A Survey of PCN-Based Admission Control and Flow Termination

Michael Menth and Frank Lehrieder, University of Würzburg, Inst. of Computer Science, Germany [†]
Bob Briscoe, Philip Eardley, and Toby Moncaster, BT Research, UK [‡]
Jozef Babiarz, Nortel Networks, Ottawa, Canada
Anna Charny and Xinyang (Joy) Zhang, Cisco, Boxborough, MA
Tom Taylor and Kwok-Ho Chan, Huawei Technologies, Canada/USA
Daisuke Satoh, NTT Advanced Technology Corporation, Japan
Georgios Karagiannis, University of Twente, The Netherlands

***Abstract*—Pre-congestion notification (PCN) provides feedback about load conditions in a network to its boundary nodes. The PCN working group of the IETF discusses the use of PCN to implement admission control (AC) and flow termination (FT) for prioritized realtime traffic in a DiffServ domain. Admission control (AC) is a well-known flow control function that blocks admission requests of new flows when they need to be carried over a link whose admitted PCN rate already exceeds an admissible rate. Flow termination (FT) is a new flow control function that terminates already some admitted flows when they are carried over a link whose admitted PCN rate exceeds a supportable rate. The latter condition can occur in spite of AC, e.g., when traffic is rerouted due to network failures.**

This survey gives an introduction to PCN in an early stage of the standardization process. It presents and discusses the multitude of architectural design options for PCN in a comprehensive and streamlined way before only a subset of them is standardized by the IETF. It brings PCN from the IETF to the research community and serves as historical record.

I. Introduction

IP networks were initially designed to perform packet forwarding without priorities. To achieve quality of service (QoS), the differentiated services (DS, DiffServ) concept introduced various service classes called per-hop behaviors (PHBs) [9]. To avoid congestion for premium traffic in a network, admission control (AC) limits the number of high-priority flows. It is a well-established flow control function for packet-switched communication networks supporting high-quality realtime applications such as voice and video. It is useful when capacity overprovisioning is difficult, too costly, or just not possible. The resource reservation protocol RSVP [10] supports admission control with per-flow reservations in each RSVP-aware node. This is a rather heavy burden for transit routers that need to keep per-flow states just to perform correct AC decisions.

AC is not enough to keep the traffic load in a DiffServ domain low. When links or nodes fail, traffic is rerouted which possibly leads to congestion on backup paths. This degrades the QoS for all flows on the congested links. In such a case, the traffic load should be quickly reduced by terminating some of the admitted flows. This is achieved by a new flow control function which is called flow termination (FT). It complements AC and is useful not only in failure cases but also in other case of overload which might be caused, e.g., by flash crowds [3], [15], [26] or unexpected rate increases of admitted flows.

The Internet Engineering Task Force (IETF) currently standardizes simple, robust, and scalable AC and FT mechanisms for DiffServ domains based on pre-congestion notification (PCN) [20]. A new prioritized traffic class for admitted PCN traffic is defined. The rate of aggregate PCN traffic is metered on all links of a DiffServ domain and packets are appropriately marked when certain rate thresholds (admissible rate, supportable rate) are exceeded. Thereby, the PCN egress nodes are notified about load conditions inside the network before congestion occurs. This information is used to perform the AC and FT decisions.

For the time being, several partly incompatible and competing proposals for PCN-based AC and FT exist. However, the objective of the standardization process is to define only one or two mechanisms to achieve compatibility among vendors. This paper develops an integrated overview of methods for metering and marking, PCN encoding, AC, and FT that were presented in different proposals. To that end, a unifying nomenclature is developed. This presentation on the level of individual concepts and features instead of packaged deployment scenarios facilitates an objective discussion of pros and cons and deepens the understanding of PCN and its associated algorithms. Thereby, it is a step forward concerning the standardization of a future PCN architecture. Moreover, the paper preserves the wealth of diverse ideas for PCN-based AC and FT beyond standardization.

The paper is structured as follows. Sect. II reviews the historic roots of PCN and related work. Sect. III introduces different types of pre-congestion, explains the basic idea of PCN, and illustrates its use in the Internet. Sect. IV presents metering and marking algorithms and Sect. V discusses how PCN marks can be encoded into the current IPv4 header. Sect. VI and Sect. VII review various AC and FT methods. Eventually, existing proposals are reviewed by Sect. VIII. Finally, Sect. IX summarizes this work.

[†] This work was partly funded by Deutsche Forschungsgemeinschaft (DFG) under grant TR257/18-2. The authors alone are responsible for the content of the paper.

[‡] This work is partly funded by Trilogy, a research project supported by the European Community under its Seventh Framework Programme. It reflects only the views of the authors.

II. Historic Roots of PCN and 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.

A. Random Early Detection (RED)

RED was originally presented in [22], and in [11] it was recommended for deployment in the Internet. It was intended to detect incipient congestion on a link and to throttle only some TCP flows early to avoid severe congestion and to improve the TCP throughput. RED measures the average buffer occupation *avg* in routers and packets are dropped or marked with a probability that increases linearly with the average queue length *avg*. Thus, a few packets are dropped before buffer overflow occurs which possibly leads to early rate reduction of some TCP flows prior to severe overload.

B. 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 [53]. Packets of non-ECN-capable flows can be differentiated by a “not-ECN-capable transport” codepoint (not-ECT, ‘00’) from packets of a ECN-capable flow which have an “ECN-capable transport” codepoint (ECT). 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 in this case. 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 [53]. 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 [59] that do not conform to the ECN semantics. The four codepoints are encoded in the (“currently unused”) bits of the DS field in the IP header which is a redefinition of the type of service octet [49]. The ECN bits can be redefined by other protocols and [21] provides guidelines for that. They are likely to be reused for encoding of PCN marks.

C. Admission Control

Recent surveys and classifications of AC methods can be found in [32], [34], [63]. We explain the problem with per-flow reservations, reservation aggregation to mitigate that problem, and show which problems still remain. We briefly review some specific AC methods that can be seen as forerunners of the PCN principle. They measure the rate of admitted traffic on each link of a network and give feedback to the network boundary if that rate exceeds a pre-configured admissible rate threshold. Thereby, no per-flow reservations need to be kept for a link and the network core remains stateless. This is a key property of PCN-based AC.

1) Aggregation of Per-Flow Reservations: Admission control can be performed in the Internet using the resource reservation protocol [10]. It sets up per-flow states in any node along the path which leads to a large number of states on links carrying many flows. The setup and maintenance of these states is a large burden for routers and makes them more complex. RSVP aggregation [6] improves this scalability concern by setting up tunnels so that individual flows need to be handled only at the edge nodes of the network. However, an n^2 scalability problem of aggregated tunnels still remains when n boundary nodes set up overlay reservations for premium communication. Forecasts predict that the average number of flows of typical edge-to-edge premium service tunnels is very low and their distribution is long-tailed [18]. As a consequence, the majority of aggregated reservations do not carry traffic most of the time but need to be supported by core nodes. Thus, other simple solutions for AC with better scaling properties in core routers are needed. PCN requires neither per-flow nor per-tunnel information in transit nodes.

2) 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 measure the rate of the observed tickets and can thereby estimate the expected load of reserved traffic. In case of a new reservation request, the ingress router sends probe tickets, intermediate routers forward them to the egress router if they have still enough capacity to support the new flow, and the egress router bounces them back to the ingress router to indicate a successful reservation. If intermediate routers do not have enough resources to carry another flow, they discard the probe tickets, the ingress router does not receive a positive response, and the reservation request is blocked. The tickets can also be encoded by a packet state. Several stateless core mechanisms work according to this idea [1], [60], [61].

3) Admission Control Based on Packet Marking: Gibbens and Kelly [23], [29] 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. The important difference to RED-like packet marking is that marking decisions are based on a virtual instead of a physical queue. This allows to limit the utilization of the link bandwidth by premium traffic to arbitrary values between 0 and 100%. Karsten and Schmitt [27], [28] 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. An early version of PCN-based AC has been reported in [58].

4) Resilient Admission Control: In resilient networks, rerouting or protection switching deviates traffic in case of a failure to backup paths. Overviews of such techniques can be found in [51] and [16]. The objective of resilient AC is to work properly even in case of failures and to avoid termination of already admitted traffic. Transit nodes of a network without reservation states seem to be a prerequisite for resilient AC. In case of a failure, traffic just needs to be rerouted but reservation states do not need to be recovered. Resilient AC admits only so much traffic that it can still be carried after rerouting in a protected failure scenario [39], [46]. It is necessary since overload occurs in wide area networks mostly due to link failures and not due to increased user activity [24]. It can be implemented with PCN by setting the admissible rate thresholds low enough so that admitted traffic is not lost due to rerouting in likely failure scenarios. In particular, the PCN traffic rate on a link after rerouting must be low enough so that flow termination is not triggered. Algorithms to configure PCN-based AC and FT for resilient AC are presented in [38]. It also optimizes IP routing to maximize the rate of admissible traffic for resilient AC.

III. PCN-Based Flow Control

This section explains the basic idea of PCN-based admission control (AC) and flow termination (FT) and discusses its application in an edge-to-edge and end-to-end context in the Internet.

A. Pre-Congestion Notification (PCN)

PCN defines a new traffic class that receives preferred treatment by PCN nodes similar to the expedited forwarding per-hop-behavior (EF PHB) in DiffServ [25]. 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 load regimes 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 rate above $AR(l)$ is *AR*-overload. In this state, no further flows should be admitted. If the PCN traffic rate $r(l)$ is above $SR(l)$, the link is *SR*-pre-congested and the rate above $SR(l)$ is *SR*-overload. In this state, some already admitted flows should be terminated to reduce the PCN rate $r(l)$ below $SR(l)$. A path is *AR*-pre-congested if at least one of its links is *AR*-pre-congested and it is *SR*-pre-congested if at least one of its links is *SR*-pre-congested; otherwise it is not pre-congested.

B. A Two-Level Architecture for PCN-Based AC and FT

PCN-based AC and FT can be described as a two-level architecture which is illustrated in Fig. 2. PCN nodes monitor the PCN rate on their links and mark packets depending on the type of pre-congestion. These mechanisms constitute the packet marking layer (PML). Different proposals exist for the PML, but within a single PCN domain, the same methods need to be implemented in all PCN nodes. PCN egress nodes or PCN endpoints evaluate the packet markings and their essence is reported to the AC and FT entities. Based on this notification, further flows are admitted or blocked and already admitted flows are terminated if necessary. The AC and FT algorithms constitute the admission control and flow termination layer (ACL, FTL). Different implementations of the ACL and FTL may be deployed within a single PCN domain as long as they coexist in a fair way, i.e. block or terminate traffic at the same PCN traffic rate.

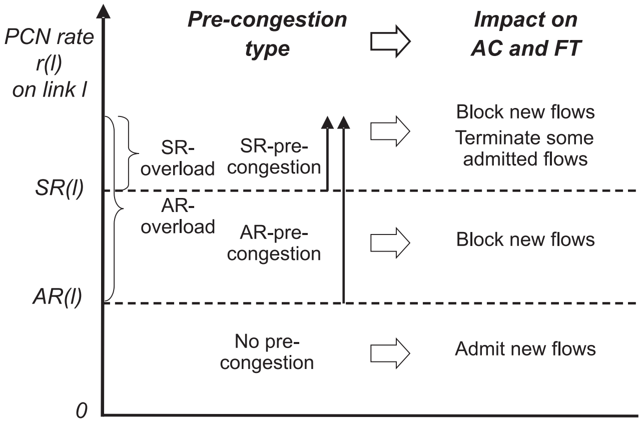

Fig. 1. The admissible and the supportable rate ($AR(l), SR(l)$) define three types of pre-congestion.

