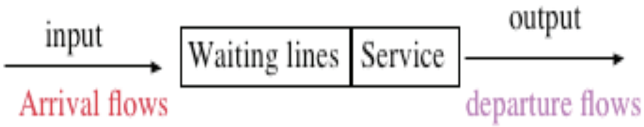
Queueing Systems

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

poisson process

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

Service

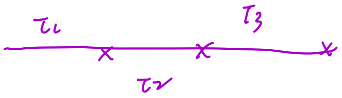
- There is a total of m spaces for both receiving service and waiting for it.
- If there is an idle server, an arriving customer is serviced immediately.
- The service time $S_j^{(i)}$ of the ith customer at the jth server is a random variable.
- When a server is serving a customer, it cannot provide 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 Discipline

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



Kendall's notation

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

M → T_i ~ Exponential & indep (iid)

G → T_i are indep (iid)

D → T_i's are fixed (deterministic)

- A describes the $\boxed{\text{arrival process}}$
  - $A = M$ (Markov) inter-arrival times are independent and exponentially-distributed. CTMC
  - $A = GI$ or (G) inter-arrival times are independent with an arbitrary distribution. not be exponentially-distributed → not Markov
  - $A = D$ inter-arrival times are deterministic.

Kendall's notation

- B describes the service process
  - $B = M$ service times are independent and exponentially-distributed.
  - $B = GI$ or (G) service times are independent with an 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.

= total # people allowed in system
(waiting for services)
+ in service

Some questions

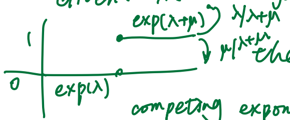
- Does a queueing system have a steady-state regime or does the queue increase unboundedly?
- What is the steady-state queue length distribution if it exists?
- What is the steady-state waiting time distribution if it exists?
- What is the average load on the server?
- What fraction of time is the server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

Exponential queueing systems: $M/M/1$

- Arrival stream: Poisson process with intensity λ
- Service: $n = 1$ server, service time $\sim \exp(\mu)$
- 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.

X_t = # customers in the system at time t

Given: $X_t = 0$ system is in state 0 for $\exp(\lambda)$ time until an arrival & then move to state 1



competing exponential clock

$1 \mapsto 0$ rate μ

$1 \mapsto 2$ rate(λ)

$(X_t)_{t\geq 0}$ is a B-D process with $\lambda_n = \lambda$ & $\mu_n = \mu$

The $M/M/1$ queue

If $X_t = 0$, the process remains at 0 for an $\exp(\lambda)$ time τ_+ 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 $\tau_+ > \tau_-$.

We see that the queue length of an $M/M/1$ queue evolves as a birth and death process with birth rates $\lambda_k = \lambda$ and death rates $\mu_k = \mu$.

Queueing Systems

Using our result from CTMCs, we see that a stationary distribution exists if $\rho \equiv \lambda/\mu < 1$, in which case, for $n \geq 1$,

for B-D process : $\pi A = 0$

$\pi_n = \pi_0 \prod_{j=1}^{n} \frac{\lambda_{j-1}}{\mu_j} \overset{\lambda_k = \lambda, \mu_k = \mu}{=} \left(\frac{\lambda}{\mu}\right)^n \pi_0$

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

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

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

which tells us that

$$\pi_0 = 1 - \rho$$

and, for $n \geq 0$,

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

$L = E[\text{Geo}, (1-\frac{\lambda}{\mu})]$

$= \frac{\frac{\lambda}{\mu}}{1-\frac{\lambda}{\mu}} = \frac{\rho}{1-\rho}$

$L_q = \sum_{n\geq 2} (n-1)\pi_n$

n people in system
n-1 waiting

$= \sum_{n\geq 0} (n-1)\rho^n(1-\rho)$

$= \rho^2(1-\rho)\sum_{n\geq 1}(n-1)\rho^{n-2}$

$\frac{d}{d\rho}[\sum_{n\geq 2}\rho^{n-1}]$

$= (1-\rho)\cdot\rho^2 \frac{d}{d\rho}(\frac{\rho}{1-\rho}) = (1-\rho)\rho^2 \cdot \frac{1}{(1-\rho)^2} = \frac{\rho^2}{1-\rho}$

Some further questions

- What is the stationary expected number L of customers in the whole system?
- What is the stationary expected number L_q of customers in just the queue?
- What is the expected waiting time of a customer?
- What is the distribution of the waiting time?

