Lecture 3: Intro to Graph & Queuing Theories

CS 5516
Computer Architecture
Networks

VA Tech

Prof. Roy M. Wnek
International Electrotechnical Commission (IEC)

- Founded in 1906 as a result of the International Electrical Congress held in St. Louis in 1904

- A 1976 agreement defines responsibilities
  - The IEC covers the field of electrical and electronic engineering
  - The ISO covers all other subject areas

- Joint technical bodies or working groups cover specific cases of mutual interest
  - The IEC and the ISO formed JTC 1 in 1986 to produce joint standards for information processing

- Mission to promote international cooperation
  - On all questions of standardization
  - Assessment of conformity to standards

- Forum for the preparation and implementation of consensus-based voluntary international standards
IEC Membership

• Membership is 52 National Committees

• Membership represents all electrotechnical interests in each country
  – Manufacturing and service industries
  – Government, scientific, R&D organizations
  – Academic, consumers, and general interest groups

• Standards "aimed at the promotion of optimum community benefits" and represent a common viewpoint of concerned parties

• A reasonable consensus standard for adopting international standards
  – Requires a degree a compromise to be reached in a reasonable amount of time

• Adoption of IEC standards by IEC members is entirely voluntary

• It is up to the National Committees to align their policies at the national level
IEC Publications

- IEC publishes international standards and technical reports
  - Serve as a basis for national standardization
  - References when drafting international contracts.
  - Published in English and French,

- IEC Multilingual Dictionary of Electricity, Electronics and Telecommunications
  - International Electrotechnical Vocabulary (IEV) database

- Technology Trend Assessments (TTAs) booklets
  - Show the state of the art or trends in emerging fields of technology
  - Assist global collaboration on standardization questions in the early stages of technological innovation

- Annual catalog of IEC Publications, Yearbook of work in progress, an Annual Report and a bimonthly newsletter
International Telegraph and Telephone Consultative Committee (CCITT)

- Publishes standards called recommendations

- Chartered by the International Telecommunications Union (ITU)
  - to produce telegraphy and telephone technical, operating, and tariff issue recommendations

- The CCITT is also a United Nations sponsored treaty organization

- U.S. is voting member in the CCITT
  - Representative of the U.S. Department of State
  - Includes technical advisors through the U. S. National Committee for the CCITT

- Publishes approved recommendations every four years in various colored books
United States National Committee for CCITT

- Administered by the U.S. Department of State

- Subcommittees which directly participate in the CCITT international study groups
  - Provide the transfer of standards information between the US and the international community

- Four U.S. CCITT study groups
  - A study group concentrates on services, operations, and tariffs
  - B study group concentrates on ISDN
  - C study group concentrates on the Telephone Network
  - D study group concentrates on data communications
Introduction to Graph Theory

- Two major design constraints
  - Reliability (graph theory)
  - Delay (queuing theory)
    - If the distances are long, propagation delay may be significant
    - As traffic intensity increases, the principal delay becomes the queuing delay

- Reliability is a significant requirements usually imposed on computer networks
  - Given unreliable nodes and lines
  - Provide redundancy
    - Allow some loss in components
    - Keep network functioning properly
    - Perhaps at a lower level of performance

- A graph consists of a set of **Nodes** connected by a set of **Arcs**
  - Node may be called a “vertex”
  - Arc may be called “link, line, branch, and edge”
  - **Adjacent nodes** are directly connected by an arc
  - Nodes which are not adjacent have one or more intermediate nodes
Graph Theory

- **Two kinds of arcs**
  - *Directed arc* indicates that information (or whatever commodity) can flow only in one direction indicated by an arrow and models a simplex (unidirectional) channel
  - *Undirected arc* indicates that information can flow both ways and models full duplex channels
  - *Directed graphs* contain only directed arcs
  - *Undirected graphs* contain only undirected arcs
  - *Mixed graphs* contain both types of arcs

- **Weights** are often assigned to arcs
  - Represents the carrying capacity of the arc
  - e.g. Bits per second in a communications network
  - Other attributes of an arc, e.g. costs, reliability, etc.

