Queueing Theory with Reneging

S. Bocquet

Defence Systems Analysis Division
Defence Science and Technology Organisation

DSTO-TR-1772

ABSTRACT

The literature on queueing with reneging is reviewed. Only random (Poisson) arrivals and unlimited capacity queues are considered, although some of the references also contain results for other arrival distributions or finite capacity queues. The paper focuses on the probability of service under steady state conditions; results for other metrics such as the expected waiting time may be found in the references. Analytic results, suitable for implementation in a spreadsheet model, are summarised. These formulae would be suitable for use in exploratory analysis, particularly in situations where queueing theory forms only a part of the model of a defence system.

RELEASE LIMITATION

Approved for public release
Queueing Theory with Reneging

Executive Summary

There is an extensive literature on queueing theory, including several texts. However, most queueing theory is concerned with queues in which all customers eventually get served. There is much less published work on queueing with impatient customers, that is, customers who renege before service is completed. In defence applications reneging is particularly important. One example is an interception situation, in which arriving ships, aircraft or missiles take a finite time to transit an area where interception is possible, and they escape if they are not intercepted within this time. In queueing theory terms the arriving entities are considered as customers who renege if they are not served (intercepted) within a finite time. Another example is a surveillance situation. In this case service represents the classification or identification of tracks in a sonar or radar system. The queueing model is different here in that the tracks can be lost at any time – reneging can occur during service as well as in the queue.

This paper contains a review of the literature on queueing with reneging. Only random arrivals and unlimited capacity queues are considered here, although some of the references also contain results for other arrival distributions or finite capacity queues. The paper focuses on the probability of service under steady state conditions; results for other metrics such as the expected waiting time may be found in the references.

The review summarises analytic results, suitable for implementation in a spreadsheet model. These formulae would be suitable for use in exploratory analysis, particularly in situations where queueing theory forms only a part of the model of a defence system.

Graphs of the probability of service as a function of traffic intensity are presented for all the queueing models with tractable solutions. These graphs should be helpful in understanding the various models, as in many cases the solutions found in the literature are presented as mathematical results only, with no graphs. In two cases involving reneging from both queue and server the results found in the literature have been extended to facilitate calculation of the probability of service for multiple servers. These calculations are presented in the appendices.
## Contents

1. INTRODUCTION ............................................................................................................... 1
2. QUEUES WITH DETERMINISTIC RENEGING .......................................................... 1
3. EXPONENTIAL WAITING TIME DISTRIBUTION ................................................... 6
4. EXPONENTIAL HOLDING TIME DISTRIBUTION .................................................. 7
5. GENERAL SERVICE TIME DISTRIBUTIONS ............................................................ 7
6. RESTRICTED ACCESS QUEUES .................................................................................. 12
7. GENERAL RENEGING DISTRIBUTIONS ..................................................................... 15
8. QUEUE DISCIPLINE AND PRIORITIES .................................................................... 15
9. CONCLUSION ................................................................................................................... 15

APPENDIX A : MULTIPLE SERVER QUEUEING SYSTEM WITH EXPONENTIAL SERVICE AND LIMITED HOLDING TIME ................................................................. 17

APPENDIX B : MULTIPLE SERVER QUEUEING SYSTEM WITH EXPONENTIAL SERVICE AND EXPONENTIAL HOLDING TIME LIMIT ................................................................. 18
1. Introduction

There is an extensive literature on queueing theory, including several texts [1,2,3,4,5]. However, most queueing theory is concerned with queues in which all customers eventually get served. There is much less published work on queueing with impatient customers, that is, customers who renege before service is completed. In defence applications reneging is particularly important. One example is an interception situation, in which arriving ships, aircraft or missiles take a finite time to transit an area where interception is possible, and they escape if they are not intercepted within this time. In queueing theory terms the arriving entities are considered as customers who renege if they are not served (intercepted) within a finite time. Another example is a surveillance situation. In this case service represents the classification or identification of tracks in a sonar or radar system. The queueing model is different here in that the tracks can be lost at any time – reneging can occur during service as well as in the queue.

This paper contains a review of the literature on queueing with reneging. Only Poisson arrivals and unlimited capacity queues are considered here, although some of the references also contain results for other arrival distributions or finite capacity queues. Apart from a brief mention of other queue disciplines in Section 8, it is assumed that the queue discipline is first come first served. The paper focuses on the probability of service under steady state conditions; results for other metrics such as the expected waiting time may be found in the references.

2. Queues with deterministic reneging

The simplest arrival pattern to model is a random one\(^1\). If the arrivals come from many independent sources, the resulting pattern is likely to be random. The random arrival pattern has the following important properties:

1. The probability of an arrival in any time interval is independent of the arrival pattern in preceding time intervals. Another way of expressing this is to say that the arrival process is a Markov process, or that it is *memoryless*.
2. The probability of \( n \) arrivals in time \( t \) follows a Poisson distribution
   \[
   p_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t},
   \]
   where \( \lambda \) is the mean arrival rate.
3. The probability distribution of the intervals between arrivals is exponential
   \[
   p(t) = \lambda e^{-\lambda t}
   \]

It can be inferred from the exponential distribution of arrival intervals that short intervals will be most common – the probability is greatest for \( t = 0 \) and decreases with increasing \( t \) – and therefore the arrivals will tend to cluster (Figure 1).

\(^1\) See Section 1.3 of reference 1 for a discussion of the properties of a random arrival pattern.
Figure 1. Random arrival pattern.

