Congestion and Pre-Congestion
Notification Working Group
Internet-Draft
Intended status: Informational
Expires: December 22, 2007

P. Eardley
BT
J. Babiarz
K. Chan
Nortel
A. Charny
Cisco Systems
R. Geib
T-Systems
G. Karagiannis
University of Twente
M. Menth
University of Wurzburg
T. Tsou
Huawei Technologies
June 20, 2007

Pre-Congestion Notification Architecture
draft-eardley-pcn-architecture-00

Status of this Memo

Copyright Notice

Abstract

The purpose of this document is to describe a general architecture for flow admission and termination based on aggregated (pre-) congestion information in order to protect the quality of service of established inelastic flows within a single DiffServ domain.

Status

Table of Contents

1. Introduction

The purpose of this document is to describe a general architecture for flow admission and termination based on aggregated (pre-) congestion information in order to protect the quality of service of flows within a DiffServ domain, RFC 2475 [13]. This document defines an architecture for implementing two mechanisms to protect the quality of service of established inelastic flows within a single DiffServ domain, where all boundary and interior nodes are PCN-enabled and trust each other for correct PCN operation. Flow admission control determines whether a new flow should be admitted and protects the QoS of existing PCN-flows in normal circumstances, by avoiding congestion occurring. However, in abnormal circumstances, for instance a disaster affecting multiple nodes and causing traffic re-routes, then the QoS on existing PCN-flows may degrade even though care was exercised when admitting those flows before those circumstances. Therefore we also propose a mechanism for flow termination, which removes enough traffic in order to protect the QoS of the remaining PCN-flows. As a fundamental building block to enable these two mechanisms, PCN-interior-nodes generate, encode and transport pre-congestion (and congestion) information towards the PCN-egress-nodes. Each link of the PCN-domain can be associated with a configured-admissible-rate and a configured-termination-rate. If PCN-traffic, that is traffic in the DiffServ class(es) subject to the PCN mechanisms, on the link exceeds these rates then PCN-packets are admission-marked or termination-marked. Another document will specify the algorithms that determine how and when a number of PCN-packets are marked, and how the markings are encoded in packet headers. PCN-egress-nodes make measurements of the packet markings and send information as necessary to the nodes that make the decision about which PCN-flows to accept/reject or terminate, based on this information. Another document will describe decision-making algorithms. Depending on the deployment scenario, the decision-making functionality could reside at the PCN-ingress-nodes or PCN-egress-nodes or at some central control node in the PCN-domain. We believe that the key benefits of the PCN mechanisms described in this document are that they are simple, scalable, and robust because:

- Per flow state is only required at the PCN-ingress-nodes - for policing purposes (to prevent non-admitted PCN traffic from entering the PCN-domain) and so on. It is not generally required that other network entities are aware of individual flows (although they may be in particular deployment scenarios).

- For each of its links a PCN-node implements either admission-marking or termination-marking behaviours or both. These markers operate on the overall PCN-traffic on the link.

- The information of these measurements is signalled to the PCN-egress-nodes by the PCN-marks in the packet headers. No additional signalling protocol is required for transporting the PCN-marks.

- The PCN-egress-nodes make separate measurements, operating on the overall PCN-traffic, for each PCN-ingress-node (ie not per flow).

- Signalling of PCN-feedback-information, from PCN-egress-node to PCN-ingress-node, is required between all pairs of PCN-boundary-nodes that have admitted PCN-flows (or prospective PCN-flows) between them. The signalled information is on the basis of the corresponding ingress-egress-aggregate.

- The configured-admissible-rates can be chosen small enough that admitted traffic can still be carried after a rerouting in most failure cases. This is an important feature as QoS violations in core networks due to link failures are more likely than QoS violations due to increased traffic volume.

- The admitted PCN-load is controlled dynamically. Therefore it adapts as the traffic matrix changes, and also if the network topology changes (eg after a link failure). Hence an operator can be less conservative when deploying network capacity, and less accurate in their prediction of the PCN-traffic matrix.

- The termination mechanism complements admission control. It allows the network to recover from sudden unexpected surges of PCN-traffic on some links, thus restoring QoS to the remaining flows. Such scenarios are expected to be rare but not impossible. They can be caused by large network failures that redirect lots of admitted PCN-traffic to other links, or by malfunction of the measurement-based admission control in the presence of admitted flows that send for a while with an atypically low rate and increase their rates in a correlated way.

- The configured-termination-rate is expected to be set above the configured-admissible-rate. It may be set below the maximum rate that PCN-traffic can be transmitted on a link, in order to trigger termination of some PCN-flows before loss of PCN-packets occurs or to keep the maximum PCN-load on a link below a level configured by the operator.

2. Terminology

- PCN-domain: a PCN-capable DiffServ domain; a contiguous set of PCN-enabled DiffServ nodes.

- PCN-boundary-node: a node that connects one PCN-domain to a node either in another PCN-domain or in a non PCN-domain.

- PCN-interior-node: a node in a PCN-domain that is not a PCN-boundary-node.

- PCN-node: a PCN-boundary-node or a PCN-interior-node

- PCN-egress-node: a PCN-boundary-node in its role in handling traffic as it leaves a PCN-domain.

- PCN-ingress-node: a PCN-boundary-node in its role in handling traffic as it enters a PCN-domain.

- PCN-traffic: A PCN-domain carries traffic of different DiffServ classes RFC 4594 [15]. Those using the PCN mechanisms are called PCN-classes and the corresponding packets are PCN-packets. The rate from PCN-traffic is the PCN-rate. The same network may carry traffic using other DiffServ classes.

