K. M. Chandy
U. Herzog
L. Woo

Parametric Analysis of Queuing Networks

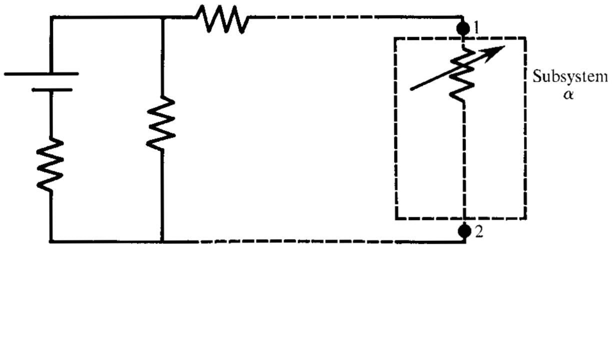
Abstract: We consider a queuing network with M exponential service stations and with N customers. We study the behavior of a subsystem σ, which has a single node as input and a single node as output, when the subsystem parameters are varied. An "equivalent" network is constructed in which all queues except those in subsystem σ are replaced by a single composite queue. We show that for certain classes of system parameters, the behavior of subsystem σ in the equivalent network is the same as in the given network. The analogy to Norton's theorem in electrical circuit theory is demonstrated. In addition, the equivalent network analysis can be applied to open exponential networks.

Introduction

Queuing models are widely used to analyze and design a variety of systems. These models are generally used to study the variation of certain system parameters such as response time as a function of the network structure and service times. In this paper, the authors determine a relationship that exists between some queuing networks and electrical networks, with customers and throughput (rate of flow of customers) being analogous to electrical charge and electrical current, respectively. The authors show how to directly apply some methods from electrical circuit theory, in particular Norton's theorem [1], to queuing networks.

Consider an electrical circuit consisting of batteries and resistors, Fig. 1. To study the behavior of a subsystem σ between terminals 1 and 2 of Fig. 1 as the subsystem parameters are varied, construct an "equivalent" circuit in which all the components of the given circuit, except those of subsystem σ, are replaced by a single current source and a parallel internal resistance, Fig. 2. The value of the current source is set equal to the current flowing between terminals 1 and 2 when subsystem σ is replaced by a "short," Fig. 3, and the value of the internal resistance is determined by the open circuit voltage. The behavior of subsystem σ in the equivalent circuit is the same as in the given circuit. This is referred to as Norton's theorem [1]. The analysis of the equivalent circuit requires less computation than the analysis of the original circuit; sensitivity analysis may hence be efficiently carried out on the equivalent circuit.

Figure 1 Electrical network.



Consider a closed queuing network with N customers. We study the behavior of subsystem σ between terminals 1 and 2, as the subsystem parameters are varied, Fig. 4. Construct an equivalent network in which all the queues, except those in subsystem σ, are replaced by a single composite queue, Fig. 5. Let $T(n)$ be the service rate for the composite queue when n is the number of customers waiting for or being served at this queue ($n = 0, 1, \cdots, N$). Set $T(n)$ equal to the rate at which customers pass (throughput) between terminals 1 and 2, when there are n customers in the given network and when subsystem σ is replaced by a "short," i.e., when the service times of all the servers in subsystem σ are set to zero, Fig. 6. The behavior of subsystem σ in the equivalent network is the same as in the given network. Although there does not exist a strict correspondence to internal resistance in the queuing network, the concept of flow rate and the shortening of the subsystem suggests an approach analogous to Norton's theorem in circuit theory. Therefore, we shall refer to it as Norton's theorem for queuing networks.

In this paper we show that Norton's theorem does hold for certain classes of queuing networks that obey local balance [2, 3]; however, the theorem does not

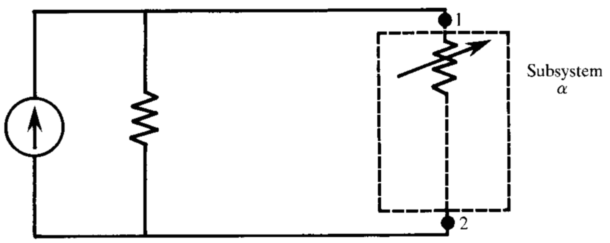


Figure 2 Equivalent electrical network.

