

TSVWG
Internet-Draft
Intended Status: Informational
Expires: August 25, 2013

K. Carlberg
G11
P. O'Hanlon
UCL
Feb 25, 2013

Reactions to Signaling from ECN Support for RTP/RTCP
<draft-carlberg-tsvwg-ecn-reactions-04.txt>

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on August 25, 2013.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Abstract

This document presents an examination of various responses to Congestion Experience (CE) notifications by real time applications that have negotiated end-to-end support of Explicit Congestion Notification (ECN). This document is a follow-on effort of [rfc6679], which specifies the signaling used to provide ECN support for RTP/RTCP flows.

1. Introduction

This document presents an examination of various responses to Congestion Experience (CE) notifications by real time applications that have negotiated end-to-end support of Explicit Congestion Notification (ECN). [rfc6679] defines the signaling for support of ECN by RTP based sessions and also covers the case where a set of nodes do not respond to CE notifications. A more detailed discussion about how back-off algorithms can be achieved, as well as other potential reactions, is viewed as out of scope of that document and may be addressed by a companion document.

1.1 Background

ECN is a mechanism used to explicitly signal the presence of congestion without relying on packet loss. It was initially designed using a dual layer signaling model; negotiation and feedback at the transport layer, and downstream notification of congestion at the network layer. For IP, a new two bit field was used to both indicate the successful negotiated support for ECN signaling, as well as indicate the presence of congestion via the CE flag. In the case of TCP [rfc3168], a new TCP header flag was defined that provides upstream end-to-end indication of congestion occurring somewhere along the downstream path.

There should be no difference in congestion response if ECN-CE marks or packet drops are detected. However it is noted that there MAY be other reactions to ECN-CE specified in the future. Such an alternative reaction MUST be specified and considered to be safe for deployment under any restrictions specified. We specify such an alternative in this document.

With respect to ECN for TCP, [rfc3168] specifies an indication of congestion, but it does so once per Round Trip Time (RTT). [rfc6679] is an effort that proposes a finer grained notification reflecting a more accurate indication of the number of ECN marked packets received within one RTT. It should be noted that there is also other on going work to provide more accurate ECN feedback information for TCP [draft-tcpm-accecn-reqs].

1.2 Terminology and Abbreviations

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC2119 [RFC2119].

2. Issues

The initial discussions and presentation of [draft-rtp-ecn] produced a consensus that the specification of signaling was to be done within the AVTcore working group, and any subsequent discussion on end-to-end reactions to the signaling would be accomplished in the Transport Services (TSV) working group. This draft satisfies the latter effort.

Another issue that needs to be recognized is that the reactions to CE in the context of [rfc6679] are the responsibility of the application. This is in contrast to ECN support for TCP, where explicit signaled feedback of, and reaction to, CE is kept transparent to the application. The issue of placing the feedback responsibility in the application is that each application needs to add specific support for that reaction. On the other hand, multiple reactions may be considered by the application. For this reason, [rfc6679] states the need for a default congestion control reaction that MUST be supported. Section 3 through 5 expands on this topic.

3. Congestion Control Algorithms

The transport of any data flow across the Internet produces a need for some form of congestion control to attain a suitable share of the capacity of the path through a network. Most of the existing work on realtime congestion control algorithms has been rooted in TCP-friendly approaches but with smoother adaptation cycles. TCP congestion control is unsuitable for interactive media for a number of reasons including the fact that it is loss-based so it maximizes the latency on a path, it changes its transmit rate to quickly for multimedia, and favors reliability over timeliness. In the case of real time media transport, one requires:

Smoother rate variation: (than for bulk data) to accommodate the underlying media flow's characteristics.

Low latency: Maintaining latencies sufficient to be usable, where 150ms is understood to be a good target [ITU.G114.2003].

Burst handling: Ability to handle bursts due to the nature of the media and codec (e.g. I-frames etc)

3.1 TCP Friendly Rate Control (TFRC)

TFRC has a smoother response to congestion than TCP-like approaches, thus making it more suitable for real-time interactive multimedia applications. It has been cited in a number of other documents within the IETF for use with UDP and media flows [rfc3714, bcpl45] and is seeing full and partial deployment in related solutions such as Empathy/Farsight, and GoogleTalk [googl].

However it should be noted that TFRC is only recommended for real-time media use with ECN response. TFRC is not recommended for non-ECN paths due to its loss based operation which leads to full queues with maximised latencies. It is assumed that ECN markings will usually occur with lower queue occupancy and thus lower latency. However it is understood that ECN marks may not provide for sufficiently low latencies in some situations so other congestion control solutions would be preferable.

[rfc4342] specifies the profile for TFRC for use in the Datagram Congestion Control Protocol (DCCP) [rfc4340] for a half connection. A DCCP half connection is defined as application data sent downstream with corresponding acknowledgements sent upstream. These half-connections can be realized in the form of one-way pre-recorded media, one-way live media, or two-way interactive. A perceived drawback in this profile concerns its application to interactive media that use small packets. [RFC4828] is an experimental protocol defining a variation of TFRC used to address this drawback and achieve the same bandwidth as a TCP flow using packets of size 1500 bytes.

[rfc6679] is an standard that specifies how RTP flows can be supported using the RTP/AVPF profile and the general RTP header extension mechanism.

3.2 Related Work

3.2.1 3GPP

Outside of this previous and on-going work with TFRC, it is understood that some parties have issues with the behavior of TFRC under certain conditions. A notable mention of this is made in the 3GPP's document on IP Multimedia Subsystem (IMS) Media handling and interaction [TR26.114], where it is mentioned:

"Note that for IMS networks, which normally have nonzero packet loss and fairly long round-trip delay, the amount of bitrate reduction specified in RFC 3448 is generally too restrictive for video and may, if used as specified, result in very low video bitrates already at (for IMS) moderate packet loss rates."

Though it is unclear exactly what the 3GPP community consider as too restrictive and whether some alteration of the response may be suitable. It should be noted that the 3GPP document only referred to an older version of TFRC defined in [RFC3448]. Given that the current version of TFRC [RFC5348] has made significant changes to the idle and data-limited responses it is unclear whether their assessment is relevant to current TFRC implementations.

Furthermore the specification [TR26.114] only outlines a rudimentary approach to congestion control, providing an example of a 60% back-off reaction to loss within an RTCP reporting period. The proposed signalling employs Temporary Maximum Media Stream Bit Rate Request (TMMBR) [RFC5104] and Codec Mode Request (CMR) [RFC4867] for video and audio respectively, which would only provide for very basic rate control if used as specified. We note that [TR26.114] specifies terminal behavior, while [TS36.300] specifies base station behaviour, though neither specify any standardised congestion control approach.

It is understood that there are a number of proprietary and patented approaches that provide more sophisticated response in the case of 3G/LTE, but since these are neither endorsed nor standardized this document advocates a standardized approach such as TFRC.

We also acknowledge that there are many congestion control algorithms available for implementers to choose from, with a subset that are specifically suited to real time media transmission. However, given a variety of real time applications and their various characteristics (sender-only broadcast, interactive unicast, etc), we need to expand the notion of how back-off can be achieved. Hence, the focus needs to be on an output that would resemble the characteristics of TFRC.

3.2.2 RTCweb

Within the RTCweb Working Group the need for a more media friendly congestion control mechanism has been made apparent. Currently, TFRC is perceived as having deficiencies (e.g. its loss-based design, lack of cross-stream congestion control functionality etc) that make it an incomplete or insufficient solution for the envisioned RTCWEB media flows. The RTP Media Congestion Avoidance Techniques (rmcat) working group has now been formed which aims to lead to the formation of a working group on these issues. The group aims to develop one or more congestion control algorithms, associated extensions, and evaluation criteria. Furthermore it has been proposed that certain practices, such as 'circuit-breaker' conditions, to provide operational limits on congestion control algorithms, and feedback messages, may be tackled in other groups such as AVTCORE and AVTEXT respectively.

Thus there is some movement to attempt to develop new algorithms better suited to media transport, but these efforts will clearly take a considerable time to reach fruition.

3.3 ECN response

As mentioned above and in accordance to [rfc3168], the actual response to the reception of an ECN-CE marked packet MUST normally be the same as that of a lost packet. However there are a number of contexts where one

may also be interested in more varied approaches. We expand on this in Section 5 below.

4. Application Layer Congestion Response

Whilst the congestion control algorithm may decide to alter the rate at which the application should operate, in the case of media applications this process is not as straightforward as the case of bulk data. The different media engines and codecs in use may only have limited adaptation ranges, thus, this limitation needs to be a consideration when adapting the rate. Furthermore the application needs to be aware of the capability of the specific codecs in terms of their ability to switch configuration mid-stream (without loss of fidelity), which may impose further limits on the modes of operation.

One approach for achieving a lower generation of data is through reduced sampling of the media (e.g., voice or video). In the case of video, this may also involve slower frame rates. Specific recommendations that describe how applications should respond to congestion in the context of supporting the algorithmic characteristics of a congestion control algorithm are outside the scope of this document.

5. Other Reactions

In addition to the activation of congestion control algorithm, other reactions can be used or leveraged by an application in response to CE. We divide these other potential reactions into three categories: signaling, fault tolerance, and reduction. In the first two cases, we note that these other reactions are considered symmetric because they require downstream peer support. We also point out that activation of other reactions represents an example of an on-demand and as-needed approach in responding to CE.

In each case, we discuss issues that should be considered when contemplating a different reaction in the presence of CE feedback.

5.1 Signaling

5.1.1 RSVP

The resource Reservation Protocol (RSVP) can be used to signal a desired set of path characteristics (e.g., bandwidth, delay) in response to CE feedback [rfc2205]. Its operation is based on the use of PATH messages sent downstream hop-by-hop from the source to a destination that specify requested forwarding characteristics. In return, the destination sends a hop-by-hop RESV message upstream towards the source confirming the resources that have been reserved for that flow.

[rfc3181] defines a priority policy element that specifies both an allocation and defending priority. This dual specification supports the use of preemption of existing reservations. [draft-priority-rsvp] is a work-in-progress that defines a new policy element that only conveys priority during reservation establishment. This latter effort also presents several reservation models, including one that describes engineered resources set aside for priority users.

5.1.1.1 Issues

As discussed in [rfc-3583], RSVP presents a difficult challenge of establishing state and effectively and efficiently migrating it during roaming in mobile environments. Its soft state design allows the protocol to attempt re-establishment of reserved resources along new path(s), but there is no guarantee that resources along the new path will be available. In addition, there is at least 1 RTT of delay and the delta in initiating a new PATH message that delays reservation establishment.

Some user groups, such as those found in the military, make a distinction between mobile and transportable environments. The former case resembles scenarios attributed to Mobile IP. The latter case is characterized by wireless hosts operating in a new location, but never moving to the extent that new paths through a network need to be established. In this latter example, the challenges of RSVP in a wireless environment are diminished. In addition, these environments tend to involve a single administrative control of both hosts and routing/forwarding nodes within a network infrastructure.

RSVP is associated with a means of retaining a minimal bound of forwarding characteristics per flow, or aggregate of flows. As such, it can be considered to run contrary to the objectives of ECN. However, in cases where some flows must be reserved, CE feedback could be used to signal the need to lower a pre-existing killer app reservation.

5.1.2 Differentiated Services

Unlike RSVP and its use of a separate signaling mechanism to reserve resources, Differentiated Services (diff-serv) uses code points within the IP header to convey the forwarding behavior of that packet [rfc2474]. This may range from various drop precedence values to a code point that signifies low delay and low loss (i.e., characteristics attributed to real time flows).

As in the case of RSVP, applications could rely on the reception of CE feedback to initiate a subsequent setting of diff-serv code points to provide additional protection or explicit association of forwarding characteristics of a given flow of packets. In addition, the setting of

diff-serv code points would be done on an as-needed basis in reaction to CE feedback. Recommendations concerning specific diff-serv values are outside the scope of this document.

5.1.2.1 Issues

Given the ease by which applications or middle boxes can set diff-serv code points, the issue of trusting values other than best effort can become problematic when hosts and routing/forwarding nodes are not associated with a single administrative authority.

As in the case of RSVP, the effectiveness of diff-serv is dependent on the number of nodes along a path that support the protocol. Thus, as opposed to a single end-point reaction to CE feedback, differentiated services requires additional support in the network to either increase or decrease the probability of traffic being forwarded to its destination.

A symbiotic capability to consider is the use of on-demand/as-needed diff-serv code points to trigger downstream actions by the network. A specific example would be a diff-serv code point sent in reaction to CE feedback that could trigger alternate path routing via MPLS.

5.2 Fault Tolerance

Fault tolerance is another category of reactions that may be used by applications in response to CE feedback. In some cases, these efforts may contribute to an increase in traffic load in order to add protection and resiliency to a flow.

Redundant Transmissions: This approach is based on a source sending duplicate payloads that can be used to compensate for lost packets. Its positive value may emerge in cases where a path has several downstream congestion points that increase the probability that a packet will be dropped instead of marked as CE and forwarded downstream.

Application Layer Forward Error Correction (FEC): This approach also adds additional overhead to the flow in order to compensate for potential packet loss. And as the case of redundant transmissions, the value of this approach can be realized when there exists multiple downstream congestion points that increase the probability of dropping packets. However, the impact of the overhead is minimized by having one (or a few) additional packet(s) used to compensate for the loss of a set of packets.

Codec Swapping: This approach involves changing codecs to either reduce load or achieve an improvement in compensating for lost packets. Depending on the codec, the reduction of load may be a simple step

function, or it may involve a gradual and variable reduction in load based on the rate of congestion feedback received by the source.

Interweaving packets: To Be Done (based on research at UCL)

5.2.1 Issues

The use of redundant transmissions or FEC produces a detrimental impact of contributing to an increase in load and the measure of congestion that triggers CE feedback. In the case of FEC, additional delay is typically incurred through the generation of X amount of erasure packets for each set of original source packets. And while an initial increase in QoS may be observed for these flows, the overall rate of congestion can be expected to increase.

Swapping codecs based on the reception of CE feedback has the positive affect of reducing load at the risk of reducing perceived QoS by the user. As in the case of all options described above regarding fault tolerance, the ability to change to a different codec is depending on end-to-end peer support. In addition, there is no assurance that the different codec reduces load in relation to the amount of congestion experienced over time.

5.3 Alternative Reaction for Emergency Communications

As mentioned in [rfc6679], the default reaction on the reception of these ECN-CE marked packets MUST be to provide the congestion control algorithm with a congestion notification that triggers the algorithm to react as if packet loss had occurred. There MAY be an alternative reaction if it is considered safe for deployment. An example of the need for an alternative reaction would be the case of Emergency Telecommunications Service (ETS) [rfc3689, rfc4190], where an improvement in QoS or a higher probability of session establishment and forwarding of traffic is of high interest.

It is proposed that certain authorized ETS flows may be permitted to employ either a substantially less aggressive back-off algorithm than the default algorithm, or some level of exemption from reacting to ECN marked packets. This alternative reaction will benefit these flows as the marks would normally be considered as equivalent to lost packets, which would effectively increase the loss level, which in turn will generally result in the reduction of flow rate. This applies to all flows that utilize some form of the rate control that is inversely proportional to the loss rate, which includes TCP-like algorithms or equation-based approaches.

Simulations of the use of ECN exemption with TFRC and have found that it has limited effect on the normal flows with low numbers of exempt flows. A half-dumbbell network was used with a RED router queue configured using the

settings recommended by Sally Floyd. The candidate flows are 1Mbit/s each with a backhaul 100Mbit/s link. In the standard case where 1% of flows would be exempt the remaining flows achieve 99.99% of the bandwidth that they would achieve without the presence of the exempt flows. This is what would be expected from the simple calculation of the allocation, given that the exempt flows achieve their full rate (1Mbit/s); With 100 normal plus 1 exempt flow, assuming that the except flow uses 1Mbit/s, the remaining capacity is 99Mbit/s which is divided between the 100 normal flows. Whilst when 101 normal flows are run over the 100Mbit/s link they would have to share it evenly, so it work

s out thus: $((99/100)/(100/101))*100=99.99\%$. In the case of 5% exempt flows then the proportion is very slightly lower at $((95/100)/(100/105))*100=99.75\%$. Bot

h these calculations are borne out in the simulation runs.

The level of exemption employed can be altered in a number of ways. Two simple approaches would be to either set a threshold number of ECN marked packets tha

t could be considered as a loss, and another approach would be to set a percentage threshold of ECN marked packet that would be considered as a loss.

It should be noted that in the simulations the end-to-end delay of the packets within the flows was monitored and the relative delay of the exempt flows apparently rises somewhat when exemption is enacted. However what is actually occurring is that the 'normal' flows are reducing their throughput and are thu

s reducing their latency somewhat. There is normally some limited latency when using loss-based techniques such as TFRC because it fills the queues to ascertain the link capacity and maintains that level of delay throughout a session. However the level of latency is clearly limited by the queue sizes in the network and on media specific links these queue sizes are typically quite small, so the resulting latency is limited.

Furthermore in the case where media flows employing TFRC, or any other congestion control algorithm (e.g. delay-based), are sharing a bottleneck link with TCP flows then the queues will be filled by the TCP flows and the latency will be kept near or at a their maximum despite any other flows.

5.3.1 Issues

To Be Done

6. IANA Considerations

This document requires no actions from IANA.

7. Security Considerations

The reliance on accurate and un-modified RTCP information means that SRTP needs to be used, or any other mechanism that helps prevent modification of RTCP feedback packets.

8. Acknowledgements

TBD

9. References

9.1 Normative

- [rfc2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [rfc2205] Braden, B., et. al., "Resource ReSerVation Protocol (RSVP) Version 1 Functional Specification", RFC 2205, September 1997
- [rfc2209] Braden, R., L. Zhang, "Resource Reservation Protocol (RSVP) Version 1 Message Processing Rules", RFC2209 September 1997
- [rfc2474] Nichols, K., et. al., "Definition of the Differentiated Services Field in the IPv4 and IPv6 Headers", RFC 2474, December 1998
- [rfc3168] Ramakrishnan, K., et. al., "The Addition of Explicit Congestion Notification (ECN) to IP", RFC 3168, September, 2001
- [rfc3181] Herzog, S., "Signaled Preemption Priority Policy Element", RFC 3181, October 2001
- [rfc3448] Handley, M., et. al., "TCP Friendly Rate Control (TFRC): Protocol Specification", RFC 3448, January 2003
- [rfc3583] Chaskar, H., "Requirements of a Quality of Service (QoS) Solution for Mobile IP", RFC 3583, September 2003
- [rfc4867] Sjöberg, J., et. al., "RTP Payload Format and File Storage Format for the AMR and AMR-WB Audio Codecs", RFC 4867, April 2007
- [rfc5104] Wenger, S., et. al., "Codec Control Messages in the RTP Audio-Visual Profile with Feedback (AVPF)", RFC 5104, February 2008
- [rfc6679] Westerlund, M., et. al., "Explicit Congestion Notification (ECN) for RTP over UDP", RFC 6679, IETF, Aug 2012

9.2 Informative

- [draft-rtp-tfrc] Gharai, L., C. Perkins, "RTP with TCP Friendly Rate Control", work-in-progress, Sept 2011
- [draft-tcpm-accecn-reqs] M. Kuehlewind, R. Scheffenegger, "Problem Statement and Requirements for a More Accurate ECN Feedback", work-in-progress, Feb 2013
- [Googl] http://code.google.com/apis/talk/call_signaling.html
- [tr26.114] "IMS; Multimedia telephony; Media Handling and Interaction", 3GPP, version 10, April 2011
- [ts36.300] "E-UTRA and E-UTRAN Overall Description, Stage 2", 3GPP, Release 10, September, 2011
- [rfc4340] Kohler, E., et. al, Datagram Congestion Control Protocol (DCCP), RFC4340, March 2006
- [rfc4342] Floyd, S., et. al., "Profile for DCCP Congestion Control ID 3: TFRC", RFC 4342, March 2006
- [rfc4828] Floyd, S., E. Kohler, "TFRC: The Small Packet Variant", RFC 4828, April 2007
- [rfc3689] Carlberg, K., Atkinson, R., "General Requirements for Emergency Telecommunications Service (ETS)", RFC 3689, February 2004
- [rfc4190] Carlberg, K. et, al., "Framework for Supporting Emergency Telecommunications Service (ETS) in IP Telephony", RFC 4190, November 2005
- [rfc3714] Floyd, S., Kempf, J., "IAB Concerns Regarding Congestion Control for Voice Traffic in the Internet", RFC 3714, March 2004
- [bcp145] Eggert, L., Fairhurst, G., "Unicast UDP Usage Guidelines for Application Designers", RFC 5405, BCP 145, November 2008
- [ITU.G114.2003]
International Telecommunications Union, "One-way transmission time", ITU-T Recommendation G.707, May 2003.

Author's Addresses

Piers O'Hanlon

University of Oxford
Oxford Internet Institute
1 St Giles
Oxford OX1 3JS
United Kingdom

Email: piers.ohanlon@oii.ox.ac.uk

Ken Carlberg
G11
1600 Clarendon Blvd
Arlington VA
USA

Email: carlberg@g11.org.uk

Internet Engineering Task Force
Internet-Draft
Intended status: Experimental
Expires: April 6, 2015

Georgios Karagiannis
Huawei Technologies
Anurag Bhargava
Cisco Systems, Inc.
October 6, 2014

Generic Aggregation of Resource ReSerVation Protocol (RSVP)
for IPv4 And IPv6 Reservations over PCN domains
draft-ietf-tsvwg-rsvp-pcn-11

Abstract

This document specifies extensions to Generic Aggregated RSVP RFC 4860 for support of the PCN Controlled Load (CL) and Single Marking (SM) edge behaviors over a Diffserv cloud using Pre-Congestion Notification.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on April 6, 2015.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

Table of Contents

1. Introduction	4
1.1. Objective	4
1.2. Overview and Motivation	5
1.3. Terminology	7
1.4. Organization of This Document	11
2. Overview of RSVP extensions and Operations	11
2.1. Overview of RSVP Aggregation Procedures in PCN domains	11
2.2. PCN Marking and encoding and transport of pre-congestion Information	13
2.3. Traffic Classification Within The Aggregation Region	13
2.4. Deaggregator (PCN-egress-node) Determination	13
2.5. Mapping E2E Reservations Onto Aggregate Reservations	13
2.6. Size of Aggregate Reservations	14
2.7. E2E Path ADSPEC update	14
2.8. Intra-domain Routes	14
2.9. Inter-domain Routes	15
2.10. Reservations for Multicast Sessions	15
2.11. Multi-level Aggregation	15
2.12. Reliability Issues	15
3. Elements of Procedure	15
3.1. Receipt of E2E Path Message by PCN-ingress-node (aggregating router)	15
3.2. Handling Of E2E Path Message by Interior Routers	16
3.3. Receipt of E2E Path Message by PCN-egress-node (deaggregating router)	16
3.4. Initiation of new Aggregate Path Message By PCN-ingress-node (Aggregating Router)	16
3.5. Handling Of new Aggregate Path Message by Interior Routers	16
3.6. Handling Of Aggregate Path Message by Deaggregating Router	16
3.7. Handling of E2E Resv Message by Deaggregating Router	17
3.8. Handling Of E2E Resv Message by Interior Routers	17

3.9. Initiation of New Aggregate Resv Message By Deaggregating Router	17
3.10. Handling of Aggregate Resv Message by Interior Routers	18
3.11. Handling of E2E Resv Message by Aggregating Router	18
3.12. Handling of Aggregated Resv Message by Aggregating Router . .	18
3.13. Removal of E2E Reservation	19
3.14. Removal of Aggregate Reservation	19
3.15. Handling of Data On Reserved E2E Flow by Aggregating Router .	19
3.16. Procedures for Multicast Sessions	19
3.17. Misconfiguration of PCN node	19
3.18. PCN based Flow Termination	19
4. Protocol Elements	20
4.1 PCN object	20
5. Security Considerations	23
6. IANA Considerations	24
7. Acknowledgments	24
8. Normative References	24
9. Informative References	25
10. Appendix A: Example Signaling Flow	26
11. Authors' Address	29

1. Introduction

1.1 Objective

Pre-Congestion Notification (PCN) can support the quality of service (QoS) of inelastic flows within a Diffserv domain in a simple, scalable, and robust fashion. Two mechanisms are used: admission control and flow termination. Admission control is used to decide whether to admit or block a new flow request, while flow termination is used in abnormal circumstances to decide whether to terminate some of the existing flows. To support these two features, the overall rate of PCN-traffic is metered on every link in the domain, and PCN-packets are appropriately marked when certain configured rates are exceeded. These configured rates are below the rate of the link, thus providing notification to boundary nodes about overloads before any congestion occurs (hence "pre-congestion" notification). The PCN-egress-nodes measure the rates of differently marked PCN traffic in periodic intervals and report these rates to the Decision Points for admission control and flow termination; the Decision Points use these rates to make decisions. The Decision Points may be collocated with the PCN-ingress-nodes, or their function may be implemented in a another node. For more details see [RFC5559], [RFC6661], and [RFC6662].

The main objective of this document is to specify the signaling protocol that can be used within a Pre-Congestion Notification (PCN) domain to carry reports from a PCN-ingress-node to a PCN Decision point, considering that the PCN Decision Point and PCN-egress-node are collocated.

If the PCN Decision Point is not collocated with the PCN-egress-node then additional signaling procedures are required that are out of the scope of this document. Moreover, as mentioned above this architecture conforms with PBAC (Policy-Based Admission Control), when the Decision Point is located in a another node then the PCN-ingress-node [RFC2753].

Several signaling protocols can be used to carry information between PCN-boundary-nodes (PCN-ingress-node and PCN-egress-node). However, since (1) both PCN-egress-node and PCN-ingress-nodes are located on the data path and (2) the admission control procedure needs to be done at PCN-egress-node, a signaling protocol that follows the same path as the data path, like RSVP (Resource Reservation Protocol), is more suited for this purpose. In particular, this document specifies extensions to Generic Aggregated RSVP [RFC4860] for support of the PCN Controlled Load (CL) and Single Marking (SM) edge behaviors over a Diffserv cloud using Pre-Congestion Notification.

This draft is intended to be published as Experimental in order to:

- o) validate industry interest by allowing implementation and deployment
- o) gather operational experience, in particular around dynamic interactions of RSVP signaling and PCN notification and

corresponding levels of performance.

Support for the techniques specified in this document involves RSVP functionality in boundary nodes of a PCN domain whose interior nodes forward RSVP traffic without performing RSVP functionality.

1.2 Overview and Motivation

Two main Quality of Service (QoS) architectures have been specified by the IETF. These are the Integrated Services (Intserv) [RFC1633] architecture and the Differentiated Services (DiffServ) architecture ([RFC2475]).

Intserv provides methods for the delivery of end-to-end Quality of Service (QoS) to applications over heterogeneous networks. One of the QoS signaling protocols used by the Intserv architecture is the Resource reSerVation Protocol (RSVP) [RFC2205], which can be used by applications to request per-flow resources from the network. These RSVP requests can be admitted or rejected by the network. Applications can express their quantifiable resource requirements using Intserv parameters as defined in [RFC2211] and [RFC2212]. The Controlled Load (CL) service [RFC2211] is a quality of service (QoS) closely approximating the QoS that the same flow would receive from a lightly loaded network element. The CL service is useful for inelastic flows such as those used for real-time media.

The DiffServ architecture can support the differentiated treatment of packets in very large scale environments. While Intserv and RSVP classify packets per-flow, Diffserv networks classify packets into one of a small number of aggregated flows or "classes", based on the Diffserv codepoint (DSCP) in the packet IP header. At each Diffserv router, packets are subjected to a "per-hop behavior" (PHB), which is invoked by the DSCP. The primary benefit of Diffserv is its scalability, since the need for per-flow state and per-flow processing, is eliminated.

However, DiffServ does not include any mechanism for communication between applications and the network. Several solutions have been specified to solve this issue. One of these solutions is Intserv over Diffserv [RFC2998] including resource-based admission control (RBAC), PBAC, assistance in traffic identification/classification, and traffic conditioning. Intserv over Diffserv can operate over a statically provisioned or a RSVP aware Diffserv region. When it is RSVP aware, several mechanisms may be used to support dynamic provisioning and topology-aware admission control, including aggregate RSVP reservations, per-flow RSVP, or a bandwidth broker. [RFC3175] specifies aggregation of Resource ReSerVation Protocol (RSVP) end-to-end reservations over aggregate RSVP reservations. In [RFC3175] the RSVP generic aggregated reservation is characterized by a RSVP SESSION object using the 3-tuple <source IP address, destination IP address, Diffserv Code Point>.

Several scenarios require the use of multiple generic aggregate reservations that are established for a given PHB from a given source

IP address to a given destination IP address, see [SIG-NESTED], [RFC4860]. For example, multiple generic aggregate reservations can be applied in the situation that multiple E2E reservations using different preemption priorities need to be aggregated through a PCN-domain using the same PHB. By using multiple aggregate reservations for the same PHB, it allows enforcement of the different preemption priorities within the aggregation region. This allows more efficient management of the Diffserv resources, and in periods of resource shortage, this allows sustainment of a larger number of E2E reservations with higher preemption priorities. In particular, [SIG-NESTED] discusses in detail how end-to-end RSVP reservations can be established in a nested VPN environment through RSVP aggregation.

[RFC4860] provides generic aggregate reservations by extending [RFC3175] to support multiple aggregate reservations for the same source IP address, destination IP address, and PHB (or set of PHBs). In particular, multiple such generic aggregate reservations can be established for a given PHB from a given source IP address to a given destination IP address. This is achieved by adding the concept of a Virtual Destination Port and of an Extended Virtual Destination Port in the RSVP SESSION object. In addition to this, the RSVP SESSION object for generic aggregate reservations uses the PHB Identification Code (PHB-ID) defined in [RFC3140], instead of using the Diffserv Code Point (DSCP) used in [RFC3175]. The PHB-ID is used to identify the PHB, or set of PHBs, from which the Diffserv resources are to be reserved.

The RSVP like signaling protocol required to carry (1) requests from a PCN-egress-node to a PCN-ingress-node and (2) reports from a PCN-ingress-node to a PCN-egress-node needs to follow the PCN signaling requirements defined in [RFC6663]. In addition to that the signaling protocol functionality supported by the PCN-ingress-nodes and PCN-egress-nodes needs to maintain logical aggregate constructs (i.e. ingress-egress-aggregate state) and be able to map E2E reservations to these aggregate constructs. Moreover, no actual reservation state is needed to be maintained inside the PCN domain, i.e., the PCN-interior-nodes are not maintaining any reservation state.

This can be accomplished by two possible approaches:

Approach (1):

- o) adapting the RFC 4860 aggregation procedures to fit the PCN requirements with as little change as possible over the RFC 4860 functionality
- o) hence performing aggregate RSVP signaling (even if it is to be ignored by PCN interior nodes)
- o) using this aggregate RSVP signaling procedures to carry PCN information between the PCN-boundary-nodes (PCN-ingress-node and PCN-egress-node).

Approach (2):

- o) adapting the RFC 4860 aggregation procedures to fit the PCN requirements with more significant changes over RFC4860 (i.e. the aspect of the procedures that have to do with maintaining aggregate states and to do with mapping the E2E reservations to aggregate constructs are kept, but the procedures that have to do with the aggregate RSVP signaling and aggregate reservation establishment/maintenance are dropped).
- o) hence not performing aggregate RSVP signaling
- o) piggy-backing of the PCN information inside the E2E RSVP signaling.

Both approaches are probably viable, however, since the RFC 4860 operations have been thoroughly studied and implemented, it can be considered that the RFC 4860 solution can better deal with the more challenging situations (rerouting in the PCN domain, failure of an PCN-ingress-node, failure of an PCN-egress-node, rerouting towards a different edge, etc.). This is the reason for choosing Approach (1) for the specification of the signaling protocol used to carry PCN information between the PCN-boundary-nodes (PCN-ingress-node and PCN-egress-node).

In particular, this document specifies extensions to Generic Aggregated RSVP [RFC4860] for support of the PCN Controlled Load (CL) and Single Marking (SM) edge behaviors over a Diffserv cloud using Pre-Congestion Notification.

This document follows the PCN signaling requirements defined in [RFC6663] and specifies extensions to Generic Aggregated RSVP [RFC4860] for support of PCN edge behaviors as specified in [RFC6661] and [RFC6662]. Moreover, this document specifies how RSVP aggregation can be used to setup and maintain: (1) Ingress Egress Aggregate (IEA) states at Ingress and Egress nodes and (2) generic aggregation of RSVP end-to-end RSVP reservations over PCN (Congestion and Pre-Congestion Notification) domains.

To comply with this specification, PCN-nodes MUST be able to support the functionality specified in [RFC5670], [RFC5559], [RFC6660], [RFC6661], [RFC6662]. Furthermore, the PCN-boundary-nodes MUST support the RSVP generic aggregated reservation procedures specified in [RFC4860] which are augmented with procedures specified in this document.

1.3. Terminology

This document uses terms defined in [RFC4860], [RFC3175], [RFC5559], [RFC5670], [RFC6661], [RFC6662].

For readability, a number of definitions from [RFC3175] as well as definitions for terms used in [RFC5559], [RFC6661], and [RFC6662] are provided here, where some of them are augmented with new meanings:

Aggregator	This is the process in (or associated with) the router at the ingress edge of the aggregation region (with respect to the end-to-end RSVP reservation) and behaving in accordance with [RFC4860]. In this document, it is also the PCN-ingress-node. It is important to notice that in the context of this document the Aggregator must be able to determine the Deaggregator using the procedures specified in Section 4 of [RFC4860] and in Section 1.4.2 of [RFC3175].
Congestion level estimate (CLE):	<p>The ratio of PCN-marked to total PCN-traffic (measured in octets) received for a given ingress-egress-aggregate during a given measurement period. The CLE is used to derive the PCN-admission-state and is also used by the report suppression procedure if report suppression is activated.</p>
Deaggregator	This is the process in (or associated with) the router at the egress edge of the aggregation region (with respect to the end-to-end RSVP reservation) and behaving in accordance with [RFC4860]. In this document, it is also the PCN-egress-node and Decision Point.
E2E	end to end
E2E Reservation	<p>This is an RSVP reservation such that:</p> <ul style="list-style-type: none">(i) corresponding RSVP Path messages are initiated upstream of the Aggregator and terminated downstream of the Deaggregator, and(ii) corresponding RSVP Resv messages are initiated downstream of the Deaggregator and terminated upstream of the Aggregator, and(iii) this RSVP reservation is aggregated over an Ingress Egress Aggregate (IEA) between the Aggregator and Deaggregator. <p>An E2E RSVP reservation may be a per-flow reservation, which in this document is only maintained at the PCN-ingress-node and PCN-egress-node. Alternatively, the E2E reservation may itself be an aggregate reservation of various types (e.g., Aggregate IP reservation, Aggregate IPsec reservation, see [RFC4860]). As per regular RSVP operations, E2E RSVP reservations are unidirectional.</p>
E2E microflow	a microflow where its associated packets are being forwarded on an E2E path.

Extended vDstPort (Extended Virtual Destination Port)

An identifier used in the SESSION that remains constant over the life of the generic aggregate reservation. The length of this identifier is 32-bits when IPv4 addresses are used and 128 bits when IPv6 addresses are used.

A sender(or Aggregator) that wishes to narrow the scope of a SESSION to the sender-receiver pair (or Aggregator-Deaggregator pair) should place its IPv4 or IPv6 address here as a network unique identifier. A sender (or Aggregator) that wishes to use a common session with other senders (or Aggregators) in order to use a shared reservation across senders (or Aggregators) must set this field to all zeros. In this document, the Extended vDstPort should contain the IPv4 or IPv6 address of the Aggregator.

ETM-rate

The rate of excess-traffic-marked PCN-traffic received at a PCN-egress-node for a given ingress-egress-aggregate in octets per second.

Ingress-egress-aggregate (IEA):

The collection of PCN-packets from all PCN-flows that travel in one direction between a specific pair of PCN-boundary-nodes. In this document one RSVP generic aggregated reservation is mapped to only one ingress-egress-aggregate, while one ingress-egress-aggregate is mapped to either one or to more than one RSVP generic aggregated reservations. PCN-flows and their PCN-traffic that are mapped into a specific RSVP generic aggregated reservation can also easily be mapped into their corresponding ingress-egress-aggregate.

Microflow:
(from [RFC2474])

a single instance of an application-to-application flow of packets which is identified by source address, destination address, protocol id, and source port, destination port (where applicable).

PCN-domain:

a PCN-capable domain; a contiguous set of PCN-enabled nodes that perform Diffserv scheduling [RFC2474]; the complete set of PCN-nodes that in principle can, through PCN-marking packets, influence decisions about flow admission and termination within the domain; includes the PCN-egress-nodes, which measure these PCN-marks, and the PCN-ingress-nodes.

PCN-boundary-node: a PCN-node that connects one PCN-domain to a node either in another PCN-domain or in a non-PCN-domain.

