Network Working Group
Internet Draft
Intended status: Proposed Standard
Expires: June 2015

S. Giacalone
Unaffiliated

D. Ward
Cisco Systems

J. Drake
Juniper Networks

A. Atlas
Juniper Networks

S. Previdi
Cisco Systems

December 01, 2014

OSPF Traffic Engineering (TE) Metric Extensions
draft-ietf-ospf-te-metric-extensions-08.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) are becoming as critical to data path selection as other metrics.

This document describes extensions to RFC 3630 'Traffic Engineering (TE) Extensions to OSPF Version 2' such that network performance information can be distributed and collected in a scalable fashion. The information distributed using OSPF TE Metric Extensions can then be used to make path selection decisions based on network performance.

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

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

This Internet-Draft will expire on June 1, 2015.

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) is becoming as critical to data path selection as other metrics.

In these networks, extremely large amounts of money rest on the ability to access market data in "real time" and to predictably make trades faster than the competition. Because of this, using metrics such as hop count or cost as routing metrics is becoming only tangentially important. Rather, it would be beneficial to be able to make path selection decisions based on performance data (such as latency) in a cost-effective and scalable way.

This document describes extensions to OSPF TE (hereafter called "OSPF TE Metric Extensions"), that can be used to distribute network performance information (viz link delay, delay variation, link loss, residual bandwidth, available bandwidth, and utilized bandwidth).

The data distributed by OSPF TE Metric Extensions is meant to be used as part of the operation of the routing protocol (e.g., by replacing cost with latency or considering bandwidth as well as cost), by enhancing CSPF, or for use by a PCE [RFC4655] or an Alto server [RFC7285]. With respect to CSPF, the data distributed by OSPF TE Metric Extensions can be used to setup, fail over, and fail back data paths using protocols such as RSVP-TE [RFC3209].

Note that 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. Example mechanisms to measure latency, delay variation, and loss in an MPLS network are given in [RFC6374]. While this document does not specify how the performance information should be obtained, the measurement of delay SHOULD NOT vary significantly based upon the offered traffic load. Thus, queuing delays and/or loss SHOULD NOT be included in any dynamic delay measurement. For links, such as Forwarding Adjacencies, care must be taken that measurement of the associated delay avoids significant queuing delay; that could be accomplished in a variety of ways, including either by measuring with a traffic class that

experiences minimal queuing or by summing the measured link delays of the components of the link's path.

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. TE Metric Extensions to OSPF TE

This document proposes new OSPF TE sub-TLVs that can be announced in OSPF TE LSAs to distribute network performance information. The extensions in this document build on the ones provided in OSPF TE [RFC3630] and GMPLS [RFC4203].

OSPF TE LSAs [RFC3630] are opaque LSAs [RFC5250] with area flooding scope. Each consists of a single TLV with one or more nested sub-TLVs, permitting the TE LSA to be readily extended. There are two OSPF TE LSA TLVs, the Router Address and the Link TLV. Like the extensions in GMPLS (RFC4203), this document proposes several additional sub-TLVs for the Link TLV:

Type	Length	Value
TBD1	4	Unidirectional Link Delay
TBD2	8	Min/Max Unidirectional Link Delay
TBD3	4	Unidirectional Delay Variation
TBD4	4	Unidirectional Link Loss
TBD5	4	Unidirectional Residual Bandwidth
TBD6	4	Unidirectional Available Bandwidth
TBD7	4	Unidirectional Utilized Bandwidth

As can be seen in the list above, the sub-TLVs described in this document carry different types of network performance information. Many (but not all) of the sub-TLVs include a bit called the Anomalous (or "A") bit. When the A bit is clear (or when the sub-TLV does not include an A bit), the sub-TLV describes steady state link performance. This information could conceivably be used to construct a steady state performance topology for initial tunnel path computation, or to verify alternative failover paths.

When network performance violates configurable link-local thresholds a sub-TLV with the A bit set is advertised. These sub-TLVs could be used by the receiving node to determine whether to fail traffic to a backup path, or whether to calculate an entirely new path. From an MPLS perspective, the intent of the A bit is to permit LSP ingress nodes to:

A) Determine whether the link referenced in the sub-TLV affects any of the LSPs for which it is ingress. If there are, then:

B) The node determines whether those LSPs still meet end-to-end performance objectives. If not, then:

C) The node could then conceivably move affected traffic to a pre-established protection LSP or establish a new LSP and place the traffic in it.

