

is the processing power assigned queue i when it contains n customers, for $1 \le i \le M$, then throughput is maximized by the assignment $r_i(n)=\frac{n}{N}R(N)$.

Proof. From Theorem 4, for maximum throughput, one must consider processing power allocations such that

$\sum_{i=1}^{M} r_i(n_i)=R(N)$ for each network state $(n_1 \ldots n_n)$. When there are N-k customers at queue one, the processing rate at queue one remains constant and independent of the distribution of the remaining k customers. Therefore, for

$$1<i\le n, \quad r_i(k)=R(N)-r_1(N-k).$$

Similarly, when there are N-k customers at queue two for $2<i\le M$,

$$r_1(k)=r_i(k)=R(N)-r_2(N-k).$$

Hence, $r_i(k)=r_j(k)$ for $1\le i,j\le M$, as long as $M>2$.

Consider the case k=2. With N-2 customers at queue one, the remaining two customers may both be at queue i, or may be at queues i and j, $1<i<j\le M$, while the processing rate at queue one remains constant.

Hence, $r_i(2)=r_i(1)+r_j(1)=2r_i(1)$.

Similarly, considering $k=3,4\ldots N$ it is seen that $r_i(k)=kr_i(1)$.

And, since $r_i(N)=Nr_i(1)$, $r_i(k)=\frac{k}{N}$ for $1\le i\le M$ and $1\le k\le N$.

The proof is complete.

## 5. Throughput with Subnetwork-State Dependent Processing Rates

Queue-state dependency of processing power allocation allows little flexibility in dynamic processor allocation. Furthermore, the results of the previous section apply only if processors may be freely switched to any queue in the network. This limits the applicability of the model to networks of which all the queues are software processors within a single multiprocessor system, in which any processor may execute any logical function; but in such case there is little scheduling difficulty. A more realistic model would limit the capability of processing power tradeoffs to subnetworks, representing processes that can be served by compatible devices located at the same processing center.

But for this case, it can be seen that queue-state dependencies cannot yield an efficient solution. Maximum throughput requires full utilization of the available processors. Since the number of customers in a subnetwork does not remain constant, either processing power must be held in reserve when the number of customers in the subnetwork is less than N, or else there is the possibility of an arrival to the subnetwork when all of the processors are busy. But this latter case precludes queue-state dependent rates. If, for example, the arriving customer joins a queue that is idle, a processor must be taken from another subnetwork queue, even though there is no change in the number of customers at that queue.

As an example of processing power tradeoffs with subnetwork-state dependent processing rates, consider the parallel and series subnetworks, as shown in Figure 2. For either case, when there are m and n customers at queues one and two, respectively, the processing rates are $\mu_1(m,n)$ and $\mu_2(m,n)$, respectively. Each of these two-queue subnetworks will have the local balance property if and only if

$$\frac{\mu_1(m,n-1)}{\mu_1(m,n)} = \frac{\mu_2(m-1,n)}{\mu_2(m,n)} . \tag{5-1}$$

A generalization for more than two queues is straightforward for the parallel case, but not for the series subnetwork case. The state probabilities for either subnetwork are expressed by

$$P(m,n) = \frac{\lambda_1^m}{\prod_{i=1}^{m} \mu_1(i,n)} \cdot \frac{\lambda_2^n}{\prod_{i=1}^{n} \mu_2(0,i)} P(0,0), \tag{5-2}$$

where $\lambda_1 = p_1\lambda$, $\lambda_2 = p_2\lambda$ for the parallel case, and $\lambda_1 = \lambda_2 = \lambda$ for the series case.

According to Norton's Theorem, the processing rate of a subnetwork that contains n customers is the same as the throughput of the subnetwork when it stands alone as a closed network containing n customers. This is because the distribution of customers in the subnetwork is the same in either case. Maximization of throughput in the subnetwork entails maximization of throughput in the two queue network (i.e., the isolated subnetwork), containing n=1,2,...N customers. But these cases are not independent, if the local balance constraint (5-1) is met. The detailed maximization, which uses the theorems of the preceding sections, results in the processing rate assignments

$$r_1(m,n) = \frac{m}{N} R(N), \; r_2(m,n) = \frac{n}{N} R(N),$$

where m and n are the number of customers at queues one and two, respectively, $N=m+n$, and $R(N)$ is the total processing power available to the subnetwork when it contains N customers.[10]

For a realistic model, in which one customer uses one processor, and there is a fixed maximum number R of processors available to the subnetwork, $R(N)=N$ for $N \le R$, and $R(N)=R$ for $N>R$.

