

Congestion and Pre-Congestion
Notification
Internet-Draft
Obsoletes: 5696 (if approved)
Intended status: Standards Track
Expires: January 11, 2012

B. Briscoe
BT
T. Moncaster
Moncaster Internet Consulting
M. Menth
University of Tuebingen
July 10, 2011

Encoding 3 PCN-States in the IP header using a single DSCP
draft-ietf-pcn-3-in-1-encoding-06

Abstract

The objective of Pre-Congestion Notification (PCN) is to protect the quality of service (QoS) of inelastic flows within a Diffserv domain. The overall rate of the PCN-traffic is metered on every link in the PCN domain, and PCN-packets are appropriately marked when certain configured rates are exceeded. Egress nodes pass information about these PCN-marks to decision points which then decide whether to admit or block new flow requests or 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 provides for up to 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. This document obsoletes RFC5696.

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 11, 2012.

Copyright Notice

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

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

Table of Contents

1.	Introduction	4
1.1.	Requirements Language	5
1.2.	Changes in This Version (to be removed by RFC Editor)	5
2.	Definitions and Abbreviations	7
2.1.	Terminology	7
2.2.	List of Abbreviations	7
3.	Definition of 3-in-1 PCN Encoding	8
4.	Requirements for and Applicability of 3-in-1 PCN Encoding	9
4.1.	PCN Requirements	9
4.2.	Requirements Imposed by Tunnelling	9
4.3.	Applicability of 3-in-1 PCN Encoding	10
5.	Behaviour of a PCN-node to Comply with the 3-in-1 PCN Encoding	10
5.1.	PCN-ingress Node Behaviour	10
5.2.	PCN-interior Node Behaviour	11
5.2.1.	Behaviour Common to all PCN-interior Nodes	11
5.2.2.	Behaviour of PCN-interior Nodes Using Two PCN-markings	11
5.2.3.	Behaviour of PCN-interior Nodes Using One PCN-marking	11
5.3.	Behaviour of PCN-egress Nodes	12
6.	Backward Compatibility	13
6.1.	Backward Compatibility with ECN	13
6.2.	Backward Compatibility with the Baseline Encoding	13
7.	IANA Considerations	13
8.	Security Considerations	14
9.	Conclusions	14
10.	Acknowledgements	14
11.	Comments Solicited	15
12.	References	15
12.1.	Normative References	15
12.2.	Informative References	15
Appendix A.	Choice of Suitable DSCPs	16
Appendix B.	Co-existence of ECN and PCN	18
Appendix C.	Example Mapping between Encoding of PCN-Marks in IP and in MPLS Shim Headers	20
Appendix D.	Rationale for different behaviours for single marking schemes	21
Authors' Addresses	21

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 generally configured with different 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 pass information about these PCN-marks to decision points which then decide whether to admit new flows or terminate existing flows [I-D.ietf-pcn-cl-edge-behaviour], [I-D.ietf-pcn-sm-edge-behaviour].

The baseline encoding defined in [RFC5696] described how two PCN marking states (Not-marked and PCN-Marked) could be encoded into the IP header using a single Diffserv codepoint. It also provided an experimental codepoint (EXP), along with guidelines for the use of that codepoint. Two PCN marking states are sufficient for the Single Marking edge behaviour [I-D.ietf-pcn-sm-edge-behaviour]. However, PCN-domains utilising the controlled load edge behaviour [I-D.ietf-pcn-cl-edge-behaviour] require three PCN marking states. This document extends the baseline encoding by redefining the EXP codepoint to provide a third PCN marking state in the IP header, still using a single Diffserv codepoint. This encoding scheme is therefore called the "3-in-1 PCN encoding". It obsoletes the baseline encoding [RFC5696], which provides only a sub-set of the same capabilities.

The full version of this encoding requires any tunnel endpoint within the PCN-domain to support the normal tunnelling rules defined in [RFC6040]. There is one limited exception to this constraint where the PCN-domain only uses the excess-traffic-marking behaviour and where the threshold-marking behaviour is deactivated. This is discussed in Section 5.2.3.1.

This document only concerns the PCN wire protocol encoding for IP headers, whether IPv4 or IPv6. It makes no changes or recommendations concerning algorithms for congestion marking or congestion response. Other documents will define the PCN wire protocol for other header types. Appendix C discusses a possible mapping between IP and MPLS.

1.1. 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].

