ANALYSIS OF QUEUE BOUNDS IN A G/G/c LOGISTICS PLANNING MODEL

Khosrow Moshirvaziri, Information Systems Department, California State University, Long Beach, CA 90840 USA. 562 985-7965, moshir@csulb.edu

Mahyar Amouzegar, RAND Corporation, 1700 Main Street, Santa Monica, CA 90407, USA. 310 393-0411, mahyar@rand.org

ABSTRACT

This paper presents a closed stochastic simulation network model and several approximation and bounding schemes for G/G/c systems. The analysis was originally conducted to verify the integrity of simulation models used to develop alternative policy options conducted on behalf of the US Air Force. We showed that the theoretical bounds could be used to approximate mean capacities at various queues. In this paper, we present results for a G/G/8 system though similar results have been obtained for other networks of queues as well.

INTRODUCTION

In this paper we consider a closed stochastic simulation system model used in the analysis of aircraft engines maintenance and repair options. In this analysis, we evaluated the cost and benefits of centralized maintenance versus a decentralized option. The usage and maintenance of engines comprises the sequence of events illustrated in Figure 1: Planes fly (sorties) from main bases and remote operating locations to meet training and other requirements. After each sortie, the planes' engines are inspected on the flight line, and depending on the accumulated flying hours, are given minor maintenance. Engines may also be removed from aircraft and sent to an intermediate maintenance facility (IMF) for major maintenance. At this facility the engines are inspected, repaired, tested, and then returned to the flight line as serviceable spares. At each operating site there is a cache of serviceable spares to replace engines sent to IMF. However, there is only a limited inventory of such spares there may be time where aircraft are grounded due to the engines availability.

The nature of this problem has lent itself to a closed loop networks of multiple servers/queues, some sequential and others parallel. A queueing system is said to be *closed* if the servicing facility processes only a given group of permanent customers. When a customer needs service, it joins the queue and it is either served based on FIFO discipline or is given priority if it meets a certain criteria (e.g., a particular engine is required in the field faster than other type). The demand for service and duration of service depends on many variables and for this study we used historical data to compute the arrival and departure rates. The complexity of this problem led to queueing model that could only be described with general arrival and service times or a G/G/c/n queueing system where n, the restriction on system capacity, varied depending on the process. G/G/c queue and its related families, M/G/c, G/G/1 are too complex to analyze mathematically and there are very few closed formed results about such systems. However, several quite useful approximate and bounding results have been obtained. We used these approximations and bounds to create a robust simulation model for a large-scale engine maintenance system. These bounds and approximations were used in evaluating the robustness of our simulation model. And their application will be the focus of this report. In the next section, we will describe some of the results associated with G/G/c.

G/G/1 SYSTEM

The G/G/1 system and its theoretical results are used to derive what is presently known about the G/G/c system and thus will be discuss it first. We consider a G/G/1 system consisting of a single server with independent and identically distributed interarrival times as well as service times and unlimited queueing capacity. Let X denote interarrival time and $f_x(x)$, $1/\lambda$, and σ_x^2 denote the pdf, the mean and the variance of X, respectively. In addition, let S, $f_s(s)$, $1/\mu$, and σ_s^2 represent those corresponding for the service times. Although there are no closed form solutions for this model, there are some useful bounds developed in recent years for the quantities L, L_q, W, and W_q (see [7] and [8]).

