Performance analysis of priority scheduling mechanisms under heterogeneous network traffic✩

Xiaolong Jin, Geyong Min *

Department of Computing, School of Informatics, University of Bradford, Bradford, BD7 1DP, UK

Received 3 October 2005; received in revised form 11 March 2006
Available online 24 February 2007

Abstract

Priority scheduling principle plays a crucial role in the Differentiated Services (DiffServ) architecture for the provisioning of Quality-of-Service (QoS) of network-based applications. Analytical modelling and performance evaluation of priority queuing systems have received significant attention and research efforts. However, most existing work has primarily focused on the analysis of priority queuing under either Short Range Dependent (SRD) or Long Range Dependent (LRD) traffic only. Recent studies have shown that realistic traffic reveals heterogeneous nature within modern multi-service networks. With the aim of investigating the impact of heterogeneous traffic on the design and performance of network-based systems, this paper proposes a novel analytical model for priority queuing systems subject to heterogeneous LRD self-similar and SRD Poisson traffic. The key contribution of the paper is to extend the application of the generalized Schilder's theorem (originally a large deviation principle for handling Gaussian processes only) to deal with heterogeneous traffic and further develop the analytical upper and lower bounds of the queue length distributions for individual traffic flows. The validity and accuracy of the model demonstrated through extensive comparisons between analytical bounds and simulation results make it a practical and cost-effective evaluation tool for investigating the performance behaviour of priority queuing systems under heterogeneous traffic with various parameter settings.

© 2007 Elsevier Inc. All rights reserved.

Keywords: Large deviation principle; Priority queuing; Self-similar traffic; Long range dependence; Short range dependence

1. Introduction

With the explosive advances in networking and communication technologies, network-based computing has become a popular paradigm for cost-effective high-performance computing [1]. More recently, Internet-based computing and grid computing have been proposed for setting up global-scale systems which are crucial in many domains, such as finance, industry, military, and telecommunications [2,3]. The provisioning of Quality-of-Service (QoS) has become an increasingly pressing demand of various network-based applications [2]. As an efficient scheme for supporting QoS, the Differentiated Services (DiffServ) architectural model [4] classifies packets into one of a small number of aggregated flows or classes which are handled with differentiated priorities. Therefore, priority scheduling policy plays

✩ This work is supported by the UK EPSRC research grant (EP/C525027/1).
* Corresponding author.

E-mail addresses: x.jin@brad.ac.uk (X. Jin), g.min@brad.ac.uk (G. Min).
a crucial role in the implementation of the DiffServ model and has a great impact on its efficiency. Many research efforts have been made on performance evaluation and analysis of the priority queuing mechanism.

Researchers originally focused on the investigation of priority queuing systems subject to Short Range Dependent (SRD) traffic [5–8]. For instance, Choi et al. [5] studied a queuing system where high and low priority traffic flows are modelled by a Markov Modulated Poisson Process (MMPP) and an ordinary Poisson process, respectively. They theoretically deduced the steady-state joint generation function of the queue length of each traffic flow. Mazzini, Rovatti, and Setti [8] investigated a queuing system where both high and low priority traffic flows are statistically in the form of a Bernoulli distribution. They derived analytical expressions for the steady-state queue length distribution as well as the average queue length. Nannersalo and Norros [7] developed practically usable approximations for the queue length distribution of priority queuing systems with Gaussian traffic. Mandjes et al. [6] further studied the asymptotics of packet loss and delay of such systems.

Many recent studies [9–13] by means of high quality, high time-resolution measurements have indicated that traffic in a variety of networks exhibits noticeable burstiness over a wide range of time scales. This fractal-like behaviour of network traffic can be much better modelled using statistically self-similar processes, which have significantly different theoretical properties from those of the conventional SRD processes. The self-similar phenomenon of traffic has been found in local-area networks [10], wide-area networks [12], World Wide Web [9], wireless networks [11,13], and Variable-Bit-Rate (VBR) video systems [14]. Subsequently, research interests in priority queuing were transferred to those systems in the presence of Long Range Dependent (LRD) self-similar traffic [15–17]. For instance, Ashour and Le-Ngoc [15] employed Multiscale Wavelet Models (MWM) to characterize the LRD input traffic of a priority queuing system and validated the analytical estimations for the queue length distributions of both high and low priority traffic. Quan and Chung [17] developed a measurement-based method for estimating the buffer overflow probability of each queue in a priority queuing multiplexer. Iacovoni and Isopi [16] analyzed a queuing system where the high priority traffic is asymptotically self-similar while the low priority traffic is exactly self-similar. They derived a lower bound of the overflow probability for the low priority queue.

