

UNIT 4: QUEUEING MODELS

4.1 Characteristics of Queueing System

- The key element's of queuing system are the “**customer and servers**”.
- **Term Customer:** Can refer to people, trucks, mechanics, airplanes or anything that arrives at a facility and requires services.
- **Term Server:** Refer to receptionists, repairperson, medical personal, retrieval machines that provides the requested services.

4.1.1 Calling Population

- The population of potential customers referred to as the “**calling population**”.
- The calling population may be assumed to be finite or infinite.
- The calling population is finite and consists
- In system with a large population of potential customers, the calling population is usually assumed to be infinite.
- The main difference between finite and infinite population models is how the arrival rate is defined.
- In an infinite population model, arrival rate is not affected by the number of customer who have left the calling population and joined the queueing.

4.1.2 System Capacity

- In many queueing system , there is a limit to the number of customers that may be in the waiting line or system.
- An arriving customer who finds the system full does not enter but returns immediately to the calling population.

4.1.3 Arrival Process

- The arrival process for “**Infinite population**” models is usually characterized in terms of interarrival time of successive customers.
- Arrivals may occur at scheduled times or at random times.
- When random times , the interarrival times are usually characterized by a probability distribution.
- Customer may arrive one at a time or in batches, the batches may be of constant size or random size.
- The second important class of arrivals is scheduled arrivals such as scheduled airline flight arrivals to an input.
- Third situation occurs when one at customer is assumed to always be present in the queue. So that the server is never idle because of a lack of customer.
- For finite population model, the arrivals process is characterized in a completely different fashion.
- Define customer as pending when that customer is outside the queueing system and a member of the calling population

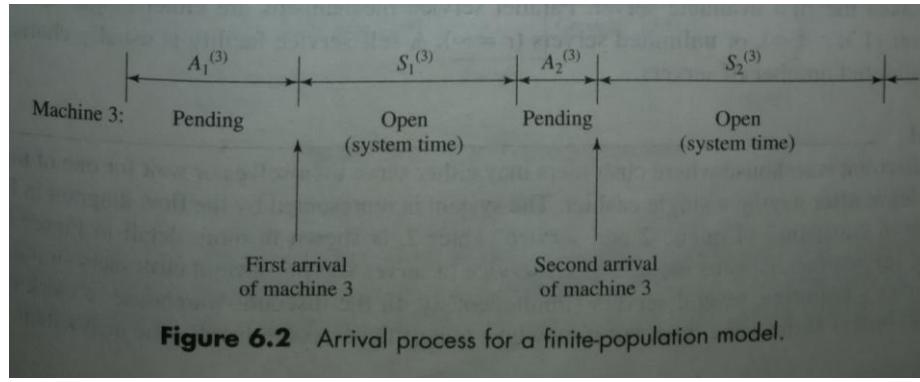


Figure 6.2 Arrival process for a finite-population model.

4.1.4 Queue Behavior and Queue Discipline

- It refers to the actions of customers while in a queue waiting for the service to begin.
- In some situations, there is a possibility that incoming customers will balk (leave when they see that the line is too long), renege (leave after being in the line when they see that the line is moving slowly), or jockey (move from one line to another if they think they have chosen a slow line).
- Queue discipline refers to the logical ordering of the customers in a queue and determines which customer will be chosen for service when a server becomes free.
- Common queue disciplines include FIFO, LIFO, service in random order (SIRO), shortest processing time first (SPT) and service according to priority (PR).

4.1.5 Service Times and Service Mechanism

- The service times of successive arrivals are denoted by s_1, s_2, s_n . They may be constant or of random duration.
- When $\{s_1, s_2, s_n\}$ is usually characterized as a sequence of independent and identically distributed random variables.
- The exponential, weibull, gamma, lognormal and truncated normal distribution have all been used successively as models of service times in different situations.
- A queueing system consists of a number of service centers and inter connecting queues. Each service center consists of some number of servers c , working in parallel.
- That is upon getting to the head of the line of customer takes the first available server.
- Parallel Service mechanisms are either single server or multiple server ($1 < c < \infty$) are unlimited servers ($c = \infty$).
- A self service facility is usually characterized as having an unlimited number of servers.

