Network Working Group X. Fu(Ed.), M. Betts, Q.
Wang
Internet Draft ZTE
Intended Status: Informational V. Manral
Expires: April 21, 2013 Hewlett-Packard Corp.
D. McDysan (Ed.), A. Malis
Verizon
S. Giacalone
Thomson Reuters
J. Drake
Juniper Networks

October 22, 2012

Loss and Delay Traffic Engineering Framework for MPLS

draft-fuxh-mpls-delay-loss-te-framework-06

Abstract

Deployment and usage of cloud based applications and services that use an underlying MPLS network are expanding and an increasing number of applications are extremely sensitive to delay and packet loss. Furthermore, in cloud computing an additional decision problem arises of simultaneously choosing the data center to host applications along with MPLS network connectivity such that the overall performance of the application is met. Mechanisms exist to measure and monitor MPLS path performance parameters for packet loss and delay, but the mechanisms work only after the path has been setup. The cloud-based and performance sensitive applications would benefit from measurement of MPLS network and potential path information that would be provided for use in the computation before LSP setup and then the selection of LSPs.

This document provides a framework and architecture to solve operator problems and requirements using current/proposed approaches, documents scalability assessment and recommendations, and identifies any needed protocol development.

Table of Contents

1. Introduction

This draft is one of two created from draft-fuxh-mpls-delay-loss-te-framework-05 in response to comments from an MPLS Review Team (RT). This draft focuses on a framework in response to the problem statement and requirements described in a peer document [DELAY-LOSS-PS].

The purpose of this draft is to summarize a framework and architecture to meet requirements using current/proposed approaches, documents scalability assessment and recommendations, and identifies any needed protocol development.

However, computing an LSP path to meet the Network Performance Objective(NPO) for delay, loss and delay variation of these QoS classes is an open problem [DELAY-LOSS-PS]. This draft describes a framework for how the MPLS TE architecture can be augmented use information on configured, measured and/or estimated delay, loss and delay variation for use in LSP path computation and selection.

1.1. Scope

A (G)MPLS network may have multiple layers of packet, TDM and/or optical network technology and an important objective is to make a prediction of end-to-end delay, loss and delay variation based upon the current state of this network with acceptable accuracy before an LSP is established.

The (G)MPLS network may cover a single IGP area/level, may be a hierarchical IGP under control of a single administrator, or may involve multiple domains under control of multiple administrators.

An MPLS architecture for Multicast with awareness of delay, loss and delay variation will be taken up in a future version of the draft.

2. 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 [RFC2119].

2.1. Acronyms

DS-TE	Differentiated Services Traffic Engineering
IGP	Interior Gateway Protocol
(G)MPLS	(Generalized) Multi-Protocol Label Switching
LSP	Label Switched Path
RSVP-TE	Resource reservation Protocol - Traffic Engineering

3. Overview of Functional Requirements

[DELAY-LOSS-PS] describes the general problem to be solved and describes a number of requirements grouped in the following subject areas for performance sensitive LSP computation and placement:

- Augment LSP Requestor Signaling with Performance Parameter Values
- Specify Criteria for Node and Link Performance Parameter Estimation, Measurement Methods
- Support Node Level Performance Information when Needed
- Augment Routing Information with Performance Parameter Estimates
- Augment Signaling Information with Concatenated Estimates
- Define Significant Performance Parameter Change Thresholds and Frequency
- Define Thresholds and Timers for Links with Unusable Performance
- Communicate Significant Performance Changes between Layers
- Support for Networks with Composite Link
- Support Performance Sensitive Restoration, Protection and Rerouting
- Support Management and Operational Requirements

The following sections describe aspects of a framework for each of the above requirement sets in terms of functions, protocols and operational scenarios for meeting the requirements. In some cases the descriptions reference current/proposed potentially applicable IETF approaches. Throughout the following sections, certain scalability challenges are identified and in most cases a potential resolution approach is described - these are summarized at the end of the document.

4. Augment LSP Requestor Signaling with Performance Parameter Values

As described in [DELAY-LOSS-PS] the LSP requestor must be able to make a request for one of two types 1) a minimum possible value or 2)a maximum acceptable valuefor each performance parameter for each LSP.

The proposed approach [EXPRESS-PATH] within a single IGP area/level, is that only the origin (or head-end) need be aware of the required performance aspects of the LSP, since the origin has performance information for all of the candidate nodes and links from a performance parameter augmented IGP [OSPF-TE-METRIC-EXT], [ISIS-TE-METRIC-EXT].

