LSR Working Group   U. Chunduri
Internet-Draft   R. Li
Intended status: Standards Track   Futurewei
Expires: April 1, 2021   R. White
                                                      Juniper Networks
                                                          J. Tantsura
                                                          Apstra Inc.
                                                         L. Contreras
                                                           Telefonica
                                                                Y. Qu
                                                            Futurewei
                                                   September 28, 2020

              Preferred Path Routing (PPR) in IS-IS
            draft-chunduri-lsr-isis-preferred-path-routing-06

Abstract

   This document specifies Preferred Path Routing (PPR), an extensible
   method of providing path based dynamic routing for a number of packet
   types including IPv4, IPv6 and MPLS. PPR uses a simple encapsulation
   to add the path identity to the packet. PPR can also be used to
   mitigate the MTU and data plane processing issues that may result
   from Segment Routing (SR) packet overhead; and also supports further
   extensions along the paths.

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 RFC2119 [RFC2119],
   RFC8174 [RFC8174] when, and only when they appear in all capitals, as
   shown here.

Status of This Memo

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 April 1, 2021.

Copyright Notice

Table of Contents

1. Introduction

   In a network implementing Segment Routing (SR), packets are steered
   through the network using Segment Identifiers (SIDs) carried in the
   packet header. Each SID uniquely identifies a segment as defined in
   [RFC8402] . SR capabilities are defined for MPLS and IPv6 data planes
   called SR-MPLS and SRv6 respectively. Appendix A.1 and Appendix A.2
   describe performance, hardware capabilities and various associated
   issues which may result in SR deployments.

   The above motivate the proposed solution, Preferred Path Routing
   (PPR). In brief, PPR involves, associating path descriptions to IS-
   IS advertised prefixes, mapping those to a data-plane identifier and
   specifying a mechanism to route packets with the abstracted
   identifier (PPR-ID), as opposed to individual segments on the packet.
   This is specified in detail in Section 2 and Section 3.

1.1. Acronyms

   EL   - Entropy Label

   ELI   - Entropy Label Indicator

   LSP   - IS-IS Link State PDU

   MPLS   - Multi Protocol Label Switching

   MSD   - Maximum SID Depth

   MTU   - Maximum Transferrable Unit

   NH   - Next-Hop

   PPR   - Preferred Path Routing/Route

   PPR-ID   - Preferred Path Route Identifier, a data plane identifier

   SID   - Segment Identifier

   SPF   - Shortest Path First

   SR-MPLS   - Segment Routing with MPLS data plane

SRH   - Segment Routing Header - IPv6 routing Extension header

SRv6   - Segment Routing with IPv6 data plane with SRH

TE   - Traffic Engineering

2. Preferred Path Routing (PPR)

PPR mitigates the issues described in Appendix A.1, while continuing
to allow the direction of traffic along an engineered path through
the network by replacing the label stack with a PPR-ID. The PPR-ID
can either be a single label or a native destination address. To
facilitate the use of a single label to describe an entire path, a
new TLV is added to IS-IS, as described below in Section 3.

A PPR could be an SR path, a traffic engineered path computed based
on some constraints, an explicitly provisioned Fast Re-Route (FRR)
path or a service chained path. A PPR can be signaled by any node,
computed by a central controller, or manually configured by an
operator. PPR extends the source routing and path steering
capabilities to native IP (IPv4 and IPv6) data planes without
hardware upgrades; see Section 5.

2.1. PPR-ID and Data Plane Extensibility

The PPR-ID describes a path through the network. A data plane type
and corresponding data plane identifier as specified in Section 3.2
is mapped to PPR-ID to allow data plane extensibility.

For SR-MPLS, PPR-ID is mapped to an MPLS Label/SID and for SRv6, this
is mapped to an IPv6-SID. For native IP data planes, this is mapped
to either IPv4 or IPv6 address/prefix.

2.2. PPR Path Description

The path identified by the PPR-ID is described as a set of Path
Description Elements (PDEs), each of which represents a segment of
the path. Each node determines its location in the path as
described, and forwards to the next segment/hop or label of the path
description (see the Forwarding Procedure Example later in this
document).

These PPR-PDEs as defined in Section 3.3, like SR SIDs, can represent
topological elements like links/nodes, backup nodes, as well as non-
topological elements such as a service, function, or context on a
particular node.

A PPR path can be described as a Strict-PPR or a Loose-PPR. In a
Strict-PPR all nodes/links on the path are described with SR SIDs for
SR data planes or IPv4/IPV6 addresses for native IP data planes. In
a Loose-PPR only some of the nodes/links from source to destination
are described. More specifics and restrictions around Strict/Loose
PPRs are described in respective data planes in Section 5. Each PDE
is described as either an MPLS label towards the Next-Hop (NH) in
MPLS enabled networks, or as an IP NH, in the case of either
"plain"/"native" IP or SRv6 enabled networks. A PPR path is related
to a set of PDEs using the TLVs as specified in Section 3.

2.3. ECMP Considerations

PPR inherently supports Equal Cost Multi Path (ECMP) for both strict
and loose paths. If a path is described using nodes, would have ECMP
NHs established for PPR-ID along the path. However, one can avoid
ECMP on any segment of the path by pinning the path using link
identifier to the next segment.

2.4. Scalability and PPR Graphs

In a network of N nodes O(N^2) total (unidirectional) paths are
necessary to establish any-to-any connectivity, and multiple (k) such
path sets may be desirable if multiple path policies are to be
supported (lowest latency, highest throughput etc.).

In many solutions and topologies, N may be small enough and/or only a
small set of paths need to be preferred paths, for example for high
value traffic (DetNet, some of the defined 5G slices), and then the
technology specified in this document can support these deployments.

However, to address the scale needed when a larger number of PPR
paths are required, the PPR TREE structure can be used [I-D.draft-ce-
ppr-graph-00]. Each PPR Tree uses one label/SID and defines paths
from any set of nodes to one destination, thus reduces the number of
entries needed in SRGB at each node (for SR-MPLS data plane
Section 5.1). These paths form a tree rooted in the destination. In
other word, PPR Tree identifiers are destination identifiers and PPR
Treed are path engineered destination routes (like IP routes) and it
scaling simplifies to linear in N i.e., O(k*N).

3. PPR Related TLVs

This section describes the encoding of PPR TLV. This TLV can be seen
as having 4 logical sections viz, encoding of the PPR-Prefix (IS-IS
Prefix), encoding of PPR-ID, encoding of path description with an
ordered PDE Sub-TLVs and a set of optional PPR attribute Sub-TLVs,
which can be used to describe one or more parameters of the path.

