

Queueing Systems with Long-Range Dependent Input Process and Subexponential Service Times *

Cathy H. Xia

IBM T.J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598, USA

cathyx@us.ibm.com

Zhen Liu

IBM T.J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598, USA

zhenl@us.ibm.com

ABSTRACT

We analyze the asymptotic tail distribution of stationary waiting times and stationary virtual waiting times in a single-server queue with long-range dependent arrival process and subexponential service times. We investigate the joint impact of the long range dependency of the arrival process and of the tail distribution of the service times. We consider two traffic models that have been widely used to characterize the long-range dependence structure, namely, the $M/G/\infty$ input model and the Fractional Gaussian Noise (FGN) model. We focus on the response times of the customers in a First-Come First-Serve (FCFS) queueing system, although the results carry through to the backlog distribution of the system with any arbitrary queueing discipline. When the arrival process is driven by an $M/G/\infty$ input model we show that if the residual service time tail distribution F_e is lighter than the residual session duration G_e , then the stationary waiting time is dominated by the long-range dependence structure, which is determined by the residual session duration G_e . If the residual service time distribution F_e is heavier than the residual session duration G_e , then the tail distribution of the stationary waiting time is dominated by that of the residual service time. When the arrival process is modeled by an FGN, we show that the waiting time tail distribution

*(Produces the permission block, copyright information and page numbering). For use with ACM_PROC_ARTICL E-SP.CLS V2.6SP. Supported by ACM.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SIGMETRICS'03, June 10–14, 2003, San Diego, California, USA.

Copyright 2003 ACM 1-58113-664-1/03/0006 ...\$5.00.

is asymptotically equal to the tail distribution of the residual service time if the latter is asymptotically heavier than Weibull distribution with shape parameter $2-2H$, where H is the Hurst parameter of the FGN. If, however, this residual service time is asymptotically lighter than Weibull distribution with shape parameter $2-2H$, then the waiting time tail distribution is dominated by the dependence structure of the arrival process so that it is asymptotically equal to Weibull distribution with shape parameter $2-2H$.

Categories and Subject Descriptors

G.3 [Probability and Statistics]: Queueing Theory

General Terms

Performance, Theory

Keywords

long-range dependency, subexponential distributions, $M/G/\infty$, FGN, asymptotic queueing analysis

1. INTRODUCTION

With more and more statistical evidence of long-range dependence (LRD) observed in communication networks (see, e.g. [21, 36]) and in Web servers (see, e.g. [10, 17, 38]), it has become increasingly important to understand the performance impact by traffic exhibiting long-range dependence structure. Previous work in this area has been mostly focused on the dependence structure while assuming deterministic service times (e.g. the so called fluid queues with LRD inputs). Two traffic input models, the $M/G/\infty$ model of Cox [11], and the Fractional Gaussian Noise (FGN) model, have been used extensively to model the long-range dependence structure. Studies using $M/G/\infty$ input model include

[18, 31, 29, 16, 22]; while studies using FGN input model include [26, 13, 40].

While 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 non-traditional (subexponential) tail distributions, [1, 10, 12, 17, 23]. Therefore, how would the performance be influenced by the LRD input process and the non-traditional service time distributions remains an open question, important both in theory and practice such as Web server performance and QoS concerns.

Our goal in this paper is to examine the asymptotic behavior of the response time tail distributions under both LRD arrival processes and general (light-tailed or subexponential) service times. That is, we consider LRD/GI/1 queues, where the arrival process is long-range dependent (LRD) and the service times are independent and identically distributed (i.i.d.) random variables, independent of the arrival process. We focus on the response times of the customers in a First-Come First-Serve (FCFS) queueing system, although the results carry through to the backlog distribution of the system with any arbitrary queueing discipline. It is worth to mention that the asymptotic behavior of response time tail distribution under other service disciplines is also a very rich subject. Studies on performance under other service disciplines such as processor sharing or generalized processor sharing can be referred to e.g. [4, 20, 41] and references therein. We shall only focus on FCFS service discipline in this paper.

Our study pertains on two models that are widely used to characterize LRD input processes, namely, the $M/G/\infty$ input model and the FGN model. We present a systematic study on the asymptotic tail distribution of the stationary waiting time which illuminates the different dominating components that influence server performance under various conditions.

Performance impact due to subexponential service times (while assuming i.i.d. interarrival times) has been explored extensively over the past decades. A fundamental result (due to Pakes [27]) shows that, for $GI/GI/1$ queues with service time distribution (d.f.) F and finite mean μ^{-1} , and traffic intensity $\rho < 1$, if the integrated (service-time) tail distribution \bar{F}_e is subexponential, then the stationary wait-

ing time W_∞ is also subexponential, and

$$\mathbf{P}[W_\infty > x] \sim \frac{\rho}{1-\rho} \bar{F}_e(x), \quad (1)$$

where F_e , the integrated tail distribution of F , is defined by

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

Note that \bar{F}_e can also be viewed as the tail distribution of the stationary *residual* service times. Therefore, in 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 [19], and for short range dependent arrival processes [2]. In particular, when the arrival process is stationary and ergodic, but can otherwise be dependent, [2] provided a sufficient condition under which (1) still holds. It was also shown that the condition is automatically satisfied when the arriving process is short range dependent, such as the stationary autoregressive process which coincides with the results developed in [18].

Our results in the paper further extend the work of [2]. We show that under long-range dependent arrival process and subexponential service times, the performance is not always dominated by the service times as in (1), and there are situations where the long-range dependence structure could dominate the performance.

When the arrival process is driven by an $M/G/\infty$ model, the so-called infinite on-off source Poisson model with session (i.e. on-period) duration distribution G , we show that the asymptotic tail distribution of the response times is dominated by the heavier one of the residual service time and the residual session duration. That is, the waiting time distribution is asymptotically equal to the tail distribution of the residual service time if the latter is heavier than that of the residual session duration time. If instead the residual service time is asymptotically lighter than the residual session duration G_e , then the long-range dependence structure dominates and the waiting time is asymptotically equal to the residual session duration G_e . Note that these results are compatible with those of [25], where this kind of asymptotic dominance was found for a centered process under some suitable scaling in an infinite-source Poisson network model.

When the arrival process is modeled by an FGN, we show that (1) holds if the residual service time is asymptotically heavier than Weibull distribution with shape parameter $2-2H$, denoted as Weibull(2-2H), where H is the Hurst parameter of the FGN; that is, the waiting time tail distribution

is dominated by the tail distribution of the residual service time. If, however, this residual service time is asymptotically lighter than Weibull(2-2H), then the waiting time tail distribution is dominated by the dependence structure of the arrival process so that it is asymptotically equal (in log scale) to Weibull(2-2H). These results further extended the preliminary lower bounds developed in [39].

