RIFT: A Novel DC Fabric Routing Protocol

draft-przygienda-rift

IETF '98

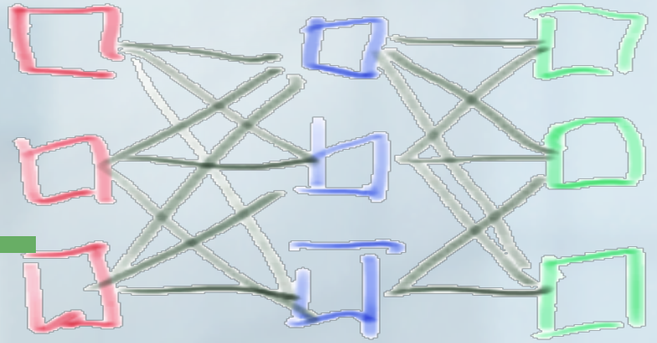
Content

- DC Fabric Routing is a Specialized Problem
- RIFT: a Novel Routing Algorithm for DC Fabric Underlay

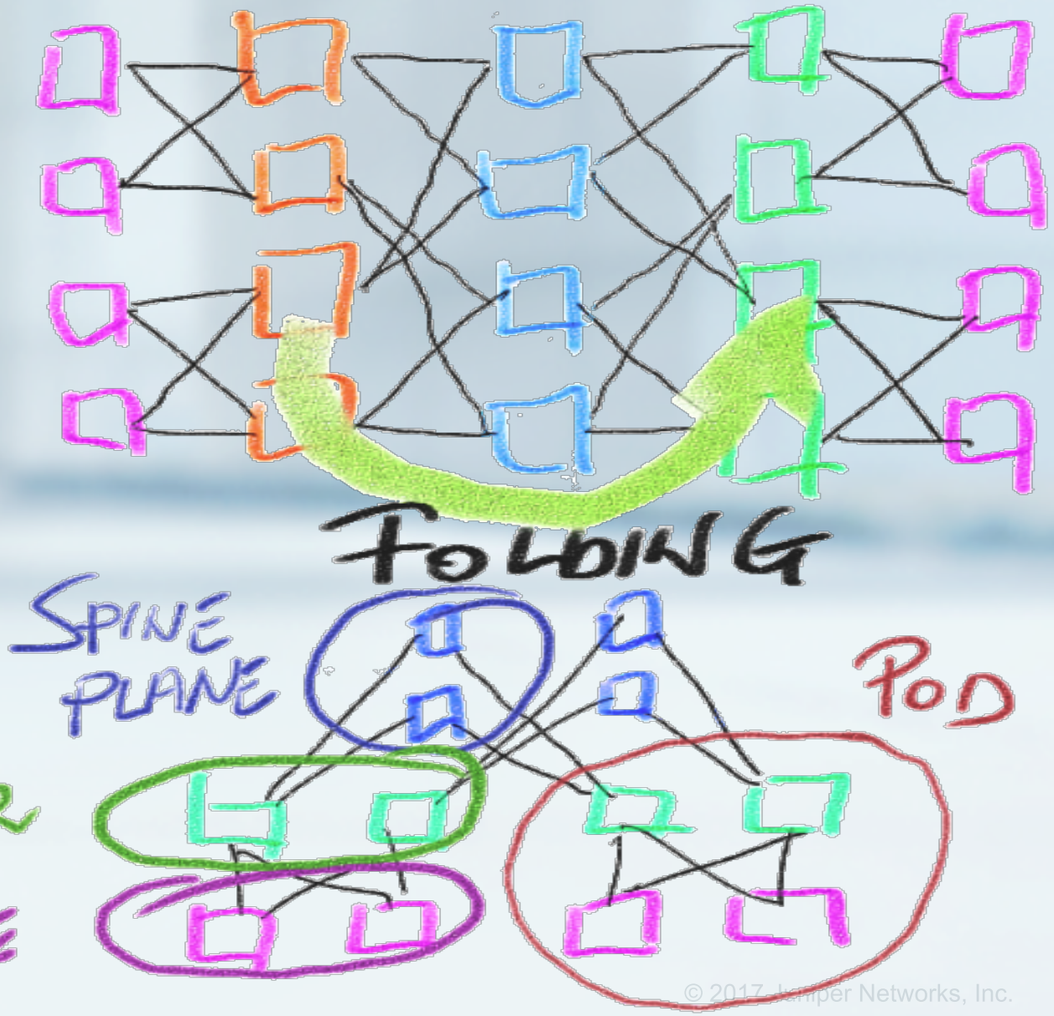
DC Fabric Routing: a Specialized Problem

