Chapter 6

Queueing Theory

6.1 Continuous Time Markov Chains

It is not always appropriate to model time using discrete steps as we did in Chapter 5. In *Continuous Time Markov Chains* (CTMCs), events (state transitions) occur at arbitrary points in time. The probability distribution underlying CTMCs is the *exponential distribution*.

Definition 6.1 (Exponential Distribution)**.** *A random variable Y with the cumulative distribution function (CDF)*

$$F_Y(t) = \Pr[Y \leq t] := \begin{cases} 1 - e^{-\lambda t} & \textit{for } t \geq 0, \textit{ and} \\ 0 & \textit{otherwise} \end{cases}$$

is ***exponentially distributed*** *with parameter λ, or $Y \sim \exp(\lambda)$ for short. The corresponding probability density function (PDF) is*

$$f_Y(t) = \frac{\mathrm{d}}{\mathrm{d}t} F_Y(t) = \lambda e^{-\lambda t}\,.$$

Remarks:

- If $Y \sim \exp(\lambda)$, then $\mathbb{E}[Y] = 1/\lambda$ and $\mathrm{Var}[Y] = 1/\lambda^2$.
- The exponential distribution is the continuous analogue to the geometric distribution—it is the only memoryless continuous distribution.
- Consider the continuous time stochastic process $\{A_t : t \in \mathbb{R}_{\geq 0}\}$ counting the number of events up to time t, where the time between two consecutive events is exponentially distributed with parameter λ. Then A_t is a *Poisson process* with *rate* λ.

Definition 6.2 (Continuous Time Markov Chain, CTMC)**.** *Let S be a finite or countably infinite set of* ***states****. A* ***Continuous Time Markov Chain (CTMC)*** *is a continuous time* ***stochastic process*** *$\{X_t : t \in \mathbb{R}_{\geq 0}\}$ with $X_t \in S$ for all t that satisfies the continuous Markov property.*

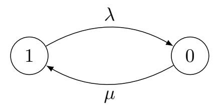

Figure 6.1: A CTMC modeling an unreliable system. In state 1 the system is working, in state 0 the system is faulty. The *failure rate*, i.e., the time until the system fails, is exponentially distributed with parameter λ. After a failure, the repair takes some time, exponentially distributed with parameter μ.

Remarks:

- The continuous Markov property, i.e., that the probability for the next state depends only on the current state, can be defined similar to the discrete Markov property, see Definition 5.2.
- We will only consider *time-homogeneous* CTMCs for which the transition probability $\Pr[X_{t_2} = j | X_{t_1} = i]$ from state i to j in the time period $[t_1, t_2)$ depends only on the difference $\Delta t = t_2 - t_1$ and not on the times t_1, t_2 themselves.
- The sojourn times (Definition 5.5) for time-homogeneous CTMCs are exponentially distributed, cf. Definition 6.1.
- What happens when a state has more than one possible following states? The following lemmas connect two different possible ways to model this case.

Lemma 6.3. *Let $Y_1, \ldots, Y_k$ be k independent exponential random variables with corresponding parameters $\lambda_1, \ldots, \lambda_k$. The random variable $Y = \min\{Y_1, \ldots, Y_k\}$ is exponentially distributed with parameter $\lambda_1 + \cdots + \lambda_k$.*

Proof. We establish the claim for $k = 2$. The general case can be derived by applying the same reasoning. By definition it holds for Y, Y_1, and Y_2 that

$$\Pr[Y > t] = \Pr[\min\{Y_1, Y_2\} > t] = \Pr[Y_1 > t, Y_2 > t]\,.$$

Since the random variables Y_1 and Y_2 are independent, this is the same as

$$\begin{aligned} \Pr[Y > t] &= \Pr[Y_1 > t] \cdot \Pr[Y_2 > t] \\ &= e^{-\lambda_1 t} \cdot e^{-\lambda_2 t} = e^{-(\lambda_1 + \lambda_2)t}\,. \end{aligned}$$

It follows that the random variable $Y = \min\{Y_1, Y_2\}$ is exponentially distributed with parameter $\lambda_1 + \lambda_2$. □

Lemma 6.4. *Let $Y_1, \ldots, Y_k$ be k independent exponential random variables with corresponding parameters $\lambda_1, \ldots, \lambda_k$. The probability $\Pr[Y_1 = \min\{Y_1, \ldots, Y_k\}]$ is $\frac{\lambda_1}{\lambda_1 + \cdots + \lambda_k}$.*

Proof. Let Z be the random variable $Z = \min\{Y_2, \ldots, Y_k\}$. Lemma 6.3 states that Z is exponentially distributed with parameter $\mu = \lambda_2 + \cdots + \lambda_k$. Applying the law of total probability we obtain that the probability for Y_1 to take on the smallest value is

$$\begin{aligned}\Pr[Y_1 < Z] &= \int_0^\infty \Pr[Y_1 < Z | Y_1 = t] \cdot f_{Y_1}(t)\,\mathrm{d}t \\ &= \int_0^\infty \Pr[t < Z | Y_1 = t] \cdot f_{Y_1}(t)\,\mathrm{d}t\,.\end{aligned}$$

