

Simple bounds for a transient queue

Takayuki Osogami

IBM Research - Tokyo

1623-14 Shimotsuruma, Yamato-shi 242-8502, Japan

Email: osogami@jp.ibm.com

Rudy Raymond

IBM Research - Tokyo

1623-14 Shimotsuruma, Yamato-shi 242-8502, Japan

Email: raymond@jp.ibm.com

Abstract—Bounds on performance of a queueing model can provide useful information to guarantee quality of service for communication networks. We study the bounds on the mean delay in a transient GI/GI/1 queue given the first two moments of the service time and the inter-arrival time, respectively. We establish a simple upper-bound, which then is used to show that the true transient mean-delay is at most four times larger than an asymptotic diffusion-approximation. We also prove that the tight lower-bound is zero as long as the service time and the inter-arrival time have finite variance and the load is below one. Tightness of the trivial lower-bound is in contrast to the stationary mean-delay, which has strictly positive lower-bound when the service time is sufficiently variable. We also show how our results can be applied to analyze the transient mean delay of packets in the real-world Internet.

Index Terms—Queue, GI/GI/1, bounds, moments.

I. INTRODUCTION

Bounds on performance metrics of queueing models have been studied extensively in the literature. The study is motivated by both practical requirements and theoretical interests. From a practical perspective, the bounds can be used to guarantee quality of service for communication networks. From a theoretical perspective, bounds on a performance metric can give an understanding of which parameters of a queueing model affect the performance metric and how. In this paper, we investigate the bounds on the mean delay of a transient GI/GI/1 queue given the first two moments of the service time, S , and the inter-arrival time, A , respectively.

Although the GI/GI/1 queue in steady state has been investigated extensively in the literature (e.g., see [16], [22]), its transient behavior is not fully understood. We believe that the study of the transient behavior is important particularly when the load is high. Communication networks often experience high load, but the high load often does not continue for a long period. Notice that, if the period of high load is long, one should add more resources to alleviate the load. Because it takes a long time for a queue at high load to approach its steady state, the queue is usually not close to the steady state during the short period of high load.

Our focus is to bound the mean delay when only the first two moments of S and A , respectively, are given. We study mean delay, because mean is one of the most fundamental statistics and allows an intuitive understanding. Bounds on mean delay are also complementary to those on other statistics, such as the tail distribution, of delay. Although bounds in terms of the first two moments might not necessarily be the best bounds,

they can complement the bounds in terms of other statistics of S and A . What we can tell about the mean delay when only the first two moments of S and A are given is also of theoretical interest.

A. Contributions

Our primary contribution is simple lower and upper bounds on the mean delay in a transient GI/GI/1 queue given the first two moments of S and A , respectively. Specifically, let W_n be the delay of the n -th job arriving at a GI/GI/1 queue under the condition that the 0-th job arrives when the queue is empty (i.e., $W_0 = 0$). Throughout, we assume that jobs are processed with the first-come-first-served policy.

We prove that

$$\mathbb{E}[W_n] \leq \frac{\pi \sqrt{n(C_A^2 + \rho^2 C_S^2)}}{2\lambda},$$

where λ is the arrival rate, ρ is the load, and C_A (respectively, C_S) is the coefficient of variation of A (respectively, S). This upper bound holds for any $n \geq 0$, but we also establish a tighter upper-bound for a large n (specifically, $n \geq \frac{C_A^2 + \rho^2 C_S^2}{16(1-\rho)^2}$).

We also prove that the tight lower-bound is 0 for any $C_A, C_S < \infty$ and $\rho < 1$. The tightness of this trivial bound is nontrivial and might be counter-intuitive, because a strictly positive lower-bound is known for the mean delay, $\mathbb{E}[W]$, in a stationary GI/GI/1 queue when S is sufficiently variable (see Equation (1) in Section II). However, our result does not contradict with this existing result, because $\mathbb{E}[W_n] < \mathbb{E}[W]$ for any finite n .

To provide an insight on how far our upper-bound is from the tight upper-bound, we compare it against a diffusion approximation of $\mathbb{E}[W_n]$ that we constructed with a standard approach. We show that the diffusion approximation is asymptotically equivalent to $\sqrt{n(C_A^2 + C_S^2)}/(\lambda\sqrt{2\pi})$ in the heavy-traffic limit of $\rho \rightarrow 1$. Hence, our upper bound is at most four times larger than the diffusion approximation in the heavy-traffic limit. We also show that the diffusion approximation well approximates $\mathbb{E}[W_n]$ for some S and A with the given first two moments although it does not provide an upper bound. This suggests that our upper bound is within a constant factor from the tight upper-bound on $\mathbb{E}[W_n]$. Our experimental results suggest that the constant factor is at most two. We will also discuss properties of a transient GI/GI/1 queue, which hint that obtaining the tight upper-bound on $\mathbb{E}[W_n]$ given the first two moments of S and A is not easy.

Finally, we show how our upper bounds can be used to predict the transient mean delay of packets in the real-world Internet traffic, where there are dependencies in inter-arrival times and in service times. For this purpose, we use the traffic datasets made public at the MAWI Working Group Traffic Archive¹. By substituting variance parameters, which capture auto-correlations, for variances in the expression of our upper bounds, we show that our upper bound is useful in predicting the transient mean delay of the real-world packets even when inter-arrival times and service times have correlations.

B. Related work

The upper bound proposed in this paper is constructed using the upper bound that we have proposed in [18]. The upper bound in [18] consists of several terms, which involve an arctangent and logarithms, and provides little intuition because of the complexity. Although bounds on $E[W_n]$ given the first two moments of S and A have only been studied in our prior work [18], [17], there is a large body of literature that studies the transient and stationary GI/GI/1 queue.

Bounds on delay and queue length in a transient GI/GI/1 queue have been proposed in various forms. We will see that our bounds are complementary to the existing ones, because the existing bounds and ours use different statistics of S and A or are on different statistics of delay. Wang [19] proposes bounds on $E[W_n]$ in terms of $E[W]$ and the distribution of the number of jobs served in a busy period. Gong and Hu [12] propose bounds on the moments of W_n , including $E[W_n]$, in terms of the moments of S and the derivatives of the density function of A evaluated at zero. Limón-Robles and Wortman [15] propose a computational approach for calculating bounds on $\Pr(Q_t \leq k)$, the distribution of the queue length, Q_t , at time t . The bounds depend on the distribution of the index of the first job that departs after time t and the distribution of the arrival time of the n -th job for each n . In [18], we derive an upper bound on $\Pr(W_n > y)$ in closed form given the first two moments of S and A , respectively, which is also refined using higher moments and evaluated numerically. Traditional exponential bounds on $\Pr(W_n > y)$ can be obtained with martingales and Wald's identity (see Section 7.5 from [10]) or with large deviations (see Section 14.2 from [15]), as is discussed in Section 4 from [17]. Such exponential bounds depend on the generating functions of S and A .

There is also a large body of literature on approximation and numerical analysis of the transient GI/GI/1 queue. For example, Abate and Whitt [2] present a diffusion approximation of the queue length as a function of time when the queue starts empty and suggest a way to extend it to a GI/GI/1 queue, which is essentially equivalent to the diffusion approximation summarized in Section 6.5 from [8]. The diffusion approximation is further studied by Wang [19]. There

¹The datasets are collected daily from the WIDE backbone network, which is a Japanese academic network connecting universities and research institutes (see [6] for the details of the datasets). The datasets are available from <http://tracer.csl.sony.co.jp/mawi>.

are many approaches for analyzing a transient GI/GI/1 queue numerically [7], [15] and with simulation [3].

Bounds on $E[W]$ (stationary mean delay) in terms of the moments of S and A have also been studied extensively in the literature. The upper bounds on $E[W]$ proposed by Kingman [13] and Daley [10] are well known. Their bounds depend on the first two moments of S and A . Notice that an upper bound on $E[W]$ is also an upper bound on $E[W_n]$, because W stochastically dominates W_n for any n . Recently, Bertsimas and Natarajan [5] propose a computational approach for calculating the bounds on the moment, $E[W^k]$, for a given k , using the moments of S and A .

There was a conjecture that, given first two moments of S and A , tight bounds on $E[W]$ are achieved with extremal distributions that have only mass probabilities at two particular points [4], [11], [20]. This conjecture is shown to hold for a limited class of the GI/GI/1 queue [9], [11], [20], but counterexamples to the conjecture have been found [20].

C. Organization

In Section II, we prove that the trivial lower-bound on $E[W_n]$ is tight and present the simple upper-bound on $E[W_n]$. In Section III, the simple upper-bound is compared against an asymptotic diffusion-approximation of $E[W_n]$ analytically. In Section IV, we further evaluate the simple upper-bound numerically and with trace-driven simulations, where we show how the upper-bound can be applied when inter-arrival times and service times have correlations. In Section V, we analyze properties of the transient GI/GI/1 queue. In particular, we show that the extremal distributions of S and A do not maximize $E[W_n]$ given the first two moments of S and A .

II. SIMPLE BOUNDS ON TRANSIENT MEAN-DELAY

Consider a GI/GI/1 queue, where A denotes the inter-arrival time and S the service time. Let $\lambda \equiv 1/E[A]$ be the arrival rate, $\mu \equiv 1/E[S]$ be the service rate, $\rho \equiv E[S]/E[A]$ be the load, $C_A \equiv \sqrt{\text{Var}[A]}/E[A]$ be the coefficient of variation for A , and $C_S \equiv \sqrt{\text{Var}[S]}/E[S]$. Let W_n be the delay of the n -th job for $n \geq 1$, given that the 0-th job arrives at an empty queue (i.e., $W_0 = 0$). In this section, we present simple lower-bound and upper-bound on $E[W_n]$.

A. Summary of results

In Section II-B, we prove that the trivial lower-bound of 0 is tight for any finite n . The tightness is nontrivial, because a nonzero lower-bound is known for the corresponding mean-delay, $E[W]$, in steady state as long as $C_S^2 > (1 - \rho)/\rho$:

$$E[W] \geq \frac{\rho^2 C_S^2 - \rho(1 - \rho)}{2\lambda(1 - \rho)} \quad (1)$$

(see (6.4.4) from [16]).