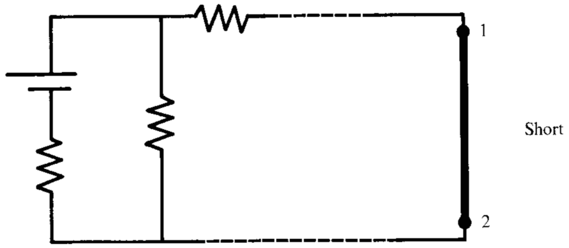


Figure 3 Subsystem σ shorted.

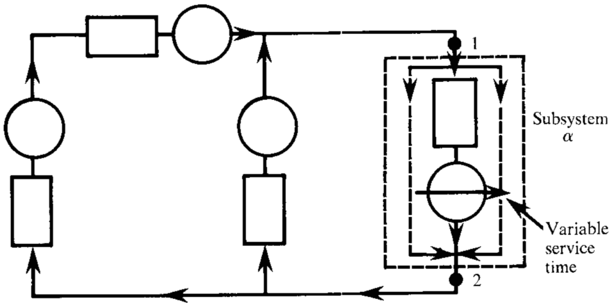


Figure 4 Queuing network.

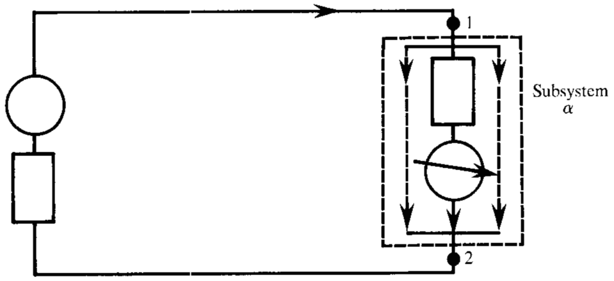


Figure 5 Equivalent queuing network.

necessarily hold for those networks in which local balance does not hold. In particular, Norton's theorem holds for the exponential networks studied by Gordon and Newell [4] and other locally balanced networks analyzed in [3]. Norton's theorem can also be applied to open networks that satisfy local balance.

Example Consider the central-server model shown in Fig. 7 in which all service times are independent exponential random variables [5]. We investigate the CPU's throughput, queue length and queue time distributions as the CPU service time is varied. We determine the throughput through the network when the CPU service time is reduced to zero, Fig. 8, and when there are n customers in the system, $n = 0, 1, \cdots, N$. The throughputs as a function of the level of multiprogramming are shown in the table in Fig. 8. The equivalent network is shown in Fig. 9. The behavior of the CPU in the equivalent network of Fig. 9 is the same as in Fig. 7.

Joint probability distribution

We restrict our attention to the class of closed queuing models with exponential servers studied by Gordon and Newell [4]. Let there be M queues in the network, which are indexed $1, 2, \cdots, M$, and N customers. The service rate for the ith queue when there are k customers in the queue is $U_i(k)$, where $i = 1, \cdots, M$ and $k = 1, \cdots, N$. The service discipline for all servers is first come, first served. When a customer has been served in queue i he joins queue j with probability p_{ij} independent of the current state of the system ($i, j = 1, \cdots, M$). The states of the system are M-tuples $(n_1, \cdots, n_M)$, where n_i is the number of customers in queue i including the customer that is being served ($i = 1, \cdots, M$). Clearly n_i, $i = 1, \cdots, M$, are nonnegative integers and $n_1 + \cdots + n_M = N$. Let $P(n_1, \cdots, n_M)$ be the probability that the system is in state $(n_1, \cdots, n_M)$, which is assumed to be a feasible state. Gordon and Newell [4] showed that $P(n_1, \cdots, n_{\mathrm{M}})$ may be expressed by

$$P(n_1, \cdots, n_M) = g(n_1, \cdots, n_M)/G, \qquad (1)$$

where

$$\begin{aligned} g(n_1, \cdots, n_M) &= \prod_{i=1}^{M} x_i(n_i) \text{ for any feasible state,} \\ &= 0 \text{ otherwise,} \qquad (2) \end{aligned}$$

and G is a normalizing constant.

