Product Form Networks of Queues with Arrival and Processing Rate Tradeoffs*

A. S. Noetzel

December, 1975 TR-53

Technical Report No. 53
Department of Computer Sciences
The University of Texas at Austin
Austin, Texas 78712

* The research reported here was partially supported by National Science Foundation Grant No. GJ-39658.

Abstract

In computer systems or networks of computers, processors or quantities of processing power are often shared among the processes within the system, on the basis of customer demand for those processes. Also, customers are often directed toward one of several processing stations, depending on the load at the stations. Here, both of these system features are modeled by queueing networks that have the local balance property, or whose state probabilities can be expressed as the product of state probabilities of the individual queues. First, the conditions for maximum throughput are derived under the constraint that the processing power allocated any queue is a function of the state of that queue only. Then the conditions are derived under which series and parallel subnetworks of queues may have the local balance property, if the processing power allocated each queue is a function of the state of the subnetwork local to the queue. Then processing power allocations of this form that yield maximum throughput are derived.

The results show that throughput in locally balanced networks or subnetworks is maximized only if all available processing power is always utilized to serve the customers at queues, and that any processing power allocation within this constraint provides maximum throughput. Ancillary theorems relating processing power to throughput in locally balanced networks are proven. It is shown that if the processing rate at each queue is a nondecreasing function of the number of customers in the queue, then the throughput of the network is a nondecreasing function of the number of customers in the network. Conditions for which the local balance property holds in networks with load-dependent branching probabilities are also derived.

1. Introduction

Queueing networks that have the product form solution are easily analyzed, whereas queueing networks without such solutions are generally intractable. It is therefore of considerable interest to determine whether queues or queueing networks with features representing certain realistic constraints yield the product-form solution. For example, one feature that has been found to be amenable to product-form solution is the queue-state dependency of service rates at queues. Gordon and Newell's classical paper (G1) considered queues that are served by a finite number of processors. The effective processing rate of each queue increases linearly with the number of customers at the queue until all processors are utilized, and then remains constant. More recently, it has been shown that the local balance (C1) or Poisson departure (M2) conditions, which give rise to the product-form solution, are possible for queues at which the total service rate is a general function of the number of customers at the queue. (C2, N1)

In this work, we first investigate the solution of networks of queues in which the service rate at each queue is a function of the state of a subnetwork local to the queue. These processing rates will be termed subnetwork-state dependent rates. They are characterized by the possibility of a change in the processing rate at one queue due to a change in the number of customers at another queue. In practice, such rate assignments occur when there are processing power tradeoffs between queues.

Processing power tradeoffs occur in several contexts within computing system and network design. For example, consider a multiprocessor system dedicated to several different programs or process types. The system may be represented as several different queues with independent processing time distributions. The queues might either be in series, representing serial execution of programs, or they might be in parallel with independent arrival streams. A processor allocation scheme without tradeoffs would assign a fixed number of processors to each queue. But this technique is inefficient, since processors of one queue may be idle while customers are waiting for service at other queues. An efficient algorithm will allocate an available processor to any queue that has a customer waiting. But the processor allocation at one queue must then change in response to the load at another queue. Hence, processing rates that depend on the state of all the queues with which processor exchanges may take place must be considered.

A second topic to be investigated is the analogous case in which the branching probabilities within the network are functions of the state of a subnetwork local to the branch point. Such branching probabilities, which will be termed state dependent branching probabilities, will allow the mean arrival rate of customers to one subnetwork to vary with the number of customers in a parallel subnetwork. In practice, the arrival rates to subnetworks are traded off on a load dependent basis. If two queues or subnetworks of queues are capable of handling the processing requirements of a customer, an efficient scheduling algorithm will direct that customer to the queue or subnetwork with the smallest load.

Sections 2 and 3 of this paper contain general theorems concerning the throughput of locally balanced queueing networks. Norton's theorem for queueing networks (C3) shows that it is possible to represent the throughput in any branch of any closed locally balanced queueing network as the throughput of an equivalent two queue network. It is important therefore to understand the throughput characteristics of the two queue network. In Section 2, the relationship between the processing rates at the queues and throughput is explored. In Section 3, it is shown that if the processing rate at every queue of an arbitrary closed network is a nondecreasing function of the number of customers at the queue, then the throughput of each subnetwork is a nondecreasing function of the number of customers in the subnetwork.

In Section 4, throughput in a closed queueing network with processing power tradeoffs and queue state dependent processing rates is considered. It is shown that any processor allocation algorithm that maximizes throughput must fully utilize all processors at all times. Furthermore, in all networks of more than two queues, the constraint of queue state dependent processing rates determines uniquely the processor allocation algorithm for maximum throughput.

In Sections 5 and 6, the constraints necessary for the product-form solution for processing rate tradeoffs in parallel subnetworks and in two queue series subnetworks, respectively, are derived. For two queue subnetworks, the constraint is seen to be the same in both the series and parallel cases.

In Section 8, the maximization of throughput in two queue subnetworks with processing power tradeoffs is demonstrated. It is shown that if the maxi-

mum processing power available to the subnetwork is always utilized in processing the customers in the subnetwork, the maximum throughput is achieved. Furthermore, the maximum throughput is independent of the number of customers in the subnetwork, or the actual distribution of processing power to the queues.

In Section 9, the case of state dependent branching probabilities is considered. The conditions for a subnetwork with state dependent branching probabilities to have a product form solution is derived. It is seen to be strongly analogous to the constraint for load dependent processing rates.

2. Throughput Related to Processing Rates

Norton's theorem for queueing networks makes it possible to obtain an equivalent two queue representation of any closed queuing network with locally balanced queues. The two queues of the equivalent network will be locally balanced. They will represent the queues within a particular branch of the network, and the queues outside the branch, respectively. The two queue representation is useful for analyzing the throughput in the selected branch of the original network. This will be done in the latter sections of this paper. But first, it will be useful to determine some general properties of the throughput in a two queue network. First, the relationship between the throughput and the individual processing rates at the queues, assumed independent of each other, will be described in the following theorem.

Theorem 2.1

Let $U(m)$ and $\mu(n)$ 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 the $U(n)$ and the $\mu(m)$ are independent of each other. for $0<m,n\leq N$. Let $\tau(N)$ be the throughput in some branch of the network when the network 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 $\mu(N-n)$.

Proof

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 $N>0$,

$$\tau(N) = \frac{G(N-1)}{G(N)} \quad . \tag{2-1}$$

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

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

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 $U(n)$.