Interestingly, the tight lower-bound on $E[W_n]$ is achieved with extremal distributions that have mass probabilities at only two particular points. Specifically, consider the parametrized

family of random variables:

$$\begin{aligned} A(\varepsilon) &\equiv \begin{cases} \frac{1}{\lambda}(1-\varepsilon) & \text{w.p. } \frac{C_A^2}{C_A^2+\varepsilon^2} \\ \frac{1}{\lambda}\left(1+\frac{C_A^2}{\varepsilon}\right) & \text{w.p. } \frac{\varepsilon^2}{C_A^2+\varepsilon^2} \end{cases} \\ S(\varepsilon) &\equiv \begin{cases} \frac{1}{\mu}(1-\varepsilon) & \text{w.p. } \frac{C_S^2}{C_S^2+\varepsilon^2} \\ \frac{1}{\mu}\left(1+\frac{C_S^2}{\varepsilon}\right) & \text{w.p. } \frac{\varepsilon^2}{C_S^2+\varepsilon^2}, \end{cases} \end{aligned} \quad (2)$$

where w.p. denotes ‘‘with probability.’’ For any $\varepsilon \in (0, 1]$, observe that $\mathbf{E}[A(\varepsilon)] = 1/\lambda$, $\sqrt{\text{Var}[A(\varepsilon)]}/\mathbf{E}[A(\varepsilon)] = C_A$, and the support of $A(\varepsilon)$ consists of two points. Analogous properties hold for $S(\varepsilon)$. For any $n, C_A, C_S < \infty$ and $\rho < 1$, we will show that $\mathbf{E}[W_n]$ can be made arbitrarily close to 0 by a pair of extremal distributions, $A = A(\varepsilon)$ and $S = S(\varepsilon)$ with $\varepsilon \rightarrow 0$. Notice that our result does not contradict with the existing result for steady state, because we assume that the queue is empty at time 0, so that the queue does not reach the steady state in finite n .

In Section II-C, we establish a simple upper-bound on $\mathbf{E}[W_n]$. Specifically, we find:

Theorem 1: For $n < \frac{C_A^2 + \rho^2 C_S^2}{16(1-\rho)^2}$, we have

$$\mathbf{E}[W_n] \leq \frac{\pi}{2\lambda} \sqrt{n(C_A^2 + \rho^2 C_S^2)}. \quad (3)$$

For $\frac{C_A^2 + \rho^2 C_S^2}{16(1-\rho)^2} \leq n \leq \frac{C_A^2 + \rho^2 C_S^2}{\varepsilon^2(1-\rho)^2}$, we have

$$\mathbf{E}[W_n] \leq \frac{e}{2\lambda} \sqrt{n(C_A^2 + \rho^2 C_S^2)}. \quad (4)$$

For succinctness, in the hereafter let us use

$$\kappa \equiv \frac{C_A^2 + \rho^2 C_S^2}{16(1-\rho)^2}.$$

At $n = 16\kappa/\varepsilon^2$, the upper bound given by (4) is equivalent to Kingman’s upper-bound [13] on $\mathbf{E}[W]$. For $n \geq 16\kappa/\varepsilon^2$, Kingman’s upper-bound can be used as an upper bound on $\mathbf{E}[W_n]$. Let u_n^{new} denote the upper bound given in Theorem 1 for $n \leq 16\kappa/\varepsilon^2$ and Kingman’s upper-bound for $n > 16\kappa/\varepsilon^2$.

Theorem 1 will be proved using more complex expressions of the upper bound on $\mathbf{E}[W_n]$, which we have established in Corollary 1 from [18] by formulating a semidefinite optimization and constructing its optimal solution in closed form. For completeness, we excerpt the relevant part of Corollary 1 from [18] in the following and provide an outline of its proof in Appendix B:

Corollary 1: [18] ... Let $\gamma \equiv (C_S^2 \rho^2 + C_A^2)/(1-\rho)^2$ If $\gamma/16 \leq n \leq \gamma/4$, then

$$\mathbf{E}[W_n] \leq \frac{\xi}{4} + \frac{1-\rho}{\lambda} \ln \left(\frac{e^2 \gamma}{4n} \right) n + \frac{2(1-\rho)n^2}{3\lambda\gamma}, \quad (5)$$

where $\xi \equiv (C_A^2 + C_S^2 \rho^2)/(2\lambda(1-\rho))$ is Kingman’s upper bound on $\mathbf{E}[W]$. Let U be the right-hand side of (5). If $n \leq \gamma/16$, then

$$\begin{aligned} \mathbf{E}[W_n] &\leq U - \frac{(1-\rho)\sqrt{\gamma n}}{\lambda} \left(\frac{1}{2} \sqrt{\phi_n} - \arctan \sqrt{\phi_n} \right) \\ &\quad - \frac{2n(1-\rho)}{\lambda} \ln \left(\sqrt{\phi_n} + \sqrt{\phi_n + 1} \right), \end{aligned} \quad (6)$$

where $\phi_n = \gamma/(16n) - 1$.

Let u_n^{org} denote the smaller of Kingman’s upper-bound and the upper-bound established in Corollary 1, where ‘‘org’’ stands for ‘‘original.’’ We will see that u_n^{new} is larger at most by a factor of 1.16 than u_n^{org} .

Observe that u_n^{new} is significantly simpler than u_n^{org} . Because of the complexity, we could not provide a theoretical analysis on the tightness of u_n^{org} . On the contrary, we will see that u_n^{new} is at a constant factor from the diffusion approximation (shown in Section III), which well approximates the true $\mathbf{E}[W_n]$ for some distributions with the given two moments of S and A . Unfortunately, at present we are not aware of the distributions of S and A with the given two moments with which $\mathbf{E}[W_n]$ becomes close to u_n^{new} . In particular, $\mathbf{E}[W_n]$ with extremal distributions, $S = S(1)$ and $A = A(1)$, is away from u_n^{new} approximately by a factor of two, although we will see in Section V that the tight upper-bound is not necessarily achieved with $S = S(1)$ and $A = A(1)$.

B. Lower bounds

When the inter-arrival time is $A(\varepsilon)$ and the service time is $S(\varepsilon)$, we find that $\mathbf{E}[W_n]$ can be made arbitrarily close to 0 by choosing a sufficiently small ε . Notice that $A(\varepsilon)$ (respectively, $S(\varepsilon)$) with $\varepsilon \rightarrow 0$ is a constant, $1/\lambda$ (respectively, $1/\mu$), almost surely and infinite with an infinitesimal probability. Notice that if it is *always* the case that $S = 1/\mu$ and $A = 1/\lambda$, then $\mathbf{E}[W_n] = 0$ because $\lambda < \mu$. It turns out that the infinite S and A have negligible impact on mean delay for a finite number of jobs. A formal statement with its rigorous proof follows.

Theorem 2: Let $W_n(\varepsilon)$ be the delay of the n -th job in a GI/GI/1 queue given that $W_0(\varepsilon) = 0$, where the inter-arrival time is $A(\varepsilon)$ and the service time is $S(\varepsilon)$, which are defined with Equation (2). Then, for any fixed n , we have

$$\lim_{\varepsilon \downarrow 0} \mathbf{E}[W_n(\varepsilon)] = 0.$$

Proof: It is well known that the delay in a GI/GI/1 queue can be expressed as the maximum of a random walk (e.g., see Section 9-1 from [22]):

$$W_n(\varepsilon) =_d \max_{0 \leq k \leq n} \sum_{i=1}^k X_i(\varepsilon),$$

where $=_d$ denotes equality in distribution, $X_i(\varepsilon) =_d S(\varepsilon) - A(\varepsilon)$ with independent $A(\varepsilon)$ and $S(\varepsilon)$, and we define $\sum_{i=1}^0 X_i \equiv 0$.

For $0 < \varepsilon < 1$, we construct a random variable, $Y_i(\varepsilon)$, that stochastically dominates $X_i(\varepsilon)$ (i.e., $X_i(\varepsilon) \leq_{\text{st}} Y_i(\varepsilon)$). Notice that X_i is dominated by $S(\varepsilon) - (1-\varepsilon)/\lambda$, because $A(\varepsilon) \geq (1-\varepsilon)/\lambda$ surely. Further, when $S(\varepsilon) = (1-\varepsilon)/\mu$, we have $S(\varepsilon) - (1-\varepsilon)/\lambda \leq 0$ surely. Therefore, $X_i(\varepsilon)$ is dominated by the following Y_i :

$$Y_i(\varepsilon) = \begin{cases} 0 & \text{w.p. } \frac{C_S^2}{C_S^2+\varepsilon^2} \\ \frac{1}{\mu} \left(1 + \frac{C_S^2}{\varepsilon} \right) - \frac{1-\varepsilon}{\lambda} & \text{w.p. } \frac{\varepsilon^2}{C_S^2+\varepsilon^2}. \end{cases}$$

Then $\mathbb{E}[W_n(\varepsilon)]$ is bounded as follows:

$$\mathbb{E}[W_n(\varepsilon)] \leq \mathbb{E}\left[\max_{0 \leq k \leq n} \sum_{i=1}^k Y_i(\varepsilon)\right] = \mathbb{E}\left[\sum_{i=1}^n Y_i(\varepsilon)\right],$$

where the equality follows from the fact that the maximum of $\sum_{i=1}^k Y_i(\varepsilon)$ for $0 \leq k \leq n$ is always achieved at $k = n$, because $Y_i \geq 0$ surely.

Because $0 \leq \mathbb{E}[W_n]$, it suffices to show that $\mathbb{E}[\sum_{i=1}^n Y_i(\varepsilon)] \rightarrow 0$ as $\varepsilon \downarrow 0$. Observe that the expectation of the sum is equal to the sum of the expectation and therefore,

$$\begin{aligned} \mathbb{E}\left[\sum_{i=1}^n Y_i(\varepsilon)\right] &= \sum_{i=1}^n \mathbb{E}[Y_i(\varepsilon)] \\ &= n \left(\frac{1}{\mu} \left(1 + \frac{C_S^2}{\varepsilon}\right) - \frac{1-\varepsilon}{\lambda} \right) \frac{\varepsilon^2}{C_S^2 + \varepsilon^2}, \end{aligned}$$

which approaches 0 as $\varepsilon \downarrow 0$. This completes the proof of the theorem. \blacksquare

C. Upper bounds

The proof of Theorem 1 is largely algebraic and hence is postponed to Appendix A. Essentially, the proof in Appendix A shows that $u_n^{\text{new}} - u_n^{\text{org}}$, is non-negative at the minimum value of n and increasing with n both for the case with $n < \kappa$ and for the case with $\kappa \leq n \leq 16\kappa/e^2$.

To gain more insights into the relation between u_n^{new} and u_n^{org} (as to why the term π appears in the upper bounds), we show the following asymptotic upper bound:

Proposition 1: For any fixed n , we have

$$\mathbb{E}[W_n] \lesssim \frac{\pi}{2\lambda} \sqrt{n(C_A^2 + \rho^2 C_S^2)},$$

where $f(\rho) \lesssim g(\rho)$ denotes that $\lim_{\rho \rightarrow 1} f(\rho)/g(\rho) \leq 1$.

