A Practical Approach to General Queuing Systems

QUEUING THEORY 1.0

Lavi Industries

Lavi Industries
27810 Avenue Hopkins
Valencia, CA 91355-3409
www.lavi.com

QUEUING THEORY 1.0

A Practical Approach
to General Queuing Systems

Table of Contents

Chapter **1**

What is a Queuing System?

1.0 A Definition....1
1.1 Arrival and Service Patterns....3
1.2 Number of Parallel Servers....5
1.3 System Capacity....5
1.4 Queue Discipline....6
1.5 Common Queuing Systems....7

Chapter **2**

Why Study Queuing Systems?....7

Chapter **3**

Relationships Between Performance Measures in General Queuing System...10

Appendix **A**

Definitions....17

Appendix **B**

References....18

Chapter 1

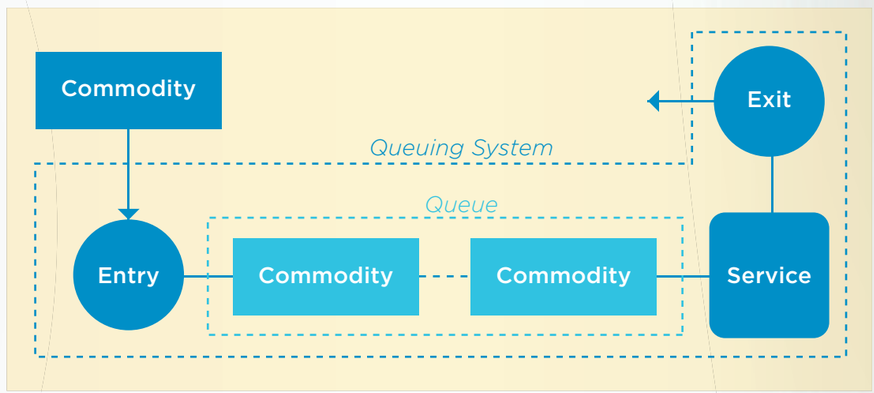
What is a Queuing System?

1.0 A Definition

A queuing system belongs to a broad range of dynamic systems that are, in turn, generally defined as flow systems in which a commodity moves through one or more channels in order to go from one point to another.[1] A queuing system consists of entry points, waiting lines, service channels, and exit points. A *commodity* enters the system at the entry point and, if no service channel is immediately available, is moved into a waiting line, also known as a *queue*. Once a service channel is available, the commodity moves from the waiting line into the service channel, and having received service, leaves the system.

Fig. 1.0

Example of a queuing system.



A commodity could be a car waiting in line to pay a toll for a bridge crossing, a part to a product in an assembly line, or a customer waiting in a checkout line. Queuing systems are applicable to a wide variety of problems in many environments such as computer network traffic flow, call flow in a call center, or patient scheduling in a medical setting. The scope of this book deals with queuing systems that describe customer flow in facilities such as retail stores, banks, hospitals, and other venues.

Real problems encountered in the world of business and industry rarely conform to the exact mathematical models.

The study of queuing systems is known as Queuing Theory. Pioneering work in understanding queues is attributed to Danish mathematician A.K. Erlang who published "The Theory of Probabilities and Telephone Conversations" in 1909. Erlang worked on a practical problem of describing the behavior of randomly rising demand in telephone traffic congestion. The random, or stochastic, nature of the queuing process was a topic of interest of E.C. Molina, Felix Pollaczek, Andrey Kolmogorov, Khintchine, Crommelin, Palm, and many others.[2] The work of the past century has provided a solid mathematical basis for exact solutions of well-formulated queuing problems; however, real problems encountered in the world of business and industry rarely conform to the exact mathematical models, and the solutions must not only draw on the mathematical scaffolding but employ simulations and computational analysis. With the advent of powerful computers, complex real-life queuing systems can now be described and optimized utilizing the mathematical basis.

To describe a particular queuing system, several key characteristics are needed: arrival pattern of the customers, service pattern, number of parallel servers, system capacity, queue discipline, and the number of service stages. Over the years of work in queuing theory, a shorthand notation, largely contributed to David Kendall, was developed:

Fig. 1.1

Shorthand notation often used in queuing theory.

A/B/X/Y/Z

A: Arrival pattern
B: Service pattern
X: Number of parallel servers
Y: System capacity
Z: Queue discipline

1.1 Arrival and Service Patterns

An arrival pattern is described as random, or stochastic, if a commodity arrival is independent of the number of commodities currently in the system, if the arrival of a commodity is independent of the previous commodity, or if the time between commodity arrivals is irregular. The number of commodities currently in a system is known as the "state of the system."

