TTC標準
Standard

The difference between TTC JT-Y1221 and ITU-T Y.1221

Traffic control and congestion control in IP based networks
(The English Edition)

Version 1.0

Published on March 27, 2013

THE TELECOMMUNICATION TECHNOLOGY COMMITTEE

Telecommunication
Technology
Committee

Contents

Introduction

This document provides the English Edition.
In case of dispute, the original to be referred is the Japanese edition of the text.

This document provides the difference between TTC standard JT-Y1221 (Version 2.0, Feb 2013) and ITU-T Recommendation Y.1221 (Version 1.0, Mar 2002).

Change History

Version	Date	Outline
1.0	March 27, 2013	Published.

Industrial Property Rights
Information regarding submittal of TTC's "The Policy for the Handling of Industrial Property Rights" is available on TTC's website

Responsible working group
NGN & FN Working Group

TTC JT-Y1221 supplements ITU-T Y.1221 with the following item as an annex.

(a) IP Traffic Specification Method to Ensure End-to-end QoS. (Annex a)

See "TTC Standard Summary" in TTC Website (http://www.ttc.or.jp/e/) for the summary of difference between TTC standards and referred international standards (ex. ITU-T recommendations).

Annex a. IP Traffic Specification Method to Ensure End-to-end QoS

(This annex forms an integral part of this standard)

a.1. Scope

a.1.1 Summary

This annex determines specifications on media traffic related to guaranteeing QoS, which is a feature of next-generation networks (NGNs)[Y.2001].

NGNs have features such as resource admission control and policing, and networks must use some or all of these functions to realize ensuring end-to-end QoS. Also, for traffic spanning two networks, the traffic on each of the networks must be regulated appropriately to guarantee end-to-end QoS. This annex determines specification of traffic necessary to guarantee end-to-end bandwidth in these sorts of environments.

a.1.2 References

The following references are used in this annex.

[Y.2001] "General overview of NGN" TTC standard JT-Y2001, version 1, The Telecommunication Technologies Committee, 2006

[RFC 3550] "RTP:A Transport Protocol for Real-Time Applications" TTC standard JF-IETF-STD64, The Telecommunication Technologies Committee, May 2005.

[RFC3556] "Session Description Protocol (SDP) Bandwidth Modifiers for RTP Control Protocol (RTCP). Bandwidth" TTC standard JF-IETF-RFC3556, version 1, The Telecommunication Technologies Committee, May 2009.

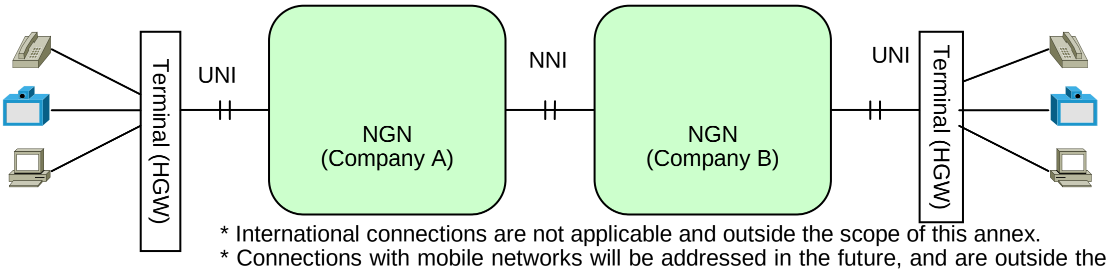
[RFC 4566] "SDP:Session Description Protocol" TTC standard JF-IETF-RFC4566, The Telecommunication Technologies Committee, Mar 2007.

[TR-1014] "Overview of the NGN architecture" TTC technical report TR-1014, version 1, The Telecommunication Technologies Committee, Jun 2006.

[National Regulation] "Regulations for Commercial telecommunications equipment," MIC Ordinance 101, Latest revision, Sept. 17, 2008.

a.1.3 Model used in this annex

In this annex, we assume a network model with the NGNs of two operators connected as shown in **Fig. a-1** for the purposes of studying end-to-end QoS guarantees. We consider international connections to be out of the scope of this model, and leave mobile networks as an issue for future consideration.



* International connections are not applicable and outside the scope of this annex.
* Connections with mobile networks will be addressed in the future, and are outside the scope of this annex.

