

Congestion and Pre-Congestion
Notification
Internet-Draft
Intended status: Standards Track
Expires: November 22, 2011

B. Briscoe
BT
T. Moncaster
Moncaster Internet Consulting
M. Menth
University of Tuebingen
May 21, 2011

Encoding 3 PCN-States in the IP header using a single DSCP
draft-ietf-pcn-3-in-1-encoding-05

Abstract

The objective of Pre-Congestion Notification (PCN) is to protect the quality of service (QoS) of inelastic flows within a Diffserv domain. On every link in the PCN domain, the overall rate of the PCN-traffic is metered, and PCN-packets are appropriately marked when certain configured rates are exceeded. Egress nodes provide decision points with information about the PCN-marks of PCN-packets which allows them to take decisions about whether to admit or block a new flow request, and to terminate some already admitted flows during serious pre-congestion.

This document specifies how PCN-marks are to be encoded into the IP header by re-using the Explicit Congestion Notification (ECN) codepoints within a PCN-domain. This encoding builds on the baseline encoding of RFC5696 and provides for three different PCN marking states using a single DSCP: not-marked (NM), threshold-marked (ThM) and excess-traffic-marked (ETM). Hence, it is called the 3-in-1 PCN encoding.

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 November 22, 2011.

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	3
1.1. Changes in This Version (to be removed by RFC Editor) . .	4
2. Requirements Language	5
2.1. Terminology	5
3. Requirements for and Applicability of 3-in-1 PCN Encoding . .	5
3.1. PCN Requirements	5
3.2. Requirements Imposed by Baseline Encoding	6
3.3. Applicability of 3-in-1 PCN Encoding	7
4. Definition of 3-in-1 PCN Encoding	7
5. Behaviour of a PCN Node Compliant with the 3-in-1 PCN Encoding	8
6. Backward Compatibility	8
6.1. Backward Compatibility with Pre-existing PCN Implementations	9
6.2. Recommendations for the Use of PCN Encoding Schemes . . .	9
6.2.1. Use of Both Excess-Traffic-Marking and Threshold-Marking	10
6.2.2. Unique Use of Excess-Traffic-Marking	10
6.2.3. Unique Use of Threshold-Marking	10
7. IANA Considerations	10
8. Security Considerations	10
9. Conclusions	11
10. Acknowledgements	11
11. Comments Solicited	11
12. References	11
12.1. Normative References	11
12.2. Informative References	12
Appendix A. Co-existence of ECN and PCN (informative)	13
Authors' Addresses	15

1. Introduction

The objective of Pre-Congestion Notification (PCN) [RFC5559] is to protect 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, to decide whether to admit or block a new flow request, and flow termination to terminate some existing flows during serious pre-congestion. To achieve this, 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 real congestion occurs (hence "pre-congestion notification").

[RFC5670] provides for two metering and marking functions that are configured with reference rates. Threshold-marking marks all PCN packets once their traffic rate on a link exceeds the configured reference rate (PCN-threshold-rate). Excess-traffic-marking marks only those PCN packets that exceed the configured reference rate (PCN-excess-rate). The PCN-excess-rate is typically larger than the PCN-threshold-rate [RFC5559]. Egress nodes monitor the PCN-marks of received PCN-packets and provide information about the PCN-marks to decision points which take decisions about flow admission and termination on this basis [I-D.ietf-pcn-cl-edge-behaviour], [I-D.ietf-pcn-sm-edge-behaviour].

The baseline encoding defined in [RFC5696] describes how two PCN marking states (Not-marked and PCN-Marked) can be encoded using a single Diffserv codepoint. It also provides an experimental codepoint (EXP), along with guidelines for use of that codepoint. To support the application of two different marking algorithms in a PCN-domain, for example as required in [I-D.ietf-pcn-cl-edge-behaviour], three PCN marking states are needed. This document describes an extension to the baseline encoding that uses the EXP codepoint to provide a third PCN marking state in the IP header, still using a single Diffserv codepoint. This encoding scheme is called "3-in-1 PCN encoding".

This document only concerns the PCN wire protocol encoding for all IP headers, whether IPv4 or IPv6. It makes no changes or recommendations concerning algorithms for congestion marking or congestion response. Other documents define the PCN wire protocol for other header types. For example, the MPLS encoding is defined in [RFC5129] and Appendix A of that document provides an informative example for a mapping between the encodings in IP and in MPLS.

1.1. Changes in This Version (to be removed by RFC Editor)

From draft-ietf-pcn-3-in-1-encoding-04 to -05:

- * Draft moved to standards track as per working group discussions.
- * Added Appendix A discussing ECN handling in the PCN-domain.
- * Clarified that this document modifies [RFC5696].
- *

From draft-ietf-pcn-3-in-1-encoding-03 to -04:

- * Updated document to reflect RFC6040.
- * Re-wrote introduction.
- * Re-wrote section on applicability.
- * Re-wrote section on choosing encoding scheme.
- * Updated author details.

From draft-ietf-pcn-3-in-1-encoding-02 to -03:

- * Corrected mistakes in introduction and improved overall readability.
- * Added new terminology.
- * Rewrote a good part of Section 4 and 5 to achieve more clarity.
- * Added appendix explaining when to use which encoding scheme and how to encode them in MPLS shim headers.
- * Added new co-author.

From draft-ietf-pcn-3-in-1-encoding-01 to -02:

- * Corrected mistake in introduction, which wrongly stated that the threshold-traffic rate is higher than the excess-traffic rate. Other minor corrections.
- * Updated acks & refs.

From draft-ietf-pcn-3-in-1-encoding-00 to -01:

- * Altered the wording to make sense if draft-ietf-tsvwg-ecn-tunnel moves to proposed standard.
- * References updated

From draft-briscoe-pcn-3-in-1-encoding-00 to draft-ietf-pcn-3-in-1-encoding-00:

- * Filename changed to draft-ietf-pcn-3-in-1-encoding.
- * Introduction altered to include new template description of PCN.
- * References updated.
- * Terminology brought into line with [RFC5670].
- * Minor corrections.

2. 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 [RFC2119].

2.1. Terminology

General PCN-related terminology is defined in the PCN architecture [RFC5559], and terminology specific to packet encoding is defined in the PCN baseline encoding [RFC5696]. Additional terminology is defined below.

PCN encoding: mapping of PCN marking states to specific codepoints in the packet header.

3. Requirements for and Applicability of 3-in-1 PCN Encoding

3.1. PCN Requirements

In accordance with the PCN architecture [RFC5559], PCN-ingress-nodes control packets entering a PCN-domain. Packets belonging to PCN-controlled flows are subject to PCN-metering and -marking, and PCN-ingress-nodes mark them as Not-marked (PCN-colouring). Any node in the PCN-domain may perform PCN-metering and -marking and mark PCN-

packets if needed. There are two different metering and marking schemes: threshold-marking and excess-traffic-marking [RFC5670]. Some edge behaviors require only a single marking scheme [I-D.ietf-pcn-sm-edge-behaviour], others require both [I-D.ietf-pcn-cl-edge-behaviour]. In the latter case, three PCN marking states are needed: not-marked (NM) to indicate not-marked packets, threshold-marked (ThM) to indicate packets marked by the threshold-marker, and excess-traffic-marked (ETM) to indicate packets marked by the excess-traffic-marker [RFC5670]. Threshold-marking and excess-traffic-marking are configured to start marking packets at different load conditions, so one marking scheme indicates more severe pre-congestion than the other. Therefore, a fourth PCN marking state indicating that a packet is marked by both markers is not needed. However a fourth codepoint is required to indicate packets that are not PCN-capable (the not-PCN codepoint).

In all current PCN edge behaviors that use two marking schemes [RFC5559], [I-D.ietf-pcn-cl-edge-behaviour], excess-traffic-marking is configured with a larger reference rate than threshold-marking. We take this as a rule and define excess-traffic-marked as a more severe PCN-mark than threshold-marked.

3.2. Requirements Imposed by Baseline Encoding

The baseline encoding scheme [RFC5696] was defined so that it could be extended to accommodate an additional marking state. It provides rules to embed the encoding of two PCN states in the IP header. Figure 1 shows the structure of the former type-of-service field. It contains the 6-bit Differentiated Services (DS) field that holds the DS codepoint (DSCP) [RFC2474] and the 2-bit ECN field [RFC3168].

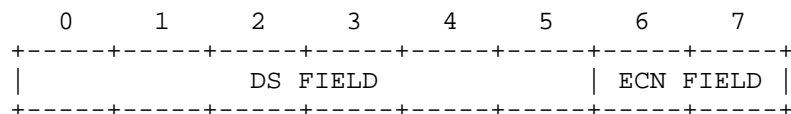


Figure 1: Structure of the former type-of-service field in IP

Baseline encoding defines that the DSCP must be set to a PCN-compatible DSCP n and the ECN-field [RFC3168] indicates the specific PCN-mark. Baseline encoding offers four possible encoding states within a single DSCP with the following restrictions.

- o Codepoint '00' (not-ECT) is used to indicate non-PCN traffic as "not-PCN". This allows both PCN and non-PCN traffic to use the same DSCP.

- o Codepoint '10' (ECT(0)) is used to indicate Not-marked PCN traffic.
- o Codepoint '11' (CE) is used to indicate the most severe PCN-mark.
- o Codepoint '01' (ECT(1)) is available for experimental use and may be re-used by other PCN encodings such as the presently defined 3-in-1 PCN encoding (subject to the rules defined in [RFC5696]).

[RFC6040] defines rules for the encapsulation and decapsulation of ECN markings within IP-in-IP tunnels. This RFC removes some of the constraints that existed when [RFC5696] was written. Happily the rules for use of the EXP codepoint are fully compatible with [RFC6040]. In particular, the relative severity of each marking is the same: CE (PM) is more severe than ECT(1) (EXP) is more severe than ECT(0) (NM). This is discussed in more detail in both the baseline encoding document [RFC5696] and in [I-D.ietf-pcn-encoding-comparison].

3.3. Applicability of 3-in-1 PCN Encoding

The 3-in-1 encoding is applicable in situations where two marking schemes are being used in the PCN-domain. In some circumstances it can also be used in PCN-domains with only a single marking scheme in use. Further guidance on choosing an encoding scheme can be found in Section 6.2. All nodes within the PCN-domain MUST be fully compliant with the ECN encapsulation rules set out in [RFC6040]. As such the encoding is not applicable in situations where legacy tunnels might exist.

4. Definition of 3-in-1 PCN Encoding

The 3-in-1 PCN encoding scheme is an extension of the baseline encoding scheme defined in [RFC5696]. The PCN requirements and the extension rules for baseline encoding presented in the previous section determine how PCN encoding states are carried in the IP headers. This is shown in Figure 2.

DSCP	Codepoint in ECN field of IP header <RFC3168 codepoint name>			
	00 <Not-ECT>	10 <ECT(0)>	01 <ECT(1)>	11 <CE>
DSCP n	Not-PCN	NM	ThM	ETM

Figure 2: 3-in-1 PCN Encoding

Like baseline encoding, 3-in-1 PCN encoding also uses a PCN compatible DSCP n and the ECN field for the encoding of PCN-marks. The PCN-marks have the following meaning.

Not-PCN: indicates a non-PCN-packet, i.e., a packet that is not subject to PCN metering and marking.

NM: Not-marked. Indicates a PCN-packet that has not yet been marked by any PCN marker.

ThM: Threshold-marked. Indicates a PCN-packet that has been marked by a threshold-marker [RFC5670].

ETM: Excess-traffic-marked. Indicates a PCN-packet that has been marked by an excess-traffic-marker [RFC5670].

5. Behaviour of a PCN Node Compliant with the 3-in-1 PCN Encoding

To be compliant with the 3-in-1 PCN Encoding, an PCN interior node behaves as follows:

- o It MUST change NM to ThM if the threshold-meter function indicates a need to mark the packet;
- o It MUST change NM or ThM to ETM if the excess-traffic-meter function indicates a need to mark the packet;
- o It MUST NOT change not-PCN to NM, ThM, or ETM;
- o It MUST NOT change a NM, ThM, or ETM to not-PCN;
- o It MUST NOT change ThM to NM;
- o It MUST NOT change ETM to ThM or to NM;

In other words, a PCN interior node MUST NOT mark PCN-packets into non-PCN packets and vice-versa, and it may increase the severity of the PCN-mark of a PCN-packet, but it MUST NOT decrease it.

6. Backward Compatibility

Discussion of backward compatibility between PCN encoding schemes and previous uses of the ECN field is given in Section 6 of [RFC5696].

6.1. Backward Compatibility with Pre-existing PCN Implementations

This encoding complies with the rules for extending the baseline PCN encoding schemes in Section 5 of [RFC5696].

The term "compatibility" is meant in the following sense. It is possible to operate nodes with baseline encoding [RFC5696] and 3-in-1 encoding in the same PCN domain. The nodes with baseline encoding MUST perform excess-traffic-marking because the 11 codepoint of 3-in-1 encoding also means excess-traffic-marked. PCN-boundary-nodes of such domains are required to interpret the full 3-in-1 encoding and not just baseline encoding, otherwise they cannot interpret the 01 codepoint.

Using nodes that perform only excess-traffic-marking may make sense in networks using the CL edge behavior [I-D.ietf-pcn-cl-edge-behaviour]. Such nodes are able to notify the egress only about severe pre-congestion when traffic needs to be terminated. This seems reasonable for locations that are not expected to see any pre-congestion, but excess-traffic-marking gives them a means to terminate traffic if unexpected overload occurs.

6.2. Recommendations for the Use of PCN Encoding Schemes

NOTE: This sub-section is informative not normative.

When deciding which PCN encoding is suitable an operator needs to take account of how many PCN states need to be encoded. The following table gives guidelines on which encoding to use with either threshold-marking, excess-traffic marking or both.

Marking schemes in use	Recommended encoding scheme
Only threshold-marking	Baseline encoding [RFC5696]
Only excess-traffic-marking	Baseline encoding [RFC5696] or 3-in-1 PCN encoding
Threshold-marking and excess-traffic-marking	3-in-1 PCN encoding

Figure 3: Guidelines for choosing PCN encoding schemes

6.2.1. Use of Both Excess-Traffic-Marking and Threshold-Marking

If both excess-traffic-marking and threshold-marking are enabled in a PCN-domain, 3-in-1 encoding should be used as described in this document.

