Segment Routing Centralized Egress Peer Engineering
draft-filsfils-spring-segment-routing-central-epe-00

Abstract

   Segment Routing (SR) leverages source routing. A node steers a
   packet through a controlled set of instructions, called segments, by
   prepending the packet with an SR header. A segment can represent any
   instruction topological or service-based. SR allows to enforce a
   flow through any topological path and service chain while maintaining
   per-flow state only at the ingress node of the SR domain.

   The Segment Routing architecture can be directly applied to the MPLS
   dataplane with no change on the forwarding plane. It requires minor
   extension to the existing link-state routing protocols.

   This document illustrates the application of Segment Routing to solve
   the Egress Peer Engineering (EPE) requirement. The SR-based EPE
   solution allows a centralized (SDN) controller to program any egress
   peer policy at ingress border routers or at hosts within the domain.
   This document is on the informational track.

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].

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 November 27, 2014.

Copyright Notice

Table of Contents

1. Introduction

   The document is structured as follows:

   o Section 1 reminds the EPE problem statement and provides the key
      references.

   o Section 2 defines the different BGP Peering Segments and the
      semantic associated to them.

   o Section 3 describes the automated allocation of BGP Peering SID's
      by the EPE-enabled egress border router and the automated
      signaling of the external peering topology and the related BGP
      Peering SID's to the collector [draft-previdi-idr-bgpls-segment-
      routing-epe-00].

   o Section 4 overviews the components of a centralized EPE
      controller. The definition of the EPE controller is outside the
      scope of this document.

   o Section 5 overviews the methods that could be used by the
      centralized EPE controller to implement an EPE policy at an
      ingress border router or at a source host within the domain. The
      exhaustive definition of all the means to program an EPE input
      policy is outside the scope of this document.

   For editorial reason, the solution is described for IPv4. A later
   section describes how the same solution is applicable to IPv6.

1.1. Segment Routing Documents

The main references for this document are:

o SR Problem Statement: [I-D.ietf-spring-problem-statement].

o SR Architecture: [I-D.filsfils-spring-segment-routing].

o Distribution of External Topology and TE Information using BGP:
   draft-previdi-idr-bgpls-segment-routing-epe-00.txt

The SR instantiation in the MPLS dataplane is described in
[I-D.filsfils-spring-segment-routing-mpls].

The SR IGP protocol extensions are defined in
[I-D.ietf-isis-segment-routing-extensions],
[I-D.psenak-ospf-segment-routing-extensions] and
[I-D.psenak-ospf-segment-routing-ospfv3-extension].

The Segment Routing PCE protocol extensions are defined in
[I-D.sivabalan-pce-segment-routing].

1.2. Problem Statement

The EPE problem statement is defined in
[I-D.ietf-spring-problem-statement].

A centralized controller should be able to instruct an ingress PE or
a content source within the domain to use a specific egress PE and a
specific external interface to reach a particular destination.

We call this solution "EPE" for "Egress Peer Engineering". The
centralized controller is called the "EPE Controller". The egress
border router where the EPE traffic-steering functionality is
implemented is called an EPE-enabled border router. The input policy
programmed at an ingress border router or at a source host is called
an EPE policy.

The requirements that have motivated the solution described in this
document are listed here below:

o The solution MUST apply to the Internet use-case where the
   Internet routes are assumed to use IPv4 unlabeled or IPv6
   unlabeled. It is not required to place the internet routes in a
   VRF and allocate labels on a per route, or on a per-path basis.

o The solution MUST NOT make any assumption on the currently
   deployed iBGP schemes (RRs, confederations or iBGP full meshes)
   and MUST be able to support all of them.

o The solution SHOULD minimize the need for new BGP capabilities at
   the ingress PE's.

o The solution MUST accommodate an ingress EPE policy at an ingress
   PE or directly at an source host within the domain.

o The solution MUST support automated FRR and fast convergence.

