TSVWG
Internet Draft
draft-briscoe-tsvwg-cl-architecture-00.txt
Expires: January 2006

B. Briscoe
G. Corliano
P. Eardley
P. Hovell
A. Jacquet
D. Songhurst
BT
July 11, 2005

An architecture for edge-to-edge controlled load service using distributed measurement-based admission control
draft-briscoe-tsvwg-cl-architecture-00.txt

Status of this Memo

Copyright Notice

Abstract

This document describes an architecture to achieve a Controlled Load (CL) service edge-to-edge, i.e. within a particular region of the Internet, by using distributed measurement-based admission control. The measurement made is of CL packets that have their Congestion Experienced (CE) codepoint set as they travel across the edge-to-edge region. Setting the CE codepoint, which is under the control of a new Per Hop Behaviour (CL-ramp-PHB, defined in draft-briscoe-tsvwg-cl-phb-00.txt), provides an "early warning" of potential congestion. This information is used by the ingress node of the edge-to-edge region to decide whether to admit a new CL microflow.

A use case is described which shows how the PHB is a fundamental building block in the edge-to-edge architecture, and in turn how this is a building block within a broader QoS architecture achieving an end-to-end CL service.

Table of Contents

1. Introduction

1.1. Summary

This document describes an architecture to achieve a controlled load service edge-to-edge, i.e. within a particular region of the Internet, using distributed measurement-based admission control. 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 streaming real-time media.

The architecture described in this document achieves edge-to-edge controlled load service using a new Per Hop Behaviour (PHB) as a fundamental building block. In turn, an end-to-end CL service would use this architecture as a building block within a broader QoS architecture. The PHB, edge-to-edge and end-to-end aspects are now briefly introduced in turn.

The new PHB, called CL-ramp-PHB, is defined in [CL-PHB]. Network nodes that implement the differentiated services (DS) enhancements to IP use a codepoint in the IP header to select a PHB as the specific forwarding treatment for that packet [RFC2474, RFC2475]. The CL-ramp-PHB is different from PHBs defined so far, in that it defines Explicit Congestion Notification (ECN) marking semantics as part of the PHB. A node in the CL-region sets the Congestion Experienced (CE) codepoint in the IP header as an "early warning" of potential congestion, and aims to do so before there is any significant build-up of CL packets in the queue.

To achieve the CL service edge-to-edge, ie within a region of the Internet - which we call CL-region (defined below) - distributed measurement-based admission control is used. All nodes within the CL-region run the CL-ramp-PHB. The measurement is of the CL packets that have had their CE codepoint set as they travel across the CL-region. Since any node in the CL-region may set the CE codepoint, the measurement is distributed. The measurement is recorded by the egress node of the CL-region. The egress node calculates the bits in these CE packets as a fraction of the bits in all the CL packets, as an exponentially weighted moving average (which we term Congestion-Level-Estimate). Depending on the value of Congestion-Level-Estimate, the ingress node of the CL-region decides whether to admit a new CL microflow. Since setting the CE codepoint is an "early warning" of potential congestion (ie before there is any significant build-up of CL packets in the queue), the admission control procedure means that previously accepted CL microflows will suffer minimal queuing delay, jitter and loss - exactly the requirements of real time traffic.

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 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 - as indeed we assume in Sections 2 and 3. [RFC2998] and our approach are compared further in Section 5.

1.2. Key features

In this section we discuss some of the key aspects of the edge-to-edge architecture.

One key feature of our approach revolves around the use of Explicit Congestion Notification (ECN) [RFC3168] to indicate that the amount of packets flowing is getting close to the engineered capacity. Note that ECN operates across the CL-region, ie edge-to-edge, and not host-to-host as in [RFC3168].

The new PHB, CL-ramp-PHB, is designed to provide an "early warning" of potential congestion. It assumes that a new microflow won't move the CL-region directly from no congestion to overload; there will always be an intermediate stage where a new CL microflow causes CL

packets to have their CE codepoint set but still be delivered without significant delay. This assumption is valid for core and backbone networks but is unlikely to be valid in access networks where the granularity of an individual call becomes significant.

