Statistical Queuing Theory with Some Applications

Manuel Alberto M. Ferreira[#1], Marina Andrade[#2], José António Filipe[#3], Manuel Pacheco Coelho[*4]

[#]*Department of Quantitative Methods*
Instituto Universitário de Lisboa (ISCTE-IUL), UNIDE – IUL
Lisboa Portugal

[*]*SOCIUS & ISEG/UTL - Portugal*
[1]manuel.ferreira@iscte.pt
[2]marina.andrade@iscte.pt
[3]jose.filipe@iscte.pt
[4]coelho@iseg.utl.pt

***Abstract*— An overview of theory of queues, single node and in network, is presented in this paper. In addition some applications are outlined. Some very well-known and others more uncommon.**

***Keywords*—** *Queues, networks of queues, applications.*

1. Queues

Consider a Service Centre at which arrive units, the customers, requiring service to other units, the servers, with or without distinction among customers and servers.

The most challenging situations in these systems study, subject matter of Statistical Queuing Theory, occur when it is assumed that:

- Customers arrivals are a stochastic process,
- Each server spent time to supply to each customer the required service is a random variable.
- Other relevant factors are:
- The number of servers that may be finite or infinite, constant or variable,
- If the number of servers is finite, some customers will have to wait to be served. The waiting capacity, that may be finite or infinite, is the maximum number of customers that may stay in the Service Centre waiting to be served. The system capacity is the maximum number of customers, being served or waiting for service, which are allowed to stay in the Service Centre simultaneously. When a customer arrives at a Service Centre with complete capacity it is considered lost to the system. So the queue systems with finite capacity are systems with losses.
- If the number of servers is infinite a customer that arrives finds immediately an available server. So there is no queue in the formal sense of the term. The queue systems with infinite servers are systems with neither waiting nor losses.
- The queue discipline is the method as the customers are selected by the servers or vice-versa. Some examples of queue disciplines are:
    - "First come-first served" (**FCFS**);
    - "Last in-first out" (**LIFO**);
    - "First in-first out" (**FIFO**);
    - "Processor sharing" (**PS**);
    - "Service in Random Order" (**SIRO**);
    - "Priority" (**PRI**);
    - "General Discipline" (**GD**).

The arrival process is usually characterized by the length of the time probability distribution between two successive arrivals of customers at the Service Centre: the inter-arrivals time. It may be deterministic or stochastic. There are models where batch arrivals are considered: the number of customers, arriving at each instant of the sequence of the arrivals instants, is a random variable R that can assume integer values greater than 1 - see, for instance, Shanbhag (1966). The arrival process may depend or not on the number of customers present at the Service Centre. Sometimes refusal situations are considered: the customer arrives and refuses to enter in the Service Centre because there are too many customers waiting to be served. And also renounce situations: the customer is already in the Service Centre and leaves it because it thinks that has waited a too long time.

The service process is specified indicating the length of the time probability distribution that a customer spends being attended by a server: the service time. There may be deterministic or stochastic service times.

A Service Centre which has associated a service process, a waiting capacity and a queue discipline is a node. A node with the respective arrival process is a queue.

The Kendall notation, see Kendall (1953), for describing queues is v/w/x/y/z where

- ***v*** denotes the arrival process (D, deterministic; M, exponential; E_k, Erlang (k); G, others),
- ***w*** denotes the service process (D, deterministic; M, exponential; E_k, Erlang (k); G, others),
- ***x*** denotes the number of servers,
- ***y*** denotes the system capacity,
- ***z*** denotes the queue discipline.

If y is not mentioned it is supposed to be infinite. If z is not mentioned it is supposed to be FCFS.

2. Networks of Queues

A network of queues is a collection of nodes, arbitrarily connected by arcs, across which the customers travel instantaneously and

- There is an arrival process associated to each node,
- There is a commutation process which commands the paths of the various costumers.

The arrival processes may be composed of exogenous arrivals, from the outside of the collection, and of endogenous arrivals, from the other collection nodes.

A network is open if any customer may enter or leave it. A network is closed if it has a fixed number of customers that travel from node to node and there are neither arrivals from the outside of the collection nor departures. A network open for some customers and closed for others is said mixed.

The commutation process rules, for each costumer that abandons a node, which node it can visit then or if it leaves the network. In a network with J nodes, the matrix

$$P=\begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1J} \\ p_{21} & p_{22} & \cdots & p_{2J} \\ \vdots & \vdots & & \vdots \\ p_{J1} & p_{J2} & \cdots & p_{JJ} \end{bmatrix}$$

is the commutation process matrix, being p_{jl} the probability of a customer, after ending its service at node j, go to node l, $j,l=1,2,\ldots,J$. The probability $q_j=1-\sum_{l=1}^{J} p_{jl}$ is the probability that a customer leaves the network from node j, $j=1,2,\ldots,J$.

