Proceedings of the 1983
Winter Simulation Conference
S. Roberts, J. Banks, B. Schmeiser (eds.)

MODELING OF QUEUING NETWORKS WITH FINITE CAPACITY QUEUES

Bahram Keramati and Joan M. Lommel
Corporate Research and Development
General Electric Company
Schenectady, NY 12301

A model is presented for the blocking phenomena associated with queuing networks with finite capacity queues. The basic idea behind the model is the calculation of the average length of time the cells are in the "blocked" state, i.e., their downstream queue is filled to capacity. The model was tested and verified against an actual simulation for a simple network. Excellent agreement for the average quantities is demonstrated between the model and the simulation for both the downstream and upstream effects resulting from blocking. The model is quite general in its applicability to M/M/c queuing networks and is easy to implement.

1. INTRODUCTION

Automation, emphasis on product quality and competitive pressures are among the contributing factors leading to the recent emphasis on manufacturing productivity. The design of modern factories and the reconfiguration of existing facilities must consider much more stringent productivity criteria than the standards of past decades if they are to lead to viable systems in the future. The flexibility that exists in the operation of a factory which, to a large extent, depends on human labor does not exist in the automated factory of the future. As a result, much more planning and front-end engineering design is required for the successful implementation of manufacturing systems than in the past. Simulation is gradually becoming a very well recognized planning and design tool for manufacturing engineering. Developments in computer science, computer hardware including personal computers and computer graphics have all contributed to the added attention that simulation has been receiving from the manufacturing community.

A very important consideration in simulations of manufacturing systems is the proper level of detail that should be represented. Many versatile simulation languages exist today which allow almost anything to be represented in a simulation. This can lead to the design of very unwieldy computer programs of little value in answering the questions faced by the manufacturing engineer. In many cases, especially at the very early stages of factory planning and design, the basic data (e.g., how many machines, type of transporters, storage capacities, etc.) are not even known to allow the construction of a simulation. What is needed at this stage of planning is an analysis tool that allows the study of alternatives in fairly broad terms. Although it is possible to use simulation in such instances, the proper design of a reconfigurable tool allowing the comparative study of alternatives is not an easy or efficient way of addressing the need.

Although queuing theory is a very mature subject in operations research, it has generally been a neglected tool in the area of manufacturing engineering. This negligence is partly due to the general tendency in manufacturing engineering against the use of analytical tools, and also due to the lack of interest in manufacturing problems on the part of queuing theory experts. The development of CAN-Q (Solberg, 1977) has been generally neglected by the manufacturing community except in rare instances, although the validity of the approach for aggregate performance predictions has been demonstrated. The major strength of CAN-Q, and queuing theory in general, is in its aggregate, time-averaged predictive ability, the basic features required of a factory planning tool. Yet, it is not uncommon to encounter expert factory planners and designers who are not aware of the existence of such analytical tools.

As a predictive tool for manufacturing systems, queuing theory is not devoid of limitations. For example, it is not possible to derive much useful information from queuing theory regarding the dynamic performance of systems. Moreover, very small systems, or systems in which most events are highly deterministic in nature, and the representation of limited storage buffers are instances where the use of state-of-the-art queuing theory must be approached with caution. Of these, the case of limited storage buffers is probably one of the most important unsolved problems in queuing theory.

It is perhaps appropriate to point out that queuing theory has many applications beyond manufacturing systems. For

CH1953-9/83/0000-0207 $01.00

example, it is quite often used in the analysis of communication and data networks where one is concerned with units of information being processed, as opposed to parts in a manufacturing system. The limitation of infinite storage in practical queuing approaches is not as serious a problem for communication networks as in manufacturing systems, as we shall discuss later.

The purpose of this paper is the presentation of a model for the blocking phenomena that result from the existence of limited storage in manufacturing systems. The theoretical background of the model, as well as previous modeling approaches will be discussed, followed by the presentation of the present model. Comparison of the model results and the results of a simulation will be given both for the upstream and downstream phenomena that can result from limited storage buffers in manufacturing systems. The paper will be concluded by suggestions for future research in this area.

