AN INVESTIGATION AND HISTORICAL OVERVIEW OF THE G/M AND M/G QUEUEING PROCESSES

I N Fabris-Rotelli[1] and C Kraamwinkel*[1]

[1]Department of Statistics, University of Pretoria, 0002, Pretoria, South Africa,
inger.fabris-rotelli@up.ac.za

ABSTRACT

We present a historical and theoretical overview of the more complicated and less used G/M and M/G queueing processes, which allow for non-specific arrival and service time distributions. Such a model provides a more general setting for model fitting of real data for which the Markov property may not hold.

1. INTRODUCTION

In real-life situations we often have to deal with queues, such as service in a bank or supermarket, airplanes waiting for permission to approach a runway, websites receiving requests from internet users, payment requests generated from EFT points in a supermarket, patients waiting in a hospital emergency room and traffic waiting at a traffic light, stop sign or roundabout. A queueing theory models these real-life queueing situations so that the behaviour of the queue can be studied mathematically (Gross & Harris, 1985). The application of queueing theory is quite diverse and includes telecommunications (Daigle, 2005), traffic engineering (Drew, 1968), computing (Menascé et al., 2004), factory shops, offices and hospital design (Graves, 1982; Green et al., 2006; Saaty, 1961) and even emergency evacuation planning (Smith, 1991), with the aim of optimizing resources. From the customer's point of view we would like to minimize the waiting time. The service provider in turn would like to prevent customers from getting frustrated and would also like to make the best possible use of resources in order to minimize costs and maximise profit.

A queueing system consists of three distinct parts namely, the service facility being, the arrival of customers to be served as well as the service process. Some of the characteristics we are interested in are the queue input, the service mechanism, queue discipline (the order in which customers are served) and the number of queues in the system (Bocharov et al., 2004; Gross & Harris, 1985). It is generally assumed that arrivals and services are independent. In 1953, David G. Kendall (1953) suggested the first and now well-used notation, known as Kendall's notation, for describing the characteristics of a queueing model.

*I N Fabris-Rotelli is the corresponding author. C Kraamwinkel is the presenting author. Thanks to Department of Statistics and SARChI at the University of Pretoria for funding.

It was first suggested as a three-factor A/B/C notation where A represents the arrival process distribution, B the service time distribution and C the number of servers. This system was extended by Lee (1966) to include factors K and D and by Taha (1971) to include a factor N. Here factor K represents the capacity of the system (those in the queue as well as in service), N the calling population (from which the arrivals originate) and D the queue's discipline. Therefore we describe a queue as A/B/C/K/N/D or just A/B/C (if $K = \infty$, $N = \infty$ and $D = FIFO$ (First In First Out)). We will only look at interarrival and service times that are exponentially distributed (M) or have general distributions (G or GI). Although G usually refers to independent service or arrival times, some authors use GI to be more explicit.

Although the first paper on queueing theory, 'Waiting times and number of calls' by Johannsen (Bhat, 1969), was published in 1907, it is generally accepted that queueing theory was invented by A.K. Erlang (Brockmeyer et al., 1948), a Danish mathematician, statistician and engineer who joined the Copenhagen Telephone Company as scientific collaborator and head of the newly established physico-technical laboratory in 1908[1]. He published his first work in 1909 in which he proved that telephone calls distributed at random followed a Poisson distribution (Erlang, 1909). Erlang (1917) is his most important work. This article contained Erlang's formulae for loss and waiting time developed on the basis of the statistical equilibrium principle. Erlang went on to develop and publish many more works on the theory of telephone traffic before his early death in 1929 at the age of 51. Almost all his works were first published in Danish but the most important were later translated into English, French and German. (Brockmeyer et al., 1948)

Erlang's work and the work done by others (Fry, 1928) in the early thirties was motivated by practical problems. One of the greatest contributors at this time, Aleksandr Khinchin, referred to queueing theory as mass-service theory. His interest in queueing theory was a result of his connection with workers of the Moscow telephone exchange. He was particularly interested in the general study of incoming calls (Khinchin, 1932, 1933). He had a great influence in the development of probability theory by investigating stationary stochastic processes and the formulations of the foundational theory leading to the application in various fields of natural science including statistical physics, queueing problems and information theory. The general study of incoming calls was also the subject of Khinchin's final monograph (Khinchin, 1955) and his last mathematical papers , where he gave the probability of k events in an interval of length t given an event at the start of the interval. (Cramér, 1962; Doob, 1961; Gnedenko, 1961)

In the following two decades, many theoreticians became interested in these and more general models to be used in more complex situations. Unfortunately this led to a wide gap between the practical and theoretical developments in the field (Bhat, 1968). The first solutions of the time-inhomogeneous problem were given by Ledermann & Reuter (1954) using spectral theory and by Bailey (1954) using generating functions. Laplace transforms have also been used on this problem but even though this is a useful technique and powerful analytically, the mathematical manipulation becomes technical. Kendall (1964) remarked that much of the detail of the queue-theoretic scene had been obscured by the Laplacian curtain.

