TSVWG
Internet Draft
draft-briscoe-tsvwg-cl-phb-03.txt
Expires: April 2006

B. Briscoe
P. Eardley
D. Songhurst
BT

F. Le Faucheur
A. Charny
V. Liatsos
Cisco Systems, Inc

J. Babiarz
K. Chan
S. Dudley
Nortel

G. Karagiannis
University of Twente / Ericsson

A. Bader
L. Westberg
Ericsson

20 October, 2006

Pre-Congestion Notification marking
draft-briscoe-tsvwg-cl-phb-03.txt

The list of Internet-Draft Shadow Directories can be accessed at
http://www.ietf.org/shadow.html

This Internet-Draft will expire on October 2006.

Abstract

Pre-Congestion Notification (PCN) builds on the concepts of RFC 3168, "The addition of Explicit Congestion Notification to IP". However, Pre-Congestion Notification aims at providing notification before any congestion actually occurs. Pre-Congestion Notification is applied to real-time flows (such as voice, video and multimedia streaming) in DiffServ networks. As described in [CL-DEPLOY], it enables "pre" congestion control through two procedures, flow admission control and flow pre-emption. The draft proposes algorithms that determine when a PCN-enabled router writes Admission Marking and Pre-emption Marking in a packet header, depending on the traffic level. The draft also proposes how to encode these markings. We present simulation results with PCN working in an edge-to-edge scenario using the marking algorithms described. Other marking algorithms will be investigated in the future.

Authors' Note (TO BE DELETED BY THE RFC EDITOR UPON PUBLICATION)

This document is posted as an Internet-Draft with the intention of eventually becoming a STANDARDS track RFC.

Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

Table of Contents

1. Overview

1.1. Introduction

Pre-Congestion Notification builds on the concepts of RFC 3168, "The addition of Explicit Congestion Notification to IP". Pre-Congestion Notification (PCN) is applied to real-time flows (such as voice, video and multimedia streaming) in DiffServ-enabled networks. The reader is referred to [CL-DEPLOY] for description of how PCN enables "pre" congestion control through two procedures, flow admission control and flow pre-emption. Flow admission control determines whether a new microflow is added into the network. Flow pre-emption reduces the current traffic load by terminating selected microflows.

Note this draft concerns the admission control and pre-emption of *flows*, not of packets.

Appendix A provides a brief summary of Explicit Congestion Notification (ECN) [RFC3168]. It specifies that a router sets the ECN field to the Congestion Experienced (CE) value as a warning of incipient congestion. RFC3168 doesn't specify a particular algorithm for setting the CE codepoint, although RED (Random Early Detection) is expected to be used. RFC3168 states that "specifications for Diffserv PHBs [RFC2475] MAY provide more specifics" on the CE marking algorithm. This document can be seen as effectively providing such "specifics" for PHBs (Per Hop Behaviours) targeting real-time services. We imagine future specifications for Diffserv PHBs MAY define their ECN marking algorithm by reference to this document. In particular we imagine a Controlled Load PHB definition would refer to

Expedited Forwarding [RFC3246] for its scheduling behaviour and to this draft for its ECN marking behaviour.

This draft does not propose to change the name of the ECN field. The term PCN is solely used for the marking process. So we say pre-congestion marking is applied to the ECN field (not to the PCN field). We also keep the names of the ECN codepoints, except wherever new codepoint semantics are required. When we talk of PCN-routers, we mean routers arranged so that they will use PCN to mark packets carrying specific, configured DSCPs (differentiated services codepoints). PCN routers may still use default ECN semantics to mark packets carrying other DSCPs.

A router enabled with Pre-Congestion Notification marks packets at a lower traffic level than an ECN-router, when there still isn't any significant build-up of real-time packets in the queue. So PCN-marked packets act as an "early warning" that the rate of packets flowing is getting close to the engineered capacity and hence indicate to the admission control system that requests to admit new real-time flows should be rejected.

In addition to admission control, another essential Quality of Service feature in deployed networks is the ability to cope with failures of routers and links. In this situation the network's capacity is reduced and selected flows may need to be terminated (pre-empted) in order to preserve the quality of service of the remaining real-time flows. Therefore PCN-routers also include the ability to PCN-mark packets to alert that the rate of packets flowing is too close, or exceeding, the engineered capacity and flow pre-emption may be needed.

So a PCN-router needs to be configured with two reference rates:

o configured-admission-rate

o configured-pre-emption-rate

Flow pre-emption should happen at a higher traffic rate than admission control for a number of reasons including:

o End-users are typically more annoyed by their established call dying than by getting a busy tone at call establishment. There may also be regulatory obligations on network operators not to drop established calls.

- A congestion notification based Admission scheme has some inherent inaccuracy because of its reactive, measurement-based nature. For example, sometimes new load may arrive so fast that the admission scheme overshoots before it can measure the effect of new sessions admitted elsewhere. Such anomalous events can usually be absorbed without any disruption, by setting a buffer zone between the configured-admission-rate and configured-pre-emption-rate. No more traffic is admitted until natural flow departures have cleared the buffer zone.

- A buffer zone also allows an operator to decide to admit an 'emergency' or 'Assured Services' call immediately, i.e. without admission control. Similarly to the previous bullet, usually the buffer zone allows the 'emergency' call to be admitted without any disruption to on-going calls. Section 5.4 of [CL-DEPLOY] discusses this option.

If the buffer zone is insufficient then the flow pre-emption mechanism will kick in; however this should very rarely happen.

Both the configured-admission-rate and the configured-pre-emption-rate will be lower than the physical line rate. ([CL-DEPLOY] Section 3.2.2 discusses the case (called implicit pre-emption alerting) where the configured-pre-emption-rate is equal to the line rate.)

Note that admission control is the primary mechanism used to prevent congestion from occurring and flow pre-emption would rarely be invoked under normal conditions; it is a safety mechanism to prevent congestion from persisting after link failures, re-routes, rare over-admission and other similar events.

Together, admission control and flow pre-emption protect the forwarding service offered to admitted and non-pre-empted flows, as well as protecting service to the traffic classes using the remainder of the link capacity.

Note well that a PCN-router does not achieve admission control or flow pre-emption on its own. Just like ECN, a PCN router requires a feedback system in order to control the load causing the congestion it is suffering. [CL-DEPLOY] describes how to achieve an end-to-end controlled load service by using, within a large region of the Internet, DiffServ and edge-to-edge distributed measurement-based admission control and flow pre-emption. Controlled load (CL) service is a quality of service (QoS) closely approximating the QoS that the same flow would receive from a lightly loaded network element [RFC2211]. The edge-to-edge region (which we call the CL-region) is a controlled environment, in that all routers in the CL-region are

enabled with Pre-Congestion Notification and packets can only enter / leave the CL-region through (enhanced) gateways. PCN-marked packets are detected by an egress gateway and associated information is sent to the relevant ingress gateway to decide whether to admit a new flow, or even pre-empt an existing flow. [CL-DEPLOY] also describes a number of assumptions about the CL-region, such as that there are a large number of real-time flows between each pair of gateways; hence the CL-region is typically the backbone of an operator.

We also would like to use PCN-routers in deployment models, such as:

o Where the CL-region spans networks run by different operators.

o End-host to end-host, i.e. a similar architecture to that described in [RTECN]

o A similar architecture to that described in [RMD]

These deployment models are for further study as some of the assumptions made about the CL-region in [CL-DEPLOY] no longer hold. We plan later drafts to describe if and how PCN can work in these frameworks.

This document describes Pre-Congestion Notification:

o (Section 2) The algorithm that determines when a packet is marked so as to warn the admission control mechanism that admission control may be needed.

o (Section 3) The algorithm that determines when a packet is marked so as to warn the pre-emption mechanism that pre-emption may be needed.

o (Section 4 & Appendix B) Simulation results that demonstrate the effectiveness of stateless admission control and flow pre-emption. The results were obtained using the algorithms of Sections 2 and 3. The pdf version of this document includes graphs of simulation results that aren't in the text version.

o (Section 5 & Appendix C) How to encode the markings, i.e. what change to make to which bits of a packet so as to convey the admission marking and pre-emption marking to the admission control and pre-emption mechanisms on the egress gateway.

Sections 2 and 3 describe the algorithms a PCN-enabled router uses to decide whether it needs to admission mark or pre-emption mark a packet. The algorithms are driven by the amount of traffic in the specified real-time service class. Note that the measurement is made on an aggregate basis, i.e. it doesn't distinguish between real-time microflows. Note also that the algorithms run separately for each outgoing link of the PCN router. We present example implementations but the same effect may be implemented in different ways. Indeed, both the admission control and pre-emption algorithms could have been implemented as variants of token buckets, but the former is implemented as a virtual queue, to present an alternative (yet still fairly similar) implementation.

