Congestion and Pre Congestion
Internet-Draft
Intended status: Historic
Expires: September 13, 2012

T. Moncaster
University of Cambridge
B. Briscoe
BT
M. Menth
University of Tuebingen
March 12, 2012

A PCN encoding using 2 DSCPs to provide 3 or more states
draft-ietf-pcn-3-state-encoding-02

Abstract

Pre-congestion notification (PCN) is a mechanism designed to protect the Quality of Service of inelastic flows within a controlled domain. 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 experimental encoding scheme specifies how three encoding states can be carried in the IP header using a combination of two DSCPs and the ECN bits. The Basic scheme only allows for three encoding states. The Full scheme provides 6 states, enough for limited end-to-end support for ECN as well.

Status

Since its original publication, the baseline encoding (RFC5696) on which this document depends has become obsolete. The PCN working Group has chosen to publish this as a historical document to preserve the details of the encoding and to allow it to be cited in other documents.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any 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 September 13, 2012.

Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

This document may contain material from IETF Documents or IETF Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF Trust the right to allow modifications of such material outside the IETF Standards Process. Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

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. 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. These configured rates are below the rate of the link thus providing notification before any congestion occurs (hence "pre-congestion notification"). The level of marking allows the boundary nodes to make decisions about whether to admit or block a new flow request, and (in abnormal circumstances) whether to terminate some of the existing flows, thereby protecting the QoS of previously admitted flows.

The baseline encoding described in [RFC5696] provides for deployment scenarios that only require two PCN encoding states. This document describes an experimental extension to the base-encoding in the IP header that adds two capabilities:

- the addition of a third PCN encoding state in the IP header
- preservation of the end-to-end semantics of the ECN field even though PCN uses the field within a PCN-region that interrupts the end-to-end path

The second of these capabilities is optional and the reasons for doing it are discussed in Section 5.

As in the baseline encoding, this extension encoding re-uses the ECN bits within the IP header within a controlled PCN-domain. This extension requires the use of two DSCPs as described later in this document. This experimental scheme is one of three that are being proposed within the PCN working group. The aim is to allow implementors to decide which scheme is most suitable for possible future standardisation.

Following the publication of new rules relating to the tunnelling of ECN marks [RFC6040], the PCN workign group decided to obsolete [RFC5696] in favour of the 3-in-1 encoding [I-D.ietf-pcn-3-in-1-encoding]. A side-effect of this decision was to make the encoding described in this document obsolete. However the PCN working group feels it is useful to have a formal historical record of this encoding. This ensures details of the encoding are not lost and also allows it to be cited in other documents.

1.1. Changes from Previous Drafts (to be removed by the RFC Editor)

From draft-ietf-pcn-3-state-encoding-01 to 02:

- Changed the document from teh experimental to the historic track
- Added notes to the Introduciton and Abstract explaining the change to historical
- Updated refs

From draft-ietf-pcn-3-state-encoding-00 to 01:

- Removed text implying the two DSCPs have different priority and added Section 6.3 specifying they must both have the same PHB.
- Made IANA considerations text more precise.
- Changed variable names for DSCP 1 & DSCP 2 to DSCP n & DSCP m to be consistent with baseline encoding.
- Updated refs

From draft-moncaster-pcn-3-state-encoding-01 to draft-ietf-pcn-3-state-encoding-00:

- Changed to WG draft. Title changed from "A three state extended PCN encoding scheme"
- Imposed new structure on document. This structure is intended to be followed by all extensions to the baseline PCN encoding scheme.
- Extensive changes throughout to ensure consistency with the baseline PCN encoding scheme.

From draft-moncaster-pcn-3-state-encoding-00 to 01:

- Checked terminology for consistency with [RFC5696]
- Minor editorial changes.

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

Most of the terminology used in this document is defined either in [RFC5559] or in [RFC5696]. The following additional terms are defined in this document:

- PCN-flow - a flow covered by a reservation but which hasn't signalled that it requires end-to-end ECN support.

- PCN-enabled-ECN-flow - a flow covered by reservation and for which the end-to-end transport has explicitly negotiated ECN support from the PCN-boundary-nodes.

- Not-marked (xxx), where xxx represents a standard ECN codepoint - packets that are PCN capable but carry no PCN mark. Abbreviated as NM(xxx). The (xxx) represents the ECN codepoint that the packet arrived with at the PCN-ingress-node e.g. NM(CE) represents a PCN capable packet that has no PCN marking but which arrived with the ECN bits set to congestion experienced.

4. The Requirement for Three PCN Encoding States

The PCN Marking Behaviours document [RFC5670] describes proposed PCN schemes that require traffic to be metered and marked using both Threshold and Excess Traffic schemes. In order to achieve this it is necessary to allow for three PCN encoding states. The constraints imposed by the way tunnels process the ECN field severely limit how to encode these states as explained in [RFC5696] and [RFC6040]. The obvious way to provide one more encoding state than the base encoding is through the use of an additional PCN-compatible DiffServ codepoint.