In the case of customers entering a waiting line to receive a service at a bank or store, each customer entry is an individual, or *discrete,* event. Some of the customers arrive shortly after the customer before them, while others arrive after a longer interval; therefore, the time between the arrivals is random. Customer arrivals can be counted over an hour, a day, or some other measurable period.

Due to the nature of the customer arrival pattern (independent discrete events occurring randomly over a fixed period of time with a known average rate), the process is well described as a particular type of discrete probability distribution called a *Poisson distribution*. Instances of customers receiving service are also independent discrete random events occurring over a period of time with a measurable average rate and, therefore, can also be characterized as a Poisson process.

The time between events, meaning inter-arrival time, and the time between service instances, is described by a probability distribution called exponential distribution. Exponential distribution describes the time between the events in a Poisson process.

Besides the exponential distribution, the inter-arrival times and service times can be described as non-random, deterministic, Erlang-type-k distributed, or general. The nomenclature used in Kendall's notation (first characteristic, A) is as follows:

Fig. 1.2

Nomenclature used for Kendall's first characteristic, A.

M: Exponentially distributed	*PH: Phase type*
Ek: Erlang-type-k distributed	*Hk: Mixture of k-exponential*
D: Deterministic	*G: General*

Upon arrival, a customer may decide that the queue is too long and leave, or *balk*. A customer may enter the queue, but become impatient and leave, behavior known as *reneging*. Or, as it often happens in supermarkets, a customer may move, or *jockey*, from a seemingly longer queue to a shorter one or the one with customers whose shopping carts are lightly loaded. Customers may become frustrated just because of the size of the queue and ultimately effect the arrival and service patterns.[2] Servers may change the pattern by speeding up the rate of service if the facility is busy, or become tired or overwhelmed and slow down. Thus, modeling a system mathematically is a good start to defining a real-life system, but computational tools may provide a more accurate description.

1.2 Number of Parallel Servers

For analytical purposes, the "quality of service" of a queuing system is determined by the number of servers. By a simple extrapolation, more servers yield quicker throughput of the system. Alternatively, servers incur a cost, and during slower periods may become idle, decreasing the overall system efficiency. Notably, a cost can be placed on the customer's waiting time as well.

1.3 System Capacity

In a typical business environment at a bank or retail store, the system capacity may be assumed to be infinite. However, there may be a case of a queuing system that places a limit on how many commodities may be allowed into the system, such as physical limitations of a facility or fire code regulations.

System capacity is noted by the integer number of allowed commodities. If it is not specified, it is assumed to be infinite.

1.4 Queue Discipline

Depending on the design of the queuing system, the facility, or the inherent behavior of the customers, the manner in which a commodity is serviced varies. The most common discipline found in a typical business is first in, first out (FIFO), also referred to as first come, first served (FCFS). For a warehouse inventory, last come, first served (LCFS) may be a preferred discipline. In a hospital queuing system, priority ordering (PRI) of patients based on critical need may be the desired choice. General ordering is noted by GD, and random selection by RSS. If the queue discipline is not specified, it is assumed to be FIFO.

FIFO: *First In, First Out*

FCFS: *First Come, First Served*

LCFS: *Last Come, First Served*

PRI: *Priority Ordering*

GD: *General Ordering*

RSS: *Random Selection*

Examples:

M/D/1: exponentially distributed inter-arrival times / deterministic service times / single service channel / infinite capacity / first in, first out

D/Ek/3/100/LCFS: deterministic inter-arrival times / Erlang-type-k distributed service times / 3 service channels / 100 units capacity / last come, first served

1.5 Common Queuing Systems

Two mathematical models which describe the most common queuing in business settings (bank, retail store, etc.) are M/M/1 and M/M/c. In both systems, the inter-arrival times and service times are exponentially distributed. The systems differ by the number of service channels, *c*.

Chapter 2

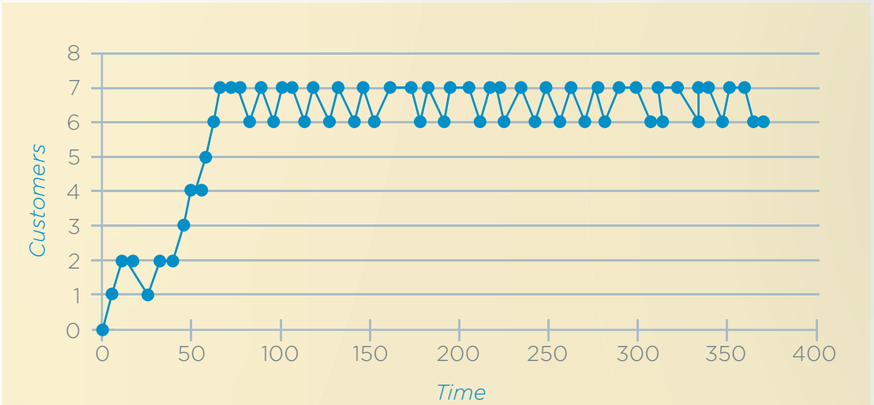
Why Study Queuing Systems?