The first two quantities can be calculated from the stationary distribution. We might guess that $L_q = L - 1$. However this is not right because the queue might be empty. In fact $L_q = E[\max(X_t - 1, 0)]$.

Waiting Times

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 $N \sim \{\pi_k\}$ (PASTA).
- If $N = 0$, then the customer will go straight into service.
- If $N > 0$, the remaining service time X_1 for the customer being served $\sim \exp(\mu)$.
- The service times $X_2, X_3, \ldots, X_N$, for those in the queue are independent $\exp(\mu)$ random variables, also independent of N.
- So the waiting time for our tagged customer is $W = \sum_{j=1}^{N} X_j$, where we interpret the empty sum as equal to 0.

$N | N \geq 1 \sim \text{Geo}_1(1-\rho)$

$\pi_0 = \pi_0$

$P(W=0) = (1-\rho)$

$P(N=0)$

given $N \geq 1$ (with prob ρ)

$W(N\geq 1) \stackrel{d}{=} \text{Gamma}(N, \mu)$

$W | N \geq 1 \sim \text{Exp}(\mu - \lambda)$ $\mu(1-\rho)$

$[P(N=k | N\geq 1] = \rho^{k-1}(1-\rho)$

Waiting Times

We can write

$$
\begin{aligned}
E[e^{-sW}] &= E[E[e^{-sW}|N]] \\
&= E\left[\left(\frac{\mu}{s+\mu}\right)^N\right] \\
&= (1-\rho)\sum_{n=0}^{\infty}\rho^n\left(\frac{\mu}{s+\mu}\right)^n \\
&= \frac{(1-\rho)(s+\mu)}{s+\mu-\lambda} \\
&= (1-\rho)+\rho\frac{\mu-\lambda}{s+\mu-\lambda},
\end{aligned}
$$

and we see that W equal to zero with probability $1-\rho$ and, with probability ρ is exponentially distributed with parameter $\mu - \lambda$.

Queueing Systems

It follows that the expected waiting time is

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

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

→ exp(μ) service

$$D = E[W] + \frac{1}{\mu} = 1/(\mu - \lambda).$$

Denoting by L, the mean number in the system, we can derive a simple relation between L and D that called Little's law.

Little's law:

$$L = \lambda D,$$

"system"

and

$$L_q = \lambda E[W].$$

"queue"

Sketch of a Proof

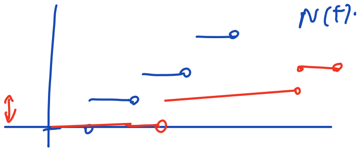
The second relationship is a queue (rather than system) version of the first.

Queuing Systems

The total number of customers to have entered the system in $[0, t]$ is $N(t)$ and the total number to have departed is $D(t)$. So the number present at time t is $L(t) = N(t) - D(t)$. Denoting the area under the function $L(t)$ by $A(t)$, we calculate $E[A(t)/t]$ in two different ways. First

$$A(t)/t = \frac{1}{t}\int_0^t L(u)du$$

which, assuming ergodicity, approaches the average number L in the system with probability one as $t \to \infty$.



distance

people in the system

Assumption: System is ergodic

$\frac{A(t)}{t}$ = area between "curve" /t

$= \frac{1}{t} \sum_{k=1}^{10} (\text{Amount of time } k \text{ people in system in } [0,t]) \cdot k$

$= \sum_{k=1}^{10} (\text{Proportion of time } k \text{ people}) \cdot k$

In system

$$\xrightarrow[t \to \infty]{} \quad \sum_{k=1}^{\infty} k \cdot \pi_k = L$$

Queueing Systems

Second, we have (approximately),

$$A(t)/t = \frac{1}{t} \sum_{n=1}^{N(t)} D_n$$

where D_n is the delay experienced by the nth customer.

