9. Elementary Queuing Analysis

In this chapter we will introduce the first analytic tools to reason about the performance of a system without simulating its step-by-step activity.

9.1 Notation

Let us first recall the notation introduced in the previous chapters. In discussing various server queues, it will be necessary to talk about various "random variables" associated with these queues. The Figure 9.1 below summarizes the notation used to describe these variables. Notice that often times, we will be interested in the expected values of random variables, thus unless specified otherwise, when we use these variables to denote the expected value of the random variable in question.

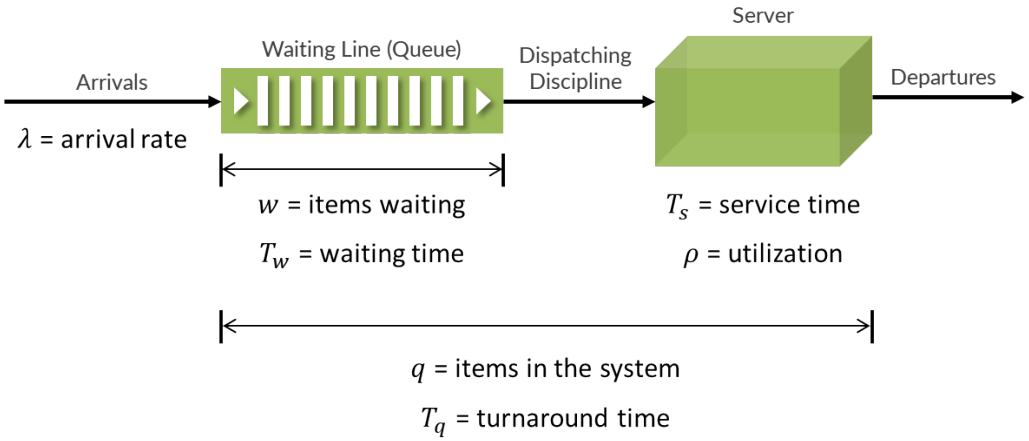

Figure 9.1: A single queue / single-server system with the corresponding notation.

9.2 Analysis of an M/M/1 Queuing System

An M/M/1 queuing system is a single-queue single-server queuing system in which arrivals are Poisson and service time is exponential. The notation M/M/1 describes the "queue" in the system as having a **M**arkovian arrival process (i.e., Poisson) and a **M**arkovian (i.e., exponential) service discipline with **1** server. Hence the name. This is visually summarized in Figure 9.2

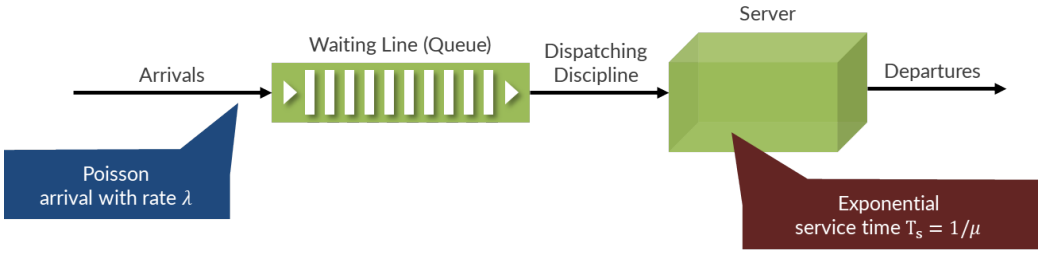

Figure 9.2: A single queue with Poisson arrivals and Exponential Service time is an M/M/1 system.

9.2.1 Birth and death probabilities for M/M/1

Consider a very small interval of time of length h. Assume that this interval of time (h) is so small that a maximum of one arrival can realistically occur in that period of time. Since the rate of arrival is λ requests per unit time, then it follows that the *rate* of arrival per interval h is λh. For instance, if λ = 100 requests/second, and h = 1 msec (i.e., 0.001 seconds), then the rate of arrivals per millisecond is $\lambda \cdot h = 100 \cdot 0.001 = 0.1$ requests/milliseconds.

