Chair
IEEE US TAG
to
ISO/IEC JTC1
SC22/WG13

But I Have Recovered

Neighbor Discovery

Randy Bush <randy@psg.com>

2018.11.10 IETF/IEEE Bangkok

Where it Started

Entering the Rabbit Hole (an example)

IIJ is Building a Second Medium Scale Data Center (MSDC) in Shiroi/Chiba Capacity of 6k Racks

OSPF OK to 500 Nodes

IS-IS good to 1,000

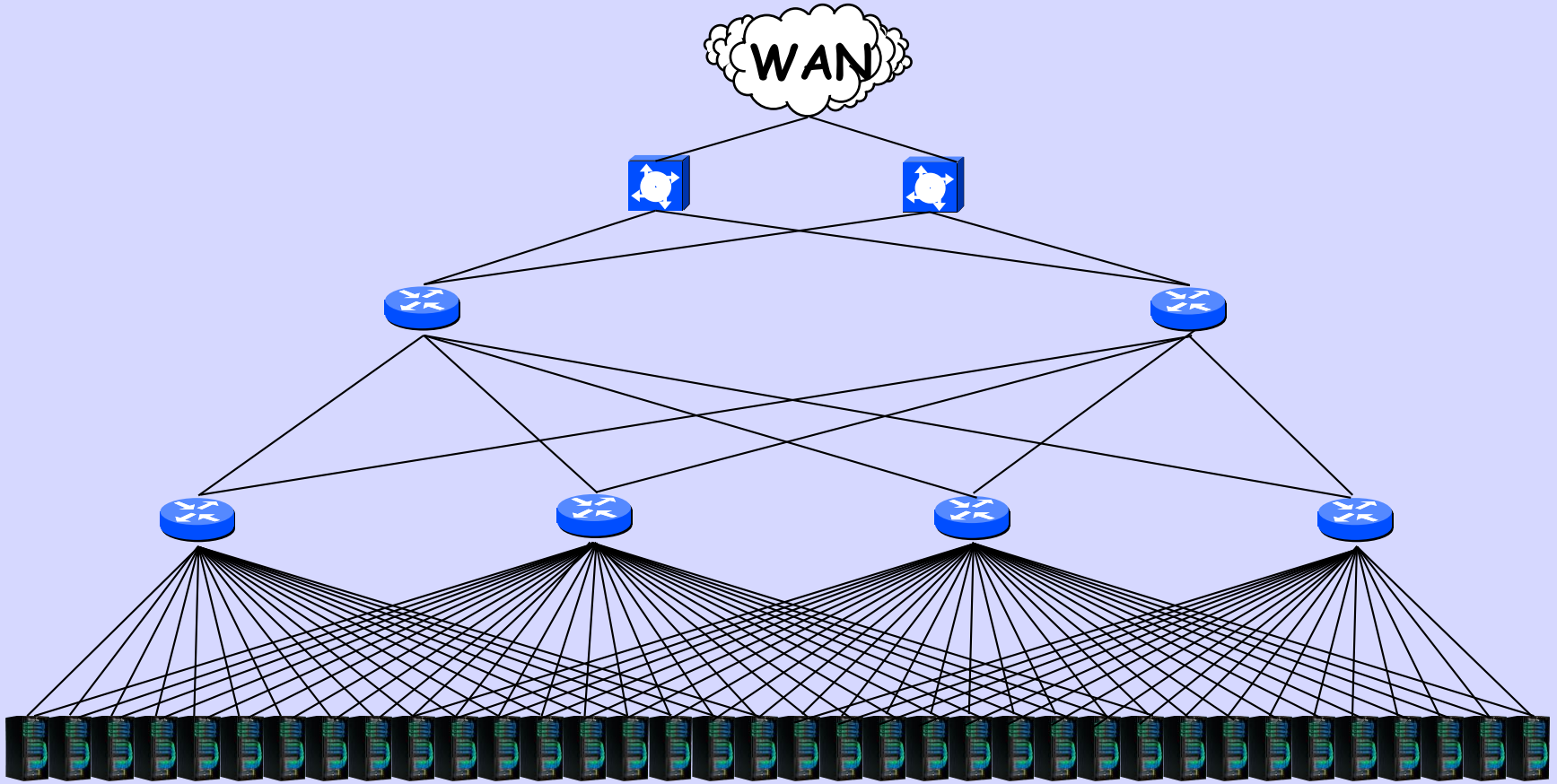
Limited Because They Repeatedly Flood Everything

Your Clos on IS-IS or OSPF

BGP Scales Because It Signals Only Changes

So BGP has become common in MSDCs

But What is the Decision Process?



