

Optimal rate for a queueing system in heavy traffic with superimposed On-Off arrivals.

Arka P. Ghosh
Iowa State University

April 22, 2012

Abstract

A rate control problem is addressed for a queueing system in heavy traffic. The arrival process is stationary heavy-tailed On-Off process and service is done at a constant-rate (controlled). With an infinite horizon discounted cost function, the main result shows the existence of an optimal rate and specifies a bound on this optimal rate. As a part of the analysis, we solve an approximating control problem driven by fractional Brownian motion. We also derive an asymptotic maximal bound on the second moment of the centered On-Off process, which is a key ingredient of the proof and is of independent interest.

MSC2000: primary 60K25, 68M20, 90B22; secondary 60G22, 90B18.

Keywords: stochastic control, controlled queueing networks, heavy traffic analysis, On-Off process, fractional Brownian motion, self-similarity, fractional Brownian control problem

1 Introduction

Self-similarity and long-range dependence are two common and important features in data arising from modern high-speed communication networks such as local area networks (LAN) etc. In an attempt to find a suitable process that captures these properties, Willinger, Taqqu, Sherman and Wilson in [25] (as well as in [24]) proposed strictly alternating On-Off process and established that a superimposition of a large number of such processes, when suitably scaled, exhibit these features in the limit. The idea here is that each user interacts with the server in alternating activity periods (On-periods) and idle periods (Off-periods) – during the On-period data/jobs are sent to the server at a constant rate and no data/jobs are sent during the Off-periods. When a large number of users interact with the server in this fashion, their cumulative behavior is modeled sufficiently well by the superimposed On-Off process with heavy-tailed On-Off periods. Figure 1 taken from [25] compares real ethernet traffic data with simulated data from traditional queueing models as well as from the On-Off process and shows that the On-Off process exhibits features similar the real internet traffic data. Since their introduction, this and similar processes have been studied extensively in

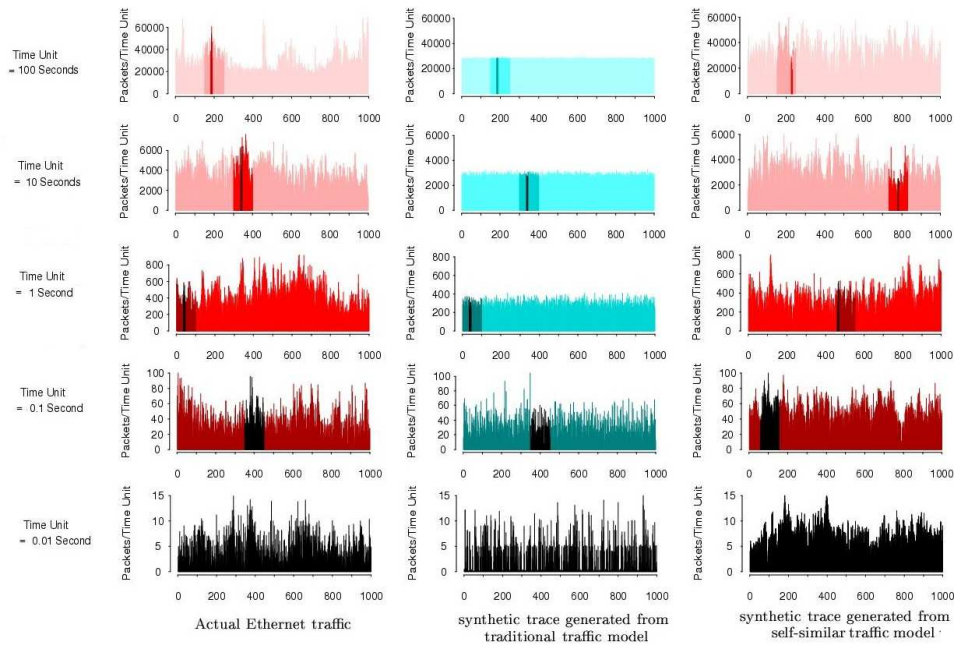


Figure 1: Taken from [25], this figure shows actual Ethernet traffic (left column), synthetic trace generated from an traditional traffic model (middle column), and synthetic trace generated from self-similar traffic model using On/Off process (right column) – on five different time scales.

the literature [7, 12, 19, 20, 25, 29, 28]). Researchers have also analyzed various aspects of queueing models associated with On-Off processes [9, 10, 13, 25, 26, 27, 4].

A natural question that comes after analysis of queueing models is *control*: how does one optimize the performance of the queueing system to ensure minimum operating cost (or maximum profit). Whereas On-Off process and associated queueing models have been studied sufficiently, the issue of control of such queueing systems has been largely missing in the literature (some works addressed control of similar models driven by fractional Brownian motion [6, 11]). One major reason for this is the non-Markovian nature of the On-Off process and fractional Brownian motion (fBm) makes these problems quite difficult to study using classical stochastic control techniques (e.g. dynamic programming). Recently in [6], using properties of fBm and convexity analysis, simple control problems of an approximate queueing model driven by fBm have been solved, but queueing control problems involving On-Off processes have not been addressed so far.

In this paper, we solve a rate control problem associated with a simple queueing system in heavy traffic where the arrivals follow the popular superimposed On-Off process and the service takes place at a constant rate which is adjustable. The controller/manager tries to find the optimal value for this service rate which minimizes a infinite horizon discounted cost function. Our main result, Theorem 2.4, guarantees the existence of an optimal rate as well as specifies an explicit bound on the possible values of this optimal rate. This explicit bound is important in real applications where one may have to do further numerical analysis to obtain the exact value of the optimal rate. Another key result in the paper (which is used

in the proof of the main result and is a result of independent interest) is Theorem 5.1 which provides an asymptotic maximal bound on the second moment of the centered cumulative On-Off process.

The analysis in this paper follows the usual steps of heavy traffic analysis for queueing control problems. As outlined by Harrison in [8], these steps are: formulating an approximating control problem (which usually is driven by Brownian motions, and is called the Brownian control problem or BCP), solving the approximating control problem, interpreting this solution to a meaningful control for the queueing network and establishing the optimality of this control for the queueing control problem. In our setup, since the arrivals are given by alternating On-Off process with strictly heavy-tailed On-periods, the associated approximating control problem is driven by a fractional Brownian motion (hence, here we refer to this approximating control problem as *fractional*-Brownian control problem or f-BCP). In [6], a solution to a control problem similar to the f-BCP has been established. We extend this analysis to obtain a bound on the range of possible values of the solution of the f-BCP. Then, we interpret this solution to propose an admissible rate for the queueing system and using weak convergence techniques, prove the optimality of the proposed rate for the queueing control problem.

In the proof of the main result, we also used a key maximal bound (stated as Theorem 5.1) on the second moment of the centered cumulative On-Off process. Theorem 5.1 states that asymptotically the supremum (over the time interval $[0, T]$) of the second moment of the centered On-Off process will be of the order $T^{3-\alpha}$, where distribution of On-periods is slowly varying with index α (Off-period distribution has lighter-tails than α). This result itself is not surprising given similar results on related quantities in the literature (cf. [16, 10]) but we believe, it does not follow from any of the known results. Moreover, since it is for any On-Off process, this bound is of independent interest for the understanding of this popular process. In this paper, this bound is used to obtain a uniform integrability of some processes to show the convergence of cost functions in the weak convergence analysis. The proof of this bound given in this paper is from the first principles and somewhat involved, and hence given separately at the end of the paper.

