Congestion and Pre-Congestion Notification
Internet-Draft
Obsoletes: 5696 (if approved)
Intended status: Standards Track
Expires: October 19, 2012

B. Briscoe
BT
T. Moncaster
Moncaster Internet Consulting
M. Menth
University of Tuebingen

April 17, 2012

Encoding 3 PCN-States in the IP header using a single DSCP
draft-ietf-pcn-3-in-1-encoding-11

Abstract

The objective of Pre-Congestion Notification (PCN) is to protect the quality of service (QoS) of inelastic flows within a Diffserv domain. 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. Egress nodes pass information about these PCN-marks to decision points which then decide whether to admit or block new flow requests or to terminate some already-admitted flows during serious pre-congestion.

This document specifies how PCN-marks are to be encoded into the IP header by re-using the Explicit Congestion Notification (ECN) codepoints within a PCN-domain. The PCN wire protocol for non-IP protocol headers will need to be defined elsewhere. Nonetheless, this document clarifies the PCN encoding for MPLS in an informational Appendix. The encoding for IP provides for up to three different PCN marking states using a single DSCP: Not-marked (NM), Threshold-marked (ThM) and Excess-traffic-marked (ETM). Hence, it is called the 3-in-1 PCN encoding. This document obsoletes RFC5696.

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 October 19, 2012.

Copyright Notice

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. Two mechanisms are used: admission control, to decide whether to admit or block a new flow request, and flow termination to terminate some existing flows during serious pre-congestion. To achieve this, the overall rate of PCN-traffic is metered on every link in the 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 to boundary nodes about overloads before any real congestion occurs (hence "pre-congestion notification").

[RFC5670] provides for two metering and marking functions that are generally configured with different reference rates. Threshold-marking marks all PCN packets once their traffic rate on a link exceeds the configured reference rate (PCN-threshold-rate). Excess-traffic-marking marks only those PCN packets that exceed the configured reference rate (PCN-excess-rate). The PCN-excess-rate is typically larger than the PCN-threshold-rate [RFC5559]. Egress nodes monitor the PCN-marks of received PCN-packets and pass information about these PCN-marks to decision points which then decide whether to admit new flows or terminate existing flows [I-D.ietf-pcn-cl-edge-behaviour], [I-D.ietf-pcn-sm-edge-behaviour].

The encoding defined in [RFC5696] described how two PCN marking states (Not-marked and PCN-Marked) could be encoded into the IP header using a single Diffserv codepoint. It defined 01 as an experimental codepoint (EXP), along with guidelines for its use. Two PCN marking states are sufficient for the Single Marking edge behaviour [I-D.ietf-pcn-sm-edge-behaviour]. However, PCN-domains utilising the controlled load edge behaviour [I-D.ietf-pcn-cl-edge-behaviour] require three PCN marking states. This document extends the RFC5696 encoding by redefining the experimental codepoint as a third PCN marking state in the IP header, but still using a single Diffserv codepoint. This encoding scheme is therefore called the "3-in-1 PCN encoding". It obsoletes the [RFC5696] encoding, which provides only a sub-set of the same capabilities.

The full version of the 3-in-1 encoding requires any tunnel endpoint within the PCN-domain to support the normal tunnelling rules defined in [RFC6040]. There is one limited exception to this constraint where the PCN-domain only uses the excess-traffic-marking behaviour and where the threshold-marking behaviour is deactivated. This is discussed in Section 5.2.3.1.

This document only concerns the PCN wire protocol encoding for IP headers, whether IPv4 or IPv6. It makes no changes or recommendations concerning algorithms for congestion marking or congestion response. Other documents will define the PCN wire protocol for other header types. Appendix C discusses a possible mapping between IP and MPLS.

1.1. Requirements Language

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].

1.2. Changes in This Version (to be removed by RFC Editor)

From draft-ietf-pcn-3-in-1-encoding-10 to -11:

* Pointed out that any DSCP re-mapping must precede PCN-ingress processing;
* Ingress behaviour for ECN-capable PCN-packets: allowed any PCN-capable encapsulation, not just IP-in-IP tunnelling. Added cautionary note about MPLS PHP;
* PCN-policing at ingress:
  + Clarified what per-flow policing entails;
  + Clarified that a DSCP of zero is ’000000’;
  + For policed packets, added SHOULD log, MAY alarm;
* Updated refs and acks.

From draft-ietf-pcn-3-in-1-encoding-09 to -10:

* Added cautionary management advice to S6.2 (backwards compatibility with RFC5696)
* Removed "emphatically" from normative "NOT RECOMMENDED" text.
* Minor editoral changes.

From draft-ietf-pcn-3-in-1-encoding-08 to -09:

* Added note about fail-safe to protect other traffic in the event of tunnel misconfiguration.

* Changed section heading to be about applicability of environments to the encoding, rather than the encoding to the environments.

* Completely re-wrote PCN-ingress Node Behaviour section.

* Changed PCN interior node to PCN-node where the term was intended to include all PCN-nodes.

* Clarified status of ECN/PCN co-existence appendix. Removed inconsistent assertion in this appendix that an admission-control DSCP alone can indicate that arriving traffic is PCN-traffic.

* A few clarifying editorial amendments and updated refs.

From draft-ietf-pcn-3-in-1-encoding-07 to -08: Editorial corrections and clarifications.

From draft-ietf-pcn-3-in-1-encoding-06 to -07:

* Clarified that each operator not the IETF chooses which DSCP(s) are PCN-compatible, and made it unambiguous that only PCN-nodes recognise that PCN-compatible DSCPs enable the 3-in-1 encoding.

* Removed statements about the PCN working group, given RFCs are meant to survive beyond the life of a w-g.

* Corrected the final para of "Rationale for Different Behaviours in Schemes with Only One Marking"

From draft-ietf-pcn-3-in-1-encoding-05 to -06:

* Draft re-written to obsolete baseline encoding [RFC5696].

* New section defining utilising this encoding for only one PCN-Marking. Added an appendix explaining an apparent inconsistency within this section.

* Moved (and updated) informative appendixes from [RFC5696] to this document. Original Appendix C was omitted as it is now redundant.

* Significant re-structuring of document.

From draft-ietf-pcn-3-in-1-encoding-04 to -05:

* Draft moved to standards track as per working group discussions.

* Added Appendix B discussing ECN handling in the PCN-domain.

* Clarified that this document modifies [RFC5696].

From draft-ietf-pcn-3-in-1-encoding-03 to -04:

* Updated document to reflect RFC6040.

* Re-wrote introduction.

* Re-wrote section on applicability.

* Re-wrote section on choosing encoding scheme.

* Updated author details.

From draft-ietf-pcn-3-in-1-encoding-02 to -03:

* Corrected mistakes in introduction and improved overall readability.

* Added new terminology.

* Rewrote a good part of Section 4 and 5 to achieve more clarity.

* Added appendix explaining when to use which encoding scheme and how to encode them in MPLS shim headers.

* Added new co-author.

From draft-ietf-pcn-3-in-1-encoding-01 to -02:

* Corrected mistake in introduction, which wrongly stated that the threshold-traffic rate is higher than the excess-traffic rate. Other minor corrections.

* Updated acks & refs.

From draft-ietf-pcn-3-in-1-encoding-00 to -01:

* Altered the wording to make sense if draft-ietf-tsvwg-ecn-tunnel moves to proposed standard.

* References updated

From draft-briscoe-pcn-3-in-1-encoding-00 to draft-ietf-pcn-3-in-1-encoding-00:

* Filename changed to draft-ietf-pcn-3-in-1-encoding.

* Introduction altered to include new template description of PCN.

* References updated.

* Terminology brought into line with [RFC5670].

* Minor corrections.

2. Definitions and Abbreviations

2.1. Terminology

The terms 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 encoding: mapping of PCN marking states to specific codepoints in the packet header.

PCN-compatible Diffserv codepoint: a Diffserv codepoint indicating packets for which the ECN field carries PCN-markings rather than [RFC3168] markings. Note that an operator configures PCN-nodes to recognise PCN-compatible DSCPs, whereas the same DSCP has no PCN-specific meaning to a node outside the PCN domain.

Threshold-marked codepoint: a codepoint that indicates a packet has been threshold-marked; that is a packet that has been marked at a PCN-interior-node as a result of an indication from the threshold-metering function [RFC5670]. Abbreviated to ThM.

Excess-traffic-marked codepoint: a codepoint that indicates packets that have been marked at a PCN-interior-node as a result of an indication from the excess-traffic-metering function [RFC5670]. Abbreviated to ETM.

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

Not-PCN codepoint: a codepoint that indicates packets that are not PCN-packets.

2.2. 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
- e2e = end-to-end
- ECN = Explicit Congestion Notification [RFC3168]
- ECT = ECN Capable Transport [RFC3168]
- EF = Expedited Forwarding [RFC3246]
- ETM = Excess-traffic-marked
- EXP = Experimental
- NM = Not-marked
- PCN = Pre-Congestion Notification
- PHB = Per-hop behaviour [RFC2474]
- ThM = Threshold-marked

3. Definition of 3-in-1 PCN Encoding

The 3-in-1 PCN encoding scheme supports networks that need three PCN-marking states to be encoded within the IP header, as well as those that need only two. The full encoding is shown in Figure 1.

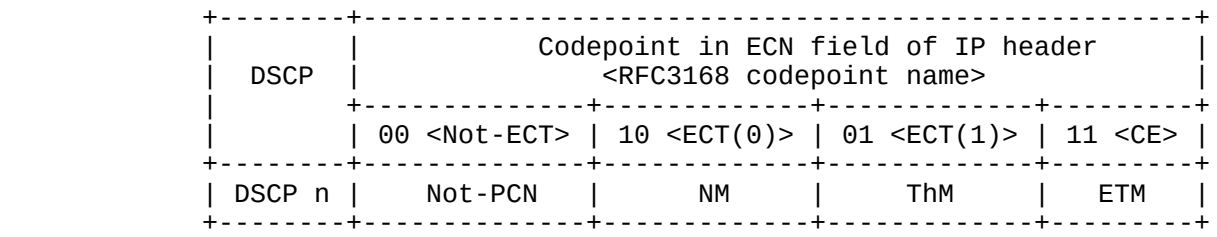


Figure 1: 3-in-1 PCN Encoding

A PCN-node will be configured to recognise certain DSCPs as PCN-compatible. Appendix A discusses the choice of suitable DSCPs. In Figure 1 'DSCP n' indicates such a PCN-compatible DSCP. In the PCN-domain, any packet carrying a PCN-compatible DSCP and with the ECN-field anything other than 00 (Not-PCN) is a PCN-packet as defined in [RFC5559].

PCN-nodes MUST interpret the ECN field of a PCN-packet using the 3-in-1 PCN encoding, rather than [RFC3168]. This does not change the behaviour for any packet with a DSCP that is not PCN-compatible, or for any node outside a PCN-domain. In all such cases the 3-in-1 encoding is not applicable and so by default the node will interpret the ECN field using [RFC3168].

When using the 3-in-1 encoding, the codepoints of the ECN field have the following meanings:

Not-PCN: indicates a non-PCN-packet, i.e., a packet that uses a PCN-compatible DSCP but is not subject to PCN metering and marking.

NM: Not-marked. Indicates a PCN-packet that has not yet been marked by any PCN marker.

ThM: Threshold-marked. Indicates a PCN-packet that has been marked by a threshold-marker [RFC5670].

ETM: Excess-traffic-marked. Indicates a PCN-packet that has been marked by an excess-traffic-marker [RFC5670].

4. Requirements for and Applicability of 3-in-1 PCN Encoding

4.1. PCN Requirements

In accordance with the PCN architecture [RFC5559], PCN-ingress-nodes control packets entering a PCN-domain. Packets belonging to PCN-controlled flows are subject to PCN-metering and -marking, and PCN-ingress-nodes mark them as Not-marked (PCN-colouring). All nodes in

the PCN-domain perform PCN-metering and PCN-mark PCN-packets if needed. There are two different metering and marking behaviours: threshold-marking and excess-traffic-marking [RFC5670]. Some edge behaviours require only a single marking behaviour [I-D.ietf-pcn-sm-edge-behaviour], others require both [I-D.ietf-pcn-cl-edge-behaviour]. In the latter case, three PCN marking states are needed: Not-marked (NM) to indicate not-marked packets, Threshold-marked (ThM) to indicate packets marked by the threshold-marker, and Excess-traffic-marked (ETM) to indicate packets marked by the excess-traffic-marker [RFC5670]. Threshold-marking and excess-traffic-marking are configured to start marking packets at different load conditions, so one marking behaviour indicates more severe pre-congestion than the other. Therefore, a fourth PCN marking state indicating that a packet is marked by both markers is not needed. However a fourth codepoint is required to indicate packets that use a PCN-compatible DSCP but do not use PCN-marking (the Not-PCN codepoint).

In all current PCN edge behaviours that use two marking behaviours [RFC5559], [I-D.ietf-pcn-cl-edge-behaviour], excess-traffic-marking is configured with a larger reference rate than threshold-marking. We take this as a rule and define excess-traffic-marked as a more severe PCN-mark than Threshold-marked.

4.2. Requirements Imposed by Tunnelling

[RFC6040] defines rules for the encapsulation and decapsulation of ECN markings within IP-in-IP tunnels. The publication of RFC6040 removed the tunnelling constraints that existed when the encoding of [RFC5696] was written (see Section 3.3.2 of [I-D.ietf-pcn-encoding-comparison]).

Nonetheless, there is still a problem if there are any legacy (pre-RFC6040) decapsulating tunnel endpoints within a PCN domain. If a PCN-node Threshold-marks the outer header of a tunnelled packet that has a Not-marked codepoint on the inner header, a legacy decapsulator will forward the packet as Not-marked, losing the Threshold-marking. The rules on applicability in Section 4.3 below are designed to avoid this problem.

Even if an operator accidentally breaks these applicability rules, the order of severity of the 3-in-1 codepoints was chosen to protect other PCN or non-PCN traffic. Although legacy pre-RFC6040 tunnels did not propagate ’01’, all tunnels pre-RFC6040 and post-RFC6040 have always propagated ’11’ correctly. Therefore ’11’ was chosen to signal the most severe pre-congestion (ETM), so it would act as a reliable fail-safe even if an overlooked legacy tunnel was suppressing 01 (ThM) signals.

4.3. Applicable Environments for the 3-in-1 PCN Encoding

The 3-in-1 encoding is applicable in situations where two marking behaviours are being used in the PCN-domain. The 3-in-1 encoding can also be used with only one marking behaviour, in which case one of the codepoints MUST NOT be used anywhere in the PCN-domain (see Section 5.2.3).

With one exception (see next paragraph), any tunnel endpoints (IP-in-IP and IPsec) within the PCN-domain MUST comply with the ECN encapsulation and decapsulation rules set out in [RFC6040] (see Section 4.2).

Operators may not be able to upgrade every pre-RFC6040 tunnel endpoint within a PCN-domain. In such circumstances a limited version of the 3-in-1 encoding can still be used but only under the following stringent condition. If any pre-RFC6040 tunnel decapsulator exists within a PCN-domain then every PCN-node in the PCN-domain MUST be configured so that it never sets the ThM codepoint. PCN-interior-nodes in this case MUST solely use the Excess-traffic-marking function, as defined in Section 5.2.3.1. In all other situations where legacy tunnel decapsulators might be present within the PCN domain, the 3-in-1 encoding is not applicable.

5. Behaviour of a PCN-node to Comply with the 3-in-1 PCN Encoding

Any tunnel endpoint implementation on a PCN-node MUST comply with [RFC6040]. Since PCN is a new capability, this is considered a reasonable requirement.

5.1. PCN-ingress Node Behaviour

If packets arrive from another Diffserv domain, any re-mapping of Diffserv codepoints MUST happen before PCN-ingress processing.

At each logical ingress link into a PCN domain, each PCN-ingress node will apply the four functions described in Section 4.2 of [RFC5559] to arriving packets. These functions are applied in the following order: PCN-classify, PCN-police, PCN-colour, PCN-rate-meter. This section describes these four steps, but only the aspects relevant to packet encoding:

1. PCN-classification: The PCN-ingress-node determines whether each packet matches the filter spec of an admitted flow. Packets that match are defined as PCN-packets.

1b. Extra step if ECN and PCN co-exist: If a packet classified as a PCN-packet arrives with the ECN field already set to a value other than Not-ECT (i.e. it is from an ECN-capable transport) then to comply with BCP 124 [RFC4774] it MUST pass through one of the following preparatory steps before the PCN-policing and PCN-colouring steps. The choice between these four actions depends on local policy:

* Encapsulate ECN-capable PCN-packets across the PCN-domain:

  + either within another IP header using an RFC6040 tunnel;

  + or within a lower layer protocol capable of being PCN marked, such as MPLS (see Appendix C).

  Encapsulation using either of these methods is the RECOMMENDED policy for ECN-capable PCN-packets, and implementations SHOULD use IP-in-IP tunnelling as the default.

  If encapsulation is used, it MUST precede PCN-policing and PCN-colouring so that the encapsulator and decapsulator are logically outside the PCN domain (see Appendix B and specifically Figure 2).

  If MPLS encapsulation is used, note that penultimate hop popping [RFC3031] is incompatible with PCN, unless the penultimate hop applies the PCN-egress node behaviour before it pops the PCN-capable MPLS label.

* If some form of encapsulation is not possible, the PCN-ingress-node can allow through ECN-capable packets without encapsulation, but it MUST drop CE-marked packets at this stage. Failure to drop CE would risk congestion collapse, because without encapsulation there is no mechanism to propagate the CE markings across the PCN-domain (see Appendix B).

  This policy is NOT RECOMMENDED because there is no tunnel to protect the e2e ECN capability, which is otherwise disabled when the PCN-egress-node zeroes the ECN field.

* Drop the packet.

  This policy is also NOT RECOMMENDED, because it precludes the possibility that e2e ECN can co-exist with PCN as a means of controlling congestion.

* Any other action that complies with [RFC4774] (see Appendix B for an example).

Appendix B provides more information about the co-existence of PCN and ECN.

2. PCN-Policing: The PCN-policing function only allows appropriate packets into the PCN behaviour aggregate. Per-flow policing actions may be required to block rejected flows and to rate-police accepted flows, but these are specified in the relevant edge-behaviour document, e.g. [I-D.ietf-pcn-sm-edge-behaviour], [I-D.ietf-pcn-cl-edge-behaviour].

   Here we only specify packet-level PCN-policing, which prevents packets that are not PCN-packets from being forwarded into the PCN-domain if PCN-interior-nodes would otherwise mistake them for PCN-packets. A non-PCN-packet will be confused with a PCN-packet if on arrival it meets all three of the following conditions:

   a) it is not classified as a PCN-packet

   b) it already carries a PCN-compatible DSCP

   c) its ECN field carries a codepoint other than Not-ECT.

   The PCN-ingress-node MUST police packets that meet all three conditions (a-c) by subjecting them to one of the following treatments:

   * re-mark the DSCP to a DSCP that is not PCN-compatible;

   * tunnel the packet to the PCN-egress with a DSCP in the outer header that is not PCN-compatible;

   * drop the packet (NOT RECOMMENDED--see below).

   The choice between these actions depends on local policy. In the absence of any operator-specific configuration for this case, by default an implementation SHOULD re-mark the DSCP to zero (000000).

   Whichever policing action is chosen, the PCN-ingress-node SHOULD log the event and MAY also raise an alarm. Alarms SHOULD be rate-limited so that the anomalous packets will not amplify into a flood of alarm messages.

   Rationale: Traffic that meets all three of the above conditions

(a-c) is not PCN-traffic, therefore ideally a PCN-ingress ought not to interfere with it, but it has to do something to avoid ambiguous packet markings. Clearing the ECN field is not an appropriate policing action, because a network node ought not to interfere with an e2e signal. Even if such packets seem like an attack, drop would be overkill, because such an attack can be neutralised by just re-marking the DSCP. And DSCP re-marking in the network is legitimate, because the DSCP is not considered an e2e signal.

3. PCN-colouring: If a packet has been classified as a PCN-packet, once it has been policed, the PCN-ingress-node:

   * MUST set a PCN-compatible Diffserv codepoint on all PCN-packets. To conserve DSCPs, Diffserv codepoints SHOULD be chosen that are already defined for use with admission-controlled traffic. Appendix A gives guidance to implementors on suitable DSCPs.

   * MUST set the PCN codepoint of all PCN-packets to Not-marked (NM).

4. PCN rate-metering: This fourth step may be necessary depending on the edge-behaviour in force. It is listed for completeness, but it is not relevant to this encoding document.

5.2. PCN-interior Node Behaviour

5.2.1. Behaviour Common to all PCN-interior Nodes

Interior nodes MUST NOT change Not-PCN to any other codepoint.

Interior nodes MUST NOT change NM to Not-PCN.

Interior nodes MUST NOT change ThM to NM or Not-PCN.

Interior nodes MUST NOT change ETM to any other codepoint.

5.2.2. Behaviour of PCN-interior Nodes Using Two PCN-markings

If the threshold-meter function indicates a need to mark a packet, the PCN-interior-node MUST change NM to ThM.

If the excess-traffic-meter function indicates a need to mark a packet:

o the PCN-interior-node MUST change NM to ETM;

- the PCN-interior-node MUST change ThM to ETM.

If both the threshold meter and the excess-traffic meter indicate the need to mark a packet, the Excess-traffic-marking rules MUST take precedence.

5.2.3. Behaviour of PCN-interior Nodes Using One PCN-marking

Some PCN edge behaviours require only one PCN-marking within the PCN-domain. The Single Marking edge behaviour [I-D.ietf-pcn-sm-edge-behaviour] requires PCN-interior-nodes to mark packets using the excess-traffic-meter function [RFC5670]. It is possible that future schemes may require only the threshold-meter function. Appendix D explains the rationale for the behaviours defined in this section.

5.2.3.1. Marking Using only the Excess-traffic-meter Function

The threshold-traffic-meter function SHOULD be disabled and MUST NOT trigger any packet marking.

The PCN-interior-node SHOULD raise a management alarm if it receives a ThM packet, but the frequency of such alarms SHOULD be limited.

If the excess-traffic-meter function indicates a need to mark the packet:

- the PCN-interior-node MUST change NM to ETM;

- the PCN-interior-node MUST change ThM to ETM. It SHOULD also raise an alarm as above.

