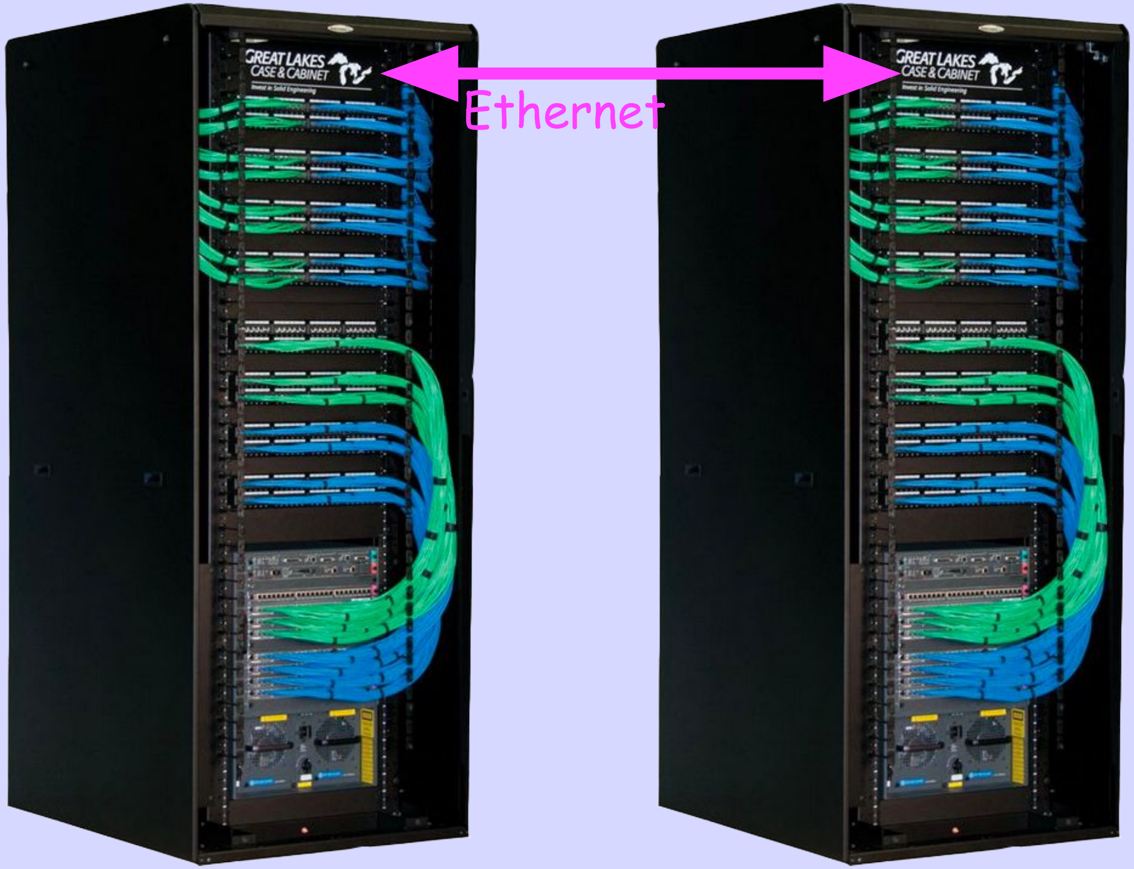
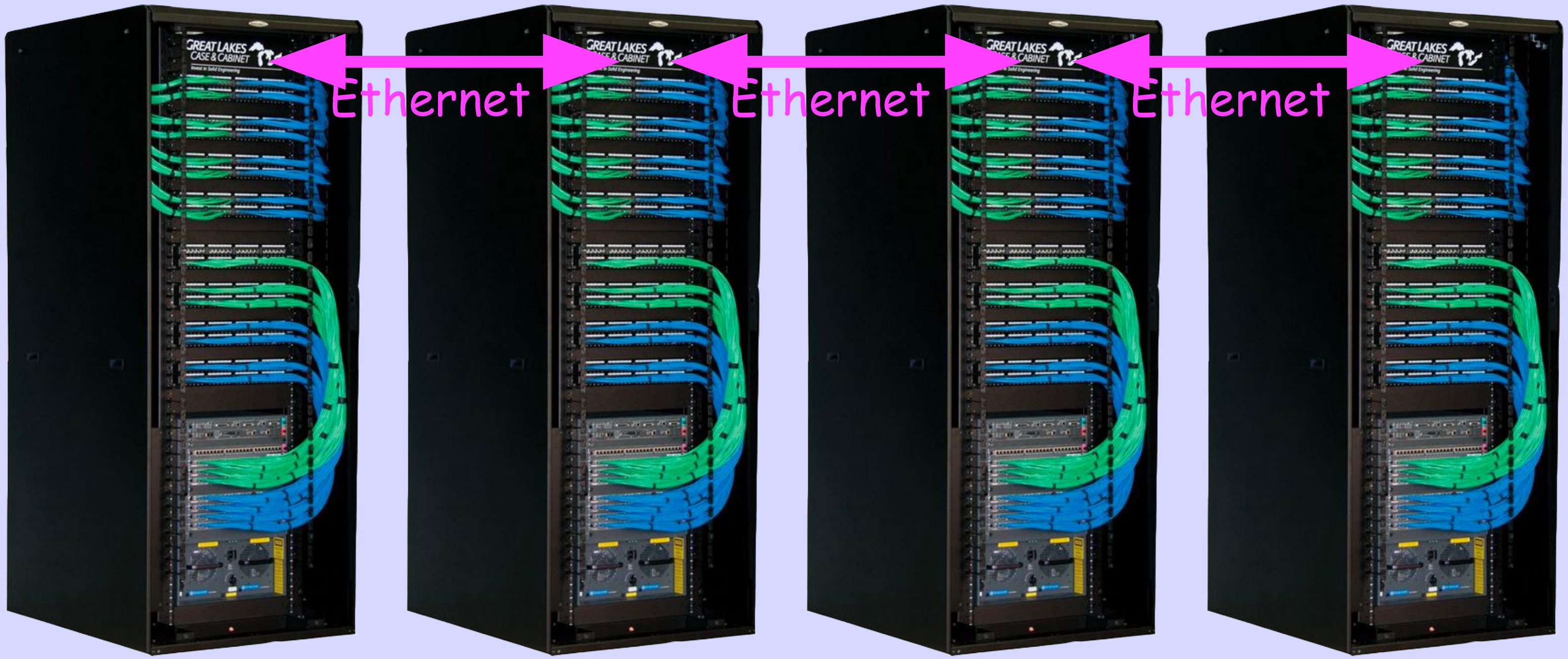
An Approach to Routing in a Clos

