On Fluid Queueing Systems with Strict Priority

Yong Liu, *Member, IEEE,* and Weibo Gong, *Fellow, IEEE*

***Abstract*— We consider priority fluid queueing systems where high priority class has strict priority access to service. Sample path analysis tools, such as Poisson Counter Driven Stochastic Differential Equation, are employed to study system queueing behavior in steady state. We are able to obtain various analytical results for different fluid traffic models and system configurations. Those results can be used as a general rule of thumb in buffer dimensioning and other traffic engineering issues.**

***Index Terms*— Priority Fluid Queue, Poisson Counter Driven Stochastic Differential Equation, Tandem Fluid System, Traffic Autocorrelation.**

I. Introduction

Fluid models are becoming increasingly important in several areas, including communication networks, manufacturing systems, transportation systems and other networks. They have been widely used as burst scale models for high speed communication networks [1], [2], [3], [4]. In a fluid model, discrete packets and cells within bursts are modeled as continuous fluid. Unlike classical queueing models, which assume renewal arrivals, fluid models can capture autocorrelation in arrival processes. The continuous nature of fluid also makes fluid queueing model more tractable analytically. Many results have been obtained for a variety of fluid queueing systems [2], [5], [6], [7]. In this paper, we use priority fluid queueing model to study the performance of routers with priority service. We employ various sample path description techniques to analytically characterize the queueing behavior of both high and low priority buffers under different traffic patterns and system configurations. This is in contrast to the tendency to move as quickly as possible away from flow equations to problem descriptions in terms of probabilities. As our results indicate, it is often more informative and effective to formulate and solve a problem of interest using sample path methods. Sample path approaches looks carefully into the dynamic behavior of the system, and naturally lead to differential equation based descriptions. Although in queuing systems most closed form solutions can only be obtained for steady state, the derivations are often started with the sample path dynamic evolution. We also hope that the differential equation based approach in this paper would bring the subject of controlling queues in closer contact with other branches of control theory [8], [9].

Traditionally, the Internet provides applications with a single class of best-effort service. With the increase of link bandwidth and processing power of routers and end hosts, it is possible for the Internet to support a large variety of applications with different quality of service requirements. Some applications, such as real audio and real video applications, are delay sensitive. It is crucial for the data to be delivered in time. At the same time, they can tolerate some loss. Some applications, like HTTP, FTP, Email, are loss sensitive. They require reliable and delay tolerant data delivery. The diversity in network applications' quality of service requirements calls for a network architecture that supports Differentiated Services (DiffServ [10]) for the end users. In such an architecture, routers at the network edge and core routers cooperate with each other to provide service differentiation. Edge routers classify packets into different service classes. They also monitor users' traffic and mark those packets falling out of users' traffic profiles. Core routers schedule packets according to their service classes and marks put on them. The simplest priority scheduling scheme is strict priority scheduling. Basically, there are two classes of services: high priority service and low priority service. Low priority traffic can only be served when there is no demand from the high priority traffic. High priority traffic has stringent quality of service requirements, e.g., delay, delay jitter, etc. Quality of service received by high priority traffic is independent of the existence of low priority traffic. It is possible to meet high priority traffic's stringent service requirements by admission control [11]. Low priority traffic has much looser requirement. It receives the "Best Effort Service". However, for the purposes of resource provisioning and dimensioning, performance analysis is still needed for low priority traffic.

In our priority fluid queueing model, high priority traffic consists of multiple ON-OFF processes with generally distributed ON-periods and exponentially distributed OFF-periods. On the other hand, the low priority traffic is modeled as a constant bit rate (CBR) flow. Various analytical results are obtained for different system configurations. For single node priority queue, we explicitly solve the stationary distribution for the low priority buffer. For the high priority buffer, we obtain its content distribution when either there is only one high priority flow or there are multiple flows whose active periods are exponentially distributed. We also investigate priority fluid queueing system with multiple nodes in tandem. At each node, we obtain buffer content's first order statistics for each priority class. Stationary distributions of individual high priority buffer and aggregated low priority buffers are also derived.

The paper is organized as follows. In Section II, we briefly introduce previous work on priority fluid queue. We then describe the mathematical model of the our priority queueing system and discuss its validation in real communication networks in Section III. In Section IV, we study single node priority fluid queue under different traffic models. In Section V, we extend our model to investigate priority fluid queues in tandem. The paper concludes in Section VI.

Yong Liu is with Computer Science Department, University of Massachusetts, email yongliu@cs.umass.edu; Weibo Gong is with Department of Electrical and Computer Engineering, University of Massachusetts, Amherst, email gong@ecs.umass.edu

This work is supported in part by DARPA under Contract DOD F30602-00-0554 and by NSF under grants EIA-0080119, ITR-0085848, and ANI-9980552.

II. Related Work

Most previous work of priority fluid queue focused on Markov Modulated Fluid Flow (MMFF) model for both high priority and low priority traffic. Zhang [12] studied a Markov modulated fluid queueing system with strict priority. The sending rates of high priority and low priority flows are regulated by a Markov chain. For two states Markov chain model, he obtained closed form buffer content distributions of both priorities. Elwalid and Mitra [13] investigated the same problem. By approximating the distribution of the non-empty period of high priority buffer with exponential distribution, they were able to develop analytical approximations for queue length distributions in two buffers. They studied the admission control problem based on that approximation. Choi [14] derived the Laplace transform of the stationary joint distribution of two buffers when the underlying Markov Chain has any finite number of states. Knessl and Tier [15] used another fluid model with one Markov ON-OFF high priority flow and one Constant Bit Rate low priority flow. They explicitly solved the joint distribution of two buffer contents and constructed approximation for its tail behavior. In contrast to these models, our model allows more general statistic behavior for high priority traffic. We model high priority flows as multiple ON-OFF processes with generally distributed ON periods and exponentially distributed OFF periods. This model is very flexible to capture a variety of traffic autocorrelation structures, from short range dependence to long range dependence. Based on this model, various analytical results are established. We will describe our model in details in the next section. Some of our work has been presented in [16].

III. Mathematical Model

In this section, we describe our model for a single node priority fluid queue. In Section V, we will extend it to multiple priority fluid queues in tandem.

We model high priority flows as ON-OFF processes with generally distributed ON periods and exponentially distributed OFF periods. The reason for choosing ON-OFF flow model for high priority traffic is that in communication network many real time applications exhibit ON-OFF traffic pattern. ON-OFF model captures traffic autocorrelation structure which will be shown in following sections to have linear impact on high priority queue. On the other hand, the queueing behavior of low priority buffer is determined by the variations in both arrival rate of low priority traffic and service rate left over by high priority traffic. Low priority flows are typically less bursty than high priority flows. We choose a simple constant bit rate (CBR) flow model for low priority traffic to focus on high priority traffic's impact on low priority buffer. The CBR flow model is a good approximation of low priority traffic whenever the variance in its aggregate rate is much smaller than the variance in its service rate. For example, if the low priority traffic comes from one upstream bottle-neck node, it can be well approximated by a CBR flow. Our model is abstract in nature and it is not our intention to model the details of real queues in the routers and switches. The purpose of this analysis is to exhibit the impact of traffic statistics on the buffer contents. For example, the mean buffer content of both the high priority buffer and the low priority buffer is shown to be proportional to the autocorrelation time constants of higher priority flows when they are Markov ON-OFF sources.