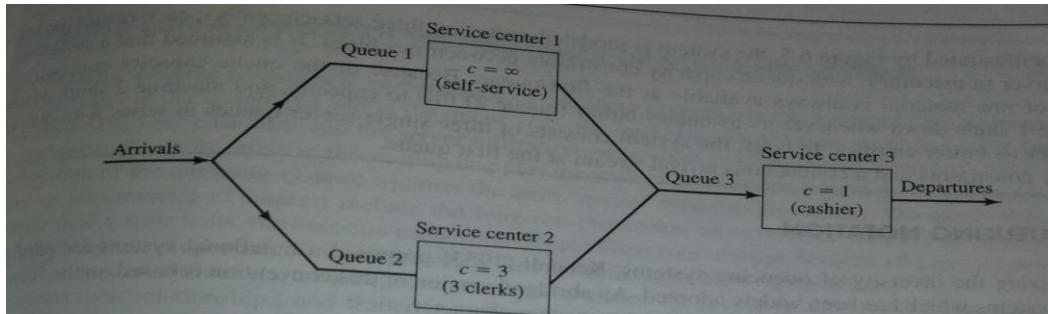


Figure 6.3 Discount warehouse with three service centers.

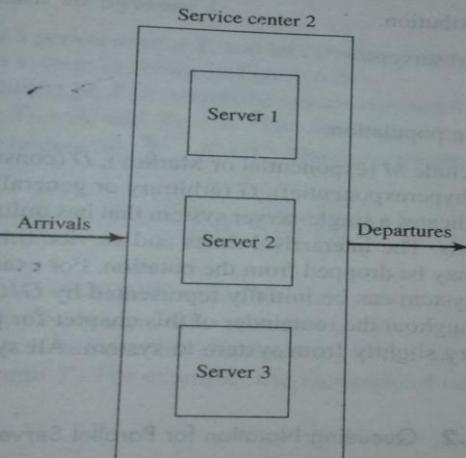


Figure 6.4 Service center 2, with $c = 3$ parallel servers.

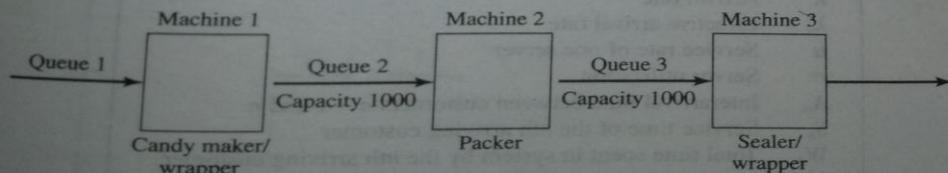


Figure 6.5 Candy-production line.

4.2 Queueing Notation(Kendal's Notation)

- Kendal's proposal a notational s/m for parallel server s/m which has been widely adopted.
- An a bridge version of this convention is based on format $A|B|C|N|K$
- These letters represent the following s/m characteristics:

- A-Represents the InterArrival Time distribution
- B-Represents the service time distribution
- C-Represents the number of parallel servers
- N-Represents the s/m capacity
- K-Represents the size of the calling populations

Common symbols for A & B include M(exponential or Markov), D(constant or deterministic), E_k (Erlang of order k), PH (phase-type), H(hyperexponential), G(arbitrary or general), & GI(general independent).

- For eg, $M|M|1|\infty|\infty$ indicates a single server s/m that has unlimited queue capacity & an infinite population of potential arrivals
- The interarrival times & service times are exponentially distributed when N & K are infinite, they may be dropped from the notation.
- For eg, $M|M|1|\infty|\infty$ is often short ended to $M|M|1$. The tire-curing s/m can be initially represented by $G|G|1|5|5$.

