PCN-Based Marked Flow Termination[☆]

Michael Menth[a,*], Frank Lehrieder[b]

[a] *University of Tübingen, Department of Computer Science, Germany*
[b] *University of Würzburg, Departement of Computer Science, Germany*

Abstract

Pre-congestion notification (PCN) uses packet metering and marking to notify boundary nodes of a Differentiated Services IP network if configured rate thresholds have been exceeded on some links. This feedback is used for PCN-based admission control and flow termination. While admission control is rather well understood, flow termination is a new flow control function and useful especially in case of failures or during flash crowds. We present marked flow termination as a new class of termination algorithms which terminate overload traffic gradually and that work well with multipath routing. We study their termination behavior, give recommendation for their configuration, and discuss their benefits and shortcomings.

Keywords: Flow termination, resilience, QoS, Differentiated Services, adaptive systems, 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 the Resource reSerVation Protocol (RSVP) [1] in terms of operation and state management. Therefore, the IETF currently standardizes admission control (AC) and flow termination (FT) for Differentiated Services IP networks based on pre-congestion notification (PCN). PCN means that routers in a so-called PCN domain meter the traffic on their links and re-mark packets if the traffic exceeds link-specific rate thresholds. Thereby, boundary nodes of the PCN domain are notified about high load conditions before congestion occurs. In contrast to RSVP, PCN scales well because the metering and marking algorithms work on aggregates and do not need to know individual flows.

The AC function admits or rejects new flows based on measured PCN feedback from the network [2] to limit the traffic load and to enforce quality of service for already admitted flows. The AC function may fail under difficult conditions, e.g. during flash crowds when the rate of admission requests rises suddenly. Moreover, overload can also appear on backup paths after traffic rerouting in case of failures. In such situations, the FT function tears down some already admitted flows and restores controlled-load service conditions [3].

While AC methods have been studied intensively in the past, PCN's FT feature is a new flow control function and only little understood. The current proposals for PCN control [4, 5] use measured rate termination (MRT) which we have investigated in [6]. MRT measures the rate of differently marked traffic per ingress-egress aggregate (IEA) over an interval and estimates the traffic rate to be terminated. Then, a suitable subset of flows from that IEA are terminated in one shot. If too little traffic was terminated, some more flows may be torn down after a safety period and another measurement period. In case of multipath routing, MRT is more complex as egress nodes need to record flows with recently re-marked packets and signal them to corresponding ingress nodes.

In this paper we propose three different algorithms for marked flow termination (MFT) and evaluate their performance. We describe their operation, analyze their termination behavior, give recommendations for their configuration, and summarize their pros and cons. In contrast to MRT, they do not terminate flows in one shot but gradually one after another. Furthermore, termination is only triggered for flows with marked packets. This facilitates the use of MFT in networks with multipath routing without additional modifications which is a strong advantage over MRT.

Section 2 reviews the current PCN dual marking architecture and gives pointers to related work. In Section 3 we propose three new MFT methods including assumptions about packet re-marking. Section 4 investigates and compares the three methods in detail. Section 5 summarizes this work and gives conclusions. A list of frequently used acronyms is provided in the appendix.

2. Overview of Pre-Congestion Notification

We give an overview of PCN, review the "Controlled Load" (CL) PCN architecture [4], and explain its flow termination in detail. To keep this paper short, we refer the interested reader for more information about PCN and related work to [7].

[☆] This work is 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.

[*] Corresponding author.

Email address: menth@informatik.uni-tuebingen.de (Michael Menth)

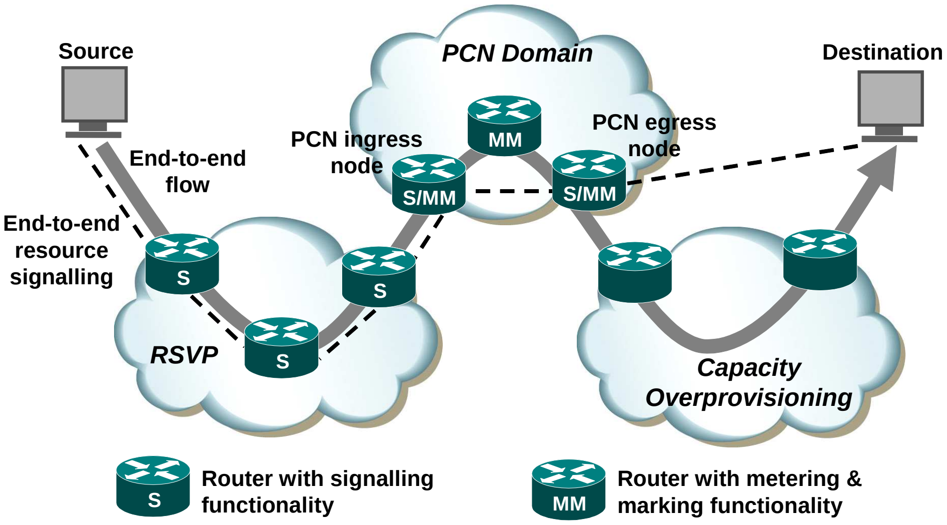

Source
PCN Domain
Destination
End-to-end flow
PCN ingress node
PCN egress node
MM
S/MM
S/MM
End-to-end resource signalling
S
S
S
RSVP
Capacity Overprovisioning
S
Router with signalling functionality
MM
Router with metering & marking functionality

Figure 1: PCN-based AC is triggered by admission requests from external signalling protocols and guarantees QoS within a single PCN domain.

2.1. *Pre-Congestion Notification (PCN)*

PCN is intended for use in a single Differentiated Services IP network, a so-called PCN domain. It defines a new traffic class that receives preferred treatment by PCN nodes and provides information to support admission control (AC) and flow termination (FT) for this traffic type. Some end-to-end signalling protocol (e.g. SIP or RSVP) requests admission for a new flow to cross the PCN domain similarly to the IntServ-over-DiffServ concept [8]. This is illustrated in Figure 1. 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 when certain configured rate thresholds are exceeded. PCN egress nodes evaluate the markings of the traffic and send a digest to the AC and FT decision points so that they can admit or block new flows or even terminate already admitted flows. The AC and FT decision points are typically collocated with the ingress nodes of a PCN domain like in the Integrated Services model or reside in a centralized node within a domain like in the IP Multimedia Subsystem (IMS). The AC and FT decisions of a PCN domain are enforced by appropriate filters and per-flow policers at the ingress nodes. 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.

