Internet Engineering Task Force
Internet-Draft
Expires: April 27, 2009

M. Menth
University of Wuerzburg

October 24, 2008

Deployment Models for PCN-Based Admission Control and Flow Termination Using Packet-Specific Dual Marking (PSDM) draft-menth-pcn-psdm-deployment-00

Status of this Memo

Abstract

This document presents different deployment models of PCN-based admission control (AC) and flow termination (FT) using packet-specific dual marking (PSDM) for encoding. Their major is that they require only a single DSCP for packet marking and that they work reliably in the presence of ingress-egress aggregates (IEAs) with only a very small average number of flows. Moreover, an advanced model even works with multipath routing.

Table of Contents

1. Introduction 4
1.1. Requirements Notation 5
2. Terminology 6
3. Admission Control Methods 7
3.1. Configuration of the Exhaustive Marker 7
3.2. Admission Control Based on AC States for IEAs (IEABAC) Using Probe Traffic 7
3.2.1. Observation-Based AC for IEAs Using Probe Packets . . 7
3.2.2. Congestion Level Estimate (CLE) Based AC for IEAs Using Probe Packets 8
3.3. Implicit per-Flow Probing 8
3.3.1. A Brief Summary of RSVP 8
3.3.2. Modification of Standard RSVP to Perform PCN-Based AC 9
4. Flow Termination Methods 10
4.1. Configuration of the Excess Marker 10
4.2. Direct Measured Rate Termination (DMRT) 10
4.2.1. Operation 10
4.2.2. Packet Dropping Policy 10
4.3. Indirect Measured Rate Termination (IMRT) 10
4.3.1. Operation 10
4.3.2. Required Packet Dropping Policy 10
4.4. Marked Flow Termination for IEAs (MFT, MFT-IEA) 10
4.4.1. Operation 10
4.4.2. Required Packet Dropping Policy 10
5. Comparison with Other Deployment Models 12
5.1. Comparison with the Single-Marking (SM) Deployment Model 12
5.2. Comparison with the Control-Load (CL) Deployment Model . . 12
6. IANA Considerations 13

7. Security Considerations 14
8. Conclusions 15
9. Comments Solicited 16
10. References 17
10.1. Normative References 17

10.2. Informative References 17
Author's Address 19
Intellectual Property and Copyright Statements 20

1. Introduction

Pre-congestion notification (PCN) supports admission control (AC) and flow termination (FT) for high priority flows in a DiffServ region in order to protect the quality of service (QoS) of inelastic flows [PCN-arch]. PCN assumes that each link of a network is associated with a PCN admissible and supportable rate (AR, SR). If its PCN traffic rate is below AR, the link is not pre-congested, if it is above AR, the link is AR-pre-congested, and if it is above SR, the link is SR-pre-congested. In case of AR-pre-congestion, AC should block new PCN flows using this link and in case of SR-pre-congestion, FT should terminate sufficiently many PCN flows using this link to reduce the PCN rate below SR.

Packet meter and markers on links of a so-called PCN domain mark packets in case of pre-congestion to notify the egress nodes about the highest pre-congestion level on their paths. The egress nodes turn this information into AC and FT decisions. Excess traffic marking marks PCN packets that exceed a certain reference rate on a link while exhaustive marking marks all PCN packets on a link when the PCN traffic rate exceeds a reference rate [PCN-marking-behaviour].

The IP header has not any unused bits. Therefore, the integration of the marks into the IP header is challenging. A DSCP is chosen to indicate PCN and the ECN field [RFC3168] is re-used to indicate the exact marking. Baseline encoding [Baseline] is able to mark packets as marked and unmarked and therefore it can support either excess traffic marking or exhaustive marking in a network. Packet-specific

dual marking (PSDM) [PSDM] supports two concurrent marking schemes in a limited way. There are two different codepoints for unmarked packets (NM1, NM2) but only one codepoint for marked packets (M). Both baseline and PSDM encoding require only a single DSCP. For the remainder of this draft we assume that packets with codepoint NM1 are subject to excess traffic marking packets with codepoint NM2 are subject to exhaustive marking. Both marker types possibly re-mark these packets to M in case of AR- or SR-pre-congestion. To determine the meaning of a M-marked packet, it is important to know whether it was NM1- or NM2-marked at the ingress node of the PCN domain. This is can be achieved by introducing two types of packets: ordinary data packets of a PCN flow and other distinguishable signalling packets. These signalling packets may be associated with a flow (e.g. RSVP signalling packets for that flow) or with an IEA (e.g. explicit probe packets to maintain its admission state). Note that both kinds of probing do not require extra packets per flow and do not delay the AC decision. Other encoding mechanisms are proposed that require two DSCPs [3state].

