Congestion and Pre Congestion
Internet-Draft
Intended status: Experimental
Expires: September 10, 2009

T. Moncaster
BT
B. Briscoe
BT & UCL
M. Menth
University of Wuerzburg
March 9, 2009

A three state extended PCN encoding scheme
draft-moncaster-pcn-3-state-encoding-01

Status of this Memo

Copyright Notice

Abstract

Pre-congestion notification (PCN) is a mechanism designed 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 baseline encoding specified how two encoding states could be encoded into the IP header. This document specified an extension to the baseline encoding that enables three encoding states to be carried in the IP header as well as enabling limited support for end-to-end ECN.

Status (to be removed by RFC Editor)

This memo is posted as an Internet-Draft with an intent to eventually be published as an experimental RFC. The PCN Working Group will be asked to adopt this memo as a Working Group document describing one of several possible experimental PCN encoding schemes. The intention is that the title of this document will change to avoid confusion with the three state marking scheme.

Changes from previous drafts

From 00 to 01:

- Checked terminology for consistency with [I-D.ietf-pcn-baseline-encoding]
- Minor editorial changes.

Table of Contents

1. Introduction

Pre-congestion notification provides information to support admission control and flow termination at the boundary nodes of a Diffserv region in order to protect the quality of service (QoS) of inelastic flows [I-D.ietf-pcn-architecture]. This is achieved by marking packets on interior nodes according to some metering function implemented at each node. Excess traffic marking marks PCN packets that exceed a certain reference rate on a link while threshold marking marks all PCN packets on a link when the PCN traffic rate exceeds a higher reference rate. These marks are monitored by the egress nodes of the PCN-domain.

The baseline encoding described in [I-D.ietf-pcn-baseline-encoding] 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 encoding 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

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 [I-D.ietf-pcn-architecture] or in [I-D.ietf-pcn-baseline-encoding]. 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. Also 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 architecture [I-D.ietf-pcn-architecture] 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 [I-D.ietf-pcn-baseline-encoding] and [I-D.ietf-tsvwg-ecn-tunnel]. The obvious way to provide one more encoding state than the base encoding is through the use of an additional PCN-compatible DiffServ codepoints. 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

as 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.

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

[I-D.sarker-pcn-ecn-pcn-usecases] suggests 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 edge 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 [I-D.ietf-pcn-baseline-encoding].

6. Encoding Three PCN States in IP

The three state PCN encoding scheme is based closely on that defined in [I-D.ietf-pcn-baseline-encoding] 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 us to carry end-to-end ECN.

6.1. Basic Three State Encoding

The following table shows how to encode the three PCN states in IP. The authors spent some time trying to establish which way round to put the two marked states before settling on this. Because it is envisaged that DSCP 2 will be of lower priority than DSCP 1 the change in marking from Threshold to Excess Traffic involves downgrading the traffic which seems to be consistent with the requirement that such changes should not be reversed.

DSCP	Not-ECT (00)	ECT(0) (10)	ECT(1) (01)	CE (11)
DSCP 1	Not-PCN	NM	CU	ThM
DSCP 2	Not-PCN	CU	CU	ETM

Where DSCP 1 is a PCN-compatible DiffServ codepoint (see

[I-D.ietf-pcn-baseline-encoding]) and DSCP 2 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 1	Not-PCN	NM(Not-ECT)	NM(CE)	ThM
DSCP 2	Not-PCN	NM(ECT(0))	NM(ECT(1))	ETM

Where DSCP 1 is a PCN-compatible DiffServ codepoint (see [I-D.ietf-pcn-baseline-encoding]) and DSCP 2 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

6.2.1. 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 table shows how the ECN field should be re-set. 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 re-instated.

DSCP	00	10	01	11
DSCP 1	Not PCN --> Not ECT	NM(Not-ECT) --> not-ECT	NM(CE) --> CE	ThM --> CE
DSCP 2	Not PCN --> Not ECT	NM(ECT(0)) --> ECT(0)	NM(ECT(1)) --> ECT(1)	ETM --> CE

Where each cell gives the incoming PCN state and the outgoing ECN state.

Table 3: Egress rules for resetting ECN field for PCN Enabled ECN Flows

For packets belonging to a PCN-flow the ECN field MUST be reset to not-ECT (00) as defined in [I-D.ietf-pcn-baseline-encoding].