- Additional notation used for parallel server queueing s/m are as follows:

Table 6.2 Queueing Notation for Parallel Server Systems	
P_n	Steady-state probability of having n customers in system
$P_n(t)$	Probability of n customers in system at time t
λ	Arrival rate
λ_e	Effective arrival rate
μ	Service rate of one server
ρ	Server utilization
A_n	Interarrival time between customers $n - 1$ and n
S_n	Service time of the n th arriving customer
W_n	Total time spent in system by the n th arriving customer
W_n^Q	Total time spent in the waiting line by customer n
$L(t)$	The number of customers in system at time t
$L_Q(t)$	The number of customers in queue at time t
L	Long-run time-average number of customers in system
L_Q	Long-run time-average number of customers in queue
w	Long-run average time spent in system per customer
w_Q	Long-run average time spent in queue per customer

4.3 Long-run Measures of performance of queueing systems

- The primary long run measures of performance of queueing system are the long run time average number of customer in s/m(L) & queue(L_Q)
- The long run average time spent in s/m(w) & in the queue(w_Q) per customer
- Server utilization or population of time that a server is busy (ρ).

4.3.1 Time average Number in s/m (L):

- Consider a queueing s/m over a period of time T & let $L(t)$ denote the number of customer I the s/m at time t.
- Let T_i denote the total time during $[0, T]$ in which the s/m contained exactly i customers.

$$\hat{L} = \sum_{i=0}^{\infty} i \left(\frac{T_i}{T} \right)$$

- where \hat{L} is the time weighted average number in a system.
- Consider an example of queuing S/m with line segment 3, 12, 4, 1. Compute the time weighted - average number in a S/m.

Soln

$$\hat{L} = \sum_{i=0}^{\infty} i \left(\frac{T_i}{T} \right)$$

$$\begin{aligned}\hat{L} &= [0(3) + 1(12) + 2(4) + 3(1)] / 20 \\ &= 23/20 \\ &= 1.15 \text{ customers.}\end{aligned}$$

4.3.2 Average Time spent in s/m per customer (w):

- Average s/m time is given as:

$$\hat{w} = \frac{1}{N} \sum_{i=1}^N w_i \quad \text{--- (1)}$$

where,

N - is the number of arrivals during $[0, T]$
 w_i - is customer spend in the S/m during $[0, T]$

- For stable s/m $N \rightarrow \infty$

$$\hat{w} \rightarrow w \quad \text{--- (2)}$$

With probability 1, where w is called the long-run average s/m time.

- Considering the equation 1 & 2 are written as,

$$\hat{w}_Q = \frac{1}{N} \sum_{i=1}^N w_i^Q \rightarrow w_Q$$

where,

w_i^Q - is the total time customer i spends waiting in queue.

\hat{w}_Q - is the observed average time spent in queue.

w_Q - is the long run average delay per customer

Example :- Consider the Queueing S/m with $N=S$

Customer arrive at $w_1 = 2$ & $w_5 = 20-16=4$ but w_2, w_3 & w_4 cannot be computed unless more is known about the S/m. Arrival occur at times 0, 3, 5, 7 & 16 & departures occur at time 2, 8, 10 & 14.

Soln

$$\hat{\omega} = \frac{1}{N} \sum_{i=1}^N w_i$$

$$w_1 = 2, w_5 = 4$$

$$w_2 = \frac{8-3}{5}$$

$$w_3 = \frac{10-5}{5} = \frac{5}{5} = 1$$

$$w_4 = \frac{14-7}{7} = 1$$

$$\hat{\omega} = \frac{2 + 5 + 5 + 7 + 4}{5} = \frac{23}{5} = 4.6 \text{ time units.}$$

