Managing Waiting Lines
Where the Time Goes

In a lifetime, the average person will spend:

- **SIX MONTHS** Waiting at stoplights
- **EIGHT MONTHS** Opening junk mail
- **ONE YEAR** Looking for misplaced objects
- **TWO YEARS** Reading E-mail
- **FOUR YEARS** Doing housework
- **FIVE YEARS** Waiting in line
- **SIX YEARS** Eating
Cultural Attitudes

- “Americans hate to wait. So business is trying a trick or two to make lines seem shorter...” *The New York Times, September 25, 1988*

- “An Englishman, even when he is by himself, will form an orderly queue of one...” *George Mikes, “How to be an Alien”*

- “In the Soviet Union, waiting lines were used as a rationing device...” *Hedrick Smith, “The Russians”*
Waiting Realities

- **Inevitability of Waiting**: Waiting results from variations in arrival rates and service rates.

- **Economics of Waiting**: High utilization purchased at the price of customer waiting. Make waiting productive (salad bar) or profitable (drinking bar).
Laws of Service

- **Maister’s First Law:**
  Customers compare expectations with perceptions.

- **Maister’s Second Law:**
  Is hard to play catch-up ball.

- **Skinner’s Law:**
  The other line always moves faster.

- **Jenkin’s Corollary:**
  However, when you switch to another other line, the line you left moves faster.
Remember Me

- I am the person who goes into a restaurant, sits down, and patiently waits while the wait-staff does everything but take my order.
- I am the person that waits in line for the clerk to finish chatting with his buddy.
- I am the one who *never comes back* and it amuses me to see money spent to get me back.
- I was there in the *first place*, all you had to do was show me some courtesy and service.

*The Customer*
Psychology of Waiting

- **That Old Empty Feeling**: Unoccupied time goes slowly
- **A Foot in the Door**: Pre-service waits seem longer than in-service waits
- **The Light at the End of the Tunnel**: Reduce anxiety with attention
- **Excuse Me, But I Was First**: Social justice with FCFS queue discipline
- **They Also Serve, Who Sit and Wait**: Avoids idle service capacity
Approaches to Controlling Customer Waiting

- **Animate**: Disneyland distractions, elevator mirror, recorded music
- **Discriminate**: Avis frequent renter treatment (out of sight)
- **Automate**: Use computer scripts to address 75% of questions
- **Obfuscate**: Disneyland staged waits (e.g. House of Horrors)
Waiting Time Concepts

- Actual waiting times: *Measurable* time from arrival to beginning of service.
- Perceived waiting times: amount of time that customers *believe* they have waited from arrival to beginning of service.
Waiting Line System

- “Queue” is another name for a waiting line.
- A waiting line system consists of two components:
  - The customer population (people or objects to be processed)
  - The process or service system
- Whenever demand exceeds available capacity, a waiting line or queue forms
- There is a tradeoff between cost and service level.
Basic Queueing Process

Arrivals → Queue → Service
Waiting Line System

A waiting line system consists of two components:

- The customer population (people or objects to be processed)
- The process or service system

Whenever demand exceeds available capacity, a waiting line or queue forms.
Components of Queuing Systems

Queuing System = Queue + Service Unit
Essential Features of Queuing Systems

- Calling population
  - Arrival process
  - Queue configuration
    - Queue discipline
    - Service process
  - Departure
    - No future need for service

- Renege
- Balk
1. Calling Population (Arrivals/ customers)
Terminology

Finite versus Infinite populations:

- Is the number of potential new customers affected by the number of customers already in queue?

For practical intent and purposes, when the population is large in comparison to the service system, we assume the source population to be infinite (e.g., in a small barber shop, 200 potential customers per day may be treated as an infinite population).
Customer Service Population Sources

- Infinite: large pool. Entry/exit has no affect

Example:

- Finite: Number of machines needing repair when a company only has three machines.
- Infinite: The number of people who could wait in a line for gasoline.

The population is the source of customers.
2. Arrival Process
Behavior of Arrivals

- Most queuing models assume arriving customer is patient customer.
- Patient customers are people or machines wait in queue until served and do not switch between lines.
- Life and decision models are complicated by fact people balk or renege.
- **Balking** refers to customers who refuse to join queue because it is too long to suit their needs.
- **Reneging** customers enter queue but become impatient and leave without completing transaction.
Arrival Process

Arrival process

Static
- Random arrivals with constant rate
- Accept/Reject

Dynamic
- Random arrival rate varying with time
- Facility-controlled
- Price
- Appointments
- Customer-exercised control
- Reneging
- Balking
Degree of Patience

No Way!