A network of queues with infinite servers in each node, with Poisson process exogenous arrivals, may be looked like an M/G/∞ queue. The service time is the sojourn time of a customer in the network. Denote S the sojourn time of a costumer in the network and S_j its service time at node j, $j=1,2,..J$ Be $G(t)$ and $G_j(t)$ the S and S_j distribution functions, respectively and $\overline{G}(s)$ and $\overline{G}_j(s)$ the Laplace Transforms. If

$$\Lambda=\begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_J \end{bmatrix}$$

is the network exogenous arrival rates vector, where the rate λ_j is the exogenous arrival rate at node j and $\sum_{j=1}^{J}\lambda_j=\lambda$, making

$$\Lambda(s)=\begin{bmatrix} \lambda_1\overline{G}_1(s) \\ \lambda_2\overline{G}_2(s) \\ \vdots \\ \lambda_J\overline{G}_J(s) \end{bmatrix}$$

and

$$P(s)=\begin{bmatrix} p_{11}\overline{G}_1(s) & p_{12}\overline{G}_2(s) & \cdots & p_{1J}\overline{G}_J(s) \\ p_{12}\overline{G}_1(s) & p_{22}\overline{G}_2(s) & \cdots & p_{2J}\overline{G}_J(s) \\ \vdots & \vdots & & \vdots \\ p_{J1}\overline{G}_1(s) & p_{J2}\overline{G}_2(2) & \cdots & p_{JJ}\overline{G}_J(s) \end{bmatrix}$$

it results

$$\overline{G}(s)=\lambda^{-1}\Lambda^{T}(s)(I-P(s))^{-1}(I-P)\mathrm{A}$$

where *A* is a column with J 1`s, for the Laplace Transform service time (Ferreira and Andrade, 2010d).

The networks of queues with infinite servers in each node have interesting applications in Logistics, based on the failures of the transport vehicles that allow computing important measures of performance. See, for instance, Ferreira and Filipe (2010a,b), Ferreira, Andrade and Filipe (2009) and Ferreira *et al* (2009).

3. Stochastic Processes in Queues

A population is a set of objects that share common characteristics. Often, in practical situations, it is important to study statistically the expansion, or the reduction, of a population in order, eventually, to control it. If N (t) is the size of the population at instant t, the states of a population process are the various values that can be assumed by N (t) and the probability that N (t) =n, n=0, 1, 2… is denoted pn (t).

There is a birth when a new member joins the population. There is a death when a member leaves the population.

A population process is a Markov process if the changing from a state to other, eventually the same,

transition probabilities depend only on the initial state and not on the mutations experienced by the process till the arrival at the present state.

The probability distribution that rules the number of births and deaths in a certain time interval, in a Markov process, depends only on the interval length and not on the initial state.

A queue system is a birth and death process with a population composed by customers receiving a service or waiting for it. There is a birth when a customer arrives at the Service Centre. There is a death when a customer abandons the Service Centre. The state of the system is the number of the customers in the Service Centre. The population process is the most important quantity of interest in the study of queues. In particular, it is important the search for a stationary distribution for it. In this situation pn (t) do not depend on time and is denoted pn. Usually pn is obtained computing $\lim_{t\to\infty} p_n(t)$. The pn (t), depending on time, characterize the queue system transient behavior. The probabilities pn characterize the queue system stationary state, also called equilibrium state.

Other important quantities, that are measures of the queuing system performance, are the waiting time, also called queue time- the time that a customer spends in the system waiting for the service - and the sojourn time- the total time that a customer spends in the system: queue time plus service time.

Often it is difficult to obtain treatable formulae for the population process, the waiting time and the sojourn time and even to make an analytic study. So numerical and simulation methods are intensively used.

Based on the transient and stationary probabilities of infinite servers systems there are interesting applications in:

- ***Financial problems***
  The study of the sustainability of a pensions fund. See, Ferreira and Andrade (2011h) and Figueira and Ferreira (1999).
- ***Energy problems***
  How to deal with motor cars in a situation of scars energy – for instance in the end of oil reserves. See Ferreira, Filipe and Coelho (2008, 2011).

