RIFT: Towards Self-Driving IP Fabrics with Zero OPEX Routing

Tony Przygienda prz@juniper.net

draft-przygienda-rift @ IETF

Disclaimers and Expectations

None of those Things Constitute Commitments to Product Specifications, Offerings or Release Dates by Juniper at this Point in Time

Agenda

- Everything in a Nutshell
- Fabric Topology
- Blitz Overview of Today's Routing
- "IP Fabric Routing" is a Specialized Problem
- RIFT: Zero OPEX Routing for Self-Driving IP Clos Fabrics

In a Nutshell

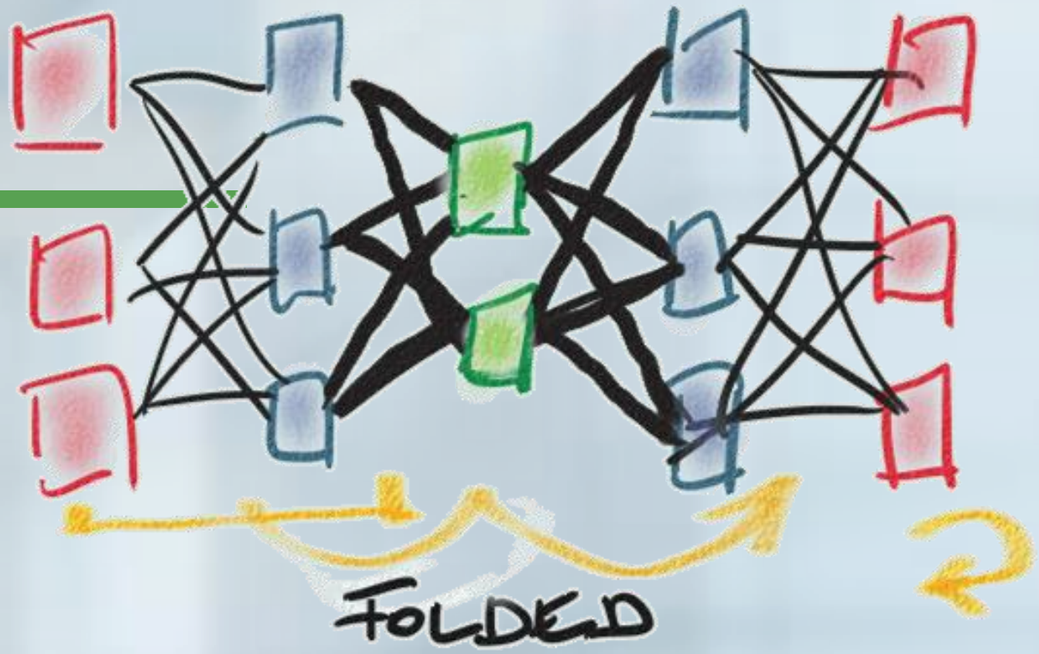
- Data Centers are a Strategic Asset for Large Corporations and Operators
  - If They Don't Own Their Critical Data Processing, Their Data Owner May Own Them in the Future
- Building Data Centers Necessitates Building Fabrics
  - Large Fabrics are Becoming IP Only and Need to Be Routed
- Routing OPEX for IP Fabrics is Significant
  - Technically Complicated, Neither Efficient nor Robust
  - Few Available Experts in the Field
  - "Self Driving" Fabrics are Needed
    - Zero Configuration, Resilience, Maximum Goodput, as "Self Driving" as Possible
- Juniper is Working on Open Standard for Zero OPEX, Self-Driving IP Fabrics Routing
- Accidentally, "Exploded Chassis Backplane" Presents the Same Problem

