Two-level Stochastic Fluid Tandem Queuing Model for Burst Impact Analysis

Yong Huang, Yong Liu, Weibo Gong, Don Towsley

Abstract— **Queuing analysis is important in providing guiding principles for packet network analysis. Stochastic fluid queueing models have been widely used as burst scale models for high speed communication networks. In this paper, we propose a novel two-level Markov On-Off source model to model the burstiness of a packet stream at different time scales. Analytical results are obtained to reveal the impact of traffic burstiness at two levels on the queue lengths in a tandem queue system. Our method combines the modeling power of the Poisson processes with that of stochastic differential equations to handle the complex interactions between the packet arrivals and the queue content. Our results for the tandem queuing network could be used to further justify the packet spacing scheme in helping deploying small buffer routers.**

I. Introduction

In the past decade, we have witnessed the dramatic growth of the Internet. On one hand, the speed of network links and routers keep increasing. On the other hand, new applications, such as Peer-to-Peer file sharing [1], [2], Voice-over-IP [3] and Video-over-IP [4], aggressively consume network resources and quickly push the Internet traffic towards its capacity. Congestion developed on network bottlenecks degrades user perceived quality of services, in terms of throughput, delay and losses. The study of network congestion is important for the management of operational networks and the design of future networks.

Queueing analysis has proven to be an efficient approach to evaluate the performance of communication networks under different traffic profiles. Classical queuing theory often requires renewal arrival assumption in order to obtain closed form results. However, traffic in modern packet networks is characterized by packet bursts. Traditional burst absorption methods rely on the use of large buffers at intermediate stations. Unfortunately optical packet switches can not afford large buffers. As a result we have to find other ways to handle the traffic bursts. For this and other reasons a quantitative analysis of the burst impact in network of queues become very critical. Towards this end, stochastic fluid queueing models have been widely used as burst scale models for high speed communication networks [5], [6]. In a fluid model, discrete packets and cells within bursts are modeled as continuous fluid. The continuous nature of fluid makes fluid models more tractable analytically. Many results have been obtained for various fluid queueing systems [6]–[9].

Recently, Markov On-Off processes have been applied to capture the correlation structure in network traffic [10]. Sample path analysis techniques, such as Poisson Counter Driven Stochastic Differential Equation [8] (PCSDE), are employed to study system queueing behavior in steady state [11], [12]. However, most previous studies assume the peak rate of a Markov On-Off source is larger than the server's capacity. This limited its modeling power. In this paper, we propose a novel two-level Markov On-Off source model to model the burstiness of a packet stream and derive analytical results that reveal the impact of the burstiness at each level. In our model, a source alternates between two states, namely "On" and "Off" states. In "On" state, it keeps sending data and at "Off" state, it sends nothing. Different from normal Markov On-Off source model, the "On" state is driven by another Markov On-Off process. Our model is motivated by the traffic pattern of TCP connections. In a regular TCP session, the sender tends to send out a burst of packets as allowed by its current congestion window. Then it keeps silent until the acknowledgments come back after one round-trip time. As studied in [13], TCP's window behaviors contribute to network traffic burstiness at multiple time scales. Our two-level Markov On-Off model essentially captures the traffic burstiness at two different time scales. By employing the PCSDE approach, we analytically study how traffic burstiness at two time scales affect the queue lengths in a tandem queue system.

Our method combines the modeling power of the Poisson processes with that of stochastic differential equations, in a way similar to the whitening filter in system theory, to handle the complex interactions between the packet arrivals and the queue content. Our analytical results demonstrate that when the average incoming rate is smaller than the server's capacity, the packet level burst plays a major role in determining the average queue length. Otherwise, the high level TCP burstiness dominates the dynamic of the average queue length. Thus, our proposed new model bridges the classical queueing theory, which focuses on the packet level burstiness, and the fluid queueing model, which is dedicated to deal with the correlation structure in network traffic. Our results for the tandem queuing network could be used to further justify the packet spacing scheme in helping deploying small buffer routers.

The paper is organized as follows. Section II presents the drawbacks of existing single layer Markov On-Off source model. Section III proposes our two levels Markov On-Off

Yong Huang and Weibo Gong are with the Department of Electrical and Computer Engineering, University of Massachusetts, Amherst, MA 01003. yhuang, gong@ecs.umass.edu.

Yong Liu is with the Department of Electrical and Computer Engineering, Polytechnic University, NY 11201. yongliu@poly.edu.

Don Towsley is with the Department of Computer Science, University of Massachusetts, Amherst, MA 01003. towsley@cs.umass.edu.

model in a tandem queue network. The network is modeled by a set of stochastic differential equations. A close form solution of average queue length is provided. Section IV presents simulation result to validate our major results. Our contributions and results are summarized in Section V.

