Congestion and Pre-Congestion	B. Briscoe	
Notification	BT	
Internet-Draft	T. Moncaster	
Intended status: Experimental	Moncaster Internet Consulting	
Expires: July 15, 2011	M. Menth	
	University of Tuebingen	
	January 11, 2011	

TOC

Encoding 3 PCN-States in the IP header using a single DSCP draft-ietf-pcn-3-in-1-encoding-04

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

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 July 15, 2011.

Table of Contents

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

---

1. Introduction

TOC

The objective of Pre-Congestion Notification (PCN) [RFC5559] (Eardley, P., “Pre-Congestion Notification (PCN) Architecture,” June 2009.) 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] (Eardley, P., “Metering and Marking Behaviour of PCN-Nodes,” November 2009.) 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] (Eardley, P., “Pre-Congestion Notification (PCN) Architecture,” June 2009.). 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] (Charny, A., Huang, F., Karagiannis, G., Menth, M., and T. Taylor, “PCN Boundary Node Behaviour for the Controlled Load (CL) Mode of Operation,” December 2010.), [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,” December 2010.). The baseline encoding defined in [RFC5696] (Moncaster, T., Briscoe, B., and M. Menth, “Baseline Encoding and Transport of Pre-Congestion Information,” November 2009.) 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] (Charny, A., Huang, F., Karagiannis, G., Menth, M., and T. Taylor, “PCN Boundary Node Behaviour for the Controlled Load (CL) Mode of Operation,” December 2010.), 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] (Davie, B., Briscoe, B., and J. Tay, "Explicit Congestion Marking in MPLS," January 2008.). Appendix A 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)

TOC

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] (Eardley, P., "Metering and Marking Behaviour of PCN-Nodes," November 2009.).

*Minor corrections.

---

2. Requirements Language

TOC

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] (Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels," March 1997.).

2.1. Terminology

TOC

General PCN-related terminology is defined in the PCN architecture [RFC5559] (Eardley, P., "Pre-Congestion Notification (PCN) Architecture," June 2009.), 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

TOC

3.1. PCN Requirements

TOC

In accordance with the PCN architecture [RFC5559] (Eardley, P., "Pre-Congestion Notification (PCN) Architecture," June 2009.), 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] (Eardley, P., "Metering and Marking Behaviour of PCN-Nodes," November 2009.). Some edge behaviors require only a single marking scheme [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," December 2010.), others require both [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," December 2010.). 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] (Eardley, P., "Metering and Marking Behaviour of PCN-Nodes," November 2009.). 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] (Eardley, P., "Pre-Congestion Notification (PCN) Architecture," June 2009.), [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," December 2010.), 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

TOC

The baseline encoding scheme [RFC5696] (Moncaster, T., Briscoe, B., and M. Menth, "Baseline Encoding and Transport of Pre-Congestion Information," November 2009.) 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 (Structure of the former type-of-service field in IP) 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] (Nichols, K., Blake, S., Baker, F., and D. Black, "Definition of the

Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers," December 1998.) and the 2-bit ECN field [RFC3168] (Ramakrishnan, K., Floyd, S., and D. Black, "The Addition of Explicit Congestion Notification (ECN) to IP," September 2001.).

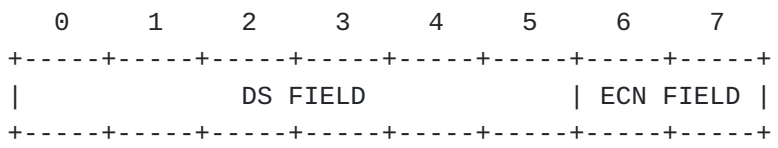

Figure 1: Structure of the former type-of-service field in IP

Baseline encoding defines that the DSCP must be set to a PCN-compatible DSCP n and the ECN-field [RFC3168] (Ramakrishnan, K., Floyd, S., and D. Black, "The Addition of Explicit Congestion Notification (ECN) to IP," September 2001.) indicates the specific PCN-mark. Baseline encoding offers four possible encoding states within a single DSCP with the following restrictions.