- **Parallel arcs may be allowed between nodes**
  - e.g., there are multiple telephone trunks running between cities
  - May be aggregated into a single arc

- **Graphs with self loops are also possible**
  - Arcs beginning and ending at the same node
Weighted, Undirected Graph

Graph Matrix

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>F</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>H</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>I</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>J</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>K</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>L</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Graph Theory Definitions

• A graph represented by a matrix, M
  – M_{ij} is equal to the capacity of the arc from i to j
  – In unweighted graphs
    • if an arc from i to j: M_{ij} = 1
    • if no arc from i to j: M_{ij} = 0
  – If the lines are halfduplex or full duplex, M_{ij} = M_{ji}.
  – If self loops are not allowed, the diagonal elements, M_{jj} = 0

• The **Degree (or valence)** of a node is the number of arcs connected to it
  – Euler theorem:
    *The sum of the degrees of all the nodes in a graph is equal to twice the number of arcs*
    • Each arc contributes to the degree of two nodes
  – A **Regular graph** of degree d is one in which all nodes have degree d

• A **Path** is a sequence of adjacent nodes
  – The **Length** of a path is the number of arcs it contains
  – May be several paths between a pair of nodes
  – Nodes may no occur twice in the sequence
  – A **Geodesic** is the shortest path between two nodes
  – The **Diameter** of a graph is the length of the longest geodesic
Graph Theory Definitions

- A **Circuit** is a sequence of adjacent nodes that comes back to the starting node
  - **Hamiltonian circuit** containing all the nodes of a graph
  - The "traveling salesman" problem consists of finding the shortest Hamiltonian circuit

- If there is a path between every pair of nodes, the graph is said to consist of **1 Component**
  - A 1 component graph is also said to be a **Connected Graph**
  - If one node is isolated from the rest of the graph, the graph would have 2 Components
  - Reliable networks should remain connected even if some nodes and links are removed

- A **Spanning Tree** of a graph is a subgraph
  - Contains all the nodes of the graph and some collection of arcs chosen so that there is exactly one path between each pair of nodes
  - In a weighted graph, the total weight of a spanning tree is the sum of the weights of all its arcs
  - The **Minimum spanning tree** is the spanning tree with minimum total weight
  - Minimum spanning trees are important in designing local access networks
Cuts and Network Flow

• Computing the information carrying capacity of a network
  – Using a cut to determine the carrying capacity

• An XY cut is a set of arcs whose removal disconnects node X from node Y
  – The number of arcs in a cut may vary from one to all the arcs in the graph

• Several minimal cuts may exist between a pair of nodes; Four AH cuts:
  – Cut 1: AB, AE
  – Cut 2: AB, ED, JF, JK
  – Cut 3: BC, FG, KL
  – Cut 4: CH, LH

• In a **Minimal Cut** the replacement of any of its members reconnects the graph
  – In a minimal cut, all of the arcs are essential
  – Arcs AB, AE, and FG form an AH cut
  – This cut is not minimal, because restoring arc FG does not reconnect node A to node H
Cuts and Network Flow

- In a weighted graph, the **Capacity of a cut** is the sum of the weights of the arcs in the cut
  - A **Minimum cut** has the minimum capacity
  - The four AH cuts have capacities 11, 23, 10, and 12, respectively

- Cuts can be used to determine the mean information flow (in bits per second) in each line of a network
A **Feasible flow** meets the following conditions

- 1. The source has no inward arcs (data move away from the source)
- 2. The sink has no outward arcs (it does not emit data)
- 3. No arc has more flow than its capacity (but may have less)
- 4. The inflow and outflow balance at all nodes except source and sink
- 5. The outflow at the source must equal the inflow at the sink
Cuts and Network Flow

• Outflow of seven units from the source, A, and an inflow of seven units into the sink, H
  – At each intermediate node, flow is conserved.

• A feasible flow, it is not the maximum possible flow from A to H

• Not possible to scale up the flow by multiplying the flow in each arc by the same constant
  – Arcs AE, GC, CH, IJ, KL and FD are already saturated

• Maximum flow can be found if all the cuts are known

• Redrawn graph is topologically identical
Cuts and Network Flow

- All flow from node A to node H must pass through the constriction indicated by cut 3.

- The capacity of the cut is 10 units:
  - No feasible flow pattern that can deliver more than 10 units from A to H.
  - Because the cut with the smallest capacity forms the network bottleneck.

- This is the "Max Flow Min Cut Theorem."

