Analysis of a RED Queue: A Singular Perturbation Approach

Eitan Altman, Konstantin Avrachenkov, Balakrishna Prabhu

Analysis of a RED Queue: A Singular Perturbation Approach

E. Altman [a], K. Avrachenkov [a], B. J. Prabhu [b]

[a] *INRIA Sophia-Antipolis, 2004 route des Lucioles, 06902 Sophia-Antipolis, France. Email:{altman, k.avrachenkov}@sophia.inria.fr*

[b] *LAAS-CNRS, 7 Avenue du Colonel Roche, 31077 Toulouse Cedex 4, France Email:bjprabhu@laas.fr*

Abstract

Several Active Queue Management (AQM) techniques for routers in the Internet have been proposed and studied during the past few years. One of the widely studied proposals, Random Early Detection (RED), involves dropping an incoming packet with some probability based on the estimated average queue length at the router. The analytical approaches to obtaining average drop probabilities in a RED enabled queue have been either based on using the instantaneous queue size for calculating the drop probability or have considered averaging with a fluid approximation. In this paper, we use a singular perturbation based approach to analyse a RED enabled queue with drop probabilities based on the estimated average queue size as has been proposed in the standard RED. The singular perturbation approach is motivated by the fact that the instantaneous and the estimated average queue lengths evolve at two different time scales. We present an analytical method to calculate the average queue size and the average drop probability for the non responsive flows. We also provide analytical expressions for the Poisson arrivals and exponential service times case. Our model is derived under several approximations, and is validated through simulations.

---

[1] This work was partially supported by the EURO NGI network of Excellence, by the INRIA's TCP ARC collaboration project and by the *Indo–French Center for Promotion of Advanced Research (IFCPAR)* under research contract number 2900–IT. The work of the third author was carried out while he was with INRIA Sophia-Antipolis.

1 Introduction

Congestion control algorithms are end-to-end algorithms designed to fully utilize the available bandwidth on the links. When used in conjunction with drop tail queues in the routers, oscillatory behaviour and synchronised packet drops have been observed at the queues in the routers [1]. Synchronized increase and decrease of the window of different sources leads to inefficient use of the outgoing link. Active Queue Management (AQM) schemes seek to improve the link utilization by sending congestion signals in anticipation of congestion [2]-[4].

To overcome the synchronised window evolution, the Random Early Detection (RED) algorithm proposes to drop packets at random before congestion actually sets in [2]. Two of the design aims of RED are to accept occasional bursts and to maintain a reasonable average queue length when the system is heavily loaded with the objective of efficient use of the system capacity. Towards this end, the routers maintain a variable corresponding to the exponentially weighted moving average of the instantaneous queue length. An incoming packet is dropped with some probability which is a function of this estimated average queue length. The performance of the RED algorithm depends on the setting of parameters and the traffic characteristics [5]. Several alternatives and modifications have been proposed [4], [6] and [7].

The aim of our study is to use singular perturbation techniques [8]-[9] to provide an analytical expression for computing the drop probabilities and average queue length in a RED enabled queue with averaging as has been proposed in RED. Furthermore, we make use of the transition probability matrices in order to preserve the stochastic nature of the system as opposed to a fluid approximation which is deterministic. The dynamics of the average and the instantaneous queue lengths is modelled as a two dimensional Markov chain, which is then approximated by a non-homogeneous (or level dependent) Quasi-Birth-Death (QBD) process [10]. This model then permits us to obtain the joint steady state distribution of the average and instantaneous queue lengths.

The use of singular perturbation technique is motivated by the fact that two different time scales are involved in the evolution of the instantaneous and the average queue length. Indeed, for very small values of the averaging parameter, the average queue length can be assumed to vary much more slowly compared to the instantaneous queue length. The use of this decomposition allows us to compute the steady state distribution and the desired performance metrics at a reduced complexity. Next, we mention the techniques which have been previously used to analyse RED, and mention the differences with our work.

1.1 Related Literature

Various authors have studied the behaviour of RED, both through analysis and through simulations. In [11], the authors analyse the performance of a RED queue using the instantaneous queue length, instead of the average queue length, to compute the drop probabilities. In [12] and [13], a fluid approximation is used to study the effect of the exponential averaging. In [14], the authors compute the joint distribution of the instantaneous and the averaged queue length of an $M/M/1/K$ queue as a solution of a set of differential equations. An analytical expression for the case of buffer size 1 and 2 is also provided. However, they consider packet drops only due to buffer overflow. The study of [15] uses a time scale decomposition to obtain in the limit a differential equation for the exponentially averaged queue length. They also provide a diffusion approximation to quantify the error due to the ODE approximation. Our method is also based on noting that average and the instantaneous queue lengths evolve at different time scales. However, unlike the ODE approximation obtained in [15], we use a discretized version of that fluid dynamics which has a simple probabilistic interpretation, which allows us then to obtain the joint distribution of the average and the instantaneous queue length (using non-homogeneous QBD). We note that this type of discretization is very frequent in numerical solutions of fluid models and of diffusions, including controlled ones and there is a huge literature that provides conditions under which the stochastic processes related to the discrete model converge (in various senses) to the original fluid model as the discretization step goes to zero, see e.g. Theorems 2.7 and 4.2 in [16] (or the Appendix of [17] as well as [18]).

The rest of the paper is organized as follows. In Section 2 we describe the RED algorithm and present the system model. In Section 3 we present key elements of singular perturbation approach to Markov chains. In Section 4 we apply the singular perturbation approach to compute the joint distribution of the average and instantaneous queue length, and to compute other performance metrics. In Section 5 we compare the results obtained using the analysis and with simulations. Finally, in Section 6 we present the conclusions.

2 System Model

2.1 The RED algorithm

The basic algorithm of RED was proposed in [2] and can be summarized as follows. The router updates an exponentially averaged queue length variable (q_n, where the subscript n denotes the nth packet arrival) at every packet arrival.

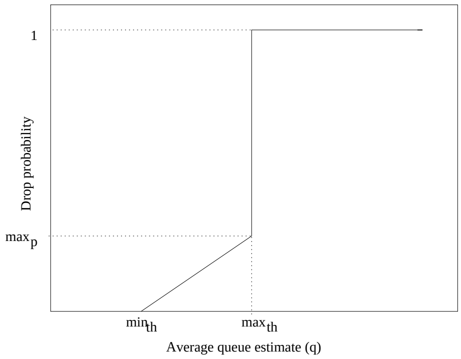

Fig. 1. RED drop probability function

In addition to q_n, the router also keeps a count (c_n) of the number of packets accepted since the last dropped packet. The incoming packet is dropped with a probability, p_d, which is computed according to the following algorithm.

`if` $q_n < min_{th}$ `then`
  $p_d = 0,\ c_n = -1,$
`else if` $min_{th} \leq q_n \leq max_{th}$ `then`
  $p_b = max_p(q_n - min_{th})/(max_{th} - min_{th})$,
  $p_d = \min(1, p_b/(1 - c_n p_b))$,
`else`
  $p_d = 1$,

where max_p (maximum drop probability), min_{th} (minimum threshold for queue), and max_{th} (maximum threshold for queue) are given constants. The variable c_n is set to zero each time an incoming packet is dropped. The estimated average queue length, q_n, is an exponentially weighted moving average of the instantaneous queue length, Q_n, and is computed as follows.

