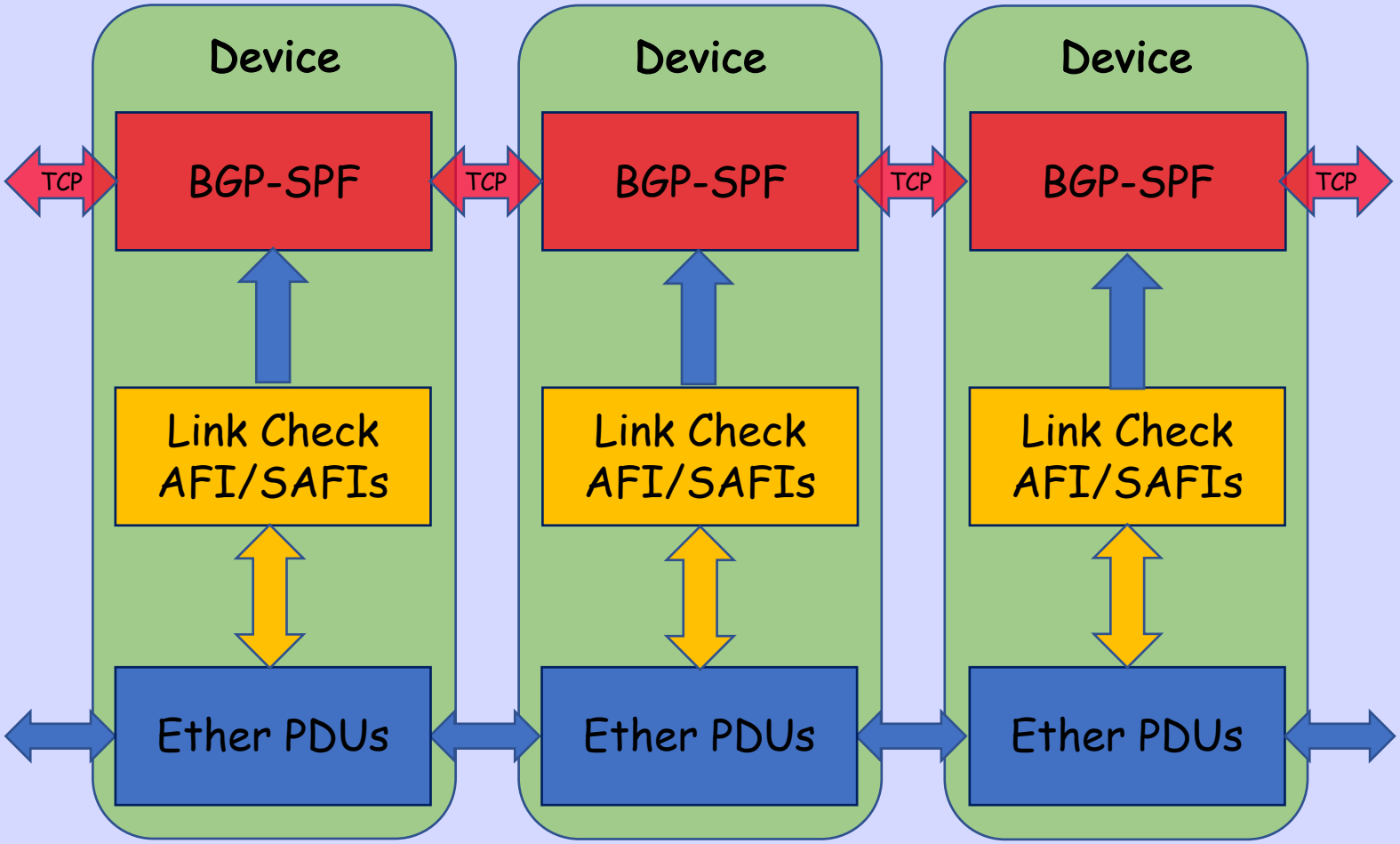
Link State Control Plane Over Ethernet

Randy Bush
Keyur Patel

Motivation

- No Central Controller, Distributed Routing!
- BGP-LS and BGP-SPF need link neighbor discovery, liveness, and addressability
- LLDP is an IEEE protocol, complex, and 'hard' 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 only discuss Ether payloads, not framing