Proof: Using $C^2 \equiv C_A^2 + \rho^2 C_S^2$ as a shorthand, we start by noting that u_n^{org} for $n \leq \kappa$ can be rewritten as $V_{n,C,\rho}/\lambda$, where

$$\begin{aligned} V_{n,C,\rho} &= \sqrt{nC^2} \arctan\left(\sqrt{\frac{\kappa}{n}} - 1\right) \\ &+ \frac{C^2}{8(1-\rho)} \left(1 - \sqrt{1 - \frac{n}{\kappa}}\right) \\ &- 2n(1-\rho) \ln\left(\sqrt{\frac{C^2}{16n}} \left(1 + \sqrt{1 - \frac{n}{\kappa}}\right)\right) \\ &+ (1-\rho)n \ln\left(\frac{e^2 C^2}{4n}\right) + \frac{2n^2(1-\rho)^3}{3C^2}. \end{aligned}$$

For any n , there exists $\rho_0 < 1$ such that $n \leq \frac{C^2}{16(1-\rho_0)^2}$, so that the bound, $V_{n,C,\rho}/\lambda$, is valid for $\rho > \rho_0$. Now, the proposition follows immediately, because, as $\rho \rightarrow 1$, the first term converges to $\pi C \sqrt{n}/2$, and the other terms converge to 0. \blacksquare

The proof of Proposition 1 implies that u_n^{new} is asymptotically equivalent to u_n^{org} (i.e., the ratio of the two bounds approach 1) in the limit of $\rho \rightarrow 1$. In the rest of this section, we discuss the gap between the two bounds for general ρ .

For $n \leq \kappa$, u_n^{new} given by (3) is derived using u_n^{org} given by (6) from [18]. We find that u_n^{new} is larger than u_n^{org} at most by a factor of $48\pi/(97 + 24\ln 4) \approx 1.157$. This factor can be shown as follows. First, we find that the gap between the two upper-bounds is increasing with n (see the proof of Theorem 1 in Appendix A). The maximum value of n is $n = \kappa$. At this n , we have

$$u_n^{\text{new}} = \frac{\pi C_A^2 + \rho^2 C_S^2}{8\lambda(1-\rho)} \quad \text{and} \quad u_n^{\text{org}} = \frac{97 + 24\ln 4}{384} \frac{C_A^2 + \rho^2 C_S^2}{\lambda(1-\rho)}.$$

The new upper-bound (4) in Theorem 1 is derived using the original upper-bound (5) in [18]. Again, the gap between the two bounds is increasing with n (see the proof of Theorem 1 in Appendix A). At the maximum value of $n = 16\kappa/e^2$, we have

$$\begin{aligned} u_n^{\text{new}} &= \frac{1}{2} \frac{C_A^2 + \rho^2 C_S^2}{\lambda(1-\rho)} \\ u_n^{\text{org}} &= \frac{16 + 3e^4 + 24e^2(4 - \ln 4)}{24e^4} \frac{C_A^2 + \rho^2 C_S^2}{\lambda(1-\rho)}. \end{aligned}$$

Thus, u_n^{new} is larger than u_n^{org} at most by a factor of $12e^4/(16 + 3e^4 + 24e^2(4 - \ln 4)) \approx 1.018$.

III. A GUARANTEE ON DIFFUSION APPROXIMATION

The upper-bound established in Theorem 1, for example, can be used to provide a guarantee on the accuracy of an approximation of $\mathbb{E}[W_n]$. In this section, we will compare u_n^{new} against a diffusion approximation, which is a standard approach to approximating the transient GI/GI/1 queue. (We should note that the diffusion approximation does not necessarily give upper or lower bounds). With a diffusion approximation, when a GI/GI/1 queue is empty at time 0, the workload, $Z(t)$, in the queue at time t is approximated with a reflected Brownian motion as follows (page 144 from [8]):

$$Z(t) \approx \hat{Z}(t) = \frac{1}{\mu} \text{RBM}(t, \lambda - \mu, \lambda(C_A^2 + C_S^2)), \quad (7)$$

where $\text{RBM}(t, \theta, \sigma^2)$ denotes the state of the reflected Brownian motion with drift θ and variance σ^2 at time t given that the state is 0 at time 0.

By applying Corollary 2.3.1 from [1] and then Corollary 1.1.1 from [1] to Equation (7), we obtain a general formula for $\rho < 1$ to approximate $\mathbb{E}[Z(t)]$ as

$$\begin{aligned} \mathbb{E}[\hat{Z}(t)] &= \frac{\lambda(C_A^2 + C_S^2)}{\mu(\mu - \lambda)} \mathbb{E}\left[\text{RBM}\left(\frac{(\mu - \lambda)^2}{\lambda(C_A^2 + C_S^2)}t, -1, 1\right)\right] \\ &= \frac{\rho^2(C_A^2 + C_S^2)}{\lambda(1-\rho)} \left(\frac{1}{2} - (\eta_\rho^2 + 1)(1 - \Phi(\eta_\rho)) + \eta_\rho \phi(\eta_\rho)\right), \end{aligned} \quad (8)$$

where $\Phi(\cdot)$ is the standard normal distribution function, $\phi(\cdot)$ is its density, and $\eta_\rho \equiv \frac{1-\rho}{\rho} \sqrt{\frac{\lambda t}{C_A^2 + C_S^2}}$.

The expression of $\mathbb{E}[\hat{Z}(t)]$ in Equation (8) is simplified in the heavy-traffic limit of $\rho \rightarrow 1$. Specifically, as $\rho \rightarrow 1$, we have $\eta_\rho \rightarrow 0$, so that $\Phi(\eta_\rho) \rightarrow 1/2$ and $\phi(\eta_\rho) \rightarrow 1/\sqrt{2\pi}$.

Thus, we have a simpler asymptotic formula to approximate $E[Z(t)]$ as

$$E[\hat{Z}(t)] \sim \frac{1}{\lambda \sqrt{2\pi}} \sqrt{\lambda t (C_A^2 + C_S^2)}, \quad (9)$$

where $f(\rho) \sim g(\rho)$ denotes $f(\rho)/g(\rho) \rightarrow 1$ as $\rho \rightarrow 1$.

Approximations for $E[W_n]$ can be constructed from $E[\hat{Z}(t)]$ by replacing the arrival time, t , of the n -th job with its expected value, n/λ . This will result in two types of approximations. Specifically, the first approximation, \tilde{w}_n , is obtained from Equation (8) by replacing λt with n , and the second approximation, \tilde{w}'_n , is obtained from the simpler asymptotic Equation (9):

$$\tilde{w}_n = \frac{\rho^2 (C_A^2 + C_S^2)}{\lambda(1-\rho)} \left(\frac{1}{2} - (\tilde{\eta}_\rho^2 + 1)(1 - \Phi(\tilde{\eta}_\rho)) + \tilde{\eta}_\rho \phi(\tilde{\eta}_\rho) \right) \quad (10)$$

$$\tilde{w}'_n = \frac{1}{\lambda \sqrt{2\pi}} \sqrt{n(C_A^2 + C_S^2)}, \quad (11)$$

$$\text{where } \tilde{\eta}_\rho \equiv \frac{1-\rho}{\rho} \sqrt{\frac{n}{C_A^2 + C_S^2}}.$$

Observe that the simpler asymptotic approximation, \tilde{w}'_n , is proportional to \sqrt{n} and quite similar to u_n^{new} (see Theorem 1). This enable us to prove that the true $E[W_n]$ is at most four times larger than \tilde{w}'_n for any S and A with the given first two moments.

Corollary 2: Let \tilde{w}'_n be as defined by Equation (11). Then

$$E[W_n] \leq \begin{cases} \frac{\pi^{3/2}}{\sqrt{2}} \tilde{w}'_n \approx 3.93 \tilde{w}'_n & \text{for } n < \kappa \\ \frac{e\sqrt{\pi}}{\sqrt{2}} \tilde{w}'_n \approx 3.40 \tilde{w}'_n & \text{for } n \geq \kappa. \end{cases} \quad (12)$$

Proof: When $n < \kappa$, the upper bound (3) is valid. Hence,

$$E[W_n] \leq \frac{\pi}{2\lambda} \sqrt{n(C_A^2 + C_S^2)} = \frac{\pi^{3/2}}{\sqrt{2}} \tilde{w}'_n,$$

because $\rho < 1$. An analogous argument can be used to show (12) for $\kappa \leq n \leq 16\kappa/e^2$, because the upper bound (4) is valid in this range of n . It turns out that (12) holds for $n > 16\kappa/e^2$ as well. Notice that the upper bound (4) is still valid for $n > 16\kappa/e^2$, because (4) is larger than Kingman's upper bound (which is an upper bound on $E[W_n]$). ■

IV. NUMERICAL EVALUATION

Next, we further study the characteristics of the diffusion approximations (\tilde{w}_n and \tilde{w}'_n) and our upper bounds (u_n^{new} and u_n^{org}) with numerical evaluation.

A. Diffusion approximation

In this section, we show by experiments that neither \tilde{w}_n nor \tilde{w}'_n give an upper bound of $E[W_n]$. Although \tilde{w}_n might approximate $E[W_n]$ well for some S and A with the given two moments, there are cases where \tilde{w}_n is far from $E[W_n]$.

Figure 1 shows \tilde{w}_n with a dashed line and \tilde{w}'_n with a dotted line, where we set $\rho = \lambda = 0.9$ and $C_A = C_S = 2$. Observe that $\tilde{w}'_n < \tilde{w}_n$ for $n \leq 560$ and $\tilde{w}'_n > \tilde{w}_n$ for $n > 560$. The figure also shows the simulated values of $E[W_n]$ for two settings of the pair of S and A , where $\rho = \lambda = 0.9$ and $C_A = C_S = 2$ are fixed. Specifically, the solid circles show $E[W_n]$ when $S = S(1)$ and $A = A(1)$, using the definitions of $A(\varepsilon)$ and $S(\varepsilon)$ in (2); the crosses show $E[W_n]$ with $S =$

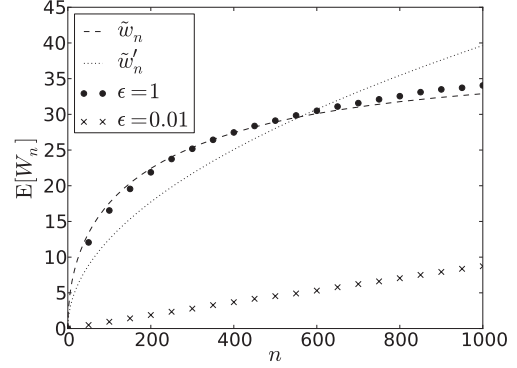


Fig. 1. The dashed line shows \tilde{w}'_n , the dotted line shows \tilde{w}_n , the solid circle shows $E[W_n]$ simulated with $S(1)$ and $A(1)$, and the cross shows $E[W_n]$ simulated with $S(0.01)$ and $A(0.01)$, where $\rho = \lambda = 0.9$ and $C_A = C_S = 2$.