Multiple instances of this TLV MAY be advertised in IS-IS LSPs with
different PPR-ID Type (data plane) and with corresponding PDE Sub-
TLVS. The PPR TLV has Type TBD (suggested value xxx), and has the
following format:

```
 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     |  PPR-Flags                    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Fragment-ID   |            MT-ID              | Algorithm      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           PPR-Prefix Sub-TLV (variable size)                 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           PPR-ID Sub-TLV (variable size)                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           PPR-PDE Sub-TLVs (variable)                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           PPR-Attribute Sub-TLVs (variable)                  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
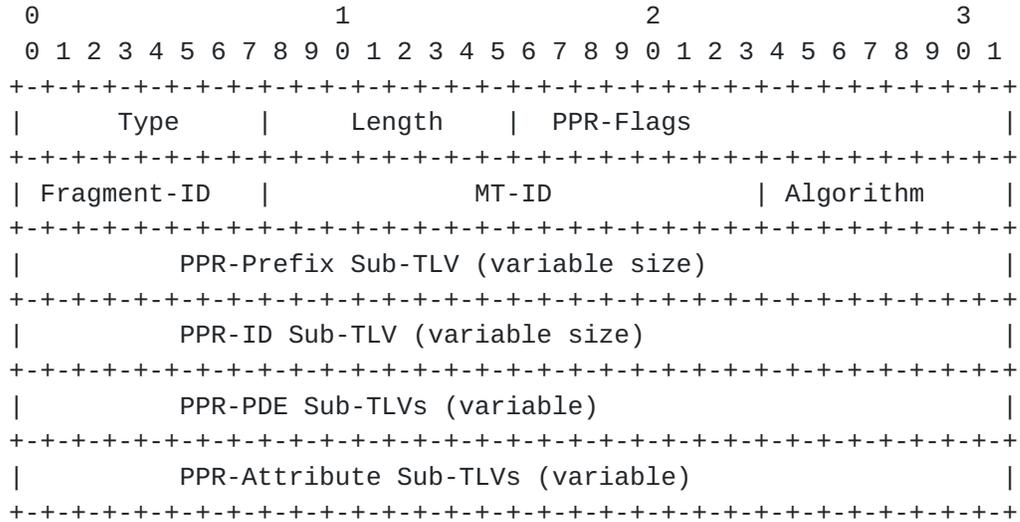
Figure 1: PPR TLV Format

o Type: 155 (Suggested Value, TBD IANA) from IS-IS top level TLV
   registry.

o Length: Total length of the value field in bytes.

o PPR-Flags: 2 Octet bit-field of flags for this TLV; described
   below.

o Fragment-ID: This is an 8-bit Identifier value (0-255) of the TLV
   fragment. If fragments are not needed to represent the complete
   path, 'U' bit MUST be set and this value MUST be set to 0.

o MT-ID: The multi-topology identifier defined in [RFC5120]; the 4
   most significant bits MUST be set to 0 on transmit and ignored on
   receive. The remaining 12-bit field contains the MT-ID.

o Algorithm: 1 octet value represents the route computation
   algorithm. Algorithm registry is as defined in
   [I-D.ietf-isis-segment-routing-extensions]. Computation towards
   PPR-ID (Section 3.2) happens per MT-ID/Algorithm pair.

o PPR-Prefix: A variable size Sub-TLV representing the destination
   of the path being described. This is defined in Section 3.1.

o PPR-ID: A variable size Sub-TLV representing the data plane or
   forwarding identifier of the PPR. Defined in Section 3.2.

o PPR-PDEs: Variable number of ordered PDE Sub-TLVs which represents
   the path. This is defined in Section 3.3.

o PPR-Attributes: Variable number of PPR-Attribute Sub-TLVs which
   represent the path attributes. These are defined in Section 3.4.

The Flags field has the following flag bits defined:

    PPR TLV Flags Format

        0 1 2 3 4 5 6 7   15
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |F|D|A|U|Reserved |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
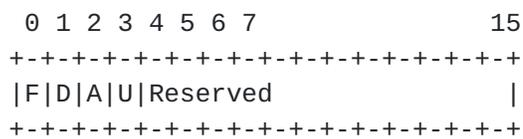
1. F: Flood bit. If set, the PPR TLV MUST be flooded across the
    entire routing domain. If the F bit is not set, the PPR TLV MUST
    NOT be leaked between IS-IS levels. This bit MUST NOT be altered
    during the TLV leaking

2. D: Down Bit. When the PPR TLV is leaked from IS-IS level-2 to
    level-1, the D bit MUST be set. Otherwise, this bit MUST be
    clear. PPR TLVs with the D bit set MUST NOT be leaked from
    level-1 to level-2. This is to prevent TLV looping across
    levels.

3. A: Attach bit. The originator of the PPR TLV MUST set the A bit
    in order to signal that the prefix and PPR-ID advertised in the
    PPR TLV are directly connected to the originators. If this bit
    is not set, this allows any other node in the network to
    advertise this TLV on behalf of the originating node of the PPR-
    Prefix. If PPR TLV is leaked to other areas/levels the A-flag
    MUST be cleared. In case if the originating node of the prefix
    must be disambiguated for any reason including, if it is a Multi
    Homed Prefix (MHP) or leaked to a different IS-IS level or
    because [RFC7794] X-Flag is set, then PPR-Attribute Sub-TLV
    Source Router ID SHOULD be included.

4. U: Ultimate fragment bit. bit MUST be set if a path has only one
    fragment or if it is the last Fragment of the path. PPR-ID value
    for all fragments of the same path MUST be same.

5. Reserved: For future use; MUST be set to 0 on transmit and
    ignored on receive.

PPR path description for each IS-IS level is computed and given to one of the nodes for L1 and L2 respectively. Similarly path information when crossing the level boundaries MUST be relevant to the destination level. If there is no path information available for the destination level PPR TLV MUST NOT be leaked regardless of F and D bits as defined above.

The following Sub-TLVs draw from a new registry for Sub-TLV numbers as specified in Section 7.1 and Section 7.2.

3.1. PPR-Prefix Sub-TLV

The structure of PPR-Prefix is:

```
    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        | Prefix Length | Mask Length   |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   //            IS-IS Prefix (variable)                          //
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
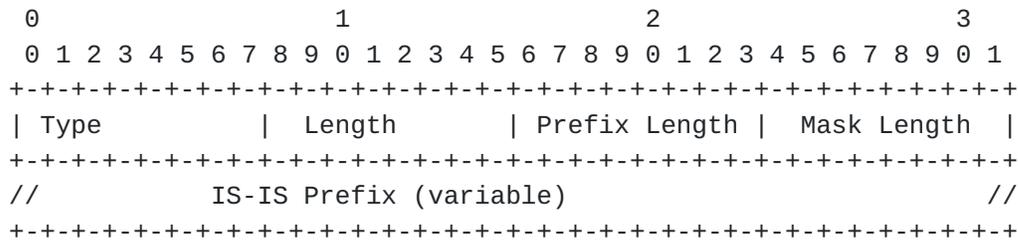
Figure 2: PPR-Prefix Sub-TLV Format

o Type: 1 (IANA to assign from Sub-TLV registry described above).

o Length: Total length of the value field in bytes.

o Prefix Length: The length of the IS-IS Prefix being encoded in bytes. For IPv4 it MUST be 4 and IPv6 it MUST be 16 bytes.

o Mask Length: The length of the prefix in bits. Only the most significant octets of the Prefix are encoded.

o IS-IS Prefix: The IS-IS prefix at the tail-end of the advertised PPR. This corresponds to a routable prefix of the originating node and it MAY have one of the [RFC7794] flags set (X-Flag/R-Flag/N-Flag) in the IS-IS reachability TLVs. Length of this field MUST be as per "Prefix Length". Encoding is same as TLV 135 [RFC5305] and TLV 236 [RFC5308] or MT-Capable [RFC5120] IPv4 (TLV 235) and IPv6 Prefixes (TLV 237) respectively.

3.2. PPR-ID Sub-TLV

This is the actual data plane identifier in the packet header and could be of any data plane as defined in PPR-ID Type field. Both PPR-Prefix and PPR-ID belongs to a same node in the network.

```
 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        |PPR-ID Flags                   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| PPR-ID Type   | PPR-ID Length |PPR-ID Mask Len|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
//                    PPR-ID (variable size)                  //
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
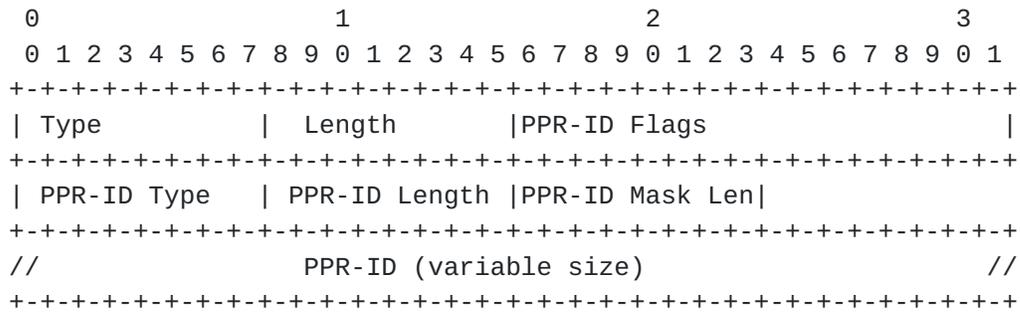
   Figure 3: PPR-ID Sub-TLV Format

o Type: 2 (IANA to assign from Sub-TLV registry described above).

o Length: Total length of the value field in bytes.

o PPR-ID Flags: 2 Octet field for PPR-ID flags:

     PPR-ID Flags Format

```
        0 1 2 3 4 5 6 7..               15
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |     Reserved                  |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
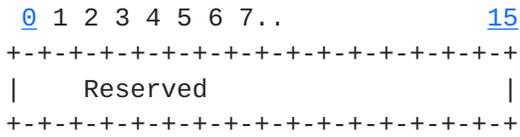
     Reserved: For future use; MUST be set to 0 on transmit and
     ignored on receive.

o PPR-ID Type: Data plane type of PPR-ID. This is a new registry
   (TBD IANA - Suggested values as below) for this Sub-TLV and the
   defined types are as follows:

     Type: 1 SR-MPLS SID/Label

     Type: 2 Native IPv4 Address/Prefix

     Type: 3 Native IPv6 Address/Prefix

     Type: 4 IPv6 SID in SRv6 with SRH

o PPR-ID Length: Length of the PPR-ID field in octets and this
   depends on the PPR-ID type.

o PPR-ID Mask Len: It is applicable for only for PPR-ID Type 2, 3
   and 4. For Type 1 this value MUST be set to zero. It contains
   the length of the PPR-ID Prefix in bits. Only the most
   significant octets of the Prefix are encoded. This is needed, if

PPR-ID followed is an IPv4/IPv6 Prefix instead of 4/16 octet
Address respectively.

o PPR-ID: This is the Preferred Path forwarding identifier that
   would be on the data packet. The value of this field is variable
   and it depends on the PPR-ID Type - for Type 1, this is encoded as
   SR-MPLS SID/Label. For Type 2 this is a 4 byte IPv4 address. For
   Type 3 this is a 16 byte IPv6 address. For Type 2 and Type 3
   encoding is similar to "IS-IS Prefix" as specified in Section 3.1.
   For Type 4, this is encoded as 16 byte SRv6 SID.

For PPR-ID Type 2, 3 or 4, PPR-ID MUST NOT be advertised as a
routable prefix in TLV 135, TLV 235, TLV 236 and TLV 237. PPR-ID
MUST belong to the node, from where the PPR-Prefix (Section 3.1) is
advertised.

