eqr366

Queues in Reliability

M. Concepcion Ausin

Department of Mathematics

Facultade de Matemáticas, Campus de Elviña

Universidade da Coruña

15071 A Coruña, Spain

Phone: +34 981 167 000 (ext. 1318)

Fax: + 34 981 167 160

mausin@udc.es*

February 21, 2007

Keywords: Finite source; machine interference; machine repair; maintenance; manufacturing; multi-programmed computers; multi-echelon inventory; retrial; spares; vacation.

[1]Text file name: eqr366.tex, prepared using LaTeX; Figure file name: eqr366_1.eps

Abstract

Queueing models can be useful in solving many complex reliability problems. Component failures are usually interpreted as the arrival of customers and the repair or replacement of failed components is typically associated with the service facility. A distinctive characteristic of queues in reliability is that requests for service are usually generated by a finite customer population because, in general, there are a limited number of units, e.g. machines which can fail, and when they are all in the system, being repaired or waiting for repair, no more can arrive. Thus the arrivals do not form a renewal process as they may depend on the number of units in the system. This is an essential difference from typical queueing systems, where the population of potential arrivals can be considered to be effectively limitless. This article overviews the main queueing models used in reliability which are illustrated using the classical machine repairmen model. Some statistical methods to estimate the main quantities of interest in a queue are also discussed.

1 Introduction

A queueing system is a mathematical model to characterize situations where certain units, called *customers*, arrive in continuous time in order to receive a service or facility provided by other units, called *servers*. When customers arrive, they immediately start to be served if there is an empty server. Otherwise, the customers must spend some time waiting in line for service until a server becomes available. Queueing theory is the methodology devoted to the stochastic description of queueing systems and the study of their performance measures,

such as the queue length, waiting times, etc. Some essential references on queueing theory are [1, 2, 3, 4, 5].

Queueing theory can be very useful in solving many complex reliability problems. Component failures or machine breakdowns can be treated as arrivals of customers and the repair or replacement of failed components may be seen as the service facility in a queueing system. The number of repair facilities or maintenance crews is equivalent to the number of servers. It is useful to be able to find the analogies between reliability and queueing theory because it is frequently found in practice that a given reliability problem has been previously studied with an equivalent problem in queueing theory [6]. For example, the number of failed components corresponds to the *queue length* and the time elapsed from a failure to repair is equivalent to the waiting time in the system or *sojourn time*. Note that the notion of repair may also refer to preventive maintenance activities performed before system failures. For example, a deteriorated machine can be repaired to an upgraded status before it fails.

A queueing model is completely described by six characteristics that, following Kendall's notation [7] are given by $A/B/c/K/m/R$, where A and B describe the arrival and service pattern, respectively, c the number of servers, K the system capacity or maximum number of customers allowed in the system, m the size of the customer population and R the discipline or the order in which customers are selected for service. Different symbols are traditionally used for the type of distribution of A and B, for example, M denotes exponential (Markovian) random variables, D degenerate (constant) variables and G general distributions. For instance, the $M/G/1$ queue denotes a single-server system with exponential inter-arrival

times and general service time distribution. Note that when nothing is said about K, m and R, it is assumed that there is an infinite capacity, an infinite customer population and a first-come-first-served discipline.

In the case of reliability, the main characteristic of queues is that requests for service are usually generated by a finite customer population, that is, $m < \infty$. This is because, in general, there is a limited number of units (machines, devices, etc.) which can fail, and when they are all in the system (being repaired or waiting for repair), no more can arrive. Then, the arrivals of requests do not form a renewal process as the arrivals may depend on the number of units staying at the system. This is an essential difference from typical queueing systems, where the population of potential arrivals can be considered to be effectively limitless. For example, in telephone operations the number of customers is usually large enough that the probability of having all of them wanting service at once is extremely small. However, in repair of machinery, the whole set of machines may be simultaneously out of order. Queueing systems with finite population are known as *finite source queues*, which have been extensively studied in the queueing literature, see e.g. [8] and the detailed bibliography provided in [9]. Finite source queueing models are used to describe the classical *machine repairmen problem* which is described in the following section.