The simplest queueing models assume that the length of time taken for service also follows an exponential distribution with mean $\mu$, and the queue discipline is first come first served. Many applications of queueing theory are primarily concerned with situations where all customers eventually get served. In the absence of reneging or baulking, this requires $\lambda < c\mu$ for $c$ service channels. Otherwise the queue length will grow indefinitely, and there is no steady state solution to the problem. However, reneging is important in a model of interception – the ‘customers’ do not wait to be intercepted. In this situation there can be a steady state solution for arbitrary arrival rate $\lambda$, and the probability of interception (service) is generally less than 1. If reneging is immediate, so that no queue forms, the probability of reneging equals the probability that all $c$ service channels are busy. For random arrivals,

$$p_c = \frac{r^c / c!}{\sum_{k=0}^{c} r^k / k!}$$

where $r = \lambda / \mu$ is the traffic intensity. This is known as Erlang’s loss formula\(^2\). It was originally derived in 1917 by the Danish engineer A.K. Erlang as a measure of the calls lost by a busy telephone exchange. Erlang’s loss formula is most readily derived for the exponential service time distribution, but in fact it applies for any service time distribution with mean $\mu$, provided the arrival pattern is random\(^3\). The probability of service $p_s = 1 - p_c$. Figure 2 shows $p_s$ as a function of $r$ for several values of $c$. For large $r$, $p_s \approx c / r$.

---

\(^2\) Reference 1, p.47; reference 3, Section 2.5; reference 4, Section 1.4.1.

\(^3\) The proof is outlined in Section 5.2.2 of reference 3.
Figure 2. Probability of service calculated from Erlang’s loss formula.

More generally the customers may wait for a maximum time \( \tau \) before reneging. In this case the loss probability \( p_l \) is the probability that the waiting time exceeds \( \tau \). For the interception problem, \( \tau \) corresponds to the transit time for a target moving through the interception area. A formula for \( p_l \) has been derived by several authors using different methods [4 (Section 1.7), 6, 7, 8]. The derivation is not straightforward, but the final result is relatively simple:

\[
p_l = \begin{cases} 
\frac{r^c e^{-\mu \tau (c-r)}}{c!} \sum_{k=0}^{\infty} \frac{r^k}{k!} & \text{for } r \neq c \\
\frac{c^c}{c!} \left( 1 + c \mu \tau \right) & \text{for } r = c 
\end{cases}
\]

Note that this reduces to Erlang’s loss formula when \( \tau = 0 \). For large \( r \), \( p_s = (1 - p_l) \approx c/r \). It turns out that \( p_s \) approaches this limit more rapidly as \( \tau \) increases. For large \( \tau \),

\[
p_s = \begin{cases} 
1 & \text{for } r \leq c \\
c/r & \text{for } r > c 
\end{cases}
\]

As \( \tau \) increases, the fluctuations in the arrival rate are smoothed out more and more, so that service becomes certain if \( r < c \), and all servers are fully occupied if \( r > c \). Figure 3 shows the effect of varying \( T = \mu \tau \), for \( c = 1 \) and \( c = 4 \).
Another situation of interest is the classification of sonar or radar contacts. In this case reneging corresponds to the loss of a contact. This can occur during service as well as in the queue, so the overall holding time is limited to $\tau$, rather than just the waiting time. The loss probability for a single server with an exponential service time distribution is [6]:

$$p_l = \frac{1 - r}{1 - re^{\mu \tau}}$$

Daley [13] and Gnedenko and Kovalenko [4, Section 1.8.1] point out that provided the arrival stream is orderly, so that the probability of two customers arriving simultaneously is negligible, each arrival obtains at least partial service – reneging always occurs from the server, not from the queue. Hence $p_l$ could also be described as the probability of partial service, and $p_s = 1 - p_l$ is the probability of full service. Figure 4 shows the full service probability for different values of $T = \mu \tau$. In this case $p_s = 0$ when $\tau = 0$, and Erlang’s loss formula does not apply. In the limit $r \rightarrow 0$ $p_s = 1 - e^{-\mu \tau}$, which is the cumulative distribution function of the service time. For large $r$ and $\tau \neq 0$, $p_s \approx 1/r$. As in the limited waiting time case, $p_s$ approaches this limit more rapidly as $\tau$ increases: for $\mu \tau \gg 1$,

$$p_s = \begin{cases} 1 & \text{for } r \leq 1 \\ 1/r & \text{for } r > 1 \end{cases}$$

Unlike the limited waiting time case, $p_s$ is quite sensitive to the form of the service time distribution, and this limit only applies for exponential service.

The loss probability for multiple servers is complicated, because it is necessary to account for the different amounts of time customers in each service channel have spent in the
queue. An upper bound on \( p_l \) can be obtained by treating each server independently, each having a separate queue with an arrival rate of \( \lambda/c \). In this situation \( p_l(c, r) = p_l(1, r/c) \). The loss probability for a single queue with multiple servers will be less than this, particularly for small \( r \), due to the possibility that when one server is busy another will be free. Thus \( p_s(c, r) > p_s(1, r/c) \) for multiple servers with a single queue. \( p_s = 0 \) when \( \tau = 0 \) as for the single server case, and in the limit of large \( \tau \)

\[
\begin{align*}
  p_s &= 1 \quad \text{for } r \leq c \\
  p_s &= c/r \quad \text{for } r > c
\end{align*}
\]

which is the same as the large \( \tau \) limit for the situation where reneging occurs only in the queue, although here the limit only applies for exponential service.

