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

CONJECTURED UPPER BOUNDS ON TRANSIENT MEAN TOTAL WAITING TIMES IN QUEUING NETWORKS

Frank D. Chance

School of Operations Research and Industrial Engineering
Cornell University
Ithaca, New York 14853

ABSTRACT

In this paper, we give conjectured upper bounds on transient total mean waiting times for jobs in initially idle single source open Jackson networks. If true, these *upper* bounds provide *lower* bounds on the time required for the transient mean to approach its steady-state value. We compute the bounds by a weighted sum of transient means from a network decomposition, and we graphically display bound performance for five and fifty node networks.

1 INTRODUCTION AND SUMMARY

This paper summarizes a few of the results given in Chance (1993b) concerning the transient behavior of queuing networks. To introduce the necessary concepts, we start with an example. Figure 1 displays one year's simulated total waiting time output for a fifty-machine, single-product factory, with re-entrant routing, rework, and machine breakdowns. Jobs are released into the factory at a steady rate, and each output point is the sum of all queuing delays for a particular job. The simulated factory initially contains no jobs, and this setting influences the output. In this paper we study the influence of this initial condition, giving conjectured bounds on transient mean total waiting times.

The simulation used to generate the output of Figure 1 is a simplified version of a semiconductor manufacturing simulation developed at IBM (see Hood, Amamoto, and Vandenberge 1989). For an analytic complement to this simulation, see Connors, Feigin, and Yao (1992). This IBM simulation has been used to model very large factories with multiple product types and complex routings. Due to the complexity of the initial state, the simulated factories are usually started with no work in progress. Transient effects in these simulations can be quite severe and are not effectively captured by summary statistics. We seek a

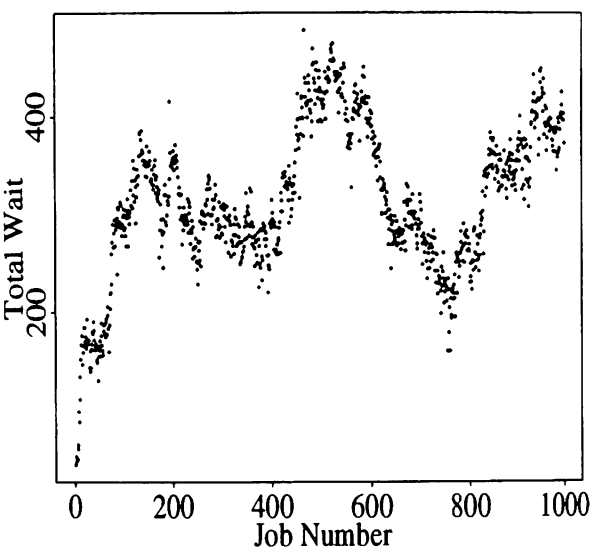

Figure 1: Example Total Waiting Time Output

prospective approach; before the simulation is run we want to approximate the extent of transient effects. To this end, we explore transient behavior in queuing network models. Although these models are an approximation of the simulation, we believe they can provide valuable insight into the simulation's transient behavior.

As an example, Figure 2 shows the layout of a typical sector inside a larger semiconductor factory model. Each block represents a group of identical tools, with the number of tools (servers) listed inside the block, and the mean service time given above the block. Jobs enter the sector and are processed through several machines. Upon completion, jobs are inspected for defects; faulty jobs are reworked, good jobs exit the sector. We model this sector with a Jackson network.