If link performance then improves beyond a configurable minimum value (reuse threshold), that sub-TLV can be re-advertised with the Anomalous bit cleared. In this case, a receiving node can conceivably do whatever re-optimization (or failback) it wishes to do (including nothing).

Note that when a sub-TLV does not include the A bit, that sub-TLV cannot be used for failover purposes. The A bit was intentionally omitted from some sub-TLVs to help mitigate oscillations. See section 7. 1. for more information.

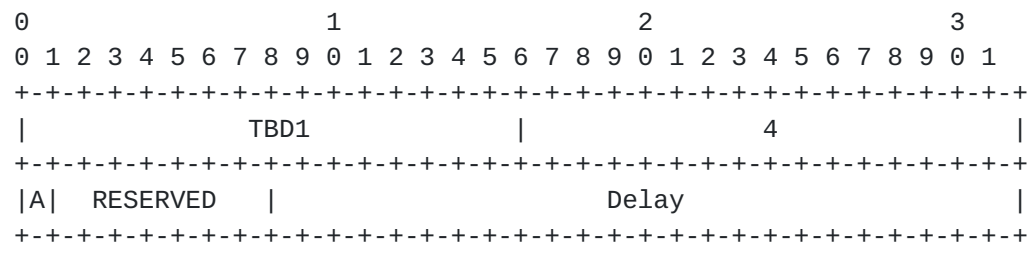
Link delay, delay variation, and link loss MUST be encoded as integers. Consistent with existing OSPF TE specifications [RFC3630], residual, available, and utilized bandwidth MUST be encoded in IEEE floating point. Link delay and delay variation MUST be in units of microseconds, link loss MUST be a percentage, and bandwidth MUST be in units of bytes per second. All values (except residual bandwidth) MUST be calculated as rolling averages where the averaging period

MUST be a configurable period of time. See section 5. for more information.

4. Sub-TLV Details

4.1. Unidirectional Link Delay Sub-TLV

This sub-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 advertising node to its neighbor (i.e., the forward path delay). The format of this sub-TLV is shown in the following diagram:



4.1.1. Type

This sub-TLV has a type of TBD1.

4.1.2. Length

The length is 4.

4.1.3. A bit

This field represents the Anomalous (A) bit. The A bit is set when the measured value of this parameter exceeds its configured maximum threshold. The A bit is cleared when the measured value falls below its configured reuse threshold. If the A bit is clear, the sub-TLV represents steady state link performance.

4.1.4. Reserved

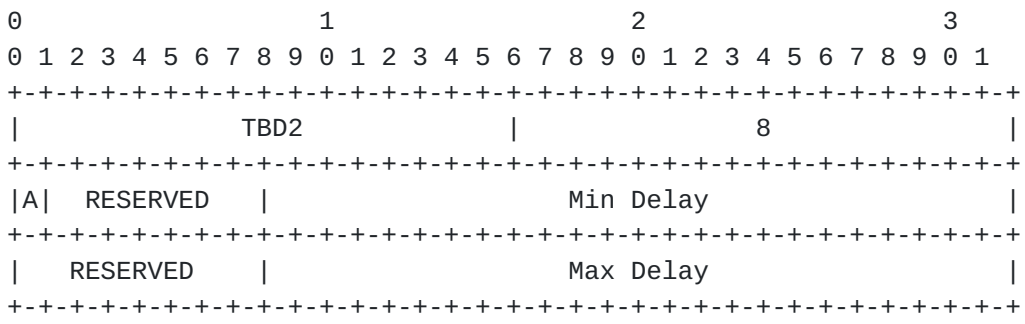
This field is reserved for future use. It MUST be set to 0 when sent and MUST be ignored when received.

4.1.5. Delay Value

This 24-bit field carries the average link delay over a configurable interval in micro-seconds, encoded as an integer value. When set to the maximum value 16,777,215 (16.777215 sec), then the delay is at least that value and may be larger. If there is no value to send (unmeasured and not statically specified), then the sub-TLV should not be sent or be withdrawn.