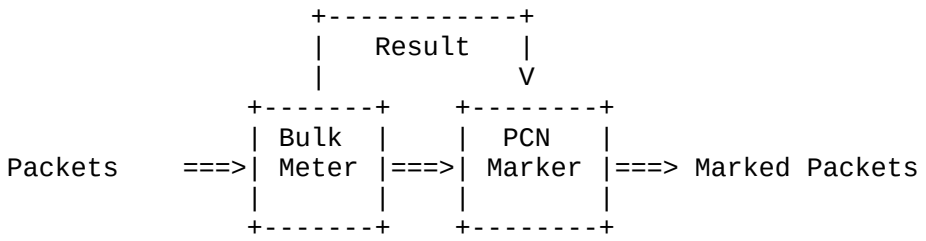


Figure 1: Block Diagram of Meter and Marker Function

Currently this draft documents pre-congestion notification algorithms that we believe are reasonably good, but not necessarily the best. On-going work will consider various alternatives and reach rough consensus on the best.

In Sections 2 and 3 we also hint at how Pre-Congestion Notification can be used within the CL-region, in order to achieve admission control and flow pre-emption "edge-to-edge" across the CL-region. Details are in [CL-DEPLOY].

Section 4 reports some simulation results obtained using these algorithms in the CL-region framework. Note that the aim of our simulations is to demonstrate to the IETF community that these PCN-based admission control and flow pre-emption mechanisms work successfully. It isn't to show that the particular marking algorithms simulated are the optimum ones; although we believe they are a reasonably good choice, on-going work will compare them with various alternatives.

Section 5 presents one possibility for how to encode the markings. Although we believe it is a reasonable choice, there are other possibilities, some of which are listed and discussed in Appendix C. We seek advice and debate as to what scheme should be standardised. Note that the choice of how to encode the markings is non-trivial because we have five things we potentially want to encode, and only have four states in the two bits of the ECN field:

- Admission Marking - the traffic level is such that the router Admission Marks the packet
- Pre-emption Marking - the traffic level is such that the router Pre-emption Marks the packet
- ECT(0) - the first ECT codepoint, for backwards compatibility with the ECN nonce
- ECT(1) - the other ECT codepoint, for backwards compatibility with the ECN nonce
- Not ECT - to indicate to a router that the traffic is not PCN-capable.

1.2. Terminology

- Pre-Congestion Notification (PCN): two new algorithms that determine when a PCN-enabled router Admission Marks and Pre-emption Marks a packet, depending on the traffic level.
- Admission Marking condition- the traffic level is such that the router Admission Marks packets. The router provides an "early warning" that the load is nearing the engineered admission control capacity, before there is any significant build-up in the queue of packets belonging to the specified real-time service class.
- Pre-emption Marking condition- the traffic level is such that the router Pre-emption Marks packets. The router warns explicitly that pre-emption may be needed.
- Configured-admission-rate - the reference rate used by the admission marking algorithm in a PCN-enabled router.
- Configured-pre-emption-rate - the reference rate used by the pre-emption marking algorithm in a PCN-enabled router.

2. Admission Marking algorithm

2.1. Outline

A PCN-enabled router monitors the aggregate traffic in the specified real-time service class. Based on this measurement, the probability that the router admission marks a packet is determined by the algorithm detailed below, configured to use the configured-admission-rate. The algorithm ensures that packets are admission marked before the actual queue builds up, but when it is in danger of doing so soon; the probability increases with the danger. Hence such packets act as an "early warning" that the engineered capacity is nearly reached, and that no more real-time flows should be admitted.

2.2. Virtual queue based algorithm for Admission Marking

In order to make the description more specific we assume a virtual queue is used; other implementations are possible. By a virtual queue we mean a *conceptual* queue - it doesn't store packets, it is just an integer. The integer represents the length of a queue that would exist if the real-time packets were drained at the configured-admission-rate instead of the real scheduling rate for the relevant PHB. Note that there is a virtual queue for each outgoing link and it operates in bulk and not per microflow, i.e. the same virtual queue is used for all the real-time packets on that link. The virtual queue could be implemented, for example, with a variation of a leaky bucket.

The virtual queue is:

- Emptied at the configured-admission-rate, which is slower (perhaps considerably slower) than the link speed and the relevant PHB scheduling rate. This provides a safety margin to minimise the chances of unnecessarily triggering the pre-emption mechanism, for instance.

- Filled when a packet arrives carrying a DSCP that has been configured for PCN (even if the packet is already admission or pre-emption marked). The amount added is the same as the number of octets in the packet.

The procedure is visualised in Figure 2:

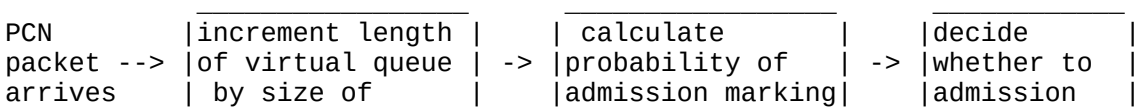


Figure 2: Router action to support admission marking

The router computes the probability that the packet should be admission marked according to the size of the virtual queue, using the following RED-like algorithm:

Size of virtual queue < min-marking-threshold, probability = 0;

min-marking-threshold < Size of virtual queue < max-marking-threshold,

probability =

(Size of virtual queue - min-marking-threshold) / (max-marking-threshold - min-marking-threshold);

Size of virtual queue > max-marking-threshold, probability = 1

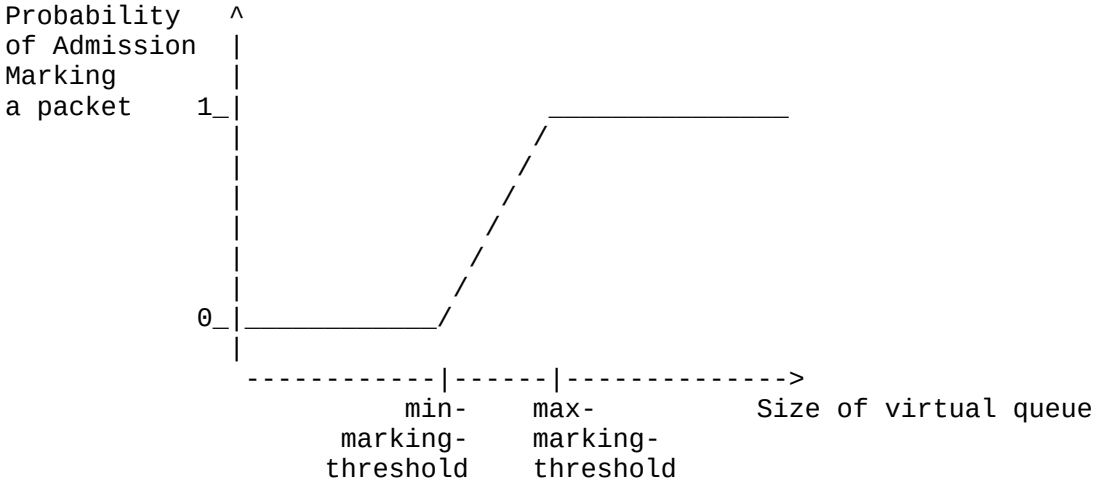


Figure 3: Probability of router admission marking a packet

If the CL traffic is sustained at a level greater than the configured-admission-rate then all packets are eventually admission marked. However, a short burst of traffic at greater than the configured-admission-rate (measured over the burst) may not trigger any admission marking if the burst is sufficiently short that the virtual queue doesn't grow beyond the min-marking-threshold.

A packet that is already pre-emption marked is never re-marked to the admission marked state. The decision whether to admission mark a particular packet is made independently of the decision for the previous packet.

2.3. Admission control within a CL-region using Pre-Congestion Notification

As an example of how the Admission Marking algorithm enables admission control, we briefly consider the edge-to-edge framework described in [CL-DEPLOY]. As real-time packets enter a CL-region, they are re-marked to enable PCN marking using the CL DSCP and the appropriate ECT field. As these CL-packets travel across the edge-to-edge CL-region, routers may admission mark packets, as determined by the algorithm described above. The egress gateway of the region measures the fraction of the real-time traffic that is in the Admission Marked state, with a separate measurement made for traffic from each ingress gateway. It calculates the fraction as an exponentially weighted moving average (which we term Congestion-Level-Estimate, or CLE). When RSVP signalling for a new flow arrives at the egress gateway, it reports the CLE to the CL-region's ingress gateway piggy-backed on the RSVP signalling. The ingress gateway only admits the new real-time microflow if the CLE is less than the CLE-threshold. Hence previously accepted microflows are protected and so suffer minimal queuing delay, jitter and loss.

