A discrete-time Markov modulated queuing system with batched arrivals

Richard G. Clegg

Department of Electrical and Electronic Engineering, University College, London WC1E 7JE

Abstract

This paper examines a discrete-time queuing system with applications to telecommunications traffic. The arrival process is a particular Markov modulated process which belongs to the class of discrete batched Markovian arrival processes. The server process is a single server deterministic queue. A closed form exact solution is given for the expected queue length and delay. A simple system of equations is given for the probability of the queue exceeding a given length.

Key words: queueing theory, D-BMAP/D/1 system, Markov-modulated process, Markov chain
PACS: 02.50.Ga, 02.50.r, 07.05.Tp

1. Introduction

This paper provides a solution for the expected queue length and probability of a given queue length for a simple discrete-time queuing system. The queuing system in question processes one unit of work in one unit of time. Work arrives in integer units according to an arrival process with the following properties.

- The system has two states, *on* and *off*. In an *off* state, no work will arrive.

- If the system is in an *off* state then, with probability f_0, in the next time unit the system is also an *off* state.

- If the system is in an *off* state then, with probability f_i, in the next time unit the system will move to an *on* state which will last for exactly i time units and then move to an *off* state.

- If the system is *on* then a non-zero integer number of work units will arrive in this time period. The number of units of work which arrive is an iid random variable with g_n as the probability that exactly n units of work will arrive in this time period.

Email address: `richard@richardclegg.org` (Richard G. Clegg)

This process can be modelled as a Markov-modulated process (MMP) which is completely characterised by the parameters f_i and g_i. This model has been studied by authors in several different areas, for example statistical physics [15], the study of dynamical systems [6] and modelling telecommunications traffic [5],[16]. In the last two papers, the model is considered as the source of input to a network and hence it is natural to consider the queuing properties of such a model. In this paper expressions are derived for the expected queue length at equilibrium and the probability that the queue has a given length at equilibrium under certain natural restrictions (for example, that the utilisation of the system is less than one). It is shown that the expected queue length is a function of only four variables, the first and second moments of the parameters f_i and g_i. From Little's law [13], the expected delay is proportional to the expected queue length.

In section 2 the model is introduced formally and some basic properties are derived. The model is related to existing work in queuing theory. In section 3 the model is solved to get equations for the expected queue length and the probability of a given queue length. In section 4 the derived equations are compared with computer simulations.

2. A Markov queuing model

Consider the discrete time process described in the introduction. The motivation behind this process is the idea that the lengths of *off* periods are memoryless (the probability that an *on* period begins is independent of the length of the *off* period so far) but *on* periods have lengths which are iid with an arbitrary distribution (within certain feasible constraints). This arrival process is then used as the input to a queue which can process one unit of work (either queued or newly arriving) per time period.

The process Y_t, the arrival process to the system, can be modelled by an MMP. The process has the underlying Markov chain shown in figure 1. At any time the chain is in an *on* state (all of the non-zero states) then a non-zero number of arrivals will occur with a given probability.

The $n+1$ state Markov chain from figure 1 has the transition matrix

$$\mathbf{P} = \begin{bmatrix} f_0 & f_1 & \cdots & f_{n-1} & f_n \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}.$$
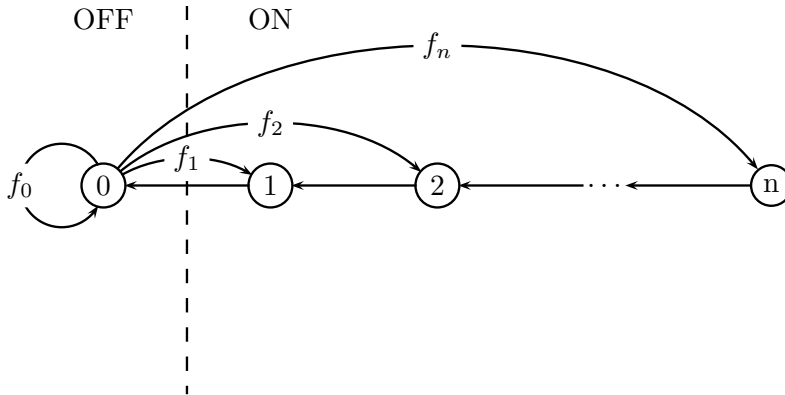
Figure 1: The Markov chain for an MMP traffic model.

Let $\{X_t : t \in \mathbb{N}\}$ be the states of a discrete-time, homogeneous Markov chain at time t. This chain is shown in figure 1 and the parameters f_i define the transition probabilities of the chain. It can be seen from the structure of the chain that f_i is the probability that, given the chain is currently in state zero, the next state will be i. This could also be thought as the probability that, given the chain is currently in an *off* state, the next state will begin an *on* period of length exactly i (f_0 is the probability the next state will be an *off* state again).

Let $g_i : i \in \mathbb{N}$ be the probability that exactly i units of work arrive when the chain is in an *on* (non-zero) state ($X_t > 0$). Let $\{Y_t : t \in \mathbb{N}\}$ be a series derived from X_t by the rule, that if $X_t = 0$ then $Y_t = 0$ and if $X_t > 0$ then for all $i \in \{1, 2, \ldots, m\}$ (where m is the maximum possible value of Y_t) $\mathbb{P}[Y_t = i] = g_i$ (for an *on* state, the possibility that $Y_t = 0$ is excluded, that is, $g_0 = 0$).

