Queuing Theory

Overview of Queuing Theory Lectures

Introduction to Queueing Systems
- Important Questions about Queueing Systems
- Components of a Queuing System

Fundamental Results for General Queueing Systems
- Terminology for General Queueing Systems
- Steady State Quantities
- Fundamental Quantities of Interest
- Basic Cost Identity
- Consequences: Little's Law and Other Useful Relationships

Properties of Poisson Arrival Processes
- Superposition and decomposition of Poisson arrivals
- Poisson arrivals see time averages

Overview of Queuing Theory Lectures, continued

Queueing Systems with Exponential Arrivals and Service

- M/M/1 queue
- M/M/S queue
- M/M/1/C queue
- Markovian queues with general state definition

examples: security checkpoint
Safeway checkout lines
two law practices
Stanford post office
shoeshine shop

Networks of Markovian Queues

- The Equivalence Property
- Open Jackson networks
- Closed Jackson networks

examples: car wash
Ward's Berry Farm
roommate network

Queueing Systems with Nonexponential Distributions

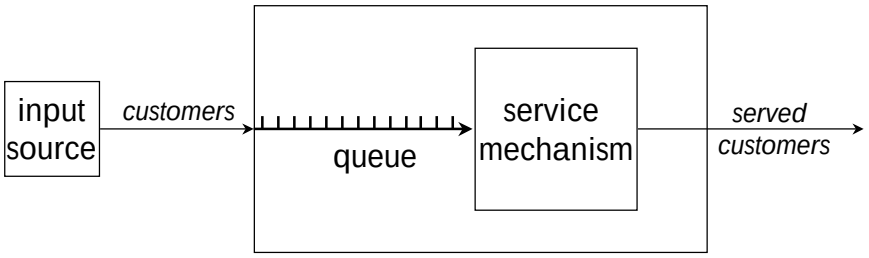
- The M/G/1 Queue
- Special Cases of M/G/1: M/D/1 and $M/E_k/1$ queues

examples: McDonald's Drive Thru
homework questions

Important Questions about Queueing Systems

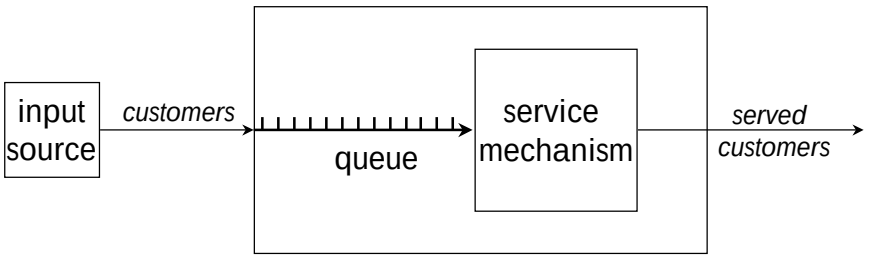
- What fraction of time is each server idle?
- What is the expected number of customers in the queue? in the queue plus in service?
- What is the probability distribution of the number of customers in the queue? in the queue plus in service?
- What is the expected time that each customer spends in the queue? in the queue plus in service?
- What is the probability distribution of a customer's waiting time in the queue? in the queue plus in service?

Components of a Queueing System



Arrival Process Queue Service Process

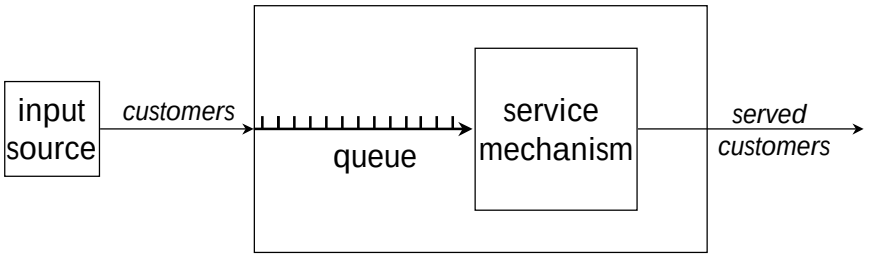
Components of a Queueing System: The Arrival Process



Characteristics of Arrival Process:

- finite or infinite calling population
- bulk or individual arrivals
- interarrival time distribution
- simple or compound arrival processes
- balking or no balking

Components of a Queueing System: The Queue



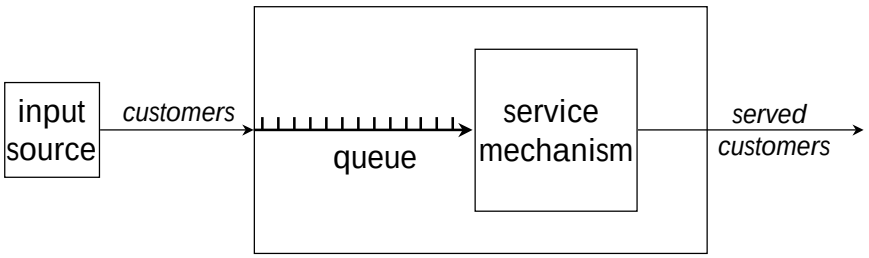
Queue Characteristics:

- finite or infinite
- queue discipline:

  FCFS = first-come-first-served *LCFS = last -come-first-served*

  Priority service order *Random service order*

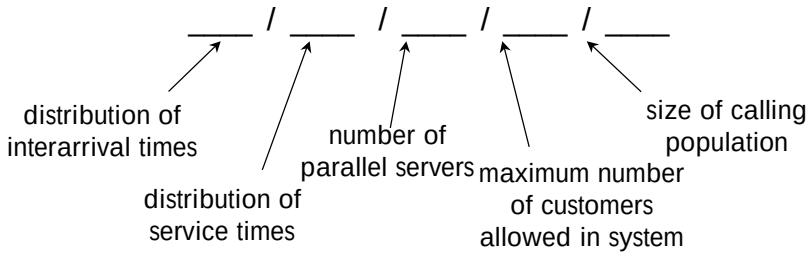
Components of a Queueing System: The Service Process



Characteristics of Service Process:

- number / configurations of servers
- batch or single service
- service time distribution
- rework

Notation for Characterizing Queueing Systems



M = iid exponential
D = iid and deterministic
$\boldsymbol{E_k}$ = Erlang with shape parameter *k*
GI = iid with general distribution

Notation for Characterizing Queueing Systems: Examples

M / M / 1 / ∞ / ∞	single server queue with Poisson arrivals and exponential service times, a.k.a. M/M/1
M / M / S / ∞ / ∞	S server queue with Poisson arrivals and exponential service times, a.k.a. M/M/S
M / M / 1 / C / ∞	same as M/M/1 with finite system capacity C
M / M / 1 / K / K	single server queue with Poisson arrivals from a population of K potential customers; and exponential service times

Terminology for General Queueing Systems

$N(t)$ = the number of customers in the system at time t: this includes both customers that are in the queue and being served

$P_n(t)$ = the probability that there are exactly n customers in the system at time t

= $P(N(t)=n)$

Note: $N(t)$ and $P_n(t)$ are difficult to compute for general t. We will be mainly interested in their values as t becomes infinite.

λ_n = mean arrival rate of entering customers when n customers are in the system

μ_n = mean service rate for overall system when n customers are in the system

Steady State Quantities

P_n = long run probability that there will be exactly n customers in the system

= $\lim_{t\to\infty} P_n(t)$

$\bar{\lambda}$ = long run average arrival rate of entering customers

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

For general queues, the steady state probabilities P_n can be difficult to compute. However, if the queue is a continuous time Markov Chain, then $P_n = \omega_n$ (if a steady-state distribution exists).

Fundamental Quantities of Interest

L = the long run average number of customers in the system

L_Q = the long run average number of customers waiting in queue

L_S = the long run average number of customers in service

$$L = L_S + L_Q$$

More Fundamental Quantities of Interest

W = the waiting time in the system for an arbitrary customer (random variable)

W_Q = the waiting time in the queue for an arbitrary customer (random variable)

W = the long run average amount of time a customer spends in the system

= $E(W)$

W_Q = the long run average amount of time a customer spends in the queue

= $E(W_Q)$

W_S = the long run average amount of time a customer spends in service

$$W = W_S + W_Q$$

Basic Cost Identity

Suppose each entering customer must pay money to the system. Then the following identity applies:

> average rate at which the system earns
> $= \bar{\lambda} \times$ average amount an entering customer pays

Many fundamental relationships about queueing system performance can be derived from this identity. It can be used not only to compute monetary income rates but also to derive relationships between the fundamental quantities we are interested in for queueing systems. For example, if each customer pays \$1 per unit time that they are in the system (i.e. either in queue or being served) then

the average rate at which the system earns = L

the average amount an entering customer pays = W

For this example, the basic cost identity tells us: $L = \bar{\lambda} W$

Little's Law

A quick diversion

A few topics I want to weave in before we proceed with Little's Law and its variations:

- The superposition and decomposition of Poisson processes
- Poisson arrivals see time averages
- Computing steady state-equations: rate in = rate out

Superposition of Poisson Arrival Processes

Often the arrival process to a queue consists of multiple different arrival processes of customers from different origins. It turns out that it is easy to deal with such compound arrival processes when each individual arrival process is Poisson and is independent of the others.

Suppose you have two independent Poisson arrival processes X and Y with respective rates λ_x and λ_y. Then the combination of the two arrival streams is also a Poisson process with rate $\lambda_x+\lambda_y$.

Why? Let T be a random variable representing the remaining time until an arrival of either type occurs. Let T_x and T_y be the remaining times until arrivals of type x and y occur. Then $T = \min\{T_x,T_y\}$. Since T_x and T_y are exponentially distributed with parameters λ_x and λ_y respectively, T must also be exponentially distributed with parameter $\lambda_x+\lambda_y$. Hence, the combined stream has exponential interarrival times, and is thus a Poisson process with rate $\lambda_x+\lambda_y$.

Decomposition of Poisson Arrival Processes

Another nice property of Poisson arrival processes allows us to disaggregate the process into independent Poisson processes. Suppose customers arrive to our system according to a Poisson process with rate λ. Suppose that an arriving customer is of type k with probability p_k, where

$$\sum_{k=1}^{n} p_k = 1.$$

Then the arrivals of customers of type k follow a Poisson process with rate $p_k\lambda$.

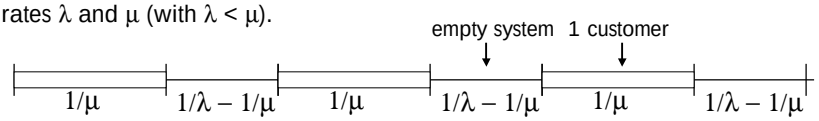
An example: suppose arriving customers to a system balk with fixed probability p independent of the number of people in the system. If the arrival process is Poisson with rate λ, the arrivals who stay follow a Poisson process with rate $(1-p)\lambda$. The customers who balk do so according to a Poisson process with rate $p\lambda$.

Poisson Arrivals See Time Averages (PASTA)

The PASTA Property

For a queuing system with Poisson arrivals, an arriving customer sees the time-average steady-state number-in-system process. In other words, the long run fraction of customers that arrive to find exactly k customers (not including him or herself) is given by ω_k.

This property is called "Poisson Arrivals See Time Averages." It depends critically on the assumption of having a Poisson arrival process. Suppose, instead, for example, that we have deterministic arrival and service processes with respective rates λ and μ (with $\lambda < \mu$).



