Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: December 29, 2010

A. Charny
Cisco Systems
F. Huang
Huawei Technologies
G. Karagiannis
U. Twente
M. Menth
University of Wuerzburg
T. Taylor, Ed.
Huawei Technologies

June 27, 2010

PCN Boundary Node Behaviour for the Controlled Load (CL) Mode of Operation
draft-ietf-pcn-cl-edge-behaviour-06

Abstract

Precongestion notification (PCN) is a means for protecting quality of service for inelastic traffic admitted to a Diffserv domain. The overall PCN architecture is described in RFC 5559. This memo 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 PCN-marked, threshold-marked, and excess-traffic-marked. This behaviour is known informally as the Controlled Load (CL) PCN edge behaviour.

Status of this Memo

Copyright Notice

Table of Contents

1. Introduction

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

Boundary node behaviours specify a detailed set of algorithms and edge node behaviours used to implement the PCN mechanisms. Since the algorithms depend on specific metering and marking behaviour at the interior nodes, it is also necessary to specify the assumptions made about interior node behaviour. Finally, because PCN uses DSCP values to carry its markings, a specification of boundary node behaviour must include the per domain behaviour (PDB) template specified in [RFC3086], filled out with the appropriate content. The present document accomplishes these tasks for the controlled load (CL) mode of operation.

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 [RFC2119].

In addition to the terms defined in [RFC5559], this document uses the following terms:

Decision Point
: The node that makes the decision about which flows to admit and to terminate. In a given network deployment, this may be the ingress node or a centralized control node. Regardless of the location of the Decision Point, the ingress node is the point where the decisions are enforced.

NM-rate
: rate of not-marked PCN traffic in octets per second. For further details see Section 3.2.1.

ThM-rate
rate of threshold-marked PCN traffic in octets per second. For further details see Section 3.2.1.

ETM-rate
rate of excess-traffic-marked PCN traffic in octets per second. For further details see Section 3.2.1.

Congestion level estimate (CLE)
A value derived from the measurement of PCN packets received at a PCN-egress-node for a given ingress-egress-aggregate, representing the ratio of marked to total PCN traffic (measured in octets) over a short period. For further details see Section 3.2.4.

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.

Admission decision threshold
A fractional value to which the Decision Point compares the CLE to determine the PCN-admission-state for a given ingress-egress aggregate. If the CLE is below the admission decision threshold the PCN-admission-state is set to "admit". If the CLE is above the admission decision threshold the PCN-admission-state is set to "block". For further details see Section 3.3.1.

2. Assumed Core Network Behaviour for CL

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

- o encoding of PCN status within individual packets is based on [RFC5696], extended to provide a third PCN encoding state. Extensions for this purpose will be specified by the IETF. Possible extensions are documented in [ID.PCN3state] or alternatively [ID.PCN3in1];

- o the domain satisfies the conditions specified in the applicable encoding extension document;

- 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 PCN-interior-nodes perform threshold-marking and excess-traffic-marking of packets according to the rules specified in [RFC5670], and any additional rules specified in the applicable encoding extension document;

According to [RFC5696], the encoding extension documents should specify the allowable transitions between marking states. However, to be absolutely clear, these allowable transitions are specified here. At any interior node, the only permitted transitions are these:

- o a PCN packet which is not-marked (NM) MAY be threshold-marked (ThM) or excess-traffic-marked (ETM);

- o a PCN packet which is threshold-marked (ThM) MAY be excess-traffic-marked (ETM).

An interior node MUST NOT perform any of the following:

- o re-mark a packet from PCN to non-PCN, or from non-PCN to PCN;

- o re-mark a PCN packet from threshold-marked (ThM) to not-marked (NM);

- o re-mark a PCN packet from excess-traffic-marked (ETM) to not-marked (NM) or threshold-marked (ThM).

3. Node Behaviours

3.1. Overview

This section describes the behaviour of the PCN ingress and egress nodes and the Decision Point (which may be collocated with the ingress node). The PCN egress node collects and reports the rates of not-marked, threshold-marked, and excess-traffic-marked PCN traffic to the Decision Point. It may also identify individual 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 admitted 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 ingress and egress nodes for a given ingress-egress-aggregate. For details, see Section 3.3.

