Network Working Group
Internet-Draft
Expires: August 21, 2008

M. Menth
F. Lehrieder
University of Wuerzburg
P. Eardley
BT
A. Charny
Cisco Systems, Inc.
J. Babiarz
Nortel
February 18, 2008

Edge-Assisted Marked Flow Termination
draft-menth-pcn-emft-00

Abstract

This document presents edge-assisted marked flow termination (EMFT) for PCN. It assumes packet-size independent excess marking, i.e. packets exceeding the supportable rate (SR) of a link are marked as "excess-traffic" (ET). EMFT terminates only flows with at least one ET-marked packet. The problem is to avoid that all flows with ET-marked packets are terminated. This draft proposes two solutions. Flow-based EMFT (F-EMFT) considers single flows separately and terminates them when sufficiently many packets of them have been received by the PCN egress node with an ET-mark. Aggregate-based EMFT (A-EMFT) considers ingress-egress-aggregates and terminates flows thereof sufficiently many ET-marked packets have been received for that aggregate.

Table of Contents

1. Introduction 3
2. Terminology 5
3. Required Marking Behavior 7
  3.1. Conventional Excess Marking 7
  3.2. Packet Size Independent Excess Marking (PSIEM) 7
4. Flow-Based Edge-Assisted Marked Flow Termination (F-EMFT) . . 8
5. Aggregate-Based Edge-Assisted Marked Flow Termination (A-EMFT) 9
6. References 10
  6.1. Normative References 10
  6.2. Informative References 10
  6.3. Other References 10
Authors' Addresses 11
Intellectual Property and Copyright Statements 13

1. Introduction

PCN defines a new PCN traffic class that receives preferred treatment by PCN nodes. It provides information to support admission control (AC) and flow termination (FT) for this traffic type. PCN introduces an admissible and a supportable rate threshold (AR(l), SR(l)) for each link l of the network which imply three different link states. If the PCN traffic rate r(l) is below AR(l), there is no pre-congestion and further flows may be admitted. If the PCN traffic rate r(l) is above AR(l), the link is AR-pre-congested and the rate above AR(l) is AR-overload. In this state, no further flows should be admitted. If the PCN traffic rate r(l) is above SR(l), the link is SR-pre-congested and the rate above SR(l) is SR-overload. In this state, some already admitted flows should be terminated. PCN nodes monitor the PCN rate on their links and they remark packets depending on their pre-congestion states. The PCN egress nodes evaluate the packet markings and their essence is reported to the AC and FT entities of the network such that they can take appropriate actions. Therefore, this concept is called pre-congestion notification. This draft proposes a new FT method.

The CL draft [I-D.briscoe-tsvwg-cl-architecture] proposes that all packets above SR are marked with "excess-traffic" (ET). Packets of the same ingress-egress aggregate (IEA) are grouped together for a joint evaluation of their markings by the PCN egress node. If packets are ET-marked, the PCN egress node signals the rate of unmarked packets to the PCN ingress node which terminates so many flows that their rate corresponds to the difference of the sent rate per IEA and the rate that was received non-ET-marked by the PCN egress node. We call this solution measured rate termination (MRT). This solution has two major drawbacks:

- At low aggregation it is hard for the ingress node to determine an appropriate set of flows to be terminated. Example: only a single flow with 1 Mbit/s in the IEA, and 500 kbit/s should be terminated. When many ingress nodes face the same problem and solve it with the same algorithm, either overtermination or undertermination occurs.

- In case of multipath routing, flows of a single IEA may take different routes. The ingress node chooses the set of flows for termination, but does not know which flows are carried over a pre-congested link. Therefore, the wrong flows are possibly terminated.

The 3sm draft [I-D.babiarz-pcn-3sm] proposes marked flow termination. If a PCN node receives an ET-marked packet, it notifies the FT entity to terminate the flow. To avoid overtermination, only a subset of

the packets above SR are ET-marked. The concept of IEA is not needed. This method is called core-assisted marked flow termination (CMFT) as only marked flows are terminated and core nodes help to identify the flows that should be terminated. This method has one major drawback:

- It requires packet size independent excess marking with marking frequency reduction (MFR) which is not yet available in today's routers.

Given the two approaches with their drawbacks, a FT method is desirable where conventional excess marking can be used by PCN nodes, that terminates only marked flows, and that is able to cope with IEAs having only a small number of flows. We present such a solution in this draft and call it edge-assisted marked flow termination (EMFT). The motivating idea for EMFT is to roll a dice at the edges to decide whether a marked packet is to be terminated instead of letting the core nodes decide. The actual solution is slightly different and saves the generation of random numbers per packet.

The next section clarifies some terminology issues. We then describe the required marking behaviour. We present flow-based and aggregate-based EMFT as new FT mechanisms and discuss security issues.

2. Terminology

The terminology used in this document conforms to the topology of [I-D.ietf-pcn-architecture].

We use the following exceptions for better readability and provide the synonyms defined in [I-D.ietf-pcn-architecture].