Note that the CL-region can potentially span multiple domains. Indeed, over time CL-regions may incrementally grow and merge, and could eventually become a single CL-region encompassing all core and backbone networks, providing Internet-wide controlled load service in concert with stateful admission control mechanisms at the very edges of the Internet.

It is also possible for a CL-region to include domains run by different operators. The border routers between operators within the CL-region only have to do bulk accounting - per microflow metering and policing is not needed. Section 4.1 discusses further.

CL-packets are marked with a Differentiated Services Codepoint (DSCP), so that nodes in the CL-region can distinguish the CL packets from non-CL ones [RFC2474] and know that the CL-ramp-PHB is required.

However, note that we do not use the traffic conditioning agreements (TCAs) of the (informational) Diffserv architecture [RFC2475], in which operators in practice rely on subscription-time Service Level Agreements (SLAs) that statically define the parameters of the traffic that will be accepted from a customer. Operators deploying our mechanism do not need to make a fixed assignment of capacity because the division of bandwidth between CL and non-CL traffic can be flexible.

Our edge-to-edge architecture uses dynamic admission control: the closed feedback loop between the ingress and egress nodes of the CL-region. The key advantage of controlling the load dynamically rather than with TCAs is that the latter can fail catastrophically. 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 a resource near the egress, even though it is within contract at the ingress [Reid]. Even though networks can be engineered to make such failures rare, when they occur all inelastic flows through the congested resource fail catastrophically. This is also why in our approach the egress node of the CL-region calculates the Congestion-Level-Estimate separately for CL packets from each ingress node.

Finally, it is assumed that the end systems react properly to non-CL packets that are dropped or have their CE codepoint set, otherwise

new CL microflows calls may get unfairly blocked. How to police this is out of scope of this document.

1.3. Benefits

We believe that the mechanism described in this document has several advantages, which we briefly explain with reference to the key features described above:

- 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 "early warning" of potential congestion. The guarantee is at least as strong as with Intserv Controlled Load (Section 5 mentions why the guarantee may be somewhat better), but without its scalability problems [RFC2208].

- 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.

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

- It can be deployed incrementally, network by network. Not all the networks on the end-to-end path need to have it deployed. Two CL-regions can be separated by a network that uses another QoS mechanism (eg MPLS), or where they are adjacent can merge to become a single CL-region.

- It can work between operators, ie the CL-region can include domains run by different operators. This is scalable because there is only bulk metering at the inter-operator interface; there is no need for per microflow accounting or policing.

1.4. Standardisation requirements

The architecture described in this document has two new standardisation requirements: for a new PHB, as described in [CL-

PHB], and for the end-to-end signalling protocol to carry the Congestion-Level-Estimate report (eg with RSVP, the RESV message must carry a new opaque object across the CL-region). Other than these two things, the arrangement uses existing standards throughout although, as mentioned above, not in their usual architecture. Section 5 discusses standardisation issues further.

This document is INFORMATIONAL.

1.5. Terminology

- Ingress node: a node which is an ingress gateway to the CL-region. A CL-region may have several ingress nodes.

- Egress node: a node which is an egress gateway from the CL-region. A CL-region may have several egress nodes.

- 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 nodes run the CL-ramp-PHB and all traffic enters/leaves through an ingress/egress node. A CL-region is a DS region (a DS region is either a single DS domain or set of contiguous DS domains), but note that the CL-region does not use the traffic conditioning agreements (TCAs) of the (informational) Diffserv architecture.

- CL-ramp-PHB: A new Per Hop Behaviour, described in [CL-PHB].

