Part II: Queueing Theory

4 Basic Queues and Little Formula

Queues are common in computer systems. There are queues of inquiries waiting to be processed by an interactive computer system, queue of data base requests, queues of I/O requests, etc. In computer networks, there are queues of packets waiting to be transmitted over a link, queues of queries in servers, queues of jobs in server farms, etc.

Typically a queue (or queueing system) has one service facility, although there may be more than one server in the service facility, and a waiting room (or buffer) of finite or infinite capacity.

Customers (from a population or *source*) enter a queueing system to receive some service. Here the word customer is used in its generic sense, and thus may be a packet in a communication network, a job or a program in a computer system, a request or an inquiry in a database system, etc.

Upon arrival a customer joins the waiting room if all servers in the service center are busy. When a customer has been served, he leaves the queueing system.

A special notation, called Kendall's notation, is used to describe a queueing system. The notation has the form

$$\text{A/B/c/K}$$

where

- A describes the *interarrival time* distribution
- B the *service time* distribution
- c the number of servers
- K the size of the system capacity (including the servers).

The symbols traditionally used for A and B are

- M for exponential distribution (M stands for Markov)
- D for deterministic distribution
- G (or GI) for general distribution.

Instances of service discipline are FIFO (first-in-first-out), LIFO (last-in-first-out), PS (processor-sharing).

When the system capacity is infinite ($K = \infty$) one simply uses the symbol $A/B/c$.

4.1 The M/M/1 queue

In this queueing system the customers arrive according to a Poisson process with rate λ. The time it takes to serve every customer is an exponential rv with parameter μ. We say that the customers have exponential service times. The service times are supposed to be mutually independent and further independent of the interarrival times.

When a customer enters an empty system his service starts at once; if the system is nonempty the incoming customer joins the queue. When a service completion occurs, a customer from the queue (we do not need to specify which one for the time being), if any, enters the service facility at once to get served.

Let $X(t)$ be the number of customers in the system at time t.

Proposition 12 (Birth and death process for an M/M/1 queue)**.** *The process $\{X(t),\, t \geq 0\}$ is a birth and death process with birth rate $\lambda_i = \lambda$ for all $i \geq 0$ and with death rate $\mu_i = \mu$ for all $i \geq 1$.* ■

This result follows from construction rule #2 (see Section 2). Indeed, in any state $i = 1, 2, \ldots$, there are two events that compete to drive the system out of i: an arrival of a new customer (birth) in which case the system will jump to state $i+1$ or a departure of a customer (death) in which case the system will jump to state $i-1$; if $i = 0$ then only an arrival will trigger a jump to state 1).

If the system is in state i ($i = 1, 2, \ldots$) at some time t, then the time to go until it leaves that state will be the minimum of two exponential rvs: the 1st rv gives the time to go until the next arrival after time t (this time is indeed distributed like an exponential rv since the interarrival times have been assumed to be exponentially distributed and that the exponential distribution is memoryless), the 2nd random variable gives the time to go until the next departure after time t (this time is also distributed like an exponential rv since we have assumed that the service times are exponentially distributed and because of the memoryless property of the exponential distribution). Since the minimum of two independent exponential rvs is an exponential rv, we see that the construction rule #2 applies to state $i = 1, 2, \ldots$. If the system is in state $i = 0$ at some time t then the time before the next event (here, necessarily an arrival) is exponentially distributed (again because the exponential distribution is memoryless and that the arrival process is Poisson). Note that the construction rule #1 may also be invoked for state 0.

We may therefore conclude from construction rule #2, that the process $\{X(t),\, t \geq 0\}$ is a Markov process. Since the only possible transitions out of state i are to enter state $i-1$ or state $i+1$ if $i = 1, 2, \ldots$ (state 1 if the system is in state 0) then we see that this Markov process is actually a birth and death process with state-space $\{0, 1, 2, \ldots\}$.

Let π_i, $i \geq 0$, be the probability distribution of the number of customers in the system in steady-state.

The balance equations for this birth and death process read

$$\begin{aligned} \lambda\, \pi_0 &= \mu\, \pi_1 \\ (\lambda + \mu)\, \pi_i &= \lambda\, \pi_{i-1} + \mu\, \pi_{i+1} \quad \forall i \geq 1. \end{aligned}$$

Define

$$\rho = \frac{\lambda}{\mu}. \tag{62}$$

The quantity ρ is referred to as the *system utilization* since it gives the fraction of time the system is busy.