2. BACKGROUND

In this paper we are concerned only with the subset of queuing networks where the arrival is a Poisson process and the service is exponential. For such systems, the theorems of Burke (1956) and Jackson (1957) are of particular importance. Burke proved that the departure process for a queue with the above characteristics (i.e., M/M/1) is also a Poisson process with the same mean as that of the arrival. Burke's results are generalizable to a queue with s servers, i.e., an M/M/s queue. Jackson further showed that in a network of open M/M/s queues ("open" signifying that each node in the system could receive arrivals from any other node, including from outside the system, and generate departures likewise), each node behaves as an independent M/M/s queue with a net arrival equaling the sum of the arrivals to that node. Jackson's theorem has profound implications in the application of queuing theorem to many manufacturing systems; in fact, he was motivated by the need for an analysis tool applicable to machine shops.

Unfortunately, the direct use of Jackson's theorem has not led to much success in the analysis of manufacturing systems. Solberg (1977) demonstrated that a very useful analysis tool for manufacturing systems can result if a closed network of M/M/s queues is considered. In this context, "closed" signifies that the total number of entities in the system remains a constant, implying that a net departure from the system is balanced by a net arrival at the same instant of time. Although this restriction may seem to be limiting in its application potential, it has been demonstrated that in many situations such an approach (e.g. CAN-Q) leads to surprisingly accurate results (Solberg, 1977).

Several explanations can be put forth, none of which theoretically convincing, as to the reason for the success of the closed network in the face of the apparent failures of the open approach of Jackson. It is suggested, for example, that in any real manufacturing system, some control is always exercised in limiting the total number of entities in the system, and, at the same time, not allowing the system to become completely depleted. A Jacksonian network, suffers from the limitation of not limiting the entities anywhere in the system. The apparent limitation of the closed approach in fixing the total number of entities at a definite value (as opposed to the real situation that only limits the total number) seems to be outweighed by the fact that the number cannot be unreasonably exceeded. In addition, the closed approach can allow the study of a given system at various levels of congestion, a very appealing feature for many manufacturing systems.

One of the limitations of the closed queuing approach is that external arrivals and departures cannot occur arbitrarily. It can be argued that if an approach existed to account for the real effects of limited storage space in a Jacksonian framework, a very general and versatile tool for the analysis of arbitrarily complex networks of M/M/s/K (K being the storage limit at each node) could result.

It is very important to cite what is meant here by the real effects of limited storage space. In manufacturing systems, limited storage space leads to the phenomenon known as blocking. Blocking of a resource (or a server) implies that the resource is prohibited from processing any additional entities until downstream storage space is available for the entity currently in the resource. This may seem rather obvious to many manufacturing engineers, but it is perhaps instructive to consider that in communication networks, the encounter of a filled buffer by an entity (or a unit of information) often leads to the total destruction of the entity, signifying a lost message. An M/M/s Jacksonian network can easily represents this situation by allowing a fraction of the parts arriving at any node to be either diverted to another node or to be completely lost to the system by perhaps leaving the network altogether. This fraction is represented by the probability that the node contains more than a specific number of entities, determined by the actual limit that exists on the node's buffer size. In manufacturing situations, parts are seldom lost (intentionally!) due to limited space. What is needed in this case is an approach that will lead to the blockage of the feeding node upon the filling of the present node's storage buffer.

Attempts at modeling blocking in the context of queuing networks have generally not been fruitful. Bell (1982) discusses the modeling approaches proposed by Boxma and Konheim (1981), Takahashi, Miyahara and Hasegawa (1980) and Hillier and Boling (1967). He concludes that although the models may be useful in some instances, especially in cases where the system under consideration is near balanced conditions, they may lead to unrealistically high throughput rates when service rates vary across the network nodes. The authors suspect that many other modeling approaches to blocking have been attempted. The following section describes a model that seems to capture the true effects of blocking in a fairly simple manner.

3. MODEL FOR BLOCKING

The central subject of this paper is the modeling of blocking due to finite queue capacities. Blocking is modeled by increasing the mean service time of the node upstream of the finite queue by a mean blocking time which is a function of queue capacity, arrival rate, and service rate of the finite node. The M/M/s queue with a finite length has been studied extensively with analytical results for utilization, queue lengths, and expected waiting times presented in the literature, Kleinrock (1975), Hillier (1980). Yet the study of a network of several M/M/s/K queues has not led to a practical modeling tool for the analysis of manufacturing systems. This model makes use of the analytical results for a single M/M/s/K queue to determine the mean blocking time for the upstream node. This mean time is then added to the mean service time of the node. The entire system is then treated as a Jacksonian network of M/M/s queues with infinite capacity.

