The Recursive Evaluation of the $M/M/r/K/K$ Queue

Ian G. Kennedy and Anthony E. Krzesinski
School of Construction, Economics and Management, University of the Witwatersrand, WITS 2050, South Africa
Department of Mathematical Sciences, University of Stellenbosch, Stellenbosch 7600, South Africa
ian.kennedy@wits.ac.za and aek1@cs.sun.ac.za

All correspondence should be addressed to Anthony E. Krzesinski (e-mail: aek1@cs.sun.ac.za, Tel: +27 21 808 4310)

Abstract—We present a recursive form of the equation for the probability that the queueing delay in an $M/M/r/K/K$ queue is larger than a given bound. The recursive form is computationally efficient and is not subject to numerical overflow. The practical implication of the recursion is that the teletraffic engineer has an accurate, practical formula for calculating the delay probability distribution for large switches where the calls are offered from a large finite population.

Index Terms—Call blocking probability; Call delay probability; Erlang-B system; Erlang-C system; Engset system; Machine repair model; Recursive computational forms.

I. INTRODUCTION

New telecommunication applications driven by multimedia and Internet activities are placing an increasing emphasis on the quality of service afforded to network users. The need for numerically stable and computationally efficient methods to calculate the call blocking and delay probabilities is as necessary today in the era of broadband multiservice networks as it was in the age of single service telephone networks.

II. HISTORICAL BACKGROUND

It is 100 years since AK Erlang joined the Copenhagen Telephone Company in 1908. The three basic teletraffic equations of the early part of the last century were his Erlang-B and Erlang-C formulae, followed later by Engset’s formula.

Jagerman [3, 4] made a major contribution to the numerical evaluation of Erlang’s equations. Further progress was made by Iversen and Sanders [2] concerning the numerical evaluation of the more general and therefore more useful Engset equation.

Boucher [1] pointed out that an equation to compute the delay probability for a finite source queueing system was missing. This shortcoming was resolved in 1997 by Stevens and Sinclair [6]. In our paper we shall refer to Eqn. (12) from [6] for the delay probability as Eqn. (S-S). The four equations can be categorized as in Table I. Up to now, all the equations except for Eqn. (S-S) have been available in a recursive form for accurate, efficient numerical evaluation.

This work was supported by grant numbers 2054027 and 2677 from the South African National Research Foundation, Nokia-Siemens Networks and Telkom SA Limited.

With reference to Table I, in comparison to the infinite population traffic formulae, the finite population equations are more general since they cater for a finite set of sources where the busy sources are removed from the pool of sources that generate new arrivals. The finite population equations are more accurate and are not over-conservative as compared to the infinite population equations. Teletraffic engineers will find the finite population equations useful and in many cases the finite population equations will replace the Erlang-B and Erlang-C equations. However, an accurate and efficient way of evaluating Eqn. (S-S) has not been published to date. This deficiency is resolved in our paper.

TABLE I

| infinite population |    | finite population |
|---------------------|--|--|----------------|
| blocking system     | calls lost | blocking system | calls delayed |
| Erlang-B            | Engset     | Erlang-C        | Eqn. (S-S)    |

With reference to Table I, in comparison to the infinite population traffic formulae, the finite population equations are more general since they cater for a finite set of sources where the busy sources are removed from the pool of sources that generate new arrivals. The finite population equations are more accurate and are not over-conservative as compared to the infinite population equations. Teletraffic engineers will find the finite population equations useful and in many cases the finite population equations will replace the Erlang-B and Erlang-C equations. However, an accurate and efficient way of evaluating Eqn. (S-S) has not been published to date. This deficiency is resolved in our paper.

A. The Erlang-B system

Erlang’s function

$$E_1(\rho, N) = \frac{\rho^N}{N!} \sum_{i=0}^{N} \frac{\rho^i}{i!}$$

(1)

where $\rho = \lambda/\mu$ is perhaps the single most important expression in the field of teletraffic theory. It was first derived by Erlang as a formula for the probability that calls arriving individually in a Poisson stream with parameter $\lambda$ to a link consisting of $N$ circuits would find all circuits occupied and therefore be lost: $1/\mu$ is the average call holding time. Erlang’s function is widely used in the field of telecommunications, in its own right and as part of more complicated analyses, both exact and approximate. The Kendall notation for the Erlang-B system is $M/M/N/N$.