4. Traffic Intensity

The traffic intensity, ρ, is the most important parameter in queues study. See, for instance, Cox and Smith (1961). It is given by

$$\rho = \lambda\alpha$$

where λ is the arrival rate of the customers and α the mean service time.

Little's formula, see also Cox and Smith (1961), is, perhaps, the most popular result in queuing theory. It is a very general formula valid for any queue system that attains the stationary state. It relates the mean number of customers in the system, N, with the mean sojourn time of a customer, W, through the arrival rate, λ:

$$N = \lambda W.$$

The Pollaczeck-Khinchine formula (Cox and Smith, 1961) is used, for the M/G/1 queue, to evaluate the mean waiting time of a customer in the system:

$$W_s = \alpha \frac{\rho}{1-\rho} \frac{1+C_s^2}{2}$$

C_s is the service time coefficient of variation.

The mean sojourn time of a customer in the system is then

$$W = W_q + \alpha .$$

5. Busy Period

The busy period of a queue system begins when a customer arrives there, finding it empty, and ends when a customer leaves the system letting it empty. Along the busy period there is always at least one customer in the system. In any queue system there is a sequence of idle periods and busy periods. In systems with Poisson arrivals the idle period length is always exponential. The statistical study of the busy period is always a very difficult task. In general the busy period length is related with the transient behavior (see, for instance, Ferreira and Andrade, 2009a,b). An idle period followed by a busy period is a busy cycle.

For a M/G/∞ queue, if the service time distribution function belongs to the collection

$$G(t) = 1 - \frac{\left(1-e^{-\rho}\right)(\lambda+\beta)}{\lambda e^{-\rho}\left(e^{(\lambda+\beta)t}-1\right)+\lambda}, t \geq 0, -\lambda \leq \beta \leq \frac{\lambda}{e^{\rho}-1}$$

the busy period length distribution function is

$$B^{\beta}(t) = 1 - \frac{\lambda+\beta}{\lambda}\left(1-e^{-\rho}\right)e^{-e^{-\rho}(\lambda+\beta)t}, t \geq 0,$$
$$-\lambda \leq \beta \leq \frac{\lambda}{e^{\rho}-1},$$

a mixture of a degenerate distribution at the origin and an exponential distribution (Ferreira and Andrade, 2009a,b).

The busy period of the $M|G|\infty$ queue may be used to model socio-economic problems as, for example

- ***Disease problems***
  An epidemic situation may be assumed as being a busy period. An idle period is the one at which there is disease absence,

- ***Unemployment situations***
  An unemployment period is a busy period, ironically, and an idle period is a full employment period.

In the modeling of these problems it is also necessary to consider the properties of the transient probabilities. See Ferreira and Andrade (2010e).

6. More Applications

Statistical Queuing Theory is applied, for example, to intelligent transportation systems, call centres (see Ferreira and Andrade, 2010a), PABXs, telecommunications networks, advanced telecommunications systems and traffic flow. The networks of queues are used to reduce the waiting times in the hospitals. Another example of application of the networks of queues are the compartment models, in which infinite servers nodes are considered, important in Biology and in the study of hierarchical systems (Ferreira, 1987).

Agner Krarup Erlang, a Danish engineer who worked for the Copenhagen Telephone Exchange, published the first paper on queuing theory in 1909 (Erlang, 1909). The famous Erlang loss formula (Erlang, 1917)

$$p_m = \frac{\rho^m}{m!}\left(\sum_{i=0}^{m}\frac{\rho^i}{i!}\right)^{-1}$$

is the stationary probability that in the M/M/m/m queue the m servers are occupied (Ferreira and Andrade, 2010a). It is very much used in call-centres management to evaluate the probability of a call lost.

Leonard Kleinrock, in the early 1960s, performed an important work on queuing theory used in modern packet switching networks (Kleinrock, 1975, 1976).

7. Product Form Equilibrium Distribution

The first important result in the network of queues area was Jackson networks - an example of open networks - for which efficient product form equilibrium distribution exists (Jackson, 1957).

In a product form solution the equilibrium state probabilities are of the form

$$\pi(\pi_1, x_2, ..x_J) = C\pi_1(x_1)\pi_2(x_2)...\pi_J(x_J)$$

where C is a normalizing constant chosen to make equilibrium state probabilities sum to 1 and $\pi_i(.)$ represents the equilibrium distribution for queue $i, i = 1,2,..J$.

The BCMP (Baskett, Chandy, Muntz and Palacios, 1975) networks are a generalization of Jackson networks, considering several classes of customers.

For the Gordon-Newell networks-that are closed networks - product form equilibrium distribution also exists (Gordon and Newell, 1967).