- Ingress-egress-aggregate: The collection of PCN-packets from all PCN-flows that travel in one direction between a specific pair of PCN-boundary-nodes.

- Configured-admissible-rate: reference rate used by the admission-marking algorithm, which is configured for each link in the PCN-domain. Roughly speaking, if the aggregate rate of PCN-traffic on any link of a path is greater than its configured-admissible-rate, then flow admission control blocks additional PCN-flows onto that path.

- Configured-termination-rate: reference rate used by the termination-marking algorithm, which may be configured for each link in the PCN-domain. Normally it is configured to be less than the maximum rate at which PCN-traffic can be forwarded on the link, so that termination-marking occurs before any significant queuing, ECN-marking or loss of PCN-packets.

- Admission-marking: the marking of PCN-packets by a PCN-node to indicate that the PCN-traffic on a link is above the configured-admissible-rate.

- Termination-marking: the marking of PCN-packets by a PCN-node to indicate that the PCN-traffic on a link is above the configured-termination-rate.

- o PCN-marking: admission-marking and/or termination-marking.

- o ECN-marking: the marking of packets according to RFC 3168, The addition of Explicit Congestion Notification to IP [16].

- o PCN-feedback-information: information signalled by PCN-egress-nodes to PCN-ingress-nodes, which is needed for the flow admission and flow termination mechanisms.

EDITOR'S NOTE: Alternative terms have been suggested:

- o Sustainable rate instead of configured-termination-rate

- o Admission-stop marking instead of admission-marking

- o Excess-traffic marking instead of termination-marking

3. Assumptions and constraints on scope

The PCN WG's charter restricts the initial scope by a set of assumptions. Here we list those assumptions and explain them.

1. these components are deployed in a single DiffServ domain, where all PCN-nodes are PCN-enabled and trust each other for correct PCN-marking and transport

2. all flows handled by these mechanisms are inelastic and constrained to a known peak rate through policing or shaping

3. the number of PCN-flows across any potential aggregation bottleneck is sufficiently large for stateless, statistical mechanisms to be effective

4. PCN-flows may have different precedence, but the applicability of the PCN mechanisms for emergency use (911, GETS, WPS, MLPP, etc.) is out of scope

After completion of the initial phase, the PCN WG may re-charter to consider applying the PCN mechanisms to additional deployment scenarios (operation over concatenated DiffServ domains, PCN-aware application mechanisms, etc.). The WG may also re-charter to develop solutions for scenarios where some of these restrictions are not in place. For example, the WG might consider other response mechanisms that act on (pre-)congestion information, for example flow-rate adaptation by elastic applications (rather than flow admission or termination); and the WG might consider operating PCN over concatenated PCN-domains that don't trust each other, using re-ECN,

I-D.briscoe-tsvwg-re-ecn-border-cheat [8] or similar techniques. The details of these work items are outside the scope of the initial phase, but the WG may consider their requirements in order to design components that are sufficiently general to support such extensions in the future - the working assumption is that the standards developed in the initial phase should not need to be modified to satisfy the solutions for when these restrictions are removed.

3.1. Assumption 1: Trust - Controlled Environment

We assume that the PCN-domain is a controlled environment, i.e. all the nodes in a PCN-domain run PCN and trust each other. There are several reasons for proposing this assumption:

- The PCN-domain has to be encircled by a ring of PCN-boundary-nodes, otherwise PCN-packets could enter the PCN-domain without being subject to admission control, which would potentially destroy the QoS of existing flows.

- Similarly, a PCN-boundary-node has to trust that all the PCN-nodes are doing PCN-marking. A non PCN-node wouldn't be able to alert that it is suffering pre-congestion, which potentially would lead to too many PCN-flows being admitted (or too few being terminated). Worse, a rogue node could perform attacks such as marking all PCN-packets so that no PCN-flows were admitted.

One way of assuring the above two points is that the entire PCN-domain is run by a single operator. Another possibility is that there are several operators but they trust each other to a sufficient level. Please note that this restriction only applies to packets in the traffic class that is subject to the PCN mechanisms.

3.2. Assumption 2: Real-Time Applications

We assume that PCN-packets come from real time applications generating inelastic traffic like voice and video requiring low delay, jitter and packet loss, for example the Controlled Load Service, RFC 2211 [17], and the Telephony service class, RFC 4594 [15]. This assumption is to help focus the effort where it looks like PCN would be most useful, ie the sorts of applications where per flow QoS is a known requirement. For instance, the impact of this assumption would be to guide simulations work.

3.3. Assumption 3: Many Flows and Additional Load

We assume that there are many flows on any bottleneck link in the PCN-domain. Measurement-based admission control assumes that the past is a reasonable reflection of the future: the network conditions

are measured at the time of a new flow request, however the actual network performance must be OK during the call some time later. One issue is that if there are only a few variable rate flows, then the aggregate traffic level may vary a lot, perhaps enough to cause some packets to get dropped. If there are many flows then the aggregate traffic level should be statistically smoothed. How many flows is enough depends on a number of things such as the variation in each flow's rate, the total PCN-rate, and the size of the "safety margin" between the traffic level at which we start admission-marking and at which packets are dropped.

We do not make explicit assumptions on how may PCN-flows are in each ingress-egress-aggregate. Performance evaluation work may clarify whether it is necessary to make any additional assumption on aggregation at the ingress-egress-aggregate level.

