Guidelines for MPLS Load Balancing

Status of this Memo

This document is an Internet-Draft and is in full conformance with all provisions of Section 10 of RFC2026.

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/ietf/1id-abstracts.txt
The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

Copyright Notice

Copyright (C) The Internet Society (2001). All Rights Reserved.

Abstract

RFC 3031 permits MPLS load balancing while making no specific representations as to implementation requirements. This has subsequently become an issue with respect to the reliability of path test mechanisms. Load balancing algorithms may separate path test probes from the path of interest. This document proposes guidelines for implementation of load balancing such that path test mechanisms are not impacted.
1. Introduction

MPLS load balancing mechanisms forward traffic with a common MPLS destination/forwarding equivalence class (FEC) over multiple parallel paths. The mechanisms to select which path are currently unspecified and frequently have payload and/or label stack dependencies. Protocols used for MPLS LSP verification that are not considered by load balancing path selection mechanisms will not produce consistent results for payload types that are considered, and vice versa. Load spreading may involve examination of network layer addressing information. Even when a common network layer protocol is employed to transport both verification messages and traffic, a verification protocol’s ability to impersonate traffic may be limited.

This document proposes guidelines for implementation of payload/label stack based load balancing such that path test mechanisms used for MPLS label switched path (LSP) verification and payload will have common forwarding behavior. This ensures that the LSP is properly verified by probe messages, and that the time to detect LSP problems is minimized.

The term "level" is significant in this document. The definition is as defined in [RFC 3031]. This document makes the further distinction of "forwarding level" which is the level in the stack exclusive of reserved labels. So, for example, the presence of a router alert label on top of some arbitrary label stack does not alter the level relationship of non-reserved labels.

2. Discussion

The MPLS Architecture [RFC 3031] and diffserv extensions [RFC3270] permit individual instances of FEC to NHLFE (FTN) and Incoming Label Map (ILM) to map to multiple Next Hop Label Forwarding Entries (NHLFEs), with the caveat that for a given packet, only one element from the set of NHLFEs must be selected for use before the packet is forwarded. The selection procedure is unspecified.

The NHLFE selection procedure should have a number of desirable
attributes:

- Minimal re-ordering of packets in a flow. This is achieved by the selection mechanism ensuring packets in the same flow use the same NHLFE.

- Path testing associated with a flow at any forwarding level use the same NHLFE as the flow itself. Otherwise, attempts to proactively detect or diagnose faults will produce inconsistent results. This is because the monitoring probes may use a different NHLFE than the monitored label switched path (LSP).

- Relative evenness in the distribution of traffic over the set of NHLFEs.

- Preservation of diffserv characteristics

One load balancing implementation selects the NHLFE based on a hash the label stack below the load shared level. It is assumed the depth of the stack is typically > 1, and the combination of stack depth, and the number of labels used at any given level will result in a reasonable distribution of traffic across the NHLFEs. Some implementations incorporate MPLS payload information into the NHLFE selection process as well.

As soon as any portion of the payload of an LSP is used as part of the NHLFE selection process, monitoring of that LSP will produce inconsistent results. This behavior is inherent to the load balancing process. The object of this draft is to provide guidelines such that operators may balance the need for testability and operational friendliness with the need for smooth randomization in load balancing.

2.1 Label Stack Entry Fields Modified by Intermediate LSRs

The ability for an LSP specific probe to follow the forwarding path of an LSP requires that some fields in the label stack must be ignored. The field value may vary from packet to packet. An example of this is time to live (TTL) when used in the uniform model [TTL]. TTL reasonably could be expected to be consistent for an IP flow in a converged network (flow being expressed as some variation of a src/dst tuple), but an LSP may aggregate a number of flows, and may use probe packets with different TTL values than the payload. More importantly, incorporating TTL into NHLFE selection would play havoc with any TTL based traceroute mechanism.

2.2 Reserved labels

Reserved labels may be used to define LSP specific functions. A simplistic hash of the stack runs into problems if the hash of the label stack also includes reserved labels for MPLS functions that currently, or in the future, may also require common forwarding with the associated LSP. MPLS reserved functions associated with a specific LSP may resolve to a different NHLFE than the LSP payload.
Reserved labels add to the stack depth (and are referred to as levels in that context), but carry functional rather than forwarding information. Examples would be the proposed OAM alert label [LABEL], the Explicit V4 label or the router alert label. Other examples may emerge in the future.

MPLS reserved labels are infrequently used. The inclusion of reserved label traffic for an LSP in the same NHLFE as the normal payload for that LSP should have negligible impact on the network engineering properties/evenness of distribution of traffic of a load-balanced LSP.

The set of reserved label values ranges from 0 to 15. A simple boolean ‘and’ of the label value with a mask should be sufficient to determine if information in that label stack entry should be included in the NHLFE selection process.