All these studies have focused on the investigation of priority queuing systems under homogeneous traffic only. More specifically, either SRD traffic or LRD traffic has been taken into account. However, realistic traffic within modern multi-service networks exhibits heterogeneous nature. To the best of our knowledge, there has been hardly any analytical model reported for priority queuing systems in the presence of heterogeneous SRD and LRD traffic. To fill in this gap, this paper proposes a novel analytical model for such systems subject to heterogeneous LRD self-similar traffic and SRD Poisson traffic based on Large Deviation Principles (LDPs). The major contributions of the paper are: (1) extending the application of the generalized Schilder’s theorem (i.e., originally an LDP for handling Gaussian processes only [7,18,19]) to deal with heterogeneous traffic; (2) deriving the analytical upper and lower bounds of the queue length distributions for individual traffic flows in priority queuing systems; and (3) using the developed analytical model, which is validated through simulation experiments, to conduct extensive performance evaluation of priority queuing systems and reaching some important performance results. The validity and accuracy of the model make it a valuable and cost-effective evaluation tool for investigating the performance behaviour of priority queuing systems under heterogeneous traffic with various parameter settings.

The rest of the paper is organized as follows. In Section 2 we briefly review the notions of Reproducing Kernel Hilbert Space (RKHS) and LDP. In particular, we introduce the generalized Schilder’s theorem, which is a special LDP defined on an RKHS generated from centered Gaussian processes. Section 3 presents the characteristics and mathematical modelling of heterogeneous self-similar and Poisson traffic. Next, we extend the application of the generalized Schilder’s theorem to heterogeneous traffic and derive the upper and lower bounds for the queue length distributions of both traffic classes in Section 4. Results obtained from extensive simulation experiments validate the accuracy of the developed model in Section 5. Finally, Section 6 concludes the paper.

2. Preliminaries

In what follows, we will briefly review the definitions and knowledge of Reproducing Kernel Hilbert Space (RKHS) and Large Deviation Principle (LDP) which are the essential fundamentals of the derivation of our analytical model.
2.1. Reproducing Kernel Hilbert Space (RKHS)

The theory of RKHS [20,21] was originally developed for the studies of integral equations and partial differential equations. Later, the use of this theory was extended to a number of fields, such as, probability and statistics, stochastic processes, and signal processing. Section 2.2 will show that the generalized Schilder’s theorem [22,23] is defined on an RKHS generated from a group of centered Gaussian processes.

Definition 1 (Reproducing Kernel Hilbert Space). A Hilbert space \(S\) of functions \(f : \mathbb{X} \to \mathbb{R}, \mathbb{X} \neq \emptyset\), is called an RKHS with inner product \(\langle \cdot, \cdot \rangle\) and norm \(\|f\| = (\langle f, f \rangle)^{1/2}\), if there exists a function \(K : \mathbb{X} \times \mathbb{X} \to \mathbb{R}\), which has the following two properties:

- For any \(f \in S\), \(\langle f, K(x, \cdot) \rangle = f(x)\);
- \(S\) is the closure of set \(\{ f | f(\cdot) = \sum_{i=1}^{m} \alpha_i K(x_i, \cdot), \forall m \in \mathbb{N}, x_1, \ldots, x_m \in \mathbb{X}, \text{and} \alpha_1, \ldots, \alpha_m \in \mathbb{R} \}\).

Particularly, function \(K\) is called the reproducing kernel of Hilbert space \(S\). \(\langle f, K(x, \cdot) \rangle = f(x)\) is called the reproducing kernel property.

From the above definition of RKHS, we can note that a Hilbert space can be generated from a given reproducing kernel function. Usually, the covariance function of a centered (i.e., zero mean) stochastic process is a reproducing kernel property.

2.2. Large Deviation Principles (LDPs)

Unlike the conventional probability theory (e.g., the law of large numbers, the central limit theorem) focusing on general events, large deviation theory was developed to address the properties of rare events, such as their frequency and most probable way of occurrence [24]. It is regarded as a refinement of the law of large numbers. It has been applied to such fields as queuing theory and network traffic engineering [18,19,25–27].

Generally speaking, an LDP characterizes limiting behaviour of a sequence of random variables in terms of a rate function. In particular, it provides the asymptotical upper and lower bounds of the probability distribution of the addressed random variables.