Randy Bush <randy@psg.com>
Arrcus & IIJ Research

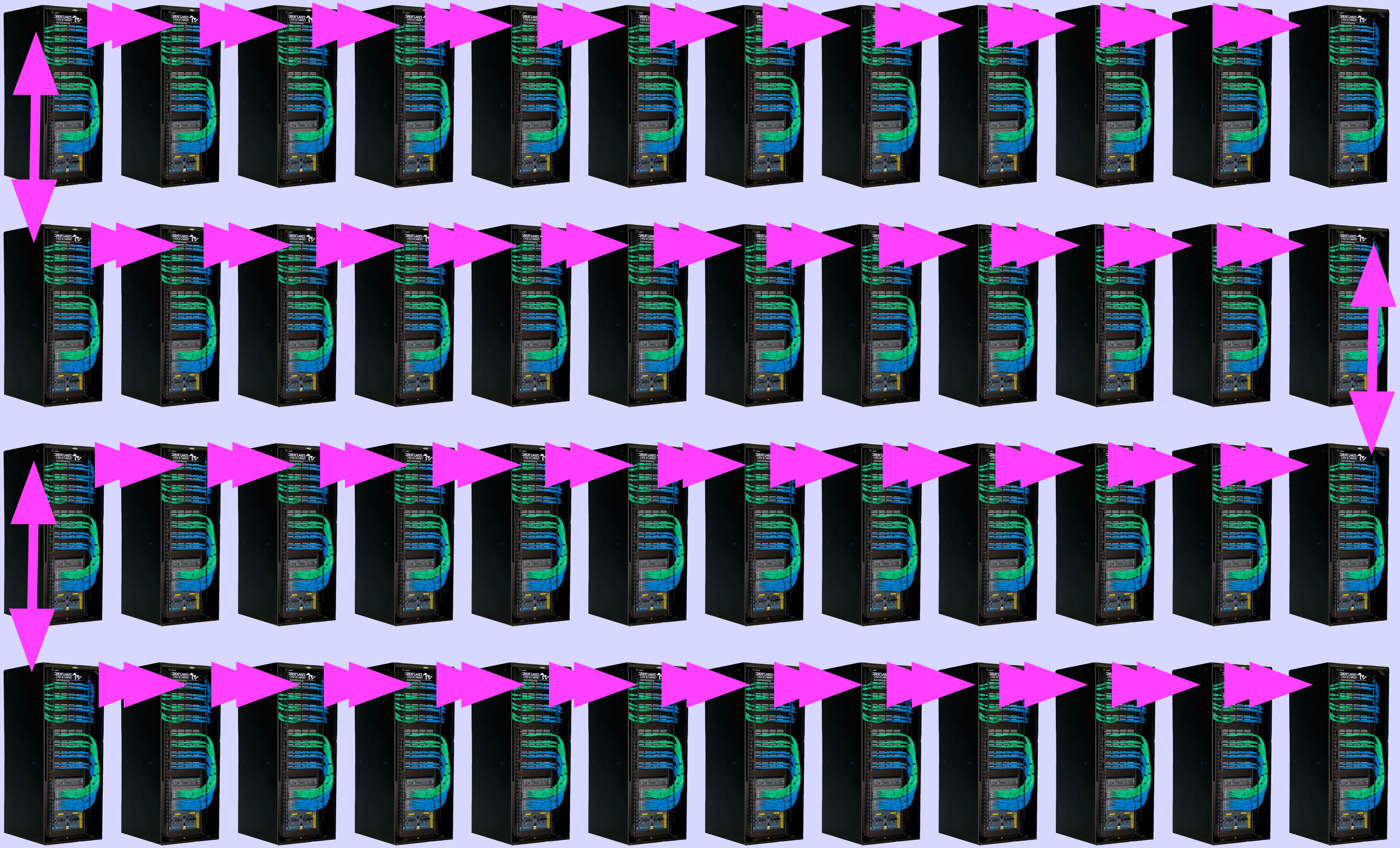
This Works



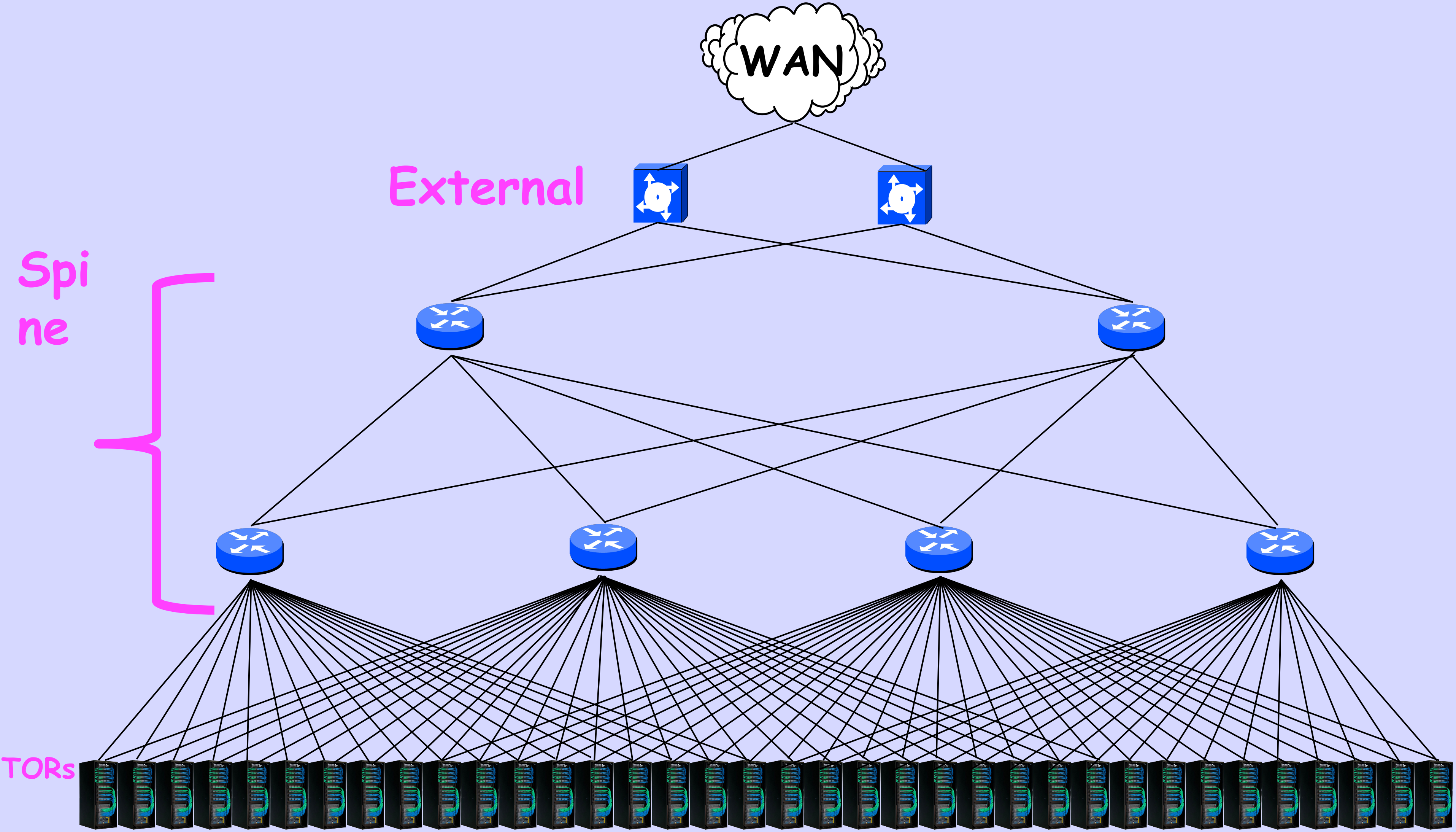
This Might Work



This Won't Work



This Works (Clos Network)



Clos is Not an Acronym

Clos, Charles (Mar 1953)
"A study of non-blocking switching networks"
Bell System Technical Journal. 32 (2): 406-424

For Example:

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

How Do You Route In Something of This Scale?

