University of Würzburg
Institute of Computer Science
Research Report Series

PCN-Based Marked Flow Termination

Michael Menth and Frank Lehrieder

Report No. 469 March 2010

University of Wuerzburg, Germany
Institute of Computer Science
Chair of Communication Networks
Am Hubland, D-97074 Würzburg, Germany
phone: (+49) 931-3186644, fax: (+49) 931-3186632
{menth|lehrieder}@informatik.uni-wuerzburg.de

PCN-Based Marked Flow Termination

Michael Menth and Frank Lehrieder
University of Wuerzburg, Germany
Institute of Computer Science
Chair of Communication Networks
Am Hubland, D-97074 Würzburg, Germany
phone: (+49) 931-3186644, fax: (+49) 931-3186632
{menth|lehrieder}@informatik.uni-wuerzburg.de

Abstract

Pre-congestion notification (PCN) uses packet metering and marking to notify network boundaries when the current traffic load on some links of a PCN domain exceeds their configured admissible or supportable rates. This feedback is used for PCN-based admission control and flow termination within single PCN domains (edge-to-edge PCN). While admission control is rather well understood, flow termination is a new flow control function and useful especially in case of failures or flash crowds. Originally, only measured rate termination was proposed. It operates on ingress-egress aggregates between ingress and egress nodes of a PCN domain, terminates overload traffic in one shot, and requires single-path routing. We present marked flow termination as a new paradigm. It works for both ingress-egress aggregates and individual flows, terminates overload traffic gradually, and works well with multipath routing. The major contribution of this paper is the presentation of several marked flow termination methods, the study of their termination behavior, recommendations for their configuration, and the discussion of their benefits and shortcomings. A secondary contribution is the introduction of end-to-end PCN which moves the control from the boundaries of possibly several PCN domains to the end systems. Although end-to-end PCN has a trust problem in an Internet context, it might be useful for large corporate networks.

Keywords: Flow control, self-adaptation, performance evaluation

1 Introduction

Network providers and manufacturers have recently recognized the need for new admission control concepts for the Internet that are simpler and more scalable than RSVP in terms of operation and state management. Therefore, the IETF has started a working group (WG) to standardize admission control (AC) and flow termination (FT) for DiffServ networks based on pre-congestion notification (PCN). The AC function admits or rejects new flows for a so-called PCN domain based on measured feedback from the network [1]. The FT function tears

This work was funded by Nortel Networks, Ottawa, and Deutsche Forschungsgemeinschaft (DFG) under grant TR257/18-2. The authors alone are responsible for the content of the paper.

down already admitted traffic in case of imminent overload which can occur in spite of AC due to rerouted traffic in failure cases or other unexpected events.

PCN defines a new admission-controlled traffic class whose traffic is preferentially forwarded. For each link l of a PCN domain an admissible rate $(AR(l))$ and a supportable rate $(SR(l))$ threshold are configured. PCN packets enter the PCN domain with a "no-pre-congestion" (NP) mark and when the PCN traffic rate $r(l)$ of a link exceeds $AR(l)$ or $SR(l)$, packets are marked with an "admission-stop" (AS) or "excess-traffic" (ET) codepoint, respectively. If an egress node of a PCN domain sees AS- or ET-marked packets, it communicates this to the AC or FT entity of the network to stop the admission of new flows or to tear down already admitted flows. This concept scales well because the metering and marking algorithm in the network core do not need to know individual flows or aggregates. The present focus of the WG is limited to a single domain (edge-to-edge PCN). So far, there is consensus on the general idea [1], but there are many open issues concerning the encoding of PCN marks in the IP header, the exact behavior of the meters and markers, and the operation of PCN edge nodes to support AC and FT based on received PCN marks.

In this work we introduce the notion of end-to-end PCN. Here, the PCN control nodes reside in the end systems instead of in boundary nodes of PCN domains as it is the case for edge-to-edge PCN. Although end-to-end PCN has a trust problem in the general Internet, it might be useful in large corporate networks.

While AC methods have been studied intensively in the past, PCN's FT feature is a new flow control function and only little understood. Early proposals use measured rate termination (MRT) to terminate flows. They estimate the rate to be terminated based on traffic measurement on an ingress-egress aggregate (IEA) basis between PCN ingress and egress nodes. Then, a suitable subset of flows from that IEA are terminated in one shot. MRT has two major shortcomings. It does not work well with multipath routing and is not applicable for end-to-end PCN.

The main contribution of this paper is the introduction and performance evaluation of marked flow termination (MFT) that is in contrast to MRT applicable for both multipath routing and end-to-end PCN. We present three partly competing MFT methods: MFT based on excess marking with marking frequency reduction (MFT-MFR), MFT based on plain excess marking for individual flows (MFT-IF) and for ingress-egress aggregates (MFT-IEA). We describe their operation, analyze their termination behavior, give recommendations for their configuration, and summarize their pros and cons.

Sect. 2 gives an overview of related work. Sect. 3 summarizes the current discussion about edge-to-edge PCN and discusses MRT. Sect. 4 proposes end-to-end PCN, and clarifies how edge-to-edge and end-to-end PCN can coexist. Sect. 5 proposes the three new MFT methods and analyzes their termination behavior by means of mathematical analysis and simulation. Sect. 6 summarizes this work and gives conclusions.

2 Related Work

We review related work regarding random early detection (RED), explicit congestion notification (ECN), and stateless core concepts for AC as they can be viewed as historic roots of

PCN.

2.1 Random Early Detection (RED)

RED was originally presented in [2], and in [3] it was recommended for deployment in the Internet. It was designed to detect incipient congestion by measuring the average buffer occupation avg in routers and to take appropriate countermeasures in order to improve the throughput of TCP connections. To that end, packets are dropped or marked to indicate congestion early to TCP senders and the probability for that action increases linearly with the average queue length avg. The value of avg relates to the physical queue size which is unlike PCN metering that relates to virtual queue sizes based on configured admissible and supportable rates.

2.2 Explicit Congestion Notification

Explicit congestion notification (ECN) is built on the idea of RED to signal incipient congestion to TCP senders in order to reduce their sending window [4]. Packets of non-ECN-capable flows can be differentiated by a "not-ECN-capable transport" (not-ECT, '00') codepoint from packets of a ECN-capable flow which have an "ECN-capable transport" (ECT) codepoint. In case of incipient congestion, RED gateways possibly drop not-ECT packets while they just switch the codepoint of ECT packets to "congestion experienced" (CE, '11') instead of discarding them. This improves the TCP throughput since packet retransmission is no longer needed. Both the ECN encoding in the packet header and the behavior of ECN-capable senders and receivers after the reception of a marked packet is defined in [4]. ECN comes with two different codepoints for ECT: ECT(0) ('10') and ECT(1) ('01'). They serve as nonces to detect cheating network equipment or receivers [5] that do not conform to the ECN semantics. The four codepoints are encoded in the (currently unused) bits of the differentiated services codepoint (DSCP) in the IP header which is a redefinition of the type of service octet [6]. The ECN bits can be redefined by other protocols and [7] gives guidelines for that.

2.3 Admission Control

We briefly review some specific AC methods that can be seen as forerunners of the PCN principle. They all have a stateless core which is a key property of PCN.

2.3.1 Admission Control Based on Reservation Tickets

To keep a reservation for a flow across a network alive, ingress routers send reservation tickets in regular intervals to the egress routers. Intermediate routers estimate the rate of the tickets and can thereby estimate the expected load. If a new reservation sends probe tickets, intermediate routers forward them to the egress router if they have still enough capacity to support the new flow. The egress router bounces them back to the ingress router indicating a successful reservation. Otherwise, the intermediate routers discard the probe tickets and the reservation request is denied. The tickets can also be marked by a packet state. Several stateless core mechanisms work according to this idea [8, 9, 10].

2.3.2 Admission Control Based on Packet Marking

Gibbens and Kelly [11, 12, 13] theoretically investigated AC based on the feedback of marked packets whereby packets are marked by routers based on a virtual queue with configurable bandwidth. This core idea is adopted by PCN. Marking based on a virtual instead of a physical queue also allows to limit the utilization of the link bandwidth by premium traffic to arbitrary values between 0 and 100%. Karsten and Schmitt [14, 15] integrated these ideas into the IntServ framework and implemented a prototype. They point out that the marking can also be based on the CPU usage of the routers instead of the link utilization if this turns out to be the limiting resource for packet forwarding. The authors of [16] presented an early version of PCN-based AC.

2.3.3 Resilient Admission Control

Resilient admission control admits only so much traffic that it still can be carried after rerouting in a protected failure scenario [17, 18]. It is necessary since overload in wide area networks mostly occurs due to link failures and not due to increased user activity [19]. It can be implemented with PCN by setting the admissible rate thresholds $AR(l)$ low enough such that the PCN rate $r(l)$ on a link l is lower than the supportable rate threshold $SR(l)$ after rerouting, at least for most likely failure events. The provided backup capacity may not suffice for rare failure events. For such cases FT functions as discussed in the PCN context are useful to quickly remove some admitted flows if overload occurs.

3 Pre-Congestion Notification for Single Domains

We explain the general concept of PCN and its application in single domains. We summarize existing proposals and point out the shortcomings of measured rate termination.

3.1 Pre-Congestion Notification (PCN)

PCN is intended for use in DiffServ networks and defines a new 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 as illustrated in Fig. 1. 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 traffic 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 AR- and SR-pre-congested and the traffic 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 re-mark packets depending on the pre-congestion states of these links. 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 admit or

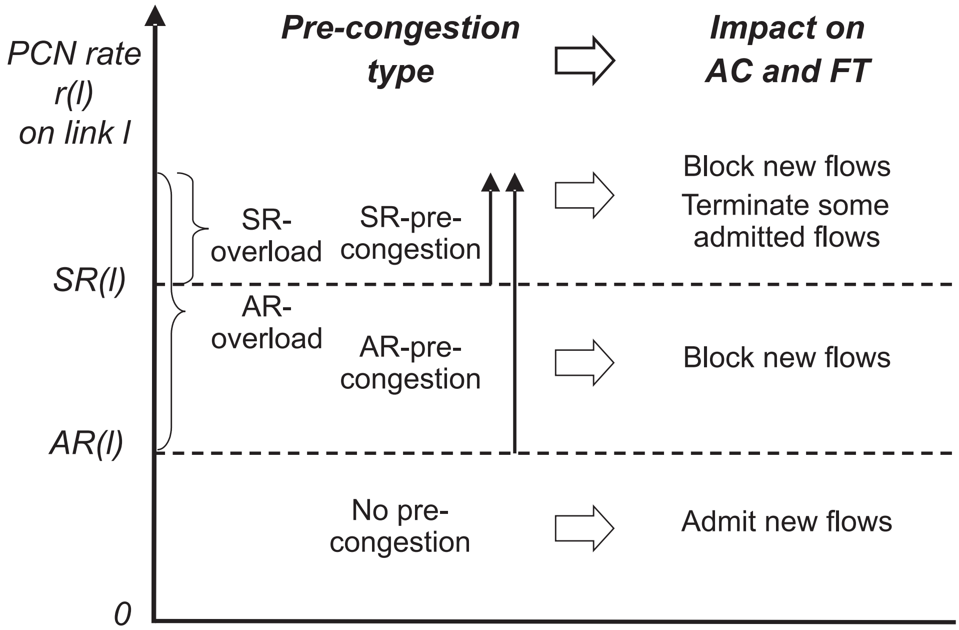


Figure 1: The admissible and the supportable rate $(AR(l), SR(l))$ define three pre-congestion states concerning the PCN traffic rate $r(l)$ on a link.

block new flows or even terminate already admitted flows. Therefore, this concept is called pre-congestion notification.

