On the dependence structure of Gaussian queues

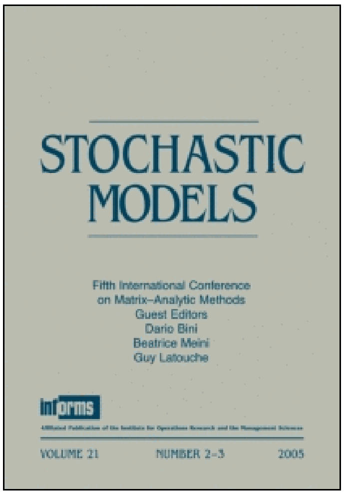
Es-Saghouani, A.; Mandjes, M.

DOI
10.1080/15326340902869887

Publication date
2009

Document Version
Final published version

Published in
Stochastic Models



Stochastic Models
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597301

On the Dependence Structure of Gaussian Queues

Abdelghafour Es-Saghouani [a];Michel Mandjes [abc]
[a] Korteweg-de Vries Institute for Mathematics, Amsterdam, TV, The Netherlands [b] CWI, Amsterdam, The Netherlands [c] EURANDOM, Eindhoven, The Netherlands

To cite this Article Es-Saghouani, Abdelghafour andMandjes, Michel(2009) 'On the Dependence Structure of Gaussian Queues', Stochastic Models, 25: 2, 221 — 247
To link to this Article: DOI: 10.1080/15326340902869887
URL: http://dx.doi.org/10.1080/15326340902869887

Stochastic Models, 25:221–247, 2009

ISSN: 1532-6349 print/1532-4214 online
DOI: 10.1080/15326340902869887

ON THE DEPENDENCE STRUCTURE OF GAUSSIAN QUEUES

Abdelghafour Es-Saghouani[1] and Michel Mandjes[1,2]

[1] *Korteweg-de Vries Institute for Mathematics, Amsterdam, TV, The Netherlands*
[2] *CWI, Amsterdam, The Netherlands; and EURANDOM, Eindhoven, The Netherlands*

□ *In this article we study Gaussian queues (that is, queues fed by Gaussian processes, such as fractional Brownian motion (fBm) and the integrated Ornstein–Uhlenbeck (iOU) process), with a focus on the dependence structure of the workload process. The main question is to what extent does the workload process inherit the dependence properties of the input process? We first present a specific notion of dependence that allows (in asymptotic regimes) explicit analysis. For the special cases of fBm and iOU, we analyze the behavior of this metric under a many sources scaling. Relying on (the generalized version of) Schilder's theorem, we are able to characterize its decay. We observe that the dependence structure of the input process essentially carries over to the workload process (in the asymptotic regime that we have chosen, in terms of our specific notion of dependence).*

Keywords Gaussian traffic; Sample-path large deviations; Schilder's theorem.

Mathematics Subject Classification Primary 60K25; Secondary 60F10, 60G15.

1. INTRODUCTION

Traffic measurement studies have provided convincing statistical evidence that in various networking environments traffic exhibits strong dependence over a wide range of time-scales. These studies, starting off in the early 1990s with the famous article by Leland et al.[16] on Ethernet traffic, showed that the traffic rate process was long-range dependent: with $X(t)$ the traffic rate at time t, the autocorrelation function $c(T)$ of

Received June 2007; Accepted January 2009
Address correspondence to Abdelghafour Es-Saghouani, Korteweg-de Vries Institute for Mathematics, Plantage Muidergracht 24, Amsterdam, TV 1018, The Netherlands; E-mail: a.es-saghouani@uva.nl

the traffic rate (i.e., the correlation coefficient between $X(0)$ and $X(T)$) vanishes extremely slow as a function of the lag T—more precisely: $c(T)$ decays so slowly that $\sum_{T\in\mathbb{N}} c(T) = \infty$.

This explains the interest in the performance evaluation of queues fed by long-range dependent traffic. Notably, the traffic models that were predominantly used till the mid-1990s did not allow for any long-range dependence; they usually corresponded to short-range dependent traffic processes (such as Poisson processes, Markov-modulated Poisson processes, or exponential on-off sources). In the late 1990s, Gaussian traffic models have gained more interest and popularity for modeling network traffic. One of their attractive features is that they cover a broad variety of dependence structures, ranging from short-range (e.g., the integrated Ornstein–Uhlenbeck process, Brownian motion) to long-range dependent (e.g., fractional Brownian motion with Hurst parameter $H > \frac{1}{2}$, see Ref.[16]). In Ref.[15] it is argued that the use of Gaussian traffic models is justified as long as the aggregation is sufficiently large, both in number of flows and time. We refer to Refs.[14,21,26] for excellent studies on network traffic modeling.

The fact that network traffic is long-range dependent is of crucial importance from the perspective of traffic engineering in communication networks. Where short-range dependent models usually lead to buffer overflow probabilities which decay exponentially in the buffer size, long-range dependent models are considerably less benign: in case of fractional Brownian motion input with Hurst parameter H, this decay is "Weibullian"[11,23] (that is, roughly like $\exp(-\alpha B^{2-2H})$, for some $\alpha > 0$, and B denoting the buffer size, which is slower than exponential for $H > \frac{1}{2}$), or even polynomial[25,30] (e.g., for on-off sources with regularly varying on-times). In other words: modeling traffic by a short-range dependent process would lead to estimates of the overflow probability that are considerably too optimistic.

For Gaussian queues (that is, queues fed by Gaussian processes), so far primary interest lay in the characterization of the buffer overflow probability. Notably, in two limiting regimes asymptotic results were obtained: in the large-buffer regime (where the buffer threshold grows large), and in the many-sources regime (in which the number of Gaussian inputs grows large, and the buffer and service speed are scaled accordingly[28]). Without exhaustively mentioning all relevant contributions, logarithmic asymptotics for the large-buffer case are due to Refs.[6,11], whereas exact asymptotics can be found in, e.g., Refs.[10,22], and the many-sources regime logarithmic asymptotics are in Refs.[1,4] and the exact asymptotics in Ref.[7].

To the best of our knowledge, hardly any attention has been paid to the characterization of the dependence structure of the workload process of

Gaussian queues. This is remarkable, as from an engineering standpoint, knowledge of the dependence structure is clearly quite relevant. Most importantly, it would give us a handle on the timescale after which it is justified to approximate transient probabilities by their steady-state counterpart. Also procedures that "learn" the characteristics of the input process by observing the workload process[19] would greatly benefit from insights into the degree of dependence between two subsequent observations. More precisely, it can be determined from what time-scale one could safely neglect the dependence between the observations.

Seen from a more mathematical angle, an interesting fundamental question is: to what extent the dependence structure of the input process is inherited by the workload process? Or put differently, does long-range dependent input give rise to a long-range dependent workload process? Our article shows that indeed for fractional Brownian motion (in the sequel abbreviated to fBm) and integrated Ornstein–Uhlenbeck (iOU) the dependence structure of the workload process strongly resembles that of the input process: both exhibit Weibullian decay for fBm and exponential decay for iOU.

The first aim would be to analyze, with Q_t denoting the workload at time t, the covariance

$$\mathbb{C}\text{ov}(Q_0, Q_T) = \mathbb{E}(Q_0 Q_T) - \mathbb{E}Q_0 \cdot \mathbb{E}Q_T = \mathbb{E}(Q_0 Q_T) - (\mathbb{E}Q_0)^2, \qquad (1)$$

or the corresponding correlation coefficient. It is not clear what methodology can be used to analyze these covariances. It is noted, for instance, that large-deviation types of results are not of any help here, as covariances are quantities related to expected values, which cannot be represented as rare-event probabilities (where we also recall that in the setting of queues with Gaussian input, apart from a few special cases, one has not even succeeded so far as to compute the mean workload $\mathbb{E}Q_0$).

To overcome this problem we have chosen the following solution:

- We choose a measure for dependence that is more tractable than (1). This new metric measures the difference between $\log \mathbb{P}(Q_0 > p, Q_T > q)$ and $\log(\mathbb{P}(Q_0 > p)\mathbb{P}(Q_T > q))$, for given $p, q > 0$; informally, the more independent $\{Q_0 > p\}$ and $\{Q_T > q\}$ are, the smaller the distance. A more specific goal is to characterize for fBm and iOU how our metric decays to 0 when T grows to infinity.
- We work in the many sources asymptotic regime[28] that was previously mentioned. As a consequence, we can use an extensive set of useful techniques, most notably (sample-path) large-deviation results, particularly (the generalized version of) Schilder's theorem.

  More precisely, the setting we consider is as follows: we let n i.i.d. Gaussian processes $A_1(\cdot), \ldots, A_n(\cdot)$ feed into a queue in which both the service speed and the buffer content are scaled by n; we denote the workload of the resulting queueing system at time t by Q_t^n. The results presented in this article are asymptotic in n.

As previously mentioned, we specialize in the important cases of fBm and iOU input. Our main conclusion is that by using the metric introduced and considering the many-sources regime, the dependence structure of the input process essentially carries over to the workload process.

