Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: January 8, 2010

A. Charny
J. Zhang
Cisco Systems
G. Karagiannis
U. Twente
M. Menth
University of Wuerzburg
T. Taylor, Ed.
Huawei Technologies
July 7, 2009

PCN Boundary Node Behaviour for the Single Marking (SM) Mode of Operation
draft-ietf-pcn-sm-edge-behaviour-00

Status of this Memo

Copyright Notice

and restrictions with respect to this document.

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 two-state measurement-based load control, known informally as Single Marking (SM).

Table of Contents

1. Introduction 3
1.1. Terminology 3
2. Assumed Core Network Behaviour for SM 4
3. Node Behaviours 5
3.1. Overview 5
3.2. Behaviour of the PCN-Egress-Node 5
3.2.1. PCN-Egress-Node Role In Flow Admission 6
3.2.2. PCN-Egress-Node Role In Flow Termination 6
3.3. Behaviour of the PCN-Ingress-Node 7
3.3.1. PCN-Ingress-Node Role In Flow Admission 7
3.3.2. PCN-Ingress-Node Role In Flow Termination 7
3.4. Possible Extension to the Basic Algorithm 7
4. Specification of Diffserv Per-Domain Behaviour 8
4.1. Applicability 8
4.2. Technical Specification 9
4.3. Attributes 9
4.4. Parameters 9
4.5. Assumptions 9
4.6. Example Uses 9
4.7. Environmental Concerns 9
4.8. Security Considerations 9
5. Security Considerations 9
6. IANA Considerations 10
7. Acknowledgements 10
8. References 10
8.1. Normative References 10
8.2. Informative References 10
Authors' Addresses 11

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 "pre-congestion" notification). The level of marking allows boundary nodes to make decisions about whether to admit or terminate. 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 Single Marking (SM) mode of operation.

1.1. Terminology

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

Policy Decision Point (PDP)
: The node that provides policy input regarding admission and termination of flows.

PCN-admission-state
: The state ("admit" or "block") derived by PCN-egress-node for a given ingress-egress-aggregate based on PCN packet marking statistics. The PCN-ingress-node admits or blocks new flows offered to the aggregate based on the current value of the PCN-admission-state. Individual decisions may be modified by policy input from the PDP. For further details see Section 3.2.1 and Section 3.3.1.

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. In this specification the CLE is an exponentially weighted moving average of the ratios observed in successive fixed-length measurement intervals. For further details see Section 3.2.1.

Admission decision threshold
A fractional value to which the CLE is compared to determine the PCN-admission-state. 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 3.2.1.

Normal regime
The operating state of the PCN-egress-node with respect to a given ingress-egress-aggregate during periods when no excess-traffic-marked packets are received within that aggregate.

Excess traffic regime
The operating state of the PCN-egress-node with respect to a given ingress-egress-aggregate during periods when excess-traffic-marked packets are being received within that aggregate. The transition from normal to excess traffic regime occurs when an excess-traffic-marked packet is received within the given ingress-egress-aggregate. The transition from excess traffic regime to normal regime occurs when a complete measurement interval passes without receipt of an excess-traffic-marked packet within the given ingress-egress-aggregate. For further details see Section 3.2.2.

2. Assumed Core Network Behaviour for SM

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

- each link has been configured with a PCN-excess-rate having a value equal to the PCN-admissible-rate for the link;
- PCN-interior-nodes perform excess-traffic-metering of packets according to the rules specified in [ID.PCN-marking].
- excess-traffic-marking of packets uses the PCN-Marked (PM) codepoint defined in [ID.PCN-baseline];

- o no link PCN-threshold-rate is configured, and PCN-interior nodes perform no threshold-metering.

3. Node Behaviours

3.1. Overview

The Single Marking (SM) mode of operation supports flow admission based on the smoothed ratio of PCN-marked to total PCN-traffic observed by the PCN-egress-node (the congestion level estimate, see Section 1.1) for each ingress-egress-aggregate. When the PCN-admission-state (see Section 1.1) for a given ingress-egress-aggregate changes from "Admit" to "Block" or vice versa, the PCN-egress-node reports this change. The PCN-ingress-node admits or blocks new PCN flows offered to a given ingress-egress-aggregate based on the PCN-admission-state, possibly modified by policy direction from the Policy Decision Point (PDP).

