September 1992 *LIDS-P-2139*

Tight Performance Bounds for $\cdot/D/1$ Queues with Leaky-bucket-regulated Arrivals [1]

Daniel C. Lee

Laboratory for Information and Decision Systems, M.I.T.
Cambridge, MA 02139

ABSTRACT

We study a single server queueing system with deterministic service time in which arrivals are regulated by a leaky-bucket control. The worst traffic of arrivals shaped by the leaky-bucket regulation is discussed. The performance measure considered is queueing delay averaged over all customers. We examine both a single stream and multiple streams of arrivals. In both cases, the worst traffic is characterized as the repetition of the following three phases: bulky arrival with bulk size related to the bucket size σ, arrival at every token generation for a specified length of interval, and then no arrival till the token bucket is full. In the case of the single stream, the average queueing delay for the worst traffic, i.e. tight performance bound, is expressed in closed form as a function of leaky bucket parameters (bucket size and arrival rate). We expect that this function will provide insights into the relationship between leaky bucket parameters and the corresponding bandwidth allocated. For the case of multiple streams, each stream is shaped by separate leaky bucket regulations, and the worst queueing delays are compared for different arrangements of token generation times for each stream.

Key words: high-speed network, congestion control, leaky bucket, admission rate, burstiness, queueing delay

[1] Research supported by the NSF under grant ECS-8552419

1 Introduction

High–speed integrated packet-switching networks are characterized by high transmission speed and variety of traffic types. The high transmission speed and the resulting high ratio of transmission speed to propagation delay make computationally simple open-loop control schemes desirable for congestion control. In an integrated network, congestion control must guarantee a certain bandwidth for real-time traffic such as voice or video. For these reasons, the leaky bucket scheme [8] is considered suitable.

A leaky bucket controller is comprised of a packet buffer and a token bucket. Packets arrive at the buffer and get queued. For a packet in the buffer to leave the controller and be admitted into the network, it must obtain a token from the token bucket. Tokens are generated in the bucket periodically with a specified rate r. The token bucket has a fixed size σ. If the token bucket is full at the time of token generation, the newly generated token is discarded. This scheme is specified by two parameters: the token generation rate r and the bucket size σ. The token generation rate quantifies the allowed rate of admissions, and the bucket size quantifies the allowed burstiness of the traffic admitted.

This scheme has drawn the attention of various authors. In [1, 2], the throughput of admitted packets and the blocking probability at the finite-buffer controller are analyzed as a function of the leaky bucket parameters. In [7], the statistics of the queue formed in the controller buffer and the interdeparture time from this buffer are quantified under the assumption that the packet arrival at the controller is modeled by a Poisson process. The relationship between the controller buffer's queue statistics and the leaky bucket parameters is thereby understood. However, the statistics of the queues formed in the network downstream from the leaky bucket controller are not analyzed. In [5], a stochastic fluid model is used to represent the continuous flow of data, whereas a point process model representing packets of data is used in the aforementioned literature. In [3, 4, 6], quantities describing the behavior of downstream queues as well as the controller are analyzed. Their formulation is drastically different from the ones used in the other literature mentioned above in that nonprobabilistic analysis is used. The worst delay over all packets and the maximal queue length that can possibly be reached at some point of time under the leaky bucket control are the primary quantities of interest in their studies.

In this paper, we analyze the leaky bucket regulation from a queueing theoretical point of view. We analyze a single-server queue with deterministic service time at which the arrival is regulated by a leaky bucket scheme. (See Figure 1.) Arrivals are modeled as a point process. These arrivals at the single-server queue are departures

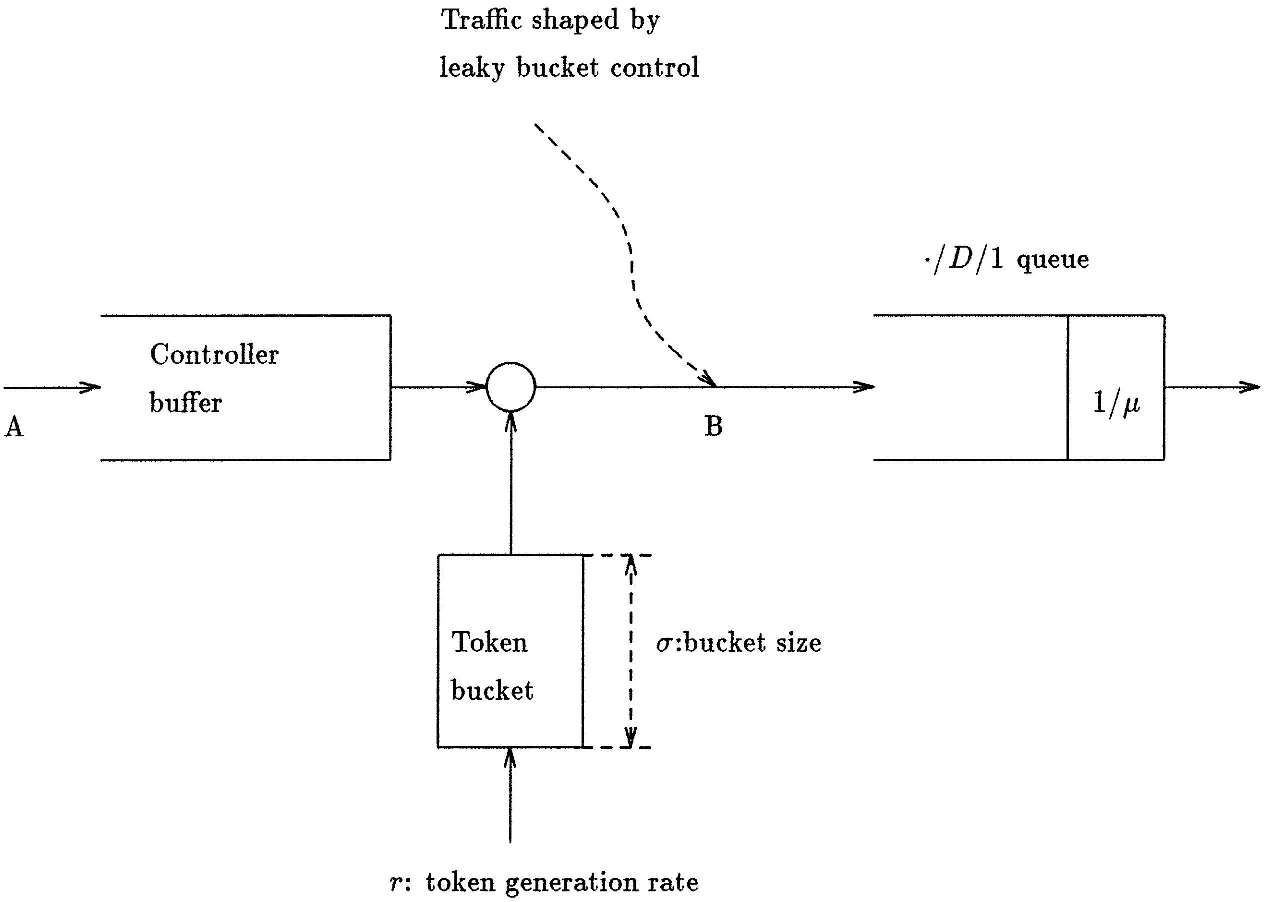


Figure 1: $\cdot/D/1$ queue with arrivals regulated by leaky bucket scheme

from the leaky bucket controller, and satisfy a certain rate and burstiness constraint. (This is the arrival process at point B in Figure 1. Throughout this paper, we will interchangeably refer to this arrival at point B as "admission".) We refer to this queueing system as a $\cdot/D/1$ queue. We will use words "packet" and "customer" interchangeably. The quantity of our interest is the queueing delay averaged over all packets for the worst arrival pattern shaped by the leaky bucket regulation. From the standpoint of networking, this paper views the whole network downstream from the leaky bucket controller as a single-server queueing system with deterministic service time. The formulation in this paper is similar to [3, 4, 6] in that a nonprobabilistic approach is taken. A major distinction between this paper and [3, 4, 6] is that the quantity of interest in this paper is the average delay rather than the delay at the peak.