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

The derivative will be nonnegative if

$$G_n^-(N-1) G_n^+(N) \geq G_n^-(N) G_n^+(N-1). \tag{2-4}$$

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 (2-4) 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 $\mu(N-k) \geq \mu(N-i)$ 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 $\mu^{-1}(m)$.

$$H_m^+(N) = \sum_{i=m}^{N} Z(N-i)X(i) \tag{2-5}$$

and

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

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

$$H_{N-n}^+(N) = G_{n+1}^-(N) \tag{2-6}$$

and

$$H_{N-n}^+(N-1) = G_n^-(N-1) \tag{2-7}$$

By the steps leading to (2-4) the condition for $\frac{\partial\tau(N)}{\partial\mu} > 0$ is determined to be

$$H_{N-n}^-(N-1)H_{N-n}^+(N) > H_{N-n}^-(N)H_{N-n}^+(N-1). \tag{2-8}$$

(2-9)

Using (2-6) and (2-7) this can be expressed as $G_n^+(N-1)G_{n+1}^-(N) > G_{n+1}^+(N)G_n^-(N-1)$. And then $G_{n+1}^+(N)$ can be related to $G_n^+(N)$,

$$G_n^+(N-1)[G_n^-(N)+Z(N-n)X(n)] > [G_{n+1}^+(N)-Z(N-n)X(n)]G_n^-(N-1), \tag{2-10}$$

which can be written

$$[G_n^+(N-1)G_n^-(N)-G_n^+(N)G_n^-(N-1)]+Z(N-n)X(n)G(N-1) > 0. \tag{2-11}$$

The inequality (2-11) must be satisfied if the term in brackets is nonnegative. But this term expresses the condition (2-4); it will be nonnegative if $\tau(N)$ is a nonincreasing function of U(n). This proves part c) of the theorem.

3. The Increase in Throughput with Load

If R(n) is the number of processors available to a queue or subnetwork when the queue or subnetwork contains n customers, R(n) will usually be a non-decreasing function of n. For example, if there are k processors at a queue, then R(n)=n for $n \leq k$ and R(n)=k for $n \leq k$. The processing rate at a queue is proportional to the number of processors in use, hence it is also a nondecreasing function of n. For the theorems that follow, it must be shown that the effect of nondecreasing throughput as load is increased holds for networks as well as queues. This is expressed in the following theorem.

Theorem 3.1

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

Proof

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

$$\tau(N+1) = \frac{G(N)}{G(N+1)} \geq \frac{G(N-1)}{G(N)} = \tau(N), \qquad (3\text{-}1)$$

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 (3-1) can be written

$$G^2(N) \geq G(N+1)G(N-1) \tag{3-2}$$

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 (3-2),

$$\sum_{i=0}^{k} Z(N-i)X(i)Z(N-k+i)X(k-i) \tag{3-3}$$

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

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

$$\sum_{i=0}^{k} [Z(N-i)Z(N-k+i)-Z(N+1-i)Z(N-1-k+i)]X(i)X(k-i) \geq 0. \tag{3-4}$$

This sum can be rewritten as two summations; first for index i=0 to $\left[\frac{k}{2}\right]$, and then for $i=k-\left[\frac{k}{2}\right]+1$ to k. If k is even, these two ranges obviously cover the range 0 to k. If k is odd, the term for $i=\left[\frac{k}{2}\right]+1$ is missing. But the term in the summation for this value of i is zero. Hence, (3-4) can be written as follows, when $j=k+1-i$ replaces i as the index of the second summation.

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

$$+ \sum_{j=1}^{\left[\frac{k}{2}\right]} [Z(N-k-1+j)Z(N+1-j)-Z(N-k+j)Z(N-j)]X(k+1-j)X(j-1) \geq 0. \tag{3-5}$$

Note that for each Z(m)Z(n) in (3-5), m+n=ZN-k. Also,

Note that for any m,n with m≥n and any j≤n,

$$Z(m)Z(n)-Z(m+j)Z(n-j)$$
$$=Z(m)Z(n-j)\left[\prod_{i=n-j+1}^{n}\frac{1}{\mu(i)}-\prod_{i=m+1}^{m+j}\frac{1}{\mu(i)}\right]\geq 0, \qquad (3\text{-}6)$$

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 $\bar{Z}(i)=Z(m)Z(n)$ and let $\bar{X}(i)=X(m)X(n)$. Then the inequality (3-5) can be rewritten by selecting the smaller index of each product.

$$\sum_{i=0}^{\left\lceil\frac{k}{2}\right\rceil}(\bar{Z}(N-k+i)-\bar{Z}(N-k+i-1))\bar{X}(i)$$
$$+\sum_{j=1}^{\left\lceil\frac{k}{2}\right\rceil}(\bar{Z}(N-k+j-1)-\bar{Z}(N-k+j))\bar{X}(j-1)\geq 0 \qquad (3\text{-}7)$$

Rearranging terms,

$$(\bar{Z}(N-k)-\bar{Z}(N-k-1))\bar{X}(0)+\sum_{i=1}^{\left\lceil\frac{k}{2}\right\rceil}[(\bar{Z}(N-k+i)-\bar{Z}(N-k+i-1))\bar{X}(i)$$
$$+(\bar{Z}(N-k+i-1)-\bar{Z}(N-k+i))\bar{X}(i-1)]\geq 0$$

or,

$$(\bar{Z}(N-k)-\bar{Z}(N-k-1))\bar{X}(0)+\sum_{i=1}^{\left\lceil\frac{k}{2}\right\rceil}[\bar{Z}(N-k+i)-\bar{Z}(N-k+i-1)][\bar{X}(i)-\bar{X}(i-1)]\geq 0. \qquad (3\text{-}8)$$

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 (3-2) 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 (3-3) is

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

$$\geq \sum_{i=0}^{N-1} Z(i+1)X(N-i)Z(N-1-i)X(i) \qquad (3\text{-}9)$$

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

$$Z(0)Z(N)X(0)X(N) + \sum_{i=1}^{N} \left[Z(N+1-i)Z(i-1)-Z(i)Z(N-i)\right] X(N-i+1)X(i-1) \geq 0. \qquad (3\text{-}10)$$

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

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

$$+ \sum_{i=1}^{\left\lceil \frac{N}{2} \right\rceil} \left[Z(j)Z(N-j)-Z(N+1-j)Z(j-1)Z(j-1)\right] X(j)X(N-j) > 0 \qquad (3\text{-}11)$$

Then, if i=min(m,n), let $\bar{Z}(i)=Z(m)Z(n)$, and $\bar{X}(i)=X(m)X(n)$, (3-11) is rewritten

$$\bar{Z}(0)\bar{X}(0) + \sum_{i=1}^{\left\lceil \frac{N}{2} \right\rceil} \left(Z(i-1)-Z(k)\right)X(i-1)+\left(Z(i)-Z(i-1)\right)X(i)\big] \geq 0.$$

Rearranging terms,

$$\bar{Z}(0)\bar{X}(0) + \sum_{i=1}^{\left[\frac{N}{2}\right]} [(\bar{Z}(i)-\bar{Z}(i-1)]\,[\,\bar{X}(i)-\bar{X}(i-1)] \geq 0. \qquad (3\text{-}12)$$

Since each term in the summation is (3-12) 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 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.

Theorem 3.2

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 power available to every queue is always fully utilized, then for all $n>0$, $\tau(n+1) \geq \tau(n)$.

Proof

The proof is by induction 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 μ be the mean processing rate of a single queue when one processor is used. If there are R processors available to the queue, then $R(n)=n$ for $n \leq R$ and $R(n)=R$ for $n>R$. If $\mu(n)$ is the processing rate at the queue when it contains n customers, then $\mu(n)=R(n)\mu$, and $\mu(n+1) \geq \mu(n)$. 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 $\mu(n+1) \geq \mu(n)$, 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 3.1. The theorem is proved.