This document proposes three AC methods that rely on exhaustive marking and three FT methods that rely on excess marking in such a way that PSDM can be used to encode the markings. Basically, any of these AC methods can be deployed in combination with any of these FT methods. The deployment of such PCN-based AC and FT has two major advantages compared to other deployment models [SM], [CL].

- They require only a single DSCP for packet marking.
- They work reliably even if the average number of flows per IEA is low.
- Some of them work even in case of multipath routing.

1.1. Requirements Notation

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

2. Terminology

Most of the terminology used in this document is defined in [PCN-arch]. The following additional terms are defined in this document:

- Exhaustive marking - generalization of threshold and ramp marking
- IEA - ingress-egress aggregate

3. Admission Control Methods

In this section we explain three AC methods that rely on PCN feedback from probe packets. We first explain the assumption on metering and marking and then present the AC methods. Two of these AC methods maintain AC states for IEAs while the third AC method implements implicit probing per flow.

3.1. Configuration of the Exhaustive Marker

The probe packets are marked with NM2 by PCN ingress nodes to indicate to PCN nodes that they are subject to exhaustive marking. The reference rate of the exhaustive markers is set to the admissible rate. It meters and marks all PCN packets and possibly re-marks NM2 marked PCN packets to M.

3.2. Admission Control Based on AC States for IEAs (IEABAC) Using Probe Traffic

IEABAC assumes that IEA contexts are a priori set up for all IEAs for which PCN flows should be admitted. An IEA context keeps an AC state K at the PCN ingress node of the IEA. The AC state K may be either "admit" or "block". Based on this AC state, the PCN ingress node can either admit or block further flows for the corresponding IEA. To update the AC state according to the pre-congestion level of the path over which the traffic of the IEA is carried, the PCN ingress node sends in regular intervals extra probe messages addressed to the corresponding PCN egress node. This signalling traffic is marked as PCN traffic with NM2 to indicate that it is subject to exhaustive marking. As exhaustive marking is independent of packet sizes, the size of the probe packets can be chosen very small probe to minimize the rate of the signalling traffic. The probe packets have a special format so that PCN egress nodes recognize them as such for the corresponding PCN ingress nodes and associate them with the correct IEA. Based on the markings of the probe packets (NM2, M), the PCN egress node derives the new AC state for the IEA and sends and admission-stop or admission-continue message to the corresponding PCN ingress in order to update the AC state of the IEA. In the following we propose two approaches for PCN egress nodes to determine the new AC state of an IEA.

3.2.1. Observation-Based AC for IEAs Using Probe Packets

When the PCN egress node receives an M-marked probe packet, it sends an admission-stop message to the corresponding PCN ingress node and sets a timer for the minimum block interval to a configurable value Tblock. The timer may be reset by consecutive arrivals of M-marked probe packets. When the timer expires, an admission-continue message

is sent to the PCN ingress node.

3.2.2. Congestion Level Estimate (CLE) Based AC for IEAs Using Probe Packets

The PCN-egress-node proceeds in measurement intervals of a configurable time MI. It tracks the number of missing missing probe packets or probe packets received with an M-mark during a measurement interval and at its end it calculates the CLE as the fraction of this number and the number of overall received and missing probe packets. If the CLE is smaller than a configurable value TadmCont, an admission-continue message is sent to the PCN ingress node. If the CLE is larger than a configurable value TadmStop, an admission-stop message is sent to the PCN ingress node.

3.3. Implicit per-Flow Probing

We briefly review RSVP and explain how its signalling messages can be re-used for implicit per-flow probing.