In section 3, the worst arrival process to the queue (or the worst departure process from the controller) is specified. The worst queueing delay for this process is derived as a function of the leaky bucket parameters: the token generation rate r, and the bucket size σ. Thus, the effect of r, σ, and their interaction effect on the worst-case average delay are specified. In section 4, the discussion is extended to the case of multiple sources of arrivals, where several streams of packets arrive at the queue, with each stream shaped by its own leaky bucket regulator.

2 Preliminaries

Before discussing main results, we note two properties of the general queueing system that we will use frequently in our analysis.

Lemma 1 *Assume equal service time for all customers. Then, the total waiting time of a busy period is increased by hastening the admission time of any customer within the busy period.*

Proof
See Figure 2. Because all the customers have an equal service time, say$1/\mu$, a service completion takes place at every $1/\mu$ time units since the beginning of the busy period. If we hasten the admission time of a customer from time t_c to t_h ($t_h < t_c$), the queue size increases by 1 in the interval between t_h and t_c. Therefore, the total waiting time increases by $t_c - t_h$. **Q.E.D.**

Lemma 1 concerns waiting times within a single busy period. Now we want to relate the waiting time averaged over all customers with the waiting time averaged within individual busy periods. We denote by w_i the i-th customer's waiting time in queue.

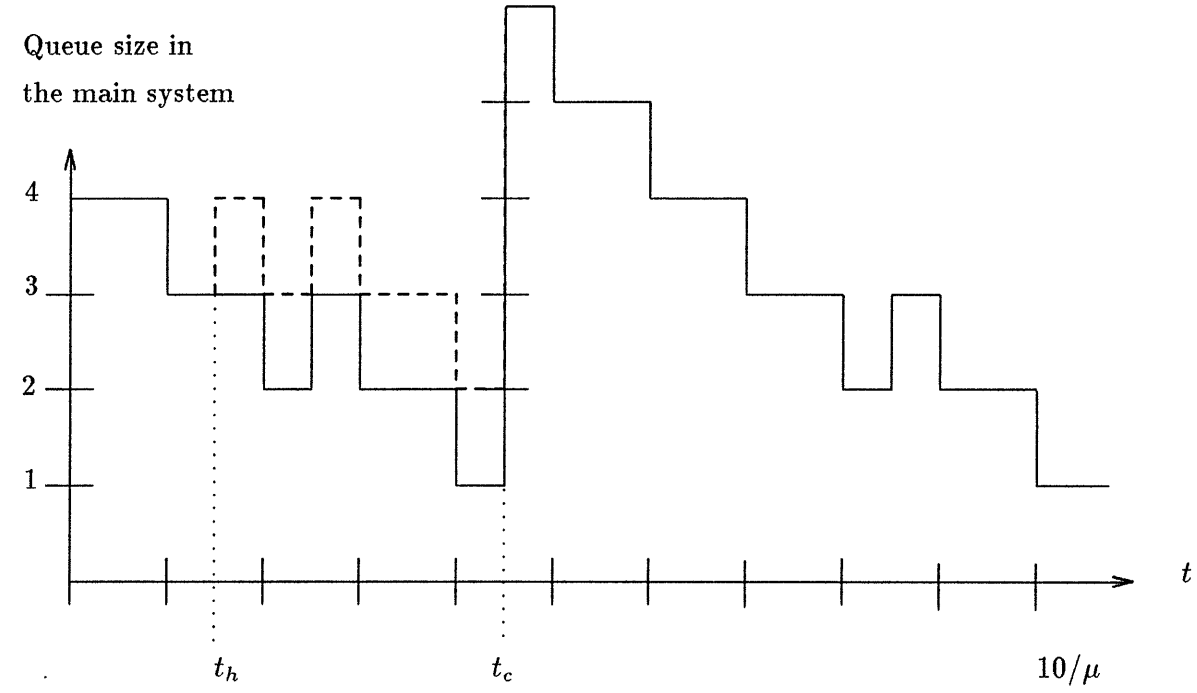


Figure 2: An admission hastened from t_c to t_h

In any admission schedule of a stable queueing system, the resulting sample path of the queue length will be a sequence of busy periods. We denote the number of admissions in the n-th busy period by a_n. We denoted by R_n the sum of the waiting times of these a_n customers. Then, the waiting time per customer averaged within the n-th busy period is R_n/a_n. The following lemma relates this quantity with the waiting time averaged over all customers.

Lemma 2 *For any input schedule, if the number of customers served in individual busy periods is bounded (i.e. $\{a_n | n = 1, 2, \cdots\}$ is bounded), and service times of customers are bounded, then we have*

$$\limsup_{M\to\infty} \frac{1}{M}\sum_{m=1}^{M} w_m \leq \sup_k \frac{R_k}{a_k}$$

Proof
Denote by $N(m)$ the number of completed busy periods up to the admission time of the m-th customer. Then, $\sum_{i=1}^{N(m)} a_i$ is the number of customers served in $N(m)$ busy periods. We then have

$$\begin{aligned}\frac{\sum_{m=1}^{M} w_m}{M} &= \frac{\sum_{m=1}^{N(M)} w_m}{M} + \frac{\sum_{m=N(M)+1}^{M} w_m}{M}\\ &\leq \frac{\sum_{m=1}^{N(M)} w_m}{\sum_{i=1}^{N(M)} a_i} + \frac{\sum_{m=N(M)+1}^{M} w_m}{M}\end{aligned}$$

Therefore,

$$\begin{aligned}\limsup_{M\to\infty} \frac{\sum_{m=1}^{M} w_m}{M} &\leq \limsup_{M\to\infty} \left[\frac{\sum_{m=1}^{N(M)} w_m}{\sum_{i=1}^{N(M)} a_i} + \frac{\sum_{m=N(M)+1}^{M} w_m}{M} \right] \\ &\leq \limsup_{M\to\infty} \frac{\sum_{m=1}^{N(M)} w_m}{\sum_{i=1}^{N(M)} a_i} + \limsup_{M\to\infty} \frac{\sum_{m=N(M)+1}^{M} w_m}{M}\end{aligned}$$

Since the number of customers served in a busy period is bounded, and service times are bounded, we have

$$\lim_{M\to\infty} \frac{\sum_{m=N(M)+1}^{M} w_m}{M} = 0 \qquad \text{and} \qquad \lim_{M\to\infty} N(M) = \infty$$

Therefore,

$$\limsup_{M\to\infty} \frac{\sum_{m=1}^{M} w_m}{M} \leq \limsup_{M\to\infty} \frac{\sum_{m=1}^{N(M)} w_m}{\sum_{i=1}^{N(M)} a_i} = \limsup_{N\to\infty} \frac{\sum_{n=1}^{N} R_n}{\sum_{n=1}^{N} a_n}$$

Hence,

$$\begin{aligned}\limsup_{M\to\infty} \frac{1}{M} \sum_{m=1}^{M} w_m \leq \limsup_{N\to\infty} \frac{\sum_{n=1}^{N} R_n}{\sum_{n=1}^{N} a_n} &= \limsup_{N\to\infty} \frac{\sum_{n=1}^{N} a_n \frac{R_n}{a_n}}{\sum_{n=1}^{N} a_n} \\ &\leq \limsup_{N\to\infty} \frac{\sum_{n=1}^{N} a_n \sup_k \frac{R_k}{a_k}}{\sum_{n=1}^{N} a_n} = \sup_k \frac{R_k}{a_k}\end{aligned}$$

Q.E.D.

This lemma enables us to focus on only one busy period in order to find the worst arrival schedule to a queueing system. Note that our model, $\cdot/D/1$ queue satisfies the assumptions of this lemma for any allowable set of arrivals, as long as the token generating rate is less than the service rate.

3 Single Source

In this section, the $\cdot/D/1$ queue illustrated in Figure 1 is analyzed. Each customer has a deterministic service time of length $1/\mu$. The arrival process in this queue satisfies a certain rate and burstiness constraint due to the preceding leaky bucket regulation. We are mainly interest in the effect of input rate and burstiness on the average queueing delay. The goal of the analysis in this section is to answer the following two questions:

- 1) What is the worst traffic of arrivals that can be allowed by the leaky bucket input regulation?
- 2) For this worst input, what is the relationship between the average queueing delay and the leaky bucket parameters?

Roughly speaking, our main result is that a certain periodic input schedule yields the worst average queueing delay. Each period is comprised of a big burst of size close to bucket size σ followed by a number of sequential admissions $1/r$ apart in time. We will also derive a closed-form expression for the average queueing delay as a function of σ and r. We assume $r \leq \mu$ for stability.