6.2.2. Unique Use of Excess-Traffic-Marking

If only excess-traffic-marking is enabled in a PCN-domain, baseline encoding or 3-in-1 encoding may be used. They lead to the same encoding because PCN-boundary nodes will interpret baseline "PCN-marked (PM)" as "excess-traffic-marked (ETM)".

6.2.3. Unique Use of Threshold-Marking

No scheme is currently proposed that solely uses threshold-marking. If such a scheme is proposed, the choice of encoding scheme will depend on whether nodes are compliant with [RFC6040] or not. Where it is certain that all nodes in the PCN-domain are compliant then either 3-in-1 encoding or baseline encoding are suitable. If legacy tunnel decapsulators exist within the PCN-domain then baseline encoding SHOULD be used.

7. IANA Considerations

This memo includes no request to IANA.

Note to RFC Editor: this section may be removed on publication as an RFC.

8. Security Considerations

The security concerns relating to this extended PCN encoding are the same as those in [RFC5696]. In summary, PCN-boundary nodes are responsible for ensuring inappropriate PCN markings do not leak into or out of a PCN domain, and the current phase of the PCN architecture assumes that all the nodes of a PCN-domain are entirely under the control of a single operator, or a set of operators who trust each other.

Given the only difference between the baseline encoding and the present 3-in-1 encoding is the use of the 01 codepoint, no new security issues are raised, as this codepoint was already available for experimental use in the baseline encoding.

9. Conclusions

The 3-in-1 PCN encoding uses a PCN-compatible DSCP and the ECN field to encode PCN-marks. One codepoint allows non-PCN traffic to be carried with the same PCN-compatible DSCP and three other codepoints support three PCN marking states with different levels of severity. The use of this PCN encoding scheme presupposes that any tunnels in the PCN region have been updated to comply with [RFC6040].

10. Acknowledgements

Thanks to Phil Eardley, Teco Boot, Kwok Ho Chan and Georgios Karaginannis for reviewing this document.

11. Comments Solicited

To be removed by RFC Editor: Comments and questions are encouraged and very welcome. They can be addressed to the IETF Congestion and Pre-Congestion working group mailing list <pcn@ietf.org>, and/or to the authors.

12. References

12.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 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.
- [RFC3168] Ramakrishnan, K., Floyd, S., and D. Black, "The Addition of Explicit Congestion Notification (ECN) to IP", RFC 3168, September 2001.
- [RFC3540] Spring, N., Wetherall, D., and D. Ely, "Robust Explicit Congestion Notification (ECN) Signaling with Nonces", RFC 3540, June 2003.
- [RFC4301] Kent, S. and K. Seo, "Security Architecture for the Internet Protocol", RFC 4301, December 2005.
- [RFC4594] Babiarz, J., Chan, K., and F. Baker, "Configuration

Guidelines for DiffServ Service Classes", RFC 4594, August 2006.

- [RFC5129] Davie, B., Briscoe, B., and J. Tay, "Explicit Congestion Marking in MPLS", RFC 5129, January 2008.
- [RFC5559] Eardley, P., "Pre-Congestion Notification (PCN) Architecture", RFC 5559, June 2009.
- [RFC5670] Eardley, P., "Metering and Marking Behaviour of PCN-Nodes", RFC 5670, November 2009.
- [RFC5696] Moncaster, T., Briscoe, B., and M. Menth, "Baseline Encoding and Transport of Pre-Congestion Information", RFC 5696, November 2009.
- [RFC5865] Baker, F., Polk, J., and M. Dolly, "A Differentiated Services Code Point (DSCP) for Capacity-Admitted Traffic", RFC 5865, May 2010.
- [RFC6040] Briscoe, B., "Tunnelling of Explicit Congestion Notification", RFC 6040, November 2010.

12.2. Informative References

- [I-D.ietf-pcn-cl-edge-behaviour]
Charny, A., Huang, F., Karagiannis, G., Menth, M., and T. Taylor, "PCN Boundary Node Behaviour for the Controlled Load (CL) Mode of Operation", draft-ietf-pcn-cl-edge-behaviour-08 (work in progress), December 2010.
- [I-D.ietf-pcn-encoding-comparison]
Karagiannis, G., Chan, K., Moncaster, T., Menth, M., Eardley, P., and B. Briscoe, "Overview of Pre-Congestion Notification Encoding", draft-ietf-pcn-encoding-comparison-05 (work in progress), April 2011.
- [I-D.ietf-pcn-sm-edge-behaviour]
Charny, A., Karagiannis, G., Menth, M., and T. Taylor, "PCN Boundary Node Behaviour for the Single Marking (SM) Mode of Operation", draft-ietf-pcn-sm-edge-behaviour-05 (work in progress), December 2010.

Appendix A. Co-existence of ECN and PCN (informative)

The PCN encoding described in this document re-uses the bits of the ECN field in the IP header. Consequently, this disables ECN within the PCN domain. Appendix B of [RFC5696] included advice on handling ECN traffic within a PCN-domain. This appendix clarifies that advice.

For the purposes of this appendix we define two forms of traffic that might arrive at a PCN-ingress node. These are Admission-controlled traffic and Non-admission-controlled traffic.

Admission-controlled traffic will be remarked to the PCN-compatible DSCP by the PCN-ingress node. Two mechanisms can be used to identify such traffic:

- a. flow signalling associates a filterspec with a need for admission control (e.g. through RSVP or some equivalent message down from a SIP server to the ingress), and the PCN-ingress remarks traffic matching that filterspec to a PCN-compatible DSCP, as its chosen admission control mechanism.
- b. Traffic arrives with a DSCP that implies it requires admission control such as VOICE-ADMIT [RFC5865] or Interactive Real-Time, Broadcast TV when used for video on demand, and Multimedia Conferencing [RFC4594][RFC5865].

All other traffic can be thought of as Non-admission-controlled. However such traffic may still need to share the same DSCP as the Admission-controlled traffic. This may be due to policy (for instance if it is high priority voice traffic), or may be because there is a shortage of local DSCPs.

ECN [RFC3168] is an end-to-end congestion notification mechanism. As such it is possible that some traffic entering the PCN-domain may also be ECN capable. The following lists the four cases for how e2e ECN traffic may wish to be treated while crossing a PCN domain:

ECN capable traffic that does not require admission control and does not carry a DSCP that the PCN-ingress is using for PCN-capable traffic. This requires no action.

ECN capable traffic that does not require admission control but carries a DSCP that the PCN-ingress is using for PCN-capable traffic. There are two options.

- * The ingress maps the DSCP to a local DSCP with the same scheduling PHB as the original DSCP, and the egress re-maps it to the original PCN-compatible DSCP.
- * The ingress tunnels the traffic, setting not-PCN in the outer header; note that this turns off ECN for this traffic within the PCN domain.

The first option is recommended unless the operator is short of local DSCPs.

ECN-capable Admission-controlled traffic: There are two options.

- * The PCN-ingress places this traffic in a tunnel with a PCN-compatible DSCP in the outer header. The PCN-egress zeroes the ECN-field before decapsulation.
- * The PCN-ingress drops CE-marked packets and the PCN-egress zeros the ECN field of all PCN packets.

The second option is not recommended unless tunnelling is not possible for some reason..

ECN-capable Admission-controlled where the e2e transport somehow indicates that it wants to see PCN marks: NOTE this is currently experimental only.

Schemes have been suggested where PCN marks may be leaked out of the PCN-domain and used by the end hosts to modify realtime data rates. Currently all such schemes are experimental and the following is for guidance only.

The PCN-ingress needs to tunnel the traffic using [RFC6040]. The PCN-egress should not zero the ECN field, and the tunnel egress should use [RFC6040] normal mode (preserving any PCN-marking). Note that this may turn ECT(0) into ECT(1) and so is not compatible with the experimental ECN nonce [RFC3540].

In the list above any form of IP-in-IP tunnel can be used unless specified otherwise. NB, We assume a logical separation of tunneling and PCN actions in both PCN-ingress and PCN-egress nodes. That is, any tunneling action happens wholly outside the PCN-domain as illustrated in the following figure:

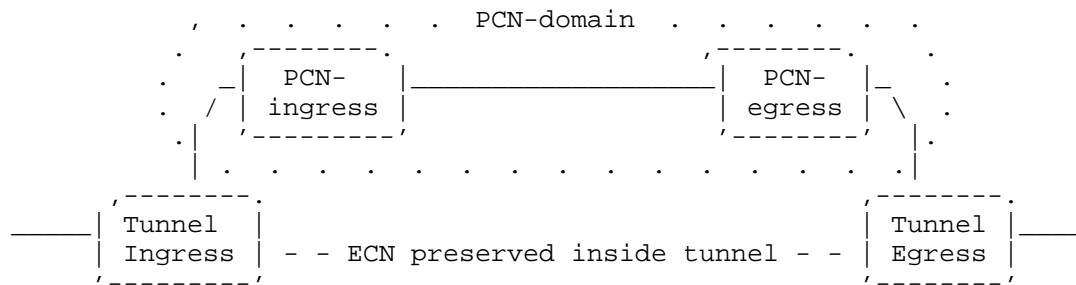


Figure 4: Separation of tunneling and PCN actions

Authors' Addresses

Bob Briscoe
 BT
 B54/77, Adastral Park
 Martlesham Heath
 Ipswich IP5 3RE
 UK

Phone: +44 1473 645196
 Email: bob.briscoe@bt.com
 URI: <http://bobbbriscoe.net/>

Toby Moncaster
 Moncaster Internet Consulting
 Dukes
 Layer Marney
 Colchester CO5 9UZ
 UK

Phone: +44 7764 185416
 Email: toby@moncaster.com
 URI: <http://www.moncaster.com/>

Michael Menth
University of Tuebingen
Sand 13
Tuebingen 72076
Germany

Phone: +49 7071 2970505
Email: menth@informatik.uni-tuebingen.de

Internet Engineering Task Force
Internet-Draft
Intended status: Experimental
Expires: December 24, 2011

A. Charny
Cisco Systems
F. Huang
Huawei Technologies
G. Karagiannis
U. Twente
M. Menth
University of Tuebingen
T. Taylor, Ed.
Huawei Technologies
June 22, 2011

PCN Boundary Node Behaviour for the Controlled Load (CL) Mode of
Operation
draft-ietf-pcn-cl-edge-behaviour-09

Abstract

Pre-congestion notification (PCN) is a means for protecting the quality of service for inelastic traffic admitted to a Diffserv domain. The overall PCN architecture is described in RFC 5559. This memo is one of a series describing possible boundary node behaviours for a PCN-domain. The behaviour described here is that for a form of measurement-based load control using three PCN marking states, not-marked, threshold-marked, and excess-traffic-marked. This behaviour is known informally as the Controlled Load (CL) PCN-boundary-node behaviour.

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 24, 2011.

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	4
1.1. Terminology	4
2. [CL-Specific] Assumed Core Network Behaviour for CL	8
3. Node Behaviours	8
3.1. Overview	8
3.2. Behaviour of the PCN-Egress-Node	9
3.2.1. Data Collection	9
3.2.2. Reporting the PCN Data	10
3.2.3. Optional Report Suppression	10
3.3. Behaviour at the Decision Point	11
3.3.1. Flow Admission	11
3.3.2. Flow Termination	12
3.3.3. Decision Point Action For Missing PCN-Boundary-Node Reports	13
3.4. Behaviour of the Ingress Node	14
3.5. Summary of Timers and Associated Configurable Durations	15
3.5.1. Recommended Values For the Configurable Durations	16
4. Identifying Ingress and Egress Nodes For PCN Traffic	17
5. Specification of Diffserv Per-Domain Behaviour	17
5.1. Applicability	17
5.2. Technical Specification	18
5.2.1. Classification and Traffic Conditioning	18
5.2.2. PHB Configuration	19
5.3. Attributes	19
5.4. Parameters	19
5.5. Assumptions	20
5.6. Example Uses	21
5.7. Environmental Concerns	21
5.8. Security Considerations	21
6. Security Considerations	21
7. IANA Considerations	21
8. Acknowledgements	21
9. References	22
9.1. Normative References	22
9.2. Informative References	22
Authors' Addresses	23

1. Introduction

The objective of Pre-Congestion Notification (PCN) is to protect 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, to decide whether to admit or block a new flow request, and (in abnormal circumstances) flow termination to decide whether to terminate some of the existing flows. To achieve this, the overall rate of PCN-traffic is metered on every link in the PCN-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 PCN-boundary-nodes about incipient overloads before any congestion occurs (hence the "pre" part of "pre-congestion notification"). The level of marking allows decisions to be made about whether to admit or terminate PCN-flows. For more details see [RFC5559].

Section 3 of this document specifies a detailed set of algorithms and procedures used to implement the PCN mechanisms for the CL mode of operation. Since the algorithms depend on specific metering and marking behaviour at the interior nodes, it is also necessary to specify the assumptions made about PCN-interior-node behaviour (Section 2). Finally, because PCN uses DSCP values to carry its markings, a specification of PCN-boundary-node behaviour MUST include the per domain behaviour (PDB) template specified in [RFC3086], filled out with the appropriate content (Section 5).

[RFC EDITOR'S NOTE: you may choose to delete the following paragraph and the "[CL-specific]" tags throughout this document when publishing it, since they are present primarily to aid reviewers. RFCyyyy is the published version of draft-ietf-pcn-sm-edge-behaviour.]

A companion document [RFCyyyy] specifies the Single Marking (SM) PCN-boundary-node behaviour. This document and [RFCyyyy] have a great deal of text in common. To simplify the task of the reader, the text in the present document that is specific to the CL PCN-boundary-node behaviour is preceded by the phrase: "[CL-specific]". A similar distinction for SM-specific text is made in [RFCyyyy].

1.1. 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 [RFC2119].

This document uses the following terms defined in Section 2 of [RFC5559]:

- o PCN-domain;
- o PCN-ingress-node;
- o PCN-egress-node;
- o PCN-interior-node;
- o PCN-boundary-node;
- o PCN-flow;
- o ingress-egress-aggregate (IEA);
- o [CL-specific] PCN-threshold-rate;
- o PCN-excess-rate;
- o PCN-admissible-rate;
- o PCN-supportable-rate;
- o PCN-marked;
- o [CL-specific] threshold-marked;
- o excess-traffic-marked.