Consult the Professor

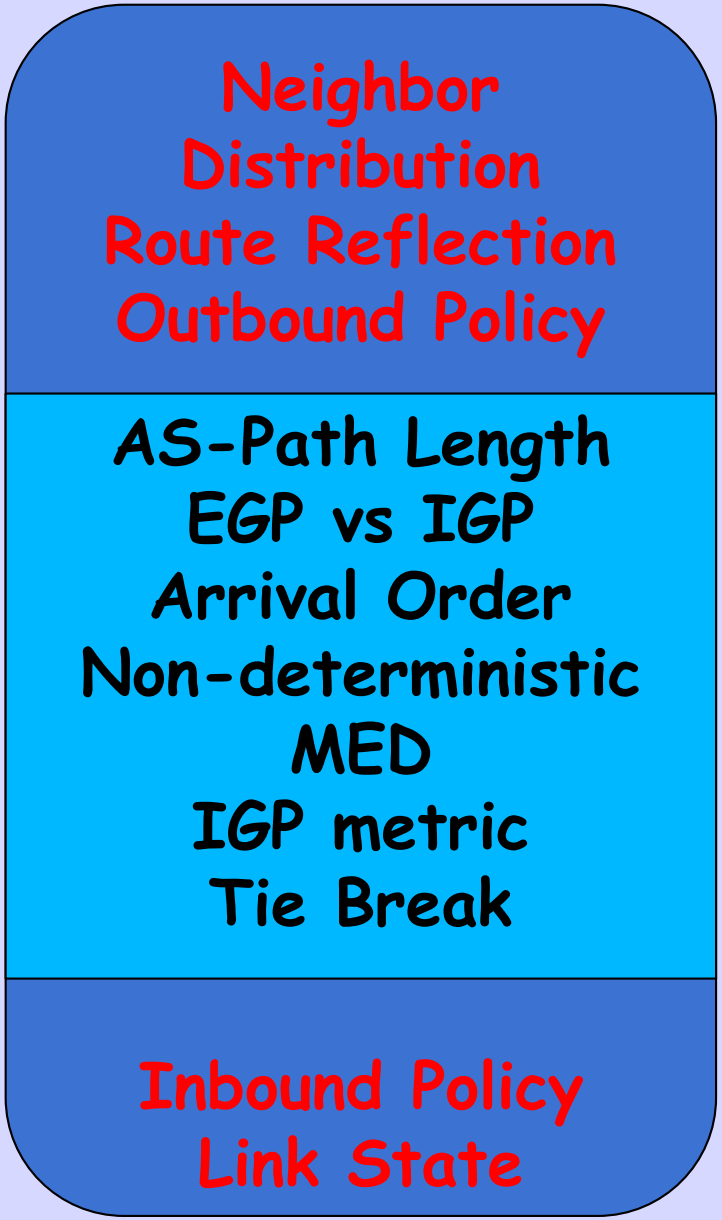
Edsger W Dijkstra
1930-2002

Shortest Path First

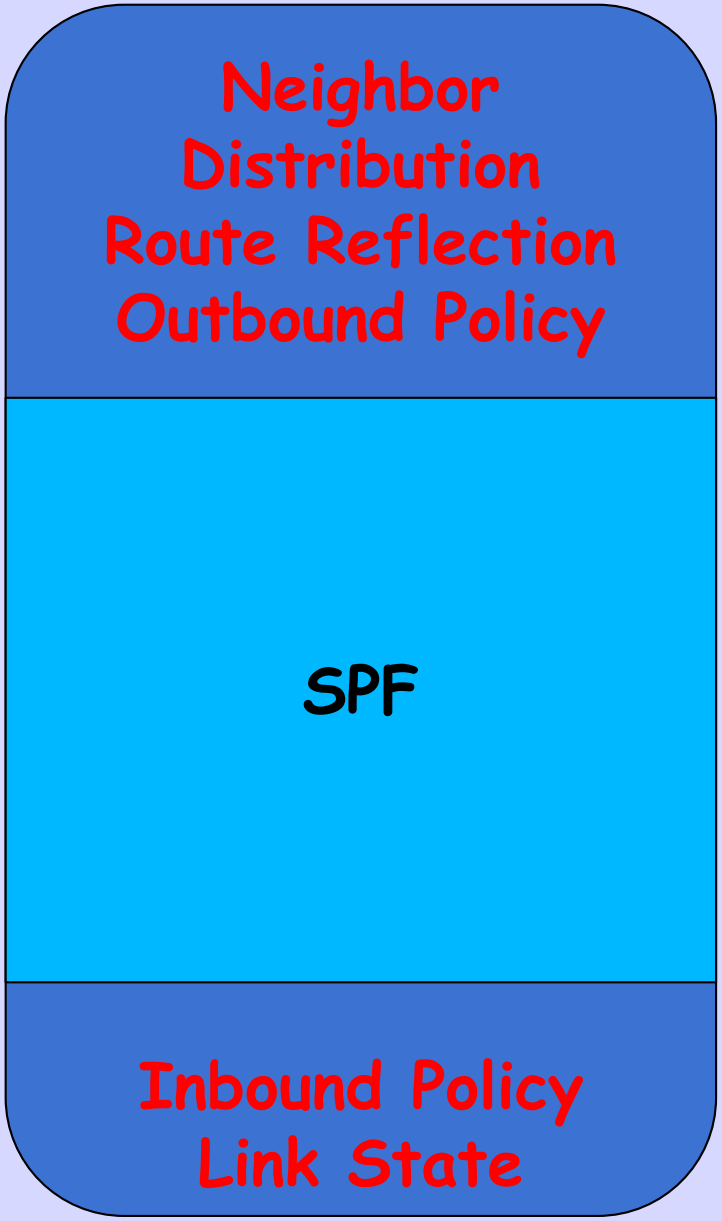
BGP-SPF

The Path Calculation of IS-IS With the Update Rate of BGP

BGP Only Signals Changes Does Not Repeatedly Flood Link State



BGP4 Classic



BGP-SPF

AS-Path Length
EGP vs IGP
Arrival Order
Non-deterministic
MED
IGP metric
Tie Break

Removed!

But ...
How Does BGP-SPF Learn Link State?

Link Discovery and Liveness

What do we really need?