Consider a fictitious adversary who schedules admissions in order to maximize the average delay. Consider how this adversary will create a busy period that yields the maximal average queueing delay per customer within that busy period. For a fixed number, a of admissions in a busy period, the way to maximize the total waiting time under the leaky bucket regulation is to inject each customer as soon as possible after the first admission (due to Lemma 1). If $a \leq \sigma$, the way to maximize the total waiting time is to wait until the token bucket is full and to admit all a customers together. The resulting total waiting time is $a(a-1)/(2\mu)$, and the average queueing delay per customer in a busy period is $(a-1)/(2\mu)$. If $a = \sigma + 1$, the adversary can still push into the queue a bulk of customers of size $\sigma + 1$ in the following manner. The adversary can wait until the token bucket is full, and further wait until the next token generation time. Immediately prior to this token generation, the adversary can send σ customers at the leaky bucket and let them be admitted. Immediately after the token generation, the adversary can send another customer and let it be admitted. This way, one can start the busy period with $\sigma + 1$ customers, and the average waiting time is still $(a-1)/(2\mu) = \sigma/(2\mu)$. If $a = \sigma + 1 + k$, $k \geq 0$, in order to maximize the total queueing delay in a busy period, one must start a busy period with $\sigma + 1$ admissions, and at the following k token generation times, the remaining k customers must be admitted. (See Figure 3 for an example.) In this schedule, the i-th customer has waiting time $(i-1)/\mu$ for $1 \leq i \leq \sigma + 1$. For the $(\sigma + 1 + l)$-th customer, the waiting time is

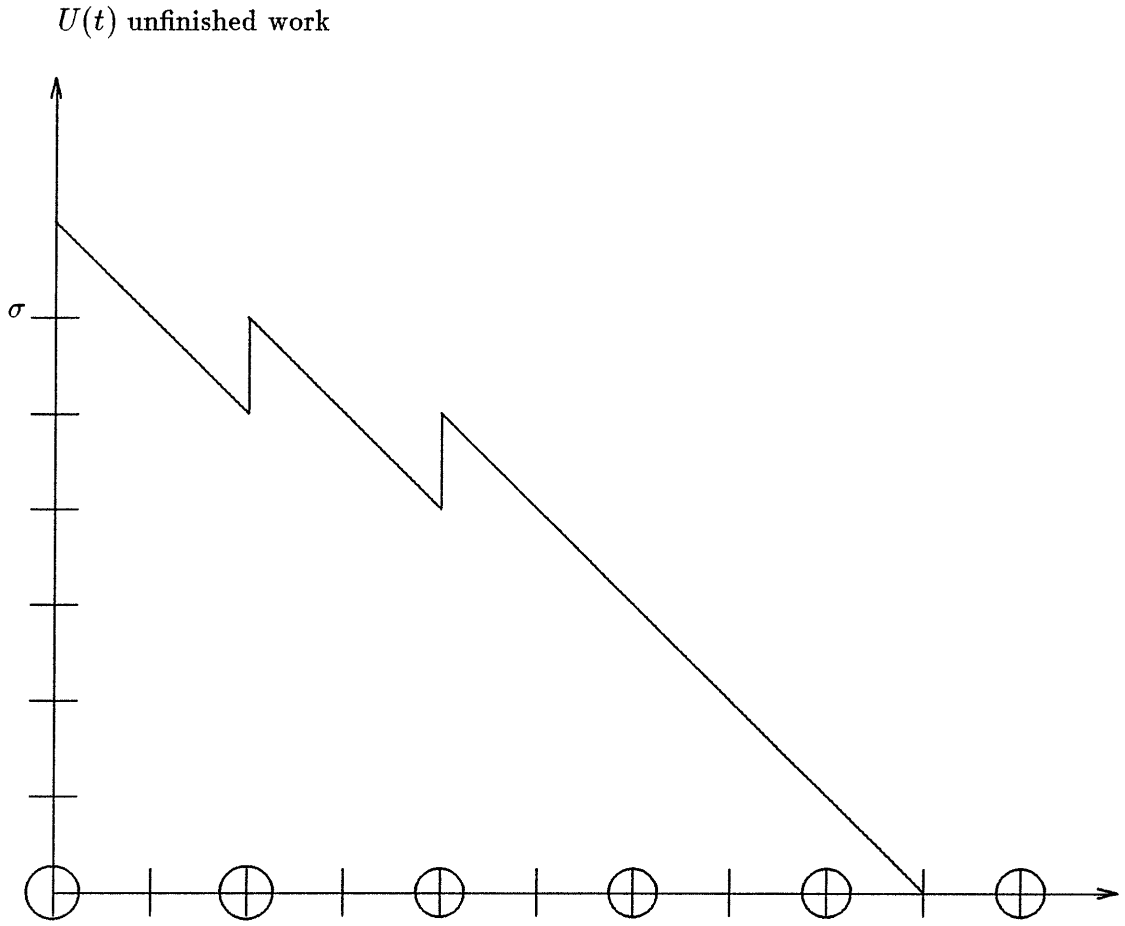
$$\sigma\frac{1}{\mu} + l(\frac{1}{\mu} - \frac{1}{r}) \quad \text{for} \quad l \leq \sigma\frac{r}{\mu - r}$$

The resulting average waiting time per customer in this busy period is

$$g(k,\sigma,r) \equiv \left[\frac{\sigma(\sigma+1)}{2}\frac{1}{\mu} + k\sigma\frac{1}{\mu} + \frac{k(k+1)}{2}(\frac{1}{\mu} - \frac{1}{r})\right]\frac{1}{\sigma+1+k} \qquad (1)$$

Notice that the average queueing delay within the busy period increases with a up to $a = \sigma + 1$. Therefore, the maximal queueing delay per customer in a busy period for the arrival schedules regulated by the leaky bucket scheme is

$$\max_{k=0,1,2,\cdots} g(k,\sigma,r)$$



○ stands for token generating epoch.

$$\frac{1}{r} = 2 \quad \sigma = 6 \quad \mu = 1 \quad k = 2$$

This figure illustrates how to maximize the total delay with a fixed number of admissions ($a = 9$ admissions for example).

Figure 3: Admission for maximal total waiting time

The maximal k for this function is

$$k^* = \begin{cases} 0 & \text{if } r \leq \mu/(1+\sigma) \\ k_l & \text{if } r > \mu/(1+\sigma) \text{ and } g(k_l,\sigma,r) \geq g(k_h,\sigma,r) \\ k_h & \text{if } r > \mu/(1+\sigma) \text{ and } g(k_l,\sigma,r) < g(k_h,\sigma,r) \end{cases} \tag{2}$$

where

$$k_l \equiv \left\lfloor -(\sigma+1) + \sqrt{\frac{\sigma(\sigma+1)}{1-r/\mu}} \right\rfloor$$

$$k_h \equiv \left\lceil -(\sigma+1) + \sqrt{\frac{\sigma(\sigma+1)}{1-r/\mu}} \right\rceil$$

See Appendix A for derivation. It turns out that $g(k^*,\sigma,r)$ is the worst queueing delay averaged over all customers. The following theorem states the result.

