

Output Analysis of a Single-Buffer Multi-Class Queue: FCFS Service.

V.G. Kulkarni

Department of Operations Research
University of North Carolina
Chapel Hill, N.C. 27599-3180

K. D. Glazebrook

School of Management,
University of Edinburgh,
Edinburgh, EH8 9JY, United Kingdom.

April 18, 2002

Abstract

We consider an infinite capacity buffer where the incoming fluid traffic belongs to K different types modulated by K independent Markovian on-off processes. The k th input process is described by three parameters: (λ_k, μ_k, r_k) , where $1/\lambda_k$ is the mean off time, $1/\mu_k$ is the mean on time and r_k is the constant peak rate during the on time. The buffer empties the fluid at rate c according to a First Come First Served discipline.

The output process of type k fluid is neither Markovian, nor on-off. We approximate it by an on-off process by defining the process to be off if no fluid of type k is leaving the buffer, and on otherwise. We compute the mean on time τ_k^{on} and mean off time τ_k^{off} . We approximate the peak output rate by a constant r_k^o so as to conserve the fluid. We approximate the output process by the three parameters $(\lambda_k^o, \mu_k^o, r_k^o)$, where $\lambda_k^o = 1/\tau_k^{off}$ and $\mu_k^o = 1/\tau_k^{on}$. In this paper we derive methods of computing τ_k^{on} , τ_k^{off} and r_k^o for $k = 1, 2, \dots, K$. They are based on the results for the computation of expected reward in a fluid queueing system during a first passage time. We illustrate the methodology by a numerical example.

In the last section we conduct a similar output analysis for a standard M/G/1 queue with K types of customers arriving according to independent Poisson processes and requiring independent generally distributed service times, and following FCFS service discipline. For this case we are able to get analytical results.

1 Introduction

In this paper we consider a single-buffer multi-class queue. We consider two cases separately: fluid queue and ordinary queue.

In the fluid case the input to the queue is fluid that belongs to K different types modulated by K independent Markovian on-off processes. The k th ($k = 1, 2, \dots, K$) modulating process is on for an $\exp(\mu_k)$ amount of time and off for $\exp(\lambda_k)$ amount of time. When it is on the fluid of type k arrives at rate $r_k > 0$, and no fluid arrives while it is off. Thus the input process of type k is completely described by three parameters: (λ_k, μ_k, r_k) .

The fluid belonging to all classes accumulates in a single buffer and is removed on a first come first served basis. In a fluid setting, the FCFS discipline is taken to mean that fluid arriving at time t is removed from the buffer only after all the fluid that arrived before time t is removed. This implies that typically the buffer will be processing fluids of many types simultaneously.

The output from such a queue is rather complex. See Figure 1 where we show the case of two-class system. The output processes of different type of fluids are neither independent, nor can be described by three parameter on-off processes. Our aim is to approximate the output process of type k fluid by a three parameter on-off process with parameters $(\lambda_k^o, \mu_k^o, r_k^o)$, $k = 1, 2, \dots, K$.

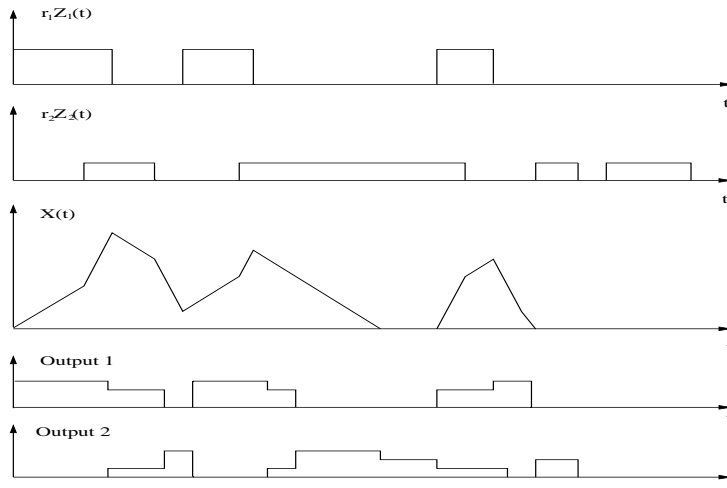


Figure 1: Sample paths of input, buffer content and output processes.

One motivation behind this analysis is in the analysis of telecommunication networks. The idea of using fluid queues in telecommunications is well established, initiated by the pioneering work by

Kosten [10] and Anick, Mitra and Sondhi [1]. There is a large literature on fluid queues, see the survey paper by Kulkarni [13]. Several researchers have also studied networks of fluid queues: see Kella [6], [7], Kella and Whitt [8], [9], Kaspi and Kella [5], etc.

Exact analysis of fluid networks is intractable, and hence one needs to find ways of approximate analysis. In this regard, earlier work done in queueing networks is a valuable guide. One promising approach is the decomposition approach, as presented by Kuehn [11], and further refined by Whitt [18] in QNA. Similar approach is also considered by Reiser and Kobayashi [16], Sevcik, et al [17], Chandy and Saur [2], and Gelenbe and Mitrani [3].

The idea is to model the telecommunication network as a network of multi-class fluid queues with FCFS discipline at all nodes. Here we assume that the external fluid inputs to all the nodes in the network are Markovian on-off processes described by three parameters. Then we approximate the output processes from a given node as independent Markovian on-off processes by using the methodology developed here. That is, we think of a node as a non-linear mapping of the input parameters to output parameters. These output processes then act as input to other queues. We seek an equilibrium solution to the non-linear system where the output parameters are consistent with the input parameters.

It is clear that the results here will be impractical to use for large values of K since the computations involves matrices of size 2^K . Hence, to create a practical fluid network analyzer, we need to develop computationally efficient approximations for computing the output parameters as a function of input parameters. The results developed here can be used to quantify the quality of these approximations. This work is under progress and will be reported at a later date.

The case of the ordinary queue is the standard M/G/1 queue with multiple type of customers. The customers belong to K different types and arrive according to K independent Poisson processes. The arrival rates and service time distributions are class dependent. The service discipline is FCFS. The output analysis of this queue is given in the last section.

2 Output Analysis

To accomplish the aim of characterizing the output of fluid type k as a three parameter on-off process, we define the output process of type k to be “on” if fluid of type k is emerging from the buffer at a positive rate, and “off” if no fluid of type k is emerging from the buffer.

The distributions of the on and off times of the output process of type k are complicated. Let τ_k^{on} and τ_k^{off} denote their respective means. Then we approximate the output on and off times by exponential random variables with parameters

$$\lambda_k^o = \frac{1}{\tau_k^{off}}, \quad (1)$$

$$\mu_k^o = \frac{1}{\tau_k^{on}}. \quad (2)$$

The actual rate at which fluid of type k emerges during the on time of the output process of type k varies. The mean input rate of fluid of type k is as given below:

$$m_k = r_k \frac{\lambda_k}{\lambda_k + \mu_k}.$$

It is known that (see Kulkarni and Rolski [14]) the fluid queue is stable if

$$m = \sum_{k=1}^K m_k < c.$$

In a stable system, the mean output rate m_k^o of type k fluid must be the same as its mean input rate. Hence if we approximate the output rate during the output on time by a constant r_k^o (which we call the effective peak rate of the output process), we must have

$$m_k^o = r_k^o \frac{\lambda_k^o}{\lambda_k^o + \mu_k^o} = m_k.$$

Using Equations 1 and 2, we get

$$r_k^o = r_k \frac{\lambda_k}{\lambda_k + \mu_k} \frac{\tau_k^{on} + \tau_k^{off}}{\tau_k^{on}}. \quad (3)$$

Equations 1, 2 and 3 imply that the output process of type k can be approximated by an on-off process with parameters $(\lambda_k^o, \mu_k^o, r_k^o)$ if we know the mean output on time τ_k^{on} and the mean off time τ_k^{off} . These are computed in Section 5. We collect some preliminary results in Sections 3 and 4.

