Network Working Group B. Davie
Internet-Draft Cisco Systems, Inc.
Intended status: Standards Track B. Briscoe
Expires: December 21, 2007 J. Tay
BT Research
June 19, 2007

Explicit Congestion Marking in MPLS
draft-ietf-tsvwg-ecn-mpls-01.txt

Status of this Memo

Copyright Notice

Abstract

RFC 3270 defines how to support the Diffserv architecture in MPLS networks, including how to encode Diffserv Code Points (DSCPs) in an MPLS header. DSCPs may be encoded in the EXP field, while other uses of that field are not precluded. RFC3270 makes no statement about how Explicit Congestion Notification (ECN) marking might be encoded in the MPLS header. This draft defines how an operator might define

some of the EXP codepoints for explicit congestion notification, without precluding other uses.

Requirements Language

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 RFC 2119 [RFC2119].

Change History

[Note to RFC Editor: This section to be removed before publication]

Changes in this version (draft-ietf-tsvwg-ecn-mpls-01.txt) relative to the last (draft-ietf-tsvwg-ecn-mpls-00.txt):

o Moved the detailed discussion of marking procedures for Pre-Congestion Notification (PCN) to an appendix.

o Removed PCN as a motivation for the efficient code-point usage in Section 2.

o Clarified the rationale for preferring the ECT-checking approach over the approach of [Floyd] in Section 9.1.

o Updated discussion of relationship to RFC3168 in Section 7

o Removed discussion of re-ECN from Security Considerations.

o Fixed typos and nits.

Changes in draft-ietf-tsvwg-ecn-mpls-00.txt relative to draft-davie-ecn-mpls-00:

o Corrected the description of ECN-MPLS marking proposed in [Shayman], which closely corresponds to that proposed in this document.

o Pre-congestion notification (PCN) marking is now described in a way that does not require normative references to PCN specifications. PCN discussion now serves only to illustrate how the ECN marking concepts can be extended to cover more complex scenarios, with PCN being an example.

o Added specification of behavior when MPLS encapsulated packets cross from an ECN-enabled domain to a domain that is not ECN-enabled.

- Clarified that copying MPLS ECN or PCN marking into exposed IP header on egress is not mandatory

- Fixed typos and nits

Table of Contents

1. Introduction

1.1. Background

[RFC3168] defines Explicit Congestion Notification for IP. The primary purpose of ECN is to allow congestion to be signalled without dropping packets.

[RFC3270] defines how to support the Diffserv architecture in MPLS networks, including how to encode Diffserv Code Points (DSCPs) in an MPLS header. DSCPs may be encoded in the EXP field, while other uses of that field are not precluded. RFC3270 makes no statement about how Explicit Congestion Notification (ECN) marking might be encoded in the MPLS header.

This draft defines how an operator might define some of the EXP codepoints for explicit congestion notification, without precluding other uses. In parallel to the activity defining the addition of ECN to IP [RFC3168], two proposals were made to add ECN to MPLS [Floyd][Shayman]. These proposals, however, fell by the wayside. With ECN for IP now being a proposed standard, and developing interest in using pre-congestion notification (PCN) for admission control and flow pre-emption [I-D.briscoe-tsvwg-cl-architecture], there is consequent interest in being able to support ECN across IP networks consisting of MPLS-enabled domains. Therefore it is necessary to specify the protocol for including ECN in the MPLS shim header, and the protocol behavior of edge MPLS nodes.

We note that in [RFC3168] there are four codepoints used for ECN marking, which are encoded using two bits of the IP header. The MPLS EXP field is the logical place to encode ECN codepoints, but with only 3 bits (8 codepoints) available, and with the same field being used to convey DSCP information as well, there is a clear incentive to conserve the number of codepoints consumed for ECN purposes. Efficient use of the EXP field has been a focus of prior drafts [Floyd] [Shayman] and we draw on those efforts in this draft as well.

We also note that [RFC3168] defines default usage of the ECN field but allows for the possibility that some Diffserv PHBs might include different specifications on how the ECN field is to be used. This draft seeks to preserve that capability.