Since Z is independent of Y_1, we can simplify to

$$\Pr[Y_1 < Z] = \int_0^\infty (1 - \Pr[Z \le t]) \cdot f_{Y_1}(t)\,\mathrm{d}t\,.$$

Recall that the probability density function of Y_1 is $f_{Y_1}(t) = \lambda_1 e^{-\lambda_1 t}$, and that the cumulative distribution function for Z is $F_Z(t) = 1 - e^{-\mu t}$. Plugging both in, we obtain

$$\begin{aligned}\Pr[Y_1 < Z] &= \lambda_1 \int_0^\infty e^{-\mu t} \cdot e^{-\lambda_1 t}\,\mathrm{d}t = \lambda_1 \int_0^\infty e^{-(\lambda_1+\mu)t}\,\mathrm{d}t \\ &= \lambda_1 \cdot \left.\frac{-e^{-(\lambda_1+\mu)t}}{\lambda_1+\mu}\right|_0^\infty = \lambda_1 \cdot \left(0 - \frac{-e^0}{\lambda_1+\mu}\right) \\ &= \frac{\lambda_1}{\lambda_1+\mu} = \frac{\lambda_1}{\lambda_1+\cdots+\lambda_k}\,,\end{aligned}$$

as desired. □

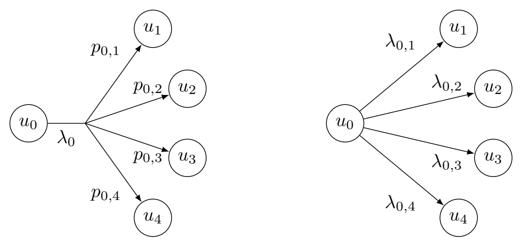

Figure 6.2: Two equivalent views: On the left, the state u_0 has a sojourn time Y exponentially distributed with parameter λ_0. After time Y has passed, the next state is chosen according to the probability distribution $(p_{0,1}, p_{0,2}, p_{0,3}, p_{0,4})$. On the right, the sojourn time in state 0 is $\min\{Y_1, Y_2, Y_3, Y_4\}$, where $Y_i \sim \exp(\lambda_{0,i})$ with parameter $\lambda_{0,i} = p_{0,i} \cdot \lambda_0$. The next state is determined by the random variable that realizes the minimum.

Remarks:

- Lemmas 6.3 and 6.4 together state that the two views in Figure 6.2 are equivalent.

- As indicated in Figure 6.2, we denote by λ_i the parameter for the exponential distribution describing the sojourn time in state i. The probability that state j is entered after leaving i is $p_{i,j}$, i.e., $p_{i,i} = 0$ and $\sum_{j\in S} p_{i,j} = 1$. The *transition rate* from state i to j is $\lambda_{i,j} = \lambda_i \cdot p_{i,j}$. Thus, for any $i \in S$ it holds that $\sum_{j\in S} \lambda_{i,j} = \lambda_i$.
- At any given moment, what is the probability that the example system from Figure 6.1 is faulty? We write $q(t)$ for the probability distribution of states at time t, and $q_i(t)$ for the probability to be in state i at time t. Let's assume that at time 0 the system is working, i.e., $q(0) = (0, 1)$.
- It turns out that the change in q can be expressed using differential equations.

Theorem 6.5. *For all $i \in S$, the change in the state probability q_i is*

$$\underbrace{\frac{\mathrm{d}}{\mathrm{d}t} q_i(t)}_{\text{Change}} = \underbrace{\sum_{j: j\neq i} q_j(t) \cdot \lambda_{j,i}}_{\text{Into } i} - \underbrace{q_i(t) \cdot \lambda_i}_{\text{Out of } i}\,.$$

Remarks:

- Theorem 6.5 follows from the memoryless property and relies on the CTMC being time homogeneous.
- Solving such differential equations for exact values of t can be a laborious task. We can look at the stationary distribution instead. In the continuous case, a stationary distribution should satisfy that $\frac{\mathrm{d}}{\mathrm{d}t} q_i(t) = 0$ "after enough time has passed". For $t \to \infty$, we obtain that π is a stationary distribution if for all $i \in S$,

$$0 = \sum_{j: j\neq i} \pi_j \cdot \lambda_{j,i} - \pi_i \cdot \lambda_i\,.$$

- Thus, one can solve above system of linear equations in order to compute the stationary distribution. Since we are interested in a probability distribution, the solution must additionally satisfy the conditions $\pi_i \ge 0$ and $\sum_i \pi_i = 1$.