The time-average steady-state fraction of time in which there are 0 customers in the system is $1 - \lambda/\mu$, and the steady-state fraction of time in which there is 1 customers is λ/μ. But 100% of arriving customers find no customers in the system! "PASTA" does not apply in this case because the arrivals are not Poisson.

Example: Poisson Arrivals See Time Averages

EXAMPLE: The M/M/1 queue

In an M/M/1 queue, what fraction of arriving customers have to wait? The fraction that finds the system in any state $k > 0$, which, according to the PASTA property, is $1 - \omega_0 = \rho$.

Computing the steady state distribution: Rate In = Rate Out

When studying CTMCs, we learned that each of the steady state equations

$$\sum_{i=0}^{S} \omega_i q_{ij} = 0 \quad \text{for} \quad j = 0,...,S$$

could be expressed as:

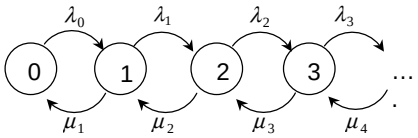
steady state rate out of state j =

$$q_j \omega_j = \sum_{i \neq j} \omega_i q_{ij}$$

= steady state rate into state j

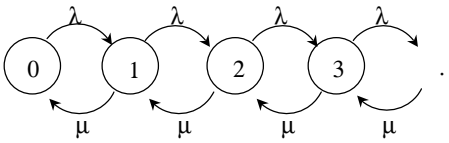
In simple queuing systems that can be modeled as birth and death processes, the "rate-in = rate-out" equations are particularly simple to write down because transitions can only occur to adjacent states. In such cases, they are a shortcut to writing down the steady state equations.

Rate In = Rate Out for the General Birth and Death Process



State	Rate In = Rate Out
0	$\mu_1 \omega_1 = \lambda_0 \omega_0$
1	$\lambda_0 \omega_0 + \mu_2 \omega_2 = (\lambda_1 + \mu_1) \omega_1$
…	…
n-1	$\lambda_{n-2} \omega_{n-2} + \mu_n \omega_n = (\lambda_{n-1} + \mu_{n-1}) \omega_{n-1}$
n	$\lambda_{n-1} \omega_{n-1} + \mu_{n+1} \omega_{n+1} = (\lambda_n + \mu_n) \omega_n$
…	…

Example: Rate In = Rate Out for the M/M/1 queue



State	Rate In = Rate Out
0	$\mu\omega_1 = \lambda\omega_0$
1	$\lambda\omega_0 + \mu\omega_2 = (\lambda + \mu)\omega_1$
…	…
n-1	$\lambda\omega_{n-2} + \mu\omega_n = (\lambda + \mu)\omega_{n-1}$
n	$\lambda\omega_{n-1} + \mu\omega_{n+1} = (\lambda + \mu)\omega_n$
…	…

Little's Law

$$L = \overline{\lambda}\, W$$

A Proof of Little's Law

number of people in system at time s.

$$L = \lim_{t\to\infty} \frac{1}{t}\int_0^t N(s)ds \qquad \bar{\lambda} = \lim_{t\to\infty} \frac{Q(t)}{t} \qquad W = \lim_{t\to\infty} \frac{1}{Q(t)}\sum_{i=1}^{Q(t)} W_i$$

number of arrivals by time t.

customer i's time in system

Let A_i and D_i be the arrival and departure times of customer i. Then $W_i = D_i - A_i$. Define the indicator function

$$I(A_i \le s \le D_i) = \begin{cases} 1 \text{ if } A_i \le s \le D_i \\ 0 \text{ otherwise} \end{cases}$$

$$\frac{1}{t}\int_0^t N(s)ds = \frac{1}{t}\int_0^t \sum_{i=1}^{Q(t)} I(A_i \le s \le D_i)ds$$

$$\approx \frac{1}{t}\sum_{i=1}^{Q(t)} \int_0^{\infty} I(A_i \le s \le D_i)ds$$

$$= \frac{1}{t}\sum_{i=1}^{Q(t)} W_i = \frac{Q(t)}{t}\frac{1}{Q(t)}\sum_{i=1}^{Q(t)} W_i$$

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t N(s)ds = \lim_{t\to\infty} \frac{Q(t)}{t}\frac{1}{Q(t)}\sum_{i=1}^{Q(t)} W_i \quad \Rightarrow \quad L = \bar{\lambda} W$$

Other Useful Relationships

Another useful relationship can be obtained from the basic cost identity by assuming each customer pays \$1 per unit time that they are in the queue. Then

the average rate at which the system earns = L_Q

the average amount an entering customer pays = W_Q

In this case, the basic cost identity tells us:

$$L_Q = \bar{\lambda} W_Q$$

Other Useful Relationships

By assuming instead that each customer pays \$1 per unit time that they are in service, we have

the average rate at which the system earns = L_S

the average amount an entering customer pays = W_S

The basic cost identity amounts to:

$$L_S = \lambda W_S$$

Exponential Queuing Models

We now show how to apply these results to queues having exponentially distributed interarrival times and service times. These types of queues are the most tractable mathematically.

Some of the queues we'll look at:

M / M / 1

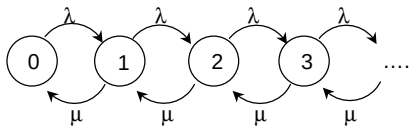
M / M / S

M / M / 1 / C (finite capacity C)

Queues with more general state definition

Exponential Queuing Models: The M/M/1 Queue

The simplest example of a queue that has exponentially distributed interarrival and service times is the by-now familiar M/M/1 queue. This is a queue for which the fundamental quantities are easy to compute. We will do so and also apply Little's law to the M/M/1 case.



$\lambda_i = \lambda \quad \text{for all } i \geq 0$

$\mu_i = \mu \quad \text{for all } i \geq 1$

Define $\rho=\lambda/\mu$. This quantity is typically called the traffic intensity for the M/M/1 queue. We assume henceforth that $\rho<1$, in which case we know a steady state distribution exists and is given by:

$$P_j = \omega_j = \rho^j(1-\rho) \quad \text{for } j = 0,1,\ldots$$

The arrival rate of entering customers is always λ; therefore $\bar{\lambda} = \lambda$.

Exponential Queuing Models: The M/M/1 Queue

The long run average number of people in the system can be computed using the steady state distribution:

$$L = \sum_{j=0}^{\infty} jP_j = \sum_{j=0}^{\infty} j\omega_j = (1-\rho)\sum_{j=0}^{\infty} j\rho^j$$

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

$$= \frac{\rho}{(1-\rho)}$$

$$= \frac{\lambda}{(\mu-\lambda)}$$

Exponential Queuing Models: The M/M/1 Queue

We can also compute the long run expected number of customers in the queue L_Q using the steady state distribution as follows:

$$L_Q = \sum_{j=1}^{\infty} (j-1) P_j = \sum_{j=1}^{\infty} (j-1)\omega_j = \sum_{j=1}^{\infty} j\omega_j - \sum_{j=1}^{\infty} \omega_j$$

$$= \sum_{j=0}^{\infty} j\omega_j - (1-\omega_0)$$

$$= L - \rho$$

$$= \frac{\lambda}{(\mu-\lambda)} - \frac{\lambda}{\mu} = \frac{\lambda^2}{\mu(\mu-\lambda)}$$

The expected number of customers in service L_S is

$$L_S = 0 \cdot \omega_0 + \sum_{j=1}^{\infty} 1 \cdot \omega_j = 1 - \omega_0 = \rho = \frac{\lambda}{\mu}$$

Verify that $L = L_S + L_Q$ for this example.

Exponential Queuing Models: The M/M/1 Queue

The distribution for the waiting time W of an arbitrary customer (in the long run) is computed as follows:

$$P(W \le a) = \sum_{n=0}^{\infty} P(W \le a \mid n \text{ in system when he arrives}) P(n \text{ in system when he arrives})$$

$$= \sum_{n=0}^{\infty} P(\text{service times of person in service} + (n-1) \text{ waiting} + \text{him} \le a) P_n$$

$$= \sum_{n=0}^{\infty} \int_0^a \mu e^{-\mu t} \frac{(\mu t)^n}{n!} dt \cdot \rho^n (1-\rho)$$

(sum of n+1 iid exponential random variables with parameter μ has an Erlang distribution with parameters (μ/(n+1), n+1) (aka gamma (n+1,μ))

$$= \int_0^a (\mu-\lambda) e^{-\mu t} \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!} dt$$

(interchanging the sum and the integral)

$$= \int_0^a (\mu-\lambda) e^{-\mu t} e^{\lambda t} dt = 1 - e^{-(\mu-\lambda)a}$$

(using the identity $\sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!} = e^{\lambda t}$)

So W is exponentially distributed with parameter (μ–λ)!

Exponential Queuing Models: The M/M/1 Queue

Now we'll derive the distribution for the waiting time in the queue W_Q of an arbitrary customer in the long run. Since an arriving customer who find no customers in the queue has no wait, $P(W_Q = 0) = P_0 = 1 - \rho$

$$P(W_Q > a) = \sum_{n=1}^{\infty} P(W_Q > a \mid n \text{ in system when he arrives}) P(n \text{ in system when he arrives})$$

$$= \sum_{n=1}^{\infty} P(\text{service times of person in service} + (n-1) \text{ waiting} > a) P_n$$

$$= \sum_{n=1}^{\infty} \int_a^{\infty} \mu e^{-\mu t} \frac{(\mu t)^{n-1}}{(n-1)!} dt \cdot \rho^n (1\text{-}\rho)$$

$$= \int_a^{\infty} \frac{\lambda(\mu - \lambda)}{\mu} e^{-\mu t} \sum_{n=1}^{\infty} \frac{(\lambda t)^{n-1}}{(n-1)!} dt$$

$$= \int_a^{\infty} \frac{\lambda(\mu-\lambda)}{\mu} e^{-\mu t} e^{\lambda t} dt$$

$$= \rho e^{-(\mu-\lambda)a}$$

W_Q is not exponentially distributed, but the conditional distribution of W_Q, given that W_Q>0, is! To see this...

$$P(W_Q > a \mid W_Q > 0) = \frac{P(W_Q > a)}{P(W_Q > 0)} = e^{-(\mu-\lambda)a}$$

Exponential Queuing Models: The M/M/1 Queue

Let's now compute W, the expected time a customer spends in the system, first using Little's Law:

$$W = \frac{L}{\bar{\lambda}} = \frac{L}{\lambda} = \frac{\lambda}{\mu - \lambda} \cdot \frac{1}{\lambda} = \frac{1}{\mu - \lambda}$$

It is easy to verify that we get the same result by taking the expected value of W.

Using $L_Q = \bar{\lambda} W_Q$ we can compute the expected time a customer spends in the queue:

$$W_Q = \frac{L_Q}{\bar{\lambda}} = \frac{L_Q}{\lambda} = \frac{\lambda^2}{\mu(\mu - \lambda)} \cdot \frac{1}{\lambda} = \frac{\lambda}{\mu(\mu - \lambda)}$$

Verify by taking the expected value of W_Q!

The expected time a customer spends in the service is:

$$W_S = \frac{L_S}{\bar{\lambda}} = \frac{\lambda}{\mu} \frac{1}{\lambda} = \frac{1}{\mu}$$ obvious!

Evidently $W = W_S + W_Q$ for this example.