Above we argued that Gaussian models (and in particular fBm) are good traffic descriptors in the setting of communication networks as long as there is sufficient aggregation[15]. We stress, however, that this is an issue that should be handled with care, as it depends very much on the situation at hand whether this is the case. Reference[21] presents a systematic assessment of this issue. There the stochastic properties of a superposition of n sources with heavy-tailed on-times (or bursts), and alternatively a corresponding M/G/∞ input model, is considered after rescaling time with T. Conditions are discussed under which the limiting process indeed looks like fBm, while in other situations α-stable Lévy motion is more suitable. More specifically, if the situation at hand is such that the rate at which bursts are generated is large in relation to the tail of the distribution of the burst duration, fBm is an appropriate approximation. For a more detailed discussion we refer to as an example Refs.[12,21].

2. PRELIMINARIES

2.1. Gaussian Processes

Let $A_i(\cdot)$ denote a sequence of i.i.d. centered Gaussian processes with continuous sample paths and stationary increments, $i = 1, \ldots, n$; it is assumed that $A_i(0) \equiv 0$ for all i. For $s < t$, we interpret $A_i(s, t) := A_i(t) - A_i(s)$ as the amount of the traffic generated by the ith Gaussian source in the time interval $(s, t]$; we let the $A_i(t)$ be "two-sided," that is, defined for all $t \in \mathbb{R}$.

We denote by $A(t)$ the generic Gaussian process corresponding to a single source, and $A(s, t) := A(t) - A(s)$. A (centered) Gaussian process is characterized by its variance function $v(\cdot)$ (which is necessarily continuous); because of the stationarity of the increments of our process, we have $\mathbb{V}\mathrm{ar}\, A(s, t) = v(t - s)$ for $s < t$.

In the sequel, we frequently work with the bivariate random variable $(A(-s, 0), A(T - t, T))$ (for large values of T). Its distribution is a bivariate normal distribution with zero mean vector and covariance matrix $\Sigma_T(s,t)$ given by

$$\Sigma_T(s,t) := \begin{pmatrix} v(s) & \Gamma_T(s,t) \\ \Gamma_T(s,t) & v(t) \end{pmatrix},$$

with $\Gamma_T(s,t) := \mathbb{C}\mathrm{ov}(A(-s,0), A(T-t,T))$. For $s > 0$ and $0 < t < T$, this covariance reduces to

$$\Gamma_T(s,t) = \frac{v(T+s) - v(T) - v(T-t+s) + v(T-t)}{2};$$

for other ranges of s and t similar expressions can be given.

Gaussian sources have the intrinsic inconvenience that in principle negative traffic can be generated: $A(s,t)$ (with $t > s$) is not necessarily nonnegative. When using the representation for the workload at time t (take for ease a queue fed by a single Gaussian source, with service rate $c > 0$)

$$Q_t := \sup_{s \geq 0} \{A(t-s,t) - cs\},$$

this turns out to not be an issue: the probabilistic properties of the above functional of the Gaussian process $A(\cdot)$ can be evaluated, irrespective of whether the input process allows negative increments.

In our study we focus, without loss of generality, on centered Gaussian processes, but it is straightforward to adapt the results to the case of noncentered Gaussian processes, as the queueing system in which the input has mean rate $m \neq 0$ and service rate c (larger than m to ensure stability) coincides with the system with centered input and service rate $c - m$.

In this article we focus on two special Gaussian processes: (standard) fractional Brownian motion (or *fBm*; $v(t) = t^{2H}$, with $H \in (0,1)$), and integrated Ornstein–Uhlenbeck (or *iOU*; $v(t) = t - 1 + e^{-t}$).

Lemma 2.1.1. *Fix $s, t > 0$; let $t < T$.*

- fBm. *For $H > \frac{1}{2}$, $\Gamma_T(s,t)$ is positive, and decreases to 0 when $T \to \infty$. For $H < \frac{1}{2}$, $\Gamma_T(s,t)$ is negative, and increases to 0 when $T \to \infty$.*
- iOU. *$\Gamma_T(s,t)$ is positive, and decreases to 0 when $T \to \infty$.*

Proof. First focus on fBm. It is immediate that

$$\Gamma_T(s,t) = \gamma_T^{(\mathrm{fBm})}(s,t) := \frac{1}{2}((T+s)^{2H} - T^{2H} - (T-t+s)^{2H} + (T-t)^{2H}).$$

Consider $H > \frac{1}{2}$. It is readily checked that in order to show that $\Gamma_T(s, t)$ is positive, we have to prove that

$$1 - (1 - t)^{2H} < (1 + s)^{2H} - (1 - t + s)^{2H},$$

or, equivalently, that $(1 + s)^{2H} - (1 - t + s)^{2H}$ increases in $s > 0$ (for all $t \in (0, 1)$). Differentiation with respect to s leads to the claim $(1 + s)^{2H-1} > (1 + s - t)^{2H-1}$, which is indeed true for $H > \frac{1}{2}$. The fact that $\Gamma_T(s, t)$ is decreasing in T (with limit 0) is proven in the same way. The case $H < \frac{1}{2}$ can be dealt with similarly.

For iOU,

$$\begin{aligned}\Gamma_T(s, t) = \gamma_T^{(\mathrm{iOU})}(s, t) &:= \frac{1}{2}\big(e^{-T-s} - e^{-T} - e^{-T+t-s} + e^{-T+t}\big)\\ &= \frac{1}{2}(1 - e^{-s})(e^{t} - 1)e^{-T},\end{aligned}$$

which is indeed positive and decreasing in T. □

2.2. Large Deviations Results

In this subsection, we give a brief description of the main results from the large-deviations theory for Gaussian processes. The proofs of the theorems presented here can be found in Refs.[8,9]; for more background see Ref.[17]. We first state Cramér's theorem, which relates to d-dimensional random variables, and then Schilder's theorem, which describes the sample-path large deviations of Gaussian processes.

Let $X \in \mathbb{R}^d$ be a d-dimensional random vector. We denote the moment generating function of X by $M(\theta) := \mathbb{E}\big(\exp(\langle\theta, X\rangle)\big)$ and its logarithm by $\Lambda(\theta) := \log M(\theta)$. Its convex conjugate $\Lambda^\star$ is defined by $\Lambda^\star(x) := \sup_{\theta\in\mathbb{R}^d}(\langle\theta, x\rangle - \Lambda(\theta))$, with $\langle\cdot,\cdot\rangle$ denoting the usual inner product: $\langle a, b\rangle := a^T b = \sum_{i=1}^d a_i b_i$. We first state (the multivariate version of) Cramér's theorem which characterizes the logarithmic rate of the convergence of the empirical mean of i.i.d. random vectors in $\mathbb{R}^d$.

Theorem 2.2.1 (Multivariate Cramér). *Let $X_i \in \mathbb{R}^d$ be i.i.d. d-dimensional random vectors, distributed as a random vector X. Then the following ldp applies*[8,9]*:*

(a) *For any closed set $F \subset \mathbb{R}^d$,*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \in F\right) \le -\inf_{x\in F} \Lambda^\star(x);$$

(b) *For any open set* $G \subset \mathbb{R}^d$,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \in G\right) \geq -\inf_{x\in G} \Lambda^\star(x),$$

where the large deviations rate function $\Lambda^\star(\cdot)$ *is as previously given.*

Remark 2.2.1. Consider the case that X has a multivariate normal distribution with mean vector 0 and $d \times d$-nonsingular covariance matrix Σ. Then using $\Lambda(\theta) = \frac{1}{2}\theta^T \Sigma \theta$ we obtain

$$\theta^\star = \Sigma^{-1}x \quad \text{and} \quad \Lambda^\star(x) = \frac{1}{2}x^T\Sigma^{-1}x, \tag{2}$$

where $\theta^\star$ is the optimizer in the definition of $\Lambda^\star$.

Before stating the generalized Schilder's theorem, we first sketch the framework of the Schilder's sample path large deviations principle as established in Ref.[3], (see also Ref.[9]). We use the same set-up and notation as in Refs.[17,18]. We consider n i.i.d. centered Gaussian processes $A_i(\cdot)$ and define the path space Ω as

$$\Omega := \left\{\omega : \mathbb{R} \to \mathbb{R},\ \text{continuous},\ \omega(0) = 0,\ \lim_{|t|\to\infty} \frac{\omega(t)}{1+|t|} = 0\right\}$$

which becomes a Banach space by equipping it with the norm

$$\|\omega\|_\Omega := \sup_{t\in\mathbb{R}} \frac{|\omega(t)|}{1+|t|}.$$

In Addie et al.[1] it is shown that $A(\cdot)$ can be realized in Ω under Assumption 2.2.1; it is clear that both fBm and iOU satisfy this requirement.

Assumption 2.2.1. The variance function $v(\cdot)$ of the process $A(\cdot)$ is continuous and it satisfies

$$\lim_{t\to\infty} \frac{v(t)}{t^\alpha} = 0 \tag{3}$$

for some $\alpha \in (0, 2)$.

