Queueing Models

Cong Ma

University of Chicago, Winter 2026

Outline

Outline

1. Queueing Basics
2. $M/M/1$ and Birth-Death Models
3. PASTA and Finite Capacity

Queueing Models

A queueing model describes "customers" arriving to receive service and then departing. 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 service rule: **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.

Kendall's notation

M = memoryless (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

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

Assume that $\{X(t), t \geq 0\}$ and $\{Q(t), t \geq 0\}$ have 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 = the average amount of time a customer spends in the system (including both waiting and service time);

W_Q = the average amount of time a customer waits in queue.

Little's Formula

Let

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

We define λ_a to 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:

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

Cost Identity (Cont'd)

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:

each customer pays \$1 per unit time while in the system.

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

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

each customer pays \$1 per unit time while in queue.

The argument is similar.

Outline

$M/M/\infty$ **Model**

In this case, customers are served immediately upon arrival. Nobody waits 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 $t \to \infty$,

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

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 that (see Example 6.14 in the book) we have shown 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

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

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

From the previous slide:

$$L = \frac{1/\mu}{1/\lambda - 1/\mu} = \frac{\mathbb{E}[\text{service time}]}{\mathbb{E}[\text{interarrival time}] - \mathbb{E}[\text{service time}]}.$$

As $\lambda \uparrow \mu$, the denominator goes to 0, so L grows quickly.

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

Let T be the time a customer spends 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,

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

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

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

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

From the previous slide,

$$\mathrm{P}(T \leq t) = 1 - e^{-(\mu-\lambda)t},$$

so $T \sim Exp(\mu - \lambda)$. Therefore

$$W = \mathbb{E}[T] = \frac{1}{\mu - \lambda},$$

which verifies Little's formula:

$$L = \lambda W = \frac{\lambda}{\mu - \lambda}.$$

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 that if the arrival rate increases by 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.

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, and suppose a customer who finds n others in the system joins with probability α_n (i.e., balks with probability $1 - \alpha_n$). This system is a birth and death process with

$$\begin{aligned} \lambda_n &= \lambda \alpha_n, \quad n \geq 0 \\ \mu_n &= \min(n, k)\mu, \quad n \geq 1 \end{aligned}$$

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 exist 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

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

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

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

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 an alternating renewal process embedded in a birth & death queueing model.
We say a renewal occurs if the system becomes empty.
Using the alternating renewal theory, the long-run proportion of time that the system is empty is $\frac{\mathbb{E}[\text{Idle}]}{\mathbb{E}[\text{Idle}] + \mathbb{E}[\text{Busy}]}$, where

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

Busy Period in a Birth & Death Queueing Model (Cont'd)

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

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

Outline

8.2.2. Steady-State Probabilities

For a general queueing model, we are interested in three different limiting probabilities:

$$P_n = \lim_{t\to\infty} \mathrm{P}(X(t) = n),$$
$$\text{where } X(t) = \#\text{ of customers in the system at time } t$$

$a_n =$ proportion of customers arrive finding n in the system

$d_n =$ proportion of customers depart leaving n behind in the system

Here we assume they exist.

8.2.2. Steady-State Probabilities (Cont'd)

Though the three are defined differently, the latter two are identical in most queueing models.

Proposition 8.1. In any system where customers arrive and depart one at a time,

$$\text{rate at which arrivals find } n = \text{rate at which departures leave } n,$$

and therefore

$$a_n = d_n.$$

Proof of Proposition 8.1

Let

$$\begin{aligned} N_{i,j}(t) &= \text{number of times the number of customers in the system} \\ &\quad\ \text{goes from } i \text{ to } j \text{ by time } t \\ A(t) &= \text{number of customers arrived by time } t \\ D(t) &= \text{number of customers departed by time } t \end{aligned}$$

Proof of Proposition 8.1 (Cont'd)

An arrival sees n whenever the system moves from n to $n+1$; a departure leaves behind n whenever the system moves from $n+1$ to n. Hence

$$\text{rate at which arrivals find } n = \lim_{t\to\infty} \frac{N_{n,n+1}(t)}{t}$$
$$\text{rate at which departures leave } n = \lim_{t\to\infty} \frac{N_{n+1,n}(t)}{t}$$
$$a_n = \lim_{t\to\infty} \frac{N_{n,n+1}(t)}{A(t)}, \quad d_n = \lim_{t\to\infty} \frac{N_{n+1,n}(t)}{D(t)}.$$

Since between any two transitions from n to $n+1$, there must be one from $n+1$ to n, and vice versa, we have

$$N_{n,n+1}(t) = N_{n+1,n}(t) \pm 1 \quad \text{for all } t.$$

Proof of Proposition 8.1 (Cont'd)

Thus

$$
\begin{aligned}
\text{rate at which arrivals find } n &= \lim_{t\to\infty} \frac{N_{n,n+1}(t)}{t} \\
&= \lim_{t\to\infty} \frac{N_{n+1,n}(t) \pm 1}{t} \\
&= \text{rate at which departures leave } n
\end{aligned}
$$

Proof of Proposition 8.1 (Cont'd)