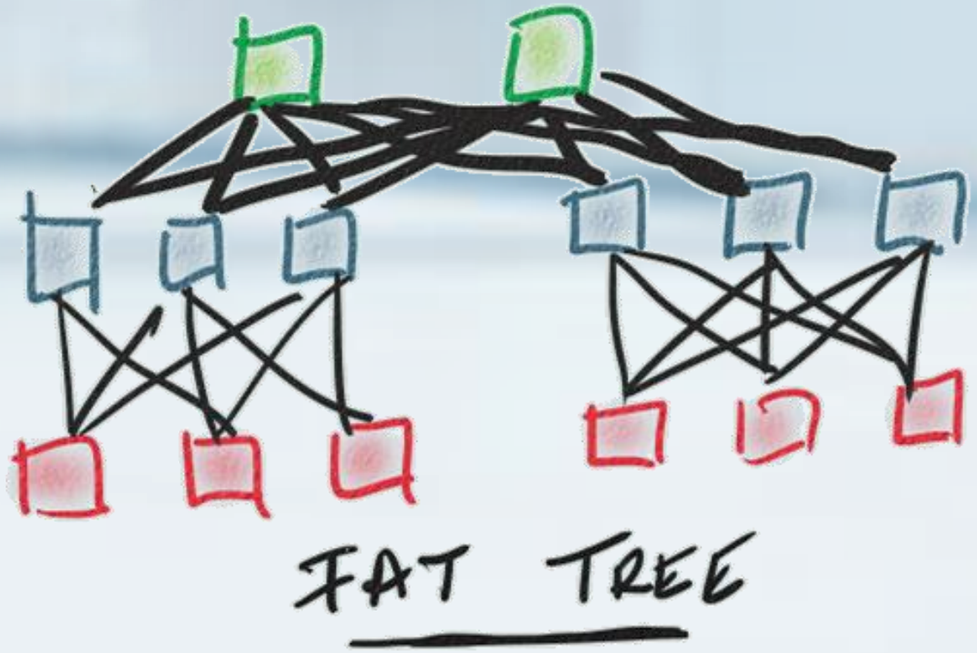
Fabric: a Specialized Topology

- Clos Topologies Are Dominant Today
- Current State of IP Fabric Routing Affairs
- Requirements Matrix for Self-Driving IP Fabric Routing

Clos Topologies

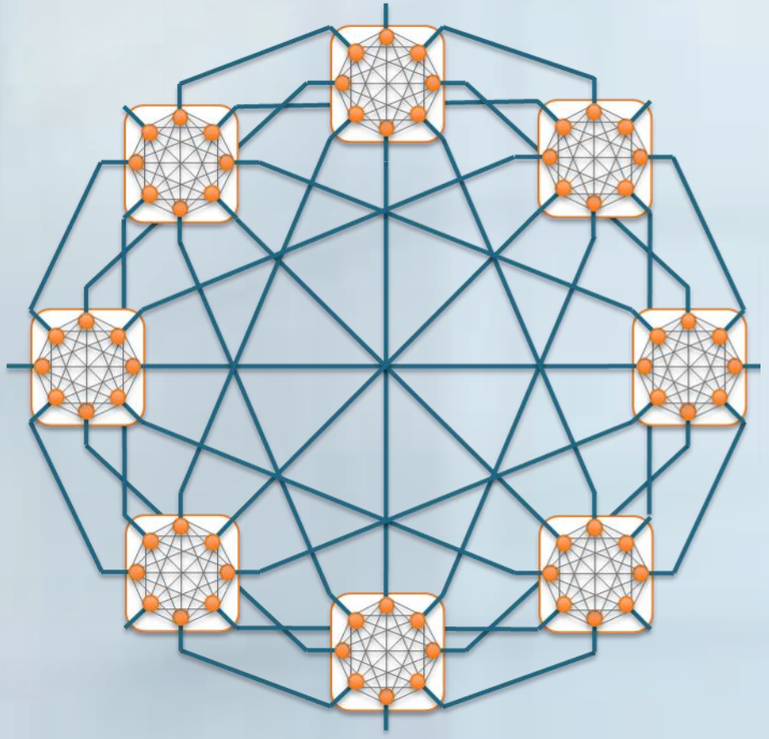
- Clos Offers Well-Understood Blocking Probabilities
- Work Done at AT&T (Bell Systems) in 1950s
- Fully Connected Clos is Dense and Expensive
- Data Centers Today Tend to Be Variations of "Folded Fat-Tree"
  - Input Stages = Output Stages
  - Clos is "Partial"
  - Links get "Fatter" Up the Tree

