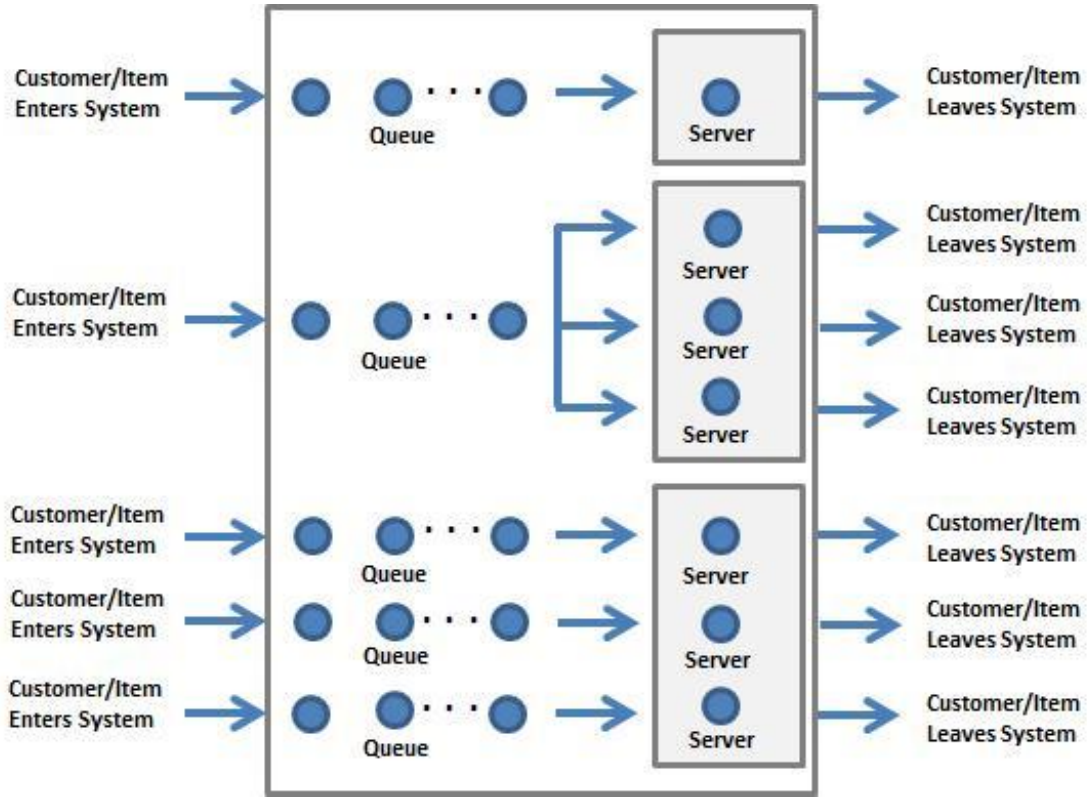
Queueing Optimization

Queueing optimization is an important science that has made many significant contributions to operations and marketing management. Queueing models have broad, practical application in numerous diverse fields. The list of potential applications is almost endless, but here are just a few examples:

- Customer service (e.g., tech support help desks or telemarketing call centers)
- Transportation
- Inventory control
- Retailing (e.g., movie theaters, banks, gas stations, supermarkets, post offices, etc.)
- Machine servicing
- Salesforce management
- Traffic flow engineering and management (e.g., traffic lights, toll booths, etc.)
- Computer networks/data communication
- Medical emergency rooms
- Production systems
- Economics

