MEF Standard
MEF 41.0.1

Amendment to MEF 41: Clarification of Generic Token Bucket Algorithm (GTBA) Behavior

July 2020

Disclaimer

The information in this publication is freely available for reproduction and use by any recipient and is believed to be accurate as of its publication date. Such information is subject to change without notice and MEF Forum (MEF) is not responsible for any errors. MEF does not assume responsibility to update or correct any information in this publication. No representation or warranty, expressed or implied, is made by MEF concerning the completeness, accuracy, or applicability of any information contained herein and no liability of any kind shall be assumed by MEF as a result of reliance upon such information.

The information contained herein is intended to be used without modification by the recipient or user of this document. MEF is not responsible or liable for any modifications to this document made by any other party.

The receipt or any use of this document or its contents does not in any way create, by implication or otherwise:

a) any express or implied license or right to or under any patent, copyright, trademark or trade secret rights held or claimed by any MEF member which are or may be associated with the ideas, techniques, concepts or expressions contained herein; nor

b) any warranty or representation that any MEF members will announce any product(s) and/or service(s) related thereto, or if such announcements are made, that such announced product(s) and/or service(s) embody any or all of the ideas, technologies, or concepts contained herein; nor

c) any form of relationship between any MEF member and the recipient or user of this document.

Implementation or use of specific MEF standards, specifications, or recommendations will be voluntary, and no Member shall be obliged to implement them by virtue of participation in MEF Forum. MEF is a non-profit international organization to enable the development and worldwide adoption of agile, assured and orchestrated network services. MEF does not, expressly or otherwise, endorse or promote any specific products or services.

Table of Contents

List of Figures

List of Tables

1 List of Contributing Members

The following members of the MEF participated in the development of this document and have requested to be included in this list.

- AT&T
- Bell Canada
- Canoga Perkins
- CenturyLink
- Cisco
- Nokia

2 Abstract

This standard amends MEF 41 [1] to describe how differences in sequences of Token Requests can affect the application of the GTR_{max}^{i} (and YTR_{max}^{i}) parameters in the Generic Token Bucket Algorithm, and how this affects the overall rate of tokens shared among Ranks and hence the overall rate of Token Requests declared Green (or Yellow) at each Rank.

3 Introduction

When the Generic Token Bucket Algorithm (GTBA) is applied by Ingress or Egress Bandwidth Profiles in a Carrier Ethernet Service, one might expect that when the average rate of traffic arrival for a set of several frames at one Rank in an envelope is less than the committed ("Green") rate for that Rank, all of the frames in the set would be declared Green. Depending on frame arrival intervals, however, the amount of traffic declared Green can be less due to the application of the CBS^i (GTV^i) parameter. Furthermore one might expect that when $CF^i = 0$ and the rate of traffic arriving at one Rank in an envelope is less than the committed ("Green") rate for that Rank, then the difference between the two rates would be shared to the next lower Rank in the envelope, up to the limit imposed by the CIR^i_{max} (GTR^i_{max}) parameter. Depending on frame arrival intervals, however, the rate shared to the lower Rank may be less than this intuitive expectation.

This amendment to MEF 41 [1] describes these behaviors in order to set expectations to match actual Token Request color declaration rates, and provides informative text to describe:

- how differences in Token Request arrival intervals can affect the application of the GTR^i_{max} (and YTR^i_{max}) parameters in the Generic Token Bucket Algorithm,
- how this affects the overall rate of tokens shared from one Rank to other Ranks, and
- how bucket sizes and other factors might affect the behavior of the algorithm.

This information includes equations to provide upper bound and lower bound on the rate of tokens shared.

The equations in Section 6 (MEF 41 Section 9) are modified to separate tokens that Bypass the token bucket (due to the limits imposed by GTR^i_{max} and YTR^i_{max}) from tokens that Overflow the token bucket (due to the limits imposed by GTV^i and YTV^i). The sum of the tokens that Bypass or Overflow the token bucket is equal to what is considered "Overflow" in MEF 41. This represents a change in the description of the GTBA, but not a change in the behavior. In particular, for any Token Request Flow (TRF) there is no change to the number of tokens added to a token bucket, converted from the Green token bucket to the Yellow token bucket, or shared to another Rank.

In this amendment, changes are shown as follows:

- Instructions for how to apply the amendment are shown in *blue italics*
- In MEF 41 sections being modified, text to be removed is shown with red strikethrough
- In MEF 41 sections being modified, text to be added is shown in red

4 Changes to Section 3, Terminology and Acronyms

The following terms in Table 1 are changed as shown:

Term	Definition	Reference
CF_0	A GTBA parameter that controls the conversion of Rank 1 Bypass and Overflow of rank 1 Green tokens to rank Rank *n* Yellow tokens.	This document
CF_i	A GTBA parameter that controls the destination of whether Bypass and Overflow TRF Rank *i* Green tokens become Rank *i-1* Green tokens or Rank *i* Yellow tokens.	This document
Burst	Given a sequence of frames, each Burst in the sequence is a maximal subsequence of one or more consecutive frames where at every point in time during the subsequence the average byte rate of the frames since the beginning of the subsequence exceeds a reference rate.	This document
Burst Magnitude	The maximum amount by which the cumulative byte count of the frames since the beginning of a Burst exceeds the reference rate multiplied by the time since the beginning of the Burst.	This document
Bypass	The action of the GTBA when tokens are not put into a token bucket because of the maximum token rate limit, regardless of whether or not the token bucket is full.	This document
Constant Bypass	Bypass that always occurs because the values of the GTBA parameters always result in exceeding the maximum token rate limit regardless of the sequence of Token Requests for any TRF.	This document
GTR_{max}^{i}	A GTBA parameter that represents the limit on the sum of the replenishing rate of new Green tokens, Bypass Green tokens, and the Overflow Green tokens for TRF *i*.	This document
Overflow	The action of the GTBA when tokens are not put into a token bucket because the token bucket is full.	This document
Transient Bypass	Bypass that is not Constant Bypass, and is dependent on the sequence of Token Requests for any TRF.	This document
YTR_{max}^{i}	A GTBA parameter that represents the limit on the sum of the replenishing rate of new Yellow tokens, Bypass Yellow tokens, and the Overflow Yellow tokens, and tokens converted from the Green token bucket for TRF *i*.	This document

Table 1 – Terminology and Acronyms

5 Changes to Section 8, Algorithm Parameters

Modify section 8 as follows:

8 Algorithm Parameters

There are two kinds of parameters that control the behavior of GTBA: General Parameters and TRF Parameters as described below.

8.1 General Parameters

There are two General Parameters:

- n a positive integer representing the number of TRFs and
- CF^0 can be 0 or 1 and controls the conversion of Bypass and Overflow of Rank 1 Green tokens to Rank n Yellow tokens.[1]

[R1] Values of n and CF^0 **MUST** be specified for each instance of a GTBA.

[R2] When $n = 1$, CF^0 **MUST** have the value of 0.

8.2 TRF Parameters

Let the n TRFs be ranked $i = 1,2,\dots,n$ where n is the highest Rank. Then for each $i \in \{1,2,\dots,n\}$, there are the following TRF Parameters:

- GTR^i a non-negative number in units of tokens per unit time that represents the replenishing rate of new Green tokens for TRF i,
- GTR^i_{max} a non-negative number in units of tokens per unit time that represents the limit on the sum of the replenishing rates of new Green tokens, Bypass Green tokens, and Overflow Green tokens for TRF i,
- GTV^i a non-negative number in units of tokens that represents the upper limit on the Green token bucket count for TRF i,
- YTR^i a non-negative number in units of tokens per unit time that represents the replenishing rate of new Yellow tokens for TRF i,
- YTR^i_{max} a non-negative number in units of tokens per unit time that represents the limit on the sum of the replenishing rates of new Yellow tokens, Bypass Yellow tokens, Overflow Yellow tokens, and tokens converted from the Green token bucket for TRF i,
- YTV^i a non-negative number in units of tokens that represents the upper limit on the Yellow token bucket count for TRF i, and
- CF^i can be 0 or 1 and controls the destination of whether Bypass and Overflow TRF Rank i Green tokens become Rank *i-1* Green tokens or Rank *i* Yellow tokens.

[R3] If $CF^0 = 1$, then CF^i **MUST** equal 0 for $i = 1,2,\dots,n$.

[1] See Figure 3 and Section 9 for the details of how CF^0 is used.

Figure 3 shows a conceptual diagram of the token flows with three TRFs.

- The solid arrows green and yellow flag-shaped icons labeled GTR^i or YTR^i represent a token source flowing into each token bucket new token sources for the TRF.
- The red funnel-shaped icons labeled GTR^i_{max} and YTR^i_{max} represent the limits on how fast tokens can be added to each token bucket.
- The green and yellow trapezoids trapezoid icons labeled GTV^i and YTV^i represent the Green and Yellow token bucket for each TRF.
- The dashed arrows show where possible token paths. An arrow that originates at a funnel represents tokens that Bypass the token bucket due to GTR^i_{max} and YTR^i_{max}. An arrow that originates at a trapezoid represents tokens that Overflow each the token bucket go when the number of tokens in the token bucket reaches the capacity (GTV^i or YTV^i).
- The circles ellipses represent summation points and/or decision points in the token path. Ellipses with two input arrows and a "+" symbol sum the tokens arriving on the two paths. Ellipses with two exit arrows represent a decision point controlled by the Coupling Flag (CF^i).
- The X's represent points at which tokens are discarded (i.e., not added to any of the token buckets).
- ~~GTR^i_{max}~~ and ~~YTR^i_{max}~~ are not shown on the figure.