The server model used is a deterministic process where exactly one unit of work is processed in one time unit.

2.1. Related work

The arrival process belongs to the processes known as discrete batch Markovian arrival process (D-BMAP) and the queuing system is a subset of D-BMAP/D/1 queues. The BMAP itself was introduced by Lucantoni [14]. Both the D-BMAP and the BMAP have previously been used as models of telecommunications traffic [3, 7, 9]. Details of the BMAP can be found in most modern books on queuing theory [4, Chapter 12] and only a brief outline is given here. It is also interesting to note that the arrival model considered here is very similar to the batch renewal process studied by Fretwell and Kouvatsos [11] in the context of internet traffic. In fact the system they study is the one in figure 1 with on and off reversed.

The structure of a generic D-BMAP is as follows. Let $\mathbf{P}$ be the transition matrix for a discrete time Markov chain with state space $E = \mathbb{N} \times 0, \ldots, n$.

3

Each state of the chain is a pair (j, i) where j is the *level* (the number of arrivals generated in that state) and i is the *phase*. The transition matrix has the structure

$$\mathbf{P} = \begin{bmatrix} \mathbf{D}_0 & \mathbf{D}_1 & \mathbf{D}_2 & \mathbf{D}_3 & \cdots \\ \mathbf{0} & \mathbf{D}_0 & \mathbf{D}_1 & \mathbf{D}_2 & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{D}_0 & \mathbf{D}_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

where $\mathbf{D}_i$ is, itself an $(n+1) \times (n+1)$ matrix. The (j, k)th entry in $\mathbf{D}_i$ is the probability of a phase transition from phase j to k given level i.

The arrival process described in the previous section could be described as a D-BMAP where the phase is the state of the chain in Figure 1 and the levels simply represent the various *on* states. The $\mathbf{D}$ are the $(n+1) \times (n+1)$ matrices

$$\mathbf{D}_0 = \begin{bmatrix} f_0 & 0 & \cdots & 0 \\ f_1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ f_n & 0 & \cdots & 0 \end{bmatrix}.$$

and

$$\mathbf{D}_i = \begin{bmatrix} 0 & g_i & 0 & \cdots & 0 \\ 0 & 0 & g_i & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & g_i \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix},$$

for $i \in \{1, 2, \ldots, m\}$.

Blondia and Cassals [3] provide a method for solving the D-BMAP/D/1/K queuing model which gives a solution in terms of a recursive series of matrix equations. The complexity of calculation is given by the author as $K^2 M^3$ where K is the buffer size (the possible number of items in the queue) and M is the number of phases ($n+1$ in the system described here).

In contrast the system studied here only works for infinite buffers and gives an answer for the expected queue length in closed form. It gives an answer for the probability of the queue having a given length as a recursive system of equations each with n terms.

2.2. Basic properties of the system

It is useful to define $\overline{f} = \sum_{i=1}^{n} i f_i$ and $\overline{f^2} = \sum_{i=1}^{n} i^2 f_i$. Similarly, define $\overline{g} = \sum_{i=1}^{m} i g_i$ and also $\overline{g^2} = \sum_{i=1}^{m} i^2 g_i$. Since $\overline{g}$ is the mean number of arrivals in an *on* state then this should be finite (otherwise the system will have an infinite mean arrival rate). This will be the case in all systems with m finite.

Let π_i be the equilibrium probability of state i. This exists when the chain is ergodic. It can be easily shown that the finite chain is ergodic if $f_0 \in (0, 1)$

and, if this is the case, the equilibrium probability of state i, π_i is given by

$$\pi_i = \pi_0 \sum_{j=i}^{n} f_j.$$

and

$$\pi_0 = 1 - \sum_{i=1}^{n} \pi_i = \frac{1}{1 + \overline{f}}$$

and rearranging gives

$$\overline{f} = \frac{1 - \pi_0}{\pi_0}.$$

It is also of interest to consider the version of this chain where $n \to \infty$. That is for any $N \in \mathbb{N}$ there exists $i > N$ such that $f_i > 0$. The mean return time to the state zero is given by $1 + \overline{f}$ which is finite if $\overline{f}$ is finite. Therefore the infinite chain is ergodic if $f_0 > 0$ and $\overline{f}$ is finite. The infinite chain is useful in studying long-range dependence [1, 5]. Obviously for finite chains $\overline{f}$ is finite.

The output Y_t is then used as input to a queuing system. Let Q_t represent the number queuing at time t and Y_t represent the number of arrivals to the system during $[t, t+1)$. One queued item is removed from the queue every time unit if there is any work in the queue or if any arrives (if $Q_t + Y_t > 0$). Therefore the queuing system is as follows

$$Q_{t+1} = [Q_t + Y_t - 1]^+ \,, \tag{1}$$

where $[X]^+$ means $\max(0, X)$.