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

From draft-ietf-pcn-3-in-1-encoding-05 to -06:

- * Draft re-written to obsolete baseline encoding [RFC5696].
- * New section defining utilising this encoding for single marking. Added an appendix explaining an apparent inconsistency relating to single marking.
- * Moved (and updated) informative appendixes from [RFC5696] to this document. Original Appendix C was omitted as it is now redundant.
- * Significant re-structuring of document.

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

- * Draft moved to standards track as per working group discussions.
- * Added Appendix B 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. Definitions and Abbreviations

2.1. Terminology

The terms PCN-domain, PCN-node, PCN-interior-node, PCN-ingress-node, PCN-egress-node, PCN-boundary-node, PCN-traffic, PCN-packets and PCN-marking are used as defined in [RFC5559]. The following additional terms are defined in this document:

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

PCN-compatible Diffserv codepoint: a Diffserv codepoint indicating packets for which the ECN field is used to carry PCN-markings rather than [RFC3168] markings (see Appendix A).

Threshold-marked codepoint: a codepoint that indicates packets that have been marked at a PCN-interior-node as a result of an indication from the threshold-metering function [RFC5670]. Abbreviated to ThM.

Excess-traffic-marked codepoint: a codepoint that indicates packets that have been marked at a PCN-interior-node as a result of an indication from the excess-traffic-metering function [RFC5670]. Abbreviated to ETM.

Not-marked codepoint: a codepoint that indicates PCN-packets but that are not PCN-marked. Abbreviated to NM.

not-PCN codepoint: a codepoint that indicates packets that are not PCN-packets.

2.2. List of Abbreviations

The following abbreviations are used in this document:

- o AF = Assured Forwarding [RFC2597]
- o CE = Congestion Experienced [RFC3168]
- o CS = Class Selector [RFC2474]
- o DSCP = Diffserv codepoint
- o ECN = Explicit Congestion Notification [RFC3168]
- o ECT = ECN Capable Transport [RFC3168]

- o EF = Expedited Forwarding [RFC3246]
- o ETM = Excess-traffic-marked
- o EXP = Experimental
- o IP = Internet protocol
- o NM = Not-marked
- o PCN = Pre-Congestion Notification
- o ThM = Threshold-marked

3. Definition of 3-in-1 PCN Encoding

The 3-in-1 PCN encoding scheme allows for two or three PCN-marking states to be encoded within the IP header. The full encoding is shown in Figure 1.

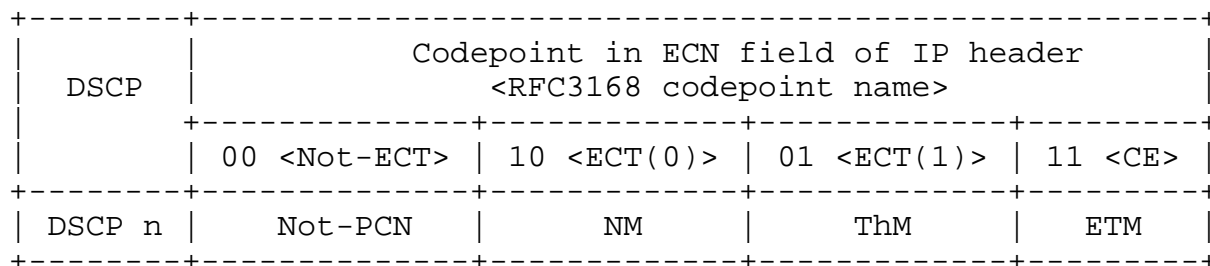


Figure 1: 3-in-1 PCN Encoding

As mentioned above, the 3-in-1 PCN encoding is an extension of the baseline encoding [RFC5696]. Like the baseline encoding it uses a combination of a PCN-compatible DSCP (DSCP n in Figure 1) and the ECN field for the encoding of PCN-marks. Appendix A discusses the choice of suitable DSCPs. 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].

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

4.1. 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 behaviours: threshold-marking and excess-traffic-marking [RFC5670]. Some edge behaviors require only a single marking behaviour [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 behaviour 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 do not use PCN (the not-PCN codepoint).

In all current PCN edge behaviors that use two marking behaviours [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.

4.2. Requirements Imposed by Tunnelling

[RFC6040] defines rules for the encapsulation and decapsulation of ECN markings within IP-in-IP tunnels. The publication of RFC6040 removed the tunnelling constraints that existed when the baseline encoding [RFC5696] was written (see section 3.3.2 of [I-D.ietf-pcn-encoding-comparison]).

