

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

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

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

Table of Contents

1. Introduction	5
1.1. Terminology	6
2. [CL-Specific] Assumed Core Network Behaviour for CL	9
3. Node Behaviours	10
3.1. Overview	10
3.2. Behaviour of the PCN-Egress-Node	11
3.2.1. Data Collection	11
3.2.2. Reporting the PCN Data	12
3.2.3. Optional Report Suppression	12
3.3. Behaviour at the Decision Point	13
3.3.1. Flow Admission	13
3.3.2. Flow Termination	14
3.3.3. Decision Point Action For Missing PCN-Boundary-Node Reports	15
3.4. Behaviour of the Ingress Node	17
3.5. Summary of Timers and Associated Configurable Durations	17
3.5.1. Recommended Values For the Configurable Durations	18
4. Specification of Diffserv Per-Domain Behaviour	19
4.1. Applicability	19
4.2. Technical Specification	19
4.2.1. Classification and Traffic Conditioning	20
4.2.2. PHB Configuration	20
4.3. Attributes	20
4.4. Parameters	20
4.5. Assumptions	20
4.6. Example Uses	21
4.7. Environmental Concerns	21
4.8. Security Considerations	21
5. Operational and Management Considerations	21
5.1. Deployment of the CL Edge Behaviour	21
5.1.1. Selection of Deployment Options and Global Parameters	21
5.1.2. Specification of Node- and Link-Specific Parameters	23
5.1.3. Installation of Parameters and Policies	24
5.1.4. Activation and Verification of All Behaviours	25
5.2. Management Considerations	26
5.2.1. Event Logging In the PCN Domain	26
5.2.1.1. Logging Loss and Restoration of Contact	26
5.2.1.2. Logging Flow Termination Events	28
5.2.2. Provision and Use of Counters	29
6. Security Considerations	30
7. IANA Considerations	30
8. Acknowledgements	31
9. References	32
9.1. Normative References	32
9.2. Informative References	32

Authors' Addresses	33
------------------------------	----

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:

- o to validate the specification of the CL edge behaviour;
- o to evaluate the effectiveness of the CL edge behaviour in preserving quality of service for admitted flows; and
- o 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]:

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

- o PCN-excess-rate;
- o PCN-admissible-rate;
- o PCN-supportable-rate;
- o PCN-marked;
- o [CL-specific] threshold-marked;
- o 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]:

- o [CL-specific] threshold-meter;
- o 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]:

- o not-PCN codepoint;
- o Not-marked (NM) codepoint;
- o [CL-specific] Threshold-marked (ThM) codepoint;
- o 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:

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

- o 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];
- o 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];
- o on each link the reference rate for the threshold-meter is configured to be equal to the PCN-admissible-rate for the link;
- o on each link the reference rate for the excess-traffic-meter is configured to be equal to the PCN-supportable-rate for the link;
- o 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`. For further information see Section 3.5.

- o NM-rate: octets per second of PCN-traffic in PCN-packets that are not-marked (i.e., marked with the NM codepoint);
- o [CL-specific] ThM-rate: octets per second of PCN-traffic in PCN-packets that are threshold-marked (i.e., marked with the ThM codepoint);
- o 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]
$$\text{CLE} = (\text{ThM-rate} + \text{ETM-rate}) / (\text{NM-rate} + \text{ThM-rate} + \text{ETM-rate})$$

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 $T_{\text{maxsuppress}}$ 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 ($T_{\text{maxsuppress}} + T_{\text{meas}}$). 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]
$$CLE = (ThM\text{-}rate + ETM\text{-}rate) / (NM\text{-}rate + ThM\text{-}rate + ETM\text{-}rate)$$

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:

$$\text{SAR} = \text{NM-rate} + \text{ThM-rate}$$

for the latest reported interval.

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

$$\text{PCN-sent-rate} - \text{SAR},$$

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]):

- 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 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.

- o 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.)
- o 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;
- o The marking rules for re-marking PCN-traffic leaving the PCN domain;
- o whether PCN-based flow admission is enabled;
- o whether PCN-based flow termination is enabled.

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

- o 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.
- o the value of the collection interval T_{meas}. For a recommended range of values see Section 3.5.1 above.
- o 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.
- o 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.
- o [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:

- o the measurement interval `T_meas` (units of ms, range 50 to 1000);
- o [CL-specific] whether specific flow identifiers must be captured when excess-traffic-marked packets are observed;
- 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);
- o 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.