The M/M/1 Queue: Example

EXAMPLE: Security Checkpoint
Airline passengers must pass through a security checkpoint consisting of a metal detector and carry-on luggage x-ray machine. Suppose that passengers arrive according to a Poisson process with a rate of λ=10 passengers per minute. The security checkpoint clears customers at a rate of μ=12 passengers per minute, with exponentially distributed clearance times. What is the probability that a passenger will have to wait before being checked?

$$P(W_Q > 0) = \rho e^0 = \rho = \frac{5}{6}$$ (or equivalently 1- $P(W_Q=0) = \rho$)

On average, how many customers are waiting in line to be checked?

$$L_Q = \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{100}{12(12-10)} = \frac{25}{6}$$

On average, how long will a customer be detained at the checkpoint?

$$W = \frac{1}{\mu - \lambda} = \frac{1}{2} \text{ minute}$$

The M/M/1 Queue: Example

EXAMPLE: Security Checkpoint, continued
The airline in question wants to determine how many checkpoints to operate to minimize the costs associated with operation and customer delays. They estimate the cost of delaying a customer 1 hour to be $4. It costs $100 per hour to staff and operate a checkpoint. Assume that a passenger is equally likely to enter each checkpoint.

Let n be the number of separate checkpoints. Then a given passenger goes to any particular checkpoint with probability $1/n$. We can model each of the n checkpoints as a separate M/M/1 queue with a Poisson arrival process having rate λ/n. For any particular checkpoint (say, checkpoint i), the average number of customers present is

$$L_i = \frac{\lambda/n}{\mu - (\lambda/n)} = \frac{\lambda}{\mu n - \lambda} = \frac{10}{12n - 10}$$

The expected average cost of operation plus passenger delays per hour that the airline faces when they have n checkpoints open is then

$$F(n) = 100n + 4L = 100n + 4(nL_i) = 100n + \frac{40n}{12n - 10}$$

where L represents the average total number of people at all checkpoints.

The M/M/1 Queue: Example

EXAMPLE: Security Checkpoint
To find the optimal number of checkpoints to operate, we need to find the minimum of the function $F(n)$.

$$F(n) = 100n + \frac{40n}{12n - 10}$$

$$F'(n) = 100 - \frac{400}{(12n - 10)^2}$$

$$F''(n) = \frac{9600}{(12n - 10)^3} \geq 0$$

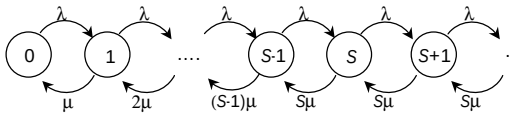
Clearly the function $F(n)$ is convex, so to find its minimum we need only to set its derivative equal to zero.

$$F'(n^*) = 0 \quad \Leftrightarrow \quad (12n^* - 10)^2 = 4 \quad \Leftrightarrow \qquad n^* = \pm 1$$

Since we must have at least one checkpoint, the optimal number of checkpoints is 1.

Exponential Queuing Models: The M/M/S Queue

Another important example of a queue with exponentially distributed interarrival and service times is the M/M/S queue. For this case we'll now compute the fundamental quantities and apply Little's law.



$$\lambda_i = \lambda \quad \text{for all } i \geq 0$$

$$\mu_i = \mu \min\{S, i\} \quad \text{for all } i \geq 1$$

Define $\rho = \lambda/S\mu$ as the traffic intensity for the M/M/S queue. To ensure the existence of a steady state distribution, we assume that $\rho < 1$. In this case:

$$P_0 = \omega_0 = (1 + K)^{-1}$$

$$P_j = \omega_j = \begin{cases} \frac{\lambda^j}{j!\mu^j}(1 + K)^{-1} & \text{for } j \leq S \\ \frac{\lambda^j}{S!S^{j-S}\mu^j}(1 + K)^{-1} & \text{for } j > S \end{cases}$$

where

$$K \equiv \sum_{i=1}^{S-1} \frac{\lambda^i}{i!\mu^i} + \frac{\lambda^S}{S!\mu^S}\left(1 - \frac{\lambda}{S\mu}\right)^{-1}$$

Again the arrival rate of entering customers is always λ; therefore $\bar{\lambda} = \lambda$.

Exponential Queuing Models: The M/M/S Queue

We can compute the long run expected number of customers in the queue L_Q using the steady state distribution as follows:

$$\begin{aligned} L_Q &= \sum_{j=S}^{\infty} (j-S) P_j = (1+K)^{-1} \sum_{j=S}^{\infty} (j-S) \frac{(\lambda/\mu)^j}{S! S^{j-S}} \\ &= (1+K)^{-1} \sum_{i=0}^{\infty} i \frac{(\lambda/\mu)^{i+S}}{S! S^i} \\ &= (1+K)^{-1} \frac{(\lambda/\mu)^S}{S!} \sum_{i=0}^{\infty} i\rho^i \\ &= (1+K)^{-1} \frac{(\lambda/\mu)^S}{S!} \frac{\rho}{(1-\rho)^2} \end{aligned}$$

Now we can apply Little's law and our other identities to compute:

$W_Q = L_Q / \bar{\lambda}$ $W = W_Q + (1/\mu)$ $L = \bar{\lambda} W$

M/M/S Queue: Example

EXAMPLE: Safeway Checkout Lines
Safeway is trying to decide how many checkout lines to keep open. An average of 18 customers arrive per hour according to a Poisson process and go to the first empty checkout line. If no checkout line is empty, suppose that arriving customers form a single line to wait for the next free checkout line. The checkout time for each customer is exponentially distributed with mean 4 minutes. It costs \$20 per hour to operate a checkout line and Safeway estimates that it costs them \$0.25 for each minute a customer waits in the cash register area. How many registers should the store have open?

The number S of registers open should certainly be enough to ensure that $\lambda/S\mu < 1$; otherwise the queue will explode. This is true only if $S > \lambda/\mu = 6/5$. Since we can only have an integer number of servers, we require S to be at least 2.

Now we want to determine the optimal number of servers; i.e., the number that minimize Safeway's total costs. We need to compute the expected costs incurred per hour.

First we consider the alternatives of either having 2 or 3 checkout lines open.

M/M/S Queue: Example

case a: two checkout lines

If there are 2 open checkout lines, then we plug S=2, λ=18 and μ=15 into the formula for K:

$$K \equiv \sum_{i=1}^{S-1} \frac{\lambda^i}{i!\,\mu^i} + \frac{\lambda^S}{S!\,\mu^S}\left(1 - \frac{\lambda}{S\mu}\right)^{-1}$$

and obtain $K = 3$. Then we can compute L_Q from:

$$L_Q = (1 + K)^{-1} \frac{(\lambda/\mu)^S}{S!} \frac{\rho}{(1-\rho)^2} \qquad \text{(Recall } \rho = \frac{\lambda}{S\mu}\text{)}$$

which yields $L_Q = 0.675$. From this we compute

$$W_Q = L_Q / \bar{\lambda} = L_Q / \lambda = 0.675/18 = 0.0375$$

From this and the fact that $W = W_Q + (1/\mu)$, we find that the average length of time a person waits in the register area is $W = 0.104$ hours, and the average number of people in the register area is $L = \bar{\lambda}W = \lambda W = 1.8738$.

Since each open checkout line costs Safeway $20 per hour, and each hour that a customer waits costs $15, the expected average costs per hour are

$$\$20S + \$15L = \$68.11$$

M/M/S Queue: Example

case b: three checkout lines

If instead there 3 servers, then after some arithmetic we find

$$K = 2.64$$

$$L_Q = 0.0879$$

$$W_Q = L_Q / \bar{\lambda} = L_Q / \lambda = 0.0879/18 = 0.0049$$

$$W = W_Q + (1/\mu) = 0.0716 \text{ hours}$$

and

$$L = \bar{\lambda}W = \lambda W = 1.2888$$

The expected average costs per hour in the 3 server case are

$$\$20S + \$15L = \$79.33$$

The additional cost of the third checkout line exceeds the benefit of shorter lines, so two checkout lines is preferable to three. In fact, 2 is the optimal number of checkout lines. (Why?)

Exponential Queuing Models: The M/M/1/C Queue

Now let's consider the variation on the M/M/1 queue in which the system has capacity *C*. In this type of queue, customers that arrive and find *C* people in the system leave. The birth and death rates are:

$$\lambda_i = \lambda \quad \text{for } i = 0,1,\ldots,C-1 \qquad \lambda_i = 0 \quad \text{for } i = C \qquad \mu_i = \mu \quad \text{for } i = 1,2,\ldots,C$$

Once again we let $\rho = \lambda/\mu$. Since the queue is finite, a steady state distribution exists regardless of whether $\rho<1$, and can be derived using Theorem 2 from CTMCs. If $\rho<1$, then the steady state distribution is:

$$P_j = \omega_j = \frac{1-\rho}{1-\rho^{C+1}}\rho^j \quad \text{for } j = 0,1,\ldots,C$$

(The case $\rho=1$ is left as an exercise for the reader.)

The average arrival rate of entering customers (when $\rho<1$) is then:

$$\bar{\lambda} = \sum_{j=0}^{C} \lambda_j P_j = \sum_{j=0}^{C-1} \lambda P_j = \lambda(1 - P_C)$$
$$= \lambda\left(1 - \frac{1-\rho}{1-\rho^{C+1}}\rho^C\right) = \lambda\left(\frac{1-\rho^C}{1-\rho^{C+1}}\right)$$

Exponential Queuing Models: The M/M/1/C Queue

The long run average number of people in the system (when $\rho<1$) is:

$$L = \sum_{j=0}^{C} jP_j = \sum_{j=0}^{C} j\omega_j = \frac{1-\rho}{1-\rho^{C+1}}\sum_{j=0}^{C} j\rho^j$$
$$= \frac{1-\rho}{1-\rho^{C+1}}\sum_{j=0}^{C} \rho\frac{d}{d\rho}(\rho^j)$$
$$= \frac{1-\rho}{1-\rho^{C+1}}\rho\frac{d}{d\rho}\left(\sum_{j=0}^{C}\rho^j\right)$$
$$= \frac{1-\rho}{1-\rho^{C+1}}\rho\frac{d}{d\rho}\left(\frac{1-\rho^{C+1}}{1-\rho}\right)$$
$$= \frac{\rho}{1-\rho} - \frac{(C+1)\rho^{C+1}}{1-\rho^{C+1}}$$

Exponential Queuing Models: The M/M/1/C Queue

The expected number of customers in the queue L_Q is:

$$L_Q = \sum_{j=1}^{\infty} (j-1) P_j = \sum_{j=1}^{\infty} (j-1)\omega_j = \sum_{j=1}^{\infty} j\omega_j - \sum_{j=1}^{\infty} \omega_j$$

$$= \sum_{j=0}^{\infty} j\omega_j - (1-\omega_0)$$

$$= L - (1-\omega_0)$$

The expected number of customers in service L_S is

$$L_S = 0 \cdot \omega_0 + \sum_{j=1}^{C} 1 \cdot \omega_j = 1 - \omega_0$$

Clearly $L = L_S + L_Q$ for this example.

From Little's law and the other identities, the quantities W, W_Q, and W_S are easily derived.

$$W_Q = L_Q / \bar{\lambda} \qquad W = W_Q + (1/\mu) \qquad L = \bar{\lambda} W$$

The M/M/1/C Queue: Example

EXAMPLE: Two Law Practices

Consider two San Francisco lawyers. Lawyer 1 works with only 1 client at a time. If a second client asks for his services while he is helping the first, he will turn that client away. He charges $20,000 per client regardless of how long the client's case takes. Lawyer 2 also helps only one client at a time, but he never turns away a client. Clients queue up to wait for his services. He charges the client he's working with a daily rate of $300 per day.