3.3.1. A Brief Summary of RSVP

Realtime flows are usually accompanied by end-to-end signalling. A popular protocol example is RSVP [RFC2205]. With RSVP, the data source issues a PATH message which is carried hop-by-hop over the same path future data packets will go. To that end, the PATH message uses the same source and destination address as future data packets and also all other header fields that are possible input for routing and load balancing decisions need to be the same. When a PATH message arrives at an RSVP-capable node, a PATH state is established pointing to the previous hop before the PATH message is forwarded further downstream. When the PATH message arrives at the destination, the destination triggers the end-to-end reservation for the flow by sending a RESV message upstream along the nodes that set up a PATH state. In these nodes, the RESV message is processed. In particular, resource admission control is performed for the new flow request and if it succeeds, the node forwards the RESV message to the previous hop recorded by the PATH state. This two pass signalling approach guarantees that the reservation is done on the downstream path of the future data flow. In contrast to PATH messages, RESV messages have the source address of the sending node and the destination address of the hop pointed to by the PATH state. That way, the information about the downstream next hop of the future data stream is conveyed to the previous hop and the flow-related information is stored in a RESV state. RSVP is a soft-state protocol, i.e., the PATH and RESV control messages are periodically sent to keep the PATH and RESV states alive and, thereby, the flow reservations. Admission control needs to be performed for a flow

only once when no RESV state is set up, yet.

3.3.2. Modification of Standard RSVP to Perform PCN-Based AC

We assume that interior nodes of a PCN domain are RSVP-disabled. That means, they just forward RSVP messages without processing them and PCN ingress and egress nodes are neighboring RSVP-capable nodes. As a consequence, PCN ingress nodes decide whether new flows can be admitted and carried through domain or not. When the initial PATH message travels downstream, it is marked with NM2 by the ingress node to indicate to PCN nodes that this packet is subject to exhaustive marking. It is possibly re-marked to M and eventually received by the PCN egress node. If no PATH state can be found for this flow at the PCN egress node, this PATH message is the first one and not a REFRESH message. If the PATH message is the first of the flow and if it is marked with M, the RSVP engine sends back a PATHERR message to reject the flow. If the PATH message is still marked with NM2, the RSVP PATH state is established at the PCN egress node and the PATH message is forwarded further downstream. REFRESH messages are just forwarded according to standard RSVP. When the PATH message arrives at the destination and a RESV message is sent back along the nodes with a PATH state. Eventually, the corresponding RESV message arrives at the PCN ingress node. When no RESV state is set up yet, this is the first RESV message and admission control must be performed. By the mere fact that the RESV message arrives, the PCN ingress node knows that the corresponding initial PATH message was not marked. Thus, it can admit any flow for which a new RESV message arrives.

Note that RSVP is only an example for a two-pass end-to-end signalling protocol and the principle can be adopted to others.

4. Flow Termination Methods

In this section we explain three FT methods that rely on PCN feedback from PCN data packets. We first explain the assumption on metering and marking and then present the FT methods.

4.1. Configuration of the Excess Marker

The PCN data packets are marked with NM1 by PCN ingress nodes to indicate to PCN nodes that they are subject to excess marking. The reference rate of the excess markers is set to the supportable rate. It meters all NM1-marked PCN packets and possibly re-marks them to M. Thus, PCN signalling traffic required for AC is not taken into account, but it is designed to have a low bitrate.

4.2. Direct Measured Rate Termination (DMRT)

4.2.1. Operation

see [Overview]

4.2.2. Packet Dropping Policy

DMRT requires preferential dropping of unmarked (NM1) PCN data packets, otherwise the termination process can be delayed [FT-PE].

4.3. Indirect Measured Rate Termination (IMRT)

4.3.1. Operation

See [Overview], same mechanism as in [CL].

4.3.2. Required Packet Dropping Policy

IMRT requires preferential dropping of marked (M) PCN data packets, otherwise significant overtermination can occur [FT-PE].

4.4. Marked Flow Termination for IEAs (MFT, MFT-IEA)

4.4.1. Operation

See [MFT-PE].

4.4.2. Required Packet Dropping Policy

MFT-IEA works optimal when unmarked (NM1) PCN data packets are preferentially dropped in case of unavoidable PCN packet loss. However, the termination process is only little delayed without

preferential dropping of any specially marked PCN packets [MFT-PE].

5. Comparison with Other Deployment Models

5.1. Comparison with the Single-Marking (SM) Deployment Model

SM [SM] uses only a single marking scheme. Therefore, baseline encoding [Baseline] can be used for PCN marking which requires also only a single DSCP. However, SM uses CLE-based AC using feedback from data packets and as a result, it cannot block when an IEA is empty. This can lead to significant overadmission [AC-PE]. Furthermore, its termination method can lead to overtermination [FT-PE] even in the case of single bottlenecks [FT-PE]. Neither the AC nor the FT part of this method reliably workw in the presence of multipath routing.