4.2. Min/Max Unidirectional Link Delay Sub-TLV

This sub-TLV advertises the minimum and maximum delay values between two directly connected OSPF neighbors. The delay advertised by this sub-TLV MUST be the delay from the advertising node to its neighbor (i.e., the forward path delay). The format of this sub-TLV is shown in the following diagram:



4.2.1. Type

This sub-TLV has a type of TBD2.

4.2.2. Length

The length is 8.

4.2.3. A bit

This field represents the Anomalous (A) bit. The A bit is set when one or more measured values exceed a configured maximum threshold. The A bit is cleared when the measured value falls below its configured reuse threshold. If the A bit is clear, the sub-TLV represents steady state link performance.

4.2.4. Reserved

This field is reserved for future use. It MUST be set to 0 when sent and MUST be ignored when received.

4.2.5. Min Delay

This 24-bit field carries minimum measured link delay value (in microseconds) over a configurable interval, encoded as an integer value.

Implementations MAY also permit the configuration of a static (non dynamic) offset value (in microseconds) to be added to the measured delay value, to facilitate the communication of operator specific delay constraints.

When set to the maximum value 16,777,215 (16.777215 sec), then the delay is at least that value and may be larger.

4.2.6. Max Delay

This 24-bit field carries the maximum measured link delay value (in microseconds) over a configurable interval, encoded as an integer value.

Implementations MAY also permit the configuration of a static (non dynamic) offset value (in microseconds) to be added to the measured delay value, to facilitate the communication of operator specific delay constraints.

It is possible for min delay and max delay to be the same value.

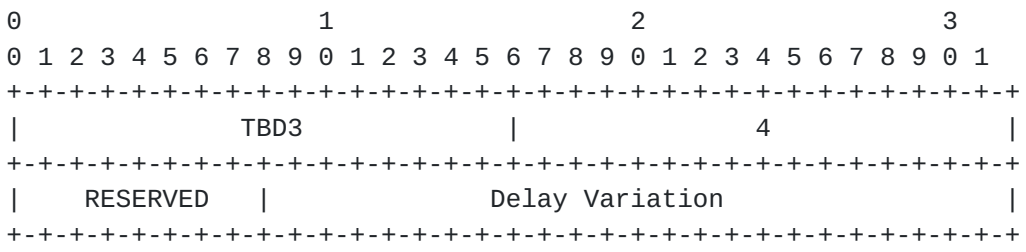
When the delay value is set to maximum value 16,777,215 (16.777215 sec), then the delay is at least that value and may be larger.

4.2.7. Reserved

This field is reserved for future use. It MUST be set to 0 when sent and MUST be ignored when received.

4.3. Unidirectional Delay Variation Sub-TLV

This sub-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 advertising node to its neighbor (i.e., the forward path delay variation). The format of this sub-TLV is shown in the following diagram:



4.3.1. Type

This sub-TLV has a type of TBD3.

4.3.2. Length

The length is 4.

4.3.3. Reserved

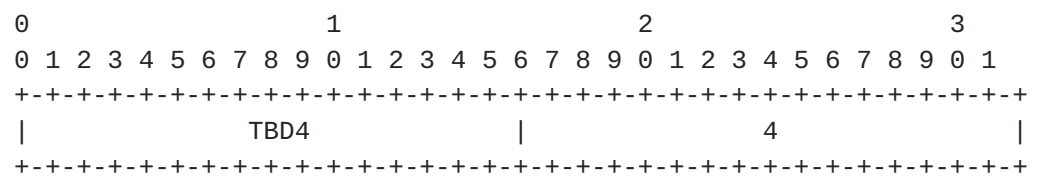
This field is reserved for future use. It MUST be set to 0 when sent and MUST be ignored when received.

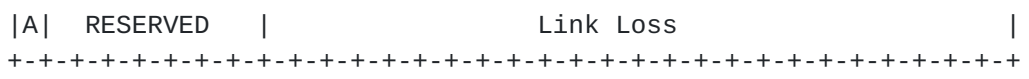
4.3.4. Delay Variation

This 24-bit field carries the average link delay variation over a configurable interval in micro-seconds, encoded as an integer value. When set to 0, it has not been measured. When set to the maximum value 16,777,215 (16.777215 sec), then the delay is at least that value and may be larger.

