

RICE UNIVERSITY

Effect of the Traffic Bursts in the Network Queue

by

Alireza KeshavarzHaddad

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Master of Science

APPROVED, THESIS COMMITTEE:

Dr. Rudolf H. Riedi
Faculty Fellow
Electrical and Computer Engineering

Dr. Richard G. Baraniuk
Professor
Electrical and Computer Engineering

Dr. Edward W. Knightly
Associate Professor
Electrical and Computer Engineering

Houston, Texas

April, 2003

ABSTRACT

This thesis studies the effect of the traffic bursts in the queue. Knowledge of the queueing behavior provides opportunity for additional control and improved performance. Most existing work on queueing today is based on Long-Range-Dependence (LRD) and Self-similarity, two well-known properties of network traffic at large scales. However, network traffic shows bursty behavior on small scales which are not captured by traditional self-similar models. We leverage a decomposition of traffic into two components. The alpha component is the bursty part of the traffic consisting of only few high bandwidth connections. The beta component collects the residual traffic and is a Gaussian LRD process. The alpha component is highly non-Gaussian and bursty. We propose two models for the alpha component, a heavy-tailed self-similar process and a high rate ON/OFF source. Our results explain how size and type of bursts affect the queueing behavior.

Acknowledgments

I thank my advisor, Dr. Rudolf H. Riedi for his support, encouragement and insightful ideas. This thesis would not have been possible without his help. My heartfelt thanks to my other thesis committee members Dr. Edward Knightly and Dr. Richard Baraniuk for consenting to be on my committee and for their helpful suggestions during the course of my research. I sincerely thank my friends and colleagues for their enjoyable company. Finally, I thank my family for their support and encouragement.

Contents

Acknowledgments	iii
List of Illustrations	vi
1 Introduction	1
2 Background	8
2.1 Classical model and self-similarity	8
2.1.1 $M/M/1$ queue	8
2.1.2 self-similarity	8
2.2 ON/OFF Model	9
2.3 Alpha-Beta decomposition	12
2.4 Exploiting self-similarity in queueing analysis	15
2.5 Multiplexing of an ON/OFF source and a background traffic	18
3 A self-similar burst model	21
3.1 Queueing analysis of self-similar burst model	23
3.2 De-multiplexing of two general flows	25
3.2.1 Lower bounds	26
3.2.2 Upper bounds	27
3.2.3 De-multiplexing parameters	29
3.3 De-multiplexing for the self-similar burst model	32
3.3.1 Lower bound	32

3.3.2	Upper bound	33
3.4	Summary of the self-similar burst model	36
4	An ON/OFF burst model	37
4.1	Markov service rate queue model	39
4.1.1	Markov service rate queue model with CBR input	40
4.1.2	Lower bound for the buffer overflow probability with CBR input	49
4.1.3	Markov service rate queue with non-constant rate input	52
4.2	Renewal service rate queue model	54
4.3	Queuing analysis of ON/OFF burst model: Particular case	60
4.4	Queueing analysis of ON/OFF burst model in the variable service queue framework	61
4.5	Summary of the ON/OFF burst model	64
5	Conclusion	65
	Bibliography	68

Illustrations

1.1	De-Multiplexing of two flows	4
1.2	A Queue with variable service rate	5
3.1	self-similar decomposition of traffic: the bursty component is modelled as sLn, the residual component by fGn	22
3.2	The queue behavior of self-similar burst model has two regimes: It has a Weibull decay like fGn traffic for small buffers, and it has power-law decay like sLn traffic for large buffers	24
3.3	Marginal density of the self-similar bursty model $Z_H(1) = \sigma_1 B_H(1) + \sigma_2 L_H(1)$	26
4.1	ON/OFF burst model in the large scale limit	38
4.2	Discretized queue and CBR traffic	41
4.3	Discretized queue and CBR traffic	46
4.4	Comparison of CBR and Poisson traffic in a queue with Markov service rate. For small variance of the Poisson traffic the two queues behave almost the same	50

Chapter 1

Introduction

Predicting queueing behavior of the Internet traffic and understanding the causes of its complex dynamics is of chief importance for network management and protocol design. The true value of the discovery of long-range dependence (LRD) in network flow dynamics, e.g., came with the identification of the source of this phenomenon — the client behavior — along with the assessment of its impact on network performance. While the concept of LRD allows to address the behavior of queues at large time scales, ideally infinitely large, and for large buffer sizes, the concept of the critical time scale (CTS) was born out of the need to handle more realistic time frames and queue sizes [BCRH⁺00].

More recent analysis of real Internet traffic traces [WSRB02, SRB02] has shown that the traffic bursts appearing at time scales of the order of round trip times (RTT) on which the predominant transfer protocol TCP operates are typically generated by just a few high bandwidth connections. This discovery has led to a decomposition of internet traffic into an alpha and a beta component. The alpha component collects all traffic generated by a few high bandwidth connections; the beta component is defined as all the residual traffic. As noted in [WSRB02, SRB02], the beta component is statistically close to fractional Gaussian noise (fGn) while the alpha component is highly non-Gaussian and bursty. The goal of this thesis is to assess the impact of the two components on queueing behavior under various network conditions.

Throughout this work, we follow [WSRB02, SRB02] and model the beta component by a fGn process. This is a well known Gaussian model with LRD which approximates network traffic produced by a large number of identical sources particularly well [Nor97, LR97, WTSW97, DO95]. Since beta connections are reported to be sending in first approximation at equal rates, an fGn model appears indeed most suitable for the beta component [WSRB02, SRB02].

However, only little indication is given in [WSRB02, SRB02] as to an appropriate choice of a model for the alpha component. In this thesis, we propose two different approaches to this end. In the first one, called the *self-similar burst model*, the alpha component is modelled as a self-similar bursty process while the second one, called the *ON/OFF burst model* represents the alpha component by one high rate ON/OFF source. Thereby, we will always assume that the alpha and beta component are statistically independent, which seems reasonable in view of the fact that the two components are generated by different types of connections.

In both modelling approaches we exploit the useful concept of the ON/OFF source. An ON/OFF source alternates between ON periods when sending or receiving a file and OFF periods when idle. It provides, thus, a simple yet effective abstraction of an actual traffic source. Thereby, an ON/OFF source will send with equal transfer rates during all its ON periods.

In its most classical form, an ON/OFF model of network traffic consists of a superposition of equal ON/OFF sources. It is notable, that the aggregate process obtained after superposing several such ON/OFF sources with identical rates has two different limiting regimes. The limit of the aggregate process is fractional Brownian motion (fBm) when the number of ON/OFF sources goes to infinity faster than the time scale. On the other hand, if the number of sources is kept finite and the time scale goes to infinity, the aggregate process of the ON/OFF sources converges to stable Levy

motion (sLm). This latter model approximates the load of a few connections which work at a fast looking pace, while the former represents the traffic of an overwhelming number of connections. The increments of the fBm process is the fGn process, which is a stationary Gaussian process with strong correlation. The increments of sLm process is stable Levy noise (sLn), a sequence of independent stable variables. We note here that the stable random variable (r.v.) has a heavy tailed distribution function which could appear reasonable as a model for traffic bursts (see [LLDH02, Whi00]).

The above discussion makes the following choice natural. In the self-similar burst model the alpha component of the traffic is modelled by a sLn process which seems appropriate since the alpha traffic is generated by a few high bandwidth connections which operate at fast clocking rate. Indeed, the study presented in [WSRB02, SRB02] revealed that alpha connections run typically over paths with very small round trip times; however, the round trip time is the basic time scale at which the control mechanism TCP works. In this modelling approach, we represent the beta component by a fGn process which again is natural since it is a superposition of the traffic of many slow sources.

First, to exploit the benefits of dealing with self-similar processes we assume in the self-similar burst model that alpha and beta components possess the same self-similarity parameters. This simplifies the analysis greatly. Our first result, which is derived purely from the self-similarity, relates that the heavy tailed traffic bursts of the alpha component may affect the queue tail strongly. More precisely, the probability of buffer overflow takes on the usual Weibull form for small buffers, governed by the beta component; yet, this behavior is contrasted by a power-law function for large buffers where the alpha component dominates.

Second, in order to study the queueing of the self-similar model we de-multiplex the incoming traffic aggregate into two flows which are served by two queues, one

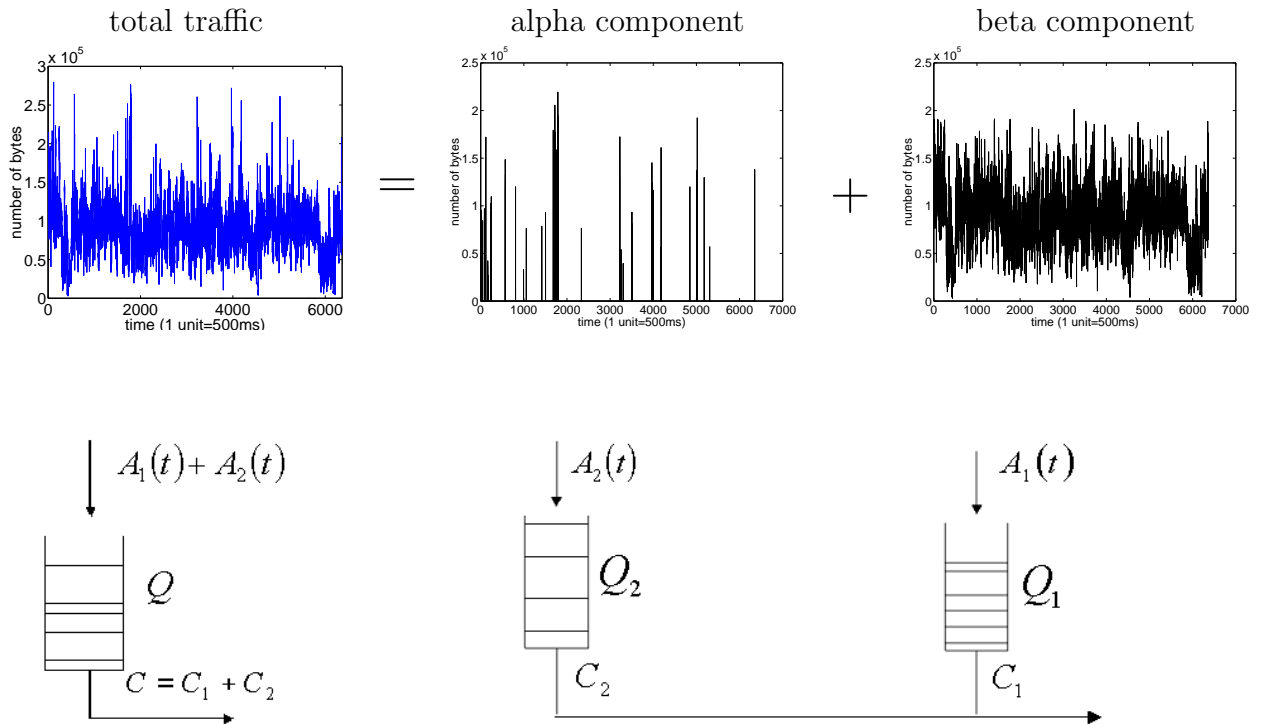


Figure 1.1 De-Multiplexing of two flows

of them serving the alpha component and the other serving the beta component. Figure (1.1) shows the single queue and de-multiplexed queues. The service rate of the queues and the buffer sizes are free parameters of the de-multiplexing operation. Our goal is to choose the parameters to achieve efficient de-multiplexing. Indeed, the sum of the sizes of the two queues must be larger than the size of the single queue. We investigate for which parameter setting the sum of the two queues becomes close to a single queue. Here, the critical time scale (CTS) emerges as an important parameter for de-multiplexing. We will show that for efficient de-multiplexing the service rates of queues and queue sizes should be chosen in such a way as to make the CTS of all queues the same.

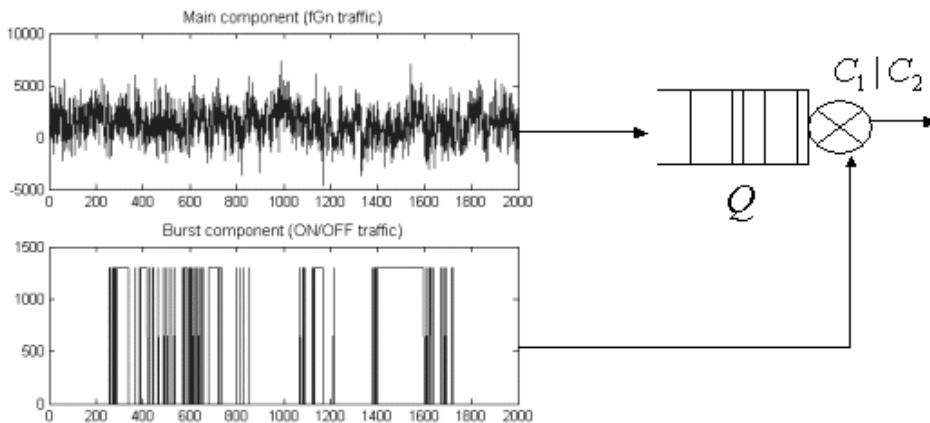


Figure 1.2 A Queue with variable service rate

In the second model, namely the *ON/OFF burst model*, the alpha component is modelled by a high rate ON/OFF source. Therefore the network traffic is the superposition of the beta component which collects all small rate connections and one high rate ON/OFF source. This is justified by the fact that typically only one alpha connection is active of a time (see [SSB02, SRB02]). The ON/OFF source increases the mean value of the traffic during the ON periods. So at large time scales the ON/OFF source produces traffic bursts. When the ON/OFF burst model is offered as the input traffic to a network queue, the ON/OFF source may or may not affect the queueing behavior.

For an analysis of the queueing behavior, we define the free capacity of the queue as the service rate of the queue minus the mean arrival rate of the beta component. If the rate of the ON/OFF source during the ON periods is less than the free capacity of the queue then the ON/OFF source does not change the asymptotic behavior of the queue. In other words it does not have much effect on the queueing behavior. However if the rate of the ON/OFF source during the ON period is larger than the

free capacity then the average queue size increases during the ON periods. So the queueing behavior can be strongly changed by the ON/OFF source.

In the presence of the beta component the high rate ON/OFF source can be viewed as providing variable service rate. The service rate for the beta component when the ON/OFF source is ON is equal to the full service rate minus the rate of the ON/OFF source and it is equal to the full service rate when the source is OFF. Therefore we prepare to study a variable service queue to analyze the queueing behavior the ON/OFF burst model.

First we analyze the queueing behavior of a Markov service rate queue, meaning discrete time queue where the service rate changes as a Markov chain. Necessarily, the ON period of the ON/OFF source in the setting has a short-tailed distribution which may be unrealistic. This leads us to study the renewal service rate queue, where the variation of the service rate is modelled by a renewal process. Here, a long-tailed ON period source maybe modelled by a renewal ON/OFF source with a long-tail distribution function for the ON periods.

Second, we analyze the queueing behavior and the effect of the ON/OFF source in the queue for several cases and we compare the queueing analysis results for the self-similar burst model and the ON/OFF the burst model.

Here is the structure of the thesis. In section 2.1 we study the self-similar model and the ON/OFF source concept and existing work on this subject. In the next section we introduce the self-similar burst model which assumes that the alpha component is well approximated by a self-similar bursty process. In section 2.3 we analyze the queueing behavior for the self-similar burst model by self-similarity techniques of queueing analysis. Next we propose a method for de-multiplexing a single flow into two flows in a general queue. By this method we de-multiplex the self-similar burst

model in order to compute the upper bound for probability of buffer overflow in section 2.7.

In the first section of chapter 3 we survey existing work on the effect of the high-rate ON/OFF source in the queue (section 3.1). Next we introduce the ON/OFF burst model where the alpha traffic is modelled by an ON/OFF source. Sections 3.4 and 3.5 explain the behavior of the variable service rate queue which we use as an abstraction of the effect of the one high rate ON/OFF source. In section 3.6 we analyze the queueing behavior of the ON/OFF burst model for several cases. The final section, 3.7 provides a summary of the queueing behavior for the ON/OFF burst model. Finally, chapter 4 presents a comparison of the self-similar burst model with the ON/OFF burst model and summarizes the important results of this thesis.

Chapter 2

Background

2.1 Classical model and self-similarity

2.1.1 $M/M/1$ queue

The $M/M/1$ model is a queueing model where both the distribution of customer arrivals and the distribution of service times are assumed to be exponential, and there is a single server. Let $f(t)$ and $g(u)$ be the probability density function of the inter-arrival times and service time respectively. In the $M/M/1$ queue $f(t)$ and $g(u)$ are exponential functions, which can be defined as:

$$f(t) = \lambda \exp(-\lambda t)$$

$$g(u) = \mu \exp(-\mu u)$$

The distribution function of $M/M/1$ queue is exponential. It is determined by the following formula

$$P[Q > b] = \exp(-(\mu - \lambda)b) \tag{2.1}$$

This equations shows the queue tail of $M/M/1$ queue is always an exponential functions which decreases very fast.