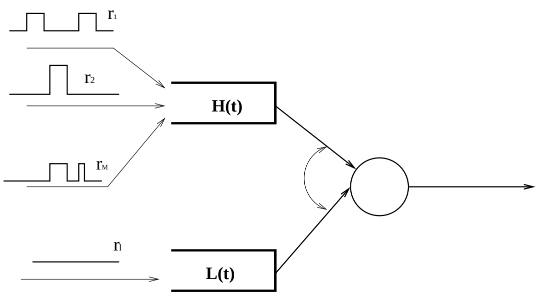


Fig. 1. Fluid queueing system with two priorities

Let's consider the fluid queue with strict priority as in Figure 1. We assume infinite capacity for both high and low priority buffer. The maximum outflow rate of the fluid server is normalized to 1. The high priority buffer is fed by M ON-OFF sources. The sending rate of source i, $i = 1, .., M$, is regulated by an ON-OFF process $x_i(t)$. When $x_i(t) = 1$, source i sends traffic to the buffer at rate $r_i > 1$. Let A_{in}, S_{in} be $x_i(t)$'s nth active period and silent period respectively. $\{A_{in}\}$ and $\{S_{in}\}$ are two sets of mutually independent $i.i.d$ random variables. A_{in} has general distribution with mean $E[A_{in}] = \alpha_i$; S_{in} is exponentially distributed with mean $1/\lambda_i$. Low priority buffer is fed by a constant bit rate (CBR) flow with rate $r_l < 1$. High priority traffic has strict priority access to service: when the high priority buffer is non-empty, it is drained out at rate 1. Low priority traffic is served only when there is no backlog in the high priority buffer. In order for the system to be stable, we enforce

$$\sum_{i=1}^{M} \frac{\alpha_i \lambda_i r_i}{1 + \alpha_i \lambda_i} + r_l < 1 \tag{1}$$

Denote by $H(t)$ and $L(t)$ the high priority buffer and low priority buffer content respectively. The sample path description of the system is :

$$\frac{d}{dt} H(t) = \sum_{i=1}^{M} r_i x_i(t) - I_{H(t)} \tag{2}$$

$$\frac{d}{dt} L(t) = r_l I_{H(t)} + (r_l - 1)(1 - I_{H(t)}) I_{L(t)}, \tag{3}$$

where $I_{f(t)}$ stands for the indicator function $1(f(t) > 0)$. We will also use this notation in following sections.

IV. Single Buffer System

In this section we study priority fluid queue at a single node. We start with presenting some previous results of single class fluid queue with both single and multiple ON-OFF sources. In Section IV-B, we briefly introduce Poisson Driven Stochastic Differential Equation and its application in fluid queue study. In Section IV-C, we study priority fluid queue fed by one high priority flow and one low priority flow. Various results are established for both high priority and low priority buffer.

Priority fluid queue fed by multiple ON-OFF high priority flows are considered in Section IV-E.

A. Results for Fluid Queue without Priorities

A fluid queue fed by single ON-OFF source has been studied thoroughly in [5], [17], [18], [19]. It is converted to an equivalent $M/G/1$ queue to get analytical results. Assume the sending rate of the source is regulated by a ON-OFF process $x(t)$. When $x(t) = 1$, the source injects fluid into the buffer at rate h. The service rate of the server is c, which is smaller than the source peak rate h. Let A_n, S_n be the length of $x(t)$'s nth active period and silent period respectively. $\{A_n\}$ and $\{S_n\}$ are two sets of mutually independent $i.i.d$ random variables. A_n has a general distribution with Laplace-Stieltjes transform $G(\theta)$ and mean $E[A_n] = \alpha$. Silent periods are exponentially distributed, with mean $1/\lambda$. Let Z be the stationary buffer content. The following Lemma is due to Corollary 3 in [19] and Lemma 3.2 in [20], which relate the buffer content of a fluid queue to the workload of a $M/G/1$ queue.

Lemma 1: Let W denote the stationary workload of a $M/G/1$ queue with arrival rate λ/c and service time distribution with Laplace-Stieltjes transform $G((h-c)\theta)$. The queue is stable if and only if $\rho \equiv \alpha\lambda h/(c+\alpha\lambda c) < 1$. When the queue is stable,

$$\begin{aligned} P\{Z > z\} &= \gamma P\{W > z\}, \forall z \geq 0 \qquad (4)\\ E[e^{-\theta Z}] &= 1 - \gamma + \gamma E[e^{-\theta W}], \qquad (5) \end{aligned}$$

where $\gamma \equiv h/((h-c)(1+\lambda\alpha))$.
It can also be derived by the generalized distributional Little's law derived in [21]. With this Lemma, we can obtain the stationary distribution of fluid buffer content Z.

The server's output process $O(t)$ is another ON-OFF process: when the buffer is non-empty, $O(t) = 1$; when the buffer is empty, $O(t) = 0$. After the buffer drains, it remains empty until the source turns on again. Because the OFF periods of the source are exponentially distributed, the silent periods of $O(t)$ are also exponentially distributed with mean $1/\lambda$. The following Lemma characterizes the busy period of the server. It is due to Proposition 1 in [5].

Lemma 2: Let B be a typical busy period of the server, (or equivalently typical active period of $O(t)$),
(i) B has the same distribution as that of a busy period in a $M/G/1$ queue with arrival rate $\lambda(1-c/h)$ and service time distribution $G(cx/h)$;
(ii) $P\{B < \infty\} = 1$ if and only if $\rho < 1$
(iii) $E[B] < \infty$ if and only if $\rho < 1$, and $E[B] = \rho/(\lambda(1-\rho))$

For the case when there are M ($M > 1$) ON-OFF sources, to our knowledge, there is no analytical formula for the stationary distribution of buffer content. The stochastic characteristics of the output process $O(t)$ have been studied by Kaspi and Rubinovitch [22], Aalto [5] and Boxma [2]. Input and service rates are normalized such that the service rate is 1 and input rate of source i is $r_i > 1$. Denote by A_{i0} the typical active period of source i, $F_i(\theta) = E[e^{-\theta A_{i0}}]$, $E[A_{i0}^k] = \alpha_i^{(k)}$, $i = 1, 2, .., M$ and $k = 1, 2, ...$ The following Lemma regarding the output process $O(t)$ of the server is a consequence of Theorem 5.2 in [2].

Lemma 3: $O(t)$ is an ON-OFF process. The typical silent period S_o is exponentially distributed with rate $\lambda = \sum_{i=1}^{M} \lambda_i$; the typical active period A_o has general distribution, which has Laplace-Stieltjes transform

$$\pi(\theta) = E[e^{-\theta A_o}] = \sum_{i=1}^{M} \frac{\lambda_i}{\lambda} \pi_i(\theta),$$

where $\{\pi_i(\theta)\}_{1 \leq i \leq M}$ is the unique solution in $[0,1]^M$ of the equations:

$$\pi_i(\theta) = F_i(r_i\theta + \lambda_i(r_i - 1)(1 - \pi_i(\theta)) + \sum_{j \neq i} \lambda_j r_i (1 - \pi_j(\theta)))$$

B. Poisson Driven Stochastic Differential Equation

In [6], [23], Poisson Driven Stochastic Differential Equation (PDSDE) has been introduced to study fluid queueing systems when sources' ON and OFF periods are exponentially distributed,

Consider the following stochastic integral equation

$$x(t) = x(0) + \int_0^t f(x(\tau), \tau) d\tau + \int_0^t g(x(\tau), \tau) dN_\tau \qquad (6)$$

where $\{N_\tau\}$ is a Poisson counting process.