Definition 2 (Rate function). A rate function \(I\) on space \(\mathbb{R}^d\) is a non-negative, lower semicontinuous mapping \(I : \mathbb{R}^d \to [0, \infty]\). Here, ‘lower semicontinuous’ implies that for any sequence \(\{z_n, n = 1, 2, \ldots\}\) and \(z_n \rightarrow z\) on \(\mathbb{R}^d\), we have

\[
\liminf_{n \to \infty} I(z_n) \geq I(z).
\]

If, further, for \(\forall b \in \mathbb{R}, \{y \mid I(y) \leq b\}\) is compact, then \(I\) is a good rate function.

Definition 3 (Large Deviation Principle). A stochastic sequence \(\{z_n, n = 1, 2, \ldots\}\) on space \(\mathbb{R}^d\) is said to satisfy an LDP with rate function \(I\), if:

- (Upper bound) for any closed set \(F \subset \mathbb{R}^d\),
  \[
  \limsup_{n \to \infty} \frac{1}{n} \log P(z_n \in F) \leq - \inf_{x \in F} I(x);
  \]
- (Lower bound)
(Lower bound) for any open set \( G \subset \mathbb{R}^d \),

\[
\liminf_{n \to \infty} \frac{1}{n} \log P(z_n \in G) \geq - \inf_{x \in G} I(x).
\] (3)

Now, let us introduce a special LDP, i.e., the generalized Schilder’s theorem [7,18,19]. This theorem has been employed to address generalized processor sharing and priority queuing systems with Gaussian traffic by Mannersalo and Norros [7,18,19]. Based on their work, we extend the application of this theorem to priority queuing systems with heterogeneous self-similar and Poisson traffic.

Let \( Z = (Z_1, Z_2, \ldots, Z_k) \) be a group of independent, centered Gaussian processes. Let \( v_1(\cdot), v_2(\cdot), \ldots, v_k(\cdot) \) and \( \Gamma_1(\cdot, \cdot), \Gamma_2(\cdot, \cdot), \ldots, \Gamma_k(\cdot, \cdot) \) represent their variance and covariance functions, respectively. The generalized Schilder’s theorem requires \( Z_i \) \((1 \leq i \leq k)\) to be a process with stationary increments and a continuous variance function satisfying the following condition:

\[
\exists \alpha < 2, \text{ s.t. } \lim_{t \to \infty} \frac{v_i(t)}{t^\alpha} = 0.
\] (4)

Since \( Z_i \) \((1 \leq i \leq k)\) is a centered Gaussian process, taking \( \Gamma_i(\cdot, \cdot) \) as a reproducing kernel, we can generate its RKHS \( S_i \).

The norm of \( S \) can be denoted as

\[
\| (f_1, \ldots, f_k) \|^2_S = \sum_{i=1}^{k} \langle f_i, f_i \rangle_S = \sum_{i=1}^{k} \langle f_i, f_i \rangle_S = \sum_{i=1}^{k} \langle f_i, f_i \rangle_S.
\] (6)

**Theorem 4 (Generalized Schilder’s theorem).** With rate function \( I: S \to [0, \infty] \),

\[
I(\omega) = \left\{ \begin{array}{ll}
\frac{1}{2} \| \omega \|^2_S, & \text{if } \omega \in S, \\
\infty, & \text{otherwise},
\end{array} \right.
\] (7)

\( Z = (Z_1, Z_2, \ldots, Z_k) \) satisfies the following LDP:

- (Upper bound) for closed set \( F \subset S \):

\[
\limsup_{k \to \infty} \frac{1}{k} \log P\left( \frac{Z}{\sqrt{k}} \in F \right) \leq - \inf_{\omega \in F} I(\omega);
\] (8)

- (Lower bound) for open set \( G \subset S \):

\[
\liminf_{k \to \infty} \frac{1}{k} \log P\left( \frac{Z}{\sqrt{k}} \in G \right) \geq - \inf_{\omega \in G} I(\omega).
\] (9)

3. Modelling of heterogeneous traffic

As aforementioned, this study focuses on performance modelling and analysis of priority queuing systems fed with heterogeneous network traffic. Figure 1 presents a schematic diagram of such a queuing system. As illustrated in this figure, we specifically address two types of traffic, namely, LRD self-similar traffic and SRD Poisson traffic. In general, self-similar processes can be used to model traffic generated by multimedia applications (e.g., VBR video), while Poisson processes are applied to model traffic generated by the traditional non-bursty text communication. As compared to the traditional text applications, multimedia applications require more stringent QoS provisioning and differentiation. Therefore, in the addressed priority queuing system, self-similar traffic is assigned with high priority.
Fig. 1. A schematic diagram of the priority queuing system addressed in this paper.

