Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 10, 2012

X. Fu
ZTE
V. Manral
Hewlett-Packard Corp.
D. McDysan
A. Malis
Verizon
S. Giacalone
Thomson Reuters
M. Betts
Q. Wang
ZTE
J. Drake
Juniper Networks
October 8, 2011

Traffic Engineering architecture for services aware MPLS
draft-fuxh-mpls-delay-loss-te-framework-02

Abstract

With more and more enterprises using cloud based services, the distances between the user and the applications are growing. A lot of the current applications are designed to work across LAN's and have various inherent assumptions. For multiple applications such as High Performance Computing and Electronic Financial markets, the response times are critical as is packet loss, while other applications require more throughput.

[RFC3031] describes the architecture of MPLS based networks. This draft extends the MPLS architecture to allow for latency, loss and jitter as properties. It describes requirements and control plane implication for latency and packet loss as a traffic engineering performance metric in today's network which is consisting of potentially multiple layers of packet transport network and optical transport network in order to make a accurate end-to-end latency and loss prediction before a path is established.

Note MPLS architecture for Multicast will be taken up in a future version of the draft.

Requirements Language

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].

Status of this Memo

This Internet-Draft will expire on April 10, 2012.

Copyright Notice

Table of Contents

1. Introduction

In High Frequency trading for Electronic Financial markets, computers make decisions based on the Electronic Data received, without human intervention. These trades now account for a majority of the trading volumes and rely exclusively on ultra-low-latency direct market access.

Extremely low latency measurements for MPLS LSP tunnels are defined in [draft-ietf-mpls-loss-delay]. They allow a mechanism to measure and monitor performance metrics for packet loss, and one-way and two-way delay, as well as related metrics like delay variation and channel throughput.

The measurements are however effective only after the LSP is created and cannot be used by MPLS Path computation engine to define paths that have the latest latency. This draft defines the architecture used, so that end-to-end tunnels can be set up based on latency, loss or jitter characteristics.

End-to-end service optimization based on latency and packet loss is a key requirement for service provider. This type of function will be adopted by their "premium" service customers. They would like to pay for this "premium" service. Latency and loss on a route level will help carriers' customers to make his provider selection decision.

2. Architecture requirements overview

2.1. Communicate Latency and Loss as TE Metric

The solution MUST provide a means to communicate latency, latency variation and packet loss of links and nodes as a traffic engineering performance metric into IGP.

Latency, latency variation and packet loss may be unstable, for example, if queueing latency were included, then IGP could become unstable. The solution MUST provide a means to control latency and loss IGP message advertisement and avoid unstable when the latency, latency variation and packet loss value changes.

Path computation entity MUST have the capability to compute one end-to-end path with latency and packet loss constraint. For example, it has the capability to compute a route with X amount of bandwidth with less than Y ms of latency and Z% packet loss limit based on the latency and packet loss traffic engineering database. It MUST also support the path computation with routing constraints combination with pre-defined priorities, e.g., SRLG diversity, latency, loss and

cost.

2.2. Requirement for Composite Link

One end-to-end LSP may traverses some Composite Links [CL-REQ]. Even if the transport technology (e.g., OTN) component links are identical, the latency and packet loss characteristics of the component links may differ.

The solution MUST provide a means to indicate that a traffic flow should select a component link with minimum latency and/or packet loss, maximum acceptable latency and/or packet loss value and maximum acceptable delay variation value as specified by protocol. The endpoints of Composite Link will take these parameters into account for component link selection or creation. The exact details for component links will be taken up seperately and are not part of this document.

2.3. Requirement for Hierarchy LSP

One end-to-end LSP may traverse a server layer. There will be some latency and packet loss constraint requirement for the segment route

in server layer.

The solution MUST provide a means to indicate FA selection or FA-LSP creation with minimum latency and/or packet loss, maximum acceptable latency and/or packet loss value and maximum acceptable delay variation value. The boundary nodes of FA-LSP will take these parameters into account for FA selection or FA-LSP creation.