During an interval h one of two things can happen: either no requests arrive during that small interval of time, or one request does arrive. We call the arrival of a request to the system a **birth** event.

We know that the probability density of the Poisson distribution is:

$$f(x) = \frac{\lambda^x}{x!} \cdot e^{-\lambda} \tag{9.1}$$

where $x = 0, 1, 2, \ldots$

Given that the rate of arrival per interval h is λh, the probability of x arrivals per interval h is:

$$f(x) = \frac{(\lambda h)^x}{x!} \cdot e^{-\lambda h} \tag{9.2}$$

where $x = 0, 1, 2, \ldots$

According to the above equation, the probability that there will be no arrivals during a given interval h is $f(0) = e^{-\lambda h}$.

Now, since we assume h to be too small for anything more than one arrival (or births), we can say that the probability of more than one arrival is negligible. Thus, the probability of one or more

arrivals in the interval h is a good approximation for the probability of exactly one arrival[1].

From the above we can say that the probability of a birth is given by:

$$P(\text{birth}) = 1 - P(\text{no arrival}) = 1 - e^{-\lambda h} \tag{9.3}$$

From the above expression, we can expand the exponential function as a Taylor series, i.e. we can write:

$$P(\text{birth}) = 1 - e^{-\lambda h} = 1 - (1 - \lambda h + \frac{(\lambda h)^2}{2!} - \frac{(\lambda h)^3}{3!} + \ldots) \tag{9.4}$$

For very small period h, we can use the first order approximation, which results in the following probability:

$$P(\text{birth}) = \lambda h \tag{9.5}$$

By reasoning in a very similar way, we can show that the probability that a customer will leave the system (i.e. a customer for whom service was finished) given that somebody is in the system in the first place is μh, where recall that $\mu = 1/T_s$. We call such an event a **death** event.

$$P(\text{death}) = \mu h \tag{9.6}$$

9.2.2 State (rate) transition diagram for M/M/1

Consider an M/M/1 system at steady state (i.e. equilibrium). Such a system will have a variable number of customers. In particular, at any point of time, a customer may be added to the system through a birth event, or a customer may be removed from the system due to a death event.

Consider the state of the system when exactly j customers are in the system. We denote such a state by S_j. In order to compute many important properties of an M/M/1 queuing system, it will be necessary to calculate the probability that at steady state, the system is in a state S_j (for any $j = 0, 1, 2, \ldots$).

Consider the system at a given instant. Let the state of the system at that instant be S_j. How could the system be in such a state? Well, to answer this question, consider the system at an interval h earlier (where h is very small). There are three scenarios that would result in the system moving into state S_j (see Figure 9.3):

1. The system was in state S_{j-1} and a *birth* occurred. The probability of that happening is λh.
2. The system was in state S_{j+1} and a *death* occurred. The probability of that happening is μh.
3. The system was in state S_j and, neither a birth nor a death occurred. The probability of that happening is $1 - \lambda h - \mu h = 1 - h(\lambda + \mu)$.

Figure 9.3 below shows the above transitions. Solid arrows denote the transitions that result from entering into state S_j.

[1] As a matter of fact, these two quantities will be equal as $h \to 0$.

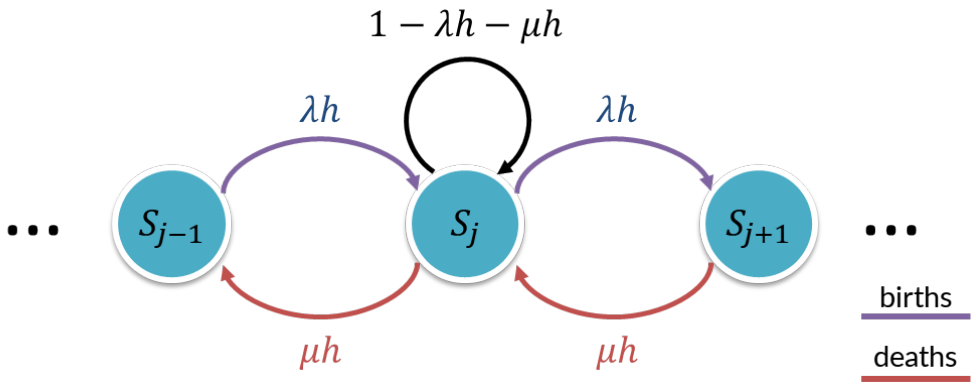

Figure 9.3: State transition diagram for an M/M/1 system, focusing on the generic state S_j.