5.2.3.2. Marking using only the Threshold-meter Function

The excess-traffic-meter function SHOULD be disabled and MUST NOT trigger any packet marking.

The PCN-interior-node SHOULD raise a management alarm if it receives an ETM packet, but the frequency of such alarms SHOULD be limited.

If the threshold-meter function indicates a need to mark the packet:

- the PCN-interior-node MUST change NM to ThM;

- the PCN-interior-node MUST NOT change ETM to any other codepoint. It SHOULD raise an alarm as above if it encounters an ETM packet.

5.3. PCN-egress Node Behaviour

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 ECN codepoints. Appendix B gives details of how such schemes might work, but such schemes are currently only tentative ideas.

If the PCN-domain is configured to use only Excess-traffic-marking, the PCN-egress-node MUST treat ThM as ETM and, if only threshold-marking is used, it SHOULD treat ETM as ThM. However it SHOULD raise a management alarm in either case since this means there is some misconfiguration in the PCN-domain.

6. Backward Compatibility

6.1. Backward Compatibility with ECN

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 encoding specified in this document uses one of those techniques; it defines PCN-compatible Diffserv codepoints as no longer supporting the default ECN semantics within a PCN domain. As such, this document is compatible with BCP 124.

There is not enough space in one IP header for the 3-in-1 encoding to support both ECN marking end-to-end and PCN-marking within a PCN-domain. The non-normative Appendix B discusses possible ways to do this, e.g. by carrying e2e ECN across a PCN-domain within the inner header of an IP-in-IP tunnel. The normative text in Section 5.1 requires one of these methods to be configured at the PCN-ingress-node and recommends that implementations offer tunnelling as the default.

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 codepoint. PCN-egress-nodes then guarantee that the ECN field of any packet leaving the PCN-domain has appropriate ECN semantics. This prevents unintended leakage of ECN marks into or out of the PCN-domain, and

thus reduces backward-compatibility issues.

6.2. Backward Compatibility with the RFC5696 Encoding

Section 5.1 of the PCN architecture gives general guidance on fault detection and diagnosis [RFC5559], including management analysis of PCN markings arriving at PCN-egress nodes to detect early signs of potential faults. Because the PCN encoding has gone through an obsoleted earlier stage [RFC5696], misconfiguration mistakes may be more likely. Therefore extra monitoring, such as in the following example, may be necessary to detect and diagnose potential problems:

> Informational example: In a controlled-load edge-behaviour scenario it could be worth the PCN-egress node detecting the onset of excess-traffic marking without any prior threshold-marking. This might indicate that an interior node has been wrongly configured to mark only ETM (which would have been correct for the single-marking edge-behaviour).

A PCN-node implemented to use the obsoleted RFC5696 encoding could conceivably have been configured so that the Threshold-meter function marked what is now defined as the ETM codepoint in the 3-in-1 encoding. However, there is no known deployment of this rather unlikely variant of RFC5696 and no reason to believe that such an implementation would ever have been built. Therefore, it seems safe to ignore this issue.

7. IANA Considerations

This memo includes no request to IANA.

Note to RFC Editor: this section may be removed on publication as an RFC.

8. 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 behaviours 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.

9. Conclusions

The 3-in-1 PCN encoding uses a PCN-compatible DSCP and the ECN field to encode PCN-marks. One codepoint allows non-PCN traffic to be carried with the same PCN-compatible DSCP and three other codepoints support three PCN marking states with different levels of severity. In general, the use of this PCN encoding scheme presupposes that any tunnel endpoints within the PCN-domain comply with [RFC6040].

10. Acknowledgements

Many thanks to Philip Eardley for providing extensive feedback, criticism and advice. Thanks also to Teco Boot, Kwok Ho Chan, Ruediger Geib, Georgios Karagiannis, James Polk, Tom Taylor, Adrian Farrel and everyone else who has commented on the document.

11. Comments Solicited

To be removed by RFC Editor: 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.

12. References

12.1. Normative References

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

[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.

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

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

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

12.2. Informative References

[I-D.ietf-pcn-cl-edge-behaviour] Charny, A., Huang, F., Karagiannis, G., Menth, M., and T. Taylor, "PCN Boundary Node Behaviour for the Controlled Load (CL) Mode of Operation", draft-ietf-pcn-cl-edge-behaviour-14 (work in progress), April 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-sm-edge-behaviour] Charny, A., Karagiannis, G., Menth, M., and T. Taylor, "PCN Boundary Node Behaviour for the Single Marking (SM) Mode of Operation", draft-ietf-pcn-sm-edge-behaviour-12 (work in progress), April 2012.

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

[RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.

[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.

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

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

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

[RFC5129] Davie, B., Briscoe, B., and J. Tay, "Explicit Congestion Marking in MPLS", RFC 5129, January 2008.

[RFC5462] Andersson, L. and R. Asati, "Multiprotocol Label Switching (MPLS) Label Stack Entry: "EXP" Field Renamed to "Traffic Class" Field", RFC 5462, February 2009.

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

[RFC5865] Baker, F., Polk, J., and M. Dolly, "A Differentiated Services Code Point (DSCP) for Capacity-Admitted Traffic", RFC 5865, May 2010.

Appendix A. Choice of Suitable DSCPs

This appendix is informative, not normative.

A single DSCP has not been defined 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]. It is suggested that admission control could be used for the following service classes (defined in [RFC4594] unless otherwise stated):

- Telephony (EF)
- Real-time interactive (CS4)
- Broadcast Video (CS3)
- Multimedia Conferencing (AF4)
- the VOICE-ADMIT codepoint defined in [RFC5865].

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].

Guidelines for conserving DSCPs by allowing non-admission-controlled-traffic to compete with PCN-traffic are given in Appendix B.1 of [RFC5670].

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.

Appendix B. Co-existence of ECN and PCN

This appendix is informative, not normative. It collects together material relevant to co-existence of ECN and PCN, including that spread throughout the body of this specification. If this results in any conflict or ambiguity, the normative text in the body of the specification takes precedence.

ECN [RFC3168] is an e2e congestion notification mechanism. As such it is possible that some traffic entering the PCN-domain may also be ECN-capable. The PCN encoding described in this document re-uses the bits of the ECN field in the IP header. Consequently, this disables ECN within the PCN domain.

For the purposes of this appendix we define two forms of traffic that might arrive at a PCN-ingress-node. These are admission-controlled traffic (PCN-traffic) and non-admission-controlled traffic (non-PCN-traffic).

Flow signalling identifies admission controlled traffic, by associating a filterspec with the need for admission control (e.g. through RSVP or some equivalent message, e.g. from a SIP server to the ingress or from a logically centralised network control system). The PCN-ingress-node re-marks admission-conrolled traffic matching that filterspec to a PCN-compatible DSCP. Note that the term flow need not imply just one microflow, but instead could match an aggregate and/or could depend on the incoming DSCP (see Appendix A).

All other traffic can be thought of as non-admission-controlled (and therefore outside the scope of PCN). However such traffic may still need to share the same DSCP as the admission-controlled traffic. This may be due to policy (for instance if it is high priority voice traffic), or may be because there is a shortage of local DSCPs.

Unless specified otherwise, for any of the cases in the list below, an IP-in-IP tunnel that complies with[RFC6040] can be used to preserve ECN markings across the PCN domain. The tunnelling action should be applied wholly outside the PCN-domain as illustrated in Figure 2. Then, by the rules of RFC6040, the tunnel egress propagates the ECN field from the inner header, because the PCN-egress will have zeroed the outer ECN field.

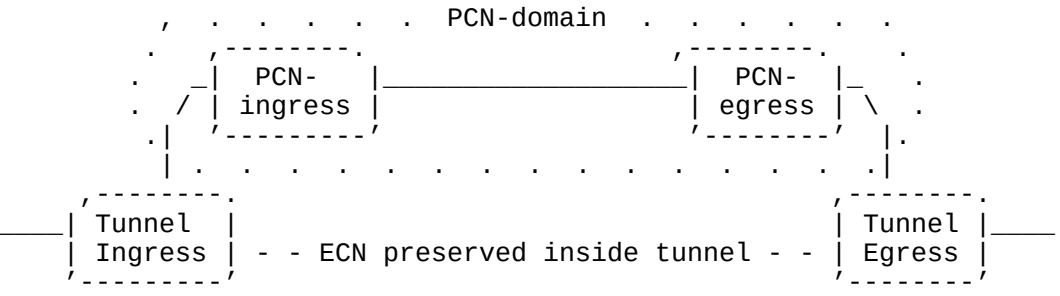


Figure 2: Separation of tunnelling and PCN actions

There are three cases for how e2e ECN traffic may wish to be treated while crossing a PCN domain:

a) Traffic that does not require PCN admission control:
   For example, traffic that does not match flow signaling being used for admission control. In this case the e2e ECN traffic is not treated as PCN-traffic. There are two possible scenarios:

   * Arriving traffic does not carry a PCN-compatible DSCP: no action required.

   * Arriving traffic carries a DSCP that clashes with a PCN-compatible DSCP. There are two options:

      1. The ingress maps the DSCP to a local DSCP with the same scheduling PHB as the original DSCP, and the egress re-maps it to the original PCN-compatible DSCP.

      2. The ingress tunnels the traffic, setting the DSCP in the outer header to a local DSCP with the same scheduling PHB as the original DSCP.

3. The ingress tunnels the traffic, using the original DSCP in the outer but setting Not-PCN in the outer header; note that this turns off ECN for this traffic within the PCN domain.

The first or second options are recommended unless the operator is short of local DSCPs.

b) Traffic that requires admission-control:
In this case the e2e ECN traffic is treated as PCN-traffic across the PCN domain. There are two options.

* The PCN-ingress-node places this traffic in a tunnel with a PCN-compatible DSCP in the outer header. The PCN-egress zeroes the ECN-field before decapsulation.

* The PCN-ingress-node drops CE-marked packets and otherwise sets the ECN-field to NM and sets the DCSP to a PCN-compatible DSCP. The PCN-egress zeroes the ECN field of all PCN packets; note that this turns off e2e ECN.

The second option is emphatically not recommended, unless perhaps as a last resort if tunnelling is not possible for some insurmountable reason.

c) Traffic that requires PCN admission control where the endpoints ask to see PCN marks:
Note that this scheme is currently only a tentative idea.

For real-time data generated by an adaptive codec, schemes have been suggested where PCN marks may be leaked out of the PCN-domain so that end hosts can drop to a lower data-rate, thus deferring the need for admission control. Currently such schemes require further study and the following is for guidance only.

The PCN-ingress-node needs to tunnel the traffic as in Figure 2, taking care to comply with [RFC6040]. In this case the PCN-egress does not zero the ECN field contrary to the recommendation in Section 5.3, so that the [RFC6040] tunnel egress will preserve any PCN-marking. Note that a PCN-interior-node may change the ECN-field from 10 to 01 (NM to ThM), which would be interpreted by the e2e ECN as a change from ECT(0) into ECT(1). This would not be compatible with the (currently experimental) ECN nonce [RFC3540].

Appendix C. Example Mapping between Encoding of PCN-Marks in IP and in MPLS Shim Headers

This appendix is informative not normative.

The 6 bits of the DS field in the IP header provide for 64 codepoints. When encapsulating IP traffic in MPLS, it is useful to make the DS field information accessible in the MPLS header. However, the MPLS shim header has only a 3-bit traffic class (TC) field [RFC5462] providing for 8 codepoints. The operator has the freedom to define a site-local mapping of the 64 codepoints of the DS field onto the 8 codepoints in the TC field.

[RFC5129] describes how ECN markings in the IP header can also be mapped to codepoints in the MPLS TC field. Appendix A of [RFC5129] gives an informative description of how to support PCN in MPLS by extending the way MPLS supports ECN. [RFC5129] was written while PCN specifications were in early draft stages. The following provides a clearer example of a mapping between PCN in IP and in MPLS using the PCN terminology and concepts that have since been specified.

To support PCN in a MPLS domain, a PCN-compatible DSCP ('DSCP n') needs codepoints to be provided in the TC field for all the PCN-marks used. That means, when for instance only excess-traffic-marking is used for PCN purposes, the operator needs to define a site-local mapping to two codepoints in the MPLS TC field for IP headers with:

- DSCP n and NM
- DSCP n and ETM

If both excess-traffic-marking and threshold-marking are used, the operator needs to define a site-local mapping to codepoints in the MPLS TC field for IP headers with all three of the 3-in-1 codepoints:

- DSCP n and NM
- DSCP n and ThM
- DSCP n and ETM

In either case, if the operator wishes to support the same Diffserv PHB but without PCN marking, it will also be necessary to define a site-local mapping to an MPLS TC codepoint for IP headers marked with:

- DSCP n and Not-PCN

The above sets of codepoints are required for every PCN-compatible DSCP. Clearly, given so few TC codepoints are available, it may be necessary to compromise by merging together some capabilities. Guidelines for conserving TC codepoints by allowing non-admission-controlled-traffic to compete with PCN-traffic are given in Appendix B.1 of [RFC5670].

Appendix D. Rationale for Difference Between the Schemes using One PCN-Marking

Readers may notice a difference between the two behaviours in Section 5.2.3.1 and Section 5.2.3.2. With only Excess-traffic-marking enabled, an unexpected ThM packet can be re-marked to ETM. However, with only Threshold-marking, an unexpected ETM packet cannot be re-marked to ThM.

This apparent inconsistency is deliberate. The behaviour with only Threshold-marking keeps to the rule of Section 5.2.1 that ETM is more severe and must never be changed to ThM even though ETM is not a valid marking in this case. Otherwise implementations would have to allow operators to configure an exception to this rule, which would not be safe practice and would require different code in the data-plane compared to the common behaviour.

Authors' Addresses

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

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

Toby Moncaster
Moncaster Internet Consulting
Dukes
Layer Marney
Colchester CO5 9UZ
UK

Phone: +44 7764 185416
EMail: toby@moncaster.com
URI: http://www.moncaster.com/

Michael Menth
University of Tuebingen
Sand 13
Tuebingen 72076
Germany

Phone: +49 7071 2970505
EMail: menth@informatik.uni-tuebingen.de

Internet Engineering Task Force
Internet-Draft
Intended status: Experimental
Expires: November 12, 2012

A. Charny
F. Huang
Huawei Technologies
G. Karagiannis
U. Twente
M. Menth
University of Tuebingen
T. Taylor, Ed.
Huawei Technologies
May 11, 2012

PCN Boundary Node Behaviour for the Controlled Load (CL) Mode of Operation
draft-ietf-pcn-cl-edge-behaviour-15

Abstract

Pre-congestion notification (PCN) is a means for protecting the quality of service for inelastic traffic admitted to a Diffserv domain. The overall PCN architecture is described in RFC 5559. This memo is one of a series describing possible boundary node behaviours for a PCN-domain. The behaviour described here is that for a form of measurement-based load control using three PCN marking states, not-marked, threshold-marked, and excess-traffic-marked. This behaviour is known informally as the Controlled Load (CL) PCN-boundary-node behaviour.

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 November 12, 2012.

Copyright Notice

Table of Contents

1. Introduction

The objective of Pre-Congestion Notification (PCN) is to protect the quality of service (QoS) of inelastic flows within a Diffserv domain, in a simple, scalable, and robust fashion. Two mechanisms are used: admission control, to decide whether to admit or block a new flow request, and (in abnormal circumstances) flow termination to decide whether to terminate some of the existing flows. To achieve this, 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 to PCN-boundary-nodes about incipient overloads before any congestion occurs (hence the "pre" part of "pre-congestion notification"). The level of marking allows decisions to be made about whether to admit or terminate PCN-flows. For more details see [RFC5559].

This document describes an experimental edge node behaviour to implement PCN in a network. The experiment may be run in a network in which a substantial proportion of the traffic carried is in the form of inelastic flows and where admission control of micro-flows is applied at the edge. For the effects of PCN to be observable, the committed bandwidth (i.e., level of non-best-effort traffic) on at least some links of the network should be near or at link capacity. The amount of effort required to prepare the network for the experiment (see Section 5.1) may constrain the size of network to which it is applied. The purposes of the experiment are:

- to validate the specification of the CL edge behaviour;
- to evaluate the effectiveness of the CL edge behaviour in preserving quality of service for admitted flows; and
- to evaluate PCN's potential for reducing the amount of capital and operational costs in comparison to alternative methods of assuring quality of service.

For the first two objectives, the experiment should run long enough for the network to experience sharp peaks of traffic in at least some directions. It would also be desirable to observe PCN performance in the face of failures in the network. A period in the order of a month or two in busy season may be enough. The third objective is more difficult, and could require observation over a period long enough for traffic demand to grow to the point where additional capacity must be provisioned at some points in the network.

Section 3 of this document specifies a detailed set of algorithms and procedures used to implement the PCN mechanisms for the CL mode of

operation. Since the algorithms depend on specific metering and marking behaviour at the interior nodes, it is also necessary to specify the assumptions made about PCN-interior-node behaviour (Section 2). Finally, because PCN uses DSCP values to carry its markings, a specification of PCN-boundary-node behaviour must include the per domain behaviour (PDB) template specified in [RFC3086], filled out with the appropriate content (Section 4).

Note that the terms "block" or "terminate" actually translate to one or more of several possible courses of action, as discussed in Section 3.6 of [RFC5559]. The choice of which action to take for blocked or terminated flows is a matter of local policy.

[RFC EDITOR'S NOTE: RFCyyyy is the published version of draft-ietf-pcn-sm-edge-behaviour.]

A companion document [RFCyyyy] specifies the Single Marking (SM) PCN-boundary-node behaviour. This document and [RFCyyyy] have a great deal of text in common. To simplify the task of the reader, the text in the present document that is specific to the CL PCN-boundary-node behaviour is preceded by the phrase: "[CL-specific]". A similar distinction for SM-specific text is made in [RFCyyyy].

1.1. Terminology

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].

This document uses the following terms defined in Section 2 of [RFC5559]:

- PCN-domain;
- PCN-ingress-node;
- PCN-egress-node;
- PCN-interior-node;
- PCN-boundary-node;
- PCN-flow;
- ingress-egress-aggregate (IEA);
- [CL-specific] PCN-threshold-rate;

- PCN-excess-rate;
- PCN-admissible-rate;
- PCN-supportable-rate;
- PCN-marked;
- [CL-specific] threshold-marked;
- excess-traffic-marked.

It also uses the terms PCN-traffic and PCN-packet, for which the definition is repeated from [RFC5559] because of their importance to the understanding of the text that follows:

PCN-traffic, PCN-packets, PCN-BA
: A PCN-domain carries traffic of different Diffserv behaviour aggregates (BAs) [RFC2474]. The PCN-BA uses the PCN mechanisms to carry PCN-traffic, and the corresponding packets are PCN-packets. The same network will carry traffic of other Diffserv BAs. The PCN-BA is distinguished by a combination of the Diffserv codepoint and the ECN field.

This document uses the following terms from [RFC5670]:

- [CL-specific] threshold-meter;
- excess-traffic-meter.

To complete the list of borrowed terms, this document reuses the following terms and abbreviations defined in Section 3 of [ID.pcn-3-in-1]:

- not-PCN codepoint;
- Not-marked (NM) codepoint;
- [CL-specific] Threshold-marked (ThM) codepoint;
- Excess-traffic-marked (ETM) codepoint.

This document defines the following additional terms:

Decision Point
: The node that makes the decision about which flows to admit and to terminate. In a given network deployment, this can be the PCN-ingress-node or a centralized control node. In either case, the

PCN-ingress-node is the point where the decisions are enforced.

NM-rate
: The rate of not-marked PCN-traffic received at a PCN-egress-node for a given ingress-egress-aggregate in octets per second. For further details see Section 3.2.1.

[CL-specific] ThM-rate
: The rate of threshold-marked PCN-traffic received at a PCN-egress-node for a given ingress-egress-aggregate in octets per second. For further details see Section 3.2.1.

ETM-rate
: The rate of excess-traffic-marked PCN-traffic received at a PCN-egress-node for a given ingress-egress-aggregate in octets per second. For further details see Section 3.2.1.

PCN-sent-rate
: The rate of PCN-traffic received at a PCN-ingress-node and destined for a given ingress-egress-aggregate in octets per second. For further details see Section 3.4.

Congestion level estimate (CLE)
: The ratio of PCN-marked to total PCN-traffic (measured in octets) received for a given ingress-egress-aggregate during a given measurement period. The CLE is used to derive the PCN-admission-state (Section 3.3.1) and is also used by the report suppression procedure (Section 3.2.3) if report suppression is activated.

PCN-admission-state
: The state ("admit" or "block") derived by the Decision Point for a given ingress-egress-aggregate based on PCN packet marking statistics. The Decision Point decides to admit or block new flows offered to the aggregate based on the current value of the PCN-admission-state. For further details see Section 3.3.1.

Sustainable aggregate rate (SAR)
: The estimated maximum rate of PCN-traffic that can be carried in a given ingress-egress-aggregate at a given moment without risking degradation of quality of service for the admitted flows. The intention is that if the PCN-sent-rate of every ingress-egress-aggregate passing through a given link is limited to its sustainable aggregate rate, the total rate of PCN-traffic flowing through the link will be limited to the PCN-supportable-rate for that link. An estimate of the sustainable aggregate rate for a given ingress-egress-aggregate is derived as part of the flow termination procedure, and is used to determine how much PCN-traffic needs to be terminated. For further details see

Section 3.3.2.

CLE-reporting-threshold
A configurable value against which the CLE is compared as part of the report suppression procedure. For further details, see Section 3.2.3.

CLE-limit
A configurable value against which the CLE is compared to determine the PCN-admission-state for a given ingress-egress-aggregate. For further details, see Section 3.3.1.

T_meas
A configurable time interval that defines the measurement period over which the PCN-egress-node collects statistics relating to PCN-traffic marking. At the end of the interval the PCN-egress-node calculates the values NM-rate, [CL-specific] ThM-rate, and ETM-rate as defined above and sends a report to the Decision Point, subject to the operation of the report suppression feature. For further details see Section 3.2.

T_maxsuppress
A configurable time interval after which the PCN-egress-node MUST send a report to the Decision Point for a given ingress-egress-aggregate regardless of the most recent values of the CLE. This mechanism provides the Decision Point with a periodic confirmation of liveness when report suppression is activated. For further details, see Section 3.2.3.

T_fail
An interval after which the Decision Point concludes that communication from a given PCN-egress-node has failed if it has received no reports from the PCN-egress-node during that interval. For further details see Section 3.3.3.