For a_n and d_n, obviously $A(t) \geq D(t)$ and hence

$$\lim_{t\to\infty} \frac{A(t)}{t} \geq \lim_{t\to\infty} \frac{D(t)}{t}$$

Combining with the fact $\lim_{t\to\infty} \frac{N_{n,n+1}(t)}{t} = \lim_{t\to\infty} \frac{N_{n+1,n}(t)}{t}$ we just shown, we obtain

$$a_n = \lim_{t\to\infty} \frac{N_{n,n+1}(t)}{A(t)} \leq \lim_{t\to\infty} \frac{N_{n+1,n}(t)}{D(t)} = d_n$$

There are two possibilities:

- if $\lim\limits_{t\to\infty} A(t)/t = \lim\limits_{t\to\infty} D(t)/t$, then obviously $a_n = d_n$ for all n
- if $\lim\limits_{t\to\infty} A(t)/t > \lim\limits_{t\to\infty} D(t)/t$, then the queue size will go to infinity, implying that $a_n = d_n = 0$. The equality is still valid.

Proof of Proposition 8.1 (Cont'd)

For a_n and d_n, obviously $A(t) \geq D(t)$ and hence

$$\lim_{t\to\infty} \frac{A(t)}{t} \geq \lim_{t\to\infty} \frac{D(t)}{t}$$

Combining with the fact $\lim_{t\to\infty} \frac{N_{n,n+1}(t)}{t} = \lim_{t\to\infty} \frac{N_{n+1,n}(t)}{t}$ we just shown, we obtain

$$a_n = \lim_{t\to\infty} \frac{N_{n,n+1}(t)\underline{\ /t\ }}{A(t)\underline{\ /t\ }} \leq \lim_{t\to\infty} \frac{N_{n+1,n}(t)\underline{\ /t\ }}{D(t)\underline{\ /t\ }} = d_n$$

There are two possibilities:

- if $\lim_{t\to\infty} A(t)/t = \lim_{t\to\infty} D(t)/t$, then obviously $a_n = d_n$ for all n
- if $\lim_{t\to\infty} A(t)/t > \lim_{t\to\infty} D(t)/t$, then the queue size will go to infinity, implying that $a_n = d_n = 0$. The equality is still valid.

Example 8.1

Here is an example where $P_n \neq a_n$. Consider a queueing model in which

- service times $= 1$, always
- interarrival times are always > 1 [e.g., Uniform(1.5,2)].

Hence, as every arrival finds the system empty and every departure leaves it empty, we have

$$a_0 = d_0 = 1$$

However, $P_0 \neq 1$ as the system is not always empty of customers.

PASTA

Proposition 8.2 (PASTA Principle). If the arrival process is Poisson, then $P_n = a_n$, and hence $P_n = d_n$.

<u>P</u>oisson <u>A</u>rrivals <u>S</u>ee <u>T</u>ime <u>A</u>verages

- By time T, the total time with n customers in the system is approximately $P_n T$
- Regardless of how many customers in the system, Poisson arrivals always arrive at rate λ. Thus by time T, the total number of arrivals that find n in the system is $\approx \lambda P_n T$.
- The overall number of customers that arrive by time T is $\approx \lambda T$.
- The proportion of arrivals that find the system in state n is

$$a_n = \frac{\lambda P_n T}{\lambda T} = P_n$$

Example 5.5 (M/M/1 Queueing w/ Finite Capacity)

- A single-server service station with i.i.d. service times $\sim Exp(\mu)$
- Poisson arrival of customers with rate λ
- Upon arrival, a customer would
  - go into service if the server is free (queue length $= 0$)
  - join the queue if 1 to $N-1$ customers in the station, or
  - walk away if N or more customers in the station

Question: What fraction of potential customers are lost?

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

$\{X(t),\ t \geq 0\}$ is a birth-death process with the birth and death rates below

$$\mu_n = \begin{cases} 0 & \text{if } n = 0 \\ \mu & \text{if } n \geq 1 \end{cases} \quad \text{and} \quad \lambda_n = \begin{cases} \lambda & \text{if } 0 \leq n < N \\ 0 & \text{if } n \geq N \end{cases}$$

Example 5.5 (M/M/1 Queueing w/ Finite Capacity)

Solving $\lambda_n P_n = \mu_{n+1} P_{n+1}$ for the limiting distribution

$$\begin{aligned}
P_1 &= (\lambda/\mu) P_0 \\
P_2 &= (\lambda/\mu) P_1 = (\lambda/\mu)^2 P_0 \\
&\vdots \\
P_i &= (\lambda/\mu)^i P_0, \qquad\qquad i = 1, 2, \ldots, N
\end{aligned}$$

Plugging $P_i = (\lambda/\mu)^i P_0$ into $\sum_{i=0}^{N} P_i = 1$, one can solve for P_0 and get

$$P_i = \frac{1 - \lambda/\mu}{1 - (\lambda/\mu)^{N+1}} (\lambda/\mu)^i$$

Answer: The fraction of customers lost is $P_N = \frac{1-\lambda/\mu}{1-(\lambda/\mu)^{N+1}} (\lambda/\mu)^N$