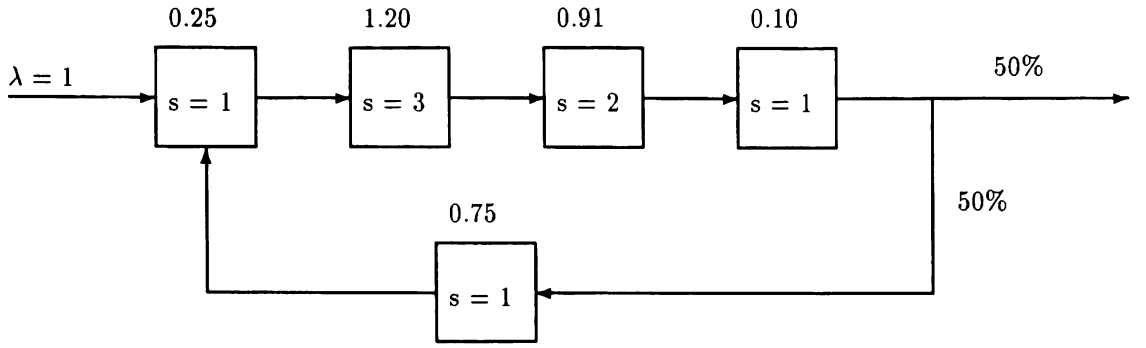

Figure 2: Single-Sector Configuration

Figure 3a displays a single realization of total waiting times for the sector, starting with no jobs in the sector. We wish to estimate the limiting mean total waiting time. Mean waiting times from the beginning of the run are much lower than the limit; from a single run it is difficult to tell the extent of this transient effect. Figures 3b, 3c, and 3d display the average total wait for 10, 50, and 100 replications. For larger numbers of replications, it becomes easier to delineate the transient period. Figure 3d displays a conjectured upper bound on the transient mean total wait. Because the conjectured bound approaches the limit faster than the mean total wait, we can use it to find a lower bound on the extent of the transient period.

To compute the bounds for networks, we use a decomposition method. These networks are as described in Jackson (1957), except we additionally restrict new jobs to enter the network at a single queue. This restriction seems typical of many semiconductor manufacturing simulations, where new jobs enter the factory at a single point. In Section 2, we give relevant assumptions and notation. In Section 3 we conjecture that mean total waiting times in an $M/M/1$ feedback queue are bounded above by those in a corresponding non-feedback $M/M/1$ queue with the same steady-state mean waiting time. In Section 4 we conjecture that mean waiting times in an $M/M/s$ queue are bounded above by those in a corresponding $M/M/1$ queue with the same steady-state mean waiting time. In Section 5 we conjecture that mean total waiting times in single source open Jackson networks are bounded above by a weighted sum of transient mean waiting times from a network decomposition. In Section 6 we propose a hypothesis for why the bound obtained through the network conjecture is loose for large networks. We consider a simple serial line, and give analytic and empirical results for the interarrival time distributions at downstream queues. It appears that downstream interarrival times are stochastically larger than those at the front of the line, even though in the limit they are distributed the same. Thus these downstream queues are initially less congested, and we hypothesize that this effect causes the network to warm up more slowly than the sum of its independent parts.

2 ASSUMPTIONS AND NOTATION

Whenever we consider a Jackson network, we mean a single-source open Jackson network with the following assumptions and notation.

Assumption 1 Queue j $(1 \leq j \leq N)$ contains s_j identical servers.

Assumption 2 New jobs enter the network at queue q_{new} according to a rate λ Poisson process.

Assumption 3 Service at each queue is first-in-first-out. Service times at queue j are i.i.d. rate μ_j exponential random variables, independent of the arrival process.

Assumption 4 After completing service in queue j, jobs proceed directly to queue k with probability p_{jk}, or exit the network with probability $1-\sum_k p_{jk}$. Routing decisions are independent of service and arrival processes.

Assumption 5 Each queue in the network has effective traffic intensity

$$\rho_j = \frac{\lambda_j}{\mu_j s_j}$$

less than 1. Let $I_{[X]}$ be 1 if condition X holds, 0 otherwise. The effective input rate λ_j satisfies