- PCN-interior-node: a node in a PCN-domain that is not a PCN-boundary-node.
- PCN-node: a PCN-boundary-node or a PCN-interior-node.
- PCN-egress-node: a PCN-boundary-node in its role in handling traffic as it leaves a PCN-domain. In this document the PCN-egress-node operates also as a Decision Point and Deaggregator.
- PCN-ingress-node: a PCN-boundary-node in its role in handling traffic as it enters a PCN-domain. In this document the PCN-ingress-node operates also as a Aggregator.
- PCN-traffic,
PCN-packets,
PCN-BA: a PCN-domain carries traffic of different Diffserv behavior aggregates (BAs) [RFC2474]. The PCN-BA uses the PCN mechanisms to carry PCN-traffic, and the corresponding packets are PCN-packets. The same network will carry traffic of other Diffserv BAs. The PCN-BA is distinguished by a combination of the Diffserv codepoint (DSCP) and ECN fields.
- PCN-flow: the unit of PCN-traffic that the PCN-boundary-node admits (or terminates); the unit could be a single E2E microflow (as defined in [RFC2474]) or some identifiable collection of microflows.
- PCN-admission-state: The state ("admit" or "block") derived by the Decision Point for a given ingress-egress-aggregate based on statistics about PCN-packet marking. The Decision Point decides to admit or block new flows offered to the aggregate based on the current value of the PCN-admission-state.
- PCN-sent-rate The rate of PCN-traffic received at a PCN-ingress-node and destined for a given ingress-egress-aggregate in octets per second.
- PHB-ID (Per Hop Behavior Identification Code)
A 16-bit field containing the Per Hop Behavior Identification Code of the PHB, or of the set of PHBs, from which Diffserv resources are to be reserved. This field must be encoded as specified in Section 2 of [RFC3140].
- RSVP generic aggregated reservation: an RSVP reservation that is identified by using the RSVP SESSION object for generic RSVP aggregated reservation. This RSVP

SESSION object is based on the RSVP SESSION object specified in [RFC4860] augmented with the following information:

- o) the IPv4 DestAddress, IPv6 DestAddress should be set to the IPv4 or IPv6 destination addresses, respectively, of the Deaggregator (PCN-egress-node)
- o) PHB-ID (Per Hop Behavior Identification Code) should be set equal to PCN-compatible Diffserv codepoint(s).
- o) Extended vDstPort should be set to the IPv4 or IPv6 destination addresses, of the Aggregator (PCN-ingress-node)

VDstPort (Virtual Destination Port)

A 16-bit identifier used in the SESSION that remains constant over the life of the generic aggregate reservation.

1.4. Organization of This Document

This document is organized as follows. Section 2 gives an overview of RSVP extensions and operations. The elements of the used procedures are specified in Section 3. Section 4 describes the protocol elements. The security considerations are given in section 5 and the IANA considerations are provided in Section 6.

2. Overview of RSVP extensions and Operations

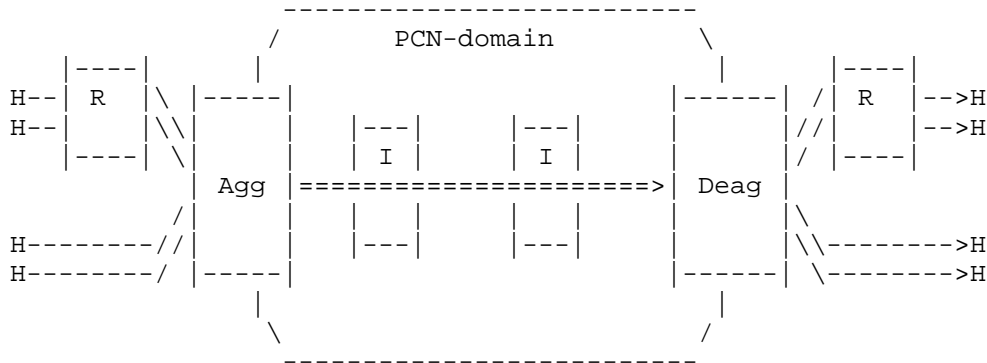
2.1 Overview of RSVP Aggregation Procedures in PCN domains

The PCN-boundary-nodes, see Figure 1, can support RSVP SESSIONS for generic aggregated reservations [RFC4860], which are depending on ingress-egress-aggregates. In particular, one RSVP generic aggregated reservation matches to only one ingress-egress-aggregate.

However, one ingress-egress-aggregate matches to either one, or more than one, RSVP generic aggregated reservations. In addition, to comply with this specification, the PCN-boundary nodes need to distinguish and process (1) RSVP SESSIONS for generic aggregated sessions and their messages according to [RFC4860], (2) E2E RSVP sessions and messages according to [RFC2205].

This document locates all RSVP processing for a PCN domain at PCN-Boundary nodes. PCN-interior-nodes do not perform any RSVP functionality or maintain RSVP-related state information. Rather, PCN-interior nodes forward all RSVP messages (for both generic aggregated reservations[RFC4860] and end to end reservations [RFC2205]) as if they were ordinary network traffic.

Moreover, each Aggregator and Deaggregator (i.e., PCN-boundary-nodes) need to support policies to initiate and maintain for each pair of PCN-boundary-nodes of the same PCN-domain one ingress-egress-aggregate.



H = Host requesting end-to-end RSVP reservations
 R = RSVP router
 Agg = Aggregator (PCN-ingress-node)
 Deag = Deaggregator (PCN-egress-node)
 I = Interior Router (PCN-interior-node)
 --> = E2E RSVP reservation
 ==> = Aggregate RSVP reservation

Figure 1 : Aggregation of E2E Reservations
 over Generic Aggregate RSVP Reservations
 in PCN domains, based on [RFC4860]

Both the Aggregator and Deaggregator can maintain one or more RSVP generic aggregated Reservations, but the Deaggregator is the entity that initiates these RSVP generic aggregated reservations. Note that one RSVP generic aggregated reservation matches to only one ingress-egress-aggregate, while one ingress-egress-aggregate matches to either one or to more than one RSVP generic aggregated reservations. This can be accomplished by using for the different RSVP generic aggregated reservations the same combinations of ingress and egress identifiers, but with a different PHB-ID value (see [RFC4860]). The procedures for aggregation of E2E reservations over generic aggregate RSVP reservations are the same as the procedures specified in Section 4 of [RFC4860], augmented with the ones specified in Section 2.5.

One significant difference between this document and [RFC4860] is the fact that in this document the admission control of E2E RSVP reservations over the PCN core is performed according to the PCN procedures, while in [RFC4860] this is achieved via first admitting aggregate RSVP reservations over the aggregation region and then admitting the E2E reservations over the aggregate RSVP reservations. Therefore, in this document, the RSVP generic aggregate RSVP reservations are not subject to admission control in the PCN-core, and the E2E RSVP reservations are not subject to admission control

over the aggregate reservations. In turn, this means that several procedures of [RFC4860] are significantly simplified in this document:

- o) unlike [RFC4860], the generic aggregate RSVP reservations need not be admitted in the PCN core.
- o) unlike [RFC4860], the RSVP aggregated traffic does not need to be tunneled between Aggregator and Deaggregator, see Section 2.3.
- o) unlike [RFC4860], the Deaggregator need not perform admission control of E2E reservations over the aggregate RSVP reservations.
- o) unlike [RFC4860], there is no need for dynamic adjustment of the RSVP generic aggregated reservation size, see Section 2.6.

2.2 PCN Marking and encoding and transport of pre-congestion information

The method of PCN marking within the PCN domain is specified in [RFC5670]. In addition, the method of encoding and transport of pre-congestion information is specified in [RFC6660]. The PHB-ID (Per Hop Behavior Identification Code) used SHOULD be set equal to PCN-compatible Diffserv codepoint(s).

2.3. Traffic Classification Within The Aggregation Region

The PCN-ingress marks a PCN-BA using PCN-marking (i.e., combination of the DSCP and ECN fields), which interior nodes use to classify PCN-traffic. The PCN-traffic (e.g., E2E microflows) belonging to a RSVP generic aggregated reservation can be classified only at the PCN-boundary-nodes (i.e., Aggregator and Deaggregator) by using the RSVP SESSION object for RSVP generic aggregated reservations, see Section 2.1 of [RFC4860]. Note that the DSCP value included in the SESSION object, SHOULD be set equal to a PCN-compatible Diffserv codepoint. Since no admission control procedures over the RSVP generic aggregated reservations in the PCN-core are required, unlike [RFC4860], the RSVP aggregated traffic need not to be tunneled between Aggregator and Deaggregator. In this document one RSVP generic aggregated reservation is mapped to only one ingress-egress-aggregate, while one ingress-egress-aggregate is mapped to either one or to more than one RSVP generic aggregated reservations. PCN-flows and their PCN-traffic that are mapped into a specific RSVP generic aggregated reservation can also easily be classified into their corresponding ingress-egress-aggregate. The method of traffic conditioning of PCN-traffic and non-PCN traffic and PHB configuration is described in [RFC6661] and [RFC6662].

2.4. Deaggregator Determination

The present document assumes the same dynamic Deaggregator determination method as used in [RFC4860].

2.5. Mapping E2E Reservations Onto Aggregate Reservations

To comply with this specification for the mapping of E2E reservations

onto aggregate reservations, the same methods MUST be used as the ones described in Section 4 of [RFC4860], augmented by the following rules:

- o) An Aggregator (also PCN-ingress-node in this document) or Deaggregator (also PCN-egress-node and Decision Point in this document) MUST use one or more policies to determine whether a RSVP generic aggregated reservation can be mapped into an ingress-Egress-aggregate. This can be accomplished by using for the different RSVP generic aggregated reservations the same combinations of ingress and egress identifiers, but with a different PHB-ID value (see [RFC4860]) corresponding to the PCN specifications. In particular, the RSVP SESSION object specified in [RFC4860] augmented with the following information:
 - o) the IPv4 DestAddress, IPv6 DestAddress MUST be set to the IPv4 or IPv6 destination addresses, respectively, of the Deaggregator (PCN-egress-node), see [RFC4860]. Note that the PCN-domain is considered as being only one RSVP hop (for Generic aggregated RSVP or E2E RSVP). This means that the next RSVP hop for the Aggregator in the downstream direction is the Deaggregator and the next RSVP hop for the Deaggregator in the upstream direction is the Aggregator.
 - o) PHB-ID (Per Hop Behavior Identification Code) SHOULD be set equal to PCN-compatible Diffserv codepoint(s).
 - o) Extended vDstPort SHOULD be set to the IPv4 or IPv6 destination addresses, of the Aggregator (PCN-ingress-node), see [RFC4860].

2.6. Size of Aggregate Reservations

Since:(i) no admission control of E2 reservations over the RSVP aggregated reservations is required, and (ii) no admission control of the RSVP aggregated reservation over the PCN core is required, the size of the generic aggregate reservation is irrelevant and can be set to any arbitrary value by the Deaggregator. The Deaggregator SHOULD set the value of a generic aggregate reservation to a null bandwidth. We also observe that there is no need for dynamic adjustment of the RSVP aggregated reservation size.

2.7. E2E Path ADSPEC update

To comply with this specification, for the update of the E2E Path ADSPEC, the same methods can be used as the ones described in [RFC4860].

2.8. Intra-domain Routes

The PCN-interior-nodes are neither maintaining E2E RSVP nor RSVP generic aggregation states and reservations. Therefore, intra-domain route changes will not affect intra-domain reservations since such reservations are not maintained by the PCN-interior-nodes.

Furthermore, it is considered that by configuration, the PCN-interior-nodes are not able to distinguish neither RSVP generic aggregated sessions and their associated messages [RFC4860], nor E2E RSVP sessions and their associated messages [RFC2205].

2.9. Inter-domain Routes

The PCN-charter scope precludes inter-domain considerations. However, for solving inter-domain routes changes associated with the operation of the RSVP messages, the same methods SHOULD be used as the ones described in [RFC4860] and in Section 1.4.7 of [RFC3175].

2.10. Reservations for Multicast Sessions

PCN does not consider reservations for multicast sessions.

2.11. Multi-level Aggregation

PCN does not consider multi-level aggregations within the PCN domain. Therefore, the PCN-interior-nodes are not supporting multi-level aggregation procedures. However, the Aggregator and Deaggregator SHOULD support the multi-level aggregation procedures specified in [RFC4860] and in Section 1.4.9 of [RFC3175].

2.12. Reliability Issues

To comply with this specification, for solving possible reliability issues, the same methods MUST be used as the ones described in Section 4 of [RFC4860].

3. Elements of Procedure

This section describes the procedures used to implement the aggregated RSVP procedure over PCN. It is considered that the procedures for aggregation of E2E reservations over generic aggregate RSVP reservations are same as the procedures specified in Section 4 of [RFC4860] except where a departure from these procedures is explicitly described in the present section. Please refer to [RFC4860] for all the below error cases:

- o) Incomplete message
- o) Unexpected objects

3.1. Receipt of E2E Path Message by Aggregating router

When the E2E Path message arrives at the exterior interface of the Aggregator, (also PCN-ingress-node in this document), then standard RSVP generic aggregation [RFC4860] procedures are used.

3.2. Handling Of E2E Path Message by Interior Routers

The E2E Path messages traverse zero or more PCN-interior-nodes. The PCN-interior-nodes receive the E2E Path message on an interior interface and forward it on another interior interface. It is considered that, by configuration, the PCN-interior-nodes ignore the E2E RSVP signaling messages [RFC2205]. Therefore, the E2E Path messages are simply forwarded as normal IP datagrams.

3.3. Receipt of E2E Path Message by Deaggregating router

When receiving the E2E Path message the Deaggregator (also PCN-egress-node and Decision Point in this document) performs the regular [RFC4860] procedures, augmented with the following rules:

- o) The Deaggregator MUST NOT perform the RSVP-TTL vs IP TTL-check and MUST NOT update the ADspec Break bit. This is because the whole PCN-domain is effectively handled by E2E RSVP as a virtual link on which integrated service is indeed supported (and admission control performed) so that the Break bit MUST NOT be set, see also [draft-lefaucheur-rsvp-ecn-01].

The Deaggregator forwards the E2E Path message towards the receiver.

3.4. Initiation of new Aggregate Path Message by Aggregating Router

To comply with this specification, for the initiation of the new RSVP generic aggregated Path message by the Aggregator (also PCN-ingress-node in this document), the same methods MUST be used as the ones described in [RFC4860].

3.5. Handling Of Aggregate Path Message By Interior Routers

The Aggregate Path messages traverse zero or more PCN-interior-nodes. The PCN-interior-nodes receive the Aggregated Path message on an interior interface and forward it on another interior interface. It is considered that, by configuration, the PCN-interior-nodes ignore the Aggregated Path signaling messages. Therefore, the Aggregated Path messages are simply forwarded as normal IP datagrams.

3.6. Handling Of Aggregate Path Message By Deaggregating Router

When receiving the Aggregated Path message, the Deaggregator (also PCN-egress-node and Decision Point in this document) performs the regular [RFC4860] procedures, augmented with the following rules:

- o) When the received Aggregated Path message by the Deaggregator contains the RSVP-AGGREGATE-IPv4-PCN-response or RSVP-AGGREGATE-IPv6-PCN-response PCN objects, which carry the PCN-sent-rate, then the procedures specified in Section 3.18 of this document MUST be followed.

3.7. Handling of E2E Resv Message by Deaggregating Router

When the E2E Resv message arrives at the exterior interface of the Deaggregator, (also PCN-egress-node and Decision Point in this document) then standard RSVP aggregation [RFC4860] procedures are used, augmented with the following rules:

- o) The E2E RSVP session associated with an E2E Resv message that arrives at the external interface of the Deaggregator is mapped/matched with an RSVP generic aggregate and with a PCN ingress-egress-aggregate.
- o) Depending on the type of the PCN edge behavior supported by the Deaggregator, the PCN admission control procedures specified in Section 3.3.1 of [RFC6661] or [RFC6662] MUST be followed. Since no admission control procedures over the RSVP aggregated reservations in the PCN-core are required, unlike [RFC4860], the Deaggregator does not perform any admission control of the E2E Reservation over the mapped generic aggregate RSVP reservation. If the PCN based admission control procedure is successful then the Deaggregator MUST allow the new flow to be admitted onto the associated RSVP generic aggregation reservation and onto the PCN ingress-egress-aggregate, see [RFC6661] and [RFC6662]. If the PCN based admission control procedure is not successful, then the E2E Resv MUST NOT be admitted onto the associated RSVP generic aggregate reservation and onto the PCN ingress-egress-aggregation. The E2E Resv message is further processed according to [RFC4860].

The way of how the PCN-admission-state is maintained is specified in [RFC6661] and [RFC6662].

3.8. Handling Of E2E Resv Message By Interior Routers

The E2E Resv messages traversing the PCN core are IP addressed to the Aggregating router and are not marked with Router Alert, therefore the E2E Resv messages are simply forwarded as normal IP datagrams.

3.9. Initiation of New Aggregate Resv Message By Deaggregating Router

To comply with this specification, for the initiation of the new RSVP generic aggregated Resv message by the Deaggregator (also PCN-egress-node and Decision Point in this document), the same methods MUST be used as the ones described in Section 4 of [RFC4860] augmented with the following rules:

- o) The size of the generic aggregate reservation is irrelevant, see Section 2.6, and can be set to any arbitrary value by the PCN-egress node. The Deaggregator SHOULD set the value of a RSVP generic aggregate reservation to a null bandwidth. We also observe that there is no need for dynamic adjustment of the RSVP generic aggregated reservation size.

- o) When [RFC6661] is used and the ETM-rate measured by the Deaggregator contains a non-zero value for some ingress-egress-aggregate, see [RFC6661] and [RFC6662], the Deaggregator MUST request the PCN-ingress-node to provide an estimate of the rate (PCN-sent-rate) at which the Aggregator (also PCN-ingress-node in this document) is receiving PCN-traffic that is destined for the given ingress-egress-aggregate.
- o) When [RFC6662] is used and the PCN-admission-state computed by the Deaggregator, on the basis of the CLE is "block" for the given ingress-egress-aggregate, the Deaggregator MUST request the PCN-ingress-node to provide an estimate of the rate (PCN-sent-rate) at which the Aggregator is receiving PCN-traffic that is destined for the given ingress-egress-aggregate.
- o) In the above two cases and when the PCN-sent-rate needs to be requested from the Aggregator, the Deaggregator MUST generate and send an (refresh) Aggregated Resv message to the Aggregator that MUST carry one of the following PCN objects, see Section 4.1, depending on whether IPv4 or IPv6 is supported:
 - o) RSVP-AGGREGATE-IPv4-PCN-request
 - o) RSVP-AGGREGATE-IPv6-PCN-request.

3.10. Handling of Aggregate Resv Message by Interior Routers

The Aggregated Resv messages traversing the PCN core are IP addressed to the Aggregating router and are not marked with Router Alert, therefore the Aggregated Resv messages are simply forwarded as normal IP datagrams.

3.11. Handling of E2E Resv Message by Aggregating Router

When the E2E Resv message arrives at the interior interface of the Aggregator (also PCN-ingress-node in this document), then standard RSVP aggregation [RFC4860] procedures are used.

3.12. Handling of Aggregated Resv Message by Aggregating Router

When the Aggregated Resv message arrives at the interior interface of the Aggregator, (also PCN-ingress-node in this document), then standard RSVP aggregation [RFC4860] procedures are used, augmented with the following rules:

- o) the Aggregator SHOULD use the information carried by the PCN objects, see Section 4, and follow the steps specified in [RFC6661], [RFC6662]. If the "R" flag carried by the RSVP-AGGREGATE-IPv4-PCN-request or RSVP-AGGREGATE-IPv6-PCN-request PCN objects is set to ON, see Section 4.1, then the Aggregator follows the steps described in Section 3.4 of [RFC6661] and [RFC6662] on calculating the PCN-sent-rate. In particular, the Aggregator MUST provide the estimated current rate of PCN-traffic received at that node and destined for a given ingress-egress-aggregate in octets per second (the PCN-sent-rate). The way this rate estimate is derived is a matter of implementation, see [RFC6661] or [RFC6662].

- o) the Aggregator initiates an Aggregated Path message. In particular, when the Aggregator receives an Aggregated Resv message which carries one of the following PCN objects: RSVP-AGGREGATE-IPv4-PCN-request or RSVP-AGGREGATE-IPv6-PCN-request, with the flag "R" set to ON, see Section 4.1, the Aggregator initiates an Aggregated Path message, and includes the calculated PCN-sent-rate into the RSVP-AGGREGATE-IPv4-PCN-response or RSVP-AGGREGATE-IPv6-PCN-response PCN objects, see Section 4.1, which that MUST be carried by the Aggregated Path message. This Aggregated Path message is sent towards the Deaggregator (also PCN-egress-node and Decision Point in this document) that requested the calculation of the PCN-sent-rate.

3.13. Removal of E2E Reservation

To comply with this specification, for the removal of E2E reservations, the same methods MUST be used as the ones described in Section 4 of [RFC4860] and [RFC4495].

3.14. Removal of Aggregate Reservation

To comply with this specification, for the removal of RSVP generic aggregated reservations, the same methods MUST be used as the ones described in Section 4 of [RFC4860] and Section 2.10 of [RFC3175]. In particular, should an aggregate reservation go away (presumably due to a configuration change, route change, or policy event), the E2E reservations it supports are no longer active. They MUST be treated accordingly.

3.15. Handling of Data On Reserved E2E Flow by Aggregating Router

The handling of data on the reserved E2E flow by Aggregator (also PCN-ingress-node in this document) uses the procedures described in [RFC4860] augmented with:

- o) Regarding, PCN marking and traffic classification the procedures defined in Section 2.2 and 2.3 of this document are used.

3.16. Procedures for Multicast Sessions

In this document no multicast sessions are considered.

3.17. Misconfiguration of PCN-node

In an event where a PCN-node is misconfigured within a PCN-domain, the desired behavior is same as described in Section 3.10.

3.18 PCN based Flow Termination

When the Deaggregator (also PCN-egress-node and Decision Point in this document) needs to terminate an amount of traffic associated with one ingress-egress-aggregate (see Section 3.3.2 of [RFC6661] and [RFC6662]), then several procedures of terminating E2E microflows can be deployed. The default procedure of terminating E2E microflows (i.e., PCN-flows) is as follows, see i.e., [RFC6661] and [RFC6662].

For the same ingress-egress-aggregate, select a number of E2E microflows to be terminated in order to decrease the total incoming amount of bandwidth associated with one ingress-egress-aggregate by the amount of traffic to be terminated, see above. In this situation the same mechanisms for terminating an E2E microflow can be followed as specified in [RFC2205]. However, based on a local policy, the Deaggregator could use other ways of selecting which microflows should be terminated. For example, for the same ingress-egress-aggregate, select a number of E2E microflows to be terminated or to reduce their reserved bandwidth in order to decrease the total incoming amount of bandwidth associated with one ingress-egress-aggregate by the amount of traffic to be terminated. In this situation the same mechanisms for terminating an E2E microflow or reducing bandwidth associated with an E2E microflow can be followed as specified in [RFC4495].

4. Protocol Elements

The protocol elements in this document are using the ones defined in Section 4 of [RFC4860] and Section 3 of [RFC3175] augmented with the following rules:

- o) the DSCP value included in the SESSION object, SHOULD be set equal to a PCN-compatible Diffserv codepoint.
- o) Extended vDstPort SHOULD be set to the IPv4 or IPv6 destination addresses, of the Aggregator (also PCN-ingress-node in this document), see [RFC4860].
- o) When the Deaggregator (also PCN-egress-node and Decision Point in this document) needs to request the PCN-sent-rate from the PCN-ingress-node, see Section 3.9 of this document, the Deaggregator MUST generate and send an (refresh) Aggregate Resv message to the Aggregator that MUST carry one of the following PCN objects, see Section 4.1, depending on whether IPv4 or IPv6 is supported:
 - o) RSVP-AGGREGATE-IPv4-PCN-request
 - o) RSVP-AGGREGATE-IPv6-PCN-request.
- o) When the Aggregator receives an Aggregate Resv message which carries one of the following PCN objects:
RSVP-AGGREGATE-IPv4-PCN-request or
RSVP-AGGREGATE-IPv6-PCN-request, with the flag "R" set to ON, see Section 4.1, then the Aggregator MUST generate and send to the Deaggregator an Aggregated Path message which carries one of the following PCN objects, see Section 4.1, depending on whether IPv4 or IPv6 is supported:
 - o) RSVP-AGGREGATE-IPv4-PCN-response,
 - o) RSVP-AGGREGATE-IPv6-PCN-response.

4.1 PCN objects

This section describes four types of PCN objects that can be carried by the (refresh) Aggregate Path or the (refresh) Aggregate Resv messages specified in [RFC4860].

These objects are:

- o RSVP-AGGREGATE-IPv4-PCN-request,
- o RSVP-AGGREGATE-IPv6-PCN-request,
- o RSVP-AGGREGATE-IPv4-PCN-response,
- o RSVP-AGGREGATE-IPv6-PCN-response.

- o) RSVP-AGGREGATE-IPv4-PCN-request: PCN request object, when IPv4 addresses are used:

Class = 248 (PCN)

C-Type = 1 (RSVP-AGGREGATE-IPv4-PCN-request)

+-----+-----+-----+-----+	
	IPv4 PCN-ingress-node Address (4 bytes)
+-----+-----+-----+-----+	
	IPv4 PCN-egress-node Address (4 bytes)
+-----+-----+-----+-----+	
	IPv4 Decision Point Address (4 bytes)
+-----+-----+-----+-----+	
R	Reserved
+-----+-----+-----+-----+	

- o) RSVP-AGGREGATE-IPv6-PCN-request: PCN object, when IPv6 addresses are used:

Class = 248 (PCN)

C-Type = 2 (RSVP-AGGREGATE-IPv6-PCN-request)

+-----+-----+-----+-----+	
	IPv6 PCN-ingress-node Address (16 bytes)
+	
+	
+-----+-----+-----+-----+	
	IPv6 PCN-egress-node Address (16 bytes)
+	
+	
+-----+-----+-----+-----+	
	Decision Point Address (16 bytes)
+	
+	
+-----+-----+-----+-----+	
R	Reserved
+-----+-----+-----+-----+	

- o) RSVP-AGGREGATE-IPv4-PCN-response: PCN object, IPv4 addresses are used:
 Class = 248 (PCN)
 C-Type = 3 (RSVP-AGGREGATE-IPv4-PCN-response)

```

+-----+-----+-----+-----+
| IPv4 PCN-ingress-node Address (4 bytes) |
+-----+-----+-----+-----+
| IPv4 PCN-egress-node Address (4 bytes) |
+-----+-----+-----+-----+
| IPv4 Decision Point Address (4 bytes) |
+-----+-----+-----+-----+
| PCN-sent-rate |
+-----+-----+-----+-----+

```

- o) RSVP-AGGREGATE-IPv6-PCN-response: PCN object, IPv6 addresses are used:
 Class = 248 (PCN)
 C-Type = 4 (RSVP-AGGREGATE-IPv6-PCN-response)

```

+-----+-----+-----+-----+
|                                     |
+                                     +
| IPv6 PCN-ingress-node Address (16 bytes) |
+                                     +
|                                     |
+-----+-----+-----+-----+
|                                     |
+                                     +
| IPv6 PCN-egress-node Address (16 bytes) |
+                                     +
|                                     |
+-----+-----+-----+-----+
|                                     |
+                                     +
| Decision Point Address (16 bytes) |
+                                     +
|                                     |
+-----+-----+-----+-----+
| PCN-sent-rate |
+-----+-----+-----+-----+

```

The fields carried by the PCN object are specified in [RFC6663], [RFC6661] and [RFC6662]:

- o the IPv4 or IPv6 address of the PCN-ingress-node (Aggregator) and the IPv4 or IPv6 address of the PCN-egress-node (Deaggregator); together they specify the ingress-egress-aggregate to which the report refers. According to [RFC6663] the report should carry the identifier of the PCN-ingress-node (Aggregator) and the identifier of the PCN-egress-node (Deaggregator) (typically their IP addresses);
- o Decision Point address specify the IPv4 or IPv6 address of the Decision Point. In this document this field MUST contain the IP address of the Deaggregator.
- o "R": 1 bit flag that when set to ON, signifies, according to [RFC6661] and [RFC6662], that the PCN-ingress-node (Aggregator) MUST provide an estimate of the rate (PCN-sent-rate) at which the PCN-ingress-node (Aggregator) is receiving PCN-traffic that is destined for the given ingress-egress-aggregate.
- o "Reserved": 31 bits that are currently not used by this document and are reserved. These SHALL be set to 0 and SHALL be ignored on reception.
- o PCN-sent-rate: the PCN-sent-rate for the given ingress-egress-aggregate. It is expressed in octets/second; its format is a 32-bit IEEE floating point number; The PCN-sent-rate is specified in [RFC6661] and [RFC6662] and it represents the estimate of the rate at which the PCN-ingress-node (Aggregator) is receiving PCN-traffic that is destined for the given ingress-egress-aggregate.

5. Security Considerations

The security considerations specified in [RFC2205], [RFC4860] and [RFC5559] apply to this document. In addition, [RFC4230] and [RFC6411] provide useful guidance on RSVP security mechanisms.

Security within a PCN domain is fundamentally based on the controlled environment trust assumption stated in Section 6.3.1 of [RFC5559], in particular that all PCN-nodes are PCN-enabled and are trusted to perform accurate PCN-metering and PCN-marking.

In the PCN domain environments addressed by this document, Generic Aggregate Resource ReSerVation Protocol (RSVP) messages specified in [RFC4860] are used for support of the PCN Controlled Load (CL) and Single Marking (SM) edge behaviors over a Diffserv cloud using Pre-Congestion Notification. Hence the security mechanisms discussed in [RFC4860] are applicable. Specifically, the INTEGRITY object [RFC2747][RFC3097] can be used to provide hop-by-hop RSVP message integrity, node authentication and replay protection, thereby protecting against corruption and spoofing of RSVP messages and PCN feedback conveyed by RSVP messages.

For these reasons, this document does not introduce significant additional security considerations beyond those discussed in

[RFC5559] and [RFC4860].

6. IANA Considerations

IANA has modified the RSVP parameters registry, 'Class Names, Class Numbers, and Class Types' subregistry, to add a new Class Number and assign 4 new C-Types under this new Class Number, as described below, see Section 4.1:

Class Number	Class Name	Reference
-----	-----	-----
248	PCN	this document
Class Types or C-Types:		
1	RSVP-AGGREGATE-IPv4-PCN-request	this document
2	RSVP-AGGREGATE-IPv6-PCN-request	this document
3	RSVP-AGGREGATE-IPv4-PCN-response	this document
4	RSVP-AGGREGATE-IPv6-PCN-response	this document

When this draft is published as an RFC, IANA should update the reference for the above 5 items to that published RFC (and the RFC Editor should remove this sentence).

7. Acknowledgments

We would like to thank the authors of [draft-lefaucheur-rsvp-ecn-01.txt], since some ideas used in this document are based on the work initiated in [draft-lefaucheur-rsvp-ecn-01.txt]. Moreover, we would like to thank Bob Briscoe, David Black, Ken Carlberg, Tom Taylor, Philip Eardley, Michael Menth, Toby Moncaster, James Polk, Scott Bradner, Lixia Zhang and Robert Sparks for the provided comments. In particular, we would like to thank Francois Le Faucheur for contributing in addition to comments also to a significant amount of text.

8. Normative References

[RFC6661] T. Taylor, A. Charny, F. Huang, G. Karagiannis, M. Menth, "PCN Boundary Node Behaviour for the Controlled Load (CL) Mode of Operation", July 2012.

[RFC6662] A. Charny, J. Zhang, G. Karagiannis, M. Menth, T. Taylor, "PCN Boundary Node Behaviour for the Single Marking (SM) Mode of Operation", July 2012.

[RFC6663] G. Karagiannis, T. Taylor, K. Chan, M. Menth, P. Eardley, "Requirements for Signaling of (Pre-) Congestion Information in a DiffServ Domain", July 2012.

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2205] Braden, R., ed., et al., "Resource ReSerVation Protocol (RSVP)- Functional Specification", RFC 2205, September 1997.
- [RFC3140] Black, D., Brim, S., Carpenter, B., and F. Le Faucheur, "Per Hop Behavior Identification Codes", RFC 3140, June 2001.
- [RFC3175] Baker, F., Iturralde, C., Le Faucheur, F., and B. Davie, "Aggregation of RSVP for IPv4 and IPv6 Reservations", RFC 3175, September 2001.
- [RFC4495] Polk, J. and S. Dhesikan, "A Resource Reservation Protocol (RSVP) Extension for the Reduction of Bandwidth of a Reservation Flow", RFC 4495, May 2006.
- [RFC4860] F. Le Faucheur, B. Davie, P. Bose, C. Christou, M. Davenport, "Generic Aggregate Resource ReSerVation Protocol (RSVP) Reservations", RFC4860, May 2007.
- [RFC5670] Eardley, P., "Metering and Marking Behaviour of PCN-Nodes", RFC 5670, November 2009.
- [RFC6660] Moncaster, T., Briscoe, B., and M. Menth, "Baseline Encoding and Transport of Pre-Congestion Information", RFC 6660, July 2012.

9. Informative References

- [draft-lefaucheur-rsvp-ecn-01.txt] Le Faucheur, F., Charny, A., Briscoe, B., Eardley, P., Chan, K., and J. Babiarz, "RSVP Extensions for Admission Control over Diffserv using Pre-congestion Notification (PCN) (Work in progress)", June 2006.
- [RFC1633] Braden, R., Clark, D., and S. Shenker, "Integrated Services in the Internet Architecture: an Overview", RFC 1633, June 1994.
- [RFC2211] J. Wroclawski, Specification of the Controlled-Load Network Element Service, September 1997
- [RFC2212] S. Shenker et al., Specification of Guaranteed Quality of Service, September 1997
- [RFC2474] Nichols, K., Blake, S., Baker, F., and D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC 2474, December 1998.
- [RFC2475] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z. and W. Weiss, "A framework for Differentiated Services", RFC 2475, December 1998.

[RFC2747] Baker, F., Lindell, B., and M. Talwar, "RSVP Cryptographic Authentication", RFC 2747, January 2000.

[RFC2753] Yavatkar, R., D. Pendarakis and R. Guerin, "A Framework for Policy-based Admission Control", January 2000.

[RFC2998] Bernet, Y., Yavatkar, R., Ford, P., Baker, F., Zhang, L., Speer, M., Braden, R., Davie, B., Wroclawski, J. and E. Felstaine, "A Framework for Integrated Services Operation Over DiffServ Networks", RFC 2998, November 2000.

[RFC3097] Braden, R. and L. Zhang, "RSVP Cryptographic Authentication -- Updated Message Type Value", RFC 3097, April 2001.

[RFC4230] H. Tschofenig, R. Graveman, "RSVP Security Properties", RFC 4230, December 2005.

[RFC5559] Eardley, P., "Pre-Congestion Notification (PCN) Architecture", RFC 5559, June 2009.

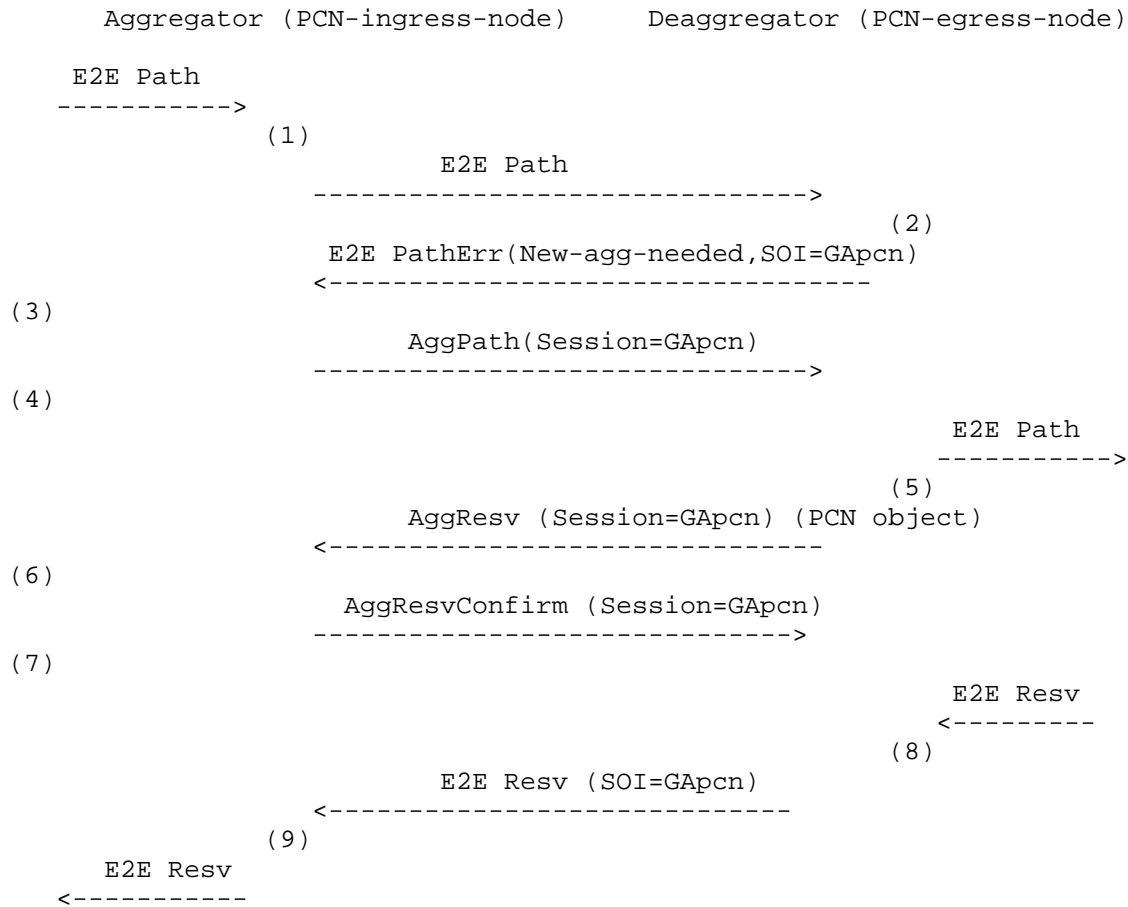
[RFC6411] M. Behringer, F. Le Faucheur, B. Weis, "Applicability of Keying Methods for RSVP Security", RFC 6411, October 2011.

