# TB/GI/1 queues with arrival traffic envelopes*

George Kesidis
School of EECS, Penn. State Univ., USA

Takis Konstantopoulos
Math. Dept, Univ. of Liverpool, UK

## ABSTRACT

Consider a queueing system where the job service times are not known upon arrival; e.g., a transmission server of a wireless channel where packet transmission times are random, or a virtual machine handling a stream of tasks whose execution times are not perfectly predictable. In this preliminary study, we give bounds on the tail of the workload distribution of a partially regulated, single server queue whose arrival processes are arbitrarily distributed stationary random point processes on the integers that satisfy token-bucket constraints expressed via an arbitrary concave function $f$, and whose job service times are independent with a common distribution.

## 1. INTRODUCTION

Token bucket (TB) or leaky bucket constraints are commonly used to limit demand for a server (or servers) when the workload (service time) of a job is known upon arrival. The tokens account for the job service times and job arrival increments are limited by a subadditive burstiness curve or traffic envelope. The use of token bucket mechanisms can guarantee that the workload in the job queue, and hence the delay to the server, is bounded.

In some cases, however, the job service times are not known when the job arrives and so it is not possible to apply a token-bucket mechanism to the arriving workload process. For example, the jobs could represent programs to be run for an input dataset and the server is a virtual machine or containerized executor. Alternatively, the jobs could be packets requiring transmission into a wireless channel (i.e., a transmission server), a process requiring a random amount of time. In TB/GI/1 queues considered in this paper, the arrival-times processes of jobs are token-bucket regulated, but the i.i.d. job service times are not, though with assumed known job service time distribution. The exact definition of these queues is in the next section.

The continuous-time TB/GI/$\infty$ (infinite server) system is considered in [8, 7]. For the stationary, continuous-time, fluid single-server queue with token-bucket constrained arrival processes and deterministic service times (TB/D/1), the tail of the workload distribution was bounded in [10, 9] (and also considered in [4]). In this preliminary study, we generalize these results to the partially regulated TB/GI/1 case with subadditive traffic envelopes.

*This work supported in part by NSF CNS grant 1717571 and a Cisco URP gift. The authors can be reached at gik2@psu.edu and takiskonst@gmail.com



## 2. BASIC ASSUMPTIONS

With $a(t)$ denoting the number of arrivals at time $t \in \mathbb{Z}$, we assume that the random process $a(t)$, $t \in \mathbb{Z}$, is stationary and

$$A(s,t] := \sum_{i=s+1}^{t} a(i) \le f(t-s), \quad -\infty < s \le t < \infty, \quad (1)$$

where $f : [0,\infty) \to [0,\infty)$ is a given concave increasing function with $f(0) = 0$. A comprehensive treatment and optimal design methodology of systems regulated by such functions can be found in [11, 1]. Let $X, X_{t,i}$ be strictly positive i.i.d. random variables representing service times, and independent of the arrival process. At time $t$ the total load brought into the system is $\sum_{i=1}^{a(t)} X_{t,i}$. We assume that $X$ is bounded: $\mathbb{P}(X \le \xi_{\max}) = 1$. Note that (1) does not affect the service times. It is easy to see that $f$ is subadditive and hence, by Fekete's lemma, $\rho = \lim_{t\to\infty} f(t)/t$ exists and is equal to $\inf_{t>0} f(t)/t$. We assume that $\rho > 0$. We denote the right derivative of $f$ by $f'$ and let $f'(\infty) = \rho$. Note that $f(1) \ge f'(0)$. So that some jobs may successfully arrive to the discrete-time system, we require that the permitted peak rate $\pi := f(1) \ge 1$. For example, we can use a dual token-bucket mechanism to achieve $f(t) = \min\{\pi t, \sigma + \rho t\}$, where the peak rate is larger than the sustainable rate, $\pi > \rho > 0$, the burst size is $\sigma > 0$, and, in order for the $\pi t$ constraint to be relevant, $\sigma + \rho > \pi$. Moreover, $\pi \ge \sigma$ permits a burst of size $\sigma$ to simultaneously arrive.

Suppose there are $n$ independent arrival processes $A_k$, $1 \le k \le n$, each satisfying (1). If the service rate is $r$ and if

$$n\rho\, \mathbb{E}X < r, \quad (2)$$

then there is a unique stationary workload process,

$$W(t) = \sup_{s\le t}\{\tilde{B}(s,t] - r(t-s)\},\ t \in \mathbb{Z}, \quad (3)$$

where $\tilde{B}(s,t] = \sum_{k=1}^{n} B_k(s,t] = \sum_{i=s+1}^{t}\sum_{k=1}^{n}\sum_{j=1}^{a_k(i)} X_{k,i,j}$.