`if` $Q_n == 0$ `then`
  $q_{n+1} = (1 - \alpha)^w q_n$,
`else`

$$q_{n+1} = (1 - \alpha)q_n + \alpha Q_{n+1}, \tag{1}$$

where α is the queue weight, and w is the average number of packets the router could have transmitted during the previous idle period. Figure 1 shows the drop function associated with the basic RED algorithm.

Packets arrive to a RED enabled router with a buffer of size K packets. Let T_n denote the sequence of the interarrival arrival time between the $(n-1)$th and nth arrival. The sequence $\{T_n, n \geq 1\}$ is assumed to be independent and identically distributed with mean λ^{-1}, distribution function $T(x)$, and Laplace-Stieltjes transform $T^*(s)$. The packet service times are assumed to be exponentially distributed with mean μ^{-1}. In the notation of queueing theory, the model corresponds to a $G/M/1/K$ queue with RED queue management scheme.

Assumption 1 *For $min_{th} \leq q_n \leq max_{th}$, an incoming packet is dropped with a probability p_d equal to p_b.*

Remark. In [2], the authors compare two methods of calculating the drop probability, p_d. In the first method, p_d is taken to be equal to p_b, and in the second method p_d is computed as in the algorithm described in Section 2.1. They conclude that, for the same number of average dropped packet, the losses are more spread out when the second method is used. Therefore, by assuming $p_d = p_b$ we can expect to obtain an upper bound on the probability of more than j successive packet losses.

Let Q_n and q_n be the instantaneous and estimated average queue lengths, respectively, just before the n^{th} arrival. The recursive equations for the evolution of the state (q_n, Q_n) are given by

$$Q_{n+1} = \begin{cases} \max(Q_n - D_n, 0) & \text{w.p.} \;\; h(q_n) \quad \text{or if} \;\; Q_n = B; \\ \max(Q_n + 1 - D_n, 0) & \text{w.p.} \;\; 1 - h(q_n), \end{cases} \tag{2}$$

$$q_{n+1} = (1-\alpha) q_n + \alpha Q_{n+1}, \tag{3}$$

where D_n is the number of packet departures during the interarrival time T_{n+1}, and $h(\cdot)$ denotes the RED packet drop probability function, i.e., $p_d = h(q_n)$.

With this formulation, the process $\{(q_n, Q_n), n \geq 0\}$ is a Markov chain. Its analysis is difficult, partly because the support of q is a countable subset of the interval $[0, B]$ where B is the maximum queue size.

We propose an approximation that reduces the support of q to a finite set denoted by $V_m = \{0, ..., mB\}$, where m is some integer. In the following discussion, $\hat{q}_n$ will refer to the discretized estimated average queue length. We assume below that α is sufficiently small such that $\alpha m B < 1$. The recursive equation for $\hat{q}_n$ is given by

$$\hat{q}_{n+1} = (1-\alpha)\hat{q}_n + \alpha m Q_{n+1}, \quad \hat{q}_n \in \{0, 1, ..., mB\}. \tag{4}$$

Then we replace equation (4) by the following approximation

$$\hat{q}_{n+1} = \hat{q}_n + C_n, \tag{5}$$

where

$$C_n = \begin{cases} 1 & \text{w.p. } \alpha(mQ_{n+1} - \hat{q}_n)^+ \ ; \\ -1 & \text{w.p. } \alpha(\hat{q}_n - mQ_{n+1})^+ \ ; \\ 0 & \qquad\qquad \text{otherwise} \ , \end{cases}$$

with $x^+ = \max(x, 0)$.

We note that with this definition, we obtain

$$\begin{aligned} E[\hat{q}_{n+1} - \hat{q}_n | \hat{q}_n, Q_{n+1}] &= 1 \cdot \alpha(mQ_{n+1} - \hat{q}_n)^+ + (-1) \cdot \alpha(\hat{q}_n - mQ_{n+1})^+ \\ &= \alpha(mQ_{n+1} - \hat{q}_n), \end{aligned} \tag{6}$$

so that (4) becomes the mean field of (5). With this approximation the transitions of $\hat{q}_n$ conditioned on Q_n become stochastic as opposed to the deterministic transitions in the original system.

The motivation for such an approximation is that we are interested in obtaining the steady state distribution of the couple $(\hat{q}_n, Q_n)$ which will allow us to study the various performance measures (e.g., average queue length, average drop probability, etc.) of the system. With this approximation we can model the $(\hat{q}_n, Q_n)$ process as a non-homogeneous QBD process for which algorithms are available to efficiently compute the steady state distributions.

In order to model the process $(\hat{q}_n, Q_n)$ as a QBD process, we first obtain the transition matrix of Q_n for a given value of $\hat{q}_n$. That is,

$$A_i = \{a_{jk}\}_i = P(Q_{n+1} = k | Q_n = j, \hat{q}_n = i), \qquad \forall i \in 0, 1, ..., mB. \tag{7}$$

The matrix A_i is a $(B+1) \times (B+1)$ matrix. Using equation (2), the probability a_{jk} can be obtained as follows.

$$a_{jk} = \begin{cases} (1-h(i))d_0 & k = \min(j+1, B); \\ (1-h(i))d_{j-k+1} + h(i)d_{j-k} & 0 < k \leq j; \\ (1-h(i))\sum_{l=\min(j+1,B)}^{\infty} d_l + h(i)\sum_{l=\min(j,B)}^{\infty} d_l & k = 0, \end{cases} \tag{8}$$

where

$$d_l = \int_0^\infty \exp(-\mu x) \frac{(\mu x)^l}{l!} dT(x) \tag{9}$$

is the average probability of l departures between two arrivals.

Next, we obtain the transition matrix for $\hat{q}_n$ conditioned on given values of Q_{n+1} and $\hat{q}_n$. Let C_i^+, C_i^-, C_i^0 denote matrices given by

$$\begin{aligned}
C_i^+ := \{c_{jj}\} &= p(\hat{q}_{n+1} = i+1 | Q_{n+1} = j, \hat{q}_n = i) \\
&= \alpha(m \cdot j - i)^+, \\
C_i^- := \{c_{jj}\} &= p(\hat{q}_{n+1} = i-1 | Q_{n+1} = j, \hat{q}_n = i) \\
&= \alpha(i - m \cdot j)^+, \\
C_i^0 := \{c_{jj}\} &= p(\hat{q}_{n+1} = i | Q_{n+1} = j, \hat{q}_n = i) \\
&= 1 - \alpha(m \cdot j - i)^+ - \alpha(i - m \cdot j)^+, \\
& \quad \forall i \in \{0, 1, ..., mB\} \qquad \forall j \in \{0, 1, ..., B\}.
\end{aligned}$$

