A Brief Intro to QUEUEING THEORY

author

Kendall's Notation

Arrival Process / Service Process / Number of Servers / Capacity of System *(omitted if infinity)*

M: Memoryless (Markovian)
E_k: Erlang-k
D: Deterministic
GI: General Interarrival times (but i.i.d.)
G: General service times (but i.i.d.)

The "Memoryless" arrival process indicates a Poisson arrival process, in which the interarrival times have an ***exponential*** distribution.

Likewise, the "Memoryless" service process indicates that the service times have an ***exponential*** distribution.

LITTLE's Queueing Formula

$$L = \lambda W$$

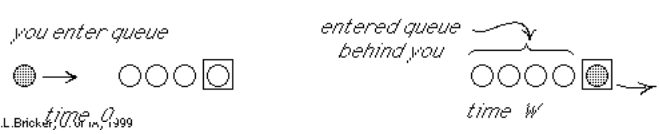
L: average number in the queueing system
λ: average arrival rate
W: average time in system per customer

☞ applies to ***any*** queueing system having a steady state distribution

LITTLE's Queueing Formula

$$L = \lambda W$$

Intuitive argument:
Suppose that you join a queue and spend **W** minutes before you have been served and leave.
During those **W** minutes, customers have been arriving and joining the queue behind you at the average rate of **λ** per minute. Thus, when you are ready to leave, you should expect to see **λW** customers remaining in the system behind you.



Most theoretical results have been obtained for the case in which both inter-arrival times and service times are ***memoryless*** (have ***exponential*** dist'n):

- ☞ M/M/1
- ☞ M/M/c (c>1)
- ☞ M/M/1/N
- ☞ M/M/1/N/N

A case in which service time is not memoryless:

- ☞ M/G/1

☞ Exercises

M/M/1

Interarrival times and service times both have exponential distributions, with parameters λ & μ, respectively.
That is, the "customers" arrive at the rate of λ per unit time, and are served at the rate μ per unit of time.
It is assumed that the queue has infinite capacity, and that $\mu > \lambda$ (so that the queue length does not tend to increase indefinitely.)
In this case, it is possible to derive the probability distribution of the number of customers in the queueing system. ↩

M/M/1

$\pi = (\pi_0, \pi_1, \pi_2, \ldots)$ denotes the "steady-state" distribution of the number of customers in this M/M/1 queueing system, i.e., 1+number in queue.
Equivalently, π_i is the probability (in steady state) that an arriving customer will find **i** customers already in the queueing system.

$$\pi_i = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^i$$

M/M/1

Using this probability distribution, we can then derive the average number of customers in the system:

$$L = \sum_{i=0}^{\infty} i\,\pi_i = \sum_{i=0}^{\infty} i\left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^i$$

$$\Longrightarrow \boxed{L = \frac{\frac{\lambda}{\mu}}{1-\frac{\lambda}{\mu}}} = \frac{\rho}{1-\rho}$$

where $\rho = \frac{\lambda}{\mu} < 1$

M/M/1

For the M/M/1 queueing system, Little's formula implies that

$$W = \frac{L}{\lambda} = \frac{\rho}{\lambda(1-\rho)}$$

$$\Longrightarrow \boxed{W = \frac{1}{\mu-\lambda}}$$

M/M/1

For the M/M/1 queueing system, then

$$W_q = W - \frac{1}{\mu} \Longrightarrow \boxed{W_q = \frac{\lambda}{\mu(\mu-\lambda)}}$$

$$L_q = \lambda W_q \Longrightarrow \boxed{L_q = \frac{\lambda^2}{\mu(\mu-\lambda)}}$$

Example

An average of 24 trucks per 8-hour day arrive to be unloaded &/or loaded, which requires an average of 15 minutes.

The loading dock can handle only a single truck at a time.

Assume that the arrival process is Poisson, and that the service times have exponential distribution.

This loading dock is modeled as an M/M/1 queue.

M/M/1

λ = arrival rate = 3/hour

μ = service rate = 4/hour

Utilization of the server $\rho = \frac{\lambda}{\mu} = 0.75$

Average number of trucks in system $L = \frac{\rho}{1-\rho} = \frac{0.75}{1-0.75} = 3$

Average time in system per truck $W = \frac{L}{\lambda} = \frac{3}{3/\text{hr}} = 1$ hr.

