5.2 $M/M/k$ Queuing Systems

An M / M / k queuing system is a multiserver extension of an M / M / 1 system.

Definition 5.1. *An* ***M / M / k queuing process*** *is a continuous-time Markov queuing process with the following characteristics:*

- *k servers;*
- *unlimited capacity;*
- *Exponential interarrival times with arrival rate λ_A;*
- *Exponential service times with service rate λ_S;*
- *independent service and arrival times, independent service times of all servers.*

Once again, we use the same approach as before, move from the discrete-time BkSQP to the continuous-time M / M / k process by letting the frame size $\Delta \to 0$. Recall that

$$\begin{aligned} p_A &= \lambda_A \Delta, \\ p_S &= \lambda_S \Delta. \end{aligned}$$

For very small Δ, we neglect terms of the form Δ^l, for $l \geq 2$, so the transition probabilities for a BkSQP become:

$$\begin{aligned}
p_{i,i+1} &= p_A(1-p_S)^n \;=\; \lambda_A\Delta(1-\lambda_S\Delta)^n \;\approx\; \lambda_A\Delta \;=\; \underline{p_A} \\
p_{i,i} &= p_A\, C_n^1\, p_S(1-p_S)^{n-1} + (1-p_A)(1-p_S)^n \\
&= \lambda_A\Delta\, n\, \lambda_S\Delta(1-\lambda_S\Delta)^{n-1} + (1-\lambda_A\Delta)(1-\lambda_S\Delta)^n \\
&\approx (1-\lambda_A\Delta)\left(1-\, C_n^1\lambda_S\Delta + \ldots\right) \\
&\approx 1-\lambda_A\Delta - n\,\lambda_S\Delta \;=\; \underline{1-p_A-np_S} \\
p_{i,i-1} &= p_A\, C_n^2\, p_S^2(1-p_S)^{n-2} + (1-p_A)\, C_n^1\, p_S(1-p_S)^{n-1} \\
&= \lambda_A\Delta\frac{n(n-1)}{2}\left(\lambda_S\Delta\right)^2\left(1-\lambda_S\Delta\right)^{n-2} \\
&+ (1-\lambda_A\Delta)\, n\, \lambda_S\Delta(1-\lambda_S\Delta)^{n-1} \\
&\approx n\,\lambda_S\Delta \;=\; \underline{np_S} \\
p_{i,j} &= 0,\ \forall j \neq i-1, i, i+1.
\end{aligned}$$

Recall that $n = \min\{i, k\}$ is the number of jobs receiving service among the total of i jobs in the system. Also, since Δ is very small, we ignored terms proportional to Δ^2, Δ^3, etc. Then, no more than one event, arrival or departure, may occur during each frame. Probability of more than one event is of the order $O(\Delta^2)$. Changing the number of jobs by 2 requires at least 2 events, and thus, such changes cannot occur during one frame. At the same time, transition from i to $i-1$ may be caused by a departure of any one of the n currently served jobs. This is why we have the departure probability p_S multiplied by n.

So, the transition probability matrix is

$$P \approx \begin{bmatrix} 1-p_A & p_A & 0 & 0 & \cdots & 0 & \cdots \\ p_S & 1-p_A-p_S & p_A & 0 & \cdots & 0 & \cdots \\ 0 & 2p_S & 1-p_A-2p_S & p_A & \cdots & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 & \cdots \\ 0 & 0 & \cdots & kp_S & 1-p_A-kp_S & 0 & \cdots \\ 0 & 0 & 0 & 0 & kp_S & 1-p_A-kp_S & \cdots \\ \vdots & \vdots & \vdots & \vdots & & \ddots & \end{bmatrix} \tag{5.1}$$

