Congestion and Pre Congestion
Internet-Draft
Intended status: Standards Track
Expires: January 12, 2009

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

July 11, 2008

Baseline Encoding and Transport of Pre-Congestion Information
draft-moncaster-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. Other documents describe extended encoding schemes that allow for three encoding states.

Status

This memo is posted as an Internet-Draft with an intent to eventually progress to standards track.

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 the PCN-domain. Two algorithms exist for that purpose. Excess traffic marking marks all PCN packets exceeding a certain reference rate on a link while threshold marking marks all PCN packets on a link when the PCN traffic rate exceeds the reference rate. 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 three. The baseline encoding described here only provides for two PCN encoding states. An extended encoding described in [PCN-3-enc-state] provides for three PCN encoding states.

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

From -01 to -02:

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

From -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 capable.
- 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-Capable codepoints - collective term for all the NM and PM codepoints.
- PCN-enabled Diffserv codepoint - a Diffserv codepoint for which PCN has been enabled on a particular machine.

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 the Type of Service field of the IP header which is a combination of the DSCP field and ECN field. 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). The following table defines how to encode these states in IP:

DSCP	not-ECT (00)	ECT(0) (10)	ECT(1) (01)	CE (11)
DSCP n	not-PCN	NM	NM	PM

Where DSCP n is a PCN-enabled DiffServ codepoint (see Section 4.2)

Table 1: Encoding PCN in IP

The following rules apply to all PCN traffic:

- PCN traffic MUST be marked with a PCN-enabled DiffServ Codepoint. That is a DiffServ codepoint that indicates that PCN is enabled. 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].

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

o Any packet that belongs to a PCN capable flow MUST have the ECN field set to one of the two ECT codepoints 10 or 01 at the PCN-ingress-node.

o Any packet that is PCN capable and has been PCN-marked by a PCN-interior-node MUST have the ECN field set to 11.

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]. Full details are contained in [pcn-enc-compare]. 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 use the ECN field set to 11 (CE codepoint). If the packet is being tunneled then only the CE codepoint gets copied into the inner header upon decapsulation. An additional constraint is the need to minimise the use of DiffServ codepoints as these are in increasingly short supply. 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.

The encoding scheme (Table 1) that best addresses the above constraints 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-Enabled DiffServ Codepoints

Equipment complying with the baseline PCN encoding MUST allow PCN to be enabled for a certain Diffserv codepoint or codepoints. This document defines the term "PCN-Enabled 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]). The scheduling behaviour used for a packet does not change whether PCN is enabled for a DSCP or not and whatever the setting of the ECN field.

4.2.1. Implications of re-using a DiffServ Codepoint

[RFC4774] requires that packets for which alternate ECN semantics (PCN semantics) are used are clearly distinguished from packets to which the default ECN semantics [RFC3168] apply. One means of doing this is using a DSCP to indicate that the ECN field is to be interpreted in a different manner. We have chosen to use this approach for PCN. Non-PCN-enabled forwarding nodes treat packets with a PCN-enabled DSCP like ECN traffic if appropriate ECN codepoints are set in the IP header. This has several consequences.

- Care must be taken to ensure that forwarding nodes do not interpret PCN encodings as ECN encodings, and that no harm is done if this were to happen. To that end, appropriate marking and re-marking is performed at the ingress and the egress of a PCN-domain.

- The re-used DSCP should be able to serve its original purpose which was not PCN support. This is achieved by marking the packets of such flows with a not-PCN codepoint.

- The scheduling behaviour is coupled with the DSCP only. Therefore, the same scheduling and buffer management rules are applied for non-PCN-capable and PCN-capable traffic using the same PCN-enabled DSCP.

- Once the ECN field of a packet is used for PCN encoding, it has lost its previous information unless this information is tunnelled through the PCN domain. Therefore, the baseline PCN encoding disables ECN for PCN-enabled DSCPs. [PCN-3-enc-state] provides end-to-end ECN support where this is needed.

4.3. Valid and Invalid Encoding Transitions at a PCN Node

PCN-boundary-node behaviour compliant with the PCN baseline encoding:

- Any packet with the ECN field already marked as CE or ECT arriving at a PCN-ingress-node SHOULD be dropped or downgraded to a lower class of service. Alternatively it MAY be tunnelled through the PCN-domain. It MUST NOT be admitted to the PCN-domain directly.