Despite its appealing analytical form, the function (1) is non-trivial to compute for large values of $N$. The problem in evaluating Eqn. (S-S) is similar to the problem encountered in trying to evaluate the Erlang-B equation: in the IEEE floating point standard, the largest number that can be represented in
double precision is about $10^{308}$. Two components in Eqn. (1) cause numerical problems: the factorial term and the power term. The factorial term $N!$ grows quickly. For example, $170!$ is about $7.26 \times 10^{308}$ and $171!$ overflows. The term $\rho^N$ also grows quickly as $N$ increases. For example if $\rho = 100$, then $\rho^N$ overflows when $N > 154$.

The Erlang-B function, and related functions such as its derivatives, are often calculated by recursive methods. An example of this is the recursion

$$E_1(\rho, N) = A/(1 + A) \quad (2)$$

where

$$A = \rho E_1(\rho, N - 1)/N$$

starting from $E_1(\rho, 0) = 1$. This recursion is commonly used to calculate $E_1(\rho, N)$ for moderately large values of $N$ – asymptotic forms are best used to compute $E_1(\rho, N)$ for large values of $N$. Eqn. (2) has computation requirements $O(N)$ and storage requirements $O(1)$. The recursion works with normalized quantities $E_1(\rho, N)$ and is therefore not subject to numerical problems such as imprecision and overflow.

B. The Erlang-C system

The Erlang-C equation computes the blocking probability for the $M/M/N$ queue where customers who find all $N$ servers busy are not lost, but are delayed in a queue with an infinite number of waiting places. Let $E_2(\rho, N)$ denote the probability that all servers are busy

$$E_2(\rho, N) = \frac{\rho^N}{(1 - \rho/N)^N} \left( \sum_{n=0}^{N-1} \frac{\rho^n}{n!} + \frac{\rho^N}{(1 - \rho/N)N!} \right)^{-1}. \quad (3)$$

Eqn. (3) has the recursive form

$$E_2(\rho, N) = A/(1 + A)$$

where

$$A = \rho E_1(\rho, N - 1)/(N - \rho)$$

starting from $E_1(\rho, 0) = 1$. The Erlang-C formula requires $\rho < N$ else the queue grows without bounds and no equilibrium distribution exists.

C. The Engset system

The Engset $M/M/N/N/K$ system describes a finite population $K$ of customers who arrive individually at the instants of a Poisson process to a service facility consisting of $N \leq K$ identical exponential servers. If all $N$ servers are busy, the arrival is dropped and resubmitted after an exponential delay. The probability that an arrival finds all $N$ servers busy is

$$E_3(\rho, K, N) = \frac{(K - 1)^N}{N!} \sum_{i=0}^{N} \binom{K - 1}{i} \rho^i,$$

which for $K \geq N > 0$ has the recursive form

$$E_3(\rho, K, N) = A/(1 + A)$$

where

$$A = \rho(K - N)E_3(\rho, K, N - 1)/N$$

starting from $E_3(\rho, K - N, 0) = 1$. The storage and computation required to compute $E_3(\rho, K, N)$ is $O(N)$ and $O(K)$ respectively.

D. The $M/M/r/K/K$ queue

The $M/M/r/K/K$ queue is well known as the machine repair model where $K$ machines are subject to failure. The time to failure is exponentially distributed. Failed machines are serviced by $r$ identical repairmen. The time to repair is exponentially distributed. If all $r$ repairmen are busy then the failed machine is queued in a FCFS waiting line. Repaired machines are immediately returned to service.

The $M/M/r/K/K$ queue can be modelled as a closed queueing network with 2 servers and $K$ customers. The machine failure process is modelled by a server with mean service rate $\mu_1(n) = (K - n)/a$ where $0 \leq n \leq K$ is the number of customers (failed machines) at the second server and $a$ is the mean time to failure. The second server models the repair facility and is a multiserver with mean service rate

$$\mu_2(n) = \begin{cases} n/h & 0 \leq n \leq r \\ r/h & r < n \leq K \end{cases}$$

where $h$ in the mean time to repair. The recursive solution of the product form network can immediately be obtained by convolution or by mean value analysis methods: see [5] and the references therein.

