Probability in the Engineering and Informational Sciences, **18**, 2004, 87–101. Printed in the U.S.A.

LOWER BOUNDS FOR LRD/GI/1 QUEUES WITH SUBEXPONENTIAL SERVICE TIMES

CATHY H. XIA, ZHEN LIU, MARK S. SQUILLANTE, AND LI ZHANG

IBM T.J. Watson Research Center
Yorktown Heights, NY 10598
E-mail: cathyx@us.ibm.com

We investigate the tail distribution of the virtual waiting times in a $LRD/GI/1$ queue where the arrival process is long-range dependent (LRD) and the service times are independent and identically distributed (i.i.d.) random variables. We present two lower bounds on the stationary waiting time tail asymptotics, which illustrate the different dominating components that influence server performance under various conditions. In particular, we show that the tail distribution of the stationary waiting time is bounded below by that of the associated $LRD/D/1$ queues resulting from replacing all random service times by the mean. This shows the performance impact purely due to the long-range dependency of the arrival process. On the other hand, when the service times are subexponential, we show that the tail distribution of the stationary waiting time is bounded below by that of the corresponding $D/GI/1$ queue by replacing the dependent arrival process with its associated independent version. This shows the minimum performance impact due to the tail distribution of the service times. The above two lower bounds indicate that the performance of $LRD/GI/1$ queues will be dominated by the heavier tail of the corresponding $LRD/D/1$ and $D/GI/1$ queues. These features are further illustrated and quantified through examples and via numerous simulation experiments.

1. INTRODUCTION

With the rapid advances in Internet technology, electronic commerce is becoming a mature business strategy. The concept of Quality of Service (QoS) is working its way to the front lines of electronic business commitments and requirements, as it

 0269-9648/04 $16.00

https://doi.org/10.1017/S0269964804181060 Published online by Cambridge University Press

plays an important role in Internet applications, services, and pricing negotiations. One needs to have a fundamental understanding of the key characteristics of the traffic patterns in commercial websites and a fundamental understanding of the impact of such traffic patterns on Web server performance, as well as the server capacity required to guarantee a certain level of QoS (e.g., request response time). Our primary focus in this article is to investigate these important research issues related to such performance metrics when the arrival traffic is long-range dependent and the service times are subexponential.

In the past decades, long-range dependence (LRD) has been detected in a wide range of applications and over multiple networking infrastructures [5,10,19]. Roughly speaking, the LRD process has a slowly decaying correlation structure that is nonintegrable. This dependence structure is expected to affect performance in a drastically negative way that is much different from that under the Poisson process. Previous work [7,13,18] in this area has been mostly focused on the dependent input process while assuming the service times of all arrivals are deterministic (i.e., $LRD/D/1$ queues or equivalently fluid queues with LRD input process). Although the assumption of deterministic service times is valid in the case of network routers serving fixed-size packets or is a reasonable approximation when the packet sizes have small variance, such an assumption is hardly justified in the case of Web servers where, in addition to the long-range dependency of the arrival process, requests vary largely in their sizes and service demands, often having nontraditional (subexponential) tail distributions [1,5,6,8,11,21]. Therefore, how the performance would be influenced by the LRD input process and the nontraditional service time distributions remains an open question, important both in theory and real practice, such as Web server performance and QoS concerns.

Performance impact due to subexponential service times (while assuming independent and identically distributed (i.i.d.) interarrival times) has been explored extensively over the past decades. A fundamental result (due to Pakes [14]) shows that for $GI/GI/1$ queues with i.i.d. interarrival times and i.i.d. service times, where it is assumed that the service times have distribution function (d.f.) F with finite mean μ^{-1} and traffic intensity $\rho < 1$, if the integrated (service time) tail distribution $\bar{F}_e$ is subexponential, then the stationary waiting time W_∞ is also subexponential and its tail distribution is asymptotically equivalent to $\bar{F}_e(x)$ as x goes to infinity. Therefore, for $GI/GI/1$ queues, when the residual service times are subexponential, the tail distribution of the service times will dominate the tail behavior of the stationary waiting times. The above result has later been generalized to Markov-modulated $G/GI/1$ queues [9] and short-range-dependent arrival processes [3].

Our goal in this article is to examine the performance asymptotics under both LRD arrival processes and i.i.d. subexponential service times. In particular, consider $LRD/GI/1$ queues, let $A(t)$ be the total number of arrivals in interval $[0, t)$, and denote by S_i the service demand by the ith arrival and by T_i the ith interarrival time. Throughout this article, we assume that the S_i's are i.i.d., and the sequence of T_i is stationary and ergodic.

https://doi.org/10.1017/S0269964804181060 Published online by Cambridge University Press

We are interested in the *stationary waiting time*, expressed in Loynes' schema as (cf. [12])

$$W_\infty \stackrel{d}{=} \left(\sup_{n \geq 1} \sum_{k=1}^{n} (S_{-k} - T_{-k}) \right)^+,$$

where $\stackrel{d}{=}$ is the equality in distribution and the sequence $\{S_k, T_k\}_{k=-\infty}^{\infty}$ is the stationary extension of the original sequence $\{S_k, T_k\}_{k=1}^{\infty}$.

We present two lower bounds on the stationary waiting time tail asymptotics, which illustrate the different dominating components that influence server performance under various conditions. Specifically, we show that for $LRD/GI/1$ queues, where the arrival process is long-range dependent and the service times are subexponential, the tail distribution of the waiting time W_∞ is asymptotically bounded below by $[\rho/(1-\rho)]\bar{F}_e(x)$, where $\bar{F}_e$ is the tail distribution of the stationary *residual* service times. Although this lower bound is known to be asymptotically exact for $GI/GI/1$ queues, this bound is the first, to the best of our knowledge, that extends to long-range dependent arrival processes. It indicates that the performance of any $G/GI/1$ queues with subexponential times is always bounded below by that of the associated $D/GI/1$ queue by replacing the (dependent) arrival process with its associated independent version. This shows the minimum performance impact due to the tail distribution of the service times.