T_crit
A configurable interval used in the calculation of T_fail. For further details see Section 3.3.3.

2. [CL-Specific] Assumed Core Network Behaviour for CL

This section describes the assumed behaviour for PCN-interior-nodes in the PCN-domain. The CL mode of operation assumes that:

- PCN-interior-nodes perform both threshold-marking and excess-traffic-marking of PCN-packets, according to the rules specified in [RFC5670];

- for IP transport, threshold-marking of PCN-packets uses the ThM codepoint defined in [ID.pcn-3-in-1]; for MPLS transport, an equivalent marking is used as discussed in Appendix C of [ID.pcn-3-in-1];

- for IP transport, excess-traffic-marking of PCN-packets uses the ETM codepoint defined in [ID.pcn-3-in-1]; for MPLS transport, an equivalent marking is used as discussed in Appendix C of [ID.pcn-3-in-1];

- on each link the reference rate for the threshold-meter is configured to be equal to the PCN-admissible-rate for the link;

- on each link the reference rate for the excess-traffic-meter is configured to be equal to the PCN-supportable-rate for the link;

- the set of valid codepoint transitions is as shown in Sections 5.2.1 and 5.2.2 of [ID.pcn-3-in-1].

3. Node Behaviours

3.1. Overview

This section describes the behaviour of the PCN-ingress-node, PCN-egress-node, and the Decision Point (which MAY be collocated with the PCN-ingress-node).

The PCN-egress-node collects the rates of not-marked, [CL-specific] threshold-marked, and excess-traffic-marked PCN-traffic for each ingress-egress-aggregate and reports them to the Decision Point. [CL-specific] It MAY also identify and report PCN-flows that have experienced excess-traffic-marking. For a detailed description, see Section 3.2.

The PCN-ingress-node enforces flow admission and termination decisions. It also reports the rate of PCN-traffic sent to a given ingress-egress-aggregate when requested by the Decision Point. For details, see Section 3.4.

Finally, the Decision Point makes flow admission decisions and selects flows to terminate based on the information provided by the PCN-ingress-node and PCN-egress-node for a given ingress-egress-aggregate. For details, see Section 3.3.

Specification of a signaling protocol to report rates to the Decision Point is out of scope of this document. If the PCN-ingress-node is chosen as the Decision Point, [I-D.tsvwg-rsvp-pcn] specifies an

appropriate signaling protocol.

Section 5.1.2 describes how to derive the filters by means of which PCN-ingress-nodes and PCN-egress-nodes are able to classify incoming packets into ingress-egress-aggregates.

3.2. Behaviour of the PCN-Egress-Node

3.2.1. Data Collection

The PCN-egress-node needs to meter the PCN-traffic it receives in order to calculate the following rates for each ingress-egress-aggregate passing through it. These rates SHOULD be calculated at the end of each measurement period based on the PCN-traffic observed during that measurement period. The duration of a measurement period is equal to the configurable value T_meas. Foir further information see Section 3.5.

- NM-rate: octets per second of PCN-traffic in PCN-packets that are not-marked (i.e., marked with the NM codepoint);
- [CL-specific] ThM-rate: octets per second of PCN-traffic in PCN-packets that are threshold-marked (i.e., marked with the ThM codepoint);
- ETM-rate: octets per second of PCN-traffic in PCN-packets that are excess-traffic-marked (i.e., marked with the ETM codepoint).

Note: metering the PCN-traffic continuously and using equal-length measurement intervals minimizes the statistical variance introduced by the measurement process itself. On the other hand, the operation of PCN is not affected if the starting and ending times of the measurement intervals for different ingress-egress-aggregates are different.

[CL-specific] As a configurable option, the PCN-egress-node MAY record flow identifiers of the PCN-flows for which excess-traffic-marked packets have been observed during this measurement interval. If this set is large (e.g., more than 20 flows), the PCN-egress-node MAY record only the most recently excess-traffic-marked PCN-flow identifiers rather than the complete set.

These can be used by the Decision Point when it selects flows for termination. In networks using multipath routing it is possible that congestion is not occurring on all paths carrying a given ingress-egress-aggregate. Assuming that specific PCN-flows are routed via specific paths, identifying the PCN-flows that are experiencing excess-traffic-marking helps to avoid termination of

PCN-flows not contributing to congestion.

3.2.2. Reporting the PCN Data

Unless the report suppression option described in Section 3.2.3 is activated, the PCN-egress-node MUST report the latest values of NM-rate, [CL-specific] ThM-rate, and ETM-rate to the Decision Point each time that it calculates them.

[CL-specific] If the PCN-egress-node recorded a set of flow identifiers of PCN-flows for which excess-traffic-marking was observed in the most recent measurement interval, then it MUST also include these identifiers in the report.

3.2.3. Optional Report Suppression

Report suppression MUST be provided as a configurable option, along with two configurable parameters, the CLE-reporting-threshold and the maximum report suppression interval T_maxsuppress. The default value of the CLE-reporting-threshold is zero. The CLE-reporting-threshold MUST NOT exceed the CLE-limit configured at the Decision Point. For further information on T_maxsuppress see Section 3.5.

If the report suppression option is enabled, the PCN-egress-node MUST apply the following procedure to decide whether to send a report to the Decision Point, rather than sending a report automatically at the end of each measurement interval.

1. As well as the quantities NM-rate, [CLE-specific] ThM-rate, and ETM-rate, the PCN-egress-node MUST calculate the congestion level estimate (CLE) for each measurement interval. The CLE is computed as:

   [CL-specific]
   CLE = (ThM-rate + ETM-rate) / (NM-rate + ThM-rate + ETM-rate)

   if any PCN-traffic was observed, or CLE = 0 if all the rates are zero.

2. If the CLE calculated for the latest measurement interval is greater than the CLE-reporting-threshold and/or the CLE calculated for the immediately previous interval was greater than the CLE-reporting-threshold, then the PCN-egress-node MUST send a report to the Decision Point. The contents of the report are described below.

   The reason for taking into account the CLE of the previous interval is to ensure that the Decision Point gets immediate

feedback if the CLE has dropped below CLE-reporting-threshold. This is essential if the Decision Point is running the flow termination procedure and observing whether (further) flow termination is needed. See Section 3.3.2.

3. If an interval T_maxsuppress has elapsed since the last report was sent to the Decision Point, then the PCN-egress-node MUST send a report to the Decision Point regardless of the CLE value.

4. If neither of the preceding conditions holds, the PCN-egress-node MUST NOT send a report for the latest measurement interval.

Each report sent to the Decision Point when report suppression has been activated MUST contain the values of NM-rate, [CL-specific] ThM-rate, ETM-rate, and CLE that were calculated for the most recent measurement interval. [CL-specific] If the PCN-egress-node recorded a set of flow identifiers of PCN-flows for which excess-traffic-marking was observed in the most recent measurement interval, then it MUST also include these identifiers in the report.

The above procedure ensures that at least one report is sent per interval (T_maxsuppress + T_meas). This demonstrates to the Decision Point that both the PCN-egress-node and the communication path between that node and the Decision Point are in operation.

3.3. Behaviour at the Decision Point

Operators can choose to use PCN procedures just for flow admission, or just for flow termination, or for both. Decision Points MUST implement both mechanisms, but configurable options MUST be provided to activate or deactivate PCN-based flow admission and flow termination independently of each other at a given Decision Point.

If PCN-based flow termination is enabled but PCN-based flow admission is not, flow termination operates as specified in this document.

Logically, some other system of flow admission control is in operation, but the description of such a system is out of scope of this document and depends on local arrangements.

3.3.1. Flow Admission

The Decision Point determines the PCN-admission-state for a given ingress-egress-aggregate each time it receives a report from the egress node. It makes this determination on the basis of the congestion level estimate (CLE). If the CLE is provided in the egress node report, the Decision Point SHOULD use the reported value. If the CLE was not provided in the report, the Decision Point MUST

calculate it based on the other values provided in the report, using the formula:

[CL-specific]
CLE = (ThM-rate + ETM-rate) / (NM-rate + ThM-rate + ETM-rate)

if any PCN-traffic was observed, or CLE = 0 if all the rates are zero.

The Decision Point MUST compare the reported or calculated CLE to a configurable value, the CLE-limit. If the CLE is less than the CLE-limit, the PCN-admission-state for that aggregate MUST be set to "admit"; otherwise it MUST be set to "block".

If the PCN-admission-state for a given ingress-egress-aggregate is "admit", the Decision Point SHOULD allow new flows to be admitted to that aggregate. If the PCN-admission-state for a given ingress-egress-aggregate is "block", the Decision Point SHOULD NOT allow new flows to be admitted to that aggregate. These actions MAY be modified by policy in specific cases, but such policy intervention risks defeating the purpose of using PCN.

A performance study of this admission control method is presented in [MeLe12].

3.3.2. Flow Termination

[CL-specific] When the report from the PCN-egress-node includes a non-zero value of the ETM-rate for some ingress-egress-aggregate, the Decision Point MUST request the PCN-ingress-node to provide an estimate of the rate (PCN-sent-rate) at which the PCN-ingress-node is receiving PCN-traffic that is destined for the given ingress-egress-aggregate.

If the Decision Point is collocated with the PCN-ingress-node, the request and response are internal operations.

The Decision Point MUST then wait, for both the requested rate from the PCN-ingress-node and the next report from the PCN-egress-node for the ingress-egress-aggregate concerned. If this next egress node report also includes a non-zero value for the ETM-rate, the Decision Point MUST determine the amount of PCN-traffic to terminate using the following steps:

1. [CL-specific] The sustainable aggregate rate (SAR) for the given ingress-egress-aggregate is estimated by the sum:

SAR = NM-rate + ThM-rate

for the latest reported interval.

2. The amount of traffic to be terminated is the difference:

PCN-sent-rate - SAR,

where PCN-sent-rate is the value provided by the PCN-ingress-node.

See Section 3.3.3 for a discussion of appropriate actions if the Decision Point fails to receive a timely response to its request for the PCN-sent-rate.

If the difference calculated in the second step is positive, the Decision Point SHOULD select PCN-flows to terminate, until it determines that the PCN-traffic admission rate will no longer be greater than the estimated sustainable aggregate rate. If the Decision Point knows the bandwidth required by individual PCN-flows (e.g., from resource signalling used to establish the flows), it MAY choose to complete its selection of PCN-flows to terminate in a single round of decisions.

Alternatively, the Decision Point MAY spread flow termination over multiple rounds to avoid over-termination. If this is done, it is RECOMMENDED that enough time elapse between successive rounds of termination to allow the effects of previous rounds to be reflected in the measurements upon which the termination decisions are based. (See [Satoh10] and sections 4.2 and 4.3 of [MeLe10].)

In general, the selection of flows for termination MAY be guided by policy. [CL-specific] If the egress node has supplied a list of identifiers of PCN-flows that experienced excess-traffic-marking (Section 3.2), the Decision Point SHOULD first consider terminating PCN-flows in that list.

The Decision Point SHOULD log each round of termination as described in Section 5.2.1.2.

3.3.3. Decision Point Action For Missing PCN-Boundary-Node Reports

The Decision Point SHOULD start a timer t_recvFail when it receives a report from the PCN-egress-node. t_recvFail is reset each time a new report is received from the PCN-egress-node. t_recvFail expires if it reaches the value T_fail. T_fail is calculated according to the following logic:

a. T_fail = the configurable duration T_crit, if report suppression is not deployed;

b. T_fail = T_crit also if report suppression is deployed and the last report received from the PCN-egress-node contained a CLE value greater than CLE-reporting-threshold (Section 3.2.3);

c. T_fail = 3 * T_maxsuppress (Section 3.2.3) if report suppression is deployed and the last report received from the PCN-egress-node contained a CLE value less than or equal to CLE-reporting-threshold.

If timer t_recvFail expires for a given PCN-egress-node, the Decision Point SHOULD notify management. A log format is defined for that purpose in Section 5.2.1.1. Other actions depend on local policy, but MAY include blocking of new flows destined for the PCN-egress-node concerned until another report is received from it. Termination of already-admitted flows is also possible, but could be triggered by "Destination unreachable" messages received at the PCN-ingress-node.

If a centralized Decision Point sends a request for the estimated value of PCN-sent-rate to a given PCN-ingress-node and fails to receive a response in a reasonable amount of time, the Decision Point SHOULD repeat the request once. [CL-specific] While waiting after sending this second request, the Decision Point MAY begin selecting flows to terminate, using ETM-rate as an estimate of the amount of traffic to be terminated in place of the quantity

PCN-sent-rate - SAR

specified in Section 3.3.2. Because ETM-rate will over-estimate the amount of traffic to be terminated due to dropping of PCN-packets by interior nodes, the Decision Point SHOULD terminate less than the full amount ETM-rate in the first pass and recalculate the additional amount to terminate in additional passes based on subsequent reports from the PCN-egress-node. If the second request to the PCN-ingress-node also fails, the Decision Point MUST select flows to terminate based on the ETM-rate approximation as just described and SHOULD notify management. The log format described in Section 5.2.1.1 is also suitable for this purpose.

The response timer t_sndFail with upper bound T_crit is specified in Section 3.5. The use of T_crit is an approximation. A more precise limit would be of the order of two round-trip times, plus an allowance for processing at each end, plus an allowance for variance in these values.

See Section 3.5 for suggested values of the configurable durations

T_crit and T_maxsuppress.

3.4. Behaviour of the Ingress Node

The PCN-ingress-node MUST provide the estimated current rate of PCN-traffic received at that node and destined for a given ingress-egress-aggregate in octets per second (the PCN-sent-rate) when the Decision Point requests it. The way this rate estimate is derived is a matter of implementation.

For example, the rate that the PCN-ingress-node supplies can be based on a quick sample taken at the time the information is required.

3.5. Summary of Timers and Associated Configurable Durations

Here is a summary of the timers used in the procedures just described:

t_meas

Where used: PCN-egress-node.

Used in procedure: data collection (Section 3.2.1).

Incidence: one per ingress-egress-aggregate.

Reset: immediately on expiry.

Expiry: when it reaches the configurable duration T_meas.

Action on expiry: calculate NM-rate, [CL-specific] ThM-rate, and ETM-rate and proceed to the applicable reporting procedure (Section 3.2.2 or Section 3.2.3).

t_maxsuppress

Where used: PCN-egress-node.

Used in procedure: report suppression (Section 3.2.3).

Incidence: one per ingress-egress-aggregate.

Reset: when the next report is sent, either after expiry or because the CLE has exceeded the reporting threshold.

Expiry: when it reaches the configurable duration T_maxsuppress.

Action on expiry: send a report to the Decision Point the next time the reporting procedure (Section 3.2.3) is invoked, regardless of the value of CLE.

t_recvFail

Where used: Decision Point.

Used in procedure: failure detection (Section 3.3.3).

Incidence: one per ingress-egress-aggregate.

Reset: when a report is received for the ingress-egress-aggregate.

Expiry: when it reaches the calculated duration T_fail. As described in Section 3.3.3, T_fail is equal either to the configured duration T_crit or to the calculated value 3 * T_maxsuppress, where T_maxsuppress is a configured duration.

Action on expiry: notify management, and possibly other actions.

t_sndFail

Where used: centralized Decision Point.

Used in procedure: failure detection (Section 3.3.3).

Incidence: only as required, one per outstanding request to a PCN-ingress-node.

Started: when a request for the value of PCN-sent-traffic for a given ingress-egress-aggregate is sent to the PCN-ingress-node.

Terminated without action: when a response is received before expiry.

Expiry: when it reaches the configured duration T_crit.

Action on expiry: as described in Section 3.3.3.

3.5.1. Recommended Values For the Configurable Durations

The timers just described depend on three configurable durations, T_meas, T_maxsuppress, and T_crit. The recommendations given below for the values of these durations are all related to the intended PCN reaction time of 1 to 3 seconds. However, they are based on

judgement rather than operational experience or mathematical derivation.

The value of T_meas is RECOMMENDED to be of the order of 100 to 500 ms to provide a reasonable tradeoff between demands on network resources (PCN-egress-node and Decision Point processing, network bandwidth) and the time taken to react to impending congestion.

The value of T_maxsuppress is RECOMMENDED to be on the order of 3 to 6 seconds, for similar reasons to those for the choice of T_meas.

The value of T_crit SHOULD NOT be less than 3 * T_meas. Otherwise it could cause too many management notifications due to transient conditions in the PCN-egress-node or along the signalling path. A reasonable upper bound on T_crit is in the order of 3 seconds.

4. Specification of Diffserv Per-Domain Behaviour

This section provides the specification required by [RFC3086] for a per-domain behaviour.

4.1. Applicability

This section quotes [RFC5559].

The PCN CL boundary node behaviour specified in this document is applicable to inelastic traffic (particularly video and voice) where quality of service for admitted flows is protected primarily by admission control at the ingress to the domain.

In exceptional circumstances (e.g., due to rerouting as a result of network failures) already-admitted flows may be terminated to protect the quality of service of the remaining flows. [CL-specific] The performance results in, e.g., [MeLe10], indicate that the CL boundary node behaviour provides better service outcomes under such circumstances than the SM boundary node behaviour described in [RFCyyyy], because CL is less likely to terminate PCN-flows unnecessarily.

[RFC EDITOR'S NOTE: please replace RFCyyyy above by the reference to the published version of draft-ietf-pcn-sm-edge-behaviour.]

4.2. Technical Specification

4.2.1. Classification and Traffic Conditioning

Packet classification and treatment at the PCN-ingress-node is described in Section 5.1 of [ID.pcn-3-in-1].

PCN packets are further classified as belonging or not belonging to an admitted flow. PCN packets not belonging to an admitted flow are "blocked". (See Section 1 for an understanding of how this term is interpreted.) Packets belonging to an admitted flow are policed to ensure that they adhere to the rate or flowspec that was negotiated during flow admission.

4.2.2. PHB Configuration

The PCN CL boundary node behaviour is a metering and marking behaviour rather than a scheduling behaviour. As a result, while the encoding uses a single DSCP value, that value can vary from one deployment to another. 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)

For a fuller discussion, see Appendix A of [ID.pcn-3-in-1].

4.3. Attributes

The purpose of this per-domain behaviour is to achieve low loss and jitter for the target class of traffic. The design requirement for PCN was that recovery from overloads through the use of flow termination should happen within 1-3 seconds. PCN probably performs better than that.

4.4. Parameters

The set of parameters that needs to be configured at each PCN-node and at the Decision Point is described in Section 5.1.

4.5. Assumptions

It is assumed that a specific portion of link capacity has been reserved for PCN-traffic.

4.6. Example Uses

The PCN CL behaviour may be used to carry real-time traffic, particularly voice and video.

4.7. Environmental Concerns

The PCN CL per-domain behaviour could theoretically interfere with the use of end-to-end ECN due to reuse of ECN bits for PCN marking. Section 5.1 of [ID.pcn-3-in-1] describes the actions that can be taken to protect ECN signalling. Appendix B of that document provides further discussion of how ECN and PCN can co-exist.

4.8. Security Considerations

Please see the security considerations in [RFC5559] as well as those in [RFC2474] and [RFC2475].

5. Operational and Management Considerations

5.1. Deployment of the CL Edge Behaviour

Deployment of the PCN Controlled Load edge behaviour requires the following steps:

- o selection of deployment options and global parameter values;
- o derivation of per-node and per-link information;
- o installation, but not activation, of parameters and policies at all of the nodes in the PCN domain;
- o activation and verification of all behaviours.

5.1.1. Selection of Deployment Options and Global Parameters

The first set of decisions affects the operation of the network as a whole. To begin with, the operator needs to make basic design decisions such as whether the Decision Point is centralized or collocated with the PCN-ingress-nodes, and whether per-flow and aggregate resource signalling as described in [I-D.tsvwg-rsvp-pcn] is deployed in the network. After that, the operator needs to decide:

- o whether PCN packets will be forwarded unencapsulated or in tunnels between the PCN-ingress-node and the PCN-egress-node. Encapsulation preserves incoming ECN settings and simplifies the PCN-egress-node’s job when it comes to relating incoming packets

to specific ingress-egress-aggregates, but lowers the path MTU and imposes the extra labour of encapsulation/decapsulation on the PCN-edge-nodes.

- which service classes will be subject to PCN control and what Diffserv code point (DSCP) will be used for each. (See [ID.pcn-3-in-1] Appendix A for advice on this topic.)

- the markings to be used at all nodes in the PCN domain to indicate Not-Marked (NM), [CL-specific] Threshold-Marked (ThM), and Excess-Traffic-Marked (ETM) PCN packets;

- The marking rules for re-marking PCN-traffic leaving the PCN domain;

- whether PCN-based flow admission is enabled;

- whether PCN-based flow termination is enabled.

The following parameters affect the operation of PCN itself. The operator needs to choose:

- the value of CLE-limit if PCN-based flow admission is enabled. [CL-specific] The operation of flow admission is not very sensitive to the value of the CLE-limit in practice, because when threshold-marking occurs it tends to persist long enough that threshold-marked traffic becomes a large proportion of the received traffic in a given interval.

- the value of the collection interval T_meas. For a recommended range of values see Section 3.5.1 above.

- whether report suppression is to be enabled at the PCN-egress-nodes and if so, the values of CLE-reporting-threshold and T_maxsuppress. It is reasonable to leave CLE-reporting-threshold at its default value (zero, as specified in Section 3.2.3). For a recommended range of values of T_maxsuppress see Section 3.5.1 above.

- the value of the duration T_crit, which the Decision Point uses in deciding whether communications with a given PCN-edge-node have failed. For a recommended range of values of T_crit see Section 3.5.1 above.

- [CL-specific] Activation/deactivation of recording of individual flow identifiers when excess-traffic-marked PCN-traffic is observed. Reporting these identifiers has value only if PCN-based flow termination is activated and Equal Cost Multi-Path (ECMP)

routing is enabled in the PCN-domain.

5.1.2. Specification of Node- and Link-Specific Parameters

Filters are required at both the PCN-ingress-node and the PCN-egress-node to classify incoming PCN packets by ingress-egress-aggregate. Because of the potential use of multi-path routing in domains upstream of the PCN-domain, it is impossible to do such classification reliably at the PCN-egress-node based on the packet header contents as originally received at the PCN-ingress-node. (Packets with the same header contents could enter the PCN-domain at multiple PCN-ingress-nodes.) As a result, the only way to construct such filters reliably is to tunnel the packets from the PCN-ingress-node to the PCN-egress-node.