Replace the existing Figure 3 with the following figure:

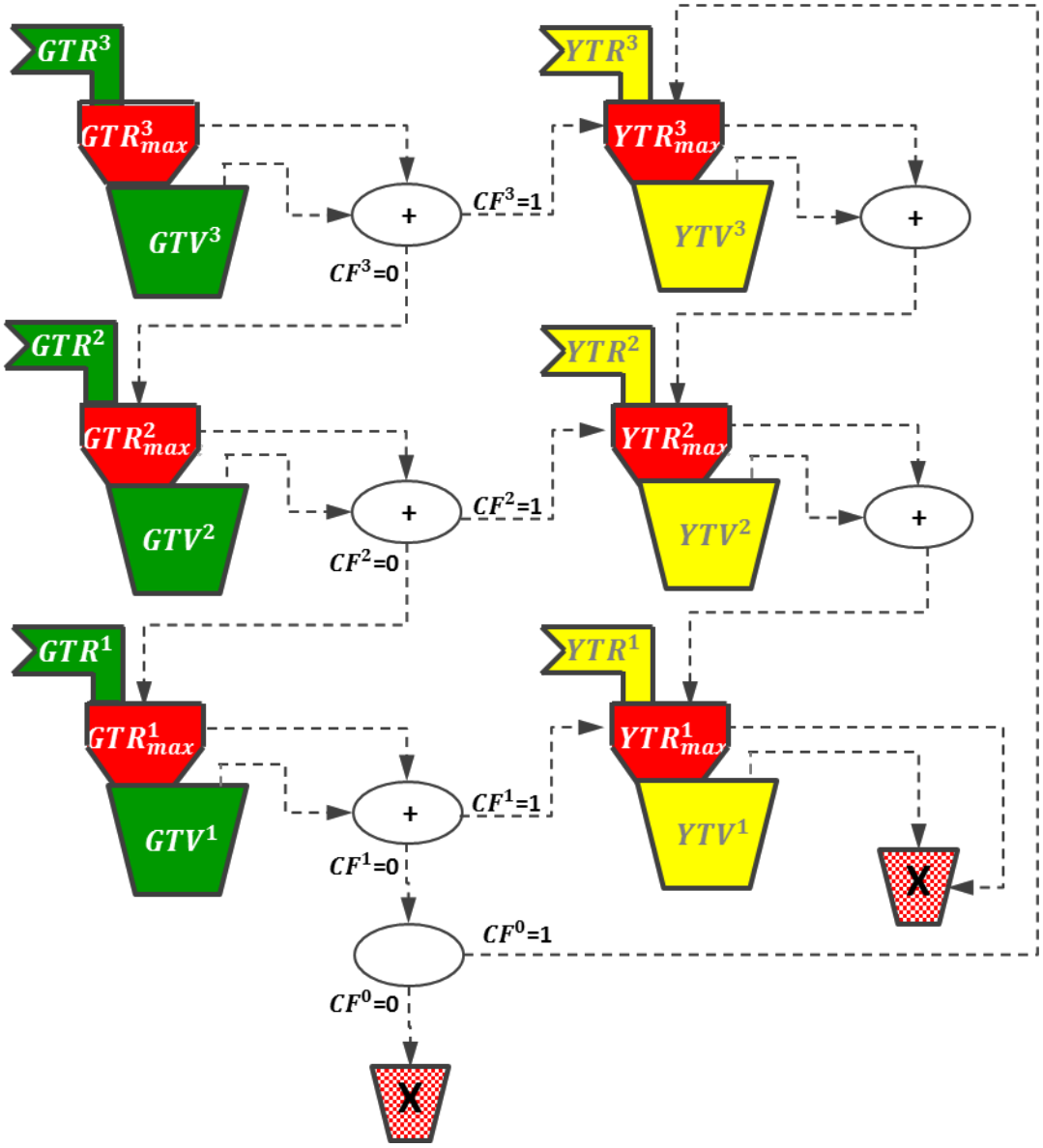

Figure 3 – Generic Token Bucket Algorithm Token Flows

6 Changes to Section 9, Token Count Updating

Modify section 9 as follows:

9 Token Count Updating

This Section specifies how the token bucket counts for each TRF are calculated. The token bucket counts for all TRFs are updated in response to each Token Request for any TRF. The times t_1 and t_2 represent the time of occurrence of consecutive Token Requests, regardless of the Rank of those Token Requests.

$C_G^i(t)$ represents the Green token bucket count at time t for TRF i, $i = 1,2,\ldots,n$ and $C_Y^i(t)$ represents the Yellow token bucket count at time t for TRF i, $i = 1,2,\ldots,n$.

$T_G^i(t_1,t_2)$ and $T_Y^i(t_1,t_2)$ represent the total maximum number of tokens available that might be added to the Green and Yellow token buckets bucket counts, respectively, for TRF i over the time interval t_1 to t_2, for $i = 1,2,\ldots,n$.

$B_G^i(t_1,t_2)$ and $B_Y^i(t_1,t_2)$ represent the number of tokens that Bypass the Green and Yellow token buckets, respectively, for TRF i over the time interval t_1 to t_2, for $i = 1,2,\ldots,n$ due to the limits imposed by GTR_{max}^i and YTR_{max}^i. Some of the effects of the maximum rate limits are discussed in Appendix B.2.

$A_G^i(t_1,t_2)$ and $A_Y^i(t_1,t_2)$ represent the number of tokens actually added to the Green and Yellow token bucket counts, respectively, for TRF i over the time interval t_1 to t_2, for $i = 1,2,\ldots,n$.

$O_G^i(t_1,t_2)$ and $O_Y^i(t_1,t_2)$ represent the number of tokens that Overflow the Green and Yellow token buckets, respectively, for TRF i over the time interval t_1 to t_2, for $i = 1,2,\ldots,n$. Tokens are said to Overflow when they cannot the number of tokens to be added to a token bucket count without causing would cause the count to exceed the upper limit imposed by GTV^i and YTV^i on the token bucket count. Some of the effects of the maximum token bucket count are discussed in Appendix B.1.

$R_G^i(t_1,t_2)$ represents the number of Green tokens that are converted by converting them to Yellow and making them available to the Yellow token bucket of TRF i over the time interval t_1 to t_2 for $i = 1,2,\ldots,n$.

$S_G^i(t_1,t_2)$ and $S_Y^i(t_1,t_2)$ represent the number of tokens available to be shared to the Green and Yellow token buckets bucket counts of TRF $i-1$, respectively, from TRF i over the time interval t_1 to t_2 for $i = 2,\ldots,n,n+1$.

Figure A1-1 shows a detailed diagram of the token flows for a TRF i with each of the above values labeled in the diagram.

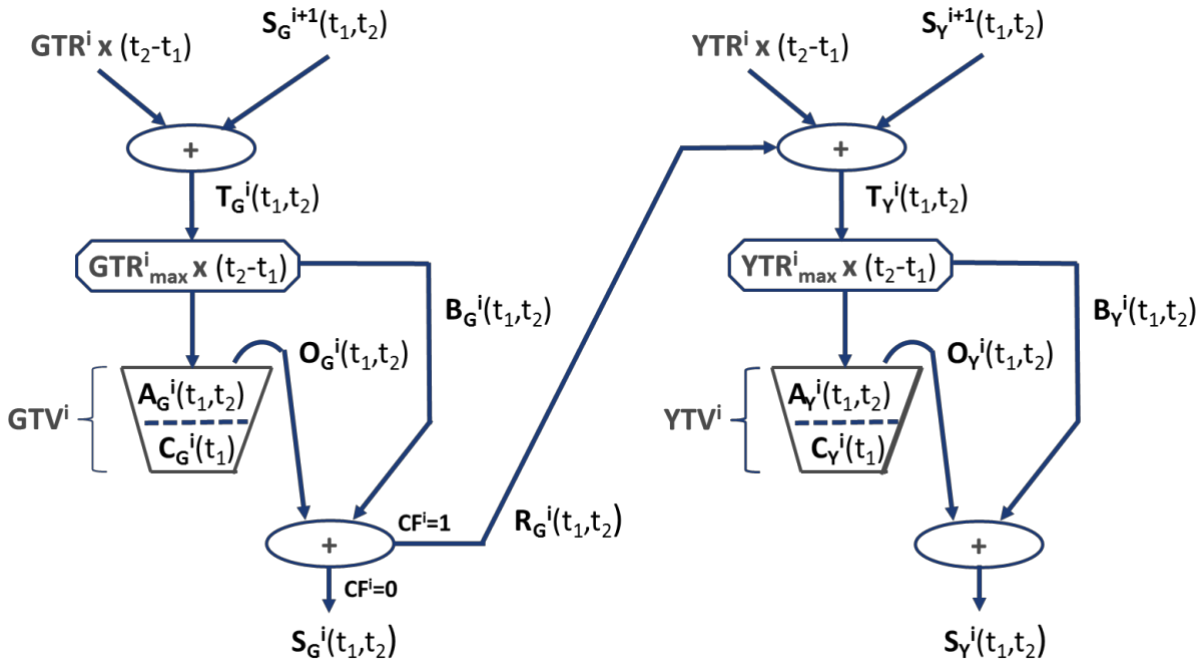

Figure A1-1 – Token Flows for TRF *i*

The total maximum number of tokens available to the that might be added to Green token bucket count, $T_G^i(t_1, t_2)$, includes Green tokens sourced at the rate GTR^i over the time interval and any Green tokens shared from the next higher Rank TRF. The total maximum number of tokens available that might be added to the Yellow token bucket count, $T_Y^i(t_1, t_2)$, includes Yellow tokens sourced at the rate YTR^i over the time interval, any Yellow tokens shared from the higher Rank TRF, and any overflow Green tokens from tokens that Overflow or Bypass the Green token bucket that are allowed and are converted to the Yellow token bucket by CF^i ($R_G^i(t_1, t_2)$). Note that for the Green token bucket count, TRF n has no tokens shared from a higher Rank since there are no TRFs with a Rank higher than n. Therefore $S_G^{n+1}(t_1, t_2) = 0$. Similarly, for the Yellow token bucket count, there are TRF n has no tokens shared from a Rank higher than n since there are no TRFs with a Rank higher than n, however, depending on CF^0, there may be tokens converted from shared from the overflow of the Green token bucket at Rank 1. Therefore $S_Y^{n+1}(t_1, t_2) = CF^0 \times \left(B_G^1(t_1, t_2) + O_G^1(t_1, t_2)\right)$. The following equations capture the above descriptions:

$$T_G^i(t_1, t_2) = GTR^i \times (t_2 - t_1) + S_G^{i+1}(t_1, t_2) \text{ for } i = 1,2, \dots, n$$

$$T_Y^i(t_1, t_2) = YTR^i \times (t_2 - t_1) + S_Y^{i+1}(t_1, t_2) + R_G^i(t_1, t_2) \sout{CF^i \times O_G^i(t_1, t_2)} \text{ for } i = 1,2, \dots, n$$