II. Drawbacks of Single Layer Markov On-Off Sources

In this section, we first introduce the related work of a tandem queue network with a single layer Markov On-Off source. We present its stochastic fluid model and the corresponding analytical results. Finally we address the drawbacks of this simple Markov On-Off model.

A. Tandem Queue Networks with a Markov On-Off Source

Fluid model is used to analyze the impact of traffic burst on network buffers. With the help of Poisson Counter Driven Stochastic Differential Equations [8], the fluid model can quantitatively analyze how much a burst traffic can impact the network buffers.

A normal Markov On-Off source has two states. At “On” state, the source sends data at a fixed rate h. The duration of “On” state follows an exponential distribution with rate μ. At “Off” state, the source sends nothing and the duration of “Off” state follows an exponential distribution with rate λ. The behavior of $x(t)$ can be expressed by a PCSDE:

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

where N_1 and N_2 are two Poisson counters indicating number of times a source turning “On” and “Off”, respectively. By taking expectation on both sides, we get:

$$\frac{dE[x(t)]}{dt} = \lambda - (\lambda + \mu)E[x(t)]. \tag{2}$$

In steady state, there is

$$E[x] = \frac{\lambda}{\lambda + \mu}. \tag{3}$$

With PCSDE, it is easy to calculate the correlation function of $x(t)$ too. Consider

$$dx(\tau)x(0) = (1 - x(\tau))x(0)dN_1(\tau) - x(\tau)x(0)dN_2(\tau) \tag{4}$$

Taking the expectation on both sides of Equation (4) leads to:

$$\frac{d}{d\tau}E[x(0)x(\tau)] = -(\lambda + \mu)E[x(0)x(\tau)] + \lambda E[x(0)]. \tag{5}$$

Given the initial condition $E[x(0)x(0)] = E[x(0)] = \lambda/(\lambda+\mu)$, we can solve the correlation function of the source

$$R_{xx}(\tau) \equiv E[x(0)x(\tau)] = \frac{\lambda}{(\lambda + \mu)^2}(\mu e^{-(\lambda+\mu)\tau} + \lambda). \tag{6}$$

The correlation of Markov On-Off traffic decays exponentially with time constant $1/(\lambda + \mu)$. This time constant is often referred as autocorrelation time constant.

A capacity decreasing tandem queue network with a single Markov On-Off source is illustrated in Fig. 1. c_i and $v_i(t)$ are the capacity and queue length of router i, respectively. The capacities satisfy the conditions that $c_1 > c_2 > \cdots > c_n > 0$. The network dynamics can be presented by a set of stochastic differential equations:

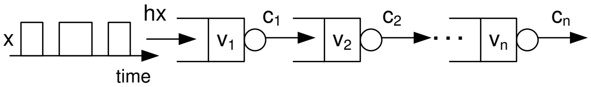


Fig. 1. A Tandem Queue Network with Single Markov On-Off Source.

$$\begin{cases} dv_1(t) & = -c_1\mathbf{I}(v_1)dt + hx(t)dt \\ dv_2(t) & = -c_2\mathbf{I}(v_2)dt + c_1\mathbf{I}(v_1)dt \\ & \vdots \\ dv_n(t) & = -c_n\mathbf{I}(v_n)dt + c_{n-1}\mathbf{I}(v_{n-1})dt. \end{cases} \tag{7}$$

where $\mathbf{I}(v_i)$ = 1 if $v_i > 0$ and equal to 0 if $v_i = 0$.

To make the problem solvable, the peak rate of On-Off source must be greater than c_1. To guarantee the system is stable, there is $hE[x] < c_n$. The average queue length is derived in [8] and we quote the results here:

$$E[v_1] = \Big(c_1 - hE[x]\Big)^{-1} \cdot \frac{h - c_1}{\lambda + \mu} \cdot hE[x], \tag{8}$$

$$E[v_n] = \frac{c_{n-1} - c_n}{c_n - hE[x]} \cdot (\frac{h}{\lambda + \mu}E[x] + E[v_{n-1}] + \cdots + E[v_1]). \tag{9}$$

An important observation is that, given $E[x]$, the average queue length at any stage linearly depends on $1/(\lambda + \mu)$, the autocorrelation time constant. The $\lambda + \mu$ illustrates how fast the correlation of $x(t)$ decays and presents the burstiness of $x(t)$. Equation (9) demonstrates the impact of source correlation on the buffer sizes of the entire network.

B. Drawbacks of Single Level Markov On-Off Sources

Equation (8) and (9) are derived with the assumption that $h > c_1$. In general, this is an unrealistic assumption. Particularly, when there are multiple On-Off sources, the peak rate of ***every*** source must be greater than c_1 as well.

Also the fluid model with a normal Markov On-Off source only considers the burst at a certain high abstraction level and cannot take care of the packet level burst at the same time.

To consider the packet arrival behavior and to relax the peak rate assumption, we propose a novel two levels Markov On-Off source in next section.

