TSVWG
Internet Draft
draft-briscoe-tsvwg-cl-architecture-02.txt
Expires: September 2006

B. Briscoe
P. Eardley
D. Songhurst
BT

F. Le Faucheur
A. Charny
Cisco Systems, Inc

J. Babiarz
K. Chan
S. Dudley
Nortel

March 6, 2006

A Framework for Admission Control over DiffServ using Pre-Congestion Notification
draft-briscoe-tsvwg-cl-architecture-02.txt

Copyright Notice

Abstract

This document describes a framework to achieve an end-to-end Controlled Load (CL) service without the scalability problems of previous approaches. Flow admission control and if necessary flow pre-emption preserve the CL service to admitted flows. But interior routers within a large DiffServ-based region of the Internet do not require flow state or signalling. They only have to give early warning of their own congestion by bulk packet marking using new pre-congestion notification marking. Gateways around the edges of the region convert measurements of this packet granularity marking into admission control and pre-emption functions at flow granularity.

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 an INFORMATIONAL RFC, rather than a standards track document.

Table of Contents

1. Introduction

1.1. Summary

This document describes a framework 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 service is a quality of service (QoS) closely approximating the QoS that the same flow would receive from a lightly loaded network element [RFC2211]. Controlled Load (CL) is useful for inelastic flows such as those for real-time media.

In line with the "IntServ over DiffServ" framework defined in [RFC2998], the CL service is supported end-to-end and RSVP signalling [RFC2205] is used end-to-end, over an edge-to-edge DiffServ region.

```
 ___    ___    _______________________________________    ____    ___
|   |  |   |  |                                       |  |    |  |   |
|   |  |   |  |Ingress         Interior         Egress|  |    |  |   |
|   |  |   |  |gateway          nodes          gateway|  |    |  |   |
|   |  |   |  |-------+  +-------+  +-------+  +------|  |    |  |   |
|   |  |   |  | PCN-  |  | PCN-  |  | PCN-  |  |      |  |    |  |   |
|   |..|   |..|marking|..|marking|..|marking|..| Meter|..|    |..|   |
|   |  |   |  |-------+  +-------+  +-------+  +------|  |    |  |   |
|   |  |   |  |  \                                 /  |  |    |  |   |
|   |  |   |  |   \                               /   |  |    |  |   |
|   |  |   |  |    \  Congestion-Level-Estimate  /    |  |    |  |   |
|   |  |   |  |     \  (for admission control)  /     |  |    |  |   |
|   |  |   |  |      --<-----<----<----<-----<--      |  |    |  |   |
|   |  |   |  |      Sustainable-Aggregate-Rate       |  |    |  |   |
|   |  |   |  |        (for flow pre-emption)         |  |    |  |   |
|___|  |___|  |_______________________________________|  |____|  |___|

Sx     Access               CL-region                   Access    Rx
End    Network                                          Network   End
Host                                                              Host
                <------ edge-to-edge signalling ----->
              (for admission control & flow pre-emption)

<-------------------end-to-end QoS signalling protocol--------------->
```

Figure 1: Overall QoS architecture (NB terminology explained later)

In Section 1.1.1 we summarise how admission of new CL microflows is controlled so as to deliver the required QoS. In abnormal circumstances for instance a disaster affecting multiple interior nodes, then the QoS on existing CL microflows may degrade even if care was exercised when admitting those microflows before those circumstances. Therefore we also propose a mechanism (summarised in Section 1.1.2) to pre-empt some of the existing microflows. Then remaining microflows retain their expected QoS, while improved QoS is quickly restored to lower priority traffic.

As a fundamental building block to support these two mechanisms, we introduce "Pre-Congestion Notification". Pre-Congestion Notification (PCN) builds on the concepts of RFC 3168, "The addition of Explicit Congestion Notification to IP". The draft [PCN] proposes the respective algorithms that determine when a PCN-enabled router marks a packet with Admission Marking or Pre-emption Marking, depending on the traffic level.

Pre-Congestion Notification can supplement any Per Hop Behaviour. In order to support CL traffic we would expect it to supplement the existing Expedited Forwarding (EF). Within the controlled edge-to-edge region, a particular packet receives the Pre-Congestion Notification behaviour if the packet's DSCP (differentiated services codepoint) is set to EF (or whatever is configured for CL traffic) and also the ECN field indicates ECN Capable Transport.

There are various possible ways to encode the markings into a packet, using the ECN field and perhaps other DSCPs, which are discussed in [PCN]. In this draft we use the abstract names Admission Marking and Pre-emption Marking.

This framework assumes that the Pre-Congestion Notification behaviour is used in a controlled environment, i.e. within the controlled edge-to-edge region.

1.1.1. Flow admission control

This document describes a new admission control procedure for an edge-to-edge region, which uses new per-hop Pre-Congestion Notification 'admission marking' as a fundamental building block. In turn, an end-to-end CL service would use this as a building block within a broader QoS architecture.

The per-hop, edge-to-edge and end-to-end aspects are now briefly introduced in turn.

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.

Pre-Congestion Notification (PCN) builds on the concepts of ECN. PCN introduces a new algorithm that Admission Marks packets before there is any significant build-up of CL packets in the queue. Admission marked packets therefore act as an "early warning" when the amount of packets flowing is getting close to the engineered capacity. Hence it can be used with per-hop behaviours (PHBs) designed to operate with very low queue occupancy, such as Expedited Forwarding (EF). Note that our use of the ECN field operates across the CL-region, i.e. edge-to-edge, and not host-to-host as in [RFC3168].

Turning next to the edge-to-edge aspect. All nodes within a region of the Internet, which we call the CL-region, apply the PHB used for CL traffic and the Pre-Congestion Notification behaviour. Traffic must enter/leave the CL-region through ingress/egress gateways, which have special functionality. Typically the CL-region is the core or backbone of an operator. The CL service is achieved "edge-to-edge" across the CL-region, by using distributed measurement-based admission control: the decision whether to admit a new microflow depends on a measurement of the existing traffic between the same pair of ingress and egress gateways (i.e. the same pair as the prospective new microflow). (See Appendix B for further discussion on "What is distributed measurement-based admission control?")

As CL packets travel across the CL-region, nodes will admission mark packets (according to the Pre-Congestion Notification algorithm) as an "early warning" of potential congestion, i.e. before there is any significant build-up of CL packets in the queue. For traffic from each remote ingress gateway, the CL-region's egress gateway measures the fraction of CL traffic that is admission marked. The egress gateway calculates the value on a per bit basis as an exponentially weighted moving average (which we term Congestion-Level-Estimate). Then it reports it to the CL-region's ingress gateway piggy-backed on the signalling for a new flow. The ingress gateway only admits the new CL microflow if the Congestion-Level-Estimate is less than the value of the CLE-threshold. Hence previously accepted CL microflows will suffer minimal queuing delay, jitter and loss.

In turn, the edge-to-edge architecture is a building block in delivering an end-to-end CL service. The approach is similar to that described in [RFC2998] for Integrated services operation over DiffServ networks. Like [RFC2998], an IntServ class (CL in our case) is achieved end-to-end, with a CL-region viewed as a single reservation hop in the total end-to-end path. Interior nodes of the CL-region do not process flow signalling nor do they hold state. We assume that the end-to-end signalling mechanism is RSVP (Section 2.2). However, the RSVP signalling may itself be originated or terminated by proxies still closer to the edge of the network, such as home hubs or the like, triggered in turn by application layer signalling. [RFC2998] and our approach are compared further in Section 6.2.

An important benefit compared with the IntServ over DiffServ model [RFC2998] arises from the fact that the load is controlled dynamically rather than with the traffic conditioning agreements (TCAs). TCAs were originally introduced in the (informational) DiffServ architecture [RFC2475] as an alternative to reservation processing in the interior region in order to reduce the burden on interior nodes. With TCAs, in practice service providers rely on subscription-time Service Level Agreements that statically define the parameters of the traffic that will be accepted from a customer. The problem arises because the TCA at the ingress must allow any destination address, if it is to remain scalable. But for longer topologies, the chances increase that traffic will focus on an interior resource, even though it is within contract at the ingress [Reid], e.g. all flows converge on the same egress gateway. Even though networks can be engineered to make such failures rare, when they occur all inelastic flows through the congested resource fail catastrophically.