It also uses the terms PCN-traffic and PCN-packet, for which the definition is repeated from [RFC5559] because of their importance to the understanding of the text that follows:

PCN-traffic, PCN-packets, PCN-BA

A PCN-domain carries traffic of different Diffserv behaviour 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 and the ECN field.

This document uses the following terms from [RFC5670]:

- o [CL-specific] threshold-meter;
- o excess-traffic-meter.

To complete the list of borrowed terms, this document reuses the following terms and abbreviations defined in Section 3 of [RFC5696]:

- o not-PCN codepoint;
- o Not-marked (NM) codepoint;
- o PCN-marked (PM) codepoint;
- o [CL-specific] Experimental (EXP) codepoint.

This document defines the following additional terms:

Decision Point

The node that makes the decision about which flows to admit and to terminate. In a given network deployment, this can be the PCN-ingress-node or a centralized control node. In either case, the PCN-ingress-node is the point where the decisions are enforced.

NM-rate

The rate of not-marked PCN-traffic received at a PCN-egress-node for a given ingress-egress-aggregate in octets per second. For further details see Section 3.2.1.

[CL-specific] ThM-rate

The rate of threshold-marked PCN-traffic received at a PCN-egress-node for a given ingress-egress-aggregate in octets per second. For further details see Section 3.2.1.

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. For further details see Section 3.2.1.

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. For further details see Section 3.4.

Congestion level estimate (CLE)

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 (Section 3.3.1) and is also used by the report suppression procedure (Section 3.2.3) if report suppression is activated.

PCN-admission-state

The state ("admit" or "block") derived by the Decision Point for a given ingress-egress-aggregate based on PCN packet marking statistics. The Decision Point decides to admit or block new flows offered to the aggregate based on the current value of the

PCN-admission-state. For further details see Section 3.3.1.

Sustainable aggregate rate (SAR)

The estimated maximum rate of PCN-traffic that can be carried in a given ingress-egress-aggregate at a given moment without risking degradation of quality of service for the admitted flows. The intention is that if the PCN-sent-rate of every ingress-egress-aggregate passing through a given link is limited to its sustainable aggregate rate, the total rate of PCN-traffic flowing through the link will be limited to the PCN-supportable-rate for that link. An estimate of the sustainable aggregate rate for a given ingress-egress-aggregate is derived as part of the flow termination procedure, and is used to determine how much PCN-traffic needs to be terminated. For further details see Section 3.3.2.

CLE-reporting-threshold

A configurable value against which the CLE is compared as part of the report suppression procedure. For further details, see Section 3.2.3.

CLE-limit

A configurable value against which the CLE is compared to determine the PCN-admission-state for a given ingress-egress-aggregate. For further details, see Section 3.3.1.

T-meas

A configurable time interval that defines the measurement period over which the PCN-egress-node collects statistics relating to PCN-traffic marking. At the end of the interval the PCN-egress-node calculates the values NM-rate, [CL-specific] ThM-rate, and ETM-rate as defined and sends a report to the Decision Point, subject to the operation of the report suppression feature. For further details see Section 3.2.

T-maxsuppress

A configurable time interval after which the PCN-egress-node MUST send a report to the Decision Point for a given ingress-egress-aggregate regardless of the most recent values of the CLE. This mechanism provides the Decision Point with a periodic confirmation of liveness when report suppression is activated. For further details, see Section 3.2.3.

T-fail

A configurable interval after which the Decision Point concludes that communication from a given PCN-egress-node has failed if it has received no reports from the PCN-egress-node during that interval. For further details see Section 3.3.3.

2. [CL-Specific] Assumed Core Network Behaviour for CL

This section describes the assumed behaviour for PCN-interior-nodes in the PCN-domain. The CL mode of operation assumes that:

- o PCN-interior-nodes perform both threshold-marking and excess-traffic-marking of PCN-packets, according to the rules specified in [RFC5670];
- o excess-traffic-marking of PCN-packets uses the PCN-Marked (PM) codepoint defined in [RFC5696];
- o threshold-marking of PCN-packets uses the EXP codepoint defined in [RFC5696];
- o the PCN-domain satisfies the conditions specified in [RFC5696];
- o on each link the reference rate for the threshold-meter is configured to be equal to the PCN-admissible-rate for the link;
- o on each link the reference rate for the excess-traffic-meter is configured to be equal to the PCN-supportable-rate for the link;
- o the set of valid codepoint transitions is as shown in Section 4.2 of [RFC5696].

3. Node Behaviours

3.1. Overview

This section describes the behaviour of the PCN-ingress-node, PCN-egress-node, and the Decision Point (which MAY be collocated with the PCN-ingress-node).

The PCN-egress-node collects the rates of not-marked, [CL-specific] threshold-marked, and excess-traffic-marked PCN-traffic for each ingress-egress-aggregate and reports them to the Decision Point. [CL-specific] It MAY also identify and report PCN-flows that have experienced excess-traffic-marking. For a detailed description, see Section 3.2.

The PCN-ingress-node enforces flow admission and termination decisions. It also reports the rate of PCN-traffic sent to a given ingress-egress-aggregate when requested by the Decision Point. For details, see Section 3.4.

Finally, the Decision Point makes flow admission decisions and

selects flows to terminate based on the information provided by the PCN-ingress-node and PCN-egress-node for a given ingress-egress-aggregate. For details, see Section 3.3.

3.2. Behaviour of the PCN-Egress-Node

3.2.1. Data Collection

The PCN-egress-node MUST meter the PCN-traffic it receives in order to calculate the following rates for each ingress-egress-aggregate passing through it. These rates SHOULD be calculated at the end of each measurement period based on the PCN-traffic observed during that measurement period. The duration of a measurement period is equal to the configurable value T-meas. For further information see Section 3.5.

- o NM-rate: octets per second of PCN-traffic in PCN-packets that are not-marked (i.e., marked with the NM codepoint);
- o [CL-specific] ThM-rate: octets per second of PCN-traffic in PCN-packets that are threshold-marked (i.e., marked with the EXP codepoint);
- o ETM-rate: octets per second of PCN-traffic in PCN-packets that are excess-traffic-marked (i.e., marked with the PM codepoint).

Informative note: metering the PCN-traffic continuously and using equal-length measurement intervals minimizes the statistical variance introduced by the measurement process itself. On the other hand, the operation of PCN is not affected if the starting and ending times of the measurement intervals for different ingress-egress-aggregates are different.

[CL-specific] As a configurable option, the PCN-egress-node MAY record flow identifiers of the PCN-flows for which excess-traffic-marked packets have been observed during this measurement interval. If this set is large (e.g., more than 20 flows), the PCN-egress-node MAY record only the most recently excess-traffic-marked PCN-flow identifiers rather than the complete set.

These can be used by the Decision Point when it selects flows for termination. In networks using multipath routing it is possible that congestion is not occurring on all paths carrying a given ingress-egress-aggregate. Assuming that specific PCN-flows are routed via specific paths, identifying the PCN-flows that are experiencing excess-traffic-marking helps to avoid termination of PCN-flows not contributing to congestion.

3.2.2. Reporting the PCN Data

Unless the report suppression option described in Section 3.2.3 is activated, the PCN-egress-node MUST report the latest values of NM-rate, [CL-specific] ThM-rate, and ETM-rate to the Decision Point each time that it calculates them.

[CL-specific] If the PCN-egress-node recorded a set of flow identifiers of PCN-flows for which excess-traffic-marking was observed in the most recent measurement interval, then it MUST also include these identifiers in the report.

3.2.3. Optional Report Suppression

Report suppression MUST be provided as a configurable option, along with two configurable parameters, the CLE-reporting-threshold and the maximum report suppression interval T-maxsuppress. The default value of the CLE-reporting-threshold is zero. The CLE-reporting-threshold MUST NOT exceed the CLE-limit configured at the Decision Point. For further information on T-maxsuppress see Section 3.5.

If the report suppression option is enabled, the PCN-egress-node MUST apply the following procedure to decide whether to send a report to the Decision Point, rather than sending a report automatically at the end of each measurement interval.

1. As well as the quantities NM-rate, [CLE-specific] ThM-rate, and ETM-rate, the PCN-egress-node MUST calculate the congestion level estimate (CLE) for each measurement interval. The CLE is computed as:

$$\begin{aligned} & \text{[CL-specific]} \\ & \text{CLE} = (\text{ThM-rate} + \text{ETM-rate}) / (\text{NM-rate} + \text{ThM-rate} + \text{ETM-rate}) \end{aligned}$$

if any PCN-traffic was observed, or CLE = 0 if all the rates are zero.

2. If the CLE calculated for the latest measurement interval is greater than the CLE-reporting-threshold and/or the CLE calculated for the immediately previous interval was greater than the CLE-reporting-threshold, then the PCN-egress-node MUST send a report to the Decision Point. The contents of the report are described below.

The reason for taking into account the CLE of the previous interval is to ensure that the Decision Point gets immediate feedback if the CLE has dropped below CLE-reporting-threshold. This is essential if the Decision Point is running the flow

termination procedure and observing whether (further) flow termination is needed. See Section 3.3.2.

3. If an interval T-maxsuppress has elapsed since the last report was sent to the Decision Point, then the PCN-egress-node MUST send a report to the Decision Point regardless of the CLE value.
4. If neither of the preceding conditions holds, the PCN-egress-node MUST NOT send a report for the latest measurement interval.

Each report sent to the Decision Point when report suppression has been activated MUST contain the values of NM-rate, [CL-specific] ThM-rate, ETM-rate, and CLE that were calculated for the most recent measurement interval. [CL-specific] If the PCN-egress-node recorded a set of flow identifiers of PCN-flows for which excess-traffic-marking was observed in the most recent measurement interval, then it MUST also include these identifiers in the report.

The above procedure ensures that at least one report is sent per interval (T-maxsuppress + T-meas). This demonstrates to the Decision Point that both the PCN-egress-node and the communication path between that node and the Decision Point are in operation.

3.3. Behaviour at the Decision Point

Operators can choose to use PCN procedures just for flow admission, or just for flow termination, or for both. A compliant Decision Point MUST implement both mechanisms, but configurable options MUST be provided to activate or deactivate PCN-based flow admission and flow termination independently of each other at a given Decision Point.

If PCN-based flow termination is enabled but PCN-based flow admission is not, flow termination operates as specified in this document.

Logically, some other system of flow admission control is in operation, but the description of such a system is out of scope of this document and depends on local arrangements.

3.3.1. Flow Admission

The Decision Point determines the PCN-admission-state for a given ingress-egress-aggregate each time it receives a report from the egress node. It makes this determination on the basis of the congestion level estimate (CLE). If the CLE is provided in the egress node report, the Decision Point SHOULD use the reported value. If the CLE was not provided in the report, the Decision Point MUST calculate it based on the other values provided in the report, using

the formula:

[CL-specific]
$$\text{CLE} = (\text{ThM-rate} + \text{ETM-rate}) / (\text{NM-rate} + \text{ThM-rate} + \text{ETM-rate})$$

if any PCN-traffic was observed, or CLE = 0 if all the rates are zero.

The Decision Point MUST compare the reported or calculated CLE to a configurable value, the CLE-limit. If the CLE is less than the CLE-limit, the PCN-admission-state for that aggregate MUST be set to "admit"; otherwise it MUST be set to "block".

[CL-specific] The outcome of the comparison is not very sensitive to the value of the CLE-limit in practice, because when threshold-marking occurs it tends to persist long enough that threshold-marked traffic becomes a large proportion of the received traffic in a given interval.

If the PCN-admission-state for a given ingress-egress-aggregate is "admit", the Decision Point SHOULD allow new flows to be admitted to that aggregate. If the PCN-admission-state for a given ingress-egress-aggregate is "block", the Decision Point SHOULD NOT allow new flows to be admitted to that aggregate. These actions MAY be modified by policy in specific cases, but such policy intervention risks defeating the purpose of using PCN.

3.3.2. Flow Termination

[CL-specific] When the report from the PCN-egress-node includes a non-zero value of the ETM-rate for some ingress-egress-aggregate, the Decision Point MUST request the PCN-ingress-node to provide an estimate of the rate (PCN-sent-rate) at which the PCN-ingress-node is receiving PCN-traffic that is destined for the given ingress-egress-aggregate.

If the Decision Point is collocated with the PCN-ingress-node, the request and response are internal operations.

The Decision Point MUST then wait, for both the requested rate from the PCN-ingress-node and the next report from the PCN-egress-node for the ingress-egress-aggregate concerned. If this next egress node report also includes a non-zero value for the ETM-rate, the Decision Point MUST determine the amount of PCN-traffic to terminate using the following steps:

1. [CL-specific] The sustainable aggregate rate (SAR) for the given ingress-egress-aggregate is estimated by the sum:

$$\text{SAR} = \text{NM-rate} + \text{ThM-rate}$$

for the latest reported interval.

2. The amount of traffic to be terminated is the difference:

$$\text{PCN-sent-rate} - \text{SAR},$$

where PCN-sent-rate is the value provided by the PCN-ingress-node.

See Section 3.3.3 for a discussion of appropriate actions if the Decision Point fails to receive a timely response to its request for the PCN-sent-rate.

If the difference calculated in the second step is positive, the Decision Point SHOULD select PCN-flows to terminate, until it determines that the PCN-traffic admission rate will no longer be greater than the estimated sustainable aggregate rate. If the Decision Point knows the bandwidth required by individual PCN-flows (e.g., from resource signalling used to establish the flows), it MAY choose to complete its selection of PCN-flows to terminate in a single round of decisions.

Alternatively, the Decision Point MAY spread flow termination over multiple rounds to avoid over-termination. If this is done, it is RECOMMENDED that enough time elapse between successive rounds of termination to allow the effects of previous rounds to be reflected in the measurements upon which the termination decisions are based. (See [IEEE-Sato] and sections 4.2 and 4.3 of [MeLe10].)

In general, the selection of flows for termination MAY be guided by policy. [CL-specific] If the egress node has supplied a list of identifiers of PCN-flows that experienced excess-traffic-marking (Section 3.2), the Decision Point SHOULD first consider terminating PCN-flows in that list.