In addition, we show that the tail distribution of the stationary waiting time of a $LRD/GI/1$ queue is always bounded below by that of the corresponding $LRD/D/1$ queue when replacing the random service times with its mean. This shows the minimum performance impact due to the long-range dependency of the input process.

The above two lower bounds further illuminate that the performance of $LRD/GI/1$ queues should be dominated by the heavier distribution between $\bar{F}_e$ and the tail distribution of W_∞ for the associated $LRD/D/1$ queues. When modeling the LRD process using the fractional Gaussian noise (FGN) with Hurst parameter H, the corresponding $FGN/D/1$ queue is well studied in [7,13], where it is known that the stationary waiting time is asymptotically equivalent to a Weibull distribution with shape parameter $2-2H$, denoted by Weibull$(2-2\mathrm{H})$. In this case, when the residual service time is heavier than Weibull$(2-2\mathrm{H})$, then the steady-state performance will be dominated by the residual service times. However, if the residual service time is lighter, then the steady-state performance is dominated by the dependence structure of the arrival process. The same results also hold for the stationary virtual waiting times. These characterizations are further illustrated and quantified via numerous simulation experiments.

The rest of the article is organized as follows. Definitions and some preliminaries are presented in Section 2. In Section 3, we show our main results, which give two important asymptotic lower bounds for the tail distribution of the stationary waiting time under general (dependent) arrival process and i.i.d. subexponential service times. These features are further illustrated and quantified via numerous

https://doi.org/10.1017/S0269964804181060 Published online by Cambridge University Press

simulation experiments in Section 4. Finally, concluding remarks are given in Section 5.

2. PRELIMINARIES

In this article, we will use the notation $a(x) \sim b(x)$ for $\lim_{x\to\infty}(a(x)/b(x)) = 1$, $a(x) \gtrsim b(x)$ for $\liminf_{x\to\infty}(a(x)/b(x)) \geq 1$, and $a(x) \lesssim b(x)$ for $\limsup_{x\to\infty}(a(x)/b(x)) \leq 1$.

For a given distribution function F on $[0,\infty\}$, designate F_e to be the integrated tail distribution of F so that

$$\bar{F}_e(x) = \mu \int_x^\infty \bar{F}(y)\, dy.$$

where $\mu = 1/\int_0^\infty \bar{F}(y)\, dy$.

DEFINITION 1: *A distribution function F on $[0,\infty)$ is said to be subexponential, denoted as $F \in \mathcal{S}$, if for any $n \geq 2$,*

$$\overline{F^{*n}}(x) \sim n\bar{F}(x),$$

where F^{*n} denotes the n-fold convolution of F.

The class of subexponential distributions was first introduced by Chistakov [4]. Intuitively, it says that the sum becomes large mainly due to one of the summands being large. Distribution functions in the subexponential family include Pareto distributions, log-normal distributions, and part of the Weibull distribution family. A Pareto distribution with shape parameter α and scale parameter β has d.f.

$$F(x) = 1 - (x/\beta)^{-\alpha}, \qquad x \geq \beta > 0, \alpha > 0. \tag{1}$$

A Weibull distribution with scale parameter α and shape parameter β has d.f.

$$F(x) = 1 - e^{-\alpha x^\beta}, \qquad x > 0. \tag{2}$$

It is subexponential when the shape parameter $\beta \in (0,1)$.

Note that $\mathcal{S}$ is a subset of $\mathcal{L}$, the class of *long-tailed* distributions such that $\bar{F}(x+y) \sim \bar{F}(x)$ as $x \to \infty$ for all y. It can be shown that any random variable (r.v.) X with d.f. $F \in \mathcal{L}$ would have an infinite moment generating function [17]: $E(e^{\theta X}) = \infty$, $\theta > 0$. We will therefore refer to a r.v. X as *light tailed* if $E(e^{\theta X}) < \infty$ for some $\theta > 0$.

Here, we quote some well-known properties that will be used later.

LEMMA 2: *Let F and G be d.f.'s on $[0,\infty)$ such that $\bar{F}(x) \sim c\bar{G}(x)$, with constant $c \in (0,\infty)$. Then, $F \in \mathcal{S}$ if and only if $G \in \mathcal{S}$.*

LEMMA 3: *If $G \in \mathcal{S}$ and F is any d.f. such that $\lim_{x\to\infty}(\bar{F}(x)/\bar{G}(x)) = c$, with constant $c \in [0,\infty)$, then*

$$\overline{F * G}(x) \sim (1+c)\bar{G}(x).$$

https://doi.org/10.1017/S0269964804181060 Published online by Cambridge University Press

3. ASYMPTOTIC LOWER BOUNDS

Consider a single queuing system where jobs arrive at random times $0 \le \Gamma_1 \le \Gamma_2 \le \cdots$, and the service times $\{S_n\}_{n\ge1}$ are i.i.d. random variables with d.f. F and finite mean μ^{-1}. Throughout this article, we assume that the service discipline is first-come first-serve (FCFS). Let $T_n = \Gamma_n - \Gamma_{n-1}$, $n = 1,2,\ldots$, be the interarrival times. We assume the arrival process is independent of the service times. Denote by $A(t)$ the cumulative number of arrivals in time $[0,t)$. We assume that process $A(t)$ is stationary and ergodic s.t. $\lim_{t\to\infty}(A(t)/t) = \lambda$. Note that the arrival process $A(t)$ could be dependent or even long-range dependent. We assume $\rho = \lambda/\mu < 1$ so that the system is stable. In addition, let $\{S_k, T_k\}_{k=-\infty}^{\infty}$ be the stationary extension of the original sequence $\{S_k, T_k\}_{k=1}^{\infty}$.