- Congestion-Level-Estimate: the bits in CL packets that have the CE codepoint set, divided by the bits in all CL packets. It is calculated as an exponentially weighted moving average. It is calculated by an egress node for CL packets from a particular ingress node.

```
   ______________________________
  /                              \
 /                                \
|-------|    |--------|    |-------|
|Ingress|----|Interior|----|Egress |
| node  |    | node   |    | node  |
|-------|    |--------|    |-------|
 \                                /
  \______________________________/

< ---------- CL-region ----------- >
```

Figure 1: Sample edge-to-edge configuration and terminology

1.6. Structure of rest of document

Section 2 describes a use case, with further details in Section 3 and extensions in Section 4. Section 5 discusses standardisation aspects.

2. Use case

In this section we outline a usage scenario to illustrate how our mechanism works. It is intended to show how the main features fit together to deliver QoS, with further details in Section 3.

Our QoS mechanism operates over a CL-region. For now we assume that it consists of one domain whilst in Section 4.1 we extend it to the multi-domain case, including where different operators run the domains. So our scenario consists of two end hosts, each connected to their own access networks, which are linked by the CL-region. We require some other method, for instance IntServ, to be used outside the CL-region to provide QoS. For now we assume that the end-to-end signalling protocol is RSVP; other protocols are considered in Section 3.2. 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 nodes but are carried transparently across all the interior nodes. Hence, the ingress and egress nodes hold per microflow state, whilst no state is kept by the interior nodes.

Section 2.1 describes a restricted scenario where the CL behaviour aggregate is assigned a fixed amount of bandwidth. This is equivalent to the case today with the DS architecture: a subscription-time Service Level Agreement (SLA) statically defines the amount of bandwidth reserved for a particular behaviour aggregate. Section 2.2 describes the more general case where there is no fixed allocation to CL traffic.