3 Multi-Class Fluid Queue

Consider a single buffer multi class fluid queue as described in Section 1. Let $Z_k(t)$ be the state of the k th input process at time t . $\{Z_k(t), t \geq 0\}$ is a Continuous Time Markov Chain (CTMC) on state space $\{0 = \text{off}, 1 = \text{on}\}$ with generator matrix

$$Q_k = \begin{bmatrix} -\lambda_k & \lambda_k \\ \mu_k & -\mu_k \end{bmatrix}.$$

Let

$$Z(t) = [Z_1(t), Z_2(t), \dots, Z_K(t)] \quad (4)$$

be the state vector of all the input processes at time t . $\{Z(t), t \geq 0\}$ is a CTMC on state space $S = \{z = (z_1, z_2, \dots, z_K) : z_k = 0, 1, k = 1, 2, \dots, K\}$ with generator matrix $Q = Q_1 \oplus Q_2 \oplus \dots \oplus Q_K$. When the Z process is in state $z \in S$, the fluid of all classes enters the buffer at rate

$$r(z) = \sum_{k=1}^K r_k z_k, \quad z \in S.$$

Let $X(t)$ be the total amount of fluid in the buffer at time t . The dynamics of $\{X(t), t \geq 0\}$ process is described by the following differential equation:

$$\frac{d}{dt} X(t) = \begin{cases} r(z) - c & \text{if } Z(t) = z, X(t) > 0, \\ \max(r(z) - c, 0) & \text{if } Z(t) = z, X(t) = 0. \end{cases}$$

It is easy to see that if

$$c > r = \sum_{k=1}^K r_k$$

the buffer eventually becomes empty and stays empty forever. The Markov nature of the bivariate process (X, Z) implies that it is a regenerative process if

$$m < c < r.$$

The regenerative cycle under this assumption is identified below.

Assume that initially all input processes are down, and the buffer is empty, i.e., $Z(0) = \mathbf{0} = (0, 0, \dots, 0)$ and $X(0) = 0$. Define

$$D = \min\{t \geq 0 : r(Z(t)) > 0\}, \quad (5)$$

$$U = \min\{t \geq 0 : Z(t + D) = \mathbf{0}, X(t + D) = 0\}. \quad (6)$$

Thus D is the duration of the first down time in the output process, i. e., output rate is zero over the interval $[0, D)$, while U is the duration of the first uptime, i. e., the output rate is positive over the interval $[D, D + U)$. Let

$$T = D + U. \quad (7)$$

Note that $Z(T) = \mathbf{0}$ and $X(T) = 0$, and that the (X, Z) process regenerates at time T .

Next define $S_k(t)$ to be 1 if the output rate of fluid of type k is positive at time t , and zero otherwise. Note that the buffer can simultaneously process multiple classes of fluid. Clearly $\{S_k(t), t \geq 0\}$ is an on-off process, but it is not necessarily an alternating renewal process. We are interested in computing τ_k^{on} , the mean on time, and τ_k^{off} , the mean off time, of the S_k process in steady state. Let

$$T_k = \int_0^T S_k(t) dt, \quad (8)$$

and let N_k be the number of times the S_k process switches from 0 to 1 during the time interval $[0, T]$.

Then the regenerative nature of the (X, Z) process implies that (See Kulkarni [12], Chapter 8)

$$\tau_k^{on} = \frac{E(T_k)}{E(N_k)}, \quad (9)$$

and

$$\tau_k^{off} = \frac{E(T) - E(T_k)}{E(N_k)}. \quad (10)$$

Thus we need to compute $E(T)$, $E(T_k)$ and $E(N_k)$ in order to compute τ_k^{on} and τ_k^{off} . In the next section we develop some preliminary results for fluid models in order to accomplish this.

4 Expected Reward in a Fluid Model

Consider a single buffer fluid queue where fluid arrives at rate $r(Z(t))$ at time t , where $\{Z(t), t \geq 0\}$ is an irreducible CTMC with state space $S = \{1, \dots, n\}$ and generator matrix Q . The buffer is drained at rate c . Let $\bar{r}(i) = r(i) - c$ be the net input rate in state $i \in S$. Let $X(t)$ be the buffer content (amount of fluid in the buffer) at time t . The dynamics of $\{X(t), t \geq 0\}$ is described by the following differential equation:

$$\frac{d}{dt}X(t) = \begin{cases} \bar{r}(i) & \text{if } Z(t) = i, X(t) > 0, \\ \max(\bar{r}(i), 0) & \text{if } Z(t) = i, X(t) = 0. \end{cases}$$

Let

$$\pi_i = \lim_{t \rightarrow \infty} P(Z(t) = i), \quad i \in S$$

be the limiting distribution of the Z process. The X process is stable if

$$m = \sum_{i=1}^n r(i)\pi_i < c.$$

Define the first passage time in the bivariate process $(X, Z) = \{(X(t), Z(t)), t \geq 0\}$ as

$$T_A = \min\{t > 0 : X(t) = 0, Z(t) \in A\}, \quad A \subseteq S. \quad (11)$$

T_A is a non defective random variable if $r(i) < c$ in at least one $i \in A$ and $m < c$. We assume that this is the case. Note that the random variable T_A is defined for any initial state $(X(0), Z(0))$.

Let $\rho(x, k)$ be the rate at which reward is earned when the (X, Z) process is in state (x, k) . The total reward earned up to time t is given by

$$Y(t) = \int_0^t \rho(X(u), Z(u)) du.$$

We are interested in computing the expected reward over $[0, T_A]$, namely

$$g_i^A(x) = E(Y(T_A) | X(0) = x, Z(0) = i).$$

The next theorem gives the equations satisfied by

$$g^A(x) = [g_1^A(x), g_2^A(x), \dots, g_n^A(x)]'. \quad (12)$$

We use the notation

$$\begin{aligned} \rho(x) &= [\rho(x, 1), \rho(x, 2), \dots, \rho(x, n)]', \\ \bar{R} &= \text{diag}(\bar{r}(1), \bar{r}(2), \dots, \bar{r}(n)). \end{aligned}$$

Theorem 1 *Assume that T_A is a non defective random variable. The function $g^A(x)$ satisfies the following non homogeneous system of linear differential equations:*

$$\bar{R} \frac{dg^A(x)}{dx} + Qg^A(x) + \rho(x) = 0, \quad x \geq 0, \quad (13)$$

with the following boundary conditions:

$$g_i^A(0) = 0, \quad i \in A, \bar{r}(i) < 0, \quad (14)$$

$$[Qg^A(0)]_i + \rho(0, i) = 0, \quad i \notin A, \bar{r}(i) < 0. \quad (15)$$

Proof: Follows along the same lines as the proof of Theorem 3.1 in Kulkarni and Tzenova [15].

We need the following special case in our applications:

$$\rho(x, i) = \begin{cases} \rho(i) & \text{if } x > 0 \\ \rho_0(i) & \text{if } x = 0. \end{cases} \quad (16)$$

In this case the above equations can be solved relatively easily following the method in Kulkarni and Tzenova [15]. We describe the result below.

