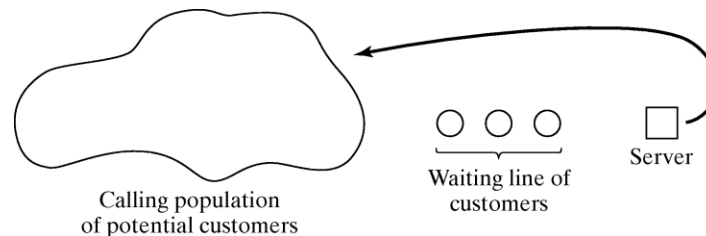




# Queueing Models

# Purpose

- Simulation is often used in the analysis of queueing models.
- A simple but typical queueing model:



- Queueing models provide the analyst with a powerful tool for designing and evaluating the performance of queueing systems.
- Typical measures of system performance:
  - Server utilization, length of waiting lines, and delays of customers
  - For relatively simple systems, compute mathematically
  - For realistic models of complex systems, simulation is usually required.

# Outline

- Discuss some well-known models (not development of queueing theories):
  - General characteristics of queues,
  - Meanings and relationships of important performance measures,
  - Estimation of mean measures of performance.
  - Effect of varying input parameters,
  - Mathematical solution of some basic queueing models.

# Characteristics of Queueing Systems

- Key elements of queueing systems:
  - Customer: refers to anything that arrives at a facility and requires service, e.g., people, machines, trucks, emails.
  - Server: refers to any resource that provides the requested service, e.g., repairpersons, retrieval machines, runways at airport.

# Calling Population

## [Characteristics of Queueing System]

- Calling population: the population of potential customers, may be assumed to be finite or infinite.
  - Finite population model: if arrival rate depends on the number of customers being served and waiting, e.g., model of one corporate jet, if it is being repaired, the repair arrival rate becomes zero.
  - Infinite population model: if arrival rate is not affected by the number of customers being served and waiting, e.g., systems with large population of potential customers.

# System Capacity

## [Characteristics of Queueing System]

- System Capacity: a limit on the number of customers that may be in the waiting line or system.
  - Limited capacity, e.g., an automatic car wash only has room for 10 cars to wait in line to enter the mechanism.
  - Unlimited capacity, e.g., concert ticket sales with no limit on the number of people allowed to wait to purchase tickets.

# Arrival Process

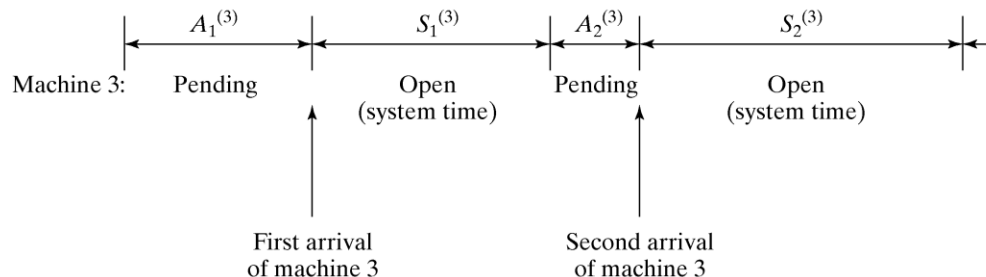
## [Characteristics of Queueing System]

- For infinite-population models:
  - In terms of interarrival times of successive customers.
  - Random arrivals: interarrival times usually characterized by a probability distribution.
    - Most important model: Poisson arrival process (with rate  $\lambda$ ), where  $A_n$  represents the interarrival time between customer  $n-1$  and customer  $n$ , and is exponentially distributed (with mean  $1/\lambda$ ).
  - Scheduled arrivals: interarrival times can be constant or constant plus or minus a small random amount to represent early or late arrivals.
    - e.g., patients to a physician or scheduled airline flight arrivals to an airport.
  - At least one customer is assumed to always be present, so the server is never idle, e.g., sufficient raw material for a machine.

# Arrival Process

## [Characteristics of Queueing System]

- For finite-population models:
  - Customer is pending when the customer is outside the queueing system, e.g., machine-repair problem: a machine is “pending” when it is operating, it becomes “not pending” the instant it demands service from the repairman.
  - Runtime of a customer is the length of time from departure from the queueing system until that customer’s next arrival to the queue, e.g., machine-repair problem, machines are customers and a runtime is time to failure.
  - Let  $A_1^{(i)}, A_2^{(i)}, \dots$  be the successive runtimes of customer  $i$ , and  $S_1^{(i)}, S_2^{(i)}$  be the corresponding successive system times:





# Queue Behavior and Queue Discipline

## [Characteristics of Queueing System]

- Queue behavior: the actions of customers while in a queue waiting for service to begin, for example:
  - Balk: leave when they see that the line is too long,
  - Renege: leave after being in the line when its moving too slowly,
  - Jockey: move from one line to a shorter line.
- Queue discipline: the logical ordering of customers in a queue that determines which customer is chosen for service when a server becomes free, for example:
  - First-in-first-out (FIFO)
  - Last-in-first-out (LIFO)
  - Service in random order (SIRO)
  - Shortest processing time first (SPT)
  - Service according to priority (PR).