2 The machine repairmen problem

Consider a repairable system (see eqr348) consisting of r machines which operate independently of each other and may eventually fail. When a machine breaks down, it has to be

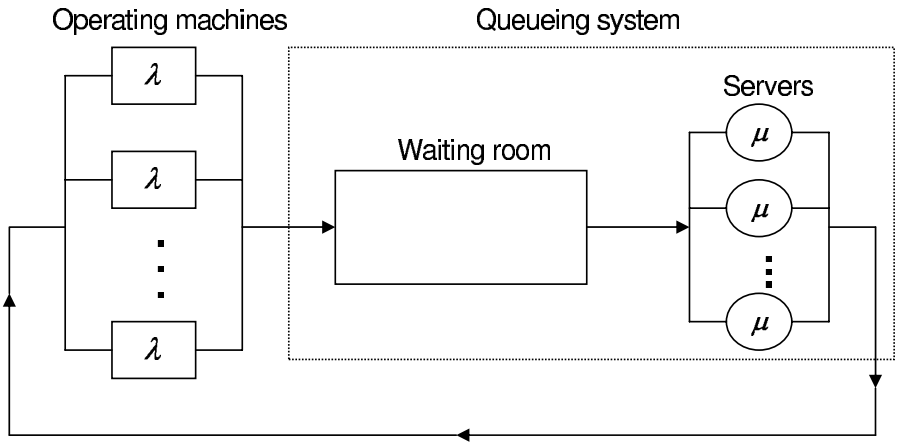

Figure 1: Machine repairmen/interference problem

repaired by one of the c repairmen, if there are any available. Otherwise, it has to wait for service until a server becomes available. Thus, the operating machines are outside the queueing system and enter the system only when they break down and require repair, as shown in Figure 1. Machine lifetimes and repair times are independent random variables with rates λ and μ, respectively. Then, the so-called mean time between failures (MTBF) and mean time to repair (MTTR) are given by $1/\lambda$ and $1/\mu$, respectively.

This model is known as the machine repairmen problem or the *machine interference problem* since the machines interfere with each other when all the servers are busy and new requests for service arrive. It can be used to describe a large variety of real situations such as manufacturing systems (see e.g. [10]), telecommunication traffic (see e.g. [11]), computer systems (see e.g. [12]), inventory models (see e.g. [13]), etc. This broad range of applications has motivated a large amount of research on this problem in the literature. For two reviews, see [14, 15].

The simplest machine repairmen model supposes that both lifetimes and service times are exponentially distributed, there is no limit in the waiting room, the repair is carried out by a first-in-first-out discipline and after having been served, each machine restarts working the same as new. The queueing model describing this system is denoted by $M/M/c/r/r$ or, equivalently, $M/M/c//r$. Note that there is an important difference from the notation used for the arrival process in infinite source queues, such as the $M/M/c$ queue, where the symbol M indicates a Poisson arrival process. However, in the $M/M/c/r/r$ queue, the time between failures of each machine is exponential but the arrival of failures from the whole set of machines does not follow a Poisson process because, as commented before, these arrivals depend on the number of machines waiting for repair. In particular, when the number of machines waiting for service increases, the arrival rate of further service demands decreases. More specifically, when n machines are "down" for repair, then, there are $(r-n)$ operating machines and the time until the next break down is the minimum of $(r-n)$ exponential distributions with rate λ, which is also an exponential distribution of rate $(r-n)\,\lambda$. Then, given that there are n machines being repaired or waiting for repair, the arrival rate or average number of fails per unit time is given by,

$$\lambda_n = \begin{cases} (r-n)\,\lambda, & \text{if } 0 \leq n \leq r, \\ 0, & \text{if } n \geq r, \end{cases} \tag{1}$$

and the average number of repaired machines per unit time is given by,

$$\mu_n = \begin{cases} n\mu, & \text{if } 0 \leq n \leq c, \\ c\mu, & \text{if } n \geq c. \end{cases} \tag{2}$$

Thus, it can be shown that the number of machines in the $M/M/c/r/r$ system follows a birth-death process where an arrival of a failure is interpreted as a birth and the completion of a repair is interpreted as a death, see e.g. [3]. Then, the long-run probability of having n machines in the system can be shown to be,

$$p_n = \begin{cases} \binom{r}{n} \left(\frac{\lambda}{\mu}\right)^n p_0, & \text{if } 1 \leq n \leq c, \\ \binom{r}{n} \frac{n!}{c^{n-c} c!} \left(\frac{\lambda}{\mu}\right)^n p_0, & \text{if } c \leq n \leq r, \end{cases} \tag{3}$$

where p_0 is obtained from $p_0 + ... + p_r = 1$. Using these probabilities, we can calculate for example the mean number of machines in the system (being repaired or waiting for repair), L, and the mean number of machines waiting in the queue for repair, L_q, using,

$$L = \sum_{n=1}^{r} n p_n \quad \text{and} \quad L_q = \sum_{n=c+1}^{r} (n-c)\, p_n. \tag{4}$$