III. Two Levels Markov On-Off Source

Currently, there are two basic types of queueing analysis models, classical queueing theory approach and stochastic fluid model. Classical queueing theory looks at the packet level and often requires renewal arrival assumption. Stochastic fluid model views the packet arrival burst as continuous fluid. As demonstrated in previous section, Markov On-Off fluid model is able to capture the impact of source correlation on the average queue size. In this section, we propose a novel two levels Markov On-Off source to consider both the impact of low level (packet level) burst and the high level (fluid level) burst.

First we describe the two levels Markov On-Off source model and then develop the average queue length of a tandem queue network with this new source.

A. Network Model

The high level On-Off process is similar to the normal On-Off model, which presents the traffic burst behavior. The low level On-Off source presents the behavior of packet arrivals and departures. In classical queueing theory, every packet arrival consumes some resources and costs delay. So we assume the peak rate of low level On-Off source is always larger than the link capacity. The low level source is modulated by the fluid level source as well.

Fig. 2 presents the two levels Markov On-Off source. High level On-Off source is denoted by stochastic process $y(t)$, where $y(t)=1$ indicates source “On” and 0 otherwise. λ_y and μ_y are Markov On, Off rates, respectively. The low level On-Off source is denoted by another stochastic process $x(t)$. $x(t)$ takes values of 1 and 0 if source is “On” and “Off”, respectively. $x(t)$ and $y(t)$ are independent processes. The modulated source is in state “On” only when both $x(t)$ and $y(t)$ are 1.

The notations are summarized as following:

- $x(t)$: stochastic process presenting the packet level Markov On-Off source;
- $y(t)$: stochastic process presenting the fluid level Markov On-Off source;
- N_{1x}, N_{1y}: Poisson counters for “On” events of $x(t)$ and $y(t)$, respectively;
- N_{2x}, N_{2y}: Poisson counters for “Off” events of $x(t)$ and $y(t)$, respectively;
- λ_x, λ_y: “On” rates of $x(t)$ and $y(t)$, respectively;
- μ_x, μ_y: “Off” rates of $x(t)$ and $y(t)$. Since the high level and low level On-Off sources are at different timescales, λ_y and μ_y are much smaller than λ_x, μ_x and $(\lambda_y+\mu_y) << (\lambda_x+\mu_x)$;
- h_x: peak rate of the source;
- c_i: capacity of router i. To make the system stable, the server’s capacity should be always larger than the average incoming rate, $c_i > h_x E[x]E[y]$;
- $v_i(t)$: queue size of router i.

Then the system can be modeled by following differential equations:

$$\begin{cases} dx(t) &= (1-x(t))dN_{1x} - x(t)dN_{2x} \\ dy(t) &= (1-y(t))dN_{1y} - y(t)dN_{2y} \\ dv_1(t) &= -c_1\mathbf{I}(v_1)dt + h_x x(t)y(t)dt \\ dv_2(t) &= -c_2\mathbf{I}(v_2)dt + c_1\mathbf{I}(v_1)dt \\ & \vdots \\ dv_n(t) &= -c_n\mathbf{I}(v_n)dt + c_{n-1}\mathbf{I}(v_{n-1})dt. \end{cases} \tag{10}$$

Since steady state average queue size is important for buffer design and network operations, we are interested in solving $E[v_i]$ at any stage of tandem queues. We start with solving the average queue length of the first queue. Then we extend our analysis to $E[v_n]$.

B. Average Queue Length $E[v_1]$

The solutions of $E[x]$ and $E[y]$ in steady state are:

$$E[x(t)] = \frac{\lambda_x}{\lambda_x+\mu_x}, \text{ and } E[y(t)] = \frac{\lambda_y}{\lambda_y+\mu_y}. \tag{11}$$

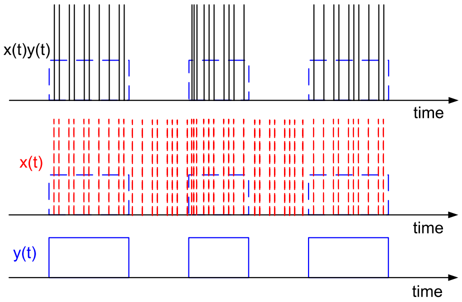


Fig. 2. Two Levels Markov On-Off Source. $y(t)$ presents the fluid level On-Off source. $x(t)$ presents the packet level On-Off source. $x(t)y(t)$ denotes the new On-Off source.