3.2 Edge-to-Edge PCN

Edge-to-edge PCN implements AC and FT for a single domain. The assumption is that some end-to-end signalling protocol (e.g. SIP or RSVP) requests admission for a new flow to cross the PCN domain similar to the IntServ-over-DiffServ concept [20]. This is illustrated in Fig. 2(a). Traffic enters the PCN domain only through PCN ingress nodes and leaves it only through PCN egress nodes. The nodes within a PCN domain are PCN nodes. They monitor the PCN traffic rate on their links and possibly re-mark the traffic in case of AR- or SR-pre-congestion. PCN egress nodes evaluate the markings of the traffic and send a digest to the AC and FT entities. There are many proposals for the technical realization of these objectives. We summarize two of them using a unified nomenclature.

3.2.1 Controlled Load (CL) Proposal

The CL proposal [21] offers a simple AC and FT to support a controlled load (CL) service [22] over DiffServ networks similar to [20]. PCN traffic enters the network unmarked, i.e. with a “no-precongestion” (NP) codepoint. If some packets exceed the supportable rate $SR(l)$ of a link, they are re-marked to the “excess-traffic” (ET) codepoint. This type of marking is called excess marking. When the admissible rate $AR(l)$ of a link is exceeded by the PCN traffic rate

$r(l)$, all non-ET-marked packets are re-marked to the "admission-stop" (AS) codepoint. This type of marking is called exhaustive marking.

For the evaluation of the markings, the PCN egress nodes map the PCN traffic which shares common PCN ingress and egress nodes into ingress-egress aggregates (IEAs). To support AC, the PCN egress node calculates the combined fraction of AS- and ET-marked packets with regard to all PCN packets per IEA. It is the so-called congestion level estimate (CLE) and the PCN egress node signals it to the PCN ingress node of the IEA. The PCN ingress node stops or continues admission of further flows if the CLE value is high or low enough. To support FT, the egress node signals the rate of non-ET marked traffic – the so-called sustainable rate – to the ingress node when it receives ET-marked packets. The ingress node compares the sustainable rate with the actually sent PCN rate and terminates a subset of flows belonging to this IEA such that their rate corresponds to the difference of the sustainable and the sent rate. We call this approach measured rate termination (MRT) because the amount of traffic to be terminated is determined by rate measurement. A minimum inter-termination time between two consecutive termination steps is required to make sure that terminated flows do not contribute anymore to the measured feedback. The AC and FT decisions of a PCN domain are enforced by appropriate filters and per flow policers. Only packets of admitted flows receive the prioritized forwarding treatment of the PCN traffic class and packets of other flows are blocked when they demand for this premium service.

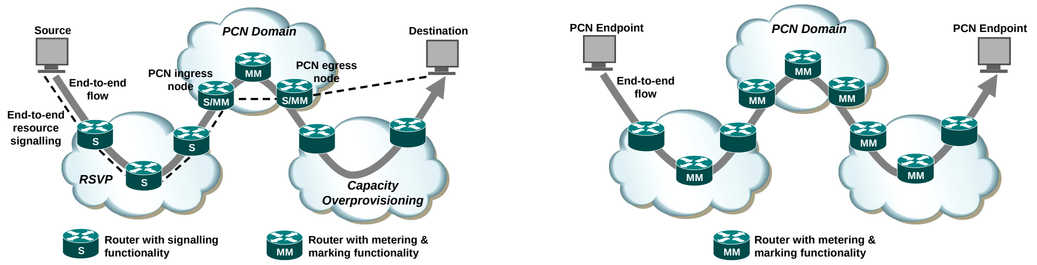
3.2.2 Single-Marking (SM) Proposal

The SM proposal [23] uses only two PCN codepoints: NP and AS, but the AS codepoint has different semantics than in the CL proposal. Packets enter the PCN domain NP-marked and are re-marked to AS when they exceed the admissible rate $AR(l)$. Excess marking is applied with respect to $AR(l)$, i.e. only the AR-overload is marked. The fact that only a single marking scheme is used explains the name of the proposal. The SM proposal requires that the supportable rate $SR(l) = AR(l) \cdot u$ is a multiples of the admissible rate on an link l whereby the factor u is a domain-wide constant.

Like in the CL proposal, PCN egress nodes signal the CLEs to PCN ingress nodes to support their AC decisions, but different threshold values are used to stop or continue admission of further flows. The PCN egress nodes measure the rate of NP-marked packets, multiply it by u, and send this value as sustainable rate to the ingress nodes when they receive AS-marked packets. Like in CL, the PCN ingress nodes measure the sent traffic. If the received sustainable rate is lower than the measured sent rate, the PCN ingress nodes terminate a subset of flows belonging to this IEA such that the rate of the remaining flows does not exceed the sustainable rate. Thus, the SM proposal also implements MRT.

3.2.3 Some Observations about Measured Rate Termination (MRT)

Both CL and SM use MRT to terminate flows. To get sufficiently accurate measurement results, the measurement interval needs to be long enough which introduces some delay. To estimate the traffic rate to be terminated, MRT requires the notion of an IEA. This limits the general applicability of MRT. For instance, it is not applicable for end-to-end PCN that is introduced in Sect. 4. The flows to be terminated need to be taken from that IEA for which SR-pre-congestion was observed. When only some ET-marked (for CL, AS-marked for SM) packets are received by the egress node, it is hard to decide whether none or one flow should be terminated if the IEA carries only a small number of flows. The FT entity needs relatively good estimates about the flow rates. Wrong estimates easily lead to overtermination or undertermination because MRT terminates the traffic in one shot. In the latter case, another termination step is required after some minimum inter-termination time. The FT entity chooses some flows of the IEA for termination. Thereby it implicitly assumes that any flow of the IEA contributes to the SR-overload on the SR-pre-congested link. This is not necessarily true. In case of multipath routing, e.g. ECMP, flows of the same IEA are possibly carried over different paths. As a consequence, MRT possibly tears down flows that do not contribute to SR-overload until also some flows are terminated that have caused the SR-pre-congestion observed for IEA. The fact that MRT does not work well with multipath routing is a rather severe constraint since multipath routing is often used for traffic engineering and resilience reasons. Marked flow termination (MFT) which is introduced, explained, and discussed in Sect. 5 does not have these problems.


Source
PCN Domain
Destination
End-to-end flow
PCN ingress node
PCN egress node
MM
S/MM
S/MM
End-to-end resource signalling
S
S
S
RSVP
Capacity Overprovisioning
S
Router with signalling functionality
MM
Router with metering & marking functionality
PCN Endpoint
PCN Domain
PCN Endpoint
End-to-end flow
MM
MM
MM
MM
MM
MM
MM
Router with metering & marking functionality

(a) Edge-to-edge PCN is triggered by admission requests from external signalling protocols and guarantees QoS within a single PCN domain.

(b) End-to-end PCN flows receive preferred treatment from PCN-capable nodes; they transparently traverse edge-to-edge PCN domains within a large trust domain and perceive them as islands with all nodes being PCN-capable.

Figure 2: Comparison of edge-to-edge and end-to-end PCN.

4 End-to-End PCN

In this section we propose end-to-end PCN. We explain its idea and benefits. We argue why it requires other AC and FT mechanisms than currently suggested for edge-to-edge PCN. We clarify where end-to-end PCN can be applied and where not due to trust issues.

4.1 Idea and Benefits

End-to-end PCN can be applied in a large network consisting of possibly multiple domains. It removes the notion of PCN domains, PCN ingress nodes, and PCN egress nodes. End systems, i.e. source and destination of PCN flows, or proxies thereof implement the control logic for PCN-based AC and FT instead of boundary nodes of PCN domains (cf. Fig. 2(b)).

As a consequence, signalling protocols like RSVP and per-flow states at domain boundary nodes become obsolete. End-to-end PCN assumes that all nodes with critical links perform PCN metering and marking. Exhaustive marking based on admissible link rates produces AS-marks and excess marking based on supportable link rates produces ET-marks. This is like in the CL proposal.

4.2 Need for New AC and FT Mechanisms

CL's and SM's CLE-based AC relies on the feedback of already admitted flows between PCN ingress and egress nodes. Since such IEAs do not exist in an end-to-end PCN context, CLE-based AC is not applicable. However, probe-based AC may be used instead. If a flow requests admission, the PCN source sends one or more probe messages to the PCN destination of the new flow and rejects the request if at least one probe message is AS-marked [24]. In a similar way CL's and SM's FT mechanisms rely on the feedback of IEAs. Therefore, end-to-end PCN requires also new FT mechanisms that decide based on the markings of a single flow whether this flow needs to be terminated or not. Unlike MRT, MFT methods fulfill this requirement. They are presented in Sect. 5.

4.3 Trust Issues

A PCN endpoint may cheat, i.e. it admits a new flow although it should block it, or it does not terminate a flow although it should. This is an issue when the endpoint is not under the control of the network operator. When end-to-end PCN is spanned over several trust domains, a cheating endpoint may affect the traffic load in another trust domain which can neither detect or track the cheating node and it is difficult to take countermeasures. Therefore, end-to-end PCN is rather a solution for large trust domains like corporate networks but probably not for the general Internet. The philosophy of end-to-end PCN is similar to the one of end-to-end ECN in the sense that core nodes believe that end systems react to the packet markings. However, anti-cheating mechanisms exist for ECN that detect cheating TCP receivers such that TCP senders can take countermeasures [5]. Possibly similar mechanisms can be developed for end-to-end PCN to allow its deployment for the general Internet.

4.4 From Edge-to-Edge to End-to-End PCN

End-to-end PCN is only a long-term vision because to be effective, it requires that most nodes of the entire trust domain implement PCN metering and marking on their links. Edge-to-edge PCN is a mid-term goal that helps to facilitate the resource management of a single DiffServ domain. Furthermore, edge-to-edge PCN prepares the ground for end-to-end PCN because it fosters the deployment of PCN-enabled nodes in the entire trust domain. When sufficiently many PCN-enabled clouds exist in the large trust domain, end-to-end PCN packets traverse their PCN nodes without receiving special treatment by the edge nodes of these clouds. This is illustrated in Fig. 2(b).

4.5 Integration of Edge-to-Edge and End-to-End PCN

When different PCN domains implement different metering and marking behavior and use different encoding of the marking, a deployment of end-to-end PCN is not possible. Therefore, metering and marking behavior of PCN nodes must have the same semantics and the encoding of the PCN marks needs to be standardized.

When different AC and FT mechanisms are used for edge-to-edge and end-to-end PCN, they should coexist in a fair way. Different AC mechanisms should have approximately the same blocking behavior for the same PCN feedback; otherwise the AC mechanism blocking at a slightly higher PCN rate can starve flows being subject to the other AC mechanism in the same network. This has been investigated in [24]. There is a similar issue with FT. Different FT mechanisms controlling traffic on a common link should have a similar termination behavior. Otherwise, the fast termination method removes the traffic under its control and the slow method does not need to react anymore. This leads to unfair termination probabilities.

5 Marked Flow Termination (MFT)

In this section we present marked flow termination (MFT) as an alternative to measured rate termination (MRT). When SR-pre-congestion occurs, packets are ET-marked. MFT methods terminate only "marked flows", i.e. those with at least one ET-marked packet. MFT has three major benefits compared to MRT. First, MFT can work on the basis of single flows and can support end-to-end PCN. Second, MFT terminates only marked flows which is an important feature in case of multipath routing. Third, MFT terminates SR-overload gradually which makes it an adaptive control whereas MRT terminates overload traffic in one shot. We propose three different methods for MFT that we also discuss in IETF for standardization [25, 26].