Application

Presentation

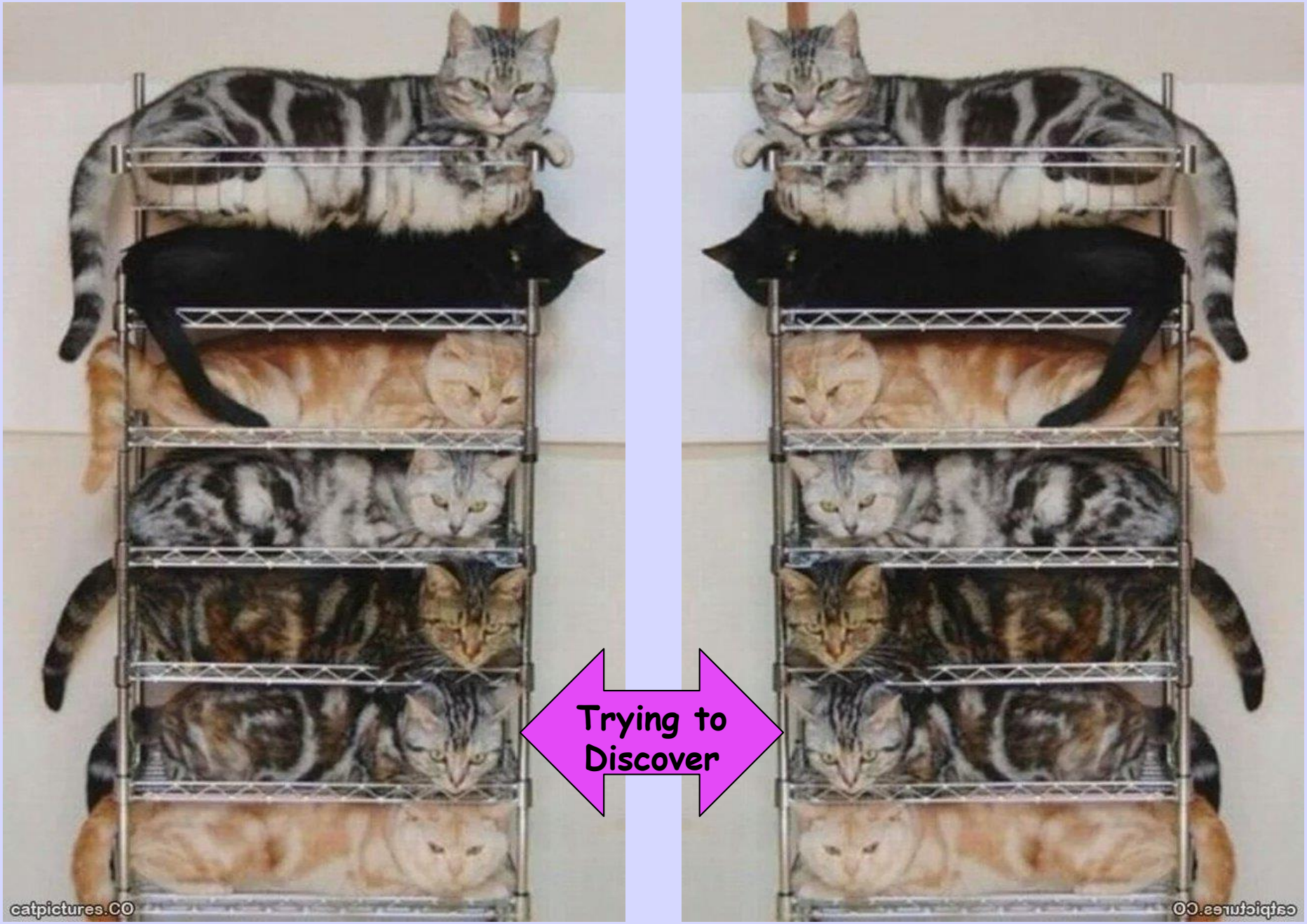
Session

Transport

Network

Data Link **We Are Here**

Physical



Two Kinds of Standards

Union – the accumulation of all the features anybody wanted

Intersection – only those things everybody absolutely had to have

Either Tony Hoare or Klaus Wirth – I can not find the quote <blush>

IETF asks the ITU

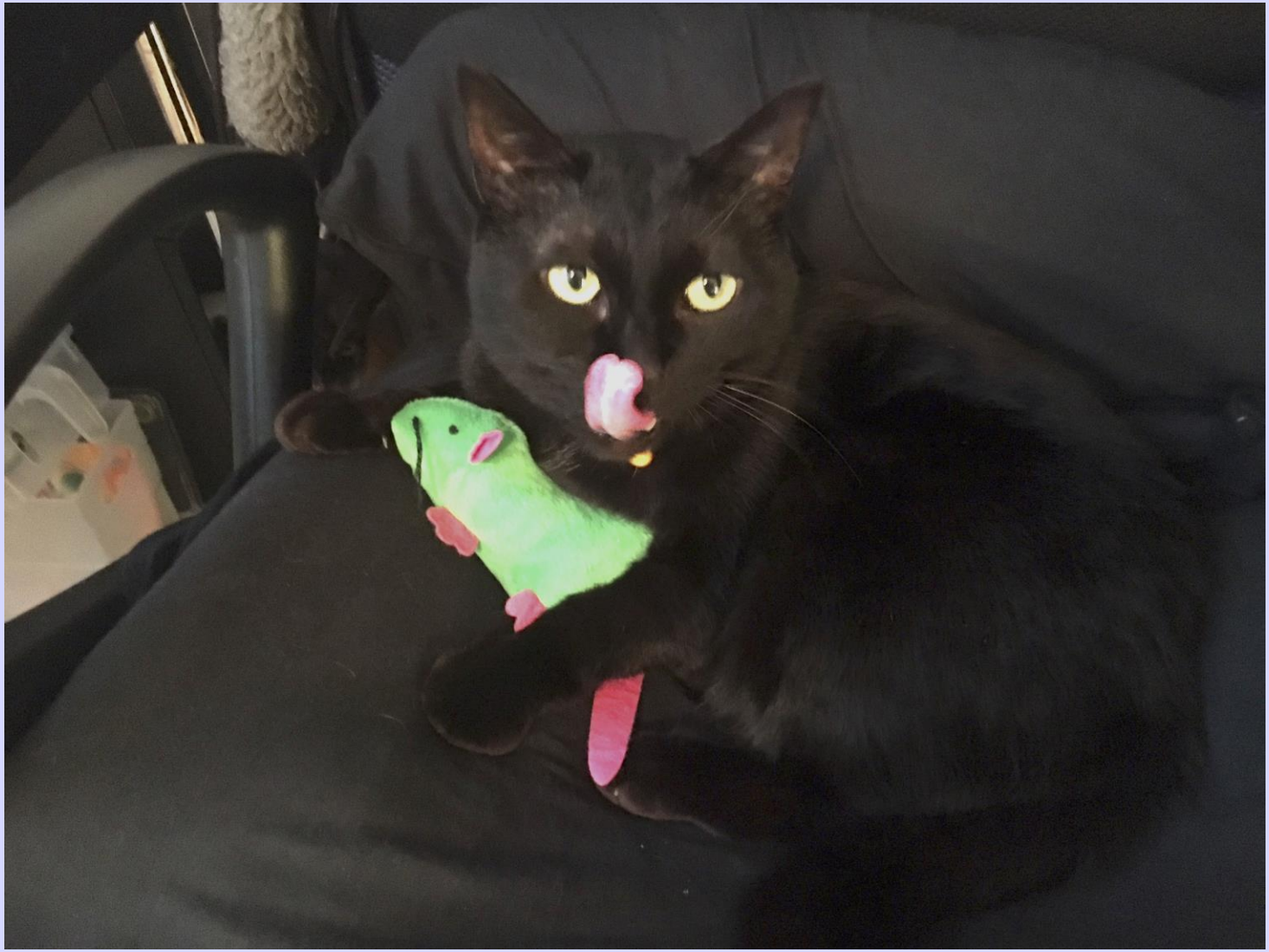
Q: So you add features until the "NO"s stop?

A: We don't like to think of it that way

Must Haves

- Discover Nodes and Links
- Discover Link Encapsulations:
  - IPv4, IPv6, MPLS4/6
- Maintain Liveness
- Northbound API to BGP-SPF

Security



Security?

- Datacenter Ops seem not to think of security at this layer (or any!)
- Do we want to add Authentication and maybe Integrity?
- One of the things which are likely to drive the size over 1,500

Non-Features

- Routing Data, BGP-SPF does that
- Access to IGP Databases, This is discovery and liveness, not routing
- Just want the Link
- Transport, not our job

Desiderata

- Discovery & Liveness for BGP-SPF
- Simple but usable in Massively Scalable networks of >10,000 nodes
- May be useful for other applications
- Simple
- Extensible (e.g. authentication, cost)
- Simple
- No IPR

Why Simple?

We are here to produce easily understood, implementable, and securable standards, not build résumés.

Why Simple?

A high goal of software engineering is to remove the need for features. It's a vital part of designing for simplicity, even invisibility. -- Rob Pike

LLDP

- IEEE Protocol
- A Little Noisy
- Beacons, not KeepAlives
- Viable but potential Show-Stopper

IS-IS Discovery

- Flooding & Noisy
- Complex enough that BGP-LS was invented so normals could get the link state database
- IS-IS not commonly implemented on MSDC devices, so would need to profile and develop

Edge Control Protocol

- It is a transport & Controlled by IEEE
- A Reliable layer two transport, on top of LLC
- Has flow control, reliable, non-reorder, … transport
- used for EVP and PD/CSP
- Reinventing TCP over 802.1

BGP Neighbor Autodiscovery

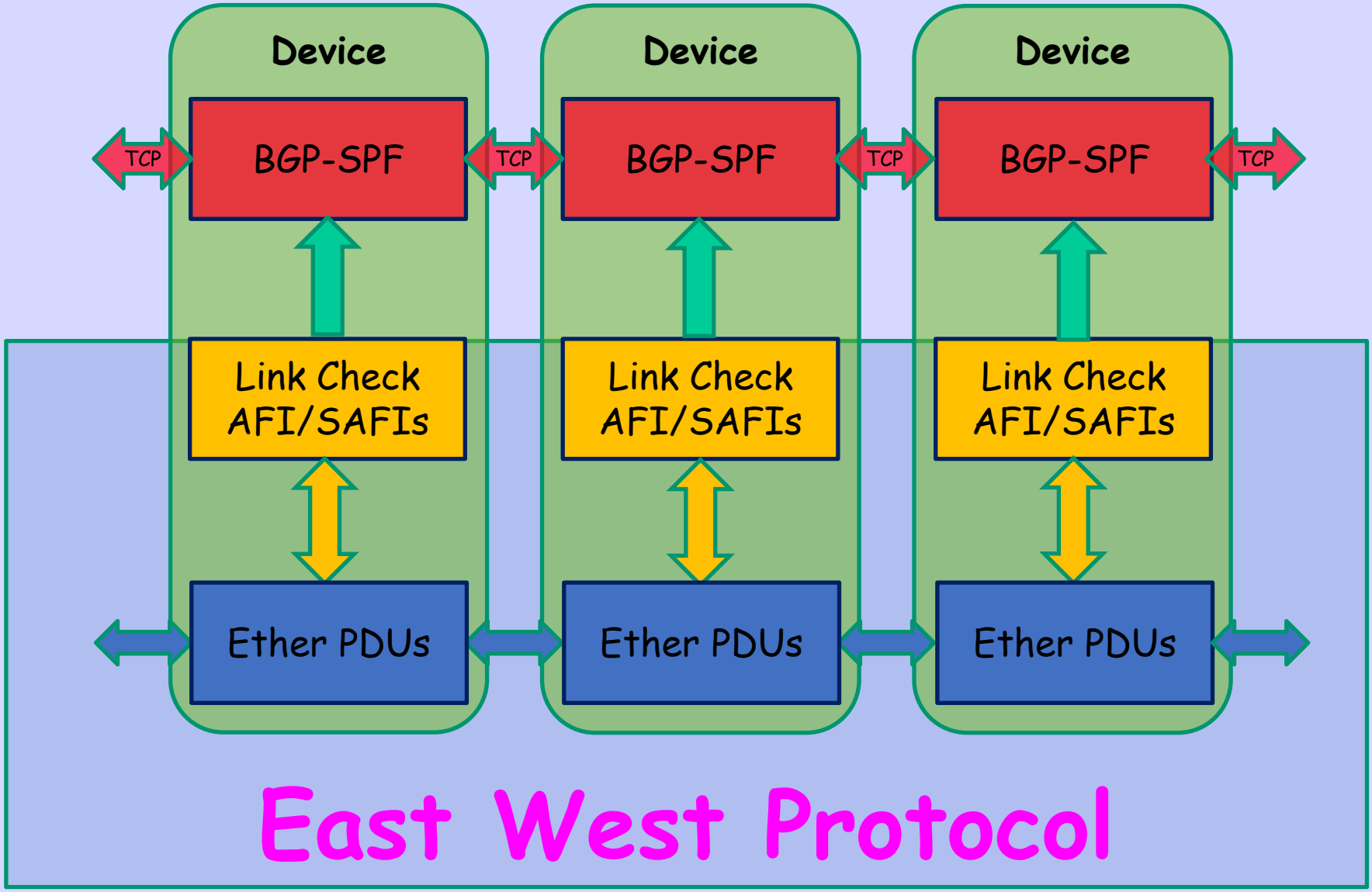
- IETF Unrealistic & Incomplete Protocol
- Very new
- Needs the peering address to get the peering address
- AS Based, can not use other idents
- Not really discovery at all, configuration
- No liveness

