Chapter 8 Queueing Models

A queueing model consists "customers" arriving to receive some service and then depart. The mechanisms involved are

- input mechanism: the arrival pattern of customers in time
- queueing mechanism: the number of servers, order of the service
- service mechanism: the time to serve one or a batch of customers

We consider queueing models that follow the most common rule of service: **first come, first served.**

Common Queueing Processes

It is often reasonable to assume

- the interarrival times of customers are i.i.d. (the arrival of customers follows a renewal process),
- the service times for customers are i.i.d. and are independent of the arrival of customers.

Notation: M = memoryless, or Markov, G = General

- $M/M/1$: Poisson arrival, service time $\sim Exp(\mu)$, 1 server = a birth and death process with birth rates $\lambda_j \equiv \lambda$, and death rates $\mu_j \equiv \mu$
- $M/M/\infty$: Poisson arrival, service time $\sim Exp(\mu)$, ∞ servers = a birth and death process with birth rates $\lambda_j \equiv \lambda$, and death rates $\mu_j \equiv j\mu$
- $M/M/k$: Poisson arrival, service time $\sim Exp(\mu)$, k servers = a birth and death process with birth rates $\lambda_j \equiv \lambda$, and death rates $\mu_j \equiv \min(j, k)\mu$

Common Queueing Processes (Cont'd)

- $M/G/1$: Poisson arrival, General service times $\sim G$, 1 server
- $M/G/\infty$: Poisson arrival, General service time $\sim G$, ∞ servers
- $M/G/k$: Poisson arrival, General service times $\sim G$, k servers
- $G/M/1$: General interarrival times, service times $\sim Exp(\mu)$, 1 server
- $G/G/k$: General interarrival times $\sim F$, General service times $\sim G$, k servers
- ...

Quantities of Interest for Queueing Models

Let

$$X(t) = \# \text{ of customers in the system at time } t$$
$$Q(t) = \# \text{ of customers waitng in queue at time } t$$

Assume that $\{X(t), t \geq 0\}$ and $\{Q(t), t \geq 0\}$ has a stationary distribution.

$$L = \lim_{t\to\infty} \frac{\int_0^t X(t)dt}{t} = \text{the average } \# \text{ of customers in the system}$$
$$L_Q = \lim_{t\to\infty} \frac{\int_0^t Q(t)dt}{t} = \text{the average } \# \text{ of customers waiting in queue}$$
$$W = \text{the average amount of time, including waiting time}$$
$$\text{and service time, a customer spends in the system;}$$
$$W_Q = \text{the average amount of time a customer waiting in queue.}$$

Little's Formula

Let

$$N(t) = \# \text{ of customers enter the system at or before time } t.$$

We define λ_a be the arrival rate of entering customers,

$$\lambda_a = \lim_{t\to\infty} \frac{N(t)}{t}$$

Little's Formula:

$$L = \lambda_a W$$
$$L_Q = \lambda_a W_Q$$

Cost Identity

Many interesting and useful relationships between quantities in Queueing models can be obtained by using the **cost identity**.

Imagine that entering customers are forced to pay money (according to some rule) to the system. We would then have the following basic cost identity:

$$\begin{aligned} &\text{average rate at which the system earns} \\ &= \lambda_a \times \text{average amount an entering customer pays} \end{aligned}$$

Proof. Let $R(t)$ be the amount of money the system has earned by time t. Then we have

$$\begin{aligned} &\text{average rate at which the system earns} \\ &= \lim_{t\to\infty} \frac{R(t)}{t} = \lim_{t\to\infty} \frac{N(t)}{t} \frac{R(t)}{N(t)} = \lambda_a \lim_{t\to\infty} \frac{R(t)}{N(t)} \\ &= \lambda_a \times \text{average amount an entering customer pays,} \end{aligned}$$

provided that the limits exist.

Proof of Little's Formula

To prove $L = \lambda_a W$:

- we use the payment rule:

$$\boxed{\text{each customer pays \$1 per unit time while in the system.}}$$

- the average amount a customer pay $= W$, the average waiting time of customers.
- the amount of money the system earns during the time interval $(t, t + \Delta t)$ is $X(t)\Delta t$, where $X(t)$ is the number of customers in the system at time t ,
- and the rate the system earns is thus $\lim_{t\to\infty} \dfrac{\int_0^t X(s)ds}{t} = L$, the formula follows from the cost identity.