D_n = time in system of nth customer

$A(t) \approx \sum_{n=1}^{N(t)} D_n$ $N(t) \approx t\lambda \to$ LLN

Assumption $N(t)$ to satisfy a LLN

$N(t) \approx \lambda t$ t larger

$\frac{A(t)}{t} \approx \sum_{n=1}^{\lambda t} D_t$

$\approx \lambda D$ $\to$ expected delay

(large t)

Assumption $\sum_{n=1}^{m} D_n \approx m D$

(m large)

Queueing Systems

Now

$$
\begin{aligned}
E[A(t)/t] &= \frac{1}{t}E[\sum_{n=1}^{N(t)} D_n] \\
&= \frac{1}{t}E[E[\sum_{n=1}^{N(t)} D_n|N(t)]] \\
&= \frac{1}{t}E[N(t)D] \\
&= \frac{\lambda t D}{t} \\
&= \lambda D.
\end{aligned}
$$

So we have $L = \lambda D$.

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.

Analysis:

- 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 $L = \rho/(1-\rho) = 2$.
- The expected time under repair is $D = L/\lambda \text{ (or } 1/(\mu - \lambda)) = 24$ minutes.
- The expected time waiting for repair is $E[W] = D - 1/\mu = 16$ minutes.
- Let T_q be the time that a machine has to wait before being repaired. Then $P(T_q > 10) = \rho e^{-(\mu-\lambda)10} = 0.44$.

Example

- Suppose that the failure rate of machines increases (eg due to aging) by 16% to $\lambda' = 1/10$, then the new traffic intensity is $\rho' = 4/5$, and $L' = 4$ with $D' = 40$ and $E[W'] = 32$.
- 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 ρ is close to 1, the effect of small changes of ρ 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

- 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?

old system
cost / minutes = $L \cdot C_D$

new system
cost / minute = $L^* \cdot C_D + C_M$

Solution

Without the new equipment

- The expected number of failed machines is $L = \rho/(1-\rho) = 2$.
- The expected cost of lost production is $LC_D = \$10$ per minute.
- With the new equipment, $\rho^* = 0.5$, and the expected cost is $\underset{1}{L^*}\,\underset{5}{C_D} + \underset{10}{C_M} = \$15$ per minute.
- We should buy the equipment if $L^* C_D + C_M < LC_D$, so we should not buy the equipment.

Fixed λ

cost $L \cdot C_1 + \mu C_2$

$= \frac{\rho}{1-\rho} \cdot C_1 + \mu C_2$

$= \frac{\lambda}{\mu-\lambda} C_1 + \mu C_2$

$= C(\mu)$

$C'(\mu) = -\frac{C_1\lambda}{(\mu-\lambda)^2} + C_2 = 0$

when $\mu^* = \sqrt{\frac{C_1\lambda}{C_2}} + \lambda$

$C''(\mu) > 0 \Rightarrow C(\mu^*)$ is a minimum

Example

At a service station the rate of service is μ cars per hour, and the rate of arrivals of cars is λ 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 μ is a control parameter.

Determine the value of μ so that the least expected cost is achieved and find the value of the latter.

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, $E[Y] = \rho/(1-\rho)$.
- The expected total cost per hour is $\$C(\mu) = C_1\rho/(1-\rho) + \mu C_2$
- To find the minimum, we find μ 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 μ_0 achieves the minimum and $C(\mu_0) = \lambda C_2 + 2\sqrt{\lambda C_1 C_2}$.

