## Introduction to Queueing Theory

Queueing theory is a set of mathematical tools for the analysis of probabilistic systems of customers and servers. Among the oldest of Management Science tools, queueing theory can be traced to the work of A. K. Erlang, a Danish mathematician who studied telephone traffic congestion in the first decade of the 20th century.

## Definitions:

Arrival Rate refers to the average number of customers who require service within a specific period of time.

A Capacitated Queue is limited as to the number of customers who are allowed to wait in line.

Customers can be people, work-in-process inventory, raw materials, incoming digital messages, or any other entities that can be modeled as lining up to wait for some process to take place.

A Queue is a set of customers waiting for service.
Queue Discipline refers to the priority system by which the next customer to receive service is selected from a set of waiting customers. One common queue discipline is first-in-first-out, or FIFO.

A Server can be a human worker, a machine, or any other entity that can be modeled as executing some process for waiting customers.

Service Rate (or Service Capacity) refers to the overall average number of customers a system can handle in a given time period.

Stochastic Processes are systems of events in which the times between events are random variables. In queueing models, the patterns of customer arrivals and service are modeled as stochastic processes based on probability distributions.

Utilization refers to the proportion of time that a server (or system of servers) is busy handling customers.

## Queueing Notation:

In the literature, queueing models are described by a series of symbols and slashes, such as $A / B / X / Y / Z$, where $A$ indicates the arrival pattern, $B$ indicates the service pattern, $X$ indicates the number of parallel servers, $Y$ indicates the
queue's capacity, and Z indicates the queue discipline. We will be concerned primarily with the $\mathrm{M} / \mathrm{M} / 1$ queue, in which the letter M indicates that times between arrivals and times between services both can be modeled as being exponentially distributed. The number 1 indicates that there is one server; we will also study some $\mathrm{M} / \mathrm{M} / \mathrm{n}$ queues, where n is some number greater than 1 .

## Symbols:

| Performance Measure | Random Variable | Expected Value |
| :--- | :---: | :---: |
| Number of Customers in the System | N | L |
| Number of Customers in the Queue | $\mathrm{N}_{\mathrm{q}}$ | $\mathrm{L}_{\mathrm{q}}$ |
| Number of Customers in Service | $\mathrm{N}_{\mathrm{s}}$ | $\mathrm{L}_{\mathrm{s}}$ |
| Time Spent in the System | T | W |
| Time Spent in the Queue | $\mathrm{T}_{\mathrm{q}}$ | $\mathrm{W}_{\mathrm{q}}$ |
| Time Spent in Service | $\mathrm{T}_{\mathrm{s}}$ | $\mathrm{W}_{\mathrm{s}}$ |


| System Parameters |  |  |
| :--- | :--- | :--- |
| Number of Servers | $S$ |  |
| Arrival Rate (number per unit of time) | $\lambda$ | (Greek letter lambda) |
| Service Rate (number per unit of time) | $\mu$ | (Greek letter mu) |
| Utilization Factor | $\rho$ | (Greek letter rho) |

## Formulas:

The utilization factor, or $\rho$ (rho), also the probability that a server will be busy at any point in time:

$$
\begin{equation*}
\rho=\frac{\lambda}{S \mu} \tag{i}
\end{equation*}
$$

Idle time, or the proportion of time servers are not busy, or the probability that a server will be idle at any given time:

$$
\begin{equation*}
1-\rho \tag{ii}
\end{equation*}
$$

The average time a customer spends in the system:

$$
\begin{equation*}
W=W_{q}+W_{s}=\frac{1}{\mu-\lambda} \tag{iii}
\end{equation*}
$$

The average number of customers in the queue:

$$
\begin{equation*}
L_{q}=\frac{\lambda^{2}}{\mu(\mu-\lambda)}=\frac{\rho^{2}}{1-\rho} \tag{iv}
\end{equation*}
$$

The average number of customers in the system:

$$
\begin{equation*}
L=L_{q}+L_{s} \tag{v}
\end{equation*}
$$

The average time spent waiting in the queue:

$$
\begin{equation*}
W_{q}=\frac{\lambda}{\mu(\mu-\lambda)}=\frac{\rho}{\mu(1-\rho)} \tag{vi}
\end{equation*}
$$

