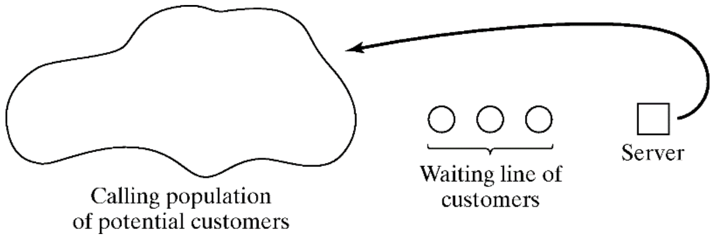
Chapter 6
Queueing Models

Banks, Carson, Nelson & Nicol

Discrete-Event System Simulation

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

Queueing System Examples

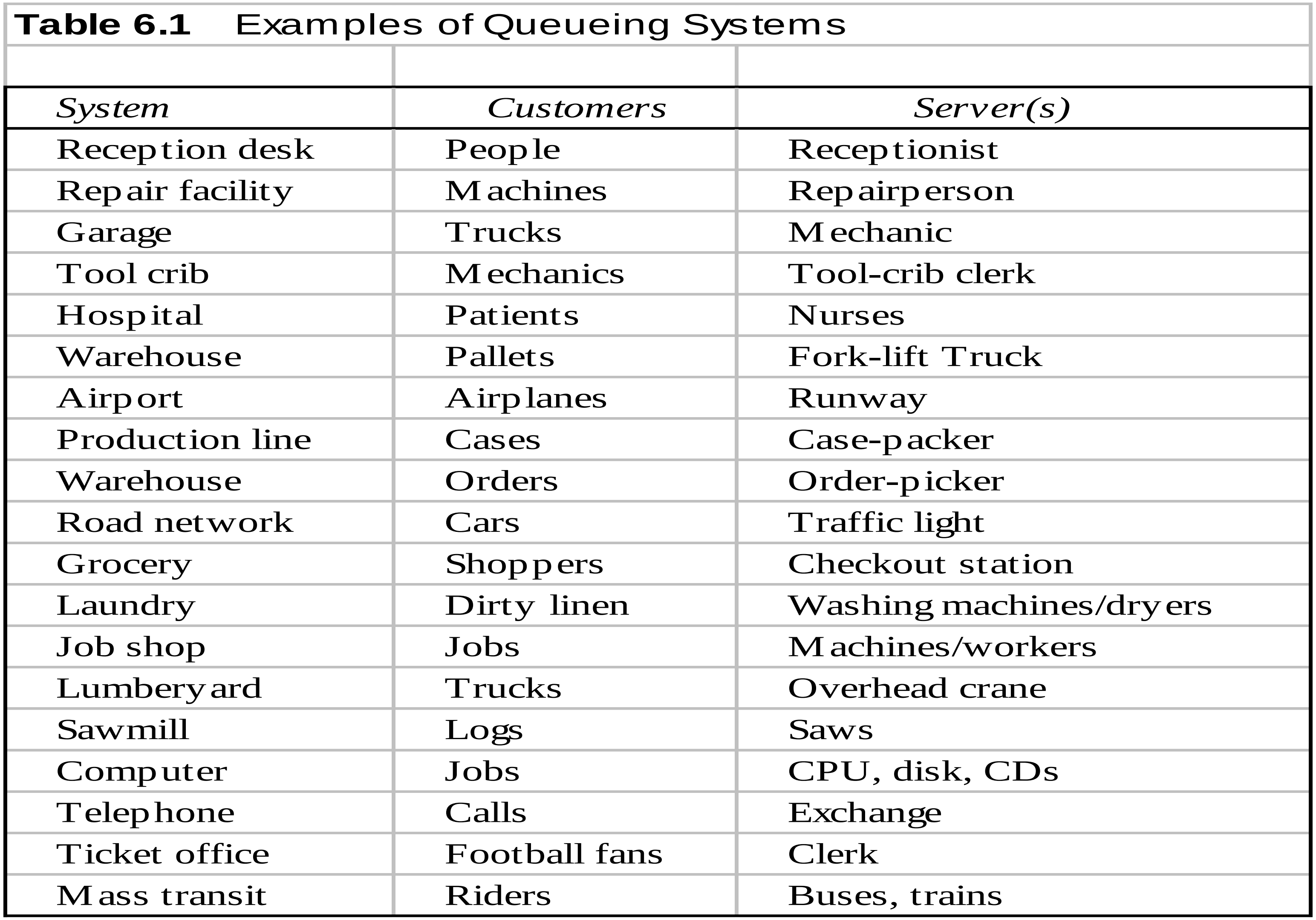
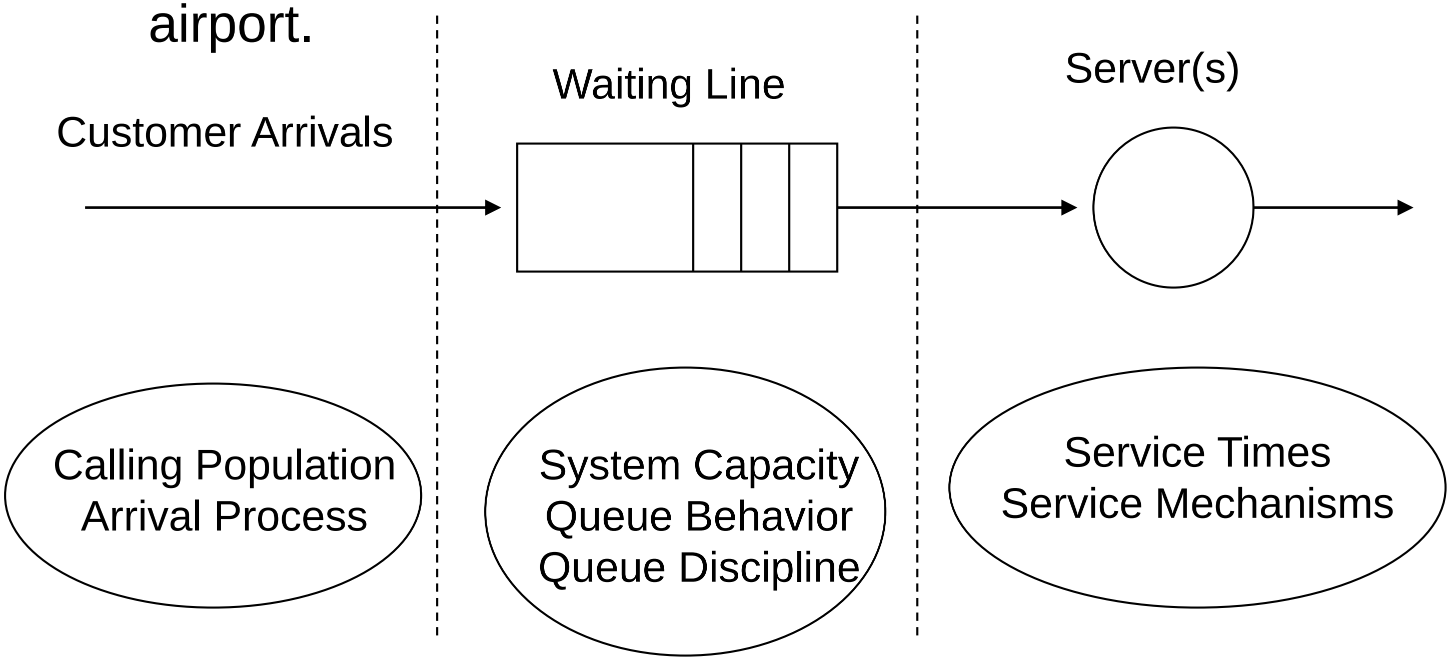


