Queuing System
Introduction

Analytic congestion models

A congestion system is system in which there is a demand for resources for a system, and when the resources become unavailable, those requesting the resources wait for them to become available. The level of congestion in such systems is usually measured by the waiting line, or queue, of resource requests (waiting line or queuing models).

Elements of a queuing system

Arrival process
Distributed of time between arrival of successive customers; Inter-arrival time could generally be exponential distributed or constant.

Exponentially distributed: If an arrival just occurred or if we have been waiting 20 minutes for an arrival, the probability of an arrival the next 5 minutes remains the same. It can be shown that the only continuous random variable that is memoryless is the exponentially distribution. It can also be shown that if the time between arrivals is exponentially distributed with mean, then the distribution of the number of arrivals during time interval is a Poisson distributed with mean $1/\theta$.

Other characteristics of an arrival process include whether the customer population is finite or infinite and whether the mean time between arrivals is constant or changing over time. If the population of customers is finite, the rate of customer arrivals as the number of customers in the system increases.

Service process

The service process is characterized by the distribution of the time to service an arrival and the number of services. Models of congestion system often assume that the service time is exponentially distributed, thereby facilitating the development of analytical models of the congestion system. The number of serviced is greater that the rate at which customers arrive.

Traffic intensity (ρ) = rate at which customers arrive/ rate customers can be served
  = mean service time/ (mean time between arrivals * Number of services)
    = arrival rate/ (service rate * Number of services)

Queuing discipline

The queuing discipline describes the order in which arrivals are serviced. Common queue disciplines include FIFO, shorter service time first, and random selection for the nest service. The queuing discipline also includes characteristics of the system such as maximum queue-

length) when the queue reaches this maximum, arrivals turn away or balk) and customer reneging (customer waiting in line become impatient and leave the system before service)

Classification of queuing system models

A/Bs/K/E   where A: specifies the arrival process
                        B: specifics the services process
                        s: specific's the number of servers
                        K: maximum number of customers allowed into the system
                        E: queue discipline

Symbols

M: exponentially-distributed service or arrival times
D: constant service or arrival times

Measures of performance of a congestion systems:

Ls: Expected number of customers in the system
Lq: Expected number of customers in the queue
Ws: Expected time a customer is in the system, including the time for service
Wq: Expected time a customer waits for service

M/M/1 queuing model

Queuing models are most easily developed when the arrival times and service times are exponentially distributed. Although the assumption of exponentially-distributed arrival and service times may seem unrealistic, this group of models has wide application and can also serve as a useful first pass in the analysis of more complex congestion systems.

Relationship between queue length, Lq and traffic intensity, ρ

As the traffic intensity gets closer to 1.0, there is a very rapid increase in the measures of congestion, Lq (expected number of customers in the queue) and Wq (expected time a customer in the system) gets closer to 1.0

Non-Markovian queuing models

When the time between customer arrival and the service times are not exponentially distributed

An integral part of simulation modeling is the identification of any analytic models that may serve as simple, initial model. The analytical model then can be used to give the analyst insights and estimates of the more complex system's behavior.