A direct application of Proposition 7 yields:

Proposition 13 (Stationary queue-length distribution function of an M/M/1 queue)**.** *If $\rho < 1$ then*

$$\pi_i = (1-\rho)\,\rho^i \tag{63}$$

for all $i \geq 0$. ■

Therefore, the *stability condition* $\rho < 1$ simply says that the system is stable if the work that is brought to the system per unit of time is strictly smaller than the processing rate (which is 1 here since there is only one server).

Proposition 13 therefore says that the distribution function of the queue-length in steady-state is a *geometric distribution* with parameter ρ.

From (63) we can compute (in particular) the mean number of customers $E[X]$ (still in steady-state). We find

$$E[X] = \frac{\rho}{1-\rho}. \tag{64}$$

Observe that $E[X] \to \infty$ when $\rho \to 1$, so that, in practice if the system is not stable, then the queue will explode. It is also worth observing that the queue will empty infinitely many times when the system is stable since $\pi_0 = 1 - \rho > 0$.

We may also be interested in the probability that the queue exceeds, say, K customers, in steady-state. From (63) we have

$$P(X \geq K) = \rho^K. \tag{65}$$

What is the throughput T of an M/M/1 in equilibrium? The answer should be $T = \lambda$. Let us check this guess.

We have

$$T = (1 - \pi_0)\,\mu.$$

Since $\pi_0 = 1 - \rho$ from (63) we see that $T = \lambda$ by definition of ρ.

Burke's theorem

Suppose an M/M/1 system starts in stead-state. Then: (i) the departure process is Poisson(λ); and (ii) the queue size at time t is independent of the sequence of departures prior to time t.

The proof of this theorem relies on the time-reversibility property of the M/M/1 queue. The departures in the forwards chain correspond to the arrivals in the reverse chain. Since the reverse chain is also an M/M/1 system with Poisson arrivals with rate λ and the sequence of arrivals *after* time t does not depend on the queue size at time t, then the theorem is true.

4.2 The M/M/1/K queue

In practice, queues are always finite. In that case, a new customer is lost when he finds the system full (e.g., telephone calls).

The M/M/1/K may accommodate at most K customers, including the customer in the service facility, if any. Let λ and μ be the rate of the Poisson process for the arrivals and the parameter of the exponential distribution for the service times, respectively.

Let π_i, $i = 0, 1, \ldots, K$, be the distribution function of the queue-length in steady-state. The balance equations for this birth and death process read

$$\begin{aligned} \lambda\,\pi_0 &= \mu\,\pi_1 \\ (\lambda + \mu)\,\pi_i &= \lambda\,\pi_{i-1} + \mu\,\pi_{i+1} \quad \text{for } i = 1, 2 \ldots, K-1 \\ \lambda\,\pi_{K-1} &= \mu\,\pi_K. \end{aligned}$$

Proposition 14 (Stationary queue-length distribution function in an M/M/1/K queue)**.** *If $\rho \neq 1$ then*

$$\pi_i = \frac{(1-\rho)\,\rho^i}{1-\rho^{K+1}} \tag{66}$$

for $i = 0, 1, \ldots, K$, and $\pi_i = 0$ for $i > K$.

If $\rho = 1$ then

$$\pi_i = \frac{1}{K+1} \tag{67}$$

for $i = 0, 1, \ldots, K$, and $\pi_i = 0$ for $i > K$. ■

Here again the proof of Proposition 14 relies on the fact that $\{X(t),\, t \geq 0\}$, where $X(t) \in \{0, 1, \ldots, K\}$ is the number of customers in the system at time t, can be modeled as a birth and death process with birth rates $\lambda_i = \lambda$ for $i = 0, 1, \ldots, K-1$ and $\lambda_i = 0$ for $i \geq K$. The proof that $\{X(t),\, t \geq 0\}$ is Markov process is analogous to the proof given above for the M/M/1 queue (the only difference is that here we only need to focus on states $i = 0, 1, \ldots, K$; note that from state $i = K$ the system may only go to state $K - 1$).

In particular, the probability that an incoming customer is rejected is π_K.

4.3 The M/M/c queue

There are $c \geq 1$ servers and the waiting room has infinite capacity. If more than one server is available when a new customer arrives (which necessarily implies that the waiting room is empty) then the incoming customer may enter any of the free servers.

