             IS-IS Traffic Engineering (TE) Metric Extensions
                draft-previdi-isis-te-metric-extensions-03

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 TE [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

   This Internet-Draft is submitted in full conformance with the

   provisions of BCP 78 and BCP 79.

   Internet-Drafts are working documents of the Internet Engineering
   Task Force (IETF). Note that other groups may also distribute
   working documents as Internet-Drafts. The list of current Internet-
   Drafts is at http://datatracker.ietf.org/drafts/current/.

   Internet-Drafts are draft documents valid for a maximum of six months
   and may be updated, replaced, or obsoleted by other documents at any
   time. It is inappropriate to use Internet-Drafts as reference
   material or to cite them other than as "work in progress."

   This Internet-Draft will expire on August 29, 2013.

Copyright Notice

Table of Contents

1. **Introduction**

   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, packet loss, residual bandwidth, and
   available bandwidth).

   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.

   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
   [I-D.ietf-alto-protocol]. 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];
   [I-D.atlas-mpls-te-express-path] describes some methods for using
   this information to compute Label Switched Paths (LSPs) at the LSP
   ingress.

   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
   ISIS Extended Reachability TLV (TLV-22) 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-ISN
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

     TBD1   Unidirectional Link Delay

     TBD2   Unidirectional Delay Variation

     TBD3   Unidirectional Packet Loss

     TBD4   Unidirectional Residual Bandwidth Sub TLV

     TBD5   Unidirectional Available Bandwidth Sub TLV

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 specifications [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,
packet 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:

```
 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     |A|   RESERVED  |    Delay      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|      Delay            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

   This sub-TLV has a type of TBD1.
   The length is 4.

Where:

"A" 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.

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

"Delay Value" is a 24-bit field carries the average link delay 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.

4.2. 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:

```
 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     |A|   RESERVED    |Delay Variation|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|      Delay Variation         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

   This sub-TLV has a type of TBD2.
   The length is 4.

Where:

"A" 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.

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

"Delay Variation" is a 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.

4.3. 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 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   Type        |    Length     |A|  RESERVED   |   Link Loss   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|        Link Loss              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```


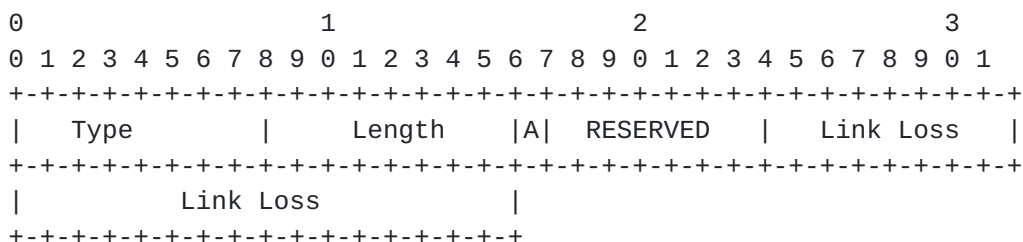   This sub-TLV has a type of TBD3.
    The length is 4.

Where:

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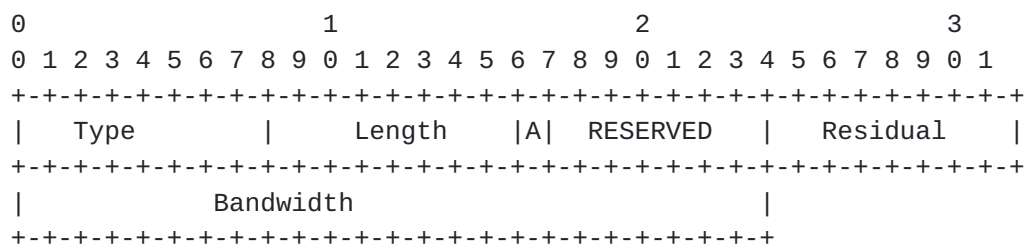when sent and MUST be ignored when received.

"Link Loss" is a 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.

4.4. 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 sub-TLV to its neighbor. 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   Type        |    Length     |A| RESERVED    |   Residual    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|            Bandwidth                          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

   This sub-TLV has a type of TBD4.
   The length is 5.