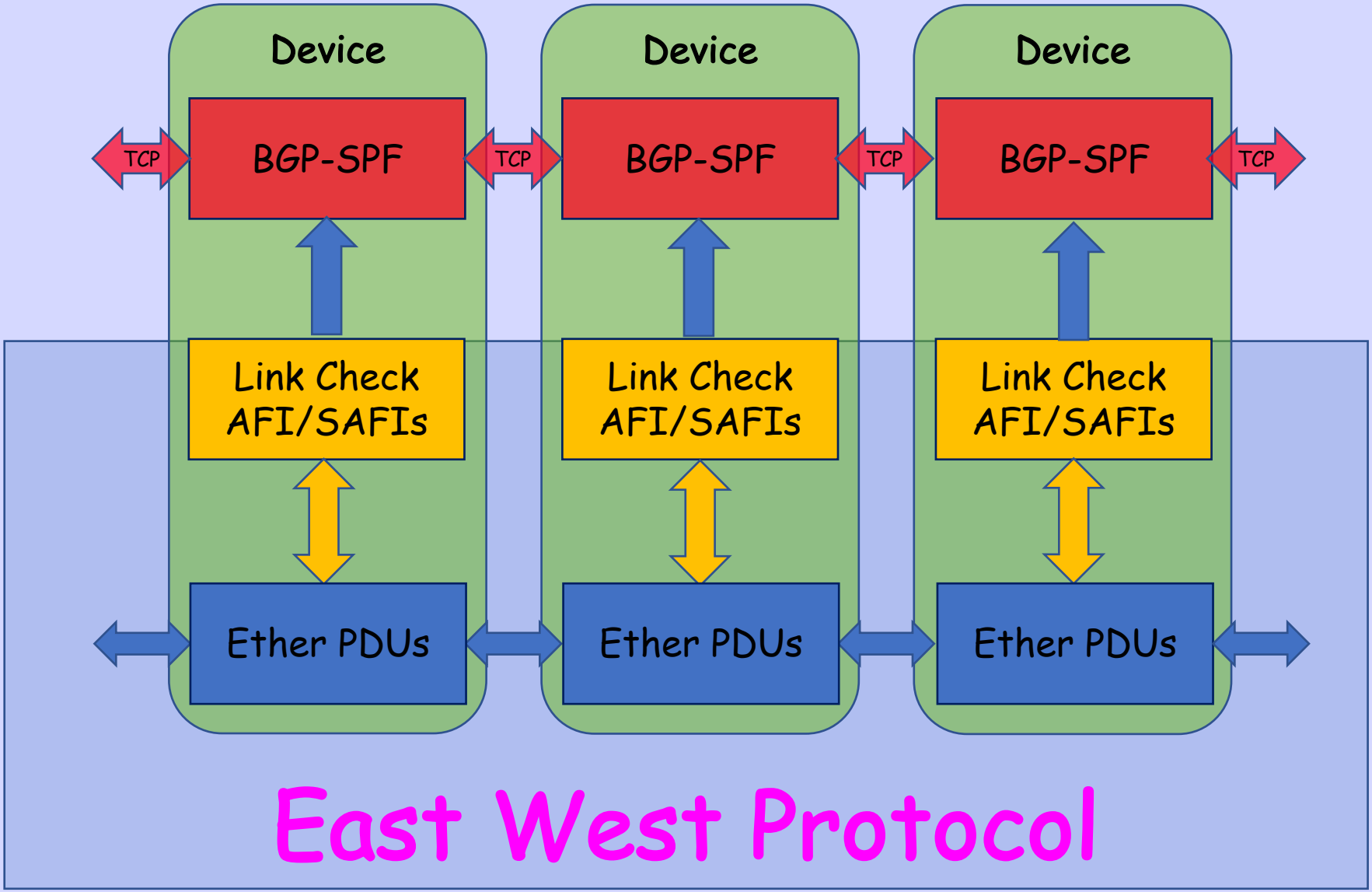
Topology / Routing Stack



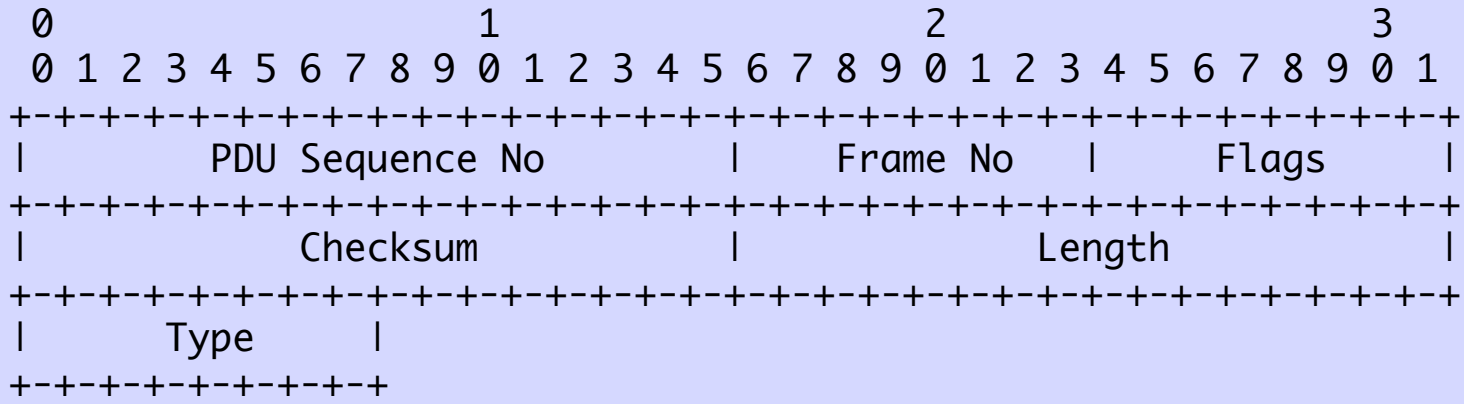
MAC Link State exchanged over raw Ethernet and pushed up stack

Add the AFI/SAFI data IP-Level Liveness Check

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

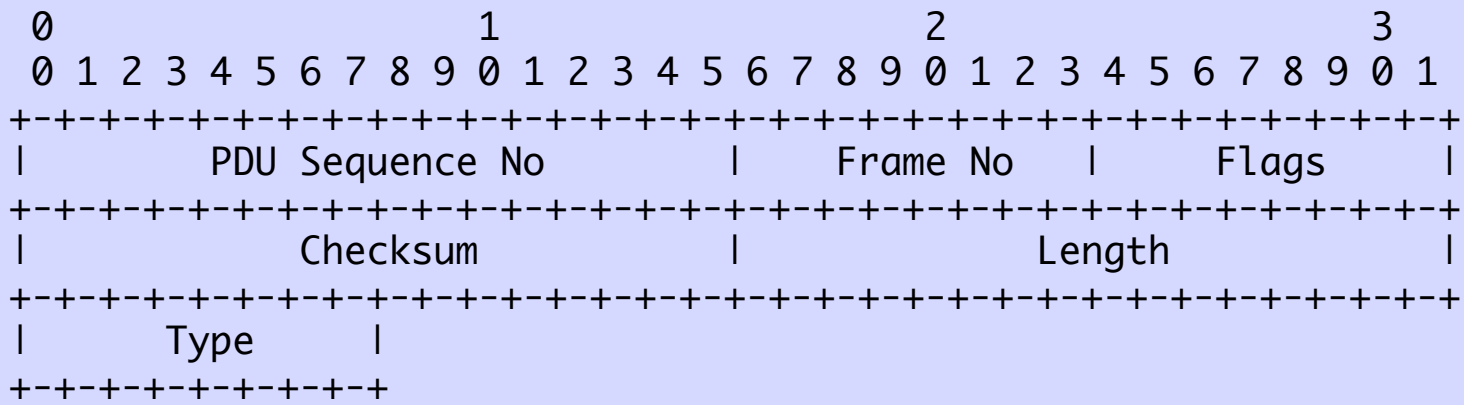


PDUs and Frames



- This is all about inter-device Link State
- A PDU is one or more Ethernet Frames
- A Frame has *PDU Sequence No* and a *Frame No* to allow assembly of out order frames
- Because BGP-SPF/IGP and Data Plane payloads are assumed to be IP over the same Ethernet, one worries about congestion

TLV, Well LTV, Frame



PDU Sequence No - Semi-unique identifier of a TLV PDU (e.g. UNIX time)

Frame No - 0..255 Frame Sequence Number Within a multi-frame PDU

Flags (bits)
- 0 – Sender has been restarted
- 1 – one of a multi-Frame sequence
- 2 – last of a multi-Frame sequence

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

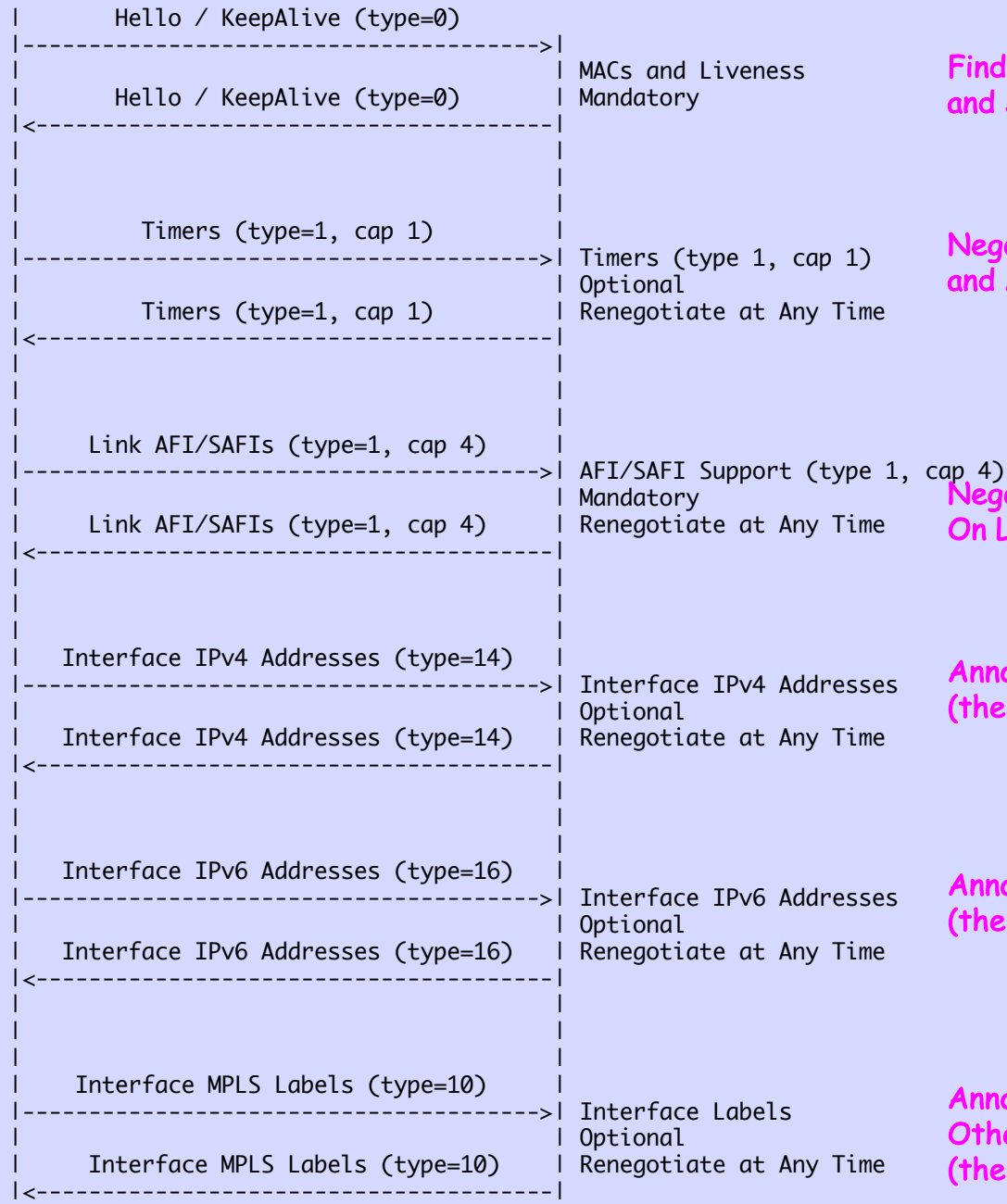
Length - Total Bytes in PDU including all frames and fields