4. Processing Power Tradeoffs in Networks with Queue-State Dependent Processing Rates

Gordon and Newell's work (G1) has been generalized (C2, B1, N1) to show that product-form state probabilities are possible for networks in which the processing rate at each queue is a general function of the number of customers in the queue. But in queueing networks of this form, processing rates at each queue must remain constant while the number of customers in the queue remains constant. Such assignments of processing power to queues will be termed queue-state dependent. The opportunities for effectively applying processing power that may be switched among the queues of such a network are limited, because 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. Here the maximization of throughput via processing power tradeoffs in networks with queue-state dependent processing rates is shown.

Theorem 4.1

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<i<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 $1 \leq i \leq M$.

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.

Proof

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 queueto 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 μ 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 $G(N) = \sum_{i=1}^{N} X(i)Z(N-i)$. (4-1)

Then $\tau_M(N) = \frac{G(N-1)}{G(N)}$ (4-2)

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

Examining (4-2) 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 $\rho(N)=\bar{\rho}(0)=1$.

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 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 nonincreasing function of $\mu'(0)$. By Theorem 2.1c, then $\tau_M(m)$ is an increasing function of $U(N')$. Therefore, $\bar{\rho}(0)=1$.

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 $0\leq U(n)\leq U\bar{\rho}(N-n)R(N)$ to maximize $\tau_M(N)$.

Suppose $\tau_M(N)$ is not an increasing function of $U(n)$, for some $n<N$. Then, by Theorem 2.1c, 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 $0\leq n\leq N'$ 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 $U(n)=U\bar{\rho}(N-n)R(N)$ must be chosen for $U(n)$, $n\leq N$ 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

$$X(j)Z(k)=\left[\mu^{j}U^{k}R^{j+k}(N)\prod_{i=1}^{j}\rho(i)\prod_{i=1}^{k}\rho(N-i)\right]^{-1}. \qquad (4\text{-}3)$$

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

$$X(j)Z(k)=\frac{1}{\mu R(N)}\left[\mu^{j-1}U^{k}R^{N-1}(N)\prod_{i=1}^{j-1}\rho(i)\prod_{i=1}^{k}\bar{\rho}(N-i)\right]^{-1}$$
$$+\frac{1}{UR(N)}\left[\mu^{j}U^{k-1}R^{N-1}(N)\prod_{i=1}^{j}\rho(i)\prod_{i=1}^{k-1}\bar{\rho}(N-i)\right]^{-1}$$
$$=\frac{1}{\mu R(N)}X(j-1)Z(k)+\frac{1}{UR(N)}X(j)Z(k-1). \qquad (4\text{-}4)$$

G(N) also includes the terms

$$X(N)Z(0)=\left[\mu^{N}R^{N}(N)\prod_{i=1}^{N-1}\rho(i)\right]^{-1}=\frac{1}{\mu R(N)}X(N-1)Z(0) \qquad (4\text{-}5a)$$

and

$$X(0)Z(N)=\left[U^{N}R^{N}(N)\prod_{i=1}^{N-1}\bar{\rho}(N-i)\right]^{-1}=\frac{1}{UR(N)}X(0)Z(N-1). \qquad (4\text{-}5b)$$

G(N) may then be expressed as follows:

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

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

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

$$= \left(\frac{1}{\mu} + \frac{1}{U}\right) \frac{1}{R(N)} G(N-1). \qquad (4\text{-}6)$$

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

$$\tau_M(N) = \frac{G(N-1)}{G(N)} = \left(\frac{1}{\mu} + \frac{1}{U}\right)^{-1} R(N). \qquad (4\text{-}7)$$

Then part b) of the theorem is proved.

Since maximum throughput is achieved when all processing power is fully utilized, strategies for optimizing throughput by holding processing power in reserve are excluded. For queue state dependent processing rate assignments, not only are processing power changes restricted to queues i and j when a customer transits from queue i to queue j, but the increment added to one queue must be that subtracted from the other. It is now shown that for closed networks with more than two queues, queue state dependent processing rates determine uniquely the processing power distribution for maximum throughput.

Theorem 4.2

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 $1 \leq i \leq M$, then throughput is maximized by the assignment $r_i(n) = \frac{n}{N} R(N)$.

Proof

From Theorem 4.1, 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

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

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

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

Hence,

$r_i(k)=r_j(k)$ for $1\leq i,j\leq M$, 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, $1<i<j\leq M$, 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..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 $1\leq i\leq M$ and $1\leq k\leq N$.

The proof is complete.

5. Processing Power Tradeoffs in Parallel Subnetworks

It is generally not the case that the processors available to a network of queues can be allocated to the queues in an unconstrained manner. The various queues may have processing requirements that are within the capabilities of only some of the available processors, or else if several queues are separated from the others by large distances or hardware constraints, they may be served only by processors within their locality.

Therefore, processing power tradeoffs within subnetworks will be considered. The queue-state dependencies considered in the previous section do not yield an efficient solution. Maximum throughput requires full utilization of the available processors. Since the number of customers in the subnetwork does not remain constant, either processing power must be held in reserve when the number of customers in the subnetwork is less than the maximum, or else there is the possibility of an arrival to the subnetwork when all of the processors are busy. The latter case precludes queue-state dependent processing rates.

Therefore, processing rates at a queue that are functions of the state of a subnetwork local to the queue, called subnetwork-state dependent processing rates, will be considered. In particular, the constraints on the subnetwork dependent rates that allow the local balance property of the subnetwork, will be determined for particular subnetwork types. First to be considered are parallel subnetworks.

Theorem 5.1

Let $\mu_1(m,n)$ and $\mu_2(m,n)$ be the processing rates of two queues connected in parallel, when the queues contain m and n customers, respectively. Then the subnetwork consisting of the two parallel queues will have the local balance property only if, for $m,n>0$.

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

Proof

Let $\mu_1(m,n)$ and $\mu_2(m,n)$ be the processing rates at parallel queues one and two, respectively, when there are m customers at queue one and n customers at queue two. Let $P_{m,n}$ be the probability of this state. Let λ_1 and λ_2 be the mean rates of the input to each queue. If the combined output of these queues is then to be a Poisson process when the inputs are Poisson processes, the balance and local balance equations for the subnetwork must be satisfied. The balance equations, with the local balance condition satisfied, will be called the departure independence equation. For all $m,n\geq 0$ they are

$$P_{m,n}(\lambda_1+\lambda_2)=P_{m+1,n}\mu_1(m+1,n)+P_{m,n+1}\mu_2(m,n+1) \qquad (5\text{-}1)$$

The local balance equations are,

for $m>0$, $\quad P_{m,0}\mu_1(m,0)=P_{m-1,0}\lambda_1,$ (5-2a)

for $n>0$, $\quad P_{0,n}\mu_2(0,n)=P_{0,n-1}\lambda_2,$ (5-2b)

and for $m,n>0$ $\quad P_{m,n}(\mu_1(m,n)+\mu_2(m,n))=P_{m-1,n}\lambda_1+P_{m,n-1}\lambda_2.$ (5-2c)

Recursive solutions to equations (5-2a) and (5-2b) are

$$P_{m,0} = \frac{\lambda_1}{\mu_1(m,0)} P_{m-1,0} = \frac{\lambda_1^n}{\prod_{i=1}^{m} \mu_1(i,0)} P_{0,0} \qquad (5\text{-}3a)$$

and

$$P_{0,n} = \frac{\lambda_2}{\mu_2(0,n)} P_{0,n-1} = \frac{\lambda_2^n}{\prod_{i=1}^{n} \mu_2(0,i)} P_{0,0}. \qquad (5\text{-}3b)$$