[SIG-NESTED] Baker, F. and P. Bose, "QoS Signaling in a Nested Virtual Private Network", Work in Progress, July 2007.

10. Appendix A: Example Signaling Flow

This appendix is based on the appendix provided in [RFC4860]. In particular, it provides an example signaling flow of the specification detailed in Section 3 and 4.

This signaling flow assumes an environment where E2E reservations are aggregated over generic aggregate RSVP reservations and applied over a PCN domain. In particular the Aggregator (PCN-ingress-node) and Deaggregator (PCN-egress-node) are located at the boundaries of the PCN domain. The PCN-interior-nodes are located within the PCN-domain, between the PCN-boundary nodes, but are not shown in this Figure. It illustrates a possible RSVP message flow that could take place in the successful establishment of a unicast E2E reservation that is the first between a given pair of Aggregator/Deaggregator.



(1) The Aggregator forwards E2E Path into the aggregation region after modifying its IP protocol number to RSVP-E2E-IGNORE

(2) Let's assume no Aggregate Path exists. To be able to accurately update the ADSPEC of the E2E Path, the Deaggregator needs the ADSPEC of Aggregate Path. In this example, the Deaggregator elects to instruct the Aggregator to set up an Aggregate Path state for the PCN PHB-ID. To do that, the Deaggregator sends an E2E PathErr message with a New-Agg-Needed PathErr code.

The PathErr message also contains a SESSION-OF-INTEREST (SOI) object. The SOI contains a GENERIC-AGGREGATE SESSION (GApcn) whose PHB-ID is set to the PCN PHB-ID. The GENERIC-AGGREGATE SESSION contains an interface-independent Deaggregator address inside the DestAddress and appropriate values inside the vDstPort and Extended vDstPort fields. In this document, the Extended vDstPort SHOULD contain the IPv4 or IPv6 address of the Aggregator.

(3) The Aggregator follows the request from the Deaggregator and

signals an Aggregate Path for the GENERIC-AGGREGATE Session (GAp cn).

- (4) The Deaggregator takes into account the information contained in the ADSPEC from both Aggregate Paths and updates the E2E Path ADSPEC accordingly. The PCN-egress-node MUST NOT perform the RSVP-TTL vs IP TTL-check and MUST NOT update the ADSpec Break bit. This is because the whole PCN-domain is effectively handled by E2E RSVP as a virtual link on which integrated service is indeed supported (and admission control performed) so that the Break bit MUST NOT be set, see also [draft-lefaucheur-rsvp-ecn-01]. The Deaggregator also modifies the E2E Path IP protocol number to RSVP before forwarding it.
- (5) In this example, the Deaggregator elects to immediately proceed with establishment of the generic aggregate reservation. In effect, the Deaggregator can be seen as anticipating the actual demand of E2E reservations so that the generic aggregate reservation is in place when the E2E Resv request arrives, in order to speed up establishment of E2E reservations. Here it is also assumed that the Deaggregator includes the optional Resv Confirm Request in the Aggregate Resv message.
- (6) The Aggregator merely complies with the received ResvConfirm Request and returns the corresponding Aggregate ResvConfirm.
- (7) The Deaggregator has explicit confirmation that the generic aggregate reservation is established.
- (8) On receipt of the E2E Resv, the Deaggregator applies the mapping policy defined by the network administrator to map the E2E Resv onto a generic aggregate reservation. Let's assume that this policy is such that the E2E reservation is to be mapped onto the generic aggregate reservation with the PCN PHB-ID=x. The Deaggregator knows that a generic aggregate reservation (GAp cn) is in place for the corresponding PHB-ID since (7). At this step the Deaggregator maps the generic aggregated reservation onto one ingress-egress-aggregate maintained by the Deaggregator (as a PCN-egress-node), see Section 3.7. The Deaggregator performs admission control of the E2E Resv onto the generic Aggregate reservation for the PCN PHB-ID (GAp cn). The Deaggregator takes also into account the PCN admission control procedure as specified in [RFC6661] and [RFC6662], see Section 3.7. If one or both the admission control procedures (PCN based admission control procedure and admission control procedure specified in [RFC4860]) are not successful, then the E2E Resv is not admitted onto the associated RSVP generic aggregate reservation for the PCN PHB-ID (GAp cn). Otherwise, assuming that the generic aggregate reservation for the PCN (GAp cn) had been established with sufficient bandwidth to support the E2E Resv, the Deaggregator adjusts its counter, tracking the unused bandwidth on the generic aggregate reservation. Then it forwards the E2E Resv to the Aggregator including a SESSION-OF-INTEREST

object conveying the selected mapping onto GApcn (and hence onto the PCN PHB-ID).

- (9) The Aggregator records the mapping of the E2E Resv onto GApcn (and onto the PCN PHB-ID). The Aggregator removes the SOI object and forwards the E2E Resv towards the sender.

11. Authors' Address

Georgios Karagiannis
Huawei Technologies
Hansaallee 205,
40549 Dusseldorf,
Germany
Email: Georgios.Karagiannis@huawei.com

Anurag Bhargava
Cisco Systems, Inc.
7100-9 Kit Creek Road
PO Box 14987
RESEARCH TRIANGLE PARK, NORTH CAROLINA 27709-4987
USA
Email: anuragb@cisco.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: October 26, 2013

P. O'Hanlon
University of Oxford
K. Carlberg
G11
April 24, 2013

Congestion control algorithm for lower latency and lower loss media
transport
draft-ohanlon-rmcat-dflow-02

Abstract

This memo provides a design for a congestion control algorithm, for media transport, which aims to provide for lower delay and lower loss communications.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on October 26, 2013.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. Conventions, Definitions and Acronyms	3
3. Background	3
3.1. TFRC	3
3.2. Delay-Based schemes	4
4. Objectives	5
5. Design Outline	6
5.1. Delay Composition	6
5.2. Delay Measurement	6
5.3. Congestion Detection	7
5.4. Slow Start	7
5.5. Loss-mode	7
6. Further Work	8
7. IANA Considerations	8
8. Security Considerations	8
9. References	8
9.1. Normative References	8
9.2. Informative References	8
Authors' Addresses	9

1. Introduction

This memo outlines DFlow, a congestion control algorithm that aims to minimise delay and loss by using delay-based techniques. The scheme is based upon TCP Friendly Rate Control (TFRC) [RFC5348], and adds a delay-based congestion detection scheme which feeds into a 'congestion event history' mechanism based upon TFRC's loss history. This then provides for a 'congestion event rate' which drives the TCP equation.

Congestion control that aims to minimise the delay is important for real-time streams as high delay can render the communication unacceptable [ITU.G114.2003]. On today's Internet a number of paths have an excess of buffering which can lead to persistent high latencies, which has become known as the Bufferbloat phenomenon. These problems are particularly apparent with loss-based congestion control schemes such as TCP, as they operate by filling the queues on a path till loss occurs, thus maximising the delay. The unfortunate consequence is that loss-based approaches not only lead to high delay for their own packets but also introduce delays and losses for all other flows that traverse those same filled queues.

Thus when competing with TCP, without the widespread deployment of Active Queue Management that aims to minimise delay, (e.g. Codel [I-D.nichols-tsvwg-codel]), it is not possible to maintain low delay

as TCP will do its best to keep the queues full and maximise the delay.

However there are many paths where the flows are not competing directly with TCP and where delay may be minimised.

The DFlow scheme can transport media with low delay and loss on paths where there is no direct competition with TCP in the same queue. Though we are currently testing some techniques to enable it compete with loss-based schemes (at the expense of delay) but they will be included in a later version of the draft. In simulations it has been seen to be reasonably fair when competing with other DFlow streams.

2. Conventions, Definitions and Acronyms

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

3. Background

Whilst the existing standard for media transport, Real-time Transport Protocol (RTP) [RFC3550], suggests that congestion control should be employed, in practice many systems tend to use fixed or variable bit rate UDP and do very little or no adaptation to their network environment. Most of the existing work on real-time congestion control algorithms has been rooted in TCP-friendly approaches but with smoother adaptation cycles. TCP congestion control is unsuitable for interactive media for a number of reasons including the fact that it is loss-based so it maximises the latency on a path, it changes its transmit rate to quickly for multimedia, and favours reliability over timeliness. Various TCP-friendly congestion control algorithms such as TFRC [RFC5348], Sisalem's LDA+ [SisalemLDA.2000], and Choi's TCP Friendly Window Control (TFWC) [ChoiTFWC.2007] have been devised for media transport, that attempt to smooth the short-term variation in sending rate. More recently there have been development of some delay-based schemes which aim to provide for low delay.

3.1. TFRC

TFRC is a rate based receiver driven congestion control algorithm which utilises the Padhye TCP equation to provide a smoothed TCP-friendly rate. The sender explicitly sets the transmission rate, using the TCP equation driven by the loss event rate which is measured and fed back by the receiver, where a loss event consists of one or more packet losses within a single RTT. It utilises a

weighted smoothed loss event rate, and EWMA smoothed RTT, as input to the TCP equation which enables it to achieve a smoother rate adaptation that provides for a more suitable transport for multimedia. TFRC was primarily aimed at streaming media delivery where a smooth rate and TCP-friendliness are more important than low latency operation.

However there are number of issues with TFRC as regards real-time media transport:

Loss-based operation: Firstly since it is a loss-based based scheme the latency is maximised which is a problem for real-time transport over heavily buffered paths. The other problem with loss-based protocols is that they rely on a certain level of packet loss which can be an issue for media traffic since lost media packet cannot usually be retransmitted in time. This problem becomes more of a concern at lower transmission rates since the TCP equation requires a corresponding increase in loss rates.

Bursty media flows: Many media flows exhibit bursty behaviour due to a number of factors. Firstly there may be negative bursts (i.e. gaps) due to silence or low motion which can lead oscillatory behaviours due to the data-limited and/or idle behaviours. Secondly there may be positive bursts (i.e. larger than normal) can also be due to the bursty nature of the media and codec (e.g. I-frames) which can be lead to drops or increased latency. Whilst the current version of TFRC [RFC5348] has attempted to address some of these issues, they are still a concern.

Small RTT environments: When operating in low RTT environments (<5ms), such as a LAN, systems implementing TFRC can have problems with scheduling packet transmissions as inter-packet timings can be lower than application level clock granularity. Whilst the current version of TFRC [RFC5348] has attempted to address these issues, they can still be a concern in some low RTT environments.

Variable packet sizes: As originally designed TFRC will only operate correctly when packet sizes are close to MTU size, and when the packet sizes are much smaller fairness issues arise. Although there have been attempts to address this problem for small packets [RFC4828], it is not clear how to deal with flows that do vary their packet sizes substantially. However this issue is only really a marked problem with lower bit rate video flows or variable packet rate audio.

3.2. Delay-Based schemes

In the last few years there has been a renewed interest in the use of delay based congestion control for media, with a slightly different emphasis to that of the history of TCP based approaches such as Jain's CARD, Wang and Crowcroft's Tri-S, Brakmo's Vegas, Tan et al's Compound TCP, and more recently Budzisz's CxTCP [BudziszCxTCP.2011]. Where the primary goal with media based transports is to actually minimise the latency of the flow, as opposed to just using delay as an early indication of loss. This is of particular relevance on paths with large queues, as is the case with a number of today's Internet paths. In 2007 Ghanbari et al [GhanbariFuzzy.2007] did some pioneering work on delay-based video congestion control using fuzzy logic based systems. Recently there has been on going activity in the IETF as part of the Low Extra Delay Background Transport (LEDBAT) Working Group which aims to provide a less than best effort delay-based transport with lower delay. However [RFC6817] specifies a one-way queuing delay target of 100ms which is quite a high baseline for interactive media, considering the recommended total one-way delay limit for a VoIP call should be less than 150ms [ITU.G114.2003].

4. Objectives

The objectives of DFlow are to provide for low delay and low loss media transport when possible. We also aim to provide (in a future version of the draft) mechanisms to provide for better burst management, and loss-mode operation.

Lower Delay: The one-way delay should be kept well within the acceptable levels of 150ms, and MUST NOT exceed 400ms [ITU.G114.2003].

Lower Loss: For media transport it is important to minimise loss as it is usually not possible to retransmit within the delay budget for many connections. Whilst modern codecs can tolerate some loss it is beneficial to avoid it. The advantage of low delay congestion control is that since it aims to operate within the queuing boundaries it generally avoids loss.

Smoothness: The media rate should aim to be smooth within the constraints of the media, codec, and the network path. A smooth rate generally provides for more palatable media consumption.

Fairness: The system should aim to be reasonably fair with itself and TCP flows. Initially we aim for self fairness, and we will aim to tackle TCP fairness when we have sufficiently robust loss-mode operation.

[Burst Management]: [Due in later rev] We are working on mechanisms to manage the bursty nature of media allowing it maintain a smoother quality.

[Loss-based mode]: [Due in later rev] We are working on mechanisms to allow the system to compete with loss-based congestion control and maintain throughput, though without additional network support it is understood that the delay (and loss) would be largely beyond control.

5. Design Outline

The DFlow scheme aims to primarily utilise delay measurements to drive the congestion control. It currently utilises some of the core aspects of TFRC, such as its rate based operation, utilisation of the TCP equation, and its rate smoothing. It also employs similar signalling mechanisms. However as the design evolves we expect that DFlow may diverge further from TFRC.

5.1. Delay Composition

The total end-to-end one-way delay (OWD) a packet incurs may be considered to consist of four elements; transmission (or serialisation), propagation, processing, and queuing delays. For our purposes the first three elements may be considered together as a largely static component, termed the base delay. The base delay generally does not change significantly unless the node is mobile or the underlying link alters due to something like a route change. The main dynamic element of the delay, which DFlow aims to utilise, is the queuing delay. Taken together with the base delay, the queuing delay provides an indication of the actual path latency and also provides an insight as to the level of congestion on the path.

5.2. Delay Measurement

The notional one-way delay is measured for each packet by comparing the sender and receiver timestamps. Whilst the clocks on the sender and receiver are unlikely to be synchronised, it is assumed that their offset is relatively constant as the clock skew is generally quite small. Thus the notional OWD may only be used in a relative context. The notional OWD is measured for each packet over two sampling periods; Firstly over the longer base_period (typically $10 \times \text{RTT}$) from which the minima are stored as the base_delay. And secondly it is sampled over a shorter period current_period (typically 50ms), which is also filtered, usually also using a minima filter, and stored as current_delay. The minima of the OWDs are used to reduce noise of the measurements, which can be beneficial in the case of variable link types such as wireless.

5.3. Congestion Detection

The delay-based detection algorithm, outlined in Figure 1, operates by comparing the `current_delay` to the `base_delay`, which gives an indication of the queuing delay. If it exceeds a set congestion detection threshold, `cd_thresh`, then the packet is considered for the next stage of detection. The `cd_thresh` sets the limit for the queuing delay incurred by the flow, and is typically set at 50ms (we are also investigating automated thresholds). Once a flow has exceeded its `cd_thresh` then it undergoes a second test which is based upon the gradient of the delay change over two `current_period`'s, indicating that delay is on the increase, if it is positive then a 'congestion event' is flagged.

```
If ((base_delay - current_delay) > cd_thresh AND
    (current_delay - prev_current_delay) > 0)
    DelayCongestionEvent = True
```

Figure 1: Congestion Detection pseudo-code

This algorithm then provides input to the 'congestion interval history' mechanism (based on TFRC's 'loss interval history'), which is combined with normal input from the TFRC packet loss detection mechanisms, from which a 'congestion event rate' is derived which is then fed into the TCP equation to determine the send rate.

Note that we currently disable TFRC's oscillation reduction mechanism from [RFC5348] (Section 4.5) as it adversely affects the delay-based operation.

We have performed a number of simulations of the above mechanism in operation and have found it to be reasonably fair to itself, providing for smooth rates at suitable RTTs.

5.4. Slow Start

The delay based congestion detection is not only used during normal the congestion avoidance phase of the protocol but it also employed during slow start allowing for rapid, lower loss, attainment of the operating rate.

5.5. Loss-mode

We are actively investigating techniques to enable competitive behaviours with loss-based protocol such as TCP. We aim to develop a solution that provides for automatic fallback between loss and delay modes.

6. Further Work

The design is still under active development and there is more work to be done. We are seeking feedback on these ideas and future directions.

7. IANA Considerations

This document makes no requests of IANA.

8. Security Considerations

With a congestion control algorithm an attacker can attempt to interfere with the protocol to cause rate changes. However encryption of the protocol will largely protect it against such threats.

9. References

9.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4828] Floyd, S. and E. Kohler, "TCP Friendly Rate Control (TFRC): The Small-Packet (SP) Variant", RFC 4828, April 2007.
- [RFC5348] Floyd, S., Handley, M., Padhye, J., and J. Widmer, "TCP Friendly Rate Control (TFRC): Protocol Specification", RFC 5348, September 2008.

9.2. Informative References

- [BudziszCxTCP.2011]
Budzisz, L., Stanojevic, R., Schlote, A., Shorten, R., and F. Baker, "On the Fair Coexistence of Loss- and Delay-Based TCP", July 2011.
- [ChoiTFWC.2007]
Choi, S. and M. Handley, "Fairer TCP-friendly congestion control protocol for multimedia streaming applications", Dec 2007.
- [GhanbariFuzzy.2007]
Jammeh, E., Fleury, M., and M. Ghanbari, "Delay-based congestion avoidance for video communication with fuzzy logic control", Nov 2007.

- [I-D.nichols-tsvwg-codel]
Nichols, K. and V. Jacobson, "Controlled Delay Active Queue Management", draft-nichols-tsvwg-codel-01 (work in progress), February 2013.
- [ITU.G114.2003]
International Telecommunications Union, "One-way transmission time", ITU-T Recommendation G.707, May 2003.
- [RFC3550] Schulzrinne, H., Casner, S., Frederick, R., and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications", STD 64, RFC 3550, July 2003.
- [RFC6817] Shalunov, S., Hazel, G., Iyengar, J., and M. Kuehlewind, "Low Extra Delay Background Transport (LEDBAT)", RFC 6817, December 2012.
- [SisalemLDA.2000]
Sisalem, D. and A. Wolisz, "LDA+: A TCP-Friendly Adaptation Scheme for Multimedia Communication", May 2000.

Authors' Addresses

Piers O'Hanlon
University of Oxford
Oxford Internet Institute
1 St Giles
Oxford OX1 3JS
United Kingdom

Email: piers.ohanlon@oii.ox.ac.uk

Ken Carlberg
G11
1600 Clarendon Blvd
Arlington VA
USA

Email: carlberg@g11.org.uk

Network WG
Internet-Draft
Intended status: Informational
Expires: January 9, 2012

James Polk
Cisco
July 9, 2012

The Problem Statement for the Standard
Configuration of DiffServ Service Classes
draft-polk-tsvwg-diffserv-stds-problem-statement-00.txt

Abstract

This document describes the problem statement on two recently proposed expansions to DiffServ. The first of these expansions proposes updating the informational RFC 4594 document to standards track status, while making the necessary changes to make it current; for example, creating more granular traffic treatments, some with new Per Hop Behaviors (PHB). The second proposal defines 6 new DiffServ Codepoints necessary from these new PHBs in the proposal within the first draft.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on January 9, 2012.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. Brief Overview of RFC 4594 and RFC 5127	3
2.1 Brief Overview of RFC 4594	3
2.2 Brief Overview of RFC 5127	4
3. Brief Discussion of the RFC 4594 Update Draft	5
4. Conclusion and What's Next	7
5. Acknowledgements	7
6. IANA Considerations	7
7. Security Considerations	8
8. References	8
8.1 Normative References	8
8.2 Informative References	8
Author's Address	9

1. Introduction

Differentiated Services (DiffServ) [RFC2474] creates an IP header marking or indicator with which intermediate nodes (i.e., routers and switches) can make policy decisions. These 6-bit values are called Differentiated Services Codepoint Point (DSCP) values. DSCP values are used to differentiate packet treatment within an intermediate node, not across a network, as the conditions affecting that marking are different within each node. This is called Per Hop Behavior (PHB). In other words, even though a packet has the same DSCP from source to destination, it can and often does experience different treatment depending on the conditions of the nodes it traverses on its journey.

The DiffServ architecture allows for DSCP values within a packet to be changed, or remarked, any number of times. In other words, a packet can have its DSCP remarked at every layer-3 hop throughout the life of that packet. This practice actually occurs infrequently, but it is allowed.

At issue is a combination of the number of networks or endpoints that are choosing to use DiffServ markings, and the number of administrative domains (called "networks" in this document) a packet traverses with different policies for how packet flows of a similar type (e.g., a voice flow, or an email flow, etc.) are to be marked.

The community presently has RFC 4594 [RFC4594], which is an informational guideline on how networks can or should mark certain packet flows with differing traffic characteristics using DiffServ. There are several reasons why this informational RFC lacks the necessary clarity and strength to reach widespread adoption:

- o confusion between RFC 4594 and RFC 5127 [RFC5127], the latter of which is for aggregating many 6-bit DSCP values into a 3-bit (8

value) field used specifically by service provider (SP) networks.

- o some believe both RFCs are for SPs, while others ignore RFC 5127 and use RFC 4594 as if it were standards track or BCP.
- o some believe RFC 5127 is for SPs only, and want RFC 4594 to reduce the number of DSCPs within its guidelines to recommend using only 3 or 4 DSCPs. This seems to stem from a manageability and operational perspective.
- o some know RFC 4594 is informational and do not follow its guidelines specifically because it is informational.
- o some use DSCP values that are not defined within RFC 4594, making mapping between different networks using similar or identical application flows difficult.
- o some believe enterprise networks should not use either RFC except at the edge of their networks, where they directly connect to SP networks.
- o some argue that the services classes guidance per class is too broad and are therefore not sure in which service class a particular application is to reside.

This document is not intended to reach RFC status. Rather, it is to stimulate discussion on both RFC 4594 and 5127 to lessen existing confusion within the community. It should be noted that RFC 4594 has an offered update within TSVWG [ID-4594-UPDATE]. This draft has created some heated discussions within that WG before and during the Paris IETF meeting.

First, we'll discuss briefly RFCs 4594 and 5127 in Section 2. Then we will discuss what the update to RFC 4594 proposes differently and what we expect to happen to RFC 5127 in Section 3.

2. Brief Overview of RFC 4594 and RFC 5127

2.1 Brief Overview of RFC 4594

Essentially, RFC 4594 is a guideline for how to choose which DSCP to use based on the traffic characteristics an application flow needs to experience within a network for optimal performance. RFC 4594 specifically points to several existing standards-track DiffServ RFCs to augment the text in each of those RFCs, without violating any of the rules within each of those documents. RFC 4594:

- o painstakingly lays out definitions and guidelines for each service class.
- o clearly indicates each service class's tolerance to delay, jitter

and packet loss.

- o details the conditioning treatments at the Differentiated Services (DS) edge.
- o categorizes traffic characteristics into 12 service classes utilizing one or more DSCPs:

Network Control	Broadcast Video
Telephony	Low-Latency Data
Signaling	OAM
Multimedia Conferencing	High-throughput Data
Realtime Interactive	Standard
Multimedia Streaming	Low-priority Data

2.2 Brief Overview of RFC 5127

At its barest, RFC 5127 recommends that, of the many service classes described within RFC 4594, each having different traffic characteristics, similar service classes be grouped or aggregated into 3, 4, or 5 markings for SP traversal. This limitation of the number of individual service classes is partly to reduce the number of separate distinctions traversing over their network because SPs have difficulty managing what is deemed 'too many' different classes. Another part for this reduction is customer expectations of meeting contractual Service Level Agreements (SLAs).

To this end, and perhaps because of it, MPLS was designed with only 8 values of priority differentiation, i.e., the 3 EXP bits. To be fair, LAN based IEEE has only a 3-bit priority field as well within its specifications, known as the Priority Code Point (PCP), as part of the 802.1Q header spec. IEEE 802.1e, which defines QoS over Wi-Fi, also only defines 8 levels (called User Priority or UP codes).

The result is to have the IETF within RFC 5127 recommend the following (which is Figure 2 within that RFC):

Treatment Aggregate Behavior	Treatment Aggregate Behavior	DSCP
Network Control	CS (RFC 2474)	CS6
Real-Time	EF (RFC 3246)	EF, CS5, AF41, AF42, AF43, CS4, CS3
Assured Elastic	AF (RFC 2597)	CS2, AF31, AF21, AF11
		AF32, AF22, AF12
		AF33, AF23, AF13
Elastic	Default (RFC 2474)	Default, (CS0)
		CS1

Figure 1: Treatment Aggregate Behavior

RFC 5127 goes on to recommend the marking and treatments on either side of the provider edge remain the same. In other words, the DSCP values remain the same and are used to determine which queue to place the packets into within the aggregates, where the packets are treated the same within that tunnel until the egress provider edge.

Many within enterprise networks do not pay attention to what RFC 5127 says because they are sufficiently removed from dealing with the constraints of very few DSCP values or the need to aggregate DSCP values into groups.

3. Brief Discussion of the RFC 4594 Update Draft

The RFC 4594 update draft [ID-4594-UPDATE] proposes to update what has occurred since RFC 4594 was written (i.e., 2006), in which more granular service classes can be differentiated by application requirements. For example, Figure 2 within RFC 4594 identifies "Telephony" as having 'Fixed-size small packets'. That is not true for today's video flow, therefore it needs to be modified. The update draft currently breaks out audio and video separately to reflect this different, as well as the ability to treat each traffic type differently within a network. Another example is gaming and TCP. The two were believed by most, and it is still believed by many that gaming requires a UDP delivery due to the requirements for timely delivery of packets and that retransmissions would cause delays and bad things to happen to gaming applications. This was proved false within [ID-TCMTF], in which the author of that document

had a presentation showing TCP was used and viable.

[RFC5865] created a new Expedited Forwarding (EF) DSCP value called VOICE-ADMIT, the second time an application is identified within the DiffServ realm. The first was the service class Broadcast Video, which is poorly used within RFC 4594 because other types of flows can be 'broadcast' other than video, such as audio. From this, [ID-4594-UPDATE] moved in two directions:

- o it called out two service classes (audio and video), even though audio and video packets are not the only types of packets within each traffic characteristic.
- o it removed "Video" from the Broadcast service class name.

From the resistance to this proposal within [ID-4594-UPDATE], perhaps other service class label names should be used.

The draft also recognizes the differences in video traffic, even though it is always carried over RTP [RFC3550]. Aside from silence suppression, video traffic varies far more than audio traffic. For example, video is

- o far more variable in bandwidth utilization within the same flow.
- o far more variable in packet size.
- o at different business priorities in some networks based on a configuration. For example, desktop video often is of less important than Telepresence video on the same network. Lacking congestion, the two are treated the same. When congestion exists, one is given priority over the other.

Consequently any service class that contains video needs to account for larger packet size variation than audio, which was equally true in 2006, but not contained in RFC 4594.

Further, with the publication of RFC 5865, the concept of 'capacity admitted' traffic flows have been defined within DiffServ, and are being expanded with the proposal within this new draft [ID-NEW-DSCPS]. There are differing opinions as to whether the realtime Treatment Aggregate in Figure 1 above should also contain these capacity admitted flows, or if 'capacity admitted' traffic flows should have their own Treatment Aggregate containing all realtime capacity admitted traffic. Mixing capacity admitted traffic with unbounded realtime traffic seems to be trouble from a predictability point of view within routers believing they individually understand exactly how much traffic will be traversing each interface and at what rate.

All this said, there is a valid argument to constrain or prevent any DSCP value from being assigned to a single application, mostly due

to the limitation of the overall number of DSCP values available for use. [ID-4594-UPDATE] provides at least several applications per service class (or DSCP); a fact many have overlooked to date.

[ID-4594-UPDATE] is not only about or because of realtime traffic. It is also an overall update to the ideas and guidelines within RFC 4594, with the intent to make that document a standards track document for interoperability purposes.

4. Conclusion and What's Next

Without attempting to fundamentally change the guidelines within RFC 5127, this effort should not be as controversial as it has been, if we understand that those networks that need more granular traffic treatments can be configured with more granularity while not violating the needs of other networks that do not wish to be made aware of the increased treatment differences.

Everyone involved in this discussion needs to have a clear understanding of the difference points of view within the RFC 4594 effort (i.e., the RFC and the update draft) as well as within RFC 5127. One focuses on defining each service class and the other focuses on determining which of the existing service classes go into which aggregate, if present.

We hope to form a BoF on this subject that will explicitly *not* form a working group or produce any documents, or even drafts, but will gather the community from several (if not all) areas, and not just within the transport area. That is the purpose of this draft: to stimulate discussion towards the goal of discussion within the community on DiffServ. If the community does not believe a BoF is necessary, the work will proceed, or not, in TSVWG. Knowing how many within the community have attended TSVWG in each meeting for the last 9 or so years, it is felt that a much wider audience is necessary, given how much impact [ID-4594-UPDATE] can potentially have.

5. Acknowledgements

The author would like to thank Gorrry Fairhurst and David Black for their positive discussions towards the formation of a BoF in Vancouver IETF. The author would also like to thank Paul Jones for doing a valuable proof read to catch points I didn't make clear, as well as identify simple nits I should have caught the nth time I reread this.

6. IANA Considerations

There are no IANA considerations as a result of this document.

7. Security Considerations

There are no security considerations within this document because it will not be progressed beyond this individual contributor stage, and all the specifying will be done in other drafts that will wholly contain all the security considerations for this goal/idea.

8. References

8.1 Normative References

There are no normative references within this document.

8.2 Informative References

- [RFC2474] K. Nichols, S. Blake, F. Baker, D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers ", RFC 2474, December 1998
- [RFC3550] Schulzrinne, H., Casner, S., Frederick, R., and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications", STD 64, RFC 3550, July 2003.
- [RFC4594] J. Babiarz, K. Chan, F Baker, "Configuration Guidelines for Diffserv Service Classes", RFC 4594, August 2006
- [RFC5127] Chan, K., Babiarz, J., and F. Baker, "Aggregation of DiffServ Service Classes", RFC 5127, February 2008.
- [RFC5865] F. Baker, J. Polk, M. Dolly, "A Differentiated Services Code Point (DSCP) for Capacity-Admitted Traffic", RFC 5865, May 2010
- [ID-4594-UPDATE] J. Polk, "Standard Configuration of DiffServ Service Classes", "work in progress", March 2012
- [ID-NEW-DSCPS] J. Polk, "New Differentiated Services Code Point Assignments for Rich Media Traffic", "work in progress", March 2012
- [ID-TCMTF] J. Saldana, D. Wing, J. Fernandez Navajas, Muthu A M. Perumal, J. Ruiz Mas, "Tunneling Compressed Multiplexed Traffic Flows (TCMTF)", "work in progress", March 2012

Authors' Address

James Polk
3913 Treemont Circle
Colleyville, Texas 76034

Phone: +1.817.271.3552
Email: jmpolk@cisco.com

Network WG
Internet-Draft
Intended status: Standards Track (PS)
Expires: August 25, 2013

James Polk
Cisco
Feb 25, 2013

New Differentiated Services Code Point Assignments
for Rich Media Traffic
draft-polk-tsvwg-new-dscp-assignments-02.txt

Abstract

This document requests five new Differentiated Services Code Point (DSCP) values (DSCP) from the Internet Assigned Numbers Authority (IANA) for new classes of rich media traffic and one additional DSCP value for the signaling of multimedia sessions.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on August 25, 2013.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. Terminology	4
3. Evolution of the Proposed DSCPs	4
4. New DSCP Assignments.	7
5. Acknowledgements	8
6. IANA Considerations	8
7. Security Considerations	9
8. References	9
8.1 Normative References	9
8.2 Informative References	10
Author's Address	10

1. Introduction

This document requests five new Differentiated Services Code Point (DSCP) values (DSCP) from the Internet Assigned Numbers Authority (IANA) for new classes of rich media traffic and one additional DSCP value for the signaling of multimedia sessions. Four of the six new DSCP values are for traffic classes that are admitted by the network using an additional Capacity-Admission signaling procedure to the normal signaling that occurs between multiple endpoints establishing a traffic flow between endpoints. The additional capacity-admission signaling procedure is offered in RFC 5865 [RFC5865], which defined the Voice-Admit per hop behavior (PHB) DSCP. Each of these four traffic classes can conform to the Expedited Forwarding Per-Hop Behavior, if configured to do so, using the Priority Queuing system such as that defined in Section 1.4.1.1 of [ID-4594-UP].

It is expected that voice and video media samples will be carried using the Real-time Transport Protocol (RTP) [RFC3550], thus making voice by itself indistinguishable from video to routers and switches, unless one of two things occurs:

- o Deep packet inspection (DPI) at the ingress of each DiffServ edge node to determine that the packet is an RTP packet with a certain codec that properly identifies it as either a voice or video packet, or
- o have a separate marking for the packets (i.e., a different DSCP).

It is certainly the case that voice samples/frames can be in the same packet as video frames, thus making the packet marked either voice or video, but that will have to be left to the application to decide if that is a good idea. For what it is worth, most current implementations of mixing the media types have the packets marked as a video.

This effort is based on the work started in RFC 5865 [RFC5865], a Differentiated Services Code Point for Capacity-Admitted Traffic voice only traffic, which recommends the classes created within RFC 4594 [RFC4594] be extended for video traffic flows of different types. Nearly all of what is requested and referenced here is based on what started in RFC 4594, but with video as the dominant application as RFC 5865 recommends. Presently, RFC 4594 is being updated by [ID-4594-UP] for many reasons, including the inclusion of these six new DSCPs.

These four new video classes differ from their existing counterparts in behavior by not being subjected to capacity admission. All of the mentioned traffic classes and subsequent DSCPs within RFC 4594 are non-binding, given that it is a non-normative RFC. RFC 4594 also did not recommend the need for capacity admission traffic classes (aka with associated DSCP values). This document is symbiotic with [ID-4594-UP] which intends to replace RFC 4594 as a standards track update which includes the new DSCP assignments created within this document.

Thus, RFC 4594 defined the need for application assignment of certain DSCPs, but only non-normatively. RFC 5865 defined updated DSCP values for a capacity-admitted voice traffic class that is normative. This document takes what was in RFC 4594, creates 4 new capacity-admitted traffic classes and associated DSCPs. This document also moves one non-capacity-admitted traffic class as well as moves the recommended audio/video signaling DSCP value to another value.

Within RFC 5865, there is the specific call for additional DSCPs for capacity-admitted traffic flows of real-time rich media (video) flows in Section 3 of that document under the heading "Summary: Changes from RFC 4594".

It should be noted here that these flows are typically video flows, and frequently include the audio with the adjoining video traffic within that flow. The details of how that gets sorted out are outside the scope of this document. DiffServ is a known and proven mechanism. This document does not change or challenge the idea that Differentiated Services is a Per Hop Behavior (PHB) mechanism, and does not create a service. Here we merely want to add new DSCP assignments because of how at least some of the world is (or wants to) differentiate video from other traffic, including other video traffic.

Section 3 will discuss some of the evolution of DSCP assignments, focusing on those aspects pertinent to the creation of these six new DSCP values. Section 4 describes and defines each of the six DSCP values being requested. Heavy reliance exists on the text of RFC 5865 for its diagrams and charts. Those were not brought into this document at this time, but could be in the future.

2. Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

CAC - defined in RFC 5865

PHB - defined in RFC 5865

DSCP - defined in RFC 5865

Queue - defined in RFC 5865

3. Evolution of the Proposed DSCPs

First of all, full consideration of PHBs and DSCPs needs to originate with RFC 2474. Section 6 of that document states the following:

"The DSCP field within the DS field is capable of conveying 64 distinct codepoints. The codepoint space is divided into three pools for the purpose of codepoint assignment and management: a pool of 32 RECOMMENDED codepoints (Pool 1) to be assigned by Standards Action as defined in [CONS], a pool of 16 codepoints (Pool 2) to be reserved for experimental or Local Use (EXP/LU) as defined in [CONS], and a pool of 16 codepoints (Pool 3) which are initially available for experimental or local use, but which should be preferentially utilized for standardized assignments if Pool 1 is ever exhausted. The pools are defined in the following table (where 'x' refers to either '0' or '1'):

Pool	Codepoint space	Assignment Policy
----	-----	-----
1	xxxxx0	Standards Action
2	xxxxx11	EXP/LU
3	xxxxx01	EXP/LU (*)

(*) may be utilized for future Standards Action allocations as Necessary"

The key part of the above quote is

"... which should be preferentially utilized for standardized assignments if Pool 1 is ever exhausted..."

which we here take to mean 'SHOULD NOT use unless you have a really good reason to use'. We propose what we consider a really good

reason to use some of the assignments from Pool 3 before Pool 1 is exhausted. One reason for assigning out of Pool 3 is to get similar marking from layer 2 technologies that only have 3 bits to use for their value, not 6 bits. Technologies such as 802.3 Ethernet, 802.11 Wireless Ethernet, and MPLS are 3 examples of technologies that only have 3 bits to use.

[Editor's Note: If this aspect of assigning DSCPs from Pool 3 before Pool 1 is exhausted requires an update to RFC 2474, please let the authors know so we can point this out to the community for additional feedback.]

Just as RFC 5865 matched the first 3 (or 4) bits with EF for Voice-Admit (101110 and 101100), we RECOMMEND the admitted DSCP for an existing value be its XXXX01 version of the non-admitted DSCP (XXXXX0). We note that the last two bits MUST NOT be x11 because that would mean the value is a Pool 2 value, which is forbidden currently by RFC 2474.