2.4. Latency Accumulation and Verification

The solution SHOULD provide a means to accumulate (e.g., sum) of latency information of links and nodes along one LSP across multi-domain (e.g., Inter-AS, Inter-Area or Multi-Layer) so that an latency validation decision can be made at the source node. One-way and round-trip latency collection along the LSP by signaling protocol and latency verification at the end of LSP should be supported.

The accumulation of the delay is "simple" for the static component i.e. its a linear addition, the dynamic/network loading component is more interesting and would involve some estimate of the "worst case". However, method of deriving this worst case appears to be more in the scope of Network Operator policy than standards i.e. the operator needs to decide, based on the SLAs offered, the required confidence level.

2.5. Restoration, Protection and Rerouting

Some customers may insist on having the ability to re-route if the latency and loss SLA is not being met. If a "provisioned" end-to-end LSP latency and/or loss could not meet the latency and loss agreement between operator and his user, the solution SHOULD support pre-defined or dynamic re-routing to handle this case based on the local policy.

If a "provisioned" end-to-end LSP latency and/or loss performance is improved (i.e., beyond a configurable minimum value) because of some segment performance promotion, the solution SHOULD support the re-routing to optimize latency and/or loss end-to-end cost.

The latency performance of pre-defined protection or dynamic re-

routing LSP MUST meet the latency SLA parameter. The difference of latency value between primary and protection/restoration path SHOULD be zero.

As a result of the change of latency and loss in the LSP, current LSP may be frequently switched to a new LSP with a appropriate latency and packet loss value. In order to avoid this, the solution SHOULD indicate the switchover of the LSP according to maximum acceptable change latency and packet loss value.

3. End-to-End Latency

Procedures to measure latency and loss has been provided in ITU-T [Y.1731], [G.709] and [ietf-mpls-loss-delay]. The control plane can be independent of the mechanism used and different mechanisms can be used for measurement based on different standards.

Latency on a path has two sources: Node latency which is caused by the node as a result of process time in each node and: Link latency as a result of packet/frame transit time between two neighbouring nodes or a FA-LSP/ Composite Link [CL-REQ].

Latency or one-way delay is the time it takes for a packet within a stream going from measurement point 1 to measurement point 2.

The architecture uses assumption that the sum of the latencies of the individual components approximately adds up to the average latency of an LSP. Though using the sum may not be perfect, it however gives a good approximation that can be used for Traffic Engineering (TE) purposes.

The total latency of an LSP consists of the sum of the latency of the

LSP hop, as well as the average latency of switching on a device, which may vary based on queuing and buffering.

Hop latency can be measured by getting the latency measurement between the egress of one MPLS LSR to the ingress of the nexthop LSR. This value may be constant for most part, unless there is protection switching, or other similar changes at a lower layer.

The switching latency on a device, can be measured internally, and multiple mechanisms and data structures to do the same have been defined. Add references to papers by Verghese, Kompella, Duffield. Though the mechanisms define how to do flow based measurements, the amount of information gathered in such a case, may become too cumbersome for the Path Computation element to effectively use.

An approximation of Flow based measurement is the per DSCP value, measurement from the ingress of one port to the egress of every other port in the device.

Another approximation that can be used is per interface DSCP based measurement, which can be an agrregate of the average measurements per interface. The average can itself be calculated in ways, so as to provide closer approximation.

For the purpose of this draft it is assumed that the node latency is a small factor of the total latency in the networks where this solution is deployed. The node latency is hence ignored for the benefit of simplicity.

The average link delay over a configurable interval should be reported by data plane in micro-seconds.

4. End-to-End Jitter

Jitter or Packet Delay Variation of a packet within a stream of packets is defined for a selected pair of packets in the stream going from measurement point 1 to measurement point 2.

The architecture uses assumption that the sum of the jitter of the individual components approximately adds up to the average jitter of an LSP. Though using the sum may not be perfect, it however gives a good approximation that can be used for Traffic Engineering (TE) purposes.

There may be very less jitter on a link-hop basis.

The buffering and queuing within a device will lead to the jitter.