The paper is organized as follows. We begin by describing the model in Section 2. This section also describes the relevant scaling, the heavy traffic assumption, the cost functional and the associated control problem. The main result of the paper, Theorem 2.4, is stated at the end of this section. The approximating limit control problem (f-BCP) driven by the fractional Brownian motion is formulated and solved (Theorem 3.1) in Section 3. The proof of the main result (Theorem 2.4) involving weak convergence analysis is presented in Section 4. This section also briefly discusses the reflection maps that are used in the proof. The proof of Theorem 5.1, which provides a key asymptotic maximal bound on the second moment of a centered On-Off process and a result of independent interest, is presented separately in Section 5.

2 Model and the Main Result

All the random variables and processes in the paper are assumed to be defined on a common probability space (Ω, \mathcal{F}, P) and the expectation under $P(\cdot)$ will be denoted by $E(\cdot)$.

The abbreviation ‘‘iid’’ will be used to denote independent and identically distributed (random variables or processes). Throughout this paper, we will use the following commonly used (c.f. [30]) notation: for any two real functions $f(\cdot)$ and $g(\cdot)$, $f(u) \sim g(u)$ would denote $\lim_{u \rightarrow \infty} f(u)/g(u) = 1$. Further, we will write $f(u) \lesssim g(u)$ to indicate that $\limsup_{u \rightarrow \infty} \frac{f(u)}{g(u)} \lesssim 1$.

2.1 Queueing System

We begin with the description of a sequence (indexed by $r \geq 1$) of queueing systems where each queueing system in the sequence has $n \geq 1$ users (n, r are integers). Figure 2 describes the r -th queueing system with n users. This model is related to the internet traffic data generated by superimposition of the service demands of multiple users as described earlier (for more detail on this model, we refer to [25] and [27, Chapters 7 and 8]). The input (of

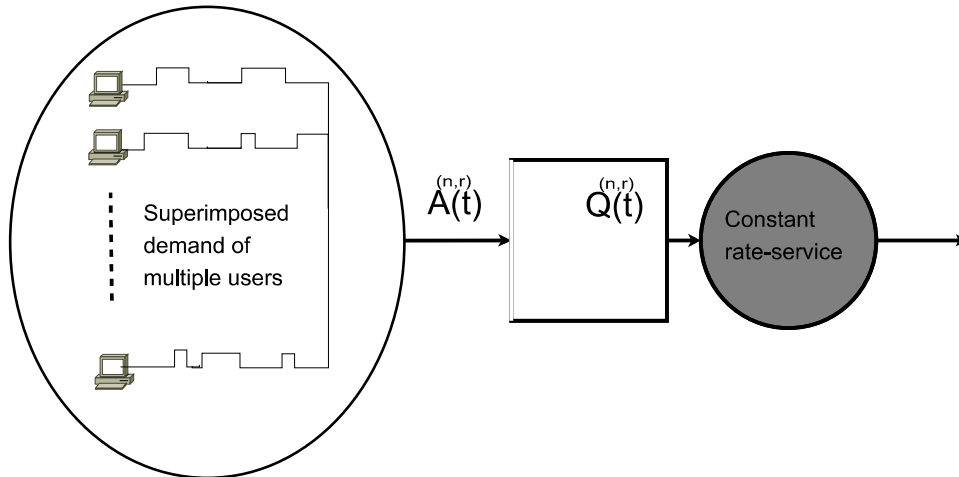


Figure 2: A typical queueing system at time t , with On-Off arrivals and constant rate service.

jobs) to the system is generated by aggregated demands of all the users in the system. Each user sends jobs to the server (represented by the shaded circle) for processing and these jobs get accumulated in the queue (represented by the open rectangle). The server completes these jobs at a constant rate $\mu^{(n,r)}$, when the queue is non-empty. At time $t \geq 0$, the total number of jobs in the queue is denoted by $Q^{(n,r)}(t)$ and the cumulative number of jobs sent to the server (up to time t) is given by $A^{(n,r)}(t)$.

Each user’s behavior is modeled by a stationary On-Off process (c.f. [25, 23]): that is, each user sends jobs to the server at a unit rate during the On intervals followed by an Off interval during which the user does not send any jobs to the server. These successive On and Off intervals are modeled by two independent set of iid non-negative random variables with distribution $F_{on}(\cdot)$ and $F_{off}(\cdot)$ respectively. In addition, we assume the following tail behavior of the distribution functions:

$$\bar{F}_{on}(x) \doteq 1 - F_{on}(x) \sim x^{-\alpha}, \quad \bar{F}_{off}(x) \doteq 1 - F_{off}(x) \sim x^{-\beta}, \quad (2.1)$$

where $1 < \alpha < 2$ and $\beta > \alpha$. That is, the Off-period distribution has lighter-tails than the distribution of the On-periods as in [23] (i.e. $\bar{F}_{off}(x) = o(\bar{F}_{on}(x))$). This, in particular,

means both the On- and Off-periods have finite mean (but On periods have infinite variance). We denote

$$\mu_{on} \doteq EX_{on}, \mu_{off} \doteq EX_{off}, \mu \doteq \mu_{on} + \mu_{off}, \lambda \doteq \frac{\mu_{on}}{\mu}. \quad (2.2)$$

The On-Off demand process generated by the i -th user ($1 \leq i \leq n$) in the r -th system is given by $W_i^{(n,r)}(\cdot)$, where $W_i^{(n,r)}(t) = 1$ if t is during an On-period and 0 otherwise. Section 5 describes the On-Off process of one user in more detail (see also [25, 23, 27]). We assume that the collection $\{W_i^{(n,r)}(\cdot), i = 1, \dots, n; n \geq 1, r \geq 1\}$ is iid. In terms of these individual On-Off processes, the cumulative arrival process $A^{(n,r)}(\cdot)$ is given as follows: for $t \geq 0$,

$$A^{(n,r)}(t) = \sum_{i=1}^n \int_0^t W_i^{(n,r)}(s) ds. \quad (2.3)$$

This represents cumulative number of jobs/packets sent to the server by all the n customers in the interval $[0, t]$.

If the initial queue-length (at time $t = 0$) is $q_0^{(n,r)}$, then queue-length equation at time $t \geq 0$ can be represented as

$$Q^{(n,r)}(t) = q_0^{(n,r)} + A^{(n,r)}(t) - \mu^{(n,r)}t + L^{(n,r)}(t), \quad (2.4)$$

where $L^{(n,r)}(\cdot)$ is a non-decreasing process that starts from 0, increases only when $Q^{(n,r)}(\cdot)$ is zero, and ensures that $Q^{(n,r)}(\cdot)$ is always non-negative. This “reflection” process has the following representation: for $t \geq 0$

$$L^{(n,r)}(t) = \mu^{(n,r)} \int_0^t \mathbf{1}_{\{Q^{(n,r)}(s)=0\}} ds,$$

where $\mathbf{1}_{\{B\}}$ denotes the indicator function of a set B . The integral in the definition of the above process represents cumulative amount of idleness in the system up to time $t \geq 0$ (so, $L^{(n,r)}(t)$ can be thought of as the average number of jobs that could not be served because of the empty queue). Physically, this implies that the server is non-idling, i.e. it serves jobs continuously as long as the buffer is non-empty. Mathematically, this means

$$\int_0^\infty Q^{(n,r)}(t) dL^{(n,r)}(t) = 0.$$

2.2 Scaling and Heavy Traffic

In this setup, $r > 1$ represents the “heavy traffic” scaling parameter, and it is well known (see [24] or Theorem 7.2.5 of [27]) that when time is scaled by r and space is divided by r^H and also by \sqrt{n} , then the queue length process in (2.4) stabilizes. The main reason for this stability is the fact that

$$r^{-H} n^{-\frac{1}{2}} \sum_{i=1}^n \int_0^{rt} (W_i^{(n,r)}(s) - \lambda) ds \Rightarrow W_H(\cdot),$$