For example, for $k = 3$ servers, the transition probability matrix is

$$P \approx \begin{bmatrix} 1-p_A & p_A & 0 & 0 & 0 & 0 & \cdots \\ p_S & 1-p_A-p_S & p_A & 0 & 0 & 0 & \cdots \\ 0 & 2p_S & 1-p_A-2p_S & p_A & 0 & 0 & \cdots \\ 0 & 0 & 3p_S & 1-p_A-3p_S & p_A & 0 & \cdots \\ 0 & 0 & 0 & 3p_S & 1-p_A-3p_S & p_A & \cdots \\ 0 & 0 & 0 & 0 & 3p_S & 1-p_A-3p_S & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Next we find the steady-state distribution, as usually, from

$$\left\{\begin{array}{rcl} \pi P & = & \pi \\ \displaystyle\sum_{i=0}^{\infty} \pi_i & = & 1, \end{array}\right.$$

again, a system of infinitely many equations with infinitely many unknowns.
The first balance equation is

$$\begin{array}{rcl} [\pi_0 \;\; \pi_1 \;\; \pi_2 \;\; \ldots] \cdot \begin{bmatrix} 1 - p_A \\ p_S \\ 0 \\ \vdots \end{bmatrix} & = & \pi_0, \;\; \text{i.e.} \\ (1 - p_A)\pi_0 + p_S \pi_1 & = & \pi_0, \;\; \text{i.e.} \\ -p_A \pi_0 + p_S \pi_1 & = & 0, \;\; \text{i.e.} \\ p_S \pi_1 & = & p_A \pi_0, \;\; \text{i.e.} \\ \lambda_S \pi_1 & = & \lambda_A \pi_0. \end{array}$$

So,

$$\pi_1 \;=\; \frac{\lambda_A}{\lambda_S}\pi_0 \;=\; r\pi_0. \tag{5.2}$$

The second balance equation is

$$\begin{array}{rcl} [\pi_0 \;\; \pi_1 \;\; \pi_2 \;\; \ldots] \cdot \begin{bmatrix} p_A \\ 1 - p_A - p_S \\ 2p_S \\ 0 \\ \vdots \end{bmatrix} & = & \pi_1, \;\; \text{i.e.} \\ p_A \pi_0 + (1 - p_A - p_S)\pi_1 + 2p_S \pi_2 & = & \pi_1, \;\; \text{i.e.} \\ p_A \pi_0 - p_A \pi_1 - p_S \pi_1 + 2p_S \pi_2 & = & 0, \;\; \text{i.e. (since } p_A \pi_0 = p_S \pi_1) \\ 2p_S \pi_2 & = & p_A \pi_1, \;\; \text{i.e.} \\ 2\lambda_S \pi_2 & = & \lambda_A \pi_1. \end{array}$$

Thus, we get

$$\pi_2 \quad = \quad \frac{1}{2}\frac{\lambda_A}{\lambda_S}\pi_1 \quad = \quad \frac{1}{2}r\pi_1 \quad = \quad \frac{1}{2}r^2\pi_0. \tag{5.3}$$

The third balance equation is

$$\begin{aligned}
[\pi_0 \;\; \pi_1 \;\; \pi_2 \;\; \pi_3 \;\;\; \ldots] \cdot \begin{bmatrix} 0 \\ p_A \\ 1 - p_A - 2p_S \\ 3p_S \\ 0 \\ \vdots \end{bmatrix} \quad &= \quad \pi_2, \;\; \text{i.e.} \\
p_A\pi_1 + (1 - p_A - 2p_S)\pi_2 + 3p_S\pi_3 \quad &= \quad \pi_2, \;\; \text{i.e.} \\
p_A\pi_1 - p_A\pi_2 - 2p_S\pi_2 + 3p_S\pi_3 \quad &= \quad 0, \;\; \text{i.e. (since } p_A\pi_1 = 2p_S\pi_2\text{)} \\
3p_S\pi_3 \quad &= \quad p_A\pi_2, \;\; \text{i.e.} \\
3\lambda_S\pi_3 \quad &= \quad \lambda_A\pi_2.
\end{aligned}$$

So,

$$\pi_3 \quad = \quad \frac{1}{3}\frac{\lambda_A}{\lambda_S}\pi_2 \quad = \quad \frac{1}{3}r\pi_2 \quad = \quad \frac{1}{2 \cdot 3}r^3\pi_0 \quad = \quad \frac{1}{3!}r^3\pi_0. \tag{5.4}$$

We see a pattern forming. The k^{th} balance equation will yield

$$\pi_k \quad = \quad \frac{1}{k}r\pi_{k-1} \quad = \quad \frac{1}{k!}r^k\pi_0. \tag{5.5}$$

Then things change. Let us see the the $(k+1)^{\text{st}}$ equation.