# Service Times and Service Mechanism

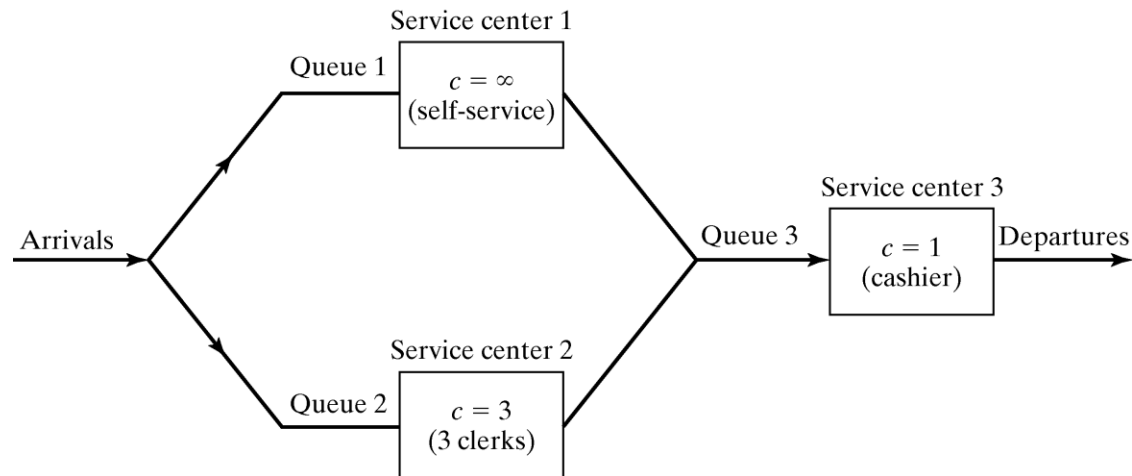
## [Characteristics of Queueing System]

- Service times of successive arrivals are denoted by  $S_1, S_2, S_3$ .
  - May be constant or random.
  - $\{S_1, S_2, S_3, \dots\}$  is usually characterized as a sequence of independent and identically distributed random variables, e.g., exponential, Weibull, gamma, lognormal, and truncated normal distribution.
- A queueing system consists of a number of service centers and interconnected queues.
  - Each service center consists of some number of servers,  $c$ , working in parallel, upon getting to the head of the line, a customer takes the  $1^{st}$  available server.

# Service Times and Service Mechanism

## [Characteristics of Queueing System]

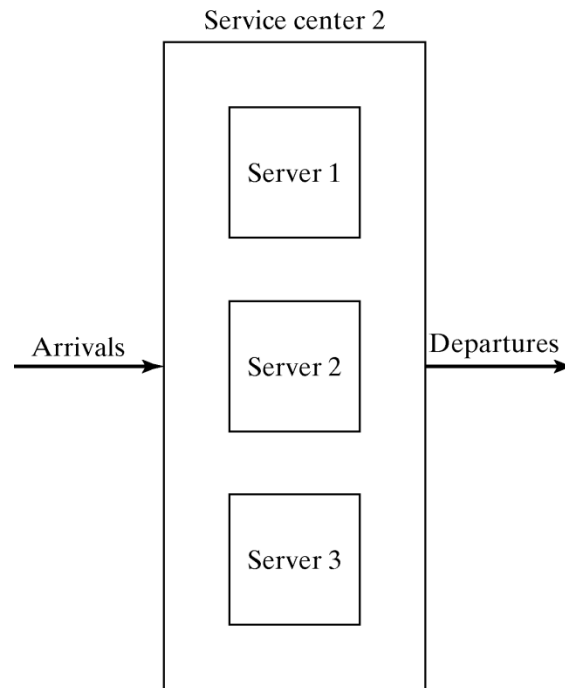
- Example: consider a discount warehouse where customers may:
  - Serve themselves before paying at the cashier:



# Service Times and Service Mechanism

## [Characteristics of Queueing System]

- Wait for one of the three clerks:



- Batch service (a server serving several customers simultaneously), or customer requires several servers simultaneously.

# Queueing Notation

## ■ A notation system for parallel server queues: $A/B/c/N/K$

□ **A** represents the Interarrival-time distribution,

- *M(Markovian, memoryless, exponential)*
- *D(constant or deterministic)*
- *$E_k$  (Erlang of order  $k$ )*
- *G(general or arbitrary)*
- *GI(general independent)*
- *H(hyper exponential)*
- *PH(phase - type)*

□ **B** represents the service-time distribution, *same as A*

□ **c** represents the number of parallel servers – *1, multiple, infinite*

□ **N** represents the system capacity or queue size – *finite, infinite*

□ **K** represents the size of the calling population - *finite, infinite*

E.g.  $G/G/1/\infty/\infty$  represents a single server system that has unlimited queue capacity and an infinite population of potential customers. The interarrival and service times are general.

# Queueing Notation

## [Characteristics of Queueing System]

- Primary performance measures of queueing systems:
  - $P_n$ : steady-state probability of having  $n$  customers in system,
  - $P_n(t)$ : probability of  $n$  customers in system at time  $t$ ,
  - $\lambda$ : arrival rate,
  - $\lambda_e$ : effective arrival rate,
  - $\mu$ : service rate of one server,
  - $\rho$ : 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.