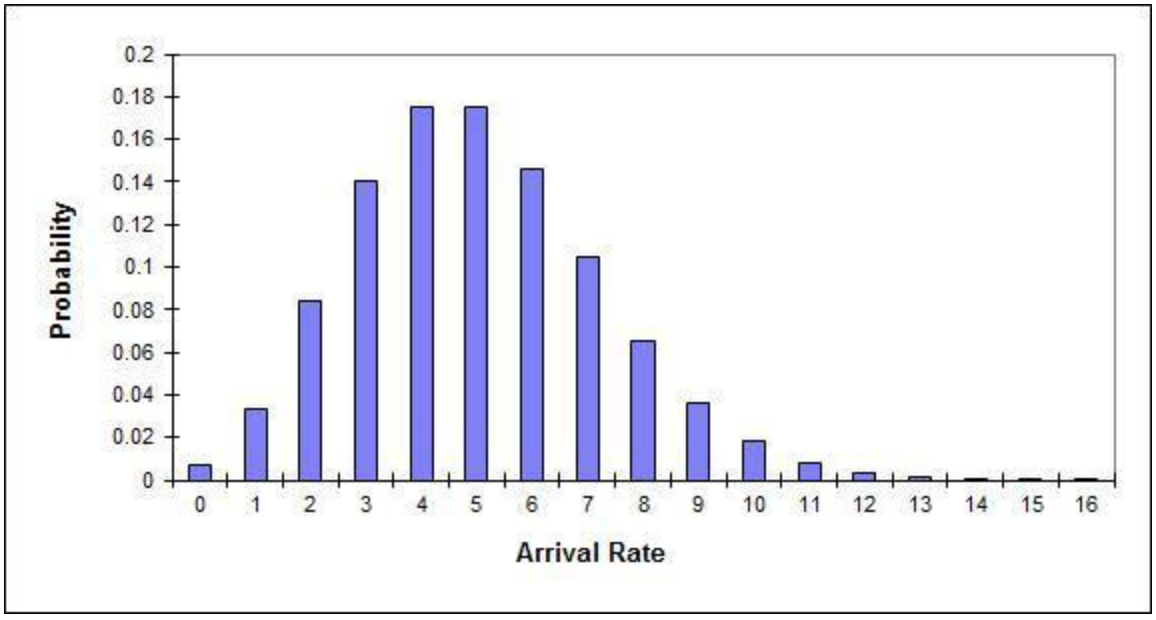
Here we present a basic introduction that will focus on operations optimization. We begin with a simple diagram that shows three typical queueing system configurations:



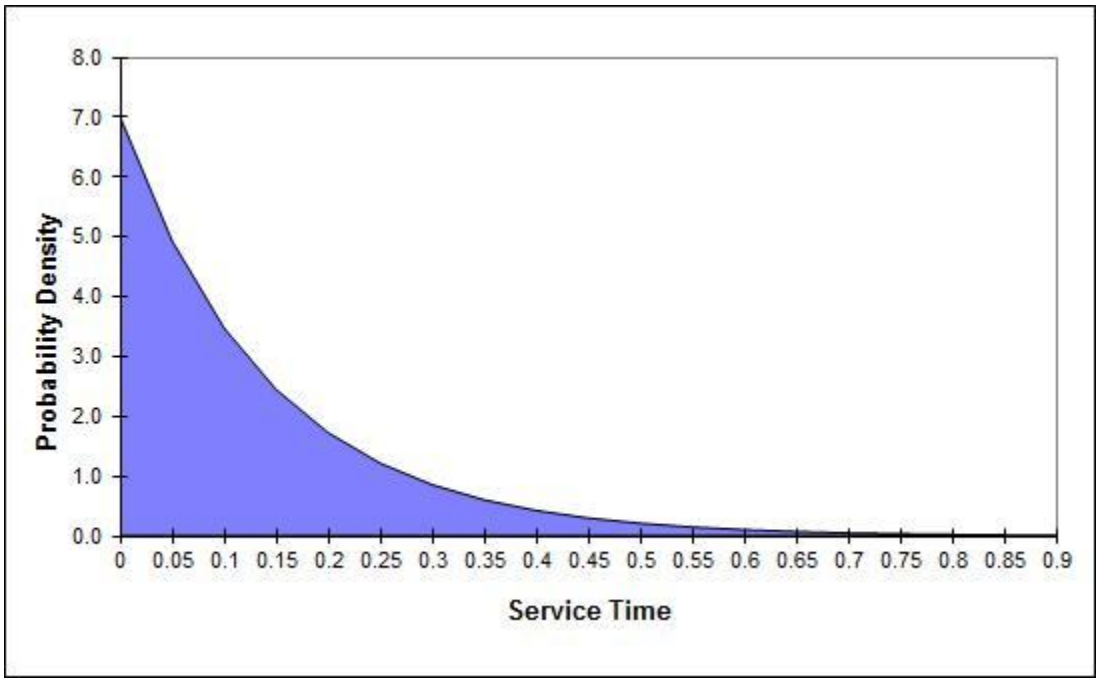
The first system above is a single-queue, single-server system typical of ATMs; the second is a single-queue, multi-server system found at most banks, post offices and airport check-in counters; and the third is a collection of single-queue, single-server systems found at some fast-food restaurants, supermarkets, etc. Each of these systems is a first-in, first-out (FIFO) system, where the first customer or object to enter the system waits until a server is available, at which time the customer or object is the first to be served, after which they leave the system.