[1]This paper was not mathematically exact and therefore Erlang's first paper on the subject is seen as historically more important.

A result of this was that many researchers took the easiest way out of a complex situation by assuming steady-state from the start or by taking both interarrival and service times as exponentially distributed without significant information loss. This situation is not adequate for modeling most real-world situations (Bhat, 1969). Other methods of dealing with the loss of the Markov property include the method of supplementary variables (Kendall, 1964; Wishart, 1961), Kendall's regeneration point technique (Kendall, 1951, 1964) and approximation methods, each with their own merits and difficulties. A comprehensive bibliography on queueing theory up to 1957 can be found in Doig (1957).

Examples of articles dealing with communications applications are Keshavamurthy & Chandra (2006), Shankar (2007) and Iftikhar et al. (2008). The theory can also be applied to insurance problems (Boxma et al. 2011b, Postan & Balobanov, 2011, Löpker & Perry, 2010). The active research field of traffic planning is discussed in Cheah & Smith (1994), Jain & Smith (1997), Vandaele et al. (2000), van Woensel & Vandaele (2006) and van Woensel et al. (2006). Also recently employed is the development of a planning model for manpower allocation to after-sales field support (Tang et al., 2008).

2. THEORY: M/G AND G/M MODELS

We no longer have a Markov process in these models but within lies an imbedded Markov chain, allowing for the use of some Markov chain theory. The results and some of the proofs for M/G and G/M queues can be found in Gross & Harris (1985); Kleinrock (1975); Giffin (1978). Let $E(N)$ denote the expected number of customers in the system, $E(W)$ ($E(W+S)$) the expected waiting time of a customer in the queue(system) and p_n the steady-state probability of n customers in the system at any point in time.

M/G Models In the M/G/1 queue, we assume a single server process with arrivals determined by a Poisson process with parameter λ and service times independently generally distributed with mean μ and variance σ_s^2. We look at the system at the points in time when a customer departs the system therefore the next customer enters service at that instant and remaining service time will not depend on the length of time already in service. The process therefore only depends on the number of customers in the system at these time points. This is known as the regeneration point technique. However in the M/G/s queue, at a service completion a customer leaves the system, but other servers will still be busy serving customers so that the remaining service times need to be considered. This can be eliminated by only looking at the size of the queue instead of system size. Let $w_q(t)$ denote the density function of the waiting time distribution in the queue, $Q(t)$ denote the number of customers in the system at time t, $Y(t)$ the number of service completions up to time t, N_q the steady-state number of customers in the queue at departure points, W_q the waiting time of a customer in the queue, π_n the steady-state probability of n customers in the system at a departure point and $\pi_n^q = P(n \text{ in the queue just after a departure})$. We note that, for M/G models, $\pi_n = p_n$. The main results for the **M/G/1 queue** are $E(N) = \rho + \frac{\rho^2+\lambda^2\sigma_S^2}{2(1-\rho)}, \rho = \frac{\lambda}{\mu}$; $E(W+S) = \frac{1}{\lambda}\left(\rho + \frac{\rho^2+\lambda^2\sigma_S^2}{2(1-\rho)}\right)$; $E(W) = \frac{\lambda(\mu^{-2}+\sigma_S^2)}{2(1-\rho)}$; $E(\text{no. cust waiting for service}) = \frac{\rho^2+\lambda^2\sigma_S^2}{2(1-\rho)}$; $\pi_i = \pi_0 k_i + \sum_{j=1}^{i+1} \pi_j k_{i-j+1},\ i = 0, 1, 2, \ldots$ where $k_n = P(n \text{ arri. during a service } S = t)$; and $E(\text{length of the busy period}) = \frac{1}{\mu-\lambda}$. For the **M/G/s**

queue we have that $E\left(N_q\right) = \lambda E\left(W_q\right)$, $\pi_n^q = \int_0^\infty \frac{e^{-\lambda t}(\lambda t)^n}{n!} w_q(t)dt$, and $E(N_q(N_q-1)\cdots(N_q-k+1))$ $= \lambda^k \int_0^\infty t^k w_q(t)dt = \lambda^k E\left(W_q^k\right)$. In the case of the **M/G/s/s queue** we have that $p_n = \frac{e^{-\rho}\rho^n}{n!}$, $\rho = \frac{\lambda}{\mu}$ and for the **M/G/∞ queue** we have $P(Q(t) = n) = \frac{e^{-\lambda tq}(\lambda tq)^n}{n!}$, $p_n = \frac{e^{-\rho}\rho^n}{n!}$, $\rho = \frac{\lambda}{\mu}$ and $P(Y(t) = n) = \frac{e^{-\lambda t(1-q)}(\lambda t(1-q))^n}{n!}$ where $q = \int_0^t P(\text{service time exceeds } t-x|\text{arrival at time } x) \times P(\text{arrival at time } x)dx$.

