Congestion and Pre Congestion
Internet-Draft
Intended status: Standards Track
Expires: August 14, 2009

T. Moncaster
BT
B. Briscoe
BT & UCL
M. Menth
University of Wuerzburg
February 10, 2009

Baseline Encoding and Transport of Pre-Congestion Information
draft-ietf-pcn-baseline-encoding-02

Status of This Memo

Copyright Notice

Abstract

Pre-congestion notification (PCN) provides information to support admission control and flow termination in order to protect the Quality of Service of inelastic flows. It does this by marking packets when traffic load on a link is approaching or has exceeded a threshold below the physical link rate. This document specifies how such marks are to be encoded into the IP header. The baseline encoding described here provides for only two PCN encoding states. It is designed to be easily extended to provide more encoding states but such schemes will be described in other documents.

Table of Contents

1. Introduction

Pre-congestion notification (PCN) provides information to support admission control and flow termination in order to protect the quality of service (QoS) of inelastic flows. This is achieved by marking packets according to the level of pre-congestion at nodes within a PCN-domain. These markings are evaluated by the egress nodes of the PCN-domain. [pcn-arch] describes how PCN packet markings can be used to assure the QoS of inelastic flows within a single DiffServ domain.

This document specifies how these PCN marks are encoded into the IP header. It also describes how packets are identified as belonging to a PCN flow. Some deployment models require two PCN encoding states, others require more. The baseline encoding described here only provides for two PCN encoding states. An extension of the baseline encoding described in [PCN-3-enc-state] provides for three PCN encoding states. Other extensions have also been suggested all of which can build on the baseline encoding. In order to ensure backward compatibility any alternative encoding schemes that claim compliance with PCN standards MUST extend this baseline scheme.

Changes from previous drafts (to be removed by the RFC Editor):

From -01 to -02:

Removed Appendix A and replaced with reference to [ecn-tunneling]

Moved Appendix B into main body of text.

Changed Appendix C to give deployment advice.

Minor changes throughout including checking consistency of capitalisation of defined terms.

Clarified that LU was deliberately excluded from encoding.

From -00 to -01:

Added section on restrictions for extension encoding schemes.

Included table in Appendix showing encoding transitions at different PCN nodes.

Checked for consistency of terminology.

Minor language changes for clarity.

Changes from previous filename

Filename changed from draft-moncaster-pcn-baseline-encoding.

Terminology changed for clarity (PCN-compatible DSCP and PCN-enabled packet).

Minor changes throughout.

Modified meaning of ECT(1) state to EXP.

Moved text relevant to behaviour of nodes into appendix for later transfer to new document on edge behaviours.

From draft-moncaster -01 to -02:

Minor changes throughout including tightening up language to remain consistent with the PCN Architecture terminology

From draft-moncaster -00 to -01:

Change of title from "Encoding and Transport of (Pre-)Congestion Information from within a DiffServ Domain to the Egress"

Extensive changes to Introduction and abstract.

Added a section on the implications of re-using a DSCP.

Added appendix listing possible operator scenarios for using this baseline encoding.

Minor changes throughout.

2. Requirements notation

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].

3. Terminology

The following terms are used in this document:

- Not-PCN - packets that are not PCN-enabled.

- o PCN-marked - codepoint indicating packets that have been marked at a PCN-interior-node using some PCN marking behaviour [pcn-marking-behaviour]. Also PM.

- o Not-marked - codepoint indicating packets that are PCN-capable but are not PCN-marked. Also NM.

- o PCN-enabled codepoints - collective term for all the NM and PM codepoints. By definition packets carrying such codepoints are PCN-packets.

- o PCN-compatible Diffserv codepoint - a Diffserv codepoint for which the ECN field is used to carry PCN markings rather than [RFC3168] markings.

In addition the document uses the terminology defined in [pcn-arch].

4. Encoding two PCN States in IP

The PCN encoding states are defined using a combination of the DSCP and ECN fields within the IP header. The baseline PCN encoding closely follows the semantics of ECN [RFC3168]. It allows the encoding of two PCN states: Not-Marked and PCN-Marked. It also allows for traffic that is not PCN capable to be marked as such (Not-PCN). Given the scarcity of codepoints within the IP header the baseline encoding leaves one codepoint free for experimental use. The following table defines how to encode these states in IP:

ECN codepoint	Not-ECT (00)	ECT(0) (10)	ECT(1) (01)	CE (11)
DSCP n	Not-PCN	NM	EXP	PM

Where DSCP n is a PCN-compatible DiffServ codepoint (see Section 4.2) and EXP means available for Experimental use. N.B. we deliberately reserve this codepoint for experimental use only (and not local use) to prevent any possible future compatability issues.

