TCSA
https://dergipark.org.tr/tr/pub/tcsa
Vol.1 No.1
May 2024: 1-1
Accepted 29 April 2024
Published 2 June 2024

Transactions on Computer Science and Applications

Sojourn distributions for particular customers in networks of queues

Russell E. King[a,*]
[a]Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, NC 27695, USA, ORCID: https://orcid.org/0000-0003-4576-6600

Corresponding author email address: king@ncsu.edu, **phone:** 1-919-515-816

Abstract

In this paper a study of the transient behavior of sojourn distributions of particular customers traversing serial networks of single-server queues is presented. It is motivated by the need to project completion times of critical customers in loaded, capacitated queueing systems. In particular, serial networks with First-Come-First-Served queueing discipline which do not allow overtaking are considered. An analytic model based on a Markovian state space is shown to be computationally prohibitive even for relatively small scenarios. Given the limitation of the exact solution, heuristic schemes, based on a characterization of the behavior of the exact solution and the Central Limit Theorem, are developed as an alternative to digital Monte-Carlo simulation. A hybrid technique combining the estimated mean from one of the heuristics and the estimated variance from another proves to be accurate and efficient in approximating the mean and variance of the sojourn distribution in a variety of application scenarios.

Keywords: Serial queueing networks sojourn time approximation phase-type distributions

1. Introduction

In this paper a study of the sojourn-time distribution for a job traversing a network of queueing stations is presented. While much work has been done on determining such distributions under "steady-state" or equilibrium conditions, the same cannot be said for "given scenario" conditions. A "given scenario" is a given state of the system at a particular instant in time (queue lengths, server statuses, etc.). The emphasis of this paper is in the presentation and evaluation of techniques for determining sojourn-time distributions for particular jobs in serial networks of single-server, First-Come-First-Served (FCFS) Markovian queues. Of particular interest are simple heuristics that can provide fast, accurate results that potentially can be used in real or near-real time application tools.

The motivation for studying queueing systems of this type stems from production problems encountered by management of repair and maintenance job shops, like the U.S. Navy's Fleet Readiness Centers (FRC's). An FRC is a large job-shop where planned, periodic depot-level maintenance (i.e., overhaul) of aircraft, engines, aircraft components, and ground support equipment is performed. Shops within an FRC, like many jobshops, are generally operated at near capacity limits so that backlogs of work-in-process almost always exist. Management's goal is to provide timely, yet cost-effective repairs so as to maintain fleet readiness. One very visible measure of performance is the ability or inability to meet scheduled completion dates. A major concern is the early detection and expeditious handling of any jobs that have fallen behind schedule. Currently available information systems can provide data on the status of any job, however, this type of information only allows managers to try to fight "fires" as they occur.

A more appropriate tool would be an information system that could, in real-time, predict the "fires" before they actually occur. Such a system would look forward in time and project the completion time for any job and provide information on potential bottlenecks, i.e., critically overloaded shops. Given such information, unsatisfactory completions and bottlenecks could be responded to by raising job priorities, buying overtime, or shifting manpower before the problem actually occurs.

It should be noted that sojourn-time results for arbitrary customers in an open network under "given scenario" conditions have been largely ignored. This is quite understandable given the severe difficulty of obtaining analytic results in other than very simple cases. The main interest of this work is the determination of the mean and variance of the sojourn times for a particular customer under "given scenario" conditions. We will consider three approaches to this problem.

One approach is based on the observation that in loaded systems the Central Limit Theorem applies, thus simple, deterministic procedures yield good approximations. The sec-

ond approach stems from characterization of the exact solution. This approach considers the paths through a Markovian state space network. The majority of the paths through the network in moderate to heavy traffic scenarios have a simple structure whose distributions can be obtained in a computationally efficient manner. The third approach is a hybrid of the first two, combining the benefits of both.

2. Literature review

Some of the first results on sojourn times in queueing networks were presented by Reich (1957). He proved that for a tandem configuration of M/M/1 queues with FCFS queue discipline, the steady-state sojourn times at each queue are independent of each other. Reich (1963) extended this result to an arbitrary number of such queues in tandem. Burke (1972) provided an alternative proof based upon a reversibility argument and the fact that in such a system customers cannot be overtaken or passed by later arriving customers. This overtaking condition was shown by Takacs (1962) to bring on dependencies among the partial sojourn times incurred at individual nodes. These dependencies lead to enormous difficulties in analysis. Simon and Foley (1979) and Mitrani (1979) also demonstrated this property in correcting assumptions made by Lemoine (1977).

Chow (1980) derived the Laplace-Stieltjes transform of the cycle time (which is equal to the sum of two successive response times) distribution for the case of exponentially distributed service times at two queues. Schassburger and Daduna (1983) generalized the results of Chow (1980) by deriving the distribution of cycle time for a closed cycle of many arbitrary single-server queues.

Walrand and Varaiya (1980) considered an open multiclass Jackson network in equilibrium. They showed that the sojourn times of a customer at the various nodes of a non-overtaking path are all mutually independent. They also give two examples to show that the non-overtaking condition may be necessary to ensure independence when there is a single customer class.

Harrison (1984) proved that in steady-state the cycle time distribution is a combination of M Erlang N density functions for an arbitrary customer in a closed cycle of M (M/M/1) queues and for tree-like M/M/1 queues where N is the total number of customers. For systems with m servers at a queue in tree-like networks the component distributions are shown to be convolutions of the Erlang N distributions with at most $m-1$ exponentials.