As noted in (1), the arrival rate given that the system is in state n is $(r-n)\,\lambda$. Then, the unconditional average rate at which machines arrive to the system (also called the effective mean arrival rate) is given by,

$$\lambda_{\text{eff}} = \sum_{n=0}^{r-1} (r-n)\,\lambda p_n = \lambda\,(r - L)\,.$$

Thus, we can also obtain the mean waiting time, W, a machine spends in the system (from failure to recovery) and the mean queueing time, W_q, a machine waits in the queue for repair using the well known Little's formulae,

$$L = \lambda_{\text{eff}} W \quad \text{and} \quad L_q = \lambda_{\text{eff}} W_q. \tag{5}$$

Moreover, it is also possible to derive the distribution functions of the waiting time and queueing time in addition to their mean values, W and W_q, see e.g. [2, 3]. Some other

performance measures such as the probability that a broken machine must wait for repair, can also be obtained. Finally, it is worth observing that these results also hold even though the lifetimes are not exponentially distributed. That is, equation (3) is also valid for the $G/M/c/r/r$ queueing system. Furthermore, equation (3) also holds for systems with exponential lifetimes, general service times and the same number of repairmen as machines, i.e. for the $M/G/c/c/c$ queue, see [3].

These results have also been extended to more complicated situations where the lifetimes and repair durations follow more general distributions than the exponential one, such as the Erlang distribution, mixtures of exponentials, Coxian and phase-type distributions (see e.g. [16] and [17]). Many other features have been introduced to extend the applicability of the machine repairmen problem including, for example, *vacation* models where the servers can take a vacation of random duration when the system is empty (see e.g. [18]), *retrial* models where machines that find all the servers busy do not enter the queue but instead reapply for service at random intervals (see e.g. [19]), unreliable servers which may themselves breakdown (see e.g. [20]), *priority* models where a group of machines are served with priority over the others (see e.g. [21]), *k-out-of-r* systems where at least k machines must be functioning for the whole system to work properly (see e.g. [22]), heterogeneous failure modes (see e.g. [23]), etc. Another important extension is the use of *spares*, also called stand-by or sparing redundancy (see eqr349), which is introduced in the following section.

2.1 The machine repairmen problem with spares

Assume now that in addition to the r working machines considered previously, there is another set of s machines standing by as spares such that, when a working machine fails, it is immediately substituted by a spare machine if any is available. When a machine is repaired, it becomes a spare unless the number of working machines is less than r, in which case the repaired machine is restarted immediately.

Consider a $M/M/c/r/r$ with s spares where the lifetimes and repair times are exponentially distributed with rates λ and μ, respectively. Then, given that there are n machines being repaired or waiting for repair, the arrival rate is now given by,

$$\lambda_n = \begin{cases} r\lambda, & \text{if } 0 \leq n \leq s, \\ (r+s-n)\,\lambda, & \text{if } s \leq n \leq s+r, \\ 0, & \text{if } n \geq s+r, \end{cases} \tag{6}$$

and μ_n is the same as given in (2). Thus, this model can also be described by a birth-death process [3]. And then, it can be shown that the probability of having n machines in the system when $c \leq s$ is given by,

$$p_n = \begin{cases} \frac{r^n}{n!}\left(\frac{\lambda}{\mu}\right)^n p_0, & \text{if } 1 \leq n \leq c, \\ \frac{r^n}{c^{n-c}c!}\left(\frac{\lambda}{\mu}\right)^n p_0, & \text{if } c \leq n \leq s, \\ \frac{r^s r!}{(r+s-n)!c^{n-c}c!}\left(\frac{\lambda}{\mu}\right)^n p_0, & \text{if } s \leq n \leq s+r, \end{cases} \tag{7}$$

and when $c > s$,

$$p_n = \begin{cases} \frac{r^n}{n!}\left(\frac{\lambda}{\mu}\right)^n p_0, & \text{if } 1 \leq n \leq s, \\ \frac{r^s r!}{(r+s-n)!n!}\left(\frac{\lambda}{\mu}\right)^n p_0, & \text{if } s \leq n \leq c, \\ \frac{r^s r!}{(r+s-n)!c^{n-c}c!}\left(\frac{\lambda}{\mu}\right)^n p_0, & \text{if } c \leq n \leq s+r. \end{cases}$$