4.3.3 Server utilization:

- Server utilization is defined as the proportion of time server is busy
- Server utilization is denoted by β is defined over a specified time interval [0,1]
- Long run server utilization is denoted by p

$$P \rightarrow P$$

as $T \rightarrow \infty$

❖ Server utilization in $G|G|C|\infty|\infty$ queues

- Consider a queuing s/m with c identical servers in parallel
- If arriving customer finds more than one server idle the customer choose a server without favoring any particular server.
- The average number of busy servers say L_s is given by,

$$L_s = \lambda / \mu$$

$$0 \leq L_s \leq C$$

- The long run average server utilization is defined by

$$P = \frac{L_s}{C} = \frac{\lambda}{c\mu} \quad \therefore 0 \leq P \leq 1$$

- The utilization P can be interpreted as the proportion of time an arbitrary server is busy in the long run

Example :

Customer arrive at random to a license bureau at a rate of $\lambda = 50$ customer per hour. Currently there are 20 clerks, each serving $\mu = 5$ customers per hour on the average. Compute long-run or steady state average utilization of a server & average number of busy server.

Sol:

Average utilization of server :

$$P = \frac{\lambda}{\mu n}$$

$$P = \frac{50}{20(5)} = 0.5$$

Average number of busy servers is :

$$L_s = \frac{\lambda}{\mu}$$

$$L_s = \frac{50}{5} = 10$$

4.4 STEADY-STATE BEHAVIOUR OF INFINITE-POPULATION MARKOVIAN MODELS

- For the infinite population models, the arrivals are assumed to follow a poisson process with rate λ arrivals per time unit
- The interarrival times are assumed to be exponentially distributed with mean $1/\lambda$
- Service times may be exponentially distributed(M) or arbitrary(G)
- The queue discipline will be FIFO because of the exponential distributed assumptions on the arrival process, these model are called "MARKOVIAN MODEL".
- The steady-state parameter L, the time average number of customers in the s/m can be computed as

$$L = \sum_{n=0}^{\infty} n P_n$$

Where P_n are the steady state probability of finding n customers in the s/m

- Other steady state parameters can be computed readily from little equation to whole system & to queue alone

$$\boxed{\begin{aligned} w &= L/\lambda \\ wQ &= w - (1/\mu) \\ LQ &= \lambda wQ \end{aligned}}$$

Where λ is the arrival rate & μ is the service rate per server

4.4.1 SINGLE-SERVER QUEUE WITH POISSON ARRIVALS & UNLIMITED CAPACITY: M|G|1

- Suppose that service times have mean $1/\mu$ & variance σ^2 & that there is one server
- If $P = \lambda / \mu < 1$, then the M|G|1 queue has a steady state probability distribution with steady state characteristics
- The quantity $P = \lambda / \mu$ is the server utilization or long run proportion of time the server is busy
- Steady state parameters of the M|G|1 are:

Notation	Description
① $P = \frac{\lambda}{\mu}$	• P is server utilization • λ is arrival rate • μ is service rate
② $L = P + \frac{P^2(1+\sigma^2\mu^2)}{2(1-P)}$	• L is long run time average number of customer in system • σ^2 is the mean service time
③ $w = \frac{1}{\mu} + \frac{\sigma^2(1/\mu^2 + \sigma^2)}{2(1-P)}$	• w is long run average time spent in system per customer
④ $w_Q = \frac{\sigma^2(1/\mu^2 + \sigma^2)}{2(1-P)}$	• w_Q is long run average time spent in queue per customer
⑤ $L_Q = \frac{P^2(1+\sigma^2\mu^2)}{2(1-P)}$	• L_Q is long run time avg no. of customer in queue
⑥ $P_0 = 1-P$	• P_0 is steady state probability of customer in system

example : Consider a candy factory for making a candy at rate $\lambda = 1.5$ per hour. Observation over several months has found by the single mlc. It's mean service time $\bar{t} = 1/2$ hour, service rate is $\mu = 2$. Compute long run time average number of customer in S/m, long run time average number of customer in queue & long run average time spent in queue per customer.