The quantity $1 - \pi_0 = \overline{f}/(1 + \overline{f})$ represents the proportion of the time the system is in an *on* state and therefore $\overline{g}(1 - \pi_0)$ is the mean arrival rate λ. Since the system can output one unit of work per time period the queue utilisation (proportion of time the queue is non-empty) $\rho = \lambda$. Both are given by,

$$\rho = \lambda = \frac{\overline{g}\overline{f}}{1 + \overline{f}}. \tag{2}$$

In order that the queue does not grow forever it is obviously a necessary condition that $\rho < 1$.

To summarise, the model is specified by the parameters f_i and g_i. The requirements on these parameters are that $f_0 > 0$ (which guarantees the underlying MC is aperiodic), that $\overline{f}$ is finite and that $\overline{g}\overline{f}/(1 + \overline{f}) < 1$ (which in turn is a requirement that $\overline{g}$ is finite). The first two requirements together guarantee that the underlying MC is ergodic the third requirement ensures that the queue utilisation is less than one.

2.3. Notation

The following notation is used in this paper and is gathered here for convenience.

- Y_t — the number of arrivals to the system during $[t, t+1)$.

- X_t — the state of the underlying Markov chain during $[t, t+1)$.

- Q_t — the number queuing at time t.

- X, Y, Q — the above quantities as random variables at some time when the system is in equilibrium.

- π_i — the equilibrium probability of the ith state of the chain.

- $\mathbf{P}$ — the transition matrix for X_t.

- f_i — the transition probabilities in $\mathbf{P}$.

- n — the highest possible value of X_t (the highest numbered state in $\mathbf{P}$).

- $\overline{f}, \overline{f^2}$ — the first and second moments of f_i, $\sum_{i=1}^{n} if_i$ and $\sum_{i=1}^{n} i^2 f_i$.

- $Q_i(z)$ — the generating function $\sum_{q=0}^{\infty} \mathbb{P}\left[Q = q, X = i\right] z^q$.

- $\mathbf{Q}(z)$ — the $n+1$ column vector $[Q_0(z), Q_1(z), \ldots, Q_n(z)]^T$.

- g_i — the probability that an amount of work i arrives in the next time unit if the system is on, $\mathbb{P}\left[Y_t = i | X_t > 0\right]$.

- m — the maximum possible value of Y_t (the largest number of units of work which may arrive in unit time).

- $\overline{g}, \overline{g^2}$ — the first and second moments of g_i, $\sum_{i=1}^{m} ig_i$ and $\sum_{i=1}^{m} i^2 g_i$.

- $g(z)$ — the generating function $\sum_{i=1}^{m} g_i z^i$.

- b_0 — the boundary condition $\mathbb{P}\left[Q = 0 | X = 0\right]$.

- ρ — the queue utilisation.

- λ — the mean arrival rate.

3. Solving the queuing model

Let $\mathbf{Q}(z) = [Q_0(z), Q_1(z), \ldots, Q_n(z)]^T$ be a column vector of the generating function for the queue in each state of the chain. That is,

$$Q_i(z) = \sum_{q=0}^{\infty} \mathbb{P}\left[Q = q, X = i\right] z^q.$$

Consider a general MMP arrival process. Let $A_i(z)$ be the generating function for the number of arrivals if the underlying chain is in state i. Let $\mathbf{B} = [b_0, b_1, \ldots, b_n]^T$ be the $(n+1)$ column vector of boundary conditions, $b_i = \mathbb{P}\left[Y = 0, Q = 0 | X = i\right]$. Following Li [12] it can be shown that the queuing

system of equation (1) using a general MMP with transition matrix $\mathbf{P}$ as input implies

$$\mathbf{Q}(z) = (z-1)[z\mathbf{I} - \mathbf{P}^T\mathbf{G}(z))]^{-1}\mathbf{P}^T\mathbf{B}, \tag{3}$$

where $\mathbf{G}(z) = \mathrm{diag}(A_0(z), A_1(z), \ldots, A_n(z))$.

For the MMP in question $\mathbf{B}$ and $\mathbf{G}(z)$ have much simpler forms. In the *off* state the generating function for arrivals is simply 1 (no arrivals occur with probability one) and in the *on* state the generating function is $g(z)$. Therefore, $\mathbf{G}(z) = \mathrm{diag}(1, g(z), g(z), \ldots, g(z))$. Since in the *on* state, the system always has at least one arrival and in the *off* state the system always has no arrivals then $\mathbf{B}$ is the $(n+1)$ column vector, $\mathbf{B} = [b_0, 0, 0, \ldots, 0]^T$ where $b_0 = \mathbb{P}[Y = 0, Q = 0|X = 0] = \mathbb{P}[Q = 0|X = 0]$. It is these simplifications which make this system soluble.

For the specific MMP being studied (3) can be rewritten,

$$\mathbf{Q}(z) = (z-1)(\mathbf{A} + \mathbf{A}')^{-1}\mathbf{P}^T\mathbf{B},$$

where $\mathbf{A}$ is the $(n+1) \times (n+1)$ matrix

$$\mathbf{A} = \begin{bmatrix} z & -g(z) & 0 & 0 & \cdots \\ 0 & z & -g(z) & 0 & \cdots \\ 0 & 0 & z & -g(z) & \cdots \\ 0 & 0 & 0 & z & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

and $\mathbf{A}'$ is the rank one $(n+1) \times (n+1)$ matrix which can be written as $\mathbf{A}' = uv$ where $u = [f_0, f_1, f_2, \ldots, f_n]^T$ and $v = [-1, 0, 0, \ldots]$.

