****

9B19E006

a note on QUEUING MODELS

[Felipe Rodrigues](https://iveypubs.my.salesforce.com/0031J00001UQbJU) wrote this note under the supervision of [Rasha Kashef](https://iveypubs.my.salesforce.com/003A000001lz6V1) solely to provide material for class discussion. The authors do not intend to illustrate either effective or ineffective handling of a managerial situation. The authors may have disguised certain names and other identifying information to protect confidentiality.

This publication may not be transmitted, photocopied, digitized, or otherwise reproduced in any form or by any means without the permission of the copyright holder. Reproduction of this material is not covered under authorization by any reproduction rights organization. To order copies or request permission to reproduce materials, contact Ivey Publishing, Ivey Business School, Western University, London, Ontario, Canada, N6G 0N1; (t) 519.661.3208; (e) cases@ivey.ca; www.iveycases.com. Our goal is to publish materials of the highest quality; submit any errata to publishcases@ivey.ca. i1v2e5y5pubs

Copyright © 2019, Ivey Business School Foundation Version: 2019-05-23

Queues (or waiting lines) are common in modern life. We wait in lines at the campus cafeteria, to board an airplane, in the emergency room, and when we call a company for customer service. We even wait for applications to be initiated and processed by our smartphone processors, and our online orders wait to be fulfilled at companies’ warehouses before being shipped to us.

How long we wait in line depends on several factors and parameters. The wait may be a result of the number of servers working to serve us and the amount of time it takes to serve each individual customer (i.e., the system’s capacity). It also depends on how many customers arrive at any given time (i.e., the system’s demand, or flow rate), but most importantly, queues form due to random variations in arrival patterns and service times.

For example, if a small call centre receives one call exactly every 10 minutes and its customer service representative takes exactly eight minutes to resolve that call, it seems intuitive that the next customer to call would not wait. After all, the customer service representative would have finished the previous call two minutes before the next customer calls. On the other hand, if the same call centre receives one call every 10 minutes *on average* and processes that call in eight minutes *on average*, a customer may call while the representative is still dealing with the previous customer. We have formulas to calculate the probability of that happening, the average wait time, and the size of this system’s queue. (The average wait may be around 32 minutes, and an average of 3.2 customers may be waiting at any given time.) What a difference randomness brings to a system!

Often, a system that has a low level of service (e.g., slow service or too few staff) may be very inexpensive to run (as the small call centre presented previously), but it may incur a high level of customer dissatisfaction, leading to lost future business and rising costs for processing complaints. Adding extra staff (thereby improving the service level) or improving service time (e.g., via automation) can be costly; however, either change will tend to reduce the wait lines and attract more consumers. Because of this trade-off, a common managerial decision is to determine a service level that optimally balances the cost of service and the cost of waiting.

BASIC QUEUING THEORY AND FORMULAS

Any queuing system (see Exhibit 1) is composed of the following key elements: a source population that generates arrivals into the system (e.g., customers, goods, calls, or patients); a queue (e.g., a wait line, inventory, or a waiting area); servers (e.g., clerks, machines, operators, or physicians); and an exit (e.g., leaving the system or returning to the source).

A common notation used to classify queuing systems is called “Kendall notation”[[1]](#footnote-1) (see Exhibit 2). A queuing system can be easily understood by using its A/P/s/K/n/D format. Where “A” stands for the arrival process, “P” for the service or process distribution, “s” for the number or servers or workstations, “K” for the number of places in the queue, “n” for the size of the population, and “D” for the arrival discipline. The arrival processes and the service time distributions are most commonly of the types described in Exhibit 3.

ELEMENTS OF WAITING LINES (QUEUING SYSTEMS)

Population

The sourcing population can be either finite or infinite. In a finite scenario, the arrival of an entity into a system can affect the rate of arrival for the remaining entities in a population. On the other hand, an infinite population is considered large enough that the arrival of an entity will not affect the rate of arrivals for the remaining population.