The PCN-ingress-node needs filters in order to place PCN packets into the right tunnel in the first instance, and also to satisfy requests from the Decision Point for admission rates into specific ingress-egress-aggregates. These filters select the PCN-egress-node, but not necessarily a specific path through the network to that node. As a result, they are likely to be stable even in the face of failures in the network, except when the PCN-egress-node itself becomes unreachable. The primary basis for their derivation will be routing policy given the packet's original origin and destination. If all PCN packets will be tunneled, the PCN-ingress-node also needs to know the address of the peer PCN-egress-node associated with each filter.

Operators may wish to give some thought to the provisioning of alternate egress points for some or all ingress-egress aggregates in case of failure of the PCN-egress-node. This could require the setting up of standby tunnels to these alternate egress points.

Each PCN-egress-node needs filters to classify incoming PCN packets by ingress-egress-aggregate, in order to gather measurements on a per-aggregate basis. If tunneling is used, these filters are constructed on the basis of the identifier of the tunnel from which the incoming packet has emerged (e.g. the source address in the outer header if IP encapsulation is used). The PCN-egress-node also needs to know the address of the Decision Point to which it sends reports for each ingress-egress-aggregate.

A centralized Decision Point needs to have the address of the PCN-ingress-node corresponding to each ingress-egress-aggregate. Security considerations require that information also be prepared for a centralized Decision Point and each PCN-edge-node to allow them to authenticate each other.

Turning to link-specific parameters, the operator needs to derive

values for the PCN-admissible-rate and [CL-specific] PCN-supportable-rate on each link in the network. The first two paragraphs of Section 5.2.2 of [RFC5559] discuss how these values may be derived.

5.1.3. Installation of Parameters and Policies

As discussed in the previous two sections, every PCN node needs to be provisioned with a number of parameters and policies relating to its behaviour in processing incoming packets. The Diffserv MIB [RFC3289] can be useful for this purpose, although it needs to be extended in some cases. This MIB covers packet classification, metering, counting, policing and dropping, and marking. The required extensions specifically include an encapsulation action following re-classification by ingress-egress-aggregate. In addition, the MIB has to be extended to include objects for marking the ECN field in the outer header at the PCN-ingress-node and an extension to the classifiers to include the ECN field at PCN-interior and PCN-egress-nodes. Finally, new objects metering algorithms may need to be defined at the PCN-interior-nodes to represent the algorithms for threshold-marking and packet-size-independent excess-traffic-marking.

Values for the PCN-admissible-rate and [CL-specific] PCN-supportable-rate on each link on a node appear as metering parameters. Operators should take note of the need to deploy meters of a given type (threshold or excess-traffic) either on the ingress side or the egress of each interior link, but not both (Appendix B.2 of [RFC5670].

The following additional information has to be configured by other means (e.g., additional MIBs, NETCONF models).

At the PCN-egress-node:

- the measurement interval T_meas (units of ms, range 50 to 1000);
- [CL-specific] whether specific flow identifiers must be captured when excess-traffic-marked packets are observed;
- whether report suppression is to be applied;
- if so, the interval T_maxsuppress (units of 100 ms, range 1 to 100) and the CLE-reporting-threshold (units of tenths of one percent, range 0 to 1000, default value 0);
- the address of the PCN-ingress-node for each ingress-egress-aggregate, if the Decision Point is collocated with the PCN-ingress-node and [I-D.tsvwg-rsvp-pcn] is not deployed.

- the address of the centralized Decision Point to which it sends its reports, if there is one.

At the Decision Point:

- whether PCN-based flow admission is enabled;
- whether PCN-based flow termination is enabled.
- the value of CLE-limit (units of tenths of one percent, range 0 to 1000);
- the value of the interval T_crit (units of 100 ms, range 1 to 100);
- whether report suppression is to be applied;
- if so, the interval T_maxsuppress (units of 100 ms, range 1 to 100) and the CLE-reporting-threshold (units of tenths of one percent, range 0 to 1000, default value 0). These MUST be the same values that are provisioned in the PCN-egress-nodes;
- if the Decision Point is centralized, the address of the PCN-ingress-node (and any other information needed to establish a security association) for each ingress-egress-aggregate.

Depending on the testing strategy, it may be necessary to install the new configuration data in stages. This is discussed further below.

5.1.4. Activation and Verification of All Behaviours

It is certainly not within the scope of this document to advise on testing strategy, which operators undoubtedly have well in hand. Quite possibly an operator will prefer an incremental approach to activation and testing. Implementing the PCN marking scheme at PCN-ingress-nodes, corresponding scheduling behaviour in downstream nodes, and re-marking at the PCN-egress-nodes is a large enough step in itself to require thorough testing before going further.

Testing will probably involve the injection of packets at individual nodes and tracking of how the node processes them. This work can make use of the counter capabilities included in the Diffserv MIB. The application of these capabilities to the management of PCN is discussed in the next section.

5.2. Management Considerations

This section focuses on the use of event logging and the use of counters supported by the Diffserv MIB [RFC3289] for the various monitoring tasks involved in management of a PCN network.

5.2.1. Event Logging In the PCN Domain

It is anticipated that event logging using SYSLOG [RFC5424] will be needed for fault management and potentially for capacity management. Implementations MUST be capable of generating logs for the following events:

- detection of loss of contact between a Decision Point and a PCN-edge-node, as described in Section 3.3.3;
- successful receipt of a report from a PCN-egress-node, following detection of loss of contact with that node;
- flow termination events.

All of these logs are generated by the Decision Point. There is a strong likelihood in the first and third cases that the events are correlated with network failures at a lower level. This has implications for how often specific event types should be reported, so as not to contribute unnecessarily to log buffer overflow. Recommendations on this topic follow for each event report type.

The field names (e.g., HOSTNAME, STRUCTURED-DATA) used in the following subsections are defined in [RFC5424].

5.2.1.1. Logging Loss and Restoration of Contact

Section 3.3.3 describes the circumstances under which the Decision Point may determine that it has lost contact, either with a PCN-ingress-node or a PCN-egress-node, due to failure to receive an expected report. Loss of contact with a PCN-ingress-node is a case primarily applicable when the Decision Point is in a separate node. However, implementations MAY implement logging in the collocated case if the implementation is such that non-response to a request from the Decision Point function can occasionally occur due to processor load or other reasons.

The log reporting the loss of contact with a PCN-ingress-node or PCN-egress-node MUST include the following content:

- The HOSTNAME field MUST identify the Decision Point issuing the log.

- A STRUCTURED-DATA element MUST be present, containing parameters identifying the node for which an expected report has not been received and the type of report lost (ingress or egress). It is RECOMMENDED that the SD-ID for the STRUCTURED-DATA element have the form "PCNNode" (without the quotes), which has been registered with IANA. The node identifier PARAM-NAME is RECOMMENDED to be "ID" (without the quotes). The identifier itself is subject to the preferences expressed in Section 6.2.4 of [RFC5424] for the HOSTNAME field. The report type PARAM-NAME is RECOMMENDED to be "RTyp" (without the quotes). The PARAM-VALUE for the RTyp field MUST be either "ingr" or "egr".

The following values are also RECOMMENDED for the indicated fields in this log, subject to local practice:

- PRI initially set to 115, representing a Facility value of (14) "log alert" and a Severity level of (3) "Error Condition". Note that loss of contact with a PCN-egress-node implies that no new flows will be admitted to one or more ingress-egress-aggregates until contact is restored. The reason a higher severity level (lower value) is not proposed for the initial log is because any corrective action would probably be based on alerts at a lower subsystem level.

- APPNAME set to "PCN" (without the quotes).

- MSGID set to "LOST" (without the quotes).

If contact is not regained with a PCN-egress-node in a reasonable period of time (say, one minute), the log SHOULD be repeated, this time with a PRI value of 113, implying a Facility value of (14) "log alert" and a Severity value of (1) "Alert: action must be taken immediately". The reasoning is that by this time, any more general conditions should have been cleared, and the problem lies specifically with the PCN-egress-node concerned and the PCN application in particular.

Whenever a loss-of-contact log is generated for a PCN-egress-node, a log indicating recovery SHOULD be generated when the Decision Point next receives a report from the node concerned. The log SHOULD have the same content as just described for the loss-of-contact log, with the following differences:

- PRI changes to 117, indicating a Facility value of (14) "log alert" and a Severity of (5) "Notice: normal but significant condition".

- MSGID changes to "RECVD" (without the quotes).

5.2.1.2. Logging Flow Termination Events

Section 3.3.2 describes the process whereby the Decision Point decides that flow termination is required for a given ingress-egress-aggregate, calculates how much flow to terminate, and selects flows for termination. This section describes a log that SHOULD be generated each time such an event occurs. (In the case where termination occurs in multiple rounds, one log SHOULD be generated per round.) The log may be useful in fault management, to indicate the service impact of a fault occuring in a lower-level subsystem. In the absence of network failures, it may also be used as an indication of an urgent need to review capacity utilization along the path of the ingress-egress-aggregate concerned.

The log reporting a flow termination event MUST include the following content:

- The HOSTNAME field MUST identify the Decision Point issuing the log.

- A STRUCTURED-DATA element MUST be present, containing parameters identifying the ingress and egress nodes for the ingress-egress-aggregate concerned, indicating the total amount of flow being terminated, and giving the number of flows terminated to achieve that objective.

  It is RECOMMENDED that the SD-ID for the STRUCTURED-DATA element have the form: "PCNTerm" (without the quotes), which has been registered with IANA. The parameter identifying the ingress node for the ingress-egress-aggregate is RECOMMENDED to have PARAM-NAME "IngrID" (without the quotes). This parameter MAY be omitted if the Decision Point is collocated with that PCN-ingress-node. The parameter identifying the egress node for the ingress-egress-aggregate is RECOMMENDED to have PARAM-NAME "EgrID" (without the quotes). Both identifiers are subject to the preferences expressed in Section 6.2.4 of [RFC5424] for the HOSTNAME field.

  The parameter giving the total amount of flow being terminated is RECOMMENDED to have PARAM-NAME "TermRate" (without the quotes). The PARAM-VALUE MUST be the target rate as calculated according to the procedures of Section 3.3.2, as an integer value in thousands of octets per second. The parameter giving the number of flows selected for termination is RECOMMENDED to have PARAM-NAME "FCnt" (without the quotes). The PARAM-VALUE for this parameter MUST be an integer, the number of flows selected.

The following values are also RECOMMENDED for the indicated fields in this log, subject to local practice:

- PRI initially set to 116, representing a Facility value of (14) "log alert" and a Severity level of (4) "Warning: warning conditions".
- APPNAME set to "PCN" (without the quotes).
- MSGID set to "TERM" (without the quotes).

5.2.2. Provision and Use of Counters

The Diffserv MIB [RFC3289] allows for the provision of counters along the various possible processing paths associated with an interface and flow direction. It is RECOMMENDED that the PCN-nodes be instrumented as described below. It is assumed that the cumulative counts so obtained will be collected periodically for use in debugging, fault management, and capacity management.

PCN-ingress-nodes SHOULD provide the following counts for each ingress-egress-aggregate. Since the Diffserv MIB installs counters by interface and direction, aggregation of counts over multiple interfaces may be necessary to obtain total counts by ingress-egress-aggregate. It is expected that such aggregation will be performed by a central system rather than at the PCN-ingress-node.

- total PCN packets and octets received for that ingress-egress-aggregate but dropped;
- total PCN packets and octets admitted to that aggregate.

PCN-interior-nodes SHOULD provide the following counts for each interface, noting that a given packet MUST NOT be counted more than once as it passes through the node:

- total PCN packets and octets dropped;
- total PCN packets and octets forwarded without re-marking;
- [CL-specific] total PCN packets and octets re-marked to Threshold-Marked;
- total PCN packets and octets re-marked to Excess-Traffic-Marked.

PCN-egress-nodes SHOULD provide the following counts for each ingress-egress-aggregate. As with the PCN-ingress-node, so with the PCN-egress-node it is expected that any necessary aggregation over

multiple interfaces will be done by a central system.

- total Not-Marked PCN packets and octets received;
- [CL-specific] total Threshold-Marked PCN packets and octets received;
- total Excess-Traffic-Marked PCN packets and octets received.

The following continuously cumulative counters SHOULD be provided as indicated, but require new MIBs to be defined. If the Decision Point is not collocated with the PCN-ingress-node, the latter SHOULD provide a count of the number of requests for PCN-sent-rate received from the Decision Point and the number of responses returned to the Decision Point. The PCN-egress-node SHOULD provide a count of the number of reports sent to each Decision Point. Each Decision Point SHOULD provide the following:

- total number of requests for PCN-sent-rate sent to each PCN-ingress-node with which it is not collocated;
- total number of reports received from each PCN-egress-node;
- total number of loss-of-contact events detected for each PCN-boundary-node;
- total cumulative duration of "block" state in hundreds of milliseconds for each ingress-egress-aggregate;
- total number of rounds of flow termination exercised for each ingress-egress-aggregate.

6. Security Considerations

[RFC5559] provides a general description of the security considerations for PCN. This memo introduces one new consideration, related to the use of a centralized Decision Point. The Decision Point itself is a trusted entity. However, its use implies the existence of an interface on the PCN-ingress-node through which communication of policy decisions takes place. That interface is a point of vulnerability which must be protected from denial of service attacks.

7. IANA Considerations

This document requests IANA to add the following entries to the

syslog Structured Data ID Values registry. RFCxxxx is this document when published.

Structured Data ID: PCNNode OPTIONAL

Structured Data Parameter: ID MANDATORY

Structured Data Parameter: Rtyp MANDATORY

Reference: RFCxxxx

Structured Data ID: PCNTerm OPTIONAL

Structured Data Parameter: IngrID MANDATORY

Structured Data Parameter: EgrID MANDATORY

Structured Data Parameter: TermRate MANDATORY

Structured Data Parameter: FCnt MANDATORY

Reference: RFCxxxx

8. Acknowledgements

The content of this memo bears a family resemblance to [ID.briscoe-CL]. The authors of that document were Bob Briscoe, Philip Eardley, and Dave Songhurst of BT, Anna Charny and Francois Le Faucheur of Cisco, Jozef Babiarz, Kwok Ho Chan, and Stephen Dudley of Nortel, Giorgios Karagiannis of U. Twente and Ericsson, and Attila Bader and Lars Westberg of Ericsson.

Ruediger Geib, Philip Eardley, and Bob Briscoe have helped to shape the present document with their comments. Toby Moncaster gave a careful review to get it into shape for Working Group Last Call.

Amongst the authors, Michael Menth deserves special mention for his constant and careful attention to both the technical content of this document and the manner in which it was expressed.

David Harrington’s careful AD review resulted not only in necessary changes throughout the document, but also the addition of the operations and management considerations (Section 5).

As part of the broader review process, the document saw further improvements as a result of comments by Joel Halpern, Brian Carpenter, Stephen Farrell, Sean Turner, and Pete Resnick.

9. References

9.1. Normative References

[ID.pcn-3-in-1]
Briscoe, B., Moncaster, T., and M. Menth, "Encoding 3 PCN-States in the IP header using a single DSCP", March 2012.

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

[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.

[RFC2475] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., and W. Weiss, "An Architecture for Differentiated Services", RFC 2475, December 1998.

[RFC3086] Nichols, K. and B. Carpenter, "Definition of Differentiated Services Per Domain Behaviors and Rules for their Specification", RFC 3086, April 2001.

[RFC3289] Baker, F., Chan, K., and A. Smith, "Management Information Base for the Differentiated Services Architecture", RFC 3289, May 2002.

[RFC5424] Gerhards, R., "The Syslog Protocol", RFC 5424, March 2009.

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

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

9.2. Informative References

[I-D.tsvwg-rsvp-pcn]
Karagiannis, G. and A. Bhargava, "Generic Aggregation of Resource ReSerVation Protocol (RSVP) for IPv4 And IPv6 Reservations over PCN domains (Work in progress)", July 2011.

[ID.briscoe-CL]
Briscoe, B., "An edge-to-edge Deployment Model for Pre-Congestion Notification: Admission Control over a DiffServ Region (expired Internet Draft)", 2006.

[MeLe10] Menth, M. and F. Lehrieder, "PCN-Based Measured Rate Termination", Computer Networks Journal (Elsevier) vol. 54, no. 13, pages 2099 - 2116, September 2010.

[MeLe12] Menth, M. and F. Lehrieder, "Performance of PCN-Based Admission Control under Challenging Conditions, IEEE/ACM Transactions on Networking, vol. 20, no. 2", April 2012.

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

[RFCyyyy] Charny, A., Zhang, J., Karagiannis, G., Menth, M., and T. Taylor, "PCN Boundary Node Behaviour for the Single Marking (SM) Mode of Operation (Work in progress)", December 2010.

[Satoh10] Satoh, D. and H. Ueno, ""Cause and Countermeasure of Overtermination for PCN-Based Flow Termination", Proceedings of IEEE Symposium on Computers and Communications (ISCC ’10), pp. 155-161, Riccione, Italy", June 2010.

Authors’ Addresses

Anna Charny
USA

Phone:
Email: anna@mwsm.com

Fortune Huang
Huawei Technologies
Section F, Huawei Industrial Base,
Bantian Longgang, Shenzhen 518129
P.R. China

Phone: +86 15013838060
Email: fqhuang@huawei.com

Georgios Karagiannis
U. Twente

Phone:
Email: karagian@cs.utwente.nl

Michael Menth
University of Tuebingen
Sand 13
Tuebingen D-72076
Germany

Phone: +49-7071-2970505
Email: menth@informatik.uni-tuebingen.de

Tom Taylor (editor)
Huawei Technologies
Ottawa, Ontario
Canada

Email: tom.taylor.stds@gmail.com

PCN
Internet-Draft
Intended status: Informational
Expires: September 08, 2012

G. Karagiannis
University of Twente
K. Chan
Consultant
T. Moncaster
University of Cambridge
M. Menth
University of Tuebingen
P. Eardley
B. Briscoe
BT

March 08, 2012

Overview of Pre-Congestion Notification Encoding
draft-ietf-pcn-encoding-comparison-09

Abstract

The objective of Pre-Congestion Notification (PCN) is to protect the quality of service (QoS) of inelastic flows within a Diffserv domain. On every link in the PCN domain, the overall rate of the PCN-traffic is metered, and PCN-packets are appropriately marked when certain configured rates are exceeded. Egress nodes provide decision points with information about the PCN-marks of PCN-packets which allows them to take decisions about whether to admit or block a new flow request, and to terminate some already admitted flows during serious pre-congestion.

The PCN Working Group explored a number of approaches for encoding this pre-congestion information into the IP header. This document provides details of all those approaches along with an explanation of the constraints that had to be met by any solution.

Status of this Memo

Copyright Notice

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. Two mechanisms are used: admission control, to decide whether to admit or block a new flow request, and flow termination to terminate some existing flows during serious pre-congestion. To achieve this, the overall rate of PCN-traffic is metered on every link in the domain, and PCN-packets are appropriately marked when certain configured rates are exceeded. These configured rates are below the rate of the link. Thus boundary nodes are notified of a potential overload before any real congestion occurs (hence "pre-congestion notification").

[RFC5670] provides for two metering and marking functions that are configured with reference rates. Threshold-marking marks all PCN packets once their traffic rate on a link exceeds the configured reference rate (PCN-threshold-rate). Excess-traffic-marking marks only those PCN packets that exceed the configured reference rate (PCN-excess-rate).

Egress nodes monitor the PCN-marks of received PCN-packets and provide information about the PCN-marks to decision points which take decisions about flow admission and termination on this basis [I-D.ietf-pcn-cl-edge-behaviour], [I-D.ietf-pcn-sm-edge-behaviour].

This PCN information has to be encoded into the IP header. This PCN information has to be encoded into the IP header. This requires at least three different codepoints: one for PCN traffic that has not been marked, one for traffic that has been marked by the threshold meter and one for traffic that has been marked by the excess-traffic-meter.
Since unused codepoints are not available for that purpose in the IP header (version 4 and 6), already used codepoints must be re-used which imposes additional constraints on design and applicability of PCN-based admission control (AC) and flow termination (FT). This document summarizes these issues as a record of the PCN WG discussions and for the benefit of the wider IETF community.

In Section 2, we briefly point out PCN encoding requirement imposed by metering and marking algorithms, and by special packet drop strategies. The Differentiated Services Codepoint (6 bits) and the ECN field (2 bits) have been selected to be re-used for encoding of PCN marks (PCN encoding). In Section 3, we briefly explain the constraints imposed by this decision. In Section 4, we review different PCN encodings supported by the PCN working group that allow different implementations of PCN-based admission control and flow termination which have different pros and cons.

2. General PCN Encoding Requirements

The choice of metering and marking algorithms and the way they are applied to PCN-based AC and FT impose certain requirements on PCN encoding.

2.1. Metering and Marking Algorithms

Two different metering and marking algorithms are defined in [RFC5670]: excess-traffic-marking and threshold-marking. They are both configured with reference rates which are termed PCN-excess-rate and PCN-threshold-rate, respectively. When traffic for PCN flows enter a PCN domain, the PCN ingress node sets a codepoint in the IP header indicating that the packet is subject to PCN metering and marking and that it is not-marked (NM). The two metering and marking algorithms possibly re-mark PCN packets as PCN and excess-traffic-marked (ETM) or threshold-marked (ThM).

Excess-traffic-marking leaves a rate of PCN traffic equal to the PCN-excess-rate to be not-ETM marked if possible. To that end, the algorithm needs to know whether a PCN packet has already been ETM marked or not. Threshold-marking re-marks all not-marked PCN traffic to ThM when the rate of PCN traffic exceeds the PCN-threshold-rate. Therefore, it does not need knowledge of the prior marking state of the packet for metering, but it needs it for packet re-marking.

2.2. Approaches for PCN-Based Admission Control and Flow Termination

We briefly review three different approaches to implement PCN-based AC and FT and derive their requirements for PCN encoding.

2.2.1. Dual Marking (DM)

The intuitive approach for PCN-based AC and FT requires that threshold and excess-traffic-marking are simultaneously activated on all links of a PCN domain and their reference rate is configured with the PCN-admissible-rate (AR) and the PCN-supportable-rate (SR), respectively. Threshold-marking meters all PCN traffic, but re-marks only not-marked traffic (NM) to ThM. Excess-traffic-marking meters only non-ETM traffic and re-marks either not-marked (NM) or threshold-marked (ThM) PCN traffic to ETM. Thus, both meters and markers need to identify PCN packets and their exact PCN codepoint. We call this marking behavior dual marking (DM) and Figure 1 illustrates all possible re-marking actions.

