10 Single-server models

The topic of this chapter is the exact and approximate analysis of $G/G/1$ queues. In the next section we first consider the model with phase-type service times. Phase-type distributions may be used to approximate any non-negative distribution arbitrarily close; see e.g. [9, 10]. The waiting time distribution for this model can be determined exactly by means of the spectral expansion method. In Section 10.2 we present some simple approximations for the first two moments and the distribution of the waiting time for the model with general (not necessarily phase-type) service times.

10.1 The $G/PH/1$ queue

We will study a single-server queue with phase-type service times and arbitrarily distributed interarrival times. The interarrival time distribution is denoted by $F_A(\cdot)$ with mean $1/\lambda$. The service times have a mixed Erlang-r distribution with scale parameter μ. This means that with probability q_n the service time is the sum of n exponential phases with the same parameter μ, $n = 1, 2, \ldots, r$. The phase representation of this distribution is shown in figure 1.

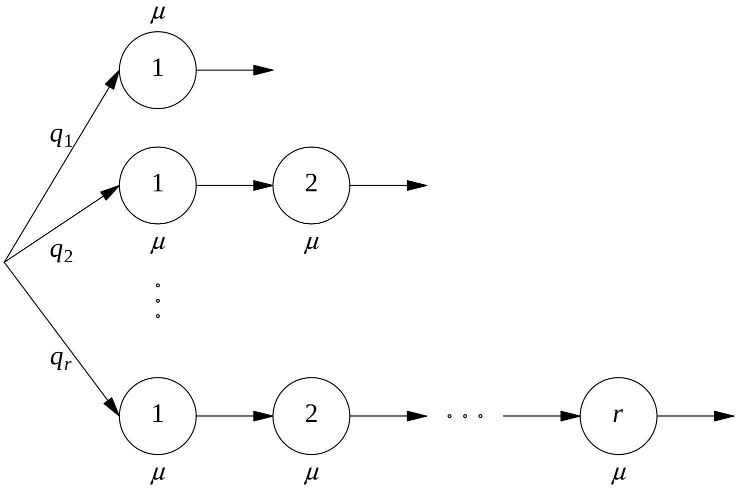

Figure 1: Phase representation of the mixed Erlang service time distribution

The system behavior will be analyzed at arrival instants. The state on arrival instants can be described by the pair (i, j), where i is the number of customers in the system and j the number of remaining service phases of the customer in service just before an arrival. This two-dimensional description leads to a $G/M/1$ type model, as studied in the previous chapter; see [8, 3, 7] for efficient algorithms for the computation of the matrix-geometric solution.

Alternatively the state on arrival instants can be described by the one-dimensional states i where i is the number of uncompleted service phases in the system. In doing so, we loose part of the information, since we cannot determine the number of customers in the system from this description (except when the service times have a pure Erlang distribution). But information on the number of uncompleted service phases is all that is needed to determine the waiting time. The transition probability p_{ij} from state $i \geq 0$ to state $j > 0$ is given by

$$p_{ij} = \begin{cases} q_1 b_{i+1-j} + q_2 b_{i+2-j} + \cdots + q_r b_{i+r-j} & i \geq 0\,, \quad 0 < j \leq i+r, \\ 0 & i \geq 0\,, \quad j > i+r, \end{cases}$$

where b_k is defined as the probability that k service stages are completed during an inter-arrival time, so

$$b_k = \int_0^\infty \frac{(\mu t)^k}{k!} e^{-\mu t} dF_A(t)\,, \qquad k \geq 0\,.$$