Type (int)
- 0 – Hello / KeepAlive
- 1 – Capability

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



Find Each Other's MACs & ASNs and Start MAC KeepAlive Checks

Negotiate Timers for KeepAlives and ACKs

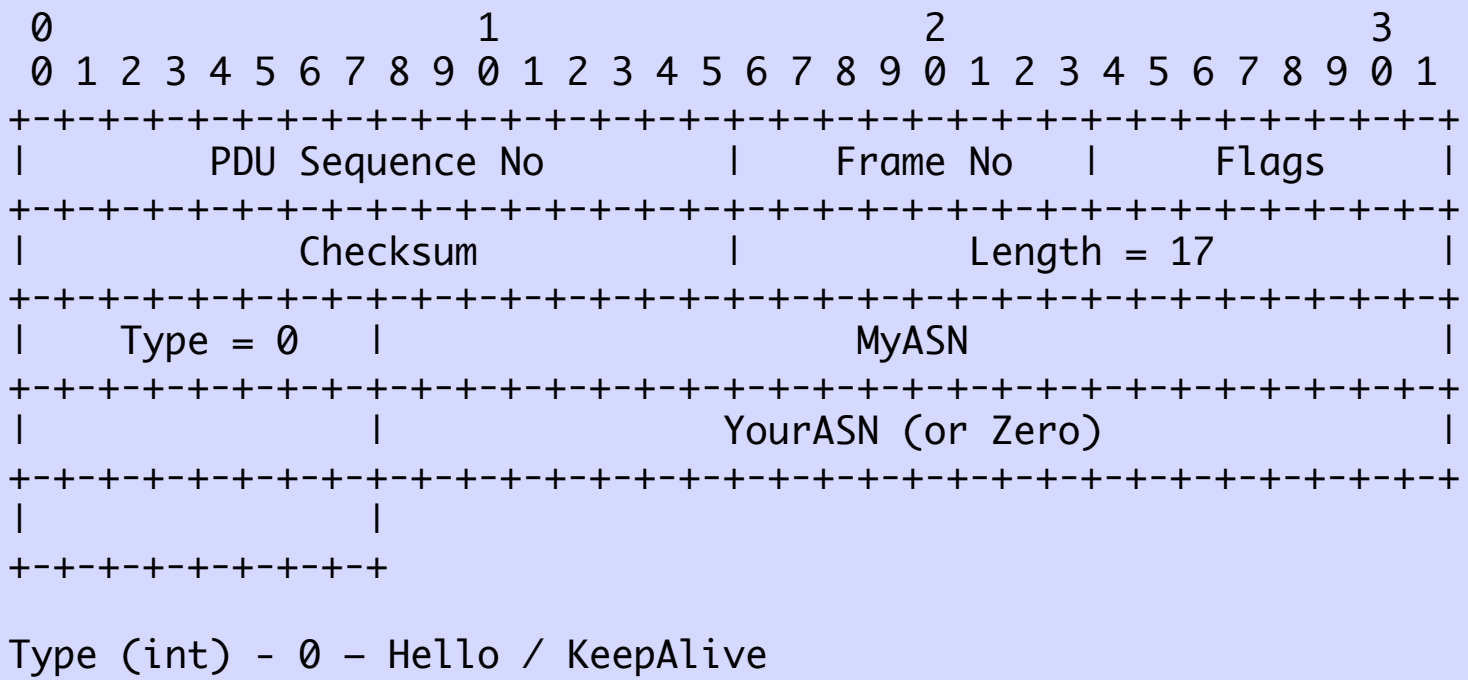
Negotiate AFIs/SAFIs Supported On Link

Announce Link IPv4 Addresses (there is an ACK)

Announce Link IPv6 Addresses (there is an ACK)

Announce Interface MPLS Labels and Other types of Tunnels (there is an ACK)

Link Hello / KeepAlive



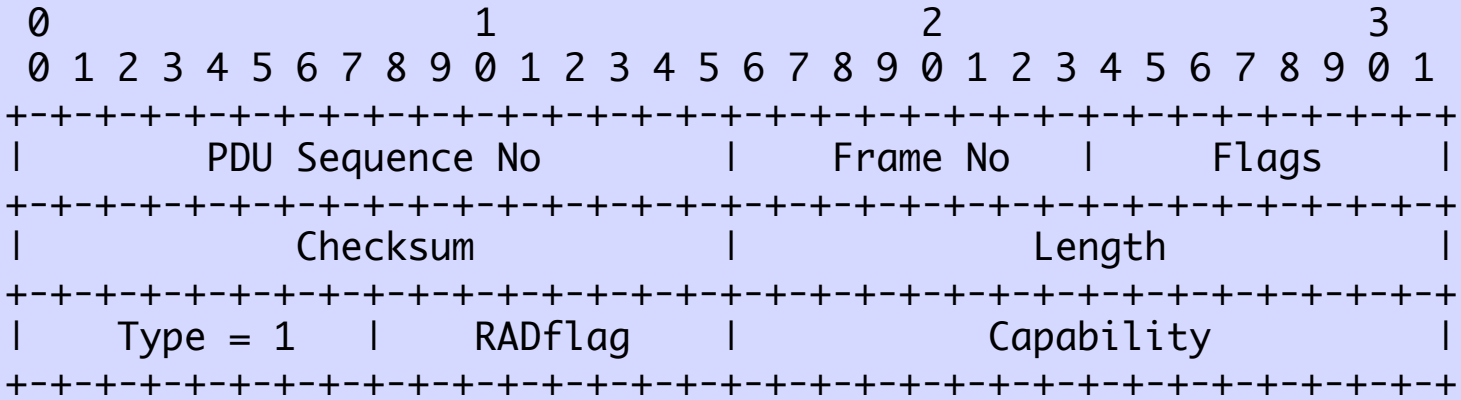
Type (int) - 0 – Hello / KeepAlive

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

Once We Know Each Other's MACs

Ether Level KeepAlives Start

Capability Negotiation (lowest MAC initiates)