A solution for the multiple server case was obtained by Kovalenko [9], but the original paper contains several errors. The correct results are presented in the first edition of the book by Gnedenko and Kovalenko [10, Section 1.6]. (The second edition of this book contains only a brief summary [4, Section 1.8.3].) The probability of full service is presented as an integral of the waiting time distribution, which is only evaluated in the single server case. In Appendix A this integral is used to derive recurrence relations for the busy server probabilities. The probability of full service is then expressed in terms of these probabilities. The results are shown in Figure 4 for 1-4 service channels.

![Figure 4. Effect of limited holding time and number of servers on full service probability.](image-url)
3. Exponential waiting time distribution

A relatively simple formula for the service probability can also be obtained if the waiting time limit, rather than being fixed, is a random variable with an exponential distribution \( p(t) = \frac{1}{\tau} e^{t/\tau} \):

\[
p_s = \frac{\sum_{k=0}^{c-1} \frac{r^k}{k!} + \frac{r^{c-1}}{(c-1)!} z}{\sum_{k=0}^{c-1} \frac{r^k}{k!} + \frac{r^c}{c!} (1 + z)} \quad \text{where } z = e^{r\mu\tau} (r\mu\tau)^{-c\mu\tau} \gamma(c\mu\tau + 1, r\mu\tau)
\]

This result was obtained by Ancker and Gafarian [11], although it is expressed somewhat differently here in order to show the relationship with Erlang’s loss formula (for \( \tau = 0 \) \( z = 0 \), and the Erlang formula is obtained). The Incomplete Gamma function \( \gamma(a,x) \) can be implemented in Microsoft Excel using the Log Gamma function and the cumulative distribution function for the Gamma distribution:

\[
\gamma(a,x) = \text{EXP}(\text{GAMMALN}(a)) \times \text{GAMMADIST}(x,a,1,\text{TRUE})
\]

Figure 5 shows the probability of service for one and four service channels with varying mean waiting time \( T = \mu\tau \). The general form is very similar to the fixed waiting time case, and the limits \( T \to 0 \) and \( T \to \infty \) are the same.

![Figure 5. Probability of service with exponentially distributed waiting time for 1 and 4 service channels.](image)
4. Exponential holding time distribution

A relatively simple result can also be obtained if the total holding time in the system is exponentially distributed. Ancker and Gafarian [12] derived the probability of complete service and various other measures for the single server case. A formula for the multiple server case can be obtained using their method with the system state probabilities given by Gnedenko and Kovalenko [4, Section 1.8.4]:

\[
p_s = \frac{\sum_{k=0}^{r-1} \frac{r^k}{k!} \left( \frac{\mu \tau}{1 + \mu \tau} \right)^{k+1}}{\sum_{k=0}^{r-1} \frac{r^k}{k!} \left( \frac{\mu \tau}{1 + \mu \tau} \right)^k} + \frac{c}{w}
\]

where

\[
w = \frac{e^{\gamma \mu \tau}(c(\mu \tau + 1), r\mu \tau)}{(c-1)(1 + \mu \tau)^{c-1}(r\mu \tau)^{\mu \tau}}
\]

For \( r = 0 \), \( p_s = \mu \tau / (1 + \mu \tau) \). Details of the solution are given in Appendix B. Figure 6 shows the full service probability for one and four service channels with varying mean holding time \( T = \mu \tau \). For \( \tau = 0 \), \( p_s = 0 \), and the limit \( T \to \infty \) is the same as in the other cases described above.

![Figure 6. Full service probability with exponentially distributed holding time for 1 and 4 service channels.](image)

5. General service time distributions

Reneging in single server queues with general service time distributions has been studied by Daley [13,14], Rao [15], Cohen [16,17], Stanford [18] and Baccelli et al. [19]. An integral
equation for the waiting time distribution can be constructed, and this equation has been solved in particular cases. For Poisson arrivals, the general solution is a distribution function \( F(x) \) whose Laplace-Stieltjes transform \( \phi(s) \) is \([13,14]\)

\[
\phi(s) = \frac{sF(0)}{s - \lambda + \lambda \beta(s)}
\]

where \( \beta(s) \) is the Laplace-Stieltjes transform of the cumulative service time distribution \( B(x) \). The waiting time distribution \( W(x) \) is obtained from \( F(x) \), with \( F(0) \) determined by the reneging behaviour. This is a generalisation of the Pollaczek-Khintchine formula for the standard queueing system with general service and no reneging, in which \( W(x) = F(x) \) for \( x \geq 0 \), and \( F(0) = W(0) = 1 - r \). (\( r = \lambda/\mu \) is the traffic intensity as before.) For deterministic reneging after a time limit \( \tau \)

\[
W(x) = F(x) \quad 0 \leq x < \tau
\]

\[
= 1 \quad x \geq \tau
\]

In the case of reneging from the queue only, \( W(0) = 1 - rW(\tau) \) \([14]\).

If reneging occurs from both queue and server, \( W(0) = 1/W(\tau) \) \([13,14]\).