3.3.3. Decision Point Action For Missing PCN-Boundary-Node Reports

The Decision Point SHOULD start a timer t-recvFail when it receives a report from the PCN-egress-node. t-recvFail is reset each time a new report is received from the PCN-egress-node. t-recvFail expires if it reaches the value T-fail. T-fail is calculated according to the following logic:

- a. T-fail = the configurable duration T-crit, if report suppression is not deployed;

- b. $T_{fail} = T_{crit}$ also if report suppression is deployed and the last report received from the PCN-egress-node contained a CLE value greater than CLE-reporting-threshold (Section 3.2.3);
- c. $T_{fail} = 3 * T_{maxsuppress}$ (Section 3.2.3) if report suppression is deployed and the last report received from the PCN-egress-node contained a CLE value less than or equal to CLE-reporting-threshold.

If timer `t-recvFail` expires for a given PCN-egress-node, the Decision Point SHOULD raise an alarm to management. A Decision Point collocated with a PCN-ingress-node SHOULD cease to admit PCN-flows to the ingress-egress-aggregate associated with the given PCN-egress-node, until it again receives a report from that node. A centralized Decision Point MAY cease to admit PCN-flows to all ingress-egress-aggregates destined to the PCN-egress-node concerned, until it again receives a report from that node.

A centralized Decision Point SHOULD start a timer `t-sndFail` when it sends a request for the estimated value of PCN-sent-rate to a given PCN-ingress-node. If the Decision Point fails to receive a response from the PCN-ingress-node before `t-sndFail` reaches the configurable value T_{crit} , the Decision Point SHOULD repeat the request but MAY also use ETM-rate as an estimate of the amount of traffic to be terminated in place of the quantity

`PCN-sent-rate - SAR`

specified in Section 3.3.2. Because this will over-estimate the amount of traffic to be terminated due to dropping of PCN-packets by interior nodes, the Decision Point SHOULD use multiple rounds of termination under these circumstances. If the second request to the PCN-ingress-node also fails, the Decision Point SHOULD raise an alarm to management.

The use of T_{crit} is an approximation. A more precise limit would be of the order of two round-trip times, plus an allowance for processing at each end, plus an allowance for variance in these values.

See Section 3.5 for suggested values of the configurable durations T_{crit} and $T_{maxsuppress}$.

3.4. Behaviour of the Ingress Node

The PCN-ingress-node 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) when the

Decision Point requests it. The way this rate estimate is derived is a matter of implementation.

For example, the rate that the PCN-ingress-node supplies MAY be based on a quick sample taken at the time the information is required.

3.5. Summary of Timers and Associated Configurable Durations

Here is a summary of the timers used in the procedures just described:

t-meas

Where used: PCN-egress-node.

Used in procedure: data collection (Section 3.2.1).

Incidence: one per ingress-egress-aggregate.

Reset: immediately on expiry.

Expiry: when it reaches the configurable duration T-meas.

Action on expiry: calculate NM-rate, [CL-specific] ThM-rate, and ETM-rate and proceed to the applicable reporting procedure (Section 3.2.2 or Section 3.2.3).

t-maxsuppress

Where used: PCN-egress-node.

Used in procedure: report suppression (Section 3.2.3).

Incidence: one per ingress-egress-aggregate.

Reset: when the next report is sent after expiry.

Expiry: when it reaches the configurable duration T-maxsuppress.

Action on expiry: send a report to the Decision Point the next time the reporting procedure (Section 3.2.3) is invoked, regardless of the value of CLE.

t-recvFail

Where used: Decision Point.

Used in procedure: failure detection (Section 3.3.3).

Incidence: one per ingress-egress-aggregate.

Reset: when a report is received for the ingress-egress-aggregate.

Expiry: when it reaches the calculated duration T-fail. As described in Section 3.3.3, T-fail is either equal to the configured duration T-crit or to the calculated value $3 * T-maxsuppress$, where T-maxsuppress is a configured duration.

Action on expiry: raise an alarm to management, and possibly other actions.

t-sndFail

Where used: centralized Decision Point.

Used in procedure: failure detection (Section 3.3.3).

Incidence: only as required, one per outstanding request to a PCN-ingress-node.

Started: when a request for the value of PCN-sent-traffic for a given ingress-egress-aggregate is sent to the PCN-ingress-node.

Terminated without action: when a response is received before expiry.

Expiry: when it reaches the configured duration T-crit.

Action on expiry: repeat the request, but use an approximation for the estimate of amount of traffic to terminate. After two failures, raise an alarm to management and stop repeating the request.

3.5.1. Recommended Values For the Configurable Durations

The timers just described depend on three configurable durations, T-meas, T-maxsuppress, and T-crit. The recommendations given below for the values of these durations are all related to the intended PCN reaction time of 1 to 3 seconds. However, they are based on judgement rather than operational experience or mathematical

derivation.

The value of T-meas is RECOMMENDED to be of the order of 100 to 500 ms to provide a reasonable tradeoff between demands on network resources (PCN-egress-node and Decision Point processing, network bandwidth) and the time taken to react to impending congestion.

The value of T-maxsuppress is RECOMMENDED to be on the order of 3 to 6 seconds, for similar reasons to those for the choice of T-meas.

The value of T-crit SHOULD NOT be less than $3 * T\text{-meas}$. Otherwise it could cause too many alarms to be raised due to transient conditions in the PCN-egress-node or along the signalling path. A reasonable upper bound on T-crit is in the order of 3 seconds.

4. Identifying Ingress and Egress Nodes For PCN Traffic

The operation of PCN depends on the ability of the PCN-ingress-node to identify the ingress-egress-aggregate to which each new PCN-flow belongs and the ability of the egress node to identify the ingress-egress-aggregate to which each received PCN-packet belongs. If the Decision Point is collocated with the PCN-ingress-node, the PCN-egress-node also needs to associate each ingress-egress-aggregate with the address of the PCN-ingress-node to which it sends its reports.

The means by which this is done depends on the packet routing technology in use in the network. The procedure to provide the required information is out of scope for this document.

5. Specification of Diffserv Per-Domain Behaviour

This section provides the specification required by [RFC3086] for a per-domain behaviour.

5.1. Applicability

This section quotes [RFC5559].

The PCN CL boundary node behaviour specified in this document is applicable to inelastic traffic (particularly video and voice) where quality of service for admitted flows is protected primarily by admission control at the ingress to the domain.

In exceptional circumstances (e.g., due to rerouting as a result of network failures) already-admitted flows MAY be terminated to protect

the quality of service of the remaining flows. [CL-specific] The performance results in, e.g., [MeLe10], indicate that the CL boundary node behaviour provides better service outcomes under such circumstances than the SM boundary node behaviour described in [RFCyyyy], because CL is less likely to terminate PCN-flows unnecessarily.

[RFC EDITOR'S NOTE: please replace RFCyyyy above by the reference to the published version of draft-ietf-pcn-sm-edge-behaviour.]

5.2. Technical Specification

5.2.1. Classification and Traffic Conditioning

This section paraphrases the applicable portions of Sections 3.6 and 4.2 of [RFC5559].

Packets at the ingress to the domain are classified as either PCN or non-PCN. Non-PCN packets MAY share the network with PCN packets within the domain. Because the encoding specified in [RFC5696] and used in this document requires the use of the ECN fields, PCN-ingress-nodes MUST prevent ECN-capable traffic that uses the same DSCP as PCN from entering the PCN-domain directly. The PCN-ingress-node can accomplish this in three ways. The choice between these depends on local policy.

- o ECN-capable traffic MAY be dropped. This policy is NOT RECOMMENDED, since it prevents the proper operation of end-to-end ECN as a means of controlling congestion.
- o ECN-capable traffic MAY be assigned a different DSCP from PCN traffic. This could mean that it is relegated to a lower-priority behaviour aggregate.
- o ECN-capable traffic MAY be tunneled across the PCN-domain. If this is done, the PCN-ingress-node MUST mark packets as either not-PCN or PCN-not-marked only after the encapsulation of the packet, including any initial setting of the ECN field, has been completed.

PCN packets are further classified as belonging or not belonging to an admitted flow. PCN packets not belonging to an admitted flow are dropped. (This assumes that requests for flow admission are signalled in advance of the arrival of the flows themselves.) Packets belonging to an admitted flow are policed to ensure that they adhere to the rate or flowspec that was negotiated during flow admission.

5.2.2. PHB Configuration

The PCN CL boundary node behaviour is a metering and marking behaviour rather than a scheduling behaviour. As a result, while the encoding uses a single DSCP value, that value MAY vary from one deployment to another. The PCN working group suggests using admission control for the following service classes (defined in [RFC4594]):

- o Telephony (EF)
- o Real-time interactive (CS4)
- o Broadcast Video (CS3)
- o Multimedia Conferencing (AF4)

For a fuller discussion, see Section A.1 of Appendix A of [RFC5696].

5.3. Attributes

The purpose of this per-domain behaviour is to achieve low loss and jitter for the target class of traffic. The design requirement for PCN was that recovery from overloads through the use of flow termination SHOULD happen within 1-3 seconds. PCN probably performs better than that.

5.4. Parameters

In the list that follows, note that most PCN-ingress-nodes are also PCN-egress-nodes, and vice versa. Furthermore, the PCN-ingress-nodes MAY be collocated with Decision Points.

Parameters at the PCN-ingress-node:

- o Filters for distinguishing PCN from non-PCN inbound traffic.
- o The markings to be applied to PCN-traffic.
- o Reference rates on each link for the [CL-specific] threshold-meter and the excess-traffic-meter; see Section 2.
- o The information needed to distinguish PCN-traffic belonging to a given ingress-egress-aggregate.

Parameters at the PCN-egress-node:

- o The measurement interval T-meas.
- o Activation/deactivation of report suppression and, if report suppression is activated, the values of the CLE-reporting-threshold and T-maxsuppress.
- o [CL-specific] Activation/deactivation of recording of individual flow identifiers when excess-traffic-marked PCN-traffic is observed.
- o The information needed to distinguish PCN-traffic belonging to a given ingress-egress-aggregate.
- o The marking rules for re-marking PCN-traffic leaving the PCN domain.

Parameters at each interior node:

- o Reference rates on each link for the [CL-specific] threshold-meter and the excess-traffic-meter; see Section 2.
- o The markings to be applied to PCN-traffic, including the identification of PCN-packets and the encodings to indicate excess-traffic-marking and [CL-specific] threshold-marking.

Parameters at the Decision Point:

- o Activation/deactivation of PCN-based flow admission.
- o Activation/deactivation of PCN-based flow termination.
- o The value of CLE-limit.
- o The maximum interval T-fail between reports from a given PCN-egress-node, for detecting failure of communications with that node.
- o The information needed to map each ingress-egress-aggregate to the corresponding PCN-ingress-node and PCN-egress-node.

5.5. Assumptions

It is assumed that a specific portion of link capacity has been reserved for PCN-traffic.

5.6. Example Uses

The PCN CL behaviour MAY be used to carry real-time traffic, particularly voice and video.

5.7. Environmental Concerns

The PCN CL per-domain behaviour can interfere with the use of end-to-end ECN due to reuse of ECN bits for PCN marking. See Appendix B of [RFC5696] for details.

5.8. Security Considerations

Please see the security considerations in [RFC5559] as well as those in [RFC2474] and [RFC2475].

6. Security Considerations

[RFC5559] provides a general description of the security considerations for PCN. This memo introduces no new considerations.

7. IANA Considerations

This memo includes no request to IANA.

8. Acknowledgements

The content of this memo bears a family resemblance to [ID.briscoe-CL]. The authors of that document were Bob Briscoe, Philip Eardley, and Dave Songhurst of BT, Anna Charny and Francois Le Faucheur of Cisco, Jozef Babiarz, Kwok Ho Chan, and Stephen Dudley of Nortel, Giorgios Karagiannis of U. Twente and Ericsson, and Attila Bader and Lars Westberg of Ericsson.

Ruediger Geib, Philip Eardley, and Bob Briscoe have helped to shape the present document with their comments. Toby Moncaster gave a careful review to get it into shape for Working Group Last Call.

Amongst the authors, Michael Menth deserves special mention for his constant and careful attention to both the technical content of this document and the manner in which it was expressed.

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.
- [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 Services", RFC 2475, December 1998.
- [RFC3086] Nichols, K. and B. Carpenter, "Definition of Differentiated Services Per Domain Behaviors and Rules for their Specification", RFC 3086, April 2001.
- [RFC5559] Eardley, P., "Pre-Congestion Notification (PCN) Architecture", RFC 5559, June 2009.
- [RFC5670] Eardley, P., "Metering and Marking Behaviour of PCN-Nodes", RFC 5670, November 2009.
- [RFC5696] Moncaster, T., Briscoe, B., and M. Menth, "Baseline Encoding and Transport of Pre-Congestion Information", RFC 5696, November 2009.

9.2. Informative References

- [ID.briscoe-CL] Briscoe, B., "An edge-to-edge Deployment Model for Pre-Congestion Notification: Admission Control over a DiffServ Region (expired Internet Draft)", 2006.
- [IEEE-Satoh] Satoh, D. and H. Ueno, "'Cause and Countermeasure of Overtermination for PCN-Based Flow Termination", Proceedings of IEEE Symposium on Computers and Communications (ISCC '10), pp. 155-161, Riccione, Italy", June 2010.
- [MeLe10] Menth, M. and F. Lehrieder, "PCN-Based Measured Rate Termination", Computer Networks Journal (Elsevier) vol. 54, no. 13, pages 2099 - 2116, September 2010.
- [RFC4594] Babiarz, J., Chan, K., and F. Baker, "Configuration Guidelines for DiffServ Service Classes", RFC 4594,

August 2006.

- [RFC6040] Briscoe, B., "Tunnelling of Explicit Congestion Notification", RFC 6040, November 2010.
- [RFCyyyy] Charny, A., Zhang, J., Karagiannis, G., Menth, M., and T. Taylor, "PCN Boundary Node Behaviour for the Single Marking (SM) Mode of Operation (Work in progress)", December 2010.