when $n \rightarrow \infty$ first and then $r \rightarrow \infty$ (see [27] for details). Here W_H is a fractional Brownian motion (fBm) with Hurst parameter

$$H = \frac{3 - \alpha}{2} \in \left(\frac{1}{2}, 1\right). \quad (2.5)$$

A more precise description of the fBm is given in Section 3. Note that scaling up time by r and dividing space by r^H is the usual scaling for heavy traffic analysis and when $\alpha = 2$ (which corresponds to $H = \frac{1}{2}$) this scaling reduces to the more common central limit theorem-type scaling. In addition, since the combined arrival rates of n users is of the order n , we divide the space by \sqrt{n} . Hence, in our heavy traffic formulation of the queueing control problem, all the processes considered will be scaled this way. More precisely, we will consider the following scaled queue length and reflection processes: for $t \geq 0$,

$$\widehat{Q}^{(n,r)}(t) \doteq \frac{Q^{(n,r)}(rt)}{r^H \sqrt{n}}, \quad \widehat{L}^{(n,r)}(t) \doteq \frac{L^{(n,r)}(rt)}{r^H \sqrt{n}}. \quad (2.6)$$

We also assume that the initial queue-lengths (scaled) converge to a limit:

Assumption 2.1. *There exists a $q_0 \geq 0$ such that $\widehat{q}_0^{(n,r)}(t) \doteq \frac{q_0^{(n,r)}}{r^H \sqrt{n}} \rightarrow q_0$ as $(n, r) \rightarrow \infty$.*

Now we state the heavy traffic assumption. The rate of arrival of a *single* On-Off process is λ defined in (2.2) (see [10], also see (5.5)). hence, that the effective rate of the system of scaled queue-lengths is given by $\widehat{u}^{(n,r)}$, where

$$\widehat{u}^{(n,r)} \doteq \frac{\mu^{(n,r)} r - \lambda n r}{r^H \sqrt{n}}. \quad (2.7)$$

Heavy traffic assumption says that this effective rate converges to a limit.

Assumption 2.2. [Heavy Traffic Assumption] *For some $u \geq 0$ and $\widehat{u}^{(n,r)}$ defined in (2.7) is positive and the following holds: $\widehat{u}^{(n,r)} \rightarrow u$ as $(n, r) \rightarrow \infty$.*

Note that Assumption 2.2 implies that $\mu^{(n,r)} \geq \lambda n$ (queue is stable) and using the fact that $H < 1$, one gets

$$\frac{\mu^{(n,r)}}{n} \rightarrow \lambda.$$

That is, asymptotically, fraction of the service rate devoted to the jobs from an individual user is equal to its demand rate of the user, on an average. That is why we call this the heavy traffic assumption.

For later use, we also define the scaled and centered arrival process

$$\widehat{A}^{(n,r)}(t) \doteq \frac{A^{(n,r)}(rt) - \lambda nrt}{r^H \sqrt{n}}, \quad \text{for } t \geq 0. \quad (2.8)$$

Note the scaled queue-length and reflection processes in (2.6) were not centered, since under the heavy traffic assumption, the average queue-length and idleness process (proportional to the reflection process) will be asymptotically zero. Finally, note that from (2.4), (2.6), (2.7) and (2.8), one gets the following representation connecting all the scaled processes:

$$\widehat{Q}^{(n,r)}(t) = \widehat{q}_0^{(n,r)} + \widehat{A}^{(n,r)}(t) - \widehat{u}^{(n,r)}t + \widehat{L}^{(n,r)}(t), \quad \text{for } t \geq 0, \quad (2.9)$$

which will be useful for our analysis in the rest of the paper.

2.3 Queueing Control Problem and the Main Result

Now we describe the control problem associated with the queueing model. Suppose that there is a system manager, who can choose the service rate $\mu^{(n,r)}$ to deal with the incoming demand. In this heavy traffic formulation, he/she can choose any sequence of non-negative numbers $\{\mu^{(n,r)}\} = \{\mu^{(n,r)} : n \geq 1, r \geq 1\}$ for the service rate as long as Assumption 2.2 is satisfied. Any such sequence will be called an *admissible* rate control sequence.

The control problem is to find the optimal sequence of rates that minimizes the following infinite horizon discounted cost functional

$$\widehat{J}(\{q_0^{(n,r)}\}, \{\mu_0^{(n,r)}\}) \doteq \liminf_{r \rightarrow \infty} \liminf_{n \rightarrow \infty} E \left(\int_0^\infty e^{-\theta t} [h(\widehat{u}^{(n,r)}) + c \widehat{Q}^{(n,r)}(t)] dt + p \int_0^\infty e^{-\theta t} d\widehat{L}^{(n,r)}(t) \right). \quad (2.10)$$

Here the \liminf 's follow the same order in which the queueing system stabilizes (as mentioned earlier in Section 2.2). The initial values $\{q_0^{(n,r)}\}$ are assumed to satisfy Assumption 2.1 with some $q_0 \geq 0$ and $\theta > 0$ is the discount factor. The constants $c > 0$, $p > 0$ represent the (linear) rates of holding cost per job waiting in the queue and idleness cost, respectively. The function $h(\cdot)$ captures the cost of exercising controls and is assumed to satisfy the following.

Assumption 2.3. *The control cost function h is a non-decreasing function on $[0, \infty)$ and is continuous and convex with $h(0) \geq 0$ and $\lim_{u \rightarrow +\infty} h(u) = +\infty$.*

The queueing control problem is to minimize the above cost functional among all the admissible rates. Before we state our main result, we introduce the following common notation for the Gamma function: for $a \geq 0$, $\text{Gamma}(a) \doteq \int_0^\infty e^{-x} x^{a-1} dx$. Our main result guarantees existence of one such optimal service rate (sequence) as well as provides a convenient bound on the possible values of the optimal rate.

Theorem 2.4 (Main Result: Optimal Service rate). *Let \bar{u} be the unique solution of the equation:*

$$h(\bar{u}) + p\bar{u} = h(0) + 2(c + \theta p) \left(q_0 + \frac{2^{\frac{3}{2}} \text{Gamma}(H + 1)}{\sqrt{\pi} \theta^H} \right), \quad \bar{u} \geq 0, \quad (2.11)$$

where H is as defined in (2.5). Then, there exists some u^* in $[0, \bar{u}]$ such that for $\{\mu^{*(n,r)}\}$ defined as

$$\mu^{*(n,r)} = n\lambda + r^{-(1-H)} \sqrt{n} u^* = n \left(\lambda + \frac{u^*}{\sqrt{nr^{(1-H)}}} \right), \quad n \geq 1, r \geq 1, \quad (2.12)$$

is optimal for the queueing control problem, that is, minimizes the cost (2.10) among all other admissible rates $\{\mu^{(n,r)}\}$.

In particular, this result specifies the optimal rate of deviation from this cumulative rate of arrival from all the users, $n\lambda$. Also note that when we have a specific form for the control cost function $h(\cdot)$, then one can easily get an explicit formula for \bar{u} in (2.11), which can be used to numerically find the optimal rate u^* , and in turn, $\{\mu^{(n,r)}\}$ for the queueing control problem. For example, in the linear control cost case: $h(u) = h_0 + h_1 u$, one gets $\bar{u} = 2(c + \theta p)K / (h_1 + p)$ where $K = (q_0 + 2^{\frac{3}{2}} \text{Gamma}(H + 1) / \sqrt{\pi} \theta^H)$ and in the quadratic control cost case it is given by the larger of the two solutions of a simple quadratic equation.