For LSPs that traverse multiple area/levels or multiple domains, what is needed in addition to [EXPRESS-PATH] is knowledge of the node and link level performance to determine a path that meets the concatenated performance estimates as described in [DELAY-LOSS-PS]. Furthermore, information available to the LSP originator (e.g., the request type (minimum possible value, maximum acceptable parameter value) may need to be carried in the RSVP_TE signaling message.

An alternative approach could make the performance information available to a (set of) Path Computation Elements (PCE), which the LSP requestor could consult. In this case, there would likely need to be extensions made to the PCE Protocol to carry LSP performance parameter information.

5. Specify Criteria for Node and Link Performance Parameter Estimation, Measurement Methods

Procedures to measure delay and loss on a path level between measurement points have been specified in ITU-T [Y.1731], [G.709] and [RFC 6374]. Ideally, a measurement point would occur within adjacent nodes to measure the delay, loss and delay variation performance for a combination of node and link performance. However, since this method is not universally deployed (and may never be deployed in some nodes), other methods of performance parameter estimation are needed to meet the requirements of [DELAY-LOSS-PS].

Important assumptions from [DELAY-LOSS-PS] are:

o the timeframe of the performance parameter estimate, which is specified as the order of minutes

o delay and loss are defined as an average and delay variation is defined based upon statistical quantiles

These assumptions could allow other methods to estimate performance parameters, such as usage of models to predict values based upon other parameters, such as load, queue thresholds and/or meters. For example, one such method could be a per QoS class based measurement from the ingress of one port to the egress of another port on a node as a function of load in a field test or laboratory to create an empirical model that could be used to insert performance parameter estimates into routing or signaling.

The switching delay on a node can be measured internally, and multiple mechanisms and data structures to do this have been defined [LEE].

6. Support Node Level Performance Information when Needed

If the IGP structure of link-level advertisements is to be used, then nodal delays can be combined with link-level performance [EXPRESS-PATH]. For example, a solution provide configuration knob to add some fixed value of a portion (e.g., one half) of node delay to link delay.

Alternatively, IGPs or a PCE information base could be extended with node-level performance parameter estimates.

7. Augment Routing Information with Performance Parameter Estimates

[DSTE-PROTO] and [EXPRESS-PATH] use information regarding bandwidth from an IGP area/level for use by performance sensitive LSPs. For a single IGP area/level, the IGP could be augmented with estimates of delay, loss and delay variation as described in [OSPF-TE-METRIC-EXT], [ISIS-TE-METRIC-EXT]. This should also apply to a Forwarding Adjacency LSP (FA-LSP) [RFC4206]. [EXPRESS-PATH] describes how to use these augmented IGP performance measures to compute explicit paths, for example, at a path computation entity.

For LSPs that cross an IGP area/level boundary and/or traverse multiple domains, some other solution is needed for LSP path computation and selection, such as augmented PCE information bases. These PCE information bases can then be used by origin or the Path Computation engine to decide paths with the desired path properties.

Routing information could use two components to represent performance, "static" and "dynamic". The dynamic component is that caused by traffic load and queuing and would be an approximate value. The static component should be fixed and independent of load (e.g., propagation delay).

8. Augment Signaling Information with Concatenated Estimates

[DELAY-LOSS-PS] cites specific sections/appendices from [ITU-T Y.1541] regarding how performance estimates are to be composed and concatenated.

For LSPs that cross an IGP area/level boundary and/or traverse multiple domains (e.g., Autonomous Systems), if detailed performance parameter information is not provided, then one approach would be to signal the requested performance parameters for the LSP in the RSVP_TE signaling message as described in [DELAY-LOSS-RSVP-TE]. If each area/level and/or domain is unaware of the composition of performance parameters from the prior area/level and/or domains, then signaling would also need to carry the concatenation of these composed performance estimates.

Signaling information could use two components to represent performance, "static" and "dynamic". The dynamic component is that caused by traffic load and queuing and would be an approximate value. The static component should be fixed and independent of load (e.g., propagation delay).

RSVP-TE signaling across multiple area/levels or domains could include recording status of previous attempts, retries and correlation with end-end LSP performance measures to improve on a trial-and-error approach.

Another approach that could meet the requirements could be a (stateful) PCE listening to each domain, communicating amongst PCEs in other domains approximating global state to reduce probing and retries to improve scalability.