Authors' Addresses

Anna Charny
Cisco Systems
300 Apollo Drive
Chelmsford, MA 01824
USA

Email: acharny@cisco.com

Fortune Huang
Huawei Technologies
Section F, Huawei Industrial Base,
Bantian Longgang, Shenzhen 518129
P.R. China

Phone: +86 15013838060
Email: fqhuang@huawei.com

Georgios Karagiannis
U. Twente

Phone:
Email: karagian@cs.utwente.nl

Michael Menth
University of Tuebingen
Sand 13
Tuebingen D-72076
Germany

Phone: +49-7071-2970505
Email: menth@informatik.uni-tuebingen.de

Tom Taylor (editor)
Huawei Technologies
1852 Lorraine Ave
Ottawa, Ontario K1H 6Z8
Canada

Phone: +1 613 680 2675
Email: toml111.taylor@bell.net

Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: December 24, 2011

A. Charny
J. Zhang
Cisco Systems
G. Karagiannis
U. Twente
M. Menth
University of Tuebingen
T. Taylor, Ed.
Huawei Technologies
June 22, 2011

PCN Boundary Node Behaviour for the Single Marking (SM) Mode of
Operation
draft-ietf-pcn-sm-edge-behaviour-06

Abstract

Pre-congestion notification (PCN) is a means for protecting the quality of service for inelastic traffic admitted to a Diffserv domain. The overall PCN architecture is described in RFC 5559. This memo is one of a series describing possible boundary node behaviours for a PCN-domain. The behaviour described here is that for a form of measurement-based load control using two PCN marking states, not-marked, and excess-traffic-marked. This behaviour is known informally as the Single Marking (SM) PCN-boundary-node behaviour.

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 24, 2011.

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	4
1.1. Terminology	4
2. [SM-Specific] Assumed Core Network Behaviour for SM	7
3. Node Behaviours	8
3.1. Overview	8
3.2. Behaviour of the PCN-Egress-Node	8
3.2.1. Data Collection	8
3.2.2. Reporting the PCN Data	9
3.2.3. Optional Report Suppression	9
3.3. Behaviour at the Decision Point	10
3.3.1. Flow Admission	10
3.3.2. Flow Termination	11
3.3.3. Decision Point Action For Missing PCN-Boundary-Node Reports	12
3.4. Behaviour of the Ingress Node	13
3.5. Summary of Timers and Associated Configurable Durations	14
3.5.1. Recommended Values For the Configurable Durations	15
4. Identifying Ingress and Egress Nodes For PCN Traffic	16
5. Specification of Diffserv Per-Domain Behaviour	16
5.1. Applicability	16
5.2. Technical Specification	17
5.2.1. Classification and Traffic Conditioning	17
5.2.2. PHB Configuration	18
5.3. Attributes	18
5.4. Parameters	18
5.5. Assumptions	19
5.6. Example Uses	20
5.7. Environmental Concerns	20
5.8. Security Considerations	20
6. Security Considerations	20
7. IANA Considerations	20
8. Acknowledgements	20
9. References	20
9.1. Normative References	20
9.2. Informative References	21

Authors' Addresses	21
------------------------------	----

1. Introduction

The objective of Pre-Congestion Notification (PCN) is to protect 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, to decide whether to admit or block a new flow request, and (in abnormal circumstances) flow termination to decide whether to terminate some of the existing flows. To achieve this, the overall rate of PCN-traffic is metered on every link in the PCN-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 PCN-boundary-nodes about incipient overloads before any congestion occurs (hence the "pre" part of "pre-congestion notification"). The level of marking allows decisions to be made about whether to admit or terminate PCN-flows. For more details see [RFC5559].

Section 3 of this document specifies a detailed set of algorithms and procedures used to implement the PCN mechanisms for the SM mode of operation. Since the algorithms depend on specific metering and marking behaviour at the interior nodes, it is also necessary to specify the assumptions made about PCN-interior-node behaviour (Section 2). Finally, because PCN uses DSCP values to carry its markings, a specification of PCN-boundary-node behaviour MUST include the per domain behaviour (PDB) template specified in [RFC3086], filled out with the appropriate content (Section 5).

[RFC EDITOR'S NOTE: you may choose to delete the following paragraph and the "[SM-specific]" tags throughout this document when publishing it, since they are present primarily to aid reviewers. RFCyyyy is the published version of draft-ietf-pcn-cl-edge-behaviour.]

A companion document [RFCyyyy] specifies the Controlled Load (CL) PCN-boundary-node behaviour. This document and [RFCyyyy] have a great deal of text in common. To simplify the task of the reader, the text in the present document that is specific to the SM PCN-boundary-node behaviour is preceded by the phrase: "[SM-specific]". A similar distinction for CL-specific text is made in [RFCyyyy].

1.1. 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 [RFC2119].

This document uses the following terms defined in Section 2 of [RFC5559]:

- o PCN-domain;
- o PCN-ingress-node;
- o PCN-egress-node;
- o PCN-interior-node;
- o PCN-boundary-node;
- o PCN-flow;
- o ingress-egress-aggregate (IEA);
- o PCN-excess-rate;
- o PCN-admissible-rate;
- o PCN-supportable-rate;
- o PCN-marked;
- o excess-traffic-marked.

It also uses the terms PCN-traffic and PCN-packet, for which the definition is repeated from [RFC5559] because of their importance to the understanding of the text that follows:

PCN-traffic, PCN-packets, PCN-BA

A PCN-domain carries traffic of different Diffserv behaviour 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 and the ECN field.

This document uses the following term from [RFC5670]:

- o excess-traffic-meter.

To complete the list of borrowed terms, this document reuses the following terms and abbreviations defined in Section 3 of [RFC5696]:

- o not-PCN codepoint;
- o Not-marked (NM) codepoint;

- o PCN-marked (PM) codepoint.

This document defines the following additional terms:

Decision Point

The node that makes the decision about which flows to admit and to terminate. In a given network deployment, this can be the PCN-ingress-node or a centralized control node. In either case, the PCN-ingress-node is the point where the decisions are enforced.

NM-rate

The rate of not-marked PCN-traffic received at a PCN-egress-node for a given ingress-egress-aggregate in octets per second. For further details see Section 3.2.1.

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. For further details see Section 3.2.1.

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. For further details see Section 3.4.

Congestion level estimate (CLE)

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 (Section 3.3.1) and is also used by the report suppression procedure (Section 3.2.3) if report suppression is activated.

PCN-admission-state

The state ("admit" or "block") derived by the Decision Point for a given ingress-egress-aggregate based on PCN packet marking statistics. The Decision Point decides to admit or block new flows offered to the aggregate based on the current value of the PCN-admission-state. For further details see Section 3.3.1.

Sustainable aggregate rate (SAR)

The estimated maximum rate of PCN-traffic that can be carried in a given ingress-egress-aggregate at a given moment without risking degradation of quality of service for the admitted flows. The intention is that if the PCN-sent-rate of every ingress-egress-aggregate passing through a given link is limited to its sustainable aggregate rate, the total rate of PCN-traffic flowing through the link will be limited to the PCN-supportable-rate for that link. An estimate of the sustainable aggregate rate for a

given ingress-egress-aggregate is derived as part of the flow termination procedure, and is used to determine how much PCN-traffic needs to be terminated. For further details see Section 3.3.2.

CLE-reporting-threshold

A configurable value against which the CLE is compared as part of the report suppression procedure. For further details, see Section 3.2.3.

CLE-limit

A configurable value against which the CLE is compared to determine the PCN-admission-state for a given ingress-egress-aggregate. For further details, see Section 3.3.1.

T-meas

A configurable time interval that defines the measurement period over which the PCN-egress-node collects statistics relating to PCN-traffic marking. At the end of the interval the PCN-egress-node calculates the values NM-rate and ETM-rate as defined and sends a report to the Decision Point, subject to the operation of the report suppression feature. For further details see Section 3.2.

T-maxsuppress

A configurable time interval after which the PCN-egress-node MUST send a report to the Decision Point for a given ingress-egress-aggregate regardless of the most recent values of the CLE. This mechanism provides the Decision Point with a periodic confirmation of liveness when report suppression is activated. For further details, see Section 3.2.3.

T-fail

A configurable interval after which the Decision Point concludes that communication from a given PCN-egress-node has failed if it has received no reports from the PCN-egress-node during that interval. For further details see Section 3.3.3.

2. [SM-Specific] Assumed Core Network Behaviour for SM

This section describes the assumed behaviour for PCN-interior-nodes in the PCN-domain. The SM mode of operation assumes that:

- o PCN-interior-nodes perform excess-traffic-marking of PCN-packets according to the rules specified in [RFC5670].

- o excess-traffic-marking of PCN-packets uses the PCN-Marked (PM) codepoint defined in [RFC5696];
- o the PCN-domain satisfies the conditions specified in [RFC5696];
- o on each link the reference rate for the excess-traffic-meter is configured to be equal to the PCN-admissible-rate for the link;
- o the set of valid codepoint transitions is as shown in Section 4.2 of [RFC5696].

3. Node Behaviours

3.1. Overview

This section describes the behaviour of the PCN-ingress-node, PCN-egress-node, and the Decision Point (which MAY be collocated with the PCN-ingress-node).

The PCN-egress-node collects the rates of not-marked and excess-traffic-marked PCN-traffic for each ingress-egress-aggregate and reports them to the Decision Point. For a detailed description, see Section 3.2.

The PCN-ingress-node enforces flow admission and termination decisions. It also reports the rate of PCN-traffic sent to a given ingress-egress-aggregate when requested by the Decision Point. For details, see Section 3.4.

Finally, the Decision Point makes flow admission decisions and selects flows to terminate based on the information provided by the PCN-ingress-node and PCN-egress-node for a given ingress-egress-aggregate. For details, see Section 3.3.

3.2. Behaviour of the PCN-Egress-Node

3.2.1. Data Collection

The PCN-egress-node MUST meter the PCN-traffic it receives in order to calculate the following rates for each ingress-egress-aggregate passing through it. These rates SHOULD be calculated at the end of each measurement period based on the PCN-traffic observed during that measurement period. The duration of a measurement period is equal to the configurable value T-meas. For further information see Section 3.5.

- o NM-rate: octets per second of PCN-traffic in PCN-packets that are not-marked (i.e., marked with the NM codepoint);
- o ETM-rate: octets per second of PCN-traffic in PCN-packets that are excess-traffic-marked (i.e., marked with the PM codepoint).

Informative note: metering the PCN-traffic continuously and using equal-length measurement intervals minimizes the statistical variance introduced by the measurement process itself. On the other hand, the operation of PCN is not affected if the starting and ending times of the measurement intervals for different ingress-egress-aggregates are different.

3.2.2. Reporting the PCN Data

Unless the report suppression option described in Section 3.2.3 is activated, the PCN-egress-node MUST report the latest values of NM-rate and ETM-rate to the Decision Point each time that it calculates them.

3.2.3. Optional Report Suppression

Report suppression MUST be provided as a configurable option, along with two configurable parameters, the CLE-reporting-threshold and the maximum report suppression interval T-maxsuppress. The default value of the CLE-reporting-threshold is zero. The CLE-reporting-threshold MUST NOT exceed the CLE-limit configured at the Decision Point. For further information on T-maxsuppress see Section 3.5.

If the report suppression option is enabled, the PCN-egress-node MUST apply the following procedure to decide whether to send a report to the Decision Point, rather than sending a report automatically at the end of each measurement interval.

1. As well as the quantities NM-rate and ETM-rate, the PCN-egress-node MUST calculate the congestion level estimate (CLE) for each measurement interval. The CLE is computed as:

[SM-specific]
$$\text{CLE} = \text{ETM-rate} / (\text{NM-rate} + \text{ETM-rate})$$

if any PCN-traffic was observed, or CLE = 0 if all the rates are zero.

2. If the CLE calculated for the latest measurement interval is greater than the CLE-reporting-threshold and/or the CLE calculated for the immediately previous interval was greater than the CLE-reporting-threshold, then the PCN-egress-node MUST send a

report to the Decision Point. The contents of the report are described below.

The reason for taking into account the CLE of the previous interval is to ensure that the Decision Point gets immediate feedback if the CLE has dropped below the CLE-reporting-threshold. This is essential if the Decision Point is running the flow termination procedure and observing whether (further) flow termination is needed. See Section 3.3.2.

3. If an interval T-maxsuppress has elapsed since the last report was sent to the Decision Point, then the PCN-egress-node MUST send a report to the Decision Point regardless of the CLE value.
4. If neither of the preceding conditions holds, the PCN-egress-node MUST NOT send a report for the latest measurement interval.

Each report sent to the Decision Point when report suppression has been activated MUST contain the values of NM-rate, ETM-rate, and CLE that were calculated for the most recent measurement interval.

The above procedure ensures that at least one report is sent per interval (T-maxsuppress + T-meas). This demonstrates to the Decision Point that both the PCN-egress-node and the communication path between that node and the Decision Point are in operation.

3.3. Behaviour at the Decision Point

Operators can choose to use PCN procedures just for flow admission, or just for flow termination, or for both. A compliant Decision Point MUST implement both mechanisms, but configurable options MUST be provided to activate or deactivate PCN-based flow admission and flow termination independently of each other at a given Decision Point.

If PCN-based flow termination is enabled but PCN-based flow admission is not, flow termination operates as specified in this document.

Logically, some other system of flow admission control is in operation, but the description of such a system is out of scope of this document and depends on local arrangements.

3.3.1. Flow Admission

The Decision Point determines the PCN-admission-state for a given ingress-egress-aggregate each time it receives a report from the egress node. It makes this determination on the basis of the congestion level estimate (CLE). If the CLE is provided in the

egress node report, the Decision Point SHOULD use the reported value. If the CLE was not provided in the report, the Decision Point MUST calculate it based on the other values provided in the report, using the formula:

```
[SM-specific]
CLE = ETM-rate / (NM-rate + ETM-rate)
```

if any PCN-traffic was observed, or CLE = 0 if all the rates are zero.

The Decision Point MUST compare the reported or calculated CLE to a configurable value, the CLE-limit. If the CLE is less than the CLE-limit, the PCN-admission-state for that aggregate MUST be set to "admit"; otherwise it MUST be set to "block".

```
[SM-specific] It is RECOMMENDED that the CLE-limit for SM be set
fairly low, in the order of 0.05.
```

If the PCN-admission-state for a given ingress-egress-aggregate is "admit", the Decision Point SHOULD allow new flows to be admitted to that aggregate. If the PCN-admission-state for a given ingress-egress-aggregate is "block", the Decision Point SHOULD NOT allow new flows to be admitted to that aggregate. These actions MAY be modified by policy in specific cases, but such policy intervention risks defeating the purpose of using PCN.

