Lecture 21
Chapter # 6 (contd.)
FUNDAMENTALS OF QUEUING MODELS

M/M/1 Analysis (Cont'd)

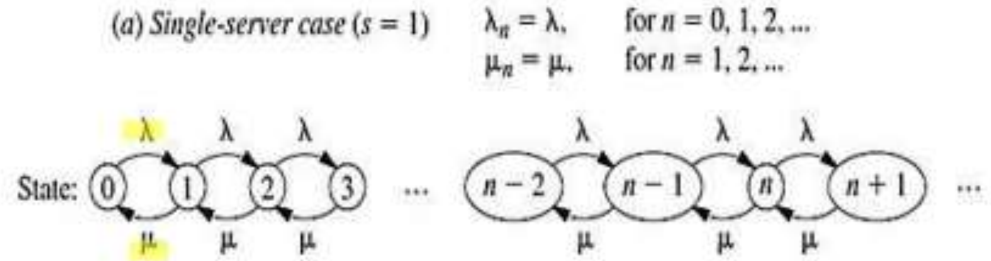
▪ Assumptions

• all inter-arrival times are independently and identically distributed according to an *exponential distribution*, all *service times* are independent and identically distributed according to *another exponential distribution* and the *number of servers* is 1.

▪ Just the special case of the birth-and-death process where

• the queuing system's *mean arrival rate* and *mean service rate* per busy server are constant (λ and μ, respectively) regardless of the state of the system.

▪ When the system has just a single server (s = 1), the implication is that the parameters for the birth-and-death process are $\lambda_n = \lambda$ (n = 0, 1, 2, ...) and $\mu_n = \mu$ (n = 1, 2, ...).


(a) Single-server case (s = 1)
λ_n = λ, for n = 0, 1, 2, ...
μ_n = μ, for n = 1, 2, ...
State:
0
1
2
3
...
n − 2
n − 1
n
n + 1
...
λ
μ

$P_0 = 1 - \rho$; ρ represents probability of busy server i.e. service utilization i.e. at least 1 customer is present in the system receiving service from the service.

Q2: What is the probability of at least k customers in the system?

P_n = probability that system is in state n = that there are n customers.

For at least k customers we can sum them starting from k to infinity.

By default, population size is infinite; there is no limit of the number of customers for the system therefore it will go till infinity.

$P_n = \rho^n P_0$

$$P[N \geq k] = \sum_{n=k}^{\infty} \overset{\rho^{n} P_{0}}{\underline{P_n}}$$
$$= \underline{\rho}^k P_0 + \rho^{k+1} P_0 + \cdots \cdot \infty$$
$$= \boldsymbol{P_0}\ (\frac{\rho^k}{1-\rho}) = (1 - \rho)\ (\frac{\rho^k}{1-\rho})$$
$$P[N \geq k] = \rho^k$$

Applying formula of some infinite geometric progression i.e. a/1-r Where a is the first term(ρ^k) and the ratio (r =ρ). Also replace p_0 =1-ρ

Q3: Determine the expected number of customers in the system.

We can determine L by little’s law as well i.e. L=λW but what if we are provided with the utilization factor. Then we will derive a relation for expected number of customers. Where N actually represents a counting process which counts the number of customers in a system i.e.

$$\begin{aligned}
L = E[N] &= \textstyle\sum_{n=0}^{\infty} n\, P_n \\
&= \textstyle\sum_{n=1}^{\infty} n\rho^{n}(1-\rho) \\
&= (1-\rho)\left(\textstyle\sum_{n=1}^{\infty} n\rho^{n}\right) \\
&= (1-\rho)\,\rho\,\frac{d}{d\rho}\left(\textstyle\sum_{n=1}^{\infty} \rho^{n}\right) \\
&= \rho(1-\rho)\,\frac{d}{d\rho}\left(\frac{\rho}{1-\rho}\right) \\
&= \rho(1-\rho)\left(\frac{1-\rho+\rho}{(1-\rho)^{2}}\right)
\end{aligned}$$

$$L = E[N] = \frac{\rho}{1-\rho}$$

$$\frac{d}{dx}x^{n} = n x^{n+1}$$

$$\frac{a}{1-r}$$

$$\frac{d}{dx}\ \frac{u}{v}$$

Prepared by: Ms. Fakhra Aftab (Lecturer, CIED, NEDUET)

Q4: Determine the expected time spent in the system. i.e. W

Starting with Little’s Law, L = λW

$$\Rightarrow W = \frac{L}{\lambda} = \frac{\rho}{1-\rho}\cdot\frac{1}{\lambda}$$

$$\Rightarrow W = \frac{\lambda/\mu}{1-\lambda/\mu}\cdot\frac{1}{\lambda}$$

$$\Rightarrow W = \frac{1}{\mu-\lambda}$$

$$\Rightarrow W = \frac{1}{\mu\,(1-\rho)} = \frac{W_s}{(1-\rho)}$$

Q5: Determine the mean waiting time.