2.1.2 self-similarity

The network traffic has self-similar scaling properties. It can be modelled by fractional Gaussian noise (fGn) process. The aggregate process of fGn process is fractional Brownian motion (fBm). The fBm is a self-similar process.

A process Z is H-self-similar with stationary increments (H-sssi) if for all $a > 0$

$$Z(at) \stackrel{d}{=} a^H Z(t) \quad (2.2)$$

where H is the self-similar parameter ($0 < H < 1$). The random process fBm satisfies (2.2). It is non-stationary process since it has stationary increments. The increments of fBm, $G(n)$ is defined by

$$G_H(n) = B_H(n\delta t) - B_H((n-1)\delta t) \quad (2.3)$$

for finite δt . The fGn satisfies in scaling property

$$m^{1-H} G^{(m)} \stackrel{d}{=} G \quad (2.4)$$

where $G^{(m)}$ is the aggregate process. It is defined as:

$$G^{(m)}(k) = \frac{1}{m} \sum_{i=(k-1)m+1}^{km} G(i) \quad (2.5)$$

Also fGn process is a Long-Range-Dependence (LRD) process when $.5 < H < 1$, which means the samples of process at different times are strongly correlated. The autocorrelation function of LRD process has a long tail so sum of the values of autocorrelation function diverges. The network traffic exhibits both self-similar scaling and long-range dependence properties. So, the fGn traffic which has these properties is a good model for network traffic.

2.2 ON/OFF Model

The connection oriented view of internet traffic considers the traffic as a superposition of independent ON/OFF sources which have heavy-tailed distributions for the ON and OFF periods. In the ON/OFF model an individual source is modelled as an alternating renewal process. When the source is active or ON, it sends packets into

the network, and when it is inactive or OFF, it is idle and does not send any packets.

Let $\{X_i(t), t \geq 0\}$ be a stationary process, where

$$X(t) = \begin{cases} 1 & \text{if } t \text{ lies in an ON period} \\ 0 & \text{if } t \text{ lies in an OFF period} \end{cases}$$

The length of the ON intervals are i.i.d., and the length of the OFF intervals are i.i.d, and also the length of ON and OFF periods are independent. Assume that $X_1(t), X_2(t), \dots, X_M(t)$ are M ON/OFF sources, then

$$\text{total traffic} = X_1(t) + X_2(t) + \dots + X_M(t)$$

Let $\bar{F}_{\text{on}}(x)$ and $\bar{F}_{\text{off}}(x)$ represent the complementary distribution function of the ON and OFF intervals. Let σ_{on} and σ_{off} denote the variance of ON and OFF interval lengths. Assume that as $x \rightarrow \infty$

$$\text{either } \bar{F}_{\text{on}}(x) = L_{\text{on}}(x)x^{-\alpha_{\text{on}}}, 1 < \alpha_{\text{on}} < 2 \quad \text{or } \sigma_{\text{on}} < \infty$$

and

$$\text{either } \bar{F}_{\text{off}}(x) = L_{\text{off}}(x)x^{-\alpha_{\text{off}}}, 1 < \alpha_{\text{off}} < 2 \quad \text{or } \sigma_{\text{off}} < \infty$$

where $L_{\text{on}}(x) > 0$ and $L_{\text{off}}(x) > 0$ are slow varying functions at infinity, which means

$$\lim_{x \rightarrow \infty} \frac{L(tx)}{L(x)} = 1$$

for any $t > 0$. For example the constant function and the Log function are slow varying functions at infinity. Here, the exponents α_{on} and α_{off} are the *tail parameters* for the ON and OFF periods respectively. Also let μ_{on} and μ_{off} be the expected values of ON and OFF interval lengths respectively.

The superposition packet count at time t is $\sum_{i=1}^M X_i(t)$. So the aggregate cumulative packet for the interval $[0, t]$ is

$$Y(t) = \int_0^t \left[\sum_{k=1}^M X_k(u) \right] du$$

The limit of the scaled random process $\{Y(Tt)\}$ has two different regimes. The limit regimes depend on the distribution function of the ON and OFF periods, and how the parameters M (number of sources) and T (scaling parameter) go to infinity.

The following theorems 2.1 and 2.2 ([TL86, TWS97, GK02]) determine the limits of the aggregate process $Y(Tt)$ when $M \rightarrow \infty$ and $T \rightarrow \infty$.

Theorem 2.1 For large M and T , the aggregate cumulative packet process $\{Y(Tt), t > 0\}$ behaves statistically like

$$TM \frac{\mu_{\text{on}}}{\mu_{\text{on}} + \mu_{\text{off}}} t + T^H \sqrt{L(T)M} \sigma B_H(t)$$

where $H = (3 - \min(\alpha_{\text{on}}, \alpha_{\text{off}}))/2$ is self-similar parameter and σ is the scale parameter which depends on the distribution function of ON and OFF periods. The $B_H(t)$ is standard fractional Brownian motion. In the limit:

$$\lim_{T \rightarrow \infty} \lim_{M \rightarrow \infty} \frac{Y_M(Tt) - TM \frac{\mu_{\text{on}}}{\mu_{\text{on}} + \mu_{\text{off}}} t}{T^H \sqrt{L(T)M}} = \sigma B_H(t)$$

where the limit converges in the sense of the finite-dimensional distributions.

Theorem 2.2 For a large T , the aggregate packet process of an ON/OFF sources $\{X(Tt), t > 0\}$ behaves statistically like the stable Levy motion (sLm) [ST94, LLDH02]. In the limit:

$$\lim_{T \rightarrow \infty} \frac{1}{T^{\alpha_{\min}}} \int_0^{Tt} [X(u) - \frac{\mu_{\text{on}}}{\mu_{\text{on}} + \mu_{\text{off}}} Tt] du = \sigma L_{\alpha, \beta}(t)$$

where $\alpha = \min(\alpha_{\text{on}}, \alpha_{\text{off}})$ and σ is the scale parameter which depends on the distribution functions of ON and OFF periods. The $L_{\alpha, \beta}(t)$ is stable Levy motion process with scaling parameter 1, and skew β . The $\beta = 1$ when $\alpha_{\text{on}} > \alpha_{\text{off}}$ and $\beta = -1$ when $\alpha_{\text{on}} < \alpha_{\text{off}}$ and $-1 < \beta < 1$ when

$\alpha_{\text{on}} = \alpha_{\text{off}}$. In the particular case that ON and OFF periods have the same distribution function, $\beta = 0$ and $L_{\alpha,\beta}(t)$ is a symmetric Levy stable motion.

2.3 Alpha-Beta decomposition

The study in [WSRB02, SRB02] on the Internet traffic shows that the traffic bursts come from huge file transmissions over high bandwidth links and concludes that the internet traffic is non-Gaussian. But if the traffic which is generated by the few high bandwidth connections is taken away, the residual traffic becomes Gaussian and non-bursty. Consequently, we define the alpha component according to [WSRB02, SRB02] to be the traffic which is generated by the few high bandwidth connections and the beta component to be the residual traffic.

$$\text{total traffic} = \text{alpha component} + \text{beta component}$$

The Alpha-Beta traffic can be modelled by ON/OFF sources. Most of the low bandwidth connections (more than 98% of connections by [WSRB02, SRB02]) can be modelled by ON/OFF sources with the same rate and the same inter-arrival distribution function. These sources are *beta-sources* which model beta-connections [WSRB02, SRB02]. The superposition of beta-sources models the beta component. The few high bandwidth connections (less than 2% of connections by [WSRB02, SRB02]) are modelled by high rate ON/OFF sources. These ON/OFF sources which model alpha connections are *alpha-sources*. The alpha connections are high bandwidth connections. So the load of the alpha-sources is more long tailed than that of the beta-sources. Because they have high bandwidth and they send and receive huge files, the ON periods of alpha-sources are more long tailed than that of the beta-sources.

By theorem 2.1 when there are many ON/OFF sources and they have almost the same rate, the aggregate cumulative process of sources in the large time scale behaves like a fBm process. So the superposition of many i.i.d. ON/OFF sources at large time scales is fractional Gaussian noise (fGn). The fGn process is the increment of fBm traffic and is a stationary and strongly correlated process. So the beta component which is the superposition of the many beta-sources is well modelled by a fGn process.

The theorem 2.2 shows that the aggregate process of an ON/OFF source converges to stable Levy motion (sLm) [ST94, Whi00] at very large time scales. So the superposition of few ON/OFF sources converges to stable Levy noise (sLn) when the time scale is large. The sLn process, the increment of sLm, which is a stationary process with heavy-tailed marginals. So the alpha component which is the superposition of few high bandwidth connections could be modelled by a sLn process. The high bandwidth connections send and receive huge files and the files have a heavy-tailed distribution function [CB97]. On the other hand, the Round-trip-time (RTT) for high bandwidth TCP connection is too small, so the connection can be considered as an ON/OFF source which has been compressed in time. So, when the time scale is large enough, the alpha component is well modelled by a sLn process.

Throughout this work ([WSRB02, SRB02]) the beta traffic is modelled by a fractional Gaussian noise (fGn) process which is a good model for beta component. In many investigations the fGn process has been introduced as a model for Internet traffic [WPRT, Nor97, LR97, WTSW97]. On the other hand, the alpha traffic can be modelled in many different ways, because the number of bursty periods is much smaller compared to the number of non-bursty periods and the heights of the bursts vary a great deal. Furthermore, the study of [CB97] shows that the internet file sizes have a heavy-tailed distribution function. So, the traffic bursts in large time scales have a heavy-tailed distribution.

If we suppose that the traffic is self-similar, then the beta component (fGn traffic) and the alpha component of the traffic will be self-similar with the same self-similarity parameters. Under those conditions (self-similar heavy-tailed arrivals) the fractional Stable noise (fSn) could be a good model for alpha traffic. The self-similar parameter of fSn is chosen to be the same as fGn and the tail of the traffic bursts determines the tail parameter of the fSn process. Also in [KH98, BM00] the internet traffic with heavy-tailed distribution in bursty points is modelled by a fSn process only, which shows that in some cases it can model the bursty traffic well by itself.

Let $\alpha_{\text{on}(\beta)}$ represent the tail parameter of beta-sources at ON period and the $\sigma_1 B_{H_1}(t)$ denote the aggregate cumulative process of beta-sources. By theorem 2.1:

$$H_1 = \frac{3 - \alpha_{\text{on}(\beta)}}{2} \quad (2.6)$$

Also, let $\alpha_{\text{on}(\alpha)}$ represent the tail parameter of alpha-sources and the $\sigma_2 L_{H_2}(t)$ denote the aggregate cumulative process of alpha-sources. By theorem 2.2:

$$H_2 = \frac{1}{\alpha_{\text{on}(\alpha)}} \quad (2.7)$$

The network traffic which is a superposition of independent alpha and beta components exhibits the self-similarity properties. If we assume that the superposition of fGn and sLn processes is self-similar, then the fGn and sLn processes have the same self-similarity parameter (because sum of two independent self-similar traffics is self-similar if and only if they have the same self-similarity parameters). This means that

$$H_1 = H_2$$

and the equations (2.6) and (2.7) imply that

$$\alpha_{\text{on}(\alpha)} < \alpha_{\text{on}(\beta)}.$$

This is consistent with the assumptions about the alpha and beta sources that the alpha sources are more long tailed than beta sources.

So by using the above results on the ON/OFF model we propose a model for the network traffic which is a sum of the fGn and sLn processes. The model is explained in chapter 3.

2.4 Exploiting self-similarity in queueing analysis

Let us set up same notation in queueing analysis. Let $A(t)$ be the total arrival into the queue from 0 to the time t and C service rate of the queue. Thus, $A(t) - A(s)$ is the amount of traffic comes into the queue during the time interval $(s, t]$ and $C(t - s)$ the maximum amount of traffic which is served by the queue in that time interval. So by the expansion of the Lindley's formula [Pra97, CY01] the backlog traffic at the time t is

$$W^{A,C}(t) = \sup_{0 \leq s \leq t} (A(t) - A(s) - C.(t - s)) \quad (2.8)$$

For a stationary reversible input traffic (the increment of a self-similar traffic has these properties), the backlog r.v. could be defined as

$$W^{A,C} = \sup_{t \geq 0} (A(t) - Ct). \quad (2.9)$$

Also the following inequality (2.10) gives a lower bound for the $P[W^{A,C} > b]$

$$P[W^{A,C} > b] = P[\sup_{t \in \Omega} W^{A,C}(t) > b] \geq \sup_{t \in \Omega} P[W^{A,C}(t) > b] \quad (2.10)$$

There are several methods to perform the queueing analysis of a self-similar traffic. In the Norros model for a self-similar traffic [Nor97, LR97, DO95, NW98], the aggregate process of traffic is defined as:

$$A(t) = Z_H(t) + mt \quad (2.11)$$

where $Z_H(t)$ is a self-similar process with self-similarity parameter H and m is mean arrival of traffic. Therefore, the queue size distribution can be computed as:

$$P[Q > b] = P[\sup_{t \geq 0} A(t) - Ct > b] = P[\sup_{t \geq 0} Z_H(t) + mt - Ct > b] \quad (2.12)$$

so by the equations (2.12), (2.9) and (2.10)

$$P[Q > b] = P[\sup_{t \geq 0} Z_H(t) - (C - m)t > b] \geq \sup_{t \geq 0} P[Z_H(t) - (C - m)t > b]. \quad (2.13)$$

Since $Z_H(t)$ is self-similar, $Z_H(t) \stackrel{d}{=} t^H Z_H(1)$ we have

$$\sup_{t \geq 0} P[Z_H(t) - (C - m)t > b] = \sup_{t \geq 0} P[Z_H(1) > \frac{b + (C - m)t}{t^H}] \quad (2.14)$$

For maximizing $P[Z_H(1) > \frac{b + (C - m)t}{t^H}]$, the value of $\frac{b + (C - m)t}{t^H}$ should be minimized.

This is achieved when $t = t^*$ where

$$t^* = \frac{Hb}{(1 - H)(C - m)}. \quad (2.15)$$

The time t^* which maximizes the probability in (2.14) is called the *Critical Time Scale* (CTS). In other words, at the CTS time the probability of overflowing of the queue is maximized. The CTS is an important parameter to find the asymptotic queueing behavior of the queue.

So, by following the basic ideas of Norros, we achieve the following lemma.

Lemma 2.1 Assume that the aggregate input traffic is defined as (2.11)

then the queue tail of traffic has following lower bound.

$$P[Q > b] \geq P[Z_H(1) > \kappa b^{1-H}] \quad (2.16)$$

where $\kappa = \frac{(C-m)^H}{H^H(1-H)^{1-H}}$ depends on the self-similar parameter (H) and the free capacity of the queue ($C - m$).

We check the inequality (2.16) for two well known self-similar processes: the fBm and sLm. Before explaining queueing analysis of each process, let define the asymptotic equality of two functions that we use in our work.

Definition 2.1 Two functions $f(x)$ and $g(x)$ are asymptotically equal when $h(x) = \frac{f(x)}{g(x)}$ is a slow varying function at infinity. The asymptotic equality is represented as: $f(x) \asymp g(x)$.

Also when the $\log f(x) \asymp \log g(x)$ then we represent then as $f(x) \stackrel{\log}{\asymp} g(x)$. It is clear, for positive functions $f(x)$ and $g(x)$, when $f(x) \asymp g(x)$ then we will have $f(x) \stackrel{\log}{\asymp} g(x)$ too.

First assume that the input traffic is a fGn process; then the aggregate traffic is a fBm process ([WPRT, Nor97, ST94, BD01]). In other words $Z_H(t) \stackrel{d}{=} \sigma B_H(t)$, where σ is the scale parameter and $B_H(t)$ is the standard fractional Brownian motion ($B_H(1) \sim \mathbf{N}(\mathbf{0}, \mathbf{1})$ Gaussian r.v.).

By the inequality (2.16)

$$\begin{aligned} P[Q > b] &\geq P[\sigma B_H(1) > \kappa b^{1-H}] \\ &\stackrel{\log}{\asymp} \exp(-(\frac{\kappa}{\sigma} b^{1-H})^2/2) \\ &= \exp(-\gamma_1 b^{2(1-H)}) \end{aligned} \tag{2.17}$$

This lower bound agrees with results of Norros and Duffield [Nor97, DO95, LR97] for fGn traffic. They have shown this lower bound is tight when $b \rightarrow \infty$. It shows that the queue tail decreases as a Weibull function with parameter $2(1-H)$. As we noted, for a self-similar LRD traffic $0.5 < H < 1$, therefore $0 < 2(1-H) < 1$, which means the queue tail decreases slower than an exponential function.