PCN introduces an admissible rate ($AR(l)$) and a supportable rate ($SR(l)$) threshold for each link l of a PCN domain. These two thresholds imply three different link states as illustrated in Figure 2. 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 SR-pre-congested and the traffic rate above $SR(l)$ is SR-overload. In this state, some already admitted flows should be terminated.

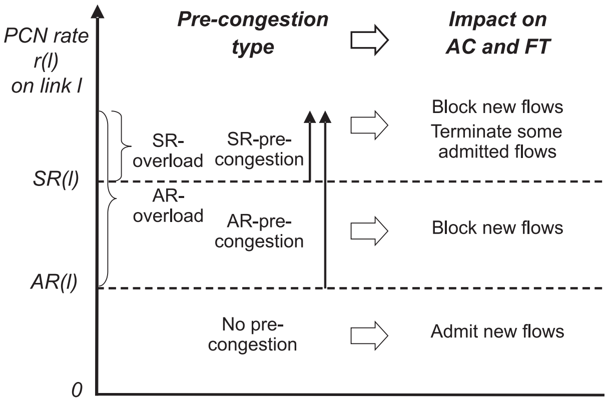


Figure 2: The admissible and the supportable rate ($AR(l), SR(l)$) define three pre-congestion states concerning the PCN traffic rate $r(l)$ on a link.

There are two metering and marking techniques for PCN nodes: threshold marking and excess traffic marking [9]. A (meter and) marker is configured with a reference rate. Ingress nodes label PCN traffic as “not-marked” (NM) to make it distinguishable from other low priority traffic. Meters measure the rate of PCN traffic and re-mark it when the metered PCN traffic rate exceeds their configured reference rate. With threshold marking, all PCN packets are re-marked as “threshold-marked” (ThM) under this condition while excess-traffic-marking re-marks only those PCN packets to “excess-traffic-marked” (ETM) that exceed the configured rate. ETM is stronger than ThM so that ThM packets may be re-marked to ETM but not vice-versa [10].

2.2. *The “Controlled Load” (CL) PCN Architecture*

We review the CL PCN architecture as defined in [4]. Threshold markers for each link l in a PCN domain are configured with the link-specific admissible rate $AR(l)$ and excess traffic markers are configured with the supportable rate $SR(l)$. In case of AR-pre-congestion, all PCN packets are re-marked to ThM. In case of SR-pre-congestion, some PCN packets are even re-marked to ETM and all others are re-marked to ThM. An ingress-egress aggregate (IEA) consists of all flows entering a PCN domain at a specific ingress and leaving it at a specific egress. Egress nodes measure the rates of differently marked PCN traffic per ingress-egress aggregate using interval-based measurement and send these rates to the AC and FT decision points, i.e., usually to the ingress node of the IEA.

For AC purposes, a decision point calculates the congestion level estimate (CLE) which is the fraction of re-marked PCN traffic. If the CLE exceeds a configured CLE limit, further flow requests for the corresponding IEA are blocked to avoid overload. If the CLE falls below the CLE limit, new flows can again be admitted.

FT works as follows. When a decision point receives a rate of ETM traffic larger than zero, it requests the rate of sent PCN traffic from the ingress node (ingress rate IR). It calculates the termination rate as the difference between the ingress rate and the non-ETM traffic rate (sum of NM traffic rate and ThM traffic rate). Then, it chooses a set of flows whose overall rate equals the termination rate and terminates these flows.

2.3. Measured Rate Termination (MRT)

We call the termination approach described above "measured rate termination" (MRT) because the amount of traffic to be terminated is determined by rate measurement. We have investigated this method in [6] and identified the following problems.

To get sufficiently accurate measurement results, the measurement interval needs to be long enough which introduces delay in the order of several hundreds milliseconds. The FT decision point 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. However, a minimum inter-termination time between two consecutive termination steps must be respected to make sure that terminated flows do not contribute anymore to the measured feedback. This further delays the termination process. Moreover, IEA-based traffic measurements are sometimes considered heavyweight and undesirable.

When an IEA carries only a small number of flows and only some ETM-packets are received by the egress node, it is hard to decide whether none or one flow should be terminated, but the result matters. We have suggested proportional termination to solve that problem. 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 observed SR-pre-congestion. This can be repaired if the egress node provides information about flows with recently marked packets to the FT decision point.

MRT requires the notion of an IEA to perform per-IEA rate measurement, but it is not yet clear how flows belonging to a specific IEA are recognized. End-to-end PCN has been introduced in [11, 7]. It allows only definition of trivial IEAs (single flows), so that MRT does not seem appropriate in that context.

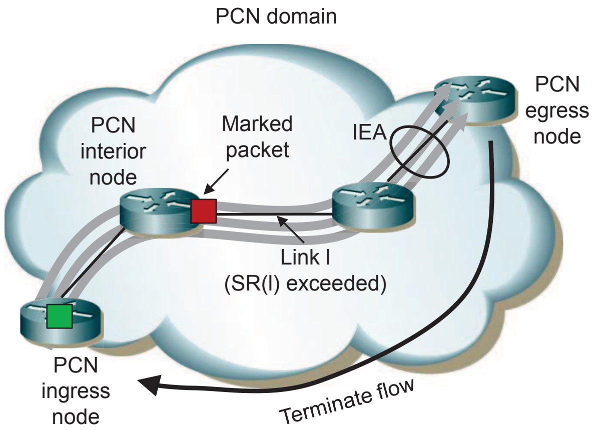


Figure 3: Basic functionality of FT mechanisms. In case of SR-pre-congestion, PCN interior nodes re-mark packets as excess-traffic-marked. PCN egress nodes evaluate these markings and possibly trigger the termination of flows.

3. Marked Flow Termination (MFT)

In this section we present marked flow termination (MFT) as an alternative to measured rate termination (MRT). Due to the shortcomings of MRT, it is of interest to explore other PCN-based flow termination methods that avoid the shortcomings of MRT by design which is the case for marked flow termination (MFT).