# Time-Average Number in System $L$

## [Characteristics of Queueing System]

- Consider a queueing system over a period of time  $T$ ,
  - Let  $T_i$  denote the total time during  $[0, T]$  in which the system contained exactly  $i$  customers, the time-weighted-average number in a system is defined by:

$$\hat{L} = \frac{1}{T} \sum_{i=0}^{\infty} iT_i = \sum_{i=0}^{\infty} i \left( \frac{T_i}{T} \right)$$

- Consider the total area under the function is  $L(t)$ , then,

$$\hat{L} = \frac{1}{T} \sum_{i=0}^{\infty} iT_i = \frac{1}{T} \int_0^T L(t) dt$$

- The long-run time-average # in system, with probability 1:

$$\hat{L} = \frac{1}{T} \int_0^T L(t) dt \rightarrow L \quad \text{as } T \rightarrow \infty$$

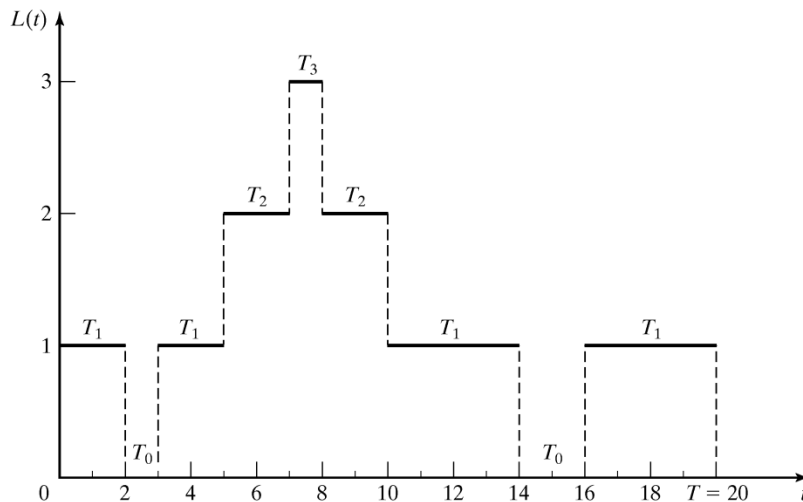
# Time-Average Number in System $L$

## [Characteristics of Queueing System]

- The time-weighted-average number in queue is:

$$\hat{L}_Q = \frac{1}{T} \sum_{i=0}^{\infty} i T_i^Q = \frac{1}{T} \int_0^T L_Q(t) dt \rightarrow L_Q \quad \text{as } T \rightarrow \infty$$

- $G/G/1/N/K$  example: consider the results from the queueing system ( $N > 4$ ,  $K > 3$ ).



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

$$L_Q(t) = \begin{cases} 0, & \text{if } L(t) = 0 \\ L(t) - 1, & \text{if } L(t) \geq 1 \end{cases}$$

$$\hat{L}_Q = \frac{0(15) + 1(4) + 2(1)}{20} = 0.3 \text{ customers}$$



# Average Time Spent in System Per Customer $w$ [Characteristics of Queueing System]

- The average time spent in system per customer, called the average system time, is:

$$\hat{w} = \frac{1}{N} \sum_{i=1}^N W_i$$

where  $W_1, W_2, \dots, W_N$  are the individual times that each of the  $N$  customers spend in the system during  $[0, T]$ .

- For stable systems:  $\hat{w} \rightarrow w$  as  $N \rightarrow \infty$
- If the system under consideration is the queue alone:

$$\hat{w}_Q = \frac{1}{N} \sum_{i=1}^N W_i^Q \rightarrow w_Q \quad \text{as } N \rightarrow \infty$$

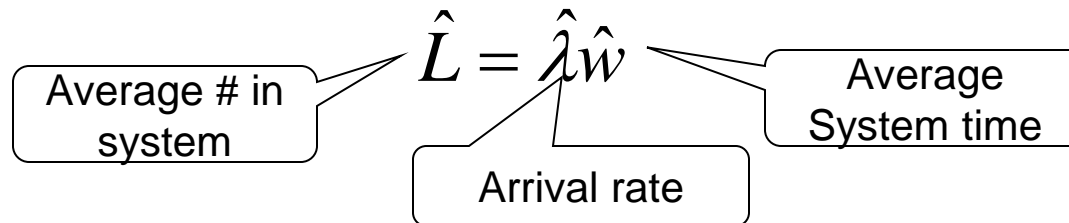
- $G/G/1/N/K$  example (cont.): the average system time is

$$\hat{w} = \frac{W_1 + W_2 + \dots + W_5}{5} = \frac{2 + (8 - 3) + \dots + (20 - 16)}{5} = 4.6 \text{ time units}$$

# The Conservation Equation

## [Characteristics of Queueing System]

- Conservation equation (a.k.a. Little's law)



$$L = \lambda w \quad \text{as } T \rightarrow \infty \text{ and } N \rightarrow \infty$$

- Holds for almost all queueing systems or subsystems (regardless of the number of servers, the queue discipline, or other special circumstances).
- *G/G/1/N/K* example (cont.): On average, one arrival every 4 time units and each arrival spends 4.6 time units in the system. Hence, at an arbitrary point in time, there is  $(1/4)(4.6) = 1.15$  customers present on average.