3.4. Assumption 4: Emergency use out of scope

The applicability of the PCN mechanisms for emergency use (911, GETS, WPS, MLPP, etc) is out of scope.

4. High-level functional architecture

The high-level approach is to split functionality between:

- PCN-interior-nodes 'inside' the PCN-domain, which monitor their own state of (pre) congestion and mark PCN-packets if appropriate. They are not flow-aware, nor aware of ingress-egress-aggregates.

- PCN-boundary-nodes at the edge of the PCN-domain, which control admission of new PCN-flows and termination of existing PCN-flows, based on information from PCN-interior-node. This information is in the form of the PCN-marked data packets and not signalling messages. PCN-ingress-nodes are flow-aware (required for policing purposes) In several deployment scenarios PCN-egress-nodes will also be flow aware. For example I-D.briscoe-tsvwg-cl-architecture [2]describes a deployment scenario where RSVP messages are processed at both the PCN-ingress-node and PCN-egress-node (but not at any PCN-interior-nodes), and both store associated per flow state.

The aim of this split is to keep the bulk of the network simple, scalable and robust, whilst confining policy, application-level and security interactions to the edge of the PCN-domain. For example the lack of flow awareness means that the PCN-interior-nodes don't care about the flow information associated with the PCN-packets that they carry, nor do the PCN-boundary-nodes care about which PCN-interior-

nodes its flows traverse.

At a high level, flow admission control works as follows. In order to generate information about the current state of the PCN-domain, each PCN-node PCN-marks packets if it is "pre-congested". Exactly how a PCN-node decides if it is "pre-congested" (the algorithm) and exactly how packets are "admission-marked" (the encoding) will be defined in a separate standards-track document, but at a high level it is expected to be as follows:

- the algorithm: a PCN-node meters the amount of PCN-traffic on each one of its outgoing links. The measurement is made as an aggregate of all PCN-packets, and not per flow. If the amount of PCN-traffic is deemed to exceed the configured-admissible-rate, then some PCN-packets are admission-marked. Note that the measurement itself may not be of a rate, for example it could be based on a (virtual) queue.

- the encoding: a PCN-node admission-marks a PCN-packet by setting fields in the header to specific values. It is expected that the ECN and/or DSCP fields will be used.

The PCN-boundary-nodes monitor the PCN-marked packets in order to extract information about the current state of the PCN-domain. Based on this monitoring, a decision is made about whether to admit a prospective new flow. Exactly how the admission control decision is made will be defined in separately (at the moment the intention is that there will be one or more informational-track RFCs), but at a high level it is expected to be as follows:

- the PCN-egress-node measures (possibly as a moving average) the fraction of the PCN-traffic that is PCN-marked. The fraction is measured for a specific ingress-egress-aggregate. If the fraction is below a threshold value then the new flow is admitted.

Note that the configured-admissible-rate is a parameter that can be configured by the operator. It will be set lower than the traffic rate at which the link becomes congested and the node drops packets or ECN-marks them. (Hence, by analogy with ECN we call our mechanism Pre-Congestion Notification.)

Note also that the admission control decision is made for a particular ingress-egress-aggregate. So it is quite possible for a new flow to be admitted between one pair of PCN-boundary-nodes, whilst at the same time another admission request is blocked between a different pair of PCN-boundary-nodes.

At a high level, flow termination control works as follows. Each

PCN-node termination-marks PCN-packets in a similar fashion to above. An obvious approach is for the algorithm to use instead a configured-termination-rate (which is higher than the configured-admissible-rate) and the encoding to use another packet marking; however there is also a proposal to use the same configured-admissible-rate and the same encoding. Several approaches have been proposed to date about how to convert this information into a flow termination decision; at a high level these are as follows:

- One approach measures the rate of unmarked PCN-traffic at the PCN-egress-node, which is the amount of that can actually be supported; and the PCN-ingress-node measures the rate of PCN-traffic that is destined for this specific PCN-egress-node and hence can calculate the excess amount that should be terminated.

- Another approach instead measures the rate of termination-marked PCN-traffic and calculates and selects the flows that should be terminated.

- Another approach terminates any PCN-flow with a termination-marked packet. It needs a different termination-marking algorithm to the first approach, otherwise far too much traffic would be terminated.

- Another approach uses admission-marking to decide not only whether to admit more PCN-flows but also whether any PCN-flows need to be terminated. It assumes that the (implicit) configured-termination-rate on all links is at the same offset from the configured-admissible-rate. This approach measures the rate of unmarked PCN-traffic at a PCN-egress-node. The PCN-ingress-node uses this measurement to compute the implicit configured-termination-rate of the bottleneck link. It then measures the rate of PCN-traffic that is destined for this specific PCN-egress-node and hence can calculate the amount that should be terminated.

Since flow termination is designed for "abnormal" circumstances, it is quite likely that some PCN-nodes are congested and hence packets are being dropped, significantly queued and/or ECN-marked. The flow termination mechanism must bear this in mind. (Hence the WG is called 'Congestion and Pre-Congestion Notification'.)

Note also that the termination control decision is made for a particular ingress-egress-aggregate. So it is quite possible for PCN-flows to be terminated between one pair of PCN-boundary-nodes, whilst at the same time none are terminated between a different pair of PCN-boundary-nodes.

Although designed to work together, flow admission and flow

termination are independent mechanisms, and the use of one does not require or prevent the use of the other (discussed further later).