Let λ and μ be the rate of the Poisson process for the arrivals and the parameter of the exponential distribution for the service times, respectively. The system utilization is defined as

$$\rho = \frac{\lambda}{c\mu}.$$

Here again the process $\{X(t),\, t \geq 0\}$ of the number of customers in the system can be modeled as a birth and death process (use construction rule #2). The birth rate is $\lambda_i = \lambda$ when $i \geq 0$. The death rate is given by

$$\begin{aligned} \mu_i &= i\mu \quad \text{for } i = 1, 2, \ldots, c-1 \\ &= c\mu \quad \text{for } i \geq c \end{aligned}$$

which can be also written as $\mu_i = \mu \min(i, c)$ for all $i \geq 1$.

Using these values of λ_i and μ_i in Proposition 7 yields

Proposition 15 (Stationary queue-length distribution function in an M/M/c queue)**.** *If $\rho < 1$ then*

$$\pi_i = \begin{cases} \pi_0 \left(\dfrac{\lambda}{\mu}\right)^i \dfrac{1}{i!} & \text{if } i = 0, 1, \ldots, c \\ \\ \pi_0 \left(\dfrac{\lambda}{\mu}\right)^i \dfrac{1}{c!} \dfrac{1}{c^{i-c}} & \text{if } i \geq c \end{cases} \tag{68}$$

where

$$\pi_0 = \left[\sum_{i=0}^{c-1} \left(\frac{\lambda}{\mu}\right)^i \frac{1}{i!} + \left(\frac{\lambda}{\mu}\right)^c \frac{1}{c!} \left(\frac{1}{1-\rho}\right)\right]^{-1}. \tag{69}$$

■

The probability that an arriving customer is forced to join the queue is given by

$$\begin{aligned} P(\text{queueing}) &= \sum_{i=c}^{\infty} \pi_i = \sum_{i=c}^{\infty} \pi_0 \left(\frac{\lambda}{\mu}\right)^i \frac{1}{c!} \frac{1}{c^{i-c}} \\ &= \frac{\left(\dfrac{\lambda}{\mu}\right)^c \dfrac{1}{c!} \left(\dfrac{1}{1-\rho}\right)}{\sum_{i=0}^{c-1} \left(\dfrac{\lambda}{\mu}\right)^i \dfrac{1}{i!} + \left(\dfrac{\lambda}{\mu}\right)^c \dfrac{1}{c!} \left(\dfrac{1}{1-\rho}\right)}. \end{aligned} \tag{70}$$

This probability is of wide use in telephony and gives the probability that no trunk (i.e., server) is available for an arriving call (i.e., customer) in a system of c trunks. It is referred to as *Erlang's C formula*.

Observe that Burke's theorem holds also for an M/M/c system.

4.4 The M/M/c/c queue

Here we have a situation when there are $c \geq 1$ available servers but no waiting room. This is a pure *loss queueing system*. Each newly arriving customer is given its private server; however, if a customer arrives when all the servers are occupied, that customer is lost. Parameters λ and μ are defined as in the previous sections.

The number of busy servers can be modelled as a birth and death process (use construction rule #2) with birth rate

$$\lambda_i = \begin{cases} \lambda & \text{if } i = 0, 1, \ldots, c-1 \\ \\ 0 & \text{if } i \geq c \end{cases}$$

and death rate $\mu_i = i\mu$ for $i = 1, 2, \ldots, c$.

We are interested in determining the limiting distribution function π_i $(i = 0, 1, \ldots, c)$ of the number of busy servers.

Proposition 16 (Stationary server occupation in an M/M/c/c queue)**.**

$$\pi_i = \pi_0 \left(\frac{\lambda}{\mu}\right)^i \frac{1}{i!} \tag{71}$$

for $i = 0, 1, \ldots, c$*,* $\pi_i = 0$ *for* $i > c$*, where*

$$\pi_0 = \left[\sum_{i=0}^{c} \left(\frac{\lambda}{\mu}\right)^i \frac{1}{i!}\right]^{-1}. \tag{72}$$

■

This system is also of great interest in telephony. In particular, π_c gives the probability that all trunks (i.e., servers) are busy, and it is given by

$$\pi_c = \frac{\left(\frac{\lambda}{\mu}\right)^c \frac{1}{c!}}{\sum_{j=0}^{c} \left(\frac{\lambda}{\mu}\right)^j \frac{1}{j!}}. \tag{73}$$

This is the celebrated *Erlang's loss formula* (derived by Agner Krarup Erlang in 1917).