Arrivals

Arrivals are an external source of variability in the system. The mean arrival rate (lambda) refers to how many arrivals enter the system in a given unit of time (e.g., customers/hour, stock-keeping units [SKUs]/minute, calls/minute, or patients/day). Arrival rates are often described as being Poisson-distributed. The inverse of the arrival rate is the inter-arrival time (), which represents the time between consecutive arrivals. The inter-arrival time, assuming a Poisson arrival rate, is exponentially distributed, leading to random independently distributed arrivals.

Queue

The maximum available size of the queue can be finite and infinite. It is more common to assume an infinite maximum queue size.

Queue Discipline

Servers may collect entities from their queues according to different rules, with “first in, first out” (FIFO) being the most common. However, priority rules may be in place depending on many factors, such as preferential status, expiry dates, severity of disease, and others.

Customer Behaviour

In addition to waiting in line, units entering the system may also choose to balk, renege, or jockey. *Balking* takes place when a unit does not join the system. For example, a student observes the cafeteria’s current wait line and decides not to join it. *Reneging* happens when a unit enters the system but, after waiting for a period of time, decides to leave the system without being serviced. For example, you call your mobile phone carrier’s customer service, wait for your turn to talk to a representative, but hang up after 10 minutes of waiting. *Jockeying* happens when units that entered the system change waiting lines due to differences in perceived size, hoping to reduce time waiting to be served. Waits at highway toll stations tend to lead to this behaviour in impatient drivers.

Service

Servers are another internal source of variability in a queuing system. Similar to the arrivals, we have the mean service rate (mu), which refers to the service capacity, or the number of entities the server can process in a given unit of time (e.g., customers/hour, SKUs/minute, calls/minute, or patients/day). Also note its inverse, the service time , which represents the time the server takes to process one entity. A measurement of the variation of the service time is often described in percentage terms as a coefficient of variation , being the ratio of the standard deviation to the mean service time . When the standard deviation is zero (σ = 0), the coefficient of variation is also zero (i.e., ), the service time has no variation, and the service rate is therefore constant. When the standard deviation is equal to the mean service time (, then , and the service time is exponentially distributed. If the coefficient or variation differs from one, we may not assume exponential service times and may instead make use of general distributions. Moreover, a queuing system can also have a single server (e.g., s = 1) or multiple servers (e.g., s = 6).

Performance Measures

The focus of this note is an infinite population and FIFO discipline infinitely sized queues, so we will omit that part of the notation. Also, we assume that customers are patient, i.e., they do not balk, renege, or jockey.

When analyzing queues, we are most often interested in determining the performance indicators related to four categories: (1) the probability of having a certain number of units in a queue or in a system, (2) the number of units in the system (e.g., how many patients are waiting in the waiting room), (3) utilization rates, and (4) wait times. We will introduce some formulas for such indicators using the following notations:

: Arrival rate (mean number of units/time-unit)

: Service rate of one server (mean number of units/time-unit)

: Utilization rate (system congestion rate)

: Average length of the queue (number of units in the queue)

: Average length of the system (number of units in the system)

: Average time spent in a queue per customer (in time units)

: Average time spent in system per customer (in time units)

: Probability of having *n* units in the system

Coefficient of variation:

At all times, we will assume that the utilization rate is below 100 per cent or . This condition is necessary to achieve what we call a “steady state,” meaning the system is stable and the parameters do not change over the foreseeable future. If the utilization rate is above 100 per cent, the arrivals (demand) are higher than capacity. In such a scenario, the intuitive conclusion is that the queue would grow indefinitely over time.

calculating SYSTEM PERFORMANCE MEASURES

Now, let’s analyze some basic queuing systems by means of the formulas available in the literature. Primarily, this analysis can be done using the intuitive and elegant Little’s Law (),[[2]](#footnote-2) which states that, in a steady state (i.e., fixed parameters over time), the average number of units in a system is equal to the arrival rate multiplied by the average time the units spend in a system. This formula can be adapted to any queuing system or any general process, where a process inventory (L), flow rates (), and cycle times () can be calculated from any combination of two of these measurements.