The rest of the paper is organized as follows. Preliminaries and some key properties and new developments for subexponential functions are presented in Section 2. In Section 3, we derive an asymptotic lower bound for stationary waiting time tail distribution under general (dependent) arrival process and i.i.d. subexponential service times. We then present our in-depth asymptotic analysis of the stationary waiting time tail behavior in Section 4 and Section 5, where Section 4 focuses on M/G/ ∞ inputs, and Section 5 focuses specifically on FGN inputs, Finally, concluding remarks are provided in Section 6.

2. PRELIMINARIES

We say that the random variable X is stochastically smaller than the random variable Y , written $X \leq_{\text{st}} Y$, if $\mathbf{P}[X > a] \leq \mathbf{P}[Y > a]$, for all a .

Given a *nonnegative* random variable X with distribution function (d.f.) F on $[0, \infty)$. We denote the tail distribution by $\overline{F}(x) = 1 - F(x)$. A d.f. G is said to have a ‘‘lighter’’ tail than d.f. F (or, equivalently, F has a ‘‘heavier’’ tail than G), if $\lim_{x \rightarrow \infty} \frac{\overline{G}(x)}{\overline{F}(x)} = 0$. Throughout the paper we use the notation $a(x) \sim b(x)$ as $x \rightarrow \infty$ to denote $\lim_{x \rightarrow \infty} \frac{a(x)}{b(x)} = 1$; we also use $a(x) \succeq b(x)$ for $\limsup_{x \rightarrow \infty} \frac{a(x)}{b(x)} \geq 1$, and $a(x) \preceq b(x)$ for $\liminf_{x \rightarrow \infty} \frac{a(x)}{b(x)} \leq 1$.

A distribution F on $[0, \infty)$ is said to belong to \mathcal{L} , the class of *long-tailed* distributions if $\overline{F}(x+y) \sim \overline{F}(x)$ as $x \rightarrow \infty$, for any fixed y . It can be proved that any r.v. $X \in \mathcal{L}$ has infinite moment generating function [33]: $E(e^{\theta X}) = \infty, \theta > 0$. We will therefore refer to any r.v. X has a *light tail* if $E(e^{\theta X}) < \infty$ for some $\theta > 0$. In other words, a light tailed distribution has its tail distribution decays in the same order or even more rapidly than exponential, where a long tailed distribution decays slower than exponential.

A d.f. F on $[0, \infty)$ is said to be *subexponential*, denoted as $F \in \mathcal{S}$, if

$$\lim_{x \rightarrow \infty} \frac{\overline{F^{*2}}(x)}{\overline{F}(x)} = 2,$$

where F^{*2} denote the 2-fold convolution of F . For convenience,

we also use $\overline{F} \in \mathcal{S}$ to denote $F \in \mathcal{S}$.

The class of subexponential distributions was first introduced by Chistakov [6]. It is known that $\mathcal{S} \subset \mathcal{L}$. Examples of subexponential distributions include Pareto, Weibull (with shape parameter in $(0,1)$) and lognormal distributions. A nice presentation for the properties of subexponential distribution functions can be referred to, e.g. [33].

We say that F is regularly varying with index α ($\alpha \geq 0$), denoted as $F \in \mathcal{R}(-\alpha)$, if $\overline{F}(x) \sim x^{-\alpha} L(x)$, as $x \rightarrow \infty$, where L is slowly varying so that $\lim_{x \rightarrow \infty} \frac{L(tx)}{L(x)} = 1$, for any $t > 0$.

Here we quote some well-known properties that will be used later, proofs can be referred to, e.g. [8].

LEMMA 1. *Let F and G be d.f.s on $[0, \infty)$ such that $\overline{F}(x) \sim c\overline{G}(x)$, with constant $c > 0$. Then $F \in \mathcal{S}$ if and only if $G \in \mathcal{S}$.*

LEMMA 2. *If $G \in \mathcal{S}$ and F is any d.f. such that $\overline{F}(x) \sim c\overline{G}(x)$ with constant $0 \leq c < \infty$, then*

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

We first establish the following lemma. Detailed proof can be referred to the technical report [37].

LEMMA 3. *Consider d.f. G on $[0, \infty)$ with density function g and hazard rate function $q(x) = g(x)/\overline{G}(x)$. Define*

$$G_\epsilon(x) = 1 - \overline{G}(x)^{1+\epsilon}.$$

If $G \in \mathcal{S}$, and $\lim_{x \rightarrow \infty} q(x)$ exists, then

$$G_\epsilon \in \mathcal{S}, \quad \text{for all } \epsilon > 0.$$

The next lemma shows convolution properties for tail distributions that are equivalent in log-scale.

LEMMA 4. *Consider d.f.s F, G and H on $[0, \infty)$. Assume similar conditions hold for G as in Lemma 3. Suppose $\log \overline{F}(x) \sim \log \overline{G}(x)$.*

a) If for all sufficiently small $\epsilon > 0$, $\overline{G}^{1-\epsilon} \in \mathcal{S}$, and

$$\lim_{x \rightarrow \infty} \frac{\overline{H}(x)}{\overline{G}(x)^{1+\epsilon}} = a_\epsilon < \infty. \quad (2)$$

Then

$$\log \overline{F * H}(x) \sim \log \overline{G}(x). \quad (3)$$

b) If $H \in \mathcal{S}$, and for all sufficiently small $\epsilon > 0$,

$$\lim_{x \rightarrow \infty} \frac{\overline{G}(x)^{1-\epsilon}}{\overline{H}(x)} = 0. \quad (4)$$

Then

$$\overline{F * H}(x) \sim \overline{H}(x). \quad (5)$$

PROOF. Since $\log \overline{F}(x) \sim \log \overline{G}(x)$, by definition, for fixed ϵ ($0 < \epsilon < 1$), there exists x_0 such that

$$\overline{G}(x)^{1+\epsilon} \leq \overline{F}(x) \leq \overline{G}(x)^{1-\epsilon}, \quad \text{for } x \geq x_0.$$

Define respectively

$$\overline{G}'_\epsilon(x) = \min\{\overline{G}(x)^{1+\epsilon}, \overline{F}(x)\},$$

and

$$\overline{G}'_{-\epsilon}(x) = \max\{\overline{G}(x)^{1-\epsilon}, \overline{F}(x)\}.$$

It follows that for large x ($\geq x_0$),

$$\overline{G}'_\epsilon(x) \sim \overline{G}(x)^{1+\epsilon}, \quad \overline{G}'_{-\epsilon}(x) \sim \overline{G}(x)^{1-\epsilon}. \quad (6)$$

In addition,

$$\overline{G}'_\epsilon(x) \leq \overline{F}(x) \leq \overline{G}'_{-\epsilon}(x), \quad \text{for all } x \geq 0.$$

Let X_ϵ , $X_{-\epsilon}$, Y and Z be mutually independent random variables on $[0, \infty)$ with tail distributions $\overline{G}'_\epsilon(x)$, $\overline{G}'_{-\epsilon}(x)$, $\overline{F}(x)$, and $\overline{H}(x)$, respectively. Then

$$X_\epsilon \leq_{\text{st}} Y \leq_{\text{st}} X_{-\epsilon}.$$

It follows from properties of stochastic ordering [34] that $X_\epsilon + Z \leq_{\text{st}} Y + Z \leq_{\text{st}} X_{-\epsilon} + Z$; or equivalently,

$$\overline{G}'_\epsilon * \overline{H}(x) \leq \overline{F} * \overline{H}(x) \leq \overline{G}'_{-\epsilon} * \overline{H}(x). \quad (7)$$