Steady-state Behavior

M/M/1

λ = arrival rate = 3/hour

μ = service rate = 4/hour

Average time in the queue $W_q = W - \frac{1}{\mu} = 1\text{ hr.} - \frac{1}{4/\text{hr}} = 0.75\text{hr.}$

Average length of the queue $L_q = \lambda W_q = (3/\text{hr})(0.75\text{hr}) = 2.25$

M/M/c

- *Arrival & Service processes are Memoryless, i.e.,*
  - *interarrival times have Exponential distribution with mean* $1/\lambda$
  - *service times have Exponential distribution with mean* $1/\mu$
- *Number of servers is* c
- *Capacity of queueing system is infinite*

M/M/c

If the arrival rate λ is less than the combined rate $c\mu$ at which the servers can work, then the system will have a ***steadystate*** distribution, given by:

$$\pi_0 = \frac{1}{\sum_{n=0}^{c-1}\frac{(c\rho)^n}{n!} + \frac{(c\rho)^c}{c!}\frac{1}{1-\rho}}$$

$$\pi_j = \frac{(c\rho)^j}{j!}\pi_0, \quad j=1,2,\ldots c$$

$$\pi_j = \frac{(c\rho)^j}{c!\,c^{j-c}}\pi_0, \quad j=c,c+1,\ldots$$

where $\rho = \frac{\lambda}{c\mu} < 1$

Probability that all servers are busy:

$$\sum_{j\geq c}^{\infty} \pi_j = \frac{(c\rho)^c}{c!(1-\rho)}\pi_0 \quad \text{where} \quad \rho = \frac{\lambda}{c\,\mu} < 1$$

This, then, is the probability that an arriving customer will be required to wait for service!

M/M/c Average Length of Queue

(not including those being served)

$$L_q = \sum_{j=0}^{\infty} j\,\pi_{c+j} = \sum_{j=0}^{\infty} j\,\pi_0 \frac{(c\rho)^{c+j}}{c!\,c^j} = \pi_0 \frac{(c\rho)^c}{c!} \sum_{j=0}^{\infty} j\,\rho^j$$

$$\rho = \frac{\lambda}{c\,\mu}$$

M/M/c Average Length of Queue

(not including those being served)

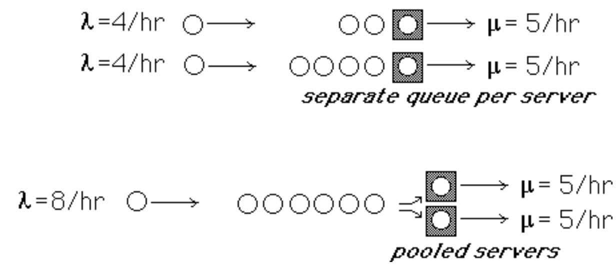
$$L_q = \frac{\rho\,(c\rho)^c}{c!}\pi_0 \left(\frac{1}{1-\rho}\right)^2$$

Once L_q is computed, then we can compute (using Little's formula)

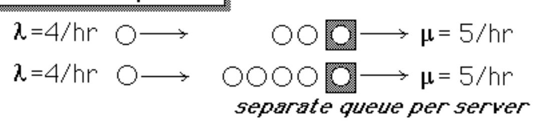
$$W_q = \frac{L_q}{\lambda}, \quad W = W_q + \frac{1}{\mu}, \ \& \quad L = \lambda W$$

Example: Pooled vs. Separate Servers

Compare two queueing systems:



two M/M/1 queues

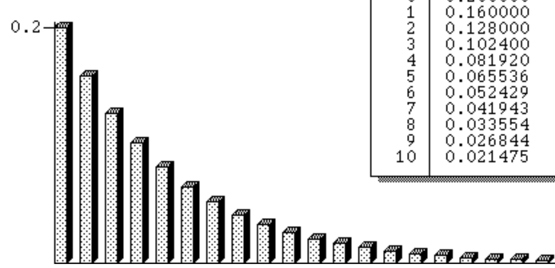


Average waiting time: $W_q = \frac{\lambda}{\mu\,(\mu - \lambda)}$

$$W_q = \frac{4/\text{hr}}{(5/\text{hr})(5-4)/\text{hr}} = 0.8\ \text{hr}$$

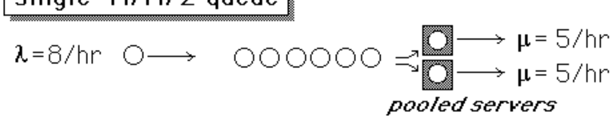
(48 minutes)