We now present two important asymptotic lower bounds for the tail distribution of the stationary waiting time W_∞. These bounds illuminate the different dominating components that influence server performance under different conditions, which shed light on the joint performance impact under both the long-range dependent arrival process and subexponential service times.

3.1. Lower Bound Based on Service Time Tail Distribution

In this subsection, we assume that the service times $S_n, n = 1,2,\ldots$ are i.i.d. subexponential. We consider a general $LRD/GI/1$ queuing system and show the minimum performance impact due to subexponential service times.

We will first present Lemma 4, which can be considered as an extended version of Lemma 3.1 in [2].

LEMMA 4: *Let D_n, $n = 1,2,\ldots$, be a sequence of random variables such that as $n \to \infty$, $D_n/n \to d$ with probability (w.p.)* 1. *For arbitrary $\epsilon, \epsilon' > 0$, there exists a constant $c > 0$ such that*

$$\boldsymbol{P}\left[\bigcap_{n\ge1}\{D_n \le n(d+\epsilon)+c\}\right] > 1-\epsilon', \tag{3}$$

and

$$\boldsymbol{P}\left[\bigcap_{n\ge1}\{D_n \ge n(d-\epsilon)-c\}\right] > 1-\epsilon'. \tag{4}$$

PROOF: Inequality (3) is basically the result of Lemma 3.1. in [2]. Noting that $-D_n/n \to -d$ w.p. 1, inequality (4) simply follows from applying (3) on the sequence $-D_n$. ■

THEOREM 5: *Consider a $LRD/GI/1$ queue with FCFS, where the input process is stationary and ergodic s.t. $\lim_{t\to\infty}(A(t)/t) = \lambda$. The service times are subexponential i.i.d. r.v.'s with mean μ^{-1} and d.f. F. Set $\rho = \lambda/\mu$ and assume $\rho < 1$. If $F_e \in \mathcal{S}$, then*

https://doi.org/10.1017/S0269964804181060 Published online by Cambridge University Press

$$\boldsymbol{P}[W_\infty > x] \geqslant \frac{\rho}{1-\rho}\,\bar{F}_e(x). \tag{5}$$

PROOF: Because the input process is stationary and ergodic, it follows immediately that $(\sum_{i=1}^{n} T_i)/n \to \lambda^{-1}$ w.p. 1. Applying Lemma 4 yields that for arbitrary $\epsilon, \epsilon' > 0$, there exists a constant $c > 0$ such that $\boldsymbol{P}[B] > 1 - \epsilon'$, where B denotes the event set $B := \bigcap_{n\geq 1}\{\sum_{k=1}^{n} T_{-k} \leq n(\lambda^{-1} + \epsilon) + c\}$. We then have

$$\begin{aligned}
\boldsymbol{P}[W_\infty > x] &\geq \boldsymbol{P}[B]\boldsymbol{P}\left[\sup_{n\geq 1}\left\{\sum_{k=1}^{n} S_{-k} - \sum_{k=1}^{n} T_{-k}\right\} > x \,|\, B\right] \\
&\geq (1-\epsilon')\boldsymbol{P}\left[\sup_{n\geq 1}\left\{\sum_{k=1}^{n}\left[S_{-k} - (\lambda^{-1}+\epsilon)\right]\right\} > x + c \,|\, B\right] \\
&= (1-\epsilon')\boldsymbol{P}\left[\sup_{n\geq 1}\left\{\sum_{k=1}^{n}\left[S_{-k} - (\lambda^{-1}+\epsilon)\right]\right\} > x + c\right] \\
&\sim (1-\epsilon')\,\frac{\mu^{-1}}{\lambda^{-1}+\epsilon-\mu^{-1}}\,\bar{F}_e(x+c),
\end{aligned}$$

where the last equality comes from the fact that the arrival process (thus, event B) is independent of the service times. The last $\sim$ equivalence is due to [14] for $D/GI/1$ queues with deterministic interarrival times $\lambda^{-1} + \epsilon$. Since $F_e \in \mathcal{S} \subset \mathcal{L}$, $\bar{F}_e(x + c) \sim \bar{F}_e(x)$, (5) follows by letting ϵ and ϵ' go to zero. ■

Note that the above argument applies for any (whether dependent or independent) stationary ergodic arrival process. Thus, from Theorem 5, we conclude that under any stationary ergodic arrival process, whenever the service times are subexponential, the tail distribution of the stationary waiting time will be at least as heavy as that of the residual service times.

Remark: Although it is well known that for $GI/GI/1$ queues where the arrival process is of renewal type and the service times are subexponential, (5) is asymptotically exact [14]; the above lower bound is the first, to the best of our knowledge, that extends to long-range-dependent arrival processes.

3.2. Lower Bound Based on Dependence Structure

We next derive a lower bound for the tail distribution of the stationary waiting time which shows the performance impact purely due to the long-range-dependent structure of the arrival process.