Remarkably enough Proposition 16 is valid for *any service time distribution* and not only for exponential service times! Such a property is called an *insensitivity property.*

Later on, we will see an extremely useful extension of this model to (in particular) several classes of customers, that has nice applications in the modeling and performance evaluation of multimedia networks.

4.5 The repairperson model

It is one of the most useful models. There are K machines and a single repairperson. Each machine breaks down after a time that is exponentially distributed with parameter α. In other words, α is the rate at which a machine breaks.

When a breakdown occurs, a request is sent to the repairperson for fixing it. Requests are buffered. It takes an exponentially distributed amount of time with parameter μ for the repairperson to repair a machine. In other words, μ is the repair rate.

We assume that "lifetimes" and repair times are all mutually independent.

What is the probability π_i that i machines are up (i.e., working properly)? What is the overall failure rate?

Let $X(t)$ be the number of machines up at time t. It is easily seen that $\{X(t),\, t \geq 0\}$ is a birth and death process with birth and death rates given by $\lambda_n = \mu$ for $n = 0, 1, \ldots, K-1$, $\lambda_n = 0$ for $n \geq K$ and $\mu_n = n\alpha$ for $n = 1, 2, \ldots, K$, respectively.

We notice that $\{X(t),\, t \geq 0\}$ has the same behavior as the queue-length process of an $M/M/K/K$ queue!

Hence, by (71) and (72) we find that

$$\pi_i = \frac{(\mu/\alpha)^i/i!}{C(K, \mu/\alpha)}$$

for $i = 0, 1, \ldots, K$, where

$$C(K, a) := \sum_{i=0}^{K} \frac{a^i}{i!}. \tag{74}$$

The overall failure rate λ_b is given by

$$\lambda_b = \sum_{i=1}^{K} \alpha\, i\, \pi_i = \alpha\, \frac{\sum_{i=1}^{K} i\, (\mu/\alpha)^i/i!}{C(K, \mu/\alpha)} = \mu\, \frac{\sum_{i=0}^{K-1} (\mu/\alpha)^i/i!}{C(K, \mu/\alpha)} = \mu\, \frac{C(K-1, \mu/\alpha)}{C(K, \mu/\alpha)}.$$

Observe that $\pi_0 = 1/C(K, \mu/\alpha)$. Hence, the mean number n_r of machines repaired by unit of time is

$$n_r = \mu\,(1 - \pi_K) = \mu \left(1 - \frac{(\mu/\alpha)^K/K!}{C(K, \mu/\alpha)}\right) = \mu \left(\frac{C(K, \mu/\alpha) - (\mu/\alpha)^K/K!}{C(K, \mu/\alpha)}\right) = \mu\, \frac{C(K-1, \mu/\alpha)}{C(K, \mu/\alpha)}.$$

4.6 Little's formula

So far we have only obtained results for the buffer occupation namely, the limiting distribution of the queue-length, the mean number of customers, etc. These performance measures are of particular interest for a system's manager. What we would like to do now is to address performance issues from a user's perspective, such as, for instance, response times and waiting times.

For this, we need to introduce the most used formula in performance evaluation.

Proposition 17 (Little's formula)**.** *Let $\lambda > 0$ be the arrival rate of customers to a queueing system in steady-state. Let $\overline{N}$ be the mean number of customers in the system and let $\overline{T}$ be the mean sojourn time of a customer (i.e., the sum of its waiting time and of its service time).*

Then,

$$\overline{N} = \lambda\, \overline{T}\,. \tag{75}$$

■

This formula is of great interest since very often one knows $\overline{N}$ and λ. It states that the average number of customers in a queueing system in steady-state is equal to the arrival rate of customers to that system, times the average time spent in that system. This result does not make any specific assumption regarding the arrival distribution or the service time distribution; nor does it depend upon the number of servers in the system or upon the particular queueing discipline within the system.

This result has an intuitive explanation: $\overline{N}/\overline{T}$ is the departure rate, which has to be equal to the input rate λ since the system is in steady-state.

We now give a proof of Little's formula in the case where the system empties infinitely often.

Proof. Starting from an empty system, let $C > 0$ be a time when the system is empty (we assume that the system is not always empty in $(0, C)$). Let $k(C)$ be the number of customers that have been served in $(0, C)$. In the following we set $k = k(C)$ for ease of notation. Let a_i be the arrival time of the i-th customer, and let d_i be the departure time of the i-th customer, $i = 1, 2, \ldots, k$.