Definition 1: $x(\cdot)$ is a solution of (6) in the Itô sense if, on an interval where N is constant, x satisfies $\dot{x} = f(x,t)$ and if, when N jumps at t_1, x changes according to

$$\lim_{t \to t_1^+} x(t) = g(\lim_{t \to t_1^-} x(t), t_1) + \lim_{t \to t_1^-} x(t)$$

and $x(\cdot)$ is taken to be left-continuous.
When this definition is in force, it is common to rewrite equation (6) as

$$dx(t) = f(x,t)dt + g(x,t)dN(t)$$

More generally, a stochastic differential equation can be driven by multiple Poisson Counting processes:

$$dx(t) = f(x,t)dt + \sum_{i=1}^{m} g_i(x,t) dN_i(t) \qquad (7)$$

It is called Poisson Driven Stochastic Differential Equation (PDSDE). Some important properties of PDSDE are listed below:

- If $\psi : R^n \to R$ is a differentiable function, then the random process $\psi(t) \triangleq \psi(x(t))$ is described by the following S.D.E

$$\begin{aligned} d\psi(t) = &\langle \frac{\partial \psi}{\partial x}, f(x) \rangle dt \\ &+ \sum_{i=1}^{m} [\psi(x(t) + g_i(x(t))) - \psi(x(t))] dN_i, \end{aligned}$$

  where $< x, y >$ stands for the dot inner-product of vector x and y
- Because $\{N_i(t)\}$ are independent of $\{x(\tau), \tau < t\}$, we have

$$\begin{aligned} \frac{d}{dt} E[\psi(t)] = &E[\langle \frac{\partial \psi}{\partial x}, f(x) \rangle] \\ &+ \sum_{i=1}^{m} E[\psi(x(t) + g_i(x(t))) - \psi(x(t))] \lambda_i \end{aligned}$$

- If $x(t)$ has a smooth density function $\rho(t,x)$, it satisfies
$$\begin{aligned}\frac{\partial\rho(t,x)}{\partial t} &= -\frac{\partial}{\partial x}[f(x)\rho(t,x)] \\ &+ \sum_{i=1}^{m}\lambda_i\left(\rho(t,\hat{g}_i^{-1}(x))\left|\det(I+\frac{\partial g_i}{\partial x})\right|^{-1} - \rho(t,x)\right)\end{aligned}$$
where λ_i is the rate of Poisson Counter N_i, $\hat{g}_i(x) = x + g_i(x)$ [23]. This equation is the counterpart of the Fokker-Planck equation for the Wiener process driven systems.

A Markov ON-OFF process $x(t)$ can be described by a simple PDSDE:

$$dx(t) = (1-x(t))dN_1(t) - x(t)dN_2(t) \quad x(0)\in\{0,1\} \tag{8}$$

where $N_1(t)$ is a Poisson Counter with rate λ, $N_2(t)$ is a Poisson Counter with rate μ. Take the expectation on both sides of (8),

$$\frac{d}{dt}E[x(t)] = (1-E[x(t)])\lambda - E[x(t)]\mu \tag{9}$$

In steady state, $E[x] = \lambda/(\lambda+\mu)$. In order to compute the temporal correlation of $x(t)$, consider

$$dx(t)x(0) = (1-x(t))x(0)dN_1(t) - x(t)x(0)dN_2(t) \tag{10}$$

Taking the expectation on both sides of equation (10):

$$\frac{d}{dt}E[x(0)x(t)] = -(\lambda+\mu)E[x(0)x(t)] + \lambda E[x(0)]$$

together with initial condition $E[x(0)x(0)] = E[x(0)] = \lambda/(\lambda+\mu)$. We can solve for the correlation function of the source

$$R_{xx}(\tau) \triangleq E[x(0)x(\tau)] = \frac{\lambda}{(\lambda+\mu)^2}(\mu e^{-(\lambda+\mu)\tau} + \lambda) \tag{11}$$

The correlation of Markov ON-OFF traffic decays exponentially with time constant $1/(\lambda+\mu)$.

The sample path description of the buffer content $Z(t)$ is

$$dZ(t) = -c\times I_Z dt + h\times x(t)dt \tag{12}$$

The moments of Z can be obtained from sample paths of quantities of the form $Z^i(t)\times x^j(t)$, where $i = 1,2,\cdots$ and $j\in\{0,1\}$ [6]. For the first order moment

$$EZ = \frac{(h-c)h\lambda}{c\lambda+c\mu-h\lambda}\times\frac{1}{\lambda+\mu} = \frac{h-c}{\frac{c}{hE[x]}-1}\times\frac{1}{\lambda+\mu} \tag{13}$$

From equation (13) we can see that the auto-correlation constant of the Markov ON-OFF source, $1/(\lambda+\mu)$, has a linear impact on the average queue length.

The nth moments of steady state Z can be obtained quite similarly and the result is

$$EZ^n = \frac{cn(h-c)}{c\lambda+c\mu-h\lambda}\times EZ^{n-1},\ \ n>1. \tag{14}$$

C. Single High Priority Flow and Single Low Priority Flow

For a priority fluid queue with only one high priority flow, we can simplify notations in Section III: denote by A_0 and S_0 the typical active and silent period of $x(t)$. A_0 has general distribution $F(x)$ and its Laplace-Stieltjes transform $F(\theta) = E[e^{-\theta A_0}]$. The kth moment of A_0 is $E[A_0^k] = \alpha^{(k)}$, $k\geq 1$. S_0 has exponential distribution and $E[S_0] = 1/\lambda$. When $x(t) = 1$, the rate of high priority flow is $r>1$.

Because high priority traffic has strict service priority, the evolution of the high priority queue is independent of low priority traffic. It can be treated as a normal fluid queue driven by one ON-OFF source. Based on Lemma 1 and Lemma 2, we are ready to present the results for priority queue fed by single high priority source and single low priority source.