Nonetheless, there is still a problem if there are any legacy (pre-RFC6040) decapsulating tunnel endpoints within a PCN domain. If a PCN node Threshold-marks the outer header of a tunnelled packet with a Not-marked codepoint on the inner header, the legacy decapsulator will revert the Threshold-marking to Not-marked. The rules on

applicability in Section 4.3 below are designed to avoid this problem.

4.3. Applicability of 3-in-1 PCN Encoding

The 3-in-1 encoding is applicable in situations where two marking behaviours are being used in the PCN-domain. The 3-in-1 encoding can also be used with only one marking behaviour, in which case one of the codepoints MUST NOT be used throughout the PCN-domain (see Section 5.2.3).

For the full 3-in-1 encoding to apply, any tunnel endpoints (IP-in-IP and IPsec) within the PCN-domain MUST comply with the ECN encapsulation and decapsulation rules set out in [RFC6040] (see Section 4.2). There is one exception to this rule outlined next.

It may not be possible to upgrade every pre-RFC6040 tunnel endpoint within a PCN-domain. In such circumstances a limited version of the 3-in-1 encoding can still be used but only under the following stringent condition. If any pre-RFC6040 tunnel endpoint exists within a PCN-domain then every PCN-node in the PCN-domain MUST be configured so that it never sets the ThM codepoint. The behaviour of PCN-interior nodes in this case is defined in Section 5.2.3.1. In all other situations where legacy tunnel endpoints might be present within the PCN domain, the 3-in-1 encoding is not applicable.

5. Behaviour of a PCN-node to Comply with the 3-in-1 PCN Encoding

As mentioned in Section 4.3 above, all PCN-nodes MUST comply with [RFC6040].

5.1. PCN-ingress Node Behaviour

PCN-traffic MUST be marked with a PCN-compatible Diffserv codepoint. To conserve DSCPs, Diffserv codepoints SHOULD be chosen that are already defined for use with admission-controlled traffic. Appendix A gives guidance to implementors on suitable DSCPs. Guidelines for mixing traffic types within a PCN-domain are given in [RFC5670].

If a packet arrives at the PCN-ingress-node that shares a PCN-compatible DSCP and is not a PCN-packet, the PCN-ingress MUST mark it as not-PCN.

If a PCN-packet arrives at the PCN-ingress-node, the PCN-ingress MUST change the PCN codepoint to Not-marked.

If a PCN-packet arrives at the PCN-ingress-node with its ECN field already set to a value other than not-ECT, then appropriate action MUST be taken to meet the requirements of [RFC4774]. The simplest appropriate action is to just drop such packets. However, this is a drastic action that an operator may feel is undesirable. Appendix B provides more information and summarises other alternative actions that might be taken.

5.2. PCN-interior Node Behaviour

5.2.1. Behaviour Common to all PCN-interior Nodes

Interior nodes MUST NOT change not-PCN to any other codepoint.

Interior nodes MUST NOT change NM to not-PCN.

Interior nodes MUST NOT change ThM to NM or not-PCN.

Interior nodes MUST NOT change ETM to any other codepoint.

5.2.2. Behaviour of PCN-interior Nodes Using Two PCN-markings

If the threshold-meter function indicates a need to mark the packet, the PCN-interior node MUST change NM to ThM.

If the excess-traffic-meter function indicates a need to mark the packet:

- o the PCN-interior node MUST change NM to ETM;
- o the PCN-interior node MUST change ThM to ETM.

If both the threshold meter and the excess-traffic meter indicate the need to mark a packet, the excess traffic marking rules MUST take priority.

5.2.3. Behaviour of PCN-interior Nodes Using One PCN-marking

Some PCN edge behaviours require only one PCN-marking within the PCN-domain. The Single Marking edge behaviour [I-D.ietf-pcn-sm-edge-behaviour] requires PCN-interior nodes to mark packets using the excess-traffic-meter function [RFC5670]. It is possible that future schemes may require only the threshold-meter function. Observant readers may spot an apparent inconsistency between the two following cases. Appendix D explains the rationale behind this inconsistency.

5.2.3.1. Marking using only the Excess-traffic-meter Function

The threshold-traffic-meter function SHOULD be disabled and MUST NOT trigger any packet marking.

The PCN-interior node SHOULD raise a management alarm if it receives a ThM packet, but the frequency of such alarms SHOULD be limited.

