A METHOD FOR APPROXIMATE ANALYSIS OF GENERAL QUEUEING NETWORKS*

D. Neuse
K. M. Chandy

TR-153 August 1980

* This work was supported in part by NSF Grant MCS74-13302.

ABSTRACT

An approximate analytical method for general queueing networks is presented. Accurate results are obtained by using composite queues which closely capture the behavior of nonlocal balance queues. The method uses an iterative approach which converges rapidly. Arbitrary numbers of queues and job classes are allowed.

KEY WORDS AND PHRASES

Approximate solution, performance evaluation, queueing networks.

1. INTRODUCTION

Queueing network models are widely used in analysing the performance of computing systems: an entire issue of the ACM Computing Surveys [1] was devoted to the subject. Markov models are the most general analytic performance models of computing systems. Unfortunately, Markov models are usually computationally intractable when there are more than two queues in the system. A class of models that are tractable are called product form networks [2]. However, queues in a product form network must have very special service disciplines called local balance disciplines [3], or they must have a very special service time density: the exponential density. Product form networks are very restrictive. For example, a network where jobs are assigned priorities does not have product form. Heuristic techniques must perforce be used to analyze many realistic models [4]. Balbo [5] and Tripathi [6] have excellent, comprehensive studies of approximation techniques, including those of Bard [7], Sauer and Chandy [8], Shum and Buzen [9], and Sevcik, Levi, Tripathi, and Zahoran [10]. Balbo's empirical analysis of several approximation algorithms showed that Marie's approach [11] was one of the best. We show how Marie's approach can be used with priority disciplines and networks with multiple job classes.

2. NORTON'S THEOREM

The algorithm is based on "Norton's theorem" of decomposition / aggregation [12]. Consider a closed local balance network Z with M queues (fig. 1) and K classes of jobs (or customer chains).

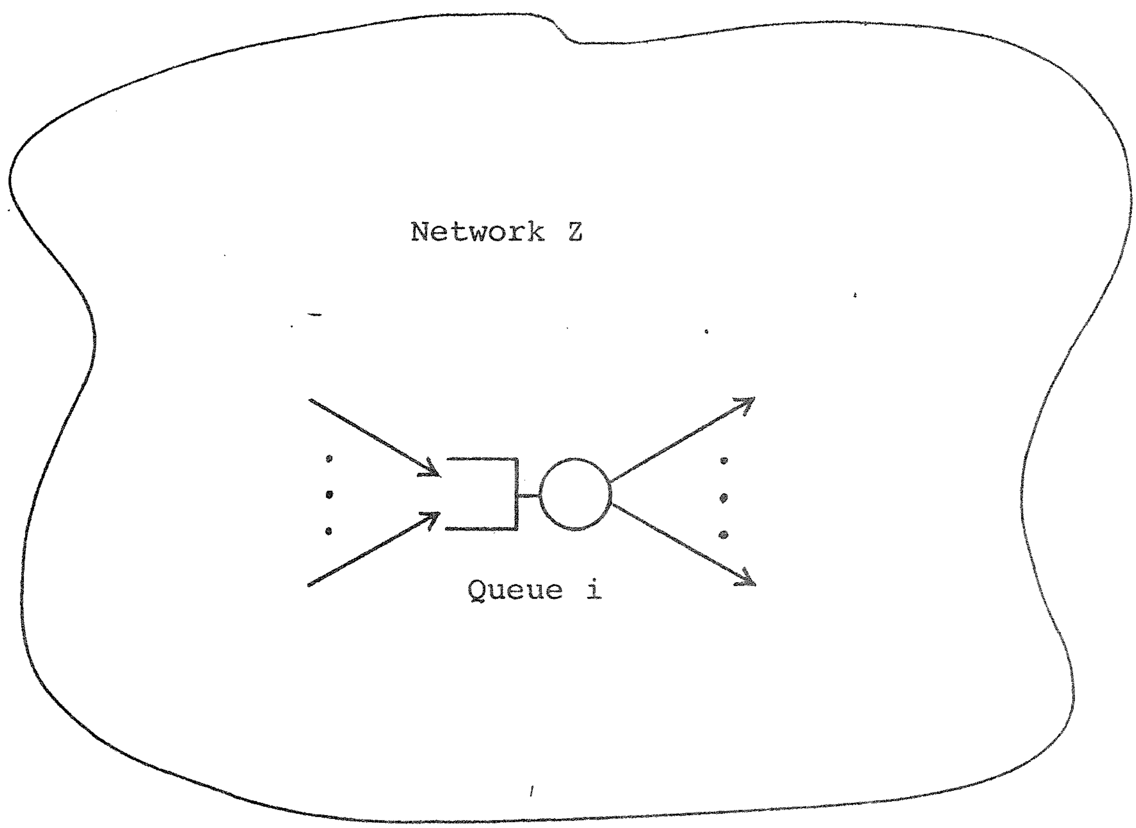

Figure 1.