Distributed measurement-based admission control avoids reservation processing (whether per flow or aggregated) on interior nodes but flows are still blocked dynamically in response to actual congestion on any interior node. Hence there is no need for accurate or conservative prediction of the traffic matrix.

1.1.2. Flow pre-emption

An essential QoS issue in core and backbone networks is being able to cope with failures of nodes and links. The consequent re-routing can cause severe congestion on some links and hence degrade the QoS experienced by on-going microflows and other, lower priority traffic. Even when the network is engineered to sustain a single link failure, multiple link failures (e.g. due to a fibre cut or a node failure, or a natural disaster) can cause violation of capacity constraints and

resulting QoS failures. Our solution uses rate-based flow pre-emption, so that sufficient of the previously admitted CL microflows are dropped to ensure that the remaining ones again receive QoS commensurate with the CL service and at least some QoS is quickly restored to other traffic classes.

The solution has two aspects. First, triggering the ingress gateway to test whether pre-emption may be needed. A router enhanced with Pre-Congestion Notification may optionally include an algorithm that sets packets into the Pre-emption Marked state. Such a packet alerts the egress that pre-emption may be needed, which in turn sends a Pre-emption Alert message to the ingress. Secondly, calculating the right amount of traffic to drop. This involves the egress gateway measuring, and reporting to the ingress gateway, the current amount of CL traffic received from that particular ingress gateway. The ingress gateway compares this measurement (which is the amount that the network can actually support, and which we thus call the Sustainable-Aggregate-Rate) with the rate that it is sending and hence determines how much traffic needs to be pre-empted.

The solution operates within a little over one round trip time - the time required for microflow packets that have experienced Pre-emption Marking to travel downstream through the CL-region and arrive at the egress gateway, plus some additional time for the egress gateway to measure the rate seen after it has been alerted that pre-emption may be needed, and the time for the egress gateway to report this information to the ingress gateway.

1.1.3. Both admission control and pre-emption

This document describes both the admission control and pre-emption mechanisms, and we suggest that an operator uses both. However, we do not require this 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 CL-region will naturally return to normal controlled load service, but with reduced capacity. The drawback of this approach would be that until flows naturally depart to relieve the congestion, all flows and lower priority services will be adversely affected. As another example, an operator could use just admission control, avoiding heavy congestion (caused by re-routing) by 'capacity planning' - by configuring admission control thresholds to lower levels than the network could accept in normal situations such that the load after failure is

expected to stay below acceptable levels even with reduced network resources.

On the other hand, an operator could just rely for admission control on the traffic conditioning agreements of the DiffServ architecture [RFC2475]. The pre-emption mechanism described in this document would be used to counteract the problem described at the end of Section 1.1.1.

1.2. Terminology

This terminology is copied from the pre-congestion notification marking draft [PCN]:

- 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 of CL packets in the queue.

- 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.

The following terms are defined here:

- Ingress gateway: node at an ingress to the CL-region. A CL-region may have several ingress gateways.

- Egress gateway: node at an egress from the CL-region. A CL-region may have several egress gateways.

- Interior node: a node which is part of the CL-region, but isn't an ingress or egress node.

- CL-region: A region of the Internet in which all traffic enters/leaves through an ingress/egress gateway and all nodes run Pre-Congestion Notification marking. A CL-region is a DiffServ region (a DiffServ region is either a single DiffServ domain or set of contiguous DiffServ domains), but note that the CL-region does not use the traffic conditioning agreements (TCAs) of the (informational) DiffServ architecture.

- CL-region-aggregate: all the microflows between a specific pair of ingress and egress gateways. Note there is no identifier unique to the aggregate.

- Congestion-Level-Estimate: the number of bits in CL packets that are admission marked, divided by the number of bits in all CL packets. It is calculated as an exponentially weighted moving average. It is calculated by an egress gateway for the CL packets from a particular ingress gateway, i.e. there is a Congestion-Level-Estimate for each CL-region-aggregate.

- Sustainable-Aggregate-Rate: the rate of traffic that the network can actually support for a specific CL-region-aggregate. So it is measured by an egress gateway for the CL packets from a particular ingress gateway.

1.3. Existing terminology

This is a placeholder for useful terminology that is defined elsewhere.

1.4. Standardisation requirements

The framework described in this document has two new standardisation requirements:

- new Pre-Congestion Notification for Admission Marking and Pre-emption Marking are required, as detailed in [PCN].

- o the end-to-end signalling protocol needs to be modified to carry the Congestion-Level-Estimate report (for admission control) and the Sustainable-Aggregate-Rate (for flow pre-emption). With our assumption of RSVP (Section 2.2) as the end-to-end signalling protocol, it means that extensions to RSVP are required, as detailed in [RSVP-ECN], for example to carry the Congestion-Level-Estimate and Sustainable-Aggregate-Rate information from egress gateway to ingress gateway.

Other than these things, the arrangement uses existing IETF protocols throughout, although not in their usual architecture.

1.5. Structure of rest of the document

Section 2 describes some key aspects of the framework: our goals, assumptions and the benefits we believe it has. Section 3 describes the architecture (including a use case), whilst Section 4 summarises the required changes to the various nodes in the CL-region. Section 5 outlines some possible extensions. Section 6 provides some comparison with existing QoS mechanisms.

2. Key aspects of the framework

In this section we discuss the key aspects of the framework:

- At a high level, our key goals, i.e. the functionality that we want to achieve
- The assumptions that we're prepared to make
- The consequent benefits they bring

2.1. Key goals

The framework achieves an end-to-end controlled load (CL) service where a segment of the end-to-end path is an edge-to-edge Pre-Congestion Notification region. CL is a quality of service (QoS) closely approximating the QoS that the same flow would receive from a lightly loaded network element [RFC2211]. It is useful for inelastic flows such as those for real-time media.

- The CL service should be achieved despite varying load levels of other sorts of traffic, which may or may not be rate adaptive (i.e. responsive to packet drops or ECN marks).
- The CL service should be supported for a variety of possible CL sources: Constant Bit Rate (CBR), Variable Bit Rate (VBR) and voice with silence suppression. VBR is the most challenging to support.
- After a localised failure in the interior of the CL-region causing heavy congestion, the CL service should recover gracefully by pre-empting (dropping) some of the admitted CL microflows, whilst preserving as many of them as possible with their full CL QoS.
- It is suggested that flow pre-emption needs to be completed within 1-2 seconds, because it is estimated that after a few seconds then many affected users will start to hang up (and then not only is a flow pre-emption mechanism redundant and possibly even counter-productive, but also many more flows than necessary to reduce congestion may hang up). Also, other, lower priority traffic classes will not be restored to partial service until the higher priority CL service reduces its load on shared links.

- The CL service should support emergency services ([EMERG-RQTS], [EMERG-TEL]) as well as the Assured Service which is the IP implementation of the existing ITU-T/NATO/DoD telephone system architecture known as Multi-Level Pre-emption and Precedence [ITU.MLPP.1990] [ANSI.MLPP.Spec][ANSI.MLPP.Supplement], or MLPP. In particular, this involves admitting new high priority sessions even when admission control thresholds are reached and new routine sessions are rejected. Similarly, this involves taking into account session priorities and properties at the time of pre-empting flows.

2.2. Key assumptions

The framework does not try to deliver the above functionality in all scenarios. We make the following assumptions about the type of scenario to be solved.

- Edge-to-edge: all the nodes in the CL-region are upgraded with Pre-Congestion Notification, and all the ingress and egress gateways are upgraded to perform the measurement-based admission control and flow pre-emption. Note that although the upgrades required are edge-to-edge, the CL service is provided end-to-end.

- Additional load: we assume that any additional load offered within the reaction time of the admission control mechanism doesn't move the CL-region directly from no congestion to overload. So it assumes there will always be an intermediate stage where some CL packets are Admission Marked, but they are still delivered without significant QoS degradation. We believe this is valid for core and backbone networks with typical call arrival patterns (given the reaction time is little more than one round trip time across the CL-region), but is unlikely to be valid in access networks where the granularity of an individual call becomes significant.