Theorem 6.6. *For finite irreducible CTMCs the limits*

$$\pi_i := \lim_{t\to\infty} q_i(t)$$

exist for all $i \in S$. Moreover, the entries in π are independent of $q(0)$.

Remarks:

- Just like in the discrete case (Definition 5.11) a CTMC is irreducible if for all states i and j it holds that j is reachable from i. That is, if there exists some $t \ge 0$ such that $\Pr[X_t = j \mid X_0 = i] > 0$.
- Because the times between two successive steps in the chain are random variables now, we do not have to worry about aperiodicity (Definition 5.14) any more!

- CTMCs for which the stationary distribution exists are called *ergodic*. For finite chains this is the same as being irreducible—as opposed to irreducible and aperiodic in the discrete case, cf. Definition 5.15. We will later see examples of irreducible infinite chains that are not ergodic.

- For our example from Figure 6.1 we obtain the following two equations:

$$0 = \mu \cdot \pi_0 - \lambda \cdot \pi_1\,, \text{ and}$$
$$0 = \lambda \cdot \pi_1 - \mu \cdot \pi_0\,.$$

  Since it must also hold that $\pi_1 + \pi_0 = 1$, we conclude that in the long run, the probability of being in the working respectively faulty state are

$$\pi_1 = \frac{\mu}{\lambda+\mu} \qquad \text{and} \qquad \pi_0 = \frac{\lambda}{\lambda+\mu}\,.$$

6.2 Kendall's Notation for Queues

Queueing theory can be a diversion to think about while queueing at the cash register, but it is also used in modeling telecommunication networks, traffic, factories, and internet servers. The queueing system we will study is illustrated in Figure 6.3.

Definition 6.7 (Jobs, Servers)**.** *A queueing system consists of a* ***queue*** *with one or more* ***servers*** *which process* ***jobs****. The queue acts as a buffer for jobs that arrived but cannot be processed yet, because the server is busy processing another job.*

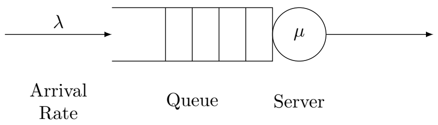

Figure 6.3: Example of a queueing system with one server. Jobs arrive at the queue from a Poisson process with rate λ, i.e., the inter-arrival time between two jobs is exponentially distributed with parameter λ. If the system is empty, the job is processed immediately, otherwise the job waits in the queue. The time it takes to process a single job is exponentially distributed with parameter μ, and after one job has been served, if there is a job waiting, the server starts to process the next job.

Remarks:

- A job may be a shopper, a phone call, a web request, etc. A server may model a checkout clerk, a factory, or the telephone network.

Definition 6.8 (Kendall's Notation)**.** *Let a and s be symbols describing the* ***arrival*** *and* ***service*** *rates, and let $m, n, j \in \mathbb{N}$. The* ***Kendall notation*** *for a queueing system Q is $a/s/m/n/j$. The symbols a and s can be D, M, or G, where*
D means that the rate distribution is ***degenerate****, i.e., of fixed length,*
M means that the arrival/service process is ***memoryless****, and*
G means that the corresponding rate stems from a ***generic*** *distribution.*
The parameter
m is the number of ***servers****,*
n is the number of ***places*** *in the system (in the queue and at servers), and*
j determines the external ***population*** *of jobs that may enter the system.*
The latter two parameters are omitted if the respective number is unbounded.

Remarks:

- Extensions to Kendall's notation include other kinds of distributions for arrival and service times. We will only consider memoryless processes, i.e., the arrival and service times are exponentially distributed.

- One reason is of course that the memoryless property allows for simpler math. But more importantly, memoryless processes turn out to be a good approximation for many real world systems, and thus memoryless queueing theory is a good tool to model such cases.

- When using this tool, one should be aware that for instance *bursty* behavior, where batches of jobs sometimes arrive in quick succession (think of a new trend appearing on Twitter) is not captured well by memoryless distributions.

- The parameter n in Kendall's notation limits how many jobs may be present in the system, and how many jobs are rejected by the queueing system. The parameter j affects the arrival rate—if a large fraction of the population is already in the queue, then jobs are less likely to arrive, and vice versa.

- Another parameter may be added to indicate the *queueing discipline*, i.e., in which order jobs are served. For our discussion this distinction is not necessary, and you may assume a *First In First Out* (FIFO) order. Other queueing disciplines are, e.g., *Last In First Out* (LIFO), random order, or queues where jobs have different priorities.

