Link State Over Ether

Randy Bush Arrcus & IIJ Lab

Keyur Patel, Arrcus

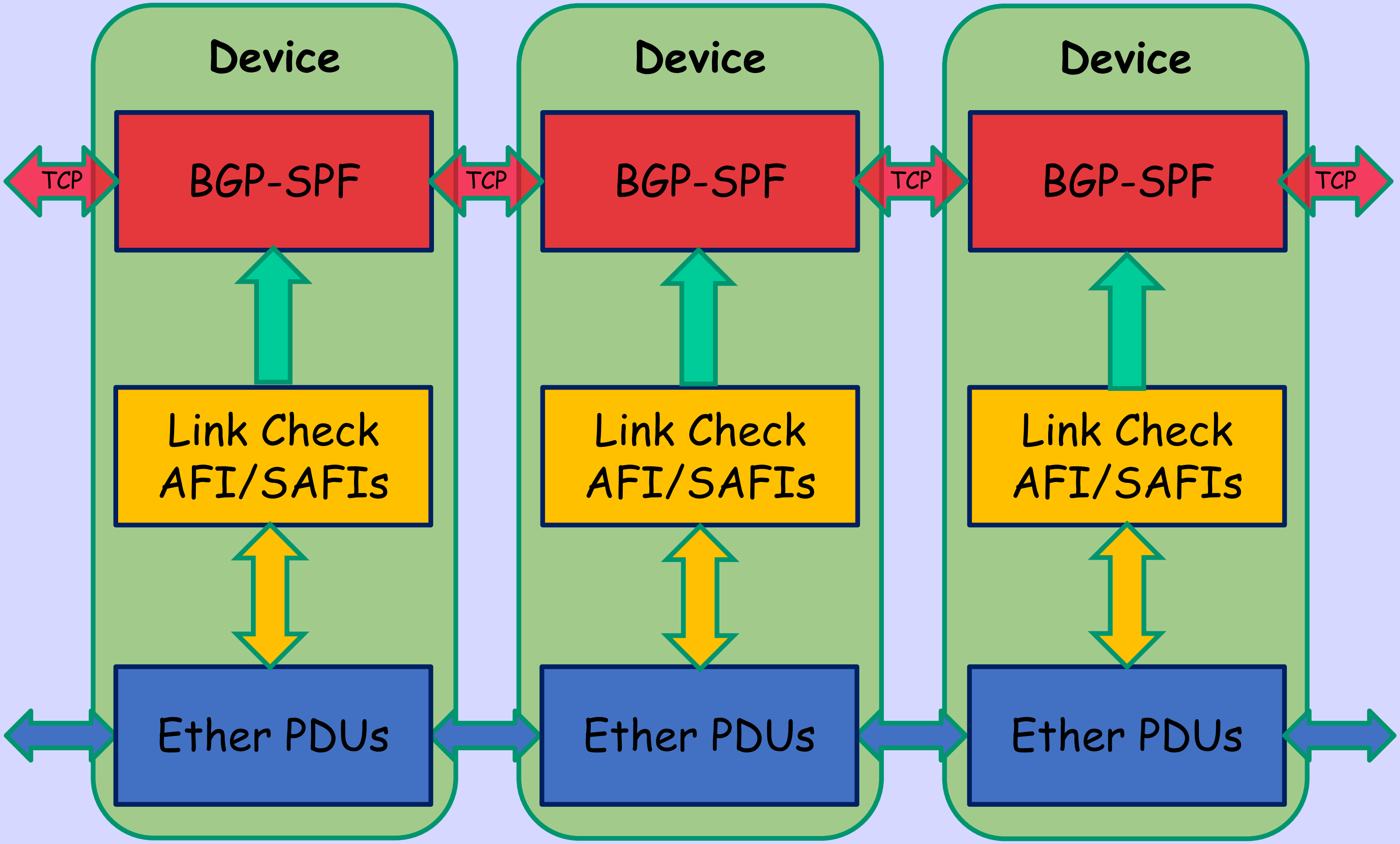
2018.10.01 LSVR Interim

How Does BGP-SPF Learn Link State?