For an Erlang service time distribution the cumulative distribution function is

\[
B(x) = 1 - e^{-k\alpha} \sum_{i=0}^{k-1} (k\alpha)^i / i!
\]

and its Laplace-Stieltjes transform is \( \beta(s) = (1 + s/k\mu)^k \)

\( \phi(s) \) may be inverted \([13,14]\) to give \( W(x) = W(0) \sum_{i=0}^{k} A_i e^{\alpha_i s\mu} \)

where \( \alpha_i \) are the roots of the polynomial \( P(z) = (1 - \gamma z)(1 + z)^k - 1 \) with \( \gamma = k/r \) and

\[
A_j = \frac{1 + \alpha_j}{(k+1)\alpha_j + 1 - r}
\]

\( \alpha_0 = 0 \) and \( A_0 = 1/(1 - r) \) for all \( k \). In the case of exponential service (\( k = 1 \)), \( \alpha_1 = r - 1 \) and \( A_1 = r/(r - 1) \). For \( k > 1 \), \( P(z) \) does not factorise, but the roots may be found using radicals for \( k = 2, 3 \) and 4.

In the case of a fixed waiting time limit (reneging from the queue only), the probability of service is \( p_s = (1 - W(0))/r \). \( W(0) \) is the probability that the waiting time is zero, in other words the probability that the server is empty. This relationship also applies to the limited holding time case (reneging from both the queue and server), but only for exponential service.
In the limit $k \to \infty$ service is deterministic, requiring a time $1/\mu$ for each customer [3, p128]. The cumulative distribution of the service time is a step function, although it is not easy to obtain this directly by taking the limit $k \to \infty$ in the Erlang distribution. The Erlang distribution is a type of gamma distribution; for large $k$ it tends to a normal distribution with variance $1/k\mu^2$, according to the central limit theorem, so the variance $\to 0$ as $k \to \infty$. Deterministic service has been studied by Hokstad [20] and De Kok and Tijms [21]. In this case the waiting time distribution is

$$W(x) = W(0)e^{\mu x} \sum_{i=0}^{\infty} \left[re^{-r}(i-\mu x)\right]^i/i!$$

$n$ is the largest integer in $\mu x$, i.e. $n \leq \mu x < (n+1)$.

Figure 7 shows the probability of service for a single server with Erlang and deterministic service and different waiting times $T = \mu \tau$. The effect of changing the service time distribution is fairly small. Note that both the limits $T \to 0$ and $T \to \infty$ are independent of the service distribution.

![Figure 7. Erlang and deterministic ($k \to \infty$) service with a single server and different waiting time limits $T = \mu \tau$.](image)

These results illustrate the effect of changing the standard deviation of the service time, but the Erlang distribution is not very flexible because the expected service time $\varepsilon = 1/\mu$
and its standard deviation $\sigma$ have the specific relationship $\sigma = \varepsilon / \sqrt{k}$, with $k$ an integer, and the results become increasingly difficult to compute as $k$ increases. De Kok and Tijms [21] propose a two-moment approximation for the loss probability $p_l$ (and other performance measures) in terms of the mean $\varepsilon$ and standard deviation $\sigma$ of the service time distribution, and the values of $p_l$ for deterministic and exponential service:

$$p_l \approx [1 - (\sigma / \varepsilon)^2] p_l^{\text{det}} + (\sigma / \varepsilon)^2 p_l^{\exp}$$

Their numerical results for several different service time distributions show that this approximation is accurate to two decimal places.

Boots and Tijms [22] give an approximation for the loss probability in the limited waiting time case, based on the relationship between the waiting times for this case and the situation without reneging. Their approximation is valid for multiple servers, but it is only applicable for $r/c < 1$ because the waiting time is undefined for $r/c > 1$ in the standard queue without reneging.

Daley [14] obtained the Laplace-Stieltjes transform of the virtual waiting time distribution for a single server queue with exponential reneging from the queue and general service, however the transform is difficult to invert except in the case of exponential service where the results of Ancker and Gafarian [11,12] are recovered.

In the case of limited holding time (reneging from the server), the probability of complete service is the probability that the sum of the waiting time and the time required for service is less than or equal to $\tau$. This can be defined in terms of the distribution of the total amount of unfinished work in the system $U(x)$ as $p_s = U(\tau)$. The unfinished work is also known as the virtual waiting time, because it is the time a hypothetical arrival would have to wait before commencing service. $U(x)$ may be obtained by integrating $W(x)$ with respect to the service time probability distribution, or vice versa:

$$U(x) = \int_0^\tau W(x - y)dB(y)$$

$$= [W(x - y)B(y)]_0^\tau - \int_0^\tau B(y)dW(x - y)$$

$$= W(0)B(x) + \int_0^\tau B(x - z)dW(z)$$

Note that $B(0) = 0$, and hence $U(0) = 0$, whereas $W(0) \neq 0$. $W(0)$ is the probability that the waiting time is zero, in other words the probability that the server is empty. Alternatively, the Laplace-Stieltjes transform of $U(x)$ can be obtained by solving an integral equation [16]. For an Erlang service time distribution the transform may be inverted by the same method used by Daley [13] to get $W(x)$, with the result

$$U(x) = W(0) \sum_{i=0}^k \frac{A_i}{(1 + \alpha_i)^k} e^{\alpha_i x}$$
$A_i$ and $a_i$ are as defined previously for $W(x)$. In the case of exponential service ($k = 1$), $U(x) = [W(x) - W(0)]/r$ and $p_s = U(\tau) = [1 - W(0)]/r$ since $W(\tau) = 1$.