Daduna (1982) considered closed multi-class Gordon-Newell networks Gordon and Newell (1967) and proved that the passage time through an overtake free path is a mixture of Erlang distributions where the mixing distribution is given by the steady state behavior of the network at arrival times at the path. Daduna (1984) extended his results to include the case where the first and last nodes can be multi-server queues.

Boxma and Donk (1982) determined the joint distribution of consecutive sojourn times of a customer in a closed cycle of two single-server exponential queues. The result is the Laplace-Stieltjes transform of the distribution which is proved using a reversibility argument. These results and the results of Schassburger and Daduna (1983) are generalized by Boxma et al. (1984) for arbitrarily many single-server queues. Kelly and Pollett (1983) extended and unified the results of Daduna (1982) and Boxma et al. (1984). They found the joint distribution of consecutive sojourn times for a customer along an overtake free path in a closed multiclass Jackson network. In this case the individual sojourn times are not independent but the joint distribution has a relatively simple representation in terms of the product form of the stationary state distribution at an arrival instant.

Grassmann (1977b) and Grassmann (1977a) applied the method of randomization to find transient solutions for the M/M/1 queue problem. Melamed (1982) considered a queueing network with types of customers, Poisson arrivals, type dependent routing and state dependent services. The main purpose was to put together some results concerning various aspects of sojourn times in a class of queueing networks of considerable generality. He made use of Little's standard formula to derive limiting mean sojourn times conditioned on customer type and path. Melamed and Yadin (1984a) proposed a methodology utilizing and generalizing the randomization procedure to approximate sojourn time distributions in discrete-state Markovian queueing networks. Melamed and Yadin (1984b) presented the computational aspects of this methodology. Also an optimal storage scheme was described for open Jackson networks.

As Grassmann (1977b) noted, the method of randomization can be "very convenient in cases where it is difficult to obtain the distributions by other methods." Such cases include the handling of arrivals to the system and customer overtaking. However, randomization does not take advantage of the structure in the case where there are no arrivals to the system and no customer overtaking. This structure, namely, the upper triangularity of the transition matrix allows for simple solutions using back substitution.

3. Exact analytic solution

The problem can be described as a series configuration of M shops (see Figure 1), each having a FCFS queue and a single server. The service times at shop j, $(j = 1, ..., M)$ form a sequence of i.i.d. random variables of the phased exponential type (generalized Erlang) having p_j phases. The rate of the k^{th} stage is constant and equal to μ_{jk}.

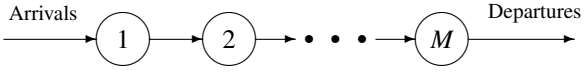


Fig. 1 Serial network of single-server queues

Given the problem as described above, the requirement is to specify the sojourn-time distribution (or remaining

sojourn-time distribution) for a particular arriving job (or any job currently in the system) which finds the system in some state $\mathbf{s}_0 \in \Omega(\mathbf{s}_0)$ = state space. Notice that any job behind the particular job will not affect its sojourn-time since the service rates are constant and service discipline is FCFS. Therefore, those jobs behind the particular job or new arrivals can be eliminated from further consideration. The sojourn-time distribution can be easily shown to be a weighted sum of generalized Erlang terms, Harrison (1980).

Given that there are M shops, there is a maximum of M events which could cause a transition out of any state $\mathbf{u}$, that is, a completion of a phase of service at any of the M shops. Thus, the time to transition out of state $\mathbf{u}$ is equal to the minimum of up to M exponential time intervals. It is well known that this time is exponentially distributed with rate $\lambda_{\mathbf{u}}$, where

$$\lambda_{\mathbf{u}} = \sum_{\substack{1 \le j \le M \\ r(j) \neq 0}} \mu_{jr(j)}$$

with $r(j)$ = phase of service in progress at shop j ($1 \le r(j) \le p_j$). Thus, $\lambda_{\mathbf{u}}$ is the holding rate of state $\mathbf{u}$.

Let $f^*_{\mathbf{u}}(s)$ be the Laplace-Stieltjes Transform ($\mathcal{LST}$) of the holding time density function of state $\mathbf{u}$, $f_{\mathbf{u}}(t)$, i.e.,

$$\mathcal{L}[f_{\mathbf{u}}(t)] = \int_0^\infty e^{-st} df_{\mathbf{u}}(t) = \frac{\lambda_{\mathbf{u}}}{s + \lambda_{\mathbf{u}}}.$$

Define the functions $\phi(\mathbf{u}, \mathbf{v})$ and $\psi(\mathbf{u}, \mathbf{v})$ such that

$\phi(\mathbf{u}, \mathbf{v})$ = number of the shop from which a completion of a phase of service causes the transition from $\mathbf{u}$ to $\mathbf{v}$, ($\mathbf{u}, \mathbf{v} \in \Omega(\mathbf{s}_0)$), which is 0 if a one-step transition $\mathbf{u} \to \mathbf{v}$ is not possible.

$\psi(\mathbf{u}, \mathbf{v})$ = number of the phase of service at shop $\phi(\mathbf{u}, \mathbf{v})$ which causes a transition from $\mathbf{u}$ to $\mathbf{v}$, which is 0 if a one-step transition $\mathbf{u} \to \mathbf{v}$ is not possible.

Then the transition probability from state $\mathbf{u}$ to $\mathbf{v}$ of the underlying Markov process is

$$T_{\mathbf{uv}} = \frac{\lambda_{\phi(\mathbf{u},\mathbf{v}),\psi(\mathbf{u},\mathbf{v})}}{\lambda_{\mathbf{u}}}$$

where $\lambda_{0,0} \equiv 0$. Let $\mathbf{T}$ be the matrix of these state transition probabilities. Then, for any row in $\mathbf{T}$ the number of non-zero entries in that row will be at most M, i.e., the number of events that cause a transition out of a given state is equal to the number of shops that are active in that state.