$$\begin{aligned}
[\ldots\ \ \pi_{k-1}\ \ \pi_k\ \ \pi_{k+1}\ \ \ldots] \cdot \begin{bmatrix} 0 \\ \vdots \\ 0 \\ p_A \\ 1 - p_A - kp_S \\ kp_S \\ 0 \\ \vdots \end{bmatrix} &= \pi_k, \ \text{ i.e.} \\
p_A\pi_{k-1} + (1 - p_A - kp_S)\pi_k + kp_S\pi_{k+1} &= \pi_k, \ \text{ i.e.} \\
p_A\pi_{k-1} - p_A\pi_k - kp_S\pi_k + kp_S\pi_{k+1} &= 0, \ \text{ i.e. (since } p_A\pi_{k-1} = kp_S\pi_k\text{)} \\
kp_S\pi_{k+1} &= p_A\pi_k, \ \text{ i.e.} \\
k\lambda_S\pi_{k+1} &= \lambda_A\pi_k,
\end{aligned}$$

which yields

$$\pi_{k+1} = \frac{1}{k} r\pi_k = \left(\frac{r}{k}\right)\frac{r^k}{k!}\pi_0. \tag{5.6}$$

All the rest of the equations will be of the same form

$$\begin{aligned}
\pi_{k+2} &= \frac{1}{k} r\pi_{k+1} = \left(\frac{r}{k}\right)^2 \frac{r^k}{k!}\pi_0 \\
\ldots
\end{aligned} \tag{5.7}$$

Now we substitute them all in the normalizing equation $\sum_{i=0}^{\infty} \pi_i = 1$. We get

$$
\begin{aligned}
1 &= \pi_0 + \pi_1 + \ldots \\
&= \big(\pi_0 + \pi_1 + \ldots\ +\ \pi_{k-1}\big) + \big(\pi_k + \pi_{k+1} + \ldots\big) \\
&= \pi_0 \left[\left(1 + r + \frac{r^2}{2!} + \ldots + \frac{r^{k-1}}{(k-1)!} \right) \ +\ \frac{r^k}{k!} \left(1 + \frac{r}{k} + \left(\frac{r}{k}\right)^2 + \ldots \right) \right] \\
&= \pi_0 \left(\sum_{i=0}^{k-1} \frac{r^i}{i!} \ +\ \frac{r^k}{k!} \cdot \frac{1}{1 - r/k} \right) \\
&= \pi_0 \left(\sum_{i=0}^{k-1} \frac{r^i}{i!} \ +\ \frac{r^k}{k!(1 - r/k)} \right),
\end{aligned}
$$

where, in the last part, the Geometric series $\sum\limits_{i=0}^{\infty} \left(\frac{r}{k}\right)^i$ is convergent and equal to $\frac{1}{1 - r/k}$, if the ratio $r/k < 1$, i.e. $r < k$. So, the M/M/k steady-state distribution of number of jobs has pdf

$$
\begin{aligned}
\pi_0 &= P(X = 0) &&= \frac{1}{\sum\limits_{i=0}^{k-1} \frac{r^i}{i!} + \frac{r^k}{k!(1 - r/k)}}, \\
\pi_x &= P(X = x) &&= \begin{cases} \frac{r^x}{x!}\pi_0, & \text{for } x < k \\ \frac{r^k}{k!}\pi_0 \left(\frac{r}{k}\right)^{x-k}, & \text{for } x \geq k \end{cases},
\end{aligned}
\tag{5.8}
$$

provided that

$$
r \ =\ \frac{\lambda_A}{\lambda_S} \ <\ k.
$$

Example 5.2. Consider again Example 4.5 (and 4.6) from Lecture 9 about the message transmission center (*Messages arrive to a communication center at random times according to a Poisson process, with an average of* 5 *messages per minute. They are transmitted through a single channel in the order they were received. On average, it takes* 10 *seconds to transmit a message*). Recall that when the number of customers increased by 10%, all the parameters of the system increased significantly, some of them even by more than 100%. Suppose now that the arrival rate has doubled to 10 messages per minute. On average, it still takes 10 seconds to transmit a message, but assume that 2 additional channels are built with the same parameters as the first channel. Evaluate the new system's performance. What percentage of messages will be sent immediately, with no waiting time?

