RIFT: Routing In Fat Trees

A modern Routing Protocol to Meet All DC Topology Requirements

Tony Przygienda

Jan 25th, 2017, IETF WG Interim

WHY ?

Requirements That Cannot Be Met with Other Attempts Easily

REQUIREMENTS BREAKDOWN

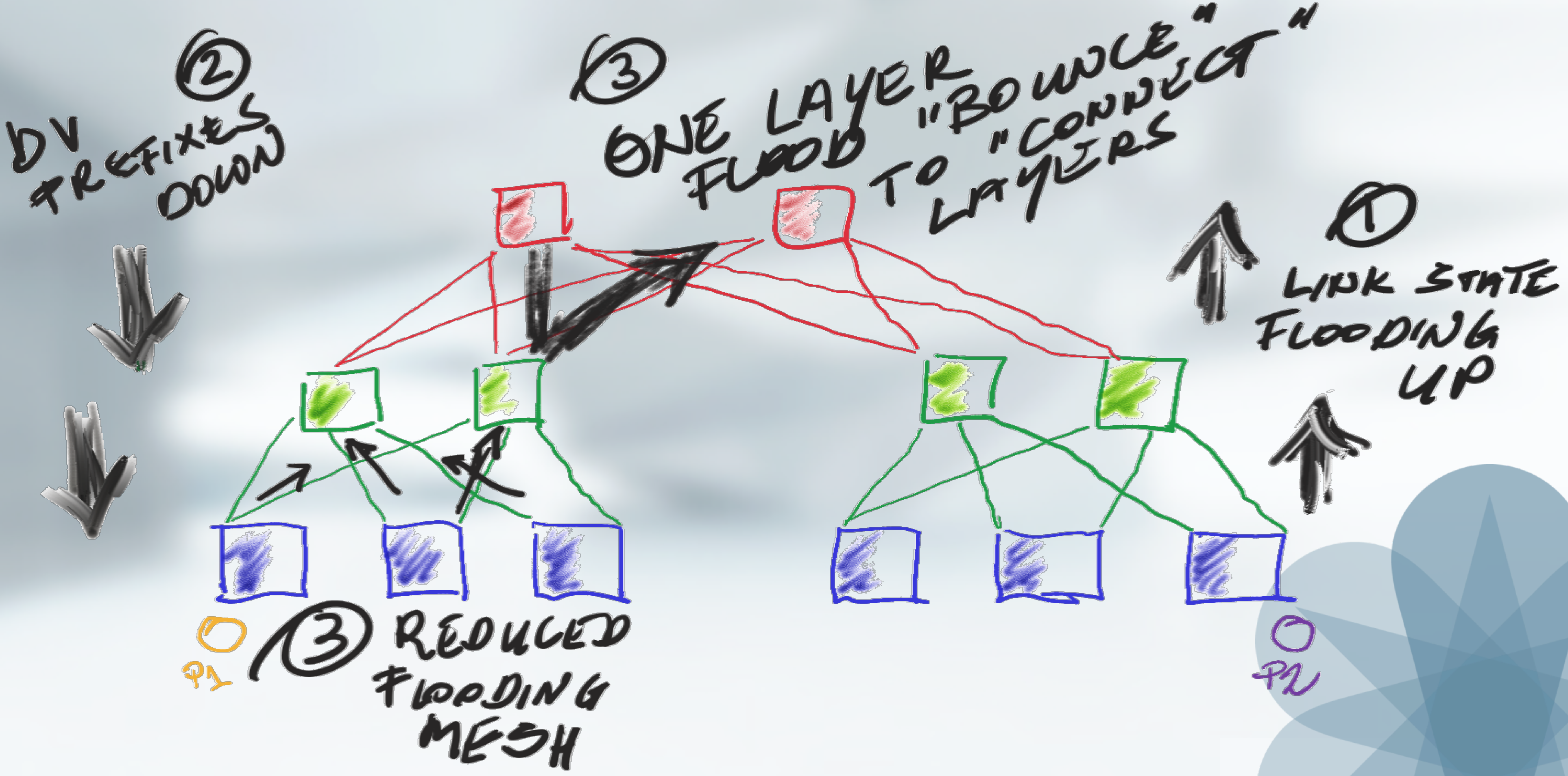
Problem / Attempted Solution	BGP modified for DC (all kind of "mods")	ISIS modified for DC (RFC7356 + "mods")	RIFT Native DC
Link Discovery/Automatic Forming of Trees/Preventing Cabling Violations	✘	⚠	✓
Minimal Amount of Routes/Information on ToRs	✘	✘	✓
High Degree of ECMP (BGP needs lots knobs, memory, own-AS-path violations) and ideally NEC and LFA	⚠	✓	✓
Traffic Engineering by Next-Hops, Prefix Modifications	✓	✘	✓
See All Links in Topology to Support PCE/SR	⚠	✓	✓
Carry Opaque Configuration Data (Key-Value) Efficiently	✘	⚠	✓
Take a Node out of Production Quickly and Without Disruption	✘	✓	✓
Automatic Disaggregation on Failures to Prevent Black-Holing and Back-Hauling	✘	✘	✓
Minimal Blast Radius on Failures (On Failure Smallest Possible Part of the Network "Shakes")	✘	✘	✓
Fastest Possible Convergence on Failures	✘	✓	✓
Simplest Initial Implementation	✓	✘	✘

HOW ?