- Clos/Fat-Tree Topology Variations
- Current State of dynamic DC Routing
- Dynamic DC Routing Requirements Matrix

Clos Variation Topologies



- Clos Offers Well-Understood Blocking Probabilities
- Work Done at AT&T (Bell Systems) in 1950s for Crossbar Scaling
- Fully Connected Clos is Dense and Expensive
- Data Centers Today Tend to Be Variations of "Folded Fat-Tree", I.E. Input Stages are Same as Output Stages and Clos is "Partial"



Current State of Affairs

- Several of Large DC Fabrics Use E-BGP with Band-Aids as IGP (RFC7938)
  - "Looping Paths" (allow-as)
  - "Relaxed Multi-Path ECMP"
  - AS Numbering Schemes to Control "Path Hunting" via Policies
  - Add Paths to Support Multi-Homing, ECMP on EBGP
  - 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
- Others Run IGP (ISIS)
- Yet Others Run BGP over IGP (Traditional Routing Architecture)
- Less Than More Successful Attempts @ Prefix Summarization, Micro- and Black-Holing
  - Works Better for Single-Tenant Fabrics Without LAN Stretch or VM Mobility

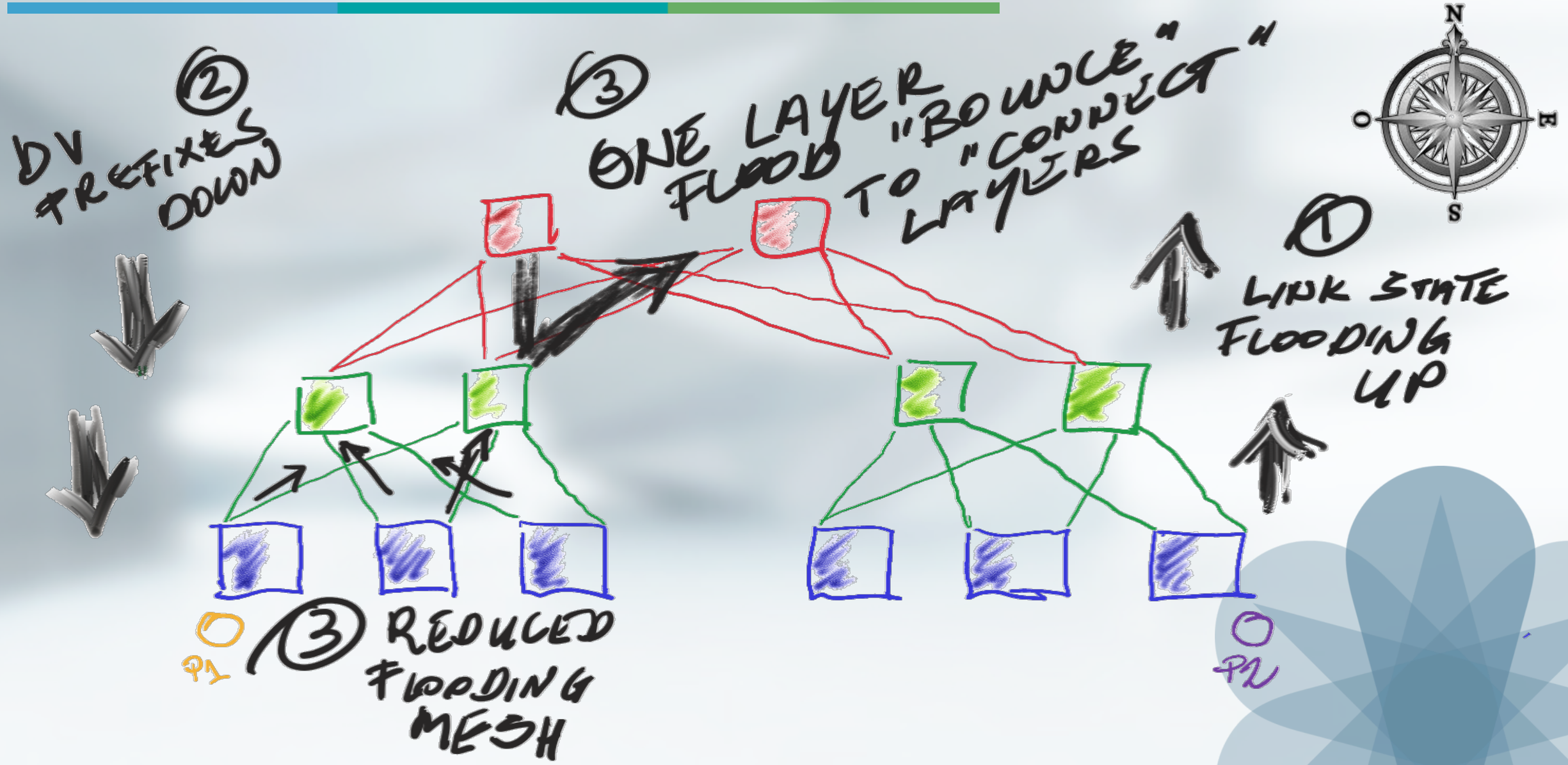
Dynamic DC Routing Requirements Breakdown (RFC7938+)