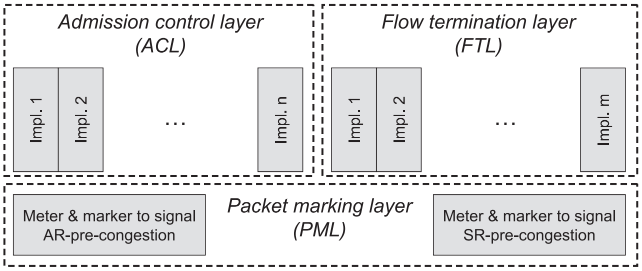
Admission control layer (ACL)
Impl. 1
Impl. 2
...
Impl. n
Flow termination layer (FTL)
Impl. 1
Impl. 2
...
Impl. m
Meter & marker to signal AR-pre-congestion
Packet marking layer (PML)
Meter & marker to signal SR-pre-congestion

Fig. 2. Packet metering and marking is performed on all interfaces of a PCN domain; the markings are evaluated at the network edges to support AC and FT.

C. Edge-to-Edge PCN

Edge-to-edge PCN assumes that some end-to-end signalling protocol (e.g. SIP or RSVP) or a similar mechanism requests admission for a new flow to cross a so-called PCN domain similar to the IntServ-over-DiffServ concept [8]. Thus, edge-to-edge PCN is a per-domain QoS mechanism and presents an alternative to RSVP clouds or extreme capacity overprovisioning. This is illustrated in Fig. 3. Traffic enters the PCN domain only through PCN ingress nodes and leaves it only through PCN egress nodes. Ingress nodes set a special header codepoint to make the packets distinguishable from other traffic and the egress nodes clear the codepoint. The nodes within a PCN domain are PCN nodes. They monitor the PCN traffic rate on their links and possibly remark 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 of the PCN domain.

D. End-to-End PCN

End-to-end PCN [42] assumes that all links providing QoS support implement PCN metering and marking. The communication endpoints, i.e. source and destination of a PCN flow or proxies thereof, react to the packet markings in a similar way as to ECN but perform AC and FT instead of rate reduction. Since PCN sources and destinations take over the functionality of PCN ingress and egress nodes, the concept

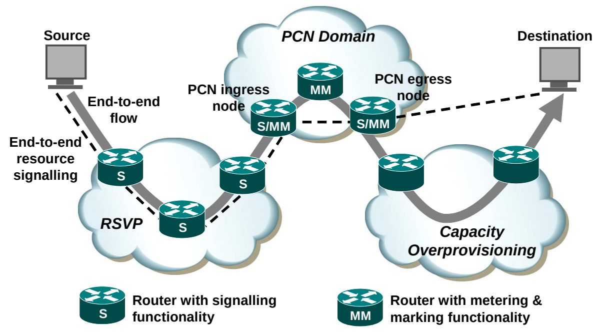
Source
PCN Domain
Destination
PCN ingress node
MM
PCN egress node
S/MM
S/MM
End-to-end flow
End-to-end resource signalling
S
S
S
RSVP
Capacity Overprovisioning
S
Router with signalling functionality
MM
Router with metering & marking functionality

Fig. 3. Edge-to-edge PCN is triggered by admission requests from external signalling protocols and guarantees QoS within a single PCN domain.

of a PCN domain is no longer needed. Packets from end-to-end PCN flows are preferentially forwarded by all upgraded PCN nodes in the Internet. When they traverse an edge-to-edge PCN domain, they do not receive special treatment by the network boundaries, but they are metered, possibly marked, and preferentially forwarded like packets from edge-to-edge PCN flows. This is illustrated in Fig. 4. Hence, the deployment of end-to-end PCN in the Internet is more attractive when sufficiently many edge-to-edge PCN islands already exist. However, end-to-end PCN is rather a solution for deployment in corporate networks than in the general Internet because of trust issues. Therefore, the current charter of the IETF WG on PCN covers only the standardization of edge-to-edge PCN.

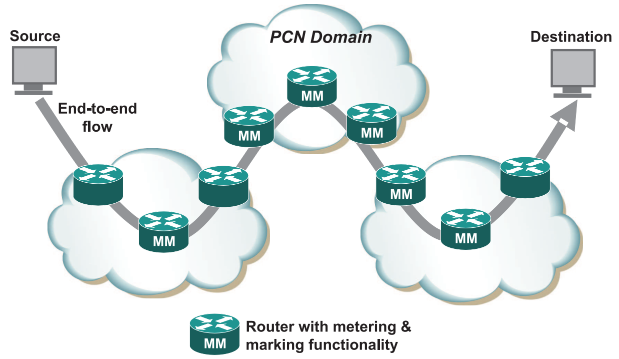

Fig. 4. End-to-end PCN flows transparently traverse edge-to-edge PCN domains and perceive them as islands with only PCN-capable nodes from which they receive preferred treatment.

Mechanisms for end-to-end PCN are more challenging than for edge-to-edge PCN. An ingress-egress aggregate (IEA) comprises all PCN flows between one PCN ingress node and another PCN egress node. With edge-to-edge PCN, the PCN egress node can evaluate the packet markings per IEA and base its AC and FT decisions on aggregated feedback of multiple flows. With end-to-end PCN, individual PCN endpoints can evaluate the markings of only their own flows. This limits the choices of applicable metering- and marking as well as AC and FT algorithms for end-to-end PCN [42].

IV. Metering and Marking

The core idea of PCN is that packets are metered and marked on the links of a PCN domain to give feedback about its pre-congestion state to its boundary nodes. Four basically different metering and marking algorithms are used to detect pre-congestion: excess marking, excess marking with marking frequency reduction, exhaustive marking, and fractional marking. In the following, we describe the metering and marking algorithms based on token buckets (TB). Other principles, e.g. virtual queues [41], can also be used for implementation.

A. Excess Marking

Excess marking [19] marks those packets that exceed a certain reference rate R on a link so that the non-marked traffic rate is at most R. When configured with the admissible or supportable rate (AR, SR) as reference rate, the rate of the excess-marked traffic is an estimate of the AR- or SR-overload.

1) Plain Excess Marking: Plain excess marking uses a TB with a bucket size S. The TB is continuously filled with tokens with a reference rate R and the variable F shows its fill state, i.e. the number of tokens in the bucket. The variable lU records the time when the TB was last updated and the global variable *now* indicates the current time.

Algorithm 1 is called for each packet. First, the fill state F of the TB is updated and so is lU. Only unmarked packets are metered and marked. If F is smaller than the packet size B, the packet is marked. Otherwise, the number of tokens in the bucket is reduced by the packet size B.

```
Input:  token bucket parameters $S$, $R$, $lU$, $F$, packet
        size $B$ and marking $M$, current time $now$
  $F = \min(S, F + (now - lU) \cdot R)$;
  $lU = now$;
  if ($M \neq$ marked) then
    if $(F < B)$ then
      $M =$ marked;
    else
      $F = F - B$;
    end if
  end if
```

Algorithm 1: Excess Marking: only those packets exceeding the reference rate R are marked.

This type of marking behavior has the great advantage that it is readily available in today's routers. It is used by various proposals [4], [5], [14], [35] that are reviewed in Sect. VIII-A, Sect. VIII-B, Sect. VIII-C, and Sect. VIII-D.

2) Excess Marking with Packet Size Independent Marking (PSIM): The marking in Algorithm 1 depends on the packet size B. This can lead to unfair treatment of flows with large packets if the packet markings are used as hints whether a certain flow should be admitted or terminated [42]. Packet size independent marking can be achieved by substituting the condition $(F < B)$ in Algorithm 1 by $(F < 0)$. As a consequence, the fill state can become negative for a while.

B. Excess Marking with Marking Frequency Reduction (MFR)

The proposals in [5] and [62] (cf. Sect. VIII-C and Sect. VIII-F) require that only a fraction of the traffic rate, that is above the reference rate R, is marked. This can be achieved by excess marking with marking frequency reduction (MFR). Simple MFR takes only the number of marked packets into account while proportional MFR takes also their size into account. We show how both options can be implemented.

1) Excess Marking with Simple MFR: Simple MFR is achieved by extending Algorithm 1 with ($\textbf{if}\ (M = marked)\ \textbf{then}\ F = \min(S, F+I)$) at its very end. Thus, a fixed increment of I tokens is added to the TB for each marked packet. Note that it is irrelevant whether the packet was marked by the current call of the algorithm or by a previous call at a preceding node.

2) Excess Marking with Proportional MFR: It was shown in [42], that MFR in proportion to the size of marked packets improves the control over some FT algorithms. It can be achieved by scaling the increment I with the size of the marked packet: $I = \beta \cdot B$ where β is a constant scaling factor.

C. Exhaustive Marking

Exhaustive marking marks all packets on a link when the metered rate exceeds its reference rate R. We present two different implementations that provide similar marking behavior.

1) Threshold Marking: The basic structure of threshold marking is similar to the one of excess marking. However, packets are marked if the fill state F of the TB is lower than a configured threshold T, i.e., marking is independent of the packet size. Moreover, the fill state F is reduced by the size of each metered packet regardless of whether it was already marked or not. Algorithm 2 explains threshold marking in detail.

```
Input:  token bucket parameters $S$, $R$, $lU$, $F$, $T$,
        packet size $B$ and marking $M$, current time
        $now$
  $F = \min(S, F + (now - lU) \cdot R)$;
  $lU = now$;
  if $(F < T)$ then
    $M =$ marked;
  end if
  $F = \max(0, F - B)$;
```

Algorithm 2: THRESHOLD MARKING: all packets are marked if the PCN rate exceeds the reference rate R.

If the metered traffic rate exceeds the reference rate R, the tokens are faster consumed than refilled and the fill state F of the TB goes to zero and remains small. Therefore, F stays below the marking threshold T and all packets are marked. Threshold marking is applied by [4], [5], [35], and [57] (cf. Sect. VIII-A, Sect. VIII-C, Sect. VIII-D, and Sect. VIII-E).

2) Ramp Marking: The intention of ramp marking is to start marking early when the fill state of the TB is still high. Packets are marked with a probability that depends on the TB fill state F. It linearly increases from an upper TB threshold T_{ramp} to a lower TB threshold T. If F is below T, all packets are marked. Ramp marking can emulate threshold marking by setting $T_{ramp} = T$. Ramp marking is clearly inspired by RED. In contrast to RED [22], the marking probability depends on the current TB fill state F instead of an exponential average thereof. Ramp marking is more complex and computationally expensive than threshold marking since it requires random numbers. Ramp marking was considered as an alternative to threshold marking in [4] (cf. Sect. VIII-A). Ramp and threshold marking have been investigated in [41], but no significant benefit of ramp marking was found.

D. Fractional Marking

In contrast to exhaustive marking, fractional marking marks only $1/N$ of the traffic when the metered rate exceeds its reference rate R. Algorithm 3 achieves that behavior. It is a simple extension of threshold marking and requires an additional byte counter Cnt. Its behavior differs from threshold marking only if the fill state F of the token bucket falls below its threshold T. In that case, the packet is marked only if the counter Cnt is negative and then the counter Cnt is increased by $N \cdot B$. Afterwards, the counter Cnt is decreased by the packet size B regardless of its value. This modification effects that only $1/N$ of the PCN traffic is marked when the metered rate exceeds the reference rate R. This algorithm also achieves packet size independent marking. The algorithm can be easily modified so that $1/N$ of the packets are marked instead $1/N$ of the data rate. Fractional marking is used in [57] (cf. Sect. VIII-E).

```
Input:  token bucket parameters $S$, $R$, $lU$, $F$, $T$,
        counter $Cnt$, denominator $N$ of fraction
        $1/N$, packet size $B$ and marking $M$, current
        time $now$
  $F = \min(S, F + (now - lU) \cdot R)$;
  $lU = now$;
  if $(F < T)$ then
    if $(Cnt < 0)$ then
      $M =$ marked;
      $Cnt = Cnt + N \cdot B$;
    end if
    $Cnt = Cnt - B$;
  end if
  $F = \max(0, F - B)$;
```

Algorithm 3: FRACTIONAL MARKING: $1/N$ of the traffic is marked if the PCN rate exceeds the reference rate R.

