4. Appendix

Page 17, **Example 4.3.**
Since $\lambda = \frac{1}{12}, \mu = \frac{1}{8}$, we have

$$L = 2, W = 24$$

Hence, the average number of customers in the system is 2, and the average time a customer spends in the system is 24 minutes. Now suppose that the arrival rate increases 20 percent to $\lambda = \frac{1}{10}$. What is the corresponding change in L and W? Similar as before, it gives

$$L = 4, W = 40$$

Hence, an increase of 20 percent in the arrival rate doubled the average number of customers in the system. From these equations we can see that when λ/μ is near 1, a slight increase in λ/μ will lead to a large increase in L and W.

Page 18, **Example 4.4**
To do so, let

$$f(\alpha) = \frac{\alpha}{\mu - \lambda\alpha}$$

and note that $W(\alpha) = f(\alpha) + f(1-\alpha)$.
Differentiation yields that

$$f'(\alpha) = \frac{\mu - \lambda\alpha + \lambda\alpha}{(\mu - \lambda\alpha)^2} = \mu(\mu - \lambda\alpha)^{-2}$$

and $f''(\alpha) = 2\lambda\mu(\mu - \lambda\alpha) - 3$
Because $\mu > \lambda\alpha$, we see that $f''(\alpha) > 0$. Similarly, because $\mu > \lambda(1-\alpha)$, we have that $f''(1-\alpha) > 0$. Hence,

$$W''(\alpha) = f''(\alpha) + f''(1-\alpha) > 0$$

Equating $W'(\alpha) = f'(\alpha) - f'(1-\alpha)$ to 0 yields the solution $\alpha = 1 - \alpha$, or $\alpha = 1/2$. Hence, $W(\alpha)$ is minimized when $\alpha = 1/2$, with minimal value

$$\min_{0 \le \alpha \le 1} W(\alpha) = W(1/2) = \frac{1}{\mu - \lambda/2}.$$

Page 18, **Remark 1**

We have used the fact that if one event occurs at an exponential rate λ, and another independent event at an exponential rate μ, then together they occur at an exponential rate $\lambda+\mu$. To check this formally, let T_1 be the time at which the first event occurs, and T_2 the time at which the second event occurs. Then

$$P\{T_1 \leq t\} = 1 - e^{-\lambda t},\ P\{T_2 \leq t\} = 1 - e^{-\mu t}$$

Now if we are interested in the time until either T_1 or T_2 occurs, then we are interested in $T = \min(T_1, T_2)$. Now,

$$P\{T \leq t\} = 1 - P\{T > t\} = 1 - P\{\min(T_1, T_2) > t\}$$

However, $\min(T_1, T_2) > t$ if and only if both T_1 and T_2 are greater than t; hence,

$$\begin{aligned} P\{T \leq t\} &= 1 - P\{T_1 > t, T_2 > t\} \\ &= 1 - P\{T_1 > t\}P\{T_2 > t\} = 1 - e^{-\lambda t}e^{-\mu t} = 1 - e^{-(\lambda+\mu)t} \end{aligned}$$

Thus, T has an exponential distribution with rate $\lambda + \mu$, and we are justified in adding the rates.

Page 18, **Remark 2**
Now ,

$$P\{N = n|W^* = t\} = \frac{f_{N,W^*}(n,t)}{f_{W^*}(t)} = \frac{P\{N = n\}f_{W^*|N}(t|n)}{f_{W^*}(t)}$$

where $f_{W^*|N}(t|n)$ is the conditional density of W^* given that $N = n$, and $f_{W^*}(t)$ is the unconditional density of W^*. Now, given that $N = n$, the time that the customer spends in the system is distributed as the sum of n+1 independent exponential random variables with a common rate μ, implying that the conditional distribution of W^* given that $N = n$ is the gamma distribution with parameters $n + 1$ and μ. Therefore, with $C = 1/f_{W^*}(t)$,

$$\begin{aligned} P\{N = n|W^* = t\} &= CP(N = n)\mu e^{-\mu t}\frac{(\mu t)^n}{n!} \\ &= C(\lambda/\mu)^n(1 - \lambda/\mu)\mu e^{-\mu t}\frac{(\mu t)^n}{n!} = K\frac{(\mu t)^n}{n!} \end{aligned}$$

where $K = C(1 - \lambda/\mu)\mu e^{-\mu t}$ does note depend on n.