Theorem 3 *For any input schedule under leaky bucket regulation with parameter σ and r, the average queueing delay has the following upper bound:*

$$\limsup_{M\to\infty} \frac{1}{M}\sum_{m=1}^{M} w_m \leq g(k^*,\sigma,r) \tag{3}$$

This upper bound is attained by the admission pattern generated by the following algorithm:

Algorithm

1. *Wait until the token bucket is full; at time $1/r$ from the moment the token bucket is full, admit $\sigma+1$ customers.*
2. *At each of the next k^* token generation times, admit a customer.*
3. *Go to 1.*

Proof

The maximal queueing delay per customer within a busy period is $g(k^*,\sigma,r)$, and it is attained by first two lines of the Algorithm above. From Lemma 2, $g(k^*,\sigma,r)$ is an upper bound for the queueing delay averaged over all customers, and the Algorithm above attains this upper bound. **Q.E.D.**

Figures 4, 5 show the relationship between the worst average delay per customer, $g(k^*,\sigma,r)$ and the leaky bucket parameters for $\mu = 1$. Figure 5 indicates that the relation between the queueing delay and σ is very close to a linear relation. Let us compute the asymptotic slope. For sufficiently large σ, we have $r > \mu/(1+\sigma)$, so the ratio of the queueing delay to σ is

$$\frac{g(k^*,\sigma,r)}{\sigma} = \frac{1}{2}\frac{1}{\mu} - \frac{k^*}{2\sigma}\left(\frac{1}{r} - \frac{1}{\mu}\right) + \frac{1}{2r}\frac{k^*}{\sigma+1+k^*}$$