Proof of a). Based on Lemma 3, $\overline{G}^{1+\epsilon} \in \mathcal{S}$. It follows from (2), (6) and Lemma 2 that, for large x ,

$$\overline{G}'_\epsilon * \overline{H}(x) \sim (1 + a_\epsilon) \overline{G}(x)^{1+\epsilon}. \quad (8)$$

In addition (2) implies that

$$\lim_{x \rightarrow \infty} \frac{\overline{H}(x)}{\overline{G}^{1-\epsilon}(x)} = \lim_{x \rightarrow \infty} \frac{\overline{H}(x)}{\overline{G}(x)^{1+\epsilon}} \cdot \overline{G}(x)^{2\epsilon} = 0.$$

Since $\overline{G}^{1-\epsilon} \in \mathcal{S}$, from (6) and Lemma 2, it follows that, for large x ,

$$\overline{G}'_{-\epsilon} * \overline{H}(x) \sim \overline{G}(x)^{1-\epsilon}. \quad (9)$$

Combine (7), (8) and (9), we then have

$$(1 + a_\epsilon) \overline{G}(x)^{1+\epsilon} \leq \overline{F} * \overline{H}(x) \leq \overline{G}(x)^{1-\epsilon}.$$

Take the logarithm of the left hand side, then yield

$$\begin{aligned} 0 &\geq \log \frac{(1 + a_\epsilon) \overline{G}(x)^{1+\epsilon}}{\overline{F} * \overline{H}(x)} \\ &= \log(1 + a_\epsilon) + (1 + \epsilon) \log \overline{G}(x) - \log \overline{F} * \overline{H}(x). \end{aligned}$$

Since $\log \overline{G}(x) \rightarrow -\infty$ as $x \rightarrow \infty$, it follows that

$$\lim_{x \rightarrow \infty} \frac{\log \overline{F} * \overline{H}(x)}{\log \overline{G}(x)} \leq 1 + \epsilon.$$

Similarly by taking the logarithm of the right hand side, we have

$$\lim_{x \rightarrow \infty} \frac{\log \overline{F} * \overline{H}(x)}{\log \overline{G}(x)} \geq 1 - \epsilon.$$

Letting ϵ go to zero, (3) then follows.

Proof of b). Since $H \in \mathcal{S}$, from (4) (6) and Lemma 2, we have

$$\overline{G}'_{-\epsilon} * H(x) \sim \overline{H}(x).$$

Similarly, since

$$\lim_{x \rightarrow \infty} \frac{\overline{G}(x)^{1+\epsilon}}{\overline{H}(x)} = \lim_{x \rightarrow \infty} \frac{\overline{G}(x)^{1-\epsilon}}{\overline{H}(x)} \cdot \overline{G}(x)^{2\epsilon} = 0,$$

from (6) and Lemma 2, it follows that

$$\overline{G}'_\epsilon * H(x) \sim \overline{H}(x).$$

Combine with (7), we then have (5). \square

3. LOWER BOUNDS UNDER SUBEXPONENTIAL SERVICE TIMES

Consider a single-server queueing system where jobs arrive at random times $0 \leq \Gamma_1 \leq \Gamma_2 \leq \dots$, and the service times $\{S_n\}_{n \geq 1}$ are i.i.d. random variables with d.f. F , independent of the arrival process. Let $T_n = \Gamma_n - \Gamma_{n-1}$, $n = 1, 2, \dots$, be the interarrival times. Throughout this paper, we shall assume that the service discipline is FCFS, and that the sequence $\{T_n\}_{n=1}^\infty$ is stationary and ergodic, which implies that the sequences $\{S_n\}_{n=1}^\infty$ and $\{T_n\}_{n=1}^\infty$ are jointly stationary and ergodic. We can then define the stationary extension of the sequences $\{S_n\}_{n=-\infty}^\infty$ and $\{T_n\}_{n=-\infty}^\infty$ which again are jointly stationary and ergodic.

Denote $A(t)$ the cumulative number of arrivals in time $[0, t)$. Under the above mentioned assumption, the process $A(t)$ is stationary and ergodic such that $\lim_{t \rightarrow \infty} \frac{A(t)}{t} = \lambda$. Note that the arrival process $A(t)$ could be dependent or even long-range dependent. In addition, arrivals could be in batches so that interarrival times for jobs within a batch are zeros. We consider only stable systems where $\rho = \frac{\lambda}{\mu} < 1$.

Let $A^r(t)$ to be the cumulative number of arrivals in time $[-t, 0)$, $U^r(t)$ to be the cumulative amount of workload ar-

rived during time $[-t, 0)$, so that

$$U^r(t) = \sum_{k=1}^{A^r(t)} S_{-k}.$$

We are interested in the tail behavior of two stationary variables, namely, the *stationary waiting time*, expressed in Loynes' schema ([24]) as

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

and the *stationary virtual waiting time* (or *stationary backlog*)

$$V_\infty = \left(\sup_{t \geq 0} (U^r(t) - t) \right)^+, \quad (10)$$

where $\stackrel{d}{=}$ denotes the equality in distribution.

The following lemma follows immediately.

LEMMA 5.

$$V_\infty \leq_{st} W_\infty \leq_{st} V_\infty + S_1. \quad (11)$$

In particular, for zero-delayed system, i.e. when $\Gamma_1 = 0$, then $V_\infty \stackrel{d}{=} W_\infty$.

We next develop lower bounds for the tail distributions of the stationary waiting time W_∞ and of the stationary virtual waiting time V_∞ . We will first need the following lemma, which can be considered as an extended version of Lemma 3.1 in [2].

LEMMA 6. Let D_n , $n = 1, 2, \dots$ be a sequence of random variables such that as $n \rightarrow \infty$, $\frac{D_n}{n} \rightarrow d$ with probability(w.p.) 1. For arbitrary $\epsilon, \epsilon' > 0$, there exists a constant $c > 0$ such that

$$\mathbf{P} \left[\bigcap_{n \geq 1} \{D_n \leq n(d + \epsilon) + c\} \right] > 1 - \epsilon'. \quad (12)$$

and

$$\mathbf{P} \left[\bigcap_{n \geq 1} \{D_n \geq n(d - \epsilon) - c\} \right] > 1 - \epsilon'. \quad (13)$$

PROOF. Inequality (12) is basically the result of Lemma 3.1. in [2]. Noticing that $\frac{-D_n}{n} \rightarrow -d$ w.p. 1, inequality (13) simply follows from applying (12) on the sequence $-D_n$. \square

