Lecture 19
Chapter # 6 (contd.)
FUNDAMENTALS OF QUEUING MODELS

Kendall Notation

Notation: A/B/C/X/Y/Z
A = inter arrival distribution of customers.
B = service time distribution of servers (probability distribution)
System capacity means that how many customers are present in the system at a particular point in time. It includes the number of customers in the queue & the number of customers present in the server or those who are actually receiving services at that point in time.
Y = population size; it's easy to keep it infinite but finite size is also possible.
Service discipline means that which discipline queue will follow.

A	Inter-arrival Time Distribution
B	Service Time Distribution
C	No. of Servers
X	System Capacity (In queue + in server)
Y	Population Size
Z	Service Discipline

A and/or B could be:

M	Markovian/Exponential
E_k	Erlang with parameter k
G	General (Sometimes GI is used rather than G)
D	Deterministic
H_k	Hyper-exponential with parameter k

Deterministic means when the arrival pattern is regular or fixed or the service pattern is also fixed but then it won't be probabilistic/random distribution.
• A, B and C are always provided. Default values of X, Y and Z are ∞, ∞ and FIFO.
• Scheduling Policies could be pre-emptive or non-preemptive. Examples are FCFS, SIRO, RR, IS (Infinite Server) and Priority-based.
In case of infinite server, the server has infinite capacity in the sense that it can serve all the jobs simultaneously it means that no queue is formed & every arriving customers goes to the server. It's not infinite in actual sense rather we can say that

the number of servers is actually approaching the value of number of customers in some systems.

Kendall Notation Examples

1) The notation **M/D/2/∞/∞/FCFS** indicates a queuing process
• with exponential inter-arrival times, deterministic service times,
• two parallel servers, no restriction on the maximum number allowed in the system and FCFS queue discipline.

• In many situation, only the first three symbols are used.
2) Thus, M/D/2 would be a queuing system with exponential input, deterministic service, two servers, no limit on system capacity and FIFO discipline.

3) The term $M^{[x]}/M/1$ denotes a single server queue with bulk Poisson arrivals and exponential service times.
This [x] actually represents the bulk arrival.
C=1 ; queue consist of just a single server.

Practice Questions

Explain the meanings of given Kendal's Notations:

• M/M/3/20/1500/FCFS
- with exponential/poison inter-arrival times, exponential service times,
- three parallel servers and FCFS queue discipline.
- 1500 population size

there are total 20 customers in the system & out of 20, 3 customers are with the servers & remaining 17 are in the queue. This 1500 is representing the population size; it means that there would be total 1500 arriving customers that needs services.

• $E_k/G^{[x]}/5/300/5000/LCFS$
- with Erlang with parameter k inter-arrival times, bulk General service times,
- 5 parallel servers and LCFS queue discipline.
- 5000 population size

• M/G/m
denotes a m server queue with Poisson arrivals and General service times.

Basic Assumptions

• ***Job flow balance*** - The number of jobs arriving at a server within a sufficiently large observation interval must be equal to the number of jobs that depart from the server.
It means that no jobs are lost with in the server or its queue & also the server cannot spontaneously generate new jobs.
This job flow balance does not hold over shorter intervals if this is the case jobs must wait in the queue while the server is busy. However, over the long term the number of jobs that are going into the servers must match the number of jobs that actually complete the service.

• ***One-step behavior*** - At any instant of time, only a single job can enter or leave a server so that the system's state changes only incrementally.
This assumption specifically disallows jobs from moving simultaneously between servers with in the system & it also disallows the coupling of arrivals & departure of jobs. Let suppose this *one-step behavior* is true for the value of C=1 if there is just a single server but if there are multiple servers & the system actually gives multiple doors or multiple output ports for exiting of customers then it may be possible that more than 1 customer are leaving simultaneously but even then jobs are not allowed to change the servers or service mechanisms while they are in the queue or while they are actually receiving the services from some specific servers.

• ***Homogeneity*** - Homogeneity means that the average job-arrival rate and the average service rate is independent of the system's state.
It means that the rate at which the customers are arriving in the system & the rate at which the servers are providing the services are not dependent on the overall queueing system's state.

• ***Exclusivity*** - A job can be present in only a single server, either waiting in the queue or receiving service. Thus, a single job cannot make a request of two servers simultaneously.
Single job can only make a request for single server. & if the server is providing services to other customers it must wait in the queue for its turn.