6.3 The $M/M/1$ Queue

Definition 6.9 (Utilization)**.** *The* ***utilization*** *of an $M/M/m$ queueing system is the fraction $\rho = \frac{\lambda}{m\mu}$.*

Theorem 6.10. *An $M/M/1$ queueing system has a stationary distribution if and only if $\rho < 1$. In that case the stationary distribution is $\pi_k = \rho^k(1-\rho)$.*

Proof. In the stationary distribution, the change in probability mass at every node must be zero. We obtain the equations

$$0 = \mu \cdot \pi_1 - \lambda\pi_0$$

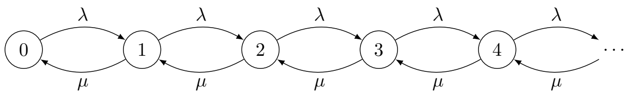

Figure 6.4: A continuous time Markov chain modeling an $M/M/1$ system. In state 0 the system is empty. When the chain is in state $i \geq 1$, then there are $i-1$ jobs in the queue, and one job is being served with rate μ. New jobs arrive with rate λ. Since the exponential distribution is memoryless, switching from state i to $i+1$ does not change the probability distribution for the service time of the currently processed job.

for state 0, and

$$0 = \lambda \cdot \pi_{k-1} + \mu \cdot \pi_{k+1} - (\lambda + \mu)\pi_k$$

for all $k \geq 1$. Rearranging yields

$$\mu \cdot \pi_{k+1} - \lambda \cdot \pi_k = \mu \cdot \pi_k - \lambda \cdot \pi_{k-1} = \cdots = \mu \cdot \pi_1 - \lambda \cdot \pi_0 = 0$$

$$\Rightarrow \mu \cdot \pi_k - \lambda \cdot \pi_{k-1} = 0 \Rightarrow \pi_k = \rho \cdot \pi_{k-1} \Rightarrow \pi_k = \rho^k \cdot \pi_0$$

In the case where $\rho \geq 1$ the only solution is $\pi = (0, 0, \ldots)$. This means that the queueing system does not converge, and that the length of the queue grows indefinitely. If on the other hand $\rho < 1$, then:

$$1 = \sum_{k=0}^{\infty} \pi_k = \pi_0 \cdot \sum_{k=0}^{\infty} \rho^k = \pi_0 \cdot \frac{1}{1-\rho} \Rightarrow \pi_0 = 1 - \rho\,.$$ □

Remarks:

- An $M/M/1$ queueing system is *stable* if $\rho = \frac{\lambda}{\mu} < 1$.
- Our model of the $M/M/1$ queueing system is an infinite irreducible CTMC, and the chain is ergodic if and only if $\rho < 1$.
- The probability that the single server in the queueing system is processing a job is $1 - \pi_0 = \rho$. This is why the fraction ρ is called *utilization*.
- In our proof of Theorem 6.10 we considered the flow of probability mass to and from a single state. It sometimes simplifies calculations to consider the flow of probability mass between *sets* of states instead. For example, in the $M/M/1$ case, first calculate the flow between state 0 and all other states, then calculate the flow from the states 0 and 1 to all other states, and so on.
- How many jobs are in the system in expectation?

Theorem 6.11. *In expectation there are $N = \frac{\lambda}{\mu - \lambda}$ jobs in an $M/M/1$ system.*

Proof. Denote by N the expected number of jobs in an $M/M/1$ queueing system. Using the stationary distribution we compute

$$\begin{aligned} N &= \sum_{k=0}^{\infty} k \cdot \pi_k = \sum_{k=0}^{\infty} k(1-\rho)\rho^k = (1-\rho)\rho \sum_{k=0}^{\infty} k\rho^{k-1} \\ &= (1-\rho)\rho \frac{1}{(1-\rho)^2} = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu - \lambda}\,, \end{aligned}$$

as claimed. □

Remarks:

- Similarly one can compute the variance as $\rho/(1-\rho)^2$.
- What is the average time a job stays in the system?

6.4 Little's Law

Definition 6.12 (Jobs in the System, Arrival Rate, Response Time)**.** *Consider any queueing system. Denote by $\overline{N}, \bar{\lambda}$, and $\overline{T}$ the random variables describing the average number of **jobs in the system**, the average **arrival rate**, and the average **time in the system** of a job (waiting time + service time), respectively.*

Theorem 6.13 (Little's Law)**.** *The three quantities from Definition 6.12 satisfy $\overline{N} = \bar{\lambda} \cdot \overline{T}$.*

Proof. For any point in time τ, we denote by $N(\tau)$ the number of jobs in the system at that time. Let t be some point in time, and for the sake of simplicity, assume that $N(0) = N(t) = 0$. Consider a realization of the queueing system, e.g., the one depicted in Figure 6.5. We write $\alpha(t)$ for the number of jobs that arrived until time t, and T_i for the response time of the i^{th} job. For any realization (Figure 6.5) it holds that

$$\sum_{i=1}^{\alpha(t)} T_i = \int_0^t N(\tau)\,\mathrm{d}\tau\,.$$

Multiplying both sides with $1/t$, and the left hand side with $1 = \alpha(t)/\alpha(t)$ we obtain by rearranging that

$$\frac{\alpha(t)}{t} \cdot \frac{1}{\alpha(t)} \sum_{i=1}^{\alpha(t)} T_i = \frac{1}{t} \int_0^t N(\tau)\,\mathrm{d}\tau\,.$$