Next we introduce the reproducing kernel Hilbert space $R \subset \Omega$, with the property that its elements are roughly as smooth as the covariance function $\Gamma(s,\cdot)$, see Adler[2] for more details. We start from a subspace $R^\star \subset \Omega$, defined by

$$R^\star := \left\{\omega \in \Omega,\ \omega(\cdot) = \sum_{i=1}^{n} a_i \Gamma(s_i,\cdot),\ a_i, s_i \in \mathbb{R},\ n \in \mathbb{N}\right\},$$

with $\Gamma(s,t) := \mathbb{C}\mathrm{ov}(A(0,s), A(0,t))$. The inner product on this space $R^\star$ is defined as follows, for $\omega_a, \omega_b \in R^\star$:

$$\langle \omega_a, \omega_b \rangle_R := \left\langle \sum_{i=1}^{n} a_i \Gamma(s_i,\cdot), \sum_{j=1}^{n} b_j \Gamma(s_j,\cdot) \right\rangle_R = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i b_j \Gamma(s_i, s_j). \tag{4}$$

Now we can introduce the norm $\|\omega\|_R := \sqrt{\langle \omega, \omega \rangle_R}$. The closure of $R^\star$ under this norm is defined as the space R. Now we can define the rate function of the sample-path large-deviations principle (ldp):

$$I(\omega) := \begin{cases} \dfrac{1}{2}\|\omega\|_R^2 & \text{if } \omega \in R; \\ \infty & \text{otherwise.} \end{cases} \tag{5}$$

For a sequence of n i.i.d. centered Gaussian processes, the following sample-path ldp holds[3,9].

Theorem 2.2.2 (Generalized Schilder)**.** *The following sample-path ldp applies:*

(a) *For any closed set $F \subset \Omega$,*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} A_i(\cdot) \in F\right) \le -\inf_{\omega \in F} I(\omega);$$

(b) *For any open set $G \subset \Omega$,*

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} A_i(\cdot) \in G\right) \ge -\inf_{\omega \in G} I(\omega).$$

3. MAIN RESULTS

As mentioned in the introduction of this article, our main interest lies in the investigation of the dependence structure of the workload process. Since only for the case of Brownian motion input the workload distribution has been found explicitly, we resort to an asymptotic framework, viz. the so-called many-sources regime. In this regime, the number of Gaussian inputs, say n, grows large, and the service rate is scaled accordingly. In this framework, the stationary workload process is given by

$$Q_t^n := \sup_{s\le t} \sum_{i=1}^n A_i(s,t) - nc(t-s) = \sup_{s\ge 0} \sum_{i=1}^n A_i(t-s,t) - ncs. \qquad (6)$$

As we wish to investigate the dependence structure of the workload process, we could try to characterize the autocorrelation

$$\delta_n(T) := \frac{\mathbb{E}Q_0^n Q_T^n - (\mathbb{E}Q_0^n)(\mathbb{E}Q_T^n)}{\sqrt{\mathbb{V}\mathrm{ar}(Q_0^n)}\sqrt{\mathbb{V}\mathrm{ar}(Q_T^n)}} = \frac{\mathbb{E}Q_0^n Q_T^n - (\mathbb{E}Q_0^n)(\mathbb{E}Q_0^n)}{\mathbb{V}\mathrm{ar}(Q_0^n)}.$$

It is evident that $\delta_n(T) \downarrow 0$ as $T \uparrow \infty$, but the question is how fast it vanishes.

Unfortunately, this notion of dependence is hard to handle—not even an explicit expression for $\mathbb{E}Q_0^n$ is known for non-Brownian Gaussian input processes. We therefore introduce an alternative notion of dependence. The following metric describes the degree of dependence between the events $\{Q_0^n > np\}$ and $\{Q_T^n > nq\}$ for positive p, q.

Definition 3.1. For given positive numbers p, q define

$$\kappa_n(T) := \frac{\mathbb{P}\big(Q_0^n > np, Q_T^n > nq\big)}{\mathbb{P}\big(Q_0^n > np\big)\mathbb{P}\big(Q_T^n > nq\big)}. \qquad (7)$$

Furthermore, let $\kappa(T)$ be the limit of $\log \kappa_n(T)/n$ as $n \to \infty$.

It is evident that various other dependence measures could be thought of. Our measure is reminiscent of quantities used when defining mixing conditions, see, e.g., Ref.[5]. For instance, with $\mathscr{A}_s^t$ defining the σ-field $\sigma(Q_u : s \le u \le t)$, and

$$\alpha(\mathscr{A}, \mathscr{B}) := \sup_{A\in\mathscr{A}, B\in\mathscr{B}} |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|,$$

we say that Q_t is strongly mixing if $\alpha(T) := \sup_s \alpha(\mathscr{A}_{-\infty}^s, \mathscr{A}_{s+T}^\infty) \to 0$ as $T \to \infty$. The relation between the decay of $\kappa(T)$ and mixing conditions is not *a priori* clear; also due to the fact that a supremum over $A \in \mathscr{A}_{-\infty}^s$ and $B \in \mathscr{A}_{s+T}^\infty$ needs to be computed, it is typically hard to characterize the decay of $\alpha(T)$ and related quantities.

Before stating the main theorems of this section, we first give the logarithmic asymptotics of the marginal probabilities involved in Definition 3.1. They are given by

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big(Q_0^n > np\big) = -\inf_{s>0} \frac{(p+cs)^2}{2v(s)} \qquad (8)$$

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}\big(Q_T^n > nq\big) = -\inf_{t>0}\frac{(q+ct)^2}{2v(t)}, \tag{9}$$

using that the queue is in stationarity at both epochs; see, for instance Ref.[1]. In Ref.[7], the following lemma was proven; it entails that the infima over s and t are attained and are unique under a specific assumption on the variance function.

Lemma 3.1. *Suppose that the standard deviation function $\sigma(t) := \sqrt{v(t)}$ of the generic input process $A(\cdot)$ is such that $\sigma(t) \in \mathscr{C}^2([0,\infty))$ is strictly increasing and strictly concave. Then the right-hand-sides of (8) and (9) have unique minimizers. Concavity of $\sigma(t)$ is equivalent to requiring that*

$$2v(t)v''(t) - (v'(t))^2 \leq 0. \tag{10}$$

We denote the minimizers by $s^\star$ and $t^\star$. It is readily checked that they solve

$$\begin{cases} 2cv(s) = (p+cs)v'(s); \\ 2cv(t) = (q+ct)v'(t). \end{cases} \tag{11}$$

Now we give the main results of this article. Theorem 3.1 states that for fBm input $\kappa(T)$ decays to zero and its decay rate is T^{2H-2} as $T\to\infty$, which indicates that the workload process has essentially the same dependence structure as the input process. As will be discussed in more detail in Section 5, this means that the workload process is (in our metric) long-range dependent if the Hurst parameter H is greater than $\frac{1}{2}$. For fBm, $\sigma(t) = t^H$ is concave, so Lemma 3.1 applies, and (11) has a unique solution; in fact, $s^\star$ and $t^\star$ can be explicitly calculated, and are given through

$$s^\star := \frac{p}{c}\frac{H}{1-H}; \quad t^\star := \frac{q}{c}\frac{H}{1-H}. \tag{12}$$

Theorem 3.1 (fBm Input). *If the input process is fBm we have the following logarithmic asymptotics for $\kappa_n(T)$:*

$$\begin{aligned}\kappa(T) = \lim_{n\to\infty}\frac{1}{n}\log\kappa_n(T) &= \frac{(p+cs^\star)(q+ct^\star)}{v(s^\star)v(t^\star)}\cdot\frac{1}{2}s^\star t^\star(2H)(2H-1)T^{2H-2} \\ &\quad + o(T^{2H-2}) \\ &= \frac{(2H-1)c^2}{H}s^{\star 2-2H}t^{\star 2-2H}T^{2H-2} + o(T^{2H-2}). \end{aligned} \tag{13}$$

It is interesting to compare this result to the dependence structure of the input process. We could look at a counterpart of $\kappa(T)$, for instance,

$$\lambda(T) := \lim_{n\to\infty} \frac{1}{n} \log\left(\frac{\mathbb{P}\big(\sum_{i=1}^n A_i(0,1) > np, \sum_{i=1}^n A_i(T,T+1) > nq\big)}{\mathbb{P}\big(\sum_{i=1}^n A_i(0,1) > np\big)\mathbb{P}\big(\sum_{i=1}^n A_i(T,T+1) > nq\big)} \right),$$

and consider its decay for T large. Denoting $\gamma_T^{(\mathrm{fBm})}(1,1)$ by $\gamma_T^{(\mathrm{fBm})}$ (see Lemma 2.1.1), we have that $\gamma_T^{(\mathrm{fBm})} \sim \frac{1}{2}T^{2H-2} \to 0$ as $T \to \infty$. Then straightforward computations show that the bivariate version of Cramér's theorem implies that

$$\lambda(T) = -\frac{1}{2}(p,q)\begin{pmatrix} v(1) & \gamma_T^{(\mathrm{fBm})} \\ \gamma_T^{(\mathrm{fBm})} & v(1) \end{pmatrix}^{-1}\begin{pmatrix} p \\ q \end{pmatrix} + \frac{p^2}{2v(1)} + \frac{q^2}{2v(1)} \sim pq\gamma_T^{(\mathrm{fBm})},$$

for T large, i.e., also decaying as T^{2H-2}! The above arguments provide support for the claim that in this metric, the workload process has essentially the same dependence structure as the input process.