The quantities $x_i(n_i)$ are defined recursively as follows:

$$x_i(0) = 1,\ x_i(k) = x_i(k-1)\ y_i/U_i(k),$$
$$\text{for } k = 1, \cdots, N;\ i = 1, \cdots, M, \qquad (3)$$

with y_i, $i = 1, \cdots, M$, being a solution of the following M linear equations:

$$\sum_{i=1}^{M} y_i p_{ij} = y_j,\ j = 1, 2, \cdots, M. \qquad (4)$$

(The set of numbers y_i is unique up to a normalizing constant [4].)

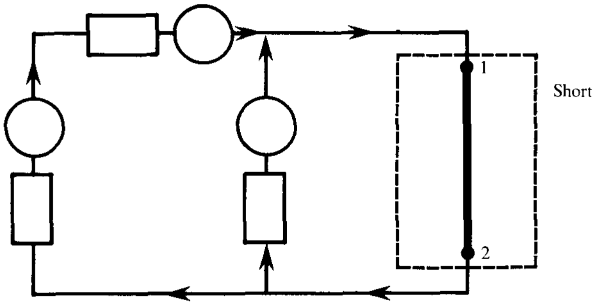


Figure 6 Subsystem σ shorted.

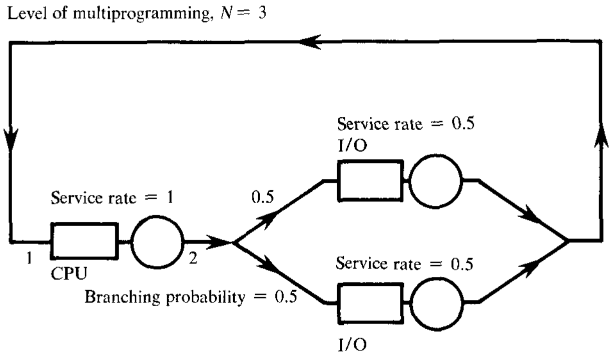


Figure 7 Central server model.

We briefly review some of the computational techniques developed by Buzen [5]. We define a "convolution" between two $(R + 1)$-dimensional vectors $\mathbf{A} = [a(0), a(1), \cdots, a(R)]$ and $\mathbf{B} = [b(0), b(1), \cdots, b(R)]$ as an $(R + 1)$ dimensional vector $\mathbf{C} = [c(0), c(1), \cdots, c(R)]$, where $c(i) = \sum_{j=0}^{R} a(j)b(i - j)$ and denote the operation by $*$. Thus $\mathbf{C} = \mathbf{A} * \mathbf{B}$. Note that the result of our convolution operation consists of the first $R + 1$ elements of the result of a regular convolution. Let us define an $(N + 1)$-dimensional vector $\mathbf{X}_i = [x_i(0), \cdots, x_i(N)]$, where $x_i(\cdot)$ is defined in equation (3). We define $M + 1$ vectors $\mathbf{G}_0, \mathbf{G}_1, \cdots, \mathbf{G}_M$ each of dimension $N + 1$:

$$\mathbf{G}_0 = (1, 0, 0, \cdots, 0); \tag{5}$$

$$\mathbf{G}_i = \mathbf{X}_1 * \mathbf{X}_2 * \cdots * \mathbf{X}_i, \text{ for } i = 1, \cdots, M, \tag{6}$$

where

$$\mathbf{G}_i = \mathbf{G}_{i-1} * \mathbf{X}_i, \text{ for } i = 1, \cdots, M. \tag{7}$$

Let $G_i(r)$ represent the rth element of $\mathbf{G}_i$, $r = 0, 1, \cdots, N$. From the definition of convolution it follows that

$$G_i(r) = \sum_{\mathscr{R}} g(n_1, \cdots, n_i, 0, \cdots, 0), \tag{8}$$

where $\mathscr{R} = \{n_i + \cdots + n_i = r\}$.