Table 6.1 Examples of Queueing Systems

System	*Customers*	*Server(s)*
Reception desk	People	Receptionist
Repair facility	Machines	Repairperson
Garage	Trucks	Mechanic
Tool crib	Mechanics	Tool-crib clerk
Hospital	Patients	Nurses
Warehouse	Pallets	Fork-lift Truck
Airport	Airplanes	Runway
Production line	Cases	Case-packer
Warehouse	Orders	Order-picker
Road network	Cars	Traffic light
Grocery	Shoppers	Checkout station
Laundry	Dirty linen	Washing machines/dryers
Job shop	Jobs	Machines/workers
Lumberyard	Trucks	Overhead crane
Sawmill	Logs	Saws
Computer	Jobs	CPU, disk, CDs
Telephone	Calls	Exchange
Ticket office	Football fans	Clerk
Mass transit	Riders	Buses, trains

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., repair persons, 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***:
  - Usually characterized in terms of *inter-arrival times* of successive customers. Arrivals may occur at random or scheduled times
  - *Random arrivals:* inter-arrival times usually characterized by a probability distribution.
    - Most important model: Poisson arrival process (with rate λ), where A_n represents the inter-arrival time between customer *n-1* and customer *n*, and is exponentially distributed (with mean $1/\lambda$).
  - *Scheduled arrivals*: inter-arrival times can be constant or constant plus or minus a small random amount (jitter) to represent early or late arrivals.
    - e.g., patients to a physician or scheduled airline flight arrivals to an airport.

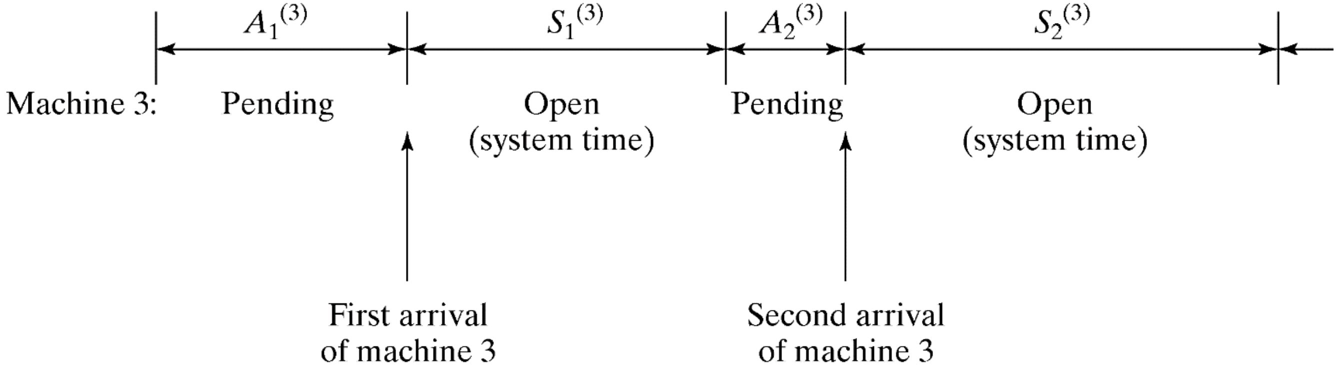
Arrival Process

[Characteristics of Queueing System]

- *For finite-population models*:
  - □ Define customer as "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 form the repairman.
  - □ Define "runtime" of a customer as 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 $A_1^{(i)}$, $A_2^{(i)}$, … be the successive runtimes of customer *i*, and $S_1^{(i)}$, $S_2^{(i)}$ be the corresponding successive system times: that is $S_n^{(i)}$ is the total time spent in the system by customer *i* during the n^{th} visit.

Arrival Process

- Finite Population Models



  - □ The total arrival process is the superposition of the arrival times of all customers.
  - □ One important application of finite models is the machine-repair problem. Machines are the customers and runtime is time to failure. When a machine fails, it “arrives” at the queueing system and remains there until it is served. Time to failure is chracterized by exponential, Weibull and Gamma distributions.

Queue Behavior and Queue Discipline

[Characteristics of Queueing System]

- ***Queue behavior:*** refers to 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:*** refers to 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). (e.g., type, class, priority)

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.
  - $\{S_1, S_2, S_3, \ldots\}$ is usually characterized as a sequence of independent and identically distributed random variables, e.g., exponential, Weibull, gamma, lognormal, and truncated normal distribution.
  - Sometimes, services are identically distributed for all customers of a given type or class or priority, where as customers of different types might have completely different service-time distributions
  - In some systems, service times depend upon the time of the day or upon the length of waiting line (e.g., servers might work faster than usual if waiting times are long, effectively reducing service times)

Service Times and Service Mechanism

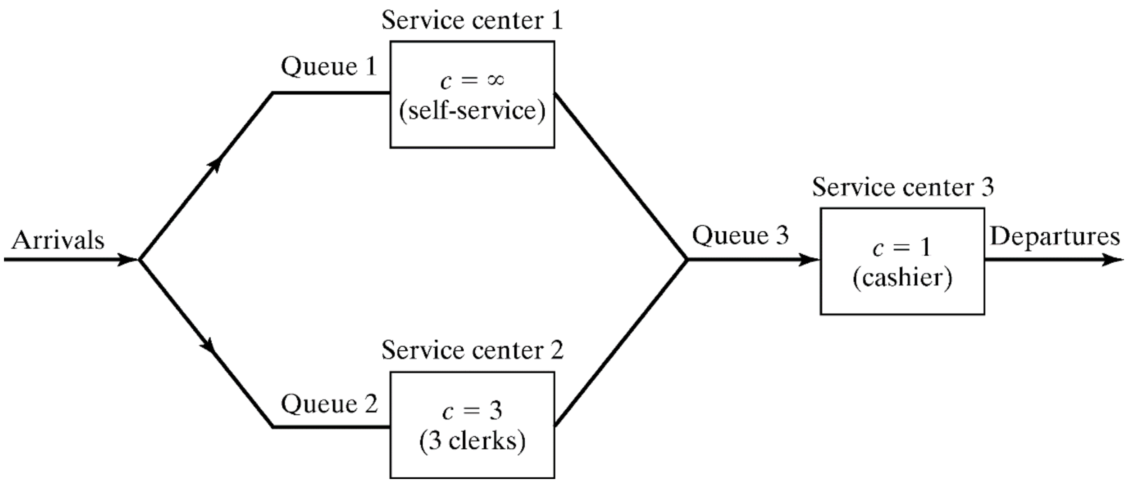
[Characteristics of Queueing System]

- 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.
  - □ Parallel service mechanisms are either single server ($c=1$), multiple server ($1<c<\infty$), or unlimited servers ($c=\infty$)
    - A self service facility is usually characterized by an unlimited number of servers

Service Times and Service Mechanism

[Characteristics of Queueing System]

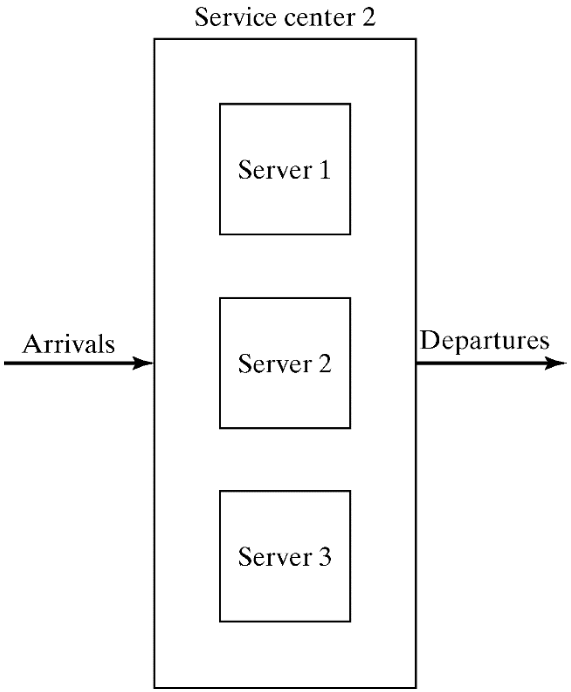
- Example: consider a discount warehouse where customers may:
  - Serve themselves before paying at the cashier:



Service Times and Service Mechanism

[Characteristics of Queueing System]

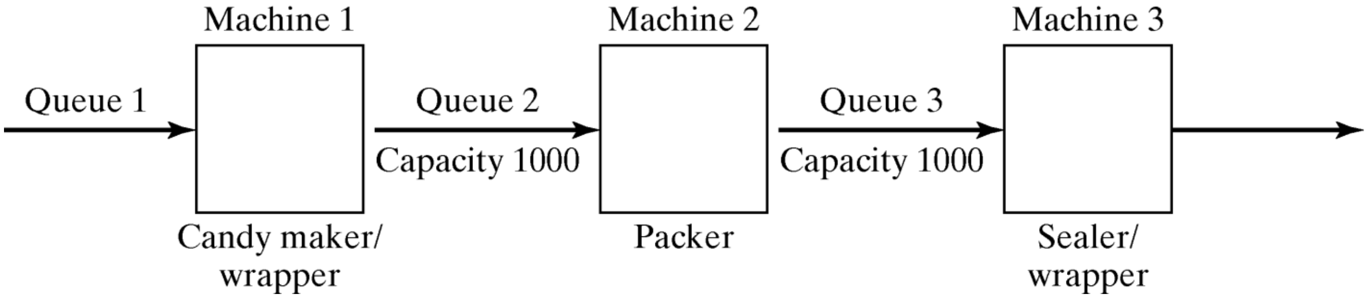
- Wait for one of the three clerks:



- **Batch service** (a server serving several customers simultaneously, e.g., small orders), or customer requires several servers simultaneously (e.g., big order or heavy item).

Example

- A candy production line has three machines separated by inventory-in-process buffers. The first machine makes and wraps individual pieces of candy, the second one packs 50 pieces in a box and third machine seals and wraps the box. The inventory buffers have capacities of 1000 boxes each. The system is modeled as three single service centers (with one server, c=1) in series and with queue capacity constraints and continuous arrival stream at first queue.



Queueing Notation

[Characteristics of Queueing System]

- Recognizing the diversity of queueing systems, Kendall proposed a notational system in 1953 that has been widely adopted.
- A notation system for parallel server queues: *A/B/c/N/K*
  - □ *A* represents the inter-arrival-time distribution,
  - □ *B* represents the service-time distribution,
  - □ *c* represents the number of parallel servers,
  - □ *N* represents the system capacity,
  - □ *K* represents the size of the calling population.
- Common symbols for *A* and *B* include *M (exponential), D (constant or deterministic),* E_k *(for Erlang order k), PH (Phase-type), H (hyper-exponential), G (arbitrary or general), and GI (general independent).*

Queueing Notation

[Characteristics of Queueing System]

- Examples:
  - *M/M/1 (also M/M/1/∞/∞)* indicates a single-server that has unlimited queue capacity and an infinite population model. The interarrivals and service times are exponentially distributed.
    - *When N and K are infinite, they are often dropped from the notation*
  - *G/G/1/5/5* represents a queueing system with general (or arbitrary) inter-arrival and service distribution with single server, with a queue capacity of 5 and finite population model of size 5
    - *General* models are used to solve the queue system with no particular distribution in mind
    - Very useful as the final results can be obtained by plugging in the values of specific distributions

Queueing Notation

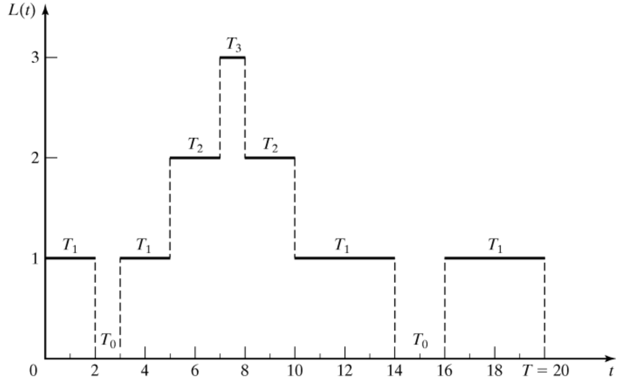
[Characteristics of Queueing System]

- ***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*,
  - λ: arrival rate,
  - λ_e: effective arrival rate,
  - μ: service rate of one server,
  - ρ: server utilization,
  - A_n: inter-arrival 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*

- 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, *time-integrated-average* is given by:

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

These two expressions are always equal for any queueing system regardless of the number of servers, the queue discipline or any other circumstances

Time-Average Number in System *L*

[Characteristics of Queueing System]

- Many queueing systems exhibit long-run stability in terms of their average performance. For such systems, as time *T* gets large, the observed time-average number in the system $\hat{L}$ approaches a limiting value *L*, called *long-run time-average*.

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

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

- The estimator $\hat{L}$ is said to be strongly consistent for *L*. If simulation run length *T* is sufficiently long, the estimator $\hat{L}$ becomes arbitrarily close to *L*. Unfortunately for $T < \infty$, it depends on the initial condition at *t=0*. (Reason to do multiple simulation runs!!)

Time-Average Number in Queue L_Q

[Characteristics of Queueing System]

- Similarly, if $L_Q(t)$ denotes the number of customers waiting in line and T_i^Q denotes the total time during [0,T] in which exactly *i* customers are waiting in line. Then,
- The *time-weighted-average* number in queue is:

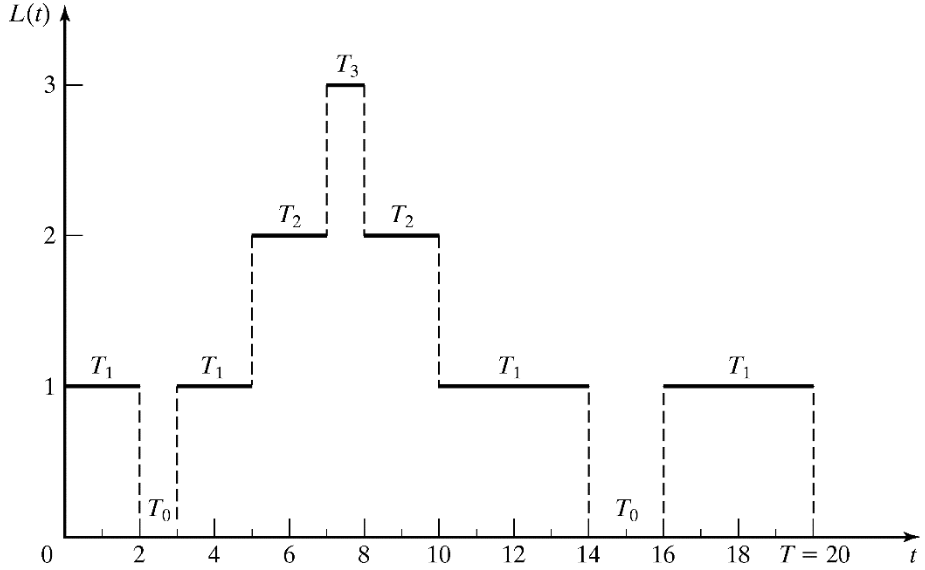
$$\hat{L}_Q = \frac{1}{T}\sum_{i=0}^{\infty} iT_i^Q = \frac{1}{T}\int_0^T L_Q(t)dt \rightarrow L_Q \quad \text{as} \quad T \rightarrow \infty$$

- Here, $\hat{L}_Q$ is the observed time-average number of customers in waiting line from time *0* to *T.*
- L_Q is the long-run time-average number waiting in line.