The maximum flow between any two nodes in any graph cannot exceed (must equal) the capacity of the minimum cut separating those two nodes.

- The flow limitation comes from the network, not from any inherent limit in the source node.
Connectivity Analysis

• Simple network can be analyzed with queuing theory
  – Average delay times
  – Average queue lengths

• Usually, not necessary or desirable that the connectivity of all nodes be the same
  – Connectivity may reflect demand-generated traffic, communications costs, etc.

• With large, irregular networks, simulation is the only recourse for connectivity analysis
  – Network augmentation questions
    • Where to add the most effective new arcs?
  – Failure conditions questions
    • When will a node lose connectivity?

• Discrete time-base simulation
  – The probability of a link or node failing during any interval is $p_l$ or $p_n$
  – Uniform time intervals
    • During each interval the status of a link or node is evaluated
  – Next event scheduling
    • Network elements are evaluated only when significant events occurs
Connectivity Analysis

• Several possible metrics that can be used for network reliability
  – Simplest is just whether or not the network is connected (probability of disconnection as a function of p)
  – Another useful metric is the fraction of nodes that can still communicate as a function of p
    • e.g. it may be tolerable that under normal conditions 99.9% of node pairs communicate
Queuing System Elements

**RESOURCES**
- Provide a service
- Permit only exclusive use
- Follow a service discipline
- Passive relative to transactions

**TRANSACTIONS**
- Demand a service
- Display persistent demand (queue)
- Follow a queue discipline
- Active relative to servers

**QUEUE**
- Holding area for transactions awaiting service
- Ordered list based on arrival and/or priority
- Characterized by a queue discipline
- Passive relative to the resource and the transactions
Informal Definitions

**QUEUING SYSTEM** - A system in which a finite service-providing resource is allocated to competing demands.

Queuing systems may be used to model the performance of:

- Transportation systems
- Manufacturing systems
- Communications systems
- Computer systems

**SERVICE** - A process of finite time duration that is provided by a resource

**QUEUE DISCIPLINE** - Rules governing the behavior of transactions who have not received service

**SERVICE DISCIPLINE** - Rules governing the behavior of the resource in allocating service
**Queuing Systems Activities**

**PRESENTATION OF DEMAND**
- Check status of resource

**ENQUEUE**
- Obey queue discipline

**SEIZE RESOURCE**
- Busy the resource
- Begin service

**RELEASE RESOURCE**
- End service
- Unbusy the resource
- Obey service discipline
  (Check contents of queue)
Queuing System Characteristics

• QUEUING Systems may configured in a number of ways
  Single systems (single server/ single queue)
  Multiple systems (multi-server/ multi-queue)
  Hierarchical systems (layered arrangements of servers and queues)

• QUEUING System can exhibit different behavior
  Simple systems (Exponentially distributed service and interarrival times with infinite transaction populations)
  Complex system (Service and interarrival times not exponentially distributed or finite transaction populations or with feedback loops)

• Expected performance of simple Queuing Systems may be readily determined using formal queuing theory

• Expected performance of more complex Queuing Systems may require computer simulation
Basic Measures Of Queue Performance

SYSTEM PERFORMANCE

\[ \text{SYSTEM PERFORMANCE} = F (\text{Arrival Rate, Service Rate, Discipline, Numbers of Servers}) \]

- TIME IN SYSTEM or QUEUE or RESOURCE
  - MINIMUM
  - MEAN
  - MAXIMUM

- NUMBER OF TRANSACTIONS IN SYSTEM or QUEUE or RESOURCE
  - MINIMUM
  - MEAN
  - MAXIMUM

- SYSTEM or QUEUE or RESOURCE UTILIZATION

- SYSTEM THROUGHPUT
Basic Queuing Systems

SINGLE QUEUING SYSTEMS

Single Server / Single Queue

MULTIPLE QUEUING SYSTEMS

Multiple Server / Single Queue

Single Server / Multiple Queue

Multiple Server / Multiple Queue
Systems Of Multiple Queues

TASK:

- **SEQUENTIAL**
  - TASK 1 → TASK 2

- **PARALLEL**
  - TASK 1 and TASK 2

- **SELECTION**
  - TASK 1 or TASK 2

- **CONCURRENT**
  - TASK 1 → TASK 2
Introduction To Queuing Theory

QUEUING THEORY - Analytic technique for modeling the performance of queuing systems.