9. Define Significant Performance Parameter Change Thresholds and Frequency

In the augmented IGP approach, performance value changes should be updated and flooded in the IGP only when there is significant change in the value. The LSP originator could determine the IGP update affects performance and can decide on whether to accept the changed value, or request another computation of the LSP.

Since performance characteristics of links, nodes and FA-LSPs may change dynamically the amount of information flooded in an augmented IGP approach could be excessive and cause instability. In order to control IGP messaging and avoid being unstable when the delay, delay variation and packet loss value changes, thresholds and a limit on rate of change should be configured in the IGP control plane.

10. Define Thresholds and Timers for Links with Unusable Performance

For the extended IGP or augmented PCE information base approaches, an acceptable and unacceptable target performance value could be configured for each link (and node, if supported). This should also apply to a Forwarding Adjacency LSP (FA-LSP) [RFC4206]. If a measured or dynamically estimated (e.g., based upon load) performance value increases above the unacceptable threshold, the link (node) could be removed from consideration for future LSP path computations. If it decreases below the acceptable target value, it can then be considered for future LSP path computations.

Performance-sensitive LSPs whose path traverses links (nodes) whose performance has been deemed unacceptable by this threshold should be notified. The LSP originator can then decide if it will accept the changed performance, or else request computation of a new path that meets the performance objective.

The frequency of a link (node) changing from an unacceptable to an acceptable state should be controlled by configurable parameters.

11. Communicate Significant Performance Changes between Layers

The generic requirement is for a lower layer network to communicate significant performance changes to a higher layer network.

An end-to-end LSP (e.g., in IP/MPLS or MPLS-TP network) may traverse a FA-LSP of a server layer (e.g., an OTN ring). The boundary nodes of the FA-LSP SHOULD be aware of the performance information for this FA-LSP.

If the FA-LSP is used to form a routing adjacency and/or used as a TE link in the client network, the composition of the performance values of the links and nodes that the FA-LSP trail traverses needs to be made available for path computation. This is especially important when the performance information of the FA-LSP changes (e.g., due to a maintenance action or failure in an OTN ring).

The frequency of a lower layer network indicating a significant performance change should be controlled by configurable parameters.

A separate end-end performance measurement could be done for an LSP after it has been established (e.g., RFC 6374) if it is a lower level FA-LSP used in an LSP hierarchy. The measurement of end-to-end LSP performance may be used to inform the higher layer network of a performance parameter change.

If the performance of FA-LSP changes, the client layer must at least be notified. The client layer can then decide if it will accept the changed performance, or else request computation of a new path that meets the performance objective.

12. Support for Networks with Composite Links

In order to assign the LSP to one of component links with different performance characteristics [CL-REQ], the RSVP-TE message could carry a indication of the request type (i.e., minimum possible value or a maximum acceptable performance parameter value) for use in component link selection or creation. The composite link should be able to take these parameters into account when assigning LSP traffic to a component link.

When Composite Links [CL-REQ] are advertised into an augmented IGP, the desirable solution is to advertise performance information for all component links into the augmented IGP [CL-FW]. Otherwise, if only partial or summarized information is advertised then the originator or a PCE cannot determine whether a computed path will meet the LSP performance objective and this could lead to crank back signaling.

13. Support Performance Sensitive Restoration, Protection and Rerouting

A change in performance of links and nodes (e.g., due to a lower level restoration action) may affect the performance of one or more end-to-end LSPs. Pre-defined protection or dynamic re-routing could be triggered to handle this case.

In the case of predefined protection, large amounts of redundant capacity may have a significant negative impact on the overall network cost. If the LSP performance objective cannot be met after a re-route is attempted, an alarm should be generated to the management plane. The solution should periodically attempt restoration for as controlled by configuration parameters to prevent excessive load on the control plane.

14. Support Management and Operational Requirements

A separate end-end performance measurement should be done for an LSP after it has been established (e.g., RFC 6374, G.709 or Y.1731). An LSP originator may re-compute a re-signal a path when the measured end-to-end performance is unacceptable. The choice by the originator to re-signal could consider a history of how accurate the performance parameter estimate is delivered by the implementation. The re-computation and re-signaling rates should be controlled by configuration parameters to prevent excessive load on the control plane.