3.3. PPR-PDE Sub-TLV

This Sub-TLV represents the PPR Path Description Element (PDE). PPR-
PDEs are used to describe the path in the form of set of contiguous
and ordered Sub-TLVs, where first Sub-TLV represents (the top of the
stack in MPLS data plane or) first node/segment of the path. These
set of ordered Sub-TLVs can have both topological elements and non-
topological elements (e.g., service segments).

```
 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        | PPR-PDE Type  | PDE-ID Type   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| PDE-ID Length | PPR-PDE Flags                 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
//               PDE-ID Value (variable size)                 //
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Sub-TLV Length |  PPR-PDE  Sub-TLVs (variable)               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
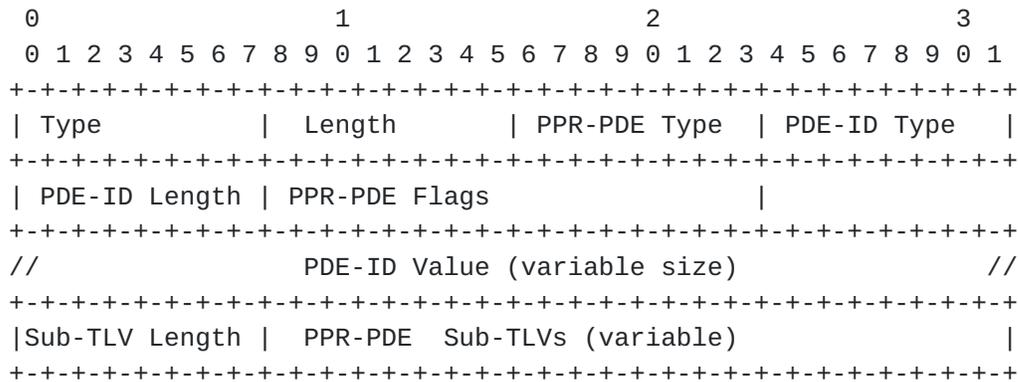
   Figure 4: PPR-PDE Sub-TLV Format

o Type: 3 (See IANA for suggested value) from IS-IS PPR TLV
   Section 3 Sub-TLV registry.

o Length: Total length of the value field in bytes.

o PPR-PDE Type: A new registry (TBD IANA) for this Sub-TLV and the
   defined types are as follows:

      Type: 1 Topological

     Type: 2 Non-Topological

  o PDE-ID Type: 1 Octet PDE-forwarding IDentifier Type. A new
     registry (Suggested Values as listed, IANA TBD) for this Sub-TLV
     and the defined types and corresponding PDE-ID Length, PDE-ID
     Value are as follows:

     Type 0: This value MUST be set only when PPR-PDE Type is Non-
     Topological. PDE-ID Length indicates the length of the PDE-ID
     Value field in bytes. For this type, PDE-ID value represents a
     service/function. This information is provisioned on the
     immediate topological PDE preceding to this PDE based on the 'E'
     bit.

     Type 1: SID/Label type as defined in
     [I-D.ietf-isis-segment-routing-extensions]. PDE-ID Length and
     PDE-ID Value fields are per Section 2.3 of the referenced
     document.

     Type 2: SR-MPLS Prefix SID. PDE-ID Length and PDE-ID Value are
     same as Type 1.

     Type 3: SR-MPLS Adjacency SID. PDE-ID Length and PDE-ID Value
     are same as Type 1.

     Type 4: IPv4 Node Loopback Address. PDE-ID Length 4 bytes and
     PDE-ID Value is full 4 bytes IPv4 address encoded as specified
     in "4-octet IPv4 address" of Sub-TLV 6/TLV 22 in [RFC5305].

     Type 5: IPv4 Interface Address. PDE-ID Length is 4 bytes and
     PDE-ID Value is full 4 bytes IPv4 address encoded as specified
     in "4-octet IPv4 address" of Sub-TLV 6/TLV 22 in [RFC5305].
     This PDE-ID in the path description represents the egress
     interface of the path segment and correponding adjacency is set
     as nexthop for the PPR-ID.

     Type 6: IPv6 Node Loopback Address. PDE-ID Length and PDE-ID
     Value are encoded as specified in "Prefix Len" and "prefix"
     portion of TLV 236 in [RFC5308] respectively.

     Type 7: IPv6 Interface Address. PDE-ID Length is 16 bytes and
     PDE-ID Value is full 16 bytes IPv6 address encoded as specified
     in "Interface Address 1" portion of TLV 232 in [RFC5308]. This
     PDE-ID in the path description represents the egress interface
     of the path segment and correponding adjacency is set as nexthop
     for the PPR-ID.

Type 8: SRv6 Node SID as defined in
[I-D.bashandy-isis-srv6-extensions]. PDE-ID Length and PDE-ID
Value are as defined in SRv6 SID from the referenced draft.

Type 9: SRv6 Adjacency-SID. PDE-ID Length and PDE-ID Values are
similar to SRv6 Node SID above.

o PPR-PDE Flags: 2 Octet bit-field of flags; described below:

   PPR-PDE Flags Format

       0 1 2 3 4 5 6 7 ..   15
      +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
      |L|N|E| Reserved |
      +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
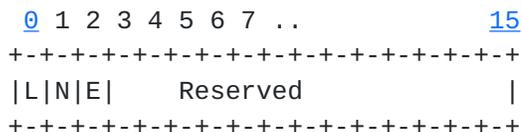
   L: Loose Bit. Indicates the type of next "Topological PDE-ID" in
   the path description. This bit MUST be set for only Node/Prefix
   PDE type. If this flag is unset, the next Topological PDE is
   Strict Type.

   N: Node Bit. By default this bit MUST be unset. This bit MUST
   be set only for PPR-PDE Type is 1 i.e., Topological and this PDE
   represents the node, where PPR-Prefix (Section 3.1) belongs to
   (if there is no further PDE specific Sub-TLVs to override PPR-
   Prefix and PPR-ID values).

   E: Egress Bit. By default this bit MUST be unset. This bit MUST
   be set only for PPR-PDE Type is 2 i.e., Non-Topological and the
   service needs to be applied on the egress side of the
   topological PDE preceding this PDE.

   Reserved: Reserved bits for future use. Reserved bits MUST be
   reset on transmission and ignored on receive.

o Sub-TLV Length: 1 byte length of all Sub-TLVs followed. It MUST
   be set to 0 if no further Sub-TLVs are present.

o PPR-PDE Sub-TLVs: These have 1 octet type, 1 octet length and
   value field is defined per the type field. Types are as defined
   in PPR-TLV Sub-TLVs (Section 7), encoded further as sub-sub-TLVs
   of PPR-PDE and the length field represents the total length of the
   value field in bytes.

   IS-IS System-ID Sub-TLV: Type 4 (IANA TBD), Length Total length
   of value field in bytes, Value: IS-IS System-ID of length "ID
   Length" as defined in [ISO.10589.1992]. This Sub-TLV MUST NOT

be present, if the PPR-PDE Type is not Topological. Though the
type for this come from the PPR Sub-TLV registry, here this is a
sub-sub-TLV and is part of PPR-ID/PPR-PDE Sub-TLV.

3.4. PPR-Attributes Sub-TLV

PPR-Attribute Sub-TLVs describe the attributes of the path. This
document defines the following optional PPR-Attribute Sub-TLVs:

o Type 5 (Suggested Value - IANA TBD): PPR-Prefix originating node's
   IPv4 Router ID Sub-TLV. Length and Value field are as specified
   in [RFC7794].

o Type 6 (Suggested Value - IANA TBD): PPR-Prefix originating node's
   IPv6 Router ID Sub-TLV. Length and Value field are as specified
   in [RFC7794].

o Type 7 (Suggested Value - IANA TBD): PPR-Metric Sub-TLV. Length 4
   bytes, and Value is metric of this path represented through the
   PPR-ID. Different nodes can advertise the same PPR-ID for the
   same Prefix with a different set of PPR-PDE Sub-TLVs and the
   receiving node MUST consider the lowest metric value.

4. PPR Processing Procedure Example

As specified in Section 2, a PPR can be a TE path, locally
provisioned by the operator or by a controller. Consider the
following IS-IS network to describe the operation of PPR TLV as
defined in Section 3:

```
                                  1
                               _____
                              /   1   \
                         +---R2-------R3---+
                        /      _____/      \
                       /           1           \
                  1 /                             \ 1
                   /         1__R13__1             \
                  /         /        \              \
              R1------R6          R7-----------R4
               \ 2       \__R14__/        2    /\
                \           2        2        /  \
               3 \                          / 3  \1
                  \                    4   /       \
                +----R8------R9-----R10------R12
                            \           1   /
                          1 \             / 1
                            +----R11---+
```

                    Figure 5: IS-IS Network

   In the (Figure 5), consider node R1 as an ingress node, or a head-end
   node, and the node R4 may be an egress node or another head-end node.
   The numbers shown on links between nodes indicate the bi-directional
   IS-IS metric as provisioned. R1 may be configured to receive TE
   source routed path information from a central entity (PCE [RFC5440],
   Netconf [RFC6241] or a Controller) that comprises of PPR information
   which relates to sources that are attached to R1. It is also
   possible to have a PPR provisioned locally by the operator for non-TE
   needs (e.g. FRR or for chaining certain services).

   The PPR TLV (as specified in Section 3) is encoded as an ordered list
   of PPR-PDEs from source to a destination node in the network and is
   represented with a PPR-ID (Section 3.2). The PPR TLV includes PPR-
   PDE Sub-TLVs Section 3.3, which represent both topological and non-
   topological elements and specifies the actual path towards a PPR-
   Prefix at R4.

   o The shortest path towards R4 from R1 are through the following
      sequence of nodes: R1-R2-R3-R4 based on the provisioned metrics.

   o The central entity can define a few PPRs from R1 to R4 that
      deviate from the shortest path based on other network
      characteristic requirements as requested by an application or
      service. For example, the network characteristics or performance
      requirements may include bandwidth, jitter, latency, throughput,
      error rate, etc.

o A first PPR may be identified by PPR-ID = 1 (value) and may
   include the path of R1-R6-R7-R4 for a Prefix advertised by R4.
   This is an example for a Loose-PPR and 'L' bit MUST be set
   appropriately at Section 3.3.

o A second PPR may be identified by PPR-ID = 2 (value) and may
   include the path of R1-R8-R9-R10-R4. This is an example for a
   Strict-PPR and 'L' bit MUST be unset appropriately at Section 3.3.
   Though this example shows PPR with all nodal SIDs, it is possible
   to have a PPR with combination of node and adjacency SIDs (local
   or global) or with PPR-PDE Type set to Non-Topological as defined
   in Section 3.3 elements along with these.

4.1. PPR TLV Processing

The first topological sub-object or PDE (Section 3.3) relative to the
beginning of PPR Path contains the information about the first node
(e.g. in SR-MPLS it's the topmost label). The last topological sub-
object or PDE contains information about the last node (e.g. in SR-
MPLS it's the bottommost label).

Each receiving node, determines whether an advertised PPR includes
information regarding the receiving node. Before processing any
further, validation MUST be done to see if any PPR topological PDE is
seen more than once (possible loop), if yes, this PPR TLV MUST be
ignored. Processing of PPR TLVs may be done, during the end of the
SPF computation (for MTID that is advertised in this TLV) and for
each prefix described through PPR TLV. For example, node R9 receives
the PPR information, and ignores the PPR-ID=1 (Section 4) because
this PPR TLV does not include node R9 in the path description/ordered
PPR-PDE list.

However, node R9 may determine that the second PPR identified by PPR-
ID = 2 does include the node R9 in its PDE list. Therefore, node R9
updates the local forwarding database to include an entry for the
destination address that R4 indicates, so that when a data packet
comprising a PPR-ID of 2 is received, forward the data packet to node
R10 instead of R11. This is done, even though from R9 the shortest
path to reach R4 via R11 (Cost 3: R9-R11-R12-R4) it chooses the NH to
R10 to reach R4 as specified in the PPR path description. Same
process happens to all nodes or all topological PDEs as described in
the PPR TLV.

In summary, the receiving node checks first, if this node is on the
path by checking the node's topological elements (with PPR-PDE Type
set to Topological) in the path list. If yes, it adds/adjusts the
PPR-ID's shortest path NH towards the next topological PDE in the
PPR's Path.

4.2. Path Fragments

   A complete PPR path may not fit into maximum allowable size of the
   IS-IS TLV. To overcome this a 7 bit Fragment-ID field is defined in
   Section 3 . With this, a single PPR path is represented via one or
   more fragmented PPR path TLVs, with all having the same PPR-ID. Each
   fragment carries the PPR-ID as well as a numeric Fragment-ID from 0
   to (N-1), when N fragments are needed to describe the PPR Graph
   (where N>1). In this case Fragment (N-1) MUST set the 'U' bit (PPR-
   Flags) to indicate it is the last fragment. If Fragment-ID is non
   zero in the TLV, then it MUST not carry PPR-Prefix Sub-TLV. The
   optional PPR Attribute Sub-TLVs which describe the path overall MUST
   be included in the last fragment only (i.e., when the 'U' bit is
   set).

5. PPR Data Plane aspects

   Data plane for PPR-ID is selected by the entity (e.g., a controller,
   locally provisioned by operator), which selects a particular PPR in
   the network. Section 3.2 defines various data plane identifier types
   and a corresponding data plane identifier is selected by the entity
   which selects the PPR.

5.1. SR-MPLS with PPR

   If PPR-ID Type is 1, then the PPR belongs to SR-MPLS data plane and
   the complete PPR stack is represented with a unique SR SID/Label and
   this gets programmed on the data plane of each node, with the
   appropriate NH computed as specified in Section 4. PPR-ID here is a
   label/index from the SRGB (like another node SID or global ADJ-SID).
   PPR path description here is a set of ordered SIDs represented with
   PPR-PDE (Section 3.2) Sub-TLVs. Non-Topological segments also
   programmed in the forwarding to enable specific function/service,