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	✓	✘	✘

RIFT: Novel Dynamic Routing Algorithm for Clos Underlay

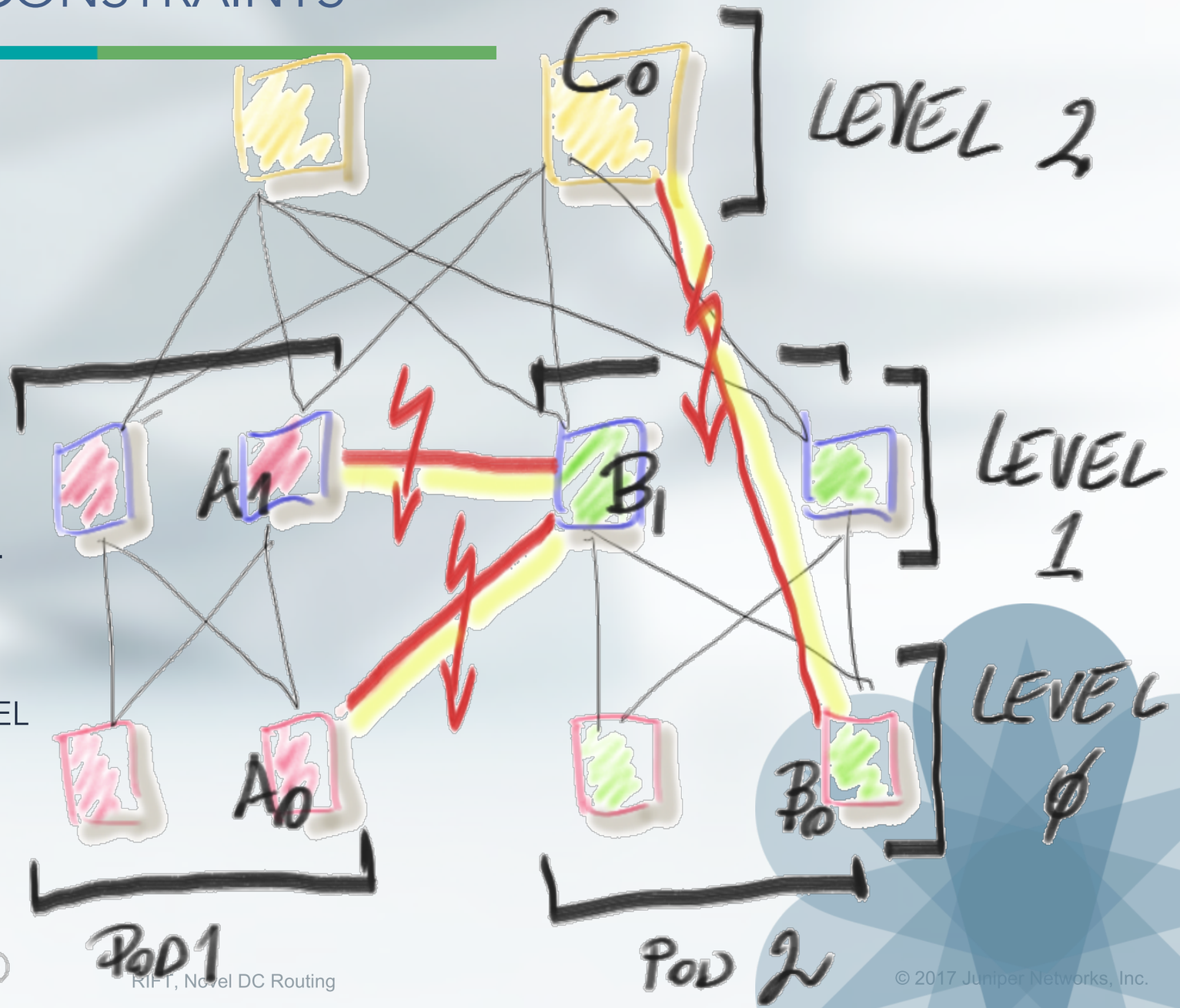
- General Concept
- Automatic Cabling Constraints
- Automatic Disaggregation on Failures
- Automatic Flooding Reduction
- More Goodies