Let $\mathbf{T}^*$ be a modified transition matrix such that:

$T^*_{\mathbf{uv}}$ = weighted $\mathcal{LST}$ of the holding time distribution of state $\mathbf{u}$ given the next state is $\mathbf{v}$,
$= T_{\mathbf{uv}} f^*_{\mathbf{u}}(s)$
$= \frac{\mu_{\phi(\mathbf{u},\mathbf{v}),\psi(\mathbf{u},\mathbf{v})}}{\lambda_{\mathbf{u}}} * \frac{\lambda_{\mathbf{u}}}{s+\lambda_{\mathbf{u}}} = \frac{\mu_{\phi(\mathbf{u},\mathbf{v}),\psi(\mathbf{u},\mathbf{v})}}{s+\lambda_{\mathbf{u}}}$.

Now, let the cumulative distribution function of the time for the system to transition from state α to state $\beta(\alpha, \beta \in \Omega(\mathbf{s}_0))$ be $G_{\alpha\beta}(t)$ and let $G^*_{\alpha\beta}(s)$ be its $\mathcal{LST}$. Further define $H_{\alpha\beta}$, such that,

$$\begin{aligned} H_{\alpha\beta} = \; & \{\mathbf{h} = (h_1, h_2, ..., h_n) \mid n \in Z^+; \} \\ & h_1 = \alpha; h_n = \beta; h_k \in \Omega(\mathbf{s}_0); T_{k,k+1} \neq 0; 1 \le k < n\} \\ & \text{where: } Z^+ \text{ is the set of positive integers} \end{aligned}$$

$H_{\alpha\beta}$ represents the set of all possible sequences of state transitions, or paths, from state α to β. Each path in $H_{\alpha\beta}$ corresponds to a unique realization of state transitions where each transition interval is exponentially distributed with known rate. Thus, the time to traverse each path is a convolution of exponential distributions, i.e., generalized Erlang.

The following is easily proved.

$$G^*_{\alpha\beta}(s) = (I - \mathbf{T}^*)^{-1}_{\alpha\beta}. \tag{1}$$

The proof involves weighting the contributions of every path in $H_{\alpha\beta}$. With an appropriate ordering of the states, it is easily seen that the transition matrix $\mathbf{T}$, and thus $\mathbf{T}^*$ and $(I - \mathbf{T}^*)$ is upper-triangular and can be solved by simple back substitution.

Thus, the distribution of $G_{\alpha\beta}(t)$ is the so-called generalized hyper-Erlang, that is, a weighted combination of generalized Erlang distributions. Each of these general Erlang distributions corresponds to a unique path from state α to β. The weight associated with each term is the probability of traversing the corresponding path.

While the back substitution solution of $(I - \mathbf{T}^*)^{-1}$ is a straightforward and simplistic procedure, there still are computational problems to overcome. The number of back substitutions required to solve $(I - \mathbf{T}^*)^{-1}_{\alpha\beta}$ is proportional to the number of states in $\Omega(\mathbf{s}_0)$. However, the number of terms in the solution is equal to the number of paths from state α to state β. This value is a function of the number of shops, the number of service phases at each shop, the initial state α, and the final state β.

For simplicity, consider the exponential service case, i.e., $p_j = 1$, for all shops j. Let $f_{H_{\alpha\beta}}(M) = \mid H_{\alpha\beta} \mid$. This value is sum of the $f_{H_{\gamma\beta}}(M)$ values for all one-step reachable states γ from state α and can be obtained recursively, i.e.,

$$f_{H_{\alpha\beta}}(M) = \sum_{\gamma \in J_\alpha} f_{H_{\gamma\beta}}(M)$$

where J_α = set of 1-step reachable states from α, i.e. $\{\gamma \mid \phi(\alpha, \gamma) > 0\}$. Recall that $\mid J_\alpha \mid \le M$.

Table 1 gives representative values of $f_{H_{\alpha\beta}}(M)$ for $\alpha = \mathbf{s}_0$ and β is the empty state, i.e., the state which represents the completion of service of the particular job at shop M. From this table it can be easily seen that the number of paths explodes as the number of customers in each queue increases and/or the number of shops grows. Therefore, a large number of calculations and a large amount of storage space is necessary to keep track of the paths.

Table 1
Number of paths in some representative scenarios

# Customers at Shop: 1	2	3	4	Number of States	Number of Paths
0	0	5	5	51	1.638×10^3
0	0	10	10	176	1.574×10^7
0	5	5	5	506	6.726×10^{10}
0	10	10	10	3311	4.541×10^{23}
5	5	5	5	5481	3.866×10^{24}
10	10	10	10	27636	1.064×10^{38}

All values in this table reflect exponential service at each shop.

4. Heuristic solution procedures based on the CLT

Perhaps the simplest heuristic scheme is to assume that all service times are deterministic (and equal to the mean of the given service distribution) and make a single pass simulation using these fixed service times. This would give an approximation for the mean sojourn time. The variance can be approximated by simply summing variances given the service times are assumed independent. This is the basic scheme of the Deterministic Simulation Heuristic (DSH).