Note that the Markov chain on arrival instants is irreducible and aperiodic. Henceforth it will be assumed that the offered load ρ, defined by

$$\rho = \lambda \left(q_1 \cdot \frac{1}{\mu} + q_2 \cdot \frac{2}{\mu} + \cdots + q_r \cdot \frac{r}{\mu} \right),$$

is less than 1. Then the equilibrium probabilities p_i of finding i customers on arrival exist. For $i > 0$ the equilibrium equations are given by

$$p_i = q_1 \sum_{k=0}^\infty p_{i-1+k} b_k + q_2 \sum_{k=0}^\infty p_{i-2+k} b_k + \cdots + q_r \sum_{k=0}^\infty p_{i-r+k} b_k, \tag{1}$$

where by convention

$$p_{-1} = p_{-2} = \cdots = p_{1-r} = 0. \tag{2}$$

Note that we do not pay attention to the equilibrium equation in state 0; since the equilibrium equations are dependent, this one will be satisfied automatically once we have satisfied (1). To solve the equilibrium equations (1) we try to find r basis solutions of the form

$$p_i = \sigma^i, \qquad i = 0, 1, 2, \ldots$$

Substitution of this form into (1) and division by σ^{i-1} yields

$$\sigma^r = \left(q_1 \sigma^{r-1} + q_2 \sigma^{r-2} + \cdots + q_r \right) \sum_{k=0}^\infty b_k \sigma^k,$$

and thus, by substituting the expression for b_k, we find the following equation for σ,

$$\sigma^r = \left(q_1 \sigma^{r-1} + q_2 \sigma^{r-2} + \cdots + q_r \right) E(e^{-\mu(1-\sigma)A}), \tag{3}$$

where the generic random variable has distribution $F_A(\cdot)$. Clearly, only solutions with $|\sigma| < 1$ are useful. By using Rouché's Theorem it can be shown that equation (3) has exactly r roots inside the unit circle (cf. [1]). We assume that these roots are all different and label them $\sigma_1, \sigma_2, \ldots, \sigma_r$. Now we take the linear combination

$$p_i = \sum_{k=1}^{r} c_k (1-\sigma_k)\sigma_k^i.$$

For any choice of the coefficients c_k, this linear combination satisfies (1); it remains to determine the coefficients c_k such that the convention (2) is satisfied. Substitution of this linear combination into (2) yields

$$c_1(1-\sigma_1)\tau_1^i + c_2(1-\sigma_2)\tau_2^i + \cdots + c_r(1-\sigma_r)\tau_r^i = 0, \qquad i = 1, 2, \ldots, r-1,$$

where $\tau_k = 1/\sigma_k$. These equations are of a VanderMonde-type and therefore, they can be solved explicitly using Cramer's rule. Then we get

$$c_k = \frac{C}{\prod_{j=1}^{r}(1-\tau_j)} \frac{\prod_{j\neq k}(1-\tau_j)}{\prod_{j\neq k}(\tau_k-\tau_j)}\,, \qquad k = 1, \ldots, r\,,$$

for some constant C. This constant follows from the normalization equation, which, by using Lagrange's interpolation formula, leads to

$$C = \prod_{j=1}^{r}(1-\tau_j).$$

Our findings are summarized in the following theorem.

Theorem 10.1 *For all $i = 0, 1, 2, \ldots$,*

$$p_i = \sum_{k=1}^{r} c_k (1-\sigma_k)\sigma_k^i,$$

where $\sigma_1, \ldots, \sigma_r$ are the roots with $|\sigma| < 1$ of equation (3) and (with $\tau_j = 1/\sigma_j$)

$$c_k = \frac{\prod_{j\neq k}(1-\tau_j)}{\prod_{j\neq k}(\tau_k-\tau_j)}\,, \qquad k = 1, \ldots, r\,.$$

The arrival probabilities p_i are of the same form as the one for the standard $G/M/1$ queue (i.e., a sum of geometric distributions). Thus the waiting time distribution can also be found along the same lines as for the $G/M/1$, yielding

$$P(W > t) = \sum_{k=1}^{r} c_k \sigma_k e^{-\mu(1-\sigma_k)t}, \qquad t \geq 0. \tag{4}$$

Based on (4) it is easy to find expressions for the moments of the (conditional) waiting time. Hence the problem of finding the waiting time distribution has been reduced to that of finding the roots σ_k of (3).