Solution. This is now an M / M / 3 system with

$$\begin{aligned} \lambda_A &= 10 \text{ / minute } = 1/6 \text{ / second}, \\ \mu_S &= 10 \text{ seconds } = \frac{1}{6} \text{ minutes}, \\ \lambda_S &= 6 \text{ / minute}, \\ r &= \frac{10}{6} = \frac{5}{3} = 1.667 > 1, \text{ but } r < 3. \end{aligned}$$

Before we proceed with the computation of $E(X)$, let us recall a few formulas related to the Geometric series:

- the Geometric series $\sum_{i=0}^{\infty} a_0 q^i$ is convergent if the ratio $|q| < 1$ and its sum is equal to

$$\sum_{i=0}^{\infty} a_0 q^i = \frac{a_0}{1-q};$$

- under the same conditions (differentiating the equation above with respect to q), the following series is also convergent

$$\sum_{i=0}^{\infty} a_0 i \, q^i = \frac{a_0 q}{(1-q)^2}.$$

The steady-state distribution, by (5.8) is given by

$$\pi_0 = \frac{1}{\sum_{i=0}^{2} \frac{r^i}{i!} + \frac{r^3}{3!(1-r/3)}} = \frac{1}{1 + r + \frac{r^2}{2!} + \frac{r^3}{3!(1-r/3)}} = 0.1727$$

and

$$\pi_x \;=\; \begin{cases} \dfrac{r^x}{x!}\pi_0, & \text{for } x = 1, 2 \\ \\ \dfrac{r^3}{3!}\pi_0 \left(\dfrac{r}{3}\right)^{x-3}, & \text{for } x = 3, 4, \ldots \end{cases} .$$

Then

$$\begin{aligned} E(X) &= \sum_{x=0}^{\infty} x\pi_x \;=\; \sum_{x=0}^{2} x\pi_x + \sum_{x=3}^{\infty} x\pi_x \\ &= \pi_0 \left(0 + 1 \cdot r + 2 \cdot \frac{r^2}{2}\right) + \pi_0 \frac{r^3}{3!} \sum_{x=3}^{\infty} x \left(\frac{r}{3}\right)^{x-3} \\ &= \pi_0 \left(r + r^2\right) + \pi_0 \frac{r^3}{3!} \sum_{j=0}^{\infty} (j+3) \left(\frac{r}{3}\right)^{j} \\ &= \pi_0 \left(r + r^2\right) + \pi_0 \frac{r^3}{3!} \left[\sum_{j=0}^{\infty} j \left(\frac{r}{3}\right)^{j} + 3 \sum_{j=0}^{\infty} \left(\frac{r}{3}\right)^{j}\right] \\ &= \pi_0 \left(r + r^2\right) + \pi_0 \frac{r^3}{6} \left[\frac{r/3}{(1 - r/3)^2} + 3 \frac{1}{1 - r/3}\right] \\ &= \pi_0 \left(r + r^2\right) + \frac{\pi_0 r^3 (9 - 2r)}{2(3 - r)^2} \\ &= \pi_0 \left(r + r^2 + \frac{r^3 (9 - 2r)}{2(3 - r)^2}\right) \\ &= 2.0418. \end{aligned}$$

Thus, the average number of messages stored in the system at any time is

$$E(X) \;=\; 2.0418.$$

By Little's law, the total time from arrival until the end of transmission has an average of

$$E(R) \;=\; E(X)/\lambda_A \;=\; 0.20418 \text{ minutes} \;=\; 12.2508 \text{ seconds}.$$

When a message arrives to the center, its average waiting time until transmission is

$$E(W) \;=\; E(R) - E(S) \;=\; E(R) - \mu_S \;=\; 12.2508 - 10 \;=\; 2.2508 \text{ seconds}.$$

Then, using Little's law again, the average number of messages waiting to be transmitted is

$$E(X_w) \;=\; \lambda_A E(W) \;=\; 1/6 \cdot 2.2508 \;=\; 0.3751.$$

Finally, the average number of messages being transmitted is

$$E(X_s) \;=\; E(X) - E(X_w) \;=\; 2.0418 - 0.3751 \;=\; 1.6667 \;=\; r.$$

Alternatively, for the last one, by Little's law,

$$E(X_s) \;=\; \lambda_A E(S) \;=\; \lambda_A \mu_S \;=\; \frac{\lambda_A}{\lambda_S} \;=\; r,$$

just like in the case of an M / M / 1 system.