Consider the simple network shown in Figure 1. Three M/M/1 queues are shown in series. The simplicity of this network is not a requirement for the applicability of the model, but is only used for the sake of clarity. Let the middle node be a finite capacity queue, allowing a maximum of K_2 entities to be waiting for service. The other notation used should be clear upon the examination of Figure 1. For a single M/M/1/K queue, theoretical results are available for the mean departure rate as a function of mean arrival rate and mean service rate. This is given by

$$\lambda' = \lambda\left[1 - \frac{1-\rho}{1-\rho^{K_2+1}}\right]\rho^{K_2} \qquad (1)$$

The basic idea behind the model consists of allocating the difference between the arrival rate to the node with the finite buffer and the departure rate from that node to a mean blocked time per entity in the previous node. The mean blocked time for node 1 in Figure 1 thus becomes

$$\tau_B = \tau_S\left[\frac{\lambda - \lambda'}{\lambda}\right], \qquad (2)$$

where τ_S and τ_B are the mean service time and the mean blocked time at node 1, respectively. Having accounted for the blocked time of node 1, it is clear that the finite buffer problem has been completely eliminated in that the addition of the blocked time of node 1 to its regular service time τ_S will allow the treatment of the problem as a queuing network with infinite buffer size at node 2. Therefore,

$$\mu_{1_{effective}} = \frac{1}{\tau_S + \tau_B} \qquad (3)$$

It is clear from the foregoing that such an approach is quite generalizable to very complex network of M/M/s/K queues. Theoretical results for the departure rate from an M/M/s/K queue are well-known. It is therefore possible, in simple cases, to start at the last node of a network and work backwards throughout the entire system and appropriately solve for all the blocked times of the nodes. For more complicated networks, a system of simulataneous equations need to be solved for the determination of blocked time at each node. The actual parameters of the network can then be solved for by assuming that a Jacksonian M/M/s network with effective service rates which now also includes the blocked rate is representative of the performance of the system.

The authors provide no proof for the proposed model. Let it suffice to say that the method has intuitive appeal, as well as a well-founded basis for recovering the relevant parameters of the finite queues, i.e., the theoretical solution for a single M/M/s/K queue. At the same time, it is not reasonable to claim general validity for the model without actual comparison of the results with a rigorous or experimental solution. Since no rigorous solutions for the general case exist, the simple network shown in Figure 1 was simulated and the results were compared with the model predictions. The comparison is discussed in the next section.

4. MODEL VERIFICATION

The system shown in Figure 1 was used to evaluate the model. The system consists of three individual nodes, each containing one server and a queue. The queue capacities, arrival time and service times all were input parameters to the model. The queues associated with the first and last

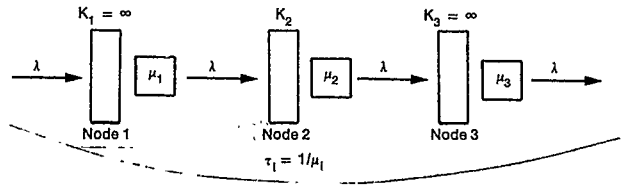

Figure 1. Network of three nodes.

nodes were assumed to have an infinite capacity. The capacity of the middle node was allowed to vary.

The model was tested against a simulation written in the event scheduling world view of SLAM II for the same situation. Exponential distributions were used in the simulation for the inter-arrival and service times.

One interesting aspect of the simulation behavior is the large number of observations required to obtain a reasonably accurate estimate of the mean behavior of such a system. This is hardly surprising for the M/M/s type of problem. We were guided by the treatise of Wilson (1979) in replicating the simulation results ten times, each with a different random number stream. In addition, each replication consisted of over 30,000 observations, leading to a population size of over a third of a million observations per point. The random number streams employed are those supplied in the SLAM-II package and are traceable to Schrage (1979). Having taken these precautions in generating reliable simulation results, the means resulting from the simulation are estimated to be within 5% of the true mean with a confidence of better than 95% for all cases shown in this paper.

The simulation properly accounted for the true effects of blocking by allowing the part to reside in its current node until room was available in the downstream queue. As part of the simulation, statistics were collected both for the fraction of time each server was blocked in addition to the fractions of time it was busy and starved.

The parameters used for comparing the model results with simulation are as follows: $\lambda = 0.5$; $\mu_1 = 0.2$; $\mu_2 = 0.4$; $\mu_3 = 0.3$. These values are, of course, arbitrary. However, as the results are obtained with varying capacities for the middle node, the comparison will encompass a range of conditions as far as traffic density for the first node is concerned.