- The first MFT method terminates a flow as soon as one of its packets is marked. To avoid overtermination, it is crucial that only some packets of the SR-overload are marked. This can be achieved by excess marking with marking frequency reduction (MFR) which requires a modification to excess marking. Therefore, we call this termination method *MFT based on excess marking with MFR (MFT-MFR)*. It is applicable for both individual flows and IEAs.

- The second MFT method is preferably applied to individual flows. It uses plain excess marking like CL and terminates a flow only when it has received already several ET-marks. We call it *MFT based on plain excess marking for individual flows (MFT-IF)*.

- The third MFT method can be applied only to IEAs. It also uses plain excess marking like MFT-IF and removes a marked flow from an IEA only when the IEA has received several ET-marked packets. We call it *MFT based on plain excess marking for ingress-egress aggregates (MFT-IEA)*. In contrast to MFT-IF, MFT-IEA has better support for termination policies.

We explain these mechanisms in detail, illustrate their termination behavior, give recommendations for their configuration, evaluate their performance, point out their shortcomings, and compare their pros and cons.

5.1 Marked Flow Termination Based on Excess Marking with Marking Frequency Reduction (MFT-MFR)

We present the basic version of MFT-MFR. It requires a modification of the basic excess metering and marking algorithm which is called marking frequency reduction (MFR). We explain the simulation setup for our performance evaluation and show termination behavior for MFT-MFR. We derive suitable configuration parameters and illustrate the impact of the aggressiveness α, the main control parameter for all MFT methods. We propose proportional marking frequency reduction (PMFR) and packet size independent marking (PSIM) as further extensions of the excess marker and show how they improve the control and the fairness of the termination process.

5.1.1 The Mechanism

With MFT-MFR, a PCN egress node or endpoint terminates a flow as soon as it receives one of its packets with an ET-mark. The duration between a PCN node marks a packet until it sees the last packet of that flow is called the flow termination delay D_T. In general it is flow-specific. For end-to-end PCN, the end-to-end round trip time is a lower bound for that value since the end systems terminate the flows. For edge-to-edge PCN, the edge-to-edge round trip time is a lower bound for that value since the PCN egress node tells the PCN ingress node to drop further packets of the terminated flow. In our study we assume $D_T = 50$ ms for local networks, $D_T = 200$ ms for national networks, and $D_T = 500$ ms for transatlantic or satellite networks.

5.1.2 Plain Excess Marking and Excess Marking with Marking Frequency Reduction

Algorithm 1 EXCESS MARKING: (packet size independent) excess marking with (proportional) marking frequency reduction.

```
Input:  token bucket parameters $S$, $R$, $F$, $lU$, packet size $B$ and marking $M$, current
        time $now$, maximum transfer unit $MTU$, increment $I$ or stretch factor $\beta_\alpha$
  $F = \min(S, F + (now - lU) \cdot R)$;
  $lU = now$;
  if $(F \geq B)$ then    {PSIM: $(F \geq MTU)$}
    $F = F - B$;
  else
    $M = ET$;
  end if
  if $(M == ET)$ then    {Marking frequency reduction}
    $F = \min(S, F + I)$;
      {PMFR: $F = \min(S, F + \beta_\alpha \cdot B)$;}
  end if
```

If the PCN rate $r(l)$ on a link l is significantly lower than $SR(l)$ after the termination, we talk about overtermination as more flows than necessary have been terminated. This occurs with MFT-MFR when too many packets are ET-marked. Therefore, the excess marker should mark only a subset of the packets that exceed SR. This can be done by excess marking with marking frequency reduction (MFR). We first explain the basic excess marking algorithm which we also call plain excess marking and then excess marking with MFR.

Plain Excess Marking Algorithm 1 provides a token bucket (TB) based formulation of the excess marker. It is called at each packet arrival. The marker has a TB which is S bytes large and constantly filled with tokens at rate R. The TB variable lU remembers the last update and helps to account for the new tokens since the last call of the algorithm. If the fill state F of the bucket is at least the size B of the arrived packet, B tokens are removed from the bucket; otherwise, the marking of the packet is set to $M = ET$. The algorithm so far describes plain excess marking.

Excess Marking with MFR MFR adds an increment of I bytes to the bucket if the packet is ET-marked – no matter whether it was ET-marked by this call of the algorithm or whether it was already marked before. This is the sub-optimal base algorithm. The alternative statements in the curly braces implement PSIM and PMFR which are covered in Sections 5.1.7 and 5.1.6.

Configuration Parameters The TB rate $R = SR(l)$ is set to the supportable rate of the monitored link l and its bucket size is set to a sufficiently large value which is $S = 50$ KB in our simulations. When the increment is set to $I = 0$, Algorithm 1 performs plain excess marking, i.e., all packets exceeding the SR of the link are marked with ET. Let $E[B]$ be the average packet size.[1] When the increment is larger than zero and a packet is marked, $\frac{I}{E[B]}$ additional packets can pass the marker without being marked compared to plain excess marking. As a result, the marking frequency is reduced by a factor of

$$\sigma_p = \frac{I}{E[B]} + 1 \tag{1}$$

in case of SR-pre-congestion.

5.1.3 Simulation of the Termination Behavior

Our objective is to investigae the termination behavior of MFT. To that end we assume sudden SR-pre-congestion on a link as it appears for instance due to rerouted traffic when fast failover mechanisms like MPLS fast reroute are used.

The inter-arrival time A and the packet size B of the flows are mostly deterministic. If not mentioned differently, they have average values of $E[A] = 20$ ms and $E[B] = 200$ bytes such that their rate is $E[R] = 80$ kbit/s. To avoid simulation artifacts due to marking synchronization for periodic traffic, we add an equally distributed random delay of up to 1 ms to the theoretic arrival instant of every packet. This traffic model is realistic because realtime applications send traffic periodically, but packets may arrive at the bottleneck link with some jitter.

[1] $E[X]$ is the mean and $c_{var}[X]$ the coefficient of variation of a random variable X.

We simulate the time-dependent PCN rate $r(l,t)$ of a link l and to study the termination process of the time-dependent SR-overload $SRO(l,t)$. The supportable link rate is $SR(l) = 8$ Mbit/s and the initial SR-overload corresponds to 100%, i.e., the initial SR-overload is also $SRO(l,0) = 8$ Mbit/s such that the initial PCN rate is $r(l,0) = 16$ Mbit/s. Hence, the initial number of flows is $n = 200$, but only $n = 100$ PCN flows can be supported.

We use a custom-made Java tool to simulate the time-dependent PCN rate $r(l,t)$ to illustrate the termination behavior.[2] This rate is calculated based on 50 ms long measurement intervals. We perform multiple experiments and report average results for the termination behavior in our figures. We run so many simulations that the 95% confidence intervals for the time-dependent PCN rate values $r(t)$ are small. However, we omit them in the figures for the sake of easier readability.[3]

5.1.4 Configuration of the Increment

As soon as the PCN rate $r(t)$ exceeds SR on a link, the token bucket empties, the PCN node starts marking packets after a while, and flows are terminated. However, the rate reduction becomes visible only after a flow termination delay D_T. Thus, for the first D_T interval the SR-overload $SRO(t)$ is the initial value $SRO(0)$, and for the second D_T interval $SRO(0)$ is still a low upper bound since the PCN traffic rate starts decreasing only at D_T. Roughly speaking, $\frac{2 \cdot D_T \cdot SRO(0)}{E[B]}$ packets are over SR within the first two D_T intervals, and $\frac{2 \cdot D_T \cdot SRO(0)}{E[B] \cdot \sigma_p}$ of them are marked (cf. Eqn. (1)) which is also the maximum number of terminated flows. To avoid overtermination, their rate should be less than the initial SR-overload, i.e. $\frac{2 \cdot D_T \cdot SRO(0)}{E[B] \cdot \sigma_p} \cdot E[R] \leq SRO(0)$. This is achieved when the marking frequency reduction is at least $\sigma_p \geq \frac{2 \cdot D_T \cdot E[R]}{E[B]}$, and the increment I is at least $I \geq 2 \cdot D_T \cdot E[R] - E[B]$. This sketch is rather a motivation than a rigid mathematical proof, but simulation results show that this inequality is sharp.

5.1.5 Termination Aggressiveness α

To control the speed of the termination process, we introduce the aggressiveness α and use it to calculate the increment

$$I_\alpha = \frac{2 \cdot E[D_T] \cdot E[R] - E[B]}{\alpha}. \tag{2}$$

The aggressiveness is defined such that the termination speed increases with α and that overtermination is avoided for $\alpha < 1$, at least for homogeneous traffic. This is illustrated by Fig. 3. The degree of overtermination also increases with α.

5.1.6 Proportional Marking Frequency Reduction (PMFR)

To keep MFT-MFR simple, the increment I_α is configured in the PCN nodes only once based on estimated values $E[B^*]$, $E[R^*]$, $E[D_T^*]$, and a desired α^*, and it is not adjusted to the

[2] Most rate variables depend on l and t, but we omit l or t when the context is clear.

[3] Even in case of strictly periodic traffic, i.e., the inter-arrival times and the sizes of the packets are constant, different runs produce different results because the first transmission of a flow within a first inter-arrival time after simulation start is random.

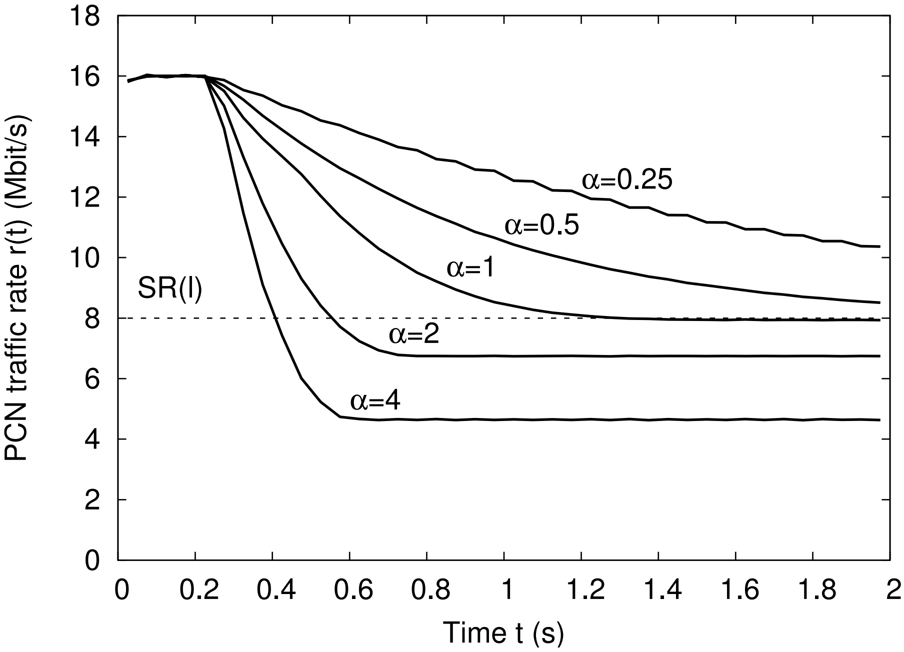


Figure 3: The aggressiveness α controls the speed of the termination process and the degree of potential overtermination.

current traffic characteristics. We configure I_α for the default values in Sect. 5.1.3 and $\alpha^* = 1$ according to Eqn. (2), but vary the actual packet sizes $E[B]$ such that also the actual flow rate $E[R]$ is affected. This leads to an actual termination aggressiveness $\alpha = \frac{2 \cdot E[D_T] \cdot E[R] - E[B]}{I^*_\alpha} = \alpha^* \cdot \frac{E[B]}{E[B^*]}$. As a result, the resulting termination behavior can be essentially derived from Fig. 3 for given $E[B]$. Hence, the termination behavior of MFT-MFR significantly depends on the average packet sizes. However, it is possible to make it independent of the packet size by applying proportional marking frequency reduction (PMFR) in Algorithm 1. We calculate the increment

$$I_\alpha = \frac{2 \cdot E[D_T] \cdot \frac{1}{E[A]} - 1}{\alpha} \cdot B = \beta_\alpha \cdot B \tag{3}$$

using the stretch factor β_α such that the increment is proportional to the size of the ET-marked packet. As a consequence, the marking algorithm is characterized by the fact that one packet is marked for

$$\sigma_b = \beta_\alpha \cdot E[B] + E[B] = \frac{2 \cdot E[D_T] \cdot E[R]}{\alpha} \tag{4}$$

bytes that have been above SR during a continuous SR-pre-congestion phase. Therefore, PMFR makes the termination behavior of MFT-MFR independent of $E[B]$. This has been validated by simulation (no figure).

