Notes on *Operational Research*
March, 2008; rev. May, 2015. IST

Queueing systems

MIGUEL A. S. CASQUILHO
IST, Universidade de Lisboa,
Ave. Rovisco Pais, IST; 1049-001 Lisboa, Portugal
Telephone: (+351) 21.841 7310; fax: (+351) 21.849 9242

Queueing systems are presented, with a brief introduction and formulas for usual practical cases. Some examples are solved and computer resolution is mentioned.

Keywords: *queueing systems, queueing theory, queue, waiting line.*

1. Fundamental and scope

The waiting phenomena, which originate the *queues*, are related to *random processes*, i.e., the models of which include random components. These are associated to probability.

The queue is almost inevitable in many situations, unless means are made available at costs possibly disproportionate to the benefits of a quick service. When circumstances impose a quick service, capable of limiting the waiting time to a reasonable level, the working conditions can be evaluated through the queueing systems theory[1].

The *queues* are frequent phenomena found in everyday life, and also in situations in economics, society, and the military. Examples: customers in a bank or post office; people waiting for a taxi or telephoning to a taxi service; cars at a (road) junction[2]; planes waiting to land or take off; broken machines waiting for repair. Several examples are given in Fig. 1. Erlang in the 1920's was one of the first to study the queueing subject applying it to the telephone system.

Arrivals	Nature of service	Servers
Customers	Sale of an article	Vendors
Ships	Unloading	Docks
Planes	Landing	Tracks
Telephone calls	Conversations	Telephone circuits
Arrival of cars	Customs control	Customs workers
Messages	Decoding	Decoders
Repair machines	Repair	Mechanics
Fires	Fire fighting	Fire brigade
Requests	Confection, repair	Repair-shop

Fig. 1 Examples of waiting phenomena.

A queue is characterized by several components: customers' population, arrival pattern, number of servers, service pattern, system capacity (size) to hold customers, and the queue discipline. Consideration of the costs of maintaining a queueing system

[1] US "waiting line"; Pt «filas de espera», «bichas»; Es «colas»; Fr «phénomènes, files d'attente»; It « fenomeni (o file) d'attesa, code»; De »Schlange(n)«.
[2] US "intersection"; Pt «cruzamento»; Fr, «carrefour».

Miguel Casquilho is Assistant Professor (ret.) in the Departament of Chemical Engineering, Instituto Superior Técnico, Universidade de Lisboa (University of Lisbon), Lisbon, Portugal. E-mail address: mcasquilho@tecnico.ulisboa.pt.

from the supplier side and the customers' side makes it an economic optimization problem. The objective of this text is to present formulas that permit that optimization.

2. Queues structure

The structure of a queueing system is addressed based on the above mentioned parameters and characteristics. A systematization of the queueing systems by the Kendall's notation is given, as well as a nomenclature.

Customers' population

The *customers' population* may be *infinite* or *finite*. It is finite if the number of possible customers is limited and known, such as the number of machines subject to failure in a factory; infinite, otherwise.

Arrival pattern

The *arrival pattern* of customers is usually specified by the *interarrival time*, the time between successive customer arrivals to the service. It may be deterministic or a random variable with a probability distribution presumed known. [Other aspects will not be considered here, such as: arriving singly or in batches; or balking (refusal to enter) or reneging (leaving the queue because the wait is too long).]

Number of servers

The *number of servers* is the number of persons, machines, tellers, gates, etc., to attend customers. These will be considered equivalent and in parallel (other cases being series or more or less complex combinations of servers in series and in parallel).

Service pattern

The *service pattern* is usually specified by the *service time*, which may be deterministic or a random variable with probability distribution assumed known. (The service time may depend on the number of customers. The customer may be attended completely by one server or any combination of servers.)

System capacity

The *system capacity* is the maximum number of customers, both those in service and those in the queue(s). Whenever a customer arrives at a facility that is full, the customer is denied entrance to the facility and not allowed to wait outside the facility, which would increase the limited capacity, and is forced to leave. Capacity is, thus, either *infinite* or *finite*.