The reason that the technique is effective lies in the symmetry of the associated distributions. It should be intuitively obvious that using mean values should lead to good estimate of the overall mean even for skewed distributions. However, for the variance, two key factors affect the performance of such a heuristic. First, for the Markovian systems under consideration, the variance of the individual shop sojourn times are proportionally additive. That is, for a fixed queue length, L, the variance of the sojourn time at a shop is simply a convolution of L identical generalized Erlang distributions which is itself generalized Erlang. It is well known that the variance of a generalized Erlang distribution is simply the sum of the variances of the individual exponentially distributed stages. The second factor affecting the performance of the heuristic is the inherent symmetry of the shop sojourn times. This is due to the normalizing effect (Central Limit Theorem) of convolving many exponentially distributed time intervals. These factors allow for averaging of variances across all possible values of L to yield an effective estimate of the true variance of the shop sojourn times.

An inherent problem with the DSH technique occurs in situations where the service variances at the individual shops are large and, according to the heuristic, the particular job finds no jobs or only a few jobs ahead of it at a given shop. These cases are potentially troublesome because of the skewed shop sojourn-time distributions that result.

The coefficient of skewness, α_3, is a measure of skewness. For symmetric distributions, like the normal, $\alpha_3 = 0$. However, for single tailed distributions (like sojourn distributions), α_3 approaches unity as skewness decreases. An N stage special Erlang, for example, is asymptotically normal (by the Central Limit Theorem) and has $\alpha_3 \approx 1$ as $N \to \infty$. This is demonstrated in Table 2 for a single, special-Erlang server at each shop. From this table it can be seen that as the queue size increases and/or the number of phases of service increases (for a constant mean service time, hence tighter variance) the coefficient of skewness decreases. If the number of exponential stages convolved is sufficiently large then $\alpha_3 \approx 1$ and hence the sojourn distribution is effectively single-tailed symmetric. The DSH technique assumes that the sojourn time distribution for the particular job at each shop in the serial network is effectively symmetric. However, if the sojourn distribution at a shop is skewed because of small queue sizes then a deterministic approximation will tend to underestimate the true shop sojourn mean.

Table 2
Coefficient of skewness for some representative distributions

Stages of Service	Number in Queue 1	2	3	4	5
1	2.12	1.63	1.44	1.34	1.28
2	1.63	1.34	1.23	1.18	1.14
3	1.44	1.23	1.16	1.12	1.10
5	1.28	1.14	1.10	1.07	1.06
10	1.14	1.07	1.05	1.04	1.03

An enhanced heuristic, DPL, was developed to account for this problem. If the number of queued jobs at a given shop j upon arrival of the particular job is large then DPL uses the DSH approximation for the mean and variance contribution at that shop. However, if this value is small, then the estimate for the shop sojourn is adjusted. The DPL heuristic incorporates a probabilistic estimate for the shop sojourn mean for those troublesome cases just described. This estimate is based on the following observations.

Consider a simple two queue (shop) network with exponential servers working at rates μ_1 and μ_2, respectively. Assume that both servers are busy. and the queue at shop 1 is empty and the queue at shop 2 contains 1 job. Let the job in service at shop 1 be considered the particular job, then the probability, P_n, that the particular customer completes service and finds a total of n jobs at shop 2, $n = 1, 2, 3$, is given in Equation 2.

$$\begin{aligned} P_3 &= \frac{(\mu_1^2+\mu_1\mu_2)}{(\mu_1+\mu_2)^2} \\ P_2 &= \frac{\mu_1\mu_2}{(\mu_1+\mu_2)^2} \\ P_1 &= \frac{\mu_2^2}{(\mu_1+\mu_2)^2} \end{aligned} \tag{2}$$

Equation 3 shows these same probabilities for a system with 2-stage special Erlang service distributions.

$$\begin{aligned} P_3 &= \frac{\mu_1^3+3\mu_1^2\mu_2}{(\mu_1+\mu_2)^3} \\ P_2 &= \frac{3\mu_1^3\mu_2^2+7\mu_1^2\mu_2^3}{(\mu_1+\mu_2)^5} \\ P_1 &= \frac{5\mu_1\mu_2^4+\mu_2^5}{(\mu_1+\mu_2)^5} \end{aligned} \tag{3}$$

Note that $P_3 \approx 0.5$ for nearly equal rates, μ_1, μ_2, in both cases. The same is true for more generalized Erlang distributions. However, to avoid the complicating factors associated

with determining the exact probabilities of the number of jobs the particular job will find at successive shops, the DPL heuristic uses a simplification of the above derivation. It adjusts the probabilities derived for the exponential case for use in those cases where staged Erlang service is assumed. The adjustment gives more weight towards the events associated with fewer jobs than the weight derived above for the exponential case. This scheme is in keeping with the goal of a simple, yet accurate, heuristic.

For the following, let L represent the number of jobs potentially at shop j upon arrival of the particular job according to the DSH technique and let L_{max} be a decision point. If the DSH technique finds that $L \geq L_{max}$, then the standard DSH estimates for the shop sojourn mean and variance at j are used. However, if this value is less than L_{max} the following is used. Let

$$P_{L_{max}} = \frac{\hat{\mu}_{j-1}}{\hat{\mu}_j + \hat{\mu}_{j-1}}$$

$$P_l = P_{L_{max}} \left(\frac{\hat{\mu}_j}{\hat{\mu}_j + \hat{\mu}_{j-1}}\right)^{p_j * (L_{max} - l)}$$

$$P_1 = 1.0 - \sum_{l=2}^{L} P_l$$

where $l = 2, ..., L_{max} - 1$ and p_j = number of phases of service at shop j and $\hat{\mu}_j$ is the total service rate of shop j.