• ***Non-blocking*** - The service provided by a server cannot be controlled by any other device in the system.
It means that when a job appears at the head of the queue the server must serve it immediately. It should exhibit a *non-blocking* behavior & it should provide services continuously.

• ***Independence*** - Jobs are not allowed to interact in any way, except for sharing space in a queue.

Jobs are not related to each other; the customers are not related to each other even if there processes are in the queue & they are waiting for the processors they are not allowed to actually share anything or interact with each other while they are in the queue. The only relation between the jobs in the queue is only that they are sharing space in the queue that's it.

Terminology and Notation

▪ Unless otherwise noted, the following standard terminology and notation will be used:

1. ***State of system*** = number of customers in queuing system.
2. ***Queue length*** = number of customers waiting for service to begin = state of system *minus* number of customers being served.
3. **N(t)** = number of customers in queuing system at time t ($t \geq 0$).

N(t) represents the Poison counting process

4. $\mathbf{P_n(t)}$ = probability of exactly n customers in queuing system at time t, given number at time 0.
5. **s** = number of servers (parallel service channels) in queuing system.
6. $\boldsymbol{\lambda}$ = mean arrival rate (expected number of arrivals per unit time)
7. $\boldsymbol{\mu}$ = mean service rate for overall system (expected number of customers completing service per unit time).

▪ Under these circumstances, $\mathbf{1/\lambda}$ and $\mathbf{1/\mu}$ are the expected inter-arrival time and the expected service time, respectively.

Little's Law

▪ Most widely used formula in queuing theory is:

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

▪ (Because John D. C. Little provided the first rigorous proof, this equation sometimes is referred to as *Little's formula*)

▪ Furthermore, the same proof also shows that

$$\mathbf{L_q = \lambda W_q}$$

where

L = Avg. no of customers in the system; L = E[N]

W = Avg. time spent in the system; W = E[W]

Let,

a(t): no. of arrivals by time t, therefore

$$\lambda_t = a(t)/t$$ (mean arrival rate)

Let,

g (t): total time spent by all the customers by time t, therefore

$$W_t = \frac{g(t)}{a(t)}$$

$$L_t = \frac{g(t)}{t} = \frac{g(t)}{a(t)} \times \frac{a(t)}{t}$$

$$L_t = \lambda_t W_t$$

This formula is independent of any kind of distribution of inter arrival and service times & there is no information of number of servers required or service discipline needed so Little's Law is a very general law &

• Can be applied to any block of queuing system.

It means that let suppose I want to know the average number of customers waiting in the queue so this L_q ; this q is actually representing the number of customers that are actually present in the queue right now.

$$L_q = \lambda W_q$$

If I want to know the number of customers that are present in the servers or the receiving services I can again use this formula.

$$L_s = \lambda W_s$$

Utilization Law

▪ $\rho = \lambda/(s\mu)$ is the utilization factor for the service facility, i.e.,

- the expected fraction of time the individual servers are busy,
- because **$1/(s\mu)$** represents the fraction of the system's service capacity ($s\mu$) that is being *utilized* on the average by arriving customers (λ).
- With a single server, s = 1 and $\rho = \lambda/\mu$
- For a stable queuing system, we must have $\lambda < \mu$ i.e. $\rho < 1$
  Because if the arrival rate of customers is higher than the service rate provided by the servers than there would be a bottleneck in the system & sometimes customers may be starved or they would have to wait for a very long duration for receiving the services.

Types of Queuing Analysis

Operational Analysis

• Operational analysis views the system being studied as a black box in which jobs arrive at one end, are processed for some period.

• Queueing models are studied using simple equations while making no assumptions about the probability distribution of the times between arrivals of jobs and the times required to service these jobs.

• Apply some simple laws to determine the system's overall, or average behavior.

• Examples: Little's or Utilization law

Stochastic Analysis

• If the times between the arrivals of new jobs, and the times required to service these jobs, follow certain probabilistic stochastic distributions, then detailed insight of the given system is possible.

• Jobs enter the system at times determined by the arrival process. If a server is available, the job can be serviced immediately. Otherwise, it must wait in the queue until one of the jobs currently being serviced completes. The time required to service each job also follows some assumed stochastic distribution.