To prove $L_Q = \lambda_a W_Q$, we use the payment rule:

$$\boxed{\text{each customer pays \$1 per unit time while in queue.}}$$

The argument is similar.

8.3.1 M/M/1 Model

Let $X(t)$ be number of customers in the system at time t.
$\{X(t), t \geq 0\}$ is a birth and death process with

$$\text{birth rates } \lambda_j \equiv \lambda, \quad \text{and death rates } \mu_j \equiv \mu.$$

Recall in Example 6.14 we have showed that the stationary distribution exists when $\lambda < \mu$, and the stationary distribution is

$$P_n = \lim_{t \to \infty} \mathrm{P}(X(t) = n) = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n, \quad n = 0, 1, \ldots$$

Thus

$$L = \lim_{t \to \infty} \mathbb{E}[X(t)] = \sum_{n=1}^{\infty} nP_n = \frac{\lambda}{\mu - \lambda} = \frac{1/\mu}{1/\lambda - 1/\mu}$$

$$= \frac{\mathbb{E}[\text{service time}]}{\mathbb{E}[\text{interarrival time}] - \mathbb{E}[\text{service time}]}$$

8.3.1 M/M/1 Model (Cont'd)

Let T be the time of a customer spend in the system.
If there are n customers in the system while this customer arrives, then T is the sum of the service times of the $n+1$ customers $\sim$ *Gamma*$(n+1,\mu)$. That is,

$$
\begin{aligned}
\mathrm{P}(T \le t) &= \sum_{n=0}^{\infty} P_n \int_0^t \frac{\mu^{n+1}}{n!} s^n e^{-\mu s} ds \\
&= \sum_{n=0}^{\infty} \left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n \int_0^t \frac{\mu^{n+1}}{n!} s^n e^{-\mu s} ds \\
&= (\mu-\lambda)\int_0^t \Bigg(\underbrace{\sum_{n=0}^{\infty}\frac{(\lambda s)^n}{n!}}_{=e^{\lambda s}}\Bigg) e^{-\mu s} ds \\
&= (\mu-\lambda)\int_0^t e^{-(\mu-\lambda)s} ds = 1-e^{-(\mu-\lambda)t}
\end{aligned}
$$

Therefore, $T \sim Exp(\mu-\lambda) \quad \Rightarrow \quad W = \mathbb{E}[T] = \dfrac{1}{\mu-\lambda}$.

This verifies Little's formula, $L = \lambda W$.

8.3.1 M/M/1 Model (Cont'd)

$$W_Q = W - \mathbb{E}[\text{service time}] = W - 1/\mu = \frac{\lambda}{\mu(\mu - \lambda)}$$

Note that

$$\#\text{ of customers in queue} = \max(0, \#\text{ of customers in system} - 1).$$

So

$$\begin{aligned} L_Q = \sum_{n=1}^{\infty}(n-1)P_n &= \underbrace{\sum_{n=1}^{\infty} nP_n}_{L} - (\underbrace{\sum_{n=1}^{\infty} P_n}_{1-P_0}) \\ &= L - 1 + P_0 \\ &= \frac{\lambda}{\mu - \lambda} - 1 + \left(1 - \frac{\lambda}{\mu}\right) \\ &= \frac{\lambda^2}{\mu(\mu - \lambda)} = \lambda W_Q \end{aligned}$$

Example 8.2

Suppose customers arrive at a Poisson rate of 1 in 12 minutes, and that the service time is exponential at a rate of one service per 8 minutes. What are L and W?
Solution. Since $\lambda = 1/12$, $\mu = 1/8$, we have

$$L = \frac{1/\mu}{1/\lambda - 1/\mu} = \frac{8}{12 - 8} = 2, \; W = \frac{1}{\mu - \lambda} = 24$$

Observe if the arrival rate increases 20% to $\lambda = 1/10$, then

$$L = 4, W = 40$$

When $\lambda/\mu \approx 1$, a slight increase in λ/μ will lead to a large increase in L and W.

$M/M/\infty$ Model

In this case, customers will be served immediately upon arrival. Nobody will be in queue. We have

$$W_Q = L_Q = 0, \quad W = \text{average service time} = 1/\mu,$$

and hence $L = \lambda W = \lambda/\mu$.

