Queueing theory

Index

Queueing systems

Queueing system:

(i) customers arrive in some random manner at service facility;
(ii) upon arrival, they might have to wait some time in queue until it is their turn to be served;
(iii) once served they leave the system.

Quantities of interest:

- L, the average number of customers in the system;
- L_Q, the average number of customers waiting in queue;
- W, the average time a customer spends in the system;
- W_Q, the average time a customer spends waiting in queue.

Cost equations

Basic cost equation. If customers pay money to the system according to some rule, then

$$\text{average rate at which the system earns} \\ = \lambda_a \times \text{average amount a customer pays},$$

where λ_a is the average arrival rate of the customers. Note that if $N(t)$ denotes the number of arrivals by time t, then $\lambda_a = \lim_{t\to\infty} \frac{N(t)}{t}$.

More equations:

- 1 SEK per unit time while in the system: $L = \lambda_a W$ (Little's formula).
- 1 SEK per unit time while in the queue: $L_Q = \lambda_a W_Q$.
- 1 SEK per unit time while in service (with service time S):

$$\text{average number of customers in service} = \lambda_a \mathbb{E}[S].$$

The models

Types of **queue**:

- The $M/M/1$ queue. Poisson (memoryless) arrivals, exponential (memoryless) service times, 1 server.
- The $M/G/1$ queue. Poisson (memoryless) arrivals, general service times, 1 server.
- The $M/M/k$ queue. Poisson (memoryless) arrivals, exponential (memoryless) service times, k servers.

Types of **service**:

- FCFS - first come, first served.
- LCFS - last come, first served: arriving customers move to the front of the queue, or start service immediately and one of the customers in service moves back to the front of the queue.
- PS - processor sharing: the capacity of the server is equally shared between the customers, which then don't have to wait at all.

Analyzing the models

- Determine the limiting probabilities $P_n = \mathbb{P}(L = n)$, for $n = 0, 1, \ldots,$ representing the long-run probabilities that the system contains exactly n customers.

  Condition for the limiting probabilities to exist:

  $$\text{mean departure time} < \text{mean arrival time},$$

  otherwise the queue size increases to infinity.

- Determine the quantities L, L_Q, W, W_Q.

Index

The $M/M/1$ queue

The $M/M/1$ **queue**:

(i) customers arrive according to a Poisson process with rate λ;

(ii) FCFS policy, i.e., if the server is free a customer goes directly into service, otherwise it joins the end of the queue; service times are i.i.d. exponential r.v.'s, $S \sim \mathrm{Exp}(\mu)$.

(iii) a customer leaves the system immediately after being served and the first in the queue (if any) enters service.

Condition: $\lambda < \mu$, i.e., the arrival rate must be smaller than the service rate.

Limiting probabilities

- Continuous-time Markov chain, **balance equations**:

$$\lambda P_0 = \mu P_1,$$
$$(\lambda + \mu) P_n = \lambda P_{n-1} + \mu P_{n+1}, \qquad n \geq 1.$$

- Rewriting, $P_1 = \frac{\lambda}{\mu} P_0$ and $P_{n+1} = \frac{\lambda}{\mu} P_n + \left(P_n - \frac{\lambda}{\mu} P_{n-1}\right)$ for $n \geq 1$. Solving in terms of P_0, for $n \geq 2$, $P_n = \frac{\lambda}{\mu} P_{n-1} + \left(P_{n-1} - \frac{\lambda}{\mu} P_{n-2}\right) = \frac{\lambda}{\mu} P_{n-1} = \left(\frac{\lambda}{\mu}\right)^n P_0$.
- Since $1 = \sum_{n=0}^{\infty} P_n = \sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^n P_0 = \frac{P_0}{1-\frac{\lambda}{\mu}}$, we get the **limiting probabilities**

$$P_0 = 1 - \frac{\lambda}{\mu},$$
$$P_n = \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right), \qquad n \geq 1.$$

Note the necessary condition $\lambda < \mu$.