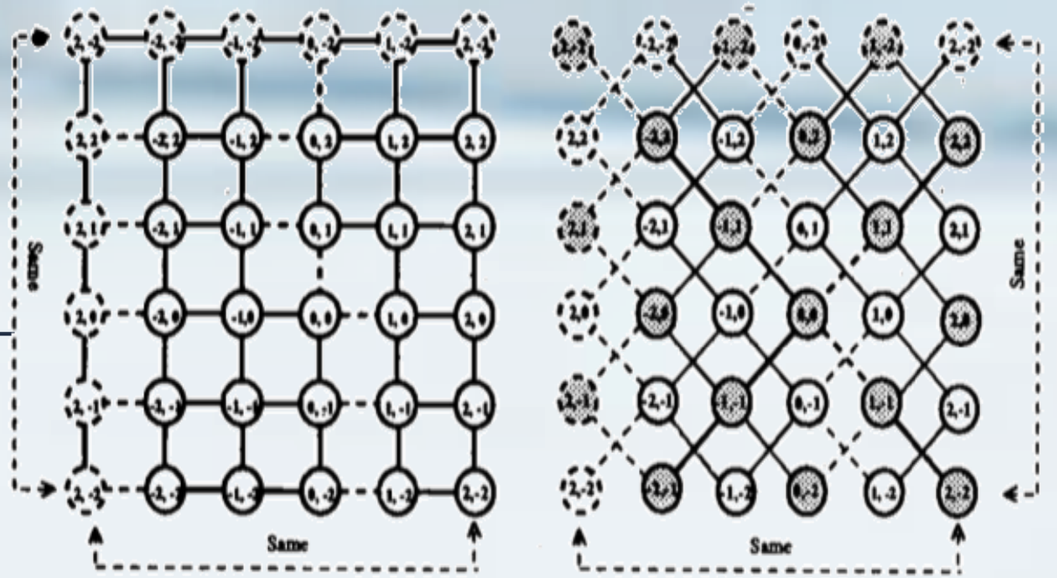




Why Not Something Else ?

- Toroidal [and Diagonal] Meshes have Long Paths, Small Bisection Width and Poor Blocking Properties
- Dragonfly is very Novel and Unproven
  - Seemingly ½ Throughput of CLOS at Same Capacity Due to Low ECMP
  - Our Suggestion Should Work Well in a Practical Modification (One Level CLOS and Dragonfly Core) if Necessary