Assume that there are N_k jobs of class k in the network, k=1,...,K. Suppose we wish to analyze the equilibrium queue-length distribution for some specific queue, say queue i in network Z. It is possible to construct a two-queue network (fig. 2), consisting of queue i and another queue, called the complement of queue i, and with the same population as in the original network, such that the equilibrium queue-length distribution for queue i in the two-queue network is the same as in the original network.

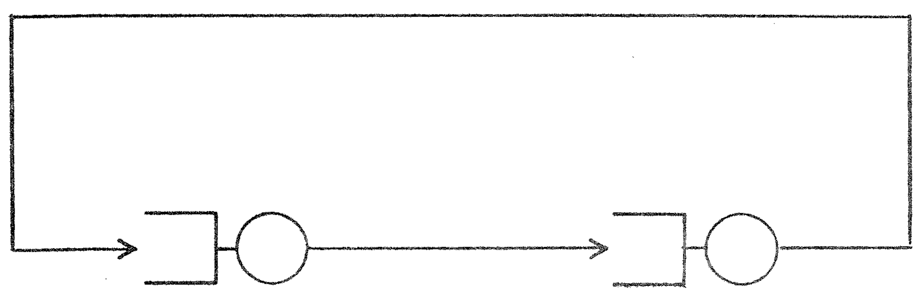

Figure 2.

The complement of queue i in network Z is denoted by COMP(i,Z). COMP(i,Z) is specified by a K dimensional matrix H_i, where the rate of service of class j jobs in the complement, when there are n_k jobs of class k, k=1,...,K, is $H_i(n_1,...,n_{j-1},n_j-1,n_{j+1},...,n_K)/H_i(n_1,...,n_K)$.

The class of decomposition methods of Marie [11], Chandy, Herzog and Woo [13], and others (see Balbo [5] or Tripathi [6]) are based on Norton's theorem.

3. ALGORITHM

Given a network A:

Let Z be the network obtained by assuming all queues in A have local balance.

1. For each queue j,
   (a) construct a two-queue network consisting of the <u>original</u> queue j and COMP(j,Z);
   (b) compute the equilibrium probability $P[j,\bar{n}]$, where $\bar{n} = (n_1,\ldots,n_K)$, of n_k jobs of class k, k=1,...K in queue j of this two-queue network by solving Markov balance equations;
   (c) construct (section 4) a local balance queue Q'(j), such that a two-queue network of Q'(j) and COMP(j,Z) yields the same $P[j,\bar{n}]$ as the network of Q(j) and COMP(j,Z).
2. Define a local balance network Z' obtained by replacing Q(j) in A by Q'(j) for every queue j.

   If the statistics for Z and Z' are close (as defined in [13]), assume that the statistics for Z' are satisfactory approximations.

   Otherwise, set Z ← Z' (i.e., call network Z' by Z) and go to step 1.

4. CONSTRUCTING THE LOCAL BALANCE EQUIVALENT QUEUE Q'(J) (step 1c of section 3)

Q'(j) is specified by a matrix A_j, where A_j has the same interpretation H_i, and [12]

$$P[j,\bar{n}] \propto A_j(\bar{n})\ H_j(\bar{N} - \bar{n}). \qquad (1)$$

Hence

$$A_j(\bar{n}) \propto P[j,\bar{n}] \ / \ H_j(\bar{N} - \bar{n}). \qquad (2)$$

Any proportionality constant can be used to compute $A_j(\bar{n})$ from eqn. 2, and the proportionality constant for eqn. 1 can be determined from

$$\sum_{\bar{n}} P[j,\bar{n}] = 1.$$

5. RESULTS

Fifty-six networks were analyzed with queues having the following disciplines: first-come-first-served, preemptive priority, nonpreemptive priority, processor sharing, and infinite server. Results obtained were uniformly good, and convergence was rapid. The algorithm typically required two or three iterations at a termination tolerance of .01. Only six networks required more than four iterations, and these were networks with more than one heavily loaded priority queue.