3.3.2. Flow Termination

```
[SM-specific] When the PCN-admission-state computed on the basis of
the CLE is "block" for the given ingress-egress-aggregate, the
Decision Point MUST request the PCN-ingress-node to provide an
estimate of the rate (PCN-sent-rate) at which the PCN-ingress-node is
receiving PCN-traffic that is destined for the given ingress-egress-
aggregate.
```

If the Decision Point is collocated with the PCN-ingress-node, the request and response are internal operations.

The Decision Point MUST then wait, for both the requested rate from the PCN-ingress-node and the next report from the PCN-egress-node for the ingress-egress-aggregate concerned. If this next egress node report also includes a non-zero value for the ETM-rate, the Decision Point MUST determine the amount of PCN-traffic to terminate using the following steps:

1. [SM-specific] The sustainable aggregate rate (SAR) for the given ingress-egress-aggregate is estimated using the formula:

$$\text{SAR} = U * \text{NM-Rate}$$

for the latest reported interval, where U is a configurable factor greater than one which is the same for all ingress-egress-aggregates. U represents the average ratio of PCN-supportable-rate to PCN-admissible-rate over all the links of the PCN-domain.

2. The amount of traffic to be terminated is the difference:

$$\text{PCN-sent-rate} - \text{SAR},$$

where PCN-sent-rate is the value provided by the PCN-ingress-node.

See Section 3.3.3 for a discussion of appropriate actions if the Decision Point fails to receive a timely response to its request for the PCN-sent-rate.

If the difference calculated in the second step is positive, the Decision Point SHOULD select PCN-flows to terminate, until it determines that the PCN-traffic admission rate will no longer be greater than the estimated sustainable aggregate rate. If the Decision Point knows the bandwidth required by individual PCN-flows (e.g., from resource signalling used to establish the flows), it MAY choose to complete its selection of PCN-flows to terminate in a single round of decisions.

Alternatively, the Decision Point MAY spread flow termination over multiple rounds to avoid over-termination. If this is done, it is RECOMMENDED that enough time elapse between successive rounds of termination to allow the effects of previous rounds to be reflected in the measurements upon which the termination decisions are based. (See [IEEE-Satoh] and sections 4.2 and 4.3 of [MeLe10].)

In general, the selection of flows for termination MAY be guided by policy.

3.3.3. Decision Point Action For Missing PCN-Boundary-Node Reports

The Decision Point SHOULD start a timer t-recvFail when it receives a report from the PCN-egress-node. t-recvFail is reset each time a new report is received from the PCN-egress-node. t-recvFail expires if it reaches the value T-fail. T-fail is calculated according to the following logic:

- a. T-fail = the configurable duration T-crit, if report suppression is not deployed;

- b. $T_{fail} = T_{crit}$ also if report suppression is deployed and the last report received from the PCN-egress-node contained a CLE value greater than CLE-reporting-threshold (Section 3.2.3);
- c. $T_{fail} = 3 * T_{maxsuppress}$ (Section 3.2.3) if report suppression is deployed and the last report received from the PCN-egress-node contained a CLE value less than or equal to CLE-reporting-threshold.

If timer $t_{recvFail}$ expires for a given PCN-egress-node, the Decision Point SHOULD raise an alarm to management. A Decision Point collocated with a PCN-ingress-node SHOULD cease to admit PCN-flows to the ingress-egress-aggregate associated with the given PCN-egress-node, until it again receives a report from that node. A centralized Decision Point MAY cease to admit PCN-flows to all ingress-egress-aggregates destined to the PCN-egress-node concerned, until it again receives a report from that node.

A centralized Decision Point SHOULD start a timer $t_{sndFail}$ when it sends a request for the estimated value of PCN-sent-rate to a given PCN-ingress-node. If the Decision Point fails to receive a response from the PCN-ingress-node before $t_{sndFail}$ reaches the configurable value T_{crit} , the Decision Point SHOULD repeat the request but MAY also use ETM-rate as an estimate of the amount of traffic to be terminated in place of the quantity

PCN-sent-rate - SAR

specified in Section 3.3.2. Because this will over-estimate the amount of traffic to be terminated due to dropping of PCN-packets by interior nodes, the Decision Point SHOULD use multiple rounds of termination under these circumstances. If the second request to the PCN-ingress-node also fails, the Decision Point SHOULD raise an alarm to management.

The use of T_{crit} is an approximation. A more precise limit would be of the order of two round-trip times, plus an allowance for processing at each end, plus an allowance for variance in these values.

See Section 3.5 for suggested values of the configurable durations T_{crit} and $T_{maxsuppress}$.

3.4. Behaviour of the Ingress Node

The PCN-ingress-node 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) when the

Decision Point requests it. The way this rate estimate is derived is a matter of implementation.

For example, the rate that the PCN-ingress-node supplies MAY be based on a quick sample taken at the time the information is required.

3.5. Summary of Timers and Associated Configurable Durations

Here is a summary of the timers used in the procedures just described:

t-meas

Where used: PCN-egress-node.

Used in procedure: data collection (Section 3.2.1).

Incidence: one per ingress-egress-aggregate.

Reset: immediately on expiry.

Expiry: when it reaches the configurable duration T-meas.

Action on expiry: calculate NM-rate, [CL-specific] ThM-rate, and ETM-rate and proceed to the applicable reporting procedure (Section 3.2.2 or Section 3.2.3).

t-maxsuppress

Where used: PCN-egress-node.

Used in procedure: report suppression (Section 3.2.3).

Incidence: one per ingress-egress-aggregate.

Reset: when the next report is sent after expiry.

Expiry: when it reaches the configurable duration T-maxsuppress.

Action on expiry: send a report to the Decision Point the next time the reporting procedure (Section 3.2.3) is invoked, regardless of the value of CLE.

t-recvFail

Where used: Decision Point.

Used in procedure: failure detection (Section 3.3.3).

Incidence: one per ingress-egress-aggregate.

Reset: when a report is received for the ingress-egress-aggregate.

Expiry: when it reaches the calculated duration T-fail. As described in Section 3.3.3, T-fail is either equal to the configured duration T-crit or to the calculated value $3 * T-maxsuppress$, where T-maxsuppress is a configured duration.

Action on expiry: raise an alarm to management, and possibly other actions.

t-sndFail

Where used: centralized Decision Point.

Used in procedure: failure detection (Section 3.3.3).

Incidence: only as required, one per outstanding request to a PCN-ingress-node.

Started: when a request for the value of PCN-sent-traffic for a given ingress-egress-aggregate is sent to the PCN-ingress-node.

Terminated without action: when a response is received before expiry.

Expiry: when it reaches the configured duration T-crit.

Action on expiry: repeat the request, but use an approximation for the estimate of amount of traffic to terminate. After two failures, raise an alarm to management and stop repeating the request.

3.5.1. Recommended Values For the Configurable Durations

The timers just described depend on three configurable durations, T-meas, T-maxsuppress, and T-crit. The recommendations given below for the values of these durations are all related to the intended PCN reaction time of 1 to 3 seconds. However, they are based on judgement rather than operational experience or mathematical

derivation.

The value of T-meas is RECOMMENDED to be of the order of 100 to 500 ms to provide a reasonable tradeoff between demands on network resources (PCN-egress-node and Decision Point processing, network bandwidth) and the time taken to react to impending congestion.

The value of T-maxsuppress is RECOMMENDED to be on the order of 3 to 6 seconds, for similar reasons to those for the choice of T-meas.

The value of T-crit SHOULD NOT be less than $3 * T\text{-meas}$. Otherwise it could cause too many alarms to be raised due to transient conditions in the PCN-egress-node or along the signalling path. A reasonable upper bound on T-crit is in the order of 3 seconds.

4. Identifying Ingress and Egress Nodes For PCN Traffic

The operation of PCN depends on the ability of the PCN-ingress-node to identify the ingress-egress-aggregate to which each new PCN-flow belongs and the ability of the egress node to identify the ingress-egress-aggregate to which each received PCN-packet belongs. If the Decision Point is collocated with the PCN-ingress-node, the PCN-egress-node also needs to associate each ingress-egress-aggregate with the address of the PCN-ingress-node to which it sends its reports.

The means by which this is done depends on the packet routing technology in use in the network. The procedure to provide the required information is out of scope for this document.

5. Specification of Diffserv Per-Domain Behaviour

This section provides the specification required by [RFC3086] for a per-domain behaviour.

5.1. Applicability

This section quotes [RFC5559].

The PCN SM boundary node behaviour specified in this document is applicable to inelastic traffic (particularly video and voice) where quality of service for admitted flows is protected primarily by admission control at the ingress to the domain.

In exceptional circumstances (e.g., due to rerouting as a result of network failures) already-admitted flows MAY be terminated to protect

the quality of service of the remaining flows. [SM-specific] The performance results in, e.g., [MeLe10], indicate that the SM boundary node behaviour is more likely to terminate too many flows under such circumstances than the CL boundary node behaviour described in [RFCyyyy].

[RFC EDITOR'S NOTE: please replace RFCyyyy above by the reference to the published version of draft-ietf-pcn-cl-edge-behaviour.]

5.2. Technical Specification

5.2.1. Classification and Traffic Conditioning

This section paraphrases the applicable portions of Sections 3.6 and 4.2 of [RFC5559].

Packets at the ingress to the domain are classified as either PCN or non-PCN. Non-PCN packets MAY share the network with PCN packets within the domain. Because the encoding specified in [RFC5696] and used in this document requires the use of the ECN fields, PCN-ingress-nodes MUST prevent ECN-capable traffic that uses the same DSCP as PCN from entering the PCN-domain directly. The PCN-ingress-node can accomplish this in three ways. The choice between these depends on local policy.

- o ECN-capable traffic MAY be dropped. This policy is NOT RECOMMENDED, since it prevents the proper operation of end-to-end ECN as a means of controlling congestion.
- o ECN-capable traffic MAY be assigned a different DSCP from PCN traffic. This could mean that it is relegated to a lower-priority behaviour aggregate.
- o ECN-capable traffic MAY be tunneled across the PCN-domain. If this is done, the PCN-ingress-node MUST mark packets as either not-PCN or PCN-not-marked only after the encapsulation of the packet, including any initial setting of the ECN field, has been completed.

PCN packets are further classified as belonging or not belonging to an admitted flow. PCN packets not belonging to an admitted flow are dropped. (This assumes that requests for flow admission are signalled in advance of the arrival of the flows themselves.) Packets belonging to an admitted flow are policed to ensure that they adhere to the rate or flowspec that was negotiated during flow admission.

5.2.2. PHB Configuration

The PCN SM boundary node behaviour is a metering and marking behaviour rather than a scheduling behaviour. As a result, while the encoding uses a single DSCP value, that value MAY vary from one deployment to another. The PCN working group suggests using admission control for the following service classes (defined in [RFC4594]):

- o Telephony (EF)
- o Real-time interactive (CS4)
- o Broadcast Video (CS3)
- o Multimedia Conferencing (AF4)

For a fuller discussion, see Section A.1 of Appendix A of [RFC5696].

5.3. Attributes

The purpose of this per-domain behaviour is to achieve low loss and jitter for the target class of traffic. The design requirement for PCN was that recovery from overloads through the use of flow termination SHOULD happen within 1-3 seconds. PCN probably performs better than that.

5.4. Parameters

In the list that follows, note that most PCN-ingress-nodes are also PCN-egress-nodes, and vice versa. Furthermore, the PCN-ingress-nodes MAY be collocated with Decision Points.

Parameters at the PCN-ingress-node:

- o Filters for distinguishing PCN from non-PCN inbound traffic.
- o The markings to be applied to PCN-traffic.
- o The reference rate on each link for the excess-traffic-meter; see Section 2.
- o The information needed to distinguish PCN-traffic belonging to a given ingress-egress-aggregate.

Parameters at the PCN-egress-node:

- o The measurement interval T-meas.
- o Activation/deactivation of report suppression and, if report suppression is activated, the values of the CLE-reporting-threshold and T-maxsuppress.
- o The information needed to distinguish PCN-traffic belonging to a given ingress-egress-aggregate.
- o The marking rules for re-marking PCN-traffic leaving the PCN domain.

Parameters at each interior node:

- o Reference rate on each link for the excess-traffic-meter; see Section 2.
- o The markings to be applied to PCN-traffic, including the identification of PCN-packets and the encoding to indicate excess-traffic-marking.

Parameters at the Decision Point:

- o Activation/deactivation of PCN-based flow admission.
- o Activation/deactivation of PCN-based flow termination.
- o The value of CLE-limit.
- o The fraction U used to derive the supportable aggregate rate (SAR) from the NM-rate;
- o The maximum interval T-fail between reports from a given PCN-egress-node, for detecting failure of communications with that node.
- o The information needed to map each ingress-egress-aggregate to the corresponding PCN-ingress-node and PCN-egress-node.

5.5. Assumptions

It is assumed that a specific portion of link capacity has been reserved for PCN-traffic.

5.6. Example Uses

The PCN SM behaviour MAY be used to carry real-time traffic, particularly voice and video.

5.7. Environmental Concerns

The PCN SM per-domain behaviour can interfere with the use of end-to-end ECN due to reuse of ECN bits for PCN marking. See Appendix B of [RFC5696] for details.

5.8. Security Considerations

Please see the security considerations in [RFC5559] as well as those in [RFC2474] and [RFC2475].

6. Security Considerations

[RFC5559] provides a general description of the security considerations for PCN. This memo introduces no new considerations.

7. IANA Considerations

This memo includes no request to IANA.

8. Acknowledgements

Ruediger Geib, Philip Eardley, and Bob Briscoe have helped to shape the present document with their comments. Toby Moncaster gave a careful review to get it into shape for Working Group Last Call.

Amongst the authors, Michael Menth deserves special mention for his constant and careful attention to both the technical content of this document and the manner in which it was expressed.

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.
- [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 Services", RFC 2475, December 1998.
- [RFC3086] Nichols, K. and B. Carpenter, "Definition of Differentiated Services Per Domain Behaviors and Rules for their Specification", RFC 3086, April 2001.
- [RFC5559] Eardley, P., "Pre-Congestion Notification (PCN) Architecture", RFC 5559, June 2009.
- [RFC5670] Eardley, P., "Metering and Marking Behaviour of PCN-Nodes", RFC 5670, November 2009.
- [RFC5696] Moncaster, T., Briscoe, B., and M. Menth, "Baseline Encoding and Transport of Pre-Congestion Information", RFC 5696, November 2009.