The following reference diagram is used throughout this document.

```
+---------+      +------+
|         |      |      |
|    H    B------D      G
|         | +---/| AS 2 |\  +------+
|         |/     +------+ \ |      |---L/8
A    AS1  C---+           \|      |
|         |\\  \  +------+ /| AS 4 |---M/8
|         | \\  +-E      |/ +------+
|    X    |  \\  |       K
|         |   +===F AS 3 |
+---------+      +------+
```
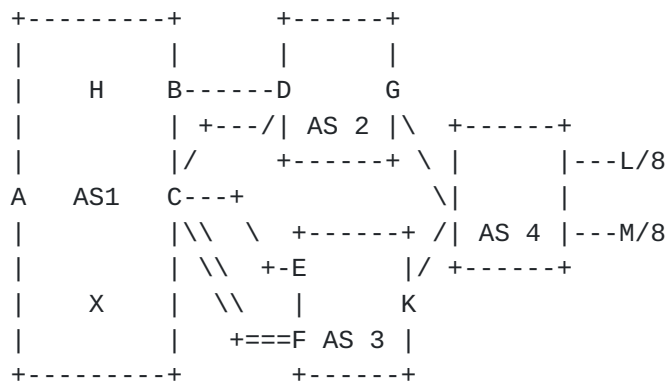
                    Figure 1: Reference Diagram

IPv4 addressing:

o C's interface to D: 1.0.1.1/24, D's interface: 1.0.1.2/24

o C's interface to E: 1.0.2.1/24, E's interface: 1.0.2.2/24

o C's upper interface to F: 1.0.3.1/24, F's interface: 1.0.3.2/24

o C's lower interface to F: 1.0.4.1/24, F's interface: 1.0.4.2/24

o Loopback of F used for eBGP multi-hop peering to C: 1.0.5.2/32

o C's loopback is 3.3.3.3/32 with SID 64

C's BGP peering:

o Single-hop eBGP peering with neighbor 1.0.1.2 (D)

o Single-hop eBGP peering with neighbor 1.0.2.2 (E)

o Multi-hop eBGP peering with F on ip address 1.0.5.2 (F)

C's resolution of the multi-hop eBGP session to F:

o Static route 1.0.5.2/32 via 1.0.3.2

o Static route 1.0.5.2/32 via 1.0.4.2

C is configured with local policy that defines a BGP PeerSet as the
set of peers (1.0.2.2 and 1.0.5.2)

X is the EPE controller within AS1 domain.

H is a content source within AS1 domain.

2. BGP Peering Segments

AS defined in [I-D.filsfils-spring-segment-routing], Segments are
defined by a Egress Peer Engineering (EPE) capable node and
corresponding to its attached peers. These segments are called BGP
peering segments or BGP Peering SIDs. They enable the expression of
source-routed inter-domain paths.

An ingress border router of an AS may compose a list of segments to
steer a flow along a selected path within the AS, towards a selected
egress border router C of the AS and through a specific peer. At
minimum, a BGP Peering Engineering policy applied at an ingress PE
involves two segments: the Node SID of the chosen egress PE and then
the BGP Peering Segment for the chosen egress PE peer or peering
interface.

[I-D.filsfils-spring-segment-routing] defines three types of BGP
peering segments/SID's: PeerNodeSID, PeerAdjSID and PeerSetSID.

The BGP extensions to signal these BGP peering segments are outlined
in the following section.

3. Distribution of External Topology and TE Information using BGP-LS

In ships-in-the-night mode with respect to the pre-existing iBGP
design, a BGPLS session is established between the EPE-enabled border
router and the EPE controller.

As a result of its local configuration and according to the behavior
described in draft-previdi-idr-bgpls-segment-routing-epe-00, node C
allocates the following BGP Peering Segments
([I-D.filsfils-spring-segment-routing]):

o A PeerNode segment for each of its defined peer (D, E and F).

o A PeerAdj segment for each recursing interface to a multi-hop peer
   (e.g.: the upper and lower interfaces from C to F in figure 1).

o A PeerSet segment to the set of peers (E and F).

C programs its forwarding table accordingly:

```
Incoming             Outgoing
Label     Operation  Interface
-----------------------------------
1012          POP    link to D
1022          POP    link to E
1032          POP    upper link to F
1042          POP    lower link to F
1052          POP    loadbalance on any link to F
1060          POP    loadbalance on any link to E or to F
```

C signals the related BGP-LS NLRI's to the EPE controller. Each such
BGP-LS route is described in the following sub-sections according to
the encoding details defined in draft-previdi-idr-bgpl-segment-
routing-epe-00.