1.2. Intent

Our intent is to specify how the MPLS shim header[RFC3032] should denote ECN marking and how MPLS nodes should understand whether the transport for a packet will be ECN capable. We offer this as a building block, from which to build different congestion notification

systems. We do not intend to specify how the resulting congestion notification is fed back to an upstream node that can mitigate congestion. For instance, unlike [Shayman], we do not specify edge-to-edge MPLS domain feedback, but we also do not preclude it. Nonetheless, we do specify how the egress node of an MPLS domain should copy congestion notification from the MPLS shim into the encapsulated IP header if the ECN is to be carried onward towards the IP receiver. But we do NOT mandate that MPLS congestion notification must be copied into the IP header for onward transmission. This draft aims to be generic for any use of congestion notification in MPLS. Support of [RFC3168] is our primary motivation; some additional potential applications to illustrate the flexibility of our approach are described in Section 8. In particular, we aim to support possible future schemes that may use more than one level of congestion marking.

1.3. Terminology

This document draws freely on the terminology of ECN [RFC3168] and MPLS [RFC3031]. For ease of reference, we have included some definitions here, but refer the reader to the references above for complete specifications of the relevant technologies:

o CE: Congestion Experienced. One of the states with which a packet may be marked in a network supporting ECN. A packet is marked in this state by an ECN-capable router, to indicate that this router was experiencing congestion at the time the packet arrived.

o ECT: ECN-capable Transport. One of the ECN states which a packet may be in when it is sent by an end system. An end system marks a packet with an ECT codepoint to indicate that the end-points of the transport protocol are ECN-capable. A router may not mark a packet as CE unless the packet was marked ECT when it arrived.

o Not-ECT: Not ECN capable transport. An end system marks a packet with this codepoint to indicate that the end-points of the transport protocol are not ECN-capable. A congested router cannot mark such packets as CE, and thus can only drop them to indicate congestion.

o EXP field. A 3 bit field in the MPLS label header [RFC3032] which may be used to convey Diffserv information (and is also used in this draft to carry ECN information).

o PHP. Penultimate Hop Popping. An MPLS operation in which the penultimate Label Switching Router (LSR) on a Label Switched Path (LSP) removes the top label from the packet before forwarding the packet to the final LSR on the LSP.

2. Use of MPLS EXP Field for ECN

We propose that LSRs configured for explicit congestion notification should use the EXP field in the MPLS shim header. However, [RFC3270] already defines use of codepoints in the EXP field for differentiated services. Although it does not preclude other compatible uses of the EXP field, this clearly seems to limit the space available for ECN, given the field is only 3 bits (8 codepoints).

[RFC3270] defines two possible approaches for requesting differentiated service treatment from an LSR.

o In the E-LSP approach, different codepoints of the EXP field in the MPLS shim header are used to indicate the packet's per hop behavior (PHB).

o In the L-LSP approach, an MPLS label is assigned for each PHB scheduling class (PSC, as defined in [RFC3260], so that an LSR determines both its forwarding and its scheduling behavior from the label.

If an MPLS domain uses the L-LSP approach, there is likely to be space in the EXP field for ECN codepoint(s). Where the E-LSP approach is used, then codepoint space in the EXP field is likely to be scarce. This draft focuses on interworking ECN marking with the E-LSP approach as it is the tougher problem. Consequently the same approach can also be applied with L-LSPs.

We recommend that explicit congestion notification in MPLS should use codepoints instead of bits in the EXP field. Since not every PHB will necessarily require an associated ECN codepoint it would be wasteful to assign a dedicated bit for ECN. (There may also be cases where a given PHB might need more than one ECN-like codepoint; see Section 8.4 for an example.)

For each PHB that uses ECN marking, we assume one EXP codepoint will be defined meaning not congestion marked (Not-CM), and at least one other codepoint will be defined meaning congestion marked (CM). Therefore, each PHB that uses ECN marking will consume at least two EXP codepoints. But PHBs that do not use ECN marking will only consume one.