Various system parameters predicted by the model can be compared with the simulation results. It is of particular interest in this case to examine the effects having to do with blocking. For this purpose, the fraction of time that the upstream node is blocked is shown in Figure 2. It is observed that the model results are indistinguishable from the simulation results if one considers the uncertainty remaining in the simulation (5%). This behavior of the model alone gives much confidence in the soundness of the modeling approach. As expected, the blocked fraction becomes less and less significant as buffer size is increased until a critical buffer size is reached, in this case approximately 11, beyond which no significant blocking is observed. Analytical tools of this type are of utmost importance in manufacturing engineering to help decide the optimum buffer size needed for a given application.

It is intuitive that as the allowed buffer size is reduced, the upstream node becomes more congested. The comparison in the expected length of the upstream queue is shown in Figure 3. As in the previous case, excellent agreement is observed, further verifying the adequacy of the model in

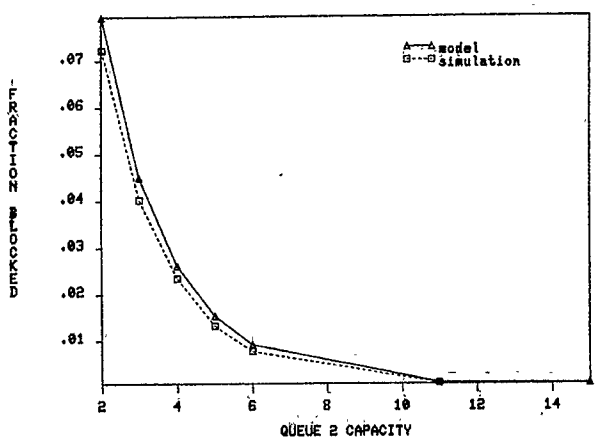

Figure 2. Blocking of cell 1 due to finite queue 2 capacity.

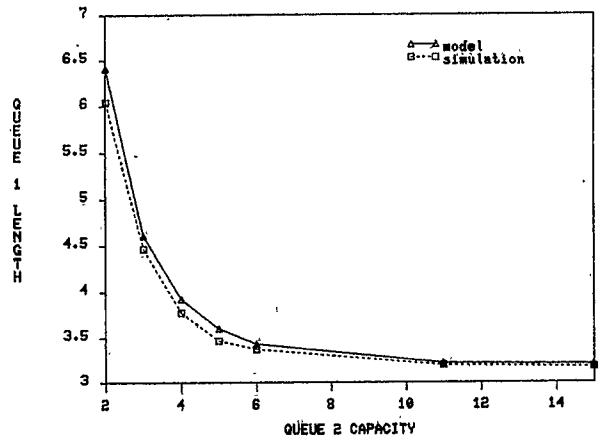

Figure 3. Effect of finite queue 2 capacity on length of queue 1.

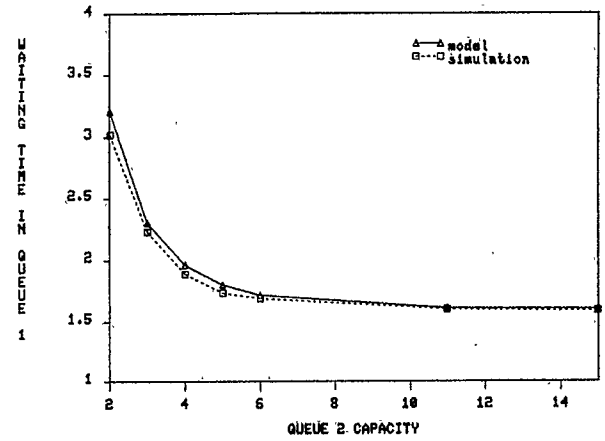

Figure 4. Effect of finite queue 2 capacity on waiting time in queue 1.