Information transport: the transport of (pre-) congestion information from a PCN-node to a PCN-egress-node is through PCN-markings in data packet headers, no signalling protocol messaging is needed. However, signalling is needed to transport PCN-feedback-information between the PCN-boundary-nodes, for example to convey the fraction of PCN-marked traffic from a PCN-egress-node to the relevant PCN-ingress-node. Exactly what information needs to be transported will be described in the future PCN WG document(s) about the boundary mechanisms. The signalling could be done by an extension of RSVP or NSIS, for instance; protocol work will be done by the relevant WG, but for example I-D.lefaucheur-rsvp-ecn [9]describes the extensions needed for RSVP.

The following are some high-level points about how PCN works:

- o There needs to be a way for a PCN-node to distinguish PCN-traffic from non PCN-traffic. This is based on the DSCP field and/or ECN field. The PCN mechanisms may be applied to more than one traffic class (which are distinguished by DSCP).

- o There may be traffic that is more important than PCN, perhaps a particular application or an operator's control messages. A PCN-node may dedicate capacity to such traffic or priority schedule it over PCN. In the latter case its traffic needs to contribute to the PCN meters.

- o There will be traffic less important than PCN. For instance best effort or assured forwarding traffic (assuming PCN is being applied to the Expedited forwarding class). It will be scheduled at lower priority than PCN, and use a separate queue or queues. However, a PCN-node may dedicate some capacity to lower priority traffic so that it isn't starved.

- o There may be other traffic with the same priority as PCN-traffic. For instance, Expedited Forwarding sessions that are originated either without capacity admission or with traffic engineering, and EF sessions that are originated using PCN admission control. In I-D.ietf-tsvwg-admitted-realtime-dscp [5] the two traffic classes are called EF and EF-ADMIT. A PCN-node could either use separate queues, or separate policers and a common queue; the draft provides some guidance when each is better, but for instance the latter is preferred when the two traffic classes are carrying the same type of application with the same jitter requirements.

5. Detailed Functional architecture

This section is intended to provide a systematic summary of the new functional architecture in the PCN-domain, which maps to the additional functionality required by the PCN-nodes, in addition to their normal router functions. The section discusses the functionality needed for both flow admission control and flow termination. It is split into:

1. functions needed at PCN-interior-nodes

2. functions needed at PCN-ingress-nodes

3. functions needed at PCN-egress-nodes

4. other functions needed for flow admission control

5. other functions needed for probing (which may be needed sometimes)

6. other functions needed for flow termination control

5.1. PCN-interior-node functions

Each link of the PCN-domain is upgraded with the following functionality:

- Packet classify - decide whether an incoming packet is a PCN-packet or not. Performed by examining the DSCP field and/or ECN field. Another PCN WG document will specify encoding.

- PCN-meter - measure the 'amount of PCN-traffic'. The measurement is made as an aggregate of all PCN-packets, and not per flow.

- PCN-mark - if the 'amount of PCN-traffic' exceeds some configurable level, then PCN-packet(s) are PCN-marked.

Another PCN WG document will specify what the 'amount of PCN-traffic' means, and how it's decided that it "exceeds some level". The same general approach of metering and PCN-marking is performed for both flow admission control and flow termination, however the algorithms and encoding may be different.

These functions are needed for each link of the PCN-region. They are therefore needed on all links of PCN-interior-nodes, and on the links of PCN-boundary-nodes that are internal to the PCN-domain. There may be more than one PCN-meter and marker installed at a given link, eg one for admission and one for termination.

5.2. PCN-ingress-node functions

Each ingress link of the PCN-domain is upgraded with the following functionality:

- Packet classify - decide whether an incoming packet is part of a previously admitted microflow, by using a filter spec (eg DSCP, source and destination addresses and port numbers)

- Police - police, by dropping or re-marking with a non-PCN DSCP, and packets received with a DSCP demanding PCN transport that do not belong to an admitted flow. Similarly, police packets that are part of a previously admitted microflow, to check that the microflow keeps to the rate agreed.

- PCN-colour - set the DSCP field or DSCP and ECN fields to the appropriate value(s) for a PCN-packet. The draft about PCN-encoding will discuss further.

- PCN-meter - make "measurements of PCN-traffic". The measurement(s) is made as an aggregate (ie not per flow) of all PCN-traffic towards a particular PCN-egress-node.

The first two are policing functions, needed to make sure that PCN-packets let into the PCN-domain belong to a flow that's been admitted (and probably also to ensure that the flow doesn't go at a faster rate than allowed by its service level agreement). The filter spec will for example come from the flow request message (outside scope of PCN WG, see I-D.briscoe-tsvwg-cl-architecture [2] for an example using RSVP). PCN-colouring allows the rest of the PCN-domain to recognise PCN-packets.

5.3. PCN-egress-node functions

Each egress link of the PCN-domain is upgraded with the following functionality:

- Packet classify - determine which PCN-ingress-node a PCN-packet has come from.

- PCN-meter - make "measurements of PCN-traffic". The measurement(s) is made as an aggregate (ie not per flow) of all PCN-packets from a particular PCN-ingress-node.

- PCN-colour - for PCN-packets, set the DSCP field or DSCP and ECN fields to the appropriate value(s) for use outside the PCN-domain.

Another PCN WG document, about boundary mechanisms, will describe

what the "measurements of PCN-traffic" are. This depends on whether the measurement is targeted at admission control or flow termination. It also depends on what encoding and PCN-marking algorithms are specified by the PCN WG.