Therefore the normalizing factor G in equation (1) is equal to $G_M(N)$.

- *Marginal probability distribution*

Let $P_i(n)$ be the marginal probability that there are n customers in queue i. Let us restrict our attention to queue M. Note that

$$p(n_1, \cdots, n_{M-1}, n) = g(n_1, \cdots, n_{M-1}, 0)\ x_M(n)/G. \tag{9}$$

Summing over all the states in which there are n customers in queue M we get

$$P_M(n) = x_M(n)\ G_{M-1}(N - n)/G \text{ for } n = 0, 1, \cdots, N. \tag{10}$$

The marginal probability for any given queue can be obtained by renumbering the queues so that the given queue is indexed M.

- *Throughput*

Let $\mathscr{T}_i$ be the throughput of queue i, i.e., the rate at which customers get serviced and leave the queue:

$$\mathscr{T}_M = \sum_{n=1}^{N} P_M(n)\ U_M(n). \tag{11}$$

From equation (3)

$$x_M(n)\ U_M(n) = x_M(n - 1)\ y_M,\ n = 1, \cdots, N. \tag{12}$$

From equations (10), (11) and (12)

$$\begin{aligned}\mathscr{T}_M &= y_M \sum_{n=1}^{N} x_M(n - 1)\ G_{M-1}(N - n)/G \\ &= y_M\ G_M(N - 1)/G.\end{aligned} \tag{13}$$

- *Queue length distribution at arrival*

Let $Q_i(n)$ be the probability that an arriving customer at queue i finds himself in the nth place in the queue. We now compute $Q_M(n)$ for $n = 1, \cdots, N$ as follows: Let the states S_i and S be represented by the tuples $(n_1, \cdots, n_i + 1, \cdots, n_M - 1)$ and $(n_1, \cdots, n_i, \cdots, n_M)$, respectively. The net rate at which the system transits from state S_i to state S when a customer arrives at queue M after being served by the i server is

$$\begin{aligned}&P(S_i)\ U_i(n_i + 1)\ p_{iM} \\ &\qquad = y_i\, p_{iM}\, g(n_1, \cdots, n_{M-1}, n_M - 1)/G.\end{aligned} \tag{14}$$

Hence the net rate at which the system transits into state $S = (n_1, \cdots, n_M)$ due to a customer's arrival at queue M is

$$\begin{aligned}r_M(S) &= \sum_{i=1}^{M} P(S_i)\ U_i(n_i + 1)\ p_{iM} \\ &= \Big(\sum_{i=1}^{M} y_i\, p_{iM}\, g(n_1, \cdots, n_{M-1}, n_M - 1\Big)/G \\ &= y_M\, g(n_1, \cdots, n_{M-1}, n_M - 1)/G.\end{aligned} \tag{15}$$

Note: For simplicity, p_{MM} is assumed to be zero; however, the same method can be applied to the general cases where $p_{MM} \neq 0$.

During a long time interval T, the number of transitions into state S due to a customer's arrival at queue M is approximately equal to $r_M(S)$ multiplied by T. Hence, considering all the arrivals of customers at queue M, the probability p that such an arrival causes a transition into state S is

$$p = r_M(S)T/\Sigma r_M(S)T = r_M(S)/\Sigma r_M(S), \qquad (16)$$

where the summation is over all the states, and therefore p is proportional to $r_M(S)$.

Let $r_M(n)$ be the total net rate at which the number of customers in queue M is increased from $n-1$ to n.