5.1.7 Packet Size Independent Marking (PSIM)

We proceed from homogeneous to heterogeneous traffic mixes. We consider constant bit rate flows with the same packet inter-arrival time $E[A] = 20$ ms and a constant packet size per flow. However, different flows may have different packet sizes $E[B]$ that also influences the flow rate R_f. We consider traffic mixes according to Table 1 where the parameter t determines the fraction of low, medium, and high bit rate flows in the traffic mix. The average flow rate in the traffic mix is $E[R] = 80$ kbit/s and the average packet size is $E[B] = 200$ bytes. However, the variability of the flow rates and flow-specific packet sizes depends on t and the coefficient of variation is $c_{var} = 1.5 \cdot \sqrt{t}$.

Table 1: Traffic mixes with $E[R] = 80$ kbit/s and $c_{var}[R] = 1.5 \cdot \sqrt{t}$. The variable t controls the proportion of low, medium, and high bit rate flows in the traffic mix.

	Flow types		
Flow type specific	low bit rate	medium bit rate	high bit rate
Proportion	$0.8 \cdot t$	$1 - t$	$0.2 \cdot t$
$E[B]$ for $E[A] = 20$ ms	50 bytes	200 bytes	800 bytes
$E[A]$ for $E[B] = 200$ bytes	80 ms	20 ms	5 ms
Rate $E[R]$	20 kbit/s	80 kbit/s	320 kbit/s

Experiments have shown that the termination behavior for highly variable traffic mixes ($t = 1$) is almost the same as for traffic with homogenous packet sizes ($t = 0$, cf. Fig. 3). However, Table 2 shows that flows with large packets have a tremendously higher termination probability than flows with small packets. Therefore, we introduce packet size independent marking (PSIM) for excess marking. This is achieved by making the marking decision in Algorithm 1 only dependent on the fill state of the token bucket but not on the packet size. With this change, low, medium, and high bit rate flows have the same termination probability and the termination behavior is still independent of the traffic mix.

5.1.8 Impact of Packet Inter-Arrival Times

Like in Sect. 5.1.6 we assume that the stretch factor β_α of Eqn. (3) is calculated only once based on estimated values $E[D_T^*]$, $E[A^*]$, and a desired aggressiveness α^*. To test the impact of packet inter-arrival times within a single flow, we use the default values in Sect. 5.1.3 and $\alpha^* = 1$ for the configuration of β_α. Varying the actual inter-arrival time $E[A]$ and keeping all other parameters constant leads to a different aggressiveness $\alpha = \frac{E[A^*]}{E[A]} \cdot \alpha^*$: increasing the actual inter-arrival time decreases the aggressiveness and vice-versa. With this knowledge, the resulting termination behavior can be essentially derived from Fig. 3 for various $E[A]$.

Hence, the termination behavior of MFT-MFR significantly depends on the packet inter-arrival times for fixed β_α. In practice we need a viable solution that reduces the SR-overload quickly while avoiding overtermination. Most realtime applications send one packet within 20 ms, some others have a period of 10 ms. Video applications are slower but possibly send

Table 2: Packet size (B) and inter-arrival time (A) dependent flow blocking probabilities for MFT-MFR.

Traffic mix	Different B, $\alpha = 1$, PMFR without PSIM		
	$E[B] = 50$ bytes	$E[B] = 200$ bytes	$E[B] = 800$ bytes
$t = 0$	-	0.501	-
$t = 0.5$	0.023	0.247	0.942
$t = 1$	0.006	-	0.625

Traffic mix	Different A, $\alpha = 0.5$, cf. Fig. 4		
	$E[A] = 80$ ms	$E[A] = 20$ ms	$E[A] = 5$ ms
$t = 0$	-	0.494	-
$t = 0.5$	0.119	0.348	0.792
$t = 1$	0.077	-	0.630

several packets for one frame. We recommend to use an aggressiveness of $\alpha = 0.5$ and an inter-arrival time of $E[A] = 20$ ms for the configuration of the stretch factor in Eqn. (3). This corresponds to an aggressiveness of $\alpha = 1$ for $E[A] = 10$ ms such that overtermination is not likely to occur with today's applications. If the actual inter-arrival time is in fact $E[A] = 20$ ms, the reduction of SR-overload to about 10% is still fast as it takes only 1.7 s (cf. Fig. 3, $\alpha = 0.5$).

We study the impact of traffic mixes consisting of different constant bit rate flows according to Table 1. The packet sizes and inter-arrival times within a single flow are constant, but different flows have different packet inter-arrival times. The average inter-arrival time over all flows is $E[A] = 20$ ms, but its variability depends on t. We configure the stretch factor β_α based on an aggressiveness $\alpha = 0.5$. Fig. 4 shows that the termination speed depends on the traffic mix: more variable inter-arrival times lead to faster termination. Table 2 shows that flows with small packet inter-arrival times have a tremendously larger flow termination probability. This is due to the fact that the probability for a flow to have one of its packets ET-marked increases when it sends more packets. Since large flows are more likely to be terminated first, the termination process for heterogeneous traffic is faster than for homogeneous traffic and prone to overtermination. However, overtermination is almost fully avoided in the experiment because the aggressiveness $\alpha = 0.5$ is chosen low enough. Unfortunately, we do not know any simple mechanism to balance the termination probability among flows with different inter-arrival times.

5.2 Marked Flow Termination Based on Plain Excess Marking for Individual Flows (MFT-IF)

We explain the operation of MFT-IF and propose a suitable initialization method. The termination process can be well controlled for heterogeneous flows when reasonable estimates of their rates are available. We show that it is possible to implement stochastic termination priorities.

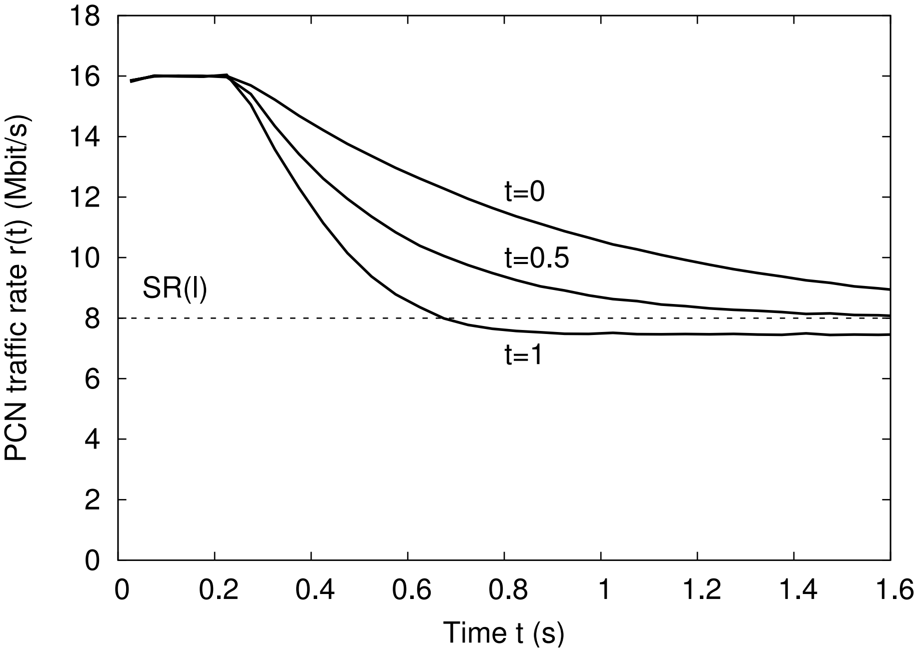


Figure 4: Traffic with more variable inter-arrival times leads to faster termination and flows with shorter inter-arrival times have higher termination probabilities.

5.2.1 The Mechanism

MFT-IF requires that PCN nodes ET-mark all packets that exceed the SR of their links. Like MFT-MFR, it also requires packet size independent excess marking (PSIM). The PCN endpoint of a flow f sets up a flow-specific credit counter C_f. If a flow's packet arrives ET-marked and its credit counter is positive, its credit counter is decreased by the size of the packet. If the counter is zero or negative at the arrival of an ET-marked packet, the flow is terminated.

5.2.2 Counter Initialization

We suggest a method for the initialization of the credit counters. We lend ideas from our analysis of MFT-MFR. MFT-MFR's termination speed is controlled by the fact that the next packet is ET-marked only after σ_b bytes have exceeded SR since the last packet was ET-marked (cf. Eqn. (4)). We mimic this fact and initialize the credit counters for MFT-IF in such a way that it achieves the same termination behavior as MFT-MFR.

We consider n flows numbered from $i = 1$ to n and having different counter initialization values C_i with $C_{i-1} < C_i$. We assume that they receive equally many ET-marked bytes in case of SR-pre-congestion. As a consequence, flows terminate in ascending order. When flow i terminates next, $n - (i - 1)$ flows are still active. To let σ_b marked bytes pass between the termination of flows $i - 1$ and i, the difference between their counters should be set to

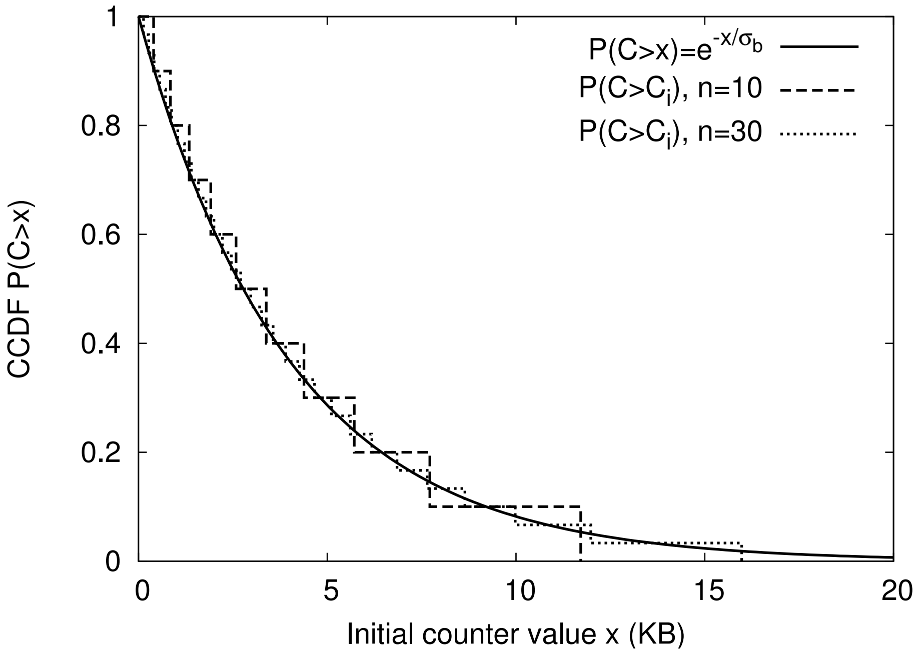


Figure 5: CCDF of the counter initialization values for a various number of n flows and their limiting function.