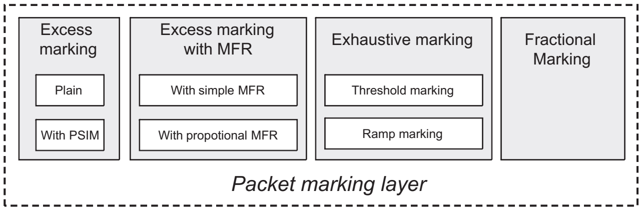

Fig. 5. Applicability of AC methods with different marking schemes.

E. Summary of PCN Marking Methods

The presented metering and marking methods are summarized in Fig. 5. Excess marking marks the metered traffic that exceeds the reference rate of the marker. There are two excess marking methods: plain excess marking has the tendency to mark larger packets with higher probability. This is different for excess marking with packet size independent marking. Excess marking with marking frequency reduction (MFR) marks traffic in proportion to the metered traffic that exceeds the reference rate. The strength of the MFR can be independent of or proportional to the size of the marked packets. Exhaustive marking marks all packets if the metered traffic exceeds the reference rate. In contrast to threshold marking, ramp marking reacts more sensitive to fluctuations of the metered traffic. In case of short-term traffic bursts, it marks more packets than threshold marking when the rate of the metered traffic is still below the reference rate, but this does not significantly impact the behavior of PCN-based AC and FT. Fractional marking is similar to threshold marking, but it marks only $1/N$ of the traffic when the metered traffic exceeds its reference rate.

V. Encoding Options for PCN Marking

PCN requires an encoding scheme to record in the IP header whether a packet belongs to a PCN flow and whether it has been re-marked by a PCN node due to pre-congestion. The difficulty is that there are almost no free bits in the IP header that can be used for that purpose so that bits which are already in use need to be reused. First, we briefly summarize general encoding issues and then we present several encoding options that are currently discussed in IETF. Finally, we present an abstraction that allows to speak about packet markings without the knowledge of the exact encoding scheme.

A. Encoding Issues with DSCPs, the ECN Field, and Tunneling

The differentiated services (DS) field in the IP header [49] is planned to be reused for PCN encoding. The type of service (TOS) octet in the IPv4 header [50] and the traffic class octet in the IPv6 header [17] were redefined to the DS field in [49]. It consist of the 6 bit DiffServ codepoint (DSCP) and the 2 bit "currently unused" (CU) field. Later, the CU field was renamed to the explicit congestion notification (ECN) field [52], [53]. Encoding in MPLS is even more challenging. To differentiate traffic, the 4 bytes shim header has only the 3 bit EXP-field for experimental use [54]. It has recently been renamed to the traffic class (TC) field [2].

In the following, we explain constraints that need to be respected when reusing the DS field for PCN encoding.

1) Problems with DSCPs: DSCPs are intended to indicate the per-hop behavior (PHB) for a packet. The PHB denotes how a packet is to be scheduled and buffered or dropped inside a DiffServ node. It has only local meaning as ingress nodes of DiffServ domains can change the DSCP of a packet. This is a potential threat to the persistence of PCN markings when PCN should ever be extended towards multiple domains. The DSCP may be reused either to just indicate that a packet belongs to a PCN-enabled flow or to indicate both whether a packet belongs to the PCN class and whether it is marked or not. The latter requires at least two DSCPs which is problematic as only very few DSCPs are available. In addition, if more than a single PCN class should ever be supported, the number of required DSCPs scales with the number of supported PCN classes.

2) Problems with the ECN Field and Tunneling: The encoding scheme must cope with tunneling within PCN domains. However, various tunneling schemes limit the persistence of the ECN field in the top-most IP header to a different degree. Two IP-in-IP tunnelling modes are defined in [53] and a third one in [55] for IP-in-IPsec tunnels.

The limited-functionality option in [53] requires that the ECN codepoint in the outer header is set to not-ECT. As a consequence, ECN routers along the tunnel drop packets instead of mark them in case of congestion. The tunnel egress just decapsulates the packet and leaves the ECN codepoints of the inner packet header unchanged. This tunneling mode is not useful for tunnels inside PCN regions because the ECN marking information from the outer ECN field is lost upon decapsulation.

The full-functionality option in [53] requires that the ECN codepoint in the outer header is copied from the inner header unless the inner header codepoint is CE. In this case, the outer header codepoint is set to ECT(0). This choice has been made for security reasons to disable the ECN fields of the outer header as a covert channel. Upon decapsulation, the ECN codepoint of the inner header remains unchanged unless the outer header ECN codepoint is CE. In this case, the inner header codepoint is also set to CE. This preserves outer header information if it is CE. However, the fact that CE marks of the inner header are not visible in the outer header is a problem for all sorts of excess marking as they take already marked traffic into account (cf. Sect. IV-A and Sect. IV-A2). Moreover, it is a problem for some FT mechanisms that require preferred dropping of marked packets to work properly (cf. Sect. VII-F2, VIII-A, and VIII-B).

Tunneling with IPSec copies the inner header ECN bits to the outer header ECN bits [55, Sect. 5.1.2.1] upon encapsulation. Upon decapsulation, CE-marks of the outer header are copied into the inner header, the other marks are ignored. With this tunneling mode, CE marks of the inner header become visible to all meters, markers, and droppers for tunneled traffic. In addition, information from the outer header can be propagated into the inner header. Therefore, only IPSec tunnels should be used inside PCN domains when ECN bits are reused for PCN encoding. However, limitations still apply. Only the CE codepoint can be used to re-mark packets as the change of one of the other codepoints in the outer header to any other codepoint is not persistent after decapsulation.

3) Problems with the ECN Field: The guidelines in [21] describe how the ECN bits can be reused while being compatible with [53]. A CE mark of a packet must never be changed to another ECN codepoint. Furthermore, a not-ECT mark of a packet must never be changed to one of the ECN-capable codepoints ECT(0), ECT(1), or CE. When the ECN field is reused for PCN marking, care must be taken that this rule is enforced when PCN packets leave the PCN domain. There are two basic options to handle ECN flows when the ECN field is reused for PCN marking in a DiffServ domain.

a) Disabling ECN: The PCN ingress node sets the appropriate ECN mark in incoming packets to indicate that they are initially unmarked. The PCN egress node resets their ECN field to not-ECT to make sure that previous not-ECT marks are not changed to any other ECN marks through the PCN domain. This disables ECN for PCN flows so that they cannot profit from both ECN and PCN. As it is prohibitive to change CE marks to not-ECT, CE-marked packets must be dropped by PCN ingress nodes.

b) Tunneling ECN Marks: Another option is tunneling ECT- or CE-marked packets through the PCN domain using the limited-functionality mode. This preserves the original ECN field so that PCN egress nodes receive PCN feedback and end systems receive ECN feedback which is not modified by the PCN domain. Moreover, CE-marked packets do not need to be dropped by the PCN ingress node.

B. Encoding Options

Different proposals for PCN-based AC and FT require a different number of codepoints to mark packets. Therefore, many encoding options have been presented and discussed in IETF. However, we review only those that use a DSCP to indicate PCN traffic, use the ECN field to indicate the marking, and conform with the limitations due to tunneling.

The VOICE-ADMIT DSCP is currently about to be standardized to indicate EF-PHB for AC-controlled flows [7]. All encoding schemes presented in this section assume that the ECT(0), ECT(1), and CE codepoints of this DSCP can be reused to mark PCN traffic and that only its not-ECT codepoint remains for the original purpose of VOICE-ADMIT. By disallowing the other ECN codepoints for this traffic type in the PCN domain, VOICE-ADMIT flows cannot profit from ECN unless their packets are tunneled through that domain and PCN marking is applied only to the outer header as described in Sect. V-A3.

1) Baseline Encoding: Baseline encoding has been presented in [48]. The meaning of the ECN field if the PCN DSCP is set is summarized in Table I. The not-ECT codepoint is used as "not-PCN" indicating that this traffic is not under PCN control. ECT(0) is reused to label "not-marked" (NM) PCN packets and CE is reused to label "marked" (M) packets. ECT(1) is reserved for "experimental use" (EXP) to allow encoding extensions. When PCN packets enter a PCN domain, they are marked with a NM codepoint and they are possibly re-marked to M by PCN nodes. Hence, this encoding scheme allows the use of a single marking scheme which may be, e.g., excess or threshold marking.

2) PCN 3-State Encoding Extension in a Single DSCP (3-in-1): 3-in-1 encoding is an extension of baseline encoding and assumes that the re-marking limitations due to tunneling (cf. Sect. V-A2) will be resolved in the future, e.g., by [12]. That means, ECT(1) and CE must be copied from the outer header to the inner header upon decapsulation. As a consequence, two different marking schemes can be concurrently used: ECT(1) indicates that packets are marked by the one scheme and CE indicates that packets are marked by the other scheme. As most proposals use threshold and excess marking, these codepoints are called ThM and EcM (cf. Table I). Since they allow re-marking of ThM-marked packets to EcM-marked packets but not vice-versa, CE is chosen for EcM to be compatible with [21].

3) Packet-Specific Dual Marking: Packet-specific dual marking (PSDM) has been presented in [36], [37] as an extension of baseline encoding. It also supports two concurrent marking schemes. However, in contrast to 3-in-1 encoding it does not assume any changes to the tunneling rules and supports only one marking scheme per packet. Table I summarizes the meaning of its ECN field. Unmarked packets that are subject to excess marking have the EcNM codepoint in their header while unmarked packets that are subject to threshold marking have the ThNM codepoint. When a packet is marked by the marking scheme it is subject to, its codepoint is set to "marked" (M). The marking algorithms must be configured so that excess marking re-marks only ExNM packets to M and threshold marking re-marks only ThNM packets to M. PSDM is useful when AC relies on probe packets (cf. Sect. VI-A and Sect. VI-C) that are subject to threshold marking and FT relies on data packets that are subject to excess marking. The benefit of PSDM is that two marking schemes are supported using only a single DSCP. When routers implement two marking schemes, but only one of them is used, the routers do not need to be configured which marking scheme applies as the packets tell them which marking scheme to use. This is another benefit of the PSDM semantics.

4) General Dual Marking: General dual marking (GDM) is an extension of baseline encoding that supports two concurrent marking schemes. In contrast to PSDM, both marking schemes can apply to all PCN packets and in contrast to 3-in-1, GDM does not assume modified tunneling rules. As only the CE codepoint can be used for re-marking, another DSCP is needed in addition to VOICE-ADMIT for which ECN is also disabled. The meaning of the combined DSCP and ECN field is summarized in Table I. When packets of a PCN flow enter a PCN domain, their DS field is set to NM. When packets are threshold- or excess-marked, their DS field is set to ThM or to EcM. Excess markers meter NM- and ThM-marked packets and possibly re-mark them to EcM. Threshold markers meter all PCN packets and possibly re-mark only NM-marked packets to ThM.