4.4. Unidirectional Link Loss Sub-TLV

This sub-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 advertising node to its neighbor (i.e., the forward path loss). The format of this sub-TLV is shown in the following diagram:





4.4.1. Type

This sub-TLV has a type of TBD4

4.4.2. Length

The length is 4.

4.4.3. A bit

This field represents the Anomalous (A) bit. The A bit is set when the measured value of this parameter exceeds its configured maximum threshold. The A bit is cleared when the measured value falls below its configured reuse threshold. If the A bit is clear, the sub-TLV represents steady state link performance.

4.4.4. Reserved

This field is reserved for future use. It MUST be set to 0 when sent and MUST be ignored when received.

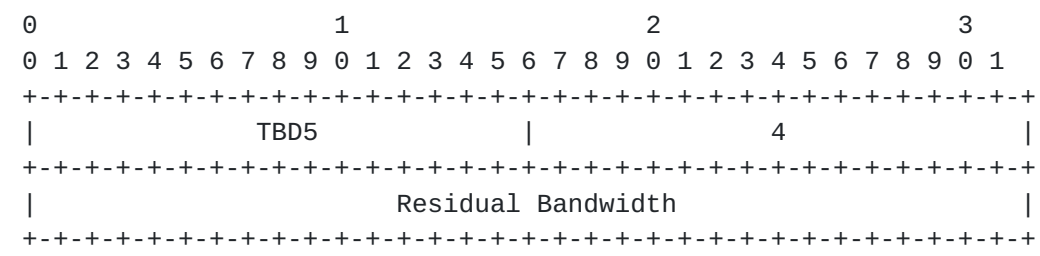
4.4.5. Link Loss

This 24-bit field carries link packet loss as a percentage of the total traffic sent over a configurable interval. The basic unit is 0.000003%, where (2^24 - 2) is 50.331642%. This value is the highest packet loss percentage that can be expressed (the assumption being that precision is more important on high speed links than the ability to advertise loss rates greater than this, and that high speed links with over 50% loss are unusable). Therefore, measured values that are larger than the field maximum SHOULD be encoded as the maximum value. When set to a value of all 1s (2^24 - 1), the link packet loss has not been measured.

4.5. Unidirectional Residual Bandwidth Sub-TLV

This sub-TLV advertises the residual bandwidth between two directly connected OSPF neighbors. The residual bandwidth advertised by this sub-TLV MUST be the residual bandwidth from the advertising node to its neighbor.

The format of this sub-TLV is shown in the following diagram:



4.5.1. Type

This sub-TLV has a type of TBD5.

4.5.2. Length

The length is 4.

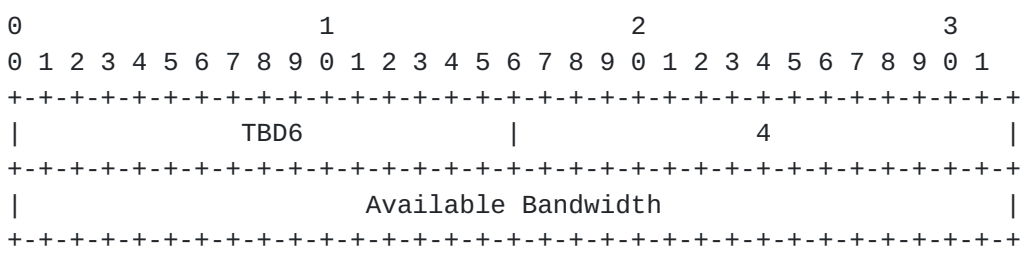
4.5.3. Residual Bandwidth

This field carries the residual bandwidth on a link, forwarding adjacency [RFC4206], or bundled link in IEEE floating point format with units of bytes per second. For a link or forwarding adjacency, residual bandwidth is defined to be Maximum Bandwidth [RFC3630] minus the bandwidth currently allocated to RSVP-TE LSPs. For a bundled link, residual bandwidth is defined to be the sum of the component link residual bandwidths.