When SR-pre-congestion occurs, packets are excess-traffic-marked. In the following, we just say "marked" for the sake of brevity. MFT methods terminate only "marked flows", i.e. those with at least one recently marked packet. In the following, we propose three different methods for MFT including modifications of the marking algorithms if needed. The basic functionality of FT mechanisms is illustrated in Fig. 3.

3.1. Marked Flow Termination Based on Excess Traffic Marking with Marking Frequency Reduction (MFT-MFR)

The idea of this approach is that the egress node triggers the termination of a flow as soon as it has received a marked packet for that flow. However, the existing algorithm for excess traffic marking re-marks so many packets to ETM that too many flows are terminated with this idea. Therefore, we reduce the marking frequency of excess traffic marking by a factor to control the termination aggressiveness of the MFT algorithm that we call MFT with marking frequency reduction (MFT-MFR). We have discussed this idea for the first time in [12].

In the following, we present the base algorithm for excess traffic marking, our modification for packet-size independent marking (PSIM), an extension for marking frequency reduction (MFR), as well as a modification for proportional MFR (PMFR).

3.1.1. Plain Excess Traffic Marking

Algorithm 1 uses a token bucket based formulation to describe the behavior of the excess traffic marker. It is called at each packet arrival. The marker has a token bucket which is S bytes large and constantly filled with tokens at rate R. The

```
Input:  token bucket parameters $S$, $R$, $F$, $lastU$,
        packet size $B$ and marking $M$, current time
        $now$, maximum transfer unit $MTU$ (only
        needed for PSIM), increment $I$ (for MFR) or
        stretch factor $\beta_\alpha$ (for PMFR)

  $F = \min(S, F + (now - lastU) \cdot R)$;
  $lastU = now$;
  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
```

Algorithm 1: EXCESS TRAFFIC MARKING: base algorithm with extension for marking frequency reduction (MFR) and modification for packet size independent marking (PSIM) as well as proportional MFR.

token bucket variable $lastU$ records the time of the last update and helps to account for the number of new tokens that have been generated 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 $M = ETM$.

3.1.2. Excess Traffic Marking with Packet Size Independent Marking (PSIM)

Plain excess traffic marking marks larger packets with a higher probability than smaller packets. This may lead to higher termination probabilities for flows with larger packets. Therefore, we propose to make the marking decision in Algorithm 1 independent of the packet size. PSIM marks a packet already if the fill state is lower than the maximum transfer unit (MTU) of the link. This assures that the marking probability is independent of the packet size B. This modification is beneficial also to other flow termination methods and has also been adopted by [9] as an explicit implementation option.

3.1.3. Excess Traffic Marking with Marking Frequency Reduction (MFR)

Plain excess traffic marking re-marks all traffic whose rate exceeds the reference rate of the marker. We present an extension that reduces the frequency at which packets are re-marked in a controllable way. It is expressed by the last if-statement in Algorithm 1. If the packet is marked, MFR adds an increment of I bytes to the bucket. This is done regardless of whether the packet was marked by the current or a previous node. This algorithm can be easily combined with PSIM.

3.1.4. Excess Traffic Marking with Proportional MFR (PMFR)

As we will see later, it is beneficial if the increment I used by excess traffic marking with MFR is proportional to the packet size B. Therefore, excess traffic marking with proportional MFR uses a stretch factor β_α which is multiplied by the packet size B to yield the increment. This is also reflected in Algorithm 1.

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

MFT-IF requires plain excess traffic marking with packet size independent marking (PSIM). Therefore, it is compatible with the currently discussed standards proposal [4]. The egress node of a flow f sets up a flow-specific credit counter C_f. If a flow's packet arrives 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 ETM packet, the flow is terminated. In contrast to MFT-MFR, this method permits to implement stochastic termination priorities by choosing larger values for the credit counter initialization for high-priority flows than for low-priority flows.

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

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 a marked packet that belongs to a flow $f \in \mathcal{F}(g)$, the packet's size in bytes is subtracted from the counter C_g. If the counter is not positive at the arrival of a 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. Like with MFT-IF, flow termination priorities can be implemented. However, termination priorities can be enforced more effectively with MFT-IEA than with MFT-MFR or MFT-IF because MFT-IEA can choose the flow to be terminated from the set of recently marked flows.

An alternative design terminates a flow already if the size of the 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. We have discussed this idea for the first time in [13].

4. Performance Evaluation and Comparison of MFT Methods

In this section we first explain our simulation methodology. We investigate the three presented MFT methods one after another and give recommendations for their configuration. We then compare the three methods under various conditions, summarize our results, and briefly comment on other work about MFT.

4.1. Simulation Setup

We investigate the termination behavior of MFT on a link that faces sudden overload as it is the case, e.g., after traffic reroutes or flash-crowd arrivals. We do not simulate complex network topologies, but abstract from an entire network to a single link that is suddenly faced with an overload condition. This simplifies and speeds up the simulation. It still permits conclusions about the termination behavior because we model the message delay which is normally caused by the missing network elements by an appropriate flow termination delay D_T. If not mentioned differently, we use the following default parameters for our experiments. We use simple constant bit rate flows because they are more appropriate to find and visualize basic effects of the mechanisms under study. The flows have deterministic inter-arrival times A with $E[A] = 20$ ms[1] and deterministic packet sizes B with $E[B] = 200$ bytes. Thus, flow rates are $E[R] = 80$ 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.

When the egress node decides to terminate a flow, it quickly informs the ingress node to reconfigure appropriate filters. We introduce the concept of the flow termination delay D_T. It is the time between the decision of the egress node until it no longer receives packets from the terminated flow. The round-trip time within a PCN domain gives a lower bound on that value but D_T may be larger due to management overhead at ingress and egress node. 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.

We simulate the time-dependent PCN traffic rate $r(t)$ of a link to study the termination process of the time-dependent SR-overload $SRO(t)$. The supportable link rate is $SR = 8$ Mbit/s and the simulation starts with $n = 200$ admitted flows which is $r(0) = 16$ Mbit/s. This corresponds to an initial SR-overload of 100%, i.e., the initial SR-overload is also $SRO(0) = 8$ Mbit/s. Thus, half of the flows need to be terminated. The token bucket 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.