In all queuing systems, we look for a way to accurately describe how the commodities accumulate and what the waiting times may be, and to determine how many serving channels have to be provided to adequately handle the inflow. In the case of a typical customer flow queuing system, improving the customer experience while minimizing the system's cost for servicing the inflow is the objective. The objective is met by studying the customer's waiting time before receiving service, the length of the queue, and the number of customers in the entire system, and then providing an adequate number of service channels to ensure that the servers remain reasonably busy. To be effective, the system needs to maintain a certain level of customer throughput to avoid congestion and to minimize customer's discomfort.

As an example, take a retail store opening in the morning. Prior to the store opening, there are zero customers in line. Once the store has been opened, a single customer approaches the checkout. Since no customers are ahead of him, he immediately is able to pay for the purchase. While the first customer is paying for the purchase, the second customer arrives into the queue and, if no other checkout counter is available, waits for the first transaction to complete. As time passes, more customers arrive and the queue begins to grow. The values of the system may change dramatically in a short period of time. To provide meaningful measurements, we consider the values of the system in the *steady state*, when the system has been in use for some time and customer flow, as well as the service rate, have stabilized.

Fig. 2.0

Queuing system approaching steady state.



Why Study Queuing Systems (cont.)

Typical steady state performance measures for a general queuing system:

- Number of customers in the system: also known as the state of the system, the average combined number of customers in queue and in service.
- Queue length: number of customers waiting for service.
- Typical waiting time: average time a customer has to wait before receiving service.
- Server idle time: average time a server spends between helping customers.
- Server utilization: ratio of the average time spent helping customers to the duration of time in question. This is also known as traffic intensity, offered work load, and the busy probability for an arbitrary server.

Chapter 3

Relationships Between Performance Measures in General Queuing Systems

The relationships between system performance measures in a general queuing system in steady state were developed by John D.C. Little. He explained how the average number of commodities in a steady state system determine wait time based on the arrival rate.

Fig. 3.0

Little's Formula

$$L = \lambda W$$

L: Mean number of commodities in the system

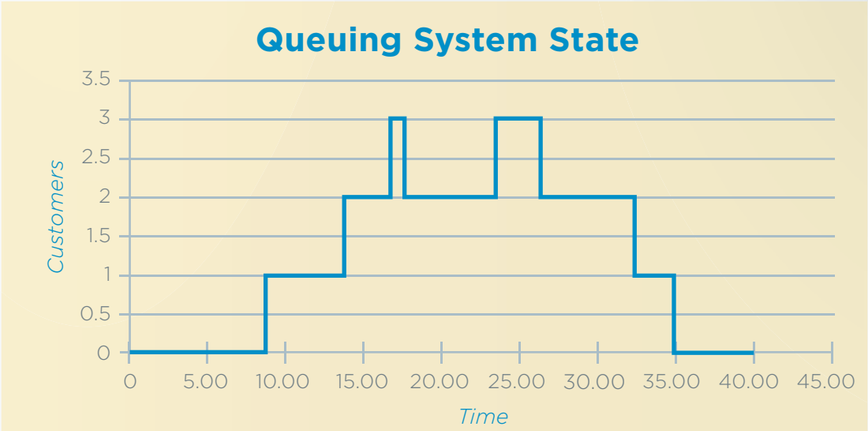
λ: Arrival rate of the commodities (i.e. 6 customers per hour)

W: Average time a commodity spends in the system

For an intuitive understanding of this concept, let's consider the following queuing system as its state changes over time[2]:

Fig. 3.1

Number of customers in a queuing system over a period of time.



The number of customers in the system, or system state, increases by one every time a new customer enters the queue and decreases by one when a customer receives service and leaves. From fig. 3.1, the first customer arrives at some time, *t1*, changing the system state to 1. The second customer arrives at *t2*, increasing the system state to 2. Customer 3 shortly follows at *t3*, increasing the system state to 3. At *t4*, customer 1 leaves, reducing the system state to 2. At *t5*, customer 4 arrives bringing the system state back to 3. At *t6*, customer 2 leaves (system state is 2), and then at *t7*, customer 3 leaves (system state is 1). Finally, customer 4 leaves the system at T restoring the state to 0.