Remark 3.1. For $H = \frac{1}{2}$, we can explicitly calculate $\kappa(T)$ for any T, relying on the formulas for the transient behavior of reflected Brownian motion, see, e.g., Ref.[13] (p. 49). It turns out that for all $T > c^{-1} \cdot (\sqrt{p} + \sqrt{q})^2$ it holds that $\kappa(T) = 0$; observe that the existence of such a threshold value could be anticipated due to the independent increments. Also note that this result is in line with Theorem 3.1.

Now consider the case of iOU input. In this case, $s^\star$ and $t^\star$ cannot be explicitly calculated. They are uniquely determined though, as can be seen as follows. Criterion (10) reduces to

$$\varphi(t) := 2te^{-t} + e^{-2t} - 1 \leq 0,$$

which is true because $\varphi(0) = 0$ and $\varphi'(t) = e^{-t}(2 - 2t - 2e^{-t}) \leq 0$.

Theorem 3.2 states that for iOU input, the speed of convergence of $\kappa(T)$ to 0 as $T \to \infty$ is e^{-T}.

Theorem 3.2 (iOU Input). *If the input process is iOU, we have the following logarithmic asymptotics for $\kappa_n(T)$:*

$$\begin{aligned}\kappa(T) = \lim_{n\to\infty}\frac{1}{n}\log\kappa_n(T) &= \frac{(p+cs^\star)(q+ct^\star)}{v(s^\star)v(t^\star)} \cdot \frac{1}{2}(1-e^{-s^\star})(e^{t^\star}-1)e^{-T} \\ &\quad + o(e^{-T}) \\ &= 2c^2 e^{-(T-t^\star)} + o(e^{-T}). \qquad (14)\end{aligned}$$

In this case, it can be verified that

$$\gamma_T^{(\mathrm{iOU})} := \gamma_T^{(\mathrm{iOU})}(1,1) \sim \frac{1}{2} e^{-T} \cdot \left(e - 2 + \frac{1}{e}\right).$$

As we have $\lambda(T) \sim pq\gamma_T^{(\mathrm{iOU})}$, it again holds that the dependence structure of the workload process essentially coincides with that of the input process (i.e., both $\kappa(T)$ and $\lambda(T)$ are roughly proportional to e^{-T}).

4. PROOFS

In this section, we give the proofs of the results that we stated in the previous section. In the first subsection we derive a number of generic results, while we specialize in fBm and iOU in the last part of the section.

4.1. General Results

The results of this subsection hold for any type of Gaussian sources (i.e., we do not restrict ourselves to fBm and iOU), the only exception being Proposition 4.1.2. We first define two sets of paths in Ω that play a crucial role in our analysis:

$$\begin{aligned}\mathscr{S}_T := \{f \in \Omega : \exists s > 0,\ \exists t > 0 : -f(-s) > p + cs,\\ f(T) - f(T-t) > q + ct\}; \qquad (15)\\ \mathscr{S}_T(s,t) := \{f \in \Omega : -f(-s) > p + cs,\ f(T) - f(T-t) > q + ct\}. \qquad (16)\end{aligned}$$

Observe that $\mathscr{S}_T$ is the union (over all $s, t > 0$) of the $\mathscr{S}_T(s,t)$. Interestingly, the set of paths $\mathscr{S}_T$ directly relates to the "joint overflow event" $\{Q_0^n > np, Q_T^n > nq\}$, as follows from the next lemma.

Lemma 4.1.1. *For any $p, q > 0$,*

$$\mathbb{P}\big(Q_0^n > np, Q_T^n > nq\big) = \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n A_i(\cdot) \in \mathscr{S}_T\right).$$

Proof. This follows by applying (6):

$$\begin{aligned}&\mathbb{P}\big(Q_0^n > np, Q_T^n > nq\big)\\ &\quad = \mathbb{P}\left(\sup_{s>0}\left\{\sum_{i=1}^n A_i(-s,0) - ncs\right\} > np,\ \sup_{t>0}\left\{\sum_{i=1}^n A_i(T-t,T) - nct\right\} > nq\right)\\ &\quad = \mathbb{P}\left(\exists s > 0 : \sum_{i=1}^n A_i(-s,0) - ncs > np,\ \exists t > 0 : \sum_{i=1}^n A_i(T-t,T) - nct > nq\right)\end{aligned}$$

$$= \mathbb{P}\left(\exists s > 0 : \sum_{i=1}^{n} \frac{A_i(-s,0)}{n} > p + cs, \exists t > 0 : \sum_{i=1}^{n} \frac{A_i(T-t,T)}{n} > q + ct\right)$$

$$= \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} A_i(\cdot) \in \mathcal{S}_T\right),$$

which proves the claimed. □

In the sequel we frequently use the following bivariate normal large-deviations rate function:

$$\Lambda_T^\star(p+cs, q+ct) := \frac{1}{2}(p+cs, q+ct)(\Sigma_T(s,t))^{-1}\begin{pmatrix} p+cs \\ q+ct \end{pmatrix}.$$

By explicitly calculating the matrix inverse, we obtain that $\Lambda_T^\star(p+cs, q+ct)$ can be written in the following alternative form:

$$\frac{1}{2}\frac{v(s)v(t)}{v(s)v(t)-\Gamma_T(s,t)^2}\left(\frac{(p+cs)^2}{v(s)} + \frac{(q+ct)^2}{v(t)} - 2\frac{(p+cs)(q+ct)\Gamma_T(s,t)}{v(s)v(t)}\right). \tag{17}$$

The next lemma determines the decay rate of the most likely path in $\mathcal{S}_T(s,t)$, for fixed values of s and t. It turns out that there are three different regimes.

Lemma 4.1.2. *For any $p, q > 0$,*

$$\inf_{f \in \mathcal{S}_T(s,t)} I(f) = \overline{\Lambda}_T^\star(p+cs, q+ct),$$

where $\overline{\Lambda}_T^\star(p+cs, q+ct)$ equals

$$\frac{(p+cs)^2}{2v(s)} \quad \text{if} \quad \frac{\Gamma_T(s,t)}{v(s)}(p+cs) > q+ct; \tag{18}$$

$$\frac{(q+ct)^2}{2v(t)} \quad \text{if} \quad \frac{\Gamma_T(s,t)}{v(t)}(q+ct) > p+cs; \tag{19}$$

$\Lambda_T^\star(p+cs, q+ct)$ *otherwise.*

Proof. Multiplication of (18) and (19) would lead to

$$\Gamma_T^2(s,t) > v(s)v(t),$$

and hence Cauchy–Schwarz implies that the conditions in (18) and (19) cannot apply simultaneously.

Then recognize

$$\frac{\Gamma_T(s,t)}{v(s)}(p+cs) = \mathbb{E}(A(T-t,T)\,|\,A(-s,0)=p+cs);$$
$$\frac{\Gamma_T(s,t)}{v(t)}(q+ct) = \mathbb{E}(A(-s,0)\,|\,A(T-t,T)=q+ct).$$

The stated now follows immediately from the bivariate version of Cramér's theorem; see the solution of Exercise 4.1.9 as given in Ref.[17] (p. 42). □

The proof of the next proposition relies on Lemma A.1, which is stated and proven in the Appendix.

Proposition 4.1.1. *For any $p, q > 0$,*

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}\big(Q_0^n > np, Q_T^n > nq\big) = -\inf_{f\in\mathscr{S}_T} I(f) = -\inf_{s,t>0}\overline{\Lambda}^{\star}_T(p+cs, q+ct).$$

Proof. From "Schilder" and Lemma 4.1.1 we have

$$-\inf_{f\in\mathscr{S}_T} I(f) \le \lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}\big(Q_0^n > np, Q_T^n > nq\big) \le -\inf_{f\in\overline{\mathscr{S}_T}} I(f).$$

We first show that the above inequalities are actually equalities, by establishing that $\mathscr{S}_T$ is an I-continuity set, that is,

$$\inf_{f\in\mathscr{S}_T} I(f) = \inf_{f\in\overline{\mathscr{S}_T}} I(f), \tag{20}$$

where the $\overline{\mathscr{S}_T}$ denotes the closure of $\mathscr{S}_T$, and is given in Lemma A.1.

This can be done in the same way as in the appendix of Ref.[20]. Choose an arbitrary path f in $\overline{\mathscr{S}_T}\cap R$, and approximate it by a path in $\mathscr{S}_T$, as follows. We use the sets $\mathscr{S}(s), \mathscr{S}_T(t), \overline{\mathscr{S}(s)}$, and $\overline{\mathscr{S}_T(t)}$ as defined in the appendix. Due to Lemma A.1 we have that $f\in\overline{\mathscr{S}(s)}\cap\overline{\mathscr{S}_T(t)}$ for some $s,t>0$. Let $\eta(\cdot)$ be a path in R that is strictly increasing and taking negative values for $u\in(-\infty,0)$ and positive values for $u\in(0,\infty)$ (for instance, $\eta(u) := \mathrm{sgn}(u)\sqrt{|u|}$ or $\arctan u$). Define

$$f_n(u) := f(u) + \frac{\eta(u)}{n}.$$

Then $f_n\in\mathscr{S}(s)\cap\mathscr{S}_T(t)$ as, for any $s>0$, it holds that

$$-f_n(-s) = -f(-s) - \frac{\eta(-s)}{n} \ge p + cs - \frac{\eta(-s)}{n} > p + cs$$

and, for any $t > 0$,

$$f_n(T) - f_n(T-t) = f(T) - f(T-t) + \frac{\eta(T) - \eta(T-t)}{n}$$
$$\geq q + ct + \frac{\eta(T) - \eta(T-t)}{n} > q + ct.$$

Moreover, we have, for $n \to \infty$,

$$\|f_n\|_R^2 = \left\| f + \frac{1}{n}\eta \right\|_R^2 \to \|f\|_R^2,$$

which proves (20) and therefore also the first equality of the proposition.