OSPF OK to 500 Nodes

IS-IS good to 1,000

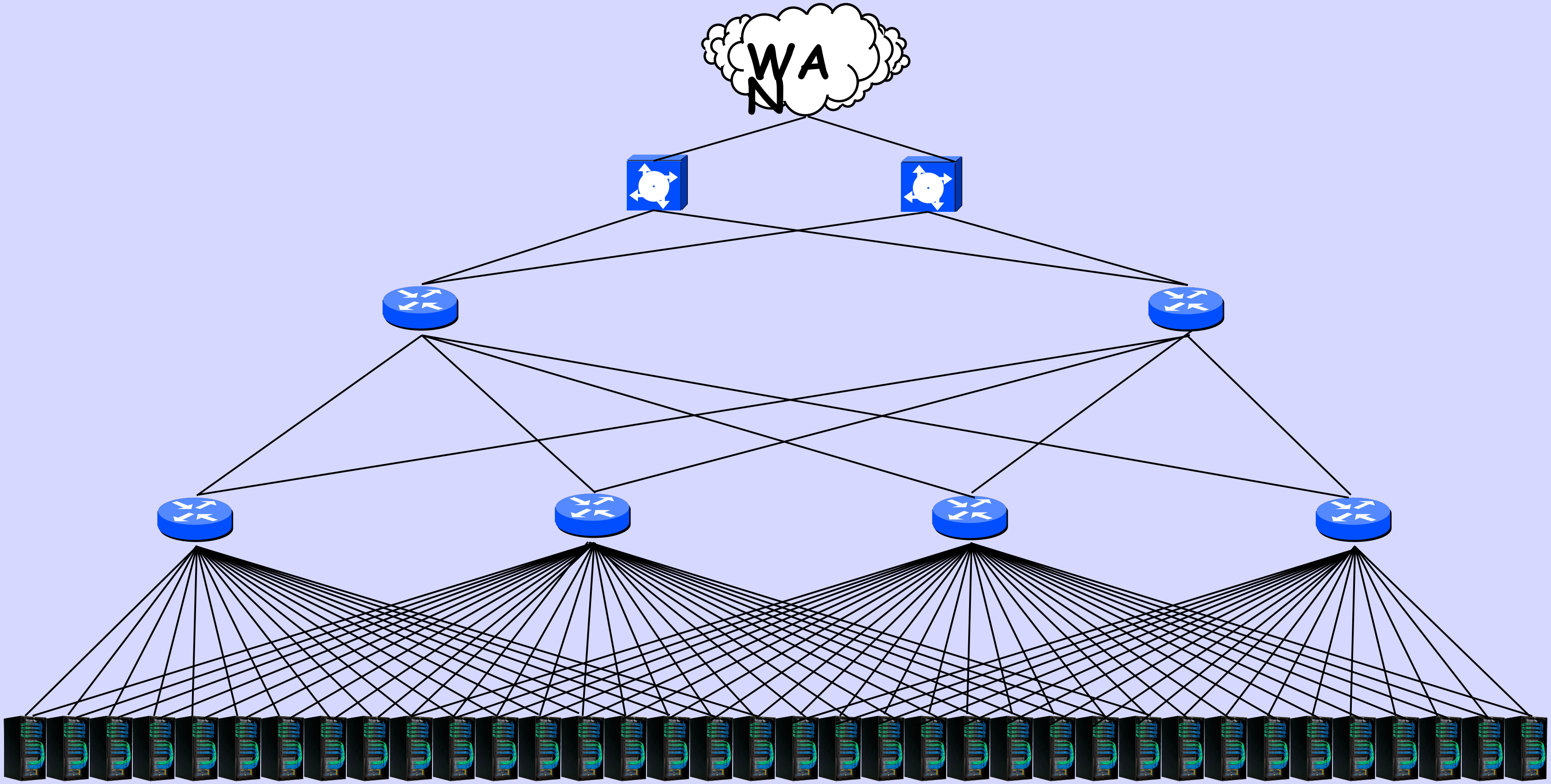
Limited Because They Repeatedly Flood Everything