A histogram of tolerance errors for the algorithm presented here (MCOMP) and for a processor-sharing approximation (see section 6) as measured against simulation results are presented below (fig. 3). Tolerance is defined as the maximum of the errors in utilization$_{m,k}$, queue length$_{m,k}$ / N_k, and wait time$_{m,k}$ / cycle time$_{m,k}$, over all queues and classes m,k. The detailed results will be found in a subsequent report.

```
                      MCOMP                                Processor Sharing

    Interval
    beginning
    value

    0.00  ****************
    0.01  *****************                               *
    0.02  ********                                        *
    0.03  ******
    0.04  ****                                            *
    0.05  **                                              **
 T  0.06  **                                              ***
 O  0.07                                                  *
 L  0.08                                                  *
 E  0.09                                                  **
 R  0.10  *                                               *******************
 A  0.20                                                  ***
 N  0.30                                                  ********
 C  0.40                                                  *
 E  0.50                                                  ***
    0.60                                                  ***
    0.70                                                  *
    0.80
    0.90
    1.00                                                  ****
    2.00
```

Figure 3.

6. EXAMPLE

The following network (fig. 4) was analyzed by simulation, by the algorithm described here (MCOMP), and by assuming processor-sharing at the priority queue.

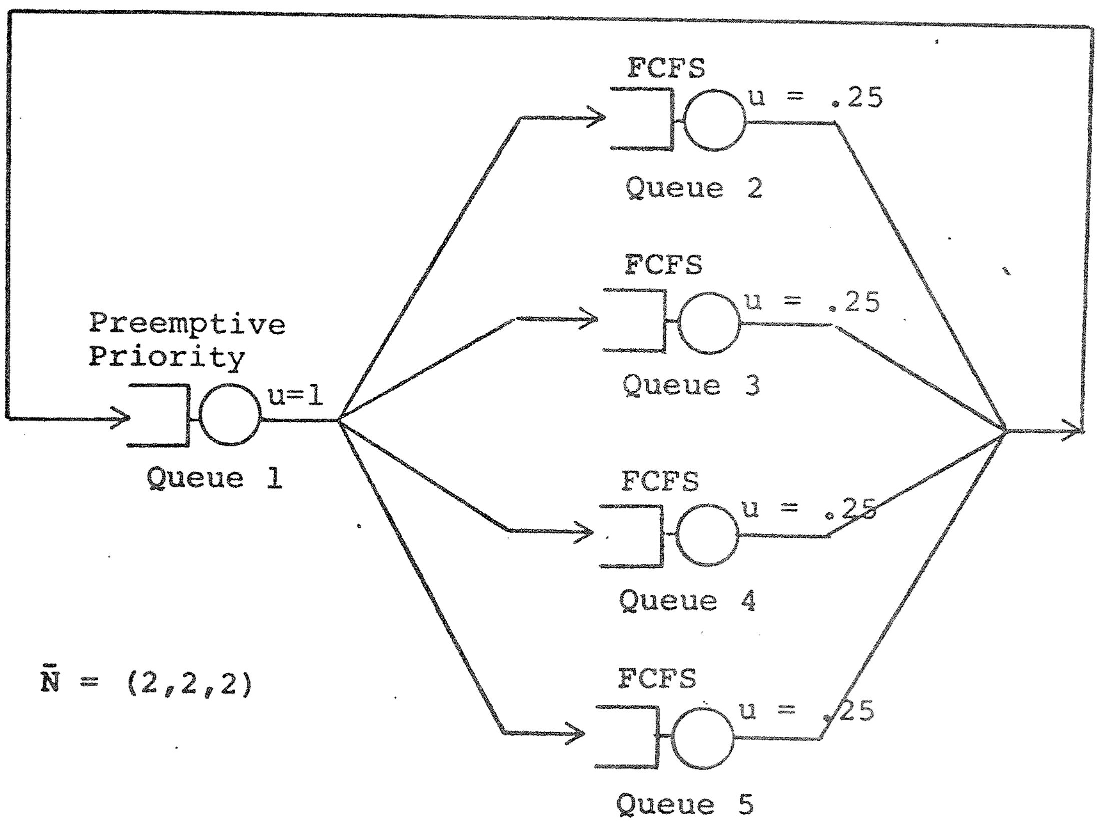

Figure 4.

