Pre-Congestion Notification (PCN) – New QoS Support for Differentiated Services IP Networks

Michael Menth [†] , University of Tübingen, Dept. of Computer Science, Germany,
Bob Briscoe [‡] , BT Research, UK, and Tina Tsou, Huawei Technologies, China

***Abstract*—Admission control is a well known technique to explicitly admit or block new flows for a domain to keep its traffic load at a moderate level and to guarantee quality of service (QoS) for admitted flows. Flow termination is a new flow control function that terminates some admitted flows when the network capacity does not suffice, e.g., in case of unexpected failures. Admission control and flow termination are useful to protect QoS for inelastic flows that require a minimum bitrate. Examples are realtime applications like voice and video. Pre-congestion notification (PCN) provides for Differentiated Services IP networks feedback about load conditions on their paths to their boundary nodes. This information is used to implement light-weight admission control and flow termination without per-flow states in interior nodes of a domain. These mechanisms are significantly simpler than explicit reservation schemes. We explain the conceptual idea of PCN-based admission control and flow termination, present recent IETF standards, and discuss benefits and limitations.**

I. Introduction

Applications like real-time voice and video conferences are called inelastic because they become unusable below a minimum bit-rate. Even in well-provisioned networks, capacity shortage may occur and cause all affected inelastic applications to fail at once. This may happen due to heavy interest in the contents of one site or due to traffic rerouting and lack of sufficient capacity on backup paths. To avoid this problem, networks often give inelastic traffic priority access to a generously provisioned logical partition of their capacity, using Differentiated Services technology.

Admission control is a well-known flow control function that prevents overload due to unexpectedly high user demands. However, it does not help against overload which occurs due to rerouted traffic in failure cases. Resilient admission control [1] reserves capacity also on backup paths so that admitted traffic can be rerouted without causing overload for a set of protected failures.

In contrast, flow termination is a new control function that quickly reduces overload by removing some admitted flows in exceptional situations. It may be useful after traffic rerouting events or when admission control has admitted too much traffic, e.g., due to flash crowds or unexpected flow behavior. Thus, with admission control and flow termination, networks can be provisioned less generously because overload situations can be avoided or quickly resolved.

Differentiated Services introduce various per-hop behaviors in IP networks to enable prioritized forwarding of appropriately marked high-priority traffic. Integrated Services implement per-hop admission control in IP networks using the Resource reSerVation Protocol (RSVP). RSVP normally requires per-flow state information in every router on a flow's path. As this is rather heavy-weight, operators and manufacturers strive for light-weight admission control that require per-flow state only in border routers of Differentiated Services networks.

The Internet Engineering Task Force (IETF) has defined pre-congestion notification (PCN) [2] for that purpose. Interior nodes of a Differentiated Services domain meter and mark data packets under PCN control to implicitly notify egress nodes whether the rate of high-priority traffic has exceeded certain rate thresholds on some links. The egress nodes provide this feedback to admission control and flow termination decision points, which use it as a base for their decisions. The IETF standardized PCN metering and marking algorithms [3], encoding of PCN information in IP headers [4],information provided from border nodes of a PCN domain to its admission control and flow termination decision points, and the way this information is interpreted by decision points depending on the applied marking model [5], [6].

We first discuss a few related techniques and point out significant differences to PCN. Then we explain the conceptual idea of PCN, illustrate recent PCN standards, and discuss benefits and limitations.

II. Related Approaches

Admission and congestion control is a wide field. We exemplarily discuss three related methods in different technologies and point out the differences to PCN.

Measurement-based admission control has successfully been used for many years in circuit-switched E1/T1 networks. More specifically, ITU-T Q.50 limits the load of voice band traffic, including compressed speech, on 64 kbit/s transmission channels towards international switching centers. In this context, capacity can be shared only among connections with the same source and destination. In contrast, with PCN-based admission control, the capacity of a link can be shared by all ingress-egress aggregates traversing it.

Load-dependent packet marking is used in ATM networks for rate control of connections in the available bit-rate (ABR) traffic contract. Nodes along a virtual path connection mark

[†] This work was partly funded by Deutsche Forschungsgemeinschaft (DFG) under grant TR257/18-2. It reflects only the views of the authors.

[‡] The research leading to these results has received funding from the European's Community Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 248567 for the 'ETICS' Project. Further information is available at www.ict-etics.eu

the Explicit Forward Congestion Indication (EFCI) bit in ATM cells in case of overload. The tail end of a connection monitors these bits and reports feedback in regular resource management cells to the head end of the connection so that the head end can adapt its sending rate. In contrast to PCN, the information is not used for admission control purposes.

PCN combines the two ideas: it uses load-dependent packet marking to support admission control. To be concise, PCN does not perform typical measurement-based admission control because the current traffic rate is not measured. Edge routers of a PCN domain rather measure the fraction of re-marked packets which is used for admission control and flow termination.