$$\lambda_j = \lambda I_{[j=q_{new}]} + \sum_{k=1}^{N} P_{kj}\lambda_k.$$

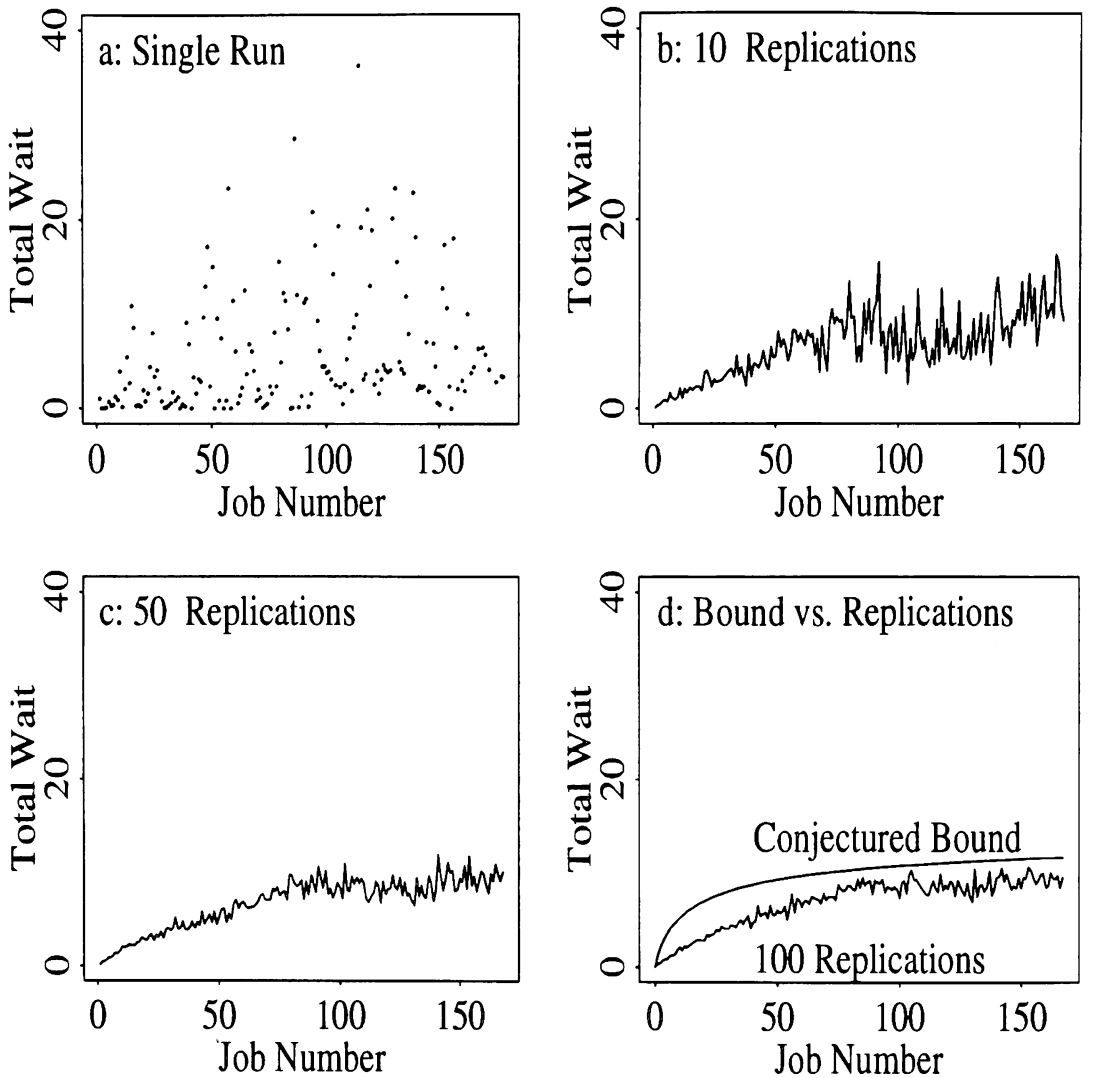

Figure 3: Output from Single Sector

Denote the steady-state expected waiting time at queue j by α_j (p. 88 of Gross and Harris 1985 gives a closed-form expression for α_j.)

3 FEEDBACK QUEUES

In this section, we consider a single $M/M/1$ queue Σ with input rate λ, service rate μ_1, and probability p of instantaneously rejoining the input queue after service completion. The effective input rate is

$$\lambda_1 = \frac{\lambda}{1-p},$$

and the effective traffic intensity is $\rho_1 = \lambda_1/\mu_1$. The service discipline is first-in-first-out. This type of system has been extensively studied, and is often called a *feedback* queue, because exiting jobs feed back into the arrival process. For more information, see the general discussion in Gross and Harris (1985, p. 235), the survey paper by Disney (1981), or the article on sojourn times by Hunter (1988).

In this section we conjecture that transient mean waiting times in the feedback $M/M/1$ queue Σ are bounded above by those from a non-feedback $M/M/1$ queue $\Sigma^\star$ with input rate λ_1 and service rate μ_1. For example, Figure 4 shows a comparison of expected waiting times in a feedback queue with input rate $\lambda = 0.5$, feedback probability $p = 0.5$, and service rate $\mu_1 = 2$ versus a non-feedback queue with input rate $\lambda_1 = 1$ and service rate μ_1.