We use a custom-made Java tool to simulate the time-dependent PCN rate $r(t)$ to illustrate the termination behavior. 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.[2]

4.2. MFT with Marking Frequency Reduction (MFR)

We investigate the performance of MFT-MFR that we presented in Section 3.1. We first motivate an appropriate value for the increment that is used by MFR to control the marking frequency to avoid overtermination. Then, we investigate the impact of traffic characteristics on the termination behavior and argue that MFT-MFR requires PSIM and PMFR to work well. Nevertheless, the method cannot cope well with flows that have different inter-arrival times.

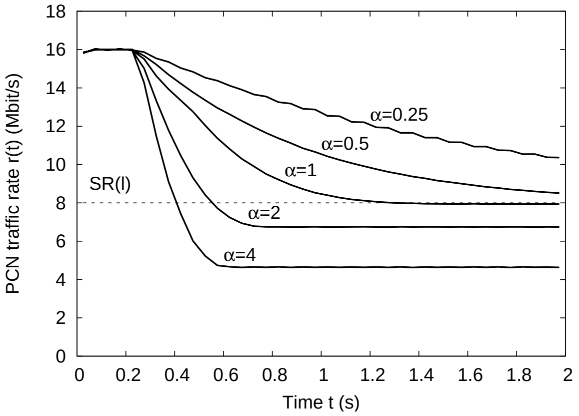


Figure 4: The aggressiveness α controls the speed of the termination process and the degree of potential overtermination.

4.2.1. Configuration of the Increment

When the increment is set to $I = 0$, Algorithm 1 performs plain excess traffic marking, i.e., all packets exceeding the SR of the link are marked. Let $E[B]$ be the average packet size. 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 traffic marking. As a result, the marking frequency is reduced by a factor of

$$\sigma_p = \frac{I}{E[B]} + 1 \qquad (1)$$

in case of SR-pre-congestion.

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 (see Equation (1)) which limits the number of terminated flows. To avoid overtermination, the rate of terminated flows 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 of the next section show that this inequality is sharp.

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

[2] 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.

4.2.2. *Termination Aggressiveness α*

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

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

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 in Figure 4. The degree of overtermination also increases with α. To keep MFT-MFR simple, the increment I_α may be 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 current traffic characteristics. In the following, we study such systems under different conditions.

4.2.3. *Impact of Different Packet Sizes – Homogeneous Traffic*

We configure I_α according to Equation (2) for the default values in Section 4.1, in particular $E[B^*] = 200$ bytes and $\alpha^* = 1$, but vary the actual packet sizes $E[B]$ which affects the actual flow rate $E[R]$. 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 Figure 4 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 in the PCN nodes proportional marking frequency reduction (PMFR) as described in Algorithm 1. The increment is then calculated by

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

using the stretch factor β_α. Thus, the increment is proportional to the size of the observed marked packet. Now, one packet is marked for

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

bytes that have been above *SR* during a continuous SR-precongestion phase. This means that also one flow is terminated for that amount of bytes that have exceeded *SR*. Therefore, PMFR makes the termination behavior of MFT-MFR independent of the packet size $E[B]$. We validated this finding by simulation but we do not show any figures. In the remainder of this work, we use PMFR for the study of MFT-MFR.

4.2.4. *Impact of Different Packet Sizes – Heterogeneous Traffic*

We consider constant bit rate flows with an average bit rate of $E[R] = 80$ kbit/s, but the bit rate of different flows varies. The flows have all the same inter-arrival time of $A = 20$ ms, but differ in packet size according to Table 1. The parameter t determines the proportion of low, medium, and high bit rate flows in the traffic mix. The parameter t controls the variability of the flow-specific packet sizes, so that the corresponding coefficient of variation is $c_{var}[R] = 1.5 \cdot \sqrt{t}$.

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. Either packet size or inter-arrival time is varied, but not both.

Flow type specific	Flow types		
	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

We conducted experiments and found that the termination behavior for highly variable traffic mixes ($t = 1$) is almost the same as for traffic with homogenous packet sizes ($t = 0$, see Figure 4). However, Table 2 shows that flows with large packets have a tremendously higher termination probability than flows with small packets. Therefore, we use from now on packet size independent marking (PSIM, see Section 3.1.2). With this change, low, medium, and high bit rate flows face the same termination probability and the termination behavior is still independent of the traffic mix.

Table 2: Flow termination probabilities depending on packet size B and inter-arrival time A for the experiments in Section 4.2.4 and Section 4.2.6.

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, $\alpha = 0.5$, see Figure 5		
	$E[A] = 80$ 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

4.2.5. *Impact of Packet Inter-Arrival Times – Homogeneous Traffic*

We configure the stretch factor β_α of Equation (3) again based on the default values given in Section 4.1, in particular $E[A^*]$ and $\alpha^* = 1$, but vary the actual packet inter-arrival time A for all flows which affects the average flow rate $E[R]$. This 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. Hence, the termination behavior of MFT-MFR significantly depends on the actual packet inter-arrival times $E[A]$ and looks like the curves in Figure 4 for different α.

In practice we need a viable solution that reduces the SR-overload quickly while avoiding overtermination. Most real-time applications send one packet within 20 ms, some others have a period of 10 ms. Video applications are slower but possibly send several packets for one frame. We recommend to use an aggressiveness of $\alpha = 0.5$ and an inter-arrival time of

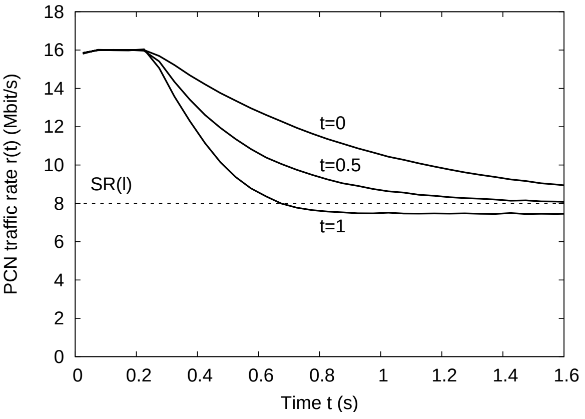


Figure 5: Traffic with more variable inter-arrival times leads to faster termination and flows with shorter inter-arrival times have higher termination probabilities.

