Response time distributions in queueing network models

Peter G. Harrison
Department of Computing
Imperial College
180 Queen's Gate
London SW7 2BZ
England

Abstract Time delays in queueing networks are assuming increasing importance with the proliferation of transaction processing and time-critical real time systems. Mean values are insufficient and it is necessary to estimate time intervals that are not exceeded with a specified probability, i.e. *quantiles*. This paper presents results on time delay distributions in single server queues of various types and extends these to networks of queues. In particular, the class of Jackson networks that permit exact solution are analysed in both the open and closed cases, and approximation techniques for more general networks are proposed.

1 Introduction

The time delays experienced by tasks passing through a sequence of processing nodes define an important class of performance measures in computer-communication systems. Their mean values provide a good overall description of performance and are readily obtained by conventional techniques, but means alone are often insufficient. For example, we may wish to predict the variability of response time in a multi-access system or various reliability measures, such as the probability that a message transmission time will exceed a given value. The importance of obtaining quantiles of distributions—i.e. time intervals that are not exceeded with a specified probability—is becoming increasingly recognised, in particular in transaction-processing systems where quantiles are specified as minimal performance requirements in international standards, such as TPC.

Queueing network models which compute queue length distributions in a steady state network are well established and from the mean queue lengths, mean passage time along a given path can be determined directly. There is now, therefore, a need to consider the more difficult problem of finding the probability distribution of passage-times along a path in a queueing network. Mathematically, the simplest type of network to analyse is open, acyclic and Markovian, i.e. has external arrivals from independent Poisson processes and fixed-rate servers with exponentially distributed service times. The arrival process at every server is then independent and Poisson. Unfortunately, even these assumptions are too weak to allow the distribution of the passage-time along an arbitrary path to be obtained in a simple form. For paths on which a task cannot be overtaken, we can consider passage time as the sum of waiting times at independent single-server (M/M/1) queues and obtain a simple solution. If any of these assumptions is violated, e.g. for any closed network of servers, independence is lost and the above approach fails. However, a more complex result can be derived for overtake-free paths in Markovian closed networks. To derive time delay distributions in more general networks requires approximate methods.

Rather than the distributions themselves, it is generally easier to work with their Laplace transforms. This is because a time delay is a sum of sojourn times (i.e. times spent in some state or at some server) and, if these are independent, the required distribution is a mixture of convolutions of sojourn time distributions. But the Laplace transform of a convolution is the product of the Laplace transforms of the constituent distributions, which is much easier to manipulate than a convolution-integral. To obtain quantiles, of course, it is necessary to be able to invert the Laplace transform of the passage time distribution so as to recover the distribution itself. In general, inversion is by numerical methods which may be difficult to implement accurately. This may be especially so at high quantiles, i.e. in the tail of a distribution—often the most important region. However, analytic inversion is possible in the solvable networks referred to above, including closed, overtake-free, Markovian networks.

This paper is organised as follows. In the next section we consider the waiting time distribution at a single-server queue, beginning with first-come-first-served queueing discipline (i.e. an M/M/1 queue) and then examining the effect of non-exponential service times (i.e. M/G/1 queue), different queueing disciplines and, very briefly, negative customers (of the Gelenbe type, [3]). We then look at passage time distributions through an open, tandem network of M/M/1 queues in section 3; this result extends immediately to open tree-like networks. The Laplace transform of passage time distribution on overtake-free paths in closed Markovian networks is given in section 4 and its analytic inversion is considered in section 5. A case study—transmission time distribution in a packet-switched, multistage interconnection network—may be found in [6]. The paper concludes in section 6 which includes discussion of approximations for Laplace transforms in more general networks. The material is presented in more detail, including proofs of theorems, in Chapter 9 of the book "Performance Modelling of Communication Networks and Computer Architectures" by Harrison and Patel, published by Addison-Wesley (1993).

2 Time delays in the single server queue

There are many intervals of time that are of interest in queueing systems. We begin with the waiting and queueing times of a customer in M/M/1 and M/G/1 queues with FCFS discipline. Another important time interval is the **busy period** (or **busy time**) of the server, i.e. the interval between successive idle periods. In fact, the analysis of busy times will prove a powerful technique and lead, in particular, to the waiting time distribution of an M/G/1 queue with LCFS queueing discipline. PS queueing discipline will also be considered, but only for M/M/1 queues where we can make use of properties of the underlying continuous time Markov chain. In fact, we will see that the method used for PS can also be used for other queueing disciplines in an M/M/1 queue.

2.1 Waiting time distribution in the M/M/1 queue

In this section we investigate the time interval between the instants at which a given customer arrives at an M/M/1 queue and departs after completing service. This random variable is called the customer's **waiting time** and is denoted by T_W; it includes the time spent being served. The corresponding interval from the arrival instant to the instant at which the customer first enters service is called the **queueing time**, denoted T'; it excludes the service time. We consider the classical M/M/1 queue with arrival rate λ and service rate μ independent of the queue length. First, we can calculate the mean waiting time (and queueing time) quite easily using Little's result as follows. We know that the mean equilibrium queue length is $\rho/(1-\rho)$, where $\rho = \lambda/\mu$, and that the mean arrival rate is λ. Hence, mean waiting time W is the ratio of these quantities, $1/(\mu-\lambda)$. For FCFS queueing discipline, we can now find the expected queueing time, Q, from the relation $T_W = T' + S$ where S is the service time random variable, i.e. exponential with parameter μ. Taking expectations gives

$$(\mu-\lambda)^{-1} = E[T_W] = E[T'] + E[S] = Q + \mu^{-1}$$

so that $Q = \rho/(\mu-\lambda)$.

