Network Working Group   S. Giacalone
Internet Draft   Thomson Reuters
Intended status: Proposed Standard
Expires: August 2013   D. Ward
                                                        Cisco Systems

                                                             J. Drake
                                                      Juniper Networks

                                                             A. Atlas
                                                      Juniper Networks

                                                           S. Previdi
                                                        Cisco Systems

                                                   February 25, 2013

                 OSPF Traffic Engineering (TE) Metric Extensions
                   draft-ietf-ospf-te-metric-extensions-03.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 OSPF TE [RFC3630] 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.

Internet-Drafts are working documents of the Internet Engineering
Task Force (IETF), its areas, and its working groups. Note that
other groups may also distribute working documents as Internet-
Drafts.

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."

The list of current Internet-Drafts can be accessed at
http://www.ietf.org/ietf/1id-abstracts.txt

The list of Internet-Draft Shadow Directories can be accessed at
http://www.ietf.org/shadow.html

This Internet-Draft will expire on August 25, 2013.

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 (such as link delay, delay variation, packet loss, residual bandwidth, and available 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 other uses such as supplementing the data used by an Alto server [Alto]. 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, 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. 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 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 extensions in GMPLS
   (RFC4203), this document proposes several additional sub-TLVs for
   the Link TE LSA:

   Type   Length   Value

   TBD1   4   Unidirectional Link Delay

   TBD2   4   Unidirectional Delay Variation

   TBD3   4   Unidirectional Packet Loss

   TBD4   4   Unidirectional Residual Bandwidth

   TBD5   4   Unidirectional Available 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
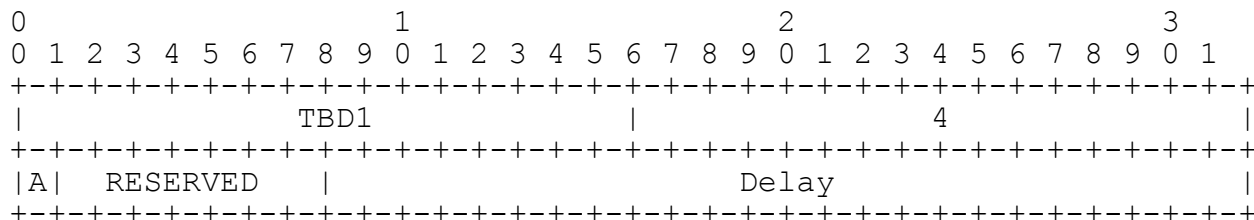 do (including nothing).

  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.

  Consistent with existing OSPF TE specifications (RFC3630), 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.

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

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.

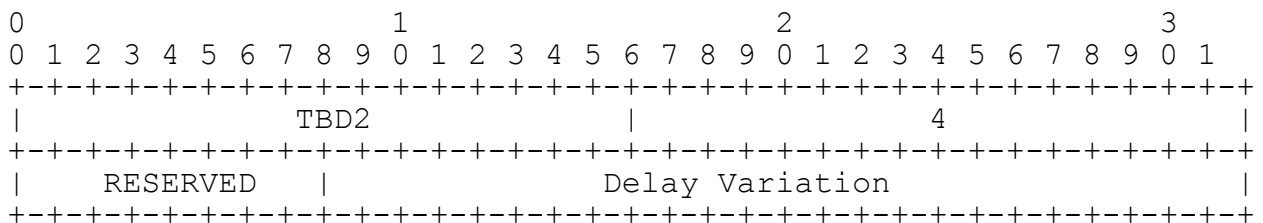### 4.1.4. Reserved

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

4.2.1. Type

   This sub-TLV has a type of TBD2.

4.2.2. Length

   The length is 4.
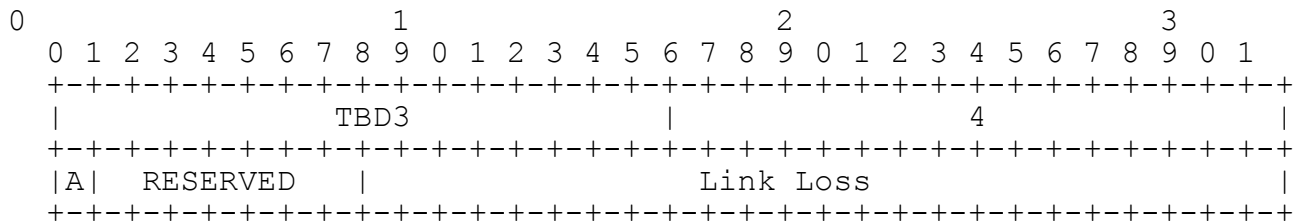
4.2.3. Reserved

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

4.2.4. 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.

