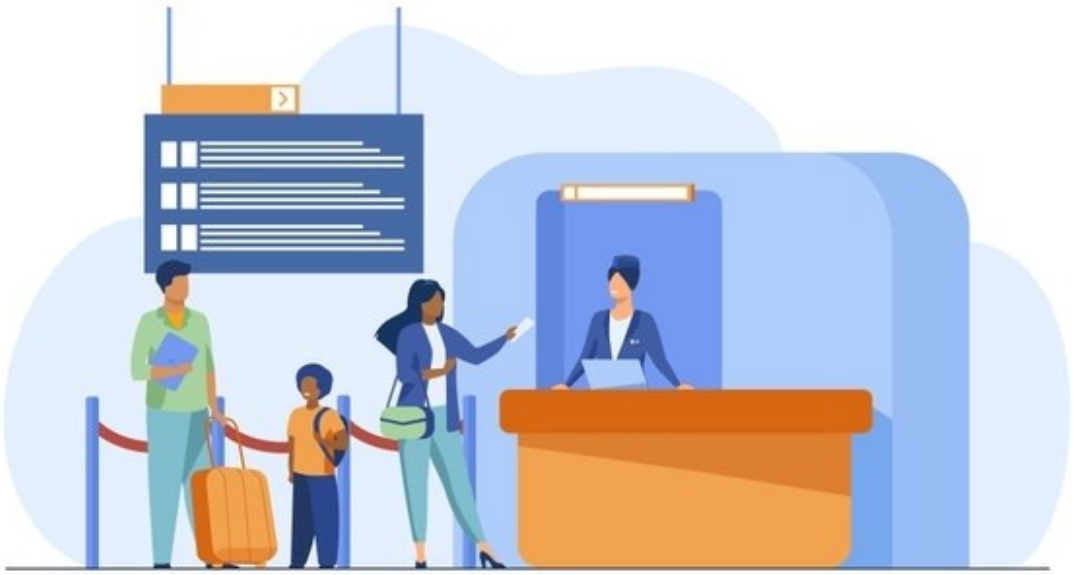
SHARIF
UNIVERSITY OF
TECHNOLOGY

Computer Simulation

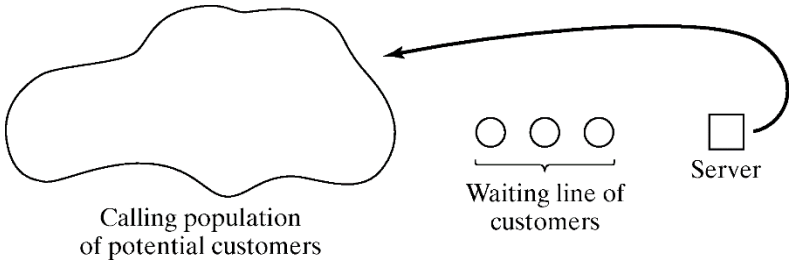
Dr. Bardia Safaei

Chapter Five: Queue Models



Purpose

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



- Queueing models provide 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

SHARIF UNIVERSITY OF TECHNOLOGY

- Discussing some well-known models
  - Not the development of queueing theories
- Explaining the general characteristics of queues
- Definition of important performance measures in queues, and determination of relationships between them
- Estimation of mean measures of performance
- Effect of varying input parameters
  - Especially on output values
- Mathematical solutions for some basic queueing models

Characteristics of Queueing Systems

SHARIF UNIVERSITY OF TECHNOLOGY

- Key elements of queueing systems:
  - **Customer:** Refers to anything that arrives at a facility and requires service
    - Example: people, machines, trucks, emails, packets, airplanes, trains, tasks, etc.
  - **Server:** Refers to any resource that provides the requested service
    - Example: repairpersons, routers, retrieval machines, runways at airport, applications, processors, etc.

Calling Population

SHARIF UNIVERSITY OF TECHNOLOGY

- In a queuing system, the population of potential customers, for **expressing their desired service** is known as calling population
- It may be assumed to be finite or infinite:
  - **Finite population model:** If arrival rate depends on the number of customers being served and waiting
    - Example: model of one corporate jet → if it is being repaired, the repair arrival rate becomes zero
      - Because there are no other planes to be repaired
  - **Infinite population model:** If arrival rate is not affected by the number of customers being served and waiting
    - Example: systems with large population of potential customers

System Capacity

SHARIF UNIVERSITY OF TECHNOLOGY

- System Capacity: The limit on the number of customers that may be in the waiting line, and being served
  - Limited capacity
    - Example: Consider an automatic **car wash**, which has room for only 10 cars to wait in line for entering the mechanism
    - The 11th car cannot wait and it must leave the area
  - Unlimited capacity
    - Example: Concert ticket sales with no limit on the number of people allowed to wait to purchase tickets
    - Note: It is true that the number of tickets is limited, but there is no restrictions in waiting for purchase
- So, similar to calling population, the system capacity could be also limited or not

Arrival Processes (1)

SHARIF UNIVERSITY OF TECHNOLOGY

- Let's consider **infinite-population models**
  - Consider **interarrival times of successive customers**
  - **Random arrivals:** interarrival times usually characterized by a probability distribution
    - Among previously studied models, the most important model is **Poisson arrival process** (with rate λ)
    - Where A_n represents the **interarrival time** between customer $n-1$ and customer n, and is **exponentially** distributed (with mean $1/\lambda$)
  - **Scheduled arrivals:** interarrivals can be **constant** or **constant plus or minus a small random amount** to represent early or late arrivals
    - Example: patients to a physician or flight arrivals to an airport
- In infinite-population models, it is assumed that at least one customer is always present, so the server is never idle
  - Example: When sufficient raw material is available for a machine

Arrival Processes (2)