The calculation of Residual Bandwidth is different than that of Unreserved Bandwidth [RFC3630]. Residual Bandwidth subtracts tunnel reservations from Maximum Bandwidth (i.e., the link capacity) [RFC3630] and provides an aggregated remainder across QoS classes. Unreserved Bandwidth [RFC3630], on the other hand, is subtracted from the Maximum Reservable Bandwidth (the bandwidth that can theoretically be reserved) [RFC3630] and provides per-QoS-class remainders. Residual Bandwidth and Unreserved Bandwidth [RFC3630] can be used concurrently, and each has a separate use case (e.g., the former can be used for applications like Weighted ECMP while the latter can be used for call admission control).

4.6. Unidirectional Available Bandwidth Sub-TLV

This TLV advertises the available bandwidth between two directly connected OSPF neighbors. The available bandwidth advertised by this sub-TLV MUST be the available bandwidth from the advertising node to its neighbor. The format of this sub-TLV is shown in the following diagram:



4.6.1. Type

This sub-TLV has a type of TBD6.

4.6.2. Length

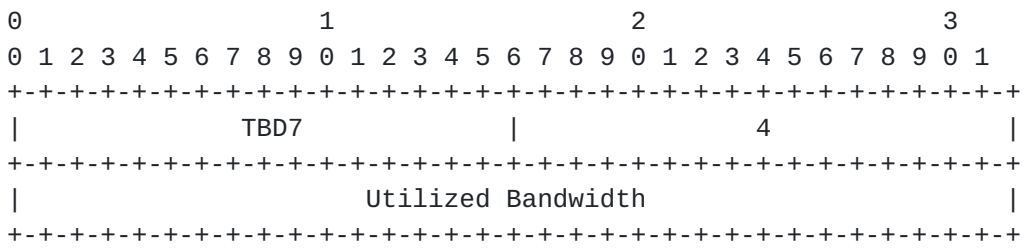
The length is 4.

4.6.3. Available Bandwidth

This field carries the available bandwidth on a link, forwarding adjacency, or bundled link in IEEE floating point format with units of bytes per second. For a link or forwarding adjacency, available bandwidth is defined to be residual bandwidth (see section 4.5.) minus the measured bandwidth used for the actual forwarding of non-RSVP-TE LSP packets. For a bundled link, available bandwidth is defined to be the sum of the component link available bandwidths.

4.7. Unidirectional Utilized Bandwidth Sub-TLV

This Sub-TLV advertises the bandwidth utilization between two directly connected OSPF neighbors. The bandwidth utilization advertised by this sub-TLV MUST be the bandwidth from the advertising node to its neighbor. The format of this Sub-TLV is shown in the following diagram:



4.7.1. Type

This sub-TLV has a type of TBD7.

4.7.2. Length

The length is 4.

4.7.3. Utilized Bandwidth

This field carries the bandwidth utilization on a link, forwarding adjacency, or bundled link in IEEE floating point format with units of bytes per second. For a link or forwarding adjacency, bandwidth utilization represents the actual utilization of the link (i.e., as measured by the advertising node). For a bundled link, bandwidth utilization is defined to be the sum of the component link bandwidth utilizations.

5. Announcement Thresholds and Filters

The values advertised in all sub-TLVs (except min/max delay and residual bandwidth) MUST represent an average over a period or be obtained by a filter that is reasonably representative of an average. For example, a rolling average is one such filter.

Min and max delay MAY be the lowest and/or highest measured value over a measurement interval or MAY make use of a filter, or other technique to obtain a reasonable representation of a min and max value representative of the interval with compensation for outliers.

The measurement interval, any filter coefficients, and any advertisement intervals MUST be configurable per sub-TLV.

In addition to the measurement intervals governing re-advertisement, implementations SHOULD provide per sub-TLV configurable accelerated advertisement thresholds, such that:

1. If the measured parameter falls outside a configured upper bound for all but the min delay metric (or lower bound for min delay metric only) and the advertised sub-TLV is not already outside that bound or,

2. If the difference between the last advertised value and current measured value exceed a configured threshold then,

3. The advertisement is made immediately.

4. For sub-TLVs which include an A-bit (except min/max delay), an additional threshold SHOULD be included corresponding to the threshold for which the performance is considered anomalous (and sub-TLVs with the A bit are sent). The A-bit is cleared when the sub-TLV's performance has been below (or re-crosses) this threshold for an advertisement interval(s) to permit fail back.