The decision to terminate flows requires measurement data from both the PCN-ingress-node and the PCN-egress-node. Hence while the the PCN-admission-state is "block", the PCN-egress-node reports the measured rate of flow of unmarked PCN-traffic it receives for each ingress-egress-aggregate. If the admitted traffic rate measured at the PCN-ingress-node exceeds the reported unmarked received PCN traffic rate multiplied by a configured factor, flows are selected for termination to reduce this difference to zero, with policy guidance from the PDP. The PCN-ingress-node ceases to admit the selected flows.

[Not sure what to do about identifying flows for ECMP]

3.2. Behaviour of the PCN-Egress-Node

For each ingress-egress-aggregate, the egress node continuously measures the following quantities over successive intervals of equal duration. That duration is suggested to be in the range of 100 to 500ms to provide a reasonable tradeoff between signalling demands on the network and the time taken to react to impending congestion.

NM-count:
: Number of octets of PCN-traffic contained in received packets which are not PCN-marked.

PM-count:
: Number of octets of PCN-traffic contained in received packets which are PCN-marked.

3.2.1. PCN-Egress-Node Role In Flow Admission

At the end of each measurement interval, the egress node calculates a ratio R. If both counts are zero for the interval, the ratio R is set to zero. Otherwise, the egress node calculates the ratio as:

R = PM-count / (NM-count + PM-count).

The egress node then updates a congestion level estimate (CLE, see Section 1.1) with this ratio using exponential smoothing:

new_CLE = k*R + (1-k)*old_CLE,

where k is a constant chosen to put most (say 80%) of the weight in the accumulated average on the most recent 1 to 3 seconds of data. The value of k thus depends on the length of the measurement interval.

The next step is to examine the relationship of old-CLE and new_CLE to a configured admission decision threshold (Section 1.1). If old_CLE is above the threshold and new_CLE is below it, the egress node reports that the PCN-admission-state is now "admit" for the ingress-egress-aggregate. If old-CLE and new-CLE are both below the threshold, no action is required. If new-CLE is above the threshold, the PCN-admission-state is now "block" for the ingress-egress-aggregate. The PCN-egress-node procedure in this case is described in Section 3.2.2.

Note: In the case of SM, the CLE is an indication of where the actual load is with respect to the PCN-admissible-rate. In fact, a admission decision threshold of x implies that the expected behavior of SM is to keep the mean load at the fraction x above the PCN-admissible-rate. Hence with SM, the admission decision threshold should be configured with a small value to avoid unintended over-admission.

3.2.2. PCN-Egress-Node Role In Flow Termination

When the PCN-egress-node determines that the PCN-admission-state computed on the basis of the updated CLE is "block", it generates a report indicating the PCN-admission-state and providing the NM-count normalized to a rate NM-rate in octets per second.

[Not sure what to do about identifying flows for ECMP.]

3.3. Behaviour of the PCN-Ingress-Node

The PCN-related functions of the PCN-ingress-node are described briefly in section 4.2 of [RFC5559]. This section focusses on the specific behaviour associated with admission and flow termination.

3.3.1. PCN-Ingress-Node Role In Flow Admission

When the PCN-ingress-node receives a report indicating that the PCN-admission-state for a given ingress-egress-aggregate is "admit", it admits new flows to that aggregate. When the PCN-ingress-node receives a report indicating that the PCN-admission-state for a given ingress-egress-aggregate is "block", it ceases to admit new flows to that aggregate. These actions may be modified by policy input from the Policy Decision Point (PDP).

3.3.2. PCN-Ingress-Node Role In Flow Termination

For each ingress-egress-aggregate, the ingress node continuously measures the following quantity over successive intervals of equal duration. That duration is suggested to be in the range of 100 to 500ms, and preferably the same as at the PCN-egress-node.