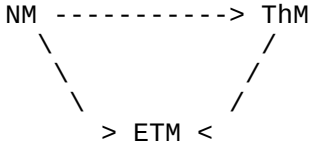


Figure 1: PCN Codepoint Re-Marking Diagram for Dual Marking (DM)

Dual marking is used to support the Controlled-Load PCN (CL-PCN) edge behavior [I-D.ietf-pcn-cl-edge-behaviour]. We briefly summarize the concept. All actions are performed on per ingress-egress-aggregate basis. The egress node measures the rate of NM-, ThM-, and ETM-traffic in regular intervals and sends them as PCN egress reports to the AC and FT decision point.

If the proportion of re-marked (ThM- and ETM-) PCN traffic is larger than a defined threshold, called CLE-limit, the decision point blocks new flow requests until new PCN egress reports are received, otherwise it admits them. With CL-PCN, AC is rather robust with regard to the value chosen for the CLE-limit. FT works as follows. If the ETM-traffic rate is positive, the decision point triggers the ingress node to send a newly measured rate of the sent PCN traffic. The decision point calculates the rate of PCN traffic that needs to be terminated by:

termination-rate = PCN-ingress-rate - (rate-of-NM-traffic + rate-of-ThM-traffic)

and terminates an appropriate set of flows. CL-PCN is accurate enough for most application scenarios and its implementation complexity is acceptable, therefore, it is a preferred implementation option for PCN-based AC and FT.

2.2.2. Single Marking (SM)

Single-marking uses only excess-traffic-marking whose reference rate is set to the PCN-admissible-rate (AR) on all links of the PCN domain. Figure 2 illustrates all possible re-marking actions.

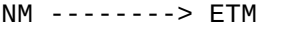


Figure 2: PCN Codepoint Re-Marking Diagram for Single Marking (SM)

Single marking is used to support the single-marking PCN (SM-PCN) edge behavior [I-D.ietf-pcn-sm-edge-behaviour]. We briefly summarize the concept. AC works essentially in the same way as with CL-PCN but AC is sensitive to the value of the CLE-limit. Also FT works similarly to CL-PCN. The PCN-supportable-rate (SR) is not configured on any link, but is implicitly:

SR=u*AR

in the PCN domain using a network-wide constant u. The decision point triggers FT only if the rate-of-NM-traffic * u < rate-of-NM-traffic + rate-of-ETM-traffic, requests the PCN-sent-rate from the corresponding PCN-ingress-node, calculates the amount of PCN traffic to be terminated by

termination-rate = PCN-sent-rate - rate-of-NM-traffic * u,

and terminates an appropriate set of flows.

SM-PCN has two major benefits: it requires only two PCN codepoints and only excess-traffic-marking is needed which means that it might be earlier to the market than CL-PCN since some chipsets do not yet support threshold-marking.

However, it only works well when ingress-egress-aggregates have a high PCN packet rate which is not always the case. Otherwise, over-admission and over-termination may occur [Menth12] [Menth10q].

2.2.3. Packet Specific Dual Marking (PSDM)

Packet-specific dual marking (PSDM) uses threshold-marking and excess-traffic-marking whose reference rates are configured with the PCN-admissible-rate and the PCN-supportable-rate, respectively. There are two different types of not-marked packets: those that are subject to threshold-marking (not-ThM) and those that are subject to excess-traffic-marking (not-ETM). Both not-Thm and not-ETM have the same NM-marking and are distinguished by higher layer information (see below). Threshold-marking meters all PCN traffic and re-marks only not-ThM packets to PCN-marked (PM). In contrast, excess-traffic-marking meters only not-ETM packets and possibly re-marks them to PM, too. Again, both meters and markers need to identify PCN packets and their exact PCN codepoint. Figure 3 illustrates all possible re-marking actions.

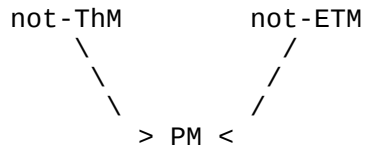


Figure 3: PCN Codepoint Re-Marking Diagram for Packet Specific Dual Marking (PSDM)

An edge behavior for PSDM has been presented in [Menth09f]. We call it PSDM-PCN. In contrast to CL-PCN and SM-PCN, AC is realized by re-using marked signaling messages for probing. The assumption is that admission requests are triggered by an external end-to-end signaling protocol, e.g. RSVP (RFC2205). Signaling traffic for a flow is also labeled as PCN traffic and if an initial signaling traverses the PCN domain and is re-marked, then the corresponding flow is blocked. This is a light-weight probing mechanism which does not generate extra traffic and does not introduce probing delay [draft-menth-pcn-marked-signaling-ac]. In PSDM-PCN, PCN-ingress-nodes label initial signaling messages as not-ThM and threshold-marking configured with admissible rates possibly re-marks them to PM. Data packets are labeled with not-ETM and excess-traffic-marking configured with supportable rates possibly re-marks them to PM, too, so that the same algorithms for FT may be used as for CL-PCN and SM-PCN.

Disadvantages of this approach are that every end-to-end signaling protocol, e.g. RSVP, needs to be adapted that it denies admission if initial request messages are re-marked to PM.
Advantages are that the AC algorithm is more accurate than the one of CL-PCN and SM-PCN [Menth12], that only a single DSCP is needed, and that the new tunneling rules in RFC6040 are not needed for deployment.

2.2.4. Preferential Packet Dropping

The termination algorithms described in [I-D.ietf-pcn-cl-edge behaviour] and [I-D.ietf-pcn-sm-edge-behaviour] require the preferential dropping of ETM-marked packets to avoid over-termination in the case of packet loss. An analysis explaining this phenomenon can be found in Section 4 of [Menth10q]. Thus, preferential dropping of ETM-marked packets is "RECOMMENDED" in [RFC5670]. As a consequence, droppers must have access to the exact marking information of PCN-packets.

3. Encoding Constraints

The PCN WG decided to use the DS field (i.e., combination of the DSCP and ECN field) for the encoding of the PCN Marks, see [RFC5696]. This section describes the criteria that are used to compare the resulting encoding options described in section 4.

3.1. Structure of the DS Field

Figure 4 shows the structure of the DS field. [RFC0793] defined the 8 bit ToS field and [RFC2474] redefined it as DS field. It consists of a 6 bit DS codepoint (DSCP, see [RFC2474]) and the 2 bit ECN field (see [RFC3168]).

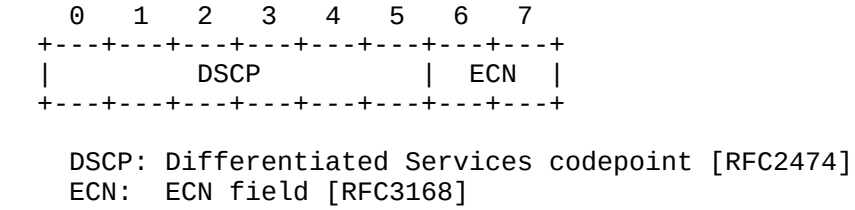


Figure 4: The Structure of the DS Field

3.2. Constraints from the DSCP

The Differentiated Services codepoint (DSCP) indicates the per-hop behavior (PHB), i.e., the treatment IP packets receive from nodes in a DS domain. Multiple DSCPs may indicate the same PHB. PCN traffic is high-priority traffic and requires a special DSCP that indicate a PHB with preferred treatment.

3.2.1. General Scarcity of DSCPs

As the number of unused DSCPs is small, PCN encoding should use only a single DSCP if possible, in any case not more than two DSCPs. Therefore, the DSCP should be used to indicate that traffic is subject to PCN metering and marking, but not to differentiate different PCN markings.

3.2.2. Handling of the DSCP in Tunneling Rules

PCN encoding must be chosen in such a way that PCN traffic can be tunneled within a PCN domain without any impact on PCN metering and re-marking. In the following, the "inner header" refers to the header of the encapsulated packet and the "outer header" refers to the encapsulating header.

[RFC2983] provides two tunneling modes for Differentiated Services networks. The uniform model copies the DSCP from the inner header to the outer header upon encapsulation and it copies the DSCP from the outer header to the inner header upon decapsulation. This assures that changes applied to the DSCP field survive encapsulation and decapsulation. In contrast, the pipe model ignores the content of the DSCP field in the outer header upon decapsulation. Therefore, decapsulation erases changes applied to the DSCP along the tunnel. As a consequence, only the uniform model may be used for tunneling PCN traffic within a PCN domain, if PCN encoding uses more than a single DSCP.

3.2.3. Restoration of Original DSCPs at the Egress Node

If PCN-marking does not alter the original DSCP, the traffic leaves the PCN-domain with its original DSCP.
However, if the PCN-marking alters the DSCP, then some additional technique is needed to restore the original DSCP. A few possibilities are discussed:

1. Each Diffserv class using PCN uses a different set of DSCPs. Therefore, if there are M DSCPs using PCN and PCN encoding uses N different DSCPs, N*M DSCPs are needed. This solution may work well in IP networks. However, when PCN is applied to MPLS networks or other layers restricted to 8 QoS classes and codepoints, this solution fails due to the extreme shortage of available DSCPs.

2. The original DSCP for the packets of a flow is signaled to the egress node. No suitable signaling protocol has been developed and therefore, it is not clear whether this approach could work.

3. PCN-traffic is tunneled across the PCN-domain. The pipe tunneling model is applied and so the original DSCP is restored after decapsulation. However, tunneling across a PCN domain adds an additional IP header and reduces the maximum transfer unit (MTU) from the perspective of the user. GRE, MPLS, or Ethernet using Pseudo-Wires are potential solutions that scale well also in backbone networks.

The most appropriate option depends on the specific circumstances an operator faces.

- o) Option 1 is most suitable unless there is a shortage of available DSCPs.

o) Option 3 is suitable where the reduction of MTU is not liable to cause issues.

3.3. Constraints from the ECN Field

This section briefly reviews the structure and use of the ECN field. The ECN field may be redefined, but certain constraints must be met [RFC4774]. The impact on PCN deployment is discussed, as well as the constraints imposed by various tunneling rules on the persistence of PCN marks after decapsulation and its impact on possible re-marking actions.

3.3.1. Structure and Use of the ECN Field

Some transport protocols, like TCP, can typically use packet drops as an indication of congestion in the Internet. The idea of Explicit Congestion Notification (ECN) [RFC3168] is that routers provide a congestion indication for incipient congestion, where the notification can sometimes be through ECN marking (and re-marking) packets rather than dropping them. Figure 5 summarizes the ECN codepoints defined [RFC3168].

```
+-----+-----+
| ECN FIELD |
+-----+-----+
   0     0         Not-ECT
   0     1         ECT(1)
   1     0         ECT(0)
   1     1         CE
```

Figure 5: ECN Codepoints within the ECN field

ECT stands for "ECN-capable transport" and indicates that the sender and receivers of a flow understand ECN semantics. Packets of other flows are labeled with not-ECT. To indicate congestion to a receiver, routers may re-mark ECT(1) or ECT(0) labeled packets to CE which stands for "congestion experienced". Two different ECT codepoints were introduced "to protect against accidental or malicious concealment of marked packets from the TCP sender" which may be the case with cheating receivers [RFC3540].

3.3.2. Redefinition of the ECN Field

The ECN field may be redefined for other purposes and [RFC4774] gives guidelines for that. Essentially, not-ECT-marked packets must never be re-marked to ECT or CE because not-ECT-capable end systems do not reduce their transmission rate when receiving CE-marked packets. This is a threat to the stability of the Internet.

Moreover, CE-marked packets must not be re-marked to not-ECT or ECT, because then ECN-capable end systems cannot reduce their transmission rate. The re-use of the ECN field for PCN encoding has some impact on the deployment of PCN. First, routers within a PCN domain must not apply ECN re-marking when the ECN field has PCN semantics. Second, before a PCN packet leaves the PCN domain, the egress nodes must either (A) reset the ECN field of the packet to the contents it had when entering the PCN domain or (B) reset its ECN field to not-ECT. According to Section 3.3.3, tunneling ECN traffic through a PCN domain may help to implement (A). When (B) applies, CE-marked packets must never become PCN packets within a PCN domain as the egress node resets their ECN field to not-ECT. The ingress node may drop such traffic instead.

3.3.3. Handling of the ECN Field in Tunneling Rules

When packets are encapsulated, the ECN field of the inner header may or may not be copied to the ECN field of the outer header and upon decapsulation, the ECN field of the outer header may or may not be copied from the ECN field of the outer header to the ECN field of the inner header. Various tunneling rules with different treatment of the ECN field exist. Two different modes are defined in [RFC3168] for IP-in-IP tunnels and a third one in [RFC4301] for IP-in-IPsec tunnels. [RFC6040] updates both these RFCs to rationalize them into one consistent approach.

3.3.3.1. Limited Functionality Option

The limited-functionality option has been defined in [RFC3168]. Upon

encapsulation, the ECN field of the outer header is generally set to not-ECT. Upon decapsulation, the ECN field of the inner header remains unchanged.

Since this tunneling mode loses information upon encapsulation and decapsulation, it cannot be used for tunneling PCN traffic within a PCN domain. However, the PCN ingress may use this mode to tunnel traffic with ECN semantics to the PCN egress to preserve the ECN field in the inner header while the ECN field of the outer header is used with PCN semantics within the PCN domain.

3.3.3.2. Full Functionality Option

The full-functionality option has been defined in [RFC3168]. Upon encapsulation, the ECN field of the inner header is copied to the outer header unless the ECN field of the inner header carries CE. In that case, the ECN field of the outer header is set to ECT(0). This choice has been made for security reasons, to disable the ECN fields of the outer header as a covert channel. Upon decapsulation, the ECN field of the inner header remains unchanged unless the ECN field of the outer header carries CE. In that case, the ECN field of the inner header is also set to CE.

This mode imposes the following constraints on PCN metering and marking. First, PCN must re-mark the ECN field only to CE because any other information is not copied to the inner header upon decapsulation and will be lost. Second, CE information in encapsulated packet headers is invisible for routers along a tunnel. Threshold marking does not require information about whether PCN packets have already been marked and would work when CE denotes that packets are marked. In contrast, excess-traffic- marking requires information about already excess-traffic-marked packets and cannot be supported with this tunneling mode. Furthermore, this tunneling mode cannot be used when marked or not-marked packets should be preferentially dropped because the PCN marking information is possibly not visible in the outer header of a packet.

3.3.3.3. Tunneling with IPSec

Tunneling has been defined in Section 5.1.2.1 of [RFC4301]. Upon encapsulation, the ECN field of the inner header is copied to the ECN field of the outer header. Decapsulation works as for the full-functionality option in Section 3.3.3.2. Tunneling with IPsec also requires that PCN re-marks the ECN field only to CE because any other information is not copied to the inner header upon decapsulation and lost. In contrast to Section 3.3.3.2, with IPsec tunnels, CE marks of tunneled PCN traffic remain visible for routers along the tunnel and to their meters, markers, and droppers.

3.3.3.4. ECN Tunneling

New tunneling rules for ECN are specified in [RFC6040], which updates [RFC3168] and [RFC4301]. These rules provide a consistent and rational approach to encapsulation and decapsulation.

With the normal mode, the ECN field of the inner header is copied to the ECN field of the outer header on encapsulation.
In compatibility mode, the ECN field of the outer header is reset to not-ECT.

Upon decapsulation, the scheme specified in [RFC6040] and shown in Figure 6 is applied. Thus, re-marking encapsulated not-ECT packets to any other codepoint would not survive decapsulation. Therefore, not-ECT cannot be used for PCN encoding. Furthermore, re-marking encapsulated ECT(0) packets to ECT(1) or CE survives decapsulation, but not vice-versa, and re-marking encapsulated ECT(1) packets to CE also survives decapsulation, but not vice-versa.
Certain combinations of inner and outer ECN fields cannot result from any transition in any current or previous ECN tunneling specification. These currently unused (CU) combinations are indicated in Figure 6 by ’(!!!)’ or ’(!)’, where ’(!!!)’ means the combination is CU and always potentially dangerous, while ’(!)’ means it is CU and possibly dangerous.

Arriving Inner Header	Arriving Outer Header: Not-ECT	ECT(0)	ECT(1)	CE
Not-ECT	Not-ECT	Not-ECT(!!!)	Not-ECT(!!!)	<drop>(!!!)
ECT(0)	ECT(0)	ECT(0)	ECT(1)	CE
ECT(1)	ECT(1)	ECT(1) (!)	ECT(1)	CE
CE	CE	CE	CE(!!!)	CE

The ECN field in the outgoing header is set to the codepoint at the intersection of the appropriate arriving inner header (row) and arriving outer header (column), or the packet is dropped where indicated. Currently unused combinations are indicated by '(!!!)' or '(!)'. ([RFC6040]: '(!!!)' means the combination is CU and always potentially dangerous, while '(!)' means it is CU and possibly dangerous.)

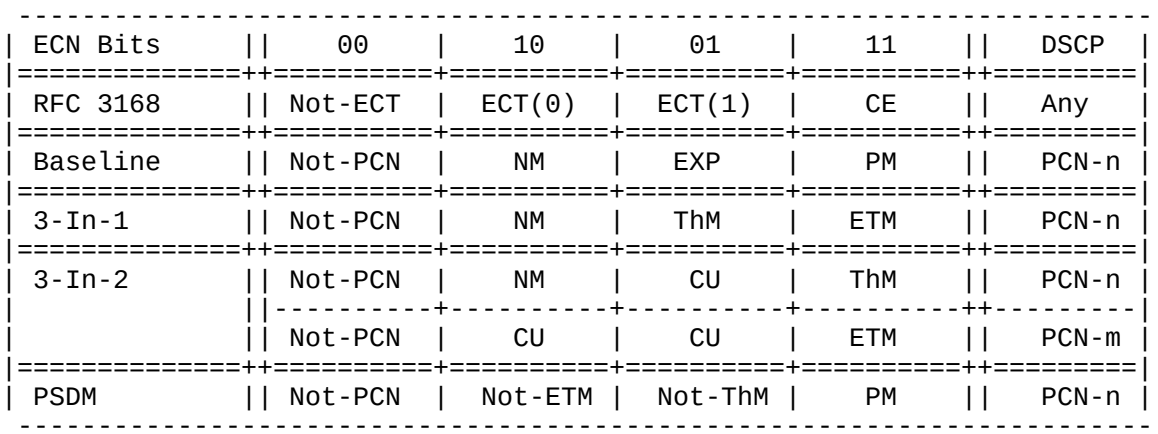
Figure 6: New IP in IP Decapsulation Behavior (from [RFC6040])

3.3.4. Restoration of the Original ECN Field at the PCN-Egress-Node

As ECN is an end-to-end service, it is desirable that the egress node of a PCN domain restores the ECN field a PCN packet had at the ingress node. There are basically two options. PCN traffic may be tunneled between ingress and egress node using limited functionality tunnels (see Section 3.3.3.1). Then, PCN marking is applied only to the outer header, and the original ECN field is restored after decapsulation. However, this reduces the MTU from the perspective of the user. Another option is to use some intelligent encoding that preserves the ECN codepoints. However, a viable solution is not known.

4. Comparison of Encoding Options

The PCN WG has studied four different PCN encodings, which redefine the ECN field. Figure 7 summarizes these PCN encodings. One or at most two different DSCPs are used to indicate PCN traffic, and only for these DSCPs the semantics of the ECN field are redefined within the PCN domain.
When a PCN-ingress-node classifies a packet as a PCN-packet it sets its PCN-codepoint to not-marked (NM). Non-PCN traffic can also to be sent with the PCN-specific DSCP, by setting the Not-PCN codepoint. Special per hop behavior, defined in [RFC5670], applies to PCN-traffic.



Notes: PCN-n, PCN-m under the DSCP column denotes PCN compatible DSCPs which may be chosen by the network operator. Not-PCN means that packets are not PCN-enabled. NM means Not-Marked to signal a not-pre-congested path. CU means Currently Unused.

Figure 7: Semantics of the ECN field for various encoding types

4.1. Baseline Encoding

With baseline encoding [RFC5696], the NM codepoint can be re-marked only to PCN-marked (PM). Excess-traffic-marking uses PM as ETM, threshold-marking uses PM as ThM, and only one of the two marking schemes can be used.

The 01-codepoint is reserved for experimental purposes (EXP) and the other defined PCN encoding schemes can be seen as extensions of baseline encoding by appropriate redefinition of EXP. Baseline encoding [RFC5696] works well with IPsec tunnels (see Section 3.3.3.3).

4.2. Encoding with 1 DSCP Providing 3 States

PCN 3-state encoding extension in a single DSCP (3-in-1 encoding, [I-D.ietf-pcn-3-in-1-encoding] extends the baseline encoding and supports the simultaneous use of both excess-traffic-marking and threshold-marking. 3-in-1 encoding well supports the preferred CL-PCN and also SM-PCN.

The problem with 3-in-1 encoding is that the 10-codepoint does not survive decapsulation with the tunneling options in Section 3.3.3.1 - 3.3.3.3. Therefore, 3-in-1 encoding may be used only for PCN domains implementing the new rules for ECN tunneling [RFC6040], see Section 3.3.3.4), or where it is known that there are no tunnels in the PCN domain. Currently it is not clear how fast the new tunneling rules will be deployed, but the applicability of 3-in-1-encoding depends on that.

4.3. Encoding with 2 DSCPs Providing 3 or More States

PCN encoding using 2 DSCPs to provide 3 or more states (3-in-2 encoding, [I-D.ietf-pcn-3-state-encoding] uses two different DSCPs to accommodate the three required codepoints NM, ThM, and ETM. It leaves some codepoints currently unused (CU) and proposes also one way how to reuse them to store some information about the content of the ECN field before the packet entered the PCN domain. 3-in-2 encoding works well with IPsec tunnels (see Section 3.3.3.3). This type of encoding can support both CL-PCN and SM-PCN schemes.

The disadvantage of 3-in-2 encoding is that it consumes two DSCPs. Moreover, the direct application of this encoding scheme to other technologies like MPLS, where even fewer bits are available for the encoding of DSCPs is more difficult.