Quantities of interest

- $L = \sum_{n=0}^{\infty} nP_n = \sum_{n=0}^{\infty} n\left(\frac{\lambda}{\mu}\right)^n\left(1-\frac{\lambda}{\mu}\right) = \frac{\lambda}{\mu-\lambda}$ (using the identity $\sum_{n=0}^{\infty} nx^n = \frac{x}{(1-x)^2}$).
- $W = \frac{L}{\lambda} = \frac{1}{\mu-\lambda}$.
- $W_Q = W - \mathbb{E}[S] = W - \frac{1}{\mu} = \frac{\lambda}{\mu(\mu-\lambda)}$.
- $L_Q = \lambda W_Q = \frac{\lambda^2}{\mu(\mu-\lambda)}$. Note that $L_Q \neq L-1$, but $L_Q = L - \frac{\lambda}{\mu}$.

Example 8.3.

Time spent in the system

- The amount of **time that a customer spends in the system** is

$$W^* \sim \mathrm{Exp}(\mu - \lambda).$$

Proof. Given the number of customers already in the system when our customer arrives $N = n$, W^* is distributed as the sum of $n+1$ i.i.d. exponential r.v.'s with rate μ, i.e., as $\Gamma(n+1, \mu)$. Write

$$\begin{aligned} f_{N|W^*=t}(n) &= \frac{f_{N,W^*}(n,t)}{f_{W^*}(t)} = \frac{f_N(n) f_{W^*|N=n}(t)}{f_{W^*}(t)} \\ &= \frac{\left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right) \mu e^{-\mu t} \frac{(\mu t)^n}{n!}}{f_{W^*}(t)} = K \frac{(\lambda t)^n}{n!}, \end{aligned}$$

with $K = \frac{(\mu - \lambda) e^{-\mu t}}{f_{W^*}(t)}$. Since
$1 = \sum_{n=0}^{\infty} f_{N|W^*=t}(n) = K \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!} = K e^{\lambda t}$, we get $K = e^{-\lambda t}$. Hence,

$$f_{W^*}(t) = \frac{(\mu - \lambda) e^{-\mu t}}{K} = (\mu - \lambda) e^{-(\mu - \lambda) t}. \qquad \square$$

The next arrival

Example 8.4.

Starting from a stationary state, what is the probability that the **next arrival** finds $N_a = n$ customers in the system?

Inspection paradox: It is not $P_n = \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right)$. Indeed, the time from the current time t to the next arrival is $\mathrm{Exp}(\lambda)$, as well as the time from t to the last arrival. Hence, time between the the last and the next arrivals is $\Gamma(2, \lambda)$. We will see that:

- $\mathbb{E}[N_a] < L$, i.e., the average number of customers seen by the next arrival is less than the average number of customers in the system;
- $\mathbb{P}(N_a = 0) > P_0$, i.e., the next arrival is more likely to find an empty system than is an average arrival.

Conditioning on the number of customers currently in the system X,

$$\begin{aligned}
\mathbb{P}(N_a = n) &= \sum_{k=0}^{\infty} \mathbb{P}(N_a = n \,|\, X = k)\mathbb{P}(X = k) \\
&= \sum_{k=0}^{\infty} \mathbb{P}(N_a = n \,|\, X = k)\left(\frac{\lambda}{\mu}\right)^k \left(1 - \frac{\lambda}{\mu}\right) \\
&= \sum_{k=n}^{\infty} \mathbb{P}(N_a = n \,|\, X = k)\left(\frac{\lambda}{\mu}\right)^k \left(1 - \frac{\lambda}{\mu}\right) \\
&= \sum_{i=0}^{\infty} \mathbb{P}(N_a = n \,|\, X = n + i)\left(\frac{\lambda}{\mu}\right)^{n+i} \left(1 - \frac{\lambda}{\mu}\right).
\end{aligned}$$