4.3. 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 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               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|A|  RESERVED   |               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. 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.

4.3.4. Reserved

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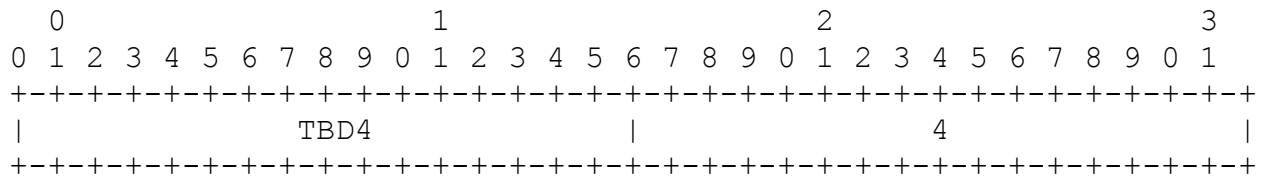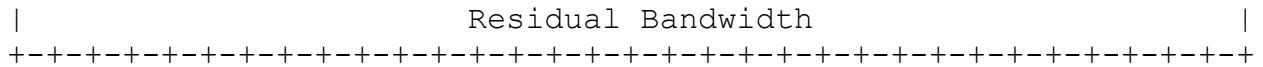
4.3.5. Link Loss

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.

4.4. Unidirectional Residual Bandwidth Sub-TLV

This TLV advertises the residual bandwidth (defined in section 4.4.3. between two directly connected OSPF neighbors. The residual bandwidth advertised by this sub-TLV MUST be the residual bandwidth from the system originating the LSA 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
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |              TBD4              |               4               |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

```
   |                      Residual Bandwidth                       |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

4.4.1. Type

   This sub-TLV has a type of TBD4.

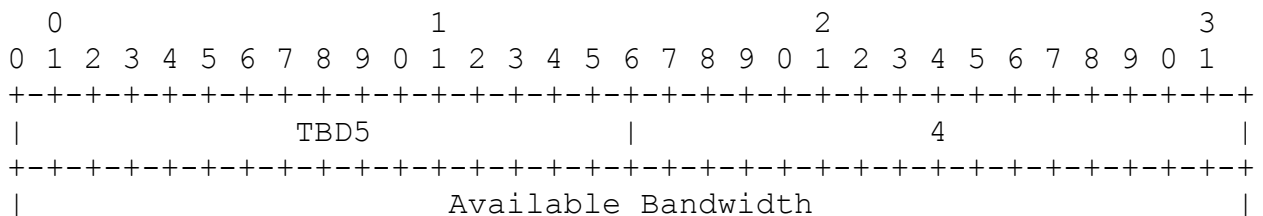## 4.4.2. Length

   The length is 4.

4.4.3. 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.

   Note that although it may seem possible to calculate Residual
   Bandwidth using the existing sub-TLVs in RFC 3630, 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 [RFC3630] can be larger than the capacity of the link,
   using it as part of an algorithm to determine the value of the
   Maximum Bandwidth [RFC3630] minus the bandwidth currently allocated
   to RSVP-TE LSPs cannot be considered reliably accurate.

4.5. Unidirectional Available Bandwidth Sub-TLV

   This TLV advertises the available bandwidth (defined in section
   4.4.6.) between two directly connected OSPF neighbors. The available
   bandwidth advertised by this sub-TLV MUST be the available bandwidth
   from the system originating the LSA 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
      +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
      |             TBD5              |               4               |
      +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
      |                      Available Bandwidth                      |
```

```
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

4.4.4. Type

This sub-TLV has a type of TBD5.

4.4.5. Length

The length is 4.

4.4.6. 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 4.4.)
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.

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-TLV 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. This threshold governs 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 be calculated as rolling averages where the averaging
   period MUST be a configurable period of time, rather than
   instantaneous measurements.

   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 (RFC3630), unrecognized 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. OSPF TE Metric
   Extensions will require one new type code per sub-TLV defined in this
   document.

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.

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

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

11.2. 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.

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

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

    [Alto]   R. Alimi R. Penno Y. Yang, "ALTO Protocol"

12. Acknowledgments

13. Author's Addresses

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

Email: Spencer.giacalone@thomsonreuters.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