Let us define $\mathcal{N} = \{n_1 + \cdots + n_{M-1} = N - n\}$. Then it follows that from Eq. (15) that

$$\begin{aligned} r_M(n) &= \sum_{\mathcal{N}} g(n_1, \cdots, n_{M-1}, n-1)\ y_M/G \\ &= [y_M\, x_M(n-1)/G] \sum_{\mathcal{N}} g(n_1, \cdots, n_{M-1}, 0) \\ &= [y_M\, x_M(n-1)/G]\ G_{M-1}(N-n) \\ &\quad \text{for } n = 1, \cdots, N. \end{aligned} \qquad (17)$$

Since $Q_M(n)$ is directly proportional to $r_M(n)$, it follows that $Q_M(n)$ is directly proportional to $x_M(n-1)G_{M-1}(N-n)$ for $n = 1, \cdots, N$. By setting $Q_M(1) + \cdots + Q_M(N) = 1$, we obtain $Q_M(n)$.

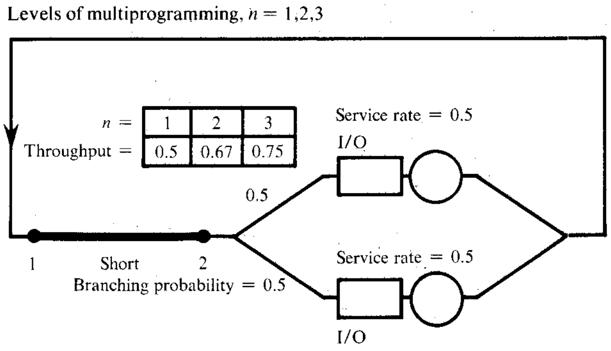


Figure 8 Central server model (CPU shorted).

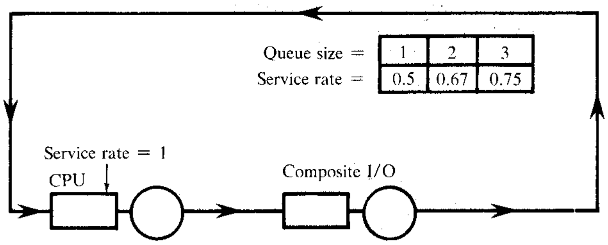


Figure 9 Central server model (equivalent network).

Norton's theorem for queuing networks

- *Closed networks*

Consider the closed network of M queues of the last section. We determine the queue length distribution for queue M as a function of the service rate parameter. We construct an equivalent network consisting of queue M and a composite queue with service rate $T(n)$; n is the number of customers in the composite queue, $n = 0, 1, \cdots, N$. Set $T(n)$ equal to the throughput of queue M when there are n customers in the given network and the service time for queue M is reduced to zero. Let $P'_M(n)$ be the marginal probability of the queue length of M in the equivalent network.

Theorem 1 The queue length distribution for queue M in the equivalent network is the same as in the given network. In other words,

$$P'_M(n) = P_M(n) \text{ for } n = 0, 1, \cdots, N.$$

Proof From Eqs. (1), (2) and (3) it follows that the probability $P'_M(n)$ of $N-n$ customers in the composite queue and n customers in queue M in the equivalent network is directly proportional to

$$\prod_{j=1}^{n} 1/U_M(j) \prod_{k=1}^{N-n} 1/T(k); \qquad (18)$$

for a detailed derivation of the above see [4]. When the mean service time of queue M is reduced to zero we have $x_M(n) = 1$ for $n = 0$ and $x_M(n) = 0$ for $n \neq 0$; in this case we get

$$T(N) = y_M\, G_{M-1}(N-1)/G_{M-1}(N). \qquad (19)$$

Similarly,

$$T(n) = y_M\, G_{M-1}(n-1)/G_{M-1}(n),\ n = 1, \cdots, N. \qquad (20)$$

Substituting (20) in (18) we find that $P'_M(n)$ is directly proportional to

$$\frac{x_M(n)}{(y_M)^N} \frac{G_{M-1}(N-n)}{G_{M-1}(0)} \text{ for } n = 0, 1, \cdots, N. \qquad (21)$$

Hence,

$$P'_M(n) \propto x_M(n)\ G_{M-1}(N-n) \text{ for } n = 0, 1, \cdots, N. \qquad (22)$$

But from (10),

$$P_M(n) \propto x_M(n)\ G_{M-1}(N-n) \text{ for } n = 0, 1, \cdots, N. \qquad (23)$$

