Lecture 23
Chapter # 6 (contd.)
FUNDAMENTALS OF QUEUING MODELS

QUEUING MODELS INVOLVING NONEXPONENTIAL DISTRIBUTIONS

• The assumption of exponential inter-arrival times implies that arrivals occur randomly (a Poisson input process).

So, the random arrival if an input is quite a reasonable approximation in many situations including the computer system applications but it will not work when the arrivals of inputs or the arrivals of customers are carefully scheduled or regulated. Means that when the arrivals are deterministic, we cannot use the exponential interarrival time or we can not use a poison process to represent the inputs.

• Furthermore, the actual service-time distribution frequently deviates greatly from the exponential form, particularly when the service requirements of the customers are quite similar.

So, if we have caught some regular or deterministic service distribution or if we want to provide equal or uniform distribution to all the customers its not good to use the exponential form for the service time distribution.

• Therefore, it is important to have available other queuing models that use alternative distributions.

While discussing Kendal's notation we have discussed about a part from Markovian distribution here actually listed down various others non-exponential distributions as well.

• Unfortunately, the mathematical analysis of queuing models with non-exponential distributions is much more difficult.

➢The M/G/1 Model

Assumptions

1. The queuing system has a *single server* and
2. A *Poisson input process* (exponential inter-arrival times) with a *fixed* mean arrival rate λ.

G means General probability distribution i.e.

3. The customers have *independent* service times with the *same* probability distribution.

Or There is no restriction on what the service time can be.

• In fact, it is only necessary to know (or estimate) the mean $1/\mu$ and variance σ^2 of this distribution.

We need to estimate mean value & there's variance value

▪ The readily available steady-state results for this general model are the following:

$$P_0 = 1 - \rho,$$

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

$$L = \rho + L_q,$$

$$W_q = \frac{L_q}{\lambda},$$

$$W = W_q + \frac{1}{\mu}.$$

We will be using these formulas.
These results are valid for also M/G/1 system
& considering the complexity in analyzing a model that actually permits any service time distribution it is really remarkable that such a simple formula has been obtained for finding out the number of customers waiting in the queue (L_q).

▪ The formula for L_q is one of the most important results in queuing theory because of
- its ease of use and
- the prevalence of *M/G/1* queuing systems in practice.

▪ This equation for L_q (or its counterpart for W_q)
- commonly referred to as ***Pollaczek-Khintchine formula***,
- named after two pioneers in the development of queuing theory
- who derived the formula independently in the early 1930s.

▪ The model does not provide a closed-form expression for P_n because of analytical intractability.

▪ For any fixed expected service time $1/\mu$, notice that L_q, L, W_q, and W all increase as σ^2(variance) is increased.

▪ This result is important because it indicates that
- the consistency of the server has a major bearing on the performance of the service facility—
- not just the server's average speed.

Performance not only depends on the server's average speed rather it depends upon the regulatory & the stability of the server as well.

▪ When the service-time distribution is exponential, $\sigma^2 = 1/\mu^2$, and the preceding results will reduce to the corresponding results for the M/M/1 model.

UNIFORM DISTRIBUTION

For the M/G/1 system it is very common that the service times are distributed uniformly among the customers.

- X is a uniform random variable if the PDF (probability density function) of X is

$$f_X(x) = \begin{cases} \dfrac{1}{(b-a)} & a \le x < b \\ 0 & otherwise \end{cases}$$

where the two parameters are **b** > **a**.

Theorem

- If X is a uniform random variable with parameters **a** and **b** > **a.**
- The CDF of X is;

$$F_X(x) = \begin{cases} 0 & x \le a \\ \dfrac{(x-a)}{(b-a)} & a < x \le b \\ 1 & x > b \end{cases}$$

- The expected value of X is E[X] = **(b+a)/2**
- The variance of X is Var[X] = **$(b-a)^2/12$**

Example Problem 1

Consider the following single-server queue: the inter-arrival time is exponentially distributed with a mean of 10 minutes and the service time has the uniform distribution with a maximum of 9 minutes and a minimum of 7 minutes, find out:
i.e. case of M/G/1
solve !

(i) mean wait in the queue,
(ii) mean number in the queue,
(iii)the mean wait in the system,
(iv) mean number in the system and
(v) proportion of time the server is idle.

Answers

i) 1.602
ii) 16.02 mins
iii) 24.02 mins
iv) 2.402
v) 0.2

Models having multiple servers in the queueing system;

The M/D/s Model

Where s represents servers in the queuing system.

For the service time distribution, we are using the deterministic distribution.

When the service essentially consists of the same routine task to be perform for all the customers there's tends to be little variation in service time required. & it actually demands regularity too so in such scenario we can actually use the deterministic distribution.

▪ The *M/D/s* model often provides a reasonable representation for this kind of situation,

where customers require essentially the same amount of service.

- because it assumes that all service times actually equal some fixed *constant* (the *degenerate* service- time distribution) and

  while discussing the review of probability distribution chapter we have discussed about a constant distribution as well where when a random variable is equal to a certain constant value only then the probability is going to be 1. Such a distribution is called the constant or degenerate distribution. So, for the service times we can allot a fixed quanta or fixed time for all the customers in the system.