3.2. Behaviour of the PCN-Egress-Node

3.2.1. Data Collection

The PCN-egress-node MUST meter received PCN traffic in order to derive periodically the following rates for each ingress-egress-aggregate passing through it:

- o NM-rate: octets per second of PCN traffic in packets which are not PCN-marked;

- o ThM-rate: octets per second of PCN traffic in PCN-threshold-marked packets;

- o ETM-rate: octets per second of PCN traffic in PCN-excess-marked packets.

It is RECOMMENDED that the interval, Tcalc, between calculation of these quantities be in the range of 100 to 500 ms to provide a reasonable tradeoff between signalling demands on the network and the time taken to react to impending congestion.

The PCN-traffic SHOULD be metered continuously and the intervals themselves SHOULD be of equal length, to minimize the statistical variance introduced by the measurement process itself.

As a configurable option, the PCN-egress-node MAY record flow identifiers of the individual flows for which excess-traffic-marked packets have been observed. 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 flows are routed via specific paths, identifying the flows that are experiencing excess-traffic-marking helps to avoid termination of flows not contributing to congestion.

3.2.2. Reporting the PCN Data

If the report suppression option described in the next sub-section is not enabled, the PCN-egress-node MUST report the latest values of NM-rate, ThM-rate, and ETM-rate to the Decision Point each time that it

calculates them.

If so configured (e.g., because multi-path routing is being used, as explained in the previous section), the PCN-egress-node MUST also report the set of flow identifiers of flows for which excess-traffic-marking was observed in the most recent measurement interval.

3.2.3. Optional Report Suppression

Report suppression MUST be provided as a configurable option. If this option is enabled, the PCN-egress-node MUST NOT send a report to the Decision Point for a given ingress-egress-aggregate whenever all of the following conditions are satisfied:

- o ThM-rate and ETM-rate were zero in the latest interval.

- o ThM-rate and ETM-rate were zero in the next most recent interval.

- o Less than time Tmaxnorep has elapsed since the last time the PCN-egress-node sent a report to the Decision Point for the given aggregate, where Tmaxnorep is a configurable value.

The above procedure ensures that at least one report is sent per period Tmaxnorep. This provides some protection against loss of egress reports and also demonstrates to the Decision Point that both the PCN-egress-node and the communication path between the two nodes are in operation. However, depending on the transport used for reporting, the operator may choose to set Tmaxnorep to an effectively infinite value. For example, the transport may include its own keep-alive signalling at a sufficient frequency that PCN keep-alive is redundant.

3.2.4. Optional Calculation and Reporting of Congestion Level Estimate

The calculation and reporting of congestion level estimates (CLE) MUST be provided as a configurable option at the PCN-egress-node. If this option is enabled, the PCN-egress-node MUST calculate the current value for CLE for each ingress-egress-aggregate in each measurement interval and include this in its report (along with the current values of NM-rate, ThM-rate and ETM-rate). The CLE is equal to the ratio:

(ThM-Rate + ETM-Rate) / (NM-rate + ThM-rate + ETM-rate)

if any PCN traffic was observed, or zero otherwise.

3.3. Behaviour at the Decision Point

Operators may choose to deploy just flow admission, or just flow termination, or both. The Decision Point 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.

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), calculated as described in Section 3.2.4. 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. The Decision Point MUST compare the reported or calculated CLE to an admission decision threshold CLElimit. If the CLE is less than the threshold, the PCN-admission-state for that aggregate MUST be set to "admit"; otherwise it MUST be set to "block".

> The outcome of the comparison is not very sensitive to the value of CLElimit in practice, because when marking occurs it tends to persist long enough that marked traffic becomes a large proportion of the received traffic in a given interval.

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.

3.3.2. Flow Termination

When the report from the egress node includes a non-zero value of the ETM-Rate for the given ingress-egress-aggregate, the Decision Point MUST request the PCN-ingress-node to provide an estimate of the rate (Admit-Rate) at which PCN-traffic is being admitted to the aggregate.