$S(0.01)$ and $A = A(0.01)$. For each setting, the simulation is repeated 100,000 times, and we have confirmed that the 99 % confidence-intervals (not shown in the figure for clarity) are sufficiently small (specifically, smaller than the size of the solid circle).

We find that \tilde{w}_n well approximates $E[W_n]$ when $S = S(1)$ and $A = A(1)$. Taking a closer look, we see that \tilde{w}_n slightly overestimates the $E[W_n]$ for $n \leq 350$ but underestimates the $E[W_n]$ for $n \geq 400$. Also, \tilde{w}'_n can underestimate or overestimate $E[W_n]$ depending on n . This shows, despite the good approximation values, neither \tilde{w}_n nor \tilde{w}'_n gives an upper bound of $E[W_n]$.

Moreover, \tilde{w}_n is a poor approximation of $E[W_n]$ when the service time is $S(0.01)$ and the interarrival time is $A(0.01)$, even though the first two moments of $S(0.01)$ (respectively, $A(0.01)$) are the same as those of $S(1)$ (respectively, $A(1)$). This is not surprising, because Theorem 2 states that $E[W_n]$ can be arbitrarily close to 0 as long as the first two moments of S and A are finite and $\rho < 1$. The limitation of approximating $E[W_n]$ with first two moments is evident in this example shown in Figure 1.

B. Bounds

Recall that u_n^{new} is provably close (within a factor of 1.16) to u_n^{org} and differs from \tilde{w}'_n at most by a factor of 4. We now compare the upper bounds and the diffusion approximations in more detail with numerical evaluations.

Figure 2 shows u_n^{new} and u_n^{org} in solid lines. Notice that the two solid-lines overlap with each other, but $u_n^{\text{org}} \leq u_n^{\text{new}}$ holds for all n . A detail is that u_n^{new} drops at $n \approx 79,000$, switching from (3) to (4). Also, u_n^{new} flattens out at $n \approx 170,000$, switching from (4) to Kingman's upper-bound. The dashed line shows \tilde{w}_n , and the dotted line shows \tilde{w}'_n . We find that $u_n^{\text{new}} \approx 2\tilde{w}_n$. Because \tilde{w}_n well approximates $E[W_n]$ with $S = S(1)$ and $A = A(1)$ as we have seen in Section IV-A, the upper bounds in Theorem 1 (u_n^{new}) could be improved at most by a factor of two. Here, we set $C_A = C_S = 8$ and $\rho = 0.99$. Such high variability and high load might be too

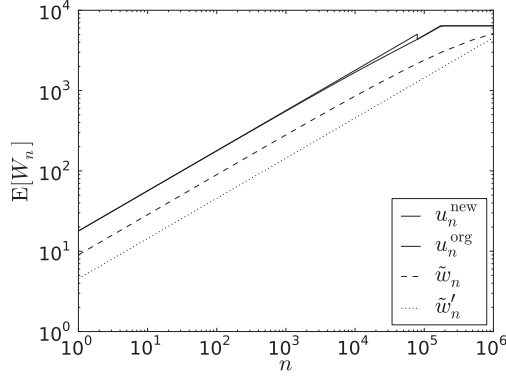


Fig. 2. Upper bounds, u_n^{new} and u_n^{org} , and approximations, \tilde{w}_n and \tilde{w}'_n , of $E[W_n]$ are shown when $C_A = C_S = 8$ and $\rho = 0.99$.

extreme for some applications; thus we next study how the results change with lower variability and lower load.

Figure 3 shows the upper bounds and the approximations similar to Figure 2. In the top two rows, $\rho = 0.99$ is fixed, and $C_A = C_S$ is varied as specified in the figure. In the bottom two rows, $C_A = C_S = 8$ is fixed, and ρ is varied. Observe that the upper bounds and the approximations have relations that are qualitatively similar to those in Figure 2. A major difference is in the range of n where u_n^{new} and u_n^{org} are non-trivial in the sense that they are strictly smaller than Kingman's upper-bound. Note that $n \leq 10^6$ is shown in Figure 2, $n \leq 10^5$ in Figure 3 (a)-(c), and $n \leq 10^4$ in Figure 3 (d).

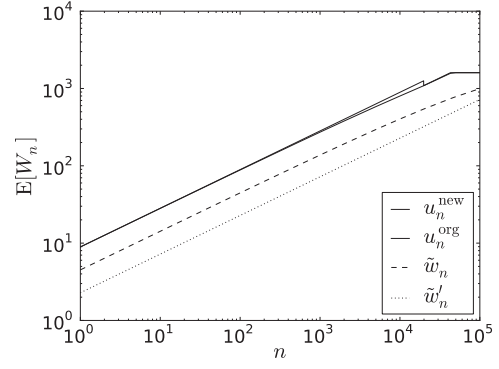
Figure 4 shows contour lines that illustrate the maximum n such that u_n^{new} is below Kingman's upper-bound. For simplicity, we assume $C_S = C_A$. The vertical axis shows $C = C_A = C_S$, and the horizontal axis shows ρ . We vary ρ between 0.8 and 1.0, because u_n^{new} is rarely useful at $\rho < 0.8$.

Note, however, that bounds on $E[W_n]$ are usually needed at high load. As n increases at low load, $E[W_n]$ converges quickly to $E[W]$, the mean delay at steady state. Hence, an existing upper-bound² on $E[W]$ should be used as an upper bound on $E[W_n]$, although, at low load, one would rarely be interested in bounding $E[W_n]$, which is known to be short. Our expectation is that u_n^{new} gives insights into the delay when high load persists for a short period, so that the steady state is not reached during the period of high load.

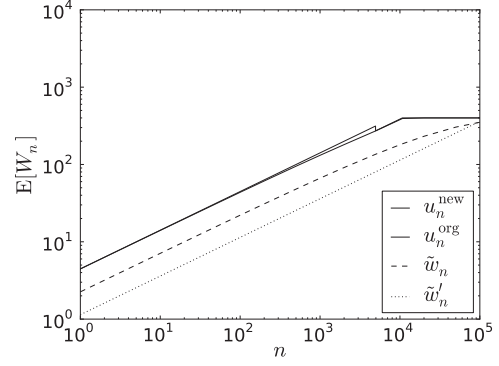
C. Transient Mean Delay of Internet packets

In this section, we show how to use our upper bounds for bounding the transient mean delay of Internet packets of the MAWI Working Group Traffic Archive, where there are dependencies in inter-arrival times and in service times. We use the daily traces at the Samplepoint-F on the 29th and 30th of November 2010 (respectively, *201011291400.dump* and *201011291400.dump* files). The dumpfiles produced by tcpdump contain various information of IP packets, but we

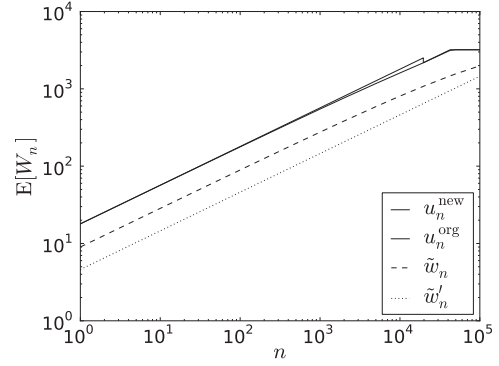
²Daley's upper-bound [10] is superior to Kingman's [13] at low load.



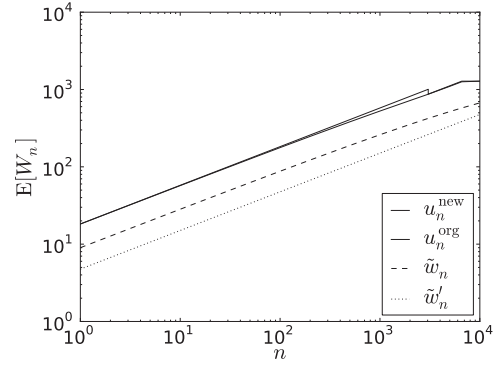
(a) $C_A = C_S = 4$, $\rho = 0.99$



(b) $C_A = C_S = 2$, $\rho = 0.99$



(c) $C_A = C_S = 8$, $\rho = 0.98$



(d) $C_A = C_S = 8$, $\rho = 0.95$

Fig. 3. Upper bounds, u_n^{new} and u_n^{org} , and approximations, \tilde{w}_n and \tilde{w}'_n , of $E[W_n]$ are shown with varying values of $C_A = C_S$ and ρ , which are as specified right below each figure.

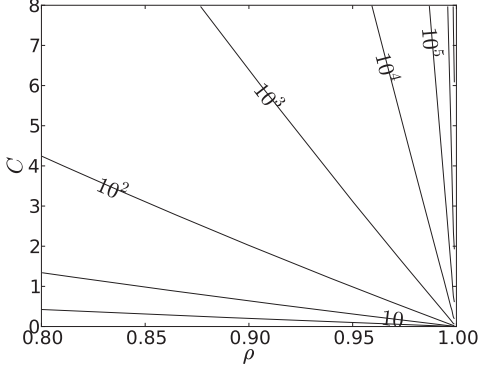


Fig. 4. Contour lines show the maximum n such that u_n^{new} is below Kingman's upper-bound for a given pair of ρ (horizontal axis) and $C = C_S = C_A$ (vertical).

only use the fields of arrival time and size of IP packets which are recorded at its header³.

Each of the dumpfile contains samples of IP traffic data for 15 minutes, where more than 40 million packets are recorded. We use the information from the first $N = 19,000,000$ packets, dividing them into 1,900 intervals each of which contains 10,000 contiguous packets⁴. For each interval i , we compute the transient delay of the n -th packet $W_{i,n}$ from its inter-arrival time $A_{i,n}$ and its service time $S_{i,n}$ according to

$$W_{i,n+1} = \max(0, W_{i,n} - A_{i,n+1} + S_{i,n}), \quad (13)$$

where $W_{i,0} = S_{i,0} = 0$. The transient mean delay, $E[W_n]$, is the average of $W_{i,n}$ over all intervals i . The service time $S_{i,n}$ is obtained from the size of the packet i divided by the bandwidth of the link, which is set to 40 Mbps, so that the load $\rho \approx 0.99$ on November 30, instead of the over-provisioned 150 Mbps to simulate the short period of high load.