Link State Over Ether

- Custom made for the job
- Very bare bones, brutally simple
- Only does discovery and liveness
- New, therefore risky, still in flux
- But so is BGP-SPF
- No measurement or monitoring tools

	LLDP	IS-IS	ECP	BNA	LSOE
Who Owns	IEEE	IETF	IEEE	IETF	IETF
Maturity	Mature	Mature	Recent	New	New
Complexity	Somewhat	Noisy	Rather	Somewhat	Almost too Simple
Discovery	Yes	Yes	Yes	Configure	Yes
Liveness	Beacons	Yes	No	No	Yes

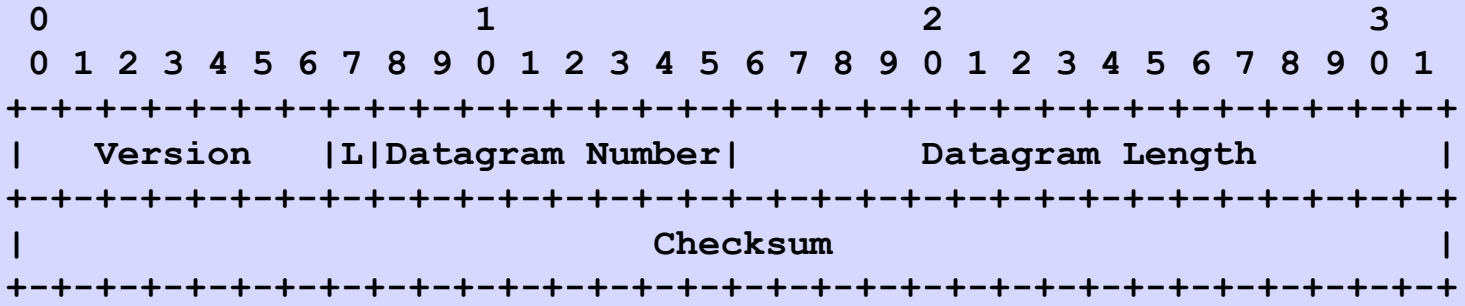
So a New Protocol



PDU

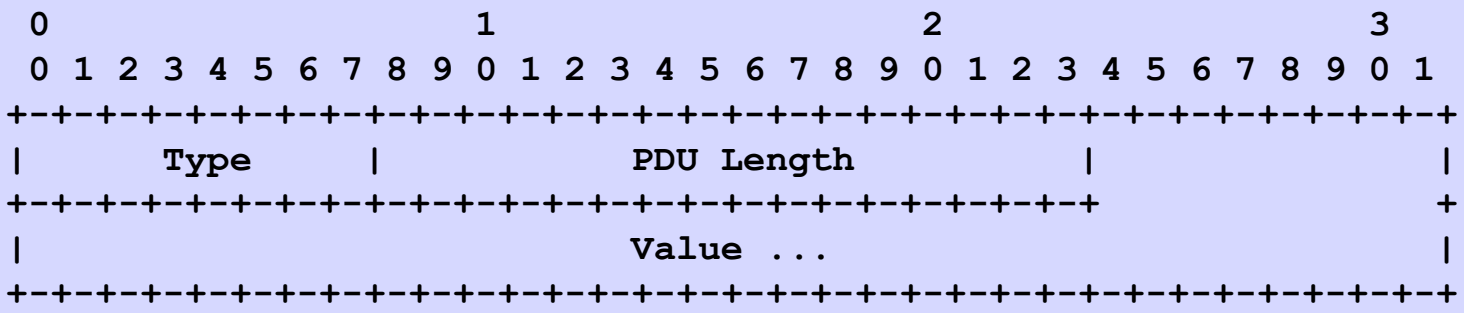
- *A PDU* (Protocol Data Unit) is an application layer message
- It may occupy multiple *Datagrams*
- Datagrams are one per Ethernet Frame

Datagram



- A Datagram is one Ethernet Frame
- A Datagram has *Number*, *Length*, and *Checksum*
- The *L* flag is set on the last datagram of an application layer PDU
- This Transport Layer assembles PDUs