For deterministic service, the cumulative distribution of the service time is a step function:

\[
B(x) = \begin{cases} 
0 & \text{for } x < \frac{1}{\mu} \\
1 & \text{for } x \geq \frac{1}{\mu}
\end{cases}
\]

Hence $B(\tau - x) = 0$ for $\tau - x < \frac{1}{\mu}$ or $x > \tau - \frac{1}{\mu}$ and

\[
p_s = W(0)B(\tau) + \int_{0}^{\tau} B(\tau - x)dW(x) = W(0)B(\tau) + \int_{0}^{\tau - \frac{1}{\mu}} dW(x) = W(0)B(\tau) + W(\tau - \frac{1}{\mu}) - W(0) = W(\tau - \frac{1}{\mu}) \quad \text{for } \tau \geq \frac{1}{\mu} = 0 \quad \text{for } \tau < \frac{1}{\mu}
\]

Figure 8 shows the probability of complete service for Erlang and deterministic service time distributions. The asymptotic behaviour for large $r$ depends strongly on $k$, unlike the limited waiting time case. For exponential service ($k = 1$) $p_s \approx 1/r$ for large $r$; for $k > 1$, $p_s$ drops to zero much more rapidly with increasing $r$. For $T < 1$, $p_s$ decreases with increasing $k$ for all $r$, whereas for $T > 1$ $p_s$ increases with increasing $k$ at low arrival rates and then crosses over to decrease more rapidly with increasing $k$. The solution for $T = 1$ and $k \to \infty$ is $p_s = e^{-r}$, which forms a kind of boundary between the two regimes. The probability of full service is zero for deterministic service and $T < 1$. 
The rapid fall off in the full service probability for $k > 1$ is a consequence of the first come – first served queue discipline: for $r > 1$ the server becomes congested, so that each successive customer has waited long enough that service is unlikely to be completed before the customer reneges. This queueing model is described in the literature as having ‘unaware’ customers. In the surveillance / classification context this description is rather misleading, because in this application it is assumed that the ‘customers’ don’t have a choice as to whether or not they join the queue, and if they did they would make that choice so as to avoid being served. In this application it might be better described as a ‘dumb server’ model – an intelligent server would not stick to the first come – first served queue discipline in overload conditions. In the exponential service ($k = 1$) case, the congestion is mitigated by the fact that a significant proportion of customers require zero service time – probably not a realistic assumption, unless many ‘customers’ are automatically dismissed as irrelevant.

6. Restricted access queues

It might be more realistic to apply a last come – first served queue discipline to the classification problem, but this is mathematically intractable, because the waiting time for a customer depends in general on the arrival and service times of both preceding and
following customers. Hence the system has ‘memory’ and cannot be modelled as a Markov process. An alternate model which is more amenable to solution is the ‘aware’ customer model, in which customers only join the queue if both waiting and service can be completed within the allowed time. In the classification problem this corresponds to automatically ignoring tracks which are about to disappear, or about to cross a ‘last chance’ threshold. In this model the probability of service is the probability of joining the queue. Solutions were found for exponential service by Gavish and Schweitzer [23], and for deterministic service by Hokstad [20]. In the case of exponential service

\[
q_s = 1 - Q b^{1-r} e^{r(b-1)} \text{ where } b = e^{-\mu r} \text{ and } \\
Q = (1-r) \left[1 - rb^{1-r} e^{r(b-1)} - (rb)^{1-r} e^{rb} [\gamma(r,r) - \gamma(r,rb)] \right]^{-1} \quad r \neq 1
\]

In the case of deterministic service

\[
q_s = D/(1+D) \text{ where } D = e^{r(\mu-1)} \sum_{i=0}^{n} \left[ r e^{-r(i+1-\mu \tau)} \right] / i! \text{ with } n \leq \mu \tau \leq (n+1)
\]

Figure 9 shows the full service probability with restricted and unrestricted access to the queue, for exponential service. Figure 10 shows the case of deterministic service. In both cases the full service probability is significantly greater with restricted access to the queue. The restriction means that server work is not wasted on partially servicing customers who renege before service is completed. The deterministic service model with restricted access and the exponential service model with unrestricted access both have the same limit as \(T \to \infty\), that is

\[
q_s = 1 \quad \text{for } r \leq 1 \\
q_s = 1/r \quad \text{for } r > 1
\]
Figure 9. Probability of full service for exponential service and limited holding time \( T = \mu \tau \) with restricted and unrestricted (asterisk) access to the queue.

Figure 10. Probability of full service for deterministic service and limited holding time \( T = \mu \tau \) with restricted and unrestricted (asterisk) access to the queue.

The restricted access queue is used as a model for buffer design in communication systems. Here the service time represents the message length, and the holding time \( \tau \) becomes the buffer capacity \( K \). The model is used to determine the buffer size required to keep the rejection probability \( 1 - p_s \) below a specified level. De Kok and Tijms [24] used the solutions for exponential and deterministic service to construct a two moment approximation for the buffer size with a general service (message length) distribution. The
rejection probability must be kept very small in order to make the communications system reliable - this approximation applies to $T > 1$ and $r < 1$.

7. General reneging distributions

Yurkevich [25] evaluated the loss probability and waiting time for reneging from the queue only with a general reneging distribution, multiple servers and exponential service. The arrival process is Poisson, with the arrival rate dependent on the number of busy servers. The solution is quite complex and contains an integral which must be evaluated numerically. Movaghar [7] has also studied this model. Baccelli et al. [19] consider general reneging from a single server queue with a general service distribution. They obtain the Laplace-Stieltjes transform of the virtual waiting time distribution for an Erlang reneging distribution. Baccelli et al. also obtain the relationship between the virtual and actual waiting time distributions for a general arrival process. In the case of exponential arrivals with reneging only from the queue the virtual waiting time equals the actual waiting time.