Your Clos on IS-IS or OSPF

BGP Scales Because It Signals Only Changes

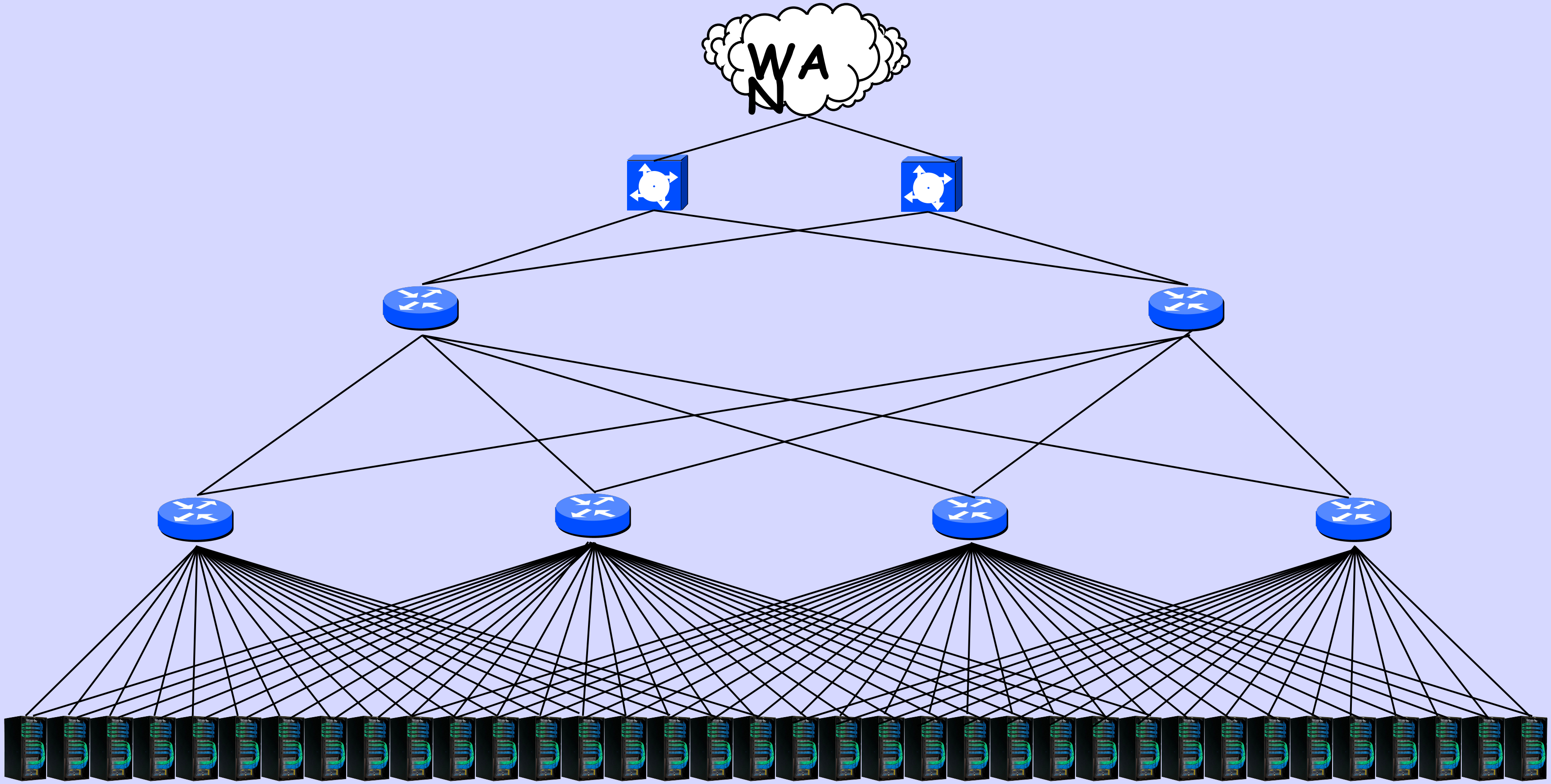
So BGP has become common in MSDCs

BGP is Quiet as Updates are Infrequent



ECMP can be Very Wide

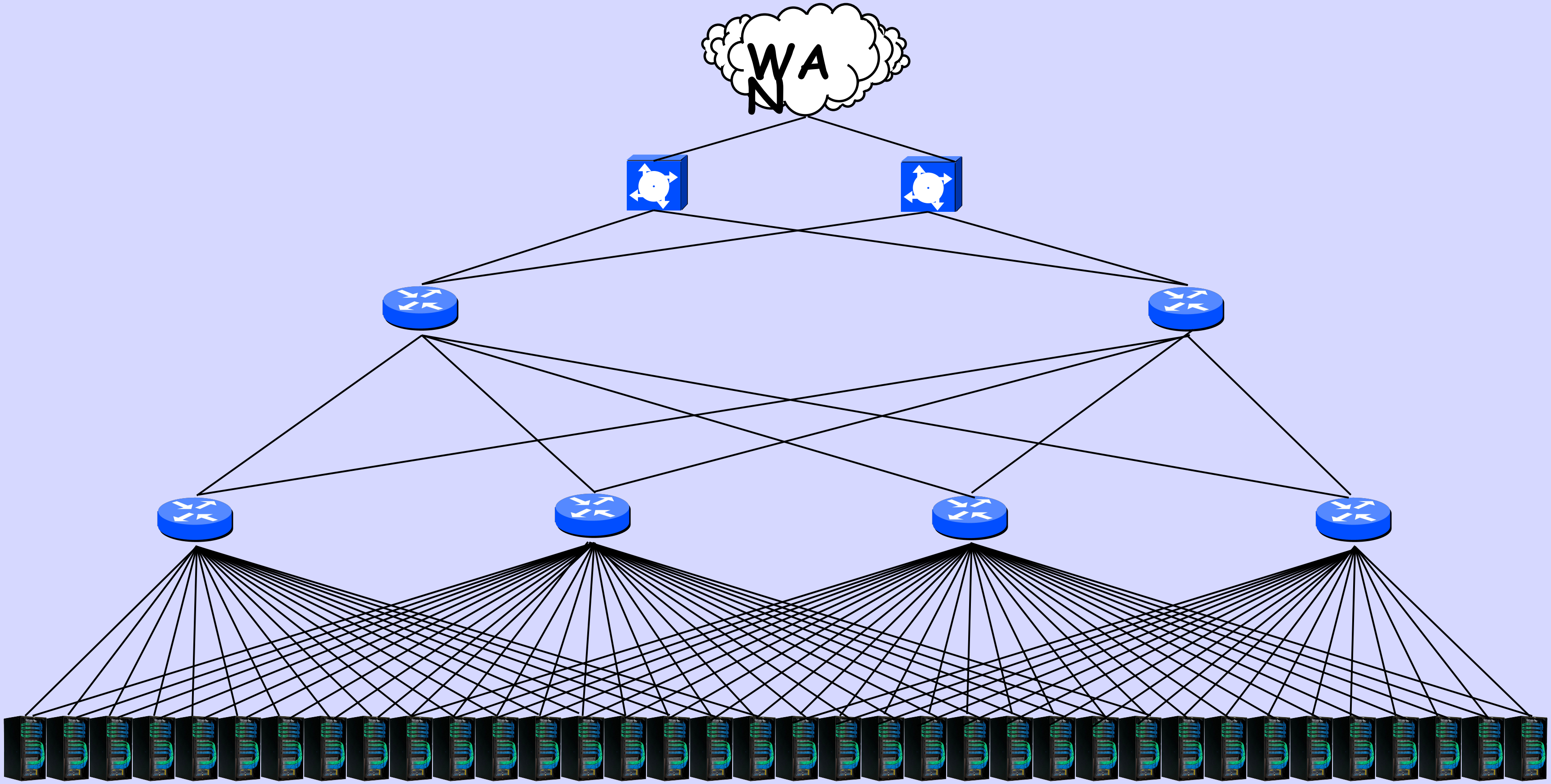
32, 64, even 128



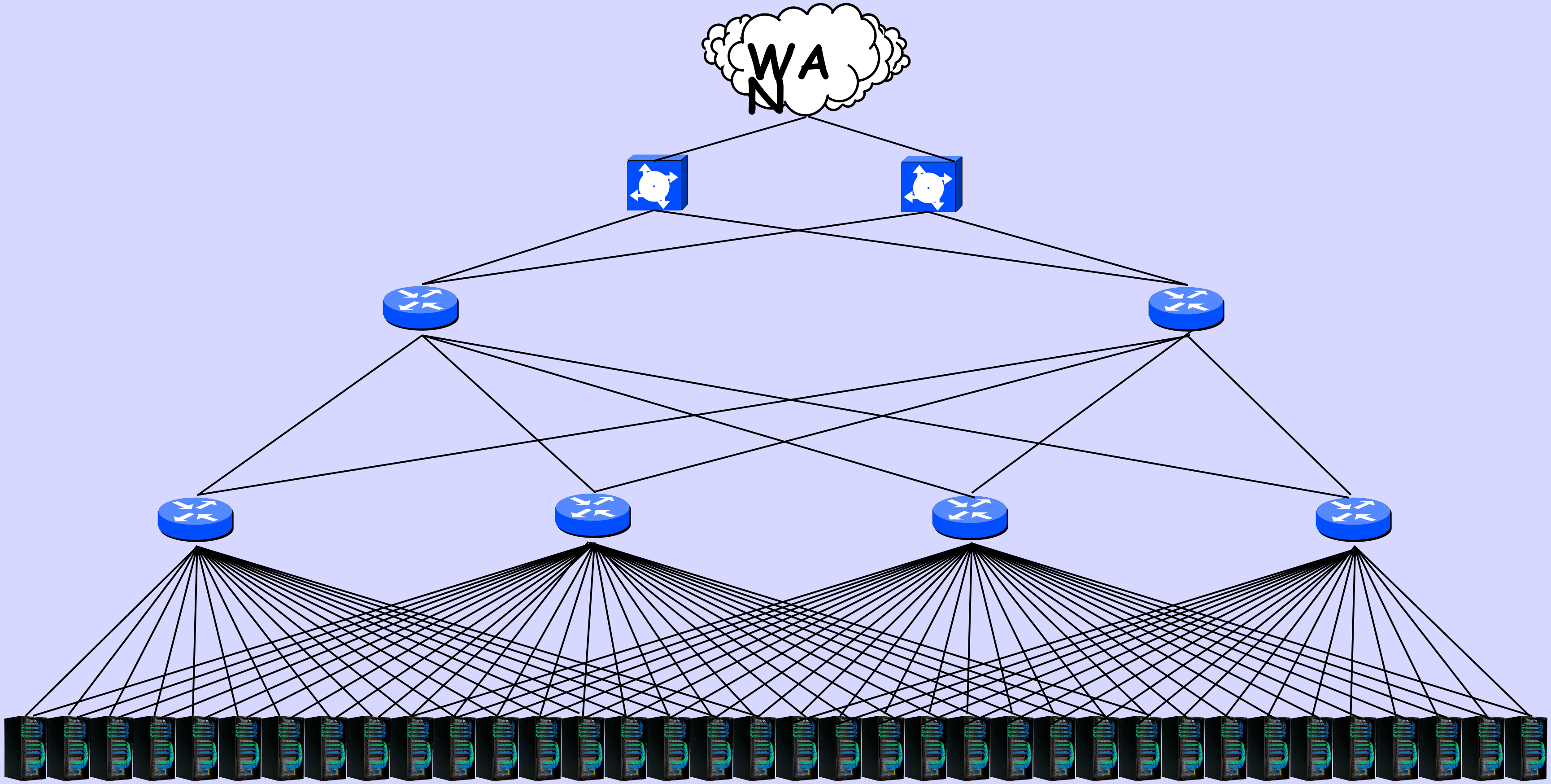
BTW, Every Rack is (often) an AS

Get Over It

But What is the Decision Process?