In the real-world Internet traffic, there are dependencies between the interarrival times (respectively, the service times), and the trace under considerations is not an exception. Figure 5 illustrates the dependencies in inter-arrival times and in service times. Specifically, the solid line with circles shows the sample variance parameter of batch size k , which is defined as follows⁵:

$$\frac{1}{[N/k]-1} \sum_{i=1}^{[N/k]} \left(\sum_{j=(i-1)k+1}^{ik} A_j - k \bar{A} \right)^2, \quad (14)$$

where A_j is the j -th inter-arrival time for $1 \leq j \leq N$, and $\bar{A} = \sum_{i=1}^N A_i / N$ is the sample average of the inter-arrival times on November 29. The batch size k is varied along the horizontal axis. Notice that (14) is an estimate of the variance

³The details of traffic trace information on November 29, 2010 are listed at <http://tracer.csl.sony.co.jp/mawi/samplepoint-F/2010/201011291400.html>

⁴The choice of the first 19 million lines is solely due to the limitation of the tcpdump/wireshark we use to retrieve the header information of the packets from the dumpfiles.

⁵Notice that the variance parameter is not normalized by the corresponding mean unlike the index of dispersion for intervals.

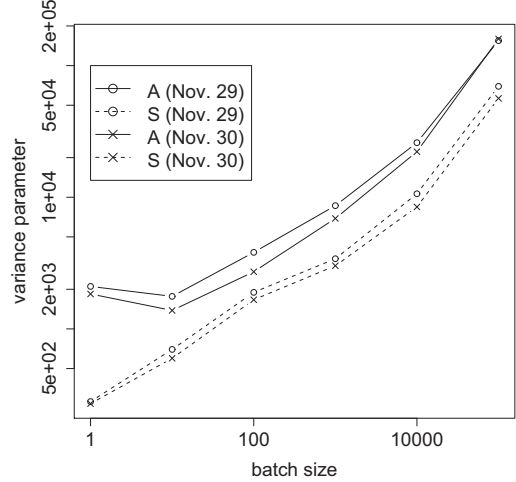


Fig. 5. The sample variance-parameter of the inter-arrival times (solid lines) and the service times (dashed lines) on November 29 (circles) and on November 30 (crosses) in 2010.

parameter of batch size k :

$$\frac{\text{Var}[A_1 + A_2 + \dots + A_k]}{k}. \quad (15)$$

If A_1, A_2, \dots were i.i.d., the variance parameter would equal to the corresponding variance, $\text{Var}[A_1]$. When the inter-arrival times have positive auto-correlations, variance parameter becomes greater than the corresponding variance, $\text{Var}[A_1]$. In Figure 5, the solid line with circles increases with k and does not appear to converge. That is, the inter-arrival times appear to have long-range dependence. The solid line with crosses shows the sample variance-parameter for the inter-arrival times on November 30. The dashed line with circles (respectively, crosses) shows the sample-variance parameter for the service times on November 29 (respectively, November 30). Observe that all of these lines are increasing with the batch size k .

Recall that our upper bounds in Section II are derived under the assumptions that the inter-arrival times and the service times, respectively, are independent. Because the strong dependency is evident from Figure 5, our upper bounds are not directly applicable to the delay that the packets experience in our settings. Therefore, in this section, we consider an approximate bound, or a quasi-bound, that appears to be effective when there are dependencies in inter-arrival times and in service times. The quasi-bound is motivated by a diffusion approximation for correlated stochastic processes.

A standard approach for a diffusion approximation of a correlated stochastic process is to use the asymptotic variance parameter of the correlated stochastic process in place of the variance in the expression of the diffusion approximation for a corresponding uncorrelated stochastic process (see Section 4.4 from [21]). The asymptotic variance parameter can be defined as (15) in the limit as $k \rightarrow \infty$. In our context, this corresponds to replacing the variance, which appears as a squared coefficient of variation (i.e., C_A^2 and C_S^2), in the

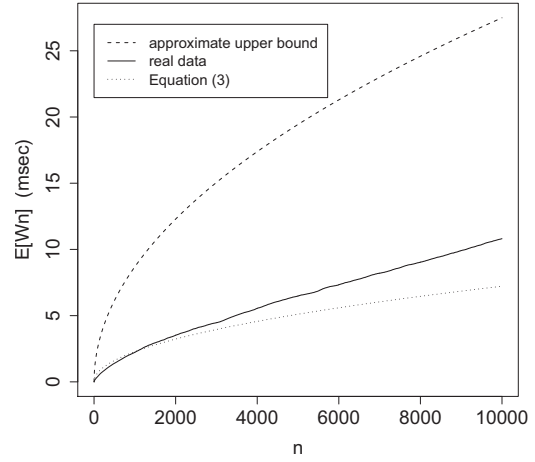
expression of the bounds in Theorem 1 with the corresponding asymptotic variance parameter. However, Figure 5 suggests that the asymptotic variance parameter is unbounded for the inter-arrival times and for the service times under considerations. Also, the delay of the first k packets is independent of the inter-arrival times and the service times of the packets that arrive after the k -th packet. Hence, it does not make intuitive sense to use the asymptotic variance parameter, which captures the correlation between infinitely many inter-arrival times, in a quasi-bound for $E[W_n]$ of a finite n .

The above arguments lead us to propose the following upper quasi-bound on $E[W_n]$ when there are dependencies in the inter-arrival times and/or in the service times. Namely, in the expressions of the bounds on $E[W_n]$ in Theorem 1, we replace the variance with the variance parameter of batch size n . For example, C_A^2 in the right-hand side of (3) is replaced with the variance parameter (15) of batch size $k = n$ divided by $E[A]^2$. We will now study the effectiveness of this quasi-bound.

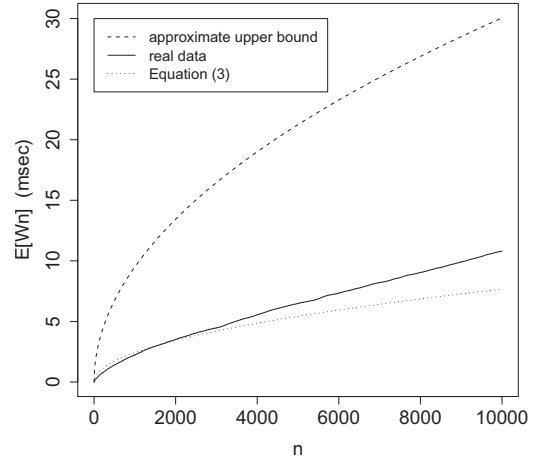
The solid line in Figure 6 (a) shows the transient delay of the 10,000 packets, averaged over the 1,900 intervals, on November 30. Recall that the delay is calculated with (13). The dotted line shows the upper bound according to Equation (3) without considering the dependency in the inter-arrival times and in the service times. The dashed line shows the upper quasi-bound, where we use the variance parameter of batch size 10,000⁶. In Figure 6 (a), all values of parameters in Equation (3) and in the corresponding quasi-bound are obtained from the corresponding sample averages, sample variances, and sample variance-parameters of size 10,000, using all of the N packets on November 30. Specifically, $E[A] = 18.73 \mu\text{sec}$, $E[S] = 18.69 \mu\text{sec}$, $\sqrt{\text{Var}[A]} = 42.92 \mu\text{sec}$, $\sqrt{\text{Var}[S]} = 16.45 \mu\text{sec}$, and the sample variance-parameters are shown in Figure 5 with circles.

Meanwhile, in Figure 6 (b) those values are obtained using the N packets on the previous day (November 29). Specifically, $E[A] = 21.76 \mu\text{sec}$, $E[S] = 18.61 \mu\text{sec}$, $\sqrt{\text{Var}[A]} = 45.85 \mu\text{sec}$, $\sqrt{\text{Var}[S]} = 16.75 \mu\text{sec}$, and the sample variance-parameters are shown in Figure 5 with crosses. From both plots in Figure 6, we can see that the upper bound shown with the dotted line is only good for approximately the first 1000 packets: the upper bound given by Equation (3) is lower than the realized average-delay for $n > 1000$, because there are dependencies that are ignored in Equation (3). We can observe that the upper quasi-bound is always greater than the realized average-delay, and the difference between the upper quasi-bound and the realized average-delay is within a factor of three. Comparing Figure 6 (a) and Figure 6 (b), we can observe that it is possible to use the values of parameters from the previous day to obtain upper bounds of the transient mean delay time because the values of the corresponding variance parameters do not vary much as shown in Figure 5.

⁶The smaller the batch size, the smaller the sample variance parameter tends to be (see Figure 5). Hence, the upper quasi-bound would be between the two lines in Figure 6 (a), if a smaller batch size is used.



(a) Parameters are estimated from the data on November 30



(b) Parameters are estimated from the data on November 29

Fig. 6. The solid line shows the realized delay of the 10,000 packets averaged over 1,900 on November 30, the dotted line shows the corresponding upper bound of the mean according to Equation (3), and the dashed line shows the corresponding upper quasi-bound.

V. PROPERTIES OF TRANSIENT GI/GI/1 QUEUE

We have shown that the trivial lower-bound on $E[W_n]$ is in fact tight (see Theorem 2), but the upper bounds in Theorem 1 and Corollary 1 (u_n^{new} and u_n^{org}), do not appear to be tight. In this section, we illustrate the difficulty in establishing the tight upper-bound on $E[W_n]$ given the first two moments of A and S , using extremal distributions that have mass probabilities at two particular points.

A. Summary of results

We will see, by constructing an example, that a pair of extremal distributions as in Equation (2), $A(1)$ and $S(1)$, do not necessarily maximize $E[W_n]$ given the first two moments of S and A . Recall that $E[W_n]$ can be made arbitrarily close to zero by the other pair of extremal distributions, $A(\varepsilon)$ and

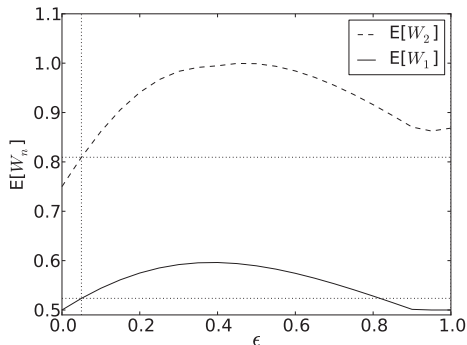


Fig. 7. The solid line shows $E[W_1]$, and the dashed line shows $E[W_2]$ as a function of ϵ , the parameter of the service time $S(\epsilon)$. The inter-arrival time is $A(1)$, and we set $\lambda = 0.95$, $\mu = 1$, and $C_A = C_S = 1$.

$S(\epsilon)$ with $\epsilon \rightarrow 0$. We will also provide other reasons why establishing a tight upper-bound is nontrivial.