Denote the total waiting time of job n in Σ by W_n, conditioned on job 0 arriving to an empty queue. Let

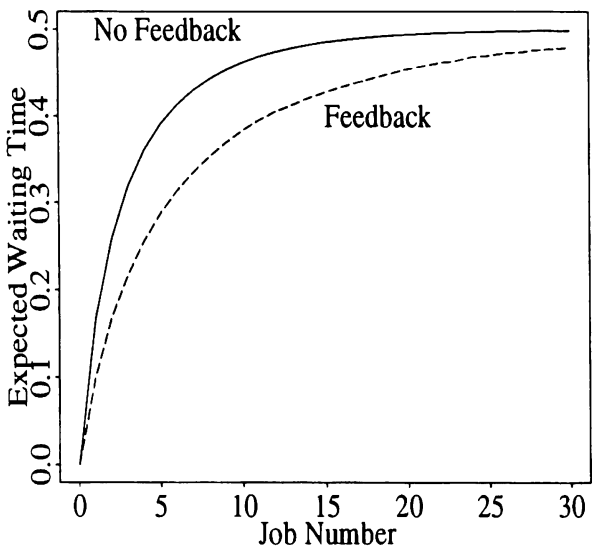

Figure 4: Comparison of $M/M/1$ Transient Mean Waiting Times with and without Feedback

$W_n^\star$ be similar for $\Sigma^\star$. We propose the following.

Conjecture 1 Transient mean total waiting times in the feedback queue Σ are bounded above by those in the corresponding non-feedback queue $\Sigma^\star$, after some small initial period. That is, there exists some n_Σ so that

$$n \geq n_\Sigma \quad \Rightarrow \quad \mathrm{E}[W_n] \leq \mathrm{E}[W_n^\star].$$

We must specify $n_\Sigma > 0$. Since job 0 in $\Sigma^\star$ never waits, $\mathrm{E}[W_0^\star] = 0$. But job 0 in Σ may feed back behind jobs 1, 2, or higher, and hence $\mathrm{E}[W_0] > 0$. In our empirical investigation, n_Σ is very small, even for very high traffic intensities (up to $\rho_1 = 0.9999$).

To provide support for this conjecture, we used the program described in Kelton and Law (1985) to give exact values for $\mathrm{E}[W_n^\star]$, and we used the simulator described in Chance (1993a) to estimate $\mathrm{E}[W_n]$ (using 1,000,000 independent replications). We tested low, moderate, and high traffic intensities ($\rho_1 = 0.10$, 0.50, and 0.90). We varied the feedback probability between 0.10, 0.50, and 0.90. For $\rho_1 = 0.10$, we experimented for jobs 0 to 5, for $\rho_1 = 0.50$, we tested jobs 0 to 25, for $\rho_1 = 0.90$, we tested jobs 0 to 100. In all cases n_Σ was less than 3.

4 MULTISERVER QUEUES

Consider an $M/M/s$ queue Σ. Arrivals occur according to a rate λ_1 Poisson process, and service times are i.i.d. rate μ_1 exponential random variables independent of the arrival process. Jobs are served first-in-first-out by s identical servers. Denote the steady-state mean waiting time for Σ by α_1. In this section we conjecture that transient mean waiting times from the $M/M/s$ queue Σ are bounded above by those those from a $M/M/1$ queue $\Sigma^\star$ having the same input rate and steady-state mean waiting time.

For example, Figure 5 compares transient mean waiting times in an $M/M/3$ queue and an $M/M/1$ queue having the same input rate and steady-state mean waiting time.

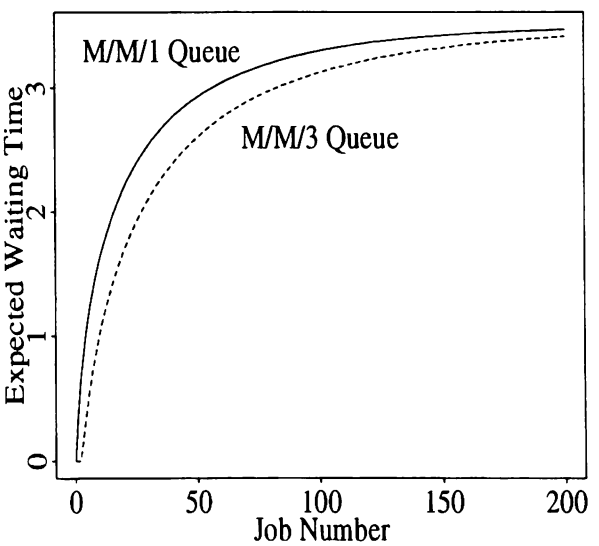

Figure 5: Comparison of $M/M/3$ and $M/M/1$ Transient Mean Waiting Times

To bound the waiting times in Σ, let the input rate to $\Sigma^\star$ be λ_1, and the service rate

$$\mu_1^\star = \frac{\lambda_1}{2}\left(1 + \sqrt{1 + 4/(\lambda_1 \alpha_1)}\right). \qquad (1)$$