9.2. Informative References

- [IEEE-Satoh] Satoh, D. and H. Ueno, "'Cause and Countermeasure of Overtermination for PCN-Based Flow Termination", Proceedings of IEEE Symposium on Computers and Communications (ISCC '10), pp. 155-161, Riccione, Italy", June 2010.
- [MeLe10] Menth, M. and F. Lehrieder, "PCN-Based Measured Rate Termination", Computer Networks Journal (Elsevier) vol. 54, no. 13, pages 2099 - 2116, September 2010.
- [RFC4594] Babiarz, J., Chan, K., and F. Baker, "Configuration Guidelines for DiffServ Service Classes", RFC 4594, August 2006.
- [RFCyyyy] Charny, A., Karagiannis, G., Menth, M., Huang, F., and T. Taylor, "PCN Boundary Node Behaviour for the Controlled Load (CL) Mode of Operation (Work in progress)", December 2010.

Authors' Addresses

Anna Charny
Cisco Systems
300 Apollo Drive
Chelmsford, MA 01824
USA

Email: acharny@cisco.com

Xinyan (Joy) Zhang
Cisco Systems
300 Apollo Drive
Chelmsford, MA 01824
USA

Georgios Karagiannis
U. Twente

Phone:
Email: karagian@cs.utwente.nl

Michael Menth
University of Tuebingen
Sand 13
Tuebingen D-72076
Germany

Phone: +49-7071-2970505
Email: menth@informatik.uni-tuebingen.de

Tom Taylor (editor)
Huawei Technologies
1852 Lorraine Ave
Ottawa, Ontario K1H 6Z8
Canada

Phone:
Email: tom111.taylor@bell.net

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: January 11, 2012

Georgios Karagiannis
University of Twente
Anurag Bhargava
Cisco Systems, Inc.
July 11, 2011

Generic Aggregation of Resource ReSerVation Protocol (RSVP)
for IPv4 And IPv6 Reservations over PCN domains
draft-karagiannis-pcn-tsvwg-rsvp-pcn-01

Abstract

This document specifies the extensions to the 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.

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 12, 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.

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	
1.1.	Terminology	
2.	Overview of RSVP extensions and Operations	
2.1	Overview of RSVP Aggregation Procedures in PCN domains	
2.1.1	PCN Marking and encoding and transport of pre-congestion Information	
2.1.2.	Traffic Classification Within The Aggregation Region	
2.1.3.	Deaggregator (PCN-egress-node) Determination	
2.1.4.	Mapping E2E Reservations Onto Aggregate Reservations	
2.1.5.	Size of Aggregate Reservations	
2.1.6.	E2E Path ADSPEC update	
2.1.7.	Intra-domain Routes	
2.1.8.	Inter-domain Routes	
2.1.9.	Reservations for Multicast Sessions	
2.1.10.	Multi-level Aggregation	
2.1.11.	Reliability Issues	
2.1.12.	Message Integrity and Node Authentication	
3.	Elements of Procedure	
3.1.	Receipt of E2E Path Message By PCN-ingress-node (aggregating router)	
3.2.	Handling Of E2E Path Message By Interior Routers	
3.3.	Receipt of E2E Path Message By PCN-egress-node (deaggregating router)	
3.4.	Initiation of new Aggregate Path Message By PCN-ingress node (Aggregating Router)	
3.5.	Handling Of new Aggregate Path Message By Interior Routers	
3.6.	Handling of E2E Resv Message by Deaggregating Router	
3.7.	Handling Of E2E Resv Message By Interior Routers	
3.8.	Initiation of New Aggregate Resv Message By Deaggregating Router	

3.9.	Handling of Aggregate Resv Message by Interior Routers	
3.10.	Handling of E2E Resv Message by Aggregating Router	
3.11.	Handling of Aggregated Resv Message by Aggregating Router	
3.12.	Removal of E2E Reservation	
3.13.	Removal of Aggregate Reservation	
3.14.	Handling of Data On Reserved E2E Flow by Aggregating Router	
3.15.	Procedures for Multicast Sessions	
4.	Protocol Elements	
4.1	PCN object	
5.	Security Considerations	
6.	IANA Considerations	
7.	Acknowledgments	
8.	Normative References	
9.	Informative References	
10.	Authors' Address	

1. Introduction

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, policy-based admission control, assistance in traffic identification/classification, and traffic conditioning.

Intserv over Diffserv can operate over a statically provisioned Diffserv region or RSVP aware. 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.

RFC 3175 [RFC3175] specifies aggregation of Resource ReSeRvation Protocol (RSVP) end-to-end reservations over aggregate RSVP reservations. In [RFC3175] the RSVP aggregated reservation is characterized by a RSVP SESSION object using the 3-tuple <source IP address, destination IP address, Diffserv Code Point>.

[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 (or set of PHBs) 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. This is among others used to specify whether the Diffserv resources belong to a single PHB or to a set of PHBs.

The main objective of Pre-Congestion Notification (PCN) is to 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, based on which they take their decisions. The decision points may be collocated with the PCN-ingress-nodes or their function may be implemented in a centralized node. For more details see[RFC5559], [draft-ietf-pcn-cl-edge-behaviour-09], [draft-ietf-pcn-sm-edge-behaviour-06]. In this document it is considered that the decision point is collocated with the PCN-ingress-node.

This document follows the PCN signaling requirements defined in [draft-ietf-pcn-signaling-requirements-06.txt] and specifies the extensions to the Generic Aggregated RSVP [RFC4860] for the support of PCN edge behaviours as specified in [draft-ietf-pcn-cl-edge-behaviour-09] and [draft-ietf-pcn-sm-edge-behaviour-06]. 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.

This document, and according to [RFC4860] MAY also be used end-to-end directly by end-systems attached to a Diffserv network.

Furthermore, this document and according to [RFC4860], in absence of e2e RSVP flows, a variety of policies (not defined in this document) can be used at the Aggregator to set the DSCP of packets passing into the aggregation region and how they are mapped onto generic aggregate reservations. These policies are not described in this document but are a matter of local configuration.

In this document it is considered that the PCN-nodes MUST be able to support the functionality specified in [RFC5670], [RFC5559], [RFC5696], [draft-ietf-pcn-cl-edge-behaviour-09], [draft-ietf-pcn-sm-edge-behaviour-06]. 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.

1.1. Terminology

This document uses terms defined in [RFC4860], [RFC3175], [RFC5559], [RFC5670], [draft-ietf-pcn-cl-edge-behaviour-09], [draft-ietf-pcn-sm-edge-behaviour-06].

For readability, a number of definitions from [RFC3175] as well as definitions for terms used in [RFC5559], [draft-ietf-pcn-cl-edge-behaviour-09], and [draft-ietf-pcn-sm-edge-behaviour-06] 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 and the decision point.

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.

E2E End to end

E2E Reservation This is an RSVP reservation such that:

- (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**

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.

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].

VDstPort (Virtual Destination Port)

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

Extended vDstPort (Extended Virtual Destination Port)

A 32-bit identifier used in the SESSION that remains constant over the life of the generic aggregate reservation. 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.

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 for the PCN-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.
- 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 Decision Point and aggregator.
- PCN-traffic,
PCN-packets,
PCN-BA: a PCN-domain carries traffic of different Diffserv behaviour 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 microflow (as defined in [RFC2474]) or some identifiable collection of microflows.
- 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. An ingress-egress-aggregate is identified by the combination of (1) fields), (2) IP addresses of the specific pair of PCN-boundary-nodes used by a ingress-egress-aggregate. In this document the ingress-egress-aggregate is associated with a RSVP generic aggregated reservation state [RFC4860].
- PCN-admission-state
The state ("admit" or "block") derived by the Decision Point (PCN-ingress-node) for a given ingress-egress-aggregate based on PCN packet marking statistics. The Decision Point decides to admit or block new flows offered to the aggregate based on the current value of the PCN-admission-state.
- Congestion level estimate (CLE)
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.

T-meas

A configurable time interval that defines the measurement period over which the PCN-egress-node collects statistics relating to PCN-traffic marking.

At the end of the interval the PCN-egress-node calculates the values NM-rate, ThM-rate, and ETM-rate as defined and sends a report to the Decision Point, subject to the operation of the Report suppression feature.

T-maxsuppress

A configurable time interval after which the PCN-egress-node MUST send a report to the Decision Point for a given ingress-egress-aggregate regardless of the most recent values of the CLE. This mechanism provides the Decision Point with a Periodic confirmation of liveness when report suppression is activated.

T-fail

A configurable interval after which the Decision Point Concludes that communication from a given PCN-egress-node has failed if it has received no reports from the PCN-egress-node during that interval.

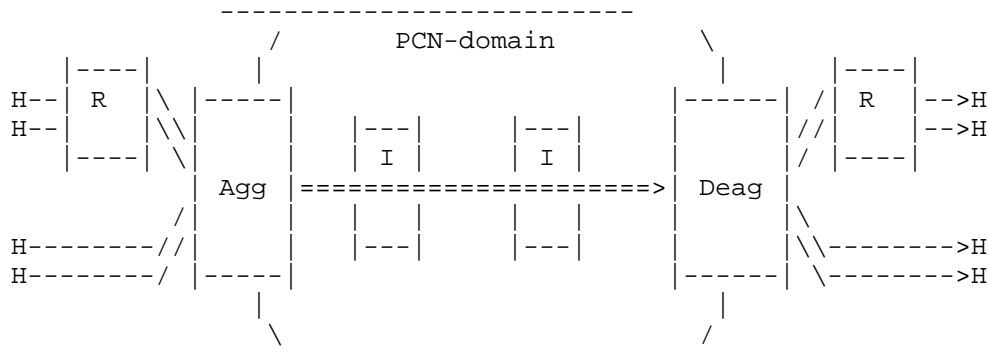
t-recvFail

An ingress-egress-aggregate timer that is used at The Decision point (in this document at the PCN-ingress-node) which when expires raises an alarm to management, and activates the PCN-ingress-node to block the admission of new PCN-flows. This timer expires when its value is equal to T-fail and is reset when a report, i.e., RSVP aggregated RESV message, is received for the ingress-egress-aggregate.

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, an ingress-egress-aggregate matches to only one RSVP SESSION for generic aggregated reservations. However, a RSVP SESSION for generic aggregated reservations can match to one or more than one ingress-egress-aggregates. This can be accomplished by using for the different ingress-egress-aggregates the same combinations of ingress and egress identifiers, but with a different PHB-ID value (see [RFC4860]).



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]

In addition, in this document it is considered that the PCN-boundary nodes are able 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].

Furthermore, it is considered that 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].

Moreover, each Aggregator and Deaggregator (i.e., PCN-boundary-nodes) MUST support policies to initiate and maintain for each combination of the PCN-boundary-node and all other PCN-boundary-nodes of the same PCN-domain one RSVP SESSION for generic aggregated reservations. Note that RSVP SESSION for generic aggregated reservations can match to one or more than one ingress-egress-aggregates. This can be accomplished by using for the different ingress-egress-aggregates the same combinations of ingress and egress identifiers, but with a different PHB-ID value (see [RFC4860]). Depending on a policy the Aggregator SHOULD be able to decide whether an e2e RSVP session can be mapped into one ingress-egress-aggregate maintained by the Aggregator (i.e., PCN-ingress-node).

The RSVP SESSION object for generic aggregate reservations, maintains the mapping and association between the PCN ingress-egress-aggregate and the PCN-flows (e2e RSVP reservation session) that travel in one direction between the specific pair of PCN-boundary-nodes specified by the ingress-egress-aggregate. Note that in this document the PCN ingress-egress-aggregate is identified by using the RSVP SESSION object for generic aggregate reservation, see [RFC4860], by using the following:

- 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)

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

The method of PCN marking within the PCN domain is based on [RFC5670]. In addition, the method of encoding and transport of pre-congestion information is based [RFC5696]. The PHB-ID (Per Hop Behavior Identification Code) used, SHOULD be set equal to PCN-compatible Diffserv codepoint(s).

2.1.1.2. Traffic Classification Within The Aggregation Region

The PCN-traffic is marked using PCN-marking and is classified using The PCN-BA (i.e., combination of the DSCP and ECN fields). The PCN-traffic belonging to an PCN aggregated session can be classified only at the PCN-boundary-nodes using the combination of (1) PCN-BA (i.e., combination of the DSCP and ECN fields), (2) IP addresses of the specific pair of PCN-boundary-nodes used by a ingress-egress-aggregate. The method of classification and traffic conditioning of PCN-traffic and non-PCN traffic and PHB configuration is described in draft-ietf-pcn-cl-edge-behaviour-09] and [draft-ietf-pcn-sm-edge-behaviour-06].

2.1.1.3. Deaggregator (PCN-egress-node) Determination

In this document it is considered that for the determination of the Deaggregator, the same methods can be used as the ones described in [RFC4860].

2.1.1.4. Mapping E2E Reservations Onto Aggregate Reservations

In this document it is considered that for the mapping of e2e reservations onto aggregate reservations, the same methods can be used as the ones described in [RFC4860], augmented by the following rules:

- o) PCN-ingress-node MUST use one or more policies to estimate whether an e2e RSVP reservation session associated with an e2e Path message that arrives at the external interface of the PCN-ingress-node can be mapped onto an existing RSVP generic aggregation reservation state, i.e., PCN ingress-egress-aggregate.

2.1.5. Size of Aggregate Reservations

In this document it is considered that for the determination of the size of the aggregate reservations, the same methods can be used as the ones described in [RFC4860].

2.1.6. E2E Path ADSPEC update

In this document it is considered that for the update of the e2e Path ADSPEC, the same methods can be used as the ones described in [RFC4860].

2.1.7. 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.

2.1.8. Inter-domain Routes

In this document it is considered that for the solving the issues caused by the inter-domain route changes, the same methods can be used as the ones described in [RFC4860].

2.1.9. Reservations for Multicast Sessions

PCN does not consider reservations for multicast sessions.

2.1.10. Multi-level Aggregation

PCN does not consider multi-level aggregations within the PCN domain.