3.1. EPE Route advertising the Peer D and its PeerNode SID

Descriptors:

o Node Descriptors (router-ID, ASN): 3.3.3.3 , AS1

o Peer Descriptors (peer ASN): AS2

o Link Descriptors (IPv4 interface address, neighbor IPv4 address):
   1.0.1.1, 1.0.1.2

Attributes:

o Adj-SID: 1012

3.2. EPE Route advertising the Peer E and its PeerNode SID

Descriptors:

o Node Descriptors (router-ID, ASN): 3.3.3.3 , AS1

o Peer Descriptors (peer ASN): AS3

o Link Descriptors (IPv4 interface address, neighbor IPv4 address):
   1.0.2.1, 1.0.2.2

Attributes:

o Adj-SID: 1022

o PeerSetSID: 1060

o Link Attributes: see section 3.3.2 of
   [I-D.ietf-idr-ls-distribution]

3.3. EPE Route advertising the Peer F and its PeerNode SID

Descriptors:

o Node Descriptors (router-ID, ASN): 3.3.3.3 , AS1

o Peer Descriptors (peer ASN): AS3

o Link Descriptors (IPv4 interface address, neighbor IPv4 address):
   3.3.3.3, 1.0.5.2

Attributes:

o Adj-SID: 1052

o PeerSetSID: 1060

3.4. EPE Route advertising a first PeerAdj to Peer F

Descriptors:

o Node Descriptors (router-ID, ASN): 3.3.3.3 , AS1

o Peer Descriptors (peer ASN): AS3

o Link Descriptors (IPv4 interface address, neighbor IPv4 address):
   1.0.3.1 , 1.0.3.2

Attributes:

o Adj-SID: 1032

o LinkAttributes: see section 3.3.2 of
   [I-D.ietf-idr-ls-distribution]

3.5. EPE Route advertising a second PeerAdj to Peer F

Descriptors:

o Node Descriptors (router-ID, ASN): 3.3.3.3 , AS1

o Peer Descriptors (peer ASN): AS3

o Link Descriptors (IPv4 interface address, neighbor IPv4 address):
   1.0.4.1 , 1.0.4.2

Attributes:

o Adj-SID: 1042

o LinkAttributes: see section 3.3.2 of
   [I-D.ietf-idr-ls-distribution]

3.6. FRR

An EPE-enabled border router should allocate a FRR backup entry on a
per BGP Peering SID basis:

o PeerNode SID

   1. If multi-hop, backup via the remaining PeerADJ SID's to the
       same peer.

   2. Else backup via local PeerNode SID to the same AS.

   3. Else pop the PeerNode SID and IP lookup (with potential BGP
       PIC fall-back).

o PeerAdj SID

   1. If to a multi-hop peer, backup via the remaining PeerADJ SID's
       to the same peer.

   2. Else backup via PeerNode SID to the same AS.

   3. Else pop the PeerNode SID and IP lookup (with potential BGP
       PIC fall-back).

o PeerSet SID

   1. Backup via remaining PeerNode SID in the same PeerSet.

2. Else pop the PeerSet SID and IP lookup (with potential BGP PIC
     fall-back).

We illustrate the different types of possible backups using the
reference diagram and considering the Peering SID's allocated by C.

PeerNode SID 1052, allocated by C for peer F:

o Upon the failure of the upper connected link CF, C can reroute all
   the traffic onto the lower CF link to the same peer (F).

PeerNode SID 1022, allocated by C for peer E:

o Upon the failure of the connected link CE, C can reroute all the
   traffic onto the link to PeerNode SID 1052 (F).

PeerNode SID 1012, allocated by C for peer D:

o Upon the failure of the connected link CD, C can pop the PeerNode
   SID and lookup the IP destination address in its FIB and route
   accordingly.

PeerSet SID 1060, allocated by C for the set of peers E and F:

o Upon the failure of a connected link in the group, the traffic to
   PeerSet SID 1060 is rerouted on any other member of the group.

For specific business reasons, the operator might not want the
default FRR behavior applied to a PeerNode SID or any of its
depending PeerADJ SID.