Note that [R3] mandates that ~~$CF^n = 0$~~ if ~~$CF^0 = 1$~~ in the equation for ~~$T_Y^n(t_1, t_2)$~~ and thus

$$\sout{T_Y^n(t_1, t_2) = \begin{cases} YTR^n \times (t_2 - t_1) + CF^n \times O_G^n(t_1, t_2), & \text{if } CF^0 = 0 \\ YTR^n \times (t_2 - t_1) + CF^0 \times O_G^1(t_1, t_2), & \text{if } CF^0 = 1 \end{cases}}$$

The number of tokens that Bypass the token bucket is the number of tokens that exceed the maximum rate at which tokens are allowed to be added (GTR^i_{max} and YTR^i_{max}) over the interval.

$$B^i_G(t_1,t_2) = \max\{0, T^i_G(t_1,t_2) - GTR^i_{max} \times (t_2 - t_1)\} \text{ for } i = 1,2,\dots,n$$

$$B^i_Y(t_1,t_2) = \max\{0, T^i_Y(t_1,t_2) - YTR^i_{max} \times (t_2 - t_1)\} \text{ for } i = 1,2,\dots,n$$

The number of tokens actually added to a token bucket count is some or all of the tokens that do not Bypass the token bucket available to be added, limited by GTV^i and YTV^i and by the maximum rate at which tokens are allowed to be added (~~GTR^i_{max}~~ and ~~YTR^i_{max}~~).

$$A^i_G(t_1,t_2) = \min\{T^i_G(t_1,t_2) - B^i_G(t_1,t_2),\ GTV^i - C^i_G(t_1)\text{~~, } GTR^i_{max} \times (t_2 - t_1)\} \quad \text{for} \quad i = 1,2,\dots,n$$

$$A^i_Y(t_1,t_2) = \min\{T^i_Y(t_1,t_2) - B^i_Y(t_1,t_2),\ YTV^i - C^i_Y(t_1)\text{~~, } YTR^i_{max} \times (t_2 - t_1)\} \quad \text{for} \quad i = 1,2,\dots,n$$

The number of tokens that Overflow each token bucket is the number of available tokens that are not actually added to the token bucket count.

$$O^i_G(t_1,t_2) = T^i_G(t_1,t_2) - B^i_G(t_1,t_2) - A^i_G(t_1,t_2) \text{ for } i = 1,2,\dots,n$$

$$O^i_Y(t_1,t_2) = T^i_Y(t_1,t_2) - B^i_Y(t_1,t_2) - A^i_Y(t_1,t_2) \text{ for } i = 1,2,\dots,n$$

The number of tokens that are converted to Yellow tokens is the number of tokens that Bypass plus the number of tokens that Overflow the Green token bucket when the Coupling Flag is set ($CF^i = 1$).

$$R^i_G(t_1,t_2) = CF^i \times \left(B^i_G(t_1,t_2) + O^i_G(t_1,t_2)\right) \text{ for } i = 1,2,\dots,n.$$

The number of tokens available to be shared from TRF i to the next lower Rank, TRF $i-1$, is the number of tokens that Bypass or Overflow the Green and Yellow token buckets and, in the case of the Green token bucket count, are not made available converted to the Yellow token bucket count by CF^i.

$$S^i_G(t_1,t_2) = (1 - CF^i) \times \left(B^i_G(t_1,t_2) + O^i_G(t_1,t_2)\right) \text{ for } i = 2,3,\dots,n.$$

$$S^i_Y(t_1,t_2) = B^i_Y(t_1,t_2) + O^i_Y(t_1,t_2) \text{ for } i = 2,3,\dots,n.$$

7 New Appendix B

Insert the content below into the document as Appendix B.

Appendix B Examples of GTBA Behavior (Informative)

This Appendix describes the behaviors of the GTBA resulting from the values of GTV^i and GTR^i_{max}. Appendix B.1 defines Burst, Burst Size, Burst Length and Burst Magnitude for a set of frames and demonstrates how two sequences of frames with the same average byte rate can have different color declarations when each frame length and arrival time is considered a Token Request for the GTBA. Appendix B.2 demonstrates how the value of GTR^i_{max} can impact the use of tokens shared from Rank *i*+1 by the TRF at Rank *i*.

B.1 GTV$_i$ and Burst Detection

Traffic in a network tends to be "bursty", meaning that the frames are typically not evenly spaced. Traffic that is excessively bursty is problematic because it can lead to frames being discarded unless there are large buffers within the network, which in turn makes it difficult to control the delay and delay variation of the transmission of frames through the network. Controlling the "burstiness" of the traffic requires a specific definition of a burst, a quantitative measurement of the burstiness of the traffic, and a way to detect when the burstiness exceeds a specified limit. The GTBA provides a parameter (GTV^i) for setting a limit on the acceptable burstiness, and a mechanism to detect when that limit is exceeded.

Intuitively, a burst within a network is a group of frames that are closer together in time than expected, or allowed, or desired, or simply closer together in time than other frames. This is obviously too general of a definition to lead to any quantitative measurement of burstiness. A more precise definition can be provided by a comparison to a reference rate, where the reference rate can be the long term average of the byte rate of the flow, or a predetermined target rate. Then the general definition becomes a sequence of frames where the byte rate of the frames exceeds the reference rate. A more precise definition is that, given a sequence of frames, each Burst in the sequence is a maximal subsequence of one or more frames where at every point in time during the Burst the average byte rate of the frames since the beginning of the subsequence exceeds the reference rate. The procedure for finding each maximal subsequence that comprises a Burst in a sequence of frames is:

1. The first frame in the sequence is always the start of a Burst.
2. For each subsequent frame, if the total byte count of all frames since the start of the current Burst (but not including this frame) is greater than the reference rate multiplied by the time since the start of the current Burst, then this frame is part of the current Burst; otherwise, it is the start of a new Burst.

Given this definition of a Burst, there are several properties of the Burst that can be measured. Three examples are:

- **Burst Size**: The sum of the byte counts of the frames in the Burst.
- **Burst Length**: The duration of the time interval from the arrival time of the first frame of the Burst to the first point in time where the cumulative byte count of the frames divided by the duration equals the reference rate.
- **Burst Magnitude**: The maximum amount by which the cumulative byte count of the frames since the beginning of the Burst exceeds the reference rate multiplied by the time since the beginning of the Burst.

While all three of these capture interesting characteristics of the Burst, only the Burst Magnitude reacts to the relative timing of the frames within the Burst. Furthermore, the Burst Magnitude is the characteristic that directly affects the amount of buffering required within a network to accommodate the sequence of frames, and the delay and delay variation experienced by the frames within the network.

A Burst can be represented graphically by a plot of the difference between the cumulative byte count of the frames and the reference rate over time. An example of this is shown in Figure A1-2. The top portion of the figure depicts the sequence of frames in time where the length of the arrows represent the number of bytes in the frame. The sequence shown forms two Bursts, the first consisting of the first five frames (at times t1, t2, t3, t4, and t5) and the second consisting of the single frame at time t6. The middle portion of the figure depicts the cumulative byte count exceeding the reference rate. The downward slope of the gray shading is the negative of the reference rate. The time of the end of the first burst is designated as t_b. The bottom portion of the figure puts this in the context of a GTBA by depicting the response of the GTBA when each frame generates a Green Token Request equal to the number of bytes in the frame and the reference rate is the rate at which tokens are added to the Green token bucket. The dark green shading represents the number of tokens in the bucket over time, while the light green shading represents the number of tokens that would be in the bucket if the token count was continuously updated rather than only updated in response to a Token Request. In this example all Token Requests are declared Green because the Burst Magnitude of the sequence of frames never exceeds GTV^i.

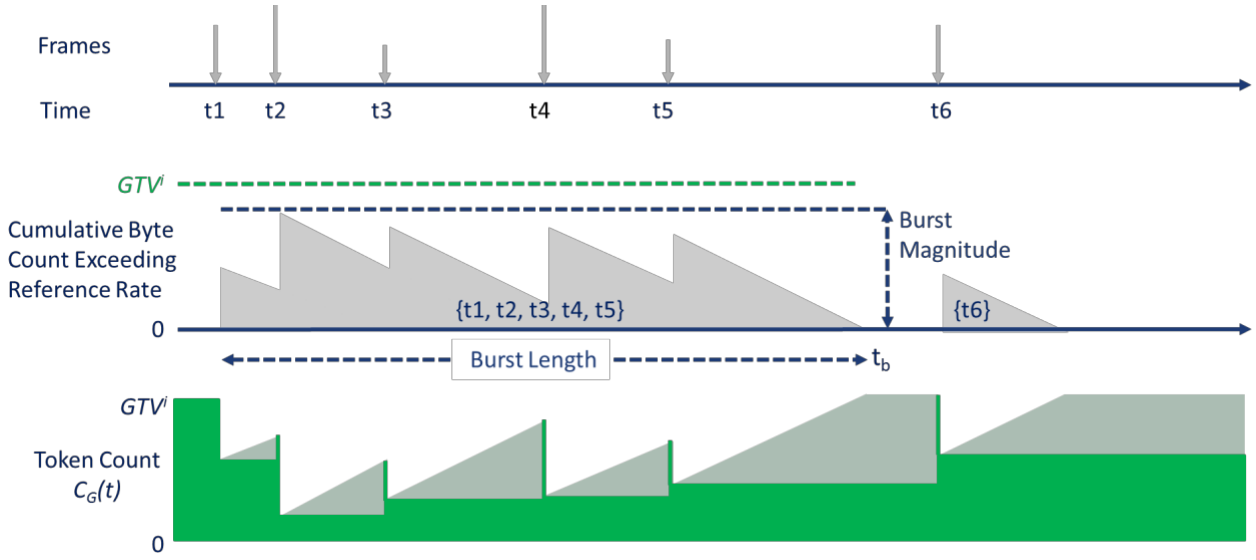

Figure A1-2 – Token Request Sequence with a Burst Magnitude less than GTV_i

Figure A1-3 modifies the sequence of frames by moving the fourth frame earlier in time (from t4 to t4'). This increases the Burst Magnitude of the sequence of frames such that it is larger than GTV^i. When Burst Magnitude of the sequence of frames exceeds GTV^i, the GTBA cannot declare all of the Token Requests generated by the frames Green without the token count dropping below zero, so the GTBA responds by declaring one or more Token Requests a different color (Yellow or Red). In this example the Token Request generated by the frame at t4' is declared Red, shown in the figure by the dashed red line that would drop the token count below zero.

