Chapter 10

Queuing Systems

D. 10. 1. (***Queuing Theory***)
Queuing theory is the branch of operations research concerned with waiting lines.

D. 10. 2. (***Queuing System***)
A *queuing system* consists of

1. a user source
2. a queue
3. a service facility with one or more identical parallel servers.

D. 10. 3. (***Queuing Network***)
A *queuing network* is a set of interconnected queuing systems.

R. 10. 1. (***Queuing Characteristics***)

1. Arrival process
2. Service time distribution
3. Number of servers
4. System capacity
5. Population size
6. Service discipline

1. *Arrival process*
   Suppose jobs arrive at times $t_1, t_2, ..., t_j$
   - Random variables $\tau_j = t_j - t_{j-1}$ are *inter-interval times*
   - There are many possible assumptions for the distribution of the τ_j ,among others:
     i. Independent
     ii. Identically distributed
     iii. Bulk arrivals
     iv. Bulking
     v. Correlated
   - For *Poisson* arrival, the interval times are
     i. independent and identically distributed (IDD)
     ii. exponentially distributed (i.e., CDF $F(x) = 1 - e^{-x/a}$)

     (Notation: M : memoryless)
   - Other common arrival time distributions include: Erlang (E), hyper-exponential (H), deterministic (D), general (G results valid for any distribution)

2. *Service time*
    - Interval spent actually receiving service (exclusive of waiting time)
    - Most common assumptions:

        i. IID random variables
        ii. Exponential service time distribution

3. *Number of servers*

    - Servers may or may not be identical
    - Service discipline determines allocation of customers to servers

4. *System capacity*

    - Maximum number of customers in the system (including those in service) may be

        i. finite
        ii. infinite

5. *Population size*

    - Maximum number of potential customers may be

        i. finite
        ii. infinite

6. *Service discipline*

    - First-come-first serve (FCFS)
    - Last-come-first-serve (LCFS)
    - Last-come-first-serve preempt resume (LCFS-PR)
    - Round robin (RR) with quantum size
    - Processor sharing (PS) with infinitesimal quantum size (PS-RR)
    - Infinite server (IS)

R. 10. 2. (*Some Applications*)

1. Airpot check-in
2. Aircraft takeoff/landing sequence
3. Automated teller machines (ATMs)
4. Traffic analysis
5. Phone switchboard
6. Toll booths
7. Police or other spatially distributed services

R.10. 3. (*Classification of Models: Kendall Notation*)
In queuing theory, *Kendall notation* is the standard system used to describe and classify queuing models. First suggested by D. G. Kedall as a three-factor system for characterising queues, it has since been extended to six items:

$$A / S / m / B / K, SD$$

A : Arrival process
S : Service time
m : Number of servers
B : Number of buffers (system capacity)
K : Population size
SD : Service discipline

Ex. 10. 1.

$$M / M / 3 / 20 / 1500 / FCFS$$

- Time between successive arrivals is exponentially distributed.
- Service times are exponentially distributed.
- Three servers.
- 20 buffers (3 in service + 20 waiting). After 20, all arriving jobs are lost.
- Total of 1500 jobs that can be serviced.
- Service discipline is first-come-first-served.

R. 10. 4. (*Single-Channeled System with Random Arrival and S*ervice *Times*)

Specifications:

λ : Average arrival rate

μ : Average service time

$\bar{n}$: Average number in the system (including the element being serviced)

W : Average time spent in the system

W_q : Average wait before service begins

$\bar{n}_q$: Average number waiting for service to begin

L_q : Average number of customers waiting in the queue

$P(0)$: Percentage of idle time

$P(x)$: Percentage of time in which exactly x elements are in the system ($x > 0$)

Formulas:

$$\bar{n} = \frac{\frac{\lambda}{\mu}}{1 - \frac{\lambda}{\mu}}$$

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

$$W_q = \frac{\lambda}{\mu} \cdot W$$

$$\bar{n}_q = \frac{\lambda}{\mu} \cdot \bar{n}$$

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

$$P(0) = 1 - \frac{\lambda}{\mu}$$

$$P(x) = \left(\frac{\lambda}{\mu}\right)^x \left(1 - \frac{\lambda}{\mu}\right)$$

$\frac{\lambda}{\mu}$ is extremely important in the queuing theory. It is sometimes referred to as the *utilisation factor*. Sometimes a separate symbol for it is used ($\rho = \frac{\lambda}{\mu}$). It is expressed in units of "Erlangs" in honour of Danish queuing theory pioneer, A. K. Erlang.

$\rho = \frac{\lambda}{\mu}$ indicates the extreme sensitivity of the system to small changes in λ or μ when λ is close to μ.

Ex. 10. 2.

An airport, under instrument conditions, can land 12 aircraft per hour on average. Aircraft arrive into the landing pattern at the average rate of 9 per hour.
Find

1. the average number of aircraft in the system,
2. the average time spent in the system,
3. the average wait for the service to begin,
4. the average number of aircraft waiting for service to begin,
5. the percentage of the idle time,
6. the percentage of time in which exactly 1 aircraft will be served,
7. the percentage of time in which exactly 2 aircraft will be served.

Solution:

$$\lambda = 9/hr, \quad \mu = 12/hr, \quad \frac{\lambda}{\mu} = \frac{9}{12} = 0.75.$$

1.

$$\bar{n} = \frac{\frac{\lambda}{\mu}}{1 - \frac{\lambda}{\mu}} = \frac{0.75}{1 - 0.75} = 3 \text{ aircraft.}$$

2.

$$W = \frac{1}{\mu - \lambda} = \frac{1}{12 - 9} = \frac{1}{3} hr = 20 \text{ min.}$$

3.

$$W_q = \frac{\lambda}{\mu} \cdot W = 0.75 \cdot 20 = 15 \text{ min.}$$

4.

$$\bar{n}_q = \frac{\lambda}{\mu} \cdot \bar{n} = 0.75 \cdot 3 = 2.25 \text{ aircraft.}$$

5.

$$P(0) = 1 - \frac{\lambda}{\mu} = 1 - 0.75 = 0.25 = 25\%.$$

6.

$$P(1) = \left(\frac{\lambda}{\mu}\right)\left(1 - \frac{\lambda}{\mu}\right) = 0.75 \cdot 0.25 = 0.1875 = 18.75\%.$$

7.

$$P(2) = \left(\frac{\lambda}{\mu}\right)^2 \left(1 - \frac{\lambda}{\mu}\right) = 0.75^2 \cdot 0.25 \approx 0.141.$$