RIFT
Routing in FAT Trees

ENOG15

Jeff Tantsura
Head of Technology Strategy Nuage Networks
IAB, IETF Routing & RIFT WG's chair

Routing in DC - some history

- 1990-2014 - DC's are L2, EIGRP/OSPF @L3
- 2010-2014 – MSDC's move to L3, first try in BGP adoption
- 2010-2015 (now)
  - Amazon - (OSPF/BGP + black magic)
  - Google develops Firepath (gRPC overlay)
  - FB develops OpenR (BGP and THRIFT overlay)
- IETF
  - 2012 - Petr Lapukhov publishes draft-lapukhov-bgp-routing-large-dc
  - After 4 years in limbo, RTGWG adopts the draft and publishes RFC7938, used by 100s of companies to implement BGP in DC
  - 2015 – RTGWG starts Routing in DC effort, 2017 initial version of requirements has been published
  - 2016 - Number of drafts, modifying OSPF/ISIS flooding have been published
  - 2016 - RIFT and BGP-SPF drafts are published
  - 2017 – Routing in DC BoF @IETF 100 and as the result 2 new WG formed:
    - RIFT - Routing in FAT TREES
    - LSVR - Link State Vector Routing

Enterprise reality - 2017 - quite some work to do

What routing protocols do you use in your network? Please select all that apply.		Back to top
OSPF	8	72.73%
EIGRP	5	45.45%
BGP	6	54.55%
Other	2	18.18%
# of people who answered question	11	

If you marked "Other", please let us know what other routing protocol you use.		Back to top
Individual Responses	2	18.18%

DC Routing protocol requirements

Jeff Tantsura
Dmitry Afanasiev
Keyur Patel
Petr Lapukhov
Tony Przygienda
Russ White
Yingzhen Qu
Jim Uttaro
Kenji Kumaki

Why DC napkin protocol design team?

Because we are long time friends ☺

Why DC napkin protocol design team?

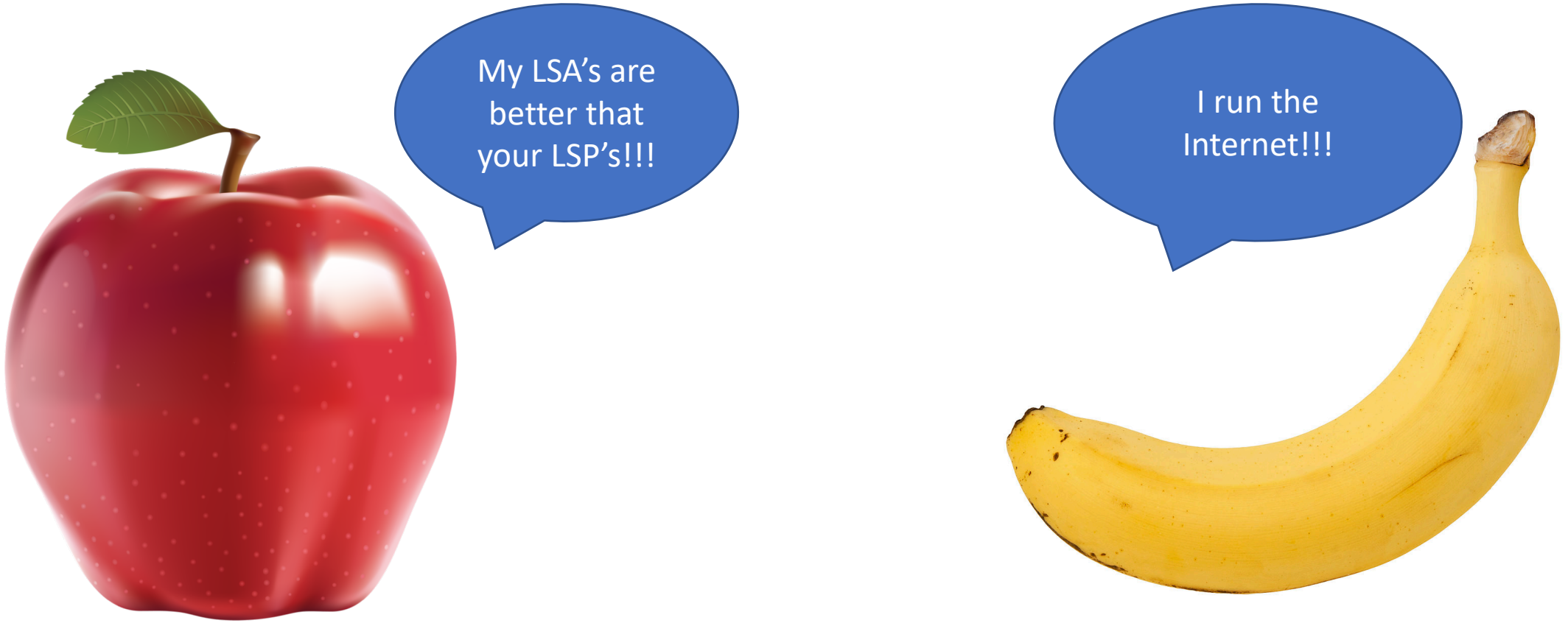
Seriously

- We know how to build routing protocols and DC's

Why DC routing protocol req's draft?