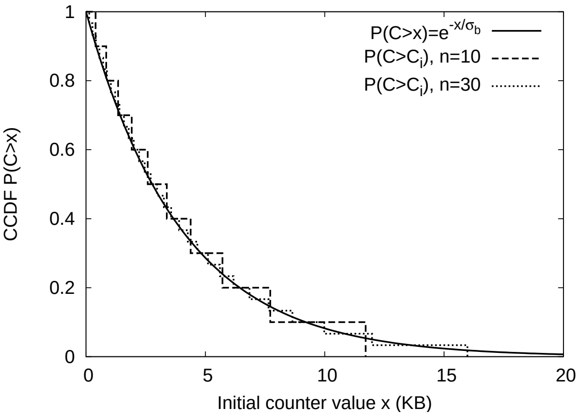


Figure 6: CCDF of the counter initialization values for a various number of n flows and their limiting function.

$E[A] = 20$ ms for the configuration of the stretch factor in Equation (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 (see Figure 4, $\alpha = 0.5$).

4.2.6. *Impact of Packet Inter-Arrival Times – Heterogeneous Traffic*

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 $\alpha = 0.5$. Figure 5 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 a marked packet 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.

4.3. *Performance Evaluation of MFT for Individual Flows (MFT-IF)*

We investigate the performance of MFT-IF that we presented in Section 3.2. We first propose a suitable initialization method for the flow-specific credit counters. 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.

4.3.1. *Counter Initialization*

We suggest a method for the initialization of the credit counters. We borrow ideas from our analysis of MFT-MFR. MFT-MFR's termination speed is controlled by the fact that the next packet is excess-traffic-marked only after σ_b bytes have exceeded SR since the last packet was marked (see Equation (4)). We mimic this behavior by initializing the credit counters for MFT-IF appropriately so that the resulting termination behavior for MFT-IF is the same as for 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 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 $C_i - C_{i-1} = \frac{\sigma_b}{n-(i-1)}$. With $C_0 = 0$, the counter initialization should be chosen

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

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.[3] Experiments with this credit counter initialization show the same termination behavior as in Figure 4.

Equation (5) can be used to initialize the credit counter of flows if all flows sharing a single bottleneck link are known. Now we develop an algorithm which allows a flow to initialize its credit counter randomly without knowing anything about other flows. Based on Equation (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

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

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

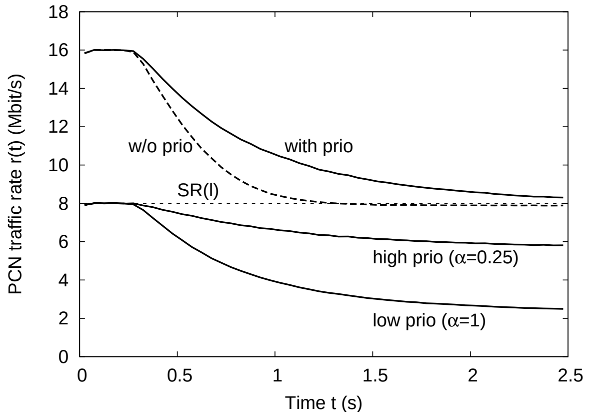


Figure 7: Termination behavior for high and low priority traffic.

for large n. Figure 6 illustrates that the exact CCDFs for various numbers of flows n converge quickly towards the limiting CCDF of Equation (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 Equation (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 Equation (5) by the stochastic initialization according to Equation (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 Section 4.5.5.

4.3.2. Impact of Packet Sizes and Inter-Arrival Times

We conducted experiments and found that the termination behavior for MFT-IF is robust against traffic mixes consisting of flows with different packet sizes and inter-arrival times. The counter initialization takes the issue of different bit rates into account by using the flow rate R_f. We also found that flows with different packet sizes or packet inter-arrival times face the same termination probabilities.

4.3.3. 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.

We consider low-priority flows for which we use $\alpha = 1$ and high-priority flows for which we use $\alpha = 0.25$. Figure 7 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.

4.4. Performance Evaluation of MFT for Ingress-Egress Aggregates (MFT-IEA)

We investigate the performance of MFT-IEA that we presented in Section 3.3. We first propose a suitable initialization method for the IEA-specific credit counters. We study the impact of the size of IEAs, packet size, and inter-arrival time on the termination process and evaluate to what extent termination policies can be enforced.

4.4.1. Configuration of MFT-IEA

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

$$I_\alpha = \sigma_b = \frac{2 \cdot D_T \cdot R_f}{\alpha} \qquad (7)$$

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

4.4.2. Impact of the Size of IEAs

We conduct experiments where the $n = 200$ flows on the bottleneck are grouped into different IEAs with $m \in \{1, 4, 20, 200\}$ flows each. They all lead to about the same termination behavior as in Figure 4. Therefore, we omit the figure for these experiments. The termination process for IEAs of size $m = 1$ just starts 20 ms later than for $m = 200$. In fact, for $m = 1$, MFT-IEA becomes MFT-IF and shows the identical termination behavior.

4.4.3. Impact of Packet Sizes and Inter-Arrival Times

We conducted experiments that show that the termination behavior of MFT-IEA and the flow termination probabilities are insensitive to the average packet size and its variation within flows.

This is slightly different for inter-arrival times. Flows with a higher packet frequency have a higher termination probability since it is more likely that one of their 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 IEAs with $m = 20$ flows. 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 (see Figure 5). We now consider the termination behavior. In contrast to MFT-MFR, with MFT-IEA the termination behavior for heterogeneous traffic hardly differs from the one of homogeneous traffic (without figure). 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 defined in Equation (7) 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, defining sub-IEAs for flows with homogenous inter-arrival times restores equal termination probabilities for all flows.

4.4.4. *Stochastic Enforcement of Termination Policies*

Stochastic termination priorities can be implemented similarly as in Section 4.3.3: 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 a 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 marked, too. Thus, MFT-IEA needs to record the set of 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 flows to be terminated have to be chosen from the set of recently marked flows, and the composition of this set is stochastic.

Table 4: Flow termination probabilities for MFT-IEA and different policies depending on the number of flows per aggregate *m*.