3 Fractional Brownian Control Problem (f-BCP):

In this section, we describe another control problem driven by a fractional Brownian motion (fBm). It is shown in Theorem 8.7.1 of [27] (see also [25]) that, under our assumptions on the control variable $\{\mu^{(n,r)}\}$, if one takes the limit as $n \rightarrow \infty$ first and then $r \rightarrow \infty$, then the scaled queue-lengths $\widehat{Q}^{(n,r)}$ in (2.9) converges weakly to a process Q satisfying:

$$Q(t) = x - ut + \sigma_H W_H(t) + L(t), \quad t \geq 0, \quad (3.1)$$

where the process L has continuous paths, and it increases at times when $Q(t) = 0$. We will verify this fact separately in Section 4 as well. Here

$$\sigma_H = \frac{2\alpha \mu_{off}}{\mu^3 \text{Gamma}(4 - \alpha)} \quad (3.2)$$

is a constant not depending on the service rate sequence $\{\mu^{(n,r)}\}$ (see Theorem 8.7.1 in [27]).

Here, by fractional Brownian motion (fBm) with Hurst parameter $H \in (0, 1)$, we mean a real-valued stochastic process $W_H = (W_H(t))_{t \geq 0}$ that has $W_H(0) = 0$ and is a continuous zero-mean Gaussian process with stationary increments and covariance function given by

$$\text{Cov}(W_H(s), W_H(t)) = \frac{1}{2} [t^{2H} + s^{2H} - |t - s|^{2H}], \quad s \geq 0, t \geq 0.$$

The fBm is a self-similar process with index H , that is for any $a > 0$ the process $\frac{1}{a^H}(W_H(at))_{t \geq 0}$ has the same distribution as $(W_H(t))_{t \geq 0}$. If $H = \frac{1}{2}$ then W_H is an ordinary Brownian motion, and if $H \in [\frac{1}{2}, 1)$ (which is the case relevant for this paper) then the increments of the process are positively correlated and the process exhibits long-range dependence. For additional properties and a more detailed description of this process we refer the readers to [15, 17, 18, 21, 22].

For the heavy traffic analysis required to solve the queueing control problem, one often follows a sequence of steps outlined by Harrison in [8] (see also [2]), which involves solving a limiting control problem that approximates (formally) the queueing control problem and then interpreting its solution to obtain meaningful control policies for the original queueing control problem. This approximating control problem, in usual heavy traffic analysis, is driven by a Brownian motion and is called the Brownian control problem or the BCP. In our case, the driving process is fBm W_H which is why we will refer to the approximating control problem as fractional Brownian control problem or the f-BCP. This control problem can be formulated as follows: for the state processes Q, L that satisfy (3.1), find optimal $u \geq 0$ which minimizes

$$J(q_0, u) \doteq E\left(\int_0^\infty e^{-\theta t} [h(u) + cQ(t)] dt + p \int_0^\infty e^{-\theta t} dL(t)\right). \quad (3.3)$$

$$= \frac{h(u)}{\theta} + E\left(\int_0^\infty e^{-\theta t} [cQ(t) + \theta p L(t)] dt\right). \quad (3.4)$$

for a fixed initial value $q_0 \geq 0$. To derive the last equality above we have used Fubini's theorem to obtain $\int_0^\infty e^{-\theta t} dL(t) = \theta \int_0^\infty e^{-\theta t} L(t) dt$.

The following result is the main result of this section, which guarantees existence of one such u^* that solves the f-BCP and also provides a bound on the possible values of u^* .

Theorem 3.1. [Solution of the f-BCP] *There exists a solution $u^* \geq 0$ for the f-BCP, which also satisfies*

$$0 \leq u^* \leq \bar{u},$$

where \bar{u} is a solution of (2.11) in Theorem 2.4.

Proof. The existence of optimal u^* follows from Theorem 5.1 of [6]. In particular, it is shown in Section 5 of [6] that $J(q_0, u)$ is a convex function of u for each initial value q_0 and $J(q_0, 0) < \infty$ and $J(q_0, u) \rightarrow \infty$ when $u \rightarrow \infty$. Hence, there is a minimizer u^* for each choice of the initial value q_0 . This optimal solution need not be unique and can possibly be zero, depending on properties of the function h . The proofs in this paper was for $\sigma_H \equiv 1$, but still apply in our setting with constant σ_H (as in (3.2), not depending on u) with minimal change.

For the second part (i.e. the upper bound) of the theorem, note that from (3.1), one gets that

$$E(L(t)) = E(Q(t)) + ut - q_0, \quad (3.5)$$

using the fact that $E(W_H(t)) = 0$. Also note that from the self similarity of W_H , it can be shown that

$$0 \leq E(Q(t)) \leq 2(q_0 + K_1 t^H), \quad t \geq 0, \quad (3.6)$$

where $K_1 = E(\sup_{0 \leq s \leq 1} |W_H(s)|)$ (see display (3.2) and preceding calculations in [6, p. 190] or [18, p. 296]). Hence, using (3.4) and (3.5) one gets the following alternative representation of the cost functional

$$J(q_0, u) = \frac{1}{\theta} (h(u) + pu) - pq_0 + (c + \theta p)g(u), \quad (3.7)$$

where

$$0 \leq g(u) = \int_0^\infty e^{-\theta t} E(Q(t)) dt.$$

Note that the function $g(\cdot)$ depends on q_0 as well, but this dependence is suppressed in the notation. Now, from (3.6) and the upper bound on the supremum of a fBm given in Theorem 1.1 (ii) of [3] (with $T = 1, H > \frac{1}{2}, \gamma = 1$ in that result), one gets the following bound

$$\begin{aligned} g(u) &\leq \int_0^\infty e^{-\theta t} 2(q_0 + K_1 t^H) dt = \frac{2}{\theta} \left(q_0 + \frac{K_1 \text{Gamma}(H+1)}{\theta^H} \right) \\ &\leq \frac{2}{\theta} \left(q_0 + \frac{2^{\frac{3}{2}} \text{Gamma}(H+1)}{\sqrt{\pi} \theta^H} \right) =: K_2. \end{aligned} \quad (3.8)$$

Hence, we have $J(q_0, u)$ a convex function of u (shown in [6], as mentioned above) and, from

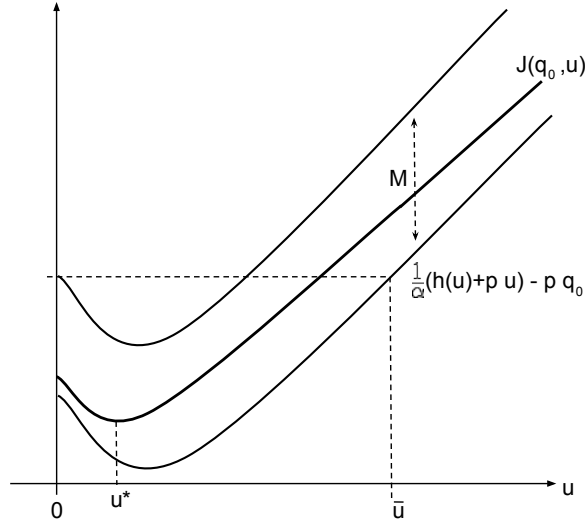


Figure 3: Cost function of f-BCP and the upper and lower convex “envelops”

(3.7) – (3.8) and with $M = (c + \theta p)K_2$, satisfies :

$$\frac{1}{\theta} (h(u) + pu) - pq_0 \leq J(q_0, u) \leq \frac{1}{\theta} (h(u) + pu) - pq_0 + M.$$

This clearly implies, as shown in Figure 3 (the two “envelops” are the two convex functions in the left and right sides of the two inequalities above), we get that the optimal solution u^* lies within $[0, \bar{u}]$ where \bar{u} satisfies

$$\frac{1}{\theta} (h(\bar{u}) + p\bar{u}) - pq_0 = \frac{1}{\theta} (h(0) + p \cdot 0) - pq_0 + M.$$

Upon canceling terms, one gets the fact that \bar{u} satisfies equation (2.11) in Theorem 2.4. This completes the proof of this theorem. \blacksquare

4 Asymptotic Optimality of the Proposed Policy

In this section, we provide the proof of the main result of the paper, Theorem 2.4. A key ingredient of this proof is an asymptotic maximal bound on the second moment of a single On-Off process. This result (stated as Theorem 5.1) has a somewhat involved proof and hence is stated and proved separately in Section 5. Another important ingredient of the proof is the reflection map (also known as the regulator map or Skorohod map) and its properties described below.