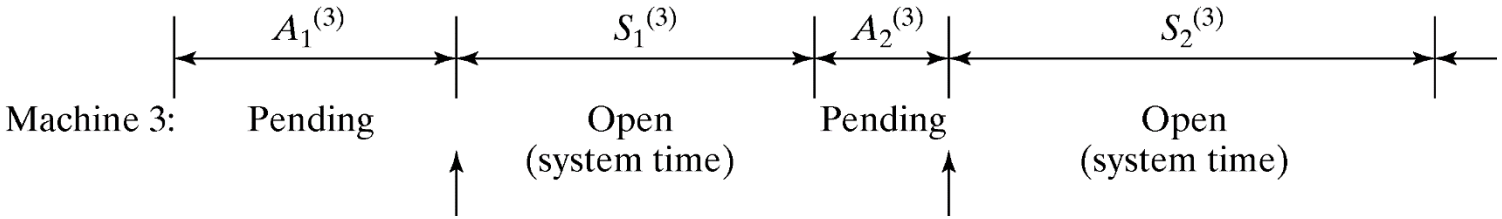
SHARIF UNIVERSITY OF TECHNOLOGY

- Now consider **finite-population models**
- The following definitions are applicable
  - Customer is **pending** when the customer is outside the queueing system
    - In other words, it is neither in the waiting line nor being served
    - Example: machine-repair problem → a machine is "pending" when it is operating normally, it becomes "not pending" the instant it demands service from the repairman and enters the system
  - When the customer enters the system, depending on the available servers, it may be served instantly or with delay (waiting)
  - After service is finished, the time duration from departing the queueing system until that customer's next arrival to the queue is called **runtime**
    - Example: machine-repair problem → Runtime is TTF

Arrival Processes (3)

SHARIF UNIVERSITY OF TECHNOLOGY

- Consider customer i
- Let $A_1^{(i)}, A_2^{(i)}, \ldots$ be the successive runtimes of customer i, and $S_1^{(i)}, S_2^{(i)}, \ldots$ be the corresponding successive system times
  - The following time diagram illustrates the pending, and not pending states for customer i=3



1st and 2nd arrivals of customer 3 to the system.
It may be served instantly or after waiting

Queue Behavior and Queue Discipline

SHARIF UNIVERSITY OF TECHNOLOGY

- **Queue behavior:** the actions of customers while in a queue waiting for service to begin, for example:
  - Balk: Do not join, 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
    - Hoping to get quickly served
    - Only applicable to multi-channel queues
- **Queue discipline:** the policy of ordering 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 (1)

SHARIF UNIVERSITY OF TECHNOLOGY

- Service times of successive arrivals are denoted by $S_1, S_2, S_3, \dots S_n$
  - They may be constant or random → Example: Airport
    - Passport check in the gate
    - Checking-in process with luggage
      - Depends on the number of items, weighting, label issuance, …
  - Depending on the application, $\{S_1, S_2, S_3, \dots\}$ is usually characterized as a sequence of *i.i.d* random variables
    - Or modeled by other distributions, e.g., exponential, Weibull, gamma, lognormal, and truncated normal distribution
- Recall: There are 4 different service mechanisms for queues
  - In every phase, all of the servers are providing service in parallel

Service Times and Service Mechanism (2)

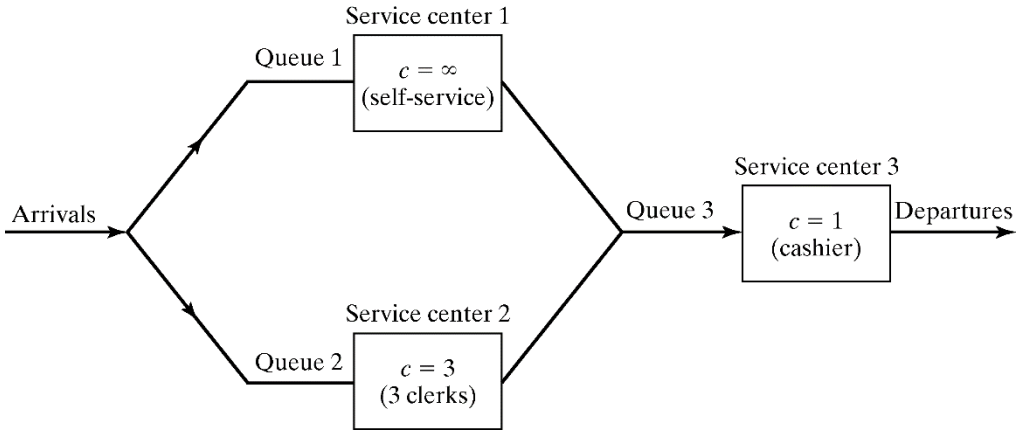
SHARIF UNIVERSITY OF TECHNOLOGY

- **Service center:**
  - Every C servers working in parallel in a phase of a queue
- Upon getting to the head of the line, a customer takes the 1st available server
- Assuming a multi-channel queue, there exists a dedicated service center for every channel
  - It is also possible to have multi-channels with single server
- A queueing system consists of a number of service centers and interconnected queues
  - According to the queueing system, we may have a single queue or a set of communicating queues

Service Times and Service Mechanism (3)

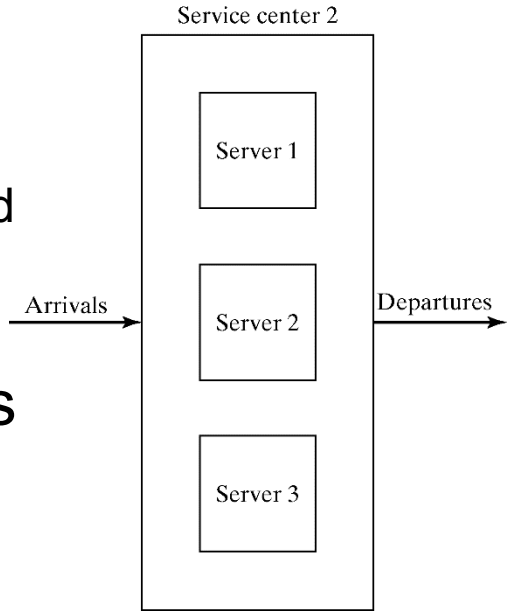
SHARIF UNIVERSITY OF TECHNOLOGY

- Let's discuss an example about service mechanism
  - Consider a discount warehouse, where customers are provided with 2 options for picking their items
    - They can serve themselves before paying at the cashier
      - Picking low-prices items
    - 3 clerks are employed to give consultation, and bring the items (gigantic in size or pricey stuff)
- This queuing system could be divided by 2 phases:
  - Pickup
  - Payment