3. Pre-emption Marking

3.1. Outline

A PCN-enabled router monitors the aggregate traffic in the specified real-time service class. Based on this measurement, when the rate of real-time traffic exceeds the configured-pre-emption-rate for some time, the router will pre-emption mark packets, as determined by the algorithm detailed below. The configured-pre-emption-rate is less than the link speed and less than the relevant PHB scheduling rate, so that Pre-emption Marked packets act as an explicit alert that the engineered capacity is nearly reached, and that some real-time flows may need to be pre-empted. This minimises the chances of a router randomly dropping packets, and hence the Quality of Service of the remaining flows is fully preserved. Also, service is preserved to traffic in other service classes using the remaining capacity.

Pre-emption Marking of packets is similar in motivation to ECN-marking of packets in [RFC3168]. With [RFC3168], feedback of an ECN-marked packet causes the TCP source to halve its effective rate, whereas in our mechanism feedback of pre-emption marking enables an upstream node to terminate real-time flow(s). Pre-emption is therefore more aggressive against selected flows, but the gain is that it enables the full QoS of the remaining flows to be preserved. Note that in [RFC3168] ECN-marking a given packet is intended to result in rate adjustment of the flow to which the packet belongs; while in this draft pre-emption marking a packet simply provides an indication that pre-emption may be needed. As described in [CL-DEPLOY] the pre-emption mechanism will then select particular flows to be pre-empted.

3.2. Token bucket based algorithm for Pre-emption Marking

In order to make the description more specific we assume a token bucket is used; other implementations are possible.

All PCN routers maintain a token bucket per outgoing link:

- Tokens are added at the configured-pre-emption-rate, which is slower than the link speed (and the relevant PHB scheduling rate).

o Usually tokens are removed when a real-time packet arrives; the amount removed is the same as the number of octets in the packet. However, if the real-time packet has already been pre-emption marked, then tokens are not removed. Also, if there are insufficient tokens (because removing them would cause a negative number of tokens in the token bucket), then tokens are not removed and the packet is pre-emption marked. This procedure is visualised in Figure 4.

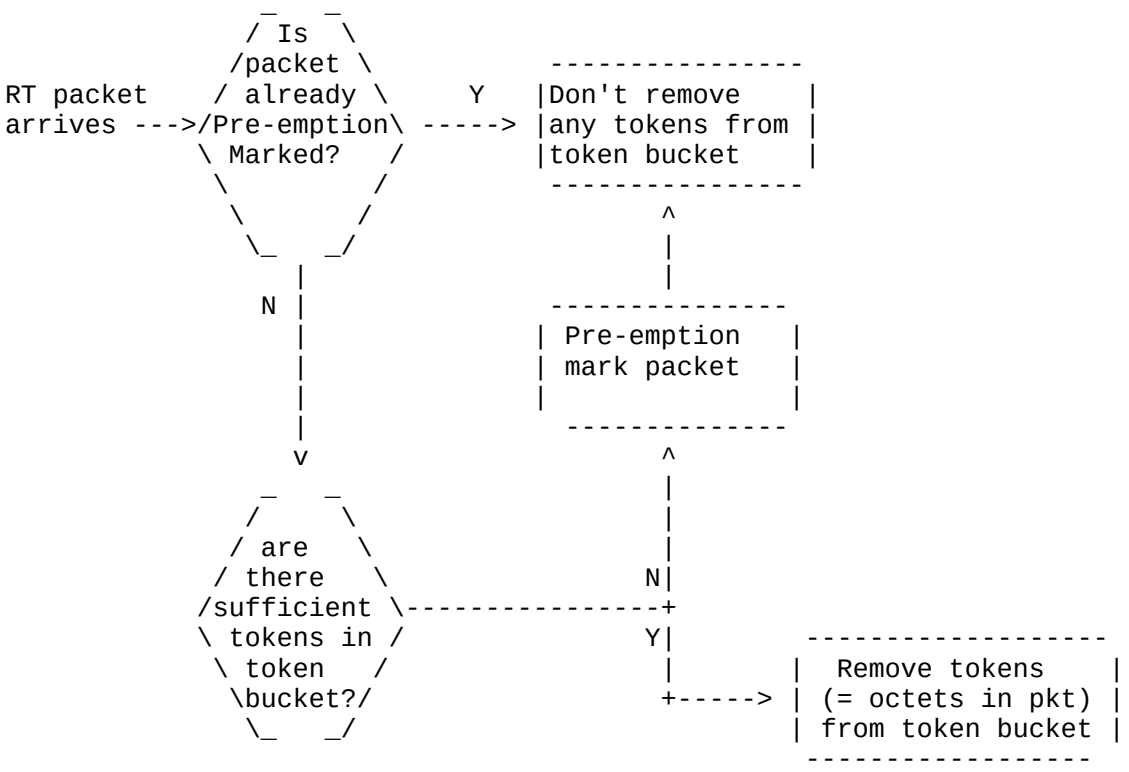


Figure 4: Router action to support pre-emption alerting

So if traffic in the specified real-time service class is sustained at a level greater than the configured-pre-emption-rate then 'non-pre-emption-marked' packet arrivals in excess of this rate are pre-emption marked, but those below it are not marked. ('Non-pre-emption-marked' means 'either unmarked or admission marked'.) The reason is that if a packet finds insufficient tokens, then no tokens are removed from the token bucket, and also the packet is pre-emption marked. Note however that a short burst of traffic at greater than the configured-pre-emption-rate (measured over the burst) may not trigger any pre-emption marking, if the burst is sufficiently short that the token bucket doesn't run out of tokens.

3.3. Flow pre-emption within a CL-region using Pre-Congestion Notification

As an example of how the Pre-emption Marking algorithm enables flow pre-emption, we briefly consider the edge-to-edge deployment model described in [CL-DEPLOY]. As real-time packets travel across the edge-to-edge CL-region, PCN-enabled routers may pre-emption mark packets, as determined by the algorithm described above.

When the egress gateway of the region detects a Pre-emption Marked packet, it measures the rate of real-time traffic *excluding* any packets that are pre-emption marked. Hence it measures the amount of traffic that the network can actually support safely (which we term Sustainable-Aggregate-Rate). The measurement is made for traffic from a particular ingress gateway, and then reported to that ingress gateway. When it receives this message, the ingress gateway measures the ingress-aggregate-rate of real-time traffic that is being sent towards the particular egress gateway. If this measured ingress-aggregate-rate exceeds the Sustainable-Aggregate-Rate, then the ingress gateway pre-empts sufficient number of real-time flow(s) to bring down the ingress-aggregate-rate to (approximately) the Sustainable-Aggregate-Rate.

Different implementations of the rate measurement (and the timescale of this measurement) at the egress and ingress gateways are possible.

4. Simulation results

We have performed an initial set of simulations of admission control and flow pre-emption mechanisms described in this document and consistent with [CL-DEPLOY].

We investigated the performance of the admission control and flow pre-emption mechanisms with traffic modelling CBR voice, on-off traffic approximating voice with silence compression, and more aggressive on-off traffic with larger packet sizes and peak and mean rates approximating that of video traffic.

In summary, both the admission control and flow pre-emption mechanisms worked well for all of these traffic types under the assumptions of [CL-DEPLOY] (in particular under the assumption that there are many micro-flows between any pair of ingress / egress gateways, which, in turn, translates in the assumption that relatively high speed links are used). Details of the simulation study are given in Appendix B. In the pdf version of this document Appendix B also include graphs of simulation results.

So far the simulations have been run with a sensible estimate of suitable parameters. While a limited amount of work has been done to evaluate sensitivity of the results to the simulation parameters (see Appendix B), investigating further the sensitivity to these parameters is the next step.

Due to time constraints, we were able to simulate a single "congestion point" only, i.e. there was a single router where pre-congestion notification for admission control and/or pre-emption was triggered. Furthermore, admission control and flow pre-emption simulations were performed independently. A study of the interaction of admission control and flow pre-emption is also a subject of future work.

A further performance evaluation study is presented in [Zhang].

5. Encoding the Admission Marked and Pre-emption Marked states

In this Section we describe one proposal for how to encode the Admission Marking and Pre-emption Marking states in a packet, i.e. what change to make to which bits of a packet.

The encoding scheme uses the two ECN (Explicit Congestion Notification) bits in the IP header. The four ECN codepoints are used as follows:

ECN FIELD		
bit 6	bit 7	
0	0	Admission Marking
0	1	ECT(1)
1	0	ECT(0)
1	1	Pre-emption Marking
Other DSCPs		Non-PCN-Capable

Figure 5: Pre-Congestion Notification's use of the ECN Field in IP

A PCN-capable environment is one in which all the devices behave in accordance with the PCN mechanisms, for packets in the specific traffic class(es). Therefore a PCN-capable environment, such as a CL-region, meets the requirements of [Floyd] for a controlled environment.

A router knows a packet should be treated with the PCN behaviour if

- Its differentiated services codepoint (DSCP) is one configured for PCN marking. Packets with this DSCP are PCN-capable whatever the ECN codepoint is.