Explicit congestion notification (ECN, RFC3168) in IP networks is a technique to record incipient congestion along a path and signal it back to a sender. It is similar to ABR in ATM and can be seen as a precursor of PCN in IP networks. In contrast to PCN, ECN is an end-to-end mechanism, it is applied to elastic traffic, and packets are re-marked only when congestion already occurs. ECN uses the two-bit ECN field in the IP header for implicit signaling of load conditions. Senders of non-ECN-capable flows mark packets with the not-ECN-capable transport codepoint (not-ECT, ‘00’), while senders of ECN-capable flows mark them with one of the two ECN-capable-transport codepoints (ECT(0), ‘10’, ECT(1), ‘01’). If the average occupation of the physical queue of an ECN router exceeds some threshold, it randomly drops not-ECT-packets and re-marks ECT-packets to the congestion-experienced codepoint (CE, ‘11’). When receivers observe CE-marked packets, they signal that information to the sender which then reduces its transmission rate in a similar way as non-ECN-capable flows react when they observe packet loss. The advantage is obvious: retransmissions are not required.

III. Conceptual Idea of PCN

We explain the basic idea of PCN, its application for PCN-based admission control and flow termination, and their use with path-coupled and path-decoupled resource signalling.

A. Pre-Congestion and Pre-Congestion Notification

In Differentiated Services IP networks, PCN traffic basically constitutes a high-priority traffic class with preferential forwarding whose flows are subject to admission control. A link is congested if packets suffer from significant queuing inside a router or if packets are dropped due to buffer overflow. Congestion mostly occurs under high-load conditions. In contrast, pre-congestion occurs on a link l if the PCN traffic rate $r(l)$ exceeds a configured link-specific PCN rate threshold $R(l)$ and burst size. This rate threshold is usually significantly smaller than the link capacity $c(l)$ so that congestion is generally not yet visible in this load regime.

A PCN domain is a network whose edge routers act as PCN ingress and egress nodes and PCN metering and marking is performed for transmission links within the PCN domain. To mark PCN traffic as such, the ingress node of a PCN domain sets an appropriate PCN codepoint in the IP headers of entering PCN packets. PCN-capable nodes of a PCN domain meter the rate of PCN traffic $r(l)$ separately for each outgoing link l and possibly re-mark packets with a different PCN codepoint if the metered rate $r(l)$ exceeds the link-specific PCN rate threshold $R(l)$. When egress nodes receive the PCN traffic, the re-marked PCN codepoints in the IP headers implicitly notify them about pre-congestion in the network. More precisely, a re-marked packet tells an egress node that at least one link is pre-congested on the path the packet has traversed, but it does not reveal the exact link.

B. PCN-Based Admission Control and Flow Termination

PCN flows are subject to admission control and flow termination. The two functions require two different rate thresholds which are specific for every link l in the PCN domain: the admissible rate threshold ($AR(l)$) and the supportable rate threshold ($SR(l)$) whereby $AR(l)$ must be smaller than $SR(l)$. The two thresholds imply three different load regimes which are illustrated in Fig. 1. If the rate $r(l)$ of PCN traffic on link l is below $AR(l)$, the link is not pre-congested and further flows may be admitted for this link. If $r(l)$ is above $AR(l)$, the link is AR-pre-congested and no further flows should be admitted for that link. If $r(l)$ is above $SR(l)$, the link is SR-pre-congested, no further flows should be admitted for that link, and some already admitted flows carried over that link should be terminated to reduce the PCN traffic rate $r(l)$ below $SR(l)$.

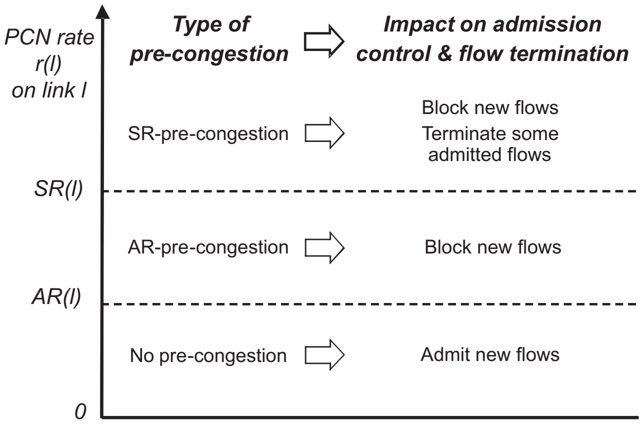
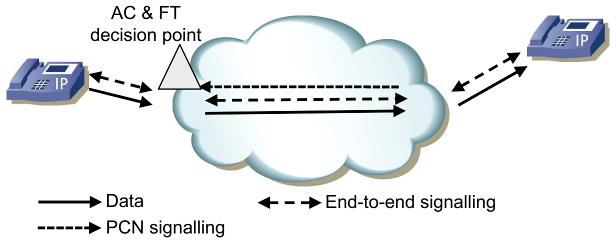


Fig. 1. The admissible and the supportable rate ($AR(l), SR(l)$) define three types of pre-congestion.