Rate	40 kbit/s	160 kbit/s	40 kbit/s	160 kbit/s	40 kbit/s	160 kbit/s
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) so that half of the traffic volume results from low and high bit rate flows. We group them equally into aggregates with $m \in \{5, 25, 250\}$ flows. 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 for larger aggregates. Thus, the effectiveness of stochastic policy enforcement depends both on the aggregation level of the IEA and the policy itself.

4.5. *Performance Comparison 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 traffic 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 split into IEAs with $m = 20$ flows. 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.

4.5.1. *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 Figure 4 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 Equation (3), to initialize all credit counters for MFT-IF and MFT-IEA in Equation (6), and to calculate the rate-dependent increments for MFT-IEA in Equation (7).

Figure 8 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. 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

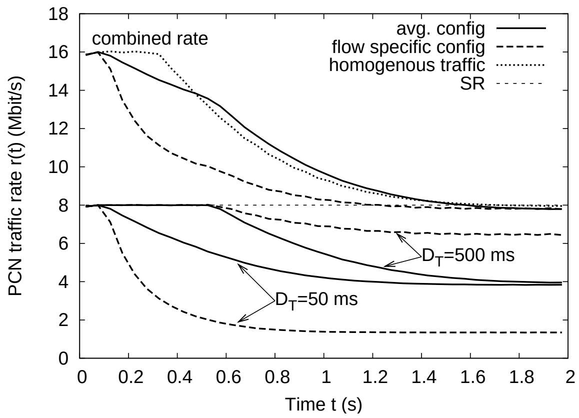


Figure 8: In spite of different flow termination delays D_T all flows have the same termination probability when all system components are configured with an average value $E[D_T]$.

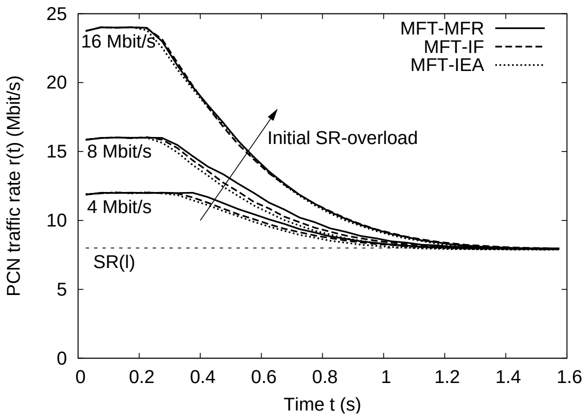


Figure 9: Impact of the initial SR-overload on the termination behavior.

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 $E[D_T]$ should be applied for the configuration of all distributed PCN egress nodes. 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 $E[D_T]$, but this value is used to configure MFR in PCN nodes.

4.5.2. 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. We observe this good scaling behavior because MFT's termination speed is proportional to the marked traffic rate which also scales with the size of the experiment in terms of flows.

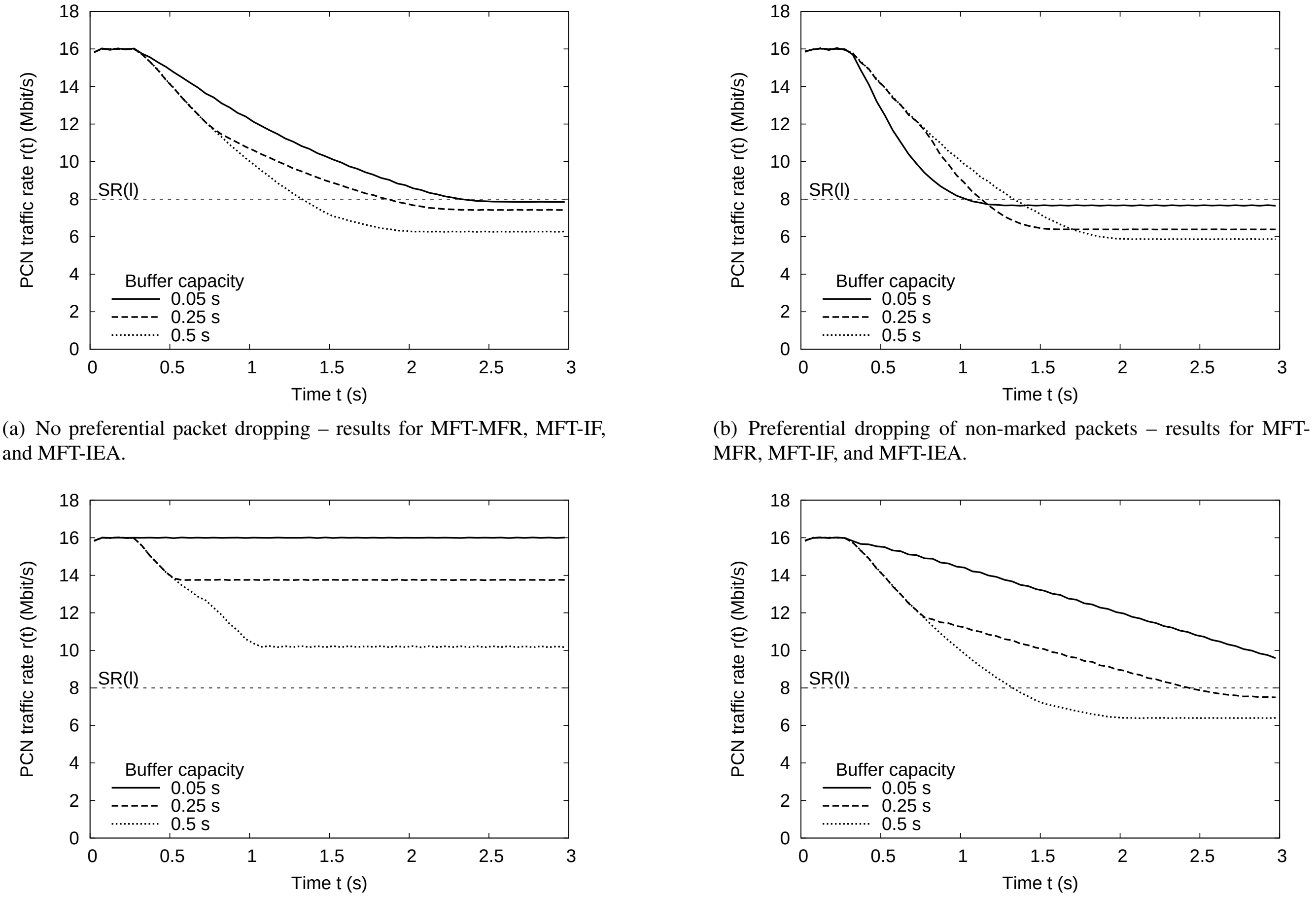
4.5.3. Impact of the SR-Overload Intensity

We set the initial PCN rate to 12, 16, and 24 Mbit/s so that the resulting SR-overload is 4, 8, and 16 Mbit/s which corresponds to an SR-overload of 50%, 100%, and 200%. Figure 9 shows that all three MFT methods yield the same termination behavior. As mentioned in Section 4.2.1, 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 4 Mbit/s overload.