## 3. BEST POSSIBLE BOUND ON THE TAIL OF THE JOBS-QUEUE DISTRIBUTION

Let $f'$ be the right derivative of $f$. For $v > 0$ let $\sigma(v) := f(v) - f'(v)v$, so that the line $g_v(t) = f'(v)t + \sigma(v)$ is tangent to $f$ at $v$. By concavity, $f(t) = \inf_{v\ge 0} g_v(t)$. Define the

auxiliary queueing system for a single job-arrival process ($n = 1$) with *job* service rate $f'(v)$:

$$Q_v(t) = \sup_{s \le t}\{A(s,t] - f'(v)(t-s)\},\ t \in \mathbb{Z}. \tag{4}$$

Note that $Q_v$ is a stationary process. Hence $\mathbb{P}(Q_v(t) \ge q)$ does not depend on $t$. Our goal is to identify the arrival process $A^*$ that satisfies (1) and results in an auxiliary queueing process $Q_v^*$ such that $\mathbb{P}(Q_v(0) \ge q) \le \mathbb{P}(Q_v^*(0) \ge q)$ for all arrival processes $A$ satisfying (1). Clearly, $Q_v(t) \le Q_\infty(t)$ for all $t$. If $v$ is such that

$$\rho < f'(v) < f(1) = \pi, \tag{5}$$

take $q \ge 0$ such that $q < \sup_{t>0}\{f(t) - f'(v)t\}$. By the above assumptions, $q < \infty$ and define

$$t_1 := \inf\{t \in \mathbb{Z}_+ :\ f(t) - f'(v)t \ge q\},$$
$$t_2 := \sup\{t \in \mathbb{Z}_+ :\ f(t_1 + t) \ge f(t_1) + f'(v)t\},$$
$$t_3 := \inf\{t \in \mathbb{Z}_+ :\ f(t_1) + f'(v)t_2 - \rho(t_1 + t_2 + t) \le 0\}.$$

Let $T = t_1 + t_2 + t_3$ and arbitrary $s_0 \in \mathbb{Z}$. Define

$$A^*(s_0, s_0+t] = \begin{cases} f(t) & 1 \le t \le t_1 \\ f(t_1) + f'(v)(t - t_1) & t_1 + 1 \le t \le t_1 + t_2 \\ 0 & t_1 + t_2 + 1 \le t \le T \end{cases}$$

and extend periodically: $A(s,t] = A(s + nT, t + nT]$ for all $n \in \mathbb{Z}$. Note that $A^*$ depends on $q, v$. Let $Q_v^*(t)$ be defined as in (4) with $A^*$ in place of $A$.

LEMMA 3.1. *The bound (1) holds for $A^*$.*

PROOF. It sufficies to show that when $A = A^*$, $Q_u(t) \le \sigma(u)$ for all $t$ and all $u \in (0, \infty)$. We take two cases for arbitrary $t \in \mathbb{Z}$: If $u \le v$, then the largest $Q$ occurs at times $t_1$ from the start of the queue busy period so that

$$Q_u(t) \le f(t_1) - f'(u)t_1 = f(t_1) - (g_u(t_1) - \sigma(u)) \quad \le \sigma(u),$$

where the first inequality is by the definition of $t_1$. If $u > v$, then the largest $Q$ occurs at times $t_1 + t_2$ from the start of the queue busy period so that

$$\begin{aligned} Q_u(t) &\le f(t_1) - f'(u)t_1 + (f'(v) - f'(u))t_2 \\ &= f(t_1) + f'(v)t_2 - f'(u)(t_1 + t_2) \\ &\le f(t_1 + t_2) - f'(u)(t_1 + t_2) \\ &\le f(t_1 + t_2) - (g_u(t_1 + t_2) - \sigma(u)) \le \sigma(u), \end{aligned}$$

where the second inequality is by the definition of $t_2$. □

The proof of the following, using the Palm inversion formula, is as that of Theorem 2.2 of [10, 9], but relies instead on the previous lemma.