4.4. Encoding for Packet Specific Dual Marking (PSDM)

PCN encoding for packet-specific dual marking (PSDM) is designed to support PSDM-PCN outlined in Section 2.2.3. It is the only proposal that supports PCN-based AC and FT with only a single DSCP [I-D.ietf-pcn-psdm-encoding] in the presence of IPsec tunnels (see Section 3.3.3.3). PSDM encoding also supports SM-PCN.

4.5. Standardized encodings

The baseline encoding described in section 4.1 was published as a draft Internet Standard [RFC5696]. The intention was to allow for experimental encodings to build upon this baseline. However, following the publication of [RFC6040], the WG decided to change approach and instead standardize only one encoding (the 3-in-1 encoding described in 4.2 [I-D.ietf-pcn-3-in-1-encoding]. Rather than defining the 3-in-1 encoding as a standards track extension to the existing baseline encoding [RFC5696], it was agreed that it was best to define a new standards track document that obsoletes [RFC5696].

5. Conclusion

This document summarizes the PCN Working Group’s exploration of a number of approaches for encoding pre-congestion information into the IP header. It is presented as an informational archive. It provides details of all those approaches along with an explanation of the constraints that have to be met. The Working Group has concluded that the "3-in-1" encoding should be published as a standards-track RFC that obsoletes the encoding specified in [RFC5696].
The reasoning is as follows. During the early life of the working group, we decided on an approach of a standardized "baseline encoding" [RFC5696] plus a series of experimental encodings that would all build on the baseline encoding and each of which would be useful in specific circumstances. However, after the tunneling of ECN was standardized in [RFC6040], the PCN WG decided on a different approach - to recommend just one encoding, the "3-in-1 encoding".

Although in theory "3-in-1" could be specified as a standards-track extension to the "baseline" encoding, the WG decided that it would be cleaner to obsolete [RFC5696] and specify "3-in-1" encoding in a new stand-alone RFC.

6. Security Implications

[RFC5559] provides a general description of the security considerations for PCN. This memo does not introduce additional security considerations.

7. IANA Considerations

This memo includes no request to IANA.

8. Acknowledgements

We would like to acknowledge the members of the PCN working group and Gorry Fairhust for the discussions that generated and improved the contents of this memo.

9. References

9.1. Normative References

[RFC0793] Postel, J., "Transmission Control Protocol", STD 7, RFC 793, September 1981.

[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.

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

9.2. Informative References

[I-D.ietf-pcn-cl-edge-behaviour] Charny, A., Huang, F., Karagiannis, G., Menth, M., and T. Taylor, "PCN Boundary Node Behaviour for the Controlled Load (CL) Mode of Operation", draft-ietf-pcn-cl-edge-behaviour-12 (work in progress), February 2012.

[I-D.ietf-pcn-sm-edge-behaviour]
Charny, A., Karagiannis, G., Menth, M., and T. Taylor, "PCN Boundary Node Behaviour for the Single Marking (SM) Mode of Operation", draft-ietf-pcn-sm-edge-behaviour-09 (work in progress), February 2012.

[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-08 (work in progress), August 2011.

[I-D.ietf-pcn-3-state-encoding]
Briscoe, B., Moncaster, T., and M. Menth, "A PCN encoding using 2 DSCPs to provide 3 or more states", draft-ietf-pcn-3-state-encoding-01 (work in progress), February 2010.

[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.

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

[RFC2983] Black, D., "Differentiated Services and Tunnels", RFC 2983, October 2000.

[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.

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

[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.

[Menth09f]
Menth, M., Babiarz, J., and P. Eardley, "Pre-Congestion Notification Using Packet-Specific Dual Marking", IEEE Proceedings of the International Workshop on the Network of the Future (Future-Net) at Dresden Germany, June 2009.

[Menth12]
Menth, M. and F. Lehrieder, " Performance of PCN-Based Admission Control under Challenging Conditions", accepted for publication IEEE/ACM Transactions on Networking in 2012.

[Menth10q]
Menth, M. and F. Lehrieder, "PCN-Based Measured Rate Termination", Computer Networks Journal, vol. 54, no. 3, Sept. 2010

Authors' Addresses

Georgios Karagiannis
University of Twente
P.O. Box 217
7500 AE Enschede,
The Netherlands

Email: g.karagiannis@utwente.nl

Kwok Ho Chan
Consultant

Email: khchan.work@gmail.com

Toby Moncaster
University of Cambridge Computer Laboratory,
William Gates Building, J J Thomson Avenue,
Cambridge, CB3 0FD.

Email Toby.Moncaster@cl.cam.ac.uk

Michael Menth
Chair of Communication Networks
University of Tuebingen
Sand 13
72076 Tuebingen
Germany

Email: menth@informatik.uni-tuebingen.de

Philip Eardley
BT
B54/77, Sirius House Adastral Park Martlesham Heath
Ipswich, Suffolk IP5 3RE
United Kingdom

Email: philip.eardley@bt.com

Bob Briscoe
BT
B54/77, Sirius House Adastral Park Martlesham Heath
Ipswich, Suffolk IP5 3RE
United Kingdom

Email: bob.briscoe@bt.com

Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: August 08, 2012

G. Karagiannis
University of Twente
T. Taylor
Huawei
K. Chan
Consultant
M. Menth
University of Tuebingen
P. Eardley
BT

February 08, 2012

Requirements for Signaling of (Pre-) Congestion Information in a DiffServ Domain
draft-ietf-pcn-signaling-requirements-08

Abstract

Precongestion notification (PCN) is a means for protecting quality of service for inelastic traffic admitted to a Diffserv domain. The overall PCN architecture is described in RFC 5559. This memo describes the requirements for the signaling applied within the PCN domain: (1) PCN-feedback-information is carried from the PCN-egress-node to the decision point;(2) the decision point may ask the PCN-ingress-node to measure, and report back, the rate of sent PCN-traffic between that PCN-ingress-node and PCN-egress-node. The decision point may be either collocated with the PCN-ingress-node or a centralized node (in the first case, (2) is not required). The signaling requirements pertain in particular to two edge behaviors, "controlled load (CL)" and "single marking (SM)".

Requirements Language

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 RFC 2119 [RFC2119].

Table of Contents

1. Introduction

The main objective of Pre-Congestion Notification (PCN) is to support the quality of service (QoS) of inelastic flows within a Diffserv domain in a simple, scalable, and robust fashion. Two mechanisms are used: admission control and flow termination. Admission control is used to decide whether to admit or block a new flow request while flow termination is used in abnormal circumstances to decide whether to terminate some of the existing flows. To support these two features, the overall rate of PCN-traffic is metered on every link in the 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 to boundary nodes about overloads before any congestion occurs (hence "pre-congestion" notification). The PCN-egress-nodes measure the rates of differently marked PCN traffic in periodic intervals and report these rates to the decision points for admission control and flow termination, based on which they take their decisions. The decision points may be collocated with the PCN-ingress-nodes or their function may be implemented in a centralized node.
For more details see [RFC5559],
[draft-ietf-pcn-cl-edge-behaviour-11],
[draft-ietf-pcn-sm-edge-behaviour-08].

This memo specifies the requirements on signaling protocols:
- to carry reports from a PCN-egress-node to the decision point,
- to carry requests, from the decision point to a PCN-ingress-node, that trigger the PCN-ingress-node to measure the PCN-sent-rate,
- to carry reports, from a PCN-ingress-node to the decision point.

The latter two messages are only needed if the decision point and PCN-ingress-node are not collocated.

2. Signaling Requirements for Messages from the PCN-Egress-Nodes to Decision Point(s)

The PCN-egress-node measures per ingress-egress-aggregate the rates of differently marked PCN-traffic in regular intervals. The measurement intervals are recommended to take a fixed value between 100 ms and 500 ms, see [draft-ietf-pcn-cl-edge-behaviour-11], [draft-ietf-pcn-sm-edge-behaviour-08]. At the end of each measurement interval, the PCN-egress-node calculates the congestion-level-estimate (CLE) based on these quantities.

The PCN-egress-node MAY be configured to record a set of identifiers of PCN-flows for which it received excess-traffic-marked packets during the last measurement interval. The latter may be useful to perform flow termination in networks with multipath routing.

At the end of each measurement interval, or less frequently if "optional report suppression" is activated, see

[draft-ietf-pcn-cl-edge-behaviour-11], [draft-ietf-pcn-sm-edge-behaviour-08], the PCN-egress-node sends a report to the decision point.

For the SM edge behavior, the report MUST contain:
- identifier of the PCN-ingress-node and the identifier of the PCN-egress-node (typically their IP addresses); together they specify the ingress-egress-aggregate to which the report refers,
- rate of not-marked PCN-traffic (NM-rate) in octets/second,
- rate of PCN-marked traffic (PM-rate) in octets/second,

For the CL edge behavior, the report MUST contain:
- identifier of the PCN-ingress-node and the identifier of the PCN-egress-node (typically their IP addresses); together they specify the ingress-egress-aggregate to which the report refers,
- rate of not-marked PCN-traffic (NM-rate) in octets/second,
- rate of threshold-marked PCN traffic (ThM-rate) in octets/second,
- rate of excess-traffic-marked traffic (ETM-rate) in octets/second,

The number format and the rate units used by the signaling protocol will limit the maximum rate that PCN can use. If signaling space is tight it might be reasonable to impose a limit, but any such limit may impose unnecessary constraints in future.

The signaling report can either be sent directly to the decision point or it can "piggy-back", i.e., be included within some other message that passes through the PCN-egress-node and then reaches the decision point.

As described in [draft-ietf-pcn-cl-edge-behaviour-11], PCN reports from the PCN-egress-node to the decision point may contain flow identifiers for individual flows within an ingress-egress-aggregate that have recently experienced excess-marking. Hence, the PCN report messages used by the PCN CL edge behavior MUST be capable of carrying sequences of octet strings constituting such identifiers."

Signaling messages SHOULD have a higher priority and a lower drop precedence than PCN-packets, see [RFC5559], to deliver them quickly and to avoid that they are dropped in case of overload.

The load generated by the signaling protocol SHOULD be minimized. We give three examples that may help to achieve that goal:
- piggy-backing the reports by the PCN-egress-nodes to the decision point(s) onto other signaling messages that are already in place,
- reducing the amount of reports to be sent by optional report suppression,
- combining reports for different ingress-egress-aggregates in a single message (if they are for the same decision point).

As PCN reports are sent regularly, additional reliability mechanisms are not needed. This also holds in the presence of optional report suppression as reports are sent periodically if actions by the decision point(s) are needed, see [draft-ietf-pcn-cl-edge-behaviour--11], [draft-ietf-pcn-sm-edge-behaviour-08].

3. Signaling Requirements for Messages between Decision Point(s) and PCN-Ingress-Nodes

Through request-response signaling between the decision point and PCN-ingress-node, the decision point requests and in response the PCN-ingress-node measures and reports the PCN-sent-rate for a specific ingress-egress-aggregate. Signaling is needed only if the decision point and PCN-ingress-node are not collocated.

The request MUST contain:
- o the identifier of the PCN-ingress-node and the identifier of the PCN-egress-node; together they determine the ingress-egress-aggregate for which the PCN-sent-rate is requested,
- o the identifier of the decision point that requests the PCN-sent-rate.

The report MUST contain:
- o the PCN-sent-rate in octets/second,
- o the identifier of the PCN-ingress-node and the identifier of the PCN-egress-node.

The request MUST be addressed to the PCN-ingress-node, and the report MUST be addressed to the decision point that requested it.

The request and the report SHOULD be sent with high priority, a lower drop precedence than PCN-packets, and reliably, because they are sent only when flow termination is needed, which is an urgent action.

Note that a complete system description for a PCN domain with centralized Decision Point includes the signaling from Decision Point to the PCN-ingress-nodes to control flow admission and termination. However, this is a known problem whose solutions were given by, for example, [RFC3084] or [RFC5431], and it lies outside the scope of the present document.

4. Security Considerations

[RFC5559] provides a general description of the security considerations for PCN. This memo relies on the security related requirements on the PCN signaling, provided in [RFC5559]. In particular, the signaling between the PCN-boundary-nodes must be protected from attacks. For example, the recipient needs to validate that the message is indeed from the node that claims to have sent it. Possible measures include digest authentication and protection against replay and man-in-the-middle attacks.

For the generic aggregate RSVP protocol, specifically, additional protection methods against security attacks are described in [RFC4860].

5. IANA Considerations

This memo includes no request to IANA.

6. Acknowledgments

We would like to acknowledge the members of the PCN working group for the discussions that produced the contents of this memo.

7. References

7.1. Normative References

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

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

[draft-ietf-pcn-cl-edge-behaviour-11] T. Taylor, A, Charny, F. Huang, G. Karagiannis, M. Menth, "PCN Boundary Node Behaviour for the Controlled Load (CL) Mode of Operation (Work in progress)", December 2011.

[draft-ietf-pcn-sm-edge-behaviour-08] A. Charny, J. Zhang, G. Karagiannis, M. Menth, T. Taylor, "PCN Boundary Node Behaviour for the Single Marking (SM) Mode of Operation (Work in progress)", December 2011.

7.2. Informative References

[RFC3084] K. Chan, J. Seligson, D. Durham, S. Gai, K. McCloghrie, S. Herzog, F. Reichmeyer, R. Yavatkar, A. Smith, "COPS Usage for Policy Provisioning (COPS-PR)", RFC 3084, March 2001.

[RFC4860] F. Le Faucheur, B. Davie, P. Bose, C. Christou, M. Davenport, "Generic Aggregate Resource ReSerVation Protocol (RSVP) Reservations", RFC 4860, May 2007.

[RFC5431] D. Sun, "Diameter ITU-T Rw Policy Enforcement Interface Application", RFC 5431, March 2009.

Authors' Addresses

Georgios Karagiannis
University of Twente
P.O. Box 217
7500 AE Enschede,
The Netherlands
EMail: g.karagiannis@utwente.nl

Tom Taylor
Huawei Technologies
1852 Lorraine Ave.
Ottawa, Ontario K1H 6Z8
Canada
Phone: +1 613 680 2675
Email: tom.taylor.stds@gmail.com

Kwok Ho Chan
Consultant

Email: khchan.work@gmail.com

Michael Menth
University of Tuebingen
Department of Computer Science
Chair of Communication Networks
Sand 13
72076 Tuebingen
Germany
Phone: +49 7071 29 70505
Email: menth@informatik.uni-tuebingen.de

Philip Eardley
BT
B54/77, Sirius House Adastral Park Martlesham Heath
Ipswich, Suffolk IP5 3RE
United Kingdom
EMail: philip.eardley@bt.com

Internet Engineering Task Force
Internet-Draft
Intended status: Experimental
Expires: October 8, 2012

A. Charny
J. Zhang
Cisco Systems
G. Karagiannis
U. Twente
M. Menth
University of Tuebingen
T. Taylor, Ed.
Huawei Technologies
April 6, 2012

PCN Boundary Node Behaviour for the Single Marking (SM) Mode of Operation
draft-ietf-pcn-sm-edge-behaviour-12

Abstract

Pre-congestion notification (PCN) is a means for protecting the quality of service for inelastic traffic admitted to a Diffserv domain. The overall PCN architecture is described in RFC 5559. This memo is one of a series describing possible boundary node behaviours for a PCN-domain. The behaviour described here is that for a form of measurement-based load control using two PCN marking states, not-marked, and excess-traffic-marked. This behaviour is known informally as the Single Marking (SM) PCN-boundary-node behaviour.

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 October 8, 2012.

Copyright Notice

Table of Contents

1. Introduction

The objective of Pre-Congestion Notification (PCN) is to protect the quality of service (QoS) of inelastic flows within a Diffserv domain, in a simple, scalable, and robust fashion. Two mechanisms are used: admission control, to decide whether to admit or block a new flow request, and (in abnormal circumstances) flow termination to decide whether to terminate some of the existing flows. To achieve this, 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 to PCN-boundary-nodes about incipient overloads before any congestion occurs (hence the "pre" part of "pre-congestion notification"). The level of marking allows decisions to be made about whether to admit or terminate PCN-flows. For more details see [RFC5559].

This document describes an experimental edge node behaviour to implement PCN in a network. The experiment may be run in a network in which a substantial proportion of the traffic carried is in the form of inelastic flows and where admission control of micro-flows is applied at the edge. For the effects of PCN to be observable, the committed bandwidth (i.e., level of non-best-effort traffic) on at least some links of the network should be near or at link capacity. The amount of effort required to prepare the network for the experiment (see Section 5.1) may constrain the size of network to which it is applied. The purposes of the experiment are:

- to validate the specification of the SM edge behaviour;
- to evaluate the effectiveness of the SM edge behaviour in preserving quality of service for admitted flows; and
- to evaluate PCN’s potential for reducing the amount of capital and operational costs in comparison to alternative methods of assuring quality of service.

For the first two objectives, the experiment should run long enough for the network to experience sharp peaks of traffic in at least some directions. It would also be desirable to observe PCN performance in the face of failures in the network. A period in the order of a month or two in busy season may be enough. The third objective is more difficult, and could require observation over a period long enough for traffic demand to grow to the point where additional capacity must be provisioned at some points in the network.

Section 3 of this document specifies a detailed set of algorithms and procedures used to implement the PCN mechanisms for the SM mode of

operation. Since the algorithms depend on specific metering and marking behaviour at the interior nodes, it is also necessary to specify the assumptions made about PCN-interior-node behaviour (Section 2). Finally, because PCN uses DSCP values to carry its markings, a specification of PCN-boundary-node behaviour must include the per domain behaviour (PDB) template specified in [RFC3086], filled out with the appropriate content (Section 4).

Note that the terms "block" or "terminate" actually translate to one or more of several possible courses of action, as discussed in Section 3.6 of [RFC5559]. The choice of which action to take for blocked or terminated flows is a matter of local policy.

[RFC EDITOR'S NOTE: RFCyyyy is the published version of draft-ietf-pcn-cl-edge-behaviour.]

A companion document [RFCyyyy] specifies the Controlled Load (CL) PCN-boundary-node behaviour. This document and [RFCyyyy] have a great deal of text in common. To simplify the task of the reader, the text in the present document that is specific to the SM PCN-boundary-node behaviour is preceded by the phrase: "[SM-specific]". A similar distinction for CL-specific text is made in [RFCyyyy].

1.1. Terminology

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].

This document uses the following terms defined in Section 2 of [RFC5559]:

- PCN-domain;
- PCN-ingress-node;
- PCN-egress-node;
- PCN-interior-node;
- PCN-boundary-node;
- PCN-flow;
- ingress-egress-aggregate (IEA);
- PCN-excess-rate;

- PCN-admissible-rate;
- PCN-supportable-rate;
- PCN-marked;
- excess-traffic-marked.

It also uses the terms PCN-traffic and PCN-packet, for which the definition is repeated from [RFC5559] because of their importance to the understanding of the text that follows:

PCN-traffic, PCN-packets, PCN-BA
: A PCN-domain carries traffic of different Diffserv behaviour aggregates (BAs) [RFC2474]. The PCN-BA uses the PCN mechanisms to carry PCN-traffic, and the corresponding packets are PCN-packets. The same network will carry traffic of other Diffserv BAs. The PCN-BA is distinguished by a combination of the Diffserv codepoint and the ECN field.

This document uses the following term from [RFC5670]:

- excess-traffic-meter.

To complete the list of borrowed terms, this document reuses the following terms and abbreviations defined in Section 3 of [ID.pcn-3-in-1]:

- not-PCN codepoint;
- Not-marked (NM) codepoint;
- Excess-traffic-marked (ETM) codepoint.

This document defines the following additional terms:

Decision Point
: The node that makes the decision about which flows to admit and to terminate. In a given network deployment, this can be the PCN-ingress-node or a centralized control node. In either case, the PCN-ingress-node is the point where the decisions are enforced.

NM-rate
: The rate of not-marked PCN-traffic received at a PCN-egress-node for a given ingress-egress-aggregate in octets per second. For further details see Section 3.2.1.

ETM-rate
: The rate of excess-traffic-marked PCN-traffic received at a PCN-egress-node for a given ingress-egress-aggregate in octets per second. For further details see Section 3.2.1.

PCN-sent-rate
: The rate of PCN-traffic received at a PCN-ingress-node and destined for a given ingress-egress-aggregate in octets per second. For further details see Section 3.4.

Congestion level estimate (CLE)
: The ratio of PCN-marked to total PCN-traffic (measured in octets) received for a given ingress-egress-aggregate during a given measurement period. The CLE is used to derive the PCN-admission-state (Section 3.3.1) and is also used by the report suppression procedure (Section 3.2.3) if report suppression is activated.

PCN-admission-state
: The state ("admit" or "block") derived by the Decision Point for a given ingress-egress-aggregate based on PCN packet marking statistics. The Decision Point decides to admit or block new flows offered to the aggregate based on the current value of the PCN-admission-state. For further details see Section 3.3.1.

Sustainable aggregate rate (SAR)
: The estimated maximum rate of PCN-traffic that can be carried in a given ingress-egress-aggregate at a given moment without risking degradation of quality of service for the admitted flows. The intention is that if the PCN-sent-rate of every ingress-egress-aggregate passing through a given link is limited to its sustainable aggregate rate, the total rate of PCN-traffic flowing through the link will be limited to the PCN-supportable-rate for that link. An estimate of the sustainable aggregate rate for a given ingress-egress-aggregate is derived as part of the flow termination procedure, and is used to determine how much PCN-traffic needs to be terminated. For further details see Section 3.3.2.

CLE-reporting-threshold
: A configurable value against which the CLE is compared as part of the report suppression procedure. For further details, see Section 3.2.3.

CLE-limit
: A configurable value against which the CLE is compared to determine the PCN-admission-state for a given ingress-egress-aggregate. For further details, see Section 3.3.1.

T_meas
A configurable time interval that defines the measurement period over which the PCN-egress-node collects statistics relating to PCN-traffic marking. At the end of the interval the PCN-egress-node calculates the values NM-rate and ETM-rate as defined above and sends a report to the Decision Point, subject to the operation of the report suppression feature. For further details see Section 3.2.

T_maxsuppress
A configurable time interval after which the PCN-egress-node MUST send a report to the Decision Point for a given ingress-egress-aggregate regardless of the most recent values of the CLE. This mechanism provides the Decision Point with a periodic confirmation of liveness when report suppression is activated. For further details, see Section 3.2.3.

T_fail
An interval after which the Decision Point concludes that communication from a given PCN-egress-node has failed if it has received no reports from the PCN-egress-node during that interval. For further details see Section 3.3.3.