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, this codepoint is either a PCN-compatible DSCP assigned by IANA for use with the baseline PCN encoding [I-D.ietf-pcn-baseline-encoding] or 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.

All nodes within a PCN-domain MUST understand and support the three PCN states of the PCN extension coding. Therefore if any PCN-node does not support three PCN encoding states, any node in the same PCN-domain MUST NOT be configured to use three PCN encoding states as defined here.

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

Transports wishing to use both a reservation 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.

7.2. PCN-boundary-node behaviour compliant with the PCN extension encoding

- If both the PCN ingress and egress nodes support end-to-end ECN, and the transport has successfully requested end-to-end ECN the flow becomes a PCN-enabled-ECN-flow.

- 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.

7.2.1. Behaviour for packets belonging to a PCN-flow

- If a packet belongs to a PCN-flow arrives at the PCN-ingress-node with its ECN field already marked as CE or ECT, it SHOULD be dropped. Alternatively it MAY be downgraded to a lower (non-PCN) service class or MAY be tunnelled through the PCN region. It MUST NOT be admitted to the PCN region directly.

- When a packet belonging to a PCN-flow carrying the not-ECT codepoint arrives at the PCN-ingress-node, the ECN field MUST be set to ECT(0) (10) and the DiffServ field set to DSCP 1.

- When a packet belonging to a PCN-flow leaves the PCN-domain through the PCN-egress-node, the ECN bits MUST be set to not-ECT (00).

7.2.2. Behaviour for packets belonging to a PCN-enabled-ECN-flow

- When a packet belonging to a PCN-enabled-ECN-flow arrives at the PCN-ingress-node, then the ECN field and DSCP MUST be set to the appropriate NM(xxx) setting as shown in Table 2.

- When a packet belonging to a PCN-enabled-ECN-flow leaves the PCN-region through a PCN-egress-node, the ECN bits MUST be set according to Table 3 and the DSCP MUST be set to the appropriate DSCP for the next hop as discussed in Section 6.2.1 above.

7.3. PCN-interior-node behaviour compliant with the PCN extension encoding

- If a PCN interior node indicates that a packet is to be threshold marked then the ThM codepoint MUST be set by changing the ECN bits to 11 and ensuring the Diffserv field is set to DSCP1.

- If a PCN interior node indicates that a packet is to be excess traffic marked then the EM codepoint MUST be set by changing the ECN bits to 11 and ensuring the Diffserv field is set to DSCP2 as defined above.

7.4. Behaviour of any PCN node compliant with the PCN extension encoding

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

- ThM MUST NOT be changed to NM.

- ETM MUST NOT be changed to ThM or to NM.

8. IANA Considerations

This document asks IANA to assign one DiffServ codepoint from Pool 2 or Pool 3 (for experimental/local use)[RFC2474]. Should any of the three encoding state experimental PCN schemes prove sufficiently successful then, at a later date, IANA will be requested in a later document to assign a dedicated DiffServ codepoint from pool 1 for standards use.

9. Security Considerations

The security concerns relating to this extended PCN encoding are essentially the same as those in [I-D.ietf-pcn-baseline-encoding].

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 support for end-to-end ECN. The encoding scheme builds on the baseline encoding described in [I-D.ietf-pcn-baseline-encoding]. 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.chan-pcn-encoding-comparison].

12. Comments Solicited

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.

13.2. Informative References

[I-D.chan-pcn-encoding-comparison] Chan, K., Karagiannis, G., Moncaster, T., Menth, M., Eardley, P., and B. Briscoe, "Pre-Congestion Notification Encoding Comparison", draft-chan-pcn-encoding-comparison-04 (work in progress),

February 2008.

[I-D.ietf-pcn-architecture] Eardley, P., "Pre-Congestion Notification (PCN) Architecture", draft-ietf-pcn-architecture-09 (work in progress), January 2009.

[I-D.ietf-pcn-baseline-encoding] Moncaster, T., Briscoe, B., and M. Menth, "Baseline Encoding and Transport of Pre-Congestion Information", draft-ietf-pcn-baseline-encoding-02 (work in progress), February 2009.

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

[I-D.sarker-pcn-ecn-pcn-usecases] Sarker, Z. and I. Johansson, "Usecases and Benefits of end to end ECN support in PCN Domains", draft-sarker-pcn-ecn-pcn-usecases-02 (work in progress), November 2008.

[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.

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