The operator should be able to associate a specific backup PeerNode
SID for a PeerNode SID: e.g. 1022 (E) must be backed up by 1012 (D)
which over-rules the default behavior which would have preferred F as
a backup for E.

[4](). EPE Controller

In this section, we provide a non-exhaustive set of inputs that an
EPE controller would likely collect such as to perform the EPE policy
decision.

The exhaustive definition is outside the scope of this document.

4.1. Valid Paths From Peers

   The EPE controller should collect all the paths advertised by all the
   engineered peers.

   This could be realized by setting an iBGP session with the EPE-
   enabled border router, with "add-path all" and original next-hop
   preserved.

   In this case, C would advertise the following Internet routes to the
   EPE controller:

   o NLRI <L/8>, nhop 1.0.1.2, AS Path {AS 2, 4}

      * X (i.e.: the EPE controller) knows that C receives a path to
         L/8 via neighbor 1.0.1.2 of AS2.

   o NLRI <L/8>, nhop 1.0.2.2, AS Path {AS 3, 4}

      * X knows that C receives a path to L/8 via neighbor 1.0.2.2 of
         AS2.

   o NLRI <L/8>, nhop 1.0.5.2, AS Path {AS 3, 4}

      * X knows that C has an eBGP path to L/8 via AS3 via neighbor
         1.0.5.2

   An alternative option consists in Adj-RIB-In BMP from EPE-enabled
   border router to the EPE collector.

4.2. Intra-Domain Topology

   The EPE controller should collect the internal topology and the
   related IGP SID's.

   This could be realized by collecting the IGP LSDB of each area or
   running a BGP-LS session with a node in each IGP area.

4.3. External Topology

   Thanks to the collected BGP-LS routes described in the section 2
   (BGPLS advertisements), the EPE controller is able to maintain an
   accurate description of the egress topology of node C. Furthermore,
   the EPE controller is able to associate BGP Peering SID's to the
   various components of the external topology.

4.4. SLA characteristics of each peer

The EPE controller might collect SLA characteristics across peers.
This requires an EPE solution as the SL A probes need to be steered
via non-best-path peers.

Uni-directional SLA monitoring of the desired path is likely
required. This might be possible when the application is controlled
at the source and the receiver side. Uni-directional monitoring
dissociates the SLA characteristic of the return path (which cannot
usually be controlled) from the forward path (the one of interest for
pushing content from a source to a consumer and the one which can be
controlled).

Alternatively, Extended Metrics, as defined in
[I-D.ietf-isis-te-metric-extensions] could also be advertised using
new bgpls attributes.

4.5. Traffic Matrix

The EPE controller might collect the traffic matrix to its peers or
the final destinations. IPFIX is a likely option.

An alternative option consists in collecting the link utilization
statistics of each of the internal and external links, also available
in current definition of [I-D.ietf-idr-ls-distribution].

4.6. Business Policies

The EPE controller should collect business policies.

4.7. EPE Policy

On the basis of all these inputs (and likely other), the EPE
Controller decides to steer some demands away from their best BGP
path.

The EPE policy is likely expressed as a two-entry segment list where
the first element is the IGP prefix SID of the selected egress border
router and the second element is a BGP Peering SID at the selected
egress border router.

A few examples are provided hereafter:

o Prefer egress PE C and peer AS AS2: {64, 1012}.

o Prefer egress PE C and peer AS AS3 via ebgp peer 1.0.2.2: {64,
   1022}.

o Prefer egress PE C and peer AS AS3 via ebgp peer 1.0.5.2: {64,
   1052}.

o Prefer egress PE C and peer AS AS3 via interface 1.0.4.2 of multi-
   hop ebgp peer 1.0.5.2: {64, 1042}.

o Prefer egress PE C and any interface to any peer in the group
   1060: {64, 1060}.

Note that the first SID could be replaced by a list of segments.
This is useful when an explicit path within the domain is required
for traffic-engineering purpose. For example, if the Prefix SID of
node B is 60 and the EPE controller would like to steer the traffic
from A to C via B then through the external link to peer D then the
segment list would be {60, 64, 1012}.

5. Programming an input policy

The detailed/exhaustive description of all the means to implement an
EPE policy are outside the scope of this document. A few examples
are provided in this section.