5) General Dual Marking with Limited ECN Support: GDM with limited ECN support (GDM-LES) is an extension of GDM [47]. It suggests to set the DS field of packets belonging to PCN-enabled flows to NM(not-ECT), NM(ECT(0)), NM(ECT(1)), or NM(CE) according to the value in the ECN field before entering the PCN domain (cf. Table I). Thus, CE-marked packets do not need to be dropped by the PCN ingress node (cf. Sect. V-A3). When PCN packets leave the PCN domain, the original ECN field of NM-marked packets is restored and the DS field of ThM- or EcM-marked packets of ECN-enabled flows is set to CE. This provides PCN-feedback to ECN-capable endpoints which may be useful in the future [56]. However, this mechanism requires signaling from the endpoints to indicate whether this combined ECN and PCN feedback is desired. Thus, GDM-LES induces significant

TABLE I
INTERPRETATION OF THE ECN FIELD FOR VARIOUS PCN ENCODING OPTIONS.

Encoding	DSCP	not-ECT (‘00’)	ECT(0) (‘10’)	ECT(1) (‘01’)	CE (‘11’)
Baseline	VOICE-ADMIT	not-PCN	NM	EXP	M
3-in-1	VOICE-ADMIT	not-PCN	NM	ThM	EcM
PSDM	VOICE-ADMIT	not-PCN	EcNM	ThNM	M
GDM	VOICE-ADMIT	not-PCN	NM	CU	ThM
GDM	DSCP 2	not-PCN	CU	CU	EcM
GDM-LES	VOICE-ADMIT	not-PCN	NM(Not-ECT)	NM(CE)	ThM
GDM-LES	DSCP 2	not-PCN	NM(ECT(0))	NM(ECT(1))	EcM

complexity.

6) Providing PCN Feedback to ECN Receivers: If ECN receivers wish to receive combined ECN feedback from outside PCN domains and PCN feedback from inside PCN domains [56], this needs to be signaled explicitly to PCN ingress and egress nodes (cf. Sect. V-B5). This behavior can be achieved when PCN ingress nodes encapsulate the packets in IPSec tunnels and PCN egress nodes decapsulate this traffic. Thus, ECN marks are saved through the PCN domain and potential PCN marks are added (cf. Sect. V-A2).

C. Encoding Abstraction

In the remainder of this paper, we abstract from the specific encoding scheme. We assume that all unmarked packets are labelled with “no-pre-congestion” (NP), packets are re-marked to “admission-stop” (AS) when the reference rate of the marker was set to the admissible rate and to “excess-traffic” (ET) when the reference rate of the marker was set to the supportable rate. When two concurrent marking schemes are in use, AS-marked packets are possibly re-marked to ET but not vice-versa.

VI. PCN-BASED ADMISSION CONTROL (AC)

When PCN markers are configured with the admissible rates of the links, they start marking traffic as soon as the PCN rate on the links exceeds that rate. Then, egress node detect AS-marked packets and this information is used to perform AC. There are basically two different approaches for PCN-based AC. Probe-based AC for individual flows relies on the feedback of probe packets that are associated only with these flows. IEA-based AC relies on the current AC state of the ingress-egress aggregate (IEA). We review both of them in the following.

A. Probe-Based AC for Individual Flows (PBAC-IF)

We explain the general concept of PBAC-IF by explicit PBAC-IF and present then how implicit PBAC-IF can do without explicit probe packets.

1) Explicit Probing: With explicit probing, the PCN ingress node generates upon admission request one or more unmarked probe packets and sends them to the appropriate PCN egress node. The egress node returns the probe packets to the PCN ingress node and if the PCN ingress node receives all of them unmarked, the new flow can be admitted, otherwise it must be blocked. This delays the probing decision by at least one round trip time of the PCN domain. Probing basically works with any marking scheme. However, with exhaustive marking, a single probe packet is enough to test whether the prospective path of the new flow is *AR*-pre-congested. With excess or fractional marking, only some packets are marked and many probe packets are needed for a reliable admission decision [44].

If the PCN ingress node does not know the corresponding PCN egress node for an admission request, the probe packets can be sent to the final destination and they are intercepted by the respective PCN egress node to avoid that they leak out of the PCN domain. In case of multipath routing, probe packets must even have the same source and destination address and port as the future data packets to guarantee that they are forwarded on the same path. This is due to the fact that routers usually apply flow-based load balancing algorithms [33].

2) Implicit Probing: Probing can also be done implicitly, e.g., in the presence of an end-to-end resource reservation protocol such as RSVP [5]. To establish a reservation, RSVP sends a PATH message to explore the path of the future data packets and each RSVP-enabled node sets up a PATH state. The destination responds with a RESV message to set up the reservation (RESV state) hop-by-hop along the explored path. PATH and RESV messages are periodically sent to refresh the flow states as they otherwise expire (soft state principle). We briefly explain how PATH and RESV messages can be reused for probing. Interior nodes of a PCN domain are usually RSVP-disabled so that PCN ingress and egress node are neighboring RSVP nodes. When the PCN egress node receives an initial PATH message, it forwards the message as usual if it is not AS-marked. Otherwise, it sends back a PATHERR message to the previous RSVP hop to indicate that the new flow should be blocked. Thus, when the PCN ingress node receives an initial RESV message, the corresponding PATH message was not AS-marked when travelling across the PCN domain and the respective flow can be admitted. In contrast to explicit probing, implicit does not required explicit probe packets and it does not delay the reservation setup.

B. Ingress-Egress-Aggregate-Based AC (IEABAC)

IEABAC assumes that all traffic from one PCN ingress to another PCN egress node takes the same path. Each IEA is associated with a single AC state K whose value is either $admit$ or $block$. When a new flow requests admission, the AC entity needs to find out which IEA the new flow belongs to and then it admits or blocks it depending on the AC state K of that IEA. More precisely, the PCN ingress node keeps the AC state K and the PCN egress node sends admission-stop and admission-continue messages to toggle the admission control state K of the PCN ingress node. In the following, we present three different methods to control the AC state K of an IEA.

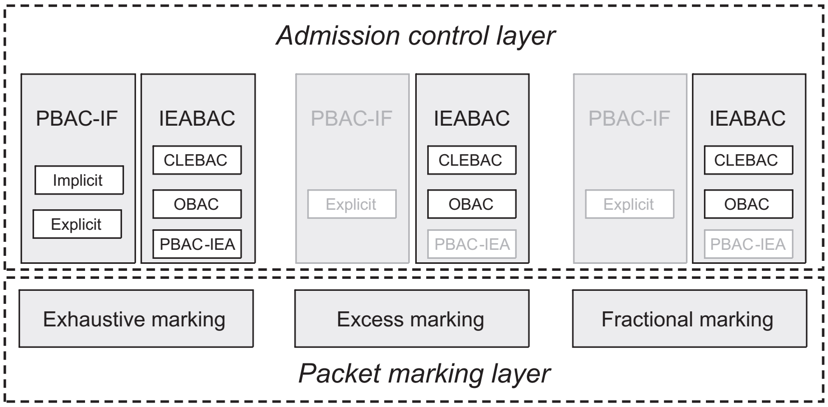

Fig. 6. Applicability of AC methods with different marking schemes; technically difficult solutions are greyed out.

1) CLE-Based AC (CLEBAC): With CLEBAC, the PCN egress node measures the rates of AS-marked and non-AS-marked data traffic (*ASR*, *nASR*) per IEA [4], [5], [62]. This is done based on measurement intervals of duration D_{MI}. Then, the congestion level estimates $CLE = \frac{ASR}{ASR+nASR}$ are calculated. If the CLE is smaller than or equal to a certain threshold T_{CLE}, the AC state K is set to *admit*; otherwise it is set to $block$. This method has two parameters: D_{MI} and T_{CLE}.

To avoid oscillations of the AC state K, the following hysteresis may be used. If the CLE value exceeds an admission-stop threshold T_{CLE}^{AStop}, the AC state K is turned to *block*; if it falls below an admission-continue threshold T_{CLE}^{ACont}, the AC state K is turned to *admit*; otherwise, the AC state K is not changed. This method depends on three parameters: D_{MI}, T_{CLE}^{AStop}, and T_{CLE}^{ACont}.

Another variant calculates the CLE based on an exponentially weighted moving average (EWMA), i.e., $CLE_{new} = w \cdot \frac{ASR}{ASR+nASR} + (1-w) \cdot CLE_{old}$ [14].

CLEBAC can be used with any marking scheme. With exhaustive marking, the admission result is rather insensitive to the value of the CLE-thresholds between 0 and 1 [44]. With excess or fractional marking, the CLE-thresholds must be set to positive values close to 0.

2) Observation-Based AC (OBAC): With OBAC, the PCN egress node observes the data traffic per IEA and turns the AC state K of an IEA to *block* when it detects an AS-marked packet [5]. It turns the state back to *admit* when it has not seen an AS-marked packet for D_{block}^{min} time. D_{block}^{min} is the only configuration parameter of OBAC. OBAC works well with exhaustive marking, excess marking, and fractional marking.

3) PBAC for IEAs (PBAC-IEA): With PBAC-IEA, the PCN ingress node sends explicit probe packets in regular intervals to the PCN egress node. This kind of probing is simpler than PBAC-IF since it does not need to make sure that probe packets take the same path as prospective data packets of an admission request. If a probe packet is missing or if it is AS-marked, it turns the AC-state K of the IEA to *block*. It turns K back to *admit* when it has not detected missing or AS-marked packets for D_{block}^{min} time. The frequency of probe packets and D_{block}^{min} are the two parameters of this method. This method can also be applied with any marking scheme. However, excess and fractional marking require a higher frequency of probe packets for reliable admission decisions than exhaustive marking.

C. Discussion of PCN-Based AC Methods

We briefly discuss the applicability of the presented AC methods with different marking schemes, their usefulness in case of low flow aggregation per IEA, their applicability with multipath routing and for end-to-end PCN, and their impact on timeliness and accuracy of AC decisions.

1) Applicability of AC Methods with Different Marking Schemes: Fig. 6 summarizes the options for PCN-based AC. Basically, any AC method can be combined with any marking scheme. However, threshold marking yields clearer feedback than excess or fractional marking and leads to faster and more reliable control of the AC state K for IEABAC. This is only an issue for IEAs with a small number of admitted PCN flows. Moreover, excess and fractional marking require more probe packets for any kind of PBAC so that explicit PBAC-IF and PBAC-IEA are impractical and implicit PBAC-IF is even impossible. The same holds for excess marking with MFR which is omitted in the figure.

Hence, PBAC methods require threshold marking to work well. In contrast, most FT method require excess marking. Therefore, the application of PBAC calls for two marking schemes which is more difficult for PCN encoding than a single marking scheme. However, it can be achieved with PSDM when probe traffic is only subject to threshold marking and data traffic is subject to excess marking.

2) Usefulness of AC Methods in Case of Low Flow Aggregation per IEA: When the average number of PCN flows per IEA is small, many IEAs are even empty. This scenario is even quite likely in the future [18] for large networks carrying realtime flows in spite of many PCN flows per link. Empty IEAs are problematic for CLEBAC and OBAC because they cannot block new admission requests. As a result, overadmission can easily occur [40]. This cannot happen with all PBAC methods including PBAC-IEA.

3) Applicability of AC Methods with Multipath Routing: All IEABAC method including PBAC-IEA cannot cope with multipath routing as the admission of a new request is taken independently of the prospective path of the associated flow. Therefore, flows are possibly admitted although their paths are already *AR*-pre-congested and they are possibly blocked although their paths are not *AR*-pre-congested. This cannot happen with implicit or explicit per-flow probing when probe packets take the same path as future data packets of the flow.

4) Applicability of AC Methods for End-to-End PCN: In case of end-to-end PCN, IEAs do not exist as end systems are the control entities of PCN flows. Therefore, all IEABAC methods are not applicable in this context and only PBAC-IF methods remain for this application scenario.