Motivation

- BGP-SPF needs link neighbor discovery, liveness, and addressability
- LLDP is an IEEE protocol, complex, and 'hard' (IPR) to extend past 1500 bytes
- We wanted something simple and saw no real need for the complexities of CLNP, …
- So we propose a new EtherType with TLVs
- We discuss Ether payloads, not framing

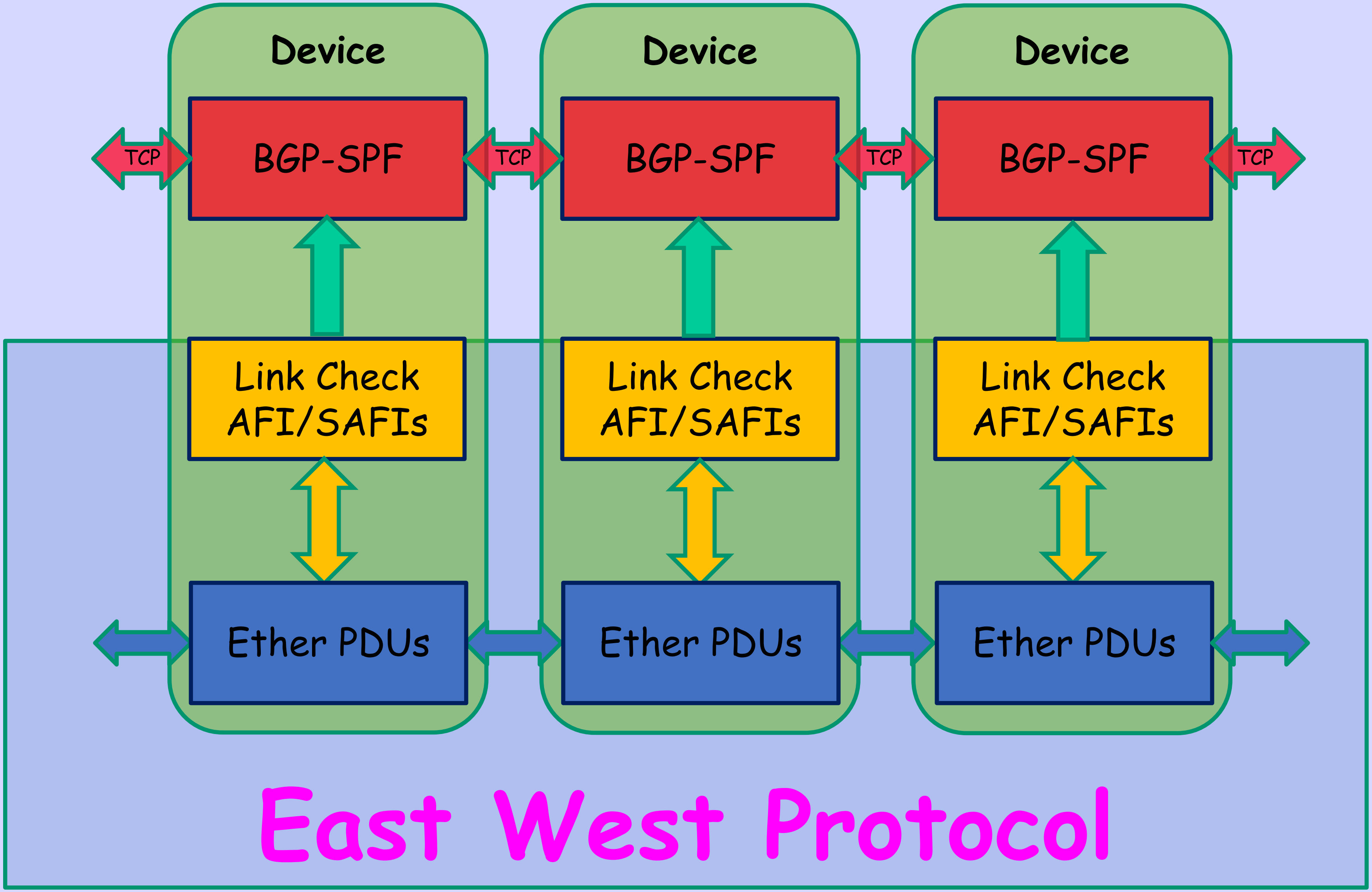
Topology / Routing Stack



MAC Link State exchanged over raw Ethernet and pushed up stack

Add the AFI/SAFI data IP-Level Liveness Check

BGP-SPF uses link data to discover and build the topology database



Device

BGP-SPF

TCP

Link Check
AFI/SAFIs

Ether PDUs

Device

BGP-SPF

TCP

Link Check
AFI/SAFIs

Ether PDUs

Device

BGP-SPF

TCP

Link Check
AFI/SAFIs

Ether PDUs

TCP

East West Protocol

PDUs and Frames

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    Version    |      Type     |           PDU Length          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            Checksum                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   PDU Number  |     Flags     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