References

[1] Andrade, M. (2010), "*A Note on Foundations of Probability"*. Journal of Mathematics and Technology, Vol. 1 (1), pp 96-98.

[2] Basket, F., Chandy, M., Muntz, R. and Palacios, J. (1975*), "Open, closed and mixed Networks of Queues with Different Classes of Customers",* Journal of ACM 22, pp 248-260.

[3] Cox, D. R. and Smith, W. L. (1961), "*Queues*". London: Methuen.

[4] Cox, D. R. and Miller, H. D. (1965), "*The Theory of Stochastic Processes*". London: Chapman and Hall.

[5] Disney, R. L. and König, D. (1985), *"Queueing Networks: a Survey of their Random Processes*". Siam Review 3, pp 335-403.

[6] Erlang, A. K. (1909), *"The Theory of Probabilities and Telephone Conversations"*. Nyt Tidsskrift for Mathematic B 20.

[7] Erlang, A. K. (1917), *"Solution of Some Problems in the Theory of Probabilities of Significance in Automatic Telephone Exchanges"*, Post Office Electrical Engineers' Journal 10, pp 189-197.

[8] Ferreira, M. A. M. (1987), *"Redes de filas de espera"*. Master thesis presented at IST-UTL.

[9] Ferreira, M. A. M. (2010), "*A Note on Jackson Networks Sojourn Times"*. Journal of Mathematics and Technology, Vol. 1, No 1, pp 91-95.

[10] Ferreira, M. A. M. and Andrade, M. (2009a), *"$M/G/\infty$ Queue System Parameters for a Particular Collection of Service Time Distributions"***.** AJMCSR-African Journal of Mathematics and Computer Science Research Vol. 2(7), pp 138-141.

[11] Ferreira, M. A. M. and Andrade, M. (2009b), *"The Ties Between the M/G/∞ Queue System Transient Behavior and the Busy Period"*. International Journal of Academic Research 1 Vol. (1), pp 84-92.

[12] Ferreira, M. A. M. and Andrade, M. (2010a), *"M/M/m/m Queue System Transient Behavior"*. Journal of Mathematics and Technology, Vol. 1 (1), pp 49-65.

[13] Ferreira, M. A. M. and Andrade, M. (2010b), *"Looking to a M/G/∞ System Occupation Through a Riccati Equation"*. Journal of Mathematics and Technology, Vol. 1 (2), pp 58-62.

[14] Ferreira, M. A. M. and Andrade, M. (2010c), *"M/G/∞ Queue Busy Period Tail"*. Journal of Mathematics and Technology, Vol. 1 (3), pp 11-16.

[15] Ferreira, M. A. M. and Andrade, M. (2010d), *"Algorithm for the Calculation of the Laplace-Stieltjes Transform of the Sojourn Time of a Customer in an Open Network of Queues with a Product Form Equilibrium Distribution, assuming Independent Sojourn Times in each Node"*. Journal of Mathematics and Technology, Vol. 1 (4), pp 31-36.

[16] Ferreira, M. A. M. and Andrade, M. (2010e), *"M/G/∞ System Transient Behavior with Time Origin at the Beginning of a Busy Period Mean and Variance"*. Aplimat- Journal of Applied Mathematics, Vol. 3 (3), pp 213-221.

[17] Ferreira, M. A. M. and Andrade, M. (2011a), *"Fundaments of Theory of Queues"*. International Journal of Academic Research, Vol. 3 (1), part II, pp 427-429.

[18] Ferreira, M. A. M. and Andrade, M. (2011b), *"Some Notes on the M/G/∞ Queue Busy Cycle Renewal Function"*. International Journal of Academic Research, Vol. 3 (6), I Part, pp179-182.

[19] Ferreira, M. A. M. and Andrade, M. (2011c), *"M/G/∞ Infinite Queue System Transient Behaviour with Time Origin at an Operation Beginning Instant and Occupation"*. Journal of Mathematics and Technology, Vol. 2 (1), pp 54-60.

[20] Ferreira, M. A. M. and Andrade, M. (2011d), *"Grouping and Reordering in a Servers Series"*. Journal of Mathematics and Technology, Vol. 2 (2), pp 4-8.

[21] Ferreira, M. A. M. and Andrade, M. (2011e), *"Non-homogeneous Networks of Queues"*. Journal of Mathematics and Technology, Vol. 2 (2), pp 24-29.

[22] Ferreira, M. A. M. and Andrade, M. (2011f), *"The M/GI/1 Queue with Instantaneous Bernoulli Feedback Stochastic Processes: A Review"*. Journal of Mathematics and Technology, Vol. 2 (3), pp 27-30.