III. A RECURSIVE FORM FOR THE $M/M/r/K/K$ DELAY PROBABILITY

The $M/M/r/K/K$ queue when applied in a machine repair context can readily be numerically solved since the machine and repairmen populations are (usually) relatively small. However, numerical problems will arise when this model is applied to compute the performance measures of a switch with a large number of circuits and a large, finite population of subscribers.

The numerical problem cannot be alleviated by using double precision or arbitrary precision arithmetic. The correct approach is to obtain a recursive form for the blocking and delay probabilities. Although a recursive form for the queue length distribution for the $M/M/r/K/K$ queue is well known [5], this is not the case for the blocking delay probability.

Consider Eqn. (12) in [6]: the probability $P_{d\geq u}(K)$ that a call is delayed longer than $u$ is

$$P_{d\geq u}(K) = \beta e^{-ru/h} \sum_{n=r}^{K} P(n, K)(K - n) \sum_{i=0}^{n-r} (ru/h)^{i}/i!$$

where $\beta = h/a$ and

$$N_s(K) = \sum_{n=1}^{K} \alpha(n)P(n, K)$$

is the mean number of calls being served (the carried traffic) where $\alpha(n) = \min(n, r)$, $r$ is the number of servers, $n = (0, 1, \ldots, K)$ represents the state of the system (the number of calls in progress), $P(n, K)$ denotes the probability of the
system being in state $n$, $a$ is the mean inter-arrival time per source and $h$ is the mean call holding time. Eqn. (4) can be written as
\[ P_{d>u}(K) = \beta e^{-x} \sum_{n=r}^{K-1} \frac{P(n, K)}{N_s(K)} (K-n)F(x, n-r) \]
where $x = ru/h$ and
\[ F(x, m) = \sum_{i=0}^{m-1} x^i/i! + x^m/m! = F(x, m-1) + xG(x, m-1)/m \] (5)
for $0 < m \leq K - r$ where $G(x, m) = x^m/m!$ and $F(x, 0) = G(x, 0) = 1$. It is assumed that $\lim_{x \to 0} F(x, m) = 1$.

Next it can be shown that
\[ \sum_{n=r}^{K-1} \frac{K!}{n!} \frac{\alpha(n)}{n} \beta^n (K-n)^! (K-n)!/m! \]
where the invariant measure $D(n, K)$ is given by
\[ D(n, K) = \left\{ \begin{array}{ll} K! \beta^n (K-n)!/m! & 0 \leq n < r \\ K! \beta^n (K-n)!/m! & r \leq n \leq K. \end{array} \right. \] (6)
for $0 \leq n < K$ and
\[ D(K, K) = \beta K D(K-1, K-1)/\min(r, K) \] (7)
for $K > 0$ where
\[ D(0, 0) = 1. \] (8)

IV. A NUMERICALLY STABLE RECURSION TO COMPUTE THE M/M/r/K/K DELAY PROBABILITY

The recursion presented above works with the un-normalised quantities $D(n, K)$ and is subject to numerical underflow. A numerically stable recursion to compute the delay probability $P_{d>u}(K)$ is as follows.
\[ P_{d>u}(K) = e^{-x} \sum_{n=r}^{K-1} \frac{P(n, K)}{N_s(K)} F(x, n-r) \]
where for $0 < n \leq K$
\[ P(n, K) = \frac{D(n, K)}{D(K)} = \frac{\beta K D(n-1, K-1) D(K-1)}{\alpha(n) D(K-1) D(K)} = hT(K)P(n-1, K-1)/\alpha(n). \] (9)

From Little’s law, the throughput $T(K)$ is
\[ T(K) = \frac{K}{a + \sum_{n=1}^{K} \alpha(n) n P(n, K)/T(K)} = \frac{K}{a + h \sum_{n=1}^{K} n P(n-1, K-1)/\alpha(n)}. \]

