Internet Draft **Francois Le Faucheur**
Anna Charny
Cisco Systems, Inc.

Bob Briscoe
Phil Eardley
BT

Joe Barbiaz
Kwok-Ho Chan
Nortel

draft-lefaucheur-rsvp-ecn-01.txt
Expires: December 2006 June 2006

RSVP Extensions for Admission Control over Diffserv using Pre-congestion Notification (PCN)

Status of this Memo

Abstract

This document specifies the extensions to RSVP for support of the Controlled Load (CL) service over a Diffserv cloud using Pre-Congestion Notification as defined in [CL-DEPLOY].

Specification of Requirements

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

1. Introduction

[RSVP] defines the Resource reSerVation Protocol which can be used by applications to request resources from the network. The network responds by explicitely admitting or rejecting these RSVP requests. Certain applications that have quantifiable resource requirements express these requirements using Intserv parameters as defined in the appropriate Intserv service specifications ([RFC2212], [RFC2211]). Controlled Load (CL) service is a quality of service (QoS) closely approximating the QoS that the same flow would receive from a lightly loaded network element [RFC2211]. CL is useful for inelastic flows such as those for real-time media.

[CL-DEPLOY] describes a deployment model to achieve a Controlled Load (CL) service ([RFC2211]) by using distributed measurement-based admission control edge-to-edge, i.e. within a particular region of the Internet. 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 Pre-congestion Marking behaviour, 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.

[CL-DEPLOY] also describes how the framework uses rate-based pre-emption to maintain the CL service to as many admitted microflows as possible even after localised failure and routing changes in the interior of the edge-to-edge region.

The edge-to-edge architecture of [CL-DEPLOY] is a building block in delivering an end-to-end CL service. The approach is similar to that described in [INTSERV-DIFFERV] for Integrated services operation over Diffserv networks. Like [INTSERV-DIFFERV], 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.

[CL-DEPLOY] assumes that the end-to-end signalling mechanism is RSVP. This document specifies the extensions to RSVP for support of the Controlled Load (CL) service over a Diffserv cloud using Pre-Congestion Notification as defined in [CL-DEPLOY].

2. Definitions

For readability, a number of definitions from [CL-DEPLOY] are repeated here:

- ingress edge (or ingress gateway): router at an ingress to the CL-region. A CL-region may have several ingress gateways.

- egress edge (or egress gateway): router at an egress from the CL-region. A CL-region may have several egress gateways.

- Interior router: a router which is part of the CL-region, but is not an ingress or egress gateway.

- CL-region: A region of the Internet in which all traffic enters/leaves through an ingress/egress gateway and all routers 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 field in the flow packet headers that uniquely identifies the aggregate.

- Congestion-Level-Estimate: the number of bits in CL packets that are admission marked (or pre-emption 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.

3. Overview of RSVP extensions and Operations

3.1. Overall QoS Architecture

The overall QoS architecture is described in [CL-DEPLOY]. For readability, the Figure of [CL-DEPLOY] illustrating this QoS architecture is reproduced below in Figure 1.

