Network Working Group
Request for Comments: 5696
Category: Standards Track

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

November 2009

Baseline Encoding and Transport of Pre-Congestion Information

Abstract

The objective of the Pre-Congestion Notification (PCN) architecture is to protect the quality of service (QoS) of inelastic flows within a Diffserv domain. It achieves this by marking packets belonging to PCN-flows when the rate of traffic exceeds certain configured thresholds on links in the domain. These marks can then be evaluated to determine how close the domain is to being congested. This document specifies how such marks are encoded into the IP header by redefining the Explicit Congestion Notification (ECN) codepoints within such domains. The baseline encoding described here provides only two PCN encoding states: Not-marked and PCN-marked. Future extensions to this encoding may be needed in order to provide more than one level of marking severity.

Status of This Memo

This document specifies an Internet standards track protocol for the Internet community, and requests discussion and suggestions for improvements. Please refer to the current edition of the "Internet Official Protocol Standards" (STD 1) for the standardization state and status of this protocol. Distribution of this memo is unlimited.

Table of Contents

1. Introduction

The objective of the Pre-Congestion Notification (PCN) architecture [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 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.

This document specifies how these PCN-marks are encoded into the IP header by reusing the bits of the Explicit Congestion Notification (ECN) field [RFC3168]. 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. However, the encoding can be easily extended to provide more states. Rules for such extensions are given in Section 5.

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 and Abbreviations

3.1. Terminology

The terms PCN-capable, PCN-domain, PCN-node, PCN-interior-node, PCN-ingress-node, PCN-egress-node, PCN-boundary-node, PCN-traffic, PCN-packets and PCN-marking are used as defined in [RFC5559]. The following additional terms are defined in this document:

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

- PCN-marked codepoint - a codepoint that indicates packets that have been marked at a PCN-interior-node using some PCN-marking behaviour [RFC5670]. Abbreviated to PM.

- Not-marked codepoint - a codepoint that indicates packets that are PCN-capable but that are not PCN-marked. Abbreviated to NM.

- not-PCN codepoint - a codepoint that indicates packets that are not PCN-capable.

3.2. List of Abbreviations

The following abbreviations are used in this document:

- AF = Assured Forwarding [RFC2597]

- CE = Congestion Experienced [RFC3168]

- CS = Class Selector [RFC2474]

- DSCP = Diffserv codepoint

- ECN = Explicit Congestion Notification [RFC3168]

- ECT = ECN Capable Transport [RFC3168]

- EF = Expedited Forwarding [RFC3246]

- EXP = Experimental

- NM = Not-marked

- PCN = Pre-Congestion Notification

- PM = PCN-marked

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

Table 1: Encoding PCN in IP

In the table above, DSCP n is a PCN-compatible Diffserv codepoint (see Section 4.4) and EXP means available for Experimental use. N.B. we deliberately reserve this codepoint for experimental use only (and not local use) to prevent future compatibility issues.

The following rules apply to all PCN-traffic:

- 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. Appendix A.1 gives guidance to implementors on suitable DSCPs. Guidelines for mixing traffic types within a PCN-domain are given in [RFC5670].

- Any packet arriving at the PCN-ingress-node that shares a PCN-compatible DSCP and is not a PCN-packet MUST be marked as not-PCN within the PCN-domain.

- If a packet arrives at the PCN-ingress-node with its ECN field already set to a value other than not-ECT, then appropriate action MUST be taken to meet the requirements of [RFC3168]. The simplest appropriate action is to just drop such packets. However, this is a drastic action that an operator may feel is undesirable. Appendix B provides more information and summarises other alternative actions that might be taken.

4.1. Marking Packets

[RFC5670] states that any encoding scheme document must specify the required action to take if one of the marking algorithms indicates that a packet needs to be marked. For the baseline encoding scheme, the required action is simply as follows:

- If a marking algorithm indicates the need to mark a PCN-packet, then that packet MUST have its PCN codepoint set to 11, PCN-marked.

4.2. Valid and Invalid Codepoint Transitions

A PCN-ingress-node MUST set the Not-marked (10) codepoint on any arriving packet that belongs to a PCN-flow. It MUST set the not-PCN (00) codepoint on all other packets sharing a PCN-compatible Diffserv codepoint.