3.1. Self-similar traffic

The innovative study of Leland et al. [10] has revealed that traffic in an Ethernet local area network environment exhibits noticeable self-similar nature (i.e., burstiness and correlations among inter-arrival intervals over many time scales) and set the groundwork for considering self-similarity as a key notion in the understanding of traffic properties, modelling and analysis of network performance. Their work has triggered an explosion of research on this subject (e.g., [28–30]). Many subsequent studies have further demonstrated traffic self-similarity be a ubiquitous phenomenon in a variety of telecommunication systems [9–13].

Many new models and techniques have been developed to characterize traffic self-similarity or generate self-similar traffic traces, such as, fractional Brownian motion (fBm)/fractional Gaussian noise (fGn) [31,32], Multiscale Wavelet Model (MWM) [15], superposition of MMPPs [28,33,34], and superposition of ON/OFF sources with heavy-tailedly distributed ON periods and/or OFF periods [35,36]. Among these models, fBm is identified as the most efficient way for modelling and generating self-similar traffic [31]. In this paper, we adopt fBm as the model of self-similar traffic.

Generally speaking, traffic arrival patterns can be modelled as a stochastic process and denoted in a cumulative arrival form,

\[ A = \{A(t)\}_{t \in \mathbb{R}}. \]

Then, \( A(s, t) = A(t) - A(s) \) denotes the amount of traffic arriving in time interval \((s, t] \).

Following [31], an fBm traffic flow can be expressed as

\[ A_f(t) = mt + Z_f(t), \]

where \( m \) is the mean arrival rate and \( Z_f(t) \) is a centered (i.e., \( E Z_f(t) = 0 \)) fBm with variance

\[ \bar{v}_f(t) = \text{Var} \bar{Z}_f(t) = t^{2H}, \]

and covariance

\[ \bar{\Gamma}_f(s, t) = \text{Cov}(\bar{Z}_f(s), \bar{Z}_f(t)) = \frac{1}{2}(\bar{v}_f(s) + \bar{v}_f(t) - \bar{v}_f(s - t)) = \frac{1}{2}(t^{2H} + s^{2H} - (t - s)^{2H}), \]

where \( H \in [0, 1] \) represents Hurst parameter which is a key measure of the persistence of a statistical self-similar phenomenon [10,12]. A value of \( H = 0.5 \) indicates the absence of self-similarity. The closer \( H \) is to 1, the greater the degree of persistence. The variance function of \( Z_f(t) \) is given by

\[ v_f(t) = \text{Var} Z_f(t) = am \bar{v}_f(t) = amt^{2H}, \]

and its covariance function is

\[ \Gamma_f(s, t) = \text{Cov}(Z_f(s), Z_f(t)) = \frac{1}{2}(v_f(s) + v_f(t) - v_f(s - t)) = \frac{1}{2}am(t^{2H} + s^{2H} - (t - s)^{2H}). \]

3.2. Poisson traffic

Poisson traffic in terms of a cumulative process can be denoted as \( A_p = \{A_p(t)\}_{t \in \mathbb{N}} \), which has expectation \( E(A_p(t)) = \lambda t \) and variance \( \text{Var}(A_p(t)) = \lambda t \).

Similar to the representation of an fBm traffic flow in Eq. (10), we can express a Poisson traffic flow \( A_p(t) \) as follows:

\[ A_p(t) = \lambda t + Z_p(t), \]
where $\lambda$ is the mean arrival rate of $A_p(t)$. $Z_p(t)$ is a stochastic process with expectation $E(Z_p(t)) = 0$ and variance $v_p(t) = \text{Var}(Z_p(t)) = \text{Var}(A_p(t)) = \lambda t$.

Further, we can derive the covariance function of $Z_p(t)$ as follows:

$$
\Gamma_p(s, t) = \lambda \min(s, t).
$$

(17)

In what follows, let us address the properties of $Z_p(t)$:

- Given $t$ is continuous, $v_p(t)$ is continuous as well. $\Gamma_p(s, t)$ is also continuous except one discontinuous point $t = s$ where $\Gamma_p(s, t)$ is left- and right-continuous.
- For $v_p(t)$ we have

$$
\lim_{t \to \infty} \frac{v_p(t)}{t^\alpha} = \lim_{t \to \infty} \frac{\lambda t}{t^\alpha} = 0, \quad \forall \ 1 < \alpha < 2.
$$

(18)

- $Z_p(t)$ can be regarded as the sum of $t$ independent, identically distributed random variables with zero expectation and limited variance. According to the central limit theorem, we have

$$
Z_p(t) \sim N(0, \sqrt{\lambda t}), \quad \text{as } t \to \infty.
$$

(19)

That is, $Z_p(t)$ approaches a centered Gaussian process with variance $\lambda t$ as $t$ tends to infinity.