Each node in the CL-region runs an algorithm to determine whether to set the CE codepoint of a particular CL packet. In our description we assume that a bulk token bucket is used (other implementations are possible), and that tokens are added when packets are queued and are consumed at a fixed rate. The idea is that an excess of tokens is seen before the queue of CL packets has got long enough to cause the CL packets to suffer a significant delay - the algorithms are explained more fully below and are slightly different in Sections 2.1 and 2.2. Note that the same token bucket is used for all the CL packets, ie it operates in bulk on the CL behaviour aggregate and not per microflow.

```
 ___    ____    _______________________________________    ____    ___
|   |  |    |  |                                       |  |    |  |   |
|   |  |    |  |Ingress   Interior   Interior    Egress|  |    |  |   |
|   |  |    |  | node      node       node       node  |  |    |  |   |
|   |  |    |  |------|   |------|   |------|   |------|  |    |  |   |
|   |  |    |  | CL-  |   | CL-  |   | CL-  |   |      |  |    |  |   |
|   |..|    |..| PHB  |...| PHB  |...| PHB  |...| Meter|..|    |..|   |
|   |  |    |  |------|   |------|   |------|   |------|  |    |  |   |
|   |  |    |  |  \                                 /  |  |    |  |   |
|   |  |    |  |   \                               /   |  |    |  |   |
|   |  |    |  |    --<------------<-----------<--     |  |    |  |   |
|   |  |    |  |                                       |  |    |  |   |
|___|  |____|  |_______________________________________|  |____|  |___|

Sx     Access               CL-region                    Access    Rx
End    Network                                           Network   End
Host                                                               Host

                <------ edge-to-edge signalling ------>
                          (admission control)

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

Figure 2: Overall QoS architecture

2.1. Configured bandwidth allocation to the CL behaviour aggregate

Each node in the CL-region has a fixed rate (bandwidth) allocated to CL traffic, under the control of management configuration. Tokens are consumed at a fixed rate that is slightly slower than the configured rate, and added when packets are queued. This means that the amount of tokens starts to increase before the actual queue builds up but when it is in danger of doing so soon; hence it can be used as an "early warning" of potential congestion. The probability that a node sets the CE codepoint of a CL packet depends on the number of tokens in the bucket. Below one threshold value of the number of tokens no packets have their CE codepoint set and above the second they all do; in between, the probability increases linearly.

We now describe how setting the CE codepoint influences admission control by the ingress node. For ease of description we imagine that packets are already flowing. Each egress meters whether a CL packet has its CE codepoint set. We assume that initially the traffic load is such that there are no CE packets.

Next a source tries to set up a new CL microflow. The RSVP PATH message is processed by the ingress and egress nodes and PATH state is installed in these two routers. When the RSVP RESV message travels back from the receiving end host, the egress node adds on an RSVP object which states that currently no CL packets have their CE codepoint set. Hence the ingress node admits the new CL microflow, and the RESV message continues on to the source.

We imagine that this new microflow results in one (or more) of the interior nodes starting to set the CE codepoint of CL packets because their arrival rate is nearing the configured rate. The egress calculates - as an exponentially weighted moving average - the fraction of CL packets from a particular ingress node that have their CE codepoint set (or rather the calculation is done according to the bits in those packets). This Congestion-Level-Estimate provides an estimate of how near the CL-region is getting to a load where the CL traffic will start suffering significant delays. Note that the metering is done separately per ingress node, because (as discussed in Section 1.2) there may be sufficient capacity on all the nodes on the path between one ingress node and a particular egress, but not from a second ingress.

The next time a source tries to set up a CL microflow, the egress informs the ingress node about the relevant Congestion-Level-Estimate; this is included as an opaque object within the RSVP RESV message. If it is greater than some threshold value then the ingress refuses the request, otherwise it is accepted and the RSVP RESV continues to the source end host.

It is also possible for an egress node to get a RSVP RESV message and not know what Congestion-Level-Estimate is. For example, if there are no CL microflows at present between the relevant ingress and egress nodes. In this case the egress requests the ingress to send probe packets, from which it can initialise its meter.

Having explained how the admission control decision is reached we now look at an on-going data microflow. The source sends CL packets, which arrive at the ingress node. The ingress uses a normal 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 PATH message.) The ingress sets the DSCP appropriately and the ECN field to ECT (ECN-Capable Transport). The CL packets now travel across the CL-region, with the CE codepoint getting set if necessary. Also, appropriate queue scheduling is needed in each node to ensure that CL traffic gets its configured bandwidth. For instance, a Weighted Round Robin scheduler could be used.