- *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.

- *Codepoint `10´ (ECT(0)) is used to indicate Not-marked PCN traffic.

- *Codepoint `11´ (CE) is used to indicate the most severe PCN-mark.

- *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] (Moncaster, T., Briscoe, B., and M. Menth, "Baseline Encoding and Transport of Pre-Congestion Information," November 2009.)).

[RFC6040] (Briscoe, B., "Tunnelling of Explicit Congestion Notification," November 2010.) 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] (Moncaster, T., Briscoe, B., and M. Menth, "Baseline Encoding and Transport of Pre-Congestion Information," November 2009.) was written. Happily the rules for use of the EXP codepoint are fully compatible with [RFC6040] (Briscoe, B., "Tunnelling of Explicit Congestion Notification," November 2010.). 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] (Moncaster, T., Briscoe, B., and M. Menth, "Baseline Encoding and Transport of Pre-Congestion Information," November 2009.) and in [I-D.ietf-pcn-encoding-comparison] (Chan, K., Karagiannis, G., Moncaster, T., Menth, M., Eardley, P., and B. Briscoe, "Pre-Congestion Notification Encoding Comparison," October 2010.).

3.3. Applicability of 3-in-1 PCN Encoding

TOC

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 (Recommendations for the Use of PCN Encoding Schemes). All nodes within the PCN-domain MUST be fully compliant with the ECN encapsulation rules set out in [RFC6040] (Briscoe, B., "Tunnelling of Explicit Congestion Notification," November 2010.). As such the encoding is not applicable in situations where legacy tunnels might exist.

4. Definition of 3-in-1 PCN Encoding

TOC

The 3-in-1 PCN encoding scheme is an extension of the baseline encoding scheme defined in [RFC5696] (Moncaster, T., Briscoe, B., and M. Menth, "Baseline Encoding and Transport of Pre-Congestion Information," November 2009.). 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 (3-in-1 PCN Encoding).

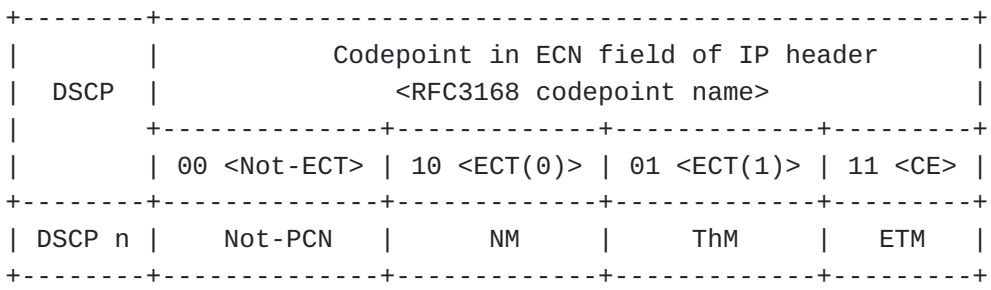

Figure 2: 3-in-1 PCN Encoding

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] (Eardley, P., "Metering and Marking Behaviour of PCN-Nodes," November 2009.).
>
> **ETM:** Excess-traffic-marked. Indicates a PCN-packet that has been marked by an excess-traffic-marker [RFC5670] (Eardley, P., "Metering and Marking Behaviour of PCN-Nodes," November 2009.).

---

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

TOC

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

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

TOC

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] (Moncaster, T., Briscoe, B., and M. Menth, “Baseline Encoding and Transport of Pre-Congestion Information,” November 2009.).

6.1. Backward Compatibility with Pre-existing PCN Implementations

TOC

This encoding complies with the rules for extending the baseline PCN encoding schemes in Section 5 of [RFC5696] (Moncaster, T., Briscoe, B., and M. Menth, “Baseline Encoding and Transport of Pre-Congestion Information,” November 2009.).
The term "compatibility" is meant in the following sense. It is possible to operate nodes with baseline encoding [RFC5696] (Moncaster, T., Briscoe, B., and M. Menth, “Baseline Encoding and Transport of Pre-Congestion Information,” November 2009.) 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] (Charny, A., Huang, F., Karagiannis, G., Menth, M., and T. Taylor, “PCN Boundary Node Behaviour for the Controlled Load (CL) Mode of Operation,” December 2010.). 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

TOC

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.