$\mu = 1$

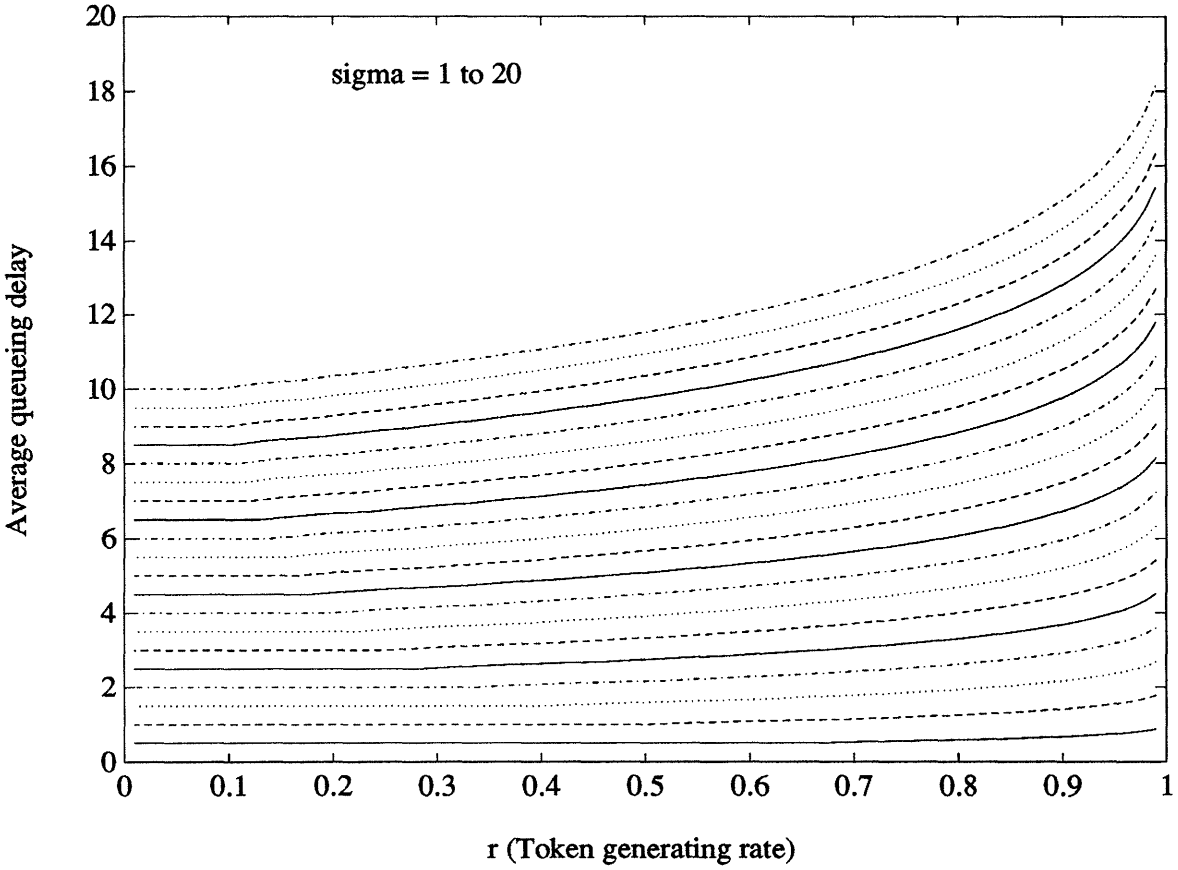


Figure 4: Average delay in the worst case vs. token generation rate

$\mu = 1$

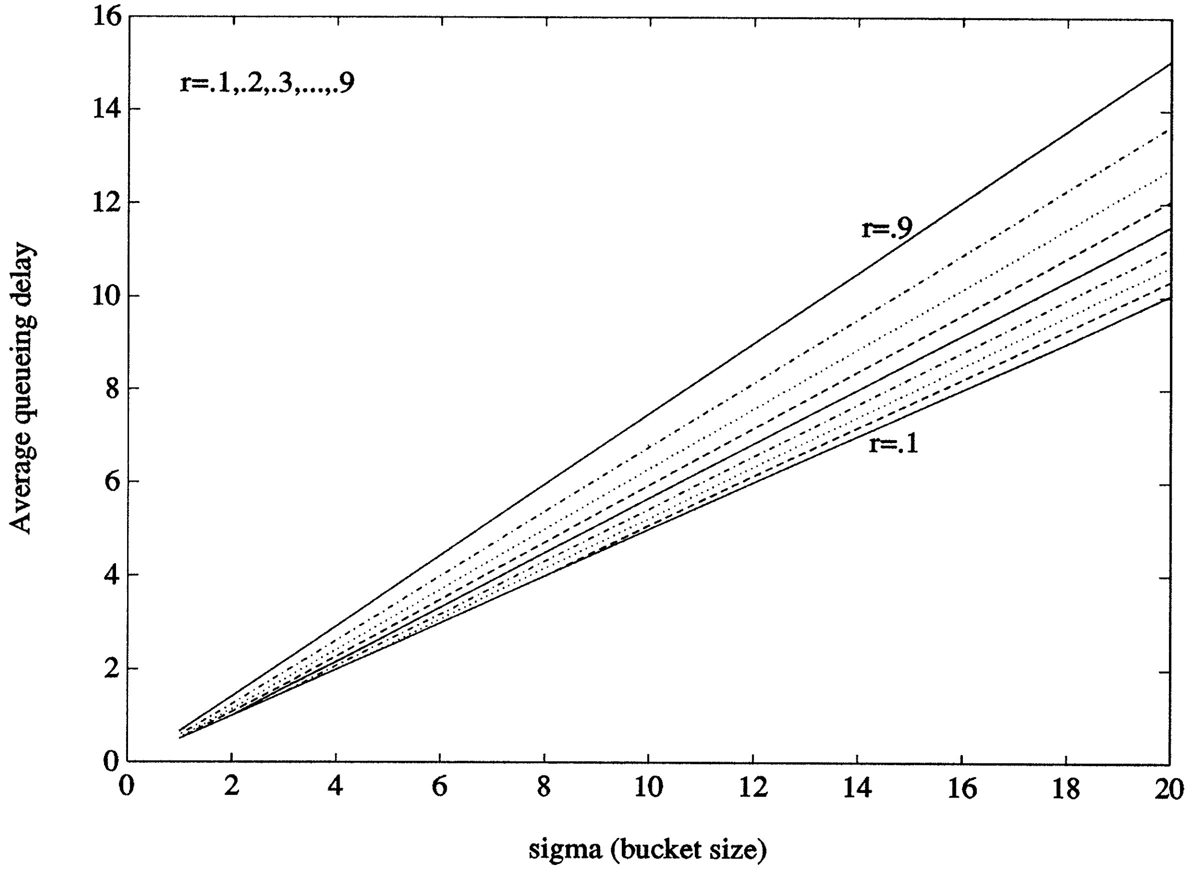


Figure 5: Average delay in the worst case vs. bucket size

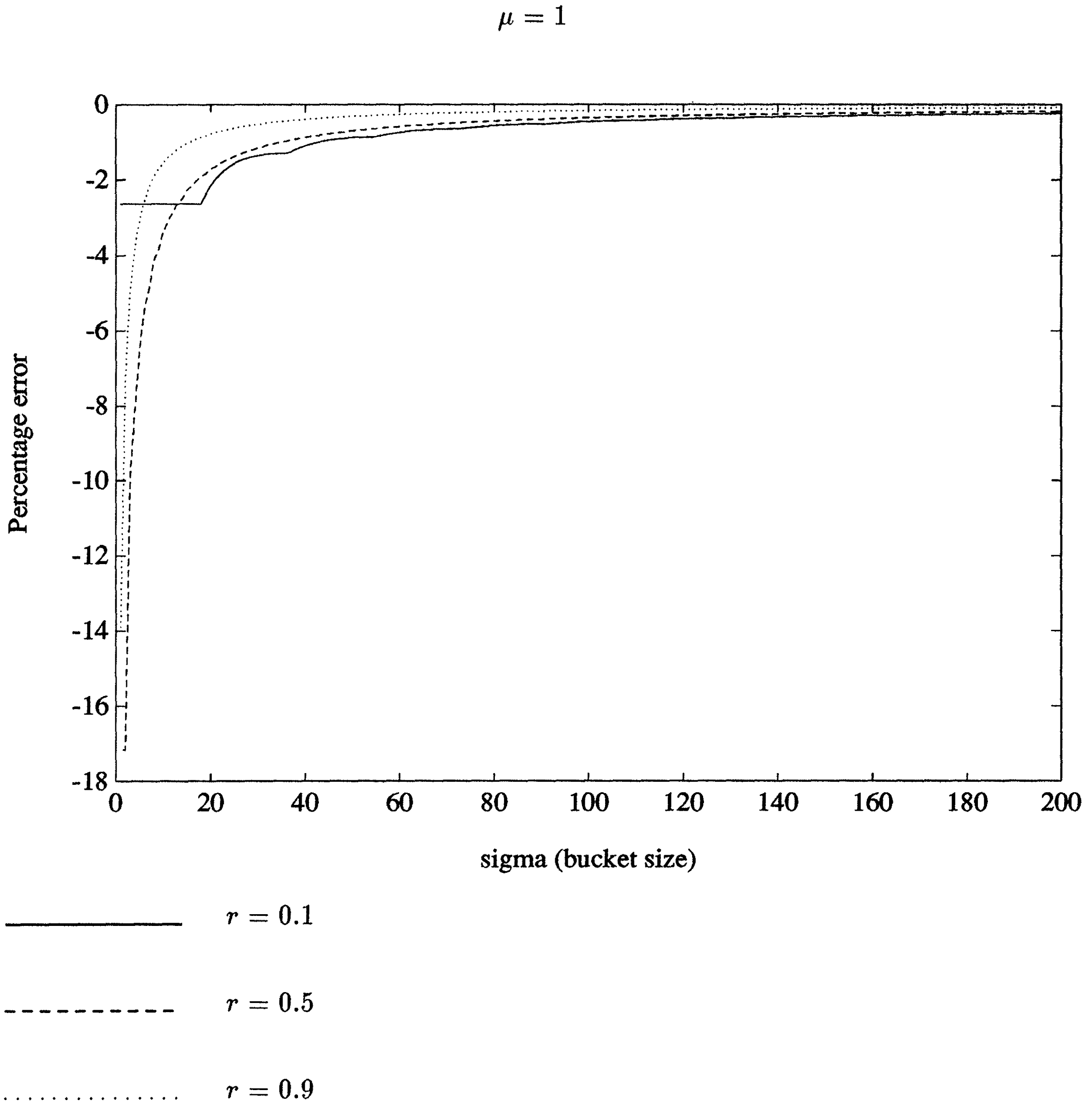


Figure 6: Percentage error

Also,

$$\lim_{\sigma\to\infty}\frac{k_l}{\sigma}=\lim_{\sigma\to\infty}\frac{k_h}{\sigma}=-1+\frac{1}{\sqrt{1-r/\mu}}=\lim_{\sigma\to\infty}\frac{k^*}{\sigma}$$

Finally, we have

$$\lim_{\sigma\to\infty}\frac{\max_k g(k,\sigma,r)}{\sigma} = \frac{1-\sqrt{1-r/\mu}}{r/\mu}\frac{1}{\mu} \tag{4}$$

Therefore, the queueing delay is approximated by the following expression:

$$\max_k g(k,\sigma,r) \simeq \frac{1-\sqrt{1-r/\mu}}{r/\mu}\frac{1}{\mu}\sigma \tag{5}$$

This equation indicates the effect of r, σ, and their interaction effect on the worst-case average delay. The percentage error of this approximation for $\mu = 1$,

$$\frac{\max_k g(k,\sigma,r)-\frac{1-\sqrt{1-r}}{r}\sigma}{\max_k g(k,\sigma,r)}$$

is plotted in Figure 6.

4 Multiple Sources

In the previous section, we considered the relationship between the worst-case average queueing delay and the leaky bucket parameters for a single source. In this section, we discuss this relationship for multiple sources. Each of S sources admits customers under a leaky bucket regulation with parameters σ/S and r/S. A new issue arises in the case of multiple sources: how to interleave the token generation times of different sources. In order to exclude the effect of fractional bucket size σ/S and to focus our attention on the effect of multiple sources, we assume that σ/S is an integer.

4.1 Perfectly Interleaved Token Generations

Suppose that the token generation times of S sources are perfectly interleaved, so that the time between the token generations of different sources is exactly $1/r$. In this case, obviously, the worst case average queueing delay is identical to the case of a single source.

4.2 Coinciding Token Generations

Suppose that all S sources generate tokens at the same time; therefore, S tokens are generated simultaneously every S/r time units. Consider the admission schedule generated by the following algorithm:

Algorithm 4

1. *Wait until all sources have a full token bucket.*

2. *Immediately prior to the next token generation, each source admits σ/S customers; immediately after this token generation, each source admits another customer.*

3. *Each source admits a customer at the next J token generation epochs; go to 1.*

For this admission pattern, the total waiting time per busy period is

$$\sum_{i=0}^{\sigma+S-1} i\frac{1}{\mu} + \sum_{j=1}^{J}\sum_{l=0}^{S-1}\left\{\sigma\frac{1}{\mu} + jS(\frac{1}{\mu}-\frac{1}{r}) + l\frac{1}{\mu}\right\}, \qquad \text{for such } J \text{ as } \sigma\frac{1}{\mu} + JS(\frac{1}{\mu}-\frac{1}{r}) \geq 0$$

The number of admitted customers in this busy period is $\sigma + S + JS$, so the average waiting time per customer in this busy period is

$$\begin{aligned} & h(J,\sigma,r,S) \\ \equiv\ & \frac{1}{\sigma+S+JS}\left[\sum_{i=0}^{\sigma+S-1} i\frac{1}{\mu} + \sum_{j=1}^{J}\sum_{l=0}^{S-1}\left\{\sigma\frac{1}{\mu} + jS(\frac{1}{\mu}-\frac{1}{r}) + l\frac{1}{\mu}\right\}\right] \\ =\ & J\frac{S}{2}(\frac{1}{\mu}-\frac{1}{r}) + \frac{\sigma+S-1}{2\mu} + \frac{\sigma}{2r} - \frac{\sigma}{2r}\left(\frac{\sigma+S}{\sigma+S+JS}\right) \qquad (6) \end{aligned}$$

Among the input patterns generated by Algorithm 4, let us consider which parameter J yields the maximal average queueing delay per customer in a busy period. By the procedure similar to the maximization of $g(k,\sigma,r)$, we can derive the maximum:

$$J^* = \begin{cases} 0 & \text{if } S/r \geq (\sigma+S)/\mu \\ J_l & \text{if } S/r < (\sigma+S)/\mu \text{ and } h(J_l,\sigma,r,S) \geq h(J_h,\sigma,r,S) \\ J_h & \text{if } S/r < (\sigma+S)/\mu \text{ and } h(J_l,\sigma,r,S) < h(J_h,\sigma,r,S) \end{cases} \qquad (7)$$

where

$$J_l \equiv -(\frac{\sigma}{S}+1) + \left\lfloor \sqrt{(\frac{\sigma}{S}+1)\frac{\sigma}{S}\frac{1}{1-r/\mu}} \right\rfloor \qquad (8)$$

$$J_h \equiv -(\frac{\sigma}{S}+1) + \left\lceil \sqrt{(\frac{\sigma}{S}+1)\frac{\sigma}{S}\frac{1}{1-r/\mu}} \right\rceil \qquad (9)$$

Theorem 5 *For S sources with coinciding token generation times with an overall rate r, and overall bucket size σ, the average queueing delay per customer has the following upper bound:*

$$\limsup_{M\to\infty} \frac{1}{M}\sum_{m=1}^{M} w_m \leq h(J^*,\sigma,r,S)$$

This upper bound is attained by Algorithm 4 *with parameter J^* of formula* (7) .