• Background
  - Era of Intuitive Understanding
  - Markov's Definition of Memory-less Properties
  - Erlang's 1917 Work on Telephone-traffic Theory
  - Extension to More Complex Systems

• Utility of Queuing Theory
  - Reliable Method for Forecasting the Performance of Simple Queuing Systems
  - No Need for Computer Support
  - Limited Utility for Forecasting the Performance of Complex Queuing Systems

• Basic Factors Considered in Queuing Theory
  - Interarrival Time Distribution
  - Service Time Distribution
  - Number Of Servers
  - (System Capacity)
  - (Queue/service Discipline)
Resource Scheduling

• Resource scheduling is a key aspect of queue or service discipline? which transaction selected for service next

• Resource scheduling also addresses? how are preempted transactions handled

Some typical resource scheduling algorithms:

• Static scheduling
  - FIRST-COME-FIRST-SERVED (FIFO)
  - LAST-COME-FIRST-SERVE (LIFO)
  - Prioritized Based On Fixed Characteristics

• Dynamic scheduling
  - SHORTEST JOB (SERVICE TIME) FIRST
  - TIME-VARIABLE PRIORITY
  - ROUND ROBIN

• System performance with each algorithm depends upon workload characteristics

• Intended service also affects algorithm quality
Queuing System Notation

Example designations:

TIME DISTRIBUTIONS  DISCIPLINES

M = MARKOVIAN  FIFO = FIRST IN FIRST OUT
G = GENERAL  LIFO = LAST IN FIRST OUT
D = DETERMINISTIC  SIRO = SERVICE IN RANDOM ORDER

The designation of a simple single-server/single-queue system with exponential distributions would be:

\[ M/M/1/\infty/FIFO \]

Arrival Rate: \( \lambda = 1 / \text{Mean Interarrival Time} \)

Service Rate: \( \mu = 1 / \text{Mean Service Time} \)

Traffic Intensity \( \rho = \text{Arrival Rate} / \text{Service Rate} \)

TRAFFIC INTENSITY is the basic measure of system busyness
Poisson Process

- Poisson Processes exhibit the Memory-less Markovian Property

- Required assumptions
  - arrivals drawn from an infinite population
  - the number of arrivals in non-overlapping time intervals is statistically independent of
    - Other time intervals
    - Service
    - Queue length or expected waiting time

- Results in a (Negative) Exponential Distribution or a Poisson Distribution

A Poisson arrival process results in a mean arrival rate of which has a Poisson distribution and a mean inter-arrival time of $1 / \lambda$ which is exponentially distributed.
The Poisson Process

- Characteristics of the Exponential Distribution
  - single parameter, the mean
  - standard deviation = the mean (worst case perspective)
  - accurate description of most human behavior in resource allocation systems at steady state
  - descriptive of any system whose behavior is independent of previous states

- Multiple Poisson processes can be combined into a single process with the same Poisson characteristics
  -- queuing systems can extended to include multiple transaction sources
  -- feeding a common queue
  -- independence of sources
  -- retain exponential distribution
  -- aggregate arrival rate $= \sum S_{\text{Arrival Rates}}$
Basic Measures Of M/M/1/∞ /FIFO Queuing System Performance

- Arrivals And Service Are Poisson Processes

- Assume Simple FIFO Queue/service Discipline
  - No Balking or Jockeying or Reneging

- Assume That The System Is In Steady State
  - Avoidance Of Opening And Closing Transients
  - System Is Operating At Statistical Equilibrium (Long-term Performance)

- Notation
  Traffic = Traffic Intensity \( \rho \)
  Arrival = Arrival Rate \( \lambda \)
  Service = Service Rate \( \mu \)
  Number = Number of Identical Servers \( C \)

- Expected Number Of Transactions In System

\[
L = \frac{\text{Traffic}}{1-\text{Traffic}} = \frac{\text{Arrival}}{\text{Service} \cdot \text{Arrival}} = \frac{\lambda}{\mu-\lambda}
\]

- Expected Number Of Transactions In Queue

\[
L_q = \frac{\text{Traffic}^2}{1-\text{Traffic}} = \frac{\text{Arrival}^2}{\text{Service} \cdot (\text{Service} - \text{Arrival})} = \frac{\lambda^2}{\mu(\mu-\lambda)}
\]
Measures Of Queuing System Performance