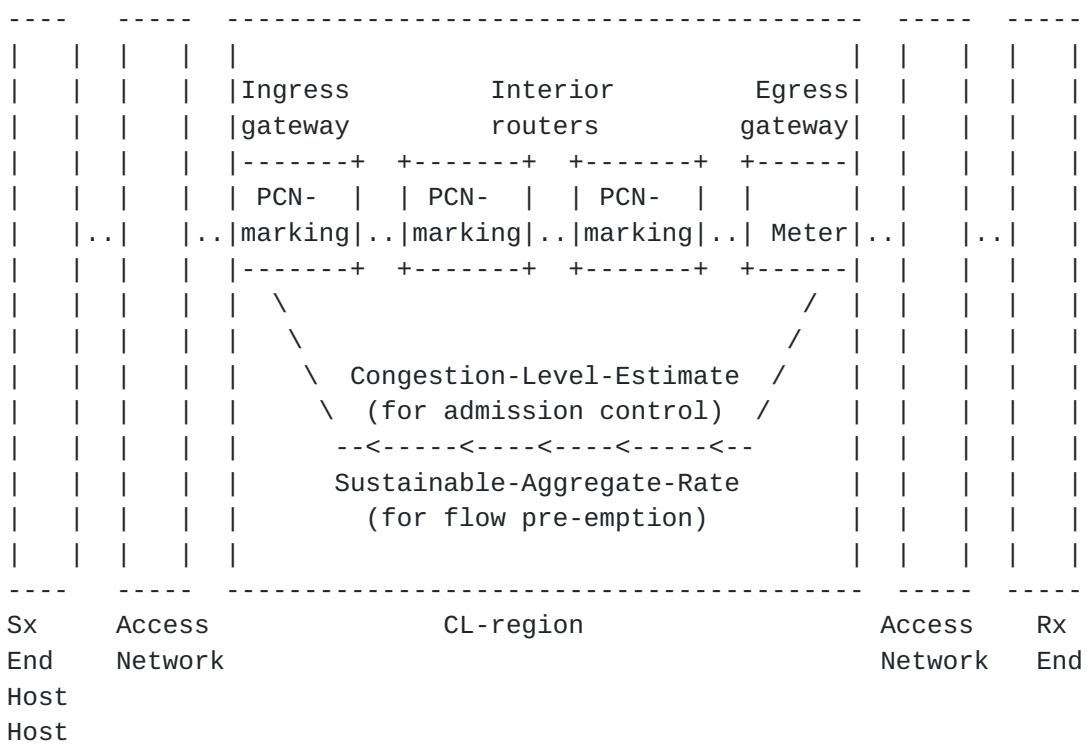


Figure 1: Overall QoS Architecture

3.2. Overview of Procedures for Admission Control of New Reservations

As mentioned earlier, [CL-DEPLOY] describes a framework to achieve a Controlled Load (CL) service by using distributed measurement-based admission control edge-to-edge, i.e. within a particular region of the Internet. This section describes RSVP operations to support such an admission control scheme relying on Pre-Congestion Notification in the eddge-to-edge region.

When a new Path message is received by Ingress Edge, the Ingress Edge does regular RSVP processing (including updating the RSVP PHOP) and forwards the Path towards destination.

All the PCN-capable Interior nodes are not RSVP-capable (or have RSVP processing disabled) and thus simply ignore the Path message.

When the Path message arrives at the Egress Edge, the Egress Edge processes it as per regular RSVP processing, augmented with the following rules:

1) The Egress Edge does NOT perform the RSVP-TTL vs IP TTL-check and does NOT update the ADspec Break bit. This is because the whole CL-region is effectively handled by RSVP as a virtual "link" on which Integrated Service is indeed supported (and admission control performed) so that the Break bit MUST not be set.

2) The Egress Edge MAY check, at the time of initial Path processing, whether it has a valid value for the corresponding Congestion-Level-Estimate and if not it MAY send a PathErr message to the Ingress Edge with a new "CL-PCN Probes Required" Error Code. This minimizes call set up time as it allows probes to be generated by the Ingress Edge and measured by the Egress Edge while the Path is traveling towards the receiver and while the Resv travels back from the receiver.

Then the Egress Edge forwards the Path message towards the receiver.

[Editor Note: discussion on Adspec update to be added]

When the Resv message is received by the Egress Edge (from the downstream side), the Egress Edge performs regular RSVP processing (including performing admission control for the segment downstream of the Egress Edge) augmented with the procedures described in this section.

The Egress Edge MUST include the new CL-PCN object in the Resv message transmitted to the RSVP PHOP (which is the Ingress Edge). The CL-PCN object MUST convey the current Pre-Congestion Notification Congestion-Level-Estimate as measured by the Egress Edge from the corresponding Ingress Edge to itself. Details for computing the Congestion-Level-estimate can be found in [CL-DEPLOY] and [PCN-MARKING].

