Network Working Group
Internet Draft
Category: Standards Track
Expiration Date: August 2007

George Swallow
Cisco Systems, Inc.

Stewart Bryant
Cisco Systems, Inc.

Loa Andersson
Acreo

February 2007

Avoiding Equal Cost Multipath Treatment in MPLS Networks

draft-ietf-mpls-ecmp-bcp-03.txt

Status of this Memo

Abstract

This document describes the Equal Cost Multipath (ECMP) behavior of currently deployed MPLS networks. This document makes best practice recommendations for anyone defining an application to run over an MPLS network that wishes to avoid the reordering that can result from

transmission of different packets from the same flow over multiple different equal cost paths.

Contents

1. Introduction

This document describes the Equal Cost Multipath (ECMP) behavior of currently deployed MPLS networks. We discuss cases where multiple packets from the same top-level LSP might be transmitted over different equal cost paths, resulting in possible mis-ordering of packets which are part of the same top-level LSP. This document also makes best practice recommendations for anyone defining an application to run over an MPLS network that wishes to avoid the resulting potential for mis-ordered packets. While disabling ECMP behavior is an option open to most operators, few (if any) have chosen to do so, and the application designer does not have control over the behavior of the networks that the application may run over. Thus ECMP behavior is a reality that must be reckoned with.

1.1. Terminology

ECMP	Equal Cost Multipath
FEC	Forwarding Equivalence Class
IP ECMP	A forwarding behavior in which the selection of the next-hop between equal cost routes is based on the header(s) of an IP packet
Label ECMP	A forwarding behavior in which the selection of the next-hop between equal cost routes is based on the label stack of an MPLS packet
LSP	Label Switched Path
LSR	Label Switching Router

2. Current ECMP Practices

The MPLS label stack and Forwarding Equivalence Classes are defined in [RFC3031]. The MPLS label stack does not carry a Protocol Identifier. Instead the payload of an MPLS packet is identified by the Forwarding Equivalence Class (FEC) of the bottom most label. Thus it is not possible to know the payload type if one does not know the label binding for the bottom most label. Since an LSR which is processing a label stack need only know the binding for the label(s) it must process, it is very often the case that LSRs along an LSP are unable to determine the payload type of the carried contents.

As a means of potentially reducing delay and congestion, IP networks

have taken advantage of multiple paths through a network by splitting traffic flows across those paths. The general name for this practice is Equal Cost Multipath or ECMP. In general this is done by hashing on various fields on the IP or contained headers. In practice, within a network core, the hashing is based mainly or exclusively on the IP source and destination addresses. The reason for splitting aggregated flows in this manner is to minimize the re-ordering of packets belonging to individual flows contained within the aggregated flow. Within this document we use the term IP ECMP for this type of forwarding algorithm.

For packets that contain both a label stack and an encapsulated IPv4 (or IPv6) packet, current implementations in some cases may hash on any combination of labels and IPv4 (or IPv6) source and destination labels.

In the early days of MPLS, the payload was almost exclusively IP. Even today the overwhelming majority of carried traffic remains IP. Providers of MPLS equipment sought to continue this IP ECMP behavior. As shown above, it is not possible to know whether the payload of an MPLS packet is IP at every place where IP ECMP needs to be performed. Thus vendors have taken the liberty of guessing what the payload is. By inspecting the first nibble beyond the label stack, existing equipment infers that a packet is not IPv4 or IPv6 if the value of the nibble (where the IP version number would be found) is not 0x4 or 0x6 respectively. Most deployed LSRs will treat a packet whose first nibble is equal to 0x4 as if the payload were IPv4 for purposes of IP ECMP.

A consequence of this is that any application which defines a FEC which does not take measures to prevent the values 0x4 and 0x6 from occurring in the first nibble of the payload may be subject to IP ECMP and thus having their flows take multiple paths and arriving with considerable jitter and possibly out of order. While none of this is in violation of the basic service offering of IP, it is detrimental to the performance of various classes of applications. It also complicates the measurement, monitoring and tracing of those flows.

New MPLS payload types are emerging such as those specified by the IETF PWE3 and AVT working groups. These payloads are not IP and, if specified without constraint might be mistaken for IP.

It must also be noted that LSRs which correctly identify a payload as not being IP, most often will load-share traffic across multiple equal-cost paths based on the label stack. Any reserved label, no matter where it is located in the stack, may be included in the com- putation for load balancing. Modification of the label stack between

packets of a single flow could result in re-ordering that flow. That is, were an explicit null or a router-alert label to be added to a packet, that packet could take a different path through the network.

Note that for some applications, being mistaken for IPv4 may not be detrimental. The trivial case where the payload behind the top label is a packet belonging to an MPLS IPv4 VPN. Here the real payload is IP and most (if not all) deployed equipment will locate the end of the label stack and correctly perform IP ECMP.