In the special case of (pure) Erlang-r service time the roots σ_k can be found very efficiently. Then (3) simplifies to

$$\sigma^r = E(e^{-\mu(1-\sigma)A}).$$

The idea is to reduce this equation for r roots to r equations for a single root, by raising both sides of the above equation to the power $1/r$. This leads to

$$\sigma = \phi F(\sigma)\,, \tag{5}$$

where ϕ satisfies $\phi^r = 1$ and

$$F(\sigma) = \sqrt[r]{E(e^{-\mu(1-\sigma)A})}\,.$$

Thus ϕ can be selected from the r unity roots $e^{2\pi i m/r}$, $m = 0, 1, \ldots, r-1$. For each choice of ϕ equation (5) is a fixed point equation. We can try to find the root of (5) with $|\sigma| < 1$ by using the iteration scheme

$$\sigma^{(k+1)} = \phi F(\sigma^{(k)})\,, \qquad k = 0, 1, \ldots$$

starting with $\sigma^{(0)} = 0$. For certain classes of interarrival time distributions it can be shown that, indeed, the sequence $\sigma^{(0)}, \sigma^{(1)}, \ldots$ converges to the desired root; see [1] for more details. These classes include deterministic, shifted exponential, gamma, mixed Erlang and hyper-exponential distributions.

10.2 The $G/G/1$ queue

Let us now consider a single-server queue with generally distributed service times and interarrival times. For this system we present some approximations.

The simplest approximation for the mean waiting time assumes that the randomness of the interarrival times has more or less the same effect on the mean waiting time as the randomness in the service times. Let $E(A)$ and $E(B)$ be the mean interarrival and mean service time, and denote their coefficient of variation by c_A and c_B, respectively. The approximation for the mean waiting time is given by (see, e.g., [6, 11])

$$E(W) = \frac{\rho}{1-\rho} \cdot \frac{c_A^2 + c_B^2}{2} \cdot E(B), \tag{6}$$

where the traffic intensity ρ is assumed to be less than 1,

$$\rho = E(B)/E(A) < 1. \tag{7}$$

Note that the approximation is exact for Poisson arrivals (for which $c_A = 1$). Also, under heavy load conditions (ρ close to 1), the waiting time distribution in the $G/G/1$ system is approximately exponentially distributed with mean given by (6); see, e.g., [4].

Now we present a simple iterative method to approximate the first two moments (and the distribution) of the waiting time. We assume that at time $t = 0$ the zeroth customer arrives at an empty system, having a service time B_0. Define for $n \geq 1$,

$$\begin{aligned} A_n &= \text{the time between the arrival of the } (n-1)\text{th and the } n\text{th customer,} \\ B_n &= \text{the service time of the } n\text{th customer,} \\ W_n &= \text{the waiting time of the } n\text{th customer,} \\ S_n &= \text{the sojourn time of the } n\text{th customer } = W_n + B_n. \end{aligned}$$

Both $\{A_n\}_{n\geq 1}$ and $\{B_n\}_{n\geq 0}$ are sequences of independent identically distributed (i.i.d.) random variables. It is readily verified that for $n \geq 1$,

$$W_n = (S_{n-1} - A_n)^+, \tag{8}$$

where $(x)^+ = \max(x, 0)$. This equation is the starting point for an iterative method to approximate the first two moments of the stationary waiting times (see [5]),

$$E(W) = \lim_{n\to\infty} E(W_n), \qquad E(W^2) = \lim_{n\to\infty} E(W_n^2),$$

where the limits exist by virtue of stability condition (7). Equation (8) relates the waiting time of the nth customer to the sojourn time of the $(n-1)$th customer. From this equation we get the following expression for the kth moment of W_n,

$$E(W_n^k) = \int_0^\infty \int_z^\infty (x-z)^k dF_{S_{n-1}}(x) dF_{A_n}(z). \tag{9}$$

Here we concentrate on the first two moments (so $k = 1, 2$). If the first two moments of the sojourn time of the $(n-1)$th customer are known and we fit a *tractable* distribution to these two moments, then the above expression with the fitted distribution can be used to compute an approximation for the first two moments of the waiting (and sojourn) time of the nth customer. For the two moment fit we may use a mixed Erlang or hyper-exponential distribution, depending on whether the coefficient of variation is less or greater than 1 (see, e.g., [12]). This procedure is then repeated for the next customer and so on. The resulting iteration scheme is presented below.