A scalar ν and a non-zero vector ϕ is called a (generalized) eigenvalue-eigenvector pair if they satisfy the equation

$$(\nu\bar{R} + Q)\phi = 0. \quad (17)$$

Now, let $k = |\{i \in S : \bar{r}(i) < 0\}|$. It is known that $\bar{r} = \sum_{i=1}^n \bar{r}(i)\pi_i < 0$ implies that there are exactly k pairs $(\nu_i, \phi_i), i = 1, 2, \dots, k$ that satisfy Equation 17 where the eigenvalues have negative real parts. The final solution is given in the following theorem.

Theorem 2 *Assume that T_A is a non defective random variable. Suppose $\rho(x, z)$ is given by Equation 16. The solution to Equation 13 subject to the boundary conditions in Equations 14 and 15 is given by*

$$g^A(x) = \sum_{j=1}^k a_j \phi_j e^{\nu_j x} - \frac{\sum \pi(i)\rho(i)}{\sum \pi(i)\bar{r}(i)} ex + b, \quad (18)$$

where e is an n -dimensional column vector of ones and b is any solution to

$$Qb = \frac{\sum \pi(i)\rho(i)}{\sum \pi(i)\bar{r}(i)} \bar{R}e - \rho. \quad (19)$$

The coefficients a_j are determined by the boundary conditions:

$$\sum_{j=1}^k a_j \phi_{ij} = -b_i, \quad i \in A, \bar{r}(i) < 0, \quad (20)$$

$$\sum_{j=1}^k a_j \nu_j \phi_{ij} = \frac{\sum \pi(i)\rho(i)}{\sum \pi(i)\bar{r}(i)} + \frac{\rho_0(i) - \rho(i)}{\bar{r}(i)}, \quad i \notin A, \bar{r}(i) < 0. \quad (21)$$

Proof: Follows along the same lines as that of Theorem 3.2 in Kulkarni and Tzenova [15]. The assumption about the non-defectiveness of T_A ensures that the number of boundary conditions uniquely determines the unknown constants a_j .

We use the above theorem to compute τ_k^{on} and τ_k^{off} in the next section.

5 Computation of τ_k^{on} and τ_k^{off}

Let $Z(t)$, U and D be as defined in Equations 4, 5, and 6. We first compute $E(T) = E(D) + E(U)$. Since $Z(0) = \mathbf{0}$ and D is the time when one of the K down processes turns on, we see that it is an exponential random variable with parameter $\lambda = \sum_{k=1}^K \lambda_k$. Hence

$$E(D) = \frac{1}{\lambda}.$$

We compute $E(U)$ in the next theorem, which yields $E(T)$.

Theorem 3 *Let $A = \{\mathbf{0}\}$ and*

$$\rho(z) = \rho_0(z) = 1, \quad z \in S. \quad (22)$$

Let $g^A(x)$ be as in Equation 12. Then

$$E(U) = \sum_{k=1}^K \frac{\lambda_k}{\lambda} g_{e_k}^A(0), \quad (23)$$

where $e_k \in S$ is a K vector with k th component 1, and all other components zero.

Proof: Recall that $X(0) = 0$ and $Z(0) = \mathbf{0}$ and let U and D be as defined in Equations 5 and 6. Now define,

$$\tilde{T}_A = \min\{t > 0 : X(t + D) = 0, Z(t + D) \in A\}, \quad (24)$$

and

$$\tilde{Y}(t) = \int_D^{D+t} \rho(X(u), Z(u)) du. \quad (25)$$

The choice of $A = \{\mathbf{0}\}$, the reward rate $\rho(x, i)$ given in Equation 16 and the expressions for $\rho(z)$ and $\rho_0(z)$ in Equation 22 imply that $\tilde{T}_A = U$ and $E(\tilde{Y}(\tilde{T}_A)) = E(U)$. Now, $Z(0) = \mathbf{0}$ implies that the Z process enters state e_k at time D with probability

$$P(Z(D) = e_k) = \frac{\lambda_k}{\lambda}, \quad 1 \leq k \leq K. \quad (26)$$

Hence

$$E(U) = E(\tilde{Y}(\tilde{T}_A)) = \sum_{k=1}^K \frac{\lambda_k}{\lambda} E(\tilde{Y}(\tilde{T}_A) | X(D) = 0, Z(D) = e_k). \quad (27)$$

From the analysis in Section 4 with general initial state for the bivariate process (X, Z) , we see that

$$g_{e_k}^A(0) = E(\tilde{Y}(\tilde{T}_A) | X(D) = 0, Z(D) = e_k). \quad (28)$$

Equations 27 and 28 yield the desired Equation 23.

Next theorem yields a method to compute $E(T_k)$.

Theorem 4 Let $A = \{\mathbf{0}\}$ and

$$\rho(z) = \begin{cases} \frac{r(z)}{c} & \text{if } z \in S, z_k = 1, \\ 0 & \text{otherwise,} \end{cases} \quad (29)$$

and

$$\rho_0(z) = \begin{cases} 1 & \text{if } z \in S, z_k = 1, \\ 0 & \text{otherwise.} \end{cases} \quad (30)$$

Let $g^A(x)$ be as in Equation 12. Then

$$E(T_k) = \sum_{i=1}^K \frac{\lambda_i}{\lambda} g_{e_i}^A(0).$$

Proof: Observe the following: For $u \in [t, t + dt)$, suppose $X(u) > 0$, and $Z(u) = z$ such that $z_k = 1$. Then the volume of type k fluid that enters the buffer during $[t, t + dt)$ is $r_k dt$. Due to the FCFS discipline, it takes $\frac{r(z)}{c} dt$ amount of time to remove that volume some time later. Similarly, for $u \in [t, t + dt)$, suppose $X(u) = 0$, and $Z(u) = z$ such that $z_k = 1$. Then it takes dt amount of

time to remove the type k fluid that enters during time $[t, t + dt)$.

We introduce the random variable \tilde{T}_A and the process $\tilde{Y}(t), t \geq 0$ as in Equations 24 and 25. Let T_k^A be the amount of time needed to process all the fluid of type k that arrived during $(D, D + \tilde{T}_A)$. The above observation implies that, almost surely

$$\begin{aligned} T_k^A &= \int_D^{D+\tilde{T}_A} \frac{r(Z(u))}{c} I\{X(u) > 0, Z_k(u) = 1\} du \\ &\quad + \int_D^{D+\tilde{T}_A} I\{X(u) = 0, Z_k(u) = 1\} du \\ &= \int_D^{D+\tilde{T}_A} \rho(X(u), Z(u)) du = \tilde{Y}(\tilde{T}_A), \end{aligned} \tag{31}$$

where $I\{E\}$ is an indicator function of the event E , $\rho(x, z)$ is as defined in Equation 16 with $\rho(z)$ and $\rho_0(z)$ given by Equations 29 and 30. If we now take $A = \{\mathbf{0}\}$, then T_k^A is stochastically identical to T_k as defined in Equation 8. Hence we have, as in the conclusion to the proof of Theorem 3, that

$$\begin{aligned} E(T_k) &= E(\tilde{Y}(\tilde{T}_A)) \\ &= \sum_{i=1}^K \frac{\lambda_i}{\lambda} E(\tilde{Y}(\tilde{T}_A) | X(D) = 0, Z(D) = e_i) \\ &= \sum_{i=1}^K \frac{\lambda_i}{\lambda} g_{e_i}^A(0). \end{aligned}$$

This yields the desired result.

