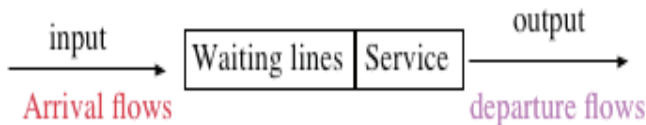


# Queueing systems

# Introduction

**Queueing theory** is the mathematical study of the operation of stochastic systems describing processing of flows of jobs. Queues occur when current demand for service exceeds the capacity of the service facility



# Arrivals

- ▶ We use the terminology 'customers', but they could be telephone calls, computer jobs, information packets, etc.
- ▶ Arrival times  $T_1, T_2, T_3, \dots$ . The **inter-arrival times** are  $\tau_1 = T_1 - T_0, \tau_2 = T_2 - T_1, \tau_3 = T_3 - T_2 \dots$
- ▶ The inter-arrival times are assumed to be i.i.d.
- ▶ Alternatively, we could use the counting process  $N_t$  giving the number of arrivals in  $[0, t]$ ,  $t \geq 0$ .

# Service

- ▶ There is a total of  $m$  spaces for both receiving service and waiting for it.
- ▶ If there is an idle server, an arriving customer is serviced immediately.
- ▶ The service time  $S_i^{(j)}$  of the  $i$ th customer at the  $j$ th server is a random variable.
- ▶ The service times are assumed to be i.i.d.
- ▶ When a server is serving a customer, it cannot provide any service to other customers.
- ▶ If all servers are busy, then the arriving customers join a queue if there is enough space, otherwise, the customer is rejected.

# Service Disciplines

This could be

- ▶ FIFO: First In - First Out (FCFS: First Come - First Served).
- ▶ Last Come - First Served (with or without pre-emption).
- ▶ Processor Sharing.
- ▶ Priority (with or without pre-emption).
- ▶ more complicated disciplines?

We will consider only FIFO in this course.

We can use such queueing systems to construct **queueing networks** by forwarding customers departing from one queue to other queues.

## Quantities of interest

$X_t$  is the number of customers in the system at time  $t$  (including those in service and those waiting to begin service)

waiting times: length of time a customer spends in the queue before her/his service commences.

sojourn times: total length of time spent in the system (waiting time plus service time).

# Kendall's notation

This was devised by David Kendall in 1953. It takes the form  $A/B/n/m$  where

- ▶  $A$  describes the arrival process
  - ▶  $A = M$  (Markov) inter-arrival times are exponentially-distributed.
  - ▶  $A = GI$  or  $(G)$  inter-arrival times have some arbitrary distribution.
  - ▶  $A = D$  inter-arrival times are deterministic.

# Kendall's notation

- ▶  $B$  describes the service process
  - ▶  $B = M$  service times are exponentially-distributed.
  - ▶  $B = GI$  or  $(G)$  service times have some arbitrary distribution.
  - ▶  $B = D$  service times are deterministic.
- ▶  $n$  gives the number of servers.
- ▶  $m$  gives the capacity of the system. When  $m = \infty$ , this is usually omitted.



# Questions

- ▶ Does a queueing system have a steady-state regime or does the queue increase unboundedly?
- ▶ What is the steady-state queue length distribution if it exists?
- ▶ What is the steady-state waiting time distribution if it exists?
- ▶ What is the average load on the server?
- ▶ What fraction of time is the server idle?

## $M/M/1$ queue

- ▶ Arrival stream: Poisson process with intensity  $\lambda$
- ▶ Service:  $n = 1$  server, service times  $\sim \exp(\mu)$
- ▶ Infinite space for waiting:  $m = \infty$
- ▶ The state  $X_t$  gives the number of customers at time  $t$ :
  - ▶ If  $X_t = 0$  the server is idle.
  - ▶ If  $X_t = k \geq 1$  one customer is being served and  $k - 1$  customers are waiting in the queue.

This is a CTMC (in fact a birth and death process) with non-zero transition rates  $q_{i,i+1} = \lambda$  and  $q_{i+1,i} = \mu$  for all  $i \in \mathbb{Z}_+$ .

Exercise: draw the transition diagram for this CTMC

## $M/M/1$ an interpretation

If  $X_t = 0$ , the process remains at 0 for an  $\exp(\lambda)$  time  $\tau_+$  until a new customer arrives, so  $X_{t+\tau_+} = 1$ .