Using the value for $P_{2,0}$ given by (5-3a) in the departure independence equation (5-1) for $P_{1,0}$, one obtains

$$P_{1,0}\lambda_2 = P_{1,1}\mu_2(1,1). \qquad (5\text{-}4a)$$

Hence,

$$P_{1,1} = \frac{\lambda_2}{\mu_2(1,1)} P_{1,0} = \frac{\lambda_2}{\mu_2(1,1)} \frac{\lambda_1}{\mu_2(0,1)} P_{0,0}. \qquad (5\text{-}4b)$$

Equating (5-4a) and (5-4b), the following constraint is determined:

$$\frac{\mu_1(1,0)}{\mu_1(1,1)} = \frac{\mu_2(0,1)}{\mu_2(1,1)}. \qquad (5\text{-}5)$$

The general solution to (5-1) and (5-2) is

$$P_{m,n} = \frac{\lambda_2}{\mu_2(m,n)} P_{m,n-1} = \frac{\lambda_1}{\mu_1(m,n)} P_{m-1,n} \quad \text{for } m,n>0. \qquad (5\text{-}6)$$

This is demonstrated by induction on $K(m,n) = \frac{(m+n+1)(m+n)}{2} + n$. The basis and the inductive step for $m=0$ or $n=0$ are given by (5-3). Assume the solution holds for all $P_{i,j}$ such that $K(i,j)<K(m,n)$. In particular, $K(m+1,n-1)<K(m,n)$, hence,

$$P_{m+1,n-1} = \frac{\lambda_1}{\mu_1(m+1,n-1)} P_{m,n-1}. \qquad (5\text{-}7)$$

Equation (5-1), written for $P_{m,n-1}$ is

$$P_{m,n-1}(\lambda_1+\lambda_2)=P_{m+1,n-1}\mu_1(m+1,n-1)+P_{m,n}\mu_2(m,n). \qquad (5\text{-}8)$$

Substituting (5-7) into (5-8) yields

$$P_{m,n-1}\lambda_2=P_{m,n}\mu_2(m,n), \qquad (5\text{-}9)$$

which provides the first equality of (5-6) for K(m,n).

The second equality may be shown by a similar induction on $K'(m,n)=\frac{(m+n+1)(m+n)}{2}+m$.

By (5-6) it is seen that

$$P_{m,n}=\frac{\lambda_2}{\mu_2(m,n)}\ \frac{\lambda_1}{\mu_1(m,n-1)}\ P_{m-1,n-1}=\frac{\lambda_1}{\mu_1(m,n)}\ \frac{\lambda_2}{\mu_2(m-1,n)}\ P_{m-1,n-1} \qquad (5\text{-}10)$$

and hence

$$\frac{\mu_2(m-1,n)}{\mu_2(m,n)}=\frac{\mu_1(m,n-1)}{\mu_1(m,n)} \qquad \text{for } m,n>0. \qquad (5\text{-}11)$$

Since this relation is derived directly from the local balance equations, it is a necessary condition for the parallel queue subnetwork to have the local balance property. The theorem is proved.

The constraint of Theorem 5.1 is also a sufficient condition for a set of subnetwork state dependent rates to afford the subnetwork the local balance property.

Theorem 5.2

Any subnetwork state dependent processing rates of a two queue parallel subnetwork that satisfy the constraints

$$\frac{\mu_1(m,n-1)}{\mu_1(m,n)} = \frac{\mu_2(m-1,n)}{\mu_2(m,n)} \quad \text{for all } m,n>0,$$

afford the subnetwork the local balance property.

Proof

The proof is by induction N=m+n. It is shown that any selection of $\mu_1(m,n)$ for $1<m<N$ satisfies (5-1) and (5-2), as long $P_{m-1,n}$ and $P_{m,n-1}$ satisfy (5-12), and $\mu_2(m,n)$ is determined by the constraint. The details of the proof are omitted.

A complete expression for the state probability is obtained from (5-6) and (5-3)

$$P(m,n)= \frac{\lambda_1^m}{\prod_{i=1}^{m} \mu_1(i,n)} \frac{\lambda_2^n}{\prod_{i=1}^{n} \mu_2(0,i)} P_{00} \qquad (5\text{-}12)$$

Note that from (5-6) the relations

$$P_{m,n}\mu_1(m,n)=\lambda_1 P_{m-1,n}$$

and

$$P_{m,n}\mu_2(m,n)=\lambda_2 P_{m,n-1}$$

express local balance conditions for queues one and two, respectively.

Generalization

The technique used to derive the constraint for the two queue case be generalized for any finite number of queues in parallel. If there are k parallel

queues, and the subnetwork state is represented as $(n_1 \ldots n_k)$, then the constraint corresponding to (5-11) is, for all $n_i, n_j > 0$,

$$\frac{\mu_j(n_1 \ldots n_{i-1} \ldots n_k)}{\mu_j(n_1 \ldots n_k)} = \frac{\mu_i(n_1 \ldots n_{j-1} \ldots n_k)}{\mu_i(n_1 \ldots n_k)} \quad . \tag{5-13}$$

Theorem 5.3

If a two queue parallel subnetwork with subnetwork dependent processing rates is within a closed network containing N customers, then the departure independence condition for the subnetwork may be satisfied for any choice of $\mu_1(N,0)$ and $\mu_2(0,N)$, independent of all other rates.

Proof

If there are only N customers in the network, then the input rates λ_1 and λ_2 are functions of the number n+m of customers in the subnetwork. In particular, if the local balance, equations (5-2a) and (5-2b) are

$$P_{N,0}\mu_1(N,0)=P_{N-1,0}\lambda_1(N-1) \tag{5-13a}$$

and

$$P_{0,N}\mu_2(0,N)=P_{0,N-1}\lambda_2(N-1) \tag{5-13b}$$

It is evident that $\mu_1(N,0)$ and $\mu_2(0,N)$ may be determined independently of all other rates in satisfying these equations.

Constraints on processing rates occur because they must satisfy the departure independence equation as well as the local balance equation. But the only departure independence equation in which rates $\mu_1(N,0)$ appears is that of equation (5-1), written for state $P_{N-1,0}$. That is,

$$P_{N-1,0}(\lambda_1(N-1;0)+\lambda_2(N-1,0))= P_{N,0}\mu_1(N,0)+P_{N-1,1}\mu_2(N-1,1)$$

But if (5-13) holds, $\mu_1(N,0)$ cancels out of this equation. Hence $\mu_1(N,0)$ is independent of all other processing rates.

The same argument holds for $\mu_2(0,N)$. The theorem is proven.

6. Processing Rate Tradeoffs in Series Subnetworks

It is possible that processing power may be traded off between two queues in series in a queueing network. We will consider constraints on this type of processing power tradeoff in order that the series subnetwork as a whole has the local balance property, without restricting the queues to have the local balance property.