Notice that the result for W holds regardless of the queueing discipline. However, we no longer have this invariance when we consider the probability distribution of waiting time. First, suppose the queueing discipline is FCFS and that immediately after a new arrival, the queue length is $n+1$; i.e. the arrival "faces" a queue of length $n \geq 0$. The arriving customer's waiting time is now a sum of $n+1$ random variables:

$$T_W = \begin{cases} U + S_1 + S_2 + \ldots + S_n & \text{if } n \geq 1 \\ S_1 & \text{if } n = 0 \end{cases}$$

Each S_i is independent and distributed as the service time, i.e. exponential with parameter μ, and U is the residual service time of the customer being served at the arrival instant. But, by the residual life (memoryless) property of exponential distributions, U has the same exponential distribution as service time. Thus, T_W is a sum of $n+1$ independent exponential random variables with parameter μ when the queue length faced on arrival is $n \geq 0$. Similarly, T' is a sum of n such random variables. Since the arrival process is Poisson, by the Random Observer Property, the probability that the queue length faced by an arrival is n is the same as the equilibrium probability that the queue length is n, here $(1-\rho)\rho^n$. Thus, by the law of total probability,

$$P(T_W \leq t) = \sum_{n=0}^{\infty}(1-\rho)\rho^n F^{(n+1)*}(t)$$

where $F(t) = 1 - e^{-\mu t}$ is the service time distribution function and F^{k*} $(k \geq 1)$ denotes the k–fold convolution of F with itself. But $F^{(n+1)*}$ is the Erlang–$(n+1)$ distribution with parameter μ and so the waiting time density function f_W is defined by

$$\begin{aligned} f_W(t) &= \sum_{n=0}^{\infty}(1-\rho)\rho^n\mu\frac{(\mu t)^n}{n!}e^{-\mu t} \\ &= (1-\rho)\mu e^{-\mu t}\sum_{n=0}^{\infty}\frac{(\rho\mu t)^n}{n!} \\ &= (\mu-\lambda)e^{-(\mu-\lambda)t} \end{aligned}$$

Waiting time is therefore exponential with parameter $\mu - \lambda$, as expected from our derivation of the mean waiting time. The fact that waiting time is exponential can actually be deduced by a purely probabilistic argument, using the memoryless properties of both the geometric distribution (of the queue length) and the exponential distribution. We then need only the mean waiting time, which we have already determined, to characterise completely the waiting time random variable. This approach is taken by [11].

We can obtain waiting time distributions for variants of the M/M/1 queue, revealing the sensitivity to different queueing disciplines, for example. If we have a load dependent server, i.e. one with rate depending on the instantaneous queue length, waiting time distribution is much more difficult to obtain. In particular, the derivation of the result for PS discipline is lengthy, even when the arrival and (total) service rates are both constant; we consider this problem below. However, we can quite easily find the waiting time density for the multi-server queue, i.e. the M/M/m queue. In this case, a new arrival has to queue iff the queue length faced on arrival is at least m. Waiting time is now given by:

$$T_W = \begin{cases} X_1 + X_2 + \ldots + X_{n-m+1} + S & \text{if } n \geq m \\ S & \text{if}\, n < m \end{cases}$$

where X_i $(1 \leq i \leq n-m+1)$ is distributed as the service time of a *single* exponential server with rate $m\mu$ and S is distributed as a single exponential server with rate μ. This follows because when the number of customers ahead of the customer being traced is $n, n-1, \ldots, m$, there are m parallel servers active and the superposition of their departure processes is a Poisson process with rate $m\mu$ (i.e. the time to the next service completion is exponential with parameter $m\mu$). But this is exactly the situation with an M/M/1 queue with service rate $m\mu$. When there are fewer than m customers ahead of the customer being traced, including when the queue length faced on arrival is $n < m$, the remaining waiting time is just one service time, S. In this way we obtain (see [7, page 181]):

$$F_Q(t) = \alpha + (1-\alpha)[1 - e^{-(m\mu-\lambda)t}] = 1 - (1-\alpha)e^{-(m\mu-\lambda)t}$$

where α is the equilibrium probability that the queue length is less than m, i.e. the equilibrium probability of not having to queue (by the random observer property of the Poisson process).

Example 2.1. A telephone exchange with holding facilities can be modelled as an M/M/m queue; calls arrive as Poisson processes with total rate λ and each has exponential duration with mean $1/\mu$. How many lines are necessary such that the probability of a caller being "on hold" for more than 1 minute is less than 0.1? We can simply use the above formula for $F_Q(t)$ since the probability of holding time exceeding 1 minute is $1 - F_Q(1)$. Thus we require

$$(1-\alpha)e^{-(m\mu-\lambda)} < 0.1$$

i.e.

$$m\mu - \lambda > log_e 10(1-\alpha)$$

i.e.

$$m > \frac{log_e 10(1-\alpha)+\lambda}{\mu}$$

We already knew that m had to be bigger than λ/μ for stability—the above inequality says by how much in order to get the required performance. Of course, α is a non-trivial function of m, and numerical methods are needed to obtain particular solutions.

2.2 Waiting time distribution in the M/G/1 queue

The waiting time distribution for FCFS discipline is readily obtained from the following observation. For $n \geq 1$, the queue, of length X_n, existing on the departure of the nth customer, C_n, comprises precisely the customers that arrived during that customer's waiting time. In equilibrium, denoting the waiting time distribution of each customer C_n by F_W, the generating function for the queue length may be expressed as:

$$\Pi(z) = E[E[z^X|W]] = E[e^{-\lambda W(1-z)}] = W^*(\lambda(1-z))$$

since X, conditional on W, has Poisson distribution. Writing $\theta = \lambda(1-z)$ so that $z = (\lambda-\theta)/\lambda$, we now have

$$W^*(\theta) = \Pi((\lambda-\theta)/\lambda) = \frac{(1-\rho)\theta B^*(\theta)}{\theta - \lambda[1-B^*(\theta)]}$$

by substituting into the Pollacek-Khintchine formula for Π.