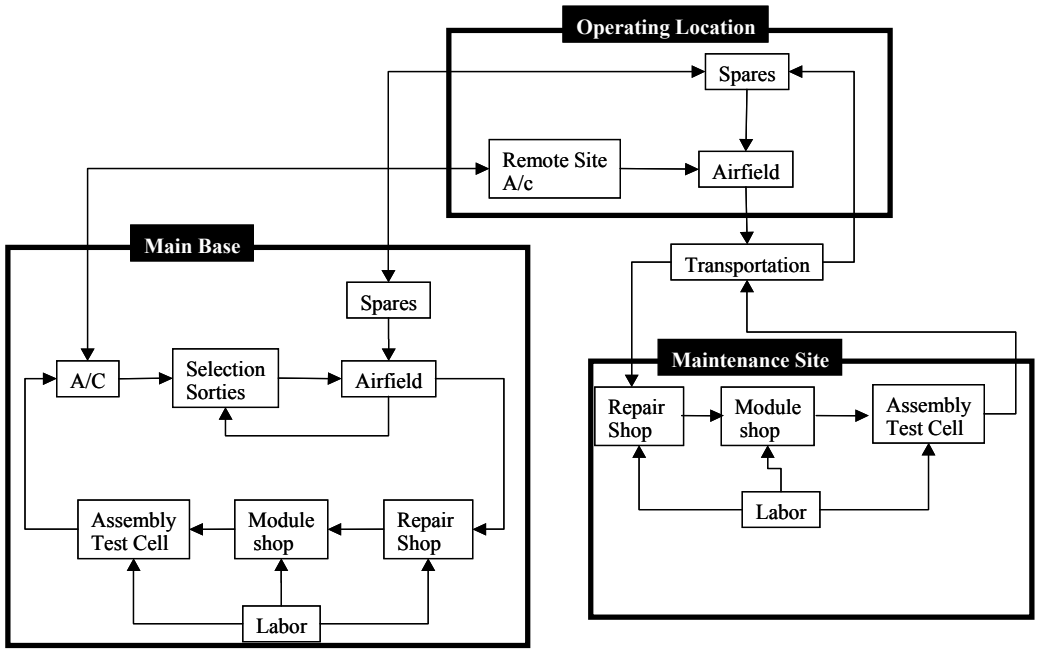

Figure 1: Operation and Maintenance Sequence

For G/G/1 systems with no restrictions on the interarrival or on the service time pdf's, several bounds have been developed (see [5] and [6]). These bounds, in essence, state that for the average steady-state waiting time in queue, W_q, we have

$$\frac{\rho^2(1+C_s^2)-2\rho}{2\lambda(1-\rho)} < W_q \le \frac{\lambda(\sigma_x^2+\sigma_s^2)}{2(1-\rho)} \qquad (1)$$

where the coefficient of variation for the service times, $C_s = \sigma_s \mu$ and, the utilization factor $\rho = \lambda/\mu$. For the stability of the system we must have $\rho < 1$. Note that the lower bound given above is not tight. This becomes obvious from the fact that, even at very high utilization rates, the bounds takes negative values, unless $C_s > 1$. But for C_s to be greater than 1, it must be that the service time pdf must be "more random" than the negative exponential pdf which has its $C_s = 1$.

Desired class property

A tight simple lower bound is given in [6] for a class of G/G/1 queues, which includes most encountered in practical cases. Thus, class requirement is that all queueing systems in it must have interarrival time

pdf, $f_x(x)$, satisfying the following property:

$$E[X-t|X>t] \le \frac{1}{\lambda} \quad \text{for all } t \ge 0 \tag{2}$$

If it is known that any given interarrival gap lasted more than a time t, then the condition above requires that the expected length of the remaining time, $X-t$, in that gap be less than the unconditional expected length of the gap, $E[X](=1/\lambda)$. This is of course true for the negative exponential variable, and in that case the condition becomes equality. When the condition holds, then we have:

$$U - \frac{1+\rho}{2\lambda} \le W_q \le U, \quad U = \frac{\lambda(\sigma_x^2 + \sigma_s^2)}{2(1-\rho)} \tag{3}$$

The upper and lower bounds may now be derived using this and by applying Little's formula, $L = \lambda W$, $L_q = \lambda W_q$ and the fact that $W = 1/\mu + W_q$. The following is easily obtained:

$$\lambda \cdot U - \frac{1+\rho}{2} \le L_q \le \lambda \cdot U \tag{4}$$

This implies that the difference between the upper and the lower bounds is $(1+\rho)/2$, but $0 < \rho < 1$, so this difference is always between 0.5 and 1. Thus, we can find the average queue length to within an accuracy of between 0.5 and 1 (depending on the value of ρ). Note that most "well-behaved" arrival time distributions satisfy the condition, including uniform, triangular or beta-type pdf's, which often are reasonably good approximations of many general interarrival time pdf's. Only a few common continuous random variables, such as those in the hyperexponential family, which are "more random" (informally speaking) than the negative exponential random variable, do not satisfy the condition.

Under heavy-traffic

Another important result that is available for the G/G/1 system is known as the heavy-traffic approximation (for more information see [3]). It applies for values of ρ near 1 and thus provides estimates for waiting times when it is known that waiting times are large. When ρ is near 1, the distribution of steady-state waiting time in queue in a G/G/1 system is approximately *negative exponential* with mean value $W_q = U$. The average waiting time for G/G/1 queueing systems is dominated by a $(1-\rho)^{-1}$ term under steady-state conditions, as the utilization ratio tends to 1. Consequently, the type of behavior that is normally seen in a simple M/M/1 system is also present for entirely general arrival- and service-time distributions, G/G/1.