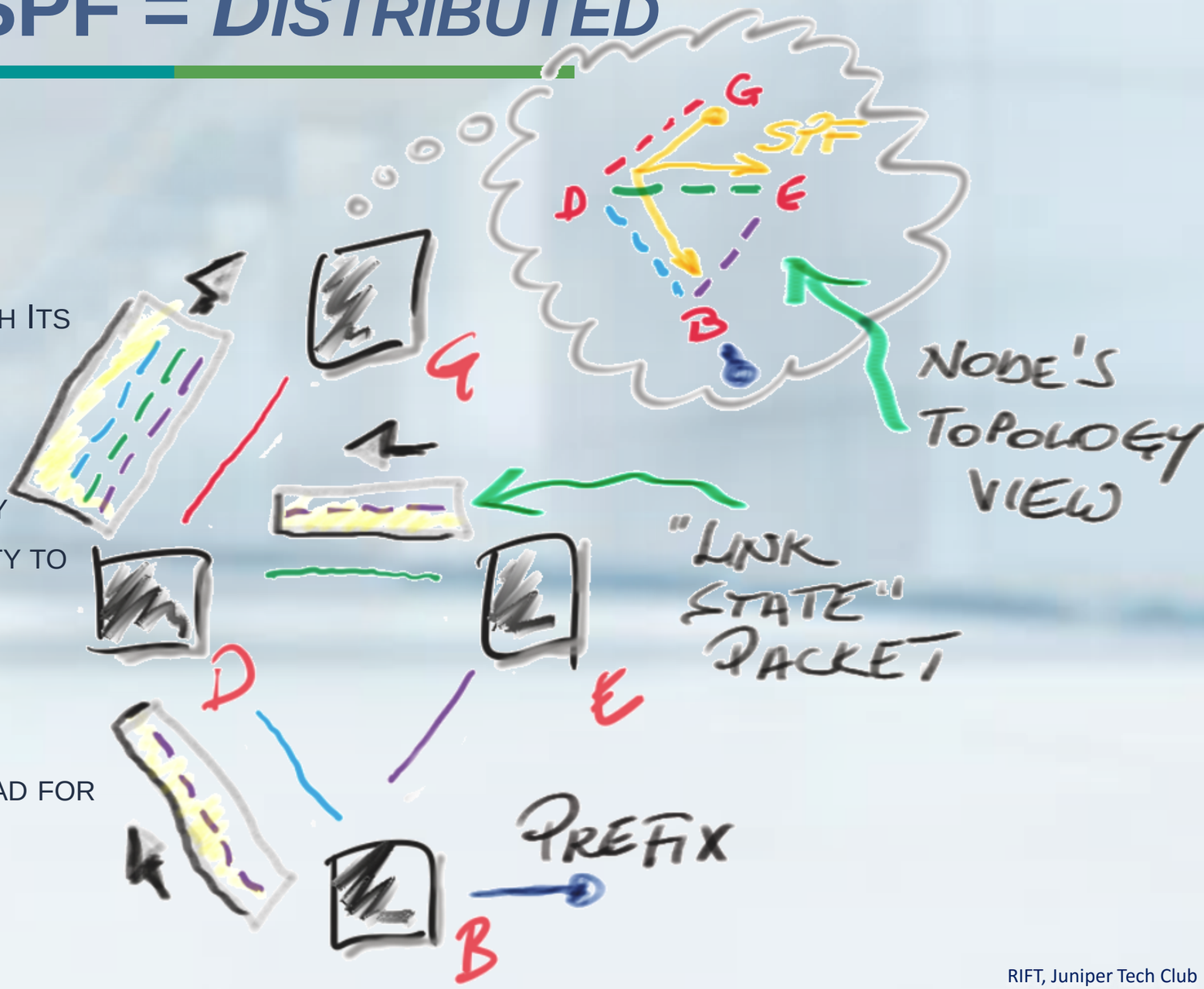


Blitz Overview of Traditional Routing

- Link State & Shortest Path First
- Distance & Path Vector

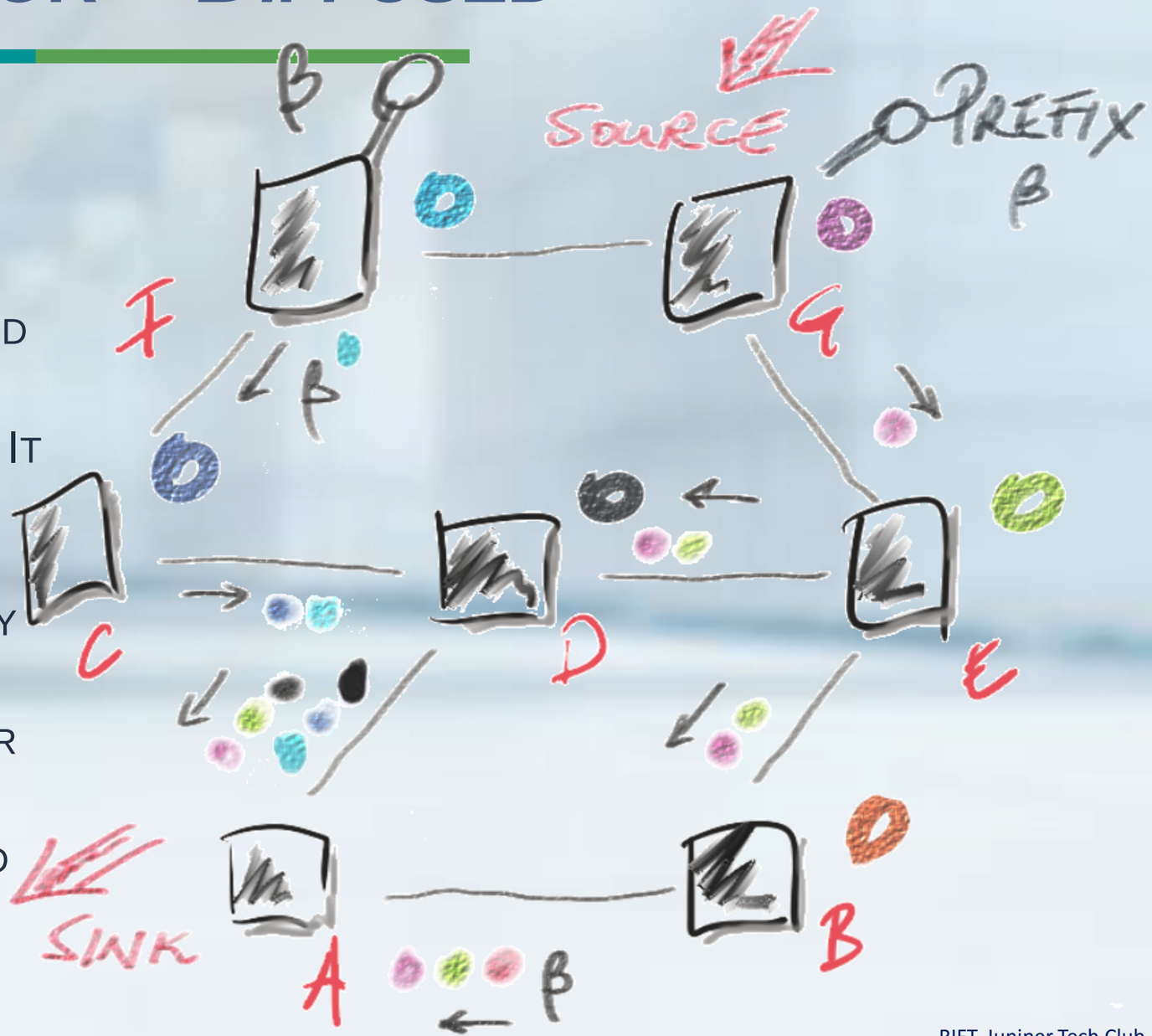
Link State and SPF = *Distributed* Computation

- Topology Elements
  - Nodes
  - Links
  - Prefixes
- Each Node Originates Packets with Its Topology Elements
- Packets are ”Flooded”
- ”Newest” Version Wins
- Each Node “Sees” Whole Topology
- Each Node “Computes” Reachability to Everywhere
- Conversion is Very Fast
- Every Link Failure Shakes Whole Network (modulo Areas)
- Flooding Generates Excessive Load for Large Adjacency Fan-Outs
- Periodic Refreshes (not Strictly Necessary)