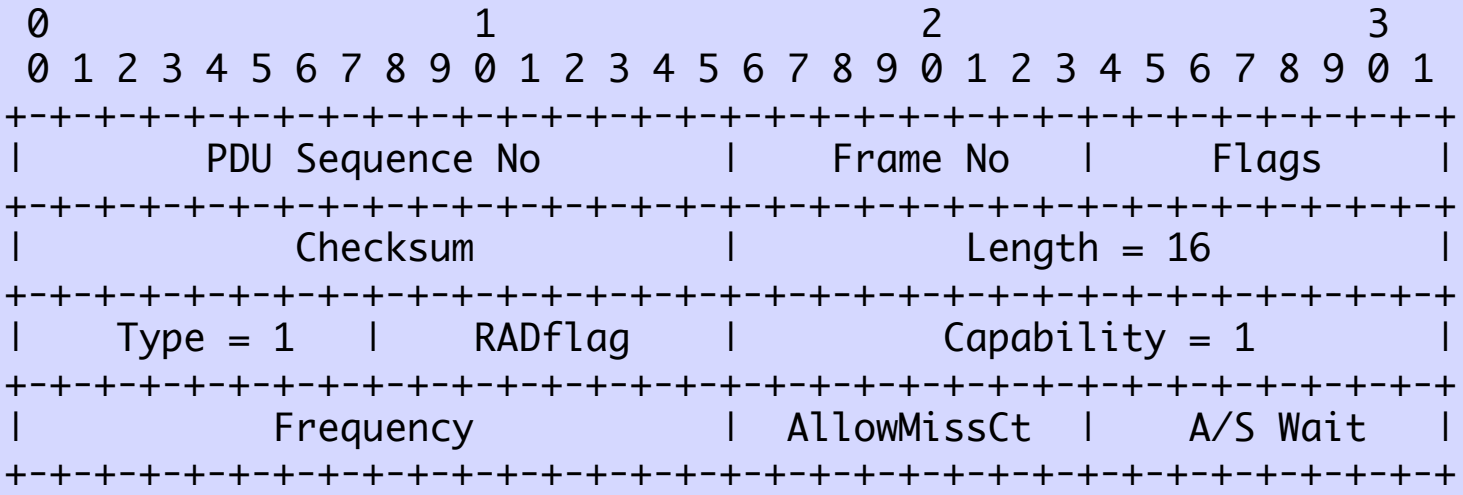


Type (int) - 1 Capability

RADflag (int) - 1 – Request
- 2 – Agree
- 3 - Deny

Capability - See Dictionary

Timer Negotiation



RADflag (int) - 1 – Request
- 2 – Agree
- 3 - Deny

Capability - 1

Frequency - Seconds/10 between KeepAlives (Default is 600)

AllowMissCt - Number of missed KeepAlives before declared down

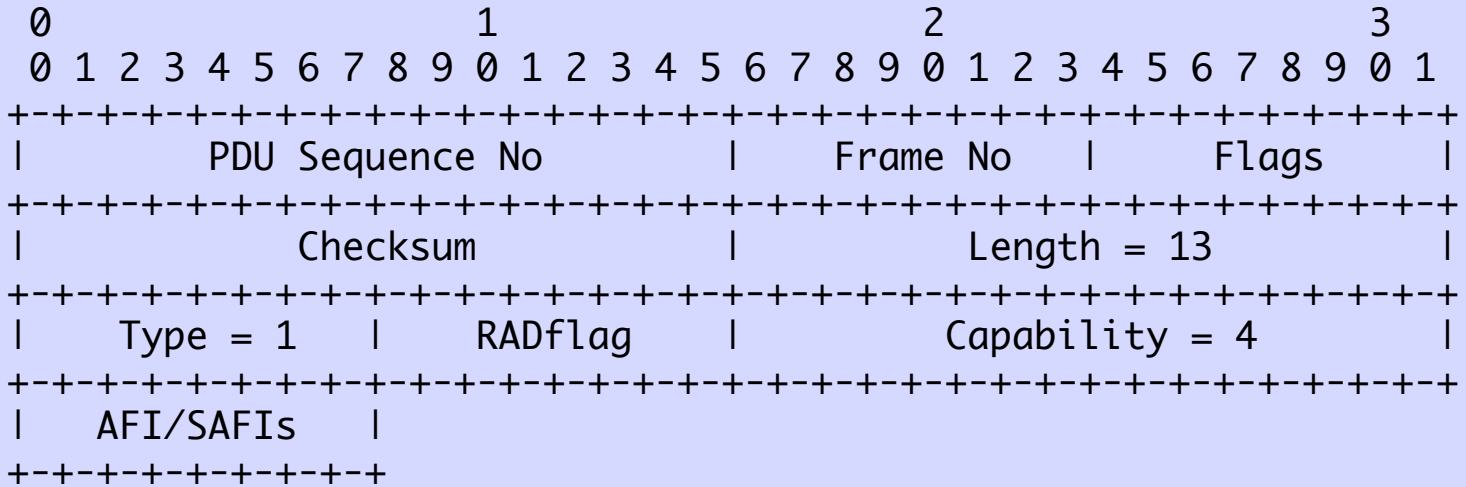
A/S Wait - AFI/SAFI ACK Timeout in Sec/10 (default 10)

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

And ASNs

Now Negotiate AFI/SAFIs of Link Interfaces

AFI/SAFI Capabilities



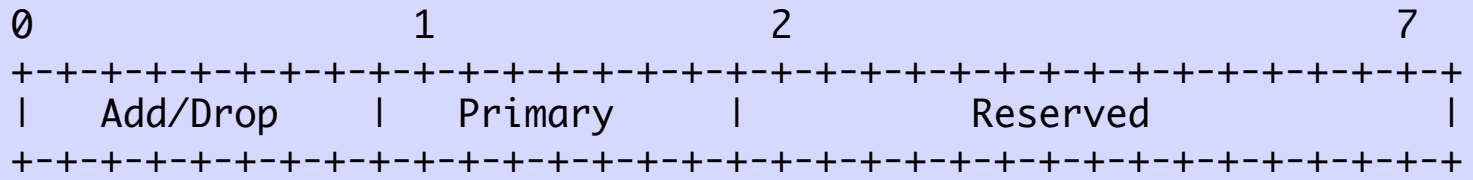
RADflag (int) - 1 – Request
- 2 – Agree
- 3 - Deny

Capability - 4

AFI/SAFI (int) – 10 – IPv4
- 11 – IPv6
- 12 – MPLS IPv4
- 13 – MPLS IPv6
- … other tunnels (e.g. GRE)

Now Both Sides Exchange Their Interface AFI/SAFI Configuration for the Negotiated AFI/SAFIs

Announce/Withdraw/Primary Flag



- An Interface may have multiple AFI/SAFIs
- For each AFI/SAFI there might be multiple Addresses
- One Address per AFI/SAFI SHOULD be marked as Primary
- It is a bit in the Announce/Withdraw Flag