In the second case, assume the input process to the queue is a sLn process ($\beta = 1$), the aggregate traffic is sLm. In other words $Z_H(t) \stackrel{d}{=} \sigma L_H(t)$, where $H = 1/\alpha$, α being

the tail parameter and σ the scale parameter ($L_H(1) \sim \mathbf{S}(\alpha, \mathbf{1}, \mathbf{1}, \mathbf{0})$ stable r.v.). By the inequality (2.16)

$$\begin{aligned} P[Q > b] &\geq P[L_H(1) > \frac{\kappa}{\sigma} b^{1-H}] \\ &\asymp \frac{\Gamma(\alpha + 1) \sin(\frac{\pi\alpha}{2})}{\pi\alpha} \left(\frac{\kappa}{\sigma} b^{1-H}\right)^{-\alpha} \\ &= M_\alpha \gamma_2 \cdot b^{-\alpha+1} \end{aligned} \tag{2.18}$$

This lower bound agrees with a result of Laskin and Lambadaris in [LLDH02](also [MDM02]). It shows that the queue size has a power-law asymptotic with the exponent $-\alpha + 1$.

The asymptotic lower bound for the queue tail for fGn and sLn processes show that when the fGn process is the input traffic of a queue, the queue tail decreases much faster than when the sLn process is the input traffic.

2.5 Multiplexing of an ON/OFF source and a background traffic

The alpha component could be modelled by an ON/OFF source in different ways and in each case it has a different effect on the network queue. In the studies [WSRB02, SRB02], which analyze traffic bursts, the Internet traffic is divided into 500ms time bins. Then, the bursty time bins are determined which contain an alpha component by using signal processing techniques. Here, we model the traffic bursts by an ON/OFF source: the source is ON in the bursty time bins and is OFF in the non-bursty time bins. Next, we study the queueing behavior of this model under the simplifying assumption that the ON/OFF source's state changes as a Markov chain from a one bin to the next. The bursts of the alpha component which is generated by an ON/OFF source in the *Markov model* have an exponential (thus

short-tailed) distribution function. For more realistic we consider bursts which have a long-tailed distribution; here, a renewal process with long-tailed distribution for the ON period can be used as a model for the bursts of the alpha component. In [JL99, Jel98, AMN99, Box96, BC00] the renewal ON/OFF source with a long-tailed distribution for the length of the ON period has been studied. Also the asymptotic queuing behavior has been analyzed when the input traffic is the superposition of the ON/OFF source and some other traffic source. Before explaining the main theorem of their studies, let define the *long-tailed* and *subexponential* random variable.

Definition 2.2 The random variable x is *long-tailed*, if for each $y \in \mathbb{R}$

$$\lim_{x \rightarrow \infty} \frac{P[X > x - y]}{P[X > x]} = 1$$

Examples: The Pareto, Weibull ($0 < \beta < 1$) and log-normal random variables are long-tailed.

Definition 2.3 The random variable x is *subexponential*, if

$$\lim_{x \rightarrow \infty} \frac{P[X + X' > x]}{P[X > x]} = 2$$

where X' is an independent copy of X . Example: The Pareto random variable is subexponential.

Definition 2.4 The integrated tail distribution of the random variable X is X^* . The complementary distribution function of X^* is defined as:

$$P[X^* > b] = \frac{1}{\mathbb{E}X} \int_b^{\infty} P[X > u] du$$

Theorem 2.3 Assume that $A_1(t)$ is the aggregate process of the traffic with mean arrival R and $A_2(t)$ is the aggregate process of the renewal

ON/OFF source with mean arrival ρr_{on} , where r_{on} is the rate during the ON periods. If $r_{\text{on}} + R > C > \rho r_{\text{on}} + R$, and if $A_1(t)$ satisfies

$$P[W^{A_1(t), R+\varepsilon} > x] = O(P[T_\alpha^* > \frac{x}{R + r_{\text{on}} - C}])$$

for every $\varepsilon > 0$, then

$$\lim_{x \rightarrow \infty} \frac{P[W^{A_1+A_2, C} > x]}{P[W^{A_2, C-R} > x]} = 1$$

(for a proof of the theorem see [AMN99]).

Chapter 3

A self-similar burst model

The traffic self-similar burst model is a self-similar bursty model for the network traffic. The aggregate traffic of self-similar burst model is the superposition of fractional Brownian motion (fBm) and stable Levy motion (sLm) ($\alpha = 1/H$ and $\beta = 1$) traffics with the same self-similar parameter. The aggregate process of self-similar burst model is defined as

$$A(t) = mt + \sigma_1 B_H(t) + \sigma_2 L_H(t) \quad (3.1)$$

where σ_1 and σ_2 are scale parameters of the fBm and sLm processes, and m is the mean value of the traffic. Figure (3.1) depicts the beta component as a fGn process and the alpha component as a sLn process. As figure (3.1) shows, in the self-similar burst model the mean value of the traffic is much bigger than the scale parameters of fGn and sLn processes. In addition the parameter β of sLn process is equal to 1. So the probability that the process takes negative values is almost zero. In the next few sections we analyze the queueing behavior of self-similar burst model by making use of the self-similarity of model.

Before starting to analyze the self-similar burst model, let explain why the model is self-similar. Next we analyze the queueing behavior of the model by using the self-similarity.

Proposition 3.1 The sum of two independent self-similar processes with the same self-similarity parameters is self-self-similar.

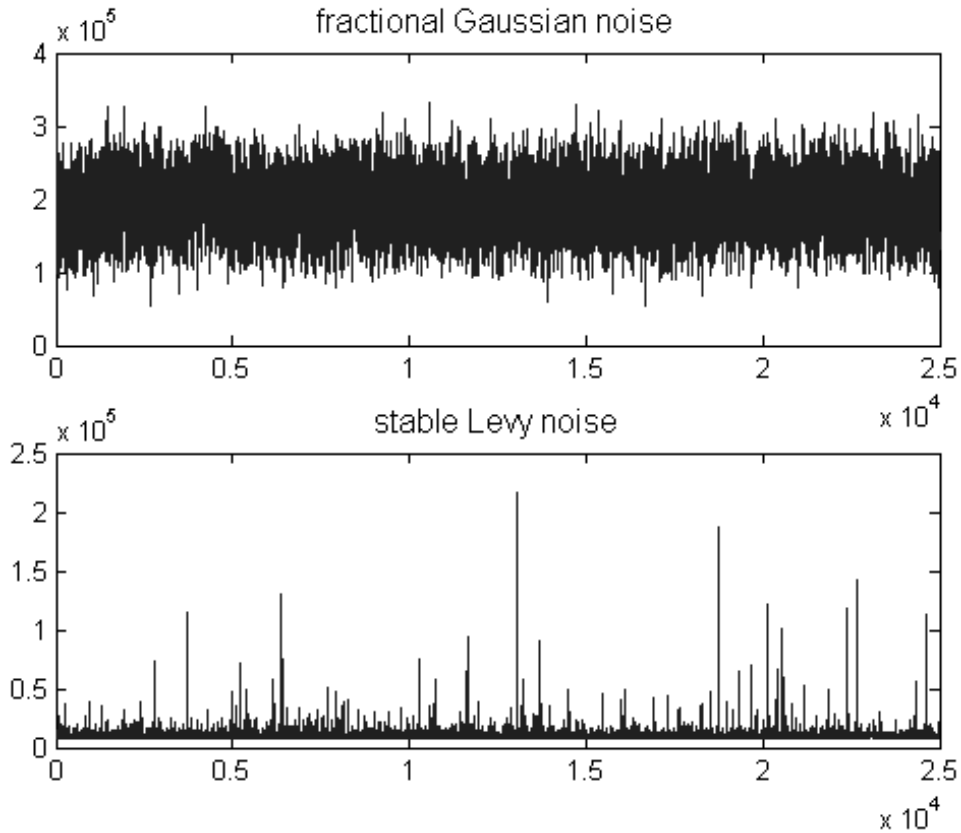


Figure 3.1 self-similar decomposition of traffic: the bursty component is modelled as sLn, the residual component by fGn

Proof of proposition 3.1 : Assume that $\sigma_1 B_H(t)$ and $\sigma_2 L_H(t)$ are two independent self-similar processes. Let

$$Z_H(t) = \sigma_1 B_H(t) + \sigma_2 L_H(t)$$

be the sum of self-similar processes. Also let $\Phi[Z_H(at_1), Z_H(at_2), \dots, Z_H(at_n)]$ be the deterministic function of finite random variables $Z_H(at_1)$ to $Z_H(at_n)$. By the assumptions that $\sigma_1 B_H(t)$ and $\sigma_2 L_H(t)$ are two independent self-similar processes we

have

$$\begin{aligned}
\phi[Z_H(at_1), \dots, Z_H(at_n)] &= \phi[\sigma_1 B_H(at_1) + \sigma_2 L_H(at_1), \dots, \sigma_1 B_H(at_n) + \sigma_2 L_H(at_n)] \\
&= \phi[\sigma_1 B_H(at_1), \dots, \sigma_1 B_H(at_n)] + \phi[\sigma_2 L_H(at_1), \dots, \sigma_2 L_H(at_n)] \\
&= \phi[a^H \sigma_1 B_H(t_1), \dots, a^H \sigma_1 B_H(t_n)] + \phi[a^H \sigma_2 L_H(t_1), \dots, a^H \sigma_2 L_H(t_n)] \\
&= \phi[a^H \sigma_1 B_H(t_1) + a^H \sigma_2 L_H(t_1), \dots, a^H \sigma_1 B_H(t_n) + a^H \sigma_2 L_H(t_n)] \\
&= \phi[a^H Z_H(t_1), \dots, a^H Z_H(t_n)]
\end{aligned}$$

This shows that $Z_H(t)$ is a self-similar process with the same same similarity parameter.

3.1 Queueing analysis of self-similar burst model

The aggregate process of self-similar burst model is the sum of the fBm and sLm processes. The fBm and sLm processes have very different statistical properties [ST94, Whi00]. In this section we explain the different effects of these processes in the network queue. When the traffic self-similar burst model is the input traffic of a queue, the fGn traffic and sLn traffics have different effects in the queue. As figure (3.2) shows, the queueing behavior of the self-similar burst model is strongly affected by the sLn process for large buffer sizes.

By equation (2.17) when the input traffic is a fGn process, the queue size has a Weibull tail for large b . But figure (3.2) shows that for very large values of b , the $P[Q > b]$ does not decrease as a Weibull function, instead it decreases as a power-law function, and when the fGn process is the only input traffic, the queue tail decreases very fast. Therefore, the power-law decay of queue tail is the effect of the sLn process in the queue.

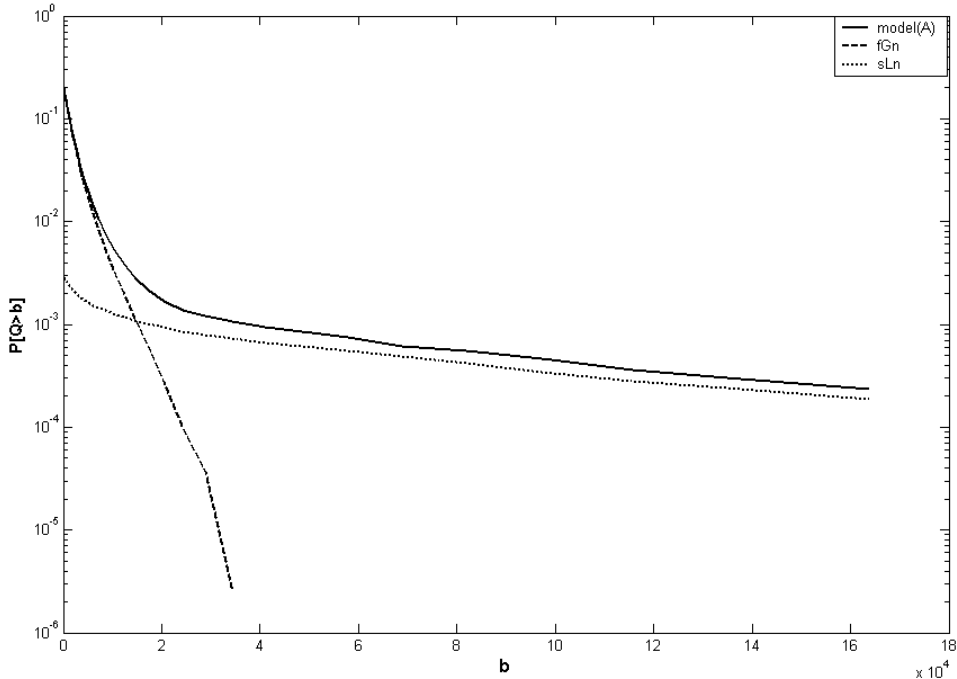


Figure 3.2 The queue behavior of self-similar burst model has two regimes: It has a Weibull decay like fGn traffic for small buffers, and it has power-law decay like sLn traffic for large buffers

The self-similar burst model is a self-similar process. So we can use the asymptotic lower bound for the self-similar traffic (equation (2.16)) to analyze the queuing behavior of self-similar burst model. The equations (2.12) and (2.16) imply

$$P[Q > b] \geq P[Z_H(1) > \kappa b^{1-H}] = P[\sigma_1 B_H(1) + \sigma_2 L_H(1) > \kappa b^{1-H}] \quad (3.2)$$

where $Z_H(1)$ is the sum of independent Gaussian and stable random variables. Therefore the density function of $Z_H(1)$ is the convolution of Gaussian and stable density functions. Figure (3.3) plots the density function of $Z_H(1)$ for a Gaussian and a non-symmetric stable random variable ($\beta = 1$). As figure (3.3) shows that the density function of $Z(1)$ is close to the Gaussian density function for small buffer sizes, and is close to the stable (heavy-tailed) density function for very large buffer sizes. This

means that the asymptotic lower-bound for $P[Q > b]$ is close to the case when the input is a fGn process (equation (2.17)) for small buffers, and the asymptotic lower bound is power-law as the case when the input process is a sLn process (equation (2.18)) for very large buffers. This agrees with simulation result in figure (3.2).

3.2 De-multiplexing of two general flows

In this section we de-multiplex the flows of a queue into two flows which feed into two multiplexed queues. This work gives us knowledge about how much each of the components affects the queueing behavior.

Figure (1.1) shows a queue with service rate C which is compared with a two queue system with service rates C_1 and C_2 where $C_1 + C_2 = C$. Our goal is to choose C_1 and C_2 such that the superposition of the de-multiplexed queue sizes is close to the queue size of the single queue, and therefore the queueing behavior of the single queue could be analyzed by using the two de-multiplexed queues. Let $A(t)$ be the aggregate process of the input traffic of the single queue, hence

$$A(t) = A_1(t) + A_2(t) \tag{3.3}$$

where $A_1(t)$ and $A_2(t)$ are the aggregate processes of the traffic(1) and the traffic(2). Let $W^{A_1, C_1}(t)$ and $W^{A_2, C_2}(t)$ be the corresponding backlog traffics of the queues Q_1 and Q_2 (equation (2.8)).

In the following sections we find some bounds for $W^{A, C}(t)$ in terms of $W^{A_1, C_1}(t)$ and $W^{A_2, C_2}(t)$ which are helpful for the queueing analysis of the single queue using the de-multiplexed queues.

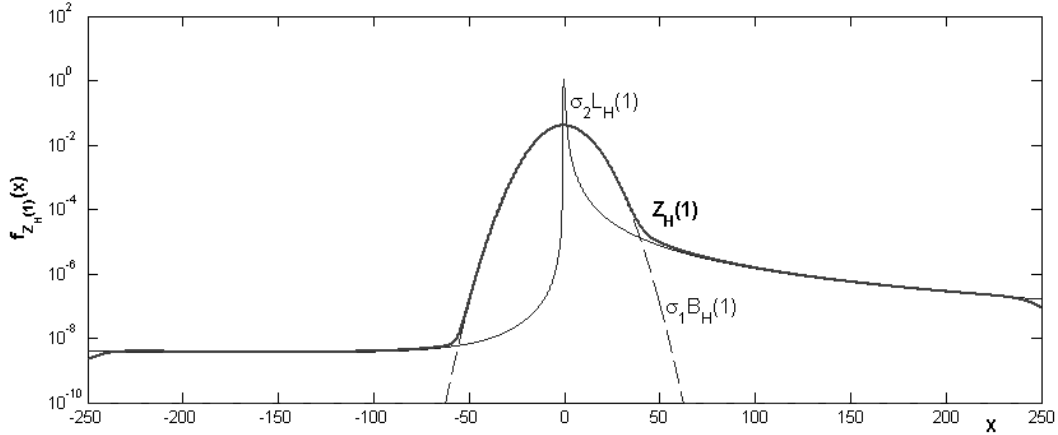


Figure 3.3 Marginal density of the self-similar bursty model $Z_H(1) = \sigma_1 B_H(1) + \sigma_2 L_H(1)$

3.2.1 Lower bounds

If we consider the single queue with only one of the two de-multiplexed flows feeding into it, then the queue size at each time is less than the case when the superposition of two traffics feeds into the queue. The proposition 3.2 determines a lower bound for $P[Q > b]$ in terms of the input components by using this idea.