G/M Models In the G/M/1 queue, we have a single server with interarrival times independently generally distributed with mean e λ and variance σ_a^2, and service times exponentially distributed with parameter μ. We follow the same regeneration point technique approach as above but instead look at time points just before a customer arrives. Generalising from 1 to s servers, everything remains the same except for b_n, the probability of n service completions during an interarrival time. The mean service rate per unit of time t will change from μ to $n\mu$ or $s\mu$ depending on the state of the system, therefore b_n will depend on the state of the system. We note that, $p_n \neq q_n$ where q_n is the steady-state probability of having n customers in the system at an arrival point. The main results for the **G/M/1 queue** are given by $q_n = (1-r_0)r_0^n$ for $n \geq 0, \rho < 1$ where r_0, $0 < r_0 < 1$, is the only root of the function $\beta(z) = z$ where $\beta(z) = \sum_{n=0}^\infty b_n z^n$ and $b_n = P(n\text{service completions during an interarrival time})$ and $W(t) = \begin{cases} 1 - r_0 e^{-\mu(1-r_0)t}, & t \geq 0 \end{cases}$. For the **G/M/s queue** we can find the limiting probabilities by solving for q_n numerically in $C = \frac{1-\sum_{n=0}^{s-1} q_n}{r_0^s(1-r_0)^{-1}}$. We can also find $W(t) = C\left(\frac{1}{C} - \frac{r_0^s}{1-r_0}e^{-\mu s(1-r_0)t}\right)$ with r_0 defined as in the G/M/1 case.

Further research into M/G/1 and G/M/1 queues has also been done. See Adan & Haviv (2009); Bae & Kim (2010); Boxma et al. (2011a, 2009, 2010); Haviv & Kerner (2011); Kahraman & Gosavi (2011); Taylor & van Houdt (2010); Wang & Huang (2009).

3. MODEL SELECTION

Arguably, the most important step in modelling using queueing theory is selecting the most appropriate distributions for the services and arrivals in the model. To do this we need to know as much as possible about the characteristics of potential distributions and the situation being modelled. We consider the service process as an example. The exponential distribution may be appropriate in a case where there is a wide variation in service required from customer to customer but it is clearly not appropriate in the case where a customer's remaining service time depends on his expended service time.

A good starting point for selecting an appropriate distribution is to make use of graphical methods. One possible option is to make use of probability plots. Probability plots compare ordered values of a variable with percentiles of a specified theoretical distribution. If the data distribution matches the theoretical distribution, the points on the plot will form a linear pattern. We should also plot the interarrival and interservice times to see whether there are non-random effects present in the data. A formal method of determining this involves the sampling distribution of 'runs' (Duncan, 1986). Once we have decided on an appropriate distribution, we can then use a test such as the χ^2 goodness-of-fit test to see if the chosen distribution fits our situation. This is however not the only available test. Other tests include the F-test, the

Kolmogorov-Smirnov test, the Anderson-Darling test and the Cramér-von Mises criterion.

Once we know which distribution fits our situation, it is important to know whether the process is time-homogeneous or not. If the process is not, we will need different arrival or service rates for different times. An example of this would be traffic on a highway. During peak hours we would expect more arrivals per time unit than during off-peak hours. One possible test for this is Bartlett's test (Bartlett, 1934, 1937; Epstein, 1960a,b).

It is important to realise it is not realistic to expect a chosen distribution to be in exact agreement with a real-life situation. A system that exactly models the situation is bound to be overly complex and it may not be possible to gain usable insight into the process. We therefore need to find a distribution that reasonably approximates the situation we are looking at. The exponential distribution is often used to model the service system since it is a conservative choice. When examining the expected waiting time for alternative choices for the service distribution, the exponential distribution will always lead to a longer waiting time as long as the coefficient of variation of the other distribution is less than one. This includes the family of gamma distributions which is often used to model the service process. If the exponential distribution is used in these circumstances, even if the service process is not exponential, we will be on the safe side when predicting the number of customers waiting to be served and their waiting times. (Giffin, 1978)

4. CONCLUSION

We have provided a historical and theoretical overview of the lesser known G/M and M/G queues allowing for general arrival and service distributions, instead of imposing a specified distribution. In addition model fitting in discussed for use in practice.