# Server Utilization

## [Characteristics of Queueing System]

- Definition: the proportion of time that a server is busy.
  - Observed server utilization,  $\hat{\rho}$ , is defined over a specified time interval  $[0, T]$ .
  - Long-run server utilization is  $\rho$ .
  - For systems with long-run stability:  $\hat{\rho} \rightarrow \rho$  as  $T \rightarrow \infty$

# Server Utilization

## [Characteristics of Queueing System]

- For  $G/G/1/\infty/\infty$  queues:
  - Any single-server queueing system with average arrival rate  $\lambda$  customers per time unit, where average service time  $E(S) = 1/\mu$  time units, infinite queue capacity and calling population.
  - Conservation equation,  $L = \lambda w$ , can be applied.
  - For a stable system, the average arrival rate to the server,  $\lambda_s$ , must be identical to  $\lambda$ .
  - The average number of customers in the server is:

$$\hat{L}_s = \frac{1}{T} \int_0^T (L(t) - L_Q(t)) dt = \frac{T - T_0}{T}$$

# Server Utilization

## [Characteristics of Queueing System]

- In general, for a single-server queue:

$$\hat{L}_s = \hat{\rho} \rightarrow L_s = \rho \text{ as } T \rightarrow \infty$$

$$\text{and } \rho = \lambda E(s) = \frac{\lambda}{\mu}$$

- For a single-server stable queue:  $\rho = \frac{\lambda}{\mu} < 1$

- For an unstable queue ( $\lambda > \mu$ ), long-run server utilization is 1.

# Server Utilization

## [Characteristics of Queueing System]

- For  $G/G/c/\infty/\infty$  queues:
  - A system with  $c$  identical servers in parallel.
  - If an arriving customer finds more than one server idle, the customer chooses a server without favoring any particular server.
  - For systems in statistical equilibrium, the average number of busy servers,  $L_s$ , is:  $L_s = \lambda E(s) = \lambda/\mu$ .
  - The long-run average server utilization is:

$$\rho = \frac{L_s}{c} = \frac{\lambda}{c\mu}, \quad \text{where } \lambda < c\mu \text{ for stable systems}$$

# Server Utilization and System Performance

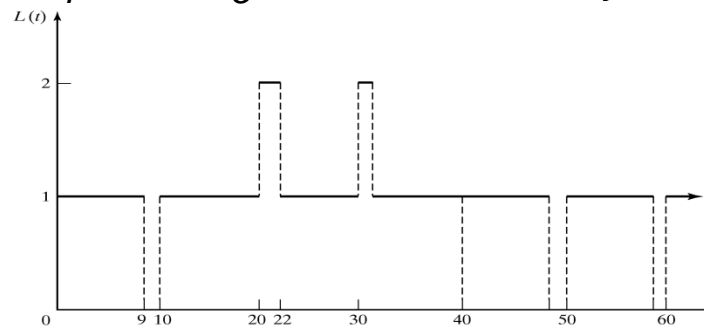
## [Characteristics of Queueing System]

- System performance varies widely for a given utilization  $\rho$ .
  - For example, a  $D/D/1$  queue where  $E(A) = 1/\lambda$  and  $E(S) = 1/\mu$ , where:
$$L = \rho = \lambda/\mu, \quad w = E(S) = 1/\mu, \quad L_Q = W_Q = 0.$$
    - By varying  $\lambda$  and  $\mu$ , server utilization can assume any value between 0 and 1.
    - Yet there is never any line.
  - In general, variability of interarrival and service times causes lines to fluctuate in length.

# Server Utilization and System Performance

## [Characteristics of Queueing System]

- Example: A physician who schedules patients every 10 minutes and spends  $S_i$  minutes with the  $i^{th}$  patient:
$$S_i = \begin{cases} 9 \text{ minutes with probability } 0.9 \\ 12 \text{ minutes with probability } 0.1 \end{cases}$$
  - Arrivals are deterministic,  $A_1 = A_2 = \dots = \lambda^{-1} = 10$ .
  - Services are stochastic,  $E(S_i) = 9.3 \text{ min}$  and  $V(S_i) = 0.81 \text{ min}^2$ .
  - On average, the physician's utilization  $= \rho = \lambda/\mu = 0.93 < 1$ .
  - Consider the system is simulated with service times:  $S_1 = 9, S_2 = 12, S_3 = 9, S_4 = 9, S_5 = 9, \dots$ . The system becomes:



- The occurrence of a relatively long service time ( $S_2 = 12$ ) causes a waiting line to form temporarily.



# Costs in Queueing Problems

## [Characteristics of Queueing System]

- Costs can be associated with various aspects of the waiting line or servers:
  - System incurs a cost for each customer in the queue, say at a rate of \$10 per hour per customer.