Proposition 3.2 (lower bound) Assume for the queue with service rate C , that $A(t) = A_1(t) + A_2(t)$ is the aggregate process of the superposition of two traffics. $A_1(t)$ and $A_2(t)$ are the aggregate traffics of each component. Then for each $b > 0$

$$\max(P[W^{A_1, C} > b], P[W^{A_2, C} > b]) \leq P[W^{A, C} > b]$$

Proof of Proposition 3.2 : $A_1(t)$ and $A_2(t)$ are non-decreasing functions. So by the definition of $W^{A,C}(t)$ in (2.8)

$$\begin{aligned}
W^{A,C}(t) &= \sup_{0 \leq s \leq t} (A(t) - A(s) - C(t-s)) \\
&= \sup_{0 \leq s \leq t} (A_1(t) - A_1(s) + A_2(t) - A_2(s) - C(t-s)) \\
&\geq \max\left\{ \sup_{0 \leq s \leq t} (A_1(t) - A_1(s) - C(t-s)), \sup_{0 \leq s \leq t} (A_2(t) - A_2(s) - C(t-s)) \right\} \\
&= \max(W^{A_1,C}(t), W^{A_2,C}(t))
\end{aligned}$$

therefore

$$\begin{aligned}
P[W^{A,C} > b] &= P[\sup_{t \geq 0} W^{A,C}(t) > b] \\
&\geq P[\sup_{t \geq 0} \max(W^{A_1,C}(t), W^{A_2,C}(t)) > b] \\
&\geq \max(P[\sup_{t \geq 0} W^{A_1,C}(t) > b], P[\sup_{t \geq 0} W^{A_2,C}(t) > b]) \\
&= \max(P[W^{A_1,C} > b], P[W^{A_2,C} > b])
\end{aligned}$$

The above lower bound for $P[Q > b]$ tells us that the queueing behavior of the superposition of two traffic has a tail longer than the case when only one of the traffics is offered as the input traffic into the same queue. For example suppose the $P[W^{A_2,C} > b]$ has a power-law lower bound, then $P[W^{A,C} > b]$ will at least have a power-law lower bound.

3.2.2 Upper bounds

When the single queue and the de-multiplexed queues are compared, the total input traffic is same in both cases. However the service rate of the single queue is $C = C_1 + C_2$ when there is load in the queue but the sum of service rates of the de-multiplexed queues could be $C = C_1 + C_2$, $C = C_1 + 0$ or $C = C_2 + 0$ at different times. Indeed, when one of the de-multiplexed queues is empty, the other one serves

the loads with a service rate (C_1 or C_2) which is less than C . So the queue size of the single queue at each time $t \geq 0$ is less than the sum of queue sizes of the de-multiplexed queues. Proposition 3.3 determines an upper bound for $P[Q > b]$ in terms of the de-multiplexed queues by using that idea.

Proposition 3.3 (upper bound) Assume that $A(t) = A_1(t) + A_2(t)$ is the aggregate process of superposition of two traffics. Also $C = C_1 + C_2$ ($C_1, C_2 > 0$), then for each $b > 0$

$$P[W^{A,C} > b] \leq P[W^{A_1,C_1} + W^{A_2,C_2} > b]$$

Proof of Proposition 3.3 : By the definition of $W^{A,C}(t)$ in (2.8)

$$\begin{aligned} W^{A,C}(t) &= \sup_{0 \leq s \leq t} (A(t) - A(s) - C(t-s)) \\ &= \sup_{0 \leq s \leq t} \{A_1(t) + A_2(t) - A_1(s) - A_2(s) - C_1(t-s) - C_2(t-s)\} \\ &\leq \sup_{0 \leq s \leq t} (A_1(t) - A_1(s) - C_1(t-s)) + \sup_{0 \leq s \leq t} (A_2(t) - A_2(s) - C_2(t-s)) \\ &= W^{A_1,C_1}(t) + W^{A_2,C_2}(t) \end{aligned}$$

therefore

$$\begin{aligned} P[W^{A,C} > b] &= P[\sup_{t \geq 0} W^{A,C}(t) > b] \\ &\leq P[\sup_{t \geq 0} (W^{A_1,C_1}(t) + W^{A_2,C_2}(t)) > b] \\ &\leq P[\sup_{t \geq 0} W^{A_1,C_1}(t) + \sup_{t \geq 0} W^{A_2,C_2}(t) > b] \\ &= P[W^{A_1,C_1} + W^{A_2,C_2} > b] \end{aligned}$$

The above upper bound of overflow probability of the single queue could be found in terms of the overflow probabilities of the de-multiplexed queues.

Proposition 3.4 Assume that Q_1 and Q_2 denote the de-multiplexed queues, then

$$P[Q > b] \leq P[Q_1 > \eta b] + P[Q_2 > (1 - \eta)b]$$

where $b > 0$ and $0 < \eta < 1$

This inequality can be interpreted as the probability that a queue with size b and input $A_1(t) + A_2(t)$ overflows. It is smaller than the sum of probabilities that one of the queues with size ηb and $(1 - \eta)b$, and with input traffics $A_1(t)$ and $A_2(t)$, overflows.

3.2.3 De-multiplexing parameters

The free parameters of the de-multiplex model $\{C_1, C_2, \eta\}$ should be chosen in such a way as to the upper bound $(P[Q_1 > \eta b] + P[Q_2 > (1 - \eta)b])$ for the probability of overflowing $(P[Q > b])$ minimize. The theorem 3.1 gives some information about the critical time scale (equation (2.14)) of the queues in the efficient de-multiplexing case. It helps us to compute the free parameters for efficient de-multiplexing when the input process is self-similar.

Theorem 3.1 Let Q be the single queue. Also Q_1 and Q_2 be the de-multiplexed queues. If the $P[Q > b]$ and $P[Q_1 > \eta b] + P[Q_2 > (1 - \eta)b]$ have same asymptotic queueing behavior. Further if Q_1 and Q_2 have only one critical time scale, then the Q , Q_1 and Q_2 will have the same critical time scale.

Proof of theorem 3.1: Let t^* , t_1^* and t_2^* represent the CTS of the single queue and de-multiplexed queues. By the inequality (2.10)

$$P[Q_1 > \eta b] \geq P[A_1(t_1^*) - C_1 t_1^* > \eta b] \tag{3.4}$$

at time t_1^* , $P[A_1(t) - Ct > \eta b]$ is maximized. So for $t = t^*$

$$\begin{aligned} P[Q_1 > \eta b] &\geq P[A_1(t_1^*) - C_1 t_1^* > \eta b] \\ &\geq P[A_1(t^*) - C_1 t^* > \eta b] \end{aligned} \quad (3.5)$$

similarly for t_2^* :

$$\begin{aligned} P[Q_2 > (1 - \eta)b] &\geq P[A_2(t_2^*) - C_2 t_2^* > (1 - \eta)b] \\ &\geq P[A_2(t^*) - C_2 t^* > (1 - \eta)b] \end{aligned} \quad (3.6)$$

By these inequalities (3.5) and (3.6)

$$\begin{aligned} P[Q_1 > \eta b] + P[Q_2 > (1 - \eta)b] &\geq P[A_1(t_1^*) - C_1 t_1^* > \eta b] + P[A_2(t_2^*) - C_2 t_2^* > (1 - \eta)b] \\ &\geq P[A_1(t^*) - C_1 t^* > \eta b] + P[A_2(t^*) - C_2 t^* > (1 - \eta)b] \\ &\geq P[A_1(t^*) - C_1 t^* + A_2(t^*) - C_2 t^* > \eta b + (1 - \eta)b] \\ &= P[A_1(t^*) + A_2(t^*) - (C_1 + C_2)t^* > b] \\ &= P[A(t^*) - Ct^* > b] \end{aligned}$$

On the other hand, the last term ($P[A(t^*) - Ct^* > b]$) is the asymptotic lower bound of the $P[Q > b]$. So if the $P[Q_1 > \eta b] + P[Q_2 > \eta b]$ and $P[Q > b]$ are asymptotically equal then $P[Q_1 > \eta b] + P[Q_2 > (1 - \eta)b] \asymp P[A(t) - Ct]$. This shows that all of the inequalities become equalities asymptotically. But by the hypothesis $Q_1(t)$ has only one CTS which is t_1^* . So if the inequalities convert to asymptotic equality, then $t^* = t_1^*$. Similarly $t^* = t_2^*$. Therefore the CTS of the queues are equal. In other words $t^* = t_1^* = t_2^*$. It makes sense because when the single queue and de-multiplexed queues have almost the same behavior, we expect the overflowing probability of all the queues to be maximized at the same time scale.

Remark: When the aggregate traffic is a self-similar process then it has only one critical time scale which depends on the self-similar parameter, the buffer size and

the free capacity of the queue (equation (2.15)). By theorem 3.1, we can determine the efficient de-multiplexing for a single queue when input traffic is a superposition of two self-similar inputs.

Example 1: Suppose that $A_1(t) = m_1 t$ is the aggregate process of a Constant Bit Rate (CBR) traffic and $A_2(t) = m_2 t + Z_H(t)$ is the aggregate process of a self-similar traffic. For Q_1 , the CTS time can be chosen arbitrary, because when $C_1 \geq m_1$ then Q_1 is always empty. And the CTS time of A_2 can be computed by equation (2.15). Therefore

$$t_2^* = \frac{H(1-\eta)b}{(1-H)(C_2 - m_2)}.$$

On the other hand

$$A_1(t) + A_2(t) = (m_1 + m_2)t + Z_H(t)$$

is the aggregate process of a self-similar traffic. Therefore

$$t^* = \frac{Hb}{(1-H)(C_1 + C_2 - m_1 - m_2)}.$$

It is clear that for efficient de-multiplexing $C_1 = m_1$ because it is the minimum value for C_1 for which Q_1 is stable and also empty. Also note that Q_1 is zero because that queue is always empty. This means that $\eta = 0$. Then we have

$$\begin{aligned} t^* &= \frac{Hb}{(1-H)(C_1 + C_2 - m_1 - m_2)} \\ &= \frac{Hb}{(1-H)(C_2 - m_2)} \\ &= \frac{H(1-\eta)b}{(1-H)(C_2 - m_2)} \\ &= t_2^*. \end{aligned}$$

This is the condition that theorem 2.3 for efficient de-multiplexing implies.

Example 2: Suppose the $A_1(t)$ and $A_2(t)$ are two independent identical self-similar process. It is clear, in this case, that efficient de-multiplexing occurs for

$C_1 = C_2 = \frac{C}{2}$ and $\eta = .5$. Then

$$\begin{aligned} t_2^* = t_1^* &= \frac{H\eta b}{(1-H)(C_1 - m_1)} \\ &= \frac{.5Hb}{.5(1-H)(C - m)} \\ &= \frac{Hb}{(1-H)(C - m)} \\ &= t^*. \end{aligned}$$

This is the condition that theorem 2.3 implies for efficient de-multiplexing.

3.3 De-multiplexing for the self-similar burst model

In the self-similar burst model, the aggregate traffic is $A(t) = mt + \sigma_1 B_H(t) + \sigma_2 L_H(t)$. Here, H is the self-similar parameter and $A(t)$ is the superposition of two traffics: one of them is fGn traffic with a mean value m_1 and the other one is sLn with another mean value m_2 . In other words

$$A(t) = A_1(t) + A_2(t) = m_1 t + \sigma_1 B_H(t) + m_2 t + \sigma_2 L_H(t) \quad (3.7)$$

By using the de-multiplexing method the queueing behavior for the self-similar burst model can be analyzed.

3.3.1 Lower bound

By the proposition 3.2 the self-similar burst model has the following lower bound

$$P[Q > b] \geq \max\{P[W^{\sigma_1 B_H(t) + m_1 t, C} > b], P[W^{\sigma_2 L_H(t) + m_2 t, C} > b]\} \quad (3.8)$$

The simulation results of the queueing behavior for the self-similar burst model and each part of the traffic are shown in figure (3.2). As the figure shows, the lower bound (equation (3.8)) can give a good approximation for $P[Q > b]$ for small as well as for

large values of b . This makes sense because for small values of b the large arrivals of fGn process cause the buffer overflow, and when b is too large only the huge bursts in the sLn process cause the buffer overflows. The following lemma determines the lower bounds of queue tail of self similar burst model.

Lemma 3.5 If the aggregate input traffic of a queue defined as (3.7) then queue tail has the following lower bounds.

$$P[Q > b] \geq \exp(-\gamma'_1 b^{2(1-H)}) \quad (3.9)$$

and for very large values of b

$$P[Q > b] \geq M_\alpha \gamma'_2 b^{-\alpha+1} \quad (3.10)$$

where α , $-\gamma'_1$, γ'_2 and M_α are computed as (2.17) and (2.18).

3.3.2 Upper bound

In the self-similar burst model the queues Q_1 and Q_2 have both self-similar inputs. So each one has only one critical time scale (CTS) which can be computed in terms of the queue parameters. Let t_1^* and t_2^* be the CTS of the queues. By equation (2.15),

$$t_1^* = \frac{H}{1-H} \frac{\eta b}{C_1 - m_1}, \quad (3.11)$$

$$t_2^* = \frac{H}{1-H} \frac{(1-\eta)b}{C_2 - m_2}. \quad (3.12)$$

On the other hand by theorem 3.1, in the efficient de-multiplexing case

$$t^* = t_1^* = t_2^*$$

So by equations (2.15), (3.11) and (3.12)

$$t^* = \frac{H}{1-H} \frac{b}{C - m} = \frac{H}{1-H} \frac{\eta b}{C_1 - m_1} = \frac{H}{1-H} \frac{(1-\eta)b}{C_2 - m_2}. \quad (3.13)$$

So we can find the de-multiplexing queue service rate in terms of η and free capacity of the single queue, via

$$C_1 - m_1 = \eta(C - m), \quad (3.14)$$

$$C_2 - m_2 = (1 - \eta)(C - m). \quad (3.15)$$

These equations also show that the free capacity of the single queue is split between de-multiplexed queues in proportion η and $1 - \eta$ between the two queues. If the asymptotic lower bound of $P[Q_1 > \eta b]$ and $P[Q_2 > (1 - \eta)b]$ is written in terms of η , then

$$\begin{aligned} P[Q_1 > \eta b] &\geq P[A_1(t_1^*) - C_1(t_1^*) > \eta b] \\ &= P[\sigma_1 B_H(1) > \frac{\eta b + (C_1 - m_1)t_1^*}{t_1^{*H}}] \\ &= P[\sigma_1 B_H(1) > \frac{(C_1 - m_1)^H (\eta b)^{1-H}}{H^H (1 - H)^{1-H}}] \\ &= P[\sigma_1 B_H(1) > \eta \frac{(C - m)^H b^{1-H}}{H^H (1 - H)^{1-H}}] \\ &= P[\sigma_1 B_H(1) > \eta \kappa b^{1-H}] \end{aligned}$$

so for the self-similar traffics $A_1(t)$ and $A_2(t)$ and $A_1(t) + A_2(t)$ we have

$$P[Q_1 > \eta b] \stackrel{\log}{\asymp} P[\sigma_1 B_H(1) > \eta \kappa b^{1-H}] \quad (3.16)$$

$$P[Q_2 > (1 - \eta)b] \asymp P[\sigma_2 L_H(1) > (1 - \eta)\kappa b^{1-H}] \quad (3.17)$$

$$P[Q > b] \asymp P[\sigma_1 B_H(1) + \sigma_2 L_H(1) > \kappa b^{1-H}] \quad (3.18)$$

The equations (3.16) and (3.17) imply

$$P[Q_1 > \eta b] + P[Q_2 > (1 - \eta)b] \stackrel{\log}{\asymp} \phi(\eta) \quad (3.19)$$

where

$$\phi(\eta) = P[\sigma_1 B_H(1) > \eta \kappa b^{1-H}] + P[\sigma_2 L_H(1) > (1 - \eta)\kappa b^{1-H}] \quad (3.20)$$

On the other hand the following equation is trivial for any $\eta \in (0, 1)$:

$$\phi(\eta) \geq P[\sigma_1 B_H(1) + \sigma_2 L_H(1) > \kappa b^{1-H}] \quad (3.21)$$

So the equations (3.18), (3.19) and (3.21) give us that the asymptotic lower bounds of the de-multiplexed queues is an upper bound for the asymptotic lower bound of the single queue. It makes sense because the inequality (3.4) for $\eta \in (0, 1)$ tells us that the superposition of the de-multiplexed queues gives an upper bound for the single queue. Now, by using the lower bound asymptotics of the de-multiplexed queues, we calculate the $\eta \in (0, 1)$ which gives the minimum asymptotic upper bound. As explained in section 2.4 the density function of $Z_H(1)$ is the convolution of the density functions of a Gaussian and a stable random variable. Figure (3.3) shows that for the small values the density function of $Z_H(1)$ is more like the Gaussian density function and for large values it is more like the stable density function. Thus, it has a tail which is the same as the one of the stable density function.