Fig. a-1/JT-Y1221 NGN connection model for this annex

a.1.4 The mechanism for ensuring end-to-end QoS and the need for traffic specification method on next-generation networks

NGNs allow services with differing conditions (traffic characteristics, quality requirements) to be offered on the same network at the same time. To achieve this, NGNs perform end-to-end (UNI-to-UNI, UNI-to-NNI) QoS control. Best-effort networks that do not guarantee communications quality do not have such a mechanism.

End-to-end QoS control consists of two functions, discussed below (**Fig. a-2**).

The first is called the Resource and Admission Control Function (RACF)[TR-1014]. In the NGN architecture, the call connection procedure determines whether the required bandwidth is available between the terminals and network for each media-traffic type and QoS class. If it is available, the NGN allocates bandwidth to that session and guarantees communication quality by prioritizing communication processing according to QoS class. If the bandwidth is not available, the NGN is unable to allocate the bandwidth needed to guarantee communications quality for that session, so the session is not admitted.

The other is policing function which provides policing on a media-traffic basis (a traffic flow monitoring function). This function monitors whether traffic exceeding the bandwidth allocated by the RACF is entering the network. If traffic exceeding the allocated bandwidth enters the network, not only can the quality of that session no longer be guaranteed, but the bandwidth allocated to other sessions can be affected. To avoid such conditions, the incoming traffic bandwidth is strictly monitored by the media-traffic policing function, and if incoming traffic exceeding the allocated bandwidth is detected, the offending packets are discarded. Accordingly, the network must transmit traffic such that the allocated bandwidth is preserved, and sessions established by terminals can receive the transmission quality provided by the network. In doing so, there will be a danger of conflict if traffic bandwidth and burstiness is not regulated consistently among terminals, the policer and admission control functions.

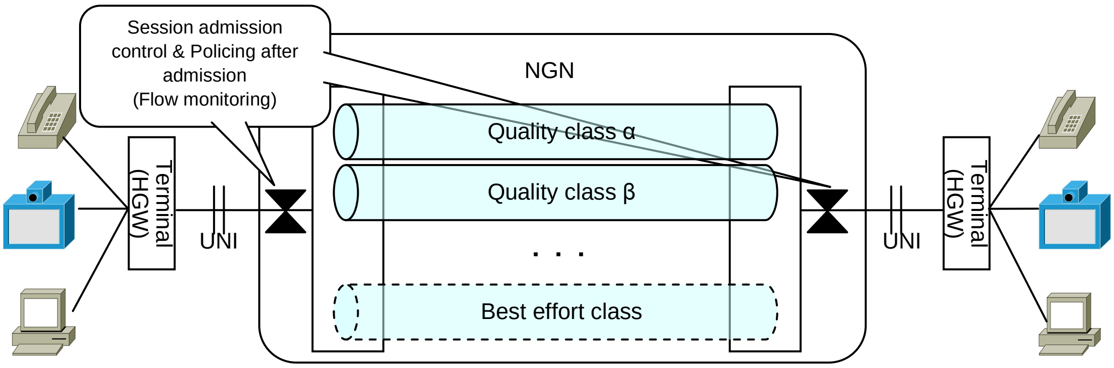


Fig. a-2/JT-Y1221 End-to-end QoS control in NGN

a.1.4.1. Resource and Admission Control Function

NGNs determine whether the bandwidth requested through call control signals is available or not on a media-traffic and quality class basis between terminals and the network. This function is called the Resource and Admission Control Function (RACF).

An overview of the RACF is shown in **Fig. a-3**. The NGN compares bandwidth requested for a new session with the available bandwidth in the QoS class, per media type. If the bandwidth is available, the session is accepted and the bandwidth is allocated. Then, transmission quality is guaranteed by prioritizing transmission processing according to QoS class. If the bandwidth is not available, the requested transmission quality cannot be guaranteed, so the session is not admitted.

The policing function, discussed in section a.1.4.2 below, monitors whether the bandwidths allocated by the RACF are maintained for each media type.