Note that we can now easily check Little's result for the M/G/1 queue since $-W^{*\prime}(0) = -\lambda^{-1}\Pi'(1)$. Notice too that we get the required exponential distribution in the case of an M/M/1 queue where Π is the generating function of the geometric random variable with parameter ρ.

Example 2.2. A rotating disk can be modelled by an M/G/1 queue as follows. Suppose that read/write requests arrive at the head as a Poisson process with parameter λ, requiring blocks of data of fixed length 1 sector, beginning at a random sector boundary. The disk spins at rate r revolutions per second and has s sectors. We make the approximation that the next request to be served always finds the head at a boundary between two sectors—this will in general be violated by arrivals to an empty queue. We require the probability that a request takes more than t time units to complete. There are essentially two problems: to find the Laplace transform of the service time distribution, $B^*(\theta)$, and then to invert the resulting expression for $W^*(\theta)$. To obtain the solution requires numerical methods and we just give the analysis. First, the service time distribution function F_S is defined by

$$F_S(t) = \begin{cases} n/s & \text{if } n \leq rst < n+1 \ (0 \leq n \leq s-1) \\ 1 & \text{if } t \geq 1/r \end{cases}$$

so that the density function is

$$f_S(t) = \frac{1}{s}\sum_{n=1}^{s} \delta\left(t - \frac{n}{sr}\right)$$

The Laplace transform of this density is therefore

$$B^*(\theta) = \frac{1}{s}\sum_{n=1}^{s} e^{-n\theta/sr}$$

from which the Laplace transform of the required waiting time density is

$$W^*(\theta) = \frac{(1-\rho)\theta\sum_{n=1}^{s} e^{-n\theta/sr}}{s\theta - \lambda\left(s - \sum_{n=1}^{s} e^{-n\theta/sr}\right)}$$

by substitution into the above formula.

2.3 Busy periods

To investigate the busy period, we first observe that its distribution is the same for all queueing disciplines that are work conserving and for which the server is never idle when the queue is non-empty. Suppose that, in equilibrium, whilst an initial customer C_1 is being served, customers $C_2, \ldots, C_{Z+1}$ arrive, where the random variable Z, conditional on service time S for C_1, is Poisson with mean λS. Without loss of generality, we assume a LCFS queueing discipline with no preemption so that, if $Z \neq 0$, the second customer to be served is C_{Z+1}. Any other customers that arrive while C_{Z+1} is being served will also be served before C_Z. Now let N be the random variable for the number of customers served during a busy period and let N_i be the number of customers served between the instants at which C_{i+1} commences service and C_i commences service ($1 \leq i \leq Z$). Then $N_1, \ldots, N_Z$ are independent and identically distributed as N. This is because the sets of customers counted by $N_Z, N_{Z-1}, \ldots, N_1$ are disjoint and (excluding $C_{Z+1}, C_Z, \ldots, C_2$ respectively) arrive consecutively after C_{Z+1}. Thus,

$$N \simeq \begin{cases} 1 + N_Z + N_{Z-1} + \ldots + N_1 & \text{if } Z \geq 1 \\ 1 & \text{if } Z = 0 \end{cases}$$

(The symbol $\simeq$ denotes "equal in distribution") Now, denoting the busy time random variable by T, its distribution function by H and the Laplace-Stieltjes transform of H by H^*, we have

$$T \simeq \begin{cases} S + T_Z + T_{Z-1} + \ldots + T_1 & \text{if } Z \geq 1 \\ S & \text{if } Z = 0 \end{cases}$$

where T_i is the length of the interval between the instants at which C_{i+1} commences service and C_i commences service ($1 \leq i \leq Z$). Moreover, the T_i are independent random variables, each distributed as T, and also independent of S. This is because the customers that arrive and complete service during the intervals T_i are disjoint Thus

$$\begin{aligned} H^*(q) &= E[E[E[e^{-\theta T}|Z,S]]|S]] \\ &= E[E[E[e^{-\theta(S+T_1+\ldots+T_Z)}|Z,S]]|S]] \\ &= E[E[e^{-\theta S}E[e^{-\theta T}]^Z|S]] \\ &= E[e^{-\theta S}E[H^*(\theta)^Z|S]] \\ &= E[e^{-\theta S}e^{-\lambda S(1-H^*(\theta))}] \end{aligned}$$

since Z (conditioned on S) is Poisson with mean λS. Thus we obtain

$$H^*(\theta) = B^*(\theta + \lambda(1 - H^*(\theta)))$$

Although this equation cannot be solved in general for $H^*(\theta)$, we can obtain the moments of busy time by differentiating at $\theta = 0$. For example, the mean busy period, m say, is given by

$$-m = H^{*\prime}(0) = B^{*\prime}(0)\{1 + \lambda[-H^{*\prime}(0)]\} = -(1 + \lambda m)\mu^{-1}$$

since $H^*(0) = 1$, and so $m = (\mu - \lambda)^{-1}$, the M/M/1 queue result. The above technique, in which a time delay is defined in terms of independent, identically distributed time delays, is often called "delay cycle analysis" and is due to [13].

2.4 Waiting times in LCFS queues

Now let us consider waiting times under LCFS disciplines. For the preemptive-resume variant, we note that a task's waiting time is independent of the queue length it faces on arrival, since the whole of the queue already there is suspended until after this task completes service. Thus without loss of generality we may assume that the task arrives at an idle server. Waiting time then becomes identical to the busy period. We therefore conclude that the waiting time distribution in a LCFS-PR M/G/1 queue has Laplace-Stieltjes transform $H^*(\theta) = B^*(\theta + \lambda(1 - H^*(\theta)))$.