- Aggregation: we assume that in normal operations, there are many CL microflows within the CL-region, typically at least hundreds between any pair of ingress and egress gateways. The implication is that the solution is targeted at core and backbone networks and possibly parts of large access networks.

- Trust: we assume that there is trust between all the nodes in the CL-region. For example, this trust model is satisfied if one operator runs the whole of the CL-region. But we make no such assumptions about the end nodes, i.e. depending on the scenario they may be trusted or untrusted by the CL-region.

- Signalling: we assume that the end-to-end signalling protocol is RSVP. Section 3 describes how the CL-region fits into such an end-to-end QoS scenario, whilst [RSVP-ECN] describes the extensions to RSVP that are required.

- Separation: we assume that all nodes within the CL-region are upgraded with the CL mechanism, so the requirements of [Floyd] are met because the CL-region is an enclosed environment. Also, an operator separates CL-traffic in the CL-region from outside traffic by administrative configuration of the ring of gateways around the region. Within the CL-region we assume that the CL-traffic is separated from non-CL traffic.

- Routing: we assume that one of the following applies:

  (same path) all packets between a pair of ingress and egress gateways follow the same path. This ensures that the Congestion-Level-Estimate used in the admission control procedure reflects the status of the path followed by the new flow's packets

  (load balanced) packets between a pair of ingress and egress gateways follow different paths but that the load balancing scheme is tuned in the CL-region to distribute load such that the different paths always receive comparable relative load. This ensures that the Congestion-Level-Estimate used in the admission control procedure (and which is computed taking into account packets travelling on all the paths) also approximately reflects the status of the actual path followed by the new microflow's packets

  (worst case assumed) packets between a pair of ingress and egress gateways follow different paths but that (i) it is acceptable for the operator to keep the CL traffic between this pair of gateways to a level dictated by the most loaded of all paths between this pair of gateways (so that CL flows may be rejected - or even pre-empted in some situations - even if one or more of the paths between the pair of gateways is operating below its engineered levels) and that (ii) it is acceptable for that operator to configure engineered levels below optimum levels to compensate for the fact that the effect on the Congestion-Level-Estimate of the congestion experienced over one of the paths may be diluted by traffic received over non-congested paths so that lower thresholds need to be used in these cases to ensure early admission control rejection and pre-emption over the congested paths.

We are investigating ways of loosening the restrictions set by some of these assumptions, for instance:

- Trust: to allow the CL-region to span multiple, non-trusting operators, using the technique of [Re-PCN] and mentioned in Section 5.1.

- Signalling: we believe that the solution could operate with another signalling protocol such as NSIS. It could also work with application level signalling as suggested in [RT-ECN].

- Additional load: we believe that the assumption is valid for core and backbone networks, with an appropriate margin between the configured-admission-rate and the capacity for CL traffic. However, in principle a burst of admission requests can occur in a short time. We expect this to be a rare event under normal conditions, but it could happen e.g. due to a 'flash crowd'. If it does, then more flows may be admitted than should be, triggering the pre-emption mechanisms. There are various approaches to how an operator might try to alleviate this issue, which are discussed in the 'Flash crowds' section 5.1 later.

- Separation: the assumption that CL traffic is separated from non-CL traffic implies that the CL traffic has its own PHB, not shared with other traffic. We are looking at whether it could share Expedited Forwarding's PHB, but supplemented with Pre-Congestion Notification. If this is possible, other PHBs (like Assured Forwarding) could be supplemented with the same new behaviours. This is similar to how RFC3168 ECN was defined to supplement any PHB.

- Routing: we are looking in greater detail at the solution in the presence of Equal Cost Multi-Path routing and at suitable enhancements. See also the "Tunnelling" section later.

2.3. Key benefits

We believe that the mechanism described in this document has several advantages:

- It achieves statistical guarantees of quality of service for microflows, delivering a very low delay, jitter and packet loss service suitable for applications like voice and video calls that generate real time inelastic traffic. This is because of its per microflow admission control scheme, combined with its dynamic on-path "early warning" of potential congestion. The guarantee is at least as strong as with IntServ Controlled Load (Section 6.1 mentions why the guarantee may be somewhat better), but without the scalability problems of per-microflow IntServ.

- It can support "Emergency" and military Multi-Level Pre-emption and Priority services, even in times of heavy congestion (perhaps caused by failure of a node within the CL-region), by pre-empting on-going "ordinary CL microflows". See also Section 4.5.

- It scales well, because there is no signal processing or path state held by the interior nodes of the CL-region.

- It is resilient, again because no state is held by the interior nodes of the CL-region. Hence during an interior routing change caused by a node failure no microflow state has to be relocated. The flow pre-emption mechanism further helps resilience because it rapidly reduces the load to one that the CL-region can support.

- It helps preserve, through the flow pre-emption mechanism, QoS to as many microflows as possible and to lower priority traffic in times of heavy congestion (e.g. caused by failure of an interior node). Otherwise long-lived microflows could cause loss on all CL microflows for a long time.

- It avoids the potential catastrophic failure problem when the DiffServ architecture is used in large networks using statically provisioned capacity. This is achieved by controlling the load dynamically, based on edge-to-edge-path real-time measurement of Pre-Congestion Notification, as discussed in Section 1.1.1.

- It requires minimal new standardisation, because it reuses existing QoS protocols and algorithms.

- It can be deployed incrementally, region by region or network by network. Not all the regions or networks on the end-to-end path need to have it deployed. Two CL-regions can even be separated by a network that uses another QoS mechanism (e.g. MPLS-TE).

- It provides a deployment path for use of ECN for real-time applications. Operators can gain experience of ECN before its applicability to end-systems is understood and end terminals are ECN capable.

3. Architecture

3.1. Admission control

In this section we describe the admission control mechanism. We discuss the three pieces of the solution and then give an example of how they fit together in a use case:

- the new Pre-Congestion Notification for Admission Marking used by all nodes in the CL-region
- how the measurements made support our admission control mechanism
- how the edge to edge mechanism fits into the end to end RSVP signalling

3.1.1. Pre-Congestion Notification for Admission Marking

This is discussed in [PCN]. Here we only give a brief outline.

To support our admission control mechanism, each node in the CL-region runs an algorithm to determine whether to set the packet into the Admission Marked state. The algorithm measures the aggregate CL traffic on the link and ensures that packets are admission marked before the actual queue builds up, but when it is in danger of doing so soon; the probability of admission marking increases with the danger. The algorithm's main parameter is the configured-admission-rate, which is set lower than the link speed, perhaps considerably so. Admission marked packets indicate that the CL traffic rate is reaching the configured-admission-rate and so act as an "early warning" that the engineered capacity is nearly reached. Therefore they indicate that requests to admit prospective new CL flows may need to be refused.

3.1.2. Measurements to support admission control

To support our admission control mechanism the egress measures the Congestion-Level-Estimate for traffic from each remote ingress gateway, i.e. per CL-region-aggregate. The Congestion-Level-Estimate is the number of bits in CL packets that are admission marked, divided by the number of bits in all CL packets. It is calculated as an exponentially weighted moving average. It is calculated by an egress node separately for the CL packets from each particular

ingress node. This Congestion-Level-Estimate provides an estimate of how near the links on the path inside the CL-region are getting to the configured-admission-rate. Note that the metering is done separately per ingress node, because there may be sufficient capacity on all the nodes on the path between one ingress gateway and a particular egress, but not from a second ingress to that same egress gateway.