Why DC routing protocol req's draft?

Avoid protocol beauty contest - Have a single set of requirements to be compared against

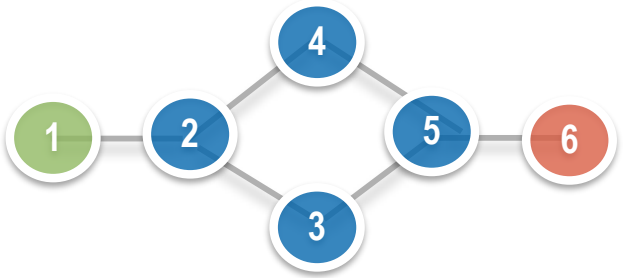


Why DC routing protocol req's draft?

We are just starting – we need your help!

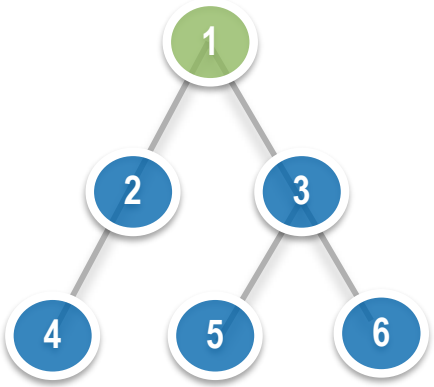
ROUTING PROTOCOLS IN OUR NETWORKS

Vectors of destination and distance
"Tell your neighbors rest of the network"



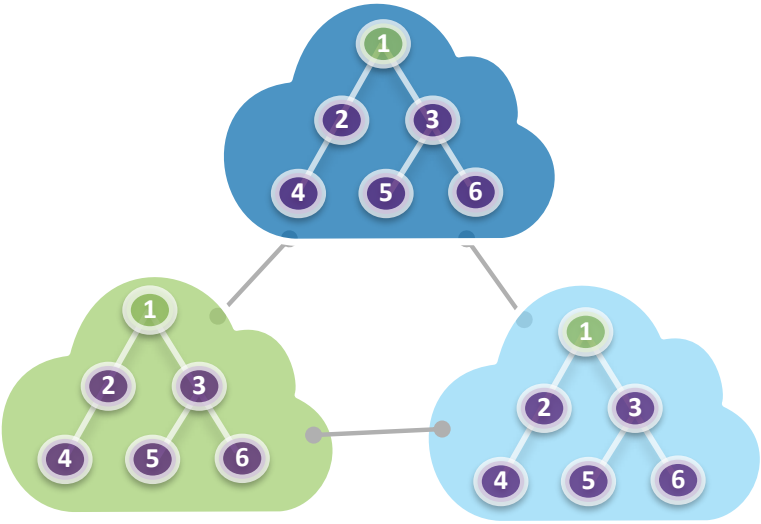
Distance-Vector
RIP

Router Announced LSDB, Dijkstra
"Tell rest of the network your neighbors"



Link-State
OSPF/ISIS

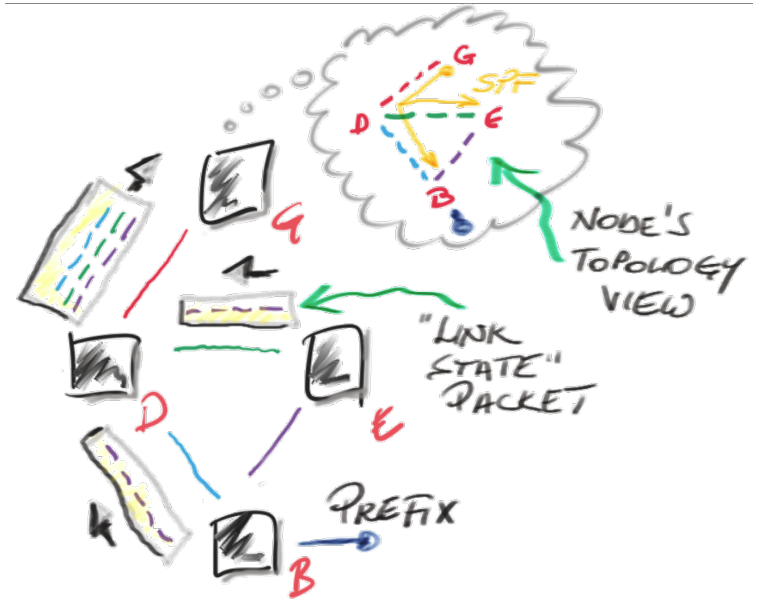
Full-paths announced in BGP.
Paths described by sequence of ASs



