Network Working Group
Internet Draft
Category: Standards Track
Expiration Date: January 2005

George Swallow
Cisco Systems, Inc.

Stewart Bryant
Cisco Systems, Inc.

Loa Andersson
Acreo

July 2004

Avoiding Equal Cost Multipath Treatment in MPLS Networks

draft-swallow-mpls-ecmp-bcp-00.txt

Status of this Memo

By submitting this Internet-Draft, the authors certify that any applicable patent or other IPR claims of which we are aware have been disclosed, and any of which we become aware will be disclosed, in accordance with RFC 3668.

This document is an Internet-Draft and is in full conformance with all provisions of Section 5 of RFC3667. Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/1id-abstracts.html

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html

Copyright Notice

Copyright (C) The Internet Society (2004).

Abstract

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

Contents

1 Introduction 3
2 Current EMCP Practices 3
3 Recommendations for Avoiding ECMP Treatment 4
4 Security Considerations 5
5 References 5
5.1 Normative References 5
6 Authors' Addresses 5
7 Full Copyright and Intellectual Property Statements 6

1. Introduction

This document describes the Equal Cost Multipath (ECMP) behavior of currently deployed MPLS networks and makes best practice recommendations for anyone defining an application to run over an MPLS network and wishes to avoid such treatment. While turning ECMP off is an option open to most operators, few (if any) have chosen to do so. Thus ECMP behavior is a reality that must be reckoned with.

2. Current EMCP Practices

The MPLS label stack and Fowarding 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.

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 in based mainly or exclusively on the IP source and destination addresses. The reason for splitting aggregated flows in this manner is to minimize the mis-ordering of flows between individual IP hosts contained with in the aggregated flow.

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 behavior. As shown above, it is not possible to know whether the payload of an MPLS packet is IP at every place where 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, it can be inferred 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 ECMP.

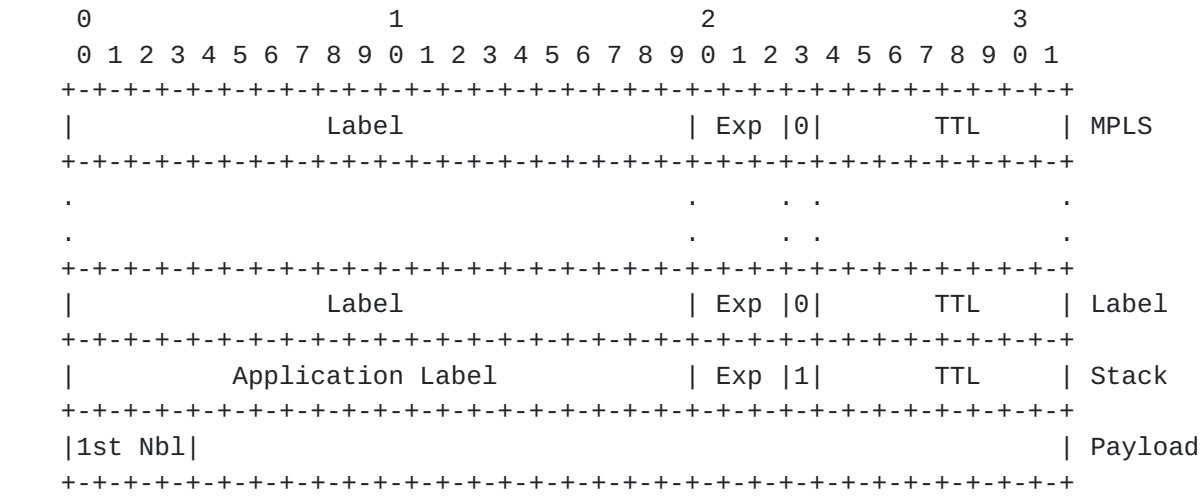
A consequence of this is that any application which defines a FEC which does not take measures to prevent the values 0x4 and 0x6 from

occurring in the first nibble of the payload may be subject to 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.

3. Recommendations for Avoiding ECMP Treatment

The field in the figure below tagged "Application Label" is a label of the FEC Type used/defined by the application. It is the bottom most label in the label stack. As such its FEC Type defines the payload which 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.



In order to avoid 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 that 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.

4. Security Considerations

This memo documents current practices. As such it creates no new security considerations.

5. References

5.1. Normative References

[RFC3031] Rosen, E. et al., "Multiprotocol Label Swithing Architecture", January 2001.

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

7. Full Copyright and Intellectual Property Statements

Copyright (C) The Internet Society (2004). This document is subject to the rights, licenses and restrictions contained in BCP 78, and except as set forth therein, the authors retain all their rights.

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.

Intellectual Property

The IETF takes no position regarding the validity or scope of any Intellectual Property Rights or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; nor does it represent that it has made any independent effort to identify any such rights. Information on the procedures with respect to rights in RFC documents can be found in BCP 78 and BCP 79.

Copies of IPR disclosures made to the IETF Secretariat and any assurances of licenses to be made available, or the result of an 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.

Acknowledgement

Funding for the RFC Editor function is currently provided by the Internet Society.