Distance/Path Vector = *Diffused* Computation

- Prefixes "Gather" Metric When Passed Along Links
- Each Sink Computes "Best" Result And Passes It on (Add-Path Changed That)
- A Sink Keeps All Copies, Otherwise It Would Have to Trigger "Re-Diffusion"
- Loop Prevention is Easy on Strictly Uniformly Increasing Metric
- Ideal for Enforcing "Policy" Rather Than Provide "Max Reachability"
- Scales When Properly Implemented to Much Higher # of Routes Than Link-State



Current State of Routing in DC Fabrics

- Several of Large DC Fabrics Use E-BGP with Band-Aids as De-Facto IGP (RFC7938)
  - Numbering Schemes to Control "Path Hunting"
    - "Looping Paths" (allow-own-as Under AS Private Numbering)
    - "Relaxed Multi-Path ECMP" Since ECMP over Different AS in EBGP Does Not Work Normally
  - Add Paths to Support Multi-Homing, N-ECMP, Prevent Oscillations
  - Efforts to Get Around 65K ASes and Limited Private AS Space
  - Proprietary Provisioning and Configuration Solutions, LLDP Extensions
  - "Violations" of FSM Like Restart Timers and Minimum-Route-Advertisement Timers
  - Emerging Work for "Peer Auto-Discovery" and "SPF" Diametrically Opposite to BGP Design Principles
  - Reliance on "Update Groups" ~ Peer Groups to Prevent Withdrawal and Path Hunting After Server Link Failures
- Many Large Corporations Run "Flat" IGP (ISIS or OSPF)
- Yet Others Run BGP over IGP (Traditional Routing Architecture)
- Less Than More Successful Attempts @ Prefix Summarization, Control of Micro- and Black-Holing

Requirements Breakdown (RFC7938+) For a "Zero OPEX Fabric"

Problem / Attempted Solution	BGP modified for DC	ISIS modified for DC	RIFT
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	✗	✓	✓
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	✗	✓	✓

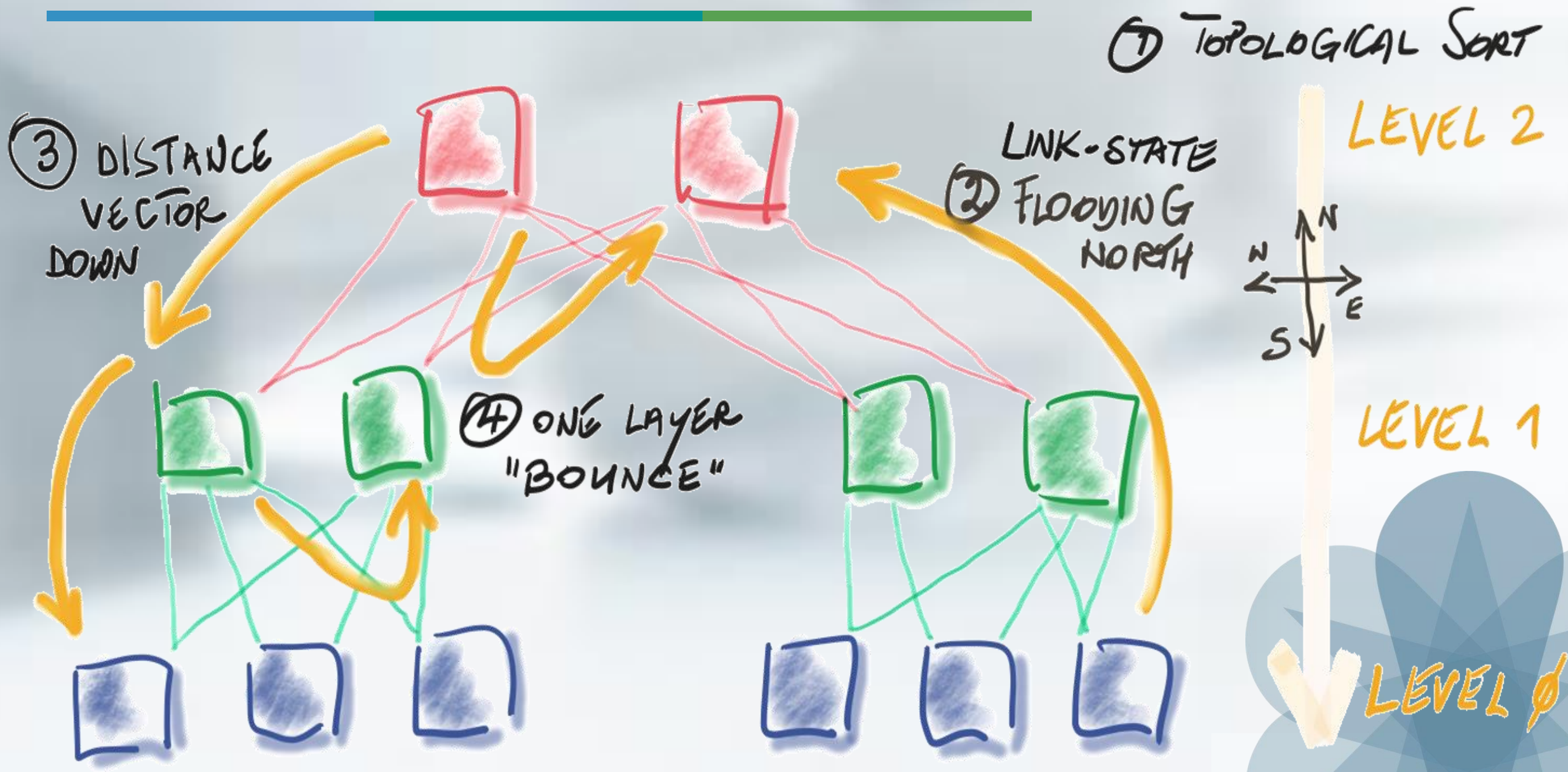
RIFT: Self Driving Routing Algorithm for Clos Underlay