THEOREM 6: *Consider a $LRD/GI/1$ queue with FCFS, where the input process is stationary and ergodic s.t.* $\lim_{t\to\infty}(A(t)/t) = \lambda$. *The service times are subexponential i.i.d. random variables with mean μ^{-1} and d.f. F. Set $\rho = \lambda/\mu$ and assume $\rho < 1$. Let $W_\infty^{LRD/D/1}$ be the stationary waiting time of the associated $LRD/D/1$*

https://doi.org/10.1017/S0269964804181060 Published online by Cambridge University Press

queues where the arrival process is the same but service times are deterministic and equal to μ^{-1}. *If for all sufficiently small* $\epsilon > 0$,

$$W_\infty^{LRD/D_{-\epsilon}/1} \in \mathcal{L}, \tag{6}$$

where $D_{-\epsilon}$ *denotes deterministic service times equal to* $\mu^{-1} - \epsilon$, *then*

$$\boldsymbol{P}[W_\infty > x] \geqslant \boldsymbol{P}[W_\infty^{LRD/D/1} > x]. \tag{7}$$

PROOF: Because S_k's are i.i.d. random variables, by the strong law of large numbers, $(\sum_{k=1}^n S_{-k})/n \to \mu^{-1}$ w.p. 1. Applying Lemma 4 entails for arbitrary $\epsilon, \epsilon' > 0$, there exists a constant $c > 0$ such that $\boldsymbol{P}[B'] > 1 - \epsilon'$, where B' denotes the event set $B' := \bigcap_{n\geq 1}\{\sum_{k=1}^n S_{-k} \geq n(\mu^{-1} - \epsilon) - c\}$. Therefore,

$$\begin{aligned}
\boldsymbol{P}[W_\infty > x] &\geq \boldsymbol{P}[B']\boldsymbol{P}\left[\sup_{n\geq 1}\left\{\sum_{k=1}^n S_{-k} - \sum_{k=1}^n T_{-k}\right\} > x | B'\right] \\
&\geq (1-\epsilon')\boldsymbol{P}\left[\sup_{n\geq 1}\left\{\sum_{k=1}^n [(\mu^{-1} - \epsilon) - T_{-k}]\right\} > x + c | B'\right] \\
&= (1-\epsilon')\boldsymbol{P}\left[\sup_{n\geq 1}\left\{\sum_{k=1}^n [(\mu^{-1} - \epsilon) - T_{-k}]\right\} > x + c\right] \\
&= (1-\epsilon')\boldsymbol{P}[W_\infty^{LRD/D_{-\epsilon}/1} > x + c],
\end{aligned}$$

where the last equality follows from the fact that the service times (thus, event B') is independent of the arrival process. Since $W_\infty^{LRD/D_{-\epsilon}/1} \in \mathcal{L}$, $\boldsymbol{P}[W_\infty^{LRD/D_{-\epsilon}/1} > x + c] \sim \boldsymbol{P}[W_\infty^{LRD/D_{-\epsilon}/1} > x]$. We then have that (7) follows by letting ϵ and ϵ' go to zero. ■

Theorem 6 shows that under both long-range-dependent arrival process and subexponential service times, the tail distribution of the stationary waiting time will be at least as heavy as that resulting from the dependence structure of the arrival process alone (i.e., as if with deterministic service times).

3.3. Example: Fractional Gaussian Noise Input Traffic

In this subsection, we concentrate on the case when the arrival traffic is FGN. The FGN model is frequently used to capture the long-range dependency of traffic mostly due to its mathematical simplicity. The use of FGN for traffic modeling is discussed in [15] and references therein.

Suppose requests arrive at discrete times $t = 0,1,2,\ldots$, where the number of arrivals at time slot t is denoted by integer a_t. Assume that $\{a_t\}$ is a stationary FGN sequence with mean λ, variance σ^2, and Hurst parameter $H \in [\frac{1}{2},1)$. In other words, $a_t = \lambda + \sigma N_t^H$, where $\{N_t^H\}$ is a zero-mean standard (fraction) Gaussian sequence with (auto)covariance function

https://doi.org/10.1017/S0269964804181060 Published online by Cambridge University Press

$$\Gamma_H(k) = \frac{1}{2}\left(|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H}\right).$$

Consider a *FGN/GI/*1 queue with arrival process $\{a_t\}$; the service times for the requests S_i's are i.i.d. random variables with finite mean μ^{-1} and general distribution G. Since jobs arrive in batches (at the beginning of each slot), here we are interested in the stationary waiting time of the first job in a batch, which can be expressed in Loynes' schema as follows:

$$W_\infty^1 \stackrel{d}{=} \left(\sup_{t=0,1,2,\ldots}(U^r(t) - t)\right)^+,$$

where $U^r(t) = \sum_{i=1}^{A^r(s)} S_{-i}$, $A^r(t) = \sum_{s=0}^{t} a_{-s}$ and the sequences $\{a_t\}_{t=-\infty}^{\infty}$ and $\{S_i\}_{i=-\infty}^{\infty}$ are respectively stationary extensions of the sequences $\{a_t\}_{t=0}^{\infty}$ and $\{S_i\}_{i=0}^{\infty}$.

Let $\mu_\epsilon^{-1} = \mu^{-1} - \epsilon$. As in the proof of Theorem 6, one can show that

$$\begin{aligned}\boldsymbol{P}[W_\infty^1 > x] &= \boldsymbol{P}\left[\sup_{t\in\mathbb{N}}\left\{\sum_{j=1}^{A^r(t)} S_{-j} - t\right\} > x\right] \\ &\geq (1-\epsilon')\boldsymbol{P}\left[\sup_{t\in\mathbb{N}}\{A^r(t)\mu_\epsilon^{-1} - t\} > x + c\right].\end{aligned}$$

Note that the last component corresponds to a fluid queue fed by LRD inputs, which has been studied in [13] assuming $\mu_\epsilon = 1$, and it is known that for large x,

$$\boldsymbol{P}\left[\sup_{t\geq 0}\{A^r(t) - t\} > x\right] \geqslant \exp\left(-\left[\frac{1}{2\sigma^2(1-H)^2}\left(\frac{(1-\lambda)(1-H)}{H}\right)^{2H}\right]x^{2-2H}\right). \tag{8}$$

This lower bound has been shown in [7] to be asymptotically exact in log scale.

