CS-417
COMPUTER SYSTEMS MODELING

Spring Semester 2020

Batch: 2016-17

(LECTURE # 19)

FAKHRA AFTAB

LECTURER

DEPARTMENT OF COMPUTER & INFORMATION SYSTEMS ENGINEERING

NED UNIVERSITY OF ENGINEERING & TECHNOLOGY

Recap of Lecture # 18

Queuing Theory - Definition

Basic Queuing Process

Input Source, Queue Behavior, Queue Discipline and Service Mechanism

An Elementary Queuing Process

Chapter # 6 (Cont'd)

FUNDAMENTALS OF QUEUING MODELS

Kendall Notation

Notation: A/B/C/X/Y/Z

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

- 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.

Kendall Notation Examples

1) The notation **M/D/2/∞/∞/FCFS** indicates a queuing process

- with exponential interarrival 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.

Practice Questions

Explain the meanings of given Kendal's Notations:

- M/M/3/20/1500/FCFS
- $E_k/G^{[x]}/5/300/5000/LCFS$
- $M/G/m$

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.
- ***One-step behaviour*** - At any instant of time, only a single job can enter or leave a server so that the system's state changes only incrementally.
- ***Homogeneity*** - Homogeneity means that the average job-arrival rate and the average service rate are independent of the 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.
- ***Non-blocking*** - The service provided by a server cannot be controlled by any other device in the system.
- ***Independence*** - Jobs are not allowed to interact in any way, except for sharing space in a queue.

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 ($t \geq 0$).

4. $P_n(t)$ = 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.

Terminology and Notation

6. λ = mean arrival rate (expected number of arrivals per unit time)
7. μ = mean service rate for overall system (expected number of customers completing service per unit time).

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

Little's Law

- Most widely used formula in queuing theory is:

$$L = \lambda W$$

- (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

$$L_q = \lambda W_q$$

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

$$\lambda_t = \frac{a\,(t)}{t}$$

Little's Law (Cont'd)

Let,

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

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

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

$$L_t = \lambda_t W_t$$

- Can be applied to any block of queuing system.

$$L_q = \lambda W_q$$

$$L_s = \lambda W_s$$

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. $\rho < 1$**

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 behaviour.
- 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.

Thank You!!

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