If the excess-traffic-meter function indicates a need to mark the packet:

- o the PCN-interior node MUST change NM to ETM;
- o the PCN-interior node MUST change ThM to ETM. It SHOULD also raise an alarm as above.

5.2.3.2. Marking using only the Threshold-meter Function

The excess-traffic-meter function SHOULD be disabled and MUST NOT trigger any packet marking.

The PCN-interior node SHOULD raise a management alarm if it receives an ETM packet, but the frequency of such alarms SHOULD be limited.

If the threshold-meter function indicates a need to mark the packet:

- o the PCN-interior node MUST change NM to ThM;
- o the PCN-interior node MUST NOT change ETM to any other codepoint. It SHOULD raise an alarm as above.

5.3. Behaviour of PCN-egress Nodes

A PCN-egress-node SHOULD set the not-PCN (00) codepoint on all packets it forwards out of the PCN-domain.

The only exception to this is if the PCN-egress-node is certain that revealing other codepoints outside the PCN-domain won't contravene the guidance given in [RFC4774]. For instance, if the PCN-ingress-node has explicitly informed the PCN-egress-node that this flow is ECN-capable, then it might be safe to expose other codepoints. Appendix B gives details of how such schemes might work, but such schemes are currently experimental.

If the PCN-domain is configured to use only excess-traffic marking, the PCN-egress node MUST treat ThM as ETM and if only threshold-marking is used it should treat ETM as ThM. However it SHOULD raise a management alarm in either instance since this means there is some

misconfiguration in the PCN-domain.

6. Backward Compatibility

6.1. Backward Compatibility with ECN

BCP 124 [RFC4774] gives guidelines for specifying alternative semantics for the ECN field. It sets out a number of factors to be taken into consideration. It also suggests various techniques to allow the co-existence of default ECN and alternative ECN semantics. The encoding specified in this document uses one of those techniques; it defines PCN-compatible Diffserv codepoints as no longer supporting the default ECN semantics. As such, this document is compatible with BCP 124.

On its own, the 3-in-1 encoding cannot support both ECN marking end-to-end (e2e) and PCN-marking within a PCN-domain. Appendix B discusses possible ways to do this, e.g. by carrying e2e ECN across a PCN-domain within the inner header of an IP-in-IP tunnel. Although Appendix B recommends various approaches over others, it is merely informative and all such schemes are beyond the normative scope of this document.

In any PCN deployment, traffic can only enter the PCN-domain through PCN-ingress-nodes and leave through PCN-egress-nodes. PCN-ingress-nodes ensure that any packets entering the PCN-domain have the ECN field in their outermost IP header set to the appropriate PCN codepoint. PCN-egress-nodes then guarantee that the ECN field of any packet leaving the PCN-domain has appropriate ECN semantics. This prevents unintended leakage of ECN marks into or out of the PCN-domain, and thus reduces backward-compatibility issues.

6.2. Backward Compatibility with the Baseline Encoding

A PCN node implemented to use the obsoleted baseline encoding could conceivably have been configured so that the Threshold-meter function marked what is now defined as the ETM codepoint in the 3-in-1 encoding. However, there is no known deployment of such an implementation and no reason to believe that such an implementation would ever have been built. Therefore, it seems safe to ignore this issue.

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

PCN-marking only carries a meaning within the confines of a PCN-domain. This encoding document is intended to stand independently of the architecture used to determine how specific packets are authorised to be PCN-marked, which will be described in separate documents on PCN-boundary-node behaviour.

This document assumes the PCN-domain to be entirely under the control of a single operator, or a set of operators who trust each other. However, future extensions to PCN might include inter-domain versions where trust cannot be assumed between domains. If such schemes are proposed, they must ensure that they can operate securely despite the lack of trust. However, such considerations are beyond the scope of this document.

One potential security concern is the injection of spurious PCN-marks into the PCN-domain. However, these can only enter the domain if a PCN-ingress-node is misconfigured. The precise impact of any such misconfiguration will depend on which of the proposed PCN-boundary-node behaviours is used, but in general spurious marks will lead to admitting fewer flows into the domain or potentially terminating too many flows. In either case, good management should be able to quickly spot the problem since the overall utilisation of the domain will rapidly fall.

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. In general, the use of this PCN encoding scheme presupposes that any tunnel endpoints within the PCN-domain comply with [RFC6040].