Since $x(t)$ and $y(t)$ are independent, we have

$$E[xy] = E[x]E[y]. \tag{12}$$

From (10), we have

$$\frac{dE[v_1^2]}{dt} = -2c_1E[v_1] + 2h_xE[xyv_1], \tag{13}$$

$$\begin{aligned}\frac{dE[xyv_1]}{dt} =& (h_x-c_1)E[x]E[y] + \lambda_yE[xv_1] + \lambda_xE[yv_1] \\ & - E[xyv_1](\lambda_x+\mu_x+\lambda_y+\mu_y).\end{aligned} \tag{14}$$

In steady state, Equation (13) and (14) lead to

$$\begin{aligned} c_1E[v_1] &= h_xE[xyv_1] \\ &= h_x\frac{(h_x-c_1)E[x]E[y]+\lambda_yE[xv_1]+\lambda_xE[yv_1]}{\lambda_x+\mu_x+\lambda_y+\mu_y}.\end{aligned} \tag{15}$$

To solve $E[v_1]$ we have to get both $E[xv_1]$ and $E[yv_1]$.

1) Solve $E[xv_1]$*:* From (10), we have

$$dxv_1 = -c_1x\mathbf{I}(v_1)dt + h_xxydt + v_1(1-x)dN_{1x} - xv_1dN_{2x}. \tag{16}$$

By taking expectation on both sides of dxv_1, we have $E[xv_1]$ in steady state:

$$E[xv_1] = \frac{\lambda_xE[v_1] - c_1E[x\mathbf{I}(v_1)] + h_xE[x]E[y]}{\lambda_x+\mu_x} \tag{17}$$

where $E[x\mathbf{I}(v_1)]$ is:

$$\begin{aligned} E[x\mathbf{I}(v_1)] &= \Pr[x=1, v_1>0] \\ &= \Pr[x=1,y=1,v_1>0] + \Pr[x=1,y=0,v_1>0] \\ &= \Pr[x=1,y=1] + \Pr[x=1,y=0,v_1>0] \\ &= E[x]E[y] + \Pr[x=1,y=0,v_1>0]. \end{aligned} \tag{18}$$

Define $P_1 = \Pr[x=1,y=0,v_1>0]$, Equation (17) can be rewritten as

$$E[xv_1] = E[x]E[v_1] + \frac{(h_x-c_1)E[x]E[y]-c_1P_1}{\lambda_x+\mu_x}. \tag{19}$$

2) Solve $E[yv_1]$*:* By symmetry of $x(t)$ and $y(t)$, we have

$$dyv_1 = -c_1y\mathbf{I}(v_1)dt + h_xxydt + v_1(1-y)dN_{1y} - yv_1dN_{2y}, \tag{20}$$

and in steady state, we have

$$\begin{aligned}
E[yv_1] &= E[y]E[v_1] + \frac{h_x E[x]E[y] - c_1 E[y\mathbf{I}(v_1)]}{\lambda_y + \mu_y} \\
&= E[y]E[v_1] + \frac{h_x E[x]E[y] - c_1 \Pr[v_1 > 0, y = 1]}{\lambda_y + \mu_y} \\
&= E[y]E[v_1] + \frac{h_x E[x]E[y] - c_1 \Pr[v_1 > 0|y = 1]E[y]}{\lambda_y + \mu_y}. \quad (21)
\end{aligned}$$

Now we need to solve $\Pr[v_1 > 0|y = 1]$. When $y(t) = 1$, if the average incoming rate $h_x E[x] \geq c_1$, the average queue keeps increasing and the queue length is non zero during almost all the time. In this scenario, $\Pr[v_1 > 0|y = 1]$ is close to 1.

If $h_x E[x] < c_1$, taking conditional expectation over $y(t) = 1$ on Equation (20) leads to

$$\begin{aligned}
\frac{dE[v_1|y = 1]}{dt} =& - c_1 E[\mathbf{I}(v_1)|y = 1] + h_x E[x|y = 1] - \\
& E[v_1|y = 1]\mu_y \qquad (22)
\end{aligned}$$

where $E[\mathbf{I}(v_1)|y = 1] = \Pr[v_1 > 0|y = 1]$, $E[v_1, y = 1] = E[v_1|y = 1]\Pr[y = 1] = E[v_1|y = 1]E[y]$. Also since $x(t)$ and $y(t)$ are independent, we have

$$E[x|y = 1] = E[x]. \qquad (23)$$

Therefore, after solving Equation (22) in steady state, we have

$$\Pr[v_1 > 0|y = 1] = \frac{h_x E[x]}{c_1} - \frac{\mu_y E[yv_1]}{c_1 E[y]}. \qquad (24)$$

Finally, by solving Equation (21) and (24) and also considering $\Pr[v_1 > 0|y = 1] \approx 1$ when $h_x E[x] < c_1$, we have

$$E[yv_1] \approx \begin{cases} E[v_1], & \text{if } h_x E[x] < c_1, \\ E[y]\left(\frac{h_x E[x] - c_1}{\lambda_y + \mu_y} + E[v_1]\right), & \text{otherwise.} \end{cases} \qquad (25)$$

3) Solve $E[v_1]$: After substituting Equation (19) and (25) into (15), we can solve $E[v_1]$.