Let \mathcal{A} be a set of random variables whose first two moments are fixed so that the mean is $1/\lambda$ and the coefficient of variation is C_A . Let \mathcal{S} be defined analogously. Let $\hat{A}, \check{A} \in \mathcal{A}$ and $\hat{S}, \check{S} \in \mathcal{S}$. Let \hat{W}_n be the n -th delay in the GI/GI/1 queue with $A = \hat{A}$ and $S = \hat{S}$, and let \check{W}_n be defined analogously.

One might expect that $E[\hat{W}_n] \geq E[\check{W}_n]$ for some n implies $E[\hat{W}_n] \geq E[\check{W}_n]$ for any n . If this were true, it would suffice to find the inter-arrival time $A^* \in \mathcal{A}$ and the service time $S^* \in \mathcal{S}$ such that $E[W_1]$ is maximized. Then the tight upper-bound on $E[W_n]$ would be obtained by evaluating $E[W_n]$ with $A = A^*$ and $S = S^*$. However, we will find an example where $E[\hat{W}_m] > E[\check{W}_m]$ and $E[\hat{W}_n] < E[\check{W}_n]$ for $m \neq n$.

One might also suspect that if $E[W_n]$ is smaller with $S = \hat{S}$ and $A = \hat{A}$ than with $S = \check{S}$ and $A = \check{A}$, then this order does not change after scaling \hat{A} and \check{A} by the same amount without changing the shape of each distribution (i.e., only the load changes). If this were true, the tight upper-bound would be obtained by finding $A^* \in \mathcal{A}$ and $S^* \in \mathcal{S}$ that maximize $E[W_n]$ at a load (e.g., at the light-traffic limit of $\rho \rightarrow 0$). Unfortunately, we will find a counter example.

B. Analysis

In this section, we consider a GI/GI/1 queue with $A = A(1)$ and $S = S(\epsilon)$, where $A(1)$ and $S(\epsilon)$ are as defined by Equation (2) with $\mu = 1$, and $C_A = C_S = 1$. We will vary λ as specified in the following. With these settings, the inter-arrival time is 0 with probability 0.5 and $2/\lambda$ with probability 0.5. In other words, a batch of jobs arrives according to a Poisson process with rate $\lambda/2$, and the size of a batch has a geometric distribution (i.e., the size is k with probability $1/2^k$ for $k \geq 1$). We study the delay, $W_n(\epsilon)$, of the n -th job in this GI/GI/1 queue. Notice that we can evaluate $E[W_n(\epsilon)]$ analytically for small n (specifically $n = 1, 2$), using Lindley's equation, $W_0 = 0$ and $W_{n+1} = \max\{W_n + S_n - A_n, 0\}$ for $n \geq 0$, in a straightforward way. Recall that the first job may experience a non-zero delay, W_1 , because the zero-th job arrives at the empty queue in our settings.

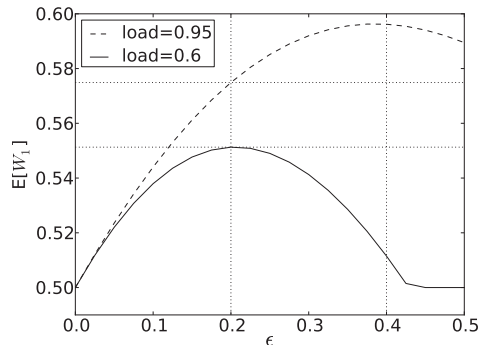


Fig. 8. The lines show $E[W_1]$ as function of ϵ , the parameter of the service time $S(\epsilon)$, where λ is varied as labeled. The inter-arrival time is $A(1)$, and we set $\mu = 1$ and $C_A = C_S = 1$.

Figure 7 shows $E[W_1(\epsilon)]$ with a solid line and $E[W_2(\epsilon)]$ with a dashed line as a function of ϵ , where we set $\lambda = 0.95$. Notice that $E[W_1(1)]$ is the mean delay with the pair of extremal distributions $A(1)$ and $S(1)$, but $E[W_1(1)]$ is smaller than $E[W_1(\epsilon)]$ for a wide range of ϵ . A similar observation can be made for $E[W_2(1)]$. Therefore, the pair of extremal distributions, $S(1)$ and $A(1)$, does not maximize $E[W_n]$ for some n , although the other pair of extremal distributions, $S(\epsilon)$ and $A(\epsilon)$ with $\epsilon \rightarrow 0$, minimizes $E[W_n]$ for any n .

Also, in Figure 7, observe that $E[W_1(0.05)] > E[W_1(1)]$ and $E[W_2(0.05)] < E[W_2(1)]$. That is, in expectation, the first job experiences more delay with $S(0.05)$ than with $S(1)$, but the second job experiences more delay with $S(1)$ than with $S(0.05)$. Here, $S(0.05)$ takes a value (i.e., 0.95) that is close to the expected value (i.e., 1) most of the time (i.e., with probability $1/(1+0.05^2) \approx 0.997$) and takes a huge value (i.e., 21) with a small probability. On the other hand, $S(1)$ takes an infinitesimal value (i.e., 0) or a large value (i.e., 2) with equal probabilities. Notice that a job of size 0 experiences a delay in the queue but does not contribute to the delay of the other jobs.

In Figure 8, we study only $E[W_1(\epsilon)] = E[\max\{S(\epsilon) - A(1), 0\}]$ but with varying λ : $\lambda = 0.6$ (low load) and $\lambda = 0.95$ (high load). Figure 8 shows $E[W_1(\epsilon)]$ with the solid line when the load is low ($\lambda = 0.6$) and with the dashed line when the load is high ($\lambda = 0.95$). Now, we investigate $E[W_1(0.2)]$ and $E[W_1(0.4)]$. At low load, $S(0.4)$ causes shorter expected-delay to the first job than $S(0.2)$ (i.e., $E[W_1(0.4)] < E[W_1(0.2)]$). At high load, however, $E[W_1(0.4)] > E[W_1(0.2)]$. Notice that $S(0.2)$ takes a value (i.e., 0.8) that is close to the expected value most of the time (i.e., with probability ≈ 0.96) and a large value (i.e., 6) with a small probability. On the other hand, $S(0.4)$ is more balanced in that $S(0.4) = 0.6$ with probability ≈ 0.86 and $S(0.4) = 3.5$ otherwise.

VI. CONCLUSION

We have studied lower and upper bounds on $E[W_n]$ given the first two moments of the service time, S , and the inter-arrival time, A , respectively. We have proved that the $E[W_n]$

can be made arbitrarily close to zero with a pair of extremal distributions, $S(\varepsilon)$ and $A(\varepsilon)$ with $\varepsilon \rightarrow 0$. Thus, the trivial lower-bound of zero is tight. On the contrary, the other pair of extremal distributions, $S(1)$ and $A(1)$, does not necessarily maximize $E[W_n]$. We have constructed counter examples, where $S(\varepsilon)$ and $A(\varepsilon)$ with $\varepsilon < 1$ can make $E[W_n]$ larger than the extremal distributions, $S(1)$ and $A(1)$. Our results with extremal distributions may be compared against the existing results with extremal distributions for the GI/GI/1 queue in steady state (see the last paragraph of Section I-B). We have also seen that tight upper-bound cannot be obtained by simply studying $E[W_n]$ for a particular n or at a particular load and extending it to the general case.

Because establishing the tight upper-bound appears to be hard, we have constructed a simple upper-bound. The new upper-bound is effective when the load is high, and S and A have high variability, which is when such bounds are needed and interesting. As a way to study the tightness of the new upper-bound, we compare it against a standard diffusion-approximation of $E[W_n]$, because the diffusion approximation has a small approximating error for some distributions with the given first two moments. Our numerical experiments suggest that the new upper-bound is within a constant factor (approximately two) from the tight upper-bound.

We emphasize that a bound is not an approximation. For example, the trivial lower-bound is useless as an approximation. However, the tightness of the trivial lower-bound implies that a diffusion approximation or another approximation based on the first two moments of S and A can overestimate the true $E[W_n]$ by an arbitrarily large factor, which might be a useful information for practitioners who use the diffusion approximation to estimate $E[W_n]$. Also, given the first two moments of S and A , our upper-bound can give a provable guarantee that $E[W_n]$ is below a certain value, a new result that cannot be obtained with approximation approaches.

When there are dependencies in inter-arrival times and in service times, we have shown that our upper bound can be modified to provide an approximate upper bound on the transient mean delay. The approximate upper bound is motivated by the use of asymptotic variance parameters in diffusion approximations for correlated stochastic processes, but we suggest that a variance parameter with a finite batch size should be used in our approximate upper bound. The effectiveness of the approximate upper bound is validated with traces from the real-world Internet traffic.

REFERENCES

- [1] J. Abate and W. Whitt. Transient behavior of regulated Brownian motion, I: Starting at the origin. *Advances in Applied Probability*, 19(3):560–598, 1987.
- [2] J. Abate and W. Whitt. Transient behavior of the M/M/1 queue: Starting at the origin. *Queueing systems*, 2(3):41–65, 1987.
- [3] S. Asmussen and C.-L. Wang. Variance reduction for simulating transient GI/G/1 behavior. *Probability in the Engineering and Informational Sciences*, 10(2):197–205, 1996.
- [4] R. Bergmann, D. J. Daley, T. Rolski, and D. Stoyan. Bounds for cumulants of waiting-times in GI/GI/1 queues. *Optimization: A Journal of Mathematical Programming and Operations Research*, 10(2):257–263, 1979.