5) Impact of AC Methods on Timeliness and Accuracy of Admission Decisions: Implicit PBAC-IF is based on recent PCN feedback and does not delay admission decision. Explicit PBAC-IF is also based on recent PCN feedback and delays admission decisions by at least one round trip time of the PCN domain which is quite short. IEABAC methods do not delay admission decisions as they are performed based on the local AC state K. However, the AC state K may have been set a while ago and does not reflect the current pre-congestion state of the associated path. The parameters to control that delay are D_{MI} for CLEBAC, D_{block}^{min} for OBAC and PBAC-IEA, as well as the frequency of probe packets for PBAC-IEA. Moreover, the use of excess or fractional marking for AC also leads to delayed control of the AC state K as only a few packets are marked in case of *AR*-pre-congestion.

VII. PCN-Based Flow Termination (FT)

FT methods use PCN feedback to detect *SR*-pre-congestion and terminate already admitted flows if necessary. There are basically three different approaches: measured-rate based flow termination (MRT), geometric flow termination (GFT), and marked-packet based flow termination (MPT). We provide some general remarks about flow termination, present the different mechanisms in detail, point out general problems with some FT methods, and finally discuss and summarize the shown mechanisms.

A. General Remarks about Flow Termination

We briefly discuss options for termination signalling, the impact of multipath routing, show some motivation for and implications of single marking schemes, and explain what we understand by over- and undertermination.

1) Options for Termination Signalling: We assume that a FT entity can terminate already admitted PCN flows if necessary. Termination implies sending a teardown message, e.g. RESVTEAR in RSVP, and modifying packet filters in the PCN ingress nodes to exclude terminated flows from prioritized forwarding. Basically, the FT entity can be collocated with PCN ingress nodes, PCN egress nodes, or it may be located in a central node. PCN ingress and egress nodes can information the FT entity to remove admitted PCN traffic in three different ways. They may signal the IDs of explicit flows that need to be terminated, they signal the PCN rate that should be terminated (termination rate TR), or they signal the PCN rate that should not be terminated (edge-to-edge supportable rate ESR). While the flows to be terminated are already determined in the first case, the two other options allow the FT entity to choose the flows to be terminated from a larger set of flows, e.g. all flows of a specific IEA. This allows to support termination policies such as low or high termination priorities which can be a useful feature for emergency calls. To work properly, the FT entity must know reliable rate information about admitted flows, e.g., through measurement results or traffic descriptors.

2) Multipath Routing: If multipath routing is used in a network, flows of a single IEA may take different paths [33]. Some of these paths may be *SR*-pre-congested, others not. Depending on the configuration of marking algorithms, a marked packet denotes that the corresponding flow is carried over an *AR*- or *SR*-pre-congested path. We call such a flow also marked. Therefore, marked flows are good candidates for termination while non-marked flows of the same IEA may be carried over non-pre-congested paths. Thus, termination of only marked flows is important for a fast reduction of *SR*-overload and the persistence of flows on non-pre-congested paths [43]. The PCN egress node can record recently marked flows and the FT entity may choose only marked flows for termination. In that case packet size independent marking (cf. Sect. IV-A2) should be used to achieve termination fairness among flows with small and large packets. Moreover, this idea requires that the FT entity is collocated with the PCN egress node or the PCN egress nodes need to communicate the information about marked flows to the FT entity.

3) AC and FT with a Single Marking Scheme: AC methods require that the reference rates of the marker on the links are configured with their admissible rates. FT methods intuitively require that the reference rates of the markers are configured with their supportable rates to provide appropriate feedback. This requires two different marking schemes and at least three codepoints (NM, AS, ET). However, PCN routers with only one marking scheme are cheaper to build than PCN routers with two marking schemes, and three appropriate codepoints are more difficult to claim than only two codepoints due to the unavailability of free codepoints in the IP header (cf. Sect. V). Therefore, single marking schemes that support both AC and FT methods are attractive.

They assume that excess marking is used on all links and that their reference rates is set to the admissible rates of the links. Furthermore, the admissible and supportable rate on all links are connected by

$$SR = u \cdot AR \qquad (1)$$

using a domain-wide constant u. As a consequence, as soon as packets are marked, *AR*-pre-congestion can be detected which is required for AC. And as soon as the proportion of marked packets is larger than $\frac{u}{u+1}$, *SR*-pre-congestion can be detected which is required for FT. In Sect. VII-B, Sect. VII-C, and Sect. VII-D we present various FT methods and show how some of them can take termination decisions based on marked *AR*-overload. These FT methods can be used in combination with single marking scheme.

4) Over- and Undertermination: A FT method is expected to terminate only so much traffic that the PCN rate on a *SR*-pre-congested link is reduced to its supportable rate. If more traffic is terminated, we talk about overtermination. If less traffic is terminated, we talk about undertermination. Inaccurate PCN feedback due to statistical variation or wrong PCN feedback due to multipath routing can cause overtermination. Undertermination can occur in combination with multipath routing and single marking schemes (cf. Sect. VII-E1).

B. Measured-Rate Based Flow Termination (MRT)

MRT requires excess marking in PCN nodes. All operations are performed per IEA. PCN egress nodes classify the received PCN traffic into IEAs and measure the rate of marked or unmarked traffic based on measurement intervals of duration D_{MI}. Flow termination is possibly triggered at the end of such measurement intervals.

1) MRT with Directly Measured Termination Rates (MRT-DTR): MRT-DTR calculates a direct estimate of the termination rate TR and signals it to the FT entity which terminates an appropriate set of flows from the IEA. To avoid overtermination, TR should not be overestimated and a minimum inter-termination time D_{term}^{inter} between consecutive termination actions is required to make sure that the new measurement results for that IEA already reflect the last termination action.

a) MRT-DTR with Marked SR-Overload: When the reference rate of the excess marker is set to the supportable rate, *SR*-overload is marked. The PCN egress node takes the measured rates of ET-marked traffic per IEA as a direct estimate of the termination rate TR. In case of packet loss, the termination rate TR is underestimated and several termination steps are needed. Preferential dropping of unmarked packets mitigates this problem.

b) MRT-DTR with Marked AR-Overload: When the reference rate of the excess marker is set to the admissible rate, *AR*-overload is marked. The PCN egress node measures the rates of AS-marked and non-AS-marked traffic ($ASR, nASR$) and calculates the termination rate by $TR = nASR + ASR - u \cdot nASR = ASR - (u-1) \cdot nASR$. The termination rate TR is overestimated when $nASR$ is underestimated. To avoid overtermination in case of packet loss, preferential dropping of marked packets is needed.

2) MRT with Edge-to-Edge Supportable Rates (MRT-ESR): MRT-ESR calculates an estimate of the edge-to-edge supportable rate ESR and signals it to the FT entity. It terminates an appropriate set of flows from the IEA so that the overall rate of the remaining flows is ESR. Traffic must be terminated only if the PCN egress node has detected *SR*-pre-congestion which needs to be signalled explicitly. To avoid overtermination, ESR should not be underestimated. A minimum inter-termination time between consecutive termination actions is not required. The advantage of MRT-ESR compared to MRT-DTR is that a single termination step suffices to remove overload even in case of severe packet loss.

a) MRT-ESR with Marked SR-Overload: The PCN egress node takes the measured rates of non-ET-marked traffic per IEA as a direct estimate of the edge-to-edge supportable rate ESR. Termination is required only if ET-marked packets have been observed. To avoid overtermination in case of packet loss, preferential dropping of marked packets is needed.

b) MRT-ESR with Marked AR-Overload: The PCN egress node measures the rates of AS-marked and non-AS-marked traffic ($ASR, nASR$) and calculates the edge-to-edge supportable rate by $ESR = u \cdot nASR$. Traffic must be terminated only if $(nASR + ASR > u \cdot nASR$ holds. To avoid overtermination in case of packet loss, preferential dropping of marked packets is needed.

3) MRT with Indirectly Measured Termination Rates (MRT-ITR): With MRT-ITR, the PCN egress node provides an estimate of the edge-to-edge supportable rate ESR and the PCN ingress node provides an estimate of the ingress rate IR per IEA. The termination rate is calculated as $TR = IR - ESR$. Appropriate signalling is required to convey the information from the PCN ingress and the PCN egress node to the FT entity together with an indication whether termination is required at all. MRT-ITR works with both marked *SR*-overload and marked *AR*-overload. The edge-to-edge supportable rate ESR as well the indication of *SR*-pre-congestion are derived as in Sect. VII-B2a and Sect. VII-B2b, respectively. To avoid overtermination in case of packet loss, preferential dropping of marked packets is required since MRT-ITR to make sure that edge-to-edge supportable rates ESR are correctly measured.

Like MRT-ESR, MRT-ITR accounts for lost PCN traffic. Its disadvantage is that measurement of IR is also required and that the rates IR and eSR must be timely correlated to avoid over- or underestimated termination rates [43].

C. Geometric Flow Termination (GFT)

GFT assumes that the reference rate of threshold marking is set the supportable rate. Furthermore, fractional marking based on the admissible rate is assumed for AC (cf. Sect. VIII-E). Thus, in case of *AR*-pre-congestion, a small fraction of the packets is marked while in case of *SR*-pre-congestion, all packets are marked. As the marking is done with the same codepoint, the PCN egress node computes CLE (cf. Sect. VI-B1) for a specific IEA to differentiate both cases. Hence, when the CLE value is larger than a certain threshold, *SR*-pre-congestion is signalled to the FT entity which terminates a fixed percentage x of the flows of the corresponding IEA. Possibly several and sufficiently spaced termination steps are required to remove the entire *SR*-overload. The PCN rate decreases like $(1-x)^k$ where k is the number of termination steps. This geometric decrease lead to the name GFT. If the termination percentage x is small, the termination process takes long. If x is large, overtermination likely occurs.

D. Marked-Packet Based Flow Termination (MPT)

With MPT, individual marked packet trigger the termination of single flows. As a result, MPT terminates flows successively and the *SR*-overload is gradually reduced which may still be fast. This is different to MRT and GFT which terminate several flows in one shot. MPT terminates only recently marked flows by communicating their flow ID to the FT entity which may be collocated with the PCN egress node. This is an important feature in networks with multipath routing (cf. Sect. VII-A2).

We first present three MPT mechanism that require the reference rates of the marker to be set to the supportable rates

[42]. Then, we present a conversion algorithms that converts marked *AR*-overload into marked *SR*-overload which makes two of the three presented MPT methods applicable in a single marking context.

1) MPT Based on Excess Marking with Marking Frequency Reduction (MPT-MFR): MPT-MFR requires excess marking with MFR and the reference rate of the marker must be set to the supportable rate of the link. A flow is terminated as soon as one of its packets is ET-marked [5]. If every packet exceeding the supportable rate is ET-marked, many flows are terminated within short time so that overtermination occurs. Therefore, MPT-MFR requires that packets are ET-marked less frequently, i.e., the PCN nodes should apply packet size independent excess marking (cf. Sect. IV-A2) with proportional MFR (cf. Sect. IV-B2). Then, only one packet is ET-marked for σ_b bytes that exceed the supportable rate on a link. The parameter σ_b controls the termination speed of MPT-MFR and its proper choice prevents overtermination [42].

2) MPT Based on Plain Excess Marking for Individual Flows (MPT-IF): With MPT-IF, PCN packets are metered and marked by plain excess marking and the reference rate of the marker is set the supportable rate. Also here, packet size independent marking (cf. Sect. IV-A2) is important to achieve termination fairness among flows with small and large packets. The PCN egress node maintains a credit counter for each flow. This counter is reduced by the size of each received marked packet. When the counter is zero or negative, the flow is terminated. The initialization of the credit counter controls the termination speed of MPT-IF in case of *SR*-pre-congestion. The credit counter needs to be set to an appropriate value when the flow is admitted to avoid slow termination or overtermination [42].