References

Adan, I. & Haviv, M. (2009). Conditional ages and residual service times in the M/G/1 queue. *Stoch Models*, 25(1), 110–128.

Bae, J. & Kim, S. (2010). The stationary workload of the G/M/1 queue with impatient customers. *Queueing Syst*, 64(3), 253–265.

Bailey, N. T. J. (1954). A continuous time treatment of a simple queue using generating functions. *J Roy Stat Soc B Met*, 16(2), 288–291.

Bartlett, M. S. (1934). The problem in statistics of testing several variances. *Math Proc Cambridge*, 30(02), 164–169.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proc R Soc Lon Ser-A*, 160(901), 268–282.

Bhat, U. N. (1968). *A Study of the Queueing Systems M/G/1 and GI/M/1*. Springer-Verlag.

Bhat, U. N. (1969). Sixty years of queueing theory. *Manage Sci*, 15(6), B280–B292.

Bocharov, P. P., D'Apice, C., Pechinkin, A. V., & Salerno, A. (2004). *Queueing Theory*. Walter de Gruyler.

Boxma, O., Perry, D., & Stadje, W. (2011a). The M/G/1+G queue revisited. *Queueing Syst*, 67(3), 207–220.

Boxma, O., Perry, D., Stadje, W., & Zacks, S. (2009). The M/G/1 queue with quasi-restricted accessibility. *Stoch Models*, 25(1), 151–196.

Boxma, O., Perry, D., Stadje, W., & Zacks, S. (2010). The busy period of an M/G/1 queue with customer impatience. *J Appl Probab*, 47(1), 130–145.

Boxma, O. J., Löpker, A., & Perry, D. (2011b). Threshold strategies for risk processes and their relation to queueing theory. *J Appl Probab*, 48A, 29–38.

Brockmeyer, E., Halstrøm, H. L., & Jensen, A. (1948). *The Life and Works of A.K. Erlang*. Academy of Technical Science.

Cheah, J. Y. & Smith, J. M. (1994). Generalizes M/G/C/C state dependent queueing models and pedestrian traffic flows. *Queueing Syst*, 15, 365–386.

Cramér, H. (1962). A. I. Khinchin's work in mathematical probability. *Ann Math Stat*, 33(4), 1227–1237.

Daigle, J. (2005). *Queueing Theory with Applications to Packet Telecommunication*. Springer.

Doig, A. (1957). A bibliography on the theory of queues. *Biometrika Trust*, 44(3/4), 490–514.

Doob, J. L. (1961). Appreciation of Khinchin. In J. Neyman (Ed.), *Proc 4th Berkeley Symp Math Statist and Prob*, volume II (pp. 17–20).: University of California Press.

Drew, D. (1968). *Traffic Flow Theory and Control*. McGraw-Hill Book Company, Inc.

Duncan, A. J. (1986). *Quality control and industrial statistics*. Homewood, Ill : Irwin, 5th edition.

Epstein, B. (1960a). Tests for the validity of the assumption that the underlying distribution of life is exponential: Part I. *Technometrics*, 2(1), 83–101.

Epstein, B. (1960b). Tests for the validity of the assumption that the underlying distribution of life is exponential: Part II. *Technometrics*, 2(2), 167–183.

Erlang, A. K. (1909). Sandsynlighedsregning og telefonsamtaler. *Nyt Tidsskrift for Matematik B*, 20, 33.

Erlang, A. K. (1917). Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *Post Office Elec Eng*, 10, 189.

Fry, T. C. (1928). *Probability and its Engineering Uses*, chapter 10, (pp. 321–388). D. van Nostrand Company, Inc., 1 edition.

Giffin, W. C. (1978). *Queueing: Basic Theory and Applications*. Grid Inc.

Gnedenko, B. V. (1961). Alexander Iacovlevich Khinchin. In J. Neyman (Ed.), *Proc 4th Berkeley Symp Math Statist and Prob*, volume II (pp. 1–15).: University of California Press.

Graves, S. C. (1982). The application of queueing theory to continuous perishable inventory systems. *Manage Sci*, 28(4), 400–406.

Green, L. V., Soares, J., Giglio, J. F., & Green, R. A. (2006). Using queueing theory to increase the effectiveness of emergency department provider staffing. *Acad Emerg Med*, 13(1), 61–68.