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

At the Decision Point:

- o whether PCN-based flow admission is enabled;
- o whether PCN-based flow termination is enabled.
- o the value of CLE-limit (units of tenths of one percent, range 0 to 1000);
- o 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 occurring 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:

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

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

- o total PCN packets and octets received for that ingress-egress-aggregate but dropped;
- o 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:

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

- o total Not-Marked PCN packets and octets received;
- o [CL-specific] total Threshold-Marked PCN packets and octets received;
- o 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:

- o total number of requests for PCN-sent-rate sent to each PCN-ingress-node with which it is not collocated;
- o total number of reports received from each PCN-egress-node;
- o 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 document requests IANA to add the following entries to the

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

Structured Data ID: PCNNode OPTIONAL

Structured Data Parameter: ID MANDATORY

Structured Data Parameter: Rtyp MANDATORY

Reference: RFCxxxxx

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: RFCxxxxx

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@mwsn.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

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 08, 2012.

Copyright Notice

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

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

Table of Contents

1. Introduction	4
2. General PCN Encoding Requirements	4
2.1. Metering and Marking Algorithms	5
2.2. Approaches for PCN Based Admission Control and Flow Termination	5
2.2.1. Dual Marking (DM)	5
2.2.2. Single Marking (SM)	6
2.2.3. Packet Specific Dual Marking (PSDM)	7
2.2.4. Preferential Packet Dropping	8
3. Encoding Constraints	8
3.1. Structure of the DS Field	8
3.2. Constraints from the DSCP	8
3.2.1. General Scarcity of DSCPs	8
3.2.2. Handling of the DSCP in Tunneling Rules	9
3.2.3. Restoration of Original DSCPs at the Egress Node	9
3.3. Constraints from the ECN Field	10
3.3.1. Structure and Use of the ECN Field	10
3.3.2. Redefinition of the ECN Field	10
3.3.3. Handling of the ECN Field in Tunneling Rules	11
3.3.4. Restoration of the Original ECN Field at the PCN-Egress-Node	13
4. Comparison of Encoding Options	13
4.1. Baseline Encoding	14
4.2. Encoding with 1 DSCP Providing 3 States	14
4.3. Encoding with 2 DSCPs Providing 3 or More States	15
4.4. Encoding for Packet Specific Dual Marking (PSDM)	15
4.5. Standardized Encodings	15
5. Conclusion	15
6. Security Implications	16
7. IANA Considerations	16
8. Acknowledgements	16
9. References	16
9.1. Normative References	16
9.2. Informative References	16

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:

$$\text{termination-rate} = \text{PCN-ingress-rate} - (\text{rate-of-NM-traffic} + \text{rate-of-ThM-traffic})$$

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.

NM -----> ETM

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 $\text{rate-of-NM-traffic} * u < \text{rate-of-NM-traffic} + \text{rate-of-ETM-traffic}$, requests the PCN-sent-rate from the corresponding PCN-ingress-node, calculates the amount of PCN traffic to be terminated by

$$\text{termination-rate} = \text{PCN-sent-rate} - \text{rate-of-NM-traffic} * u,$$

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

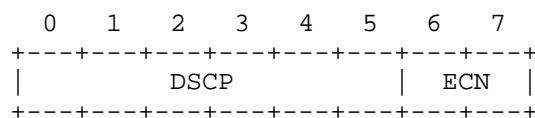
The termination algorithms described in [I-D.ietf-pcn-cl-edge behaviour] and [I-D.ietf-pcn-sm-edge-behaviour] require the preferential dropping of ETM-marked packets to avoid over-termination in the case of packet loss. An analysis explaining this phenomenon can be found in Section 4 of [Menth10q]. Thus, 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]).



DSCP: Differentiated Services codepoint [RFC2474]

ECN: ECN field [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.

ECN Bits	00	10	01	11	DSCP
RFC 3168	Not-ECT	ECT(0)	ECT(1)	CE	Any
Baseline	Not-PCN	NM	EXP	PM	PCN-n
3-In-1	Not-PCN	NM	ThM	ETM	PCN-n
3-In-2	Not-PCN	NM	CU	ThM	PCN-n
	Not-PCN	CU	CU	ETM	PCN-m
PSDM	Not-PCN	Not-ETM	Not-ThM	PM	PCN-n

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 Fairhurst 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: 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