$C_i - C_{i-1} = \frac{\sigma_b}{n-(i-1)}$. With $C_0 = 0$, the counter initialization should be chosen

$$C_i = \sum_{0<k\leq i} \frac{\sigma_b}{n-(k-1)} = \sigma_b \cdot (H_n - H_{n-i}) = \sigma_b \cdot \ln(\frac{n}{n-i}) \tag{5}$$

with $H_i = \sum_{0<k\leq i} \frac{1}{k}$ being the i-th harmonic number for which the approximation $H_i \approx \ln(i) - \gamma$ holds when i is finite.[4] Experiments with this credit counter initialization show the same termination behavior as in Fig. 3.

Eqn. (5) can be used to initialize the credit counter of flows when all flows sharing a single bottleneck link are known. Now we develop an algorithm that allows a flow to initialize its credit counter randomly without knowing anything about other flows. Based on Eqn. (5), the complementary cumulative distribution function (CCDF) of the counter initialization values for n flows is $P(C > C_i) = P(C > \sigma_b \cdot \ln(\frac{n}{n-i})) = \frac{n-i}{n}$. Substituting $\sigma_b \cdot \ln(\frac{n}{n-i})$ by x we get

$$P(C > x) = \exp\left(\frac{-x}{\sigma_b}\right) = \exp\left(\frac{-x \cdot \alpha}{2 \cdot E[D_T] \cdot E[R]}\right) \tag{6}$$

for large n. Fig. 5 illustrates that the exact CCDFs for various numbers of flows n converge quickly towards the limiting CCDF of Eqn. (6). Therefore, we propose that a new flow f takes its own rate R_f as an estimate for $E[R]$ and randomly initializes its credit counter according to

[4] $\gamma = 0.57721...$ is the Euler-Mascheroni constant.

Eqn. (6). It picks a uniformly distributed random number $0 < y < 1$ and sets its credit counter to $C_f = -\frac{2 \cdot E[D_T] \cdot R_f}{\alpha} \cdot \ln(y)$.

When we substitute the deterministic initialization according to Eqn. (5) by the stochastic initialization according to Eqn. (6), we expect less control or at least more variance of the termination behavior. However, we tested this issue and found that the deviation from the average termination behavior is rather small. More evidence on the variability of the termination behavior is given in Sect. 5.4.5.

5.2.3 Impact of Packet Sizes and Inter-Arrival Times

With MFT-IF, the termination behavior is robust against traffic mixes consisting of flows with different packet sizes and inter-arrival times since the counter initialization takes this issue into account by using the flow rate R_f. Flows with different packet sizes or packet inter-arrival times have also the same termination probabilities.

5.2.4 Implementation of Stochastic Flow Termination Priorities

The initialization value of its credit counter heavily impacts the termination probability of a flow in case of SR-overload. Therefore, high priority flows should be assigned larger initial credit counters than low priority flows to have a better chance to survive SR-pre-congestion. We achieve that by using a smaller aggressiveness α to initialize the credit counters of high-priority flows.

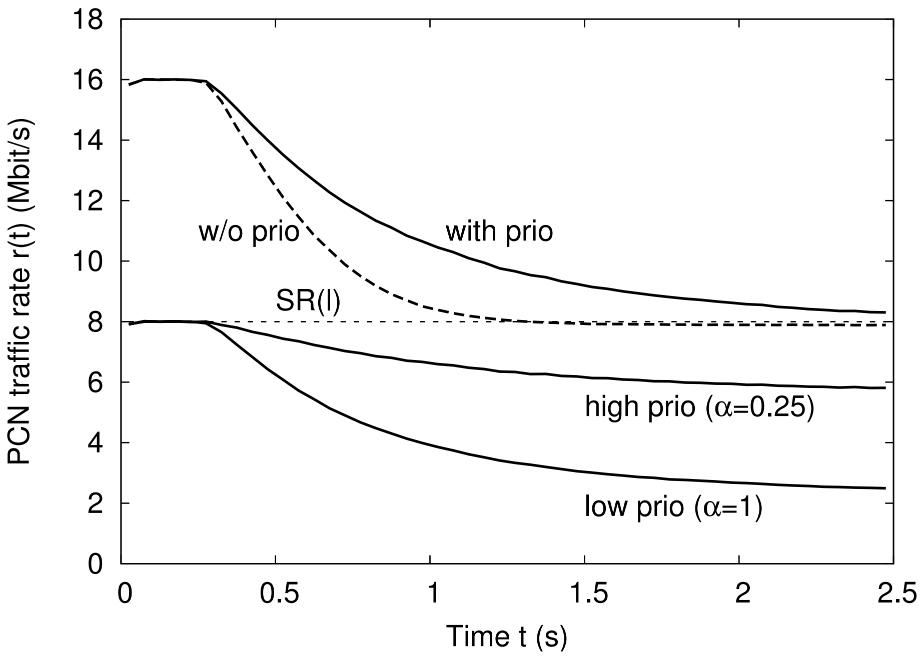


Figure 6: Termination behavior for high and low priority traffic.

We consider low-priority flows for which we use $\alpha = 1$ and high-priority flows for which we use $\alpha = 0.25$. Fig. 6 shows their individual and combined termination behavior. While the aggregate rate of low-priority flows is significantly reduced, the aggregate rate of high-priority flows is less decreased. Thus, high-priority flows have indeed a lower termination probability than low-priority flows. The dashed line is the termination behavior without prioritized flows ($\alpha = 1$). It shows that prioritization prolongs the duration of the termination process.

5.3 Marked Flow Termination Based on Plain Excess Marking for Ingress-Egress Aggregates (MFT-IEA)

With edge-to-edge PCN IEAs are available. We develop MFT-IEA as an enhanced version of MFT-IF that takes advantage of IEAs and works with the same marking behavior, i.e. packet size independent excess marking (cf. Sect. 5.2.1). The idea is to use MFT-IF for end-to-end PCN and MFT-IEA for edge-to-edge PCN. We study the impact of the aggregation level per IEA, of packet sizes and inter-arrival times, and illustrate stochastic enforcement of termination policies.

5.3.1 The Mechanism

MFT-IEA groups flows sharing a common PCN ingress and egress node into a common IEA. We denote the flow set of such an IEA g by $\mathcal{F}(g)$. The PCN egress node has a credit counter C_g for each of its IEAs g. When the PCN egress node receives an ET-marked packet that belongs to a flow $f \in \mathcal{F}(g)$, its size in bytes is subtracted from the counter C_g. If the counter is not positive at the arrival of an ET-marked packet, the flow f is terminated. In this case, the credit counter is decreased by the packet size and increased by an increment $I_\alpha = \sigma_b$ which is proportional to the flow rate R_f. An alternative design terminates a flow already if the size of the ET-marked packet is larger than the credit counter. On the one hand this is simpler, but on the other hand it leads to packet size dependent termination probabilities that we want to avoid. Hence, our design complements PSIM in the core and also influenced the design of the MFT-IF mechanism in Sect. 5.2.1.

5.3.2 Configuration of MFT-IEA

When a first flow joins the IEA g after system start, Eqns. (4) and (6) may be used to randomly initialize the credit counter C_g. To implement a similar control as for MFT-MFR, we choose an increment of

$$I_\alpha = \sigma_b = \frac{2 \cdot D_T \cdot R_f}{\alpha} \tag{7}$$

when a flow is terminated. Note that this equation differs from Eqn. (3) by the fact that the increment is proportional to the flow rate R_f instead of the packet size.

5.3.3 Impact of the Number of IEAs Sharing a Bottleneck Link

We conduct an experiment with $n = 200$ flows that share a bottleneck link and belong to m different IEAs with $\frac{n}{m}$ flows each. The average termination behavior is about the same for

$m \in \{1, 10, 50, 200\}$ IEAs. We omit the figure for this experiment. The termination process for $m = 200$ IEAs just starts 20 ms later than for $m = 1$ IEA. Note that we always run our experiments many times. The interval between the 5% and 95%-quantiles for the time-dependent aggregate rate $r(t)$ is small for all studied m and only slightly larger for many IEAs than for only one IEA. The curves for $m = 10$ and highly variable flow rates are shown in Fig. 10. Hence, this method also works well with rather small IEAs. Moreover, for $m = 200$ each IEA contains only a single flow such that the termination behavior of MFT-IEA becomes the one of MFT-IF. Hence, the experiment also shows that MFT-IF can fairly coexist with MFT-IEA as both MFT methods have the same termination behavior.

5.3.4 Impact of Packet Sizes and Inter-Arrival Times

With MFT-IEA, the termination behavior on the bottleneck link and the flow termination probabilities are insensitive to the average packet size and its variation within flows. The termination behavior is also rather independent of the average inter-arrival time since the increment defined in Eqn. (7) is based only on the flow rate.

We now consider traffic mixes of flows with different inter-arrival times. Flows with a higher packet frequency have a higher termination probability since it is more likely that one of their ET-marked packets sees a non-positive credit counter at their arrival compared to flows with a lower packet frequency. We show this phenomenon by an experiment. We consider traffic mixes of flows having different inter-arrival times according to Table 1 and flows of different types are equally assigned to $m = 10$ IEAs. Table 3 illustrates that the termination probabilities of high bit rate flows are larger than those for low bit rate flows. This is similar to MFT-MFR where different flow termination probabilities also impact the termination behavior (cf. Fig. 4). In contrast to MFT-MFR, with MFT-IEA the termination behavior for heterogeneous traffic hardly differs from the one of homogeneous traffic. This is due to the fact that MFT-MFR's increment is only proportional to the packet size of the terminated flow while MFT-IEA's increment is proportional to its rate.

Table 3: Flow termination probabilities for MFT-IEA depending on the traffic mix. All flows have a fixed packet size of 200 bytes but different inter-arrival times.

Rate	20 kbit/s	80 kbit/s	320 kbit/s
$E[A]$	80 ms	20 ms	5 ms
$t = 0$	-	0.507	-
$t = 0.5$	0.096	0.317	0.861
$t = 1$	0.060	-	0.647

When we group the heterogeneous flows in such a way that IEAs have only flows with equal inter-arrival times, the effect of different termination probabilities vanishes. Thus, edge-to-edge PCN might define separate sub-IEAs for flows with low and high packet frequency to achieve fair termination probabilities.

5.3.5 Stochastic Enforcement of Termination Policies

Stochastic termination priorities can be implemented similarly as in Sect. 5.2.4: low and high priority flows are grouped into different IEAs that are configured with larger and smaller aggressiveness. In addition to such termination priorities, we propose stochastic enforcement of termination policies. When an ET-marked packet arrives and the credit counter is not positive, a flow must be terminated. However, this is not necessarily the flow to which the newly arrived packet belongs to. Basically, any other flow from the same IEA can be terminated. However, to cope with multipath routing, the other flow must have been recently ET-marked, too. Thus, MFT-IEA needs to remember the set of ET-marked flows and can choose a flow from this set according to some policy when a flow needs to be terminated. We call this stochastic policy enforcement because the policies can only be applied to the recently ET-marked flows which is stochastic.

Table 4: Flow termination probabilities for MFT-IEA and different policies depending on the number of flows $\frac{n}{m}$ per aggregate.