The result, that maximum throughput is achieved by sharing the processing power proportionately based on the load at each queue, can be generalized for the case of any fixed number of queues in parallel. This will serve as a local balance model of a multiprocessor system handling parallel queues in a computer network.

# REFERENCES

1. Baskett, F., K. M. Chandy, R. R. Muntz and F. Palacios-Gomez, "Open, Closed and Mixed Networks of Queues with Different Classes of Customers", JACM 22 2(1975), pp. 248-260.

2. Buzen, J. P., "Computational Algorithms for Closed Queueing Networks with Exponential Servers" CACM 16, September 1973.

3. Chandy, K. M., "The Analysis and Solutions for General Queueing Networks", Proc. Sixth Annual Princeton Conf. on Information Sciences and Systems, Princeton Univ., Princeton, N. J., (March 1972), pp. 219-224.

4. Chandy, K. M., U. Herzog, and L. Woo, "Parametric Analysis of Queueing Network Models" IBM Journal of Research and Development, 19,1 (January 1975), pp. 36-42.

5. Chandy, K. M., U. Herzog, and L. Woo, "Approximate Analysis of General Queueing Networks" IBM Journal of Research and Development, 19, 1 (January 1975).

6. Gordon, W. J. and G. F. Newell, "Closed Queueing Systems with Exponential Servers", Oper. Res. 15, 2 (1967), pp. 252-267.

7. Keller, T. W., "ASQ Manual" Dept. of Computer Sciences Report TR-27, University of Texas, Austin, Texas, 1973.

8. Kleinrock, L., Queueing Systems, Vol. 1: Theory. John Wiley and Sons, N. Y., 1975.

9. Noetzel, A. S., "Analysis of Discrete-Time Queues for Random Output Processes and Network Solutions" Proc. 1977 Conf. on Information Sciences and Systems, The John Hopkins University, March 1977.

10. Noetzel, A. S., "Product-Form Queueing Networks with Processing-Rate and Arrival-Rate Tradeoffs" Technical Report 53, Department of Computer Sciences, The University of Texas at Austin, December 1975.

11. Reiser, M., and H. Kobayashi, "Recursive Algorithms for General Queueing Networks with Exponential Servers" IBM Res. Report. RC-4254, March 1973.

12. Reiser, M., "QNET4 User's Guide". IBM Research Report RA71, June 1975.

13. Reiser, M. and H. Kobayashi, "On the Convolution Algorithm for Separable Queueing Networks" Proceedings, 1976 International Symposium on Computer Performance Modeling, Measurement, and Evaluation, Harvard University, Cambridge, Mass.

## Appendix A.

Proof of Theorem 1.

The throughput $\tau(N)$ of the locally balanced two queue network can be expressed as the ratio of the normalization constants with N-1 and N customers in the network. Hence, the theorem is proved if

$$\tau(N+1) = \frac{G(N)}{G(N+1)} \geq \frac{G(N-1)}{G(N)} = \tau(N), \qquad \text{(A-1)}$$

where

$$G(N) = \sum_{i=0}^{n} Z(N-i)X(i), \text{ and } Z(n) = \prod_{i=1}^{n} \frac{1}{\mu(i)} \text{ and } X(n) = \prod_{i=1}^{n} \frac{1}{U(i)}$$

The inequality (A-1) can be written

$$G^2(N) \geq G(N+1)G(N-1) \qquad \text{(A-2)}$$

The terms of $G^2(N)$ contain the factors $X(i)X(j)$ for $0 \leq i,j \leq N$. The terms of $G(N+1)G(N-1)$ contain the factors $X(i)X(j)$ for $0 \leq i \leq N-1$, $0 \leq j \leq N+1$. In both cases $0 \leq i+j \leq 2N$. The inequality is demonstrated by grouping the terms into 2N+1 inequalities. Inequality k will have all the terms with factors $X(i)X(j)$ such that $i+j=k$.