The matrix $\mathbf{A}$ can be inverted giving the $(n+1) \times (n+1)$ matrix

$$\mathbf{A}^{-1} = \frac{1}{z} \begin{bmatrix} (g(z)/z)^0 & (g(z)/z)^1 & (g(z)/z)^2 & (g(z)/z)^3 & \cdots \\ 0 & (g(z)/z)^0 & (g(z)/z)^1 & (g(z)/z)^2 & \cdots \\ 0 & 0 & (g(z)/z)^0 & (g(z)/z)^1 & \cdots \\ 0 & 0 & 0 & (g(z)/z)^0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Note that since $g_0 = 0$ then $g(z)/z = \sum_{i=1}^m g_i z^i / z = \sum_{i=1}^m g_i z^{i-1}$. If $z \in [0, 1]$ then $\sum_{i=1}^m g_i z^{i-1} \leq \sum_{i=1}^m g_i = 1$. Therefore $g(z)/z \in [0, 1]$ if $z \in [0, 1]$ Hence $(g(z)/z)^n$ remains bounded as $n \to \infty$. The Sherman–Morrison formula (see, for example, [2]) states that,

$$(\mathbf{A} + \mathbf{A}')^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}uv\mathbf{A}^{-1}}{1 + v\mathbf{A}^{-1}u}.$$

Now,

$$1 + v\mathbf{A}^{-1}u = 1 - 1/z \sum_{i=0}^n (g(z)/z)^i f_i.$$

Define
$$a_i(z) = \frac{1}{z}\sum_{j=i}^{n} f_j \left(\frac{g(z)}{z}\right)^{j-i}.$$

Then
$$(\mathbf{A} + \mathbf{A}')^{-1} = \left[\frac{[1 - a_0(z)]\mathbf{I} - \mathbf{A}^{-1}uv}{1 - a_0(z)}\right]\mathbf{A}^{-1},$$

whence,
$$\mathbf{Q}(z) = (z-1)\frac{[1 - a_0(z)]\mathbf{I} - \mathbf{A}^{-1}uv}{1 - a_0(z)}\mathbf{A}^{-1}\mathbf{P}^T\mathbf{B}.$$

Multiplying the matrices gives
$$Q_i(z) = \frac{z-1}{1 - a_0(z)}b_0 a_i(z).$$

From the definition of $Q_i(z)$, $\mathrm{E}\left[z^Q\right] = \sum_{i=0}^{n} Q_i(z)$ and therefore

$$\mathrm{E}\left[z^Q\right] = b_0\frac{(z-1)\sum_{i=0}^{n} f_i \sum_{j=0}^{i}(g(z)/z)^j}{z - \sum_{i=0}^{n} f_i(g(z)/z)^i} = \frac{b_0 N(z)}{D(z)}, \tag{4}$$

where $N(z) = (z-1)\sum_{i=0}^{n} f_i \sum_{j=0}^{i}(g(z)/z)^j$ and $D(z) = z - \sum_{i=0}^{n} f_i(g(z)/z)^i$.

3.1. Calculating b_0

Now it is necessary to calculate b_0. Note that $\lim_{z\to 1} z^Q = 1$, whence

$$b_0 = \lim_{z\to 1}\frac{D(z)}{N(z)}.$$

Since, $\lim_{z\to 1} D(z) = \lim_{z\to 1} N(z) = 0$, by L'Hôpital's rule,

$$b_0 = \lim_{z\to 1}\frac{D'(z)}{N'(z)}.$$

But
$$D'(z) = 1 - (g(z)/z)'\sum_{i=1}^{n} i f_i(g(z)/z)^{i-1},$$

and
$$(g(z)/z)' = z^{-2}\sum_{i=1}^{m}\left(i g_i z^i - g_i z^i\right).$$

Hence $\lim_{z\to 1}(g(z)/z)' = \bar{g} - 1$, which implies,

$$\lim_{z\to 1} D'(z) = 1 - (\bar{g} - 1)\sum_{i=1}^{n} i f_i = 1 + (1 - \bar{g})\bar{f}. \tag{5}$$

8

Similarly,

$$N'(z) = \sum_{i=0}^{n} f_i \sum_{j=0}^{i} (g(z)/z)^j + (z-1) \sum_{i=0}^{n} (g(z)/z)' f_i \sum_{j=0}^{i} j(g(z)/z)^{j-1}.$$

Providing the sum at the right hand side remains finite (which it will for all finite n) then the factor of $(z-1)$ will cancel this term as $z \to 1$. This gives

$$\lim_{z \to 1} N'(z) = \sum_{i=0}^{n} (i+1) f_i = 1 + \overline{f}.$$

Finally, therefore,

$$b_0 = 1 - \frac{\overline{g}\overline{f}}{1 + \overline{f}} = 1 - \rho. \tag{6}$$