8. Queue discipline and priorities

So far it has been assumed that the queue discipline is ‘first come first served’. Barrer [6] compared this with a situation where the customers are served at random, and found that the steady state probability of service is slightly less for random selection. Another situation of interest has two classes of customer with different priorities. Choi et al. [8] obtained analytic results for a model with a single server, exponential service and priority customers with either limited waiting time or limited holding time.

9. Conclusion

The theory of queueing with reneging has been reviewed, with two situations relevant to defence analysis in mind. The first situation is an interception problem, where ships, aircraft or missiles must be intercepted within a limited time. This case is described by fairly straightforward formulae which can be implemented in a spreadsheet, and the results are only weakly dependent on the form of the service time distribution. The second situation is a surveillance problem, where contacts must be identified, and tracks may be lost at any time. In this situation a simple mathematical description is more difficult, because the result is strongly dependent on the form of the service time distribution and on the queue discipline. A restricted access queue, where tracks which cannot be processed in the available time are ignored, may provide the most realistic simple model for this case. The relatively simple formulae are particularly useful for exploratory analysis in situations where queueing theory describes only part of the problem, as they can be readily combined with other mathematical models describing the other aspects.
Graphs of the probability of service as a function of traffic intensity are presented for all the queueing models with tractable solutions. These graphs should be helpful in understanding the various models, as in many cases the solutions found in the literature are presented as mathematical results only, with no graphs. In two cases involving reneging from both queue and server the results found in the literature have been extended to facilitate calculation of the probability of service for multiple servers. These calculations are presented in the appendices.
Appendix A: Multiple server queueing system with exponential service and limited holding time

The solution is expressed in terms of the probabilities $\pi_k^n$ that $k$ out of $n$ servers are busy. For $k < n$, at least one server is free, and there is no waiting [10, Section 1.6, eqn (10)]:

$$\pi_k^n = \frac{\pi_0^n}{k!} (1 - e^{-r\mu})^k, \ k < n$$

The probability that all servers are busy equals the probability that a customer has to wait any amount of time, that is

$$\pi_n^n = \int_0^\infty w(x) \, dx$$

The waiting time probability density is [4, Section 1.8.3; 10, Section 1.6, eqn (12)]:

$$w(x) = \pi_0^n \frac{\mu^n x^{n-1} e^{(r-1)\mu x} (e^{-\mu x} - e^{-r\mu x})^{n-1}}{(n-1)!}$$

The single server case is easily evaluated:

$$\pi_1^n = \pi_0^n \frac{1}{r+1} (e^{(r-1)\mu x} - 1), \ r \neq 1$$

$$\pi_1^n = \pi_0^n \mu x, \ r = 1$$

The general formula for $n$ servers ($n > 1$) is complicated:

$$\pi_n^n = \frac{\pi_0^n}{(n-1)!} \left\{ \sum_{k=0}^{n-1} \frac{(n-2)!}{k!(n-2-k)!} \left[ \frac{e^{(r-k-2)\mu x}}{r-k-2} - (1 - e^{-r\mu})^{n-1} \right] \right\}$$

This formula has singularities at $r = 1$ and $r = k+2$. Recurrence relations can be derived from the integral for $\pi_n^n$ which are easier to use. After a change of variable

$$\pi_n^n = \pi_0^n \frac{\mu^n x^{n-1} e^{(r-1)\mu x} (e^{-\mu x} - e^{-r\mu x})^{n-1}}{(n-1)!} \int_0^\infty e^{(r-1)\mu x} (e^{-\mu x} - e^{-r\mu x})^{n-1} \, dx$$

$$= \pi_0^n \frac{r^n}{(n-1)!} \int_0^a u^{n-1} \frac{du}{(u+a)^r} \text{ where } u = e^{-\mu x} - e^{-r\mu x} \text{ and } a = e^{-r\mu}$$

and the use of integral tables [26, §2.111] the following recurrence relations for $\phi_n^n = \pi_n^n / \pi_0^n$ are obtained:

$$\phi_{n+1}^{n+1}(r) = \frac{r}{n t_{n+1}} [\phi_n^{n+1}(r) - e^{-r\mu} \phi_n^n(r)]$$

for $r \neq n+1$, and
\[ \phi_{n+1} (n + 1) = \left( \frac{n+1}{n} \right)^{n+1} [\phi_n^*(n) - \phi_{n+1}^*(n)] \]

for \( r = n+1 \). Note that \( \phi_{n+1}^* \) is obtained from the formula for \( \pi_n^* \) above. The empty system probability \( \pi_0^* \) is obtained from the normalisation condition

\[ \sum_{k=0}^{n} \pi_k^* = 1 \]

The probability of full service \( p_s \) equals the probability that at least one server is free and service is completed within the time limit \( \tau \), plus the probability that the customer waits a time \( x \) and service is completed within the remaining time \( \tau - x \), which may be expressed in terms of \( \pi_n^* \):

\[
\begin{align*}
p_s &= (1 - \pi_n^*)(1 - e^{-\mu\tau}) + \int_0^\tau w(x)(1 - e^{-\mu(\tau-x)})dx \\
&= 1 - e^{-\mu\tau} (1 - \pi_n^*) - e^{-\mu\tau} \int_0^\tau e^{\mu x} w(x)dx \\
&= 1 - e^{-\mu\tau} [1 - \pi_n^*(r) + \left( \frac{\mu}{e}\right)^n \pi_n^*(r+1)]
\end{align*}
\]

**Appendix B: Multiple server queueing system with exponential service and exponential holding time limit**

The probabilities for \( k \) customers present in a system with \( c \) servers are [4, Section 1.8.4]

\[
\begin{align*}
p_k &= p_0 \frac{r^k}{k!} \left( \frac{\mu\tau}{1 + \mu\tau} \right)^k \text{ for } k < c \\
p_k &= p_0 \frac{(r\mu\tau)^k}{c!(1 + \mu\tau)^c \prod_{j=c+1}^{\infty} (c\mu\tau + j)} \\
&= p_0 \frac{(r\mu\tau)^k \Gamma(c\mu\tau + c + 1)}{c!(1 + \mu\tau)^c \Gamma(c\mu\tau + k + 1)} \text{ for } k \geq c
\end{align*}
\]

The empty system probability \( p_0 \) is determined by the normalisation \( \sum_{k=0}^{\infty} p_k = 1 \).