Further, we wish to use minimal space in the MPLS shim header to tell interior LSRs whether each packet will be received by an ECN-capable transport (ECT). Nonetheless, we must ensure that an end-point that would not understand an ECN mark will not receive one, otherwise it will not be able to respond to congestion as it should. In the past, three solutions to this problem have been proposed:

- One possible approach is for congested LSRs to mark the ECN field in the underlying IP header at the bottom of the label stack. Although many commercial LSRs routinely access the IP header for other reasons (ECMP), there are numerous drawbacks to attempting to find an IP header beneath an MPLS label stack. Notably, there is the challenge of detecting the absence of an IP header when non-IP packets are carried on an LSP. Therefore we will not consider this approach further.

- In the scheme suggested by [Floyd] ECT and CE are overloaded into one bit, so that a 0 means ECT while a 1 might either mean Not-ECT or it might mean CE. A packet that has been marked as having experienced congestion upstream, and then is picked out for marking at a second congested LSR, will be dropped by the second LSR since it cannot determine whether the packet has previously experienced congestion or if ECN is not supported by the transport.

  While such an approach seemed potentially palatable, we do not recommend it here for the following reasons. In some cases we wish to be able to use ECN marking long before actual congestion (e.g. pre-congestion notification). In these circumstances, marking rates at each LSR might be non-negligible most of the time, so the chances of a previously marked packet encountering an LSR that wants to mark it again will also be non-negligible. In the case where CE and not-ECT are indistinguishable to core routers, such a scenario could lead to unacceptable drop rates. If the typical marking rate at every router or LSR is p, and the typical diameter of the network of LSRs is d, then the probability that a marked packet will be chosen for marking more than once is 1-[Pr(never marked) + Pr(marked at exactly one hop)] = 1- [(1-p)^d + dp(1-p)^(d-1)]. For instance, with 6 LSRs in a row, each marking ECN with 1% probability, the chances of a packet that is already marked being chosen for marking a second time is 0.15%. The bit overloading scheme would therefore introduce a drop rate of 0.15% unnecessarily. Given that most modern core networks are sized to introduce near-zero packet drop, it may be unacceptable to drop over one in a thousand packets unnecessarily.

- A third possible approach was suggested by [Shayman]. In this scheme, interior LSRs assume that the endpoints are ECN-capable, but this assumption is checked when the final label is popped. If an interior LSR has marked ECN in the EXP field of the shim header, but the IP header says the endpoints are not ECN capable, the edge router (or penultimate router, if using penultimate hop popping) drops the packet. We recommend this scheme, which we call `per-domain ECT checking', and define it more precisely in the following section. Its chief drawback is that it can cause

packets to be forwarded after encountering congestion only to be dropped at the egress of the MPLS domain. The rationale for this decision is given in Section 9.1.

3. Per-domain ECT checking

For the purposes of this discussion, we define the egress nodes of an MPLS domain as the nodes that pop the last MPLS label from the label stack, exposing the IP (or, potentially non-IP) header. Note that such a node may be the ultimate or penultimate hop of an LSP, depending on whether penultimate hop popping (PHP) is employed.

In the per-domain ECT checking approach, the egress nodes take responsibility for checking whether the transport is ECN capable. This draft does not specify how these nodes should pass on congestion notification, because different approaches are likely in different scenarios. However, if congestion notification in the MPLS header is copied into the IP header, the procedure MUST conform to the specification given here.

If congestion notification is passed to the transport without first passing it onward in the IP header, the approach used must take similar care to check that the transport is ECN capable before passing it ECN markings. Specifically, if the transport for a particular congestion marked MPLS packet is found not to be ECN-capable, the packet MUST be dropped at this egress node.

In the per-domain ECT checking approach, only the egress nodes check whether an IP packet is destined for an ECN-capable transport. Therefore, any single LSR within an MPLS domain MUST NOT be configured to enable ECN marking unless all the egress LSRs surrounding it are already configured to handle ECN marking.

We call a domain surrounded by ECN-capable egress LSRs an ECN-enabled MPLS domain. This term only implies that all the egress LSRs are ECN-enabled; some interior LSRs may not be ECN-enabled. For instance, it would be possible to use some legacy LSRs incapable of supporting ECN in the interior of an MPLS domain as long as all the egress LSRs were ECN-capable. Note that if PHP is used, the "penultimate hop" routers which perform the pop operation do need to be ECN-enabled, since they are acting in this context as egress LSRs.