Iteration scheme

1. Initially set $E(W_0) = E(W_0^2) = 0$ and set $n = 1$;

2. Fit a tractable distribution to the first two moments of S_{n-1}. The fitted distribution $\tilde{F}_{S_{n-1}}(\cdot)$ is a mixture of two Erlang distributions with the same scale parameter if the coefficient of variation is less than 1, and otherwise it is a mixture of two exponential distributions with balanced means.

3. Compute $E(W_n)$ and $E(W_n^2)$ according to (9), with $F_{S_{n-1}}(\cdot)$ replaced by the fitted distribution $\tilde{F}_{S_{n-1}}(\cdot)$.
4. If $|E(W_n)-E(W_{n-1})|$ and $|E(W_n^2)-E(W_{n-1}^2)|$ are sufficiently small, then stop and use $E(W_n)$ and $E(W_n^2)$ as approximation for $E(W)$ and $E(W^2)$; otherwise set $n=n+1$ and go to step 2.

In general, the approximations for $E(W)$ and $E(W^2)$, produced by this algorithm, are excellent; see [5] for more details. Of course, the distribution of the waiting time can be approximated by fitting a distribution on the (approximate) first two moments.

Note that the iteration method can be used to gain insight in the *transient* behavior of the queueing system; e.g., What is the time (in terms of number of arrivals) needed to reach stationarity? Further, the method may also be used in case of *discrete* interarrival and service time distributions (cf. [2]).

References

[1] I.J.B.F. Adan, Y. Zhao, *Analyzing $GI/E_r/1$ queues.* Operations Research Letters, 19 (1996), pp. 183–190.

[2] I.J.B.F. Adan, M.J.A. van Eenige, J.A.C. Resing, *Fitting discrete distributions on the first two moments.* Probability in the Engineering and Informational Sciences, 9 (1995), pp. 623–632

[3] W.K. Grassman, *The $GI/PH/1$ queue: a method to find the transition matrix.* Infor., 20 (1982), pp. 144–156.

[4] J.C. Kingman, *On queues in heavy traffic*, J. Roy. Statist. Soc., Ser. B, 24 (1962), pp. 383–392.

[5] A.G. de Kok, *A Moment-iteration method for approximating the waiting-time characteristics of the GI/G/1 queue.* Probability in the Engineering and Informational Sciences, 3 (1989), pp. 273–287

[6] P.J. Kuehn, *Approximate analysis of general queueing networks by decomposition*, IEEE Trans. Comm., 27 (1979), 113–126.

[7] D.M. Lucantoni, *Efficient algorithms for solving the non-linear matrix equations arising in phase type queues.* Stochastic Models, 1 (1985), pp. 29–51.

[8] M.F. Neuts, *Matrix-geometric solutions in stochastic models.* The John Hopkins University Press, Baltimore, 1981.

[9] R.S. Schassberger, *On the waiting time in the queueing system $GI/G/1$*, Ann. Math. Statist., 41 (1970), pp. 182–187.

[10] R.S. Schassberger, *Warteschlangen,* Springer-Verlag, Berlin, 1973.

[11] J.G. Shanthikumar, J.A. Buzacott, *On the approximations to the single-server queue*, Internat. J. Prod. Res., 18 (1980), pp. 761–773.

[12] H.C. Tijms *Stochastic models: an algorithmic approach.* John Wiley & Sons, Chichester, 1994.