Thus, a DSCP value commonly traverses a layer 2 device by ignoring the last 3 bits of the DSCP value, i.e., taking EF, which is 101110, and reducing it to 101 only, and transmitting this over the layer 2 infrastructure.

RFC 4954, and its intended replacement document [ID-4594-UP], create several service classes primarily intended for video traffic with slightly different characteristics. It was stated there that not all video DSCP values from RFC 4594 are expected to be within the same network, but that could be the case.

RFC 4594 listed these voice and video services classes:

- o "Telephony" using the EF DSCP
- o "Realtime Interactive" using the CS4 DSCP
- o "Multimedia Conferencing" using the AF4X DSCP
- o "Multimedia Streaming" using the AF3X DSCP
- o "Broadcast Video" using the CS3 DSCP

Plus, for Telephony Signaling

- o "Signaling" using the CS5 DSCP

[ID-4594-UP] lists these 'non-admitted' voice and video services classes (some with changed service names, as well as some DSCPs changed):

- o Audio using the EF DSCP

- o Video using the AF4X DSCP
- o Hi-Res using the CS4 DSCP
- o Realtime-Interactive using the CS5 DSCP
- o Multimedia Streaming using the AF3X DSCP
- o Broadcast using the CS3 DSCP

The Multimedia Conferencing purpose and meaning has been changed within [ID-DSCP-UP], as has its DSCPs, which will be listed in the next set of bullets and defined within this document.

RFC 5865 created the new capacity-admitted Voice-Admit, which mentions specifically that a reservation protocol, "such as RSVP" is used to establish those sessions or traffic flows.

This document creates six additional services classes that are incorporated into [ID-4594-UP]:

- o Hi-Res-Admit using the CS4-Admit (100001) DSCP
- o Realtime-Interactive-Admit using the CS5-Admit (101001) DSCP
- o Multimedia Conferencing using the MC (011101) DSCP
- o Multimedia Conferencing-Admit using the MC-Admit (100101) DSCP
- o Broadcast-Admit using the CS3-Admit (011001) DSCP

Plus, for Conversational Signaling (a term described in [ID-4594-UP]), which is no longer to use the CS5 DSCP,

- o "A/V-Sig" using the 010001 DSCP

The results of this are that the

- CS4-Admit is the xxxxxx1 version of CS4.
- CS5-Admit is the xxxxxx1 version of CS5.
- CS3-Admit is the xxxxxx1 version of CS3.

MC-Admit is not the xxxxxx1 version of the new MC DSCP value (100101), because there are no more 100xxx values that are available, outside of the two x11 values from Pool 2, which cannot be assigned for public use.

[Editor's Note: The author is open to suggestions from the community for how to resolve this issue, if anyone considers it an issue.]

The new goal for the signaling service class is to not be starved. It has been shown that mission critical voice and video call set-up does not require expedited forwarding as a PHB. However, this service class MUST NOT be starved, and so it is RECOMMENDED to use a codepoint similar in characteristics to the RFC 4594 (and [ID-4594-UP] defined Low-Latency Data service class of 010xxx.

4. New DSCP Assignments

4.1 The CS5-Admit PHB

'CS5-Admit' MUST be used with a capacity-admission signaling procedure similar to what is required of 'Voice-Admit' [RFC5865]. RSVP [RFC2205] and NSIS [RFC4080] are two good examples for data-path signaling for capacity-admission. Neither is mandatory, but one of them SHOULD be used.

CS5-Admit has traffic characteristics described in [ID-4594-UP].

The DSCP value requested for CS5-Admit is 101001.

4.2 The CS4-Admit DSCP

'CS4-Admit' MUST be used with a capacity-admission signaling procedure similar to what is required of 'Voice-Admit' [RFC5865]. RSVP [RFC2205] and NSIS [RFC4080] are two good examples for data-path signaling for capacity-admission. Neither is mandatory, but one of them SHOULD be used.

CS4-Admit has traffic characteristics described in [ID-4594-UP].

The DSCP value requested for CS4-Admit is 100001.

4.3 The CS3-Admit DSCP

'CS3-Admit' MUST be used with a capacity-admission signaling procedure similar to what is required of 'Voice-Admit' [RFC5865]. RSVP [RFC2205] and NSIS [RFC4080] are two good examples for data-path signaling for capacity-admission. Neither is mandatory, but one of them SHOULD be used.

CS3-Admit has traffic characteristics described in [ID-4594-UP].

The DSCP value requested for CS3-Admit is 011001.

4.4 The MC DSCP

'MC' SHOULD NOT use a capacity-admission signaling procedure. Rather, the MC-Admit is used with a capacity-admission signaling procedure if needed. This PHB MUST be non-admitted.

MC has traffic characteristics described in [ID-4594-UP].

The DSCP value requested for MC is 011001.

4.5 The MC-Admit DSCP

'MC-Admit' MUST be used with a capacity-admission signaling procedure similar to what is required of 'Voice-Admit' [RFC5865]. RSVP [RFC2205] and NSIS [RFC4080] are two good examples for data-path signaling for capacity-admission. Neither is mandatory, but one of them SHOULD be used.

MC-Admit has traffic characteristics described in [ID-4594-UP].

The DSCP value requested for MC-Admit is 100101.

4.6 The Conversational Signaling (A/V-Sig) DSCP

'A/V-Sig' MUST be used with a capacity-admission signaling procedure similar to what is required of 'Voice-Admit' [RFC5865]. RSVP [RFC2205] and NSIS [RFC4080] are two good examples for data-path signaling for capacity-admission. Neither is mandatory, but one of them SHOULD be used.

A/V-Sig has traffic characteristics described in [ID-4594-UP].

The DSCP value requested for A/V-Sig is 010001.

5. Acknowledgements

The author would like to thank Paul Jones, Glen Lavers, Mo Zanaty, David Benham, Michael Ramalho for their comments and questions about this effort that ultimately helped shape this document.

6. IANA Considerations

IANA is requested to make the following registry assignments from Pool 1 and Pool 3 from the dscp-parameters section within IANA. Justification for assigning from Pool 3 is in Section 3 of this document, and are the only possible parallel assignments to existing assignments of similar registries - very much for the reason Voice-Admit [RFC5865] was assigned a codepoint similar to EF. That

aspect is the main point of this document.

6.1 DSCP Assignments from Pool 1

The code point described in this document is requested to be added to the Pool 1 Codepoint table as follows:

Sub-registry: Pool 1 Codepoints

Reference: [RFC2474]

Registration Procedures: Standards Action

Registry: Name	Space	Reference
-----	-----	-----
A/V-Sig	010010	[this document]

6.2 DSCP Assignments from Pool 3

A new "Pool 3 Codepoints" table is requested to be built by IANA similar to the Pool 1 Codepoint table in the form:

Sub-registry: Pool 3 Codepoints

Reference: [RFC2474]

Registration Procedures: Standards Action

Registry: Name	Space	Reference
-----	-----	-----
CS5-Admit	101001	[this document]
CS4-Admit	100001	[this document]
CS3-Admit	011001	[this document]
MC-Admit	100101	[this document]
MC	011001	[this document]

7. Security Considerations

The Security Considerations are identical to those of RFC 5865.

Every newly proposed DSCP (save A/V-Sig) serves the same security risk and properties of the Voice-Admit DSCP. Section 3 of this document discusses why these DSCP values are should be parallel to their non-admitted counterparts, just as Voice-Admit states in RFC 5865 it is parallel to the existing (at the time) EF.

The A/V-Sig merely has a new DSCP name, RFC 4594 currently has this service class called "Signaling", serving the same purpose.

8. References

8.1 Normative References

- [ID-4594-UP] J. Polk, "Standard Configuration of DiffServ Service Classes", "work in progress", February 2013
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2205] R. Braden, Ed., L. Zhang, S. Berson, S. Herzog, S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, September 1997
- [RFC2474] K. Nichols, S. Blake, F. Baker, D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers ", RFC 2474, December 1998
- [RFC5865] F. Baker, J. Polk, M. Dolly, "A Differentiated Services Code Point (DSCP) for Capacity-Admitted Traffic", RFC 5865, May 2010

8.2 Informative References

- [RFC3550] Schulzrinne, H., Casner, S., Frederick, R., and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications", STD 64, RFC 3550, July 2003.
- [RFC4080] R. Hancock, G. Karagiannis, J. Loughney, S. Van den Bosch, "Next Steps in Signaling (NSIS): Framework", RFC 4080, June 2005
- [RFC4594] J. Babiarez, K. Chan, F Baker, "Configuration Guidelines for Diffserv Service Classes", RFC 4594, August 2006

Author's Addresses

James Polk
3913 Treemont Circle
Colleyville, Texas 76034

Phone: +1.817.271.3552
Email: jmpolk@cisco.com

Network WG
Internet-Draft
Intended status: Standards Track (PS)
Obsoletes: RFC 4594
Updates: RFC 5865
Expires: August 25, 2013

James Polk, ed.
Cisco
Feb, 2013

Standard Configuration of DiffServ Service Classes
draft-polk-tsvwg-rfc4594-update-03.txt

Abstract

This document describes service classes configured with DiffServ and identifies how they are used and how to construct them using Differentiated Services Code Points (DSCPs), traffic conditioners, Per-Hop Behaviors (PHBs), and Active Queue Management (AQM) mechanisms. There is no intrinsic requirement that particular DSCPs, traffic conditioners, PHBs, and AQM be used for a certain service class, but for consistent behavior under the same network conditions, configuring networks as described here is appropriate.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on August 25, 2013.

Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in

Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
1.1. Requirements Notation	
1.2. Expected Use in the Network	
1.3. Service Class Definition	
1.4. Key Differentiated Services Concepts	
1.4.1. Queuing	
1.4.1.1. Priority Queuing	
1.4.1.2. Rate Queuing	
1.4.2. Active Queue Management	
1.4.3. Traffic Conditioning	
1.4.4. Differentiated Services Code Point (DSCP)	
1.4.5. Per-Hop Behavior (PHB)	
1.5. Key Service Concepts	
1.5.1. Default Forwarding (DF)	
1.5.2. Assured Forwarding (AF)	
1.5.3. Expedited Forwarding (EF)	1
1.5.4. Class Selector (CS)	1
1.5.5. Admission Control	1
1.6 What Changes are Proposed Here from RFC 4594?.....	1
2. Service Differentiation	1
2.1. Service Classes	1
2.2. Categorization of User Oriented Service Classes	1
2.3. Service Class Characteristics	1
2.4. Service Classes vs. Treatment Aggregate (from RFC 5127)...	2
2.4.1 Examples of Service Classes in Treatment Aggregates...	2
3. Network Control Traffic	2
3.1. Current Practice in the Internet	2
3.2. Network Control Service Class	2
3.3. OAM Service Class	2
4. User Oriented Traffic	3
4.1. Conversational Service Class Group	3
4.1.1 Audio Service Class	3
4.1.2 Video Service Class	3
4.1.3 Hi-Res Service Class	3
4.2. Realtime-Interactive Service Class	3
4.3. Multimedia Conferencing Service Class	3
4.4. Multimedia Streaming Service Class	3
4.5. Broadcast Video Service Class	4
4.6. Low-Latency Data Service Class	4
4.7. Conversational Signaling Service Class	4
4.8. High-Throughput Data Service Class	4
4.9. Standard Service Class	4
4.10. Low-Priority Data	4
5. Additional Information on Service Class Usage	4
5.1. Mapping for NTP	5

5.2. VPN Service Mapping	5
6. Security Considerations	5
7. Contributing Authors	5
8. Acknowledgements	5
9. References	5
9.1. Normative References	5
9.2. Informative References	5
Author's Address	5
Appendix A - Changes	5

1. Introduction

Differentiated Services [RFC2474][RFC2475] provides the ability to mark/label/classify IP packets differently to distinguish how individual packets need to be treated differently through (or throughout) a network on a per hop basis. Local administrators are who configure each router for which Differentiated Services Code Points (DSCP) are to be treated differently, which are to be ignored (i.e., no differentiated treatment), and which DSCPs are to have their packets remarked (to different DSCPs) as they pass through a router. Local administrators are also who assign which applications, or traffic types, should use which DSCPs to receive the treatment the administrators expect within their network.

What most people fail to understand is that DSCPs provide a per hop behavior (PHB) through that router, but not the previous or next router. In this way of understanding PHB markings, one can understand that Differentiated Services (DiffServ) is not a Quality of Service (QoS) mechanism, but rather a Classification of Service (CoS) mechanism.

For instance, there are 64 possible DSCP values, i.e., using 6 bits of the old Type of Service (TOS) byte [RFC0791]. Each can be configured locally to have greater or less treatment relative to any other DSCP with two exceptions*.

- * Expedited Forwarding (EF) [RFC3246] DSCPs have a treatment requirement that any packet marked within an EF class has to be the next packet transmitted out its egress interface. If there are more than one EF marked packet in the queue, obviously the queue sets the order they are transmitted. Further, if there are more than one EF DSCP, local configuration determines if each are treated the same or differently relate to each other EF DSCP. Currently, there are two Expedited Forwarding DSCPs: EF (101110) [RFC3246] and VOICE-ADMIT (101100) [RFC5865].

- * Class Selector 6 (CS6) [RFC2474] is for routing protocol traffic. There are deemed important because if the network does not transmit and receive its routing protocol traffic in a timely manner, the network stops operating properly.

Not all are configured to mean anything other than best effort forwarding by local administrators of a network. Let us say there are 5 DSCPs configured within network A. Network A's administrator chooses and configures which order (obeying the two exceptions noted above) which application packets are treated differently than any other packets within that network (A). The DSCPs are not fixed to a linear order for relative priority on a per hop basis. Further, and this is often the case, there might be packets with the same DSCP arriving at multiple interfaces of a node, each egressing that node out the same interface. At ingress to this node, everything was fine, with no poor behavior or noticeably excessive amount of packets with the same DSCP. However, at the egress interface, there might not be enough capacity to satisfy the load, thus the departing packets transmit at their maximum rate for that DSCP, but have additional latency due to the overload within that one node. This is called fan-in congestion (or problem). By itself, DiffServ will not remedy this problem for the application that is intolerant to added latency because DiffServ only functions within 1 node at a time.

An additional mechanism is needed to ensure each flow or session receives the amount of packets at its destination that the application requires to perform properly; a mechanism such as IntServ, by way of RSVP [RFC2205] or NSIS [RFC4080]. With this added capability to be session aware, something DiffServ is not, the packets transmitted within a single session have a very good probability of arriving in such a way the receiving application can make full use of each. That said, signaling reservations for each session or flow adds complexity, which creates more work for those who maintain and administer such a network. Adding bandwidth and using DiffServ marking is an easier pill to swallow. The deployment of not few, but more and more audio and (particularly bandwidth hogging) video codecs and their respective application rigidity has caused some to conclude that throwing bandwidth at the problem is no longer acceptable.

With this in mind, this document incorporates five of the six new DSCPs from [ID-DSCP] identified as capacity-admitted DSCPs for most of the service classes in this document. As explained in [ID-DSCP], the five new capacity-admitted DSCPs are from Pool 3. [ID-DSCP] goes further to explain that many layer 2 technologies use fewer bits for marking and prioritization. Instead of six bits like DiffServ, they have three bits, which yields a maximum of 8 values, which tend to line up quite well with the TOS field values. Thus, aggregation of DSCPs is typically accomplished by simply ignoring or reducing the number of bits used to the most significant ones available, such as

EF is 101110, at layer 2 this is merely 101;

Broadcast is 011000, at layer 2 this is merely 011.

However, that was not a premise DiffServ was built upon, to merely

reduce the number of bits. In other words, within DiffServ, XXX is not the same as XXX000 (where XXX is the same binary value in both cases).

This document is originally built upon the RFC 4594 effort, while updating some of the usages and expanding the scope for newer applications that are in use today. The idea in RFC 4594 remains true here, to define a set of service classes, each having unique traffic characteristics, and assigning one or more DSCPs to each service class. As much as the focus could be on the DSCP values, it is not. The focus of this document is the unique traffic characteristics of each service class.

There are many services classes defined in this document, not all will be used in each network at any period of time. This consistency packet markings we talk about is for several reasons, including in a network that does not currently implement a certain service class because they do not have that type of traffic in their network, or that the network merely gives that traffic best effort service. Having a solid guideline to know where to progress or reconfigure a network and endpoints to, say from best effort for a particular traffic type, is a very good thing to do more uniformly than not. A fair amount of burden is placed at DS boundaries needing to keep up with which markings turn into which other markings at both ingress and egress to a network. The same holds true for application developers choosing a default DSCP for their application, lacking a guideline means everyone picks for themselves - and usually with a highly inflated sense of self importance for their application or service.

Another point to make is that there are 20+ service classes defined within the IETF, and that is far too many for most service providers to manage effectively. So, they have formed groups around certain aggregation solutions of service classes. One such aggregation group is based on RFC 5127, which defines what it calls a treatment aggregate, which is taking RFC 4594's service classes and placing them each into one of four treatment aggregates for service providers to handle as a group. SG12 within the ITU-T has an alternative that has nine aggregate groups, so there is work to be done to harmonize aggregates of service classes. This discussion is articulated more in section 2.4. At the end of Section 2.4 we have introduced a series of example configurations which provide examples of how only a few service classes - yet still most treatment aggregates - can be configured in example networks.

Does RFC 4594 need updating? That document is an informational guideline on how networks can or should mark certain packet flows with differing traffic characteristics using DiffServ. There are several reasons why this informational RFC lacks the necessary clarity and strength to reach widespread adoption:

- o confusion between RFC 4594 and RFC 5127 [RFC5127], the latter of

which is for aggregating many 6-bit DSCP values into a 3-bit (8 value) field used specifically by service provider (SP) networks.

- o some believe both RFCs are for SPs, while others ignore RFC 5127 and use RFC 4594 as if it were standards track or BCP.
- o some believe RFC 5127 is for SPs only, and want RFC 4594 to reduce the number of DSCPs within its guidelines to recommend using only 3 or 4 DSCPs. This seems to stem from a manageability and operational perspective.
- o some know RFC 4594 is informational and do not follow its guidelines specifically because it is informational.
- o some use DSCP values that are not defined within RFC 4594, making mapping between different networks using similar or identical application flows difficult.
- o some believe enterprise networks should not use either RFC except at the edge of their networks, where they directly connect to SP networks.
- o some argue that the services classes guidance per class is too broad and are therefore not sure in which service class a particular application is to reside.
- o time has shown that video has become a dominant application on the Internet, and many believe it now requires to be treated uniquely in environments that want to. Video also does not always plan nice with audio, so knowing the two use the same transport (RTP) [RFC3550], a means of separation is in order.

Service class definitions are based on the different traffic characteristics and required performance of the applications/services. There are a greater number of service classes in this document than there were when RFC 4594 [RFC4594] was published (the RFC this document intends to obsolete). The required performance of applications/services has also changed since the publication of RFC 4594, specifically in the area of conversational real time communications. As a result, this document has a greater number of real time applications with more granular set of DSCPs due to their different required performances. Like RFC 4594 before, this approach allows those applications with similar traffic characteristics and performance requirements to be placed in the same service class.

The notion of traffic characteristics and required performance is a per application concept, therefore the label name of each service class remains the same on an end-to-end basis, even if we understand that DiffServ is only a PHB and cannot guarantee anything, even packet delivery at the intended destination node. That said, several applications can be configured to have the same DSCP, or

each have different DSCPs that have the same treatment per hop within a network.

Since RFC 4594 was first published, a new concept has been introduced that will appear throughout this document, including DSCP assignments -- the idea of "admitted" traffic, initially introduced into DiffServ within RFC 5865 [RFC5865]. The VOICE-ADMIT Expedited Forwarding class differentiates itself from the EF Expedited Forwarding by having the packets marked be for admitted traffic. This concept of "admitted" traffic is spread throughout the real time traffic classes.

Thus, the document flow is as follows:

- o maintain the general format of RFC 4594;
- o augment the content with the concept of capacity-admission;
- o incorporate more video into this document, as it has become a dominant application in enterprises and other managed networks, as well as on the open public Internet;
- o reduce the discussion on voice and its examples;
- o articulate the subtle differences learned since RFC 4594 was published.

The goal here is to provide a standard configuration for DiffServ DSCP assignments and expected PHBs for enterprises and other managed networks, as well as towards the public Internet with specific traffic characteristics per Service class/DSCP, and example applications shown for each.

This document describes service classes configured with DiffServ and defines how they can be used and how to construct them using Differentiated Services Code Points (DSCPs), and recommends how to construct them using traffic conditioners, Per-Hop Behaviors (PHBs), and Active Queue Management (AQM) mechanisms. There is no intrinsic requirement that particular traffic conditioners, PHBs, and AQM be used for a certain service class, but as a policy and for interoperability it is useful to apply them consistently.

We differentiate services and their characteristics in Section 2. Network control traffic, as well as user oriented traffic are discussed in Sections 3 and 4, respectively. We analyze the security considerations in Section 6. Section 7 offers a tribute to the authors of RFC 4594, from which this document is based. It is in its own section, and not part of the normal acknowledgements portion of each IETF document.

1.1. Requirements Notation

The key words "SHOULD", "SHOULD NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119] when they appear in ALL CAPS. These words may also appear in this document in lower case as plain English words, absent their normative meanings.

1.2. Expected Use in the Network

In the Internet today, corporate LANs and ISP WANs are increasingly utilized, to the point in which network congestion is affecting performance of applications. For this reason, congestion, loss, and variation in delay within corporate LANs and ISP backbones is becoming known to the users collectively as "the network is slow for this application" or just "right now" or "for today". Users do not directly detect network congestion. They react to applications that run slow, or to downloads that take too long in their mind(s). The explosion of video traffic on the internet recently has cause much of this, and is often the application the user is using when they have this slowness.

In the past, application slowness occurred for three very good reasons.

- o the networks the user oriented traffic traverses moves through cycles of bandwidth boom and bandwidth bust, the latter of which become apparent with the periodic deployment of new bandwidth-hungry applications.
- o In access networks, the state is often different. This may be because throughput rates are artificially limited or over-subscribed, or because of access network design trade-offs.
- o Other characteristics, such as database design on web servers (that may create contention points, e.g., in filestore) and configuration of firewalls and routers, often look externally like a bandwidth limitation.

The intent of this document is to provide a standardized marking, plus a conditioning and packet treatment strategy so that it can be configured and put into service on any link that is itself congested.

1.3. Service Class Definition

A "service class" represents a similar set of traffic characteristics for delay, loss, and jitter as packets traverse routers in a network. For example, "High-Throughput Data" service class for store-and-forward applications, or a "Broadcast" service

class for minimally time-shifted IPTV or Internet radio broadcasts. Such a service class may be defined locally in a Differentiated Services (DS) domain, or across multiple DS domains, possibly extending end to end. A goal of this document is to have most/all networks assign the same type of traffic the same for consistency.

A service class is a naming convention which is defined as a word, phrase or initialism/acronym representing a set of necessary traffic characteristics of a certain type of data flow. The necessary characteristics of these traffic flows can be realized by the use of defined per-hop behavior that started with [RFC2474]. The actual specification of the expected treatment of a traffic aggregate within a domain may also be defined as a per-domain behavior (PDB) [RFC3086].

Each domain will locally choose to

- o implement one or more service classes with traffic characteristics as defined here, or
- o implement one or more service classes with similar traffic characteristics as defined here, or
- o implement one or more service classes with similar traffic characteristics as defined here and to aggregate one or more service classes to reduce the number of unique DSCPs within their network, or
- o implement one or more non-standard service classes with traffic characteristics not as defined here, or
- o not use DiffServ within their domain.

For example, low delay, low loss, and minimal jitter may be realized using the EF PHB, or with an over-provisioned AF PHB. This must be done with care as it may disrupt the end-to-end performance required by the applications/services. If the packet sizes are similar within an application, but different between two applications, say small voice packets and large video packets, these two applications may not realize optimum results if merged into the same aggregate if there are any bottlenecks in the network. We provide for this flexibility on a per hop or per domain basis within this document.

This document provides standardized markings for traffic with similar characteristics, and usage expectations for PHBs for specific service classes for their consistent implementation.

The Default Forwarding "Standard" service class is REQUIRED; all other service classes are OPTIONAL. That said, each service class lists traffic characteristics that are expected when using that type of traffic. It is RECOMMENDED that applications and protocols that fit a certain traffic characteristic use the appropriate service

class mark, i.e., the DSCP, for consistent behavior. It is expected that network administrators will base their endpoint application and router configuration choices on the level of service differentiation they require to meet the needs of their customers (i.e., their end-users).

1.4. Key Differentiated Services Concepts

In order to fully understand this document, a reader needs to familiarize themselves with the principles of the Differentiated Services Architecture [RFC2474]. We summarize some key concepts here only to provide convenience for the reader, the referenced RFCs providing the authoritative definitions.

1.4.1. Queuing

A queue is a data structure that holds packets that are awaiting transmission. A router interface can only transmit one packet at a time, however fast the interface speed is. If there is only 1 queue at an interface, the packets are transmitted in the order they are received into that queue - called FIFO, or "first in, first out". Sometimes there is a lag in the time between a packets arrives in the queue and when it is transmitted. This delay might be due to lack of bandwidth, or if there are multiple queues on that interface, because a packet is low in priority relative to other packets that are awaiting to transmit. The scheduler is the system entity that chooses which packet is next in line for transmission when more than one packet are awaiting transmission out the same router interface.

1.4.1.1 Priority Queuing

A priority queuing system is a combination of a set of queues and a scheduler that empties the queues (of packets) in priority sequence. When asked for a packet, the scheduler inspects the highest priority queue and, if there is data present, returns a packet from that queue. Failing that, it inspects the next highest priority queue, and so on. A freeway onramp with a stoplight for one lane that allows vehicles in the high-occupancy-vehicle lane to pass is an example of a priority queuing system; the high-occupancy-vehicle lane represents the "queue" having priority.

In a priority queuing system, a packet in the highest priority queue will experience a readily calculated delay. This is proportional to the amount of data remaining to be serialized when the packet arrived plus the volume of the data already queued ahead of it in the same queue. The technical reason for using a priority queue relates exactly to this fact: it limits delay and variations in delay and should be used for traffic that has that requirement.

A priority queue or queuing system needs to avoid starvation of lower-priority queues. This may be achieved through a variety of means, such as admission control, rate control, or network engineering.

1.4.1.2. Rate Queuing

Similarly, a rate-based queuing system is a combination of a set of queues and a scheduler that empties each at a specified rate. An example of a rate-based queuing system is a road intersection with a stoplight. The stoplight acts as a scheduler, giving each lane a certain opportunity to pass traffic through the intersection.

In a rate-based queuing system, such as Weighted Fair Queuing (WFQ) or Weighted Round Robin (WRR), the delay that a packet in any given queue will experience depends on the parameters and occupancy of its queue and the parameters and occupancy of the queues it is competing with. A queue whose traffic arrival rate is much less than the rate at which it lets traffic depart will tend to be empty, and packets in it will experience nominal delays. A queue whose traffic arrival rate approximates or exceeds its departure rate will tend not to be empty, and packets in it will experience greater delay. Such a scheduler can impose a minimum rate, a maximum rate, or both, on any queue it touches.

1.4.2 Active Queue Management

Active Queue Management, or AQM, is a generic name for any of a variety of procedures that use packet dropping or marking to manage the depth of a queue. The canonical example of such a procedure is Random Early Detection (RED), in that a queue is assigned a minimum and maximum threshold, and the queuing algorithm maintains a moving average of the queue depth. While the mean queue depth exceeds the maximum threshold, all arriving traffic is dropped. While the mean queue depth exceeds the minimum threshold but not the maximum threshold, a randomly selected subset of arriving traffic is marked or dropped. This marking or dropping of traffic is intended to communicate with the sending system, causing its congestion avoidance algorithms to kick in. As a result of this behavior, it is reasonable to expect that TCP's cyclic behavior is desynchronized and that the mean queue depth (and therefore delay) should normally approximate the minimum threshold.

A variation of the algorithm is applied in Assured Forwarding PHB [RFC2597], in that the behavior aggregate consists of traffic with multiple DSCP marks, which are intermingled in a common queue. Different minima and maxima are configured for the several DSCPs separately, such that traffic that exceeds a stated rate at ingress is more likely to be dropped or marked than traffic that is within its contracted rate.

1.4.3 Traffic Conditioning

In addition, at the first router in a network that a packet crosses, arriving traffic may be measured and dropped or marked according to a policy, or perhaps shaped on network ingress, as in "A Rate Adaptive Shaper for Differentiated Services" [RFC2963]. This may be used to bias feedback loops, as is done in "Assured Forwarding PHB" [RFC2597], or to limit the amount of traffic in a system, as is done in "Expedited Forwarding PHB" [RFC3246]. Such measurement procedures are collectively referred to as "traffic conditioners". Traffic conditioners are normally built using token bucket meters, for example with a committed rate and burst size, as in Section 1.5.3 of the DiffServ Model [RFC3290]. The Assured Forwarding PHB [RFC2597] uses a variation on a meter with multiple rate and burst size measurements to test and identify multiple levels of conformance.

Multiple rates and burst sizes can be realized using multiple levels of token buckets or more complex token buckets; these are implementation details. The following are some traffic conditioners that may be used in deployment of differentiated services:

- o For Class Selector (CS) PHBs, a single token bucket meter to provide a rate plus burst size control.
- o For Expedited Forwarding (EF) PHB, a single token bucket meter to provide a rate plus burst size control.
- o For Assured Forwarding (AF) PHBs, usually two token bucket meters configured to provide behavior as outlined in "Two Rate Three Color Marker (trTCM)" [RFC2698] or "Single Rate Three Color Marker (srTCM)" [RFC2697]. The two-rate, three-color marker is used to enforce two rates, whereas the single-rate, three-color marker is used to enforce a committed rate with two burst lengths.

1.4.4 Differentiated Services Code Point (DSCP)

The DSCP is a number in the range 0..63 that is placed into an IP packet to mark it according to the class of traffic it belongs in. These are divided into 3 groups, or pools, defined in RFC 2474, arranged as follows:

- o Pool-1 has 32 values designated for standards assignment (of the form 'xxxxx0').
- o Pool-2 has 16 values designated for experimental or local use only (EXP/LU) assignment (of the form 'xxxx11').
- o Pool-3 has 16 values designated for experimental or local use (EXP/LU) assignment (of the form 'xxxx01').

However, pool-3 is allowed to be assigned for one of two reasons,

#1 - if the values in pool-1 are exhausted, or

#2 - if there is a justifiable reason for assigning a pool-3 DSCP prior to pool-1's exhaustion.

1.4.5 Per-Hop Behavior (PHB)

In the end, the mechanisms described above are combined to form a specified set of characteristics for handling different kinds of traffic, depending on the needs of the application. This document seeks to identify useful traffic aggregates and to specify what PHB should be applied to them.

1.5 Key Service Concepts

While Differentiated Services is a general architecture that may be used to implement a variety of services, three fundamental forwarding behaviors have been defined and characterized for general use. These are basic Default Forwarding (DF) behavior for elastic traffic, the Assured Forwarding (AF) behavior, and the Expedited Forwarding (EF) behavior for real-time (inelastic) traffic. The facts that four code points are recommended for AF and that one code point is recommended for EF are arbitrary choices, and the architecture allows any reasonable number of AF and EF classes simultaneously. The choice of four AF classes and one EF class in the current document is also arbitrary, and operators MAY choose to operate more or fewer of either.

The terms "elastic" and "real-time" are defined in [RFC1633], Section 3.1, as a way of understanding broad-brush application requirements. This document should be reviewed to obtain a broad understanding of the issues in quality of service, just as [RFC2475] should be reviewed to understand the data plane architecture used in today's Internet.

1.5.1 Default Forwarding (DF)

The basic forwarding behaviors applied to any class of traffic are those described in [RFC2474] and [RFC2309]. Best-effort service may be summarized as "I will accept your packets" and is typically configured with some bandwidth guarantee. Packets in transit may be lost, reordered, duplicated, or delayed at random. Generally, networks are engineered to limit this behavior, but changing traffic loads can push any network into such a state.

Application traffic in the internet that uses default forwarding is expected to be "elastic" in nature. By this, we mean that the sender of traffic will adjust its transmission rate in response to

changes in available rate, loss, or delay.

For the basic best-effort service, a single DSCP value is provided to identify the traffic, a queue to store it, and active queue management to protect the network from it and to limit delays.

1.5.2 Assured Forwarding (AF)

The Assured Forwarding PHB [RFC2597] behavior is explicitly modeled on Frame Relay's Discard Eligible (DE) flag or ATM's Cell Loss Priority (CLP) capability. It is intended for networks that offer average-rate Service Level Agreements (SLAs) (as FR and ATM networks do). This is an enhanced best-effort service; traffic is expected to be "elastic" in nature. The receiver will detect loss or variation in delay in the network and provide feedback such that the sender adjusts its transmission rate to approximate available capacity.

For such behaviors, multiple DSCP values are provided (two or three, perhaps more using local values) to identify the traffic, a common queue to store the aggregate, and active queue management to protect the network from it and to limit delays. Traffic is metered as it enters the network, and traffic is variously marked depending on the arrival rate of the aggregate. The premise is that it is normal for users occasionally to use more capacity than their contract stipulates, perhaps up to some bound. However, if traffic should be marked or lost to manage the queue, this excess traffic will be marked or lost first.

1.5.3. Expedited Forwarding (EF)

The intent of Expedited Forwarding PHB [RFC3246] is to provide a building block for low-loss, low-delay, and low-jitter services. It can be used to build an enhanced best-effort service: traffic remains subject to loss due to line errors and reordering during routing changes. However, using queuing techniques, the probability of delay or variation in delay is minimized. For this reason, it is generally used to carry voice and for transport of data information that requires "wire like" behavior through the IP network. Voice is an inelastic "real-time" application that sends packets at the rate the codec produces them, regardless of availability of capacity. As such, this service has the potential to disrupt or congest a network if not controlled. It also has the potential for abuse.

To protect the network, at minimum one SHOULD police traffic at various points to ensure that the design of a queue is not overrun, and then the traffic SHOULD be given a low-delay queue (often using priority, although it is asserted that a rate-based queue can do this) to ensure that variation in delay is not an issue, to meet application needs.

1.5.4 Class Selector (CS)

Class Selector, those DSCPs that end in zeros (xxx000), provide support for historical codepoint definitions and PHB requirement. The CS fields provide a limited backward compatibility with legacy practice, as described in [RFC2474], Section 4. Backward compatibility is addressed in two ways,

- First, there are per-hop behaviors that are already in widespread use (e.g., those satisfying the IPv4 Precedence queuing requirements specified in [RFC1812]), and
- this document will continue to permit their use in DS-compliant networks.

In addition, there are some DSCPs that correspond to historical use of the IP Precedence field,

- CS0 (000000) will remain 'Default Forwarding' (also known as 'Best Effort')
- 11xxxx will remain for routing traffic

and will map to PHBs that meet the general requirements specified in [RFC2474], Section 4.2.2.2.

No attempt is made to maintain backward compatibility with the "DTR" or Type of Service (TOS) bits of the IPv4 TOS octet, as defined in [RFC0791] and [RFC1349].

A DS-compliant network can be deployed exclusively by using one or more CS-compliant PHB groups. Thus, for example, codepoint '011000' would map to the same PHB as codepoint '011010'.

1.5.5 Admission Control

Admission control (including refusal when policy thresholds are crossed) can ensure high-quality communication by ensuring the availability of bandwidth to carry a load. Inelastic real-time flows such as Voice over Internet Protocol (VoIP) (audio) or video conferencing services can benefit from use of an admission control mechanism, as generally the audio or video service is configured with over-subscription, meaning that some users may not be able to make a call during peak periods.

For VoIP (audio) service, a common approach is to use signaling protocols such as SIP, H.323, H.248, MEGACO, along with Resource Reservation Protocol (RSVP) to negotiate admittance and use of network transport capabilities. When a user has been authorized to send voice traffic, this admission procedure has verified that data rates will be within the capacity of the network that it will use.

Many RTP voice and video payloads are inelastic and cannot react to loss or delay in any substantive way. For these payload types, the network needs to police at ingress to ensure that the voice traffic stays within its negotiated bounds. Having thus assured a predictable input rate, the network may use a priority queue to ensure nominal delay and variation in delay.

1.5.5.1 Capacity Admitted (*-Admit)

This is a newer group of traffic types that started with RFC 5865 and the Voice-Admit service type. Voice-Admit is an EF class marking but has capacity-admission always applied to it to ensure each of these flows are managed through a network, though not necessarily on an end-to-end basis. This depends on how many networks each flow transits and the load on each transited network. There are a series of new DSCPs proposed in [ID-DSCP], each specifying unique characteristics necessitating a separate marking from what existing before that document.

This document will import in four new '*-Admit' DSCPs from [ID-DSCP], 2 others that are new but not capacity-admitted, one from RFC 5865, and change the existing usage of 2 DSCPs from RFC 4594. This is discussed throughout the rest of this document.

1.6 What Changes are Proposed Here from RFC 4594?

Changing an entire network DiffServ configuration has proven to be a painful experience for both individuals and companies. It is not done very often, and for good reason. This effort is based on experience learned since the publication of RFC 4594 (circa 2006). Audio, once thought to be ok grouped with video, needs to be in separate service classes. Collaboration has taken off, mostly because of mobility, but also because of a worldwide recession that has limited physical travel, and relying on people to do more with their computers. With that in mind, there has been an explosion in application development for the individual (seems everyone has an "app-store"). The following set of bullets has this world - that needs a robust layer 3 - in mind.