Table 1: Encoding PCN in IP

The following rules apply to all PCN traffic:

- o 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, such as the Voice-Admit codepoint defined in [voice-admit]. Guidelines for mixing traffic-types within a PCN-domain are given in [pcn-marking-behaviour].

- o Any packet that is not PCN-enabled (Not-PCN) but which shares the same DiffServ codepoint as PCN-enabled traffic MUST have the ECN field equal to 00.

The following table sets out the valid and invalid codepoint transitions at PCN-nodes for this baseline encoding. Extension encodings may have different rules regarding the validity of the transitions. Note that this table assumes there is a functional separation between a PCN-boundary-node and a PCN-interior-node such that PCN-boundary-nodes do not perform packet metering or marking functions. PCN-nodes MUST follow the encoding transition rules set out in this table (e.g. they MUST NOT set invalid codepoints on packets they forward). This table only applies to PCN-packets.

PCN node type	Codepoint in	Valid codepoint out	Invalid codepoint out
ingress	Any	NM (or Not-PCN)	PM
interior	NM	NM or PM	Not-PCN or EXP
interior	EXP +	EXP or PM	Not-PCN
interior	Not-PCN	Not-PCN	Any other codepoint
interior	PM	PM	Any other codepoint
egress	Any	00	Any other codepoint *

+ This SHOULD cause an alarm to be raised at a higher layer. The packet MUST be treated as if it were NM.

* Except where the egress node knows that other marks may be safely exposed outside the PCN-domain (e.g. [PCN-3-enc-state]).

Table 2: Valid and Invalid Codepoint Transitions for PCN-packets at PCN-nodes

If a pcn-interior-node compliant with this baseline encoding receives a

4.1. Rationale for Encoding

The exact choice of encoding was dictated by the constraints imposed by existing IETF RFCs, in particular [RFC3168], [RFC4301] and [RFC4774]. One of the tightest constraints was the need for any PCN encoding to survive being tunnelled through either an IP in IP tunnel or an IPSec Tunnel. [ecn-tunneling] explains this in more detail. The main effect of this constraint is that any PCN marking has to carry the 11 codepoint in the ECN field. If the packet is being tunneled then only the 11 codepoint gets copied into the inner header upon decapsulation. An additional constraint is the need to minimise the use of DiffServ codepoints as there is a limited supply of standards track codepoints remaining. Section 4.2 explains how we have minimised this still further by reusing pre-existing Diffserv codepoint(s) such that non-PCN traffic can still be distinguished from PCN traffic. There are a number of factors that were considered before deciding to set 10 as the NM state. These included similarity to ECN, presence of tunnels within the domain, leakage into and out of PCN-domain and incremental deployment.

The encoding scheme above seems to meet all these constraints and ends up looking very similar to ECN. This is perhaps not surprising given the similarity in architectural intent between PCN and ECN.

4.2. PCN-Compatible DiffServ Codepoints

Equipment complying with the baseline PCN encoding MUST allow PCN to be enabled for certain Diffserv codepoints. This document defines the term "PCN-compatible Diffserv codepoint" for such a DSCP. Enabling PCN for a DSCP switches on PCN marking behaviour for packets with that DSCP, but only if those packets also have their ECN field set to indicate a codepoint other than Not-PCN.

Enabling PCN marking behaviour disables any other marking behaviour (e.g. enabling PCN disables the default ECN marking behaviour introduced in [RFC3168]). All traffic scheduling and conditioning behaviours are discussed in [pcn-marking-behaviour]. This ensures compliance with the BCP guidance set out in [RFC4774].

5. Rules for Experimental Encoding Schemes

Any experimental encoding scheme MUST follow these rules to ensure backward compatibility with this baseline scheme:

- The 00 codepoint in the ECN field MUST mean Not-PCN.

- The 11 codepoint in the ECN field MUST mean PCN-marked (though this doesn't exclude other codepoints from carrying the same meaning).

- Once set the 11 codepoint in the ECN field MUST NOT be changed to any other codepoint.

- Any experimental scheme MUST include details of all valid and invalid codepoint transitions at any PCN nodes.

6. Backwards Compatibility

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 baseline encoding specified in this document defines PCN-compatible DiffServ codepoints as no longer supporting the default ECN semantics. As such this document is compatible with BCP 124. It should be noted that this baseline encoding effectively disables end-to-end ECN except where mechanisms are put in place to tunnel such traffic across the PCN-domain.

7. IANA Considerations

This document makes no request to IANA.

8. Security Considerations

Packets claim entitlement to be PCN marked by carrying a PCN-Compatible DSCP and a PCN-Enabled ECN codepoint. This encoding document is intended to stand independently of the architecture used to determine whether specific packets are authorised to be PCN marked, which will be described in a future separate document on PCN edge-node behaviour.