In One Picture: Direction, Link-State Up, Distance Vector Down & a Bounce

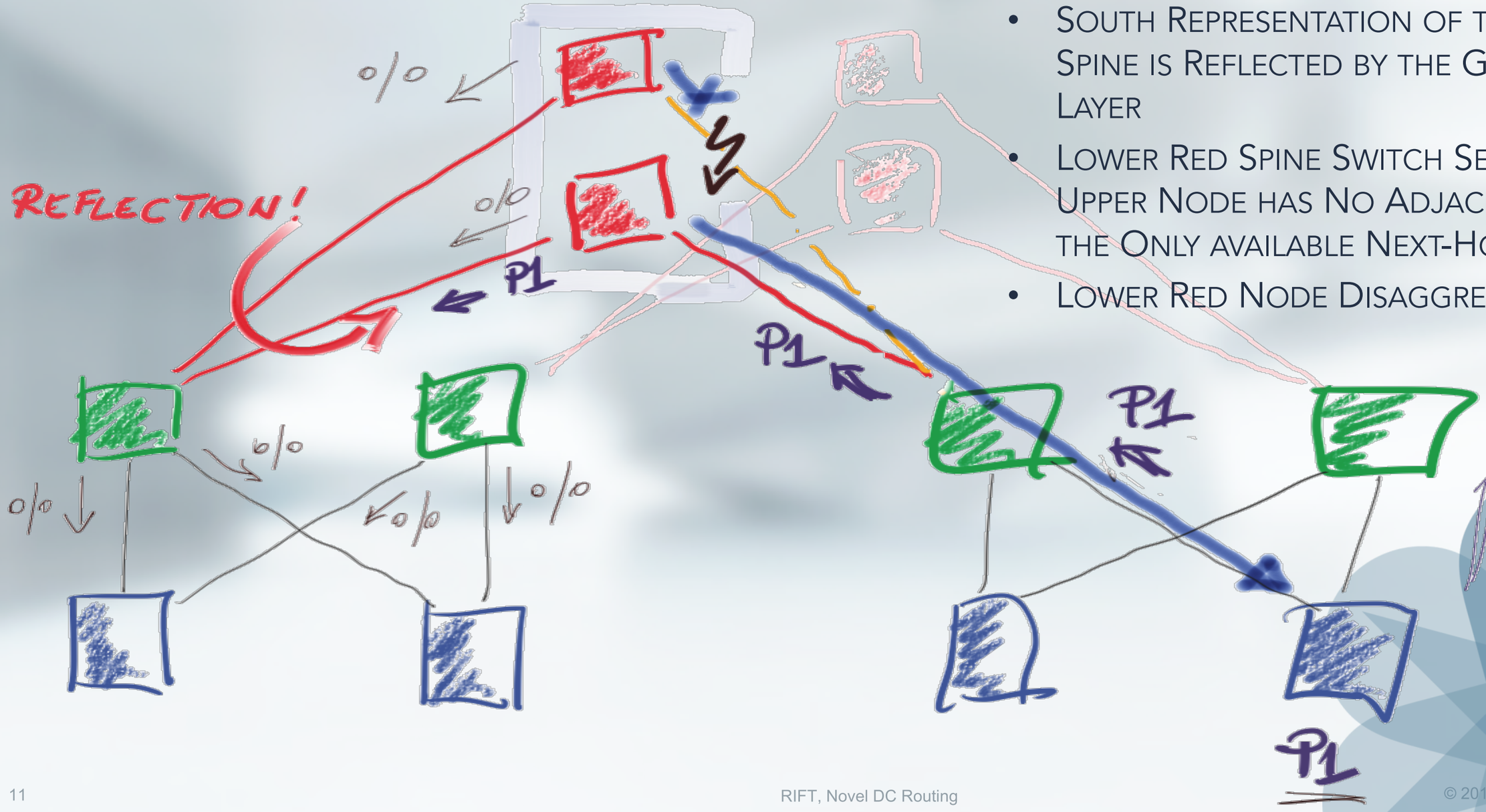


Automatic Topology Constraints

- Level 0 = Leaf
- Pod 0 = Any Pod
- Automatic Rejection of Adjacencies Based on Minimum Configuration
- A1 to B1 Forbidden Due to POD Mismatch
- A0 to B1 Forbidden Due to POD Mismatch (A0 Already Formed A0-A1 Even If POD Not Configured on A0)
- B0 to C0 Forbidden Based on Level Mismatch
- Could Form Other Topology Variations as Well