4.5.4. Impact of Packet Loss

We point out again that marked means excess-traffic-marked in this section. MFT requires marked packets to trigger the termination process. In case of packet loss, 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 so 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-marked packets, and preferential drop of 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 other PCN proposals (see [6] for more). Figures 10(a)–10(d) illustrate the results of the experiments.

Figures 10(a) and 10(b) show the termination behavior for MFT-IEA without preferential packet dropping and with preferential dropping of non-marked packets. Without preferential packet dropping, the termination process is visibly slower than with preferential dropping of non-marked packets because lost marked packets are missing triggers for flow termination.



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

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

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

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

Figure 10: Impact of packet drop policies, buffer sizes, and MFT methods on the termination behavior.

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 marked packets that take effect when the buffer empties and the SR-overload is already removed. With preferential dropping of non-marked packets the termination process is faster with small buffers than with large buffers because short buffers lead to more dropped non-marked packets and to a faster delivery of 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 marked packets leads to different results for MFT-MFR compared to MFT-IF and MFT-IEA. Figure 10(c) shows them for MFT-MFR. MFT-MFR uses marking frequency reduction and, hence, only a small fraction of packets is marked. If they are lost, no flows are terminated. If the buffer is large, packet loss is delayed and within that time marked packets still arrive and terminate flows. Therefore, the termination process stops without being completed for small buffers earlier than for large buffers.

Figure 10(d) shows that preferential dropping of 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 marked packets arrive in case of SR-overload and guarantee that the termination process continues. Although 87.5% of all marked packets are initially lost, the SR-overload is removed after 3.5 s.

4.5.5. 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. Figure 11 shows the mean values of the PCN rate $r(t)$ 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-IEA.

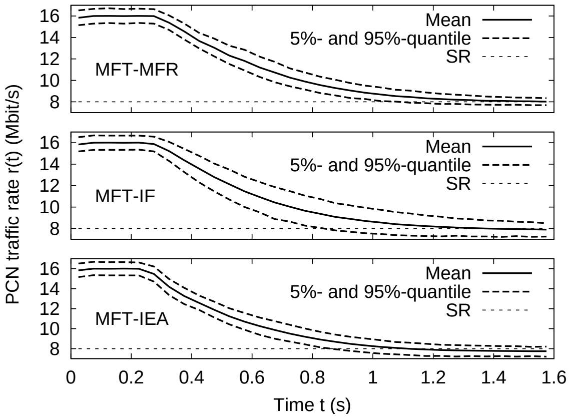


Figure 11: The fluctuation of the termination behavior for all three MFT methods is similar.

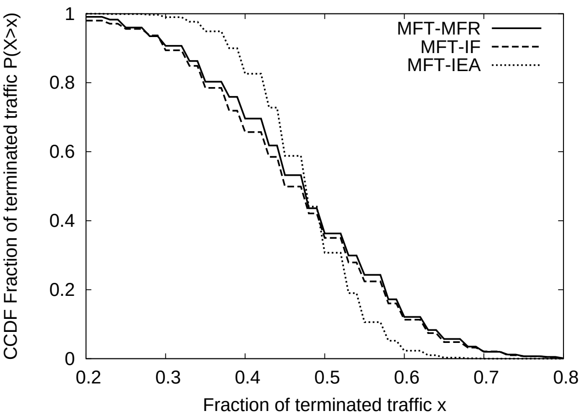


Figure 12: CCDF of the fraction of terminated traffic per (virtual) IEA for MFT-MFR, MFT-IF, and MFT-IEA.

4.5.6. 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 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 IEAs with $m = 25$ flows each. We expect that 50% of the traffic is removed per IEA. Figure 12 shows the CCDF for the fraction of terminated traffic per IEA. 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 larger number of flows per aggregate m, the CCDF is steeper around 50% termination while for a smaller number of flows per aggregate m, the CCDF is more flat (both without figures). With an increasing number of flows per aggregate, the packet rate increases but the marking probability stays about the same. One can show by means of the binomial distribution that the coefficient of variation for the fraction of marked packets per aggregate decreases with increasing packet rate. For homogenous traffic all curves are rather steep (also without figure) because the terminated traffic rate of a virtual aggregate depends only on the number of terminated flows but not on their individual bit rate which is equal for all flows.

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

4.6. Summary

We have investigated three different MFT methods. All of them terminate PCN flows gradually one after another until overload is removed. This differentiates them from measured rate termination (MRT) which terminates a large set of flows in one shot, possibly several times if needed. Moreover, they all terminate only marked flows so that flows are terminated only if they are carried over SR-pre-congested links. Therefore, MFT works well in the presence of multipath routing by design. MRT requires additional signalling to achieve that.

With MFT-MFR a parameter for the increment of the marking mechanism is needed, for MFT-IF a parameter for the initial counter value is needed, and for MFT-IEA a parameter for the initial counter value and for a counter increment. We proposed configurations for all three mechanisms so that the termination speed can be controlled by a termination aggressiveness value α and that they have about the same termination behavior. For $\alpha = 1$ or smaller they avoid overtermination.

The termination behavior of all MFT methods depends on correct estimates for average packet size B, inter-arrival time A, or flow rate R. Moreover, the flow termination delay D_T must be known. Wrong estimates immediately yield a different aggressiveness so that termination takes longer than needed or overtermination occurs. Disadvantages of MFT-MFR are that it does not work with the standardized excess traffic marking [9] and that its termination speed heavily depends on the packet frequency of flows. MFT-IF does not suffer from these problems, it can even support termination policies to some extent. The same holds for MFT-IEA. In addition, it supports stochastic enforcement of termination policies which goes beyond that of MFT-IF, but it may lead to larger termination probabilities for flows with higher packet frequency within a single IEA.