With this choice of $\mu_1^\star$, the steady-state mean waiting time for $\Sigma^\star$ works out to be α_1, the same as for Σ. Denote the waiting time of job n in Σ by W_n, conditioned on job 0 arriving to an empty queue. Let $W_n^\star$ be similar for $\Sigma^\star$.

On the basis of experimentation with the program described in Kelton and Law (1985), which computes exact values for $\mathrm{E}[W_n]$ and $\mathrm{E}[W_n^\star]$, we propose the following.

Conjecture 2 Transient mean waiting times in the $M/M/s$ queue Σ are bounded above by those in the $M/M/1$ queue $\Sigma^\star$:

$$n \geq 0 \quad \Rightarrow \quad \mathrm{E}[W_n] \leq \mathrm{E}[W_n^\star].$$

Bhaskaran (1986) proves a similar result, but does not cover the choice of $\mu_1^\star$ we use. To provide support for this conjecture, we used the program described in Kelton and Law (1985) to give exact values for $E[W_n]$ and $E[W_n^\star]$. It is only necessary to experiment for $\lambda_1 = 1$, and a range of traffic intensities $\rho = \lambda_1/(s\mu_1)$; otherwise, we can rescale time by dividing input and service rates by the input rate.

We tested very low and very high intensities (0.01 and 0.99) as approximate boundary cases. We tested low, middle, and high intensities (0.10, 0.50, and 0.90) to cover the range of intermediate intensities. Second, the lowest non-trivial number of servers is two, so we tested that as a boundary case. For intensities 0.01 and 0.10, going beyond three servers resulted in extremely small waiting times, so we tested only two and three servers for these intensities. For higher intensities, we tested a reasonable range of servers (2, 3, 5, 10, and 25) Finally, for all intensities except 0.99, we tested a range of job numbers so that at the upper end of the range the waiting times were within 1% of the limiting value α_1. For intensity 0.99, we chose the upper limit so as to make the computation time reasonable (approximately twelve hours on a Sun SPARCstation for each choice of s). For all these cases, Conjecture 2 holds.

5 NETWORKS

Consider a Jackson network Σ composed of N queues as described in Section 2. In this section we conjecture that the total mean waiting times for this network are bounded above by a weighted sum of transient mean waiting times from a network decomposition. For $j = 1, \ldots, N$, let $\Sigma_j^\star$ denote an $M/M/1$ queue, independent of Σ and all $\Sigma_k^\star, k \neq j$, with input rate λ_j and service rate $\mu_j^\star$, where $\mu_j^\star$ is chosen in an analogous fashion to Equation (1). Denote the waiting time for job n in $\Sigma_j^\star$ by $W_{n,j}^\star$, given that job 0 arrives to find $\Sigma_j^\star$ empty.

Conjecture 3 After some small initial period, transient mean total waiting times in the Jackson network Σ are bounded above by the weighted sum of expected waiting times in queues $\Sigma_j^\star$. That is, there exists some n_Σ so that

$$n \geq n_\Sigma \quad \Rightarrow \quad E[W_n] \leq \sum_{j=1}^{N} \frac{\lambda_j}{\lambda} E[W_{n,j}^\star]. \qquad (2)$$

To provide support for this conjecture, we experimented for two network models: small (five nodes) and large (fifty nodes). We estimated the expected network waiting times $E[W_n]$ from simulation, and calculated the exact expected waiting times $E[W_{n,j}^\star]$ using the program described in Kelton and Law (1985). The small network is the single sector configuration displayed in Figure 2. This network is meant to mimic one sector of a larger semiconductor manufacturing line, where jobs are processed sequentially on machines one through four, then with probability 0.5 are reworked at machine five and started again at machine one. Figure 6 compares the conjectured upper bound (2) with the simulated values $E[W_n]$.