C_i^+i, C_i^-, and C_i^0 are $(B+1) \times (B+1)$ matrices which denote the transition probability of $\hat{q}_{n+1} = \hat{q}_n + 1$, $\hat{q}_{n+1} = \hat{q}_n - 1$, and $\hat{q}_{n+1} = \hat{q}_n$, respectively, when $Q_{n+1} = j$. We note that C_i's are diagonal matrices which satisfy the equality

$$C_i^+ + C_i^- + C_i^0 = I. \tag{10}$$

Let $\tilde{P}$ denote the transition probability matrix for the two dimensional Markov chain $(\hat{q}_n, Q_n)$ with $\tilde{\pi}(\alpha)$ as its stationary distribution. Using the identity

$$\begin{aligned}
p(\hat{q}_{n+1}, Q_{n+1} | \hat{q}_n, Q_n) &= p(\hat{q}_{n+1} | \hat{q}_n, Q_n, Q_{n+1}) \cdot p(Q_{n+1} | \hat{q}_n, Q_n) \\
&= p(\hat{q}_{n+1} | \hat{q}_n, Q_{n+1}) \cdot p(Q_{n+1} | \hat{q}_n, Q_n),
\end{aligned}$$

we can write $\tilde{P}$ as

$$\tilde{P} = \begin{bmatrix}
A_0(C_0^- + C_0^0) & A_0 C_0^+ & 0 & \dots & \dots \\
A_1 C_1^- & A_1 C_1^0 & A_1 C_1^+ & 0 & \dots \\
0 & A_2 C_2^- & A_2 C_2^0 & A_2 C_2^+ & \dots \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & \dots & \dots & A_{mB} C_{mB}^- & A_{mB}(C_{mB}^0 + C_{mB}^+)
\end{bmatrix}. \tag{11}$$

The diagonal rows entries of $\tilde{P}$ are matrices which give the transition probability matrix of Q for a given value of $\hat{q}_n$, and $\hat{q}_{n+1} = \hat{q}_n$ whereas the off-diagonal entries give the transition probability matrix of Q for a given value of $\hat{q}_n$ and $\hat{q}_{n+1} = \hat{q}_n + 1$, or $\hat{q}_{n+1} = \hat{q}_n - 1$. Using equation(10), we can rewrite $\tilde{P}$ as

$$\tilde{P} = P + PC, \tag{12}$$

where P is given by

$$P = \begin{bmatrix} A_0 & 0 & \dots & 0 \\ 0 & A_1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & A_{mB} \end{bmatrix},$$

and C is given by

$$C = \begin{bmatrix} -C_0^+ & C_0^+ & 0 & \dots \\ C_1^- & -(C_1^- + C_1^+) & C_1^+ & \dots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & C_{mB}^- & -C_{mB}^- \end{bmatrix}.$$

We note that $\tilde{P}$ is a function of the averaging parameter α. In [19], the recommended value of α is of the order of 0.002. This suggests that the averaging parameter is, in general, small. This assumption on α allows us to obtain the steady state probabilities of $\tilde{P}$ using a singular perturbation approach which we outline below and which is computationally less expensive than algorithms to compute the invariant vectors of matrices.

3 Review of Singular Perturbation Approach

Let $P \in R^{n \times n}$ be a transition stochastic matrix representing transition probabilities in a Markov chain. Suppose that the structure of the underlying Markov chain is aperiodic. Let $P^* = \lim_{t \to \infty} P^t$ which is well-known to exist for aperiodic process. The matrix P^* is called ergodic projection. In the case when the process is also ergodic, P^* has identical rows, each of which is the stationary distribution of P, denoted by π. Let Y be the deviation matrix of P which is defined by $Y = (I - P + P^*)^{-1} - P^*$. It is well known that Y exists and it is the unique matrix satisfying

$$Y(I - P) = (I - P)Y = I - P^*,$$

and

$$P^*Y = YE = 0,$$

(where E is a matrix full of 1's) making it the group inverse of $I - P$. Finally, $Y = \lim_{T\to\infty} \sum_{t=0}^{T}(P^t - P^*)$.

We consider (linear) perturbations of the matrix P and their impact on the structure of the process. Specifically, for a scalar ε, $0 < \varepsilon < \varepsilon_{max}$, and for some zero rowsum matrix Q, we look at the set of perturbed stochastic matrices $P(\varepsilon) = P + \varepsilon Q$, which are assumed to be unichain for any ε in the above mention region. A unichain Markov chain is a Markov chain which has a unique stationary distribution. We emphasize that unichain assumption is not assumed with regard to $P = P(0)$. Actually, the case when $P(0)$ contains some unrelated chains (with or without transient states) is our main focus. Here we survey some results on series expansions for $\pi(\varepsilon)$ and $Y(\varepsilon)$, which denote the stationary distribution and the deviation matrix, respectively, of $P(\varepsilon)$ for $0 < \varepsilon < \varepsilon_{max}$ and consider their relationship to the corresponding entities in the unperturbed Markov chain for $P = P(0)$.

Let us concentrate on so-called Nearly Completely Decomposable (NCD) case, in which the state space under the unperturbed process is decomposed into a number of unichain Markov chains. Specifically, let $P(0) \in R^{n\times n}$ be a stochastic matrix representing transition probabilities in a completely decomposable Markov chain. By the latter we mean that there exists a partition Ω of te state space into $p, p \geq 2$, subsets $\Omega = \{I_1, ..., I_p\}$ each of which being an ergodic class. We assume that the order of the rows and of the columns of P is compatible with Ω, i.e., for p stochastic matrices, $P_{I_1}, ..., P_{I_p}$,

$$P = \begin{pmatrix} P_{I_1} & 0 & \cdots & 0 \\ 0 & P_{I_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & P_{I_p} \end{pmatrix}.$$

Note that we assume above that P_{I_k} represents a unichain Markov chain.

For small values of ε, $P(\varepsilon) = P + \varepsilon Q$ is called nearly completely decomposable or sometimes nearly uncoupled.

Let us define the limiting stationary distribution $\pi^{(0)}$ as

$$\pi^{(0)} = \lim_{\varepsilon\to 0} \pi(\varepsilon).$$

For any subset $I \subset \Omega$, let

$$\kappa_I = \sum_{i\in I} \pi_i^{(0)}.$$

Also, let γ_I be the subvector of $\pi^{(0)}$ corresponding to subset I rescaled so as its entry-sum is now one. Then, γ_I is the unique stationary distribution of

P_I. Note that computing γ_I is relatively easy as only the knowledge of P_I is needed.

Next we define the matrix $S \in R^{p\times p}$ which is usually referred to as the aggregate transition matrix. Each row, and likewise each column in S corresponds to a subset in Ω. Then, for subsets I and J, $I \neq J$, let

$$S_{IJ} = \sum_{i\in I}(\gamma_I)_i \sum_{j\in J} Q_{ij},$$

and let

$$S_{II} = 1 + \sum_{i\in I}(\gamma_I)_i \sum_{j\in J} Q_{ij} = 1 - \sum_{J\neq I} S_{IJ}.$$

Without loss of generality, assume that S_{II} is non-negative for all subsets I and hence S is easily seen to be a stochastic matrix. Note that the matrix S can be divided by any constant and ε can be multiplied by this constant leading to the same $n \times n$ transition matrix. Taking this constant small enough guarantees the stochasticity of S and hence this is assumed without loss of generality. In particular, the stationary distribution of S is invariant with respect to the choice of this constant. Moreover, by assumption S is a unichain matrix.