The above entails that the decay rate of our interest equals

$$\inf_{s,t>0} \inf_{f \in (\mathscr{S}(s) \cap \mathscr{S}_T(t))} I(f).$$

Recall from (15) and (16) that $\mathscr{S}_T$ is the union over all $s \geq 0$ and $t \geq 0$ of the $\mathscr{S}_T(s,t)$, and observe that $\mathscr{S}_T(s,t) = \mathscr{S}(s) \cap \mathscr{S}_T(t)$. The second equality of the proposition now follows directly from Lemma 4.1.2. □

Proposition 4.1.2. *Consider fBm or iOU. For any $p, q > 0$, and T large enough*

$$\inf_{s,t>0} \overline{\Lambda}_T^\star(p+cs, q+ct) = \inf_{s,t>0} \Lambda_T^\star(p+cs, q+ct). \tag{21}$$

Proof. As, for any $s, t > 0$ and any $T > 0$, it holds that $\overline{\Lambda}_T^\star(s,t) \leq \Lambda_T^\star(s,t)$, it suffices to prove that, for T sufficiently large,

$$\inf_{s,t>0} \overline{\Lambda}_T^\star(p+cs, q+ct) \geq \inf_{s,t>0} \Lambda_T^\star(p+cs, q+ct).$$

We prove that in a number of steps.

- **Step 1.** As, evidently,

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n A_i(-s,0) > p + cs, \frac{1}{n} \sum_{i=1}^n A_i(T-t,0) > q + ct \right)$$
$$\leq \min\left\{ \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n A_i(-s,0) > p + cs \right), \right.$$
$$\left. \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n A_i(T-t,0) > q + ct \right) \right\},$$

we have that

$$\overline{\Lambda}_T(p+cs, q+ct) \geq \max\left\{\frac{(p+cs)^2}{2v(s)}, \frac{(q+ct)^2}{2v(t)}\right\}.$$

- **Step 2.** Lemma 2.1.1 states that, for any fixed s, t, $\Gamma_T(s,t) \to 0$ as $T \to \infty$. It can be checked that this implies that also $\overline{\Lambda_T^\star}(p+cs, q+ct) \to \Lambda_\infty^\star(p+cs, q+ct)$ as $T \to \infty$, where

$$\Lambda_\infty^\star(p+cs, q+ct) = \frac{(p+cs)^2}{v(s)} + \frac{(q+ct)^2}{v(t)}$$

(to this end, observe that for any fixed s, t, the conditions in (18) and (19) are not fulfilled for T sufficiently large). It is clear that, when taking the infimum of $\Lambda_\infty^\star(p+cs, q+ct)$ over $s, t > 0$, the expression decouples into the sum of an infimum over s and an infimum over t. Both individual infima have a unique minimizer, namely, $s^\star$ and $t^\star$ as introduced earlier. In the remainder of the proof, we use the notation $\ell := \Lambda_\infty^\star(p+cs^\star, q+ct^\star)$. It is clear that the above implies that for T sufficiently large

$$\inf_{s,t>0} \overline{\Lambda_T^\star}(p+cs, q+ct) \leq \overline{\Lambda_T^\star}(p+cs^\star, q+ct^\star) \leq 2\ell. \tag{22}$$

- **Step 3.** Using Step 1, for any $t > 0$, both as $s \downarrow 0$ and as $s \to \infty$, uniformly in T,

$$\overline{\Lambda_T^\star}(p+cs, q+ct) \geq \frac{(p+cs)^2}{2v(s)} \to \infty;$$

likewise, for any $s > 0$, both as $t \downarrow 0$ and as $t \to \infty$, we have that $\overline{\Lambda_T^\star}(p+cs, q+ct) \to \infty$. It implies that we can find $\underline{\varepsilon}, \bar{\varepsilon} \in (0, \infty)$, independent of T, such that for all $s, t \notin [\underline{\varepsilon}, \bar{\varepsilon}]$ it holds that $\overline{\Lambda_T^\star}(p+cs, q+ct) \geq 3\ell$.
- **Step 4.** Using (22) and Step 3, we conclude that we can restrict ourselves, for T sufficiently large, to $s, t \in [\underline{\varepsilon}, \bar{\varepsilon}]$. Again using that $\Gamma_T(s,t) \to 0$ as $T \to \infty$ (by virtue of Lemma 2.1.1), it is seen that for T large enough, for all $s, t \in [\underline{\varepsilon}, \bar{\varepsilon}]$ the conditions in (18) and (19) are not satisfied, and therefore we have that $\overline{\Lambda_T^\star}(p+cs, q+ct) = \Lambda_T^\star(p+cs, q+ct)$. This entails that, for T sufficiently large,

$$\begin{aligned}\inf_{s,t>0} \overline{\Lambda_T^\star}(p+cs, q+ct) &= \inf_{s,t\in[\underline{\varepsilon},\bar{\varepsilon}]} \overline{\Lambda_T^\star}(p+cs, q+ct) = \inf_{s,t\in[\underline{\varepsilon},\bar{\varepsilon}]} \Lambda_T^\star(p+cs, q+ct) \\ &\geq \inf_{s,t>0} \Lambda_T^\star(p+cs, q+ct).\end{aligned}$$

This concludes the proof. $\square$

In view of the fact that $\Lambda^\star_T(p+cs, q+ct) \to \Lambda^\star_\infty(p+cs, q+ct)$, we now also have that a sequence of local optimizers of the right-hand side of (21), say $(s^\star_T, t^\star_T)$, converges to $(s^\star, t^\star)$ as $T \to \infty$. Relying on Taylor expansions around $(s^\star, t^\star)$, the vector $(s^\star_T, t^\star_T)$ at which the function $\Lambda^\star_T(p+cs, q+ct)$ is minimum solves the following system:

$$\begin{aligned}(p+cs)&(2cv(s)-(p+cs)v'(s))\\ &= 2\left(\frac{q+ct}{v(t)}\right)\left((cv(s)-(p+cs)v'(s))\Gamma_T(s,t)+(p+cs)v(s)\frac{\partial\Gamma_T}{\partial s}(s,t)\right);\end{aligned} \tag{23}$$

$$\begin{aligned}(q+ct)&(2cv(t)-(q+ct)v'(t))\\ &= 2\left(\frac{p+cs}{v(s)}\right)\left((cv(t)-(q+ct)v'(t))\Gamma_T(s,t)+(q+ct)v(t)\frac{\partial\Gamma_T}{\partial t}(s,t)\right)\end{aligned} \tag{24}$$

where the partial derivatives of $\Gamma_T(s,t)$ with respect to s and t are given by

$$\frac{\partial\Gamma_T}{\partial s}(s,t) = \frac{1}{2}(v'(T+s)-v'(T-t+s));$$
$$\frac{\partial\Gamma_T}{\partial t}(s,t) = \frac{1}{2}(v'(T-t+s)-v'(T-t)).$$

In the next two subsections, we study the system (23)–(24), for both fBm and iOU, by analyzing the behavior of $s^\star_T, t^\star_T$ in detail. This yields the desired information, needed in order to characterize the decay rate $\kappa(T)$ for T large.

4.2. Proof for fBm Input

As we have seen in the proof of Lemma 2.1.1, for $T \to \infty$,

$$\gamma_T^{(\mathrm{fBm})}(s,t) = st \cdot H(2H-1)\cdot T^{2H-2} + o(T^{2H-2}).$$

For large T, we obtain in the same way

$$\frac{\partial\gamma_T^{(\mathrm{fBm})}}{\partial s} = t\cdot H(2H-1)\cdot T^{2H-2} + o(T^{2H-2});$$
$$\frac{\partial\gamma_T^{(\mathrm{fBm})}}{\partial t} = s\cdot H(2H-1)\cdot T^{2H-2} + o(T^{2H-2}).$$

Inserting these into (23)–(24) we obtain

$$(2cs - 2H(p+cs)) = \frac{2H(2H-1)(q+ct)(cs-(2H-1)(p+cs))st}{t^{2H}(p+cs)}T^{2H-2} + o(T^{2H-2}); \tag{25}$$

$$(2ct - 2H(q+ct)) = \frac{2H(2H-1)(p+cs)(ct-(2H-1)(q+ct))st}{s^{2H}(q+ct)}T^{2H-2} + o(T^{2H-2}). \tag{26}$$

Note that if we let $T \to \infty$ in the last system, we retrieve (11), which has a unique solution (12). Observe that in the system of equations (25)–(26), the right-hand-side of the equations decays to 0 with speed T^{2H-2} as T grows to infinity. This observation, in conjunction with $(s_T^\star, t_T^\star)$ converging to $(s^\star, t^\star)$, entails that we can express $s_T^\star, t_T^\star$ as follows:

$$\begin{cases} s_T^\star = s^\star + f(s^\star, t^\star)T^{2H-2} + o(T^{2H-2}); \\ t_T^\star = t^\star + g(s^\star, t^\star)T^{2H-2} + o(T^{2H-2}). \end{cases}$$