Finally, we develop a method of computing $E(N_k)$. We first fix a k and construct a modified process $\{Z^k(t), t \geq 0\}$ with state space $S \cup \{\Delta\}$ as follows. $Z^k(t) = Z(t)$ as long as $Z(t) \neq \mathbf{0}$. If $Z(t) = \mathbf{0}$ and the Z process entered this state from any state other than e_k , then $Z^k(t) = \mathbf{0}$, else $Z^k(t) = \Delta$. Thus $\{Z^k(t), t \geq 0\}$ is a CTMC with generator matrix Q^k matrix as given below:

$$[Q^k]_{i,j} = \begin{cases} Q_{i,j} & \text{if } i \notin \{e_k, \Delta\}, j \neq \Delta, \\ \mu_k & \text{if } i = e_k \text{ and } j = \Delta, \\ -\lambda & \text{if } i, j = \Delta, \\ Q_{\mathbf{0},j} & \text{if } i = \Delta \text{ and } j \notin \{\Delta, \mathbf{0}\}, \\ 0 & \text{otherwise.} \end{cases}$$

The rate matrix R^k is a diagonal matrix with diagonal entries as given below:

$$R_{i,i}^k = \begin{cases} r(i) - c & \text{if } i \in S, \\ 0 & \text{if } i = \Delta. \end{cases}$$

The next theorem gives a method of computing $E(N_k)$ using the modified CTMC $\{Z^k(t), t \geq 0\}$ with generator matrix Q^k and rate matrix R^k .

Theorem 5 Let $A = \{\mathbf{0}, \Delta\}$ and, let ρ and ρ_0 be as given below

$$\rho(z) = \begin{cases} \lambda_k & \text{if } z \in \{S : z_k = 0\}, \\ 0 & \text{otherwise.} \end{cases}$$

$$\rho_0(z) = \begin{cases} \lambda_k & \text{if } z \in \{S : z_k = 0\} \cup \{\Delta\}, \\ 0 & \text{otherwise.} \end{cases}$$

Let $g^A(x)$ be as in Equation 12, using the process (X, Z^k) . Then

$$E(N_k) = \frac{\lambda_k}{\lambda} + \sum_{i=1}^K \frac{\lambda_i}{\lambda} g_{e_i}^A(0).$$

Proof: Let $N_k(t)$ be the number of on periods of type k during $[0, t]$, including the one that may be continuing at time t . The counting process $N_k(\cdot)$ jumps up by one at time t if source k turns on at time t (i.e. $Z_k(t-) = 0$ and $Z_k(t) = 1$) in the following three cases:

1. $X(t-) = 0$,

2. $X(t-) > 0$ and $Z(t-) \neq \mathbf{0}$,

3. $X(t-) > 0$, $Z(t-) = \mathbf{0}$ and $Z^-(t-) \neq e_k$, where $Z^-(t)$ is the most recent state visited by the CTMC Z before it entered the current state $Z(t)$.

Now define the sequences of positive-valued random variables $\{\bar{T}_n, n \geq 1\}$ and $\{X_n, n \geq 1\}$ as follows. First define

$$\bar{T}_1 = \min\{t > 0 : Z_k(t) = 0 \text{ and we do not have } X(t-) > 0, Z(t-) = \mathbf{0} \text{ and } Z^-(t-) = e_k\}.$$

Hence \bar{T}_1 is the first time upon which one of the conditions 1-3 above applies. We now define

$$X_1 = \min\{t > 0 : Z_k(t + \bar{T}_1) = 1\}.$$

Plainly, $X_1 \sim \exp(\lambda_k)$ and it is easy to check that throughout the interval $[\bar{T}_1, \bar{T}_1 + X_1)$, one of the conditions 1-3 will apply. Hence $N_k(\bar{T}_1 + X_1) = 1$. We now develop our sequences of random variables inductively by

$$\bar{T}_n = \min\{t > \bar{T}_{n-1} + X_{n-1} : Z_k(t) = 0 \text{ and we do not have } X(t-) > 0, Z(t-) = \mathbf{0} \text{ and } Z^-(t-) = e_k\},$$

and

$$X_n = \min\{t > 0 : Z_k(t + \bar{T}_n) = 1\}.$$

The sequence $\{X_n, n \geq 1\}$ is of iid $\exp(\lambda_k)$ random variables and

$$N_k(t) = n, \quad \text{for } \bar{T}_n + X_n \leq t < \bar{T}_{n+1} + X_{n+1}.$$

Now fix $t \in [\bar{T}_n + X_n, \bar{T}_{n+1} + X_{n+1})$. It is clear from the above construction that the total time up to t for which one of the conditions 1-3 applies is bounded above and below by $\sum_1^{n+1} X_i$ and

$\sum_1^n X_i$, respectively. For general t , we express this mathematically as

$$\begin{aligned}
\sum_{i=1}^{N_k(t)} X_i &\leq \int_0^t I\{X(u) = 0, Z_k(u) = 0\}du + \int_0^t I\{X(u) > 0, Z(u) \neq \mathbf{0}, Z_k(u) = 0\}du \\
&\quad + \int_0^t I\{X(u) > 0, Z(u) = \mathbf{0}, Z^-(u) \neq e_k\}du \\
&= \int_0^t I\{X(u) = 0, Z(u) = 0\}du + \int_0^t I\{Z(u) \neq \mathbf{0}, Z_k(u) = 0\}du \\
&\quad + \int_0^t I\{X(u) > 0, Z(u) = \mathbf{0}, Z^-(u) \neq e_k\}du \\
&\leq \sum_{i=1}^{N_k(t)+1} X_i.
\end{aligned} \tag{32}$$

Here the last equality follows because

$$\begin{aligned}
&I\{X(u) = 0, Z_k(u) = 0\} + I\{X(u) > 0, Z(u) \neq \mathbf{0}, Z_k(u) = 0\} \\
= &I\{X(u) = 0, Z_k(u) = 0, Z(u) = \mathbf{0}\} + I\{X(u) = 0, Z_k(u) = 0, Z(u) \neq \mathbf{0}\} + \\
&I\{X(u) > 0, Z(u) \neq \mathbf{0}, Z_k(u) = 0\} = I\{X(u) = 0, Z(u) = \mathbf{0}\} + I\{Z(u) \neq \mathbf{0}, Z_k(u) = 0\}.
\end{aligned}$$

Taking expectations throughout the above inequality, and dividing by t we obtain

$$\begin{aligned}
\frac{E(N_k(t))}{t} &\leq \lambda_k E\left[\frac{1}{t} \int_0^t I\{X(u) = 0, Z(u) = \mathbf{0}\}du + \frac{1}{t} \int_0^t I\{Z(u) \neq \mathbf{0}, Z_k(u) = 0\}du\right. \\
&\quad \left. + \frac{1}{t} \int_0^t I\{X(u) > 0, Z(u) = \mathbf{0}, Z^-(u) \neq e_k\}du\right] \\
&\leq \frac{E(N_k(t)) + 1}{t}.
\end{aligned} \tag{33}$$