If the Egress Edge does not have a current value for the Congestion-Level-estimate for the corresponding Ingress Edge (because there was no traffic received by the Egress Edge from that Ingress Edge) and it has not already requested the Ingress Edge to generate CL-PCN probes, the Egress Edge:

1) triggers a timer and puts the Resv message processing on hold

2) sends a PathErr message towards the Ingress Edge with the new Error Code of "CL-PCN Probes Required" specified in this document, in order to instruct the Ingress Edge to generate the necessary probe traffic to enable the Egress Edge to compute the Congestion-Level-Estimate from that Ingress Edge

3) When timer expires the Resv processing resumes. Assuming the Congestion-Level-Estimate is now available, the Egress Edge can include it in the CL-PCN object and complete Resv processing. If the Congestion-Level-Estimate is still available, the Egress Edge may loop again a few times through step 1) and 2). After a given number of times, the Egress Edge MUST send a ResvErr towards the receiver with existing ErrorCode "Admission Control Failure"

[Editor note: approach in previous paragraph may be revisited to try avoid having to "put Resv message processing on hold".]

The Egress Edge will then forward the Resv message to the PHOP signaled earlier in the Path message and which identifies the Ingress Edge. Since the Resv message is directly addressed to the Ingress Edge and does not carry the Router Alert option (as per regular RSVP Resv procedures), the Resv message is hidden from the Interior nodes which handle the E2E Resv message as a regular IP packet.

When receiving the Resv message, the Ingress Edge processes the Resv message as per regular RSVP with the following exceptions:

1) if the CL-PCN object is absent from the Resv message, this means that the RSVP Next Hop is not CL-PCN capable and hence proper admission control can not be achieved for that reservation over the PCN cloud. Thus, the Ingress Edge MUST send a ResvErr message towards the receiver with an new Error Code "Inconsistent Admission Control Behaviour across Ingress and Egress Edge" and an Error Value of "Egress Edge Router not CL-PCN capable". The Ingress Edge MAY also generate an alarm to the network operator.
   Note that in the case where the RSVP Next Hop is not CL-PCN capable, this RSVP hop would have (most probably) performed the RSVP-TTL vs IP-TTL check when processing the initial Path message and as a result would have set the Break bit in the Adspec (assuming there is at least one Interior node on the path from the Ingress Edge to the RSVP Next Hop). Thus, the sender would already have been notified in the first place that the QoS could not be guaranteed end-to-end.

2) The Ingress Edge MUST carry out the admission control decision

(for admission of the reservation over the path from Ingress

Edge to Egress Edge through the PCN cloud) taking into account the congestion information provided in the CL-PCN object of the Resv message in accordance with the procedures of [CL-DEPLOY] and [PCN-MARKING]. For example, if the Congestion Level Estimate conveyed in the CL-PCN object exceeds a configured threshold, the Ingress Edge may decide to reject this new reservation. Once the admission control decision is taken by the Ingress Edge, regular RSVP procedures are followed to either proceed with the reservation (and forward the Resv towards the sender) or tear down the reservation (and, in particular, send a ResvErr towards the receiver with existing Error Code "Admission Control failure".

3) In case the Ingress Edge forwards the Resv message upstream, the Ingress Edge MUST remove the CL-PCN object

When generating a refresh for a Resv message towards the Ingress Edge, the Egress Edge SHOULD NOT include the current value of the Congestion-Level-Estimate in the CL-PCN object, but rather SHOULD include the value which was included in the previous refresh. This is for implementation reasons, to facilitate detection by the Ingress Edge that this message is a mere refresh even if the value of the actual Congestion-Level-Estimate has changed since the previous refresh.