Rate	40 kbit/s	160 kbit/s	40 kbit/s	160 kbit/s	40 kbit/s	160 kbit/s
$\frac{n}{m}$	No priorities		Large flows first		Small flows first	
250	0.286	0.721	0.037	1.000	0.987	0.067
25	0.275	0.741	0.029	0.986	0.962	0.132
5	0.186	0.827	0.047	0.929	0.809	0.379

We perform some experiments to show the effectiveness of stochastic policy enforcement. In the first experiment, we consider 200 flows with 40 kbit/s ($E[A] = 40$ ms) and 50 flows with 160 kbit/s ($E[A] = 10$ ms) such that half of the traffic volume results from low and high bit rate flows. We group them equally into m different aggregates with $\frac{250}{m}$ flows each. Table 4 shows that when no policy is applied, large flows have a significantly higher termination probability due to their larger packet frequency. When large flows are terminated first, only 2.9%–4.7% of the small flows are terminated but 92.9%–100% of the large flows. In contrast, when small flows are terminated first, 6.7%–37.9% of the large flows are still terminated and 80.9%–98.7% of the small flows. The table also shows that stochastic policy enforcement is more effective on larger aggregates. Thus, the effectiveness of stochastic policy enforcement depends both on the aggregation level of the IEA and the policy itself.

5.4 General Performance of MFT Methods

In this section, we study aspects that are common to all three MFT methods: MFT with marking frequency reduction (MFT-MFR), MFT with plain excess marking for individual flows (MFT-IF) and for IEAs (MFT-IEA). For MFT-IEA we assume in our simulations that 200 flows on the bottleneck link are evenly split among $m = 10$ IEAs. We study the impact on the termination behavior of the flow termination delay D_T, the aggregation level on the bottleneck link, the degree of SR-overload, packet loss, the variability of the termination process, per aggregate fairness, and various traffic characteristics.

5.4.1 Impact of Flow Termination Delays

We study the impact of the duration of the flow termination delay D_T on the termination behavior, of wrong D_T, and of different D_T. The results are the same for all MFT methods.

Duration of Flow Termination Delays The time to terminate the overload increases linearly with D_T for all MFT methods when configured appropriately. This result is almost trivial and we do not illustrate it by a figure.

Wrong Flow Termination Delays We assume that MFT-MFR, MFT-IF, and MFT-IEA are configured for an expected flow termination delay of $E[D_T^*] = 200$ ms and a target aggressiveness $\alpha^* = 1$ using the configuration formulae in Eqns. (4), (6), and (7). If the actual flow termination delay $E[D_T]$ is different from $E[D_T^*]$, the actual aggressiveness is $\alpha = \frac{E[D_T]}{E[D_T^*]} \cdot \alpha^*$. Thus, the actual aggressiveness is proportional to the actual flow termination delay $E[D_T]$. With this knowledge, the resulting termination behavior can be derived from Fig. 3 for various $E[D_T]$.

Different Flow Termination Delays We assume that half of the flows on a bottleneck link have a flow termination delay of $D_T = 50$ ms and the other half has $D_T = 500$ ms. We choose this very extreme setting to make the impact of different D_T clearly visible. We use the average value $E[D_T] = 275$ ms to configure the stretch factor β_α of the marking algorithm for MFT-MFR in Eqn. (3), to initialize all credit counters for MFT-IF and MFT-IEA in Eqn. (6), and to calculate the rate-dependent increments for MFT-IEA in Eqn. (7).

Fig. 7 illustrates the termination behavior of MFT-MFR. The time-dependent aggregate rate of the flows with $D_T = 50$ ms starts decreasing early while the one of the flows with $D_T = 500$ ms starts decreasing rather late (solid lines). However, they both converge to their fair share of 4 Mbit/s. The reason for that phenomenon is that the packets of all flows passing the SR-pre-congested link experience the same marking probabilities. Therefore, with MFT-MFR the termination probability of flows is independent of D_T. Surprisingly, we get the same results for MFT-IF and MFT-IEA. Like with MFT-MFR, the marking probability of the packets is independent of the flow termination delay D_T. Therefore, no compensation for large or small D_T is needed for the initialization of the credit counters or the calculation of the rate-dependent increments and they work well with $E[D_T]$.

The combined time-dependent rate of flows with short and long D_T reveals a different shape but a very similar termination speed compared to the same number of flows with a homogeneous flow termination delay of $D_T = 275$ ms (dotted line).

For MFT-IF and MFT-IEA, we have the option to use the flow-specific D_T for the initialization of the credit counters and the rate-dependent increment. In that case, the rate of flows with short D_T drops extremely fast and the rate of flows with long D_T drops very slowly (dashed lines). Their combined rate decays faster than those in the experiments above. The rates converge to different values. This is unfair as it entails different termination probabilities for flows with small and large D_T. Thus, for the sake of fairness, the same average value

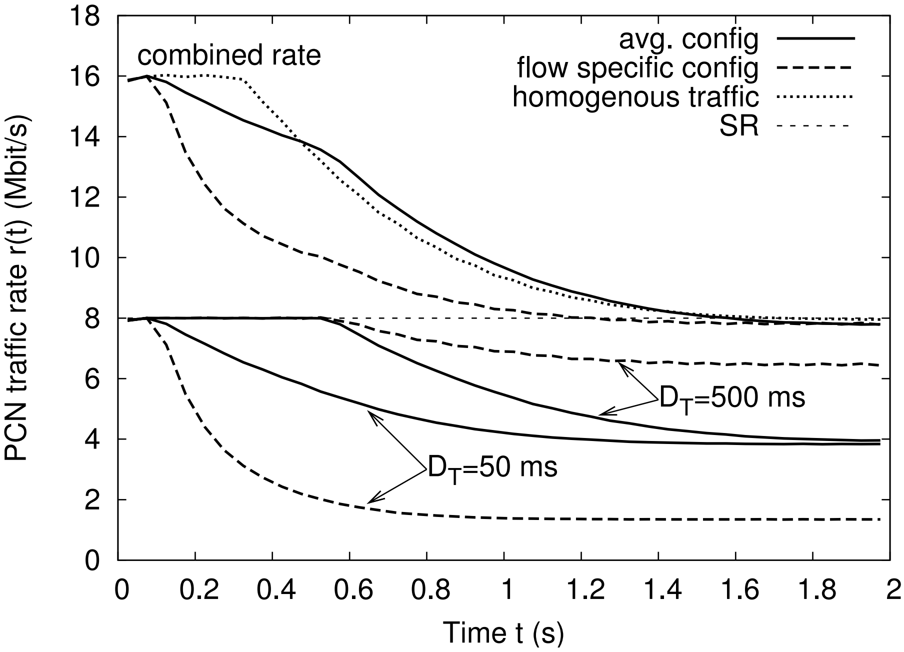


Figure 7: In spite of different flow termination delays D_T all flows have the same termination probability when all system components are configured with an average value $E[D_T]$.

$E[D_T]$ should be applied for the configuration of all distributed PCN egress nodes or endpoints. However, the choice of this network-wide or global value needs to be taken carefully because it influences the actual aggressiveness and thereby the termination speed and the degree of potential overtermination. There is no such debate with MFT-MFR as its edge systems act independently of $E[D_T]$.

5.4.2 Impact of the Aggregation Level

We consider $n \in \{20, 200, 2000\}$ flows on the bottleneck link and scale the supportable rate SR of the link and its marking parameters accordingly. We apply $\alpha = 1$ to achieve fastest overload reduction without overtermination. We perform one experiment series using flows with homogeneous traffic rates and another using flows with heterogeneous traffic rates (different packet sizes). We omit the figures with the simulation results but report the findings. The relative shape of the termination behavior is the same for all experiments and for all considered MFT methods except for low aggregation. In particular the time to reduce the overload is the same and there is no significant overtermination. For low aggregation we observe a slightly delayed termination process and in addition some small overtermination for heterogeneous traffic.

5.4.3 Impact of the SR-Overload Intensity

We set the initial PCN rate to 12, 16, and 24 Mbit/s such that the resulting SR-overload is 4, 8, and 16 Mbit/s which corresponds to an SR-overload of 50%, 100%, and 200%. Fig. 8 shows that all three MFT methods yield the same termination behavior. This is an important finding because it shows that MFT-IF applied for end-to-end PCN and MFT-IEA applied for edge-to-edge PCN can coexist in a fair way, i.e., flows controlled by MFT-IF or MFT-IEA have the same blocking probabilities. We again observe that overtermination does not occur for $\alpha = 1$. As mentioned in Sect. 5.1.4, with $\alpha = 1$ about half of the SR-overload is terminated within a single D_T. Therefore, the termination of 8 Mbit/s and 16 Mbit/s SR-overload takes about D_T and $2 \cdot D_T$ longer than the termination of of 4 Mbit/s overload.

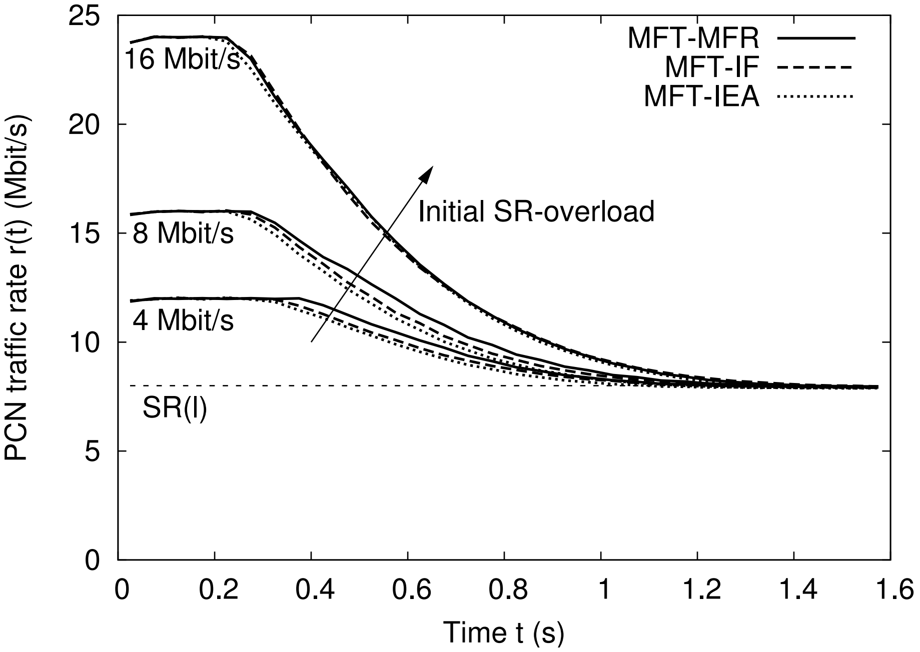
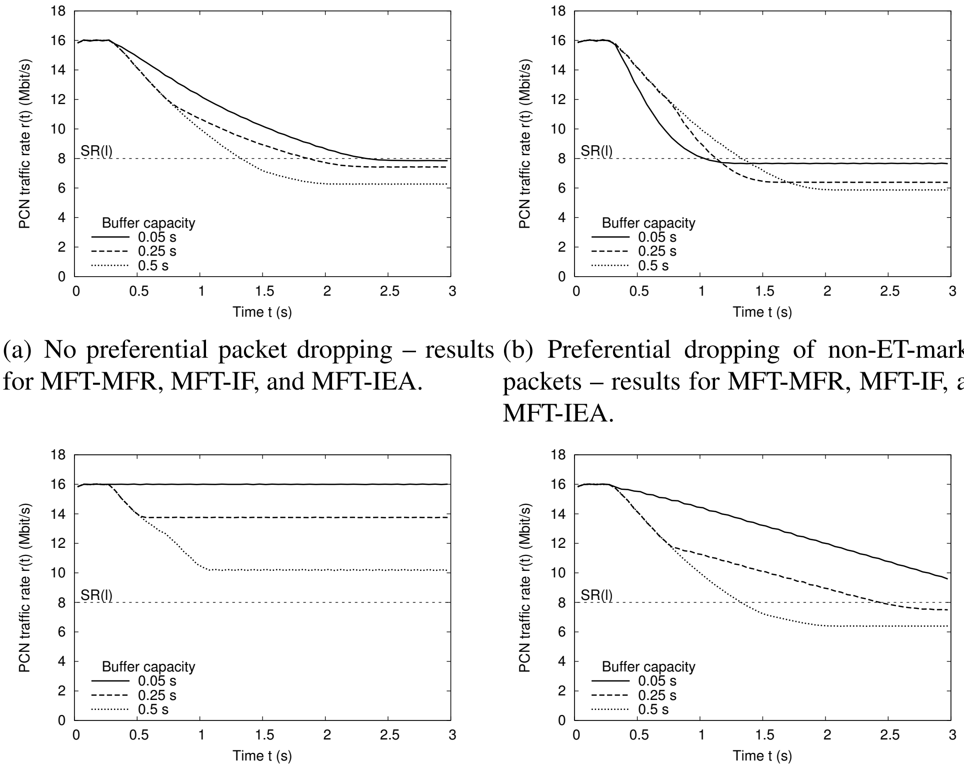


Figure 8: Impact of the SR-overload SRO on the termination behavior.