Proof
Lemma 1 states that for a fixed number of customers, say a, admissions must take place as soon as possible in order to maximize the total waiting time. Therefore, the adversary admits these a customers according to a strategy similar to Algorithm 4. Namely, wait until all the token buckets are full. Immediately prior to the next token generation time, admit $\min(a, \sigma)$ customers. If $\sigma < a < \sigma + S$, immediately after this token generation time, admit the rest $a - \sigma$ customers. If $a > \sigma + S$, immediately after this token generation time, admit S customers, ending up with $\sigma + S$ admissions at the beginning of the busy period. From there on, at every token generation time, admit up to S customers until all a customers are exhausted. We claim that the number of admissions that maximizes the average waiting time per customer satisfies

$$a^* = \sigma + S + JS \qquad \text{for some integer } J \ . \tag{10}$$

This is proved in Appendix B. Therefore, the maximal average waiting time per customer in a busy period is $h(J, \sigma, r, S)$ for some J. Hence, the maximal average waiting time per customer in a busy period is $\max_J h(J, \sigma, r, S) = h(J^*, \sigma, r, S)$. From Lemma 2, for any admission schedule,

$$\limsup_{M \to \infty} \frac{1}{M} \sum_{m=1}^{M} w_m \ \leq \ h(J^*, \sigma, r, S) \ ,$$

and this bound is attained by Algorithm 4. **Q.E.D.**

4.3 General Token Generating Patterns

Recall function $g(k, \sigma, r)$ defined in (1) and maximum, k^* of formula (2).

Theorem 6 *The average queueing delay in the worst case under any token generation pattern lies between the worst-case bounds for the 'perfectly interleaved' and the 'coinciding' token generation patterns. That is,*

$$g(k^*, \sigma, r) \leq \liminf_{M \to \infty} \frac{1}{M} \sum_{m=1}^{M} w_m \leq \limsup_{M \to \infty} \frac{1}{M} \sum_{m=1}^{M} w_m \leq h(J^*, \sigma, r, S)$$

Proof
See Appendix C.

5 Comparison between Multiple Sources and Single Source

We have established that multiple sources exhibit worse delay performance than a single source except in the case of a perfectly interleaved token generating pattern.

For this pattern, the performance is identical to the single source. Now we show that the asymptotic growth of the maximal average queueing delay as a function of σ is independent of the number of sources. From equations (6), (8) and (9),

$$\lim_{\sigma \to \infty} \frac{J_l}{\sigma} = \lim_{\sigma \to \infty} \frac{J_h}{\sigma} = \frac{1}{S}\left(-1 + \frac{1}{\sqrt{1 - r/\mu}}\right) , \qquad \text{so}$$

$$\begin{aligned}
&\lim_{\sigma \to \infty} \frac{\max_J h(J, \sigma, r, S)}{\sigma} \\
= \;& \frac{1}{2}(\frac{1}{\mu} - \frac{1}{r})\left(-1 + \frac{1}{\sqrt{1 - r/\mu}}\right) + \frac{1}{2}(\frac{1}{\mu} + \frac{1}{r}) - \frac{1}{2r}\sqrt{1 - r/\mu} \\
= \;& \frac{1 - \sqrt{1 - r/\mu}}{r/\mu}\frac{1}{\mu}
\end{aligned}$$

This quantity is identical to formula (4) of the single source. Therefore, the asymptotic growth does not depend upon the number of sources.

6 Conclusion

We have studied the performance of a single-server queue with deterministic service time where arrivals are regulated by a leaky bucket scheme. We have discussed the cases of both a single source of arrivals and multiple sources. In both cases, we have characterized the worst arrival pattern that passes through the leaky bucket regulation. In the case of a single arrival source, we have also specified the average queueing delay per customer as a function of the leaky bucket parameters.

For a leaky bucket regulation with bucket size σ and token generation rate r, the arrival pattern that maximizes the average queueing delay per customer is characterized as the repetition of the following three phases: bulky admission with bulk size related to σ, admission at every token generation for a specified length of interval, and then no admission till the token bucket is full. For the case of the single source, the maximal average queueing delay is closely approximated by

$$\frac{1 - \sqrt{1 - r/\mu}}{r/\mu}\frac{1}{\mu}\sigma$$

For the case of multiple sources, the arrangement of the token generation times affects the delay averaged over all customers. The worst arrangement is when the token generation times for all sources coincide. The best arrangement is when the token generations for different sources are perfectly interleaved.

A Derivation of $\max g(k, \sigma, r)$

Let us extend the function g for real values of k and consider the partial derivative with respect to k. We have

$$\frac{\partial g}{\partial k} = \frac{1}{2} \frac{(\frac{1}{\mu} - \frac{1}{r})}{(\sigma + 1 + k)^2} \left[k^2 + 2(\sigma+1)k + (\sigma+1)^2 + \frac{\sigma(\sigma+1)}{r/\mu - 1} \right]$$

The roots of this partial derivative are

$$-(\sigma+1) \pm \sqrt{\frac{\sigma(\sigma+1)}{1 - r/\mu}}$$

Function $g(k, \sigma, r)$ is nondecreasing in k for,

$$k \in \left[\ -(\sigma+1) - \sqrt{\frac{\sigma(\sigma+1)}{1 - r/\mu}} \ , \ -(\sigma+1) + \sqrt{\frac{\sigma(\sigma+1)}{1 - r/\mu}} \ \right] \ ,$$

and nonincreasing in k for

$$k \in \left(\ -(\sigma+1) + \sqrt{\frac{\sigma(\sigma+1)}{1 - r/\mu}} \ , \ \infty \ \right)$$

For the case, $r \leq \mu/(1+\sigma)$, we have

$$-(\sigma+1) + \sqrt{\frac{\sigma(\sigma+1)}{1 - r/\mu}} \leq 0 \ ,$$

so $g(k, \sigma, r)$ on nonnegative integer domain is maximized at $k = 0$. For the case $r > \mu/(1+\sigma)$,

$$-(\sigma+1) + \sqrt{\frac{\sigma(\sigma+1)}{1 - r/\mu}} > 0 \ ,$$

so $g(k, \sigma, r)$ on nonnegative integer domain is maximized either at

$$\begin{aligned} k_l &\equiv \left\lfloor \ -(\sigma+1) + \sqrt{\frac{\sigma(\sigma+1)}{1 - r/\mu}} \ \right\rfloor \qquad \text{or} \\ k_h &\equiv \left\lceil \ -(\sigma+1) + \sqrt{\frac{\sigma(\sigma+1)}{1 - r/\mu}} \ \right\rceil \end{aligned}$$

Hence,

$$k^* = \begin{cases} 0 & \text{if } \ r \leq \mu/(1+\sigma) \\ k_l & \text{if } \ r > \mu/(1+\sigma) \ \text{ and } \ g(k_l, \sigma, r) \geq g(k_h, \sigma, r) \\ k_h & \text{if } \ r > \mu/(1+\sigma) \ \text{ and } \ g(k_l, \sigma, r) < g(k_h, \sigma, r) \end{cases}$$

Q.E.D.

B Proof of formula (10)

Because a^* maximizes the average waiting time per customer in a busy period,

$$\frac{\sum_{i=1}^{a^*-1} w_i}{a^*-1} \leq w_{a^*} \tag{11}$$

Suppose $a^* = \sigma + S + (J-1)S + l$, $1 \leq l \leq S-1$ as the example in Figure 7. Then, we can admit another customer at the same time as the a^*-th customer, and $w_{a^*+1} = w_{a^*} + (1/\mu)$. We now compare the waiting time the (a^*+1)-st customer would have, w_{a^*+1} and the waiting time averaged up to a^*-th customer, $(\sum_{i=1}^{a^*} w_i)/a^*$. We have

$$\begin{aligned} w_{a^*+1} - \frac{\sum_{i=1}^{a^*} w_i}{a^*} &= w_{a^*} + \frac{1}{\mu} - \frac{\sum_{i=1}^{a^*} w_i}{a^*} \\ &= \frac{(a^*-1)w_{a^*} - \sum_{i=1}^{a^*-1} w_i}{a^*} + \frac{1}{\mu} \\ &= \frac{a^*-1}{a^*}[w_{a^*} - \frac{\sum_{i=1}^{a^*-1} w_i}{a^*-1}] + \frac{1}{\mu} \\ &> 0 \quad \text{using inequality (11)} \end{aligned}$$