Time-Average Number in System *L*

[Characteristics of Queueing System]

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



$$\hat{L} = [0(3) + 1(12) + 2(4) + 3(1)] / 20$$
$$= 23/20 = 1.15 \text{ cusomters}$$

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

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

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

- If we simulate the queueing system for a period, say *T* and then record the time each customer spends in the system during *[0,T]*, say W_1, W_2, …., W_N where N is the number of arrivals in *[0,T]*.
- The average time spent in system *per customer*, called the average system time, is:

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

- For stable systems: $\hat{w} \to w$ as $N \to \infty$ with probability 1, where *w* is called the long-run average system time..
- If the system under consideration is the queue alone (with W_i^Q time customer *i* spends in the queue):

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

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

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

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

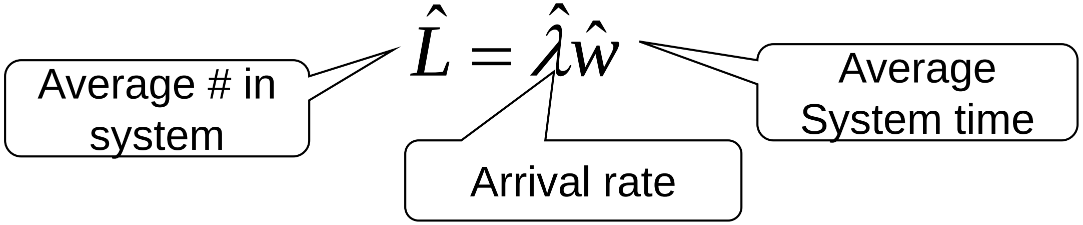
- Assumptions are single server, FIFO queue discipline.
- Time spent in waiting line is:

$$\hat{w}_Q = \frac{W_1^{\ Q} + W_2^{\ Q} + \ldots + W_5^{\ Q}}{5} = \frac{0+0+3+3+0}{5} = 1.2 \text{ time units}$$

The Conservation Equation (Little’s Law)

[Characteristics of Queueing System]

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



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

Average # in system

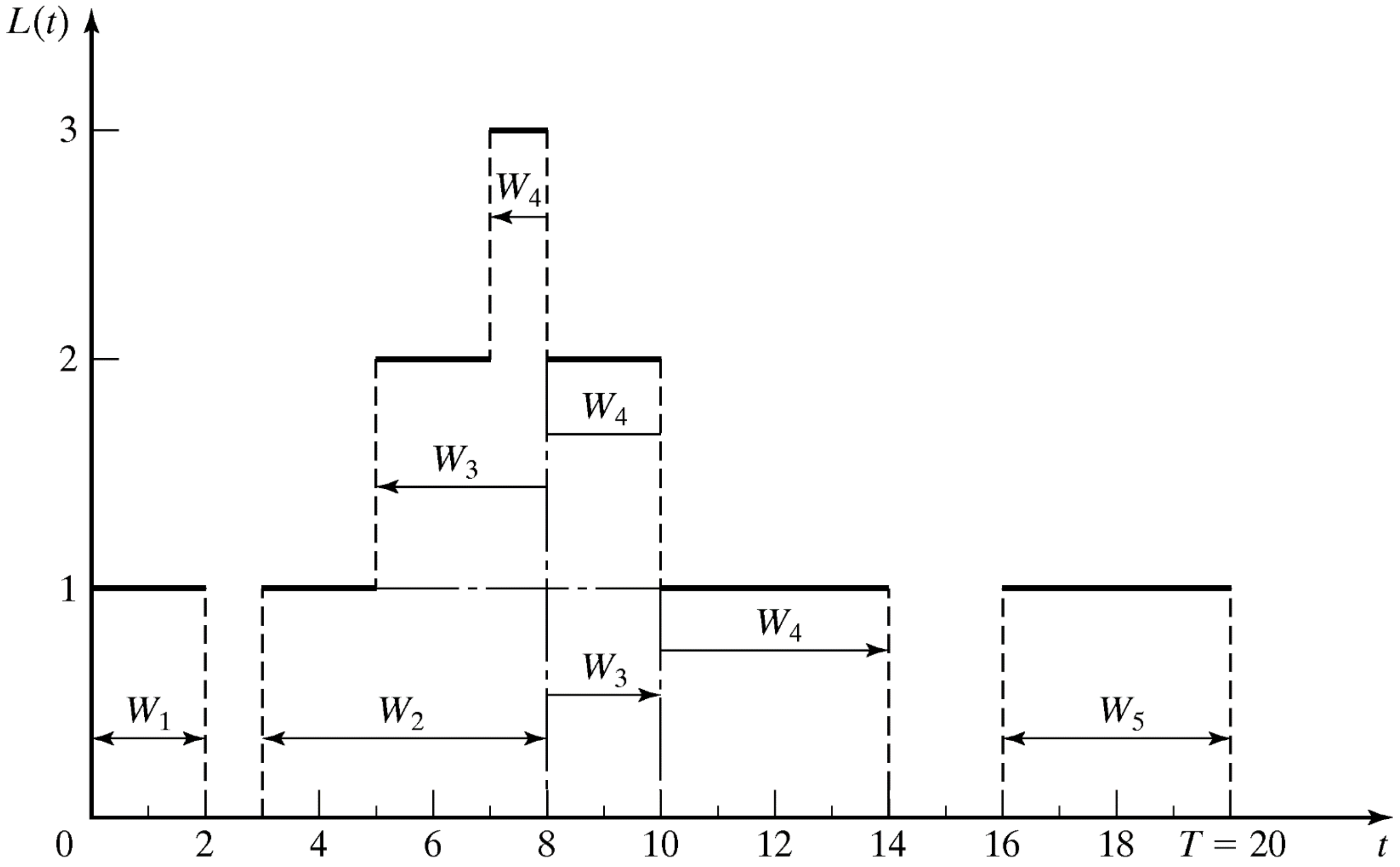
Arrival rate

Average System time

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

- □ 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.

System Time



The Conservation Equation

[Characteristics of Queueing System]

- Total system time for all customers is given by the total area under the number in system function *L(t).*

$$\sum_{i=1}^{N} W_i = \int_0^T L(t)dt$$

- Therefore,

$$\hat{L} = \frac{1}{T}\int_0^T L(t)dt = \frac{1}{T}\frac{N}{N}\sum_{i=1}^{N} W_i = \hat{\lambda}\hat{w}$$

$$\text{where } \hat{\lambda} = N/T$$

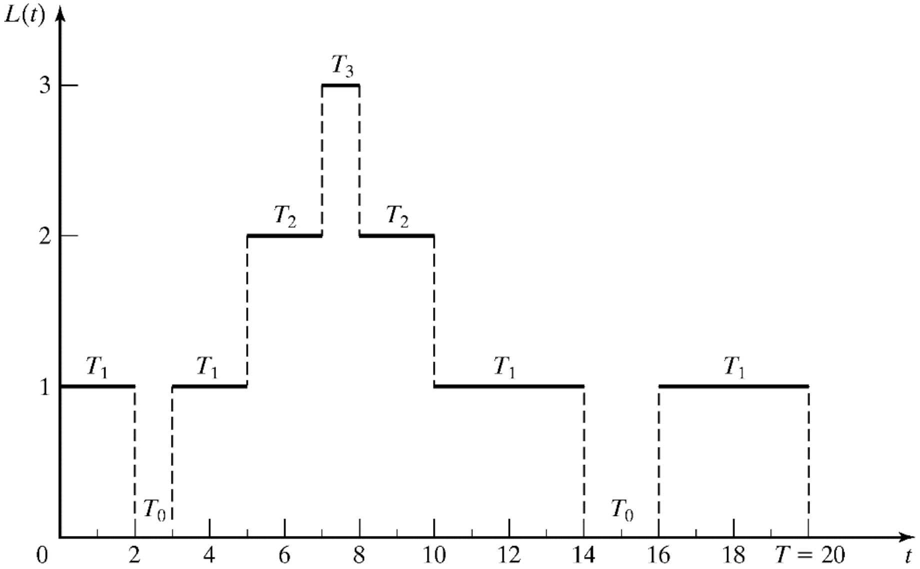
Little's Law

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 ρ.
  - □ For systems with long-run stability: $\hat{\rho} \rightarrow \rho \quad \text{as} \quad T \rightarrow \infty$