Table 1

From	To	System State
t1	t2	1
t2	t3	2
t3	t4	3
t4	t5	2
t5	t6	3
t6	t7	2
t7	T	1

Table 2

Customer	Entry	Exit
1	t1	t4
2	t2	t6
3	t3	t7
4	t5	T

Think of the curve in fig. 3.1 as a function *f(t)* with multiple values *f(t1)....f(tn)*. To determine the average value of *f(t)*, we have to add all values of *f(t)* and divide that sum by n:

Equation 1

$$L = E[n] = \frac{f(t1) + f(t2) + \cdots + f(tn)}{n}$$

To ensure that the duration of the value is considered, *n* is a number such that duration of each value *f(x1)...f(xn)* is the same. In our case, we can use the time from *t3* to *t4*, or $\Delta t = t4\text{-}t3$, as it is the shortest period with the same value. Therefore, *n* is the number of Δt intervals in *T*.

$$n = \frac{T}{\Delta t}$$

Substituting $\frac{T}{\Delta t}$ for *n* in eq. 1 yields:

$$L = E[n] = \frac{\Delta t(f(t1) + f(t2) + \cdots + f(tn))}{T} = \frac{1}{T}\Sigma_{i=1}^{n} f(t)\Delta t$$

Note that expression $\frac{1}{T}\Sigma_{i=1}^{n} f(t)\Delta t$ is a *Reimann Sum.*

Thus, from the graph in fig. 3, the average number of customers in the system over period *T*:

$$L = \frac{1(t2-t1) + 2(t3-t2) + 3(t4-t3) + 2(t5-t4) + 3(t6-t5) + 2(t7-t6) + 1(T-t7)}{T}$$

Equation 2

$$L = \frac{\textit{area under the curve}}{T} = \frac{-t1 - t2 - t3 + t4 - t5 + t6 + t7 + T}{T}$$

To determine the average time spent in the system, *W*, we add the individual time for each customer and divide the sum by the total number of customers, N_C. From the fig. 3 and table 2, the average time spent in the system, W:

$$W = \frac{(t4-t1)+(t6-t2)+(t7-t3)+(T-t5)}{N_c} = \frac{-t1-t2-t3+t4-t5+t6+t7+T}{N_c}$$

Equation 3

$$W = \frac{\textit{area under the curve}}{N_c}$$

Solving for the area under the curve in fig. 3 in eq. 2 and eq. 3:

$$\textit{area under the curve} = LT$$

$$\textit{area under the curve} = W\ N_c$$

Equation 4

$$\text{Then, } LT = W\ N_c \text{ or } L = \frac{W\ N_c}{T}$$

By definition, the arrival rate of customers, λ, is the total number of customers, N_C, over a period of time, *T*:

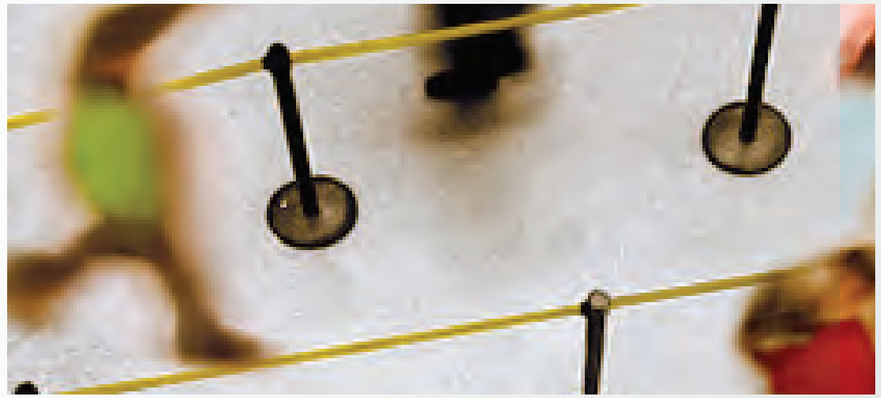
Equation 5

$$\lambda = \frac{N_c}{T}$$

Substituting, λ for $\frac{N_c}{T}$ in eq. 4, we have Little's formula:

$$L = \lambda W$$

Note: For a complete proof, please see Little, 1961.