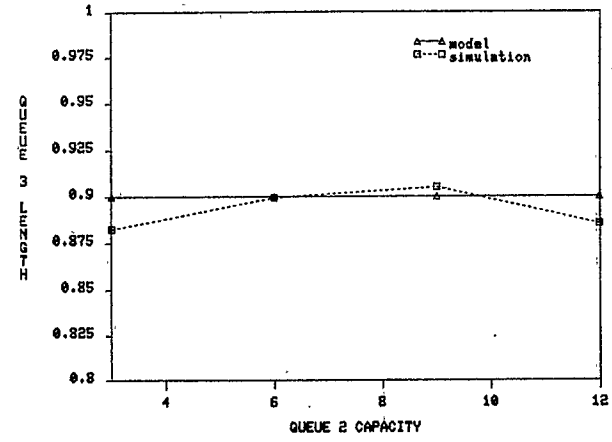

Figure 5. Effect of finite queue 2 capacity on length of queue 3.

the representation of the upstream effects of blocking. Finally, the average waiting time in the upstream queue is shown in Figure 4 with the same degree of agreement between the model and simulation results.

A very interesting feature of the model is that all the effects of blocking are represented only by adding to the service time of the UPSTREAM node a blocking time, as described in the previous section. No downstream effects are modeled. In fact, the upstream propagation of blocking stops upon the encounter of the first infinite queue. The authors suspect that these observations may in fact have a theoretical foundation for a network of M/M/s/K queues, although they are not aware of the existence of such a proof. The upstream effects of blocking predicted by the model have already been successfully compared with simulation results. Figure 5 shows that, as predicted by the basic structure of the present model, the downstream effects of blocking as far as the expected length of the downstream node is concerned is negligible. Similar agreement exists between the model and simulation results for other parameters associated with the downstream node.

Though validation of the model described above was performed for a system of three single server nodes, it is postulated that these results can be extended to a more complex system such as one with multiple nodes and servers. However further testing of the model with such systems is necessary before such a claim can be proved. Investigation of a multi-server, multi-node system may be done by changing the appropriate input parameters. Such a simple adaptation is possible because all the equations utilized in calculations by the model are applicable to multiserver systems.

A logical application of the model is in complex manufacturing situations, such as a job shop. Provided that the probability of a given item following a certain route is known, adaptation of the model is possible to represent this situation by including an additional factor in the blocking time. If the input to node c comes from nodes a and b, the routing factor for node a is the arrival rate to node c from node a divided by the total arrival rate to node c from nodes a and b. The blocked time in this case is equal to the product of the routing factor and the blocked fraction divided by the service rate of the blocked node. Once again the calculations would proceed backwards through the system until the first node was reached. As in the case presented in this paper, once the assignment of the blocked times to the appropriate nodes is completed, all the relevant parameters of the system are calculated by using the results for a network of M/M/s/∞ queues, as given by Jackson.

5. CONCLUSIONS

A model was presented for the representation of the mean performance of queuing networks with finite capacity queues. It was shown that the model produced excellent mean results as compared to a simulation of a simple network of three nodes. Although the case of multi-server nodes with complex routings was not compared with simulation, both the upstream and downstream effects resulting from blocking due to finite capacity queues was shown to

agree quite well with simulation results. It is therefore reasonable to expect that the modeling approach will prove correct in arbitrarily complex situations. However, this expectation needs to be verified by actual simulation experiments.

REFERENCES

Bell PC (1982), The use of decomposition techniques for the analysis of open queuing networks, *Operation Research Letters,* Vol. 1, No. 6, pp. 230-235.

Boxma OJ and Konheim AG (1981), An approximate analysis of exponential queuing systems with blocking, *Acta Informatica,* Vol. 15, pp. 19-66.

Burke PJ (1956), The output of a queuing system, *Operations Research,* Vol. 4, pp. 699-704.

Hillier FS and Boling RW (1967), Finite queue in series with exponential or Erlang service times — a numerical approach, *Operations Research,* Vol. 15, pp. 286-303.

Hillier FS and Lieberman GJ (1980) *Introduction to operations research,* Holden-Day, Inc., San Francisco.

Jackson JR (1957), Networks of waiting lines, *Operations Research,* Vol. 5, pp. 518-521.

Kleinrock L (1975), *Queuing systems,* Vol. 1, Wiley-Interscience, New York.

Solberg JJ (1977), A mathematical model of computerized manufacturing systems. In: *Proceedings, 4th International Conference on Production Research,* Tokyo, August 1977.

Schrage L (1979), A more portable Fortran random number generator, *ACM Transaction on Mathematical Software,* Vol. 5, pp. 132-138.

Takahashi Y, Miahara H and Hasegawa T (1980), An approximation method for open restricted queuing networks, *Operations Research,* Vol. 28, pp. 594-602.

Wilson JR (1979), *Variance Reduction Techniques for The Simulation of Queuing Networks,* Ph.D. Thesis, Purdue University.