Therefore, $\sup_{t\geq 0}(A^r(t)\mu_\epsilon^{-1} - t)$ has its tail distribution asymptotically equivalent to a Weibull distribution with shape parameter $2-2H$, which is clearly long tailed, thus satisfying condition 6.

As the waiting times of other customers in a batch are larger than that of the first customer, we obtain the lower bound of the tail distribution of the stationary waiting time.

COROLLARY 7: *Consider a FGN/GI/1 queue, where the arrival process is FGN and the service times are i.i.d. r.v.s with mean* μ^{-1}. *Then, for large x,*

$$\boldsymbol{P}[W_\infty > x] \geqslant e^{-\delta x^\beta}, \tag{9}$$

where

$$\beta = 2 - 2H, \qquad \delta = \frac{1}{2\rho^2\gamma^2(1-H)^2}\left(\frac{(1-\rho)(1-H)}{H}\right)^{2H}, \tag{10}$$

with $\rho = \lambda/c\mu$ *and* $\gamma^2 = \sigma^2/\lambda^2$. *Note that* γ^2 *is simply the coefficient of variance of the arrival process.*

https://doi.org/10.1017/S0269964804181060 Published online by Cambridge University Press

Clearly, lower bound (9) only shows the impact on performance by the long-range-dependence arrival process.

3.4. Further Insights

The results from previous subsections show that the joint impact on performance by a long-range dependent arrival process and subexponential would be bounded below by that of the associated queues when replacing the (dependent) arrival process with its independent version, and by the tail distribution of the residual service times. Thus, the heavier tail of (9) and (5) dominates. We then have the following corollary.

COROLLARY 8: *For a LRD/GI/1 queue, where the input process is LRD and the service times are i.i.d. subexponential with finite mean μ^{-1}, then for large x, the tail distribution of the stationary waiting time is bounded below by the heavier tail of (5) and (7).*

COROLLARY 9: *Consider FGN/GI/1 queues as introduced in Section 3.3, The service times are i.i.d. subexponential with finite mean μ^{-1}. Then, the lower bound of the stationary waiting time will be dominated by the heavier tail of the residual service time and Weibull distribution with shape parameter $2-2H$ and scale parameter δ, which is given by (10).*

In fact, similar results hold for the *stationary virtual waiting time*

$$V_\infty \stackrel{d}{=} \left(\sup_{t\in\mathbb{R}^+} \left(\sum_{i=1}^{A^r(t)} S_{-i} - t \right) \right)^+.$$

This is because of the following stochastic comparisons between stationary virtual waiting time and stationary waiting time:

$$W_\infty \leq_{\text{st}} V_\infty \leq_{\text{st}} W_\infty + S_e, \tag{11}$$

where $S_e \sim F_e$ and is independent of W_∞. The stochastic inequality $X \leq_{\text{st}} Y$ means that $\boldsymbol{P}[X > x] \leq \boldsymbol{P}[Y > x]$ for all $x \in \mathbb{R}$.

In order to show (11), let $t_n = \sum_{k=1}^n T_{-k}$ for $n \geq 1$. Note that

$$\begin{aligned} V_\infty &\stackrel{d}{=} \left(\sup_{t\geq 0} [U^r(t) - t] \right)^+ \geq \left(\sup_{n\geq 1} [U^r(t_n) - t_n] \right)^+ \\ &= \left(\sup_{n\geq 1} \sum_{k=1}^n [S_{-k} - T_{-k}] \right)^+ \stackrel{d}{=} W_\infty. \end{aligned}$$

Thus, the left-hand side of (11) follows immediately.

The right-hand side of (11) follows from the known fact [2] that for general $G/GI/1$ queues with FCFS and $\rho < 1$,

$$\boldsymbol{P}[V_\infty > x] = \rho \boldsymbol{P}[W_\infty + S_e > x], \tag{12}$$

for all x, where S_e is distributed as F_e and is independent of W_∞.

https://doi.org/10.1017/S0269964804181060 Published online by Cambridge University Press

We immediately have the following corollary.

COROLLARY 10: *The two lower bounds given by (5) and (9) hold asymptotically for the stationary virtual waiting time V_∞. In addition, the heavier one of both the stationary waiting times W_∞ dominates.*

Therefore, when (5) is heavier than (9), the tail distributions of the stationary waiting time and virtual waiting time are dominated by the subexponential service times; that is, the impact by the dependence structure of the input process becomes relatively small. If instead, the lower bound (9) is heavier than (5), then the tail distributions of the stationary waiting time and virtual waiting time are dominated by the dependence structure. These features are further illustrated by the simulation studies in the next section.

We conjecture that the lower bound of Corollary 8 actually provides an asymptotically exact solution.

4. COMPARISON WITH SIMULATIONS

To compare with the theoretical bounds developed in the last section, we simulate the system under generated self-similar input sequence and explore the performance under different service time distributions. The self-similar input process is generated using the drill-down techniques introduced in [21] using the FGN model, which is based on the fast Fourier transform algorithm proposed in [16]. Specifically, we set the parameters of the FGN model to be $H = 0.7505$, $\lambda = 30.1471$ and $\sigma^2 = 86.6037$, which are obtained from a real set of commercial website data dated August 9, 2001, hour 16–21.2. The service times are generated (based on the inversion formula) under different distributions. In all cases, we assume the traffic intensity equals $\rho = 0.8$. The resulting empirical stationary waiting time tail distributions $\boldsymbol{P}[W_\infty > x]$ are then calculated and plotted. Note that $W_\infty \geq_{\text{st}} V_\infty$; we further compare $\boldsymbol{P}[W_\infty > x]$ with the lower bounds developed in Section 3. We denote the lower bound (5) simply as Poisson since it corresponds to the performance under the same service times but with Poisson input process. Therefore, it shows the performance impact due to the subexponential service times. Similarly, we denote as LB-H the lower bound given by (9) since it shows the performance impact under the long-range-dependent structure. Specifically, LB-H is set to $\rho e^{-\delta x^{2-2H}}$ for $x > 0$ since $P(W > 0) = \rho$.