Sent-count:
Number of octets of PCN-traffic contained in PCN packets which are admitted to the PCN domain.

When the PCN-ingress-node receives a report containing a value for the unmarked PCN traffic rate NM-rate for a given ingress-egress-aggregate, it takes the most recently observed value of Sent-count and normalizes it to a rate Sent-rate in octets per second. It then calculates the difference

Sent-rate - U * NM-rate,

where U is a configured network-wide constant. If this difference is positive, it indicates a required reduction in the rate of admission of PCN traffic to that ingress-egress-aggregate. Flows are selected for termination with policy input from the PDP. The PCN-ingress-node ceases to admit the selected flows.

If the computed difference is negative, the PCN-ingress-node takes no further action.

3.4. Possible Extension to the Basic Algorithm

The termination mechanisms of SM and CL as described in [I-D.pcn-CL-edge-behaviour] are both based on excess-rate metering and marking,

however, there is a subtle difference between the two mechanisms stemming from the fact that in SM, the bottleneck condition with respect to the PCN-supportable-rate is not directly conveyed through the markings. SM meters against the PCN-admissible- rate and infers the bottleneck condition based on excess-marked traffic. The inference process is vulnerable to inaccuracies, such as non-uniformity in the marking distribution, and may result in over-termination, especially when ingress-egress aggregation is low (< 50 flows).

If SM is used in a low IE-aggregation enviroment, to mitigate this problem, a possible extension to the basic algorithm is to implement an additional control, based on smoothing, to counter the inaccuracy in the interval measurements and to safeguard the triggering of termination. One such control can be implemented with a CLE-like value (referred to as CLE-t). Note, the CLE-t is computed in exactly the same way as described in Section 3.2.1, only with a different value of k (so that the termination is independent of the admission decision). The CLE-t is then compared to the value (U-1)/U. If CLE-t is smaller, no termination should be applied, even if the computed U * NM-rate is smaller than the Sent-rate. Otherwise, the aggregate compares the U * NM-rate to Sent-rate to see if (and how much) to terminate as described in Section 3.2.2.

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 draws heavily upon points made in the PCN architecture document, [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 network failures) already-admitted flows may be terminated to protect the quality of service of the remainder. The SM boundary node behaviour is more likely to terminate too many flows under such circumstances than some alternative PCN boundary node behaviours.

Single-Marking requires no extension to the baseline PCN encoding described in [ID.PCN-baseline], thus reducing the work expected to be performed in the data path of the high-speed routing equipment, and

saving valuable real estate in the packet header.

4.2. Technical Specification

The technical specification of the PCN SM per domain behaviour is provided by the contents of [RFC5559], [ID.PCN-baseline], [ID.PCN-marking], and the present document.

4.3. Attributes

TBD -- basically low loss, low jitter. Low delay would be nice but has to be quantified

4.4. Parameters

TBD. Don't think RFC 3068 is looking for the list of configurable parameters given in the architecture document.

4.5. Assumptions

Assumed that a specific portion of link capacity has been reserved for PCN traffic. Assumed that recovery from overloads by flow termination should happen within 1-3 seconds.

4.6. Example Uses

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

4.7. Environmental Concerns

In some markets, traffic preemption is considered to be impermissible. In such environments, flow termination would not be enabled.

4.8. Security Considerations

Please see the security considerations in Section 5 as well as those in [RFC2474] and [RFC2475].

5. Security Considerations

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

6. IANA Considerations

This memo includes no request to IANA.

7. Acknowledgements

Excluding the appendices, the content of this memo is drawn from [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.

8. References

8.1. Normative References

[ID.PCN-baseline]
Moncaster, T., Briscoe, B., and M. Menth, "Baseline Encoding and Transport of Pre-Congestion Information (Work in progress)", May 2009.

[ID.PCN-marking]
Eardley, P., "Metering and marking behaviour of PCN-nodes (Work in progress)", June 2009.

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

[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.

8.2. Informative References

[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.

[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

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

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: tom.taylor@rogers.com