To determine the values of $f(s^\star, t^\star)$ and $g(s^\star, t^\star)$, we proceed as follows. Using Taylor expansions we obtain for the left-hand-side of (25), after tedious calculus,

$$\begin{aligned}(p+cs)&(2cv(s) - (p+cs)v'(s)) \\ &= 2H(p+cs^\star)s^{\star 2H-2}(cs^\star - (2H-1)(p+cs^\star))f(s^\star, t^\star)T^{2H-2} + o(T^{2H-2}),\end{aligned}$$

and for the right-hand-side

$$\begin{aligned}2&\left(\frac{q+ct}{v(t)}\right)\left((cv(s) - (p+cs)v'(s))\Gamma_T(s,t) + (p+cs)v(s)\frac{\partial \Gamma_T}{\partial s}(s,t)\right) \\ &= 2H(2H-1)(q+ct^\star)s^{\star 2H}t^{\star 1-2H}(cs^\star - (2H-1)(p+cs^\star))T^{2H-2} \\ &\quad + o(T^{2H-2}).\end{aligned}$$

Doing the same for (26), and inserting (12), we find the following expressions for f and g at $s^\star, t^\star$:

$$f(s^\star, t^\star) = (2H-1)\frac{q}{p}s^{\star 2}t^{\star 1-2H} = (2H-1)s^\star t^{\star 2-2H};$$

$$g(s^\star, t^\star) = (2H-1)\frac{p}{q}t^{\star 2}s^{\star 1-2H} = (2H-1)t^\star s^{\star 2-2H}.$$

Inserting these expressions into $\Lambda_T^\star(p+cs, q+ct)$, we can evaluate the components of (17):

$$\begin{aligned}
\frac{(p+cs)^2}{s^{2H}} &= \frac{(p+cs^\star)^2}{s^{\star 2H}}\left(1+2\left(\frac{c}{(p+cs^\star)}-\frac{H}{s^\star}\right)f(s^\star,t^\star)T^{2H-2}+o(T^{2H-2})\right)\\
&= \frac{(p+cs^\star)^2}{s^{\star 2H}}+o(T^{2H-2});\\
\frac{(q+ct)^2}{t^{2H}} &= \frac{(q+ct^\star)^2}{t^{\star 2H}}\left(1+2\left(\frac{c}{(q+ct^\star)}-\frac{H}{t^\star}\right)g(s^\star,t^\star)T^{2H-2}+o(T^{2H-2})\right)\\
&= \frac{(q+ct^\star)^2}{t^{\star 2H}}+o(T^{2H-2});
\end{aligned}$$

$$\begin{aligned}
&2H(2H-1)\cdot(p+cs)(q+ct)(st)^{1-2H}\cdot T^{2H-2}+o(T^{2H-2})\\
&\quad = 2H(2H-1)(p+cs^\star)(q+ct^\star)s^{\star 1-2H}t^{\star 1-2H}T^{2H-2}+o(T^{2H-2})\\
&\quad = 2\frac{(2H-1)c^2}{H}s^{\star 2-2H}t^{\star 2-2H}T^{2H-2}+o(T^{2H-2}).
\end{aligned}$$

We thus obtain the desired result, i.e.,

$$\kappa(T)=\lim_{n\to\infty}\frac{1}{n}\log\kappa_n(T)=\frac{(2H-1)c^2}{H}s^{\star 2-2H}t^{\star 2-2H}T^{2H-2}+o(T^{2H-2}).$$

4.3. Proof for iOU Input

As in the fBm case, denote by $s^\star, t^\star$ the minimizing point when there is independence, i.e., the solution of (11). We follow the same arguments as in the case of fBm. For fixed (s,t) the covariance $\Gamma_T(s,t)$ is decreasing exponentially in T. The solution of system (23)–(24), say $s_T^\star, t_T^\star$, converges to $s^\star, t^\star$, and its convergence speed is of the order e^{-T} for large T. These observations entail that

$$\begin{cases} s_T^\star = s^\star + k(s^\star,t^\star)e^{-T}+o(e^{-T});\\ t_T^\star = t^\star+\ell(s^\star,t^\star)e^{-T}+o(e^{-T}).\end{cases}$$

To determine k and ℓ at $s^\star, t^\star$, we proceed as in the above subsection. We find

$$\begin{aligned}
&k(s^\star,t^\star)\\
&\quad=\frac{q+ct^\star}{p+cs^\star}(e^{t^\star}-1)\frac{cv(s^\star)(1-e^{-s^\star})-(p+cs^\star)(v'(s^\star)(1-e^{-s^\star})-v(s^\star)e^{-s^\star})}{(cv'(s^\star)-(p+cs^\star)v''(s^\star))v(t^\star)}\\
&\quad=\frac{(q+ct^\star)v'(t^\star)}{v''(t^\star)(p+cs^\star)v(t^\star)}\cdot\frac{cv(s^\star)v'(s^\star)-(p+cs^\star)v'(s^\star)^2+(p+cs^\star)v(s^\star)v''(s^\star)}{(cv'(s^\star)-(p+cs^\star)v''(s^\star))}
\end{aligned}$$

$$= \frac{2cv(s^\star)}{v''(t^\star)(p+cs^\star)} \cdot \frac{(-cv'(s^\star)+(p+cs^\star)v''(s^\star))}{(cv'(s^\star)-(p+cs^\star)v''(s^\star))} = -\frac{v'(s^\star)}{v''(t^\star)};$$

$$\begin{aligned}
&\ell(s^\star, t^\star)\\
&= \frac{p+cs^\star}{q+ct^\star}(1-e^{-s^\star})\frac{cv(t^\star)(e^{t^\star}-1)-(q+ct^\star)(v'(t^\star)(e^{t^\star}-1)-v(t^\star)e^{t^\star})}{(cv'(t^\star)-(q+ct^\star)v''(t^\star))v(s^\star)}\\
&= \frac{(p+cs^\star)v'(s^\star)}{(q+ct^\star)v(s^\star)} \cdot \frac{cv(t^\star)v'(t^\star)e^{t^\star}-(q+ct^\star)v'(t^\star)^2e^{t^\star}+(q+ct^\star)v(t^\star)e^{t^\star}}{(cv'(t^\star)-(q+ct^\star)v''(t^\star))}\\
&= \frac{2cv(s^\star)}{v''(t^\star)(q+ct^\star)} \cdot \frac{(-cv'(t^\star)+(q+ct^\star))}{(cv'(t^\star)-(p+ct^\star)v''(t^\star))}\\
&= \frac{v'(t^\star)}{v''(t^\star)} \cdot \frac{q+cv(t^\star)}{(cv'(t^\star)-(q+ct^\star)v''(t^\star))}.
\end{aligned}$$

Now we insert this in the objective function (17), and similarly to the fBm case we obtain

$$\begin{aligned}
&\frac{(p+cs)^2}{v(s)} = \frac{(p+cs^\star)^2}{v(s^\star)}(1+o(e^{-T}));\\
&\frac{(q+ct)^2}{v(t)} = \frac{(q+ct^\star)^2}{v(t^\star)}(1+o(e^{-T}));\\
&\frac{(p+cs)(q+ct)(1-e^{-s})(e^t-1)e^{-T}}{v(s)v(t)}\\
&\quad = \frac{(p+cs^\star)(q+ct^\star)(1-e^{-s^\star})(e^{t^\star}-1)e^{-T}}{v(s^\star)v(t^\star)} + o(e^{-T}).
\end{aligned}$$

Thus we get for iOU input the desired result:

$$\begin{aligned}
\kappa(T) &= \lim_{n\to\infty}\frac{1}{n}\log\kappa_n(T)\\
&= \frac{(p+cs^\star)(q+ct^\star)}{v(s^\star)v(t^\star)} \cdot \frac{1}{2}(1-e^{-s^\star})(e^{t^\star}-1)e^{-T} + o(e^{-T}),
\end{aligned}$$

which simplifies to $2c^2e^{-(T-t^\star)}$.