Server Utilization Example



- As per the above picture, the server utilization is

$$\hat{\rho} = (\text{total busy time})/T = \left(\sum_{i=1}^{\infty} T_i\right)\Big/T = \left(T - T_0\right)/T = 17/20$$

T_0 is the total idle time

Server Utilization

[Characteristics of Queueing System]

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

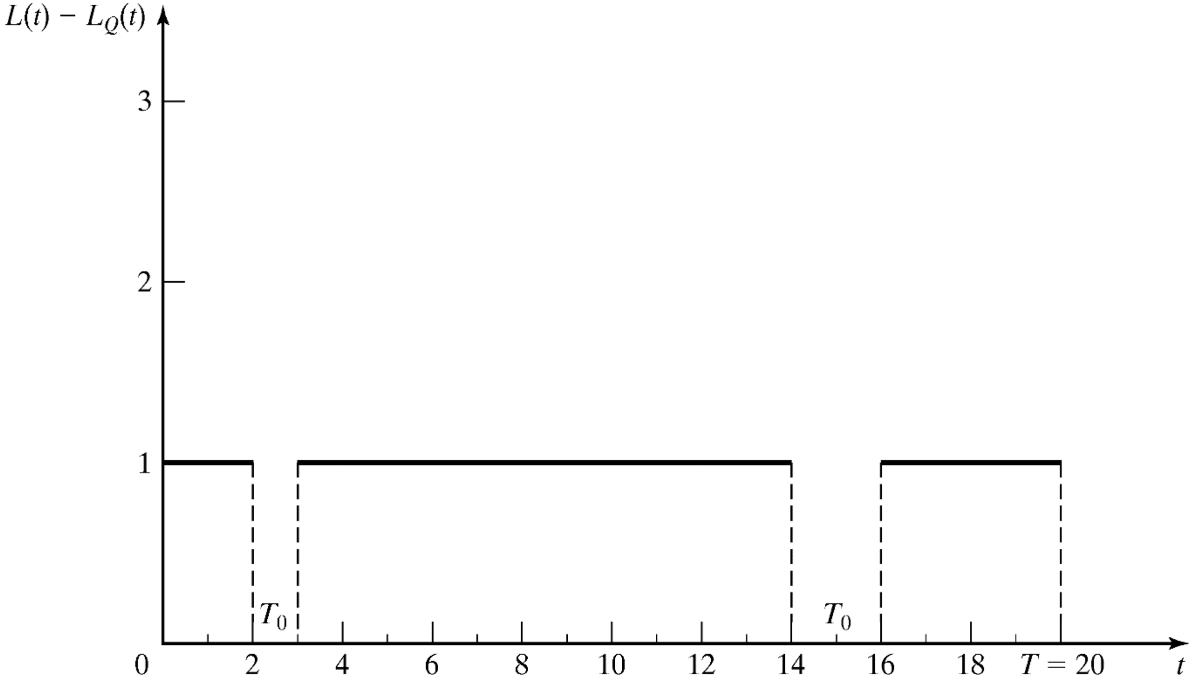
$$\hat{L}_s = \frac{1}{T}\int_0^T \left(L(t) - L_Q(t)\right) dt = \frac{T - T_0}{T}$$

$\hat{\rho}$

What is this??

For a single-server case,
*the **average number of customers** being served at any arbitrary point in time is equal to **server utilization**!!*

Server Utilization



The actual number of customers in the server subsystem is either 0 or 1

Server Utilization

[Characteristics of Queueing System]

- In general, for a single-server queue:

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

and combining this into $L = \lambda w$, for the server sub system, $\rho = \lambda E(s) = \dfrac{\lambda}{\mu}$

- For a single-server stable queue:
  - The arrival rate λ must be less than the service rate μ; $(\lambda < \mu)$

$$\rho = \frac{\lambda}{\mu} < 1$$

- *For an unstable queue* $(\lambda > \mu)$
  - The server is always busy
  - Waiting line tend to grow in length at an average rate of $(\lambda - \mu)$ customers per time unit and long run average queue length is ∞
  - long-run server utilization is *1*.

Server Utilization

[Characteristics of Queueing System]

- For *G/G/c/∞/∞* queues:
  - □ A system with *c* identical servers in parallel. Arrivals occur at a rate λ and each server works at a rate μ customers per time unit.
  - □ 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 Example

- Customers arrive at random to a license bureau at a rate of λ=50 customers per hour. Currently, there are 20 clerks, each serving μ=5 customers per hour on the average.
- Therefore, the long-run or steady state average utilization of a server is

$$\rho = \frac{\lambda}{c\mu} = \frac{50}{20(5)} = 0.5$$

- The average number of busy servers is: $L_s = \frac{\lambda}{\mu} = \frac{50}{5} = 10$
- A typical clerk is busy only 50% of the time!!
- Can we decrease the number of servers??
  - For system to be stable, the number of servers must be

$$c > \frac{\lambda}{\mu} = \frac{50}{5} = 10$$

Possibilities are *c≥11, 12 etc*

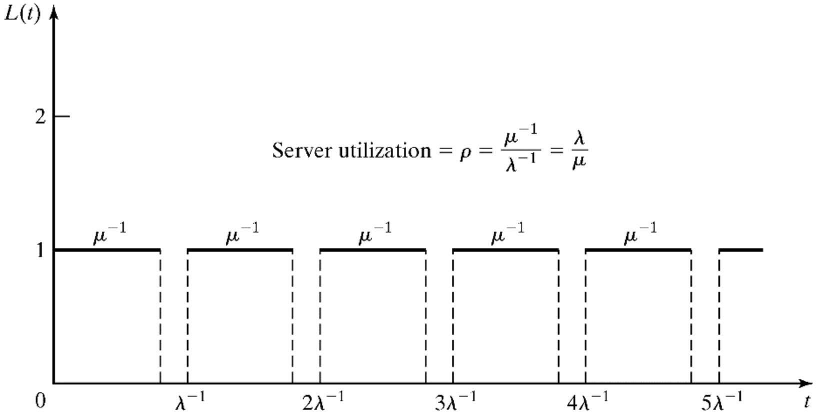
Server Utilization and System Performance

[Characteristics of Queueing System]

- System performance varies widely for a given utilization ρ.
  - For example, a *D/D/1* queue where $E(A) = 1/\lambda$ and $E(S) = 1/\mu$, where:

$$L = \rho = \lambda/\mu, \quad w = E(S) = 1/\mu, \quad L_Q = W_Q = 0.$$

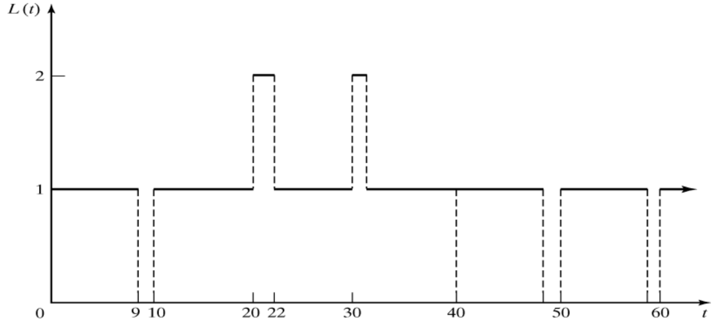
    - By varying λ and μ, server utilization can assume any value between *0* and *1*.
    - Yet there is never any line.
  - In general, variability of inter-arrival and service times causes lines to fluctuate in length.