These three properties of $Z_p(t)$ make it possible for us to extend the generalized Schilder’s theorem to Poisson traffic.

### 3.3. Queuing system

Let $A_{\{f,p\}}(t) = A_f(t) + A_p(t)$ denote the cumulative amount of fBm and Poisson traffic arriving at the priority queuing system at time $t$. $A_{\{f,p\}}(s, t) = A_f(s, t) + A_p(s, t)$ is then the total amount of fBm and Poisson traffic arriving in time interval $(s, t]$. Further, the aggregated queue length of the fBm and Poisson traffic at time $t$ can be readily denoted as follows [31]:

$$
Q_{\{f,p\}}(t) = \sup_{s \leq t} \{ A_{\{f,p\}}(s, t) - c(t - s) \},
$$

(20)

where $c$ represents the service rate of the server. Later on, we employ $Q_f(t)$ and $Q_p(t)$ to denote the individual queue lengths corresponding to fBm and Poisson traffic at time $t$, respectively.

### 4. Upper and lower bounds for queue length distributions

We have known that LDPs can be used to address rare events. The rare events considered in this paper are buffer overflow, represented by $\{ Q_{\{f,p\}}(0) > x \}$. In what follows, we will extend the generalized Schilder’s theorem to heterogeneous fBm and Poisson traffic and consequently derive the most probable path vector of $\{ Q_{\{f,p\}}(0) > x \}$.

**Proposition 5.** The most probable path vector in set $\{ Q_{\{f,p\}}(0) \geq x \}$ is:

$$
Z(t) = -x + (c - m - \lambda)t_x \Gamma_f(t_x, t_x) + \Gamma_p(t_x, t_x) (\Gamma_f(t_x, t), \Gamma_p(t_x, t)),
$$

(21)

where $t_x < 0$ and $t_x = \arg \min Y(t)$,

$$
Y(t) = \frac{(-x + (c - m - \lambda)\kappa)^2}{\Gamma_f(t, t) + \Gamma_p(t, t)}.
$$

(22)

**Proof.** We can analyze the events of buffer overflow as follows:

$$
\{ Q_{\{f,p\}}(0) \geq x \} = \bigcup_{t \leq 0} \{ A_{\{f,p\}}(t, 0) - c(0 - t) \geq x \} = \bigcup_{t \leq 0} \{ Z_{\{f,p\}}(t, 0) + (c - m - \lambda)t \geq x \}.
$$

(23)
If
\[ Z \in \{ Z_{(f,p)}(t,0) + (c - m - \lambda)t \geqslant x \} \cap S, \] (24)
then \( Z \in S \) and
\[ Z_f(t) + Z_p(t) \leqslant -x + (c - m - \lambda)t. \] (25)

Further, according to the reproducing kernel property, we have
\[ \{(Z_f(\cdot), Z_p(\cdot)), (\Gamma_f(t,\cdot), \Gamma_p(t,\cdot))\}_R \leqslant -x + (c - m - \lambda)t. \] (26)

Therefore, the problem to find the most probable path transforms to minimizing the norm of
\[ (Z_f(\cdot), Z_p(\cdot)) = \frac{-x + (c - m - \lambda)t}{\Gamma_f(t,t) + \Gamma_p(t,t)} (\Gamma_f(t,\cdot), \Gamma_p(t,\cdot)), \] (27)
with respect to \( t \), i.e.,
\[ \| (Z_f(\cdot), Z_p(\cdot)) \|^2 = \left( \frac{-x + (c - m - \lambda)t}{\Gamma_f(t,t) + \Gamma_p(t,t)} \right)^2 \left( (\Gamma_f(t,\cdot), \Gamma_p(t,\cdot)), (\Gamma_f(t,\cdot), \Gamma_p(t,\cdot)) \right) \]
\[ = \left( \frac{-x + (c - m - \lambda)t}{\Gamma_f(t,t) + \Gamma_p(t,t)} \right)^2 \left( \Gamma_f(t,t), \Gamma_p(t,t) \right) = \frac{(-x + (c - m - \lambda)t)^2}{\Gamma_f(t,t) + \Gamma_p(t,t)} \overset{\text{def}}{=} Y(t). \] (28)