- The average cost per customer is:

$$\sum_{j=1}^N \frac{\$10 * W_j^Q}{N} = \$10 * \hat{w}_Q$$

$W_j^Q$  is the time customer j spends in queue

- If  $\hat{\lambda}$  customers per hour arrive (on average), the average cost per hour is:

$$\left( \hat{\lambda} \frac{\text{customer}}{\text{hour}} \right) \left( \frac{\$10 * \hat{w}_Q}{\text{customer}} \right) = \$10 * \hat{\lambda} \hat{w}_Q = \$10 * \hat{L}_Q / \text{hour}$$

- Server may also impose costs on the system, if a group of  $c$  parallel servers ( $1 \leq c \leq \infty$ ) have utilization  $r$ , each server imposes a cost of \$5 per hour while busy.
  - The total server cost is:  $\$5 * c\rho$ .

# Steady-State Behavior of Infinite-Population Markovian Models

- Markovian models: exponential-distribution arrival process (mean arrival rate =  $\lambda$ ).
- Service times may be exponentially distributed as well ( $M$ ) or arbitrary ( $G$ ).
- A queueing system is in statistical equilibrium if the probability that the system is in a given state is not time dependent:
$$P( L(t) = n ) = P_n(t) = P_n.$$
- Mathematical models in this chapter can be used to obtain approximate results even when the model assumptions do not strictly hold .
- Simulation can be used for more refined analysis (more faithful representation for complex systems).

# Steady-State Behavior of Infinite-Population Markovian Models

- For the simple model studied in this chapter, the steady-state parameter,  $L$ , the time-average number of customers in the system is:

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

- Apply Little's equation to the whole system and to the queue alone:

$$w = \frac{L}{\lambda}, \quad w_Q = w - \frac{1}{\mu}$$

$$L_Q = \lambda w_Q$$

- $G/G/c/\infty/\infty$  example: to have a statistical equilibrium, a necessary and sufficient condition is  $\lambda/(c\mu) < 1$ .

# M/G/1 Queues

[Steady-State of Markovian Model]

- Single-server queues with Poisson arrivals & unlimited capacity.
- Suppose service times have mean  $1/\mu$  and variance  $\sigma^2$  and  $\rho = \lambda/\mu < 1$ , the steady-state parameters of M/G/1 queue:

$$\begin{aligned}\rho &= \lambda / \mu, \quad P_0 = 1 - \rho \\ L &= \rho + \frac{\rho^2 (1 + \sigma^2 \mu^2)}{2(1 - \rho)}, \quad L_Q = \frac{\rho^2 (1 + \sigma^2 \mu^2)}{2(1 - \rho)} \\ w &= \frac{1}{\mu} + \frac{\lambda(1/\mu^2 + \sigma^2)}{2(1 - \rho)}, \quad w_Q = \frac{\lambda(1/\mu^2 + \sigma^2)}{2(1 - \rho)}\end{aligned}$$

# M/G/1 Queues

[Steady-State of Markovian Model]

- No simple expression for the steady-state probabilities  $P_0, P_1, \dots$
- $L - L_Q = \rho$  is the time-average number of customers being served.
- Average length of queue,  $L_Q$ , can be rewritten as:

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

- If  $\lambda$  and  $\mu$  are held constant,  $L_Q$  depends on the variability,  $\sigma^2$ , of the service times.

# M/G/1 Queues

## [Steady-State of Markovian Model]

- Example: Two workers competing for a job, Able claims to be faster than Baker on average, but Baker claims to be more consistent,
  - Poisson arrivals at rate  $\lambda = 2$  per hour ( $1/30$  per minute).
  - Able:  $1/\mu = 24$  minutes and  $\sigma^2 = 20^2 = 400$  minutes<sup>2</sup>:

$$L_Q = \frac{(1/30)^2 [24^2 + 400]}{2(1 - 4/5)} = 2.711 \text{ customers}$$

- The proportion of arrivals who find Able idle and thus experience no delay is  $P_0 = 1 - \rho = 1/5 = 20\%$ .
  - Baker:  $1/\mu = 25$  minutes and  $\sigma^2 = 2^2 = 4$  minutes<sup>2</sup>:

$$L_Q = \frac{(1/30)^2 [25^2 + 4]}{2(1 - 5/6)} = 2.097 \text{ customers}$$

- The proportion of arrivals who find Baker idle and thus experience no delay is  $P_0 = 1 - \rho = 1/6 = 16.7\%$ .
  - Although working faster on average, Able's greater service variability results in an average queue length about 30% greater than Baker's.