Queue discipline

The *queue discipline* is the order in which customers are served. This can be on a first-in, first-out (FIFO) basis (i.e., service in order of arrival, the usual one), a last-in, first-out (LIFO) basis, a random basis or a priority basis (as in hospital emergency services).

To make queue classification simpler, the so-called Kendall's notation is usually employed.

Kendall's notation

The Kendall's notation indicates ({1}): v, the arrival pattern; w, the service

$$v / w / x / y / z \qquad \{1\}$$

pattern; x, the number of servers; y, the system's capacity; and z, the queue discipline, as in Table 1. If y or z is not specified, it is taken to be ∞ or FIFO, respectively.

Table 1 Kendall's notation

	Queue characteristic	Symbol	Meaning
v, w	Interarrival time or service time	D	Deterministic
		M	Exponential
		E_k	Erlang-type (k = 1, 2 …)
		G	Any other
x	Number of servers	*Number*	∞ if not specified
y	System's capacity	*Number*	∞ if not specified
z	Queue discipline	FIFO	First in, first out
		LIFO	Last in, first out
		SIRO	Service in random order
		PRI	Priority ordering
		GD	Any other ordering

The initials are related to: D, deterministic or "degenerate" (a deterministic variable being a constant, a degenerate random variable); M, Markovian (Markovian "birth and death" process, typically with Poissonian arrivals); G, general.

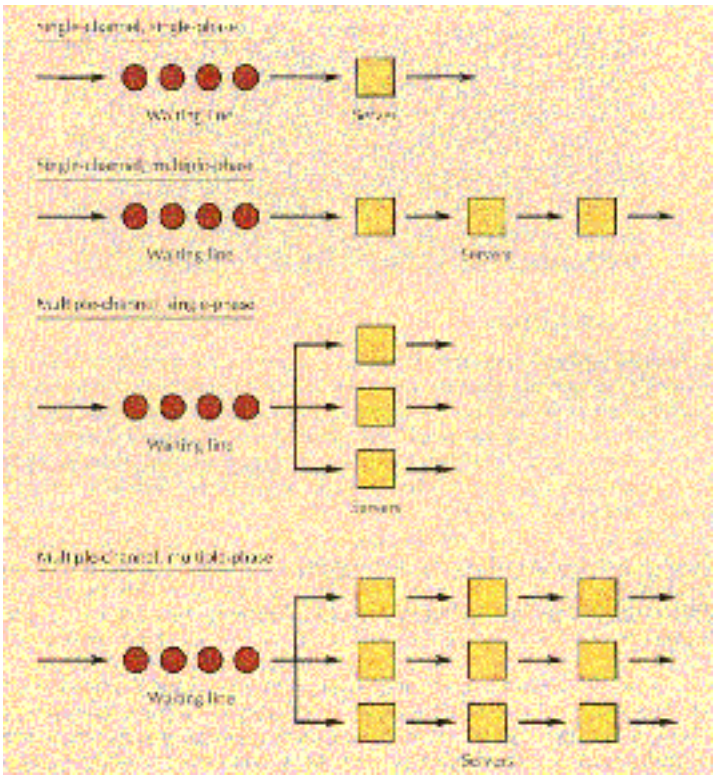

Fig. 2 Simplified queue taxonomy.

Let it be noted that in the frequent M/M/*s* case of more than one server, $s > 1$, the customers (and the selling entity) benefit from a *single* queue (which is rarely the case in large stores) [Ravindran *et al.*, 1987, 329]. This can be easily accomplished by making available *numbered tickets* (as in post offices and usually pharmacies, in Portugal).

For a simplified taxonomy of queues, see Nemetz-Mills [2008], from whom Fig. 2 was taken. This author mentions "single or multiple channel", i.e., single or

multiple servers —here, M/M/1 and M/M/s— and "single or multiple-phase". A multiple-phase queueing system (2.nd and 4.th rows in the figure) is a ("pure") mixture of parallel and series servers, a complex case having a better resolution by Monte Carlo simulation.