- If $h_x E[x] \geq c_1$: Consider that λ_x and μ_x are much larger than λ_y and μ_y, $E[v_1]$ can be approximated as

$$E[v_1] \approx \frac{h_x E[x]E[y]}{c_1 - h_x E[x]E[y]} \cdot \left(\frac{h_x - c_1}{\lambda_x + \mu_x} + \frac{h_x E[x] - c_1}{\lambda_y + \mu_y}\right) \qquad (26)$$

  where $1/(\lambda_x + \mu_x)$ and $1/(\lambda_y + \mu_y)$ are autocorrelation time constants of $x(t)$ and $y(t)$, respectively. Therefore, Equation (26) shows that $E[v_1]$ depends on the burstiness of both $x(t)$ and $y(t)$.
- If $h_x E[x] < c_1$:

$$E[v_1] \approx \frac{h_x E[x]E[y]}{c_1 - h_x E[x]} \cdot \left(\frac{h_x - c_1}{\lambda_x + \mu_x}\right). \qquad (27)$$

  In this scenario $E[v_1]$ only depends on the burstiness of $x(t)$. For fixed $E[x]$ and $E[y]$, $E[v_1]$ linearly changes with $\frac{1}{\lambda_x + \mu_x}$.

Special cases When $E[x] = 1$, since $h_x > c_1$, the result fits into the case that $h_x E[x] > c_1$.

$$E[v_1] \approx \frac{h_x E[y]}{c_1 - h_x E[y]}\left(\frac{h_x - c_1}{\lambda_y + \mu_y}\right), \qquad (28)$$

which is exactly the result when there is only high level On-Off source.

When $E[y] = 1$, to make the system stable we require that $h_x E[x] < c_1$. Then according to Equation (27),

$$E[v_1] \approx \frac{h_x E[x]}{c_1 - h_x E[x]}\left(\frac{h_x - c_1}{\lambda_x + \mu_x}\right) \qquad (29)$$

which matches the result when there is only one level Markov On-Off source of $x(t)$.

In summary, Equation(26) and (27) show that the average queue length $E[v_1]$ depends on burst of $x(t)$ and $y(t)$. The importance of their impacts varies according to the difference between $h_x E[x]$ and c_1.

If $h_x E[x] \geq c_1$, $E[v_1]$ depends on the burstiness of both $x(t)$ and $y(t)$. Because the average incoming rate is larger than the capacity, queue length is not zero when $y(t) = 0$. In this scenario, the off period of $y(t)$ plays an essential role in dequeuing. Since the timescale of $y(t)$ is much larger than that of $x(t)$, $E[v_1]$ is mainly dominated by the burstiness of $y(t)$ in this scenario.

If $h_x E[x] < c_1$, $E[v_1]$ is only dominated by the burstiness of $x(t)$. When $y(t)$ is on, the average input traffic is less than the system capacity. Since the timescale of $y(t)$ is much larger than that of $x(t)$, the queue can be absorbed mostly when $y(t) = 1$. In other words, when $y(t) = 0$, the queue is empty most of time. Therefore, the burstiness of $y(t)$ affects $E[v_1]$ very little.

C. Average Queue Length $E[v_n]$

In the previous part, we derived the $E[v_1]$. Now we can continue to calculate the $E[v_n]$, where $n > 1$.

First we solve $E[v_2]$. An easy way to do it is to view the first two queues as a single queue. Now, the new queue has input $h_x xy$ and capacity c_2. We define the $E[v_2']$ as the average queue length of this new queue. From the analysis of previous section, we know

$$E[v_2'] \approx \begin{cases} \text{If } h_x E[x] \geq c_2\text{:} \frac{h_x E[x]E[y]}{c_2 - h_x E[x]E[y]}\left(\frac{h_x - c_2}{\lambda_x + \mu_x} + \frac{h_x E[x] - c_2}{\lambda_y + \mu_y}\right), \\ \text{If } h_x E[x] < c_2\text{: } \frac{h_x E[x]E[y]}{c_2 - h_x E[x]} \cdot \frac{h_x - c_2}{\lambda_x + \mu_x}. \end{cases} \qquad (30)$$

Since the packets in the new queue are actually either in the first queue or in the second queue, we have $E[v_2'] = E[v_1] + E[v_2]$. Thus, $E[v_2]$ is

$$E[v_2] \approx \begin{cases} \text{If } h_x E[x] < c_2\text{:} \frac{h_x E[x]E[y]h_x(1 - E[x])(c_1 - c_2)}{(\lambda_x + \mu_x)(c_2 - h_x E[x])(c_1 - h_x E[x])}, \\ \text{If } c_2 \leq h_x E[x] < c_1\text{:} \frac{h_x E[x]E[y]}{c_2 - h_x E[x]E[y]} \cdot \\ \left(\frac{h_x - c_2}{\lambda_x + \mu_x} + \frac{h_x E[x] - c_2}{\lambda_y + \mu_y}\right) - \frac{h_x E[x]E[y](h_x - c_1)}{(c_1 - h_x E[x])(\lambda_x + \mu_x)}, \\ \text{If } c_1 \leq h_x E[x]\text{:} \frac{h_x E[x]E[y]h_x(c_1 - c_2)}{(c_2 - h_x E[x]E[y])(c_1 - h_x E[x]E[y])} \cdot \\ \left(\frac{1 - E[x]E[y]}{\lambda_x + \mu_x} + \frac{E[x](1 - E[y])}{\lambda_y + \mu_y}\right). \end{cases} \qquad (31)$$