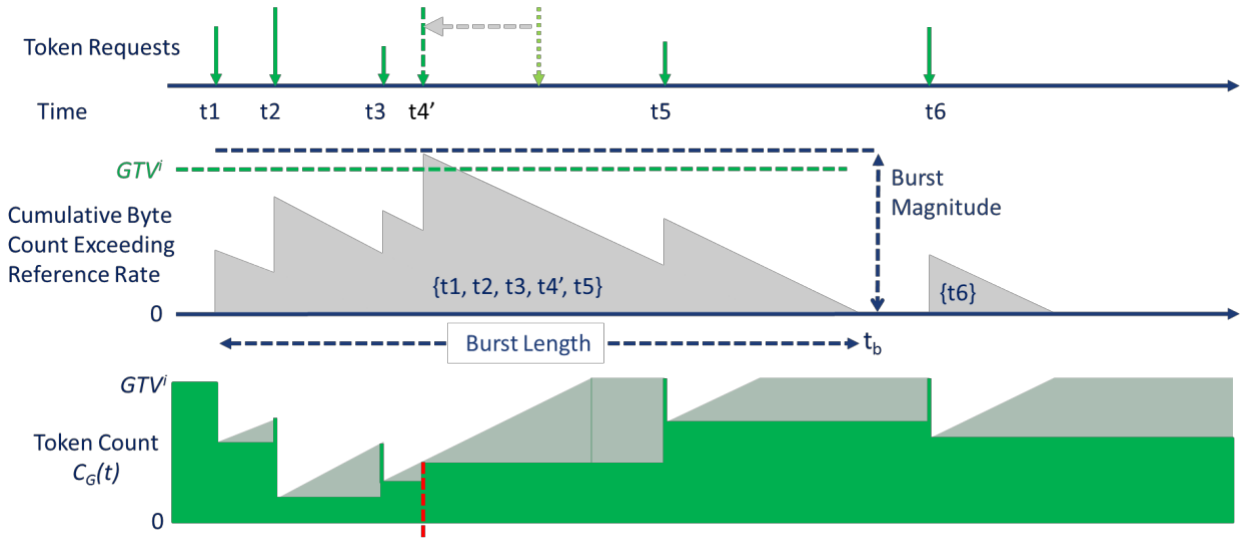

Figure A1-3 – Token Request Sequence with a Burst Magnitude greater than GTV_i

The sequence of frames from t1 to t5 in both Figure A1-2 and Figure A1-3 have the same Burst Size and Burst Length. Therefore the byte rate of both sequences equals the reference rate at the same point in time (t_b). Yet in one case all Token Requests generated by the frames are declared

Green and in the other case one of the Token Requests is declared Yellow or Red. The difference is that in Figure A1-3 the first four frames are closer together in time, to the point that the Burst Magnitude exceeds the limit set by GTV^i. Declaring the Token Request generated by the frame at t4' in Figure A1-3 Red, but not the Token Request generated by the frame at t4 in Figure A1-2, is an example that shows the GTBA can detect that the Burst Magnitude of a sequence of frames exceeds a desired limit (GTV^i), even when the average byte rate of the frames does not exceed the reference rate (GTR^i).

These examples have assumed that the reference rate used to delineate a Burst is a constant equal to the number of tokens sourced at a given Rank by the GTBA (GTR^i). With token sharing the rate of tokens added to the token bucket is not necessarily constant, however it is bounded between a lower bound equal to the minimum of GTR^i and GTR^i_{max}, and an upper bound equal to GTR^i_{max}. In this case two statements can be made about the relationship between a Burst Magnitude and the GTBA. First, when using the minimum of GTR^i and GTR^i_{max} as a reference rate to delineate a Burst, and the resulting Burst Magnitude is less than GTV^i, all Token Requests generated by the frames within the Burst will be declared Green. Second, when using GTR^i_{max} as a reference rate to delineate a Burst, and the resulting Burst Magnitude exceeds GTV^i, one or more Token Requests generated by the frames in the Burst will not be declared Green. Analogous statements can be made regarding the actions of the Yellow token bucket in the GTBA, where the Token Requests being considered includes all Yellow Token Requests and all Green Token Requests that were not declared Green. In this case the rate of tokens added to the Yellow token bucket can include tokens from a Yellow token source at that Rank, tokens shared from another Rank, or Green tokens converted to Yellow at the same Rank.

B.2 GTR$_{imax}$ and Token Bypass

The maximum token rate parameters (GTR^i_{max} and YTR^i_{max}) divide the total tokens available at a Rank into tokens that are added to the token bucket (or Overflow if the bucket is full), and tokens that Bypass the token bucket at that Rank (regardless of whether the token bucket is full or not). This section discusses how the maximum token rate parameters affect the algorithm behavior under two different circumstances. The first occurs when tokens Bypass the token bucket purely as a result of the GTBA parameter values, regardless of the sequence of Token Requests. The rate of this type of Bypass does not vary between consecutive Token Requests, and is therefore referred to as Constant Bypass. The other occurs when tokens Bypass the token bucket as a result of a specific sequence of Token Requests. The rate of this type of Bypass can vary considerably between consecutive Token Requests, and is therefore referred to as Transient Bypass. These two types of Bypass are discussed separately in the following sections. The examples in each section assume the token flows shown in Figure A1-4.

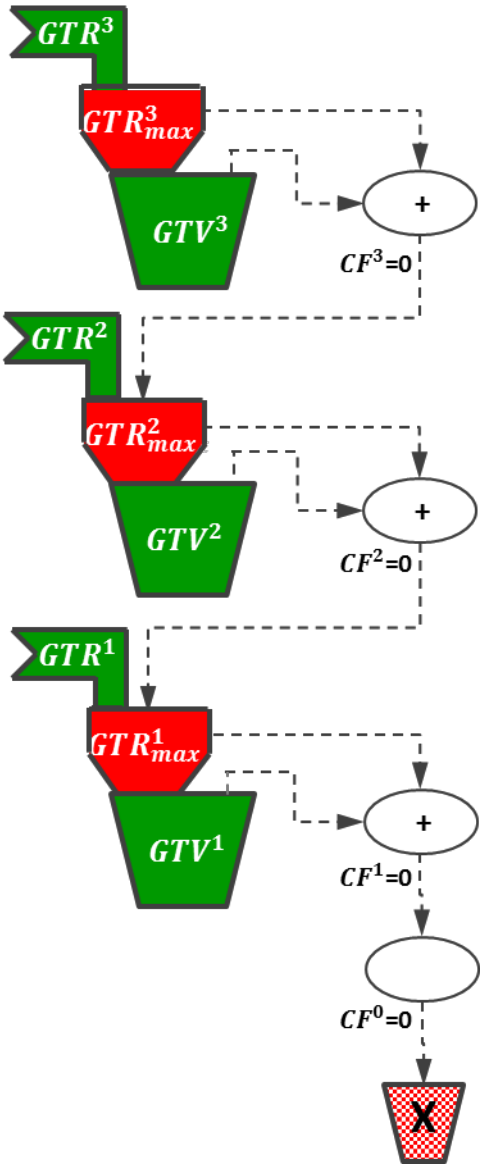

Figure A1-4 – GTBA Token Paths for Constant and Transient Bypass Examples

B.2.1 Constant Bypass Example

For Green token buckets, Constant Bypass occurs at a given Rank when the combination of the new token rate at that Rank (GTR^i) plus the Constant Bypass Rate from a higher Rank (if any) exceeds the maximum token rate at that Rank (GTR^i_{max}). This is exemplified in the tables of GTBA parameter values shown in Table A1-1. (All rates are in units of tokens per second.)

Rank (i)	GTR^i	GTR^i_{max}	Constant Bypass Rate
3	100	20	80
2	0	30	50
1	0	100	0

Table A1-1(a)

Rank (i)	GTR^i	GTR^i_{max}	Constant Bypass Rate
3	20	20	0
2	30	30	0
1	50	100	0

Table A1-1(b)

Table A1-1 – Constant Bypass Examples

In Table A1-1(a) the GTR^3 of the highest Rank exceeds GTR^3_{max}, resulting in a Constant Bypass rate at Rank 3 of 80. These tokens are shared to Rank 2, where they exceed the GTR^2_{max} of Rank 2, resulting in a Constant Bypass rate of 50. These tokens are shared to Rank 1.

For Yellow token buckets, Constant Bypass occurs when the maximum token rate at that Rank (YTR^i_{max}) is exceeded by the combination of the new token rate at that Rank (YTR^i), plus any Constant Bypass from the Green token bucket at the same Rank converted to Yellow (if $CF^i = 1$), plus any Constant Bypass from the Yellow token bucket at a higher Rank (or, in the case of Rank n, any Constant Bypass from the Rank 1 Green token bucket converted to Yellow by $CF^0 = 1$).

In general, the Constant Bypass Rate for the Green and Yellow token buckets for Rank i, CBR^i_G and CBR^i_Y respectively, can be calculated using the following equations:

$$CBR^i_G = \max\{0,\ GTR^i + (1 - CF^{i+1}) \times CBR^{i+1}_G\ -\ GTR^i_{max}\} \text{ for } i = 1,2,\dots,n$$

where $CBR^{n+1}_G = 0$ and $CF^{n+1} = 0$.

$$CBR^i_Y = \max\{0,\ YTR^i + CBR^{i+1}_Y + CF^i \times CBR^i_G\ -\ YTR^i_{max}\} \text{ for } i = 1,2,\dots,n$$

where $CBR^{n+1}_Y = (1 - CF^0) \times CBR^1_G$.