2.2. Flexible bandwidth allocation to CL behaviour aggregate

The set-up is similar to the previous sub-section, except that nodes in the CL-region do not allocate a fixed bandwidth to CL flows. As a consequence, the algorithm for setting the CE codepoint is slightly altered.

Tokens are consumed at a fixed rate that is slightly slower than the (total) outgoing service rate, and added when packets are queued. The probability that a node sets the CE codepoint of a CL packet depends on the number of tokens in the bucket *plus* the number of queued non-CL packets. Below one threshold value of this sum no packets have their CE codepoint set and above the second they all do; in between, the probability increases linearly.

Note that the probability reflects the load of both CL and non-CL traffic. The reason is to ensure a 'fair balance' between the two classes, by rejecting CL session requests if non-CL demand is very high. Alternatively, if the number of queued non-CL packets is not

included, then the admission of a CL microflow is independent of the amount of non-CL traffic.

The admission control procedure is as in the previous sub-section. As regards queue scheduling, CL packets are always scheduled ahead of non-CL ones, in order to minimise their delay and jitter, and FIFO (First In First Out) queuing is used to prevent reordering within a CL microflow. This is more restrictive than in the previous sub-section, which we believe is necessary now the arrival rate of CL packets is unknown.

3. Details

In this section we first concentrate on the details about packet processing in nodes in the CL-region, before looking more briefly at issues associated with the signalling for admission control.

3.1. Packet processing

A network operator upgrades normal IP routers by:

o Adding functionality related to admission control to all its ingress and egress nodes

o Adding appropriate queuing and scheduling behaviour to its nodes, including the ability to set the CE codepoint "early".

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

3.1.1. Ingress nodes

Ingress nodes perform the following tasks:

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

o Police - check that the microflow is conformant with what has been agreed (ie the flow keeps to its agreed data rate). If necessary, the suggested action is that packets are marked to Best Effort.

o Packet colouring - for CL microflows, set the DSCP appropriately and set the ECN field to ECT(0) or ECT(1)

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

3.1.2. Interior nodes

Interior nodes do the following tasks:

- o Examine the DSCP - to see if it's a CL packet

- o Enqueue - CL and non-CL packets are put into logically separate queues; if required, a CL packet can pre-empt non-CL packet(s) in the total buffer (see below).

- o Non-CL packets are handled as usual. A RED algorithm [RFC2309] is used to decide whether to drop packets or, if they are ECN-capable, set their CE codepoint.

- o CL packets have their CE codepoint set according to what is essentially a token bucket algorithm (see below).

- o Dequeue - any CL packet is always dequeued before a non-CL packet. Within the CL class scheduling is FIFO. There may be a hierarchy of non-CL classes, this is out of scope.

Queuing:

Although CL and non-CL packets are put into logically separate queues, implementations in practice share the same buffer space. If the buffer is full then an incoming non-CL packet is dropped, whilst an incoming CL packets is queued and sufficient of the newest non-CL packet(s) are dropped. In the unlikely event that the buffer is full of CL packets, then the newest CL packet is discarded (ie tail drop). Because of the admission procedure this should be rare, but it is needed to protect the network in case of misconfiguration for instance.

Setting the CE codepoint:

Tokens are added when CL packets are queued and are consumed at a fixed rate related to the outgoing service rate.

When a CL packet arrives the token bucket is updated as follows:

[CL-bucket-level]n+1 = [CL-bucket-level]n + CL-packet-size - (service-bit-rate * time * safety-factor)

Where

CL-bucket-level is the amount of tokens in the token bucket. It is constrained to lie between 0 and a fixed upper limit

time is the time elapsed since CL-bucket-level was last updated

safety-factor is > 1 and gives the "early warning" of potential congestion

service-bit-rate is

  either the configured bit rate for CL traffic - for the fixed bandwidth case (ie Section 2.1),

  or the outgoing service rate for all traffic - for the flexible bandwidth case (ie Section 2.2).

CL packets have their CE codepoint set with a probability that depends on the number of non-CL packets in the queue, as well as the number of tokens in a token bucket.

When a CL packet arrives, the probability that the node sets its CE codepoint is determined as follows:

if [CL-bucket-level]n+1 + (A * smoothed-non-CL-queue-length) < min-threshold

  Probability-CE-codepoint-set = 0

if [CL-bucket-level]n+1   + (A * smoothed-non-CL-queue-length) > max-threshold

  Probability-CE-codepoint-set = 1

otherwise

  Probability-CE-codepoint-set = (CL-bucket-level - min-threshold) / (max-threshold - min-threshold)

Where

max-threshold > min-threshold

max-threshold <= the fixed upper limit of CL-bucket-level

smoothed-non-CL-queue-length is the number of bits in packets in the non-CL queue, smoothed as an exponentially weighted moving average (EWMA)

A is either 0 or 1:

A = 0 for the fixed bandwidth case (ie Section 2.1),