THEOREM 7 (LOWER BOUNDS). Consider a $G/GI/1$ queue with FCFS, where the input process is stationary and ergodic

such that $\lim_{t \rightarrow \infty} \frac{A(t)}{t} = \lambda$. The service times are subexponential i.i.d. random variables with mean μ^{-1} and d.f. F . Denote $\rho = \frac{\lambda}{\mu}$, and assume $\rho < 1$.

a). If $F_e \in \mathcal{S}$, then

$$\mathbf{P}[W_\infty > x] \succeq \frac{\rho}{1 - \rho} \overline{F}_e(x). \quad (14)$$

In this case, the same (corresponding) result also holds for V_∞ .

b). Let $V_\infty^{G/D/1}$ be the stationary virtual waiting time of the associated $G/D/1$ queues where the arrival process is the same but service times are deterministic and equal to μ^{-1} . If for all sufficiently small $\epsilon > 0$,

$$V_\infty^{G/D-\epsilon/1} \in \mathcal{L}, \quad (15)$$

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

$$\mathbf{P}[W_\infty > x] \geq \mathbf{P}[V_\infty > x] \succeq \mathbf{P}[V_\infty^{G/D/1} > x]. \quad (16)$$

PROOF. a). Since the input process is stationary and ergodic, it follows immediately that $\frac{\sum_{i=1}^n T_i}{n} \rightarrow \lambda^{-1}$ w.p. 1. Applying Lemma 6 yields that for arbitrary $\epsilon, \epsilon' > 0$, there exists a constant $c > 0$ such that $\mathbf{P}[B] > 1 - \epsilon'$, where B denotes the event set

$$B := \bigcap_{n \geq 1} \left\{ \sum_{k=1}^n T_{-k} \leq n(\lambda^{-1} + \epsilon) + c \right\}.$$

We then have

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

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 obtained by applying (1) to

D/GI/1 queues with deterministic interarrival times $\lambda^{-1} + \epsilon$. Since $F_e \in \mathcal{S} \subset \mathcal{L}$, $\overline{F}_e(x+c) \sim \overline{F}_e(x)$. Hence, (14) follows by letting ϵ and ϵ' go to 0.

Based on Lemma 5 and the fact that $\lim_{x \rightarrow \infty} \frac{\overline{F}(x)}{\overline{F}_e(x)} = 0$, the same corresponding result for V_∞ follows immediately. b). Since S_k 's are i.i.d. random variables, by the strong law of large numbers, $\frac{\sum_{k=1}^n S_{-k}}{n} \rightarrow \mu^{-1}$ w.p. 1 as $n \rightarrow \infty$. Applying Lemma 6 entails that for arbitrary $\epsilon, \epsilon' > 0$, there exists a constant $c > 0$ such that $\mathbf{P}[B'] > 1 - \epsilon'$, where B' denotes the event set

$$B' := \bigcap_{n \geq 1} \left\{ \sum_{k=1}^n S_{-k} \geq n(\mu^{-1} - \epsilon) - c \right\}.$$

Therefore,

$$\begin{aligned} & \mathbf{P}[V_\infty > x] \\ &= \mathbf{P} \left[\sup_{t \geq 0} \left\{ \sum_{j=1}^{A^r(t)} S_{-j} - t \right\} > x \right] \\ &\geq \mathbf{P}[B'] \cdot \mathbf{P} \left[\sup_{t \geq 0} \left\{ \sum_{j=1}^{A^r(t)} S_{-j} - t \right\} > x \mid B' \right] \\ &\geq (1 - \epsilon') \mathbf{P} \left[\sup_{t \geq 0} \{A^r(t)(\mu^{-1} - \epsilon) - t\} > x + c \mid B' \right] \\ &= (1 - \epsilon') \mathbf{P} \left[V_\infty^{G/D-\epsilon/1} > x + c \right], \end{aligned}$$

where the last equality follows from the fact that the service times (thus event B') is independent of the arrival process. Since $V_\infty^{G/D-\epsilon/1} \in \mathcal{L}$, we have

$$\mathbf{P} \left[V_\infty^{G/D-\epsilon/1} > x + c \right] \sim \mathbf{P} \left[V_\infty^{G/D-\epsilon/1} > x \right].$$

By letting ϵ and ϵ' go to 0, (16) then follows. \square

Theorem 7 a) shows that under dependent arrival process and subexponential service times, the tail distribution of the stationary waiting time will be at least as heavy as that by replacing the arrival process with its independent version. Similar lower bounds were obtained for the case of deterministic service times in [35] via supermodular ordering for FGN, on-off sources and M/G/ ∞ models.

Theorem 7 b) shows that 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. with deterministic service times) if the impact is non-negligible. The next immediate question, how would the performance be under both the impact of subexponential

service times and long-range dependence arrival process? In the next two sections, we answer this question separately for FGN input and M/G/ ∞ input process.

4. M/G/ ∞ INPUTS

A frequently used model to characterize dependent traffic is the so-called Cox model, or M/G/ ∞ input model [11]. A queueing system with M/G/ ∞ inputs is the one where customer arrivals characterized by the number of busy servers of an infinite server queueing system M/G/ ∞ with Poisson arrivals and i.i.d. service times with distribution G . The M/G/ ∞ arrival model is a versatile process as it can be applied to generate both short-range and long-range dependent traffic by properly selecting the service time distribution G , which is also referred to as the session time distribution (see remark below).

In this section, we consider a single-server queue fed by M/G/ ∞ arrivals, where each arrival incurs a random service demand S_i . We assume the service times S_i 's are i.i.d. random variables with finite mean μ^{-1} and distribution F . For convenience, we simply denote such systems as $MG^\infty/GI/1$ queues and we are interested in the asymptotic tail distribution of response times in such queues.

More formally, define an M/G/ ∞ input process with Poisson arrival process at rate λ_0 and i.i.d. session times with mean ν^{-1} and common distribution G . Denote b_t the number of busy servers at time t . Suppose each busy server in the M/G/ ∞ system sends out r requests (in batch) at the beginning of each time slot, where r is an integer and $r \geq 1$. Then the number of arrivals at time t is

$$a_t = r b_t, \quad t = 0, 1, 2, \dots \quad (17)$$

Note that the process is completely specified by the triplet (λ_0, G, r) . We will simply say that this process is an M/G/ ∞ input process with parameters (λ_0, G, r) .

Remark: Consider the above M/G/ ∞ model in the context of infinite on-and-off sources, the distribution G is then the distribution of the on-period or the session duration. We will simply call G the *session duration distribution*.

It is known that in steady-state, $b_t \sim \text{Poisson}(\lambda_0/\nu)$. Therefore the $MG^\infty/GI/1$ system should have request arrival rate $\lambda = r\lambda_0/\nu$, and traffic intensity $\rho = \frac{\lambda}{\mu}$. Let

$$A(t) = \sum_{s=1}^t a_s \quad \text{and} \quad U(t) = \sum_{n=1}^{A(t)} S_n$$

be respectively the cumulative number of requests and the cumulative amount of workload in time $[0, t)$. Let $A^r(t)$ and

$U^r(t)$ be the corresponding quantities for the time interval $[-t, 0)$.