5. DISCUSSION AND CONCLUDING REMARKS

5.1. Generalizations

Theorems 3.1 and 3.2 suggest that our results can be generalized considerably, in that it can be expected that their counterparts can be stated for a substantially broader class of Gaussian processes with stationary increments. First observe that expression (17) can alternatively be written as

$$\frac{1}{2}\left(\frac{(p+cs)^2}{v(s)}+\frac{(q+ct)^2}{v(t)}-2\frac{(p+cs)(q+ct)\Gamma_T(s,t)}{v(s)v(t)}\right)+o(\Gamma_T(s,t)). \quad (27)$$

Now suppose we wish to evaluate $\inf_t(f(t)+\varepsilon g(t))-f(t^\star)$, where $t^\star$ is minimizer of $f(\cdot)$. A Taylor expansion of $f'(t)+\varepsilon g'(t)$ in $t^\star_\varepsilon = t^\star+\varepsilon\bar{t}$ reads

$$\begin{aligned} &f'(t^\star)+\varepsilon\bar{t}f''(t^\star)+\varepsilon g'(t^\star)+O(\varepsilon^2)\\ &\quad = f'(t^\star)+\varepsilon(\bar{t}f''(t^\star)+g'(t^\star))+O(\varepsilon^2), \end{aligned}$$

so that we obtain $\bar{t} = -g'(t^\star)/f''(t^\star)$. Hence, under appropriate regularity conditions,

$$\inf_t(f(t)+\varepsilon g(t)) = f(t^\star)-\varepsilon\frac{g'(t^\star)}{f''(t^\star)}f'(t^\star)+\varepsilon g(t^\star)+O(\varepsilon^2).$$

Now using $f'(t^\star)=0$ it follows that

$$\inf_t(f(t)+\varepsilon g(t))-f(t^\star) = \varepsilon g(t^\star)+O(\varepsilon^2). \quad (28)$$

In the same way, a 2-dimensional counterpart of (28) can be stated. Now suppose that (for large T) $\Gamma_T(s,t)$ decouples as $\varphi(s,t)\cdot\varepsilon(T)$; here $\varepsilon(T)$ does not depend on s and t, and converges to 0 as $T\to\infty$. Applying then the two-dimensional version of (28) to (27),

$$\begin{aligned} \kappa(T) &= \lim_{n\to\infty}\frac{1}{n}\log\kappa_n(T) = \frac{(p+cs^\star)(q+ct^\star)}{v(s^\star)v(t^\star)}\varphi(s^\star,t^\star)\varepsilon(T)+o(\varepsilon(T))\\ &= 4c^2\frac{\varphi(s^\star,t^\star)}{v'(s^\star)v'(t^\star)}\varepsilon(T)+o(\varepsilon(T)). \end{aligned}$$

5.2. Long-Range Dependence

Based on Theorems 3.1 and 3.2, one may conjecture that long-range dependence of the input process carries over to workload process. We now provide additional heuristic support for this claim.

First consider fBm. Heuristically reasoning, Theorem 3.1 entails that, for some constant κ_0,

$$\kappa_n(T)\approx\exp(n\kappa_0T^{2H-2}),$$

and hence we have that the correlation coefficient of the indicator functions $1\{Q_0^n>np\}$ and $1\{Q_T^n>nq\}$ roughly equals (using that

$x(1-x) \approx x$ for x small)

$$\frac{\mathbb{P}\big(Q_0^n > np, Q_T^n > nq\big) - \mathbb{P}\big(Q_0^n > np\big)\mathbb{P}\big(Q_T^n > nq\big)}{\sqrt{\mathbb{P}\big(Q_0^n > np\big)\mathbb{P}\big(Q_T^n > nq\big)}}$$
$$\approx (e^{n\kappa_0 T^{2H-2}} - 1) \cdot \sqrt{\mathbb{P}\big(Q_0^n > np\big)\mathbb{P}\big(Q_0^n > nq\big)}.$$

Using $e^x \approx 1 + x$ for x small, we find that for T large, the above display is of the form $\psi(n)T^{2H-2}$, where the function $\psi(\cdot)$ does not depend on T. Observe that the latter expression is nonsummable (over T) for $H > \frac{1}{2}$. This intuitive argument suggests that the long-range dependence of the input process propagates to the queueing process.

Likewise, for iOU we find that the correlation coefficient previously introduced is roughly proportional to e^{-T}, and hence corresponds to a short-range dependent process.

Further research on this issue could make use of the concept of Hurstiness, as introduced in Ref.[29]. Hurstiness is a property of the queue's input process (closely related to long-range dependence), and it is shown that the Hurstiness is preserved by several fundamental operators; for instance, the Hurstiness of the departure process equals that of the arrival process. It is not immediately clear, however, whether results as those presented in the present article, can be found relying on the notion of Hurstiness. As there is a clear relation between the departure process and the workload dynamics, one would think so, but a technical issue is that Hurstiness relates to cumulative processes, such as arrival and departure processes, whereas our focus is on the dependence between "instantaneous values" of the workload at time 0 and T. Also, Hurstiness relates to just the rate of decay, and in view of this it is not likely that it would help to (for instance) find the constant in front of T^{2H-2} in Theorem 3.1.

5.3. Remarks on Asymptotics for iOU

It may be surprising, at first glance, that the asymptotics of $\kappa(T)$ for iOU, that is $2c^2e^{-(T-t^\star)}$, depend on q, but do not depend on p. This can be understood as follows.

First observe that for iOU input (unlike for fBm input) there is a notion of a traffic rate process $X(\cdot)$, where $X(t) = A'(t)$. It can be checked easily that (i) $X(t)$ is normally distributed with mean 0 and variance $\frac{1}{2}$, (ii) $\mathbb{C}\mathrm{ov}(X(0), X(T)) = \frac{1}{2}e^{-T}$, (iii) the conditional distribution of $A(T-t, T)$ given $X(0) = x$ is normal with mean and variance, respectively,

$$\mu_T(t \mid x) = \mathbb{E}(A(T-t, T) \mid X(0) = x) = x(e^t - 1)e^{-T},$$
$$v_T(t \mid x) = \mathbb{V}\mathrm{ar}(A(T-t, T) \mid X(0) = x) = v(t) - e^{-2T}(e^t - 1)^2,$$

as follows from standard formula for conditional normal distributions (cf. Section 4.3 in Ref.[18]).

Also, rewrite $\kappa_n(t)$ as the ratio of $\mathbb{P}(Q_T^n > nq \mid Q_0^n > np)$ and $\mathbb{P}(Q_0^n > np)$. The decay rate of the latter probability is given by (8). Now focus on the decay rate of the former (i.e., conditional) probability. Realize that, as the condition $Q_0^n > np$ is binding, the most likely path (in the "Schilder sense") must be such that the traffic rate at time 0 is c (which means that the aggregate input process is generating traffic at a rate nc); otherwise, the queue grows even beyond np. Also notice that the most likely path is such that the buffer has been empty between 0 and T. These observations, in conjunction with the Markovian nature of the rate process of iOU, entail that all the information about the system at time 0 which has an impact on the system at time T, is contained in the fact that the rate is (most likely) nc at time 0. To find the decay rate of $\mathbb{P}(Q_T^n > nq \mid Q_0^n > np)$, we therefore have to solve

$$\inf_{t>0} \frac{(q + ct - \mu_T(t \mid c))^2}{2v_T(t \mid c)}.$$

The above formula for the conditional mean and variance entail that this optimization problem reduces to

$$\inf_{t>0} \left(\frac{(q+ct)^2}{2v(t)} - \frac{(q+ct)c(e^t - 1)e^{-T}}{v(t)} + o(e^{-T}) \right).$$

Applying Equation (28) once again, inserting (11), and using that $v'(t) = 1 - e^{-t} = e^{-t}(e^t - 1)$, we indeed obtain that $\kappa(T)$ equals $2c^2 e^{-(T-t^\star)} + o(e^{-T})$, as expected. The above reasoning explains why the decay rate does not depend on p; as an aside we mention that also $\ell(s^\star, t^\star)$ does not depend on p.

5.4. Further Research

In this article we have focused on the metric $\kappa(T)$ that relates to the many-sources scaling, and that was intended to express the level of correlation between the workloads at time 0 and T. Then we studied the asymptotics of $\kappa(T)$ for large T. Evidently, many other measures for correlation can be thought of. One could, for instance, consider similar measures, but then in the large-buffer regime.

In this respect, we could consider a queue fed by a single Gaussian input, emptied at a constant rate $C > 0$. Then an interesting measure could be, for fixed p, q, T,

$$\bar{\kappa}_B := \frac{\mathbb{P}(Q_0 > pB, Q_{TB} > qB)}{\mathbb{P}(Q_0 > pB)\mathbb{P}(Q_{TB} > qB)} = \frac{\mathbb{P}(Q_0 > pB, Q_{TB} > qB)}{\mathbb{P}(Q_0 > pB)\mathbb{P}(Q_0 > qB)},$$

and its asymptotics for large B. The analysis of $\bar{\kappa}_B$ is radically different from that of $\kappa_n(T)$; the reason for this is that in the many-sources regime, the most likely time-scales to overflow are more or less constant in the scaling parameter (i.e., n), whereas in the large-buffer, one would expect that these time-scales are roughly proportional to the scaling parameter (i.e., B).

In this case we expect, when analyzing $\mathbb{P}(Q_0 > pB, Q_{TB} > qB)$, different regimes. More precisely: for B large it is not always true that, in the most likely scenario, both constraints are tightly met; for some values of p, q, T this will be the case, while for others just one constraint will be tightly met (and the other event "comes for free"). In case both constraints are tightly met, again two cases can be distinguished: a first in which the queue has not become empty between 0 and TB (which we expect is the case for T smaller than some critical timescale $T^\star$), and a second in which epochs 0 and TB lie in different busy periods (for T larger than $T^\star$), cf., Ref.[27] (Section 11.2).