The estimate of the mean shop sojourn time at each shop j, $MSJ(j)$, and the shop variance, $SV(j)$, are then computed from the following.

$$MSJ(j) = \sum_{l=1}^{L} \frac{lP_l}{\hat{\mu}_j}$$

$$SV(j) = \sum_{l=1}^{L} \frac{lP_l r_j}{\hat{\mu}_j}$$

Based upon empirical evidence and the coefficients of skewness shown in Table 2, setting $L_{max} = 3$ provides the best estimates. Defining $\mathcal{T}$ as the sojourn time for a particular customer, then

5. Experimental design

An experimentation plan was developed in order to evaluate the effectiveness of the heuristics presented. The plan consists of a set of example scenarios which was designed to test the heuristics under conditions similar to those found in the heavy traffic facilities described in Section 1. Scenarios which did not fall into this category were also included to test the robustness of the heuristics.

Under steady-state conditions heavy traffic is defined as ρ values near unity, however, for transient "given scenario" conditions the meaning is unclear. For the purposes of this research *heavy traffic* is defined as the condition under which there is a very low probability that the particular customer will arrive at a shop and find it empty. *Light traffic* will refer to cases where this probability is high.

Experimentation was designed around four key parameters: the number of shops, the initial queue lengths, the shop service rates, and the service distributions. The number of shops was set at four levels; 2, 3, 5, and 10. The queue lengths were varied from 1 to 12. The service rates were normalized to a base value of 1.0. For purposes of emulating heavy traffic facilities, rates between 0.8 and 1.0 were used. However, rates as low as 0.2 were also included to test robustness. These light traffic examples were included in the two and three shop scenarios. A total of 595 example scenarios were developed.

Each scenario was solved for each of 5 different service time distributions for a total of nearly 3000 problems. The service distributions included the exponential distribution and four special Erlang distributions where the number of stages was 2, 3, 5 and 10, respectively. These distributions provided a means to determine the effect of the variance on the heuristics. Henceforth, the example datasets with the exponential service times are referred to as the Model 1 datasets, those with a 2-stage Erlang distribution are Model 2, those with a 3-stage are Model 3, those with a 5-stage are Model 5, and finally those with a 10-stage are Model 10.

6. Results for CLT-based heuristics

In this section the results of the heuristic DSH and DPL techniques are compared to those from a Monte-Carlo simulation model for the examples just described. These results are broken down by mean comparison and standard deviation comparison.

6.1. Comparison of means

Due to the computational problems of the exact technique, a Monte-Carlo simulation model was employed for comparison purposes for all datasets. The number of iterations of the simulation was gauged to provide estimates within 1% of the exact values at the $\alpha = 0.05$ level.

As mentioned previously both heavy and light traffic scenarios were tested to evaluate the robustness of the heuristics. Table 3 summarizes the effectiveness of both heuristics. As expected the DPL technique outperforms DSH in the light traffic scenarios since skewed sojourn times occur more often. The difference in the estimate of the mean for DSH is nearly twice that of DPL in the high variance model (Model 1). The difference between the two tapers off with the service variance as the shop sojourn times become less skewed. Under light traffic the percentage differences for the DPL technique are very good ranging from a worst case of 8.73% down to 1.09%.

Notice that in the heavy traffic examples the difference between DSH and DPL is insignificant. Both techniques provide very good approximations generally within about 2% of the simulated results for all but the highest service variance cases (Models 1 and 2).

6.2. *Comparison of standard deviations*

The standard deviation from the heuristic was compared with that from the Monte-Carlo simulation for the entire dataset. The comparisons summarized by average absolute difference in Table 4.

Both techniques tend to overestimate the total sojourn variance. Empirical evidence indicates that there is an inherent negative correlation between successive shop sojourn times. This makes sense intuitively since it would be expected that a particularly long stay at a shop would result in a shorter stay at the next shop since the extra time would allow more jobs at that next shop to complete before the particular job arrives. Neither heuristic attempted to account for this correlation due to the complex nature of this phenomenon.

Contrary to the estimates of the mean, DSH tends to provide slightly better estimates in the light traffic scenarios. This is understandable given the way the variance is estimated. With both techniques, the mean estimate influences the estimate of the variance through the estimated number of completions at each shop. Therefore, a higher estimate of the mean produces a higher estimate of variance. The DSH technique tends to underestimate the true mean in light traffic. DPL provides a better and hence larger estimate of the mean. Thus, the DSH estimate of the variance is smaller and, given the correlation, generally better. However, in the heavy traffic cases there is little difference because situations which cause the heuristics to produce different estimates occur infrequently.

7. Heuristic techniques based on the analytic solution

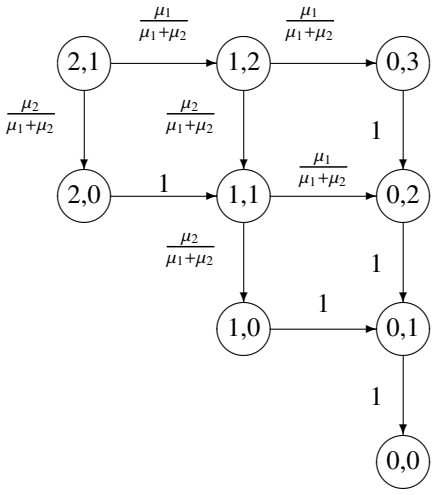
The Path Heuristic was developed to solve the tandem queue problem through the analysis of "paths" through a Markovian state space network. Before proceeding further, it is useful to define a few terms and consider a Markovian state transition diagram.