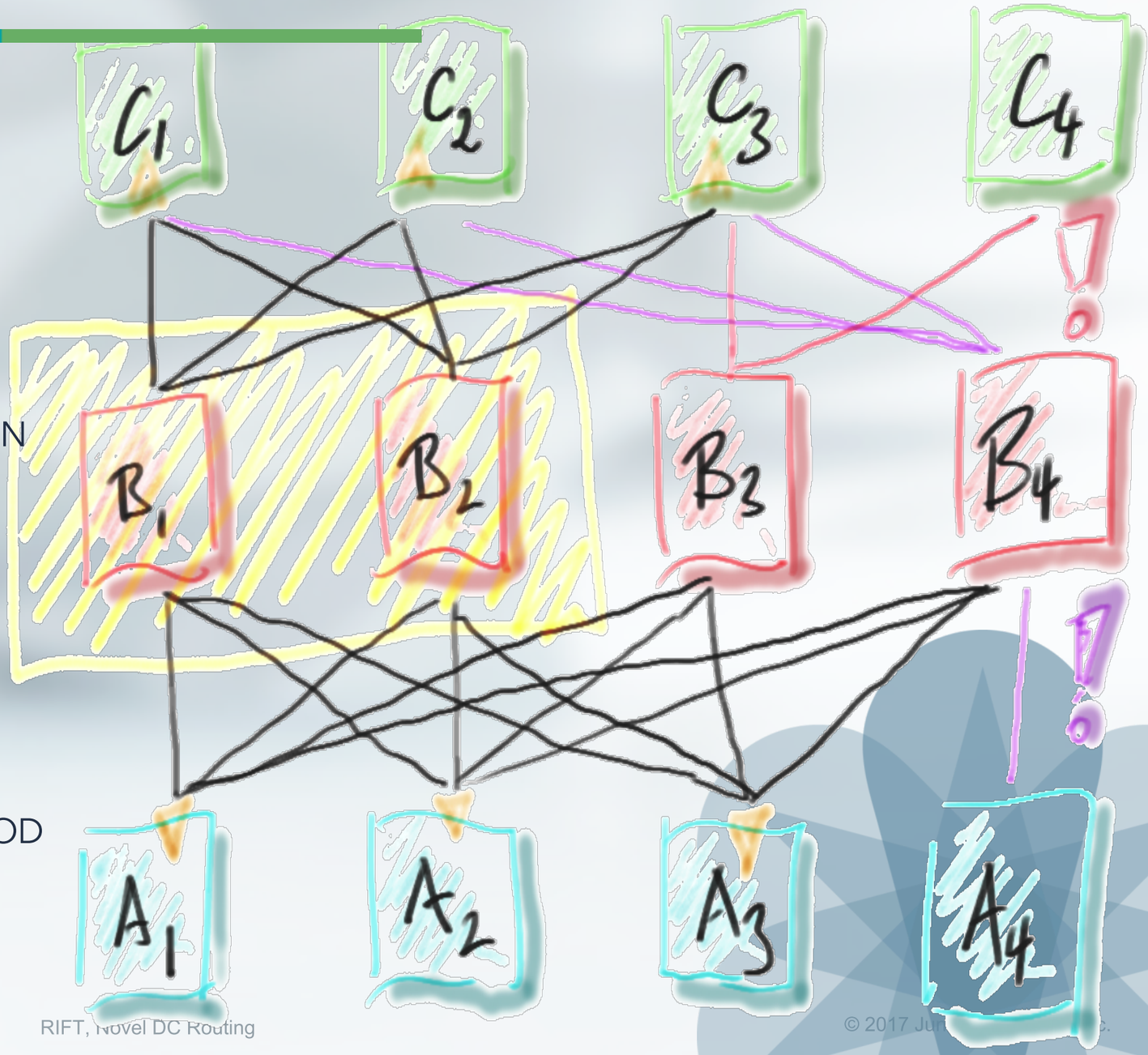
Automatic De-Aggregation



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

Automatic Flooding Reduction

- Each "B" Node Computes From Reflected South Representation of Other "B" Nodes
  - Set of South Neighbors
  - Set of North Neighbors
- Nodes Having Both Sets Matching Consider Themselves "Flood Reduction Group" and Load-Balance Flooding
- Fully Distributed, Unsynchronized Election
- In this Example Case B1 & B2
- Each Node Chooses Based on Hash Computation which Other Nodes' Information it Forwards on ***First*** Flood Attempt
- Similar to DF Election in EVPN but Much Faster



Policy Guided Prefixes (PGP)

- South and North Variant since the "Propagation Direction" is Fixed
  - Avoids the "Colliding Diffused Computation Fronts" Problems
- Propagate like Distance Vector but Based on Flooding
  - No Necessity to Build Specialized Updates "per Peer"
- Ingress Policies Can be Applied on PGPs
  - No Need for "Refreshes" on Policy Changes
- Uses
  - Traffic Engineering like SR