Define W_t as the waiting time of the first job arrived at time t (since arrivals can occur in batches). Similarly we can define V_t to be the virtual waiting time observed by the first job arrived at time t . We know from [24] that the stationary virtual waiting time satisfies

$$V_\infty \stackrel{d}{=} \left(\sup_{t \geq 0} (U^r(t) - t) \right)^+.$$

In order to state our main results, we need to introduce the notion of *intermediately regularly varying* distributions. A distribution function F is intermediate regular varying $F \in \mathcal{IR}$ if

$$\lim_{a \downarrow 1} \liminf_{t \rightarrow \infty} \frac{\overline{F}(at)}{\overline{F}(t)} = 1.$$

Various properties of \mathcal{IR} distributions can be found in [9]. Basic properties of \mathcal{IR} include: $\mathcal{IR} \subset \mathcal{S}$, $\mathcal{R} \subset \mathcal{IR}$. Also, $F \in \mathcal{IR}$ and $\int_0^\infty \overline{F}(y) dy < \infty$, implies $F_e \in \mathcal{IR}$. Pareto distributions are well known examples of \mathcal{IR} . In addition, if $F \in \mathcal{IR}$, then there exists $\alpha \geq 0$ and a finite constant C such that for all $x > 0$, $\overline{F}(x) \leq \frac{C}{x^\alpha}$.

The next Lemma summarizes some known results on fluid queueing models (i.e. when all arrivals have constant service demands). Studies on this model can be found in [18, 31, 29, 16, 22], etc. Here we quote the latest result based on [31, 18]. We restate the result here in the context of queues.

LEMMA 8. Consider an $MG^\infty/D/1$ queue with deterministic service times $S \equiv \mu^{-1}$. The arrival process is an $M/G/\infty$ input process with parameters (λ_0, G, r) . If $G \in \mathcal{IR}$, $\alpha > 1$, and

$$\rho < 1 < r/\mu + \rho, \quad (18)$$

then the stationary virtual waiting time

$$V_\infty = \sup_{t \geq 0} (\mu^{-1} A^r(t) - t)$$

must satisfy,

$$\mathbf{P}[V_\infty > x] \sim \frac{\rho}{1-\rho} \cdot \overline{G}_e\left(\frac{x}{r/\mu + \rho - 1}\right). \quad (19)$$

Now consider random i.i.d. service times $\{S_i\}$ with d.f. F . We claim the following:

THEOREM 9. Consider an $MG^\infty/GI/1$ queue where the service times $\{S_i\}$'s are i.i.d. random variables with mean μ^{-1} and distribution F . We assume similar conditions hold as in Lemma 8.

a) If F is light-tailed or if $F_e \in \mathcal{S}$ such that

$$\lim_{x \rightarrow \infty} \frac{\overline{F}_e(x)}{\overline{G}_e(x)} = 0. \quad (20)$$

Then, (19) holds.

b) If $F_e \in \mathcal{S}$ and it is heavier than G_e , i.e.

$$\lim_{x \rightarrow \infty} \frac{\overline{G}_e(x)}{\overline{F}_e(x)} = 0, \quad (21)$$

then

$$\mathbf{P}[V_\infty > x] \sim \frac{\rho}{1-\rho} \overline{F}_e(x). \quad (22)$$

c) If $F_e \in \mathcal{S}$ and it is tail equivalent to G_e , i.e. $\overline{F}_e(x) \sim c \overline{G}_e(x)$ for some constant $c > 0$, then there exists $b > 0$, such that

$$\frac{\rho}{1-\rho} \overline{F}_e(x) \preceq \mathbf{P}[V_\infty > x] \preceq b \cdot \overline{F}_e(x). \quad (23)$$

PROOF. *Proof of a).* If we can show that condition (15) holds, then the \succeq part of (19) will follow immediately from Theorem 7 part b).

Let $\mu_{-\epsilon} = 1/(\mu^{-1} - \epsilon)$. Denote $V_\infty^{MG^\infty/D-\epsilon/1}$ the stationary virtual waiting time of the associated with $MG^\infty/D-\epsilon/1$ queues where the service times are deterministic and equal to $1/\mu_{-\epsilon}$. By choosing ϵ small enough so that $\mu_{-\epsilon}$ satisfies (18), we obtain from Lemma 8 that

$$\mathbf{P}\left[V_\infty^{MG^\infty/D-\epsilon/1} > x\right] \sim \frac{\rho_{-\epsilon}}{1-\rho_{-\epsilon}} \cdot \overline{G}_e\left(\frac{x}{r/\mu_{-\epsilon} + \rho_{-\epsilon} - 1}\right),$$

where $\rho_{-\epsilon} = \lambda/\mu_{-\epsilon}$. Since $G \in \mathcal{IR}$ and $\alpha > 1$, we have $\int_0^\infty \overline{G}(y) dy < \infty$; thus $G_e \in \mathcal{IR} \subset \mathcal{S} \subset \mathcal{L}$. Therefore condition (15) holds.

It remains to show the \preceq part of (19). We use the following decomposition

$$\begin{aligned} V_\infty &\stackrel{d}{=} \left(\sup_{t \geq 0} (U^r(t) - t) \right)^+ \\ &= \left(\sup_{t \geq 0} \left\{ \sum_{i=1}^{A_t^r} (S_{-i} - \mu_\epsilon^{-1}) + \left(\sum_{i=1}^{A_t^r} \mu_\epsilon^{-1} - t \right) \right\} \right)^+ \\ &\leq M_\epsilon + Y_\epsilon, \end{aligned}$$

where $\mu_\epsilon^{-1} = \mu^{-1} + \epsilon$ for some small $\epsilon > 0$, and

$$M_\epsilon = \left(\sup_{n \geq 1} \sum_{i=1}^n (S_{-i} - \mu_\epsilon^{-1}) \right)^+,$$

and

$$Y_\epsilon = \left(\sup_{t \geq 0} \left(\sum_{i=1}^{A_t^r} \mu_\epsilon^{-1} - t \right) \right)^+.$$

It follows that

$$\mathbf{P}[V_\infty > x] \leq \mathbf{P}[M_\epsilon + Y_\epsilon > x], \quad (24)$$

Note that Y_ϵ corresponds to the stationary virtual waiting time of a $MG^\infty/D/1$ queue with service rate μ_ϵ . Thus, by choosing ϵ small enough so that μ_ϵ satisfies (18), and by applying Lemma 8, we obtain

$$\mathbf{P}[Y_\epsilon > x] \sim \frac{\rho_\epsilon}{1 - \rho_\epsilon} \cdot \bar{G}_\epsilon\left(\frac{x}{r/\mu_\epsilon + \rho_\epsilon - 1}\right), \quad (25)$$

where $\rho_\epsilon = \lambda/\mu_\epsilon$.