3.1.3. How edge-to-edge admission control supports end-to-end QoS signalling

Consider a scenario that consists of two end hosts, each connected to their own access networks, which are linked by the CL-region. A source tries to set up a new CL microflow by sending an RSVP PATH message, and the receiving end host replies with an RSVP RESV message. Outside the CL-region some other method, for instance IntServ, is used to provide QoS. From the perspective of RSVP the CL-region is a single hop, so the RSVP PATH and RESV messages are processed by the ingress and egress gateways but are carried transparently across all the interior nodes; hence, the ingress and egress gateways hold per microflow state, whilst no state is kept by the interior nodes. So far this is as in IntServ over DiffServ [RFC2998]. However, in order to support our admission control mechanism, the egress gateway adds to the RESV message an opaque object which states the current Congestion-Level-Estimate for the relevant CL-region-aggregate. Details of the corresponding RSVP extensions are described in [RSVP-ECN].

3.1.4. Use case

To see how the three pieces of the solution fit together, we imagine a scenario where some microflows are already in place between a given pair of ingress and egress gateways, but the traffic load is such that no packets from these flows are admission marked as they travel across the CL-region. A source wanting to start a new CL microflow sends an RSVP PATH message. The egress gateway adds an object to the RESV message with the Congestion-Level-Estimate, which is zero. The ingress gateway sees this and consequently admits the new flow. It then forwards the RSVP RESV message upstream towards the source end host. Hence, assuming there's sufficient capacity in the access networks, the new microflow is admitted end-to-end.

The source now sends CL packets, which arrive at the ingress gateway. The ingress uses a five-tuple filter to identify that the packets are part of a previously admitted CL microflow, and it also polices the microflow to ensure it remains within its traffic profile. (The ingress has learnt the required information from the RSVP messages.)

When forwarding a packet belonging to an admitted microflow, the ingress sets the packet's DSCP and ECN fields to the appropriate values configured for the CL region. The CL packet now travels across the CL-region, getting admission marked if necessary.

Next, we imagine the same scenario but at a later time when load is higher at one (or more) of the interior nodes, which start to set CL packets into the Admission Marked state, because their load on the outgoing link is nearing the configured-admission-rate. The next time a source tries to set up a CL microflow, the ingress gateway learns (from the egress) the relevant Congestion-Level-Estimate. If it is greater than some CLE-threshold value then the ingress refuses the request, otherwise it is accepted.

It is also possible for an egress gateway to get a RSVP RESV message and not know what the Congestion-Level-Estimate is. For example, if there are no CL microflows at present between the relevant ingress and egress gateways. In this case the egress requests the ingress to send probe packets, from which it can initialise its meter. RSVP Extensions for such a request to send probe data can be found in [RSVP-ECN].

3.2. Flow pre-emption

In this section we describe the flow pre-emption mechanism. We discuss the two parts of the solution and then give an example of how they fit together in a use case:

- How an ingress gateway is triggered to test whether flow pre-emption may be needed

- How an ingress gateway determines the right amount of CL traffic to drop

The mechanism is defined in [PCN] and [RSVP-ECN].

3.2.1. Alerting an ingress gateway that flow pre-emption may be needed

Alerting an ingress gateway that flow pre-emption may be needed is a two stage process: a router in the CL-region alerts an egress gateway that flow pre-emption may be needed; in turn the egress gateway alerts the relevant ingress gateway. Every router in the CL-region

has the ability to alert egress gateways, which may be done either explicitly or implicitly:

- Explicit - the router per-hop behaviour is supplemented with a new Pre-emption Marking behaviour, which is outlined below. Reception of such a packet by the egress gateway alerts it that pre-emption may be needed.
- Implicit - the router behaviour is unchanged from the Admission Marking behaviour described earlier. The egress gateway treats a Congestion-Level-Estimate of (almost) 100% as an implicit alert that pre-emption may be required. ('Almost' because the Congestion-Level-Estimate is a moving average, so can never reach exactly 100%.)

To support explicit pre-emption alerting, each node in the CL-region runs an algorithm to determine whether to set the packet into the Pre-emption Marked state. The algorithm measures the aggregate CL traffic and ensures that packets are pre-emption marked before the actual queue builds up. The algorithm's main parameter is the configured-pre-emption-rate, which is set lower than the link speed (but higher than the configured-admission-rate). Thus pre-emption marked packets indicate that the CL traffic rate is reaching the configured-pre-emption-rate and so act as an "early warning" that the engineered capacity is nearly reached. Therefore they indicate that it may be advisable to pre-empt some of the existing CL flows in order to preserve the QoS of the others.

Note that the explicit mechanism only makes sense if all the routers in the CL-region have the functionality so that the egress gateways can rely on the explicit mechanism. Otherwise there is the danger that the traffic happens to focus on a router without it, and egress gateways then have also to watch for implicit pre-emption alerts.

When one or more packets in a CL-region-aggregate alert the egress gateway of the need for flow pre-emption, whether explicitly or implicitly, the egress puts that CL-region-aggregate into the Pre-emption Alert state. For each CL-region-aggregate in alert state it measures the rate of traffic at the egress gateway (i.e. the traffic rate of the appropriate CL-region-aggregate) and reports this to the relevant ingress gateway. The steps are:

- o Determine the relevant ingress gateway - for the explicit case the egress gateway examines the pre-emption marked packet and uses the state installed at the time of admission to determine which ingress gateway the packet came from. For the implicit case the egress gateway has already determined this information, because the Congestion-Level-Estimate is calculated per ingress gateway.

- o Measure the traffic rate of CL packets - as soon as the egress gateway is alerted (whether explicitly or implicitly) it measures the rate of CL traffic from this ingress gateway (i.e. for this CL-region-aggregate). Note that pre-emption marked packets are excluded from that measurement. It should make its measurement quickly and accurately, but exactly how is up to the implementation.

- o Alert the ingress gateway - the egress gateway then immediately alerts the relevant ingress gateway about the fact that flow pre-emption may be required. This Alert message also includes the measured Sustainable-Aggregate-Rate, i.e. the egress rate of CL-traffic for this ingress gateway. The Alert message is sent using reliable delivery. Procedures for support of such an Alert using RSVP are defined in [RSVP-ECN].

```
                                   _     _
             --------------       /       \          -----------------
CL packet   |Update        |     / Is it a \   Y    | Measure CL rate |
arrives --->|Congestion-   |--->/pre-emption\-----> | from ingress and|
            |Level-Estimate|    \  marked   /       | alert ingress   |
             --------------      \ packet? /         -----------------
                                  \_     _/
```

Figure 2: Egress gateway action for explicit Pre-emption Alert

```
                                   _     _
             --------------       /       \          -----------------
CL packet   |Update        |     /  Is     \   Y    | Measure CL rate |
arrives --->|Congestion-   |--->/  C.L.E.   \-----> | from ingress and|
            |Level-Estimate|    \ (nearly)  /       | alert ingress   |
             --------------      \ 100%?   /         -----------------
                                  \_     _/
```

Figure 3: Egress gateway action for implicit Pre-emption Alert

3.2.2. Determining the right amount of CL traffic to drop

The method relies on the insight that the amount of CL traffic that can be supported between a particular pair of ingress and egress gateways, is the amount of CL traffic that is actually getting across the CL-region to the egress gateway without being re-marked to the Pre-emption Marked state. Hence we term it the Sustainable-Aggregate-Rate.

So when the ingress gateway gets the Alert message from an egress gateway, it compares:

- The traffic rate that it is sending to this particular egress gateway (which we term ingress-aggregate-rate)

- The traffic rate that the egress gateway reports (in the Alert message) that it is receiving from this ingress gateway (which is the Sustainable-Aggregate-Rate)

If the difference is significant, then the ingress gateway pre-empts some microflows. It only pre-empts if:

```
Ingress-aggregate-rate > Sustainable-Aggregate-Rate + error
```

The "error" term is partly to allow for inaccuracies in the measurements of the rates. It is also needed because the ingress-aggregate-rate is measured at a slightly later moment than the Sustainable-Aggregate-Rate, and it is quite possible that the ingress-aggregate-rate has increased in the interim due to natural variation of the bit rate of the CL sources. So the "error" term allows for some variation in the ingress rate without triggering pre-emption.

The ingress gateway should pre-empt enough microflows to ensure that:

```
New ingress-aggregate-rate < Sustainable-Aggregate-Rate - error
```

The "error" term here is used for similar reasons but in the other direction, to ensure slightly more load is shed than seems necessary, in case the two measurements were taken during a short-term fall in load.