Theorem 6.1

Let $\mu_1(m,n)$ and $\mu_2(m,n)$ be the processing rates of two queues connected in series, when the queues contain m and n customers, respectively. Then the subnetwork consisting of the series queues has the local balance property if and only if for all $m,n>0$,

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

Proof

Let $\mu_1(m,n)$ and $\mu_2(m,n)$ be the processing rates at queues one and two, respectively, when there are m customers in queue one and n customers at queue two. All customers entering the series subnetwork receive service at queue one and then queue two. Let λ be the rate at which customers arrive at the subnetwork. If the rate at which customers depart from the subnetwork is to be a Poisson process whenever the input is Poisson, the local balance and departure independence equations for the subnetwork must be satisfied.

The departure independence condition for the subnetwork for m,n>0, is

$$P_{m,n}\lambda=P_{m,n+1}\mu_2(m,n+1) \tag{6-1}$$

The local balance conditions are, for m>0,

$$P_{m,0}\mu_1(m,0)=P_{m-1,0}\lambda \tag{6-2a}$$

for n>0

$$P_{0,n}\mu_2(0,n)=P_{1,n-1}\mu_1(1,n-1) \tag{6-2b}$$

and, for m,n>0

$$P_{m,n}(\mu_1(m,n)+\mu_2(m,n))=P_{m+1,n-1}\mu_1(m+1,n-1)+P_{m-1,n}\lambda. \tag{6-2c}$$

For all m>0, the departure independence condition (6-1) for state (m,0), together with the local balance condition (6-2a) for state (m+1,0) yields

$$P_{m+1,0}\mu_1(m+1,0)=P_{m,1}\mu_2(m,1). \tag{6-3}$$

This may be used as the basis for an induction on n, showing

$$P_{m+1,n-1}\mu_1(m+1,n-1)=P_{m,n}\mu_2(m,n). \tag{6-4}$$

Substituting (6-4) into (6-2c) produces

$$P_{m,n}\mu_1(m,n)=P_{m-1,n}\lambda. \tag{6-5}$$

Substituting (6-5) into the departure independence condition (6-1) written for state (m-1,n) yields, for all m>0,

$$P_{m,n}\mu_1(m,n)=P_{m-1,n+1}\mu_2(m-1,n+1), \tag{6-6}$$

which completes the inductive step.

Note that (6-4) expresses a local balance condition for queue two and (6-5) is a local balance condition for queue one.

From (6-1), written for state (m,n-1) one obtains

$$P_{m,n}=\frac{\lambda}{\mu_2(m,n)}P_{m,n-1}. \tag{6-7}$$

And (6-6) written for state (m,n-1) is

$$P_{m,n-1}\mu_1(m,n-1)=P_{m-1,n}\mu_2(m-1,n). \tag{6-8}$$

Substituting the value for $P_{m,n-1}$ from (6-8) into (6-7)

$$P_{m,n}= \frac{\lambda}{\mu_2(m,n)} \cdot \frac{\mu_2(m-1,n)}{\mu_1(m,n-1)} P_{m-1,n}. \tag{6-9}$$

But from (4-5)

$$P_{m,n}= \frac{\lambda}{\mu_1(m,n)} P_{m-1,n}, \tag{6-10}$$

hence, equating (6-9) and (6-10), one obtains

$$\frac{\mu_1(m,n-1)}{\mu_1(m,n)} = \frac{\mu_2(m-1,n)}{\mu_2(m,n)} \tag{6-11}$$

This is the same condition as (5-11), for parallel queues with processing rate tradeoffs. The necessity of the constraint is proved.

It can also be shown, through induction on N=m+n, that any rates $\mu_1(m,n)$ and $\mu_2(m,n)$ that are chosen to satisfy the constraint (6-11) satisfy the local balance and departure independence conditions. The details will be omitted.

From (6-5) and (6-7) the solution for the state probabilities of the subnetwork are obtained

$$P_{m,n} = \prod_{i=1}^{m} \frac{\lambda^m}{\mu_1(i,n)} \prod_{i=1}^{n} \frac{\lambda^n}{\mu_2(0,i)} P_{0,0} \tag{6-12}$$

This is quite similar to the state probability for the parallel subnetwork, given by (6-12). The relationship between the series and parallel two queue tradeoff subnetworks will be elucidated in the next section. However, it is not possible to generalize the processing rate tradeoff constraint for more than two queues in series, as in (6-13) for extended parallel subnetworks.

7. Processing Rates in Tradeoff Subnetworks

Subnetworks have the local balance property if the processing rates of queues that exchange processing power are constrained by (6-11). We now consider the possible assignments of processing power within this constraint that allow maximizing the processing rates in the subnetwork, and maximizing throughput of the network. The equivalent queue construction by Norton's theorem for queueing networks will be used to determine network throughput.

Since Norton's theorem has been proven for all locally balanced queues, (C1) it is unnecessary to prove it for queueing subnetworks with processing power tradeoff that meet the local balance conditions. Instead, a sufficient condition for the equivalent queue construction of Norton's theorem will be shown. This condition provides some insight for the equivalent queue construction, and elucidates relationships needed for the optimization of throughput.

Theorem 7.1

Let $P_{m,n|N}$ be the conditional probability that a two queue tradeoff subnetwork, which is part of a closed network of locally balanced queues containing $M \geq N$ customers, is in state (m,n), given that there are $N=m+n$ customers in the subnetwork. Let $Q_{m,n}$ be the probability of subnetwork state (m,n) when the subnetwork is itself a closed network (with output connected to input) containing N customers. Then $Q_{m,n}=P_{m,n|N}$.

Proof:

Let $\mu_1(m,n)$ and $\mu_2(m,n)$ be the processing rates at the queues of a series tradeoff subnetwork when the subnetwork state is (m,n).

$$\text{Let } Y(m,n) = \prod_{i=1}^{m} \frac{1}{\mu_1(i,n)} \prod_{i=1}^{n} \frac{1}{\mu_2(0,i)} \qquad (7\text{-}1)$$

and let $X(N)= \sum_{m+n=N} Y(m,n)$. (7-2)

Then from (6-12), $P_{m,n}=\lambda^{m+n}Y(m,n)P_{0,0}$. The term $Y(m,n)$ is used in the product form state probability for a network containing the tradeoff subnetwork. Consider the three queue network consisting of the tradeoff subnetwork in series with queue zero, an exponential server that has processing rate $U(N)$ when it contains N customers. Queue zero may represent the equivalent queue of some network and hence the three queue network represents the tradeoff subnetwork in an arbitrary closed network of locally balanced queues.

Let $Z(K)= \prod_{i=0}^{k} \frac{1}{U(k)}$. Let $Q_{k,m,n}$ be the probability of k customers in queue zero, and tradeoff subnetwork state (m,n). It is straightforward to show that

$$Q_{k,m,n}= \frac{1}{G_3(M)} Z(k)Y(m,n)$$

satisfies the balance equation for the network when it contains $M=k+m+n$ customers, and $G_3(M) = \sum_{k+m+n=M} Z(k)Y(m,n)$ is the normalization constant.