Those who would commune more directly with the Arab mathematician al Jebr (whose name is the source for our word algebra) will note that there is another formula for N :

$$
N_{q}+N_{s}=\frac{\rho^{2}}{1-\rho}+\rho=\frac{\frac{\lambda}{\mu}\left(\frac{\lambda}{\mu}\right)}{1-\frac{\lambda}{\mu}}+\frac{\lambda}{\mu}=\frac{\left\{\mu\left[\frac{\lambda}{\mu}\left(\frac{\lambda}{\mu}\right)\right]\right\}+\left\{\left(1-\frac{\lambda}{\mu}\right)(\lambda)\right\}}{\mu\left(1-\frac{\lambda}{\mu}\right)}=\frac{\frac{\lambda}{\mu}}{1-\frac{\lambda}{\mu}}=\frac{\rho}{1-\rho}=N
$$

Therefore:

$$
\begin{equation*}
L=\frac{\rho}{1-\rho} \tag{vii}
\end{equation*}
$$

The single most important formula in queueing theory is called Little's Law:

$$
\begin{equation*}
L=\lambda W \tag{viii}
\end{equation*}
$$

Little's Law can be restated in several ways; for example, here is a formula for W :

$$
\begin{equation*}
W=\frac{L}{\lambda} \tag{ix}
\end{equation*}
$$

There are also some formulas that are based on the special characteristics of the exponential distribution. These generally are beyond the scope of this course, but one useful example is offered here. In an $M / M / n$ system, the probability of a customer spending longer than some specific time period $t$ in the system is given by the following formula. $e$ is the base of the natural logarithms (roughly 2.718):

$$
\begin{equation*}
P\{T \geq t\}=e^{(\lambda-\mu) t} \tag{x}
\end{equation*}
$$

## What Every Manager Should Know about Queueing Theory

1. Little's Law describes the fundamental trade-off between customer service and server utilization. To reduce the amount of time a customer spends waiting in your system, you must increase the amount of time your server is idle. Consider the following graph, which uses Little's Law to illustrate the relationship between rho and the expected number of people in the system:

2. Great economies of scale with respect to waiting times can be achieved by joining multiple queues into a single queue with multiple servers. For example, supermarket customers spend less time waiting in a single queue served by 4 cashiers than in one of four separate queues, each served by one cashier.

This seeming paradox stems from the fact that, with four independent queues, it is possible for one cashier to be idle while the other three have long lines of people waiting. Although supermarkets are set up with separate queues for each cashier, the customers intuitively understand and make adjustments; they won't stand in line for one cashier when another cashier is idle.
3. Lambda (the arrival rate of customers) is rarely stationary. For example, customers arrive at a hospital emergency room more frequently from 12:00 to 1:00 AM than from 9:00 to 10:00 AM. A common error is to use a long-run average to calculate lambda, without accounting for changes in the arrival rate over time. Depending on the nature of the operation being planned, it may be best to use a "worst-case" estimate of lambda, based on the arrival rate during peak hours.
4. A simple rule of thumb for deciding how many servers to provide is given by Peter Kolesar. Let us suppose that you want only a $5 \%$ chance that any given customer will have to wait in line; call this probability of waiting $\alpha$ (alpha). Using the normal table, we can estimate the number of servers needed to provide this level of service using the following formula:

$$
s \approx \frac{\lambda}{\mu}+z_{1-\alpha} \sqrt{\frac{\lambda}{\mu}}
$$

If, for example, we have a system in which we expect 75 customers per hour, and in which each server can handle about 10 customers per hour, we can achieve the $5 \%$ probability of waiting by providing 12 servers:

$$
\begin{aligned}
& s \approx \frac{75}{10}+z_{1-.05} \sqrt{\frac{75}{10}} \\
& =7.5+z_{1-.05} \sqrt{7.5} \\
& =7.5+1.64(2.74) \\
& =11.99
\end{aligned}
$$