- **Expected Time In System**

  "Little's Formula" : \( L = \text{Arrival} \times \text{Wait} \)

  \[
  \text{Wait} = \frac{L}{\text{Arrival}} = \frac{1}{\text{Service} - \text{Arrival}} = \frac{1}{\mu - \lambda}
  \]

- **Expected Time In Queue**

  \[
  \text{Wait}_q = \frac{\text{Arrival}}{\text{Service} (\text{Service} - \text{Arrival})} = \frac{\lambda}{\mu (\mu - \lambda)}
  \]

  "Little's Formula" for queues \( L_q = \text{Arrival} \times \text{Wait}_q \)

- **Expected System Utilization**

  \[
  \text{UTILIZATION} = \frac{\text{Traffic}}{1} = \frac{\text{Arrival}}{\text{Service}} = \frac{\lambda}{\mu}
  \]

- **Expected System Throughput (Stable System)**

  \[
  \text{THROUGHPUT} = \frac{\text{Number} \times \text{Utilization}}{1} = \frac{C \times \rho}{\mu}
  \]

- **Expected System Throughput (Unstable System)**

  \[
  \text{THROUGHPUT} = \frac{\text{Number} \times \frac{1}{\text{Arrival}}}{\text{Utilization}} = \frac{C \times \frac{1}{\lambda}}{\rho}
  \]
• Throughput Measures the Flow Capacity of a System

• Saturated and Unsaturated Systems
  - Relative values of the arrival and service rates
  - $\rho > 1$ is a saturated system

• Stable and Unstable Systems
  - Bounds to the growth of the length of the queue
  - A finite length queue is artificially stable

Throughput in an Unsaturated Systems = Arrival Rate

Throughput in a Saturated Systems = Service Rate
Queuing System Performance

Expected Waiting Time

Expected Number of Transactions
Probability of Arrivals as a Function of Time to a M/M/1 System

\[ \frac{1}{\lambda} = 7 \text{ or } 9 \text{ time units} \]
with discrete Poisson probability distribution

Mean = 7 time units
Mean = 9 time units
Probability of Arrivals as a Function of Time to a M/M/1 System

\[ 1/\lambda = 15 \text{ or } 30 \text{ time units} \]
with discrete Poisson probability distribution
Simulated Queue Length as a Function of Time for a M/M/1 System

\[
\frac{1}{\mu} = 10 \text{ time units} \\
\frac{1}{\lambda} = 7 \text{ or } 9 \text{ time units}
\]
Markov Chains

A Markov Chain is a Sequence of Events or States Generated by a Stochastic Process each of which is Dependant only on the Previous State and the Generating Process.

- Given a Sequence of Random Variables \(X_0, X_1, X_2, \ldots\) which take on values in the set \(\{0,1,2,\ldots\}\)

The Sequence \(X\) is a Markov Chain if for all \(N\) and all Possible Values of \(X_i\)

\[
P(X_n = x \mid X_0 = i_0, X_1 = i_1, X_2 = i_2, \ldots X_n = i_n) = P(X_n = j \mid X_{n-1} = i_{n-1})
\]

- A Markov chain can describe arrivals and service in an idealized Queuing System.

- A Markov Chain is said to be in Steady State if the Probability of entering any State is Equal to the Probability of leaving that State.

\[
P(X=x) = P_x
\]

- If \(X\) is an Random Variable that Counts the number of transactions in a Queuing System when at Steady State.
A Markov Chain

- State N :
  (N Transactions in System)

- Arrival Process :
  (Arrival Rate = $\lambda$ Transactions / Unit Time)

- Departure Process :
  (Service Rate = $\mu$ Transactions / Unit Time)
Derivation Of Steady State Equations Using Stochastic Balance

• Begin with the Expression of Stochastic Balance

Stochastic Balance Exists Between the Inputs and Output of a Given State

\[ \text{In} \ (S_n) = \text{Out} \ (S_n) \]

Into State \( N \):

\[ \Pr (S_n) = \lambda \Pr (S_{n-1}) + \mu \Pr (S_{n+1}) \]

Out of State \( N \):

\[ \Pr (S_n) = \lambda \Pr (S_n) + \mu \Pr (S_n) \]