This equation already states $\bar{\lambda} \cdot \overline{T} = \overline{N}$, as desired. □

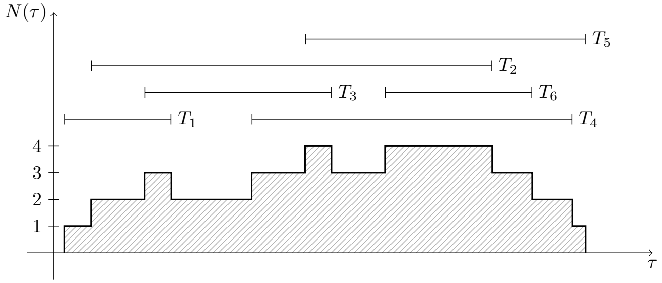

Figure 6.5: A possible realization of the random process describing a queueing system. The jobs and their response times are depicted as segments with lengths T_i, and the number of jobs in the system is given by the curve $N(\tau)$. In the proof of Little's Law the hatched area is measured in two ways. On the one hand, the hatched area can be obtained by taking the integral of the function $N(\tau)$. On the other hand, the hatched area is the sum $T_1 + \cdots + T_6$ due to the definition of $N(\tau)$.

Remarks:

- The simplifying assumption made in our proof, i.e., that at times 0 and t the system is empty, is not necessary for Theorem 6.13 to hold.
- Little's Law in the above form connects the random variables taking on *average* properties of a queueing system, and holds regardless of the probability distributions that describe the arrival and service times.
- It also holds for the *expected* values of $\overline{N}$, $\bar{\lambda}$, and $\overline{T}$. In many cases, for $t \to \infty$, the expected values are equal to the limit of the random variables with probability 1.
- So far we suggested a FIFO (first in first out) queueing discipline. To prove Little's Law this assumption was not required, i.e., Theorem 6.13 also holds for systems other than $M/M/1$ queues.
- Applying Little's Law we conclude that in the steady state the average time in the system (sometimes called *response time*) is $\overline{T} = \frac{N}{\lambda} = \frac{1}{\mu-\lambda}$, since for $M/M/1$ queueing systems we know that $N = \frac{\rho}{1-\rho}$.

Definition 6.14 (Waiting Time, Jobs in the Queue)**.** *We denote by $\overline{W}$ the average* ***waiting time*** *of a job (time spent in the queue) and by $\overline{N}_Q$ the average number of* ***jobs in the queue****.*

Remarks:

- Similar to the time in the system, for $M/M/1$ queueing systems the average waiting time of a job is $\overline{W} = \overline{T} - \frac{1}{\mu} = \frac{\rho}{\mu-\lambda}$.
- The average number of jobs in the queue is $\overline{N}_Q = \bar{\lambda}\overline{W} = \frac{\rho^2}{1-\rho}$.

6.5 Birth-Death Processes

Our CTMC for the $M/M/1$ queueing system is a special case of a so-called *Birth-Death Process.*

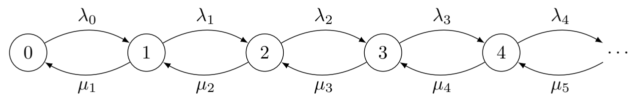

Figure 6.6: A generic Birth-Death Process.

Remarks:

- As before we can compute the stationary distribution. We obtain

$$\pi_0 = \frac{1}{1 + \sum_{k\geq 1} \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}}, \quad \text{and}$$
$$\pi_k = \pi_0 \cdot \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \quad \text{for } k \geq 1\,.$$

$M/M/m$ Queues

What if there is a single queue for multiple servers, e.g., in a service hotline? In Kendall's notation such systems are written as $M/M/m$ systems, where m denotes the number of servers.

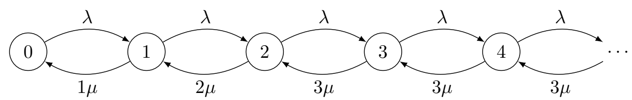

Figure 6.7: Birth-Death process modeling an $M/M/3$ queueing system. If there are less than 3 jobs, then the number of active servers is the number of jobs in the system. When 3 or more jobs are in the system all servers are active.