Do You Want to Write BGP Policy for Massive ECMP?



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

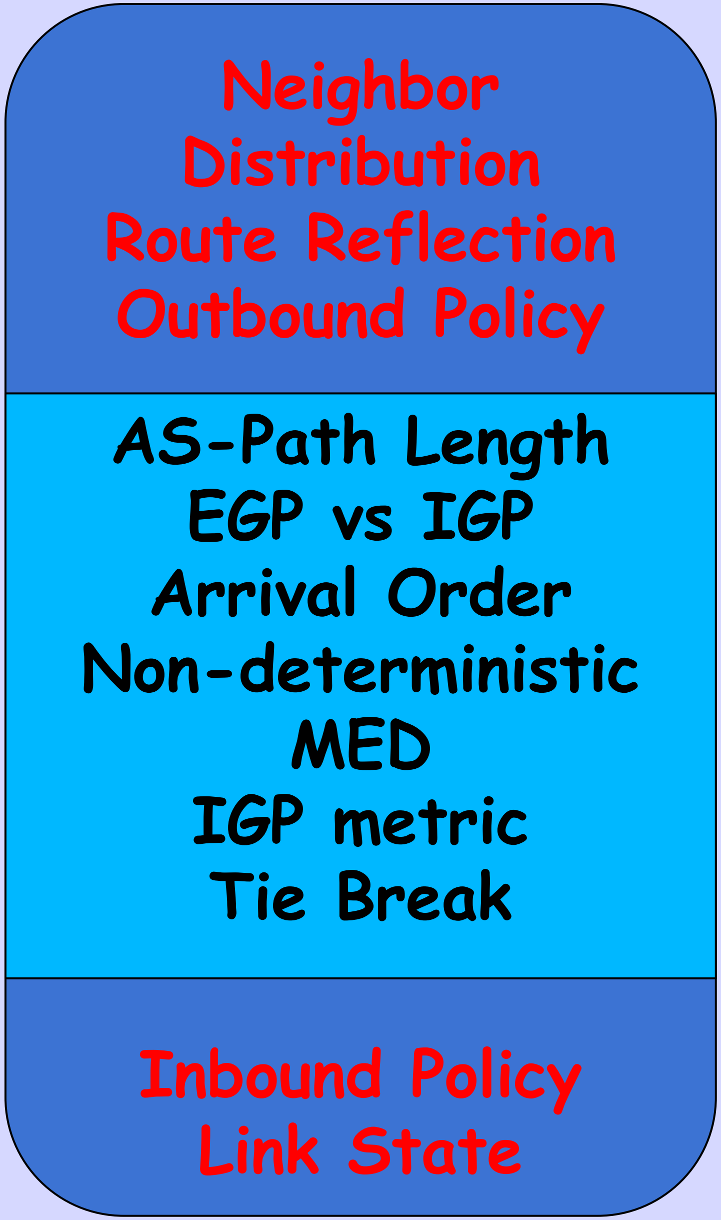
SPF?

I thought BGP was path vector, not link state!

s/Best Path/SPF/

- New SAFI
- NLRI format exactly same as BGP LS (RFC 7752) Address Family to carry link state information
- BGP runs Dijkstra instead of Best Path Decision process
- BGP MP (new SAFI) and BGP-LS Node attribute for compatibility
- Peering Models: eBGP, iBGP, RR

BGP4
Classic

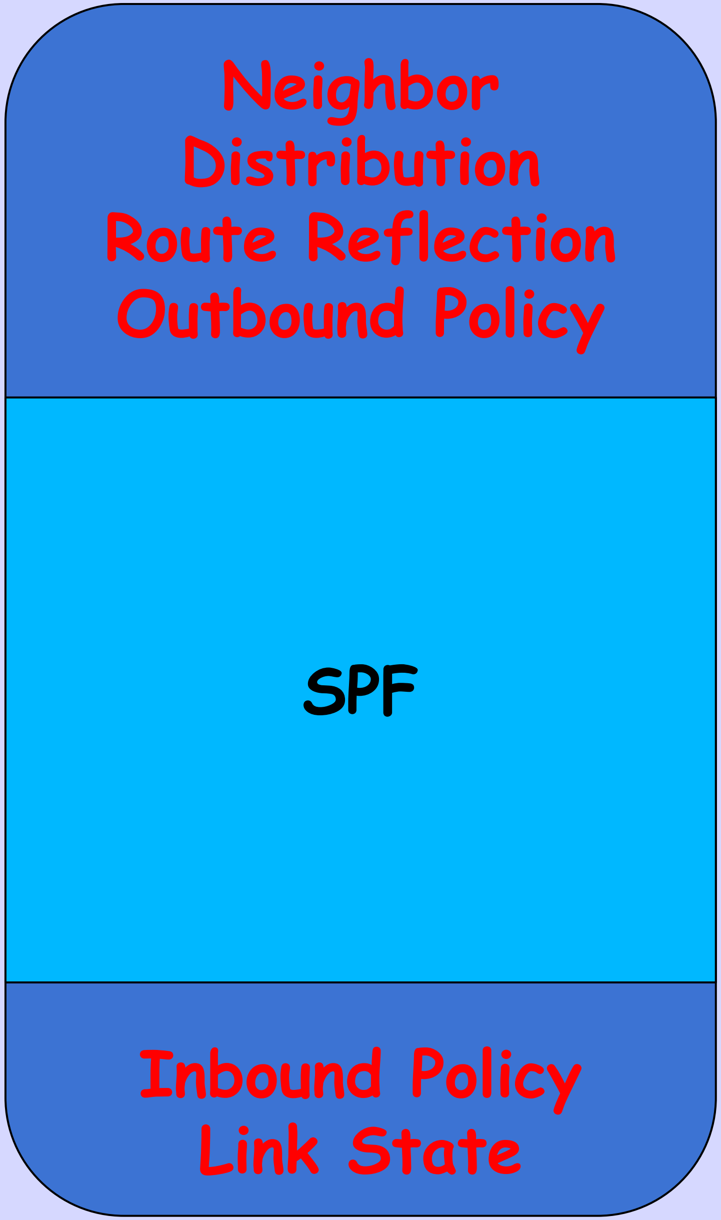


Neighbor
Distribution
Route Reflection
Outbound Policy

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

Inbound Policy
Link State

BGP-SPF



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

Removed!