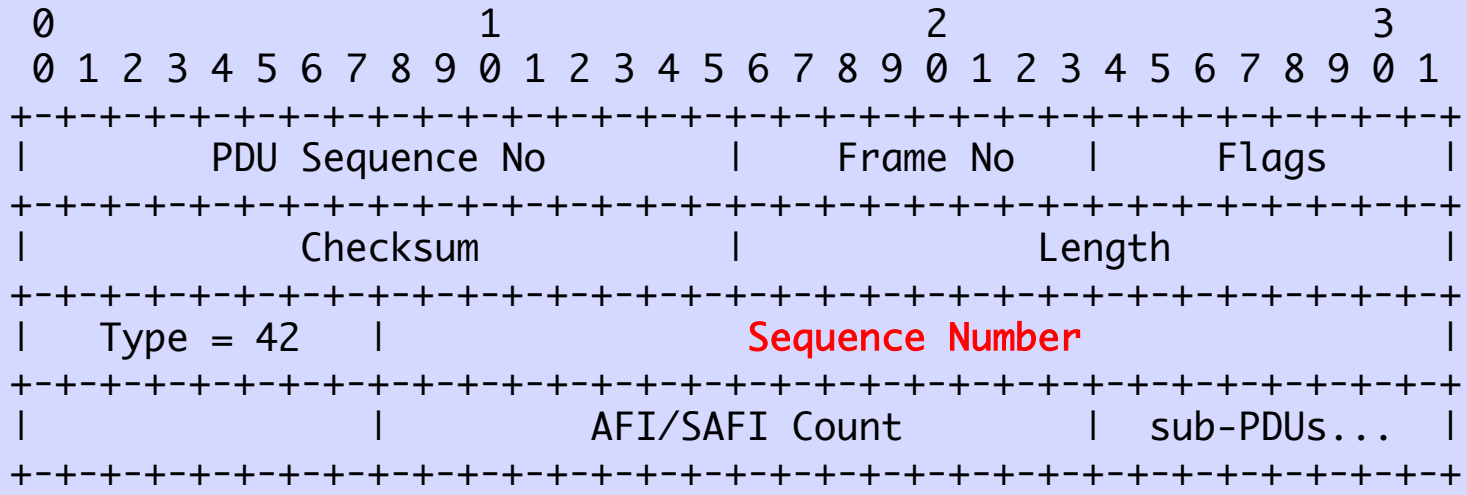
The AFI/SAFI Exchange

Is Over an Unreliable Transport

So There are Sequence Numbers

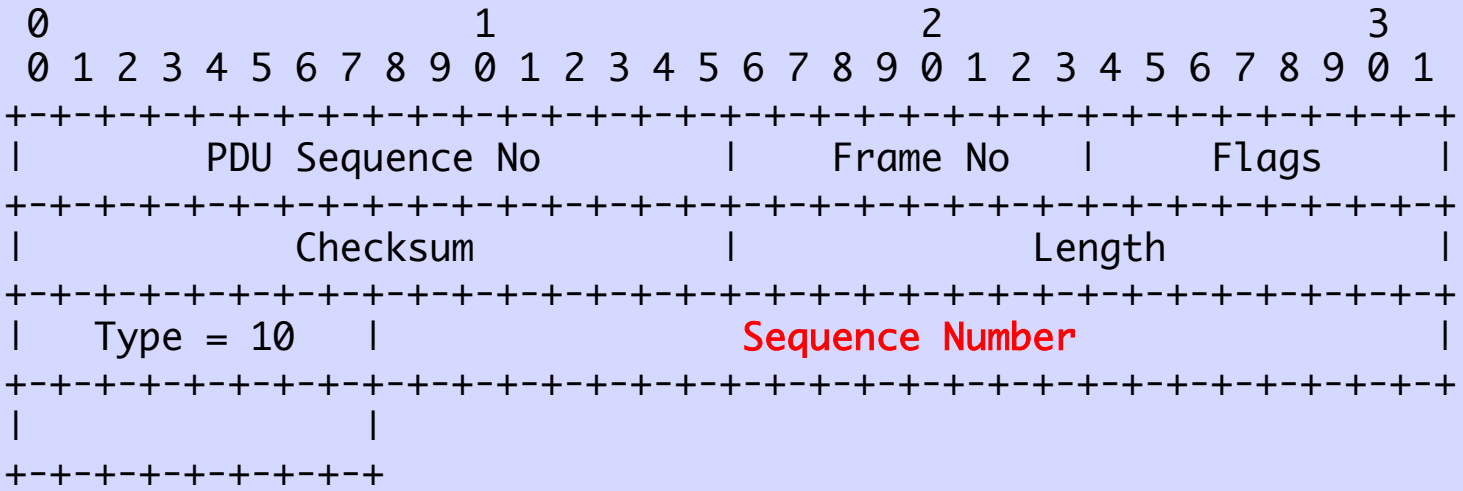
and ACKs

AFI/SAFI PDU Sequence Number



- The Sequence Number is a p2p link Announcement Counter
- The Receiver will ACK it with a Type=10
- If the Sender does not receive an ACK in one second, they retransmit. Other delay negotiated in Timing Capability