Soln

↳ long run time average number of customer in S/m

$$L = \rho + \frac{\rho^2(1 + \sigma^2\mu^2)}{2(1-\rho)}$$

$$\rho = \frac{\lambda}{\mu} = \frac{1.5}{2} = 0.75$$

$$L = 0.75 + \frac{0.75(1 + (0.5)^2(2)^2)}{2(1 - 0.75)}$$

$$= 3.75$$

↳ long run time average number of customer in queue

$$L_Q = \frac{\rho^2(1 + \sigma^2\mu^2)}{2(1-\rho)}$$

$$L_Q = \frac{(0.75)^2(1 + (0.5)^2(2)^2)}{2(1 - 0.75)}$$

$$= 2.25$$

↳ long run average time spent in queue per customer:

$$W_Q = \frac{\lambda(1/\mu^2 + \sigma^2)}{2(1-\rho)}$$

$$W_Q = \frac{1.5(1/(2)^2 + (0.5)^2)}{2(1 - 0.75)}$$

$$= 1.5$$

- Steady state parameters of the m/m/1 queue

Notation	Description
$L = \frac{\rho}{1-\rho}$	• L is long run time average number of customer in s/m
$\omega = \frac{1}{\mu(1-\rho)}$	• ρ is server utilization • ω is long run average time spent is s/m per customer
$\omega_q = \frac{\rho}{\mu(1-\rho)}$	• μ is service rate
$L_q = \frac{\rho^2}{1-\rho}$	• ω_q is long run average time spent in queue per customer
$P_n = (1-\rho)\rho^n$	• L_q is long run time average number of customer in queue • P_n is steady state probability of n customer in s/m

4.4.2 MULTISERVER QUEUE: M|M|C| ∞ | ∞

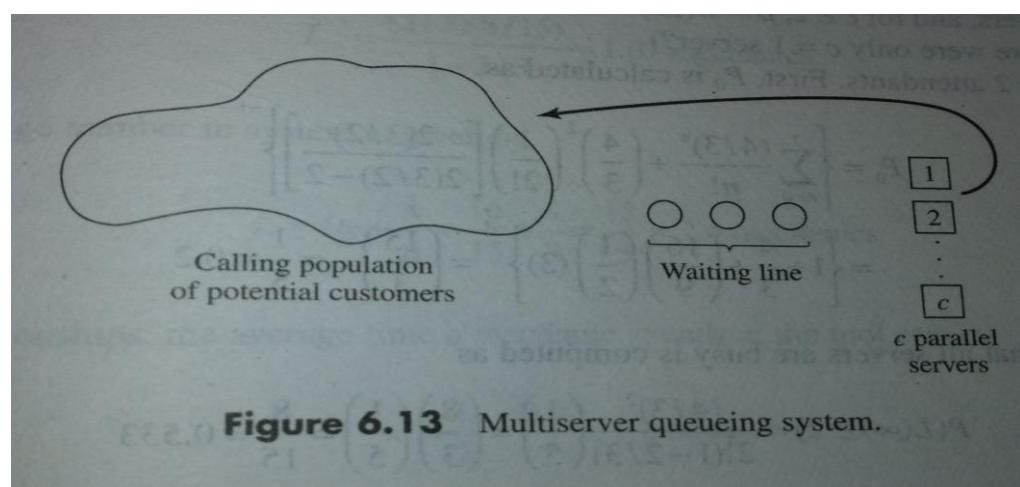


Figure 6.13 Multiserver queueing system.

- Suppose that there are c channels operating in parallel
- Each of these channels has an independent & identical exponential service time distribution with mean $1/\mu$
- The arrival process is poisson with rate λ . Arrival will join a single queue & enter the first available service channel

- For the M|M|C queue to have statistical equilibrium the offered load must satisfy $\lambda/\mu < c$ in which case $\lambda/(c\mu) = P$ the server utilization.