One aim of this document is to allow for experiments to show whether such schemes are better than those that only employ two PCN encoding states. As such, the additional DSCP will be taken from the EXP/LU pools defined in [RFC2474]. If the experiments demonstrate that PCN schemes employing three encoding states are significantly better than those only employing two, then at a later date IANA might be asked to assign a new PCN enabled DSCP from pool 1. Note that there are other experimental encoding schemes being considered which only use one DSCP but require either alternative tunnel semantics ([I-D.ietf-pcn-3-in-1-encoding]) or additional signalling ([I-D.ietf-pcn-psdm-encoding])in order to work.

5. Adding Limited End-to-End ECN Support to PCN

There are a number of use-cases where explicit preservation of end-to-end ECN semantics might be needed across a PCN domain. One of the use-cases suggests that the end-nodes might be running rate-adaptive codecs that would respond to ECN marks by reducing their transmission rate. If the sending transport sets the ECT codepoint, the setting of the ECN field as it arrives at the PCN ingress node will need to be re-instated as it leaves the PCN egress node.

If a PCN region is starting to suffer pre-congestion then it may make sense to expose marks generated within the PCN region by forwarding CE marks from the PCN egress to such a rate-adaptive endpoint. They would be in addition to any CE marks generated elsewhere on the end-to-end path. This would allow the endpoints to reduce the traffic rate. This will in turn help to alleviate the pre-congestion, potentially averting any need for call blocking or termination. However, the 'leaking' of CE marks out of the PCN region is potentially dangerous and could violate [RFC4774] if the end hosts don't understand ECN (see section 18.1.4 of [RFC3168]).

Therefore, a PCN region can only support end-to-end ECN if the PCN-boundary-nodes are sure that the end-to-end transport is ECN-capable. That way the PCN-egress-nodes can ensure that they only expose CE marks to those receivers that will correctly interpret them as a notification of congestion. The end-points may indicate they are ECN-capable through some higher-layer signalling process that sets up their reservation with the PCN boundary nodes. The exact process of negotiation is beyond the scope of this document but is likely to involve explicit two way signalling between the end-host and the PCN-domain.

In the absence of such signalling the default behaviour of the PCN egress node will be to clear the ECN field to 00 as in the baseline PCN encoding [RFC5696].

6. Encoding Three PCN States in IP

The three state PCN encoding scheme is based closely on that defined in [RFC5696] so that there will be no compatibility issues if a PCN-domain changes from using the baseline encoding scheme to the experimental scheme described here. There are two versions of the scheme. The basic three state scheme allows for carrying both Threshold-marked (ThM) and Excess-traffic-marked (ETM) traffic. The full scheme additionally allows end-to-end ECN to be carried across the PCN-domain.

6.1. Basic Three State Encoding

Table 1 below shows how to encode the three PCN states in IP.

DSCP	Not-ECT (00)	ECT(0) (10)	ECT(1) (01)	CE (11)
DSCP n	Not-PCN	NM	CU	ThM
DSCP m	Not-PCN	CU	CU	ETM

(where DSCP n is a PCN-compatible DiffServ codepoint (see [RFC5696]) and DSCP m is a PCN-compatible DSCP from the EXP/LU pools as defined in [RFC2474])

Table 1: Encoding three PCN states in IP

6.2. Full Three State Encoding

Table 2 shows how to additionally carry the end-to-end ECN state in the IP header.

DSCP	Not-ECT (00)	ECT(0) (10)	ECT(1) (01)	CE (11)
DSCP n	Not-PCN	NM(Not-ECT)	NM(CE)	ThM
DSCP m	Not-PCN	NM(ECT(0))	NM(ECT(1))	ETM

(where DSCP n is a PCN-compatible DiffServ codepoint (see [RFC5696]) and DSCP m is a PCN-compatible DSCP from the EXP/LU pools as defined in [RFC2474])

Table 2: Encoding three PCN states in IP

The four different Not-marked (NM) states allow for the addition of limited end-to-end ECN support as explained in the previous section.

WARNING: In order to comply with this encoding all the nodes within the PCN-domain MUST be configured with this encoding scheme. However there may be operators who choose not to be fully compliant with the scheme. If an operator chooses to leave some PCN-interior-nodes that only support two marking states (the baseline encoding [RFC5696]), then they must be aware of the following: Ideally such nodes would be configured to indicate pre-congestion or congestion using the ETM state since this would ensure they could notify worst-case congestion, however this is not possible since it requires the packets to be re-marked to DSCP

m (hence altering the baseline encoding). This means that such nodes will only be able to indicate ThM traffic.