Just like latency measurements, jitter measurements can be

appproximated as either per DSCP per port pair (Ingresss and Egress) or as per DSCP per egress port.

For the purpose of this draft it is assumed that the node latency is a small factor of the total latency in the networks where this solution is deployed. The node latency is hence ignored for the benefit of simplicity.

The jitter is measured in terms of 10's of nano-seconds.

5. End-to-End Loss

Loss or Packet Drop probability of a packet within a stream of packets is defined as the number of packets dropped within a given interval.

The architecture uses assumption that the sum of the loss of the individual components approximately adds up to the average loss of an LSP. Though using the sum may not be perfect, it however gives a good approximation that can be used for Traffic Engineering (TE) purposes.

There may be very less loss on a link-hop basis, except in case of physical link issues.

The buffering and queuing mechanisms within a device will decide which packet is to be dropped. Just like latency and jitter measurements, the loss can best be appproximated as either per DSCP per port pair (Ingresss and Egress) or as per DSCP per egress port.

The loss is measured in terms of the number of packets per million packets.

6. Protocol Considerations

The protocol metrics above can be sent in IGP protocol packets RFC 3630. They can then be used by the Path Computation engine to decide paths with the desired path properties.

As Link-state IGP information is flooded throughout an area, frequent changes can cause a lot of control traffic. To prevent such flooding, data should only be flooded when it crosses a certain configured maximum.

A seperate measurement should be done for an LSP when it is UP. Also

LSP's path should only be recalculated when the end-to-end metrics changes in a way it becomes more than desired.

7. Control Plane Implication

7.1. Implications for Routing

The latency and packet loss performance metric MUST be advertised into path computation entity by IGP (etc., OSPF-TE or IS-IS-TE) to perform route computation and network planning based on latency and packet loss SLA target.

Latency, latecny variation and packet loss value MUST be reported as a average value which is calculated by data plane.

Latency and packet loss characteristics of these links and nodes may change dynamically. In order to control IGP messaging and avoid being unstable when the latency, latency variation and packet loss value changes, a threshold and a limit on rate of change MUST be configured to control plane.

If any latency and packet loss values change and over than the threshold and a limit on rate of change, then the latency and loss change of link MUST be notified to the IGP again. The receiving node detrimines whether the link affects any of these LSPs for which it is ingress. If there are, it must determine whether those LSPs still meet end-to-end performance objectives.

A minimum value MUST be configured to control plane. If the link performance improves beyond a configurable minimum value, it must be re-advertised. The receiving node detrimines whether a "provisioned" end-to-end LSP latency and/or loss performance is improved because of some segment performance promotion.

It is sometimes important for paths that desire low latency is to avoid nodes that have a significant contribution to latency. Control plane should report two components of the delay, "static" and "dynamic". The dynamic component is always caused by traffic loading and queuing. The "dynamic" portion SHOULD be reported as an approximate value. It should be a fixed latency through the node without any queuing. Link latency attribute should also take into account the latency of node, i.e., the latency between the incoming port and the outgoing port of a network element. Half of the fixed node latency can be added to each link.

When the Composite Links [CL-REQ] is advertised into IGP, there are following considerations.

- o One option is that the latency and packet loss of composite link may be the range (e.g., at least minimum and maximum) latency value of all component links. It may also be the maximum latency value of all component links. In both cases, only partial information is transmited in the IGP. So the path computation entity has insufficient information to determine whether a particular path can support its latency and packet loss requirements. This leads to signaling crankback.

- o Another option is that latency and packet loss of each component link within one Composite Link could be advertised but having only one IGP adjacency.

One end-to-end LSP (e.g., in IP/MPLS or MPLS-TP network) may traverse a FA-LSP of server layer (e.g., OTN rings). The boundary nodes of the FA-LSP SHOULD be aware of the latency and packet loss information of this FA-LSP.

If the FA-LSP is able to form a routing adjacency and/or as a TE link in the client network, the total latency and packet loss value of the FA-LSP can be as an input to a transformation that results in a FA traffic engineering metric and advertised into the client layer routing instances. Note that this metric will include the latency and packet loss of the links and nodes that the trail traverses.