- This is all about inter-device Link State
- A PDU is one Ethernet Frame
- A Frame has *PDU Number* and *Flags* to allow assembling Messages needing more than one PDU
- Flags:
    - Bit 0 – One of a Multi-PDU Message
    - Bit 1 – Last of a Multi-PDU Message

Every Frame a TLV

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    Version    |      Type     |           PDU Length          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            Checksum                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   PDU Number  |     Flags     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

```
Version / Type  - Version = 0; Type is the PDU Type

Length          - Total Bytes in PDU including all fields

Checksum        - one’s complement over Frame, detect bit flips

PDU Sequence No – Monotonically increasing, wraps around

Flags (bits)    - 0 – one of a multi-Frame sequence
                - 1 – last of a multi-Frame sequence
```

Checksum

- There is a reason conservative folk use a checksum in UDP
- And when the op stretches to jumbo frames ...
- One's complement is a bit silly, though trivial to implement
- Sum up either 16-bit shorts in a 32-bit int, or 32-bit ints in a 64-bit long, then take the high-order section, shift it right, rotate, add it in, repeat until zero. -- smb off the top of his head

Inter-Link Ether Protocol

```
|             HELLO               |
|<--------------------------->| Mandatory
|                             |
|             OPEN            | MACs, IDs
|<--------------------------->| Mandatory
|             OPEN            |
|                             |
|   Interface IPv4 Addresses  | Interface IPv4 Addresses
|<--------------------------->| Optional
|                             |
|   Interface IPv6 Addresses  | Interface IP4 Addresses
|<--------------------------->| Optional
|                             |
|   Interface MPLSv4 Labels   | MPLS IPv4 Addresses
|<--------------------------->| Optional
|                             |
|   Interface MPLSv6 Labels   | MPLS IPv6 Addresses
|<--------------------------->| Optional
|                             |
|      Layer 2 KeepAlives     | Layer 2 Liveness
|<--------------------------->|
```

Link HELLO

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Version = 0  |    Type = 0   |        PDU Length = 14        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            Checksum                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                         MyMAC Address                         |
+                               +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

- HELLO is Multicast, à la LLDP

- Each device learns the other's MAC from its HELLO whining. All devices on a wire/interface know each others MACs and learn each other’s IDs

- Respond with OPEN

- A multi-point topology is a set of point-to-point links

OPEN

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    Version    |    Type = 1   |           PDU Length          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            Checksum                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   PDU Number  |     Flags     |                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+                               +
|                                                               |
+                            Local ID                           +
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
+                                                               +
|                      Remote ID (or Zero)                      |
+                               +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                               |   AttrCount   |               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|       Attribute List ...      |          Auth Length          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    Authentication Data ...                    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Local/Remote IDs

Might be

- an ASN with high order bits zero
- a classic RouterID with high order bits zero
- a catenation of the two
- a 80-bit ISO System-ID
- or any other identifier unique to a single device in the current routing space

Attributes

A node may have zero or more user-defined attributes, e.g. spine, leaf, backbone, route reflector, arabica, ...

Nodes exchange their attributes only in the OPEN message

Authentication Data

- Specific to the Operational Environment
- Might be Certificate derived from Op's CA
- Failure to authenticate is a failure to start the LSOE association, and HELLOs MUST BE restarted.

Once We Know
Each Other's MACs

Layer Two KeepAlives
May be Started

L2 KEEPALIVE

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    Version    |    Type = 2   |           Length = 8          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            Checksum                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

This is in addition to L3 BFD etc.