For $n > 0$, the next arrival finds n customers if there are i services before the arrival and then the arrival before the next service, i.e.,

$$\begin{aligned}\mathbb{P}(N_a = n) &= \sum_{i=0}^{\infty}\left(\frac{\mu}{\lambda+\mu}\right)^i \frac{\lambda}{\lambda+\mu}\left(\frac{\lambda}{\mu}\right)^{n+i}\left(1-\frac{\lambda}{\mu}\right) \\ &= \left(\frac{\lambda}{\mu}\right)^n\left(1-\frac{\lambda}{\mu}\right)\frac{\lambda}{\lambda+\mu}\sum_{i=0}^{\infty}\left(\frac{\lambda}{\lambda+\mu}\right)^i = \left(\frac{\lambda}{\mu}\right)^{n+1}\left(1-\frac{\lambda}{\mu}\right).\end{aligned}$$

Note that $\mathbb{E}[N_a] = \sum_{n=0}^{\infty} n\mathbb{P}(N_a = n) = \frac{\lambda}{\mu}\sum_{n=1}^{\infty} nP_n = \frac{\lambda}{\mu}L < L$.

For $n = 0$, the next arrival finds the system empy if there are i services before the arrival, i.e.,

$$\begin{aligned}\mathbb{P}(N_a = 0) &= \sum_{i=0}^{\infty}\left(\frac{\mu}{\lambda+\mu}\right)^i\left(\frac{\lambda}{\mu}\right)^i\left(1-\frac{\lambda}{\mu}\right) \\ &= \left(1-\frac{\lambda}{\mu}\right)\sum_{i=0}^{\infty}\left(\frac{\lambda}{\lambda+\mu}\right)^i = \left(1-\frac{\lambda}{\mu}\right)\left(1+\frac{\lambda}{\mu}\right).\end{aligned}$$

Note that $\mathbb{P}(N_a = 0) > P_0 = 1 - \frac{\lambda}{\mu}$.

The $M/M/1$ queue with capacity K

There can be **no more than K customers** in the system at any time.

- Continuous-time Markov chain, balance equations:

$$\begin{aligned} \lambda P_0 &= \mu P_1, \\ (\lambda+\mu)P_n &= \lambda P_{n-1} + \mu P_{n+1}, \qquad 1 \le n \le K-1, \\ \mu P_K &= \lambda P_{K-1}. \end{aligned}$$

- Rewriting in terms of P_0, we obtain $P_1 = \frac{\lambda}{\mu} P_0$ and $P_n = \left(\frac{\lambda}{\mu}\right)^n P_0$.
- Using $\sum_{n=0}^{K} P_n = 1$, we get the limiting probabilities

$$P_n = \frac{\left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right)}{1 - \left(\frac{\lambda}{\mu}\right)^{K+1}}, \qquad n = 0, 1, \ldots, K.$$

No need to impose the condition $\lambda < \mu$, since the queue size is bounded and it cannot increase indefinitely.

The $M/M/k$ queue

If any of the k **servers** are free, a customer goes directly into service, otherwise it joins the end of the queue.

- If there are $n = 1, \ldots, k$ customers in the system, the time until a departure is the minimum of n i.i.d. exponentials with rate μ, hence it will be exponential with rate $n\mu$.
- Continuous-time Markov chain, balance equations:

$$\begin{aligned} \lambda P_0 &= \mu P_1, & \\ (\lambda + n\mu)P_n &= \lambda P_{n-1} + (n+1)\mu P_{n+1}, \qquad & n < k, \\ (\lambda + k\mu)P_n &= \lambda P_{n-1} + k\mu P_{n+1}, \qquad & n \geq k. \end{aligned}$$

- The limiting probabilities are given in Example 8.6. Note the necessary condition $\lambda < k\mu$.

Birth and death queueing models

Birth and death queueing model: arrival rate λ_n and departure rate μ_n depend on the number of customers in the system $n = 0, 1, \ldots$.