5.1. At a Host

A static IP/MPLS route can be programmed at the host H. The static
route would define a destination prefix, a next-hop and a label stack
to push. The global property of the IGP Prefix SID is particularly
convenient: the same policy could be programmed across hosts
connected to different routers.

5.2. At a router - SR Traffic Engineering tunnel

The EPE controller can configure the ingress border router with an SR
traffic engineering tunnel T1 and a steering-policy S1 which causes a
certain class of traffic to be mapped on the tunnel T1.

The tunnel T1 would be configured to push the require segment list.

The tunnel and the steering policy could be configured via PCEP
according to [I-D.sivabalan-pce-segment-routing] and
[I-D.ietf-pce-pce-initiated-lsp] or via Netconf ([RFC6241]).

Example: at A

Tunnel T1: push {64, 1042}
IP route L/8 set nhop T1

5.3. At a Router - BGP3107 policy route

The EPE Controller could build a BGP3107 ([RFC3107]) route (from
scratch) and send it to the ingress router:

o NLRI: the destination prefix to engineer: e.g. L/8.

o Next-Hop: the selected egress border router: C.

o Label: the selected egress peer: 1042.

o AS path: reflecting the valid AS path of the selected.

o Some BGP policy to ensure it be selected as best by the ingress
   router.

This BGP3107 policy route "overwrites" an equivalent or less-specific
"best path". As the best-path is changed, this EPE input policy
option influences the path propagated to the upstream peer/customers.

5.4. At a Router - VPN policy route

The EPE Controller could build a VPNv4 route (from scratch) and send
it to the ingress router:

o NLRI: the destination prefix to engineer: e.g. L/8.

o Next-Hop: the selected egress border router: C.

o Label: the selected egress peer: 1042.

o Route-Target: selecting the appropriate VRF at the ingress router.

o AS path: reflecting the valid AS path of the selected.

o Some BGP policy to ensure it be selected as best by the ingress
   router in the related VRF.

The related VRF must be pre-configured. A VRF fall-back into main
FIB might be beneficial to avoid replicating all the "normal"
internet paths in each VRF.

5.5. At a Router - Flowspec route

EPE Controller builds a FlowSpec route and sends it to the ingress
router to engineer:

o Dissemination of Flow Specification Rules ([RFC5575].

   o Destination/Source IP Addresses, IP Protocol, Destination/Source
      port (+1 component).

   o ICMP Type/Code, TCP Flags, Packet length, DSCP, Fragment.

6. IPv6

   The described solution is applicable to IPv6, either with MPLS-based
   or IPv6-Native segments. In both cases, the same three steps of the
   solution are applicable:

   o BGP-LS-based signaling of the external topology and BGP Peering
      Segments to the EPE controller.

   o Collection of various inputs by the EPE controller to come up with
      a policy decision.

   o Programming at an ingress router or source host of the desired EPE
      policy which consists in a list of segments to push on a defined
      traffic class.

7. Benefits

   The EPE solutions described in this document has the following
   benefits:

   o No assumption on the iBGP design with AS1.

   o Next-Hop-Self on the internet routes propagated to the ingress
      border routers is possible. This is a common design rule to
      minimize the number of IGP routes and to avoid importing external
      churn into the internal domain.

   o Consistent support for traffic-engineering within the domain and
      at the external edge of the domain.

   o Support host and ingress border router EPE policy programming.

   o EPE functionality is only required on the EPE-enabled egress
      border router and the EPE controller: an ingress policy can be
      programmed at the ingress border router without any new
      functionality.

   o Ability to deploy the same input policy across hosts connected to
      different routers (global property of the IGP prefix SID).

8. IANA Considerations

   TBD

9. Manageability Considerations

   TBD

10. Security Considerations

   TBD

11. Acknowledgements

   TBD

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.

   [RFC3107] Rekhter, Y. and E. Rosen, "Carrying Label Information in
              BGP-4", RFC 3107, May 2001.

   [RFC5575] Marques, P., Sheth, N., Raszuk, R., Greene, B., Mauch, J.,
              and D. McPherson, "Dissemination of Flow Specification
              Rules", RFC 5575, August 2009.

   [RFC6241] Enns, R., Bjorklund, M., Schoenwaelder, J., and A.
              Bierman, "Network Configuration Protocol (NETCONF)", RFC
              6241, June 2011.