10. Acknowledgements

Many thanks to Phil Eardley for providing extensive feedback, criticism and advice. Thanks also to Teco Boot, Kwok Ho Chan, Ruediger Geib, Georgios Karaginannis and everyone else who has commented on the 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.
- [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.
- [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-09 (work in progress), June 2011.
- [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-06 (work in progress), June 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-06
(work in progress), June 2011.
- [RFC2597] Heinanen, J., Baker, F., Weiss, W., and J. Wroclawski,
"Assured Forwarding PHB Group", RFC 2597, June 1999.
- [RFC3246] Davie, B., Charny, A., Bennet, J., Benson, K., Le Boudec,
J., Courtney, W., Davari, S., Firoiu, V., and D.
Stiliadis, "An Expedited Forwarding PHB (Per-Hop
Behavior)", RFC 3246, March 2002.
- [RFC3540] Spring, N., Wetherall, D., and D. Ely, "Robust Explicit
Congestion Notification (ECN) Signaling with Nonces",
RFC 3540, June 2003.
- [RFC4594] Babiarz, J., Chan, K., and F. Baker, "Configuration
Guidelines for DiffServ Service Classes", RFC 4594,
August 2006.
- [RFC4774] Floyd, S., "Specifying Alternate Semantics for the
Explicit Congestion Notification (ECN) Field", BCP 124,
RFC 4774, November 2006.
- [RFC5127] Chan, K., Babiarz, J., and F. Baker, "Aggregation of
DiffServ Service Classes", RFC 5127, February 2008.
- [RFC5129] Davie, B., Briscoe, B., and J. Tay, "Explicit Congestion
Marking in MPLS", RFC 5129, January 2008.
- [RFC5462] Andersson, L. and R. Asati, "Multiprotocol Label Switching
(MPLS) Label Stack Entry: "EXP" Field Renamed to "Traffic
Class" Field", RFC 5462, February 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.

Appendix A. Choice of Suitable DSCPs

This appendix is informative, not normative.

The PCN working group chose not to define a single DSCP for use with PCN for several reasons. Firstly, the PCN mechanism is applicable to a variety of different traffic classes. Secondly, Standards Track DSCPs are in increasingly short supply. Thirdly, PCN is not a scheduling behaviour -- rather, it should be seen as being a marking behaviour similar to ECN but intended for inelastic traffic. The choice of which DSCP is most suitable for a given PCN-domain is dependent on the nature of the traffic entering that domain and the link rates of all the links making up that domain. In PCN-domains with sufficient aggregation, the appropriate DSCPs would currently be those for the Real-Time Treatment Aggregate [RFC5127]. 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)

CS5 is excluded from this list since PCN is not expected to be applied to signalling traffic. PCN can also be applied to the VOICE-ADMIT codepoint defined in [RFC5865].

PCN-marking is intended to provide a scalable admission-control mechanism for traffic with a high degree of statistical multiplexing. PCN-marking would therefore be appropriate to apply to traffic in the above classes, but only within a PCN-domain containing sufficiently aggregated traffic. In such cases, the above service classes may well all be subject to a single forwarding treatment (treatment aggregate [RFC5127]). However, this does not imply all such IP traffic would necessarily be identified by one DSCP -- each service class might keep a distinct DSCP within the highly aggregated region [RFC5127].

Additional service classes may be defined for which admission control is appropriate, whether through some future standards action or through local use by certain operators, e.g., the Multimedia Streaming service class (AF3). This document does not preclude the use of PCN in more cases than those listed above.

Note: The above discussion is informative not normative, as operators are ultimately free to decide whether to use admission control for certain service classes and whether to use PCN as their mechanism of choice.

Appendix B. Co-existence of ECN and PCN

This appendix is informative, not normative.

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] (obsoleted) included advice on handling ECN traffic within a PCN-domain. This appendix reiterates and 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 a 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, e.g. 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 (and therefore outside the scope of PCN). 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.