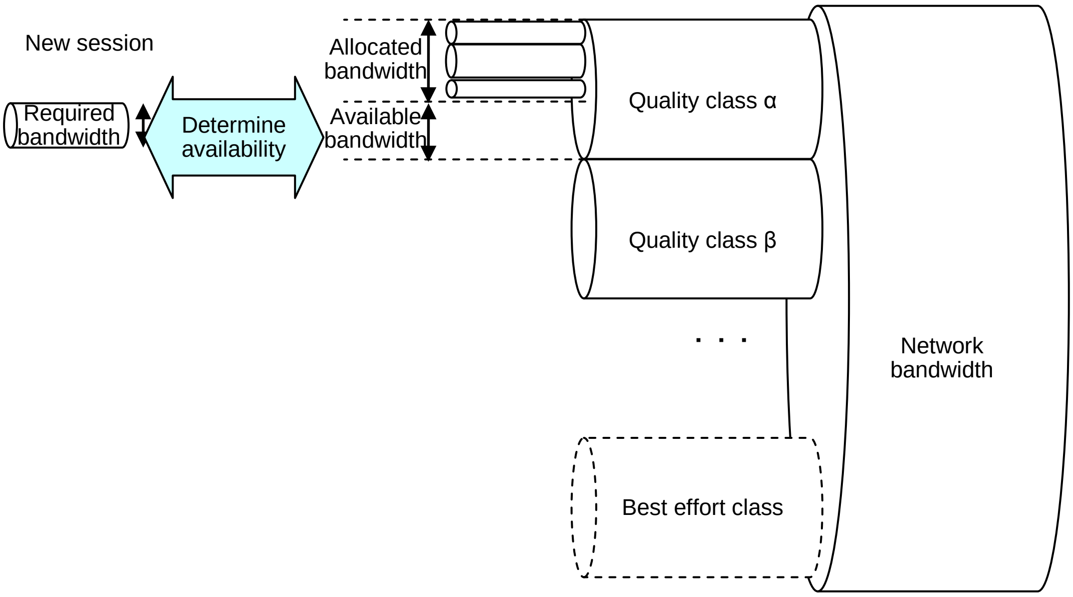


Fig. a-3/JT-Y1221 Session admission control function

a.1.4.2. Policing function

The policing function (traffic flow monitoring function) monitors traffic flow. On NGNs, it monitors whether the traffic flowing into the NGN at UNI and NNI is according to the bandwidths allocated for each media type by the RACF, as described in section a.1.4.1.

a.1.5 Bandwidth specification using a token bucket model

A token bucket model is used to specify bandwidth and burst characteristics of traffic by media type (Appendix I). With the token bucket model, the bandwidth and burst characteristics of traffic can be expressed with two parameters: the token bucket rate and size.

a.1.6 Scope of traffic specification

This annex gives provisions for expressing traffic bandwidth and burst characteristics that can be shared among UNI, between terminal and network, NNI, between network and network, and for terminal bandwidth reports between terminal and server. It uses the token bucket model as a reference in determining methods for specifying traffic bandwidth and burst characteristics. End-to-end transmission quality can be guaranteed by having terminals and network each send and transmit traffic correctly according to their responsibilities in this reference model. Traffic specification using the token bucket model in this annex applies to the NGN traffic in general. However, specific token bucket parameter values are determined based on units of media traffic used in SIP/SDP admission and control.

a.2. Determining token bucket parameters

a.2.1 Token bucket parameters

With the token bucket model, traffic bandwidth and burstiness are specified using two parameters: the token bucket rate and size. The token bucket rate expresses the bandwidth, while the token bucket size expresses the burstiness of the traffic. Also, for the parameters needed to determine that all packets of the same traffic type conform, the token bucket rate can be increased and the token bucket size can be reduced. Also, there is a trade-off relationship between them, in that by increasing the token bucket size, the token bucket rate can be reduced.

On the other hand, to guarantee transmission quality on the network, it is not adequate to simply guarantee bandwidth; each piece of transmission equipment must have buffers adequate to allow for the burstiness of the traffic. In other words, to guarantee end-to-end transmission quality, both the bandwidth and buffering resources of each piece of equipment must be managed appropriately. To do so, when regulating the bandwidth and burstiness of a given traffic type, care must be taken that both parameters are set appropriately, ensuring that neither the token bucket rate or size are too large.

a.2.2 Determining token bucket speed

The token bucket rate parameter expresses the average characteristics of the traffic. In the token bucket model, communication does not exceed this rate over a set time interval. Thus, when reporting bandwidth to the resource and admission control function with the SIP/SDP b= line, it is appropriate to consider the average rate per media type as the applicable rate. However, most transmission equipment on a network transmits data in IP packet units, so the bandwidth reported on the SDP b= line includes overhead such as IP headers, and not just the payload bandwidth, and is the token bucket rate.

a.2.2.1. Specifying RTP bandwidth

For media traffic using RTP, the token bucket rate is set to the b=AS line of the applicable traffic media type.