3) MPT Based on Plain Excess Marking for IEAs (MPT-IEA): MPT-IEA is a modification of MPT-IF for IEAs and assumes the same marking behavior. The motivation is to choose flows to be terminated from a larger set to support termination policies. The egress node of an IEA maintains a credit counter for that IEA which is reduced by the size of each received ET-marked packet belonging to the IEA [45]. When a packet arrives and the counter is already zero or negative, a recently marked flow f of the IEA is terminated. Then, the credit counter is incremented by the product of that flow's rate R_f and some time constant T_{inc}. The choice of this constant determines the speed of the *SR*-overload reduction, but it should not be too small to avoid overtermination [42].

4) Marking Conversion from AR-Overload to SR-Overload: The two algorithms MPT-IF and MPT-IEA assume excess marking with the reference rate set to the supportable rate. To support single marking, they should also work when the reference rate is set to the admissible rate. In [30] an algorithm was presented that converts and AS-marked stream into an ET-marked stream by unmarking some AS-marked packets. That means marked *AR*-overload is converted into marked *SR*-overload. When preprocessing an AS-marked packet stream with that algorithm, MPT-IF and MPT-IEA can be used as termination method without any modification.

The conversion algorithm is called for each packet arrival and either converts an existing AS-mark into an ET-mark or clears it. The algorithm a counter *Cnt* with maximum value Cnt_{max} and is explained in Algorithm 4. The counter *Cnt* indicates how many AS-marked bytes can be re-marked to unmarked before a next AS-marked packet will not be re-marked. For each non-AS-marked byte, the counter *Cnt* is incremented by $u-1$, but it cannot exceed Cnt_{max}. When a packet arrives AS-marked and if the counter *Cnt* is not negative, the packet is re-marked to unmarked and the counter *Cnt* is reduced by the packet size *B*. Otherwise, the packet remains marked which is then interpreted as ET-mark.

```
Input:  counter $Cnt$, maximum counter size $Cnt_{max}$,
        packet size $B$ and marking $M$
  if $(M == \text{unmarked})$ then
    $Cnt = \min(Cnt_{max}, Cnt + (u-1) \cdot B)$;
  else if $(Cnt \geq 0)$ then    $\{(M == \text{AS})\}$
    $Cnt = Cnt - B$;
    $M = \text{unmarked}$;
  else
    $M = \text{ET}$;
  end if
```

Algorithm 4: MARKING CONVERSION: converts a stream with AS- and non-AS-marked packets into a stream with ET- and non-ET-marked packets.

The conversion algorithm implements packet size independent re-marking as the re-marking decisions are taken independently of the packet size. A sufficiently large maximum Cnt_{max} for the counter is needed to tolerate short-term variations of packet markings, i.e. a burst of S AS-marked bytes should not be ET-marked. However, this tolerance also delays initial re-marking. The performance of MPT based on *AR*-overload using marking conversion was also studied in [30].

E. General Problems of FT Methods

Like overtermination expresses the fact that more traffic than needed is terminated, undertermination means that less traffic is removed than necessary. In case of multipath routing, over- and undertermination possibly occur for IEA-based FT methods (MRT and MPT-IEA). In scenarios with multiple bottlenecks, overtermination occurs for all FT methods. We briefly illustrate these two fundamental problems in the following.

1) Over- and Undertermination due to Multipath Routing: With multipath routing, flows of the same IEA possibly take different paths from the ingress to the egress node of the PCN domain. Fig. 7 shows that these paths can experience different levels of pre-congestion.

MRT and MPT-IEA are IEA-based FT methods. While the termination of only marked flows is an important feature of MPT-IEA, MRT is mostly discussed without this feature. Therefore, we focus in the following on the more specific MRT method. With MRT based on *SR*-overload, the egress node detects *SR*-pre-congestion by received ET-marked packets. Thus, *SR*-overload can be recognized when at least one flow is carried over a *SR*-pre-congested path which triggers FT.

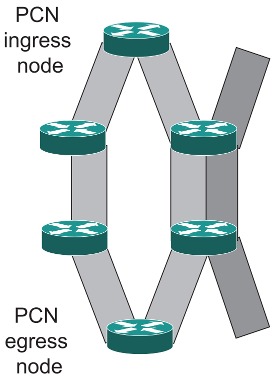

Fig. 7. A multipath can consist of non-pre-congested and *AR*- or *SR*-pre-congested paths. IEA-based FT methods possibly lead to overtermination when they react to marked *SR*-overload. They possibly lead to over- and undertermination when they react to marked *AR*-overload.

FT terminates flows, but possibly also non-marked flows. The termination process continues until enough flows on the *SR*-pre-congested paths are terminated. Several termination steps are required because flows on non-*SR*-pre-congested paths are possibly also terminated. This possibly leads to overtermination. MPT does not suffer from this problem as it terminates flows only if at least one of their packets was ET-marked. This guarantees that only flows of *SR*-pre-congested paths are terminated.

This is different with MRT based on *AR*-overload. Packets are AS-marked so that egress nodes recognize *AR*-pre-congestion when they receive marked packets and only if the fraction of received AS-marked packets is large enough, *SR*-pre-congestion is detected. Thus, if a single path is *SR*-pre-congested and the other paths are not, the egress node possibly cannot detect *SR*-pre-congestion. If the egress node detects *SR*-pre-congestion, admitted flows are removed until *SR*-pre-congestion cannot be recognized anymore, i.e., until the fraction of AS-marked packets is small enough. This may be a case where one path is not pre-congested at all and another path is even *SR*-pre-congested. When flows are removed, flows from non-*SR*-pre-congested paths are possibly also removed. Thus, undertermination may be observed on some paths while overtermination is observed on other paths when the termination process has completed.

With MPT-IF, packet markings are evaluated per flow and so end systems can detect whether a flow runs over an *SR*-pre-congested path. This is different with MPT-IEA when marking conversion is used to cope with marked *AR*-overload. The marking conversion algorithm is applied to the overall traffic. If there is substantial traffic from only lightly pre-congested paths, the conversion algorithm possibly receives too few AS-markings to produce ET-markings so that *SR*-pre-congestion cannot be detected and undertermination occurs. If *SR*-pre-congestion is detected, overtermination can occur although only ET-marked flows are terminated because the ET-markings can result from AS-marked packets carried on *AR*- or *SR*-pre-congested paths.

We briefly consider GFT. On the one hand, *SR*-pre-congestion cannot be detected when the fraction of marked packets is smaller than a certain CLE threshold. Then undertermination occurs. On the other hand, GFT is usually applied with fractional marking based on the admissible rate and threshold marking based on the supportable rate. Then, marked flows were possibly marked due to *AR*-pre-congestion only instead of *SR*-pre-congestion. Hence, the condition that a flow is marked is not a sufficient condition that is carried over an *SR*-pre-congested path.

A detailed study of over- and undertermination due to multipath routing is provided in [43] and [30].

2) Overtermination due to Multiple Bottlenecks: When a link or node fails, flows are possibly rerouted over a backup path and the backup traffic cause simultaneous pre-congestion on several links which we call multiple bottlenecks. We consider the multiple bottleneck scenario in Fig. 8. There are 2, 3, and 4 serial links. Aggregate 0 represents backup traffic and the other aggregates provide cross traffic for each link. We assume that the backup traffic turns all links into *SR*-pre-congestion so that traffic is terminated. This problem has been studied in [31]. The packets of aggregate 0 are marked on all links and, therefore, its percentage of marked packets is larger than after just crossing the most pre-congested link. As a result, too much traffic is terminated and overtermination occurs. This effect of increased marking percentage is so strong, that MRT based on marked *AR*-overload starts terminating already when none of the links is *SR*-pre-congested. The strength of the overtermination depends on the traffic load on the links relative to the supportable rate *SR*, the fraction of backup traffic, the number of pre-congested links, and the parameter *u* which controls $SR = u \cdot AR$ for MRT based on *AR*-overload. For MPT the same phenomenon is observed. Thus, it is common to all known FT methods, but it is significantly stronger when they trigger termination based on *AR*-overload.

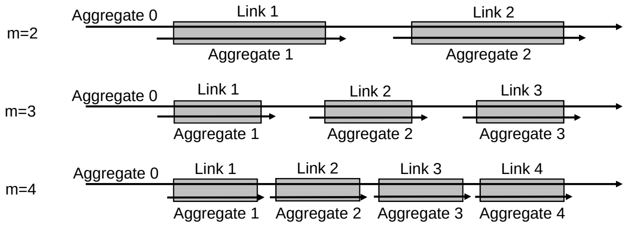

Fig. 8. Multiple bottleneck scnearios: all links are pre-congested, aggregate 0 represents backup traffic while other aggregates provide cross traffic. Overtermination occurs since traffic of aggregate 0 receives more markings than appropriate.

F. Summary and Discussion of FT Methods

We briefly summarize the presented FT methods and compare their behavior under packet loss, their requirements regarding packet drop policies, their behavior with a small number of flows per IEA, and their ability to support multipath routing, termination policies, and end-to-end PCN.

1) Summary of FT-Methods: FT methods work with different marking schemes. The most intuitive marking scheme for FT purposes is excess marking with the reference rate set to the supportable rate as the marked traffic provides an estimate for the *SR*-overload in the absence of traffic loss. It is the base for measured-rate based flow termination (MRT) as well as for marked-packet based flow termination (MPT) for individual flows (MPT-IF) or for IEAs (MPT-IEA). To allow for a single marking that supports both AC and FT, excess marking with the admissible rate as reference rate is required. All MRT methods and MPT for individual flows and IEAs can be adapted for that purpose. MPT with marking frequency reduction (MFR) requires excess marking with MFR with the reference rate set to the supportable rate. Finally, geometric flow termination (GFT) works with threshold marking whose reference rate is set to the supportable rate. MRT and MPT methods cannot work with threshold marking as they need some feedback that is proportional to the *SR*-overload to control the termination rate. Conversely, it does not make sense to use GFT when such information is available as GFT cannot profit from it.

2) Behavior under Packet Loss and Required Packet Drop Policies: GFT terminates a fixed fraction of the admitted traffic. Therefore, its termination speed is independent of the strength of the *SR*-overload. However, the time to reduce the *SR*-overload increases with *SR*-overload regardless whether packets are lost. GFT is used only with threshold marking which marks all packets or none. Therefore, the dropping policy does not impact the termination behavior.

As MPT-MFR uses excess marking with MFR, only a few packets are marked, and every marked packet terminates a flow. If marked packets are lost, the termination process is significantly delayed. If all marked packets are lost, termination does not work anymore. Hence, MPT-MFR benefits from preferential dropping of unmarked packets in case of packet loss and it is broken when all marked packets are lost when their dropping is preferred in case of packet loss (cf. [42]).

MPT-IF and MPT-IEA use excess marking. When marked packets are lost, the per flow or per IEA credit counters are decremented more slowly and the termination process is delayed. Hence, MPT-IF and MPT-IEA benefit from preferential dropping of unmarked packets. Preferential dropping of marked packets can delay the termination process significantly, but it does not break it as long as some marked packets remain. Thus, the difference between supportable rate and link bandwidth must be sufficiently large.

MRT-DTR with *SR*-overload benefits from preferential dropping of non-ET-marked packets in case of packet loss since this maximizes its termination speed. All other MRT methods require preferential dropping of marked packets to avoid overtermination in case of packet loss. MRT-ESR and MRT-ITR terminate very fast even in the presence of large traffic loss.