If total latency and packet loss information of the FA-LSP changes (e.g., due to a maintenance action or failure in OTN rings), the boundary node of the FA-LSP will receive the TE link information advertisement including the latency and packet value which is already changed and if it is over than the threshold and a limit on rate of change, then it will compute the total latency and packet value of the FA-LSP again. If the total latency and packet loss value of FA-LSP changes, the client layer MUST also be notified about the latest value of FA. The client layer can then decide if it will accept the increased latency and packet loss or request a new path that meets the latency and packet loss requirement.

7.2. Implications for Signaling

In order to assign the LSP to one of component links with different latency and loss characteristics, RSVP-TE message needs to carry a indication of request minimum latency and/or packet loss, maximum acceptable latency and/or packet loss value and maximum acceptable delay variation value for the component link selection or creation. The composite link will take these parameters into account when assigning traffic of LSP to a component link.

One end-to-end LSP (e.g., in IP/MPLS or MPLS-TP network) may traverse

a FA-LSP of server layer (e.g., OTN rings). There will be some latency and packet loss constraint requirement for the segment route in server layer. So RSVP-TE message needs to carry a indication of request minimum latency and/or packet loss, maximum acceptable latency and/or packet loss value and maximum acceptable delay variation value. The boundary nodes of FA-LSP will take these parameters into account for FA selection or FA-LSP creation.

RSVP-TE needs to be extended to accumulate (e.g., sum) latency information of links and nodes along one LSP across multi-domain (e.g., Inter-AS, Inter-Area or Multi-Layer) so that an latency verification can be made at end points. One-way and round-trip latency collection along the LSP by signaling protocol can be supported. So the end points of this LSP can verify whether the total amount of latency could meet the latency agreement between operator and his user. When RSVP-TE signaling is used, the source can determine if the latency requirement is met much more rapidly than performing the actual end-to-end latency measurement.

Restoration, protection and equipment variations can impact "provisioned" latency and packet loss (e.g., latency and packet loss increase). For example, restoration/provisioning action in transport network that increases latency seen by packet network observable by customers, possibly violating SLAs. The change of one end-to-end LSP latency and packet loss performance MUST be known by source and/or sink node. So it can inform the higher layer network of a latency and packet loss change. The latency or packet loss change of links and nodes will affect one end-to-end LSPs total amount of latency or packet loss. Applications can fail beyond an application-specific threshold. Some remedy mechanism could be used.

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. Service provider may have many layers of pre-defined restoration for this transfer, but they have to duplicate restoration resources at significant cost. Solution should provides some mechanisms to avoid the duplicate restoration and reduce the network cost. Dynamic re-routing also has to face the risk of resource limitation. So the choice of mechanism MUST be based on SLA or policy. In the case where the latency SLA can not be met after a re-route is attempted, control plane should report an alarm to management plane. It could also try restoration for several times which could be configured.

8. IANA Considerations

No new IANA consideration are raised by this document.

9. Security Considerations

This document raises no new security issues.

10. Acknowledgements

TBD.

11. References

11.1. Normative References

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

[RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.

[RFC3473] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.

[RFC3477] Kompella, K. and Y. Rekhter, "Signalling Unnumbered Links in Resource ReSerVation Protocol - Traffic Engineering (RSVP-TE)", RFC 3477, January 2003.

[RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, September 2003.

[RFC4203] Kompella, K. and Y. Rekhter, "OSPF Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4203, October 2005.

11.2. Informative References

[CL-REQ] C. Villamizar, "Requirements for MPLS Over a Composite Link", draft-ietf-rtgwg-cl-requirement-04 .

[EXPRESS-PATH]

S. Giacalone, "OSPF Traffic Engineering (TE) Express Path", draft-giacalone-ospf-te-express-path-01 .

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

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

[ietf-mpls-loss-delay]
D. Frost, "Packet Loss and Delay Measurement for MPLS Networks", draft-ietf-mpls-loss-delay-03 .

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
URI:

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
URI:

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