Service Times and Service Mechanism (4)

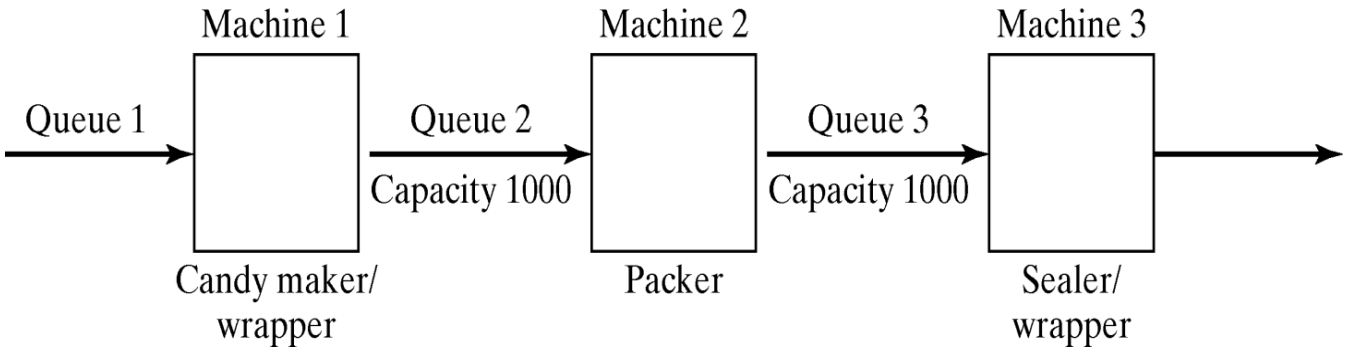
- The pickup phase has 2 queues:
  - One with the self-service providing infinite channels ($C = \infty$)
    - It is not important how many customers are in queue 1 getting service
    - Any new customers can enter without waiting in the line (length of the queue equals to 0)
  - One with 3 clerks ($C = 3$)
    - These clerks are forming a service center providing parallel service to 3 customers
    - Unlike queue 1, the capacity in queue 2 is limited
      - According to FIFO policy, head of the queue gets service from the 1st server, which becomes idle



- After getting service in phase 1, customers enter phase 2 (queue 3) for payment
  - A single-channel/single-phase queue

Group Services

- Some possible service management strategies:
  - If a single server serves several customers simultaneously, it is called as a **batch service**
  - Also, it is possible for a customer to require service from several servers simultaneously
- The following depicts a batch service, where every machine can serve (produce) several items



Queueing Notation

SHARIF UNIVERSITY OF TECHNOLOGY

- For representing queuing systems, especially those with parallel servers, a standard notation system is used
- This notation is denoted as A/B/c/N/K
  - A represents the interarrival-time distribution for customers
  - B represents the service-time distribution
  - c represents the number of parallel servers in a service center
    - Usually is a single phase of the queue
  - N represents the system capacity
  - K represents the size of the calling population
    - Finite or infinite
- Notations for A, and B
  - **M** → Exponential or Markovian, **D** → Deterministic, $\mathbf{E_k}$ → k-order Erlang, **PH** → Phase distribution, **H** → Hyper exponential, **G** → General, etc.

Performance Measures

SHARIF UNIVERSITY OF TECHNOLOGY

- Primary performance measures of queueing systems:
  - $\mathbf{P_n}$: Steady-state probability of having n customers in system
    - As we saw in Chapter 4, we calculate P_n as $t \to \infty$
    - For instance P_0 indicates the probability of having 0 customers in the system
  - $\mathbf{P_n(t)}$: Probability of having n customers in system at time t
    - This is the transient responses for the system
  - $\boldsymbol{\lambda}$: Arrival rate for jobs, requests, customers, etc.
  - $\boldsymbol{\lambda_e}$: Effective arrival rate (discussed in examples)
  - $\boldsymbol{\mu}$: Service rate of one server
    - Use indices in case we have several servers with different service rates (unless that have identical rates)
  - $\boldsymbol{\rho}$: Server utilization
    - Indicating the fraction of time a server is busy

Performance Measures

SHARIF UNIVERSITY OF TECHNOLOGY

- Primary performance measures of queueing systems:
  - $\mathbf{A_n}$: Interarrival time between customers n-1 and n
  - $\mathbf{S_n}$: Service time of the n^{th} arriving customer
  - $\mathbf{W_n}$: Total time spent in **system** by the n^{th} arriving customer
  - $\mathbf{W_n^Q}$: Total time spent in the waiting line by customer n
  - $\mathbf{L(t)}$: The number of customers in **system** at time t
  - $\mathbf{L_Q(t)}$: The number of customers in queue at time t
  - $\mathbf{L}$: Long-run time-average number of customers in system
  - $\mathbf{L_Q}$: Long-run time-average number of customers in queue
  - $\mathbf{w}$: Long-run average time spent in system per customer
  - $\mathbf{w_Q}$: Long-run average time spent in queue per customer

Time-Average Number of Customers L (1)

SHARIF UNIVERSITY OF TECHNOLOGY

- Consider a queueing system, which we have analyzed its behavior over a period of time T
  - We know that number of customers changes during time
  - Assume a time interval as a time unit
  - Let T_i denote the total time during [0,T] in which the system contained exactly i customers
    - Example: $T_1 = 12$ indicates that in the entire time interval [0,T], in 12 time intervals, only 1 customer was present in the system
- The time-weighted-average number of customers in the system is defined by:

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

Time-Average Number of Customers L (2)

- If we plot the number of customers in the system in every instance of time (interval), the L(t) diagram is obtained
  - Then, if we divide the plot into rectangles with intervals (T_i) as their width, and number of customers (i) as their length:

Sum of rectangle areas $$\sum_{i=0}^{\infty} iT_i = \int_0^T L(t)dt$$ Area under the curve

- Considering the previous equations:

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

- The **long-run** time-average number of customers in the system:
  - Steady-state behavior

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

Time-Average Number of Customers $\hat{L}_Q$ (1)

- A similar equation exists for the time-weighted-average number of customers in the queue, which is:

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

  - As we seen before, with having a long period observation (simulation), $\hat{L}_Q$ moves towards L_Q

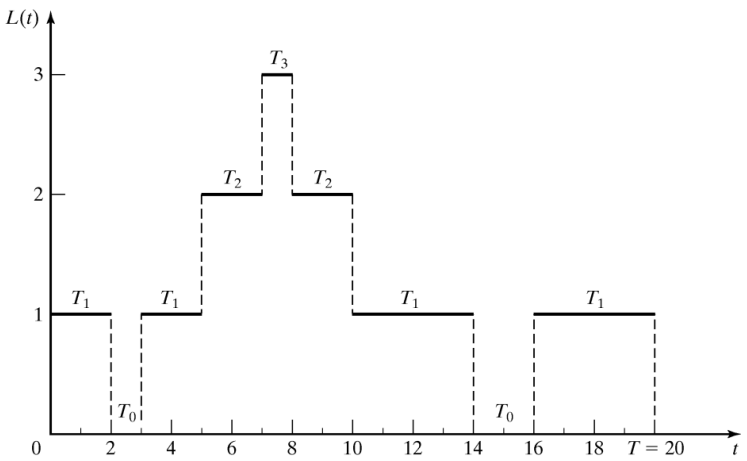
- Example: Consider a G/G/1/N/K queuing systems
  - $N \geq 3$, and $K \geq 3$
  - The simulation table for this system is depicted

Costumer	Arrival Time	Service Time	System Time	Finish Time
1	0	2	2	2
2	3	5	5	8
3	5	2	5	10
4	7	4	7	14
5	16	4	4	20

Time-Average Number of Customers $\hat{L}_Q$ (2)

- It has been assumed that the number of customers in the queue follows the following equation:

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



- The number of customers in the system is depicted as a function of t (L(t))

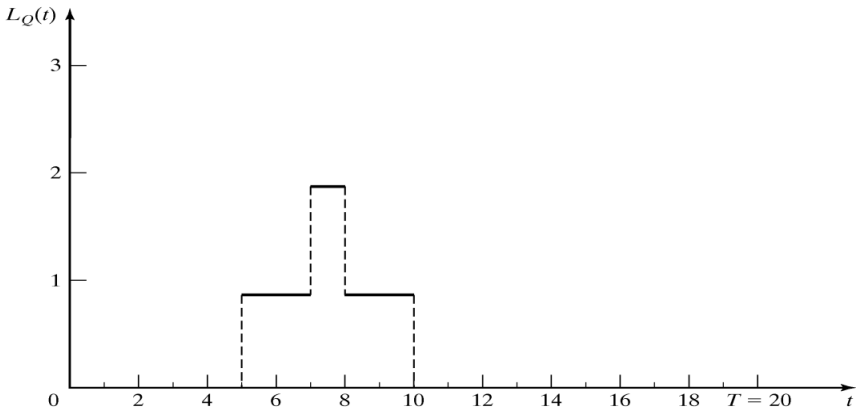
$$\hat{L} = \frac{1}{T}\sum_{i=0}^{\infty} iT_i = [0(3) + 1(12) + 2(4) + 3(1)] / 20$$

$$= 23 / 20 = 1.15 \text{ cusomters}$$

Time-Average Number of Customers $\hat{L}_Q$ (3)

SHARIF UNIVERSITY OF TECHNOLOGY

- Let's talk about $\hat{L}_Q$
- According to the relation between $\hat{L}$, and $\hat{L}_Q$, the diagram for $\hat{L}_Q$ could be obtained by decreasing the values by 1 unit
  - In those cases, where L(t)=0, the value of $\hat{L}_Q$ remains 0
  - The plot could have been also obtained by the simulation table
    - How?
- The number of customers in the queue is depicted as a function of t ($L_Q(t)$)



$$\hat{L}_Q = \frac{1}{T}\sum_{i=0}^{\infty} iT_i^Q = \frac{0(15)+1(4)+2(1)}{20}$$

$$= 0.3 \text{ customers}$$

Average Time Spent in the System Per Customer w

SHARIF UNIVERSITY OF TECHNOLOGY

- The average time spent in the system per customer, called the **average system** time, is: $\hat{w} = \frac{1}{N}\sum_{i=1}^{N} W_i$
  - Where $\{W_1, W_2, \ldots, W_N\}$ are the individual times that each of the N customers spend in the system during [0,T]
  - Example: Consider the system time column in table of Slide 21 for our previous G/G/1/N/K queuing system

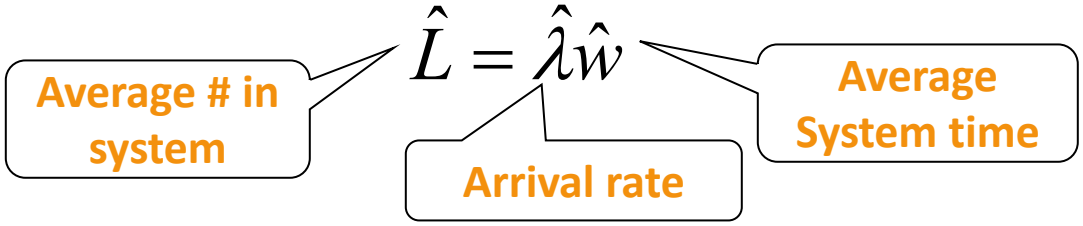
$$\hat{w} = \frac{W_1 + W_2 + \ldots + W_5}{5} = \frac{2 + (8-3) + \ldots + (20-16)}{5} = 4.6 \text{ time units}$$