Similarly, M_ϵ corresponds to the stationary waiting time of a $D/GI/1$ queue with the service time distribution F . If F is light-tailed, then so is M_ϵ (see, e.g. [15]). If $F_\epsilon \in \mathcal{S}$, then the results in [27] implies that

$$\mathbf{P}[M_\epsilon > x] \sim \frac{\rho(M_\epsilon)}{1 - \rho(M_\epsilon)} \bar{F}_\epsilon(x), \quad (26)$$

where $\rho(M_\epsilon) = \frac{1}{1 + \mu_\epsilon}$.

In either cases, the tail of M_ϵ is lighter than that Y_ϵ due to condition (20). The convolution rule in Lemma 2 therefore implies that

$$\mathbf{P}[M_\epsilon + Y_\epsilon > x] \sim \mathbf{P}[Y_\epsilon > x].$$

Letting ϵ goes to 0, we then have the \preceq part of (19).

Proof of b). Since $F_\epsilon \in \mathcal{S}$, from Theorem 7 part a), we know that (14) holds for V_∞ .

To show the upper bound, note that

$$\mathbf{P}[V_\infty > x] \leq \mathbf{P}[M'_\epsilon + Y'_\epsilon > x],$$

where

$$M'_\epsilon = \left(\sup_{n \geq 1} \left\{ \sum_{k=1}^n [S_{-k} - \lambda_\epsilon^{-1}] \right\} \right)^+,$$

$$Y'_\epsilon = \left(\sup_{t \geq 0} \left(\sum_{i=1}^{A_t^r} \lambda_\epsilon^{-1} - t \right) \right)^+,$$

and $\lambda_\epsilon^{-1} = \lambda^{-1} - \epsilon$ for some small $\epsilon > 0$.

Observe that Y'_ϵ corresponds to the stationary virtual waiting time of a $MG^\infty/D/1$ queue with service rate λ_ϵ . By choosing ϵ small enough so that (18) holds for $\mu = \lambda_\epsilon$, and by applying Lemma 8, we obtain

$$\mathbf{P}[Y'_\epsilon > x] \sim \frac{\rho'_\epsilon}{1 - \rho'_\epsilon} \cdot \bar{G}_\epsilon\left(\frac{x}{r/\lambda_\epsilon + \rho'_\epsilon - 1}\right). \quad (27)$$

where $\rho'_\epsilon = \lambda/\lambda_\epsilon = 1 - \lambda\epsilon$.

Since $F_\epsilon \in \mathcal{S}$, Pakes' result [27] implies that

$$\mathbf{P}[M'_\epsilon > x] \sim \frac{\rho(M'_\epsilon)}{1 - \rho(M'_\epsilon)} \bar{F}_\epsilon(x), \quad (28)$$

where $\rho(M'_\epsilon) = \lambda_\epsilon/\mu$. Applying (21) and the convolution rule in Lemma 2 yields

$$\mathbf{P}[M'_\epsilon + Y'_\epsilon > x] \sim \frac{\rho(M'_\epsilon)}{1 - \rho(M'_\epsilon)} \bar{F}_\epsilon(x).$$

By further letting ϵ go to 0, we then obtain the desired upper bound.

Proof of c). The lower bound for c) follows similarly as in the proof for b). Since $\bar{F}_\epsilon(x) \sim c\bar{G}_\epsilon(x)$ for $c > 0$, the constant b in upper bound follows from the convolution rule of M'_ϵ and Y'_ϵ based on (28) and (27). \square

Note that, thanks to Lemma 5, the statements of Theorem 9 hold true when the stationary virtual waiting time V_∞ is replaced by the stationary waiting time W_∞ .

In summary, for $MG^\infty/GI/1$ queues, where the service times are i.i.d. r.v.s with distribution F , and the arrival process is driven by the queue length process of an $M/G/\infty$ queue with parameter (λ_0, G, r) , Theorem 9 says that the asymptotic tail behavior of the stationary waiting time and the stationary virtual waiting time are of two folds: if the residual service time tail distribution F_ϵ is lighter than the residual session duration G_ϵ , then the tail distributions of the stationary waiting time and of the stationary virtual waiting time are dominated by the residual session duration G_ϵ , i.e. the dependent input process dominates the performance and impact due to the subexponential service times becomes negligible; on the other hand, when the residual service time distribution F_ϵ is heavier than the residual session duration G_ϵ , then the tail distributions of the stationary waiting time and of the stationary virtual waiting time are dominated by that of the residual service time, in which case, the performance impact due to the input traffic dependence structure becomes negligible.

Based on the computation for the integrated tail distributions $F_\epsilon(\cdot)$ for different subexponential distribution families (refer to, e.g. [22]), we can easily check whether condition (20) or (21) is satisfied. The next corollary is immediate.

COROLLARY 10. *Consider $MG^\infty/GI/1$ queues, where the service-times are i.i.d. random variables with distribution F . Assume similar conditions hold for the input process as in Theorem 9.*

i) If F is light-tailed, lognormal, Weibull, or $F \in \mathcal{R}(-\tilde{\alpha})$ with $\tilde{\alpha} > \alpha$, then (20) holds, and the tail distributions of the

stationary waiting time and of the stationary virtual waiting time are dominated by the long-range dependence of the input process.

ii) If $F \in \mathcal{R}(-\tilde{\alpha})$ with $1 < \tilde{\alpha} < \alpha$, then (21) holds, and the tail distributions of the stationary waiting time and of the stationary virtual waiting time are dominated by the tail of the residual service times.

5. FGN INPUTS

Besides M/G/ ∞ model, fractional Gaussian noise (FGN) is another popular model that is frequently used to model long-range dependent traffic, mostly due to its mathematical simplicity. In this section, we concentrate on the case when the arrival traffic is FGN. A detailed treatment of FGN processes can be found in [32]. Its use for traffic modeling is discussed in [26] and in [28] (and references therein).

Suppose requests arrive at discrete times $t = 0, 1, 2, \dots$, 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

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

Consider a FGN/GI/1 queue with arrival process $\{a_t\}_{t=-\infty}^{\infty}$, the service times for the requests $\{S_i\}_{i=-\infty}^{\infty}$'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 as follows:

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

where $U^r(t) = \sum_{i=1}^{A^r(t)} S_{-i}$ and $A^r(t) = \sum_{s=0}^t a_{-s}$.

In what follows, we study the asymptotic performance of the FGN/GI/1 queue, where the arrival process is FGN as defined above, and the service times are subexponential i.i.d. random variables with finite mean μ^{-1} and general distribution F .

5.1 FGN Inputs with Deterministic Service Times

Note that the cumulative number of arrivals of the FGN process is the so-called FGN process, denoted as $B(t)$, which can be expressed as follows,

$$B(t) = \lambda t + \sigma Z(t), \quad t \in (-\infty, \infty),$$

where $Z(\cdot)$ is a normalized fractional Gaussian process, with mean 0 and variance 1, it also has a covariance function

Fluid queues with FGN inputs have been studied in [26, 13]. A lower bound on the tail probability of the stationary workload was first obtained in [26]. It has been shown later in [13] to be asymptotically exact in log scale using large deviation principle. Their results can be restated as follows.