THEOREM 3.1. *For all $v$ such that (5) holds,*

$$\mathbb{P}(Q_v(0) \ge q) \le \frac{t_2}{t_1 + t_2 + t_3} = \mathbb{P}(Q_v^*(0) \ge q).$$

Note that $\frac{t_2}{t_1+t_2+t_3}$ is maximized by minimizing $t_1, t_3$ and maximizing $t_2$,consistent with their definitions. Also, note that $\mathbb{P}(Q_v \ge q) > 0$ only if $q < \sup_{t>0}\{f(t) - f'(v)t\}$.

# 4. BOUNDS ON THE TAIL OF THE WORKLOAD DISTRIBUTION

## 4.1 Single arrival process

Consider the stationary workload process $W$ of (3) with $n = 1$. Since $X \le \xi_{\max}$ with probability 1, we have

$$\sup_{t \ge 0} W(t) \le \sup_{t \ge 0}\{f(t)\xi_{\max} - rt\}^+.$$

So as to avoid a trivial system with zero workload, we need to assume $\pi > r/\xi_{\max}$. Let $v$ be such that $f'(v) = r/\xi_{\max}$ which implies that $W(t) \le \xi_{\max} Q_v(t)$, for all $t$. Thus, for stability of both (4) and (3) with $n = 1$, we require

$$r = f'(v)\xi_{\max} > \quad \rho\xi_{\max} \tag{6}$$

which implies (2) with $n = 1$.

COROLLARY 4.1. *If (6) holds then*

$$\mathbb{P}(W(0) \ge q\xi_{\max}) \le \mathbb{P}(Q_v(0) \ge q) \le \frac{t_2}{t_1 + t_2 + t_3}. \tag{7}$$

Though the second inequality of (7) is achieved by $A^*$[1], the first may not be tight particularly when the variance in the service-time distribution $X$ is large.

## 4.2 Multiple arrival processes

We now consider a queueing system with plural ($n \ge 1$) independent arrival processes each satisfying (1). Let $W$ be defined by (3). The following results are easily extended to the case where constraints on each arrival process are different, i.e., $n\rho \to \sum_{k=1}^n \rho_k$ and $nf \to \sum_{k=1}^n f_k$.

PROPOSITION 4.1. *If $\mathbb{P}(X > r) > 0$, then*

$$\mathbb{P}(W(0) > x) \le \exp\big[-\sup_{\theta>0}\{\theta x - \log(1 - \alpha + \alpha \sup_{s>0} h(s,\theta))\}\big], \tag{8}$$

*where $\alpha := n\rho\mu/r$, $\mu := \mathbb{E}X$, $\Phi_X(a) := \mathbb{E}e^{aX}$, and*

$$h(s,\theta) := \frac{1}{\mathbb{P}(X > r)} \cdot \frac{1}{s}\sum_{t=1}^{s} \Phi_X(\theta n f(t)) e^{-\theta r t}.$$

PROOF. Let time $S_k$, respectively $V_k$, be the beginning of the $k^{\text{th}}$ busy, respectively idle, period of $W$ for $k \in \mathbb{Z}$. So $\tau_k = V_k - S_k > 0$ and $\tau_k' = S_{k+1} - V_k > 0$ are respectively the duration of the $k^{\text{th}}$ busy and idle period of $W$. The duration of the $k^{\text{th}}$ busy cycle is $\tau_k + \tau_k' = S_{k+1} - S_k$. Define $\mathbb{P}^\sharp(A) := \mathbb{P}(A | S_k = 0 \text{ for some } k)$. Since $\mathbb{P}$ can be recovered from $\mathbb{P}^\sharp$ (see, e.g., [12, Theorem 1])

$$\mathbb{E}\mathrm{e}^{\theta W(0)} = \frac{\mathbb{E}^\sharp \sum_{t=S_0}^{V_0-1} \mathrm{e}^{\theta W(t)} + \mathbb{E}^\sharp \tau_0'}{\mathbb{E}^\sharp \tau_0 + \mathbb{E}^\sharp \tau_0'} = \frac{y + z}{1 + z} \tag{9}$$

where

$$y = \frac{1}{\mathbb{E}^\sharp \tau_0} \mathbb{E}^\sharp \sum_{t=S_0}^{V_0-1} \mathrm{e}^{\theta W(t)}, \qquad z = \frac{\mathbb{E}^\sharp \tau_0'}{\mathbb{E}^\sharp \tau_0}, \tag{10}$$

and $\mathbb{E}^\sharp$ is expectation with respect to $\mathbb{P}^\sharp$. Note that $y$ is the $\mathbb{P}$-expectation of $\mathrm{e}^{\theta W(0)}$ conditional on $W(0) > 0$. Also note that the ratio that $(y + z)/(1 + z)$ is increasing in $y$ and, since $y > 1$, the same ratio is decreasing in $z$. Thus, our objective is to maximize $y$ and minimize $z$.