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

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

The Conservation Equation

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



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

- This is not accidental, and 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, 1 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

SHARIF UNIVERSITY OF TECHNOLOGY

- Definition: the proportion of time that a server is busy
  - During a specific time interval [0,t], which the system is being analyzed, the server utilization is denoted with $\hat{\rho}$
  - In long-run, the server utilization is indicated by ρ
  - For systems with long-run stability:

$$\hat{\rho} \to \rho \quad \text{as} \quad T \to \infty$$

- Example: In our previous G/G/1/N/K queuing system, there is only a single server
  - Utilization of this server is obtained as:

$$\hat{\rho} = \frac{Total\ time\ server\ is\ busy}{Total\ simulation\ time} = \frac{\sum_{i=1}^{\infty} T_i}{T} = \frac{T - T_0}{T} = \frac{17}{20} = 0.85$$

    - In 85% of the time, the server is in operation

Server Utilization For G/G/1/∞/∞ queues (1)

SHARIF UNIVERSITY OF TECHNOLOGY

- Consider a single server queueing system with following assumptions:
  - Average arrival rate λ customers per time unit
  - Average service time $E(S) = 1/\mu$ time units
  - Infinite queue capacity and calling population
- The conservation equation, $L = \lambda w$ can be applied
- For a stable system, the average arrival rate to the server is less than the service rate
  - No lines will be created → Why?
    - Average time spent in the system is identical to average service time
    - $w = E(S) = 1/\mu$
- Therefore, based on the conservation law, $L = \lambda/\mu$

Server Utilization For G/G/1/∞/∞ queues (2)

SHARIF UNIVERSITY OF TECHNOLOGY

- In such queuing system, the average number of customers in the system (server) is calculated as:

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

- As we seen in the previous example, in a single-server queue, $\frac{T-T_0}{T} = \rho$
  - Accordingly, $\hat{L}_s = \hat{\rho} \rightarrow L_s = \rho$ as $T \rightarrow \infty$ and $\rho = \lambda E(s) = \frac{\lambda}{\mu}$
- For a single-server **stable** queue: $\rho = \frac{\lambda}{\mu} < 1$
- For an unstable queue, where $\lambda > \mu$, the long-run server utilization is 1

Server Utilization For G/G/c/∞/∞ queues

SHARIF UNIVERSITY OF TECHNOLOGY

- Now, consider a system with c identical servers in parallel
  - The distribution of arrivals and services are the same as before
  - The queue capacity and calling population are also infinite
  - If an arriving customer finds more than one server idle, the customer chooses a server without favoring any particular server
- Since we have c parallel servers, the rate of service will be multiplied by c → $c\mu$
- Therefore, the long-run average server utilization is:

$$\rho = \frac{\lambda}{c\mu}$$

- Similarly, in order to have a stable system, the rate of arrival λ must be less than $c\mu$, unless, the line will grow

Server Utilization and System Performance (1)

- Different performance metrics in a queuing system depend on the value of ρ
  - Example: average number of customers in the system/queue
  - Example: average time a customer spends in the system/queue
- Roll back to the **stable** G/G/1/∞/∞ queuing system, where $E(A) = 1/\lambda$, and $E(S) = 1/\mu$
  - We know $L = \rho = \lambda/\mu, \quad w = E(S) = 1/\mu, \quad L_Q = W_Q = 0$
    - By varying λ and μ, server utilization can assume any value between 0 and 1
    - So as the number of customers in the system
- In general, variability of interarrival and service rates causes lines to fluctuate in length
  - Applicable to queuing systems with any number of servers

Server Utilization and System Performance (2)

SHARIF UNIVERSITY OF TECHNOLOGY

- Let's consider an example:
  - A physician schedules patients every 10 minutes, and spends S_i minutes with the i^{th} patient:

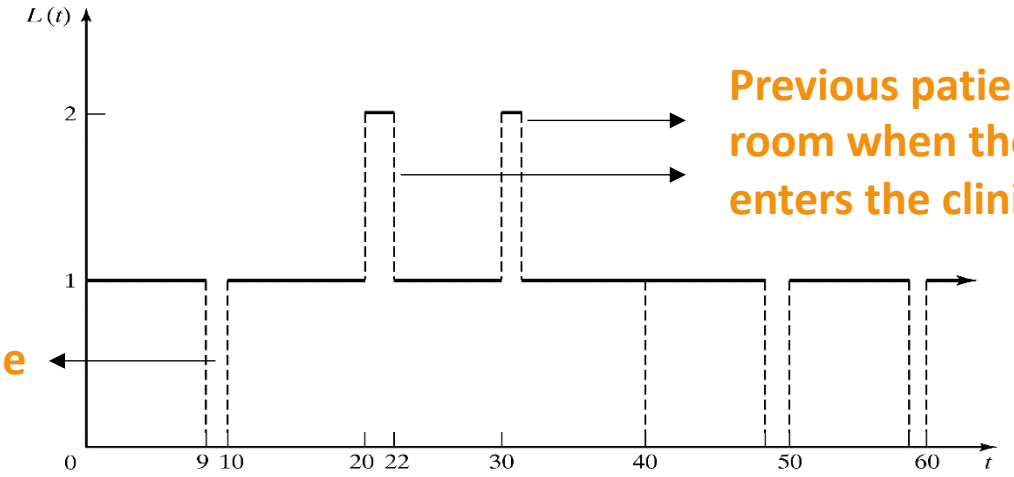
$$S_i = \begin{cases} 9 \text{ minutes with probability } 0.9 \\ 12 \text{ minutes with probability } 0.1 \end{cases}$$