4. ECN-enabled MPLS domain

In the following subsections we describe various operations affecting the ECN marking of a packet that may be performed at MPLS edge and

core LSRs.

4.1. Pushing (adding) one or more labels to an IP packet

On encapsulating an IP packet with an MPLS label stack, the ECN field must be translated from the IP packet into the MPLS EXP field. The Not-CM (not congestion marked) state is set in the MPLS EXP field if the ECN status of the IP packet is "Not ECT" or ECT(1) or ECT(0). The CM state is set if the ECN status of the IP packet is "CE". If more than one label is pushed at one time, the same value should be placed in the EXP value of all label stack entries.

4.2. Pushing one or more labels onto an MPLS labelled packet

The EXP field is copied directly from the topmost label before the push to the newly added outer label. If more than one label is being pushed, the same EXP value is copied to all label stack entries.

4.3. Congestion experienced in an interior MPLS node

If the EXP codepoint of the packet maps to a PHB that uses ECN marking and the marking algorithm requires the packet to be marked, the CM state is set (irrespective of whether it is already in the CM state).

If the buffer is full, a packet is dropped.

4.4. Crossing a Diffserv Domain Boundary

If an MPLS-encapsulated packet crosses a Diffserv domain boundary, it may be the case that the two domains use different encodings of the same PHB in the EXP field. In such cases, the EXP field must be rewritten at the domain boundary. If the PHB is one that supports ECN, then the appropriate ECN marking should also be preserved when the EXP field is mapped at the boundary.

If an MPLS-encapsulated packet that is in the CM state crosses from a domain that is ECN-enabled (as defined in Section 3) to a domain that is not ECN-enabled, then it is necessary to perform the egress checking procedures at the egress LSR of the ECN-enabled domain. This means that if the encapsulated packet is not ECN capable, the packet MUST be dropped. Note that this implies the egress LSR must be able to look beneath the MPLS header without popping the label stack.

The related issue of Diffserv tunnel models is discussed in Section 4.7.

4.5. Popping an MPLS label (not the end of the stack)

When a packet has more than one MPLS label in the stack and the top label is popped, another MPLS label is exposed. In this case the ECN information should be transferred from the outer EXP field to the inner MPLS label in the following manner. If the inner EXP field is Not-CM, the inner EXP field is set to the same CM or Not-CM state as the outer EXP field. If the inner EXP field is CM, it remains unchanged whatever the outer EXP field. Note that an inner value of CM and an outer value of not-CM should be considered anomalous, and SHOULD be logged in some way by the LSR.

4.6. Popping the last MPLS label in the stack

When the last MPLS label is popped from the packet, its payload is exposed. If that packet is not IP, and does not have any capability equivalent to ECT, it is assumed Not-ECT and treated as such. That means that if the EXP value of the MPLS header was CM, the packet MUST be dropped.

Assuming an IP packet was exposed, we have to examine whether that packet is ECT or not. A Not-ECT packet MUST be dropped if the EXP field is CM.

For the remainder of this section, we describe the behavior that is required if the ECN information is to be transferred from the MPLS header into the exposed IP header for onward transmission. As noted in Section 1.2, such behavior is not mandated by this document, but may be selected by an operator.

If the inner IP packet is Not-ECT, its ECN field remains unchanged if the EXP field is Not-CM. If the ECN field of the inner packet is set to ECT(0), ECT(1) or CE, the ECN field remains unchanged if the EXP field is set to Not-CM. The ECN field is set to CE if the EXP field is CM. Note that an inner value of CE and an outer value of not-CM should be considered anomalous, and SHOULD be logged in some way by the LSR.

4.7. Diffserv Tunneling Models

[RFC3270] describes three tunneling models for Diffserv support across MPLS Domains, referred to as the uniform, short pipe, and pipe models. The differences between these models lie in whether the Diffserv treatment that applies to a packet while it travels along a particular LSP is carried to the last hop of the LSP and beyond the last hop. Depending on which mode is preferred by an operator, the EXP value or DSCP value of an exposed header following a label pop may or may not be dependent on the EXP value of the label that is

removed by the pop operation. We believe that in the case of ECN marking, the use of these models should only apply to the encoding of the Diffserv PHB in the EXP value, and that the choice of codepoint for ECN should always be made based on the procedures described above, independent of the tunneling model.