Path-Vector
BGP

Link State and SPF = Distributed Computation

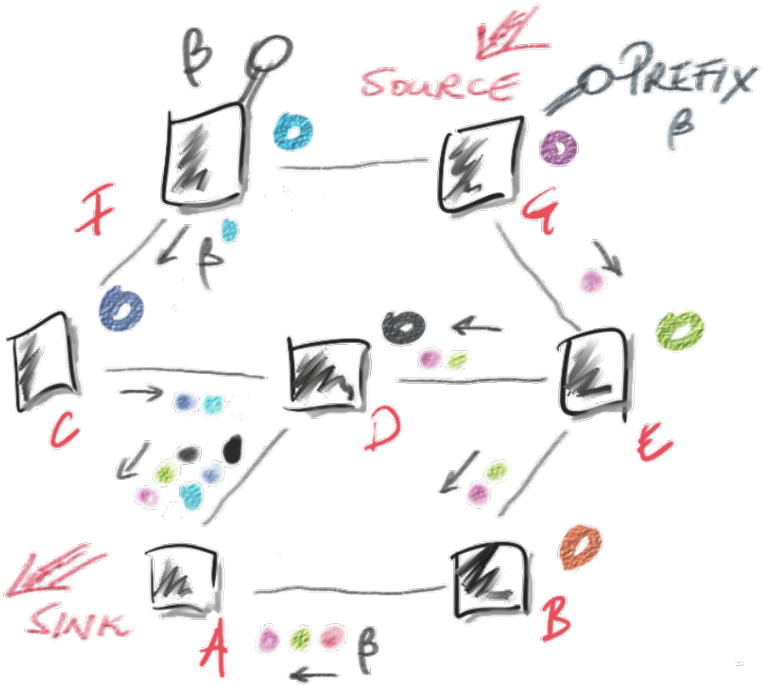
- Advantages
  - Topology elements - nodes, links, prefixes
  - Each node originates packets with its elements
  - Packets are ”flooded” across the network
  - ”Newest” version wins
  - Each node “sees” whole topology
  - Each node “computes” reachability to everywhere
  - Conversion is very fast
- Disadvantages
  - Every link failure shakes whole network
  - Flooding generates excessive load for large average connectivity
  - Periodic re-flooding (refreshes)



Examples: OSPF, IS-IS, PNNI, TRILL, RBridges

DISTANCE/PATH VECTOR = DIFFUSED COMPUTATION(DBF)

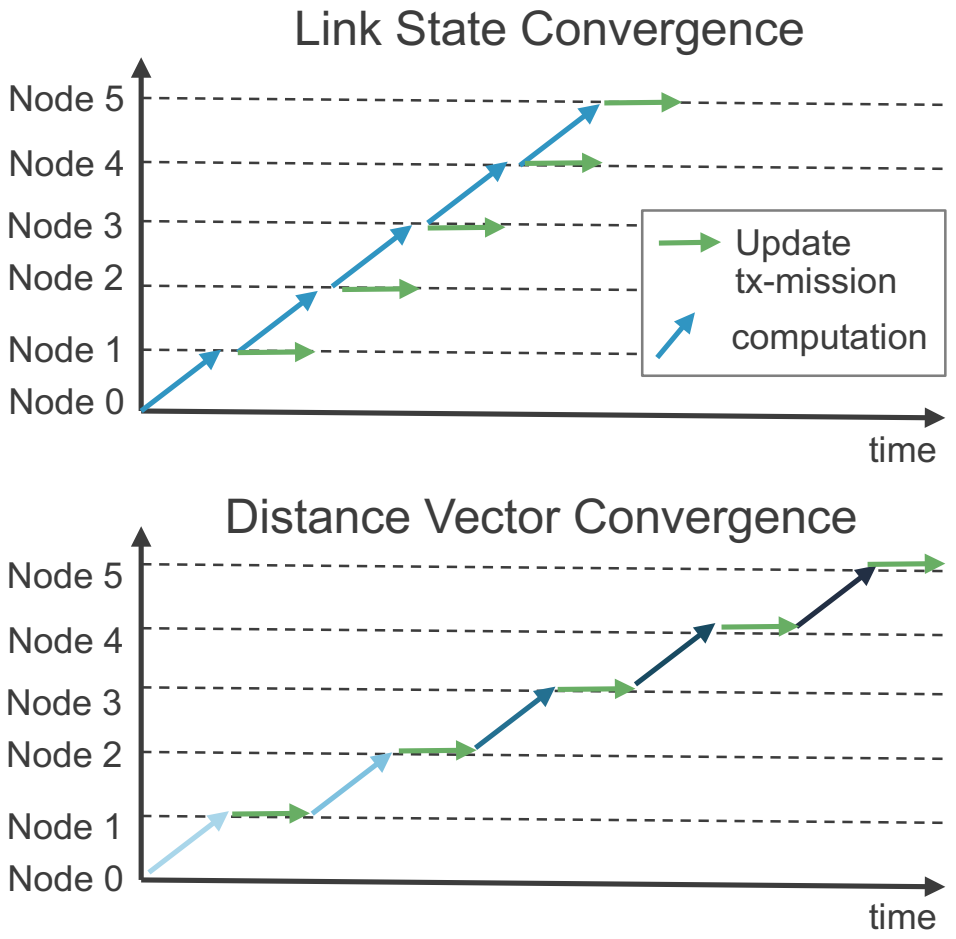
- Prefixes “gather” metric when passed along links
- Each node computes “single best” result and passes it on (*Add-Path* added “multiple best” results)
- A node keeps all copies, otherwise it would have to trigger “re-diffusion”
- Loop prevention is easy on strictly uniformly increasing metric.
- Ideal for “policy” rather than “reachability”
- Scales when properly implemented to much higher # of routes than Link-State
- Slow convergence