**To minimize** $z$, first define $Y(t) = \sup_{-\infty < s \le t}\{\tilde{B}(s,t] -$

[1]i.e., the TB/D/1 case, again see [10, 9] for the special case of $f(s) = \min\{\pi s, \sigma + \rho s\}$

$n\rho\mu(t-s)\}$ and observe that $Y \geq W$ by (2). Note that $\tilde{B}(S_0-1, V_0-1] \geq r\tau_0$. Since the busy periods of $W$ are entirely contained in those of $Y$,

$$\begin{aligned} Y(V_0-1) &= Y(S_0-1) + \tilde{B}(S_0-1, V_0-1] - n\rho\mu\tau_0 \\ &\geq Y(S_0-1) + (r-n\rho\mu)\tau_0. \end{aligned}$$

Also, note that

$$\begin{aligned} Y(S_1-1) &\geq Y(V_0-1) + \tilde{B}(V_0-1, S_1-1] - n\rho\mu\tau_0' \\ &\geq Y(V_0-1) - n\rho\mu\tau_0' \\ &= Y(S_0-1) + (r-n\rho\mu)\tau_0 - n\rho\mu\tau_0'. \end{aligned}$$

Finally, since $\mathbb{E}^\sharp Y(S_1-1) = \mathbb{E}^\sharp Y(S_0-1)$, we have, by taking $\mathbb{P}^\sharp$-expectations in the above display, $0 \geq (r-n\rho\mu)\mathbb{E}^\sharp\tau_0 - n\rho\mu\mathbb{E}^\sharp\tau_0'$. Substituting in (10) we obtain

$$z \geq \frac{r}{n\rho\mu} - 1 = \frac{1}{\alpha} - 1 \quad \text{(where } \alpha < 1 \text{ by (6))}.$$

**To maximize** $y$: For $0 \leq t < \tau_0 = V_0 - S_0$ we can write

$$\begin{aligned} W(S_0+t) &= \tilde{B}(S_0-1, S_0+t] - r(t+1) \\ &= \sum_{i=S_0}^{S_0+t} \sum_{k=1}^{n} \sum_{j=1}^{a_k(i)} X_{k,i,j} - r(t+1), \end{aligned} \tag{11}$$

because the process $W$ is positive on the interval $[S_0, V_0)$. Define

$$\tilde{W}(S_0+t) = \sum_{i=1}^{nf(t+1)} \tilde{X}_i - r(t+1), \tag{12}$$

where $\tilde{X}_i$ are i.i.d. copies of $X$. Note that under $\mathbb{P}^\sharp$, $W(S_0) > 0$ and $S_0 + \tau_0 = V_0 > S_0$ is the first time $t$ after $S_0$ that $\tilde{B}(S_0-1, S_0+t] - r(t+1) \leq 0$.
Now, $X_{1,1,1} > r$ implies $W(S_0) > 0$, and, by hypothesis, $\mathbb{P}(X > r) > 0$. So, by (1), the $\mathbb{P}^\sharp$ distribution of $W$ is dominated by that of $\tilde{W}$ divided by $\mathbb{P}(X > r)$.[2] Therefore,

$$y = \mathbb{E}^\sharp \sum_{t=0}^{\tau_0-1} \mathrm{e}^{\theta W(S_0+t)} \leq \frac{1}{\mathbb{P}(X>r)} \mathbb{E}^\sharp \sum_{t=0}^{\tau_0-1} \mathrm{e}^{\theta \tilde{W}(S_0+t)}.$$

By substituting (12) and conditioning on $\tau_0$, we get

$$\begin{aligned} y &\leq \frac{1}{\mathbb{P}(X>r)} \mathbb{E}^\sharp \sum_{t=0}^{\tau_0-1} \Phi_X(\theta n f(t+1)) \mathrm{e}^{-\theta r(t+1)} \\ &= \mathbb{E}^\sharp \tau_0 h(\tau_0, \theta) \;\leq\; (\mathbb{E}^\sharp \tau_0) \sup_{s>0} h(s,\theta). \end{aligned}$$

Substituting the above bounds on $z$ and $y$ into (9) gives

$$\mathbb{E}\mathrm{e}^{\theta W(0)} \;\leq\; 1 - \alpha + \alpha \sup_{s>0} h(s,\theta). \tag{13}$$