- [5] D. Bertsimas and K. Natarajan. A semidefinite optimization approach to the steady-state analysis of queueing systems. *Queueing Systems: Theory and Applications*, 56:27–39, 2007.
- [6] P. Borgnat, G. Dewaele, K. Fukuda, P. Abry, and K. Cho. Seven years and one day: Sketching the evolution of Internet traffic. In *Proceedings of the 28th Conference on Computer Communications (IEEE INFOCOM 2009)*, pages 711–719, April 2009.
- [7] L. Breuer. Transient and stationary distributions for the GI/G/k queue with Lebesgue-dominated inter-arrival time distribution. *Queueing Systems*, 45(1):47–57, 2003.
- [8] H. Chen and D. D. Yao. *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization*. Springer, 2000.
- [9] J. W. Cohen. *The Single Server Queue*. Elsevier Science Publishing Company, 1969.
- [10] D. J. Daley. Inequalities for moments of tails of random variables, with a queueing application. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 41:139–143, 1977.
- [11] D. J. Daley and C. D. Trengove. Bounds for mean waiting times in single-server queues: A survey. Technical report, Department of Statistics, The Australian National University, 1977.
- [12] W.-B. Gong and J.-Q. Hu. The Maclaurin series for the GI/G/1 queue. *Journal of Applied Probability*, 29(1):176–184, 1992.
- [13] J. F. C. Kingman. Some inequalities for the GI/GI/1 queue. *Biometrika*, 49:315–324, 1962.
- [14] H. Kobayashi and B. L. Mark. *System Modeling and Analysis: Foundations of System Performance Evaluation*. Prentice Hall, 2008.
- [15] J. Limón-Robles and M. A. Wortman. On the time-dependent occupancy distribution of the G/G/1 queueing system. *Probability in the Engineering and Informational Sciences*, 23(2):261–280, 2009.
- [16] A. Muller and D. Stoyan. *Comparison Methods for Stochastic Models and Risks*. John Wiley & Sons, 2002.
- [17] T. Osogami and R. Raymond. Semidefinite optimization for analysis of queues in closed forms. Technical Report RT0896, IBM Research - Tokyo, March 2010. <http://www.research.ibm.com/tr/people/osogami/paper/RT0896.pdf>.
- [18] T. Osogami and R. Raymond. Semidefinite optimization for transient analysis of queues. In *Proceedings of the ACM SIGMETRICS 2010*, pages 363–364, June 2010.
- [19] C.-L. Wang. An identity of the GI/G/1 transient delay and its applications. *Probability in the Engineering and Informational Sciences*, 16(1):47–66, 2002.
- [20] W. Whitt. On approximations for queues, I: Extremal distributions. *AT&T Bell Laboratories Technical Journal*, 63(1):115–138, 1984.
- [21] W. Whitt. *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues*. Springer, 2002.
- [22] R. W. Wolff. *Stochastic Modeling and the Theory of Queues*. Prentice Hall, 1989.

APPENDIX

A. Proof of Theorem 1

We begin with a simple proof for the case $\kappa \leq n \leq 16\kappa/e^2$. From the bound of (5) in Corollary 1, we have

$$\frac{\lambda}{(1-\rho)} E[W_n] \leq 2\kappa + n \ln \frac{4e^2\kappa}{n} + \frac{n^2}{24\kappa} \equiv v_n^{\text{org}}.$$

For $\kappa \leq n \leq 16\kappa/e^2$, We will show that v_n^{org} , the right-hand side of the above inequality, is at most v_n^{new} , defined as

$$v_n^{\text{new}} \equiv \frac{e}{2(1-\rho)} \sqrt{n(C_A^2 + C_S^2 \rho^2)} = 2e\sqrt{n\kappa}.$$

This is accomplished by showing that $f(n) \equiv v_n^{\text{new}} - v_n^{\text{org}}$ is nonnegative when $\kappa \leq n \leq 4\kappa$ (notice that the range of $\kappa \leq n \leq 4\kappa$ is wider than what is needed, i.e. $\kappa \leq n \leq 16\kappa/e^2$, but the expressions appearing in the following are simplified by considering this wider range of n). First, it is straightforward to verify that

$$\frac{d^3 f(n)}{dn^3} = \frac{3e\sqrt{\kappa n} - 4n}{4n^3} > 0$$

for $\kappa \leq n \leq 16\kappa/e^2$, so that $\frac{d^2 f(n)}{dn^2}$ is increasing for this range of n . Now, because

$$\left. \frac{d^2 f(n)}{dn^2} \right|_{n=4\kappa} = -\frac{3e-8}{48\kappa} < 0,$$

we have $\frac{d^2 f(n)}{dn^2} < 0$ for $\kappa \leq n \leq 4\kappa$, so that $\frac{df(n)}{dn}$ is decreasing for this range of n . Then because

$$\left. \frac{df(n)}{dn} \right|_{n=4\kappa} = \frac{3e-8}{6} > 0,$$

we have $\frac{df(n)}{dn} > 0$ for $\kappa \leq n \leq 4\kappa$, so that $f(n)$ is increasing for this range of n . Finally, because

$$f(\kappa) = \kappa \left(2e - 4 - \ln 4 - \frac{1}{24} \right) > 0,$$

it follows that $f(n) > 0$ for $\kappa \leq n \leq 4\kappa$.

Next, we analyze the more complicated case when $n \leq \kappa$. For this range of n , recall that u_n^{org} is $V_{n,C,\rho}/\lambda$, where $V_{n,C,\rho}$ is as defined in the proof of Proposition 1. Letting $\phi_n \equiv \kappa/n - 1$, we can rewrite $\mathbf{E}[W_n] \leq V_{n,C,\rho}/\lambda$ as follows:

$$\begin{aligned} \frac{\lambda \mathbf{E}[W_n]}{1-\rho} &\leq 2\kappa + n \ln(4e^2(\phi_n + 1)) + \frac{n}{24(\phi_n + 1)} \\ &\quad - 4\sqrt{n\kappa} \left(\frac{1}{2}\sqrt{\phi_n} - \arctan \sqrt{\phi_n} \right) \\ &\quad - 2n \ln \left(\sqrt{\phi_n} + \sqrt{\phi_n + 1} \right) \\ &\equiv v_n^{\text{org}}. \end{aligned}$$

Here, notice that the term, $n \ln(4e^2(\phi_n + 1))$, in v_n^{org} does not directly correspond to the term, $n \ln(e^2 C^2/4n)$, yielded from $V_{n,C,\rho}/(1-\rho)$. The former term can be expanded into $n \ln(e^2 C^2/4n) - 2n \ln(1-\rho)$, the second term of which is canceled out with a term that yields analogously from the term, $2n \ln(\sqrt{\phi_n} + \sqrt{\phi_n + 1})$, in v_n^{org} .

Similar to the previous case, in the hereafter we will show that v_n^{org} is at most

$$v_n^{\text{new}} \equiv \frac{\pi \sqrt{n(C_A^2 + C_S^2 \rho^2)}}{2(1-\rho)} = 2\pi \sqrt{n\kappa}$$

by proving the nonnegativity of $g(n) \equiv v_n^{\text{new}} - v_n^{\text{org}}$. The nonnegativity of $g(n)$ is derived from the following three observations: (i) $g(\kappa) \geq 0$, (ii) $g(1) \geq 0$, and (iii) $\frac{d^2 g(n)}{dn^2} \leq 0$. We give formal proofs below.

First, it is easy to notice that $g(\kappa) > 0$ since by substitution,

$$g(\kappa) = \kappa \left(2\pi - 4 - \ln 4 - \frac{1}{24} \right) > 0.$$

Second, to prove $g(1) \geq 0$, let $\phi_1 = \kappa - 1$. Then $g(1)$ can be considered as a function, $\bar{g}(\phi_1)$, of ϕ_1 as the following:

$$\begin{aligned} \bar{g}(\phi_1) &= 2\pi\sqrt{\phi_1 + 1} - 2(\phi_1 + 1) - \ln 4 - 2 - \ln(\phi_1 + 1) \\ &\quad - \frac{1}{24(\phi_1 + 1)} + 2\sqrt{\phi_1(\phi_1 + 1)} \\ &\quad - 4\sqrt{\phi_1 + 1} \arctan \sqrt{\phi_1} + 2\ln \left(\sqrt{\phi_1} + \sqrt{\phi_1 + 1} \right). \end{aligned}$$

Notice that ϕ_1 can take a value in $[0, \infty)$, because $\kappa \geq 1$ for the case under consideration. Hence, it suffices to show that $\bar{g}(0) \geq 0$ and $\frac{d\bar{g}(\phi_1)}{d\phi_1} \geq 0$. We can observe by substitution that

$$\bar{g}(0) = 2\pi - 4 - \ln 4 - 1/24 > 0.$$

To show $\frac{d\bar{g}(\phi_1)}{d\phi_1} \geq 0$, observe that

$$\begin{aligned} \frac{d\bar{g}(\phi_1)}{d\phi_1} &= \frac{\pi - 2 \arctan \sqrt{\phi_1}}{\sqrt{\phi_1 + 1}} - 2 \left(1 - \sqrt{\frac{\phi_1}{\phi_1 + 1}} \right) \\ &\quad - \frac{1}{\phi_1 + 1} \left(1 - \frac{1}{24(\phi_1 + 1)} \right), \end{aligned}$$

so that $\frac{d\bar{g}(\phi_1)}{d\phi_1} \rightarrow 0$ as $\phi_1 \rightarrow \infty$. Hence, $\frac{d\bar{g}(\phi_1)}{d\phi_1}$ is nonnegative if it is non-increasing for $\phi_1 \geq 0$. In fact, $\frac{d\bar{g}(\phi_1)}{d\phi_1}$ is non-increasing, because

$$\begin{aligned} &\frac{d \left(\sqrt{\phi_1 + 1} \frac{d\bar{g}(\phi_1)}{d\phi_1} \right)}{d\phi_1} \\ &= \frac{\sqrt{\phi_1(\phi_1 + 1)} - (\phi_1 + \frac{1}{2})}{(\phi_1 + 1)^{3/2}} - \frac{1}{16(\phi_1 + 1)^{5/2}} < 0. \end{aligned}$$

where the last inequality holds, because $\phi_1 \geq 0$ and

$$\left(\sqrt{\phi_1(\phi_1 + 1)} \right)^2 - \left(\phi_1 + \frac{1}{2} \right)^2 = -\frac{1}{4} < 0.$$

Third, we prove $\frac{d^2 g(n)}{dn^2} \leq 0$. We start by computing $\frac{dg(n)}{dn}$ as follows:

$$\begin{aligned} \frac{dg(n)}{dn} &= 2\sqrt{\phi_n + 1} \left(\frac{\pi}{2} - \arctan \sqrt{\phi_n} \right) \\ &\quad - 1 - \ln 4 - \ln(\phi_n + 1) - \frac{1}{12(\phi_n + 1)} \\ &\quad + 2 \ln \left(\sqrt{\phi_n} + \sqrt{\phi_n + 1} \right) \\ &\equiv h(\phi_n), \end{aligned}$$

where recall that $\phi_n = \kappa/n - 1$, and notice that $h(\phi_n)$ is a function of ϕ_n . By the chain rule, we have

$$\frac{d^2 g(n)}{dn^2} = \frac{dh(\phi_n)}{dn} = \frac{d\phi_n}{dn} \frac{dh(\phi_n)}{d\phi_n}.$$