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

The Decision Point MUST then wait, for both the requested rate from the ingress node and the next report from the egress node. If this next egress node report also includes a non-zero value for the ETM-

Rate, the Decision Point MUST determine an amount of flow to terminate in the following steps:

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

   SAR = NM-Rate + ThM-Rate

   for the latest reported interval.

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

   Admit-Rate - SAR,

   where Admit-Rate is the value provided by the ingress node.

If the difference calculated in the second step is positive, the Decision Point SHOULD select 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 flows (e.g., from resource signalling used to establish the flows), it MAY choose to complete its selection of 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 [I-D.satoh-pcn-performance-termination] and sections 4.2 and 4.3 of [Menth08-sub-9]).

If the egress node has supplied a list of flow identifiers (Section 3.2), the Decision Point SHOULD first consider terminating flows in that list. In general, the selection of flows for termination MAY be guided by policy.

3.3.3. Decision Point Action For Missing Egress Node Reports

If the Decision Point fails to receive reports from a given egress node for a configurable interval Tfail, it SHOULD cease to admit flows to that aggregate and raise an alarm to management. This provides some protection against the case where congestion is preventing the transfer of reports from the egress node to the Decision Point. If a report is subsequently received from the egress node concerned, the Decision Point MUST restart failure timing and resume making admission and termination decisions based on the

reports it receives.

3.4. Behaviour of the Ingress Node

The PCN-ingress-node MUST provide the estimated current rate of admitted PCN traffic (octets per second) for a specific ingress-egress-aggregate 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 MAY be based on a quick sample taken at the time the information is required. It is RECOMMENDED that such a sample be based on observation of at least 30 PCN packets to achieve reasonable statistical reliability.

3.5. Summary of Timers

Table 1 summarizes the timers implied by the preceding procedures. Tcol and Trep are reset upon expiry. Tmon is reset by management action or by receipt of a report from the egress node concerned.

Timer	Location	Incidence	Limit	Action on Expiry
Tcol	Egress node	One per node	Tcalc	Calculate and possibly report NM-rate, ThM-rate, ETM-rate and optionally CLE for each IEA.
-	-	-	-	-
Trep	Egress node	One per IEA if report suppression is enabled.	Tmaxnorep	Send a report for that IEA at the next expiry of Tcol.
-	-	-	-	-
Tmon	Decision point	One per egress node	Tfail	Assume failure and cease to admit flows passing through that egress node.

IEA = ingress-egress-aggregate

Table 1: Timers Used For the CL Edge Behaviour

The value of Tcalc SHOULD be configurable, and is RECOMMENDED to be

of the order of 100 to 500 ms.

Trep is active only when report suppression is enabled. The value of Tmaxnorep SHOULD be configurable. The appropriate value depends on the transport used to carry the egress node reports. For unreliable transport, Tmaxnorep is RECOMMENDED to be of the order of one second.

The value of Tfail MUST be configurable. When unreliable transport is used, the value of Tfail is RECOMMENDED to be of the order of 3 * Tmaxnorep if report suppression is enabled, and of the order of 3 * Tcalc if report suppression is not enabled. When reliable transport is used, the operator may choose to provide similar values for Tfail or may choose to disable report timing by setting an effectively infinite value for Tfail.

4. Identifying Ingress and Egress Nodes for PCN Traffic

The operation of PCN depends on the ability of the ingress node to identify the ingress-egress-aggregate to which each new flow belongs and the ability of the egress node to identify the aggregate to which each received PCN packet belongs. If the Decision Point is collocated with the ingress node, the egress node also needs to associate each aggregate with the address of the ingress node to which it must send its reports.

The means by which this is done depends on the packet routing technology in use in the network. In general, classification of individual packets at the ingress node (for enforcement and metering of admission rates) and at the egress node must use the content of the outer packet header. The process may well require configuration of routing information in the ingress and egress nodes.

5. Specification of Diffserv Per-Domain Behaviour

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

5.1. Applicability

This section draws heavily upon points made in the PCN architecture document, [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 network failures) already-admitted flows may be terminated to protect the quality of service of the remaining flows. The CL boundary node behaviour is less likely to terminate too many flows under such circumstances than the SM boundary node behaviour ([I-D.SM-edge-behaviour]).