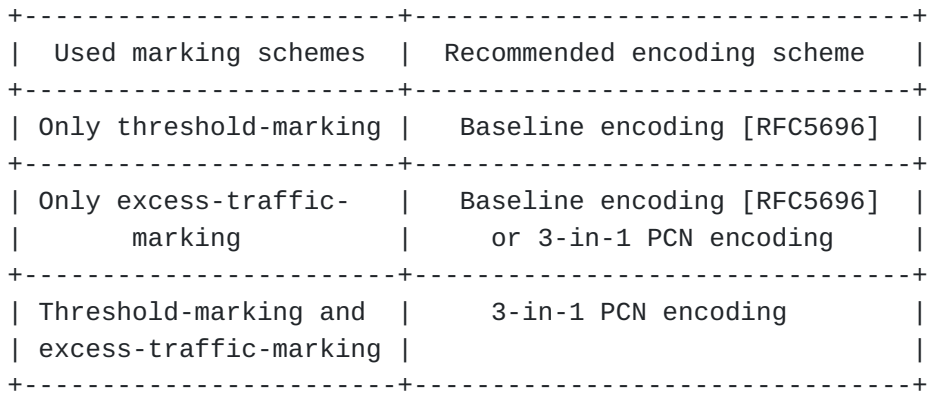

Figure 3: Guidelines for choosing PCN encoding schemes

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

TOC

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

TOC

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

TOC

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] (Briscoe, B., “Tunnelling of Explicit Congestion Notification,” November 2010.) 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

TOC

This memo includes no request to IANA.
Note to RFC Editor: this section may be removed on publication as an RFC.

8. Security Considerations

TOC

The security concerns relating to this extended PCN encoding are the same as those in [RFC5696] (Moncaster, T., Briscoe, B., and M. Menth, "Baseline Encoding and Transport of Pre-Congestion Information," November 2009.). 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

TOC

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] (Briscoe, B., "Tunnelling of Explicit Congestion Notification," November 2010.).

10. Acknowledgements

TOC

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

11. Comments Solicited

TOC

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

TOC

12.1. Normative References

TOC

[RFC2119]	Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels," BCP 14, RFC 2119, March 1997 (TXT, HTML, XML).
[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 (TXT, HTML, XML).
[RFC3168]	Ramakrishnan, K., Floyd, S., and D. Black, "The Addition of Explicit Congestion Notification (ECN) to IP," RFC 3168, September 2001 (TXT).
[RFC4301]	Kent, S. and K. Seo, "Security Architecture for the Internet Protocol," RFC 4301, December 2005 (TXT).
[RFC5129]	Davie, B., Briscoe, B., and J. Tay, "Explicit Congestion Marking in MPLS," RFC 5129, January 2008 (TXT).
[RFC5559]	Eardley, P., "Pre-Congestion Notification (PCN) Architecture," RFC 5559, June 2009 (TXT).
[RFC5670]	Eardley, P., "Metering and Marking Behaviour of PCN-Nodes," RFC 5670, November 2009 (TXT).
[RFC5696]	Moncaster, T., Briscoe, B., and M. Menth, "Baseline Encoding and Transport of Pre-Congestion Information," RFC 5696, November 2009 (TXT).
[RFC6040]	Briscoe, B., "Tunnelling of Explicit Congestion Notification," RFC 6040, November 2010 (TXT).

12.2. Informative References

TOC

[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 (TXT).
[I-D.ietf-pcn-encoding-comparison]	Chan, K., Karagiannis, G., Moncaster, T., Menth, M., Eardley, P., and B. Briscoe, "Pre-Congestion Notification Encoding Comparison," draft-ietf-pcn-encoding-comparison-03 (work in progress), October 2010 (TXT).
[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 (TXT).

Authors' Addresses

TOC

	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