Theorem 4: As long as $\alpha^{(1)}\lambda r/(1+\alpha^{(1)}\lambda) + r_l < 1$, both the high priority and low priority queues are stable. Let H and L be the stationary buffer content of high priority and low priority buffer respectively, then:
(i) For the high priority buffer $H(t)$,

$$E[e^{-\theta H}] = 1-\gamma+\gamma\frac{(1-(r-1)\lambda\alpha^{(1)})\theta}{\theta-\lambda+\lambda F((r-1)\theta)} \tag{15}$$

$$E[H] = \frac{\alpha^{(2)}(r-1)\rho_h}{2\alpha^{(1)}(1+\lambda\alpha^{(1)})(1-\rho_h)} \tag{16}$$

$$\begin{aligned}E[H^2] &= \frac{\alpha^{(3)}(r-1)^2\rho_h}{3\alpha^{(1)}(1+\lambda\alpha^{(1)})(1-\rho_h)} \\ &+ \frac{(\alpha^{(2)}\rho_h)^2(r-1)^3}{2r(1+\lambda\alpha^{(1)})(\alpha^{(1)}(1-\rho_h))^2},\end{aligned} \tag{17}$$

where $\rho_h \equiv \alpha^{(1)}\lambda r/(1+\alpha^{(1)}\lambda)$, $\gamma\equiv r/((r-1)\times(1+\lambda\alpha^{(1)}))$;
(ii) For the low priority buffer $L(t)$,

$$E[e^{-\theta L}] = 1-\eta+\eta\frac{(1-r_l-r_l\lambda\beta^{(1)})\theta}{(1-r_l)\theta-\lambda+\lambda B(r_l\theta)} \tag{18}$$

$$E[L] = \frac{\beta^{(2)}r_l\rho_l}{2\beta^{(1)}(1+\lambda\beta^{(1)})(1-\rho_l)} \tag{19}$$

$$\begin{aligned}E[L^2] &= \frac{\beta^{(3)}r_l^2\rho_l}{3\beta^{(1)}(1+\lambda\beta^{(1)})(1-\rho_l)} \\ &+ \frac{(\beta^{(2)}\rho_l)^2 r_l^3}{2(1+\lambda\beta^{(1)})(\beta^{(1)}(1-\rho_l))^2},\end{aligned} \tag{20}$$

where $B(\theta)$ is the Laplace-Stieltjes transform of typical busy period of a $M/G/1$ queue with arrival rate $\lambda(1-1/r)$ and service time distribution $F(x/r)$; $\beta^{(k)}$ is the kth moment of the busy period; $\rho_l\equiv\lambda\beta^{(1)}/((1-r_l)(1+\lambda\beta^{(1)}))$; $\eta\equiv 1/(r_l\times(1+\lambda\beta^{(1)}))$.

Proof: If $\alpha^{(1)}\lambda r/(1+\alpha^{(1)}\lambda)+r_l<1$, for the high priority queue, the average load $\alpha^{(1)}\lambda r/(1+\alpha^{(1)}\lambda)$ is less than available service rate 1, therefore it is stable. Low priority queue is also stable since the input rate r_l is less than average available service rate $1-\alpha^{(1)}\lambda r/(1+\alpha^{(1)}\lambda)$.

(i) Buffer content of high priority flow is independent of low priority traffic. Let W denote the stationary buffer content of a $M/G/1$ queue with arrival rate λ and service time distribution with Laplace-Stieltjes transform $F((r-1)\theta)$. From Pollaczek-Khintchine formula, we know

$$E[e^{-\theta W}] = \frac{(1-(r-1)\lambda\alpha^{(1)})\theta}{\theta-\lambda+\lambda F((r-1)\theta)}$$

Then from Lemma 1,

$$E[e^{-\theta H}] = 1 - \gamma + \gamma E[e^{-\theta W}]$$

Based on equation (4), we have $E[H^k] = \gamma E[W^k]$. Then the first two moments of H can be derived directly from $E[W]$ and $E[W^2]$.

(ii) Let $Y(t)$ be the high priority buffer 's output process. $Y(t)$ is another ON-OFF process. Silent periods of $Y(t)$ are exponentially distributed with rate λ. From Lemma 2, we can calculate the Laplace-Stieltjes transform of $Y(t)$'s active period distribution as $B(\theta)$.

The sample path description of $L(t)$ based on $Y(t)$ is:

$$\frac{d}{dt}L(t) = \begin{cases} r_l & Y(t) = 1 \\ r_l - 1 & Y(t) = 0, L(t) > 0 \\ 0 & Y(t) = 0, L(t) = 0 \end{cases} \tag{21}$$

Let's consider a single source fluid queue with service rate $1 - r_l$. If the source is regulated by $Y(t)$ and source peak rate is 1, it has exactly the same sample path as (21). Then the stationary distribution of $L(t)$ can be solved as a single class fluid queue. According to Lemma 1, we can get $E[e^{-\theta L}]$, $E[L]$ and $E[L^2]$ by setting $h = 1$, $c = 1 - r_l$ and $F(\theta) = B(\theta)$. ■

The sample path description of the combined buffer content is:

$$\frac{d}{dt}(H(t) + L(t)) = \begin{cases} r + r_l - 1 & x(t) = 1 \\ r_l - 1 & x(t) = 0, H(t) + L(t) > 0 \\ 0 & x(t) = 0, H(t) + L(t) = 0 \end{cases}$$

It is exactly the sample path of the buffer content of a fluid queue with service rate $1 - r_l$ and fed by an ON-OFF source regulated by $x(t)$ with peak rate r. Another way to explain this is that the combined buffer content is independent of service discipline as long as server is work conservative. If we invert the service priority, the ON-OFF flow is served by residual service rate $(1 - r_l)$ left by the CBR flow. Denote by $\hat{H}(t)$ and $\hat{L}(t)$ the buffer content of the ON-OFF flow and the CBR flow respectively in the new priority queue. Now the CBR flow has service priority and its rate is less than its service rate. Therefore $\hat{L}(t) = 0$. Then we have

$$H(t) + L(t) = \hat{H}(t) + \hat{L}(t) = \hat{H}(t).$$

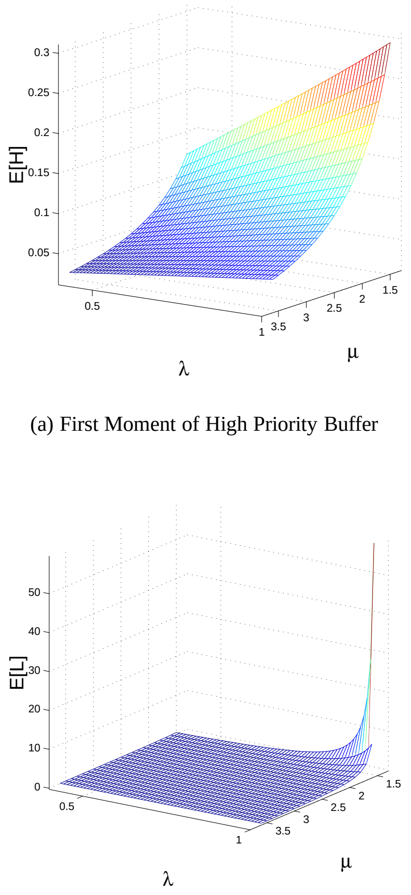
Using Lemma 1, we can obtain the distribution of $\hat{H}(t)$. Then the correlation between $H(t)$ and $L(t)$ can be derived as:

$$E[HL] = \frac{1}{2}(E[(H+L)^2] - E[H^2] - E[L^2]),$$

where $E[H^2]$ and $E[L^2]$ have been calculated in (17) and (20),

$$E[(H+L)^2] = \frac{\alpha^{(3)}(r + r_l - 1)^2\hat{\rho}}{3\alpha^{(1)}(1 + \lambda\alpha^{(1)})(1 - \hat{\rho})} + \frac{(\alpha^{(2)}\hat{\rho})^2(r + r_l - 1)^3}{2r(1 + \lambda\alpha^{(1)})(\alpha^{(1)}(1 - \hat{\rho}))^2},$$

where $\hat{\rho} \equiv \alpha\lambda r/((1 + \alpha\lambda)(1 - r_l))$.