When the routers in the CL-region are using explicit pre-emption alerting, the ingress gateway would normally pre-empt microflows whenever it gets an alert (it always would if it were possible to set

"error" equal to zero). For the implicit case however this is not so. It receives an Alert message when the Congestion-Level-Estimate reaches (almost) 100%, which is roughly when traffic exceeds the configured-admission-rate. However, it is only when packets are indeed dropped en route that the Sustainable-Aggregate-Rate becomes less than the ingress-aggregate-rate so only then will pre-emption will actually occur on the ingress router.

Hence with the implicit scheme, pre-emption can only be triggered once the system starts dropping packets and thus the QoS of flows starts being significantly degraded. This is in contrast with the explicit scheme which allows flow pre-emption to be triggered before any packet drop, simply when the traffic reaches the configured-pre-emption-rate. Therefore we believe that the explicit mechanism is superior. However it does require new functionality on all the routers (although this is little more than a bulk token bucket - see [PCN] for details).

3.2.3. Use case for flow pre-emption

To see how the pieces of the solution fit together in a use case, we imagine a scenario where many microflows have already been admitted. We confine our description to the explicit pre-emption mechanism. Now an interior router in the CL-region fails. The network layer routing protocol re-routes round the problem, but as a consequence traffic on other links increases. In fact let's assume the traffic on one link now exceeds its configured-pre-emption-rate and so the router pre-emption marks CL packets. When the egress sees the first one of the pre-emption marked packets it immediately determines which microflow this packet is part of (by using a five-tuple filter and comparing it with state installed at admission) and hence which ingress gateway the packet came from. It sets up a meter to measure the traffic rate from this ingress gateway, and as soon as possible sends a message to the ingress gateway. This message alerts the ingress gateway that pre-emption may be needed and contains the traffic rate measured by the egress gateway. Then the ingress gateway determines the traffic rate that it is sending towards this egress gateway and hence it can calculate the amount of traffic that needs to be pre-empted.

The ingress gateway could now just shed random microflows, but it is better if the least important ones are dropped. The ingress gateway could use information stored locally in each reservation's state (such as for example the RSVP pre-emption priority) as well as information provided by a policy decision point in order to decide which of the flows to shed (or perhaps which ones not to shed). The

ingress gateway then initiates RSVP signalling to instruct the relevant destinations that their session has been terminated, and to tell (RSVP) nodes along the path to tear down associated RSVP state. To guard against recalcitrant sources, normal IntServ policing will block any future traffic from the dropped flows from entering the CL-region. Note that - with the explicit Pre-emption Alert mechanism - since the configured-pre-emption-rate may be significantly less than the physical line capacity, flow pre-emption may be triggered before any congestion has actually occurred and before any packet is dropped.

We extend the scenario further by imagining that (due to a disaster of some kind) further routers in the CL-region fail during the time taken by the pre-emption process described above. This is handled naturally, as packets will continue to be pre-emption marked and so the pre-emption process will happen for a second time.

Flow pre-emption also helps emergency/military calls by taking into account the corresponding call priorities when selecting calls to be pre-empted, which is likely to be particularly important in a disaster scenario.

4. Details

This section is intended to provide a systematic summary of the new functionality required by the routers in the CL-region.

A network operator upgrades normal IP routers by:

- Adding functionality related to admission control and flow pre-emption to all its ingress and egress gateways

- Adding Pre-Congestion Notification for Admission and Pre-emption Marking to all the nodes in the CL-region.

We consider the detailed actions required for each of the types of node in turn.

4.1. Ingress gateways

Ingress gateways perform the following tasks:

- Classify incoming packets - decide whether they are CL or non-CL packets. This is done using an IntServ filter spec (source and destination addresses and port numbers), whose details have been gathered from the RSVP messaging.

- Police - check that the microflow conforms with what has been agreed (i.e. it keeps to its agreed data rate). If necessary, packets which do not correspond to any reservations, packets which are in excess of the rate agreed for their reservation, and packets for a reservation that has earlier been pre-empted may be policed. Policing may be achieved via dropping or via re-marking of the packet's DSCP to a value different from the CL behaviour aggregate.

- Packet ECN colouring - for CL microflows, set the ECN field appropriately (see [PCN] for some discussion of encoding)

- Perform 'interior node' functions (see next sub-section)

- Admission Control - on new session establishment, consider the Congestion-Level-Estimate received from the corresponding egress gateway and most likely based on a simple configured CLE-threshold decide if a new call is to be admitted or rejected (taking into account local policy information as well as optionally information provided by a policy decision point).

- Probe - if requested by the egress gateway to do so, the ingress gateway generates probe traffic so that the egress gateway can compute the Congestion-Level-Estimate from this ingress gateway. Probe packets may be simple data addressed to the egress gateway and require no protocol standardisation, although there will be best practice for their number, size and rate.

- Measure - when it receives an Alert message from an egress gateway, it determines the rate at which it is sending packets to that egress gateway

- Pre-empt - calculate how much CL traffic needs to be pre-empted; decide which microflows should be dropped, perhaps in consultation with a Policy Decision Point; and do the necessary signalling to drop them.

4.2. Interior nodes

Interior nodes do the following tasks:

- Classify packets - examine the DSCP and ECN field to see if it's a CL packet

- Non-CL packets are handled as usual, with respect to dropping them or setting their CE codepoint.

- Pre-Congestion Notification - CL packets are Admission Marked and Pre-emption Marked according to the algorithm detailed in [PCN] and outlined in Section 3.

4.3. Egress gateways

Egress gateways do the following tasks:

- Classify packets - determine which ingress gateway a CL packet has come from. This is the previous RSVP hop, hence the necessary details are obtained just as with IntServ from the state associated with the packet five-tuple, which has been built using information from the RSVP messages.

- o Meter - for CL packets, calculate the fraction of the total number of bits which are in Admission marked packets. The calculation is done as an exponentially weighted moving average (see Appendix). A separate calculation is made for CL packets from each ingress gateway. The meter works on an aggregate basis and not per microflow.

- o Signal the Congestion-Level-Estimate - this is piggy-backed on the reservation reply. An egress gateway's interface is configured to know it is an egress gateway, so it always appends this to the RESV message. If the Congestion-Level-Estimate is unknown or is too stale, then the egress gateway can request the ingress gateway to send probes.

- o Packet colouring - for CL packets, set the DSCP and the ECN field to whatever has been agreed as appropriate for the next domain. By default the ECN field is set to the Not-ECT codepoint. See also the discussion in the Tunnelling section later.

- o Measure the rate - measure the rate of CL traffic from a particular ingress gateway (i.e. the rate for the CL-region-aggregate), when alerted (either explicitly or implicitly) that pre-emption may be required. The measured rate is reported back to the appropriate ingress gateway [RSVP-ECN].

4.4. Failures

If an 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 congestion from 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 pre-emption mechanism will kick in.

If a gateway fails then we would like regular RSVP procedures [RFC2205] to take care of things. With the local repair mechanism of [RFC2205], when a route changes the next RSVP PATH refresh message will establish path state along the new route, and thus attempt to re-establish reservations through the new ingress gateway. Essentially the same procedure is used as described earlier in this document, with the re-routed session treated as a new session request.

In more detail, consider what happens if an ingress gateway of the CL-region fails. Then RSVP routers upstream of it do IP re-routing to

a new ingress gateway. The next time the upstream RSVP router sends a PATH refresh message it reaches the new ingress gateway which therefore installs the associated RSVP state. The next RSVP RESV refresh will pick up the congestion-level-estimate from the egress gateway, and the ingress compares this with its threshold to decide whether to admit the new session. This could result in some of the flows being rejected, but those accepted will receive the full QoS.

An issue with this is that we have to wait until a PATH and RESV refresh messages are sent - which may not be very often - the default value is 30 seconds. [RFC2205] discusses how to speed up the local repair mechanism. First, the RSVP module is notified by the local routing protocol module of a route change to particular destinations, which triggers it to rapidly send out PATH refresh messages. Further, when a PATH refresh arrives with a previous hop address different from the one stored, then RESV refreshes are immediately sent to that previous hop. Where RSVP is operating hop-by-hop, ie on every router, then triggering the PATH refresh is easy as the node can simply monitor its local link. Thus, this fast local repair mechanism can be used to deal with failures upstream of the ingress gateway, with failures of the ingress gateway and with failures downstream of the egress gateway.