First, consider the case $0 \leq k < N$. Collecting terms from (A-2),

$$\sum_{i=0}^{k} Z(N-i)X(i)Z(N-k+i)X(k-i)$$
$$\geq \sum_{i=0}^{k} Z(N+1-i)X(i)Z(N-1-k+i)X(k-i). \qquad \text{(A-3)}$$

Grouping coefficients of X(i)X(k-i),

$$\sum_{i=0}^{k} [Z(N-i)Z(N-k+i)-Z(N+1-i)Z(N-1-k+i)]X(i)X(k-i) \geq 0. \qquad \text{(A-4)}$$

This sum can be rewritten as two summations; first for index i=0 to $[\frac{k}{2}]$, and then for $i=k-[\frac{k}{2}]+1$ to k. If k is even, these two ranges obviously cover the range 0 to k. If k is odd, the term for $i = [\frac{k}{2}] + 1$ is missing. But the term in the summation for this value of i is zero. Hence, (A-4) can be written as follows, when j=k+1-i replaces i as the index of the second summation.

$$\sum_{i=0}^{[\frac{k}{2}]} [Z(N-i)Z(N-k+i)-Z(N+1-i)Z(N-1-k+i)]X(i)X(k-i)$$

$$+ \sum_{j=1}^{[\frac{k}{2}]} (Z(N-k-1+j)Z(N+1-j)-Z(N-k+j)Z(N-j)]X(k+1-j)X(j-1) \geq 0.$$

$$\text{(A-5)}$$

Note that for each Z(m)Z(n) in (A-5), m+n=2N-k. Also, note that for any m,n with m≥n and any j≤n,

$$Z(m)Z(n)-Z(m+j)Z(n-j)$$

$$=Z(m)Z(n-j) \left[ \prod_{i=n-j+1}^{n} \frac{1}{\mu(i)} - \prod_{i=m+1}^{m+j} \frac{1}{\mu(i)} \right] \geq 0, \qquad \text{(A-6)}$$

since all of the indices i, and hence rates $\mu(i)$ in the second product are greater than those of the first product. Therefore, the products $Z(m)Z(n)$ for all $m+n=2N-k$ are ordered inversely as $|m-n|$, or directly as $\min(m,n)$. If $i=\min(m,n)$ let $\overline{Z}(i)=Z(m)Z(n)$ and let $\overline{X}(i)=X(m)X(n)$. Then the inequality (A-5) can be rewritten by selecting the smaller index of each product.

$$\sum_{i=0}^{[\frac{k}{2}]} (\overline{Z}(N-k+i)-\overline{Z}(N-k+i-1))\overline{X}(i)$$

$$+\sum_{j=1}^{[\frac{k}{2}]} (\overline{Z}(N-k+j-1)-\overline{Z}(N-k+j))\overline{X}(j-1) \geq 0 \qquad \text{(A-7)}$$

Rearranging terms,

$$(\overline{Z}(N-k)-\overline{Z}(N-k-1))\overline{X}(0) + \sum_{i=1}^{[\frac{k}{2}]} [(\overline{Z}(N-k+i)-\overline{Z}(N-k+i-1))\overline{X}(i)$$

$$+ (\overline{Z}(N-k+i-1)-\overline{Z}(N-k+i))\overline{X}(i-1)]\geq 0$$

or

$$(\overline{Z}(N-k)-\overline{Z}(N-k-1))\overline{X}(0) + \sum_{i=1}^{[\frac{k}{2}]} [\overline{Z}(N-k+i)-\overline{Z}(N-k+i-1)][\overline{X}(i)-\overline{X}(i-1)]\geq 0.$$

(A-8)

It is seen that each factor of every term of the summation is nonnegative. Hence, the inequality is demonstrated.