If appropriately configured, the termination behaviors of the three mechanisms are very similar in most considered scenarios. Therefore, the most suitable MFT algorithm for practical use depends on aspects like implementation complexity, robustness to incorrectly estimated parameters, and support of termination priorities. In that respect, out of the three MFT methods, MFT-IF is most interesting from our perspective. It works with plain excess traffic marking, it leads to fair termination probabilities for flows with different packet sizes, and it does not require the notion of IEAs. It may be combined with probe-based

Table 5: List of frequently used acronyms.

Acronym	Meaning
AC	admission control
AR	admissible rate
CCDF	complementary cumulative distribution function
CL	“Controlled Load” PCN architecture [4]
CLE	congestion level estimate
ETM	excess-traffic marked
FT	flow termination
IEA	ingress-egress aggregate
MFR	marking frequency reduction
MFT	marked flow termination
MFT-IEA	MFT for IEAs
MFT-IF	MFT for individual flows
MFT-MFR	MFT with MFR
MRT	measured rate termination
NM	not-marked
PCN	pre-congestion notification
PMFR	proportional MFR
PSIM	packet size independent marking
RSVP	Resource reSerVation Protocol
SR	supportable rate
ThM	threshold-marked

admission control for PCN [14, 7] so that PCN-based AC and FT do not need IEAs at all. We refer to the special probe-based admission control which requires only a single probe packet per flow to be sent from ingress node to egress node. The PCN interior nodes perform threshold marking based on the admissible rate AR so that all PCN packets are re-marked in case of pre-congestion. The new flow is admitted only if the probe packet arrives at the egress node without being re-marked. This combination is of interest as the definition of IEA depends on the networking technology and it is not always clear how to realize them which is currently an unsolved problem for MRT. An additional advantage of all MFT methods over MRT is that they work well with multipath routing because they guarantee that flows are terminated only if they traverse SR-pre-congested links.

4.7. More Results on MFT

We further studied the termination behavior of MFT in the presence of multiple bottlenecks in [15]. In [16] we proposed a method that makes MFT applicable if excess traffic marking is configured with the admissible rate instead of the supportable rate which is useful if only a single marking scheme can be applied for both AC and FT in the PCN domain.

5. Conclusion

Pre-congestion notification (PCN) allows simple implementation of admission control (AC) and flow termination (FT) for Differentiated Services domains. While current FT algorithms require measurement of differently marked PCN traffic per ingress-egress aggregate, we proposed the concept of marked flow termination (MFT), where flows are terminated because one or more of their packets have been marked. Therefore, MFT works well with multipath routing by design, which is not the case with established FT methods. We elaborated three different variants of MFT, gave recommendations for their configuration, and investigated their termination behavior. All MFT methods terminate overload traffic rather quickly within one or two seconds. We found that their termination behaviors are similar under many conditions but there are also situations where they differ. We also pointed out weaknesses and showed that even termination priorities can be enforced to different degrees.

Acknowledgements

The authors thank Joe Babiarz, Bob Briscoe, Kwok Chan, Anna Charny, Phil Eardley, and Alfons Martin for fruitful discussions.

Appendix

See Table 5.

References

[1] B. Braden, L. Zhang, S. Berson, S. Herzog, S. Jamin, RFC2205: Resource ReSerVation Protocol (RSVP) - Version 1 Functional Specification (Sep. 1997).

[2] P. Eardley (Ed.), RFC5559: Pre-Congestion Notification (PCN) Architecture (Jun. 2009).

[3] J. Wroclawski, RFC2211: Specification of the Controlled-Load Network Element Service (Sep. 1997).

[4] A. Charny, F. Huang, G. Karagiannis, M. Menth, T. Taylor, PCN Boundary Node Behavior for the Controlled Load (CL) Mode of Operation, http://tools.ietf.org/html/draft-ietf-pcn-cl-edge-behaviour (Dec. 2010).

[5] A. Charny, J. Zhang, G. Karagiannis, M. Menth, T. Taylor, PCN Boundary Node Behavior for the Single-Marking (SM) Mode of Operation, http://tools.ietf.org/html/draft-ietf-pcn-sm-edge-behaviour (Dec. 2010).

[6] M. Menth, F. Lehrieder, PCN-Based Measured Rate Termination, Computer Networks 54 (13) (2010) 2099 – 2116.

[7] M. Menth, F. Lehrieder, B. Briscoe, P. Eardley, T. Moncaster, J. Babiarz, A. Charny, X. J. Zhang, T. Taylor, K.-H. Chan, D. Satoh, R. Geib, G. Karagiannis, A Survey of PCN-Based Admission Control and Flow Termination, IEEE Communications Surveys & Tutorials 12 (3).

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

[9] P. Eardley (Ed.), RFC5670: Metering and Marking Behaviour of PCN Nodes (Nov. 2009).

[10] B. Briscoe, T. Moncaster, M. Menth, Encoding 3 PCN-States in the IP Header Using a Single DSCP, http://www.ietf.org/internet-drafts/draft-ietf-pcn-3-in-1-encoding-04.txt (Jan. 2011).

[11] M. Menth, F. Lehrieder, PCN-Based Marked Flow Termination, Technical Report, No. 469, University of Würzburg, Institute of Computer Science (Mar. 2010).

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

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

[14] T. Cicic, A. F. Hansen, A. Kvalbein, M. Hartmann, R. Martin, M. Menth, Relaxed Multiple Routing Configurations for IP Fast Reroute, in: IEEE Network Operations and Management Symposium (NOMS), Salvador de Bahia, Brazil, 2008.

[15] F. Lehrieder, M. Menth, PCN-Based Flow Termination with Multiple Bottleneck Links, in: IEEE International Conference on Communications (ICC), Dresden, Germany, 2009.

[16] F. Lehrieder, M. Menth, Marking Conversion for Pre-Congestion Notification, in: IEEE International Conference on Communications (ICC), Dresden, Germany, 2009.