From Eqs. (14) and (17), we can see that both \( \Gamma_f(t,t) \) and \( \Gamma_p(t,t) \) are differentiable with respect to \( t \). Therefore, minimizing \( Y(t) \) can be transformed to find \( t_x \) satisfying the equation
\[ \frac{\Gamma_f(t,t) + \Gamma_p(t,t)}{\Gamma_f(t,t) + \Gamma_p(t,t)} = \frac{1}{2} \left( t - \frac{x}{c - m - \lambda} \right). \] (29)

According to Proposition 5 and Theorem 4, we have the following basic approximation for the distribution of the aggregated queue length:
\[ P(Q_{(f,p)}>x) \approx \exp \left( -\frac{1}{2} \| Z(t) \|^2_S \right) = \exp \left( -\frac{1}{2} Y(t_x) \right). \] (30)

Through late comparisons to simulation results, we can see that this basic approximation is actually an upper bound of \( P(Q_{(f,p)}>x) \).

In what follows, let us derive a lower bound for \( P(Q_{(f,p)}>x) \):
\[ P(Q_{(f,p)}>x) = P \left( Q_{(f,p)}(0)>x \right) = P \left( \sup_{t \leq 0} Z_{(f,p)}(t,0) + (c - m - \lambda)t > x \right) \]
\[ \geq \sup_{t_x} P \left( \sup_{t \leq 0} Z_{(f,p)}(t) > x - (c - m - \lambda)t_x \right) = P \left( Z_{(f,p)}(-t_x) \geq x - (c - m - \lambda)t_x \right) \]
\[ = \Phi \left( \frac{x - (c - m - \lambda)t_x}{\sqrt{\Gamma_f(t_x,t_x) + \Gamma_p(t_x,t_x)}} \right), \] (31)
where \( t_x \) minimizes function \( Y(t) \) in Eq. (22) and \( \Phi(\cdot) \) is the residual distribution function of the standard Gaussian distribution.

Following the well-known method of the empty buffer approximation proposed by Berger and Whitt [37], the total queue in a priority queuing system is almost exclusively composed of the low priority traffic. As a consequence, the total queue length distribution can be reasonably used to approximate that of the low priority traffic. The empty buffer approximation has been widely adopted as an efficient method for modelling priority queuing systems [19,38,39]. Furthermore, our simulation experiments have demonstrated that the queue length distribution of the low priority Poisson traffic (i.e., \( P(Q_p>x) \)) is indistinguishable from that of the total queue (i.e., \( P(Q_{(f,p)}>x) \)). This further verifies the feasibility of the empty buffer approximation for analyzing priority queuing systems. Therefore, Eqs. (30) and (31) can be used to represent the upper and lower bounds of the queue length distribution of the low priority Poisson traffic.
The high priority traffic in a priority queuing system is served in a manner that the low priority traffic seems to be inexistent. Consequently, we may derive the following upper and lower bounds for the fBm queue length distribution by setting the amount of the arriving Poisson traffic to be zero in Eqs. (30) and (31):

\[ P(Q_f > x) \leq \exp \left( -\frac{1}{2} \frac{(-x + (c - m)t_x)^2}{\Gamma_f(t_x, t_x)} \right), \quad (32) \]

\[ P(Q_f > x) \geq \Phi \left( \frac{x - (c - m)t_x}{\sqrt{\Gamma_f(t_x, t_x)}} \right), \quad (33) \]

where \( t_x \) is the value that minimizes:

\[ Y'(t) = \frac{(-x + (c - m)t)^2}{\Gamma_f(t, t)}. \quad (34) \]

5. Validation and performance analysis

In this section, we investigate the accuracy of the upper and lower bounds of the queue length distributions we derived for priority queuing systems subject to heterogeneous LRD self-similar and SRD Poisson traffic. For this purpose, we developed a simulator for such systems using the C++ programming language and compared the analytical upper and lower bounds with the performance results obtained from extensive simulation experiments. In our simulation, an LRD fBm traffic flow and an SRD Poisson traffic flow were fed into the addressed queuing systems with service rate \( c = 120 \) where the high priority was assigned to the fBm traffic flow. The conditionalized Random Midpoint Displacement algorithm (RMD3,3) [40,41] was adopted to generate fBm self-similar traffic traces owing to its ability of producing real-time fBm traffic without prior knowledge of simulation trace length. Moreover, the computational complexity of this algorithm is linear as the trace length increases, thus keeping the time complexity of simulation at a reasonable level.

