Networking Working Group
Internet-Draft
Intended status: Standards Track
Expires: December 18, 2015

S. Previdi, Ed.
Cisco Systems, Inc.
S. Giacalone
Unaffiliated
D. Ward
Cisco Systems, Inc.
J. Drake
A. Atlas
Juniper Networks
C. Filsfils
Cisco Systems, Inc.
Q. Wu
Huawei

June 16, 2015

IS-IS Traffic Engineering (TE) Metric Extensions
draft-ietf-isis-te-metric-extensions-07

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 IS-IS Traffic Engineering Extensions (RFC5305) such that network performance information can be distributed and collected in a scalable fashion. The information distributed using ISIS 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.

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 [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.

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) 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 IS-IS Extended Reachability TLV defined in [RFC5305] (hereafter called "IS-IS TE Metric Extensions"), that can be used to distribute network performance information (such as link delay, delay variation, link loss, residual bandwidth, and available bandwidth).

The data distributed by the TE Metric Extensions proposed in this document 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 Constrained-SPF (CSPF), or for other uses such as supplementing the data used by an ALTO server [RFC7285]. With respect to CSPF, the data distributed by ISIS 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, such as [RFC6375], 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 SHOULD NOT be included in the 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. TE Metric Extensions to IS-IS

This document proposes new IS-IS TE sub-TLVs that can be announced in TLVs 22, 23, 141, 222, and 223 in order to distribute network performance information. The extensions in this document build on the ones provided in IS-IS TE [RFC5305] and GMPLS [RFC4203].

IS-IS Extended Reachability TLV 22 (defined in [RFC5305]), Inter-AS reachability information TLV 141 (defined in [RFC5316]) and MT-ISIS TLV 222 (defined in [RFC5120]) have nested sub-TLVs which permit the TLVs to be readily extended. This document proposes several additional sub-TLVs:

Type	Value
33 (Suggested)	Unidirectional Link Delay
34 (Suggested)	Min/Max Unidirectional Link Delay
35 (Suggested)	Unidirectional Delay Variation
36 (Suggested)	Unidirectional Link Loss
37 (Suggested)	Unidirectional Residual Bandwidth
38 (Suggested)	Unidirectional Available Bandwidth
39 (Suggested)	Unidirectional Bandwidth Utilization

As can be seen in the list above, the sub-TLVs described in this document carry different types of network performance information. The new 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) Determine 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 5 for more information.

Consistent with existing IS-IS TE specification [RFC5305], the bandwidth advertisements defined in this draft MUST be encoded as IEEE floating point values. The delay and delay variation advertisements defined in this draft MUST be encoded as integer values. Delay values MUST be quantified in units of microseconds, link loss MUST be quantified as a percentage of packets sent, and bandwidth MUST be sent as 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.

3. Interface and Neighbor Addresses

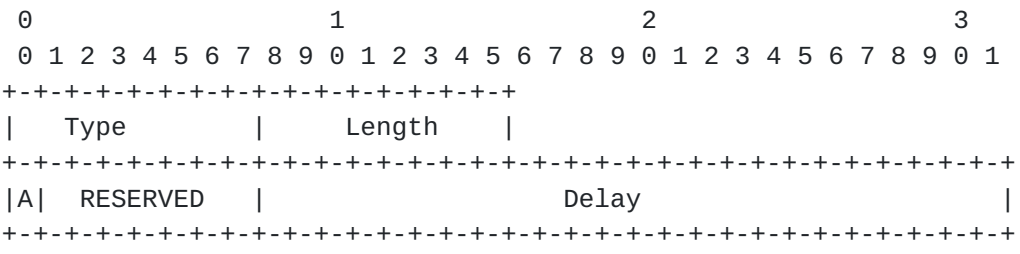
The use of TE Metric Extensions SubTLVs is not confined to the TE context. In other words, IS-IS TE Metric Extensions SubTLVs defined in this document can also be used for computing paths in the absence of a TE subsystem.