Using these probabilities, we can obtain the mean number of machines in the system and in the queue, L and L_q, respectively, as in (4). Also, we can obtain the effective mean arrival rate to the system by,

$$\lambda_{\text{eff}} = \sum_{n=0}^{s-1} r\lambda p_n + \sum_{n=s}^{s+r} (r+s-n)\,\lambda p_n = \lambda\left(r - \sum_{n=s}^{s+r} (n-s)\,p_n\right),$$

which allows us to obtain the mean waiting time in the system and the mean queueing time, W and W_q, respectively, using the Little's formula given in (5).

An interesting particular case appears when the number of spares, s, is very large. Then, the arrival rate λ_n given in (6) will be essentially constant and equal to λr, so that the arrival process will be well approximated by a Poisson process of rate λr. Thus, the system will converge to an infinite source $M/M/c$ model whose stationary distribution can be obtained as the limit of (7) when s goes to infinity [2, 3], that is,

$$p_n = \begin{cases} \frac{(\lambda r)^n}{n!\mu^n} p_0, & \text{if } n \leq c, \\ \frac{(\lambda r)^n}{c^{n-c} c!\mu^n} p_0, & \text{if } n \geq c. \end{cases}$$

As shown in this case, infinite source queues are in general much simpler than finite source queues and consequently, queueing theory on infinite systems is much more developed. Thus, in practice, it may be convenient to assume an infinite source if the customer population is finite but large. Note that this will be a good approximation when the arrival process depends only in a negligible way on the number of customers already in the system.

3 Inference for queues

Failure and repair times are random variables and the distributions and parameters describing their behavior are unknown in practice. Thus, the use of statistical methods (see eqr356) becomes necessary in order to estimate the quantities of interest in the system. Firstly, an experiment to collect data from the system must be designed. If we assume independence between the lifetimes and repair durations, a simple experiment providing complete information about the system consists in observing independently n lifetimes $\{x_1, ..., x_n\}$ and m repair times $\{y_1, ..., y_m\}$. Using these data, the traditional approach to queueing inference is based on maximum likelihood estimation of the failure and service parameters. For example, the likelihood function of the parameters in the $M/M/c/r/r$ queue is given by,

$$l(\lambda, \mu) \propto \lambda^n \exp\left(-\sum_{i=1}^{n} x_i\right) \mu^n \exp\left(-\sum_{i=1}^{m} y_i\right), \tag{8}$$

and then, the maximum likelihood estimators of λ and μ are given by $\hat{\lambda} = \bar{x}^{-1}$ and $\hat{\mu} = \bar{y}^{-1}$, respectively, where $\bar{x}$ and $\bar{y}$ denote the means of the lifetimes and repair times, respectively. These estimators can be plugged in to obtain estimations of the performance measures. For example, the stationary probabilities, p_n, can be estimated by replacing λ and μ by $\hat{\lambda}$ and $\hat{\mu}$, respectively, in (3). However, using this classical approach, it is often difficult to obtain measures of the uncertainty in the estimated performance measures.

An alternative approach is the use of the Bayesian methodology (see eqr364) which is specially well suited for queues [24]. Consider again the $M/M/c/r/r$ model and assume independent gamma prior distributions for λ and μ; say $\lambda \sim G(\alpha_0, \beta_0)$ and $\mu \sim G(a_0, b_0)$. From (8), it is easy to see that the posterior distributions are also independent gamma

distributions, $\lambda|data \sim G(\alpha, \beta)$ and $\mu|data \sim G(a, b)$, where $\alpha = \alpha_0 + n$; $\beta = \beta_0 + \sum_{i=1}^{n} x_i$; $a = a_0 + m$; $b = b_0 + \sum_{j=1}^{m} y_j$. Using these posterior distributions, we can approximate the posterior predictive distributions of any performance measure. For example, we can simulate K values $\{\lambda^{(k)}, \mu^{(k)}\}_{k=1}^{K}$ from the joint posterior distribution of (λ, μ), and approximate the posterior mean of p_n with,

$$E\left[p_n \mid data\right] \approx \frac{1}{K} \sum_{k=1}^{K} p_n^{(k)},$$

where $p_n^{(k)}$ is the value of p_n for $\lambda = \lambda^{(k)}$ and $\mu = \mu^{(k)}$. Analogously, we can easily approximate the posterior variance, percentiles and credible intervals of p_n. For further details of Bayesian methods for the $M/M/c/r/r$ model, see [25]. Also, Bayesian analysis of more general queues, where the exponential assumption for arrivals and services is relaxed, can be found in e.g. [26] and the references therein.