These dates form an increasing sequence of times $(t_n)_{n=1}^{2k}$ such that

$$a_1 = t_1 < t_2 < \cdots < t_{2k-1} < t_{2k} = d_k.$$

The mean sojourn time $\overline{T}$ of a customer in $(0, C)$ is by definition

$$\overline{T} = \frac{1}{k} \sum_{i=1}^{k} (d_i - a_i)$$

since $d_i - a_i$ is the time spent in the system by the i-th customer.

Let us now compute $\overline{N}$, the mean number of customers in the system in $(0, C)$. Denote by $N(t)$ the number of customers at time t. Then,

$$\overline{N} \;=\; \frac{1}{C} \int_0^C N(t)\, dt \;=\; \frac{1}{C} \sum_{i=1}^{2k-1} N(t_i^+)\,(t_{i+1} - t_i)$$

where $N(t^+)$ is the number of customers in the system *just after* time t.

It is not difficult to see (draw the queue size $N(t)$ and compute the area under it) that

$$\underbrace{\sum_{i=1}^{k} (d_i - a_i)}_{\text{sum of horizontal blocks}} \qquad = \qquad \underbrace{\sum_{i=1}^{2k-1} N(t_i^+)\,(t_{i+1} - t_i)}_{\text{sum of vertical blocks}}.$$

Hence,

$$\overline{N} = \frac{k}{C}\,\overline{T}.$$

The proof is concluded as follows: since the system empties infinitely often we can choose C large enough so that k/C is equal to the arrival rate λ. Hence, $\overline{N} = \lambda\,\overline{T}$. ★

Example 5. Consider an M/M/1 queue with arrival rate λ and service rate μ. Let $\overline{T}$ (resp. $\overline{W}$) be the mean customer sojourn time, also referred to as the mean customer response time (resp. waiting time).

If $\rho := \lambda/\mu < 1$ (i.e., if the queue is stable) then we know that the mean number of customers N is given by $\overline{N} = \rho/(1-\rho)$ (see 64).

Therefore, by Little's formula,

$$\overline{T} = \frac{\rho}{\lambda\,(1-\rho)} = \frac{1}{\mu\,(1-\rho)} = \frac{1}{\mu-\lambda} \tag{76}$$

$$\overline{W} = \overline{T} - \frac{1}{\mu} = \frac{\rho}{\mu\,(1-\rho)}\,. \tag{77}$$

Observe that both $\overline{T} \to \infty$ and $\overline{W} \to \infty$ when $\rho \to 1$. ♦

Example 6 (Comparing different multiprocessor systems)**.** While designing a multiprocessor system we may wish to compare different systems.

The first system is an M/M/2 queue with arrival rate 2λ and service rate μ.

The second system is an M/M/1 queue with arrival rate 2λ and service rate 2μ.

Note that the comparison is fair since in both systems the utilization is λ/μ.

Which system yields the smallest expected customer response time?

Let $\overline{T}_1$ and $\overline{T}_2$ be the expected customer response time in systems 1 and 2, respectively.

Computation of $\overline{T}_1$:

Denote $\overline{N_1}$ by the mean number of customers in the system. From (68) and (69) we get that

$$\pi_i = 2\,\pi_0\,\rho^i \quad \forall i \geq 1,$$

if $\rho < 1$ (stability condition), from which we deduce that

$$\pi_0 = \frac{1-\rho}{1+\rho}.$$

Thus, for $\rho < 1$,

$$\overline{N}_1 = \sum_{i=1}^{\infty} i\,\pi_i = 2\left(\frac{1-\rho}{1+\rho}\right)\sum_{i=1}^{\infty} i\,\rho^i = \frac{2\rho}{(1-\rho)\,(1+\rho)}$$

by using the well-known identity $\sum_{i=1}^{\infty} i\,z^{i-1} = 1/(1-z)^2$ for all $0 \leq z < 1$.

From Little's formula we deduce that

$$\overline{T}_1 = \frac{\rho}{\lambda\,(1-\rho)\,(1+\rho)}$$

under the stability condition $\rho < 1$.

Computation of $\overline{T}_2$:

For the M/M/1 queue with arrival rate 2λ and service rate 2μ we have already seen in Example 5 that

$$\overline{T}_2 = \frac{\rho}{2\lambda\,(1-\rho)}$$

under the stability condition $\rho < 1$.

It is easily seen that $\overline{T_2} < \overline{T_1}$ when $\rho < 1$.