Every Datagram a TLV



0 - HELLO

1 - OPEN

2 - KEEPALIVE

3 - ACK

4 - IPv4 Announcement

5 - IPv6 Announcement

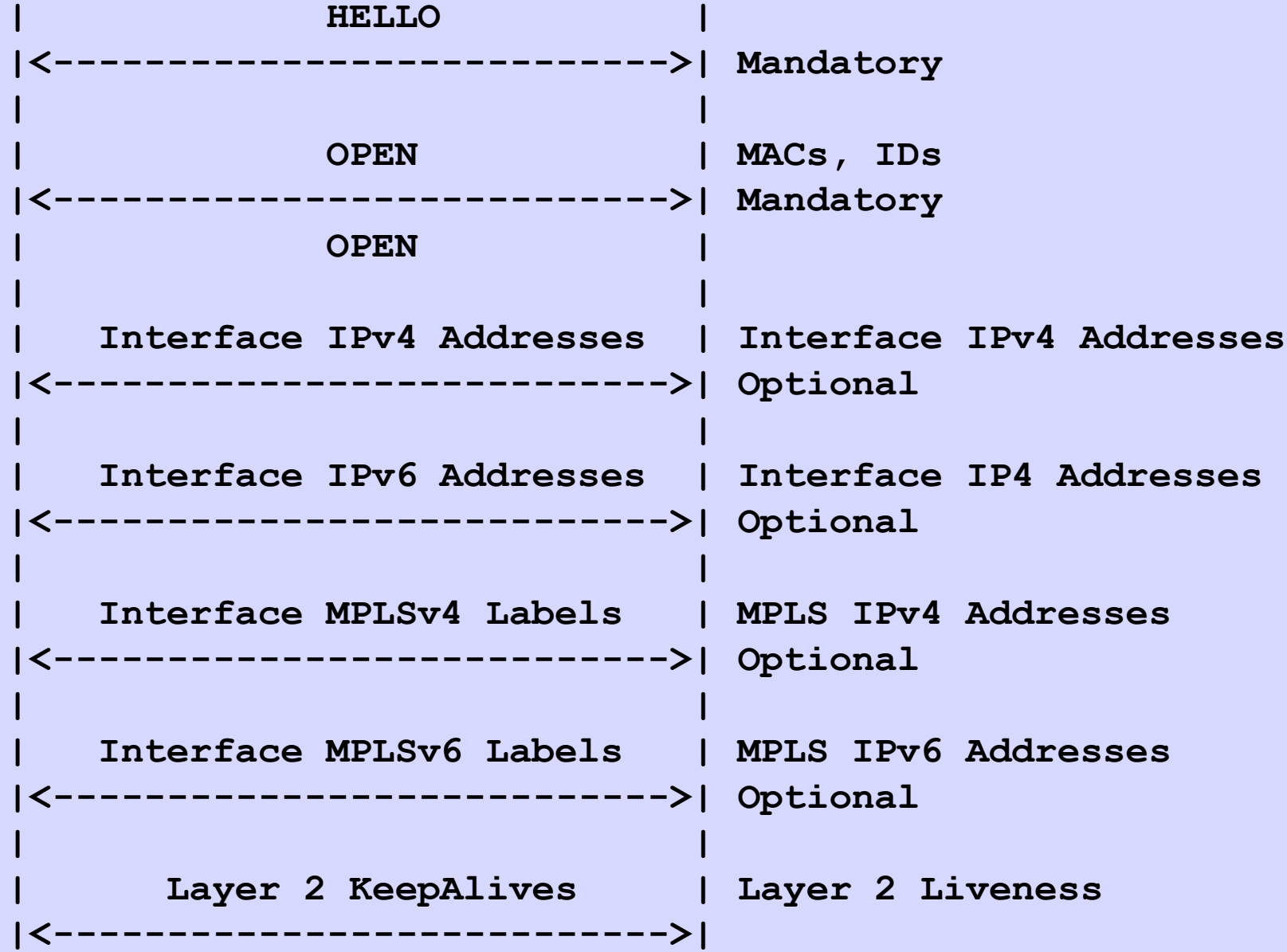
6 - MPLS IPv4 Announcement

7 - MPLS IPv6 Announcement

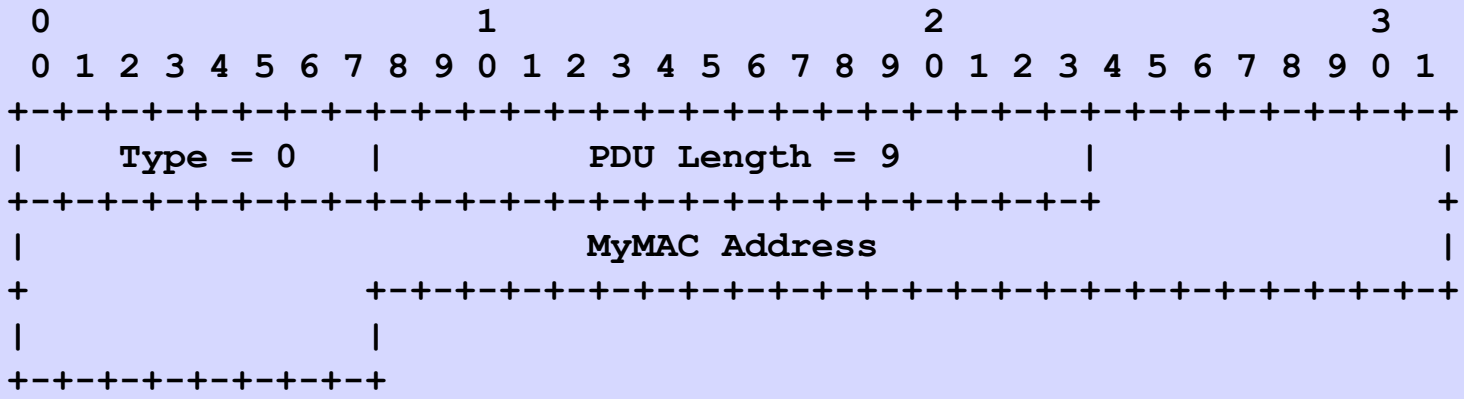
Sessions are Pretty Clear

OPEN and Encapsulation PDUs are ACKed

Inter-Link Ether Protocol

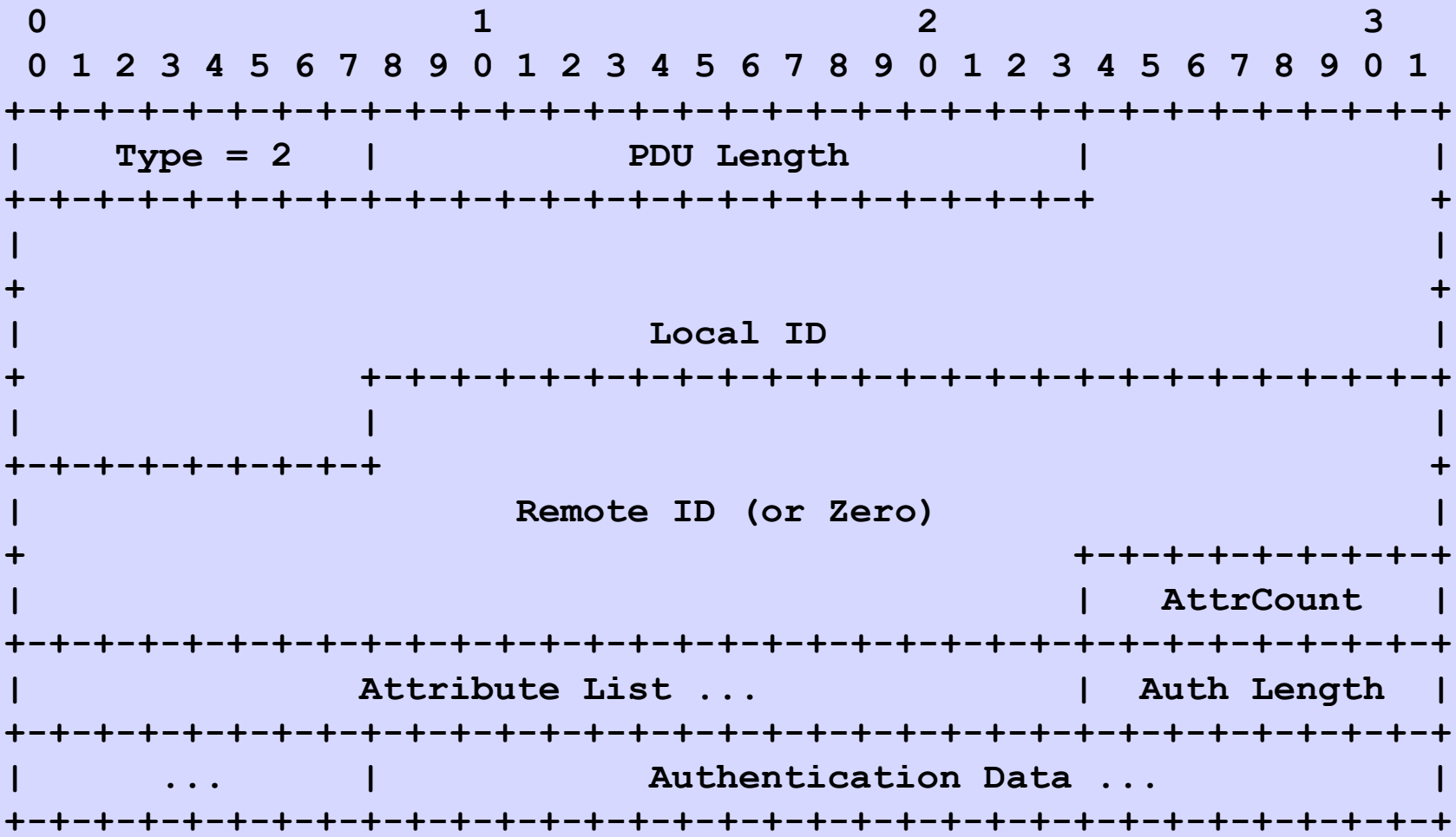


Link HELLO



- 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
- Config to be link-only or piercing switches

OPEN



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

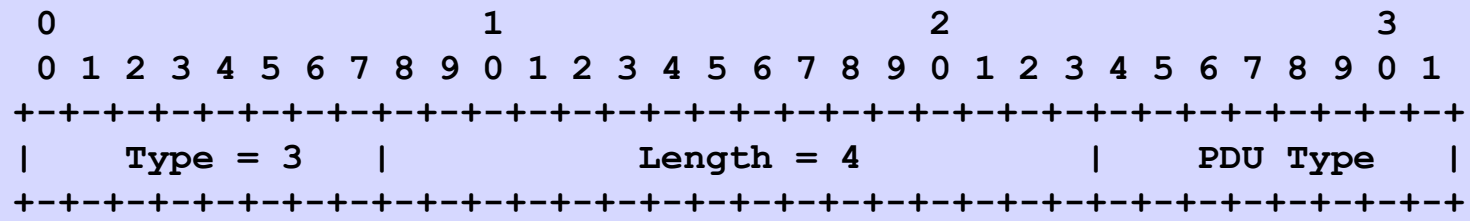
A node may have zero or more user-defined attributes, e.g. spine, leaf, backbone, route reflector, arabica, ...