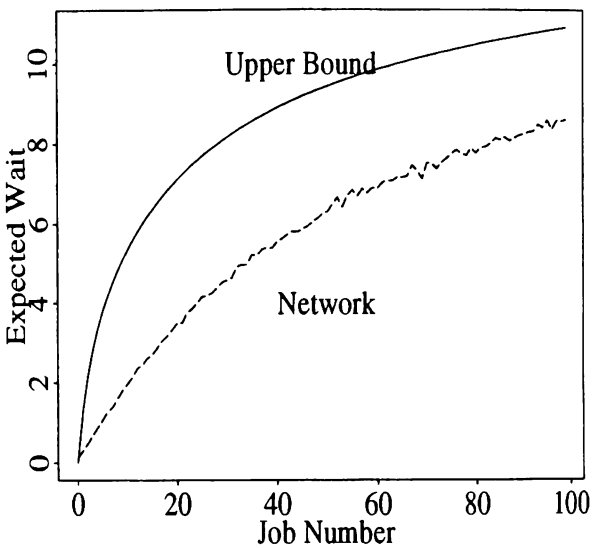

Figure 6: Comparison of Transient Mean Waiting Times and Conjectured Upper Bound in Small Network

Figure 7 displays the large network configuration. For simplicity, all $\mu_j = 1$, and $s_j = 1$ unless otherwise noted in Figure 7. This network is meant to mimic a larger manufacturing environment, including scrap, parallel routing, rework, and multi-level assembly. Figure 8 displays the simulated expected values $E[W_n]$ and the conjectured upper bound (2). The conjectured bound does not appear to be very tight, based on Figures 6 and 8. It does appear to be valid, however, after a very small n_Σ. See Chance (1993b) for more applications of Conjecture 3, including bottleneck analysis.

6 WHY THE CONJECTURED BOUND MIGHT BE LOOSE FOR LARGE NETWORKS

In the previous section, we gave numerical results indicating that the large network example warms up

