1979

Alternative Parameterizations of Product-Form Queueing Networks

James P. Bouhana

Report Number:
79-320

CSD-TR-320

ALTERNATIVE PARAMETERIZATIONS OF PRODUCT-FORM QUEUEING NETWORKS

James P. Bouhana

Purdue University
Computer Sciences Department
West Lafayette, IN 47907

November, 1979

ABSTRACT

Product-form queueing networks are considered which allow for conceptual job initiations and terminations (such as a central server model). It is shown that the product-form expression can be parameterized with either mean job resource usages, total server busy times, or overall server utilizations. Measurement and computational efficiencies resulting from these alternative parameterizations are discussed. None of the attenatives presented require that mean per-request service times or inter-server routing frequencies be determined.

1. INTRODUCTION

The customary specification of an M-server queueing network's product-form solution is expressed as a function of M parameters: $F(X_1, X_2, \ldots, X_M)$, where each parameter depends on a server's mean per-request service time and the inter-server routing frequencies. Since the basic quantities upon which the parameters depend may be difficult to easily or accurately measure, alternative solutions parameterized with more tractable data are desirable. Subsequent sections discuss three such alternatives, based on mean per-job resource usage, total server busy time, and server utilization. In addition to requiring less measurement work, the alternative parameterizations require less computations than is customary.

Two examples of the types of network considered are shown below.