Using regenerative process theory, we get

$$\begin{aligned}
\lim_{t \rightarrow \infty} \frac{E(N_k(t))}{t} &= \frac{E(N_k)}{E(T)}, \\
\lim_{t \rightarrow \infty} \frac{1}{t} E\left(\int_0^t I\{X(u) = 0, Z(u) = \mathbf{0}\}du\right) &= \frac{1}{E(T)} E\left(\int_0^T I\{X(u) = 0, Z(u) = \mathbf{0}\}du\right), \\
\lim_{t \rightarrow \infty} \frac{1}{t} E\left(\int_0^t I\{Z(u) \neq \mathbf{0}, Z_k(u) = 0\}du\right) &= \frac{1}{E(T)} E\left(\int_0^T I\{Z(u) \neq \mathbf{0}, Z_k(u) = 0\}du\right), \\
\lim_{t \rightarrow \infty} \frac{1}{t} E\left(\int_0^t I\{X(u) > 0, Z(u) = \mathbf{0}, Z^-(u) \neq e_k\}du\right) &= \frac{1}{E(T)} E\left(\int_0^T I\{X(u) > 0, Z(u) = \mathbf{0}, Z^-(u) \neq e_k\}du\right).
\end{aligned}$$

All the expectations above are carried out under the initial condition $X(0) = 0, Z(0) = \mathbf{0}$. Hence, taking limits as $t \rightarrow \infty$ in 33 and using the above results from the regenerative process theory, we get

$$\begin{aligned}
E(N_k) &= E\left(\int_0^T \lambda_k I\{X(u) = 0, Z(u) = \mathbf{0}\}du\right) + \\
&\quad E\left(\int_0^T [\lambda_k I\{Z(u) \neq \mathbf{0}, Z_k(u) = 0\} + \lambda_k I\{X(u) > 0, Z(u) = \mathbf{0}, Z^-(u) \neq e_k\}]du\right)
\end{aligned} \tag{34}$$

Now, it is clear from the definition that the (X, Z) process stays in the state $(0, \mathbf{0})$ for an exponential amount of time with parameter λ during $[0, T]$. Hence,

$$E\left(\int_0^T I\{X(u) = 0, Z(u) = \mathbf{0}\}du\right) = 1/\lambda.$$

On the other hand

$$\begin{aligned} & E\left(\int_0^T [\lambda_k I\{Z(u) \neq \mathbf{0}, Z_k(u) = 0\} + \lambda_k I\{X(u) > 0, Z(u) = \mathbf{0}, Z^-(u) \neq e_k\}]du\right) \\ &= E\left(\int_D^T [\lambda_k I\{Z(u) \neq \mathbf{0}, Z_k(u) = 0\} + \lambda_k I\{X(u) > 0, Z(u) = \mathbf{0}, Z^-(u) \neq e_k\}]du\right). \end{aligned}$$

As in the proofs of Theorems 3 and 4, we now write

$$\tilde{T}_A = \min\{t > 0 : X(t + D) = 0, Z^k(t + D) \in A\},$$

and observe that with $A = \{\mathbf{0}, \Delta\}$, $D + \tilde{T}_A$ is stochastically identical to T . We also write

$$\tilde{Y}(t) = \int_D^{D+t} \rho(X(u), Z^k(u))du.$$

Furthermore, using the CTMC Z^k as defined earlier, we see that

$$\{Z(u) = \mathbf{0}, Z^-(u) = e_k\} \Leftrightarrow \{Z^k(u) = \Delta\}.$$

Hence the last expectation can be computed as

$$\begin{aligned} & E\left(\int_D^{D+\tilde{T}_A} [\lambda_k I\{Z(u) \neq \mathbf{0}, Z_k(u) = 0\} + \lambda_k I\{X(u) > 0, Z(u) = \mathbf{0}, Z^-(u) \neq e_k\}]du\right) \\ &= E\left(\int_D^{D+\tilde{T}_A} [\lambda_k I\{Z^k(u) \notin \{\mathbf{0}, \Delta\}, Z_k(u) = 0\} + \lambda_k I\{X(u) > 0, Z^k(u) = \mathbf{0}\}]du\right) \\ &= E\left(\int_D^{D+\tilde{T}_A} [\lambda_k I\{X(u) = 0, Z^k(u) \notin \{\mathbf{0}, \Delta\}, Z_k(u) = 0\} + \lambda_k I\{X(u) > 0, Z^k(u) \notin \{\mathbf{0}, \Delta\}, Z_k(u) = 0\} \right. \\ &\quad \left. + \lambda_k I\{X(u) > 0, Z^k(u) = \mathbf{0}\}]du\right) \\ &= E\left(\int_D^{D+\tilde{T}_A} [\lambda_k I\{X(u) = 0, Z_k(u) = 0\} + \lambda_k I\{X(u) > 0, Z^k(u) \neq \Delta, Z_k(u) = 0\}]du\right) \\ &= E\left(\int_D^{D+\tilde{T}_A} [\rho_0(Z^k(u))I\{X(u) = 0\} + \rho(Z^k(u))I\{X(u) > 0\}]du\right) \\ &= E\left(\int_D^{D+\tilde{T}_A} \rho(X(u), Z^k(u))du\right) \\ &= E(\tilde{Y}(\tilde{T}_A)) \\ &= \sum_{i=1}^K \frac{\lambda_i}{\lambda} E(\tilde{Y}(\tilde{T}_A) \mid X(D) = 0, Z^k(D) = e_i) \\ &= \sum_{i=1}^K \frac{\lambda_i}{\lambda} g_{e_i}^A(0). \end{aligned}$$

Here the third equality follows because $D \leq u \leq D + \tilde{T}_A$, $X(u) = 0$ implies $Z^k(u) \notin \{\mathbf{0}, \Delta\}$; the fourth equality follows from the use of the functions ρ_0 and ρ given in the theorem, and the fifth

k	λ_k	μ_k	r_k
1	1	1	2
2	1	2	3
3	1	3	4
4	1	4	5

Table 1: The Parameters of the Input Processes

equality follows from the definition of the function ρ given in Equation 16. The final equality is as in the conclusions to the proofs of Theorems 3 and 4. Substituting in Equation 34 yields the theorem.

Thus $E(T)$ can be computed using the results of Theorem 2 with the reward rate function given in Theorem 3, $E(T_k)$ can be computed using the results of Theorem 2 with the reward rate function given in Theorem 4, and $E(N_k)$ can be computed using the results of Theorem 2 with the reward rate function given in Theorem 5. The required quantity τ_k^{on} can now be computed by using Equation 9, and τ_k^{off} by using Equation 10. We illustrate the computations with an example in the next section.

6 Example

We consider a multi-class fluid queue with $K=4$ classes. The three parameters of the four input streams are given in Table 1. The mean input rate from each source is 1. However, the burstiness of source k increases with k . The total mean input rate is $m = 4$. The maximum peak rate (when all sources are on simultaneously) is $r = 14$. We plot τ_k^{on} , τ_k^{off} and r_k^o for $k = 1, 2, 3, 4$ as a function of c over $m = 4 < c < 14 = r$ in Figures 2, 3, and 4 below. Several interesting features can be gleaned from these graphs.

From Figure 2 we see that the mean on times of the output of source k decreases with c . This is as expected. It converges to $1/\mu_k$ as c increases to r . In fact we can see that for $c \geq r$ (=14 in this case), the fluid passes through the buffer uninterrupted and hence the output on-time periods are the same as the input on-time periods, which have mean $1/\mu_k$. The mean output burst length is never less than the mean input burst length.

Figure 3 shows that the mean off-times increase as a function of c , although this is not always true. For some parameter values they can be non-monotonic functions of c . We do not show these examples to save space. The limit of the mean output off-time of type k , as c increases to r , is $1/\lambda_k$, which is the mean input off-time of type k . The explanation for this is same as in the case of mean on-times.