- Arrivals are **deterministic**, $A_1 = A_2 = \cdots = \lambda^{-1} = 10$
- 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 = $\rho = \frac{\lambda}{\mu} = 0.93 < 1$
  - Physician works 93% of the time
  - The system is stable in long term

Server Utilization and System Performance (3)

SHARIF UNIVERSITY OF TECHNOLOGY

- What about **short term**?
  - As long as the patients are served every 9 minutes, the system stays stable (no lines), but since there is a 10% chance that they be served in12 minutes, the clinic may get unstable
  - Assume that the system is simulated with service times: $S_1 = 9,\ S_2 = 12,\ S_3 = 9,\ S_4 = 9,\ S_5 = 9, \ldots$
  - The number of patients in the system (at t) becomes:



Previous patient is still in the room when the new patient enters the clinic

Physician is idle

Server Utilization and System Performance (4)

SHARIF UNIVERSITY OF TECHNOLOGY

- The average waiting time for each of our patients in the queue is as follows:

$$w_1^Q = w_2^Q = 0$$

$$w_3^Q = finishing\ time\ of\ 2\ - arrival\ of\ 3 = 22 - 20 = 2$$

$$w_4^Q = 31 - 30 = 1$$

$$w_5^Q = 0$$

- The occurrence of a relatively long service time ($S_2 = 12$) causes a waiting line to form temporarily
- Find a scenario, which a line with four patients is formed in the clinic
  - Plot L(t)

Costs in Queueing Problems (1)

SHARIF UNIVERSITY OF TECHNOLOGY

- Costs can be associated with various aspects of the waiting line or servers:
  - Example: A 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:

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

W_j^Q is the time customer j spends in queue

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

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

      - This has been achieved according to the conservation law
  - To obtain long term costs, $N \to \infty$, and $T \to \infty$

Costs in Queueing Problems (2)

SHARIF UNIVERSITY OF TECHNOLOGY

- Servers may also impose costs to the system
  - Consider a group of c parallel servers ($1 \leq c < \infty$)
  - Each has a utilization ρ
  - Every server imposes a cost of $5 per hour while operating
    - The total cost of this service center is $\$5 \times c\rho$ per hour
- In case that the **idle state** of these servers imposes cost to the provider, what would be the total cost?

$$\$5c \times (1 - \rho)$$

Fraction of time the server is idle

Steady-State Behavior of Infinite-Population Markovian Models (1)

SHARIF UNIVERSITY OF TECHNOLOGY

- Recall: In Markovian models, the distribution of interarrival times is exponential with mean rate = λ
  - Meanwhile, the service times may be exponentially distributed (M) or arbitrary (G)
- A queueing system is in **statistical equilibrium** if the probability of being in a given state is not time dependent:

$$P(\,L(t) = n\,) = P_n(t) = P_n$$

- Mathematical models in the rest of this chapter can be used to obtain **approximate results** regarding the behavior of the system in statistical equilibrium
  - Even if the model assumptions do not strictly hold

Steady-State Behavior of Infinite-Population Markovian Models (2)

SHARIF UNIVERSITY OF TECHNOLOGY

- For the simple models studied in this chapter, the steady-state value for the time-average number of customers in the system is:

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

  - Where, P_n represents the probability of having exactly n customers in the system
- We mentioned that L plays an important role in specifying the performance of the system in steady-state
  - In other words, if L is known, other metrics could be obtained by applying the little's law
    - Either for the system, or for the queue:

For the system $$L = w\lambda \rightarrow w = \frac{\mathrm{L}}{\lambda}$$

Steady-State Behavior of Infinite-Population Markovian Models (4)

SHARIF UNIVERSITY OF TECHNOLOGY

- The time spent in the queue could be obtained by subtracting service time from the system time:

For the queue $$w_Q = w - E(S) \rightarrow w_Q = w - \frac{1}{\mu} \Rightarrow$$

$$L_Q = \lambda w_Q$$

- In G/G/c/∞/∞ queues, to have a statistical equilibrium, the necessary and sufficient condition is:

$$\frac{\lambda}{c\mu} < 1$$

Steady-State Behavior for M/G/1 Queues (1)

SHARIF UNIVERSITY OF TECHNOLOGY

- Single-server queues with Poisson arrivals, unlimited capacity, and unlimited population
- **Suppose** service times have mean $1/\mu$, and variance σ^2, and $\rho = \frac{\lambda}{\mu} < 1$
  - So, the system will be in equilibrium state
- The steady-state parameters of M/G/1 queue will be calculated as follows:

$$\rho = \lambda / \mu, \quad P_0 = 1 - \rho$$

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

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

Steady-State Behavior for M/G/1 Queues (2)

- Unlike other parameters in the table, there is no simple expression for the steady-state probabilities $P_1, \ldots, P_n$
  - Do not confuse with P_0, which is simply equal to the fraction of time the server is idle $(1 - \rho)$
- In **single server queues**, $L - L_Q = \rho$, gives us the time-average number of customers **being served**
  - This equation could assist us in many analysis
- The average length of the queue, L_Q in previous slide, can be rewritten in 2 parts:

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

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

M/G/1 Queue Performance Analysis Example (1)

- Two workers competing for a job, Able claims to be faster than Baker on average, but Baker claims to be more consistent
  - Baker has lower σ^2
- Arrivals are Poisson at rate λ=2 per hour (1/30 per minute)
- If the employer decides to employ based on the **created length of the queue**, which worker will be selected?
  - Able: $1/\mu = 24$ minutes, and $\sigma^2 = 20^2 = 400\ \text{minutes}^2$:

$$\rho = \frac{\lambda}{\mu} = \frac{24}{30} = 0.8 \qquad L_Q = \frac{\rho^2(1+\sigma^2\mu^2)}{2(1-\rho)} = \frac{0.64(1+\frac{400}{576})}{2(0.2)} = 2.71$$