We assume that one or more Encapsulation addresses will be used to ping, BFD, or whatever the operator configures

We Know MAC/Ether Link State of This Device & Neighbor

And Node IDs (often ASNs)

Now Announce Encapsulations of the Link Interfaces

Encaps PDU Header

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    Version    |      Type     |           PDU Length          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            Checksum                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   PDU Number  |     Flags     |      Encapsulation Count      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Encapsulation List...                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

The Encapsulation Exchange

Is Over an Unreliable Transport

So There Are

Sequence Numbers and ACKs

Encapsulation PDU ACK

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    Version    |    Type = 3   |           Length = 11         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            Checksum                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   PDU Number  |     Flags     |   Encap Type  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

- The PDU Number is a p2p link Announcement Counter
- The Receiver will ACK it with a Type=3
- If the Sender does not receive an ACK in one second, they retransmit. Operator configured failure count

IPv4 Encapsulations

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    Version    |    Type = 4   |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            Checksum                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   PDU Number  |     Flags     |          Encaps Count         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| PrimLoop Flags|                  IPv4 Address                 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|               |   PrefixLen   | PrimLoop Flags|               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                  IPv4 Address                 |   PrefixLen   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            more ...                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

An Encapsulation message describes zero or more addresses of the encapsulation type.

An Encapsulation message of Type T replaces all previous encapsulations of Type T