(a) First Moment of High Priority Buffer

(b) First Moment of Low Priority Buffer

Fig. 2. Priority Queue with Single Markov ON-OFF High Priority Flow

D. Example

For the special case when the high priority flow is a Markov ON-OFF source with mean length of ON period $1/\mu$, the first moment of both high and low priority buffer can be calculated directly from equation (13)

$$E[H] = \frac{(r-1)r\lambda}{\lambda + \mu - r\lambda} \times \frac{1}{\lambda + \mu} \tag{22}$$

$$E[L] = \left(\frac{r + r_l - 1}{(\lambda + \mu)(1 - r_l) - r\lambda} - \frac{r - 1}{\lambda + \mu - r\lambda}\right) \times \frac{r\lambda}{\lambda + \mu} \tag{23}$$

In Figure 2, we plot the first moment of high and low priority buffer. The parameters are: $r = 1.5$, $r_l = 0.4$, $\lambda \in [0.4, 1]$ and $\mu \in [1.5, 3.6]$. When system utilization increases, the average queue length of both high and low priority buffer grow. The low priority buffer grows dramatically when system utilization approaches 1.

E. Multiple High Priority Flows and Single Low Priority Flow

When $M > 1$, let A_{i0} be the typical active period of high priority flow i, $F_i(\theta) = E[e^{-\theta A_{i0}}]$, $E[A_{i0}^k] = \alpha_i^{(k)}$,

$i = 1, 2, .., M$ and $k = 1, 2, ...$ The following Theorem solves the stationary distribution for low priority buffer.

Theorem 5: For low priority traffic,

$$E[e^{-\theta L}] = 1 - \zeta + \zeta \frac{(1 - r_l - r_l \lambda \pi^{(1)})\theta}{(1 - r_l)\theta - \lambda + \lambda \pi(r_l \theta)} \quad (24)$$

$$E[L] = \frac{\pi^{(2)} r_l \rho}{2\pi^{(1)}(1 + \lambda\pi^{(1)})(1 - \rho)} \quad (25)$$

$$E[L^2] = \frac{\pi^{(3)} r_l^2 \rho}{3\pi^{(1)}(1 + \lambda\pi^{(1)})(1 - \rho)} + \frac{(\pi^{(2)}\rho)^2 r_l^3}{2(1 + \lambda\pi^{(1)})(\pi^{(1)}(1 - \rho))^2} \quad (26)$$

where $\pi(\theta)$ is defined in Lemma 3, $\pi^{(k)}$ is the kth moment of $O(t)$'s active period. $\zeta \equiv 1/(r_l \times (1 + \lambda\pi^{(1)}))$ and $\rho \equiv \lambda\pi^{(1)}/((1 - r_l)(1 + \lambda\pi^{(1)}))$

Proof: As in the single high priority flow case, the high priority queue evolves independently of low priority traffic. It is equivalent to a single fluid queue with multiple ON-OFF sources. Unfortunately, to our knowledge, there is no analytical formula for the stationary distribution of its buffer content. However, according to Lemma 3, the output process $O(t)$ of high priority buffer can be characterized by an ON-OFF process. Following the same procedure as the proof of Theorem 4(ii), we can solve stationary distribution of low priority buffer. ∎

If A_{i0} are all exponentially distributed, we can explicitly solve $H(t)$'s stationary distribution by constructing the generator matrix of the underlying Markov chain which regulates the aggregate sending rate of ON-OFF sources. Based on that, we can get ordinary differential equations to obtain the stationary distribution of $H(t)$. We are not going to present the procedure here. Interested readers can check for details in [1], [6], [13] If we are only interested in the first few moments of the buffer content, we can use PDSDE technique introduced in Section IV-B to get the moments of $H(t)$ directly. Let $\{N_{i1}, N_{i2}\}$ be the pair of Poisson Counters which drive source i. Their rates are λ_i and $1/\alpha_i$ respectively. Then the sample path description of the system is :

$$\begin{cases} dx_i(t) = (1 - x_i(t))dN_{i1}(t) - x_i(t)dN_{i2}(t) \\ dH(t) = -I_{H(t)}dt + \sum_{i=1}^{M} r_i x_i(t)dt \end{cases} \quad (27)$$

It follows that

$$d[\frac{H(t)^2}{2}] = -H(t)dt + \sum_{i=1}^{M} r_i H(t)x_i(t)dt \quad (28)$$

$$d[H(t)x_i(t)] = H(t)(1 - x_i(t))dN_{i1}(t) - H(t)x_i(t)dN_{i2}(t) + x_i(t)\sum_{k=1}^{M} r_k x_k(t)dt - x_i(t)I_{H(t)}dt \quad (29)$$

By taking expectation on both sides of (28), (29) and let $t \to \infty$, we have:

$$E[H] = \sum_{i=1}^{M} r_i E[Hx_i] \quad (30)$$

$$\lambda_i E[H] = (\lambda_i + \frac{1}{\alpha_i})E[Hx_i] + (r_i - 1 + \sum_{k=1,k\neq i}^{M} r_k E[x_k])E[x_i] \quad (31)$$

There are $M+1$ equations for $M+1$ unknowns, we can solve it for $E[H]$ and $E[Hx_i]$ as

$$E[H] = \frac{1}{1 - \sum_{i=1}^{M}\nu_i}\sum_{i=1}^{M}\nu_i\tau_i(r_i - 1 + \sum_{k=1,k\neq i}^{M}\nu_k) \quad (32)$$

$$E[Hx_i] = \lambda_i\tau_i(E[H] + \tau_i(r_i - 1 + \sum_{k=1,k\neq i}^{M}\nu_k)), \quad (33)$$

where $\tau_i = \alpha_i/(1 + \alpha_i\lambda_i), \nu_i = r_i E x_i = r_i\lambda_i\tau_i$.

Similarly, from the S.D.E.s for $H(t)x_i(t)x_j(t)$, $H(t)^2 x_i(t)$, $H(t)^3$, we can get:

$$E[H^2] = \sum_{i=1}^{M} r_i E[H^2 x_i] \quad (34)$$

$$E[Hx_i x_j] \times (\lambda_i + \lambda_j + \frac{1}{\alpha_i} + \frac{1}{\alpha_j}) - \lambda_i E[Hx_j] = (r_i + r_j - 1 + \sum_{k=1,k\neq i,j}^{M}\nu_k)E[x_i]E[x_j] + \lambda_j E[Hx_i] \quad (35)$$

$$(\lambda_i + \frac{1}{\alpha_i})E[H^2 x_i] = 2(r_i - 1)E[Hx_i] + 2\sum_{j=1,j\neq i}^{M} E[Hx_i x_j] + \lambda_i E[H^2] \quad (36)$$

In (35), $E[Hx_ix_j]$ can be directly solved based on (33). Then, there are $M+1$ equations with $M+1$ unknowns in equations (34) and (36). We can get $E[H^2]$ and $E[H^2 x_i]$.
Following the similar procedure, all the moments of H can be obtained from sample paths of $H^n(t) \times \prod_{i=1}^{M} x^{m_i}(t)$, where $n = 1, 2, \cdots$ and $m_i \in \{0, 1\}$.