BGP-SPF

- Next-Hop and Path Attributes come for free with BGP Link-State Address Family
  - Needed for RFC 4271 error handling
- Decision Process Phases 1 and 2 (best path) replaced by SPF algorithm (AKA Dijkstra)
- Decision Process Phase 3 (tie break) may be skipped as NLRI is unique per BGP speaker
- Need to assure the most recent version of NLRI is always used and re-advertised
  - Augmented with sequence numbers

BGP-SPF

- Starting with greatly simplified SPF with P2P only links in single area (i.e., SPT)
- Should scale very well to many use cases
- Could support computation of LFAs, Segment Routing SIDs, and other IGP features
  - BGP-LS format includes necessary Link-State
- Link-State AF is dual-stack AF since both IPv4 and IPv6 addresses/prefixes advertised
  - BGP-LS format also supports VPNs but SPF behavior not defined
  - Work needed to define interaction with existing unicast AFs
    - Matter of local implementation policy

Peering Model

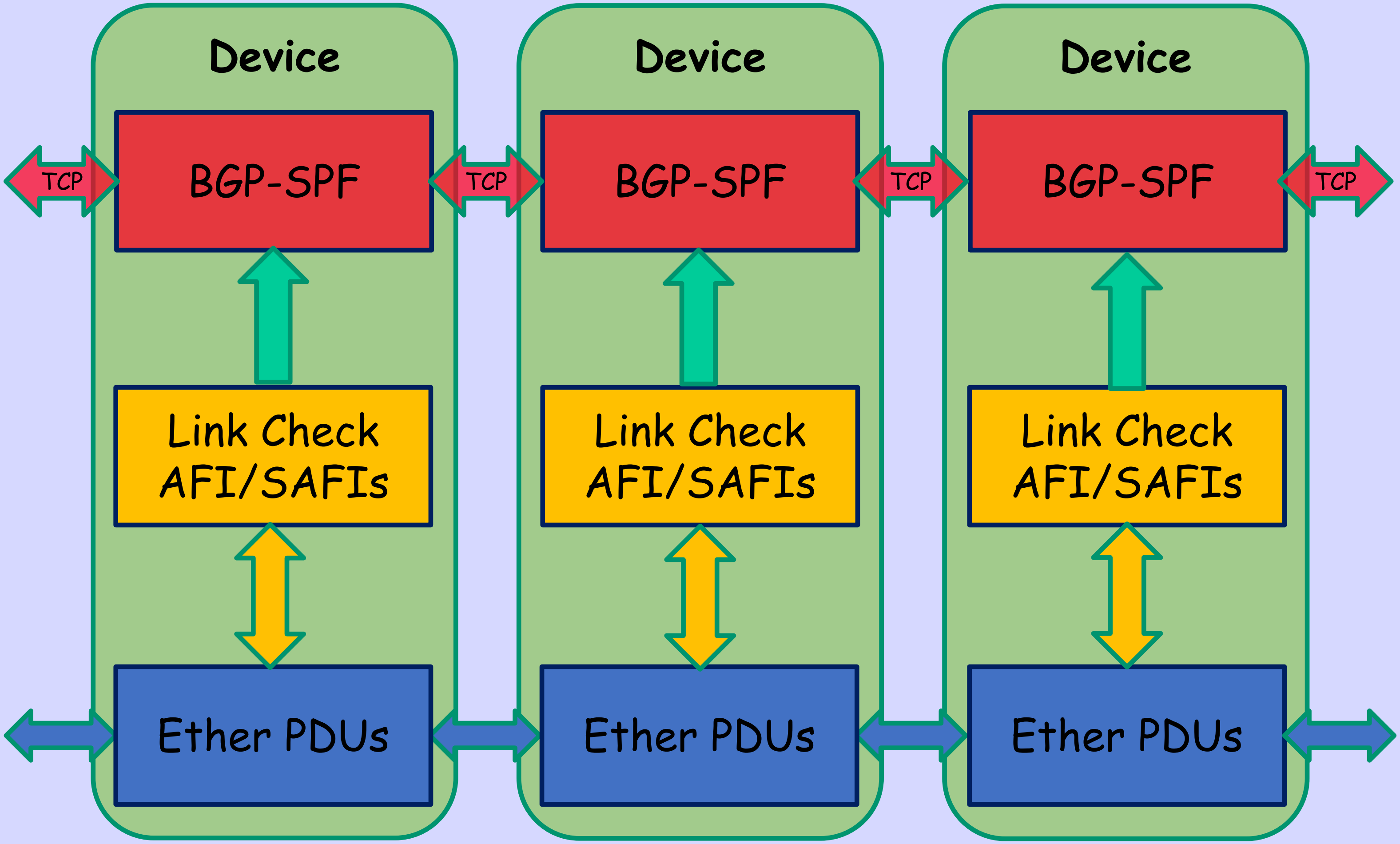
- BGP sessions, optionally with Route-Reflector or controller hierarchy
  - Link discovery/liveliness detection outside of BGP
- RR hierarchy can be less than fully connected but must provide redundancy
  - Must not be dependent on SPF for connectivity
- Controller could learn the expected topology through some other means and inject it
  - SPF Computation is distributed though
  - Similar to "Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenter Network"

How Does BGP-SPF Learn Link State so it can Build the Topology?

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

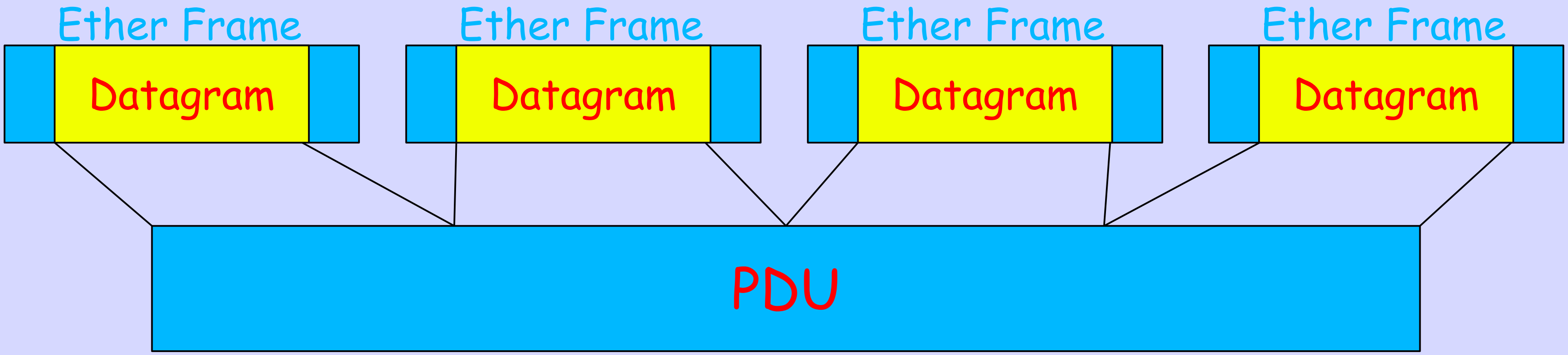


MAC Link State exchanged over raw Ethernet and pushed up stack

Add the AFI/SAFI data IP-Level Liveness Check

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

```
 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    | Transmission Sequence Number |L|  Datagram    ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    Number     |        Datagram Length         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                 Checksum                       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                 Payload...                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```



```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   PDU Type    |           Payload Length            ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~               |         Payload ...            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   Sig Type    |     Signature Length          |         ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+         +
~                    Signature                   ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Why not Just Use TCP?