Then, $P_{m,n|N} = \frac{Q_{k,m,n}}{\sum_{m'+n'=N} Q(k,m',n')} = \frac{Y(m,n)}{X(N)}$. (7-3)

But the state probabilities for the network consisting of the tradeoff subnetwork alone, containing N customers, are the limiting values of the three-queue network state probabilities with N customers, as the processing rates at queue zero increase beyond all bounds. Let

$$G_2(N)= \lim_{\substack{U(n)\to\infty \\ n>0}} G_3(N).$$

Since $\lim_{\substack{U(n)\to\infty \\ n>0}} Z(k)=0$ for all $k>0$, and $Z(0)=1$, then $G_2(N)=\sum_{m+n=N} Y(m,n)=X(N)$. Hence,

$$Q_{m,n}=\lim_{U(n)\to\infty} Q_{k,m,n}=\begin{cases}\dfrac{Y(m,n)}{X(N)}, & k=0\\ 0, & k\neq 0\end{cases} \qquad (7\text{-}4)$$

Comparing (7-3) and (7-4) shows that the proof is complete.

The interesting point with respect to this theorem is that the tradeoff subnetwork state probabilities are expressed in terms of rates $\mu_i(m,n)$ for $m+n<N$, but these rate obviously do not apply when the subnetwork stands alone containing N customers. Of course, the local balance condition (5-11) can be used to express these probabilities in terms of $\mu_i(m,n)$, $m+n=N$ only.

The mean processing rate U(N) of the tradeoff subnetwork when it contains N customers can be compute as

$$U(N)=\sum_{\substack{m+n=N \\ n\neq 0}} \mu_2(m,n)Q_{k,m,n}$$

Alternatively, U(N) is the throughput of the tradeoff subnetwork when its output is connected to its input and it contains N customers.

$$U(N)=\frac{G_2(N-1)}{G_2(N)}=\frac{X(N-1)}{X(N)} \qquad (7\text{-}5)$$

For $N>1$, each term $Y(m,n)$ of $G_2(N)$ has exactly one factor $\mu_1^{-1}(1,0)$ or $\mu_2^{-1}(0,1)$. Hence, multiplying numerator and denominator of (7-5), $N>1$, by

$\dfrac{\mu_1(0,1)}{\mu_1(1,1)}=\dfrac{\mu_2(1,0)}{\mu_2(1,1)}$ shows that U_N can be expressed in a form that is independent of rates $\mu(i,j)$, where $i+j=1$.

Similarly, multiplying numerator and denominator of (7-5) by the factor

$\prod_{k=1}^{N-1} \frac{\mu_2(0,k)}{\mu_2(N-k,k)}$ or its equivalent forms by the relation (6-11), shows that U(N) can be expressed as a function of rates $\mu_i(j,k)$ where j+k=N only. In particular, let

$$Y'(m,n)=\prod_{i=1}^{m} \frac{1}{\mu_1(i,N-i)} \prod_{i=1}^{n} \frac{1}{\mu_2(N-i,k)}, \qquad (7\text{-}6)$$

and $G_2'(N) = \sum_{m+n=N} Y'(m,n)$.

Then $U_N = \frac{G_2'(N-1)}{G_2'(N)}$.

In the next sections, the maximization of throughput in tradeoff subnetworks will be demonstrated. These results and those just shown are derived for series subnetworks. But the results are applicable to parallel subnetworks as well, by means of the transformation described in the following theorem.

Theorem 7.2 Let p_1 and $p_2 = 1-p_1$, be the probabilities that customers arriving at a locally balanced two queue parallel tradeoff subnetwork receive service at queues one and two, respectively. Let $\mu_1(m,n)$ and $\mu_2(m,n)$ be the processing rates of the queues when they contain m and n customers, respectively, and let $P_{m,n}$ be the conditional probability of that state. Then the state probabilities and the throughput of the parallel subnetwork are the same as those of a series two queue tradeoff subnetwork whose queues have service rates $\mu_1(m,n)/p_1$ and $\mu_2(m,n)/p_2$, respectively.

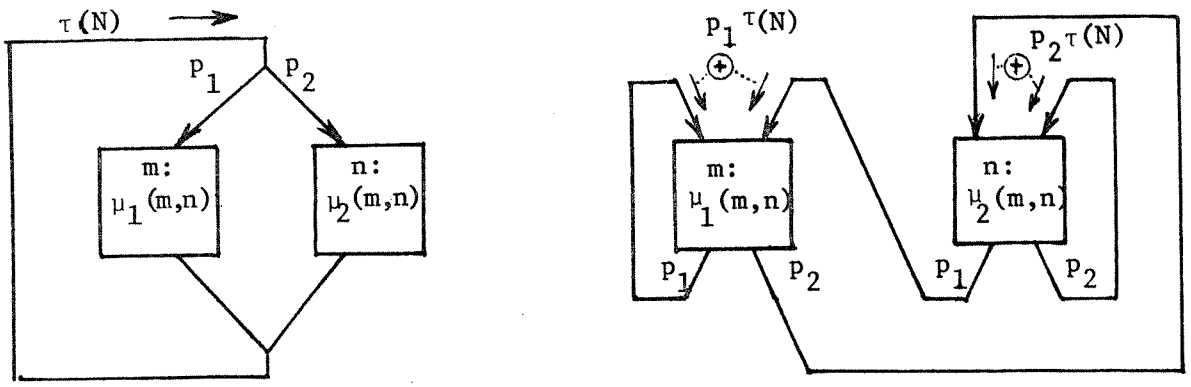
Proof

The theorem is proved straightforwardly by noting that the state probability for the parallel subnetwork, which is given by (5-12), with $\lambda_i=p_i\lambda$, is the same as that of the series subnetwork, (6-12) with each occurence of a $\mu_i(i,j)$ multiplied by p_i^{-1}.

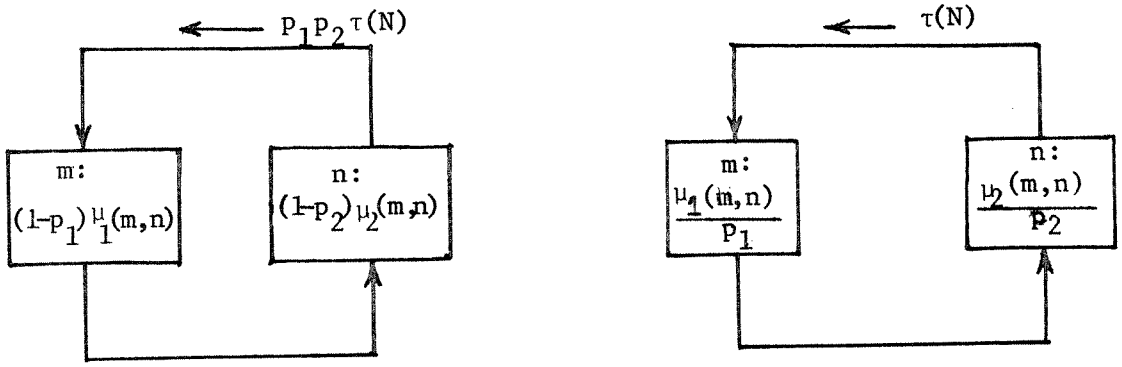
But more insight into these relationships is provided by use of Norton's theorem. By Norton's theorem, the total processing rate U'(N) of the parallel

subnetwork when it contains N customers is the throughput in the link connecting output to input of the subnetwork when there are N customers in the closed network obtained by making this connection. (See Figure 1a). But at each departure from queue i in this network, a customer returns to queue i with probability p_i. This is shown explicitly by redrawing the network as in Figure 1b. The effective processing time of a customer at a queue is the sum of k processing periods, the length of each described by an exponential distribution with mean time $\mu_i^{-1}(m,n)$. But k is geometrically distributed with parameter p_i. The total effective processing time is then known to be exponential, with mean rate $(1-p_i)\mu_i$. Replacing the parallel queues with the return loops by queues with rate $(1-p_i)\mu_i$ connected in series, does not alter the conditional probabilities $P_{m,n|N}, N=m+n$ of the network. The network of Figure 1c is equivalent to that of 1b.