Two instances of the GTBA are said to have identical color declaration behavior when any sequence of Token Requests yields identical color declarations when presented to each GTBA instance. The GTBA instance in Table A1-1(b) has identical color declaration behavior as that in Table A1-1(a), but has the token sources adjusted so that all Ranks have a Constant Bypass Rate equal to zero and GTR^i_{max} greater than or equal to GTR^i. There can be many GTBA instances that have identical color declaration behavior, but only one of those will have zero Constant Bypass Rates at all Ranks. This is referred to as the "normalized" GTBA instance. Working with the normalized GTBA instance is useful for analyzing Transient Bypass, and will be used extensively in the Section B.2.2.3. It is possible to transform any instance of the GTBA that has a non-zero Constant Bypass Rate at one or more Ranks to the normalized instance of the GTBA that has the same color declaration behavior but has a Constant Bypass Rate of zero at all Ranks. This is done by changing the value of the token source rates GTR^i and YTR^i to the normalized token source rate values GTR^i_{nrm} and YTR^i_{nrm}, respectively, for $i = 1,2,\dots,n$ where:

$$GTR^i_{nrm} = min\{\, GTR^i + (1 - CF^{i+1}) \times CBR^{i+1}_G,\ GTR^i_{max}\},$$

$$YTR^i_{nrm} = min\{\, YTR^i + CBR^{i+1}_Y + CF^i \times CBR^i_G,\ YTR^i_{max}\}.$$

B.2.2 Transient Bypass Example

Transient Bypass is a more subtle effect of GTR^i_{max} than Constant Bypass, because it depends not just on the GTBA parameter values but also on the Token Requests. Table A1-2 shows example GTBA parameter values and average Token Request rates, $TRR^n_{avg,G}$, at each Rank. It is assumed in the discussion of these examples that the Token Requests at each Rank can be bursty, but that the GTV^i values are sufficiently large that no Token Requests would fail to be declared Green if there were no Transient Bypass. This table will be used to demonstrate the effect of Transient Bypass on the average rate of Token Requests declared Green at Ranks 2 and 3.

Rank (i)	GTR^i	GTR^i_{max}	Constant Bypass Rate	Average Token Request Rate ($TRR^i_{avg,G}$)	Average Overflow Rate	Intuitive Rate of Token Requests Declared Green	Average Transient Bypass Rate ($TBR^i_{avg,G}$)	Actual Rate of Token Requests Declared Green
3	20	20	0	10	10	10	0	10
2	30	40	0	40	0	40	0 - 5	35 - 40
1	0	50	0	5	0	0	0	0 - 5

Table A1-2 – Transient Bypass Example

In this example the average rate of Token Requests at Rank 3 is 10 tokens per second, and both GTR^3 and GTR^3_{max} are 20 tokens per second. Therefore the average Overflow rate is also 10 tokens per second.

The average rate of Token Requests at Rank 2 is 40 tokens per second. It might appear that with an average Overflow rate from Rank 3 of 10 tokens per second, a GTR^2 of 30 tokens per second, and a GTR^2_{max} of 40 tokens per second, all Token Requests at Rank 2 would be declared Green, leaving no tokens to Overflow to Rank 1. With no Overflow or Constant Bypass from Rank 2 to Rank 1, and GTR^1 equal zero, the rate of Token Requests declared Green at Rank 1 would be zero. This is shown in the column labeled "Intuitive Rate of Token Requests Declared Green". However, this intuition is based on the average Overflow rate from Rank 3 to Rank 2. On any given interval between Token Requests, the actual Overflow from Rank 3 could be higher or lower than the average. When it is higher it can exceed the value of GTR^2_{max}, resulting in some tokens that Bypass the token bucket at Rank 2. This is Transient Bypass. The effect of Transient Bypass is that not all of the Token Requests at Rank 2 will be declared Green, and some Token Requests at Rank 1 can be declared Green.

Given a sequence of Token Requests, the average Transient Bypass rate at each Rank can be calculated by summing the number of Transient Bypass tokens calculated for that Rank at each token count update (i.e. at each Token Request), and dividing by the duration of the sequence. The average Transient Bypass rate for Rank 2 in Table A1-2 is shown as a range of values from

0 to 5. This is because even when all the GTBA Parameters and all the average Token Request rates are fixed, different sequences of Token Requests can result in different average Transient Bypass rates. The remainder of this section focuses on determining the bounds on the average Transient Bypass rate as a function of the GTBA Parameters and the average Token Request rates.

- Section B.2.2.1 describes a specific sequence of Token Requests that results in the highest possible average Transient Bypass rate (i.e. the least upper bound) at Rank 2 given the GTBA Parameter and average Token Request rate values in Table A1-2.
- Section B.2.2.2 describes a specific sequence of Token Requests that results in a lower value for the average Transient Bypass rate at Rank 2, and extrapolates this to the lowest possible average Transient Bypass rate (i.e. the greatest lower bound) at Rank 2 given the GTBA Parameter and average Token Request rate values in Table A1-2.
- Section B.2.2.3 develops general equations for bounds on the average Token Bypass rate as a function of the GTBA Parameters and average Token Request rates.
- Section B.2.2.4 plots the average Transient Bypass rate at Rank n-1 as a function of the average Token Request rate at Rank n, and makes observations about how the average Transient Bypass rate is affected by changing values of the GTBA Parameters and/or the average Token Request rates.

B.2.2.1 Average Transient Bypass Rate Least Upper Bound Example

Suppose the Token Requests at Rank 3 are uniformly spaced at 1s intervals, and each request is for 10 tokens. Further suppose that between these there are nine Token Requests in other Ranks every 0.1s. Finally suppose that eight of these are requests at Rank 2 for 5 tokens each, and one is a request at Rank 1 for 5 tokens. Recall that the token count updates occur for all Ranks each time a Token Request occurs at any Rank, and therefore in this example the token count updates occur every 0.1s. This sequence of Token Requests is diagrammed in Figure A1-5. The figure also shows the timeline of the token count at each Rank.

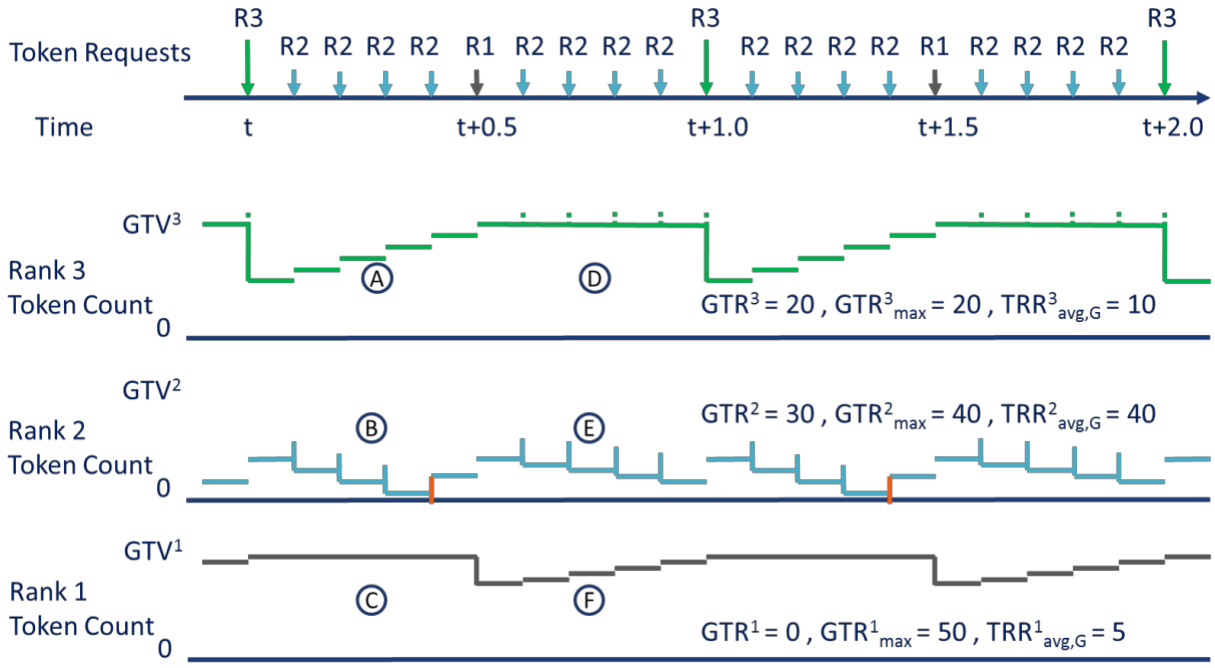

Figure A1-5 – Transient Bypass Least Upper Bound Example

For the first 0.5s after a Token Request is received at Rank 3, the Rank 3 bucket will not be full, and so it will be replenished at 20 tokens per second (GTR^3), or 2 tokens every 0.1s (reference point A in Figure A1-5). During this 0.5s period no tokens will Overflow from Rank 3 to Rank 2, the Rank 2 bucket will be replenished at 3 tokens every 0.1s, and 5 tokens will be removed if the Rank 2 Token Request is declared Green (reference point B). Also during this 0.5s period there is no change to the Rank 1 bucket until 5 tokens are removed when the Rank 1 Token Request is declared Green (reference point C). After 0.5s, the Rank 3 bucket will be full again, and so for the next 0.5s tokens will Overflow to Rank 2 at a rate of 20 tokens per second, or 2 every 0.1s (reference point D). For each 0.1s interval during this period there will be 5 tokens available at Rank 2 (3 from GTR^2 plus 2 overflowing from Rank 3), however, GTR^2_{max} limits the number that can be added to the Rank 2 bucket to 4. Therefore 4 tokens will be added to the bucket, 5 tokens will be removed if the Rank 2 Token Request is declared Green, and 1 token will Bypass Rank 2 and be available at Rank 1 (reference point E). This single token is added to the Rank 1 bucket every 0.1s (reference point F).