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

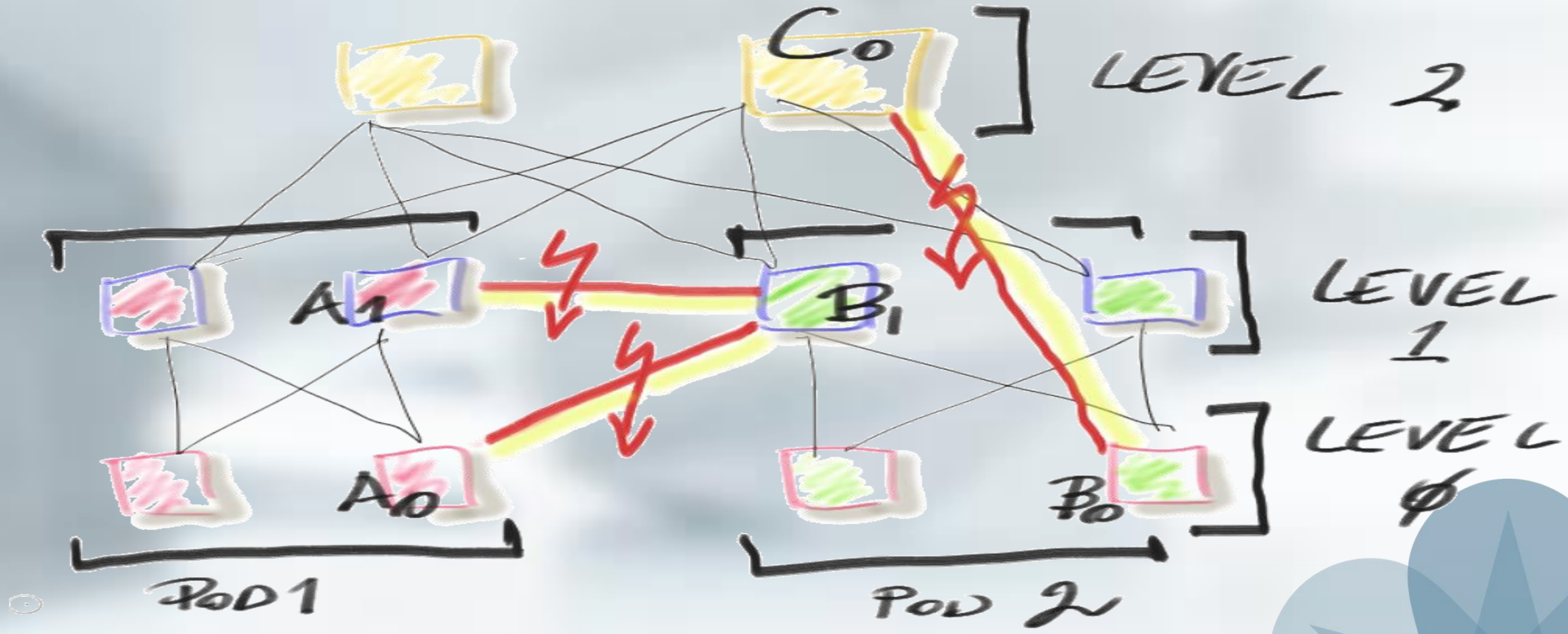
“Just because the standard provides a cliff in front of you, you are not necessarily required to jump off it.”