The value of η which minimizes the function $\phi(\eta)$ can be found by numerical methods for each $b > 0$. By the equation (3.16) the value of $P[\sigma_1 B_H(1) > \eta \kappa b^{1-H}]$ decreases as a Weibull function when b increases. However, equation (3.17) shows that the value of $P[\sigma_2 L_H(1) > (1 - \eta) \kappa b^{1-H}]$ decreases as a power-law function. Therefore for large values of b , η is small positive number. So

$$\begin{aligned} \phi(\eta) &\asymp P[\sigma_2 L_H(1) > (1 - \eta) \kappa b^{1-H}] \\ &\asymp P[\sigma_2 L_H(1) > \kappa b^{1-H}] \\ &\asymp P[Z_H(1) > \kappa b^{1-H}] \end{aligned}$$

So when $b \rightarrow \infty$ the asymptotic upper bound and asymptotic lower bound of $P[Q > b]$ have the same power-law exponent. So for very large values of b

Theorem 3.2 Assume that (3.7) holds then queue tail for large buffer sizes is a power-law function.

$$P[Q > b] \asymp M_\alpha \gamma'_2 \cdot b^{-\alpha+1} \quad b \rightarrow \infty \quad (3.22)$$

Therefore the tail of the queue in the self-similar burst model has a power-law decay with exponent $-\alpha + 1$. It means that only the alpha component affects the queue overflowing probability for large buffer sizes.

3.4 Summary of the self-similar burst model

In the self-similar burst model the alpha component is modelled by a stable Levy noise (sLn) process and the beta component is modelled by a fractional Gaussian noise (fGn) process. Since the sum of the alpha and beta component is assumed to be self-similar as well, the two independent components must have the same self-similarity parameters.

The queueing analysis of the self-similar burst model shows that the buffer overflow probability is approximately a Weibull function for small buffer sizes and a power-law function for large buffer sizes. So, the asymptotic queueing behavior of the queue tail for the self-similar burst model is a power-law.

Also, it is possible to de-multiplex the flow into two flows feeding into two buffers. This allows to for proposes of compute the upper bound of queue tail. When the input traffic obeys the self-similar burst model, the parameters of the de-multiplexed queues are chosen such they have the same critical time scale (CTS) and efficient de-multiplexing. By using the de-multiplexing method we showed that for very large buffer sizes, only the alpha-traffic causes the buffer overflows.

Chapter 4

An ON/OFF burst model

The alpha component which generates the bursts in the small time scales could be considered as the ON periods of a high rate ON/OFF source. In this model the alpha component is modelled by an ON/OFF source (not necessarily a renewal ON/OFF source). By the connection-level analysis, when the highest bandwidth connection is sending or receiving a huge file it increases the average traffic for duration of the file transfer. So, it generates the traffic bursts in the large time scales. The rate of such a connection has been observed to be high and is almost constant [WSRB02, SRB02]. Therefore, the highest bandwidth connection would be modelled by a high rate ON/OFF source. In the ON/OFF burst model the aggregate process is

$$A(t) = A_1(t) + A_2(t) = mt + \sigma B_H(t) + \int_0^t r_{\text{on}} \cdot X(s) ds \quad (4.1)$$

where $A_1(t)$ is the aggregate process of the beta component and $A_2(t)$ is the aggregate process of a high rate ON/OFF source that models the alpha component. The $R(s) = r_{\text{on}} \cdot X(s)$ represents a high rate ON/OFF source, i.e.,

$$R(s) = \begin{cases} r_{\text{on}} & \text{if } s \text{ lies in an ON period} \\ 0 & \text{if } s \text{ lies in an OFF period} \end{cases} \quad (4.2)$$

Figure (4.1) depicts a simulation result for the ON/OFF burst model. As the figure shows, the ON periods of the high rate ON/OFF appear as traffic bursts in the large time scales.

When the alpha traffic is modelled by an ON/OFF source, it has a different effect on the network queue than the previously discussed self-similar burst model. In the

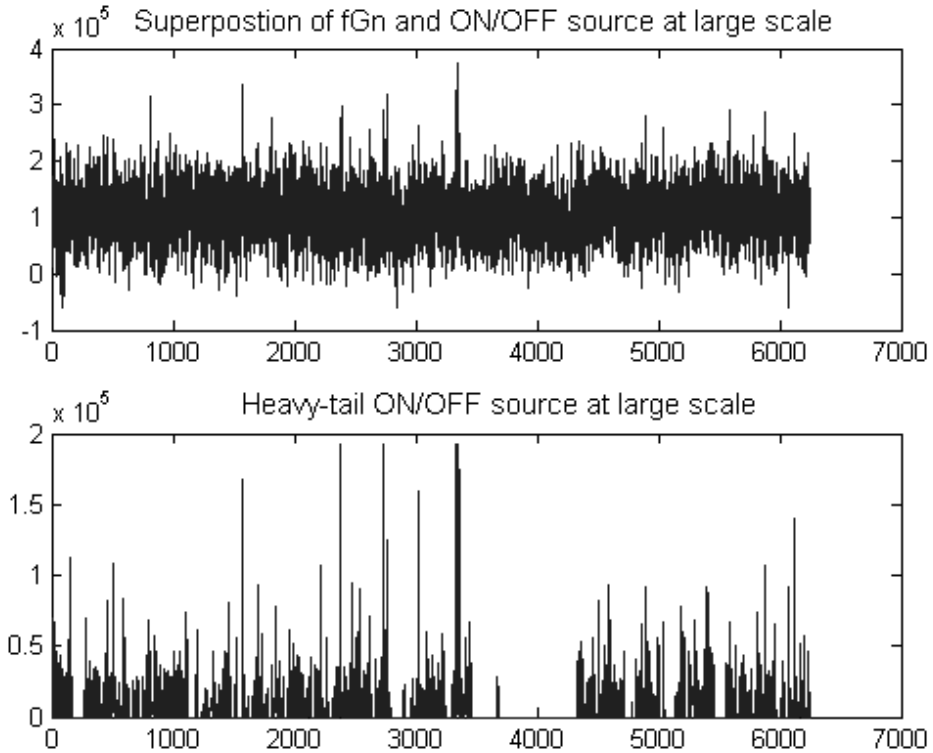


Figure 4.1 ON/OFF burst model in the large scale limit

network queue, when the ON/OFF source along with the background traffic feeds into the queue, the ON/OFF source reduces the queue capacity during the ON periods. Therefore, the queue behaves as a variable service rate queue, which decreases the service rate during the ON periods to the ON/OFF source. Figure (1.2) shows how a variable service rate queue models effect of an ON/OFF source in a constant service rate queue. So, we may consider a queue with variable capacity $C(s) = C - R(s)$:

$$C(s) = \begin{cases} C - r_{\text{on}} & \text{if } s \text{ lies in an ON period} \\ C & \text{if } s \text{ lies in an OFF period} \end{cases} \quad (4.3)$$

In this chapter we first prepare the ground to study the queueing behavior of a variable service rate queue for general models by considering simple cases. Next we

determine the queueing behavior of the particular case where the alpha component of the traffic is modelled by a high rate ON/OFF source and the beta component is modelled by the fGn traffic. Then, we analyze the queueing behavior for the ON/OFF burst model in the different cases. Finally we compare queueing behavior of those cases of the ON/OFF burst model.

4.1 Markov service rate queue model

In this section we use the idea of [WSRB02, SRB02], which was explained in section 2.3. We model the occurrence of the bursty time bins (which have alpha traffic) and the non-bursty time bins (which have only beta traffic) as a Markov chain. So the alpha traffic is modelled by a high rate ON/OFF source, where the ON periods correspond to the bursty time bins.

In the Markov chain model, time bins have two states. A time bin is in the beta-state when there is no alpha traffic in the time bin, and it is in the alpha-state when there is alpha traffic in the time bin. When the superposition of the alpha and beta traffics is feed into a queue, the queue can be considered as a variable service rate queue with a decreased service rate during the alpha-state time periods.

Let C_1 and C_2 represent the service rates of the queue in the beta-state and the alpha-state time bins ($C_1 \geq C_2$). So the service rates of the queue (C_1 and C_2) change as a Markov chain in the different time bins. We analyze the queueing behavior of this model for some input traffic models. Let $P_{\alpha\alpha}$, $P_{\alpha\beta}$, $P_{\beta\alpha}$ and $P_{\beta\beta}$ denote the transition probabilities of the Markov Chain. Also, let T represent the size of time bins and $Q(t)$ represent the size of the queue at time t .

4.1.1 Markov service rate queue model with CBR input

For the queueing analysis of the Markov service rate queue, we first analyze the Constant Bit Rate (CBR) traffic as the input traffic of the queue. Let R be the rate of CBR traffic. When $R < C_2$, the queue can forward input packets as soon as they come inside the queue. This means that even in the alpha-state time bins, the queue has the ability to serve the packets that come into the queue as soon as they arrive. So in this case the queue size is zero ($Q(t) = 0$).

However when the input rate of queue is larger than the service rate during the alpha-state time bins, the queue size increases in time in the alpha-state time bins. The stability of the queue implies that $R < C$, where C is average service rate [Pra97, CY01]. In other words:

$$R < C = C_1P_\beta + C_2P_\alpha \quad (4.4)$$

where P_α and P_β represent the probabilities that a time bin is in the alpha-state or in the beta-state. In this case the queue size increases with the rate $R - C_2$ in the alpha-state time bins and it decreases with the rate $C_1 - R$ in the beta-state time bins (when queue is not empty). Figure (4.2) shows that the variation of the queue size in different time bins. It has linear increase and linear decrease. In Theorem 4.1 (which will be explained later) we will show, when $C_2 < R < C$ the queue size has a short-tailed (exponential) distribution function, which can be computed in terms of the queue parameters. Before stating the theorem we will state Lemmas 4.1 and 4.2 for the Markov service rate queue model.

Lemma 4.1 Let C_1 and C_2 be the service rates of the Markov service rate queue and $R(t)$ be the rate of the input traffic at time t . Assume

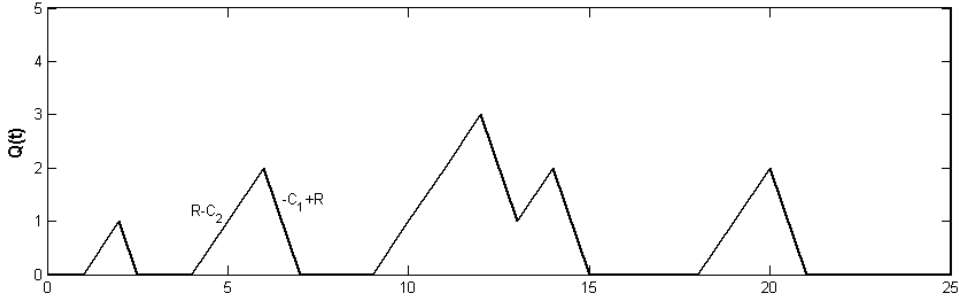


Figure 4.2 Discretized queue and CBR traffic

that (4.4) holds and $\lim_{t \rightarrow \infty} \frac{R(t)}{t} = R$, then

$$\lim_{t \rightarrow \infty} \frac{Q(t)}{t} = 0$$

Proof of lemma 4.1 : If $\lim_{t \rightarrow \infty} \frac{Q(t)}{t} \neq 0$ then there exists $\varepsilon > 0$ and infinite discrete time instance t_k such that $\frac{Q(t_k)}{t_k} > \varepsilon$ for $k = 1, 2, 3, \dots$. It implies that the sequence $\{Q(t_k)\}_{k=0}^{\infty}$ is an unbounded sequence. So there exists a subsequence $\{l_k\}_{k=0}^{\infty}$ of the sequence $\{t_k\}_{k=0}^{\infty}$ such that for each positive integer number $n : Q(l_n) > Q(t_k)$ for all $0 \leq t_k < l_n$.

Consider a positive integer number n that $l_n \gg \max(T, R/\varepsilon)$, where T is the size of time bin. The Lindley's equation [Pra97, CY01] for the discrete queue gives the recursive formula of the queue size,

$$Q(t) = \max(Q(t-1) + R(t) - C(t), 0). \quad (4.5)$$

If we expand this formula from $t-1$ to $t-k$, then

$$Q(t) = \max\left\{Q(t-k) + \sum_{\tau=t-k+1}^t R(\tau) - C(\tau), \sum_{\tau=t-k+2}^t R(\tau) - C(\tau), \dots, R(t) - C(t), 0\right\}. \quad (4.6)$$

The equation (4.6) for $t = l_n$ and $t - k = l'_n$ implies the following inequality

$$\frac{Q(l'_n)}{l_n} + \frac{\sum_{t=l'_n+1}^{l_n} R(t) - C(t)}{l_n} \leq \frac{Q(l_n)}{l_n} \quad (4.7)$$

where $l'_n = l_n(1 - \varepsilon/R)$. But the maximum increase of the queue size from the time $t = l'_n$ to the time $t = l_n$ is equal to $\sum_{t=l'_n+1}^{l_n} R(t) - C_2(l_n - l'_n)$. By hypothesis of lemma, $\lim_{t \rightarrow \infty} \frac{R(t)}{t} = R$. So for large enough n

$$\frac{\sum_{t=l'_n+1}^{l_n} R(t) - C_2}{l_n} \simeq (\varepsilon/R)(R - C_2) < \varepsilon$$

This inequality shows that the queue size cannot be zero at the time $l'_n < t < l_n$. So by the equation (4.6) at the times $t = l'_n$ and $t = l_n$ we have:

$$\begin{aligned} Q(l_n) &= \max\left\{Q(l'_n) + \sum_{\tau=l'_n+1}^t R(\tau) - C(\tau), \sum_{\tau=l_n+2}^t R(\tau) - C(\tau), \right. \\ &\quad \left. \sum_{\tau=l_n+3}^t R(\tau) - C(\tau), \dots, R(t) - C(t), 0\right\} \\ &= Q(l'_n) + \sum_{\tau=l'_n+1}^t R(\tau) - C(\tau) \end{aligned}$$

In other words

$$\frac{Q(l'_n)}{l_n} + \frac{\sum_{t=l'_n+1}^{l_n} R(t) - C(t)}{l_n} = \frac{Q(l_n)}{l_n}$$

Also the Markov chain property implies for enough large n 's:

$$\begin{aligned} \frac{\sum_{t=l'_n+1}^{l_n} C_t(l_n - l'_n)}{l_n} &= C \frac{l_n - l'_n}{l_n} \\ &= C(\varepsilon/R), \end{aligned}$$

where C is the average queue service rate. So

$$\begin{aligned} \frac{Q(l'_n)}{l_n} + \frac{\sum_{t=l'_n+1}^{l_n} R(t) - C(t)}{l_n} &\simeq \frac{Q(l'_n)}{l_n} + (\varepsilon/R)(R - C) \\ &= \frac{Q(l_n)}{l_n}. \end{aligned}$$

But by the hypothesis of Lemma $C > R$ so $\frac{Q(l'_n)}{l'_n} > \frac{Q(l_n)}{l_n}$. On the other hand, the property of $\{l_k\}_1^\infty$ sequence implies, $Q(l_n) > Q(t)$ for $0 < t < l_n$. It shows there is no $\varepsilon > 0$ and infinite points in the discrete time $\frac{Q(t_k)}{t_k}$ with those conditions. This means $\lim_{t \rightarrow \infty} \frac{Q(t)}{t} = 0$.

Lemma 4.2 Let C_1 and C_2 be the service rates of the Markov service rate queue and R be the rate of the CBR input. If (4.4) holds and $R > C_2$, then

$$P[Q > 0] = \frac{C_1 - C_2}{C_1 - R} P_\alpha$$

Proof of lemma 4.2 : By the Lemma (4.1) $\lim_{t \rightarrow \infty} \frac{Q(t)}{t} = 0$. This expression can be written in the integral form: $\lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t dQ(\tau) = 0$. For the CBR input $dQ(t)$ can have three values, Case(1): when t is in the alpha-state time bin, the queue size increases at the time t and $dQ(t) = R - C_2$. Case(2): when t is in the beta-state time bin and $Q(t) > 0$, the queue size decreases at the time t and $dQ = R - C_1$. Case(3): when t is in the beta-state time bin and $Q(t) = 0$ then Queue size does not change at time t and $dQ(t) = 0$. So the integral expression can be written as:

$$\lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t dQ(\tau) = \lim_{t \rightarrow \infty} \frac{1}{t} ((R - C_2)t_{\text{inc}} - (C_1 - R)t_{\text{dec}}) \quad (4.8)$$

where

$$t_{\text{inc}} = \#\{0 < \tau < t : dQ(\tau) > 0\}$$

$$t_{\text{dec}} = \#\{0 < \tau < t : dQ(\tau) < 0\}.$$

Let $P(Q \uparrow)$ and $P(Q \downarrow)$ be probabilities that queue size is increasing; respectively decreasing at the time τ . Then we have

$$\lim_{t \rightarrow \infty} \frac{t_{\text{inc}}}{t} = P(Q \uparrow) \quad (4.9)$$

$$\lim_{t \rightarrow \infty} \frac{t_{\text{dec}}}{t} = P(Q \downarrow) \quad (4.10)$$

Therefore $\lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t dQ(\tau) = (R - C_2)P(Q \uparrow) - (C_1 - R)P(Q \downarrow) = 0$. By this equation $P(Q \downarrow)$ can be computed in terms of $P(Q \uparrow)$

$$P(Q \downarrow) = \frac{R - C_2}{C_1 - R} P(Q \uparrow) \quad (4.11)$$

On the other hand, the queue size increases only in the alpha-state time bins. Therefore they all have the same probability, i.e.,

$$P(Q \uparrow) = P_\alpha \quad (4.12)$$

By the equations (4.11) and (4.12)

$$\begin{aligned} P(Q > 0) = P(Q \uparrow) + P(Q \downarrow) &= P(Q \uparrow) \left(1 + \frac{R - C_2}{C_1 - R}\right) \\ &= P_\alpha \frac{C_1 - R + R - C_2}{C_1 - R} \\ &= \frac{C_1 - C_2}{C_1 - R} P_\alpha \end{aligned}$$

Theorem 4.1 Let C_1 and C_2 be the service rates of a Markov service rate queue and R be the rate of CBR input. If (4.4) holds and $R > C_2$, then the queue size has a short-tailed (sum of exponentials) distribution function. Also $P[Q > b]$ can be computed explicitly in terms of the queue parameters.