- The proportion of arrivals who find Able idle and thus experience no delay is $1 - \rho = 20\%$

M/G/1 Queue Performance Analysis Example (2)

- The same approach could be applied to Baker:
  - Baker: $1/\mu = 25$ minutes, and $\sigma^2 = 2^2 = 4\ \text{minutes}^2$:

$$\rho = \frac{\lambda}{\mu} = \frac{25}{30} = \frac{5}{6} \qquad L_Q = \frac{\rho^2(1+\sigma^2\mu^2)}{2(1-\rho)} = \frac{0.69(1+\frac{4}{625})}{2(\frac{1}{6})} = 2.1$$

- The proportion of arrivals who find Baker idle and thus experience no delay is $P_0 = 1 - \rho = 1/6 = 16.7\%$
- Comparing two workers:
  - Although by employing Able, the arriving customers may get immediate service without waiting in the queue with higher probability, but due to more variability of service times (less consistency) by Able, length of the queue will be longer against Baker (nearly 30%)

Steady-State Behavior for M/M/1 Queues (1)

SHARIF UNIVERSITY OF TECHNOLOGY

- A special case for M/G/1 queue is the one, which service times are following the exponential distribution instead of having a general distribution
  - They are called M/M/1
  - Suppose the service times are exponentially distributed with mean $1/\mu$, and the variance is $\sigma^2 = \frac{1}{\mu^2}$
    - In case of M/G/1 we were **not allowed** to specify σ^2 based on μ
      - Because the general distribution is not known
- Accordingly, M/M/1 queues are a useful approximate modeling when:
  - **Standard deviation** of service times are approximately equal to their **means**

Steady-State Behavior for M/M/1 Queues (2)

- Since we use Markov process with exponential distribution for service times
  - Unlike M/G/1, in M/M/1 we can determine P_n values (for any $n \geq 0$) based on a closed form equation
    - Recall: in M/G/1, only p_0 had an equation
- In order to obtain steady-state parameters for M/M/1 queues, you can use M/G/1 equations with replacing σ^2 with $\frac{1}{\mu^2}$

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

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

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

Example

- Consider a M/M/1 queue with service rate $\mu = 10\ \frac{Customers}{hour}$
  - We want to analyze that with increasing the arrival rate (λ) from 5 to 10, what happens to w, and L values
  - Assume the steps to be 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	∞

- Analysis:
  - If $\rho = \lambda/\mu$ gets close to 1, waiting lines tend to grow in length
    - As a result, the time the customers spend in the system will also grow
- In M/M/1 queues do not approach ρ to 1
  - Because w, and L in system are highly nonlinear as a function of ρ

Effect of Utilization and Service Variability (1)

SHARIF UNIVERSITY OF TECHNOLOGY

- 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)
- How could we reduce utilization?
  - Reduce the rate of arrival
  - Increase rate of service
  - Increase parallel servers in service centers
- There exists a measure for the **variability of a distribution**
  - Known as the coefficient of variation (CV):
  - The larger CV gets, the more variable is the distribution relative to its **expected value**
  - Why we use CV?
    - We had variance before

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

$$cv = \sqrt{\frac{V(x)}{E(X)^2}} = \frac{\sigma}{E(X)}$$

Effect of Utilization and Service Variability (2)

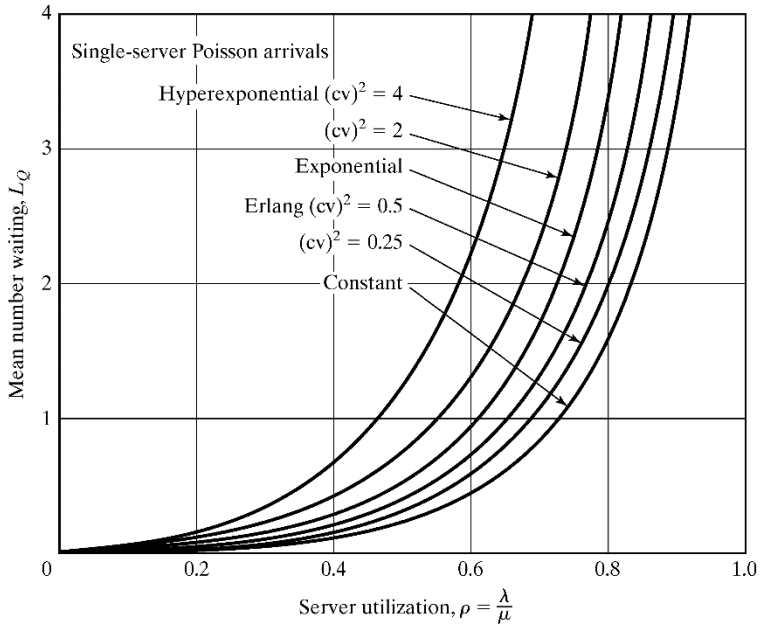
- Bring back the L_Q equation in M/G/1 queues
  - We could rewrite this equation based on CV

$$CV = \frac{\sigma^2}{1/\mu^2} = \sigma^2\mu^2$$

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

L_Q for M/M/1 queues

Is multiplied to the entire equation to consider the non-exponential service time distribution in M/G/1



- Based on this equation, independent from ρ, reducing CV will decrease L_Q (as in reducing ρ) → **Check the plot**

Steady-State Behavior for Multiserver Queues (1)

SHARIF UNIVERSITY OF TECHNOLOGY

- Assume a system with c servers working in parallel
  - Denoted with M/M/c/∞/∞
- Each channel has an independent and identical exponential service-time distribution, with mean $1/\mu$
- Arrival of customers support Poisson with rate λ
  - Customers in the line enter service as soon as one server becomes idle
  - If number of customers $n \leq c$, customers do not wait and get immediate service
    - Otherwise, customers should wait
- To achieve statistical equilibrium, the offered load (λ/μ) must be less than c, where $\frac{\lambda}{c\mu} = \rho$ is the server utilization