$$\Rightarrow W_q = W - Ws$$

$$\Rightarrow \mathrm{W}_q = \frac{1}{\mu-\lambda} - \frac{1}{\mu}$$

$$\Rightarrow \mathrm{W}_q = \frac{\lambda}{\mu(\mu-\lambda)}$$

$$\Rightarrow \mathrm{W}_q = \rho\, W = \frac{\rho}{1-\rho} . \frac{1}{\mu}$$

$$\Rightarrow \mathrm{W}_q = \frac{L}{\mu}$$

Q6: Determine the expected number of customers in the queue.

$$\Rightarrow L_q = \lambda W_q$$

$$\Rightarrow L_q = \frac{\lambda}{\mu} . L$$

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

Q7: Determine the mean number of customers in the service facility.

$$\Rightarrow L_s = L - L_q$$

$$\Rightarrow L_s = \frac{\rho}{1-\rho} - \frac{\rho^2}{1-\rho}$$

$$\Rightarrow L_s = \rho$$

Q8: Give the formulae for the distribution of total time, service time & waiting time.

By hypothesis;

$W_s(t) = P\ \{s \leq t\} = 1 - e^{-\mu t} = 1 - e^{-t/Ws}$

Similarly,

Probability distribution of total time can also be evaluated using exponential distribution.

$W(t) = P\ \{W \leq t\} = 1 - e^{-t/W}$

Also,

$W_q(t) = P\ \{q \leq t\} = 1 - \rho e^{-t/W}$ (When queue discipline is FCFS)

If a customer arrives and he finds n customer already in the system, then he will have to wait n+1 exponential service time.

Let's suppose there are different independent service time random variables let say T1, T2, T3 they are actually associated with the customers that are already present in the queue. & they will first receive the service since we are actually following the FCFS queue so eventually the customers that are already present will receive their services first with the parameter μ. It means that for the customer that has just arrived it has to wait for some ample amount of time. & that amount can be represented by T1+T2+T3 till Tn+1.

This total time actually represents the conditional variant time provided that n customers are already in the system. This total time has known to have Erlang distribution. You can consider it as the generalization of the exponential distribution. Erlang has different parameters like scale shape etc.

All the derivations for M/M/1 analysis that we have performed are for steady state & the results are not applicable to the initial few customers who arrives soon after opening. Also, the results do not apply when the arrival rate or service rate varies in time as in case for example; when there's a rush of customers at a particular day of time. So, all the formula are applicable for steady-state & we will not perform any calculation for the transient state.

PASTA (Poisson Arrivals See Time Averages) Theorem

- The state of an M/M/1 queue is the number of customers in the system.
- More general queueing systems have a more general state that may include how much service each customer has already received.
- For Poisson arrivals, the arrivals in any future increment of time is independent of those in past increments and for many systems of interest, independent of the

present state S(t) (true for M/M/1, M/M/m, and M/G/1).
If we are actually considering the poison arrival, it suggests that the arriving customer will find on average the same situation in the queueing system as in outside observer looking at an arbitrary point in time. It means that the fraction of customers finding on arrival the system at some state A is exactly the same as fraction of time the system is in state A. in simple words the poison arrival will always see the steady-state probabilities.

- In steady state, arrivals see steady state probabilities.

& the long-term probability that in arriving job let's suppose see k's job in the system is exactly equal to the probability of k jobs that are available in the queueing system.

M/M/1 Analysis – Example Problem

A computer system has tasks arriving on average every 0.4 sec and requires service of 0.3 sec CPU time to process each job.
Assume an exponential distribution to process each job and their random arrivals.
Find out:
a) Utilization of CPU
b) Avg. System Time
c) Avg. Waiting Time
d) Avg. number of jobs in the queue
e) Avg. number of jobs inside the system
f) Probability that there are 5 or more tasks waiting for service.
g) If inter-arrival time is reduced to 0.25 sec, what will happen to the queuing system?

Answers:

a) 0.75
b) 1.2 sec
c) 0.9 sec
d) 2.25
e) 3 tasks
f) 0.1779
g) Unstable system

Task

The time between requests to a Web server is found to approximately follow an exponential distribution with a mean time between requests of 8 msec. The time required by the server to process each request is also found to roughly follow an exponential distribution with an average service time of approximately 5 msec.

(Use M/M/1 queueing analysis)

Single web server means just a single serving facility is available in the system.

The arrival time of request is also given, the average arrival time is given & it follows the exponential distribution.

The service time distribution is also provided which is again exponential
& the mean service time is 5ms.

Be careful you are not provided with the arrival rates & service rates. You are actually provided with the time values.

a) What is the average response time observed by the users making requests on this server?

It includes the time when you actually submit your request & wait for the response. It includes waiting time & service time both.

b) How much faster must the server process request to halve this average response time?

Result achieved in part a, reduce it by 50%.

Calculate a new service time in order to reduce the response time by 50 %.

Answers:

a) 13.33 msec

b) 275 / sec