Introduction to Queueing Theory

A pioneer:
Agner Krarup Erlang
(1878-1929)

Queueing theory - basics

- Important queuing models with FIFO discipline
  - The M/M/1 model
  - The M/M/c model
  - The M/M/c/K model (limited queuing capacity)
  - The M/M/c/∞/N model (limited calling population)
- Priority-discipline queuing models
- Application of Queueing Theory to system
design and decision making
What is queueing theory?

- Mathematical analysis of queues and waiting times in stochastic systems.
  - Used extensively to analyze production and service processes exhibiting random variability in market demand (arrival times) and service times.
- Queues arise when the short term demand for service exceeds the capacity
  - Most often caused by random variation in service times and the times between customer arrivals.
  - If long term demand for service > capacity the queue will explode!

Why is queueing theory important?

- Capacity problems are very common in industry and one of the main drivers of process redesign
  - Need to balance the cost of increased capacity against the gains of increased productivity and service
- Queuing and waiting time analysis is particularly important in service systems
  - Large costs of waiting and of lost sales due to waiting

Prototype Example – ER at County Hospital
- Patients arrive by ambulance or by their own accord
- One doctor is always on duty
- More and more patients seeks help ⇒ longer waiting times
  - Question: Should another MD position be instated?
A cost/capacity tradeoff model

Components of basic a queueing process

**Input Source**
- Calling Population
- Arrival Process

**The Queuing System**
- Queue
- Service Mechanism
- Queue Discipline
- Service Process

Served Jobs leave the system.
Components of a queueing process

- The calling population
  - The population from which customers/jobs originate
  - The size can be finite or infinite (the latter is most common)
  - Can be homogeneous (only one type of customers/jobs) or heterogeneous (several different kinds of customers/jobs)

- The Arrival Process
  - Determines how, when and where customer/jobs arrive to the system
  - Important characteristic is the customers’/jobs’ inter-arrival times
  - To correctly specify the arrival process requires data collection of interarrival times and statistical analysis.

Components of a queueing process

- The queue configuration
  - Specifies the number of queues
    - Single or multiple lines to a number of service stations
  - Their location
  - Their effect on customer behavior
    - Balking and reneging
  - Their maximum size (# of jobs the queue can hold)
    - Distinction between infinite and finite capacity
## Example – Two Queue Configurations

### Multiple Queues

- Servers
- Multiple lines

### Single Queue

- Servers
- Single line

## Multiple v.s. Single Customer Queue Configuration

<table>
<thead>
<tr>
<th>Multiple Line Advantages</th>
<th>Single Line Advantages</th>
</tr>
</thead>
</table>
| 1. The service provided can be differentiated  
  - Ex. Supermarket express lanes   | 1. Guarantees fairness  
  - FIFO applied to all arrivals |
| 2. Labor specialization possible            | 2. No customer anxiety regarding choice of queue |
| 3. Customer has more flexibility          | 3. Avoids “cutting in” problems |
| 4. Balking behavior may be deterred        | 4. The most efficient set up for minimizing time in the queue |
|   - Several medium-length lines are less intimidating than one very long line          | 5. Jockeying (line switching) is avoided       |
Components of a Basic Queuing Process

- The Service Mechanism
  - Can involve one or several service facilities, with one or several parallel service channels (servers) - Specification is required
  - The service provided by a server is characterized by its service time
    - Specification is required and typically involves data gathering and statistical analysis.
    - Most analytical queuing models are based on the assumption of exponentially distributed service times, with some generalizations.

- The queue discipline
  - Specifies the order by which jobs in the queue are being served.
  - Most commonly used principle is FIFO.
  - Other rules are, for example, LIFO, SPT, EDD...
  - Can entail prioritization based on customer type.

Mitigating Effects of Long Queues

1. Concealing the queue from arriving customers
   - Ex. Restaurants divert people to the bar or use pagers, amusement parks require people to buy tickets outside the park, banks broadcast news on TV at various stations along the queue, casinos snake night club queues through slot machine areas.

2. Use the customer as a resource
   - Ex. Patient filling out medical history form while waiting for physician

3. Making the customer’s wait comfortable and distracting their attention
   - Ex. Complementary drinks at restaurants, computer games, internet stations, food courts, shops, etc. at airports

4. Explain reason for the wait

5. Provide pessimistic estimates of the remaining wait time
   - Wait seems shorter if a time estimate is given.

6. Be fair and open about the queuing disciplines used
Queueing Modeling and System Design

- Two fundamental questions when designing (queuing) systems
  - Which service level should we aim for?
  - How much capacity should we acquire?
- The cost of increased capacity must be balanced against the cost reduction due to shorter waiting time
  - Specify a waiting cost or a shortage cost accruing when customers have to wait for service or...
  - … Specify an acceptable service level and minimize the capacity under this condition
- The shortage or waiting cost rate is situation dependent and often difficult to quantify
  - Should reflect the monetary impact a delay has on the organization where the queuing system resides

Analyzing Design-Cost Tradeoffs

- Given a specified shortage or waiting cost function the analysis is straightforward
- Define
  - WC = Expected Waiting Cost (shortage cost) per time unit
  - SC = Expected Service Cost (capacity cost) per time unit
  - TC = Expected Total system cost per time unit
- The objective is to minimize the total expected system cost

\[
\text{Min } TC = WC + SC
\]