Congestion and Pre Congestion
Internet-Draft
Intended status: Standards Track
Expires: April 17, 2009

T. Moncaster
BT
B. Briscoe
BT & UCL
M. Menth
University of Wuerzburg

October 14, 2008

Baseline Encoding and Transport of Pre-Congestion Information
draft-ietf-pcn-baseline-encoding-01

Status of This Memo

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 -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.
- PCN-marked - codepoint indicating packets that have been marked at a PCN-interior-node using some PCN marking behaviour. Also PM.
- Not-marked - codepoint indicating packets that are PCN-capable but are not PCN-marked. Also NM.
- PCN-enabled codepoints - collective term for all the NM and PM codepoints.
- 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.

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.

4.1. Rationale for Encoding

The exact choice of encoding was dictated by the constraints imposed by existing IETF RFCs, in particular [RFC3168] 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. Appendix A explains this in 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].

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.

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 blocks 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 (see Appendix B).

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.

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

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.

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

[pcn-marking-behaviour] Eardley, P., "Marking behaviour of PCN-nodes", draft-ietf-pcn-marking-behaviour-00 (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, "DSCPs for Capacity-Admitted Traffic", draft-ietf-tsvwg-admitted-realtime-dscp-04 (work in progress), February 2008.

Appendix A. Tunnelling Constraints

The rules that govern the behaviour of the ECN field for IP-in-IP tunnels were defined in [RFC3168]. This allowed for two tunnel modes. The limited functionality mode sets the outer header to not-ECT, regardless of the value of the inner header, in other words disabling ECN within the tunnel. The full functionality mode copies the inner ECN field into the outer header if the inner header is not-ECT or either of the 2 ECT codepoints. If the inner header is CE then the outer header is set to ECT(0). On decapsulation, if the CE codepoint is set on the outer header then this is copied into the inner header. Otherwise the inner header is left unchanged. The stated reason for blocking CE from being copied to the outer header was to prevent this from being used as a covert channel through IPSec tunnels.

The IPSec protocol [RFC4301] changed the ECN tunnelling rule to allow IPSec tunnels to simply copy the inner header into the outer header. On decapsulation the outer header is discarded and the ECN field is only copied down if it is set to CE.

Because of the possible existence of tunnels, only CE (11) can be used as a PCN marking as it is the only mark that will always survive decapsulation. However there is a need for caution with all tunneling within the PCN-domain. RFC3168 full functionality IP in IP tunnels are expected to set the ECN field to ECT(0) if the inner ECN field is set to CE. This leads to the possibility that some packets within the PCN-domain that have already been marked may have that mark concealed further into the domain. This is undesirable for many PCN schemes and thus the PCN working group needs to decide whether to advise against the use of full functionality RFC3168 IP in IP tunnels within a PCN-domain to support the ongoing work within the Transport Area to rationalise the behaviour of IP in IP tunnels in respect to the ECN field and bring them in line with the behaviour of IPSec

tunnels [ecn-tunnelling].

Appendix B. PCN Node Behaviours

The following table of valid and invalid transitions, while necessary for the correct functioning of PCN they is not strictly part of the encoding scheme. The PCN working group needs to decide whether to include this in this baseline encoding or whether to transfer it to an alternative document.

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
interior	Not-PCN	Not-PCN	Any other codepoint
egress	Any	00	Any other codepoint *

* 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 Transitions at PCN nodes

It is also necessary to define a safe behaviour for baseline-compliant nodes to follow should they unexpectedly encounter a packet carrying the EXP (01) codepoint. The obvious safe behaviour would be to treat this as if it were a NM packet but to raise an alarm at a higher layer to check why the packet was there. An alternative safe approach is to treat it as a not-PCN packet but this might jeopardise partial deployment of any future experimental encoding scheme.

Appendix C. Deployment Scenarios for PCN Using Baseline Encoding

This appendix illustrates possible PCN deployment scenarios where the baseline encoding can be used and also explain a case for which baseline encoding is not sufficient. {Note this appendix is provided for information only}.

1. an operator requires only admission control. Then admission control is triggered from PCN-packets that are threshold-marked and this baseline encdoding scheme suffices.

2. an operator requires only flow termination. Then flow termination is triggered from PCN-packets that are excess-traffic-marked and this baseline encdoding scheme suffices.

3. an operator requires both admission control and flow termination. If both admission control and flow termination are triggered from PCN-packets that are excess-traffic-marked then this baseline encoding scheme suffices.

4. an operator requires both admission control triggered by packets that are threshold-marked and flow termination triggered by packets that are excess-traffic-marked. In this case the baseline encoding provides insufficient encoding states to achieve this.

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

Acknowledgement

This document was produced using xml2rfc v1.33 (of http://xml.resource.org/) from a source in RFC-2629 XML format.