Examples: BGP, RIP, IGRP

Link State vs Distance/Path Vector

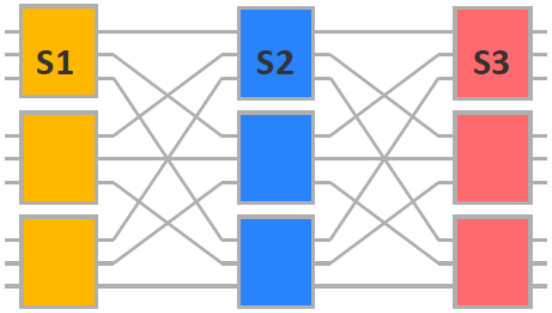
- Link State
  - Topology view → TE enabler
- Distance/Path Vector
  - Every computation could enforce policy – granular control – TE
- Both protocols types (LS and Distance/Path Vector) are frequently used in todays networks



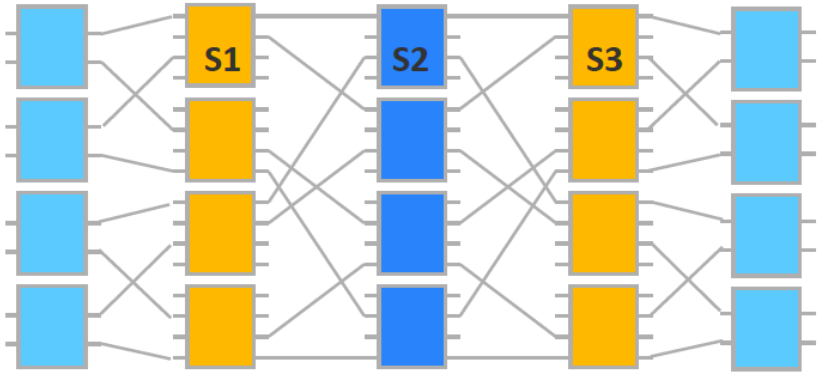
CLOS TOPOLOGIES

- Clos Offers Well-Understood non-Blocking Probabilities, Work Done at AT&T (Bell Systems) in 1950s
- Fully Connected Clos is Dense and Expensive. Data Centers Today Tend to Be Variations of "Folded Fat-Tree"

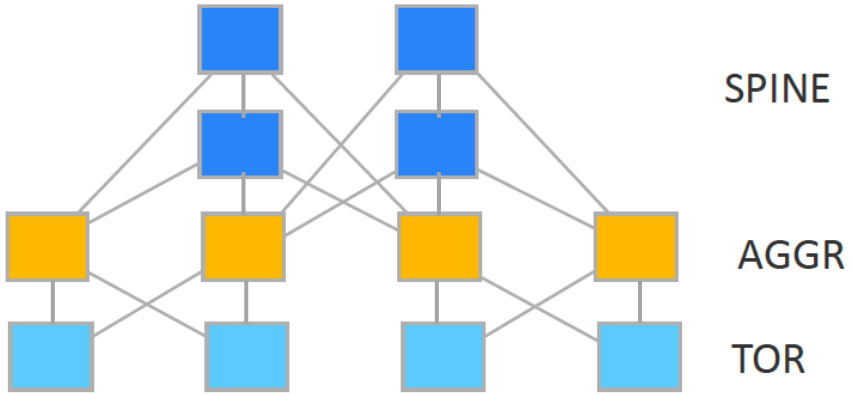
CLOS



Fat-Tree



Folded Fat-Tree



RIFT: Routing Protocol for Clos Underlay

- General Concept
- Automatic Disaggregation
- Optional Horizontal Links
- And More Beyond That