# M/M/1 Queues

[Steady-State of Markovian Model]

- Suppose the service times in an  $M/G/1$  queue are exponentially distributed with mean  $1/\mu$ , then the variance is  $\sigma^2 = 1/\mu^2$ .
  - $M/M/1$  queue is a useful approximate model when service times have standard deviation approximately equal to their means.
  - The steady-state parameters:

$$\rho = \lambda / \mu, \quad P_n = (1 - \rho) \rho^n$$

$$L = \frac{\lambda}{\mu - \lambda} = \frac{\rho}{1 - \rho}, \quad L_Q = \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{\rho^2}{1 - \rho}$$

$$w = \frac{1}{\mu - \lambda} = \frac{1}{\mu(1 - \rho)}, \quad w_Q = \frac{\lambda}{\mu(\mu - \lambda)} = \frac{\rho}{\mu(1 - \rho)}$$

# M/M/1 Queues

[Steady-State of Markovian Model]

- Example: *M/M/1* queue with service rate  $\mu=10$  customers per hour.
  - Consider how  $L$  and  $w$  increase as arrival rate,  $\lambda$ , increases from 5 to 8.64 by increments of 20%:

$\lambda$	5.0	6.0	7.2	8.64	10.0
$\rho$	0.500	0.600	0.720	0.864	1.000
$L$	1.00	1.50	2.57	6.35	$\infty$
$w$	0.20	0.25	0.36	0.73	$\infty$

- If  $\lambda/\mu \geq 1$ , waiting lines tend to continually grow in length.
- Increase in average system time ( $w$ ) and average number in system ( $L$ ) is highly nonlinear as a function of  $\rho$ .



# Effect of Utilization and Service Variability

[Steady-State of Markovian Model]

- For almost all queues, if lines are too long, they can be reduced by decreasing server utilization ( $\rho$ ) or by decreasing the service time variability ( $\sigma^2$ ).
- A measure of the variability of a distribution, coefficient of variation (cv):

$$(cv)^2 = \frac{V(X)}{[E(X)]^2}$$

- The larger cv is, the more variable is the distribution relative to its expected value

# Effect of Utilization and Service Variability

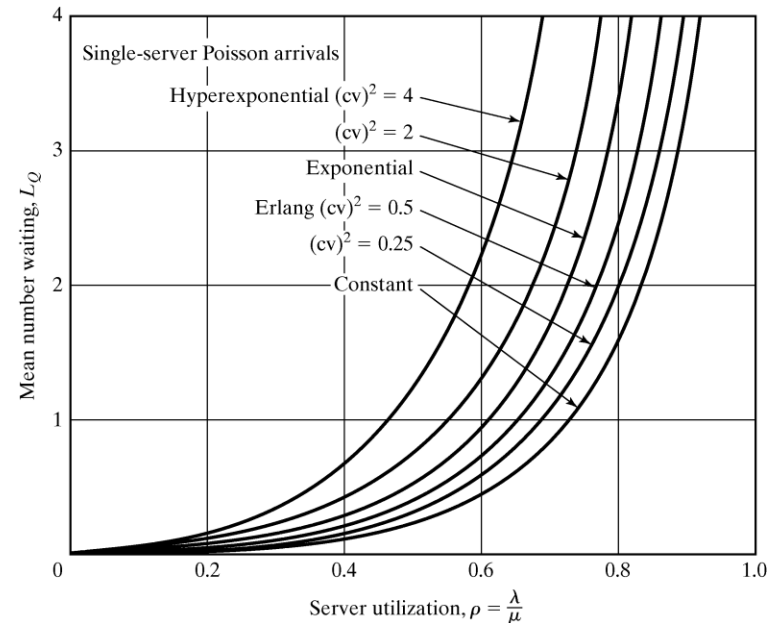
[Steady-State of Markovian Model]

- Consider  $L_Q$  for any  $M/G/1$  queue:

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

$L_Q$  for  $M/M/1$  queue

Corrects the  $M/M/1$  formula to account for a non-exponential service time dist'n



# Multiserver Queue [Steady-State of Markovian Model]

- $M/M/c/\infty/\infty$  queue:  $c$  channels operating in parallel.
  - Each channel has an independent and identical exponential service-time distribution, with mean  $1/\mu$ .
  - To achieve statistical equilibrium, the offered load ( $\lambda/\mu$ ) must satisfy  $\lambda/\mu < c$ , where  $\lambda/(c\mu) = \rho$  is the server utilization.
  - Some of the steady-state probabilities:

$$\rho = \lambda / c\mu$$

$$P_0 = \left\{ \left[ \sum_{n=0}^{c-1} \frac{(\lambda / \mu)^n}{n!} \right] + \left[ \left( \frac{\lambda}{\mu} \right)^c \left( \frac{1}{c!} \right) \left( \frac{c\mu}{c\mu - \lambda} \right) \right] \right\}^{-1}$$

$$L = c\rho + \frac{(c\rho)^{c+1} P_0}{c(c!)(1-\rho)^2} = c\rho + \frac{\rho P(L(\infty) \geq c)}{1-\rho}$$

$$w = \frac{L}{\lambda}$$

# Multiserver Queue [Steady-State of Markovian Model]

- Other common multiserver queueing models:
  - $M/G/c/\infty$ : general service times and  $c$  parallel server. The parameters can be approximated from those of the  $M/M/c/\infty/\infty$  model.
  - $M/G/\infty$ : general service times and infinite number of servers, e.g., customer is its own system, service capacity far exceeds service demand.
  - $M/M/C/N/\infty$ : service times are exponentially distributed at rate  $m$  and  $c$  servers where the total system capacity is  $N \geq c$  customer (when an arrival occurs and the system is full, that arrival is turned away).