However, as for the TE case, Interface Address and Neighbor Address SubTLVs (IPv4 or IPv6) MUST be present. The encoding is defined in [RFC5305] for IPv4 and in [RFC6119] for IPv6.

4. Sub TLV Details

4.1. Unidirectional Link Delay Sub-TLV

This sub-TLV advertises the average link delay between two directly connected IS-IS 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:



where:

Figure 1

Type: TBA (suggested value: 33).

Length: 4.

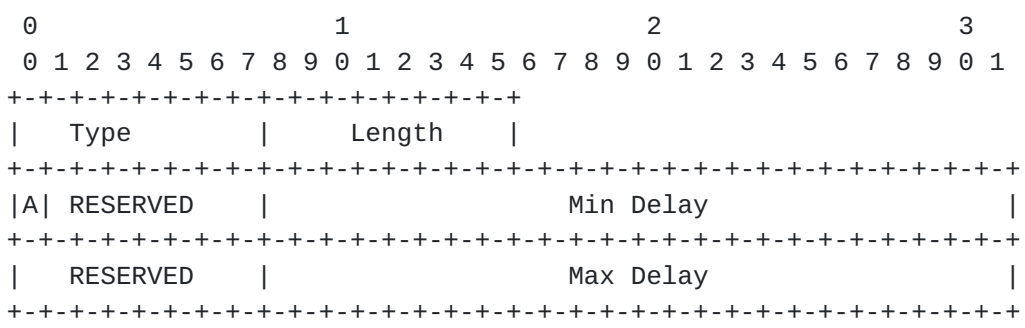
A-bit. The A-bit represents the Anomalous (A) bit. The A-bit is set when the measured value of this parameter exceeds its configured maximum threshold. The A bit is cleared when the measured value falls below its configured reuse threshold. If the A-bit is clear, the sub-TLV represents steady state link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

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

Delay. 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.

4.2. Min/Max Unidirectional Link Delay Sub-TLV

This sub-TLV advertises the minimum and maximum delay values between two directly connected IS-IS 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:



where:

Figure 2

Type: TBA (suggested value: 34).

Length: 8.

A-bit. The A-bit 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.

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

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

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

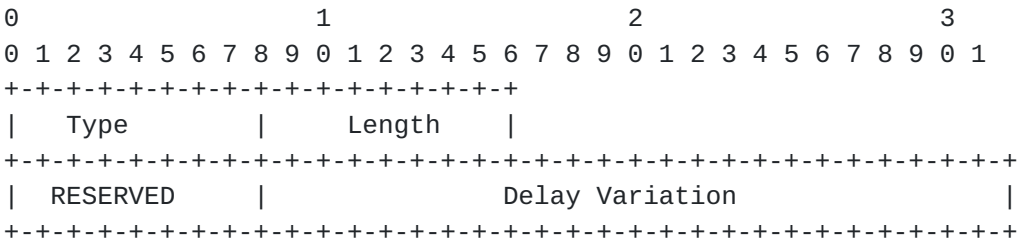
Implementations MAY also permit the configuration of an offset value (in microseconds) to be added to the measured delay value, to facilitate the communication of operator specific delay constraints.

It is possible for the Min and Max delay to be the same value.

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

4.3. Unidirectional Delay Variation Sub-TLV

This sub-TLV advertises the average link delay variation between two directly connected IS-IS 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:



where:

Figure 3

Type: TBA (suggested value: 35).

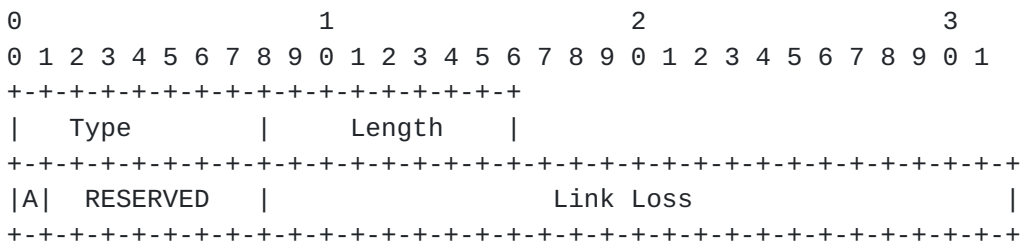
Length: 4.