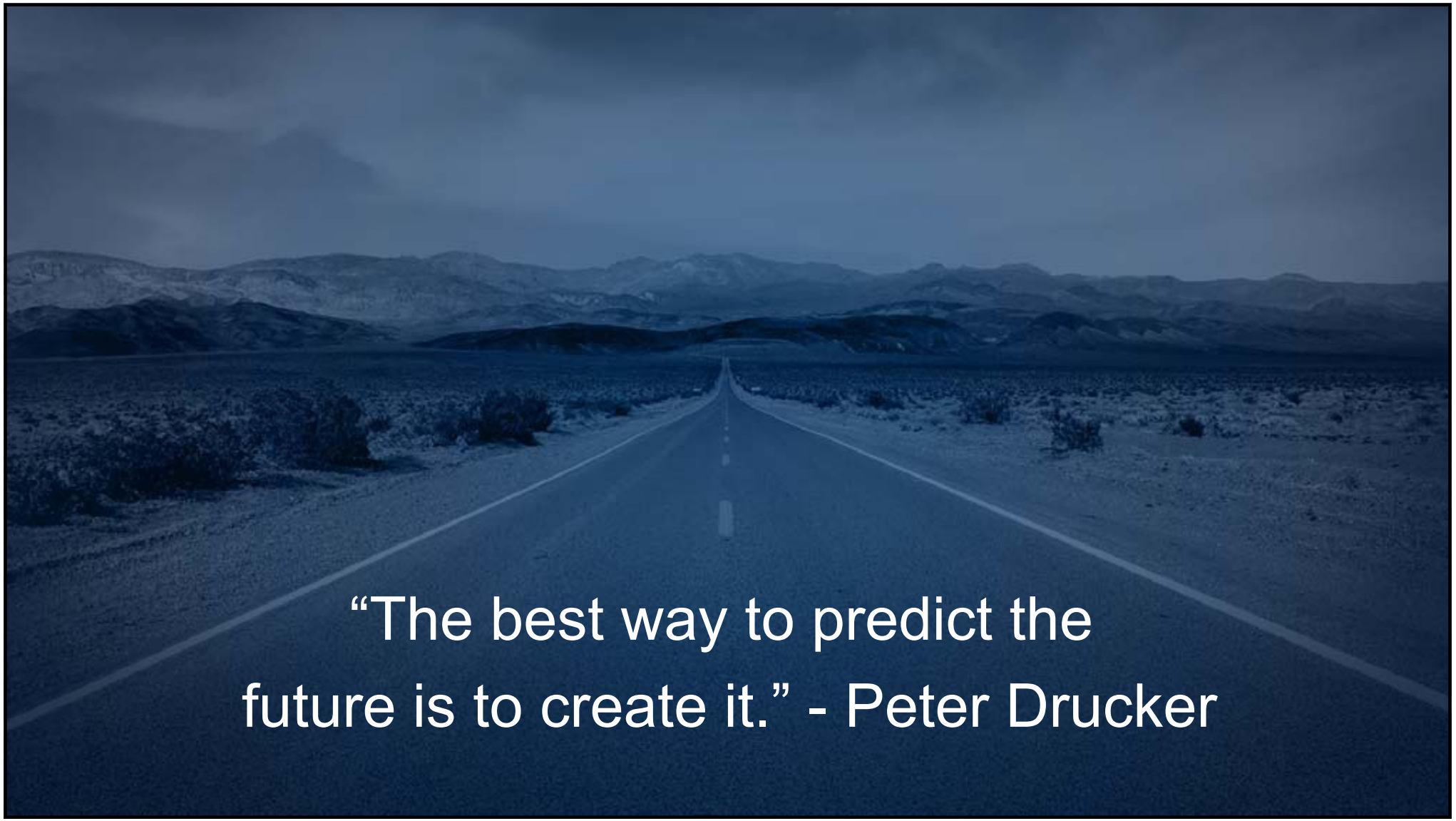
But it's so New …

"Man cannot discover new oceans unless he has the courage to lose sight of the shore." --- Andre Gide

Well, You Must Be …

"The reasonable man adapts himself to the world: the unreasonable one persists in trying to adapt the world to himself. Therefore all progress depends on the unreasonable man." --- Bernard Shaw

RIFT - A TRY TO CREATING THE FUTURE!



RIFT VS. DRAFT-DT-RTGWG-DCROUTING-REQUIREMENTS

Problem / Attempted Solution	Vs. draft-dt-rtgwg-dcrouting-requirements
01. As Close to Zero Necessary Configuration as Possible (Contradicts 02)	
02. Peer Discovery/Automatic Forming of Trees/Preventing Cabling Violations (Contradicts 01)	
03. Minimal Amount of Routes/Information on ToRs	
04. High Degree of ECMP (BGP needs lots knobs, memory, own-AS-path violations) and ideally NEC and LFA	
05. Traffic Engineering by Next-Hops, Prefix Modifications	
06. See All Links in Topology to Support PCE/SR	
07. Carry Opaque Configuration Data (Key-Value) Efficiently	
08. Take a Node out of Production Quickly and Without Disruption	(do we need GR?)
09. Automatic Disaggregation on Failures to Prevent Black-Holing and Back-Hauling	
10. Minimal Blast Radius on Failures (On Failure Smallest Possible Part of the Network "Shakes")	
11. Fastest Possible Convergence on Failures	

General Terminology

-Spine/Aggregation/Leaf Levels: Traditional names for Level 2, 1 and 0 respectively.

-Point of Delivery (PoD): A self-contained vertical slice of a Clos or Fat Tree network containing normally only level 0 and level 1 nodes. It communicates with nodes in other PoDs via the spine.