— Norman Diamond

Link-State Up, Distance Vector Down & Bounce

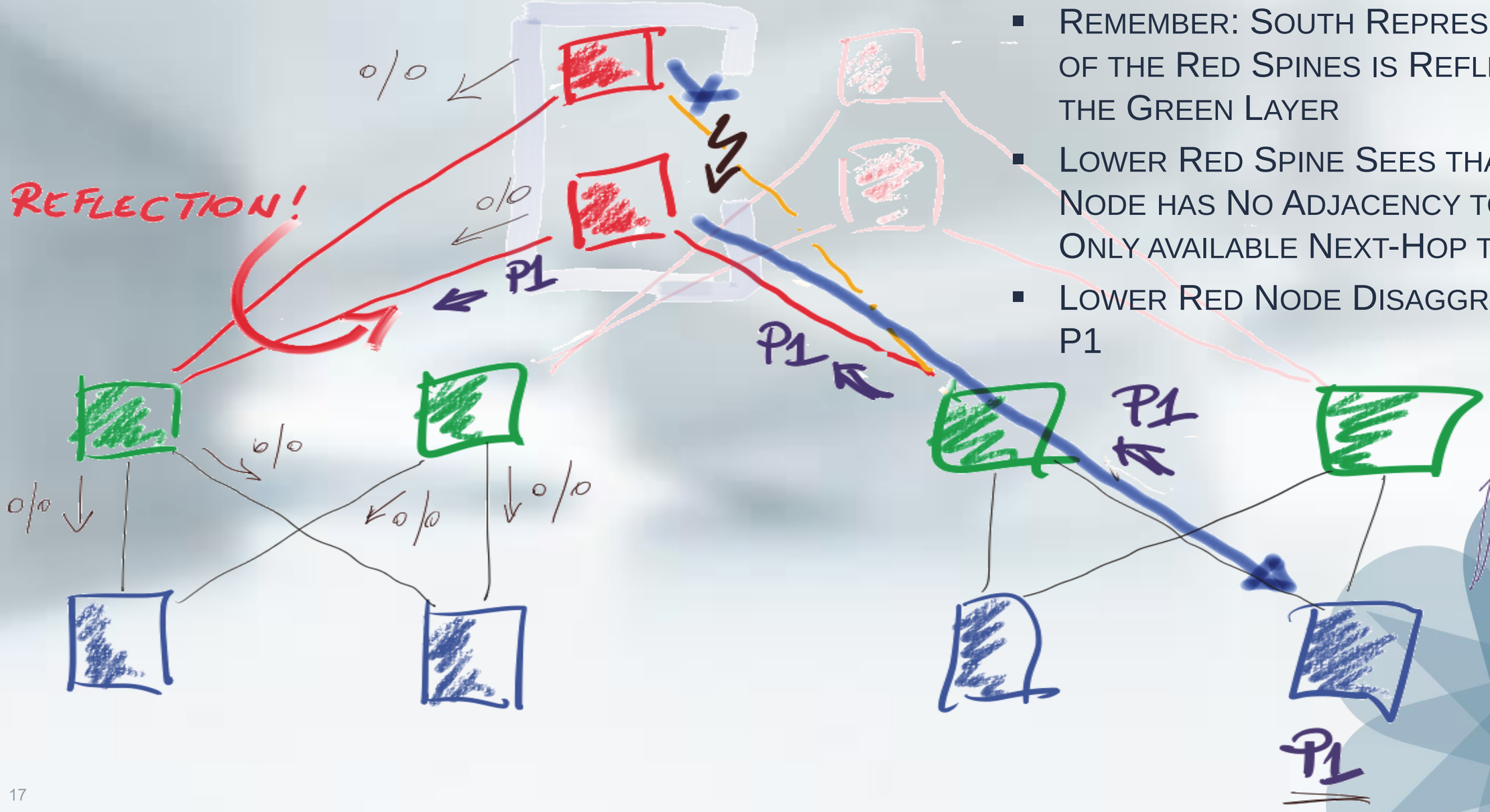


Automatic Mis-Cabling Constraints



- Automatic Rejection of Adjacencies Based on Minimum Configuration
- Protocol Will Work as Well If Level 0 is Allowed to Connect to Level 2 but Optimal Routing Would Need Larger FIBs on Leafs