The only valid codepoint transitions within a PCN-interior-node are from NM to PM (which should occur if either meter indicates a need to PCN-mark a packet [RFC5670]) and from EXP to PM. PCN-nodes that only implement the baseline encoding MUST be able to PCN-mark packets that arrive with the EXP codepoint. This should ease the design of experimental schemes that want to allow partial deployment of experimental nodes alongside nodes that only implement the baseline encoding. The following table gives the full set of valid and invalid codepoint transitions.

	Codepoint Out			
Codepoint in	not-PCN(00)	NM(10)	EXP(01)	PM(11)
not-PCN(00)	Valid	Not valid	Not valid	Not valid
NM(10)	Not valid	Valid	Not valid	Valid
EXP(01)*	Not valid	Not valid	Valid	Valid
PM(11)	Not valid	Not valid	Not valid	Valid

* This MAY cause an alarm to be raised at a management layer. See paragraph above for an explanation of this transition.

Table 2: Valid and Invalid Codepoint Transitions for PCN-Packets at PCN-Interior-Nodes

The codepoint transition constraints given here apply only to the baseline encoding scheme. Constraints on codepoint transitions for future experimental schemes are discussed in Section 5.

A PCN-egress-node SHOULD set the not-PCN (00) codepoint on all packets it forwards out of the PCN-domain. The only exception to this is if the PCN-egress-node is certain that revealing other codepoints outside the PCN-domain won't contravene the guidance given in [RFC4774]. For instance, if the PCN-ingress-node has explicitly informed the PCN-egress-node that this flow is ECN-capable, then it might be safe to expose other codepoints.

4.3. 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-TUN] 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 since this is the only codepoint that is guaranteed to be copied down into the forwarded header upon decapsulation. An additional constraint is the need to minimise the use of Diffserv codepoints because there is a limited supply of Standards Track codepoints remaining. Section 4.4 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 choosing to set 10 as the NM state instead of 01. These included similarity to ECN, presence of tunnels within the domain, leakage into and out of the PCN-domain, and incremental deployment (see Appendix A.2).

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.

4.4. 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. To be clear, any packets with such a DSCP will be PCN-enabled only if they are within a PCN-domain and have their ECN field set to indicate a codepoint other than not-PCN.

Enabling PCN-marking behaviour for a specific DSCP disables any other marking behaviour (e.g., enabling PCN replaces the default ECN marking behaviour introduced in [RFC3168]) with the PCN-metering and -marking behaviours described in [RFC5670]). This ensures compliance with the Best Current Practice (BCP) guidance set out in [RFC4774].

The PCN working group has chosen not to define a single DSCP for use with PCN for several reasons. Firstly, the PCN mechanism is applicable to a variety of different traffic classes. Secondly, Standards Track DSCPs are in increasingly short supply. Thirdly, PCN is not a scheduling behaviour -- rather, it should be seen as being

essentially a marking behaviour similar to ECN but intended for inelastic traffic. More details are given in the informational Appendix A.1.

4.4.1. Co-Existence of PCN and Not-PCN Traffic

The scarcity of pool 1 DSCPs, coupled with the fact that PCN is envisaged as a marking behaviour that could be applied to a number of different DSCPs, makes it essential that we provide a not-PCN state. As stated above (and expanded in Appendix A.1), the aim is for PCN to re-use existing DSCPs. Because PCN redefines the meaning of the ECN field for such DSCPs, it is important to allow an operator to still use the DSCP for non-PCN-traffic. This is achieved by providing a not-PCN state within the encoding scheme. Section 3.5 of [RFC5559] discusses how competing-non-PCN-traffic should be handled.

5. Rules for Experimental Encoding Schemes

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

- All PCN-interior-nodes within a PCN-domain MUST interpret the 00 codepoint in the ECN field as not-PCN and MUST NOT change it to another value. Therefore, a PCN-ingress-node wishing to disable PCN-marking for a packet with a PCN-compatible Diffserv codepoint MUST set the ECN field to 00.

- The 11 codepoint in the ECN field MUST indicate that the packet has been PCN-marked as the result of one or both of the meters indicating a need to PCN-mark a packet [RFC5670]. The experimental scheme MUST define which meter(s) trigger this marking.