For LCFS without preemption we can modify the busy period analysis. First, if a task arrives at an empty queue, its waiting time is the same as a service time. Otherwise, its queueing time Q is the sum of the residual service time R of the customer in service and the service times of all other tasks that arrive before it commences service. This definition is almost the same as that of a busy period given above. The only differences are that the time spent in service by the initial customer C_1' (C_1 above) is not a service time but a residual service time and the random variable Z' (Z above) is the number of customers that arrive whilst C_1' is in (residual) service. Proceeding as before, we obtain

$$Q \simeq \begin{cases} R + T_Z + T_{Z-1} + \ldots + T_1 & \text{if } Z \geq 1 \\ R & \text{if } Z = 0 \end{cases}$$

We can now derive the Laplace-Stieltjes transform Q^* of the distribution function of Q similarly to obtain:

$$Q^*(\theta) = R^*(\theta + \lambda(1 - H^*(\theta)))$$

where R^* denotes the Laplace-Stieltjes transform of the probability distribution of R. But since R is a forward recurrence time, $R^*(\theta) = \mu[1 - B^*(\theta)]/\theta$. Thus,

$$Q^*(\theta) = \frac{\mu(1 - B^*(\theta + \lambda(1 - H^*(\theta))))}{\theta + \lambda(1 - H^*(\theta))} = \frac{\mu(1 - H^*(\theta))}{\theta + \lambda(1 - H^*(\theta))}$$

Finally, since a customer arrives at an empty queue with probability $1 - \rho$ in equilibrium, we obtain for the transform of the waiting time distribution

$$\begin{aligned} W^*(\theta) &= (1 - \rho)B^*(\theta) + \rho B^*(\theta)Q^*(\theta) \\ &= B^*(\theta)\left(1 - \rho + \frac{\lambda(1 - H^*(\theta))}{\theta + \lambda(1 - H^*(\theta))}\right) \end{aligned}$$

since waiting time is the sum of queueing time and service time and these two random variables are independent.

Example 2.3. Let us compare the response time variability in a computer system, modelled by an M/G/1 queue, with FCFS and LCFS scheduling policies. We can do this to a great extent by comparing the first two moments which are obtained by differentiating the respective formulae for $W^*(\theta)$ at $\theta = 0$. We obtain the same result for the mean waiting time, which is as expected from Little's result since the mean queue lengths are the same under each discipline. However, it turns out that the second moment of waiting time for FCFS discipline is $1 - \rho$ times that for LCFS. Thus, LCFS discipline suffers a much greater variability as ρ approaches 1, i.e. as the queue begins to saturate. The qualitative result is quite obvious, but the preceding analysis enables the load at which the effect becomes serious to be estimated quantitatively.

2.5 Waiting times with Processor-Sharing discipline

The problem with PS discipline is that the rate at which a customer receives service during his sojourn at a server varies as the queue length changes due to new arrivals and other departures. Thus, we begin by analysing the waiting time density (or rather its Laplace transform) in an M/M/1 queue of a customer with a given service time requirement.

Proposition 2.1 *In a PS M/M/1 queue with fixed arrival rate λ and fixed service rate μ, the Laplace transform of the waiting time density, conditional on a customer's service time being x is*

$$W^*(s|x) = \frac{(1-\rho)(1-\rho r^2)e^{-[\lambda(1-r)+s]x}}{(1-\rho r)^2 - \rho(1-r)^2 e^{-(\mu/r-\lambda r)x}}$$

where r is the smaller root of the equation $\lambda r^2 - (\lambda + \mu + s)r + \mu = 0$ and $\rho = \lambda/\mu$.

This result, proved in [7], was first derived by [1]. We can obtain the Laplace transform of the unconditional waiting time density as

$$W^*(s) = \int_0^\infty W^*(s|x)\mu e^{-\mu x} dx$$

The essential technique used in the proof of Proposition 1 splits the waiting time in an M/M/1 queue into an infinitesimal initial interval and the remaining waiting time. In fact the technique is quite general, applying to more disciplines than PS. In particular, it can be used to find the Laplace transform of the waiting time density in an M/M/1 queue with random discipline or FCFS discipline with certain queue length dependent service rates and in M/M/1 queues with "negative customers", [8].

3 Time delays in open networks of queues

Networks of queues present an entirely different kettle of fish to the case of a single server queue—even a stationary Markovian network. This is because, although we know the distribution of the queue lengths at the time of arrival of a given (tagged) customer at the first queue in his path (by the Random Observer Property or the Job Observer Property), we cannot assume this stationary distribution exists upon arrival at subsequent queues. The reason is that the arrival times at the subsequent queues are only finitely later than the arrival time at the first queue. Hence, the state existing at the subsequent arrival times must be conditioned on the state that existed at the time of arrival at the first queue. Effectively, a new time origin is set at the first arrival time, with known initial joint queue length probability distribution—the stationary distribution. Even in open networks with no feedback, where it is easy to see that all arrival processes are Poisson, this conditioning cannot be overlooked and we cannot assume all queues on a path are independent and in an equilibrium state at the arrival times of the tagged customer. The situation appears even more hopeless in open networks with feedback and closed networks.

However, things are not quite as bad as they seem when we have fixed arrival and service rates. First, we can prove that the FCFS queues in an **overtake-free** path in a Markovian open network behave as if they were independent and in equilibrium when observed at the successive arrival times of a tagged customer. By an overtake-free path, or a path with **no overtaking**, we mean that a customer following this path will depart from its last queue before any other customer that joins any queue in that path after the said custromer. Surprisingly, a similar result holds for overtake-free paths in closed networks, e.g. all paths in networks with a tree-like structure—see Figure 2. In the next subsection, we consider those open networks for which a solution for the time delay density along a given path can be derived. This is followed by a discussion of the problems that confront us when we attempt to generalise the network structure. Closed networks are considered in the next main section.

3.1 Tandem networks

The simplest open network we can consider is a pair of queues in series. However, it is almost as easy to analyse a tandem series of any number of queues, as shown in Figure 1. In fact, we can be more general than this, as we will see shortly.