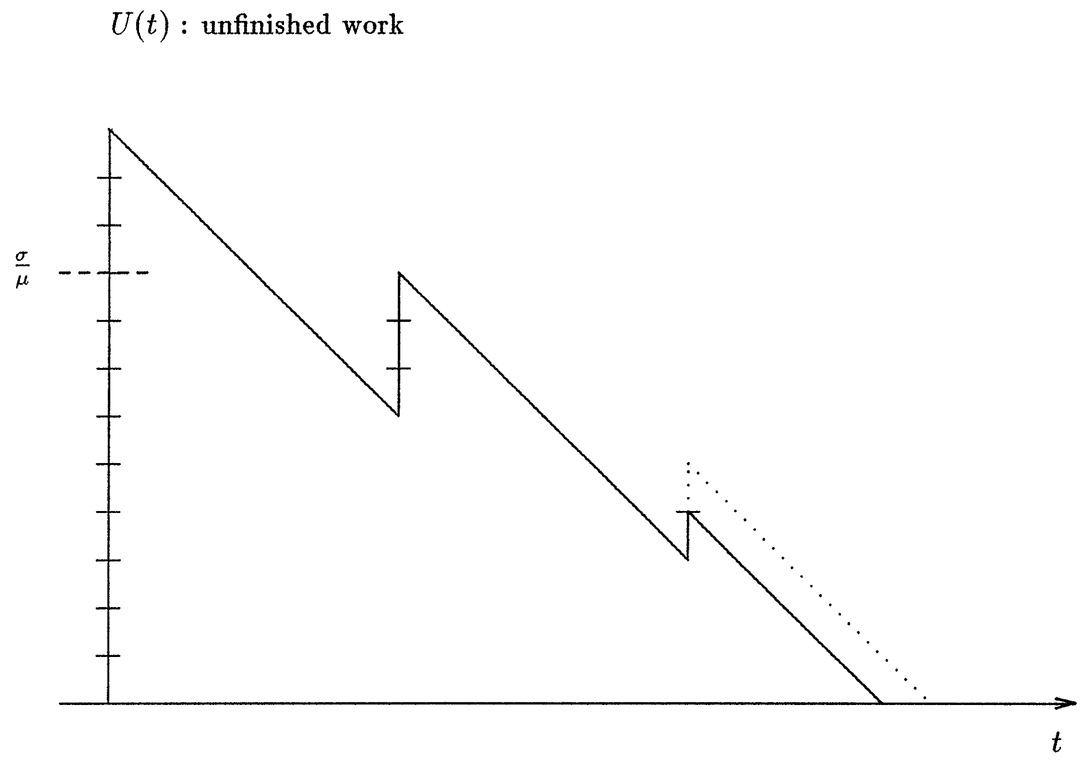
This implies that

$$\frac{\sum_{i=1}^{a^*} w_i}{a^*} < \frac{\sum_{i=1}^{a^*+1} w_i}{a^*+1}$$

This contradicts the maximality of a^*. Therefore, $a^* = \sigma + S + JS$ for some integer J. **Q.E.D.**

C Proof of Theorem 6

Suppose that we try to make the total waiting time of a busy period as large as possible with a fixed number of admissions, a. From Lemma 1, for any token generating pattern, we will wait until all the sources have a full token bucket. That way, we can admit at least σ customers at the beginning of the busy period. For a coinciding pattern, we can admit S more customers at the beginning of the busy period by admitting the σ customers immediately prior to the token generation and S customers immediately after the token generation. For any token generation pattern, one and only one token generation for each source happens in the interval $[t_b, t_b + S/r)$, where t_b is the time at which busy period starts with at least σ admissions. Therefore, the S customers for the case of the coinciding pattern admitted at the beginning of the busy period in addition to the σ customers can be viewed as admitted earlier than S customers admitted in the interval $[t_b, t_b + S/r)$ for the case of an arbitrary token generation pattern. The same statement holds true for subsequent bulky admissions of



$\sigma = 9, S = 3, J = 2, l = 1$

Figure 7: Algebraic structure of the maximal number of admissions

the coinciding token generation pattern and the admissions of the subsequent intervals of length S/r for the case of an arbitrary token generation pattern. (See Figure 8.) Therefore, for any fixed number of admissions, a, the coinciding pattern ends up with the largest total waiting time in one busy period. Therefore, the coinciding pattern gives rise to the largest queueing delay per customer in one busy period. Therefore, using Lemma 2, we prove

$$\limsup_{M\to\infty} \frac{1}{M} \sum_{m=1}^{M} w_m \leq h(J^*, \sigma, r, S)$$

For a perfectly interleaved token generation, tokens for each source are generated with period S/r. Since the token generations for S sources are perfectly interleaved, tokens are generated with period $1/r$. Therefore, the maximal average waiting time per customer is identical to the single source case. Let us compare this system with an arbitrarily interleaved token generation. Without loss of generality, let us say that a token for source 0 is generated at time 0, and that the following token generation for source $1, 2, \cdots, S-1$ occurs at time $0 \leq t_1 \leq t_2 \leq \cdots \leq t_{S-1} < S/r$, respectively. Each source generates a token with period S/r, so source k generates tokens at

$$m\frac{S}{r} + t_k \qquad m = 0, 1, 2, \cdots$$

Without loss of generality, source k of the system with perfectly interleaved token generation generates a token at times

$$m\frac{S}{r} + \frac{k}{r} \qquad m = 0, 1, 2, \cdots$$

Note that for a fixed number of admissions within a busy period, in order to maximize the total waiting time, the busy period should start after all sources have a full token bucket. Define $\tau_k \equiv t_k - k/r$. Suppose $\tau_k \leq 0$ for $k = 0, 1, \cdots, S-1$. Then, if we start busy periods for both token generating patterns at time 0, all the admissions of this arbitrarily interleaved token generation system can be viewed as the hastened admission from the perfectly interleaved one. Suppose $\tau_k > 0$ for some k. Take the largest τ_k and define

$$\tau_{k^*} \geq \tau_k \qquad \text{for all } k$$

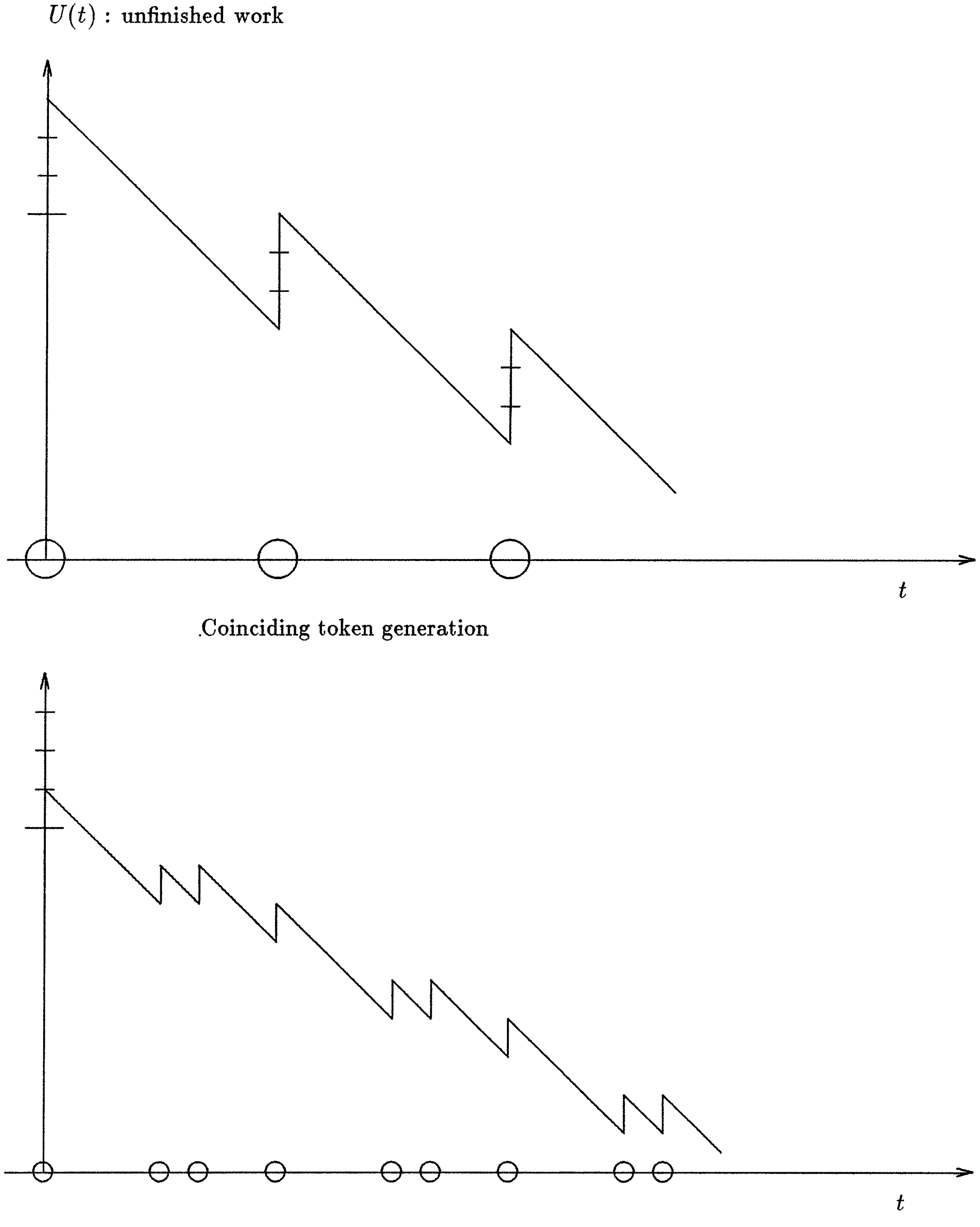
If we start busy periods for both token generating patterns at time t_{k^*}, an admission at any subsequent token generation of this arbitrarily interleaved token generating system can be viewed as a hastened admission from the 'perfectly interleaved' system. This statement can be explained pictorially in Figure 9. For the purpose of comparison with the 'perfectly interleaved' system, starting a busy period at time t_{k^*} is viewed as shifting the time axis of the system with an arbitrary token generating pattern so that t_{k^*} coincides with a token generation time of the 'perfectly interleaved' system. Due to this shift and the periodic nature of the token generation, the time difference between the token generation time for each source k for the system with an arbitrary token generation pattern and the corresponding token generation time of the 'perfectly interleaved' system is