A surprising feature of Figures 2 and 3 is that the mean on-times and the mean off-times for the output processes do *not* tend to infinity as c decreases to m (=4 in this case). However, the mean output burst of the superposition of all the output processes does tend to infinity as c decreases to m . Indeed the output process analysis is rather complicated in the case of an unstable fluid queue, and we do not have complete answers for this case. We have been unable to find a simple expression for the limit of the mean on and off times as c decreases to m .

Figure 2

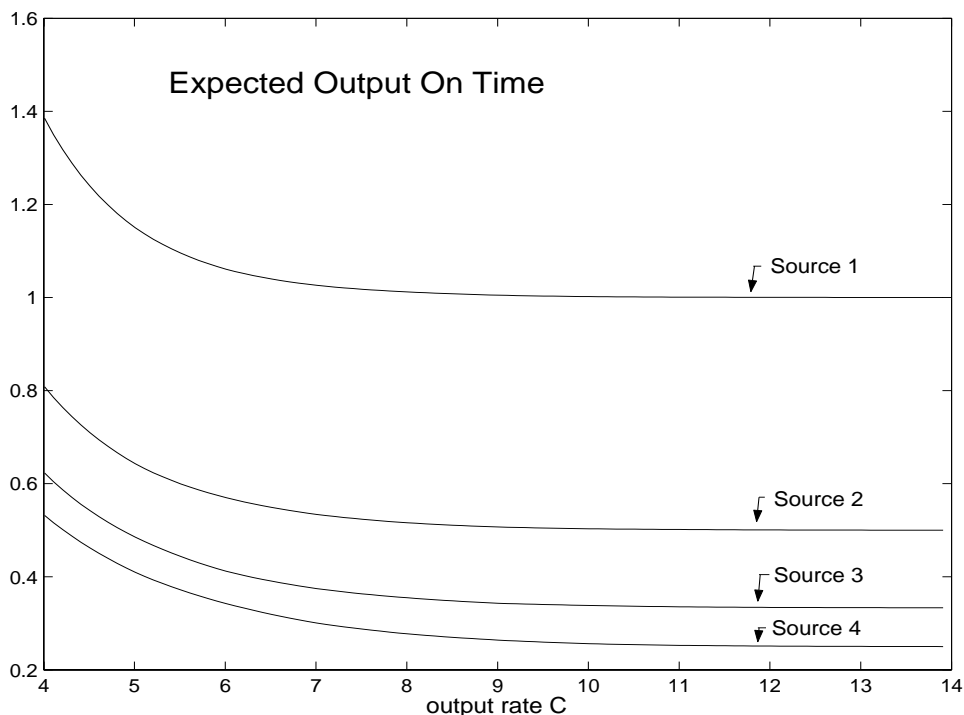


Figure 4 shows the effective peak rate as a function of c for the four sources. As expected, it is an increasing function of c and approaches the input peak rate as c increases to r . In fact, for $c \geq r$ we have $r_k^o = r_k$ by the same argument as before. Note that the sum of the effective peak rates of the four sources can be more than c . This is a consequence of the definition of the output peak rate, and it shows the effect of statistical multiplexing of traffic. It is interesting to note that for source 4 (with peak input rate 5), the effective output peak rate increases linearly for $4 < c < 5$. In fact our numerical experimentation shows that

$$r_k^o = \frac{r_k}{r_k - m_k + m} c, \quad m < c \leq r_k.$$

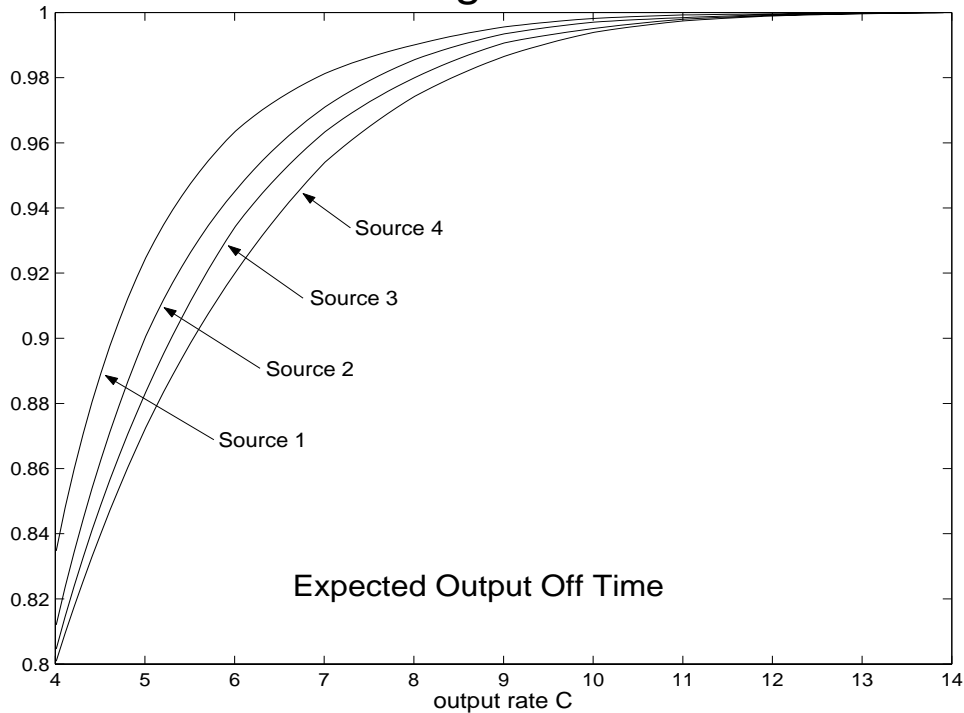
and that

$$\lim_{c \downarrow m} r_k^o = \frac{r_k}{r_k - m_k + m} m, \quad 1 \leq k \leq K.$$

We do not have formal proofs of these results. However, the first relation is proved to hold for $0 \leq c \leq \max(r_k, m)$ by entirely different methods in Hirasawa [4], and the second follows from this.

As mentioned in the introduction, the motivation for this research is to approximate the output processes of the individual sources by on-off processes with exponential on and off times whose means are computed by the methods given here. A natural question is: how good is this approximation? Here we show two simple simulation results that show that the approximation is indeed excellent, at least for the parameters chosen. The parameters of the four sources are as before, and c is chosen to be $22/3$. The simulation was run until the output processes of each source contained at least 10,000 on times. The numerically computed mean on time was .294, and mean off time was .967 for source 4, and its output process contained 16,113 on and off times. Figure 5 shows the exponential pdf with mean .294 superimposed with the pdf produced by the simulation using the

Figure 3



16,113 observations of on-times from the output stream of source 4 clubbed into 100 bins. Figure 6 shows the exponential pdf with mean .967 superimposed with the pdf produced by the simulation using the 16,113 observations of off-times from the output stream of source 4 clubbed into 100 bins.

The complete evaluation of the quality of the proposed approximation requires a rather extensive simulation experimentation that is underway as part of the Multi-Class Fluid Network analysis effort, and will be published subsequently.