Moreover

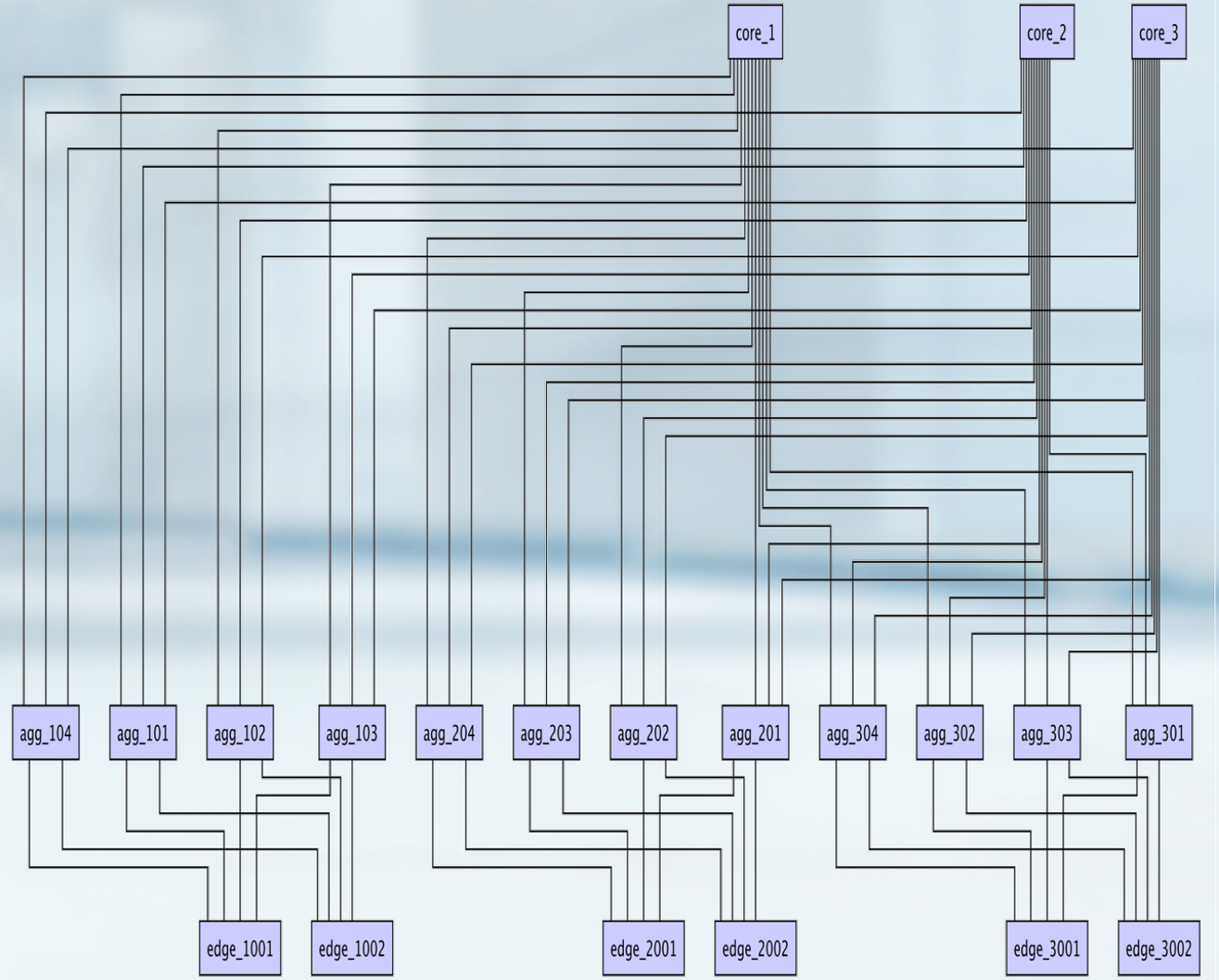
- Traffic Engineering, SR is Included via PGP
- Packet Formats are Completely Model Based
- Channel Agnostic Delivery, Could Be QUICK, TCP, UDP, UDT
- Prefixes are Mapped to Flooding Element Based on Local Hash Functions
  - One Extreme Point is a Prefix Per Flooded Element = BGP Update
- Purging (Given Complexity) is Omitted
- Key-Value Store is Supported (e.g. Service Configuration During Flooding) Including Policies and "Best Copy Tie-Breaking"

Summary of RIFT Advantages

- Advantages of Link-State and Distance Vector
  - Fastest Possible Convergence
  - Automatic Detection of Topology
  - Minimal Routes on TORs
  - Easy to Achieve High Degree of ECMP/N-ECMP
  - Minimal Blast Radius on Failures
  - Fast De-commisioning of Nodes
- No Disadvantages of Link-State or Distance Vector
  - Reduced Flooding
  - Automatic Neighbor Detection
- And Some Neither Can Do
  - Automatic Disaggregation on Failures
  - Scope Controlled Key-Value Store

Sample Comparison to IGP

- 21 Nodes
- 60 links
- 600 Prefixes
- All Run on a Single 4 cores Low End I7
- Comparison RIFT to Equivalent IGP
  - Avg. Node CPU Use: 3x Better
  - Convergence (RIB): 4x Faster
  - Flooding: 4x Less Transmissions