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.

ACK

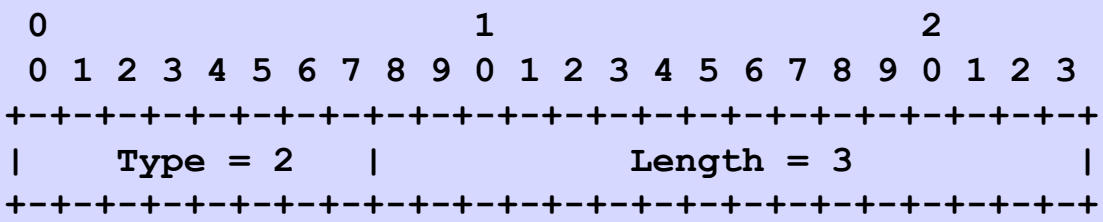


- The Receiver ACKs with a Type=3
- PDU Type is the Type of the PDU being ACKed
- Might add PDU Number being ACKed
- If the Sender does not receive an ACK in one second, they retransmit. Operator configured failure count.

Once We Know Each Other's MACs

Layer Two KeepAlives May be Started

L2 KEEPALIVE



This is in addition to L3 BFD etc.

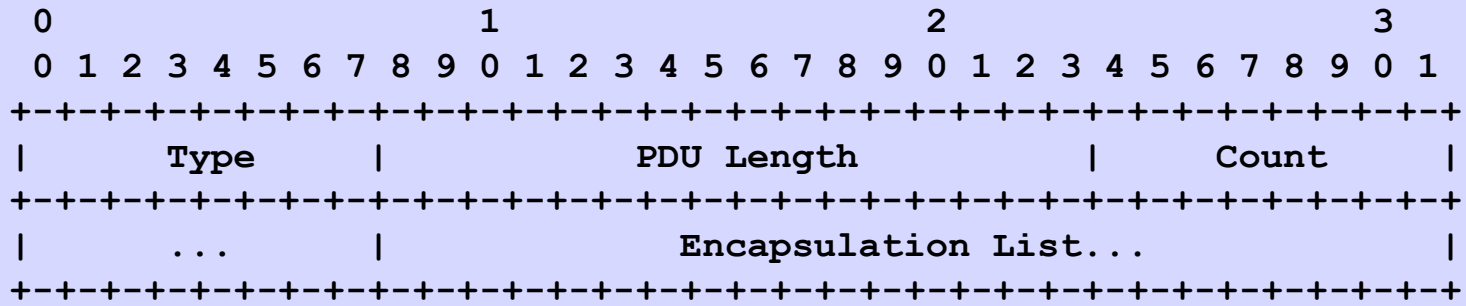
We assume that one or more Encapsulation addresses will be used to ping, BFD, or whatever the operator configures

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

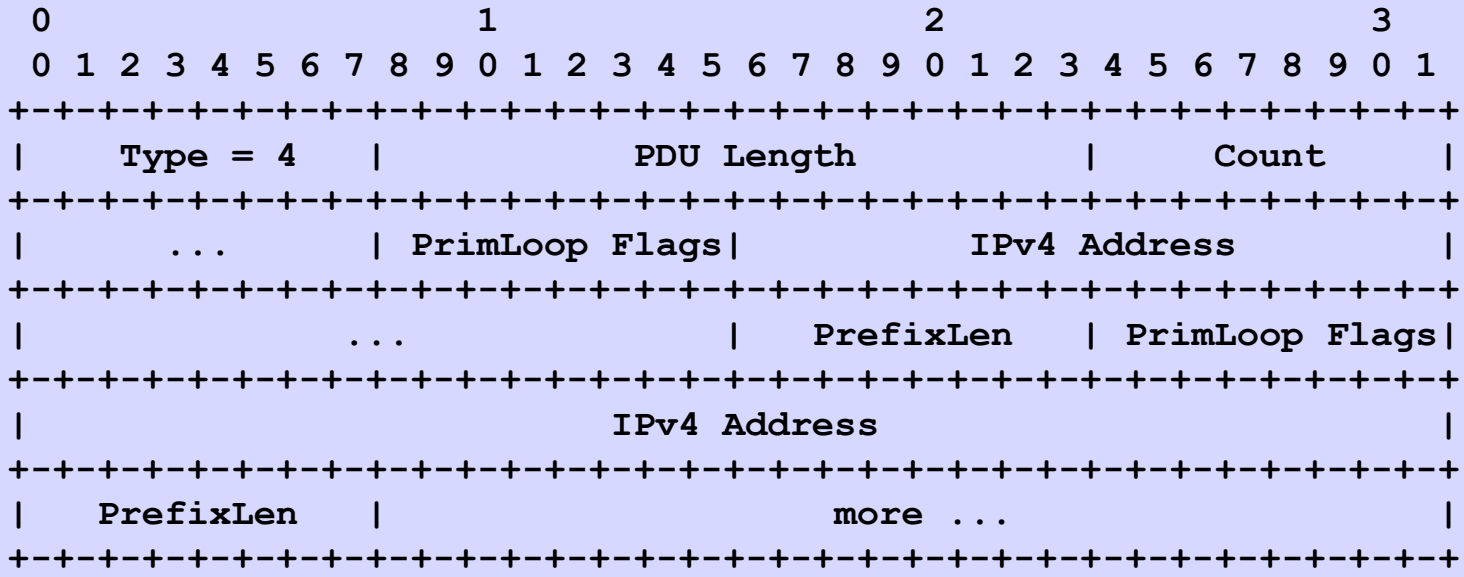
And Node IDs (often ASNs)