Examples:

- $M/M/1$: $\lambda_n = \lambda$ for $n \geq 0$ and $\mu_n = \mu$ for $n \geq 1$.
- $M/M/1$ with capacity K: $\lambda_n = \begin{cases} \lambda, & n < K, \\ 0, & n \geq K, \end{cases}$ and $\mu_n = \mu$ for $n \geq 1$.
- $M/M/k$: $\lambda_n = \lambda$ for $n \geq 0$ and $\mu_n = \begin{cases} n\mu, & n \leq k, \\ k\mu, & n \geq k. \end{cases}$

Example 8.7: $M/M/1$ with impatient customers.

Idle and busy periods in birth and death models

The system alternates between **idle periods** when there are no customers and **busy periods** in which there is at least one customer.

- Idle periods are i.i.d. $I \sim \mathrm{Exp}(\lambda_0)$, hence $\mathbb{E}[I] = \frac{1}{\lambda_0}$.
- Busy periods are also i.i.d., hence the long-run proportion of time in which the system is empy is

$$P_0 = \frac{\mathbb{E}[I]}{\mathbb{E}[I] + \mathbb{E}[B]} = \frac{1}{1 + \lambda_0 \mathbb{E}[B]},$$

which gives

$$\mathbb{E}[B] = \frac{1 - P_0}{\lambda_0 P_0}.$$

Note that in the $M/M/1$ queue, we get $\mathbb{E}[B] = \frac{\frac{\lambda}{\mu}}{\lambda\left(1 - \frac{\lambda}{\mu}\right)} = \frac{1}{\mu - \lambda}$.

Index

Steady-state probabilities

- If $X(t)$ is the number of customers in the system at time t, the **limiting probabilities**

$$P_n = \lim_{t\to\infty} \mathbb{P}(X(t) = n) = \mathbb{P}(L = n), \qquad n \geq 0,$$

  represent the proportion of time that the system contains exactly n customers.
- Let a_n be the **proportion of arriving customers** that find n customers already in the system.
- Let d_n be **proportion of departing customers** that leave behind n other customers in the system.

They are not always equal: *Example 8.1* with $a_0 = d_0 = 1 \neq P_0$.

Arrivals and departures see the same number of customers

Theorem

The rate at which arrivals find n customers equals the rate at which departures leave n customers, and

$$a_n = d_n, \qquad n \geq 0.$$

Proof. Note that,

$$a_n = \frac{\text{rate at which arrivals find } n \text{ customers}}{\text{overall arrival rate}}$$

and

$$d_n = \frac{\text{rate at which departures leave } n \text{ customers}}{\text{overall departure rate}}.$$

Since the number of transitions from n to $n+1$ must equal to within 1 the number of transitions from $n+1$ to n, the numerators are equal. If the denominators are equal, then $a_n = d_n$, and note that the result holds even if they are not (special cases). □

The PASTA principle

On average, arrivals and departures always see the same number of customers, but they do not in general see time averages (Example 8.1).

Theorem (The PASTA principle)

Poisson Arrivals See Time Averages. In particular, $P_n = a_n$.

Proof 1. Since the Poisson process has independent increments, knowing that an arrival occurs at time t gives us no information about what occurred prior to t. Hence, an arrival would just see the system according to the limiting probabilities, i.e., $a_n = P_n$. $\square$

Proof 2. The total time the system has exactly n customers by time T is $P_n T$. Then the number of arrivals in $[0, T]$ that find n customers is $\lambda P_n T$. In the long-run (as $T \to \infty$), the rate at which arrivals find n customers is λP_n. Since λ is the overall arrival rate, it follows that the proportion of arrivals that find n customers is $a_n = \frac{\lambda P_n}{\lambda} = P_n$. $\square$

Example

Example 8.2: people at a bus stop.

Index

Work and cost equations

Define the **work** in the system at any time t as the sum of the remaining service times of all customers in the system at time t. Let V denote the average work in the system.