Unless specified otherwise, for any of the cases in the list below, an IP-in-IP tunnel can be used to preserve ECN markings across the PCN domain. However the tunnelling action should be applied wholly outside the PCN-domain as illustrated in the following figure:

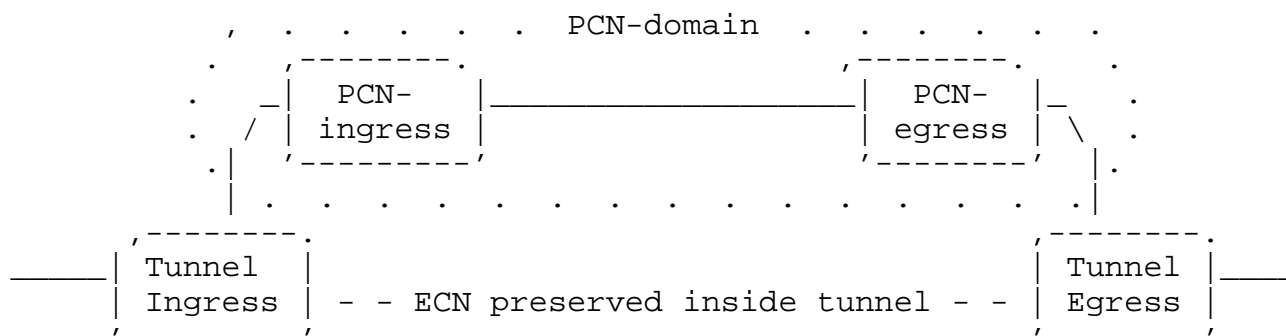


Figure 2: Separation of tunneling and PCN actions

There are 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 emphatically not recommended, unless perhaps as a last resort if tunnelling is not possible for some insurmountable reason.

ECN-capable Admission-controlled traffic where the e2e transport somehow indicates that it wants to see PCN marks: NOTE this is currently tentative 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 require further study and the following is for guidance only.

The PCN-ingress needs to tunnel the traffic, taking care to comply with [RFC6040]. In this case the PCN-egress should not zero the ECN field, and then the [RFC6040] tunnel egress will preserve any PCN-marking. Note that a PCN interior node may turn ECT(0) into ECT(1), which would not be compatible with the (currently experimental) ECN nonce [RFC3540].

Appendix C. Example Mapping between Encoding of PCN-Marks in IP and in MPLS Shim Headers

This appendix is informative not normative.

The 6 bits of the DS field in the IP header provide for 64 codepoints. When encapsulating IP traffic in MPLS, it is useful to make the DS field information accessible in the MPLS header. However, the MPLS shim header has only a 3-bit traffic class (TC) field [RFC5462] providing for 8 codepoints. The operator has the freedom to define a site-local mapping of a subset of the 64 codepoints of the DS field to the 8 codepoints in the TC field.

[RFC5129] describes how ECN markings in the IP header can also be mapped to codepoints in the MPLS TC field. Appendix A of [RFC5129] gives an informative description of how to support PCN in MPLS by extending the way MPLS supports ECN. But [RFC5129] was written while PCN specifications were in early draft stages. The following provides a clearer example of a mapping between PCN in IP and in MPLS using the PCN terminology and concepts that have since been specified.

To support PCN in a MPLS domain, codepoints for all used PCN-marks need to be provided in the TC field. That means, when e.g. only excess-traffic-marking is used for PCN purposes, the operator needs to define a site-local mapping to codepoints in the MPLS TC field for IP headers with:

- o DSCP n and ECT(0)
- o DSCP n and CE

If both excess-traffic-marking and threshold-marking are used, the operator needs to define a site-local mapping to codepoints in the MPLS TC field for IP headers with all three of the 3-in-1 codepoints:

- o DSCP n and ECT(0)
- o DSCP n and ECT(1)
- o DSCP n and CE

In either case, if the operator wishes to support the same Diffserv PHB but without PCN marking, it will also be necessary to define a site-local mapping to an MPLS TC codepoint for IP headers marked with:

- o DSCP n and Not-ECT

Appendix D. Rationale for different behaviours for single marking schemes

Readers may notice an apparent discrepancy between the two single marking behaviours in Section 5.2.3.1 and Section 5.2.3.2. For the excess-traffic only marking an unexpected ThM marked packet is remarked as ETM. For the threshold only marking, an unexpected ETM marked packet is simply ignored (apart from an optional management alarm).

There are two reasons for having these seemingly contradictory requirements. Firstly these behaviours conform with the expected behaviour where both metering functions are being used for marking-- ETM is always a more severe marking than ThM and so should never be re-marked. Secondly the threshold-metering behaviour in [RFC5670] uses the current marking state of the arriving packet to determine what action to take. Consequently, in the ETM only marking it would be potentially unsafe to allow ThM packets to propagate forward in the network as they may adversely affect the threshold-metering function.

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://bobbriscoe.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