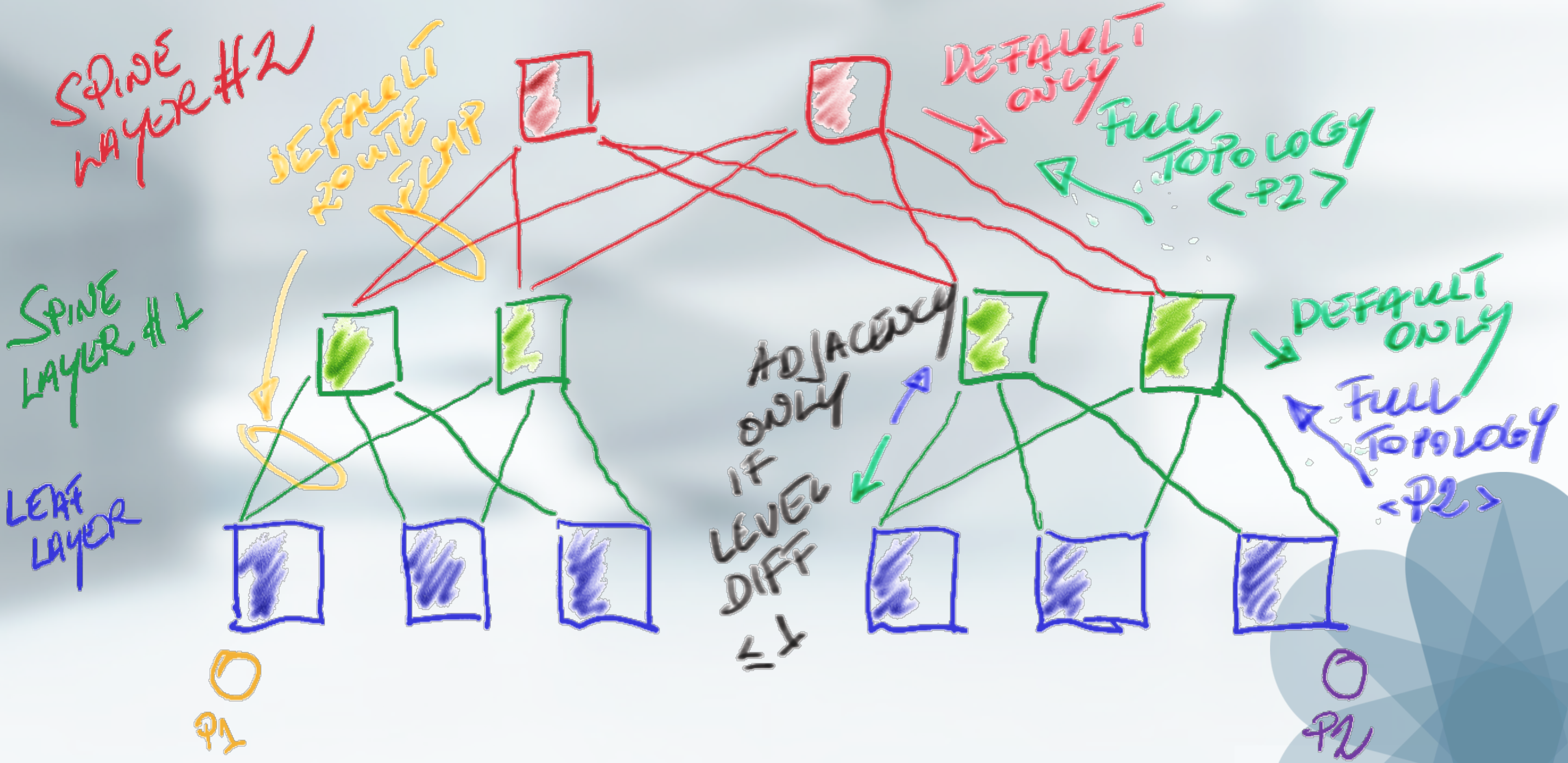
The Way It's Put Together

10K Feet Overview

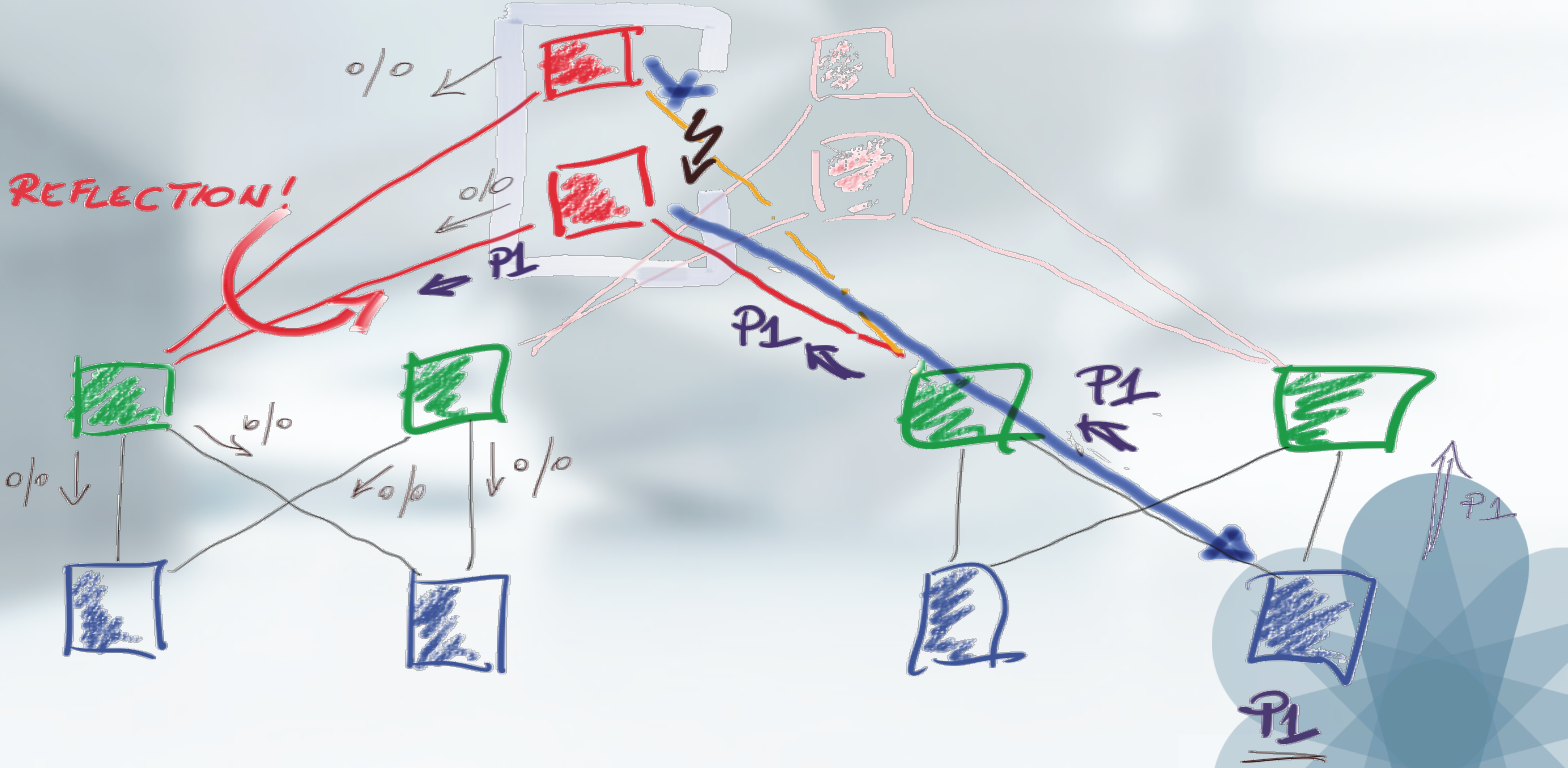
10K FEET VIEW: "LINK-STATE UP, DISTANCE VECTOR DOWN & ONE BOUNCE"



10K FEET VIEW: "AUTOMATIC CABLING AND MINIMAL TOPOLOGY EXCHANGE"



10K FEET VIEW: "AUTOMATIC DE-AGGREGATION"



WHAT ELSE ?

Rest of Things Worth Mentioning

AND MOREOVER …

- Traffic Engineering via “Flooded DV Overlay” With Policies
- Easy to Support NECMP and W(N)ECMP
- Automatic Robust Flooding Reduction Without CDS or a Synchronous Distributed Protocol
- Time Moved On and Things Progressed
  - Time to Loose Hand-Crafted Packets
    - Model Based Packet Formats
    - Channel agnostic, LIEs over UDP, Flooding Could Be QUICK, TCP, UDP
  - Build Prefix TIEs Based on Hash Functions
    - One Extreme Point is TIE per Prefix
  - Purging (Given Complexity) Omitted Granted Today’s Memory Footprints
  - Key-Value Store Support

WHILE IT REMAINS TO DECIDE …

- Leaf-2-Leaf ?
- NBMA ?
- Parallel Links ?
- Interactions
  - BFD ?
  - LFA ? = NECMP for RIFT ?
  - FRR ?

THANKS