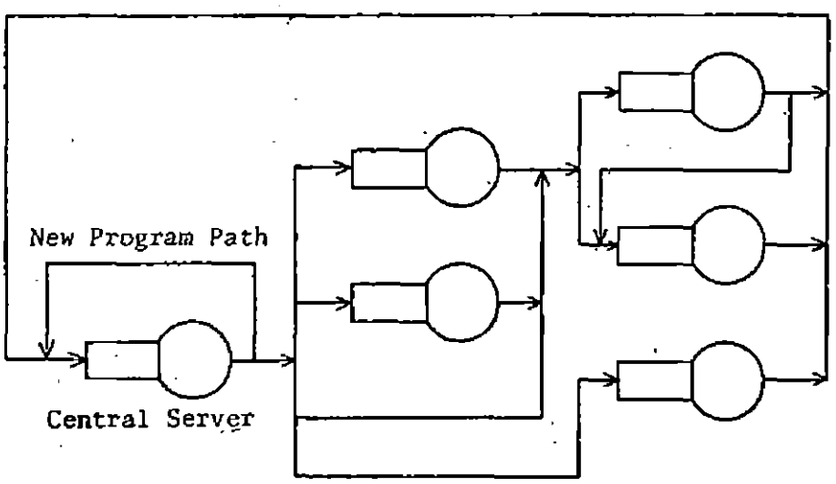

Figure 1: A Closed Centralized Network

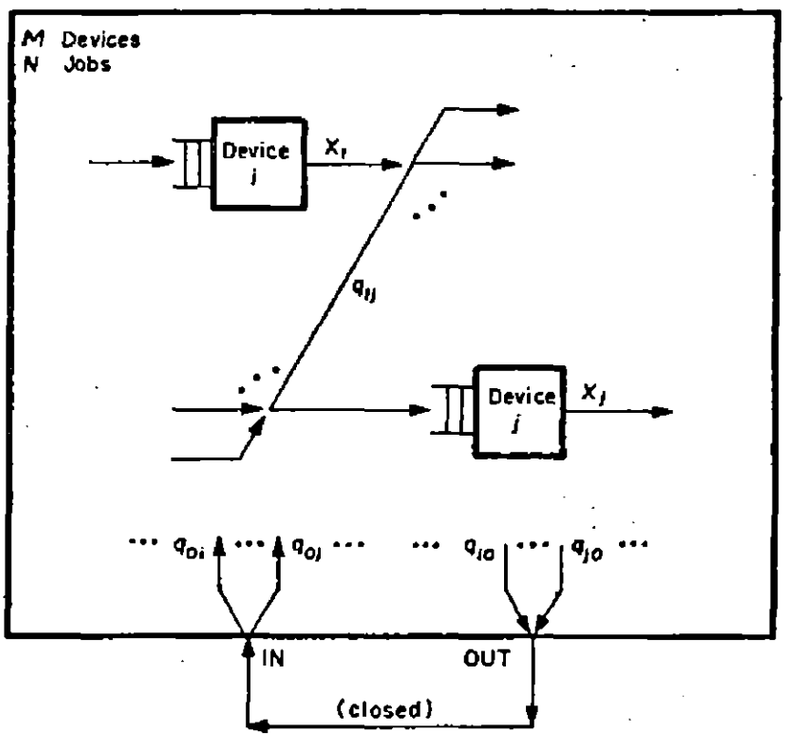

Figure 2: Two Devices in a Network with a Single Input/Output Port

Figure 1 depicts a closed network containing a fixed number of jobs. Job initiations and terminations can be conceived as occurring whenever a job traverses a self-loop on a designated server. Closed networks having this property are central server models [BUZEN71] and centralized models [BOUHA78]. The designated server is called the "central server" and it usually represents the CPU, modeling the real-life situation in which a CPU is the first and last server utilized by a job. An "open" centralized network is shown in Figure 2; jobs may exit the network from any server but they immediately reenter the network as a new job. The number of jobs in the network itself is thus constant.

The restriction to networks containing a fixed number of customers does not preclude modeling actual systems in which the multiprogramming level varies. Most of the performance measures discussed here are stated in terms of the total number of jobs, and this parameter can be varied as needed. Also, the other measures are stated as measurements made over an entire interval of observation -- they are insensitive to the proportions of time that specific multiprogramming levels persist.

After having completed service at server i, a job next visits server j with routing frequency q_{ij}. The mean per-request service time of the i-th server is S_i. Servers are assumed to be load-independent; their mean service times do not vary according to their queue length. All servers are simple, having only one unit of a resource present. For brevity, simple load-independent servers will be called SLI servers

We assume that a network satisfies either the stochastic assumptions [BASKE75] or the operational assumptions [DENNI78], [BOUHA78] required for a product-form solution. (A network's solution is its distribution of customer apportionments $(n_1, n_2, \ldots, n_M) = \underline{n}$ where M is the number of servers and n_i is the number of jobs either enqueued or in service at the

i-th server. For closed networks, the sum of the n_i ($1 \leq i \leq M$) is constant).

As stated by Gordon and Newell [GORDO67], the solution for an M-server networks having SLI servers and containing N customers is:

$$p(\underline{n}) = F(X_1, X_2, \ldots, X_M) = \prod_{i=1}^{M} (X_i)^{n_i} / G(N) \qquad (1)$$

where the X_i depend upon S_i and q_{ij} as follows

$$S_j^{-1} X_j = \sum_{i=1}^{M} S_i^{-1} X_i q_{ij} \quad (1 \leq j \leq M) \qquad (2)$$

Equation (2) is an eigenvector equation for the matrix of routing frequencies. Using an eigenvector routine especially suited to work with routing matrices [ROBIN70], all X_i can be determined in $o(M^2)$ operations. In Equation (2), G(N) is a normalizing constant defined as:

$$G(N) = \sum_{n \in S(N,M)} \prod_{i=1}^{M} (X_i)^{n_i}$$

where S(N,M) is the set of all customer apportionment states. The constant G(N), as well as the intermediate values G(1), ..., G(N-1) can be computed in $o(2NM)$ operations [BUZEN73].

Since Equation (2) is an eigenvector equation, it has infinitely many solutions. Thus, any constant multiple of the X_i ($1 \leq i \leq M$) is also a solution parameter. This may be verified by noting that any such constant, c, appears on both sides of Equation (2); it thus concels. Subsequent derivations utilize the fact that:

$$F(X_1, X_2, \ldots, X_M) = F(cX_1, cX_2, \ldots, cX_M).$$

1.1 Marginal Distributions

Expressions for several marginal distributions of usual interest to performance analysts are derivable from the solution. The following expressions were derived by Buzen [BUZEN73]. They give the proportion of time that a condition exists, contingent on having N customers in a network:

Minimum number of jobs: The proportion of time that there are at least k jobs present at the i-th server (both enqueued and in service) when there are N jobs in an M-server network is:

$$P(n_i \geq K, N) = (X_i)^k \frac{G(N-k)}{G(N)} \tag{3}$$

Server utilization is given by the proportion of time that there is at least one job at the server:

$$P(n_i \geq 1, N) = (X_i) \frac{G(N-1)}{G(N)} \tag{4}$$

Exact number of jobs. The proportion of time that there are exactly k jobs at the i-th server is:

$$P(n_i = k, N) = \frac{(X_i)^k}{G(N)} (G(N-k) - (X_i)G(N-k-1)) \tag{5}$$

Mean number of jobs. The mean number of jobs both enqueued and in service at the i-th server is:

$$E(n_i,N) = \sum_{k=1}^{N} (X_i)^k \frac{G(N-k)}{G(N)} \tag{6}$$

Server overlap. The proportion of time that the i-th server and the j-th server are simultaneously busy is:

$$P(n_i \geq 1 \ \& \ n_j \geq 1, N) = X_i X_j \frac{G(N-2)}{G(N)} \tag{7}$$

The above expressions are stated only in terms of the solution parameters and the G(i) values ($1 \leq i \leq N$) that are functions of them. Thus, any network solution stated in terms of X_i that are directly measurable obviates both the need to solve an eigenvector equation and the need to determine mean service times and routing frequencies.

2. MEAN RESOURCE USAGE

Let R_i^k denote the service time delivered by the i-th server to the k-th job totaled over all visits that the k-th job makes to server i. The quantities R_i^k represent the resource usages of jobs for servers; on most medium to large scale computers, resource usages are reported in the job accounting log that is maintained by the operating system. Mean resource usages are straightforwardly calculated given per-job resource usages and the number of completed jobs. Letting K denote the number of job completions and R_i denote the mean resource usage of jobs for server i, we have:

$$R_i = \frac{1}{K} \sum_{j=1}^{K} R_i^j \tag{8}$$

Mean resource usages are also representable in terms of the mean number of visits, V_i, that a job makes to server i, and the mean per-visit service time. That is,

$$R_i = V_i S_i$$

Mean visit counts also satisfy the eigenvector Equation (2) [KLEIN75]; that is:

$$V_j = \sum_{i=1}^{M} V_i \, q_{ij} \quad (1 \leq j \leq M)$$

Since there is but one independent vector satisfying Equation (2), we necessarily have:

$$V_i = c \, S_i^{-1} \, X_i$$
$$X_i = \frac{1}{c} V_i \, S_i = \frac{1}{c} R_i \quad (1 \leq i \leq M)$$

This last equation shows that mean resource usages can be used to parameterize a queueing network solution:

$$p(\underline{n}) = F(R_1, R_2, \ldots, R_M) = \frac{1}{G(N)} \prod_{i=1}^{M} (R_i)^{n_i}$$

The validity of using mean resource usages as solution parameters has been recognized before although it has not widely been noted in the literature. Some commercially available modeling software use mean resource usages obtained from accounting data [BUZEN78].

3. TOTAL SERVER BUSY TIME

Let $S_i(*)$ denote the total time that the i-th server is busy during an interval of observation. For some servers, total busy time may be more easily determined than per-job resource usage, especially if the usage of a server is not reported in the accounting log. An example is a software queue,

such as that induced by a serially reusable, non-reentrant routine of an operating system (e.g. a file assignment or a directory update routine). The usage of these types of servers can be measured by clocking on and off at the instances of service initiation and termination. Clocking can be done by the operating system or by an event-driven monitor. Aggregating the clocked intervals yields total busy time. Also, some hardware devices, especially CPU's and peripheral units, have digital clocks which record the device's busy time.