# Steady-State Behavior of Finite-Population Models

- When the calling population is small, the presence of one or more customers in the system has a strong effect on the distribution of future arrivals.
- Consider a finite-calling population model with  $K$  customers ( $M/M/c/K/K$ ):
  - The time between the end of one service visit and the next call for service is exponentially distributed, (mean =  $1/\lambda$ ).
  - Service times are also exponentially distributed.
  - $c$  parallel servers and system capacity is  $K$ .

# Steady-State Behavior of Finite-Population Models

- Some of the steady-state probabilities:

$$P_0 = \left\{ \sum_{n=0}^{c-1} \binom{K}{n} \left( \frac{\lambda}{\mu} \right)^n + \sum_{n=c}^K \frac{K!}{(K-n)!c!c^{n-c}} \left( \frac{\lambda}{\mu} \right)^n \right\}^{-1}$$

$$P_n = \begin{cases} \binom{K}{n} \left( \frac{\lambda}{\mu} \right)^n P_0, & n = 0, 1, \dots, c-1 \\ \frac{K!}{(K-n)!c!c^{n-c}} \left( \frac{\lambda}{\mu} \right)^n, & n = c, c+1, \dots, K \end{cases}$$

$$L = \sum_{n=0}^K nP_n, \quad w = L / \lambda_e, \quad \rho = \lambda_e / c\mu$$

where  $\lambda_e$  is the long run effective arrival rate of customers to queue (or entering/exiting service)

$$\lambda_e = \sum_{n=0}^K (K-n)\lambda P_n$$

# Steady-State Behavior of Finite-Population Models

- Example: two workers who are responsible for 10 milling machines.
  - Machines run on the average for 20 minutes, then require an average 5-minute service period, both times exponentially distributed:  $\lambda = 1/20$  and  $\mu = 1/5$ .

- All of the performance measures depend on  $P_0$ :

$$P_0 = \left\{ \sum_{n=0}^{2-1} \binom{10}{n} \left( \frac{5}{20} \right)^n + \sum_{n=2}^{10} \frac{10!}{(10-n)!2!2^{n-2}} \left( \frac{5}{20} \right)^n \right\}^{-1} = 0.065$$

- Then, we can obtain the other  $P_n$ .
- Expected number of machines in system:

$$L = \sum_{n=0}^{10} nP_n = 3.17 \text{ machines}$$

- The average number of running machines:

$$K - L = 10 - 3.17 = 6.83 \text{ machines}$$

# Networks of Queues

- Many systems are naturally modeled as networks of single queues: customers 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:
  - 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).
  - If customers arrive to queue  $i$  at rate  $\lambda_i$ , 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 the long run).



# Networks of Queues

- The overall arrival rate into queue  $j$ :

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

Arrival rate  
from outside  
the network

Sum of arrival rates  
from other queues in  
network

- If queue  $j$  has  $c_j < \infty$  parallel servers, each working at rate  $\mu_j$ , then the long-run utilization of each server is  $\rho_j = \lambda_j / (c_j \mu_j)$  (where  $\rho_j < 1$  for stable queue).
- If arrivals from outside the network form a Poisson process with rate  $a_j$  for each queue  $j$ , and if there are  $c_j$  identical servers delivering exponentially distributed service times with mean  $1/\mu_j$ , then, in steady state, queue  $j$  behaves like an  $M/M/c_j$  queue with arrival rate  $\lambda_j = a_j + \sum_{\text{all } i} \lambda_i p_{ij}$

# Network of Queues

## ■ Discount store example:

- Suppose customers arrive at the rate 80 per hour and 40% choose self-service. Hence:

- Arrival rate to service center 1 is  $\lambda_1 = 80(0.4) = 32$  per hour
- Arrival rate to service center 2 is  $\lambda_2 = 80(0.6) = 48$  per hour.

- $c_2 = 3$  clerks and  $\mu_2 = 20$  customers per hour.

- The long-run utilization of the clerks is:

$$\rho_2 = 48/(3 \cdot 20) = 0.8$$

- All customers must see the cashier at service center 3, the overall rate to service center 3 is  $\lambda_3 = \lambda_1 + \lambda_2 = 80$  per hour.

- If  $\mu_3 = 90$  per hour, then the utilization of the cashier is:

$$\rho_3 = 80/90 = 0.89$$

# Summary

- Introduced basic concepts of queueing models.
- Show how simulation, and some times mathematical analysis, can be used to estimate the performance measures of a system.
- Commonly used performance measures:  $L$ ,  $L_Q$ ,  $w$ ,  $w_Q$ ,  $\rho$ , and  $\lambda_e$ .
- When simulating any system that evolves over time, analyst must decide whether to study transient behavior or steady-state behavior.
  - Simple formulas exist for the steady-state behavior of some queues.
- Simple models can be solved mathematically, and can be useful in providing a rough estimate of a performance measure.



## ■ Problem Statement 1

- Arrivals at a telephone booth are considered to be Poisson, with an average time of 10 minutes between one arrival and the next. The length of a phone call is assumed to be distributed exponentially with mean 3 minutes. Find:
- The probability that an arrival finds that four persons are waiting for their turn;
- The average number of persons waiting and making telephone calls; and
- The average length of the queue that is formed from time to time.