SINGLE-SERVER WAITING LINE MODEL (M/M/1)

This system assumes exponential inter-arrival and service times with a single server, no restrictions on queue size, and a FIFO discipline.

* Average System Utilization Rate:

The ratio between the arrival rate and the service rate: < 1 for stability

* Probability of having 0 units in the system:

(100% capacity subtracted by the average utilization rate)

* Probability of having units in the system:

(Probability of having empty system multiplied by the utilization rate of as many independent arrivals you expect)

* Average time spent in system: or

The total time in the system (also known as the process cycle time) is related to the difference in service rate (capacity) and arrival rate (demand). Better yet, the wait time has an exponential distribution with parameter , which means the wait time is related to the service time (capacity) and the probability of an empty system (i.e., idle time).

* Average time spent in queue: or

The total time in the system minus the service time yields the time waiting for service. Operationally, the time waiting (non-value-added time) is the difference between the process cycle time and its run time.

* Average length of the system: or

Using Little’s Law, the average number of units in the system is calculated by multiplying the arrival rate by the average time spent in the system. In other words, the total inventory of an operation is calculated by multiplying its flow rate (demand) by the process’s cycle time.

* Average length of the queue: or or

Using Little’s Law, the average number of units in the queue is calculated by multiplying the arrival rate by the average time spent in the queue. This formula also expresses the idea of work-in-process inventory in an operation being the result of the flow rate (demand) times the non-value-added time.

As an example, students arrive at the parking lot entrance, where they must proceed through a pay station to enter the parking lot and park their car. The students patiently form a single line of cars in front of the pay station to wait their turn. Students are served based on a first-come, first-served priority rule (the order in which the cars arrive). On average, 12 students per hour arrive at the paying station. Student arrivals are best described using a Poisson distribution. The pay station can collect payments from an average of 15 students per hour, with the service rate being described by an exponential distribution. The characteristics of the parking lot system are calculated as in Exhibit 4, and the probabilities of *n* students in the system are shown in Exhibit 5.

MULTIPLE-SERVER MARKOVIAN WAITING LINE MODEL (M/M/s), where *s* = Number of Servers

This system assumes exponential inter-arrival and service times with multiple servers “*s*,” no restrictions on queue size, and a FIFO discipline. A performance calculation example can be found in Exhibit 6.

* Average system utilization rate:

We need to multiply the service rate by the number of available servers to determine the system capacity. This calculation accounts for system utilization depending on the number of parallel servers. Before using the formulas, ensure that the total service rate is greater than the arrival rate, that is, s, i.e., the system is in a steady state. As seen before, if s ≤ (capacity is smaller than the demand), the waiting line would eventually grow infinitely large.

* Probability of having 0 units (patients) in the system:
* Average length of the queue:
* Average length of the system:
* Average queue wait time:
* Average system total time:

M/G/1 Queuing Models

This queuing system assumes exponential inter-arrival times but does not require the service time distribution to be exponential. In fact, no assumptions are needed regarding the service time distributions.

The relaxation of the assumption of the exponential service time can be very useful for systems in which the standard deviation ( differs from the mean service time (a characteristic of the exponential distribution). As such, one needs only to estimate either the standard deviation ( or the coefficient of variation ( and the mean service time . The service time is therefore considered “general independent,” with a single server, no restrictions on queue size, and a FIFO discipline.

Pollaczek-Khintchine’s famous P-K formula,[[3]](#footnote-3) , finds the length of the queue and demonstrates the negative effect of variation in all performance measures. Note that if variance is equal to the service time, this formula reduces to the M/M/1 formula for . An example of system performance measures can be found in Exhibit 7.

* Probability of having 0 units at the queue:
* Length of the queue: (Pollaczek-Khintchine formula) or
* Intuition: the larger the variation in the service time, the larger the queue.
* Queue wait time: or
* Length of the system: or
* System total time: or

M/D/1 Queuing Models

This is a special case of the M/G/1 model, where there is no variation in service time ( making it a constant or deterministic service. An example of system performance measures can be found in Exhibit 8.