If  $X_t = k > 0$ , the process remains at  $k$  for a time  $\tau = \min(\tau_+, \tau_-)$  where

- ▶  $\tau_+ \sim \exp(\lambda)$  is the time until the next arrival after  $t$
- ▶  $\tau_- \sim \exp(\mu)$  is the time until the end of service of the customer in service at  $t$ .

$X_{t+\tau} = k + 1$  if  $\tau_+ < \tau_-$  and  $X_{t+\tau} = k - 1$  if  $\tau_+ > \tau_-$ .

## $M/M/1$ stationary distribution

Using our results from CTMCs, we see that a stationary distribution for  $(X_t)_{t \geq 0}$  exists if (and only if) the chain is positive recurrent. This is equivalent to  $\rho \equiv \lambda/\mu < 1$ , in which case, for  $n \geq 1$ ,

$$\pi_n = (\lambda/\mu)^n \pi_0.$$

Using the normalisation condition  $\sum_{i=0}^{\infty} \pi_i = 1$ , we see that

$$\pi_0 \sum_{i=0}^{\infty} (\lambda/\mu)^i = 1$$

which tells us that

$$\pi_0 = 1 - \rho$$

and, for  $n \geq 0$ ,

$$\pi_n = (1 - \rho) \rho^n.$$

So the stationary distribution for the number of customers in the system is geometric $^*(1 - \rho)$ . (Note that this geometric takes values in  $\mathbb{Z}_+$ ).

## $M/M/1$ further questions

- ▶ What is the stationary expected number  $\ell$  of customers in the whole system?
- ▶ What is the stationary expected number  $\ell_q$  of customers in just the queue?
- ▶ What is the expected waiting time of a customer in stationarity?
- ▶ What is the distribution of the waiting time?

The first two quantities can be calculated from the stationary distribution. We might guess that  $\ell_q = \ell - 1$ . However this is not right because the queue might be empty.

## $M/M/1$ waiting times in stationarity

Assume that an  $M/M/1$  queue is operating under a FCFS discipline.

- ▶ In the stationary regime, a tagged arriving customer will find a random number  $N$  of customers where  $N \sim (\pi_k)_{k \in \mathbb{Z}_+}$  (PASTA).
- ▶ If  $N = 0$ , then the customer will go straight into service.
- ▶ If  $N > 0$ , the remaining service time  $S_1$  for the customer being served  $\sim \exp(\mu)$ .
- ▶ The service times  $S_2, S_3, \dots, S_N$ , for those in the queue are independent  $\exp(\mu)$  random variables, also independent of  $N$ .
- ▶ So the waiting time for our tagged customer is  $W = \sum_{j=1}^N S_j$ , where we interpret the empty sum as equal to 0.

## $M/M/1$ waiting times in stationarity

The distribution of a non-negative random variable  $Y$  is characterized by its Laplace transform  $M_Y(-s) = \mathbb{E}[e^{-sY}]$  for  $s > 0$ .

We can write

$$\begin{aligned}\mathbb{E}_\pi[e^{-sW}] &= \mathbb{E}_\pi[\mathbb{E}_\pi[e^{-sW} | N]] = \mathbb{E}_\pi[\mathbb{E}_\pi[e^{-s \sum_{j=1}^N S_j} | N]] \\&= \mathbb{E}_\pi[(\mathbb{E}_\pi[e^{-sS_1}])^N] \\&= \mathbb{E}_\pi \left[ \left( \frac{\mu}{s + \mu} \right)^N \right] = M_N(\log(\mu/(s + \mu))) \\&= (1 - \rho) \sum_{n=0}^{\infty} \rho^n \left( \frac{\mu}{s + \mu} \right)^n \\&= \frac{(1 - \rho)(s + \mu)}{s + \mu - \lambda} \\&= (1 - \rho) + \rho \frac{\mu - \lambda}{s + \mu - \lambda},\end{aligned}$$

and we see that the distribution of  $W$  is a mixture of a 0-random variable and an exponential( $\mu - \lambda$ ) random variable. To be precise,

$$\mathbb{P}(W = 0) = 1 - \rho, \quad \mathbb{P}(W > x) = \rho e^{-x(1-\rho)\mu}, \quad \text{for } x > 0.$$

## $M/M/1$ waiting times in stationarity

It follows that the expected waiting time is

$$\mathbb{E}[W] = \frac{\rho}{\mu - \lambda}.$$

Once we have the expected waiting time, we can calculate the expected total time  $d$  in the system via the formula

$$d = \mathbb{E}[W] + \frac{1}{\mu} = \frac{1}{\mu - \lambda}.$$



## Little's law:

Recall that  $\ell$  is the expected number of customers in the system, while  $\ell_q$  is the expected number of customers waiting for service (both at stationarity).