Each lawyer receives inquiries from prospective clients according to a Poisson process with rate λ=0.01 per day. Also, for each lawyer, the time to finish a case is exponentially distributed with mean 50 days.

Which lawyer makes more money?

We use the basic cost identity to determine the (long-run) average daily income of each lawyer:

average rate lawyer earns = $\bar{\lambda}$ x average amount an accepted client pays

The M/M/1/C Queue: Example

EXAMPLE: Two Law Practices

Let $\bar{\lambda}_i$ be the average arrival rate of clients and F_i be the average daily income of lawyer i, i=1,2.

Since we know for an M/M/1/C queue $\bar{\lambda} = \lambda \left(\frac{1 - \rho^C}{1 - \rho^{C+1}} \right)$ we have for lawyer 1:

$$\bar{\lambda}_1 = \lambda(1-P_1) = \lambda \left(\frac{1 - \rho}{1 - \rho^2} \right) = \lambda \left(\frac{1}{1 + \rho} \right) = \left(\frac{\lambda \mu}{\mu + \lambda} \right) = \left(\frac{0.0002}{0.03} \right)$$

$$F_1 = \bar{\lambda}_1 \times \$20{,}000 = \$133.33 \text{ per day}$$

Lawyer 2's system is an M/M/1 queue:

$$F_2 = \bar{\lambda}_2 \times (\$300/\mu) = \lambda \times (\$300/\mu) = \$150 \text{ per day}$$

Lawyer 2 makes more money.

Exponential Queuing Models with More General State Definition

In all of the examples we have seen so far, we defined the state of the system to be the number of customers in the system. This worked because in these examples, the rate of departure from the system was dependent only on the number of customers in the system. For example, in the M/M/S server queue, if the number of people in the system is j>S, all servers are busy and j-S people are in line. If j<S, then j servers are busy. Since all servers are assumed to be indistinguishable, we do not care which j servers are busy.

In some service systems, we may have multiple *nonidentical* servers. In such systems it is not sufficient to define the state as simply the number of customers in the system. The departure rates from the system depend on which specific servers are busy. We now consider two examples of exponential queues in which an extended state definition is required. We will see that all of the techniques for analyzing CTMCs and standard exponential queues still apply.

Exponential Queuing Models with More General State Definition

EXAMPLE: Stanford Post Office

Suppose the Stanford post office has 2 employees working at the counter, one faster than the other. Customers arrive according to a Poisson process with rate λ and form one line. Assume there is room for at most 4 customers in the Post Office. Server 1 has service times that are exponentially distributed with rate μ_1 and server 2 completes service in times that are exponentially distributed with rate μ_2. When both servers are free, arriving customers choose the faster server.

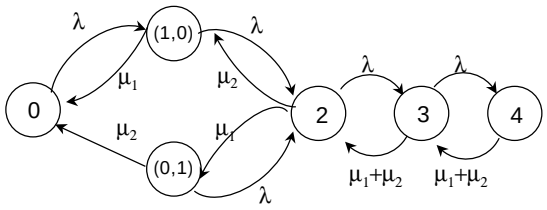
How do we define the number of people in the system in this problem? If there are 2 or more customers in the system at a particular time, then we know both servers are busy. If there are no customers in the system, then no servers are busy. But if there is exactly one customer in the system, which server is busy? We need to distinguish between the two possibilities in that case.

Define states of the process to be:

0 : no customers present
(1,0) : 1 customer present, being served by server 1
(0,1) : 1 customer present, being served by server 2
n : *n* customers present, *n*=2,3,4

Exponential Queuing Models with More General State Definition

EXAMPLE: Stanford Post Office, continued



$$Q = \begin{array}{c} \\ 0 \\ (1,0) \\ (0,1) \\ 2 \\ 3 \\ 4 \end{array} \begin{bmatrix} -\lambda & \lambda & 0 & 0 & 0 & 0 \\ \mu_1 & -(\lambda+\mu_1) & 0 & \lambda & 0 & 0 \\ \mu_2 & 0 & -(\lambda+\mu_2) & \lambda & 0 & 0 \\ 0 & \mu_2 & \mu_1 & -(\lambda+\mu_1+\mu_2) & \lambda & 0 \\ 0 & 0 & 0 & (\mu_1+\mu_2) & -(\lambda+\mu_1+\mu_2) & \lambda \\ 0 & 0 & 0 & 0 & (\mu_1+\mu_2) & -(\mu_1+\mu_2) \end{bmatrix}$$

Exponential Queuing Models with More General State Definition

EXAMPLE: Stanford Post Office, continued

$$
\begin{aligned}
0 &= -\lambda P_0 + \mu_1 P_{(1,0)} + \mu_2 P_{(0,1)} \\
0 &= \lambda P_0 - (\lambda + \mu_1) P_{(1,0)} + \mu_2 P_2 \\
0 &= -(\lambda + \mu_2) P_{(0,1)} + \mu_1 P_2 \\
0 &= \lambda P_{(1,0)} + \lambda P_{(0,1)} - (\lambda + \mu_1 + \mu_2) P_2 + (\mu_1 + \mu_2) P_3 \\
0 &= \lambda P_2 - (\lambda + \mu_1 + \mu_2) P_3 + (\mu_1 + \mu_2) P_4 \\
0 &= \lambda P_3 - (\mu_1 + \mu_2) P_4
\end{aligned}
$$

$$1 = P_0 + P_{(1,0)} + P_{(0,1)} + P_2 + P_3 + P_4$$

Suppose $\lambda = \mu_2 = 1$ per minute and $\mu_1 = 2$ per minute. In this case, solving with Matlab gives us:

$$(P_0, P_{(1,0)}, P_{(0,1)}, P_2, P_3, P_4) = (0.5294,\ 0.2118,\ 0.1059,\ 0.1059,\ 0.0353,\ 0.0118)$$

Exponential Queuing Models with More General State Definition

EXAMPLE: Stanford Post Office, continued

From the steady state distribution, we can answer a number of interesting questions:

(1) What proportion of time is each server busy?

server 1's proportion of busy time is $P_{(1,0)} + P_2 + P_3 + P_4 = 0.3647$
server 2's proportion of busy time is $P_{(0,1)} + P_2 + P_3 + P_4 = 0.2588$

(2) What is the average number of customers in service at the Post Office?

$$L = 0 \cdot P_0 + 1(P_{(0,1)} + P_{(0,1)}) + 2(P_2 + P_3 + P_4) = 0.6824$$

(3) What is the average time a customer spends in service at the Post Office?

$$\bar{\lambda} = \lambda(P_0 + P_{(0,1)} + P_{(1,0)} + P_2 + P_3) = 1(1 - P_4) = 0.8941$$
$$W = L / \bar{\lambda} = 0.6824 / 0.8941 = 0.7632 \quad \text{minutes}$$

Exponential Queuing Models with More General State Definition

EXAMPLE: Shoeshine Shop

Consider a shoeshine shop with 2 chairs. An entering customer goes to chair 1 to get his shoes polished. When the polishing is done, he will either go on to chair 2 to have his shoes buffed if that chair is empty, or else wait in chair 1 until chair 2 becomes empty. A potential customer will enter the shop only if chair 1 is empty. Potential customers arrive according to a Poisson process with rate λ. The service time in chair i exponentially distributed with rate μ_i for i=1,2.

We'd like to know:

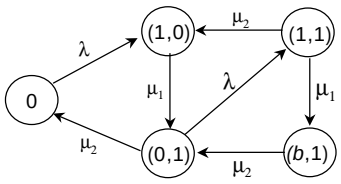
(1) What proportion of potential customers enter the shop?
(2) What is the mean number of customers in the shop?
(3) What is the average amount of time an entering customer spends in the shop?

The state of the system must include more information than simply the number of customers in the shop, because the arrival and departure rates depend on *where* in the shop the customers are.

Exponential Queuing Models with More General State Definition

EXAMPLE: Shoeshine Shop, continued

Define states of the process to be:
0 : no customers present
(1,0) : 1 customer present, in chair 1
(0,1) : 1 customer present , in chair 2
(1,1) : two customers present, both being served
(b,1) : two customers present, but the customer in chair 1 is waiting for chair 2 to become available



Exponential Queuing Models with More General State Definition

EXAMPLE: Shoeshine Shop, continued

The transition rate matrix for this process is

$$Q = \begin{array}{c} \\ 0 \\ (1,0) \\ (0,1) \\ (1,1) \\ (b,1) \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & (1,0) & (0,1) & (1,1) & (b,1) \end{array} \\ \begin{bmatrix} -\lambda & \lambda & & & \\ & -\mu_1 & \mu_1 & & \\ \mu_2 & & -(\lambda+\mu_2) & \lambda & \\ & \mu_2 & & -(\mu_1+\mu_2) & \mu_1 \\ & & \mu_2 & & -\mu_2 \end{bmatrix} \end{array}$$

which yields the steady state equations.

$$\begin{aligned} 0 &= -\lambda P_0 + \mu_2 P_{(0,1)} \\ 0 &= \lambda P - \mu_1 P_{(1,0)} + \mu_2 P_{(1,1)} \\ 0 &= \mu_1 P_{(1,0)} - (\lambda + \mu_2) P_{(0,1)} + \mu_2 P_{(b,1)} \\ 0 &= \lambda P_{(0,1)} - (\mu_1 + \mu_2) P_{(1,1)} \\ 0 &= \mu_1 P_{(1,1)} - \mu_2 P_{(b,1)} \\ P_0 &+ P_{(0,1)} + P_{(1,0)} + P_{(1,1)} + P_{(b,1)} = 1 \end{aligned}$$

Exponential Queuing Models with More General State Definition

EXAMPLE: Shoeshine Shop, continued

We are particularly interested in what proportion of potential customers enter the shop:

$$P_0 + P_{(0,1)}$$

the average number of customers in the shop:

$$L = P_{(0,1)} + P_{(1,0)} + 2(P_{(1,1)} + P_{(b,1)})$$

as well as the average time a customer spends in the shop:

$$W = L / \overline{\lambda} = \frac{P_{(0,1)} + P_{(1,0)} + 2(P_{(1,1)} + P_{(b,1)})}{\lambda(P_0 + P_{(0,1)})}$$

Exponential Queuing Models with More General State Definition

EXAMPLE: Shoeshine Shop, continued

Let's suppose that the arrival rate λ=1 customer per minute. Consider first the case that μ_1=1 and μ_2 =2, i.e., the server for chair 2 is twice as fast. In that case, the steady state equations are:

$$P_0 = 2P_{(0,1)}$$
$$P_{(1,0)} = P_0 + 2P_{(1,1)}$$
$$3P_{(0,1)} = P_{(1,0)} + 2P_{(b,1)}$$
$$P_{(0,1)} = 3P_{(1,1)}$$
$$P_{(1,1)} = 2P_{(b,1)}$$
$$P_0 + P_{(0,1)} + P_{(1,0)} + P_{(1,1)} + P_{(b,1)} = 1$$

Solution:

$$P_0 = \frac{12}{37},\ P_{(1,0)} = \frac{16}{37}$$
$$P_{(0,1)} = \frac{6}{37},\ P_{(1,1)} = \frac{2}{37}$$
$$P_{(b,1)} = \frac{1}{37}$$

From the steady state distribution, we can answer our questions:

(1) the proportion of potential customers entering the shop = $P_0 + P_{(0,1)} = \frac{18}{37}$

(2) $L = \frac{22}{37} + 2(\frac{3}{37}) = \frac{28}{37}$

(3) $W = \frac{L}{\lambda(P_0 + P_{(0,1)})} = \frac{28}{37}\frac{37}{18} = \frac{14}{9}$

Exponential Queuing Models with More General State Definition

EXAMPLE: Shoeshine Shop, continued

Now suppose that μ_1=2 and μ_2 =1, i.e., the server for chair 1 is twice as fast. In that case, the steady state equations are:

$$P_0 = P_{(0,1)}$$
$$2P_{(1,0)} = P_0 + P_{(1,1)}$$
$$2P_{(0,1)} = 2P_{(1,0)} + P_{(b,1)}$$
$$P_{(0,1)} = 3P_{(1,1)}$$
$$2P_{(1,1)} = P_{(b,1)}$$
$$P_0 + P_{(0,1)} + P_{(1,0)} + P_{(1,1)} + P_{(b,1)} = 1$$

Solution:

$$P_0 = \frac{3}{11},\ P_{(1,0)} = \frac{2}{11}$$
$$P_{(0,1)} = \frac{3}{11},\ P_{(1,1)} = \frac{1}{11}$$
$$P_{(b,1)} = \frac{2}{11}$$

(1) the proportion of potential customers entering the shop = $P_0 + P_{(0,1)} = \frac{6}{11}$

(2) $L = \frac{5}{11} + 2(\frac{3}{11}) = 1$

(3) $W = \frac{L}{\lambda(P_0 + P_{(0,1)})} = 1\frac{11}{6} = \frac{11}{6}$

Having the server for the second chair server be faster leads to loss of more potential customers, but shorter average waits and fewer customers in the shop on average.

Networks of Markovian Queues

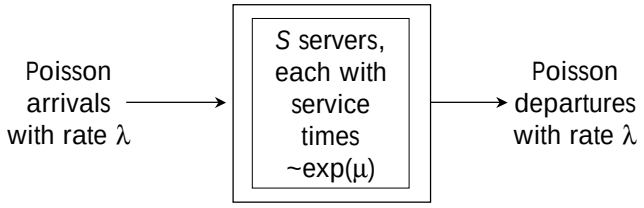
Queueing systems often provide more than a single service. Many real-world systems involve customers receiving a number of different services, each at a different service station, where each station has a queue for service. We can model systems of this kind using networks of queues. There are two types of queueing networks: open and closed. An open queueing network is one in which customers can enter and leave the system. A closed queueing network is one in which there are a fixed number of customers that never leave; new customers never enter.

We will be concerned with networks of Markovian queues called Jackson networks. These networks have exponentially distributed service times at each service station and exponentially distributed interarrival times of new customers. The sequence of stations visited is governed by a probability matrix. We'll consider first open and then closed Jackson networks. We'll start off with an important result that will help with the analysis.

The Equivalence Property

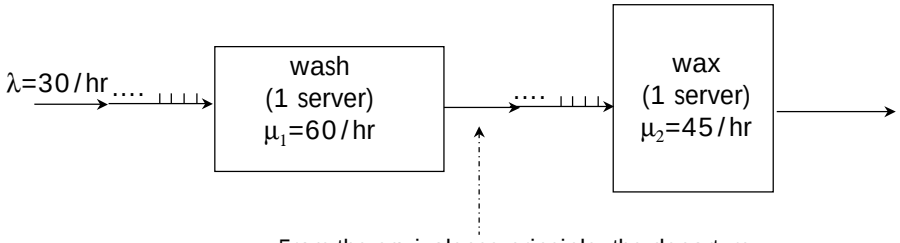
The Equivalence Property

In an M/M/S queue which is positive recurrent (i.e., $\lambda < S\mu$), the steady state output of the system is a Poisson process with the same rate as the input process. Note: S may be infinite.



Open Jackson Networks: Example

Example: Car Wash (a tandem network)



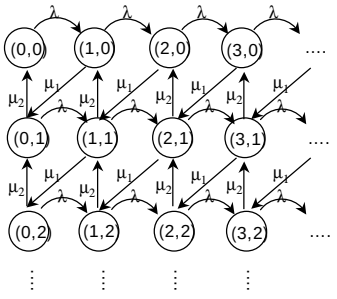
From the equivalence principle, the departure process of the wash station is a Poisson process with rate $\lambda=30/hr$ in steady state. Thus, in steady state both stations are effectively M/M/1 queues with arrival rate $\lambda=30/hr$.

Open Jackson Networks: Example

Example: Car Wash, continued

In order to analyze this system, we need to keep track of the number of cars at each station. Thus our state definition is a vector with 2 elements $X(t)=(n_1, n_2)$ where n_1 represents the number of cars being washed or in queue to be washed and n_2 represents the number of cars being waxed or in queue to be waxed.

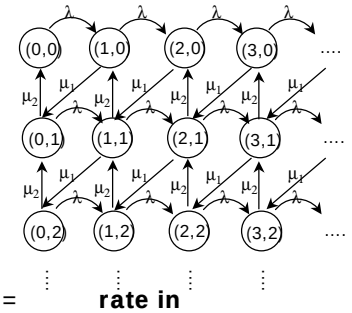
Even for such a simple network, the rate diagram is a mess.



We would like to find the steady state distribution $\omega=\{\omega_{n_1,n_2}\}$ for $X(t)=(n_1, n_2)$. To do this we would need to solve the system $\omega Q=0$, $\omega e=1$. But the Q matrix is difficult to write down.To make our lives easier, recall that the equations $\omega Q=0$ simply state that the rate into any state equals the rate out of that state.

Open Jackson Networks: Example

Example: Car Wash, continued



The equations $\omega Q=0$ are given by

state	**rate out**		**rate in**
(0,0)	$\lambda\omega_{0,0}$	=	$\mu_2\omega_{0,1}$
$(n_1,0)$; $n_1>0$	$(\lambda+\mu_1)\omega_{n_1,0}$	=	$\mu_2\omega_{n_1,1} + \lambda\omega_{n_1-1,0}$
$(0,n_2)$; $n_2>0$	$(\lambda+\mu_2)\omega_{0,n_2}$	=	$\mu_2\omega_{0,n_2+1} + \mu_1\omega_{1,n_2-1}$
(n_1,n_2); $n_1,n_2>0$	$(\lambda+\mu_1+\mu_2)\omega_{n_1,n_2}$	=	$\mu_2\omega_{n_1,n_2+1} + \mu_1\omega_{n_1+1,n_2-1} + \lambda\omega_{n_1-1,n_2}$

To compute the steady state probabilities $P_{n_1,n_2} = \omega_{n_1,n_2}$ we must solve the above equations along with

$$\sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty} P_{n_1,n_2} = \sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty} \omega_{n_1,n_2} = 1$$

Open Jackson Networks: Example

Example: Car Wash, continued

Rather than solve the steady state equations directly, let's guess at a solution. First, by the equivalence principal, we know station i is an M/M/1 queue with arrival rate λ and service rate μ_i. Assuming $\rho_i = \lambda/\mu_i < 1$ for $i=1,2$, we know

$$P(n_1 \text{ at wash station}) = \left(\frac{\lambda}{\mu_1}\right)^{n_1}\left(1-\frac{\lambda}{\mu_1}\right) = \rho_1^{n_1}(1-\rho_1)$$

$$P(n_2 \text{ at wax station}) = \left(\frac{\lambda}{\mu_2}\right)^{n_2}\left(1-\frac{\lambda}{\mu_2}\right) = \rho_2^{n_2}(1-\rho_2)$$

Now, if it were true that the number of cars at the wash and wax stations at any time were independent random variables, then we would know that

$$\begin{aligned} P_{n_1,n_2} &= P(n_1 \text{ at wash station and } n_2 \text{ at wax station}) \\ &= P(n_1 \text{ at wash station})\,P(n_2 \text{ at wax station}) = \rho_1^{n_1}(1-\rho_1)\rho_2^{n_2}(1-\rho_2) \end{aligned}$$

Open Jackson Networks: Example

Example: Car Wash, continued

As it turns out, our guess is correct. Verify for yourself that $\omega_{n_1,n_2} = P_{n_1,n_2}$ given by

$$P_{n_1,n_2} = \rho_1^{n_1}(1-\rho_1)\rho_2^{n_2}(1-\rho_2)$$

satisfies the steady state equations

state	**rate out**		**rate out**
(0,0)	$\lambda P_{0,0}$	=	$\mu_2 P_{0,1}$
$(n_1,0)$; $n_1>0$	$(\lambda+\mu_1)P_{n_1,0}$	=	$\mu_2 P_{n_1,1} + \lambda P_{n_1-1,1}$
$(0,n_2)$; $n_2>0$	$(\lambda+\mu_2)P_{0,n_2}$	=	$\mu_2 P_{0,n_2+1} + \mu_1 P_{1,n_2-1}$
(n_1,n_2); $n_1,n_2>0$	$(\lambda+\mu_1+\mu_2)P_{n_1,n_2}$	=	$\mu_2 P_{n_1,n_2+1} + \mu_1 P_{n_1+1,n_2-1} + \lambda P_{n_1-1,n_2}$

and $\sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty} P_{n_1,n_2} = 1$

Open Jackson Networks: Example

Example: Car Wash, continued

The result $P_{n_1,n_2} = \rho_1^{n_1}(1-\rho_1)\rho_2^{n_2}(1-\rho_2)$

is called the product form solution for the steady state distribution of the state of the queueing network.

What is its significance? It tells us that the numbers of cars at the wash and wax stations at any given time are, in fact, independent random variables!

This result depends heavily on the fact that the queue in front of each station has infinite capacity. If not, the numbers of cars at the two stations would not be independent.

Open Jackson Networks: Example

Example: Car Wash, continued
The product form solution of the the steady state distribution makes it easy to calculate the expected number of cars in the system:

$$L = \sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty}(n_1+n_2)P_{n_1,n_2} = \sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty}(n_1+n_2)\rho_1^{n_1}(1-\rho_1)\rho_2^{n_2}(1-\rho_2)$$

$$= \sum_{n_1=0}^{\infty} n_1\rho_1^{n_1}(1-\rho_1)\sum_{n_2=0}^{\infty}\rho_2^{n_2}(1-\rho_2) + \sum_{n_2=0}^{\infty} n_2\rho_2^{n_2}(1-\rho_2)\sum_{n_1=0}^{\infty}\rho_1^{n_1}(1-\rho_1)$$

$$= \sum_{n_1=0}^{\infty} n_1\rho_1^{n_1}(1-\rho_1) + \sum_{n_2=0}^{\infty} n_2\rho_2^{n_2}(1-\rho_2)$$

$$= (1-\rho_1)\frac{\rho_1}{(1-\rho_1)^2} + (1-\rho_2)\frac{\rho_2}{(1-\rho_2)^2}$$

$$= \frac{\rho_1}{1-\rho_1} + \frac{\rho_2}{1-\rho_2}$$

$$= \frac{\lambda}{\mu_1-\lambda} + \frac{\lambda}{\mu_2-\lambda}$$

$$= \frac{30}{30} + \frac{30}{15} = 3$$

Notice that this equals the sum of the expected numbers of cars at the two stations, respectively.