5.4.4 Impact of Packet Loss

MFT requires marked packets to trigger the termination process. In case of packet loss, ET-marked packets may be lost which possibly delays the termination process. We consider a bottleneck link with $SR = 8$ Mbit/s, a limited capacity of 9 Mbit/s, and an initial PCN traffic rate of 16 Mbit/s such that 43.75% is lost. Before packet loss occurs, the packet buffer fills up. We set the buffer size such that it can accommodate the amount of traffic that can be sent within 0.05 s, 0.25 s, or 0.5 s at the bottleneck bandwidth of 9 Mbit/s. The termination aggressiveness is set to $\alpha = 1$ and the average flow termination delay is $E[D_T] = 0.2$ s. We consider three packet drop options: no preferential packet drop, preferential drop of non-ET-marked packets,



(a) No preferential packet dropping – results for MFT-MFR, MFT-IF, and MFT-IEA.

(b) Preferential dropping of non-ET-marked packets – results for MFT-MFR, MFT-IF, and MFT-IEA.

(c) Preferential dropping of ET-marked packets – results for MFT-MFR.

(d) Preferential dropping of ET-marked packets – results for MFT-IF and MFT-IEA.

Figure 9: Impact of packet drop policies, buffer sizes, and MFT methods on the termination behavior.

and preferential drop of ET-marked packets. The first option is relevant because it is mostly default, the second option is beneficial to MFT, and the third option is required by some other PCN proposals [21, 23]. Figs. 9(a)–9(d) illustrate the results of the experiments.

Figs. 9(a) and 9(b) show the termination behavior for MFT-IEA without preferential packet dropping and with preferential dropping of non-ET-marked packets. Without preferential packet dropping, the termination process is visibly slower than with preferential dropping of non-ET-marked packets because lost ET-marked packets are missing triggers for flow termination. However, the SR-overload is removed after 2 s. The figures also show that overtermination occurs in spite of $\alpha = 1$ and increases with the buffer size. A large buffer stores ET-marked packets that take effect when the buffer empties and the SR-overload is already removed. With preferential dropping of non-ET-marked packets the termination process is faster with small buffers than with large buffers because short buffers lead to more dropped non-ET-marked packets and to a faster delivery of ET-marked packets which expedites the termination process. This is different for other the packet dropping policies. The same simulation results are obtained for MFT-MFR, MFT-IF, and MFT-IEA.

Preferential dropping of ET-marked packets leads to different results for MFT-MFR compared to MFT-IF and MFT-IEA. Fig. 9(c) shows them for MFT-MFR. MFT-MFR uses marking frequency reduction and, hence, only a small fraction of packets is ET-marked. If they are lost, no flows are terminated. If the buffer is large, packet loss is delayed and within that time ET-marked packets still arrive and terminate flows. Therefore, the termination process stops without being completed for small buffers earlier than for large buffers.

Fig. 9(d) shows that preferential dropping of ET-marked packets also slows down the termination process for MFT-IF and MFT-IEA, but it does not stop it before completion. As long as the supportable rate SR is lower than the bottleneck bandwidth, at least some ET-marked packets arrive in case of SR-overload and guarantee that the termination process continues. Although 87.5% of all ET-marked packets are lost initially, the SR-overload is removed after 3.5 s.

5.4.5 Variability of the Termination Process

As MFT depends on stochastic packet marks, the termination behavior is variable, i.e., sometimes the termination process is faster, sometimes slower. We explore that issue by using highly variable packet sizes according to Table 1 ($t = 1$) to provoke well visible variations and set $\alpha = 1$. In our simulation we performed multiple runs of the same experiment with different seeds. Fig. 10 shows the mean values of the PCN rate $r(l, t)$ and the 5%- and 95%-quantiles to characterize its variability. Some variability is due to the stochastic variability of the traffic. This is well visible before termination starts. The distance between the 5%- and 95%-quantiles is rather small and, hence, the termination behavior is rather predictable. The termination behavior for MFT-IF is more variable than for MFT-MFR and MFT-IAE.

5.4.6 Termination Fairness among Aggregates

In a provider network, a link carries usually the traffic of different customers. With MFT-IEA, the traffic of each customer is likely to be explicitly grouped by a single IEA while there is no

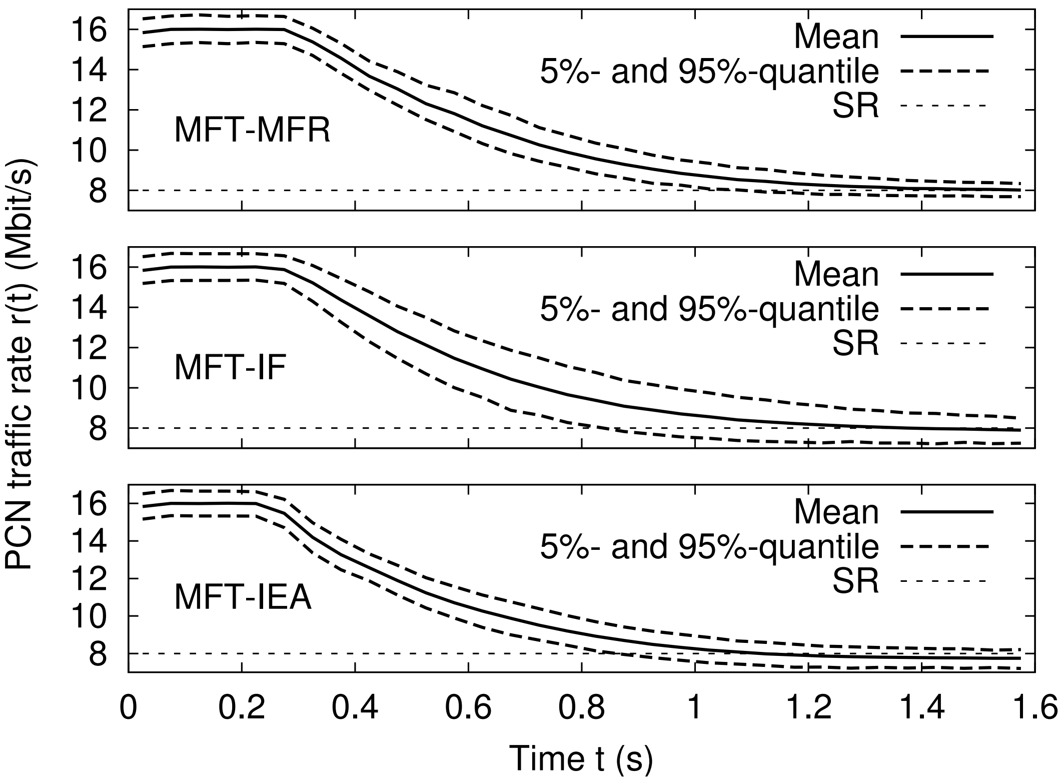


Figure 10: The fluctuation of the termination behavior for all three MFT methods is similar.

explicit grouping with MFT-MFR or MFT-IF. When 50% of the traffic needs to be terminated, it is desirable to have 50% reduction for each customer aggregate. For our next experiment, we use 200 flows with 40 kbit/s and 50 flows with 160 kbit/s that have the same $E[A] = 20$ ms and group them proportionally into $m = 10$ aggregates. We expect that 50% of the traffic is removed per IEA. Fig. 11 shows the CCDF for the fraction of terminated traffic per IEA. The curve illustrates that less than 40% or more than 60% of the traffic of an aggregate is terminated with a certain probability. We derive the same curve for MFT-MFR and MFT-IF based on virtual aggregates since these mechanisms do not require explicit aggregates. The probability to terminate less than 40% or more than 60% of the traffic is significantly larger than with MFT-IEA. Thus, MFT-IEA terminates the traffic of different aggregates in a fairer way than MFT-MFR or MFT-IF. For a smaller number of aggregates m and more flows per aggregate $\frac{n}{m}$, the CCDF is steeper around 50% termination while for a larger number of aggregates and fewer flows per aggregate $\frac{n}{m}$, the CCDF is more flat. For homogenous traffic all curves are rather steep.

5.4.7 Impact of Traffic Characteristics

We studied the impact of strongly varying packet sizes and inter-arrival times, but they had a rather negligible impact on the termination behavior. The same holds for on/off traffic with

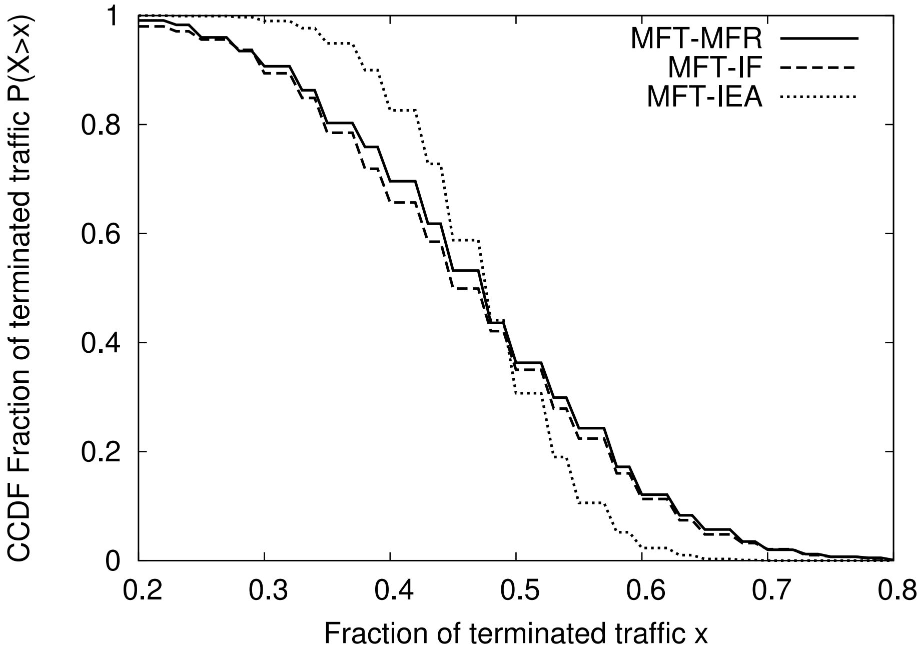


Figure 11: CCDF of the fraction of terminated traffic per (virtual) IEA for MFT-MFR, MFT-IF, and MFT-IEA.

exponentially distributed on/off phase durations and for different average values of these durations.