Now consider the case $k=N$. In collecting all terms of (A-2) with factors $X(i)X(j)$ where $i+j=N$, $G(N-1)$ contributes terms with factors $X(i)$ for $0 \leq i \leq N-1$. Hence, $G(N+1)$ contributes terms with factors $X(j)$ for $1 \leq j \leq N$. The inequality corresponding to (A-3) is

$$\sum_{i=0}^{n} Z(N-i)X(i)Z(i)X(N-i)$$

$$\geq \sum_{i=0}^{n-1} Z(i+1)X(N-i)Z(N-1-i)X(i) \qquad \text{(A-9)}$$

Collecting terms, and then adjusting the index of the summation to range from 1 to N, this inequality is written as follows:

$$Z(0)Z(N)X(0)X(N) + \sum_{i=1}^{n} \left[Z(N+1-i)Z(i-1)-Z(i)Z(N-i)\right]X(N-i+1)X(i-1) \geq 0. \qquad \text{(A-10)}$$

The summation can be expressed as two summations, first with index $i=1$ to $\left[\frac{N}{2}\right]$, then with $i=N+1-\left[\frac{N}{2}\right]$ to N, noting that if N is odd, the term for $i = \left[\frac{N}{2}\right] + 1$ disappears. Then rewriting the second summation with index $j=N+1-i$ one obtains

$$Z(0)Z(N)X(0)X(N)+\sum_{i=1}^{[\frac{N}{2}]}[Z(N+1-i)Z(i-1)-Z(i)Z(N-i)]X(N+1-i)X(i-1)$$

$$+\sum_{j=1}^{[\frac{N}{2}]}[Z(j)Z(N-j)-Z(N+1-j)Z(j-1)]X(j)X(N-j)\geq 0 \qquad \text{(A-11)}$$

Then, if $i=\min(m,n)$, let $\overline{Z}(i)=Z(m)Z(n)$, and $\overline{X}(i)=X(m)X(n)$, (A-11) is rewritten

$$\overline{Z}(0)\overline{X}(0)+\sum_{i=1}^{[\frac{N}{2}]}[(\overline{Z}(i-1)-\overline{Z}(i))\overline{X}(i-1)+(\overline{Z}(i)-\overline{Z}(i-1))\overline{X}(i)]\geq 0.$$

Rearranging terms,

$$\overline{Z}(0)\overline{X}(0)+\sum_{i=1}^{[\frac{N}{2}]}[(\overline{Z}(i)-\overline{Z}(i-1)][\overline{X}(i)-\overline{X}(i-1)]\geq 0. \qquad \text{(A-12)}$$

Since each term in the summation is (A-12) positive, the inequality is demonstrated.

Last, the case for $N<k\leq 2N+1$ must be considered. But from the symmetry of Z and X in the definition of the function G, inequality k of A-2 is exactly inequality 2N+1-k with the roles of Z and X exchanged. Hence, it has been demonstrated in the first case. The theorem is proved.

## Appendix B: Proof of Theorem 3

Let $Z(n) = \prod_{i=1}^{n} \frac{1}{U(i)}$ and $X(n) = \prod_{i=1}^{n} \frac{1}{\mu(i)}$ for $0<n\leq N$. Let $G(N) = \sum_{i=0}^{N} Z(i)X(N-i)$ be the normalization constant for the state probabilities of the two queue network with N customers. Then for all $N>0$,

$$\tau(N) = \frac{G(N-1)}{G(N)} . \tag{B-1}$$

For $0<n\leq N$ let $G_n^+(N)$ be all of the terms of $G(N)$ that have the factor $U^{-1}(n)$

$$G_n^+(N) = \sum_{i=n}^{N} Z(i)X(N-i), \tag{B-2}$$

and $G_n^-(N)=G(N)-G_n^+(N)$.

Then, differentiating $G(N)$ with respect to $U(n)$,

$$\frac{\partial G(N)}{\partial U(n)} = -\frac{1}{U(n)}G_n^+(N).$$

and differentiating $\tau(N)$ with respect to $U(n)$.

$$\begin{aligned}\frac{\partial \tau(N)}{\partial U(n)} &= \left[G(N)\frac{\partial G(N-1)}{\partial U(n)} - G(N-1)\frac{\partial G(N)}{\partial U(n)}\right]\frac{1}{G^2(N)}\\ &= \Big[-(G_n^-(N)+G_n^+(N))G_n^+(N-1)U(n)^{-1}\\ &\qquad +G_n^-(N-1)+G_n^+(N-1))G_n^+(N)U(n)^{-1}\Big]\frac{1}{G^2(N)}\\ &= \left[-G_n^-(N)G_n^+(N-1)+G_n^-(N-1)G_n^+(N)\right]\frac{1}{U(n)G^2(N)}\end{aligned} \tag{B-3}$$

The derivative will be nonnegative if

$$G_n^-(N-1)G_n^+(N) \geq G_n^-(N)G_n^+(N-1). \tag{B-4}$$

But each term of this inequality has exactly one factor $U^{-1}(n)$. Hence it may be cancelled out of the inequality. As a function of $U(n)$, $\tau(N)$ is therefore either always increasing, always decreasing, or is constant. This proves part a) of the theorem.

