Network Working Group
Internet Draft
Intended status: Proposed Standard
Expires: September 2011

S. Giacalone
Thomson Reuters

D. Ward
Juniper Networks

J. Drake
Juniper Networks

A. Atlas
Juniper Networks

March 4, 2011

OSPF Traffic Engineering (TE) Express Path
draft-giacalone-ospf-te-express-path-00.txt

Abstract

In certain networks, such as, but not limited to, financial information networks (e.g. stock market data providers), network performance criteria (e.g. latency) have become (or are becoming) as (or more) critical to data path selection than other metrics.

This document describes extensions to OSPF TE (RFC3630) such that network performance information can be distributed and collected in a scalable fashion. The information collected from OSPF TE Express Path can then be used to make path selection decisions. Additionally, the information passed in these extensions will permit granular network performance monitoring.

Note that this document only covers the mechanisms with which network performance information is distributed. The mechanisms for measuring network performance or acting on that information, once distributed, are outside the scope of this document.

Status of this Memo

Copyright Notice

Table of Contents

1. Introduction

In certain networks, such as, but not limited to, financial information networks (e.g. stock market data providers), network performance information (e.g. latency) have (or are becoming) as (or more) critical to data path selection than other metrics. In many of these networks, bandwidth is relatively rich and homogeneous (e.g. a core network of all 10 or 20 Gigabit Ethernet links, or greater), however path length (and therefore latency) can vary in between end-points (e.g. PE nodes), and segment length or latency can change based on the path protection scheme used. In these networks, extremely large amounts of money rest on the ability to predictably make trades faster than the competition and the ability to access real time market data.

In certain financial services networks, hop count, cost, and bandwidth are only tangentially important. Rather, it would be beneficial to be able to granularly monitor network performance and/or make path selection decisions based on performance data (such as latency) in a cost-effective and scalable way. In addition, since these networks may be built as overlays on top of multiple service provider networks, strict link-by-link service level agreement monitoring and enforcement mechanisms are needed.

This document describes extensions to OSPF TE (hereafter called "OSPF TE Express Path"), that can be used to distribute various pieces of network performance information (such as link latency). The mechanisms described in this document only disseminate performance information. The methods for initially gathering that performance information, or acting on it once it is distributed are outside the scope of this document. OSPF Express Path provides a number of benefits:

The data distributed by OSPF TE Express Path can be used to make path selection decisions. Using the link-by-link performance information data distributed by OSPF TE Express Path, end-to-end path selection can be performed based on performance metrics, as part of the normal operation of various routing protocols (e.g. by replacing cost with latency) or by using "second order" control plane protocols such as CSPF, RSVP-TE [RFC3209], etc.

OSPF TE Express Path enables a scalable, open mechanism for link-by-link SLA compliance monitoring, which is an important issue in large, diverse networks that use transport services from various providers. In networks like this, end-to-end latency is not always useful for enforcement of "underlying" SLAs (since various links from different providers may make up a path). This link-by-link performance monitoring data could easily be gathered by looking at a routing protocol's state database (on any router in an area, depending on what is being monitoring and disseminated by the routing protocol), using SNMP [RFC1441] on a per device basis, or in other ways.

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].

In this document, these words will appear with that interpretation only when in ALL CAPS. Lower case uses of these words are not to be interpreted as carrying RFC-2119 significance.

3. Express Path Extensions to OSPF TE

The extensions in this document build on the ones provided in OSPF TE (RFC3630) and GMPLS (RFC4203) to permit path selection and network monitoring based on various network performance items. As such, this document proposes new OSPF TE sub-TLVs that can be announced in OSPF TE LSAs. OSPF TE LSAs (RFC3630) are opaque LSAs (RFC5250) with area flooding scope. Each TLV has one or more nested sub-TLVs which permit the TE LSA to be readily extended. There are two main types of OSPF TE LSA; the Router Address or Link TE LSA. Like the GMPLS extensions (RFC4203), this document proposes additional sub-TLVs for the Link TE LSA. As background, all OSPF TE TLVs and sub-TLVs use the same general format (RFC3630):

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|              Type             |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            Value...                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

As per (RFC3630) the Length field defines the length of the value portion of the sub-TLV in octets (thus a TLV with no value portion would have a length of zero). TLVs are padded to four-octet alignment; padding is not included in the length field (so a three octet value would have a length of three, but the total size of the TLV would be eight octets). Unrecognized types are ignored.

OSPF TE Express Path defines several new sub-TLVs. These sub-TLVs fall into 2 distinct categories; “Routine” or “Significant”. Routine and Significant sub-TLVs are intended to be used for different purposes (i.e. monitoring or control plane manipulation, respectively). The technical differences between Routine and Significant sub-TLVs are related to the averaging periodicity and announcement frequency of each category of sub-TLV. More information on this subject can be found in section 5.