- The 01 Experimental codepoint in the ECN field MAY mean PCN-marked or it MAY carry some other meaning. However, any experimental scheme MUST define its meaning in the context of that experiment.

- If both the 01 and 11 codepoints are being used to indicate PCN-marked, then the 11 codepoint MUST be taken to be the more severe marking and the choice of which meter sets which mark MUST be defined.

- 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. Backward 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.

On its own, this baseline encoding cannot support both ECN marking end-to-end (e2e) and PCN-marking within a PCN-domain. It is possible to do this by carrying e2e ECN across a PCN-domain within the inner header of an IP-in-IP tunnel, or by using a richer encoding such as the proposed experimental scheme in [PCN-ENC].

In any PCN deployment, traffic can only enter the PCN-domain through PCN-ingress-nodes and leave through PCN-egress-nodes. PCN-ingress-nodes ensure that any packets entering the PCN-domain have the ECN field in their outermost IP header set to the appropriate PCN codepoint. PCN-egress-nodes then guarantee that the ECN field of any packet leaving the PCN-domain has the correct ECN semantics. This prevents unintended leakage of ECN marks into or out of the PCN-domain, and thus reduces backward-compatibility issues.

7. Security Considerations

PCN-marking only carries a meaning within the confines of a PCN-domain. This encoding document is intended to stand independently of the architecture used to determine how specific packets are authorised to be PCN-marked, which will be described in separate documents on PCN-boundary-node behaviour.

This document assumes the PCN-domain to be entirely under the control of a single operator, or a set of operators who trust each other. However, future extensions to PCN might include inter-domain versions where trust cannot be assumed between domains. If such schemes are proposed, they must ensure that they can operate securely despite the lack of trust. However, such considerations are beyond the scope of this document.

One potential security concern is the injection of spurious PCN-marks into the PCN-domain. However, these can only enter the domain if a PCN-ingress-node is misconfigured. The precise impact of any such misconfiguration will depend on which of the proposed PCN-boundary-node behaviour schemes is used, but in general spurious marks will lead to admitting fewer flows into the domain or potentially terminating too many flows. In either case, good management should

be able to quickly spot the problem since the overall utilisation of the domain will rapidly fall.

8. Conclusions

This document defines the baseline PCN encoding, utilising a combination of a PCN-compatible 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 does not require ECN support within the PCN-domain. The encoding scheme is conformant with [RFC4774]. The working group has chosen not to define a single DSCP for use with PCN. The rationale for this decision along with advice relating to the choice of suitable DSCPs can be found in Appendix A.1.

9. 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 and Gorry Fairhurst for providing detailed comments on this document.

10. References

10.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

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

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

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

10.2. Informative References

[ECN-TUN] Briscoe, B., "Tunnelling of Explicit Congestion Notification", Work in Progress, July 2009.

[PCN-ENC] Moncaster, T., Briscoe, B., and M. Menth, "A PCN encoding using 2 DSCPs to provide 3 or more states", Work in Progress, April 2009.

[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.

[RFC2597] Heinanen, J., Baker, F., Weiss, W., and J. Wroclawski, "Assured Forwarding PHB Group", RFC 2597, June 1999.

[RFC3246] Davie, B., Charny, A., Bennet, J., Benson, K., Le Boudec, J., Courtney, W., Davari, S., Firoiu, V., and D. Stiliadis, "An Expedited Forwarding PHB (Per-Hop Behavior)", RFC 3246, March 2002.

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

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

[RFC4594] Babiarz, J., Chan, K., and F. Baker, "Configuration Guidelines for DiffServ Service Classes", RFC 4594, August 2006.