LEMMA 11. The variable $V_{\infty}^B := \left(\sup_{t \geq 0} \{\mu^{-1} B(t) - t\} \right)^+$ satisfies

$$\lim_{x \rightarrow \infty} x^{-\beta} \log \mathbf{P} \left[V_{\infty}^B > x \right] = -\delta, \quad (29)$$

where $\beta = 2 - 2H$, $\rho = \frac{\lambda}{\mu}$,

$$\delta = \frac{1}{2\rho^2\gamma^2(1-H)^2} \left(\frac{(1-\rho)(1-H)}{H} \right)^{2H}, \quad (30)$$

and $\gamma^2 = \frac{\sigma^2}{\lambda^2}$ is the coefficient of variation of $B(\cdot)$.

The above lemma says that under input process $A(t)$ and deterministic service times, the queueing behavior is basically dominated by the long-range dependence, and the tail distributions of the stationary waiting times and stationary virtual waiting times are in log-scale equivalent to Weibull distribution with shape parameter $\beta = 2 - 2H$.

5.2 FGN Inputs with Subexponential Service Times

In this section, we focus on systems under the FGN arrival process and subexponential service times. Again, we are interested in two quantities:

$$\overline{F}_w(x) := \mathbf{P} \left[\sup_{n \geq 1} \left\{ \sum_{k=1}^n (S_k - T_k) \right\} > x \right],$$

and

$$\overline{F}_v(x) := \mathbf{P} \left[\sup_{t \geq 0} \left\{ \sum_{k=1}^{A(t)} S_k - t \right\} > x \right],$$

which are, according to Loynes' schema [24], the stationary waiting time and the stationary virtual waiting time distributions.

We show that the tail distribution of stationary waiting times is of two folds. When $\overline{F}_e(x)$, the tail distribution of the residual service times, is heavier than the Weibull distribution with shape parameter $\beta = 2 - 2H$, (1) still holds; i.e. the service time dominates the performance. However, when $\overline{F}_e(x)$ is lighter than Weibull(2-2H), the tail behavior of the stationary waiting time is then dominated by the long-range dependence of the arrival process.

THEOREM 12. Consider an FGN/GI/1 queue with Hurst parameter $H \in (\frac{1}{2}, 1)$. The service times $\{S_i\}$'s are i.i.d. with distribution F and $F_e \in \mathcal{S}$. Let $\beta = 2 - 2H$ and δ defined by (30).

a) If there exists $\eta > 0$, such that

$$\lim_{x \rightarrow \infty} \frac{\overline{F}_e(x)}{e^{-(\delta+\eta)x^\beta}} = a_\eta < \infty, \quad (31)$$

then the tail distributions of the stationary waiting time must satisfy (29).

b) If

$$\lim_{x \rightarrow \infty} \frac{\exp(-\eta x^\beta)}{\overline{F}_e(x)} = 0, \quad (32)$$

for all sufficiently small $\eta > 0$, then (1) holds, i.e. the tail distribution of the stationary waiting time is dominated by the residual service time distribution.

In either case, the same (corresponding) results also hold for V_∞ .

PROOF. The proof is based on similar techniques as used for proving Theorem 9 in Section 4. More details can be referred to the technical report [37]. \square

Remark: Part b) of Theorem 12 can be considered as a special case of the result in [2]. Using a similar decomposition approach as in the proof of Theorem 12, [2] provides a sufficient condition under which (1) still holds so that the stationary waiting time tail distribution is dominated by the residual service time.

From Theorem 12, we see that if F_e is in the same order or lighter than Weibull($2 - 2H + \epsilon$) for arbitrarily small ϵ , then the queueing performance is dominated by the LRD properties of the input process. On the other hand, if F_e is in the same order or heavier than Weibull($2 - 2H - \epsilon$) for arbitrarily small ϵ , then the queueing performance is dominated by the residual service time.

The following corollary is immediate.

COROLLARY 13. Consider an FGN/GI/1 queues, where the arrival process is long-range dependent with Hurst parameter $H \in (\frac{1}{2}, 1)$, and the service-times are i.i.d. subexponential.

i) If the service time distribution is regularly varying with finite mean, or lognormal, or Weibull with parameter $\nu \in (0, 2 - 2H)$, then (1) holds; that is, the tail distributions of the stationary waiting time and of the stationary virtual

waiting time are dominated by the tail of the residual service time distribution;

ii) If the service time distribution is Weibull with shape parameter $\nu \in (2 - 2H, 1)$, or $\nu = 2 - 2H$ and $a > \delta$, where δ is given by (30), then (29) holds; that is, the tail distributions of the stationary waiting time and of the stationary virtual waiting time are dominated by the long-range dependence.

6. CONCLUDING REMARKS

In this paper, we presented a study of the asymptotics of the tail distributions of the stationary waiting time and of the stationary virtual waiting time of LRD/GI/1 queues, where two widely-used LRD models, namely the FGN model and the M/G/ ∞ model, were used to capture the long-range dependence structure of the arrival process. By looking into different families of tail distributions of the service times, we show the combined impact on the tail probabilities of the stationary waiting time and of the stationary virtual waiting time by both the long-range dependence structure of the input traffic and the tail properties of the service times.

For MG^∞ /GI/1 queues, where the service times are i.i.d. r.v.s with distribution F , and the arrival process is driven by a M/G/ ∞ input model, the asymptotic tail behavior of these stationary performance metrics is of two folds: if the residual service time tail distribution F_e is lighter than the residual session duration G_e , then they are dominated by the residual session duration G_e . On the other hand, when the residual service time distribution F_e is heavier than the residual session duration G_e , then they are dominated by that of the residual service time.

When the arrival process is modeled by a fractional Gaussian noise model, we show that if the service times are subexponential, and the integrated service time is asymptotically heavier than Weibull($2-2H$), then the tail distributions of the stationary waiting time and of the stationary virtual waiting time are dominated by the residual service time distribution. On the other hand, if the service times are light-tailed, or if the service times are subexponential, but its integrated tail distribution is asymptotically lighter than Weibull($2-2H$), then the tail distributions of the stationary waiting time and of the stationary virtual waiting time are asymptotically equivalent to Weibull($2-2H$) (in log-scale), i.e., the long-range dependence structure of the arrival process dominates the performance.

Although these results were established only for two LRD models, we conjecture that such asymptotic dominance by

either the arrival process or the service time holds for other models as well. The service discipline that was considered in this paper is FCFS. It would be interesting to study LRD/GI/1 queues with other service disciplines such as Processor Sharing. Some preliminary investigations show that the analysis of processor sharing queueing systems with dependent arrivals tends to be much more involved.