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, $A_1 = A_2 = \ldots = \lambda^{-1} = 10$.
  - □ Services are stochastic, $E(S_i) = 9.3$ min and $V(S_i) = 0.81$ min^2.
  - □ On average, the physician's utilization = $\rho = \lambda/\mu = 0.93 < 1$.
  - □ Consider the system is simulated with service times: $S_1 = 9$, $S_2 = 12$, $S_3 = 9$, $S_4 = 9$, $S_5 = 9$, …. 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 ≤ c ≤ ∞*) have utilization ρ, 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 = λ).
- Service times may be exponentially distributed (*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 (as a rough guide).
- 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/∞/∞* example: to have a statistical equilibrium, a necessary and sufficient condition is $\lambda/(c\mu) < 1$.

M/G/1 Queues

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

$$\rho = \lambda / \mu, \quad P_0 = (1-\rho) \text{ is the probability server is idle (no customers)}$$

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

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

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

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

M/G/1 Queues [Steady-State of Markovian Model]

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

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

  - *If λ and μ are held constant, L_Q depends on the variability, σ^2, of the service times.*

M/G/1 Queues [Steady-State of Markovian Model]

- ***Example:*** Widget making machines malfunction at random and then require mechanic's attention. Malfunctions occur according to a poisson process at a rate λ = *1.5* per hour. Observation over several months has found that repair times by the single mechanic take an average time of 30 minutes with a standard deviation of 20 minutes.
- Therefore, the mean service time is *1/μ = ½ =0.5 hour,* the service rate μ = 2 per hour and the variance σ^2 = 20^2 *minutes = 1/9 hour.*
- The customers are the widget makers and appropriate model is *M/G/1* because only the mean and variance of the service times are known, not the distribution
- Therefore, $\rho=\lambda/\mu=1.5/2 = 0.75$.
- Steady state time average number of broken machines is:

$$L = 0.75 + \frac{(1.5)^2[0.5^2 + 1/9]}{2(1-0.75)} = 2.375\,\text{customers}$$

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^2:

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

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

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

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

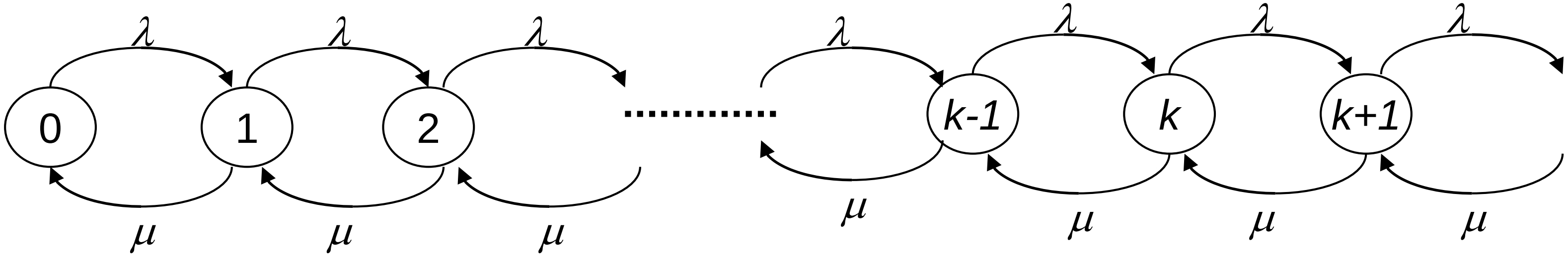
M/M/1 Queues [Steady-State of Markovian Model]

- 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:

$$\boxed{\begin{gathered}\rho = \lambda / \mu, \quad P_0 = (1 - \rho), \quad P_n = (1 - \rho)\rho^n \\ L = \frac{\lambda}{\mu - \lambda} = \frac{\rho}{1 - \rho}, \quad L_Q = \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{\rho^2}{1 - \rho} \\ w = \frac{1}{\mu - \lambda} = \frac{1}{\mu(1 - \rho)}, \quad w_Q = \frac{\lambda}{\mu(\mu - \lambda)} = \frac{\rho}{\mu(1 - \rho)}\end{gathered}}$$

M/M/1 Queues [Steady-State of Markovian Model]

- State transition diagram of *M/M/1 queue*



$$\text{Flow Rate into } E_k = \lambda_{k-1}P_{k-1} + \mu_{k+1}P_{k+1}$$

$$\text{Flow Rate out of } E_k = \lambda_k P_k + \mu_k P_k$$

$$\text{In equilibrium, these two must be the same} \Rightarrow \lambda_{k-1}P_{k-1} + \mu_{k+1}P_{k+1} = (\lambda_k + \mu_k)P_k$$

$$\text{For M/M/1: } \lambda_k = \lambda; \mu_k = \mu$$

$$\therefore P_1 = \frac{\lambda}{\mu}P_0; \qquad P_2 = \frac{\lambda}{\mu}P_1 = \frac{\lambda^2}{\mu^2}P_0; \ldots\ldots; P_k = \frac{\lambda^k}{\mu^k}P_0 = \left(\frac{\lambda}{\mu}\right)^k P_0$$

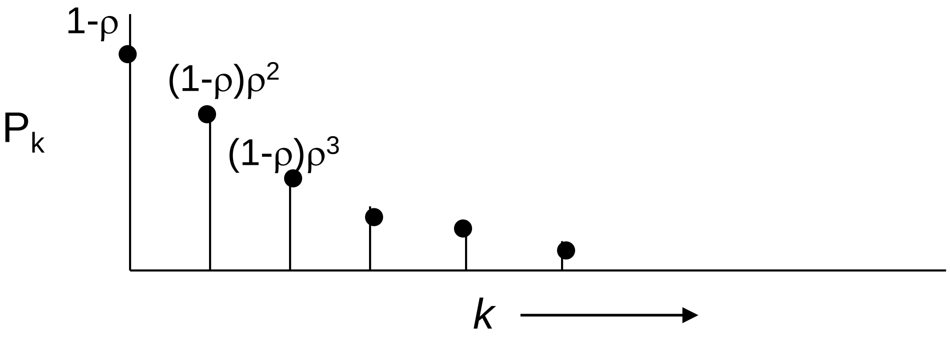
$$\text{We also know that } \sum_{k=0}^{\infty} P_k = 1 \qquad \Rightarrow P_0 = 1 - \frac{\lambda}{\mu} = 1 - \rho$$

$$\therefore P_k = (1-\rho)\rho^k$$

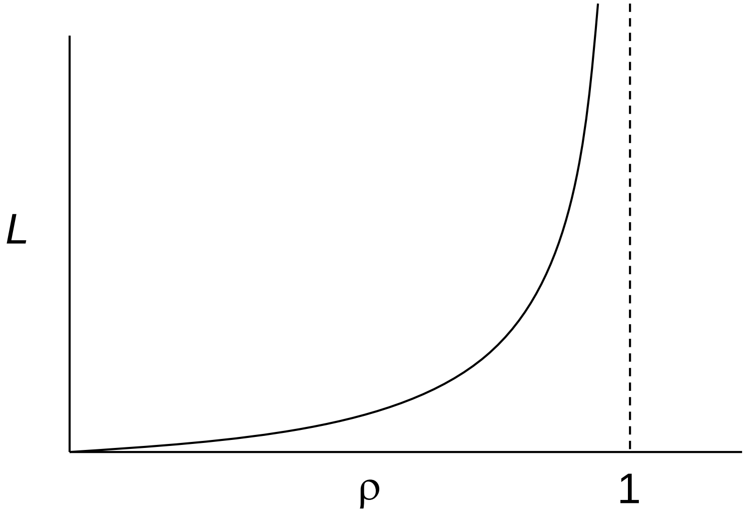
M/M/1 Queues

[Steady-State of Markovian Model]