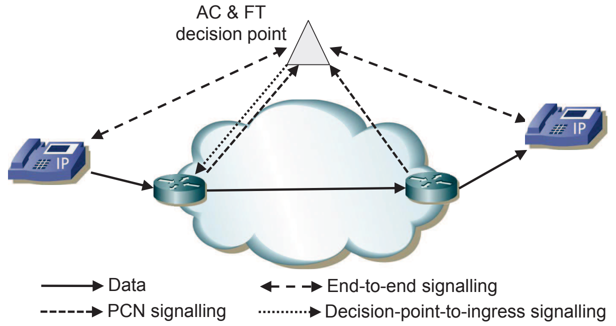
PCN nodes meter the PCN traffic on a link l and re-mark it appropriately if the PCN traffic rate $r(l)$ exceeds the admissible or supportable rate thresholds. The egress nodes monitor the packet markings and are thereby notified about the load conditions inside the PCN domain. This information is provided to the admission control and flow termination decision points in a suitable way so that they can use it to admit or block admission requests for new flows or to terminate admitted flows.

C. PCN-Based QoS – The Big Picture

PCN-based QoS control can be applied in various environments. Currently, its deployment is discussed in the context of the Internet and the IP Multimedia Subsystem (IMS). They use path-coupled and path-decoupled resource signalling, respectively. In both scenarios, some end-to-end protocol (e.g. RSVP or SIP) is needed to request admission for new flows in a domain.



(a) Path-coupled signalling.



(b) Path-decoupled signalling.

Fig. 2. PCN-based admission control and flow termination with path-coupled and path-decoupled resource signalling.

1) PCN-Based QoS for Path-Coupled Resource Signalling: In an Internet context, the ingress node is in charge of admitting or blocking reservation requests for new flows. Thus, each ingress node serves as a decision point for admission control and flow termination and requires for its decisions PCN egress reports from all other egress nodes of the domain. This is depicted in Fig. 2(a). For example, RSVP may be used for resource signalling. Then, only the ingress and egress nodes of a PCN domain process RSVP messages and the ingress node admits or blocks requests for transmission across the domain based on received PCN feedback. After admission or termination of a flow, the ingress node locally reconfigures the policers so that only PCN packets of admitted flows can enter the network.

2) PCN-Based QoS for Path-Decoupled Resource Signalling: IMS is an architectural control framework for delivering multimedia services over multiple IP-based wireline and wireless technologies, e.g., DSL, cable modem, Ethernet, W-CDMA, CDMA2000, GSM, GPRS, or UMTS. The Call Session Control Function (CSCF) is part of the control plane in IMS. It interacts with a policy server (Resource Admission Control Subsystem, RACS) that provides a Resource and Admission Control Function (RACF). The user equipment (UE) issues admission requests for new flows and RACF decides whether to admit them. To that end, RACS requires sufficient information about the prospective paths of new flows and about the resource conditions on these paths. PCN may be used to provide that information. The RACS serves as centralized decision point and takes admission control and flow termination decisions based on received PCN ingress and egress reports (see Fig. 2(b)). When RACF admits or terminates a flow, some additional signalling is needed to reconfigure the policers at the ingress node of the respective flow.

IV. The Standards

We summarize the PCN standards. We explain two different metering and marking algorithms [3] and show how PCN marks are encoded in IP headers [4]. Then we present the two different experimental standards for PCN-based flow control: the “Controlled Load (CL)” edge behavior (CL-PCN) [5] and the “Single Marking (SM)” edge behavior (SM-PCN) [6]. They describe the operation of edge nodes in a PCN domain. CL-PCN works more accurately than SM-PCN but it is more difficult to implement and deploy than SM-PCN. This justifies the definition of two different standards and the market may decide which of them will prevail.

A. Traffic Meters and Markers

Two different metering and marking methods are defined [3]: excess-traffic-marking and threshold-marking. When PCN packets enter a PCN domain, the ingress node marks them with “not-marked” (NM). The excess-traffic-marker possibly re-marks them to “excess-traffic-marked” (ETM) while the threshold-marker possibly re-marks them to “threshold-marked” (ThM).

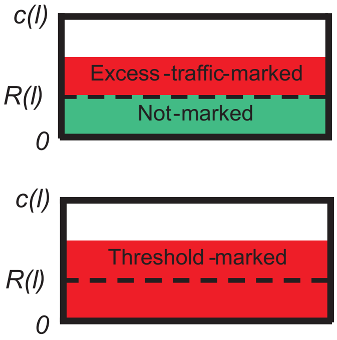


Fig. 3. Excess-traffic-marking re-marks only the PCN traffic exceeding its reference rate $R(l)$ while threshold-marking re-marks all PCN traffic when its reference rate $R(l)$ is exceeded.