Remarks:

- In $M/M/m$ queueing systems, the utilization ρ is the average fraction of active servers.
- If $\rho = \frac{\lambda}{m\mu} < 1$, then the stationary distribution is

$$\pi_k = \begin{cases} \pi_0 \cdot \frac{(\rho m)^k}{k!} & \text{for } 1 \le k \le m \\ \pi_0 \cdot \frac{\rho^k m^m}{m!} & \text{for } k \ge m\,. \end{cases}$$

  and

$$\pi_0 = \frac{1}{\sum_{k=0}^{m-1} \frac{(\rho m)^k}{k!} + \frac{(\rho m)^m}{m!(1-\rho)}}\,.$$

- The probability that in the stationary distribution an arriving job has to wait in the queue is

$$\begin{aligned} P_Q &= \sum_{k=m}^{\infty} \pi_k = \sum_{k=m}^{\infty} \frac{\pi_0 \rho^k m^m}{m!} \\ &= \frac{\pi_0 (\rho m)^m}{m!} \sum_{k=m}^{\infty} \rho^{k-m} = \frac{\pi_0 (\rho m)^m}{m!(1-\rho)}\,. \end{aligned}$$

  Plugging in π_0 we obtain the following expression, which is also known as the *Erlang C Formula*:

$$P_Q = \frac{(\rho m)^m / (m!(1-\rho))}{\sum_{k=0}^{m-1} \frac{(\rho m)^k}{k!} + \frac{(\rho m)^m}{m!(1-\rho)}} \qquad (\text{for } \rho < 1)$$

- The average number of jobs in the queue $\overline{N}_Q$ can be calculated in a similar fashion. With P_Q the number can be expressed as

$$\overline{N}_Q = P_Q \cdot \frac{\rho}{1-\rho}\,.$$

The $M/M/m/n$ Queue

Often, the space in the queue is bounded, i.e., the system is $M/M/m/n$. Recall that n is the number of places in the system, so the maximum length of the queue is $n - m$.

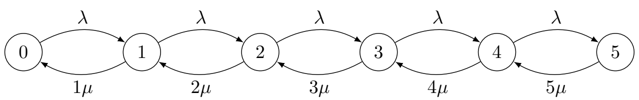

Figure 6.8: Birth-Death process modeling an $M/M/5/5$ queueing system.

Remarks:

- The case $m = n$ is often used to model communication networks. Such a system can accommodate m simultaneous calls, and the duration of a call is distributed with $\exp(\mu)$. One can calculate that in this case

$$\pi_k = \pi_0 \cdot \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!} \quad \text{for } 1 \le k \le m$$

  Using that $\sum_{k=0}^{m} \pi_k = 1$ yields that the probability to be in state 0 is

$$\pi_0 = \frac{1}{\sum_{k=0}^{m} \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}}$$

- The *blocking probability*, i.e., the probability that an arriving job is rejected, is thus

$$\pi_m = \frac{\left(\frac{\lambda}{\mu}\right)^m \frac{1}{m!}}{\sum_{k=0}^{m} \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}}$$

  This so-called *Erlang-B formula* also holds for $M/G/m/m$ systems where the service times are $1/\mu$ in expectation, regardless of their distribution.

The $M/M/n/m/m$ Queue

In telephone networks the population is assumed to be much larger than the number of places in the system. Thus, it is justified to assume that the arrival rate is independent of the number of jobs in the system. Cases where this assumption cannot be made can be modeled as $M/M/n/m/j$ systems.

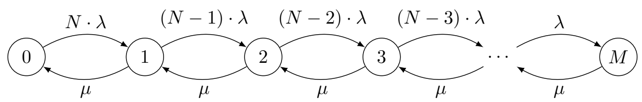

Figure 6.9: Birth-Death process modeling an $M/M/1/m/m$ queueing system.

Remarks:

- For $M/M/1/m/m$ systems, one can calculate that

$$\begin{aligned} \pi_k &= \pi_0 \cdot \prod_{i=0}^{k-1} \frac{\lambda(m-i)}{\mu} \quad \text{for } 1 \le k \le m \\ \pi_0 &= \frac{1}{\sum_{k=0}^{m} \left(\frac{\lambda}{\mu}\right)^k \cdot m^{\underline{k}}}\,, \end{aligned}$$

  where $m^{\underline{k}} := m(m-1)(m-2) \cdot \ldots \cdot (m-k+1)$.

6.6 Queueing Networks

Sometimes, systems consist of more than a single queueing system. Consider, for instance, a support call center where calls are initially handled by first-line support. Customers with problems that cannot be solved by the first-line support are handed over to technicians with a separate queue. See Figure 6.10 for an illustration.