From the above, we have the following relationship:

$$P(S_j) = \lambda h \cdot P(S_{j-1}) + \mu h \cdot P(S_{j+1}) + (1 - \lambda h - \mu h) \cdot P(S_j) \tag{9.7}$$

By rearranging the various terms, we get:

$$\mu \cdot P(S_{j+1}) = (\lambda + \mu) \cdot P(S_j) - \lambda \cdot P(S_{j-1}) \tag{9.8}$$

And by diving both sides by μ we get:

$$P(S_{j+1}) = (1 + \rho) \cdot P(S_j) - \rho \cdot P(S_{j-1}) \tag{9.9}$$

Note that S_0 is obviously a special case – since there is no S_{j-1} as it is impossible to have a negative number of customers in the M/M/1 system – as illustrated in Figure 9.4.

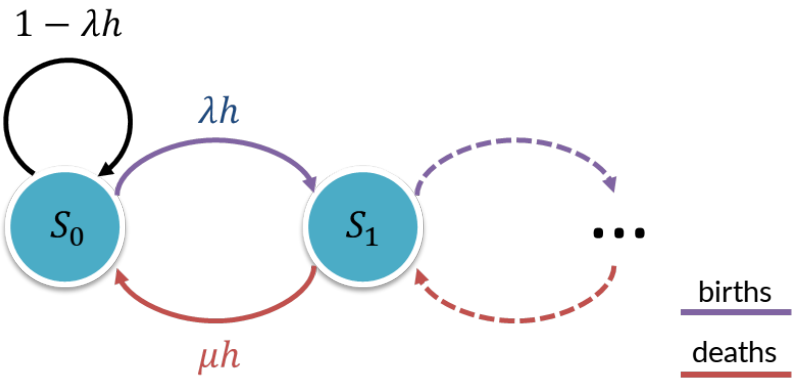

Figure 9.4: State transition for M/M/1 when number of customers in the system is 0, i.e. the system is in S_0.

As such, we can write:

$$P(S_0) = \mu \cdot P(S_1) + (1 - \lambda) \cdot P(S_0) \tag{9.10}$$

And by re-applying the same steps as before, we can write:

$$P(S_1) = \rho \cdot P(S_0) \tag{9.11}$$

Now we can obtain $P(S_2)$ by unfolding Equation 9.11 into Equation 9.9 with $j = 1$. We have:

$$P(S_2) = (1+\rho) \cdot P(S_1) - \rho \cdot P(S_0) = (1+\rho) \cdot (\rho \cdot P(S_0)) - \rho \cdot P(S_0) = \rho^2 \cdot P(S_0) \tag{9.12}$$

And we can now compute $P(S_3)$ with the same approach:

$$P(S_3) = (1+\rho) \cdot P(S_2) - \rho \cdot P(S_1) = (1+\rho) \cdot \rho^2 \cdot P(S_0) - \rho^2 \cdot P(S_0) = \rho^3 \cdot P(S_0) \tag{9.13}$$

It is easy to see how this formula can be generalized for a generic state S_j, as:

$$P(S_j) = \rho^j \cdot P(S_0) \tag{9.14}$$

Moreover, since the overall probability density must add up to 1, we get:

$$\sum_{i=0}^{\infty} P(S_i) = \sum_{i=0}^{\infty} \rho^i \cdot P(S_0) = 1 \tag{9.15}$$