Use the same argument as in Section IV-C, the combined buffer content $H(t) + L(t)$ is equivalent to the buffer content $\hat{H}(t)$ of a single class fluid queue with service rate $(1 - r_l)$ and fed by multiple Markov ON-OFF flows $\{r_i, x_i(t)\}$. Then we have

$$E[\hat{H}] = \frac{1}{1 - r_l - \sum_{i=1}^{M} \nu_i} \sum_{i=1}^{M} \nu_i \tau_i (r_i - 1 + r_l + \sum_{k=1, k \neq i}^{M} \nu_k) \tag{37}$$

$$\begin{aligned} E[L] &= E[H+L] - E[H] = E[\hat{H}] - E[H] \\ &= \frac{1}{1 - r_l - \sum_{i=1}^{M} \nu_i} \sum_{i=1}^{M} \nu_i \tau_i (r_i - 1 + r_l + \sum_{k=1, k \neq i}^{M} \nu_k) \\ &\quad - \frac{1}{1 - \sum_{i=1}^{M} \nu_i} \sum_{i=1}^{M} \nu_i \tau_i (r_i - 1 + \sum_{k=1, k \neq i}^{M} \nu_k) \end{aligned} \tag{38}$$

From (32) and (38), we can see the high priority traffic autocorrelation constants $\{\tau_i\}$ have linear impact on the queue length of both high priority and low priority buffer. We verify our results for both high priority and low priority buffer on a priority fluid queue feeded by two Markov ON-OFF high priority flows and one CBR low priority flow. The capacity of the server is scaled to 1. The rate of the low priority flow, r_l, is set to be 0.25. The parameters for the first high priority flow are: $r_1 = 1.5$, $\lambda_1 = 0.5$, $\mu_1 = 3$. For the second high priority flow, we fix $r_2 = 1.5$, $\lambda_2 = 0.5$ and vary the average length of its ON periods, $\theta = 1/\mu_2$ from 0.4 to 0.8 with step size 0.01. For each θ, we run 100 simulations with different random sequences. Each simulation simulate the priority fluid queue for 2500 seconds. Averages are taken for the mean queue lengths of both the high priority buffer and the low priority buffer. The experimental results are compared with our analytical results in Figure 3.

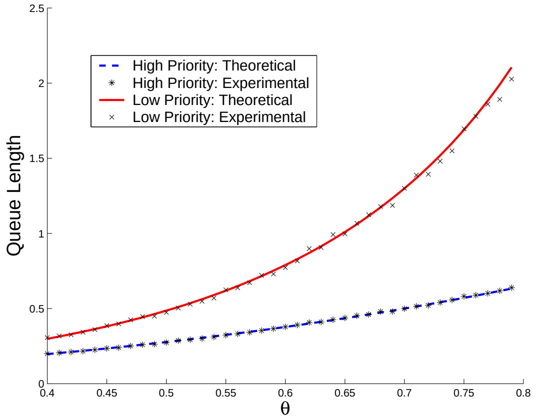


Fig. 3. Priority Fluid Queue with Two High Priority Flows and Single Priority Flow

Similar to (34) and (36) for $E[H^2]$, we can get a group of equations to obtain the second moment of $\hat{H}$. Together with $E[L^2]$ obtained in Theorem 5, the correlation between $H(t)$ and $L(t)$ can be solved as:

$$E[HL] = \frac{1}{2}(E[\hat{H}^2] - E[H^2] - E[L^2])$$

V. Tandem Priority Queueing System

In this section, we study tandem priority queueing system with strict priority. As shown in Figure 4, the system has V fluid servers in tandem. Service rate of server i is c_i. To make the problem nontrivial, we set $c_1 > \cdots > c_V$. We take the high and low priority traffic models in Section III as traffic model at the first node. We assume for high priority traffic $r_j > c_1$ and low priority traffic $r_l < c_V$. At node i, denote by $H_i(t)$ the high priority buffer content, $L_i(t)$ the low priority buffer, $O_i(t)$ the output process of high priority buffer and $Q_i(t)$ the output process of low priority buffer. A similar tandem fluid queueing system without priority has been studied in [20].

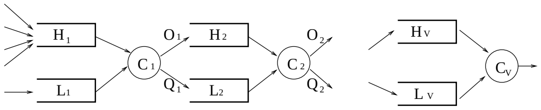


Fig. 4. Tandem Fluid Queueing System with Two Priorities

The sample path description for the first node is:

$$\frac{d}{dt} H_1(t) = \sum_{j=1}^{M} r_j x_j(t) dt - I_{H_1(t)} \tag{39}$$

$$\frac{d}{dt} L_1(t) = r_l I_{H_1(t)} + (r_l - 1)(1 - I_{H_1(t)}) I_{L_1(t)} \tag{40}$$

The solution for $H_1(t)$ and $L_1(t)$ has been studied in Section IV-C.

The sample path description for node i ($i \geq 2$) is:

$$\begin{cases} O_{i-1}(t) = c_{i-1} I_{H_{i-1}(t)} \\ Q_{i-1}(t) = (1 - I_{H_{i-1}(t)})(c_{i-1} I_{L_{i-1}(t)} + r_l(1 - I_{L_{i-1}(t)})) \\ \frac{d}{dt} H_i(t) = O_{i-1}(t) - c_i I_{H_i(t)} \\ \frac{d}{dt} L_i(t) = Q_{i-1}(t) I_{H_i(t)} - (c_i - Q_{i-1}(t)) I_{L_i(t)} (1 - I_{H_i(t)}) \end{cases} \tag{41}$$

For the ith high priority buffer, it is a single class fluid queue driven by a M/G type of ON-OFF process $O_{i-1}(t)$. Given the distribution of $O_{i-1}(t)$, by using Lemma 1 and Lemma 2, we can obtain the stationary distribution of $H_i(t)$ and $O_i(t)$. According to Lemma 3, we can solve for the distribution of $O_1(t)$. Therefore we can get the stationary distribution of $\{H_i(t)$,$O_i(t), 2 \leq i \leq V\}$ one by one. For low priority buffer $L_i(t)$, since its input process $Q_{i-1}(t)$ depends on its service process (the residual service left by high priority traffic is dependent with $H_{i-1}(t)$ thus $Q_{i-1}(t)$), its stationary distribution is difficult to get.

A better way to solve $\{H_i(t), L_i(t)\}$ is to explore the equivalence between the tandem priority fluid queueing system under study and a single node priority queue, which is established in the following theorem.

Theorem 6: For $2 \leq k \leq V$, $\sum_{i=1}^{k} H_i(t) \sim \hat{H}_k(t)$, $\sum_{i=1}^{k} L_i(t) \sim \hat{L}_k(t)$ and $O_k(t) \sim \hat{O}_k(t)$, where $\hat{H}_k(t)$ and $\hat{L}_k(t)$ are the buffer contents of a single node priority fluid queue with service rate c_k and fed by the same high and low priority traffic as the tandem system; $\hat{O}_k(t)$ is the output process of the single node priority queue's high priority buffer.