The PCN working group has initially been chartered to only consider a PCN-domain to be entirely under the control of one operator, or a set of operators who trust each other [PCN-charter]. However there is a requirement to keep inter-domain scenarios in mind when defining the PCN encoding. One way to extend to multiple domains would be to concatenate PCN-domains and use PCN-boundary-nodes back to back at borders. Then any one domain's security against its neighbours would be described as part of the proposed edge-node behaviour document.

One proposal on the table allows one to extend PCN across multiple domains without PCN-boundary-nodes back-to-back at borders [re-PCN]. It is believed that the encoding described here would be compatible with the security framework described there.

9. Conclusions

This document defines the baseline PCN encoding utilising a combination of a PCN-enabled DSCP and the ECN field in the IP header. This baseline encoding allows the existence of two PCN encoding states, not-Marked and PCN-Marked. It also allows for the co-existence of competing traffic within the same DSCP so long as that traffic doesn't require end-to-end ECN support. The encoding scheme is conformant with [RFC4774].

10. Acknowledgements

This document builds extensively on work done in the PCN working group by Kwok Ho Chan, Georgios Karagiannis, Philip Eardley, Anna Charny, Joe Babiarz and others. Thanks to Ruediger Geib for providing detailed comments on this document.

11. Comments Solicited

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.

[RFC4774] Floyd, S., "Specifying Alternate Semantics for the Explicit Congestion Notification (ECN) Field", BCP 124, RFC 4774, November 2006.

12.2. Informative References

[PCN-3-enc-state] Moncaster, T., Briscoe, B., and M. Menth, "A three state extended PCN encoding scheme", draft-moncaster-pcn-3-state-encoding-00 (work in progress), June 2008.

[PCN-charter] IETF, "IETF Charter for Congestion and Pre-Congestion Notification Working Group".

[RFC3168] Ramakrishnan, K., Floyd, S., and D. Black, "The Addition of Explicit Congestion Notification (ECN) to IP", RFC 3168, September 2001.

[RFC4301] Kent, S. and K. Seo, "Security Architecture for the Internet Protocol", RFC 4301, December 2005.

[RFC5127] Chan, K., Babiarz, J., and F. Baker, "Aggregation of DiffServ Service Classes", RFC 5127, February 2008.

[ecn-tunneling] Briscoe, B., "Layered Encapsulation of Congestion Notification", draft-ietf-tsvwg-ecn-tunnel-01 (work in progress), October 2008.

[pcn-arch] Eardley, P., "Pre-Congestion Notification (PCN) Architecture", draft-ietf-pcn-architecture-07 (work in progress), September 2008.

[pcn-marking-behaviour] Eardley, P., "Marking behaviour of PCN-nodes", draft-ietf-pcn-marking-behaviour-01 (work in progress), October 2008.

[re-PCN] Briscoe, B., "Emulating Border Flow Policing using Re-ECN on Bulk Data", draft-briscoe-re-pcn-border-cheat-00 (work in progress), July 2007.

[voice-admit] Baker, F., Polk, J., and M. Dolly, "DSCP for Capacity-Admitted Traffic", draft-ietf-tsvwg-admitted-realtime-dscp-05 (work in progress), November 2008.

Appendix A. PCN Deployment Considerations

A.1. Choice of Suitable DSCPs

The choice of which DSCP is most suitable for the PCN-domain is dependant 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 uniformly high link rates, the appropriate DSCPs would currently be those for the Real Time Traffic Class [RFC5127]. If the PCN domain includes lower speed links it would also be appropriate to use the DSCPs of the other traffic classes that [voice-admit] defines for use with admission control, such as the three video classes CS4, CS3 and AF4 and the Admitted Telephony Class.

A.2. Rationale for Using ECT(0) for Not Marked

The choice of which ECT codepoint to use for the Not Marked state was based on the following considerations:

- o [RFC3168] full functionality tunnel within PCN-domain: Either ECT is safe.

- o Leakage of traffic into PCN-domain: ECT(1) is less often correct.

- o Leakage of traffic out of PCN-domainL Either ECT is equally unsafe (since this would incorrectly indicate the traffic was ECN capable outside the controlled PCN-domain).

- o Incremental deployment: Either ECT is suitable as long as they are used consistently.

- o Conceptual consistency with other schemes: ECT(0) is conceptually consistent with [RFC3168].

Authors' Addresses

Toby Moncaster
BT
B54/70, Adastral Park
Martlesham Heath
Ipswich IP5 3RE
UK

Phone: +44 1473 648734
EMail: toby.moncaster@bt.com

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

Phone: +44 1473 645196
EMail: bob.briscoe@bt.com

Michael Menth
University of Wuerzburg
room B206, Institute of Computer Science
Am Hubland
Wuerzburg D-97074
Germany

Phone: +49 931 888 6644
EMail: menth@informatik.uni-wuerzburg.de