AFI/SAFI ACK

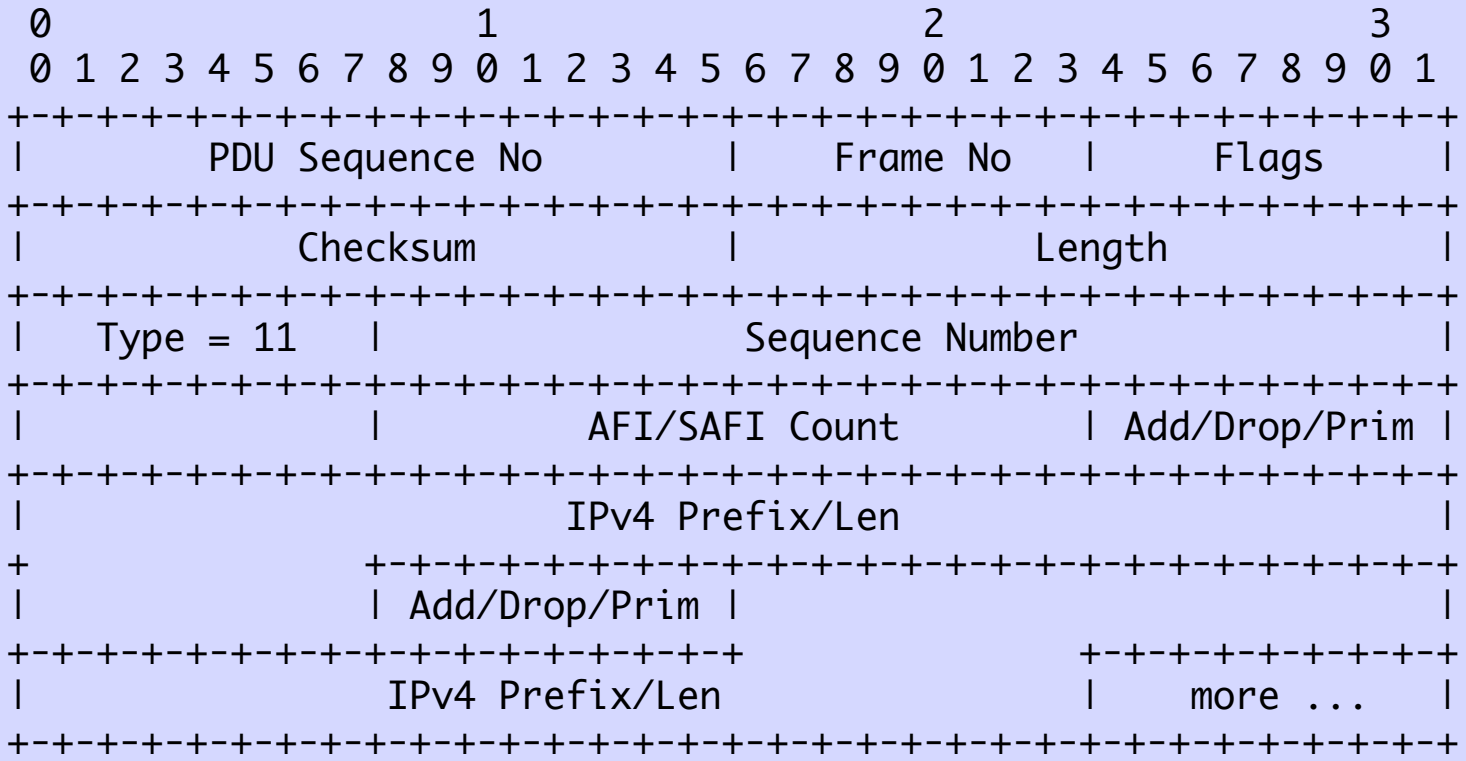


The sequence Number is the one being ACKed

Any PDU with a Sequence Number needs an ACK from the other end

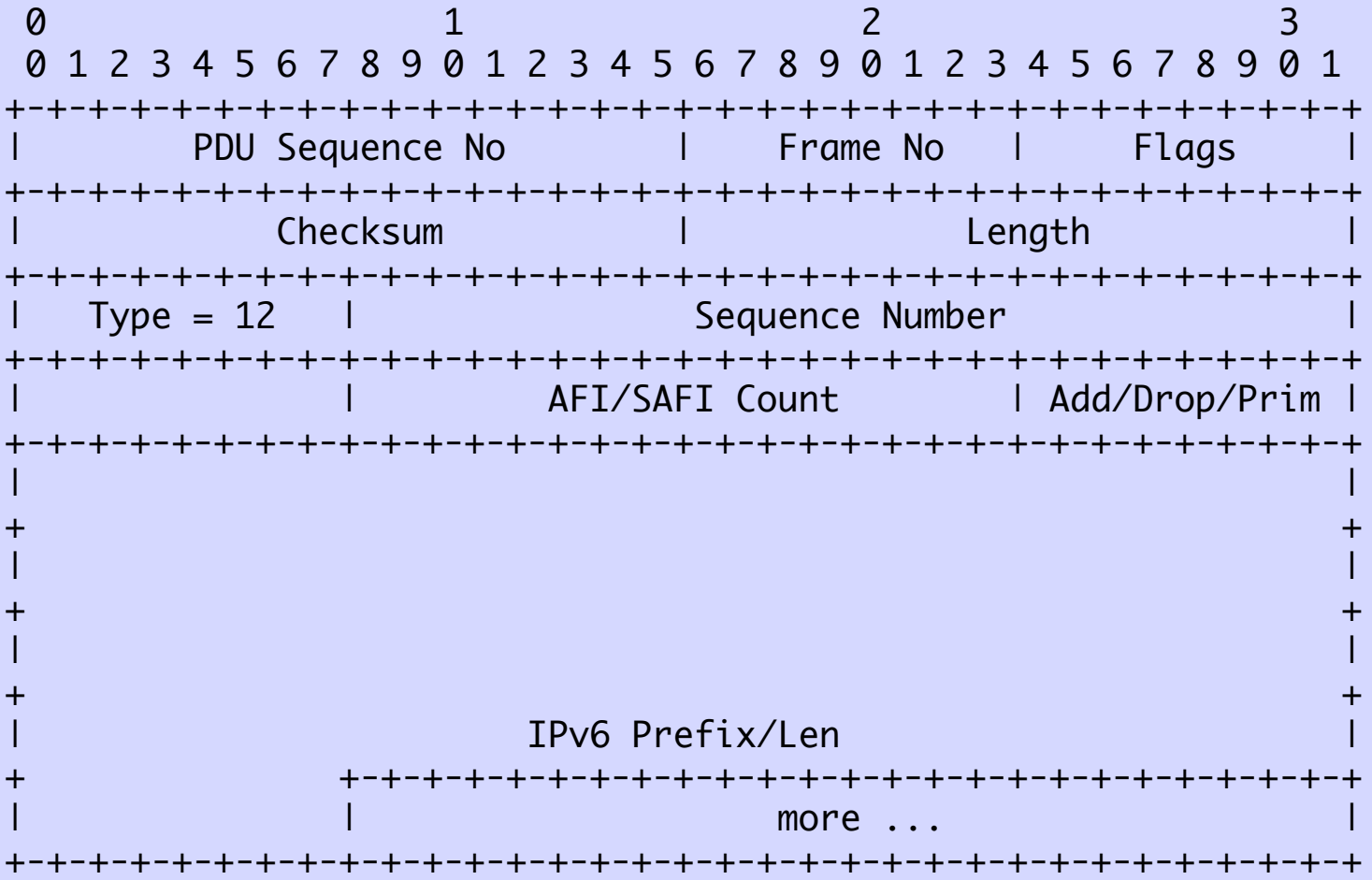
Any alignment freaks in the readership? It gets worse ☺

IPv4 Announce / Withdraw

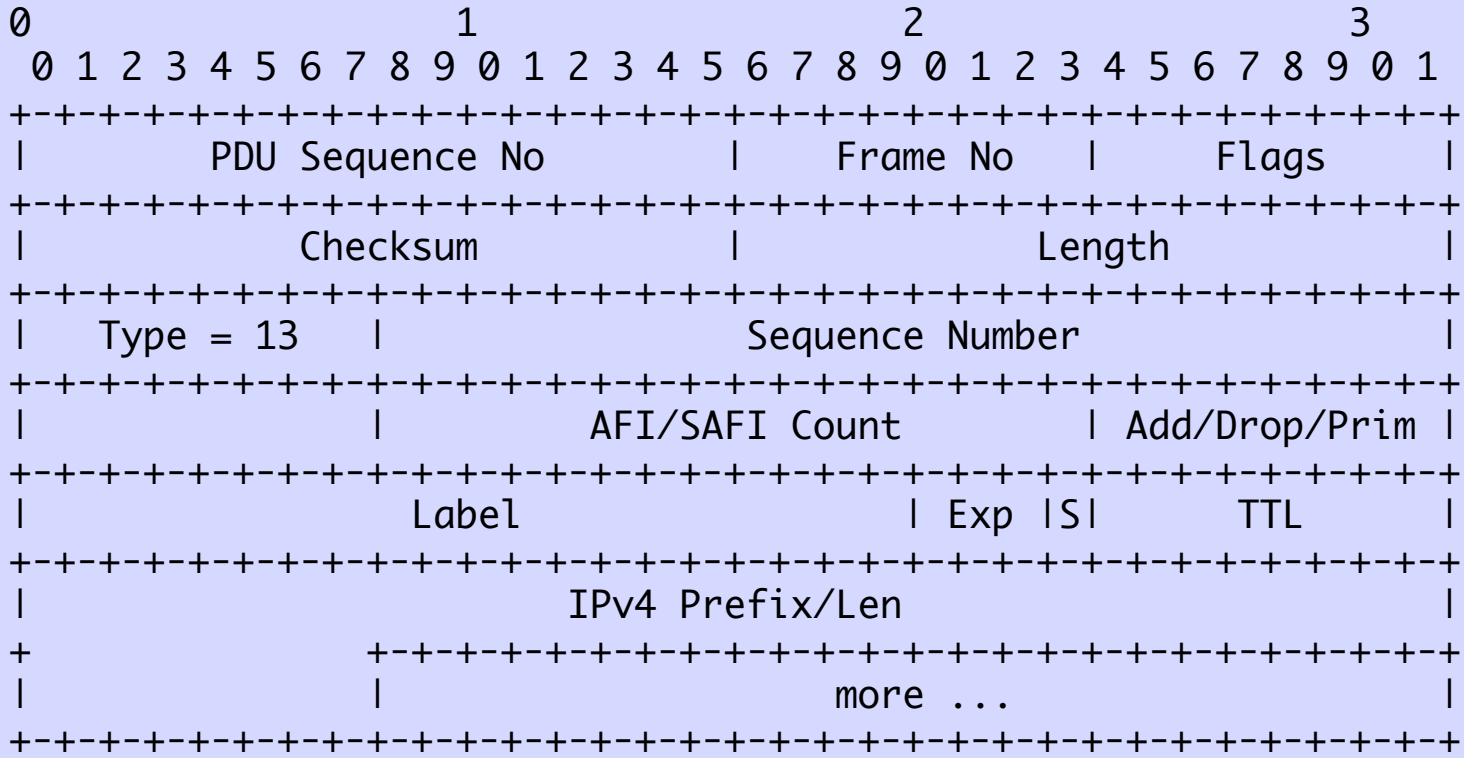


Add/Drop/Prim (bits) - 0 Announce / Withdraw
- 1 Primary of this AFI/SAFI on this Interface
- 2-7 Reserved