To answer the last question, a message does not wait at all if there is an idle server (channel) to service (transmit) it. That happens when the number of jobs in the system is *less* than the number of servers $k = 3$. So,

$$\begin{aligned} P(W=0) &= P(X<3) \;=\; P(X=0 \text{ or } X=1 \text{ or } X=2) \\ &= \pi_0 + \pi_1 + \pi_2 \\ &= \pi_0 \left(1 + r + \frac{r^2}{2!}\right) \\ &= \frac{73}{18}\pi_0 \;=\; 0.7004. \end{aligned}$$

Or, we can directly compute

$$\begin{aligned} \pi_1 &= \frac{r}{1!}\pi_0 \;=\; 0.2878, \\ \pi_2 &= \frac{r^2}{2!}\pi_0 \;=\; 0.2398, \\ P(W=0) &= \pi_0 + \pi_1 + \pi_2 \;=\; 0.7004. \end{aligned}$$

That means that 70% of the messages are transmitted immediately, with no waiting time.

■

6 $M/M/\infty$ Queuing Systems

Let us now consider an unlimited number of servers $k = \infty$. That *completely* eliminates the waiting time. Whenever a job arrives, there will always be servers available to handle it and thus,

$$\begin{aligned} X &= X_s, \text{ the number of jobs in the system is the number of jobs receiving service,} \\ R &= S, \text{ response time consists of service time only,} \\ X_w &= 0, \text{ no jobs waiting in queue,} \\ W &= 0, \text{ no waiting time.} \end{aligned}$$

All the formulas we derived for M / M / k systems apply to M / M / ∞ systems, by letting the number of servers $k \to \infty$. Let us see what we get.

First off, the number of jobs will always be less than the number of servers ($i < k$), so we always have $n = i$. That is, with i jobs in the system, exactly i servers are busy.

The transition probability matrix for the number of jobs in the system, X, is given by

$$P = \begin{bmatrix} 1-p_A & p_A & 0 & 0 & 0 & 0 & \ldots \\ p_S & 1-p_A-p_S & p_A & 0 & 0 & 0 & \ldots \\ 0 & 2p_S & 1-p_A-2p_S & p_A & 0 & 0 & \ldots \\ 0 & 0 & 3p_S & 1-p_A-3p_S & p_A & 0 & \ldots \\ 0 & 0 & 0 & 4p_S & 1-p_A-4p_S & p_A & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Let us see what becomes the steady-state distribution. The first component, π_0, becomes

$$\pi_0 = P(X=0) = \frac{1}{\sum_{i=0}^{\infty} \frac{r^i}{i!} + \lim_{k\to\infty}\left(\frac{r^k}{k!}\cdot\frac{1}{1-r/k}\right)}.$$

Now,

$$\lim_{k\to\infty}\left(\frac{r^k}{k!}\cdot\frac{1}{1-r/k}\right) \;=\; \lim_{k\to\infty}\frac{r^k}{k!} \;=\; 0,$$

because the factorial converges faster to ∞ than the exponential function. The other term is the Taylor series of the function e^r, so the steady-state distribution is

$$\begin{aligned}\pi_0 &= e^{-r},\\ \pi_i &= \frac{r^i}{i!}e^{-r},\ \ \forall i \geq 1.\end{aligned} \tag{6.1}$$

So the pdf of $X(t)$, the number of concurrent jobs in an M / M / ∞ system at time t, is

$$X(t)\ \left(\begin{array}{c} i \\ \dfrac{r^i}{i!}e^{-r} \end{array}\right)_{i=0,1,\ldots}, \tag{6.2}$$

a Poisson distribution with parameter $r = \dfrac{\lambda_A}{\lambda_S}$ (which can be arbitrarily large), with mean and variance

$$E(X) \;=\; V(X) \;=\; r. \tag{6.3}$$

Remark 6.1. Clearly, nobody can physically build an *infinite* number of devices. In practice, having an unlimited number of servers simply means that any number of concurrent jobs can be served simultaneously. Example: internet service providers or telephone companies (which allow virtually any number of concurrent connections), an unlimited number of people can watch a TV channel or listen to a radio station, etc. A model with infinitely many servers is a reasonable approximation for a system where jobs typically don't wait and get their service immediately. This may be appropriate for a computer server, a grocery store, Facebook, etc.