7. REFERENCES

- [1] M.F. Arlitt and C.L. Williamson. Internet Web servers: Workload characterization and performance implications. *IEEE/ACM Transactions on Networking*, 5(5):631-645, Oct. 1997.
- [2] S. Asmussen, H. Schmidli and V. Schmidt. Tail probabilities for non-standard risk and queueing processes with subexponential jumps, *Adv. in Appl. Probab.* 31 (1999) 422-447.
- [3] F. Baccelli and A. M. Makowski. Queueing Models for Systems with Synchronization Constraints *Proceedings of the IEEE*, 77, Special Issue on Dynamics of Discrete Event Systems, 1989, pp. 138-161.
- [4] S.C. Borst, O.J. Boxma and R.N. Queija. Heavy tails: the effect of the service discipline. *Computer Performance Evaluation/TOOLS*, 2002: 1-30.
- [5] C.S. Chang. *Performance Guarantees in Communication Networks*. Springer-Verlag, 2000.
- [6] V.P. Chistakov. A theorem on sums of independent positive random variables and its application to branching random process. In *Theor. Probab. Appl.*, 9:640:648, 1964.
- [7] G.L. Choudhury and W. Whitt. Long-tail buffer-content distributions in broadband networks. *Performance Evaluation*, 30 (1997) 177-190.
- [8] D.B.H. Cline. Convolution tails, product tails and domains of attraction. *Prob. Theory Related Fields*, 72 (1986) 529-557.
- [9] D.B.H. Cline. Intermediate regular and π variation. *Proc. London Math. Soc.*, 1994.
- [10] M.E. Corvella and A. Bestavros. Self-similarity in World Wide Web traffic: evidence and possible causes. *Performance Evaluation Review*, 24 (1996) 160-169.
- [11] D.R. Cox. Long-range dependence: A review. *Statistics: An Appraisal*. H.A. David and H.T. David, Eds., The Iowa State University Press, Ames. (IA), 1984, pp 55-74.
- [12] A. Downey. The structural cause of file size distributions. In *Proceedings of the International Symposium on Modeling Analysis and Simulation of Computer and Telecommunication Systems*, Aug. 2001.
- [13] N.G. Duffield and N. O'Connell. Large deviation and overflow probabilities for the general single-server queue, with applications. In *Mathematical Proceedings of the Cambridge Philosophical Society*, 118 (1995) 363-375.
- [14] P. Embrechts, C. Klüppelberg and T. Mikosch, *Modelling Extremal Events for Insurance and Finance*. Springer, Heidelberg, 1997.
- [15] P.V. Glynn and W. Whitt. Longarithmic asymptotics for steady-state tail probabilities in a single-server queue. *J. Appl. Prob.*, 31A (1994) 131-156.
- [16] F.Guillemain, R. Mazumdar and A. Simonian. On heavy traffic approximations for transient characteristics of M/M/ ∞ queues, *J. Appl. Prob.*, 33, No. 2, 1996, pp. 490-506.
- [17] A.K. Iyengar, M.S. Squillante, and L. Zhang. Analysis and Characterization of Large-Scale Web Server Access Patterns and Performance. *World Wide Web*, 2 (1999) 85-100.
- [18] P. Jelenkovic. On the asymptotic behavior of a fluid queue with a heavy-tailed M/G/ ∞ arrival process. June 2000, preprint.
- [19] P. Jelenkovic and A.A. Lazar. Subexponential asymptotics of a Markov-modulated random walk with queueing applications, *J. Appl. Prob.*, June 1998.
- [20] P. Jelenkovic and P. Momcilovic. Resource Sharing with Subexponential Distributions. In *Proceedings of IEEE INFOCOM*, June 2002, Vol. 3, 1316-1325.
- [21] W.E. Leland, M.S. Taqqu, W. Willinger, and D.V. Wilson. On the self-similarity nature of Ethernet traffic (Extended version). In *IEEE/ACM Trans. on Networking*, Vol. 2, No. 1, pp 1-15, Feb. 1994.
- [22] Z. Liu, P. Nain, D. Towsley and Z.-L. Zhang. Asymptotic behavior of a multiplexer fed by a long-range dependent process. *J. Appl. Probab.* 36 (1999) 105-118.
- [23] Z. Liu, N. Niclausse, C. Jalpa-Villanueva. Traffic model and performance evaluation of Web servers. *Performance Evaluation*, Vol. 46, No. 2-3, pp. 77-100, 2001.
- [24] R.M. Loynes. The stability of a queue with non-independent inter-arrival and service times. In *Proc. Cambridge Philos. Soc.*, 58 (1968) 497-520.

- [25] K. Maulik, S. Resnick, and H. Rootzen. A network traffic model with random transmission rate. Preprint, 2000.
- [26] I. Norros. A storage model with self-similar input. *Queueing Systems*, 16:387-396, 1994.
- [27] A. Pakes. On the tails of waiting time distributions, *J. Appl. Probab.* 12 (1975) 555-564.
- [28] K. Park and W. Willinger (Eds.), *Self-Similar Network Traffic and Performance Evaluation*, John Wiley & Sons, New York (NY), 2000.
- [29] M. Parulekar and A.M. Makowski. Tail probabilities for a multiplexer with self-similar traffic. *Proc. IEEE INFOCOM*, 1996.
- [30] E.J.G. Pitman, Subexponential distribution functions. *J. Austral. Math. Soc. Ser. A* 29 (1980) 337-347.
- [31] S. Resnick and G. Samorodnitsky. Steady state distribution of the buffer content fro $M/G/\infty$ input fluid queues. preprint, 1999.
- [32] G. Samorodnitsky and M.S. Taqqu, *Stable Non-Gaussian Random Processes: Stochastic Models With Infinite Variance*, Chapman and Hal, New York (NY), 1994.
- [33] K. Sigman. Appendix: A primer on heavy-tailed distributions, *Queueing Systems*, 33 (1999) 261-275.
- [34] D. Stoyan. *Comparison Methods for Queues and Other Stochastic Models*. (English translation ed. D. J. Daley). Wiley, New York, 1983.
- [35] S. Vanichpun and A.M. Makowski. Positive correlations and buffer occupancy: lower bounds via supermodular ordering. *Proc. of IEEE INFOCOM*, June 2002, Vol. 3, 1298-1306.
- [36] W. Willinger, M.S. Taqqu, R. Sherman and D.V. Wilson. Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level. In *IEEE/ACM Trans. on Networking*, **5**(1), pp. 71-86, 1997.
- [37] C.H. Xia and Z. Liu. Queueing Systems with Long-range Dependent Input Process and Subexponential Service Times. *IBM Technical Report*, 2002.
- [38] C.H. Xia, Z. Liu, M.S. Squillante, L. Zhang and N. Malouch. Analysis of performance impact of drill-down techniques for Web traffic models. Dec. 2002, submitted for publication.
- [39] C.H. Xia, Z. Liu, M.S. Squillante, L. Zhang. Lower bounds for LRD/GI/1 queues with subexponential service times. Dec. 2002, submitted for publication.
- [40] A.J. Zeevi and P.W. Glynn. On the maximum workload of a queue fed by fractional Brownian motion. *Ann. Appl. Probab.* 10 (2000), no. 4, 1084-1099.
- [41] A.P. Zwart and O.J. Boxma. Sojourn time asymptotics in the $M/G/1$ processor sharing queue. *Queueing Systems*, 35 (2000) 141-166.