3.2. Queuing results

The next stage is to get a function for the expectation of the queue size.

$$\lim_{z \to 1} \frac{d\mathrm{E}\left[z^Q\right]}{dz} = \lim_{z \to 1} \sum_{q=1}^{\infty} q\mathbb{P}\left[Q = q\right] z^{q-1} = \sum_{q=1}^{\infty} q\mathbb{P}\left[Q = q\right] = \mathrm{E}\left[Q\right].$$

Since b_0 is constant, from (4),

$$\mathrm{E}\left[Q\right] = b_0 \lim_{z \to 1} \frac{d(N(z)/D(z))}{dz} = b_0 \lim_{z \to 1} \frac{N'(z)D(z) - N(z)D'(z)}{D(z)^2}.$$

Similarly, $\lim_{z \to 1} D(z) = 0$ and, from L'Hôpital's rule,

$$\mathrm{E}\left[Q\right] = b_0 \lim_{z \to 1} \frac{N''(z)D(z) - N(z)D''(z)}{2D(z)D'(z)}$$

and, since $\lim_{z \to 1} N(z)/D(z) = 1/b_0$,

$$\mathrm{E}\left[Q\right] = b_0 \lim_{z \to 1} \frac{N''(z) - D''(z)}{2D'(z)}. \tag{7}$$

It is now necessary to find expressions for $N''(z)$ and $D''(z)$. Firstly,

$$N''(z) = [2(g(z)/z)' + (z-1)(g(z)/z)''] \sum_{i=1}^{n} f_i \sum_{j=1}^{i} j(g(z)/z)^{j-1}$$
$$+ (z-1)(g(z)/z)'^2 \sum_{i=1}^{n} f_i \sum_{j=1}^{i} j(j-1)(g(z)/z)^{j-2}.$$

Hence, if all the sums remain finite (which they will if n is finite),

$$\lim_{z \to 1} N''(z) = [2(\overline{g} - 1)] \sum_{i=1}^{n} i(i+1) f_i/2 = (\overline{g} - 1)(\overline{f^2} + \overline{f}). \tag{8}$$

9

Similarly

$$D''(z) = - (g(z)/z)'' \sum_{i=1}^{n} i f_i (g(z)/z)^{i-1}$$

$$- (g(z)/z)'^2 \sum_{i=1}^{n} i(i-1) f_i (g(z)/z)^{i-2},$$

and

$$\lim_{z \to 1} (g(z)/z)'' = \lim_{z \to 1} z^{-3} \sum_{i=1}^{m} (i^2 - 3i + 2) g_i z^i = \overline{g^2} - 3\overline{g} + 2,$$

therefore,

$$\lim_{z \to 1} D''(z) = [\overline{g} + \overline{g}^2 - \overline{g^2} - 1]\overline{f} - (\overline{g} - 1)^2 \overline{f^2}. \tag{9}$$

Substituting (5), (6), (8) and (9) into (7) gives

$$\mathrm{E}\,[Q] = \frac{\overline{g}(\overline{g} - 1)\left[\overline{f^2} - \overline{f}^2\right] + \overline{f}(1 + \overline{f})\left[\overline{g^2} - \overline{g}^2\right]}{2(1 + \overline{f})[1 + \overline{f} - \overline{f}\overline{g}]}. \tag{10}$$

Note that $\overline{g} \geq 1$, $\overline{f^2} \geq \overline{f}^2$ and $\overline{g^2} \geq \overline{g}^2$ by their respective definitions and $1 + \overline{f} > \overline{f}\overline{g}$ for a system with utilisation less than one by equation (2). All bracketed terms in the numerator are therefore positive or zero and the denominator is strictly positive. Note, however that $\overline{f^2}$ and $\overline{g^2}$ are only guaranteed finite for systems with finite n and m respectively. That is to say that some systems with a mean arrival rate less than one and an utilisation less than one will still have an no finite value for the expected queue length. An example of such a system are the systems with the f_i parameters given in [5, 15] which both have $\overline{f^2}$ as a non-convergent series. Such systems are of interested to those studying long-range dependence and heavy-tailed distributions.

Interpreting $\overline{f^2} - \overline{f}^2$ and $\overline{g^2} - \overline{g}^2$ as the variance of f and g respectively then (10) could also be written as

$$\mathrm{E}\,[Q] = \frac{\overline{g}(\overline{g} - 1)\mathrm{var}\,(f) + \overline{f}(1 + \overline{f})\mathrm{var}\,(g)}{2(1 + \overline{f})^2(1 - \rho)}.$$

From (2), the expected delay is given from Little's law

$$\mathrm{E}\,[T] = \frac{\mathrm{E}\,[Q]}{\lambda} = \frac{\overline{g}(\overline{g} - 1)\mathrm{var}\,(f) + \overline{f}(1 + \overline{f})\mathrm{var}\,(g)}{2(1 + \overline{f})^2\rho(1 - \rho)}.$$

An implication of these equations is that, assuming the mean traffic level is fixed (that is $\overline{f}$ and $\overline{g}$ cannot be changed) then the queueing delay of the system would be minimised if the variance in the lengths of the *on* periods was minimised and the variance of the amount of traffic arriving in an *on* period was minimised. This has an interesting analogy to the well-known Pollaczek-Khinchin result [8] that for an M/G/1 queue the waiting time is proportional to the variance in the service time.