Where:

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.

"Residual Bandwidth" is the residual bandwidth in IEEE floating point
format in units of bytes per second. The link may be a single link,
forwarding adjacency [RFC4206], or bundled link. For a link or
forwarding adjacency, residual bandwidth is defined to be Maximum
Link 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.

Note that although it may seem possible to calculate Residual
Bandwidth using the existing sub-TLVs in [RFC5305], this is not a
consistently reliable approach and hence the Residual Bandwidth sub-
TLV has been added here. For example, because the Maximum Reservable

Bandwidth [RFC5305] can be larger than the capacity of the link,
using it as part of an algorithm to determine the value of the
Maximum Link Bandwidth [RFC5305]minus the bandwidth currently
allocated to RSVP-TE Label Switched Paths cannot be considered
reliably accurate.

4.5. Unidirectional Available Bandwidth Sub-TLV

This TLV advertises the available bandwidth between two directly
connected IS-IS neighbors. The available bandwidth advertised in
this sub-TLV MUST be the available bandwidth from the originating
system to its neighbor. 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   Type        |    Length     |A|  RESERVED   |   Available   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|             Bandwidth                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

   This sub-TLV has a type of TBD5.
   The length is 5.

   Where:

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.

"Available Bandwidth" is a field that 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 4.4) 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.

5. Announcement Thresholds and Filters

The values advertised in all sub-TLVs MUST be controlled using an
exponential filter (i.e. a rolling average) with a configurable

   measurement interval and filter coefficient.

   Implementations are expected to provide separately configurable
   advertisement thresholds. All thresholds MUST be configurable on a
   per sub-TLV basis.

   The announcement of all sub-TLVs that do not include the A bit SHOULD
   be controlled by variation thresholds that govern when they are sent.

   Sub-TLVs that include the A bit are governed by several thresholds.
   Firstly, a threshold SHOULD be implemented to govern the announcement
   of sub-TLVs that advertise a change in performance, but not an SLA
   violation (i.e. when the A bit is not set). Secondly,
   implementations MUST provide configurable thresholds that govern the
   announcement of sub-TLVs with the A bit set (for the indication of a
   performance violation). Thirdly, implementations SHOULD provide
   reuse thresholds. These thresholds govern sub-TLV re-announcement
   with the A bit cleared to permit fail back.

6. Announcement Suppression

   When link performance average values change, but fall under the
   threshold that would cause the announcement of a sub-TLV with the A
   bit set, implementations MAY suppress or throttle sub-TLV
   announcements. All suppression features and thresholds SHOULD be
   configurable.

7. Network Stability and Announcement Periodicity

   To mitigate concerns about stability, all values (except residual
   bandwidth) MUST NOT be instantaneous measurements. The period to
   compute statistics, whether rolling average, rolling 98th percentile,
   or network custom, MUST be a configurable period of time.
   Announcements MUST also be able to be throttled using configurable
   inter-update throttle timers. The minimum announcement periodicity
   is 1 announcement per second.

8. Compatibility

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

9. Security Considerations

   This document does not introduce security issues beyond those

discussed in [RFC3630] and [RFC5329].

10. 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.

11. Acknowledgements

The authors would like to recognize Ayman Soliman and Les Ginsberg
for their contributions.

12. References

12.1. 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.

[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.

[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.

12.2. Informative References

   [I-D.atlas-mpls-te-express-path]
              Atlas, A., Drake, J., Giacalone, S., Ward, D., Previdi,
              S., and C. Filsfils, "Performance-based Path Selection for
              Explicitly Routed LSPs",
              draft-atlas-mpls-te-express-path-02 (work in progress),
              February 2013.

   [I-D.ietf-alto-protocol]
              Alimi, R., Penno, R., and Y. Yang, "ALTO Protocol",
              draft-ietf-alto-protocol-13 (work in progress),
              September 2012.

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

Authors' Addresses

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

   Email: sprevidi@cisco.com

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

Email: Spencer.giacalone@thomsonreuters.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