Open Jackson Networks: Example

Example: Car Wash, continued

Using this last result and Little's Law, we can compute the average time a car spends at the car wash:

$$W = \frac{L}{\lambda} = \frac{L}{\lambda} = \frac{1}{\mu_1-\lambda} + \frac{1}{\mu_2-\lambda}$$

$$= \frac{1}{30} + \frac{1}{15} = \frac{1}{10}\ \text{hour} = 6\ \text{min}$$

Notice that this equals the sum of the average time a car spends at each of the two stations, respectively.

What made the computation of L and W easy in this example was the product form solution of the steady state distribution. It turns out that there is a much more general framework under which a product form solutions are available. These are called Jackson networks.

Open Jackson Networks

An **Open Jackson Network** is a network of m service stations, where station i has

(1) an infinite queue

(2) customers arriving from outside the system according to a Poisson process with rate a_i

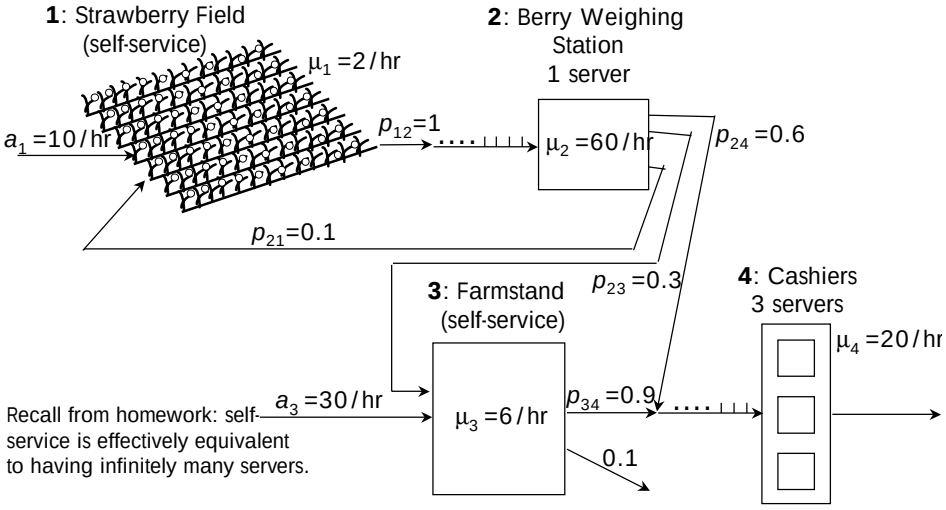
(3) s_i identical servers, each with an exponential service time distribution with rate μ_i

(4) the probability that a customer exiting station i goes to station j is p_{ij}

(5) a customer exiting station i departs the system with probability $1 - \sum_{j=1}^{m} p_{ij}$

Open Jackson Networks: Example

Example: Ward's Berry Farm



Open Jackson Networks: Traffic Equations

Define:

λ_i = total arrival rate to station i
(including external and internal arrivals)

The values $\{\lambda_i, i=1,2\ldots,m\}$ must satisfy the following equations:

$$\lambda_i = a_i + \sum_{j=1}^{m} \lambda_j p_{ji} \quad \text{for } i = 1,2,\ldots,m$$

total arrival rate into station i — external arrival rate into station i — total arrival rate from other stations

Traffic Equations for Open Jackson Networks

$$\lambda_i = a_i + \sum_{j=1}^{m} \lambda_j p_{ji} \quad \text{for } i = 1,2,\ldots,m$$

Open Jackson Networks: Example

Example: Ward's Berry Farm, continued
In this example, our traffic equations look like:

$$\lambda_1 = a_1 + \sum_{j=1}^{m} \lambda_j p_{j1} = 10 + (0.1)\lambda_2$$

$$\lambda_2 = a_2 + \sum_{j=1}^{m} \lambda_j p_{j2} = \lambda_1$$

$$\lambda_3 = a_3 + \sum_{j=1}^{m} \lambda_j p_{j3} = 30 + (0.3)\lambda_2$$

$$\lambda_4 = a_4 + \sum_{j=1}^{m} \lambda_j p_{j4} = (0.6)\lambda_2 + (0.9)\lambda_3$$

Solving these simultaneously gives us:

$$\lambda_1 = \lambda_2 = \frac{100}{9}$$

$$\lambda_3 = 30 + \frac{3}{10}\frac{100}{9} = \frac{100}{3}$$

$$\lambda_4 = \frac{6}{10}\frac{100}{9} + \frac{9}{10}\frac{100}{3} = \frac{110}{3}$$

Open Jackson Networks: The Product Form Solutions

The state of the system is a vector of length m:

$X(t)=(n_1, n_2, n_3, \ldots, n_m)$ means n_1 customers at station 1,
n_2 customers at station 2,
…
n_m customers at station m.

It turns out that if $\lambda_i / s_i\mu_i < 1$ for each station $i=1,2,\ldots,m$, then, as in the car wash example, the steady state distribution has a product form solution:

Product Form Solution for Open Jackson Networks

$$P_{n_1,n_2,\ldots,n_m} = P(n_1 \text{ at station } 1,\ n_2 \text{ at station } 2,\ldots\ n_m \text{ at station } m)$$
$$= P(n_1 \text{ at station } 1)\,P(n_2 \text{ at station } 2)\cdots\ P(n_m \text{ at station } m)$$

Open Jackson Networks: Example

Example: Ward's Berry Farm, continued
To determine the steady state distribution for the entire network in this example, we first recall what we know about the M/M/1, M/M/3 and M/M/∞ queues to determine whether each of the 4 stations has a steady state distribution of its own:

stations 1 and 3 have steady state distributions because the M/M/∞ queue always does!

station 2: $\lambda_2 = (100/9) < \mu_2 = 60$ ⇨ station 2 has a steady state distribution

station 4: $\lambda_4 = (110/3) < s_4\mu_4 = 3(20) = 60$ ⇨ station 4 has a steady state distribution

Open Jackson Networks: Example

Example: Ward's Berry Farm, continued
We compute each station's own steady state distribution using what we already know about the M/M/1, M/M/3 and M/M/∞ queues:

$$P(j \text{ at station } 1) = \frac{(\lambda_1/\mu_1)^j}{j!} e^{-\lambda_1/\mu_1}$$

$$P(j \text{ at station } 2) = (1-\lambda_2/\mu_2)(\lambda_2/\mu_2)^j$$

$$P(j \text{ at station } 3) = \frac{(\lambda_3/\mu_3)^j}{j!} e^{-\lambda_3/\mu_3}$$

$$P(j \text{ at station } 4) = \begin{cases} \dfrac{(\lambda_4/\mu_4)^j}{j!}(1+K)^{-1} & \text{for } 0 \le j \le 3 \\ \dfrac{(\lambda_4/\mu_4)^j}{3!\,3^{j-3}}(1+K)^{-1} & \text{for } j \ge 4 \end{cases}$$

$$\text{where } K = \sum_{i=0}^{2} \frac{(\lambda_4/\mu_4)^i}{i!} + \frac{(\lambda_4/\mu_4)^3}{3!}\left(\frac{1}{1-(\lambda_4/3\mu_4)}\right)$$

Open Jackson Networks: Example

Example: Ward's Berry Farm, continued
We express the steady state distribution of the entire network in product form:

$$P_{n_1,n_2,n_3,n_4}$$
$$= P(n_1 \text{ at station } 1)\,P(n_2 \text{ at station } 2)\,P(n_3 \text{ at station } 3)\,P(n_4 \text{ at station } 4)$$

$$= \frac{(\lambda_1/\mu_1)^{n_1}}{n_1!} e^{-\lambda_1/\mu_1} \cdot (1-\lambda_2/\mu_2)(\lambda_2/\mu_2)^{n_2}$$

$$\cdot \frac{(\lambda_3/\mu_3)^{n_3}}{n_3!} e^{-\lambda_3/\mu_3} \cdot \begin{cases} \dfrac{(\lambda_4/\mu_4)^{n_4}}{n_4!}(1+K)^{-1} & \text{if } 0 \le n_4 \le 3 \\ \dfrac{(\lambda_4/\mu_4)^{n_4}}{3!\,3^{n_4-3}}(1+K)^{-1} & \text{if } n_4 \ge 4 \end{cases}$$

This looks cumbersome but is easily evaluated for any particular state vector (n_1, n_2, n_3, n_4). An example is (n_1, n_2, n_3, n_4) = (0,0,0,0):

$$P_{0,0,0,0} = e^{-\lambda_1/\mu_1}(1-\lambda_2/\mu_2)e^{-\lambda_3/\mu_3}(1+K)^{-1}$$
$$= e^{-5.556}(1-0.185)e^{-5.556}(1+7.11)^{-1}$$
$$= 1.4962 \times 10^{-6}$$

Open Jackson Networks

Steps to Analyze an Open Jackson Network

(1) Solve the traffic equations $\lambda_i = a_i + \sum_{j=1}^{m} \lambda_j p_{ji} \quad \text{for } i = 1,2,\ldots,m$

(2) Check that $\lambda_i < s_i \mu_i$ for each station i=1,…,m.
If **NO**, the number of customers in the network blows up,
so there is no steady state distribution.
If **YES**, go to step 3.

(3) For each station i, calculate the steady state distribution $\omega^i = (\omega_0^i, \omega_1^i, \omega_2^i, \ldots)$
for an M/M/s_i queue with arrival rate λ_i and the service rate μ_i.

(4) The steady state probability that there are n_1 customers at station 1,
n_2 customers at station 2,…., n_m customers at station m is just

$$P_{n_1,n_2,\ldots,n_m} = P(n_1 \text{ at station } 1)\, P(n_2 \text{ at station } 2) \cdots P(n_m \text{ at station m})$$
$$= \omega^1_{n_1} \cdot \omega^2_{n_2} \cdots \omega^m_{n_m}$$

Open Jackson Networks

In the special case that all n stations have a single server (s_i=1 for all i) then the analysis is particularly easy:

(1) Solve the traffic equations $\lambda_i = a_i + \sum_{j=1}^{m} \lambda_j p_{ji} \quad \text{for } i = 1,2,\ldots,m$

(2) Check that $\rho_i = \lambda_i/\mu_i < 1$ for each station i=1,…,m.
If **NO**, there is no steady state distribution.
If **YES**, go to step 3.

(3) For each station i, the steady state distribution $\omega^i = (\omega_0^i, \omega_1^i, \omega_2^i, \ldots)$
for an M/M/1 queue with arrival rate λ_i and the service rate μ_i is:

$$\omega_j^i = (1-\rho_i)(\rho_i)^j$$

(4) The steady state probability that there are n_1 customers at station 1,
n_2 customers at station 2,…., n_m customers at station m is just

$$P_{n_1,n_2,\ldots,n_m} = \omega^1_{n_1} \cdot \omega^2_{n_2} \cdots \omega^m_{n_m} = (1-\rho_1)(\rho_1)^{n_1}(1-\rho_2)(\rho_2)^{n_2} \cdots (1-\rho_m)(\rho_m)^{n_m}$$
$$= \prod_{i=1}^{m} (1-\rho_i)(\rho_i)^{n_i}$$

Open Jackson Networks: Example

Example: Ward's Berry Farm, continued
Ward's Berry Farm makes, on average, $15 per hour that a single customer spends picking berries. What is their average hourly income from the berry field? It's

$$\$15\ L_1$$

where

$$L_1 = \frac{\lambda_1}{\mu_1} = \frac{100/9}{2} = 5.56$$

Thus, their average hourly income from the berry field alone is $83.40.