More generally, we have

$$E[v_n] = E[v_n'] - E[v_{n-1}']. \qquad (32)$$

So, for $n > 1$, the steady state average queue length is

$$E[v_n] \approx \begin{cases} \text{If } h_x E[x] < c_n\text{:} \frac{h_x E[x] E[y] h_x (1-E[x])(c_{n-1}-c_n)}{(\lambda_x+\mu_x)(c_n-h_x E[x])(c_{n-1}-h_x E[x])}, \\ \\ \text{If } c_n \le h_x E[x] < c_{n-1}\text{:} \frac{h_x E[x] E[y]}{c_n - h_x E[x] E[y]} \cdot \\ \left(\frac{h_x - c_n}{\lambda_x+\mu_x} + \frac{h_x E[x] - c_n}{\lambda_y+\mu_y}\right) - \frac{h_x E[x] E[y] (h_x - c_{n-1})}{(c_{n-1}-h_x E[x])(\lambda_x+\mu_x)}, \\ \\ \text{If } c_{n-1} \le h_x E[x]\text{:} \frac{h_x E[x] E[y] h_x (c_{n-1}-c_n)}{(c_n - h_x E[x] E[y])(c_{n-1}-h_x E[x] E[y])} \cdot \\ \left(\frac{1-E[x]E[y]}{\lambda_x+\mu_x} + \frac{E[x](1-E[y])}{\lambda_y+\mu_y}\right). \end{cases} \tag{33}$$

Formula (33) summarizes our major result of average queue length of a tandem queue network. Depending on different traffic conditions, the average queue length at every stage in a tandem queue network may be dominated by the burstiness of either level or both.

When the network is light loaded (in the case of $h_x E[x] < c_n$), the average queue length only depends on the packet level burst. For fixed $E[x]$ and $E[y]$, queue length at every stage of the network linearly changes with $\frac{1}{\lambda_x+\mu_x}$.

When the network is heavy loaded (in the case of $h_x E[x] \geq c_1$), the high level burstiness plays a major role.

When the network traffic condition is in the middle of them, the average queue length depends on the burstiness of both levels.

In either case, for fixed $E[x]$ and $E[y]$, when we simultaneously increase both the $\lambda_x + \mu_x$ and $\lambda_y + \mu_y$ linearly, the average queue length at every stage of the tandem queue network will linearly decrease. Therefore, packet spacing at both levels helps the entire tandem queues performance gracefully.

IV. Simulation Validation

In this section, we use simulation to validate our main results Equation (26), (27) and (33).

A. Validation of $E[v_1]$

Fig. 3 shows the simulation results when $h_x E[x] \geq c_1$. Fig. 3(a) and Fig. 3(b) depict $E[v_1]$ with different $E[x]$'s and $E[y]$'s. They demonstrate that the $E[v_1]$ linearly changes with $1/(\lambda_y + \mu_y)$, which is consistent with our model (26). Fig. 3(c) presents the relationship between $E[v_1]$ and $1/(\lambda_x+\mu_x)$. It validates that $1/(\lambda_x+\mu_x)$ has little impact on $E[v_1]$ in this scenario.

Fig. 4 validates the result when $h_x E[x] < c_1$. Fig. 4(a) and 4(b) show the $E[v_1]$ with different $E[x]$'s and h_x's. They validate that $E[v_1]$ linearly increases with $1/(\lambda_x+\mu_x)$ in this scenario. However, when $\lambda_x+\mu_x$ becomes small, the model over estimates the impact of $1/(\lambda_x+\mu_x)$. The reason is that as $\lambda_x+\mu_x$ decreases, the assumption that $\lambda_x, \mu_x >> \lambda_y, \mu_y$ is getting weak. Fig. 4(c) illustrates that the burstiness of $y(t)$ doesn't affect $E[v_1]$ in this scenario.

B. Validation of $E[v_n]$

Fig. 5 validates Formula (33) when $h_x E[x] > c_1$. Fig. 5(a), 5(b) and 5(c) illustrate that $E[v_1]$, $E[v_2]$ and $E[v_3]$ are all linearly changing with $1/(\lambda_y + \mu_y)$. In this scenario, the average incoming rate is bigger than capacities of all queues. Thus, the burstiness of $y(t)$ dominates the average queue behavior.