G/G/c SYSTEM

The only general results on G/G/c system [2] that have been obtained to date are in the form of quite relaxed upper and lower bounds on average steady-state queueing characteristics. These bounds are often computed by, first, comparing a G/G/c system with a G/G/1 system that has the same "service behavior" as the G/G/c system. That is, the single server in G/G/1 works *c* times as fast as each of the servers in G/G/c and by applying the earlier results on G/G/1, given in the previous section. The most useful and applicable bounds on the average waiting time in queue which have been derived to date for G/G/c systems, based on those of G/G/1 is

$$W_q^1 - \frac{(c-1)\mu E[S^2]}{2c} \le W_q \le \frac{[\sigma_X^2 + (1/c)\sigma_S^2 + ((c-1)/c^2)(1/\mu^2)]\lambda}{2(1-\lambda/c\mu)} \tag{5}$$

where for each of the *c* servers, μ, σ_S^2, and $E[S^2]$ are the rate, variance, and the second moment of

service time, respectively. W_q^1 denotes the mean waiting time for a G/G/1 system with a service time denoted by a random variable $S^1 = S/c$ with service c times faster than that of each of the c servers in the G/G/c system, but with an identical arrival process. If W_q^1 is known or is computed using the results discussed above, we can substitute an exact expression. Note that for the general M/G/1 system we have the following well-known results, which can be used in deriving the G/G/c approximation bounds:

$$P_\circ = 1-\rho \qquad L = \rho + \frac{\rho^2 + \lambda^2\sigma_S^2}{2(1-\rho)}$$

$$W = \frac{L}{\lambda} = \frac{1}{\mu} + \frac{\rho^2 + \lambda^2\sigma_S^2}{2(1-\rho)} \qquad W_q = W - \frac{1}{\mu} = \frac{\rho^2 + \lambda^2\sigma_S^2}{2(1-\rho)} = \frac{\lambda[(1/\mu^2) + \sigma_S^2]}{2(1-\rho)}$$

$$L_q = \lambda W_q = \frac{\rho^2 + \lambda^2\sigma_S^2}{2(1-\rho)}$$

Thus, for example, for the M/G/c queueing system, one should use the exact expression for W_q^1 given above with $1/c\mu$ and σ_S^2/c^2, for the expected value and variance of the service times, respectively.
The corresponding heavy-traffic approximation for G/G/c systems has been derived [4]. This result implies that: For $\lambda c\mu$ approaching 1 in a G/G/c system, the waiting time in queue under steady-state conditions assumes a distribution that is approximately *negative exponential* with mean value

$$W_q = \frac{[\sigma_X^2 + (\sigma_S^2/c)]\lambda}{2(1-\lambda/c\mu)} \qquad (6)$$

Note once more that expected waiting time is dominated by a $(1-\rho)$ term, as ρ approaches 1 ($\rho = \lambda/c\mu$ for multiserver systems). We used the above results for G/G/c to have a point of reference for the simulation and test the results against these theoretical backdrops.

AN OVERVIEW OF THE SIMULATION MODEL

In terms of modeling, we are interested in the flow of entities (e.g., spares, personnel), the state of the system (e.g., engine not serviceable, spares inventory), and the processes (e.g., service time, sortie rates). The structure of the model is based on a set of hierarchical, functional blocks that generate and modify entities, processes and attributes. These blocks represent main bases, airfields, and intermediate maintenance shops. In general, the simulation is based on the following sequence of events: aircraft are flown from main bases or remote sites to meet certain flying requirements. After each mission, the aircraft and their engines are inspected at the airfield and in most cases they are fully operational within hours. However, when engines accumulate enough flying hours, or when unscheduled maintenance is required, engines are removed from the planes, and sent to an intermediate maintenance facility. They are then inspected, repaired, tested, and returned to service. The first requirement for the model is the number and types of aircraft, and the number and the age of installed engines. The aircraft and engines are combined to form fully operational aircraft. They are sorted, based on the age of the engine, and are then queued for flying. After each sortie, the aircraft is sent to the airfield block where it is inspected and maintained. Each aircraft that passes the inspection is sent back to the pool of available aircraft. Some aircraft require minor repair, which is performed on the flight line. The number of engines pulled from the aircraft is a function of the age and the type of the engine. The detached engines are tagged according to the removal type (i.e., scheduled or unscheduled) and are sent to the IMF shop. Aircraft are then identified as not operational and are queued for the next available serviceable engine. These

aircraft are either put back to service immediately, if there are serviceable spares available, or they await the arrival of engines from the maintenance shop.
At the maintenance facility, engines are queued in two parallel lines, the first is for the engines that require parts that are not available and the other is for engines that await maintenance. The modular engines that have been processed by the IMF shop are sent to the module shops. Engines that enter the module shop are separated into five modules. Engines that leave the module shop are sent to the assembly and test cell. In this section, engines are queued for assembly, the test cell and the final inspection. After assembly and test cell, engines are sent to the spare engines pool to be installed on the aircraft to create fully operations aircraft. These aircraft leave this section to join the pool of other aircraft and the whole cycle starts again. Figure 1 illustrates this process for only one main and operating base. The model, however, has taken into account a problem with several such bases (for more information on the simulation model see [1]).