   when the data packet hits with corresponding PPR-ID.

   Based on 'L' flag in PPR-ID Flags (Section 3.2), for SR-MPLS data
   plane either 1 label or 2 labels need to be provisioned on individual
   nodes on the path description. For the example network in Section 4,
   for PPR-ID=1, which is a loose path, node R6 programs the bottom
   label as PPR-ID and the top label as the next topological PPR-PDE in
   the path, which is a node SID of R7. The NH computed at R6 would be
   the shortest path towards R7 i.e., the interface towards R13. If 'L'
   flag is unset only PPR-ID is programmed on the data plane with NH set
   to the shortest path towards next topological PPR-PDE.

5.2. PPR Native IP Data Planes

   If PPR-ID Type is 2 then source routing and packet steering can be
   done in IPv4 data plane (PPR-IPv4), along the path as described in
   PPR Path description. This is achieved by setting the destination IP
   address as PPR-ID, which is an IPv4 address in the data packet
   (tunneled/encapsulated). There is no data plane change or upgrade
   needed to support this.

   Similarly for PPR-ID Type is 3, then source routing and packet
   steering can be done in IPv6 data plane (PPR-IPv6), along the path as
   described in PPR Path description. Whatever specified above for IPv4
   applies here too, except that destination IP address of the data
   packet is an IPv6 Address (PPR-ID). This doesn't require any IPv6
   extension headers (EH), if there is no metadata/TLVs need to be
   carried in the data packet.

   Based on 'L' flag in PPR-ID Flags (Section 3.2), for PPR-ID Type 2 or
   3 (Native IPv4 or IPv6 data planes respectively) the packet has to be
   encapsulated using the capabilities (either dynamically signaled
   through [I-D.ietf-isis-encapsulation-cap] or statically provisioned
   on the nodes) of the next loose PDE in the path description.

   For the example network in Section 4, for PPR-ID=1, which is a loose
   path, node R6 programs to encapsulate the data packet towards the
   next loose topological PPR-PDE in the path, which is R7. The NH
   computed at R6 would be the shortest path towards R7 i.e., the
   interface towards R13. If 'L' flag is unset only PPR-ID is
   programmed on the data plane with NH set to the shortest path towards
   next topological PPR-PDE, with no further encapsulation of the data
   packet.

5.3. SRv6 with PPR

   If PPR-ID Type is 4, the PPR belongs to SRv6 with SRH data plane and
   the complete PPR stack is represented with IPv6 SIDs and this gets
   programmed on the data plane with the appropriate NH computed as
   specified in Section 4. PPR-ID here is a SRv6 SID. PPR path
   description here is a set of ordered SID TLVs similar to as specified
   in Section 5.1. One way PPR-ID would be used in this case is by
   setting it as the destination IPv6 address and SL field in SRH would
   be set to 0; however SRH [I-D.ietf-6man-segment-routing-header] can
   contain any other TLVs and non-topological SIDs as needed.

6. Acknowledgements

   Thanks to Alex Clemm, Lin Han, Toerless Eckert, Asit Chakraborti,
   Stewart Bryant and Kiran Makhijani for initial discussions on this
   topic. Thanks to Kevin Smith and Stephen Johnson for various
   deployment scenarios applicability from ETSI WGs perspective.
   Authors also acknowledge Alexander Vainshtein for detailed
   discussions and few suggestions on this topic.

   Earlier versions of draft-ietf-isis-segment-routing-extensions have a
   mechanism to advertise EROs through Binding SID.

7. IANA Considerations

   This document requests the following new TLV in IANA IS-IS TLV code-
   point registry.

         TLV #   Name
         -----   --------------
         155   PPR TLV (Suggested Value, IANA TBD)

7.1. PPR Sub-TLVs

   This document requests IANA to create a new Sub-TLV registry for PPR
   TLV Section 3 with the following initial entries (suggested values).
   Though these are defined as Sub-TLVs of PPR TLV, these can be part of
   another Sub-TLV as a nested sub-sub-TLV (e.g. IS-IS System-ID).

   Sub-TLV #   Sub-TLV Name
   ---------   --

    1   PPR-Prefix (Section 3.1)

    2   PPR-ID (Section 3.2)

    3   PPR-PDE (Section 3.3)

    4   IS-IS System-ID (Section 3.3)

    5   PPR-Prefix Source IPv4 Router ID (Section 3.4)

    6   PPR-Prefix Source IPv6 Router ID (Section 3.4)

    7   PPR-Metric (Section 3.4)

7.2. IGP Parameters

   This document requests additional IANA registries in an IANA managed
   registry "Interior Gateway Protocol (IGP) Parameters" for various PPR
   TLV parameters. The registration procedure is based on the "Expert
   Review" as defined in [RFC8126]. The suggested registry names are:

   o "PPR-Type" - Types are an unsigned 8 bit numbers. Values are as
      defined in Section 3 of this document.

   o "PPR-Flags" - 1 Octet. Bits as described in Section 3 of this
      document.

   o "PPR-ID Type" - Types are an unsigned 8 bit numbers. Values are
      as defined in Section 3.2 of this document.

   o "PPR-ID Flags" - 1 Octet. Bits as described in Section 3.2 of
      this document.

   o "PPR-PDE Type" - Types are an unsigned 8 bit numbers. Values are
      as defined in Section 3.3 of this document.

   o "PPR-PDE Flags" - 1 Octet. Bits as described in Section 3.3 of
      this document.

   o "PDE-ID Type" - Types are an unsigned 8 bit numbers. Values are
      as defined in Section 3.3 of this document.

8. Security Considerations

   Security concerns for IS-IS are addressed in [RFC5304] and [RFC5310].
   Further security analysis for IS-IS protocol is done in [RFC7645]
   with detailed analysis of various security threats and why [RFC5304]
   should not be used in the deployments. Advertisement of the
   additional information defined in this document introduces no new
   security concerns in IS-IS protocol. However, for extensions related
   ro SR-MPLS and SRH data planes, those particular data plane security
   considerations does apply here.

9. References

9.1. Normative References

[ISO.10589.1992]
          International Organization for Standardization,
          "Intermediate system to intermediate system intra-domain-
          routing routine information exchange protocol for use in
          conjunction with the protocol for providing the
          connectionless-mode Network Service (ISO 8473)",
          ISO Standard 10589, 1992.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate
           Requirement Levels", BCP 14, RFC 2119,
           DOI 10.17487/RFC2119, March 1997,
           <https://www.rfc-editor.org/info/rfc2119>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC
           2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174,
           May 2017, <https://www.rfc-editor.org/info/rfc8174>.

9.2. Informative References

[I-D.bashandy-isis-srv6-extensions]
          Psenak, P., Filsfils, C., Bashandy, A., Decraene, B., and
          Z. Hu, "IS-IS Extensions to Support Routing over IPv6
          Dataplane", draft-bashandy-isis-srv6-extensions-05 (work
          in progress), March 2019.

[I-D.ietf-6man-segment-routing-header]
          Filsfils, C., Dukes, D., Previdi, S., Leddy, J.,
          Matsushima, S., and D. Voyer, "IPv6 Segment Routing Header
          (SRH)", draft-ietf-6man-segment-routing-header-26 (work in
          progress), October 2019.

[I-D.ietf-isis-encapsulation-cap]
          Xu, X., Decraene, B., Raszuk, R., Chunduri, U., Contreras,
          L., and L. Jalil, "Advertising Tunnelling Capability in
          IS-IS", draft-ietf-isis-encapsulation-cap-01 (work in
          progress), April 2017.

[I-D.ietf-isis-mpls-elc]
          Xu, X., Kini, S., Psenak, P., Filsfils, C., Litkowski, S.,
          and M. Bocci, "Signaling Entropy Label Capability and
          Entropy Readable Label Depth Using IS-IS", draft-ietf-
          isis-mpls-elc-13 (work in progress), May 2020.

[I-D.ietf-isis-segment-routing-extensions]
          Previdi, S., Ginsberg, L., Filsfils, C., Bashandy, A.,
          Gredler, H., and B. Decraene, "IS-IS Extensions for
          Segment Routing", draft-ietf-isis-segment-routing-
          extensions-25 (work in progress), May 2019.

   [I-D.ietf-mpls-sfc]
              Farrel, A., Bryant, S., and J. Drake, "An MPLS-Based
              Forwarding Plane for Service Function Chaining", draft-
              ietf-mpls-sfc-07 (work in progress), March 2019.

   [I-D.xuclad-spring-sr-service-chaining]
              Clad, F., Xu, X., Filsfils, C., daniel.bernier@bell.ca,
              d., Li, C., Decraene, B., Ma, S., Yadlapalli, C.,
              Henderickx, W., and S. Salsano, "Segment Routing for
              Service Chaining", draft-xuclad-spring-sr-service-
              chaining-01 (work in progress), March 2018.

   [RFC5120] Przygienda, T., Shen, N., and N. Sheth, "M-ISIS: Multi
              Topology (MT) Routing in Intermediate System to
              Intermediate Systems (IS-ISs)", RFC 5120,
              DOI 10.17487/RFC5120, February 2008,
              <https://www.rfc-editor.org/info/rfc5120>.

   [RFC5304] Li, T. and R. Atkinson, "IS-IS Cryptographic
              Authentication", RFC 5304, DOI 10.17487/RFC5304, October
              2008, <https://www.rfc-editor.org/info/rfc5304>.

   [RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic
              Engineering", RFC 5305, DOI 10.17487/RFC5305, October
              2008, <https://www.rfc-editor.org/info/rfc5305>.

   [RFC5308] Hopps, C., "Routing IPv6 with IS-IS", RFC 5308,
              DOI 10.17487/RFC5308, October 2008,
              <https://www.rfc-editor.org/info/rfc5308>.

   [RFC5310] Bhatia, M., Manral, V., Li, T., Atkinson, R., White, R.,
              and M. Fanto, "IS-IS Generic Cryptographic
              Authentication", RFC 5310, DOI 10.17487/RFC5310, February
              2009, <https://www.rfc-editor.org/info/rfc5310>.

   [RFC5440] Vasseur, JP., Ed. and JL. Le Roux, Ed., "Path Computation
              Element (PCE) Communication Protocol (PCEP)", RFC 5440,
              DOI 10.17487/RFC5440, March 2009,
              <https://www.rfc-editor.org/info/rfc5440>.

   [RFC6241] Enns, R., Ed., Bjorklund, M., Ed., Schoenwaelder, J., Ed.,
              and A. Bierman, Ed., "Network Configuration Protocol
              (NETCONF)", RFC 6241, DOI 10.17487/RFC6241, June 2011,
              <https://www.rfc-editor.org/info/rfc6241>.

   [RFC6790] Kompella, K., Drake, J., Amante, S., Henderickx, W., and
              L. Yong, "The Use of Entropy Labels in MPLS Forwarding",
              RFC 6790, DOI 10.17487/RFC6790, November 2012,
              <https://www.rfc-editor.org/info/rfc6790>.

   [RFC7645] Chunduri, U., Tian, A., and W. Lu, "The Keying and
              Authentication for Routing Protocol (KARP) IS-IS Security
              Analysis", RFC 7645, DOI 10.17487/RFC7645, September 2015,
              <https://www.rfc-editor.org/info/rfc7645>.

   [RFC7794] Ginsberg, L., Ed., Decraene, B., Previdi, S., Xu, X., and
              U. Chunduri, "IS-IS Prefix Attributes for Extended IPv4
              and IPv6 Reachability", RFC 7794, DOI 10.17487/RFC7794,
              March 2016, <https://www.rfc-editor.org/info/rfc7794>.

   [RFC8126] Cotton, M., Leiba, B., and T. Narten, "Guidelines for
              Writing an IANA Considerations Section in RFCs", BCP 26,
              RFC 8126, DOI 10.17487/RFC8126, June 2017,
              <https://www.rfc-editor.org/info/rfc8126>.

   [RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L.,
              Decraene, B., Litkowski, S., and R. Shakir, "Segment
              Routing Architecture", RFC 8402, DOI 10.17487/RFC8402,
              July 2018, <https://www.rfc-editor.org/info/rfc8402>.

   [RFC8491] Tantsura, J., Chunduri, U., Aldrin, S., and L. Ginsberg,
              "Signaling Maximum SID Depth (MSD) Using IS-IS", RFC 8491,
              DOI 10.17487/RFC8491, November 2018,
              <https://www.rfc-editor.org/info/rfc8491>.

Appendix A. Appendix

A.1. Challenges with Increased SID Depth

   SR label stacks carried in the packet header create challenges in the
   design and deployment of networks and networking equipment.
   Following examples illustrates the need for increased SID depth in
   various use cases:

   (a). Consider the following network where SR-MPLS data plane is in
   use and with same SRGB (5000-6000) on all nodes i.e., A1 to A11 and
   B1 to B7 for illustration:
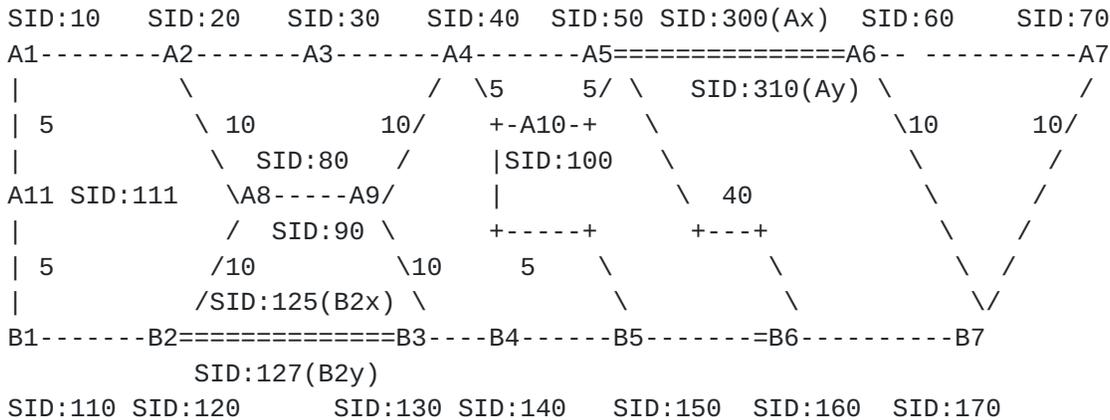
```
 SID:10   SID:20   SID:30   SID:40  SID:50 SID:300(Ax)  SID:60     SID:70
 A1--------A2-------A3-------A4-------A5==============A6-- ----------A7
 |          \                 /  \5     5/ \    SID:310(Ay) \          /
 | 5         \ 10         10/    +-A10-+   \                \10      10/
 |            \  SID:80   /     |SID:100    \                \       /
 A11 SID:111   \A8-----A9/      |            \   40          \     /
 |            /  SID:90 \      +-----+      +---+            \    /
 | 5        /10         \10     5    \          \            \  /
 |         /SID:125(B2x) \            \          \            \/
 B1-------B2==============B3----B4------B5-------=B6----------B7
           SID:127(B2y)
 SID:110 SID:120      SID:130 SID:140   SID:150  SID:160  SID:170