Steady-State Distribution



i	Pi	CDF
0	0.200000	0.200000
1	0.160000	0.360000
2	0.128000	0.488000
3	0.102400	0.590400
4	0.081920	0.672320
5	0.065536	0.737856
6	0.052429	0.790285
7	0.041943	0.832228
8	0.033554	0.865782
9	0.026844	0.892626
10	0.021475	0.914101

single M/M/2 queue

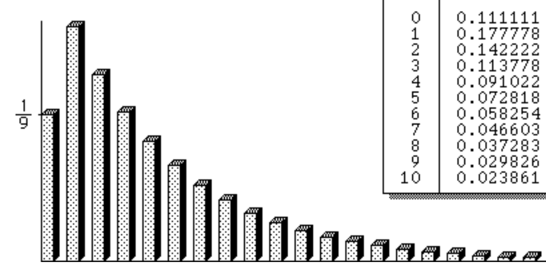


Rather than maintaining a separate queue for each server, customers enter a common queue.

$$\rho = \frac{\lambda}{2\mu} = \frac{8/\text{hr}}{2\times 5/\text{hr}} = 0.8 < 1$$

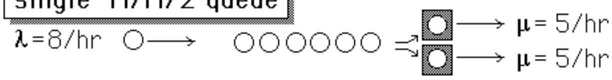
which implies that a steady state exists!

Steady-State Distribution



i	Pi	CDF
0	0.111111	0.111111
1	0.177778	0.288889
2	0.142222	0.431111
3	0.113778	0.544889
4	0.091022	0.635911
5	0.072818	0.708729
6	0.058254	0.766983
7	0.046603	0.813586
8	0.037283	0.850869
9	0.029826	0.880695
10	0.023861	0.904556

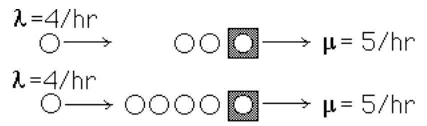
single M/M/2 queue



$$L_q = \frac{\rho}{1-\rho} P\{\text{both servers busy}\}$$

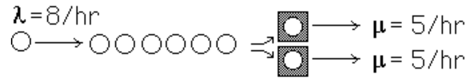
$$= \frac{0.8}{0.2}\ (0.71111111) \quad = 2.844444444$$

$$W_q = \frac{L_q}{\lambda} \quad = 0.35156 \text{ hr.} = 21.1 \text{ minutes}$$



separate queue per server

$W_q = 0.8$ hr. $= 48$ min.



pooled servers

$W_q = 0.352$ hr. $= 21.1$ min.

By pooling the servers, the average waiting time per customer is reduced by approximately 56%

M/M/1/N

- *Arrival & Service processes are **M**emoryless, i.e.,*
  interarrival times have Exponential distribution with mean $1/\lambda$
  service times have Exponential distribution with mean $1/\mu$
- *Single server*
- ***Capacity** of queueing system is **finite**:* N *(including customer currently being served)*
- *Arriving customers **balk** when queue is full.*

M/M/1/N — *Steadystate Distribution*

$$\pi_0 = \frac{1-\rho}{1-\rho^{N+1}}$$

$$\pi_j = \rho^j \pi_0 = \rho^j \left(\frac{1-\rho}{1-\rho^{N+1}}\right)$$

where $\rho = \frac{\lambda}{\mu} \neq 1$

Note that ρ *is not restricted to be less than 1 for steady state to exist!*

Average Number of Customers in System

$$L = \sum_{j=0}^{N} j\,\pi_j$$

$$L = \frac{\rho\left[1-(N+1)\rho^N + N\rho^{N+1}\right]}{(1-\rho^{N+1})(1-\rho)}$$

where $\rho = \frac{\lambda}{\mu} \neq 1$

M/M/1/N

Special Case: $\lambda = \mu$, i.e., $\rho = \frac{\lambda}{\mu} = 1$

Arrival rate = Service rate

$$\pi_j = \frac{1}{N+1}$$

$$L = \frac{N}{2}$$

All states are equally likely!

System is, on average, half-full!