Simulation results for exponential service times are plotted in Figure 1, where in (a) the tail probability of the waiting times $P[W > x]$ is plotted as a function of x, and in (b) the $\log P[W > x]$ versus x plot is displayed. Note that when the arrival process is Poisson, the resulting queue is a $M/M/1$ queue, where its steady-state waiting time has a known exponential tail distribution [20] such that $\boldsymbol{P}[W > x] = \rho e^{-\mu(1-\rho)x}$ for $x > 0$. Clearly, empirical tail probabilities obtained from simulation stay very close to the lower bound LB-H and far from the tail distribution under Poisson inputs. This suggests that when the service times have exponential tails, the performance is es-

https://doi.org/10.1017/S0269964804181060 Published online by Cambridge University Press

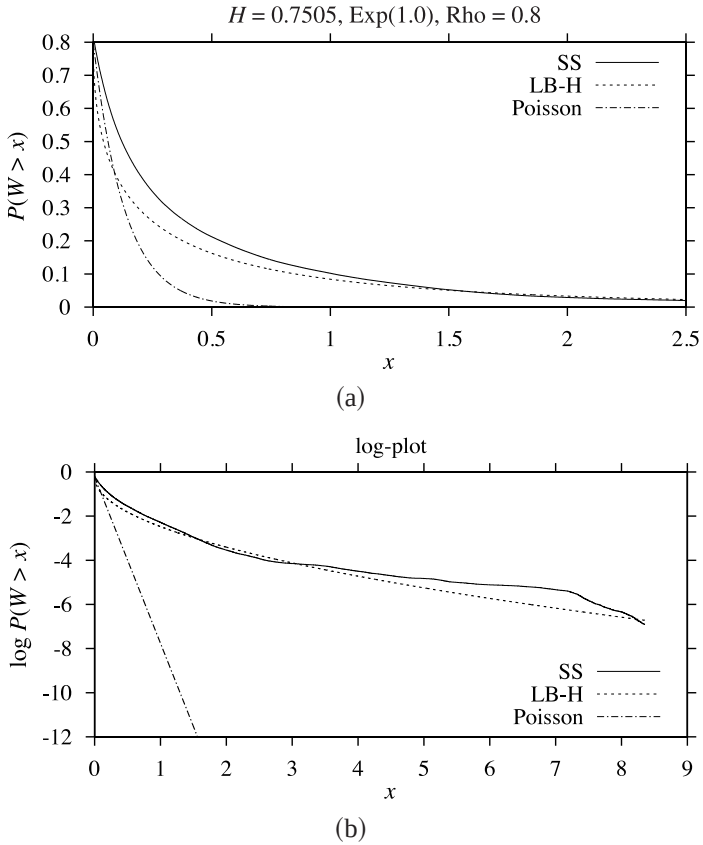


FIGURE 1. (a) Waiting time tail probabilities under exponential service times with self-similar arrival process. (b) The corresponding plot in log scale.

sentially dominated by the long-range dependence and the steady-state waiting time is asymptotically equivalent to a Weibull distribution.

Figures 2 and 3 show the simulation results when the service times are of the Weibull distributions given by (2), where two cases were considered, with $\beta = 0.8$ and $\beta = 0.2$, respectively; the corresponding parameter α is chosen so that the mean service time[1] is equal to ρ/m with $\rho = 0.8$. Note that in both cases the service times are subexponential, however, Weibull(0.8) is lighter than LB-H which is Weibull$(2-2\mathrm{H})$ with $H = 0.7505$, while Weibull(0.2) is heavier.

Observe that under Weibull(0.8) service times, the tail distribution under self-similar (SS) inputs is very close to the lower bound LB-H, suggesting that the dependence structure dominates the performance. However, when the service times

[1]The mean of a Weibull r.v. with parameter (α, β) is $\alpha^{-1/\beta}\Gamma(1 + 1/\beta)$, where $\Gamma(y) = \int_0^\infty e^{-x}x^{y-1}\,dx$.

https://doi.org/10.1017/S0269964804181060 Published online by Cambridge University Press

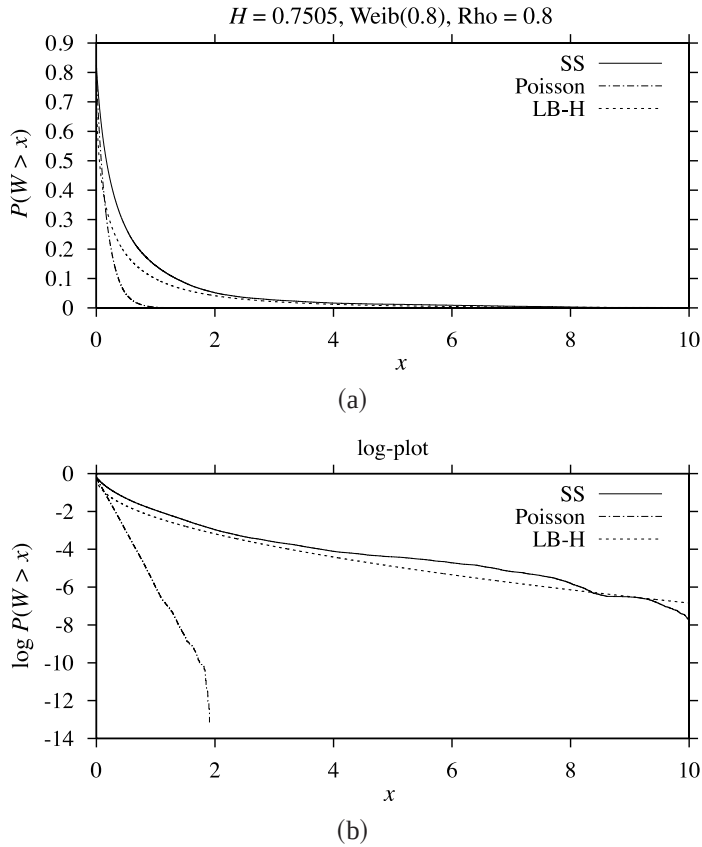


FIGURE 2. (a) Waiting time tail probabilities under Weibull($\beta = 0.8$) service times. (b) The corresponding log plot.

are Weibull(0.2), the performance deviates greatly from LB-H and stays closer to that under the Poisson inputs, suggesting that the service times dominated the performance instead when the service time tail distribution is heavier than Weibull(2−2H).