Proof: The sample path for the single node priority fluid queue is:

$$\frac{d}{dt}\hat{H}_k(t) = \sum_{j=1}^{M} r_j x_j(t)dt - c_k I_{\hat{H}_k(t)} \tag{42}$$

$$\frac{d}{dt}\hat{L}_k(t) = r_l I_{\hat{H}_k(t)} + (r_l - c_k)(1 - I_{\hat{H}_k(t)}) I_{\hat{L}_k(t)} \tag{43}$$

$$\hat{O}_k(t) = c_k I_{\hat{H}_k(t)} \tag{44}$$

For $\forall k \geq 2$,

$$\begin{aligned}\frac{d}{dt}\sum_{i=2}^{k} H_i(t) &= \sum_{i=2}^{k}(c_{i-1} I_{H_{i-1}(t)} - c_i I_{H_i(t)}) \\ &= c_1 I_{H_1(t)} - c_k I_{H_k(t)}\end{aligned}$$

Then we have

$$\frac{d}{dt}\sum_{i=1}^{k} H_i(t) = \sum_{j=1}^{M} r_j x_j(t)dt - c_k I_{H_k(t)} \tag{45}$$

Because $c_1 > \cdots > c_V$, it is easy to see $\sum_{i=1}^{k} H_i(t) > 0$ if and only if $H_k(t) > 0$, so $I_{\sum_{i=1}^{k} H_i(t)} = I_{H_k(t)}$. Equation (45) can be rewritten as

$$\frac{d}{dt}\sum_{i=1}^{k} H_i(t) = \sum_{j=1}^{M} r_j x_j(t)dt - c_k I_{\sum_{i=1}^{k} H_i(t)} \tag{46}$$

Comparing (46) with (42), we will have $\sum_{i=1}^{k} H_i(t) \sim \hat{H}_k(t)$ and $O_k(t) \sim \hat{O}_k(t)$.

For the low priority flow,

$$\frac{d}{dt}\sum_{i=1}^{k} L_i(t) = r_l - (1 - I_{H_k(t)})(c_k I_{L_k(t)} + r_l(1 - I_{L_k(t)})) \tag{47}$$

When $H_k(t) = 0$, $L_k(t) = 0$ if and only if $\sum_{i=1}^{k} L_i(t) = 0$. Thus (47) can be rewritten as:

$$\begin{aligned}\frac{d}{dt}\sum_{i=1}^{k} L_i(t) &= r_l I_{H_k(t)} + (r_l - c_k) I_{L_k(t)}(1 - I_{H_k(t)}) \\ &= r_l I_{\sum_{i=1}^{k} H_i(t)} + (r_l - c_k) I_{\sum_{i=1}^{k} L_i(t)}(1 - I_{\sum_{i=1}^{k} H_i(t)})\end{aligned} \tag{48}$$

Compared with (43), it is the same as the sample path of $\hat{L}_k(t)$. So we have $\sum_{i=1}^{k} L_i(t) \sim \hat{L}_k(t)$. ■

From this theorem, to solve the distribution of any high priority buffer $H_i(t), i \geq 2$, we first solve the $\hat{O}_{i-1}(t)$ according to Lemma 2. Since $O_{i-1}(t)$ is the only source of the ith high priority buffer and $O_{i-1}(t) \sim \hat{O}_{i-1}(t)$, the distribution of $H_i(t)$ is readily to get through Lemma 1. Also we can get the distribution of combined length of all the low priority buffers $\sum_{i=1}^{k} L_i$. From Theorem 5, the first moment of $\hat{L}_k(t), 1 \leq k \leq V$ can be solved. Then the first moment of $L_k(t), 2 \leq k \leq V$ is:

$$E[L_k] = E[\sum_{i=1}^{k} L_i] - E[\sum_{i=1}^{k-1} L_i] = E[\hat{L}_k] - E[\hat{L}_{k-1}] \tag{49}$$

Similarly, when the first moment of $\hat{H}_k(t), 1 \leq k \leq V$ can be solved, e.g., single high priority flow or multiple Markov ON-OFF high priority flows, the first moment of $H_k(t), 2 \leq k \leq V$ can be quickly solved as:

$$E[H_k] = E[\sum_{i=1}^{k} H_i] - E[\sum_{i=1}^{k-1} H_i] = E[\hat{H}_k] - E[\hat{H}_{k-1}] \tag{50}$$

A. Example

For the special case when high priority traffic are Markov ON-OFF flows, we can solve for both $\sum_{i=1}^{k} H_i$ and $\sum_{i=1}^{k} L_i$. Let A_{i0} be the typical active period of high priority flow i, and $E[A_{i0}] = \alpha_i$. Using results in Section IV-E, we immediately have

$$E[\sum_{i=1}^{k} H_i] = \frac{1}{c_k - \sum_{i=1}^{M} \nu_i} \sum_{i=1}^{M} \nu_i \tau_i (r_i - c_k + \sum_{k=1, k \neq i}^{M} \nu_k)$$

$$\begin{aligned}E[\sum_{i=1}^{k} L_i] = &- \frac{1}{c_k - \sum_{i=1}^{M} \nu_i} \sum_{i=1}^{M} \nu_i \tau_i (r_i - c_k + \sum_{k=1, k \neq i}^{M} \nu_k) + \\ &\frac{1}{c_k - r_l - \sum_{i=1}^{M} \nu_i} \sum_{i=1}^{M} \nu_i \tau_i (r_i - c_k + r_l + \sum_{k=1, k \neq i}^{M} \nu_k)\end{aligned}$$

Together with (50) and (49), $\{E[H_i], E[L_i], 1 \leq i \leq V\}$ can be solved.

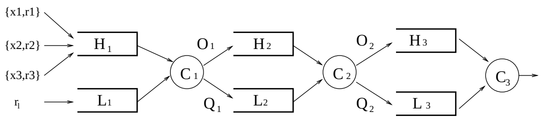
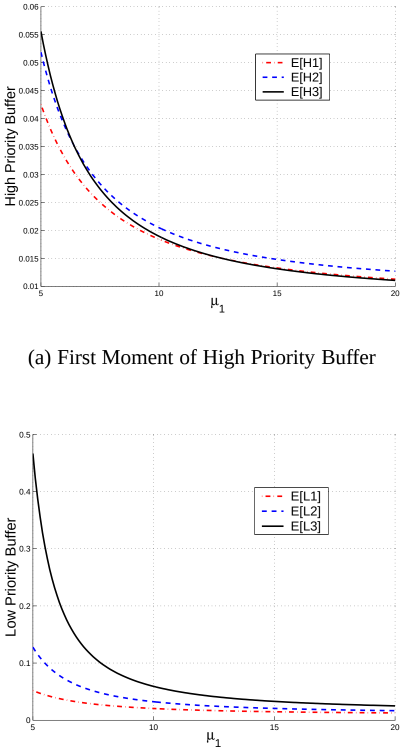


Fig. 5. Tandem Priority Queueing System with Three Nodes and Four Sources

For the system in Figure 5, we have 3 priority fluid queues in tandem. The service rate for queues are $c_1 = 1$, $c_2 = 0.8$ and $c_3 = 0.7$. The rate for low priority CBR flow is 0.2. There are 3 Markov ON-OFF high priority flows with peak rates $r_i = 1.2$, $1 \leq i \leq 3$. The jump up rates of those Markov ON-OFF sources are set to be $\lambda_1 = 2$, $\lambda_2 = 2.4$ and $\lambda_3 = 3$. We fix $\mu_2 = 20$ and $\mu_3 = 30$ and vary μ_1 from 5 to 20. Figure 6 plots the average queue length for high and low priority buffer at each node.