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

Delay Variation. This 24-bit field carries the average link delay variation over a configurable interval in microseconds, 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 IS-IS 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:



This sub-TLV has a type of TBD3.
The length is 4.

where:

Type: TBA (suggested value: 36).

Length: 4.

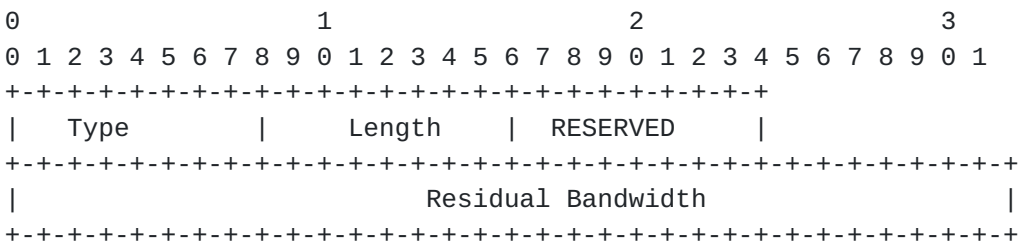
A-bit. The A-bit represents the Anomalous (A) bit. The A-bit is set when the measured value of this parameter exceeds its configured maximum threshold. The A bit is cleared when the measured value falls below its configured reuse threshold. If the A-bit is clear, the sub-TLV represents steady state link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

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

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.

4.5. Unidirectional Residual Bandwidth Sub-TLV

This TLV advertises the residual bandwidth between two directly connected IS-IS neighbors. The residual bandwidth advertised by this sub-TLV MUST be the residual bandwidth from the system originating the LSA to its neighbor.



where:

Type: TBA (suggested value: 37).

Length: 4.

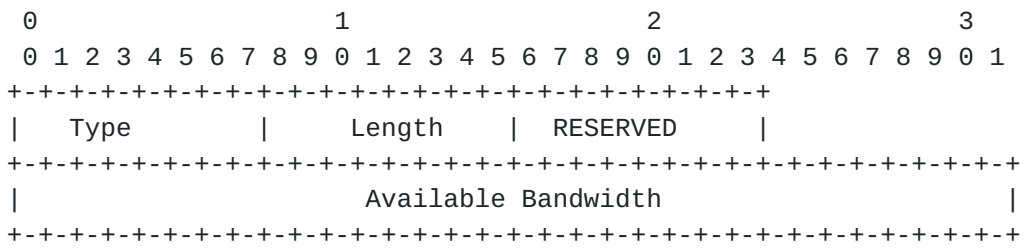
RESERVED. This field is reserved for future use. It MUST be set to 0 when sent and MUST be ignored when received.

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 [RFC5305] 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 [RFC5305]. Residual Bandwidth subtracts tunnel reservations from Maximum Bandwidth (i.e. the link capacity) [RFC5305] and provides an aggregated remainder across priorities. Unreserved Bandwidth, on the other hand, is subtracted from the Maximum Reservable Bandwidth (the bandwidth that can theoretically be reserved) and provides per priority remainders. Residual Bandwidth and Unreserved Bandwidth [RFC5305] 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 Sub-TLV advertises the available bandwidth between two directly connected IS-IS neighbors. The available bandwidth advertised by this sub-TLV MUST be the available bandwidth from the system originating this Sub-TLV. The format of this Sub-TLV is shown in the following diagram:



where:

Figure 4

Type: TBA (suggested value: 38).