T_crit
A configurable interval used in the calculation of T_fail. For further details see Section 3.3.3.

2. [SM-Specific] Assumed Core Network Behaviour for SM

This section describes the assumed behaviour for PCN-interior-nodes in the PCN-domain. The SM mode of operation assumes that:

- PCN-interior-nodes perform excess-traffic-marking of PCN-packets according to the rules specified in [RFC5670].
- for IP transport, excess-traffic-marking of PCN-packets uses the excess-traffic-marked (ETM) codepoint defined in [ID.pcn-3-in-1]; for MPLS transport, an equivalent marking is used as discussed in [ID.pcn-3-in-1] Appendix C;
- on each link the reference rate for the excess-traffic-meter is configured to be equal to the PCN-admissible-rate for the link;
- the set of valid codepoint transitions is as shown in Sections 5.2.1 and 5.2.3.1 of [ID.pcn-3-in-1].

3. Node Behaviours

3.1. Overview

This section describes the behaviour of the PCN-ingress-node, PCN-egress-node, and the Decision Point (which MAY be collocated with the PCN-ingress-node).

The PCN-egress-node collects the rates of not-marked and excess-traffic-marked PCN-traffic for each ingress-egress-aggregate and reports them to the Decision Point. For a detailed description, see Section 3.2.

The PCN-ingress-node enforces flow admission and termination decisions. It also reports the rate of PCN-traffic sent to a given ingress-egress-aggregate when requested by the Decision Point. For details, see Section 3.4.

Finally, the Decision Point makes flow admission decisions and selects flows to terminate based on the information provided by the PCN-ingress-node and PCN-egress-node for a given ingress-egress-aggregate. For details, see Section 3.3.

Specification of a signaling protocol to report rates to the Decision Point is out of scope of this document. If the PCN-ingress-node is chosen as the Decision Point, [I-D.tsvwg-rsvp-pcn] specifies an appropriate signaling protocol.

Section 5.1.2 describes how to derive the filters by means of which PCN-ingress-nodes and PCN-egress-nodes are able to classify incoming packets into ingress-egress-aggregates.

3.2. Behaviour of the PCN-Egress-Node

3.2.1. Data Collection

The PCN-egress-node needs to meter the PCN-traffic it receives in order to calculate the following rates for each ingress-egress-aggregate passing through it. These rates SHOULD be calculated at the end of each measurement period based on the PCN-traffic observed during that measurement period. The duration of a measurement period is equal to the configurable value T_meas. For further information see Section 3.5.

- NM-rate: octets per second of PCN-traffic in PCN-packets that are not-marked (i.e., marked with the NM codepoint);

- ETM-rate: octets per second of PCN-traffic in PCN-packets that are excess-traffic-marked (i.e., marked with the ETM codepoint).

  Note: metering the PCN-traffic continuously and using equal-length measurement intervals minimizes the statistical variance introduced by the measurement process itself. On the other hand, the operation of PCN is not affected if the starting and ending times of the measurement intervals for different ingress-egress-aggregates are different.

3.2.2. Reporting the PCN Data

Unless the report suppression option described in Section 3.2.3 is activated, the PCN-egress-node MUST report the latest values of NM-rate and ETM-rate to the Decision Point each time that it calculates them.

3.2.3. Optional Report Suppression

Report suppression MUST be provided as a configurable option, along with two configurable parameters, the CLE-reporting-threshold and the maximum report suppression interval T_maxsuppress. The default value of the CLE-reporting-threshold is zero. The CLE-reporting-threshold MUST NOT exceed the CLE-limit configured at the Decision Point. For further information on T_maxsuppress see Section 3.5.

If the report suppression option is enabled, the PCN-egress-node MUST apply the following procedure to decide whether to send a report to the Decision Point, rather than sending a report automatically at the end of each measurement interval.

1. As well as the quantities NM-rate and ETM-rate, the PCN-egress-node MUST calculate the congestion level estimate (CLE) for each measurement interval. The CLE is computed as:

   ```
   [SM-specific]
   CLE = ETM-rate / (NM-rate + ETM-rate)
   ```

   if any PCN-traffic was observed, or CLE = 0 if all the rates are zero.

2. If the CLE calculated for the latest measurement interval is greater than the CLE-reporting-threshold and/or the CLE calculated for the immediately previous interval was greater than the CLE-reporting-threshold, then the PCN-egress-node MUST send a report to the Decision Point. The contents of the report are described below.

The reason for taking into account the CLE of the previous interval is to ensure that the Decision Point gets immediate feedback if the CLE has dropped below the CLE-reporting-threshold. This is essential if the Decision Point is running the flow termination procedure and observing whether (further) flow termination is needed. See Section 3.3.2.

3. If an interval T_maxsuppress has elapsed since the last report was sent to the Decision Point, then the PCN-egress-node MUST send a report to the Decision Point regardless of the CLE value.

4. If neither of the preceding conditions holds, the PCN-egress-node MUST NOT send a report for the latest measurement interval.

Each report sent to the Decision Point when report suppression has been activated MUST contain the values of NM-rate, ETM-rate, and CLE that were calculated for the most recent measurement interval.

The above procedure ensures that at least one report is sent per interval (T_maxsuppress + T_meas). This demonstrates to the Decision Point that both the PCN-egress-node and the communication path between that node and the Decision Point are in operation.

3.3. Behaviour at the Decision Point

Operators can choose to use PCN procedures just for flow admission, or just for flow termination, or for both. Decision Points MUST implement both mechanisms, but configurable options MUST be provided to activate or deactivate PCN-based flow admission and flow termination independently of each other at a given Decision Point.

If PCN-based flow termination is enabled but PCN-based flow admission is not, flow termination operates as specified in this document.

Logically, some other system of flow admission control is in operation, but the description of such a system is out of scope of this document and depends on local arrangements.

3.3.1. Flow Admission

The Decision Point determines the PCN-admission-state for a given ingress-egress-aggregate each time it receives a report from the egress node. It makes this determination on the basis of the congestion level estimate (CLE). If the CLE is provided in the egress node report, the Decision Point SHOULD use the reported value. If the CLE was not provided in the report, the Decision Point MUST calculate it based on the other values provided in the report, using the formula:

[SM-specific]
CLE = ETM-rate / (NM-rate + ETM-rate)

if any PCN-traffic was observed, or CLE = 0 if all the rates are zero.

The Decision Point MUST compare the reported or calculated CLE to a configurable value, the CLE-limit. If the CLE is less than the CLE-limit, the PCN-admission-state for that aggregate MUST be set to "admit"; otherwise it MUST be set to "block".

If the PCN-admission-state for a given ingress-egress-aggregate is "admit", the Decision Point SHOULD allow new flows to be admitted to that aggregate. If the PCN-admission-state for a given ingress-egress-aggregate is "block", the Decision Point SHOULD NOT allow new flows to be admitted to that aggregate. These actions MAY be modified by policy in specific cases, but such policy intervention risks defeating the purpose of using PCN.

A performance study of this admission control method is presented in [MeLe12].

3.3.2. Flow Termination

[SM-specific] When the PCN-admission-state computed on the basis of the CLE is "block" for the given ingress-egress-aggregate, the Decision Point MUST request the PCN-ingress-node to provide an estimate of the rate (PCN-sent-rate) at which the PCN-ingress-node is receiving PCN-traffic that is destined for the given ingress-egress-aggregate.

If the Decision Point is collocated with the PCN-ingress-node, the request and response are internal operations.

The Decision Point MUST then wait, for both the requested rate from the PCN-ingress-node and the next report from the PCN-egress-node for the ingress-egress-aggregate concerned. If this next egress node report also includes a non-zero value for the ETM-rate, the Decision Point MUST determine the amount of PCN-traffic to terminate using the following steps:

1. [SM-specific] The sustainable aggregate rate (SAR) for the given ingress-egress-aggregate is estimated using the formula:

   SAR = U * NM-Rate

   for the latest reported interval, where U is a configurable factor greater than one which is the same for all ingress-egress-

aggregates. In effect, the value of the PCN-supportable-rate for each link is approximated by the expression

U*PCN-admissible-rate

rather than being calculated explicitly.

2. The amount of traffic to be terminated is the difference:

PCN-sent-rate - SAR,

where PCN-sent-rate is the value provided by the PCN-ingress-node.

See Section 3.3.3 for a discussion of appropriate actions if the Decision Point fails to receive a timely response to its request for the PCN-sent-rate.

If the difference calculated in the second step is positive, the Decision Point SHOULD select PCN-flows to terminate, until it determines that the PCN-traffic admission rate will no longer be greater than the estimated sustainable aggregate rate. If the Decision Point knows the bandwidth required by individual PCN-flows (e.g., from resource signalling used to establish the flows), it MAY choose to complete its selection of PCN-flows to terminate in a single round of decisions.

Alternatively, the Decision Point MAY spread flow termination over multiple rounds to avoid over-termination. If this is done, it is RECOMMENDED that enough time elapse between successive rounds of termination to allow the effects of previous rounds to be reflected in the measurements upon which the termination decisions are based. (See [Satoh10] and sections 4.2 and 4.3 of [MeLe10].)

In general, the selection of flows for termination MAY be guided by policy.

The Decision Point SHOULD log each round of termination as described in Section 5.2.1.2.

3.3.3. Decision Point Action For Missing PCN-Boundary-Node Reports

The Decision Point SHOULD start a timer t_recvFail when it receives a report from the PCN-egress-node. t_recvFail is reset each time a new report is received from the PCN-egress-node. t_recvFail expires if it reaches the value T_fail. T_fail is calculated according to the following logic:

a. T_fail = the configurable duration T_crit, if report suppression is not deployed;

b. T_fail = T_crit also if report suppression is deployed and the last report received from the PCN-egress-node contained a CLE value greater than CLE-reporting-threshold (Section 3.2.3);

c. T_fail = 3 * T_maxsuppress (Section 3.2.3) if report suppression is deployed and the last report received from the PCN-egress-node contained a CLE value less than or equal to CLE-reporting-threshold.

If timer t_recvFail expires for a given PCN-egress-node, the Decision Point SHOULD notify management. A log format is defined for that purpose in Section 5.2.1.1. Other actions depend on local policy, but MAY include blocking of new flows destined for the PCN-egress-node concerned until another report is received from it. Termination of already-admitted flows is also possible, but could be triggered by "Destination unreachable" messages received at the PCN-ingress-node.

If a centralized Decision Point sends a request for the estimated value of PCN-sent-rate to a given PCN-ingress-node and fails to receive a response in a reasonable amount of time, the Decision Point SHOULD repeat the request once. [SM-specific] If the second request to the PCN-ingress-node also fails, the Decision Point SHOULD notify management. The log format defined in Section 5.2.1.1 is also suitable for this case.

> The response timer t_sndFail with upper bound T_crit is specified in Section 3.5. The use of T_crit is an approximation. A more precise limit would be of the order of two round-trip times, plus an allowance for processing at each end, plus an allowance for variance in these values.

See Section 3.5 for suggested values of the configurable durations T_crit and T_maxsuppress.

3.4. Behaviour of the Ingress Node

The PCN-ingress-node MUST provide the estimated current rate of PCN-traffic received at that node and destined for a given ingress-egress-aggregate in octets per second (the PCN-sent-rate) when the Decision Point requests it. The way this rate estimate is derived is a matter of implementation.

> For example, the rate that the PCN-ingress-node supplies can be based on a quick sample taken at the time the information is required.

3.5. Summary of Timers and Associated Configurable Durations

Here is a summary of the timers used in the procedures just described:

t_meas

Where used: PCN-egress-node.

Used in procedure: data collection (Section 3.2.1).

Incidence: one per ingress-egress-aggregate.

Reset: immediately on expiry.

Expiry: when it reaches the configurable duration T_meas.

Action on expiry: calculate NM-rate and ETM-rate and proceed to the applicable reporting procedure (Section 3.2.2 or Section 3.2.3).

t_maxsuppress

Where used: PCN-egress-node.

Used in procedure: report suppression (Section 3.2.3).

Incidence: one per ingress-egress-aggregate.

Reset: when the next report is sent, either after expiry or because the CLE has exceeded the reporting threshold.

Expiry: when it reaches the configurable duration T_maxsuppress.

Action on expiry: send a report to the Decision Point the next time the reporting procedure (Section 3.2.3) is invoked, regardless of the value of CLE.

t_recvFail

Where used: Decision Point.

Used in procedure: failure detection (Section 3.3.3).

Incidence: one per ingress-egress-aggregate.

Reset: when a report is received for the ingress-egress-aggregate.

Expiry: when it reaches the calculated duration T_fail. As described in Section 3.3.3, T_fail is equal either to the configured duration T_crit or to the calculated value 3 * T_maxsuppress, where T_maxsuppress is a configured duration.

Action on expiry: notify management, and possibly other actions.

t_sndFail

Where used: centralized Decision Point.

Used in procedure: failure detection (Section 3.3.3).

Incidence: only as required, one per outstanding request to a PCN-ingress-node.

Started: when a request for the value of PCN-sent-traffic for a given ingress-egress-aggregate is sent to the PCN-ingress-node.

Terminated without action: when a response is received before expiry.

Expiry: when it reaches the configured duration T_crit.

Action on expiry: as described in Section 3.3.3.

3.5.1. Recommended Values For the Configurable Durations

The timers just described depend on three configurable durations, T_meas, T_maxsuppress, and T_crit. The recommendations given below for the values of these durations are all related to the intended PCN reaction time of 1 to 3 seconds. However, they are based on judgement rather than operational experience or mathematical derivation.

The value of T_meas is RECOMMENDED to be of the order of 100 to 500 ms to provide a reasonable tradeoff between demands on network resources (PCN-egress-node and Decision Point processing, network bandwidth) and the time taken to react to impending congestion.

The value of T_maxsuppress is RECOMMENDED to be on the order of 3 to 6 seconds, for similar reasons to those for the choice of T_meas.

The value of T_crit SHOULD NOT be less than 3 * T_meas. Otherwise it

could cause too many management notifications due to transient conditions in the PCN-egress-node or along the signalling path. A reasonable upper bound on T_crit is in the order of 3 seconds.

4. Specification of Diffserv Per-Domain Behaviour

This section provides the specification required by [RFC3086] for a per-domain behaviour.

4.1. Applicability

This section quotes [RFC5559].

The PCN SM boundary node behaviour specified in this document is applicable to inelastic traffic (particularly video and voice) where quality of service for admitted flows is protected primarily by admission control at the ingress to the domain.

In exceptional circumstances (e.g., due to rerouting as a result of network failures) already-admitted flows may be terminated to protect the quality of service of the remaining flows. [SM-specific] The performance results in, e.g., [MeLe10], indicate that the SM boundary node behaviour is more likely to terminate too many flows under such circumstances than the CL boundary node behaviour described in [RFCyyyy].

[RFC EDITOR'S NOTE: please replace RFCyyyy above by the reference to the published version of draft-ietf-pcn-cl-edge-behaviour.]

4.2. Technical Specification

4.2.1. Classification and Traffic Conditioning

Packet classification and treatment at the PCN-ingress-node is described in Section 5.1 of [ID.pcn-3-in-1].

PCN packets are further classified as belonging or not belonging to an admitted flow. PCN packets not belonging to an admitted flow are "blocked". (See Section 1 for an understanding of how this term is interpreted.) Packets belonging to an admitted flow are policed to ensure that they adhere to the rate or flowspec that was negotiated during flow admission.

4.2.2. PHB Configuration

The PCN SM boundary node behaviour is a metering and marking behaviour rather than a scheduling behaviour. As a result, while the

encoding uses a single DSCP value, that value can vary from one deployment to another. The PCN working group suggests using admission control for the following service classes (defined in [RFC4594]):

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

For a fuller discussion, see Appendix A of [ID.pcn-3-in-1].

4.3. Attributes

The purpose of this per-domain behaviour is to achieve low loss and jitter for the target class of traffic. The design requirement for PCN was that recovery from overloads through the use of flow termination should happen within 1-3 seconds. PCN probably performs better than that.

4.4. Parameters

The set of parameters that needs to be configured at each PCN-node and at the Decision Point is described in Section 5.1.

4.5. Assumptions

It is assumed that a specific portion of link capacity has been reserved for PCN-traffic.

4.6. Example Uses

The PCN SM behaviour may be used to carry real-time traffic, particularly voice and video.

4.7. Environmental Concerns

The PCN SM per-domain behaviour could theoretically interfere with the use of end-to-end ECN due to reuse of ECN bits for PCN marking. Section 5.1 of [ID.pcn-3-in-1] describes the actions that can be taken to protect ECN signalling. Appendix B of that document provides further discussion of how ECN and PCN can co-exist.

4.8. Security Considerations

Please see the security considerations in [RFC5559] as well as those in [RFC2474] and [RFC2475].

5. Operational and Management Considerations

5.1. Deployment of the SM Edge Behaviour

Deployment of the PCN Single Marking edge behaviour requires the following steps:

- selection of deployment options and global parameter values;
- derivation of per-node and per-link information;
- installation, but not activation, of parameters and policies at all of the nodes in the PCN domain;
- activation and verification of all behaviours.

5.1.1. Selection of Deployment Options and Global Parameters

The first set of decisions affects the operation of the network as a whole. To begin with, the operator needs to make basic design decisions such as whether the Decision Point is centralized or collocated with the PCN-ingress-nodes, and whether per-flow and aggregate resource signalling as described in [I-D.tsvwg-rsvp-pcn] is deployed in the network. After that, the operator needs to decide:

- whether PCN packets will be forwarded unencapsulated or in tunnels between the PCN-ingress-node and the PCN-egress-node. Encapsulation preserves incoming ECN settings and simplifies the PCN-egress-node’s job when it comes to relating incoming packets to specific ingress-egress-aggregates, but lowers the path MTU and imposes the extra labour of encapsulation/decapsulation on the PCN-edge-nodes.
- which service classes will be subject to PCN control and what Diffserv code point (DSCP) will be used for each. (See [ID.pcn-3-in-1] Appendix A for advice on this topic.)
- the markings to be used at all nodes in the PCN domain to indicate Not-Marked (NM) and Excess-Traffic-Marked (ETM) PCN packets;
- The marking rules for re-marking PCN-traffic leaving the PCN domain;

- whether PCN-based flow admission is enabled;
- whether PCN-based flow termination is enabled.

The following parameters affect the operation of PCN itself. The operator needs to choose:

- the value of CLE-limit if PCN-based flow admission is enabled. [SM-specific] It is RECOMMENDED that the CLE-limit for SM be set fairly low, in the order of 0.05.
- the value of the collection interval T_meas. For a recommended range of values see Section 3.5.1 above.
- whether report suppression is to be enabled at the PCN-egress-nodes and if so, the values of CLE-reporting-threshold and T_maxsuppress. It is reasonable to leave CLE-reporting-threshold at its default value (zero, as specified in Section 3.2.3). For a recommended range of values of T_maxsuppress see Section 3.5.1 above.
- the value of the duration T_crit, which the Decision Point uses in deciding whether communications with a given PCN-edge-node have failed. For a recommended range of values of T_crit see Section 3.5.1 above.
- [SM-specific] The factor U that is used in the flow termination procedure (Section 3.3.2). An operational definition for U is given in that section, but it may be thought of as a contingency factor providing a buffer to handle flow peaks above the aggregate levels expected when flows are admitted. A reasonable value for U is between 1.2 and 2. Larger values of U tend to cause more over-termination of traffic during peaks, but raise the average link utilization level.

5.1.2. Specification of Node- and Link-Specific Parameters

Filters are required at both the PCN-ingress-node and the PCN-egress-node to classify incoming PCN packets by ingress-egress-aggregate. Because of the potential use of multi-path routing in domains upstream of the PCN-domain, it is impossible to do such classification reliably at the PCN-egress-node based on the packet header contents as originally received at the PCN-ingress-node. (Packets with the same header contents could enter the PCN-domain at multiple PCN-ingress-nodes.) As a result, the only way to construct such filters reliably is to tunnel the packets from the PCN-ingress-node to the PCN-egress-node.

The PCN-ingress-node needs filters in order to place PCN packets into the right tunnel in the first instance, and also to satisfy requests from the Decision Point for admission rates into specific ingress-egress-aggregates. These filters select the PCN-egress-node, but not necessarily a specific path through the network to that node. As a result, they are likely to be stable even in the face of failures in the network, except when the PCN-egress-node itself becomes unreachable. If all PCN packets will be tunneled, the PCN-ingress-node also needs to know the address of the peer PCN-egress-node associated with each filter.

Operators may wish to give some thought to the provisioning of alternate egress points for some or all ingress-egress aggregates in case of failure of the PCN-egress-node. This could require the setting up of standby tunnels to these alternate egress points.

Each PCN-egress-node needs filters to classify incoming PCN packets by ingress-egress-aggregate, in order to gather measurements on a per-aggregate basis. If tunneling is used, these filters are constructed on the basis of the identifier of the tunnel from which the incoming packet has emerged (e.g. the source address in the outer header if IP encapsulation is used). The PCN-egress-node also needs to know the address of the Decision Point to which it sends reports for each ingress-egress-aggregate.

A centralized Decision Point needs to have the address of the PCN-ingress-node corresponding to each ingress-egress-aggregate. Security considerations require that information also be prepared for a centralized Decision Point and each PCN-edge-node to allow them to authenticate each other.

Turning to link-specific parameters, the operator needs to derive a value for the PCN-admissible-rate on each link in the network. The first two paragraphs of Section 5.2.2 of [RFC5559] discuss how these values may be derived. ([SM-specific] Confusingly, "PCN-admissible-rate" in the present context corresponds to "PCN-threshold-rate" in the cited paragraphs.)

5.1.3. Installation of Parameters and Policies

As discussed in the previous two sections, every PCN node needs to be provisioned with a number of parameters and policies relating to its behaviour in processing incoming packets. The Diffserv MIB [RFC3289] can be useful for this purpose, although it needs to be extended in some cases. This MIB covers packet classification, metering, counting, policing and dropping, and marking. The required extensions specifically include an encapsulation action following re-classification by ingress-egress-aggregate. In addition, the MIB has

to be extended to include objects for marking the ECN field in the outer header at the PCN-ingress-node and an extension to the classifiers to include the ECN field at PCN-interior and PCN-egress-nodes. Finally, a new object may need to be defined at the PCN-interior-nodes to represent the packet-size-independent excess-traffic-marking metering algorithm.