References

[1] Kleinrock L. Queueing Systems. Wiley: New York, 1976.

[2] Allen AO. Probability, statistics, and queueing theory with computer science applications (2nd ed). Academic Press: Boston, MA, 1976.

[3] Gross D, Harris CM. Fundamentals of Queueing Theory (3rd ed). Wiley: New York, 1998.

[4] Trivedi KS. Probability and Statistics with Reliability, Queueing, and Computer Science Applications (2nd ed). Wiley: New York, 2002.

[5] Medhi J. Stochastic Models in Queueing Theory (2nd ed). Academic Press: Boston, MA, 2002.

[6] Kovalenko IN, Kuznetsov NY, Pegg PA. Mathematical theory of reliability of time dependent systems with practical applications. Wiley: New York, 1997.

[7] Kendall DG. Stochastic processes occurring in the theory of queues and their analysis by the method of imbedded Markov chains. Annals of Mathematical Statistics 1953 24: 338-354.

[8] Takagi H. Queueing Analysis: A Foundation of Performance Evaluation Finite Systems (vol 2). North-Holland: Amsterdam, 1993.

[9] Sztrik J. Finite source queueing systems and their applications: A bibliography. Working paper 2002, Institute of Mathematics & Informatics, University of Debrecen. Available from: `http://it.math.klte.hu/user/jsztrik/research/fsqreview.pdf`.

[10] Buzacott J, Shanthikumar J. Stochastic Models of Manufacturing Systems. Prentice Hall: Englewood Cliffs, NJ, 1993.

[11] Daigle JN. Queueing theory with applications to packet telecommunication. Springer: New York, 2004.

[12] Fortier, PJ. Computer systems performance evaluation and prediction. Digital Press: Burlington, MA, 2003.

[13] Perlman Y, Mehrez A, Kaspi M. Setting repair policy in a multi-echelon repairable-item inventory system with limited repair capacity. Journal of the Operational Research Society 2001 52: 198-209.

[14] Stecke KE, Aronson JE. Review of operator/machine interference models. International Journal of Production Research 1985 23: 129-151.

[15] Haque L, Armstrong MJ. A survey of the machine interference problem. European Journal of Operational Research 2006, forthcoming.

[16] Neuts MF, Meier KS. On the use of phase type distributions in reliability modelling of systems with two components. Operation Research Spectrum 1981 2: 227-234.

[17] Carmichael DG. Machine interference with general repair and running times. Mathematical Methods of Operations Research 1987 31: B115-B133.

[18] Gupta SM. Machine interference problem with warm spares, server vacations and exhaustive service. Performance Evaluation 1997 29: 195-211.

[19] Falin G, Templeton J. Retrial queues. Chapman and Hall: London, 1997.

[20] Wang KH, Hsu LY. Cost analysis of the machine repair problem with r non-reliable service stations. Microelectronics Reliability 1995 35: 923-934.

[21] Posafalvi A, Sztrik J. On the heterogeneous machine interference problem with priority and ordinary machines. European Journal of Operational Research 1989 41: 54- 63.

[22] Krishnamoorthy A, Ushakumari PV. K-out-of-n:G system with repair: The D-policy. Computers & Operations Research 2001 28: 973-981.

[23] Palesano J, Chandra J. A machine interference problem with multiple types of failures. International Journal of Production Research 1986 24: 567-582.

[24] Armero C, Bayarri MJ. Dealing with uncertainties in queues and networks of queues: A Bayesian approach. Multivariate Analysis, Design of Experiments, and Survey Sampling, Ghosh S (ed). Marcel Dekker: New York, 1999; 579-608.

[25] Castellanos ME, Morales J, Mayoral AM, Fried R, Armero C. On Bayesian design in finite source queues. COMPSTAT 2006 Proceedings in Computational Statistics, Rizzi A, Vichi M (eds). Physica-Verlag: Heidelberg, 2006; 1381-1388.

[26] Ausin MC, Lillo RE, Wiper MP. Bayesian control of the number of servers in a GI/M/c queueing system. Journal of Statistical Planning and inference 2006, forthcoming.

Related articles

eqr341, eqr342, eqr348, eqr349, eqr356, eqr364