Average Time in System per Customer

Little's Formula: $L = \underline{\lambda} W$ ← *average arrival rate*

$$\underline{\lambda} = \sum_{j=0}^{N-1} \lambda\,\pi_j = \lambda \sum_{j=0}^{N-1} \pi_j = \lambda\,(1-\pi_N)$$

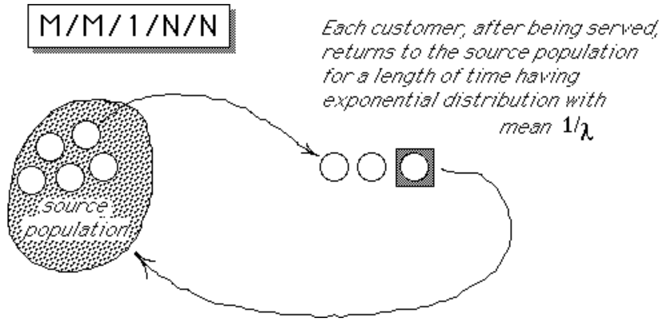
since arrival rate is zero when there are N in system

$$W = \frac{L}{\underline{\lambda}} = \frac{L}{\lambda(1-\pi_N)}$$

M/M/1/N/N

- *Single server*
- ***Finite Source Population** of size* N
- *Arrival & Service processes are **M**emoryless, i.e.,*
  service times have Exponential distribution with mean $1/\mu$
- *A departing customer returns to the queue after a time having an Exponential distribution with mean* $1/\lambda$

M/M/1/N/N



M/M/1/N/N — Steadystate Distribution

$$\pi_0 = \frac{1}{\sum_{j=0}^{N} \frac{N!}{(N-j)!}\rho^j}$$

$$\pi_j = \frac{N!}{(N-j)!}\rho^j \pi_0$$

First calculate the probability π_0 that the server is idle.

Other probabilities are then multiples of π_0

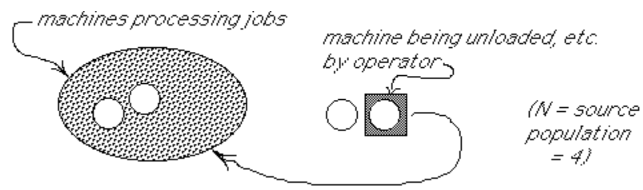
where $\rho = \frac{\lambda}{\mu}$

Example

An operator can be assigned to service (load, unload, adjust, etc.) several automatic machines in a factory

- Running time of each machine before it must be serviced has exponential distribution, with mean 120 minutes.
- Service time has an exponential distribution with mean 12 minutes.

To achieve a desired utilization of ≥ 87.5% for the machines, how many machines should be assigned to the operator?



This can be modeled as a M/M/1 queueing system with finite source population.
Machine operator = server
Machines = customers
μ = 5/hour
λ = 0.5/hour

Steadystate Distribution

$$\frac{1}{\pi_0} = \sum_{j=0}^{3} \frac{3!}{(3-j)!}(0.1)^j$$

$$= 1 + 0.3 + 0.06 + 0.006$$

$$= 1.366$$

$$\pi_0 = 1/1.366 = 0.732965$$

i.e., operator will be idle about 73% of the time!

$$\pi_1 = 0.3\,\pi_0 = 0.2196$$
$$\pi_2 = 0.06\,\pi_0 = 0.0439$$
$$\pi_3 = 0.006\,\pi_0 = 0.0044$$

$\pi_0 = 0.732965$
$\pi_1 = 0.2196$
$\pi_2 = 0.0439$
$\pi_3 = 0.0044$

If 0 machines are in system, then 3 are busy processing jobs;
if 1 machine is in system, then 2 are busy processing jobs, etc.

Average utilization of the machines will be

$$\frac{3\,\pi_0 + 2\,\pi_1 + 1\,\pi_2 + 0\,\pi_3}{3} = 89.3\%$$

M/G/1

- *Arrival process is **M**emoryless, i.e., interarrival times have Exponential distribution with mean* $1/\lambda$
- *Single server*
- *Service times are independent, identically-distributed, but not necessarily exponential. Mean service time is* $1/\mu$ *with variance* σ^2
- *Queue capacity is infinite*

↵

M/G/1 — Steadystate Characteristics

A steadystate distribution exists if $\rho = \frac{\lambda}{\mu} < 1$
i.e., if service rate exceeds the arrival rate.

$\pi_0 = 1 - \rho$ *= probability that server is idle*
$1 - \pi_0 = \rho$ *= probability that server is busy i.e., utilization of server*

There is no convenient formula for the probability of j customers in system when j > 0.