The value set on the b=AS line is the RTP bandwidth, not including RTCP bandwidth.

To set the RTCP bandwidth, the b=RR and b=RS lines can be used, as stipulated in JF-IETF-RFC3556[RFC3556].

When using the b=RR and b=RS lines, they are used for the RTCP token bucket rate value. If the b=RR and b=RS lines are not specified, using 5% of the RTP bandwidth for the RTCP bandwidth is recommended, as stipulated in section 6.2 of JF-IETF-STD64[RFC3550].

a.2.2.2. Low-layer overhead considerations

Overhead such as low-layer headers must also be considered, as indicated in section 5.8 of JF-IETF-RFC4566[RFC4566].

As indicated in section 6.2 of JF-IETF-STD64, the bandwidth specified on the b=AS line includes the layer 4 and layer 3 bandwidths. Specifically, the bandwidth on the b=AS line includes RTP, UDP, and IP headers, but it does not include layer 2 protocol overhead, such as Ethernet frame headers.

a.2.2.3. Burstiness considerations

On conventional best-effort networks, bandwidth is usually handled as an average of a longer time period (on the order of seconds). In contrast, with NGNs, bandwidth is managed based on average rates over shorter time periods (on the order of 10 ms), depending on the token bucket policer. When designing NGN terminals, any discrepancy between these long-term average rates and the short-term average rates specified on the b=AS line must be considered.

The characteristic of issuing continuous, large amounts of traffic for short periods of time is called burstiness. Even if the long-term average rate is less than the rate specified on the SDP b=AS line, care must be taken because packets may be discarded by the token packet policer if the traffic issued is very bursty.

Points for consideration regarding video communications, which tends to be quite bursty, are given below.

- Video is encoded in frame units, but sending encoded data in single frame units increases burstiness. When sending RTP packets, traffic should be shaped to transmit it steadily.
- Video codecs generally use inter-frame compression techniques, and frames with inter-frame compression tend to use less data than those with compression only within the frame, so this can increase burstiness. When encoding, bit-rate allocation should be adjusted for each frame to equalize it, or shaping should be used when sending RTP packets to smooth out the traffic.

a.2.3 Determining token bucket size

Token bucket size is the parameter that expresses burstiness of the traffic. With current SIP/SDP, it is difficult to specify values for token bucket size directly, so we specify rules for determining token bucket sizes from existing attributes described by SDP. The burstiness of traffic is expected to vary according to application, but generally, if the average bandwidth is large, the traffic will also tend to be more bursty. Thus, a positive correlation between the two is expected, with token bucket sizes tending to be larger for higher token bucket rates, and we treat them as being linearly proportional (as in Appendix IV, Example 2). However, when applying this proportional relationship, the token bucket size becomes small when the token bucket rate is small, and can result in refusal to transmit even a single packet. Also, for large token bucket rates, the proportional token bucket size is large, so the size of buffer resources needed to permit bursts on the network also becomes large, and could exceed the buffers implemented in equipment. Thus, it is not practical to apply this proportional relationship to token bucket rates in all domains. On the other hand, in order to define quantitatively, the range over which this proportional relationship can be applied, factors such as IP packet lengths, which depend heavily on the application, and network buffer resources, which depend heavily on equipment implementations, must be considered. It is difficult to define such conditions within this annex, so we assume it will be applied for traffic with a token bucket rate in the range from 100 kbit/s to several Mbit/s. In operation, if it is difficult to use this proportional relationship upon consideration of the IP packet lengths for individual applications, or buffer implementations in equipment, token bucket sizes a different method will need to be used.

a.2.4 Definition of rate factor

When there is a proportional relationship between token bucket rate and size, the slope corresponds to the time required for the bucket to fill up from zero to the maximum number of tokens. This interval is defined as the rate factor. In other words, the token bucket size, which is a token bucket parameter, can be calculated by (token bucket rate) × (rate factor). Thus, even if the token bucket rate is the same, a large rate factor corresponds to a large token bucket size, and a small rate factor corresponds to a small token bucket size. Rate factor is shown in **Fig. a-4**.