We describe how both markers work and explain later in this paper how they are used in the PCN context. Fig. 3 illustrates the effect of both marking algorithms. Both excess-traffic-markers and threshold-markers are configured with a reference rate $R(l)$ which may be either $AR(l)$ or $SR(l)$. The excess-traffic-marker meters only non-excess-traffic-marked PCN traffic. If the metered rate exceeds the reference rate, all metered traffic exceeding that rate is excess-traffic-marked. The threshold-marker meters all PCN traffic. If the input traffic rate of the threshold-marker exceeds its reference rate, all not-marked packets are threshold-marked. Both marking schemes can be easily described by token bucket based algorithms which provide configurable bounds on rate variation so that small traffic bursts do not immediately cause re-marking. While excess-traffic-marking is already available in modern routers for some time, implementations for threshold-marking have only recently been released.

B. Encoding of PCN Marks in IP Headers

As the IP header is already overpopulated, the integration of new codepoints is difficult. Therefore, the two bits of the explicit congestion notification (ECN) field are re-used for that purpose but this redefinition applies only to the so-called PCN-compatible Differentiated Services codepoints (DSCPs).

To support both PCN-based admission control and flow termination in an intuitive way, three different codepoints are needed for PCN traffic: not-marked (NM), excess-traffic-marked (ETM), and threshold-marked (ThM). A fourth codepoint, not-PCN, is used to carry non-PCN traffic with a PCN-compatible DSCP. This encoding is called 3-in-1 and defined in [4]. It is applicable only in PCN domains where all intra-domain tunnels comply with normal mode tunnels defined in RFC6040 [7]. PCN domains using other legacy tunnel types for intra-domain tunnels may use baseline encoding which provides for only two PCN codepoints (NM and either ETM or ThM) and not-PCN, so that only a single marking algorithm can be supported.

With both encoding options, ingress nodes set the ECN field of incoming PCN traffic to NM. To avoid interpretation of PCN marks as ECN marks outside PCN domains, egress nodes reset the ECN field in the IP header of PCN packets to not-ECT before they leave the PCN domain. Appropriate actions, e.g. tunneling, may be taken at the ingress node to support the restoration of the original ECN field at the egress node.

C. The “Controlled Load (CL)” Edge Behavior (CL-PCN)

We describe the required marking behavior for networks implementing CL-PCN, the generation of PCN egress reports by PCN edge nodes, and how admission control and flow termination are supported.

1) Marking Behavior for CL-PCN: With CL-PCN, PCN nodes perform threshold-marking and excess-traffic-marking on each link l of a PCN domain. The reference rate of the threshold-marker is configured with the link-specific admissible rate $AR(l)$ and the reference rate of the excess-traffic-marker is configured with the link-specific supportable rate $SR(l)$. As soon as PCN traffic is carried over an AR- or SR-pre-congested link, all PCN packets are threshold-marked or excess-traffic-marked which is a very clear signal to stop admission of further flows. If PCN traffic is carried over an SR-pre-congested link, some PCN packets are excess-traffic-marked.

CL-PCN requires threshold-marking and 3-in-1 encoding which limits its applicability: so far, there are only a few chip sets that implement threshold-marking and due to 3-in-1 encoding CL-PCN can be deployed only in networks where all tunnel decapsulators comply with RFC6040 [7].

2) Generation of PCN Egress Reports with CL-PCN: Egress nodes evaluate the markings of received PCN packets. They classify them into ingress-egress aggregates and measure the rates of differently marked PCN traffic per ingress-egress aggregate in periodic intervals. Those rates are the rate of not-marked PCN traffic (NMR), the rate of threshold-marked PCN traffic (TMR), and the rate of excess-traffic-marked PCN traffic (EMR) in octets/second. At the end of each measurement interval, the egress node generates PCN egress reports containing the measured rates for different ingress-egress aggregates and sends them to appropriate decision points. We also refer to this information as PCN feedback.

The classification of the packets into ingress-egress aggregates is not trivial and depends on the networking environment. If available, the previous-hop information of the RSVP state in the egress node may be used for that purpose, or tunnels may be used to facilitate the classification.

3) Admission Control with CL-PCN: The decision point keeps an admission control state for each ingress-egress aggregate. This state may be set to *admit* or *block*. In case of *admit*, the decision point admits flow requests and blocks them in case of *block*. The decision point updates the admission control state whenever it receives a new PCN egress report. It calculates a congestion level estimate by

$$CLE = \frac{TMR+EMR}{NMR+TMR+EMR} \qquad (1)$$

whereby NMR, TMR, and EMR are provided by the PCN egress report. If they are all zero, the CLE is also defined zero. If the CLE is smaller than a defined CLE limit $0 < L_{CLE} < 1$, the admission control state is set to *admit*, otherwise it is set to *block*.

With CL-PCN, the admission decisions are not very sensitive to the value of L_{CLE} thanks to threshold-marking. If an ingress-egress aggregate is not carried over a pre-congested link, its CLE is expected to be zero. If it is carried over a pre-congested link, its CLE is expected to be one. This provides a clear signal for admission control.