Often it is convenient to express the aggregate transition matrix S in matrix terms. Namely, let $V \in R^{p\times n}$ be such that its i-th row is full of zeros except for γ_{I_i} at the entries corresponding to subset I_i, and where $W \in R^{n\times p}$ is such that its j-th column is full of zeros except for 1's in the entries corresponding to subset I_j. Note that $VW \in R^{p\times p}$ is the identity matrix. Moreover, V and W correspond to bi-orthonormal sets of eigenvectors of $P(0)$ belonging to the eigenvalue 1, V as left eigenvectors and W as right eigenvectors. Then, we can write

$$S = I + VQW.$$

The aggregate stochastic matrix S represents transition probabilities between subsets which in this context are sometimes referrred to as macro-states. However, although the original process among states is Markovian, this is not necessarily the case with the process among macro-states. Yet, as the following Theorem indicates, much can be learned about the original process from the analysis of the aggregate matrix.

Theorem 1 *Let the perturbed Markov chain be nearly completely decomposable. The stationary distribution $\pi(\varepsilon)$ admits a Maclaurin series expansion in a punctured neighborhood of zero. Namely, for some vectors $\{\pi^{(m)}\}_{m=0}^{\infty}$ with $\pi^{(0)}$ being a probability vector satisfying $\pi^{(0)} = \pi^{(0)}P(0)$, and for some zerosum vectors $\pi^{(m)}$, $m \geq 1$,*

$$\pi(\varepsilon) = \sum_{m=0}^{\infty} \varepsilon^m \pi^{(m)}.$$

Moreover, for $I \in \Omega$, $\pi_I^{(0)} = \kappa_I \gamma_I$. Also, the sequence $\{\pi^{(m)}\}_{m=0}^{\infty}$ is geometric, *i.e., for some square matrix U, $\pi^{(m)} = \pi^{(0)}U$ for any $m \geq 0$. Actually,*

$$U = QY(0)(I + QWDV),$$

where D is the deviation matrix of the aggregate transition matrix S. Alternatively,

$$U = QY^{(0)},$$

where $Y^{(0)}$ is the first regular term of the Laurent series expansion for $Y(\varepsilon)$ (see also the next theorem). Finally, the validity of the series expansion holds for any ε, $0 \leq \varepsilon < \min\{\varepsilon_{max}, \rho^{-1}(U)\}$ where $\rho(U)$ is the spectral radius of U.

Recall that for ε, $0 \leq \varepsilon < \varepsilon_{max}$, we denote by $Y(\varepsilon)$ the deviation matrix of $P(\varepsilon)$. This matrix is uniquely defined and the case $\varepsilon = 0$ is no exception. Yet, as we see shortly, there is no continuity of $Y(\varepsilon)$ at $\varepsilon = 0$. We note that $Y(0)$ has the same shape as P has, namely

$$Y(0) = \begin{pmatrix} Y_{I_1} & 0 & \cdots & 0 \\ 0 & Y_{I_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & Y_{I_p} \end{pmatrix},$$

where Y_{I_i} is the deviation matrix of P_{I_i}, $1 \leq i \leq p$.

Theorem 2 *In the case of NCD Markov chains, the matrix $Y(\varepsilon)$ admits a Laurent series expansion in a punctured neighborhood of zero with the order of the pole being exactly one. Namely, for some matrices $\{Y^{(m)}\}_{m=-1}^{\infty}$ with $Y^{-1} \neq 0$, we have*

$$Y(\varepsilon) = \frac{1}{\varepsilon}Y^{(-1)} + Y^{(0)} + \varepsilon Y^{(1)} + \varepsilon^2 Y^{(2)} + \ldots$$

for $0 < \varepsilon < \varepsilon_{max}$. Moreover,

$$Y^{-1} = WDV,$$

and in a component form,

$$Y_{ij}^{(-1)} = D_{IJ}(\gamma_J)_j, \quad i \in I, \quad j \in J.$$

An interested reader can find more details and references on singularly perturbed Markov chains in the survey [9].

4 Analysis

To use the singular perturbation approach outlined in the previous section, we first note that P is the matrix containing the transition probability matrices of the instantaneous queue length for a given value of the estimated average queue length when the averaging parameter $\alpha = 0$. Thus, P has mB ergodic classes each corresponding to a particular value of $\hat{q}$. The stationary distribution of each of these classes can be computed separately. This system of decomposable Markov chains is the unperturbed system which can be analyzed separately for different values of $\hat{q}$. As noted in the previous section, in general, the stationary distribution of the unperturbed system (i.e., $\alpha = 0$) is not the same as the stationary distribution of the original system when $\alpha \to 0$ [20]. This motivates us to use singular perturbation technique to determine the stationary distribution of our system when $\alpha \to 0$.

Later, we also show that the distributions of the instantaneous queue length and the average queue length can be computed for the general case (i.e. $\alpha \nrightarrow 0$). We note that the use of the approximation given in (5) has resulted in a level dependent Quasi-Birth-Death (QBD) type of transition matrix which is as given in (11). We can now use the algorithm, given in [21], for computing the steady state probability vector for level dependent QBD.

4.1 Limiting case

In this subsection we obtain the steady state probability vector of $\tilde{P}$ when $\alpha \to 0$. We note that all the elements of the matrices C_i^+ and C_i^- have α as a common multiple and thus we can replace C_i^+ by αD_i^+ and C_i^- by αD_i^- where D_i^+ and D_i^- have elements independent of α. Thus we can rewrite equation (12) as

$$\tilde{P} = P + \alpha P D. \tag{13}$$

With $\varepsilon = \alpha$ and $Q = PD$ we recover the standard formulation for the singularly perturbed Markov chains. The singular perturbation approach allows us to find the stationary distribution of $\tilde{P}$ as $\lim_{\alpha \to 0} \tilde{\pi}(\alpha)$.

Let π_i denote the stationary distribution of A_i, and let V be a $(mB + 1) \times (mB+1)(B+1)$ matrix such that in the ith row the entries, corresponding to the columns $\hat{q}_n = i$, are given by the rows of π_i, and is zero elsewhere. π_i is the stationary distribution vector of the unperturbed Markov chain corresponding to the instantaneous queue length when packets are dropped with a probability depending on $h(i)$. Let W be a $(mB+1)(B+1) \times (mB+1)$ matrix such that in the ith column the entries corresponding to the rows $\hat{q}_n = i$ are 1, and is zero elsewhere.

$$V = \begin{bmatrix} \pi_0 & \underline{0} & \dots & \underline{0} \\ \underline{0} & \pi_1 & \dots & \underline{0} \\ \vdots & \vdots & \ddots & \vdots \\ \underline{0} & \dots & \dots & \pi_{mB} \end{bmatrix}, \qquad W = \begin{bmatrix} \underline{1}' & \underline{0} & \dots & \underline{0} \\ \underline{0} & \underline{1}' & \dots & \underline{0} \\ \vdots & \vdots & \ddots & \vdots \\ \underline{0} & \dots & \dots & \underline{1}' \end{bmatrix},$$

where $\underline{0}$ and $\underline{1}$ are $(B+1)\times 1$ vectors of 0 and 1, respectively.