PrimLoop Flags

```
 0               1               2               3    ...       7
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Primary      |   Loopback    |  Reserved ...                 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

- An Interface may have multiple Encapsulations
- For each Encapsulation there might be multiple Addresses
- One Address per Encapsulation SHOULD be marked as Primary
- An Address may be marked as a loopback

IPv6 Encapsulations

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    Version    |    Type = 5   |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            Checksum                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   PDU Number  |     Flags     |          Encaps Count         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| PrimLoop Flags|                                               |
+-+-+-+-+-+-+-+-+                                               +
|                                                               |
+                                                               +
|                                                               |
+                                                               +
|                          IPv6 Address                         |
+               +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|               |   PrefixLen   |            more ...           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

MPLS IPv4 Encapsulations

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    Version    |   Type = 6    |            Length             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            Checksum                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   PDU Number  |     Flags     |          Encaps Count         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| PrimLoop Flags|              MPLS Label List ...              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|               |                  IPv4 Address                 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|               |   PrefixLen   | PrimLoop Flags|               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|              MPLS Label List ...              |               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                  IPv4 Address                 |   Prefix Len  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            more ...                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

MPLS Label List

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Label Count  |                 Label                 | Exp |S|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                 Label                 | Exp |S|    more ...   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Use Multiple MPLS Label Encapsulations to Allow One Label to be Associated with Multiple AFI/SAFIs and/or Multiple IP Addresses

MPLS IPv6 Encapsulartions

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    Version    |    Type = 7   |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            Checksum                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   PDU Number  |     Flags     |          Encaps Count         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| PrimLoop Flags|   MPLS Label  |    List ...   |               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+               +
|                                                               |
+                                                               +
|                                                               |
+                                                               +
|                          IPv6 Address                         |
+                                               +-+-+-+-+-+-+-+-+
|                                               |   Prefix Len  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            more ...                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Layer-3 IP/Label Liveness Should Also be Tested

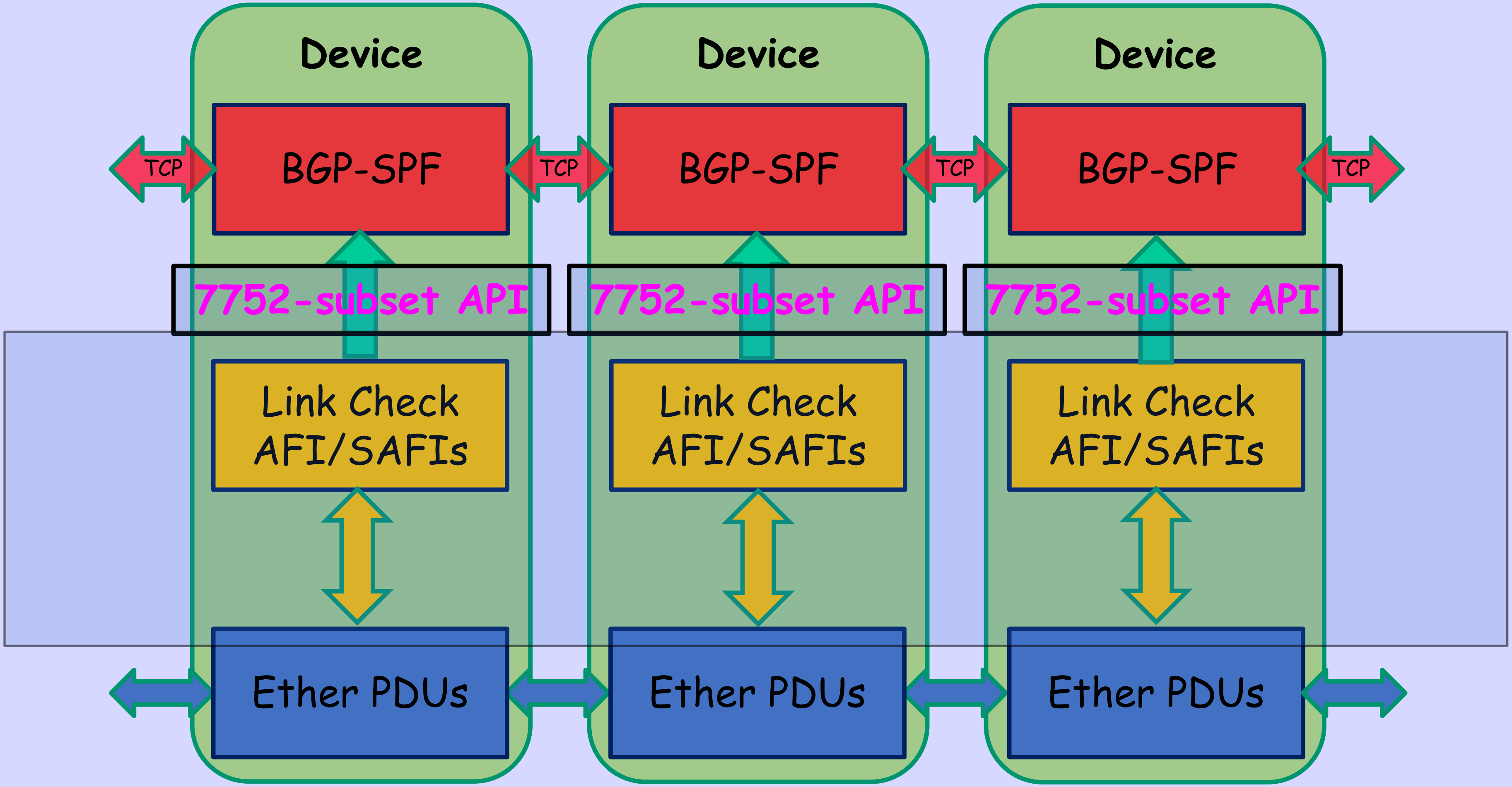
One or more Discovered AFI/SAFI Addresses Are Used to Ping, BFD, ... to Assure Layer-3 Liveness

We now know all links, IDs, Encapsulation Types, and Addresses of this Device

Now Present an API to Topology and Dijkstra Layers

BGP-LS (RFC 7752) an extension to BGP to distribute the network's link-state (LS) topology

North/South Protocol



Node Descriptors

- Similarly to BGP-SPF, the BGP protocol is used in the Protocol-ID field specified in table 1 of draft-ietf-idr-bgpls-segment-routing-epe.
- The local and remote node descriptors for all NLRI are the ID's described in Section 5.3.
- This is equivalent to an adjacency SID or a node SID if the address is a loopback address.

IPvX Links

TLVs 259 and 260 are used. And for IPv6 links, TLVs 261 and 262. If there are multiple addresses on a link, multiple TLV pairs are pushed North, having the same ID pairs.

MPLS Links

Label Sub-TLVs from draft-ietf-idr-bgp-ls-segment-routing-ext Section 2.1.1, are used to associate one or more MPLS Labels with a link.

And Bob's Your Uncle

Open Questions

Should HELLO go <u>Through</u> an intermediate Layer Two Switch

Are HELLO and KEEPALIVE Redundant?

BTW,
There is No IPR