Open Jackson Networks: Example

Example: Ward's Berry Farm, continued
What is the average number of people at Ward's Berry Farm at a given point in time in steady state? By the product form solution for Open Jackson Networks, we know the answer is

$$L = L_1 + L_2 + L_3 + L_4$$

where L_i is the average number of people at station i in steady state.

$$L_1 = \frac{\lambda_1}{\mu_1} = \frac{100/9}{2} = 5.56$$

$$L_2 = \frac{\lambda_2}{(\mu_2 - \lambda_2)} = \frac{100/9}{(60 - (100/9))} = .28$$

$$L_3 = \frac{\lambda_3}{\mu_3} = \frac{100/3}{6} = 5.56$$

$$L_4 = (1 + K_4)^{-1} \frac{(\lambda_4/\mu_4)^3}{3!} \frac{\rho_4}{(1-\rho_4)^2} = .20 \quad \text{where} \quad \rho_4 = \frac{\lambda_4}{S\mu_4} = \frac{\lambda_4}{3\mu_4}$$

$$L = L_1 + L_2 + L_3 + L_4 = 11.6$$

is the average steady state number of people at the farm.

Open Jackson Networks: Example

Example: Ward's Berry Farm, continued
What is the average duration time a visiting customer spends at the farm? It turns out we can't just add average times they spend at each station, because the customer doesn't necessarily visit each station exactly once. Instead, we can just use Little's Law:

$$L = \overline{\lambda} W$$

where $\overline{\lambda}$ represents the average total external arrival rate to the entire system, i.e.,

$$\overline{\lambda} = a_1 + a_2 + a_3 + a_4 = 10 + 30 = 40$$

So a visiting customer spends, on average,

$$W = L / \overline{\lambda} = 11.6 / 40 = .29$$ hours at the farm.

Closed Jackson Networks

A **closed Jackson network of queues** is a network with a fixed number n of customers and m service stations, where station i has

(1) an infinite queue

(2) s_i identical servers, each with an exponential service time distribution with rate μ_i

(3) the probability that a customer exiting station i goes to station j is p_{ij}

(4) no customers exiting the system, i.e., $\sum_{j=1}^{m} p_{ij} = 1$

(5) no arrivals of customers from outside the system

Closed Jackson Networks

As in the open network case, we let

λ_i = total arrival rate to station i

where this time there are no external arrivals.

Now the values $\{\lambda_i, i=1,2\ldots,m\}$ must satisfy:

Traffic Equations for Closed Jackson Networks

$$\lambda_i = \sum_{j=1}^{m} \lambda_j p_{ji} \quad \text{for } i = 1,2,\ldots,m$$

$$\sum_{i=1}^{m} \lambda_i = 1$$

Closed Jackson Networks

Steps to Analyze a Closed Jackson Network

(1) Solve the traffic equations $\lambda_i = \sum_{j=1}^{m} \lambda_j p_{ji}$ for $i = 1,2,\ldots,m$, $\sum_{i=1}^{m} \lambda_i = 1$

If a solution exists, go to step 2.
If not, there is no steady state distribution.

(2) For each station i, calculate the steady state distribution $\omega^i = (\omega_0^i, \omega_1^i, \omega_2^i, \ldots)$ for an $M/M/s_i$ queue with arrival rate λ_i and the service rate μ_i.
(If λ_i exceeds μ_i, and s_i=1, then it's okay to use the steady state distribution anyway.)

continued....

Closed Jackson Networks

Steps to Analyze a Closed Jackson Network

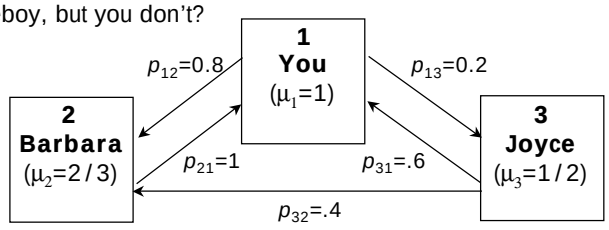
(3) The steady state probability that there are n_1 customers at station 1, n_2 customers at station 2,…., n_m customers at station m is just

$$P_{n_1,n_2,\ldots,n_m} = \begin{cases} 0 & \text{if } n_1 + n_2 + \ldots n_m \neq n \\ \dfrac{\omega_{n_1}^{1} \cdot \omega_{n_2}^{2} \cdots \omega_{n_m}^{m}}{\sum_{\substack{j_1, j_2, \ldots, j_m \geq 0 \\ j_1 + j_2 + \ldots + j_m = n}} \omega_{j_1}^{1} \cdot \omega_{j_2}^{2} \cdots \omega_{j_m}^{m}} & \text{if } n_1 + n_2 + \ldots n_m = n \end{cases}$$

Closed Jackson Networks: Example

Example: Roommate Network

Suppose that between yourself and your two roommates, you own 2 Gameboys. When you get your hands on one, you tend to use it for a time that's exponentially distributed with mean of 1 hour. When you're tired of it, you give it to your roommate Barbara with with probability 0.8 or your other roommate Joyce with probability 0.2. The lengths of time that Barbara and Joyce keep a Gameboy is exponentially distributed with mean of 1.5 hours and 2 hours, respectively. Barbara will always give the one she's been playing with to you when she's through, whereas Joyce give it to you only 60% of the time; the rest of the time she'll let Barbara have it next. What's the probability that each of your roommates has Gameboy, but you don't?



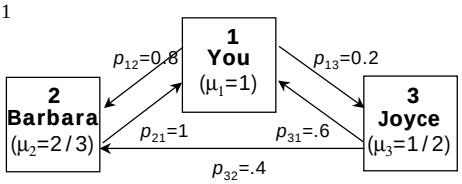
Closed Jackson Networks: Example

Example: Roommate network, continued

(1) First we solve the traffic equations to obtain the arrival rates are each "station":

$$\lambda_i = \sum_{j=1}^{m} \lambda_j p_{ji} \quad \text{for } i = 1,2,\ldots,m, \quad \sum_{i=1}^{m} \lambda_i = 1$$

$$\lambda_1 = \lambda_2 + (0.6)\lambda_3$$
$$\lambda_2 = (0.8)\lambda_1 + (0.4)\lambda_3$$
$$\lambda_3 = (0.2)\lambda_1$$
$$\lambda_1 + \lambda_2 + \lambda_3 = 1$$



The resulting arrival rates are: $\lambda_1 = \frac{25}{52}$, $\lambda_2 = \frac{22}{52}$, $\lambda_3 = \frac{5}{52}$

Closed Jackson Networks: Example

Example: Roommate network, continued

(2) We now calculate the steady state distribution for each station *i* by modeling it as an M/M/1 queue with arrival rate λ_i and the service rate μ_i.

station 1 (you): $\omega_j^1 = \left(\frac{27}{52}\right)\left(\frac{25}{52}\right)^j$

station 2 (Barbara): $\omega_j^2 = \left(\frac{19}{52}\right)\left(\frac{33}{52}\right)^j$

station 3 (Joyce): $\omega_j^3 = \left(\frac{42}{52}\right)\left(\frac{10}{52}\right)^j$

(3) Now we want to compute the steady state probability that each of your roommates has a Gameboy. This probability is denoted by $P_{0,1,1}$

Closed Jackson Networks: Example

Example: Roommate network, continued

The steady state probability that there are n_1 Gameboys with you, n_2 Gameboys with Barbara, and n_3 Gameboys with Joyce is

$$P_{n_1,n_2,n_3} = \begin{cases} 0 & \text{if } n_1 + n_2 + n_3 \neq 2 \\ \dfrac{\omega^1_{n_1} \cdot \omega^2_{n_2} \cdot \omega^3_{n_3}}{\sum_{\substack{j_1, j_2, j_3 \geq 0 \\ j_1 + j_2 + j_3 = 2}} \omega^1_{j_1} \cdot \omega^2_{j_2} \cdot \omega^3_{j_3}} & \text{if } n_1 + n_2 + n_3 = 2 \end{cases}$$

Let's first compute the denominator: $\sum_{\substack{j_1, j_2, j_3 \geq 0 \\ j_1 + j_2 + j_3 = 2}} \omega^1_{j_1} \cdot \omega^2_{j_2} \cdot \omega^3_{j_3}$

Closed Jackson Networks: Example

Example: Roommate network, continued

There are six feasible combinations of indices (j_1, j_2, j_3) that appear in the sum, namely those satisfying $j_1, j_2, j_3 \geq 0,\ j_1 + j_2 + j_3 = 2$

(j_1, j_2, j_3)	$\omega^1_{j_1} \cdot \omega^2_{j_2} \cdot \omega^3_{j_3}$
(2, 0, 0)	$\omega^1_2 \omega^2_0 \omega^3_0 = \left(\frac{27}{52}\right)\left(\frac{25}{52}\right)^2\left(\frac{19}{52}\right)\left(\frac{42}{52}\right)$
(0, 2, 0)	$\omega^1_0 \omega^2_2 \omega^3_0 = \left(\frac{27}{52}\right)\left(\frac{19}{52}\right)\left(\frac{33}{52}\right)^2\left(\frac{42}{52}\right)$
(0, 0, 2)	$\omega^1_0 \omega^2_0 \omega^3_2 = \left(\frac{27}{52}\right)\left(\frac{19}{52}\right)\left(\frac{42}{52}\right)\left(\frac{10}{52}\right)^2$
(1, 1, 0)	$\omega^1_1 \omega^2_1 \omega^3_0 = \left(\frac{27}{52}\right)\left(\frac{25}{52}\right)\left(\frac{19}{52}\right)\left(\frac{33}{52}\right)\left(\frac{42}{52}\right)$
(1, 0, 1)	$\omega^1_1 \omega^2_0 \omega^3_1 = \left(\frac{27}{52}\right)\left(\frac{25}{52}\right)\left(\frac{19}{52}\right)\left(\frac{42}{52}\right)\left(\frac{10}{52}\right)$
(0, 1, 1)	$\omega^1_0 \omega^2_1 \omega^3_1 = \left(\frac{27}{52}\right)\left(\frac{19}{52}\right)\left(\frac{33}{52}\right)\left(\frac{42}{52}\right)\left(\frac{10}{52}\right)$

we're using:

$$\omega^1_j = \left(\frac{27}{52}\right)\left(\frac{25}{52}\right)^j$$

$$\omega^2_j = \left(\frac{19}{52}\right)\left(\frac{33}{52}\right)^j$$

$$\omega^3_j = \left(\frac{42}{52}\right)\left(\frac{10}{52}\right)^j$$

Closed Jackson Networks: Example

Example: Roommate network, continued

$$\sum_{\substack{j_1, j_2, j_3 \geq 0 \\ j_1 + j_2 + j_3 = 2}} \omega_{j_1}^1 \cdot \omega_{j_2}^2 \cdot \omega_{j_3}^3 = \omega_2^1 \omega_0^2 \omega_0^3 = \left(\frac{27}{52}\right)\left(\frac{25}{52}\right)^2\left(\frac{19}{52}\right)\left(\frac{42}{52}\right)$$

$$+ \omega_0^1 \omega_2^2 \omega_0^3 = \left(\frac{27}{52}\right)\left(\frac{19}{52}\right)\left(\frac{33}{52}\right)^2\left(\frac{42}{52}\right)$$

$$+ \omega_0^1 \omega_0^2 \omega_2^3 = \left(\frac{27}{52}\right)\left(\frac{19}{52}\right)\left(\frac{42}{52}\right)\left(\frac{10}{52}\right)^2$$

$$+ \omega_1^1 \omega_1^2 \omega_0^3 = \left(\frac{27}{52}\right)\left(\frac{25}{52}\right)\left(\frac{19}{52}\right)\left(\frac{33}{52}\right)\left(\frac{42}{52}\right)$$

$$+ \omega_1^1 \omega_0^2 \omega_1^3 = \left(\frac{27}{52}\right)\left(\frac{25}{52}\right)\left(\frac{19}{52}\right)\left(\frac{42}{52}\right)\left(\frac{10}{52}\right)$$

$$+ \omega_0^1 \omega_1^2 \omega_1^3 = \left(\frac{27}{52}\right)\left(\frac{19}{52}\right)\left(\frac{33}{52}\right)\left(\frac{42}{52}\right)\left(\frac{10}{52}\right)$$

$$= 0.1824$$

Closed Jackson Networks: Example

Example: Roommate network, continued

$$P_{n_1, n_2, n_3} = \begin{cases} 0 & \text{if } n_1 + n_2 + n_3 \neq 2 \\ \dfrac{\omega_{n_1}^1 \cdot \omega_{n_2}^2 \cdot \omega_{n_3}^3}{\displaystyle\sum_{\substack{j_1, j_2, j_3 \geq 0 \\ j_1 + j_2 + j_3 = 2}} \omega_{j_1}^1 \cdot \omega_{j_2}^2 \cdot \omega_{j_3}^3} & \text{if } n_1 + n_2 + n_3 = 2 \end{cases}$$