Proof of the theorem 4.1 : For proving the theorem we discretize the queue size and time by the following method. First assume $\frac{R - C_2}{C_1 - R} = \frac{k_2}{k_1}$, where k_1 and k_2 are positive integers which do not have a common divisor greater than 1. By the assumptions the queue size increases by $(R - C_2)T$ in the alpha-state time bins, and it decreases by $(C_1 - R)T$ in the beta-state time bins when $Q(t) \geq (C_1 - R)T$. This fact shows that if we consider the queue size at the discrete times $t = 0, T, 2T, 3T, \dots$ the queue size

is $(R - C_2)T.X - (C_1 - R)T.Y$ where X and Y are some integers. Also we define the parameter

$$D = \frac{(R - C_2)T}{k_2} = \frac{(C_1 - R)T}{k_1}. \quad (4.13)$$

By the definition of D , the queue size at the times $t = 0, T, 2T, \dots$ is an integer multiple of D ($Q(0)=0$, because the queue is empty at the time $t = 0$). In other words for arbitrary integers X and Y there exists an integer Z such that

$$(R - C_2)T.X - (C_1 - R)T.Y = D.Z$$

Therefore the queue size at times $t = 0, T, 2T, \dots$ takes the values from the set $\{0, D, 2D, 3D, \dots\}$. We define the probabilities α_k and β_k by the set $A = \{0, T, 2T, \dots\}$

$$\alpha_k = P[Q(t) = kD | t \in A, (t, t + T) : \text{alpha-state}] \quad (4.14)$$

$$\beta_k = P[Q(t) = kD | t \in A, (t, t + T) : \text{beta-state}] \quad (4.15)$$

The recursive formula for the α_k and β_k can be found in terms of the transition probabilities of the Markov chain. Figure (4.3) shows how α_k and β_k are related to each other by the transition probabilities of the Markov Chain. So the recursive formulas of α_k and β_k are

$$\alpha_n = P_{\alpha\alpha}\alpha_{n-k_2} + P_{\beta\alpha}\beta_{n+k_1}, \quad (4.16)$$

$$\beta_n = P_{\alpha\beta}\alpha_{n-k_2} + P_{\beta\beta}\beta_{n+k_1}. \quad (4.17)$$

The initial conditions for the recursion are

i) $\alpha(n) = 0$ for $n < 0$

ii) $P_\alpha = \alpha_0 + \alpha_1 + \alpha_2 + \dots$

The recursive formula of each parameter α_n and β_n can be found in this way:

$$\beta_{n+k_1} = \frac{\alpha_n - P_{\alpha\alpha}\alpha_{n-k_2}}{P_{\beta\alpha}}$$

so if we write β_{n+k_1} and β_n in terms of α_n and put it in the

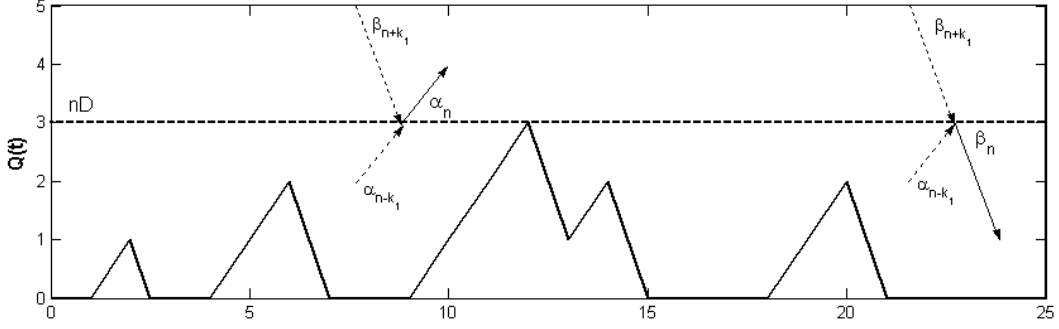


Figure 4.3 Discretized queue and CBR traffic

second recursive formula, then

$$P_{\beta\beta}\alpha_n + (P_{\beta\alpha}P_{\alpha\beta} - P_{\alpha\alpha}P_{\beta\beta})\alpha_{n-k_2} + P_{\alpha\alpha}\alpha_{n-k_2-k_1} - \alpha_{n-k_1} = 0. \quad (4.18)$$

By the same method we can show that

$$P_{\beta\beta}\beta_{n+k_1+k_2} + (P_{\beta\alpha}P_{\alpha\beta} - P_{\alpha\alpha}P_{\beta\beta})\beta_{n+k_1} + P_{\alpha\alpha}\beta_n - \beta_{n+k_2} = 0. \quad (4.19)$$

The linear recursive sequence theorems imply that α_n and β_n can be written in the form:

$$\alpha_n = a_1 r_1^n + a_2 r_2^n + \dots + a_{k_1+k_2} r_{k_1+k_2}^n,$$

$$\beta_n = b_1 r_1^n + b_2 r_2^n + \dots + b_{k_1+k_2} r_{k_1+k_2}^n,$$

where a_k and b_k are the constant coefficients which depend on the initial conditions of the recursion. The r_k 's are the roots of this equation (the polynomial is obtained by replacing α_n in the recursive formula (4.18) by X^n , where X is the variable). Thus,

$$f(X) = P_{\beta\beta}X^{k_1+k_2} + (P_{\beta\alpha}P_{\alpha\beta} - P_{\alpha\alpha}P_{\beta\beta})X^{k_1} - X^{k_2} + P_{\alpha\alpha} = 0 \quad (4.20)$$

As we know, $\alpha_n \rightarrow 0$ and $\beta_n \rightarrow 0$ when $n \rightarrow \infty$. So the coefficients of the roots r_k , with $|r_k| \geq 1$ must be zero ($a_k = b_k = 0$). So only the roots of the polynomial

$f(X)$ which are inside the unit circle of the complex plane should be considered in the expansion of α_n and β_n .

The queue size discrete distribution function can be computed in terms of α_n and β_n , the in this way:

$$P(nD \leq Q < (n+1)D) = \frac{\alpha_n + \alpha_{n-1} + \dots + \alpha_{n-k_2+1}}{k_2} + \frac{\beta_{n+1} + \beta_{n+2} + \dots + \beta_{n+k_1}}{k_1} \quad (4.21)$$

So $P(Q \geq nD) = \sum_{k=n}^{\infty} P(kD \leq Q < (k+1)D)$ is a linear combination of $r_1^n, r_2^n, \dots, r_{k_1+k_2}^n$ can be written as

$$P(Q \geq nD) = q_1 r_1^n + q_2 r_2^n + \dots + q_{k_1+k_2} r_{k_1+k_2}^n \quad (4.22)$$

Thus, $P[Q > b]$ is sum of some exponential functions with negative exponents since only roots within the unit circle have non-zero coefficients. So, the queue size has a short-tailed distribution function (which means $\lim_{x \rightarrow \infty} \frac{P[Q > x+a]}{P[Q > x]} \neq 1$).

Remark 1: When r is a root of $f(X)$ with multiplicity $m > 1$ then $r^n, nr^n, \dots, n^{m-1}r^n$ appear in the expansion formula of α_n and β_n instead of m roots. Again for this case $P(Q > n.D)$ decreases as $q_1 r^n + q_2 n r^n + \dots + q_m n^{m-1} r^n$ for $|r| < 1$ and large n . This shows that the queue size of this model always has a short-tailed distribution function.

Remark 2: When there are many states (more than 2) and the Markov chain has higher order, the parameters D and $\alpha_k, \beta_k, \gamma_k, \delta_k, \dots$ could be defined in the similar way. By similar arguments it follows that the queue has a short-tailed distribution function.

Remark 3: The function $f(X)$ has some useful properties which help determine the locations of the roots in the complex plane:

- i) $f(1) = 0$
- ii) $f(0) = P_{\alpha\alpha}$

iii)

$$\begin{aligned}
f'(1) &= (k_1 + k_2)P_{\beta\beta} + k_1(P_{\beta\alpha}P_{\alpha\beta} - P_{\alpha\alpha}P_{\beta\beta}) - k_2 \\
&= -(k_1 + k_2)P_{\beta\alpha} + k_1(P_{\beta\alpha} + P_{\alpha\beta}) \\
&= k_1(P_{\beta\alpha} + P_{\alpha\beta})(1 - P(Q > 0)) > 0
\end{aligned}$$

These equations show that $f(X)$ has at least one root in the interval $(0, 1)$. For some cases, this is the only root of polynomial $f(X)$ in the unit circle of the complex plane. For those cases, let r represent the root of $f(X)$ in the interval $(0, 1)$. The initial conditions imply then that $\alpha_n = P_\alpha(1 - r)r^n$ and $\beta_n = \frac{R - C_2}{C_1 - R}P_\alpha(1 - r)r^n$.

So by the Theorem 4.2, the coefficient r could be found in this case as:

$$P(Q > n.D) = P(Q > 0).r^n = P(Q > 0).z^{n.D} \quad (4.23)$$

where $z = r^{1/D}$ and for $b > 0$,

$$P(Q > b) = P(Q > 0).z^b = P_\alpha \frac{C_1 - C_2}{C_1 - R}.z^b \quad (4.24)$$

We present such a special case.

Corollary 4.1 Let C_1 and C_2 be the service rates of the Markov service rate queue and let R be rate of the CBR input. If (4.4) holds and $R = \frac{C_1 + C_2}{2}$ then

$$P(Q > b) = 2P_\alpha \left(\frac{P_{\alpha\alpha}}{P_{\beta\beta}} \right)^{2b/(C_1 - C_2)T}$$

Proof of Corollary 4.1 : By the Theorem 4.1 in this case $k_1 = k_2 = 1$ and $D = T(C_1 - C_2)/2$. So

$$\begin{aligned}
f(X) &= P_{\beta\beta}X^2 + (P_{\beta\alpha}P_{\alpha\beta} - P_{\alpha\alpha}P_{\beta\beta} - 1)X + P_{\alpha\alpha} \\
&= P_{\beta\beta}X^2 - (P_{\alpha\alpha} + P_{\beta\beta})X + P_{\alpha\alpha}
\end{aligned}$$

The roots of $f(X)$ are equal to 1 and $\frac{P_{\alpha\alpha}}{P_{\beta\beta}}$. The condition $P(Q > 0) < 1$ implies that $\frac{P_{\alpha\alpha}}{P_{\beta\beta}} < 1$. This is the root of $f(X)$ in the interval $(0, 1)$ and this case it is the only root of $f(X)$ that is inside the unit circle of the complex plane. So by Remark (2)

$$\begin{aligned} P(Q > b) &= P(Q > 0) \cdot z^b \\ &= P_{\alpha} \frac{C_1 - C_2}{C_1 - R} \cdot z^b \end{aligned}$$

where $z = \left(\frac{P_{\alpha\alpha}}{P_{\beta\beta}}\right)^{1/D}$, and $\frac{C_1 - C_2}{C_1 - R} = 2$. Therefore

$$P(Q > b) = 2P_{\alpha} \cdot \left(\frac{P_{\alpha\alpha}}{P_{\beta\beta}}\right)^{2b/(C_1 - C_2)T}$$

4.1.2 Lower bound for the buffer overflow probability with CBR input

In this section we give a formula for the Markov service rate queue, which provides a lower bound for $P[Q > b]$ in terms of the service rates of the queue and the transition probabilities of the Markov chain. It is possible to use theorem 4.1 to find the queue size distribution function, but depending on the values of k_1 and k_2 it may need a lot of computations. This lower bound gives an exponential lower bound that depends on the queue parameters in some cases is close to the distribution function of queue size.

Theorem 4.2 Let C_1 and C_2 be service rates of a Markov service rate queue and R be rate of CBR input. If (4.4) holds and $R > C_2$, then

$$P[Q > b] \geq P_{\alpha} \frac{C_1 - C_2}{C_1 - R} \cdot P_{\alpha\alpha}^{b/T(R - C_2)}$$

Proof of theorem 4.2 : We define the consecutive alpha-state time bins as an alpha-period. The size of the alpha-period is a random variable. It can be one time bin

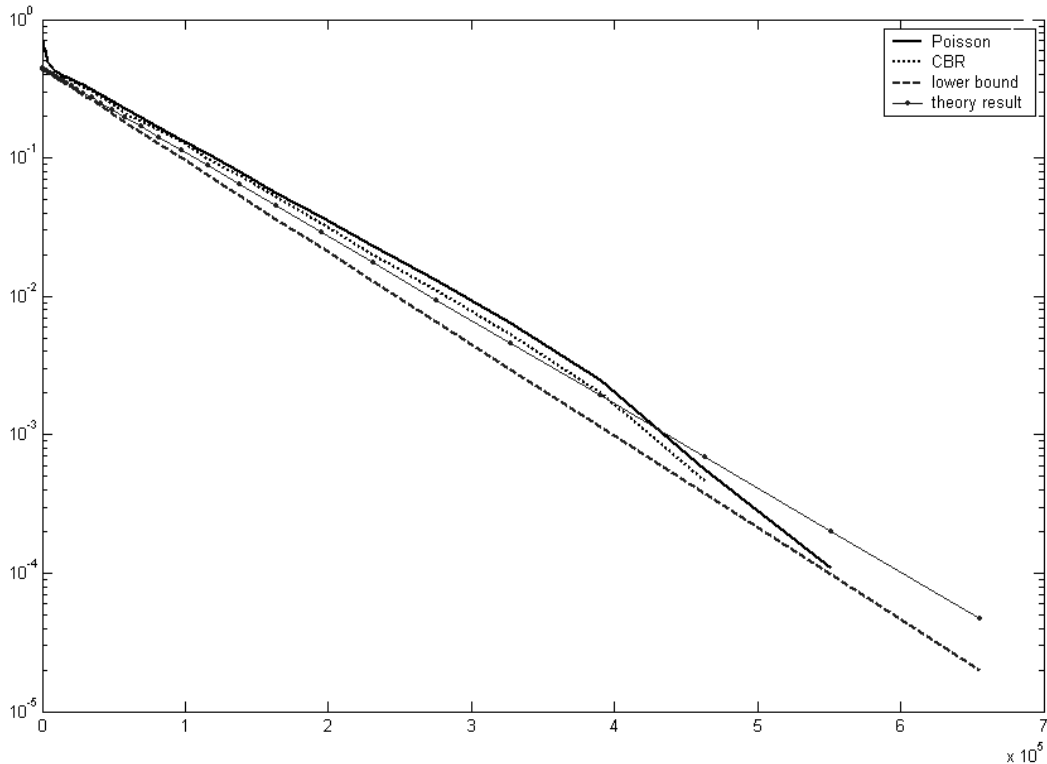


Figure 4.4 Comparison of CBR and Poisson traffic in a queue with Markov service rate. For small variance of the Poisson traffic the two queues behave almost the same

or more than one time bin. The size of the alpha-period is a random variable in terms of the number of time bins. To approach the problem at hand we introduce an approximation Q_{LS} to the true queue Q which provides always a lower bound. To this end, we define Q_{LS} to be the queue length of a queue with the same input as the given Q , however, with a different policy: The queue Q_{LS} will disregard and drop any packet of an alpha source if there are still packets in the queue which entered during a previous alpha period. In other words, the queue Q_{LS} will first relax and empty completely before accepting new alpha packets.