By definition, total busy time is the sum of resource usages over all jobs, K:

$$S_i(*) = \sum_{k=1}^{K} R_i^k \qquad (1 \leq i \leq M)$$

Combining the above equation with Equation (8):

$$S_i(*) = K\, R_i \qquad (1 \leq i \leq M)$$

Since the $S_i(*)$ are constant multiples of the R_i, they constitute valid solution parameters:

$$p(\underline{n}) = F(S_1(*), S_2(*), \ldots, S_M(*)) = \frac{1}{G(N)} \prod_{i=1}^{M} (S_i(*))^{n_i}$$

4. SERVER UTILIZATION

Let $U_i(k)$ denote the utilization of server i when there are k jobs in a network. In contrast to the previously discussed measures, utilizations

depend upon a system's total job loading. (For closed models applied to real systems in which the multiprogramming level (MPL) varies, performance measures are derived by taking the performance measures calculated with a fixed-load model and appropriately weighting them according to the proportion of time that each MPL occurs.) However, we can define an overall server utilization, independent of mix level, during an observation period of length T to be the ratio of total server busy time and T:

$$U_i(*) = S_i(*)/T$$

For some performance data collection situations, utilizations may be the most readily available measure. An example is a hardware monitor that displays device utilization as an overall average computed over a moving time interval window. The average of all of the moniter's observations yields the overall utilization for an entire interval of observation.

Since $U_i(*)$ is a scalar multiple of the previously established solution parameter $S_i(*)$, we immediately have:

$$p(\underline{n}) = F(U_1(*), U_2(*), \ldots, U_M(*)) = \frac{1}{G(N)} \prod_{i=1}^{M} (U_i(*))^{n_i}$$

The concept of server utilization is applicable to a wider class of networks than is discussed in previous sections. The restriction to a centralized network can be lifted. If Y_i $(1 \leq i \leq M)$ denote relative server throughputs (request completions divided by the length of an observation interval), then a product-form network satisfies:

$$Y_j = \sum_{i=1}^{M} Y_i \, q_{ij} \qquad (1 \leq j \leq M)$$

Using the relation, $Y_j = c\, S_j^{-1}\, X_j$, we get:

$$X_j = \frac{1}{c}\, Y_j\, S_j$$

By the Utilization Law [DENNE78], the quantity $Y_j\, S_j$ in the above equation is the overall utilization $U_j(*)$.

4.1 Marginal Distributions

Parameterizing a product form solution with utilizations is especially attractive, since the expressions for marginal distributions (Equations (3) - (7)) may be reformulated to include utilizations only. To see this, we note that for general parameters X_i $(1 \leq i \leq M)$, we have from Equation (3) the following utilization expression for a network containing N customers:

$$U_i(N) = (X_i)\, \frac{G(N-1)}{G(N)}\,.$$

Rearranging terms yields the following expressions for G(1), G(2), ..., G(N):

$$G(1) = \frac{(X_i)G(0)}{U_i(1)} = \frac{X_i}{U_i(1)}$$

$$G(2) = \frac{(X_i)G(1)}{U_i(2)} = \frac{(X_i)^2}{U_i(1)U_i(2)}$$

$$\vdots \qquad\qquad \vdots \qquad\qquad \vdots$$

$$G(N) = \frac{(X_i)G(N-1)}{U_i(N)} = \frac{(X_i)^N}{\prod_{j=1}^{N} U_i(j)}$$

Substituting this last equation for G(N) into Equations (3), (5)-(7) yields:

Minimum number of jobs.

$$P(n_i \geq k, N) = \prod_{j=N-k+1}^{N} U_i(j)$$

Exact number of jobs.

$$P(n_i = k, N) = \left[\prod_{j=N-k+1}^{N} U_i(j) \right] \cdot [1 - U_i(N-k)]$$

Mean number of jobs.

$$E(n_i, N) = \sum_{k=1}^{N} \prod_{j=N-k+1}^{N} U_i(j) = \sum_{k=1}^{N} \prod_{j=k}^{N} U_i(j)$$

Server overlap.

$$P(n_i \geq 1 \ \& \ n_j \geq 1, N) = \frac{X_i}{X_j} U_i(N) U_i(N-1)$$

$$= \frac{X_j}{X_i} U_j(N) U_j(N-1)$$

Note that only the expression for server overlap contains terms representing the original solution parameters. For this expression, the ratio of either mean resource usages, total server busy times, or overall server utilizations may be used -- these ratios have the same value for any pair of servers.