The consequence of this sequence of Token Requests is that there are 35 tokens added to the Rank 2 bucket every second, and 40 tokens removed from the bucket (assuming all Rank 2 Token Requests are declared Green). Of course this will eventually empty the Rank 2 bucket, and will reach a steady state scenario where only 7 of the 8 Token Requests at Rank 2 are declared Green, and one is declared Red. Meanwhile Rank 1 receives 5 tokens per second due to the Transient Bypass from Rank 2, and therefore declares the one Token Request per second Green.

This sequence of Token Requests is an example that results in the maximum possible value for the average Transient Bypass rate at Rank 2 (i.e. the least upper bound). This is because there is one Token Request that occurs at just the right time for the Rank 3 bucket to be exactly filled,

which means that for all the Token Requests there will be either no tokens Overflowing the Rank 3 bucket, or all the tokens sourced by GTR^3 will Overflow. Therefore when tokens do Overflow, they do so at the maximum possible rate which maximizes the amount by which they exceed GTR^2_{max}, and thus maximizes Transient Bypass at Rank 2. If there is no Token Request that exactly fills the Rank 3 bucket, but instead there is a Token Request where some tokens are added to the bucket (filling it) and some Overflow, then for those Overflow tokens the rate of Overflow would be less than GTR^3, resulting in less Transient Bypass at Rank 2 and bringing down the average Transient Bypass rate.

B.2.2.2 Average Transient Bypass Rate Greatest Lower Bound Example

If the Rank 2 average Transient Bypass rate least upper bound is the result when a Token Request comes at just the right point in time to fill the Rank 3 bucket, it stands to reason that the Rank 2 average Transient Bypass greatest lower bound would be the result when all Token Requests were as far from this point in time as possible. Consider a sequence of Token Requests similar to that described in Section B.2.2.1, but instead of the Rank 2 and Rank 1 Token Requests being evenly distributed at 0.1s intervals, they all occur within the first 0.1s after a Rank 3 Token Request. This sequence of Token Requests, with the timeline of the token count at each Rank, is diagrammed in Figure A1-6.

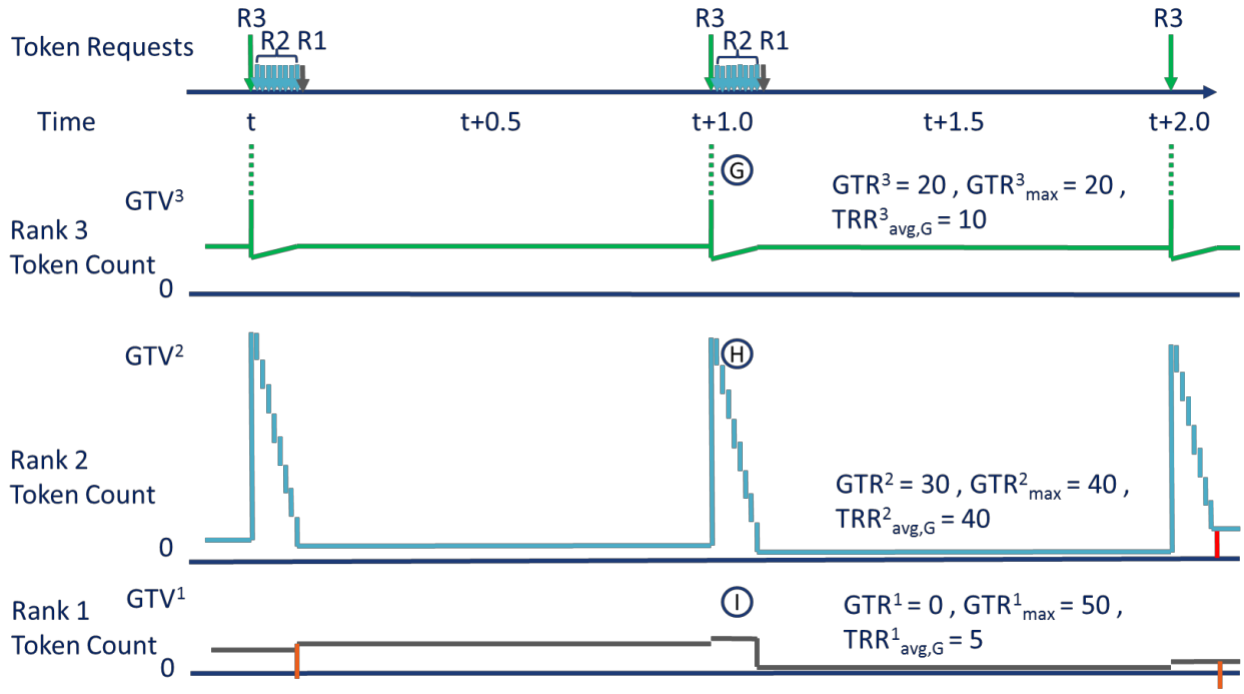

Figure A1-6 – Transient Bypass Example Near Greatest Lower Bound

Each Rank 3 Token Request occurs after a 0.9s gap, so there will be 18 tokens sourced by GTR^3, 8 of which will be added to the Rank 3 bucket (filling it) and 10 Overflow. 10 tokens are removed when the Rank 3 Token Request is declared Green. Over the next 0.1s there will be 2 tokens added to the Rank 3 bucket, and no tokens Overflow (reference point G in Figure A1-6). At Rank 2 there will be 37 tokens available (27 from GTR^2 plus the 10 Overflowing from Rank 3) when the Rank 3 Token Request occurs, however, GTR^2_{max} limits the number that can be

added to the Rank 2 bucket to 36. Therefore one token will Bypass Rank 2 and be available at Rank 1. Over the next 0.1s there will be 3 tokens added to the Rank 2 bucket, and 5 tokens removed for each Rank 2 Token Request declared Green (reference point H). The single token that Bypassed the Rank 2 bucket will be added to the Rank 1 bucket, and 5 tokens will be removed if the Rank 1 Token Request is declared Green (reference point I).

This sequence of Token Requests results in an average Transient Bypass rate at Rank 2 of 1 token per second, whereas the previous example resulted in an average Transient Bypass rate at Rank 2 of 5 tokens per second. Compressing the Rank 2 and Rank 1 Token Requests into an even smaller interval following the Rank 3 Token Requests would result in an even lower average Transient Bypass rate at Rank 2, approaching the minimum possible value (i.e. the greatest lower bound) which occurs when all Token Requests are simultaneous. In this example the greatest lower bound of the Rank 2 average Transient Bypass rate is zero.

B.2.2.3 Calculating Bounds on the Average Transient Bypass Rate

The analysis in this section uses the normalized GTBA. Thus Constant Bypass is zero and GTR^i_{max} is greater than or equal to GTR^i_{nrm} at all ranks (Section B.2.1).

The examples in Section B.2.2.1 and Section B.2.2.2 show that the average Transient Bypass rate at Rank i ($TBR^i_{avg,G}$) varies depending on when token count updates occur, and therefore on the arrival of Token Requests at any Rank. However the value of the average Transient Bypass rate is bounded above and below. The greatest lower bound on the average Transient Bypass rate can be computed by observing that during any time interval the average rate of tokens added to the token bucket at Rank i (i.e. the average rate of tokens shared from Rank $i+1$, denoted by $S^{i+1}_{avg,G}$, plus GTR^i_{nrm} minus the average Transient Bypass rate) can be less than GTR^i_{max} but cannot exceed GTR^i_{max}. Therefore the following inequality is true for any time interval:

$$TBR^i_{avg,G} \geq max\{0, S^{i+1}_{avg,G} + GTR^i_{nrm} - GTR^i_{max}\}.$$

In the case of Rank $n-1$, the average rate of tokens shared from Rank n is the average rate of Overflow from Rank n. The greatest lower bound on the average rate of overflow is either zero (when the average Token Request rate, $TRR^n_{avg,G}$, exceeds GTR^n_{nrm}), or GTR^n_{nrm} minus the average Token Request rate (assuming the value of GTV^n is large enough that all Token Requests are declared Green as long as the average Token Request rate is less than GTR^n_{nrm}). Therefore the greatest lower bound on the average Transient Bypass rate at Rank $n-1$ is:

$$TBR^{n-1}_{avg,G} \geq max\{0, (1 - CF^n) \times max\{0, GTR^n_{nrm} - TRR^n_{avg,G}\} + GTR^{n-1}_{nrm} - GTR^{n-1}_{max}\}$$

In the general case of Rank i, the average rate of tokens shared from Rank $i+1$ depends upon the rate of tokens shared from Rank $i+2$. Tokens shared from Rank $i+2$ can only increase the tokens shared from Rank $i+1$, which can only increase the average Transient Bypass rate at Rank i. Therefore a lower bound on the average Transient Bypass rate at Rank i can be calculated by taking the above equation and replacing "$n-1$" with "i":

$$TBR^i_{avg,G} \geq max\{0, (1 - CF^{i+1}) \times max\{0, GTR^{i+1}_{nrm} - TRR^{i+1}_{avg,G}\} + GTR^i_{nrm} - GTR^i_{max}\}$$

for $i = 1, .., n-1$.

Note that this inequality provides a lower bound, but not necessarily the greatest lower bound. When the rate of tokens shared from Rank $i+2$ is zero the lower bound calculated by this equation is the greatest lower bound. When the rate of tokens shared from Rank $i+2$ is large and GTR^i_{max} is small, the lower bound calculated by this equation may significantly under estimate the greatest lower bound.