Let S denote the generator matrix of the aggregated Markov chain. When α goes to 0, the stochastic sequence $\hat{q}_n$ converges weakly to a Markov chain induced by the aggregated transition matrix. Then S is given by

$$S = VPDW.$$

Since π_i is the stationary distribution of A_i, VP reduces to V, i.e.,

$$S = VDW. \tag{14}$$

Denote

$$\begin{aligned} f_i^+ &= \pi_i D_i^+ \underline{1}' && (15)\\ &= \sum_{j=\lceil \frac{i}{m} \rceil}^{B} \pi_i(j)(mj-i). && (16)\\ f_i^- &= \pi_i D_i^- \underline{1}' && (17)\\ &= \sum_{j=0}^{\lfloor \frac{i}{m} \rfloor} \pi_i(j)(i-mj). && (18)\\ & && (19) \end{aligned}$$

where $\pi_i(j)$ is the j^{th} element of π_i.

Using equations (15) and (17) we can rewrite S as

$$S = \begin{bmatrix} -f_0^+ & f_0^+ & 0 & \dots \\ f_1^- & -(f_1^- + f_1^+) & f_1^+ & \dots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & f_{mB}^- & -f_{mB}^- \end{bmatrix}.$$

The stationary distribution of this generator matrix can be obtained by solving $\gamma S = 0$. The stationary distribution, γ, is given by

$$\gamma := \{\gamma(i)\} = \gamma(0)\frac{\prod_{j=0}^{i-1} f_j^+}{\prod_{j=1}^{i} f_j^-}, \qquad i \in \{0, 1, ..., mB\}, \tag{20}$$

where

$$\gamma(0) = \left(\sum_{i=0}^{mB} \frac{\prod_{j=0}^{i-1} f_j^+}{\prod_{j=1}^{i} f_j^-} \right)^{-1}.$$

Proposition 1 *Let $\tilde{\pi}(\alpha)$ be the stationary distribution of $\tilde{P}$ for a given value of α, and let π_i denote the stationary distribution of A_i. Then, the limiting stationary distribution of $\tilde{P}$, $\tilde{\pi}(0)$ is given by*

$$\tilde{\pi}(0) = [\gamma_1 \pi_1 \; \gamma_2 \pi_2 \; ... \; \gamma_{mB} \pi_{mB}], \tag{21}$$

where γ_i is given in (20).

Proof: Follows from Theorem 1.

We note that the states $q > m \cdot max_{th}$ are transient and the steady state probability of $q > m \cdot max_{th}$ is 0. This can be shown as follows. From equation (16) we have

$$f_i^+ = \sum_{j=\lceil \frac{i}{m} \rceil}^{B} \pi_i(j)(mj - i).$$

Since the packet drop probability is 1 for $q \geq m \cdot max_{th}$, the steady state probability, $\pi_i(j)$, becomes

$$\pi_i(j) = \begin{cases} 1 & j = 0; \\ 0 & j > 0; \end{cases} \qquad \forall i \geq m \cdot max_{th}. \tag{22}$$

This is true since all the packets arriving at the queue would be dropped and the queue would be empty. Since, $\pi_i(j)$ is as given in equation (22), and $i \geq m \cdot max_{th} > 0$, we get that $f_i^+ = 0, \quad \forall i \geq m \cdot max_{th}$. From equation (18) we get $f_i^- > 0 \quad \forall i$. From equation (20) it follows that $\gamma(i) = 0 \quad \forall i \geq m \cdot max_{th}$.

Remark. The stationary probability $\pi_i(j)$ is the steady state probability of an arrival finding j customers in a $G/M/1/K$ queue. These probabilities can be computed recursively. We note that in the limit $\alpha \to 0$, the instantaneous queue moves on a much faster time scale than the estimated average queue length, and hence approaches stationarity. Therefore, if we had exponential interarrival times and general service times (i.e., $M/G/1/K$) instead of general interarrival times and exponential interarrival times (i.e., $G/M/1/K$) we could use the steady state distribution of the $M/G/1/K$ system to compute the performance measures. In the limiting case, the performance measures of both $G/M/1/K$ and $M/G/1/K$ queues with RED queue management scheme could be obtained using this approach. However, for the general case ($\alpha \nrightarrow 0$), using our method, we can only obtain the performance measures of the $G/M/1/K$ queue with RED queue management scheme.

If the arrival process to the queue is Poisson with rate λ and the service times are exponentially distributed with mean μ^{-1}, then the stationary distribution of A_i, π_i, is given by

$$\pi_i := \{\pi_i(j)\} = \pi_i(0)\rho_i^j, \tag{23}$$

where

$$\pi_i(0) = \left(\sum_{j=0}^{B} \rho_i^j\right)^{-1},$$

and $\rho_i = \lambda \cdot (1 - h(\hat{q}_n = i)) \cdot \mu^{-1}$. Here $h(\cdot)$ is the given drop function.

From the above analysis, using equation (20), we are able to obtain the stationary distribution of the average queue length variable, q. We use the PASTA property which states that the arrivals see this average queue length distribution. Hence, we can calculate the average packet drop probability as

$$P_{avg} = \sum_{i=0}^{mB} \gamma(i)h(i), \tag{24}$$

For batch arrivals, we make the assumption that every packet in a batch is dropped with same drop probability. With this assumption, we can write the probability of dropping at least one packet in batch as

$$P_{avg} = \sum_{i=0}^{m(B-N)} (1 - (1-\gamma(i))^N) \cdot h(i) + \sum_{i=m(B-N)+1}^{mB} h(i), \tag{25}$$

4.2 Approximate analysis

We present an approximate analysis which allows us to obtain the mean value of q without having to find the distribution.

The equilibrium value of q, q^*, can be obtained by solving

$$q^* = mE(Q(q^*)), \tag{26}$$

where $E(Q(q^*))$ is the expected queue length when we drop with constant probability corresponding to q^*.