Figure 1: A tandem series of queues

Now, the distribution of the time delay of a customer passing through a tandem series of queues is the convolution of the stationary waiting time distributions of each queue in the series considered in isolation. This follows from the following result.

Proposition 3.1 *In a series of stationary M/M/1 queues with FCFS discipline, the waiting times of a given customer in each queue are independent.*

Proof

First we claim that the waiting time of a tagged customer, C say, in a stationary M/M/1 queue is independent of the departure process before the departure of C. This is a direct consequence of reversibility since C's waiting time is clearly independent of the arrival process after C's arrival under FCFS discipline at a single server. Applying this property to the stochastically identical reversed process, a corresponding customer C' arrives at the negative time of departure of C and departs at the negative time of arrival of C. It therefore has the same waiting time as C and the claim follows in the original process by the duality between the processes.

To complete the proof, let A_i, T_i denote C's time of arrival and waiting time respectively at queue i in a series of m queues $(1 \leq i \leq m)$. Certainly, by our claim, T_1 is independent of the arrival process at queue 2 before A_2 and so of the queue length faced by C on arrival at queue 2. Now, we can ignore customers that leave queue 1 after C since they cannot arrive at any queue in the series before C, again because all queues have single servers and FCFS discipline. Thus, T_2 is independent of T_1 and similarly T_1 is independent of the arrival process at queue i before A_i and so of T_i for $2 \leq i \leq m$. Similarly, T_j is independent of T_k for $2 \leq j < k \leq m$. ♣

From this proposition it follows that, since the waiting time probability density at the stationary queue i, considered in isolation $(1 \leq i \leq m)$, has Laplace transform $(\mu_i - \lambda)/(s + \mu_i - \lambda)$, the density of the time to pass through the whole series of m queues is the convolution of these densities, with Laplace transform $\prod_{i=1}^{m}(\mu_i - \lambda)/(s + \mu_i - \lambda)$.

There is one obvious generalisation of this result: the final queue in the series need not be M/M/1 since we are not concerned with its output. Also, the same result holds, by the same reasoning, when the final queue is M/G/n for $n \geq 1$. Moreover, Proposition 3.1 generalises to **treelike networks** which are defined as follows, and illustrated in Figure 2. A treelike network consists of:

- a linear **trunk segment** containing one or more queues in tandem, the first being called the **root** queue;
- a number (greater than or equal to zero) of disjoint **subtrees**, i.e. treelike subnetworks, such that customers can pass to the roots of the subtrees from the last queue in the trunk segment or else leave the network with specified routing probabilities (which sum to 1).

The **leaf** queues (or **leaves**) are those from which customers leave the network.

The proof of Proposition 3.1, extended to treelike networks, carries through unchanged since every path in the network is overtake-free. Hence we can ignore the customers that leave any queue on the path after the tagged customer. Indeed, we can generalise further to overtake-free paths in any Markovian open network for the same reason. Conditional on the choice of path of queues, numbered, without loss of generality, $1, \ldots, m$, the Laplace transform of the passage time density is the same as for the tandem queue of m servers considered above.

To generalise the network structure further leads to serious problems and solutions have been obtained only for very special cases. The simplest case of a network with overtaking is the following three-queue network.

In this network, the path of queues numbered $\{1, 3\}$ is overtake-free and so the passage time density can be obtained as described above. However, overtaking is possible on the path $\{1, 2, 3\}$ since when the tagged customer C is at queue 2, any customers departing queue 1 (after C) can reach queue 3 first. The arrival processes to every queue

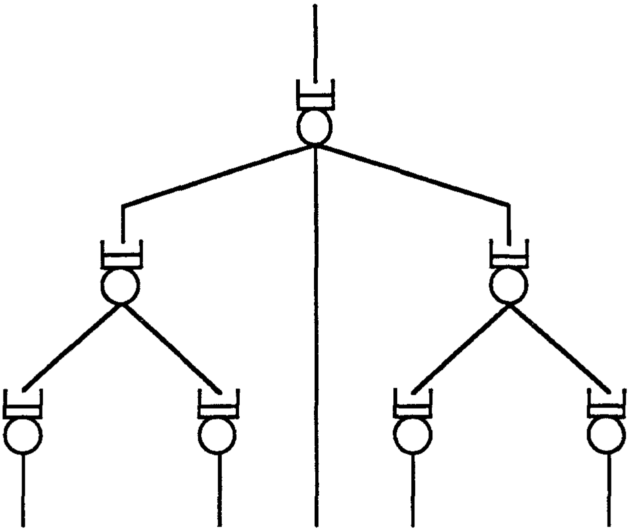

Figure 2: An open tree-like network

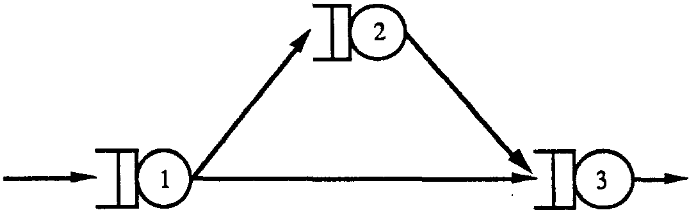

Figure 3: A three-node network with overtaking

in this network are independent Poisson, by Burke's theorem together with the decomposition and superposition properties of Poisson processes. However, this is not sufficient for the passage time distribution to be the convolution of the stationary sojourn time distributions at each queue on a path with overtaking: the proof of Proposition 3.1 breaks down. This particular problem has been solved, by considering the state of the system at the departure instant of the tagged customer from server 1 and using complex variable methods; see [10]. A similar analysis is required—for similar reasons—to analyse a tandem pair of queues with negative customers, [12]. In this case, negative arrivals at the second queue allow the first queue to influence the sojourn time of a tagged customer in the second; departures from the first queue offer a degree of "protection". More general networks appear intractable.