#Markov arrival
↑ Markov service

The $M/M/a$ Queue

servers

This system has

B-D process

$\lambda_k = \lambda \to$ Poisson(λ) arrivals

- $a \geq 1$ servers,
- a Poisson arrival process with rate λ,
- a FIFO service discipline, and
- 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 time until the next departure is an $\exp(k\mu)$ random variable.

$$\mu_k = \begin{cases} k\mu & \text{if } 1 \leq k \leq a \\ a\mu & \text{if } k > a \end{cases}$$

$a+j$, $j \geq 0$ customer in system

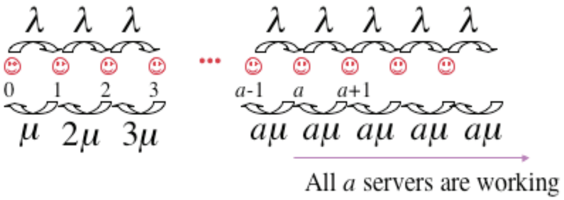
$\Rightarrow$ all a servers are working

(a customer in service & j customers waiting for service)

Queueing Systems

B-D process will be ergodic when $\pi \Lambda = 0$ has a pos sol

The transition diagram is



$$\pi_k = \pi_0 \prod_{j=1}^{k} \frac{\lambda_{j-1}}{\mu_j}$$

$$= \begin{cases} \pi_0 \prod_{j=1}^{k} \frac{\lambda}{\mu_j} & k \le a \\ \pi_0 \prod_{j=1}^{a} \frac{\lambda}{\mu_j} \prod_{l=a+1}^{k} \left(\frac{\lambda}{\mu a}\right) & k > a \end{cases}$$

The number in the queue follows a birth-and-death process with $\lambda_j = \lambda$ for $j = 0, 1, 2, \ldots$ and $\mu_j = j\mu$ for $j = 1, 2, \ldots, a$ and $\mu_j = a\mu$ for $j = a+1, a+2, \ldots$.

$$= \begin{cases} \pi_0 \left(\frac{\lambda}{\mu}\right)^k \cdot \frac{1}{k!} & k \le a \\ \pi_0 \left(\frac{\lambda}{\mu}\right)^a \cdot \frac{1}{a!} \left(\frac{\lambda}{\mu a}\right) & k > a \end{cases}$$

Queueing Systems

When is the $M/M/a$ queue ergodic?

$$K_j \equiv \frac{\lambda_0 \cdots \lambda_{j-1}}{\mu_1 \cdots \mu_j} = \begin{cases} (\lambda/\mu)^j/j! & \text{if } j \le a \\ \frac{(\lambda/\mu)^j}{a!a^{j-a}} & \text{if } j > a. \end{cases}$$

We know that

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

$$\sum_{j=a}^{\infty} \left(\frac{\lambda}{\mu}\right)^a \cdot \frac{1}{a!} \left(\frac{\lambda}{\mu a}\right)^{j-a}$$

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

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

$$\left(\frac{\lambda}{\mu}\right)^a \cdot \frac{1}{a!} \left(\frac{1}{1-\frac{\lambda}{\mu a}}\right)$$

$$\Rightarrow \frac{\lambda}{\mu a} < 1.$$

So the $M/M/a$ queue is ergodic if and only if arrival rate λ 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

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

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

more than a people in the system

We have

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

$$\sum_{k\geq a} \pi_0 \left(\frac{\lambda}{\mu}\right)^a \frac{1}{a!} \left(\frac{\lambda}{\mu a}\right)^{k-a}$$

$$= \pi_0 \left(\frac{\lambda}{\mu}\right)^a \frac{1}{a!} \left(\frac{1}{1-\frac{\lambda}{\mu a}}\right)$$

The expected queue length is

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

$$= \sum_{k\geq a} (k-a)\pi_k = \pi_0 \left(\frac{\lambda}{\mu}\right)^a \frac{1}{a!} \sum_{k\geq a} (k-a) \left(\frac{\lambda}{\mu a}\right)^{k-a}$$ (make it like derivative)

$$= \pi_0 \left(\frac{\lambda}{\mu}\right)^a \frac{1}{a!} \left(\frac{\lambda}{\mu a}\right) \sum_{k\geq a} (k-a) \left(\frac{\lambda}{\mu a}\right)^{k-a-1}$$