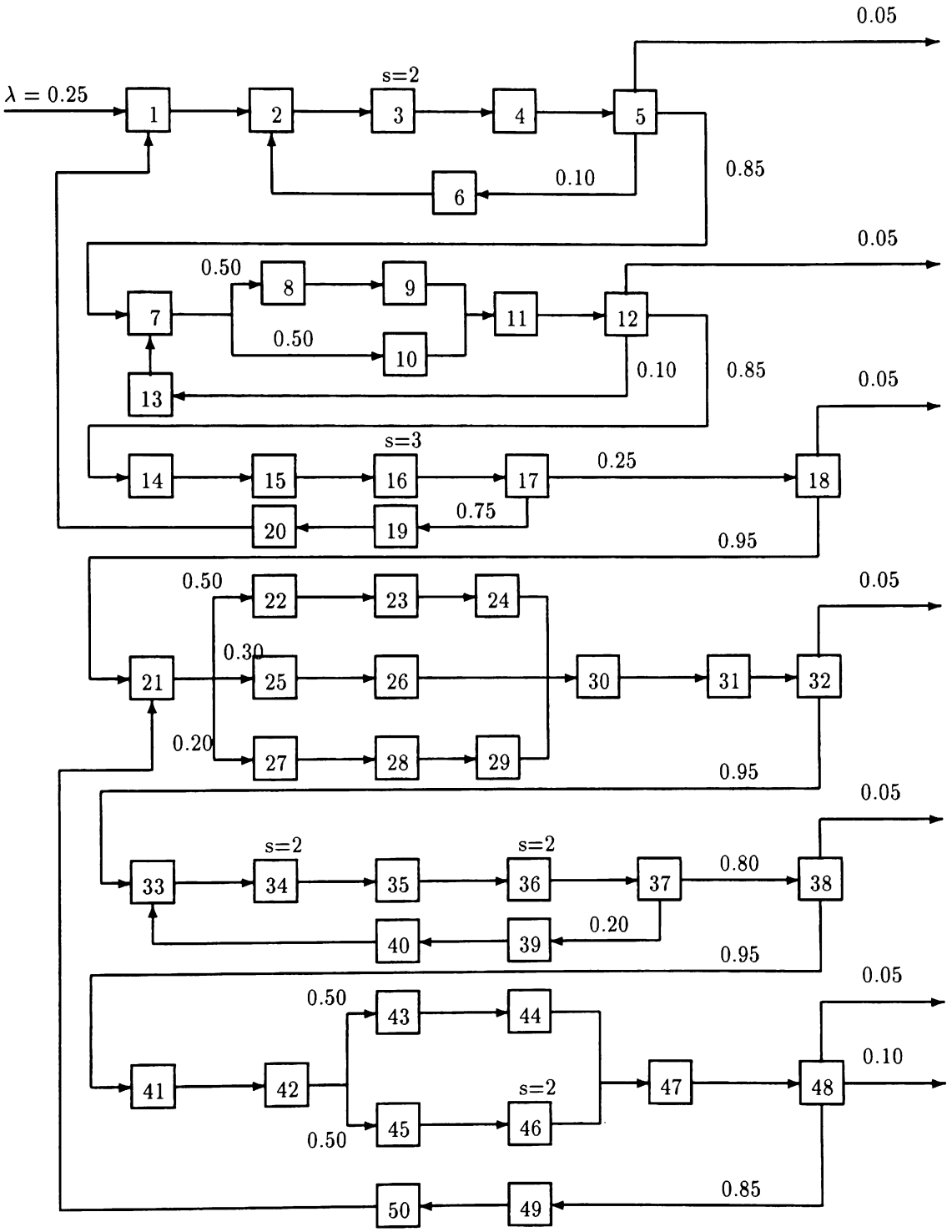

Figure 7: Large Network Configuration

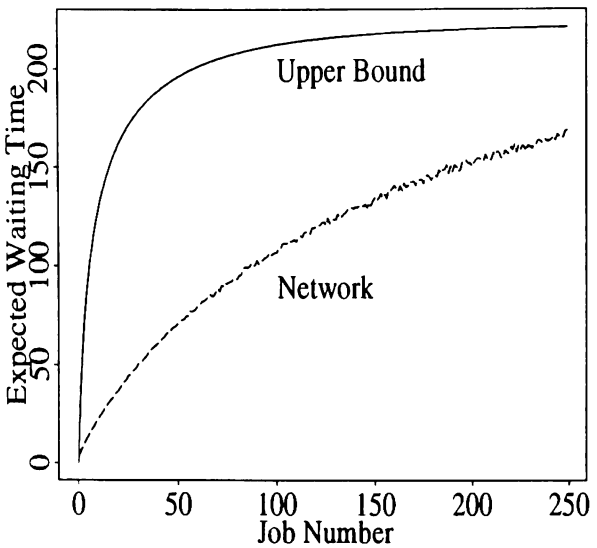

Figure 8: Comparison of Transient Mean Waiting Times and Conjectured Upper Bound in Large Network

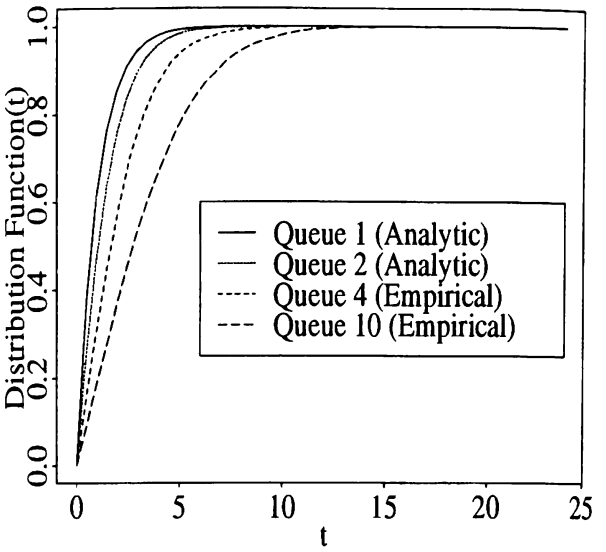

Figure 9: Comparison of First Interarrival Time Distributions for Queues in a Tandem Jackson Network

much more slowly than the sum of its independent queues. In this section we seek to understand this behavior. We consider a tandem Jackson network of N single server queues, with the assumptions and terminology of Section 2. New jobs arrive at queue 1 according to a rate λ Poisson process. After completing service at queue j, jobs join queue $j+1$ if $j < N$, or exit the system if $j = N$. We briefly present some analytic and empirical results suggesting that the transient interarrival times at downstream queues are stochastically larger than at the first queue, and thus the transient input processes are not identical. From these results, we propose that these downstream queues warm up more slowly than if they were supplied by independent, rate λ Poisson arrival processes.

Denote the time between the arrival of job n and job $n+1$ to queue j by A_n^j. For $j = 1$, the A_n^j are i.i.d. rate λ exponential random variables. In steady-state all the A_n^j are i.i.d. rate λ exponential random variables (see Gross and Harris 1985, pp. 221-223). For finite n, this is not the case. For example, Figure 9 shows the distribution functions of the first interarrival times for queues one, two, four, and ten in a ten queue tandem Jackson network. The distribution functions for downstream queues appear to be dominated by those of upstream queues, and hence downstream interarrival times will tend to be larger than those upstream.