A = 1 for the flexible bandwidth case (ie Section 2.2).

3.1.3. Egress nodes

Egress nodes do the following tasks:

o Metering - for CL packets, calculating the fraction of the total bits which are in CE packets. The calculation is done as an exponentially weighted moving average. A separate calculation is made for CL packets from each ingress router.

o Packet colouring - for CL packets, set the DSCP and the ECN field to whatever has been agreed as appropriate for the next domain.

An egress node getting a CL packet first determines which ingress node that packet has come from. The necessary details are gathered from the RSVP PATH message (previous RSVP hop, ie ingress node, vs. filter spec). It then updates the two meters associated with that ingress node. The meters work on an aggregate basis, and not per microflow.

For every CL packet arrival:

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

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

[Congestion-Level-Estimate]n+1 = [EWMA-CE-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-CE-bits is the total number of bits in CL packets where the packet has its CE codepoint set, again calculated as an EWMA.

B is either 0 or 1:

B = 0 if the CL packet does not have its CE codepoint set

B = 1 if the CL packet has its CE codepoint set

w is the exponential weighting factor.

Varying the value of the weight trades off between the smoothness and responsiveness of the estimate of the percentage of CE packets. 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 probing by the ingress.
For packet colouring, by default the ECN field is set to the Not-ECT codepoint. Note that this results in the loss of the end-to-end meaning of the ECN field. It can usually be assumed that end-to-end congestion control is unnecessary within an end-to-end reservation. But if a genuine need is identified for end-to-end ECN semantics within a reservation, then an alternative is to tunnel CL packets across the CL-region, or to agree an extension to end-to-end signalling to indicate that the microflow uses an ECN-capable transport. We do not recommend such apparently unnecessary complexity.

3.2. Signalling

The admission control procedure involves signalling between the ingress and egress nodes. The following new messages are needed:-

- Egress to ingress: piggy-backed on reservation reply: this is the current value of Congestion-Level-Estimate. An egress node is configured to know it is an egress node, so it always appends this to the reservation response. A flag in this message can indicate the value is unknown, in order to trigger probing by the ingress.

- Ingress to egress: probe: this is a probe packet

The description in the earlier sections has assumed that RSVP signalling is used. In this case, the first bullet requires standardisation so that the RSVP RESV message can carry a new opaque object with the load report.

However, there are several other possible signalling protocols, for instance using NSIS. It would therefore be sensible to ensure that the new signalling messages do not constrain the choice of end-to-end QoS mechanism nor how the end-to-end and edge-to-edge (ie ingress-to-egress) mechanisms interact. As an example on the latter point, with RSVP the PATH message is forwarded immediately to the next domain, with the Congestion-Level-Estimate report only being calculated when the RESV returns, at which point it can be piggy-backed on to the RESV and sent to the ingress. In other cases, it may be that admission control is performed before the signalling message is forwarded to the next domain.

4. Extensions

4.1. Multi-domain and multi-operator usage

The CL-region can consist of multiple domains. Then only the ingress and egress nodes of the CL-region take part in the admission control procedure, ie at the ingress to the first domain and the egress from the final domain. Note that domain border nodes within the CL-region do not take part in signal processing or hold path state.

The multiple domains can even be run by different operators. The border routers between operators within the CL-region only have to do bulk accounting - per microflow metering and policing is not needed [Briscoe]. This is possible even when the operators do not trust each other. In a later version of the draft we will explain how a downstream domain can police that its upstream domain does not 'cheat' by admitting traffic when the downstream path is over-congested [Re-feedback].

4.2. Variable bit rate sources

So far we have assumed that the real time inelastic sources operate at a constant bit rate. We have determined under what conditions it is possible to handle variable bit rate (VBR) sources. The simplest approach is an algorithm that decides whether to set the CE codepoint using a service rate much less than the real service rate (ie allowing an extra safety margin); the network can still operate efficiently when resources are shared between CL and non-CL flows. This approach assumes that the sources are statistically independent.

4.3. Starvation prevention

According to the particular traffic levels it may sometimes be possible for either the non-CL or CL traffic to be starved. An algorithm to prevent starvation will be documented in a future draft.

5. Relationship to other QoS mechanisms

5.1. Standardisation requirements

Standardisation of two functions is needed:

- First, a new per hop behaviour is required (CL-ramp-PHB), which is described in [CL-PHB]. The corresponding DSCP needs to be RECOMMENDED rather than EXP/LU (experimental / local use), to enable multi-domain operation and vendor interoperability. This document is a use case of CL-ramp-PHB.