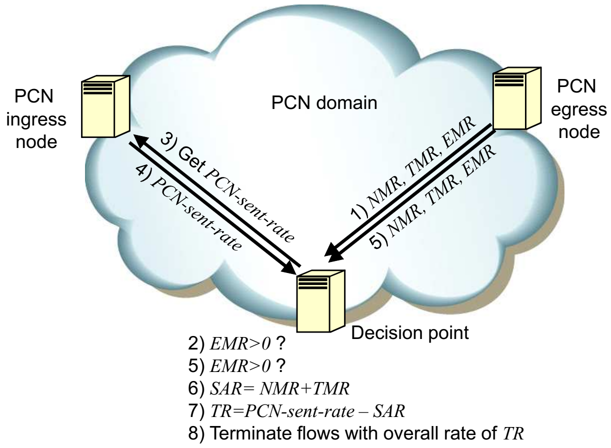


Fig. 4. Flow termination with CL-PCN.

4) Flow Termination with CL-PCN: The flow termination process with CL-PCN is illustrated in Fig. 4. When the decision point receives a PCN egress report with a rate of excess-traffic-marked PCN traffic larger than zero, the respective ingress-egress aggregate was carried over a path on which the PCN traffic rate exceeded the supportable rate on at least one link. Thus, flows need to be terminated. Therefore, the decision point requests the rate of sent PCN traffic (*PCN-sent-rate*) of the respective ingress-egress aggregate from the ingress node. The ingress node then measures the desired *PCN-sent-rate* and responds it to the decision point. When the decision point receives this PCN ingress report, it has already received another PCN egress report. If the rate of excess-traffic-marked PCN traffic EMR is still larger than zero, the decision point calculates the sustainable aggregate rate $SAR = NMR + TMR$ based on the latest PCN egress report. If the PCN traffic rates of all ingress-egress aggregates are reduced to their sustainable aggregate rates, SR-pre-congestion is removed on the bottleneck link. To that end, the decision point computes the termination rate $TR = PCN\text{-}sent\text{-}rate - SAR$ and triggers the termination of an appropriate set of admitted flows with an overall rate of TR. The decision point has access to the traffic descriptors of admitted flows. These descriptors provide upper bounds on the flow rates, but usually overestimate them. The decision point uses them as rate estimates to compose an appropriate flow set for termination, but then the actual rate of the chosen flows is likely to be smaller than TR. Therefore, another termination step may be needed to fully remove SR-pre-congestion on the bottleneck link.

If the decision point underestimates the sustainable aggregate rate SAR, it overestimates the termination rate TR and possibly terminates too much traffic. This may happen if not-marked and threshold-marked packets are dropped due to packet loss on some link. To avoid this source of over-termination, a router should preferably drop threshold-marked and excess-traffic-marked packets in case of packet loss.

D. The "Single Marking (SM)" Edge Behavior (SM-PCN)

We describe the required marking behavior for networks implementing SM-PCN and how admission control and flow termination are supported. PCN egress reports are generated in the same way as for CL-PCN.

1) Marking Behavior with SM-PCN: With single marking, PCN nodes perform only excess-traffic-marking on all links of a PCN domain and configure their reference rates with their admissible rates. The link-specific supportable rate $SR(l)$ depends on admissible rate $AR(l)$ and is determined for each link l by a constant u that is consistent throughout a PCN domain:

$$SR(l) = u \cdot AR(l). \qquad (2)$$

SM-PCN requires only excess-traffic-marking and it may use baseline encoding. Therefore, it can be built with existing hardware and deployed in networks with tunnel decapsulators that do not comply with RFC6040 [7].

2) Admission Control with SM-PCN: Admission control with SM-PCN works as for CL-PCN from an algorithmic point of view. However, admission decisions are more sensitive to the value of the CLE limit L_{CLE}. When PCN traffic exceeds the admissible rate of a link, only a fraction of PCN packets are excess-traffic-marked. Therefore, a small CLE value may already be a sign of serious AR-pre-congestion. To block new admission requests in such situations, the CLE limit L_{CLE} must be set to a small value.

3) Flow Termination with SM-PCN: Flow termination with SM-PCN works similarly as with CL-PCN. However, if the excess-traffic-marked PCN traffic rate EMR is larger than zero, this just indicates AR-pre-congestion for SM-PCN so that termination of traffic is not necessarily required. When packets are re-marked on link l due to pre-congestion, the fraction of not-marked PCN traffic of an ingress-egress aggregate can be approximated by $\frac{NMR}{NMR+EMR} \approx \min\left(1, \frac{AR(l)}{r(l)}\right)$ whereby $r(l)$ is the PCN traffic rate on link l. If PCN traffic is carried over an SR-pre-congested link l, this fraction is smaller than $\frac{AR(l)}{SR(l)} = \frac{1}{u}$. The decision point uses this observation. It detects SR-pre-congestion when $u \cdot NMR < NMR + EMR$ holds and calculates the sustainable aggregates rate by $SAR = u \cdot NMR$. Apart from that, flow termination for SM-PCN works in the same way as for CL-PCN. SM-PCN also requires preferential dropping of excess-traffic-marked traffic to avoid over-termination in case of packet loss.

V. Benefits of PCN-Based Flow Control

PCN-based admission control and flow termination provide multiple benefits for network operators.