- o Scope of document is changed to tighten it up for standards track consideration.
- o This document explicitly states there is a fundamental requirement that a particular DSCP(s) be used for each service class, each with a recommended set of applications to be used by that service class - at least on that individual's externally facing (public) interface.
- o Created the Conversational group of service classes to focus on realtime, mostly bidirectional communications (unless multicast is

used).

- o "Realtime-Interactive"
Moved to (near) realtime TCP-based apps

Why the change? TCP based transports have proven, in certain environments, to be a bidirectional realtime transport, e.g., for multiplayer gaming and virtual desktops applications.

- o "Audio"
Same as Telephony (which is now gone), adds Voice-Admit for capacity-admitted traffic

Why the change? RFC 5865 (Voice-Admit) needed to be added to the Audio service class. Video needed to be separate from audio, hence the name change from Telephony (which includes video) to just audio.

- o "Video"
NEW for video and audio/video conferencing, was in Multimedia-Conferencing service classification

Why the change? Many networks are using the AF4X for video, but others are throwing anything "multimedia" into the same service class (like elastic TCP flows). Video has become so dominant that it should be what mostly goes into one service class.

- o "Hi-Res"
NEW for video and audio/video conferencing

Why the change? This entirely new service class is for local policy based higher end video (think Telepresence). Without congestion, this service class has the same treatment as Video, but if there is any pushback from the network, Hi-Res (note: not married to the name) has a better PHB.

- o "Multimedia-Conferencing"
Now without audio or human video

Why the change? The change is taking bidirectional human audio and video out of this service class. This is all about non-realtime collaboration - even in conjunction with an audio and/or video flow.

- o "Broadcast"
Remains the same, added CS3-Admit for capacity-admitted

Why the change? Removing the "-Video" from the name because there are so many more flows that are Broadcast in realtime than video.

- o "Low-Latency Data"
Remains the same, adds IM & Presence traffic explicitly

Why the change? Merely explicitly stating a place for some

additional traffic types that otherwise could go elsewhere.

- o "Conversational Signaling" (A/V-Sig)
Was 'Signaling'

Why the change? This change is merely a renaming of a service class, and acknowledgement that some of the previous authors inaccurate beliefs that DSCPs were linearly ordered with those values having a higher value definitely getting better treatment than lower values.

2. Service Differentiation

There are practical limits on the level of service differentiation that should be offered in the IP networks. We believe we have defined a practical approach in delivering service differentiation by defining different service classes that networks may choose to support in order to provide the appropriate level of behaviors and performance needed by current and future applications and services. The defined structure for providing services allows several applications having similar traffic characteristics and performance requirements to be grouped into the same service class. This approach provides a lot of flexibility in providing the appropriate level of service differentiation for current and new, yet unknown applications without introducing significant changes to routers or network configurations when a new traffic type is added to the network.

2.1 Service Classes

Traffic flowing in a network can be classified in many different ways. We have chosen to divide it into two groupings, network control and user/subscriber traffic. To provide service differentiation, different service classes are defined in each grouping. The network control traffic group can further be divided into two service classes (see Section 3 for detailed definition of each service class):

- o "Network Control" for routing and network control function.
- o "OAM" (Operations, Administration, and Management) for network configuration and management functions.

The user/subscriber traffic group is broken down into ten service classes to provide service differentiation for all the different types of applications/services (see Section 4 for detailed definition of each service class):

- o Conversational service group consists of three service classes:
 - Audio, which includes both 'admitted' and 'unadmitted' audio

service classes, is for non-one way (i.e., generally bidirectional) audio media packets between human users of smaller size and at a constant delivery rate.

- Hi-Res Video, which includes both 'admitted' and 'unadmitted' Hi-Res Video service classes, is for video traffic from higher end endpoints between human users necessitating different treatment than from desktop or video phone endpoints. This has a clearly business differentiation, and not a technical differentiation - as both Hi-Res-Video and Video will be treated similarly on the wire when no congestion occurs.
- Video, which includes both 'admitted' and 'unadmitted' video service classes, is for video traffic from lower end endpoints between human users necessitating different treatment than from higher end (i.e., Telepresence) endpoints. This has a clearly business differentiation, and not a technical differentiation - as both Hi-Res-Video and Video will be treated similarly on the wire when no congestion occurs.
- o Conversational Signaling service class is for peer-to-peer and client-server signaling and control functions using protocols such as SIP, H.323, H.248, and Media Gateway Control Protocol (MGCP). This traffic needs to not be starved on the network.

Editor's note: RFC 4594 had this DSCP marking as CS5, but with clearly different characteristics (i.e., no sensitivity to jitter or (unreasonable) delay), this DSCP has been moved to a more appropriate (new) value, defined in [ID-DSCP].

- o Real-Time Interactive, which includes both 'admitted' and 'unadmitted' Realtime-Interactive service class, is for bidirectional variable rate inelastic applications that require low jitter and loss and very low delay, such as interactive gaming applications that use RTP/UDP streams for game control commands, and Virtualized Desktop applications between the user and content source, typically in a centralized data center.
- o Multimedia Conferencing, which includes both 'admitted' and 'unadmitted' multimedia conferencing service class, is for applications that require minimal delay, but not like those of realtime application requirements. This service class can be bursty in nature, as well as not transmit packets for some time. Applications such as presentation data or collaborative application sharing will use this service class.
- o Multimedia Streaming, which includes both 'admitted' and 'unadmitted' multimedia streaming service class, is for one-way bufferable streaming media applications such as Video on Demand (VOD) and webcasts.

- o Broadcast, which includes both 'admitted' and 'unadmitted' broadcast service class, is for inelastic streaming media applications that may be of constant or variable rate, requiring low jitter and very low packet loss, such as broadcast TV and live events, video surveillance, and security.
- o Low-Latency Data service class is for data processing applications such as client/server interactions or Instant Messaging (IM) and Presence data.
- o Conversational Signaling (A/V-Sig) service class is for all signaling messages, whether in-band (i.e., along the data path) or out-of-band (separate from the data path), for the purposes of setting up, maintaining, managing and terminating bi- or multi-directional realtime sessions.
- o High-Throughput Data service class is for store and forward applications such as FTP and billing record transfer.
- o Standard service class, commonly called best effort (BE), is for traffic that has not been identified as requiring differentiated treatment.
- o Low-Priority Data service class, which some could call the scavenger class, is for packet flows where bandwidth assurance is not required.

2.2 Categorization of User Oriented Service Classes

The ten defined user/subscriber service classes listed above can be grouped into a small number of application categories. For some application categories, it was felt that more than one service class was needed to provide service differentiation within that category due to the different traffic characteristic of the applications, control function, and the required flow behavior. Figure 1 provides a summary of service class grouping into four application categories.

Application Control Category

- o The Conversational Signaling service class is intended to be used to control applications or user endpoints. Examples of protocols that would use this service class are SIP, XMPP or H.323 for voice and/or video over IP services. User signaling flows have similar performance requirements as Low-Latency Data, they require a separate DSCP to be distinguished other traffic and allow for a treatment that is unique.

Media-Oriented Category

Due to the vast number of new (in process of being deployed) and already-in-use media-oriented services in IP networks, seven service

classes have been defined.

- o Audio service class is intended for Voice-over-IP (VoIP) services. It may also be used for other applications that meet the defined traffic characteristics and performance requirements.
- o Video service class is intended for Video over IP services. It may also be used for other applications that meet the defined traffic characteristics and performance requirements.
- o Hi-Res service class is intended for higher end video services that have the same traffic characteristics as the video service class, but have a business requirement(s) to be treated differently. One example of this is Telepresence video applications.
- o Realtime-Interactive service class is intended for inelastic applications such as desktop virtualization applications and for interactive gaming.
- o Multimedia Conferencing service class is for everything about or within video conferencing solutions that does not include the voice or (human) video components. Several examples are
 - the presentation data part of an IP conference (call).
 - the application sharing part of an IP conference (call).
 - the whiteboarding aspect of an IP conference (call).

Each of the above can be part of a lower end web-conferencing application or part of a higher end Telepresence video conference. Each also has the ability to reduce their transmission rate on detection of congestion. These flows can therefore be classified as rate adaptive and most often more elastic than their voice and video counterparts.

- o Broadcast Video service class is to be used for inelastic traffic flows specifically with minimal buffering expected by the source or destination, which are intended for broadcast HDTV service, as well as for transport of live video (sports or concerts) and audio events.
- o Multimedia Streaming service class is to be used for elastic multimedia traffic flows where buffering is expected. This is the fundamental difference between the Broadcast and multimedia streaming service classes. Multimedia streaming content is typically stored before being transmitted. It is also buffered at the receiving end before being played out. The buffering is sufficiently large to accommodate any variation in transmission rate that is encountered in the network. Multimedia entertainment over IP delivery services that are being developed

can generate both elastic and inelastic traffic flows; therefore, two service classes are defined to address this space, respectively: Multimedia Streaming and Broadcast Video.

Data Category

The data category is divided into three service classes.

- o Low-Latency Data for applications/services that require low delay or latency for bursty but short-lived flows.
- o High-Throughput Data for applications/services that require good throughput for long-lived bursty flows. High Throughput and Multimedia Streaming are close in their traffic flow characteristics with High Throughput being a bit more bursty and not as long-lived as Multimedia Streaming.
- o Low-Priority Data for applications or services that can tolerate short or long interruptions of packet flows. The Low-Priority Data service class can be viewed as "don't care" to some degree.

Best-Effort Category

- o All traffic that is not differentiated in the network falls into this category and is mapped into the Standard service class. If a packet is marked with a DSCP value that is not supported in the network, it SHOULD be forwarded using the Standard service class.

Figure 1, below, provides a grouping of the defined user/subscriber service classes into four categories, with indications of which ones use an independent flow for signaling or control; type of flow behavior (elastic, rate adaptive, or inelastic); and the last column provides end user Class of Service (CoS) rating as defined in ITU-T Recommendation G.1010.

Application Categories	Service Class	Signaled	Flow Behavior	G.1010 Rating
Application Control	A/V Sig	Not applicable	Inelastic	Responsive
Media-	Realtime Interactive	Yes	Inelastic	Interactive
	Audio	Yes	Inelastic	Interactive
	Video	Yes	Inelastic	Interactive
	Hi-Res	Yes	Inelastic	Interactive
	Multimedia	Yes	Rate	Moderately

Oriented	Conferencing		Adaptive	Interactive
	Broadcast	Yes	Inelastic	Responsive
	Multimedia Streaming	Yes	Elastic	Timely
Data	Low-Latency Data	No	Elastic	Responsive
	Conversational Signaling	No	Elastic or Inelastic	Timely
	High-Throughput Data	No	Elastic	Timely
	Low-Priority Data	No	Elastic	Non-critical
Best Effort	Standard	Not Specified		Non-critical

Figure 1. User/Subscriber Service Classes Grouping

Here is a short explanation of the end user CoS category as defined in ITU-T Recommendation G.1010. User oriented traffic is divided into four different categories, namely, interactive, responsive, timely, and non-critical. An example of interactive traffic is between two humans and is most sensitive to delay, loss, and jitter. Another example of interactive traffic is between two servers where very low delay and loss are needed. Responsive traffic is typically between a human and a server but can also be between two servers. Responsive traffic is less affected by jitter and can tolerate longer delays than interactive traffic. Timely traffic is either between servers or servers and humans and the delay tolerance is significantly longer than responsive traffic. Non-critical traffic is normally between servers/machines where delivery may be delay for period of time.

2.3. Service Class Characteristics

This document specifies what network administrators are to expect when configuring service classes identified by their differing characteristics. Figure 2 identifies these service classes along with their characteristics, as well as the tolerance to loss, delay and jitter for each service class. Properly engineered networks to these PHBs will achieve expected results. That said, not all of the identified service classes are expected in each operator's network.

Service Class Name	Traffic Characteristics	Tolerance to		
		Loss	Delay	Jitter
Network Control	Variable size packets, mostly inelastic short messages, but traffic can also burst (BGP)	Low	Low	Yes
Realtime Interactive	Inelastic, mostly variable rate	Low	Very Low	Low
Audio	Fixed-size small packets, inelastic	Very Low	Very Low	Very Low
Video	Fixed-size small-large packets, inelastic	Very Low	Very Low	Very Low
Hi-Res A/V	Fixed-size small-large packets, inelastic	Very Low	Very Low	Very Low
Multimedia Conferencing	Variable size packets, constant transmit interval, rate adaptive, reacts to loss	Low - Medium	Low - Medium	Low - Medium
Multimedia Streaming	Variable size packets, elastic with variable rate	Low - Medium	Medium	High
Broadcast	Constant and variable rate, inelastic, non-bursty flows	Very Low	Medium	Low
Low-Latency Data	Variable rate, bursty short-lived elastic flows	Low	Low - Medium	Yes
Conversational Signaling	Variable size packets, some what bursty short-lived flows	Low	Low	Yes
OAM	Variable size packets, elastic & inelastic flows	Low	Medium	Yes
High-Throughput Data	Variable rate, bursty long-lived elastic flows	Low	Medium - High	Yes
Standard	A bit of everything	Not Specified		
Low-Priority Data	Non-real-time and elastic	High	High	Yes

Figure 2. Service Class Characteristics

Notes for Figure 2: A "Yes" in the jitter-tolerant column implies that received data is buffered at the endpoint and that a moderate level of server or network-induced variation in delay is not expected to affect the application. Applications that use TCP or SCTP as a transport are generally good examples. Routing protocols and peer-to-peer signaling also fall in this class; although loss can create problems in setting up calls, a moderate level of jitter merely makes call placement a little less predictable in duration.

Service classes indicate the required traffic forwarding treatment in order to meet user, application, and/or network expectations. Section 3 defines the service classes that MAY be used for forwarding network control traffic, and Section 4 defines the service classes that MAY be used for forwarding user oriented traffic with examples of intended application types mapped into each service class. Note that the application types are only examples and are not meant to be all-inclusive or prescriptive. Also, note that the service class naming or ordering does not imply any priority ordering. They are simply reference names that are used in this document with associated QoS behaviors that are optimized for the particular application types they support. Network administrators MAY choose to assign different service class names to the service classes that they will support. Figure 3 defines the RECOMMENDED relationship between service classes and DS codepoint assignment with application examples. It is RECOMMENDED that this relationship be preserved end to end.

Service Class Name	DSCP Name	DSCP Value	Application Examples
Network Control	CS6&CS7	11xxxx	Network routing
Realtime Interactive	CS5, CS5-Admit	101000, 101001	Remote/Virtual Desktop and Interactive gaming
Audio	EF Voice-Admit	101110 101100	Voice bearer
Hi-Res A/V	CS4, CS4-Admit	100000, 100001	Conversational Hi-Res Audio/Video bearer
Video	AF41,AF42 AF43	100010,100100 100110	Audio/Video conferencing bearer
Multimedia Conferencing	MC, MC-Admit	011101, 100101	Presentation Data and App Sharing/Whiteboarding
Multimedia Streaming	AF31,AF32 AF33	011010,011100 011110	Streaming video and audio on demand

Broadcast	CS3, CS3-Admit	011000, 011001	Broadcast TV, live events & video surveillance
Low-Latency Data	AF21,AF22 AF23	010010,010100 010110	Client/server trans., Web- based ordering, IM/Pres
Conversational Signaling	A/V-Sig	010001	Conversational signaling
OAM	CS2	010000	OAM&P
High-Throughput Data	AF11,AF12 AF13	001010,001100 001110	Store and forward applications
Low-Priority Data	CS1	001000	Any flow that has no BW assurance
Best Effort	CS0	000000	Undifferentiated applications

Figure 3. DSCP to Service Class Mapping

Notes for Figure 3:

- o Default Forwarding (DF) and Class Selector 0 (CS0) (i.e., Best Effort) provide equivalent behavior and use the same DS codepoint, '000000'.
- o RFC 2474 identifies any DSCP with a value of 11xxxx to be for network control. This remains true, while it removes 12 DSCPs from the overall pool of 64 available DSCP values (the 4 that are x11 from this group are within pool 2 of RFC 2474, and remain as only experimentally assignable/useable).
- o All PHB names that say "-Admit" are to be used only when a capacity-admission protocol is utilized for that or each traffic flow.

Changes from table 3 of RFC 4594 are as follows:

- o The old term "Signaling" was using CS5 (101000), now is exclusively for the "Conversational Signaling" service group using the DSCP name of "A/V-Sig" (010001), which is newly defined in [ID-DSCP]. This is because CS5 aggregates into the 101xxx aggregate when using layer 2 technologies such as 802.3 Ethernet, 802.11 Wireless Ethernet MPLS, etc - each of which only have 3 bits to mark with. A traffic type that can have very large packets and is not delay sensitive (within reason) is not appropriate for have a 101xxx marking. A REQUIRED behavior for this PHB is that it not be starved in any node.

- o "Conversational" is a new term to include all interactive audio and video. The Conversational service group consists of the audio service class, the video service class and the new Hi-Res service class.
- o "Audio" obsoletes the term "Telephony", which has generally not retained the "video" aspect within the IETF, where video is still commonly called out as a separate thing. Audio retains the nonadmitted traffic PHB of EF (101110), while capacity-admitted audio has been added via the RFC 5865 defined PHB Voice-Admit.
- o "Video" now is AF4x, with AF41 specifically for capacity-admitted video traffic, while AF42 and AF43 are nonadmitted video traffic.
- o "Hi-Res A/V", part of the Conversational service group, is created by [ID-DSCP] for an additional business differentiation interactive video marking for higher end traffic. It is within the 100xxx as CS4 (for nonadmitted traffic) and CS4-Admit (100001) (for capacity-admitted traffic).
- o "Realtime Interactive" is now using CS5 (for nonadmitted traffic), but adds a capacity-admitted DSCP CS5-Admit (101001).
- o "Multimedia Conferencing" is no longer using the AF4x DSCPs, rather it will use the new PHB MC (100101) (for capacity-admitted) and MC-Admit (011101) (for nonadmitted traffic).
- o "Multimedia Streaming" retains using AF3x, however, AF31 is now used for capacity-admitted traffic, while AF32/33 are nonadmitted.
- o "Broadcast" replaces "Broadcast Video" using CS3 (for nonadmitted traffic), and adds a capacity-admitted PHB CS3-Admit (011001).

It is expected that network administrators will base their choice of the service classes that they will support on their need.

Figure 4 provides a summary of DiffServ CoS mechanisms that MUST be used for the defined service classes that are further detailed in Sections 3 and 4 of this document. According to what applications/services need to be differentiated, network administrators MAY choose the service class(es) that need to be supported in their network.

Service Class	DSCP	Conditioning at DS Edge	PHB Used	Queuing	AQM
Network Control	CS6/CS7	See Section 3.1	RFC2474	Rate	Yes
Realtime	CS5,	Police using sr+bs	RFC2474	Rate	No

Interactive	CS5- Admit*					
Audio	EF, Voice- Admit*	Police using sr+bs	RFC3246 RFC5865	Priority	No	
Hi-Res A/V	CS4, CS4- Admit*	Police using sr+bs	RFC2474 [ID-DSCP]	Priority	No	
Video	AF41*, AF42 AF43	Using two-rate, three-color marker (such as RFC 2698)	RFC2597	Rate	Yes per DSCP	
Multimedia Conferencing	MC, MC- Admit*	Police using sr+bs	[ID-DSCP] [ID-DSCP]	Rate	No	
Multimedia Streaming	AF31*, AF32 AF33	Using two-rate, three-color marker (such as RFC 2698)	RFC2597	Rate	Yes per DSCP	
Broadcast	CS3, CS3- Admit*	Police using sr+bs	RFC2474 [ID-DSCP]	Rate	No	
Low- Latency Data	AF21 AF22 AF23	Using single-rate, three-color marker (such as RFC 2697)	RFC2597	Rate	Yes per DSCP	
Conversational Signaling	AV-Sig	Police using sr+bs	[ID-DSCP]	Rate	No	
OAM	CS2	Police using sr+bs	RFC2474	Rate	Yes	
High- Throughput Data	AF11 AF12 AF13	Using two-rate, three-color marker (such as RFC 2698)	RFC2597	Rate	Yes per DSCP	
Standard	DF	Not applicable	RFC2474	Rate	Yes	
Low-Priority Data	CS1	Not applicable	RFC3662	Rate	Yes	

Figure 4. Summary of CoS Mechanisms Used for Each Service Class

* denotes each DSCP identified for capacity-admission traffic only.

Notes for Figure 4:

- o Conditioning at DS edge means that traffic conditioning is performed at the edge of the DiffServ network where untrusted user devices are connected to two different administrative DiffServ networks.
- o "sr+bs" represents a policing mechanism that provides single rate with burst size control.
- o The single-rate, three-color marker (srTCM) behavior SHOULD be equivalent to RFC 2697, and the two-rate, three-color marker (trTCM) behavior SHOULD be equivalent to RFC 2698.
- o The PHB for Realtime-Interactive service class SHOULD be configured to provide high bandwidth assurance. It MAY be configured as another EF PHB (one capacity-admitted and one non-capacity-admitted, if both are to be used) that uses relaxed performance parameters and a rate scheduler.
- o The PHB for Multimedia Conferencing service class SHOULD be configured to provide high bandwidth assurance. It MAY be configured as another EF PHB (one capacity-admitted and one non-capacity-admitted, if both are to be used) that uses relaxed performance parameters and a rate scheduler.
- o The PHB for Broadcast service class SHOULD be configured to provide high bandwidth assurance. It MAY be configured as another EF PHB (one capacity-admitted and one non-capacity-admitted, if both are to be used) that uses relaxed performance parameters and a rate scheduler.

2.4. Service Classes vs. Treatment Aggregates (from RFC 5127)

There are misconceptions about the differences between RFC 4594 specified service classes, and RFC 5127 specified treatment aggregates. Often the two are conflated, and more often the phrase service class is used to mean both definitions. Almost all of the text previous to this section is used in defining service classes, and how one service class is different than another service class (based on traffic characteristics of the applications). Treatment aggregates are groupings of service classes with similar, but not identical, traffic characteristics to give similar treatment from a SP's network.

Below is taken from appendix of RFC 5127 as its recommended groupings of service classes into aggregates based in RFC 4594 specified traffic characteristic expectations.

+-----+			
Treatment	Treatment	DSCP	
Aggregate	Aggregate		
	Behavior		

Network Control	CS (RFC 2474)	CS6
Real-Time*	EF (RFC 3246)	EF, CS5, AF41, AF42, AF43, CS4, CS3
Assured Elastic	AF (RFC 2597)	CS2, AF31, AF21, AF11 AF32, AF22, AF12 AF33, AF23, AF13
Elastic	Default (RFC 2474)	Default, (CS0) CS1

Figure 5: RFC 5127 Defined Treatment Aggregate Behavior**

*NOTE: The RFC 5865 created VOICE-ADMIT is absence from the above figure because VOICE-ADMIT was created far later than this recommendation was. VOICE-ADMIT is appropriate for the Realtime Traffic Aggregate.

**NOTE: Figure 5 is directly from the appendix of RFC 5127 as that RFC's recommendation for configuration. This draft does not directly affect RFC 5127. That is left for an update to RFC 5127 itself. Based on the WG's take on this draft, RFC 5127 will necessitate an update to match this document's new service classes and additional DSCPs. The number of treatment aggregates are not expected to change in the RFC 5127 Update draft though, with the possible exception of a new treatment aggregate for capacity admitted flows; meaning there *might* be a 5th treatment aggregate proposed.

Treatment Aggregates are designed to nicely fit into technologies that do not have many different treatment levels to use. Here are 3 examples of technologies limited to an 8-value field,

- MPLS with its 3 Traffic Class (TC) bits [RFC5462].
- IEEE LANs with its 8-value Priority Code Point (PCP) field, as part of the 802.1Q header spec [IEEE1Q].
- IEEE 802.1e, which defines QoS over Wi-Fi, also only defines 8 levels (called User Priority or UP codes) [IEEE1E].

Treatment Aggregates are dependent on service classes to exist. Therefore many service classes can exist without the need to consider the use of treatment aggregates or their 8-value technologies. For example, a Layer 3 VPN can be all that is needed

to transit traffic flows, regardless of desired treatment, between enterprise LAN campuses. From this reality, the number of treatment aggregates has no direct bearing on the number of service classes.

2.4.1 Examples of Service Classes in Treatment Aggregates

It is **not** expected that all traffic characteristics are to be experienced across an SP's network for any given customer. For example, if VOICE-ADMIT is added to the Realtime Treatment Aggregate in Figure 5, there are 8 different service classes within the Realtime Treatment Aggregate. It is not expected that all 8 service classes will be deployed by customer networks traversing SP networks. RFC 5127's Treatment Aggregates are a table to configure which service class goes into which treatment aggregate. If there are 8 services classes in the Realtime treatment aggregate, there is very little difference than if there were one service within that same Realtime treatment aggregate - it would still be necessary to configure that treatment aggregate. Thus, it becomes a question of not

"how many service classes are there that go into treatment aggregates?"

but

"how many treatment aggregates have one or more services classes requiring configuration"?

Of the 4 treatment aggregates shown in Figure 5, if there are existing service classes in only 3 of the aggregates, then only 3 treatment aggregates are necessary. Of the 3 following examples, notice that examples 2 and 3 have the same number of treatment aggregates, but example 3 has more applications in their own service classes.

Examples 2 and 3 are made under the following assumptions:

- this draft's Service Classes and DSCP assignments are utilized.
- the new AF-Sig DSCP in the Assured Elastic treatment aggregate.
- the Audio, Video service classes are in the EF treatment aggregate.
- the VOICE-ADMIT DSCP is in the EF treatment aggregate.

2.4.1.1 Example 1 - Simple Voice Configuration/SLA

For example 1, we have an SP running MPLS and has an SLA to deliver Network Control, Voice and everything else is Best Effort. The

following table would apply to this configuration/SLA:

Applications	Service Class	DSCP(s)	Treatment Aggregate
Network Control	Network Control	CS6	Network Control
Voice	Audio	EF	Realtime
Everything else	DF	Default (CS0)	Elastic

Figure 6. Example 1 Configuration

Insert different treatments for this example
(i.e., AQM, RED, WFQ, colors, etc from above charts)

2.4.1.2 Example 2 - Voice/Video/Surveillance Configuration/SLA

For example 1, we have an SP running MPLS and has an SLA to deliver Control, audio, video, surveillance, audio & video signaling, and everything else is BE

Applications	Service Class	DSCP(s)	Treatment Aggregate
Network Control	Network Control	CS6	Network Control
Voice, video, surveillance	Audio, Video, Broadcast	EF, AF42, CS3	Realtime
audio & video signaling	Conversational Signaling	AV-Sig	Assured Elastic
Everything else	DF	Default (CS0)	Elastic

Figure 7. Example 2 Configuration

Insert different treatments for this example
(i.e., AQM, RED, WFQ, colors, etc from above charts)

2.4.1.2 Example 3 - Complex CAC realtime/Surveillance/+apps Configuration/SLA

For example 1, we have an SP running MPLS and has an SLA to deliver

Control, voice, CAC voice, CAC video, streaming, signaling, LL data, Network Mgmt., and everything else is BE (including non-CAC video because it is not authorized or authenticated on network)

Applications	Service Class	DSCP(s)	Treatment Aggregate
Network Control	Network Control	CS6	Network Control
Voice, CAC-Voice CAC-video, surveillance	Audio, Video, Broadcast	Voice-Admit EF, AF41 CS3	Realtime
audio & video signaling, VOD (streaming), Network Mgmt.	Conversational Signaling, Low- Latency Data, Multimedia Streaming, OAM	AV-Sig AF21 AF31 CS2	Assured Elastic
Everything else	DF	Default (CS0)	Elastic

Figure 8. Example 3 Configuration

Insert different treatments for this example
(i.e., AQM, RED, WFQ, colors, etc from above charts)

3. Network Control Traffic

Network control traffic is defined as packet flows that are essential for stable operation of an administered network, as well as the information exchanged between neighboring networks across a peering point where SLAs are in place. Network control traffic is different from user application control (signaling) that may be generated by some applications or services. Network control traffic is mostly between routers and network nodes (e.g., routing or mgmt protocols) that are used for operating, administering, controlling, or managing whole networks, network parts or just network segments. Network Control Traffic may be split into two service classes, i.e., Network Control and OAM.

3.1. Current Practice in the Internet

Based on today's routing protocols and network control procedures that are used in the Internet, we have determined that CS6 DSCP value SHOULD be used for routing and control and that CS7 DSCP value SHOULD be reserved for future use, specifically if needed for future

routing or control protocols. Network administrators MAY use a Local/Experimental DSCP, any value that contains 11xx11; therefore, they may use a locally defined service class within their network to further differentiate their routing and control traffic.

RECOMMENDED Network Edge Conditioning for CS7 DSCP marked packets:

- o Drop or remark 111xxx packets at ingress to DiffServ network domain.
- o 111xxx marked packets SHOULD NOT be sent across peering points. Exchange of control information across peering points SHOULD be done using CS6 DSCP and the Network Control service class.
- o any internally defined 11xxx1 values, valid within that network domain, be remarked to CS6 upon egress at network peering points.

3.2. Network Control Service Class

The Network Control service class is used for transmitting packets between network devices (routers) that require control (routing) information to be exchanged between similar devices within the administrative domain, as well as across a peering point between adjacent administrative domains. Traffic transmitted in this service class is very important as it keeps the network operational, and it needs to be forwarded in a timely manner.

The Network Control service class SHOULD be configured using the DiffServ CS6 PHB, defined in [RFC2474]. This service class MUST be configured so that the traffic receives a minimum bandwidth guarantee, to ensure that the packets always receive timely service. The configured forwarding resources for Network Control service class MUST be such that the probability of packet drop under peak load is very low. The Network Control service class SHOULD be configured to use a Rate Queuing system such as defined in Section 1.4.1.2 of this document.

The following are examples of protocols and applications that MUST use the Network Control service class if present in a network:

- o Routing packet flows: OSPF, BGP, ISIS, RIP.
- o Control information exchange within and between different administrative domains across a peering point where SLAs are in place.
- o LSP setup using CR-LDP and RSVP-TE.

The following protocols and applications MUST NOT use the Network Control service class:

- o User oriented traffic is not allowed to use this service class.

By user oriented traffic, we mean packet flows that originate from user-controlled end points that are connected to the network.

- o even if originating from a server or a device acting on behalf of a user or endpoint,
- o even if it is application or in-band signaling to establish a connection wholly within a single network or across peering points of/to adjacent networks (e.g., creating a tunnel such as a VPN, or data path control signaling).

The following are traffic characteristics of packet flows in the Network Control service class:

- o Mostly messages sent between routers and network servers.
- o Variable size packets, normally one packet at a time, but traffic can also burst (BGP, OSPF, etc).
- o IGMP, hen is used only for the normal multicast routing purpose.

The REQUIRED DSCP marking is CS6 (Class Selector 6).

RECOMMENDED Network Edge Conditioning:

- o At peering points (between two DiffServ networks) where SLAs are in place, CS6 marked packets MUST be policed, e.g., using a single rate with burst size (sr+bs) token bucket policer to keep the CS6 marked packet flows to within the traffic rate specified in the SLA.
- o CS6 marked packet flows from untrusted sources (for example, end user devices) MUST be dropped or remarked at ingress to the DiffServ network. What a network admin remarks this user oriented traffic to is a matter of local policy, and inspection of the packets can determine which application is used for proper marking to a more appropriate DSCP, such as from table 3. of this document.
- o Packets from users/subscribers are not permitted access to the Network Control service classes.

The fundamental service offered to the Network Control service class is enhanced best-effort service with high bandwidth assurance. Since this service class is used to forward both elastic and inelastic flows, the service SHOULD be engineered so that the Active Queue Management (AQM) [RFC2309] is applied to CS6 marked packets.

If RED [RFC2309] is used as an AQM algorithm, the min-threshold specifies a target queue depth, and the max-threshold specifies the queue depth above which all traffic is dropped or ECN marked. Thus,

in this service class, the following inequality should hold in queue configurations:

- o min-threshold CS6 < max-threshold CS6
- o max-threshold CS6 <= memory assigned to the queue

Note: Many other AQM algorithms exist and are used; they should be configured to achieve a similar result.

3.3. OAM Service Class

The OAM (Operations, Administration, and Management) service class is RECOMMENDED for OAM&P (Operations, Administration, and Management and Provisioning) using protocols such as Simple Network Management Protocol (SNMP), Trivial File Transfer Protocol (TFTP), FTP, Telnet, and Common Open Policy Service (COPS). Applications using this service class require a low packet loss but are relatively not sensitive to delay. This service class is configured to provide good packet delivery for intermittent flows.

The OAM service class SHOULD use the Class Selector (CS) PHB defined in [RFC2474]. This service class SHOULD be configured to provide a minimum bandwidth assurance for CS2 marked packets to ensure that they get forwarded. The OAM service class SHOULD be configured to use a Rate Queuing system such as defined in Section 1.4.1.2 of this document.

The following applications SHOULD use the OAM service class:

- o Provisioning and configuration of network elements.
- o Performance monitoring of network elements.
- o Any network operational alarms.

The following are traffic characteristics:

- o Variable size packets.
- o Intermittent traffic flows.
- o Traffic may burst at times.
- o Both elastic and inelastic flows.
- o Traffic not sensitive to delays.

RECOMMENDED DSCP marking:

- o All flows in this service class are marked with CS2 (Class Selector 2).

Applications or IP end points SHOULD pre-mark their packets with CS2 DSCP value. If the end point is not capable of setting the DSCP value, then the router topologically closest to the end point SHOULD perform Multifield (MF) Classification, as defined in [RFC2475].

RECOMMENDED conditioning performed at DiffServ network edge:

- o Packet flow marking (DSCP setting) from untrusted sources (end user devices) SHOULD be verified at ingress to DiffServ network using Multifield (MF) Classification methods, defined in [RFC2475].
- o Packet flows from untrusted sources (end user devices) SHOULD be policed at ingress to DiffServ network, e.g., using single rate with burst size token bucket policer to ensure that the traffic stays within its negotiated or engineered bounds.
- o Packet flows from trusted sources (routers inside administered network) MAY not require policing.
- o Normally OAM&P CS2 marked packet flows are not allowed to flow across peering points. If that is the case, then CS2 marked packets SHOULD be policed (dropped) at both egress and ingress peering interfaces.

The fundamental service offered to "OAM" traffic is enhanced best-effort service with controlled rate. The service SHOULD be engineered so that CS2 marked packet flows have sufficient bandwidth in the network to provide high assurance of delivery. Since this service class is used to forward both elastic and inelastic flows, the service SHOULD be engineered so that Active Queue Management [RFC2309] is applied to CS2 marked packets.

If RED [RFC2309] is used as an AQM algorithm, the min-threshold specifies a target queue depth for each DSCP, and the max-threshold specifies the queue depth above which all traffic with such a DSCP is dropped or ECN marked. Thus, in this service class, the following inequality should hold in queue configurations:

- o min-threshold CS2 < max-threshold CS2
- o max-threshold CS2 <= memory assigned to the queue

Note: Many other AQM algorithms exist and are used; they should be configured to achieve a similar result.

4. User Oriented Traffic

User oriented traffic is defined as packet flows between different users or subscribers, or from servers/nodes on behalf of a user. It is the traffic that is sent to or from end-terminals and that

supports a very wide variety of applications and services, to include traffic about a user or application that assists a user communicate. User oriented traffic can be classified in many different ways. What we have articulated throughout this document is a series of non-exhaustive list of categories for classifying user oriented traffic. We differentiated user oriented traffic that is real-time versus non-real-time, elastic or rate-adaptive versus inelastic, sensitive versus insensitive to loss as well as considering whether the traffic is interactive vs. one way communication, its responsiveness, whether it requires timely delivery, and critical versus non-critical. In the final analysis, we used all of the above for service differentiation, mapping application types that seemed to have different sets of performance sensitivities, and requirements to different service classes.

Network administrators can categorize their applications according to the type of behavior that they require and MAY choose to support all or a subset of the defined service classes. At the same time, we include a public facing default DSCP value, with its associated PHB, that is expected for each traffic type to ensure common or pervasive performance. Figure 3 provides some common applications and the forwarding service classes that best support them, based on their performance requirements.

4.1. Conversational Service Class Group

The Conversational Service Class Group consists of 3 different service classes, audio, video, and Hi-Res. We are describing the media sample, or bearer, packets for applications (e.g., RTP from [RFC3550]) that require bi-directional real-time, very low delay, very low jitter, and very low packet loss for relatively constant-rate traffic sources (inelastic traffic sources). It is RECOMMENDED that RTCP feedback use the same service class and be marked with the same DSCP as the bearer traffic for that (audio and/or video) call. This ensures comparable treatment within the network between endpoints.

The signaling to set-up these bearer flows is part of the Conversational Signaling service group that will be discussed later in Section 4. The following 3 subsections will detail what is expected within each bearer service class.

4.1.1 Audio Service Class

This service class MUST be used for IP Audio service.

The fundamental service offered to traffic in the Audio service class is minimum jitter, delay, and packet loss service up to a specified upper bound. There are two PHBs, both EF based, for the Audio service class:

Nonadmitted Audio traffic - MUST use the EF DSCP [RFC3246], and

is for traffic that has not had any capacity admission signaling performed for that flow or session.

Capacity-Admitted Audio traffic - MUST use the Voice-Admit DSCP [RFC5865], and is for traffic that has had any capacity admission signaling performed for that flow or session, e.g., RSVP [RFC2205] or NSIS [RFC4080].