Only M/M/1/∞/FIFO and M/M/s/∞/FIFO systems, i.e., for short,

$$\text{M/M/1} \qquad \text{M/M/}s \qquad \{2\}$$

will be addressed in the following sections.

3. Single and multiple server queues

The set of a queue (or queues) and the servers constitutes the *waiting system* or simply the *system*. In the cases where it is supposed to have several queues, the *customers* place themselves either automatically in the shortest queue or according to a priority. (The term "customer" will be used instead of the more general "unit", whether it is a person or any other entity.) These priorities make the queue discipline (hospitals, restaurants).

With the given structure of a waiting phenomenon, the notation in Table will be used:

Table 2 Notation

	Meaning
m	Number of existing customers (population size)
n	Number of customers in the system (waiting or being served)
α	Arrival rate (T^{-1}, customers / time unit)
μ	Service rate (T^{-1}, services / time unit)
ψ	Utilization factor, or traffic intensity, $\alpha/(s\mu)$
ν	Number of customers in queue
j	N. of customers being served
s	N. of servers

So, it is

$$\begin{aligned} j &= n \quad \text{if} \quad n \le s \\ \nu + j &= n \qquad\quad n > s \end{aligned} \qquad \{3\}$$

The values n, ν and j are random. If it is

$$p_n = \Pr(n \text{ customers in the system}) \qquad \{4\}$$

then, p_n, a probability, represents the fraction of the time the system is in state n.

3.1 Single server queues

The basic variables for a single server queue system, $s = 1$, will now be determined for the simpler and usual case of an infinite population, i.e., $m = \infty$.

The Poisson process is often used to model the situation in which a count is made on the number of events occurring in a given time, here the arrival of customers to a service facility: $p_{\text{Poi}}(j) = [\exp(-\alpha t)](\alpha t)^j / j!$, $j = 0..\infty$ ($[\alpha] = T^{-1}$). The time

between events in a Poisson process follows an Exponential[3] distribution with the same parameter α, $f(t)=\frac{1}{\tau}\exp\left(-\frac{t}{\tau}\right)$, with mean $\tau = 1/\alpha$ ($[\tau]$ = T). The parameter τ is the expected time between events.

To find p_n, consider its evolution during an instant, from time t to $t + \mathrm{d}t$, with dt small enough so that no two (or more) events can occur.

$$p_0(t+\mathrm{d}t) = p_0(t)\underset{\text{no change}}{(1-\alpha\,\mathrm{d}t)} + p_1(t)\underset{\text{departure}}{(\mu)}\,\mathrm{d}t \qquad \{5a\}$$

$$p_n(t+\mathrm{d}t) = p_{n-1}(t)\underset{\text{arrival}}{(\alpha)}\mathrm{d}t + p_{n+1}(t)\underset{\text{departure}}{(\mu)}\,\mathrm{d}t + p_n(t)\underset{\text{no change}}{[1-(\alpha+\mu)\mathrm{d}t]} \qquad \{5b\}$$

This becomes

$$\frac{p_0(t+\mathrm{d}t)-p_0(t)}{\mathrm{d}t} = -\alpha p_0(t) + \mu p_1(t) \qquad \{6a\}$$

$$\frac{p_n(t+\mathrm{d}t)-p_n(t)}{\mathrm{d}t} = \alpha p_{n-1}(t) + \mu p_{n+1}(t) - (\alpha+\mu)p_n(t) \qquad \{6b\}$$

Introducing the *utilization factor* [H&L, 2005, 770] or *traffic intensity* [Ravindran *et al.*, 1987, 320]

$$\psi = \frac{\alpha}{S\mu} \qquad \{7\}$$

which is here simply $\psi = \frac{\alpha}{\mu}$, and in the limit as d$t$ goes to zero, it is

$$\frac{1}{\mu}p_0'(t) = -\psi p_0(t) + p_1(t) \qquad \{8a\}$$

$$\frac{1}{\mu}p_n'(t) = \psi p_{n-1}(t) + p_{n+1}(t) - (1+\psi)p_n(t) \qquad \{8b\}$$