* Probability of having 0 units in the queue:
* Length of the queue: or

Intuition: without variation, the length of the queue in the deterministic case is smaller than in the M/M/1 and M/G/1 models. For example, compare the M/D/1 model’s to the P-K formula.

* Queue wait time: or
* Length of the system: or
* System total time: or

G/G/1 Queuing Models

With greater variation now introduced into the previous M/G/1 system, we may now have two sources of variation, the inter-arrival time and the service time distribution. As such, the G/G/1 model tends to perform worse than the M/G/1 model when both arrival and service times have great variation. This system makes no assumption in terms of arrival and service distributions; thus, we have general independent inter-arrival and service times with single server, no restrictions on queue size, and a FIFO discipline. There are approximations for some performance indicators as below, now requiring the coefficient of variation of both inter-arrival time and service time .

* Probability of having 0 units at the queue:
* Length of the queue (Marchal’s equation[[4]](#footnote-4))

Intuition: compare it to the P-K formula and you will see how the extra variation influences the length of the queue.

* Queue wait time: (Kingman’s equation)
* Or

Intuition: In congested systems, the queue wait time in the G/G/1 queue is approximately the wait time of the M/M1 queue multiplied by the average variation in the system.

* Length of the system:
* System total time:

G/G/s Queuing Models

The G/G/S queuing models represent, by far, the most general case, where we may add another parameter to the Kingman equation, the number of servers, as done by Hopp and Spearman’s cycle time approximation formula.[[5]](#footnote-5) This approximation formula provides fast and reliable queue wait-time approximations, particularly in congested systems. The other performance metrics rely on Little’s Law. Hopp and Spearman’s formula is a “quick and dirty” approximation often used to provide estimated performance metrics, and due to Little’s Law, it is often use in operations management for cycle time and work-in-process inventory calculations.

* Queue wait time:

Intuition: The wait time in the queue is composed of three factors. The first, is the utilization factor. The busier the system (as measured by the longer the wait time. The second factor is a capacity factor, The more resources (as measured by *s* number of servers), the shorter the wait. The final factor, is the variation factor. The higher the variation of arrivals and/or service times, the longer the wait. Operationally, this formula is useful to calculate the non-value-added time of work-in-process inventory. Observe that Hopp and Spearman’s formula converges to the M/M/1 formula for when the number of servers is one and the coefficient of variation for both arrival () and service () is also one.

* Length of the queue: (Little’s Law)

Intuition: Work-in-process inventory is approximately the multiplication of the non-value-added time of a process by its flow rate (or demand).

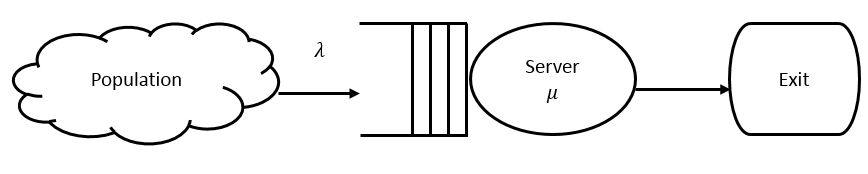
* System wait time:

Intuition: Total system wait time is the sum of the wait in queue and the service time. Operationally, the cycle time is equal to the non-value-added time plus the run time.

* Length of the system: (Little’s Law)

Intuition: Similar to the other models, the total inventory in a process is the result of cycle time multiplied by the flow rate (demand) in that process.

**exhibit 1:** **Queuing system elements**



Note: = arrival rate; service rate.

Source: Created by the case authors.

**exhibit 2:** **Kendall notation used to classify queuing systems**

|  |  |
| --- | --- |
| *A* | Arrival process |
| *P* | Service time distribution |
| *s* | Number of servers |
| *K* | Number of places in the queue |
| *n* | Size of the population |
| *D* | Queue discipline |

Source: Created by the case authors.