## ■ Problem Statement 2:

The Service Station has a central store where service mechanics arrive to take spare parts for the jobs they work upon. The mechanics wait in queue if necessary and are served on a first-come-first-served basis. The store is manned by one attendant who can attend 8 mechanics in an hour on an average. The arrival rate of the mechanics averages 6 per hour. Assuming that the pattern of mechanics arrivals is Poisson distributed and the servicing time is exponentially distributed, determine  $W_s$ ,  $W_q$  and  $L_q$  where the symbols carry their usual meaning.



### **Problem Statement 3:**

- Customers arrive at the First Class Ticket counter of a Theatre at a rate of 12 per hour. There is one clerk serving the customers at a rate of 30 per hour
- What is the probability that there is no customer in counter (i.e. that the system is idle)?
- What is the probability that there are more than 2 customers in the counter?
- What is the probability that there is no customer waiting to be served?
- What is the probability that a customer is being served and nobody is waiting?



## Problem Statement 4:

- The rate of arrival of customer at a public telephone follows Poisson distribution, with an average time of ten minutes between one customer and the next. The duration of a phone call is assume to follow exponential distribution with a mean time of three minutes
- What is the probability that person arriving at the booth will have a wait?
- What is the average length of the queue?
- The Mahangar Telephone Nigam Ltd. will install another booth when it is convinced that the customers would have a wait for at least three minutes for their turn to make a call. How much should be the flow of customers in order to justify a second booth?



## Problem Statement 5:

- A warehouse has only one loading dock manned by a three person crew. Trucks arrive at the loading dock at an average rate of 4 trucks per hour and the arrivals are Poisson distributed. The loading of a truck takes 10 minutes on an average and the loading time can be assumed to be exponentially distributed about this average. The operating cost of a truck is Rs. 100 per hour and the members of the loading crew are paid at a rate of Rs 25 per hour. Assuming that the addition of new crew members would reduce the loading time to 7.5 minutes, would you advise the truck owner to add another crew of three persons?





## **Problem Statement 6:**

- (a) A typist at an office receives, on an average, 22 letters per day of typing. The typist works 8 hours a day and it takes on an average 20 minutes to type a letter. The company has determined that the cost of a letter waiting to be mailed (opportunity cost) is 80 paisa per hour and the equipment operating cost plus the salary of the typist will be Rs. 40 per day.
- What is the typist's utilization rate?
- What is the average number of letters waiting to be typed?
- What is the average waiting time needed to have letter typed?
- What is the total cost of waiting letters to be mailed?
- Forced to improve the letter typing service, the above company is planning to take lease of one of the two models of an automated typewriter available in the market. The daily costs and the resulting increase in the typist's efficiency are displayed in the table given below:

- (b) Forced to improve the letter typing service, the above company is planning to take lease of one of the two models of an automated typewriter available in the market. The daily costs and the resulting increase in the typist's efficiency are displayed in the table given below:

Model	Additional Cost per Day	Increase in Typist's Efficiency
I	Rs. 37	50%
II	Rs. 39	75%

- What action should the company take to minimize the total daily costs of waiting letters to be mailed?



## Problem Statement 7:

- On an average, 4 patients per hour require the service of an emergency clinic. Also, on an average, a patient requires 10 minutes of active attention. The clinic can handle only one emergency at a time. Suppose that it costs the clinic Rs. 300 per patient treated to obtain an average servicing time of 10 minutes and that each minute of decrease in this average time would cost Rs. 50 per patient treated. How much would have to be budgeted by the clinic to decrease the average size of the queue from  $\frac{4}{3}$  patients to  $\frac{1}{2}$  patients?



## ■ Problem Statement 8

### ■ Hospital Emergency Department

Patients arrive at an emergency department following a random pattern influenced by time of day, with typically higher rates during evening hours. The time each patient spends receiving treatment varies widely, influenced by the nature of their emergencies.

Task: Determine the most suitable queue model to analyze patient wait times and resource utilization. Justify your choice based on the distribution characteristics of arrival and service times.



## ■ Problem Statement 9

### ■ Retail Store Checkout

Scenario: A retail store has checkout counters where customers arrive randomly but with increased frequency during sales or holiday seasons. Each cashier works at a different speed, and some have fixed times due to the complexity of transactions.

Task: Choose a queuing model to evaluate the efficiency of the checkout process during peak and off-peak hours. Discuss the factors that influenced your choice of model and how it impacts the analysis of service efficiency and customer wait times.



## ■ Problem Statement 10

### ■ Software Update Downloads

Scenario: Users download a critical software update released simultaneously worldwide. The download requests arrive at servers at a rate that sharply increases and then stabilizes. The server handles each request with varying times depending on the user's internet speed and server load.

Task: Identify the best queuing model to study the download wait times and server response times under varying loads. Provide reasons for your choice and consider the nature of both arrival and service processes.



## ■ Problem Statement 11

### ■ Comparative Analysis of Different Distributions:

Consider a simulation of customer arrivals at a retail store. If you model the arrivals using a Poisson distribution versus a normal distribution, how might the results differ in terms of customer flow and queue management? What factors would influence your choice of distribution?