4.1 The Reflection Map

Let $\mathcal{C}([0, \infty), \mathbb{R})$ be the space of continuous functions with domain $[0, \infty)$. The standard reflection mapping $\Gamma : \mathcal{C}([0, \infty), \mathbb{R}) \rightarrow \mathcal{C}([0, \infty), \mathbb{R})$ is defined by

$$\Gamma(f)(t) = f(t) + \hat{\Gamma}(f)(t), \quad \text{where } \hat{\Gamma}(f)(t) = \sup_{s \in [0, t]} (-f(s))^+, t \geq 0, \quad (4.1)$$

for $f \in \mathcal{C}([0, \infty), \mathbb{R})$, and a^+ denotes $\max\{0, a\}$. For a detailed discussion of this map and its properties, we refer the readers to [14, 27]. The following two properties will be used in the proof of Theorem 2.4. It is well known (c.f. [27]) that for $f \in \mathcal{C}([0, \infty), \mathbb{R})$ and $T \geq 0$

$$\sup_{t \in [0, T]} |\Gamma(f)(t)| \leq 2 \sup_{t \in [0, T]} |f(t)|, \quad \sup_{t \in [0, T]} |\hat{\Gamma}(f)(t)| \leq 2 \sup_{t \in [0, T]} |f(t)|, \quad (4.2)$$

$$\text{and, both } \Gamma(\cdot) \text{ and } \hat{\Gamma}(\cdot) \text{ are continuous functions on } \mathcal{C}([0, \infty), \mathbb{R}). \quad (4.3)$$

More properties of these maps are used in the analysis in [6].

Proof of Theorem 2.4:

First note that the cost functional defined in (2.10) can alternatively written as

$$\hat{\mathcal{J}}(\{q_0^{(n,r)}\}, \{\mu_0^{(n,r)}\}) \doteq \liminf_{r \rightarrow \infty} \liminf_{n \rightarrow \infty} \left[\frac{h(\hat{u}^{(n,r)})}{\theta} + E \left(\int_0^\infty e^{-\theta t} \left\{ c \hat{Q}^{(n,r)}(t) + \theta p \hat{L}^{(n,r)}(t) \right\} dt \right) \right] \quad (4.4)$$

The proof involves Fubini’s theorem (see, for example, the proof of Lemma 4.3 of [5]).

Next, note that from the definition of the scaled-queue-length process $\widehat{Q}^{(n,r)}$ in (2.9) and of the state process Q in (3.1) in the approximating f-BCP as well as the definition of the reflection map above it follows that for all $t \geq 0$

$$\begin{aligned}\widehat{Q}^{(n,r)}(t) &= \Gamma \left(\widehat{q}_0^{(n,r)} - \widehat{u}^{(n,r)} e(\cdot) + \widehat{A}^{(n,r)}(\cdot) \right) (t), \quad \widehat{L}^{(n,r)}(t) = \widehat{\Gamma} \left(\widehat{q}_0^{(n,r)} - \widehat{u}^{(n,r)} e(\cdot) + \widehat{A}^{(n,r)}(\cdot) \right) (t), \\ Q(t) &= \Gamma (q_0 - ue(\cdot) + \sigma_H W_H(\cdot)) (t), \quad L(t) = \widehat{\Gamma} (q_0 - ue(\cdot) + \sigma_H W_H(\cdot)) (t),\end{aligned}\tag{4.5}$$

where $e(t) \equiv t, t \geq 0$ denotes the identity map. Now, from [24] (see also [25, 12, 27]) we know that the centered arrival process $\widehat{A}(\cdot)$ converges weakly to a fBm $\sigma_H W_H(\cdot)$. Also, from the heavy traffic assumption (Assumption 2.2), we get that the drift-term $\widehat{u}^{(n,r)}$ converges to some u and the initial queue-lengths converge. Hence, by the continuity property (4.3) of reflection maps, we get that

$$\widehat{Q}^{(n,r)}(\cdot) \Rightarrow Q(\cdot), \quad \widehat{L}^{(n,r)}(\cdot) \Rightarrow L(\cdot), \quad \text{as } n \rightarrow \infty \text{ first, and then } r \rightarrow \infty,\tag{4.7}$$

as long as the scaled drifts match, i.e. $\widehat{u}^{(n,r)} \rightarrow u$ as $(n, r) \rightarrow \infty$.

From (4.7) and definition of the cost functions (4.4) and (3.4) we get from Fatou's lemma that for any choice of admissible policy $\{\mu^{(n,r)}\}$ satisfying $(\widehat{q}_0^{(n,r)}, \widehat{u}^{(n,r)}) \rightarrow (q_0, u)$ as $(n, r) \rightarrow \infty$, we have

$$\widehat{J}(\{q_0^{(n,r)}\}, \{\mu_0^{(n,r)}\}) \geq J(q_0, u).\tag{4.8}$$

Now, for the proposed optimal rates $\{\mu^{*(n,r)}\}$ in Theorem 2.4 in (2.12), it follows that the corresponding scaled-drift $\widehat{u}^{*(n,r)} = u^*$, where u^* is the optimal solution of the f-BCP in Theorem 3.1. Hence, using the same set of calculation, one can see that

$$\widehat{J}(\{q_0^{(n,r)}\}, \{\mu_0^{*(n,r)}\}) \geq J(q_0, u^*).$$

Moreover, we will now show that this inequality is indeed an equality. Note that for $t \geq 0$,

$$\widehat{A}^{(n,r)} = \frac{1}{r^H \sqrt{n}} \sum_{i=1}^n V_i^{(n,r)}(t), \quad \text{where } V_i^{(n,r)}(t) = \int_0^t \left(W_i^{(n,r)}(s) - \lambda \right) ds,\tag{4.9}$$

and $\{V_i^{(n,r)} : i = 1, \dots, n\}$ are iid. So, we get from Theorem 5.1 in Section 5, that there exists $n_0 \geq 1, r_0 \geq 1$ such that for $n \geq n_0, r \geq r_0$

$$E \left[\sup_{0 \leq t \leq t} |\widehat{A}^{(n,r)}(s)|^2 \right] \leq \frac{1}{r^{2H} n} \sum_{i=1}^n E \left[\sup_{0 \leq s \leq t} |V_i^{(n,r)}(s)|^2 \right] \leq \frac{1}{r^{2H}} (C(rt)^{3-\alpha}) = Ct^{2H}, \quad \text{for } t \geq 0.\tag{4.10}$$

where $C > 0$ is as in Theorem 5.1. Here, we have also used the fact that $2H = 3 - \alpha$. Now, since $\{\widehat{q}_0^{(n,r)}\}$ is convergent and $\widehat{u}^{*(n,r)} = u^*$ is constant, it follows from (4.10) that for $n \geq n_0, r \geq r_0$ (and hence for all n, r) the following holds.