If necessary the router re-sets the ECN field to '00' to indicate Admission Marking and to '11' to indicate Pre-emption Marking. Packets with Admission Marking may be re-marked to Pre-emption Marking, but not vice-versa.

For the deployment model of [CL-DEPLOY] an ingress gateway knows, as part of the RSVP signalling set-up, whether a microflow is to be treated with the CPN behaviour by the CL-region. If necessary it sets the DSCP to a PCN-capable DSCP. It also sets the ECN field to either ECT(0) or ECT(1) as it chooses.

Other deployment models would be very similar. For example, in a framework where Pre-Congestion Notification operates from one end-

host to another, then the sending end-host would set the ECN field to either ECT(0) or ECT(1).One advantage of this encoding scheme is that it allows the (partial) use of the ECN nonce, thus providing similar protection against a cheater as [RFC3540]. However, a drawback is that if PCN marking is used with a pre-existing scheduling behaviour (such as EF), and some traffic still uses the legacy (EF) behaviour, then a new DSCP would be required to distinguish PCN-capable packets from ones that aren't PCN-capable.

Note that although we believe the encoding scheme is reasonable, it is not our final proposal. Alternatives are listed and discussed in Appendix C. We welcome advice and comments as to the most appropriate scheme.

6. Acknowledgements

This work has evolved from several previous independent efforts:

- Guaranteed QoS Synthesis [Hovell], which evolved from the Guaranteed Stream Provider developed in the M3I project [GSPa, GSP-TR], which in turn was based on the theoretical work of Gibbens and Kelly [DCAC]

- RTECN (Real-Time Explicit Congestion Notification) [RTECN]

- RMD (Resource Management in DiffServ) [RMD] and [Westberg]

7. Comments solicited

Comments and questions are encouraged and very welcome. They can be sent to the Transport Area Working Group's mailing list, tsvwg@ietf.org, and/or to the authors.

8. Changes from earlier version of the draft

The main changes are:

From -01 to -02:

Minor clarifications and corrections throughout.

From -00 to -01

The description of how to use pre-congestion notification marking in a CL-region is now described in [CL-DEPLOY].

Only one admission marking algorithm is now described.

A pre-emption marking scheme has been added.

Various options for encoding the marking are described and discussed in Appendix C.

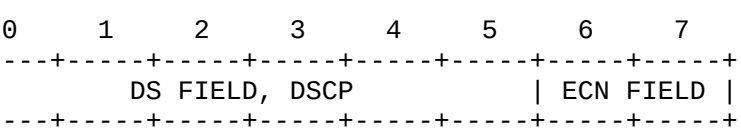
Simulation results are described in Appendix B and summarised in Section 4.

9. Appendix A: Explicit Congestion Notification

This Appendix provides a brief summary of Explicit Congestion Notification (ECN).

[RFC3168] specifies the incorporation of ECN to TCP and IP, including ECN's use of two bits in the IP header. It specifies a method for indicating incipient congestion to end-nodes (e.g. as in RED, Random Early Detection), where the notification is through ECN marking packets rather than dropping them.

ECN uses two bits in the IP header of both IPv4 and IPv6 packets:



DSCP: differentiated services codepoint
ECN: Explicit Congestion Notification

Figure A.1: The Differentiated Services and ECN Fields in IP.

The two bits of the ECN field have four ECN codepoints, '00' to '11':

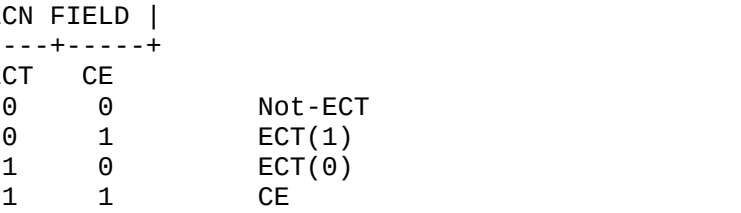


Figure A.2: The ECN Field in IP.

The not-ECT codepoint '00' indicates a packet that is not using ECN.

The CE codepoint '11' is set by a router to indicate congestion to the end nodes. The term 'CE packet' denotes a packet that has the CE codepoint set.

The ECN-Capable Transport (ECT) codepoints '10' and '01' (ECT(0) and ECT(1) respectively) are set by the data sender to indicate that the end-points of the transport protocol are ECN-capable. Routers treat the ECT(0) and ECT(1) codepoints as equivalent. Senders are free to use either the ECT(0) or the ECT(1) codepoint to indicate ECT, on a packet-by-packet basis. The use of both the two codepoints for ECT is

motivated primarily by the desire to allow mechanisms for the data sender to verify that network elements are not erasing the CE codepoint, and that data receivers are properly reporting to the sender the receipt of packets with the CE codepoint set.

ECN requires support from the transport protocol, in addition to the functionality given by the ECN field in the IP packet header. [RFC3168] addresses the addition of ECN Capability to TCP, specifying three new pieces of functionality: negotiation between the endpoints during connection setup to determine if they are both ECN-capable; an ECN-Echo (ECE) flag in the TCP header so that the data receiver can inform the data sender when a CE packet has been received; and a Congestion Window Reduced (CWR) flag in the TCP header so that the data sender can inform the data receiver that the congestion window has been reduced.

The transport layer (e.g. TCP) must respond, in terms of congestion control, to a *single* CE packet as it would to a packet drop.

The advantage of setting the CE codepoint as an indication of congestion, instead of relying on packet drops, is that it allows the receiver(s) to receive the packet, thus avoiding the potential for excessive delays due to retransmissions after packet losses.

10. Appendix B - Details of simulations

The results of the simulation study referred to in Section 4 are presented below. Further evaluation can be found in [Zhang].

10.1. Network and signalling model

In most simulations, the network is modelled as a single link between an ingress and an egress node, all flows sharing the same link. Figure B.1 shows the modelled network. A is the ingress node and B is the egress node.

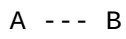


Figure B.1: Simulated Single Link Network.

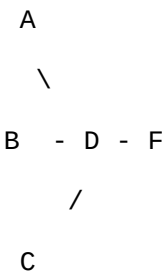


Figure B.2: Simulated Multi Link Network.

A subset of simulations uses a network structured similarly to the network shown on figure B.2. A set of ingresses (A,B,C) connected to an interior node in the network (D) with links of different propagation delay. This node in turn is connected to the egress (F). In this topology, different sets of flows between each ingress and the egress converge on the single link, where pre-congestion notification algorithm is enabled. In our simulations, the network has 100 ingress nodes, each connected to the interior node with a different propagation delay (1ms to 100ms). The point of congestion is taken to be the link (D-F) connecting the interior node to the

egress node. This link is modelled with a 10ms propagation delay. Therefore the range of RTTs is from 22ms to 220ms.

The simple network topology was due to a lack of time for the simulations.

Our simulations concentrated primarily on the range of capacities of 'bottleneck' links with sufficient aggregation - above 10 Mbps for voice and 622 Mbps for "video", up to 1 Gbps. But we also investigated slower 'bottleneck' links down to 512 kbps.

In the simulation model, a call request arrives at the ingress and immediately sends a message to the egress. The message arrives at the egress after the propagation time plus link processing time (but no queuing delay). When the egress receives this message, it immediately responds to the ingress with the current Congestion-Level-Estimate. If the Congestion-Level-Estimate is below the specified CLE-threshold, the call is admitted, otherwise it is rejected.

The life of a call outside the domain described above is not modelled. Propagation delay from source to the ingress and from destination to the egress is assumed negligible and is not modelled.

10.2. Simulated Traffic types

Three types of traffic were simulated (CBR voice, on-off traffic approximating voice with silence compression, and on-off traffic with higher peak and mean rates (we termed the latter "video" as the chosen peak and mean rate was similar to that of an mpeg video stream, although no attempt was made to match any other parameters of this traffic to those of a video stream). The distribution of flow duration was chosen to be exponentially distributed with mean 2min, regardless of the traffic type. In most of the experiments flows arrived according to a Poisson distribution with mean arrival rate chosen to achieve a desired amount of overload over the configured-pre-emption-rate or configured-admission-limit in each experiment. Overloads in the range 2x to 5x have been investigated.

In addition, some experiments investigated a batch Poisson model. Here the batch represented a set of calls arriving at almost the same time. The batch arrival process was Poisson, and the batch size was geometrically distributed with a mean of up to 5 calls per batch.

For on-off traffic, on and off periods were exponentially distributed with the specified mean.

Traffic parameters for each flow are summarized below:

10.2.1. Voice CBR

* Average rate 64 Kbps,

* Packet length 160 bytes

* packet inter-arrival time 20ms

10.2.2. On-off traffic approximating voice with silence compression

* Packet length 160 bytes

* Long-term average rate 21.76 Kbps

* On Period mean duration 340ms; during the on period traffic is sent with the CBR voice parameters described above

* Off Period mean duration 660ms; no traffic is sent during the off period.