When receiving a PathErr message with the new Error Code of "CL-PCN Probes Required", the Ingress Edge MUST generate CL-PCN probes as described in [CL-DEPLOY] towards the Egress Edge which sent the PathErr Message, and MUST not propagate the PathErr message further upstream.

[Editor Note: discuss RSVP Authentication between ingress and egress edges]

3.3. Removal of E2E reservations

E2E reservations are removed in the usual RSVP way via PathTear, ResvTear, timeout, or as the result of an error condition. This does not directly affect CL-PCN operations.

3.4. Overview of Procedures for Preemption of Existing Reservations

As mentioned earlier, [CL-DEPLOY] describes how rate-based pre-emption can be used to maintain the CL service to as many admitted microflows as possible, even after localised failure and routing changes in the interior of the edge-to-edge region. The solution

involves the egress edge alerting the ingress edge of the CL-region-aggregate that preemption may be needed and conveying to the ingress edge the measured Sustainable-Aggregate-Rate. [CL-DEPLOY] also identifies that this information needs to be transferred reliably. This section describes the corresponding RSVP extensions.

Section 3.2.1 "Alerting the Ingress Edge that pre-emption may be

Let us assume that a number of reservations are established and transit through a given Ingress Edge Ei and a given Egress Edge Ee. Let us now assume that Ee is alerted that preemption may be needed and that Ee has measured the Sustainable-Aggregate-Rate for the CL-region-aggregate from Ei to Ee.

Then, Ee MUST arbitrarily select one of the reservations whose Previous Hop is Ei and address to Ei a Resv message for that reservation with a CL-PCN object containing the current Sustainable-Aggregate-Rate for the relevant CL-region-aggregate.

To avoid the risk that this Resv message gets lost and in turn that the Ingress Edge is not made aware in a timely manner that preemption may be needed, the RSVP reliable messaging procedures specified in [RSVP-REFRESH] SHOULD be used.

Note that, even when reliable messaging is used, there is a very small risk that the alert that preemption may be needed does not make it to the Ingress Edge. For example, this could happen because there could be a race condition whereby the corresponding RSVP reservation may get torn down around the same time where the Resv message with the CL-PCN object is transmitted, resulting in the Ingress Edge ignoring the whole Resv message. However, the probability for this to occur is low and could also be mitigated by the Egress Edge sending the alert on more than one reservation.

[Editor Note: optional use of a Notify message will be investigated. Can this solve the race condition problem mentioned above?]

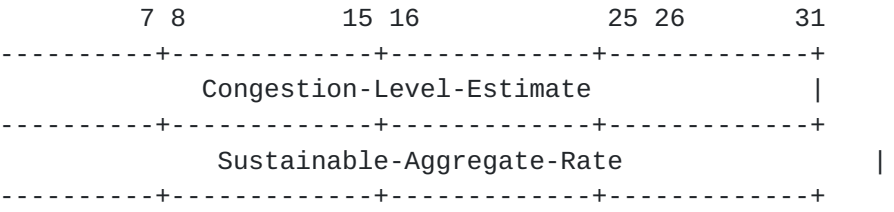
On receipt of the Resv message Ei will detect that this message is not just a refresh because the content of the CL-PCN object has changed and will immediately trigger its preemption logic. This will assess whether some reservations need to be dropped in accordance with the [CL-DEPLOY] and [PCN-MARKING] scheme. In case some do, those will be torn down as per regular RSVP procedures (in particular a ResvErr message is then sent to the receiver).

4. RSVP Object and Error Code Definition

This document defines a new object and two new error codes.

4.1. CL-PCN Object

o Class = To be allocated by IANA
C-Type = 1



The CL-PCN Object may only be used in Resv messages.

Let us refer:
- to the Egress Edge which generated the Resv message containing the CL-PCN object as Ee
- to the RSVP Previous HOP (Ingres Edge) for the corresponding reservation as Ei.

