Congestion and Pre-Congestion Notification	B. Briscoe
Internet-Draft	BT
Obsoletes: 5696 (if approved)	T. Moncaster
Intended status: Standards Track	Moncaster Internet Consulting
Expires: February 19, 2012	M. Menth
	University of Tuebingen
	August 18, 2011

Encoding 3 PCN-States in the IP header using a single DSCP
draft-ietf-pcn-3-in-1-encoding-08

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

Copyright Notice

Table of Contents

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 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 defined 01 as an experimental codepoint (EXP), along with guidelines for its use. 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 RFC5696 encoding by redefining the experimental codepoint as 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 [RFC5696] encoding, 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 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-07 to -08: Editorial corrections and clarifications.

From draft-ietf-pcn-3-in-1-encoding-06 to -07: Clarified that each operator not the IETF chooses which DSCP(s) are PCN-compatible, and made it unambiguous that only PCN-nodes recognise that PCN-compatible DSCPs enable the 3-in-1 encoding.

- Removed statements about the PCN working group, given RFCs are meant to survive beyond the life of a w-g.

- Corrected the final para of "Rationale for Different Behaviours in Schemes with Only One Marking"

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 only one PCN-Marking. Added an appendix explaining an apparent inconsistency within this section.

- 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 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 carries PCN-markings rather than [RFC3168] markings. Note that an operator configures PCN-nodes to recognise PCN-compatible DSCPs, whereas the same DSCP has no PCN-specific meaning to a node outside the PCN domain.

Threshold-marked codepoint: a codepoint that indicates a packet has been threshold-marked; that is a packet that has 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 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:

- AF = Assured Forwarding [RFC2597]
- CE = Congestion Experienced [RFC3168]
- CS = Class Selector [RFC2474]
- DSCP = Diffserv codepoint
- ECN = Explicit Congestion Notification [RFC3168]
- ECT = ECN Capable Transport [RFC3168]
- EF = Expedited Forwarding [RFC3246]
- ETM = Excess-traffic-marked
- EXP = Experimental
- IP = Internet protocol
- NM = Not-marked
- PCN = Pre-Congestion Notification
- 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.

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

A PCN-node (i.e. a node within a PCN-domain) will be configured to recognise certain DSCPs as PCN-compatible. Appendix Appendix A discusses the choice of suitable DSCPs. In Figure 1 'DSCP n' indicates such a PCN-compatible DSCP. Within the PCN-domain, any packet carrying a PCN-compatible DSCP and with the ECN-field anything other than 00 (Not-PCN) is a PCN-packet as defined in [RFC5559].
PCN-nodes MUST interpret the ECN field of a PCN-packet using the 3-in-1 PCN encoding, rather than [RFC3168]. This does not change the behaviour for any packet with a DSCP that is not PCN-compatible, or for any node outside a PCN-domain. In all such cases the 3-in-1 encoding is not applicable and so by default the node will interpret the ECN field using [RFC3168].
When using the 3-in-1 encoding, the codepoints of the ECN field have the following meanings:

Not-PCN: indicates a non-PCN-packet, i.e., a packet that uses a PCN-compatible DSCP but 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 use a PCN-compatible DSCP but do not use PCN-marking (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 encoding of [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 that has a Not-marked codepoint on the inner header, a legacy decapsulator will forward the packet as Not-marked, losing the threshold marking. 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).

With one exception (see next paragraph), 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).

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. PCN-interior nodes in this case MUST solely use the Excess Traffic marking function, as 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

Any tunnel endpoint implementation on a PCN-node MUST comply with [RFC6040]. Since PCN is a new capability, this is considered a reasonable requirement.

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 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-node MUST mark it as not-PCN.

If a PCN-packet arrives at the PCN-ingress-node, the PCN-ingress-node 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 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 a packet, the PCN-interior node MUST change NM to ThM.

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

* the PCN-interior node MUST change NM to ETM;

* 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 precedence.

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. Appendix Appendix D explains the rationale for the behaviours defined in this section.

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:

*the PCN-interior node MUST change NM to ETM;

*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:

*the PCN-interior node MUST change NM to ThM;

*the PCN-interior node MUST NOT change ETM to any other codepoint. It SHOULD raise an alarm as above if it encounters an ETM packet.

5.3. PCN-egress Node Behaviour

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 Appendix B gives details of how such schemes might work, but such schemes are currently only tentative ideas.
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 within a PCN domain. 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 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 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 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 RFC5696 Encoding

A PCN node implemented to use the obsoleted RFC5696 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.L. 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

[RFC2597]	Heinanen, J., Baker, F., Weiss, W. and J. Wroclawski, "Assured Forwarding PHB Group", RFC 2597, June 1999.
[RFC3246]	Davie, B., Charny, A., Bennet, J.C.R., Benson, K., Le Boudec, J.Y., 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.
[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", Internet-Draft draft-ietf-pcn-cl-edge-behaviour-10, October 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", Internet-Draft draft-ietf-pcn-sm-edge-behaviour-06, 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", Internet-Draft draft-ietf-pcn-encoding-comparison-06, June 2011.