Clearly, $Q_{LS}(t)$ is always smaller than $Q(t)$ since it accepts fewer packets. On the other hand, Q_{LS} provides a good approximation to Q if the queue will relax after each alpha period with high probability. A first such case is found when $P_{\beta\beta} \simeq 1$. Here, the beta-periods are long and thus, two consecutive alpha periods are far from each other. Also in the case that $C_1 - R \gg R - C_2$, the consecutive alpha-periods do not effect each other and we have $Q_{LS}(t) \simeq Q(t)$.

The distribution function of Q_{LS} could be found in terms of the Markov service rate queue in the following way

$$\begin{aligned} P[Q_{LS} > b] &= \lim_{M \rightarrow \infty} \frac{1}{MT} \sum_{k=1}^{\infty} M.P[\text{period} = kT].N(k, b) \\ &= \sum_{k=1}^{\infty} P[\text{period} = kT].N(k, b)/T \\ &= \sum_{k=b'}^{\infty} P_{\beta}P_{\beta\alpha}P_{\alpha\alpha}^{k-1}P_{\alpha\beta}(N(k, b)/T) \end{aligned}$$

Here, $N(b, k)$ denotes the number of consecutive time bins during which the queue size exceeds b during and after an alpha period of k time bins. In other words, N measures how long it takes the queue to relax back to the level b once it exceeded the level b due to an alpha connection. Note that $N(k, b) = 0$ for $k < b'$, where $b'T$ is the minimum period size during which the queue size could be larger than b in some period of the time ($b' = \frac{b}{(R-C_2)T}$). Also we can show $\frac{N(k, b)}{T} = \frac{C_1 - C_2}{C_1 - R}(k - b')$.

$$\begin{aligned} P[Q_{LS} > b] &= \sum_{k=b'}^{\infty} P_{\beta}P_{\beta\alpha}P_{\alpha\alpha}^{k-1}P_{\alpha\beta} \frac{C_1 - C_2}{C_1 - R}(k - b') \\ &= P_{\beta}P_{\beta\alpha}P_{\alpha\beta} \frac{C_1 - C_2}{C_1 - R} \sum_{k=b'}^{\infty} P_{\alpha\alpha}^{k-1}(k - b') \\ &= P_{\alpha} \frac{C_1 - C_2}{C_1 - R} P_{\alpha\beta}^2 \sum_{k=b'}^{\infty} \{b'.P_{\alpha\alpha}^{b'-1}/(1 - P_{\alpha\alpha}) + P_{\alpha\alpha}^{b'}/(1 - P_{\alpha\alpha})^2 - b'.P_{\alpha\alpha}^{b'-1}/(1 - P_{\alpha\alpha})\} \\ &= P_{\alpha} \frac{C_1 - C_2}{C_1 - R} .P_{\alpha\alpha}^{b'} \\ &= P(Q > 0)P_{\alpha\alpha}^{b/(R-C_2)T} \end{aligned}$$

If the input traffic is CBR or any other traffic with a tail shorter than $P_{\alpha\alpha}^{b/T(R-C_2)}$ the queue size distribution can be approximated by theorems 4.1 and 4.2. Figure (4.4) compare the queue tails of simulated CBR and Poisson traffics with results of theorems 4.1 and 4.2. The figure shows that the queueing behavior of Poisson and CBR traffic is almost the same as in the Markov service rate queue model and it can be estimated by theorem 4.1. In the next section we analyze the queueing behavior of non-CBR traffics in the Markov service rate queue.

4.1.3 Markov service rate queue with non-constant rate input

When the Markov service rate queue is fed by a non-constant rate input, the distribution function of the queue size depends on the queue parameters and the statistical properties of the input process. In this section we compute some bounds for the queue size distribution by using the de-multiplexing method (section 3.2). The theorem 4.3 determines an upper bound and a lower bound for the queue tail of a non-CBR traffic in the variable service rate queue.

Theorem 4.3 Let C_1 and C_2 be service rates of the variable service rate queue and $R(t)$, the input rate. If (4.4) holds and $R > C_2$, then for arbitrary F, η such that $R < F < C$ and $0 < \eta < 1$

$$P[W^{A(t), C_1} > b] \leq P[W^{A(t), (C_1, C_2)} > b] \leq P[W^{A(t), F} > \eta b] + P[W^{Ft, (C_1, C_2)} > (1-\eta)b]$$

where $A(t)$ is the aggregate process of $R(t)$ and $W^{A(t), F}$ as defined in (2.9).

Proof of theorem 4.3 : By the expansion of the Lindley formula (4.6) for the discrete queue (here $C(t)$ is not constant) we find

$$\begin{aligned}
P[W^{A(t),(C_1,C_2)} > b] &= P[\max_{s \geq 0} \sum_{k=0}^s \{R(t-k) - C(t-k)\} > b] \\
&= P[\max_{s \geq 0} \sum_{k=0}^s \{R(t-k) - F + F - C(t-k)\} > b] \\
&\leq P[\max_{s \geq 0} \sum_{k=0}^s \{R(t-k) - F\} + \\
&\quad \max_{s \geq 0} \sum_{k=0}^s \{F - C(t-k)\} > \eta b + (1-\eta)b] \\
&\leq P[\max_{s \geq 0} \sum_{k=0}^s \{R(t-k) - F\} > \eta b] + \\
&\quad P[\max_{s \geq 0} \sum_{k=0}^s \{F - C(t-k)\} > (1-\eta)b] \\
&= P[W^{A(t),F}(t) > \eta b] + P[W^{Ft,(C_1,C_2)}(t) > (1-\eta)b]
\end{aligned}$$

This note that the above inequality is true for all time t . Therefore,

$$P[Q(t) > b] \leq P[W^{A(t),F} > \eta b] + P[W^{Ft,(C_1,C_2)} > (1-\eta)b]$$

Note: This inequality shows that for a Markov service rate queue with non-constant input we can find an upper bound for the overflowing probability, which is sum of two overflow probability, one for the same queue with CBR input traffic and a one for constant service rate queue with the same non-constant rate input traffic.

The lower bound inequality: By the hypothesis that $C(t)$ is equal to C_1 or C_2 at each time, we find $C(t) \leq C_1$. Thus,

$$\begin{aligned}
P[W^{A(t),(C_1,C_2)} > b] &= P[\max_{s \geq 0} \sum_{k=0}^s \{R(t-k) - C(t-k)\} > b] \quad (4.25) \\
&\geq P[\max_{s \geq 0} \sum_{k=0}^s \{R(t-k) - C_1\} > b] \\
&= P[W^{A(t),C_1} > b]
\end{aligned}$$

In the Markov service rate queue, if $P[W^{A(t),C} > b]$ has a longer tailed distribution than the exponential function, then the upper bound and lower bound are asymptotically the same. Therefore the asymptotic of $P[W^{A(t),(C_1,C_2)(t)} > b]$ is determined. For example, when the input process is fGn with $1 > H > .5$ then $A(t)$ is fBm. By the equation (3.9),

$$P[W^{A(t),C_1} > b] \stackrel{\log}{\asymp} \exp(-\gamma_1 b^{2-2H})$$

Also

$$\begin{aligned} P[W^{A(t),F} > \eta b] + P[W^{Ft,(C_1,C_2)} > (1-\eta)b] &\stackrel{\log}{\asymp} \exp(-\gamma'_1 b^{2-2H}) + z^b \\ &\stackrel{\log}{\asymp} \exp(-\gamma'_1 b^{2-2H}) \end{aligned}$$

Therefore

$$P[W^{A(t),(C_1,C_2)(t)} > b] \stackrel{\log}{\asymp} \exp(-b^{2-2H})$$

So the Markov service rate queue does not change the asymptotic queueing behavior of input traffic with long-tailed queue tail.

Note: If we use the de-multiplexing technique which was explained in subsection 3.2 for the two components of the ON/OFF burst model, we get the same bound which we found in theorem 4.3. In other words theorem 4.3 for variable service rate queue is equivalent to proposition 3.2 and proposition 3.3 for the ON/OFF burst model traffic.

4.2 Renewal service rate queue model

In the renewal service rate queue, the service rate changes in time as a renewal process. The Markov service rate queue cannot model the variation of the queue service rate for heavy-tailed bursts because its alpha-period has an exponential distribution. For

more flexibility and realistic case introduce the renewal service rate queue and focus on heavy-tailed distribution of the alpha-periods.

In this model the service rates of the queue (C_1 and C_2) changes like a renewal ON/OFF source. We define the state of the queue as follows: when it has full capacity ($C(t) = C_1$) the queue is in the beta-period and when traffic burst takes a part of the queue capacity ($C(t) = C_2 < C_1$) the queue is in the alpha-period. Let τ_α and τ_β represent the expected values for the alpha-period and beta-period ($\tau_\alpha, \tau_\beta < \infty$). In the steady state (when $t \rightarrow \infty$) the service rate queue process becomes a stationary process. The average service rate can be computed by the renewal theorem:

$$\begin{aligned} C = \mathbb{E}[C(t)] &= P[C(t) = C_1]C_1 + P[C(t) = C_2]C_2 \\ &= \frac{\tau_\beta C_1 + \tau_\alpha C_2}{\tau_\alpha + \tau_\beta} \end{aligned} \quad (4.26)$$

Let $R(t)$ represent the input traffic of the queue. For the stability of the queue we require

$$R = \mathbb{E}[R(t)] < C = \mathbb{E}[C(t)]$$

Therefore

$$R = \mathbb{E}[R(t)] < C = \frac{\tau_\beta C_1 + \tau_\alpha C_2}{\tau_\alpha + \tau_\beta} \quad (4.27)$$

Assume that the input traffic is a CBR traffic. If $R < C_2$ the queue forwards the packets as soon as they enter the queue. So the queue size is almost zero. But if $R > C_2$, the queue size increases in the alpha-periods. So the queue has a different behavior which depends on the distribution function of the length of the alpha-periods and beta-periods. Also for the non-CBR traffic we analyze the variable queue as CBR traffic in two cases ($R < C_2$ and $R > C_2$) separately.

Case 1: When $R < C_2$, the variation in service rate does not affect the asymptotic queueing behavior. This is because the average of input traffic even during the alpha-periods, is less than the service rate. Denote by $Q_L(t)$ and $Q_U(t)$ to be the queue

length of the two queues with service rates C_1 and C_2 , and let $Q(t)$ be the size of the variable service rate queue at time t . By assumption $C_2 \leq C(\tau) \leq C_1$ for $0 \leq \tau$. So when Q , Q_L and Q_U receive the same input traffic then

$$Q_L(t) \leq Q(t) \leq Q_U(t) \quad (4.28)$$

This inequality helps us to analyze $P[Q > b]$ asymptotically. The distribution function of the queue size depends on the type of input traffic. So the queues Q_L and Q_U have the same type of distribution functions which helps us to find the asymptotic queueing behavior of Q . The following theorem develops the asymptotic equality of the queue tails.

Theorem 4.4 Let C_1 and C_2 be service rates of a variable service rate queue and $R(t)$, the input rate. If $R = \mathbb{E}[R(t)] < C_2 < C_1$ and if $P[W^{A(t), C_1} > b]$ and $P[W^{A(t), C_2} > b]$ are equal asymptotically, then $P[W^{A(t), (C_1, C_2)} > b]$ is equal to both asymptotically.

Proof of theorem 4.4 : By the definition of backlog traffic:

$$\begin{aligned} P[W^{A(t), (C_1, C_2)} > b] &= P[\max_{s \geq 0} \sum_{k=0}^s \{R(t-k) - C(t-k)\} > b] \quad (4.29) \\ &\leq P[\max_{s \geq 0} \sum_{k=0}^s \{R(t-k) - C_2\} > b] \\ &= P[W^{A(t), C_2} > b] \end{aligned}$$

By inequalities (4.26) and (4.30) we have:

$$P[W^{A(t), C_2} > b] \leq P[W^{A(t), (C_1, C_2)} > b] \leq P[W^{A(t), C_1} > b]$$

So, the Sandwich theorem implies that if $P[W^{A(t), C_2} > b]$ and $P[W^{A(t), C_1} > b]$ are asymptotically equal, then $P[W^{A(t), C_2} > b]$ is asymptotically equal to them.

For example, let's assume that the input traffic is an fGn process with a positive mean value less than C_2 . It is known [Nor97, LR97, DO95] that the Q_L and Q_U processes have Weibull distribution functions with different scaling parameters. So distribution function of the queue Q is bounded between the Weibull distribution functions. This shows that Q has a Weibull distribution function asymptotically.

Case 2: When the average input traffic is larger than the service rate during the alpha-periods ($R > C_2$), the average queue size increases in alpha-periods. In this case the queueing behavior could be strongly changed, and it depends on the distribution function of the alpha-period. In [JL99, SSB02, AMN99, Box96, BC00], the queueing behavior constant service rate queue with multiplexing input has been analyzed. In these studies the input traffic is modelled as a superposition of an arbitrary traffic and high rate ON/OFF traffic and found the asymptotic queue tail of the input traffic in terms of the distribution of the ON time period. Below, we will use some of the these results cast in our frame work (variable service rate queue).

Proposition 4.3 Let C_1 and C_2 be the service rates of the renewal service rate queue and R be the rate of the CBR input. Also let $\tau_\alpha < \infty$ and $\tau_\beta < \infty$ be the expected values of the alpha-period ($C(t) = C_2$) and beta-period ($C(t) = C_1$). If $C_2 < R < C = \mathbb{E}[C(t)]$ then

$$P[Q > 0] = \frac{C_1 - C_2}{C_1 - R} \frac{\tau_\alpha}{\tau_\alpha + \tau_\beta}$$

Proof of Proposition 4.3: As we showed in the proof of Lemma 4.2, in equation (4.11)

$$\begin{aligned} P[Q > 0] &= P(Q \downarrow) + P(Q \uparrow) \\ &= \frac{C_1 - C_2}{C_1 - R} P(Q \uparrow) \end{aligned}$$

and by the renewal theorem $P(Q \uparrow) = P[t \in \text{alpha-period}] = \tau_\alpha / (\tau_\alpha + \tau_\beta)$, so

$$\begin{aligned} P[Q > 0] &= \frac{C_1 - C_2}{C_1 - R} P(Q \uparrow) \\ &= \frac{C_1 - C_2}{C_1 - R} \frac{\tau_\alpha}{\tau_\alpha + \tau_\beta} \end{aligned}$$

Theorem 4.5 Let C_1 and C_2 be the service rates of a renewal service rate queue and R be the rate of the CBR input. Also let $\tau_\alpha < \infty$ and $\tau_\beta < \infty$ be the expected values of the alpha-period ($C(t) = C_2$) and beta-period ($C(t) = C_1$). If $C_2 < R < C = \mathbb{E}[C(t)]$ and if the size of the alpha-period (T_α) is a long-tailed random variable such that T_α^* is a subexponential random variable then:

$$P[Q > b] \asymp \frac{\tau_\beta}{(\tau_\alpha + \tau_\beta)^2} \frac{C_1 - C_2}{C - R} P[T_\alpha^* > \frac{b}{R - C_2}]$$

Remark: In the study [JL99, Box96] the queueing behavior of a single ON/OFF source has been explained. In the case that T_α is a long-tailed random variable and T_α^* is a subexponential random variable, the queue tail can be computed in terms of distribution functions of the ON periods and OFF periods asymptotically. If the variation of queue input in the ON and OFF periods is modelled by an renewal service rate queue then by using their result, the queue tail can be computed in terms of distribution functions of the alpha-periods and beta-periods asymptotically. Also by considering proposition 4.3 we have:

$$P[Q > b] \asymp \frac{1 - P[Q > 0]}{\tau_\alpha + \tau_\beta} P[T_\alpha^* > \frac{b}{R - C_2}] \quad (4.30)$$

We can determine the asymptotic queueing behavior of a non-CBR traffic in the renewal service rate queue. We compare the tail of $P[T_\alpha^* > x]$ with the tail of

$P[W^{A,C} > x]$. If $P[T_\alpha^* > x]$ has a tail longer than the tail of $P[W^{A,C} > x]$ then, the renewal service rate queue with input $A(t)$ behaves like a renewal service rate queue with CBR input ($A(t) = Rt$), in the queue tail. It makes sense because when the alpha-period is too long and the variation of input process is small, the size of the queue increases almost with rate $R - C_2$ in the alpha-periods and it decreases almost with the rate $C_1 - R$ in the beta-periods. So for large queues, only the mean value of input process affects the queueing behavior.

By having some conditions on $A(t)$ and the distribution of T_α we have the following asymptotic equality.

$$\lim_{x \rightarrow \infty} \frac{P[W^{A(t), (C_1, C_2)}]}{P[W^{Rt, (C_1, C_2)}]} = 1 \quad (4.31)$$

Thereby, we use some of the results of [JL99, SSB02, AMN99, Box96] to find the asymptotic queueing behavior of the renewal service rate queue. The following theorem explains a particular case that (4.31) holds.

Theorem 4.6 Assume that T_α has a heavy-tailed distribution function ($T_\alpha \sim L(x).x^{-\gamma}$ where $1 < \gamma < 2$ and $L(x)$ is a slowly varying function), and for $\varepsilon > 0$

$$P[W^{A(t), R+\varepsilon} > x] = O(x^{-\gamma+1}). \quad (4.32)$$

If $C_2 < R < C$, then (4.31) holds.