In the steady state the probability of service is the average number of busy servers divided by \( r \):

\[
p_s = \frac{1}{r} \left( \sum_{k=1}^{c-1} kp_k + c \sum_{k=c}^{\infty} p_k \right)
\]
(The method used by Ancker and Gafarian [12] to calculate $p_s$ is more involved, but the result is the same.) The infinite sum may be evaluated using an identity for the incomplete gamma function [11, Eq. 20] (see also [12, Eq. 12; 26, §8.356]):

$$
\sum_{k=0}^{n-1} \frac{x^k}{\Gamma(z+k+1)} = e^x x^{-z} \left[ \gamma(z,x) - \frac{\gamma(z+n,x)}{\Gamma(z)} \right]
$$

The second term on the right hand side goes to zero as $n \to \infty$. Hence

$$
\sum_{k=0}^{\infty} \frac{x^k}{\Gamma(z+k+1)} = \sum_{k=0}^{\infty} \frac{x^k}{\Gamma(z+k+1)} - \sum_{k=0}^{n-1} \frac{x^k}{\Gamma(z+k+1)} = e^x x^{-z} \frac{\gamma(z+n,x)}{\Gamma(z+n)}
$$

which can be used to evaluate the probability all servers are busy:

$$
\sum_{k=c}^{\infty} p_k = \frac{p_0}{c!(1+\mu\tau)^c} \Gamma(c\mu\tau+c+1) \sum_{k=c}^{\infty} \frac{(r\mu\tau)^k}{\Gamma(c\mu\tau+k+1)}
$$

$$
= \frac{p_0}{c!(1+\mu\tau)^c} \Gamma(c\mu\tau+c+1) e^{r\mu\tau} (r\mu\tau)^{-c\mu\tau} \frac{\gamma(c\mu\tau+c,r\mu\tau)}{\Gamma(c\mu\tau+c)}
$$

$$
= p_0 \frac{e^{r\mu\tau} \gamma(c\mu\tau+c,r\mu\tau)}{(c-1)!(1+\mu\tau)^{-1}(r\mu\tau)^{c\mu\tau}}
$$

$$
= p_0 w
$$

The final result is obtained by substitution into the expression for $p_0$, with

$$
p_0 = \left[ \sum_{k=0}^{\infty} \frac{x^k}{k!} \left( \frac{\mu\tau}{1+\mu\tau} \right)^k + w \right]^{-1}
$$
References

2. P.M. Morse, *Queues, inventories and maintenance*, Wiley 1958
# DISTRIBUTION LIST

Queueing Theory with Reneging

S. Bocquet

**AUSTRALIA**

## DEFENCE ORGANISATION

<table>
<thead>
<tr>
<th>Task Sponsor</th>
<th>No. of copies</th>
</tr>
</thead>
<tbody>
<tr>
<td>DGAD</td>
<td>1</td>
</tr>
</tbody>
</table>

### S&T Program

<table>
<thead>
<tr>
<th>Position</th>
<th>No. of copies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chief Defence Scientist</td>
<td>1</td>
</tr>
<tr>
<td>Deputy Chief Defence Scientist, Policy</td>
<td>1</td>
</tr>
<tr>
<td>AS Science Corporate Management</td>
<td>1</td>
</tr>
<tr>
<td>Director General Science Policy Development</td>
<td>1</td>
</tr>
<tr>
<td>Counsellor Defence Science, London</td>
<td></td>
</tr>
<tr>
<td>Counsellor Defence Science, Washington</td>
<td></td>
</tr>
<tr>
<td>Scientific Adviser to MRDC, Thailand</td>
<td></td>
</tr>
<tr>
<td>Scientific Adviser Joint</td>
<td>1</td>
</tr>
<tr>
<td>Navy Scientific Adviser</td>
<td></td>
</tr>
<tr>
<td>Scientific Adviser - Army</td>
<td></td>
</tr>
<tr>
<td>Air Force Scientific Adviser</td>
<td>1</td>
</tr>
<tr>
<td>Scientific Adviser to the DMO</td>
<td></td>
</tr>
<tr>
<td>Systems Sciences Laboratory</td>
<td></td>
</tr>
<tr>
<td>Information Sciences Laboratory</td>
<td></td>
</tr>
<tr>
<td>DSTO Library and Archives</td>
<td></td>
</tr>
<tr>
<td>Information Sciences Laboratory</td>
<td></td>
</tr>
<tr>
<td>Capability Development Group</td>
<td></td>
</tr>
</tbody>
</table>