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

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

Table of Contents

1. Introduction	5
1.1. Terminology	6
2. [SM-Specific] Assumed Core Network Behaviour for SM	9
3. Node Behaviours	10
3.1. Overview	10
3.2. Behaviour of the PCN-Egress-Node	10
3.2.1. Data Collection	10
3.2.2. Reporting the PCN Data	11
3.2.3. Optional Report Suppression	11
3.3. Behaviour at the Decision Point	12
3.3.1. Flow Admission	12
3.3.2. Flow Termination	13
3.3.3. Decision Point Action For Missing PCN-Boundary-Node Reports	14
3.4. Behaviour of the Ingress Node	15
3.5. Summary of Timers and Associated Configurable Durations	16
3.5.1. Recommended Values For the Configurable Durations	17
4. Specification of Diffserv Per-Domain Behaviour	18
4.1. Applicability	18
4.2. Technical Specification	18
4.2.1. Classification and Traffic Conditioning	18
4.2.2. PHB Configuration	18
4.3. Attributes	19
4.4. Parameters	19
4.5. Assumptions	19
4.6. Example Uses	19
4.7. Environmental Concerns	19
4.8. Security Considerations	20
5. Operational and Management Considerations	20
5.1. Deployment of the SM Edge Behaviour	20
5.1.1. Selection of Deployment Options and Global Parameters	20
5.1.2. Specification of Node- and Link-Specific Parameters	21
5.1.3. Installation of Parameters and Policies	22
5.1.4. Activation and Verification of All Behaviours	24
5.2. Management Considerations	24
5.2.1. Event Logging In the PCN Domain	24
5.2.1.1. Logging Loss and Restoration of Contact	25
5.2.1.2. Logging Flow Termination Events	26
5.2.2. Provision and Use of Counters	27
6. Security Considerations	29
7. IANA Considerations	29
8. Acknowledgements	29
9. References	30
9.1. Normative References	30
9.2. Informative References	30

Authors' Addresses	31
------------------------------	----

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:

- o to validate the specification of the SM edge behaviour;
- o to evaluate the effectiveness of the SM edge behaviour in preserving quality of service for admitted flows; and
- o 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]:

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

- o PCN-admissible-rate;
- o PCN-supportable-rate;
- o PCN-marked;
- o 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]:

- o 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]:

- o not-PCN codepoint;
- o Not-marked (NM) codepoint;
- o 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:

- o PCN-interior-nodes perform excess-traffic-marking of PCN-packets according to the rules specified in [RFC5670].
- o 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;
- o on each link the reference rate for the excess-traffic-meter is configured to be equal to the PCN-admissible-rate for the link;
- o 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.

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

- o 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]
$$\text{CLE} = \text{ETM-rate} / (\text{NM-rate} + \text{ETM-rate})$$

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 $T_{\text{maxsuppress}}$ 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 ($T_{\text{maxsuppress}} + T_{\text{meas}}$). 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]
 $CLE = ETM\text{-}rate / (NM\text{-}rate + ETM\text{-}rate)$

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:

$$SAR = U * NM\text{-}Rate$$

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

$$U * \text{PCN-admissible-rate}$$

rather than being calculated explicitly.

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

$$\text{PCN-sent-rate} - \text{SAR},$$

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 [Sato10] 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:

- 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.
- o 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.)
- o the markings to be used at all nodes in the PCN domain to indicate Not-Marked (NM) and Excess-Traffic-Marked (ETM) PCN packets;
- o The marking rules for re-marking PCN-traffic leaving the PCN domain;

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

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

- o 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.
- o the value of the collection interval T_{meas} . For a recommended range of values see Section 3.5.1 above.
- o whether report suppression is to be enabled at the PCN-egress-nodes and if so, the values of CLE-reporting-threshold and $T_{\text{maxsuppress}}$. 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 $T_{\text{maxsuppress}}$ see Section 3.5.1 above.
- o 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.
- o [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:

- o the measurement interval `T_meas` (units of ms, range 50 to 1000);
- 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);
- o 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.
- o the address of the centralized Decision Point to which it sends its reports, if there is one.

At the Decision Point:

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

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

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

- o total PCN packets and octets received for that ingress-egress-aggregate but dropped;
- o 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:

- o total PCN packets and octets dropped;
- o total PCN packets and octets forwarded without re-marking;
- o 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.

- o total Not-Marked PCN packets and octets received;
- o 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:

- o total number of requests for PCN-sent-rate sent to each PCN-ingress-node with which it is not collocated;
- o total number of reports received from each PCN-egress-node;
- o 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@mwsn.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