We have conducted extensive simulation experiments under various scenarios. For the sake of space limitation, in what follows we will present the results of four typical scenarios, which correspond to four different values of Hurst parameter \( H \), i.e., \( H \in \{0.6, 0.7, 0.8, 0.9\} \). Under each scenario, we will further test three most representative cases where the mean input rates, \( m \), of fBm flows and, \( \lambda \), of Poisson flows are set as follows:

- Case I: \( m = 90 \) and \( \lambda = 20 \);
- Case II: \( m = 55 \) and \( \lambda = 55 \);
- Case III: \( m = 20 \) and \( \lambda = 90 \).

In Case I, the fBm traffic dominates the input of the queuing system as \( m = 4.5 \lambda \). On the contrary, the Poisson traffic dominates in Case III. In Case II, the fBm traffic and the Poisson traffic are comparative. Simulation and analytical results are shown in Figs. 2–5 where the horizontal axis represents queue lengths and the vertical axis denotes their probability distributions. In these figures, we employed the solid (dashed) curves with signs ‘○,’ ‘▽,’ ‘△’ to represent the simulation results and the corresponding analytical upper and lower bounds of the fBm (Poisson) traffic, respectively.

5.1. Scenario 1: \( H = 0.6 \)

First of all, we can note that in all three cases (see Figs. 2(a), (b), and (c)), the curves representing the simulation results of both fBm and Poisson traffic are well situated within the scopes between the corresponding upper and lower bounds.

We can also note that the queue length distribution curves (i.e., analytical upper and lower bounds and simulation results) for both fBm and Poisson traffic appear in Fig. 2(a), corresponding to Case I. However, only the curves for the Poisson traffic can be plotted in Figs. 2(b) and (c), corresponding to Cases II and III, respectively. This phenomenon is due to the following reasons:
(a) Case I: $m = 90$ and $\lambda = 20$.  
(b) Case II: $m = 55$ and $\lambda = 55$.  
(c) Case III: $m = 20$ and $\lambda = 90$.  

Fig. 2. Comparison between analytical upper and lower bounds and simulation results of the queue length distribution with $H = 0.6$.  

- From the perspective of simulation, the arrival rate of the fBm traffic in Case I is considerably large, as compared to the service rate of the priority queuing system. Therefore, the queue of the fBm traffic could be non-empty even if such traffic is served with high priority. On the other hand, the arrival rates of the fBm traffic in Cases II and III are much smaller than the service rate of the queuing system. As a result, the queue of the fBm traffic is almost empty in simulation. Therefore, no empirical curve is obtained.  
- From the perspective of the analytical model, those curves corresponding to the upper and lower bounds in Cases II and III do exist. However, their values are so small as to exceed the scales of the vertical axis.  

5.2. Scenario 2: $H = 0.7$  

It is worth noting that the aforementioned observations obtained from Fig. 2 also hold in Fig. 3 which depicts the performance results for the larger Hurst parameter $H = 0.7$. Furthermore, a new sign can be found from this figure. As compared to Fig. 2, the scales of the horizontal axis representing queue lengths in Fig. 3 become larger for all three cases. That is to say, the probability of the larger queue length for both fBm and Poisson traffic increases as Hurst parameter increases. This phenomenon can be readily explained as follows: the larger Hurst parameter $H$, the higher the probability that traffic bursts are followed by each other. As a result, such extreme traffic burstiness spanning over many time scales gives rise to extended periods of large queue build-ups and also leads to long queue.  

Moreover, the scale of the horizontal axis in Fig. 3(a) is almost twice of that in Fig. 2(a), while the difference between the scales of the horizontal axis in Figs. 2(c) and 3(c) is around ten only. This phenomenon can be ascribed to the following reason. The high priority fBm traffic in Case I dominates the input of the queuing system, while in
Case III the low priority Poisson traffic dominates the input. Therefore, the change of the Hurst parameter in Case I causes the greater impact on simulation results and analytical bounds than that in Case III.

5.3. Scenario 3: $H = 0.8$

The phenomena we observed in Figs. 2 and 3 exist in Fig. 4 as well. Another interesting phenomenon, as illustrated by Figs. 4(a) and (b), is that the queue length distribution curves corresponding to simulation results in Cases I and II drop and fall below the analytical lower bounds at their right-hand ends. In principle, this phenomenon is due to the slow convergence of simulations subject to LRD self-similar traffic and will be discussed in detail in the next subsection.