$$E \left(\int_0^\infty e^{-\theta t} \sup_{0 \leq s \leq t} \left| \widehat{q}_0^{(n,r)} + \widehat{A}^{(n,r)}(s) - \widehat{u}^{(n,r)} s \right| dt \right) < \tilde{K},\tag{4.11}$$

for some constant $\tilde{K} > 0$ which does not depend on (n, r) . Hence, using the representation (4.5) and property of the reflection map in (4.2) that

$$\sup_{n \geq 1, r \geq 1} E \left(\int_0^\infty e^{-\theta t} \left[\sup_{0 \leq s \leq t} [h(\hat{u}^{(n,r)}) + c\hat{Q}^{(n,r)}(s) + \theta p \hat{L}^{(n,r)}(s)]^2 \right] dt < \infty. \quad (4.12)$$

This provides the necessary uniform integrability condition to conclude from (4.7) (with $\{\mu^{*(n,r)}\}$ and u^*) that

$$\hat{J}(\{q_0^{(n,r)}\}, \{\mu_0^{*(n,r)}\}) = J(q_0, u^*). \quad (4.13)$$

Finally, from (4.8) and (4.13) and the fact that $J(q_0, u^*) \leq J(q_0, u)$ for any other $u \geq 0$ (Theorem 3.1), the proof is complete. \blacksquare

5 An Asymptotic Maximal Inequality for Superimposed On-Off Process

Let $\{X_{on}, X_n : n \geq 1\}$ be iid nonnegative random variables representing the On-periods, with the common distribution $F_{on}(\cdot)$. Similarly, $\{Y_{off}, Y_n : n \geq 1\}$ be another independent set of iid random variables with common distribution $F_{off}(\cdot)$, that represents the Off-periods and satisfy tail conditions (2.1) described in Section 2 (see also [10]). Let

$$\{S_n : n \geq 0\} \doteq \left\{ D, D + \sum_{i=1}^n (X_i + Y_i), n \geq 1 \right\}$$

be a *stationary* renewal sequence, where the non-negative delay variable D is as described in [10]. More precisely, D is defined using Y_{off} and three new variables $X_{on}^{(0)}$, $Y_{off}^{(0)}$ and B which are independent of $\{Y_{off}, X_n, Y_n : n \geq 1\}$. Here B is a Bernoulli ($\frac{\mu_{on}}{\mu}$) random variable and distributions of $X_{on}^{(0)}$, $Y_{on}^{(0)}$ are given by,

$$P(X_{on}^{(0)} > x) = \frac{1}{\mu_{on}} \int_x^\infty \bar{F}_{on}(s) ds, \quad P(Y_{off}^{(0)} > x) = \frac{1}{\mu_{off}} \int_x^\infty \bar{F}_{off}(s) ds, \quad (5.1)$$

for $x > 0$. Then the delay variable D is defined as (see [10] for more detail)

$$D = B(X_{on}^{(0)} + Y_{off}) + (1 - B)Y_{off}^{(0)}. \quad (5.2)$$

Now, the On-Off process $\{W(t) : t \geq 0\}$ can be precisely described as

$$W(t) = B 1_{[0, X_{on}^{(0)})}(t) + \sum_{n=0}^{\infty} 1_{[S_n, S_n + X_{n+1})}(t), \quad (5.3)$$

so that for $t \geq D$, $W(t) = 1$ if t is within an On period (i.e. $S_n \leq t < S_n + X_{n+1}$, for some n) and $W(t) = 0$ if t is within an Off-period (i.e. $S_n + X_{n+1} \leq t < S_{n+1}$). Note that, in

this definition, $W(t)$ takes values in $\{0, 1\}$ for $0 \leq t \leq D$. Also define the renewal counting process

$$\xi(t) = \sum_{n=0}^{\infty} 1_{[0,t]}(S_n), \quad t \geq 0. \quad (5.4)$$

As shown in [10], $\{W(t) : t \geq 0\}$ is strictly stationary and

$$E[W(t)] = \frac{\mu_{on}}{\mu} = \lambda, \quad E[\xi(t)] = \frac{t}{\mu}. \quad (5.5)$$

Our main result in this section is for the centered cumulative On-Off process:

$$V(t) = \int_0^t (W(s) - \lambda) ds, \quad t \geq 0, \quad (5.6)$$

where $\lambda = \frac{\mu_{on}}{\mu}$. It is well-known (see [16, Section 3.1]) that $EV(t)^2 \sim t^{3-\alpha}$. The following result provides an asymptotic maximal bound for the process $\{V(t) : t \geq 0\}$.

Theorem 5.1. *For the cumulative On-Off process, the following holds for all $T > 0$*

$$E \left[\sup_{0 \leq t \leq T} |V(t)|^2 \right] \lesssim CT^{3-\alpha},$$

for some constant $C > 0$ (free of T).

Proof. Since the result is about the *rate* of growth of $E(\sup_{0 \leq t \leq T} |V(t)|^2)$, we will use $C > 0$ to denote a generic constant (free of T), whose value may change from one line to another. For the rest of the proof, we fix a (large) value of $T > 0$.

First consider the truncated variables: for $n \geq 1$,

$$\begin{aligned} X_n^T &= \min(X_n, T), \quad Y_n^T = \min(Y_n, T), \quad Y_{on}^{(0),T} = \min(Y_{on}^{(0)}, T), \\ X_{on}^{(0),T} &= \min(X_{on}^{(0)}, T), \quad Y_{off}^T = \min(Y_{off}, T), \quad D^T = \min(D, T). \end{aligned} \quad (5.7)$$

Using the fact that for any nonnegative random variable Z , $EZ^2 = 2 \int_0^\infty zP(Z > z)dz$ and Karamata's Theorem (see [1, Section 16]) we get from 2.1 that

$$E[(X_1^T)^2] = 2 \int_0^T x \bar{F}_{on}(x) dx \sim CT^{2-\alpha}. \quad (5.8)$$

Similar calculation yields that

$$E[(Y_1^T)^2] \sim CT^{2-\beta} \quad \text{and} \quad E[(Y_{off}^T)^2] \sim CT^{2-\beta}. \quad (5.9)$$

Note that from (5.1), we see that $X_{on}^{(0)}, Y_{off}^{(0)}$, have different tail index than X_1, Y_1 . In fact, using Karamata's theorem, we see that

$$P(Y_{on}^{(0)} > x) \sim Cx^{1-\alpha}, \quad P(Y_{off}^{(0)} > x) \sim Cx^{1-\beta}. \quad (5.10)$$

Next consider the delay variable D . From (5.2), one gets that

$$P(D > T) \leq P\left(X_{on}^{(0)} \geq \frac{T}{3}\right) + P\left(Y_{off} \geq \frac{T}{3}\right) + P\left(Y_{off}^{(0)} \geq \frac{t}{3}\right),$$

which together with (5.10), (2.1) and the fact that $T^a = o(T^b)$ for any $b > a$ implies

$$P(D > T) \sim CT^{1-\alpha}. \quad (5.11)$$

Thus, using calculatinos similar to those in deducing (5.8) we get

$$E[(D^T)^2] \sim CT^{3-\alpha}. \quad (5.12)$$

Now, notice that for $t \geq D$, $V(t)$ grows linearly with slope $1 - \frac{\mu_{on}}{\mu} = \frac{\mu_{off}}{\mu}$ over On-intervals and decreases linearly with slope $(-\frac{\mu_{on}}{\mu})$ over the Off-intervals. So the supremum of $|V(t)|$ over $[0, T]$ is attained either at $t = T$ or at the end-points of On or Off periods (see Figure 4). At these points the value of $V(\cdot)$ can be written explicitly: $V(S_0) = V(D)$ and for $n \geq 0$,

$$V(S_n + X_{n+1}) = V(S_n) + \frac{\mu_{off}}{\mu} X_{n+1}, \quad V(S_{n+1}) = V(S_n + X_{n+1}) - \frac{\mu_{on}}{\mu} Y_{n+1}. \quad (5.13)$$

