Proceedings of the 1993 Winter Simulation Conference
G.W. Evans, M. Mollaghasemi, E.C. Russell, W.E. Biles (eds.)

ESTIMATING SERVICE CAPACITY USING NONHOMOGENEOUS WORK ARRIVALS

Michael P. Bailey

Department of Operations Research Naval Postgraduate School
Monterey, CA 93943-5000

ABSTRACT

We explore experimental procedures for comparing the capabilities of complex discrete event service systems. Instead of measuring system capability by analyzing or simulating the system with a constant rate of arriving work, system capability is measured as the maximum rate of work arrival for which the system has a steady state. Hence, we seek the arrival rate which causes the system to be at full capacity. This rate is arguably the best indication of the service system's capability.

1 INTRODUCTION

As industrial engineers, applied probabilists, simulationists, and systems analysts, we are often called upon to evaluate systems which service input traffic and produce finished products. These systems are sometimes traditional queues or networks of queues, but are often systems with queue-like characteristics which cannot accurately be modeled as traditional queuing systems. In practice and in the literature, this evaluation is traditionally based on exercising a model of the service system by subjecting it to a stream of input traffic and estimating or calculating some expected system performance measure.

We feel that this typical experimental design is lacking, and that the shortcomings stem from the arbitrary choice of the distribution of the input process. Especially problematic are cases where the service system being modeled does not currently exist, where worst-case behavior is sought, or where we wish to evaluate the system in situations which are not accessible for data collection. Practical service system analysis is interesting only when the service system is in an environment where the workload is high relative to the system's capability to serve. In all that follows, we are interested in finding the intensity of the input process that taxes the service system to the extremes of its capabilities, and using this intensity as a measure to compare systems.

2 THE GENERAL SERVICE MODEL

The service systems considered all have the following features:

1. a centralized, controlable, nonlattice process which generates tasks at a rate λ per unit time;
2. tasks are admitted upon generation and processed by the system;
3. a completed task is ejected from the system;
4. the system has the capability to process as many as μ tasks per unit time on average.

We will call such systems Discrete Event Service Systems (DESSs), see figure 1. In this work, we study the behavior of systems where there is a one-to-one correspondence between the tasks we submit for processing and the finished tasks the service system produces. Work-conserving queuing models do not allow

- tasks to expire while in service;
- tasks to create other tasks while in service;
- tasks to be split or combined;
- tasks which never finish service.

Work-conserving queuing system models are common in both the practice and literature of applied probability. In a typical experiment, we generate input to the system at a constant rate, monitor the performance of the system either at fixed intervals or upon departure from the system, and employ well-known methods of steady-state analysis to estimate the steady-state average of the performance measure.

A maxim of the analysis of service systems is that the system will have stationary long-run behavior if

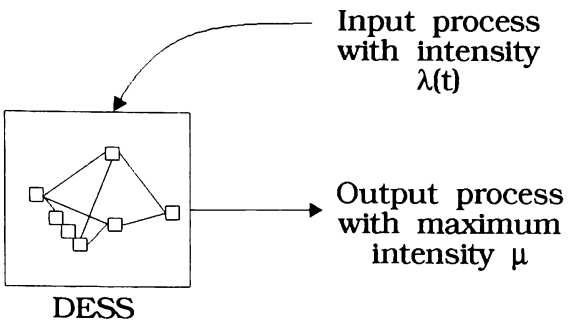

Figure 1: A simple DESS.

and only if the number of arriving tasks are, on average, less than the number of tasks the system is capable of processing. If our overall system can work at a maximum of μ tasks per unit time, we can input as many as μ per unit time and the system will remain stationary. If λ is our arrival rate for the system, we wish to manipulate λ to expose μ.

3 GENERATING DATA

There are two ways we can generate data from a work-conserving system which will reveal the maximum processing rate in the system. They are:

- input tasks to the system at a rate known to be much higher than the system can handle;
- fill the system, then input a new task every time that a task completes.