A less obvious case is when the packets of a given flow happen to have constant values in the fields upon which IP ECMP would be performed. For example if an ethernet frame immediately follows the label and the LSR does not do ECMP on IPv6, then either the first nibble will be 0x4 or it will be something else. If the nibble is not 0x4 then no IP ECMP is performed, but Label ECMP may be performed. If it is 0x4, then the constant values of the MAC addresses overlay the fields that would have been occupied by the source and destination addresses of an IP header. As a result the ECMP algorithm would be feed a constant value and thus would always return the same result.

3. Recommendations for Avoiding ECMP Treatment

We will use the term "Application Label" to refer to a label that has been allocated with a FEC Type that is defined (or simply used) by an application. Such labels necessarily appear at the bottom of the label stack, that is, below labels associated with transporting the packet across an MPLS network. The FEC Type of the Application label defines the payload that follows. Anyone defining an application to be transported over MPLS is free to define new FEC Types and the format of the payload which will be carried.

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                Label                  | Exp |0|       TTL     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
.                                       .     . .               .
.                                       .     . .               .
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                Label                  | Exp |0|       TTL     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|          Application Label            | Exp |1|       TTL     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|1st Nbl|                                                       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

In order to avoid IP ECMP treatment it is necessary that an application take precautions to not be mistaken as IP by deployed equipment that snoops on the presumed location of the IP Version field. Thus, at a minimum, the chosen format must disallow the values 0x4 and 0x6 in the first nibble of their payload.

It is strongly recommended, however, that applications restrict the first nibble values to 0x0 and 0x1. This will ensure that that their traffic flows will not be affected if some future routing equipment does similar snooping on some future version of IP.

For an example of how ECMP is avoided in Pseudowires, see [RFC4385].

4. Security Considerations

This memo discusses the conditions under which MPLS traffic associated with a single top-level LSP either does or does not have the possibility of being split between multiple paths, implying the possibility of mis-ordering between packets belonging to the same top-level LSP. From a security point of view, the worse that could result from a security breach of the mechanisms described here would be mis-ordering of packets, and possible corresponding loss of throughput (for example, TCP connections may in some cases reduce the window size in response to mis-ordered packets). However, in order to create even this limited result, a hacker would need to either change the configuration or implementation of a router, or change the bits on the wire as transmitted in a packet.

Other security issues in the deployment of MPLS are outside of the scope of this document, but are discussed in other MPLS specifications such as RFCs 3031, 3036, 3107, 3209, 3478, 3479, 4206, 4220, 4221, 4378, AND 4379.

5. References

5.1. Normative References

[RFC3031] Rosen, E. et al., "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.

5.2. Informative References

[RFC3036] Andersson, L., et. al., "LDP Specification", RFC 3036, January 2001.

[RFC3107] Rekhter, Y. and E. Rosen, "Carrying Label Information in BGP-4", RFC 3107, May 2001.

[RFC3209] Awduche, D., et. al., "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.

[RFC3478] Leelanivas, M., et. al., "Graceful Restart Mechanism for Label Distribution Protocol", RFC 3478, February 2003.

[RFC3479] Farrel, A., "Fault Tolerance for the Label Distribution Protocol (LDP)", RFC 3479, February 2003.

[RFC4206] Kompella, K. and Y. Rekhter, "Label Switched Paths (LSP) Hierarchy with Generalized Multi-Protocol Label Switching (GMPLS) Traffic Engineering (TE)", RFC 4206, October 2005.

[RFC4220] Dubuc, M., et. al., "Traffic Engineering Link Management Information Base", RFC 4220, November 2005.

[RFC4221] Nadeau, T., et. al., "Multiprotocol Label Switching (MPLS) Management Overview", RFC 4221, November 2005.

[RFC4378] Allan, D. and T. Nadeau, "A Framework for Multi-Protocol Label Switching (MPLS) Operations and Management (OAM)", RFC 4378, February 2006.

[RFC4379] Kompella, K. and G. Swallow, "Detecting Multi-Protocol Label Switched (MPLS) Data Plane Failures", RFC 4379, February 2006.

[RFC4385] Bryant, S., et. al., "Pseudowire Emulation Edge-to-Edge (PWE3) Control Word for Use over an MPLS PSN", RFC 4385, February 2006.

6. Authors' Addresses

Loa Andersson
Acreo

Email: loa@pi.se

Stewart Bryant
Cisco Systems
250, Longwater,
Green Park,
Reading, RG2 6GB, UK

Email: stbryant@cisco.com

George Swallow
Cisco Systems, Inc.
1414 Massachusetts Ave
Boxborough, MA 01719

Email: swallow@cisco.com

ipr@ietf.org.

Full Copyright Notice