Let

i = the number of entities initially at shop 1
j = the number of entities initially at shop 2
i' = the number of entities at shop 1 in some arbitrary state
j' = the number of entities at shop 2 in some arbitrary state.

The example in Figure 2 is a tandem arrangement of two exponential queues with initial state ($i = 2, j = 1$) and service rates μ_1 and μ_2 for servers 1 and 2, respectively.

An example path for this problem starts from state (2,1) and proceeds through (1,2), (1,1), (0,2), and (0,1) to (0,0). The probability of traversing a path is simply the product of the probabilities of traversing the arcs between the states. For example, the path described above has probability $(\frac{\mu_1}{\mu_1+\mu_2})^2(\frac{\mu_2}{\mu_1+\mu_2})$. Note that the probability of the majority of the paths take on the general form $(\frac{\mu_1}{\mu_1+\mu_2})^i(\frac{\mu_2}{\mu_1+\mu_2})^k$, where



(NOTE: Arc values represent traversal probabilities)

Fig. 2 Diagram of state transition probabilities

k = the number of server 2 completions in the path before reaching a state with $i' = 0$.

There are some paths, those through $j' = 0$ states, that do not follow this general form. For example, the path (2,1), (2,0), (1,1), (0,2), (0,1) to (0,0) has probability $(\frac{\mu_1}{\mu_1+\mu_2})(\frac{\mu_2}{\mu_1+\mu_2})$. Although these paths do not follow the general form, if $i \ll j$ or $\mu_2 \ll \mu_1$ they tend to contribute little to the overall sojourn distribution since few paths of this nature exist and they have small traversal probabilities. In these cases, the sum of the probabilities of all paths is approximately 1, i.e.

$$\sum_{k=0}^{i+j-1} P_{i,j}^k (\frac{\mu_1}{\mu_1+\mu_2})^i (\frac{\mu_2}{\mu_1+\mu_2})^k \approx 1$$

where $i + j - 1$ = the maximum level for k, $P_{i,j}^k$ = the total number of level k paths from the initial state (i, j).

With the path probabilities in hand, we proceed to capture the sojourn times at each level, k. There are three components which comprise the sojourn time of a given path. The first component occurs in the situation where the first and second server are both busy, i.e., an (i', j') state where $i' \leq i, i' \neq 0$ and $j' \leq j, j \neq 0$. There are an estimated $i + k$ of these states. It is an estimate of $i + k$ due to the addition of the second component. The second component occurs in the situation where only the first server is busy, an $(i', 0)$ state. As mentioned previously, there are few occurrences of this state in the set of all paths. Therefore, we have omitted this component and added it to the first component, thus the $i + k$ estimate of occurring states. Thus, the $\mathcal{LST}$ of this component is $(\frac{\mu_1+\mu_2}{s+\mu_1+\mu_2})^{i+k}$.

The third component occurs in the situation where only the second server is busy, a $(0, j')$ state. For a level k path, there are $i + j - k$ of these states and the $\mathcal{LST}$ of this component is $(\frac{\mu_2}{s+\mu_2})^{i+j-k}$.

The weighted combination of the $\mathcal{LST}$ of these sojourn distributions with weights corresponding to their respective traversal probabilities yields the following sojourn distribu-

Table 3
Average absolute percent difference of the mean for DSH and DPL heuristics

Model (Service Stages)	Heuristic Technique	2 Shops Scenarios		3 Shops Scenarios		5 Shops Scenarios	10 Shops Scenarios
		Light	Heavy	Light	Heavy	Heavy	Heavy
1	DSH	5.633	2.119	16.675	4.472	4.65	6.08
	DPL	2.912	1.598	8.730	4.624	4.63	4.26
2	DSH	3.555	0.872	11.736	2.204	1.83	2.34
	DPL	2.638	0.872	7.795	2.148	1.14	1.67
3	DSH	2.612	0.589	9.399	1.435	1.01	1.41
	DPL	2.289	0.589	7.632	1.402	1.00	1.28
5	DSH	1.772	0.302	6.959	0.863	0.45	0.74
	DPL	2.337	0.302	6.565	0.849	0.37	1.24
10	DSH	1.092	0.190	4.595	0.488	0.16	0.29
	DPL	1.091	0.190	4.583	0.488	0.17	1.31

Table 4
Average absolute percent difference for the standard deviation for DSH and DPL heuristics

Model (Service Stages)	Heuristic Technique	2 Shops Scenarios		3 Shops Scenarios		5 Shops Scenarios	10 Shops Scenarios
		Light	Heavy	Light	Heavy	Heavy	Heavy
1	DSH	6.233	5.264	4.006	7.883	13.64	17.49
	DPL	7.198	5.264	8.481	8.107	13.50	17.02
2	DSH	6.925	5.128	4.728	7.559	10.88	12.53
	DPL	7.576	5.128	6.985	7.494	10.72	11.99
3	DSH	7.421	5.054	5.334	7.066	9.74	9.83
	DPL	8.187	5.054	6.085	6.884	9.46	8.82
5	DSH	7.691	4.407	6.565	6.553	7.99	7.50
	DPL	9.174	4.407	6.543	6.309	7.82	6.96
10	DSH	8.816	4.184	10.616	6.441	6.44	5.78
	DPL	10.116	4.184	9.802	6.176	6.28	5.54

tion, $\mathcal{T}$, of the time from an (i, j) state to the (0,0) state, i.e.

$$\begin{aligned} F_{\mathcal{T}}(s) &= \sum_{k=0}^{i+j-1} P_{i,j}^{k} \xi^{i} (1-\xi)^{k} \\ &* \quad [(\tfrac{\mu_1+\mu_2}{s+\mu_1+\mu_2})^{i+k} (\tfrac{\mu_2}{s+\mu_2})^{i+j-k}] \end{aligned} \tag{4}$$

where $\xi = (\frac{\mu_1}{\mu_1+\mu_2})$.