The terms on the right of the inequality (B-4) have factors $Z(k)Z(i)$, $0<k<n$, $n\leq i<N$, and the terms on the left have factors $Z(k)Z(i)$, $0<k<n$, $n\leq i\leq N$. The coefficient of each $Z(k)Z(i)$ that appears on the right is $X(N-k)X(N-1-i)$, and the coefficient of that term on the left is $X(N-1-k)X(N-i)$. Hence if $X(N-1-k)X(N-i)\geq X(N-k)X(N-1-i)$, the inequality holds. But since $i>k$, $X(N-1-i)X(N-1-k)$ can be factored out of this inequality, leaving $\frac{1}{\mu(N-i)} \geq \frac{1}{\mu(N-k)}$. Therefore, if $\mu(N-k)\geq\mu(N-1)$ for all $0<k<n\leq i\leq N$, $\tau(N)$ is a nondecreasing function of $U(n)$. This proves part b) of the theorem.

In particular, it should be noted that if $\mu(i)\geq\mu(j)$ for all $i>j$, which is the usual case, then $\tau(N)$ is a nondecreasing function of $U(n)$, for all n.

Now let $H_m^+(N)$ be the sum of all of the terms of $G(N)$ that have the factor $\mu^{-1}(m)$.

$$H_m^+(N) = \sum_{i=m}^{N} Z(N-i)X(i) \tag{B-5}$$

and

$$H_m^-(N) = G(N)-H_m^+(N).$$

Then, from the definition of $H^+_m(N)$, the following relationships are noted

$$H^+_{N-n}(N) = G^-_{n+1}(N) \tag{B-6}$$

and

$$H^+_{N-n}(N-1) = G^-_n(N-1) \tag{B-7}$$

By the steps leading to (B-4) the condition for $\frac{\partial\tau(N)}{\partial\mu(N-n)} > 0$ is determined to be $H^-_{N-n}(N-1)H^+_{N-n}(N) > H^-_{N-n}(N)H^+_{N-n}(N-1)$. (B-8)

Using (B-6) and (B-7) this can be expressed as

$$G^+_n(N-1)G^-_{n+1}(N) > G^+_{n+1}(N)G^-_n(N-1). \tag{B-9}$$

And then $G^+_{n+1}(N)$ can be related to $G^+_n(N)$,

$$G^+_n(N-1)[G^-_n(N)+Z(N-n)X(n)] > [G^+_{n+1}(N)-Z(N-n)X(n)]G^-_n(N-1), \tag{B-10}$$

which can be written

$$[G^+_n(N-1)G^-_n(N)-G^+_n(N)G^-_n(N-1)] + Z(N-n)X(n)G(N-1) > 0 \tag{B-11}$$

The inequality (B-11) must be satisfied if the term in brackets is nonnegative. But this term expresses the condition (B-4); it will be nonnegative if $\tau(N)$ is a nonincreasing function of $U(n)$. This proves part c) of the theorem.

## Appendix C: Proof of Theorem 4

The proof is by induction on M. First, consider all networks with only one locally balanced queue. A network with only one queue may have several paths from the output of the queue to the input. Let p be the probability that a customer leaving the queue uses a particular branch b in returning to the queue. Let the processing rate at the queue be $\mu$ when a single processor is assigned the queue. If $r(N)$ processors are assigned the queue when the network contains N customers, the throughput in branch b is $\tau_1(N) = pr(N)\mu$. The throughput is maximum when $r(N)=R(N)$. Then $\tau_1(N)=p\mu R(N)$, which satisfies the theorem.

Suppose the theorem holds for all networks of M-1 queues. Let $\tau_{M-1}(n)$ be the throughput at some branch b of the M-1 queue network when there are n customers in the M-1 queue network.