But where RSVP is not operating hop-by-hop (as is the case within the CL-region), it is not so easy to trigger the PATH refresh.

Unfortunately, this problem applies if an egress gateway fails, since it's very likely that an egress gateway is several IP hops from the ingress gateway. (If the ingress is several IP hops from its previous RSVP node, then there is the same issue.) The options appear to be:

- the ingress gateway has a link state database for the CL-region, so it can detect that an egress gateway has failed or became unreachable

- there is an inter-gateway protocol, so the ingress can continuously check that the egress gateways are still alive

- (default) do nothing and wait for the regular PATH/RESV refreshes (and, if needed, the pre-emption mechanism) to sort things out.

4.5. Admission of 'emergency / higher precedence' session

Section 4.1 describes how if the Congestion-Level-Estimate is greater than the CLE-threshold all new sessions are refused. But it is

unsatisfactory to block emergency calls, for instance. Therefore it is recommended that an 'emergency / higher precedence' call is admitted immediately even if the CLE-threshold is exceeded. Usually the network can actually handle the additional microflow, because there is a safety margin between the configured-admission-rate and the configured-pre-emption-rate. Normal call termination behaviour will soon bring the traffic level down below the configured-admission-rate. However, in exceptional circumstances the 'emergency / higher precedence' call may cause the traffic level to exceed the configured-pre-emption-rate; then the usual pre-emption mechanism will pre-empt enough (non 'emergency / higher precedence') microflows to bring the total traffic back under the configured-pre-emption-rate.

4.6. Tunnelling

It is possible to tunnel all CL packets across the CL-region. Although there is a cost of tunnelling (additional header on each packet, additional processing at tunnel ingress and egress), there are three reasons it may be interesting.

ECMP:

If the CL-region uses Equal Cost Multipath Routing (ECMP), then traffic between a particular pair of ingress and egress gateways may follow several different paths.

Why? An ECMP-enabled router runs an algorithm to choose between potential outgoing links, based on a hash of fields such as the packet's source and destination addresses - exactly what depends on the proprietary algorithm. Packets are addressed to the CL flow's end-point, and therefore different flows may follow different paths through the CL-region.

The problem is that if one of the paths is congested such that packets are being admission marked, then the Congestion-Level-Estimate measured by the egress gateway will be diluted by unmarked packets from other non-congested paths. Similarly, the measurement of the Sustainable-Aggregate-Rate will also be diluted.

One solution is to tunnel across the CL-region. Then the destination address (and so on) seen by the ECMP algorithm is that of the egress gateway, so all flows follow the same path.

Ingress gateway determination:

If packets are tunnelled from ingress gateway to egress gateway, the egress gateway can very easily determine in the datapath which ingress gateway a packet comes from (by simply looking at the source address of the tunnel header). This can facilitate operations such as computing the Congestion-Level-Estimate on a per ingress gateway basis.

End-to-end ECN:

The ECN field is used for PCN marking (see [PCN] for details), and so it needs to be re-set by the egress gateway to whatever has been agreed as appropriate for the next domain. Therefore if a packet arrives at the ingress gateway with its ECN field already set (ie not '00'), it may leave the egress gateway with a different value. Hence the end-to-end meaning of the ECN field is lost.

It is open to debate whether end-to-end congestion control is ever necessary within an end-to-end reservation. But if a genuine need is identified for end-to-end ECN semantics within a reservation, then one solution is to tunnel CL packets across the CL-region. When the egress gateway decapsulates them the original ECN field is recovered.

5. Potential future extensions

5.1. Mechanisms to deal with 'Flash crowds'

There is a time lag between the admission control decision (which depends on the Congestion-Level-Estimate during RSVP signalling during call set-up) and when the data is actually sent (after the called party has answered). In PSTN terms this is the time the phone rings. Normally the time lag doesn't matter much because (1) in the CL-region there are many flows and they terminate and are answered at roughly the same rate, and (2) the network can still operate safely when the traffic level is some margin above the configured-admission-rate.

A 'flash crowd' occurs when something causes many calls to be initiated in a short period of time - for instance a 'televote'. So there is a danger that a 'flash' of calls is accepted, but when the calls are answered and data flows the traffic overloads the network. There are various possible ways an operator could try to address the problem.

The simplest option is to do nothing; an operator relies on the pre-emption mechanism if there is a problem. This doesn't seem a good choice, as 'flash crowds' are reasonably common on the PSTN, unless the operator can ensure that nearly all "flash crowd" events are blocked in the access network and so do not impact on the CL-region.

A second option is to send 'dummy data' as soon as the call is admitted, thus effectively reserving the bandwidth whilst waiting for the called party to answer. Reserving bandwidth in advance means that the network cannot admit as many calls. For example, suppose sessions last 100 seconds and ringing for 10 seconds, the cost is a 10% loss of capacity. It may be possible to offset this somewhat by increasing the configured-admission-rate in the routers, but it would need further investigation.

A concern with this 'dummy data' option is that it may allow an attacker to initiate many calls that are never answered (by a cooperating attacker), so eventually the network would only be carrying 'dummy data'. The attack exploits that charging only starts when the call is answered and not when it is dialled. It may be possible to alleviate the attack at the session layer - for example, when the ingress gateway gets an RSVP PATH message it checks that the source has been well-behaved recently.

A third option is that the egress gateway limits the rate at which it sends out the Congestion-Level-Estimate, or limits the rate at which

calls are accepted by replying with a Congestion-Level-Estimate of 100% (this is the equivalent of 'call gapping' in the PSTN). There is a trade-off, which would need to be investigated further, between the degree of protection and possible adverse side-effects like slowing down call set-up.

A final option is to re-perform admission control before the call is answered. The ingress gateway monitors Congestion-Level-Estimate updates received from each egress. If it notices that a Congestion-Level-Estimate has risen above the CLE-threshold, then it terminates all unanswered calls through that egress (eg by instructing the session protocol to stop the 'ringing tone'). For extra safety the Congestion-Level-Estimate could be re-checked when the call is answered. A potential drawback for an operator that wants to emulate the PSTN is that the PSTN network never drops a 'ringing' PSTN call.

5.2. Multi-domain and multi-operator usage

This potential extension would eliminate the trust assumption (Section 2.2), so that the CL-region could consist of multiple domains run by different operators that did not trust each other. Then only the ingress and egress gateways of the CL-region would take part in the admission control procedure, i.e. at the ingress to the first domain and the egress from the final domain. The border routers between operators within the CL-region would only have to do bulk accounting - they wouldn't do per microflow metering and policing, and they wouldn't take part in signal processing or hold path state [Briscoe]. [Re-feedback] explains how a downstream domain can police that its upstream domain does not 'cheat' by admitting traffic when the downstream path is over-congested. [Re-PCN] proposes how to achieve this with the help of another recently proposed extension to ECN, involving re-echoing ECN feedback [Re-ECN].

5.3. Adaptive bandwidth for the Controlled Load service

The admission control mechanism described in this document assumes that each router has a fixed bandwidth allocated to CL flows. A possible extension is that the bandwidth is flexible, depending on the level of non-CL traffic. If a large share of the current load on a path is CL, then more CL traffic can be admitted. And if the greater share of the load is non-CL, then the admission threshold can be proportionately lower. The approach re-arranges sharing between classes to aim for economic efficiency, whatever the traffic load

matrix. It also deals with unforeseen changes to capacity during failures better than configuring fixed engineered rates. Adaptive bandwidth allocation can be achieved by changing the admission marking behaviour, so that the probability of admission marking a packet would now depend on the number of queued non-CL packets as well as the size of the virtual queue. The adaptive bandwidth approach would be supplemented by placing limits on the adaptation to prevent starvation of the CL by other traffic classes and of other classes by CL traffic.