Appendix A. Choice of Suitable DSCPs

This appendix is informative, not normative.

A single DSCP has not been defined 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]. It is suggested that admission control could be used for the following service classes (defined in [RFC4594] unless otherwise stated):

- *Telephony (EF)
- *Real-time interactive (CS4)
- *Broadcast Video (CS3)
- *Multimedia Conferencing (AF4)
- *the VOICE-ADMIT codepoint defined in [RFC5865].

CS5 is excluded from this list since PCN is not expected to be applied to signalling traffic.

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 re-marked to a PCN-compatible DSCP by the PCN-ingress-node. Two mechanisms can be used to identify such traffic:

a. Flow signalling, which 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); the PCN-ingress-node re-marks traffic matching that filterspec to a PCN-compatible DSCP.

b. Traffic arrives with a DSCP that implies it requires admission control such as VOICE-ADMIT [RFC5865] or Real-Time Interactive, Broadcast Video when used for video on demand, and multimedia conferencing [RFC4594][RFC5865] (see Appendix Appendix A).

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. The tunnelling action should be applied wholly outside the PCN-domain as illustrated in the following figure:

```
              ,  .  .  .  .  .  PCN-domain  .  .  .  .  .  .
             .   ,--------.                   ,--------.    .
            .   _|  PCN-   |___________________|  PCN-  |_   .
            .  / | ingress |                   | egress | \  .
             .|  '---------'                   '--------'  |.
              | .  .  .  .  .  .  .  .  .  .  .  .  .  .  .|
         ,--------.                                      ,--------.
    _____| Tunnel  |                                     | Tunnel |____
         | Ingress | - - ECN preserved inside tunnel - - | Egress |
         '---------'                                     '--------'
```

There are three cases for how e2e ECN traffic may wish to be treated while crossing a PCN domain:

a) Traffic that does not require admission control:
For example, traffic that does not match flow signaling being used for admission control. In this case the e2e ECN traffic is not treated as PCN-traffic. There are two possible scenarios:

* Arriving traffic does not carry a PCN-compatible DSCP: no action required.

* Arriving traffic carries a DSCP that clashes with a PCN-compatible DSCP. There are two options:

  1. 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.

  2. 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.

b) Traffic that requires admission-control:
In this case the e2e ECN traffic is treated as PCN-traffic across the PCN domain. There are two options.

* The PCN-ingress-node 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-node drops CE-marked packets and otherwise sets the ECN-field to NM and sets the DCSP to a PCN-compatible DSCP. The PCN-egress zeroes the ECN field of all PCN packets; note that this turns off e2e ECN.

The second option is emphatically not recommended, unless perhaps as a last resort if tunnelling is not possible for some insurmountable reason.

c) Traffic that requires admission control where the endpoints ask to see PCN marks:

Note that this scheme is currently only a tentative idea.

For real-time data generated by an adaptive codec, schemes have been suggested where PCN marks may be leaked out of the PCN-domain so that end hosts can drop to a lower data rate, thus deferring the need for admission control. Currently such schemes require further study and the following is for guidance only.

The PCN-ingress-node needs to tunnel the traffic as in Figure 2, taking care to comply with [RFC6040]. In this case the PCN-egress does not zero the ECN field contrary to the recommendation in Section 5.3, so that the [RFC6040] tunnel egress will preserve any PCN-marking. Note that a PCN interior node may change the ECN-field from 10 to 01 (NM to ThM), which would be interpreted by the e2e ECN as a change from ECT(0) into ECT(1). This 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 the 64 codepoints of the DS field onto 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. [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, a PCN-compatible DSCP ('DSCP n') needs codepoints to be provided in the TC field for all the PCN-marks used. That means, when for instance only excess-traffic-marking is used for PCN purposes, the operator needs to define a site-local mapping to two codepoints in the MPLS TC field for IP headers with:

- DSCP n and NM
- DSCP n and ETM

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:

- DSCP n and NM
- DSCP n and ThM
- DSCP n and ETM

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:

- DSCP n and Not-PCN

The above sets of codepoints are required for every PCN-compatible DSCP. Clearly, given so few TC codepoints are available, it may be necessary to compromise by merging together some capabilities.

Appendix D. Rationale for Difference Between the Schemes using One PCN-Marking

Readers may notice a difference between the two behaviours in Section 5.2.3.1 and Section 5.2.3.2. With only excess-traffic marking enabled, an unexpected ThM packet can be re-marked to ETM. However, with only Threshold-marking, an unexpected ETM packet cannot be re-marked to ThM. This apparent inconsistency is deliberate. The behaviour with only threshold marking keeps to the rule of Section 5.2.1 that ETM is more severe and must never be changed to ThM even though ETM is not a valid marking in this case. Otherwise implementations would have to allow operators to configure an exception to this rule, which would not be safe practice and would require different code in the data plane compared to the common behaviour.

Authors' Addresses

Bob Briscoe 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 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 Menth University of Tuebingen Sand 13 Tuebingen, 72076 Germany Phone: +49 7071 2970505 EMail: menth@informatik.uni-tuebingen.de