For the former, we can find the value of $P(S_0)$ by writing:

$$P(S_0) = \frac{1}{\sum_{i=0}^{\infty} \rho^i} \tag{9.16}$$

Note that the series $\sum_{i=0}^{\infty} \rho^i$ is a geometric series, and since $\rho < 1$ it always converges to $1/(1-\rho)$. With this, we can re-write the expression in Equation 9.16 as:

$$P(S_0) = 1 - \rho \tag{9.17}$$

And finally we have that the probability of having j customers in the system is given by:

$$P(S_j) = \rho^j \cdot (1-\rho) \tag{9.18}$$

The above relationship gives us the PMF for the number of customers in the M/M/1 system, which is the same as a **geometric distribution with parameter** ρ. To get a feel for the above relationship, let us look at the probability of having j customers in the system (i.e. $P(S_j)$) as a function of utilization. This is shown in Figure 9.5 for j=0, 1, 2, 4, 8, and 16.

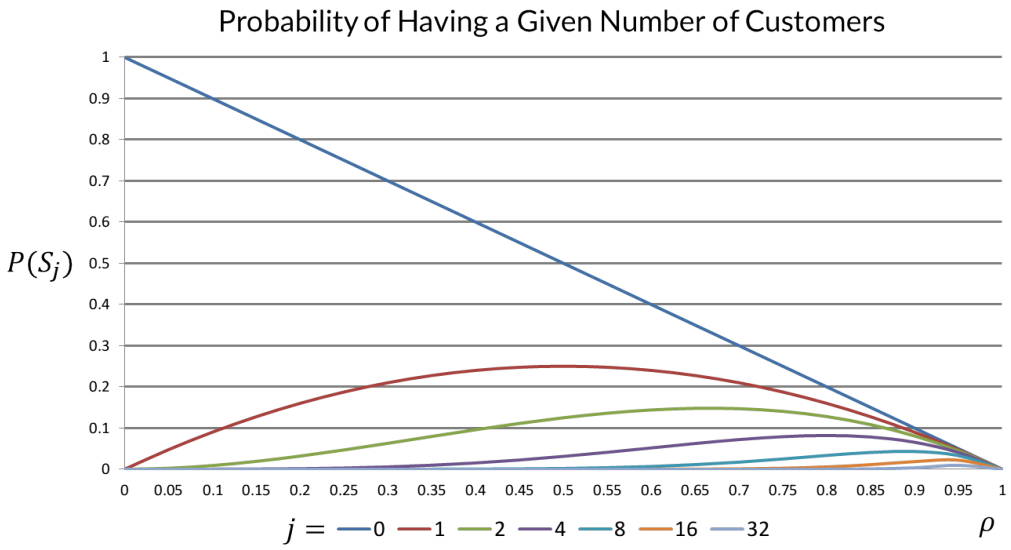

Figure 9.5: Probability (y-axis) of having j customers (different lines) in the system as a function of utilization (x-axis).

9.2.3 Average number of customer in an M/M/1 System

By using what derived in Equation 9.18, we can now compute the average number of customers q in an M/M/1 system. This can be computed as follows:

$$q = E[\text{number of customers}] = \sum_{i=0}^{\infty} i \cdot P(S_i) = \sum_{i=0}^{\infty} i \cdot \rho^i \cdot (1-\rho) \tag{9.19}$$

It can be shown that the expression above converges to a final value (see Box 9.2.1 for more details). The resulting result follows:

$$q = \frac{\rho}{1-\rho} \tag{9.20}$$

Notice that the above result also follows directly from the mean of the geometric distribution, with parameter ρ. To appreciate the above relationship, let us plot the number of pending requests (i.e. q) as a function of the utilization of the system ρ. This is shown in Figure 9.6.