6.3. Common Diffserv Per-Hop Behaviour

Packets carrying Diffserv codepoint 'DSCP n' or 'DSCP m' MUST all be treated with the same Diffserv PHB [RFC2474]. The choice of PHB is discussed in [RFC5559] and [RFC5696].

Two DSCPs are merely used to provide sufficient PCN encoding states, there is no need or intention to provide different scheduling or drop preference for each row in the table of PCN codepoints. Specifically:

- Both DSCPs MUST be served in the same queue to prevent reordering within an application flow.

- Both DSCPs MUST be assigned the same drop preference. Note that [RFC5670] already provides for preferential drop of excess-rate-marked packets, so assigning additional drop preference at the coarser granularity of each DSCP would be incorrect.

6.4. Valid and invalid codepoint transitions at PCN-ingress-nodes

A PCN-ingress-node operating the Basic version of the 3-State Encoding scheme MUST set the Not-marked codepoint on any arriving packet that belongs to a PCN-flow. It MUST set the not-PCN codepoint on any other packet.

A PCN-ingress-node operating the Full version of the 3-State Encoding scheme MUST establish whether a packet is a member of a PCN-enabled-ECN-flow. If it is, the PCN-ingress-node MUST set the appropriate NM(xxx) codepoint depending on the value carried in the ECN field of that packet. To be clear:

- A packet carrying the not-ECT codepoint in the ECN field MUST be assigned the NM(not-ECT) codepoint

- A packet carrying the ECT(0) codepoint in the ECN field MUST be assigned the NM(ECT(0)) codepoint

- A packet carrying the ECT(1) codepoint in the ECN field MUST be assigned the NM(ECT(1)) codepoint

- A packet carrying the CE codepoint in the ECN field MUST be assigned the NM(CE) codepoint

If it is not a member of such a flow then the behaviour MUST be the

same as for the Basic version of the Encoding scheme.

6.5. Valid and invalid codepoint transitions at PCN-interior-nodes

A PCN-interior-node MUST obey the following rules:

- It MUST NOT change the not-PCN codepoint to any other codepoint.
- It MAY change any Not-marked codepoint to either the Threshold-marked or Excess-traffic-marked codepoints.
- It MUST NOT change a Not-marked codepoint to the not-PCN codepoint.
- A Not-marked codepoint MUST NOT be changed to any other Not-marked codepoint.
- It MAY change the ThM codepoint to the ETM codepoint but it MUST NOT change the ThM codepoint to any other codepoint.
- It MUST NOT change the ETM codepoint to any other codepoint.

Obviously in every case a codepoint can remain unchanged. The precise rules governing which valid transition to use are set out in [RFC5670]

6.6. Forwarding traffic out of the PCN-domain

As each packet exits the PCN-domain, the PCN-egress-node MUST check whether it belongs to a PCN-enabled-ECN-flow. If it belongs to such a flow then the following rules dictate how the ECN field should be reset:

- A packet carrying the not-PCN codepoint MUST be given the not-ECT codepoint.
- A packet carrying the NM(not-ECT) codepoint MUST be assigned the not-ECT codepoint.
- A packet carrying the NM(ECT(0)) codepoint MUST be assigned the ECT(0) codepoint.
- A packet carrying the NM(ECT(1)) codepoint MUST be assigned the ECT(1) codepoint.
- A packet carrying the NM(CE) codepoint MUST be assigned the CE codepoint.

- A packet carrying the ThM codepoint MUST be assigned CE codepoint.

- A packet carrying the ETM codepoint MUST be assigned CE codepoint.

If the packet is part of a PCN-flow then it MUST be assigned the not-ECT codepoint regardless of which PCN-codepoint it carried.

In addition all packets should have their DSCP reset to the appropriate DSCP for the next hop. If the next hop is not another PCN region this will not be a PCN-compatible DSCP, and by default will be the best-efforts DSCP. Alterntively, higher layer signalling mechanisms may allow the DSCP that packets entered the PCN-domain with to be reinstated.

7. PCN-domain support for the PCN extension encoding

PCN traffic MUST be marked with a DiffServ codepoint that indicates PCN is enabled. To comply with the PCN extension encoding, codepoint 'DSCP n' MUST be a PCN-compatible DSCP assigned by IANA for use with the baseline PCN encoding [RFC5696] while 'DSCP m' can be a DSCP from pools 2 or 3 for experimental and local use [RFC2474]. The exact choice of DSCP may vary between PCN-domains but MUST be fixed within each PCN-domain.