2.1.11. Reliability Issues

In this document it is considered that for solving possible reliability issues, the same methods can be used as the ones described in [RFC4860].

2.1.12. Message Integrity and Node Authentication

In this document it is considered that for message integrity and node authentication, the same methods can be used as the ones described in [RFC4860] and [RFC5559].

3. Elements of Procedure

This section describes the procedures used to implement the aggregated RSVP procedure over PCN.

3.1. Receipt of E2E Path Message By PCN-ingress-node (aggregating router)

When the e2e RSVP message arrives at the exterior interface of the aggregator, i.e., PCN-ingress-node, then standard RSVP generic aggregation [RFC4860] procedures are used, augmented with the following rules:

- o) The e2e RSVP reservation session associated with an e2e Path message that arrives at the external interface of the PCN-ingress-node is mapped onto an existing RSVP generic aggregation reservation state (i.e., PCN ingress-egress-aggregate).
- o) If the timer t-recvFail expires for a given PCN-egress-node, the Decision Point (i.e., PCN-ingress-node) SHOULD NOT allow the e2e RSVP flow to be admitted to that ingress-egress-aggregate. This procedure is defined in detail in: [draft-ietf-pcn-cl-edge-behaviour-09] and [draft-ietf-pcn-sm-edge-behaviour-06].

Depending on a local policy the Aggregator SHOULD decide whether this situation is considered of being an error, or whether the e2e reservation session SHOULD be mapped to another ingress-egress-aggregate maintained by the same RSVP SESSION for aggregated reservations.

If the Aggregator is not able to map the requesting e2e RSVP session into another ingress-egress-aggregate, then the Aggregator SHOULD NOT admit the e2e RSVP session and it SHOULD generate an e2e PathErr message using standard e2e RSVP procedures [RFC2205]. This e2e PathErr message is sent to the originating sender of the e2e Path message.

- o) If the timer t-recvFail does NOT expire for a given PCN-egress-node, then:
 - *) If the PCN-admission state for the ingress-egress-aggregate associated with the received e2e Path is "admit", the Decision Point (i.e., PCN-ingress-node) SHOULD allow new flows to be admitted to that aggregate. The e2e Path message is then forwarded towards destination.

- *) If the PCN-admission-state for the same PCN aggregation state is "block", the Aggregator using the same policy as mentioned above SHOULD either map the incoming e2e RSVP session to another ingress-egress-aggregate associated with the same generic aggregated RSVP session, or the flow SHOULD NOT be admitted and an e2e PathErr message SHOULD be generated, using standard e2e RSVP procedures [RFC2205], [RFC4495].

This e2e PathErr message is sent to the originating sender of the e2e Path message, using standard e2e RSVP procedures [RFC2205], [RFC4495]. A new error code "PCN-domain rejects e2e reservation" MUST be augmented to the RSVP error codes to inform the sender that a PCN domains rejects the e2e reservation request.

The way of how the PCN-admission-state is maintained is specified in [draft-ietf-pcn-cl-edge-behaviour-09] and [draft-ietf-pcn-sm-edge-behaviour-06].

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. The e2e Path messages are simply forwarded as normal IP datagrams.

3.3. Receipt of E2E Path Message By PCN-egress-node (deaggregating router)

When receiving the e2e Path message the PCN-egress-node (deaggregating router) performs main regular [RFC4860] procedures, augmented with the following rules, see also [draft-lefaucheur-rsvp-ecn-01]:

- o) 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.

The PCN-egress-nodes forwards the e2e Path message towards the receiver.

3.4. Initiation of new Aggregate Path Message By PCN-ingress node (Aggregating Router)

In this document it is considered that for the initiation of the new RSVP aggregated Path message by the PCN-ingress-node (Aggregation Router), the same methods can be used as the ones described in [RFC4860].

3.5. Handling Of new Aggregate Path Message By Interior Routers

The Aggregate 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. The Aggregated Path messages are simply forwarded as normal IP datagrams.

3.6. Handling of E2E Resv Message by Deaggregating Router

When the e2e Resv message arrives at the exterior interface of the Deaggregating router, i.e., PCN-egress-node, then standard RSVP aggregation [RFC4860] procedures are used.

3.7. Handling Of E2E Resv Message By Interior Routers

The e2e Resv messages traverse zero or more PCN-interior-nodes. The PCN-interior-nodes receive the e2e Resv message on an interior interface and forward it on another interior interface. The e2e Resv messages are simply forwarded as normal IP datagrams.

3.8. Initiation of New Aggregate Resv Message By Deaggregating Router

In this document it is considered that for the initiation of the new RSVP aggregated Resv message by the PCN-ingress-node (Aggregation Router), the same methods can be used as the ones described in [RFC4860] augmented with the following rules:

- o) At the end of each t-meas measurement interval, or less frequently if "optional report suppression" is activated, see [draft-ietf-pcn-cl-edge-behaviour-09], and [draft-ietf-pcn-sm-edge-behaviour-06], the PCN-egress-node MUST include the new PCN object that will be sent to the associated Decision Point (i.e., PCN-ingress-node). The PCN object is specified in this document and is used to report of the data measured by the PCN-egress-node, for a particular ingress-egress-aggregate, see [draft-ietf-pcn-cl-edge-behaviour-09], and [draft-ietf-pcn-sm-edge-behaviour-06]. The address of the PCN-ingress-node is the one specified in the same ingress-egress-aggregate.

3.9. Handling of Aggregate Resv Message by Interior Routers

The Aggregated Resv messages traverse zero or more PCN-interior-nodes. The PCN-interior-nodes receive the Aggregated Resv message on an interior interface and forward it on another interior interface. The Aggregated Resv messages are simply forwarded as normal IP datagrams.

3.10. Handling of E2E Resv Message by Aggregating Router

When the e2e Resv message arrives at the interior interface of the Aggregating router, i.e., PCN-ingress-node, then standard RSVP aggregation [RFC4860] procedures are used.

3.11. Handling of Aggregated Resv Message by Aggregating Router

When the Aggregated Resv message arrives at the interior interface of the Aggregating router, i.e., PCN-ingress-node, then standard RSVP aggregation [RFC4860] procedures are used, augmented with the following rules:

- o) the Decision Point (i.e., the PCN-ingress-node) SHOULD use the information carried by the PCN object as specified in [draft-ietf-pcn-cl-edge-behaviour-09], [draft-ietf-pcn-sm-edge-behaviour-06].

When the Aggregator (i.e., PCN-ingress-node) needs to terminate an amount of traffic associated to one ingress-egress-aggregate (see bullet 2 in Section 3.3.2 of [draft-ietf-pcn-cl-edge-behaviour-09] and [draft-ietf-pcn-sm-edge-behaviour-06]), then the following procedure is followed. Based on a local policy, the Aggregator SHOULD select one of the following options:

- o) for the same ingress-egress-aggregate, select a number of e2e RSVP sessions 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 RSVP flow can be followed as specified in [RFC4495].
- o) for the same ingress-egress-aggregate, select a number of e2e RSVP sessions 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, see above. In this situation the same mechanisms for terminating an e2e RSVP flow or reducing bandwidth associated with an e2e RSVP flow can be followed as specified in [RFC4495].

3.12. Removal of E2E Reservation

In this document it is considered that for the removal of e2e reservations, the same methods can be used as the ones described in [RFC4860] and [RFC4495].

3.13. Removal of Aggregate Reservation

In this document it is considered that for the removal of aggregated reservations, the same methods can be used as the ones described in [RFC4860].

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

The handling of data on the reserved e2e Flow by Aggregating Router is using the procedures described in [RFC4860] augmented with:

- o) Regarding, PCN marking and traffic classification the procedures defined in Section 2.1.1 and 2.1.3 of this document are used.

3.15. Procedures for Multicast Sessions

In this document no multicast sessions are considered.

4. Protocol Elements

The protocol elements in this document are using the protocol Elements defined in [RFC4860], augmented with the following rules:

- o) A PCN-egress-node (i.e., deaggregator) SHOULD send periodically and at the end of each t-meas measurement interval, or less frequently if "optional report suppression" is activated, an (refresh) aggregated RSVP message to the PCN-ingress-node (i.e. aggregator).
- o) the DSCP value included in the SESSION object, SHOULD be set equal to a PCN-compatible Diffserv codepoint.
- o) An aggregated Resv message MUST carry a PCN object to report the data measured by an PCN-egress-node (i.e., Deaggregator).

4.1 PCN object

The PCN object reports data measured by an PCN-egress-node.

PCN objects are defined for different PCN edge behavior drafts. This document defines several types of PCN objects.

- o) Single Marking (SM) PCN object, when IPv4 addresses are used:
Class = PCN
C-Type = RSVP-AGGREGATE-IPv4-PCN-SM

```

+-----+-----+-----+-----+
| IPv4 PCN-ingress-node Address (4 bytes) |
+-----+-----+-----+-----+
| IPv4 PCN-egress-node Address (4 bytes) |
+-----+-----+-----+-----+
| Congestion-Level-Estimate |
+-----+-----+-----+-----+
| rate of not marked PCN-traffic (NM-rate) |
+-----+-----+-----+-----+
| rate of PCN-marked PCN-traffic (PM-rate) |
+-----+-----+-----+-----+

```

- o) Single Marking (SM) PCN object, when IPv6 addresses are used:
 Class = PCN
 C-Type = RSVP-AGGREGATE-IPv6-PCN-SM

```

+-----+-----+-----+-----+
|                                     |
+                                     +
| IPv6 PCN-ingress-node Address (16 bytes) |
+                                     +
|                                     |
+-----+-----+-----+-----+
|                                     |
+                                     +
| IPv6 PCN-egress-node Address (16 bytes) |
+                                     +
|                                     |
+-----+-----+-----+-----+
| Congestion-Level-Estimate |
+-----+-----+-----+-----+
| rate of not marked PCN-traffic (NM-rate) |
+-----+-----+-----+-----+
| rate of PCN-marked PCN-traffic (PM-rate) |
+-----+-----+-----+-----+

```

- o) Controlled (CL) PCN object, IPv4 addresses are used:
 Class = PCN
 C-Type = RSVP-AGGREGATE-IPv4-PCN-CL

```

+-----+-----+-----+-----+
| IPv4 PCN-ingress-node Address (4 bytes) |
+-----+-----+-----+-----+
| IPv4 PCN-egress-node Address (4 bytes) |
+-----+-----+-----+-----+
| Congestion-Level-Estimate |
+-----+-----+-----+-----+
| rate of not marked PCN-traffic (NM-rate) |
+-----+-----+-----+-----+
| rate of threshold-marked PCN-traffic (ThM-rate) |
+-----+-----+-----+-----+
| rate of excess-traffic-marked PCN-traffic (ETM-rate) |
+-----+-----+-----+-----+

```

- o) Controlled (CL) PCN object, IPv6 addresses are used:

Class = PCN

C-Type = RSVP-AGGREGATE-IPv6-PCN-CL

```

+-----+-----+-----+-----+
|
+
|
+   IPv6 PCN-ingress-node Address (16 bytes)
+
|
+-----+-----+-----+-----+
|
+
|
+   IPv6 PCN-egress-node Address (16 bytes)
+
|
+-----+-----+-----+-----+
| Congestion-Level-Estimate |
+-----+-----+-----+-----+
| rate of not marked PCN-traffic (NM-rate) |
+-----+-----+-----+-----+
| rate of threshold-marked PCN-traffic (ThM-rate) |
+-----+-----+-----+-----+
| rate of excess-traffic-marked PCN-traffic (ETM-rate) |
+-----+-----+-----+-----+

```

The fields carried by the PCN object are specified in [draft-ietf-pcn-signaling-requirements-06.txt], [draft-ietf-pcn-cl-edge-behaviour-09] and [draft-ietf-pcn-sm-edge-behaviour-06]:

- o the IPv4 or IPv6 address of the PCN-ingress-node and the the IPv4 or IPv6 address of the PCN-egress-node; together they specify the ingress-egress-aggregate to which the report refers;

- o rate of not-marked PCN-traffic (NM-rate) in octets/second; its format is a 32-bit IEEE floating point number;
- o rate of PCN-marked traffic (PM-rate) in octets/second; its format is a 32-bit IEEE floating point number;
- o congestion-level-estimate, which is a number between zero and one; its format is a 32-bit IEEE floating point number;
- o rate of threshold-marked PCN traffic (ThM-rate) in octets/second; its format is a 32-bit IEEE floating point number;
- o rate of excess-traffic-marked traffic (ETM-rate) in octets/second; its format is a 32-bit IEEE floating point number;

5. Security Considerations

The same security considerations specified in [RFC4860] and [RFC5559] apply also to this document.

6. IANA Considerations

This document makes the following requests to the IANA:

- o allocate a new Object Class (PCN Object), see Section 4.1.
- o allocate a "PCN-domain rejects e2e reservation" Error Code that may appear only in e2e PathErr messages, see Section 3.1.

Error Value for "PCN-domain rejects e2e reservation" = To be allocated by IANA

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 Tom Taylor, Francois Le Faucheur and James Polk for the comments provided on the 00 version of this draft.

8. Normative References

[draft-ietf-pcn-cl-edge-behaviour-09] T. Taylor, A. Charny, F. Huang, G. Karagiannis, M. Menth, "PCN Boundary Node Behaviour for the Controlled Load (CL) Mode of Operation (Work in progress)", June 2011.

[draft-ietf-pcn-sm-edge-behaviour-06] A. Charny, J. Zhang, G. Karagiannis, M. Menth, T. Taylor, "PCN Boundary Node Behaviour for the Single Marking (SM) Mode of Operation (Work in progress)", June 2011.

[draft-ietf-pcn-signaling-requirements-06] G. Karagiannis, T. Taylor, K. Chan, M. Menth, P. Eardley, " Requirements for Signaling of (Pre-) Congestion Information in a DiffServ Domain(Work in progress)", July 2011.

[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.

[RFC5696] Moncaster, T., Briscoe, B., and M. Menth, "Baseline Encoding and Transport of Pre-Congestion Information", RFC 5696, November 2009.

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.

[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.

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

10. Authors' Address

Georgios Karagiannis
University of Twente
P.O. Box 217
7500 AE Enschede,
The Netherlands
EMail: g.karagiannis@utwente.nl

Anurag Bhargava
Cisco Systems, Inc.
USA
Email: anuragb@cisco.com