Typical Steady State Performance Measures Formula:

• **Number of customers in the system:** also known as the state of the system. The average combined number of customers in queue and in service.

$$L = \lambda W$$

• **Queue length:** number of customers waiting for service.

$L_q = \lambda W_q$ where W_q is the time spent in the queue

• **Typical waiting time:** average time a customer has to wait before receiving service.

$$W = W_q + \frac{1}{\mu},$$

μ is the service rate - customers served per unit of time.

• **Server utilization:** ratio of the average time spent helping customers to the duration of time in question. This is also known as traffic intensity, offered work load, and busy probability for an arbitrary server:

$$P_b = \frac{\textit{time server is busy}}{\textit{total time}} =$$

$$= \frac{\textit{arrival rate} \;\; x \;\; \textit{total time} \;\; x \;\; \textit{service time per customer}}{\textit{total time}} =$$

$$= \frac{\lambda T \frac{1}{c\mu}}{T} = \frac{\lambda}{c\mu},$$

where c is number of service channels[3].

• **Server idle time:** average time a server spends between helping customers:

$$\textit{system's expected time spent in service} =$$

$$= \textit{expected service time} \;\; x \;\; \textit{number of customers} =$$

$$= \frac{1}{\mu} N_c = \frac{1}{\mu} \lambda T \text{ (from eq. 5)} = \frac{\lambda}{\mu} T,$$

where N_C is number of customers.

Per service channel, busy time is reduced by the number of servers:

$$\frac{\lambda}{c\mu} T = P_b T$$

Expected idle time per server: $T - P_b T = T(1 - P_b)$

Appendix A Definitions

Balking: leaving the system before entering the waiting line.

Commodity: a discrete unit that is expected to be serviced.

Discrete: refers to an individual event or a commodity.

Exponential distribution: describes the time between the events in a Poisson process.

Jockeying: moving from one queue to another.

Poisson distribution: describes the probability of discrete events occurring with a known average rate, independent of the time since the last event and the system state.

Reneging: leaving the system after having spent some time in the waiting line.

Steady state: describes the system which has been in use for some time where customer flow as well as service rate have stabilized.

Stochastic process: a random process.

System state: total number of commodities in the queuing system.

Queue: waiting line where the commodity remains until a service channel becomes available.

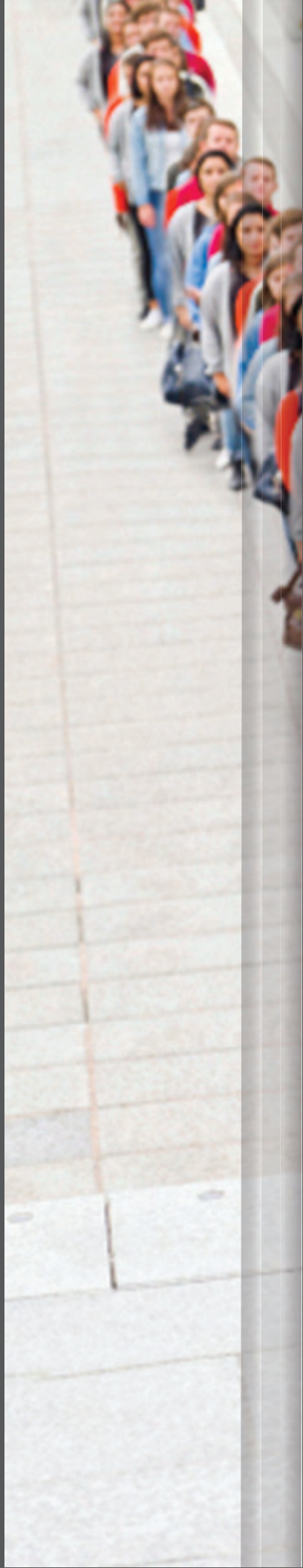
Queuing system: a type of flow system in which a commodity enters the system, waits for service if no service channel is available, receives the service, and exits the system.

Appendix B References

1. Kleinrock, Leonard. “Queuing Systems” v.1 Theory, John Wiley & Sons, 1975

2. Gross, Donald and Carl M. Harris. “Fundamentals of Queuing Theory.” Wiley Inter-Science, 1998

3. Pinker. “Urban Operations Research.” Supplementary Notes. Updated by Kang. Available on MIT Open Courseware: *http://ocw.mit.edu/courses/civil-and-environmental-engineering/1-203j-logistical-and-transportation-planning-methods-fall-2004/lecture-notes/class12mg1q.pdf*



Lavi Industries

Lavi Industries
27810 Avenue Hopkins
Valencia, CA 91355-3409
www.lavi.com