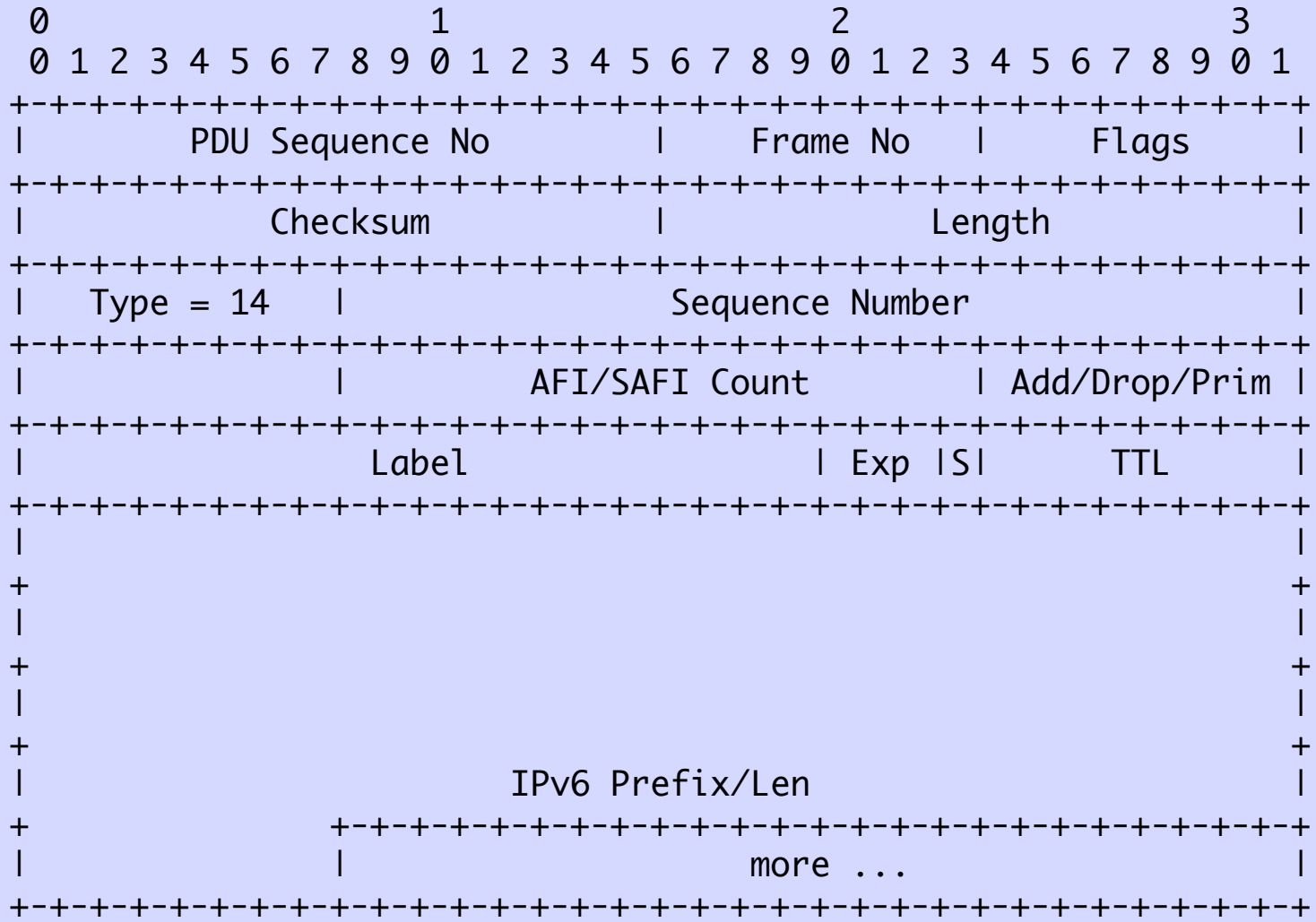
IPv6 Announce / Withdraw



MPLS IPv4 Announce / Withdraw



MPLS IPv6 Announce / Withdraw



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

Now IP/Label Liveness
may be Tested

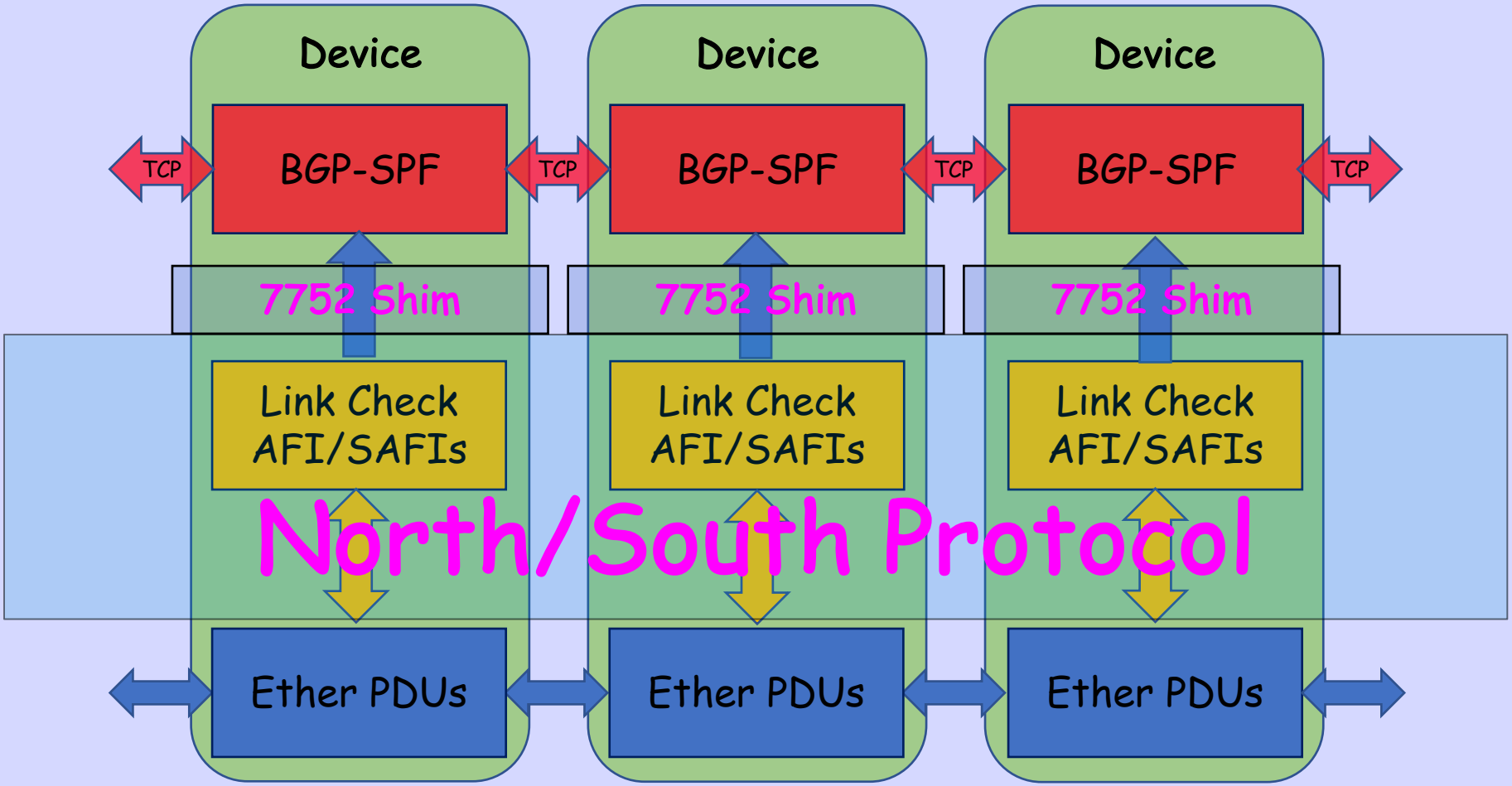
Assume one or more AFI/SAFI
Addresses Will be Used to
Ping, BFD, or Whatever

Routing Done by BGP-SPF

- Thus far, we have a one-hop point-to-point link discovery protocol
- We know what ASNs and AFI/SAFIs are on each Link Interface
- At the Ethernet layer we did not want to do topology discovery and Dijkstra à la IS-IS
- So the link ASNs, AFI/SAFIs, and state changes are passed up the stack to BGP-SPF which discovers the topology, runs Dijkstra, and builds the routing database