4 Time delays in closed networks

As for the case of open networks, we begin with the simplest case, a cyclic network that comprises a tandem network with departures from its last queue fed back into the first queue. There are no external arrivals and hence a constant population. Again, all service disciplines are FCFS and all service rates are constant.

We solve for the Laplace transform of the cycle time density by considering a dual network, viz. the tandem, open network consisting of the same servers $1, \ldots, m$ with no external arrivals. Eventually, therefore, the dual network has no customers, i.e. its state is $e = (0, 0, ..., 0)$, the empty state, with probability 1. All other states with one or more customers are transient. Now, given that the state immediately after the arrival of the tagged customer at queue 1 is i, the ensuing cycle time in the closed network is the same as the time interval between the dual network entering states i and e—the (first) passage time from i to e. This is so because there is no overtaking and service rates are constant. Thus the progress of the tagged customer in its cycle cannot be influenced by any customer behind it. We only need consider customers ahead of the tagged customer and can ignore those recycling after leaving the last queue. Observe that if service rates varied with queue length, we could not ignore customers behind the tagged customer, even though they could not overtake, because they would influence the service rate received by the tagged customer.

We therefore seek the density of the first passage time from state i to e in the dual network, $f(t|i)$, where i is a state of the form $(i_1, \ldots, i_m)$ with $i_1 > 0$, corresponding to the tagged customer having just arrived at server 1. We know the probability distribution of the state seen by the tagged customer on arrival at the first queue by the Job Observer Property and so can calculate the cycle time density by deconditioning f.

Given a cyclic network of population n, let the state space of the dual network be $S_n = \{(u_1, \ldots, u_m) | 0 \leq u_i \leq n, 1 \leq i \leq m; \sum_{i=1}^{m} u_i \leq n\}$ and define, for $u \in S_n$,

$$\lambda_u = \sum_{i=1}^{m} \mu_i \epsilon(u_i)$$

where μ_i is the service rate of server i, $\epsilon(n) = 1$ if $n > 0$ and $\epsilon(0) = 0$. Thus λ_u is the total service rate in state u, i.e. the instantaneous transition rate out of state u in the Markov process defining the queueing network. The holding time in state u is an exponential random variable with parameter λ_u and so has a density with Laplace transform $\lambda_u/(s+\lambda_u)$. Now, given that the network next enters state v after u, the passage time from u to e is the sum of the holding time in state u and the passage time from v to e. Thus, the density of the passage time from u to e has Laplace transform $L(s|u)$ given by the equations

$$\begin{aligned} L(s|u) &= \sum_{v\in S_n} q_{uv}\frac{\lambda_u}{s+\lambda_u}L(s|v) \qquad\qquad u\neq e \\ L(s|e) &= 1 \end{aligned}$$

where q_{uv} is the one-step transition probability from state u to v. Now let $\mu(u,v)$ denote the rate of the server from which a departure causes the state transition $u \to v$. Then $q_{uv} = \mu(u,v)/\lambda_u$. Thus, writing $q^*_{uv} = \mu(u,v)/(s+\lambda_u)$, we have the matrix equation

$$\mathbf{L} = \mathbf{Q}^*\mathbf{L} + \mathbf{1_e}$$

where $\mathbf{L} = (L(s|u)|u \in S_n)$, $\mathbf{Q}^* = (q^*_{uv}|u,v \in S_n)$ and $\mathbf{1_e}$ is the vector with component corresponding to state e equal to 1 and the rest 0.

Using this equation and deconditioning the state u seen on arrival via the Job Observer Property, we can obtain a product form for the Laplace transform of the cycle time probability density function. More generally, however, we consider cycle times in closed tree-like queueing networks. Such networks are defined in the same way as open tree-like networks except that customers departing from leaf-queues next visit the root queue. Clearly such networks have the no-overtaking property and if paths are restricted to start at one given server (here the root), they define the most general class for which it holds.

Now let Z denote the set of all paths through a closed tree-like network A, i.e. sequences of servers entered in passage through A. For all $z = (z_1, \ldots, z_k) \in Z$, $z_1 = 1$, z_k is a leaf-server and the order of Z is the number of leaf servers since there is only one path from the root to a given leaf in a tree. The probability of choosing path z is equal to the product of the routing probabilities between successive component centres in z. The Laplace transform of cycle time density is given by the following Proposition. The most general result, viz. the multidimensional Laplace transform of the joint density of the sojourn times spent by the tagged customer at each server on any overtake-free path in a network with multiple classes is given by [9]. The proof given in [7] is simpler, being based on the recursive properties of trees. At the same time the result is almost as general in that any overtake-free path must be tree-like (although several such intersecting paths could exist in the whole network) and the extension to multiple classes and joint sojourn times is straightforward.

Proposition 4.1 *For the closed tree-like network A of M servers, the Laplace transform of cycle time density, conditional on choice of path $z \in Z$ is*

$$L(s|z) = \frac{1}{G(n-1)}\sum_{u\in S(n-1)}\prod_{i=1}^{M}\left(\frac{e_i}{\mu_i}\right)^{u_i}\prod_{j=1}^{|z|}\left(\frac{\mu_{z_j}}{s+\mu_{z_j}}\right)^{u_{z_j}+1}$$

where $|z|$ is the number of servers in path z, $S(k)$ is the state space of the network when its population is $k \geq 1$, e_i and μ_i are the respective visitation rate and service rate of server i, and G is the network's normalising constant function.

In fact Proposition 4.1 holds for any overtake-free path in an arbitrary closed Jackson queueing network (recall the preceding discussion) and this form of the result is used in [6].