As a verification, observe that $\{X(t), t \geq 0\}$ is a birth and death process with

$$\text{birth rates } \lambda_j \equiv \lambda, \quad \text{and death rates } \mu_j \equiv j\mu.$$

The stationary distribution is

$$P_n = \frac{\lambda^n}{n!\mu^n} P_0 = \frac{\lambda^n}{n!\mu^n} \frac{1}{\sum_{n=0}^{\infty} \frac{\lambda^n}{n!\mu^n}} = e^{-\lambda/\mu} \frac{(\lambda/\mu)^n}{n!}, \quad n = 0, 1, \ldots$$

Therefore $X(t) \sim Poisson(\lambda/\mu)$ as $t \to \infty$,

$$L = \mathbb{E}[X(t)] = \lambda/\mu.$$

Birth & Death Queueing Models

In addition to $M/M/1$ and $M/M/\infty$ models, a more general family of birth & death queueing models is the following:

$M/M/k$ **Queueing System with Balking**
Consider a $M/M/k$ system, but suppose a customer arrives finding n others in the system will only join the system with probability α_n, i.e., he balks (walks away) w/ prob. $1 - \alpha_n$. This system is a birth and death process with

$$\lambda_n = \lambda\alpha_n, \quad n \geq 0$$
$$\mu_n = \min(n, k)\mu, \quad n \geq 1$$

A special case of $M/M/k$ queueing system with balking is the $M/M/k$ system with finite capacity N, where

$$\alpha_n = \begin{cases} 1 & \text{if } n < N \\ 0 & \text{if } n \geq N \end{cases}$$

Birth & Death Queueing Models

For a birth & death queueing model, the stationary distribution of the number of customers in the system is given by

$$P_k = \lim_{t\to\infty} \mathrm{P}(X(t) = k) = \frac{\lambda_0\lambda_1\cdots\lambda_{k-1}/(\mu_1\mu_2\cdots\mu_k)}{1+\sum_{n=1}^{\infty}\frac{\lambda_0\lambda_1\cdots\lambda_{n-1}}{\mu_1\mu_2\cdots\mu_n}}, \quad k \geq 1$$

The necessary and sufficient condition for such a stationary distribution to exists is that

$$\sum_{n=1}^{\infty}\frac{\lambda_0\lambda_1\cdots\lambda_{n-1}}{\mu_1\mu_2\cdots\mu_n} < \infty.$$

With $\{P_n\}$, the average number of customers in the system is simply

$$L = \sum_{n=0}^{\infty} nP_n.$$

Birth & Death Queueing Models (Cont'd)

With balking, the rate that customers enter the system is not λ (since not all customers enter the system), but

$$\lambda_a = \sum_{n=0}^{\infty} \lambda_n P_n.$$

Consequently, the average waiting time is

$$W = L/\lambda_a = \frac{\sum_{n=0}^{\infty} nP_n}{\sum_{n=0}^{\infty} \lambda_n P_n},$$

and the average amount of time waiting in queue (W_Q) and average number of customers in queue (L_Q) are respectively

$$W_Q = W - \mathbb{E}[\text{service time}] = W - (1/\mu),$$
$$L_Q = \lambda_a W_Q$$

Busy Period in a Birth & Death Queueing Model

There is a alternating renewal process embedded in a birth & death queueing model.

We say a renewal occurs if the system become empty.

Using the alternating renewal theory, the long-run proportion of time that the system is empty is $\dfrac{\mathbb{E}[\text{Idle}]}{\mathbb{E}[\text{Idle}] + \mathbb{E}[\text{Busy}]}$, where

$$\mathbb{E}[\text{Idle}] = \text{expected length of an idle period}$$
$$\mathbb{E}[\text{Busy}] = \text{expected length of a busy period}$$

Also note that the long-run proportion of time that the system is empty is simply $P_0 = \lim_{t\to\infty} \mathrm{P}(X(t) = 0)$. Since the length of an idle period $\sim Exp(\lambda_0)$, we have $\mathbb{E}[\text{Idle}] = 1/\lambda_0$. In summary, we have that

$$P_0 = \frac{1/\lambda_0}{(1/\lambda_0) + \mathbb{E}[\text{Busy}]}$$

or

$$\mathbb{E}[\text{Busy}] = \frac{1 - P_0}{\lambda_0 P_0}$$