The capacity-admitted Audio traffic operation is similar to an ATM CBR service, which has guaranteed bandwidth and which, if it stays within the negotiated rate, experiences nominal delay and no loss.

The nonadmitted Audio traffic, on the other hand, has had no such explicit guarantee, but has a favorable PHB ensuring high probability of delivery as well as nominal delay and no loss - implicitly assuming there is not too much like marked traffic between users within a flow.

There are two typical scenarios in which audio calls are established, on the public open Internet using protocols such as SIP, XMPP or H.323, or in more managed networks like enterprises or certain service providers which offer a audio service with some feature benefits and take part in the call signaling. These SPs or enterprises also use protocols like SIP, XMPP, H.323, but also use H.248/MEGACO and MGCP.

On the open Internet, typically there is no SP actively involved in the session set-up of calls, and therefore no servers providing assistance or features to help one user contact another user. Often, this traffic is marked or remarked with the DF (i.e., Best Effort) DSCP.

In more managed networks in which one of more operators have active servers aiding the audio call set-up, where DiffServ can be used and preserved to differentiate traffic, networks are offering a service, therefore need to do some, or a lot of engineering to ensure that capacity offered to one or more applications does not exceed the load to the network. Otherwise, the operator will have unhappy users, at least for that application's usage. This is true for any application, but is especially true for inelastic applications in which the application is rigid in its delivery requirements. Audio bearer traffic is typically such an application, video is another such application, but we will get to video in the next subsection.

When a user in a managed network has been authorized to send Audio traffic (i.e., call initiation via the operator's servers was not rejected), the call admission procedure should have verified that the newly admitted flow will be within the capacity of the Audio service class forwarding capability in the network. Capacity verification is a non-trivial thing, and can either be implicitly assumed by the call server(s) based on the operator's network design, or it can be explicitly signaled from an in-data-path

signaling mechanism that verifies the capacity is available now for this call, for each call made within that network. In the latter case, those that do not have verifiable network capacity along the data path are rejected. An in between means method is for call servers to count calls between two or more endpoints. By topologically understanding where the caller and called party is and have configured a known maximum it will allow between the two locations. This is especially true over WAN links that have far less capacity than LAN links or core parts of a network. Network operators will need to understand the topology between any two callers to ensure the appropriate amount of bandwidth is available for an expected number of simultaneous audio calls.

Once more than one bandwidth amount can be used for audio calls, for example - by allowing more than one codec with different bandwidths per codec for such calls, network engineering becomes more difficult. Since the inelastic nature of RTP payloads from this class do not react well to loss or significant delay in any substantive way, the Audio service class MUST forward packets as soon as possible.

The Audio service class that does not have capacity admission performed in the data path MUST use the Expedited Forwarding (EF) PHB, as defined in [RFC3246], so that all packets are forwarded quickly. The Audio service class that does have capacity admission performed in the data path MUST use the Voice-Admit PHB, as defined in [RFC5865], so that all packets are forwarded quickly. The Audio service class SHOULD be configured to use a Priority Queuing system such as that defined in Section 1.4.1.1 of this document.

The following applications SHOULD use the Audio service class:

- o VoIP (G.711, G.729, iLBC and other audio codecs).
- o Voice-band data over IP (modem, fax).
- o T.38 fax over IP.
- o Circuit emulation over IP, virtual wire, etc.
- o IP Virtual Private Network (VPN) service that specifies single-rate, mean network delay that is slightly longer than network propagation delay, very low jitter, and a very low packet loss.

The following are traffic characteristics:

- o Mostly fixed-size packets for VoIP (30, 60, 70, 120 or 200 bytes in size).
- o Packets emitted at constant time intervals.

- o Admission control of new flows is provided by Audio call server, media gateway, gatekeeper, edge router, end terminal, access node or in-data-path signaling that provides flow admission control function.

Applications or IP end points SHOULD pre-mark their packets with EF or Voice-Admit DSCP value, whichever is appropriate. If the end point is not capable of setting the DSCP value, then the router topologically closest to the end point SHOULD perform Multifield (MF) Classification, as defined in [RFC2475].

The RECOMMENDED DSCP marking is EF for nonadmitted audio flows, and Voice-Admit for capacity-admitted flows for the following applications:

- o VoIP (G.711, G.729 and other codecs).
- o Voice-band data over IP (modem and fax).
- o T.38 fax over IP.
- o Circuit emulation over IP, virtual wire, etc.

RECOMMENDED Network Edge Conditioning:

- o Packet flow marking (DSCP setting) from untrusted sources (end user devices) SHOULD be verified at ingress to DiffServ network using Multifield (MF) Classification methods, defined in [RFC2475]. If untrusted, the network edge SHOULD know if capacity-admission has been applied, since the edge router will have taken part in the admission signaling; therefore will know whether EF or Voice-Admit is the proper marking for that flow.
- o Packet flows from untrusted sources (end user devices) SHOULD be policed at ingress to DiffServ network, e.g., using single rate with burst size token bucket policer to ensure that the Audio traffic stays within its negotiated bounds.
- o Policing is OPTIONAL for packet flows from trusted sources whose behavior is ensured via other means (e.g., administrative controls on those systems).
- o Policing of Audio packet flows across peering points where SLA is in place is OPTIONAL as Audio traffic will be controlled by admission control mechanism between peering points.

The fundamental service offered to "Audio" traffic is enhanced best-effort service with controlled rate, very low delay, and very low loss. The service MUST be engineered so that EF marked packet flows have sufficient bandwidth in the network to provide guaranteed delivery. Otherwise, the service will have in place an explicit capacity-admission signaling protocol such as RSVP or NSIS and thus

mark the packets within the flow as Voice-Admit. Normally traffic in this service class does not respond dynamically to packet loss. As such, Active Queue Management [RFC2309] SHOULD NOT be applied to EF marked packet flows.

4.1.2 Video Service Class

The Video service class is for bidirectional applications that require real-time service for both constant and rate-adaptive traffic. SIP and H.323/V2 (and later) versions of video conferencing equipment with constant and dynamic bandwidth adjustment are such applications. The traffic sources in this service class either have a fixed bandwidth requirement (e.g., MPEG2, etc.), or have the ability to dynamically change their transmission rate (e.g., MPEG4/H.264, etc.) based on feedback from the receiver. This feedback SHOULD be accomplished using RTCP [RFC3550]. One approach for this downspeeding has the receiver detect packet loss, thus signaling in an RTCP message to the source the indication of lost (or delayed or out of order) packets in transit. When necessary the source then selects a lower rate encoding codec. When available, the source merely sends less data, resulting in lower resolution of the same visual display.

The Video service class is not for video downloads, webcasts, or single directional video or audio/video traffic of any kind. It is for human-to-human visual interaction between two users, or more if an MTP is used.

Typical video conferencing configurations negotiate the setup of audio/video session using protocols such as SIP and H.323. Just as with networks that have audio traversing them, video typically traverses the same two types of networks: the open big "I" Internet, in which most every type of traffic is best effort (DF), or on a more managed network such as an enterprise or SP's managed network in which servers within either network take part in the call signaling, thereby offering the video service.

When a user in a managed network has been authorized to send video traffic (i.e., call initiation via the operator's servers was not rejected), the call admission procedure should have verified that the newly admitted flow will be within the capacity of the video service class forwarding capability in the network. Capacity verification is a non-trivial thing, and can either be implicitly assumed by the call server(s) based on the operator's network design, or it can be explicitly signaled from an in-data-path signaling mechanism that verifies the capacity is available now for this call, for each call made within that network. In the latter case, those that do not have verifiable network capacity along the data path are rejected. An in between means method is for call servers to count calls between two or more endpoints. By topologically understanding where the caller and called party is and

have configured a known maximum it will allow between the two locations. Video is larger in bandwidth than audio, and the difference can be significant. For example, for a single G.711 audio call that is 80kbps, an associated video bandwidth for the same call can easily be 4Mbps. This is especially true over WAN links that have far less capacity than LAN links or core parts of a network. Network operators will need to understand the topology between any two callers to ensure the appropriate amount of bandwidth is available for an expected number of simultaneous video and/or audio/video calls.

Note that it is OPTIONALLY the case in these networks that the accompanying audio for the video call will be marked as the video is marked (i.e., using the same DSCP), but not always. One reason this has been done is for lip-sync.

The Video service class MUST use the Assured Forwarding (AF) PHB, defined in [RFC2597]. This service class MUST be configured to provide a bandwidth assurance for AF41, AF42, and AF43 marked packets to ensure that they get forwarded. The Video service class SHOULD be configured to use a Rate Queuing system for AF42 and AF43 traffic flows, such as that defined in Section 1.4.1.2 of this document. However, AF41 MUST be designated as the DSCP for use when capacity-admission signaling has been used, such as RSVP or NSIS, to guarantee delivery through the network. AF42 and AF43 will be used for non-admitted video calls, as well as overflows from AF41 sources that send more packets than they have negotiated bandwidth for that call.

The following applications MUST use the Video service class:

- o SIP and H.323/V2 (and later) versions of video conferencing applications (interactive video).
- o Video conferencing applications with rate control or traffic content importance marking.
- o Interactive, time-critical, and mission-critical applications.

NOTE with regards to the above bullet: this usage SHOULD be minimized, else the video traffic will suffer - unless this is engineered into the topology.

The following are traffic characteristics:

- o Variable size packets (i.e., small to large in size).
- o The higher the resolution or change rate between each image, the higher the duration of large packets.
- o Usually constant inter-packet time interval.

- o Can be Variable rate in transmission.
- o Source is capable of reducing its transmission rate based on being told receiver is detecting packet loss (e.g., via RTCP).

Applications or IP end points SHOULD pre-mark their packets with DSCP values as shown below. If the end point is not capable of setting the DSCP value, then the router topologically closest to the end point SHOULD perform Multifield (MF) Classification, as defined in [RFC2475] and mark all packets as AF4x. Note: In this case, the two-rate, three-color marker will be configured to operate in Color-Blind mode.

Mandatory DSCP marking when performed by router closest to source:

- o AF41 = up to specified rate "A", which is dedicated to non-Hi-Res capacity-admitted video traffic.

Note the audio of an A/V call can be marked AF41 as well.

- o AF42 = all non-Hi-Res video traffic marked AF41 in excess of specified rate "A", or new non-admitted video traffic but below specified rate "B".
- o AF43 = in excess of specified rate "B".
- o Where "A" < "B".

Note: One might expect "A" to approximate the peak rates of sum of all admitted video flows, plus the sum of the mean rates and "B" to approximate the sum of the peak rates of those same two flows.

Mandatory DSCP marking when performed by SIP or H.323/V2 videoconferencing equipment:

- o AF41 = SIP or H.323 video conferencing audio stream RTP.
- o AF41 = SIP or H.323 video conferencing video control RTCP.
- o AF41 = SIP or H.323 video conferencing video stream up to specified rate "A".
- o AF42 = SIP or H.323 video conferencing video stream in excess of specified rate "A" but below specified rate "B".
- o AF42 = SIP or H.323 video conferencing video control RTCP, for those video streams that were generated using AF42.
- o AF43 = SIP or H.323 video conferencing video stream in excess of specified rate "B".

- o AF43 = SIP or H.323 video conferencing video control RTCP, for those video streams that were generated using AF43.
- o Where "A" < "B".

Mandatory conditioning performed at DiffServ network edge:

- o The two-rate, three-color marker SHOULD be configured to provide the behavior as defined in trTCM [RFC2698].
- o If packets are marked by trusted sources or a previously trusted DiffServ domain and the color marking is to be preserved, then the two-rate, three-color marker SHOULD be configured to operate in Color-Aware mode.
- o If the packet marking is not trusted or the color marking is not to be preserved, then the two-rate, three-color marker SHOULD be configured to operate in Color-Blind mode.

The fundamental service offered to nonadmitted "Video" traffic is enhanced best-effort service with controlled rate and delay. The fundamental service offered to capacity-admitted "Video" traffic is a guaranteed service using in-data-path signaling to ensure expected delivery in a timely manner. For a non-admitted video conferencing service, if a 1% packet loss detected at the receiver triggers an encoding rate change, thus dropping to the next lower provisioned video encoding rate then Active Queue Management [RFC2309] SHOULD be used primarily to switch the video encoding rate under congestion, changing from high rate to lower rate, i.e., 1472 kbps to 768 kbps. This rule applies to all AF42 and 43 flows. The probability of loss of AF41 traffic MUST NOT exceed the probability of loss of AF42 traffic, which in turn MUST NOT exceed the probability of loss of AF43 traffic.

Capacity-admitted video service should not result in packet loss. However, administratively this MAY be allowed to cause a purposeful downspeeding event (i.e., a change in resolution or a change in codec) to occur due to congestion.

If RED [RFC2309] is used as an AQM algorithm, the min-threshold specifies a target queue depth for each DSCP, and the max-threshold specifies the queue depth above which all traffic with such a DSCP is dropped or ECN marked. Thus, in this service class, the following inequality should hold in queue configurations:

- o min-threshold AF43 < max-threshold AF43
- o max-threshold AF43 <= min-threshold AF42
- o min-threshold AF42 < max-threshold AF42
- o max-threshold AF42 <= min-threshold AF41

- o min-threshold AF41 < max-threshold AF41
- o max-threshold AF41 <= memory assigned to the queue

Note: This configuration tends to drop AF43 traffic before AF42 and AF42 before AF41. Many other AQM algorithms exist and are used; they should be configured to achieve a similar result.

4.1.3 Hi-Res Service Class

The Hi-Res service class is for higher end (i.e., deemed 'more important') bidirectional applications that require real-time service for both constant and rate-adaptive traffic. There are two PHBs, both EF based, for the Hi-Res video conferencing service class:

Nonadmitted Hi-Res traffic - MUST use the CS4 DSCP [RFC2474], and is for traffic that has not had any capacity admission signaling performed for that flow or session.

Capacity-Admitted Hi-Res traffic - MUST use the CS4-Admit DSCP [ID-DSCP], and is for traffic that has had any capacity admission signaling performed for that flow or session, e.g., RSVP [RFC2205] or NSIS [RFC4080].

The capacity-admitted Hi-Res video conferencing traffic operation is similar to an ATM CBR service, which has guaranteed bandwidth and which, if it stays within the negotiated rate, experiences nominal delay and no loss.

SIP and H.323/V2 (and later) versions of video conferencing equipment with constant and dynamic bandwidth adjustment are such applications. The traffic sources in this service class either have a fixed bandwidth requirement (e.g., MPEG2), or have the ability to dynamically change their transmission rate (e.g., MPEG4/H.264) based on feedback from the receiver. This feedback SHOULD be accomplished using RTCP [RFC3550]. One approach for this downspeeding has the receiver detect packet loss, thus signaling in an RTCP message to the source the indication of lost (or delayed or out of order) packets in transit. When necessary the source then selects a lower rate encoding codec. When available, the source merely sends less data, resulting in lower resolution of the same visual display.

The Hi-Res service class, as with the Video service class, is not for video downloads, webcasts, or single directional video or audio/video traffic of any kind. It is for human-to-human visual interaction between two users, or more if a video conference bridge is used.

Typical Hi-Res video conferencing configurations negotiate the setup

of audio/video session using protocols such as SIP and H.323. Hi-Res video conferencing is generally not over the big "I" Internet, rather nearly exclusively over more managed networks such as an enterprise or special purpose SP's managed network in which servers within either network take part in the call signaling, thereby offering the video service. In addition, typically this type of audio/video service has high business expectations for minimized packet loss, pixilation or other issues with the audio/video experience. In the recent past, entire T3s have been dedicated to a signal Hi-Res call; sometimes one T3 per site of a multi-site video conference.

Hi-Res video conferencing often has larger in bandwidth than the typical video call. The audio portion can be increased as well, as stereo capabilities are often necessary to provide an in-room experience from a distance. The difference can be significant (or another step up from just a typical video service). For example, for a single G.711 audio call that is 80kbps, a Hi-Res conference usually runs G.722 wideband audio at 256kbps. Typical video delivery is up to 4Mbps, whereas a Hi-Res conference can have three 1080p/30fps widescreen displays requiring at least 12Mbps, with a burst capability of much more.

If there were no congestion on the wire, the expected treatment between a video service and a Hi-Res conference would be the same. However, it is typically the case that the Hi-Res conferencing flows have more rigid requirements for quality and business-wise, need to be experience far less errors than the regular video service on the same network.

Note that it is likely the case in these networks that the accompanying audio to the Hi-Res video call will be marked as the Hi-Res video is marked (i.e., using the same DSCP).

The Hi-Res service class MUST use the Class Selector 5 (CS4) PHB, defined in [RFC2474], for non-capacity-admitted conferences. While the capacity-admitted Hi-Res conferences MUST use the CS4-Admit PHB, defined in [ID-DSCP]. This service class MUST be configured to provide a bandwidth assurance for CS4 and CS4-Admit marked packets to ensure that they get forwarded. The Hi-Res service class SHOULD be configured to use a Priority Queuing system such as that defined in Section 1.4.1.1 of this document. Further, CS4-Admit will be designated as the DSCP for use when capacity-admission signaling has been used, such as RSVP or NSIS, to guarantee delivery through the network. CS4 will be used for non-admitted Hi-Res conferences, as well as overflows from CS4-Admit sources that send more packets than they have negotiated bandwidth for that call.

The following applications MUST use the Hi-Res service class:

- o SIP and H.323/V2 (and later) versions of Hi-Res video conferencing applications (interactive Hi-Res video).

- o Video conferencing applications with rate control or traffic content importance marking.

The following are traffic characteristics:

- o Variable size packets.
- o The higher the resolution or change rate between each image, the higher the duration of large packets.
- o Usually constant inter-packet time interval.
- o Can be Variable rate in transmission.
- o Source is capable of reducing its transmission rate based on being told receiver is detecting packet loss.

Applications or IP end points SHOULD pre-mark their packets with DSCP values as shown below. If the end point is not capable of setting the DSCP value, then the router topologically closest to the end point SHOULD perform Multifield (MF) Classification, as defined in [RFC2475] and mark all packets as AF4x.

Mandatory DSCP marking when performed by router closest to source:

- o CS4-Admit = up to specified rate "A", which is dedicated to capacity-admitted Hi-Res traffic.

Note the audio of an A/V call can be marked CS4-Admit as well.

- o CS4 = all video traffic marked CS4-Admit in excess of specified rate "A", or new non-admitted video traffic but below specified rate "B".
- o Where "A" < "B".

Note: One might expect "A" to approximate the peak rates of sum of all admitted video flows, plus the sum of the mean rates and "B" to approximate the sum of the peak rates of those same two flows.

Mandatory DSCP marking when performed by SIP or H.323/V2 videoconferencing equipment:

- o CS4-Admit = SIP or H.323 video conferencing audio stream RTP/UDP.
- o CS4-Admit = SIP or H.323 video conferencing video control RTCP/TCP.
- o CS4-Admit = SIP or H.323 video conferencing video stream up to specified rate "A".

- o CS4 = SIP or H.323 video conferencing video stream in excess of specified rate "A" but below specified rate "B".
- o Where "A" < "B".

Mandatory conditioning performed at DiffServ network edge:

- o The two-rate, three-color marker SHOULD be configured to provide the behavior as defined in trTCM [RFC2698].
- o If packets are marked by trusted sources or a previously trusted DiffServ domain and the color marking is to be preserved, then the two-rate, three-color marker SHOULD be configured to operate in Color-Aware mode.
- o If the packet marking is not trusted or the color marking is not to be preserved, then the two-rate, three-color marker SHOULD be configured to operate in Color-Blind mode.

The fundamental service offered to nonadmitted "Hi-Res" traffic is enhanced best-effort service with controlled rate and delay. The fundamental service offered to capacity-admitted "Hi-Res" traffic is a guaranteed service using in-data-path signaling to ensure expected or timely delivery. Capacity-admitted video service SHOULD NOT result in packet loss. However, administratively this MAY be allowed to cause a purposeful downspeeding event (i.e., a change in resolution or a change in codec) to occur.

4.2. Realtime-Interactive Service Class

The Realtime-Interactive service class is for bidirectional applications that require low loss and jitter and very low delay for constant or variable rate inelastic traffic sources. Interactive gaming applications that do not have the ability to change encoding rates or to mark packets with different importance indications is one good example of such an application. Another set of applications is virtualized desktop applications in which a remote user has a keyboard, mouse and display monitor, but the desktop is virtualized with the memory/processor/applications back in a common data center, requiring near instantaneous feedback on the user's monitor of any changes caused by the application or an action by the user. Rich media protocols for voice and video MUST NOT use the Realtime-Interactive service class, but rather the appropriate service class from the Conversational service group discussed early in Section 4.1.

The Realtime-Interactive service class will use two PHBs:

Nonadmitted Realtime-Interactive traffic - MUST use the CS5 DSCP [RFC2474], and is for traffic that has not had any capacity

admission signaling performed for that flow or session.

Capacity-Admitted Realtime-Interactive traffic - MUST use the CS5-Admit DSCP [ID-DSCP], and is for traffic that has had any capacity admission signaling performed for that flow or session, e.g., RSVP [RFC2205] or NSIS [RFC4080].

The capacity-admitted Realtime-Interactive traffic operation is similar to an ATM CBR service, which has guaranteed bandwidth and which, if it stays within the negotiated rate, experiences nominal delay and no loss.

Either of the above service classes can be configured as EF based by using a relaxed performance parameter and a rate scheduler.

When a user/endpoint has been authorized to start a new session (i.e., joins a networked game or logs onto a virtualized workstation), the admission procedure should have verified that the newly admitted data rates will be within the engineered capacity of the Realtime-Interactive service class. The bandwidth in the core network and the number of simultaneous Realtime-Interactive sessions that can be supported SHOULD be engineered to control traffic load for this service.

This service class SHOULD be configured to provide a high assurance for bandwidth for CS5 PHB, defined in [RFC2474], or CS5-Admit [ID-DSCP] for guaranteed service through a capacity-admission signaling protocol. The Realtime-Interactive service class SHOULD be configured to use a Rate Queuing system such as that defined in Section 1.4.1.2 of this document. Note that either Realtime-Interactive PHB MAY be configured as another EF PHB, specifically CS5-Admit, that uses a relaxed performance parameter and a rate scheduler, in the priority queue as defined in Section 1.4.1.1 of this document.

The following applications MUST use the Realtime-Interactive service class:

- o Interactive gaming and control.
- o Remote Desktop applications
- o Virtualized Desktop applications.
- o Application server-to-application server non-bursty data transfer requiring very low delay.
- o Inelastic, interactive, time-critical, and mission-critical applications requiring very low delay.

The following are traffic characteristics:

- o Variable size packets.
- o Variable rate, though sometimes bursty, which will require engineering of the network to accommodate.
- o Application is sensitive to delay variation between flows and sessions.
- o Lost packets, if any, are usually ignored by application.

RECOMMENDED DSCP marking:

- o All non-admitted flows in this service class are marked with CS5 (Class Selector 5).
- o All capacity-admitted flows in this service class are marked with CS5-Admit.

Applications or IP end points SHOULD pre-mark their packets with CS5 or CS5-Admit DSCP value, depending on whether a capacity-admission signaling protocol is used for a flow. If the end point is not capable of setting the DSCP value, then the router topologically closest to the end point SHOULD perform Multifield (MF) Classification, as defined in [RFC2475].

RECOMMENDED conditioning performed at DiffServ network edge:

- o Packet flow marking (DSCP setting) from untrusted sources (end user devices) SHOULD be verified at ingress to DiffServ network using Multifield (MF) Classification methods defined in [RFC2475].
- o Packet flows from untrusted sources (end user devices) SHOULD be policed at ingress to DiffServ network, e.g., using single rate with burst size token bucket policer to ensure that the traffic stays within its negotiated or engineered bounds.
- o Packet flows from trusted sources (application servers inside administered network) MAY not require policing.
- o Policing of packet flows across peering points MUST adhere to the Service Level Agreement (SLA).

The fundamental service offered to nonadmitted "Realtime-Interactive" traffic is enhanced best-effort service with controlled rate and delay. The fundamental service offered to capacity-admitted "Realtime-Interactive" traffic is a guaranteed service using in-data-path signaling to ensure expected or timely delivery. Capacity-admitted Realtime-Interactive service SHOULD NOT result in packet loss. The service SHOULD be engineered so that CS5 marked packet flows have sufficient bandwidth in the network to provide high assurance of delivery. Normally, traffic in this

service class does not respond dynamically to packet loss. As such, Active Queue Management [RFC2309] SHOULD NOT be applied to CS5 marked packet flows.

4.3. Multimedia Conferencing Service Class

The Multimedia Conferencing service class is for applications that have a low to medium tolerance to delay, and are rate adaptive to lost packets in transit from sources. Presentation Data applications that are operational in conjunction with an audio/video conference is one good example of such an application. Another set of applications is application sharing or whiteboarding applications, also in conjunction to an A/V conference. In either case, the audio & video part of the flow MUST NOT use the Multimedia Conferencing service class, rather the more appropriate service class within the Conversational service group discussed earlier in Section 4.1.

The Multimedia Conferencing service class will use two PHBs:

- Nonadmitted Multimedia Conferencing traffic - MUST use the (new) MC DSCP [ID-DSCP], and is for traffic that has not had any capacity admission signaling performed for that flow or session.

- Capacity-Admitted Multimedia Conferencing traffic - MUST use the (new) MC-Admit DSCP [ID-DSCP], and is for traffic that has had any capacity admission signaling performed for that flow or session, e.g., RSVP [RFC2205] or NSIS [RFC4080].

The capacity-admitted Multimedia Conferencing traffic operation is similar to an ATM CBR service, which has guaranteed bandwidth and which, if it stays within the negotiated rate, experiences nominal delay and no loss.

When a user/endpoint initiates a presentation data, application sharing or whiteboarding session, it will typically be part of an audio or audio/video conference such as web-conferencing or an existing Telepresence call. The authorization procedure SHOULD be controlled through the coordinated effort to bind the A/V call with the correct Multimedia Conferencing packet flow through some use of identifiers not in scope of this document. The managed network this flow traverse and the number of simultaneous Multimedia Conferencing sessions that can be supported SHOULD be engineered to control traffic load for this service.

The non-capacity admitted Multimedia Conferencing service class SHOULD use the new MC PHB, defined in [ID-DSCP]. This service class SHOULD be configured to provide a high assurance for bandwidth for CS5 marked packets to ensure that they get forwarded. The Multimedia Conferencing service class SHOULD be configured to use a

Rate Queuing system such as that defined in Section 1.4.1.2 of this document. Note that this service class MAY be configured as another EF PHB that uses a relaxed performance parameter, a rate scheduler, and MC-Admit DSCP value, which MUST use the priority queue as defined in Section 1.4.1.1 of this document.

The following applications MUST use the Multimedia Conferencing service class:

- o Presentation Data applications, which can utilize vector graphics, raster graphics or video delivery.
- o Virtualized Desktop applications.
- o Application server-to-application server non-bursty data transfer requiring very low delay.

The following are traffic characteristics:

- o Variable size packets.
- o Variable rate, though sometimes bursty, which will require engineering of the network to accommodate.
- o Application is sensitive to delay variation between flows and sessions.
- o Lost packets, if any, can be ignored by the application.

RECOMMENDED DSCP marking:

- o All non-admitted flows in this service class are marked with the new MC DSCP.
- o All capacity-admitted flows in this service class are marked with MC-Admit.

Applications or IP end points SHOULD pre-mark their packets with the MC DSCP value. If the end point is not capable of setting the DSCP value, then the router topologically closest to the end point SHOULD perform Multifield (MF) Classification, as defined in [RFC2475].

RECOMMENDED conditioning performed at DiffServ network edge:

- o Packet flow marking (DSCP setting) from untrusted sources (end user devices) SHOULD be verified at ingress to DiffServ network using Multifield (MF) Classification methods defined in [RFC2475].
- o Packet flows from untrusted sources (end user devices) SHOULD be policed at ingress to DiffServ network, e.g., using single rate with burst size token bucket policer to ensure that the traffic

stays within its negotiated or engineered bounds.

- o Packet flows from trusted sources (application servers inside administered network) MAY not require policing.
- o Policing of packet flows across peering points MUST adhere to the Service Level Agreement (SLA).

The fundamental service offered to nonadmitted "Multimedia Conferencing" traffic is enhanced best-effort service with controlled rate and delay. The fundamental service offered to capacity-admitted "Multimedia Conferencing" traffic is a guaranteed service using in-data-path signaling to ensure expected or timely delivery. Capacity-admitted Multimedia Conferencing service SHOULD NOT result in packet loss. The service SHOULD be engineered so that Multimedia Conferencing marked packet flows have sufficient bandwidth in the network to provide high assurance of delivery. Normally, traffic in this service class does not respond dynamically to packet loss. As such, Active Queue Management [RFC2309] SHOULD NOT be applied to MC or MC-Admit marked packet flows.

4.4. Multimedia Streaming Service Class

The Multimedia Streaming service class is RECOMMENDED for applications that require near-real-time packet forwarding of variable rate elastic traffic sources that are not as delay sensitive as applications using the Broadcast service class. Such applications include streaming audio and video, some video (movies) on-demand applications, and non-interactive webcasts. In general, the Multimedia Streaming service class assumes that the traffic is buffered at the source/destination; therefore, it is less sensitive to delay and jitter.

The Multimedia Streaming service class MUST use the Assured Forwarding (AF3x) PHB, defined in [RFC2597]. This service class MUST be configured to provide a minimum bandwidth assurance for AF31, AF32, and AF33 marked packets to ensure that they get forwarded. The Multimedia Streaming service class SHOULD be configured to use Rate Queuing system for AF32 and AF33 traffic flows, such as that defined in Section 1.4.1.2 of this document. However, AF31 MUST be designated as the DSCP for use when capacity-admission signaling has been used, such as RSVP or NSIS, to guarantee delivery through the network. AF32 and AF33 will be used for non-admitted streaming flows, as well as overflows from AF31 sources that send more packets than they have negotiated bandwidth for that call.

The following applications SHOULD use the Multimedia Streaming service class:

- o Buffered streaming audio (unicast).

- o Buffered streaming video (unicast).
- o Non-interactive Webcasts.
- o IP VPN service that specifies two rates and is less sensitive to delay and jitter.

The following are traffic characteristics:

- o Variable size packets.
- o The higher the rate, the higher the density of large packets.
- o Variable rate.
- o Elastic flows.
- o Some bursting at start of flow from some applications, as well as an expected stepping up and down on the rate of the flow based on changes in resolution due to network conditions.

Applications or IP end points SHOULD pre-mark their packets with DSCP values as shown below. If the end point is not capable of setting the DSCP value, then the router topologically closest to the end point SHOULD perform Multifield (MF) Classification, as defined in [RFC2475], and mark all packets as AF3x. Note: In this case, the two-rate, three-color marker will be configured to operate in Color-Blind mode.

RECOMMENDED DSCP marking:

- o AF31 = up to specified rate "A".
- o AF32 = all traffic marked AF31 in excess of specified rate "A", or new AF32 traffic but below specified rate "B".
- o AF33 = in excess of specified rate "B".
- o Where "A" < "B".

Note: One might expect "A" to approximate the peak rates of sum of all streaming flows, plus the sum of the mean rates and "B" to approximate the sum of the peak rates of those same two flows.

RECOMMENDED conditioning performed at DiffServ network edge:

- o The two-rate, three-color marker SHOULD be configured to provide the behavior as defined in trTCM [RFC2698].
- o If packets are marked by trusted sources or a previously trusted DiffServ domain and the color marking is to be preserved, then

the two-rate, three-color marker SHOULD be configured to operate in Color-Aware mode.

- o If the packet marking is not trusted or the color marking is not to be preserved, then the two-rate, three-color marker SHOULD be configured to operate in Color-Blind mode.

The fundamental service offered to nonadmitted "Multimedia Streaming" traffic is enhanced best-effort service with controlled rate and delay. The fundamental service offered to capacity-admitted "Multimedia Streaming" traffic is a guaranteed service using in-data-path signaling to ensure expected delivery in a reasonable manner. The service SHOULD be engineered so that AF31 marked packet flows have sufficient bandwidth in the network to provide high assurance of delivery. Since the AF3x traffic is elastic and responds dynamically to packet loss, Active Queue Management [RFC2309] SHOULD be used primarily to reduce forwarding rate to the minimum assured rate at congestion points, unless AF31 has had a capacity-admission signaling protocol applied to the flow, such as RSVP or NSIS.

If a capacity-admission signaling protocol applied to the AF31 flow, which SHOULD be the case always, the AF31 PHB MAY be configured as another EF PHB that uses a relaxed performance parameter and a rate scheduler, in the priority queue as defined in Section 1.4.1.1 of this document.

The probability of loss of AF31 traffic MUST NOT exceed the probability of loss of AF32 traffic, which in turn MUST NOT exceed the probability of loss of AF33.

If RED [RFC2309] is used as an AQM algorithm, the min-threshold specifies a target queue depth for each DSCP, and the max-threshold specifies the queue depth above which all traffic with such a DSCP is dropped or ECN marked. Thus, in this service class, the following inequality MUST hold in queue configurations:

- o min-threshold AF33 < max-threshold AF33
- o max-threshold AF33 <= min-threshold AF32
- o min-threshold AF32 < max-threshold AF32
- o max-threshold AF32 <= min-threshold AF31
- o min-threshold AF31 < max-threshold AF31
- o max-threshold AF31 <= memory assigned to the queue

Note#1: this confirmation MUST be modified if AF31 has a capacity-admission signaling protocol applied to those flows, and the above will only apply to AF32 and AF33, while

AF31 (theoretically) has no packet loss.

Note#2: This configuration tends to drop AF33 traffic before AF32 and AF32 before AF31. Note: Many other AQM algorithms exist and are used; they SHOULD be configured to achieve a similar result.

4.5. Broadcast Service Class

The Broadcast service class is RECOMMENDED for applications that require near-real-time packet forwarding with very low packet loss of constant rate and variable rate inelastic traffic sources that are more delay sensitive than applications using the Multimedia Streaming service class. Such applications include broadcast TV, streaming of live audio and video events, some video-on-demand applications, and video surveillance. In general, the Broadcast service class assumes that the destination end point has a dejitter buffer, for video application usually a 2 - 8 video-frame buffer (66 to several hundred of milliseconds), thus expecting far less buffering before play-out than Multimedia Streaming, which can buffer in the seconds to minutes (to hours).

The Broadcast service class will use two PHBs:

Nonadmitted Broadcast traffic - MUST use the CS3 DSCP [RFC2474], and is for traffic that has not had any capacity admission signaling performed for that flow or session.

Capacity-Admitted Broadcast traffic - MUST use the CS3-Admit DSCP [ID-DSCP], and is for traffic that has had any capacity admission signaling performed for that flow or session, e.g., RSVP [RFC2205] or NSIS [RFC4080].

The capacity-admitted Broadcast traffic operation is similar to an ATM CBR service, which has guaranteed bandwidth and which, if it stays within the negotiated rate, experiences nominal delay and no loss.

Either of the above service classes can be configured as EF based by using a relaxed performance parameter and a rate scheduler.

When a user/endpoint initiates a new Broadcast session (i.e., starts an Internet radio application, starts a live Internet A/V event or a camera comes online to do video-surveillance), the admission procedure should be verified within the application that triggers the flow. The newly admitted data rates will SHOULD be within the engineered capacity of the Broadcast service class within that network. The bandwidth in the core network and the number of simultaneous Broadcast sessions that can be supported SHOULD be engineered to control traffic load for this service.

This service class SHOULD be configured to provide high assurance for bandwidth for CS3 marked packets to ensure that they get forwarded. The Broadcast service class SHOULD be configured to use Rate Queuing system such as that defined in Section 1.4.1.2 of this document. Note that either Broadcast PHB MAY be configured as another EF PHB, specifically CS3-Admit, that uses a relaxed performance parameter and a rate scheduler, in the priority queue as defined in Section 1.4.1.1 of this document.

The following applications SHOULD use the Broadcast service class:

- o Video surveillance and security (unicast).
- o TV broadcast including HDTV (likely multicast, but can be unicast).
- o Video on demand (unicast) with control (virtual DVD).
- o Streaming of live audio events (both unicast and multicast).
- o Streaming of live video events (both unicast and multicast).

The following are traffic characteristics:

- o Variable size packets.
- o The higher the rate, the higher the density of large packets.
- o Mixture of variable rate and constant rate flows.
- o Fixed packet emission time intervals.
- o Inelastic flows.

RECOMMENDED DSCP marking:

- o All non-admitted flows in this service class are marked with CS3 (Class Selector 3).
- o All capacity-admitted flows in this service class are marked with CS3-Admit.
- o In some cases, such as those for security and video surveillance applications, it is NOT RECOMMENDED, but allowed to use a different DSCP marking.

If so, then locally user definable (EXP/LU) codepoints in the range '011x11' MAY be used to provide unique traffic identification. The locally administrator definable (EXP/LU, from pool 2 of RFC 2474) codepoint(s) MAY be associated with the PHB that is used for CS3 or CS3-Admit traffic. Furthermore, depending on the network scenario, additional network edge

conditioning policy MAY be needed for the EXP/LU codepoint(s) used.

Applications or IP end points SHOULD pre-mark their packets with CS3 or CS3-Admit DSCP value. If the end point is not capable of setting the DSCP value, then the router topologically closest to the end point SHOULD perform Multifield (MF) Classification, as defined in [RFC2475].