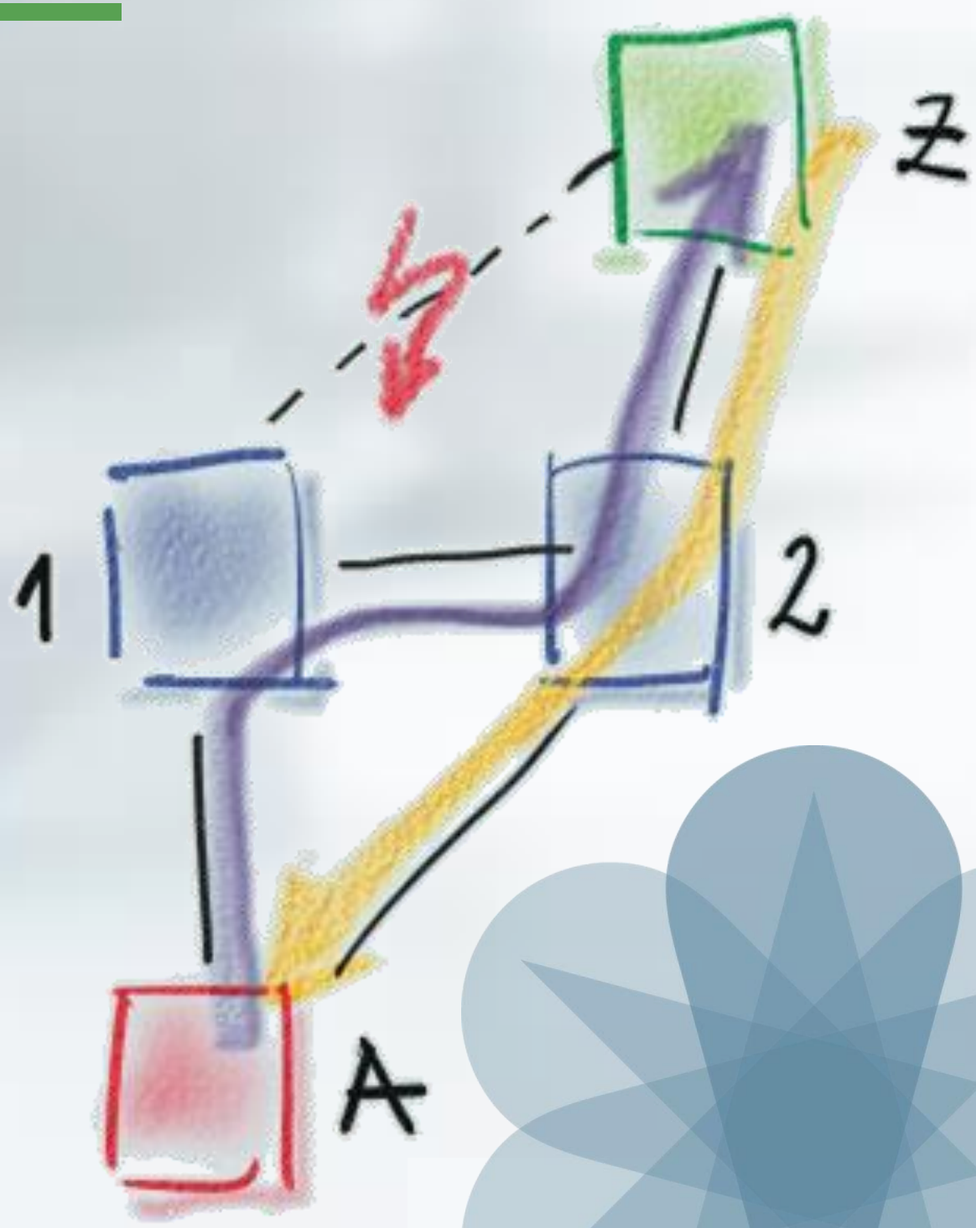
Automatic Disaggregation



- 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 Layer 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 <u>for Failure Protection Only</u>
  - 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



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

Summary of RIFT Advantages

- Open IETF Standard
- **Advantages of Both Link-State and Distance Vector**
  - Fastest Possible Convergence
  - Automatic Detection of Topology
  - Minimal Routes on TORs
  - High Degree of ECMP
  - Fast De-comissioning of Nodes
  - Maximum Propagation Speed with Flexible # Prefixes in an Update
- **No Disadvantages of Neither Link-State nor Distance Vector**
  - Reduced Flooding
  - Automatic Neighbor Detection
- **Unique RIFT Advantages**
  - Automatic Disaggregation on Failures
  - Key-Value Store
  - Horizontal Links Used for Protection
  - Minimal Blast Radius on Failures
  - Can Utilize **All** Paths Through Fabric Without Looping

Is There More Than Bits Over PowerPoint ?

- Detailed Draft in IETF
  - https://datatracker.ietf.org/doc/draft-przygienda-rift/
- Pre-Production Code Available Under NDA
  - Please Talk to Your Friendly Systems or Resident Engineer
- Singapore IETF is Hosting a “DC Routing BoF” Session

THANKS