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 requirements of implementation. 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 I-D proposes guidelines for implementation of load balancing such that path test mechanisms are not impacted.

Sub-IP ID Summary

WHERE DOES IT FIT IN THE PICTURE OF THE SUB-IP WORK

Fits in the MPLS box.

WHY IS IT TARGETED AT THIS WG

It proposes guidelines for packet MPLS specific load balancing techniques.
The WG should consider this document as it addresses an under-specified area of the MPLS architecture.

Table of Contents

1. Conventions used in this document........................................2
2. Discussion.............................................................................2
   2.1 Label Stack Entry Fields Modified by Intermediate LSRs........3
   2.2 Reserved labels..........................................................3
   2.3 Diffserv.................................................................4
   2.4 Monitored LSPs..........................................................4
3. Normative Rules.....................................................................5
4. Security Considerations.....................................................6
5. References............................................................................6
6. Acknowledgements.............................................................6
7. Author’s Address....................................................................6

1. Conventions used in this document

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 [1].

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] permits individual instances of FTN and ILM to map to multiple 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, and the selection procedure is unspecified.

It is well understood that the 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 LSP.

- Relative evenness in the distribution of traffic over the set
- Preservation of diffserv characteristics

One solution commonly implemented is to select 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 is sufficiently large that a reasonable distribution of traffic across the NHLFEs is achieved. Some implementations also examine packet payload (packets with stack depth of zero) and incorporate payload information into the NHLFE selection process as well.

It is understood that as soon as the payload of an LSP (be it another LSP or a packet) is incorporated into the NHLFE selection process, monitoring of that LSP will produce inconsistent results and that 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

A number of label stack entry fields for an LSP may not have consistent values, and therefore could result in the mapping of an LSPs traffic to multiple NHLFEs. An example of this is TTL when used in the uniform model \[\text{TTL}\]. 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 therefore a variety of TTL values may be encountered by the load sharing hash function. This results in the LSPs traffic being distributed across the set of candidate NHLFEs.

2.2 Reserved labels

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. 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 \[\text{LABEL}\], or use of the Explicit V4 label with \[\text{LSP-PING}\]. Other examples may emerge in the future. MPLS reserved functions associated with a specific LSP may resolve to a different NHLFE than the LSP payload.

MPLS reserved labels are infrequently used, therefore 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.

To avoid the reserved label issue, the hash of a label stack should only include label stack entries that specifically pertain to forwarding (the forwarding levels). Reserved labels defining MPLS
specific functions and associated stack indications should be excluded and have no influence on the NHLFE selected.

The set of reserved labels is 0-15 therefore a simple boolean ‘and’ of the label value with a mask should be sufficient to determine if the label should be included in the hash. Similarly, the ‘S’ bit, indicating bottom of stack, does not uniquely identify the presence of forwarding information (it may indicate the presence of a reserved label) therefore it should not be incorporated into the selection process.

2.3 Diffserv

The existence of 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 DSCP (as per [RFC3270]). They merely must be excluded from any simple randomization of packet forwarding across a multiplex group.

2.4 Monitored LSPs

A simplistic hash of all the forwarding labels in the stack can introduce problems if other than the payload carrying LSP is monitored or requires diagnosis of a problem. The hashing approach will only guarantee common forwarding for flows that have identical forwarding components in their label stacks. All LSPs at forwarding levels above the bottom of the stack may be inverse multiplexed arbitrarily across the set of LSPs used for load sharing, which implies partial failure or degradation of all these LSP levels can occur. 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" and therefore requiring common treatment (both unscalable and operationally intractable).

- a range of label values is designated to specifically identify monitored LSPs (significant backwards compatibility issues)

3. Normative Rules
The following set of rules will permit load balancing to co-exist with path oriented verification tools. Although the discussion in this draft has focused on hashing based NHLFE selection, the rules are sufficiently general to have broader applicability. These are:

1) A NHLFE selection procedures MUST NOT include 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 was no reserved label(s) present.

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

3) NHLFE selection procedures MUST NOT include the ‘S’ bits from any label stack entries.

4) NHLFE selection procedures MUST NOT include the TTL field from any label stack entries.

5) NHLFE selection procedures MUST NOT include 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 MUST be able to be administratively configured. Levels with reserved labels do not contribute to depth establishment, nor are they included as per rule 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 administrative 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 rules 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.

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


[TTL] Agarwal et.al. "Time to Live (TTL) Processing in MPLS Networks (Updates RFC 3032)", draft-ietf-mpls-ttl-03 work in progress, June 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
3500 Carling Ave.
Ottawa, Ontario, CANADA
Phone: 1-613-763-6362
Email: dallan@nortelnetworks.com

Expires January 2003