VI. Conclusions and Future Work

In this paper, we studied the fluid queueing system with two priorities. We use ON-OFF flows with exponential silent periods and general active periods to model high priority traffic; CBR flow for low priority traffic. Analytical results are obtained for different system configurations. In a single node system, when there is only one high priority flow or multiple flows with exponential active periods, marginal buffer distributions and moments are explicitly solved, correlations between two buffers are calculated; for multiple flows with arbitrary active periods, buffer content distribution of low priority traffic is solved. We also extend our results to tandem priority fluid queueing system. By relating it to equivalent



(a) First Moment of High Priority Buffer

(b) First Moment of Low Priority Buffer

Fig. 6. First Moments of Tandem Priority Queue

problems in single fluid queue, individual high priority buffer distributions and combined low priority buffer distribution are solved.

Sample path description tools are repeatedly used to obtain analytical results. By looking into the queue evolutions, we have established the equivalence between fluid queue and discrete queue, tandem queues and single node queue, prioritized queue and single class queue. As a new sample path tool, Poisson Driven Stochastic Differential Equation is shown in this paper to be very efficient in the study of fluid queueing system driven by Markov ON-OFF flows. It enables us to get moments of queues by solving a group of linear sample path equations. Compared with traditional methods, which involve solving probability density equations, this method is more direct and easier to get closed form solutions.

We have shown that the autocorrelation function of Markov ON-OFF sources has linear impact on congestion within fluid queue. We also show this impact can traverse tandem queues and priority classes. All these suggest that traffic autocorrelation is in a sense as important as its average rate, and not as a "second order factor" as commonly perceived. This could be regarded as a general rule of thumb in buffer dimensioning and other traffic engineering issues, even if flows can only be modelled as general ON-OFF processes.

References

[1] D. Anick, D. Mitra, and M.M. Sondhi, "Stochastic theory of a data-handling system with multiple sources," *The Bell System Technical Journal*, vol. 61, no. 8, pp. 1871–1894, 1982.

[2] O.J. Boxma and Dumas, "The busy period in the fluid queue," in *Proceedings of Sigmetrics/Performance'98*, 1998, pp. 100–110.

[3] V. Misra, W.B. Gong, and D. Towsley, "Fluid based analysis of a network of AQM routers supporting TCP flows with an application to RED," in *Proceedings of ACM/SIGCOMM*, 2000.

[4] S. Shakkottai and R. Srikant, "Mean FDE models for Internet congestion control under a many-flows regime," Tech. Rep., Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, 2001.

[5] S. Aalto, "Output of a multiplexer loaded by heterogeneous on-off sources," *Communication Stochastical Models*, vol. 14, pp. 993–1005, 1998.

[6] R.W. Brockett, W.B. Gong, and Y. Guo, "Stochastic analysis for fluid queueing systems," in *Proceedings of IEEE CDC'99*, 1999, pp. 3077–3082.

[7] T. Konstantopoulos and M. Zazanis, "Conservation laws and reflection mappings with an application to multi-class mean value analysis for stochastic fluid queues," *Stochastic Processes And Their Application*, vol. 65, no. 1, pp. 139–146, 1997.

[8] C.V. Hollot, V. Misra, D. Towsley, and W.B. Gong, "On designing improved controllers for AQM routers supporting TCP flows," in *Proceedings of IEEE/INFOCOM*, 2001.

[9] C.V. Hollot, V. Misra, D. Towsley, and W.B. Gong, "Analysis and design of controllers for AQM routers supporting TCP flows," *IEEE Transactions on Automatic Control*, vol. 47, no. 6, June 2002.

[10] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An architecture for Differentiated Services, RFC-2475," Tech. Rep., IETF, December 1998.

[11] D. Clark, S. Shenker, and L. Zhang, "Supporting real-time applications in an integrated services packet network: architecture and mechanism," in *Proceedings of ACM SIGCOMM'93*, 1992, pp. 14–26.

[12] J. Zhang, "Performance study of markov modulated fluid flow models with priority traffic," in *Proceedings of IEEE INFOCOM*, 1993, pp. 10–17.

[13] A.I. Elwalid and D. Mitra, "Analysis approximations and admission control of a multi-service multiplexing system with priorities," in *Proceedings of IEEE INFOCOM'95*, 1995, pp. 463–472.

[14] B.D. Choi and K.B. Choi, "A markov modulated fluid queueing system with strict priority," *Telecommunication Systems*, vol. 9, pp. 79–95, 1998.

[15] C. Knessl and C. Tier, "A simple fluid model for servicing priority traffic," Tech. Rep., Mathematics Department, University of Chicago, June 1999.

[16] Y. Liu and W.B. Gong, "On fluid queueing system with strict priority," in *Proceedings of 40th IEEE Conference on Decision and Control*, December 2001, pp. 1923–1928.

[17] R.G. Miller, "Continuous time stochastic storage processes with random linear inputs and outputs," *J. Math. Mech*, vol. 12, pp. 275–291, 1963.

[18] N.U. Prabhu, Ed., *Queues and inventories*, Wiley New York, 1965.

[19] O. Kella and W. Whitt, "A storage model with a two state random environment," *Operations Research*, vol. 40, pp. S257S262, 1992.

[20] S. Aalto and W. R. W. Scheinhardt, "Tandem fluid queues fed by homogeneous on-off sources," *Operations Research Letters*, vol. 27, pp. 73–82, 2000.

[21] T. Konstantopoulos and G. Last, "On the dynamics and performance of stochastic fluid systems," *Journal of Applied Probability*, vol. 37, no. 3, pp. 652–667, 2000.

[22] H. Kaspi and M. Rubinovitch, "The stochastic behavior of a buffer with non-identical input lines," *Stochastic Processes And Their Applications*, vol. 3, pp. 73–88, 1975.

[23] R.W. Brockett, "Lecture notes: Stochastic control," Tech. Rep., Harvard University, 1983.

PLACE
PHOTO
HERE

Yong Liu Yong Liu received his Ph.D. degree from University of Massachusetts, Amherst in 2002 and has been a postdoctoral research associate in the Computer Science Department of the same university since then. His research interests include modeling and control for communication networks and stochastic models of complex systems. He is a member of IEEE and ACM.

PLACE
PHOTO
HERE

Weibo Gong Weibo Gong received his Ph.D. degree from Harvard University in 1987 and has been with the Department of Electrical and Computer Engineering at the University of Massachusetts, Amherst since then. He is also an adjunct professor of the Department of Computer Science. His research interests include network modeling and control, large scale optimization, and architectural issues of complex engineering systems. He is a recipient of the IEEE Transactions on Automatic Control George Axelby Outstanding paper award.