Little's law says that

$$\ell = \lambda d,$$

and

$$\ell_q = \lambda \mathbb{E}[W].$$

## Sketch proof of Little's law

Let  $N_t$  and  $D_t$  denote the number of customers who have *entered* and *departed* from the system in  $[0, t]$  respectively. So the number in the system at time  $t$  is  $X_t = N_t - D_t$ . Denoting the area under the function  $X_s$  for  $s \leq t$  by  $A_t$ , we calculate  $\mathbb{E}[A_t/t]$  in two different ways. First

$$\mathbb{E}\left[\frac{A_t}{t}\right] = \mathbb{E}\left[\frac{1}{t} \int_0^t X_u du\right]$$

which approaches the average number  $\ell$  in the system as  $t \rightarrow \infty$ .

## Sketch proof of Little's law

Second, we have (approximately),

$$\frac{A_t}{t} = \frac{1}{t} \sum_{n=1}^{N_t} D_n$$

where  $D_i$  is the time spent in the system by the  $i$ th customer.

## Sketch proof of Little's law

Now

$$\begin{aligned}\mathbb{E}\left[\frac{A_t}{t}\right] &= \frac{1}{t}\mathbb{E}\left[\sum_{n=1}^{N_t} D_n\right] \\ &= \frac{1}{t}\mathbb{E}\left[\mathbb{E}\left[\sum_{n=1}^{N_t} D_n \middle| N_t\right]\right] \\ &= \frac{1}{t}\mathbb{E}[N_t d] \\ &= \frac{\lambda t d}{t} \\ &= \lambda d.\end{aligned}$$

So we have  $\ell = \lambda d$ .

## Example

A repairperson is assigned to service a bank of machines in a shop. Assume that failure times occur according to a Poisson process with rate  $\lambda = 1/12$  per minute and the repair rate is  $\mu = 1/8$  per minute.

## Example

- ▶ The traffic intensity is  $\rho = 2/3 < 1$ , so a stationary distribution exists.
- ▶ For  $k \geq 0$ , the stationary distribution is  $\pi_k = (1 - \rho)\rho^k$ .
- ▶ The repairperson is idle with prob  $1 - \rho = 1/3$ .
- ▶ The expected number of machines under repair is  $\ell = \rho/(1 - \rho) = 2$ .
- ▶ The expected time under repair is  $d = \ell/\lambda$  (or  $1/(\mu - \lambda)$ ) = 24 minutes.
- ▶ The expected time waiting for repair is  $\mathbb{E}[W] = d - 1/\mu = 16$  minutes.
- ▶ Let  $T_q$  be the time that a machine has to wait before being repaired. Then  $\mathbb{P}(T_q > 10) = \rho e^{-(\mu - \lambda)10} = 0.44$ .

## Example

- ▶ Suppose that the failure rate of machines increases (e.g. due to aging) by 16% to  $\lambda' = 1/10$ , then the new traffic intensity is  $\rho' = 4/5$ , and  $\ell' = 4$  with  $d' = 40$  and  $\mathbb{E}[W'] = 32$ .
- ▶ A 16% increase in arrival time rate has drastically increased the expected number of failed machines and doubled the time that they have to wait before getting repaired.
- ▶ We see that, when  $\rho$  is close to 1, the effect of small changes of  $\rho$  is profound: if a queueing system has long waiting times and lines, a rather modest increase in the service rate can bring about a dramatic reduction in waiting times.

## Costs example

- ▶ Suppose there is a new piece of equipment that will increase the repair rate from  $\mu = 1/8$  to  $\mu^* = 1/6$ , that is, decrease the expected repair time from 8 minutes to 6 minutes.
- ▶ The increase in maintenance cost for the new equipment is  $c_M = \$10$  per minute.
- ▶ The cost of lost production when a machine is out of order is  $c_D = \$5$  per minute.
- ▶ Should we purchase the new equipment?



## Costs example solution

Without the new equipment

- ▶ The expected number of failed machines is  $\ell = \rho/(1 - \rho) = 2$ .
- ▶ The expected cost of lost production is  $\ell c_D = \$10$  per minute.
- ▶ With the new equipment,  $\rho^* = 0.5$ , and the expected cost is  $\ell^* c_D + c_M = \$15$  per minute.
- ▶ We should buy the equipment if  $\ell^* c_D + c_M < \ell c_D$ , so we should not buy the equipment.