10.2.3. High-rate on-off traffic

* Long term average rate 4 Mbps

* On Period mean duration 340ms; during the on-period the packets are sent at 12 Mbps (1500 byte packets, packet inter-arrival: 1ms)

* Off Period mean duration 660ms

10.3. Admission Control Simulations

10.3.1. Summary of the key parameters for CAC

10.3.1.1. Virtual Queue settings

Most of the simulations were run with the following Virtual Queue thresholds:

* min-marking-threshold: 5ms at link speed,

* max-marking-threshold: 15ms at link speed,

* virtual-queue-upper-limit: 20ms at link speed.

The virtual-queue-upper-limit puts an upper bound on how much the virtual queue can grow.

Note that the virtual queue is drained at a configured rate smaller than the link speed. Most of the simulations were set with the configured-admission-rate of the virtual queue at half the link speed.

Note that as long as there is no packet loss, the admission control scheme successfully keeps the load of admitted flows at the desired level regardless of the actual setting of the configured-admission-limit. However, it is not clear if this remains true when the configured-admission-rate is close to the link speed/actual queue service rate. Further work is necessary to quantify the performance of the scheme with smaller service rate/virtual queue rate ratio, where packet loss may be an issue.

10.3.1.2. Egress measurement parameters.

In our simulations, the CLE-threshold was chosen as 0.5. The CLE is computed as an exponential weighted moving average (EWMA) with a weight of 0.01. The CLE is computed on a per-packet basis.

10.3.2. Overview of the Admission Control Results

We found that on links of capacity from 10Mbps to OC3, congestion control for CBR voice and ON_OFF voice traffic work reliably with the range of parameters we simulated, both with Poisson and Batch call arrivals. As the performance of the algorithm was quite good at these speeds, and generally becomes the better the higher the degree of aggregation of traffic, we chose to not investigate higher link speeds for CBR and on-off voice, within the time constraints of this effort.

For higher-rate on-off "video" traffic, due to time limitations we simulated 1Gbps and OC12 (622 Mbps) links and Poisson arrivals only. Note that due to the high mean and peak rates of this traffic model, slower links are unlikely to yield sufficient level of aggregation of this type of traffic to satisfy the flow aggregation assumptions of [CL-ARCH]. Our simulations indicated that this model also behaved quite well, although the deviation from the configured-admission-rate is slightly higher in this case than for the less bursty traffic models.

For these link speeds and traffic models, we investigated the demand overload of 2x-5x.

Table B.1 below summarizes the worst case difference between the admitted load vs. configured-admission-rate. The worst case difference was taken over all experiments with the corresponding range of link speeds and demand overloads. In general, the higher the demand, the more challenging it is for the admission control algorithm due to a larger number of near-simultaneous arrivals at higher overloads, and as a result the worst case results in Table B.1 correspond to the 5x demand overload experiments.

Link type	traffic type	call arrival process	diff between mean admitted load & conf-adm-rate	standard deviation
T3,100Mbps,OC3	CBR	POISSON	0.5%	0.5%
T3,100Mbps,OC3	ON-OFF V	POISSON	2.5%	2.5%
T3,100Mbps,OC3	CBR	BATCH	1.0%	1.0%
T3,100Mbps,OC3	ON-OFF V	BATCH	3.0%	3.0%
1Gbps	"Video"	POISSON	2.0%	8.0%
OC12	"Video	POISSON	0.0%	10.0%

Table B.1. Summary of the admission control results for links above T3 speeds
Note: T1 = 1.5Mbps, T3 = 45Mbps, OC3 = 155Mbps, OC12 = 622Mbps

Sample simulation graphs for the experiments summarized in Table 6.1 can be viewed in the PDF version of this draft.

Below are sample results for admission control experiments. Graphs a) and b) show results for a 155 Mbps link with the CBR voice, Poisson and Batch call arrival models respectively. Graphs c) and d) show results for an 155 Mbps link with on-off voice, Poisson and Batch arrival model respectively. Graph e) shows the results for a 1Gbps

link with on-off-video traffic, Poisson call arrival model. All these results were obtained with min-marking-threshold = 5 ms, max-marking-threshold = 15 ms, virtual-queue-upper-limit=20ms.

Graphs a) and b) show results for a 155 Mbps link with the CBR voice, Poisson and Batch call arrival models respectively.

Graphs c) and d) show results for an 155 Mbps link with on-off voice, Poisson and Batch arrival model respectively.

Graph e) shows the results for a 1Gbps link with on-off-video traffic, Poisson call arrival model.

On slower links, accuracy of admission control algorithm was lower with Poisson arrivals, and was especially challenging with burstier Batch arrivals. This is described in section 6.3.3 below.

In general, we find that the admission control algorithm perform the better the larger degree of aggregation of traffic on the link. The algorithm performs well in the range of link speeds we expect to see in a CL region.

10.3.3. Sensitivity to Poisson Arrivals assumption

We investigated whether making the call arrival process burstier than Poisson has an effect on the performance of the admission control algorithm. To that end we investigated the comparative performance of the algorithm with Poisson and Batch call arrival processes, described in section 10.2. The mean call arrival rate was the same for both processes, with the demand overloads ranging from 2x to 5x.

We found that the admission control algorithm works reliably for both CBR and VBR at links of 1Mbps and above for up to 5x overloads for both Poisson and Batch call arrivals. We also found that the admission control algorithm only works reasonably well at links of 1 Mb/s if we assume CBR traffic and Poisson arrival. At T1 speeds and below, Batch arrivals resulted in over-admission, the degree of which increased on slower links.

Table B.2 below summarizes the difference between the admitted load and the configured-admission-rate for CBR Voice in the case of

Poisson and Batch arrivals. Table B.3 provides a similar summary for on-off traffic simulating voice with silence compression. The results in the tables correspond to the worst case across all overload factors (and when multiple links speeds are listed, across all those link speeds).

Link type	arrival model	diff between mean admitted load & conf-adm-rate	standard deviation
1Mbps, T1	BATCH	30.0%	30.0%
10 Mbps	BATCH	5.0%	8.0%
T3,100Mbps,OC3	BATCH	1.0%	1.0%
1Mbps, T1	POISSON	5.0%	10.0%
10 Mbps	POISSON	1.0%	2.0%
T3,100Mbps,OC3	POISSON	0.5%	0.5%

Table B.2. Comparison of Poisson and Batch call arrival models for CBR voice. Note: T1 = 1.5Mbps, T3 = 45Mbps, OC3 = 155Mbps, OC12 = 622Mbps

Link type	arrival model	diff between mean admitted load & conf-adm-rate	standard deviation
1Mbps, T1	BATCH	40.0%	30.0%
10 Mbps	BATCH	8.0%	6.0%
T3,100Mbps,OC3	BATCH	3.0%	3.0%
1Mbps, T1	POISSON	15.0%	20.0%
10 Mbps	POISSON	7.0%	6.0%
T3,100Mbps,OC3	POISSON	2.5%	2.5%

Table B.3. Comparison of Poisson and Batch call arrival models for on-off voice with silence compression.
Note: T1 = 1.5Mbps, T3 = 45Mbps, OC3 = 155Mbps, OC12 = 622Mbps

10.3.4. Sensitivity to marking parameters

The behaviour of the congestion control algorithm in all simulation experiments did not substantially differ depending on whether the marking was "ramp", i.e. whether a separate min-marking-threshold and max-marking-threshold were used, with linear marking probability between these thresholds, or whether the marking was "step" with the min-marking-threshold and max-marking-threshold collapsed at the max-marking-threshold value, and marking all packets with probability 1 above this collapsed threshold.

However, the difference between "ramp" and "step" may be more visible in the multiple congestion point case (recall that only a single congestion point experiments were performed so far).

Another possible reason for this apparent lack of difference between "ramp" and "step" may relate to the choice of the egress measurement parameters and a relatively high CLE threshold of 50%. Choosing a lower CLE-acceptance threshold and a faster measurement timescale may result in a better sensitivity to lower levels of marked traffic.

Investigating the interaction between settings of the marking thresholds, the CLE-threshold, and the measurement parameters at the egress is an area of future investigation.

In contrast, the limited number of simulation experiments we performed indicate that the choice of the absolute value of the min-marking-threshold, the max-marking-threshold and the virtual-queue-upper-limit can have an effect on the algorithm performance. Specifically, choosing the min-marking-threshold and the max-marking-threshold too small may cause substantial underutilization, especially on the slow links. However, at larger values of the min-marking-threshold and the max-marking-threshold, preliminary experiments suggest the algorithm's performance is insensitive to their values. The choice of the virtual-queue-upper-limit affects the amount of over-admission (above the configured-admission-rate threshold) in some cases, although this effect is not consistent throughout the experiments.