Page 18, **Remark 2** Continuous
Summing over n yields

$$1 = \sum_{n=0}^{\infty} P(N = n|T = t) = K \sum_{n=0}^{\infty} \frac{(\mu t)^n}{n!} = Ke^{\lambda t}$$

Thus, $K = e^{-\lambda t}$, showing that

$$P\{N = n|W^* = t\} = e^{-\lambda t}\frac{(\lambda t)^n}{n!}$$

Therefore, the conditional distribution of the number seen by an arrival who spends a total of t time units in the system is the Poisson distribution with mean λt.

In addition, as a by-product of our analysis, we have

$$f_{W^*}(t) = 1/C = \frac{1}{K}(1 - \lambda/\mu)\mu e^{-\mu t} = (\mu - \lambda)e^{-(\mu-\lambda)t}$$

In other words, W^*, the amount of time a customer spends in the system, is an exponential random variable with rate $\mu - \lambda$. (As a check, we note that $E[W^*] = 1/(\mu - \lambda)$.)

Page 18, **Remark 3**

If we let N denote the number of customers in the system as seen by an arrival, then this arrival will spend $N+1$ service times in the system before departing. Now,

$$P\{N+1=j\} = P\{N=j-1\} = (\lambda/\mu)^{j-1}(1-\lambda/\mu),\ j \geq 1$$

In words, the number of services that have to be completed before the arrival departs is a geometric random variable with parameter $1-\lambda/\mu$. Therefore, after each service completion our customer will be the one departing with probability $1-\lambda/\mu$. Thus, no matter how long the customer has already spent in the system, the probability he will depart in the next h time units is $\mu h + o(h)$, the probability that a service ends in that time, multiplied by $1-\lambda/\mu$. That is, the customer will depart in the next h time units with probability $(\mu-\lambda)h + o(h)$, which says that the hazard rate function of W^* is the constant $\mu-\lambda$. But only the exponential has a constant hazard rate, and so we can conclude that W^* is exponential with rate $\mu-\lambda$.

Page 19, **Example 4.5**

Although it might initially seem, by the PASTA principle, that this probability should just be $(\lambda/\mu)^n(1-\lambda/\mu)$, we must be careful. Because if t is the current time, then the time from t until the next arrival is exponentially distributed with rate λ, and is independent of the time from t since the last arrival, which (in the limit, as t goes to infinity) is also exponential with rate λ. Thus, although the times between successive arrivals of a Poisson process are exponential with rate λ, the time between the previous arrival before t and the first arrival after t is distributed as the sum of two independent exponentials. (This is an illustration of the inspection paradox, which results because the length of an interarrival interval that contains a specified time tends to be longer than an ordinary interarrival interval)

Let N_a denote the number found by the next arrival, and let X be the number currently in the system. Conditioning on X yields

$$\begin{aligned} P\{N_a = n\} &= \sum_{k=0}^{\infty} P(N_a = n|X = k)P(X = k) \\ &= \sum_{k=0}^{\infty} P(N_a = n|X = k)(\lambda/\mu)^k(1-k/\mu) \\ &= \sum_{k=n}^{\infty} P(N_a = n|X = k)(\lambda/\mu)^k(1-\lambda/\mu) \\ &= \sum_{i=0}^{\infty} P(N_a = n|X = n+i)(\lambda/\mu)^{n+i}(1-\lambda/\mu) \end{aligned}$$

Page 19, **Example 4.5** Continuous

Now, for $n > 0$, given there are currently $n + i$ in the system, the next arrival will find n if we have i services before an arrival and then an arrival before the next service completion. By the lack of memory property of exponential interarrival random variables, this gives

$$P(N_a = n | X = n + i) = \left(\frac{\mu}{\lambda + \mu}\right)^i \frac{\lambda}{\lambda + \mu}, n > 0$$

