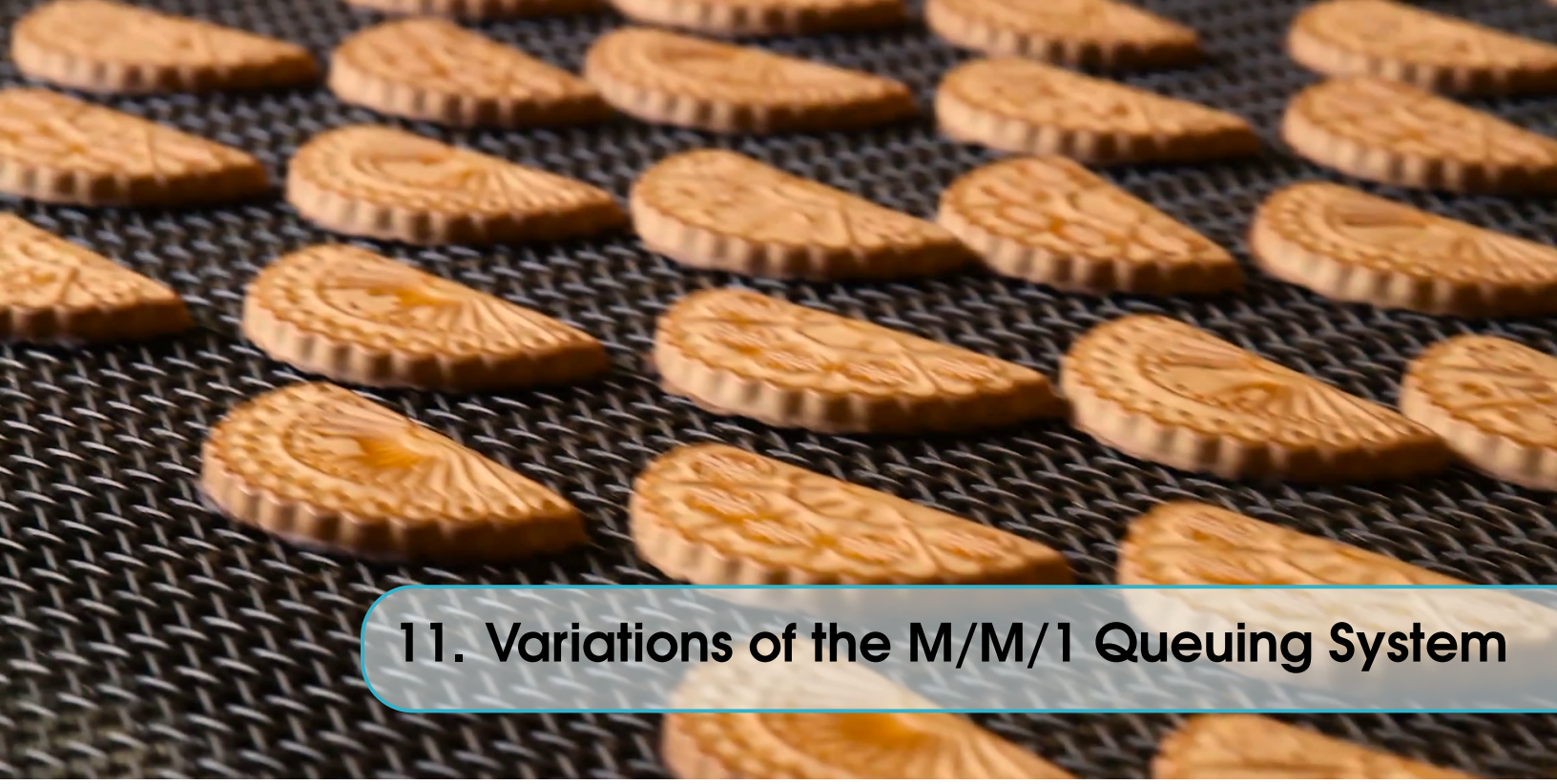

11. Variations of the M/M/1 Queuing System

In the analysis of the M/M/1 system, we have been concerned with exponential service times. Recall that an exponential distribution is memoryless (i.e. the service time for one customer could be thought of as totally independent of the service time of other customers). What if we relax this condition? What if the service time is NOT exponential? In this chapter we will explore a few variations on the M/M/1 model introduced in Chapter 9.