5.4. Admission control functions

Specific admission control functions can be performed at a PCN-boundary-node (PCN-ingress-node or PCN-egress-node) or at a centralised node, but not at normal PCN-interior-nodes. The functions are:

- Make decision about admission - compare the "the required measurements of PCN-traffic" (output of the PCN-egress-node's PCN-meter function) with some reference level, and hence decide whether to admit the potential new PCN-flow. As well as the PCN measurements, the decision takes account of policy and application layer requirements.

- Communicate decision about admission - signal the decision to the node making the admission control request (which may be outside the PCN-region), and to the policer (PCN-ingress-node function)

There are various possibilities for how the functionality can be distributed:

- The decision is made at the PCN-egress-node and signalled to the PCN-ingress-node

- The decision is made at the PCN-ingress-node, which requires that the PCN-egress-node signals to the PCN-ingress-node about the measurement of "the required measurements of PCN-traffic"

- The decision is made at a centralised node, which requires that the PCN-egress-node signals to the centralised node about "the required measurements of PCN-traffic" (and that the centralised node learns about policy and application layer requirements), and that the centralised node signals to the PCN-ingress-node about the decision about admission control. It would be possible for the centralised node to be one of the PCN-boundary-nodes, when clearly the signalling would sometimes be replaced by a message internal to the node.

5.5. Probing functions

Probing functions are optional. Admission control, as described in the previous section, is a measurement-based decision. Therefore it is possible that there may be insufficient traffic for a PCN-egress-

node to accurately make the "required measurements of PCN-traffic". Then it requests that the PCN-ingress-node generates probe traffic. Probe packets may be simple data addressed to the PCN-egress-node and require no protocol standardisation, although there will be best practice for their number, size and rate. The functions are:

- Make decision that probing is needed - the PCN-egress-node decides that probe traffic is needed
- Communicate request that probing is needed - the PCN-egress-node signals to the PCN-ingress-node that probe traffic is needed
- Generator of probe traffic - the PCN-ingress-node generates the probe traffic

5.6. Flow termination functions

Specific termination control functions can be performed at a PCN-boundary-node (PCN-ingress-node or PCN-egress-node) or at a centralised node, but not at normal PCN-interior-nodes. The functions are:

- Make decision about flow termination - use the "measurements of PCN-traffic" to decide which PCN-flow or PCN-flows to terminate
- Communicate decision about flow termination - signal the decision to the node that is able to terminate the flow (which may be outside the PCN-region), and to the policer (PCN-ingress-node function)
- (possibly) PCN-meter - make required measurements of PCN-traffic. The measurement(s) is made as an aggregate (ie not per flow) of all PCN-packets being sent towards a particular PCN-egress-node.

There are various possibilities for how the functionality can be distributed, similar to those discussed above in the Admission control section.

6. Design goals and challenges

Prior work on PCN and similar mechanisms has thrown up a number of considerations about PCN's design goals (things PCN should be good at) and some issues that have been hard to solve in a fully satisfactory manner. Taken as a whole it represents a list of trade-offs (it's unlikely that they can all be 100% achieved) and perhaps as evaluation criteria to help an operator (or the IETF) decide between options.

I-D.chan-pcn-problem-statement [10] considers the following as key design goals, ie why PCN is interesting:

- The PCN-enabled packet forwarding network should be simple, scalable and robust
- Compatibility with other traffic (i.e. a proposed solution should work well when non-PCN traffic is also present in the network)
- Support of different types of real-time traffic (eg should work well with CBR and VBR voice and video sources)
- Reaction time of the mechanisms should be commensurate with the desired application-level requirements (e.g. a termination mechanism needs to terminate flows before significant QoS issues are experienced by all real-time traffic, and before a user hangs up)
- Compatibility with different precedence levels of real-time applications (e.g. preferential treatment of higher precedence calls over lower precedence calls, ITU-MLPP [20].

The following are open issues. They are taken from I-D.briscoe-tsvwg-cl-architecture [2] which also describes some possible solutions (potential solutions are out of scope for this document). Note that some may be considered unimportant in general or in specific deployment scenarios.

- ECMP (Equal Cost Multi-Path) Routing: The level of pre-congestion is measured on a specific ingress-egress-aggregate. However, if the PCN-domain runs ECMP, then traffic on this ingress-egress-aggregate may follow several different paths. The problem is that if just one of the paths is pre-congested such that packets are being PCN-marked, then the pre-congestion level as measured by the PCN-egress-node will be diluted by unmarked packets from other non-congested paths. This could lead to a new flow being admitted (because the measured level is below the threshold) but its packets will travel through the pre-congested router (so really it shouldn't be admitted).
- Bi-Directional Sessions: Many applications have bi-directional sessions - hence there are two flows that should be admitted (or terminated) as a pair - for instance a bi-directional voice call only makes sense if flows in both directions are admitted. However, PCN's mechanisms concern admission and termination of a single flow, and coordinating about the decision for both flows is a matter for the signalling protocol and out of scope of PCN. One possible example would use SIP pre-conditions; there are others.

- Global Coordination: PCN makes its admission decision based on PCN-markings on a particular pair of ingress-egress-aggregate. Decisions about flows through a different ingress-egress-aggregate are made independently. However, one can imagine network topologies and traffic matrices where from a global perspective it would be better to make a coordinated decision across all the ingress-egress-aggregates for the whole PCN-domain. For example, to block (or even terminate) flows on one ingress-egress-aggregate so that more important flows through a different ingress-egress-aggregate could be admitted. Mechanisms to solve these problems may well be out of scope.