- Solution for P_k



- The average number in the system *M/M/1*



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, λ, increases from *5* to *8.64* by increments of *20%*:

λ	*5.0*	*6.0*	*7.2*	*8.64*	*10.0*
ρ	*0.500*	*0.600*	*0.720*	*0.864*	*1.000*
L	*1.00*	*1.50*	*2.57*	*6.35*	∞
w	*0.20*	*0.25*	*0.36*	*0.73*	∞

  - □ 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 non-linear as a function of ρ.

M/M/1 Queues [Steady-State of Markovian Model]

- Example: If arrivals are occurring at rate $\lambda = 10$ per hour. Management has a choice of two servers, one who works at rate $\mu_1=11$ customers per hour and second at rate $\mu_2=12$ customers per hour
- Respective utilizations are:
  - □ $\rho_1=\lambda/\mu_1=10/11=0.909$; $\rho_2=\lambda/\mu_2=10/12=0.833$
- Average number in the system is:
  - □ $L_1= \rho_1/(1- \rho_1)=10$; $L_2= \rho_2/(1- \rho_2)=5$
- Thus, a decrease in service rate from 12 to 11 customers per hour, a mere 8.3% decrease would result in an increase in average number in system from 5 to 10, which is 100% increase!!!

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 (ρ) or by decreasing the service time variability (σ^2).
- A measure of the variability of a distribution is coefficient of variation (*cv*):

$$(cv)^2 = \frac{V(X)}{\left[E(X)\right]^2} \quad \text{or} \quad cv = \frac{SD}{E(X)}$$

  - □ The larger *cv* is, the more variable is the distribution relative to its expected value
- For deterministic servers $V(X)=0 \Rightarrow cv = 0$
- For exponential servers, $E(X)=1/\mu$; $V(X)=1/\mu^2 \Rightarrow cv = 1$;

Effect of Utilization and Service Variability

[Steady-State of Markovian Model]

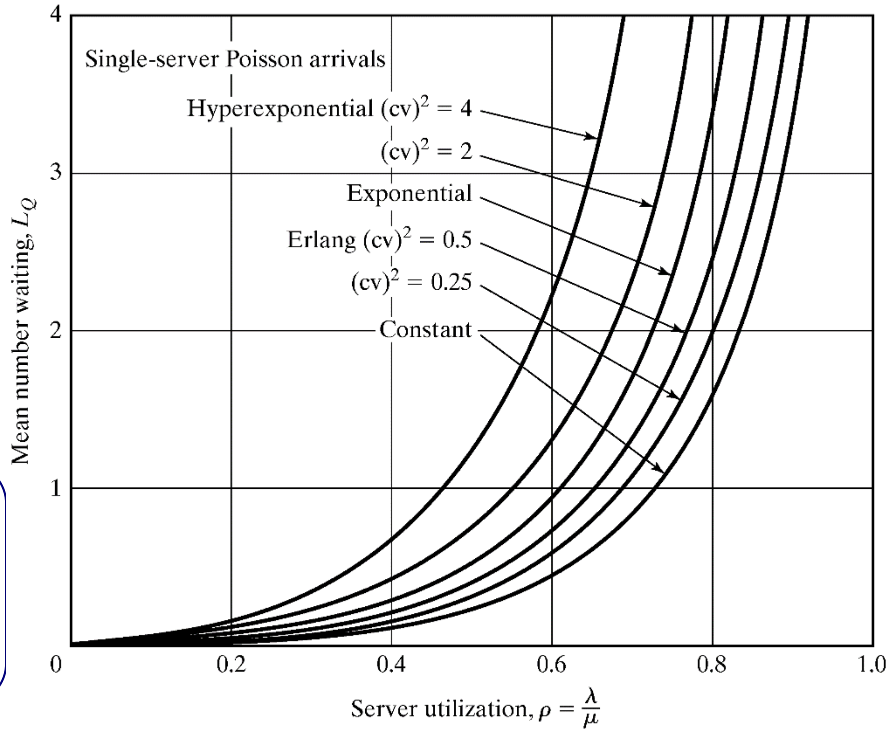
- Consider L_Q for any *M/G/1* queue (see Slide 41):

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

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

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/∞/∞* 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 (λ/μ) must satisfy $\lambda/\mu < c$, where $\lambda/(c\mu) = \rho$ is the server utilization.
  - Some of the steady-state probabilities are given in terms of
    - P_0, the probability that the system is empty
    - $P(L(\infty)\geq c)$ where $L(\infty)$ is a random variable representing the number in the system in statistical equilibrium
      - Implies the probability of all servers busy or $\sum_{n=c}^{\infty} P_n$

Multiserver Queue [Steady-State of Markovian Model]

- *M/M/c/∞/∞* queue: *c* channels operating in parallel:

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

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

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

$$P(L(\infty) \geq c) = \frac{(\lambda/\mu)^c P_0}{c!(1-\lambda/c\mu)} = \frac{(c\rho)^c P_0}{c!(1-\rho)}$$