## Systems Sciences Laboratory

<table>
<thead>
<tr>
<th>Person</th>
<th>No. of copies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matthew Fewell (MOD)</td>
<td>1 Printed</td>
</tr>
<tr>
<td>Anthony Ween (MOD)</td>
<td>1</td>
</tr>
<tr>
<td>Genevieve Mortiss (MOD)</td>
<td>1</td>
</tr>
</tbody>
</table>

## Information Sciences Laboratory

<table>
<thead>
<tr>
<th>Position</th>
<th>No. of copies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chief of Defence Systems Analysis Division</td>
<td></td>
</tr>
<tr>
<td>Research Leader Integrated Capabilities</td>
<td></td>
</tr>
<tr>
<td>Task Manager: Thea Clark</td>
<td>1</td>
</tr>
<tr>
<td>Author: S. Bocquet</td>
<td></td>
</tr>
<tr>
<td>Michael Ling</td>
<td>1 Printed</td>
</tr>
<tr>
<td>Nigel Perry</td>
<td>1 Printed</td>
</tr>
<tr>
<td>Daniel Hall</td>
<td>1</td>
</tr>
<tr>
<td>Ian Grivell (IND)</td>
<td>1</td>
</tr>
</tbody>
</table>

## DSTO Library and Archives

<table>
<thead>
<tr>
<th>Location</th>
<th>No. of copies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Library Fishermans Bend</td>
<td></td>
</tr>
<tr>
<td>Library Edinburgh</td>
<td>1</td>
</tr>
<tr>
<td>Defence Archives</td>
<td>1</td>
</tr>
</tbody>
</table>

## Capability Development Group

<table>
<thead>
<tr>
<th>Position</th>
<th>No. of copies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Director General Maritime Development</td>
<td></td>
</tr>
<tr>
<td>Director General Capability and Plans</td>
<td></td>
</tr>
<tr>
<td>Assistant Secretary Investment Analysis</td>
<td></td>
</tr>
</tbody>
</table>
Director Capability Plans and Programming

**Chief Information Officer Group**
- Deputy CIO
- Director General Information Policy and Plans
- AS Information Strategy and Futures
- Director General Australian Defence Simulation Office
- Director General Information Services

**Strategy Group**
- Director General Military Strategy
- Director General Preparedness
- Assistant Secretary Strategic Policy
- Assistant Secretary Governance and Counter-Proliferation

**Navy**
- Maritime Operational Analysis Centre, Building 89/90 Garden Island Sydney
  - Deputy Director (Operations)
  - Deputy Director (Analysis)
- Director General Navy Capability, Performance and Plans, Navy Headquarters
- Director General Navy Strategic Policy and Futures, Navy Headquarters

**Air Force**
- SO (Science) - Headquarters Air Combat Group, RAAF Base, Williamtown NSW 2314

**Army**
- ABCA National Standardisation Officer, Land Warfare Development Sector, Puckapunyal
- SO (Science) - Land Headquarters (LHQ), Victoria Barracks NSW
- SO (Science), Deployable Joint Force Headquarters (DJFHQ) (L), Enoggera QLD

**Joint Operations Command**
- Director General Joint Operations
- Chief of Staff Headquarters Joint Operations Command
- Commandant ADF Warfare Centre
- Director General Strategic Logistics

**Intelligence and Security Group**
- DGSTA Defence Intelligence Organisation
- Manager, Information Centre, Defence Intelligence Organisation
- Assistant Secretary Capability Provisioning
- Assistant Secretary Capability and Systems
- Assistant Secretary Corporate, Defence Imagery and Geospatial Organisation
Defence Materiel Organisation
  Deputy CEO  Doc Data Sheet
  Head Aerospace Systems Division  Doc Data Sheet
  Head Maritime Systems Division  Doc Data Sheet
  Chief Joint Logistics Command  Doc Data Sheet

OTHER ORGANISATIONS
  National Library of Australia  1
  NASA (Canberra)  1

UNIVERSITIES AND COLLEGES
  Australian Defence Force Academy
    Library  1
    Head of Aerospace and Mechanical Engineering  1
  Serials Section (M list), Deakin University Library, Geelong, VIC  1
  Hargrave Library, Monash University  Doc Data Sheet
  Librarian, Flinders University  1

OUTSIDE AUSTRALIA

INTERNATIONAL DEFENCE INFORMATION CENTRES
  US Defense Technical Information Center  1
  UK Dstl Knowledge Services  1
  Canada Defence Research Directorate R&D Knowledge & Information Management (DRDKIM)  1
  NZ Defence Information Centre  1

ABSTRACTING AND INFORMATION ORGANISATIONS
  Library, Chemical Abstracts Reference Service  1
  Engineering Societies Library, US  1
  Materials Information, Cambridge Scientific Abstracts, US  1
  Documents Librarian, The Center for Research Libraries, US  1

SPARES  5 Printed

Total number of copies: Printed 9  PDF 30  =  39
Queueing Theory with Reneging

S. Bocquet

DSTO Defence Science and Technology Organisation
506 Lorimer St
Fishermans Bend, Victoria 3207 Australia

DSTO-TR-1772
Technical Report
September 2005

20
26


Chief, Defence Systems Analysis Division

Approved for public release

The literature on queueing with reneging is reviewed. Only random (Poisson) arrivals and unlimited capacity queues are considered, although some of the references also contain results for other arrival distributions or finite capacity queues. The paper focuses on the probability of service under steady state conditions; results for other metrics such as the expected waiting time may be found in the references. Analytic results, suitable for implementation in a spreadsheet model, are summarised. These formulae would be suitable for use in exploratory analysis, particularly in situations where queueing theory forms only a part of the model of a defence system.