- Each customer pays at a rate of y per unit time when his remaining service time is y, whether he is in the queue or in service. In other words, the rate at which the system earns is the work in the system.
- Recall the basic cost equation:
  $\mathbb{E}[\text{rate at which the system earns}] = \lambda_a \mathbb{E}[\text{amount paid by a customer}]$.
- A customer pays at rate S per time unit while he is in queue and at rate $S - x$ after being in service for time x. Hence, if W^* is the time a given customer spends in queue,

$$V = \lambda_a \mathbb{E}\left[S W_Q^* + \int_0^S (S - x)\, dx \right] = \lambda_a \mathbb{E}[S W_Q^*] + \frac{\lambda_a \mathbb{E}[S^2]}{2}.$$

- If the service time is independent of the waiting time of customers in the queue, then $V = \lambda_a \mathbb{E}[S] W_Q + \frac{\lambda_a \mathbb{E}[S^2]}{2}$.

Quantities of interest

- The time a customer waits in the queue equals the work that he sees in the system when he arrives, since there is only a single server.

  Taking expectations, W_Q equals the average work seen by an arrival, which, due to Poisson arrivals, equals the average work in the system. Hence, $W_Q = V$.

  Pollaczek-Khintchine formula:

$$W_Q = \frac{\lambda \mathbb{E}[S^2]}{2(1 - \lambda \mathbb{E}[S])}.$$

- The other quantities of interest can be obtained:

$$L_Q = \lambda W_Q, \qquad W = W_Q + \mathbb{E}[S], \qquad L = \lambda W.$$

- Note the necessary condition $\lambda < \frac{1}{\mathbb{E}[S]}$, i.e., the arrival rate must be smaller than the service rate.

Idle and busy periods

- **Idle periods** are i.i.d. $I \sim \mathrm{Exp}(\lambda)$, hence $\mathbb{E}[I] = \frac{1}{\lambda}$.
- **Busy periods** are also i.i.d., hence the long-run proportion of time in which the system is empty is

$$P_0 = \frac{\mathbb{E}[I]}{\mathbb{E}[I] + \mathbb{E}[B]} = \frac{1}{1 + \lambda\mathbb{E}[B]},$$

which gives

$$\mathbb{E}[B] = \frac{1 - P_0}{\lambda P_0}.$$

Since the average number of customers in service is $\lambda\mathbb{E}[S]$ (cost equation) and is also $0 \cdot P_0 + 1 \cdot (1 - P_0) = 1 - P_0$, we get $P_0 = 1 - \lambda\mathbb{E}[S]$. Hence,

$$\mathbb{E}[B] = \frac{\mathbb{E}[S]}{1 - \lambda\mathbb{E}[S]}.$$

Number of customers in a busy period

What is the **number of customers C served in a busy period**?

- On average, for every $\mathbb{E}[C]$ arrivals exactly one will find the system empty (the first one), hence $a_0 = \frac{1}{\mathbb{E}[C]}$. Since $a_0 = P_0 = 1 - \lambda\mathbb{E}[S]$, we get
$$\mathbb{E}[C] = \frac{1}{1 - \lambda\mathbb{E}[S]}.$$
- Recall the RRT: if $R(t)$ is the reward earned by time t, then
$$\lim_{t\to\infty} \frac{\mathbb{E}[R(t)]}{t} = \frac{\mathbb{E}[\text{reward earned during a cycle}]}{\mathbb{E}[\text{length of a cycle}]}.$$
If the reward for a customer is 1, then we get
$$\lambda = \frac{\mathbb{E}[\text{customers per cycle}]}{\mathbb{E}[\text{length of a cycle}]} = \frac{\mathbb{E}[C]}{\frac{1}{\lambda} + \mathbb{E}[B]}.$$
Hence,
$$\mathbb{E}[C] = \lambda\mathbb{E}[B] + 1 = \frac{1}{1 - \lambda\mathbb{E}[S]}.$$

Exercises

Session 6. Chapter 8: 1, 6, 8 (do part c before part b), 12a-b.

Session 7. Chapter 8: 23 (for questions c,d,e, express the answer in P_S's, where S is the state, but you do not have to compute the P_S's), 28, 36, 37, 40.