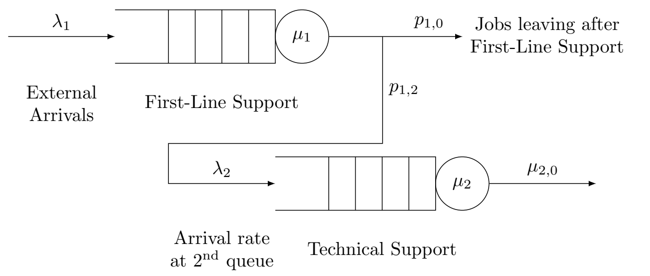

Figure 6.10: A queueing network modeling a two-tier support hotline. Jobs arrive from the outside with rate λ_1 and enter the queue for first-line support. After the first-line support served the job, with rate μ_1, a $p_{1,0} = (1 - p_{1,2})$ fraction of the jobs are satisfied and leave the system. The remaining $p_{1,2}$ fraction of jobs need in-depth technical assistance, which is provided by a technician in the technical support queue. Technical support takes time exponentially distributed with parameter μ_2, and afterwards the job leaves the system.

Remarks:

- Before looking at the whole network, let us look at a single queueing system. If the queueing system is stable, i.e., if $\rho < 1$, what is the inter-departure time between consecutive departing jobs?

Theorem 6.15 (Burke's Theorem)**.** *Consider a $M/M/m$ queue for arbitrary $m \in \mathbb{N}_0 \cup \{\infty\}$ with arrival rate λ and service rate μ. If the system is stable, then in the steady state the time between two departures is exponentially distributed with parameter λ.*

Proof. Consider any point in time, and let T be the random variable for the time until the next job leaves the queueing system. Denoting by ρ the probability that the system is not empty, we can write

$$\Pr[T \le t] = \rho \cdot \Pr[T \le t \mid \text{system not empty}] + (1-\rho) \cdot \Pr[T \le t \mid \text{system empty}]$$

When the queueing system is not empty, we know that $T \sim \exp(\mu)$. For the empty case, recalling that the arrival and service rates are exponentially distributed, the term can be rewritten as

$$\Pr[T \le t] = \rho \cdot \Pr[T \le t] + (1-\rho) \cdot \Pr[A + S \le t]\,,$$

where $A \sim \exp(\lambda)$ and $S \sim \exp(\mu)$ are random variables describing the arrival and service time of the next arriving job, respectively. By conditioning on S we obtain

$$\begin{aligned}\Pr[T \le t] &= \rho \cdot \Pr[T \le t] + (1-\rho) \cdot \int_0^t \Pr[A + S \le t \mid S = \tau] \cdot f_S(\tau)\,\mathrm{d}\tau \\ &= \rho \cdot \Pr[T \le t] + (1-\rho) \cdot \int_0^t \Pr[A \le t - \tau] \cdot f_S(\tau)\,\mathrm{d}\tau\,.\end{aligned}$$

Plugging in the probability density and distribution function and solving the integral yields

$$\Pr[T \le t] = \rho \cdot (1 - e^{-\mu t}) + (1-\rho) \cdot (1 - e^{-\mu t}) - (1-\rho) \cdot \mu \cdot \left(\frac{e^{-\lambda t} - e^{-\mu t}}{\lambda - \mu} \right).$$

By rearranging we get that $\Pr[T \le t] = 1 - e^{-\lambda t}$, which means that T is exponentially distributed with parameter λ, as desired. □

Remarks:

- Burke's theorem simplifies the analysis of $M/M/m$ queueing systems in the stationary case. Perhaps surprisingly, the departure process does not depend on the time it takes to serve a job, but just on the rate of arrivals.
- The stochastic process counting the number of arrivals or departures from a memoryless queuing system up to time t is a Poisson process.
- What about networks of queues?

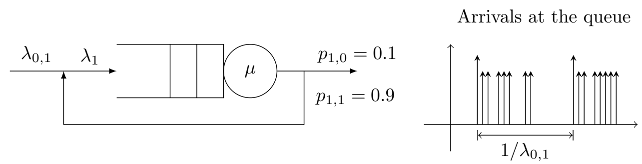

Figure 6.11: A queueing network exhibiting bursty behaviour. External jobs arrive with rate $\lambda_{0,1}$, and jobs leaving the queueing system immediately re-enter with probability $p_{1,1} = 0.9$. Thus, in expectation, each job enters the queue 10 times before leaving the system. The external arrivals at the queue, indicated as long arrows in the right graph, are a Poisson process. However, the total arrivals at the queue are not Poisson—after serving an external arrival the job is likely to loop back to the queue a few times.

Definition 6.16 (Queueing Network)**.** *A queueing network is a directed graph in which nodes represent queueing systems and edges direct jobs from one queueing system towards the next one. The network is **open** if external jobs arrive and depart the network, and **closed** if jobs never enter or leave the network.*

Remarks:

- In a closed network, the number of jobs is constant. For the following, we consider an open network containing $M/M/m$ queueing systems. Let us denote the number of queues (nodes) in the network by n.

- We assume that the external arrivals come from a Poisson distribution with some rate λ_0. The nodes in the graph are identified with positive integers, and an external arrival joins queueing system i with probability $p_{0,i}$, i.e., with rate $\lambda_{0,i} = \lambda_0 \cdot p_{0,i}$.