Simulation set up

In this section, we will present some of the input and output parameters used in our analysis. We will illustrate these parameters by running a scenario with 36 single engines aircraft and 66 two-engines aircraft. The model is run for about two simulated years.
Engines are typically set on a rail and require a 5-person team per shift. The regular shift is about 8 hours and the shops operate at 2 shifts a day. During peak demand period, the shops may shift their operations to 24 hours a day, seven days a week with each shift as long as 12 hours. The capacity of the IMF is determined by the combination of rails and the personnel, a "rail team". Other shops have different architecture but all are bounded by number of staff and the equipment. Airfields and the transportation network are bounded by the capacity of the flight line and the number of transporters, respectively.
There are three smaller main bases with three-rail team capacity and one large one with 7-rail teams capacity. In other words, 3 and 7 parallel servers, respectively. There is also a remote facility with 8 rail teams. The other parts of the shop (e.g., the module shop) are sized accordingly.

Simulation Results

On average about 119 customers entered the system (with variance of 253 and an standard deviation of 15). At the end of the simulation run, about 105 customers were served. The IMF shop at the remote site (with 8 servers) reported an average wait time of 5.538461538462 days. Although the arrival and the service times varied widely, as they depend heavily on the other parts of the system, the reported wait time seemed reasonable and was consistent with the theoretical bounds. Using the Poisson distribution, we get a wait time of 2.06 days and 4.13 using the Exponential distribution. d 10 days for the average wait.
Table 1 illustrates the theoretical bounds for a single server process in the inspection shop. The simulation model reported an average of 0.808499576845 for the queue length and 10 days for the average wait.

Table 1: Sample Results for a G/G/1 System

Distribution	Utilization Factor	Var (X)	Var (S)	Wait Time		Length	
				LB	UB	LB	UB
Poisson	1.32	0.33	0.25	-	-	0.06130	-
Exponential	0.757576	0.1089	0.0625	0.78125	1.07125	3.125	3.24621
Uniform	0.08	0.020833	0.003333	-	0.01050	-	0.00840
Normal	0.18018	0.900901	5	2.58725	3.24225	2.51103	2.92094

Customers enter the last server.

Table 2 illustrates the arrival and departure rates for the sequence of servers in the maintenance process. Some customers bypass the first queue and enter the second queue with multiple servers. After the service, some customers, again, bypass the next server. In this section, there are five parallel servers and customers depending on their requirement must enter a particular server queue. Finally all customers enter the last server.

Table 2: Arrive and Departure in the JEIM Shop

Server (Single)			Server (multiple)		Server (Single) 5 Parallel			Server (Single)	
A	D	B	A	D	A	D	B	A	D
33	25	86	111	104	94	92	10	102	96

Table 3 illustrates the theoretical versus simulated bounds for the first queue, in the eight-server scenario discussed above.

Table 3: Sample Results for a G/G/8 System

Distribution	Queue Wait Time		Length		Simulation Results	
	LB	UB	LB	UB	W	L
Poisson Arrivals	—	—	—	—	—	—
Exponential	3.21107	4.97622	1.22465	1.64215	5.04683	5.66852
Uniform	0.68181	1.3561	.02367	.15627	3.97198	3.20338
Normal	2.58725	3.24225	2.51103	2.92094	5.04280	5.89083

CONCLUDING REMARKS

In this paper we presented a closed stochastic simulation network model and several approximation and bounding options available in a G/G/c system. The analysis was conducted to verify the integrity of the simulation model used to developed alternative policy options conducted on behalf of the US Air Force and presented in [1]. We showed that the theoretical bounds could be used to approximate mean capacities at various queues. In this paper only the results for G/G/8 was presented in order to avoid lengthy tables of results. However, such consistency was observed amongst the other queues.

REFERENCES

[1] Amouzegar, M.A., et al., Supporting Expeditionary Aerospace Forces: An Analysis of Jet Engine Intermediate Maintenance Options, *RAND, MR-1431-AF*, 2001.

[2] Brumelle, S. L., Some Inequalities for Parallel Server Queues, *Operations Research*, 1971, 402–413.

[3] Kingman, J. F. C., On Queues in Heavy Traffic, *Journal of the Royal Statistical Society*, Series B, 1962, 383–392

[4] Kollerstrom, L., Heavy Traffic Theory for Queues with Several Servers: l, *Journal of Applied Probability*, 1974, 544–552

[5] Marshall, W. G., Some Simpler Bounds on the Mean Queueing Time, *Operations Research*, 1978, 1083–1088

[6] Marshall, K. T., Some Inequalities in Queueing, *Operations Research*, 1968, 651–665

[7] Larson, R. and Odoni, A., Urban Operations Research, Prentice Hall, 1981

[8] Gross, D. and Harris, C. M., Fundamentals of Queueing Theory, John Wiley & Sons, 1998