-Spine: The set of nodes that provide inter-PoD communication. These nodes are also organized into levels (typically one, three, or five levels).

-Leaf: A node without southbound adjacencies. Its level is 0.

Directions:
-Northbound Link: A link to a node one level up/ one level further north.
-Southbound Link: A link to a node one level down/ one level further south.
-East-West Link: A link between two nodes at the same level.
East- West links are normally not part of Clos or "fat-tree" topologies.

RIFT TERMINOLOGY

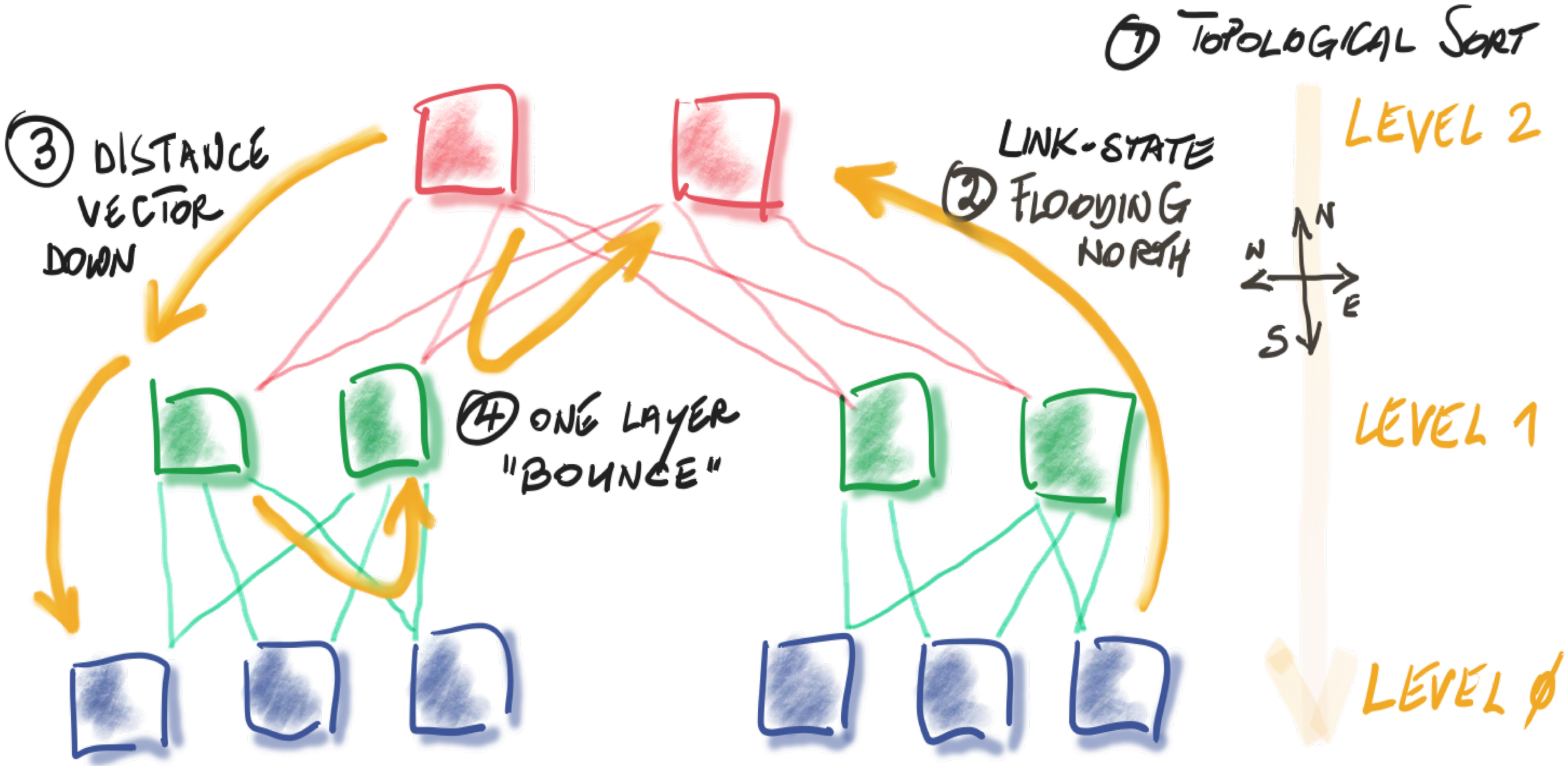
-TIE: Topology Information Element (S-TIE != N-TIE)
-TIEs are exchanged between RIFT nodes to describe parts of a network such as links and address prefixes. It can be thought of as largely equivalent to ISIS LSPs or OSPF LSA.

-Node TIE: equivalent to OSPF Node LSA
-Prefix TIE: contains all prefixes directly attached to this node in case of a N-TIE and in case of S-TIE the necessary default and de-aggregated prefixes the node passes southbound.
-Key Value TIE: A S-TIE that is carrying a set of key value pairs.
It can be used to distribute information in the southbound direction within the protocol.