In Steady State:

\[ \Pr (S_{n+1}) = \frac{\lambda + \mu}{\mu} \Pr (S_n) - \frac{\lambda}{\mu} \Pr (S_{n-1}) \]

• The Steady State of a System is related to its Traffic Intensity

\[ \Pr (S_1) = \frac{\lambda}{\mu} \Pr (S_0) \]
Steady State Equations prescribe the Probability that a system will be in a Given State

- Forecasting the Probability that a System will be in a Given State

\[
P_0'(t) = -\lambda P_0 + \mu P_1(t)
\]

\[
P_n'(t) = -\lambda P_{n-1} - (\lambda + \mu) P_n(t) + \mu P_{n+1}(t)
\]

for \( n \geq 1 \)

- State Equations are More Readily Solved if the System is in Equilibrium; Time is no longer a Factor

\[
\frac{dP_n(t)}{dt} = 0
\]

- Steady State has been reached if the probability of finding N Transactions in the system approaches a Limiting Value

\[
P_1 = \frac{\lambda}{\mu} P_0
\]

\[
P_{n+1} = \frac{(\lambda + \mu)}{\mu} P_n - \frac{\lambda}{\mu} P_{n-1}(t)
\]

for \( n \geq 1 \)
Solution Of Steady State Equations

- Solution using Iterative Techniques involves inductive reasoning to solve for $P_n$

\[
P_n = \left( \frac{\lambda}{\mu} \right)^n P_0
\]

Assume: $\rho = \left( \frac{\lambda}{\mu} \right) < 1$

which is a necessary condition of Steady State, Then $P_0 = (1 - \rho)$

\[
P_n = \rho^n (1 - \rho) \quad \text{for } n \geq 0
\]

- Solution is also possible by obtaining probabilities with generating functions
Multiserver Queuing Systems
(M/M/C)

• Assume FIFO queue/service discipline
• Assume that the system is in steady state
• Assume identical servers and service rates
• In stable systems \( \rho < 1 \) the aggregate mean service rate is
  
  \[ c \cdot \text{mean service rate} \]
  
  if \( c \) or more transactions are in system and

  \[ k \cdot \text{mean service rate} \]
  
  when \( k \) (less than \( c \)) transactions in system.

• The probability that \( k \) transactions are in the system is

\[
P_k = \frac{\lambda^k}{k! \mu^k} P_0 \quad \text{where } 1 \leq k \leq C
\]

\[
P_k = \frac{\lambda^k}{C^{k-C} C! \mu^k} P_0 \quad \text{where } k > C
\]
Multiserver Queuing Systems
(M/M/C)

• Calculating the Probability that Zero transactions are in system is based on the condition that 0 or more Transactions will certainly be In the system

\[ \sum_{k=0}^{\infty} P_k = 1 \]

• The probability that zero transactions are in the system is based on this certainty

\[
P_0 \left[ \sum_{k=0}^{c-1} \frac{\lambda^k}{k! \mu^k} + \sum_{k=c}^{\infty} \frac{\lambda^k}{C^k \cdot C! \cdot \mu^k} \right] = 1
\]

Solving for \( P_0 \)

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

• M/M/C Queuing Systems (C>1) are inherently more stable than M/M/1 queuing systems
  – greater capacity to serve
  – stable Traffic Intensities permissible
    (Traffic Intensity, \( \rho = \frac{\lambda}{C\mu} \))
  – queue storage capacity limitations become critical
  – service discipline
Basic Measures Of M/M/C Queuing System Performance

- As restrictions on M/M/1 queuing systems are removed, the analytical forecasting of their performance becomes more difficult.

- Expected Number of transactions in System

\[
L = \frac{\lambda}{\mu} + \left( \frac{\left(\frac{\lambda}{\mu}\right)^c \lambda \mu}{(C-1)!(C \mu - \lambda)^2} \right) * P_0
\]

- Expected Number of Transactions in Queue

\[
L_q = \left( \frac{\left(\frac{\lambda}{\mu}\right)^c \lambda \mu}{(C-1)!(C \mu - \lambda)^2} \right) * P_0
\]

- Expected Time in System

\[
W = \frac{1}{\mu} + \left( \frac{\left(\frac{\lambda}{\mu}\right)^c \mu}{(C-1)!(C \mu - \lambda)^2} \right) * P_0
\]