APPENDIX

In this appendix, we prove a lemma that is needed to establish Proposition 4.1.1. We first determine the closure of the set $\mathscr{S}_T$. We define

$$\mathscr{S}(s) := \{f \in \Omega : -f(-s) > p + cs\};$$
$$\mathscr{S}_T(t) := \{f \in \Omega : f(T) - f(T-t) > q + ct\};$$

also

$$\overline{\mathscr{S}(s)} := \{f \in \Omega : -f(-s) \geq p + cs\};$$
$$\overline{\mathscr{S}_T(t)} := \{f \in \Omega : f(T) - f(T-t) \geq q + ct\}.$$

Notice that evidently

$$\mathscr{S}_T = \bigcup_{s,t>0} (\mathscr{S}(s) \cap \mathscr{S}_T(t)).$$

Lemma A.1. *For any T, we have that the closure $\overline{\mathscr{S}_T}$ of $\mathscr{S}_T$ is given by*

$$\bigcup_{s,t>0} \left(\overline{\mathscr{S}(s)} \cap \overline{\mathscr{S}_T(t)}\right).$$

Proof. The proof is similar to those in Refs.[20,24]. We prove both inclusions separately.

- We show first the inclusion "$\subseteq$". For any $f \in \overline{\mathscr{S}_T}$ there exists a sequence $f_n \in \mathscr{S}_T$ such that $\|f_n - f\|_\Omega \to 0$ as $n \to \infty$. Now since $f_n \in \mathscr{S}_T$ there is an $s_n > 0$ and a $t_n > 0$ such that $f_n \in \mathscr{S}(s_n) \cap \mathscr{S}_T(t_n)$, so that we have

$$-f_n(-s_n) > p + cs_n \quad \text{and} \quad f_n(T) - f_n(T - t_n) > q + ct_n.$$

The sequence s_n is bounded, because if not we would have a subsequence satisfying

$$0 = \lim_{n\to\infty} \|f - f_n\|_\Omega \geq \lim_{n\to\infty} \frac{f(-s_n) - f_n(-s_n)}{1 + s_n} \geq \lim_{n\to\infty} \left(\frac{f(-s_n)}{1 + s_n} + \frac{p + cs_n}{1 + s_n} \right) = c,$$

(use that $f \in \Omega$!), which gives a contradiction (recall that $c > 0$). Along the same lines it can be shown that t_n is bounded. Hence there are subsequences $s_{n_k} \to s_0$ and $t_{n_k} \to t_0$, for finite s_0 and t_0. We conclude that for large enough k

$$-f_{n_k}(-s_0) \geq p + cs_0 \quad \text{and} \quad f_{n_k}(T) - f_{n_k}(T - t_0) \geq q + ct_0.$$

We conclude that

$$f \in \left(\overline{\mathscr{S}(s_0) \cap S_T(t_0)}\right) \subseteq \left(\overline{\mathscr{S}(s_0)} \cap \overline{\mathscr{S}_T(t_0)}\right).$$

- For the other inclusion, "$\supseteq$", let

$$f \in \bigcup_{s,t>0} \left(\overline{\mathscr{S}(s)} \cap \overline{\mathscr{S}_T(t)}\right).$$

Then there exist $s_0, t_0 > 0$ such that $f \in \overline{\mathscr{S}(s_0)} \cap \overline{\mathscr{S}_T(t_0)}$. Let $\eta(\cdot)$ be a path in R that is strictly increasing and taking negative values for $u \in (-\infty, 0)$ and positive values for $u \in (0, \infty)$ (for instance $\eta(u) := \mathrm{sgn}(u)\sqrt{|u|}$ or $\arctan u$). Define

$$f_n(u) := f(u) + \frac{\eta(u)}{n}.$$

Then $f_n \in \mathscr{S}(s_0) \cap \mathscr{S}_T(t_0)$ as

$$-f_n(-s_0) = -f(-s_0) - \frac{\eta(-s_0)}{n} \geq p + cs_0 - \frac{\eta(-s_0)}{n} > p + cs_0$$

and

$$\begin{aligned} f_n(T) - f_n(T - t_0) &= f(T) - f(T - t_0) + \frac{\eta(T) - \eta(T - t_0)}{n} \\ &\geq q + ct_0 + \frac{\eta(T) - \eta(T - t_0)}{n} > q + ct_0. \end{aligned}$$

Moreover, we have, for $n \to \infty$, that $\|f_n - f\|_\Omega \to 0$ (use that $\eta \in R \subset \Omega$), and hence

$$f \in \left(\overline{\mathcal{S}(s_0) \cap S_T(t_0)}\right) \subseteq \left(\overline{\mathcal{S}(s_0)} \cap \overline{\mathcal{S}_T(t_0)}\right) \subseteq \overline{\mathcal{S}_T}.$$

This proves the second inclusion. $\square$

ACKNOWLEDGMENTS

The second author wants to thank Peter Glynn (Stanford, CT) and Ilkka Norros (VTT, Finland) for stimulating discussions. The authors also thank two anonymous referees for their careful assessment of the article.

REFERENCES

1. Addie, R.; Mannersalo, P.; Norros, I. Most probable paths and performance formulae for buffers with Gaussian input traffic. European Trans. Telecommun. **2002**, *13*, 183–196.
2. Adler, R. An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes. Lecture Notes-Monograph Series, Vol. 12. Institute of Mathematical Statistics, Hayward, CA, 1990.
3. Bahadur, R.; Zabell, S. Large deviations of the sample mean in general vector spaces. Ann. Prob. **1979**, *7*, 587–621.
4. Botvich, D.; Duffield, N. Large deviations, the shape of the loss curve, and econimies of large scale multiplexers. Queueing Syst. **1995**, *20*, 293–320.
5. Bradley, R. Basic properties of strong mixing conditions – a survey and some open questions. Prob. Surveys **2005**, *2*, 107–144.
6. Dębicki, K. A note on LDP for supremum of Gaussian processes over infinite horizon. Stat. Prob. Lett. **1999**, *44*, 211–220.
7. Dębicki, K.; Mandjes, M. Exact overflow asymptotics for queues with many Gaussian inputs. J. Appl. Prob. **2003**, *40*, 702–720.
8. Dembo, A.; Zeitouni, O. *Large Deviations Techniques and Applications*; 2nd ed.; Springer: New York, 1998.
9. Deuschel, J.-D., Stroock, D. *Large Deviations*; Academic Press: Boston, MA, 1989.
10. Dieker, T. Extremes of Gaussian processes over an infinite horizon. Stoch. Proc. Appl. **2005**, *115*, 207–248.
11. Duffield, N.; O'Connell, N. Large deviations and overflow probabilities for general single-server queue, with applications. Math. Proc. Camb. Phil. Soc. **1995**, *118*, 363–374.
12. Gaigalas, R.; Kaj, I. Convergence of scaled renewal processes and a packet arrival model. Bernoulli **2003**, *9*, 671–703.
13. Harrison, J. *Brownian Motion and Stochastic Flow Systems;* Wiley: New York, 1985.
14. Hernandez-Campos, F.; Jeffay, K.; Park, C.; Marron, J.; Resnick, S. Extremal dependence: Internet traffic applications. Stoch. Mod. **2005**, *21*, 1–35.
15. Kilpi, J.; Norros, I. Testing the Gaussian approximation of aggregate traffic. In *Proc. 2nd Internet Measurement Workshop*, 2002, pp. 49–61.
16. Leland, W.; Taqqu, M.; Willinger, W.; Wilson, D. On the self-similar nature of Ethernet traffic (extended version). IEEE/ACM Trans. Network **1994**, *2*, 1–15.
17. Mandjes, M. *Large Deviations for Gaussian Queues*; Wiley: Chichester, UK, 2007.
18. Mandjes, M.; Mannersalo, P.; Norros, I.; van Uitert, M. Large deviations of infinite intersections of events in Gaussian processes. Stoch. Proc. Appl. **2006**, *116*, 1269–1293.
19. Mandjes, M.; van de Meent, R. Resource provisioning through buffer sampling. Submitted 2008.
20. Mandjes, M.; van Uitert, M. Sample-path large deviations for tandem and priority queues with Gaussian inputs. Ann. Appl. Prob. **2005**, *15*, 1193–1226.

21. Mikosch, T.; Resnick, S.; Rootzén, H.; Stegeman, A. Is network traffic approximated by stable Lévy motion or fractional Brownian motion? Ann. Appl. Prob. **2002**, *12*, 23–68.
22. Narayan, O. Exact asymptotic queue length distribution for fractional Brownian traffic. Adv. Perf. Anal. **1998**, *1*, 39–63.
23. Norros, I. A storage model with self-similar input. Queueing Syst. **1994**, *16*, 387–396.
24. Norros, I. Busy periods of fractional Brownian storage: a large deviations approach. Adv. Perf. Anal. **1999**, *2*, 1–20.
25. Pakes, A. On the tails of waiting-time distributions. J. Appl. Prob. **1975**, *12*, 555–564.
26. Samorodnitsky, G. Long memory and self-similar processes. Ann. de la Faculté des Sciences de Toulouse Série 6, **2006**, *15*, 107–123.
27. Shwartz, A.; Weiss, A. *Large Deviations for Performance Analysis. Queues, Communications, and Computing*; Chapman & Hall: London, UK, 1995.
28. Weiss, A. A new technique for analyzing large traffic systems. Adv. Appl. Prob. **1986**, *18*, 506–532.
29. Wischik, D.; Ganesh, A. The calculus of Hurstiness. Queueing Syst. **2006**.
30. Zwart, B.; Borst, S.; Mandjes, M. Exact asymptotics for fluid queues fed by multiple heavytailed on-off sources. Ann. Appl. Prob. **2004**, *14*, 903–957.