-TIDE: Topology Information Description Element, equivalent to CSNP in ISIS
-TIRE: Topology Information Request Element, equivalent to PSNP in ISIS.
-PGP: Policy-Guided Prefixes allow to support traffic engineering that cannot be achieved by the means of SPF computation

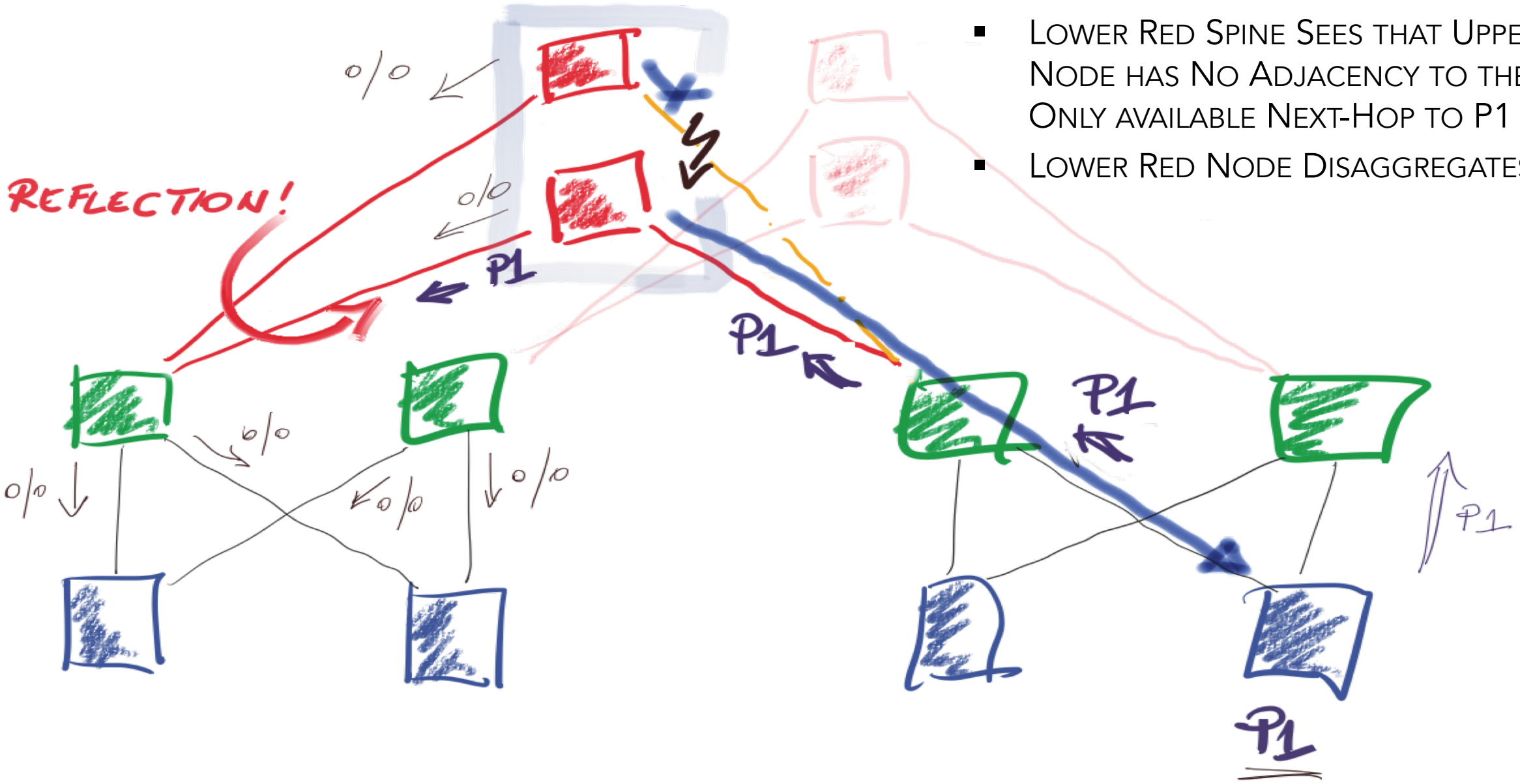
-LIE: equivalent to HELLOs in IGPs and exchanged over all the links between systems running RIFT to form adjacencies.
-BAD: This is an acronym for Bandwidth Adjusted Distance.