Because $\frac{d\phi_n}{dn} < 0$, we have $\frac{d^2 g(n)}{dn^2} \leq 0$ iff $\frac{dh(\phi_n)}{d\phi_n} \geq 0$. Now, observe that

$$\begin{aligned} \frac{dh(\phi_n)}{d\phi_n} &= \frac{\pi/2 - \arctan \sqrt{\phi_n}}{\sqrt{\phi_n + 1}} + \frac{1}{12(\phi_n + 1)^2} - \frac{1}{(\phi_n + 1)} \\ &\geq 0, \end{aligned}$$

where the last equality holds because $\lim_{\phi_n \rightarrow \infty} \frac{dh(\phi_n)}{d\phi_n} = 0$, and $\frac{dh(\phi_n)}{d\phi_n}$ is non-increasing as can be shown from

$$\frac{d \left(\sqrt{\phi_n + 1} \frac{dh(\phi_n)}{d\phi_n} \right)}{d\phi_n} = \frac{\sqrt{\phi_n}(3 + 4\phi_n) - 4(1 + \phi_n)^{\frac{3}{2}}}{8\sqrt{\phi_n}(1 + \phi_n)^{\frac{5}{2}}} < 0,$$

where the last equality follows from

$$\begin{aligned} &\left(\sqrt{\phi_n}(3 + 4\phi_n) \right)^2 - \left(4(1 + \phi_n)^{\frac{3}{2}} \right)^2 \\ &= -(16 + 39\phi_n + 24\phi_n^2) < 0. \end{aligned}$$

B. Proof outline of Corollary 1

Our analysis of $\mathbf{E}[W_n]$ hinges on an analysis of a random walk, $T_i = X_1 + \dots + X_i$ for $i \geq 1$ and $T_0 = 0$, where X_1, \dots, X_i are independent and their first two moments are respectively identical. We study the bound on $\pi_{b,n}$, the probability that T_i crosses b in n steps. We say that T_i crosses b , when $T_{i-1} < b \leq T_i$. Let $x_1 = \mathbf{E}[X_j]$ and $x_2 = \mathbf{E}[X_j^2]$ for any j . We define $\sigma_X^2 \equiv x_2 - x_1^2 > 0$ and $\eta \equiv 2x_1(n x_1 - b)/\sigma_X^2$.

Our proof is organized as follows. We will first derive equalities that hold among the moments of the measures that we will define below. The problem of finding an upper bound on $\pi_{b,n}$, will then be formulated as a semidefinite programming (SDP), a primal problem. We will then formulate the dual problem of the SDP, so that a feasible solution to the dual problem provides an upper bound on the solution to the primal problem. Finally, we will construct a feasible solution to the dual problem. The upper bound on $\pi_{b,n}$ can be translated into an upper bound on the tail probability of W_n , which will then be integrated into an upper bound on $\mathbf{E}[W_n]$.

We first derive equalities that hold among the moments of some measures. Let $\Psi_{p,q} \equiv \{0, \dots, p\} \times \{0, \dots, q\} \setminus \{(p, q)\}$. Then, for any $1 \leq i \leq n$, we have

$$\begin{aligned} i^p T_i^q &= ((i-1)+1)^p (T_{i-1} + X_i)^q \\ &= \sum_{(k,\ell) \in \Psi_{p,q}} \binom{p}{k} \binom{q}{\ell} (i-1)^k X_i^{q-\ell} T_{i-1}^\ell + (i-1)^p T_{i-1}^q \\ &= 0^{p+q} + \sum_{j=0}^{i-1} \sum_{(k,\ell) \in \Psi_{p,q}} \binom{p}{k} \binom{q}{\ell} j^k X_{j-1}^{q-\ell} T_j^\ell, \end{aligned} \quad (16)$$

where the last equality follows from taking a telescoping sum. Let $N \equiv \min\{i \mid T_i \geq b \text{ or } i = n\}$ be a stopping time, so that $\Pr(T_N \geq b) = \pi_{b,n}$. Substituting $i = N$ into (16) and taking expectations, we can derive, for $p = 0, 1, 2$ and $q = 0, 1, 2$,

$$\mathbf{E}[N^p T_N^q] = 0^{p+q} + \sum_{(k,\ell) \in \Psi_{p,q}} \binom{p}{k} \binom{q}{\ell} x_{q-\ell} \mathbf{E}\left[\sum_{j=0}^{N-1} j^k T_j^\ell\right]. \quad (17)$$

More details about the derivation of (17) are provided in Section 2.2.2 from [17]. For $0 \leq m < n$ and $y < b$, we define $\mu_c(m, y) \equiv \mathbf{E}[\sum_{i=0}^{N-1} \mathbf{1}\{i \leq m \text{ \& } T_i \leq y\}]$ be the expected time that (i, T_i) spends in region $[0, m] \times (-\infty, y]$ before T_i crosses b or i reaches n . For $0 \leq m \leq n$ and $-\infty < y < \infty$, let $\mu_s(m, y) \equiv \Pr(N \leq m \text{ \& } T_N \leq y)$ be the probability that (N, T_N) hits a point in $[0, m] \times (-\infty, y]$. For $k = 0, 1, 2$ and $\ell = 0, 1, 2$, we define the following moments:

$$\begin{aligned} m_{k,\ell}^{(c)} &\equiv \sum_{i=0}^{n-1} \int_{-\infty}^b (i/n)^k y^\ell d\mu_c(i, y), \quad m_\ell^{(r)} \equiv \int_{-\infty}^b y^\ell d\mu_s(n, y), \\ m_{k,\ell}^{(u)} &\equiv \sum_{i=0}^n \int_b^\infty (i/n)^k y^\ell d\mu_s(i, y). \end{aligned}$$

For $p = 0, 1, 2$ and $q = 0, 1, 2$, we can express the expected values appearing in (17) using the moments:

$$m_{p,q}^{(u)} + m_q^{(r)} = 0^{p+q} + \sum_{(k,\ell) \in \Psi_{p,q}} n^{k-p} \binom{p}{k} \binom{q}{\ell} x_{q-\ell} m_{k,\ell}^{(c)}. \quad (18)$$

Next, we consider the positive semidefinite (PSD) conditions that follow from the fact that $m_q^{(r)}$ (respectively, $m_q^{(c)}$, and $m_q^{(u)}$) for $q = 0, 1, 2$, are moments of some measures with given supports. Specifically, the following matrices must be PSD (see Section 2.2.2 from [17] for more details):

$$\begin{aligned} \hat{M}^{(r)} &\equiv \begin{pmatrix} \hat{m}_0^{(r)} & \hat{m}_1^{(r)} \\ \hat{m}_1^{(r)} & \hat{m}_2^{(r)} \end{pmatrix}, & L^{(r)} &\equiv b m_0^{(r)} - m_1^{(r)}, \\ M_2^{(c)} &\equiv \begin{pmatrix} m_{0,0}^{(c)} & m_{0,1}^{(c)} & m_{1,0}^{(c)} \\ m_{0,1}^{(c)} & m_{0,2}^{(c)} & m_{1,1}^{(c)} \\ m_{1,0}^{(c)} & m_{1,1}^{(c)} & m_{2,0}^{(c)} \end{pmatrix}, & L^{(c,x)} &\equiv \frac{n-1}{n} m_{1,0}^{(c)} - m_{2,0}^{(c)}, \\ M_2^{(u)} &\equiv \begin{pmatrix} m_{0,0}^{(u)} & m_{0,1}^{(u)} & m_{1,0}^{(u)} \\ m_{0,1}^{(u)} & m_{0,2}^{(u)} & m_{1,1}^{(u)} \\ m_{1,0}^{(u)} & m_{1,1}^{(u)} & m_{2,0}^{(u)} \end{pmatrix}, & L^{(c,y)} &\equiv b m_{0,0}^{(c)} - m_{0,1}^{(c)}, \\ & & L^{(u,x)} &\equiv m_{1,0}^{(u)} - m_{2,0}^{(u)}, \\ & & L^{(u,y)} &\equiv -b m_{0,0}^{(u)} + m_{0,1}^{(u)}. \end{aligned}$$

Now, observe that $m_0^{(r)} = \Pr(T_N < b)$, the probability that i reaches n before T_i crosses b . Therefore, the solution to the following SDP gives an upper bound on $\pi_{b,n}$:

$$\begin{aligned} \max. \quad & 1 - m_0^{(r)} \\ \text{s.t.} \quad & M^{(c)}, M^{(r)}, M^{(u)}, L^{(c,x)}, L^{(c,y)}, L^{(r)}, L^{(u,x)}, L^{(u,y)} \succeq 0, \end{aligned} \quad (19)$$

where we have eliminated $m_{p,q}^{(u)}$ from the expressions of $M^{(u)}$, $L^{(u,x)}$, and $L^{(u,y)}$ using Equation (18).

The following lemma shows an upper bound on the objective value, $1 - m_0^{(r)}$, of the optimal solution to (19).

Lemma 1: Consider the $1 - m_0^{(r)}$ of any feasible solution to (19). Let $\sigma_X^2 \equiv x_2 - x_1^2 > 0$. If $x_1 < 0$, then

$$1 - m_0^{(r)} \leq \frac{1}{2} + \sqrt{\left(\frac{1}{2} + \frac{2x_1 \delta}{\sigma_X^2}\right)^2 + \frac{4bx_1}{\sigma_X^2} - \frac{2x_1 \delta}{\sigma_X^2}}, \quad (20)$$

where $\delta \equiv nx_1 - b$. If $x_1 < 0$ and $\sigma_X^2 \leq 4nx_1^2 - 2bx_1$, then

$$1 - m_0^{(r)} \leq 1 - \frac{b^2}{2n\sigma_X^2} \left(\sqrt{1 + \frac{4n\sigma_X^2}{b^2}} - 1 \right). \quad (21)$$

Lemma 1 can be proved by formulating the dual problem of (19) and constructing a feasible solution of the dual problem (see Appendix A.1 from [17]).

The expression of an upper bound on $1 - m_0^{(r)}$ can be used to derive the bounds on the tail probability of W_n . Let $X_i = S_i - A_i$, where A_i is the interarrival time between the i -th job and the $(i+1)$ -st job, and S_i is the service time of the i -th job for $i = 0, 1, \dots$. We assume that the 0-th job arrives when the system is idle (i.e., its delay is $W_0 = 0$). Then we have $\Pr(W_n > b) = 1 - m_0^{(r)}$ (see Section 4.5 from [14]).

We can integrate the bound on $\Pr(W_n > b)$ to obtain the bound on $\mathbf{E}[W_n]$. Specifically, we obtain (5) by integrating (20) from $b = 0$ to $b = \beta \equiv -\sigma_X^2/(2x_1) + 2nx_1$ and (21) from $b = \beta$ to $b = \infty$. We obtain (6) in the same way as we obtain (5) except that we integrate (21), instead of (20), from $b = b_1$ to $b = b_2$, where b_1 and b_2 are the solutions of the quadratic equation specified in the proof. We show these integrations more formally in Appendix B of [17].