THANK YOU …

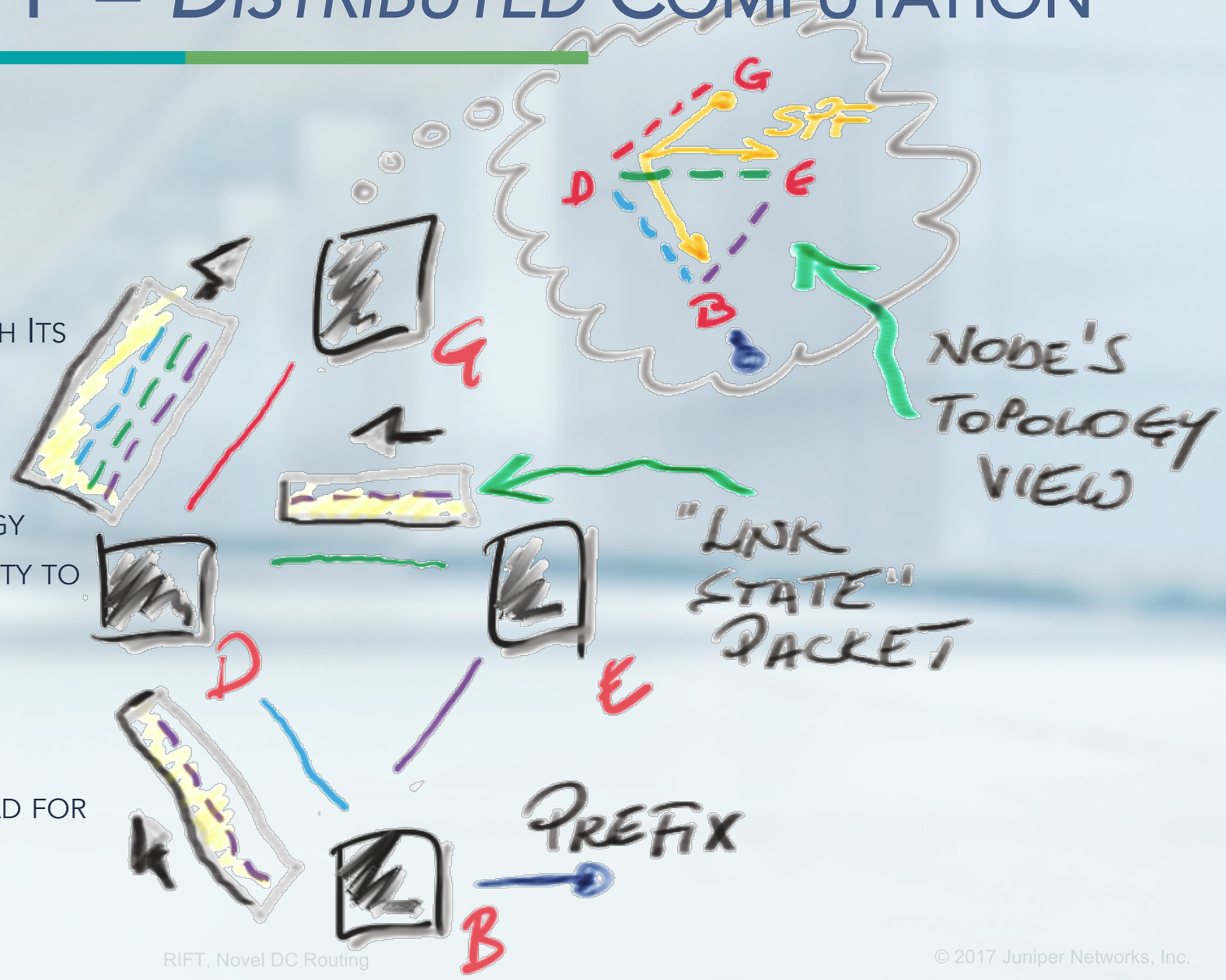
Backup Material

Blitz Overview of Today's Routing

- Link State & SPF
- Distance/Path Vector

Link State and SPF = *Distributed* Computation

- Topology Elements
  - Nodes
  - Links
  - Prefixes
- Each Node Originates Packets with Its 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
- Flooding Generates Excessive Load for Large Average Connectivity
- Periodic Refreshes



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 "Policy" Rather Than "Reachability"
- Scales When Properly Implemented to Much Higher # of Routes Than Link-State