A. Simplicity

With PCN-based admission control, decision points take admission decisions on behalf of the entire network and only ingress routers need to know the flow descriptors of admitted flows for policing purposes. Interior nodes of a PCN domain just meter and possibly re-mark PCN traffic without knowing individual flows. This makes PCN nodes simple and scalable compared to routers that keep per-flow reservation states.

B. Robustness

PCN-based admission control is performed only at ingress nodes which makes it robust against link or node failures. In case of a failure, traffic may just be rerouted and admission states of affected flows do not need to be modified. This is different in other admission-controlled systems, e.g., Integrated Services (IntServ), where reservations are bound to specific paths. In IntServ, after a failure, admitted traffic is rerouted and possibly policed on other links due to missing reservations. To restore reservations on backup paths, signalling is needed which is time-consuming and burdens routers with additional load in a critical state. Within that time, the quality of service of admitted flows is severely impacted.

C. Capacity Savings

Networks using PCN-based admission control and flows termination require less capacity compared to networks using conventional admission control. They achieve that through the use of measured feedback for admission control and by the ability to terminate flows under extreme conditions. This is an economic incentive for network operators to use PCN technology.

Conventional admission control uses traffic descriptors for resource allocation. As these traffic descriptors are also used for policing, applications tend to specify generous upper bounds. Therefore, substantial over-reservation of capacity occurs which leads to inefficient use of transmission capacity. To increase efficiency, effective bandwidths may be calculated which require additional assumptions on traffic characteristics and introduce substantial mathematical complexity. This is different with PCN-based admission control: it relies on measured feedback from the network and stops admission of new flows only if the actual rate of admitted PCN traffic exceeds the admissible rate threshold on some link in the network. Thereby it makes better use of transmission capacities than conventional admission control.

To make networks robust against network failures, rerouting mechanisms are used. Moreover, admission control needs to allocate sufficient capacity on all links so that traffic can be carried during normal operation and in most probable failure cases [1]. With PCN-based flow termination, overload after rerouting can be resolved by terminating some admitted flows so that generous provision of backup capacity may not be needed.

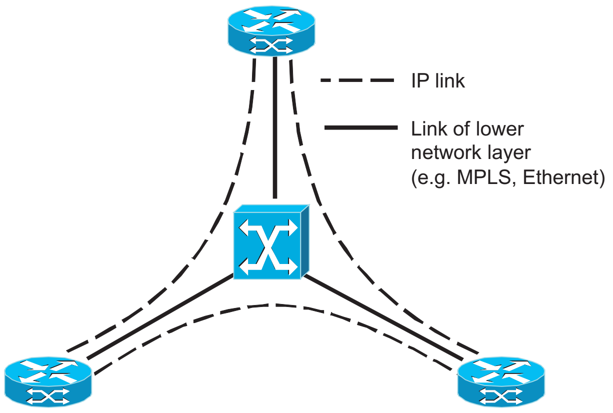


Fig. 5. IP links may share the capacity of the same physical links.

D. Extensibility of PCN to Multiple Networking Layers

Fig. 5 shows virtual IP links which are realized as paths of a lower-layer packet-switched network. Their capacity is not fixed but shared with other IP links. For such links there are no meaningful admissible and supportable rate thresholds. If layer-2 links have fixed capacity, admissible and supportable rates may be assigned to them, and PCN metering and re-marking may be applied to them. Appropriate codepoints for MPLS have already been proposed [4].

Lower layer equipment may just perform metering and marking. The the marking information is propagated to the IP layer upon decapsulation so that admission control and flow termination can still be carried out on the IP layer. This hardly implies changes to the existing proposals. However, it requires that PCN marking information is propagated from lower layers to the IP layer upon decapsulation [8]. This is illustrated in Fig. 6. The encapsulating node copies the marking information from the IP header to the layer-2 header and the decapsulating node copies the marking information from the layer-2 header to the IP header.

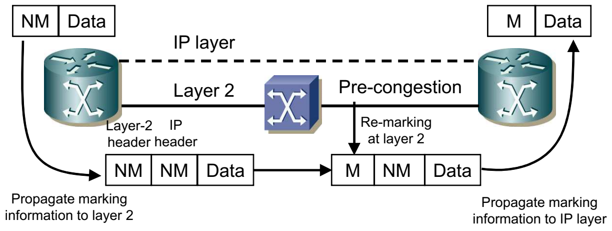


Fig. 6. PCN metering and re-marking may be done at layer 2 and marking information may be propagated from the layer 2 frame header to the IP datagram header upon decapsulation.

E. Support of Termination Priorities

Flow termination is a drastic measure and should be avoided if possible. However, if the restoration of a controlled load condition is needed, it is desirable that low-priority flows are terminated before high-priority flows are removed. The existing architecture for PCN-based flow termination allows decision points to choose a subset of flows from a specific ingress-egress aggregate for termination so that termination priorities can be respected to some extent.

VI. Limitations of PCN-Based Flow Control

PCN-based flow control can be applied only under certain conditions. These limitations are due to the very nature of admission control, to the fact that PCN-based flow control is a feedback-based system, and to the simple implementation of CL-PCN and SM-PCN.