- The service rate of queueing system i is μ_i. After being served at queueing system i, a job leaves the system with probability $p_{i,0}$, and joins queueing system j with probability $p_{i,j}$.

- Due to Burke's theorem we know that in the stationary case the departures from a queueing system have the same distribution as the arrivals. We can thus compute the arrival rate λ_i at queue i by solving the linear equations

$$\lambda_i = \lambda_{0,i} + \sum_{j=1}^{n} \lambda_j \cdot p_{j,i} \,.$$

- The utilization ρ_i of a station is $\lambda_i/(m_i \cdot \mu_i)$, where m_i is the number of servers at the i^{th} queueing system.

Theorem 6.17 (Jackson's Theorem)**.** *Consider an open queueing network with n nodes where each node v_i, $i \in \{1, \ldots, n\}$, represents an $M/M/m_i$ queueing system. If all queues v_i are stable, then the steady state of the network is*

$$\pi(k_1, \ldots, k_n) = \prod_{i=1}^{n} \pi_i(k_i) \,.$$

Here $\pi(k_1, \ldots, k_n)$ denotes the stationary distribution for the network, i.e., the probability that k_i jobs are in queueing system i; and $\pi_i(k_i)$ is the probability that k_i jobs are in v_i when considering v_i as a single $M/M/m_i$ queue with arrival rate λ_i, i.e., the corresponding entry in v_i's stationary distribution.

Remarks:

- Jackson's Theorem allows us to compute the stationary distribution of an open queueing network containing memoryless queues. The distribution is obtained by computing the product of each queue's stationary distribution when considered in isolation (with arrival rate λ_i as above).

- Before applying the theorem, one needs to check that each queue is stable. This is done by computing the values λ_i and checking that each $\rho_i = \lambda_i/(m_i \cdot \mu_i) < 1$.

- Little's law also applies to networks of queueing systems as a whole.

- For closed networks the stationary distribution can be computed as follows.

Theorem 6.18 (Gordon, Newell)**.** *Consider a closed queueing network with total population K and n nodes, where each node v_i, $i \in \{1, \ldots, n\}$, represents an $M/M/m_i/n_i$ queue. If all queues v_i are stable, then the steady state of the network is*

$$\pi(k_1, \ldots, k_n) = \frac{1}{G(K)} \prod_{i=1}^{n} \rho_i^{k_i} \,,$$

where $G(K)$ is the normalizing constant

$$G(K) = \sum_{\substack{(k_1,\ldots,k_n) \\ k_i \le n_i, \sum k_i = K}} \prod_{i=1}^{n} \rho_i^{k_i} \,,$$

and the values ρ_i are obtained from the λ_i satisfying the equations

$$\lambda_i = \sum_{j=1}^{n} \lambda_j \cdot p_{i,j} \,.$$

Chapter Notes

The founder of queueing theory is Agner Karup Erlang (1878–1929), who wanted to understand how the telephone network needs to be dimensioned. He already described the stationary solutions to $M/M/m$ and $M/M/m/n$ queues, also referred to as Erlang C and Erlang B models, respectively, and was particularly interested in the probability that the system loses a call [1]. Since then many other kinds of queues were studied, and in 1953 Kendall introduced the notation described in Definition 6.8 to better categorize previous results [4].

For a long time Little's Law (Theorem 6.13) was believed to be true without a formal proof. In a book from 1958 Morse challenged his readers to find a counterexample [7], but Little found a proof for the statement instead [6]. A series of papers studied variants and extensions, thus widening the applicability of the law. Fifty years later Little summarized the progress in [5].

Jackson's Theorem for open networks (Theorem 6.17) was a first step in understanding networks of queues [3]. The stationary distribution for the closed network case (Theorem 6.18) was described by Gordon and Newell [2].

This chapter was written in collaboration with Jochen Seidel.

Bibliography

[1] Agner Karup Erlang. Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *Post Office Electrical Engineers' Journal*, 10:189–197, 1917.

[2] William J. Gordon and Gordon F. Newell. Closed queuing systems with exponential servers. *Operations Research*, 15(2):254–265, 1967.

[3] James R. Jackson. Jobshop-like queueing systems. *Management Science*, 10(1):131–142, November 1963.

[4] David G. Kendall. Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded markov chain. *Ann. Math. Statist.*, 24(3):338–354, 09 1953.

[5] John D. C. Little. OR FORUM—Little's Law as viewed on its 50th anniversary. *Operations Research*, 59(3):536–549, 2011.

[6] John D.C. Little. A PROOF FOR THE QUEUING FORMULA: $L = \lambda W$. *Operations Research*, 9(3):383–387, 1961.

[7] Philipp M. Morse. *Queues, inventories, and maintenance: the analysis of operational system with variable demand and supply.* Operations Research Society of America. Wiley, 1958.