- When this runs, there are no IP Addresses
- This protocol is to Learn IP Addresses
- So it is a cheap TCP-like protocol
- Reassembly of out of order Datagrams
- Retransmission with Back-off
- PDUs are ACKnowledged
- …

Fully Stateful Session Per Peer

Graceful Restart

State May Be Resumed
á la BGP

Encaps etc PDUs

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   PDU Type    |              Payload Length           ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~          |               Count              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                 Serial Number                  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|          Encapsulation List...           |   Sig Type   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|       Signature Length      |       Signature ...      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

OPEN PDU

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  PDU Type = 1 |              Payload Length              ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~               |                  Nonce                  ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~               |  LLEI Length  |          My LLEI          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-~
~                               |   AttrCount   |           ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~      Attribute List ...       |   Auth Type   |  Key Length  ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~               |                  Key ...                  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                         Serial Number                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   Sig Type    |       Signature Length       | Signature ... |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Announce/Withdraw

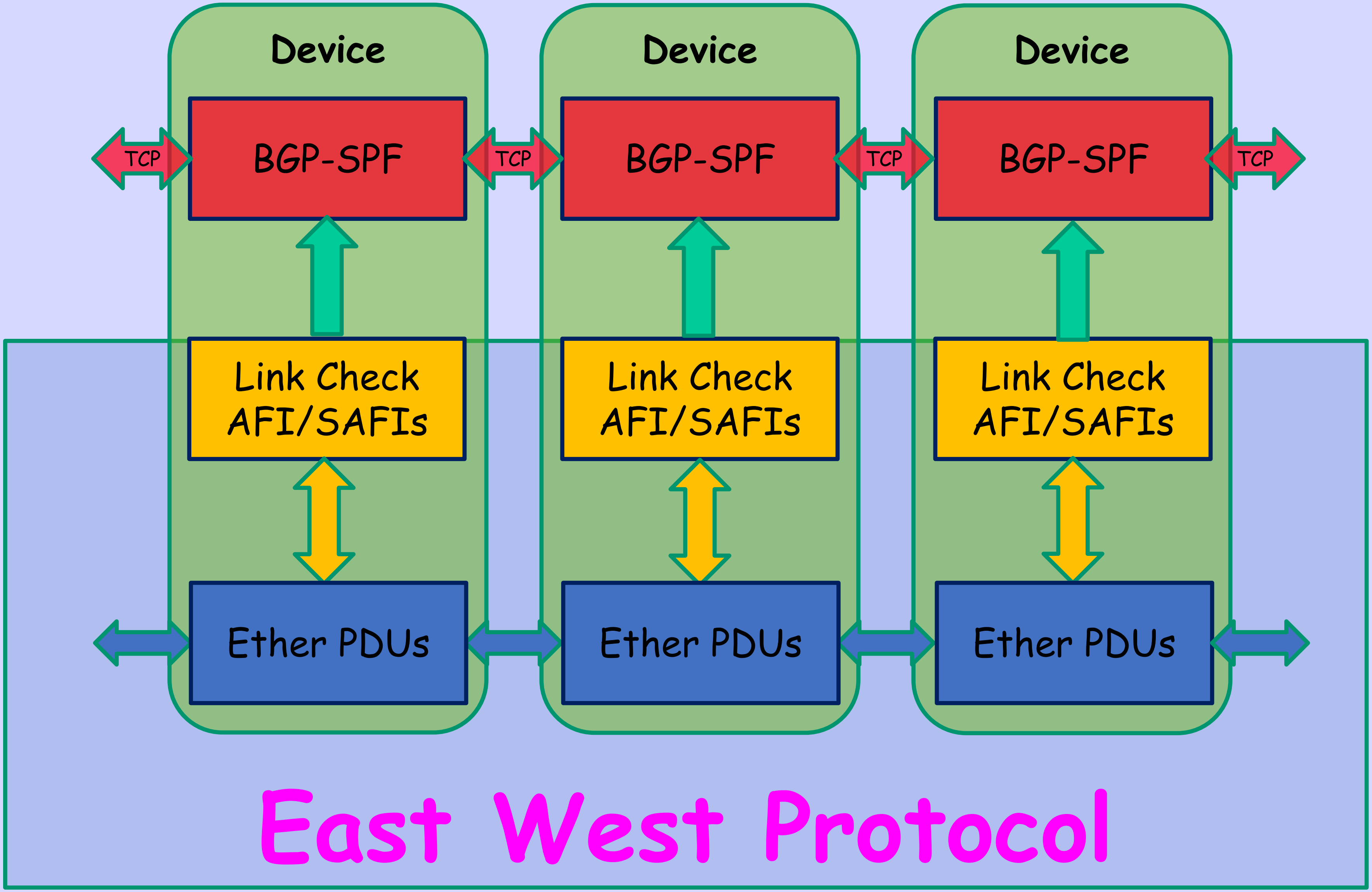
```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  PDU Type = 4 |             Payload Length            ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~             |                  Count                  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      Serial Number                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Encaps Flags  |             IPv4 Address              ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~             |  PrefixLen  |   more ...   |  Sig Type  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|       Signature Length       |       Signature ...      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