Straightforward computation of the moments of this function can be decomposed into the contributions from each term in the summation. This yields moments for all paths of level k as

$$\begin{aligned} M_1(k) &= P_{i,j}^{k} \xi^{i} (1-\xi)^{k} [(\tfrac{(i+j-k)}{\mu_2}) + (\tfrac{(i+k)}{\mu_1+\mu_2})] \\ M_2(k) &= P_{i,j}^{k} \xi^{i} (1-\xi)^{k} [(\tfrac{(i+j-k)}{\mu_2}) + (\tfrac{(i+k)}{\mu_1+\mu_2})]^2 \\ &+ \quad [(\tfrac{(i+j-k)}{(\mu 2)^2}) + \tfrac{(i+k)}{(\mu 1+\mu 2)^2})]. \end{aligned}$$

The mean of time spent in the system is simply the sum of the first moment contributions at each level, i.e.

$$E[\mathcal{T}] = \sum_{k=0}^{i+j-1} M_1(k). \tag{5}$$

The variance of the time spent in the system is calculated by subtracting the sum of the square of the first moment for all levels of k from the sum of the second moment for all levels of k, i.e.

$$V[\mathcal{T}] = \sum_{k=0}^{i+j-1} M_2(k) - [\sum_{k=0}^{i+j-1} M_1(k)]^2. \tag{6}$$

In order to account for the lost traversal probabilities and lost holding times of the paths through $j' = 0$ states, a "weighted" path is added, i.e., a weighted worse case path (path through all $j' = 0$ states). The weighted value of this path is based on the total lost traversal probability, the service rate of each server, and the total number of $(i', 0)$ states (i of them). Thus, the weighted value is

$$[1 - (\sum_{k=0}^{i+j+1} P_{i,j}^{k} \xi^{i} (1-\xi)^{k})](1-\xi)(i) \tag{7}$$

This factor is rounded up or down about the 0.5 value, since it represents a number of actual $(i', 0)$ states.

The first and second moments of the weighted path are calculated and added to the previous first and second moments to generate the mean and variance of the system.

To test the technique, the two queue scenarios described in Section 5 were solved. In general, the Path Heuristic generates better estimates of the sojourn variance than the CLT-based heuristics while mean estimates are not quite as good. The mean estimates are consistently in the same range while the variance increases with the number of Erlang stages. The Path Heuristic gives its worse estimates in those cases where the service rate at the second server is much faster than the first. Such scenarios would increase the probability of entering an $(i', 0)$ state, thus increasing the error of the estimate.

As with the exact solution, storage problems arise when solving large problems due to the large number of paths through a network (see Table 1).

In order to take advantage of the improved estimates of the variance that this technique gives, a hybrid technique, Deterministic Path Heuristic (DPH), was developed. The DPH technique combines features of the Path, DSH, and DPL techniques.

The DPH technique uses the DPL heuristic to estimate the mean since it yields the best estimates. The DSH and Path Heuristic are combined in order to estimate the variance. This was accomplished by decomposing the M queue model into a series of paired queues. The solution is embedded at queue completion times for the particular customer. In other words, the technique solves for the sojourn time between successive service completions for the particular customer.

Using the Path technique, the sojourn time is found iteratively from the initial (i, j) state to a $(0, j')$ for each pair of queues considered. The DSH heuristic is used to deterministically approximate the number of entities at the next queue, thus setting up another two queue scenario. When the algorithm reaches the final two queues, the Path technique is used to solve for the remaining sojourn time, i.e., it solves them completely to the (0,0) state as before.

For simplicity, each of the two queue problems are assumed separate and independent systems, therefore the total sojourn mean and variance is approximated as the sum of the means of the individual problems. Although these two queue scenarios are not truly independent of the others, this assumption yields good results if the true dependency is small.

8. Results for heuristics based on the analytic solution

As previously mentioned, a Monte-Carlo simulation model was employed for comparison purposes for all datasets. The number of replicates of the simulation model provided confidence interval estimates within 1% of the exact values at the $\alpha = 0.05$ level.

Based on observation of Tables 5 and 6, one can see that the DPL/DPH methods yields the best overall estimates for the mean, and the DPL results in the lowest variance of the test cases. The mean percentage differences range from a worse case of 5.8% down to 0.17% with an overall average percentage difference of 2.58% for all cases. The standard deviation percentage differences range from a worse case of 9.43% to 2.76% with an overall average percentage difference of 5.64%. With 2,268 different scenarios being considered, these values are quite good.

9. Conclusions

This paper has been concerned with the solution of sojourn-time distributions for particular jobs in serial networks of queues. The emphasis was the attainment of such

Table 5
Average absolute percent difference for the mean for all heuristics

Model Number	Heuristic	Number of Shops			
		2	3	5	10
1	DSH	4.91	7.96	4.65	6.08
	DPL and DPH	2.64	5.80	4.63	4.26
2	DSH	3.00	4.93	1.83	2.34
	DPL and DPH	1.76	3.30	1.14	1.67
3	DSH	2.20	3.71	1.01	1.41
	DPL and DPH	1.94	3.18	1.00	1.28
5	DSH	1.47	2.60	0.45	0.74
	DPL and DPH	1.56	2.37	0.37	1.24
10	DSH	0.91	1.66	0.16	0.29
	DPL and DPH	1.64	2.47	0.17	1.31

distributions in a computationally efficient manner for use in real-time or near real-time applications. An exact technique was developed but shown to be computationally prohibitive. The limitations of the exact technique led to the development of heuristic methodologies which could provide a computationally attractive alternative to straightforward Monte-Carlo simulation.