5 Inversion of the Laplace transforms

The majority of results on distributions of time delays in queueing networks and passage times in more general stochastic processes are given as Laplace (or Laplace-Stieltjes) transforms. The preceding is no exception. In general, numerical methods must be used to invert the Laplace transform which can be expensive to implement and are sometimes unreliable. However, in certain cases, analytical inversion is possible, typically when a stochastic model is based on exponential distributions. The result of Proposition 4.1 is a good example. First, we can simplify the summation giving $L(s|z)$ by partitioning the sum over $S(n-1)$ according to the total number of customers, p, at servers in the overtake-free path $1, 2, \ldots, m$ (say, without loss of generality). Now, the Laplace transforms in the inner sum are products of the Laplace transforms of Erlang densities. Moreover, their coefficients are geometric. Such transforms can be inverted analytically. In the simplest case, all the servers on the overtake-free path are identical, i.e. have the same rate, and the inversion can be done by inspection. In the case that the μ_i are all distinct ($1 \leq i \leq m$), the density function is derived in [5] and the question of degenerate μ_i is addressed in [6]. These results are considered in the next two sections.

5.1 Overtake-free paths with identical servers

When all the rates μ_i in the path are the same, equal to μ say, the above Laplace transform is a mixed sum of terms of the form $[\mu/(s+\mu)]^{p+m}$ since in the inner summation $\sum_{i=1}^{m} u_i + 1 = p + m$. Each term can therefore be inverted by inspection to give a corresponding mixture of Erlangians for the passage time density. We therefore have:

Proposition 5.1 *If the centres in overtake-free path* $1, 2, \ldots, m$ *in the network of Proposition 4.1 all have service rate* μ*, the path's time delay density function is*

$$\frac{\mu^m e^{-\mu t}}{G(n-1)} \sum_{p=0}^{n-1} G_m(n-p-1) G^m(p) \mu^p \frac{t^{p+m-1}}{(p+m-1)!}$$

where $G^m(k)$ *is the normalising constant for the subnetwork comprising servers* $1, \ldots, m$ *only, with population* $k \geq 0$*, and* $G_m(k)$ *is the normalising constant of the whole network with servers* $1, \ldots, m$ *removed and population* $k \geq 0$.

From this result we can immediately obtain formulae for moments higher than the mean of a customer's transmission time.

Corollary

For a path of equal rate servers, message transmission time has kth moment equal to

$$\frac{1}{\mu^k G(n-1)} \sum_{p=0}^{n-1} G_m(n-p-1) G^m(p)(p+m)\ldots(p+m-k+1)$$

5.2 Overtake-free paths with distinct servers

The case of paths with equal rate servers is easy, involving only some algebraic manipulaton of summations. However, even when the rates are different, the inversion can be done analytically to give a closed form result. The analysis is now rather more difficult, however, and we just state the main result after giving a sketch of its derivation. The result was first derived by the author, [5], for the case where all the service rates on the overtake-free path are distinct, the opposite extreme to the previous section. The first step is to invert the Laplace transform $L(\mathbf{n}, s) = \prod_{i=1}^{m} [\mu_i/(s+\mu_i)]^{n_i}$ where $\mathbf{n} = (n_1, \ldots, n_m)$, $n_i \geq 1$ and the μ_is are distinct. This yields the density function

$$f(n,t) = \prod_{i=1}^{m} \mu_i^{n_i} \sum_{j=1}^{m} D_j(\mathbf{n},t)$$

where the $D_j(\mathbf{n},t)$ are given by the following recurrence on $\mathbf{n}$:

$$(n_j - 1) D_j(\mathbf{n},t) = t D_j(\mathbf{n}_j,t) - \sum_{k \neq j} n_k D_j(\mathbf{n}_j^k,t)$$

with boundary condition

$$D_j(\mathbf{n},t) = \frac{e^{-\mu_j t}}{\prod_{i \neq j} (\mu_i - \mu_j)^{n_i}} \qquad (n_i \geq 1, n_j = 1)$$

where $\mathbf{n}_j = (n_1, \ldots, n_j - 1, \ldots, n_m)$ and $\mathbf{n}_j^k = (n_1, \ldots, n_j - 1, \ldots, n_k + 1, \ldots, n_m)$.

Next, given real numbers $a_1, \ldots, a_M$ for integer $M \geq m$, define

$$H_{jm}(\mathbf{z}) = \sum_{\mathbf{n} \in S(M+m)} D_j(\mathbf{n},t) \prod_{i=1}^{M} (a_i z_i)^{n_i - 1}$$

so that passage time density is obtained from the $H_{jm}(1, \ldots, 1)$ with $a_i = e_i/\mu_i$. The central result is an expression for $H_{jm}(\mathbf{z})$ from which follows:

Proposition 5.2 *If the servers in an overtake-free path* $1, 2, \ldots, m$ *have distinct service rates* $\mu_1, \mu_2, \ldots, \mu_m$*, the passage time density function, conditional on the choice of path, is*

$$\frac{\prod_{i=1}^{m} \mu_i}{G(n-1)} \sum_{p=0}^{n-1} G_m(n-p-1) \sum_{j=1}^{m} \frac{e^{-\mu_j t}}{\prod_{i \neq j} (\mu_i - \mu_j)}$$
$$\times \sum_{i=0}^{p} \frac{(e_j t)^{p-i}}{(p-i)!} \sum_{\mathbf{n} \in S_m(m+i), n_j = 1} \prod_{1 \leq k \neq j \leq m} \left(\frac{e_k - e_j}{\mu_k - \mu_j} \right)^{n_k - 1}$$

where $S_m(k)$ *denotes the state space for the subnetwork of servers* $1, \ldots, m$ *with population* k*.*

The summations over $S_m(m+i)$ are just normalising constants that may be computed efficiently along with the $G_m(n-p-1)$ and $G(n-1)$ by Buzen's algorithm.