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

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

Appendix A. PCN Deployment Considerations (Informative)

A.1. Choice of Suitable DSCPs

The PCN working group chose not to define a single DSCP for use with PCN for several reasons. Firstly, the PCN mechanism is applicable to a variety of different traffic classes. Secondly, Standards Track DSCPs are in increasingly short supply. Thirdly, PCN is not a scheduling behaviour -- rather, it should be seen as being a marking behaviour similar to ECN but intended for inelastic traffic. The choice of which DSCP is most suitable for a given PCN-domain is dependent 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 sufficient aggregation, the appropriate DSCPs would currently be those for the Real-Time Treatment Aggregate [RFC5127]. The PCN working group suggests using admission control for the following service classes (defined in [RFC4594]):

- Telephony (EF)
- Real-time interactive (CS4)
- Broadcast Video (CS3)
- Multimedia Conferencing (AF4)

CS5 is excluded from this list since PCN is not expected to be applied to signalling traffic.

PCN-marking is intended to provide a scalable admission-control mechanism for traffic with a high degree of statistical multiplexing. PCN-marking would therefore be appropriate to apply to traffic in the above classes, but only within a PCN-domain containing sufficiently aggregated traffic. In such cases, the above service classes may well all be subject to a single forwarding treatment (treatment aggregate [RFC5127]). However, this does not imply all such IP traffic would necessarily be identified by one DSCP -- each service class might keep a distinct DSCP within the highly aggregated region [RFC5127].

Additional service classes may be defined for which admission control is appropriate, whether through some future standards action or through local use by certain operators, e.g., the Multimedia Streaming service class (AF3). This document does not preclude the use of PCN in more cases than those listed above.

Note: The above discussion is informative not normative, as operators are ultimately free to decide whether to use admission control for

certain service classes and whether to use PCN as their mechanism of choice.

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:

- [RFC3168] full-functionality tunnel within the PCN-domain: Either ECT is safe.
- Leakage of traffic into PCN-domain: Because of the lack of take-up of the ECN nonce [RFC3540], leakage of ECT(1) is less likely to occur and so might be considered safer.

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

- Incremental deployment: Either codepoint is suitable, providing that the codepoints are used consistently.

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

Overall, this seemed to suggest that ECT(0) was most appropriate to use.

Appendix B. Co-Existence of PCN and ECN (Informative)

This baseline encoding scheme redefines the ECN codepoints within the PCN-domain. As packets with a PCN-compatible DSCP leave the PCN-domain, their ECN field is reset to not-ECT (00). This is a problem for the operator if packets with a PCN-compatible DSCP arrive at the PCN-domain with any ECN codepoint other than not-ECN. If the ECN-codepoint is ECT(0) (10) or ECT(1) (01), resetting the ECN field to 00 effectively turns off end-to-end ECN. This is undesirable as it removes the benefits of ECN, but [RFC3168] states that it is no worse than dropping the packet. However, if a packet was marked with CE (11), resetting the ECN field to 00 at the PCN egress node violates the rule that CE-marks must never be lost except as a result of packet drop [RFC3168].

A number of options exist to overcome this issue. The most appropriate option will depend on the circumstances and has to be a decision for the operator. The definition of the action is beyond the scope of this document, but we briefly explain the four broad categories of solution below: tunnelling the packets, using an extended encoding scheme, signalling to the end systems to stop using ECN, or re-marking packets to a different DSCP.

- Tunnelling the packets across the PCN-domain (for instance, in an IP-in-IP tunnel from the PCN-ingress-node to the PCN-egress-node) preserves the original ECN marking on the inner header.

- An extended encoding scheme can be designed that preserves the original ECN codepoints. For instance, if the PCN-egress-node can determine from the PCN codepoint what the original ECN codepoint was, then it can reset the packet to that codepoint. [PCN-ENC] partially achieves this but is unable to recover ECN markings if the packet is PCN-marked in the PCN-domain.

- Explicit signalling to the end systems can indicate to the source that ECN cannot be used on this path (because it does not support ECN and PCN at the same time). Dropping the packet can be thought of as a form of silent signal to the source, as it will see any ECT-marked packets it sends being dropped.

- Packets that are not part of a PCN-flow but which share a PCN-compatible DSCP can be re-marked to a different local-use DSCP at the PCN-ingress-node with the original DSCP restored at the PCN-egress. This preserves the ECN codepoint on these packets but relies on there being spare local-use DSCPs within the PCN-domain.

Authors' Addresses

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

Phone: +44 7918 901170
EMail: toby.moncaster@bt.com

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

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

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

Phone: +49 931 318 6644
EMail: menth@informatik.uni-wuerzburg.de