7 Multi-class M/G/1 Queue

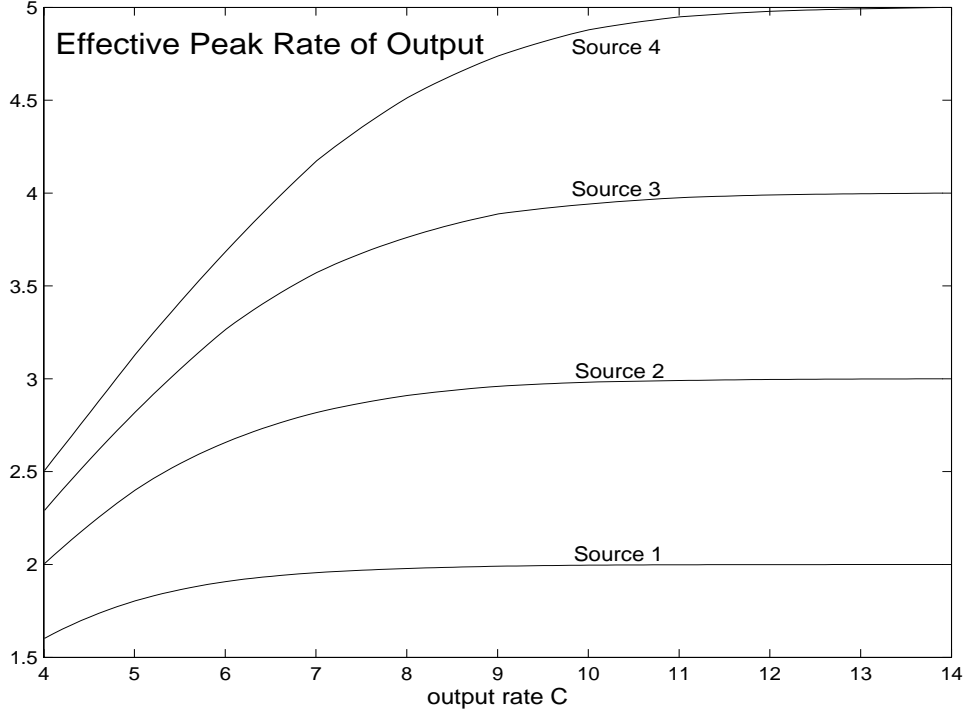
The output analysis of the previous sections can be done even more simply for a standard multi-class M/G/1 queue, although the motivation of the fluid networks does not apply here. We report the results here as they may be of independent interest.

Consider a single server queue with K classes of customer. Customers of class k ($1 \leq k \leq K$) arrive according to a $PP(\lambda_k)$ and form a single line. They are served by the single server according to a first come first served discipline. The service times of class k customers are iid with mean ν_k . The classes are independent of each other.

A multi-class M/G/1 queue with an FCFS service discipline may be regarded as a standard single-class M/G/1 queue with arrival rate

$$\lambda = \sum_{k=1}^K \lambda_k, \quad (35)$$

Figure 4



and mean service time

$$\nu = \sum_{k=1}^K \lambda_k \nu_k / \lambda. \quad (36)$$

Let

$$\rho_k = \lambda_k \nu_k. \quad (37)$$

The queue is stable if

$$\rho = \sum_{k=1}^K \rho_k < 1. \quad (38)$$

Define $S(t)$ to be the state of the server as follows: $S(t) = 0$ if the server is idle at time t , and $S(t) = k$ if the server is serving a customer of class k at time t . (Unlike in the fluid case, here the server cannot serve more than one type of customer at a time.) Clearly the $\{S(t), t \geq 0\}$ process is a regenerative process with state space $\{0, 1, \dots, K\}$, and it regenerates whenever it enters state 0. In this section we compute τ_k , the expected sojourn time in state k , and τ_{kk} , the expected inter visit time to state k . The main result is given in the following theorem:

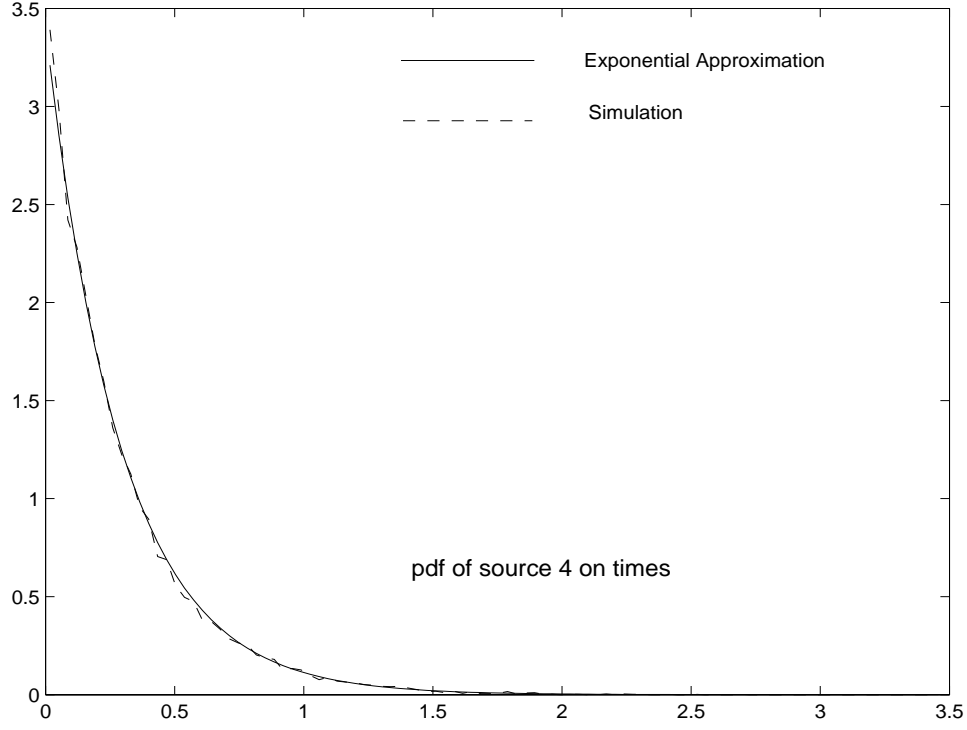
Theorem 6 *Assume that the queue is stable. Then,*

$$\tau_k = \frac{\lambda \nu_k}{(1 - \rho) \lambda_k + (\lambda - \lambda_k)}, \quad 1 \leq k \leq K, \quad (39)$$

and

$$\tau_{kk} = \frac{\lambda}{(1 - \rho) \lambda_k^2 + \lambda_k (\lambda - \lambda_k)} = \frac{\tau_k}{\rho_k}, \quad 1 \leq k \leq K, \quad (40)$$

Figure 5



Proof: Since the queue is assumed to be stable, ρ_k is the long run fraction of the time the server is busy serving customers of class k . Suppose $S(0) = 0$, and define T to be the first time the S process re-enters state 0. Furthermore, define T_k to be the total time spent in state k by the S process during $(0, T]$, and N_k be the total number of visits to state k during $(0, T]$. Results from regenerative processes imply that

$$\rho_k = \frac{E(T_k)}{E(T)}, \quad (41)$$

$$\tau_k = \frac{E(T_k)}{E(N_k)}, \quad (42)$$

$$\tau_{kk} = \frac{E(T)}{E(N_k)}. \quad (43)$$

Since $E(T)$ is the expected length of a busy cycle in an M/G/1 queue with arrival rate λ and mean service times ν as given in Equations 35 and 36, we have