Using equations (16) and (18) we get

$$f_i^+ - f_i^- = m\sum_{j=0}^{B} j\pi_i(j) - i = mE(Q_i) - i, \tag{27}$$

where $E(Q_i)$ is the expected queue length in the unperturbed Markov chain with $\hat{q}_n = i$. The term $f_i^+ - f_i^-$ can be thought of as an indicator for transitions of the average queue length, q. We note that in states where $mE(Q_i) > i$, $f_i^+ - f_i^-$ is positive, hence indicating a tendency to increase q. Similarly, when $mE(Q_i) > i$, $f_i^+ - f_i^-$ is negative, hence indicating a tendency to decrease q. The equilibrium value of q will then occur at the point where $f_i^+ - f_i^- = 0$. Hence, q^* can be obtained by solving (26). For the existence and uniqueness of the solution, we assume that the drop probability function (this is the case in the gentle variant of RED [22]) is a continuous non-decreasing function, $h(\hat{q})$, on the estimated average queue length. The effective offered load, ρ^*, for a given equilibrium value of q^* is then given by

$$\rho^* = \rho(1 - h(q^*)). \tag{28}$$

Since ρ^* is a continuous non-increasing function on q^*, we can express q^* as

$$q^* = h^{-1}\Big(1 - \frac{\rho^*}{\rho}\Big) := g(\rho^*). \tag{29}$$

We note that g is continuous non-increasing function on ρ^*. Since the average queue length, $E(Q(\rho^*))$, in the stationary regime is a continuous non-decreasing function of the offered load, then we can say that $E(Q(q^*))$, is a continuous non-increasing function of q^*. As q^* and $EQ(\cdot)$ the map to the same interval i.e., $[0, B]$, we can say that the solution to equation (26) exists and is unique.

Remark. The fixed point equation was also obtained in [15] as the stationary solution of an ordinary differential equation.

4.3 General Case

In section 4.1 we presented an analysis for the limiting case $\alpha \to 0$. In this section we outline an algorithm with which we can compute the steady state probabilities when $\alpha \nrightarrow 0$. However, α still has to satisfy the condition $\alpha m B < 1$.

The transition probability matrix of the two dimensional Markov chain $(\hat{q}_n, Q_n)$ given in equation (11) can be written as

$$\tilde{P} = \begin{bmatrix} L_0 & F_0 & 0 & \dots & \dots \\ B_1 & L_1 & F_1 & 0 & \dots \\ 0 & B_2 & L_2 & F_2 & \dots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & B_{mB} & L_{mB} \end{bmatrix}, \tag{30}$$

where B_i, L_i, F_i denote the backward, local, and forward transition matrices, respectively. This is the transition matrix for a level dependent QBD process. Let the $\tilde{\pi}$ be stationary distribution of $\tilde{P}$. Let $\tilde{\pi} = [\tilde{\pi}_0 \tilde{\pi}_1 ... \tilde{\pi}_{mB}]$, where $\tilde{\pi}_i$'s are vectors of size $1 \times B$. $\tilde{\pi}_i$ is the steady state probability vector of Q for $q = i$. $\tilde{\pi}$ can be found using the following algorithm [21].

(1) Compute the S_i matrices using the following recursion

$$\begin{aligned} S_0 &= L_0 \\ S_i &= L_i + B_i(I - S_{i-1})^{-1} F_{i-1}, \qquad 1 \leq i \leq mB. \end{aligned}$$

(2) $\tilde{\pi}_{mB}$ is computed by solving

$$\begin{aligned} \tilde{\pi}_{mB} &= \tilde{\pi}_{mB} S_{mB}, \\ \tilde{\pi}_{mB} \cdot \underline{1} &= 1. \end{aligned}$$

(3) $\tilde{\pi}_i$, $0 \leq i \leq mB$ is computed using the recursion

$$\tilde{\pi}_i = \tilde{\pi}_{i+1} B_{i+1} (I - S_n)^{-1},$$

(4) $\tilde{\pi}$ is found by normalization

$$\tilde{\pi} = \frac{\tilde{\pi}}{\tilde{\pi} \cdot \underline{1}}$$

5 Numerical results

In this section we present the results obtained through the analysis as described in the previous section and the results obtained through simulations. Our aim to see the effect of the approximations that we made in the model and to verify that these effects are indeed negligible. The arrival process is assumed to be Poisson batch arrival process with rate λ and fixed batch sizes, N. The service times are assumed to be exponential with mean $1/\mu$. The offered load, ρ, is defined as $\rho = \frac{\lambda N}{\mu}$.

Remark. The validity of exponential interarrival time has been widely debated. Various authors have shown that for different scenarios the traffic is self-similar and not Poissonian [23]-[25]. However, in [26], the authors argue that under heavy load conditions the traffic tends towards a Poisson process. In a recent article [27], the main findings suggest that, when the level of multiplexing is large, packet arrivals appear Poisson at small time scales, non-stationary at medium time scales, and long-range dependent at large time scale. Thus, as a preliminary study, we present some results related to the performance of a RED queue when the interarrival time is exponentially distributed.

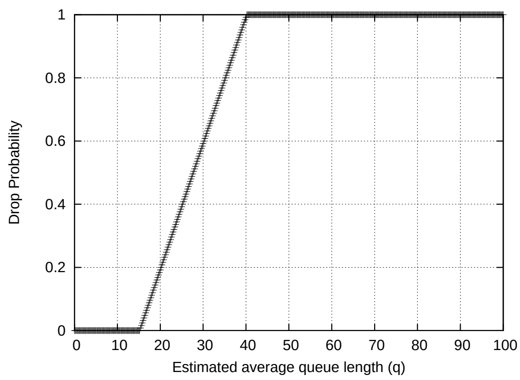

Fig. 2. RED drop probability function

The analytical results were obtained by numerically calculating the values using MATLAB. The buffer length, B, is taken to be 100 whereas the discretization parameter, m, is taken to be 5. The drop probability function is taken to be of the form

$$P_{drop} = \begin{cases} 0, & \hat{q} < min_{th} \\ \frac{\hat{q} - min_{th}}{max_{th} - min_{th}}, & min_{th} \leq \hat{q} \leq max_{th} \\ 1, & \hat{q} > max_{th} \end{cases} \tag{31}$$

We note that all the values of min_{th}, max_{th}, and $\hat{q}$ are scaled by a factor m in the analysis in order to discretize the estimated average queue size. The unscaled values for min_{th} and max_{th} are 15 and 40, respectively, and the drop probability function versus the unscaled estimated average queue length is as shown in figure 2. We shall use the average drop probability and the probability of at least one drop in a burst as performance measures. The simulations were performed using C, and approximately 10^8 packets were generated during the simulations.

5.1 Limiting case

First, we present the results for the limiting case. In the simulations, the averaging parameter, α, was taken to be 0.0001. Figures 3 shows the analytical as well as the simulation results for average packet drop probability as a function of the offered load for different values of burst sizes. There is a fairly good correspondence between the analytical and simulation results.

In [11] it has been argued that the number of drops becomes infinite with positive probability as $\alpha \to 0$. However, for the drop function as shown in Figure 1, the number of consecutive drops appears to follow a geometric behaviour for small values of the number of consecutive drops. We plot the probability

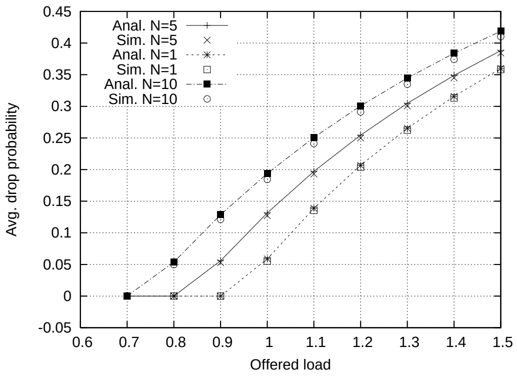

Fig. 3. Average packet drop probability versus offered load

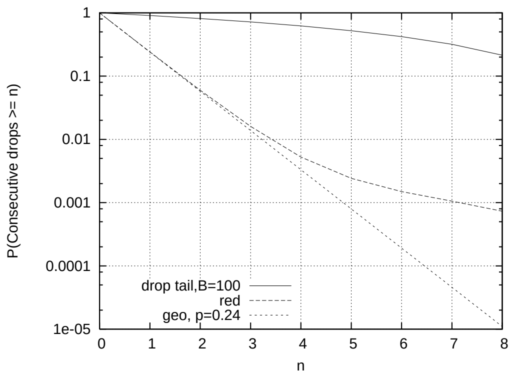

Fig. 4. Probability of greater than or equal to n consecutive drops. $\rho = 1.1$. Burst size = 10.