Then by Norton's Theorem, the M-1 queue network may be represented by an equivalent queue with respect to branch b. If $U(n)$ is the processing rate of the equivalent queue when the queue contains n customers, then $U(n)=\tau_{M-1}(N)$. By the inductive hypothesis, if processing power $R_1(n)$ is available to the M-1 queue network, the maximum throughput at branch b is $\tau_{M-1}(n)=UR_1(n)$, where U is a constant, and is achieved when processing power $R_1(n)$ is maximally utilized. This is also the maximum processing rate of the equivalent queue when it contains n customers and has available processing power $R_1(n)$.

Let $\tau_M(n)$ be the throughput in the two queue network consisting of the equivalent queue in series with queue M. Then $\tau_M(n)$ is equal to the throughput in branch b of the M-1 queue network with queue M inserted in branch b. Suppose the processing rate at queue M, when it contains n customers and has processing power $r(n)$, is $\mu(n)=r(n)\mu$. The throughput of the two queue network is determined as follows.

Let $X(n) = \prod_{i=1}^{n} \frac{1}{\mu(i)}$ and let $Z(n) = \prod_{i=1}^{n} \frac{1}{U(i)}$.

Let $G(N) = \sum_{i=1}^{N} X(i)Z(N-i)$. (C-1)

Then $\tau_M(N) = \frac{G(N-1)}{G(N)}$ (C-2)

Let R(N) be the processing power available to the M queue network when it contains N customers. And let $\rho(n) = \frac{r(n)}{R(N)}$ be the optimum fraction of the available processing power to be used by queue M when it contains $n<N$ customers, in order to maximize $\tau_M(N)$. Then $\mu(n)=\rho(n)R(N)\mu$ are the processing rates at queue M that maximize $\tau_M(n)$.

Let $\overline{\rho}(n)=1-\rho(n)$. Then, for maximum $\tau_M(N)$, processing power $\overline{\rho}(n)R(N)$ is available to be allocated to the equivalent queue when it contains N-n customers.

Examining (C-2) shows that for maximum $\tau_M(N)$, U(N) and $\mu(N)$ are to be maximized. Clearly, all available processing power is used for these rates, so that $\rho(N)=\overline{\rho}(0)=1$.

Note that $\rho(n)>0$ for all $n>0$ may be assumed. For if this is not the case, let m be the largest integer for which $\rho(m)=0$. Then at least m customers will always be at queue M. Let $N'=N-m$ and $\mu'(n)=\mu(n-m)$. Maximization of $\tau_M(N)$ is accomplished in this case by considering only the rates $U(N'-n)$ and $\mu'(n)$, for $n\leq N'$. The result will be the same as maximization with $\rho(n)>0$ for $n\leq N$, if it is shown that processing power is fully utilized when there are m customers in queue M. But note that $\rho(m)=0$ only if $\tau_M(N)$ is a non-increasing function of $\mu'(0)$. By Theorem 3C, then $\tau_M(m)$ is an increasing function of $U(N')$. Therefore, $\overline{\rho}(0)=1$.

With the rates $\rho(n)R(N)$ for queue M fixed at the values required for maximum $\tau_M(n)$, the rates $U(n)$ may be selected within the range $0 \le U(n) \le U\overline{\rho}(N-n)R(N)$ to maximize $\tau_M(N)$.

Suppose $\tau_M(N)$ is not an increasing function of $U(n)$, for some $n<N$. Then, by Theorem 3C, it must be an increasing function of $\mu(N-n)$. Then $\rho(N-m)=1$, and therefore $U(m)=0$. If m is the largest integer for which $\tau_M(N)$ is not an increasing function of $U(m)$, then there will never be less than m customers at the equivalent queue. Hence, letting $N'=N-m$ and $U'(n)=U(n-m)$, only rates $U'(n)$ for $0 \le n \le N'$ must be considered in maximizing $\tau_M(N)$. And this maximization will yield the same result as maximization with $\tau_M(n)$ an increasing function of $U(n)$ for all $n>0$. Therefore, the maximum value $U(n)=U\overline{\rho}(N-n)R(N)$ must be chosen for $U(n)$, $n \le N$ in order to maximize $\tau_M(n)$. This proves part a) of the theorem. Then each term $X(j)Z(k)$ of $G(N-1)$, where $j+k=N-1$, can be written

$$X(j)Z(k) = \left[\mu^j U^k R^{j+k}(N) \prod_{i=1}^{j} \rho(i) \prod_{i=1}^{k} \rho(N-i)\right]^{-1}. \qquad \text{(C-3)}$$