$= \pi_0 \left(\frac{\lambda}{\mu}\right)^a \cdot \frac{1}{a!} \left(\frac{\lambda}{\mu a}\right) \left(\frac{d}{dx} \sum_{k \geq a} x^{k-a} \Big|_{x=\frac{\lambda}{\mu a}}\right).$

$= \pi_0 \left(\frac{\lambda}{\mu}\right)^a \cdot \frac{1}{a!} \left(\frac{\lambda}{\mu a}\right) \left(\frac{1}{1-\frac{\lambda}{\mu a}}\right)^2$

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

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

The expected number N_b of busy servers is

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

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

$$L = N_b + L_q = \frac{\lambda}{\mu} + \frac{\lambda}{a\mu - \lambda} P_q$$

busy server

= # servers nasservice

waiting for service

$E[N_0]$

$$= \sum_{k=1}^{a} k \cdot \pi_k + \sum_{k=a+1}^{\infty} a \cdot \pi_k$$

$$= \sum_{k=1}^{a-1} k \cdot \pi_k + \sum_{k=a}^{\infty} a \pi_k$$

$$= \pi_0 \sum_{k=1}^{a-1} k \cdot \left(\frac{\lambda}{\mu}\right)^k \cdot \frac{1}{k!} + a \pi_0 \sum_{k=a}^{\infty} \left(\frac{\lambda}{\mu}\right)^a \cdot \frac{1}{a!} \cdot \left(\frac{\lambda}{\mu a}\right)^{k-a}$$

$$= \pi_0 \cdot \left(\sum_{k=1}^{a-1} \left(\frac{\lambda}{\mu}\right)^k \cdot \frac{1}{(k-1)!} + \frac{\lambda}{\mu} \sum_{k=a}^{\infty} \left(\frac{\lambda}{\mu}\right)^a \cdot \frac{1}{a!} \left(\frac{\lambda}{\mu}\right)^{k-a-1} \right)$$

$$= \frac{\lambda}{\mu} \underbrace{\left(\sum_{k=1}^{a-1} \underbrace{\left(\frac{\lambda}{\mu}\right)^{k-1} \cdot \frac{1}{(k-1)!} \cdot \pi_0}_{\pi_{k-1}} + \sum_{k=a}^{\infty} \left(\frac{\lambda}{\mu}\right)^a \cdot \underbrace{\frac{1}{a!} \left(\frac{\lambda}{\mu}\right)^{k-a-1}}_{\pi_{k-1}} \right)}_{=1}$$

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

Queueing Systems

system
$L = \lambda D \to$ expected total time in system
↑ arrival rate
mean number of people

By Little's Law, the expected waiting time is

$$E[W] = L_q/\lambda = P_q/(a\mu - \lambda).$$

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

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

queue
$L_q = \lambda E(W)$

if you show up, all servers are busy but no one waiting, you are waiting for the first person to leave exp(aμ)

a+i customer
wait i+1 server time

The expected delay is

$$D = E[W] + \frac{1}{\mu} = \frac{P_q}{a\mu - \lambda} + \frac{1}{\mu}.$$

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?

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 P(adjuster is busy) by noticing that

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

So, by symmetry,

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

proportion of time given the adjuster is busy is

$$\frac{Nb}{3} = \frac{\lambda}{\mu \cdot 3} = \frac{5}{9}$$

$$\# \text{ hours} / \text{week} = \frac{40 \cdot 5}{9} = \frac{200}{9} \approx 22.2$$

φ_0 = # hour per week

Substituting the parameter values, we calculate that each adjuster spends 22.2 hours a week on claims and that $\pi_0 = 24/139$, $P_q = 125/417$ 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 $\pi_0 = 1/11$, $P_q = 25/33$ 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 λ, and compare

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

more power when fewer people In the system

Queueing Systems

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

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

and the expected time in the system is

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

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

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

With some work, we can show that $P_q + (a\mu - \lambda)/\mu > 1$, 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.