## Another costs example

At a service station the rate of service is  $\mu$  cars per hour, and the rate of arrivals of cars is  $\lambda$  per hour. The cost incurred by the service station due to delaying cars is  $\$c_1$  per car per hour and the operating and service costs are  $\$\mu c_2$  per hour. The rate of service  $\mu$  is a control parameter. Determine the value of  $\mu$  so that the least expected cost is achieved and find the value of the latter.

## Another costs example solution

- ▶ If there are  $Y$  cars in the service station, the cost is  $\$c_1 Y + \mu c_2$  per hour.
- ▶ In the stationary regime,  $\mathbb{E}[Y] = \rho/(1 - \rho)$ .
- ▶ The expected total cost per hour is  $\$c(\mu) = c_1 \rho/(1 - \rho) + \mu c_2$
- ▶ To find the minimum, we find  $\mu$  such that  $c'(\mu) = 0$  and since  $\mu > \lambda$ , we have a solution  $\mu_0 = \lambda + \sqrt{\lambda c_1 / c_2}$ .
- ▶ We can check that  $\mu_0$  achieves the minimum and  $c(\mu_0) = \lambda c_2 + 2\sqrt{\lambda c_1 c_2}$ .

## M/M/a Queue

This system has the following properties

- ▶  $a \geq 1$  servers,
- ▶ a Poisson arrival process with rate  $\lambda$ ,
- ▶ a FIFO service discipline,
- ▶ independent  $\exp(\mu)$  service times,
- ▶ when an arrival finds more than one idle server, it chooses one at random,
- ▶ when  $k$  servers are working, the total service rate is  $k\mu$ .

## $M/M/a$ Queue

The transition rates are  $q_{i,i+1} = \lambda$ , for  $i \geq 0$  and  $q_{i,i-1} = \mu \times \min(a, i)$  for  $i \geq 1$ .

Exercise: draw the transition diagram

This is a birth-and-death process with  $\nu_i = \lambda$  for  $i = 0, 1, 2, \dots$  and  $\mu_i = i\mu$  for  $i = 1, 2, \dots, a$  and  $\mu_i = a\mu$  for  $i > a$ .

## $M/M/a$ ergodicity

$$\kappa_j \equiv \frac{\nu_0 \cdots \lambda_{j-1}}{\mu_1 \cdots \mu_j} = \begin{cases} (\lambda/\mu)^j / j! & \text{if } j \leq a \\ \frac{(\lambda/\mu)^j}{a! a^{j-a}} & \text{if } j > a. \end{cases}$$

We know that

$$\sum_{j=0}^{\infty} \kappa_j < \infty \iff \sum_{j=a}^{\infty} \kappa_j < \infty.$$

This occurs if  $\lambda < a\mu$ , in which case

$$\sum_{j=0}^{\infty} \kappa_j = \sum_{k=0}^{a-1} \frac{\lambda^k}{k! \mu^k} + \frac{\lambda^a}{a! \mu^a} \frac{a\mu}{a\mu - \lambda}.$$

## $M/M/a$ ergodicity

So the  $M/M/a$  queue is ergodic if and only if arrival rate  $\lambda$  is less than the maximum service rate  $a\mu$ .

In this case, the stationary distribution is given by

$$\pi_k = \begin{cases} \pi_0(\lambda/\mu)^k/k! & \text{if } k < a \\ \pi_0(\lambda/\mu)^k/(a!a^{k-a}) & \text{if } k \geq a, \end{cases}$$

where

$$\pi_0 = \left( \sum_{j=0}^{\infty} \kappa_j \right)^{-1}.$$

## $M/M/a$ busy servers

For what proportion of time  $\delta_q$  are all the servers busy? This is the same as the probability that an arriving customer will have to wait. (Why?)