Gross, D. & Harris, C. M. (1985). *Fundamentals of Queueing Theory*. John Wiley & Sons, 2nd edition.

Haviv, M. & Kerner, Y. (2011). The age of the arrival process in the M/G/1 and G/M/1 queues. *Math Method Oper Res*, 73(1), 139–152.

Iftikhar, M., Singh, T., Landfeldt, B., & Caglar, M. (2008). Multiclass G/M/1 queueing system with self-similar input and non-preemptive priority. *Comput Commun*, 31(5), 1012 – 1027.

Jain, R. & Smith, J. M. (1997). Modeling vehicular traffic flow using M/G/C/C state dependent queueing models. *Transport Sci*, 31(4), 324–336.

Kahraman, A. & Gosavi, A. (2011). On the distribution of the number stranded in bulk-arrival, bulk-service queues of the M/G/1 form. *Eur J Oper Res*, 212(2), 352–360.

Kendall, D. (1951). Some problems in the theory of queues. *J Roy Stat Soc B Met*, 13(2), 151–185.

Kendall, D. G. (1953). Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. *Ann Math Stat*, 24(3), 338–354.

Kendall, D. G. (1964). Some recent work and further problems in the theory of queues. *Theor Probab Appl+*, 9(1), 1–13.

Keshavamurthy, S. & Chandra, K. (2006). Multiplexing analysis for dynamic spectrum access. In *IEEE MILCOM* (pp. 1–7).

Khinchin, A. A. (1932). Mathematical theory of a stationary queue. *Mat Sb*, 39(4), 73–84.

Khinchin, A. I. (1933). On the mean time of non-attendance of stations. *Mat Sb*, 40, 119–123.

Khinchin, A. I. (1955). Mathematical methods in the theory of mass service. *Trudy Mat Inst Akad Nauk*, 49, 1–123.

Kleinrock, L. (1975). *Queueing Systems*, volume 1: Theory. John Wiley & Sons.

Ledermann, W. & Reuter, G. E. H. (1954). Spectral theory for the differential equations of simple birth and death processes. *Philos Tr R Soc S-A*, 246(914), 321–369.

Lee, A. M. (1966). *Applied Queueing Theory*. Macmillan.

Löpker, A. & Perry, D. (2010). The idle period of the finite G/M/1 queue with an interpretation in risk theory. *Queueing Syst*, 64, 395–407.

Menascé, D., Almeida, V., & Dowdy, L. (2004). *Performance by Design: Computer Capacity Planning by Example*. Prentice Hall.

Postan, M. Y. & Balobanov, O. O. (2011). Method of evaluation of insurance expediency of stevedoring company's responsibility for cargo safety. *Int J on Mar Nav and Safety of Sea Transport.*, 5(4), 479–482.

Saaty, T. (1961). *Elements of Queueing Theory with Applications*. McGraw-Hill Book Company, Inc.

Shankar, S. N. (2007). Squeezing the most out of cognitive radio: A joint MAC/PHY perspective. In *ICASSP 2007*, volume 4 (pp. 1361–1364).

Smith, J. M. (1991). State-dependent queueing models in emergency evacuation networks. *Transport Res B-Meth*, 25(6), 373 – 389.

Taha, H. A. (1971). *Operations Research: An Introduction*. Collier-Mac.

Tang, Q., Wilson, G. R., & Perevalov, E. (2008). An approximation manpower planning model for after-sales field service support. *Comput Oper Res*, 35(11), 3479 – 3488.

Taylor, P. G. & van Houdt, B. (2010). On the dual relationship between Markov chains of GI/M/1 and M/G/1 type. *Adv Appl Probab*, 42(1), 210–225.

van Woensel, T. & Vandaele, N. (2006). Empirical validation of a queueing approach to uninterrupted traffic flows. *4OR-Q J OPER RES*, 4(1), 59–72.

van Woensel, T., Wuyts, B., & Vandaele, N. (2006). Validating state-dependent queueing models for uninterrupted traffic flows using simulation. *4OR-Q J OPER RES*, 4(2), 159–174.

Vandaele, N., van Woensel, T., & Verbruggen, A. (2000). A queueing based traffic flow model. *Transport Res D-Tr E*, 5(2), 121 – 135.

Wang, K. H. & Huang, K. B. (2009). A maximum entropy approach for the <p,N>-policy M/G/1 queue with a removable and unreliable server. *Appl Math Model*, 33(4), 2024–2034.

Wishart, D. M. G. (1961). An application of ergodic theorems in the theory of queues. In J. Neyman (Ed.), *Proc 4th Berkeley Symp Math Statist and Prob*, volume II (pp. 581–592).: University of California Press.