Now Announce Encapsulations of the Link Interfaces

Encapsulation PDU Header



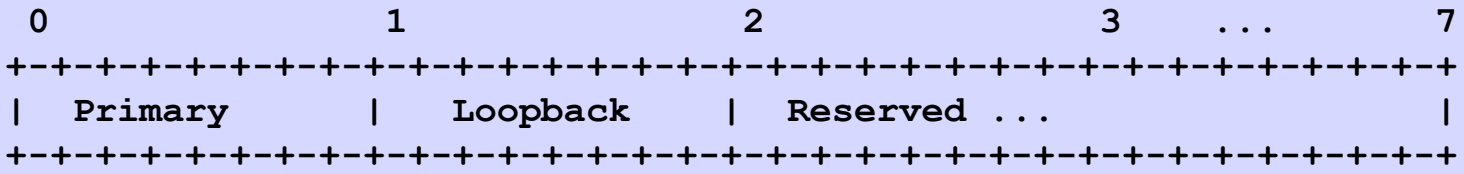
IPv4 Encapsulations



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

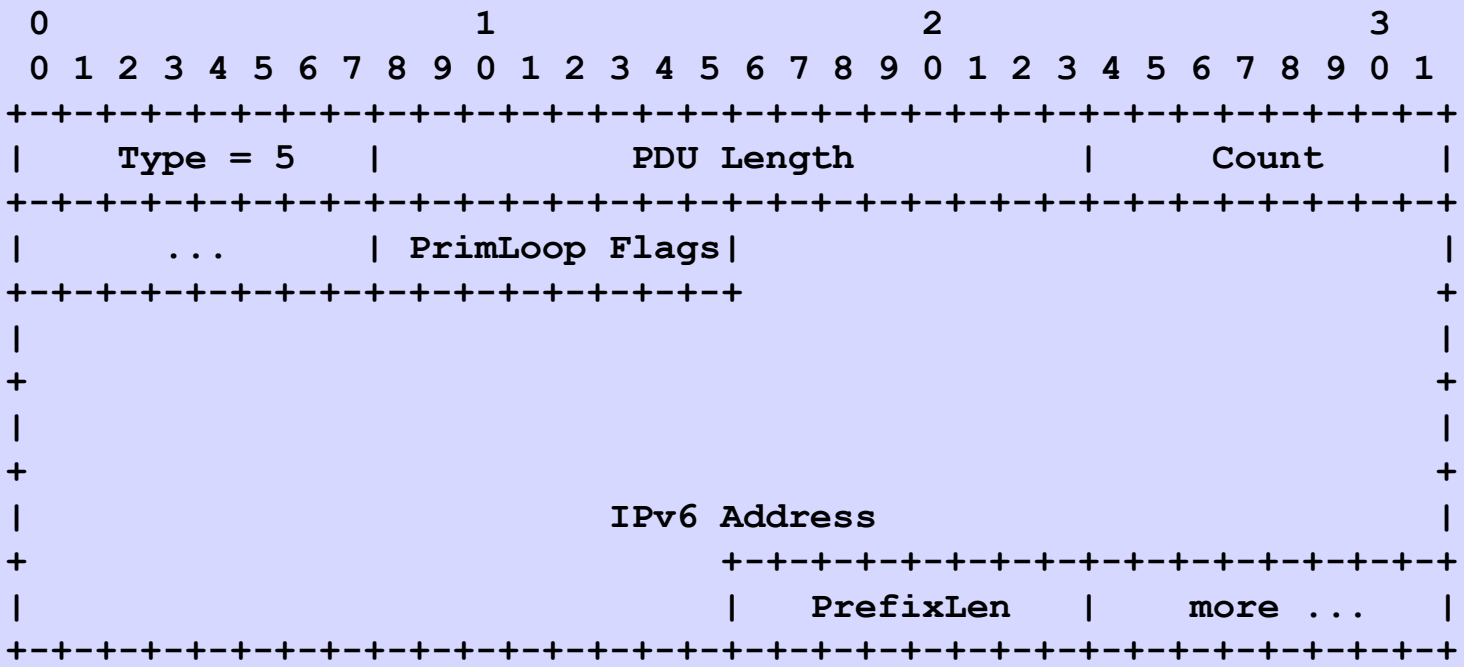
An Encapsulation message of Type T replaces all previous encapsulations of Type T

PrimLoop Flags

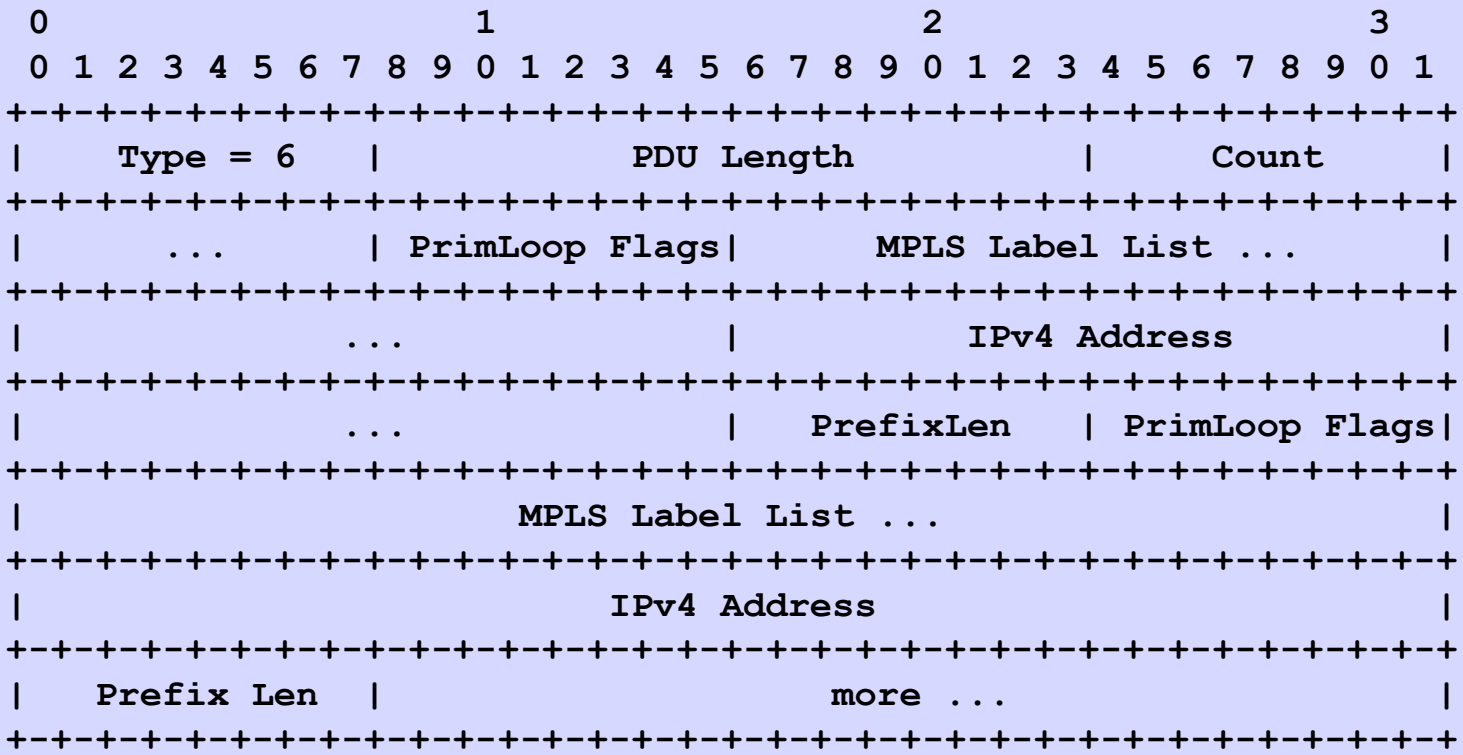


- An Interface may have multiple Encapsulations
- For each Encapsulation there might be multiple Addresses
- One Address per Encapsulation SHOULD be marked as Primary
- An Address may be marked as a loopback