The Table B.4 below gives a summary of the difference between the admitted load and the configured-admission-rate as a function of the virtual queue parameters, for the 4 Mbps on-off traffic model. The results in the table represent the worst case result among the experiments with different degree of demand overloads in the range of 2x-5x. Typically, higher deviation of admitted load from the configured-admission-rate occurs for the higher degree of demand overload.

Link type	min-threshold, max-threshold, upper-limit(ms)	diff between mean admitted load & conf-adm-rate	standard deviation
1Gbps	5, 15, 20	6.0%	8.0%
1Gbps	1, 5, 10	2.0%	7.0%
1Gbps	5, 15, 45	2.0%	8.0%
OC12	5, 15, 20	5.0%	11.0%
OC12	1, 5, 10	2.0%	13.0%
OC12	5, 15, 45	0.0%	10.0%

Table B.4. Sensitivity of 4 Mbps on-off "video" traffic to the virtual queue settings.
Note: T1 = 1.5Mbps, T3 = 45Mbps, OC3 = 155Mbps, OC12 = 622Mbps

Impact of the virtual queue parameter setting is a subject of further study.

10.3.5. Sensitivity to RTT

We performed a limited amount of sensitivity of the admission control algorithm used to the range of round trip propagation time (which is the dominant component of the control delay in the typical environment using pre-congestion notification).

Specifically, we studied the case when different groups of flows sharing a single bottleneck link in the network have a range of roundtrip delays between 22 and 220 ms, as shown in Figure B.2.

The results were good for all types of traffic tested, implying that the admission control algorithm is not sensitive to the either the absolute value of the round-trip propagation time or relative value of the round-trip propagation time, at least in the range of values tested. We expect this to remain true for a wider range of round-trip propagation times.

10.3.6. Future Work for Admission Control Experiments

Areas of future investigation include extending the study of sensitivity to multiple congestion points and topologies, further investigation of sensitivity to factors such as marking parameters, implementation details and time scale of egress measurements, the CLE-threshold. Also variations on the marking algorithm will be studied.

Another area of investigation is to understand the sensitivity to the ratio of configured-admission-rate to the actual queue service rate/link speed, and specifically study how close the configured-admission-rate can be to the actual queue draining rate. A related investigation is to understand the effect of packet loss on the admission control mechanisms. Packet loss can occur if the configured-admission-rate is sufficiently close to the actual queue rate.

More realistic Video modelling and the mix of video and voice traffic in the same queue is also an area of further study.

10.4. Flow Pre-emption Simulations

10.4.1. Flow Pre-emption Model and key parameters

The same single-congestion-point network model as described in section 10.1 for admission control is used for flow pre-emption. Flow arrival and traffic models are also the same as for CAC admission control simulations.

In all flow pre-emption simulations, flows arrive at the ingress according to a Poisson distribution, with the mean load of "unrestricted" arrivals exceeding the pre-emption threshold by a factor of 2 to 5. However, as explained below, the pre-emption simulation involve a very sudden surge of traffic to simulate a network failure scenario.

In the simulation, the router implementing PCN Pre-emption Marking operates as described in section 3, marking packets which find no token in the token bucket. When an egress gateway receives a marked packet from the ingress, it will start measuring its Sustainable-

Aggregate-Rate for this ingress, if it is not already in the pre-emption mode.

If a marked packet arrives while the egress is already in the pre-emption mode, the packet is ignored.

The measurement is interval based, with 100ms measurement interval chosen in all simulations.

At the end of the measurement interval, the egress sends the measured Sustainable-Aggregate-Rate to the ingress, and leaves the pre-emption mode.

When the ingress receives the sustainable rate from the egress, it starts its own interval immediately (unless it is already in a measurement interval), and measures its sending rate to that egress. Then at the end of that measurement interval, it pre-empts the necessary amount of traffic. The ingress then leaves the pre-emption mode until the next time it receives the sustainable rate estimate from the egress.

Due to time limitations, in all our simulations the ingress used the same length of the measurement interval as the egress. Investigation of the impact of different measurement intervals is an important area of future work.

To avoid excessive pre-emption due to the rate measurement errors, we used two error factors, Error1 and Error2 to trigger decisions on when to pre-empt and how much to pre-empt at the ingress. To that end, the ingress did not trigger pre-emption unless the sending rate it measured was greater than SAR + Error1 (SAR=Sustainable Aggregate Rate). Similarly, the ingress pre-empted enough flows to reduce its sending rate to SAR - Error2. Both Error1 and Error2 in all simulations were in the range of 2-5%.

The configured-pre-emption-rate was set to 50% of link speed. Token bucket depth was set to 64 packets for CBR and 128 packets for on-off traffic.

We only tested on the network shown in Figure B.1 and we experimented with different propagation delay values: 10ms, 50ms and 100ms.

Due to time limitation, only links above T3 rate were simulated in Pre-emption experiments.

In all pre-emption experiments, we simulated the base load of traffic below pre-emption threshold. At some point during the experiment, the

load was suddenly increased to simulate sudden overload such that might occur after a link failure causes rerouting of some traffic to a previously un-congested link. In order to model the fact that a link failure may cause flows rerouting to a particular link over a period of time, we simulated a "one-wave" traffic surge, where the extra flows arrived near simultaneously, and a "three-wave" traffic surge, where there are two surges of traffic arriving close together (within one measurement interval), followed by a third surge at a later time.

10.4.2. Summary of Flow Pre-emption Experiments.

Our initial simulations demonstrated that in general performance of the flow pre-emption mechanism was good, and the appropriate amount of traffic was pre-empted in all simulated cases, as long as the depth of the pre-emption token bucket was set appropriately (64 packets for CBR, 128 or higher for on-off traffic). The pre-emption always occurred very fast (in particular, in the simulation graphs shown in the pdf version of this document with time granularity of 1 second, pre-emption looks instantaneous).

Perhaps the most useful result of the simulation experiments we were able to run so far was the importance of choosing the token bucket depth deep enough to accommodate the expected burstiness on CL traffic. If the token bucket depth is too small, instantaneous bursts may cause false pre-emption events. Note that if traffic load is stable or decreasing, then marking some packets erroneously during a an unexpected short burst does not cause any false pre-emption, because the rate measurement of the sustained rate is not affected by a small amount of pre-emption-marked packets. However, if the traffic load is increasing (while still remaining below pre-emption level on the average), a packet marked for pre-emption because it found no tokens in the too-shallow token bucket, may cause a false pre-emption event.

Below are sample results for pre-emption experiments with CBR voice, on-off voice and on-off "video" traffic, and a Poisson call arrival model. In all these graphs a single overload event occurs in the middle of a simulation run, triggering pre-emption. Graphs a) and b) show pre-emption simulations on voice traffic (CBR and on-off) on a 155Mbps link, with the pre-emption token bucket depth of 64 packets. Graph c) shows pre-emption of on-off "video" traffic on a 1Gbps link, with the pre-emption token bucket depth of 128 packets. All three experiments use Error1=Error2=5%, and the configured-pre-emption-rate set to 50% of the link rate.

Graphs a) and b) show pre-emption simulations on voice traffic (CBR and on-off) on a 155Mbps link, with the pre-emption token bucket depth of 64 packets.

Graph c) shows pre-emption of on-off "video" traffic on a 1Gbps link, with the pre-emption token bucket depth of 128 packets.

10.4.3. Future Work on Flow Pre-emption Experiments

Further work is required to study potential ways of reducing sensitivity of the algorithm to the token bucket depth. Potential approaches may be to smooth out pre-emption signal by requiring a certain amount of pre-emption-marked packets to arrive to the egress before measurement of the sustainable rate is triggered. An obvious trade-off to be quantified is the corresponding increase in the reaction time to receiving a pre-emption-marked packet.

Further quantification of the sensitivity to traffic burstiness and rate measurement implementation and time scales is an important area for future work.

More realistic Video modelling and the mix of video and voice traffic in the same queue is also an area of further study.

Another area of further investigation is the interaction of flow pre-emption and admission control, and specifically understanding of how close the admission and pre-emption rates can be on one link. A related topic is the interaction of flow pre-emption and admission control triggered by different links for the same ingress-egress pair.

The exact algorithm for selecting which flows to pre-empt in the case of variable rate flows and mixture of traffic profile is subject of further study.

Representative graphs for pre-emption experiments are presented in the PDF version of this draft.

11. Appendix C - Alternative ways of encoding the Admission Marked and Pre-emption Marked States

In this Appendix we list and discuss alternative ways of encoding the Admission Marked and Pre-emption Marked states. We ignore minor variants such as swapping the encoding for the Admission Marked and Pre-emption Marked states.