```
 0          1          2          3          4  ...      7
+-----------+-----------+-----------+-----------+-----------+
|  Ann/With  |  Primary  | Under/Over |  Loopback  | Reserved ..|
+-----------+-----------+-----------+-----------+-----------+
```

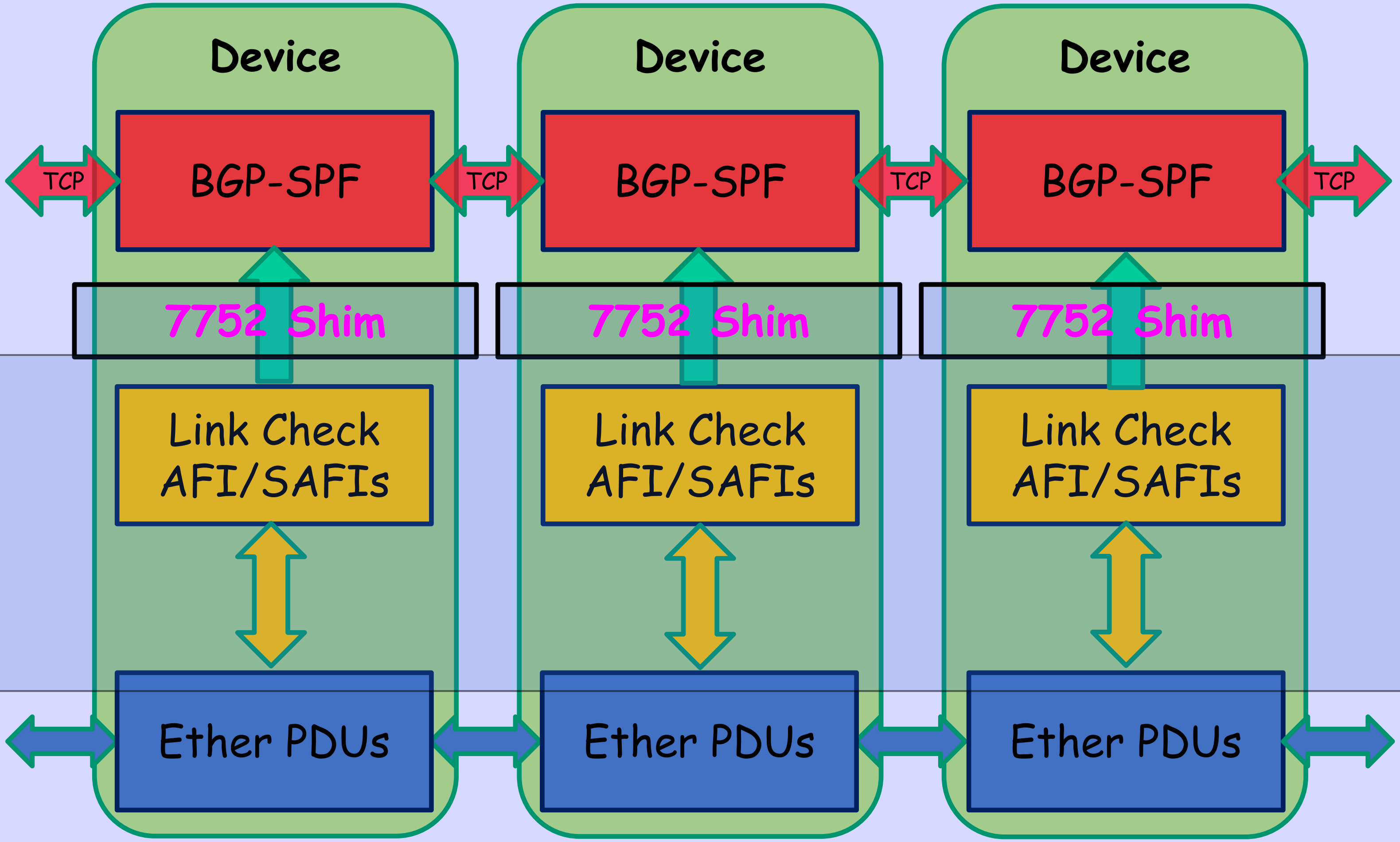
Explicit ACK/ERROR

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  PDU Type = 3 |            Payload Length = 5            ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~               |  ACKed PDU   | EType |       Error Code      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|          Error Hint           |    Sig Type   |Signature Leng.~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~               |              Signature ...                   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

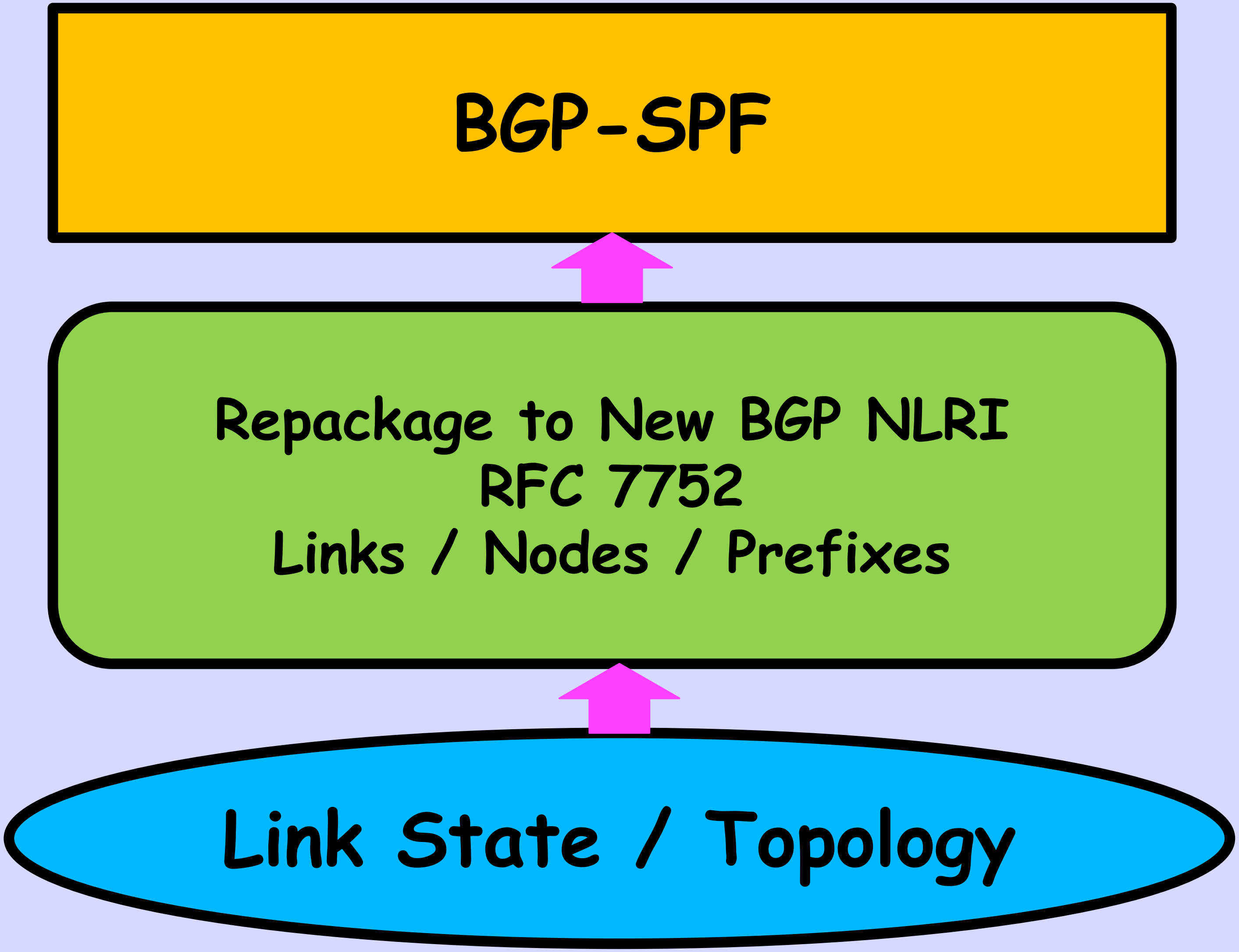
EType, Error Code, Error Hint = 0   /* no error, just an ACK */


Device
Device
Device
TCP
BGP-SPF
TCP
BGP-SPF
TCP
BGP-SPF
TCP
Link Check
AFI/SAFIs
Link Check
AFI/SAFIs
Link Check
AFI/SAFIs
Ether PDUs
Ether PDUs
Ether PDUs
East West Protocol

North/South Protocol



BGP-LS for BGP-SPF



BGP-SPF

Repackage to New BGP NLRI
RFC 7752
Links / Nodes / Prefixes

Link State / Topology

How Does BGP-SPF Start?

- For BGP-SPF to build topology and state, need to peer with BGP-SPF neighbors
- But we do not want to configure it more than necessary
- Why not extend L3DL to pass BGP config data to my peer?

L3DL-ULPC

Upper Layer Protocol Configuration

draft-ymbk-lsvr-l3dl-ulpc

L3DL PDU for ULPC

```
 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 = 9   |               Payload Length                  ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~               |   ULPC Type   |   AttrCount   |               ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~       Attribute List ...      |    Sig Type   | Signature Len ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~               |           Signature ...            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Provide the minimal set of configuration parameters for BGP OPEN to succeed

Not to replace or conflict with data exchanged by BGP OPEN

Multiple sources of truth are a recipe for complexity and pain

ULPC for BGP

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Attr Type = 1 | Attr Len = 48 |         My ASN          ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Attr Type = 2 | Attr Len = 56 |  My IPv4 Peering Address      ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                               |  Prefix Len   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Attr Type = 4 |   Attr Len    |                    ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+                        ~
~           BGP Authentication Data ...                 ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

BGP Can Now OPEN

And that will exchange all the rest

There is Running Code

- Open Source Python3 for LSOE, an early version of L3DL
- GoLang Source for current L3DL. We hope to be allowed to open source

BTW, There is No IPR in these Standards Proposals