10

3.3. Finding the queue distribution function

In order to be able to ask questions about, for example, buffer overflow probabilities, it would be useful to be able to ask questions about the probability of a given queue size ($\mathbb{P}[Q = i]$) or the probability that the queue is more than a given size ($\mathbb{P}[Q > i]$).

The probability that the queue is zero is given by

$$\mathbb{P}[Q = 0] = \lim_{z \to 0} \mathrm{E}\left[z^Q\right],$$

and, more generally, the probability that the queue length is q can be found by differentiating q times and taking the limit as $z \to 0$.

$$\mathbb{P}[Q = q] = \lim_{z \to 0} \frac{d^q \mathrm{E}\left[z^Q\right]}{q! d^q z}.$$

This can be solved computationally by repeated symbolic differentiation. However, this is computationally intensive and the algorithm is numerically unstable. Another approach is to produce a recursive formula for the coefficient of z^i in $\mathrm{E}\left[z^Q\right]$. This can be done using standard techniques for formal power series from, for example, Knuth [10].

By definition, the coefficient of z^i in $\mathrm{E}\left[z^Q\right]$ is $\mathbb{P}[Q = i]$. Let N_i and D_i be the coefficients of z^i in $N(z)$ and $D(z)$. Since (4) is true for all z, therefore standard techniques for division of power series give

$$\sum_{i=0}^{k} \mathbb{P}[Q = i] D_{k-i} = b_0 N_k,$$

which rearranges to

$$\mathbb{P}[Q = k] = \frac{1}{D_0} \left[N_k b_0 - \sum_{i=0}^{k-1} \mathbb{P}[Q = i] D_{k-i} \right]. \tag{11}$$

This recursive formula expresses $\mathbb{P}[Q = k]$ in terms of b_0 which can be evaluated with (6), coefficients N_i and D_i and $\mathbb{P}[Q = j]$ for $j < k$. Therefore, the $\mathbb{P}[Q = k]$ can be calculated in turn beginning with $\mathbb{P}[Q = 0]$ which is given by

$$\mathbb{P}[Q = 0] = \frac{b_0 N_0}{D_0}.$$

The coefficients N_i and D_i can be easily calculated. Let $G_{i,j}$ be the coefficient of z^i in $(g(z)/z)^j$. The coefficients above can be expressed as

$$D_i = \delta_{i-1} - \sum_{j=0}^{n} f_j G_{i,j} \tag{12}$$

where δ_i is the Kronecker delta function ($\delta_{i-1} = 1$ if $i = 1$ and 0 otherwise) and

$$N_i = \begin{cases} -\sum_{j=0}^{n} G_{i,j} \sum_{k=j}^{n} f_k & i = 0 \\ \sum_{j=0}^{n} [G_{i-1,j} - G_{i,j}] \sum_{k=j}^{n} f_k & i > 0. \end{cases} \tag{13}$$

11

The coefficients $G_{i,j}$ can be calculated by another recurrence relation. Since

$$g(z)/z = \sum_{i=0}^{m-1} g_{i+1}z^i,$$

and as $G_{i,0} = \delta_i$, this gives the recurrence relation

$$G_{i,j+1} = \sum_{k=0}^{i} G_{k,j}g_{i+1-k}.$$

3.4. A simpler model — the constant batch size model

A simplification occurs when the batch size is fixed. Assume that, work must arrive in units of exactly r where $r > 1$ ($r = 1$ is the uninteresting system where no queue ever forms). This means that $g(z)/z = z^{r-1}$ and $(g(z)/z)^j = z^{j(r-1)}$. In turn this gives $G_{i,j} = \delta_{i-j(r-1)}$ where here, and throughout this section, δ is the Kronecker delta function. Obviously $\overline{g} = r$ and $\overline{g^2} = \overline{g}^2 = r^2$, hence from (10)

$$\mathrm{E}\,[Q] = \frac{r(r-1)(\overline{f^2} - \overline{f}^2)}{2(1+\overline{f})(1+\overline{f}-r\overline{f})}.$$

It can also be shown that $\mathbb{P}\,[Q = 0] = b_0/f_0$ and for $k > 0$,

$$\mathbb{P}\,[Q = k] = \frac{1}{f_0}\left[\mathbb{P}\,[Q = k-1] - \sum_{j=0}^{k-1}\delta_{j/(r-1)-\lfloor j/(r-1)\rfloor}f_j\mathbb{P}\,[Q = j]\right.$$

$$- b_0\delta_{(k-1)/(r-1)-\lfloor(k-1)/(r-1)\rfloor}\sum_{j=(k-1)/(r-1)}^{n} f_j$$

$$\left.+ b_0\delta_{k/(r-1)-\lfloor k/(r-1)\rfloor}\sum_{j=k/(r-1)}^{n} f_j\right],$$

where $\lfloor x \rfloor$ is the floor function and with the notational conveniences that $f_j = 0$ for $j > n$ and that $\sum_j^n f_j = 0$ for $j > n$. Note that the delta functions here are simply testing if a given expression, for example $k/(r-1)$, is integer. From (6),