5. ECN-disabled MPLS domain

If ECN is not enabled on all the egress LSRs of a domain, ECN MUST NOT be enabled on any LSRs throughout the domain. If congestion is experienced on any LSR in an ECN-disabled MPLS domain, packets MUST be dropped, NOT marked. The exact algorithm for deciding when to drop packets during congestion (e.g. tail-drop, RED, etc.) is a local matter for the operator of the domain.

6. The use of more codepoints with E-LSPs and L-LSPs

[RFC3270] gives different options with E-LSPs and L-LSPs and some of those could potentially provide ample EXP codepoints for ECN. However, deploying L-LSPs vs E-LSPs has many implications such as platform support and operational complexity. The above ECN MPLS solution should provide some flexibility. If the operator has deployed one L-LSP per PHB scheduling class, then EXP space will be a non-issue and it could be used to achieve more sophisticated ECN behavior if required. If the operator wants to stick to E-LSPs and uses a handful of EXP codepoints for Diffserv, it may be desirable to operate with a minimum number of extra ECN codepoints, even if this comes with some compromise on ECN optimality. See Section 8 for discussion of some possible deployment scenarios.

7. Relationship to tunnel behavior in RFC 3168

[RFC3168] defines two modes of encapsulating ECN-marked IP packets inside additional IP headers when tunnels are used. The two modes are the "full functionality" and "limited functionality" modes. In the full functionality mode, the ECT information from the inner header is copied to the outer header at the tunnel ingress, but the CE information is not. In the limited functionality mode, neither ECT nor CE information is copied to the outer header, and thus ECN cannot be applied to the encapsulated packet.

The behavior that is specified in Section 4 of this document resembles the "full functionality" mode in the sense that it conveys some information from inner to outer header, and in the sense that it enables full ECN support along the MPLS LSP (which is analogous to an

IP tunnel in this context). However it differs in one respect, which is that the CE information is conveyed from the inner header to the outer header. Our original reason for this different design choice was to give interior routers and LSRs more information about upstream marking in multi-bottleneck cases. For instance, the flow pre-emption marking mechanism proposed for PCN works by only considering packets for marking that have not already been marked upstream. Unless existing pre-emption marking is copied from the inner to the outer header at tunnel ingress, the mechanism doesn't pre-empt enough traffic in cases where anomalous events hit multiple domains at once. [RFC3168] does not give any reasons against conveying CE information from the inner header to the outer in the "full functionality" mode. Furthermore, [RFC4301] specifies that the ECN marking should be copied from inner header to outer header in IPSEC tunnels, consistent with the approach defined here. [Briscoe] discusses this issue in more detail. In summary, the approach described in Section 4 appears to be both a sound technical choice and consistent with the current state of thinking in the IETF.

8. Example Uses

8.1. RFC3168-style ECN

[RFC3168] proposes the use of ECN in TCP and introduces the use of ECN-Echo and CWR flags in the TCP header for initialization. The TCP sender responds accordingly (such as not increasing the congestion window) when it receives an ECN-Echo (ECE) ACK packet (that is, an ACK packet with ECN-Echo flag set in the TCP header), then the sender knows that congestion was encountered in the network on the path from the sender to the receiver.

It would be possible to enable ECN in an MPLS domain for Diffserv PHBs like AF and best efforts that are expected to be used by TCP and similar transports (e.g. DCCP [RFC4340]). Then end-to-end congestion control in transports capable of understanding ECN would be able to respond to approaching congestion on LSRs without having to rely on packet discard to signal congestion.

8.2. ECN Co-existence with Diffserv E-LSPs

Many operators today have deployed Diffserv using the E-LSP approach of [RFC3270]. In many cases the number of PHBs used is less than 8, and hence there remain available codepoints in the EXP space. If an operator wished to support ECN for single PHB, this can be accomplished by simply allocated a second codepoint to the PHB for the "CM" state of that PHB and retaining the old codepoint for the "not-CM" state. An operator with only four deployed PHBs could of

course enable ECN marking on all those PHBs. It is easy to imagine cases where some PHBs might benefit more from ECN than others - for example, an operator might use ECN on a premium data service but not on a PHB used for best effort internet traffic.