Hence,

$$P'_M(n) = P_M(n) \text{ for } n = 0, \cdots, N. \qquad (24)$$

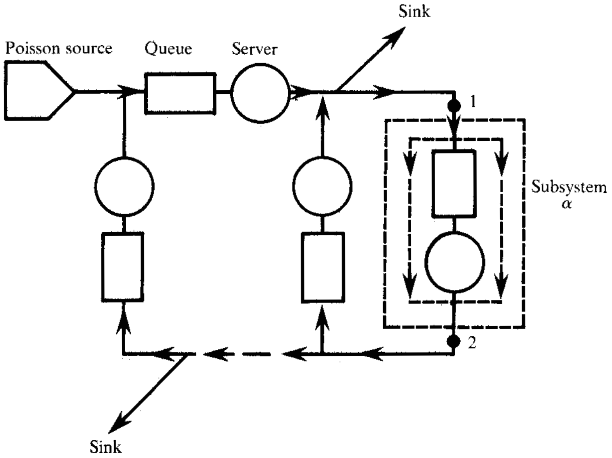


Figure 10 Open queuing network.

Theorem 2 The queue length distribution at arrival for queue M in the equivalent network is the same as in the given network:

$Q'_M(n) = Q_M(n)$ for $n = 1, \cdots, N$.

Proof Following the same argument as that for Theorem 1, we observe that $Q_M(n)$ is directly proportional to $x_M(n-1)\ G_{M-1}(N-m)$.

Corollary The queue time distribution of arrival for queue M in the equivalent network is the same as in the given network.

Discussion Consider a closed exponential queuing network with M queues indexed $1, 2, \cdots, M$ and N customers. Consider an equivalent closed network consisting of some queue i, $1 \leq i \leq M$, and a composite queue. The service rate $T(n)$ for the composite queue, when there are n customers in it, is set equal to the throughput of queue i in the original network when there are n customers in it and when the mean service time of queue i is reduced to zero. The queue length and queue time distributions of queue i in the equivalent network are the same as in the given network. This follows from Theorems 1 and 2 and the Corollary by renumbering the queues so that queue i is renumbered M.

From the above discussion we can conclude that Norton's theorem holds for closed networks when the subsystem σ is a single queue. The proof for the general case follows the same line and is omitted.

• *Open networks*

Consider an open queuing network consisting of servers with exponential service times and Poisson arrivals (Fig. 10) for which equilibrium conditions exist. Let subsystem σ consist of one or more queues of the given network such that all customers enter subsystem σ at point 1 and leave subsystem σ at point 2. Construct an equivalent open network which consists of a composite Poisson source and subsystem σ, Fig. 11. Let T be the rate at which customers are generated by the composite Poisson source; T is set equal to the throughput of subsystem σ in the given network when the service times of all queues in the subsystem are reduced to zero. The queue length distributions for all queues in subsystem σ in the equivalent network are the same as in the given network. This is stated without proof since the arguments are very similar to those invoked in the case of closed networks.

• *Networks which satisfy local balance*

We extend the results of the previous section to a class of networks which satisfy local balance [2, 3]. We restrict our attention to this class of networks because they form a natural extension, for investigation, to networks studied by Gordon and Newell [4].

In the following we discuss the general queuing networks in which customers from more than one class are being served. In order to be specific about the class of customers being served we have to introduce the term *stage of service*. Stage of service is defined as the ordered-pair server-class. Stage (i, j) implies that server i is serving customers of class j.

From Markov process analysis we know that the number of states and the complexity of the balance equations increase rapidly with the complexity of the queuing networks. In the analysis of networks which satisfy local balance, the problem can be partitioned into stages and the computation is greatly simplified [2, 3].

Basically, networks are said to satisfy local balance if the rate of flow into a state caused by customers *entering a stage of service* is equal to the rate of flow out of this state caused by customers *leaving this stage of service*.