The mean throughput of the equivalent series network, U'(N) is the same as the rate of flow from queue one to queue two in the parallel network . The flow rate into queue one is p_1U_N, and so $p_1p_2U_N$ is the flow rate between queues one and two. Hence, $U(N) = p_1p_2U_N$. Therefore, dividing the processing rate of each queue of this equivalent series network by p_1p_2 does not alter the conditional probabilities $P_{m,n|N}$, but does adjust the processing rate for N customers to be the same as that of the parallel subnetwork. This is the network shown in Figure 1d. Because both processing rates and conditional probabilities for the series subnetwork with rates $\mu_i(m,n)/p_i$ are the same as those of the parallel subnetwork, the series subnetwork has the same state probabilities as well.



a) b)



c) d)

Figure 1
The equivalence of series and
parallel two queue networks.

8. Subnetwork Processing-Power Tradeoffs to Maximize Throughput

Network throughput can be maximized by optimum subnetwork dependent processing rate assignments. The process at queue i of the tradeoff subnetwork will be considered to have mean rate μ_i when a single processor is assigned the queue. The actual processing rate is proportional to the number of processors assigned the queue. If $r_i(m,n)$ processors are assigned queue i when the subnetwork state is (m,n), then $\mu_i(m,n) = r_i(m,n)\ \mu_i$. The total number of processors available to the subnetwork when it contains N customers is R(N). Subnetworks that have the local balance property will be considered.

Theorem 8.1

Let R(n) be the maximum processing power that may be distributed to the queues of a locally balanced tradeoff subnetwork when the subnetwork contains n customers, and let $\mu(n)$ be the processing rate of the subnetwork when it contains n customers. If the subnetwork is contained in a network of locally balanced queues containing N customers, then $\mu(N)$ is maximized within the local balance constraint by full utilization of all the R(N) available processors; that is, $r_1(m,n)+r_2(m,n) = R(N)$ for all $m+n=N$, with $r_1(0,N) = r_2(N,0)=0$. Further, the maximum processing rate is independent of the actual distribution of processing power R(N) to the subnetwork queues.

Proof

By the Norton's theorem construction, the processing rate $\mu(N)$ of the locally balanced tradeoff subnetwork when it contains N customers is the throughput in the link from output to input of the tradeoff subnetwork, when there are N customers in the network formed by making this connection. But in a two queue closed network, queue state dependent processing rates are the same as subnetwork dependent rates. Hence, Theorem 2.1 applies, and $\mu(N)$ is maximized when the R(N) processors are fully utilized, and if $\mu(N)$ is maximum,

$$\mu(N) = \left(\frac{1}{\mu_1} + \frac{1}{\mu_2} \right) R(N).$$

The proof is complete.

Theorem 8.2

Let R(N) be the total processing power available to the queues of a tradeoff subnetwork when it contains N customers, and is within a closed network of locally balanced queues containing M≥N customers. Let $r_1(m,n)$ and $r_2(m,n)$ be the processing rates assigned to each subnetwork queue when the subnetwork state is (m,n), m+n=N. Then if the processing power available to each queue of the network outside the tradeoff subnetwork is maximally utilized, the throughput of the subnetwork is maximized by the processing power assignments $r_1(m,n) = \frac{m}{N} R(N)$ and $r_2(m,n) = \frac{n}{N} R(N)$.

Proof

Let U(n) be the processing rate of the equivalent queue for the closed network with respect to the branch containing the processing tradeoff subnetwork. Then by Theorem 3.2, U(n+1) ≥ U(n).

Let $\mu(n)$ be the processing rate of the equivalent queue for the tradeoff subnetwork when it contains n customers. Then by Theorem 2.1b, $\mu(n)$ is to be maximized for each n, in order to maximize throughput in the network consisting of the equivalent queues, if the $\mu(n)$, 1<n<N are independent of each other.

By Theorem 8.1, $\mu(n)$ is maximized by the choice $r_1(N,0) = r_2(0,N) = R(N)$ and is independent of the choice $r_i(m,n)$ for m,n>0, as long as $r_1(m,n)+r_2(m,n)=R(N)$. Hence, by Theorem 4.3, the optimal processing power assignments for $\mu_i(m,n)$ where m+n=N place no contraints on $\mu_i(m,n)$ for m+n<N. Hence, $\mu(N-1)$ can be maximized following the same procedure, resulting in the assignments $r_1(0,N-1) = r_2(N-1,0) = 0$ and $r_1(m,n) + r_2(m,n) = R(N-1)$ or all m+n=N-1. Maximizing each U(n), n-N,N-1 ... 1 by the same procedure, $\mu_1(1,0)=R(1)\mu_1$, and $\mu_2(0,1)=R(1)\mu_2$.

Then by (5-11), $\frac{\mu_1(1,0)}{\mu_1(1,1)} = \frac{R(1)}{r_1(1,1)} = \frac{R(1)}{r_2(1,1)} = \frac{\mu_2(0,1)}{\mu_2(1,1)}$

and $r_1(1,1) = r_2(1,1) = r_2(1,1) = \frac{R(2)}{2}$. By (5-11) again,

$$\frac{\mu_1(2,0)}{\mu_1(2,1)} = \frac{R(2)}{r_1(2,1)} = \frac{R(2)/2}{r_2(2,1)} = \frac{\mu_2(1,1)}{\mu_2(2,1)}$$

and hence $r_1(2,1) = 2r_2(2,1) = 2R(3)$. Repeated application of 5-11 shows $r_1(m,n) = \frac{m}{m+n} R(m,n)$ and $r_2(m,n) = \frac{n}{m+n} R(m,n)$.

The proof is complete.

9. Load Dependent Branching Probabilities

Closely analogus to the case in which processing rates may be traded off due to load, is the case in which arrival rates at queues may be traded off. In the usual model of subnetworks with parallel branches, arriving customers are probabalistically directed toward one of the branches. For maximizing throughput, the probability of a customer being directed to a particular branch should increase as the number of customers in that branch decreases relative to the other branches of the subnetwork. As the number of customers in one branch of the subnetwork changes, the arrival rate to all branches are modified. Thus, we must consider local dependent branching probabilities, or subnetwork state dependent arrival rates.

Theorem 9.1

Let $p_1(m,n)$ and $p_2(m,n)=1-p_1(m,n)$ be the probabilities that a customer arriving at a subnetwork of two locally balanced queues, is directed to queue one or queue two, respectively, when there are m customers at queue one and n customers at queue two. Then the subnetwork has the local balance property if and only if, for all $m,n \geq 0$,

$$\frac{p_1(m,n+1)}{p_1(m,n)} = \frac{p_2(m+1,n)}{p_2(m,n)}$$

Proof

Let $\lambda(N)$ be the mean rate of arrivals to the two queue subnetwork when there are N customers in the subnetwork, and let μ_1 and μ_2 be the mean processing

rates of the queues. Let $\lambda_1(m,n)=p_1(m,n)\lambda(N)$ and $\lambda_2(m,n)=p_2(m,n)\lambda(N)$ be the input rates to the respective queues when the subnetwork state is (m,n) and $m+n=N$.