As an illustrative example of how the EXP field might be used in this case, consider the example of an operator who is using the aggregated service classes proposed in [I-D.ietf-tsvwg-diffserv-class-aggr]. He may choose to support ECN only for the Assured Elastic Treatment Aggregate, using the EXP codepoint 010 for the not-CM state and 011 for the CM state. All other codepoints could be the same as in [I-D.ietf-tsvwg-diffserv-class-aggr]. Of course any other combination of EXP values can be used according to the specific set of PHBs and marking conventions used within that operator's network.

8.3. Congestion-feedback-based Traffic Engineering

Shayman's traffic engineering [Shayman] proposed the use of ECN by an egress LSR feeding back congestion to an ingress LSR to mitigate congestion by employing dynamic traffic engineering techniques such as shifting flows to an alternate path. It proposed a new RSVP TUNNEL CONGESTION message which was sent to the ingress LSR and ignored by transit LSRs.

8.4. PCN flow admission control and flow pre-emption

[I-D.briscoe-tsvwg-cl-architecture] proposes using pre-congestion notification (PCN) on routers within an edge-to-edge Diffserv region to control admission of new flows to the region and, if necessary, to pre-empt existing flows in response to disasters and other anomalous routing events. In this approach, the current level of PCN marking is picked up by the signalling used to initiate each flow in order to inform the admission control decision for the whole region at once. As an example, a minor extension to RSVP signalling has been proposed [I-D.lefaucheur-rsvp-ecn] to carry this message, but a similar approach has also been proposed that uses NSIS signalling [I-D.ietf-nsis-rmd].

If it is possible for LSRs to signify congestion in MPLS, PCN marking could be used for admission control and flow pre-emption across a Diffserv region, irrespective of whether it contained pure IP routers, MPLS LSRs, or both. Indeed, the solution could be somewhat more efficient to implement if aggregates could identify themselves by their MPLS label. Appendix A describes the mechanisms by which the necessary markings for PCN could be carried in the MPLS header.

As an illustrative example of how the EXP field might be used in this case, consider the example of an operator who is using the aggregated

service classes proposed in [I-D.ietf-tsvwg-diffserv-class-aggr]. He may choose to support PCN only for the Real Time Treatment Aggregate, using the EXP codepoint 100 for the not-marked (NM) state, 101 for the Admission Marked (AM) state, and 111 for the Pre-emption Marked (PM) state. All other codepoints could be the same as in [I-D.ietf-tsvwg-diffserv-class-aggr]. Of course any other combination of EXP values can be used according to the specific set of PHBs and marking conventions used within that operator's network.

It might also be possible to deploy a similar solution using PCN marking over MPLS for just admission control alone, or just flow pre-emption alone, particularly if codepoint space was at a premium in the MPLS EXP field. However, the feasibility of deploying one without the other would require further study. We also note that an approach to deploying PCN using only a single marking codepoint to support both pre-emption and admission control has been proposed[I-D.charny-pcn-single-marking].

9. Deployment Considerations

9.1. Marking non-ECN Capable Packets

What are the consequences of marking a packet that is not ECN-capable? Even if it will be dropped before leaving the domain, doesn't this consume resources unnecessarily?

The problem only arises if there is congestion downstream of an earlier congested queue in the same MPLS domain. Downstream congested LSRs might forward packets already marked, even though they will be dropped later when the inner IP header is found to be Not-ECT on decapsulation. Such packets might cause the downstream LSRs to mark (or drop) other packets that they would otherwise not have had to.

We expect congestion will typically be rare in MPLS networks, but it might not be. The extra unnecessary load at downstream LSRs will not be more than the fraction of marked packets from upstream LSRs, even in the worst case where no transports are ECN capable. Therefore the amount of unnecessary marking (or drop) on an LSR will not be more than the product of its local marking rate and the marking rate due to upstream LSRs within the same domain - typically the product of two small (often zero) probabilities.