4.2 Computational Efficiency

If the stratified utilizations $U_i(j)$ $(i \leq i \leq M)$, $(j \geq 1)$ are known, then all marginal distributions discussed above, except server overlap, can be computed directly. If the proportion of time that each MPL persists is also known, then $U_i(*)$ can be determined, allowing computation of server overlap. Equivalently, if the R_i or $S_i(*)$ are known in addition to stratified server utilizations, then all marginal distributions discussed above can be determined. For these situations, there is no need to compute values of the normalizing constant G(N).

Note that, except for server overlap, the marginal distribution expressions depend only upon performance measures for a single server. Thus an analyst wishing to study a single device may find it attractive to invest in the appropriate instrumentation (e.g. a hardware monitor) needed to measure the stratified utilizations for the device of interest. Stratified utilizations may also be approximated by accounting log data, provided that the initiation and termination times of jobs is recorded in addition to their resource usages [BOUHA76].

By examining the product form appearing in the mean queue length expression (Equation (9)), one notes that it can be calculated using only one multiplication for each iteration of the summation. Also, intermediate values of that product form appear as subexpressions for the other marginal distributions. Thus knowledge of stratified utilizations permits calculating these marginal distributions for any specific server is o(N) operations.

If the stratified utilizations are not known, then they may be calculated from overall utilizations. A procedure for doing this is given in [BOUHA78]. Its computational complexity is o(2NM), which is the same as is required for computing G(N) values. The array space required for computing stratified

utilizations is 2M+N, whereas the G(N) computation requires only M+N array locations (including the space required for the parameters).

5. CONCLUSIONS

Depending on which performance measures are most easily available, a performance analyst may wish to use mean resource usages, total server busy times, or overall utilizations as solution parameters to a queueing network model. All of these measures are more computationally efficient to use in calculating marginal distributions (compared with using mean per-request service times and routing frequencies), since an eigenvector equation need not be solved. If stratified utilizations are determined, then a device can be studied in isolation and the computation of its marginal distributions is further simplified.

REFERENCES

[BASKE75] Baskett, F., Chandy, K.M., Muntz, R.R., and Palacios, F.G., "Open, Closed, and Mixed Networks of Queues with Different Classes of Customers," Jour. ACM, Vol. 22, No. 2, (April, 1975), 248-260.

[BOUHA76] Bouhana, J. "A Family of Mix Characterisitics Curves," Proc. 12-th Computer Performance Evaluation Users Group Meeting, San Diego, Calif., (Nov., 1976), 181-188.

[BOUHA78] Bouhana, J.P. "Operational Aspects of Centralized Queueing Networks," Ph.D. Thesis, Univ. of Wisconsin-Madison, (January, 1978), (also U.W. Academic Computing Center Report TR-50, January, 1978).

[BUZEN71] Buzen, J.P. "Queueing Network Models of Multiprogramming," Ph.D. Thesis, Div. of Engineering and Applied Physics, Harvard Univ., Cambridge, Mass., (May, 1971), (also NTIS Report AD731575, August, 1971).

[BUZEN73] Buzen, J.P., "Computational Algorithms for Closed Queueing Networks with Exponential Servers," Comm. ACM, Vol. 16, No. 9, (Sept., 1973), 527-531.

[BUZEN78] Buzen, J.P., et al. "BEST/1--Design of a tool for Computer System Capacity Planning," Proc. 1978 AFIPS National Computer Comb., Vol. 47, AFIPS Press, Montvale, N.J., 447-455.

[DENNI78] Denning, P.J., and Buzen, J.P. "The Operational Analysis of Queueing Network Models," Computing Surveys, Vol. 10, No. 2, (Sept., 1978), 225-261.

[GORDO67a] Gordon, W.J., and Newell, G.F. "Closed Queueing Systems with Exponential Servers," Operations Research, Vol. 15, No. 2, (April, 1967), 254-265.

[KLEIN75] Kleinrock, L., Queueing Systems, Vol. 1, John Wiley and Sons, N.Y., (1975).

[ROBIN70] Robinson, S.M., and Nickel, K. "Computation of the Perron Root and Vector of a Non-negative Matrix," Mathematics Research Center Technical Summary Report No. 1100, Univ. of Wisconsin-Madison, (Sept., 1970).