Figure 4 corresponds to service times of the Pareto distribution given by (1), where $\alpha = 1.5$ and β is chosen so that the mean service time[2] is equal to ρ/m with traffic intensity $\rho = 0.8$. In this case, the Pareto variable is heavy tailed with infinite variance. In Figure 4, the tail probabilities of the waiting times under SS inputs is compared with that under Poisson arrival process (refer to, e.g., [9,14]) and the two lower bounds developed in Section 3. Observe that the tail asymptotics under SS inputs significantly deviates from LB-H while staying very close to that under Poisson inputs. This suggests that under heavy-tailed service times, the impact on performance by the dependence structure of the input process becomes minor and the

[2]The mean of a Pareto r.v. with parameter (α, β) is $\beta\alpha/(\alpha - 1)$.

https://doi.org/10.1017/S0269964804181060 Published online by Cambridge University Press

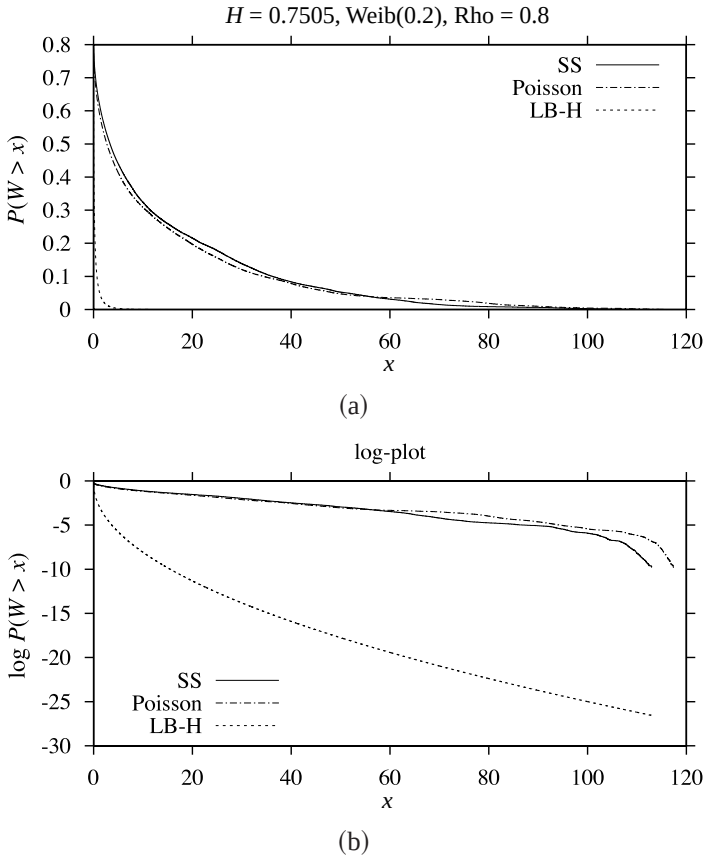


FIGURE 3. (a) Waiting time tail probabilities under Weibull ($\beta = 0.2$) service times. (b) The corresponding log plot.

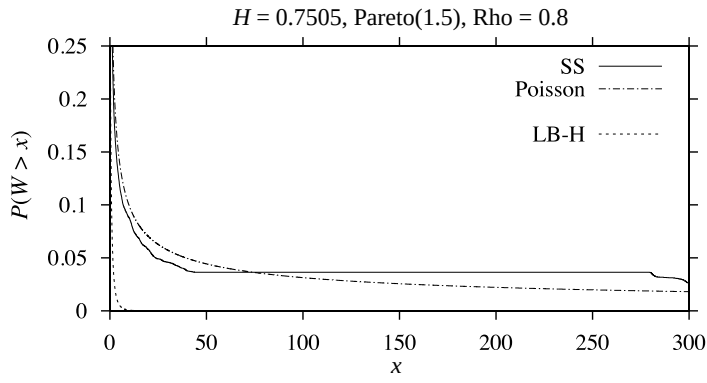
tail behavior is essentially dominated by the tail behavior of the service time distribution.

Informally, we can summarize the performance issues as follows. Under FGN arrival process and exponential tailed service times, the waiting time tail distribution is dominated by the dependence structure of the arrival process. When the service times are subexponential, if the residual service time (viz. lower bound (5)) is heavier than Weibull$(2-2H)$, then the waiting time distribution is dominated largely by the service time tail properties; otherwise, the dependence structure dominates the performance.