This is why we decided to use the per-domain ECT checking approach - because the most likely effect would be a very slightly increased marking rate, which would result in very slightly higher drop only for non-ECN-capable transports. We chose not to use the [Floyd]

alternative which introduced a low but persistent level of unnecessary packet drop for all time, even for ECN-capable transports. Although that scheme did not carry traffic to the edge of the MPLS domain only to be dropped on decapsulation, we felt our minor inefficiency was a small price to pay. And it would get smaller still if ECN deployment widened.

A partial solution would be to preferentially drop packets arriving at a congested router that were already marked. There is no solution to the problem of marking a packet when congestion is caused by another packet that should have been dropped. However, the chance of such an occurrence is very low and the consequences are not significant. It merely causes an application to very occasionally slow down its rate when it did not have to.

9.2. Non-ECN capable routers in an MPLS Domain

What if an MPLS domain wants to use ECN, but not all legacy routers are able to support it?

If the legacy router(s) are used in the interior, this is not a problem. They will simply have to drop the packets if they are congested, rather than mark them, which is the standard behavior for IP routers that are not ECN-enabled.

If the legacy router were used as an egress router, it would not be able to check the ECN capability of the transport correctly. An operator in this position would not be able to use this solution and therefore MUST NOT enable ECN unless all egress routers are ECN-capable.

10. IANA Considerations

This document makes no request of IANA.

Note to RFC Editor: this section may be removed on publication as an RFC.

11. Security Considerations

We believe no new vulnerabilities are introduced by this draft.

We have considered whether malicious sources might be able to exploit the fact that interior LSRs will mark packets that are Not-ECT, relying on their egress LSR to drop them. Although this might allow sources to engineer a situation where more traffic is carried across

an MPLS domain than should be, we figured that even if we hadn't introduced this feature, these sources would have been able to prevent these LSRs dropping this traffic anyway, simply by setting ECT in the first place.

An ECN sender can use the ECN nonce [RFC3540] to detect a misbehaving receiver. The ECN nonce works correctly across an MPLS domain without requiring any specific support from the proposal in this draft. The nonce does not need to be present in the MPLS shim header. As long as the nonce is present in the IP header when the ECN information is copied from the last MPLS shim header, it will be overwritten if congestion has been experienced by an LSR. This is all that is necessary for the sender to detect a misbehaving receiver.

12. Acknowledgments

Thanks to K.K. Ramakrishnan and Sally Floyd for getting us thinking about this in the first place and for providing advice on tunneling of ECN packets, and to Sally Floyd, Joe Babiarz, Ben Niven-Jenkins, Phil Eardley, Ruediger Geib, and Magnus Westerlund for their comments on the draft.

Appendix A. Extension to Pre-Congestion Notification

This appendix describes how the mechanisms decribed in the body of the document can be extended to support PCN [I-D.briscoe-tsvwg-cl-architecture]. Our intent here is to show that the mechanisms are readily extended to more complex scenarios than ECN, particulary in the case where more codepoints are needed, but this appendix may be safely ignored if one is interested only in supporting ECN. Note that the PCN standards are still very much under development at the time of writing, hence the precise details contained in this appendix may be subject to change, and we stress that this appendix is for illustrative purposes only.

The relevant aspects of PCN for the purposes of this discussion are:

- PCN uses 3 states rather than 2 for ECN - these are referred to as admission marked (AM), pre-emption marked (PM) and not marked (NM) states. (See Section 8.4 for further discussion of PCN and the possibility of using fewer codepoints.)

- A packet can go from NM to AM, from NM to PM, or from AM to PM, but no other transition is possible.

- The determination of whether a packet is subject to PCN is based on the PHB of the packet.

Thus, to support PCN fully in an MPLS domain for a particular PHB, a total of 3 codepoints need to be allocated for that PHB. These 3 codepoints represent the admission marked (AM), pre-emption marked (PM) and not marked (NM) states. The procedures described in Section 4 above need to be slightly modified to support this scenario. The following procedures are invoked when the topmost DSCP or EXP value indicates a PHB that supports PCN.

Appendix A.1. Label Push onto IP packet

If the IP packet header indicates AM, set the EXP value of all entries in the label stack to AM. If the IP packet header indicates PM, set the EXP value of all entries in the label stack to PM. For any other marking of the IP header, set the EXP value of all entries in the label stack to NM.