11.1 Constant Service Time – M/D/1

Here we assume that the service time is constant. In other words, all customers require the same amount of service. Examples of constant service time include the transmission time of constant-size cells in ATM networks, or the time it takes to get a single car through a car-wash, etc. In this case, the resulting system goes under the name of **M/D/1**. One can show that under such assumption we get[1]:

$$q = \frac{\rho^2}{2 \cdot (1-\rho)} + \rho \tag{11.1}$$

and:

$$w = q - \rho = \frac{\rho^2}{2 \cdot (1-\rho)} \tag{11.2}$$

At this point using Little's Law (see Equation 5.2) we can derive T_q and T_w as:

$$T_q = \frac{\rho}{2 \cdot \mu \cdot (1-\rho)} + T_s \tag{11.3}$$

[1] To prove the formulas that follow, it is enough to repeat the steps we used to derive the results for M/M/1 considering a constant service time D instead of the average of an exponential distribution.

and:

$$T_w = \frac{\rho}{2 \cdot \mu \cdot (1-\rho)} \tag{11.4}$$

11.2 General Service Time – M/G/1

If we do not know the distribution of service time, but (1) we know the average service time T_s, as well as the service time standard deviation σ_{T_s}, or (2) the normalized value of the service time standard deviation σ_{T_s}/T_s (i.e. the ratio of the service time standard deviation and the service time), then we can use the following formulae that we present without proof. First, we compute the parameter A as follows:

$$A = \frac{1}{2} \cdot \left[1 + \left(\frac{\sigma_{T_s}}{T_s}\right)^2\right] \tag{11.5}$$

Next, we can compute q and w as follows:

$$q = \frac{\rho^2 \cdot A}{1-\rho} + \rho \tag{11.6}$$

and:

$$w = q - \rho = \frac{\rho^2 \cdot A}{1-\rho} \tag{11.7}$$

Once again using Little's Law (see Equation 5.2) we can derive T_q and T_w as:

$$T_q = \frac{1}{\mu}\left(\frac{\rho \cdot A}{1-\rho} + 1\right) \tag{11.8}$$

and:

$$T_w = \frac{1}{\mu} \cdot \frac{\rho \cdot A}{(1-\rho)} \tag{11.9}$$

Let us do a few sanity-check considerations.

1. We know that for an exponential distribution, the standard is equal to the mean. It follows that the ratio $\sigma_{T_s}/T_s = 1$. As such, the value of A from Equation 11.5 turns out to be: $A = 1$. Hence we have:

$$q = \frac{\rho^2 \cdot A}{1-\rho} + \rho = \frac{\rho^2}{1-\rho} + \rho = \frac{\rho}{1-\rho} \tag{11.10}$$

Unsurprisingly, this is the expression for q when we consider a standard M/M/1 system.

2. Similarly, if the service time is constant, then its standard deviation is $\sigma_{T_s} = 0$. It follows that $A = 1/2$. As such we can write:

$$q = \frac{\rho^2 \cdot A}{1-\rho} + \rho = \frac{\rho^2}{2(1-\rho)} + \rho \tag{11.11}$$

Once again, unsurprisingly, this is the expression for q in case of M/D/1 system, as described in Equation 11.1.

Generally speaking, the normalized value of the standard deviation of the service time gives us a good indication of the expected performance of the system. The plot below shows the number of customers in an M/G/1 system as we change this quantity. Notice that if we assume an exponential distribution instead of a distribution for which the aforementioned ratio is less than 1, then we will be "over estimating" the load in the system. Thus, the exponential distribution is a safe assumption to make about service time if we know that the above ratio is less than 1.

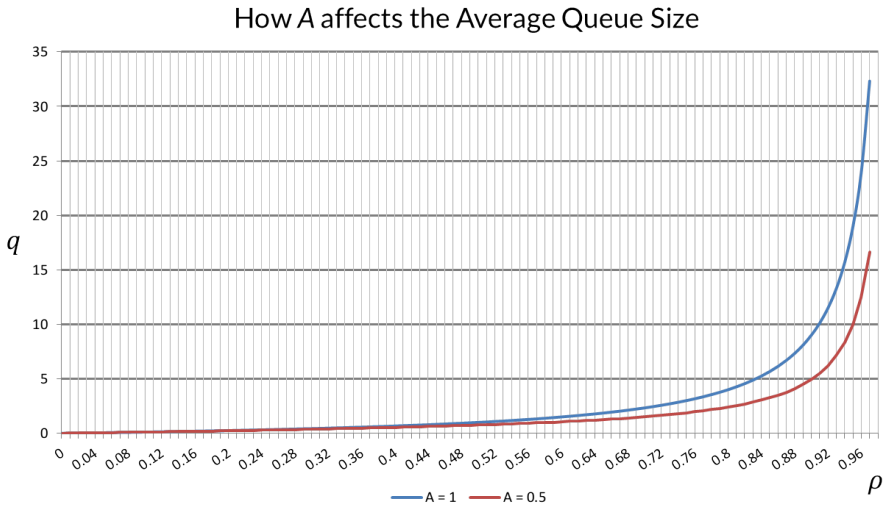

Figure 11.1: Variation in the trend of average queue size q with two different values of A, i.e. $A = 1$ (M/M/1) and $A = 0.5$ (M/D/1).