The ‘S’ bit, indicating bottom of stack, does not uniquely identify the presence of forwarding information as it may indicate the presence of a reserved label. The ‘S’ bit should not be incorporated into the selection process.

2.3 Diffserv

The existence of EXP-inferred LSPs (E-LSPs) means that a tested LSP may transport a number of diffserv class types. It would be desirable to be able to test/monitor only the LSP and not have to uniquely test/monitor each class type. To avoid inverse multiplexing of class types, EXP bits must be excluded from the selection process. Note that at a level ingress (either FTN or ILM) the EXP bits (or packet TOS bits) must be interpreted to ensure correct mapping of the diff-serv code point (DSCP as per [RFC3270]). They merely must be excluded from any simple randomization of packet forwarding across a multiplex group.

2.4 Monitored LSPs

An operator may want to monitor LSPs that transport MPLS LSPs. For example, a packet switched network (PSN) trunk for pseudo-wires (PWs). A simplistic hash of all the forwarding labels in the stack will mean that the monitored LSP and the payload of the monitored LSP may not have common forwarding. The hashing approach will only guarantee common forwarding for flows that have identical forwarding components in their label stacks. If the packet payload is also incorporated into the NHLFE selection process, the payload carrying LSP (bottom of the stack) may exhibit similar behavior.

Accommodating this while providing for monitored LSPs is difficult, either:

- Specific LSPs at arbitrary forwarding levels need to be able to be administratively designated as "monitored". This would be both unscalable and operationally intractable.

- A range of label values is designated to specifically identify monitored LSPs (significant backwards compatibility issues)
- The depth of the label stack (and payload) incorporated into the load balancing NHLFE selection process must be able to be administratively set. This trades off some evenness of distribution of traffic for testability and also means un-monitored LSPs will get the same treatment although they do not require it. Operationally this appears to be the most straightforward solution.

3. Guidelines

The following set of guidelines will permit load balancing to co-exist with path oriented verification tools that use reserved labels. It also permits IP tools to exercise load balancing constructs in a fixed amount of time.

Although the discussion in this draft has focused on hashing-based NHLFE selection, the guidelines are sufficiently general to have broader applicability. These are:

1) A NHLFE selection procedures excludes the MPLS stack entries for any MPLS reserved labels [RFC 3032]. NHLFE selection procedures must resolve to the same NHLFE as they would if there were no reserved label(s) present.

2) The NHLFE selection procedures for a stack that contains only reserved labels below the load-balanced forwarding level always resolves to a common NHLFE.

3) NHLFE selection procedures excludes the ’S’ bits from any label stack entries.

4) NHLFE selection procedures excludes the TTL field from any label stack entries.

5) NHLFE selection procedures exclude the EXP bits for the labels incorporated into the selection process beyond ensuring that the selected NHLFE entry supports the outgoing PHB of the forwarded packet (FTN case) and the set of outgoing PHBs required by the ILM (ILM case).

6) The depth of forwarding levels below the top label that is included in NHLFE selection procedures can be administratively configured. Levels with reserved labels do not contribute to depth establishment, nor are they included as per guideline 1 above. Implementations may include label stack forwarding information or packet payload in the selection process providing the depth does not exceed the administratively set boundary. If the level is administratively set to ‘n’, then forwarding labels at level ‘n’ or higher, or the packet payload of level ‘n+1’ or higher may be incorporated into the selection process.

The scenarios supported by these guidelines are:

1) When NHLFE selection input is administratively limited to the top of the stack or unlabelled packet, then testing/monitoring of all
LSPs will produce consistent results. This will be true for both Y.1711 and LSP-PING (this eliminates the need to randomly manipulate the destination address to achieve fate sharing with the LSP under test).

2) When NHLFE selection input is limited to the label stack, and the payload of an individual LSP is either another LSP or an unlabelled packet but not both, then testing/monitoring of all packet carrying LSPs (forwarding depth equals one) will produce consistent results.

3) When NHLFE selection input is limited to the label stack, and the payload of an individual LSP can be another LSP or an unlabelled packet, then testing/monitoring of all LSPs at and below the administratively set level will produce consistent results.

4) When NHLFE selection input may include the label stack and payload then testing/monitoring of all LSPs at and below the administratively set level will produce consistent results.

4. Security Considerations

This draft introduces no new security issues into the MPLS architecture.

5. References


[RFC 3270] Le Faucheur et.al., "MPLS Support of Differentiated Services", IETF RFC 3270, May 2002


6. Acknowledgements

Thanks to Shahram Davari and Neil Harrison for their detailed review of this draft.

7. Author’s Address
David Allan
Nortel Networks          Phone: 1-613-763-6362
3500 Carling Ave.        Email: dallan@nortelnetworks.com
Ottawa, Ontario, CANADA