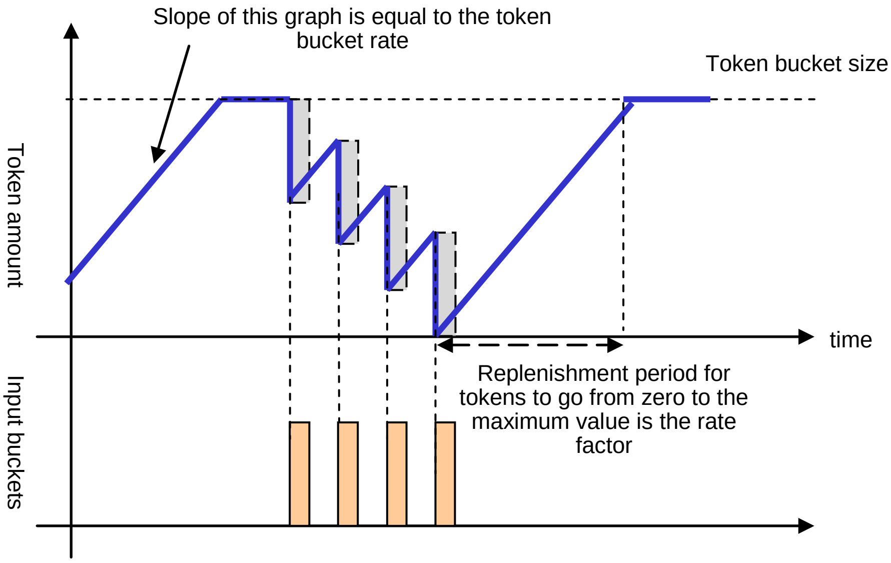

Slope of this graph is equal to the token bucket rate
Token bucket size
Token amount
time
Replenishment period for tokens to go from zero to the maximum value is the rate factor
Input buckets

Fig. a-4/JT-Y1221 Token bucket speed and rate factors

a.2.5 Concrete values of rate factors

In this section we determine some concrete rate factors. Regulating rate factor values has benefits in that it clarifies the performance requirements for terminals, because it determines the traffic conditions accepted by the

NGN, and it clarifies design and control conditions for network equipment, because it determines traffic conditions of transmission quality that must be guaranteed by network operators. We also assume that QoS requirements and traffic handling within the network will differ by transmission QoS class, and application service or application, so a rate factor value, which regulates burstiness, is needed for each transmission QoS class. In this annex, we assume two types of application on NGN, represented by telephone-type and distribution-type applications, and we discuss the QoS classes for providing these applications as QoS class α and QoS class β, respectively. We stipulate rate factors for QoS class α, which assumes telephone-type two-way communication in this annex, and stipulation of rate factors for QoS class β, which assumes distribution-type bi-directional communication is left for future work.

a.2.5.1. Rate factor for QoS class α, telephone-type two-way traffic (UNI)

This class assumes telephone-type two-way communication, so it requires high communication quality. When comparing video and sound media assumed to be used with this QoS class, video media traffic is more bursty, and will require a higher rate factor to allow for bursts. Because of this, the rate factor must be set with attention to the periodic nature of the traffic, especially for video communication such as video calling. If we assume 30 frame/sec coding as standard for video communication such as video calls (and conferencing), the frame interval is 33 ms. To allow data to be sent in single-frame units, the rate factor must be at least 33 ms (**Fig. a-5**). However, to only allow for traffic bursts of a single cycle is somewhat strict and will not allow for any real variance, so a rate factor of 70 ms is set for QoS class α, allowing for bursts of approximately two frames in telephone-type two-way communication.