BALKING

Line too long?
Customer balks (never enters queue)

No Way!

RENEGING

Line too long?
Customer reneges (abandons queue)
Arrival

Pattern

- Controllable
- Uncontrollable

Size of arrival

- Single
- Batch

Distribution

- Exponential or Poisson
- Erlang
- Other
- Patient (in line and stay)

Degree of patience

- Impatient

Arrive, view, and leave
Arrive, wait awhile, then leave
Distribution of Arrivals

- Arrival rate: mean (average) number of customers or units arriving per time period (e.g., per hour).
  - Constant—rate has no variance.
  - Variable—Poisson distribution frequently applied.

- Inter-arrival time: mean (average) time between arrivals.
  - Variable—Exponential distribution frequently applied.
Arrival process distributions

- Two basic distributions
  - Exponential
    - Inter-arrival time: mean (average) time between arrivals
    - Continuous prob. function
  - Poisson
    - Arrival rate: mean (average) number of customers or units arriving per time period (e.g., per hour).
    - Rate has no variance.
    - Discrete probability function
POISSON ARRIVAL PROCESS

REQUIRED CONDITIONS

- Orderliness
  - at most one customer will arrive in any small time interval of $\Delta t$

- Stationarity
  - for time intervals of equal length, the probability of $n$ arrivals in the interval is constant

- Independence
  - the time to the next arrival is independent of when the last arrival occurred
NUMBER OF ARRIVALS IN TIME $t$

- Assume $\lambda$ = the average number of arrivals per a given interval of time (THE ARRIVAL RATE)
- For a Poisson process, the probability of $x$ arrivals in $t$ time periods has the following Poisson distribution:

$$P(x) = P(x \text{ arrivals in time } t) = \frac{(\lambda t)^x e^{-\lambda t}}{x!}, \quad x = 0, 1, 2, \ldots$$

$$\mu = \lambda t \quad \sigma^2 = \lambda t$$
Time Between Arrivals

- The average time between arrivals is \(1/\lambda\)
- For a Poisson process, the time between arrivals (t) has the following exponential distribution:

\[
f(t) = \lambda e^{-\lambda t}, \quad t \geq 0
\]

\[
F(t) = 1 - e^{-\lambda t}
\]

\[
\mu = 1/\lambda
\]

\[
\sigma^2 = 1/\lambda^2
\]
Poisson and Exponential Equivalence

Poisson distribution for number of arrivals per hour (top view)

<table>
<thead>
<tr>
<th>Arrival</th>
<th>Arrivals</th>
<th>Arrivals</th>
<th>Arrival</th>
</tr>
</thead>
<tbody>
<tr>
<td>62 min.</td>
<td>40 min.</td>
<td>123 min.</td>
<td></td>
</tr>
</tbody>
</table>

Exponential distribution of time between arrivals in minutes (bottom view)
Distribution of Arrivals

- Relation of Poisson/exponential distribution
  - Poisson is based on rate, mean is $\lambda$. Say that 15 customers arrive per hour, with the Poisson distribution, $\lambda = 15$ customers/hour.
  - Exponential is based on time. If 15 customers arrive per hour, the mean inter-arrival time is 4 minutes. This can be determined as $1/\lambda = 1/15$ customers/hour = $1/15$ hour x 60 minutes/hour = 4 minutes.
  - Work the other way: inter-arrival time is 10 minutes or 1/6 hour. So $1/\lambda = 1/6$, thus $\lambda = 6$ customers/hour.
Examples Poisson Distribution for Arrival Times

Probability: \[ P(X) = \frac{e^{-\lambda} \lambda^x}{x!} \]

\( \lambda = 2 \) Distribution

\( \lambda = 4 \) Distribution
Distribution of Patient Interarrival Times (the mean times between arrivals is 2.4 minutes, which implies that $\lambda = 0.4167$ arrival per minute)
3. Queue Configuration
Queue Characteristics

- **Length** of queue either *limited* (finite) or *unlimited* (infinite).
- Queue limited when it cannot increase to infinite length due to physical or other restrictions.
- Queue defined as unlimited when size is unrestricted.
- Assume queue lengths are *unlimited*.
- **Queue discipline.** Rule by which customers in line receive service.
- Most systems use queue discipline known as *first-in, first-out* rule (FIFO).
Queue Configurations

Multiple Queue

Take a Number

Single queue

Enter
Number of Lines

- Waiting lines systems can have single or multiple queues.
  - Single queues avoid jockeying behavior & all customers are served on a first-come, first-served fashion (perceived fairness is high)
  - Multiple queues are often used when arriving customers have differing characteristics (e.g.: paying with cash, less than 10 items, etc.) and can be readily segmented
4. Queue Discipline

Queue
Priority Rule

- The priority rule determines which customer to serve next.