Two simple procedures were developed which were founded in the Central Limit Theorem. Both of these techniques were shown to give good approximations for the mean and variance of the desired distribution for a range of scenarios. In particular, they were very good for the cases which closely resembled the operation of heavy traffic jobshops. A third method considered paths through a Markovian state space network. This method proved to give even better approximations of the variance. This led to the development of a hybrid technique, Deterministic Path Heuristic (DPH), which combined the concepts of both approaches into a single technique.

The DPH technique yielded the best mean estimates by using the DPL method and yielded the best variance estimates by using a combination of the DSH and Path Heuristic. Thus, the DPH gives the best overall results for the 1932 different scenarios considered (see Tables 3 and 6).

Limitations of the DPH lie in the large amount of storage space required for large problems. This is due to the large number of paths that are possible for bigger problems (see Table 1).

Possible future endeavors include modifying the DPH by enhancing the Path heuristic. One possibility is to eliminate the $j' = 0$ paths to see if better estimates can be obtained. This may hold true since these paths add to the error of the current version of the DPH. Another possibility is based upon the observation that the mean calculation used by the DSH, Path Heuristic combination is generally underestimated for each two queue scenario, which leads to overestimates of the variance; this can be seen from Equation 6. These mean calculations can be improved by incorporating the DPL technique for the paired queue calculations. As a result, better variance estimates should be obtainable.

Table 6
Average absolute percent difference for the standard deviation for all heuristics

Model Number	Heuristic	Number of Shops			
		2	3	5	10
1	DSH	6.03	6.78	13.64	17.49
	DPL	6.80	8.21	13.50	17.02
	DPH	2.76	3.79	6.88	9.43
2	DSH	6.56	6.75	10.88	12.53
	DPL	7.47	8.44	10.72	11.99
	DPH	3.63	4.40	7.84	9.38
3	DSH	6.93	6.57	9.74	9.83
	DPL	7.54	6.66	9.46	8.82
	DPH	4.60	4.40	7.85	9.42
5	DSH	7.02	6.56	7.99	7.50
	DPL	7.96	8.13	7.82	6.96
	DPH	5.61	4.70	7.23	6.43
10	DSH	7.86	7.63	6.44	5.78
	DPL	8.82	9.24	6.28	5.54
	DPH	5.84	4.93	5.09	5.27

References

Boxma, O., Donk, P., 1982. On response time and cycle time distributions in a two-stage cyclic queue. Performance Eval 2, pp. 181–194.

Boxma, O., Kelly, F., Konheim, A., 1984. The product form for the sojourn time distribution in cyclic exponential queues. JACM 31, pp. 128–133.

Burke, P., 1972. Output processes and tandem queues. Proc. Symp. Comp.-Comm. Networks and Teletraffic, pp. 419–428.

Chow, W., 1980. The cycle time distribution of exponential cyclic queues. JACM 27, pp. 281–286.

Daduna, H., 1982. Passage times for overtake-free paths in Gordon-Newell networks. Adv Appl Prob 14, pp. 672–686.

Daduna, H., 1984. Burke's theorem on passage time in Gordon-Newell networks. Adv Appl Prob 16, pp. 867–886.

Gordon, W., Newell, G., 1967. Closed queueing systems with exponential servers. Operations Research 15, pp. 254–265.

Grassmann, W., 1977a. Transient solutions in Markovian queueing systems. Computers and OR 4, pp. 47–56.

Grassmann, W., 1977b. Transient solutions in Markovian queues. European Journal of OR 1, pp. 392–402.

Harrison, P., 1980. Distributions of time delays in queueing networks. Unpublished manuscript.

Harrison, P., 1984. A note on cycle times in tree-like queueing networks. Adv. Appl. Prob 16, pp. 216–219.

Kelly, F., Pollett, P., 1983. Sojourn times in closed queueing networks. Adv Appl Prob 15, pp. 638–656.

Lemoine, A., 1977. Networks of queues – a survey of equilibrium analysis. Management Science 24, pp. 464–481.

Melamed, B., 1982. Sojourn times in queueing networks. Math of OR 7, pp. 223–244.

Melamed, B., Yadin, M., 1984a. Numerical computation of sojourn-time distributions in queueing networks. JACM 31, pp. 839–854.

Melamed, B., Yadin, M., 1984b. Randomization procedures in the computation of cumulative-time distributions over discrete state Markov processes. Operations Research 32, pp. 926–944.

Mitrani, I., 1979. A critical note on a result by Lemoine. Management Science 25, pp. 1026–1027.

Reich, E., 1957. Waiting times when queues are in tandem. Ann Math Stat 28, pp. 768–773.

Reich, E., 1963. Note on queueing in tandem. Ann Math Stat 34, pp. 338–341.

Schassburger, R., Daduna, H., 1983. The time for a round exponential queues. JACM 30, pp. 146–150.

Simon, B., Foley, R., 1979. Some results on sojourn times in acyclic Jackson networks. Management Science 25, pp. 1027–1034.

Takacs, L., 1962. Stochastic Processes. John Wiley and Sons, Inc.

Walrand, J., Varaiya, P., 1980. Sojourn times and the overtaking condition in Jacksonian networks. Advances in Applied Probability 12, pp. 1000–1018.