Figure 4 shows evolution of a sample path of $V(\cdot)$. From (5.13), it follows that with

$$Z_n \doteq \frac{1}{\mu}(\mu_{off} X_n + \mu_{on} Y_n), \quad n \geq 1, \quad (5.14)$$

one gets that for all $n \geq 1$

$$\begin{aligned} |V(S_n) - V(D)| &\leq \sum_{i=1}^n Z_i, \\ |V(S_{n-1} + X_n) - V(D)| &\leq \sum_{i=1}^{n-1} Z_i + \frac{\mu_{off}}{\mu} X_n \leq \sum_{i=1}^n Z_i. \end{aligned} \quad (5.15)$$

Next, note that only the X_i 's and Y_i 's that have values less than T contribute towards the values of $V(s)$ (except for the last On-Off cycle). More precisely, the fact $\xi_T \geq 1$ implies

$$D_i \leq T, \quad X_i \leq T, \quad Y_i \leq T, \quad \text{for } i = 1, \dots, (\xi_T - 1).$$

In other words, with $Z_i^T = \min(Z_i, T)$, $i \geq 1$,

$$\{\xi_T \geq 1\} \subseteq \{D_i = D_i^T, Z_i = Z_i^T, \quad i = 1, \dots, \xi_T - 1\}. \quad (5.16)$$

Note that by definition (see (5.4)), $(\xi_T - 1)$ counts the number of completed On-Off cycles, not including the delay interval D (in Figure 4, $\xi_T - 1 = n - 1$). In the last (incomplete) On-Off cycle $[S_{\xi_T-1}, T]$, we can have two cases based on whether T belongs to an On period or Off period. If T is within an On-period, then

$$\begin{aligned} \sup_{S_{\xi_T-1} \leq t \leq T} |V(t) - V(S_{\xi_T-1})| &\leq \frac{\mu_{off}}{\mu} \min(X_{\xi_T}, T - S_{\xi_T-1}) \\ &\leq \frac{\mu_{off}}{\mu} [\min(X_{\xi_T}, T)] + \frac{\mu_{on}}{\mu} [\min(Y_{\xi_T}, T)]. \end{aligned} \quad (5.17)$$

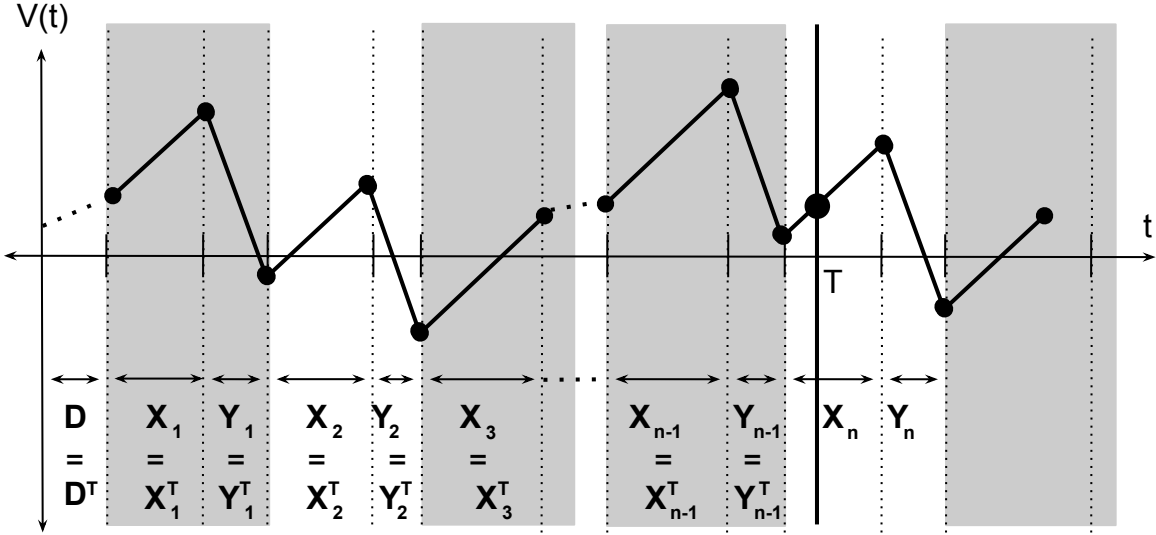


Figure 4: Evolution of a centered On-Off process over time

And if T falls within an Off-period, then the last incomplete cycle has a completed On-period, i.e. $X_{\xi_T} \leq T$ and so,

$$\begin{aligned} \sup_{S_{\xi_T-1} \leq t \leq T} |V(t) - V(S_{\xi_T-1})| &\leq \frac{\mu_{off}}{\mu} \min(X_{\xi_T}, T) + \frac{\mu_{on}}{\mu} \min(Y_{\xi_T}, T - (S_{\xi_T-1} + X_{\xi_T})) \\ &\leq \frac{\mu_{off}}{\mu} [\min(X_{\xi_T}, T)] + \frac{\mu_{on}}{\mu} [\min(Y_{\xi_T}, T)]. \end{aligned} \quad (5.18)$$

Hence,

$$\begin{aligned} E \left[\sup_{0 \leq t \leq T} |V(t)|^2 \right] &\leq E \left[\sup_{0 \leq t \leq T} |V(t)|^2 1_{\{\xi_T=0\}} \right] + E \left[\sup_{0 \leq t \leq T} |V(t)|^2 1_{\{\xi_T \geq 1\}} \right] \\ &\leq E \left[\sup_{0 \leq t \leq T} |V(t)|^2 1_{\{\xi_T=0\}} \right] + 2E \left[\sup_{0 \leq t \leq D} |V(t)|^2 1_{\{\xi_T \geq 1\}} \right] \\ &\quad + E \left[\sup_{D \leq t \leq S_{\xi_T-1}} |V(t) - V(D)|^2 1_{\{\xi_T \geq 1\}} \right] + E \left[\sup_{S_{\xi_T-1} \leq t \leq T} |V(t) - V(S_{\xi_T-1})|^2 1_{\{\xi_T \geq 1\}} \right]. \end{aligned} \quad (5.19)$$

Now we treat each term of (5.19) separately. For the first term, note that

$$|V(t)| \leq Ct, \quad (5.20)$$

for some $C > 0$ (see definition (5.6) and use the fact that $\mu_{on} \leq \mu$). Also, $\xi_T = 0$ is equivalent to $D \geq T$. So,

$$E \left[\sup_{0 \leq t \leq T} |V(t)|^2 1_{\{\xi_T=0\}} \right] \leq (C)^2 T^2 P(D \geq T) \sim CT^{3-\alpha}, \quad (5.21)$$

using (5.11). Again using (5.20) above with (5.12), we get

$$E \left[\sup_{0 \leq t \leq D} |V(t)|^2 1_{\{\xi_T \geq 1\}} \right] \leq CE[(D^T)^2] \sim CT^{3-\alpha}. \quad (5.22)$$

Here we have also used the observation (5.16). Now, recall Z_i 's defined in (5.14) and set $Z_i^T = \min(Z_i, T)$, $i \geq 1$. Then using (5.16) and (5.15) we get

$$E \left[\sup_{D \leq t \leq S_{\xi_T-1}} |V(t) - V(D)|^2 1_{\{\xi_T \geq 1\}} \right] \leq 2E \left[\sum_{i=1}^{\xi_T-1} Z_i^T \right], \quad (5.23)$$

and from (5.17) – (5.18), we get from the definition of Z_i^T ,

$$E \left[\sup_{S_{\xi_T-1} \leq t \leq T} |V(t) - V(S_{\xi_T})|^2 1_{\{\xi_T \geq 1\}} \right] \leq E \left[(Z_{\xi_T}^T)^2 \right]. \quad (5.24)$$