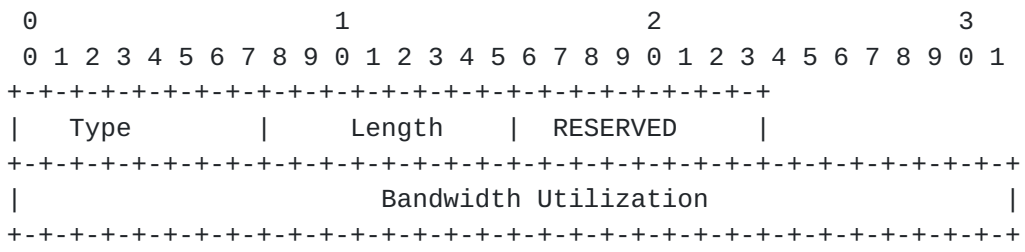
Length: 4.

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

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 minus the measured bandwidth used for the actual forwarding of non-RSVP-TE Label Switched Paths 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 IS-IS neighbors. The bandwidth utilization advertised by this sub-TLV MUST be the bandwidth from the system originating this Sub-TLV. The format of this Sub-TLV is shown in the following diagram:



where:

Figure 5

Type: TBA (suggested value: 39).

Length: 4.

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

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 readvertisement 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 5 may shorten the re-advertisement interval. All suppression and re-advertisement interval backoff timer features SHOULD be configurable.

7. Network Stability and Announcement Periodicity

Section 5 and Section 6 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 for each sub-TLV in order to simplify migration and to mitigate scenarios where dynamic measurements are not possible.

10. Compatibility

As per [RFC5305], unrecognized Sub-TLVs should be silently ignored.

11. Security Considerations

This document does not introduce security issues beyond those discussed in [RFC5305] and [RFC5329].

12. IANA Considerations

IANA maintains the registry for the sub-TLVs. IS-IS TE Metric Extensions will require one new type code per sub-TLV defined in this document in the following sub-TLV registry: TLVs 22, 23, 141, 222, and 223:

Type	Value
33	(Suggested) Unidirectional Link Delay
34	(Suggested) Min/Max Unidirectional Link Delay
35	(Suggested) Unidirectional Delay Variation
36	(Suggested) Unidirectional Link Loss
37	(Suggested) Unidirectional Residual Bandwidth
38	(Suggested) Unidirectional Available Bandwidth
39	(Suggested) Unidirectional Bandwidth Utilization

13. Acknowledgements

The authors would like to recognize Ayman Soliman, Nabil Bitar, David McDysan, Les Ginsberg, Edward Crabbe, Don Fedyk, Hannes Gredler and Uma Chunduri for their contributions.

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

14. References

14.1. 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.

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

[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.

[RFC5120] Przygienda, T., Shen, N., and N. Sheth, "M-ISIS: Multi Topology (MT) Routing in Intermediate System to Intermediate Systems (IS-ISs)", RFC 5120, February 2008.

[RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, October 2008.

[RFC5316] Chen, M., Zhang, R., and X. Duan, "ISIS Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS Traffic Engineering", RFC 5316, December 2008.

[RFC5329] Ishiguro, K., Manral, V., Davey, A., and A. Lindem, "Traffic Engineering Extensions to OSPF Version 3", RFC 5329, September 2008.

[RFC6119] Harrison, J., Berger, J., and M. Bartlett, "IPv6 Traffic Engineering in IS-IS", RFC 6119, February 2011.

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

14.2. Informative References

[RFC6375] Frost, D. and S. Bryant, "A Packet Loss and Delay Measurement Profile for MPLS-Based Transport Networks", RFC 6375, September 2011.

[RFC7285] Alimi, 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.

Authors' Addresses

Stefano Previdi (editor)
Cisco Systems, Inc.
Via Del Serafico 200
Rome 00191
IT

Email: sprevidi@cisco.com

Spencer Giacalone
Unaffiliated

Email: spencer.giacalone@gmail.com

Dave Ward
Cisco Systems, Inc.
3700 Cisco Way
SAN JOSE, CA 95134
US

Email: wardd@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

Clarence Filsfils
Cisco Systems, Inc.
Brussels
Belgium

Email: cfilsfil@cisco.com

Qin Wu
Huawei
101 Software Avenue, Yuhua District
Nanjing, Jiangsu 210012
China

Email: sunseawq@huawei.com