SOS POLITICAL SCIENCE AND PUBLIC ADMINISTRATION

MBA FA 204

SUBJECT NAME: OPERATION RESEARCH

UNIT-IV

TOPIC NAME: QUEUING THEORY

What Is Queuing Theory?

Queuing theory is the mathematical study of the congestion and delays of waiting in line. Queuing theory (or "queueing theory") examines every component of waiting in line to be served, including the arrival process, service process, number of servers, number of system places, and the number of customers—which might be people, data packets, cars, etc.

As a branch of operations research, queuing theory can help users make informed business decisions on how to build efficient and cost-effective workflow systems. Real-life applications of queuing theory cover a wide range of applications, such as how to provide faster customer service, improve traffic flow, efficiently ship orders from a warehouse, and design of telecommunications systems, from data networks to call centers.

How Queuing Theory Works

Queues happen when resources are limited. In fact, queues make economic sense; no queues would equate to costly overcapacity. Queuing theory helps in the design of balanced systems that serve customers quickly and efficiently but do not cost too much to be sustainable. All queuing systems are broken down into the entities queuing for an activity.

At its most elementary level, queuing theory involves the analysis of arrivals at a facility, such as a bank or fast food restaurant, then the service requirements of that facility, e.g., tellers or attendants.

Queuing Theory?

Queuing theory is the study of queues and the random processes that characterize them. It deals with making mathematical sense of real-life scenarios. For example, a mob of people queuing up at a bank or the tasks queuing up on your computer’s back end.

In queuing theory we often want to find out how long wait times or queue lengths are, and we can use models to do this. These models are typically important in business and software applications, and queueing theory is often considered a part of operations research.

About Queuing

Any queuing activity can be summarized as entities (customers in your supermarket queue, or jobs in a computer queue) trying to get through an activity (waiting to be served). Queues happen when we can't all access the activity at the same time: when it is not economically efficient to have enough checkout lines for everyone to go right through as soon as they were ready, or there isn't enough server space to do an unlimited amount of computer tasks at one moment.

In queueing theory a queue does not refer simply to a neat row which is always first come, first served. This is one example of a queue, but not the only kind. A mob trying to rush for the door on Black Friday is considered a queue as well, as is a group of job applicants waiting for interviews who are picked randomly, one by one, to be interviewed.

Types of Queues and Types of Service

First In First Out, or First Come First Served, is fairly common in banking and commerce. It is the type of queue you get when you have people politely lined up, waiting for their turn.

Last In First Out is the opposite scheme; whoever has been waiting for the shortest time is served first. This type of queue management is common in asset management, where assets produced or acquired last are the ones used or disposed of first. For example: the most recent employees are often the ones laid off first.

Priority is where customers are served based on their priority level; these levels could be based on status, task urgency, or some other criteria.

Shortest Job First is when whoever needs the shortest amount of service gets taken care of first

Processor Sharing is when everyone gets served, or half-served, at the same time; service capacity is distributed evenly among everyone waiting.

There may be a single server, where a line of people or items must go through a single bottleneck, or parallel servers, where the same line is served by several servers. Or there may be a tandem queue, where each of multiple servers has their own queue or line.

Balking when a customer decides not to wait for service because the wait time threatens to be too long. Reneging is similar, but when a customer who has waited already decides to leave because they've wasted too much time. Jockeying is when a customer switches between queues in a tandem queue system, trying to orchestrate the shortest wait possible.

Standard Notation for Queueing Theory

To make life easier, there's standard notation for queueing theory that is used across the board. These standard symbols include

- λ: the mean arrival rate.
- μ: the mean service rate.
- n: the number of people in the system.
- A: the arrival process probability distribution.
- B: the service process probability distribution.
- C: the number of servers.
- D: the maximum number of customers allowed in the system at any given time, waiting or being served (without getting bumped).
- E: the maximum number of customers total.

change by Agner Krarup Erlang, a Danish engineer, statistician and, mathematician. His work led to the Erlang theory of efficient networks and the field of telephone network analysis.

Queues are not necessarily a negative aspect of a business, as their absence suggests overcapacity.

Benefits of Queuing Theory

By applying queuing theory, a business can develop more efficient queuing systems, processes, pricing mechanisms, staffing solutions, and arrival management strategies to reduce customer wait times and increase the number of customers that can be served.

Queuing theory as an operations management technique is commonly used to determine and streamline staffing needs, scheduling, and inventory, which helps improve overall customer service. It is often used by Six Sigma practitioners to improve processes.