M/G/1

Steadystate Characteristics

$$L_q = \frac{\lambda^2 \sigma^2 + \rho^2}{2(1-\rho)}$$

average number of customers waiting

After calculating L_q, Little's Formula allows us to compute:

$$W_q = \frac{L_q}{\lambda}, \quad W = W_q + \frac{1}{\mu},$$

$$\& \quad L = \lambda W = L_q + \rho$$

For the M/M/1 queue, the standard deviation equals the mean service time, i.e., $\sigma = 1/\mu$ and the coefficient of variation equals 1.0

Using these formulae for the M/G/1 queueing system with $\sigma^2 = 1/\mu^2$ will give results consistent with the formulae for M/M/1.

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

$$L_q = \frac{\lambda^2 \sigma^2 + \rho^2}{2(1-\rho)}$$

Keeping the mean service time fixed, it is clear that the length of the queue is proportional to the variance of the service time.
The more ***regular*** the service time distribution, i.e., the smaller the coefficient of variation, the ***shorter*** the queue.

$$L_q = \frac{\lambda^2 \sigma^2 + \rho^2}{2(1-\rho)}$$

The average number in the queue will be ***minimized*** when the service time is ***constant***, i.e., $\sigma^2 = 0$.
In this case, the average number in the queue will be exactly ***half*** of that for the exponential dist'n:

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

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

- The UI Dept. of Public Safety has 5 patrol cars.
- A patrol car breaks down and requires service once every 30 days.
- The dept. has 2 mechanics, each of whom takes an average of 3 days to repair a car.
- Time between breakdowns & repair times have exponential distribution.

What is...
the average # of patrol cars in good condition
the average down time for a car that needs repair

A small bank is trying to determine how many tellers to employ.

- Total cost of employing a teller is $100/day.
- A teller can serve an average of 60 customers per day (i.e., 8 minutes/customer).
- An average of 50 customers per day visit the bank.
- Arrivals form a Poisson process & service times have exponential distribution.

If delay cost per customer is $100/day (i.e., about 21¢/minute), how many tellers should be employed?

An average of 40 cars/hr. are tempted to use the drive-in window at the Hot Dog King.

- If 5 cars (including the one at the window) are in line, no car will join the line.
- It takes an average of 4 minutes to serve each car (with time having exponential dist'n)

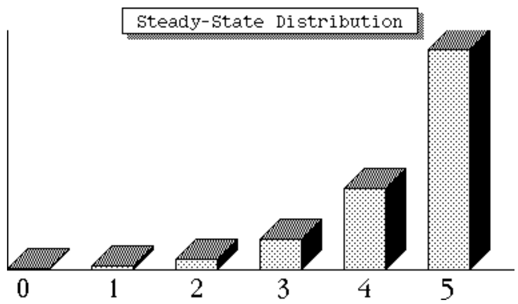
What is...
... average # of cars waiting in line?
... # cars per hour served?
... average waiting time per car?

Solution

Steady-State Distribution

i	ρ	Pi	CDF
0	2.666667	0.004648	0.004648
1	2.666667	0.012394	0.017042
2	2.666667	0.033051	0.050093
3	2.666667	0.088136	0.138228
4	2.666667	0.235029	0.373257
5	2.666667	0.626743	1.000000

The mean number of customers in the system (including the one being served) is 4.41673



- An average of 10 persons/hour arrive at the YMCA intending to swim laps.
- Each swimmer intends to swim an average of 30 minutes.
- The Y has 3 lanes open for lap swimming. Each lane can handle 2 swimmers.
- If all 3 lanes are occupied by 2 swimmers, a prospective swimmer becomes disgusted and goes running.

What fraction of the time will all 3 lanes be filled?

On the average, how many persons will be swimming?

How many lanes should be allocated to lap swimming to ensure that at most 5% of all prospective swimmers will be turned away?

- The manager of an office must decide whether to rent a second copier.
- The cost of a machine is $40 per 8-hour day, whether used or not.
- An average of 4 workers/hour need to use the copier, and each uses it for an average of 10 minutes.
- Interarrival times & copying times are exponentially distributed.
- Employees are paid $8/hour, which is assumed to be the cost to the firm of a worker waiting in line for the copier.

How many copiers should be rented?