5.4. Controlled Load service with end-to-end Pre-Congestion Notification

It may be possible to extend the framework to parts of the network where there are only a low number of CL microflows, i.e. the aggregation assumption (Section 2.2) doesn't hold. In the extreme it may be possible to operate the framework end-to-end, i.e. between end hosts. One potential method is to send probe packets to test whether the network can support a prospective new CL microflow. The probe packets would be sent at the same traffic rate as expected for the actual microflow, but in order not to disturb existing CL traffic a router would always schedule probe packets behind CL ones (compare [Breslau00]); this implies they have a new DSCP. Otherwise the routers would treat probe packets identically to CL packets. In order to perform admission control quickly, in parts of the network where there are only a few CL microflows, the Pre-Congestion marking behaviour for probe packets would switch from admission marking no packets to admission marking them all for only a minimal increase in load.

5.5. MPLS-TE

It may be possible to extend the framework for admission control of microflows into a set of MPLS-TE aggregates (Multi-protocol label switching traffic engineering). However it would require that the MPLS header could include the ECN field, which is not precluded by RFC3270.

6. Relationship to other QoS mechanisms

6.1. IntServ Controlled Load

The CL mechanism delivers QoS similar to Integrated Services controlled load, but rather better as queues are kept empty by driving admission control from a bulk virtual queue on each interface that can detect a rise in load before queues build, sometimes termed a virtual queue [AVQ, vq]. It is also more robust to route changes.

6.2. Integrated services operation over DiffServ

Our approach to end-to-end QoS is similar to that described in [RFC2998] for Integrated services operation over DiffServ networks. Like [RFC2998], an IntServ class (CL in our case) is achieved end-to-end, with a CL-region viewed as a single reservation hop in the total end-to-end path. Interior routers of the CL-region do not process flow signalling nor do they hold state. Unlike [RFC2998] we do not require the end-to-end signalling mechanism to be RSVP, although it can be.

Bearing in mind these differences, we can describe our architecture in the terms of the options in [RFC2998]. The DiffServ network region is RSVP-aware, but awareness is confined to (what [RFC2998] calls) the "border routers" of the DiffServ region. We use explicit admission control into this region, with static provisioning within it. The ingress "border router" does per microflow policing and sets the DSCP and ECN fields to indicate the packets are CL ones (i.e. we use router marking rather than host marking).

6.3. Differentiated Services

The DiffServ architecture does not specify any way for devices outside the domain to dynamically reserve resources or receive indications of network resource availability. In practice, service providers rely on subscription-time Service Level Agreements (SLAs) that statically define the parameters of the traffic that will be accepted from a customer. The CL mechanism allows dynamic reservation of resources through the DiffServ domain and, with the potential extension mentioned in Section 5.1, it can span multiple domains without active policing mechanisms at the borders (unlike DiffServ). Therefore we do not use the traffic conditioning agreements (TCAs) of the (informational) DiffServ architecture [RFC2475].

[Johnson] compares admission control with a 'generously dimensioned' DiffServ network as ways to achieve QoS. The former is recommended.

6.4. ECN

The marking behaviour described in this document complies with the ECN aspects of the IP wire protocol RFC3168, but provides its own edge-to-edge feedback instead of the TCP aspects of RFC3168. All nodes within the CL-region are upgraded with the admission marking and pre-emption marking of Pre-Congestion Notification, so the requirements of [Floyd] are met because the CL-region is an enclosed environment. The operator prevents traffic arriving at a node that doesn't understand CL by administrative configuration of the ring of gateways around the CL-region.

6.5. RTECN

Real-time ECN (RTECN) [RTECN, RTECN-usage] has a similar aim to this document (to achieve a low delay, jitter and loss service suitable for RT traffic) and a similar approach (per microflow admission control combined with an "early warning" of potential congestion through setting the CE codepoint). But it explores a different architecture without the aggregation assumption: host-to-host rather than edge-to-edge. We plan to document such a host-to-host framework in a parallel draft to this one, and to describe if and how [PCN] can work in this framework.

6.6. RMD

Resource Management in DiffServ (RMD) [RMD] is similar to this work, in that it pushes complex classification, traffic conditioning and admission control functions to the edge of a DiffServ domain and simplifies the operation of the interior nodes. One of the RMD modes uses measurement-based admission control, however it works differently: each interior node measures the user traffic load in the PHB traffic aggregate, and each interior node processes a local RESERVE message and compares the requested resources with the available resources (maximum allowed load minus current load).

Hence a difference is that the CL architecture described in this document has been designed not to require interaction between interior nodes and signalling, whereas in RMD all interior nodes are QoS-NSLP aware. So our architecture involves less processing in interior nodes, is more agnostic to signalling, requires fewer changes to existing standards and therefore works with existing RSVP as well as having the potential to work with future signalling protocols like NSIS.

RMD introduced the concept of Severe Congestion handling. The pre-emption mechanism described in the CL architecture has similar objectives but relies on different mechanisms.

It is planned to work together with the authors of [RMD] and that the next version of this draft and [PCN] will be co-authored with them.

6.7. RSVP Aggregation over MPLS-TE

Multi-protocol label switching traffic engineering (MPLS-TE) allows scalable reservation of resources in the core for an aggregate of many microflows. To achieve end-to-end reservations, admission control and policing of microflows into the aggregate can be achieved using techniques such as RSVP Aggregation over MPLS TE Tunnels as per [AGGRE-TE]. However, in the case of inter-provider environments, these techniques require that admission control and policing be repeated at each trust boundary or that MPLS TE tunnels span multiple domains.

7. Security Considerations

To protect against denial of service attacks, the ingress gateway of the CL-region needs to police all CL packets and drop packets in excess of the reservation. This is similar to operations with existing IntServ behaviour.

For pre-emption, it is considered acceptable from a security perspective that the ingress gateway can treat "emergency/military" CL flows preferentially compared with "ordinary" CL flows. However, in the rest of the CL-region they are not distinguished (nonetheless, our proposed technique does not preclude the use of different DSCPs at the packet level as well as different priorities at the flow level.). Keeping emergency traffic indistinguishable at the packet level minimises the opportunity for new security attacks. For example, if instead a mechanism used different DSCPs for "emergency/military" and "ordinary" packets, then an attacker could specifically target the former in the data plane (perhaps for DoS or for eavesdropping).

Further security aspects to be considered later.

8. Acknowledgements

The admission control mechanism evolved from the work led by Martin Karsten on 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]. Kennedy Cheng, Gabriele Corliano, Carla Di Cairano-Gilfedder, Kashaf Khan, Peter Hovell, Arnaud Jacquet and June Tay (BT) helped develop and evaluate this approach.

9. 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.

10. Changes from earlier versions of the draft

The main changes are:

From -00 to -01

The whole of the Pre-emption mechanism is added.

There are several modifications to the admission control mechanism.

From -01 to -02

The pre-congestion notification algorithms for admission marking and pre-emption marking are now described in [PCN].

There are new sub-sections in Section 4 on Failures, Admission of 'emergency / higher precedence' session, and Tunnelling; and a new sub-section in Section 5 on Mechanisms to deal with 'Flash crowds'.