Consequently, for $n > 0$

$$\begin{aligned} P(N_a = n) &= \sum_{i=0}^{\infty} \left(\frac{\mu}{\lambda + \mu}\right)^i \frac{\lambda}{\lambda + \mu} \left(\frac{\lambda}{\mu}\right)^{n+i} (1 - \lambda/\mu) \\ &= (\lambda/\mu)^n (1 - \lambda/\mu) \frac{\lambda}{\lambda + \mu} \sum_{i=0}^{\infty} \left(\frac{\lambda}{\lambda + \mu}\right)^i \\ &= (\lambda/\mu)^{n+1} (1 - \lambda/\mu) \end{aligned}$$

On the other hand, the probability that the next arrival will find the system empty when there are currently i in the system, is the probability that there are i services before the next arrival. Therefore, $P\{N_a = 0 | X = i\} = (\frac{\mu}{\lambda+\mu})^i$, giving

$$\begin{aligned} P(N_a = 0) &= \sum_{i=0}^{\infty} \left(\frac{\mu}{\lambda + \mu}\right)^i \left(\frac{\lambda}{\mu}\right)^i (1 - \lambda/\mu) \\ &= (1 - \lambda/\mu) \sum_{i=0}^{\infty} \left(\frac{\lambda}{\lambda + \mu}\right)^i = (1 + \lambda/\mu)(1 - \lambda/\mu) \end{aligned}$$

Page 19, **Example 4.5** Continuous

As a check, note that

$$\sum_{i=0}^{\infty} P\{N_a = n\} = (1-\lambda/\mu)\left[1 + \lambda/\mu + \sum_{n=1}^{\infty}(\lambda/\mu)^{n+1}\right] = (1-\lambda/\mu)\sum_{i=0}^{\infty}(\lambda/\mu)^{i} = 1$$

Note that $P\{N_a = 0\}$ is larger than $P_0 = 1 - \lambda/\mu$, showing that the next arrival is more likely to find an empty system than is an average arrival, and thus illustrating the inspection paradox that when the next customer arrives the elapsed time since the previous arrival is distributed as the sum of two independent exponentials with rate λ. Also, we might expect because of the inspection paradox that $\mathrm{E}[N_a]$ is less than L, the average number of customers seen by an arrival. That this is indeed the case is seen from

$$\mathrm{E}[N_a] = \sum_{n=1}^{\infty} n(\lambda/\mu)^{n+1}(1 - \lambda/\mu) = \frac{\lambda}{\mu}L < L$$

Page 23, **Example 4.6** To solve this, suppose that we use rate μ. Let us determine the amount of money coming in per hour and subtract from this the amount going out each hour. This will give us our profit per hour, and we can choose μ so as to maximize this. Now, potential customers arrive at a rate λ. However, a certain proportion of them do not join the system—namely, those who arrive when there are N customers already in the system. Hence, since P_N is the proportion of time that the system is full, it follows that entering customers arrive at a rate of $\lambda(1-P_N)$. Since each customer pays \$A, it follows that money comes in at an hourly rate of $\lambda(1-P_N)A$ and since it goes out at an hourly rate of $c\mu$, it follows that our total profit per hour is given by

$$\begin{aligned}\text{profit per hour} &= \lambda(1-P_N)A - c\mu = \lambda A\left[1 - \frac{(\lambda/\mu)^N(1-\lambda/\mu)}{1-(\lambda/\mu)^{N+1}}\right] - c\mu \\ &= \frac{\lambda A[1-(\lambda/\mu)^N]}{1-(\lambda/\mu)^{N+1}} - c\mu\end{aligned}$$

For instance if $N=2, \lambda=1, A=10, c=1$, then

$$\text{profit per hour} = \frac{10[1-(1/\mu)^2]}{1-(1/\mu)^3} - \mu = \frac{10(\mu^3-\mu)}{\mu^3-1} - \mu$$

Page 23, **Example 4.6** Continuous
In order to maximize profit we differentiate to obtain

$$\frac{d}{d\mu}[\text{profit per hour}] = 10\frac{(2\mu^3 - 3\mu^2 + 1)}{(\mu^3 - 1)} - 1$$

The value of μ that maximizes our profit now can be obtained by equating to zero and solving numerically.