- Admissible rate: PCN-lower-rate
- Supportable rate: PCN-upper-rate
- Admission-stop marking: first encoding or PCN-lower-rate-marking
- Excess-traffic marking: second encoding or PCN-upper-rate-marking

New terminology

- Flow termination (FT): function to terminate flows in case of SR-pre-congestion
- No-pre-congestion (NP) marking: marking for packets that have not yet experience any form of pre-congestion
- Packet size independent marking (PSIM): marks all packets exceeding a certain rate, but the marking probability of a packet is independent of its size. This is in contrast to pure excess marking. May be implemented by a threshold marking algorithm.
- MFT: marked flow termination terminates only flows with at least one ET-marked packet; guarantees that terminated flow traverses an AR-pre-congested link.
- CMFT: core-assisted MFT: core nodes apply marking frequency reduction to control termination speed of MFT
- EMFT: edge-assisted MFT: egde nodes control the termination speed of MFT
- F-EMFT: flow-based EMFT
- A-EMFT: aggregate-based EMFT
- IEA: ingress-egress aggregate
- Flow termination delay D_T: duration of the interval between the decision for the termination of a flow at the PCN egress node and the time the PCN egress node does not receive packets of that flow

anymore.

3. Required Marking Behavior

EMFT works with conventional excess marking, but for the sake of fairness, packet-size independent excess marking is preferred. We describe both marking behaviours in the following.

3.1. Conventional Excess Marking

Conventional excess marking is based on a token bucket with size S and Rate R. When a packet arrives, and the number of tokens in the bucket is at least the packet size, the number of tokens is reduced by the packet size. If the number of tokens in the bucket is smaller than the packet size, the packet is marked.

Larger packets have a higher probability to be marked. Therefore, marked flow termination (MFT) algorithms terminate flows sending larger packets with a higher probability than flows sending small packets.

3.2. Packet Size Independent Excess Marking (PSIEM)

PSIEM addresses the above problem and makes the marking probability independent of the packet size. To that end, a marking threshold T is introduced which is set to the maximum transfer unit (MTU). If a packet arrives and the number of tokens in the bucket is T or larger, the number of tokens in the bucket is reduced by the packet size. If the number of tokens in the bucket is smaller than the threshold T, it remains unchanged, but the packet is marked.

4. Flow-Based Edge-Assisted Marked Flow Termination (F-EMFT)

The PCN egress node keeps a credit counter C for each flow. When an ET-marked packet arrives for a flow, the corresponding credit counter is reduced by the size of that packet. If the credit counter is non-positive at the arrival of a marked packet, the flow is terminated.

The difficulty is the suitable initialization of the credit counter when a reservation is set up for a new flow. In [Menth08-PCN-MFT] we have shown that the initial counter size should be exponentially distributed with mean 2*R_f*E[DT]/alpha where R_f is the rate of the flow f, E[DT] is a global average value for the flow termination delay, and alpha is a knob to control the termination speed. The parameter alpha should be set at most 1 to avoid that flows are terminated too fast such that overtermination occurs. Smaller alpha results in a longer time to reduce SR-overload. The impact of these parameters is also studied in [Menth08-PCN-MFT].

Statistical flow termination priorities can be implemented by granting larger initial credit counters to more important flows.

We give an example for a potential technical implementation of the exponentially distributed credit counter size distribution. The end system generates a random number x between 0 and 1. Then it determines the initial size of the credit counter by C=-ln(x)*2*R_f*E[D_T]/alpha.

5. Aggregate-Based Edge-Assisted Marked Flow Termination (A-EMFT)

If it is easy for the PCN egress node to identify all packets of the same PCN ingress node, the packet markings can be evaluated on an aggregate basis. Then, the following algorithm may be used. A credit counter is associated with each IEA and initialized similarly as for F-EMFT, i.e. by an exponential distribution with average value 2*E[R]*E[DT]/alpha where E[R] is the average rate of the current flows in the IEA. Usually, E[R] is the rate R_f of the first flow when the system starts with a single flow.

When ET-marked packets arrive and the credit counter is positive, the size of the credit counter C is reduced by the packet size. If the credit counter C is not positive, a flow f of the aggregate is terminated and a deterministic increment of I=2*R_f*E[DT]/alpha is added to the credit counter, i.e., the increment is proportional to the rate of the terminated flow f. With this configuration, F-EMFT and A-EMFT lead to the same termination behaviour.

Note that the flow f to be terminated can be the flow to which the last ET-marked packet belongs to, but it may also be any other flow for which an ET-marked packet recently arrived. This allows the enforcement of termination policies. For instance, high priority flows may be later terminated than low priority flows.

6. References

6.1. Normative References

6.2. Informative References

[I-D.babiarz-pcn-3sm]
Babiarz, J., "Three State PCN Marking", draft-babiarz-pcn-3sm-00 (work in progress), July 2007.

[I-D.briscoe-tsvwg-cl-architecture]
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.

[I-D.ietf-pcn-architecture]
Eardley, P., "Pre-Congestion Notification Architecture", draft-ietf-pcn-architecture-01 (work in progress), October 2007.

6.3. Other References

[Menth08-PCN-MFT]
Menth, M. and F. Lehrieder, "Termination Methods for End-to-End PCN-Based Flow Control", February 2008, <http://www3.informatik.uni-wuerzburg.de/staff/menth/Publications/Menth08-PCN-MFT.pdf>.

Authors' Addresses

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

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

Frank Lehrieder
University of Wuerzburg
Am Hubland
Wuerzburg D-97074
Germany

Phone: +49-931-888-6634
Email: lehrieder@informatik.uni-wuerzburg.de

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

Email: philip.eardley@bt.com

Anna Charny
Cisco Systems, Inc.
1414 Mass. Ave.
Boxborough, MA 01719
USA

Email: acharny@cisco.com

Jozef Z. Babiarz
Nortel
3500 Carling Avenue
Ottawa, Ont. K2H 8E9
Canada

Phone: +1-613-763-6098
Email: babiarz@nortel.com

Acknowledgment

Funding for the RFC Editor function is provided by the IETF Administrative Support Activity (IASA).