Steady-State Behavior for Multiserver Queues (2)

SHARIF UNIVERSITY OF TECHNOLOGY

- If $\lambda > c\mu$, our multisever queue starts moving towards instability with long waiting lines with rate of $\lambda - c\mu$
- Some of the steady-state probabilities:

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

$$L = c\rho + \frac{(c\rho)^{c+1} P_0}{c(c!)(1-\rho)^2} = c\rho + \frac{\rho\, P\left(L(\infty) \geq c\right)}{1-\rho}, \qquad w = \frac{L}{\lambda}$$

- As you can see, obtaining P_0 is the **first step** towards calculating other parameters in multiserver queues

Other Common Multiserver Queueing Models

- M/G/c/∞
  - General service times and c parallel servers
  - The parameters can be **approximated** from those of the M/M/c/∞/∞ model
    - Because general distributions have pretty complex equations
- M/G/∞
  - General service times and infinite number of servers
  - Service (system) capacity **far exceeds** service demand
    - The self service example with no lines
- M/M/c/N/∞
  - Service times are exponentially distributed at rate μ and c servers
  - Obviously, the total system capacity $N \geq c$ customers **(when an arrival occurs to a full system, that arrival is turned away)**

Steady-State Behavior of Finite-Population Models (1)

SHARIF UNIVERSITY OF TECHNOLOGY

- In many real-world applications, assuming an infinite calling population leads into wrong results in our modeling
- When the calling population is finite, 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 denoted with M/M/c/K/K
- Main characteristics of M/M/c/K/K:
  - The time between the end of one service visit and the next call for service is exponentially distributed with mean $1/\lambda$
    - This rate could change as customers leave the calling population and enter the system
  - Service times are also exponentially distributed with mean $1/\mu$
  - c parallel servers and system capacity is K

Steady-State Behavior of Finite-Population Models (2)

SHARIF UNIVERSITY OF TECHNOLOGY

- The steady-state parameters are obtained via following:

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

P_0 again has an important role in calculating other parameters!

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

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

- According to these equations, we have used λ_e instead of λ
  - Where λ_e is the **long run effective arrival rate** of customers to the queue (or rate of entering/exiting service or calling population)

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

Use λ_e only in w, and ρ. Other equations use λ

Example (1)

- Two workers are responsible for working with10 milling machines
  - Machines run on the average for 20 minutes
  - Then, they require an average 5-minute service period
  - Both times are exponentially distributed: $\lambda = 1/20$, and $\mu = 1/5$
- Obtain the performance of this workshop
  - First, specify the queue model?
    - M/M/2/K/K
  - All of the performance measures depend on P0:

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

  - For K<<, we can use pen & paper, but for K>>, use programming
    - Checkout page 262

Example (2)

SHARIF UNIVERSITY OF TECHNOLOGY

- With having P_0, then, we can obtain the other P_n
- Expected number of machines in the system:

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

- Expected number of machines waiting for repair in the queue:

$$L_Q = \sum_{n=3}^{10} (n-2)P_n = 1.46 \text{ machines}$$

  - Why we should start from n=3? We are calculating the average number of machines **in the queue** → no presence doesn't make any sense
  - Why we use n-2?
    - There are 2 repairman in the service center

Example (3)

- The effective rate of arrivals in long-term:

$$\lambda_e = \sum_{n=0}^{10} (10-n)\left(\frac{1}{20}\right) P_n = 0.342 \ \frac{machine}{\text{s}}$$

  - Notice the difference with $\lambda = 0.05$

- After calculating, L_Q, and λ_e, we can calculate the average waiting time for a machine in the queue to get repaired:

$$w_Q = \frac{L_Q}{\lambda_e} = 4.27 \ Minutes$$

- The average number of correctly running machines:

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

  - If $c = 3 \rightarrow K - L = 7.74$, and $if\ c = 1 \rightarrow K - L = 3.98$

Networks of Queues (1)

SHARIF UNIVERSITY OF TECHNOLOGY

- Many systems are naturally modeled as networks of single queues
  - Indicating that customers may depart from one queue and route to another one
- Assuming a stable system with infinite calling population, and no limit on the system capacity, the following could be expected for our system:
  - 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 $0 \leq P_{i,j} \leq 1$, of them are routed to queue j upon departure
    - Then, the arrival rate from queue i to queue j is $\lambda_i P_{i,j}$ (over the long run)

Networks of Queues (2)

SHARIF UNIVERSITY OF TECHNOLOGY

- The other issues, which are expected:
  - The overall arrival rate into queue j:

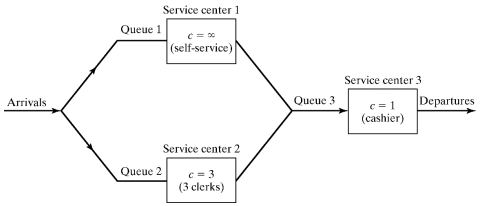
$$\lambda_j = a_j + \sum_{\text{all } i} \lambda_i p_{ij}$$

Arrival rate from outside the network

Sum of arrival rates from other queues in the network

  - If queue j has $c_j < \infty$ parallel servers, each working at rate μ_j, then the long-run utilization of each server is $\rho_j = \lambda_j / c\mu_j$
    - 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 $\frac{1}{\mu_j}$
    - Then, in steady state, queue j behaves likes an M/M/c_j queue with the same arrival rate mentioned above

Example



- Recall our discount warehouse example:
  - Customers arrive with 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
  - Now assume that each of the 3 clerks in service center 2 are providing service with rate $\mu_2 = 20$ customers per hour
- The long-run utilization of the clerks is:
  - $\rho_2 = 48/(3\times 20) = 0.8$
- **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:
  - $\rho_3 = 80/(1\times 90) = 0.89$

Summary

SHARIF UNIVERSITY OF TECHNOLOGY

- Introduced basic concepts of queueing models
- Showed 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 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
  - Are useful in providing a rough estimate of a performance measure