15. Major Architectural and Scaling Challenges

As described in the preceding sections, there are a several scaling and architectural challenges, with proposed resolution as described below:

o Frequency of performance parameter value changes limited to the order of minutes by definition

- Augmented IGP flooding performance parameter change frequency within one area/level controlled by configuration parameters

- Augmented PCE information base performance parameter change frequency within one area/level controlled by configuration parameters

- Re-computation and re-signaling of LSPs whose composition of performance parameter values changes to unacceptable controlled by configuration parameters

- Declaration of links, nodes, FA-LSPs as unacceptable/acceptable controlled by configuration parameters

- Frequency of a lower layer network indicating a significant performance change controlled by configuration parameters

- Re-computation and re-signaling of LSPs whose measured end-end performance is unacceptable controlled by configuration parameters

16. Approaches Considered but not Taken

One approach would be for the PCE to compute paths for use by the LSP originator for signaling. Some measurement method (e.g., RFC 6374) could then be used to measure the performance of this path. If the measurement indicates that the performance is not met then another request is made to the PCE for a different path, the originator signals for the LSP to be set up and then measured again. This "trial and error" process is very inefficient and a more predictable method is required.

17. IANA Considerations

No new IANA consideration are raised by this document.

18. Security Considerations

This document raises no new security issues.

19. References

19.1. Normative References

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

19.2. Informative References

[DELAY-LOSS-PS] X.Fu, D. McDysan et al., "Delay and Loss Traffic Engineering Problem Statement for MPLS," draft-fuxh-mpls-delay-loss-te-problem-statement

[DSTE-PROTO] Le Faucheur, F., Ed., "Protocol Extensions for Supportof Diffserv-aware MPLS Traffic Engineering", RFC 4124, June 2005.

[ISIS-TE-METRIC-EXT] S. Previdi, "IS-IS Traffic Engineering (TE) Metric Extensions", draft-previdi-isis-te-metric-extensions.

[OSPF-TE-METRIC-EXT] S. Giacalone, "OSPF Traffic Engineering (TE)
Metric Extensions", draft-ietf-ospf-te-metric-extensions.

[EXPRESS-PATH] A. Atlas et al, "Performance-based Path Selection for Explicitly Routed LSPs", draft-atlas-mpls-te-express-path.

[Y.1731] ITU-T Recommendation Y.1731, "OAM functions and mechanisms for Ethernet based networks", Feb 2008.

[G.709] ITU-T Recommendation G.709, "Interfaces for the Optical Transport Network (OTN)", December 2009.

[RFC 6374] D. Frost, S. Bryant, "Packet Loss and Delay Measurement for MPLS Networks," RFC 6374, September 2011.

[DELAY-LOSS-RSVP-TE] X. Fu, "RSVP-TE extensions for Delay and Loss Traffic Engineering", draft-fuxh-mpls-delay-loss-rsvp-te-ext.

[ITU-T.Y.1541] ITU-T, "Network performance objectives for IP-based services", 2011, <http://www.itu.int/rec/T-REC-Y.1541/en>.

[CL-REQ] C. Villamizar, "Requirements for MPLS Over a Composite Link", draft-ietf-rtgwg-cl-requirement

[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.

[CL-FW] C. Villamizar et al, "Composite Link Framework in Multi Protocol Label Switching (MPLS)", draft-ietf-rtgwg-cl-framework

[LEE] Myungjin Lee , Sharon Goldberg , Ramana Rao Kompella , George Varghese "Fine-Grained Latency and Loss Measurements in the Presence of Reordering," http://www.cs.bu.edu/fac/goldbe/papers/sigmet2011.pdf

20. Acknowledgments

This document was prepared using 2-Word-v2.0.template.dot.

The authors would like to thank the MPLS Review Team of Stewart Bryant, Daniel King and He Jia for their many helpful comments suggestions in July 2012.

```
This code was derived from IETF RFC [insert RFC number]. Please
reproduce this note if possible.
```

Authors' Addresses

Xihua Fu
ZTE
Email: fu.xihua@zte.com.cn

Vishwas Manral
Hewlett-Packard Corp.
191111 Pruneridge Ave.
Cupertino, CA 95014
US
Phone: 408-447-1497
Email: vishwas.manral@hp.com

Dave McDysan
Verizon
Email: dave.mcdysan@verizon.com

Andrew Malis
Verizon
Email: andrew.g.malis@verizon.com

Spencer Giacalone
Thomson Reuters
195 Broadway
New York, NY 10007
US
Phone: 646-822-3000
Email: spencer.giacalone@thomsonreuters.com

Malcolm Betts
ZTE
Email: malcolm.betts@zte.com.cn

Qilei Wang
ZTE
Email: wang.qilei@zte.com.cn

John Drake
Juniper Networks
Email: jdrake@juniper.net