<u>The steady state parameters for the m m/c queue</u>	
Notation	Description
$P = \frac{\lambda}{\mu}$	<ul style="list-style-type: none"> P is server utilization λ arrival rate μ service rate
$P_0 = \left[\sum_{n=0}^{c-1} \frac{c^n \lambda^n}{n!} + \left(\frac{c \lambda}{\mu} \right)^c \left(\frac{1}{c!} \right) \right]$	<ul style="list-style-type: none"> • steady state for probability of customers in s/m
$L = CP + \frac{PPL(\infty) > c}{(1-P)}$	<ul style="list-style-type: none"> L is long run time average number of customers in s/m
$\omega = \frac{L}{\lambda}$	<ul style="list-style-type: none"> ω is long run average time spent in s/m per customer
$\omega_q = \omega - \frac{1}{\mu}$	<ul style="list-style-type: none"> ω_q is long run average time spent in queue per customer
$L_q = \frac{PPL(\infty) > c}{(1-P)}$	<ul style="list-style-type: none"> L_q is long run time average number of customers in queue
$L - L_q = CP$	

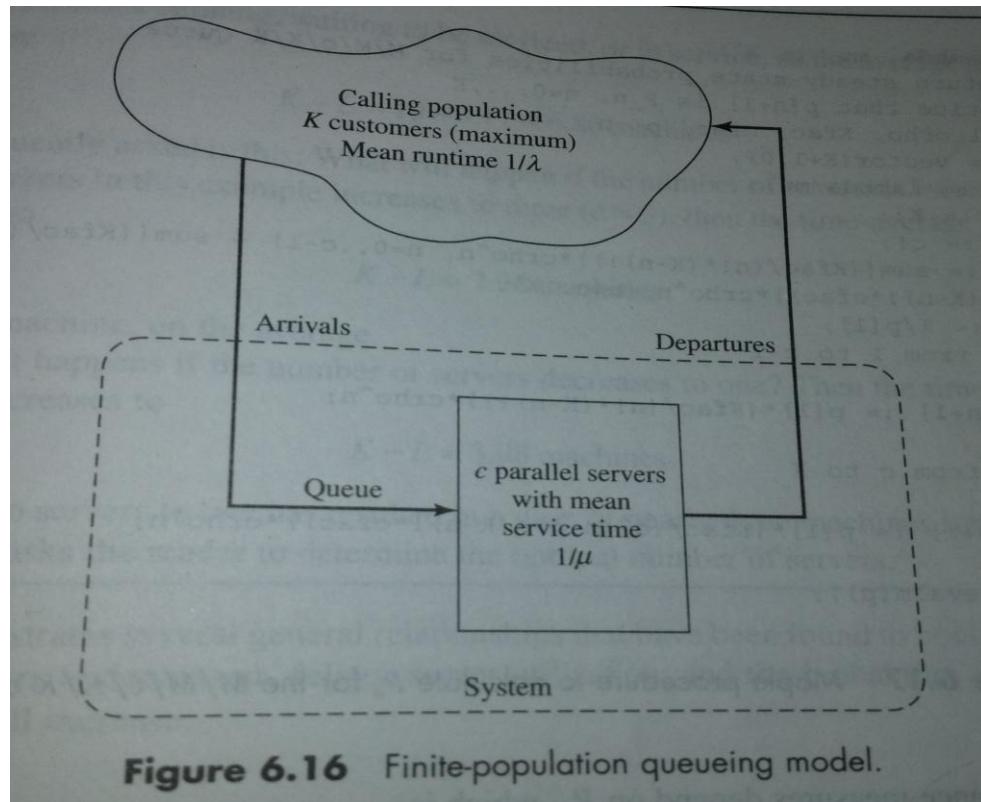
WHEN THE NUMBER OF SERVERS IS INFINITE ($M|c|\infty$)

- There are at least three situations in which it is appropriate to treat the number of servers as infinite
 - When each customer is its own server in other words in a self service s/m
 - When service capacity far exceeds service demand as in a so called ample server s/m
 - When we want to know how many servers are required so that customer will rarely be delayed.