The system will be studied only in the steady state (null derivatives), so it is

$$-\psi p_0 + p_1 = 0 \qquad \{9a\}$$

$$\psi p_{n-1} + p_{n+1} - (1+\psi)p_n = 0 \qquad \{9b\}$$

or [from $p_2 = (1+\psi)p_1 - \psi p_0$, $p_3 = (1+\psi)p_2 - \psi p_1$, etc.]

$$p_1 = \psi p_0 \qquad \{10a\}$$

[3] Also called "negative exponential" [Ravindran *et al.*, 1987, 293].

$$p_2 = (1+\psi)(\psi p_0) - \psi p_0 = (\psi^2 + \psi - \psi)p_0 = \psi^2 p_0$$
$$p_3 = (1+\psi)(\psi p_2) - \psi p_1 = (\psi^3 + \psi^2 - \psi^2)p_0 = \psi^3 p_0 \qquad \{10b\}$$
etc.

So, in general, it is

$$p_n = p_0 \psi^n \qquad \{11\}$$

The population size is $m = \infty$. As the probabilities must, of course, add to one, and recognizing the sum of a geometric series (it is $\psi < 1$), it is

$$1 = \sum_{n=0}^{\infty} p_n = p_0 \sum_{n=0}^{\infty} \psi^n = \frac{p_0}{1-\psi} \qquad \{12\}$$

Thus, it is

$$p_0 = 1 - \psi \qquad \{13\}$$

and generally

$$p_n = (1-\psi)\psi^n \qquad \{14\}$$

[The *geometric distribution* can be recognized in Eq. {14}: $p(n) = r(1-r)^n$, $n = 0..\infty$, with parameter $r = 1 - \psi$, mean $\mu = (1-r)/r$, i.e., $\psi/(1-\psi)$.]

The probability p_0 is the fraction of time the system is ***idle*** (empty), and the parameter ψ can be taken as the fraction of time the server is ***busy*** [Ravindran *et al.*, 1987, 320].

The *mean* or *expected value* of the **number of customers in the system** is, by the definition of mean,

$$\begin{aligned} \bar{n} &= \sum_{n=0}^{\infty} n p_n = (1-\psi)\sum_{n=0}^{\infty} n\psi^n = (1-\psi)\sum_{n=1}^{\infty} n\psi^n = (1-\psi)\psi\sum_{n=1}^{\infty} n\psi^{n-1} = \\ &= (1-\psi)\psi \frac{\mathrm{d}}{\mathrm{d}\psi}\sum_{n=1}^{\infty}\psi^n = (1-\psi)\psi\frac{\mathrm{d}}{\mathrm{d}\psi}(1-\psi)^{-1} = (1-\psi)\psi(1-\psi)^{-2} \end{aligned} \qquad \{15\}$$

or

$$L \equiv \bar{n} = \frac{\psi}{1-\psi} \qquad \{16\}$$

The mean **number of customers in the queue,** or mean queue length (with a queue of zero if there are 0 or 1 customers in the system), is

$$\begin{aligned} L_q \equiv \bar{\nu} &= \sum_{n=2}^{\infty}(n-1)p_n = (1-\psi)\sum_{n=2}^{\infty}(n-1)\psi^n = \\ &= (1-\psi)\psi^2\sum_{n=2}^{\infty}(n-1)\psi^{n-2} = (1-\psi)\psi^2\frac{\mathrm{d}}{\mathrm{d}\psi}\sum_{n=2}^{\infty}\psi^{n-1} = \frac{\psi^2}{1-\psi} \end{aligned} \qquad \{17\}$$

The difference between L (customers in the system) and L_q (customers in the queue) should be, and is, the mean number of busy servers, ψ:

$$L - L_q = \frac{\psi}{1-\psi} - \frac{\psi^2}{1-\psi} = \frac{\psi(1-\psi)}{1-\psi} = \psi \qquad \{18\}$$

An equation known as **Little's formula** (cited in most queueing literature) relates L to W, the mean waiting time in system:

$$L = (1 - p_N)\,\alpha W \qquad \{19\}$$

When it is $m = \infty$, as in the cases presented, the formula reduces to $L = \alpha W$ (as the probability p_m obviously tends to zero). This permits easily finding the **mean time in the queue**, W, and **mean time in the system**, W_q. The formulas for the M/M/1 case are shown in Table 3. As α and μ are rates (times per unit time), the expressions with $1/\alpha$ or $1/\mu$ represent, indeed, time.