11. Appendices

11.1. 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 (eg 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:

```
     0     1     2     3     4     5     6     7
  +-----+-----+-----+-----+-----+-----+-----+-----+
  |          DS FIELD, DSCP           | ECN FIELD |
  +-----+-----+-----+-----+-----+-----+-----+-----+

    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':

```
  +-----+-----+
  | ECN FIELD |
  +-----+-----+
    ECT   CE
     0     0         Not-ECT
     0     1         ECT(1)
     1     0         ECT(0)
     1     1         CE
```

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.

11.2. Appendix B: What is distributed measurement-based admission control?

This Appendix briefly explains what distributed measurement-based admission control is [Breslau99].

Traditional admission control algorithms for 'hard' real-time services (those providing a firm delay bound for example) guarantee QoS by using 'worst case analysis'. Each time a flow is admitted its traffic parameters are examined and the network re-calculates the remaining resources. When the network gets a new request it therefore knows for certain whether the prospective flow, with its particular parameters, should be admitted. However, parameter-based admission control algorithms result in under-utilisation when the traffic is bursty. Therefore 'soft' real time services - like Controlled Load - can use a more relaxed admission control algorithm.

This insight suggests measurement-based admission control (MBAC). The aim of MBAC is to provide a statistical service guarantee. The

classic scenario for MBAC is where each node participates in hop-by-hop admission control, characterising existing traffic locally through measurements (instead of keeping an accurate track of traffic as it is admitted), in order to determine the current value of some parameter e.g. load. Note that for scalability the measurement is of the aggregate of the flows in the local system. The measured parameter(s) is then compared to the requirements of the prospective flow to see whether it should be admitted.

MBAC may also be performed centrally for a network, it which case it uses centralised measurements by a bandwidth broker.

We use distributed MBAC. "Distributed" means that the measurement is accumulated for the 'whole-path' using in-band signalling. In our case, this means that the measurement of existing traffic is for the same pair of ingress and egress gateways as the prospective microflow.

In fact our mechanism can be said to be distributed in three ways: all nodes on the ingress-egress path affect the Congestion-Level-Estimate; the admission control decision is made just once on behalf of all the nodes on the path across the CL-region; and the ingress and egress gateways cooperate to perform MBAC.

11.3. Appendix C: Calculating the Exponentially weighted moving average (EWMA)

At the egress gateway, for every CL packet arrival:

[EWMA-total-bits]n+1 = (w * bits-in-packet) + ((1-w) * [EWMA-total-bits]n)

[EWMA-AM-bits]n+1 = (B * w * bits-in-packet) + ((1-w) * [EWMA-AM-bits]n)

Then, per new flow arrival:

[Congestion-Level-Estimate]n+1 = [EWMA-AM-bits]n+1 / [EWMA-total-bits]n+1

where

EWMA-total-bits is the total number of bits in CL packets, calculated as an exponentially weighted moving average (EWMA)

EWMA-AM-bits is the total number of bits in CL packets that are Admission Marked, again calculated as an EWMA.

B is either 0 or 1:

B = 0 if the CL packet is not admission marked

B = 1 if the CL packet is admission marked

w is the exponential weighting factor.

Varying the value of the weight trades off between the smoothness and responsiveness of the Congestion-Level-Estimate. However, in general both can be achieved, given our original assumption of many CL microflows and remembering that the EWMA is calculated on the basis of aggregate traffic between the ingress and egress gateways.
There will be a threshold inter-arrival time between packets of the same aggregate below which the egress will consider the estimate of the Congestion-Level-Estimate as too stale, and it will then trigger generation of probes by the ingress.

The first two per-packet algorithms can be simplified, if their only use will be where the result of one is divided by the result of the other in the third, per-flow algorithm.

[EWMA-total-bits]'n+1 = bits-in-packet + (w' * [EWMA- total-bits]n)

[EWMA-AM-bits]'n+1 = (B * bits-in-packet) + (w' * [EWMA-AM-bits]n)

where w' = (1-w)/w.

If w' is arranged to be a power of 2, these per packet algorithms can be implemented solely with a shift and an add.

12. References

A later version will distinguish normative and informative references.

[AGGRE-TE] Francois Le Faucheur, Michael Dibiasio, Bruce Davie, Michael Davenport, Chris Christou, Jerry Ash, Bur Goode, 'Aggregation of RSVP Reservations over MPLS TE/DS-TE Tunnels', draft-ietf-tsvwg-rsvp-dste-00 (work in progress), July 2005

[ANSI.MLPP.Spec] American National Standards Institute, "Telecommunications- Integrated Services Digital Network (ISDN) - Multi-Level Precedence and Pre-emption (MLPP) Service Capability", ANSI T1.619-1992 (R1999), 1992.

[ANSI.MLPP.Supplement] American National Standards Institute, "MLPP Service Domain Cause Value Changes", ANSI ANSI T1.619a-1994 (R1999), 1990.

[AVQ] S. Kunniyur and R. Srikant "Analysis and Design of an Adaptive Virtual Queue (AVQ) Algorithm for Active Queue Management", In: Proc. ACM SIGCOMM'01, Computer Communication Review 31 (4) (October, 2001).

[Breslau99] L. Breslau, S. Jamin, S. Shenker "Measurement-based admission control: what is the research agenda?", In: Proc. Int'l Workshop on Quality of Service 1999.

[Breslau00] L. Breslau, E. Knightly, S. Shenker, I. Stoica, H. Zhang "Endpoint Admission Control: Architectural Issues and Performance", In: ACM SIGCOMM 2000

[Briscoe] Bob Briscoe and Steve Rudkin, "Commercial Models for IP Quality of Service Interconnect", BT Technology Journal, Vol 23 No 2, April 2005.

[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).

[EMERG-RQTS] Carlberg, K. and R. Atkinson, "General Requirements for Emergency Telecommunication Service (ETS)", RFC 3689, February 2004.

[EMERG-TEL] Carlberg, K. and R. Atkinson, "IP Telephony Requirements for Emergency Telecommunication Service (ETS)", RFC 3690, February 2004.

[Floyd] S. Floyd, 'Specifying Alternate Semantics for the Explicit Congestion Notification (ECN) Field', draft-floyd-ecn-alternates-02.txt (work in progress), August 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)

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

[Johnson] DM Johnson, 'QoS control versus generous dimensioning', BT Technology Journal, Vol 23 No 2, April 2005

[PCN] B. Briscoe, P. Eardley, D. Songhurst, F. Le Faucheur, A. Charny, V. Liatsos, S. Dudley, J. Babiarz, K. Chan. 'Pre-Congestion Notification marking', draft-briscoe-tsvwg-cl-phb-01 (work in progress), March 2006.

[Re-ECN] Bob Briscoe, Arnaud Jacquet, Alessandro Salvatori, 'Re-ECN: Adding Accountability for Causing Congestion to TCP/IP', draft-briscoe-tsvwg-re-ecn-tcp-01 (work in progress), March 2006.

[Re-feedback] Bob Briscoe, Arnaud Jacquet, Carla Di Cairano-Gilfedder, Andrea Soppera, 'Re-feedback for Policing Congestion Response in an Inter-network', ACM SIGCOMM 2005, August 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.

[Reid]   ABD Reid, 'Economics and scalability of QoS solutions', BT Technology Journal, Vol 23 No 2, April 2005

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

[RFC2309]   Braden, B., et al., "Recommendations on Queue Management and Congestion Avoidance in the Internet", RFC 2309, April 1998.

[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, 'A framework 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.

[RFC2998]   Bernet, Y., Yavatkar, R., Ford, P., 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.

[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.

[RFC3270]   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.

[RMD] Attila Bader, Lars Westberg, Georgios Karagiannis, Cornelia Kappler, Tom Phelan, 'RMD-QOSM - The Resource Management in DiffServ QoS model', draft-ietf-nsis-rmd-03 Work in Progress, June 2005.

[RSVP-ECN] Francois Le Faucheur, Anna Charny, Bob Briscoe, Philip Eardley, Joe Barbiaz, Kwok-Ho Chan, 'RSVP Extensions for Admission Control over DiffServ using Pre-congestion Notification', draft-lefaucheur-rsvp-ecn-00 (work in progress), October 2005.

[RTECN] Babiarz, J., Chan, K. and V. Firoiu, 'Congestion Notification Process for Real-Time Traffic', draft-babiarz-tsvwg-rtecn-04 Work in Progress, July 2005.

[RTECN-usage] Alexander, C., Ed., Babiarz, J. and J. Matthews, 'Admission Control Use Case for Real-time ECN', draft-alexander-rtecn-admission-control-use-case-00, Work in Progress, February 2005.

[vq] Costas Courcoubetis and Richard Weber "Buffer Overflow Asymptotics for a Switch Handling Many Traffic Sources" In: Journal Applied Probability 33 pp. 886--903 (1996).

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

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
300 Apollo Drive
Chelmsford, MA 01824
USA
Email: acharny@cisco.com

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

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

Intellectual Property Statement

Disclaimer of Validity

Copyright Statement