5.4. Scenario 4: $H = 0.9$

Figure 5 reveals that the queue length distribution curves corresponding to simulation results are by and large situated below the scope of the analytical upper and lower bound curves for Cases I and II (see Figs. 5(a) and (b)), except their left-hand ends. In particular, these curves drop drastically at their right-hand ends. These phenomena are caused by the large Hurst parameter of self-similar traffic with which the convergence of simulations to a steady state becomes very slow, and even impossible under some extreme scenarios. This important observation highlights the significance of developing and using analytical models rather than adopting the simulation approach to quantitatively and accurately evaluating the effects of self-similar traffic on network performance.
However, it can be seen that the curve representing the simulation result in Case III (see Fig. 5(c)) is still well situated within the scope except its right-hand end. This is because the Poisson traffic dominates the input of the system under this case and thus the negative effect of large $H$ on the simulation is degraded.

5.5. Summary

It can be observed that for both fBm and Poisson traffic, their simulation results are well situated within the scopes of their corresponding analytical upper and lower bounds. This observation suggests that the developed analytical performance model has a good degree of accuracy in predicting the probability distributions of queue length in priority queuing systems under various working conditions. Moreover, the performance results reached in this section can be summarised as follows:

- A large Hurst parameter implies a high probability that the bursty arrivals of self-similar traffic are followed by each other. Thus, a large Hurst parameter gives rise to extended periods of queue build-ups and causes long queue. For this reason, as Hurst parameter $H$ increases from 0.6 in Scenario I to 0.9 in Scenario III, the maximum queue lengths of both fBm and Poisson traffic increase in an approximately exponential manner. For instance, the maximum queue length of Poisson traffic increases exponentially from less than 110 in Fig. 2(a) to around 3500 in Fig. 5(a).
- A large Hurst parameter leads to slow convergence of simulations to steady states. As a consequence, when Hurst parameter $H$ is small, it is relatively easy to characterize the queue length distribution. However, if Hurst
parameter $H$ is large, it becomes considerably difficult to obtain the accurate queue length distribution using a simulation approach.

- As compared to the service rate of the queuing system, the arrival rate of the self-similar traffic in Case I is considerably large. As a result, although the self-similar traffic is served with high priority, the queue of such traffic may not keep empty. Thus, it is possible to depict the curves representing the queue length distribution of the self-similar traffic in Case I (see subfigures (a) in Figs. 2–5). On the other hand, in Cases II and III the arrival rate of self-similar traffic is relatively small compared to the service rate. Therefore, the queue corresponding to high priority self-similar traffic keeps empty and no curves representing queuing length of self-similar traffic can be depicted in subfigures (b) and (c) of Figs. 2–5.

- In Case I the self-similar traffic dominates the input of the priority queuing system, while in Case III Poisson traffic dominates the input. As a consequence, the change of the Hurst parameter in Case I causes a greater impact on the accuracy of simulation results than in Case III.

6. Conclusions

As the rapid development of network-based applications, the provisioning of differentiated Quality-of-Service (QoS) has become a crucial issue for network design and performance analysis. In the DiffServ architecture, classified traffic flows are handled with priority scheduling policy. To the best of our knowledge, most existing work on the analysis of priority queuing systems has primarily considered homogeneous traffic only, i.e., either Short Range Dependent (SRD) or Long Range Dependent (LRD) traffic. Motivated by the heterogeneous nature of traffic within multi-service networks, this paper has presented an analytical model for priority queuing systems subject to LRD self-
similar and SRD Poisson traffic. More specifically, we have adopted and extended the generalized Schilder’s theorem to derive the analytical upper and lower bounds for the queue length distributions of individual traffic flows. Results obtained through simulation experiments have revealed that the model exhibits a good degree of accuracy in predicting the probability distributions of queue length under various working conditions. The validity and accuracy of the model make it a practical and cost-effective evaluation tool to study the performance behaviour of priority queuing systems with heterogeneous traffic under different parameter settings. The findings of performance analysis of this study can be summarised as follows:

- When the high priority LRD self-similar traffic dominates the input of the priority queuing system or it is comparable to the low priority SRD Poisson traffic, traffic Hurst parameter has a great impact on the accuracy of simulation results. More specifically, if the length of simulation time is fixed, a large Hurst parameter will debase its accuracy. This is because the convergence of simulations to a steady state becomes very slow with large Hurst parameter, and even impossible under some extreme scenarios. This finding highlights the importance of developing analytical models rather than adopting the simulation approach to quantitatively and accurately evaluating the effects of self-similar traffic on the performance of modern large-scale network-based systems.
- When the SRD Poisson traffic dominates the input of the priority queuing system, the developed upper and lower bounds of queue length distributions and simulation results are matched considerably well, even if a large Hurst parameter for its counterpart self-similar traffic is set.

Acknowledgments

We thank Dr. Petteri Mannersalo for his valuable suggestions. Thanks also go to Dr. Ton Dieker for providing us with some codes used for generating artificial fBm traces.

References