$$b_0 = \frac{1+\overline{f}-r\overline{f}}{1+\overline{f}}.$$

4. Simulation tests

The system needs to be tested against simulation to see if it can be practically used. While the equations from the previous sections are correct, they are not useful if the numerical stability of the recursive system of equations is poor. For the expected queue length calculations this is not an issue but it is for the probability of higher queue lengths since it is likely that $\mathbb{P}\,[Q = i]$ will become

extremely small as i becomes large. This, in turn, will make the calculation of (11) problematic as i becomes large. Calculating the potential effects of errors in the system of equations given by (11), (12) and (13) is non-trivial. The answers presented here are tested against simulation methods and appear valid for smaller queue sizes but become obviously incorrect (negative probabilities for example) for larger queue sizes. The simulations here were done in python. The same calculations have been tried using arbitrary precision arithmetic libraries. This enabled slightly larger queue sizes to be calculated and give reasonable answers but at the expense of greatly increased run time.

The method is simply to replicate the Markov chain and queuing system described earlier and simulate it. This is an exact simulation of the queuing system described in the paper. The simulation can then run for a set number of iterations and the queue measured at each point to sample $\mathrm{E}\,[Q]$ or get a sample of the probabilities $\mathbb{P}\,[Q = i]$ by measuring the proportion of the iterations where the queue has the value i in simulation. Tables 1 or 2 show the parameter sets for two different simulation scenarios with the first representing a lightly loaded system and the second representing much heavier loading. In both scenarios $\mathrm{E}\,[Q]$ matches well between theory and experiment as expected and no results are presented here.

i	0	1	2	3
f_i	0.8	0.1	0.05	0.05
g_i	0	0.4	0.4	0.2

Table 1: Parameter set 1 for simulation.

i	0	1	2	3	4
f_i	0.6	0.2	0.1	0.05	0.05
g_i	0	0.2	0.6	0.1	0.1

Table 2: Parameter set 2 for simulation.

Figures 2 and 3 show the theory from section 3.3 plotted against ten simulation runs for each parameter set, each run being 10^8 iterations. The simulation results are shown as a mean for the ten simulation runs and upper and lower 95% confidence intervals. The plots are on a logscale and therefore, obviously, values which are zero or negative do not show up. This happens with lower confidence intervals and when there are rounding errors in the theoretical calculation.

For parameter set one, rounding errors make the theoretical calculation obviously unreliable (this is obvious when the numbers become negative) at around $i = 40$ (where $\mathbb{P}\,[Q = i]$ is around 1×10^{-16}) although it is likely that some figures before this are rendered inaccurate due to rounding — the negative numbers are omitted from the plot because of the logscale, all others are included. In fact Note that because only 10^8 iterations were simulated the lowest sample probability that can be assigned in simulation is 1×10^{-8}. For parameter set two,
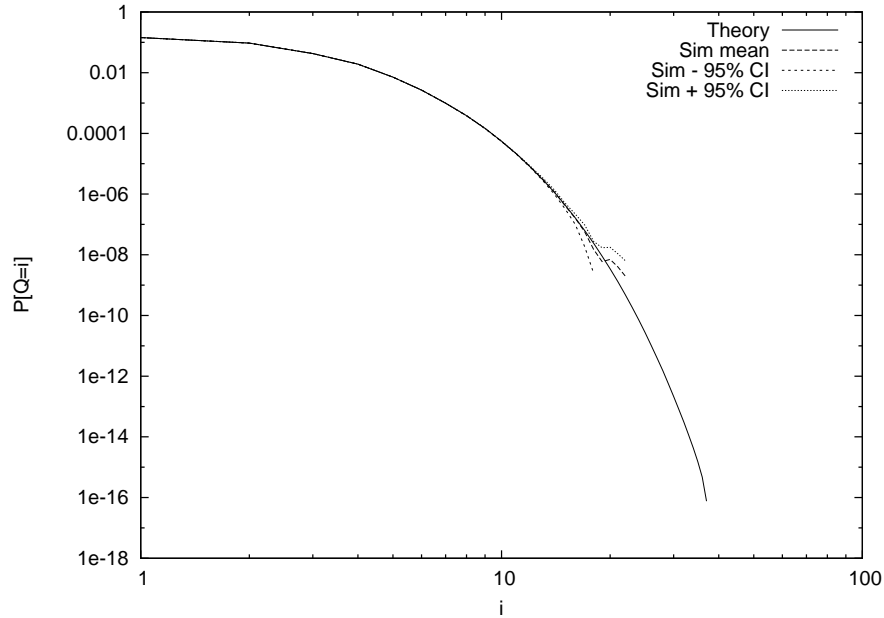
Figure 2: Theory versus simulation for parameter set 1.

it can be seen from figure 3 that the calculation remains reliable until above $i = 110$ when, again, it becomes obviously inaccurate due to rounding at a probability again around 1×10^{-16}. As has been mentioned, by using arbitrary precision arithmetic libraries, the numerical stability can be increased slightly but at the expense of greatly increased run times.