- Aggregate Traffic Characteristics: Even when the number of flows is stable, the traffic level through the PCN-domain will vary because the sources vary their traffic rates. PCN works best when there's not too much variability in the total traffic level at a node's interface (ie in the aggregate traffic from all sources). Too much variation means that a node may (at one moment) not be doing any PCN-marking and then (at another moment) drop packets because it's overloaded. This makes it hard to tune the admission control scheme to stop admitting new flows at the right time.

- Flash crowds and Speed of Reaction: PCN is a measurement-based mechanism and so has a limited speed of reaction. For example, potentially if a big burst of admission requests occurs in a very short space of time (eg prompted by a televote), they could all get admitted before enough PCN-marks are seen to block new flows. In other words, any additional load offered within the reaction time of the mechanism mustn't move the PCN-domain directly from no congestion to overload.

- Compatibility of PCN-encoding with ECN-encoding, as described in RFC 4774 [12].

7. Deployment scenarios

Operators of networks will want to use the PCN mechanisms in various arrangements, for instance depending on how they are performing admission control outside the PCN-domain (users after all are concerned about QoS end-to-end), their particular goals and assumptions, and so on. Several deployment models are possible:

- IntServ over DiffServ RFC 2998 [18]. The DiffServ region is PCN-enabled, RSVP signalling is used end-to-end and the PCN-domain is a single RSVP hop, ie only the gateways process RSVP messages. Outside the PCN-domain RSVP messages are processed on each hop. This is described in I-D.briscoe-tsvwg-cl-architecture [2]

- Similar to previous bullet but NSIS signalling is used instead of RSVP.

- There are several PCN-domains on the end-to-end path, each operating PCN mechanisms independently. NOTE: A possibility after re-chartering is to consider operating PCN over concatenated DiffServ domains that don't trust each other (ie weakens Assumption 1 about trust)

- RSVP signalling is originated and/or terminated by proxies, with application-layer signalling between the end user and the proxy. For instance SIP signalling with a home hub.

- The PCN-domain extends to the end users. NOTE: This could be considered after re-chartering; it breaks Assumption 3 (aggregation); it doesn't necessarily break Assumption 1 (trust), because in some environments, eg corporate, the end user may have a controlled configuration and so be trusted. This is described in I-D.babiarz-pcn-sip-cap [6].

- Pseudowire: PCN may be used as a congestion avoidance mechanism for end-user deployed pseudowires (collaborate with the PWE3 WG in investigation of this possibility).

- MPLS: RFC 3270 [19], defines how to support the DiffServ architecture in MPLS networks. I-D.ietf-tsvwg-ecn-mpls [7] describes how to add PCN for admission control of microflows into a set of MPLS-TE aggregates (Multi-protocol label switching traffic engineering). PCN-marking is done in MPLS's EXP field.

- Similarly, it may be possible to extend PCN into Ethernet networks, where PCN-marking is done in the Ethernet header.

- The actual decision about admission and termination may be made at the ingress gateway, egress gateway or at some other 'centralised' node, according to the operator's preferences.

8. Operations and Management

EDITOR'S NOTE: The PCN WG Charter says that the architecture document should include security, manageability and operational considerations. Help is requested to achieve this. Also, should the section specifically list a set of OAM parameters that need to be collected?

This Section considers operations and management issues, under the FCAPS headings: OAM of Faults, Configuration, Accounting, Performance

and Security.

8.1. Fault OAM

If a PCN-interior-node fails, then the regular IP routing protocol will re-route round it. If the new route can carry all the admitted traffic, flows will gracefully continue. If instead this causes early warning of pre-congestion on the new route, then admission control based on pre-congestion notification will ensure new flows will not be admitted until enough existing flows have departed. Finally re-routing may result in heavy congestion, when the flow termination mechanism will kick in.

If a PCN-boundary-node fails then we would like the regular QoS signalling protocol to take care of things. As an example I-D.briscoe-tsvwg-cl-architecture [2] considers what happens if RSVP is the QoS signalling protocol. However, such mechanisms are out of scope of PCN.

8.2. Configuration OAM

Perhaps the most important consideration here is that the level of detail of the standardisation affects what can be configured. We would like different implementations and configurations (eg choice of parameters) that are compliant with the PCN standard to work together successfully.

Obvious configuration parameters are the configured-admissible-rate and configured-termination-rate. A higher configured-admissible-rate enables more PCN-traffic to be admitted on a link, hence improving capacity utilisation. A configured-termination-rate set further above the configured-admissible-rate allows greater increases in traffic (whether due to natural fluctuations or some unexpected event) before any flows are terminated, ie to minimise the chances of unnecessarily triggering the termination mechanism. A greater gap between the maximum rate at which PCN-traffic can be forwarded on a link, and the configured-admissible-rate and configured-termination-rate increases the 'safety margin' - which can cover unexpected surges in traffic due to a re-routing event for instance. For instance an operator may want to design their network so that it can cope with a failure of any single PCN-node without terminating any flows. Setting the rates will therefore depend on things like: the operator's requirements, the link's capacity, the typical number of flows and perhaps their traffic characteristics, and so on.