The measurements obtained from these analyses are given below (fig. 5).

MCOMP: Termination tolerance = .001. Number of iterations = 2.

Simulation: Number events = 100000. Simulated time = 83445 sec.

	Queue	Class	Simul. Meas.	MCOMP Meas.	Diff.	Proc. Meas.	Sharing Diff.
Utilization	1-5	1	.232	.223	.009	.200	.032
		2	.201	.202	.001	.200	.001
		3	.166	.175	.009	.200	.034
Queue Length	1	1	.262	.254	.008	.400	.138
		2	.382	.382	.000	.400	.018
		3	.543	.565	.022	.400	.143
	2-5	1	.435	.437	.002	.400	.035
		2	.405	.405	.000	.400	.005
		3	.364	.359	.005	.400	.036
Throughput (jobs/sec.)	1	1	.232	.223	.009	.200	.032
		2	.201	.202	.001	.200	.001
		3	.166	.175	.009	.200	.034
	2-5	1	.0580	.0558	.0022	.05	.0080
		2	.0503	.0504	.0001	.05	.0003
		3	.0415	.0438	.0023	.05	.0085
Wait Time (sec.)	1	1	1.13	1.14	.01	2.00	0.87
		2	1.90	1.89	.01	2.00	0.10
		3	3.27	3.23	.04	2.00	1.27
	2-5	1	7.50	7.82	.32	8.00	0.50
		2	8.06	8.02	.04	8.00	0.06
		3	8.78	8.20	.58	8.00	0.78

Maximum Differences

	Diff.	Sharing Diff.
Utilization	.009	.034
Queue Length / N_k	.011	.072
Wait Time / cycle $\text{time}_{m,k}$	.012	.105
Overall (Tolerance)	.012	.105

Figure 5.

1. Computing Surveys vol.10,3 (Sept. 1978).

2. Chandy K.M. "The Analysis and Solutions for General Queueing Networks," Proc. Sixth Annual Princeton Conf. on Inform. Sci. and Systems, Princeton U., Princeton, N.J. March 1972, 219-224.

3. Basket F., Chandy K.M., Muntz R.R., and Palacios F.G. "Open, Closed and Mixed Networks of Queues with Different Classes of Customers," J.ACM vol.22,2 (April 1975), 248-260.

4. Chandy K.M. and Sauer C.H. "Approximate Methods for Analyzing Queueing Network Models of Computing Systems," Computing Surveys vol.10,3 (Sept. 1978), 281-317.

5. Balbo G. "Approximate Solutions of Queueing Network Models of Computer Systems," Ph.D. Thesis, Purdue University, December 1979.

6. Tripathi S.K. "On Approximate Solution Techniques for Queueing Network Models of Computer Systems," Ph.D. Thesis, University of Toronto, 1979.

7. Bard Y. "Some Extensions to Multiclass Queueing Network Analysis," Proc. 4-th International Symposium on Modelling and Performance Evaluation of Computer Systems, Vienna, Austria, (February 1979).

8. Sauer C.H., and Chandy K.M. "Approximate Analysis of Central Server Models," IBM Journal of Research and

Development vol. 3, (May 1975), 301-313.

9. Shum A.W.C., and Buzen J.P. "The EPF Technique: a Method for Obtaining Approximate Solutions to Closed Queueing Networks with General Service Times," Proc. 3rd International Symposium on Modeling and Performance Evaluation of Computer Systems, North-Holland Publishing Co., (October 1977).

10. Sevcik K.C., Levy A.I., Tripathi S.L., and Zahorjan J.L. "Improving Approximations of Aggregated Queueing Network Subsystems," Proc. International Symposium on Computer Performance Modeling, Measurement, and Evaluation, North-Holland Publishing Co., (August 1977), 1-22.

11. Marie R.A. "An Approximate Analytical Method for General Queueing Networks," IEEE Transactions on Software Engineering vol.SE-5, 5 (Sept. 1979), 530-538.

12. Chandy K.M., Herzog U., and Woo L.S. "Parametric Analysis of Queueing Networks," IBM Journal of Research and Development vol. 19, (Jan. 1975), 36-42.

13. Chandy K.M., Herzog U., and Woo L.S. "Approximate Analysis of General Queueing Networks," IBM Journal of Research and Development vol.19, (Jan, 1975), 43-49.