[23] Ferreira, M. A. M. and Andrade, M. (2011g), *"The Fundamental Theorem in Queuing Networks"*. Journal of Mathematics and Technology, Vol. 2 (3), pp 48-53.

[24] Ferreira, M. A. M. and Andrade, M. (2011h), *"An infinite servers Nodes Network in the Study of a Pensions Fund"*. International Journal of Latest Trends in Finance and Economic Sciences, Vol. 1 (2), pp 91-94.

[25] Ferreira, M. A. M. and Filipe, J. A. (2010a), *"Solving Logistics Problems using M/G/∞ Queue Systems Busy Period"*. Aplimat- Journal of Applied Mathematics, Vol. 3 (3), pp 207-212.

[26] Ferreira, M. A. M. and Filipe, J. A. (2010b), *"Economic Crisis: Using M/G/oo Queue System Busy Period to Solve Logistics Problems in an Organization"*. *China-USA Business* Review, Vol. 9 (9), pp 59-63.

[27] Ferreira, M. A. M., Andrade, M. and Filipe, J. A. (2008), *"The Ricatti Equation in the M|G|∞ System Busy Cycle Study"*. Journal of Mathematics, Statistics and Allied Fields 2(1).

[28] Ferreira, M. A. M., Andrade, M. and Filipe, J. A. (2009), *"Networks of Queues with Infinite Servers in Each Node Applied to the Management of a Two Echelons Repair System"*. China-USA Business Review 8(8), pp 39-45 and 62.

[29] Ferreira, M. A. M., Filipe, J. A. and Coelho, M. (2008), *"A Queue Model for Motor Vehicles Dismantling and Recycling"*. Aplimat- Journal of Applied Mathematics, Vol. 1 (1), pp 337-344.

[30] Ferreira, M. A. M., Filipe, J. A. and Coelho, M. (2011), *"A Queue model- Recycling and Dismantling Motor Vehicles"*. International Journal of Latest Trends in Finance and Economic Sciences, Vol. 1 (3), pp 137-141.

[31] Ferreira, M. A. M., Andrade, M., Filipe, J. A. and Selvarasu, A. (2009), "*The Management of a Two Echelons Repair System Using Queuing Networks with Infinite Servers Queues*". Annamalai International Journal of Business and Research, Vol. 1, pp 132-137.

[32] Figueira, J. and Ferreira, M. A. M. (1999), *"Representation of a Pensions Fund by a Stochastic Network with Two Nodes: an Exercise"*. Portuguese Review of Financial Markets, Vol. 2, No 1, pp 75-81.

[33] Gordon, W. J. and Newell, G. F. (1967), *"Closed Queueing Systems with Exponential Servers"*. Operations Research 15, pp 254-265.

[34] Jackson, J. R. (1957), *"Networks of Waiting Lines"*. Operations Research 5, pp 518-521.

[35] Kelly, F. P. (1979), *"Reversibility and Stochastic Networks"*. New York: John Wiley and Sons.

[36] Kendall, D. G. (1953), *"The Analysis of Economic Time-Series-Part I: Prices"*. Journal of the Royal Statistical Society. A (General) 116(1), pp 11-34.

[37] Kleinrock, L. (1975), *"Queueing Systems"*, Vol. 1, Wiley, New York.

[38] Kleinrock, L. (1976), *"Queueing Systems"*, Vol. 2, Wiley, New York.

[39] Mathew, L. and Smith, D. (2006), *"Using Queuing Theory to Analyze Completion Times in Accident and Emergency Departments in the Light of the Government 4-hours Target"*. Cass Business School retrieved on 2008-05-20.

[40] Shanbhag, D. N. (1966), *"On Infinite Servers Queues with Batch Arrivals"*. Journal of Applied Probability, 3, pp 274-279.

[41] Syski, R. (1960), *"Introduction to Congestion Theory in Telephone Systems"*. Oliver and Boyd. London.

[42] Syski, R. (1986), *"Introduction to Congestion Theory in Telephone Systems"*. North Holland. Amsterdam.

[43] Takács, L. (1962), *"An Introduction to Queueing Theory"*. Oxford University Press". New York.

[44] Tijms, H. C. (2003), *"Algorithmic Analysis of Queues"*. Chapter 9 in "A First Course in Stochastic Models", Wiley, Chichester.

[45] Walrand, J. (1988), *"An Introduction to Queueing Networks*. New Jersey: Prentice-Hall, Inc.