3) Behavior with a Small Number of Flows per IEA: MRT methods terminate a desired fraction of the traffic. However, if the number of flows is very small like 0-3 flows per IEA, MRT cannot always terminate the exact desired fraction. This can lead to over- or undertermination depending on the strategy [43]. For MRT based on *AR*-overload, significant overtermination can occur even for 10 flows per IEA. Due to fluctuations of the percentage of marked packets in case of *AR*-pre-congestion, increased percentages of marked packets can occur that can be interpreted as termination signals. MPT methods work well even with a small number of flows per IEA as flows are terminated successively one after another and termination stops if the *SR*-pre-congestion is removed [42].

4) Support of Multipath Routing: MPT-MFR and MPT-IF terminate only flows that are carried over *SR*-pre-congested paths even if they react to marked *AR*- or *SR*-overload. With MPT-IEA and all MRT methods, termination decisions can basically be taken at the PCN egress node so that local information about recently marked flows can be respected. However, current proposals choose to have the FT entity collocated with the PCN ingress nodes so that support for multipath routing requires additional signalling. If MPT-IEA and MRT react to marked *SR*-overload, marked flows are always safe candidates for termination. This is different when these FT methods react to *AR*-overload since then under- and overtermination possibly occurs (cf. Sect. VII-E1). GFT alone works well with multipath routing. However, it was designed for scenarios with fractional marking based on the admissible rate, threshold marking based on the supportable rate, and baseline encoding (cf. Sect. VIII-E). Therefore, marked flows can result from *AR*- or *SR*-pre-congested paths. Under these circumstances, it is not possible to guarantee correct flow termination decisions in networks with multipath routing.

5) Support of Termination Policies: If the FT entity can select flows to be terminated from a larger set, then termination policies can be enforced. This works well for all IEA-based FT methods, i.e. for all MRT methods, for GFT and for MPT-IEA. MPT-MFR and MPT-IF decide only whether a particular flow is terminated. Therefore, termination policies cannot be enforced.

6) Support of End-to-End PCN: End-to-end PCN requires FT mechanism that can decide whether an admitted flow should be terminated when only the packet markings of that flow are given. MRT and GFT are not applicable as they tend to terminate a traffic fraction which is either proportional to the strength of the observed *SR*-overload or fixed. Therefore, they fail when they are applied to individual flows. MPT-IEA basically becomes MPT-IF if applied to individual flows instead to IEAs. Hence, only MPT-IF and MPT-MFR remain for application with end-to-end PCN and work well for that purpose.

VIII. Existing Proposals

Various proposals for PCN-based AC and FT were presented in individual drafts in the PCN WG with different nomenclature. They all implement the edge-to-edge PCN concept. We briefly review their marking as well as their AC and FT methods using the nomenclature presented in this paper. In addition, we highlight their benefits and shortcomings.

A. “Controlled Load” (CL) PCN

An early draft [4] describes a first PCN architecture to support a controlled load service within a single domain. The detailed algorithms are documented in [13]. CL uses threshold marking based on admissible rates and excess marking based on supportable rates. General dual marking is used which

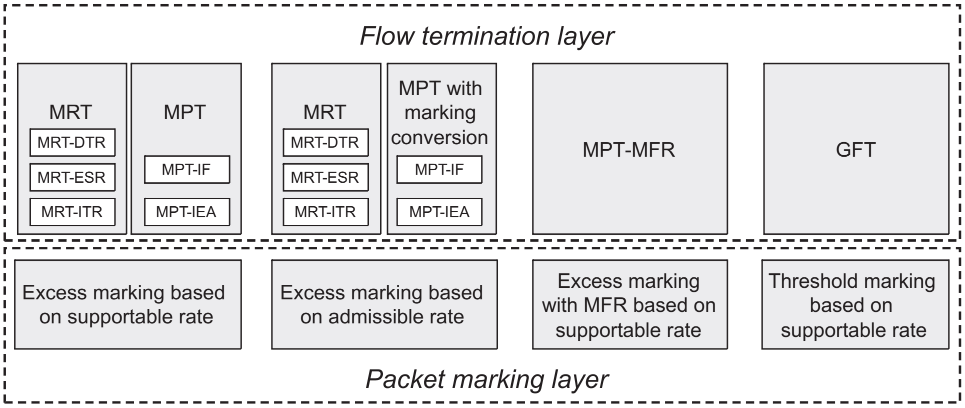
Flow termination layer
MRT
MRT-DTR
MRT-ESR
MRT-ITR
MPT
MPT-IF
MPT-IEA
MRT
MRT-DTR
MRT-ESR
MRT-ITR
MPT with marking conversion
MPT-IF
MPT-IEA
MPT-MFR
GFT
Excess marking based on supportable rate
Excess marking based on admissible rate
Excess marking with MFR based on supportable rate
Threshold marking based on supportable rate
Packet marking layer

Fig. 9. Applicability of FT methods with different marking schemes.

requires two DSCPs. CLEBAC based on ThM- and EcM-marked packets is used for AC (cf. Sect. VI-B1) and MRT-ITR is used for FT (cf. Sect. VII-B3). Therefore, preferential dropping of ThM- and EcM-marked packets is needed to prevent overtermination in case of packet loss.

CL requires two DSCPs for PCN encoding, it cannot block admission requests for empty IEAs, IEABAC and the specific description of MRT-ITR do not work with multipath routing, and MRT in general does not work well with a small number of flows per IEA. However, threshold marking gives clear feedback about *AR*-pre-congestion so that AC works already well for a small number of flows per IEA.

B. *"Single Marking" (SM)*

The SM proposal has been presented in [14] and evaluated in [64], [65]. SM uses excess marking based on admissible rates as a single marking scheme. It uses baseline encoding which requires only a single DSCP. It implements CLEBAC for AC and MRT-ITR based on *AR*-overload for FT (cf. Sect. VII-B3). Therefore, it requires preferential dropping of marked packets to avoid overtermination in case of packet loss.

The benefits of SM are that only a single marking scheme is needed and that only a single DSCP is used. Shortcomings are the fact that CLEBAC cannot block admission requests for empty IEAs, CLEBAC and the specific description of MRT-ITR do not work with multipath routing, and SM requires a large number of flows per IEA that MRT and CLEBAC based on excess marking work well.

C. *"Three State Marking" (3sm)*

3sm has been presented in [5]. It uses threshold marking based on admissible rates and excess marking with MFR based on supportable rates. General dual marking is used which requires two DSCPs. CLEBAC or OBAC is used for AC (cf. Sect. VI-B1, Sect. VI-B2, Sect. VI-A) and explicit or implicit PBAC-IF (cf. Sect. VI-A) may be applied as an alternative. MPT-MFR is used for FT (cf. Sect. VII-D1). Therefore, preferential dropping of non-EcM-marked packets is beneficial for fast termination, but it is not required. However, preferential dropping of EcM-marked packets is detrimental.

Shortcomings of 3sm are the fact that it requires two DSCPs for PCN encoding. When used with probing, AC and FT in 3sm work well with multipath routing and with a small number of flows per IEA. 3sm is able to block admission requests for empty IEAs. Moreover, 3sm can be easily adapted for end-to-end PCN.

D. *"Packet-Specific Dual Marking" (PSDM)*

PSDM has been proposed in [35] and [36]. It uses threshold marking based on admissible rates to possibly re-mark probe packets and excess marking based on supportable rates to possibly re-mark data packets. PSDM encoding is used to mark the packets (cf. Sect. V-B3), which requires the reuse of only a single DSCP. In an early stage, PBAC-IEA can be used as it is easy to implement (cf. Sect. VI-C) which allows to block admission requests even for empty IEAs. In a later stage, explicit and implicit PBAC-IF may be used to cope with multipath routing (cf. Sect. VI-A1 and Sect. VI-A2). Any flow termination method may be used that reacts to marked *SR*-overload. It should be chosen such that multipath routing can be well supported. Preferred packet dropping policies depend on the choice of the FT method.

PSDM requires only a single DSCP, it can work with small numbers of flows per IEA, it can block admission requests for empty IEAs if necessary, and it works well with multipath routing when the enhanced PBAC methods are used. It also supports end-to-end PCN when CFT-IF is used for FT.

E. *"Fractional and Threshold Marking PCN" (FTM-PCN)*

FTM-PCN has been proposed in [57]. It uses fractional marking based on the admissible rate and threshold marking based on the supportable rate for marking purposes. Both marking schemes use baseline encoding so that only a single DSCP needs to be reused for PCN. CLEBAC is used for AC and GFT is used for FT.

The benefit of FTM-PCN is that only a single DSCP is required for PCN marking. Drawbacks are the fact that its AC method does not work well with a small numbers of flows per IEA and it cannot block traffic for empty IEAs. Its FT method is either slow or leads to overtermination. Neither AC nor FT work with multipath routing.

F. “Load Control PCN” (LC-PCN)

In contrast to other proposals, LC-PCN [62] uses rate measurement on PCN links instead of metering algorithms to detect *AR*- and *SR*-pre-congestion. In case of *AR*-pre-congestion, a traffic rate proportional to the *AR*-overload is AS-marked and CLEBAC is used to perform AC. In addition, LC-PCN also supports PBAC-IF. To make it work with a single probe packet in spite of excess marking, probe packets are recognized by the marking algorithm and explicitly AS-marked in case of *AR*-pre-congestion. LC-PCN implements MRT-DTR based on *AR*-overload (cf. Sect. VII-B1b)). To cope better with multipath routing, the marking algorithm is expected to re-mark all non-AS-marked packets to “affected” in case of *SR*-pre-congestion so that the flows to be removed can be chosen from a large set of either AS- or affected-marked flows. LC-PCN optionally AS-marks only a fraction $\frac{1}{N}$ of the *AR*-overload on PCN links, and the PCN egress nodes multiplies the rate of AS-marked packets by N. This marking reduction allows to implicitly track lost excess traffic when non-AS-marked packets are preferentially dropped; however, MRT-DTR-AR requires preferential dropping of AS-marked packets to avoid overtermination. More details are in the draft [62].

LC-PCN works with multipath routing and admission requests can be blocked for empty IEAs when PBAC-IF is used. While AC works well for a small number of flows per IEA when probing is used, FT works not well in that case as MRT is used. The major drawbacks of LC-PCN are its complex marking algorithms and the fact that three codepoints are needed which requires the reuse of two DSCPs.

IX. Summary

In this paper, we have presented a simplified description of pre-congestion notification (PCN) in an edge-to-edge and end-to-end context. We provided compact formulations of various marking behaviors, gave insights into problems and solutions with PCN encoding, and provided an ontology of admission control (AC) and flow termination (FT) algorithms. We discussed how they can be combined with different marking behaviors and different configurations thereof and compared their pros and cons. Existing proposals were summarized in the unified PCN terminology of the paper and their benefits and shortcomings were discussed.

The paper provides an overview of most PCN ideas, it improves their understanding by a streamlined nomenclature, clarifies commonalities and differences of existing approaches, and helps to think in terms of design options rather than in terms of fixed-package proposals which fosters the consensus building process in IETF. The paper preserves the wealth of PCN concepts that will be strongly limited by the standardization process.