**exhibit 3:** **SYMBOLS AND DESCRIPTIONS OF Queuing system arrival processes and service time distributions**

|  |  |  |  |
| --- | --- | --- | --- |
| **Symbol** | **Name** | **Description** | **Example** |
| M | Markovian or Memoryless | Exponential service time or inter-arrival time | M/M/1/∞/∞/FIFO |
| Ek | Erlang distribution | An Erlang distribution with *k* as the shape parameter | M/Ek/2/10/∞/FIFO |
| D | Degenerate distribution | A deterministic or fixed time | M/D/1/∞/∞/FIFO |
| G | General distribution | General independent time | G/M/2/∞/∞/FIFO |

Note: ∞ = infinite; FIFO = first in, first out.

Source: Created by the case authors.

**exhibit 4:** **Spreadsheet for single-server operating characteristics**

|  |  |
| --- | --- |
| **M/M/1 Queue** | |
| Infinite population, Poisson arrivals, FIFO, Exponential service time, Unlimited queue length | |
| **Inputs** | |
| Units of time | hour |
| Units of arrivals | students |
| Arrival rate (lambda) | 12 students per hour |
| Service rate (mu) | 15 students per hour |
| Coefficient of variation (lambda) | 100% |
| Coefficient of variation (mu) | 100% |
| **Outputs** | |
| Direct outputs from inputs |  |
| Mean time between arrivals | 0.08333 hour |
| Mean time per service | 0.06667 hour |
| Standard deviation arrival time | 0.083333333 |
| Standard deviation service time | 0.066666667 |
| **Performance Indicators** | |
| Utilization rate of server | 80.00% |
| Average number of units waiting in line (Lq) | 3.20 students |
| Average number of units in system (L) | 4.00 students |
| Average time waiting in line (Wq) | 0.267 hour |
| Average time in system (W) | 0.333 hour |
| Probability of no units in system (P0) | 20% |

Note: FIFO = first in, first out.

Source: Created by the case authors.

**exhibit5:** **Probability of having “*n”* students in the system**

Source: Created by the case authors.

**exhibit 6:** **Spreadsheet for multiple-server operating characteristics**

|  |  |
| --- | --- |
| **M/M/s Queue** | |
| Multiple servers, Infinite population, Poisson arrival, FIFO, | |
| Exponential service time, Unlimited waiting room | |
| **Inputs** | |
| Unit of time | day |
| Arrival rate (lambda) | 4 units per day |
| Service rate (mu) | 2.5 units per day |
| Number of identical servers (s) | 2 servers |
| **Outputs** | |
| Direct outputs from inputs |  |
| Mean time between arrivals | 0.25 day |
| Mean time per service | 0.4 day |
| Traffic intensity | 0.8 |
| **Performance Indicators** |  |
| Average utilization rate of server | 0.8 |
| Average number of units waiting in line (Lq) | 2.844 units |
| Average number of units in system (L) | 4.444 units |
| Average time waiting in line (Wq) | 0.711 day |
| Average time in system (W) | 1.111 day |
| Probability of no units in system (P0) | 11.11% |
| Probability that all servers are busy | 71.11% |
| Probability that at least one server is idle | 28.89% |
| **Distribution of Number of Units in System** | |
| n (units) | P(n in system) |
| 4 | 9.10% |
| **Distribution of Time in Queue** |  |
| t (time in queue) | P(wait > t) |
| 0.25 | 55.38% |

Note: FIFO = first in, first out.

Source: Created by the case authors.