As discussed in Section B.2.2.1, the least upper bound on the average Transient Bypass rate at Rank $n-1$ occurs when there is a token count update in response to a Token Request at a Rank other than n that exactly fills the Rank n bucket. Then, during the interval when the Rank n bucket is full, all the tokens sourced at Rank n Overflow and there can be Transient Bypass at Rank $n-1$. During the interval when the Rank n bucket is not full none of the tokens sourced at Rank n will Overflow and there will not be any Transient Bypass at Rank $n-1$. The least upper bound on the average Transient Bypass rate at Rank $n-1$ can be determined by multiplying the Transient Bypass rate when the Rank n bucket is full by the fraction of time that the Rank n bucket is full. This fraction is one minus the ratio of the average Token Request rate at Rank n to the token source rate at Rank n. The Transient Bypass at Rank $n-1$ during the interval when the Rank n bucket is full is the sum of the tokens sourced at Ranks n and $n-1$, minus the maximum token rate at Rank $n-1$. Therefore the least upper bound on the average Transient Bypass rate at Rank $n-1$ is:

$$TBR^{n-1}_{avg,G} \leq max\left\{0, 1 - \frac{TRR^n_{avg,G}}{GTR^n_{nrm}}\right\} \times max\{0, (1 - CF^n) \times GTR^n_{nrm} + GTR^{n-1}_{nrm} - GTR^{n-1}_{max}\}$$

In the general case of Rank i, the maximum Transient Bypass occurs during intervals where the token buckets at all higher Ranks are full. During these intervals the Transient Bypass rate is the amount by which the sum of the token source rates at Rank i and all higher Ranks exceeds GTR^i_{max} (assuming $CF^i = 0$ at all higher Ranks). The Transient Bypass rate will be lower during any other intervals, with the actual value depending upon which of the higher Rank token buckets are full. Therefore calculating the least upper bound for Ranks lower than n-1 is complex and beyond the scope of this document. However, an upper bound on the average Transient Bypass rate at Rank i can be calculated by assuming the token buckets at all higher Ranks fill at the same time, and that maximal rate of Token Bypass is sustained for the longest time that the token bucket at any Rank remains full:

$$TBR^i_{avg,G} \leq max\left\{0, 1 - \frac{TRR^{i+1}_{avg,G}}{GTR^{i+1}_{nrm}}, .., 1 - \frac{TRR^n_{avg,G}}{GTR^n_{nrm}}\right\} \times max\left\{0, \left(\sum_{j=i}^{n} GTR^j_{nrm}\right) - GTR^i_{max}\right\}$$

for $i = 1, .., n$ and where $CF^i = 0$.

Note that this equation provides an upper bound, but not necessarily the least upper bound. The upper bound calculated by this equation will be the least upper bound only in two cases. One of these cases is that the token buckets at all higher Ranks are full for the same percentage of time, i.e. the ratio of the average Token Request rate to the normalized token source rate is the same at all Ranks. The other of these cases is when GTR^i_{max} is greater than or equal to the sum of the normalized token source rates at all Ranks, in which case the average Transient Bypass rate is

zero. In all other circumstances the upper bound calculated by this equation will over estimate the least upper bound. When there is wide variation in the ratio of average Token Request rate to normalized token source rate at different Ranks, the over estimate can be quite large.

The greatest lower bound and least upper bound for the average Transient Bypass rate at Rank $n-1$ can be applied to the example in Table A1-2. First note that the Constant Bypass rate at each Rank is zero, so GTR^i_{nrm} is equal to GTR^i at all Ranks. Looking at the least upper bound, the ratio of the $TRR^3_{avg,G}$ to GTR^3 is 1/2, and the sum of GTR^2 and GTR^3 exceeds GTR^2_{max} by 10 tokens per second, so the average Transient Bypass rate is a maximum of 5 tokens per second. Looking at the greatest lower bound, the overflow from Rank 3 plus GTR^2 equals GTR^2_{max}, so the average Transient Bypass rate in this case is a minimum of zero. The consequence of this Token Bypass rate is that the actual rate of tokens declared Green at Rank 2 will be up to 5 tokens per second less than the intuitive rate of tokens declared Green, and the actual rate of tokens declared Green at Rank 1 will be up to 5 tokens per second more than the intuitive rate.

B.2.2.4 Observations Regarding the Average Transient Bypass Rate

Figure A1-7 shows a plot of the least upper bound and greatest lower bound of the average Transient Bypass rate at Rank $n-1$ versus the average Token Request rate at Rank n. The shaded area between the bounds represents possible values of the average Transient Bypass rate. The exact value of the average Token Request rate is uncertain because it depends on the specific sequence of Token Requests at all Ranks. Nonetheless it is possible to use Figure A1-7 to make some observations about what value of the average Transient Bypass rate to expect for given GTBA parameters and average Token Request rate.

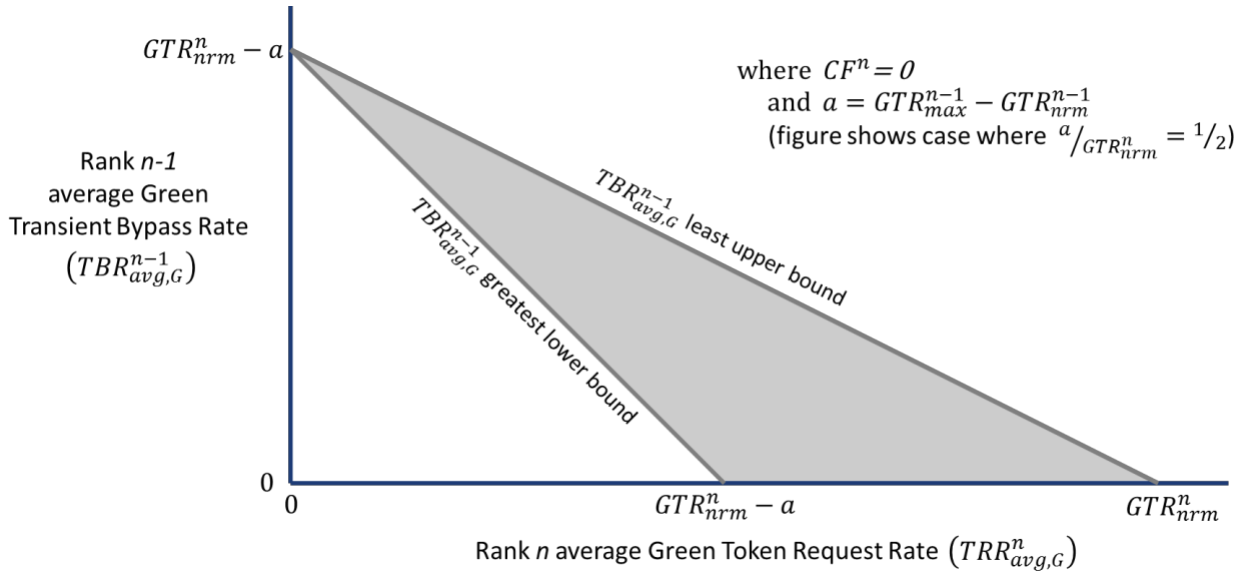

Figure A1-7 – Example of the Rank *n-1* Average Token Bypass Rate Bounds

Figure A1-7 assumes the Coupling Flag at Rank n is zero. As can be seen in Figure A1-1, the tokens shared from Rank $i+1$ is always equal to zero when $CF^{i+1} = 1$. This leads to the first, somewhat trivial, observation:

1. When the Coupling Flag at Rank $i+1$ is one, there will be no tokens shared from Rank $i+1$ to Rank i, and therefore there will be no Transient Bypass at Rank i.

Figure A1-7 also assumes that $(GTR_{max}^{n-1} - GTR_{nrm}^{n-1})/GTR_{nrm}^{n} = \frac{1}{2}$, however observations 2 through 5 are true for any values of the GTBA parameters. As the average Token Request rate at Rank n decreases, the number of tokens Overflowing the token bucket at Rank n increases. Therefore the average rate of tokens shared to Rank $n-1$ increases, causing the average Transient Bypass rate at Rank $n-1$ to increase.

2. The value of $TBR_{avg,G}^{n-1}$ increases as $TRR_{avg,G}^{n}$ approaches zero. When $TRR_{avg,G}^{n} = 0$, the value of $TBR_{avg,G}^{n-1} = GTR_{nrm}^{n} - (GTR_{max}^{n-1} - GTR_{nrm}^{n-1})$.

As the average Token Request rate at Rank n approaches the new token source rate at Rank n, the average rate of tokens shared to Rank $n-1$ decreases, and therefore the average Transient Bypass rate at Rank $n-1$ decreases.

3. The value of $TBR_{avg,G}^{n-1}$ decreases as $TRR_{avg,G}^{n}$ approaches GTR_{nrm}^{n}. When $TRR_{avg,G}^{n} \geq GTR_{nrm}^{n}$, $TBR_{avg,G}^{n-1} = 0$.

The uncertainty (U) in the value of the average Transient Bypass rate at Rank $n-1$ is the difference between the least upper bound and greatest lower bound lines in Figure A1-7. U can be calculated as follows:

$$U = \begin{cases} \left(1 - \dfrac{TRR_{avg,G}^{n}}{GTR_{nrm}^{n}}\right)(GTR_{nrm}^{n} - a) - \left(GTR_{nrm}^{n} - TRR_{avg,G}^{n} - a\right) & when\ TRR_{avg,G}^{n} < GTR_{nrm}^{n} - a \\ \left(1 - \dfrac{TRR_{avg,G}^{n}}{GTR_{nrm}^{n}}\right)(GTR_{nrm}^{n} - a) & when\ TRR_{avg,G}^{n} \geq GTR_{nrm}^{n} - a \\ a - \dfrac{a^2}{GTR_{nrm}^{n}} & when\ TRR_{avg,G}^{n} = GTR_{nrm}^{n} - a \end{cases}$$

$$where\ CF^{n} = 0\ \ and\ \ a = GTR_{max}^{n-1} - GTR_{nrm}^{n-1}$$

Figure A1-8 shows the uncertainty in the value of the average Transient Bypass rate at Rank $n-1$ as the average Token Request rate at Rank n varies from zero to GTR_{nrm}^{n} where the Coupling Flag at Rank n is zero and $(GTR_{max}^{n-1} - GTR_{nrm}^{n-1})/GTR_{nrm}^{n} = \frac{1}{2}$.