5.5 Comparison of MFT Methods

We highlight the key benefits of MFT and discuss the pros and cons of MFT-MFR, MFT-IF, and MFT-IEA under challenging conditions.

5.5.1 Key Benefits of MFT

MFT consecutively terminates only ET-marked flows. This has several advantages compared to measured rate termination (MRT) as suggested in the CL and SM proposal (cf. Sect. 3.2.1 and 3.2.2). (1) MFT-MFR and MFT-IF do not require IEAs and are, therefore, applicable for end-to-end PCN. If IEAs are available, MFT-IEA can take advantage of them. (2) MFT works well with multipath routing. (3) MFT does not require that egress nodes or endpoints take rate measurement of marked and unmarked traffic. This is an error-prone process due to stochastic variations in case of low aggregation and short measurement intervals. (4) MFT decreases the SR-overload only gradually. It is self-adaptive in the sense that wrong rate estimates of terminated flows are compensated by more or less frequent termination of further flows. With MRT, under- or overestimated flow rates lead either to overtermination or to significant delay of the termination process as a minimum inter-termination time must be respected.

5.5.2 Unknown Traffic Characteristics

MFT-MFR requires estimates for the average packet inter-arrival time within flows $E[A]$ and the average flow termination delay $E[D_T]$ for the configuration of the marking algorithm of PCN nodes (cf. Eqn. (3)). MFT-IF and MFT-IEA need only an estimate for $E[D_T]$ when we assume that the rates of the flows are known by the PCN egress nodes or endpoints (cf. Eqns. (6) and (7)). Therefore, the termination behavior is harder to control for MFT-MFR than for MFT-IF and MFT-IEA.

5.5.3 Implementation and Configuration Complexity

MFT-MFR and MFT-IF are simple to implement in the sense that they do not need IEAs. This is an advantage since IEAs need extra data structures and it is sometimes difficult to associate flows with correct IEAs because it is not trivial to derive the PCN ingress and egress node for a flow. The termination function of MFT-MFR is simple while MFT-IF and MFT-IEA need initialization and maintenance of credit counters per flow or aggregate.

With MFT-MFR, the stretch factor β_α in the marking frequency reduction part in Algorithm 1 requires an estimate of the mean packet inter-arrival time $E[A]$ within flows and the mean flow termination delay $E[D_T]$ (cf. Eqn. (3)). The parameters may be different in different nodes. In contrast, MFT-IF and MFT-IEA require only an estimate for $E[D_T]$ in egress nodes or endpoints for the initialization of credit counters and the calculation of the increments. However, they globally require the same values for the sake of fair termination probabilities. This is especially difficult for MFT-IF if many distributed endpoints are under the control of a user instead of an operator. It may be more feasible for MFT-IEA as PCN egress nodes are under the control of operators. The setting of the aggressiveness α raises similar security issues.

5.5.4 Fairness Issues, Termination Priorities, and Policies

With MFT-MFR and MFT-IEA, flows with a higher packet rate than others have a higher termination probability even if they have the same bit rate. In contrast, MFT-IF leads to fair termination (cf. Sect. 5.1.8, 5.2.3, and 5.3.4). MFT-IEA may use sub-IEAs for flows with small and large packet rates to overcome this problem. In addition, MFT-IEA leads to more equal termination probabilities among IEAs than MFT-MFR and MFT-IF among virtual IEAs (cf. Sect. 5.4.6). Simple stochastic termination priorities can be implemented with both MFT-IF and MFT-IEA by modifying the aggressiveness α for a set of flows. MFT-IEA supports stochastic enforcement of general termination policies.

5.5.5 Controllability of End-to-End PCN Flows by PCN Egress Nodes

It is desirable that edge nodes of an edge-to-edge PCN domain can control whether end-to-end PCN flows behave correctly. When MFT-MFR is used, a single marked packet indicates the termination of a flow. When a PCN egress node recognizes an ET-marked packet for an end-to-end PCN flow, it can signal the PCN ingress node to set up a filter and block further packets of that flow. This is not possible with MFT-IF as a single ET-marked packets does not necessarily mean that a flow will be terminated.

5.5.6 Compatibility with Existing Hardware

Current hardware offers simple excess marking, but not marking frequency reduction (MFR), proportional MFR (PMFR) or packet size independent marking (PSIM) as required by PCN nodes to support MFT-MFR. Thus MFT-MFR needs new metering and marking features in routers. MFT-IF and MFT-IEA require excess marking with PSIM, but PSIM is only required to equalize termination probabilities for flows with different packet sizes. Therefore, the roll-out of MFT-IF and MFT-IEA could start without waiting for new router features to be deployed and PSIM may be added as an improvement by future updates.

6 Conclusion

Pre-congestion notification (PCN) allows simple implementation of admission control (AC) and flow termination (FT) for single DiffServ domains (edge-to-edge PCN). As an alternative to edge-to-edge PCN, we suggested to move the PCN control entities to the end systems to have an end-to-end AC and FT solution similar to ECN-based congestion control [4]. In spite of trust issues for the general Internet, end-to-end PCN can be useful for corporate networks and it can coexist with edge-to-edge PCN.

The major contribution of this work is the definition and investigation of marked flow termination (MFT). We proposed *MFT based on excess marking with marking frequency reduction* which can be applied both to individual flows and ingress-egress aggregates (IEAs) and *MFT based on plain excess marking for individual flows (MFT-IF) and for IEAs (MFT-IEA)*. The major benefits of MFT compared to existing measured rate termination (MRT) methods are that they applicable both in an edge-to-edge and end-to-end PCN context and in networks using multipath routing. We analyzed the termination behavior for all MFT methods and provided recommendations for their configuration. All MFT methods terminate overload traffic rather quickly and can be configured that they yield the same termination behavior under most considered conditions. This is important for a fair coexistence of edge-to-edge and end-to-end PCN. We identified challenges for the MFT under non-trivial conditions and compared benefits and shortcomings of the different MFT methods.

Acknowledgements

The authors would like to thank Joe Babiarz, Bob Briscoe, Kwok Chan, Anna Charny, Phil Eardley for fruitful discussions.

References

[1] P. Eardley (ed.), “Pre-Congestion Notification Architecture.” http://tools.ietf.org/id/draft-ietf-pcn-architecture-10.txt, Mar. 2009.

[2] S. Floyd and V. Jacobson, “Random Early Detection Gateways for Congestion Avoidance,” *IEEE/ACM Transactions on Networking*, vol. 1, pp. 397–413, Aug. 1993.

[3] B. Braden et al., “RFC2309: Recommendations on Queue Management and Congestion Avoidance in the Internet,” Apr. 1998.

[4] K. Ramakrishnan, S. Floyd, and D. Black, “RFC3168: The Addition of Explicit Congestion Notification (ECN) to IP,” Sept. 2001.

[5] N. Spring, D. Wetherall, and D. Ely, “RFC3540: Robust Explicit Congestion Notification (ECN),” June 2003.

[6] K. Nichols, S. Blake, F. Baker, and D. L. Black, “RFC2474: Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers,” Dec. 1998.

[7] S. Floyd, “RFC4774: Specifying Alternate Semantics for the Explicit Congestion Notification (ECN) Field,” Feb. 2007.

[8] W. Almesberger, T. Ferrari, and J.-Y. Le Boudec, “SRP: A Scalable Resource Reservation for the Internet,” *Computer Communications*, vol. 21, pp. 1200–1211, Nov. 1998.

[9] I. Stoica and H. Zhang, “Providing Guaranteed Services without per Flow Management,” in *ACM SIGCOMM*, (Boston, MA), Sept. 1999.

[10] R. Szábó, T. Henk, V. Rexhepi, and G. Karagiannis, “Resource Management in Differentiated Services (RMD) IP Networks,” in *International Conference on Emerging Telecommunications Technologies and Applications (ICETA 2001)*, (Kosice, Slovak Republic), Oct. 2001.

[11] R. J. Gibbens and F. P. Kelly, “Resource Pricing and the Evolution of Congestion Control,” *Automatica*, vol. 35, no. 12, pp. 1969–1985, 1999.

[12] R. J. Gibbens and F. P. Kelly, “Distributed Connection Acceptance Control for a Connectionless Network,” in 16^{th} *International Teletraffic Congress (ITC)*, (Edinburgh, UK), pp. 941 – 952, June 1999.

[13] F. Kelly, P. Key, and S. Zachary, “Distributed Admission Control,” *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 12, pp. 2617–2628, 2000.

[14] M. Karsten and J. Schmitt, “Admission Control based on Packet Marking and Feedback Signalling – Mechanisms, Implementation and Experiments,” Technical Report 03/2002, Darmstadt University of Technology, 2002.

[15] M. Karsten and J. Schmitt, “Packet Marking for Integrated Load Control,” in *IFIP/IEEE Symposium on Integrated Management (IM)*, 2005.

[16] D. J. Songhurst, P. Eardley, B. Briscoe, C. di Cairano Gilfedder, and J. Tay, “Guaranteed QoS Synthesis for Admission Control with Shared Capacity,” technical report TR-CXR9-2006-001, BT, Feb. 2006.

[17] M. Menth, R. Martin, and J. Charzinski, “Capacity Overprovisioning for Networks with Resilience Requirements,” in *ACM SIGCOMM*, (Pisa, Italy), Sept. 2006.

[18] M. Menth, S. Kopf, J. Charzinski, and K. Schrodi, “Resilient Network Admission Control,” *Computer Networks*, vol. 52, pp. 2805–2815, Oct. 2008.

[19] S. Iyer, S. Bhattacharyya, N. Taft, and C. Diot, “An Approach to Alleviate Link Overload as Observed on an IP Backbone,” in *IEEE Infocom*, (San Francisco, CA), April 2003.

[20] Y. Bernet, P. Ford, R. Yavatkar, F. Baker, L. Zhang, M. Speer, R. Braden, B. Davie, J. Wroclawski, and E. Felstaine, “RFC2998: A Framework for Integrated Services Operation over Diffserv Networks,” Nov. 2000.

[21] B. Briscoe et al., “An Edge-to-Edge Deployment Model for Pre-Congestion Notification: Admission Control over a DiffServ Region.” http://tools.ietf.org/id/draft-briscoe-tsvwg-cl-architecture-04.txt, Oct. 2006.

[22] J. Wroclawski, “RFC2211: Specification of the Controlled-Load Network Element Service,” Sept. 1997.

[23] A. Charny, F. L. Faucheur, V. Liatsos, and J. Zhang, “Pre-Congestion Notification Using Single Marking for Admission and Pre-emption.” http://tools.ietf.org/id/draft-charny-pcn-single-marking-03.txt, Nov. 2007.

[24] M. Menth and F. Lehrieder, “Performance Evaluation of PCN-Based Admission Control,” in *International Workshop on Quality of Service (IWQoS)*, (Enschede, The Netherlands), June 2008.

[25] J. Babiarz, X.-G. Liu, K. Chan, and M. Menth, “Three State PCN Marking.” http://tools.ietf.org/html/draft-babiarz-pcn-3sm, Nov. 2007.

[26] M. Menth, F. Lehrieder, P. Eardley, A. Charny, and J. Babiarz, “Edge-Assisted Marked Flow Termination.” http://tools.ietf.org/html/draft-menth-pcn-emft, Feb. 2008.