In the former, the rate of outgoing jobs eventually converges to μ. Instead of choosing a very high input rate and dealing with the problems of exploding buffer contents and a nonrecurrent system, we will simply close off the system and recirculate the tasks which finish. Hence, we take the second approach.

Thus, we examine a special kind of closed queuing network – one with a single loop-back which all tasks traverse. Let $\lambda(t)$ be the time-dependent rate of recirculation of tasks in the system. So long as the system contains enough tasks to keep it working at capacity, we have $\lambda(t) \rightarrow \mu$ as $t \rightarrow \infty$. Kelly (1979) and Walrand (1988) both show this for exponentially distributed service, and Disney and Kiessler (1987) make the extension to Jackson networks. The result can be extended in the obvious way by treating Phase-type distributions for service times, to produce the result we seek ($\lambda(t) \rightarrow \mu$ as $t \rightarrow \infty$) for generally distributed service.

4 DETECTING TRANSITION TO STEADY STATE

Let us simulate the completion of the first N customers serviced by the closed system for M independent replications. Let $T_{i,j}$ be the j^{th} time between recirculation during the i^{th} replication. Thus, $T_{i,j}, i = 1, 2, \ldots, M$ is a set of *iid* samples. Let $\bar{T}_j = \sum_{i=1}^{M} T_{i,j}/M$ be the average recirculation time process. We seek the index N^* such that $ET_{i,j} = E\bar{T}_j = \mu$ for all $j > N^*$. Hence, we are in the setting of a traditional initial transient detection problem.

There exist many ways to tackle this problem, including

- cross-replication confidence intervals, Welsh (1983)
- tests for significant drift;
- standardized time series (STS), Schruben (1982).

In our experiences, we have found STS useful, especially in a slightly modified version we have developed, which we call *ratio STS* (RSTS).

4.1 Ratio STS

The method of standardized time series (STS), Schruben (1982) produces confidence intervals from autocorrelated, stationary data. This method was used in Schruben, Singh, and Tierney (1983) to detect the existence of initialization bias in simulation output, and was sharpened to produce optimal tests when the functional form of the initialization bias is known.

Suppose that we have M independent samples of n points each, with $Y_{i,j}$ being the j^{th} point in the i^{th}

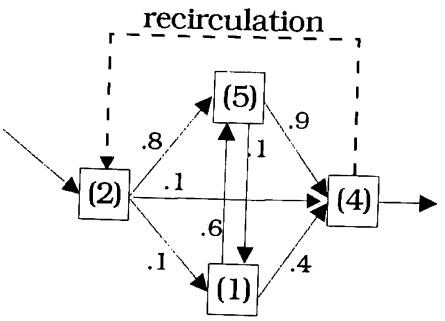

Figure 2: A Jackson Network designed to have a maximum service rate of 0.5. The numbers in parentheses are the number of servers at each station, and the routing probabilities are shown on the workstation connections. All servers have unit service time, and all buffers are infinite. The dashed line shows the recirculation route added to force the system to serve at the maximum rate.