<u>Steady state parameters for the m/g/c queue</u>	
description	
P_0	P_0 - probability of customer 0 in system
$\omega = \frac{1}{\mu}$	ω - long run average time spent in system
$w_q = 0$	w_q - long run average time spent in queue
$L = \lambda/\mu$	L - long run time average no. of customers in system
$W_i = 0$	
$P_n = \frac{e^{-\lambda/\mu} (\lambda/\mu)^n}{n!}$	

4.5 STEADY STATE BEHAVIOR OF FINITE POPULATION MODELS (M|M|C|K|K)

- In many practical problems, the assumption of an infinite calling population leads to invalid results because the calling population is, in fact small.
- When the calling population is small, the presence of one or more customers in the system have a strong effect on the distribution of future arrivals and the use of an infinite population model can be misleading.
- Consider a finite calling population model with k customers. The time between the end of one service visit and the next call for service for each member of the population is assumed to be exponentially distributed with mean $1/\lambda$ time units.
- Service times are also exponentially distributed, with mean $1/\mu$ time units. There are c parallel servers and system capacity is so that all arrivals remain for service. Such a system is shown in figure.



The effective arrival rate λ_e has several valid interpretations:

- Λ_e = long-run effective arrival rate of customers to queue
- = long-run effective arrival rate of customers entering service
- = long-run rate at which customers exit from service
- = long-run rate at which customers enter the calling population
- = long-run rate at which customers exit from the calling population.

Table 6.8 Steady-State Parameters for the $M/M/c/K/K$ Queue

P_0	$\left[\sum_{n=0}^{c-1} \binom{K}{n} \left(\frac{\lambda}{\mu} \right)^n + \sum_{n=c}^K \frac{K!}{(K-n)! c! c^{n-c}} \left(\frac{\lambda}{\mu} \right)^n \right]^{-1}$
P_n	$\begin{cases} \binom{K}{n} \left(\frac{\lambda}{\mu} \right)^n P_0, & n = 0, 1, \dots, c-1 \\ \frac{K!}{(K-n)! c! c^{n-c}} \left(\frac{\lambda}{\mu} \right)^n P_0, & n = c, c+1, \dots, K \end{cases}$
L	$\sum_{n=c}^K n P_n$
L_Q	$\sum_{n=c+1}^K (n-c) P_n$
λ_e	$\sum_{n=0}^K (K-n) \lambda P_n$
w	L / λ_e
w_Q	L_Q / λ_e
ρ	$\frac{L - L_Q}{c} = \frac{\lambda_e}{c\mu}$

4.6 NETWORKS OF QUEUE

- Many systems are naturally modeled as networks of single queues in which customer departing from one queue may be routed to another
 - The following results assume a stable system with infinite calling population and no limit on system capacity.
- 1) Provided that no customers are created or destroyed in the queue, then the departure rate out of a queue is the same as the arrival rate into the queue over the long run.
 - 2) If customers arrive to queue i at rate λ_i and a fraction $0 \leq p_{ij} \leq 1$ of them are routed to queue j upon departure, then the arrival rate from queue i to queue j is $\lambda_i p_{ij}$ over long run
 - 3) The overall arrival rate into queue j , λ_j is the sum of the arrival rate from all source. If customers arrive from outside the network at rate a_j then

$$\lambda_j = a_j + \sum_{\text{all } i} \lambda_i p_{ij}$$

- 4) If queue j has $c_j < \infty$ parallel servers, each working at rate μ_j , then the long run utilization of each server is

$$p_j = \frac{\lambda_j}{c_j \mu_j}$$

& $p_j < 1$ is required for queue to be stable

- 5) If, for each queue j , arrivals from outside the network form a poisson process with rate a_j and if there are c_j identical services delivering exponentially distributed service times with mean $1/\mu_j$ then in steady state queue j behaves like a $M|M|C$; queue with arrival rate

$$\lambda_j = a_j + \sum_{\text{all } i} \lambda_i p_{ij}$$