- On leaving the PCN-domain the ECN bits of every PCN-packet MUST be set to 00 (not-ECT).

PCN-interior-node behaviour compliant with the PCN baseline encoding:

- PCN-interior-nodes MUST NOT change not-PCN to another codepoint and they MUST NOT change a PCN-Capable codepoint to not-PCN.

- o PCN-interior-nodes that are in a pre-congestion state above the configured level MUST set the PM codepoint by changing the ECN bits of NM marked packets to 11.

- o The PM codepoint MUST NOT be changed to NM.

5. Backwards Compatability

BCP 124 [RFC4774] gives guidelines for specifying alternative semantics for the ECN field. It sets out a number of factors that must be taken into consideration. It also suggests various techniques to allow the co-existence of default ECN and alternative ECN semantics. The alternative semantics specified here are compliant with this BCP:

- o they use a DSCP to allow routers to distinguish that traffic uses the alternate ECN semantics;

- o these semantics are defined for use within a controlled domain;

- o ECN marked traffic is blocked from entering the PCN-domain directly (though it might be tunnelled through the PCN-domain).

- o All traffic leaving the controlled domain is re-marked as not-ECT.

6. IANA Considerations

This document makes no request to IANA. It does however suggest a change to the default ([RFC3168]) behaviour for the ECN field for the Voice-Admit [voice-admit] DSCP.

7. Security Considerations

Packets claim entitlement to be PCN marked by carrying a PCN-enabled DSCP and a PCN-Capable 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 edge-node behaviour document as above. 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.

8. 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 traffic that is not PCN-capable within the same DSCP so long as theat traffic doesn't require end-to-end ECN support. The encoding scheme is conformant with [RFC4774].

9. Acknowledgements

This document builds extensively on work done in the PCN working group by Kwok Ho Chan, Georgios Karagiannis, Philip Eardley and others. Full details of the alternative schemes that were considered for adoption can be found in the document [pcn-enc-compare]. Thanks to Ruediger Geib for providing detailed comments on this document.

10. 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.

11. References

11.1. 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.

11.2. 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-arch]
Eardley, P., "Pre-Congestion Notification Architecture", draft-ietf-pcn-architecture-03 (work in progress), February 2008.

[PCN-charter]
"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.

[pcn-enc-compare]
Chan, K., Karagiannis, G., Moncaster, T., Menth, M., Eardley, P., and B. Briscoe, "Pre-Congestion Notification Encoding Comparison", draft-chan-pcn-encoding-comparison-03 (work in progress), February 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 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 standard IP in IP tunnels SHOULD NOT be used within a PCN-domain. Further work is needed within the Transport Area to rationalise the behaviour of tunnels in respect to the ECN field.

Appendix B. 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 may wish to use PCN-based admission control only. To that end, threshold marking based on admissible rates might be used as the only PCN metering and marking algorithm. As a consequence, the PM marks on the packets are interpreted as admission-stop (AS) marks. The admission-control algorithm is based on "admissible-rate overload".

2. An operator may wish to use PCN-based flow termination only. To that end, excess rate marking based on supportable rates might be used as the only PCN metering and marking algorithm. As a consequence, the PM marks on the packets are interpreted as excess-traffic (ET) marks. The flow termination algorithm is

based on "supportable-rate overload".

3. An operator may wish to use both PCN-based admission control and flow termination. To that end, excess rate marking based on admissible rates may be used as the only PCN metering and marking algorithm. As a consequence, the PM marks on the packets are interpreted as admission-stop (AS) marks. Both the admission control and the flow termination algorithm are based on "admissible-rate overload".

4. An operator may wish to implement admission control based on threshold marking at admissible rates and flow termination based on excess rate marking at supportable rates because these methods are believed to work better with small ingress-egress aggregates. Then two different markings are needed. Such a deployment scenario is not supported by the PCN baseline encoding.

Authors' Addresses

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

Phone: +44 1473 648734
Email: toby.moncaster@bt.com
URI: http://www.cs.ucl.ac.uk/staff/B.Briscoe/

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

Full Copyright Statement

Intellectual Property

Acknowledgments

Funding for the RFC Editor function is provided by the IETF Administrative Support Activity (IASA). This document was produced using xml2rfc v1.32 (of http://xml.resource.org/) from a source in RFC-2629 XML format.