Page 29, **Example 4.7**
Using that $\frac{\lambda^n}{\mu^n k! k^{n-k}} = (\lambda/k\mu)^n k^k/k!$ we see that

$$\begin{aligned} P_0 &= \frac{1}{1+\sum_{n=1}^{k}(\lambda/\mu)^n/n! + \sum_{n=k+1}^{\infty}(\lambda/k\mu)^n k^k/k!,} \\ P_n &= P_0(\lambda/\mu)^n/n!, \text{ if } n \le k \\ P_n &= P_0(\lambda/k\mu)^n k^k/k!, \text{ if } n > k \end{aligned}$$

It follows from the preceding that the condition needed for the limiting probabilities to exist is $\lambda < k\mu$. Because $k\mu$ is the service rate when all servers are busy, the preceding is just the intuitive condition that for limiting probabilities to exist the service rate needs to be larger than the arrival rate when there are many customers in the system.

Page 29, **Example 4.8**

Letting $\mu_2 = 2\mu$, the long run proportions for the M/M/2 system can be expressed as

$$P_n = 2(\lambda/\mu_2)^n P_0, n \geq 1$$

This yields that

$$1 = \sum_{n=0}^{\infty} P_n = P_0 \left(1 + 2 \sum_{n=1}^{\infty} (\lambda/2\mu_2)^n \right) = P_0 \left(1 + \frac{\lambda/\mu}{1 - \lambda/\mu_2} \right) = P_0 \left(\frac{1 + \lambda/\mu_2}{1 - \lambda/\mu_2} \right)$$

Thus

$$P_0 = \frac{1 - \lambda/\mu_2}{1 + \lambda/\mu_2}$$

To determine W, we first compute L. This gives

$$L = \sum_{n=1}^{\infty} nP_n = 2P_0 \sum_{n=1}^{\infty} n(\lambda/\mu_2)^n = 2P_0 \frac{\lambda/\mu_2}{(1 - \lambda/\mu_2)^2} = \frac{\lambda/\mu}{(1 - \lambda/\mu_2)(1 + \lambda/\mu_2)}$$

Because $L = \lambda W$, the preceding gives

$$W = \frac{1}{(\mu - \lambda/2)(1 + \lambda/\mu_2)}$$

Page 29, **Example 4.8** continuous

It is interesting to contrast the average time in the system when there is a single queue as in the M/M/2, with when arrivals are randomly sent to be served by either server. As shown in Example 4.4, the average time in the system in the latter case is minimized when each customer is equally likely to be sent to either server, with the average time being equal to $\frac{1}{\mu-\lambda/2}$ in this case. Hence, the average time that a customer spends in the system when using a single queue as in the M/M/2 system is $\frac{1}{1+\lambda/\mu_2}$ multiplied by what it would be if each customer were equally likely to be sent to either server's queue. For instance, if $\lambda = \mu = 1$, then $\lambda/\mu_2 = 1/2$, and the use of a single queue results in the customer average time in the system being equal to $2/3$ times what it would be if two separate queues were used. When $\lambda = 1.5\mu$, the reduction factor becomes $4/7$; and when $\lambda = 1.9\mu$, it is 20/39.

Page 29, **Example 4.9**

This system can be modeled as a birth and death process with birth and death rates

$$\begin{aligned} \lambda_n &= \lambda, n \geq 0 \\ \mu_n &= \mu + (n-1)\alpha, n \geq 1 \end{aligned}$$

Using the previously obtained limiting probabilities enables us to answer a variety of questions about this system. For instance, suppose we wanted to determine the proportion of arrivals that receive service. Calling this quantity π_s, it can be obtained by letting λ_s be the average rate at which customers are served and noting that

$$\pi_s = \frac{\lambda_s}{\lambda}$$

To verify the preceding equation, let $N_a(t)$ and $N_s(t)$ denote, respectively, the number of arrivals and the number of services by time t. Then,

$$\pi_s = \lim_{t\to\infty} \frac{N_s(t)}{N_a(t)} = \lim_{t\to\infty} \frac{N_s(t)/t}{N_a(t)/t} = \frac{\lambda_s}{\lambda}$$