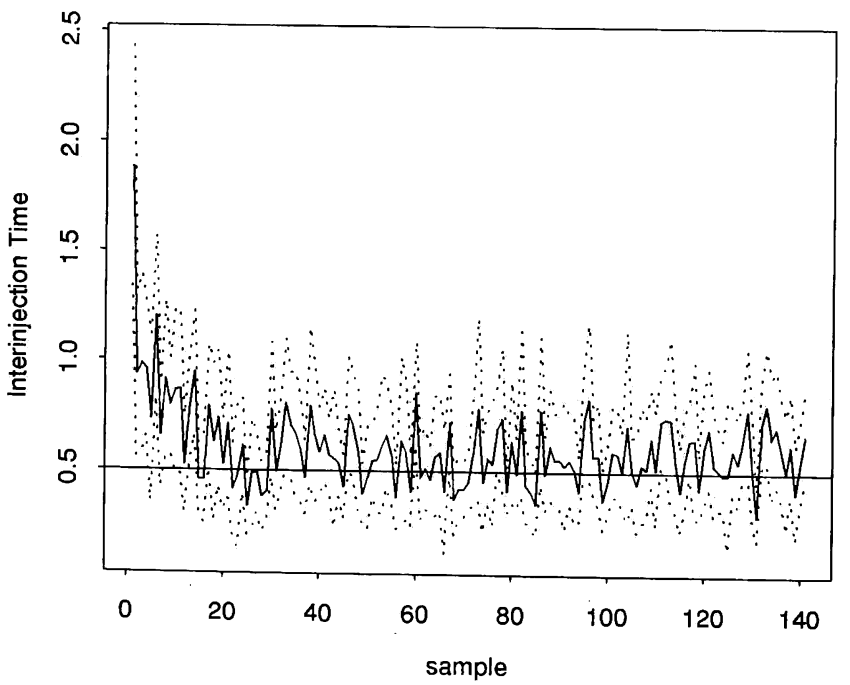

Figure 3: Trajectory of the mean process $\bar{T}_j$ and the confidence intervals.

independent sample. Let

$$\bar{Y}_{i,j} = j^{-1} \sum_{k=1}^{j} Y_{i,k} \tag{1}$$

for $i = 1, 2, \ldots, M$, and with $\bar{Y}_{i,0} = 0$ for each i. The time series $S_i(k), k = 1, 2, \ldots, n$ is constructed for each independent replication i as

$$S_i(k) = \begin{cases} \bar{Y}_{i,n} - \bar{Y}_{i,k} & for\ 0 < k < n \\ 0 & k = 0, n. \end{cases} \tag{2}$$

Let σ be the variance of Y_i. If $S_i(k)$ is divided by $\sigma\sqrt{n}/k$ and scale the index k so that the result resides in the unit interval $[0, 1]$, the resulting time series $T_i(t), 0 \leq t \leq 1$ is known to approximate a Brownian bridge as $n \rightarrow \infty$. This is the fundamental result of Schruben (1982), and the theoretical basis of this sequential procedure.

Schruben shows that scaling and summing $T_i(t)$,

$$A_i = \sigma\sqrt{n} \sum_{k=1}^{n} T_i(kn) \tag{3}$$

results in a normal random variable A_i with variance given by

$$VAR(A_i) = \frac{\sigma^2 n(n^2 - 1)}{12}. \tag{4}$$

Note that, except for a factor of σ^2, $VAR(A_i)$ is independent of the data, it relies only on the parameters of the experiment. Hence, for any integer $d < M$,

$$\chi_d^2 \sim \sum_{i=1}^{d} \left(\frac{A_i}{\sqrt{VAR(A_i)}}\right)^2 \tag{5}$$

$$= \frac{12}{\sigma^2 n(n^2 - 1)} \sum_{i=1}^{d} A_i^2. \tag{6}$$

The original STS used to detect initial transients used χ_d^2 as a test statistic for stationarity of the mean response. If we form a ratio of χ_d^2 and χ_{M-d}^2, we can eliminate the need to estimate σ^2, forming

$$F_{d,M-d} \sim \frac{d^{-1} \sum_{i=1}^{d} A_i^2}{(M-d)^{-1} \sum_{i=M-d}^{M} A_i^2}. \tag{7}$$

This test statistic, which we call the RSTS test statistic, is easy to use in all of the applications where STS is applied. In particular, if we are interested in determining the onset of steady state, we can form the backward-moving sequences $A_{i,j}, j = n-1, n-2, \ldots, 1$ for each replication i, where $A_{i,j}$ is formed from the subsequence $Y_{i,k}, k = j, j+1, \ldots, n$, the portion of the i^{th} replication between j and n. Thus, we form the sequence of F-statistics

$$F_{d,M-d}(j) \sim \frac{d^{-1} \sum_{i=1}^{d} A_{i,j}^2}{(M-d)^{-1} \sum_{i=M-d}^{M} A_{i,j}^2}. \tag{8}$$

If we assume that the system is in steady state when each of the $A_{i,n}$ are collected, then we can detect the transition of the system into steady state by looking at the first index N^* where $F_{d,M-d}(N^*)$ exceeds the critical value for an F random variable with identical degrees of freedom. This method is demonstrated in the following example.