Therefore, combining (5.23)-(5.24), we get

$$\begin{aligned} E \left[\sup_{0 \leq t \leq S_{\xi_T-1}} |V(t) - V(D)|^2 1_{\{\xi_T \geq 1\}} \right] &+ E \left[\sup_{S_{\xi_T-1} \leq t \leq T} |V(t) - V(S_{\xi_T-1})|^2 1_{\{\xi_T \geq 1\}} \right] \\ &\leq 2E \left[\sum_{i=1}^{\xi_T} Z_i^T \right] = E[\xi_T]E[Z_1^T], \end{aligned} \quad (5.25)$$

using Wald's theorem for random sums, since $Z_i^T \leq T$ are iid and ξ_T is a stopping time (see [10]) with respect to the filtration $\{\mathcal{F}_n\}_{n \geq 1}$, where $\mathcal{F}_n = \sigma\{D, X_i, Y_i : 1 \leq i \leq n\}$ and $E|\xi_T| < \infty$. In fact, from (5.5), $E(\xi_T) = \frac{T}{\mu}$ and from (5.8)-(5.9) and definition of Z_i^T , it follows that $E[Z_1^T] \sim CT^{2-\alpha}$. Thus,

$$E[\xi_T]E[Z_1^T] \sim CT^{3-\alpha}. \quad (5.26)$$

Thus, by substituting (5.21), (5.22), (5.25) and (5.26) in (5.19), we get that

$$E \left[\sup_{0 \leq t \leq T} |V(t)|^2 \right] \lesssim CT^{3-\alpha}.$$

This completes the proof of the theorem. ■

References

- [1] N.H. Bingham, C.M. Goldie and J.L. Teugels, Regular Variation. *Cambridge Univ. Press*, 1987.
- [2] M. Bramson and R.J. Williams, On dynamic scheduling of stochastic networks in heavy traffic and some new results for the workload process. In *Proceedings of the 39th IEEE Conference on Decision and Control*. IEEE, New York, 2000.

- [3] K. Dębicki and A. Tomanek, Estimates for moments of supremum of reflected fractional Brownian motion. *preprint.*, 2010, ARXIV <http://arxiv.org/abs/092.3117>
- [4] R. Delgado, A reflected fBm limit for fluid models with ON/OFF sources under heavy traffic. *Stochastic Processes and their Applications*, 117 (2), 188–201, 2007.
- [5] A.P. Ghosh and A. Weerasinghe, Optimal buffer size and dynamic rate control for a queueing system with impatient customers in heavy traffic. *Stochastic Processes and their Applications*, 120(11), 2103–2141, 2010.
- [6] A.P. Ghosh, A. Roitershtein, and A. Weerasinghe, Optimal control of a stochastic processing system driven by a fractional Brownian motion input. *Advances in Applied Probability* 42(1), 183–209, 2010.
- [7] W.B. Gong, Y. Liu, V. Misra, and D Towsley, Self-similarity and long range dependence on the Internet: a second look at the evidence, origins and implications, *Computer Networks*, 48, 377–399, 2005.
- [8] J.M. Harrison, A broader view of Brownian networks, *Annals of Applied Probability* 13, 1119–1150, 2003.
- [9] D. Heath, S. Resnick, and G. Samorodnitsky, Patterns of buffer overflow in a class of queues with long memory in the input stream, *Annals of Applied Probability* 7, 1021–1057, 1997.
- [10] D. Heath, S. Resnick and G. Samorodnitsky, Heavy tails and long range dependence in on/off processes and associated fluid models. *Mathematics of Operations Research*, 23 145–165, 1998.
- [11] Y. Hu, and X.Y. Zhou, Stochastic control for linear systems driven by fractional noises. *SIAM Journal of Control and Optimization*, 43, 2245–2277, 2005.
- [12] W.E. Leland, M.S. Taquq, W. Willinger, and D.V. Wilson, On the self-similar nature of Ethernet traffic (extended version), *IEEE/ACM Transactions on Networking*, 2, 1–15, 1994.
- [13] T. Konstantopoulos and S. Lin, Fractional Brownian approximation of queueing networks, In *Stochastic Networks*, Lecture Notes in Statistics 117, Springer, New York, 257–273, 1996.
- [14] L. Kruk, J. Lehoczky, K. Ramanan, and S. Shreve, An explicit formula for the Skorohod map on $[0, a]$, *Annals of Probability* 35, 1740–1768, 2007.
- [15] B.B. Mandelbrot and J.W. Van Ness, Fractional Brownian motions, fractional noises and applications, *SIAM Review*, 10, 422–437, 1968.
- [16] T. Mikosch and G. Samorodnitsky, Scaling limits for cumulative input process. *Mathematics of Operations Research*, 32, 890–919, 2007.

- [17] I. Norros, A storage model with self-similar input, *Queueing Systems*, 16, 387–396, 1994.
- [18] D. Nualart, The Malliavin Calculus and Related Topics, *Springer’s series in Probability and its Applications*, 2nd ed., Springer, 2006.
- [19] V. Paxson and S. Floyd, Wide-area traffic: the failure of Poisson modeling, *IEEE/ACM Transactions on Networking*, 3, 226–244, 1995.
- [20] Z. Sahinoglu and S. Tekinay, Self-similar traffic and network performance, *IEEE Communications Magazine*, 37, 48–52, 1999.
- [21] G. Samorodnitsky and M.S. Taqqu, Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance, *Chapman and Hall*, New York, 1994.
- [22] A.N. Shiryaev, Essentials of Stochastic Finance, *World Scientific*, Singapore, 1999.
- [23] A. Stegeman, Extremal behavior of heavy-tailed ON-periods in a superposition of ON/OFF processes. *Advances in Applied Probability*, 34, 179–204, 2002
- [24] M. Taqqu, W. Willinger, and R. Sherman, Proof of a fundamental result in self-similar traffic modeling, *Computer Comm. Rev.*, 27, 5–23, 1997.
- [25] M.S. Taqqu, W. Willinger, R. Sherman, and D.V. Wilson, Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level, *IEEE/ACM Transactions on Networking*, 5, 71–86, 1997.
- [26] W. Whitt, An overview of Brownian and non-Brownian FCLTs for the single-server queue, *Queueing Systems*, 36, 39–70, 2000.
- [27] W. Whitt, Stochastic-Process Limits. An Introduction to Stochastic-Process Limits and their Application to Queues, *Springer*, 2002.
- [28] W. Willinger, V. Paxson, and M. S. Taqqu, Self-similarity and heavy tails: structural modeling of network traffic, In *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*, R. Adler, R. Feldman, and M. S. Taqqu editors, Birkhauser, 1998.
- [29] W. Willinger, M. S. Taqqu, and A. Erramilli, A Bibliographical Guide to Self-Similar Traffic and Performance Modeling for Modern High-Speed Networks Stochastic Networks: Theory and Applications, *Royal Statistical Society Lecture Notes Series*, 4, Oxford University Press, 1996.
- [30] B. Zwart, S.C. Borst and K.G. Dębicki, Subexponential asymptotics of hybrid fluid and ruin problems. *Annals of Applied Probability*, 15, 500–517, 2005.

Arka P. Ghosh
 3216 Snedecor Hall, Department of Statistics
 Iowa State University, Ames, IA 50011, USA
 apghosh@iastate.edu