5.2. Comparison with the Control-Load (CL) Deployment Model

CL [CL] uses two marking schemes and requires that packets can be re-marked to two different codepoints. This cannot be achieved with a single DSCP [3state]. Spending more than a single DSCP for PCN encoding seems very expensive. Therefore, the AC mechanism of CL should be modified, e.g. using one of the mechanisms presented in this document, so that a single DSCP suffices for PCN encoding.

Like SM, CL is prone to overadmission when the number of expected flows per IEA is small [AC-PE]. CL's flow termination method is

prone to overtermination in case of quickly changing traffic rates [FT-PE]. Neither its AC nor its FT method works with multipath routing.

6. IANA Considerations

{ToDo}

7. Security Considerations

{ToDo}

8. Conclusions

This document presented three methods for admission control and three methods for flow termination. The AC methods rely on probe packets

which are subject to exhaustive marking and the FT methods rely on data packets which are subject to excess marking so that packet-specific dual marking (PSDM) can be used to encode PCN marks which requires only a single DSCP for PCN. All AC and FT methods work with when the expected number of flows per ingress-aggregate is small. Each of the AC methods is compatible with each of the FT methods. One AC method and one FT method even works with multipath routing. Those are significant advantages compared to the deployment models presented in [SM], [CL].

9. Comments Solicited

Comments and questions are encouraged and very welcome. They can be addressed to the IETF PCN working group mailing list <pcn@ietf.org>, and/or to the authors.

10. References

10.1. Normative References

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

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

[RFC3168] Ramakrishnan, K., Floyd, S., and D. Black, "The Addition of Explicit Congestion Notification (ECN) to IP", RFC 3168, September 2001.

10.2. Informative References

[3state] Moncaster, T., "A three state extended PCN encoding scheme", draft-moncaster-pcn-3-state-encoding-00 (work in progress), June 2008.

[AC-PE] Menth, M. and F. Lehrieder, "Applicability of PCN-Based Admission Control", <http://www3.informatik.uni-wuerzburg.de/staff/menth/Publications/papers/Menth08-Sub-8.pdf>.

[Baseline] Moncaster, T., Briscoe, B., and M. Menth, "Baseline Encoding and Transport of Pre-Congestion Information", draft-moncaster-pcn-baseline-encoding-02 (work in progress), July 2008.

[CL] Briscoe, B., "An edge-to-edge Deployment Model for Pre-Congestion Notification: Admission Control over a DiffServ Region", draft-briscoe-tsvwg-cl-architecture-04 (work in progress), October 2006.

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

[MFT-PE] Menth, M. and F. Lehrieder, "PCN-Based Marked Flow Termination", <http://www3.informatik.uni-wuerzburg.de/staff/menth/Publications/papers/Menth08-PCN-MFT.pdf>.

[Overview] Menth, M. and et al., "A Survey of PCN-Based Admission Control and Flow Termination", <http://

www3.informatik.uni-wuerzburg.de/staff/menth/Publications/papers/Menth08-PCN-Overview.pdf>.

[PCN-arch] Eardley, P., "Pre-Congestion Notification Architecture", draft-ietf-pcn-architecture-03 (work in progress), February 2008.

[PCN-marking-behaviour] Eardley, P., "Marking behaviour of PCN-nodes", draft-eardley-pcn-marking-behaviour-01 (work in progress), June 2008.

[PSDM] Menth, M., Babiarz, J., Moncaster, T., and B. Briscoe, "PCN Encoding for Packet-Specific Dual Marking (PSDM)", Internet-Draft menth-pcn-psdm-encoding-00, June 2008.

[SM] Charny, A., Zhang, X., Faucheur, F., and V. Liatsos, "Pre-Congestion Notification Using Single Marking for Admission and Termination", draft-charny-pcn-single-marking-03 (work in progress), November 2007.

Author's Address

Michael Menth
University of Wuerzburg
Am Hubland
Wuerzburg D-97074
Germany

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

Full Copyright Statement

Intellectual Property

The IETF takes no position regarding the validity or scope of any Intellectual Property Rights or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; nor does it represent that it has made any independent effort to identify any such rights. Information on the procedures with respect to rights in RFC documents can be found in BCP 78 and BCP 79.

Copies of IPR disclosures made to the IETF Secretariat and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the IETF on-line IPR repository at http://www.ietf.org/ipr.

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement this standard. Please address the information to the IETF at ietf-ipr@ietf.org.