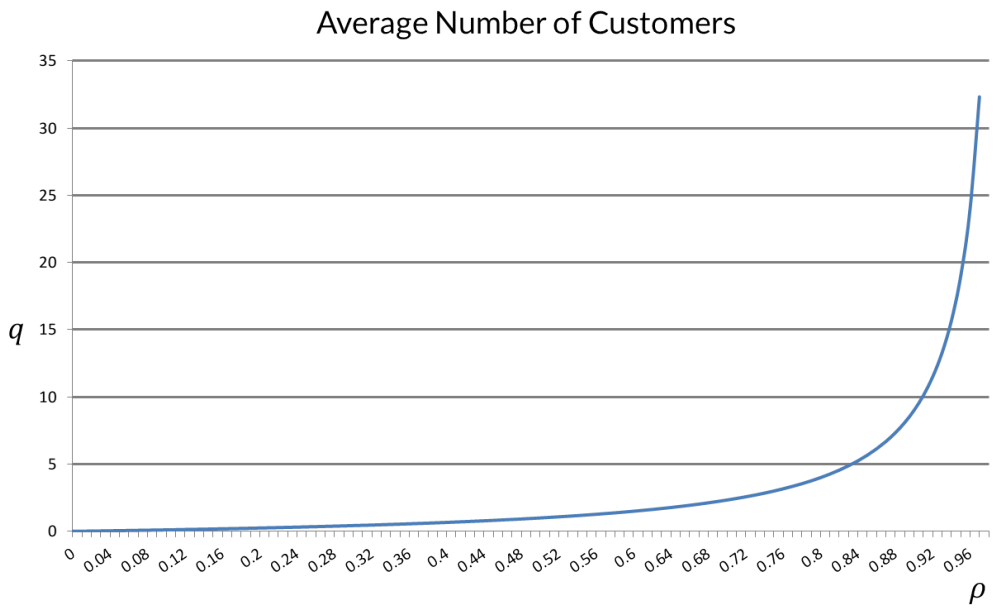

Figure 9.6: Average number of customers in the system q as function of the system utilization ρ.

Box 9.2.1 Expected Value of the Number of Customers

We have already determined that the value of q is given by the following expression (see Equation 9.19):

$$q = \sum_{i=0}^{\infty} i \cdot \rho^i \cdot (1-\rho) = (1-\rho) \sum_{i=0}^{\infty} i \cdot \rho^i \tag{9.21}$$

In order to simplify this expression, we first introduce a class of functions, namely *polylogarithms*[a]. In short, a polylogarithm is a function of order $s \in \mathbb{C}$ and argument $z \in \mathbb{C}$ indicated with the shorthand notation $\mathrm{Li}_s(z)$. The generic polylogarithm is defined as:

$$\mathrm{Li}_s(z) = \sum_{i=1}^{\infty} \frac{z^i}{i^s} = z + \frac{z^2}{2^s} + \frac{z^3}{3^s} + \ldots \tag{9.22}$$

Looking at Equation 9.21, note that we can rewrite the right-hand side of the equality as polylogarithm of order $s = -1$ and argument $z = \rho$, multiplied by a constant. In fact, we have:

$$q = (1-\rho) \sum_{i=0}^{\infty} \frac{\rho^i}{i^{-1}} = (1-\rho)\mathrm{Li}_{-1}(\rho) \tag{9.23}$$

Since $\rho < 1$, the value of the expression $\mathrm{Li}_{-1}(\rho)$ is given by:

$$\mathrm{Li}_{-1}(\rho) = \frac{\rho}{(1-\rho)^2} \tag{9.24}$$

Thus it follows that:

$$q = (1-\rho)\mathrm{Li}_{-1}(\rho) = \frac{\rho}{1-\rho} \tag{9.25}$$

[a] They take this name because $\mathrm{Li}_1(z) = -\ln(1-z)$.

9.2.4 Average number of customers waiting for service in a M/M/1 system

Now that we have an expression for q, we can use that to compute the average for all the other metrics of interest. If we want to compute w, we can proceed as follows. First, recall that $q = w + \rho$ (see Equation 5.4). Next we can write:

$$w = q - \rho = \frac{\rho^2}{1-\rho} \tag{9.26}$$

9.2.5 Average Time in an M/M/1 System

In order to compute the average time spent in total in the M/M/1 system, i.e. T_q, we can use Little's Law (see Equation 5.2):

$$T_q = \frac{q}{\lambda} = \frac{1}{\mu \cdot (1-\rho)} \tag{9.27}$$

9.2.6 Average Time Waiting in an M/M/1 System

Once again in order to compute the average time waiting in the M/M/1 system, i.e. T_w, we can use Little's Law (see Equation 5.2):