Continuing with the work-conserving system example, suppose that we

- start the system with 25 tasks enqued at workstation 1 at time 0.0;
- simulate $N = 500$ customer recirculations;
- replicate $M = 20$ times.

Figure 3 shows the trajectory of $\bar{T}_j$ and the associated confidence interval process for the first 140 recirculations. Clearly, by sample $N^* = 80$ we have passed the criteria for being in steady state according to Welsh's cross-replication confidence interval method. Furthermore, we can see that any detectable slope in the mean process is negilible. When tested for our 20 independent samples, the drift of tested to be insignificant (H_0: no drift has p-value ≈ 0.4).

The mean time between recirculations in $100, 101, \ldots, 140$ was $\bar{T} = 0.56$, (confidence interval $(0.55330, 0.56691)$), clearly not as fast as the $\mu = 0.50$ which we know to be the system's capacity. Performing unweighted RSTS in the first 140 samples showed *no transition to steady state detectable* - the procedure seemed to be accurate enough to discern that the transition had not yet occured, see figure 4.

When we extend the length of the runs we consider to the full 500 samples, we see that RSTS was able to indicate a strong transition to steady state around the $N^* = 330$ sample, see figure 5. Averaging the samples collected in $350, 351, \ldots, 500$, we observe an overall average of $\bar{T} = 0.501$ (confidence interval $(0.49816, 0.50389)$).

RSTS clearly dominated the other traditional initial transient methods. In the case of the recirculating jobs, we clearly have a very gradual descent to the steady-state average. The detection method is not important to our overall theme, though we must issue a general caution: The choice of N^* should be made very conservatively.

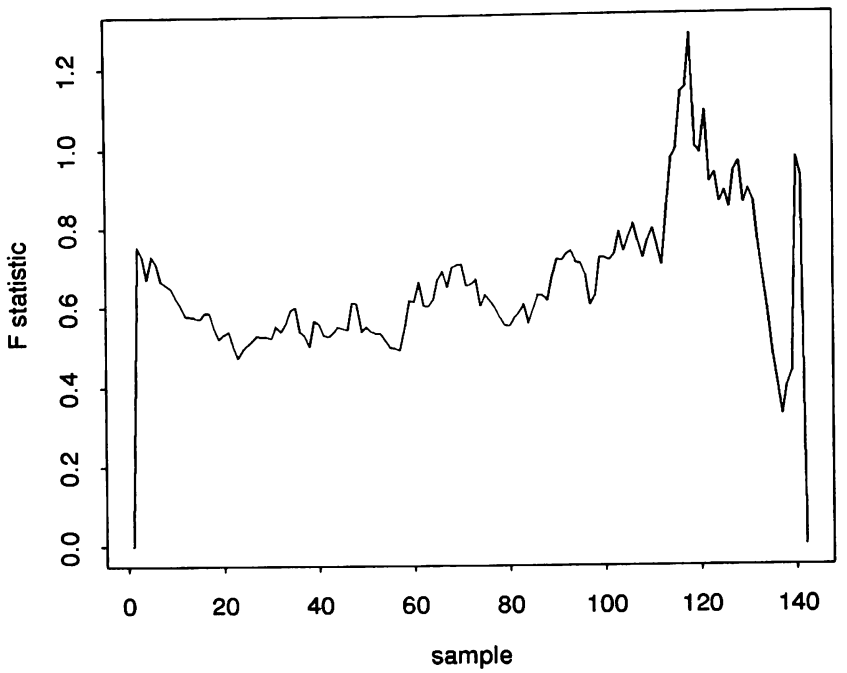

Figure 4: RSTS performed on the first 140 sample recirculation times, using 12 numerator degrees of freedom and 8 denominator degrees of freedom. Conclusion: no transition to steady state was evident.