The following sub-TLVs are defined in OSPF TE Express Path:

Value	Length	Name

TBD1	4	Routine Unidirectional Link Delay
TBD2	4	Routine Unidirectional Delay Variation
TBD3	4	Routine Unidirectional Link Loss
TBD4	4	Significant Unidirectional Link Delay
TBD5	4	Significant Unidirectional Link Loss

4. Sub TLV Details

4.1. Routine Unidirectional Link Delay Sub-TLV

This TLV advertises the average link delay between two directly connected OSPF neighbors. The delay advertised by this sub TLV MUST be the delay from the local neighbor to the remote one (i.e. the forward path latency). The format of this sub-TLV is shown in the following diagram:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|             TBD1              |               4               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                             Delay                             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

4.1.1. Type

This sub-TLV has a type of TBD1

4.1.2. Length

The length is 4

4.1.3. Delay Value

This field carries the average link delay over a configurable interval in micro-seconds, encoded as an IEEE floating point single precision value.

4.2. Routine Unidirectional Delay Variation Sub-TLV

This TLV advertises the average link delay variation between two directly connected OSPF neighbors. The delay variation advertised by this sub-TLV MUST be the delay from the local neighbor to the remote one (i.e. the forward path latency). The format of this sub-TLV is shown in the following diagram:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|              TBD2             |               4               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                        Delay Variation                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

4.2.1. Type

This sub-TLV has a type of TBD2

4.2.2. Length

The length is 4

4.2.3. Delay Variation

This field carries the average link delay variation over a configurable interval in micro-seconds, encoded as an IEEE floating point single precision value.

4.3. Routine Unidirectional Link Loss Sub TLV

This TLV advertises the loss (as a packet percentage) between two directly connected OSPF neighbors. The link loss advertised by this sub-TLV MUST be the packet loss from the local neighbor to the remote one (i.e. the forward path loss). The format of this sub-TLV is shown in the following diagram:

```
  0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|              TBD3             |               4               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                           Link Loss                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

4.3.1. Type

This sub-TLV has a type of TBD3

4.3.2. Length

The length is 4

4.3.3. Link Loss

This field carries the link packet loss as a percentage of the total traffic sent over a configurable interval, encoded as an IEEE floating point single precision value.

4.4. Significant Unidirectional Link Delay Sub-TLV

This TLV advertises the average link delay between two directly connected OSPF neighbors. This TLV is announced when either a configurable maximum average delay or a configurable reuse delay threshold is passed. The delay advertised by this sub TLV MUST be the delay from the local neighbor to the remote one (i.e. the forward path latency). The format of this sub-TLV is shown in the following diagram:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|              TBD4             |               4               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                             Delay                             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

4.4.1. Type

This sub-TLV has a type of TBD4

4.4.2. Length

The length is 4

4.4.3. Delay Value

This field carries the average link delay over a configurable interval in micro-seconds, encoded as an IEEE floating point single precision value.

4.5. Significant Unidirectional Link Loss Sub TLV

This TLV advertises the loss (as a packet percentage) between two directly connected OSPF neighbors. This TLV is announced when either a configurable loss threshold or a configurable loss reuse threshold is passed. The link loss advertised by this sub-TLV MUST be the packet loss from the local neighbor to the remote one (i.e. the forward path loss). The format of this sub-TLV is shown in the following diagram:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|              TBD3             |               4               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                          Link Loss                            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

4.5.1. Type

This sub-TLV has a type of TBD5

4.5.2. Length

The length is 4

4.5.3. Link Loss

This field carries the link packet loss as a percentage of the total traffic sent over a configurable interval, encoded as an IEEE floating point single precision value.

5. Announcement Periodicity

Routine announcements are intended to announce data for trending applications (e.g. advertising small variations in performance occurring over a longer period of time). Significant sub-TLVs are intended to announce the occurrence of more dramatic events that affect network performance (e.g. protection switching). A primary function of Significant sub-TLVs are to manipulate the control plane.

Since Routine and Significant sub-TLVs have generally different goals, implementations SHOULD permit them to be announced using different thresholds and filtering (i.e. rolling average) parameters.

6. Announcement Suppression

Implementations MAY suppress Routine announcements when performance metrics averages do not change by more than a certain amount. These suppression thresholds SHOULD be configurable.

Significant announcements MUST only be sent when configurable thresholds are surpassed.

7. Compatibility

As per (RFC3630), unrecognized TLVs should be silently ignored

8. Security Considerations

This document does not introduce security issues beyond those discussed in [RFC3630] and [RFC5329].

9. IANA Considerations

IANA maintains the registry for the sub-TLVs. OSPF TE Express Path will require one new type code per sub-TLV defined in this document.

10. References

10.1. Normative References

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

10.2. Informative References

[RFC2328] Moy, J, "OSPF Version 2”, RFC 2328, April 1998

[RFC3031] Rosen, E., Viswanathan, A., Callon, R., “Multiprotocol Label Switching Architecture”, January 2001

[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.

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

[RFC5250] Berger, L., Bryskin I., Zinin, A., Coltun, R., "The OSPF Opaque LSA Option", RFC 5250, July 2008.

11. Acknowledgments

The authors would like to recognize Ayman Soliman for his contributions.

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

12. Author's Addresses

Spencer Giacalone
Thomson Reuters
195 Broadway
New York NY 10007, USA

Email: Spencer.giacalone@thomsonreuters.com

Dave Ward
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089, USA

Email: dward@juniper.net

John Drake
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089, USA

Email: jdrake@juniper.net

Alia Atlas
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089, USA

Email: akatlas@juniper.net