7.1. End-to-End transport behaviour compliant with the PCN extension encoding

Transports wishing to use both PCN and end-to-end ECN MUST establish that their path supports this combination. Support of end-to-end ECN by PCN-boundary-nodes is OPTIONAL. Therefore transports MUST check with both the PCN-ingress-node and PCN-egress-node for each flow. The sending of such a request MUST NOT be taken to mean the request has been granted. The PCN-boundary-nodes MAY choose to inform the end-node of a successful request. The exact mechanism for such negotiation is beyond the scope of this document. A transport that receives no response or a negative response to a request to support end-to-end ECN within a flow reservation MUST set the ECN field of all subsequent packets in that flow to Not-ECT if it wishes to guarantee that the flow will receive PCN treatment.

If a domain wishes to use the full scheme described in Table 2 all nodes in that domain MUST be configured to understand the full scheme.

If either of a PCN ingress-egress pair does not support end-to-end ECN or if the end-to-end transport does not request support for end-to-end ECN then the PCN-boundary-nodes MUST assume the packet belongs

to a PCN-flow.

8. IANA Considerations

This document asks IANA to assign one DiffServ codepoint from Pool 2 or Pool 3 (for experimental/local use)[RFC2474]. Should this experimental PCN scheme prove sufficiently successful then IANA will be requested in a later document to assign a dedicated DiffServ codepoint from pool 1 for standards use and the experimental codepoint will be returned to its IANA pool.

9. Security Considerations

The security concerns relating to this extended PCN encoding are essentially the same as those in [RFC5696].

This extension coding gives end-to-end support for the ECN nonce [RFC3540], which is intended to protect the sender against the receiver or against network elements concealing a congestion experienced marking or a lost packet. PCN-based reservations combined with end-to-end ECN are intended for partially inelastic traffic using rate-adaptive codecs. Therefore the end-to-end transport is unlikely to be TCP, but at this time the nonce has only been defined for TCP transports.

10. Conclusions

This document describes an extended encoding scheme for PCN that provides for three encoding states as well as optional support for end-to-end ECN. The encoding scheme builds on the baseline encoding described in [RFC5696]. Using this encoding scheme it is possible for operators to conduct experiments to check whether the addition of an extra encoding state will significantly improve the performance of PCN. It will also allow experiments to determine whether there is a need for end-to-end ECN support within the PCN-domain (as against end-to-end ECN support through the use of IP-in-IP tunnelling or by downgrading the traffic to a lower service class).

11. Acknowledgements

This document builds extensively on work done in the PCN working group by Kwok Ho Chan, Georgios Karagiannis, Philip Eardley, Joe Babiarz and others. Full details of alternative schemes that were considered for adoption can be found in the document

[I-D.ietf-pcn-encoding-comparison].

12. Comments Solicited

(Section to be removed by RFC_Editor) Comments and questions are encouraged and very welcome. They can be addressed to the IETF Transport Area working group mailing list <tsvwg@ietf.org>, and/or to the authors.

13. References

13.1. 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.

[RFC5670] Eardley, P., "Metering and Marking Behaviour of PCN-Nodes", RFC 5670, November 2009.

[RFC5696] Moncaster, T., Briscoe, B., and M. Menth, "Baseline Encoding and Transport of Pre-Congestion Information", RFC 5696, November 2009.

13.2. Informative References

[I-D.ietf-pcn-3-in-1-encoding]
Briscoe, B., Moncaster, T., and M. Menth, "Encoding 3 PCN-States in the IP header using a single DSCP", draft-ietf-pcn-3-in-1-encoding-09 (work in progress), March 2012.

[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", draft-ietf-pcn-encoding-comparison-09 (work in progress), March 2012.

[I-D.ietf-pcn-psdm-encoding]
Menth, M., Babiarz, J., Moncaster, T., and B. Briscoe, "PCN Encoding for Packet-Specific Dual Marking (PSDM Encoding)", draft-ietf-pcn-psdm-encoding-01 (work in

progress), March 2010.

[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.

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

[RFC3540] Spring, N., Wetherall, D., and D. Ely, "Robust Explicit Congestion Notification (ECN) Signaling with Nonces", RFC 3540, June 2003.

[RFC5559] Eardley, P., "Pre-Congestion Notification (PCN) Architecture", RFC 5559, June 2009.

[RFC6040] Briscoe, B., "Tunnelling of Explicit Congestion Notification", RFC 6040, November 2010.

Authors' Addresses

Toby Moncaster
University of Cambridge
Computer Laboratory
JJ Thomson Avenue
Cambridge CB3 0FD
UK

Phone: +44 1223 763654
Email: toby@moncaster.com

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

Phone: +44 1473 645196
Email: bob.briscoe@bt.com
URI: http://www.bobbriscoe.net

Michael Menth
University of Tuebingen
Department of Computer Science
Sand 13
Tuebingen D-72076
Germany

Phone: +49 07071 29 70505
Email: menth@informatik.uni-tuebingen.de
URI: http://www.kn.inf.uni-tuebingen.de