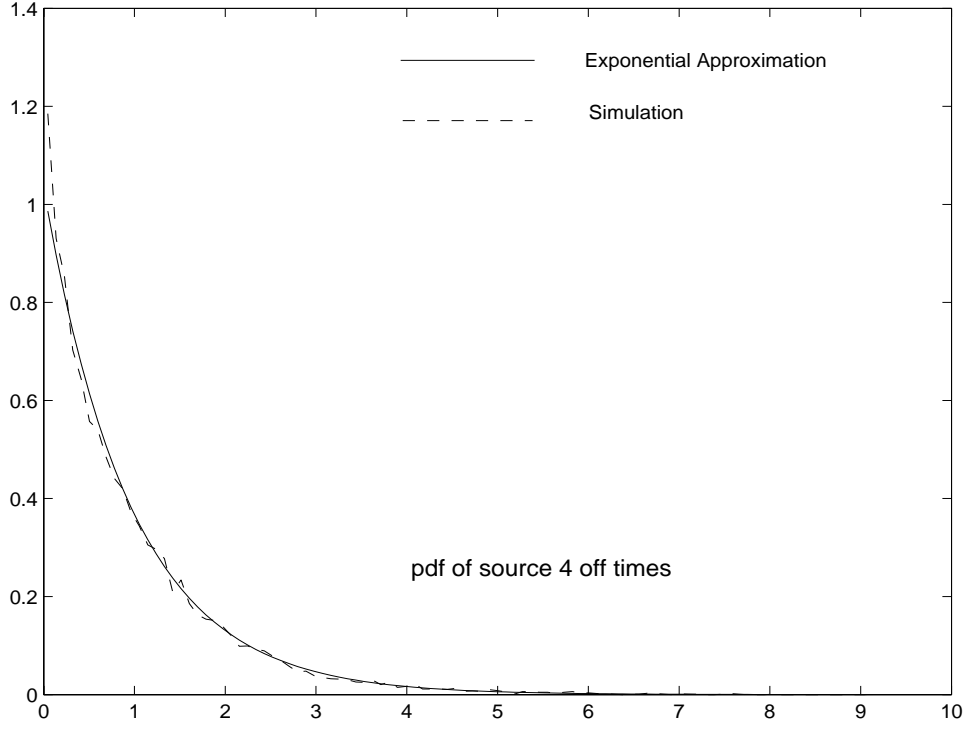
$$E(T) = \frac{1}{\lambda} + \frac{\nu}{1 - \rho} = \frac{1}{\lambda(1 - \rho)}, \quad (44)$$

where ρ is as defined in Equation 38. Using Equation 41 we get

$$E(T_k) = \frac{\rho_k}{\lambda(1 - \rho)}. \quad (45)$$

where ρ_k is as defined by Equation 37.

Figure 6



Next we compute $E(N_k)$. Let N be the total number of customers served during the first busy cycle, and let X_n be the type of the n th customer ($1 \leq n \leq N$). Then, FCFS service discipline implies that

$$N_k = 1\{X_1 = k\} + \sum_{n=2}^N 1\{X_{n-1} \neq k, X_n = k\}.$$

Hence,

$$E(N_k) = P(X_1 = k) + E\left(\sum_{n=2}^N 1\{X_{n-1} \neq k, X_n = k\}\right).$$

Using the fact that $\{X_n, n \geq 1\}$ is a sequence of iid random variables with common pmf

$$P(X_n = k) = \frac{\lambda_k}{\lambda}, \quad 1 \leq k \leq K,$$

we get

$$E(N_k) = \frac{\lambda_k}{\lambda} + (E(N) - 1) \frac{\lambda_k}{\lambda} \left(1 - \frac{\lambda_k}{\lambda}\right).$$

From the results for standard M/G/1 queue, we know that

$$E(N) = \frac{1}{1 - \rho}.$$

Using this we get

$$E(N_k) = \frac{\lambda_k^2}{\lambda^2} + \frac{1}{1 - \rho} \cdot \frac{\lambda_k}{\lambda} \left(1 - \frac{\lambda_k}{\lambda}\right). \quad (46)$$

Using Equations 46 and 45 in Equations 42 and 43 we get Equations 39 and 40. This completes the proof.

Note that τ_k and τ_{kk} do *not* tend to infinity as $\rho \rightarrow 1$, although $E(T)$ does. This phenomenon was observed in the fluid case as well.

8 Conclusions

In this paper we have presented the analysis of the output processes from a multi-class queue, both the fluid case and the ordinary case. In both cases we have derived results about the mean on and off times of the output processes of the different types. In the fluid case these results take the form of algorithms to compute these quantities, while in the regular case we have derived explicit results.

9 Acknowledgments

The authors wish to thank Dr. Ivo Adan for fruitful discussions that helped clarify some subtle points in the calculation of $E(N_k)$ in Section 5.

References

- [1] Anick, D., D. Mitra, and M. M. Sondhi. (1982) Stochastic theory of a data-handling system with multiple sources. *The Bell Syst. Tech. J.*, **61**, 1871–1894.
- [2] Chandy, K. M. and C. H. Saur. (1978) Approximate Methods for Analyzing Queueing Network Models of Computing Systems. *ACM Computing Surveys*, **10**, 281-317.
- [3] Gelenbe, E. and I. Mitrani. (1980) *Analysis and Synthesis of Computer Systems*. Academic Press, NY.
- [4] Hirasawa, Y. (2000) *Approximating Traffic Parameters in Multi-Class Fluid Networks*. Ph.D. Thesis, Department of Operations Research, University of North Carolina, Chapel Hill, NC 27599-3180.
- [5] Kaspi, H. and O. Kella. (1996) Stability of Feed-Forward Fluid Networks with Levy Input. *J. Appl. Prob.*, **33**, 513-532.
- [6] Kella, O. (1993) Parallel and Tandem Fluid Networks with Dependent Levy Inputs. *Ann. Appl. Prob.* **3**, 682-695.
- [7] Kella, O. (1996) Stability and Non-product Form of Stochastic Fluid Networks with Levy Inputs. *Ann. Appl. Prob.* **6**, 186-199.
- [8] Kella, O. and W. Whitt. (1990) A Tandem Fluid Network With Levy Input. *Technical Report 90-16*, Yale University, CT 06511.
- [9] Kella, O. and W. Whitt. (1990) Linear Stochastic Fluid Networks. *J. Appl. Prob.*, **36**, 244-260.
- [10] Kosten, L. (1974) Stochastic theory of a multi-entry buffer, part 1. *Delft Progress Report*, 10-18.
- [11] Kuehn, P. J. (1979) Approximate Analysis of General Queueing Networks by Decomposition. *IEEE Trans. Commun.*, **COM-27**, 113-126.
- [12] Kulkarni, V. G. (1995) *Modeling and Analysis of Stochastic Systems*. CRC Press, UK.

- [13] Kulkarni, V. G. (1997) Fluid models for single buffer systems. *Frontiers in Queueing: Models and Applications in Science and Engineering*, 321-338, Ed. J. H. Dshalalow, CRC Press.
- [14] Kulkarni, V. G. and T. Rolski. (1994) Fluid model driven by an Ornstein-Uhlenbeck process. *Prob. Eng. Inf. Sci.*, **8**,403-417.
- [15] Kulkarni, V. G. and E. Tzenova. (2001) Mean First Passage Times in Fluid Queues. To appear in *Oper. Res. Lett.*
- [16] Reiser, M. and H. Kobayashi. (1974) Accuracy of the Diffusion Approximation for Some Queueing Systems. *IBM J. Res. Dev.*, **18**, 110-124.
- [17] Sevcik, K. C., A. I. Levy, S. K. Tripathi and J. L. Zahorjan. (1977) Improving Approximations of Aggregated Queueing Network Subsystems. *Computer Performance*, Eds: Chandy and Reiser, Amsterdam, 1-22.
- [18] Whitt, W. (1983) The Queueing Network Analyzer. *The Bell Syst. Tech. J.*, **62**, 2779-2815.