11.1. Alternative 1

The first alternative is the one given in Section 5 above.

ECN FIELD		
bit 6	bit 7	
0	0	Admission Marking
0	1	ECT(1)
1	0	ECT(0)
1	1	Pre-emption Marking
Other DSCPs		Not ECN capable

Figure C.1: Encoding scheme Alternative 1

11.2. Alternative 2

In the second alternative, both Admission Marking and Pre-emption Marking are encoded as '11', depending on the original ECT marking:

- Setting the ECN field of an ECT(1) packet to '11' indicates Admission Marking
- Setting the ECN field of an ECT(0) packet to '11' indicates Pre-emption Marking

```
        +-----+-----+
        | ECN FIELD |
        +-----+-----+
        bit 6  bit 7
           0     0         Not-ECT
           0     1         ECT(1/A)  re-mark ECT(1) to '11' to encode
                                     Admission Marking
           1     0         ECT(0/P)  re-mark ECT(0) to '11' to encode
                                     Pre-emption Marking
           1     1         Admission Marking or Pre-emption Marking
```

Figure C.2: Encoding scheme Alternative 2

11.3. Alternative 3

The third alternative is a combination of the previous two schemes.

```
        +-----+-----+
        | ECN FIELD |
        +-----+-----+
        bit 6  bit 7
           0     0         Admission Marking
           0     1         ECT(1/A)  re-mark ECT(1) to '00' to encode
                                     Admission Marking
           1     0         ECT(0/P)  re-mark ECT(0) to '11' to encode
                                     Pre-emption Marking
           1     1         Pre-emption Marking

        Other DSCPs        Not ECN capable
```

Figure C.3: Encoding scheme Alternative 3

11.4. Alternative 4

In the fourth alternative a packet is re-marked with a new DSCP to indicate Pre-emption Marking.

```
        +-----+-----+
        | ECN FIELD |
        +-----+-----+
        bit 6  bit 7
           0     0         Not ECN capable
           0     1         ECT(1)
           1     0         ECT(0)
           1     1         Admission Marking

           New DSCP        Pre-emption Marking
```

Figure C.4: Encoding scheme Alternative 4

11.5. Alternative 5

The fifth alternative doesn't include the ECN nonce.

```
        +-----+-----+
        | ECN FIELD |
        +-----+-----+
        bit 6  bit 7
           0     0         Not ECN capable
           0     1         PCN capable
           1     0         Admission Marking
           1     1         Pre-emption Marking
```

Figure C.5: Encoding scheme Alternative 5

11.6. Comparison of Alternatives

In this section we compare the encoding alternatives against various criteria. No scheme is perfect. We would like feedback and advice from the IETF community as to which is most suitable. The choice of how to encode the markings is non-trivial because we have five things we want to encode, and only have four states available in the two bits of the ECN field:

- Admission Marking - the traffic level is such that the router Admission Marks the packet

- Pre-emption Marking - the traffic level is such that the router Pre-emption Marks the packet

- o ECT(0) - the first ECT codepoint, for backwards compatibility with the ECN nonce
- o ECT(1) - the other ECT codepoint, for backwards compatibility with the ECN nonce
- o Not ECN - to indicate to a router that the traffic is not ECN-capable, and indeed not PCN-capable.

Some of the issues won't be relevant in particular scenarios. For example, with the CL-region framework[CL-ARCH], the edge-to-edge region is a controlled environment so an ECN (RFC3168) packet should never encounter a PCN-enabled router.

Occasionally we use the terminology of the CL-region framework. This is merely to make the language more specific.

11.6.1. How compatible is the encoding scheme with RFC 3168 ECN?

All the encoding schemes for Pre-Congestion Notification use the ECN field, so there will be interactions between PCN and ECN. Three aspects are:

- o What happens if an ECN (RFC3168) packet encounters a PCN-enabled router?
- o What happens if a PCN-capable packet encounters an ECN-enabled router?
- o What happens if a flow that has been admitted, using the PCN-based admission control mechanism, wants to use ECN (i.e. from end-point to end-point as in RFC3168)?

The first two bullets are about an "unusual" situation, perhaps where re-routing means that a PCN-enabled packet gets routed onto an ECN router - or perhaps where one of the CL-regions ingress gateways is misconfigured so that it allows in ECN packets into the CL traffic class. The third bullet is when the end-point wants its flow, which has been reserved using PCN-based admission control, to also use ECN-congestion control. There has been some discussion (and disagreement) about whether this is a realistic requirement [Floyd] [tsvwg-ml].

- What happens if an ECN (RFC3168) packet encounters a PCN-enabled router?

The main issue here is if traffic at the PCN-router is above the admission or pre-emption threshold, and what then happens when the ECN packet reaches the RFC3168 ECN end-point.

Alternative 2 and 4 are very safe. If the PCN-router Admission Marks a packet ('11'), the ECN end-point interprets this as the CE codepoint. The admission threshold is lower (perhaps much lower) than an ECN threshold would be.

Alternative 3 is also safe. If the PCN-router Pre-emption Marks a packet ('11'), the ECN end-point interprets this as the CE codepoint. The pre-emption threshold is likely to be lower than an ECN threshold would be, and is definitely lower than the traffic level at which packets would start to be dropped.

Alternative 5 is probably OK. However if the level of RFC3168 traffic is above the PCN router's configured-admission-rate but below its configured-pre-emption-rate, then packets are admission marked (to '10') but not pre-emption marked (to '11'). Therefore the ECN traffic would tend to block new PCN flows, but not reduce its own rate. This would be safer with the encodings for admission marking and pre-emption marking swapped.

With Alternatives 1 and 3, if traffic is above the admission threshold then packets will be re-marked to '00'. A subsequent ECN router will therefore think the packet isn't ECN-capable.

With Alternative 5 packets are admission marked to '10', which could confuse an ECN RFC3168 end-point using the ECN nonce.

- What happens if a PCN-capable packet encounters an ECN-enabled router?

The main issue is if the ECN-router is becoming congested, so it changes the ECN field to '11', to indicate Congestion Experienced (CE).

With Alternatives 1, 3 and 5 '11' will be interpreted as Pre-emption Marking, so the pre-emption mechanism will be triggered.

With Alternative 2 either the pre-emption or admission mechanism would be triggered (depending whether it was originally a '10' or '01' packet).

With Alternative 4 the admission control mechanism will be triggered.

Interpretation of '11' as pre-emption marking is probably safer than interpreting it as admission marking, because it then pre-empts flows going through a congested ECN router. However, it isn't clear-cut what 'safe' means in this context.

o What happens if a flow that has been admitted, using the PCN-based admission control mechanism, wants to use ECN (i.e. from end-point to end-point as in RFC3168)?

For instance with the CL-region framework, it isn't clear what the ingress gateway should do if it gets a packet with the CE codepoint, '11'. All the PCN encoding schemes have the same issue. Some options:

- the ingress gateway could re-set a '11' packet to one of the ECT codepoints. However, as far as the ECN-end-point is concerned, the CE information is lost.

- The ingress gateway could pre-empt the flow. This is safer, but perhaps harsh as the flow would now be handled by the non-PCN-capable class within the CL-region, and by the non-ECN-capable class after that.

- Tunnelling between the ingress and egress gateways, e.g. all PCN-capable traffic could be tunnelled. This preserves both the ECN and PCN functionality, but at the cost of the tunnelling.

11.6.2. Does the encoding scheme allow an "ECN-nonce"?

The Explicit Congestion Notification (ECN)-nonce is an optional addition to ECN that protects against accidental or malicious concealment of marked packets from the TCP sender. It uses the two ECN-Capable Transport (ECT) codepoints in the ECN field of the IP header. It improves the robustness of congestion control by enabling co-operative senders to prevent receivers from exploiting ECN to gain an unfair share of network bandwidth.

Pre-Congestion Notification is targeted at real-time traffic, which we'd expect to use UDP or DCCP rather than TCP. However, we imagine an "ECN-nonce" could be defined for DCCP and perhaps UDP with similar functionality to the ECN-nonce.

Analysing the encoding schemes in the context of an ECN-nonce:

- Alternatives 2 and 4 would allow an ECN-nonce
- Alternatives 1 and 3 would party allow an ECN-nonce - in terms of the edge-to-edge framework, an egress gateway would be able to detect a cheating ingress gateway, but it wouldn't detect an interior router re-marking the ECN field from '11' to '00'.
- Alternative 5 wouldn't allow an ECN-nonce

An alternative scheme intended to prevent cheating when using ECN for admission control is proposed in [Re-PCN]. This scheme claims to provide protection against a much wider range of cheating strategies than the ECN-Nonce, including against cheating ingress nodes or senders. Whereas the ECN-nonce requires the sender to be trusted. This scheme uses a bit outside the ECN field, so Alternative 5 combined with that scheme could solve the problem of fitting five states into four codepoints.