To prevent oscillations, only the high threshold or the low threshold (but not both) may be used to trigger any given sub-TLV that supports both.

Additionally, once outside of the bounds of the threshold, any re-advertisement of a measurement within the bounds would remain governed solely by the measurement interval for that sub-TLV.

6. Announcement Suppression

When link performance values change by small amounts that fall under thresholds that would cause the announcement of a sub-TLV, implementations SHOULD suppress sub-TLV readvertisement and/or lengthen the period within which they are refreshed.

Only the accelerated advertisement threshold mechanism described in section 6 may shorten the re-advertisement interval.

All suppression and re-advertisement interval back-off timer features SHOULD be configurable.

7. Network Stability and Announcement Periodicity

Sections 6 and 7 provide configurable mechanisms to bound the number of re-advertisements. Instability might occur in very large networks if measurement intervals are set low enough to overwhelm the

processing of flooded information at some of the routers in the topology. Therefore care SHOULD be taken in setting these values.

Additionally, the default measurement interval for all sub-TLVs SHOULD be 30 seconds.

Announcements MUST also be able to be throttled using configurable inter-update throttle timers. The minimum announcement periodicity is 1 announcement per second. The default value SHOULD be set to 120 seconds.

Implementations SHOULD NOT permit the inter-update timer to be lower than the measurement interval.

Furthermore, it is RECOMMENDED that any underlying performance measurement mechanisms not include any significant buffer delay, any significant buffer induced delay variation, or any significant loss due to buffer overflow or due to active queue management.

8. Enabling and Disabling Sub-TLVs

Implementations MUST make it possible to individually enable or disable each sub-TLV based on configuration.

9. Static Metric Override

Implementations SHOULD permit the static configuration and/or manual override of dynamic measurements data on a per sub-TLV, per metric basis in order to simplify migrations and to mitigate scenarios where measurements are not possible across an entire network.

10. Compatibility

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

11. Security Considerations

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

12. IANA Considerations

IANA maintains the registry for the Link TLV sub-TLVs. OSPF TE Metric Extensions will require one new type code per sub-TLV defined in this document, as follows:

Type	Description
TBD1	Unidirectional Link Delay
TBD2	Min/Max Unidirectional Link Delay
TBD3	Unidirectional Delay Variation
TBD4	Unidirectional Link Loss
TBD5	Unidirectional Residual Bandwidth
TBD6	Unidirectional Available Bandwidth
TBD7	Unidirectional Utilized Bandwidth

13. References

13.1. Normative References

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

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

13.2. Informative References

[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.

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

[RFC4206] Kompella, K., Rekhter, Y., "Label Switched Paths (LSP) Hierarchy with Generalized Multi-Protocol Label Switching (GMPLS) Traffic Engineering (TE)", RFC 4206, October 2005.

[RFC4655] Farrel, A., Vasseur, J.-P., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, August 2006.

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

[RFC6374] Frost, D. and S. Bryant, "Packet Loss and Delay Measurement for MPLS Networks", RFC 6374, September 2011.

[RFC7285] Almi, R., Penno, R., Yang, Y., Kiesel, S., Previdi, S., Roome, W., Shalunov, S., and R. Woundy, "Application-Layer Traffic Optimization (ALTO) Protocol", RFC 7285, September 2014.

14. Acknowledgments

The authors would like to recognize Ayman Soliman, Nabil Bitar, David McDysan, Edward Crabbe, and Don Fedyk for their contributions.

The authors also recognize Curtis Villamizar for significant comments and direct content collaboration.

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

15. Author's Addresses

Spencer Giacalone
Unaffiliated

Email: spencer.giacalone@gmail.com

Dave Ward
Cisco Systems
170 West Tasman Dr.
San Jose, CA 95134, USA

Email: dward@cisco.com

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

Stefano Previdi
Cisco Systems
Via Del Serafico 200
00142 Rome
Italy

Email: sprevidi@cisco.com