- Signalling between the ingress and egress nodes and its interaction with the end-to-end QoS mechanism, for instance RSVP or NSIS. For instance, given RSVP's capabilities to carry opaque objects, define an object to carry the Congestion-Level-Estimate report. Probe packets are simply data addressed to the egress gateway and require no protocol standardisation, although best practice is required for their number, size and rate.

5.2. 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 bulk token buckets 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.

5.3. 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. Also, we do not use the DS architecture (see Section 5.4).

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 either static provisioning or explicit signalling (corresponding to the configured and flexible bandwidth cases of Sections 2.1 and 2.2 respectively). The ingress "border router" does per microflow policing and sets the correct DSCP (ie we use router marking rather than host marking).

5.4. Differentiated Services

The DS 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 and unlike Diffserv it can span multiple domains without active mechanisms at the borders. 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.

5.5. ECN

CL 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 ECN. All nodes within a particular CL-region are upgraded with the CL mechanism, so the requirements of [Floyd] are met. The operator prevents traffic arriving at a node that doesn't understand CL by administrative configuration of the ring of gateways around the region. Where a region of nodes that understand CL spans multiple domains, the operators contract with each other to surround the

region by gateways to prevent CL traffic being handled by nodes that do not understand it.

5.6. 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 has a different architecture: host-to-host (rather than edge-to-edge). [CL-PHB] defines a new PHB, CL-step-PHB, that should be suitable; its algorithm is similar to CL-ramp-PHB, but setting the CE codepoint is either 'on' or 'off'. Only probe packets use the CL-step-PHB, whilst data uses the Expedited Forwarding PHB [RFC3246].

5.7. 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 DS 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 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.

5.8. MPLS-TE

Multi-protocol label switching traffic engineering (MPLS-TE) allows reservation of resources for an aggregate of many flows. However, it still requires admission control and policing (using a bandwidth manager) of microflows into the aggregate. This must be repeated at each trust boundary. The present technique could be used for admission control of microflows into a set of MPLS-TE aggregates. They may span multiple domains without requiring per-microflow processing at the trust boundaries. However it would require that the MPLS header could include the ECN field.

6. Security Considerations

To protect against denial of service attacks, the ingress node of the CL-region needs to police all CL packets and drop packets in excess of the reservation.

Further security aspects to be considered later.

7. Acknowledgements

We thank Joe Babiarz for very helpful discussion about this document and [RTECN].

This work 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].

8. 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 (either individually or collectively at gqs@jungle.bt.co.uk).

9. References

A later version will distinguish normative and informative references.

[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).

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

[CL-PHB] B. Briscoe, G. Corliano, P. Eardley, P. Hovell, A. Jacquet, D. Songhurst, "The Controlled Load per hop behaviour", draft-briscoe-tsvwg-cl-phb-00.txt (work in progress), July 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. 941ù952 (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)

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

[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.

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

[RFC2208] F. Baker et al, "Resource ReSerVation Protocol (RSVP) --- Version 1 Applicability Statement; Some Guidelines on Deployment" RFC2208 (January, 1997)

[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, "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.

[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.

[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.

[RTECN] Babiarz, J., Chan, K. and V. Firoiu, 'Congestion Notification Process for Real-Time Traffic', draft-babiarz-tsvwg-rtecn-03" Work in Progress, February 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

Peter Hovell
BT Research
B54/69, Sirius House
Adastral Park
Martlesham Heath
Ipswich, Suffolk
IP5 3RE
United Kingdom
Email: peter.hovell@bt.com

Gabriele Corliano
BT Research
B54/70, Sirius House
Adastral Park
Martlesham Heath
Ipswich, Suffolk
IP5 3RE
United Kingdom
Email: gabriele.2.corliano@bt.com

Arnaud Jacquet
BT Research
B54/70, Sirius House
Adastral Park
Martlesham Heath
Ipswich, Suffolk
IP5 3RE
United Kingdom
Email: arnaud.jacquet@bt.com

Intellectual Property Statement

Disclaimer of Validity

This document and the information contained herein are provided on an "AS IS" basis and THE CONTRIBUTOR, THE ORGANIZATION HE/SHE REPRESENTS OR IS SPONSORED BY (IF ANY), THE INTERNET SOCIETY AND THE INTERNET ENGINEERING TASK FORCE DISCLAIM ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Copyright Statement