Remark: For a heavy-tailed T_α :

$$P[T_\alpha^* > x] \asymp \frac{L(x)}{\tau_\alpha} x^{-\gamma+1} \quad (4.33)$$

Therefore for many types of input processes, (4.31) holds. For example when the input process is fGn, $W^{A(t), R+\varepsilon}$ has a Weibull distribution function. So it satisfies the condition of theorem 4.6.

4.3 Queuing analysis of ON/OFF burst model: Particular case

The effect of the ON/OFF source in the queue depends on the rate of the ON/OFF source during the ON periods and the free capacity of the queue. The free capacity of the queue is the service rate of the queue minus the mean value of beta component. It is the free capacity of the queue that can be used by the alpha traffic. Let r_{on} be the rate of the ON/OFF source during the ON periods. Let τ_{on} and τ_{off} be the expected values of the length of the ON and OFF periods. Also let R be the average rate of beta traffic and C the service rate of the queue. For the stability of the queue, the rate of input traffic should be less than the service rate of the queue. Therefore

$$C > R + \frac{\tau_{\text{on}} r_{\text{on}}}{\tau_{\text{on}} + \tau_{\text{off}}} \quad (4.34)$$

If $r_{\text{on}} < C - R$, the queue has enough capacity during the ON periods, and so the ON/OFF source only changes the rate of input traffic and it does not change the asymptotic queueing behavior of queue. This fact will be explained in next section by the variable service rate queue concept.

However when $r_{\text{on}} > C - R$, the queueing behavior strongly depends on the distribution function of the length of the ON period. Since in this case the rate of the input process is larger than the service rate during the ON periods, the size of the queue increases during the ON periods. Therefore depending on the distribution function of the length of the ON period the queueing behavior could be very different. Theorem 4.6 determines the asymptotic queueing behavior of the queue when the ON/OFF source has a heavy-tailed distribution for the ON period and the beta traffic is modelled by fGn traffic. The following corollary determines the asymptotic queueing behavior of special case by using the theorems 4.4 and 4.6.

Corollary 4.2 Assume that T_α (length of ON periods) has a heavy-tailed distribution function and $A(t) = \sigma B_H(t) + Rt$ is the aggregate of fGn traffic, if $R + r_{\text{on}} > C$ and (4.34) holds then

$$P[Q > b] \asymp M_\alpha \gamma_2 b^{-\alpha+1} \quad (4.35)$$

and if $R + r_{\text{on}} < C$

$$P[Q > b] \stackrel{\log}{\asymp} \exp(-\gamma_1 b^{2(1-H)}) \quad (4.36)$$

The asymptotic queueing behavior of the self-similar burst model (equation (3.22)) and the ON/OFF burst model (equation (4.2)) shows that if the ON/OFF traffic has a rate larger than free capacity of the queue during the ON periods then it behaves like a heavy-tailed arrival traffic.

Also when there are N ON/OFF sources with heavy-tailed ON periods, and if $r_{\text{on}_1} + r_{\text{on}_2} + \dots + r_{\text{on}_N}$ is larger than the free capacity, then the queue tail is a power-law function for large buffers. But when $r_{\text{on}_1} + r_{\text{on}_2} + \dots + r_{\text{on}_N}$ is less than the free capacity, the queue tail is described by a Weibull function.

4.4 Queueing analysis of ON/OFF burst model in the variable service queue framework

When the input traffic obeys the ON/OFF burst model, the queue behaves like a variable service rate queue. Assume that an ON/OFF source with a rate r_{on} during the ON periods is feed into the queue with service rate C_1 . So the service rate of queue is $C_1 - 0 = C_1$ during the OFF periods and it is $C_2 = C_1 - r_{\text{on}}$ during the ON periods. If the rate of ON/OFF source is smaller than free capacity of the queue

($r_{\text{on}} < C - R$), the asymptotic behavior of the queue can be determined by theorem 4.4. So we wonder to focus in the case that the rate of the ON/OFF source during the ON periods is larger than free capacity of the queue ($r_{\text{on}} > C - R$) and also for the stability of the queue (4.34) holds.

We consider two cases for ON/OFF source with the rate larger than free capacity of the queue during the ON periods. First when the ON/OFF source satisfies the moment conditions, for example when the transition states is a Markov chain in the constant size time bins and second when ON/OFF source has a long-tailed distribution in the ON periods, for example when ON/OFF source is renewal process and ON period has a heavy-tailed distribution function.

Case 1: When the ON/OFF source satisfies the moment conditions (4.37) and (4.38), it can affect the queueing behavior of short-tailed traffic. For example when the ON/OFF periods can be modelled by a Markov chain in the constant size time bins and the other part of traffic is CBR, the queueing behavior could be analyzed by theorems 4.1 and 4.2. Also by theorem 4.3 when the other part of the traffic has a queue tail longer than the exponential function then these ON/OFF sources do not affect the queueing behavior very much.

Definition 4.1 The moment conditions for an ON/OFF source are

$$\mathbb{E}[\exp(\theta T_{\text{on}})] < \infty \tag{4.37}$$

$$\mathbb{E}[\exp(\theta r_{\text{on}})] < \infty \tag{4.38}$$

for $0 < \theta < \theta_0$, where T_{on} and r_{on} are the length and height of the ON period respectively. For example the Poisson process with constant or exponential distribution for the heights satisfies the moment conditions.

Proposition 4.4 When the superposition of a CBR traffic with rate R and Markov ON/OFF sources with rate r_{on} during the ON periods is the input traffic of a queue with service rate C_1 and if (4.34) holds and $R > C_2 = C_1 - r_{\text{on}}$, the queue size has short-tailed distribution and it can be computed by theorem 4.1.

Proposition 4.5 When the superposition of a long-tailed traffic with a finite number of Markov ON/OFF sources is the input traffic of a queue and the stability condition of the queue holds, then the queueing behavior is asymptotically the same as in the case when the long-tailed traffic and a CBR traffic with the rate equal to the mean values of ON/OFF sources (instead of the ON/OFF sources) are feed into the queue.

(proof by theorem 4.3)

Case 2: When the ON/OFF source does not satisfy the moment condition, then it can not be modelled by a Markov ON/OFF source. We may still model it by a renewal process. Let T_α represent the ON period random variable and also let $\tau_\alpha, \tau_\beta < \infty$ represent the expectation values of the length of the ON and OFF periods. The proposition 4.6 determines the queue tail for the long-tailed ON/OFF source.

Proposition 4.6 Assume that T_α has a heavy-tailed distribution function ($T_\alpha \sim L(x).x^{-\gamma}$ where $1 < \gamma < 2$ and $L(x)$ is a slowly varying function), and for $\varepsilon > 0$

$$P[W^{A(t), R+\varepsilon} > x] = O(x^{-\gamma+1}). \quad (4.39)$$

If $r_{\text{on}} > C - R$ and (4.34) holds, then

$$P[Q > b] \asymp M.b^{-\gamma+1}$$

4.5 Summary of the ON/OFF burst model

In the traffic ON/OFF burst model the alpha component is modelled by a high rate ON/OFF source. We showed that when the rate of the ON/OFF source during the ON periods is less than the free capacity (which is the service rate of the queue minus mean arrival of the beta component) the ON/OFF source does not change the asymptotic quality of the queueing behavior of the queue. However when the rate is larger than the free capacity, the queueing behavior could be strongly changed by the ON/OFF source.

If the rate of the ON/OFF source is larger than the free capacity during the ON periods then the asymptotic queueing behavior for the traffic obeying the ON/OFF burst model is determined by the beta traffic or the distribution function of ON periods, depending on which of the components has a longer tailed queueing behavior. For example when the ON/OFF source is a Markov process and if the beta component is CBR or Poisson with small variance then the queueing behavior is determined by the Markov chain parameters. But if the beta traffic is fGn ($1/2 < H < 1$) or any other long-tailed traffic then the asymptotic queueing behavior is determined by the beta traffic. In the particular case when the alpha component is modelled by a renewal ON/OFF source with heavy-tailed distribution for the ON period and the beta component is modelled by a fGn process, the queue tail has a Weibull distribution function when the rate of ON/OFF source during the ON periods is less than the free capacity, and it has a power-law distribution function when the rate is larger than the free capacity.

Chapter 5

Conclusion

In this thesis, we studied the effect of traffic bursts in the network queue, building on the alpha-beta decomposition of internet traffic. The alpha component is the traffic which is generated by a few high bandwidth sources. The beta traffic is the residual traffic and makes up the bulk of the traffic. The beta traffic is Gaussian and it is well modelled by the fractional Gaussian noise (fGn) process. But the alpha traffic is non-Gaussian and bursty. In this work, the alpha traffic is modelled in two different ways.

First the alpha traffic is modelled by a self-similar bursty process. The full traffic is the superposition of a fGn traffic which models the beta component and stable Levy motion (sLn) which models the alpha component. So the self-similar burst model has bursts with a heavy-tailed distribution function. The queueing analysis of this model shows that the components affect the network queue differently. The fGn traffic (beta component) affects the queueing behavior only for small buffer sizes; when the buffer size is large, it is strongly affected by the sLn traffic (alpha component).

In the second model (ON/OFF burst model) the alpha traffic is modelled by a high rate ON/OFF source. At large time scales, a long ON period of this source produces a burst in the traffic. When the alpha component is feed into a queue it consumes a part of the queue capacity. So the queueing behavior depends on the amount of capacity which the ON/OFF source takes up.

When the rate of the ON/OFF source during the ON periods is less than the free capacity of the queue (which is capacity of the queue minus mean value of the beta component) the ON/OFF source does not change the asymptotic behavior of the traffic. But when it is larger than the free capacity, the queueing behavior could be very different. The queueing behavior in this case is analyzed in the two cases of ON/OFF source with short-tailed and long-tailed ON duration.

When the ON/OFF source has a short-tailed distribution for the ON periods or when the ON and OFF state transitions occur as a Markov chain in the constant size time bins then it does not affect the asymptotic queueing behavior of long-tailed traffic. It just affects the queueing behavior in the presence of CBR or any other short-tailed traffic. For example when the ON/OFF source changes as a Markov chain in the constant size time bins and the beta traffic is fGn traffic ($1/2 < H < 1$) then the ON/OFF source does not change the asymptotic behavior of the queue, and the queue tail has a Weibull decay in this case.

When the ON periods have a long-tailed distribution function, the source can be modelled by a renewal ON/OFF source. If the rate of the ON/OFF source during the ON periods is larger than the free capacity, the average queue size increases during the ON periods. Therefore the queueing behavior depends on the distribution of ON periods. For example, when the beta component is fGn traffic and the ON time period has a heavy-tailed distribution function, the queue tail has a power-law decay.

The queue tails in both presented models have the same asymptotic behavior assume that the ON time periods in the ON/OFF burst model have a heavy-tailed distribution which is the same as the distribution of bursts in the self-similar burst model; assume also that the rate of the ON/OFF source during the ON periods is larger than the free capacity, then. It shows that if we consider the traffic bursts as huge TCP windows with heavy-tailed distribution or as heavy-tailed ON periods of

an ON/OFF source with a rate larger than the free capacity during the ON periods, the queue tail has the same asymptotic behavior. This makes sense because the self-similar burst model is a large scale version of the ON/OFF burst model; therefore they should have the same queueing behavior. However, the ON/OFF burst model gives more accurate results for various types of alpha and beta traffics. In particular the queue tail remains Weibullian for a fGn beta traffic component and an ON/OFF alpha component with high rate yet not too high as to consume not more than the free capacity.

Bibliography

- [AMN99] R. Agrawal, A. M. Makowski, and P. Nain. On a reduced load equivalence for fluid queues under subexponentiality. *QUESTA*, 33(1-3):5–41, 1999.
- [BC00] O.J. Boxma and J.W. Cohen. The single server queue: Heavy tails and heavy traffic. *Self-similar Network Traffic and Performance Evaluation*, Wiley, 2000.
- [BCRH⁺00] V. Bolotin, J. Coombs-Reyes, D. Heyman, Y. Levy, and D. Liu. Ip traffic characterization for planning and control. *AT&T Labs*, 2000.
- [BD01] A. Budhiraja and P. Dupuis. Large deviations for the empirical measures of reflecting brownian motion and related constrained processes in r_+ . *Division of Applied Mathematics*, 2001.
- [BM00] S. Bates and S. McLaughlin. The estimation of stable distribution parameters from teletraffic data. *IEEE Trans. Signal Proc.*, 48(3):865–870, 2000.
- [Box96] O.J. Boxma. Fluid queues and regular variation. *BS-R9608, ISSN 0924-0659*, 1996.
- [CB97] M. Crovella and A. Bestavros. Self-similarity in World Wide Web traffic. Evidence and possible causes. *IEEE/ACM Transactions on Networking*, 5:835–846, December 1997.

- [CY01] H. Chen and D. D. Yao. *Fundamentals of Stochastic Networks*. Springer, 2001.
- [DO95] N. Duffield and N. O’Connell. Large deviations and overflow probabilities for the general single-server queue, with applications. *Math. Proc. Cambr. Phil. Soc.*, 118:363–374, 1995.
- [GK02] R. Gaigalas and I. Kaj. Convergence of scaled renewal processes and a packet arrival model. *U.U.D.M. Report 2002:2*, Uppsala University, 2002.
- [Jel98] P. R. Jelenkovic. Asymptotic results for queues with subexponential arrivals. *Self-Similar Network Traffic and Performance Evaluation*, Wiley, 1998.
- [JL99] P. R. Jelenkovic and A. A. Lazar. Asymptotic results for multiplexing subexponential on-off processes. *Advances in Applied Probability*, 31(2), 1999.
- [KH98] A. Karasaridis and D. Hatzinakos. Broadband heavy-traffic modeling using stable self-similar processes. Technical report, 2nd Canadian Conference on Broadband Research, Ottawa, Ontario, June 1998.
- [LLDH02] N. Laskin, I. Lambadaris, M. Devetsikiotis, and F. Harmantzis. Fractional levy motion and its application to traffic modeling. *Math. Proc. Cambr. Phil. Soc.*, 40:363–375, 2002.
- [LR97] J. Lévy Véhel and R. Riedi. Fractional Brownian motion and data traffic modeling: The other end of the spectrum. *Fractals in Engineering*, pages 185–202, Springer 1997.

- [MDM02] S. Molnar, T. D. Dang, and I. Maricza. On the queue tail asymptotics for general multifractal. 2002.
- [Nor97] I. Norros. Four approaches to the fractional Brownian storage. *Fractals in Engineering*, pages 154–169, 1997.
- [NW98] A. L. Neidhardt and J. L. Wang. The concept of relevant time scales and its application to queuing analysis of self-similar traffic. *SIGMETRICS '98/Performance '98*, 1998.
- [Pra97] N. U. Prabhu. *Foundation of Queueing Theory*. Kluwer Academic Publication, 1997.
- [SRB02] S. Sarvotham, R. Riedi, and R. Baraniuk. Multiscale connection-level analysis of network traffic. *Proceedings Proceedings of the 36th Conference on "Signals, Systems and Computers", Asilomar, CA., Nov 2002*.
- [SSB02] R. Riedi S. Sarvotham and R. Baraniuk. Connection-level modeling of network traffic. Technical report, ECE Dept, Rice University, Houston, Texas, 2002.
- [ST94] G. Samorodnitsky and M. S. Taqqu. *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*. Chapman and Hall, New York, 1994.
- [TL86] M. Taqqu and J. Levy. *Using renewal processes to generate LRD and high variability*. In: Progress in probability and statistics, E. Eberlein and M. Taqqu eds., volume 11. Birkhaeuser, Boston, 1986. pp 73–89.

- [TWS97] M. S. Taqqu, W. Willinger, and R. Sherman. Proof of a fundamental result in self-similar traffic modeling. *Computer Communication Review*, 17:5–23, 1997.
- [Whi00] W. Whitt. *Stochastic-Process Limits: An Introduction to Stochastic-Process limits and Their Application to Queue*. Springer, 2000.
- [WPRT] W. Willinger, V. Paxson, R. Riedi, and M. Taqqu. *Long range dependence : theory and applications*, chapter Long range dependence and Data Network Traffic. Doukhan, Oppenheim and Taqqu eds.
- [WSRB02] X. Wang, S. Sarvotham, R. Riedi, and R. Baraniuk. Connection-level modeling of network traffic. *Proceedings DIMACS Workshop on Internet and WWW Measurement, Mapping and Modeling, Rutgers, NJ*, Feb 2002.
- [WTSW97] W. Willinger, M. Taqqu, R. Sherman, and D. Wilson. Self-similarity through high-variability: Statistical analysis of Ethernet LAN traffic at the source level. *IEEE/ACM Trans. Networking (Extended Version)*, 5(1):71–86, Feb. 1997.