Fig. 6 illustrates $E[v_n]$ when $c_2 < h_x E[x] < c_1$. Fig. 6(a) shows that when $h_x E[x] < c_1$, $E[v_1]$ increases as $\frac{1}{(\lambda_x+\mu_x)}$ increases. The simulation result is close to the model output. However, Fig. 6(b) shows big errors on $E[v_2]$. The reason is that model output of $E[v_1]$ over estimates the impact of $\lambda_x + \mu_x$ as shown in the right end of Fig. 4(a) and 6(a). After removing the impact of $E[v_1]$, we get a close match to the simulation result. This indicates that when the average incoming rate $h_x E[x]$ is between c_j and c_{j-1}, we should carefully evaluate the impact of $E[v_{j-1}]$ when $\lambda_x+\mu_x$ is small. Fig. 6(c) shows that $E[v_3]$ doesn't change with $1/(\lambda_x+\mu_x)$. This is consistent with our model because $E[v_3]$ is dominated by burst of $y(t)$ only when $h_x E[x] > c_2 > c_3$.

V. Conclusion

In this paper, we propose a novel two-level Markov On-Off source model to model the burstiness of a packet stream at different time scales. Analytical results are obtained to reveal the impact of traffic burstiness at two levels on the queue lengths in a tandem queue system. Both our analytical and simulation results demonstrate that the burst of both levels have linear impact on the average queue size throughout the entire tandem queue network. Depending on the traffic load conditions, the importance of the two level burst may vary. Our proposed new model bridges the classical queueing theory, which focuses on the packet level burstiness, and the fluid queueing model, which is dedicated to deal with the correlation structure in network traffic. Our results for the tandem queuing network could be used to further justify the packet spacing scheme in helping deploying small buffer routers.

Acknowledgments

This work is supported in part by the National Science Foundation grants CNS-0524323, CNS-0519922, CNS-0312563 and CNS-0519998. We also appreciate the helpful discussions with Yan Cai and Jie Sun during the initial stage of this project.

References

[1] Bittorrent. BitTorrent web site. http://www.bittorrent.com/.

[2] Kazaa. Kazaahomepage. http://www.kazaa.com.

[3] Skype. Skype homepage. http://www.skype.com/.

[4] Youtube. Youtube homepage. http://www.youtube.com.

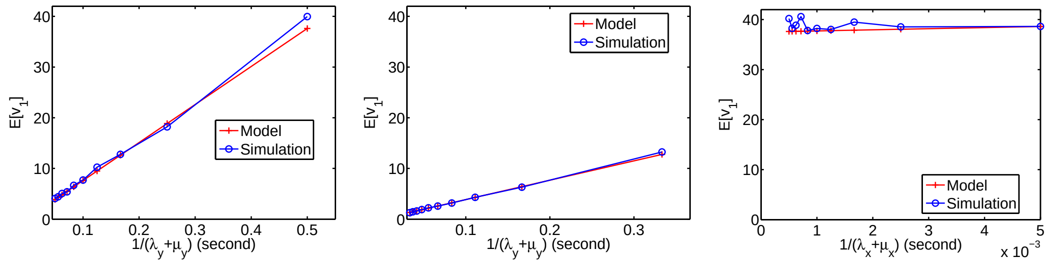
[5] D. Anick, D. Mitra, and M.M. Sondhi. Stochastic theory of a data-handling system with multiple sources. *The Bell System Technical Journal*, 61(8):1871–1894, 1982.

[6] O.J. Boxma and Dumas. The busy period in the fluid queue. In *SIGMETRICS '98/PERFORMANCE '98: Proceedings of the 1998 ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems*, pages 100–110, 1998.

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

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

[9] 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*, 65(1):139–146, 1997.

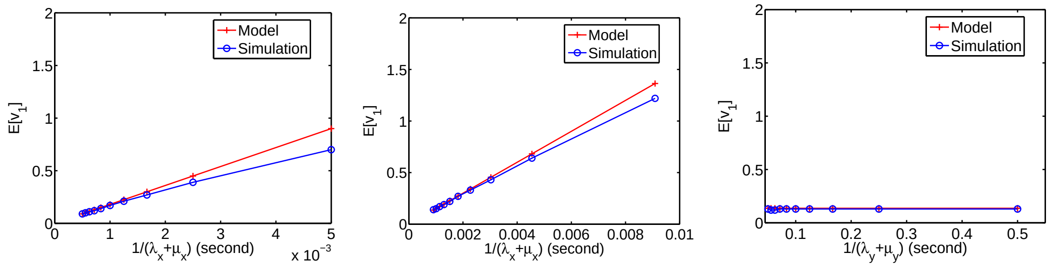


(a) $c_1 : 50,\ h_x : 150,\ \lambda_x : 1000,\ \mu_x : 1000,\ \lambda_y : 1-10,\ \mu_y : 1-10$

(b) $c_1 : 50,\ h_x : 150,\ \lambda_x : 1200,\ \mu_x : 1000,\ \lambda_y : 1-10,\ \mu_y : 2-20$

(c) $c_1 : 50,\ h_x : 150,\ \lambda_x : 100-1000,\ \mu_x : 100-1000,\ \lambda_y : 1,\ \mu_y : 1$

Fig. 3. Validation of $E[v_1]$ When $h_xE[x] \geq c_1$: $E[v_1]$ linearly increases with $\frac{1}{\lambda_y+\mu_y}$.