CL-PCN Congestion-Level-Estimate:
This contains the value of the Congestion-Level-Estimate (defined in [CL-DEPLOY] and [PCN-MARKING]) computed by Ee for the CL-region-aggregate from Ei to Ee. When generating a refresh for a Resv message towards the Ingress Edge, the Egress Edge SHOULD NOT include the current value of the Congestion-Level-Estimate in the CL-PCN object, but rather SHOULD include the value which was included in the previous refresh.
[Editor Note: Encoding details to be added]

Sustainable-Aggregate-Rate:
This contains:
- When Ee is not alerted that preemption is needed for the CL-region-aggregate from Ei to Ee, this field is set to 0,
- When Ee is alerted that preemption is (or may be) needed for the CL-region-aggregate from Ei to Ee, the Sustainable-Aggregate-Rate for the relevant CL-region-aggregate (defined in [CL-DEPLOY] and [PCN-MARKING]) computed by Ee for this CL-region-aggregate.

[Editor Note: Encoding details to be added]

4.2. "CL-PCN Probes Required" Error Code

The "CL-PCN Probes Required" Error Code may appear only in PathErr messages.

Error Code = To be allocated by IANA

4.3. "Inconsistent Admission Control Behaviour across Ingress and Egress Edge" Error Code

The "Inconsistent Admission Control Behaviour across Ingress and Egress Edge" may appear only in ResvErr messages.
[Editor note: should we allow it in PathErr messages too so that notification can also be provided to the sender?]

Error Code for "Inconsistent Admission Control Behaviour across Ingress and Egress Edge"= To be allocated by IANA

Error Value for "Egress Edge Router not CL-PCN capable"= To be allocated by IANA

5. Security Considerations

To be added

6. IANA Considerations

This document makes the following requests to the IANA:
- allocate a new Object Class (CL-PCN Object)
- allocate a new Error Code ("CL-PCN Probes Required") and manage the corresponding Error Value range
- allocate a new Error Code ("Inconsistent Admission Control Behaviour across Ingress and Egress Edge"), manage the corresponding Error Value range, and allocate the "Egress Edge Router not CL-PCN capable" under that range.

7. Acknowledgments

We would like to thank Carol Iturralde for her input into this document.

8. Normative References

[RSVP] Braden, R., ed., et al., "Resource ReSerVation Protocol (RSVP)- Functional Specification", RFC 2205, September 1997.

[CL-DEPLOY] B. Briscoe, P. Eardley, D. Songhurst, F. Le Faucheur, A. Charny, S. Dudley, J. Babiarz, K. Chan, G. Karagiannis, A. Bader., L Westberg. "A Deployment Model for Admission Control over DiffServ

using Pre-Congestion Notification, draft-briscoe-tsvwg-cl-architecture-03.txt", (work in progress), June 2006

[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.

[PCN-MARKING] B. Briscoe, P. Eardley, D. Songhurst, F. Le Faucheur, A. Charny, S. Dudley, J. Babiarz, K. Chan, G. Karagiannis, A. Bader., L Westberg. "Pre-Congestion Notification marking" draft-briscoe-tsvwg-cl-phb-02.txt (work in progress), June 2006.

[RSVP-REFRESH] Burger et al, "RSVP Refresh Overhead Reduction Extensions", RFC2961, April 2001

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

[RFC2212] S. Shenker et al., Specification of Guaranteed Quality of Service, September 1997

9. Informative References

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

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

10. Authors' Address

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

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

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

IPR Statements

attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the IETF on-line IPR repository at http://www.ietf.org/ipr.

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement this standard.
Please address the information to the IETF at ietf-ipr@ietf.org.

Disclaimer of Validity

Copyright Notice

Appendix A - Example RSVP Signaling Flow for Admission Control

To be added. Shows RSVP message flow in case of admission control of new reservations.

Appendix B - Example Signaling Flow for Preemption

To be added. Shows RSVP message flow in case of preemption of existing reservations.