Appendix A.2. Pushing Additional MPLS Labels

The procedures of Section 4.2 apply.

Appendix A.3. Admission Control or Pre-emption Marking inside MPLS domain

The EXP value can be set to AM or PM according to the same procedures as described in [I-D.briscoe-tsvwg-cl-phb]. For the purposes of this document, it does not matter exactly what algorithms are used to decide when to set AM or PM; all that matters is that if a router would have marked AM (or PM) in the IP header, it should set the EXP value in the MPLS header to the AM (or PM) codepoint.

Appendix A.4. Popping an MPLS Label (not end of stack)

When popping an MPLS Label exposes another MPLS label, the AM or PM marking should be transferred to the exposed EXP field in the following manner:

- If the inner EXP value is NM, then it should be set to the same marking state as the EXP value of the popped label stack entry.

- If the inner EXP value is AM, it should be unchanged if the popped EXP value was AM, and it should be set to PM if the popped EXP value was PM. If the popped EXP value was NM, this should be logged in some way and the inner EXP value should be unchanged.

- If the inner EXP value is PM, it should be unchanged whatever the popped EXP value was, but any EXP value other than PM should be logged.

Appendix A.5. Popping the last MPLS Label to expose IP header

When popping the last MPLS Label exposes the IP header, there are two cases to consider:

- the popping LSR is NOT the egress router of the PCN region, in which case AM or PM marking should be transferred to the exposed IP header field; or

- the popping LSR IS the egress router of the PCN region.

In the latter case, the behavior of the egress LSR is defined in [I-D.briscoe-tsvwg-cl-architecture] and is beyond the scope of this document. In the former case, the marking should be transferred from the popped MPLS header to the exposed IP header as follows:

- If the inner IP header value is neither AM nor PM, and the EXP value was NM, then the IP header should be unchanged. For any other EXP value, the IP header should be set to the same marking state as the EXP value of the popped label stack entry.

- If the inner IP header value is AM, it should be unchanged if the popped EXP value was AM, and it should be set to PM if the popped EXP value was PM. If the popped EXP value was NM, this should be logged in some way and the inner IP header value should be unchanged.

- If the IP header value is PM, it should be unchanged whatever the popped EXP value was, but any EXP value other than PM should be logged.

13. References

13.1. Normative References

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

[RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.

[RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack

Encoding", RFC 3032, January 2001.

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

[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.

[RFC4301] Kent, S. and K. Seo, "Security Architecture for the Internet Protocol", RFC 4301, December 2005.

13.2. Informative References

[Briscoe] "Layered Encapsulation of Congestion Notification", June 2007.

Work in progress.

[Floyd] "A Proposal to Incorporate ECN in MPLS", 1999.

Work in progress. http://www.icir.org/floyd/papers/draft-ietf-mpls-ecn-00.txt

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

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

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

[I-D.ietf-nsis-rmd] Bader, A., "RMD-QOSM - The Resource Management in Diffserv QOS Model", draft-ietf-nsis-rmd-09 (work in progress), March 2007.

[I-D.ietf-tsvwg-diffserv-class-aggr]
Chan, K., "Aggregation of DiffServ Service Classes", draft-ietf-tsvwg-diffserv-class-aggr-02 (work in progress), March 2007.

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

[RFC3260] Grossman, D., "New Terminology and Clarifications for Diffserv", RFC 3260, April 2002.

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

[RFC4340] Kohler, E., Handley, M., and S. Floyd, "Datagram Congestion Control Protocol (DCCP)", RFC 4340, March 2006.

[Shayman] "Using ECN to Signal Congestion Within an MPLS Domain", 2000.

Work in progress. http://www.ee.umd.edu/~shayman/papers.d/draft-shayman-mpls-ecn-00.txt

Authors' Addresses

Bruce Davie
Cisco Systems, Inc.
1414 Mass. Ave.
Boxborough, MA 01719
USA

Email: bsd@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

June Tay
BT Research
B54/77, Sirius House
Adastral Park
Martlesham Heath
Ipswich
Suffolk IP5 3RE
United Kingdom

Email: june.tay@bt.com

Full Copyright Statement

Intellectual Property

Acknowledgment

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