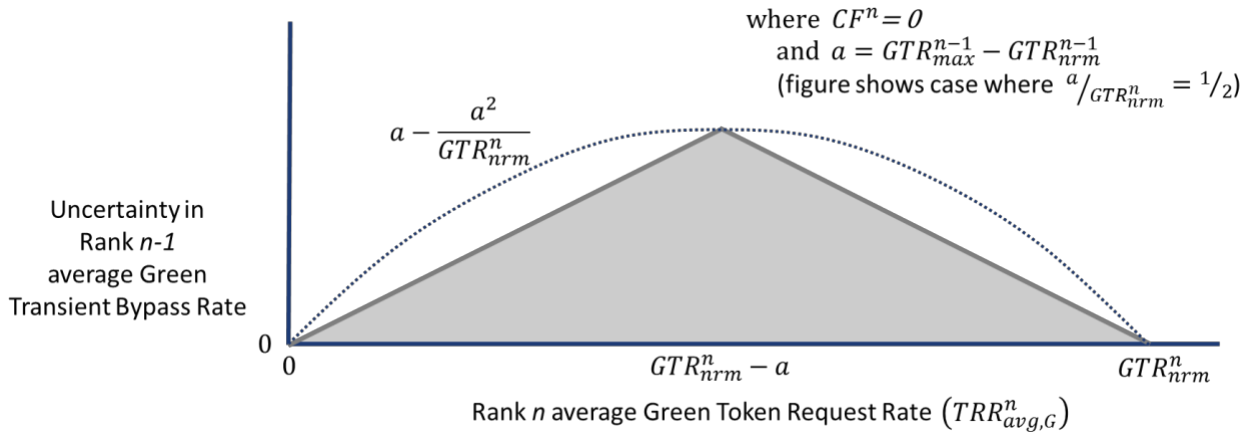

Figure A1-8 – Uncertainty in the Rank *n-1* Average Token Bypass Rate

The smallest uncertainty occurs when then average Token Request rate at Rank n is either zero or greater than or equal to the new token source rate at Rank n. The largest uncertainty occurs when the greatest lower bound line intersects the horizontal axis. This is when the average Overflow rate from Rank n (the difference between the new token source rate and the average Token Request rate at Rank n) is equal to the difference between the maximum token rate and the new token source rate at Rank $n-1$. At this point the uncertainty is equal to the least upper bound.

4. The uncertainty in $TBR^{n-1}_{avg,G}$ decreases as $TRR^n_{avg,G}$ approaches zero, and as $TRR^n_{avg,G}$ approaches GTR^n_{nrm}. The uncertainty in the $TBR^{n-1}_{avg,G}$ is zero when $TRR^n_{avg,G} = 0$ or $TRR^n_{avg,G} \geq GTR^n_{nrm}$.
5. The maximum uncertainty in the $TBR^{n-1}_{avg,G}$ occurs when $TRR^n_{avg,G} = GTR^n_{nrm} - (GTR^{n-1}_{max} - GTR^{n-1}_{nrm})$. At this point the greatest lower bound of $TBR^{n-1}_{avg,G}$ is equal to zero and the least upper bound is equal to $a - a^2/GTR^n_{nrm}$ where $a = GTR^{n-1}_{max} - GTR^{n-1}_{nrm}$.

Figure A1-7 and Figure A1-8 have been drawn using a ratio of "a" to GTR^n_{nrm} of ½, which is the same as in the example in Table A1-2. These figures can also help visualize how the average Transient Bypass rate is affected by changes in the GTBA parameters. Consider what happens to Figure A1-7 when the difference between the maximum token rate and the new token source rate at Rank $n-1$ (labeled "a" in Figure A1-7) changes. As the difference gets smaller, the intersection of both the least upper bound and greatest lower bound lines with the vertical axis moves up, and the intersection of the greatest lower bound line with the horizontal axis moves right. The shaded area representing the possible values of the average Transient Bypass rate becomes an increasingly steep narrow triangle. When the difference equals zero the triangle becomes a line where the average Transient Bypass rate at Rank $n-1$ equals the new token source rate at Rank n minus the average Token Request rate at Rank n. This makes sense since as the difference between the maximum token rate and the new token source rate gets small, very few tokens shared from the higher Rank can be added to the token bucket, and almost all shared tokens bypass the token bucket. Furthermore, as the intersection of the greatest lower bound line with the horizontal axis in Figure A1-7 moves right, the peak of the triangle of the uncertainty in Figure A1-8 follows the curved dashed line down and right until it also reaches zero.

6. The value of $TBR_{avg,G}^{n-1}$ increases and the uncertainty in $TBR_{avg,G}^{n-1}$ decreases as $GTR_{max}^{n-1} - GTR_{nrm}^{n-1}$ approaches zero. When $GTR_{max}^{n-1} - GTR_{nrm}^{n-1} = 0$, the uncertainty in $TBR_{avg,G}^{n-1}$ is zero and $TBR_{avg,G}^{n-1} = GTR_{nrm}^{n} - TRR_{avg,G}^{n}$.

As the difference between the maximum token rate and the new token source rate at Rank $n-1$ increases, the intersection of both the least upper bound and greatest lower bound lines with the vertical axis moves down, and the intersection of the greatest lower bound line with the horizontal axis moves left. The shaded area representing the possible values of the average Transient Bypass rate becomes an increasingly long narrow triangle just above the horizontal axis. When the difference equals the new token source rate at Rank n, the triangle becomes a line where the average Transient Bypass rate equals zero. This makes sense since as the difference between the maximum token rate and the new token source rate at Rank $n-1$ gets close to the new token source rate at Rank n, almost all tokens shared from the higher Rank can be added to the token bucket, and very few shared tokens bypass the token bucket. Furthermore, as the intersection of the greatest lower bound line with the horizontal axis in Figure A1-7 moves left, the peak of the triangle of the uncertainty in Figure A1-8 follows the curved dashed line down and left until it also reaches zero. As an example, Figure A1-9 shows the uncertainty plot when the ratio of "a" to GTR_{nrm}^{n} is changed to ¾.

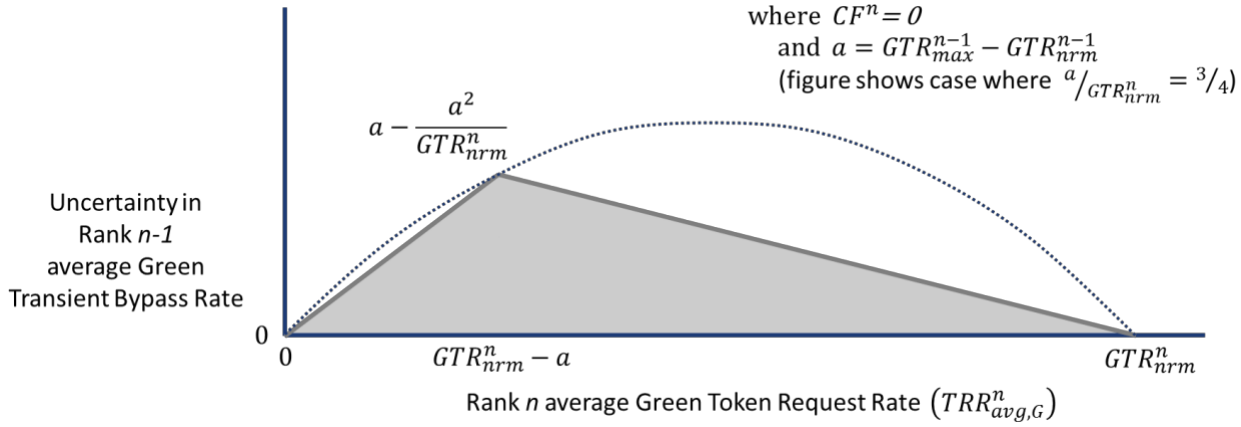

Figure A1-9 – Uncertainty in the Rank *n-1* Average Token Bypass Rate

7. The value of $TBR_{avg,G}^{n-1}$ decreases and the uncertainty in $TBR_{avg,G}^{n-1}$ decreases as $GTR_{max}^{n-1} - GTR_{nrm}^{n-1}$ approaches GTR_{nrm}^{n}. When $GTR_{max}^{n-1} - GTR_{nrm}^{n-1} \geq GTR_{nrm}^{n}$, the uncertainty in $TBR_{avg,G}^{n-1}$ is zero and $TBR_{avg,G}^{n-1} = 0$.

The largest uncertainty in the average Transient Bypass rate at Rank $n-1$ is always where the greatest lower bound line intersects the horizontal axis, which is where $TRR_{avg,G}^{n} = GTR_{nrm}^{n} - (GTR_{max}^{n-1} - GTR_{nrm}^{n-1})$. As $GTR_{max}^{n-1} - GTR_{nrm}^{n-1}$ varies from zero to GTR_{nrm}^{n}, the peak of the uncertainty triangle in Figure A1-8 follows the dotted line shown in the figure. The overall maximum occurs when the average Token Request rate at Rank n and the difference between the maximum token rate and the new token source rate at Rank $n-1$ are both equal to $GTR_{nrm}^{n}/2$.

8. The maximum uncertainty in the average Transient Bypass rate at Rank $n-1$ over any value of the average Token Request rate at Rank n and any value of the difference

between the the maximum token rate and the new token source rate at Rank $n-1$ occurs when $TRR^{n}_{avg,G} = (GTR^{n-1}_{max} - GTR^{n-1}_{nrm}) = GTR^{n}_{nrm}/2$. In this case the greatest lower bound of the average Transient Bypass rate is zero, and the least upper bound is equal to $GTR^{n}_{nrm}/4$.

Looking back at the example in Table A1-2, the average Overflow rate from Rank 3 is 10 tokens per second, and the difference between the maximum token rate and the new token source rate at Rank 2 is 10 tokens per second, meeting the criteria for observation 5. One half of the new token source rate at Rank 3 is also 10 tokens per second, meeting the criteria for observation 8. Therefore this is an example where the overall uncertainty in the average Transient Bypass rate is at a maximum, with the greatest lower bound at zero and the least upper bound at one fourth the new token source rate at Rank 3, or 5 tokens per second.

A final observation can be made regarding the likelihood of observing an average Transient Bypass rate at Rank $n-1$ near the least upper bound, as opposed to near the greatest lower bound, when the GTBA is used for an ingress bandwidth profile at a UNI or ENNI. In this case Token Requests correspond to frames that arrive sequentially at the interface. The probability of having a token count update that exactly fills the Rank n bucket increases as the number of frames (at any Rank) increases, and as the rate of tokens requested approaches the bandwidth of the interface. Therefore the likelihood of observing an average Transient Bypass rate at Rank $n-1$ near the least upper bound increases at heavily utilized interfaces.

8 References

[1] MEF 41, *Generic Token Bucket Algorithm*, October 2013.