The probability that each of your roommates has a Gameboy and you don't is

$$P_{0,1,1} = \frac{\omega_0^1 \cdot \omega_1^2 \cdot \omega_1^3}{\displaystyle\sum_{\substack{j_1, j_2, j_3 \geq 0 \\ j_1 + j_2 + j_3 = 2}} \omega_{j_1}^1 \cdot \omega_{j_2}^2 \cdot \omega_{j_3}^3} = \frac{0.0187}{0.1824} = 0.1025$$

Queuing Models with Nonexponential Service Distribution

The exponential distribution is a very convenient choice for modeling service and interarrival time distributions. As we have seen, it leads to very tractable results in characterizing system behavior. In some contexts, however, it is not a realistic choice of distribution. In some service systems, the service time might be known to have a different distribution. One important example of this situation is when the service times are completely predictable (i.e., deterministic). We will consider the following types of non-exponential queues.

- M/G/1 queue
- M/D/1 queue
- $M/E_k/1$ queue

Nonexponential Service: The M/G/1 Queue

- Poisson input process with rate λ
- single server
- arbitrary service time distribution (iid for all customers) with mean $1/\mu$ and variance σ^2
- infinite capacity queue

For this class of queues, if $\rho=\lambda/\mu<1$, then $P_0 = 1-\rho$ and $L_Q = \frac{\lambda^2\sigma^2+\rho^2}{2(1-\rho)}$

POLLACZEK-KHINTCHINE EQUATION for M/G/1

For an M/G/1 queue,

$$L_Q = \frac{\lambda^2\sigma^2+\rho^2}{2(1-\rho)} \qquad \text{if } \rho=\lambda/\mu<1.$$

Additional quantities of interest (W, W_Q, L, L_S) can be derived from L_Q and P_0.

The M/G/1 Queue: Example

EXAMPLE: McD's Drive Thru

Suppose that customers arrive at the McDonald's drive-thru at a rate of 30 per hour. The time to until any given car completes service is uniformly distributed on the interval [0,3] minutes.

(1) What is the average amount of time a car spends waiting to be served?
(2) What is the average number of cars being served or waiting to be served?

The mean service time is 1.5 minute (obtained by taking the mean of a uniform [0,3] random variable) so the average service rate μ is 2/3 cars/minute. Thus $\rho=\lambda/\mu=3/4$. The variance of the service time is $\sigma^2 = (3)^2/12 = 3/4$. The average waiting time in queue for each car, measured in minutes, is computed using the P-K equation and our useful relationship between L_Q and W_Q :

$$W_Q = \frac{L_Q}{\lambda} = \frac{L_Q}{\lambda} = \frac{\lambda^2\sigma^2 + \rho^2}{2\lambda(1-\rho)} = \frac{(0.5)^2(0.75) + (0.75)^2}{2(0.5)(0.25)} = 3 \text{ minutes}$$

The M/G/1 Queue: Example

EXAMPLE: McD's Drive Thru

The average number of cars being served or waiting to be served is

$$L = L_Q + L_S$$

Since $L_S = (1-P_0) = \rho$ for an M/G/1 queue, L equals

$$L = L_Q + L_S = \frac{\lambda^2\sigma^2 + \rho^2}{2(1-\rho)} + \rho$$
$$= \frac{(0.5)^2(0.75) + (0.75)^2}{2(0.25)} + (0.75) = 2.25 \text{ cars}$$

Nonexponential Service: The M/D/1 Queue

- Poisson input process with rate λ
- single server
- service time is deterministic and equal to $1/\mu$
- infinite capacity queue

This is a special case of the M/G/1 queue in which $\sigma^2=0$. As in the general case,if $\rho=\lambda/\mu<1$, then $P_0 = 1-\rho$. Moreover, in this case the P-K equation reduces to

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

POLLACZEK-KHINTCHINE EQUATION for M/D/1

For an M/D/1 queue,

$$L_Q = \frac{\rho^2}{2(1-\rho)} \quad \text{where } \rho=\lambda/\mu$$

The M/D/1 Queue: Example

EXAMPLE: McD's Drive Thru, again

Let's compare two different employees at the drive thru window. Ann's service time is uniformly distributed on the interval [0,3] minutes. Jim is slower on average, but more consistent: he completes service for every car in exactly 1.55 minutes every time. Which server makes customers wait less time from entry in the queue until completing service?

We've already seen that Ann's service performance leads to:

$$W_Q = \frac{L_Q}{\lambda} = \frac{L_Q}{\lambda} = \frac{\lambda^2\sigma^2+\rho^2}{2\lambda(1-\rho)} = \frac{(0.5)^2(0.75)+(0.75)^2}{2(0.5)(0.25)} = 3 \text{ minutes}$$

so $$W = W_Q + W_S = 3 + 1.5 = 4.5 \text{ minutes}$$

The M/D/1 Queue: Example

EXAMPLE: McD's Drive Thru, continued

Since Jim's service time is deterministic, the variance of his service time is zero. In his case, $\rho=\lambda/\mu=0.775$. By contrast, Jim has people wait in queue on average

$$W_Q = \frac{L_Q}{\lambda} = \frac{L_Q}{\lambda} = \frac{\rho^2}{2\lambda(1-\rho)} = \frac{(.775)^2}{2(0.5)(.225)} = 2.6694 \text{ minutes}$$

$$W = W_Q + W_S = 2.6694 + 1.55 = 4.2194 \text{ minutes}$$

Although Ann's average service time is faster, Jim's consistency leads to shorter waits.

Nonexponential Service: The M/E_k/1 Queue

Now we consider another special case of the M/G/1 queue, namely the M/ E_k/1 queue. The notation E_k stands for the Erlang distribution with shape parameter *k.* To motivate this type of queue, we first mention some important facts about the Erlang distribution.

The Erlang Distribution

A random variable X having an Erlang distribution with parameters (μ, k) has probability density function

$$f(x) = \frac{(\mu k)^k x^{k-1} e^{-k\mu x}}{(k-1)!} \quad \text{for } x \geq 0$$

and mean and variance $E(X) = \frac{1}{\mu}$ $\quad \text{Var}(X) = \frac{1}{k\mu^2}$

The Erlang (μ,k) random variable is also called a Gamma (k,μk) random variable.

Note: if $k = 1$, then X is exponentially distributed with parameter μ

The Erlang Distribution

Useful Fact About the Erlang Distribution

Let $Y_1, Y_2, \ldots, Y_n$ be independent and identically distributed exponential random variables with parameter $n\mu$. Then the random variable

$$Y = Y_1 + Y_2 + \ldots + Y_n$$

has an Erlang distribution with parameters (μ, n).

The Erlang distribution is of particular significance in queueing theory because of this useful fact. To see why, imagine a queuing system where the server (or servers) performs not just a single task that takes an exponential length of time but instead a sequence of n tasks, where each task takes an exponential length of time. When each task's time has parameter $n\mu$ (mean $1/n\mu$) then the length of time to complete all tasks has an Erlang distribution with parameters (μ, n).

Nonexponential Service: The $M/E_k/1$ Queue

- Poisson input process with rate λ
- single server
- Erlang (μ,k) service time distribution; mean $1/\mu$ and variance $1/k\mu^2$
- infinite capacity queue

As for the general M/G/1 case, if $\rho=\lambda/\mu<1$, then $P_0 = 1 - \rho$.
Applying the P-K equation to this case yields

$$L_Q = \frac{\lambda^2 / k\mu^2 + \rho^2}{2(1-\rho)} = \frac{\rho^2((1/k)+1)}{2(1-\rho)}$$

POLLACZEK-KHINTCHINE EQUATION for $M/E_k/1$

For an $M/E_k/1$ queue,

$$L_Q = \frac{\rho^2((1/k)+1)}{2(1-\rho)} \quad \text{if } \rho=\lambda/\mu<1.$$

The $M/E_k/1$ Queue: Example

EXAMPLE: Homework Questions

Students working on their EESOR 121 homework send email asking for help at a rate of 1 email per hour. When an email arrives I get right to work answering that person's email if I am not busy answering another student's email at the time. Each student's email asks one question for each of the 7 problems on the homework.) The time each question takes me to answer is exponentially distributed with mean 3 minutes, so I am capable of replying to one email every 21 minutes on average.

(1) What is the average time a student waits for a reply?

(2) If I would like to spend only 10% of of my time answering homework questions, how many homework problems should I assign each week?

The $M/E_k/1$ Queue: Example

EXAMPLE: Homework, continued

It is given that the time to answer each question is exponentially distributed with parameter 3. Since there are n=7 questions one each homework set, letting $1/n\mu$=3 implies that the time it takes me to answer one entire email has an Erlang distribution with parameters (μ, n)=(1/21,7).

In this example $\rho=\lambda/\mu = 7/20$. The P-K equation tells us

$$L_Q = \frac{\rho^2((1/n)+1)}{2(1-\rho)} = \frac{(7/20)^2((1/7)+1)}{2(13/20)} = \frac{7}{65}$$

We also know $L_S = 1(1-P_0) = \rho = \frac{7}{20}$

(1) The average time a student waits for a reply is

$$W = L/\bar{\lambda} = (L_S + L_Q)/\lambda = 60\left(\frac{7}{20} + \frac{7}{65}\right) = 60\left(\frac{119}{260}\right)$$
$$= 27.46 \text{ minutes} = 0.4577 \text{ hours}$$

The $M/E_k/1$ Queue: Example

EXAMPLE: Homework, continued

(2) Let n be the number of questions on the homework. My rate of replying to emails is then $\mu = 1/3n$ per hour. If I would like to spend at most 10% of of my time answering homework questions, then I require

$$1 - P_0 \leq 0.1$$

This holds only if $$1 - P_0 = \rho = \frac{\lambda}{\mu} = \frac{3n}{60} = \frac{n}{20} \leq 0.1$$

or $$n \leq 2$$

I should assign at most 2 problems per week.