$$(t_k - \tau_{k^*}) - \frac{k}{r} = \tau_k - \tau_{k^*} \le 0$$

Therefore, the subsequent admissions of the arbitrarily interleaved system can be viewed as hastened admissions of the perfectly interleaved one. Therefore, for each fixed number of admissions, a, the total waiting time in a busy period for the arbitrarily interleaved system is no smaller than the perfectly interleaved one from Lemma 1. Therefore, using Lemma 2, we prove that the maximal average waiting time of an arbitrarily interleaved token generating pattern is no smaller than the one from the perfectly interleaved system. Therefore,

$$g(k^*, \sigma, r) \le \liminf_{M \to \infty} \frac{1}{M} \sum_{m=1}^{M} w_m$$

Q.E.D.



.Coinciding token generation

Arbitrarily interleaved token generation

Figure 8: Comparison between token generation time arrangements

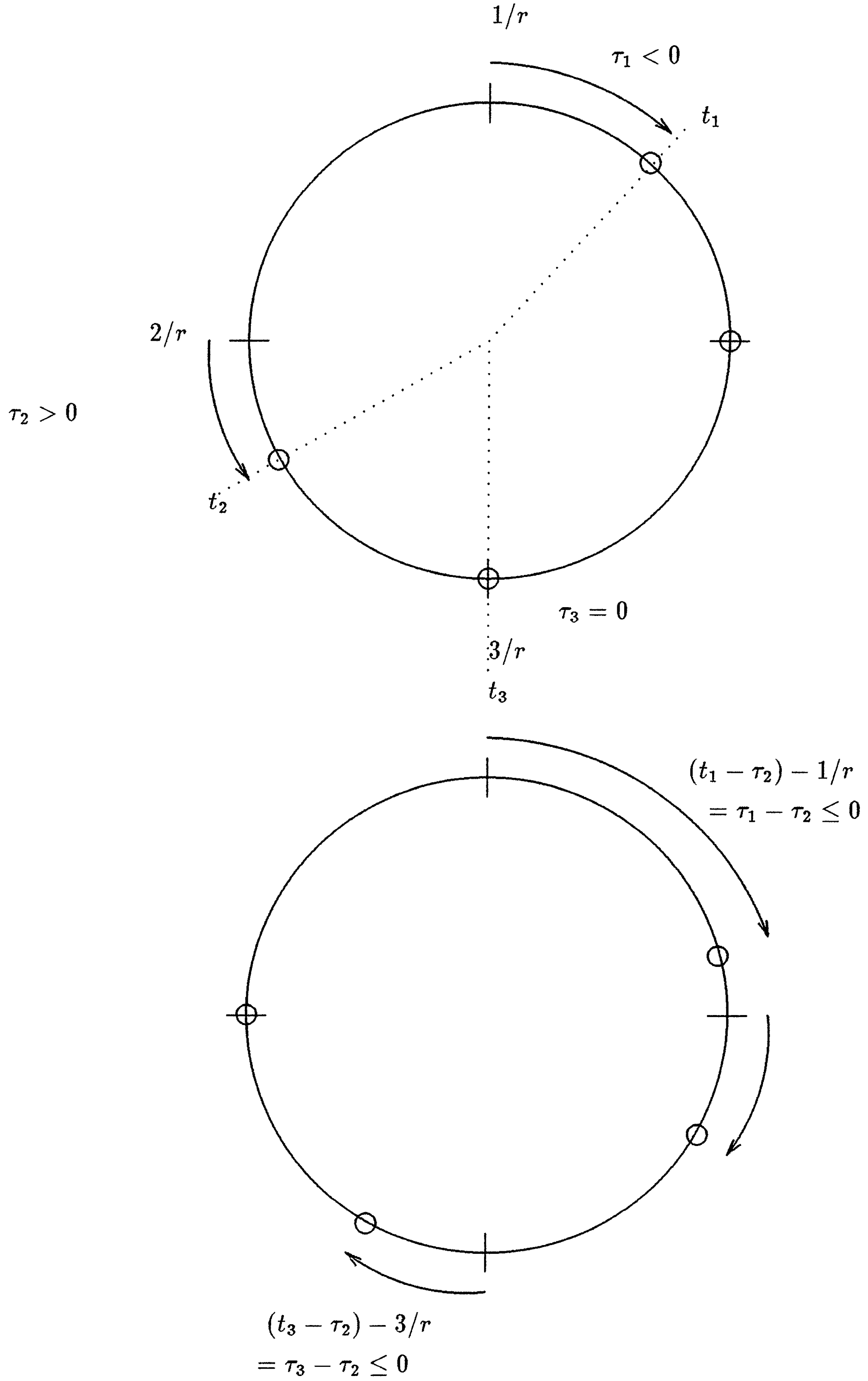


Figure 9: Circular shift of an arbitrarily interleaved token generation

References

[1] Arthur W. Berger. Performance analysis of a rate-control throttle where tokens and jobs queue. *IEEE Journal on Selected Areas in Communications*, 9(2):165–170, February 1991.

[2] Arthur W. Berger and Ward Whitt. A multi-class input-regulation throttle. In *Proceedings of the 29th IEEE Conference on Decision and Control*, pages 2106–2111, Honolulu, Hawaii, December 1990.

[3] Rene L. Cruz. Calculus for network delay – part I: Network elements in isolation. *IEEE Transactions on Information Theory*, 37(1):114–131, January 1991.

[4] R.L. Cruz. Calculus for network delay – part II: Network analysis. *IEEE Transactions on Information Theory*, 37(1):132–141, January 1991.

[5] A.I. Elwalid and D. Mitra. Rate-based congestion control. *Queueing Systems*, 9, 1991.

[6] A. Parekh. *A generalized processor sharing approach to flow control in integrated service networks.* PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, 1992. Dept. of Electrical Engineering and Computer Science.

[7] M. Sidi, W.Z. Liu, I. Cidon, and I. Gopal. Congestion control through input rate regulation. In *Proceedings of GLOBECOM'89*, Dallas, TX, 1989. volume 3.

[8] Jonathan S. Turner. New directions in communications (or which way to the information age?). *IEEE Communications Magazine*, October 1986.