The proof then follows by the Chernoff bound [2, XIII(2.2)]. □

[2]To see why, suppose random variables $Z_i \geq 0$ are i.i.d. and $\mathbb{P}(Z_i > r) > 0$. For all $a > 0$, $\mathbb{P}(Z_1+Z_2+Z_3 > a \mid Z_1+Z_2 > r) \leq \frac{\mathbb{P}(Z_1+Z_2+Z_3>a)}{\mathbb{P}(Z_1>r)}$. So, for any increasing, non-negative function $G$, $\mathbb{E}(G(Z_1,Z_2,Z_3) \mid Z_1+Z_2>r) = \int_0^\infty \mathbb{P}(G(Z_1+Z_2+Z_3) > g \mid Z_1+Z_2>r)\mathrm{d}g \leq \frac{\mathbb{E}G(Z_1+Z_2+Z_3)}{\mathbb{P}(Z_1>r)}$.

# 5. DISCUSSION

We derived bounds on the tail of the stationary distribution of the workload (virtual queueing-delay process) of partially regulated TB/GI/1 queues.

The results of [4], using the union bound of [3], can be similarly extended.

In addition to different traffic envelopes $f$, we can also generalize the above results to allow for different service time distributions for each of the independent arrival processes.

We can generalize the above results to a lower service curve (instead of constant-rate service $r$) [6]. Also, we can generalize to a queueing system without "cut-through," $W(t) = (W(t-1)-r)^+ + \sum_{j=1}^{a(t)} X_{t,j}$.

Note that a continuous-time queue with discrete arrivals requires $f(0) \geq 1$; e.g., $f(t) = \min\{f(0) + \pi t, \sigma + \rho t\}$ where $1 \leq f(0) < \sigma$ and $0 < \rho < \pi$. Of course, here the definitions of $t_k$ for Theorem 3.1 would need to be changed to optimize over the positive reals, and the definition of $h$ in Prop. 4.1 would involve an integral. (These definitions would also work to extend the results of [9, 10] for continuous-time fluid queues to more general traffic envelopes.)

# 6. REFERENCES


[1] V. Anantharam and T. Konstantopoulos. A methodology for the design of optimal traffic shapers in communication networks. *IEEE Trans. Aut. Control* **44**:583-586, 1999.

[2] S. Asmussen. *Applied Probability and Queues.* Springer-Verlag, New York, 2003.

[3] R. Boorstyn, A. Burchard, J. Liebeherr, and C. Oottamakorn. Statistical multiplexing gain of link scheduling algorithms in QoS networks. Tech. Report CS-99-21, Univ. of Virginia, 1999.

[4] C.-S. Chang. Stochastic bounds for multiplexing independent regulated inputs. *ACM SIGMETRICS Performance Evaluation Review* **29**, June 2003.

[5] R.L. Cruz. Quality of service guarantees in virtual circuit switched networks. *IEEE J. Selected Areas Com.* **13**, no. 6. 1048–1056, 1995.

[6] M. Fidler. Survey of deterministic and stochastic service curve models in the network calculus. *IEEE Communications Surveys & Tutorials* **12**, 2010.

[7] G. Kesidis. Overbooking Lambda functions in the cloud. In *Proc. Workshop on Containers*, U.C. Davis, Davis, CA, USA, Dec. 2019.

[8] G. Kesidis, K. Chakraborty, and L. Tassiulas. Traffic shaping for a loss system. *IEEE Communication Letters*, Vol. 4, No. 12:pp. 417–419, Dec. 2000.

[9] G. Kesidis and T. Konstantopoulos. Extremal traffic and worst-case performance for a queue with shaped arrivals. In *Analysis of Communication Networks* edited by D.R. McDonald and S.R.E. Turner, Fields Institute Communications/AMS, 2000.

[10] G. Kesidis and T. Konstantopoulos. Shape-controlled traffic patterns that maximize overflow probabilities in high-speed networks. In *Proc. IEEE CDC*, Dec. 1998.

[11] T. Konstantopoulos and V. Anantharam. Optimal flow control schemes that regulate the burstiness of traffic. *IEEE/ACM Trans. Networking* **3**:423-432, 1995.

[12] T. Konstantopoulos and M. Zazanis. A discrete time proof of Neveu's exchange formula. *J. Appl. Prob.* **32**:917-921, 1995.