Notice that in the equation for the expression A the normalized value of the standard deviation (i.e. σ_{T_s}/T_s) is an indication of the *unpredictability* of (or variability in) the service time. The larger this value, the more likely it is for the system's performance to deteriorate faster as the system utilization gets closer to 1. This means that, to get better performance, it is desirable to design systems with **predictable" service time** (i.e. less variability).

11.3 Finite Queue Systems – M/M/1/K

In all the analysis we have done so far, we assumed that the queue size was infinite, in the sense that we were always guaranteed that an incoming request would find a place to be queued. Obviously, this is not a realistic assumption. Now assume that the buffer size (i.e. number of place holders that

one can use for requests waiting for or currently receiving service) is finite. Specifically, assume that there are only K such spaces. It is possible to analyze a system with Poisson Arrivals and Exponential service times with a finite queue. For the purposes of this class, we will not derive these relationships, but before presenting them, let us first discuss some of the consequences of having limited queuing space.

In the M/M/1 system, every request was guaranteed to be serviced eventually. We ensured this by insisting that the utilization of the system be strictly less than 1 — i.e., the arrival rate had to be less than the maximum departure (or service) rate. This assured us that waiting time (in the queue) would be finite. Now, if the buffer size is limited, then we cannot give a guarantee that every service request will be serviced simply because there is a possibility that when a request arrives to the system, there will not be any place to "queue" that request. The result is that the request must be rejected.

The possibility that requests could be "rejected" allows us (in fact) to relax the assumption we made in M/M/1 systems — namely, that the system utilization be strictly less than 1. Therefore, in M/M/1/K systems, it is possible for the arrival rate to be **more** than the maximum departure (or service) rate.

Just like we did for M/M/1 queues, let's first review the probability of being in a given state S_j, namely $P(S_j)$ i.e. the probability of having $j \in \{0,\ldots,K\}$ requests in the system. Clearly, if $j > 0$, then $j-1$ are the requests waiting to receive service. But it can never happen that $j > K$ because the cap on the total number of requests in the system is K. $P(S_j)$ can be computed as follows.

$$P(S_j) = \begin{cases} \frac{(1-\rho)\rho^j}{1-\rho^{K+1}} & \text{if } \rho \neq 1 \\ \\ \frac{1}{K+1} & \text{if } \rho = 1 \end{cases} \tag{11.12}$$

Note that in the above equation there is an expression both for when utilization is exactly equal to 1, as well as when it is either larger or smaller than 1. The same will be true for the relations that follow. We are now ready to review the formula for the total average number of requests in the system q.

$$q = \begin{cases} \frac{\rho}{1-\rho} - \frac{(K+1)\cdot\rho^{K+1}}{(1-\rho^{K+1})} & \text{if } \rho \neq 1 \\ \\ \frac{K}{2} & \text{if } \rho = 1 \end{cases} \tag{11.13}$$

Here, ρ is defined as always, i.e. as $\rho = \lambda \cdot T_s$.

As we mentioned before, in an M/M/1/K system, there is a possibility that a request will be "rejected" (i.e. no space to queue it). This probability follows directly from Equation 11.12 when $j = K$ and is given by:

$$P(\text{"rejection"}) = P(S_K) = \begin{cases} \frac{(1-\rho)\cdot\rho^K}{(1-\rho^{K+1})} & \text{if } \rho \neq 1 \\ \\ \frac{1}{K+1} & \text{if } \rho = 1 \end{cases} \tag{11.14}$$

As we have done repeatedly before, once we have obtained a value for the total number of requests in a system (q), we can use Little's law to find out the value of the response time T_q for that system. Recall that Little's law states that the number of requests in the system is the product of the response time and the rate of arrivals (see Equation 5.2).

For an M/M/1/K system we can compute q (as shown above). The question is **what is the rate of arrivals for an M/M/1/K system?** Is it simply λ? Not really! Remember that a percentage of the requests will be *rejected*, so in effect the rate of requests that will experience the response time must exclude the **rate of rejection**. This gives us a way to calculate the **effective rate** with which requests are processed. Namely:

$$\lambda' = \lambda \cdot (1 - P(S_K)) \tag{11.15}$$

Now using this effective rate, we can use Little's law to find the response time for requests that were not rejected. Namely:

$$T_q = \frac{q}{\lambda'} \tag{11.16}$$