- Expected Time in Queue

\[
W_q = \left( \frac{\left(\frac{\lambda}{\mu}\right)^c \mu}{(C-1)!(C \mu - \lambda)^2} \right) * P_0
\]
Limited Capacity Queuing Systems
(M/M/1/MAX)

- Ultimately more realistic than queuing systems that permit unlimited occupancy
- Assume simple FIFO queue/service discipline
- Assume that the system is in Steady State
- For less than "MAX" number of transactions, the behavior of the system is the same as an M/M1/∞
- Max + 1 number of transactions is not allowed; further transactions are turned away
- A stable system will be achieved regardless of the traffic intensity, since the system is self-limiting (constrained by queue length)
- Performance measurement will involve consideration of the effective service rate
Solving The Steady State Equations For M/M/1/MAX Systems

Steady State Equations Are:

\[
\begin{align*}
P_1 &= \frac{\lambda}{\mu} P_0 \\
P_{j+1} &= \frac{\lambda + \mu}{\mu} P_j - \frac{\lambda}{\mu} P_{j-1}, \quad 1 \leq j \leq k - 1 \\
P_k &= \frac{\lambda}{\mu} P_{k-1}
\end{align*}
\]

Since the Probability of the System Being in a Given State Depends upon the Initial State and the Traffic Intensity:

\[
P_j = \rho^j P_0, \quad 1 \leq j \leq k
\]

\[
\sum_{j=0}^{k} P_j = 1 \Rightarrow \sum_{j=0}^{k} \rho^j P_0 = 1
\]

and \( P_0 = \frac{1}{\sum_{j=0}^{k} \rho^j} \)

The Summation of the System's Utilization at Each State Is:

\[
\sum_{j=0}^{k} \rho^j = 1 - \rho^{k+1}, \quad \rho \neq 1
\]

\[
\sum_{j=0}^{k} \rho^j = k + 1, \quad \rho = 1
\]
Limited Capacity Queuing Systems
(M/M/1/MAX)

- Probability of 0 Transactions In System

\[ P_0 = \frac{1 - \rho}{1 - \rho^{k+1}}, \quad \rho \neq 1 \]

\[ P_0 = \frac{1}{k + 1}, \quad \rho = 1 \]

- Probability Of N Transactions In System

\[ P_n = \frac{(1 - \rho)\rho^n}{1 - \rho^{k+1}}, \quad \rho \neq 1 \]

\[ P_n = \frac{1}{k + 1}, \quad \rho = 1 \quad n = 0, 1, 2, \ldots, k \]
Basic Measures Of M/M/1/MAX Queuing System Performance

- Expected Number Of Transactions In System
  \[ L = \frac{\rho [1 - (k + 1)\rho^k + k\rho^{k+1}]}{(1 - \rho^{k+1})(k - \rho)}, \quad \rho \neq 1 \]
  \[ L = \frac{k}{2}, \quad \rho = 1 \]

- Expected Number Of Transactions In Queue
  \[ L_q = L - (1 - P_0) \]

- Expected Time In System
  \[ W = \frac{L}{\lambda} = \frac{L}{\lambda(1 - P_k)} \]
  The Effective Arrival Rate takes into account the probability of saturation:
  \[ \lambda' = \lambda(1 - P_k) \]

- Expected Time In Queue
  \[ W_q = W - \frac{1}{\mu} = \frac{L_q}{\lambda'} = \frac{L_q}{\lambda(1 - P_k)} \]
Priority-based Queue Disciplines

• Non-FIFO Queue Disciplines often rely on a Prioritization scheme to tailor system performance

• Motivations for Non-FIFO Queue Disciples
  – reduce Average Cost Of Providing Service
  – reduce Average Time In Queue
  – reduce Average Number In Queue

• Priority-based Queue Disciplines Fall into two General Classes
  – Static Priority (Transactions Are Assigned
    • an Initial Priority Which Does Not Change)
  – Dynamic Priority (Transaction Priority
    • may Change During Its Time In The System)

• Two Types of Priority-based Queue Disciplines
  – Non-preemptive (Transactions Are Not Removed From Service)
  – Preemptive (Transactions Can Be Removed From Service By One With A Higher Priority)

• Priority-based Queue Disciplines must provide for tie-breaking between Transactions of equal priority