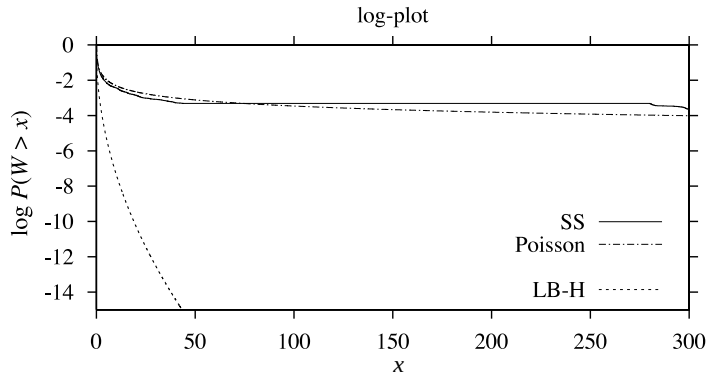
5. CONCLUSIONS

Based on both analytical results and simulations, we investigated various issues concerning the joint impact on the asymptotic behavior of the stationary waiting

https://doi.org/10.1017/S0269964804181060 Published online by Cambridge University Press



(a)



(b)

FIGURE 4. (a) Waiting time tail probabilities under Pareto service times with $\alpha = 1.5$. (b) The corresponding plot in log scale.

time and virtual waiting time by the long-range-dependent arrival process and subexponential service times. We presented two important lower bounds on the stationary virtual waiting time tail asymptotics, which illuminates the different dominating components that influence server performance under various conditions. In particular, we showed that the tail distributions of the stationary waiting time and virtual waiting time of an $LRD/GI/1$ queue with subexponential service times are bounded below by that of the associated $D/GI/1$ queue by replacing the dependent arrival process with its associated independent version. This shows the performance impact purely due to the tail distribution of the service times. In addition, tail distributions of the stationary waiting time and virtual waiting time are also bounded below by that of the corresponding $LRD/D/1$ queues, which shows the performance impact [13] purely due to the long-range dependency of the arrival process when replacing the random service times with its mean. These features are further illustrated and quantified via numerous simulation experiments.

https://doi.org/10.1017/S0269964804181060 Published online by Cambridge University Press

Although the analytical solutions provide only asymptotic lower bounds for the tail distribution of the response times, we believe that these bounds are asymptotically exact. This is the subject of our ongoing investigation.

Acknowledgment

The authors would like to thank the reviewer for pointing out some omissions in the original proofs and the helpful comments and suggestions.

References

1. Arlitt, M.F. & Williamson, C.L. (1997). Internet Web servers: Workload characterization and performance implications. *IEEE/ACM Transactions on Networking* 5(5): 631–645.
2. Asmussen, S. (1998). Subexponential asymptotics for stochastic processes: Extremeal behaviour, stationary distributions and first passage probabilities. *Annals of Applied Probability* 8: 354–374.
3. Asmussen, S., Schmidli, H., & Schmidt, V. (1999). Tail probabilities for non-standard risk and queueing processes with subexponential jumps. *Advances in Applied Probability* 31: 422–447.
4. Chistakov, V.P. (1964). A theorem on sums of independent positive random variables and its application to branching random process. *Theory of Probability and Its Applications* 9: 640–648.
5. Corvella, M.E. & Bestavros, A. (1996). Self-similarity in World Wide Web traffic: Evidence and possible causes. *Performance Evaluation Review* 24: 160–169.
6. Downey, A. (2001). The structural cause of file size distributions. *Proceedings of the International Symposium on Modeling Analysis and Simulation of Computer and Telecommunication Systems*.
7. Duffield, N.G. & O'Connell, N. (1995). Large deviation and overflow probabilities for the general single-server queue, with applications. *Mathematical Proceedings of the Cambridge Philosophical Society* 118: 363–375.
8. Iyengar, A.K., Squillante, M.S., & Zhang, L. (1999). Analysis and characterization of large-scale web server access patterns and performance. *World Wide Web* 2: 85–100.
9. Jelenkovic, P. & Lazar, A.A. (1998). Subexponential asymptotics of a Markov-modulated random walk with queueing applications. *Journal of Applied Probility* 35(2): 325–347.
10. Leland, W.E., Taqqu, M.S., Willinger, W., & Wilson, D.V. (1994). On the self-similarity nature of Ethernet traffic (extended version). *IEEE/ACM Transactions on Networking* 2(1): 1–15.
11. Liu, Z., Niclausse, N., & Jalpa-Villanueva, C. (2001). Traffic model and performance evaluation of Web servers. *Performance Evaluation* 46(2–3): 77–100.
12. Loynes, R.M. (1968). The stability of a queue with non-independent inter-arrival and service times. *Proceedings of Cambridge Philosophical Society* 58: 497–520.
13. Norros, I. (1994). A storage model with self-similar input. *Queueing Systems* 16: 387-396.
14. Pakes, A. (1975). On the tails of waiting time distributions. *Journal of Applied Probability* 12: 555–564.
15. Park, K. & Willinger, W. (eds.). (2002). *Self-similar network traffic and performance evaluation*. New York: Wiley.
16. Paxson, V. (1995). Fast approximation of self-similar network traffic. Technical Report LBL-36750/UC-405, Lawrence Berkeley Laboratory.
17. Sigman, K. (1999). Appendix: A primer on heavy-tailed distributions. *Queueing Systems* 33: 261–275.
18. Vanichpun, S. & Makowski, A.M. (2002). Positive correlations and buffer occupancy: Lower bounds via supermodular ordering. *Proceedings of IEEE INFOCOM 2002*, pp. 1298–1306.
19. Willinger, W., Taqqu, M.S., Sherman, R., & Wilson, D.V. (1997). Self-similarity through high-variability: Statistical analysis of ethernet LAN traffic at the source level. *IEEE/ACM Transactions on Networking* 5(1): 71–86.
20. Wolff, R.W. (1988). *Stochastic modeling and the theory of queues*. Englewood Cliffs, NJ: Prentice-Hall.
21. Xia, C.H., Liu, Z., Squillante, M.S., Zhang, L., & Malouch, N. (2003). Analysis of performance impact of drill-down techniques for Web traffic models. *Proceedings of the 18th International Teletraffic Congress* (ITC-18), pp. 1–10.