11.6.3. Does the encoding scheme require new DSCP(s)?

- Alternatives 2 and 5 do not.
- Alternative 1 does not allow indication of a non-PCN-capable transport within the same DSCP as used by PCN-capable transports. Therefore, if the PCN-routers are used with a pre-existing scheduling behaviour (such as EF) an extra DSCP would have to be used to indicate the combination of PCN marking with EF scheduling.
- Alternative 4 needs a new DSCP so a PCN-router can Pre-emption Mark a packet.

In Section 5 we suggested that the Expedited Forwarding DSCP might be used to indicate to a PCN-router that a packet is part of a PCN-capable flow. However PCN could be used similarly to add admission control and flow pre-emption to other DSCP classes. With Alternative 4 a new DSCP would be needed for each PCN-enabled class.

It's not clear to what extent the requirement for extra DSCP(s) matters. DSCPs are plentiful in an IP network, but scarce in an MPLS

network where the DSCP/ECN byte is mapped to the three MPLS header EXP bits [MPLS/EXP]. However, note that there is at least no need to encode the ECN-nonce in the MPLS EXP field, as it is sufficient to encode the ECN-nonce in the underlying IP header.

11.6.4. Impact on measurements

With some of the Alternatives, the measurements by the egress gateway for instance, have to be modified:

With Alternative 2 and 3, it has to measure the rate of ECT(1/A) in order to deduce the total number of bits in admission marked packets.

With Alternative 2, the egress moves into the pre-emption alert state if the rate of ECT(0/P) is significantly less than 50%. This is slower than the other Alternatives which are triggered by a single pre-emption marked packet. It also makes it more likely that the egress moves into the pre-emption alert state when the traffic level actually doesn't justify this.

With Alternative 4 the egress has to monitor the new DSCP in order to measure pre-emption marked packets.

11.6.5. Other issues

With Alternatives 2 and 3, Admission Marking means re-marking the ECN field of a '01' packet and Pre-emption Marking means re-marking a '10' packet. Therefore extra work is required compared with the other Alternatives; exactly what the work is depends on the details of the framework using PCN.

With Alternatives 1 and 5 Pre-emption Marking overwrites Admission Marking.

With Alternative 4 Pre-emption Marking is indicated by a new DSCP. Some ECMP (Equal Cost Multipath Routing) algorithms use the DSCP field as one of the input fields used to calculate which link to forward a packet on. Therefore, with a network running ECMP there is a danger that a Pre-emption Marked packet might be forwarded on a different path to other PCN-capable packets. The extent that this matters is for further study. It is not an issue for the other encoding Alternatives.

12. References

A later version will distinguish normative and informative references.

[CL-DEPLOY] B. Briscoe, P. Eardley, D. Songhurst, F. Le Faucheur, A. Charny, S. Dudley, J. Babiarz, K. Chan, G. Karagiannis, A. Bader. A Deployment Model for Admission Control over DiffServ using Pre-Congestion Notification, draft-briscoe-tsvwg-cl-architecture-03.txt", (work in progress), October 2006

[DCAC] Richard J. Gibbens and Frank P. Kelly "Distributed connection acceptance control for a connectionless network", In: Proc. International Teletraffic Congress (ITC16), Edinburgh, pp. 941?952 (1999).

[Floyd] S. Floyd, 'Specifying Alternate Semantics for the Explicit Congestion Notification (ECN) Field', draft-floyd-ecn-alternates-00.txt (work in progress), April 2005

[GSPa] Karsten (Ed.), Martin "GSP/ECN Technology \& Experiments", Deliverable: 15.3 PtIII, M3I Eu Vth Framework Project IST-1999-11429, URL: http://www.m3i.org/ (February, 2002) (superseded by [GSP- TR])

[GSP-TR] Martin Karsten and Jens Schmitt, "Admission Control Based on Packet Marking and Feedback Signalling ?-- Mechanisms, Implementation and Experiments", TU-Darmstadt Technical Report TR-KOM-2002-03, URL: http://www.kom.e-technik.tu-darmstadt.de/publications/abstracts/KS02-5.html (May, 2002)

[Hovell] P. Hovell, R. Briscoe, G. Corliano, "Guaranteed QoS Synthesis - an example of a scalable core IP quality of service solution", BT Technology Journal, Vol 23 No 2, April 2005

[Re-PCN] B. Briscoe, "Emulating Border Flow Policing using Re-ECN on Bulk Data", draft-briscoe-tsvwg-re-ecn-border-cheat-00 (work in progress), February 2006

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC2211] J. Wroclawski, Specification of the Controlled-Load Network Element Service, September 1997

[RFC2474] Nichols, K., Blake, S., Baker, F. and D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC 2474, December 1998

[RFC2475] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z. and W. Weiss, "An Architecture for Differentiated Services", RFC 2475, December 1998.

[RFC2597] Heinanen, J., Baker, F., Weiss, W. and J. Wrocklawski, "Assured Forwarding PHB Group", RFC 2597, June 1999.

[RFC3168] Ramakrishnan, K., Floyd, S. and D. Black "The Addition of Explicit Congestion Notification (ECN) to IP", RFC 3168, September 2001.

[RFC3246] B. Davie, A. Charny, J.C.R. Bennet, K. Benson, J.Y. Le Boudec, W. Courtney, S. Davari, V. Firoiu, D. Stiliadis, 'An Expedited Forwarding PHB (Per-Hop Behavior)', RFC 3246, March 2002.

[RFC3540] N. Spring, D. Wetherall, D. Ely, 'Robust Explicit Congestion Notification (ECN) Signaling with Nonces', RFC 3540, June 2003.

[RMD] A Bader, L Westberg, G Karagiannis, C Kappler, T Phelan, 'RMD-QOSM - The Resource Management in DiffServ QoS model', draft-ietf-nsis-rmd-06 Work in Progress, February 2006

[RTECN] Babiarz, J., Chan, K. and V. Firoiu, 'Congestion Notification Process for Real-Time Traffic', draft-babiarz-tsvwg-rtecn-05 Work in Progress, October 2005.

[tsvwg-ml] Discussion on the TSVWG mailing list, Nov/Dec 2005.

[Westberg] L. Westberg, Z. R. Turanyi, D. Partain, A. Bader, G. Karagiannis, "Load Control of Real-Time Traffic", draft-westberg-loadcntr-04.txt (Work in progress), Dec 2005

[Zhang] J. Zhang, A. Charny, V. Liatsos, F. Le Faucheur, "Performance Evaluation of CL-PHB Admission and pre-emption Algorithms", draft-zhang-pcn-performance-evaluation.txt (Work in progress), October 2005

Authors' Addresses

Bob Briscoe
BT Research
B54/77, Sirius House
Adastral Park
Martlesham Heath
Ipswich, Suffolk
IP5 3RE
United Kingdom
Email: bob.briscoe@bt.com

Dave Songhurst
BT Research
B54/69, Sirius House
Adastral Park
Martlesham Heath
Ipswich, Suffolk
IP5 3RE
United Kingdom
Email: dsonghurst@jungle.bt.co.uk

Philip Eardley
BT Research
B54/77, Sirius House
Adastral Park
Martlesham Heath
Ipswich, Suffolk
IP5 3RE
United Kingdom
Email: philip.eardley@bt.com

Vassilis Liatsos
Cisco Systems, Inc.
1414 Massachusetts Avenue
Boxborough,
MA 01719,
USA
Email: vliatsos@ciscoyahoo.com

Francois Le Faucheur
Cisco Systems, Inc.
Village d'Entreprise Green Side - Batiment T3
400, Avenue de Roumanille
06410 Biot Sophia-Antipolis
France
Email: flefauch@cisco.com

Anna Charny
Cisco Systems, Inc.
14164 Massachusetts Ave
Boxborough,
MA 01719
USA
Email: acharny@cisco.com

Jozef Babiarz
Nortel Networks
3500 Carling Avenue
Ottawa, Ont. K2H 8E9
Canada
Email: babiarz@nortel.com

Kwok Ho Chan
Nortel Networks
600 Technology Park Drive
Billerica, MA 01821
USA
Email: khchan@nortel.com

Stephen Dudley
Nortel Networks
4001 E. Chapel Hill Nelson Highway
P.O. Box 13010, ms 570-01-0V8
Research Triangle Park, NC 27709
USA
Email: smdudley@nortel.com

Georgios Karagiannis
University of Twente
P.O. BOX 217
7500 AE Enschede,
The Netherlands
EMail: g.karagiannis@ewi.utwente.nl

Attila Bá¤¥r
attila.bader@ericsson.com

Lars Westberg
Ericsson AB
SE-164 80 Stockholm
Sweden
EMail: Lars.Westberg@ericsson.com

Intellectual Property Statement

Disclaimer of Validity

Copyright Statement