Another configuration decision is whether to operate both the admission control and termination mechanisms. Although we suggest that an operator uses both, this isn't required and some operators

may want to implement only one. For example, an operator could use just admission control, solving heavy congestion (caused by re-routing) by 'just waiting' - as sessions end, existing microflows naturally depart from the system over time, and the admission control mechanism will prevent admission of new microflows that use the affected links. So the PCN-domain will naturally return to normal operation, but with reduced capacity. The drawback of this approach would be that until PCN-flows naturally depart to relieve the congestion, all PCN-flows as well as lower priority services will be adversely affected. On the other hand, an operator could just rely for admission control on statically provisioned capacity per ingress (regardless of the egress of a flow), as is typical in the DiffServ architecture RFC 2475 [13]. Such traffic conditioning agreements can lead to focused overload: many flows happen to focus on a particular link and then all flows through the congested link fail catastrophically. The flow termination mechanism would be used to counteract such a problem.

A different possibility is to configure only the configured-admissible-rate and only do admission-marking. This is suggested in I-D.charny-pcn-single-marking [4] which gives some of the pros and cons of this approach.

Another PCN WG document will specify PCN-marking, in particular how many PCN-packets get PCN-marked according to what measure of PCN-traffic. For instance an algorithm relating current PCN-rate to probability of admission-marking a packet. Depending on how tightly it is decided to specify this, there are potentially quite a few configuration choices, for instance:

- does the probability go from 0% at one PCN-rate (the configured-admissible-rate) to 100% at a slightly higher rate, or 'ramp up' gradually (as in RED)? Does the standard allow both?

- how is the current PCN-rate measured? Rate cannot be measured instantaneously, so how is this smoothed? a sliding window or exponentially weighted moving average?

- is the configured-admissible-rate a fixed parameter? An idea raised in Songhurst [21] is that the configured-admissible-rate on each router should be flexible depending on the current amount of non-PCN-traffic; the aim is that resource allocation reflects the traffic mix - for instance more PCN-traffic could be admitted if the fraction of PCN-traffic was higher. Is this allowed?

Another question is whether there are any configuration parameters that have to be set 'globally' over the whole PCN-domain (as required by some proposals). This may increase operational complexity and the

chances of interoperability problems between kit from different vendors.

8.3. Accounting OAM

Accounting OAM considerations are out of scope of the PCN WG.

8.4. Performance OAM

Performance characteristics that an operator might want to understand could include:

- how quickly do the PCN mechanisms react to a major problem? For example a 'flash crowd' where there could be a big burst of admission requests in a very short space of time. Therefore the reaction time of the admission control mechanism needs to be understood (how long is it before enough admission-marks are seen to block new flows?). This is the 'vulnerability period', and may impact at the application level, for instance QoS requests are not handled any faster than the vulnerability period.

- can the operator identify 'hot spots' in the network (links which most often do PCN-marking)? This would help them plan to install extra capacity where it is most needed.

- what is the rate at which flows are admitted and terminated (for each pair of PCN-boundary-nodes)? Such information would be useful for fault management, networking planning and service level monitoring.

8.5. Security OAM

Security considerations essentially come from Assumption 1, that all nodes in the PCN-domain trust each other. PCN splits functionality between PCN-interior-nodes and PCN-boundary-nodes, and the security considerations are somewhat different for them, mainly because PCN-boundary-nodes are flow-aware and PCN-interior-nodes are not.

- because the PCN-boundary-nodes are flow-aware, they are trusted to use that awareness correctly. The degree of trust required depends on the kinds of decisions it has to make and the kinds of information it needs to make them. For example when the PCN-boundary-node needs to know the contents of the sessions for making the admission and termination decisions (perhaps based on the MLPP precedence), or when the contents are highly classified, then the security requirements for the PCN-boundary-nodes involved will also need to be high.

- The PCN-ingress-nodes police packets to ensure a flow sticks within its agreed limit, and to ensure that only flows which have been admitted contribute PCN-traffic into the PCN-domain. The policer must drop (or perhaps re-mark) any PCN-packets received that are outside this remit. This is similar to the existing IntServ behaviour. Between them the PCN-boundary-nodes must encircle the PCN-domain, otherwise PCN-packets could enter the PCN-domain without being subject to admission control, which would potentially destroy the QoS of existing flows.

- PCN-interior-nodes aren't flow-aware. For example, PCN-packets from normal and higher MLPP precedence sessions aren't distinguishable by PCN-interior-nodes. This prevents an attacker specifically targeting, in the data plane, higher precedence packets (perhaps for DoS or for eavesdropping).

- PCN-marking by the PCN-interior-nodes along the packet forwarding path needs to be trusted, because the PCN-boundary-nodes rely on this information. For instance a non PCN-node wouldn't be able to alert that it's suffering pre-congestion, which potentially would lead to too many PCN-flows being admitted (or too few being terminated). Worse, a rogue node could perform attacks such as PCN-marking all packets so that no flows were admitted.

- the PCN-boundary-nodes should be able to deal with DoS attacks and state exhaustion attacks based on fast changes in per flow signalling.

- The signalling between the PCN-boundary-nodes (and possibly a centralised decision making node) must be protect from attacks. Possible measures include digest authentication, and protection against replay and man-in-the-middle attacks.

9. IANA Considerations

This memo includes no request to IANA.

