An Overview of Queueing Theory

A queue is any system that can be thought of as a sequence of customers arriving at a service installation and receiving service. A great variety of real scenarios fit this broad description, including telephone call centers, fast food restaurants, court dockets, medical practice schedules, and maintenance operations in factories. Queueing theory seeks to tie the observable properties of a system to the properties that characterize the process that generate new customers and the process that describes the service completion times. As such, it does not necessarily involve a decision, as is the case for most topics in operations research. However, as a practical matter one would likely initiate a queueing theory study only with the intention of using the results to inform a decision. We'll see that queueing-based decisions require models that consist of a pure queueing theory component (Sections 1–5) along with a component that associates the emergent properties of the system with a cost or value (Section 6).

Structure of Queueing Models

In general, there are a number of features that must be identified in order to specify a queueing system model. The principal unknowns in the model are determined by some analytical method or simulation, and then these principal unknowns are used to calculate the important emergent properties of the queueing system.

Model Specification

The state of a queueing system at any time is the number of customers in the system, including customers being served as well as customers waiting for service. Any particular queueing system is specified by assumptions about the process by which customers arrive, the service process, and any special restrictions on the queue itself.

- Customers arrive according to some random process characterized by a probability distribution of arrival times. The arrival time distribution is almost always taken to be exponential, which is generally a good assumption. The primary quantitative property is the mean arrival rate λ, measured in customers per convenient time unit. Usually the population of potential customers is taken to be infinite, although in some cases it is necessary to indicate a finite calling population size.

- The service station houses one or more servers, which are generally considered to be identical. The service process is characterized by a distribution of service times with mean μ for each server. One should always do an observational study to obtain an empirical service time distribution before deciding what distribution to use in the model. Often the service process is taken for mathematical convenience to be exponential, but that is generally not a good assumption, and one must always consider whether the simplicity of exponential distribution models justifies the potential error caused by using them in place of a more realistic choice.

- Usually the queue itself is assumed to be unlimited, although in practice it is likely that customers will balk (leave the system without receiving service) if the line is too large or the wait is too long. It is assumed that servers are busy whenever there are enough customers, so that the number of customers being served at any time is the smaller of the number of customers in the system (n) and the number of servers (s).

There is a standard descriptive system for queueing systems. Systems with an infinite calling population and no limit on queue length are classified by a system of the form "$A/S/s$," where A indicates the type of distribution for arrival times, S indicates the type of distribution for service times, and s is the number of identical servers in the system. Common choices for A and S are

- M - Markovian (exponential)
- D - degenerate (constant)
- E_k - Erlang (a generalization of the exponential distribution)
- G - general (no specific distribution type)

The arrival and service distributions always have a specified mean rate (λ and μ). Erlang distributions require an additional shape factor k and formulas that apply to general distributions require a standard deviation σ.

Model Equations

Since arrivals are unscheduled, the state of a queueing system changes in time. While it is *not* possible to predict the state of the system at some future time, it *is* possible to predict some aggregate properties of the system, such as the average state over time. The principle unknowns in a queueing model are the probabilities $P_n(t)$ that the system is in state n at time t. For the special case where the arrival and service distribution times are exponential, we can write down a set of differential equations that govern the changes in the probabilities P_n. Most queueing studies only consider the steady-state case, in which the probabilities are constants P_n. The steady-state versions of the differential equations allow for each P_n to be determined in terms of the previous one, so that ultimately they are all given as multiples of P_0. The correct value of P_0 then follows from the requirement that the sum of the probabilities is 1. When the probability distributions are not both exponential, then the probabilities can only be determined by simulation.

Emergent System Properties

Usually we are not actually interested in the probabilities P_n. Instead, we use them to determine emergent system properties of general interest. Chief among these are L, the average number of customers in the system, L_q, the average number of customers waiting to be served, W, the average amount of time customers spend in the system, and W_q, the average amount of time customers spend waiting for service to begin. Sometimes the underlying probability distribution of waiting times is also important.

Analysis of Queueing Models

The method of analysis for queueing models depends on the nature of the model.

1. When both the arrival times and the service times are exponentially distributed, the steady-state system properties can usually be determined analytically (Sections 3–4). Unfortunately, service times are generally not exponentially distributed; however, results for this case are often a good approximation of a more realistic case.
2. Some analytical results are available for arbitrary service time distributions, provided there is only one server. (Section 2)

3. In general, steady-state properties must be determined by simulation.

4. Transient properties always need to be determined by simulation.

Using Queueing Theory to Make Decisions

A decision problem (of any kind) has four components:

1. A set of **decision variables**, which can be chosen from some continuous or discrete set of options;

2. A set of **parameters**, which have values that cannot be controlled;

3. An **objective function**, which defines the quantity to be maximized (value) or minimized (cost);

4. A mathematical model that defines the quantities needed to compute the objective function in terms of the decision variables and parameters.

In queueing theory, it is usually possible to make choices in the number of servers s, and it may also be possible to make choices that modify the mean arrival rate λ and the mean single-server service rate μ, possibly in terms of other parameters. The objective function is usually a total cost that includes both the direct cost of operating the system and indirect costs, such as the cost of losing customers to faster competitors. Increasing s or μ or decreasing λ increase the direct cost while decreasing the indirect cost. Usually service increases become progressively more expensive while yielding progressively less decrease in indirect cost; hence, there will be an optimal strategy that minimizes total cost.

The required mathematical model has two components. First, an appropriate queueing model determines the key output quantities, such as the distribution of waiting times and the expected queue size. Second, a cost model associates these output characteristics with the indirect costs and prescribes the direct costs in terms of the decision varibles. Because the number of decision variables is generally small, there is usually no difficulty in determining the optimal solution once the results of the queueing model are known. In general, an optimal strategy is only as good as the quality of the model from which it is obtained, so practitioners of queueing theory need to be careful in modeling the objective function as well as the queueing system.