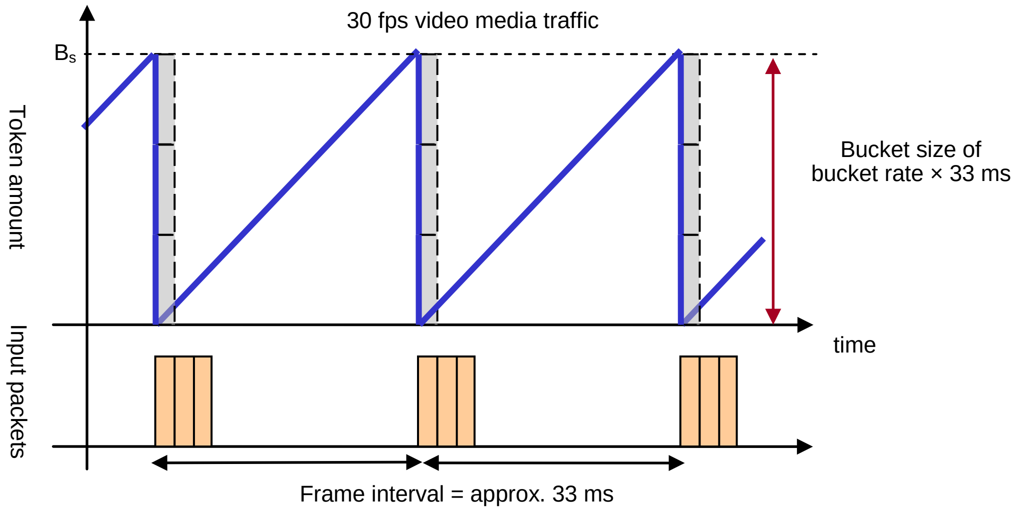

30 fps video media traffic
B_S
Token amount
Bucket size of bucket rate × 33 ms
time
Input packets
Frame interval = approx. 33 ms

Fig. a-5/JT-Y1221 Operation of a token packet policer permitting only a single frame amount

a.2.5.2. Rate factor for QoS class α, telephone-type two-way traffic (NNI)

Burstiness of traffic at an NNI consists of two factors; the burstiness of the source traffic, and that resulting from varying transmission delays within the network. On the other hand, NNI must naturally permit the traffic received on a UNI, so the rate factor must be set, adding to the rate factor for UNI to account for variation in transmission delay within the network. Also, voice traffic is generally more sensitive to delay than video, although they are both transported in this QoS class. Since only one QoS level is possible for a given QoS class, quality suitable for voice, which is more sensitive to delay, must be provided. Moreover, the MIC ordinance, "Commercial communications equipment regulations [National Regulation]," regulates the QoS standard for IP telephony in number 0AB-J, such that the variance in transmission delay between UNI-NNI must be within 10 ms. Considering these issues, the NNI rate factor for QoS class α is set to 80 ms, adding 10 ms to the UNI rate factor of 70 ms to allow for delay variation within the network.

a.2.5.3. Rate factor for QoS class β, distribution-type one-way traffic (UNI)

Since the traffic assumed for QoS class β is one-way distribution, it does not require the level of quality needed by QoS class α, which is for telephone-type two-way communication, so it is appropriate to permit larger bursts in the network. However, in setting concrete values, values for UNI and NNI must be considered together. As discussed below, the NNI rate factor is left for future consideration in this annex, so the UNI rate will also be left for future study.

a.2.5.4. Rate factor for QoS class β, distribution-type one-way traffic (NNI)

As with the QoS class for telephone-type bi-directional communication, it is desirable that the NNI rate factor be determined by adding an amount to allow for variation in delay within the network. However, regulations for variation in delay permitted for QoS class β within the network and service requirements are not well-defined, so this is left for future consideration.

a.3. Token bucket parameter specifications

a.3.1 scope

In this section, we specify token bucket parameters for the case with packetization at 20 ms intervals using G.711 μ-law codec, which is the most common use for IP telephony services.

a.3.2 Token bucket rate specification

The token bucket rate parameter expresses the average characteristics of the traffic. If it is exceeded over long periods, communication cannot continue, so a value greater than the average speed of the traffic over a long period of time must be used. Also, the packetization interval is a fixed condition, and the case of packets with maximal length gives the highest average speed, so we assume the case with the longest packet length, and with all optional IPv4 header fields used.

Average speed (bps) = packet size × packets/second ×8×RTCP overhead (*)
= (IP header length + UDP header length + RTP header length + RTP payload)×50×8×1.05
= (60+8+12+160)×50×8×1.05
= 100.8 kbps
(*) RTCP overhead is set at 5% [RFC 3550]

There is room for fluctuation in the RTCP overhead and some safety margin is needed, so the token bucket rate is set at 105 kbps.

a.3.3 Token bucket size specification

We now calculate the required bucket size needed from the above token bucket rate and the NNI rate factor assumed for telephone-type services (80 ms). We compare this value with the minimum bucket size required to pass a single packet (up to 1 500 bytes) and specify the larger as the token bucket size.

Required bucket size = 80 ms × 105 kbps = 8 400 bits (1 050 bytes)

However, the maximum packet size is assumed to be 1 500 bytes, and the bucket size must be larger than this to allow packets to pass, so the token bucket size is set to 12 000 bits (1 500 bytes).