The value for the PCN-admissible-rate on each link on a node appears as a metering parameter. Operators should take note of the need to deploy excess-traffic meters either on the ingress side or the egress of each interior link, but not both (Appendix B.2 of [RFC5670].

The following additional information has to be configured by other means (e.g., additional MIBs, NETCONF models).

At the PCN-egress-node:

- the measurement interval T_meas (units of ms, range 50 to 1000);
- whether report suppression is to be applied;
- if so, the interval T_maxsuppress (units of 100 ms, range 1 to 100) and the CLE-reporting-threshold (units of tenths of one percent, range 0 to 1000, default value 0);
- the address of the PCN-ingress-node for each ingress-egress-aggregate, if the Decision Point is collocated with the PCN-ingress-node and [I-D.tsvwg-rsvp-pcn] is not deployed.
- the address of the centralized Decision Point to which it sends its reports, if there is one.

At the Decision Point:

- whether PCN-based flow admission is enabled;
- whether PCN-based flow termination is enabled.
- the value of CLE-limit (units of tenths of one percent, range 0 to 1000);
- [SM-specific] the value of the factor U used in the flow termination procedure;
- the value of the interval T_crit (units of 100 ms, range 1 to 100);

- o whether report suppression is to be applied;
- o if so, the interval T_maxsuppress (units of 100 ms, range 1 to 100) and the CLE-reporting-threshold (units of tenths of one percent, range 0 to 1000, default value 0). These MUST be the same values that are provisioned in the PCN-egress-nodes;
- o if the Decision Point is centralized, the address of the PCN-ingress-node (and any other information needed to establish a security association) for each ingress-egress-aggregate.

Depending on the testing strategy, it may be necessary to install the new configuration data in stages. This is discussed further below.

5.1.4. Activation and Verification of All Behaviours

It is certainly not within the scope of this document to advise on testing strategy, which operators undoubtedly have well in hand. Quite possibly an operator will prefer an incremental approach to activation and testing. Implementing the PCN marking scheme at PCN-ingress-nodes, corresponding scheduling behaviour in downstream nodes, and re-marking at the PCN-egress-nodes is a large enough step in itself to require thorough testing before going further.

Testing will probably involve the injection of packets at individual nodes and tracking of how the node processes them. This work can make use of the counter capabilities included in the Diffserv MIB. The application of these capabilities to the management of PCN is discussed in the next section.

5.2. Management Considerations

This section focuses on the use of event logging and the use of counters supported by the Diffserv MIB [RFC3289] for the various monitoring tasks involved in management of a PCN network.

5.2.1. Event Logging In the PCN Domain

It is anticipated that event logging using SYSLOG [RFC5424] will be needed for fault management and potentially for capacity management. Implementations MUST be capable of generating logs for the following events:

- o detection of loss of contact between a Decision Point and a PCN-edge-node, as described in Section 3.3.3;
- o successful receipt of a report from a PCN-egress-node, following detection of loss of contact with that node;

- o flow termination events.

All of these logs are generated by the Decision Point. There is a strong likelihood in the first and third cases that the events are correlated with network failures at a lower level. This has implications for how often specific event types should be reported, so as not to contribute unnecessarily to log buffer overflow. Recommendations on this topic follow for each event report type.

The field names (e.g., HOSTNAME, STRUCTURED-DATA) used in the following subsections are defined in [RFC5424].

5.2.1.1. Logging Loss and Restoration of Contact

Section 3.3.3 describes the circumstances under which the Decision Point may determine that it has lost contact, either with a PCN-ingress-node or a PCN-egress-node, due to failure to receive an expected report. Loss of contact with a PCN-ingress-node is a case primarily applicable when the Decision Point is in a separate node. However, implementations MAY implement logging in the collocated case if the implementation is such that non-response to a request from the Decision Point function can occasionally occur due to processor load or other reasons.

The log reporting the loss of contact with a PCN-ingress-node or PCN-egress-node MUST include the following content:

- o The HOSTNAME field MUST identify the Decision Point issuing the log.

- o A STRUCTURED-DATA element MUST be present, containing parameters identifying the node for which an expected report has not been received and the type of report lost (ingress or egress). It is RECOMMENDED that the SD-ID for the STRUCTURED-DATA element have the form "PCNNode" (without the quotes), which has been registered with IANA. The node identifier PARAM-NAME is RECOMMENDED to be "ID" (without the quotes). The identifier itself is subject to the preferences expressed in Section 6.2.4 of [RFC5424] for the HOSTNAME field. The report type PARAM-NAME is RECOMMENDED to be "RTyp" (without the quotes). The PARAM-VALUE for the RTyp field MUST be either "ingr" or "egr".

The following values are also RECOMMENDED for the indicated fields in this log, subject to local practice:

- o PRI initially set to 115, representing a Facility value of (14) "log alert" and a Severity level of (3) "Error Condition". Note that loss of contact with a PCN-egress-node implies that no new

flows will be admitted to one or more ingress-egress-aggregates until contact is restored. The reason a higher severity level (lower value) is not proposed for the initial log is because any corrective action would probably be based on alerts at a lower subsystem level.

- o APPNAME set to "PCN" (without the quotes).
- o MSGID set to "LOST" (without the quotes).

If contact is not regained with a PCN-egress-node in a reasonable period of time (say, one minute), the log SHOULD be repeated, this time with a PRI value of 113, implying a Facility value of (14) "log alert" and a Severity value of (1) "Alert: action must be taken immediately". The reasoning is that by this time, any more general conditions should have been cleared, and the problem lies specifically with the PCN-egress-node concerned and the PCN application in particular.

Whenever a loss-of-contact log is generated for a PCN-egress-node, a log indicating recovery SHOULD be generated when the Decision Point next receives a report from the node concerned. The log SHOULD have the same content as just described for the loss-of-contact log, with the following differences:

- o PRI changes to 117, indicating a Facility value of (14) "log alert" and a Severity of (5) "Notice: normal but significant condition".
- o MSGID changes to "RECVD" (without the quotes).

5.2.1.2. Logging Flow Termination Events

Section 3.3.2 describes the process whereby the Decision Point decides that flow termination is required for a given ingress-egress-aggregate, calculates how much flow to terminate, and selects flows for termination. This section describes a log that SHOULD be generated each time such an event occurs. (In the case where termination occurs in multiple rounds, one log SHOULD be generated per round.) The log may be useful in fault management, to indicate the service impact of a fault occuring in a lower-level subsystem. In the absence of network failures, it may also be used as an indication of an urgent need to review capacity utilization along the path of the ingress-egress-aggregate concerned.

The log reporting a flow termination event MUST include the following content:

- The HOSTNAME field MUST identify the Decision Point issuing the log.

- A STRUCTURED-DATA element MUST be present, containing parameters identifying the ingress and egress nodes for the ingress-egress-aggregate concerned, indicating the total amount of flow being terminated, and giving the number of flows terminated to achieve that objective.

  It is RECOMMENDED that the SD-ID for the STRUCTURED-DATA element have the form: "PCNTerm" (without the quotes), which has been registered with IANA. The parameter identifying the ingress node for the ingress-egress-aggregate is RECOMMENDED to have PARAM-NAME "IngrID" (without the quotes). This parameter MAY be omitted if the Decision Point is collocated with that PCN-ingress-node. The parameter identifying the egress node for the ingress-egress-aggregate is RECOMMENDED to have PARAM-NAME "EgrID" (without the quotes). Both identifiers are subject to the preferences expressed in Section 6.2.4 of [RFC5424] for the HOSTNAME field.

  The parameter giving the total amount of flow being terminated is RECOMMENDED to have PARAM-NAME "TermRate" (without the quotes). The PARAM-VALUE MUST be the target rate as calculated according to the procedures of Section 3.3.2, as an integer value in thousands of octets per second. The parameter giving the number of flows selected for termination is RECOMMENDED to have PARAM-NAME "FCnt" (without the quotes). The PARAM-VALUE for this parameter MUST be an integer, the number of flows selected.

The following values are also RECOMMENDED for the indicated fields in this log, subject to local practice:

- PRI initially set to 116, representing a Facility value of (14) "log alert" and a Severity level of (4) "Warning: warning conditions".

- APPNAME set to "PCN" (without the quotes).

- MSGID set to "TERM" (without the quotes).

5.2.2. Provision and Use of Counters

The Diffserv MIB [RFC3289] allows for the provision of counters along the various possible processing paths associated with an interface and flow direction. It is RECOMMENDED that the PCN-nodes be instrumented as described below. It is assumed that the cumulative counts so obtained will be collected periodically for use in debugging, fault management, and capacity management.

PCN-ingress-nodes SHOULD provide the following counts for each ingress-egress-aggregate. Since the Diffserv MIB installs counters by interface and direction, aggregation of counts over multiple interfaces may be necessary to obtain total counts by ingress-egress-aggregate. It is expected that such aggregation will be performed by a central system rather than at the PCN-ingress-node.

- total PCN packets and octets received for that ingress-egress-aggregate but dropped;
- total PCN packets and octets admitted to that aggregate.

PCN-interior-nodes SHOULD provide the following counts for each interface, noting that a given packet MUST NOT be counted more than once as it passes through the node:

- total PCN packets and octets dropped;
- total PCN packets and octets forwarded without re-marking;
- total PCN packets and octets re-marked to Excess-Traffic-Marked.

PCN-egress-nodes SHOULD provide the following counts for each ingress-egress-aggregate. As with the PCN-ingress-node, so with the PCN-egress-node it is expected that any necessary aggregation over multiple interfaces will be done by a central system.

- total Not-Marked PCN packets and octets received;
- total Excess-Traffic-Marked PCN packets and octets received.

The following continuously cumulative counters SHOULD be provided as indicated, but require new MIBs to be defined. If the Decision Point is not collocated with the PCN-ingress-node, the latter SHOULD provide a count of the number of requests for PCN-sent-rate received from the Decision Point and the number of responses returned to the Decision Point. The PCN-egress-node SHOULD provide a count of the number of reports sent to each Decision Point. Each Decision Point SHOULD provide the following:

- total number of requests for PCN-sent-rate sent to each PCN-ingress-node with which it is not collocated;
- total number of reports received from each PCN-egress-node;
- total number of loss-of-contact events detected for each PCN-boundary-node;

- o total cumulative duration of "block" state in hundreds of milliseconds for each ingress-egress-aggregate;
- o total number of rounds of flow termination exercised for each ingress-egress-aggregate.

6. Security Considerations

[RFC5559] provides a general description of the security considerations for PCN. This memo introduces one new consideration, related to the use of a centralized Decision Point. The Decision Point itself is a trusted entity. However, its use implies the existence of an interface on the PCN-ingress-node through which communication of policy decisions takes place. That interface is a point of vulnerability which must be protected from denial of service attacks.

7. IANA Considerations

This memo includes no request to IANA.

8. Acknowledgements

Ruediger Geib, Philip Eardley, and Bob Briscoe have helped to shape the present document with their comments. Toby Moncaster gave a careful review to get it into shape for Working Group Last Call.

Amongst the authors, Michael Menth deserves special mention for his constant and careful attention to both the technical content of this document and the manner in which it was expressed.

David Harrington’s careful AD review resulted not only in necessary changes throughout the document, but also the addition of the operations and management considerations (Section 5).

Finally, reviews by Joel Halpern and Brian Carpenter helped to clarify how ingress-egress-aggregates are distinguished (Joel) and handling of packets that cannot be carried successfully as PCN-packets (Brian). They also made other suggestions to improve the document, as did Stephen Farrell, Sean Turner, and Pete Resnick.

9. References

9.1. Normative References

[ID.pcn-3-in-1]
Briscoe, B., Moncaster, T., and M. Menth, "Encoding 3 PCN-States in the IP header using a single DSCP", March 2012.

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

[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.

[RFC2475] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., and W. Weiss, "An Architecture for Differentiated Services", RFC 2475, December 1998.

[RFC3086] Nichols, K. and B. Carpenter, "Definition of Differentiated Services Per Domain Behaviors and Rules for their Specification", RFC 3086, April 2001.

[RFC3289] Baker, F., Chan, K., and A. Smith, "Management Information Base for the Differentiated Services Architecture", RFC 3289, May 2002.

[RFC5424] Gerhards, R., "The Syslog Protocol", RFC 5424, March 2009.

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

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

9.2. Informative References

[I-D.tsvwg-rsvp-pcn]
Karagiannis, G. and A. Bhargava, "Generic Aggregation of Resource ReSerVation Protocol (RSVP) for IPv4 And IPv6 Reservations over PCN domains (Work in progress)", July 2011.

[MeLe10] Menth, M. and F. Lehrieder, "PCN-Based Measured Rate Termination", Computer Networks Journal (Elsevier) vol. 54, no. 13, pages 2099 - 2116, September 2010.

[MeLe12] Menth, M. and F. Lehrieder, "Performance of PCN-Based Admission Control under Challenging Conditions, IEEE/ACM

Transactions on Networking, vol. 20, no. 2", April 2012.

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

[RFCyyyy] Charny, A., Karagiannis, G., Menth, M., Huang, F., and T. Taylor, "PCN Boundary Node Behaviour for the Controlled Load (CL) Mode of Operation (Work in progress)", February 2012.

[Satoh10] Satoh, D. and H. Ueno, ""Cause and Countermeasure of Overtermination for PCN-Based Flow Termination", Proceedings of IEEE Symposium on Computers and Communications (ISCC '10), pp. 155-161, Riccione, Italy", June 2010.

Authors' Addresses

Anna Charny
Cisco Systems
USA

Email: anna@mwsm.com

Xinyan (Joy) Zhang
Cisco Systems
300 Apollo Drive
Chelmsford, MA 01824
USA

Email: joyzhang@cisco.com

Georgios Karagiannis
U. Twente

Phone:
Email: karagian@cs.utwente.nl

Michael Menth
University of Tuebingen
Sand 13
Tuebingen D-72076
Germany

Phone: +49-7071-2970505
Email: menth@informatik.uni-tuebingen.de

Tom Taylor (editor)
Huawei Technologies
Ottawa, Ontario
Canada

Email: tom.taylor.stds@gmail.com

Congestion and Pre-Congestion
Internet-Draft
Intended status: Experimental
Expires: September 1, 2011

M. Menth
University of Tuebingen
R. Geib
Deutsche Telekom
February 28, 2011

Admission Control Using PCN-Marked Signaling
draft-menth-pcn-marked-signaling-ac-00

Abstract

Pre-congestion notification (PCN) is a means for protecting quality of service for inelastic traffic admitted to a Diffserv domain. The overall PCN architecture is described in RFC5559. This memo is one of a series describing possible boundary node behaviours for a PCN domain.

This document proposes an admission control method. It assumes that PCN nodes perform threshold-marking configured with the PCN-admissible-rate on any link. The PCN marking state of an initial signaling message of a flow is used to determine whether the flow should be admitted or blocked.

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 1, 2011.

Table of Contents

1. Introduction

The objective of Pre-Congestion Notification (PCN) is to protect the quality of service (QoS) of inelastic flows within a Diffserv domain, in a simple, scalable, and robust fashion. Two mechanisms are used: admission control, to decide whether to admit or block a new flow request, and flow termination to decide whether to terminate some already admitted flows during serious congestion. To achieve this, the overall rate of PCN-traffic is metered on every link in the domain, and PCN-packets are appropriately remarked when certain configured rates are exceeded. These configured rates are below the rate of the link thus providing notification to boundary nodes about overloads before any congestion occurs (hence the "pre" part of pre-congestion notification). For more details see [RFC5559].

This document presents PCN-marked signaling as a method to perform admission control based on PCN information. It requires that all PCN-ingress-nodes perform threshold marking [RFC5670] configured with the PCN-admissible-rate as reference rate, and uses the marking state of initial signaling messages to determine whether flows should be admitted or blocked. It neither describes a corresponding flow termination behavior nor does it preclude flow termination.

The proposed method has several benefits: it does not require any measurement, it blocks very quickly as soon as pre-congestion occurs [Menth08-Sub-8], and it works well with multipath routing if signaling messages are carried on the same path as future data packets.

1.1. Terminology

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].

The terminology defined in [RFC5559] applies.

2. Assumed Core Network Behaviour for PCN-marked signaling

Admission control using PCN-marked signaling requires that nodes of a PCN-domain perform threshold marking [RFC5670]. The reference rate must be set to the PCN-admissible-rate of a link. Either Baseline Encoding [RFC5696] or 3-in-1 Encoding [I-D.ietf-pcn-3-in-1-encoding] may be used to distinguish re-marked signaling packets from unmarked signaling packets.

3. Edge Node Behaviours

This section explains the behavior of PCN-ingress-nodes and PCN-egress-nodes.

3.1. Prerequisites

PCN-marked signaling assumes that admission control is triggered by a signaling message at the PCN-ingress-node and that this signaling message is carried across the PCN-domain to the PCN-egress-node on the same path as future data packets of the associated flow. These signaling messages are processed only by PCN-ingress-nodes and PCN-egress-nodes. An example for such a signaling is the Resource ReServation Protocol [RFC2205]. PCN-marked signaling is relatively simple to implement when either PCN-ingress-node or PCN-egress-node are involved in the signaling anyway.

3.2. Behavior of PCN-Ingress-Nodes

The PCN-ingress-node re-marks signaling messages to PCN not-marked (NM) so that they are subject to metering and re-marking by PCN-interior-nodes. Note that signaling packets need to be marked as PCN not-marked (NM) only as long as the flow is not yet admitted.

In case of RSVP, the PCN-ingress-node performs the following non-standard actions. If the PCN-ingress-node receives a PATH message, it re-marks it to NM. If the PCN-ingress-node receives an initial RESV message, it admits the flow for the hop over the PCN domain and forwards the RESV message to the previous RSVP-hop on the path.

3.3. Behavior of PCN-Egress-Nodes

The PCN-egress-node detects signaling messages. As long as the flow is not yet admitted, the PCN-egress-node evaluates the PCN codepoint of received signaling messages. If the codepoint is NM, it takes actions so that the flow can be admitted; otherwise it takes actions so that the flow will be blocked. Finally, the PCN-egress-node resets the PCN codepoint to not-PCN.

In case of RSVP, the PCN-egress-node performs the following non-standard actions. If the PCN-egress-node receives an initial not-marked PATH message, the PCN-egress-node forwards the message as usual. If the PCN-egress-node receives an initial re-marked PATH message, the PCN-egress-node drops the PATH message and returns a PATH TEAR message to the previous RSVP hop indicating insufficient resources.

4. IANA Considerations

This document makes no request to IANA.

5. Security Considerations

Please see the security considerations in [RFC2205], [RFC2474], and [RFC2475]. [RFC5559] provides a general description of the security considerations for PCN.

6. Conclusions

The PCN-based admission control method proposed in this document has several benefits. It does not require any measurement and does not require any parameters except for threshold metering and re-marking. Implicit probing blocks very quickly as soon as pre-congestion occurs [Menth08-Sub-8] and leads to less over-admission than PCN-based admission control that calculates congestion level estimates per ingress-egress aggregate to derive admission decisions. Moreover, Implicit Probing works well with multipath routing when the signaling message is carried on the same path as future data packets [Menth08-Sub-8]. Admission control using PCN-marked signaling as proposed in this document is simple provided that the admission of flows is requested by a path-coupled signaling protocol (e.g. RSVP). In contrast to other approaches [I-D.ietf-pcn-cl-edge-behaviour], [I-D.ietf-pcn-sm-edge-behaviour] PCN-egress-nodes neither need to measure PCN traffic nor need to signal PCN-feedback. In particular, pcn-egress-nodes no longer need to map packets to corresponding ingress-egress-aggregates. Moreover, the presented method blocks very quickly as soon as pre-congestion occurs [Menth08-Sub-8], minimizing over-admission. It also works well with multipath routing if signaling messages are carried on the same path as future data packets [Menth08-Sub-8], minimizing under-admission.

7. Acknowledgements

Joe Babiarz presented the idea documented in this memo for the first time in [I-D.babiarz-pcn-3sm]. It was further developed to be useful for restricted tunneling rules which called for a special encoding [I-D.ietf-pcn-psdm-encoding], [I-D.menth-pcn-psdm-deployment], [Menth09f].

8. References

8.1. Normative References

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

[RFC2205] Braden, B., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, September 1997.

[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.

[RFC2475] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., and W. Weiss, "An Architecture for Differentiated Services", RFC 2475, December 1998.

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

[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.

8.2. Informative References

[I-D.babiarz-pcn-3sm]
Babiarz, J., Liu, X., Chan, K., and M. Menth, "Three State PCN Marking", draft-babiarz-pcn-3sm-01 (work in progress), November 2007.

[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-04 (work in progress), January 2011.

[I-D.ietf-pcn-cl-edge-behaviour]
Charny, A., Huang, F., Karagiannis, G., Menth, M., and T. Taylor, "PCN Boundary Node Behaviour for the Controlled Load (CL) Mode of Operation", draft-ietf-pcn-cl-edge-behaviour-08 (work in progress), December 2010.

[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.

[I-D.ietf-pcn-sm-edge-behaviour]
Charny, A., Karagiannis, G., Menth, M., and T. Taylor, "PCN Boundary Node Behaviour for the Single Marking (SM) Mode of Operation", draft-ietf-pcn-sm-edge-behaviour-05 (work in progress), December 2010.

[I-D.menth-pcn-psdm-deployment]
Menth, M., "Deployment Models for PCN-Based Admission Control and Flow Termination Using Packet-Specific Dual Marking (PSDM)", draft-menth-pcn-psdm-deployment-00 (work in progress), October 2008.

[Menth08-Sub-8]
Menth, M. and F. Lehrieder, "Performance of PCN-Based Admission Control", currently under submission, University of Wuerzburg, Germany, 2011.

[Menth09f]
Menth, M., Babiarz, J., and P. Eardley, "Pre-Congestion Notification Using Packet-Specific Dual Marking", in Proceedings of the International Workshop on the Network of the Future (Future-Net), IEEE, Dresden, Germany, June 2009.

Authors' Addresses

Michael Menth
University of Tuebingen
Department of Computer Science
Chair of Communication Networks
Sand 13
Tuebingen 72076
Germany

Phone: +49 7071 29 70505
Email: menth@informatik.uni-tuebingen.de

Ruediger Geib
Deutsche Telekom
Heinrich-Hertz-Strasse 3-7
Darmstadt 64295
Germany

Phone: +49 6151 628 2747
Email: ruediger.geib@telekom.de