A. General Limitations of Admission Control

Inelastic flows have a maximum bitrate and mostly require a minimum bitrate to work properly. Therefore, they benefit from high-priority transmission which may be combined with admission control for efficiency reasons. Examples are realtime voice and video traffic. Most admission control approaches require the maximum bitrate for policing purposes to verify that flows do not send more traffic than negotiated. PCN-based admission control does not explicitly use this upper bound, but it implicitly needs that admitted flows do not increase their bitrates to arbitrarily high values. This may be controlled by additional policing functions at PCN ingress nodes which this is not subject to standardization of PCN.

B. Limitations of PCN-Based Flow Control

Some shortcomings of PCN technology are due to the fact that it uses measured feedback for control. PCN-based admission control causes over-admission if it admits too many flows and it causes under-admission if it blocks too many flows. In a similar way, PCN-based flow termination causes over-termination if it removes too many flows and it causes under-termination if it removes too few flows. We summarize findings from literature [9], [10], discuss under which conditions over- or under-admission or -termination may happen, and suggest workarounds.

1) Inability to Support Advance Reservations: Consider the transmission of a popular live event starting at 8 pm. Flows may be successfully admitted at 7.50 pm, but transmission starts only at 8 pm. Then over-admission occurs if too many flows have been admitted before 8 pm. To work properly, PCN-based admission control requires that a flow starts sending traffic immediately after admission so that its effect on the network load is visible and can be reflected by PCN feedback. A solution to support advance reservation for PCN-based admission control is the generation of dummy traffic from the instant of admission of a flow until it starts sending real traffic. However, this is not efficient.

2) Susceptibility to Flash Events and Delayed Media: A flash crowd provides an unexpectedly high rate of admission requests. This is problematic for PCN-based admission control because under these conditions many new flows are admitted before the influence of previously admitted flows is reflected in the PCN feedback based on which the new flows are admitted. Therefore, too many flows may be admitted so that over-admission can occur. Similar effects can be observed when flows are admitted but start their data transmission significantly later. In that case, the admitted traffic rate may even oscillate to some extent.

3) Need for Sufficient Link Bandwidth: With PCN-based flow control, new flows are admitted until AR-pre-congestion occurs. This is not problematic if flow rates are small compared to admissible rates since then the relative over-admission is small. If flow rates are rather large, a non-pre-congested link may become even SR-pre-congested through the admission of a single flow so that flow termination is needed. Hence, PCN technology is applicable only for links with sufficient bandwidth. In particular, the difference between the admissible rate and supportable rate of a link must be clearly larger than the largest supported flow rate.

C. Limitations due to Implementation Specifics

Both CL-PCN and SM-PCN rely on PCN feedback measured per ingress-egress aggregate. This keeps operation simple but also causes performance issues in the absence of admitted traffic or in the presence of multipath routing. In addition, SM-PCN may suffer over-admission and over-termination because its PCN feedback is impacted by statistical noise.

1) Over-Admission with Empty Ingress-Egress Aggregates: With CL-PCN and SM-PCN, a new flow is admitted unless the PCN feedback of the ingress-egress aggregate the new flow belongs to indicates pre-congestion. If the ingress-egress aggregate does not carry any flow at that time, pre-congestion cannot be indicated so that the new flow must be accepted. This effectively disables admission control in this situation and may lead to over-admission. On the one hand, this situation seems unrealistic as we already postulated links with sufficient bandwidth. On the other hand, such links may carry the traffic of multiple ingress-egress pairs, each of them having a very small average number of flows. Then, empty ingress-egress aggregates are quite likely.

Probing can solve this problem. At the arrival of a new flow request, probe packets are sent from ingress to egress and the new flow is admitted if all of them arrived without being re-marked. Probing is rather efficient and simple to adopt for CL-PCN but not for SM-PCN. With CL-PCN, probe packets are re-marked in the case of pre-congestion. Therefore, a single probe packet already suffices to detect pre-congestion. This avoids heavy probe traffic and substantial probing delay. If end-to-end signalling messages are re-used for probing (implicit probing), extra probe traffic can be fully avoided.

This is different for SM-PCN: only a fraction of probe packets are re-marked in the case of pre-congestion. Therefore, multiple probe packets are needed to detect pre-congestion with a sufficiently high probability. They generate additional probe traffic and cause long probing delays.

2) Inability to Cope with Multipaths: With multipath routing, flows of a single ingress-egress aggregate are possibly forwarded over different paths between an ingress-egress pair of edge routers. As a consequence, PCN feedback per ingress-egress aggregate is not a reliable load estimate for the prospective path of a new flow. We discuss several problems that are caused by this fact.

a) Under-Admission: If pre-congestion occurs on a link which is part of a multipath between an ingress-egress pair, PCN feedback triggers the decision point to stop admission of new flows. New flows that would be carried only over non-pre-congested links of the multipath are blocked in the same way as those whose prospective paths include pre-congested links. Therefore, less flows are admitted than possible so that under-admission occurs. Probing can overcome this problem.

b) Over-Termination: If SR-pre-congestion occurs on a link which is part of a multipath between an ingress-egress pair of edge routers, then PCN feedback triggers the decision point to terminate admitted flows. As the decision point does not know which flows are carried over the SR-pre-congested link, it might terminate also other flows until SR-pre-congestion stops. This causes over-termination.