12.2. Informative References

   [I-D.filsfils-spring-segment-routing]
              Filsfils, C., Previdi, S., Bashandy, A., Decraene, B.,
              Litkowski, S., Horneffer, M., Milojevic, I., Shakir, R.,
              Ytti, S., Henderickx, W., Tantsura, J., and E. Crabbe,
              "Segment Routing Architecture", draft-filsfils-spring-
              segment-routing-01 (work in progress), May 2014.

   [I-D.filsfils-spring-segment-routing-mpls]
             Filsfils, C., Previdi, S., Bashandy, A., Decraene, B.,
             Litkowski, S., Horneffer, M., Milojevic, I., Shakir, R.,
             Ytti, S., Henderickx, W., Tantsura, J., and E. Crabbe,
             "Segment Routing with MPLS data plane", draft-filsfils-
             spring-segment-routing-mpls-01 (work in progress), April
             2014.

   [I-D.ietf-idr-ls-distribution]
             Gredler, H., Medved, J., Previdi, S., Farrel, A., and S.
             Ray, "North-Bound Distribution of Link-State and TE
             Information using BGP", draft-ietf-idr-ls-distribution-05
             (work in progress), May 2014.

   [I-D.ietf-isis-segment-routing-extensions]
             Previdi, S., Filsfils, C., Bashandy, A., Gredler, H.,
             Litkowski, S., and J. Tantsura, "IS-IS Extensions for
             Segment Routing", draft-ietf-isis-segment-routing-
             extensions-00 (work in progress), April 2014.

   [I-D.ietf-isis-te-metric-extensions]
             Previdi, S., Giacalone, S., Ward, D., Drake, J., Atlas,
             A., Filsfils, C., and W. Wu, "IS-IS Traffic Engineering
             (TE) Metric Extensions", draft-ietf-isis-te-metric-
             extensions-03 (work in progress), April 2014.

   [I-D.ietf-pce-pce-initiated-lsp]
             Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "PCEP
             Extensions for PCE-initiated LSP Setup in a Stateful PCE
             Model", draft-ietf-pce-pce-initiated-lsp-00 (work in
             progress), December 2013.

   [I-D.ietf-spring-problem-statement]
             Previdi, S., Filsfils, C., Decraene, B., Litkowski, S.,
             Horneffer, M., Geib, R., Shakir, R., and R. Raszuk,
             "SPRING Problem Statement and Requirements", draft-ietf-
             spring-problem-statement-00 (work in progress), May 2014.

   [I-D.psenak-ospf-segment-routing-extensions]
             Psenak, P., Previdi, S., Filsfils, C., Gredler, H.,
             Shakir, R., and W. Henderickx, "OSPF Extensions for
             Segment Routing", draft-psenak-ospf-segment-routing-
             extensions-04 (work in progress), February 2014.

   [I-D.psenak-ospf-segment-routing-ospfv3-extension]
              Psenak, P., Previdi, S., Filsfils, C., Gredler, H.,
              Shakir, R., and W. Henderickx, "OSPFv3 Extensions for
              Segment Routing", draft-psenak-ospf-segment-routing-
              ospfv3-extension-01 (work in progress), February 2014.

   [I-D.sivabalan-pce-segment-routing]
              Sivabalan, S., Medved, J., Filsfils, C., Crabbe, E., and
              R. Raszuk, "PCEP Extensions for Segment Routing", draft-
              sivabalan-pce-segment-routing-02 (work in progress),
              October 2013.

Authors' Addresses

   Clarence Filsfils (editor)
   Cisco Systems, Inc.
   Brussels
   BE

   Email: cfilsfil@cisco.com

   Stefano Previdi (editor)
   Cisco Systems, Inc.
   Via Del Serafico, 200
   Rome 00142
   Italy

   Email: sprevidi@cisco.com

   Keyur Patel
   Cisco Systems, Inc.
   US

   Email: keyupate@cisco.com

   Ebben Aries
   Facebook
   US

   Email: exa@fb.com

Daniel Ginsburg
Yandex
RU

Email: dbg@yandex-team.ru

Dmitry Afanasiev
Yandex
RU

Email: fl0w@yandex-team.ru