We Now Know all Link State of This Device

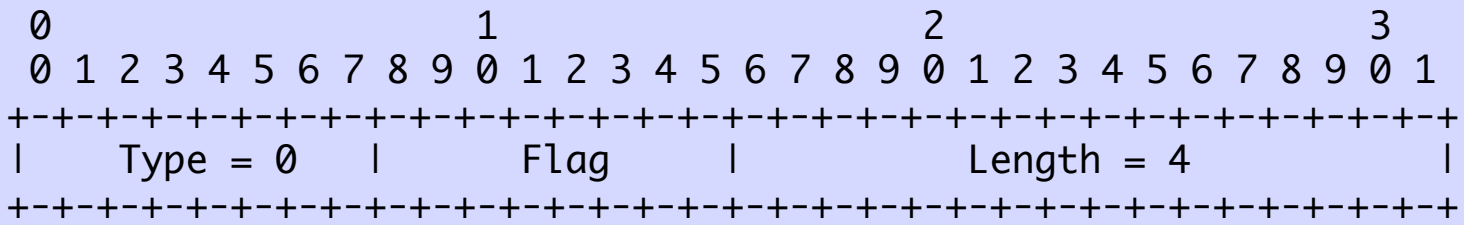
Now Push it Up to Topology and Dijkstra Layers



North/South Protocol

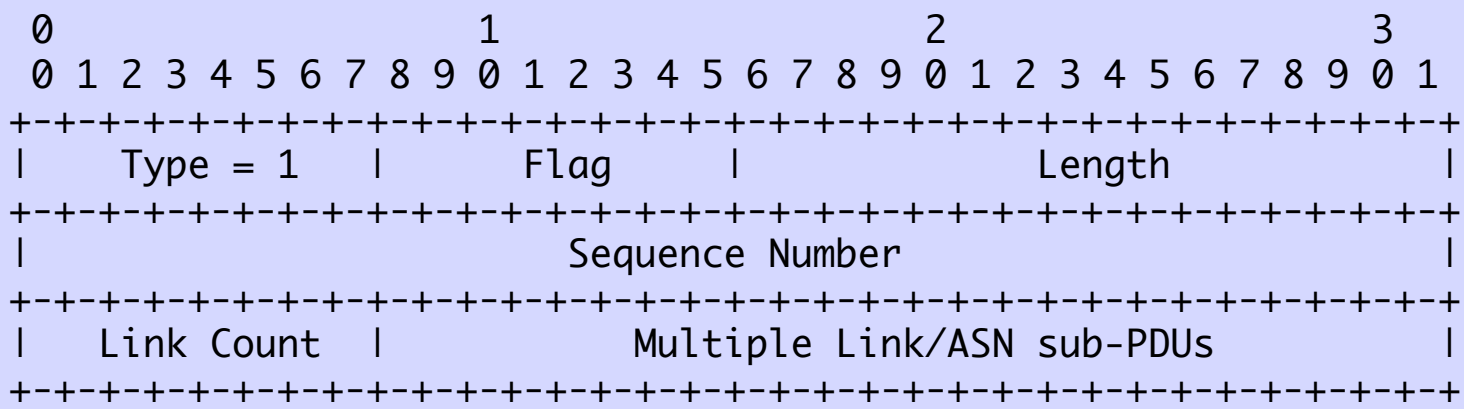
- We assume a reliable intra-device transport, so no ACKs
- We assume a PDU capable of 64k
- The protocol is [re]started by a request from the 7752 Topology Shim Layer
- The Ether Layer then sends the full topology, its full link neighbor state, North
- The Ether Layer sends incremental updates as links and/or addressing change

Topology Request for Full State From 7752 Shim to Link Layer



Flag - 0 – Request Full Ethernet Topology (i.e. restart)

PDU from Link Layer to Shim

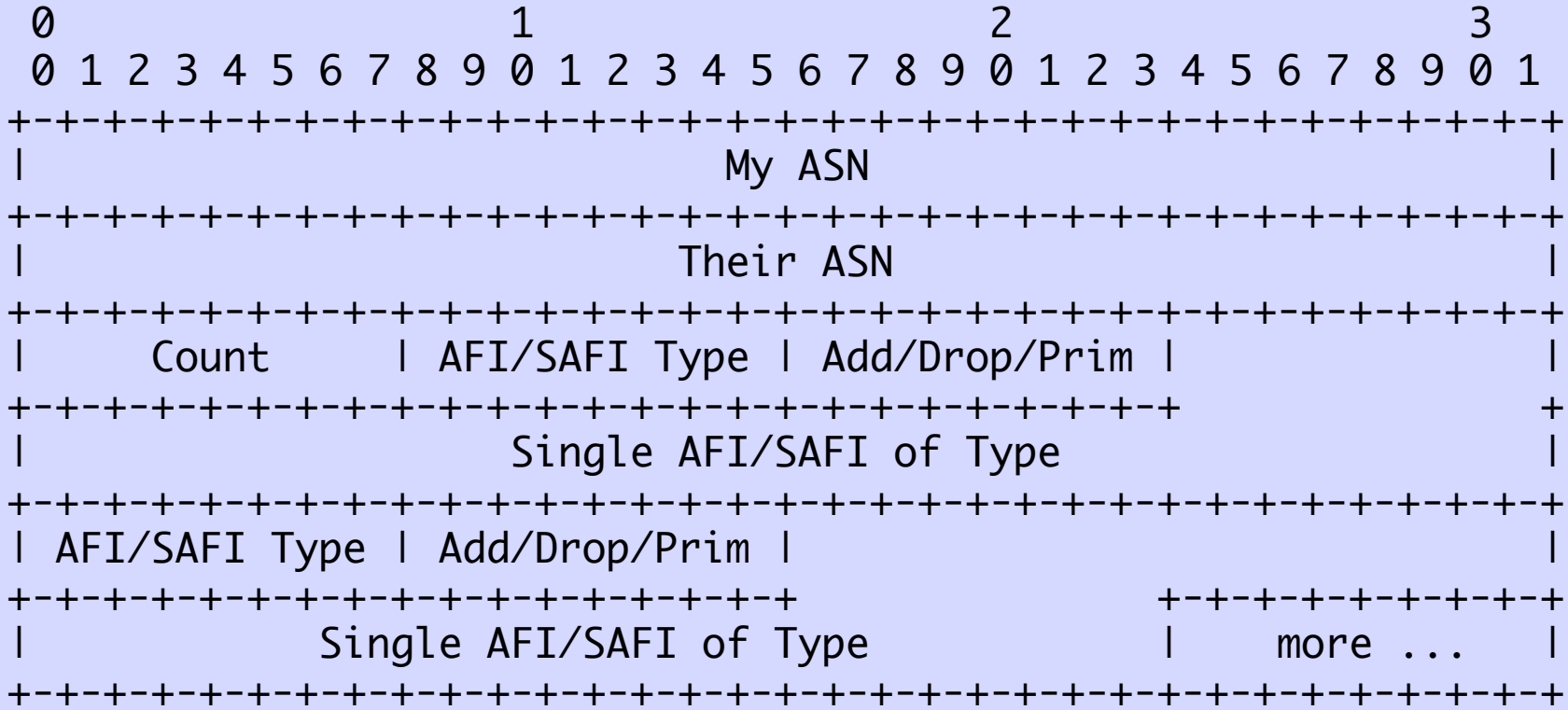


Flag
- 0 – This is the start of a Full State transfer
- 1 – Continuation PDU
- 2 – Last PDU of transfer
- 3 - This is the start of a Update for a state change

Link Count – Number of Link/ASN sub-PDUs to follow

Multiple Link/ASN LSAs, see following

Link/ASN sub-PDU



Count of AFI/SAFIs in this sub-PDU

AFI/SAFI Type - 11-IPv4, 12-IPv6, 13-MPLSv4, 14-MPLSv6, …

Add/Drop/Prim (bits) - 0 Announce / Withdraw
- 1 Primary
- 2-7 Reserved

Addressing and Routing Are Done in Upper Layers

- The 7752 Shim takes the above and translates to BGP-LS/SPF PDUs
- The Topology Layer, BGP-SPF, can now construct the full link state and IP/MPLS topology of the network
- Routing protocols such as BGP-SPF can then talk between devices to exchange reachability

And Bob's Your Uncle