In most of these queueing systems, customers or objects arrive at the system in a random pattern or frequency distribution. (The exception would be production systems in which objects tend to arrive at and move through the system in a nonrandom or deterministic fashion.) There are three key variables of interest to us when analyzing queueing systems: arrival rate, service rate, and length or capacity of the queue in which customers or objects can wait for service or processing. Total time spent in the system is the sum of waiting time plus service time.

Random arrival rates (e.g., number of arrivals per hour) typically follow a Poisson distribution, where the horizontal axis represents number of arrivals per time period and the vertical axis represents the probability with which a particular arrival rate occurs:



A given arrival rate will necessarily imply a particular probability distribution for **interarrival times** (the time periods elapsing between arrivals). If an arrival rate follows a Poisson probability distribution, then the resulting interarrival times will follow an exponential distribution. Service times may also follow an exponential distribution:



As the graphs above indicate, many queueing models assume that most customers or objects that pass through the system arrive at relatively modest time intervals, and fewer have relatively long interarrival times; and that most customers or objects are served with relatively modest service times, but a few require relatively long service times. In addition, queue length can be assumed to be either finite or infinite. When dealing with customers, a longer queue space will allow more customers to wait in the queue; but if the queue gets too full, then

customers will begin to "balk" or fail to enter the system at all. Such an outcome is undesirable, and queueing models can be used to attempt to find suitable parameter values (i.e., queue lengths, service times and numbers of servers) that minimize balking.

Because there are many possible queueing models, a notational system, Kendall Notation, has been developed to describe a given model. Using this notation system, queueing models can be described with three parameters using the general format of
1/2/3. (There is actually a more complicated version of the Kendall Notation system that allows for six parameters instead of three, but that is beyond the scope of this discussion.)

The first parameter represents **interarrival times**, and is abbreviated as **M** for Markovian interarrival times (following an exponential random distribution); or as **G** for Generally distributed interarrival times (non-exponential random); or as **D** for Deterministic (non-random) interarrival times. The second parameter represents **service times**, and can be similarly abbreviated as either **M** for Markovian; or as **G** for General; or as **D** for Deterministic. The third parameter, **S**, represents the number of servers available in the system.

For example, a notation of M/M/1 represents a queueing model in which interarrival times and service times both follow an exponential distribution, and there is one server in the system. A notation of M/G/2 represents a queueing model in which interarrival times are distributed exponentially, service times follow a general (nonexponential) distribution, and there are two servers. And so forth.

This link takes you to an example of a simple M/M/S queueing model in which we can try to adjust the queue length and number of servers (S) to handle shorter interarrival times that tend to occur at heavier customer traffic times.

This link leads to an example of a G/G/S queueing model that helps a supermarket reduce the amount of time customers have to spend waiting in check-out lines.

This link goes to an example of a M/G/S queueing model showing how a fast-food restaurant can improve service and its bottom line by installing a self-service soft-drink dispenser.

Here is a M/D/S queueing model example showing how a partially automated car wash can benefit from replacing post-washing cloth drying with automated blow drying.