Denote the number in system (waiting plus in service) just after the nth departure at queue j by X_n^j, and the service time of job n at queue j by S_n^j. For queue $j > 1$, the nth interarrival time depends on X_n^{j-1} in the following manner

$$A_n^j = \begin{cases} S_{n+1}^{j-1} & X_n^{j-1} > 0, \\ S_{n+1}^{j-1} + Y & X_n^{j-1} = 0, \end{cases}$$

where Y is the length of an idle period at queue $j-1$ conditioned on the system being empty after the departure of job n.

In particular, for $j = 2$ and $n = 0$, Y is a rate λ exponential random variable independent of S_1^1, S_0^1, and A_0^1; hence

$$\begin{aligned} P[A_0^2 \le t] &= P[S_1^1 \le t | X_0^1 > 0] P[X_0^1 > 0] + \\ &\quad P[S_1^1 + Y \le t | X_0^1 = 0] P[X_0^1 = 0]. \end{aligned}$$

Using the algebraic manipulator MACSYMA to evaluate $P[S_1^1 + Y \le t]$ and to simplify the resulting expression, we find

$$P[A_0^2 \le t] = 1 - \frac{e^{-\lambda t}}{1-\rho_1^2} + \frac{\rho_1^2 e^{-\mu_1 t}}{1-\rho_1^2}.$$

Thus the finite-time arrival process to queue 2 is not Poisson, since it does not have exponentially distributed interarrival times.

Starting from the assumption $\rho_1 < 1$, it is a matter of algebraic manipulation to show

$$P[A_0^2 \le t] < P[A_0^1 \le t],$$

and hence the first interarrival time at queue two is stochastically larger than the first interarrival time at queue one. Conceptually, we can think of queue one as operating on its input process, in this case by *stretching* it. Thus even though in the limit these queues have identically distributed input processes, for finite times the interarrival times are larger for downstream queues. When we approximate the network by a sum of independent queues, we are ignoring this effect. It appears that this effect increases with the size of the network.

ACKNOWLEDGMENTS

The author would like to thank Sarah Hood and Gerry Feigin for their advice on this line of study, Lee Schruben and David Goldsman for direction and encouragement, and Robin Roundy and Stuart Carr for their helpful comments on a draft of this paper. Research supported by the Advanced Research Projects Agency.

REFERENCES

Bhaskaran, B.G. 1986. Almost Sure Comparison of Birth and Death Processes with Application to $M/M/s$ Queueing Systems. *Queueing Systems* 1: 103-127.

Chance, F. 1993a. Delphi: A C-Based Queuing Network Simulator. Technical Report No. 1045, School of Operations Research and Industrial Engineering, Cornell University.

Chance, F. 1993b. *The Indifference Approach to the Analysis of Transient Effects in Queuing Networks and Simulations.* PhD thesis, Cornell University.

Connors, D., G. Feigin, and D. Yao. 1992. A Queueing Network Model for Semiconductor Manufacturing. Research Report, IBM T. J. Watson Research Center.

Disney, R.L. 1981. Queueing Networks. *Proceedings of Symposia in Applied Mathematics* 25: 53-83.

Gross, D. and C.M. Harris. 1985. *Fundamentals of Queueing Theory.* New York: John Wiley & Sons.

Hood, S.J., A.E.B. Amamoto, and A.T. Vandenberge. 1989. A Modular Structure for a Highly Detailed Model of Semiconductor Manufacturing. In: *Proceedings of the 1989 Winter Simulation Conference*, 811-817. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Hunter, J.J. 1988. Sojourn Time Problems in Feedback Queues. Technical Report No. '88-127, Mathematical Sciences Institute, Cornell University.

Jackson, J.R. 1957. Networks of Waiting Lines. *Operations Research* 5: 518-521.

Kelton, W.D. and A.M. Law. 1985. The Transient Behavior of the $M/M/s$ Queue, with Implications for Steady-State Simulation. *Operations Research* 33: 378-396.

AUTHOR BIOGRAPHY

FRANK D. CHANCE is a visiting scientist in the School of Operations Research and Industrial Engineering at Cornell University. His research interests are manufacturing and queuing network simulations.