Then for the subnetwork to have the local balance property, the local balance and balance equations for the subnetwork must be satisfied.

The local balance conditions are,

for all $m>0, P_{m,0}\mu_1 = \lambda_1(m-1,0)P_{m-1,0}$ (9-1a)

for all $n>0, P_{n,0}\mu_2 = \lambda_2(0,n-1)P_{0,n-1}$ (9-1b)

and for all $m,n>0, P_{m,n}(\mu_1+\mu_2)=\lambda_2(m,n-1)P_{m,n-1}+\lambda_1(m-1,n)P_{m-1,n}$ (9-1c)

The balance equations for the two queue subnetwork will be called the departure independence equations when the local balance conditions (9-1) are satisfied. For all $m,n\geq 0$, the departure independence equations are

$$P_{m,n}(\lambda_1(m,n) + \lambda_2(m,n)) = P_{m+1,n}\mu_1+P_{m,n+1}\mu_2. \quad (9-2)$$

Recursive solutions to (9-1a) are, for $m > 0$,

$$P_{m,0} = \frac{\lambda_1(m-1,0)}{\mu_1} P_{m-1,0} = \frac{\prod_{i=0}^{m-1} \lambda_1(i,0)}{\mu_1^m} P_{0,0} \quad (9-3a)$$

and, for $n>0$,

$$P_{0,n} = \frac{\lambda_2(0,n-1)}{\mu_2} P_{0,n-1} = \frac{\prod_{i=0}^{n-1} \lambda_2(0,i)}{\mu_2^m} P_{0,0}. \quad (9-3b)$$

The solution to (9-1) and (9-2) is completed by the relation

$$P_{m,n} = \frac{\lambda_2(m,n-1)}{\mu_2} P_{m,n-1} = \frac{\lambda_1(m-1,n)}{\mu_1} P_{m-1,n} \quad (9-4)$$

whenever $m,n > 0$. The complete solution is demonstrated by induction on $k(m,n) = \frac{(m+n+1)(m+n)}{2} + n$. The basis, and the inductive steps whenever $m=0$ or $n=0$ are given by (9-3). Assume the solution holds for all $P_{i,j}$ such that $k(i,j) < k(m,n)$. In particular, $k(m+1,n-1) < k(m,n)$ for $n>0$, hence

$$P_{m+1,n-1} = \frac{\lambda_1(m,n-1)}{\mu_1} P_{m,n-1} \tag{9-5}$$

Equation (9-2), written for $P_{m,n-1}$ is

$$P_{m,n-1}(\lambda_1(m,n-1)+\lambda_2(m,n-1)) = P_{m+1,n-1}\mu_1+P_{m,n}\mu_2. \tag{9-6}$$

Substituting (9-5) into (9-6) yields

$$P_{m,n-1}\ \lambda_2(m,n-1) = P_{m,n}\mu_2, \tag{9-7}$$

which provides the first equality of (9-4) for $k(m,n)$.

The second equality of (9-4) may be shown by a similar induction on $k'(m,n) = \frac{(m+n+1)(m+n)}{2} + m$.

By (9-4), it is seen that for $m,n>0$

$$P_{m,n} = \frac{\lambda_2(m,n-1)}{\mu_2}\ \frac{\lambda_1(m-1,n-1)}{\mu_1} P_{m-1,n-1} = \frac{\lambda_1(m,n-1)}{\mu_1} \cdot \frac{\lambda_2(m-1,n-1)}{\mu_2} P_{m-1,n-1} \tag{9-8}$$

Since $\lambda_1(m,n) = p_1(m,n)\lambda$ and $\lambda_2(m,n) = p_2(m,n)\lambda$, the relation

$$p_2(m,n-1)p_1(m-1,n-1) = p_1(m-1,n)p_2(m-1,n-1)$$

is derived from (9-8). For all $m,n \geq 0$, this condition is written

$$\frac{p_2(m+1,n)}{p_2(m,n)} = \frac{p_1(m,n+1)}{p_1(m,n)} \tag{9-9}$$

Since (9-9) was derived directly from the balance equations and local balance conditions, it is seen to be a necessary condition for local balance.

It can also be shown to be sufficient condition by induction on N=m+n.

If (9-1) and (9-2) are satisfied for $P_{m,n}$ defined by $p_1(i,j)$ and $p_2(i,j)$ for $i+j<m+n$, then they are satisfied for $P_{m+1,n}$ and $p_{m,n+1}$ defined by $p_1(i,j)$ and $p_2(i,j)$ for $i+j=m+n$, if these probabilities satisfy the constraint (9-9). The theorem is proved.

From (9-4) and (9-3), the state probabilities of a subnetwork with load-dependent branching probabilities are seen to be

$$P(m,n) = \frac{\prod_{i=0}^{m-1} \lambda_1(i,n)}{\mu_1^m} \cdot \frac{\prod_{i=0}^{n-1} \lambda_2(0,i)}{\mu_2^n} P_{0,0}. \qquad (9\text{-}10)$$

Generalization

The technique used to determine the constraints on the branching probabilities for the two queue case is easily generalized to the case of any finite number of queues in parallel. For load-dependent branching to any of k parallel queues, the constraint corresponding to (9-9) is, for $1 \leq i,j \leq k$.

$$\frac{p_i(n_1,\ldots n_j+1,\ldots n_k)}{p_i(n_1,\ldots n_k)} = \frac{p_j(n_1\ldots n_i+1,\ldots n_k)}{p_j(n,\ldots n_k)}. \qquad (9\text{-}11)$$

References

B1. 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.

C1. 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.

C2. Chandy, K. M., Herzog, U., and Woo, L., "Parametric Analysis of Queueing Network Models" IBM Journal of Research and Development, 19,1 (January 1975), pp. 36-42.

C3. Chandy, K. M., Howard, J. H., and Towsley, D. F., "Product Form and Local Balance in Queueing Networks," to appear in JACM.

G1. Gordon, W. J. and G. F. Newell, "Closed queueing systems with exponential servers", Oper. Res. 15, 2 (1967), pp. 252-267.

J1. Jackson, J. R., "Job-shop like queueing systems", Man. Sci., 10, (1963), pp. 131-142.

M1. Mirasol, N. M., "The Output of an M/G/∞ Queue is Poisson", Oper. Res. 11, (1968), pp. 282-284.

M2. Muntz, R. R., "Poisson departure processes and queueing networks", IBM Research Report, RC-4145, Yorktown Heights, New York, 1972.

N1. Noetzel, A. S., "A Complete Class of M/G/n Queueing Disciplines with the MM property". Technical Report 50, Department of Computer Sciences, The University of Texas at Austin. June, 1975.

N2. Noetzel, A. S., "Product-Form Queuing Networks with Processing-Rate and Arrival-Rate Tradeoffs" Technical Report 53, Department of Computer Sciences, The University of Texas at Austin, December 1975.

T1. Towsley, D. F., "Local Balance Models of Computer Systems", Ph.D. Dissertation, The University of Texas at Austin, December 1975.