- Most service systems use the first-come, first-serve (FCFS) rule. Other priority rules include:
  - Earliest promised due date (EDD)
  - Customer with the shortest expected processing time (SPT)

- **Preemptive discipline**: A rule that allows a customer of higher priority to interrupt the service or another customer.
Customer Selection

Queue discipline

- First come, first served
- Shortest processing time
- Reservations first
- Emergencies first
- Limited needs
- Other
Queue discipline

Static (FCFS rule)
- selection based on status of queue
  - Number of customers waiting
  - Round robin

Dynamic
- Selection based on individual customer attributes
  - Priority
  - Preemptive
  - Processing time of customers (SPT or cpu rule)
5. Service Process
Service Facility Characteristics

- Important to examine two basic properties:
  - (1) Configuration of service facility.
  - (2) Pattern of service times.

- Number of service channels in queuing system is number of servers.

- Single-phase means customer receives service at only one station before leaving system.

- Multiphase implies two or more stops before leaving system.
Service Facility Structure

- Single
  - Single phase
  - Multiphase
- Multichannel
  - Single phase
  - Multiphase
- Mixed
  - Multi-to-single channel
  - Alternate paths, such as:
## Service Facility Arrangements

<table>
<thead>
<tr>
<th>Service facility</th>
<th>Server arrangement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parking lot</td>
<td>Self-serve</td>
</tr>
<tr>
<td>Cafeteria</td>
<td>Servers in series</td>
</tr>
<tr>
<td>Toll booths</td>
<td>Servers in parallel</td>
</tr>
<tr>
<td>Supermarket</td>
<td>Self-serve, first stage; parallel servers, second stage</td>
</tr>
<tr>
<td>Hospital</td>
<td>Many service centers in parallel and series, not all used by each patient</td>
</tr>
</tbody>
</table>
Service Rate

- Capacity of server—how many customers can be serviced per time-unit (e.g., per hour).
  - Constant—rate has no variance.
  - Variable—Poisson distribution frequently applied.

- Service time: mean (average) time per service.
  - Constant—time has no variance.
  - Variable—negative exponential distribution frequently applied.

- Note that the same Poisson/negative exponential relationship shown in arrival rate applies to the service rate.
Service Pattern

Constant

Example: Items coming down an automated assembly line.

Variable

Example: People spending time shopping.

Same classification as arrival process
### The Queuing System

#### Queue Discipline
- First in, first out (FIFO)
- First in, last out (LIFO)
- Various priorities

#### Service Time Distribution
- Constant inter-arrival times
- Random
- Event dependent

#### Number of Lines & Line Structures
- Single Q, single S
- Single Q, multiple S
- Multiple Qs, multiple Ss, w/ Q switching

#### Length

<table>
<thead>
<tr>
<th>Queuing System</th>
<th>Single Q, single S</th>
<th>Single Q, multiple S</th>
<th>Multiple Qs, multiple Ss, w/ Q switching</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Lines &amp; Line Structures</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Queue Discipline</td>
<td>First in, first out (FIFO)</td>
<td>First in, last out (LIFO)</td>
<td>Various priorities</td>
</tr>
<tr>
<td>Service Time Distribution</td>
<td>Constant inter-arrival times</td>
<td>Random</td>
<td>Event dependent</td>
</tr>
</tbody>
</table>
Exit

- May need service again (particularly affects finite populations).
- Low probability of [immediate] re-service.
Common Performance Measures

- The average number of customers waiting in queue
- The average number of customers in the system (multiphase systems)
- The average waiting time in queue
- The average time in the system
- The system utilization rate (% of time servers are busy)
# Thrifty Car Rental

<table>
<thead>
<tr>
<th>Queue Features</th>
<th>Customer Counter</th>
<th>Garage</th>
<th>Car Wash</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calling Population</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arrival Process</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Queue Configuration</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Queue Discipline</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Service Process</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>