- that we have a *Poisson* input process with a fixed mean arrival rate λ.

▪ When there is just a single server, the M/D/1 model is just the special case of the *M/G/1* model where $\sigma^2 = 0$, so that the *Pollaczek-Khintchine formula* reduces to

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

• where L, W_q, and W are obtained from L_q as just shown.

• Notice that these L_q and W_q are exactly *half* as large as those for the exponential service-time case of the *M/M/1* model, where $\sigma^2 = 1/\mu^2$, so decreasing σ^2 can greatly improve the measures of performance of a queuing system.

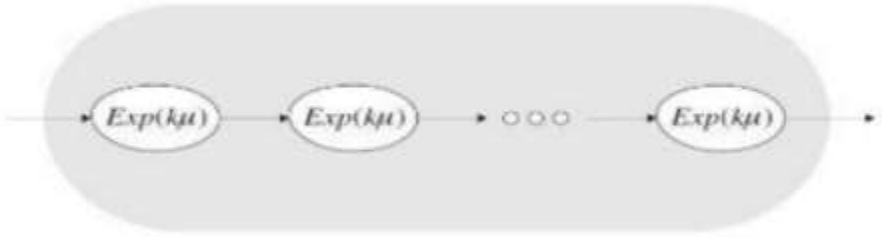
• For the multiple-server version of this model *(M/D/s)*, a complicated method is available for deriving the steady-state probability distribution of the number of customers in the system and its mean [assuming $\rho = \lambda/(s\mu) < 1$].

M/D/s or M/D/1 systems; its applications include in the wide area in network design. Let's suppose there is a single central processing unit & it is placed there to read the headers of the packets that are arrived in a random or exponential fashion & then it computes the next adaptor on which each packet should go. That's central unit could be a router as well. Here the service time is the processing of packet's header & let's suppose for checking its cyclic redundancy and these two tasks are actually independent of the length of each arriving packet.

Processing of packet header & the cyclic redundancy check would take the equal amount of time for every packet. Such a system can actually be modeled as M/D/1 queue.

Erlang-n Distribution

• Think of putting exponential distribution in series.
• If a random variable X is the sum of n-identical exponential random variables of service times n/μ, then X is said to have an Erlang-n distribution. Service time of a single server is given by 1/μ.
If we talk about Erlang-n distribution & if there are n stages then the service time is given by n/μ.
• The customer has to visit each stage in the facility to complete the service.
Let's suppose you want to get your degree & in order to get your degree you need to get few things done & that too in the series fashion. First of all, you need to get the clearance done from the library then you need to get your character certificate from the CSA office, & then you have to submit these both documents to some other departments & after that you may get your degree. So, all these services are arranged in series & in order to complete the service you have to visit each stage separately.



• A generalized Erlang Distribution is the sum of exponential random variables with different rates (also called a hypo-exponential distribution)
So, if the service rates are different for every server then such a distribution is called generalized Erlang Distribution or also called a hypo-exponential distribution.

• An Erlang random variable X with scale parameter α and n stages has probability density function:

$$f(x) = \frac{x^{n-1}e^{-x/\alpha}}{\alpha^n(n-1)!} \qquad x > 0.$$

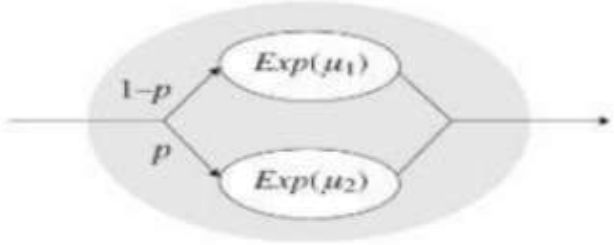
• The cumulative distribution function on the support of X is:

$$F(x) = P(X \le x) = 1 - \sum_{i=0}^{n-1} \frac{e^{-x/\alpha}x^n}{\alpha^n n!} \qquad x > 0.$$

Let's suppose we need to calculate waiting time of some repair facility then we can use this formula.
• The population mean and variance are given by E[X] = nα V [X] = $n\alpha^2$ respectively. α is given by 1/μ.

Hyper-Exponential Distribution

• Think of putting exponential distribution in parallel.



K represents number of stages here it is k=2 in the fig

• A random variable X is hyper-exponentially distributed if X is with probability p_i, i = 1, . . ., k an exponential random variable X_i with mean $1/\mu_i$. For this random variable we use the notation H_k (p_1, . . ., p_k; μ_1,..., μ_k), or simply H_k. The density is given by:

$$f(t) = \sum_{i=1}^{k} p_i \mu_i e^{-\mu_i t}, \qquad t > 0,$$

• The cumulative distribution function on the support of X is:

$$F(x) = P(X \le x) = 1 - \sum_{i=1}^{n} p_i e^{-x/\alpha_i} \qquad x > 0.$$

α is given by 1/μ.