   === = Path with Parallel Adjacencies and ADJ-SIDs
   --- = Shortest Path Nodal SID
```

   Figure 6: SR-MPLS Network

     Global ADJ-SIDs are provisioned between A5-A6 and B2-B3 (with
     parallel adjecencies). All other SIDs shown are nodal SID
     indices.

     All metrics of the links are set to 1, unless marked otherwise.

     Shortest Path from A1 to A7: A2-A3-A4-A5-A6-A7

     Path-x: From A1 to A7 - A2-A8-B2-B2x-A9-A10-Ax-A7; Pushed Label
     Stack @A1: 5020:5080:5120:5125:5090:5100:5300:5070 (where B2x is a
     local ADJ-SID and Ax is a global ADJ-SID).

     In this example, the traffic engineered path is represented with a
     combination of Adjacency and Node SIDs with a stack of 8 labels.
     However, this value can be larger, if the use of entropy label
     [RFC6790] is desired and based on the Readable Label Depth
     (Appendix A.2) capabilities of each node and additional labels
     required to insert ELI/EL at appropriate places.

     Though above network is shown with SR-MPLS data plane, if the
     network were to use SRv6 data plane, path size would be increased
     even more because of the size of the IPv6 SID (16 bytes) in SRH.

   (b). Apart from the TE case above, when deploying
   [I-D.ietf-mpls-sfc] or [I-D.xuclad-spring-sr-service-chaining], with
   the inclusion of services, or non-topological segments on the label
   stack, can also make the size of the stack much larger.

Overall the additional path overhead in various SR deployments may
cause the following issues:

a. HW Capabilities: Not all nodes in the path can support the
    ability to push or read label stack (with additional non-
    topological and special labels) needed to satisfy user/operator
    requirements. Alternate paths, which meet these user/operator
    requirements may not be available.

b. Line Rate: Potential performance issues in deployments, which use
    data plane with extension header as both size of the SIDs in the
    extension header and the fixed extension header size itself needs
    to be factored by the hardware.

c. MTU: Larger SID stacks on the data packet can cause potential
    MTU/fragmentation issues (SRH).

d. Header Tax: Some deployments, such as 5G, require minimal packet
    overhead in order to conserve network resources. Carrying 40 or
    50 octets of data in a packet with hundreds of octet of header
    would be an unacceptable use of available bandwidth.

With the solution proposed in this document (Section 2), for Path-x
in Figure 6 above, SID stack would be reduced from 8 SIDs to a single
SID witout any additional overhead.

A.2. Mitigation with MSD

The number of SIDs in the stack a node can impose is referred as
Maximum SID Depth (MSD) capability [RFC8491], which must be taken
into consideration when computing a path to transport a data packet
in a network implementing segment routing. [I-D.ietf-isis-mpls-elc]
defines another MSD type, Readable Label Depth (RLD) that is used by
a head-end to insert Entropy Label pair (ELI/EL) at appropriate
depth, so it could be read by transit nodes. There are situations
where the source routed path can be excessive as path represented by
SR SIDs need to describe all the nodes and ELI/EL based on the
readability of the nodes in that path. Registries setforth in
[RFC8491] applicable for MPLS data plane and for IPv6 data plane with
SRH.

MSDs (and RLD type) capabilities advertisement help mitigate the
problem for a central entity to create the right source routed path
per application/operator requirements. However the availability of
actual paths meeting these requirements are still limited by the
underlying hardware and their MSD capabilities in the data path.

Authors' Addresses

    Uma Chunduri
    Futurewei
    2330 Central Expressway
    Santa Clara, CA 95050
    USA

    Email: umac.ietf@gmail.com

    Richard Li
    Futurewei
    2330 Central Expressway
    Santa Clara, CA 95050
    USA

    Email: richard.li@futurewei.com

    Russ White
    Juniper Networks
    Oak Island, NC 28465
    USA

    Email: russ@riw.us

    Jeff Tantsura
    Apstra Inc.
    333 Middlefield Road
    Menlo Park, CA 94025
    USA

    Email: jefftant.ietf@gmail.com

    Luis M. Contreras
    Telefonica
    Sur-3 building, 3rd floor
    Madrid 28050
    Spain

    Email: luismiguel.contrerasmurillo@telefonica.com

Yingzhen Qu
Futurewei
2330 Central Expressway
Santa Clara, CA 95050
USA

Email: yingzhen.qu@futurewei.com