Link-State Up, Distance Vector Down & Bounce



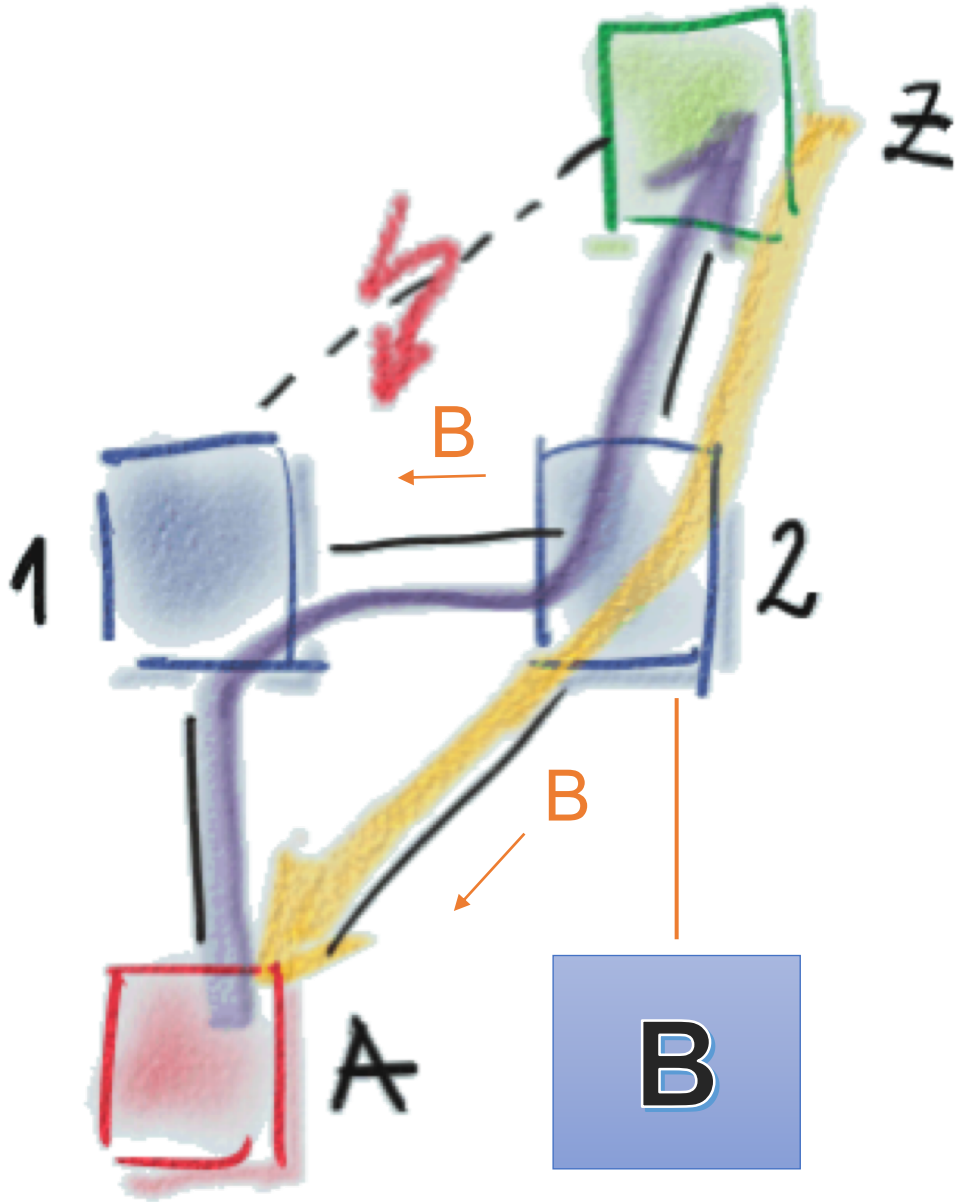
Automatic Disaggregation

- Remember: South Representation of the Red Spines is Reflected by the Green Layer
- Lower Red Spine Sees that Upper Node has No Adjacency to the Only available Next-Hop to P1
- Lower Red Node Disaggregates P1



Optional Horizontal Links for Failure Protection

- Levels Can Install Optional Horizontal Links
- Level 0 is Special:
  - Leaf-2-Leaf Connection that Cannot Be Used Except for Leaf-2-Leaf Traffic
- Level > 0 Uses Horizontal Links for Failure Protection Only
  - Single Node Protection: Node that Lost Northbound Links but has Neighbors that Can Reach Higher Layers Uses the Horizontal Link
  - N:N-1 Protection: Full Mesh in a Level can Provide Up to N-2 Northbound Protection
  - Horizontal Disaggregation Can Heal Complex Failures (Not Different from SouthBound Disaggregation)



RIFT Does On Top

- Automatic Flood Reduction
- Leaf-To-Leaf Bi-Directional Shortcuts
- Possible Traffic Engineering via "Flooded DV Overlay" With Policies
- Completely Model Based Packet Formats
- Channel Agnostic Delivery, Could Be QUICK, TCP, UDP
- Prefixes to Topology Element Mapping Based on Hash Functions Local to Each Node
  - One Extreme Point is Prefix Per Flooded Element = BGP Update
- Purging (Given Complexity) is Omitted
- Policy Controlled Key-Value Store Support

RIFT STATUS IN THE INDUSTRY

Standardization

- Individual contribution to IETF Routing WG
- draft-przygienda-rift -> draft-ietf-rift-rift-01

Implementation

- Prototype reference code exist
- PoC Test runs, performance data collected
- Cooperation

Join work at IETF WG

- Contact authors, share opinion
- The data structures for packet are public (GPB)

Questions?