**exhibit 7:** **Spreadsheet for M/G/1 characteristics**

|  |  |
| --- | --- |
| **M/G/1 Queue** | |
| Single Server, Infinite Population, Poisson Arrival, FIFO, | |
| General Service Time, Unlimited Queue Length | |
| **Inputs** | |
| Unit of time | day |
| Arrival rate (lambda) | 4.000 units per day |
| Average service rate (mu) | 5.000 units per day |
| Coefficient of variation (mu) | **110%** of the Std. Dev. |
| **Outputs** | |
| Mean time between arrivals | 0.2500 day |
| Mean time per service | 0.2000 day |
| Traffic intensity | 80.00% |
| Service time standard deviation | 0.2200 day |
| **Performance Indicators** | |
| Utilization rate of server | 80.00% |
| Average number of units waiting in line (Lq) | 3.5 units |
| Average number of units in system (L) | 4.3 units |
| Average time in queue (Wq) | 0.884 day |
| Average time in system (W) | 1.084 day |
| Probability of no units in system (P0) | 20.00% |

Note: FIFO = first in, first out; Std. Dev. = standard deviation.

Source: Created by the case authors.

**exhibit 8:** **Spreadsheet for M/D/1 characteristics**

|  |  |  |
| --- | --- | --- |
| **M/D/1 Queue** | | |
| Single Server, Infinite Population, Poisson Arrival, | | |
| FIFO, Constant Service Time, Unlimited Queue Length | | |
| **Inputs** |  |  |
| Unit of time | days | |
| Arrival rate (lambda) | 4 units per days | |
| Service rate (mu) | 5 units per days | |
| **Outputs** | | |
| Mean time between arrivals | 0.250 days | |
| Exact time per service (no variability) | 0.2 days | |
| Traffic intensity | 0.8 | |
| **Performance indicators** | | |
| Utilization rate of server | 80.0% | |
| Average number of units waiting in line (Lq) | 1.6 units | |
| Average number of units in system (L) | 2.4 units | |
| Average time waiting in line (Wq) | 0.4 days | |
| Average time in system (W) | 0.6 days | |
| Probability of no units in system (P0) | 20.0% | |

Note: FIFO = first in, first out.

Source: Created by the case authors.

**exhibit 9:** **Spreadsheet for G/G/1 characteristics**

|  |  |
| --- | --- |
| Single Server, Infinite Population, General Service Arrival, FIFO, | |
| General Service Time, Unlimited Queue Length | |
| Unit of time | day |
| Arrival rate (lambda) | 4 units per day |
| Average service rate (mu) | 5 units per day |
| Arrival time standard deviation | 0.275 day |
| Service time standard deviation | 0.22 day |
| Mean time between arrivals | 0.250 day |
| Mean time per service | 0.2 day |
| Traffic intensity | 0.8 |
| Coefficient of variation (lambda) | 1.1 |
| Coefficient of variation (mu) | 1.1 |
| Parameter k (lambda) | 0.004726563 |
| Parameter k (mu) | 0.001936 |
| Utilization rate of server | 80.0% |
| Average number of units waiting in line (Lq) | 3.954 units |
| Average number of units in system (L) | 4.754 units |
| Average time waiting in line (Wq) | 0.968 day |
| Average time in system (W) | 1.189 day |
| Probability of no units in system (P0) | 20.0% |

Note: FIFO = first in, first out.

Source: Created by the case authors.

1. David G. Kendall, “Stochastic Processes Occurring in the Theory of Queues and Their Analysis by the Method of the Imbedded Markov Chain,” *Annals of Mathematical Statistics* 24, no. 3 (1953): 338. [↑](#footnote-ref-1)
2. John D.C. Little, “A Proof for the Queueing Formula: L = λW,” *Operations Research* 9, no. 3 (1961): 383–387. [↑](#footnote-ref-2)
3. Felix Pollaczek, “Über Eine Aufgabe der Wahrscheinlichkeitstheorie,” *Mathematische Zeitschrift* 32, no. 1 (1930): 64–100. [↑](#footnote-ref-3)
4. William G. Marchal, “An Approximate Formula for Waiting Time in Single Server Queues,” *AIIE Transactions* 8, no. 4 (1976): 473–474. [↑](#footnote-ref-4)
5. Wallace J. Hopp and Mark L. Spearman, *Factory Physics: Foundations of Manufacturing Management*, 2nd ed. (New York: McGraw-Hill, 2001). [↑](#footnote-ref-5)