It remains to compute the probability $P(0, K)$. We have
\[ P(0, K) = f_1(K)/f(K) \]
where
\[ f(K) = \sum_{n=0}^{K} f_1(K-n)f_2(n) \]
\[ f_1(m) = a f_1(m-1)/m \text{ and } f_2(m) = h f_2(m-1)/\alpha(m) \text{ for } 0 < m \leq K \text{ where } f_1(0) = f_2(0) = 1. \]

A numerically stable recursion for $P(0, K)$ is given by
\[ P(0, K) = f_1(K)/f(K) = \frac{f_1(K-1) f_1(K) f(K-1)}{f(K-1) f_1(K-1) f(K)} = P(0, K-1)(a/K)T(K) \]
1: \( P(0, 0) = 1 \)
2: \( \text{for } (k = 1; k \leq K; k = k + 1) \text{ begin} \)
3: \( T(k) = \frac{a + h \sum_{n=k}^{K} nP(n-1, k-1)/\alpha(n)}{k} \)
4: \( \text{for } (n = 1; n \leq k; n = n + 1) \text{ begin} \)
5: \( P(n, k) = hT(k)P(n-1, k-1)/\alpha(n) \)
6: \( \text{end for} \)
7: \( P(0, k) = aT(k)P(0, k-1)/k \)
8: \( \text{end for} \)
9: \( X = 0; F = 1 \)
10: \( \text{for } (n = r; n < K; n = n + 1) \text{ begin} \)
11: \( X = X + F P(n, K-1) \)
12: \( F = xF/(n - r + 1) \)
13: \( \text{end for} \)
14: \( P_{d>u}(K) = e^{-\pi} X \)

Fig. 1. The recursive calculation of the delay probability \( P_{d>u}(K) \)

where \( P(0, 0) = 1 \).

An algorithm to compute the delay probability \( P_{d>u}(K) \) is shown in Fig. 1. The computational and storage requirements of the algorithm in Fig. 1 are \( O(K^2) \). A more efficient memory allocation scheme for storing the values of \( P(n, k) \) reduces the storage requirement to \( O(K) \).

Fig. 2 shows the performance measures of the \( M/M/r/K/K \) queue for \( r = 750 \) servers and \( K = 1000 \) sources. As \( \beta \) increases, the average number \( N_s(K) \) of busy servers approaches 750 (right hand scale) and the probability \( P_d(K) \) of being delayed approaches 1 (left hand scale). The probability \( P_{d>h/10}(K) \) of being delayed for a period of time longer than \( h/10 \) where \( h \) is the mean call holding time also tends to one (left hand scale).

V. CONCLUSIONS

A traditional traffic engineering problem is to quickly and correctly dimension all circuit groups in a national telecommunications network, whether the number of switches be small \((N \sim 10)\) or large \((N \sim 1000)\). To do so requires the accurate and rapid computation of the blocking and delay probabilities to ensure that no excessive end-to-end delay occurs anywhere in the network.

Queueing systems have the advantage, at the expense of a small queueing delay, of queueing blocked calls rather than losing blocked calls. This paper presents a computationally efficient and numerically stable recursion to compute the probability that the queueing delay in an \( M/M/r/K/K \) queue is larger than a given bound. Data networks will benefit from fast computational methods for determining whether end-to-end delays in the network satisfy the national standard.

ACKNOWLEDGEMENTS

Thanks are expressed to V Iversen for reviewing an early version of this paper. Thanks are also expressed to IGK’s 2008 teletraffic class for discussions and references.

REFERENCES


Ian G. Kennedy (MSAIEE) obtained the PhD, MSc (Eng) and BSc (Eng) from the University of the Witwatersrand, Johannesburg, South Africa. He is presently a Senior Research Officer at the University of the Witwatersrand. His current and previous research interests include construction engineering, education, e-education and teletraffic.

Anthony Krzesinski obtained the MSc from the University of Cape Town and the PhD from Cambridge University, England. He is presently a Professor of Computer Science at the University of Stellenbosch. His research interests centre on the performance evaluation of communication networks.