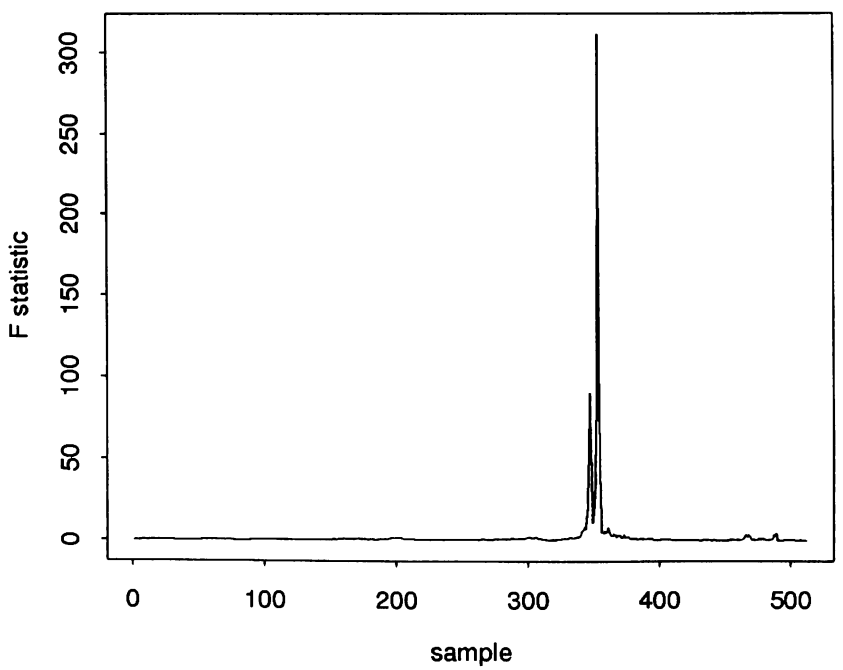

Figure 5: RSTS performed on the first 500 sample recirculation times.

5 CONCLUSION

In this work, we have described the analysis problems which arise when we are attempting to determine the service capacity of a *black box* service system. We concentrated this investigation on simple queuing systems which conserve work. In this case, we showed the advantages of closing the system so that output from the system was recirculated, as the time between recirculations converges to the system's service rate. We investigated ways to detect this convergence, and showed how difficult this is in a simple Jackson network example.

The wider significance of this work is the beginning of an exploration for empirical methods for determining the capacity of a service system. This exploration is done not by representing the system using a queuing model which we know how to analyze *a priori*, but by using a realistic model of the system and measuring its performance in terms the *user* has in mind.

REFERENCES

Disney, R. L. and P. C. Kiessler. 1987. *Traffic Processes in Queuing Networks, A Markov Renewal Approach.* Baltimore, Maryland: Johns Hopkins University Press.

Kelly, F. P. 1979. *Reversibility and Stochastic Networks.* New York: John Wiley and Sons.

Schruben, L. 1982. Detecting initialization bias in simulation experiments, *Operations Research* **30**, p. 569-590.

Schruben, L., H. Singh, and L. Tierney. 1983. Optimal tests for the initialization bias in simulation output. *Operations Research* **31**, p. 1167-78.

Walrand, J. 1988. *An Introduction to Queueing Networks.* Englewood Cliffs, New Jersey: Prentice Hall.

Welsh, P. 1983. The statistical analysis of simulation results. *Computer Performance Modeling Handbook.* New York: Academic Press.

AUTHOR BIOGRAPHIES

MICHAEL BAILEY is an associate professor of operations research at the Naval Postgraduate School. He actively pursues research in the fields of simulation modeling, experimental design for simulations, and applied stochastic analysis. He also has expertise in in several military applications areas, such as electronic warfare, C^4I, weapon system reliability, and transportation.