Many networks fall into this category. Chandy et al. [3] show that for networks which satisfy local balance, the steady state probability has the product form of Eq. (27). Consider a closed queuing network with M queues indexed $1, \cdots, M$. Let there be V classes of customers indexed $1, \cdots, v, \cdots, V$. Let $N(v)$ be the total number of customers of class v.

Let $p_{ij}(v)$ be the probability that a customer of class v joins queue j after being served by server i. Let $y_i(v)$, $i = 1, \cdots, M$ and $v = 1, \cdots, V$, be a set of numbers such that

$$y_j(v) = \sum_{i=1}^{M} y_i(v)\ p_{ij}(v) \text{ for all } j,\ v. \tag{25}$$

Let the event that there are $n_i(v)$ customers of class v in queue i be represented by the matrix $\{n_{iv}\}$, $i = 1, \cdots, M$ and $v = 1, \cdots, V$. We say that $\{n_{iv}\}$ is a feasible state if

$n_i(v) \geq 0 \; i = 1,\cdots, M$ and $v = 1,\cdots, V$,

and

$$\sum_{i=1}^{M} n_i(v) = N(v) \text{ for } v = 1,\cdots, V. \tag{26}$$

If $\{n_{iv}\}$ is feasible, the probability of the event $\{n_{iv}\}$ has the product form

$$P[\{n_{iv}\}] = \frac{1}{G} \prod_{i=1}^{M} x_i[n_i(1),\cdots, n_i(V)], \tag{27}$$

where $x_i[\cdot]$ is a function of $n_i(1),\cdots, n_i(V)$ for $i = 1,\cdots, M$ and G is a normalizing constant. In analogy to the previous section, we define a real function g over the matrix $\{n_{iv}\}$ where

$$g[\{n_{iv}\}] = \prod_{i=1}^{M} x_i[n_i(1),\cdots, n_i(V)]. \tag{28}$$

In a manner similar to the single class problem, we define $\mathscr{X}_i$ as an array of dimension V and with respective lengths of components $1+N(1), 1+N(2),\cdots, 1+N(V)$. In other words, for the component v, the index ranges from 0 to $N(v)$. The elements of the array are, respectively, $x_i[n(1), n(2),\cdots n(V)]$.

Consider two arrays $\mathscr{A}$ and $\mathscr{B}$, both of dimension V and of the same component lengths as shown above. Again as in the case of a single class problem, we define the convolution of $\mathscr{A}$ and $\mathscr{B}$ as an array $\mathscr{C}$ with the same dimension and structure as $\mathscr{A}$ or $\mathscr{B}$. The elements of $\mathscr{C}$ are as follows:

$$\begin{aligned} &c[n(1), n(2),\cdots n(V)] \\ &= \sum_{m(V)=0}^{n(V)} \cdots \sum_{m(2)=0}^{n(2)} \cdots \sum_{m(1)=0}^{n(1)} a[m(1), m(2),\cdots m(V)] \\ &\times b[n(1)-m(1), n(2)-m(2),\cdots, n(V)-m(V)]. \end{aligned} \tag{29}$$

We represent the convolution as

$$\mathscr{C} = \mathscr{A} * \mathscr{B}. \tag{30}$$

Define M arrays $\mathscr{G}_i$, $i = 1,\cdots, M$, where

$$\mathscr{G}_i = \mathscr{X}_1 * \cdots * \mathscr{X}_i,\; i = 1,\cdots, M. \tag{31}$$

Note that

$$\mathscr{G}_i = \mathscr{G}_{i-1} * \mathscr{X}_i \text{ for } i = 1,\cdots, M, \tag{32}$$

where $\mathscr{G}_0$ is an array containing all zero elements except for $g_0(0,\cdots, 0) = 1$.