Example 5.1. For a *cyclic* network of M exponential servers and population N, cycle time distribution is

$$\frac{\left(\prod_{i=1}^{M} \mu_i\right) t^{N-1}}{(N-1)!G(N)} \sum_{j=1}^{M} \frac{e^{-\mu_j t}}{\prod_{i \neq j}(\mu_i - \mu_j)}$$

This follows by setting $e_1 = ... = e_M = 1$ in Proposition 5.2, so that all terms are zero in the rightmost sum except when $n_k = 1$ for all k, i.e. when $i = 0$. Finally, note there is only one partition of the state space, namely the one with all $N-1$ customers at the servers $1, \ldots, M$. Thus we have $G_M(n) = 1$ if $n = 0$ and $G_M(n) = 0$ if $n > 0$, so that only terms with $p = N-1$ give a non-zero contribution.

Proposition 5.2 can be generalised to allow arbitrary service rates at the nodes on an overtake-free path: not necessarily all the same nor all distinct. Essentially, we start with the case of distinct rates and successively combine any two servers with equal rates. The combination inolves manipulation of the summations and reduces the problem to two similar problems on networks with one less node in the overtake-free path. Thus, in each step, one degenerate server is removed until all the remaining problems are on paths with distinct rate servers. The details may be found in [6].

6 Conclusion

We have seen that finding time delay densities is a hard problem, often with complex and computationally expensive solutions when they can be found at all. Consequently, in most practical applications, the performance engineer requires approximate methods. There is no single established methodology for such approximation and most of the techniques used are *ad hoc*. In increasing order of sophistication, the following techniques have been used.

- A particular form is prescribed for the required distribution and its parameters are determined by matching moments. Moments may be predicted by an analytical model or estimated by simulation or actual measurement. Typical distributions include Coxian (with a small number of phases), Generalised Exponential and (mixtures of) Erlang. Although adequate for some purposes, involving probabilities near the median, for example, this approximation lacks a cause and effect relationship and is likely to be poor in the tail region in particular.

- A common simplifying assumption is that the queues in the path of a tagged customer in a queueing network behave as if independent, isolated and in equilibrium at the times of arrival of that customer; often called the **independence approximation.** The assumption is always true for the first queue in the path by the arrival theorem (with one less customer in the case of a closed network) but approximate for all the other queues, except in the case of simple open networks of the type we considered in section 3. The approximation is poorest when the ordering of customers in the network is most highly constrained, since then the independence assumption is clearly invalid. For example, in a 2-node cyclic network with FCFS queues and population N, it is known with probability one that if there are k customers at server 1 at any time, then there are $N-k$ at server 2. In particular, suppose server 1 is fast and its queue is empty on arrival of the tagged customer. Then it is very unlikely that queue 2 will be empty on arrival there and very likely that it will contain $N-1$ or $N-2$ customers. It does turn out that cyclic networks with FCFS queues give poor results under the independence approximation, but in networks where the ordering of customers has few constraints, for example richly connected networks or networks with PS discipline at many queues, it is usually quite accurate.

- An enhancement of the independence approximation admits limited dependence of the queue lengths faced by the tagged customer at successive servers. It is assumed that the queue length faced at any queue entered after the first in the path (which is independent by the arrival theorem) depends only on that faced at the previous node. This is called the **Paired Centre Approximation** and gives accurate results in a variety of queueing networks [4].
- Finally, it might be possible to use maximum entropy methods in continuous time to give the "least surprising" density function for a time delay subject to the constraints imposed by its moments. The maximum entropy method has been used mainly for discrete random variables in computer performance modelling and has produced accurate approximations for the state space distributions in a variety of networks. A continuous time analogue exists and appears well suited to predicting time delay distributions efficiently, given the expected values of certain functions of the state random variable. As usual, the most important step would be to identify and estimate the crucial constraints, but this is an open problem. The reader is referred to Kouvatsos's tutorial on this subject.

As with any approximate model, the above methods are subject to validation. The exact results described in the previous sections provide valuable benchmarks for this purpose. An approximation that passes these tests should be subjected to simulation testing and compared with real observations before being accepted as a performance engineering tool.

References

[1] E.G. Coffman Jnr, R.R. Muntz, H. Trotter
Waiting time distribution for processor-sharing systems
JACM 17, pp123—30, 1970

[2] H. Daduna
Passage times for overtake-free paths in Gordon-Newell networks
Adv. Appl. Prob. 14, pp672—86, 1982

[3] E. Gelenbe, P. Glynn, K. Sigman
Queues with negative arrivals
J. Appl. Prob. 28, pp245—50, 1991

[4] P.G. Harrison
An enhanced approximation by pair-wise analysis of servers for time delay distributions in queueing networks
IEEE Transactions on Computers C-35,1, pp54—61, 1986

[5] P.G. Harrison
Laplace transform inversion and passage time distributions in Markov processes
J. Appl. Prob. 27, pp74—87, 1990

[6] P.G. Harrison
On non-uniform packet switched delta networks and the hot-spot effect
IEE Proceedings E 138, 3, pp123—30, 1991

[7] P.G. Harrison, N.M. Patel
Performance Modelling of Communication Networks and Computer Architectures
Addison-Wesley, 1993

[8] P.G. Harrison, E. Pitel
Sojourn times in single server queues with negative customers
J. Appl. Prob., 1993 (to appear)

[9] F.P. Kelly, P.K. Pollett
Sojourn times in closed queueing networks
Adv. Appl. Prob. 15, 638—56, 1983

[10] I. Mitrani
Response time problems in communication networks
J. Roy. Stat. Soc. B-47, 3, pp396-406, 1985

[11] I. Mitrani
Modelling of Computer and Communication Systems
Cambridge University Press, 1987

[12] E. Pitel
Queues with negative customers and their applications
PhD Thesis, Department of Computing, Imperial College, University of London, 1994 (in preparation)

[13] L. Takacs
Introduction to the theory of queues
Oxford University Press, 1962