The terms of $G(N)$ have the same form, but $j+k=N$. Each term $X(j)Z(k)$ of $G(N)$ with $j,k>0$ contains the product

$$\frac{1}{\rho(j)} \cdot \frac{1}{\bar{\rho}(j)} = \frac{1}{\bar{\rho}(j)} + \frac{1}{\rho(j)} ,$$

Hence, it can be written

$$X(j)Z(k) = \frac{1}{\mu R(N)} \left[ \mu^{j-1} U^{k} R^{N-1}(N) \prod_{i=1}^{j-1} \rho(i) \prod_{i=1}^{k} \bar{\rho}(N-i) \right]^{-1}$$

$$+ \frac{1}{UR(N)} \left[ \mu^{j} U^{k-1} R^{N-1}(N) \prod_{i=1}^{j} \rho(i) \prod_{i=1}^{k-1} \bar{\rho}(N-i) \right]^{-1}$$

$$= \frac{1}{\mu R(N)} X(j-1)Z(k) + \frac{1}{UR(N)} X(j)Z(k-1). \qquad \text{(C-4)}$$

G(N) also includes the terms

$$X(N)Z(0) = \left[ \mu^{N} R^{N}(N) \prod_{i=1}^{N-1} \rho(i) \right]^{-1} = \frac{1}{\mu R(N)} X(N-1)Z(0) \qquad \text{(C-5a)}$$

and

$$X(0)Z(N) = \left[ U^{N} R^{N}(N) \prod_{i=1}^{N-1} \bar{\rho}(N-i) \right]^{-1} = \frac{1}{UR(N)} X(0)Z(N-1). \qquad \text{(C-5b)}$$

G(N) may then be expressed as follows:

$$G(N) = \sum_{j+k=N} X(j)Z(k)$$

$$= X(N)Z(0) + \sum_{\substack{j+k=N \\ j,k>0}} X(j)Z(k) + X(0)Z(N)$$

$$= \frac{1}{\mu R(N)} X(N-1)Z(0) + \sum_{\substack{j+k=N \\ j,k>0}} \frac{1}{\mu R(N)} X(j-1)Z(k)$$

$$+ \frac{1}{UR(N)} X(0)Z(N) + \sum_{\substack{j+k=N \\ j,k>0}} \frac{1}{UR(N)} X(j)Z(k-1)$$

$$= \frac{1}{\mu R(N)} \sum_{j+k=N-1} X(j)Z(k) + \frac{1}{UR(N)} \sum_{j+k=N-1} X(j)Z(k)$$

$$= \left(\frac{1}{\mu} + \frac{1}{U}\right) \frac{1}{R(N)} \; G(N-1). \tag{C-6}$$

Therefore, when $\tau_M(N)$ is maximum,

$$\tau_M(N) = \frac{G(N-1)}{G(N)} = \left(\frac{1}{\mu} + \frac{1}{U}\right)^{-1} R(N). \tag{C-7}$$

Then part b) of the theorem is proved.