RECOMMENDED conditioning performed at DiffServ network edge:

- o Packet flow marking (DSCP setting) from untrusted sources (end user devices) SHOULD be verified at ingress to DiffServ network using Multifield (MF) Classification methods defined in [RFC2475].
- o Packet flows from untrusted sources (end user devices) SHOULD be policed at ingress to DiffServ network, e.g., using single rate with burst size token bucket policer to ensure that the traffic stays within its negotiated or engineered bounds.
- o Packet flows from trusted sources (application servers inside administered network) MAY not require policing.
- o Policing of packet flows across peering points MUST be performed to the Service Level Agreement (SLA) of those peering entities.

The fundamental service offered to "Broadcast" traffic is enhanced best-effort service with controlled rate and delay. The fundamental service offered to capacity-admitted "Broadcast" traffic is a guaranteed service using in-data-path signaling to ensure expected or timely delivery. Capacity-admitted Broadcast service SHOULD NOT result in packet loss. The service SHOULD be engineered so that CS3 and CS3-Admit marked packet flows have sufficient bandwidth in the network to provide high assurance of delivery. Normally, traffic in this service class does not respond dynamically to packet loss. As such, Active Queue Management [RFC2309] SHOULD NOT be applied to CS3 marked packet flows.

4.6. Low-Latency Data Service Class

The Low-Latency Data service class is RECOMMENDED for elastic and responsive typically client-/server-based applications. Applications forwarded by this service class are those that require a relatively fast response and typically have asymmetrical bandwidth need, i.e., the client typically sends a short message to the server and the server responds with a much larger data flow back to the client. The most common example of this is when a user clicks a hyperlink (~ few dozen bytes) on a web page, resulting in a new web page to be loaded (Kbytes or MBs of data). This service class is configured to provide good response for TCP [RFC1633] short-lived flows that require real-time packet forwarding of variable rate

traffic sources.

The Low-Latency Data service class SHOULD use the Assured Forwarding (AF) PHB, defined in [RFC2597]. This service class SHOULD be configured to provide a minimum bandwidth assurance for AF21, AF22, and AF23 marked packets to ensure that they get forwarded. The Low-Latency Data service class SHOULD be configured to use a Rate Queuing system such as that defined in Section 1.4.1.2 of this document.

The following applications SHOULD use the Low-Latency Data service class:

- o Client/server applications.
- o Systems Network Architecture (SNA) terminal to host transactions (SNA over IP using Data Link Switching (DLSw)).
- o Web-based transactions (E-commerce).
- o Credit card transactions.
- o Financial wire transfers.
- o Enterprise Resource Planning (ERP) applications (e.g., SAP/BaaN).
- o VPN service that supports Committed Information Rate (CIR) with up to two burst sizes.
- o Instant Messaging and Presence protocols (e.g., SIP, XMPP).

The following are traffic characteristics:

- o Variable size packets.
- o Variable packet emission rate.
- o With packet bursts of TCP window size.
- o Short traffic bursts.
- o Source capable of reducing its transmission rate based on detection of packet loss at the receiver or through explicit congestion notification.

Applications or IP end points SHOULD pre-mark their packets with DSCP values as shown below. If the end point is not capable of setting the DSCP value, then the router topologically closest to the end point SHOULD perform Multifield (MF) Classification, as defined in [RFC2475] and mark all packets as AF2x. Note: In this case, the single-rate, three-color marker will be configured to operate in Color-Blind mode.

RECOMMENDED DSCP marking:

- o AF21 = flow stream with packet burst size up to "A" bytes.
- o AF22 = flow stream with packet burst size in excess of "A" but below "B" bytes.
- o AF23 = flow stream with packet burst size in excess of "B" bytes.
- o Where "A" < "B".

RECOMMENDED conditioning performed at DiffServ network edge:

- o The single-rate, three-color marker SHOULD be configured to provide the behavior as defined in srTCM [RFC2697].
- o If packets are marked by trusted sources or a previously trusted DiffServ domain and the color marking is to be preserved, then the single-rate, three-color marker SHOULD be configured to operate in Color-Aware mode.
- o If the packet marking is not trusted or the color marking is not to be preserved, then the single-rate, three-color marker SHOULD be configured to operate in Color-Blind mode.

The fundamental service offered to "Low-Latency Data" traffic is enhanced best-effort service with controlled rate and delay. The service SHOULD be engineered so that AF21 marked packet flows have sufficient bandwidth in the network to provide high assurance of delivery. Since the AF2x traffic is elastic and responds dynamically to packet loss, Active Queue Management [RFC2309] SHOULD be used primarily to control TCP flow rates at congestion points by dropping packets from TCP flows that have large burst size. The probability of loss of AF21 traffic MUST NOT exceed the probability of loss of AF22 traffic, which in turn MUST NOT exceed the probability of loss of AF23. Explicit Congestion Notification (ECN) [RFC3168] MAY also be used with Active Queue Management.

If RED [RFC2309] is used as an AQM algorithm, the min-threshold specifies a target queue depth for each DSCP, and the max-threshold specifies the queue depth above which all traffic with such a DSCP is dropped or ECN marked. Thus, in this service class, the following inequality should hold in queue configurations:

- o min-threshold AF23 < max-threshold AF23
- o max-threshold AF23 <= min-threshold AF22
- o min-threshold AF22 < max-threshold AF22
- o max-threshold AF22 <= min-threshold AF21

- o min-threshold AF21 < max-threshold AF21
- o max-threshold AF21 <= memory assigned to the queue

Note: This configuration tends to drop AF23 traffic before AF22 and AF22 before AF21. Many other AQM algorithms exist and are used; they should be configured to achieve a similar result.

4.7. Conversational Signaling Service Class

The Signaling service class is MUST be limited to delay-sensitive signaling traffic only, and then only applying to signaling that involves the Conversational service group. Audio signaling includes signaling between IP phone and soft-switch, soft-client and soft-switch, and media gateway and soft-switch as well as peer-to-peer using various protocols. Video and Hi-Res signaling includes video endpoint to video endpoint, as well as to Media transfer Point (MTP), to call control server(S), etc. This service class is intended to be used for control of voice and video sessions and applications. Protocols using this service class require a relatively fast response, as there are typically several messages of different sizes sent for control of the session. This service class is configured to provide good response for short-lived, intermittent flows that require real-time packet forwarding. This is not the service class for Instant Messaging (IM), that's within the bounds of the Low-Latency Data service class. The Conversational Signaling service class MUST be configured so that the probability of packet drop or significant queuing delay under peak load is very low in IP network segments that provide this interface.

The Conversational Signaling service class MUST use the new A/V-Sig PHB, defined in [ID-DSCP]. This service class MUST be configured to provide a minimum bandwidth assurance for A/V-Sig marked packets to ensure that they get forwarded. In other words, this service class MUST NOT be starved from transmission within a reasonable timeframe, given that the entire Conversational service group depends on these signaling messages successful delivery. Network engineering SHOULD be done to ensure there is roughly 1-4% available per node interface that audio and video traverse. Local conditions MUST be considered when determining exactly how much bandwidth is given to this service class. The Conversational Signaling service class SHOULD be configured to use a Rate Queuing system such as that defined in Section 1.4.1.2 of this document.

The following applications SHOULD use the Conversational Signaling service class:

- o Peer-to-peer IP telephony signaling (e.g., SIP, H.323, XMPP).
- o Peer-to-peer signaling for multimedia applications (e.g., SIP, H.323, XMPP).

- o Peer-to-peer real-time control function.
- o Client-server IP telephony signaling using H.248, MEGACO, MGCP, IP encapsulated ISDN, or other proprietary protocols.
- o Signaling to control IPTV applications using protocols such as IGMP.
- o Signaling flows between high-capacity telephony call servers or soft switches using protocol such as SIP-T. Such high-capacity devices may control thousands of telephony (VoIP) calls.
- o Signaling for one-way video flows, such as RTSP [RFC2326].
- o IGMP, when used for multicast session control such as channel changing in IPTV systems.
- o OPTIONALLY, this service class can be used for on-path reservation signaling for the traffic flows that will use the "admitted" DSCPs. The alternative is to have the on-path signaling (for reservations) use the DSCP within that service class. This provides a similar treatment of the signaling to the data flow, which might be desired.

The following are traffic characteristics:

- o Variable size packets, normally one packet at a time.
- o Intermittent traffic flows.
- o Traffic may burst at times.
- o Delay-sensitive control messages sent between two end points.

RECOMMENDED DSCP marking:

- o All flows in this service class are marked with A/V-Sig.

Applications or IP end points SHOULD pre-mark their packets with A/V-Sig DSCP value. If the end point is not capable of setting the DSCP value, then the router topologically closest to the end point SHOULD perform Multifield (MF) Classification, as defined in [RFC2475].

RECOMMENDED conditioning performed at DiffServ network edge:

- o Packet flow marking (DSCP setting) from untrusted sources (end user devices) SHOULD be verified at ingress to DiffServ network using Multifield (MF) Classification methods defined in [RFC2475].

- o Packet flows from untrusted sources (end user devices) SHOULD be policed at ingress to DiffServ network, e.g., using single rate with burst size token bucket policer to ensure that the traffic stays within its negotiated or engineered bounds.
- o Packet flows from trusted sources (application servers inside administered network) MAY not require policing.
- o Policing of packet flows across peering points in which each peer is participating in the call set-up MUST be performed to the Service Level Agreement (SLA).

The fundamental service offered to "Conversational Signaling" traffic is enhanced best-effort service with controlled rate and delay. The service SHOULD be engineered so that A/V-Sig marked packet flows have sufficient bandwidth in the network to provide high assurance of delivery and low delay. Normally, traffic in this service class does not respond dynamically to packet loss. As such, Active Queue Management [RFC2309] SHOULD NOT be applied to A/V-Sig marked packet flows.

4.8. High-Throughput Data Service Class

The High-Throughput Data service class is RECOMMENDED for elastic applications that require timely packet forwarding of variable rate traffic sources and, more specifically, is configured to provide good throughput for TCP longer-lived flows. TCP [RFC1633] or a transport with a consistent Congestion Avoidance Procedure [RFC2581] [RFC3782] normally will drive as high a data rate as it can obtain over a long period of time. The FTP protocol is a common example, although one cannot definitively say that all FTP transfers are moving data in bulk.

The High-Throughput Data service class SHOULD use the Assured Forwarding (AF) PHB, defined in [RFC2597]. This service class SHOULD be configured to provide a minimum bandwidth assurance for AF11, AF12, and AF13 marked packets to ensure that they are forwarded in a timely manner. The High-Throughput Data service class SHOULD be configured to use a Rate Queuing system such as that defined in Section 1.4.1.2 of this document.

The following applications SHOULD use the High-Throughput Data service class:

- o Store and forward applications.
- o File transfer applications (e.g., FTP, HTTP, etc).
- o Email.
- o VPN service that supports two rates (committed information rate

and excess or peak information rate).

The following are traffic characteristics:

- o Variable size packets.
- o Variable packet emission rate.
- o Variable rate.
- o With packet bursts of TCP window size.
- o Source capable of reducing its transmission rate based on detection of packet loss at the receiver or through explicit congestion notification.

Applications or IP end points SHOULD pre-mark their packets with DSCP values as shown below. If the end point is not capable of setting the DSCP value, then the router topologically closest to the end point SHOULD perform Multifield (MF) Classification, as defined in [RFC2475], and mark all packets as AF1x. Note: In this case, the two-rate, three-color marker will be configured to operate in Color-Blind mode.

RECOMMENDED DSCP marking:

- o AF11 = up to specified rate "A".
- o AF12 = in excess of specified rate "A" but below specified rate "B".
- o AF13 = in excess of specified rate "B".
- o Where "A" < "B".

RECOMMENDED conditioning performed at DiffServ network edge:

- o The two-rate, three-color marker SHOULD be configured to provide the behavior as defined in trTCM [RFC2698].
- o If packets are marked by trusted sources or a previously trusted DiffServ domain and the color marking is to be preserved, then the two-rate, three-color marker SHOULD be configured to operate in Color-Aware mode.
- o If the packet marking is not trusted or the color marking is not to be preserved, then the two-rate, three-color marker SHOULD be configured to operate in Color-Blind mode.

The fundamental service offered to "High-Throughput Data" traffic is enhanced best-effort service with a specified minimum rate. The service SHOULD be engineered so that AF11 marked packet flows have

sufficient bandwidth in the network to provide assured delivery. It can be assumed that this class will consume any available bandwidth and that packets traversing congested links may experience higher queuing delays or packet loss. Since the AF1x traffic is elastic and responds dynamically to packet loss, Active Queue Management [RFC2309] SHOULD be used primarily to control TCP flow rates at congestion points by dropping packets from TCP flows that have higher rates first. The probability of loss of AF11 traffic MUST NOT exceed the probability of loss of AF12 traffic, which in turn MUST NOT exceed the probability of loss of AF13. In such a case, if one network customer is driving significant excess and another seeks to use the link, any losses will be experienced by the high-rate user, causing him to reduce his rate. Explicit Congestion Notification (ECN) [RFC3168] MAY also be used with Active Queue Management.

If RED [RFC2309] is used as an AQM algorithm, the min-threshold specifies a target queue depth for each DSCP, and the max-threshold specifies the queue depth above which all traffic with such a DSCP is dropped or ECN marked. Thus, in this service class, the following inequality should hold in queue configurations:

- o min-threshold AF13 < max-threshold AF13
- o max-threshold AF13 <= min-threshold AF12
- o min-threshold AF12 < max-threshold AF12
- o max-threshold AF12 <= min-threshold AF11
- o min-threshold AF11 < max-threshold AF11
- o max-threshold AF11 <= memory assigned to the queue

Note: This configuration tends to drop AF13 traffic before AF12 and AF12 before AF11. Many other AQM algorithms exist and are used; they should be configured to achieve a similar result.

4.9. Standard Service Class

The Standard service class is RECOMMENDED for traffic that has not been classified into one of the other supported forwarding service classes in the DiffServ network domain. This service class provides the Internet's "best-effort" forwarding behavior. This service class typically has minimum bandwidth guarantee.

The Standard service class MUST use the Default Forwarding (DF) PHB, defined in [RFC2474], and SHOULD be configured to receive at least a small percentage of forwarding resources as a guaranteed minimum. This service class SHOULD be configured to use a Rate Queuing system such as that defined in Section 1.4.1.2 of this document.

The following applications SHOULD use the Standard service class:

- o Network services, DNS, DHCP, BootP.
- o Any undifferentiated application/packet flow transported through the DiffServ enabled network.

The following is a traffic characteristic:

- o Non-deterministic, mixture of everything.

The RECOMMENDED DSCP marking is DF (Default Forwarding) '000000'.

Network Edge Conditioning:

There is no requirement that conditioning of packet flows be performed for this service class.

The fundamental service offered to the Standard service class is best-effort service with active queue management to limit overall delay. Typical configurations SHOULD use random packet dropping to implement Active Queue Management [RFC2309] or Explicit Congestion Notification [RFC3168], and MAY impose a minimum or maximum rate on the queue.

If RED [RFC2309] is used as an AQM algorithm, the min-threshold specifies a target queue depth, and the max-threshold specifies the queue depth above which all traffic is dropped or ECN marked. Thus, in this service class, the following inequality should hold in queue configurations:

- o min-threshold DF < max-threshold DF
- o max-threshold DF <= memory assigned to the queue

Note: Many other AQM algorithms exist and are used; they should be configured to achieve a similar result.

4.10. Low-Priority Data

The Low-Priority Data service class serves applications that run over TCP [RFC0793] or a transport with consistent congestion avoidance procedures [RFC2581] [RFC3782] and that the user is willing to accept service without guarantees. This service class is specified in [RFC3662] and [QBSS].

The following applications MAY use the Low-Priority Data service class:

- o Any TCP based-application/packet flow transported through the DiffServ enabled network that does not require any bandwidth assurances.

The following is a traffic characteristic:

- o Non-real-time and elastic.

Network Edge Conditioning:

There is no requirement that conditioning of packet flows be performed for this service class.

The RECOMMENDED DSCP marking is CS1 (Class Selector 1).

The fundamental service offered to the Low-Priority Data service class is best-effort service with zero bandwidth assurance. By placing it into a separate queue or class, it may be treated in a manner consistent with a specific Service Level Agreement.

Typical configurations SHOULD use Explicit Congestion Notification [RFC3168] or random loss to implement Active Queue Management [RFC2309].

If RED [RFC2309] is used as an AQM algorithm, the min-threshold specifies a target queue depth, and the max-threshold specifies the queue depth above which all traffic is dropped or ECN marked. Thus, in this service class, the following inequality should hold in queue configurations:

- o min-threshold CS1 < max-threshold CS1
- o max-threshold CS1 <= memory assigned to the queue

Note: Many other AQM algorithms exist and are used; they should be configured to achieve a similar result.

5. Additional Information on Service Class Usage

In this section, we provide additional information on how some specific applications should be configured to use the defined service classes.

5.1. Mapping for NTP

From tests that were performed, indications are that precise time distribution requires a very low packet delay variation (jitter) transport. Therefore, we suggest that the following guidelines for Network Time Protocol (NTP) be used:

- o When NTP is used for providing high-accuracy timing within an administrator's (carrier's) network or to end users/clients, the audio service class SHOULD be used, and NTP packets should be marked with EF DSCP value.

- o For applications that require "wall clock" timing accuracy, the Standard service class should be used, and packets should be marked with DF DSCP.

5.2. VPN Service Mapping

"Differentiated Services and Tunnels" [RFC2983] considers the interaction of DiffServ architecture with IP tunnels of various forms. Further to guidelines provided in RFC 2983, below are additional guidelines for mapping service classes that are supported in one part of the network into a VPN connection. This discussion is limited to VPNs that use DiffServ technology for traffic differentiation.

- o The DSCP value(s) that is/are used to represent a PHB or a PHB group SHOULD be the same for the networks at both ends of the VPN tunnel, unless remarking of DSCP is done as ingress/egress processing function of the tunnel. DSCP marking needs to be preserved along the tunnel, end to end.
- o The VPN MAY be configured to support one or more service classes. It is left up to the administrators of the two networks to agree on the level of traffic differentiation that will be provided in the network that supports VPN service. Service classes are then mapped into the supported VPN traffic forwarding behaviors that meet the traffic characteristics and performance requirements of the encapsulated service classes.
- o The traffic treatment in the network that is providing the VPN service needs to be such that the encapsulated service class or classes receive comparable behavior and performance in terms of delay, jitter, and packet loss and that they are within the limits of the service specified.
- o The DSCP value in the external header of the packet forwarded through the network providing the VPN service can be different from the DSCP value that is used end to end for service differentiation in the end network.
- o The guidelines for aggregation of two or more service classes into a single traffic forwarding treatment in the network that is providing the VPN service is for further study.

6. Security Considerations

This document discusses policy and describes a common policy configuration, for the use of a Differentiated Services Code Point by transports and applications. If implemented as described, it should require that the network do nothing that the network has not already allowed. If that is the case, no new security issues should arise from the use of such a policy.

It is possible for the policy to be applied incorrectly, or for a wrong policy to be applied in the network for the defined service class. In that case, a policy issue exists that the network SHOULD detect, assess, and deal with. This is a known security issue in any network dependent on policy-directed behavior.

A well-known flaw appears when bandwidth is reserved or enabled for a service (for example, voice and/or video transport) and another service or an attacking traffic stream uses it. This possibility is inherent in DiffServ technology, which depends on appropriate packet markings. When bandwidth reservation or a priority queuing system is used in a vulnerable network, the use of authentication and flow admission is recommended. To the author's knowledge, there is no known technical way to respond to an unauthenticated data stream using service that it is not intended to use, and such is the nature of the Internet.

The use of a service class by a user is not an issue when the SLA between the user and the network permits him to use it, or to use it up to a stated rate. In such cases, simple policing is used in the Differentiated Services Architecture. Some service classes, such as Network Control, are not permitted to be used by users at all; such traffic should be dropped or remarked by ingress filters. Where service classes are available under the SLA only to an authenticated user rather than to the entire population of users, authentication and authorization services are required, such as those surveyed in [AUTHMECH].

7. Contributing Authors

This section specifically calls out the authors of RFC 4594, from which this document is based on.

Jozef Babiarez
Nortel Networks

Kwok Ho Chan
Nortel Networks
Email: khchan.work@gmail.com

Fred Baker
Cisco Systems
EMail: fred@cisco.com

Of note, two of the three mentioned authors above worked for Nortel Networks at the time of writing RFC 4594, a company that no longer exists. This author has not seen or heard from those two in many, many years or IETF meetings - as a result of not knowing their new email addresses (or phone numbers).

While much of this document has been rewritten with either edited or

brand new material, there are many short paragraphs that remain as they were from RFC 4594, as well as many sentences that were also left unchanged. Additionally, there were no new graphs, charts, diagrams, or tables introduced, meaning the first 4 tables within this document existed in RFC 4594, created by those authors. Presently, each of those tables contain modified and new information. The last 3 tables, specifically tables 5, 6, & 7 were removed because the examples section was removed.

This author believes there must be proper credit given for all the contributions, including the framework this document retains from that RFC. Periodically, throughout this document, what was written remains the best way of conveying a thought, rule, or otherwise stated behavior or mechanism. Because RFC 4594 was rather large, there is no realistic way of identifying each part that was left untouched. Further, properly quoting that RFC and leaving those sentences embedded in this document would render this document highly unreadable. Another application could be used to show the changes, deletions and additions - but not one that the IETF accepts presently.

This author has created this "Contributing Authors" section as a way of properly identifying those 3 individuals that provided text within this document. We will let the community judge if this is 'good enough' (i.e., rough consensus), or if another way is better.

8. Acknowledgements

The author would like to thank Paul Jones, Glen Lavers, Mo Zanaty, David Benham, Michael Ramalho, Gorrry Fairhurst, David Black, Brian Carpenter, Al Morton, Ruediger Geib and Shitanshu Shah for their comments and questions about this effort that ultimately helped shape this document.

Below are the folks that were acknowledged in RFC 4594, and this author does not want to lose their recognition of contributions to the original effort.

"The authors thank the TSVWG reviewers, David Black, Brian E. Carpenter, and Alan O'Neill for their review and input to this document.

The authors acknowledge a great many inputs, most notably from Bruce Davie, Dave Oran, Ralph Santitoro, Gary Kenward, Francois Audet, Morgan Littlewood, Robert Milne, John Shuler, Nalin Mistry, Al Morton, Mike Pierce, Ed Koehler Jr., Tim Rahrer, Fil Dickinson, Mike Fidler, and Shane Amante. Kimberly King, Joe Zebarth, and Alistair Munroe each did a thorough proofreading,

and the document is better for their contributions."

9. References

9.1. Normative References

- [RFC0791] Postel, J., "Internet Protocol", STD 5, RFC 791, September 1981.
- [RFC0793] Postel, J., "Transmission Control Protocol", STD 7, RFC 793, September 1981.
- [RFC1349] Almquist, P., "Type of Service in the Internet Protocol Suite", RFC 1349, July 1992.
- [RFC1812] Baker, F., "Requirements for IP Version 4 Routers", RFC 1812, June 1995.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2309] Braden, B., Clark, D., Crowcroft, J., Davie, B., Deering, S., Estrin, D., Floyd, S., Jacobson, V., Minshall, G., Partridge, C., Peterson, L., Ramakrishnan, K., Shenker, S., Wroclawski, J., and L. Zhang, "Recommendations on Queue Management and Congestion Avoidance in the Internet", RFC 2309, April 1998.
- [RFC2474] Nichols, K., Blake, S., Baker, F., and D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC 2474, December 1998.
- [RFC2475] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., and W. Weiss, "An Architecture for Differentiated Service", RFC 2475, December 1998.
- [RFC2597] Heinanen, J., Baker, F., Weiss, W., and J. Wroclawski, "Assured Forwarding PHB Group", RFC 2597, June 1999.
- [RFC3246] Davie, B., Charny, A., Bennet, J.C., Benson, K., Le Boudec, J., Courtney, W., Davari, S., Firoiu, V., and D. Stiliadis, "An Expedited Forwarding PHB (Per-Hop Behavior)", RFC 3246, March 2002.
- [RFC3550] Schulzrinne, H., Casner, S., Frederick, R., and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications", STD 64, RFC 3550, July 2003.
- [RFC3662] Bless, R., Nichols, K., and K. Wehrle, "A Lower Effort Per-Domain Behavior (PDB) for Differentiated Services",

RFC 3662, December 2003.

- [RFC5865] F. Baker, J. Polk, M. Dolly, "A Differentiated Services Code Point (DSCP) for Capacity-Admitted Traffic", RFC 5865, May 2010

9.2. Informative References

- [AUTHMECH] Rescorla, E., "A Survey of Authentication Mechanisms", Work in Progress, September 2005.
- [QBSS] "QBone Scavenger Service (QBSS) Definition", Internet2 Technical Report Proposed Service Definition, March 2001.
- [IEEE1Q] IEEE, 802.1Q Specification
- [IEEE1E] IEEE, 802.1E Wireless LAN User Priority Specification
- [RFC1633] Braden, R., Clark, D., and S. Shenker, "Integrated Services in the Internet Architecture: an Overview", RFC 1633, June 1994.
- [RFC2205] Braden, R., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, September 1997.
- [RFC2581] Allman, M., Paxson, V., and W. Stevens, "TCP Congestion Control", RFC 2581, April 1999.
- [RFC2697] Heinanen, J. and R. Guerin, "A Single Rate Three Color Marker", RFC 2697, September 1999.
- [RFC2698] Heinanen, J. and R. Guerin, "A Two Rate Three Color Marker", RFC 2698, September 1999.
- [RFC2963] Bonaventure, O. and S. De Cnodder, "A Rate Adaptive Shaper for Differentiated Services", RFC 2963, October 2000.
- [RFC2983] Black, D., "Differentiated Services and Tunnels", RFC 2983, October 2000.
- [RFC2996] Bernet, Y., "Format of the RSVP DCLASS Object", RFC 2996, November 2000.
- [RFC3086] Nichols, K. and B. Carpenter, "Definition of Differentiated Services Per Domain Behaviors and Rules for their Specification", RFC 3086, April 2001.
- [RFC3168] Ramakrishnan, K., Floyd, S., and D. Black, "The Addition of Explicit Congestion Notification (ECN) to IP", RFC

3168, September 2001.

- [RFC3175] Baker, F., Iturralde, C., Le Faucheur, F., and B. Davie, "Aggregation of RSVP for IPv4 and IPv6 Reservations", RFC 3175, September 2001.
- [RFC3290] Bernet, Y., Blake, S., Grossman, D., and A. Smith, "An Informal Management Model for Diffserv Routers", RFC 3290, May 2002.
- [RFC3782] Floyd, S., Henderson, T., and A. Gurtov, "The NewReno Modification to TCP's Fast Recovery Algorithm", RFC 3782, April 2004.
- [RFC5462] L. Andersson, R. Asati, "Multiprotocol Label Switching (MPLS) Label Stack Entry: EXP Field Renamed to Traffic Class Field", RFC 5462, February 2009

Authors' Address

James Polk
3913 Treemont Circle
Colleyville, Texas 76034

Phone: +1.817.271.3552
Email: jmpolk@cisco.com

Appendix A - Changes

Here is a list of all the changes that were captured during the editing process. This will not be a complete list, and others are free to point out what the authors missed, and we'll include that in the next release.

A.1 Since Individual -02 to -03

- o Inserted section 1.6 to explain fundamentally what has changed since RFC 4594, and why changes are necessary.

A.2 Since Individual -01 to -02

- o Added text to the Intro section on the justification from DiffServ Problem Statement draft, as to more of why this update is necessary.
- o Added text to the Intro section expanding on the concept of service classes vs. treatment aggregates (from RFC 5127).

A.3 Since Individual -00 to -01

- o Added Section 2.4 which covers the conflation issues regarding the differences between service classes and treatment aggregates.
- o Added example operational configurations of treatment aggregates applied to this draft's new set of service classes and additional DSCPs.
- o Added references RFC 5865, RFC 5462, IEEE 802.1E and IEEE 802.1Q.

A.4 Since RFC 4594 to Individual Update -00

- o rewrote Intro to emphasize current topics
- o Created a Conversational Service group, comprising the audio, video and Hi-Res service classes - because they have similar characteristics.
- o Incorporated the 6 new DSCPs from [ID-DSCP].
- o moved the example section, en mass, to an appendix that might not be kept for this version. We're not sure it accomplishes what it needs to, and might not provide any real usefulness.
- o Moved 'Realtime-Interactive' service class to CS5, from CS4
- o Changed 'Broadcast Video' service class to 'Broadcast' service class
- o Changed AF4X to 'Video' service class, replacing 'Multimedia Conferencing' service class
- o Moved 'Multimedia Conferencing' service class to different DSCPs
- o Added the 'Hi-Res' service class
- o Removed section 5.1 on signaling choices. It has been included in the main body of the text.
- o Changed document title
- o ...

Network WG
Internet-Draft
Expires: January 16, 2013
Intended Status: Standards Track
Updates: RFC 2872 (if accepted)

James Polk
Subha Dhesikan
Cisco Systems
July 16, 2012

Resource Reservation Protocol (RSVP) Application-ID
Profiles for Voice and Video Streams
draft-polk-tsvwg-rsvp-app-id-vv-profiles-04

Abstract

RFC 2872 defines an Resource Reservation Protocol (RSVP) object for application identifiers. This document uses that App-ID and gives implementers specific guidelines for differing voice and video stream identifications to nodes along a reservation path, creating specific profiles for voice and video session identification.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on January 16, 2013.

Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. Application ID Template	3
3. The Voice and Video Application-ID Profiles	4
3.1 The Broadcast video Profile	4
3.2 The Real-time Interactive Profile	5
3.3 The Multimedia Conferencing Profile	5
3.4 The Multimedia Streaming Profile	6
3.5 The Conversational Profile	6
4. Security considerations	7
5. IANA considerations	7
5.1 New RSVP Policy Element (P-Type)	7
5.2 Application Profiles	7
5.2.1 Broadcast Profiles IANA Registry	8
5.2.2 Realtime-Interactive Profiles IANA Registry	8
5.2.3 Multimedia-Conferencing Profiles IANA Registry	9
5.2.4 Multimedia-Streaming Profiles IANA Registry	10
5.2.5 Conversational Profiles IANA Registry	10
6. Acknowledgments	12
7. References	12
7.1. Normative References	12
7.2. Informative References	13
Authors' Addresses	13
Appendix	14

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC 2119].

1. Introduction

RFC 2872 [RFC2872] describes the usage of policy elements for providing application information in Resource Reservation Protocol (RSVP) signaling [RFC2205]. The intention of providing this information is to enable application-based policy control. However, RFC 2872 does not enumerate any application profiles. The absence of explicit, uniform profiles leads to incompatible handling of these values and misapplied policies. An application profile used by a sender might not be understood by the intermediaries or receiver in a different domain. Therefore, there is a need to enumerate application profiles that are universally understood and applied for correct policy control.

Call control between endpoints has the ability to bind or associate many attributes to a reservation. One new attribute currently being defined is to establish the type of traffic contained that reservation. This is accomplished via assigning a traffic label to

the call (or session or flow) [ID-TRAF-CLASS].

This document takes the application traffic classes from [ID-TRAF-CLASS] and places those strings in the APP-ID object defined in RFC 2872. Thus, the intermediary devices (e.g., routers) processing the RSVP message can learn the identified profile within the Application-ID policy element for a particular reservation, and possibly be configured with the profile(s) to understand them correctly, thus performing the correct admission control.

Another goal of this document is to the ability to signal an application profile which can then be translated into a DSCP value as per the choice of each domain. While the DCLASS object [RFC2996] allows the transfer of DSCP value in an RSVP message, it does not allow the flexibility of having different domains choosing the DSCP value for the traffic classes that that they maintain.

How these labels indicate the appropriate Differentiated Services Codepoint (DSCP) is out of scope for this document.

This document will break out each application type and propose how the values in application-id template should be populated for uniformity and interoperability.

2. Application ID Template

The template from RFC 2872 is as follows:

0	1	2	3
PE Length (8)		P-type = AUTH_APP	
Attribute Length		A-type = POLICY_LOCATOR	Sub-type = ASCII_DN
Application name as ASCII string (e.g. SAP.EXE)			

In line with how this policy element is constructed in RFC 2872, the A-type will remain "POLICY_LOCATOR".

The P-type field is first created in [RFC2752]. This document creates the new P-type "APP_TC" for application traffic class, which is more appropriately named for the purpose described in this extension.

The first Sub-type will be mandatory for every profile within this document, and will be "ASCII_DN". No other Sub-types are defined by

any profile within this document, but MAY be included by individual implementations - and MUST be ignored if not understood by receiving implementations along the reservation path.

RFC 2872 states the #1 sub-element from RFC 2872 as the "identifier that uniquely identifies the application vendor", which is optional to include. This document modifies this vendor limitation so that the identifier need only be unique - and not limited to an application vendor (identifier). For example, this specification now allows an RFC that defines an industry recognizable term or string to be a valid identifier. For example, a term or string taken from another IETF document, such as "conversational" or "avconf" from [ID-TRAF-CLASS]. This sub-element is still optional to include.

The following subsections will define the values within the above template into specific profiles for voice and video identification.

3. The Voice and Video Application-ID Profiles

This section contains the elements of the Application ID policy object which is used to signal the application classes defined in [ID-TRAF-CLASS].

3.1 The Broadcast Profiles

Broadcast profiles are for minimally buffered one-way streaming flows, such as video surveillance, or Internet based concerts or non-VOD TV broadcasts such as live sporting events.

There will be Broadcast profiles for

- Broadcast IPTV for audio and video
- Broadcast Live-events for audio and video
- Broadcast Surveillance for audio and video

Here is an example profile for identifying Broadcast Video-Surveillance

```
APP_TC, POLICY_LOCATOR, ASCII_DN,  
"GUID=http://www.ietf.org/internet-drafts/  
    draft-ietf-mmusic-traffic-class-for-sdp-01.txt,  
APP=broadcast.video.surveillance, VER="
```

Where the Globally Unique Identifier (GUID) indicates the documented reference that created this well known string [ID-TRAF-CLASS], the APP is the profile name with no spaces, and the "VER=" is included, but has no value at this time.

3.2 The Realtime Interactive Profiles

Realtime Interactive profiles are for on-line gaming, and both remote and virtual avconf applications, in which the timing is particularly important towards the feedback to uses of these applications. This traffic type will generally not be UDP based, with minimal tolerance to RTT delays.

There will be Realtime Interactive profiles for

- Realtime-Interactive Gaming
- Realtime-Interactive Remote-Desktop
- Realtime-Interactive Virtualized-Desktop

Here is the profile for identifying Realtime-Interactive Gaming

```
APP_TC, POLICY_LOCATOR, ASCII_DN,  
"GUID=http://www.ietf.org/internet-drafts/  
    draft-ietf-mmusic-traffic-class-for-sdp-01.txt,  
APP=realtime-interactive.gaming, VER="
```

Where the Globally Unique Identifier (GUID) indicates the documented reference that created this well known string [ID-TRAF-CLASS], the APP is the profile name with no spaces, and the "VER=" is included, but has no value, but MAY if versioning becomes important.

3.3 The Multimedia Conferencing Profiles

There will be Multimedia Conferencing profiles for presentation data, application sharing and whiteboarding, where these applications will most often be associated with a larger Conversational (audio and/or audio/video) conference. Timing is important, but some minimal delays are acceptable, unlike the case for Realtime-Interactive traffic.

- Multimedia-Conferencing presentation-data
- Multimedia-Conferencing application-sharing
- Multimedia-Conferencing whiteboarding

Here is the profile for identifying Multimedia-Conferencing Application-sharing

```
APP_TC, POLICY_LOCATOR, ASCII_DN,  
"GUID=http://www.ietf.org/internet-drafts/  
    draft-ietf-mmusic-traffic-class-for-sdp-01.txt,  
APP=multimedia-conferencing.application-sharing, VER="
```

Where the Globally Unique Identifier (GUID) indicates the RFC reference that created this well known string [ID-TRAF-CLASS], the APP is the profile name with no spaces, and the "VER=" is included, but has no value, but MAY if versioning becomes important.

3.4 The Multimedia Streaming Profiles

Multimedia Streaming profiles are for more significantly buffered one-way streaming flows than Broadcast profiles. These include...

There will be Multimedia Streaming profiles for

- Multimedia-Streaming multiplex
- Multimedia-Streaming webcast

Here is the profile for identifying Multimedia Streaming webcast

```
APP_TC, POLICY_LOCATOR, ASCII_DN,  
"GUID=http://www.ietf.org/internet-drafts/  
    draft-ietf-mmusic-traffic-class-for-sdp-01.txt,  
APP=multimedia-streaming.webcast, VER="
```

Where the Globally Unique Identifier (GUID) indicates the documented reference that created this well known string [ID-TRAF-CLASS], the APP is the profile name with no spaces, and the "VER=" is included, but has no value, but MAY if versioning becomes important.

3.5 The Conversational Profiles

Conversational category is for realtime bidirectional communications, such as voice or video, and is the most numerous due to the choices of application with or without adjectives. The number of profiles is then doubled because there needs to be one for unadmitted and one for admitted. The IANA section lists all that are currently proposed for registration at this time, therefore there will not be an exhaustive list provided in this section.

There will be conversational profiles for

- Conversational Audio
- Conversational Audio Admitted
- Conversational Video
- Conversational Video Admitted
- Conversational Audio Avconf
- Conversational Audio Avconf Admitted
- Conversational Video Avconf
- Conversational Video Avconf Admitted
- Conversational Audio Immersive
- Conversational Audio Immersive Admitted
- Conversational Video Immersive
- Conversational Video Immersive Admitted

Here is an example profile for identifying Conversational Audio:

```
APP_TC, POLICY_LOCATOR, ASCII_DN,  
"GUID=http://www.ietf.org/internet-drafts/  
    draft-ietf-mmusic-traffic-class-for-sdp-01.txt,  
APP=conversational.audio, VER="
```

Where the Globally Unique Identifier (GUID) indicates the documented reference that created this well known string [ID-TRAF-CLASS], the APP is the profile name with no spaces, and the "VER=" is included, but has no value, but MAY if versioning becomes important.