Recall that $\{n_{iv}\}$ is a matrix. Analagous to the expression for the single class problem as shown in Eq. (8), we derive the expression as $G_j[r(1),\cdots r(V)]$, where

$$G_j[r(1),\cdots r(V)] = \sum_{\mathscr{R}_v} g[\{n_{iv}\}], \tag{33}$$

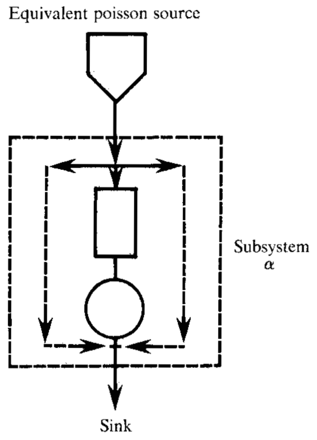


Figure 11 Equivalent open queuing network.

where

$$\mathscr{R}_v = \{\{n_{iv}\} \in \{r(1), r(2),\cdots, r(V)\}\}.$$

The set $\{r(1),\cdots, r(V)\}$ includes those states $\{n_{iv}\}$ which satisfy the defining relation similar to Eq. (8) for every component v; i.e.,

$$\sum_{i=1}^{j} n_i(v) = r(v) \text{ and } n_i(v) = 0 \text{ for } i > j$$

for $v = 1,\cdots V$.

Let $P_i[n(1),\cdots, n(V)]$ be the marginal probability that there are $n(v)$ customers of class v in queue i, $v = 1, \cdots, V$. Using the same arguments as in the previous sections we get

$$\begin{aligned} P_M[n(1),\cdots, n(V)] &= x_M[n(1),\cdots, n(V)] \\ &\times G_{M-1}[N(1)-n(1),\cdots\cdots, N(V)-n(V)]/G. \end{aligned} \tag{34}$$

Let $\mathscr{T}_M(v)$ be the throughput of customers of class v of queue M. By arguments similar to that of the previous section we obtain

$$\begin{aligned} \mathscr{T}_M = y_M(v)\; G_M[N(1),\cdots, N(v-1), N(v)-1, N(v \\ N(v+1),\cdots, N(V)]/G. \end{aligned} \tag{35}$$

It can be shown that Norton's theorem holds for the class of networks which satisfy local balance, using an approach identical to that of the previous section. The statement of Norton's theorem for this class of networks is as follows: For a closed system the service rate for customers of type s in the composite queue, when there are $n(v)$ customers in the composite queue, is set equal to the throughput of customers of class s through the short when there are $n(v)$ customers of class v in the shorted network, $v = 1, \cdots, V$, for any s in $\{1, \cdots, V\}$.

In an open network, all the queues in the system, except in the subsystem under consideration, are replaced by a single composite Poisson source which generates customers of all classes, where each class is generated independently. The rate at which the composite source generates customers of class v is set equal to the throughput of customers of class v through the short of the subsystem under consideration.

Conclusion

We have shown that a theorem analogous to Norton's theorem from the theory of electrical networks holds for a class of queuing networks which satisfy local balance. In certain design problems where a subsystem can be selected for parametric analysis, Norton's theorem can be applied in order to simplify the amount of computation.

Acknowledgments

The authors are grateful to K. Hazeghi for his valuable comments. The authors also express their gratitude to P. E. Green and D. T. Tang for their perceptive readings of the manuscript and many helpful suggestions.

References

1. C. E. Smith, *Communication Circuit Fundamentals*, McGraw-Hill Book Co., Inc., New York 1949.
2. K. M. Chandy, "The Analysis and Solutions for General Queueing Networks," *Proceedings of Sixth Annual Princeton Conference on Information Sciences and Systems*, Princeton University, Princeton, NJ 1972.
3. F. Baskett, K. M. Chandy, R. R. Muntz, and F. Palacios-Gomez, "Open, Closed and Mixed Networks of Queues with Different Classes of Customers," to be published in *J. ACM*.
4. W. J. Gordon and G. F. Newell, "Closed Queueing Systems with Exponential Servers," *Oper. Res.* **15**, 254 (1967).
5. J. Buzen, "Queueing Network Models of Multiprogramming," Ph.D. Thesis, Harvard University, Cambridge, MA 1971.

Received March 15, 1974

K. M. Chandy is located at the University of Texas, Austin, Texas 78712; U. Herzog is at the University of Stuttgart, Stuttgart, West Germany; and L. Woo is at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598.