Baseline Encoding and Transport of Pre-Congestion Information
draft-ietf-pcn-baseline-encoding-00

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 extensible 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.

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

From -02 to -03:

   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 -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:

o Not-PCN - packets that are not PCN capable.

o PCN-marked - codepoint indicating packets that have been marked at
   a PCN-interior-node using some PCN marking behaviour. Also PM.

o Not-Marked - codepoint indicating packets that are PCN capable but
   are not PCN-marked. Also NM.

o PCN-enabled codepoints - collective term for all the NM and PM
   codepoints.

o PCN-compatible Diffserv codepoint - a Diffserv codepoint for which
   the ECN field is used to carry PCN markings rather thatn [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:

```
+---------------------+-----------+----------+----------+--------+
| DSCP \ RFC3168 ECN  |  not-ECT  |  ECT(0)  |  ECT(1)  |   CE   |
|     codepoint       |   (00)    |   (10)   |   (01)   |  (11)  |
+---------------------+-----------+----------+----------+--------+
|       DSCP n        |  not-PCN  |    NM    |   EXP    |   PM   |
+---------------------+-----------+----------+----------+--------+
```

   Where DSCP n is a PCN-enabled 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. That is a DiffServ codepoint that indicates that PCN
      could be enabled by setting the appropriate value in the ECN
      field. 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-enabled (not-PCN) but which shares the
      same DiffServ codepoint as PCN-enabled traffic MUST have the ECN
      field set 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]. 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-Compatible 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-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]). The scheduling behaviour is discussed in [pcn-marking-behaviour].

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 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.

6. IANA Considerations

This document makes no request to IANA.

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 (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 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, "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-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.

   [pcn-marking-behaviour]
              Eardley, P., "Marking behaviour of PCN-nodes",
              draft-eardley-pcn-marking-behaviour-01 (work in progress),
              June 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 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 Behvaiours

Any packet that belongs to a PCN capable flow MUST have the ECN field
set to indicate a NM state at the PCN-ingress-node.

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

Any packet leaving the PCN-domain SHOULD have the ECN field reset to
00. The only exception is where the egress node knows the end-hosts
will react safely to any PCN marks they receive.

B.1. Valid and Invalid Encoding Transitions at a PCN Node

o PCN-interior-nodes MUST NOT change not-PCN to another codepoint
   and they SHOULD NOT change a PCN-Capable codepoint to not-PCN
   except where they need to downgrade the packet to a lower class of
   service.

o PCN-interior-nodes that are in a pre-congestion state above the
   configured level MUST set a PM codepoint as defined in Table 1 or
   in any local/experimental scheme running within the PCN-domain.

o Packets carrying the 01 ECT(1) codepoint are for local/
   experimental use only and their unexpected presence SHOULD cause
   an alarm to be raised at the management level. However, to allow
   for the possibility of misconfiguration they SHOULD be treated as
   NM packets.

o The PM codepoint MUST NOT be changed to NM.

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 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
    meaning the ingress should stop admitting new traffic.

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
    meaning the ingress shoudl start terminating appropriate flows.

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 might be used as the only PCN metering and
    marking algorithm. The level of marks will be used to determine
    when the ingress shoudl stop admitting new traffic and whether

        the ingress should terminate any flows.

   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

   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

Acknowledgment

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