$$T_w = \frac{w}{\lambda} = \frac{\rho}{\mu \cdot (1-\rho)} \tag{9.28}$$

9.2.7 Slowdown

Slowdown is defined as the ratio between the typical (or average) length of time it takes to get serviced and the absolute minimum amount of time in which the service can be rendered. For an M/M/1 system, the average time it takes to get serviced is the response time (queuing time + service time). As we have seen above, due to queuing, the response time (T_q) for a service is larger than the service time (T_s). Thus, the "slowdown" due to queuing is a good indicator as to how much degradation in service a customer is experiencing due to queuing (compared to what it would have

taken had the customer been the only customer). Thus, we call the ratio T_q/T_s the slowdown of the M/M/1 system, which is given by:

$$\text{Slowdown} = \frac{T_q}{T_s} = \frac{\frac{1}{\mu\cdot(1-\rho)}}{\frac{1}{\mu}} = \frac{1}{1-\rho} \tag{9.29}$$

9.2.8 Variability Measures

The above measures average-case behaviors. As we explained earlier, it would be interesting to estimate the variability around these averages. Using similar derivations, one can derive the standard deviations of the above metrics. In particular, we can calculate the standard deviation in the number of customers in an M/M/1 system (i.e. q), as well as the total time in the system (i.e. T_q). The result is provided in the following two equations:

$$\sigma_q = \frac{\sqrt{\rho}}{1-\rho} \tag{9.30}$$

$$\sigma_{T_q} = \frac{1}{\mu\cdot(1-\rho)} = \frac{T_s}{1-\rho} \tag{9.31}$$

9.3 An M/M/1 Example

Consider a system, where the following holds:

1. Hits to a single-process, single-threaded web server follow a Poisson arrival process with mean **10 requests per second**. The service time (per request) follows an exponential distribution with **mean 9 msec**. Calculate the response time of this web server.
2. Due to a super bowl ad, hits to the web server went up 10-fold. Calculate the response time under such a condition.
3. Compute the relative slowdown of the system before and after the super bowl ad.
4. How much faster should the web server be so that the slowdown after the super bowl ad would be comparable to that before the ad?

Solution: To answer the first part, notice that we are considering a single-threaded server. This means that the server can only serve one request at a time. Also, the arrival process of the traffic is Poisson, and the service time is exponentially distributed. Hence, it maps directly to an M/M/1 queue. The arrival rate λ is 10 (requests per second) and the mean service time T_s is 9 msec = 0.009 seconds (always make sure the units for the rate and times are consistent!) This gives us the following for the utilization ρ of the server:

$$\rho = \lambda \cdot T_s = 0.09 = 9\%$$

Applying what we know about the average number of customers q in an M/M/1 system, we get:

$$q = \frac{\rho}{1-\rho} = \frac{0.09}{1-0.09} = 0.099$$

Using Little's law (see Equation 5.2), we can figure out the response time (in seconds), which would be:

$$T_q = \frac{q}{\lambda} = \frac{0.099}{10} = 0.0099$$

To answer the second part, we consider the new arrival rate $\lambda' = 100$ requests per second. We get:

$$\rho' = \lambda' \cdot T_s = 0.9 = 90\%$$

$$q' = \frac{\rho'}{1-\rho'} = \frac{0.9}{1-0.9} = 9$$

$$T'_q = \frac{q'}{\lambda'} = \frac{9}{100} = 0.09$$

For the third part, the slowdown is simply the ratio of the response time to the service time. So, before the super-bowl ad, the slowdown is 1.09 – i.e., the response time was 9% higher than service time. But afterwards, it is 10.0 – i.e., the response time is 900% higher than service time! That is a whopping 100-fold increase in slowdown, when the increase in arrival rate was only 10-fold higher!

Finally, for the last part, it follows that to bring the slowdown back to what it was before, we need to bring the service rate to what it was before. In other words, we can do this by cutting down the average service time by 10-fold – i.e., from 9 msec to 0.9 msec. Or as we will see later, add 9 more servers to work in parallel, each of which is processing requests at the old service time.