IPv6 Encapsulations



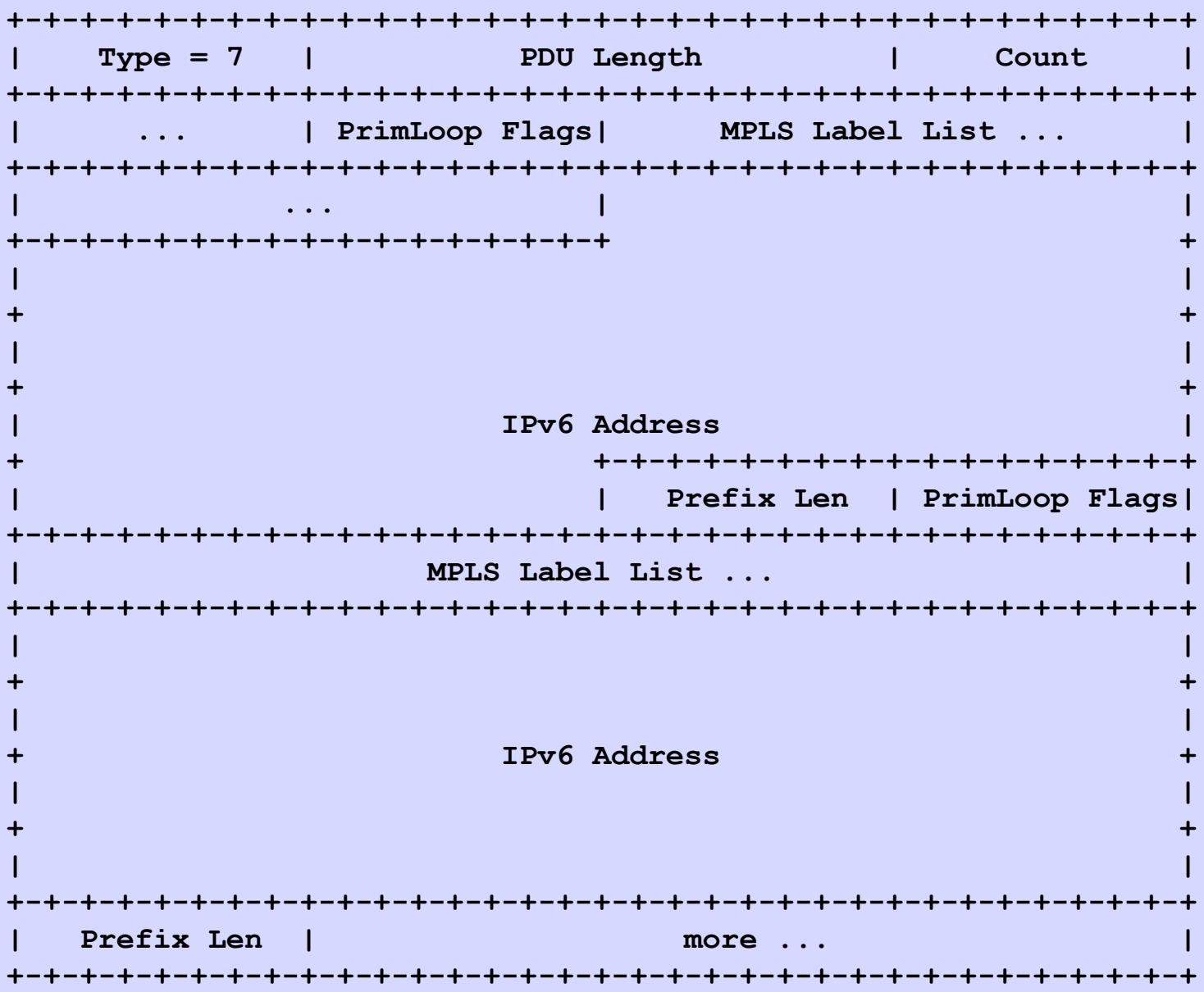
MPLS IPv4 Encapsulations



MPLS Label List

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

MPLS IPv6 Encapsulations



We're Looking at Security

Are you the Droid I was talking to Earlier?

Thinking of Security

- OPEN has public key plus …
- Signed with private key, proves possession
- All PDUs signed with 512-bit suffix
- KEEPALIVE could get a sequence number to reduce replay attack window
- Maybe a later Proof of Possession Challenge / Response PDU pair

Layer-3 IP/Label Liveness Should Also be Tested

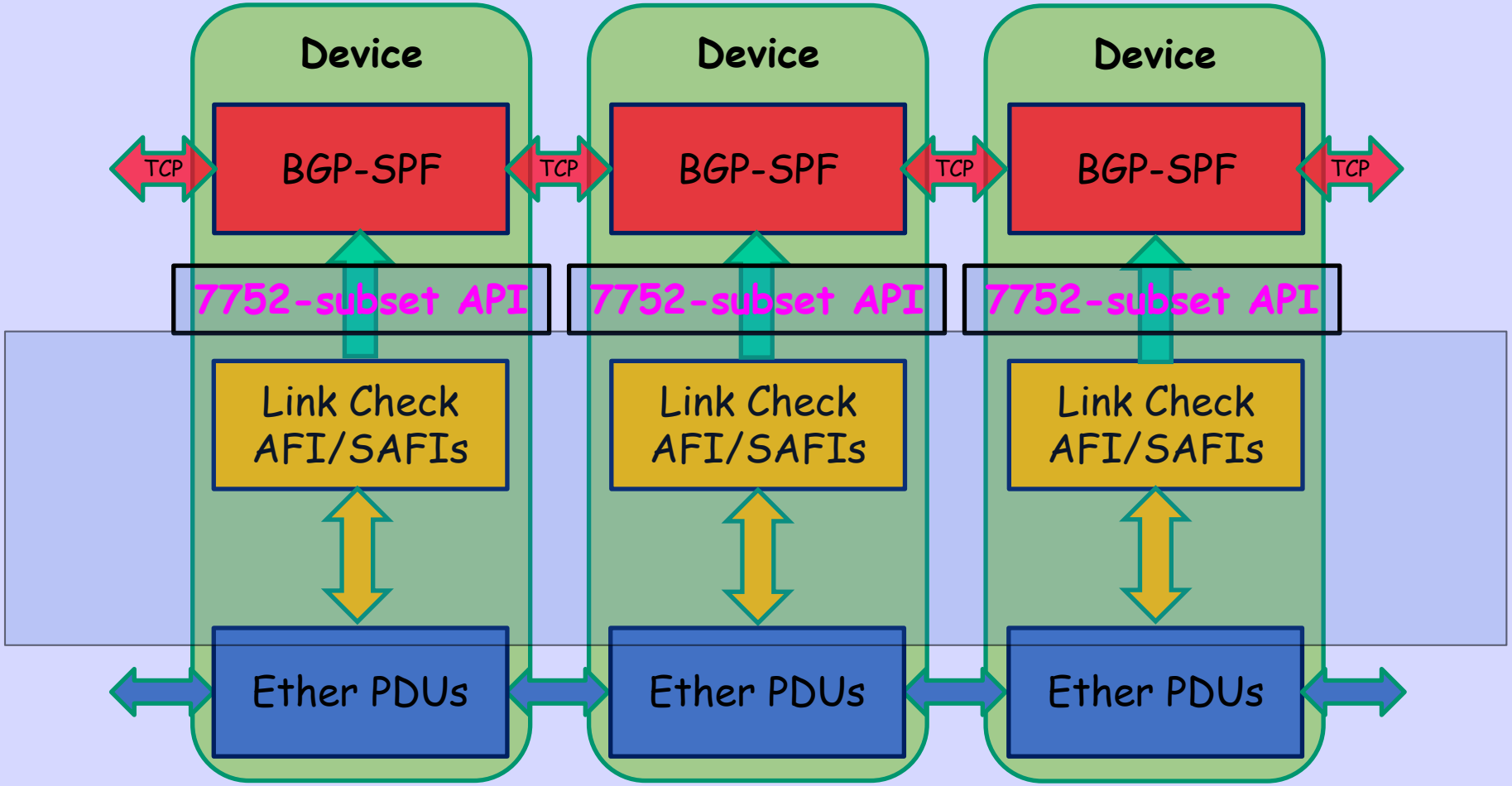
One or more Discovered AFI/SAFI Addresses Are Used to Ping, BFD, … to Assure Layer-3 Liveness

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



Open Questions

Should HELLO go Through an intermediate Layer Two Switch

Are HELLO and KEEPALIVE Redundant?

Should PDUs be Numbered so ACKs are not Ambiguous

Should the Version Number be Fail on MisMatch?

A Python3 Implementation is in Progress