5. Conclusions

The arrival system described is quite general and could be useful in any system when work arrives at discrete times in discrete batches. The solutions given provide mean queue lengths and delays for the given arrival process. In addition a system of equations has been given which can calculate the probability that the queue has a given length from the system parameters and the probabilities of smaller lengths. The numerical stability of the recursive system of equations giving the probability distribution has been assessed via simulation.

References

[1] M. Barenco and D.K. Arrowsmith. The autocorrelation of double intermittency maps and the simulation of computer packet traffic. *Dynamical Systems*, 19(1):61–74, 2004.
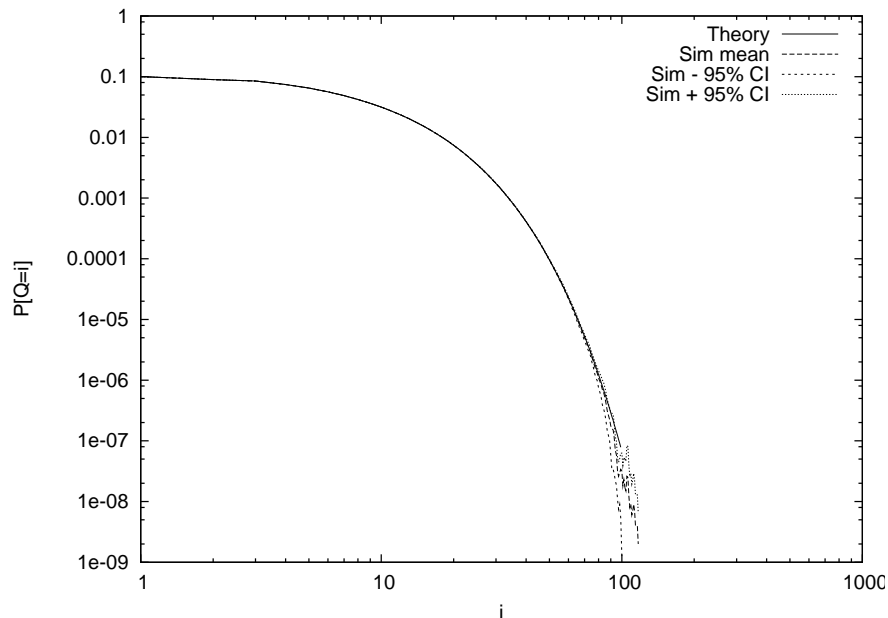
Figure 3: Theory versus simulation for parameter set 2.

[2] D. S. Bernstein and C. F. Van Loan. Rational matrix functions and rank-1 updates. *SIAM J. Matrix Anal. Appl.*, 22(1):145–154, 2000.

[3] C. Blondia and O. Casals. Statistical multiplexing of VBR sources: a matrix- analytic approach. *Performance evaluation*, 16(1–3):5–20, 1992.

[4] L. Breuer and D. Baum. *An introduction to queueing theory and matrix-analytic methods.* Springer, 2005.

[5] R. G. Clegg and M. M. Dodson. A Markov based method for generating long-range dependence. *Phys. Rev. E*, 72:026118, 2005. Available online at:
www.richardclegg.org/pubs/rgcpre2004.pdf.

[6] Z. Coelho. *Asymptotic laws for symbolic dynamical systems*, pages 123–165. Number 279 in LMS Lecture Notes Series. Cambridge University Press, 2000.

[7] Christoph Herrmann. Analyis of the discrete time SMP/D/1/s finite buffer queue with applications in ATM. *Proc. IEEE INFOCOM*, 1993.

[8] A. Y. Khinchin. Mathematical theory of stationary queues. *Mat. Sbornik*, 39:73–84, 1932.

[9] Alexander Klemm, Christoph Lindemann, and Marco Lohmann. Modeling IP traffic using the batch Markovian arrival process. *Performance Evaluation*, 54:149–173, 2003.

[10] D. E. Knuth. *Seminumerical Algorithms*, volume 2 of *The Art of Computer Programming*, section 4.7, pages 525–533. Addison-Wesley, Reading, Massachusetts, third edition, 1997.

[11] Demetres D. Kouvatsos and Rod J. Fretwell. Batch renewal process: exact model of traffic correlation. In *Proceedings of the 2nd International Workshop on Architecture and Protocols for High Performance Networks*, pages 285–304, Deventer, The Netherlands, The Netherlands, 1996. Kluwer, B.V.

[12] S. Li and C. Hwang. Queue response to input correlation functions: Discrete spectral analysis. *IEEE/ACM Trans. on Networking*, 1(5):522–533, October 1993.

[13] J. Little. A proof of the queueing formula $L = \lambda W$. *Oper. Res. J.*, 18:172–174, 1961.

[14] D. M. Lucantoni. New results on the single server queue with a batch markovian arrival process. *Stochastic models*, 7, 1991.

[15] X. J. Wang. Statistical physics of temporal intermittency. *Phys. Rev. A*, 40(11):6647–6661, 1989.

[16] M. Woolf, D. K. Arrowsmith, R. J. Mongragón, and J. M. Pitts. Optimization and phase transitions in a chaotic model of data traffic. *Phys. Rev. E*, 66:046106, 2002.