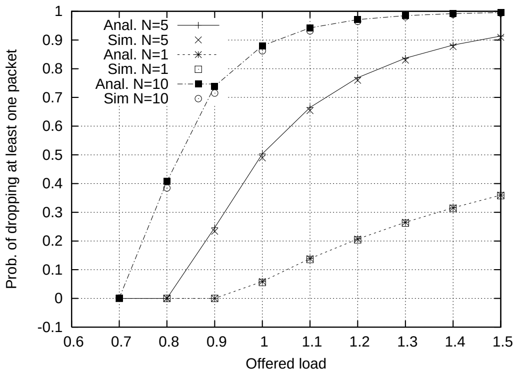

Fig. 5. Probability of drop of at least one packet in a burst versus offered load

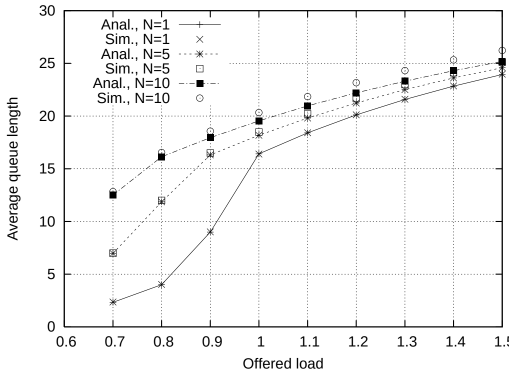

Fig. 6. Average queue length versus offered load

that there are greater than or equal to n consecutive drops as a function of n for a drop tail and a RED queue in Figure 4. The decay rate of the tail is much faster in the case of a RED queue when compared to that of a drop tail queue. For small values of n, the decay rate is close to p^n, where p is the average packet drop probability. Most of the mass can be seen to be concentrated within small values of n and, so the geometric approximation appears to be a reasonable one. A continuous drop function ensures that the drops are independent and, thus, reduces the synchronisation of drops for different flows. This argument supports using gentle RED. The tail drop probability during this simulation scenario was 2×10^{-4} which also suggests that synchronised losses are reduced.

As mentioned above, the probability of dropping consecutive packets becomes independent when the drop function is continuous. This suggests that the probability of drops in a burst is binomially distributed. The probability of at least one drop in a burst can thus be approximated by $1 - (1 - p)^N$, where p is the average drop probability and N is the burst size. This behaviour is seen in Figure 5 in which the probability of dropping at least one packet in a burst is plotted as a function of the offered load.

Next, we validate the approximate analysis for obtaining the average queue length. In order to obtain the average queue length we need to solve the fixed point equation as given in equation (26). For example, the average queue length as observed by a incoming packet in a $M/M/1$ queue with finite buffer B is given by

$$E(Q_{q^*}) = \frac{\rho_{q^*}}{(1-\rho_{q^*})} - (B+1)\frac{\rho_{q^*}^{B+1}}{(1-\rho_{q^*}^{B+1})}, \tag{32}$$

where

$$\rho_{q^*} = \rho\left(1 - \frac{q^* - min_{th}}{max_{th} - min_{th}}\right).$$

We can now obtain the equilibrium value of average queue length by solving

$$q^* = \frac{\rho_{q^*}}{(1-\rho_{q^*})} - (B+1)\frac{\rho_{q^*}^{B+1}}{(1-\rho_{q^*}^{B+1})}. \tag{33}$$

Similarly, for the batch arrivals we use the expression for the infinite buffer case as an approximation. The equilibrium value for batch arrivals is obtained by solving

$$q^* = \frac{\rho_{q^*}}{(1-\rho_{q^*})}\frac{N+1}{2}. \tag{34}$$

For the values of min_{th}, max_{th}, and B assumed in this section, we plot in Fig. 6 the average queue as obtained from solving equations (33) and (34) numerically using MATLAB and the results obtained through simulations.

5.2 General Case

Next, we consider the case when $\alpha \nrightarrow 0$. In the rest of this section we assume that the discretization parameter, m, is 1. The drop function is the same as in the previous section. We need that $\alpha m B$ be less than 1. Thus, α has to be chosen such that $\alpha < 0.01$ in order to use the QBD algorithm. We choose $\alpha = 0.009$ which is close to 0.01. In Figure 7 and Figure 8 we plot the average drop probability and the average queue length, respectively, as a function of the offered load. We note that the behaviour is similar to that observed in the limiting case.

In order to see the effect of choosing an averaging parameter which is not necessarily small, we plot in Figure 9 the distribution function of the average queue length, q, using the QBD analysis, simulations, and the singular perturbation (SP) analysis. The offered load was 1.2 and the burst size was taken to be 10. As can be observed from the figure, the effect of increasing a is to

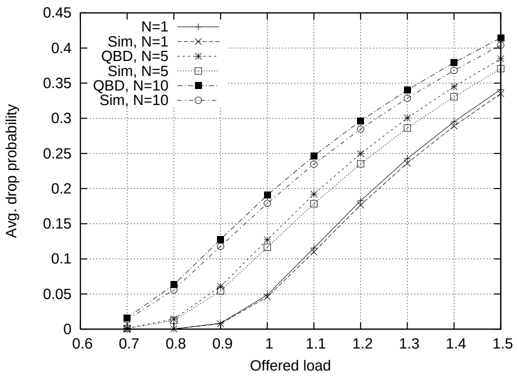

Fig. 7. Average drop probability versus offered load.

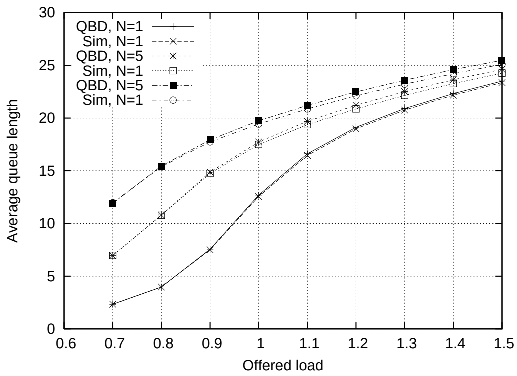

Fig. 8. Average queue length versus offered load.