Table 3 Synopsis for M/M/1

Variable and formula	
Probability of 0 customers in the system	
$p_0 = 1 - \psi \quad$ with $\quad \psi = \frac{\alpha}{\mu} < 1$	(a)
Probability of n customers in the system	
$p_n = (1-\psi)\psi^n$ $P_n = \sum_{j=0}^{n} p_j = 1 - \psi^{n+1}$	(b)
Mean of no. of customers in the queue (waiting)	
$L_q = \frac{\psi^2}{1-\psi}$	(c)
Mean of no. of customers in the system	
$L = \frac{\psi}{1-\psi} = L_q + \psi$	(d)
Mean of time in the queue (a customer waiting)	
$W_q = \frac{1}{\alpha}\frac{\psi^2}{1-\psi} = \frac{\psi}{\mu - \alpha}$	(e)
Mean of time in the system (a customer spending)	
$W = \frac{L}{\alpha} = \frac{1}{\alpha}\frac{\psi}{1-\psi} = \frac{1}{\mu-\alpha} = W_q + \frac{1}{\mu}$	(f)

The probabilities of waiting at least t (with $t \geq 0$) are given [H&L, 1995, 681] by

$$\begin{aligned} \Pr(\text{wait} > t) &= \exp[-\mu(1-\psi)t] \\ \Pr(\text{wait}_q > t) &= \psi \Pr[\text{wait} > t] \end{aligned} \qquad \{20\}$$

[the first expression an exponential distribution with parameter $\mu(1-\psi)$] which lead to (and confirm) $W = 1/(\mu - \alpha)$ and $W_q = \psi/(\mu - \alpha)$.

3.2 Multiple server queues

For this case, similar but more laborious derivations can be made. The results only are presented in Table 4. A *single* queue for customers waiting and steady state are also supposed.

In the particular case of $s = \infty$, it is

$$p_0^{-1} = \frac{1}{1-\psi}\lim_{s\to\infty}\frac{(s\psi)^s}{s!} + \exp(s\psi) = \exp(s\psi) = \exp\left(\frac{\alpha}{\mu}\right) \qquad \{21\}$$

Eq. {21} comes from the fact that *(i)* the sum (from 0 to $s-1$) can be recognized as the Taylor series development of the exponential function and *(ii)* the other term goes to zero. So, p_0 becomes a constant:

$$p_0 = \exp(-\alpha/\mu) \qquad \{22\}$$

The remaining variables will have the following values:

$$p_n = p_0\frac{\rho^n}{n!}$$
$$L_q = 0 \qquad L = L_q + \rho = \rho = \frac{\alpha}{\mu}; \qquad W_q = 0 \qquad W = W_q + \frac{1}{\mu} = \frac{1}{\mu} \qquad \{23\}$$

Indeed, p_0 is not 1 (a value that might be intuitive), as there are customers arriving; L_q is zero (zero customers waiting), but L is not zero, as they are being served (spending useful time); and W_q is zero (no wait in queue), but W is the inevitable service time, $1/\mu$ (not zero). This may be the case of a self-service situation if there are "many" servers, enough for all the arriving customers.

Table 4 Synopsis for M/M/*s*

Variable and formula
Probability of 0 customers in the system
$p_0^{-1} = \frac{(s\psi)^s}{s!(1-\psi)} + \sum_{n=0}^{s-1}\frac{(s\psi)^n}{n!} \qquad \text{with} \quad \psi = \frac{\alpha}{s\mu} < 1$ (a)
Remark: p_0^{-1}, not p_0
Probability of *n* customers in the system
$p_n = \begin{cases} \frac{(s\psi)^n}{n!}p_0 & 0 \le n \le s \\ \frac{s^s}{s!}\psi^n p_0 & n \ge s \end{cases}$ (b)
$P_n = P_{s-1} + \sum_{j=n}^{\infty}\frac{s^s}{s!}\psi^j p_0 = P_{s-1} + p_0\frac{s^s}{s!(1-\psi)}\psi^n \qquad n \ge s$
Mean of no. of customers in the queue (waiting)
$L_q = p_0\frac{s^s\psi^{s+1}}{s!(1-\psi)^2}$ (c)