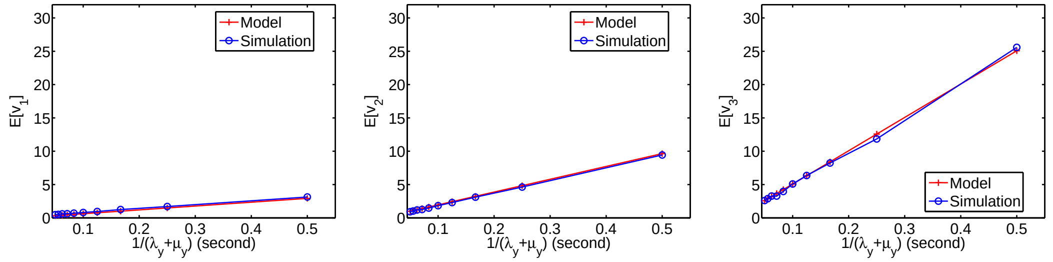


(a) $c_1 : 50,\ h_x : 90,\ \lambda_x : 100-1000,\ \mu_x : 100-1000, \lambda_y : 1,\ \mu_y : 1$

(b) $c_1 : 50,\ h_x : 300,\ \lambda_x : 10-100,\ \mu_x : 100-1000,\ \lambda_y : 1,\ \mu_y : 1$

(c) $c_1 : 50,\ h_x : 300,\ \lambda_x : 100,\ \mu_x : 1000,\ \lambda_y : 1-10,\ \mu_y : 1-10$

Fig. 4. Validation of $E[v_1]$ When $h_xE[x] < c_1$: $E[v_1]$ linearly increases with $\frac{1}{\lambda_x+\mu_x}$.

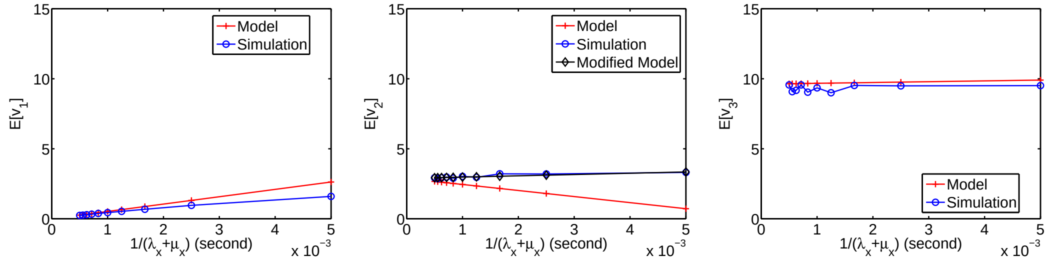


(a) Average Queue Length of v_1

(b) Average Queue Length of v_2

(c) Average Queue Length of v_3

Fig. 5. Validation of $E[v_n]$ When $h_xE[x] > c_1$: $E[v_n]$ at any stage linearly increases with $\frac{1}{\lambda_y+\mu_y}$. $h_x = 150$, $c_1 = 70$, $c_2 = 60$, $c_3 = 50$. $\lambda_x = 1000, \mu_x = 1000,\ \lambda_y : 1-10, \mu_y : 1-10$.



(a) Average Queue Length of v_1

(b) Average Queue Length of v_2

(c) Average Queue Length of v_3

Fig. 6. Validation of $E[v_n]$ When $c_2 < h_xE[x] < c_1$. $E[v_1]$ changes with $\frac{1}{\lambda_x+\mu_x}$, but $E[v_3]$ doesn't. $h_x = 150$, $c_1 = 80$, $c_2 = 70$, $c_3 = 60$. $\lambda_x : 100-1000, \mu_x : 100-1000,\ \lambda_y : 1, \mu_y : 1$.

[10] V. Misra and W. Gong. A hierarchical model for teletraffic. In *Proceedings of the 37th Annual IEEE CDC*, pages 1674–1679, 1998.

[11] Y. Liu and W. Gong. On fluid queueing system with strict priority. *IEEE Transactions on Automatic Control*, 12, 2003.

[12] Y. Wu and W. Gong. Error analysis of burst level modeling of active-idle sources. *ACM Trans. Model. Comput. Simul.*, 14(3):278–304, 2004.

[13] D. R. Figueiredo, B. Liu, V. Misra, and D. Towsley. On the autocorrelation structure of tcp traffic. *Comput. Networks*, 40(3):339–361, 2002.