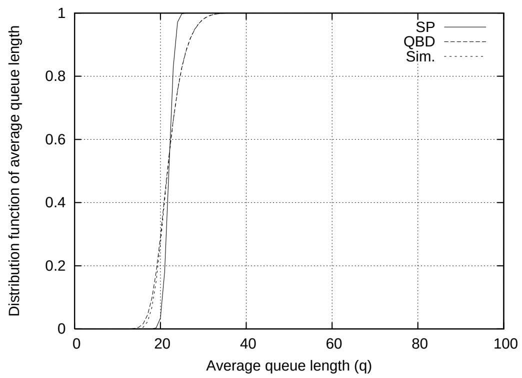

Fig. 9. Distribution function of the average queue length. $N = 10$ and offered load is 1.2. $a = 0.009$.

increase the variance of q. We cannot, however, reduce the variance of q by taking the limit. There is a limiting distribution of q (as was seen from the limiting case) and hence a limiting variance which is different from 0.

6 Conclusions

We used a singular perturbation based approach to find the limiting stationary distribution of the average queue length and the packet drop probability in a RED enabled queue. For Poisson arrival and exponential service times the limiting stationary distribution was given as closed form expressions. The equilibrium point of the average queue length was found to be the fixed point solution of a function depending on the average queue length of a system with no averaging. The analytical results were observed to match fairly accurately with results obtained through simulations. We also showed that by using a continuous drop function we can reduce the number of consecutive losses thereby avoiding synchronised losses. We showed that for the case when the averaging parameter is not small, the stationary distribution can be computed using an algorithm for level dependent QBDs. However, for small values of the averaging parameter, the computational complexity of using this algorithm was much greater than that by using the singular perturbation approach. In this work, we considered scenario in which packet generation process is independent of packet drops. Future work will involve the study of a RED queue with closed loop TCP traffic sources.

References

[1] L. Zhang and D. Clark. Oscillating Behavior of Network Traffic: a Case Study Simulation. *Internetworking:research and experience*, 1(2):101–112, 1990.

[2] S. Floyd and V. Jacobson. Random Early Detection gateways for Congestion Avoidance. *IEEE/ACM Transactions on Networking*, 1(4):397–413, August 1993.

[3] R. Pan, B. Prabhakar, and K. Psounis. CHOKe: A Stateless Active Queue Management Scheme for Approximating Fair Bandwidth Allocation. In *Proc. of the IEEE INFOCOM*, 2000.

[4] W. Feng, D. Kandlur, D. Saha, and K. Shin. The Blue Queue Management Algorithms. *IEEE/ACM Transactions on Networking*, 10(4):249–264, August 2002.

[5] M. Christiansen, K. Jeffay, D. Ott, and F.D. Smith. Tuning RED for Web Traffic. *IEEE/ACM Transactions on Networking*, 9(3):249–264, June 2001.

[6] S. Athuraliya, V. H. Li, S. H. Low, and Q. Yin. REM: Active Queue Management. *IEEE Network*, 15(3):48–53, May/June 2001.

[7] T. J. Ott, T. V. Lakshman, and L. Wong. Sred: Stabilized red. In *Proc. of IEEE INFOCOM*, pages 1346 – 1355, 1999.

[8] P.J. Schweitzer. Perturbation Series Expansion of Nearly Completely-decomposable Markov Chains. In O.J. Boxma, J.W. Cohen, and H.C. Tijms, editors, *Teletraffic Analysis and Computer Performance Evaluation.* Elsevier Science Publishers B.V. (North Holland), 1986.

[9] K.E. Avrachenkov, J.A. Filar, and M. Haviv. Singular Perturbations of Markov Chains and Decision Processes. A Survey. In E.A. Feinberg and A. Shwartz, editors, *Handbook of Markov Decision Processes: Methods and Applications*, International Series in Operations Research and Management Science, pages 113–153. Kluwer Academic Publishers, January 2002.

[10] G. Latouche and V. Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modeling.* ASA-SIAM Series on Statistics and Applied Probability. SIAM, 1999.

[11] T. Bonald, M. May, and J. Bolot. Analytic Evaluation of RED Performance. In *Proc. of the IEEE INFOCOM*, 2000.

[12] V. Mishra, W. Gong, and D. Towsley. Fluid based Analysis of a Network of AQM Routers Supporting TCP Flows with an Application to RED. In *Proc. of the ACM SIGCOMM*, 2000.

[13] C. Hollot, V. Misra, D. Towsley, and W. Gong. A Control Theoretic Analysis of RED. In *Proc. of the IEEE INFOCOM*, 2001.

[14] E. Kuumola, J.A.C. Resing, and J. Virtamo. Joint Distribution of Instantaneous and Averaged Queue Length in an M/M/1/K System. In P. Tran-Gia and J. Roberts, editors, *Proc. of the 15th ITC Specialist Seminar "Internet Traffic Engineering and Traffic Management"*, pages 58–67, July 2002.

[15] V. Sharma, J. Virtamo, and P. Lassila. Performance Analysis of the RED Algorithm. *Probability in the Engineering and Information Sciences*, 16:367–388, 2002.

[16] F. Camilli and M. Falcone. *Nonsmooth Analysis and Geometric Methods in Deterministic Optimal Control*, chapter Approximation of optimal control problems with state constraints: estimates and applications. Springer, 1196.

[17] M. Bardi and I. Capuzzo-Dolcetta. *Optimal control and viscosity solutions of Hamilton-Jacobi-Bellman Equations.* Birkhäuser, 1997.

[18] H. J. Kushner and P. G. Dupuis. *Numerical Methods for Stochastic Control Problems in Continuous Time.* Springer, 1992.

[19] S. Floyd. RED: Discussions of Setting Parameters, November 1997. Available at http://www.icir.org/floyd/red.html.

[20] P.J. Schweitzer. Perturbation Theory and Finite Markov Chains. *J. Appl. Probability*, 5:410–413, 1968.

[21] D.P. Gaver, P.A. Jacobs, and G. Latouche. Finite Birth-and-death Models in Randomly Changing Environments. *Adv. Appl. Probability*, 16:715–731, 1984.

[22] S. Floyd. Recommendation on using the gentle variant of RED, March 2000. Available at http://www.icir.org/floyd/red.html.

[23] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the self-similar nature of ethernet traffic. *IEEE/ACM Transactions Networking*, 2(1):1–15, 1994.

[24] V. Paxson and S. Floyd. Wide-Area Traffic: The Failure of Poisson Modeling. *IEEE/ACM Transactions on Networking*, 3(3):226–244, June 1995.

[25] Mark E. Crovella and Azer Bestavros. Self-similarity in World Wide Web traffic: Evidence and possible causes. *IEEE/ACM Transactions on Networking*, 5(6):835–846, December 1997.

[26] J. Cao, W. S. Cleveland, D. Lin, and D. X. Sun. *Nonlinear Estimation and Classification*, volume 171 of *Lecture Notes in Statistics*, chapter Internet Traffic Tends Toward Poisson and Independent as the Load Increases, pages 83–109. Springer, 2003.

[27] T. Karagiannis, M. Molle, M. Faloutsos, and A. Broido. A Nonstationary Poisson View of Internet Traffic. In *Proc. of IEEE INFOCOM*, March 2004.