10. Conclusions

{ToDo:}

11. Acknowledgements

This document is the result of discussions in the PCN WG and forerunner activity in the TSVWG. A number of previous draft were

presented to TSVWG: I-D.chan-pcn-problem-statement [10], I-D.briscoe-tsvwg-cl-architecture [2], I-D.briscoe-tsvwg-cl-phb [3], I-D.charny-pcn-single-marking [4], I-D.babiarz-pcn-sip-cap [6], I-D.lefaucheur-rsvp-ecn [9]. The authors of them were: B, Briscoe, P. Eardley, D. Songhurst, F. Le Faucheur, A. Charny, J. Babiarz, K. Chan, S. Dudley, G. Karagiannis, A. Bader, L. Westberg, J. Zhang, V. Liatsos, X-G. Liu.

12. Comments Solicited

Comments and questions are encouraged and very welcome. They can be addressed to the IETF PCN working group mailing list <pcn@ietf.org>.

13. References

13.1. Normative References

[1] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

13.2. Informative References

[2] Briscoe, B., "An edge-to-edge Deployment Model for Pre-Congestion Notification: Admission Control over a DiffServ Region", draft-briscoe-tsvwg-cl-architecture-04 (work in progress), October 2006.

[3] Briscoe, B., "Pre-Congestion Notification marking", draft-briscoe-tsvwg-cl-phb-03 (work in progress), October 2006.

[4] Charny, A., "Pre-Congestion Notification Using Single Marking for Admission and Pre-emption", draft-charny-pcn-single-marking-01 (work in progress), March 2007.

[5] Baker, F., "DSCPs for Capacity-Admitted Traffic", draft-ietf-tsvwg-admitted-realtime-dscp-01 (work in progress), March 2007.

[6] Babiarz, J., "SIP Controlled Admission and Preemption", draft-babiarz-pcn-sip-cap-00 (work in progress), October 2006.

[7] Davie, B., "Explicit Congestion Marking in MPLS", draft-ietf-tsvwg-ecn-mpls-01 (work in progress), June 2007.

[8] Briscoe, B., "Emulating Border Flow Policing using Re-ECN on

Bulk Data", draft-briscoe-tsvwg-re-ecn-border-cheat-01 (work in progress), June 2006.

[9] Faucheur, F., "RSVP Extensions for Admission Control over Diffserv using Pre-congestion Notification (PCN)", draft-lefaucheur-rsvp-ecn-01 (work in progress), June 2006.

[10] Chan, K., "Pre-Congestion Notification Problem Statement", draft-chan-pcn-problem-statement-01 (work in progress), October 2006.

[11] Chan, K., "Pre-Congestion Notification Encoding Comparison", draft-chan-pcn-encoding-comparison-00 (work in progress), June 2007.

[12] Floyd, S., "Specifying Alternate Semantics for the Explicit Congestion Notification (ECN) Field", BCP 124, RFC 4774, November 2006.

[13] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., and W. Weiss, "An Architecture for Differentiated Services", RFC 2475, December 1998.

[14] Davie, B., Charny, A., Bennet, J., Benson, K., Le Boudec, J., Courtney, W., Davari, S., Firoiu, V., and D. Stiliadis, "An Expedited Forwarding PHB (Per-Hop Behavior)", RFC 3246, March 2002.

[15] Babiarz, J., Chan, K., and F. Baker, "Configuration Guidelines for DiffServ Service Classes", RFC 4594, August 2006.

[16] Ramakrishnan, K., Floyd, S., and D. Black, "The Addition of Explicit Congestion Notification (ECN) to IP", RFC 3168, September 2001.

[17] Wroclawski, J., "Specification of the Controlled-Load Network Element Service", RFC 2211, September 1997.

[18] Bernet, Y., Ford, P., Yavatkar, R., Baker, F., Zhang, L., Speer, M., Braden, R., Davie, B., Wroclawski, J., and E. Felstaine, "A Framework for Integrated Services Operation over Diffserv Networks", RFC 2998, November 2000.

[19] Le Faucheur, F., Wu, L., Davie, B., Davari, S., Vaananen, P., Krishnan, R., Cheval, P., and J. Heinanen, "Multi-Protocol Label Switching (MPLS) Support of Differentiated Services", RFC 3270, May 2002.

[20] "Multilevel Precedence and Pre-emption Service (MLPP)", ITU-T Recommendation I.255.3, 1990.

[21] "Guaranteed QoS Synthesis for Admission Control with Shared Capacity", BT Technical Report TR-CXR9-2006-001, Feburary 2006, < http://www.cs.ucl.ac.uk/staff/B.Briscoe/projects/ipe2eqos/gqs/papers/GQS_shared_tr.pdf>.

Authors' Addresses

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

Email: philip.eardley@bt.com

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

Email: babiarz@nortel.com

Kwok Ho Chan
Nortel
600 Technology Park Drive
Billerica, MA 01821
USA

Email: khchan@nortel.com

Anna Charny
Cisco Systems
14164 Massachusetts Ave
Boxborough, MA 01719
USA

Email: acharny@cisco.com

Ruediger Geib
T-Systems
Deutsche-Telekom-Allee 7
Darmstadt, - 64297
Germany

Email: Ruediger.Geib@t-systems.com

Georgios Karagiannis
University of Twente
P.O. Box 217
7500 AE Enschede,
The Netherlands

Email: g.karagiannis@ewi.utwente.nl

Michael Menth
University of Wurzburg
Institute of Computer Science
Room B206
Am Hubland, Wuerzburg D-97074
Germany

Email: menth@informatik.uni-wuerzburg.de

Tina Tsou
Huawei Technologies
F3-5-089S, R&D Center,
Longgang District
Shenzhen, - 518129
China

Email: tena@huawei.com

Full Copyright Statement

Intellectual Property

Acknowledgment

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