Example 6.2. A certain powerful server can afford practically any number of concurrent users. Users connect to the server at random times, every 3 minutes, on the average, according to a Poisson counting process. Each user spends an Exponential amount of time on the server with an average of 1 hour and disconnects from it, independently of other users. Find
a) the fraction of time when no users are connected to the server;
b) the expected number of concurrent users at any time;
c) if a message is sent to all users, the probability that 15 or more users will receive this message immediately.

Solution. This fits the description of an M / M / ∞ system with

$$\begin{aligned}
\mu_A &= \text{3 minutes, so} \\
\lambda_A &= 1/\mu_A \;=\; 1/3 \text{ / minute,} \\
\mu_S &= \text{1 hour} \;=\; 60 \text{ minutes, so} \\
\lambda_S &= 1/\mu_S \;=\; 1/60 \text{ / minute,} \\
r &= \frac{\lambda_A}{\lambda_S} \;=\; \frac{1/3}{1/60} \;=\; 20.
\end{aligned}$$

The number of concurrent users has $Poiss(20)$ distribution.

a)

$$P(X=0) \;=\; \pi_0 \;=\; e^{-20} \;=\; 2.06 \cdot 10^{-9} \;=\; 0.$$

This server is practically *never* idle.

b) The expected number of concurrent users is

$$E(X) \;=\; r \;=\; 20 \text{ users.}$$

Also, if an urgent message is sent to all the users, then 20 users, on the average, will see it immediately.

c) Fifteen or more users will receive a message immediately if 15 or more users are connected, so, with probability

$$\begin{aligned}
P(X \geq 15) &= 1 - P(X < 15) \;=\; 1 - P(X \leq 14) \\
&= 1 - \mathit{poisscdf}(14, 20) \;=\; 0.8951.
\end{aligned}$$

■

7 Simulation of Queuing Systems

We developed a theory and understood how to analyze and evaluate rather basic queuing systems: Bernoulli and M / M / k. Most of the results were obtained from the Markov property of the considered queuing processes. For these systems, we derived a steady-state distribution of the number of concurrent jobs and computed the vital performance characteristics from it.

In practice, however, many queuing systems have a rather complex structure. Jobs may arrive according to a non-Poisson process, often the rate of arrivals changes during the day (there is a rush hour on highways or on the internet, etc.). Service times may have different distributions and they are not always memoryless, thus the Markov property may not be satisfied. The number of servers may also change during the day (additional servers may turn on during rush hours). Some customers may get dissatisfied with a long waiting time and quit in the middle of their queue. And so on. Queuing theory does not cover all the possible situations. On the other hand, we can simulate the behavior of almost any queuing system and study its properties by Monte Carlo methods.

A queuing system is Markov only when its interarrival and service times are memoryless. Then the future can be predicted from the present without relying on the past. It can be simulated using the algorithm given for Markov chains (Algorithm 2.13 in Lecture 5). To study long-term characteristics of a queuing system, the initial distribution of X_0 typically does not matter, so we may start this algorithm with 0 jobs in the system and then “switch on” the servers.

Even when the system is Markov, some interesting characteristics do not follow from its steady-state distribution directly, but they can be estimated from a Monte Carlo study.

Performance of more complicated and advanced queuing systems can be evaluated by Monte Carlo methods. One needs to simulate arrivals of jobs, assignment of servers and service times and to keep track of all variables of interest. Monte Carlo methods of Chapter 2 let us simulate and evaluate rather complex queuing systems far beyond Bernoulli and M/M/k. As long as we know the distributions of interarrival and service times, we can generate the processes of arrivals and services. To assign jobs to servers, we keep track of servers that are available each time when a new job arrives. When all the servers are busy, the new job will enter a queue. As we simulate the work of a queuing system, we keep records of events and variables that are of interest to us. After a large number of Monte Carlo runs, we average our records in order to estimate probabilities by long-run proportions and expected values by long-run averages.