5.2. Technical Specification

The technical specification of the PCN CL per domain behaviour is provided by the contents of [RFC5559], [RFC5696], [RFC5670], the specification of the encoding extension (e.g. [ID.PCN3state], [ID.PCN3in1]), and the present document.

5.3. Attributes

The purpose of this per-domain behaviour is to achieve low loss and jitter for the target class of traffic. Recovery from overloads through the use of flow termination should happen within 1-3 seconds.

5.4. Parameters

In the list that follows, note that most PCN-ingress-nodes are also egress nodes, and vice versa. Furthermore, the ingress nodes may be collocated with Decision Points.

Parameters at the PCN-ingress-node:

- o Filters for distinguishing PCN from non-PCN inbound traffic.
- o The DSCP(s) to be used to mark PCN traffic.
- o Reference rates on each inward link for the PCN-threshold-rate and PCN-excess-rate; see Section 2.
- o The information needed to distinguish PCN traffic belonging to a given ingress-egress-aggregate.

Parameters at the PCN-egress-node:

- o The calculation interval Tcalc.
- o Whether report suppression is enabled and, if so, the value of Tmaxnorep, the maximum interval between reports for a given ingress-egress-aggregate.
- o Whether calculation and reporting of congestion level estimates is enabled at the PCN-egress-node.

- o Whether individual flow identifiers must be reported for excess-traffic-marked PCN traffic.

- o The information needed to distinguish PCN traffic belonging to a given ingress-egress-aggregate.

- o The marking rules for re-marking PCN traffic leaving the PCN domain.

Parameters at each interior node:

- o Reference rates on each link for the PCN-threshold-rate and PCN-excess-rate; see Section 2.

Parameters at the Decision Point:

- o Activation/deactivation of PCN-based flow admission.

- o Activation/deactivation of PCN-based flow termination.

- o The admission decision threshold CLElimit.

- o The maximum interval Tfail between reports from a given egress node, for detecting failure of communications with that node.

- o The information needed to map between each ingress-egress-aggregate and its edgepoints, particularly the corresponding ingress node.

5.5. Assumptions

Assumed that a specific portion of link capacity has been reserved for PCN traffic.

5.6. Example Uses

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

5.7. Environmental Concerns

The PCN CL per-domain behaviour may interfere with the use of end-to-end ECN due to reuse of ECN bits for PCN marking. See the applicable PCN marking specifications for details.

5.8. Security Considerations

Please see the security considerations in Section 6 as well as those in [RFC2474] and [RFC2475].

6. Security Considerations

[RFC5559] provides a general description of the security considerations for PCN. This memo introduces no new considerations.

7. IANA Considerations

This memo includes no request to IANA.

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.

9. References

9.1. Normative References

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

[RFC2474] Nichols, K., Blake, S., Baker, F., and D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC 2474, December 1998.

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

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

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

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

9.2. Informative References

[I-D.SM-edge-behaviour]
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)", June 2010.

[I-D.satoh-pcn-performance-termination]
Satoh, D., Ueno, H., and M. Menth, "Performance Evaluation of Termination in CL-Algorithm (Work in progress)", July 2009.

[ID.PCN3in1]
Briscoe, B., "PCN 3-State Encoding Extension in a single DSCP (Work in progress)", February 2010.

[ID.PCN3state]
Moncaster, T., Briscoe, B., and M. Menth, "A PCN encoding using 2 DSCPs to provide 3 or more states (Work in progress)", February 2010.

[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.

[Menth08-sub-9]
Menth, M. and F. Lehrieder, "PCN-Based Measured Rate Termination", July 2009, <http://www3.informatik.uni-wuerzburg.de/~menth/Publications/papers/Menth08-Sub-9.pdf>.

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

Authors' Addresses

Anna Charny
Cisco Systems
300 Apollo Drive
Chelmsford, MA 01824
USA

Email: acharny@cisco.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 Wuerzburg
Am Hubland
Wuerzburg D-97074
Germany

Phone: +49-931-888-6644
Email: menth@informatik.uni-wuerzburg.de

Tom Taylor (editor)
Huawei Technologies
1852 Lorraine Ave
Ottawa, Ontario K1H 6Z8
Canada

Phone: +1 613 680 2675
Email: tom111.taylor@bell.net