Page 29, **Example 4.9** continuous

Because the service departure rate is 0 when the system is empty and is μ when the system is nonempty, it follows that $\lambda_s = \mu(1 - P_0)$, yielding that

$$\pi_s = \frac{\mu(1 - P_0)}{\lambda}$$

Page 35, **Example 4.10**

If we let λ_j denote the total arrival rate of customers to server j, then the λ_j can be obtained as the solution of

$$\lambda_j = r_j + \sum_{i=1}^{k} \lambda_i P_{ij}, i = 1, \ldots, k$$

The above equation follows since r_j is the arrival rate of customers to j coming from outside the system and, as λ_i is the rate at which customers depart server i (rate in must equal rate out), $\lambda_i P_{ij}$ is the arrival rate to j of those coming from server i. The total arrival rates to servers 1 and 2—call them λ_1 and λ_2—can be obtained from the above eqaution. That is, we have

$$\begin{aligned} \lambda_1 &= 4 + \frac{1}{4}\lambda_2, \\ \lambda_2 &= 5 + \frac{1}{2}\lambda_1 \end{aligned}$$

Page 35, **Example 4.10** Continuous

implying $\lambda_1 = 6, \lambda_2 = 8$. Hence

$$P(n \text{ at server 1}, \ m \text{ at server 2}) = \left(\frac{3}{4}\right)^n \frac{1}{4} \left(\frac{4}{5}\right)^m \frac{1}{5} = \frac{1}{20} \left(\frac{3}{4}\right) \left(\frac{4}{5}\right)$$

and

$$L = \frac{6}{8-6} + \frac{8}{10-8} = 7$$

$$W = \frac{L}{9} = \frac{7}{9}$$

Page 40, **Example 4.11**

First note that this model is a special case of the M/G/1 model, where if S is the service time of a customer, than the service distribution G is obtained by conditioning on the type of the customer:

$$G(x) = P(S \leq x) = \sum_{i=1}^{n} P(S \leq x | \text{customer is type } i)\alpha_i = \sum_{i=1}^{n} F_i(x)\alpha_i$$

To compute $\mathrm{E}[S]$ and $\mathrm{E}[S^2]$, we condition on the customer's type. This yields

$$\mathrm{E}[S] = \sum_{i=1}^{n} \mathrm{E}[S|\text{type } i]\alpha_i = \sum_{i=1}^{n} \mu_i \alpha_i$$

and

$$\mathrm{E}[S^2] = \sum_{i=1}^{n} \mathrm{E}[S^2|\text{type } i]\alpha_i = \sum_{i=1}^{n} (\mu_i^2 + \sigma_i^2)\alpha_i$$

where the final equality used that $\mathrm{E}[X^2] = \mathrm{E}^2[X] + \mathrm{Var}(X)$. Now, because the time that a customer spends in queue is equal to the work in the system when that customer arrives, it follows that the average time that a type j customer spends in queue, call it $W_Q(j)$, is equal to the average work seen by a time j arrival. However, because type j customers arrive according to a Poisson process with rate $\lambda\alpha_j$ it follows, from the PASTA principle, that the work seen by a type j arrival has the same distribution as the work as it averages over time, and thus the average work seen by a type j arrival is equal to V . Consequently,

$$W_Q(j) = V = \frac{\lambda \mathrm{E}[S^2]}{2(1-\lambda \mathrm{E}[S])} = \frac{\lambda \sum_{i=1}^{n} (\mu_i^2 + \sigma_i^2)\alpha_i}{2(1-\lambda \sum_{i=1}^{n} \mu_i \alpha_i)}$$

With $W(j)$ being the average time that a type j customer spends in the system, we have

$$W(j) = W_Q(j) + \mu_j$$

Page 40, **Example 4.11** continuous

Finally, using that the average number of type j customers in the system is the average arrival rate of type j customers times the average time they spend in the system ($L = \lambda_a W$ applied to type j customers), we see that $L(j)$, the average number of type j customers in the system, is

$$L(j) = \lambda \alpha_j W(j)$$