4. Security considerations

The security considerations section within RFC 2872 sufficiently covers this document, with one possible exception - someone using the wrong template values (e.g., claiming a reservation is Multimedia Streaming when it is in fact Real-time Interactive). Given that each traffic flow is within separate reservations, and RSVP does not have the ability to police the type of traffic within any reservation, solving for this appears to be administratively handled at best. This is not meant to be a 'punt', but there really is nothing this template creates that is going to make things any harder for anyone (that we know of now).

5. IANA considerations

5.1 New RSVP Policy Element (P-Type)

In line with the convention created in RFC 3182, the following P-Type is created in the RSVP Policy Element registry [TBD]:

4	APP_TC	Traffic Class identification of applications
---	--------	--

[Editor's note: Unfortunately, RFC 2750 specified the creation of the "RSVP Policy Element" IANA registration, which does not appear at the <http://www.iana.org/assignments/rsvp-parameters> page, therefore it appears this registry does not yet exist. We will get with the chairs to work on this.]

5.2 Application Profiles

This document requests IANA create a new registry for the application identification classes similar to the following table within the Resource Reservation Protocol (RSVP) Parameters registry:

Registry Name: RSVP APP-ID Profiles
Reference: [this document]
Registration procedures: Standards Track document [RFC5226]

5.2.1 Broadcast Profiles IANA Registry

Broadcast Audio IPTV Profile

P-type = APP_TC
A-type = POLICY_LOCATOR
Sub-type = ASCII_DN
Conformant policy locator =
"GUID=http://www.ietf.org/internet-drafts/
draft-ietf-mmusic-traffic-class-for-sdp-01.txt,
APP=broadcast.audio.iptv, VER="

Reference: [this document]

Broadcast Video IPTV Profile

P-type = APP_TC
A-type = POLICY_LOCATOR
Sub-type = ASCII_DN
Conformant policy locator =
"GUID=http://www.ietf.org/internet-drafts/
draft-ietf-mmusic-traffic-class-for-sdp-01.txt,
APP=broadcast.video.iptv, VER="

Reference: [this document]

Broadcast Audio Live-events Profile

P-type = APP_TC
A-type = POLICY_LOCATOR
Sub-type = ASCII_DN
Conformant policy locator =
"GUID=http://www.ietf.org/internet-drafts/
draft-ietf-mmusic-traffic-class-for-sdp-01.txt,
APP=broadcast.audio.live-events, VER="

Reference: [this document]

Broadcast Audio Live-events Profile

P-type = APP_TC
A-type = POLICY_LOCATOR
Sub-type = ASCII_DN
Conformant policy locator =
"GUID=http://www.ietf.org/internet-drafts/
draft-ietf-mmusic-traffic-class-for-sdp-01.txt,
APP=broadcast.video.live-events, VER="

Reference: [this document]

Broadcast Audio-Surveillance Profile

P-type = APP_TC
A-type = POLICY_LOCATOR
Sub-type = ASCII_DN
Conformant policy locator =
"GUID=http://www.ietf.org/internet-drafts/
draft-ietf-mmusic-traffic-class-for-sdp-01.txt,
APP=broadcast.audio.surveillance, VER="

Reference: [this document]

Broadcast Video-Surveillance Profile

P-type = APP_TC

A-type = POLICY_LOCATOR

Sub-type = ASCII_DN

Conformant policy locator =

"GUID=http://www.ietf.org/internet-drafts/
draft-ietf-mmusic-traffic-class-for-sdp-01.txt,
APP=broadcast.video.surveillance, VER="

Reference: [this document]

5.2.2 Realtime-Interactive Profiles IANA Registry

Realtime-Interactive Gaming Profile

P-type = APP_TC

A-type = POLICY_LOCATOR

Sub-type = ASCII_DN

Conformant policy locator =

"GUID=http://www.ietf.org/internet-drafts/
draft-ietf-mmusic-traffic-class-for-sdp-01.txt,
APP= realtime-interactive.gaming, VER="

Reference: [this document]

Real-time Interactive Remote-Desktop Profile

P-type = APP_TC

A-type = POLICY_LOCATOR

Sub-type = ASCII_DN

Conformant policy locator =

"GUID=http://www.ietf.org/internet-drafts/
draft-ietf-mmusic-traffic-class-for-sdp-01.txt,
APP=realtime-interactive.remote-desktop, VER="

Reference: [this document]

Real-time Interactive Virtualized-Desktop Profile

P-type = APP_TC

A-type = POLICY_LOCATOR

Sub-type = ASCII_DN

Conformant policy locator =

"GUID=http://www.ietf.org/internet-drafts/
draft-ietf-mmusic-traffic-class-for-sdp-01.txt,
APP=realtime-interactive.virtualized-desktop,
VER="

Reference: [this document]

5.2.3 Multimedia-Conferencing Profiles IANA Registry

Multimedia-Conferencing Presentation-Data Profile

P-type = APP_TC

A-type = POLICY_LOCATOR

Sub-type = ASCII_DN
Conformant policy locator =
 "GUID=http://www.ietf.org/internet-drafts/
 draft-ietf-mmusic-traffic-class-for-sdp-01.txt,
 APP= multimedia-conferencing.presentation-data,
 VER="

Reference: [this document]

Multimedia-Conferencing Application-Sharing Profile

P-type = APP_TC
A-type = POLICY_LOCATOR
Sub-type = ASCII_DN
Conformant policy locator =
 "GUID=http://www.ietf.org/internet-drafts/
 draft-ietf-mmusic-traffic-class-for-sdp-01.txt,
 APP= multimedia-conferencing.application-sharing,
 VER="

Reference: [this document]

Multimedia-Conferencing Whiteboarding Profile

P-type = APP_TC
A-type = POLICY_LOCATOR
Sub-type = ASCII_DN
Conformant policy locator =
 "GUID=http://www.ietf.org/internet-drafts/
 draft-ietf-mmusic-traffic-class-for-sdp-01.txt,
 APP= multimedia-conferencing.whiteboarding, VER="

Reference: [this document]

5.2.4 Multimedia-Streaming Profiles IANA Registry

Multimedia-Streaming Multiplex Profile

P-type = APP_TC
A-type = POLICY_LOCATOR
Sub-type = ASCII_DN
Conformant policy locator =
 "GUID=http://www.ietf.org/internet-drafts/
 draft-ietf-mmusic-traffic-class-for-sdp-01.txt,
 APP=multimedia-streaming.multiplex, VER="

Reference: [this document]

Multimedia-Streaming Webcast Profile

P-type = APP_TC
A-type = POLICY_LOCATOR
Sub-type = ASCII_DN
Conformant policy locator =
 "GUID=http://www.ietf.org/internet-drafts/
 draft-ietf-mmusic-traffic-class-for-sdp-01.txt,
 APP=multimedia-streaming.webcast, VER="

Reference: [this document]

5.2.5 Conversational Profiles IANA Registry

Conversational Audio Profile

P-type = APP_TC

A-type = POLICY_LOCATOR

Sub-type = ASCII_DN

Conformant policy locator =

"GUID=http://www.ietf.org/internet-drafts/
draft-ietf-mmusic-traffic-class-for-sdp-01.txt,
APP=conversational.audio, VER="

Reference: [this document]

Conversational Audio Admitted Profile

P-type = APP_TC

A-type = POLICY_LOCATOR

Sub-type = ASCII_DN

Conformant policy locator =

"GUID=http://www.ietf.org/internet-drafts/
draft-ietf-mmusic-traffic-class-for-sdp-01.txt,
APP=conversational.audio.aq:admitted, VER="

Reference: [this document]

Conversational Video Profile

P-type = APP_TC

A-type = POLICY_LOCATOR

Sub-type = ASCII_DN

Conformant policy locator =

"GUID=http://www.ietf.org/internet-drafts/
draft-ietf-mmusic-traffic-class-for-sdp-01.txt,
APP=conversational.video, VER="

Reference: [this document]

Conversational Video Admitted Profile

P-type = APP_TC

A-type = POLICY_LOCATOR

Sub-type = ASCII_DN

Conformant policy locator =

"GUID=http://www.ietf.org/internet-drafts/
draft-ietf-mmusic-traffic-class-for-sdp-01.txt,
APP=conversational.video.aq:admitted, VER="

Reference: [this document]

Conversational Audio Avconf Profile

P-type = APP_TC

A-type = POLICY_LOCATOR

Sub-type = ASCII_DN

Conformant policy locator =

"GUID=http://www.ietf.org/internet-drafts/
draft-ietf-mmusic-traffic-class-for-sdp-01.txt,
APP=conversational.audio.avconf, VER="

Reference: [this document]

Conversational Audio Avconf Admitted Profile

P-type = APP_TC
A-type = POLICY_LOCATOR
Sub-type = ASCII_DN
Conformant policy locator =
"GUID=http://www.ietf.org/internet-drafts/
draft-ietf-mmusic-traffic-class-for-sdp-01.txt,
APP=conversational.audio.avconf.aq:admitted,
VER="

Reference: [this document]

Conversational Video Avconf Profile

P-type = APP_TC
A-type = POLICY_LOCATOR
Sub-type = ASCII_DN
Conformant policy locator =
"GUID=http://www.ietf.org/internet-drafts/
draft-ietf-mmusic-traffic-class-for-sdp-01.txt,
APP=conversational.video.avconf, VER="

Reference: [this document]

Conversational Video Avconf Admitted Profile

P-type = APP_TC
A-type = POLICY_LOCATOR
Sub-type = ASCII_DN
Conformant policy locator =
"GUID=http://www.ietf.org/internet-drafts/
draft-ietf-mmusic-traffic-class-for-sdp-01.txt,
APP=conversational.video.avconf.aq:admitted,
VER="

Reference: [this document]

Conversational Audio Immersive Profile

P-type = APP_TC
A-type = POLICY_LOCATOR
Sub-type = ASCII_DN
Conformant policy locator =
"GUID=http://www.ietf.org/internet-drafts/
draft-ietf-mmusic-traffic-class-for-sdp-01.txt,
APP=conversational.audio.immersive, VER="

Reference: [this document]

Conversational Audio Immersive Admitted Profile

P-type = APP_TC
A-type = POLICY_LOCATOR
Sub-type = ASCII_DN
Conformant policy locator =
"GUID=http://www.ietf.org/internet-drafts/
draft-ietf-mmusic-traffic-class-for-sdp-01.txt,
APP=conversational.audio.immersive.aq:admitted,
VER="

Reference: [this document]

Conversational Video Immersive Profile

P-type = APP_TC

A-type = POLICY_LOCATOR

Sub-type = ASCII_DN

Conformant policy locator =

"GUID=http://www.ietf.org/internet-drafts/
draft-ietf-mmusic-traffic-class-for-sdp-01.txt,
APP=conversational.video.immersive, VER="

Reference: [this document]

Conversational Video Immersive Admitted Profile

P-type = APP_TC

A-type = POLICY_LOCATOR

Sub-type = ASCII_DN

Conformant policy locator =

"GUID=http://www.ietf.org/internet-drafts/
draft-ietf-mmusic-traffic-class-for-sdp-01.txt,
APP=conversational.video.immersive.ag:admitted,
VER="

Reference: [this document]

7. Acknowledgments

To Francois Le Faucheur, Paul Jones and Glen Lavers for their helpful comments and encouragement.

8. References

8.1. Normative References

- [RFC2119] S. Bradner, "Key words for use in RFCs to Indicate Requirement Levels", RFC 2119, March 1997
- [RFC2205] R. Braden, Ed., L. Zhang, S. Berson, S. Herzog, S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, September 1997
- [RFC2474] K. Nichols, S. Blake, F. Baker, D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers ", RFC 2474, December 1998
- [RFC2750] S. Herzog, "RSVP Extensions for Policy Control", RFC 2750, January 2000
- [RFC2872] Y. Bernet, R. Pabbati, "Application and Sub Application Identity Policy Element for Use with RSVP", RFC 2872, June 2000

- [RFC2996] Y. Bernet, "Format of the RSVP DCLASS Object", RFC 2996, November 2000
- [RFC3182] S. Yadav, R. Yavatkar, R. Pabbati, P. Ford, T. Moore, S. Herzog, R. Hess, "Identity Representation for RSVP", RFC 3182, October 2001
- [RFC5226] T. Narten, H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", RFC 5226, May 2008
- [ID-TRAF-CLASS] J. Polk, S. Dhesikan, P. Jones, " The Session Description Protocol (SDP) 'trafficclass' Attribute ", work in progress, Oct 2011

8.2. Informative References

- [RFC4594] J. Babiarz, K. Chan, F Baker, "Configuration Guidelines for Diffserv Service Classes", RFC 4594, August 2006

Authors' Addresses

James Polk
3913 Treemont Circle
Colleyville, Texas, USA
+1.817.271.3552

mailto: jmpolk@cisco.com

Subha Dhesikan
170 W Tasman St
San Jose, CA, USA
+1.408-902-3351

mailto: sdhesika@cisco.com

Appendix - Changes to ID

A.1 - Changes from Individual -03 to -04

The following changes were made in this version:

- clarified security considerations section to mean RSVP cannot police the type of traffic within a reservation to know if a traffic flow should be using a different profile, as defined in this document.
- changed existing informative language regarding "... other Sub-types ..." from 'can' to normative 'MAY'.

- editorial changes to clear up minor mistakes

A.2 - Changes from Individual -02 to -03

The following changes were made in this version:

- Added [ID-TRAF-CLASS] as a reference
- Changed to a new format of the profile string.
- Added many new profiles based on the new format into each parent category of Section 3.
- changed the GUID to refer to draft-ietf-mmusic-traffic-class-for-sdp-01.txt
- changed 'desktop' adjective to 'avconf' to keep in alignment with draft-ietf-mmusic-traffic-class-for-sdp-01.txt
- Have a complete IANA Registry proposal for each application-ID discussed in this draft.
- General text clean-up of the draft.

TSVWG
Internet-Draft
Intended status: Informational
Expires: December 31, 2012

Lishun. Sun
Wendong. Wang
BUPT
Fang. Yu
Huawei Technologies
June 29, 2012

Flow-based Performance Measurement
draft-sun-tsvwg-flowbased-pm-00

Abstract

The performance measurements of service flow are becoming significant important for administrators monitoring the fitness of the network. This memo defines an end-to-end application-based performance measurement method, which is achieved by generating synthetic measurement packets, injecting them to the network and analyzing the statistics carried in the measurement packets.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on December 31, 2012.

Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must

include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

This document may contain material from IETF Documents or IETF Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF Trust the right to allow modifications of such material outside the IETF Standards Process. Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

Table of Contents

1. Introduction	3
2. Conventions and Terminology	3
2.1. Conventions Used in This Document	3
2.2. Terminology	3
3. Overview	3
3.1. Motivation	4
3.2. Protocol overview	4
3.3. Logical Model	5
4. Measurement Process	6
4.1. Connection Activation	6
4.2. Measurement Process	8
4.3. Connection Deactivation	10
5. Statistics	11
5.1. Delay Calculation	11
5.2. Jitter Calculation	12
5.3. Loss Calculation	12
6. Exception Handling	13
6.1. FM/BR Packet Loss	13
6.2. Packet Mis-ordering	13
7. Use Case	13
8. Security Considerations	14
9. IANA Considerations	14
10. Acknowledgments	14
11. References	14
11.1. Normative Reference	14
11.2. Informative References	14
Authors' Addresses	15

1. Introduction

The IETF IP Performance Metrics (IPPM) working group has defined a series of standard metrics that can be applied to the quality, performance and reliability of Internet data delivery services.

In some cases, it needs to monitor the various time-varying performance indexes of the IP network, the performance measurement should be based on real service stream and reflect the real performance of the network. The average performance indexes measured by the active measurement method may not be suitable in these cases.

This memo proposes an IP performance measurement method which can support on-the-spot measurement, that is, measurement is performed online when service is running. This method is an end-to-end flow-based method which can obtain measurement data simply and accurately. It injects the OAM packets to the network which are used to carry some parameters related to service flow and some statistic information. The OAM packets are processed using the same method as its corresponding service flow, so the measurements can reflect the network status suffered by the service flow more accurately.

This IP performance measurement method can monitor the real time change of service and reflect the running situation of actual IP service. It can play an important role in the fault diagnosis and statistics of the service in IP transmission network.

2. Conventions and Terminology

2.1. Conventions Used in This Document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

2.2. Terminology

PM:performance monitoring

FM:Forward Monitoring

BR:Backward Reporting

3. Overview

3.1. Motivation

It is required to provide a reasonable estimation measurement of delay, jitter and packet loss in IP network (such as backhaul network). The above parameters are functions of time, and are stochastic in nature. Therefore, the mechanism is required to provide statistical condition estimations of the IP link status. Moreover, since measurement injects some OAM packets to the network, it is necessary that the frequency of packet generation is kept at a minimum. Moreover, these packets should not lead to an excessive computational overload on the measure device. In other words, the process of generation of these measurement packets should be simple and must not overload the transport interface. The packets must be small and infrequent so as not to cause unnecessary overload on the network bandwidth or influence the service running on the network.

3.2. Protocol overview

Firstly we make some statement for this method. We define a measurement method that can be used in some scene. The method proposed here involves three steps#: connection activation#, measurement process and connection deactivation.

Two types of logical entities are defined, the PM Initiator and the PM Responder.

Firstly the PM Initiator sends a request to the PM Responder to set up the PM connection activation. The PM Responder responds with an ACK packet including some parameters based on the request. When the PM Initiator receives the ACK, it will prepare for starting the measurement process.

During the measurement process, the PM Initiator periodically generates Forward Monitor (FM) packets with the source and destination IP addresses, and other classification information (for example DSCP class) of the service packets, which are sent to the PM Responder. The generation and transmission of FM packets can be periodical with a specific time interval, or a certain number of traffic packets should be sent between two contiguous FM packets.

The PM Responder receives the FM packets and sends Backward Reporting (BR) packets which is constructed according to the FM packets.

The path performance such as the delay, jitter and loss rate etc. are calculates by the PM Initiator according to the information in the BR packets.

The FM packets have the same source and destination IP addresses, even the same DSCP class in some cases with the data packets. So they are carried through the transport network just most like the data packets, and delay, jitter and packet loss encountered by them resembles the performance as seen by the packets of the actual traffic flows. The FM and BR packets used in this method are small enough to produce influence to actual service flow as little as possible.

3.3. Logical Model

The role and definition of the logical entities and measurement packets in this method are defined as follows.

This method is an end-to-end measurement, so two logical entities are defined.

- o PM Initiator: PM Initiator serves as the sending endpoint, and charges for generating and sending the request to initiate a PM connection. It could also send FM packets to collect measurement data and generate statistical report.
- o PM Responder: PM Responder serves as the receiving endpoint, and charges for responding the request of initiating a link. It could also send back a BR packet to the sending endpoint once it receives the FM package from the PM Initiator.

One possible scenario of relationships between these roles is shown below.

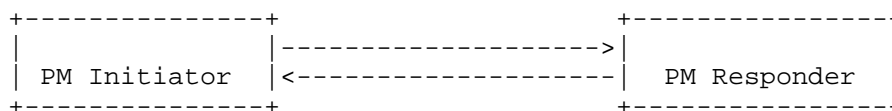


Figure 1: One possible relationship between PM Initiator and PM Responder

There are six types of packets in total, which include four types of control packets and two types of measurement packets.

Note that a new port number is introduced to be assigned by the IANA. The assigned port number is used as the destination port number of the control packets and measurement packets#, and the source port number can be random.

Control packet:

- o ACT: It is sent from the PM Initiator to a specific UDP port on PM Responder, carries parameters used in negotiation process when initiating a PM connection.
- o ACT-ACK: It is a response for ACT sent by the PM Responder to the PM Initiator.
- o DEA: It is sent by the PM Initiator to the PM Responder for disconnecting the PM connection.
- o DEA-ACK: It is a response for DEA sent by the PM Responder to the PM Initiator.

Measurement packet:

- o FM (Forward Monitoring): It is sent by the PM Initiator. The format of FM packet payload as defined by this document will be shown below. FM packet header is constructed in accordance with the data packet except the destination port.
- o BR (Backward Reporting): It is sent by the PM Responder. The format of BR packet payload as defined by this document will be shown below. It is a response for FM.

4. Measurement Process

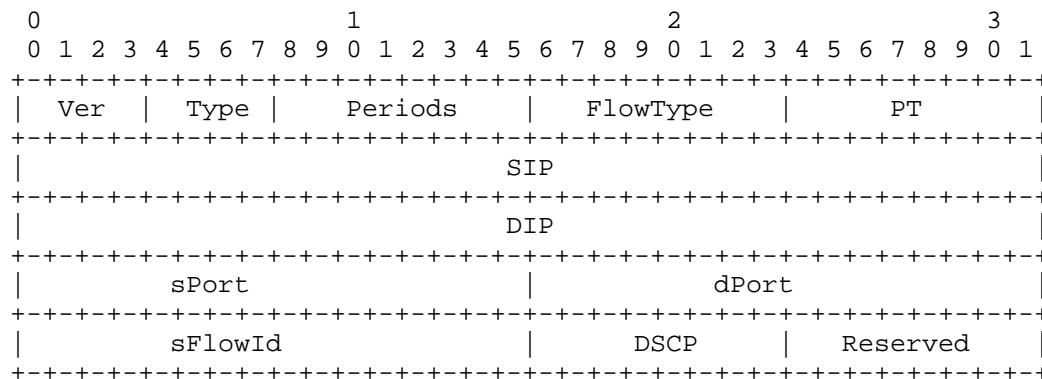
4.1. Connection Activation

In the PM connection activation process, both the PM Initiator and the PM Responder are assigned a Flow ID for a defined flow (the Flow ID is unique in a connection between the PM Initiator and the PM Responder). It should specify how to define the Flow corresponding to the measurement instance. Flow can be defined by different combinations of source IP address (SIP), destination IP address (DIP), protocol type (PT), DSCP, source port number (sPort) and destination port number (dPort). Three types of combinations are suggested: (SIP, DIP, PT), or (SIP, DIP, PT, DSCP) or (SIP, DIP, PT, sPort, dPort). The more the combinational dimensions are, the more fine-grained can be the monitoring of data flow.

Before starting the measurement, a connection should be established. When the PM Initiator wants to start the measurement process, it enables the measurement capabilities to the PM Responder by sending ACT packet to the specific UDP port on the PM Responder. When the PM Responder receives the ACT, it enables its measurement function and responses to the Sender with ACT-ACK packet. The connection activation process is finished after the PM Initiator receiving the

ACT-ACK packet from the PM Responder, then the PM Initiator can send FM packet after one cycle. The definition of flow, FlowID, and the sending period of FM packets must be consulted by two ends during the connection activation process.

The format of ACT packet is defined as follows:



Ver and Type existed in all packets in this memo indicate the version and type of packet. Type in this packets MUST be 0x1 indicates that this is an ACT packet.

Periods defined by PM Initiator indicates the sending period of FM packets.

FlowType indicates how a flow is defined. 0x0 in this filed is for (SIP, DIP, PT, sPort, dPort), while 0x1 is for (SIP, DIP, PT, DSCP) and 0x2 is for (SIP, DIP, PT). The other values are not defined.

PT is the protocol type value of the service flow needed to be measured. It may be UDP, TCP, SCTP or other types.

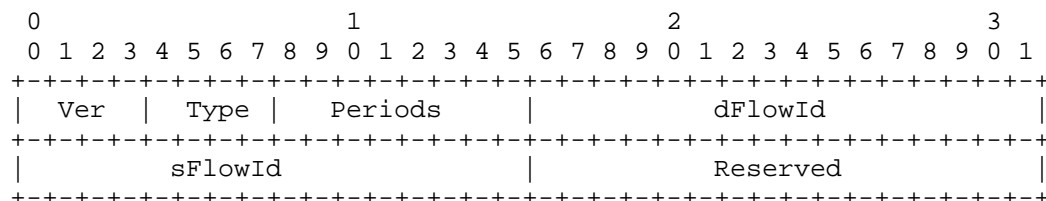
SIP is the source IP address of the service flow, and DIP is the destination IP address of the service flow. SPort and DPort which are valid only when the flow is defined by (SIP, DIP, PT, sPort, dPort) indicate source/destination port number of the ACT packets. If the FlowType is not defined by (SIP, DIP, PT, sPort, dPort), this filed is 0xFF.

sFlowId is the flow id defined by the PM Initiator.

DSCP is valid only when the flow is defined by (SIP, DIP, PT, DSCP). It indicates the value of the DSCP filed in IP header of service flow.

Reserved is reserved for extensions in future and MUST be set to 0x0 currently.

The format of ACT-ACK packet is defined as follows:



Type of 0X2 indicates ACT-ACK.

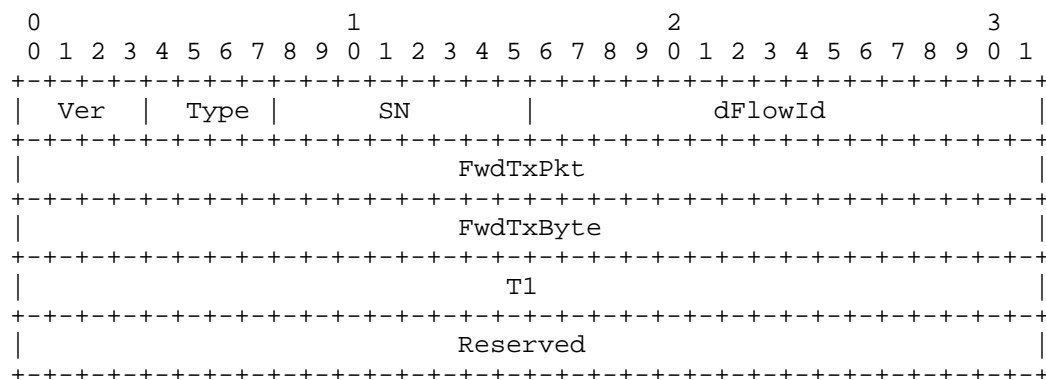
dFlowId is copied from the sFlowId field of the ACT packets.

sFlowId is the flow id defined by the PM Responder.

4.2. Measurement Process

When the connection is established successfully, the PM Initiator sends FM packets according to the given time-interval, and the PM Responder responds by sending BR packets after receiving FM packets from the PM Initiator. The destination port number in the FM packet header must be set as the specific number.

The format of FM packet is defined as follows:



Type of 0X3 indicates FM.

SN is the sequence number of the flow, which distinguishes the

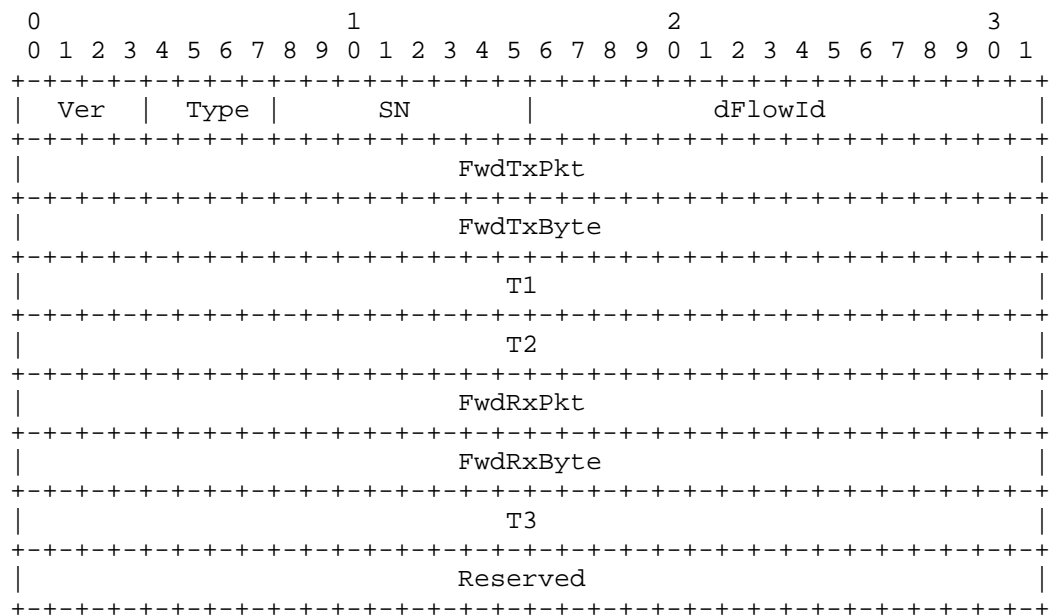
different FM packets and indicates the correspondence between FM packets and BR packets. Each PM flow should maintain a set of sequence numbers (SN).

dFlowId is the flow id of the PM Responder, which is copied from sFlowId in ACT-ACK.

FwdTxPkt is the accumulation of the number of the packets sent by the PM Initiator. FwdTxByte is the accumulation of the number of bytes sent by the PM Initiator. In order to determine the value of the fields of FwdTxPkt and FwdTxByte, the PM Initiator maintains two counters, SPN and SBN, for each PM flow that is incremented every time a traffic packet is sent. When the FM packets are to be sent, the FwdTxPkt and FwdTxByte are set to the then value of the counters respectively.

T1 is the timestamp when the PM Initiator sends FM packets. There is no requirement for synchronization between the PM Initiator and the PM Responder.

The format of BR packet is defined as follows:



Type of 0X4 indicates BR.

SN is copied from the SN field of the corresponding FM packet.

dFlowId is the flow id of the PM Initiator, which is copied from sFlowId in ACT.

The FwdTxPkt and FwdTxByte are copied from the corresponding FM packet. FwdRxPkt is the accumulation of the number of the packets received by the PM Responder. FwdRxByte is the accumulation of the number of bytes received by the PM Responder. In order to determine the value of the fields of FwdRxPkt and FwdRxByte, the PM Responder maintains two counters, RPN and RBN, for each PM flow that is incremented every time a traffic packet is received. When the BR packets are to be sent, the FwdRxPkt and FwdRxByte are set to the then value of the counters respectively.

T1 is copied from the T1 field of the FM packets, T2 is the timestamp when the PM Responder receives the FM packets, and T3 is the timestamp when the PM Responder sends the BR packets.

When the PM Responder receives a FM packet, it copies the value of FwdTxPkt, FwdTxByte and T1 in FM packet into the corresponding fields of the BR packet, and sets the fields of T3, FwdRxPkt and FwdRxByte and sends the BR packet.

Note that the PM Initiator could start multiple measurement engines; each engine is corresponding to an active logical path (with a different Flow). These measurement engines operate in parallel, and send FM packets with the flow id of the logical path, collect the corresponding BR packets, and maintain the collected statistical values.

4.3. Connection Deactivation

When the PM Initiator wants to stop the measurement, it sends Connection Deactivation request packet called DEA to the PM Responder. The PM Responder sends DEA-ACK packets back to the PM Initiator after it receives the DEA packets.

The format of DEA packet is defined as follows:

0										1										2										3									
0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+																																							
Ver										Type										Reserved										dFlowId									
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+																																							

Type of 0X5 indicates DEA.

dFlowId is the flow id defined by the PM Responder, which is copied

from sFlowId in ACT-ACK.

The format of DEA-ACK packet is defined as follows:

```

      0               1               2               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
| Ver  | Type |   Reserved   |                dFlowId                |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

Type of 0X6 indicates DEA-ACK.

dFlowId is the flow id defined by the PM Initiator, which is copied from sFlowId in ACT. .

5. Statistics

5.1. Delay Calculation

In order to determine the delay sample for a given FM and BR packets, we assume that the paths are symmetric; that is the delay would be same for FM and BR packets. Therefore, the traffic delay is calculated as:

$$T_d = \frac{(T_4 - T_1) - (T_3 - T_2)}{2}$$

Where t_d is the one-way delay for the d th BR received, t_4 is the time that the PM Initiator received the d th BR packet, t_3 is the time that the PM Initiator sent the d th FM packet. t_2 is the time that the PM Responder received d th FM packet, and t_1 is the time that PM Responder sent the d th BR packet(d th is a SN of a flow).

Let's assume that during the current reporting interval D , N BR packets were received at the PM Initiator, the delay reported to the network management entity is determined as

$$TD = \frac{1}{N} * \sum_{d=M}^{M+N-1} t_d$$

Where TD is the delay metric reported corresponding to the interval D to the network management entity.

5.2. Jitter Calculation

Jitter is defined as the Packet Delay Variation, and is not calculated for each arriving BR packet. Instead, td values are considered over several seconds, and the associated jitter value is calculated. Let's consider that N consecutive delay values were used to determine the dth jitter value, jp as:

$$jp = \sqrt{\frac{1}{N} \sum_{d=M}^{M+N-1} (TD - Td)^2}$$

jp is the variance of delay

Note that hence calculated jitter value needs to be aggregated over the reporting interval. Let's assume that during the current reporting interval D, N such jitter calculations were made at the PM Initiator, therefore the jitter reported to the network management entity is determined as:

$$jD = \frac{1}{N} \sum_{p=1}^N jp$$

5.3. Loss Calculation

When the dth BR packet is received at the PM Initiator, the loss rate plr based on the dth BR packet and (d-1)th BR packet is calculated as:

$$plr_d = \frac{(SPN(d)-SPN(d-1))-(RPN(d)-RPN(d-1))}{SPN(d)-SPN(d-1)}$$

(SPN(d)-SPN(d-1)) indicates the number of service packets sent by the PM Initiator during dth measurement, and (RPN(d)-RPN(d-1)) indicates the actual number of service packets received by the PM Responder during dth measurement.

The loss rate needs to be aggregated over the reporting interval. Let's assume that N BR packets were received during the dth reporting interval. Therefore, the packet loss rate for that interval can be calculated as:

$$PLRd = \frac{1}{N} * \sum_{d=1}^N plrd$$

6. Exception Handling

6.1. FM/BR Packet Loss

In some cases the FM or BR packet may be lost in transit, then no statistics can be obtained from this round of measurement. So the loss rate of the mth measurement can be calculated as:

$$plrm = \frac{(SPN(m)-SPN(n))-(RPN(m)-RPN(d-n))}{SPN(m)-SPN(n)}$$

where m is the SN of the BR packet currently received and n is the SN of the latest BR received.

6.2. Packet Mis-ordering

In the receive side if the received packets are out of order, the FM packet may arrive earlier than the last service packet sent before it, or later than the first service packet sent after it. Then statistical error of packet loss will be result in. Note that the packet loss calculation is based on sample statistic, so the occasional packet mis-ordering may make less impact on the packet loss statistic. And the mis-ordering can be solved by some private method which is out-of-scope of this document.

7. Use Case

This section describes a typical scene using the measurement method. The wireless mobile backhaul networks based on IP, share the available capacity between the connected eNodeBs. Compared to the traditional SDH/ATM transport network, in IP-RAN, the data transfer speed is unstable and data transfer is lacks of QoS guarantee and there is no perfect testing method on packet loss, delay and jitter. So it is necessary for the nodes in the RAN side to detect the network quality of the connection between RNC and NodeB or eNodeB and SAE.

Take the eNodeB and SAE for example; in order to make sure that the amount of generated traffic is aligned with the available capacity,

it is important that the eNodeB probes the backhaul network to determine the actual delay, jitter and packet loss encountered by typical packets. The proposed method in this document can be used to detect the IP Performance of the connections between the eNB and S-GW.

At the eNodeB, FM OAM packets are generated periodically with the source and destination IP addresses, and DSCP class. At the S-GW, after receiving the FM packets, BR packets are constructed. They are then forwarded back to the eNodeB. Upon receiving the BR packets at the logical port exactly knows the current congestion extent in transport network. The bandwidth of the logical port is reduced if congestion is detected according to the measurement result; otherwise, the bandwidth is increased slowly.

8. Security Considerations

To be defined.

9. IANA Considerations

The destination port number of the newly defined packets for measurement needs to be assigned by the IANA.

10. Acknowledgments

The authors gratefully acknowledge reviews and contributions from Peter McCann and Anthony Chan.

11. References

11.1. Normative Reference

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

11.2. Informative References

[RFC2330] Paxson, V., Almes, G., Mahdavi, J., and M. Mathis, "Framework for IP Performance Metrics", RFC 2330, May 1998.

[RFC2679] Almes, G., Kalidindi, S., and M. Zekauskas, "A One-way Delay Metric for IPPM", RFC 2679, September 1999.

- [RFC2680] Almes, G., Kalidindi, S., and M. Zekauskas, "A One-way Packet Loss Metric for IPPM", RFC 2680, September 1999.
- [RFC4656] Shalunov, S., Teitelbaum, B., Karp, A., Boote, J., and M. Zekauskas, "A One-way Active Measurement Protocol (OWAMP)", September 2006.

Authors' Addresses

Lishun Sun
Beijing University of Posts and Telecommunications
Xitucheng road 10
Haidian District, Beijing 100876
P. R. China

Email: lishunsun@Gmail.com

Wendong Wang
Beijing University of Posts and Telecommunications
Xitucheng road 10
Haidian District, Beijing 100876
P. R. China

Email: wdwang@bupt.edu.cn

Fang Yu
Huawei Technologies
Huawei Building, Q20 No.156 Beiqing Rd.Z-park
Haidian District, Beijing 100095
P. R. China

Email: grace.yufang@huawei.com