Solution

Steady State Dist'n

servers = 1

i	ρ	Pi	CDF
0	0.666667	0.333333	0.333333
1	0.666667	0.222222	0.555556
2	0.666667	0.148148	0.703704
3	0.666667	0.098765	0.802469
4	0.666667	0.065844	0.868313
5	0.666667	0.043896	0.912209
6	0.666667	0.029264	0.941472
7	0.666667	0.019509	0.960982
8	0.666667	0.013006	0.973988
9	0.666667	0.008671	0.982658
10	0.666667	0.005781	0.988439
11	0.666667	0.003854	0.992293
12	0.666667	0.002569	0.994862
13	0.666667	0.001713	0.996575
14	0.666667	0.001142	0.997716
15	0.666667	0.000761	0.998478

 # servers = 1

Mean Queue Length (L) = 1.3333
Mean # Servers Busy = 0.66667
P{# idle servers > 1} = 0.3333

Steady State Dist'n

servers = 2

i	ρ	Pi	CDF
0	0.400000	0.666667	0.666667
1	0.400000	0.266667	0.933333
2	0.400000	0.053333	0.986667
3	0.400000	0.010667	0.997333
4	0.400000	0.002133	0.999467
5	0.400000	0.000427	0.999893
6	0.400000	0.000085	0.999979
7	0.400000	0.000017	0.999996
8	0.400000	0.000003	0.999999
9	0.400000	0.000001	1.000000
10	0.400000	0.000000	1.000000
11	0.400000	0.000000	1.000000
12	0.400000	0.000000	1.000000
13	0.400000	0.000000	1.000000
14	0.400000	0.000000	1.000000
15	0.400000	0.000000	1.000000

 # servers = 2

Mean Queue Length (L) = 0.016667
Mean # Servers Busy = 0.4
P{# idle servers > 1} = 0.93333

Steady State Dist'n

servers = 3

i	ρ	Pi	CDF
0	0.400000	0.670103	0.670103
1	0.400000	0.268041	0.938144
2	0.400000	0.053608	0.991753
3	0.400000	0.007148	0.998900
4	0.400000	0.000953	0.999853
5	0.400000	0.000127	0.999980
6	0.400000	0.000017	0.999997
7	0.400000	0.000002	1.000000
8	0.400000	0.000000	1.000000
9	0.400000	0.000000	1.000000
10	0.400000	0.000000	1.000000
11	0.400000	0.000000	1.000000
12	0.400000	0.000000	1.000000
13	0.400000	0.000000	1.000000
14	0.400000	0.000000	1.000000
15	0.400000	0.000000	1.000000

 # servers = 3

Mean Queue Length (L) = 0.0012688
Mean # Servers Busy = 0.4
P{# idle servers > 1} = 0.99175

- An automated car wash will wash a car in 10 minutes.
- Arrivals occur an average of 15 minutes apart (exponentially distributed).
- On the average, how many cars are waiting in line for the car wash?

If the car wash could be speeded up, what wash time would reduce the average wait to 5 minutes?

Solution

Steady State Dist'n

i	Pi	CDF
0	0.333333	0.333333
1	0.222222	0.555556
2	0.148148	0.703704
3	0.098765	0.802469
4	0.065844	0.868313
5	0.043896	0.912209
6	0.029264	0.941472
7	0.019509	0.960982
8	0.013006	0.973988
9	0.008671	0.982658
10	0.005781	0.988439
11	0.003854	0.992293
12	0.002569	0.994862

Mean Queue Length (L) = 1.3333
Mean number of servers busy = 0.66667
Probability that at least one server is idle = 0.33333

- Each airline passenger & his/her luggage must be checked to prevent weapons carried onto the plane.
- At the local airport, 10 passengers/minute arrive at the checkpoint.
- A checkpoint can check 12 passengers/minute (with exponential distribution).

What is the probability that an arriving passenger must wait to be checked?
What is the average time that a passenger spends at the checkpoint?

Solution

Steady State Dist'n

i	Pi	CDF
0	0.166667	0.166667
1	0.138889	0.305556
2	0.115741	0.421296
3	0.096451	0.517747
4	0.080376	0.598122
5	0.066980	0.665102
6	0.055816	0.720918
7	0.046514	0.767432
8	0.038761	0.806193
9	0.032301	0.838494
10	0.026918	0.865412
11	0.022431	0.887843
12	0.018693	0.906536
13	0.015577	0.922113
14	0.012981	0.935095
15	0.010818	0.945912

Mean Queue Length (L) = 4.1667