We have

$$\delta_q = \sum_{k=a}^{\infty} \pi_k = \pi_0 \frac{\lambda^a}{\mu^a a!} \frac{a\mu}{a\mu - \lambda}.$$

The expected queue length is

$$\ell_q = \mathbb{E}[\max(X_t - a, 0)] = \frac{\lambda}{a\mu - \lambda} \delta_q.$$



## $M/M/a$ busy servers

In stationarity, the expected number  $b_s$  of busy servers is

$$\mathbb{E}_\pi[\min(X_t, a)] = \frac{\lambda}{\mu}.$$

Note that, provided that  $\lambda < a\mu$ , this does not depend on  $a$ .  
The expected number of customers is

$$\ell = b_s + \ell_q = \frac{\lambda}{\mu} + \frac{\lambda}{a\mu - \lambda} \delta_q$$

## $M/M/a$ waiting times

By Little's Law, the expected waiting time is

$$\mathbb{E}_{\pi}[W] = \ell_q / \lambda = \delta_q / (a\mu - \lambda).$$

As a check: if there are  $a$  customers present, the average waiting time of an arriving customer is  $1/(a\mu)$  (the time until the first server becomes free) and if there are  $a + i$  customers present, the average waiting time will be  $(i + 1)/(a\mu)$ . So

$$\mathbb{E}[W] = \sum_{i=0}^{\infty} (i + 1)/(a\mu) \times \pi_{a+i} = (a\mu)^{-1} [\mathbb{E}[\max\{X_t - a, 0\}] + \delta_q].$$

The expected delay is

$$d = \mathbb{E}[W] + \frac{1}{\mu} = \frac{\delta_q}{a\mu - \lambda} + \frac{1}{\mu}.$$

## $M/M/a$ Example

An insurance company has 3 claim adjusters in its branch office. Claims against the company arrive according to a Poisson process at an average rate of 20 per 8 hour day. The amount of time an adjuster spends with a claimant is exponentially-distributed with mean service time 40 minutes.

- ▶ How many hours a week can an adjuster expect to spend with claimants?
- ▶ How much time, on average, does a claimant spend in the branch office?

## $M/M/a$ example solution

- ▶ The arrival rate is  $\lambda = 20/8 = 2.5$  per hour.
- ▶ The service rate is  $\mu = 1.5$  per hour.
- ▶  $\lambda/(a\mu) = 5/9 < 1$ , so a stationary distribution exists.
- ▶ We get  $\mathbb{P}(\text{adjuster is busy})$  by noticing that

$$\mathbb{E}[\text{number of busy adjusters}] = \sum_{i=1}^3 \mathbb{E}[i \text{th adjuster is busy}]$$

So, by symmetry,

$$\begin{aligned}\mathbb{E}[\text{number of busy adjusters}] &= 3\mathbb{E}[1 \text{a given adjuster is busy}] \\ &= 3\mathbb{P}(\text{a given adjuster is busy}).\end{aligned}$$

## $M/M/a$ example solution

Substituting the parameter values, we calculate that each adjuster spends 22.2 hours a week on claims and that  $\pi_0 = 24/139$ ,  $\delta_q = 125/417$  and  $d = 0.817$  hours (which corresponds to 49 minutes).

If there were only two adjusters, we could similarly calculate

- ▶ An adjuster will spend on average 33.3 hours with claimants.
- ▶ We can calculate that  $\pi_0 = 1/11$ ,  $\delta_q = 25/33$  and  $d = 2.18$  hours.

We can use this information to quantify the trade-off between the cost of an extra adjuster and the extra level of service that is produced.

## Single or multiple servers?

Which is better? A single fast server or several smaller ones with the same total productivity?

Assume that the arrival process is Poisson with rate  $\lambda$ , and compare

- ▶ A single server with service rate  $a\mu$ , and
- ▶  $a$  servers with service rate  $\mu$  each.

A heuristic argument tells us that

- ▶ if  $X_t \geq a$ , both systems work with the same rate, but
- ▶ if  $X_t = k < a$  the rate for the  $a$  server queue is  $k\mu$ , which is less than the rate  $a\mu$  for the single server.

So we might conclude that the single server is better. This is easy to prove via the technique of *coupling*.

## Single or multiple servers?

We saw that, for the  $M/M/a$  queue, the expected number in the system is

$$= \frac{\lambda}{\mu} + \frac{\lambda}{a\mu - \lambda} \delta_q$$

and the expected time in the system is

$$= \frac{1}{\mu} + \frac{1}{a\mu - \lambda} \delta_q.$$

For the  $M/M/1$  queue with service rate  $a\mu$ , the expected number in the system is

$$= \frac{\lambda}{a\mu - \lambda}$$

and the expected time in the system is

$$= \frac{1}{a\mu - \lambda}.$$

## Single or multiple servers?

With some work, we can show that  $\delta_q + (a\mu - \lambda)/\mu > 1$ , so both the expected number in the system and the expected time in the system are smaller for the  $M/M/1$  queue, which proves our conjecture.

As an exercise, think about the waiting time, rather than the time in the system, for each of the systems.