With CL-PCN, this shortcoming can be easily repaired. The egress node communicates to the decision point the set of flows for which it has recently observed excess-traffic-marked packets. These flows are carried over an SR-pre-congested link and are safe candidates for termination.

With SM-PCN, excess-traffic-marked packets are a sign of general pre-congestion and not specifically of SR-pre-congestion. Therefore, terminating only flows with recently excess-traffic-marked packets may reduce over-termination, but it cannot safely avoid it.

c) Under-Termination: Another problem occurs only with SM-PCN. It terminates flows only if the fraction of excess-traffic-marked packets is sufficiently large. If one link of a multipath is SR-pre-congested but others are not, the fraction of re-marked packets received by the egress node may not be large enough for the decision point to trigger termination.

3) Additional Problems with SM-PCN: SM-PCN suffers additional over-admission and over-termination because it lacks clear signals for AR- and SR-pre-congestion if the number of packets per measurement interval is small. Hence, SM-

PCN can be applied only for large ingress-egress aggregates.

a) Over-Admission: With SM-PCN, only a small fraction of packets are re-marked in the case of AR-pre-congestion. If the PCN traffic rate on a link exceeds its admissible rate by 5%, only 1 out of 20 packets is excess-traffic-marked. As a consequence, AR-pre-congestion is difficult to detect. If the egress node receives only 10 or 20 packets per measurement interval, the probability that none of these packets are marked is 60% or 35%. Therefore, the decision point cannot reliably block new flows and over-admission occurs.

b) Over-Termination: With SM-PCN, excess-traffic-marked packets can be a sign of light or severe pre-congestion. The decision point uses the fraction of excess-traffic-marked PCN traffic to detect whether flows must be removed. This fraction is subject to statistical fluctuations and can well exceed the threshold $\frac{1}{u}$ in the absence of SR-pre-congestion so that flows are terminated although not needed. The resulting over-termination can be significant even for 500 packets per measurement interval.

VII. Conclusion

Pre-congestion notification (PCN) implements admission control and flow termination for Differentiated IP networks to support quality of service for realtime traffic like voice and video. While admission control is a well-known flow control method, flow termination is new. The combination of both methods allows economic provisioning of transport networks. PCN-based admission control avoids overload caused by increased user activity in such a way that an admissible rate can effectively be reached on a bottleneck link. This facilitates efficient multiplexing of flows with overestimated traffic characteristics. Congestion can occur in spite of admission control under extreme conditions, e.g., due to rerouted traffic after link or node failures. Normally, sufficient capacity needs to be provided for such cases. PCN's flow termination function gives a new perspective on network provisioning: capacity may be deployed less generously since congestion can be resolved by removing admitted flows.

Important features of PCN technology are simplicity, extensibility, and universal applicability. Interior nodes of a PCN domain are unaware of any flows which is important for scalability reasons. PCN technology can easily be extended to lower network layers which makes it applicable to multi-layer networks such as IP-over-MPLS or IP-over-Ethernet. Moreover, it may support network architectures with path-coupled and path-decoupled resource signalling. PCN technology comes with new and attractive features, but is not perfect. Performance results have shown that it requires sufficient traffic aggregation to work as desired. Two different standards exist: CL-PCN and SM-PCN. While SM-PCN can be built with existing hardware, CL-PCN raises fewer performance issues and is, therefore, applicable to a wider range of networking scenarios.

References

[1] M. Menth, R. Martin, and J. Charzinski, "Capacity Overprovisioning for Networks with Resilience Requirements," in *ACM SIGCOMM*, Pisa, Italy, Sep. 2006.

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

[3] ——, "RFC5670: Metering and Marking Behaviour of PCN Nodes," Nov. 2009.

[4] B. Briscoe, T. Moncaster, and M. Menth, "Encoding 3 PCN-States in the IP Header Using a Single DSCP," http://tools.ietf.org/html/draft-ietf-pcn-3-in-1-encoding-04.txt, Jan. 2011.

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

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

[7] B. Briscoe, "RFC6040: Tunnelling of Explicit Congestion Notification," Nov. 2010.

[8] ——, "Guidelines for Adding Congestion Notification to Protocols that Encapsulate IP," http://tools.ietf.org/html/draft-briscoe-tsvwg-ecn-encap-guidelines, Mar. 2011.

[9] M. Menth and F. Lehrieder, "Applicability of PCN-Based Admission Control," University of Würzburg, Institute of Computer Science, Technical Report, No. 468, Mar. 2010.

[10] ——, "PCN-Based Measured Rate Termination," *Computer Networks*, vol. 54, no. 13, pp. 2099 – 2116, Sep. 2010.