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

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

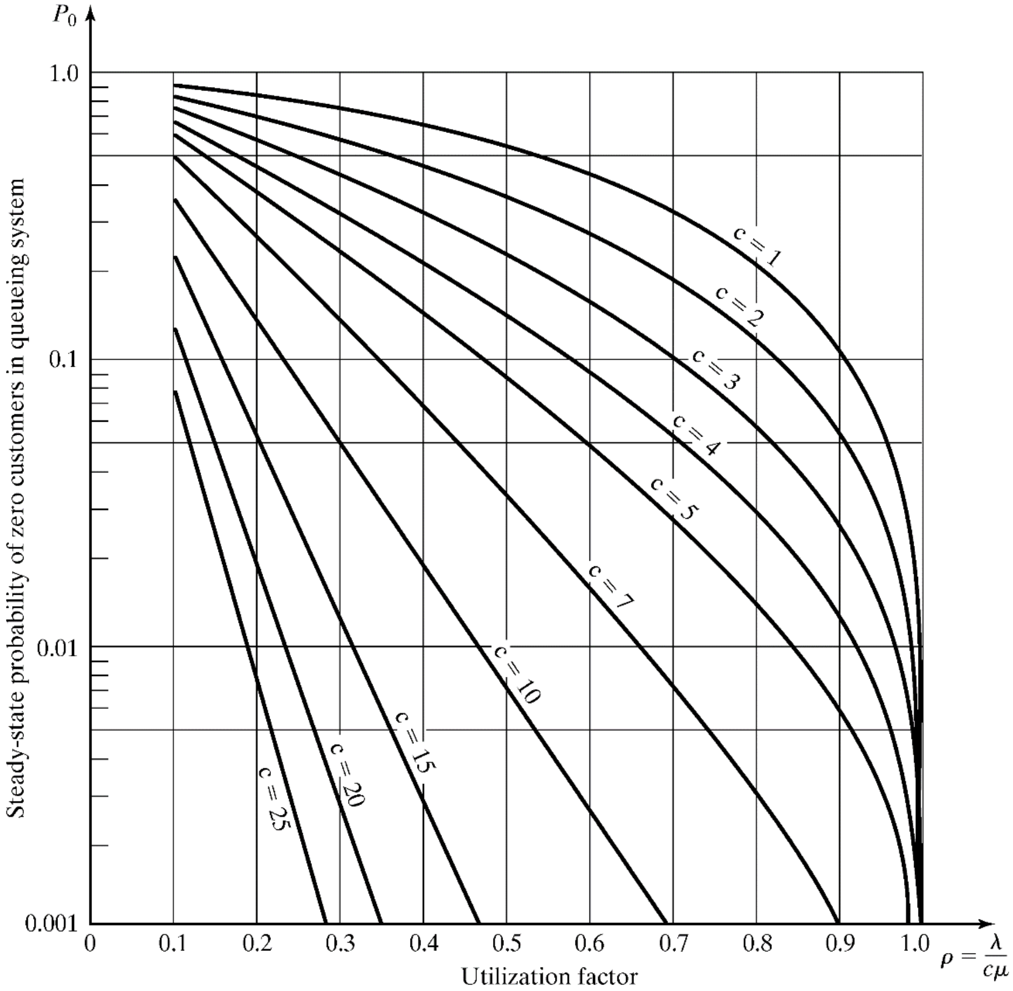
$$w_Q = w - \frac{1}{\mu}$$

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

$$L - L_Q = \frac{\lambda}{\mu} = c\rho$$

Multiserver Queue [Steady-State of Markovian Model]

- P_0 Steady state probability that system is empty depends only on c & ρ



Multiserver Queue [Steady-State of Markovian Model]

- ***Example:*** Attendants manage the tool cribs, as mechanics, assumed to be from an infinite calling population, arrive for service
- Assume Poisson arrivals at rate 2 mechanics per minute and exponentially distributed service times with mean 40 seconds
- $\lambda=2$ per minute $\mu=60/40=1.5$ per minute $\rightarrow$ offered load > 1
- Let there be $c=2$ attendants
- Probability $P_0 = 0.2$
- Probability that all servers are busy $P(L(\infty)\geq 2) = 0.533$
- Time average length of waiting line L_Q = 1.07 mechanics
- Time average number in the system $L = L_Q+\lambda/\mu$ = 2.4 mechanics
- From little's formula $w = L/\lambda$ = 1.2 minutes
- Average time spent waiting $w_Q = w - 1/\mu$ = 0.533 minutes
- Server utilization $\lambda/(c\mu) = \rho$ = 0.667

Multiserver Queue [Steady-State of Markovian Model]

- Other common multiserver queueing models:
  - □ *M/G/c/∞*: general service times and c parallel server. The parameters can be approximated from those of the *M/M/c/∞/∞* model.
    - Use the similar correction factor $(1+cv^2)/2$ that was used in *M/G/1*
  - □ *M/G/∞:* 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/∞*: 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).

Multiserver Queue [Steady-State of Markovian Model]

- Other common multiserver queueing models:
  - □ *M/G/c/∞*: general service times and c parallel server. The parameters can be approximated from those of the *M/M/c/∞/∞* model.
    - Use the similar correction factor *(1+cv²)/2* that was used in *M/G/1*
  - □ *M/G/∞:* general service times and infinite number of servers, e.g., customer is its own system, service capacity far exceeds service demand

$$P_0 = e^{-\lambda/\mu}$$

$$w = 1/\mu;\ w_Q = 0$$

$$L = \lambda/\mu;\ L_Q = 0$$

$$P_n = \frac{e^{-\lambda/\mu}\left(\lambda/\mu\right)^n}{n!},\ n = 0,1,\ldots$$

Multiserver Queue [Steady-State of Markovian Model]

- *M/M/c/N/*∞ queue: *c* channels operating in parallel; Limited Capacity (*N* $\geq$ *c*), when an arrival occurs and the system is full, that arrival is turned away. Service times are exponentially distributed.

$$\rho = \lambda / c\mu; \quad a = \lambda / \mu$$

$$P_0 = \left[1 + \sum_{n=1}^{c} \frac{a^n}{n!} + \frac{a^c}{c!} \sum_{n=c+1}^{N} \rho^{n-c} \right]^{-1}$$

$$P_N = \frac{a^N P_0}{c! c^{N-c}}$$

$$L_Q = \frac{\rho a^c P_0}{c!(1-\rho)^2} \left[1 - \rho^{N-c} - (N-c) \rho^{N-c} (1-\rho) \right]$$

$$\lambda_e = \lambda (1 - P_N)$$

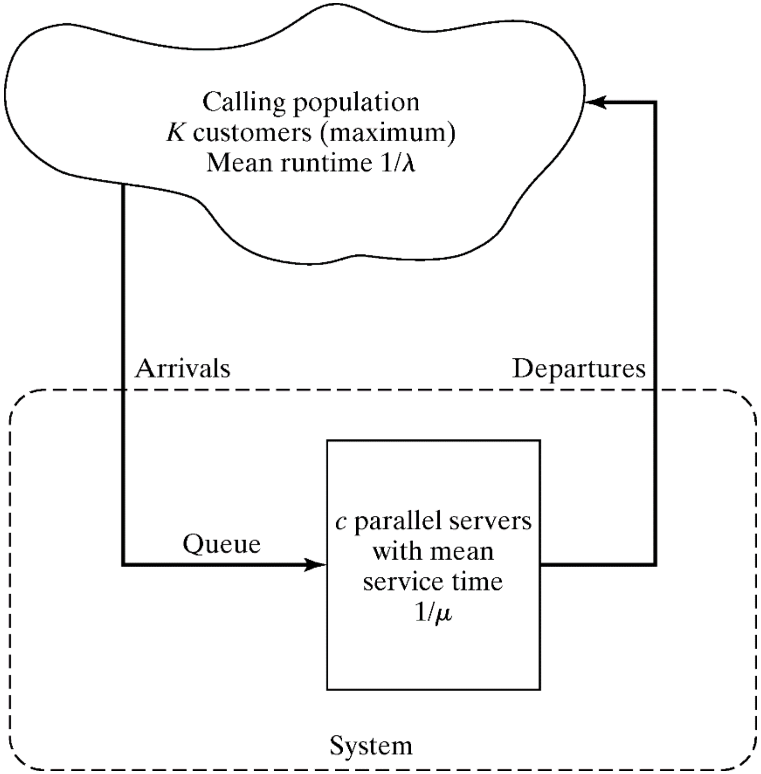
$$w_Q = \frac{L_Q}{\lambda_e}$$

$$w = w_Q + \frac{1}{\mu}$$

$$L = \lambda_e . w$$

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 for *(M/M/c/K/K):*

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

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

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

where λ_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 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: λ = *1/20 and* μ = *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 λ_i, and a fraction $0 \leq p_{ij} \leq 1$ of them are routed to queue *j* upon departure, then the arrival rate form 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 $c_i < \infty$ parallel servers, each working at rate μ_i, then the long-run utilization of each server is $\rho_i = \lambda_j/(c_j \mu_i)$ (where $\rho_i < 1$ 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_i identical servers delivering exponentially distributed service times with mean $1/\mu_j$, then, in steady state, queue *j* behaves likes 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 (see Slide # 14):
  - 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:

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

  - 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, ρ, and λ_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.