Mean of no. of customers in the system

$$L = \alpha W = L_q + s\psi = \alpha\left(W_q + \frac{1}{\mu}\right) \qquad \text{(d)}$$

Mean of time in the queue (a customer waiting)

$$W_q = \frac{L_q}{\alpha} \qquad \text{(e)}$$

Mean of time in the system (a customer spending)

$$W = W_q + \frac{1}{\mu} \qquad \text{(f)}$$

The probabilities of waiting at least t (with $t \geq 0$) are given [H&L, 1995, 684] by

$$\Pr(\text{wait} > t) = \mathrm{e}^{-\mu t}\left\{1 + p_0 \frac{(s\psi)^s}{s!(1-\psi)} \frac{1-\exp\left[-\mu t(s-1-\rho)\right]}{s-1-\rho}\right\} \qquad \{24\}$$

$$\Pr(\text{wait}_q > t) = (1 - P_{s-1})\exp\left[-s\mu(1-\rho)t\right]$$

which leads to (and confirms) $W = 1 / (\mu - \alpha)$.

4. Illustrative examples

Suppose α = 10 hr^{-1} and μ = 15 hr^{-1} (data from Baker's [2006, 2] pharmacy example). For s = 1, s = 2 (in the reference), and s = 100, the results are given in Table 5. (In the reference, ρ is used for ψ.) In this case, with α / μ = 0.667, they show, namely, little difference from 2 to 100 servers.

Table 5 Results for growing s (other data constant)

	s = 1	s = 2	s = 100
ψ (or ρ)	0.667	0.333	0.007
p_0	0.333	0.500	0.513
L_q	1.333	0.083	0.000
L	2.000	0.750	0.667
W_q	0.133	0.008	0.000
Service time, $1 / \mu$	0.067	0.067	0.067
W	0.200	0.075	0.067

Various examples can be run on the author's Internet page [Casquilho, 2008]. Also, an economic optimization of s can be made there.

5. Conclusions

The theory applicable to queueing systems —provided that the underlying conditions are met, namely, steady state— can lead to useful results, permitting significant control on the behaviour of such systems. The calculations are cumbersome, adequate to computer treatment.

Acknowledgements

This study was made within the author's teaching and research activities in the "Centro de Processos Químicos" (Centre for Chemical Processes), Department of

Chemical and Biological Engineering, Instituto Superior Técnico, Universidade Técnica de Lisboa (Technical University of Lisbon). The computing and Internet publishing was made at the "Centro de Informática do IST" (IST Informatics Centre).

References

– BAKER, Samuel L., 2006, Internet page: ("%20" means blank) (visited Feb. 2008) *http://hspm.sph.sc.edu/Courses/J716/pdf/716-8%20Queuing%20Theory%20II.pdf*.
– CASQUILHO, Miguel, 2008, Internet page: *http://web.ist.utl.pt/mcasquilho/compute/or/Fx-queues.php*.
– (H&L) HILLER, Frederick S., Gerald J, LIEBERMAN, 2005 (2001, 1995, 1990, 1986, 1980, 1974, 1967), "Introduction to Operations Research", 8.th ed., McGraw-Hill, New York, NY (USA), (xxv+1062 pp), ISBN 007-123828-X.
– NEMETZ-MILLS, Patricia, Internet page: *http://www.cbpa.ewu.edu/~pnemetzmills* [Feb, 2008]
– RAVINDRAN, A., Don T. PHILLIPS, James J. SOLBERG, 1987, "Operations Research: principles and practice", 2.nd ed., John Wiley & Sons, New York, NY (USA), (xviii+637 pp), ISBN 0-471-85980-X.

❖