TABLE II
List of Abbreviations

Acronym	Meaning
AC	admission control
ACL	admission control layer
AR	admissible rate
AS	admission-stop
ASR	rate of AS-marked traffic
CE	congestion experienced
CL	Controlled Load (proposal)
CLE	congestion level estimate
CLEBAC	CLE-based AC
CL-PSDM	modified CL based on PSDM encoding (proposal)
DS	differentiated services
DSCP	DS codepoint
ECMP	equal-cost multipath
ECN	explicit congestion notification
EcM	excess-traffic marked
EcNM	excess-traffic not-marked
ECT	ECN-capable transport
EhM	exhaustive marked
EhNM	exhaustive not-marked
ESR	edge-to-edge supportable rate
ET	excess traffic
ETR	rate of ET-marked traffic
EWMA	exponentially weighted moving average
EXP	experimental use
FT	flow termination
FTL	flow termination layer
GDM	general dual marking
GDM-LES	GDM with limited ECN support
IEA	ingress-egress aggregate
IEABAC	IEA-based AC
IETF	Internet Engineering Task Force
IR	ingress rate
LC-PCN	Load Control PCN (proposal)
M	marked
MFR	marking frequency reduction
MPT	marked-packet based flow termination
MPT-IF	MPT for individual flows
MPT-IEA	MPT for IEAs
MPT-MFR	MPT with MFR
MRT	measured-rate based flow termination
MRT-DTR	MRT with directly measured termination rates
MRT-ESR	MRT with edge-to-edge supportable rates
MRT-ITR	MRT with indirectly computed termination rates
MRT-X-AR	MRT variant X based on marked *AR*-overload
MRT-X-SR	MRT variant X based on marked *SR*-overload
MTU	maximum transfer unit
nASR	rate of not-AS-marked traffic
$nETR$	rate of not-ET-marked traffic
NM	not marked
OBAC	observation-based AC
PBAC	probe-based AC
PBAC-IEA	probe-based AC for IEAs
PBAC-IF	probe-based AC for individual flows
PCN	pre-congestion notification
PML	packet marking layer
PSDM	packet-specific dual marking
QoS	quality of service
RED	random early detection
RSVP	Resource reSerVation Protocol
SM	Single-Marking (proposal)
SR	supportable rate
TB	token bucket
TR	termination rate
VOICE-ADMIT	name of a standardized DSCP
3sm	Three-State Marking (proposal)

References

[1] W. Almesberger, T. Ferrari, and J.-Y. Le Boudec. SRP: A Scalable Resource Reservation for the Internet. *Computer Communications*, 21(14):1200–1211, Nov. 1998.

[2] L. Andersson and R. Asati. RFC5462: Multiprotocol Label Switching (MPLS) Label Stack Entry: “EXP” Field Renamed to "Traffic Class" Field, Feb. 2009.

[3] I. Ari, B. Hong, E. L. Miller, S. A. Brandt, and D. D. E. Long. Managing Flash Crowds on the Internet. In *International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS)*, Orlando, FL, USA, Oct. 2003.

[4] 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.

[5] J. Babiarz, X.-G. Liu, K. Chan, and M. Menth. Three State PCN Marking. http://tools.ietf.org/id/draft-babiarz-pcn-3sm-01.txt, Nov. 2007.

[6] F. Baker, C. Iturralde, F. Le Faucheur, and B. Davie. RFC3175: Aggregation of RSVP for IPv4 and IPv6 Reservations, Sept. 2001.

[7] F. Baker, J. Polk, and M. Dolly. DSCPs for Capacity-Admitted Traffic. http://www.ietf.org/internet-drafts/draft-ietf-tsvwg-admitted-realtime-dscp-05.txt, Nov. 2008.

[8] 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.

[9] S. Blake, D. L. Black, M. A. Carlson, E. Davies, Z. Wang, and W. Weiss. RFC2475: An Architecture for Differentiated Services, Dec. 1998.

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

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

[12] B. Briscoe. Layered Encapsulation of Congestion Notification. http://tools.ietf.org/id/draft-briscoe-tsvwg-ecn-tunnel-01.txt, Oct. 2008.

[13] B. Briscoe, P. Eardley, D. Songhurst, F. L. Faucheur, A. Charny, J. Babiarz, K. Chan, S. Dudley, G. Karagiannis, A. Bader, and L. Westberg. Pre-Congestion Notification Marking. http://www.ietf.org/internet-drafts/draft-briscoe-tsvwg-cl-phb-03.txt, Oct. 2006.

[14] 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.

[15] X. Chen and J. Heidemann. Flash Crowd Mitigation via Adaptive Admission Control Based on Application-Level Observation. *ACM Transactions on Internet Technology*, 5(3):532–562, Aug. 2005.

[16] P. Cholda, A. Mykkeltveit, B. E. Helvik, O. J. Wittner, and A. Jajszczyk. A Survey of Resilience Differentiation Frameworks in Communication Networks. *IEEE Communications Surveys & Tutorials*, 9(4), 2007.

[17] S. Deering and R. Hinden. RFC2460: Internet Protocol Version 6 (IPv6) Specification, Dec. 1998.

[18] P. Eardley. Traffic Matrix Scenario. http://www.ietf.org/mail-archive/web/pcn/current/msg00831.html, Oct. 2007.

[19] P. Eardley. Marking Behaviour of PCN Nodes. http://tools.ietf.org/id/draft-eardley-pcn-marking-behaviour-02.txt, Mar. 2009.

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

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

[22] S. Floyd and V. Jacobson. Random Early Detection Gateways for Congestion Avoidance. *IEEE/ACM Transactions on Networking*, 1(4):397–413, Aug. 1993.

[23] R. J. Gibbens and F. P. Kelly. Distributed Connection Acceptance Control for a Connectionless Network. In 16th *International Teletraffic Congress (ITC)*, pages 941 – 952, Edinburgh, UK, June 1999.

[24] 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.

[25] V. Jacobson, K. Nichols, and K. Poduri. RFC2598: An Expedited Forwarding PHB, June 1999.

[26] J. Jung, B. Krishnamurthy, and M. Rabinovich. Flash Crowds and Denial of Service Attacks: Characterization and Implications for CDNs and Web Sites. In *International World Wide Web Conference (WWW)*, Honolulu, Hawaii, USA, May 2002.

[27] 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.

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

[29] F. Kelly, P. Key, and S. Zachary. Distributed Admission Control. *IEEE Journal on Selected Areas in Communications*, 18(12):2617–2628, 2000.

[30] F. Lehrieder and M. Menth. Marking Conversion for Pre-Congestion Notification. In *IEEE International Conference on Communications (ICC)*, Dresden, Germany, June 2009.

[31] F. Lehrieder and M. Menth. PCN-Based Flow Termination with Multiple Bottleneck Links. In *IEEE International Conference on Communications (ICC)*, Dresden, Germany, June 2009.

[32] S. R. Lima, P. Carvalho, and V. Freitas. Admission Control in Multiservice IP Networks: Architectural Issues and Trends. *IEEE Communications Magazine*, 45(4):114 – 121, Apr. 2007.

[33] R. Martin, M. Menth, and M. Hemmkeppler. Accuracy and Dynamics of Hash-Based Load Balancing Algorithms for Multipath Internet Routing. In *IEEE International Conference on Broadband Communication, Networks, and Systems (BROADNETS)*, San Jose, CA, USA, Oct. 2006.

[34] M. Menth. *Efficient Admission Control and Routing in Resilient Communication Networks*. PhD thesis, University of Würzburg, Faculty of Computer Science, Am Hubland, July 2004.

[35] M. Menth. Deployment Models for PCN-Based Admission Control and Flow Termination Using Packet-Specific Dual Marking (PSDM). http://tools.ietf.org/id/draft-menth-pcn-psdm-deployment-00.txt, Oct. 2008.

[36] M. Menth, J. Babiarz, and P. Eardley. Pre-Congestion Notification Using Packet-Specific Dual Marking. In *International Workshop on the Network of the Future (Future-Net)*, Dresden, Germany, June 2009.

[37] M. Menth, J. Babiarz, T. Moncaster, and B. Briscoe. PCN Encoding for Packet-Specific Dual Marking (PSDM). http://tools.ietf.org/id/draft-menth-pcn-psdm-encoding-00.txt, July 2008.

[38] M. Menth and M. Hartmann. Threshold Configuration and Routing Optimization for PCN-Based Resilient Admission Control. *accepted for Computer Networks*, 2009.

[39] M. Menth, S. Kopf, J. Charzinski, and K. Schrodi. Resilient Network Admission Control. *Computer Networks*, 52(14):2805–2815, Oct. 2008.

[40] M. Menth and F. Lehrieder. Applicability of PCN-Based Admission Control. In *currently under submission*, 2008.

[41] M. Menth and F. Lehrieder. Comparison of Marking Algorithms for PCN-Based Admission Control. In 14th *GI/ITG Conference on Measuring, Modelling and Evaluation of Computer and Communication Systems (MMB)*, pages 77–91, Dortmund, Germany, Mar. 2008.

[42] M. Menth and F. Lehrieder. PCN-Based Marked Flow Termination. In *currently under submission*, 2008.

[43] M. Menth and F. Lehrieder. PCN-Based Measured Rate Termination. In *currently under submission*, 2008.

[44] 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.

[45] M. Menth, F. Lehrieder, P. Eardley, A. Charny, and J. Babiarz. Edge-Assisted Marked Flow Termination. http://tools.ietf.org/id/draft-menth-pcn-emft-00.txt, Feb. 2008.

[46] M. Menth, R. Martin, and J. Charzinski. Capacity Overprovisioning for Networks with Resilience Requirements. In *ACM SIGCOMM*, Pisa, Italy, Sept. 2006.

[47] T. Moncaster, B. Briscoe, and M. Menth. A Three State Extended PCN Encoding Scheme. http://tools.ietf.org/id/draft-moncaster-pcn-3-state-encoding-00.txt, June 2008.

[48] T. Moncaster, B. Briscoe, and M. Menth. Baseline Encoding and Transport of Pre-Congestion Information. http://tools.ietf.org/id/draft-moncaster-pcn-baseline-encoding-01.txt, June 2008.

[49] 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.

[50] J. Postel. RFC791: Internet Protocol, Aug. 1981.

[51] S. Rai, B. Mukherjee, and O. Deshpande. IP Resilience within an Autonomous System: Current Approaches, Challenges, and Future Directions. *IEEE Communications Magazine*, 43(10):142–149, Oct. 2005.

[52] K. Ramakrishnan and S. Floyd. RFC2481: A Proposal to add Explicit Congestion Notification (ECN) to IP, Jan. 1999.

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

[54] E. Rosen, D. Tappan, G. Fedorkow, Y. Rekhter, D. Farinacci, T. Li, and A. Conta. RFC3032: MPLS Label Stack Encoding, Jan. 2001.

[55] S. Kent and K. Seo. RFC4301: Security Architecture for the Internet Protocol, Dec. 2005.

[56] Z. Sarker and I. Johansson. Usecases and Benefits of end to end ECN support in PCN Domains. http://www.ietf.org/internet-drafts/draft-sarker-pcn-ecn-pcn-usecases-01.txt, May 2008.

[57] D. Satoh, M. Ishizuka, O. Phanachet, and Y. Maeda. Single PCN Threshold Marking by Using PCN Baseline Encoding for Both Admission and Termination Controls. http://tools.ietf.org/id/draft-satoh-pcn-st-marking-01.txt, Mar. 2009.

[58] 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.

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

[60] I. Stoica and H. Zhang. Providing Guaranteed Services without per Flow Management. In *ACM SIGCOMM*, Boston, MA, Sept. 1999.

[61] 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.

[62] L. Westberg, A. Bhargava, A. Bader, G. Karagiannis, and H. Mekkes. LC-PCN: The Load Control PCN Solution. http://tools.ietf.org/id/draft-westberg-pcn-load-control-05.txt, Nov. 2008.

[63] S. Wright. Admission Control in Multi-Service IP Networks: A Tutorial. *IEEE Communications Surveys & Tutorials*, 9(1), 2007.

[64] J. Zhang, A. Charny, V. Liatsos, and F. L. Faucheur. Performance Evaluation of CL-PHB Admission and Pre-emption Algorithms. http://www.ietf.org/internet-drafts/draft-zhang-pcn-performance-evaluation-02.txt, July 2007.

[65] X. Zhang and A. Charny. Performance Evaluation of Pre-Congestion Notification. In *International Workshop on Quality of Service (IWQoS)*, Enschede, The Netherlands, June 2008.