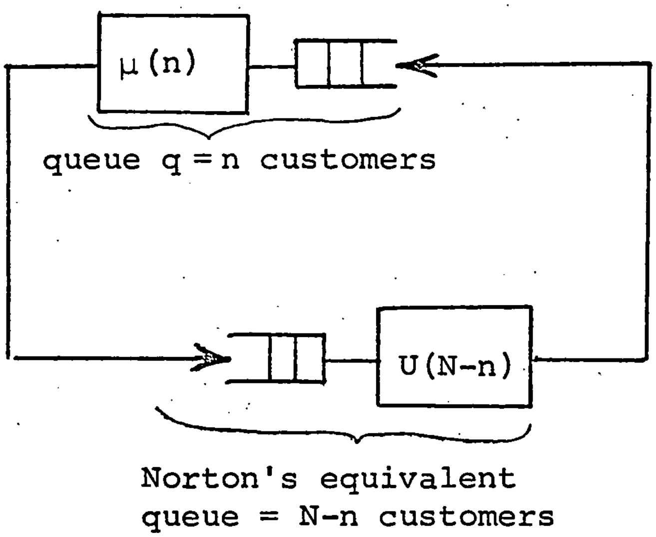

Figure 1

Two Queue Network

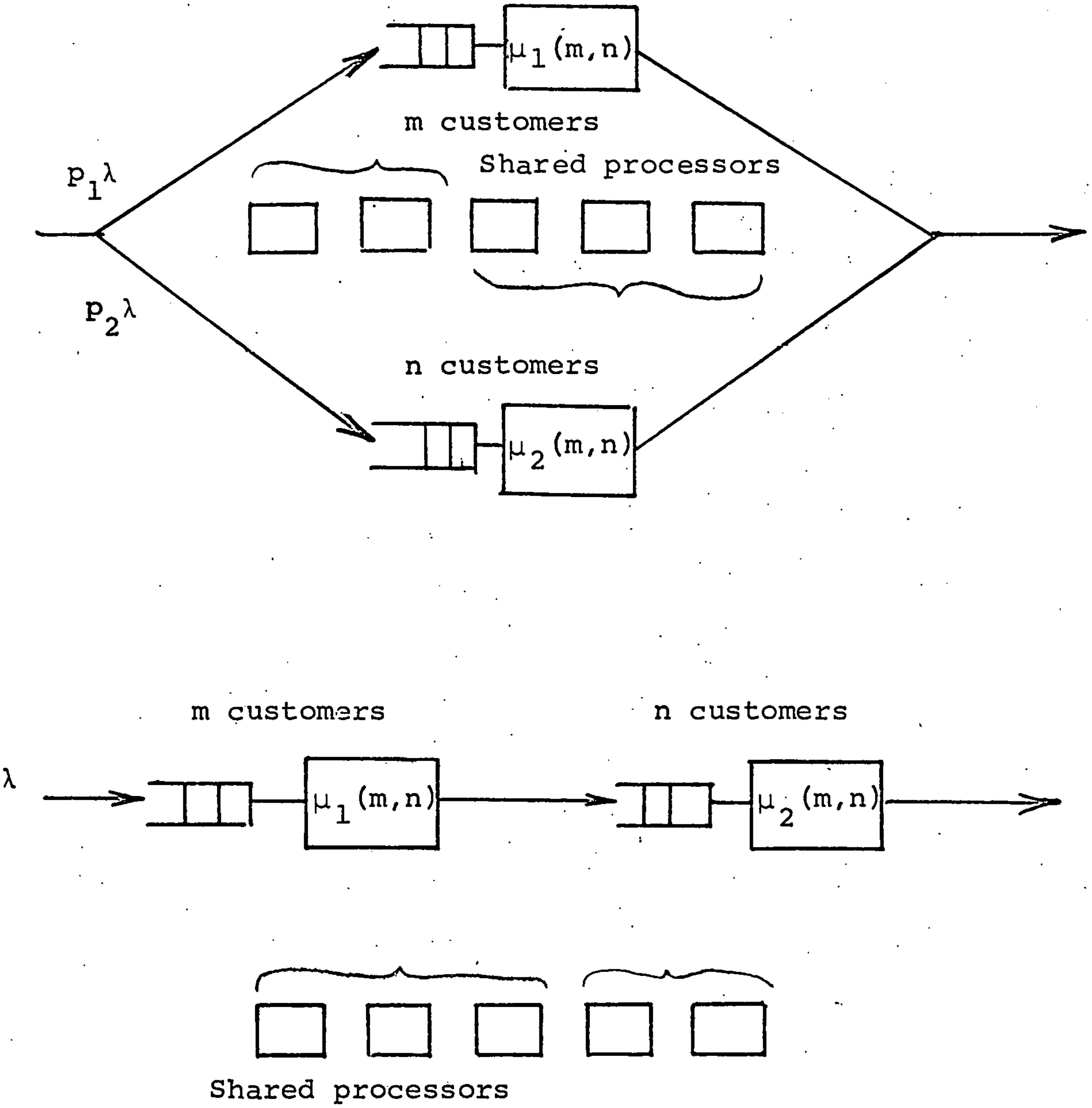

Figure 2

Parallel and series networks
with processing power tradeoffs