Routing in Fat Trees
Engineering Simplicity in IP Fabric

Melchior Aelmans

MELCHIOR@JUNIPER.NET

AGENDA

- IP Fabrics: The Big Picture
  - Standardized Form-Factor of Network Capacity
  - Next-Gen Evolution Drivers
- Next Gen Underlay Routing Protocol Requirements
- RIFT Basic Concept
- RIFT Advantages
- Standards Status
- Product Status

The BIG Picture

Next Gen IP Fabric Evolution Drivers

IP Fabrics: The Big Picture

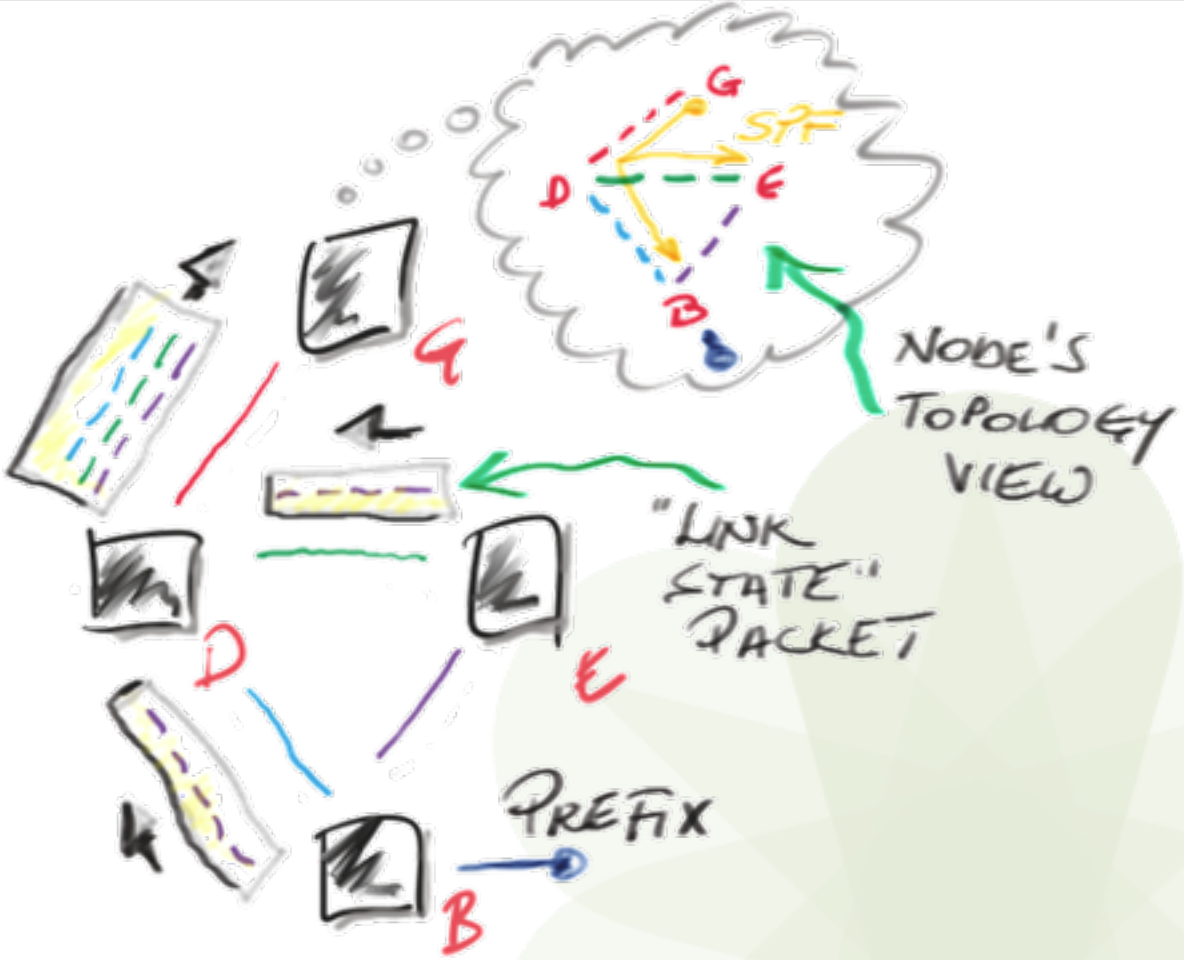
- The Global Data Center Market Size is Expected to Reach Revenues of around $174 Billion by 2023
    - growing at a CAGR of about 4% 2018–2023
- Topologies & Technologies are Unifying Towards a "IP Fabric Substrate" in Proximity to Producers and Consumers of Content
- Fortune 1000 Generate and Host Content and are Building Own IP Fabrics to Consume Larger and Larger Amounts of "Local" Networking Capacity
- Building IP Fabrics Today is Largely High OPEX Artisanal Activity
- Most Fortune 1000 Need a "Standardized Form Factor" for Bandwidth Just Like Storage
- We Lack a Standardized ZTP Underlay Technology to Make IP Forwarding in Fabrics as Simple as Ethernet Switching

What and Why ?

- Hyper-scalers are Extrapolating the Things to Come
  - Vast Amount of Bandwidth Close to Producer & Consumer Are Provisioned
    - IP Fabrics in DC (Server Farms)
    - Metro (Caches and Access)
    - Disaggregated Chassis Architectures in Backbone with Regular Fabrics Holding Leaves Together
  - Those Topologies are Uniform, Local and Regular
    - At the End, Economics of Interconnecting of Crossbars Done in 1950 in Bell Labs are Still Valid
  - WAN-Style Traffic Engineering & Protection is Being Replaced by Wide Fan-Out & Distributed Systems Redundancy (Rather Than Chassis & FRR)
    - Simpler is Cheaper in Opex _and_ Capex
  - Hyper-Scalers are Building Customized High-Opex Solutions to Manage those Fabrics
- IP Fabric is Becoming the New “RAM Chip” to Consume Bandwidth
  - No’one Configures SSD Wear-Leveling, RAM Banks and CAS/RAS Manually in Every Laptop
  - IP Fabrics HW is Largely Commodity Already
    - L3 Forwarding is the New L2
  - IP Fabrics Must and Will “OPEX Commoditize”
- Customers are Hosting Their Content & Critical Business Processes
  - Hybrid Cloud for Many Reasons, One of Them to Keep Real-Estate from Hyper-scalers
  - Need to Build Own Fabrics
  - Hard to **Sustain** Proprietary OPEX Efforts

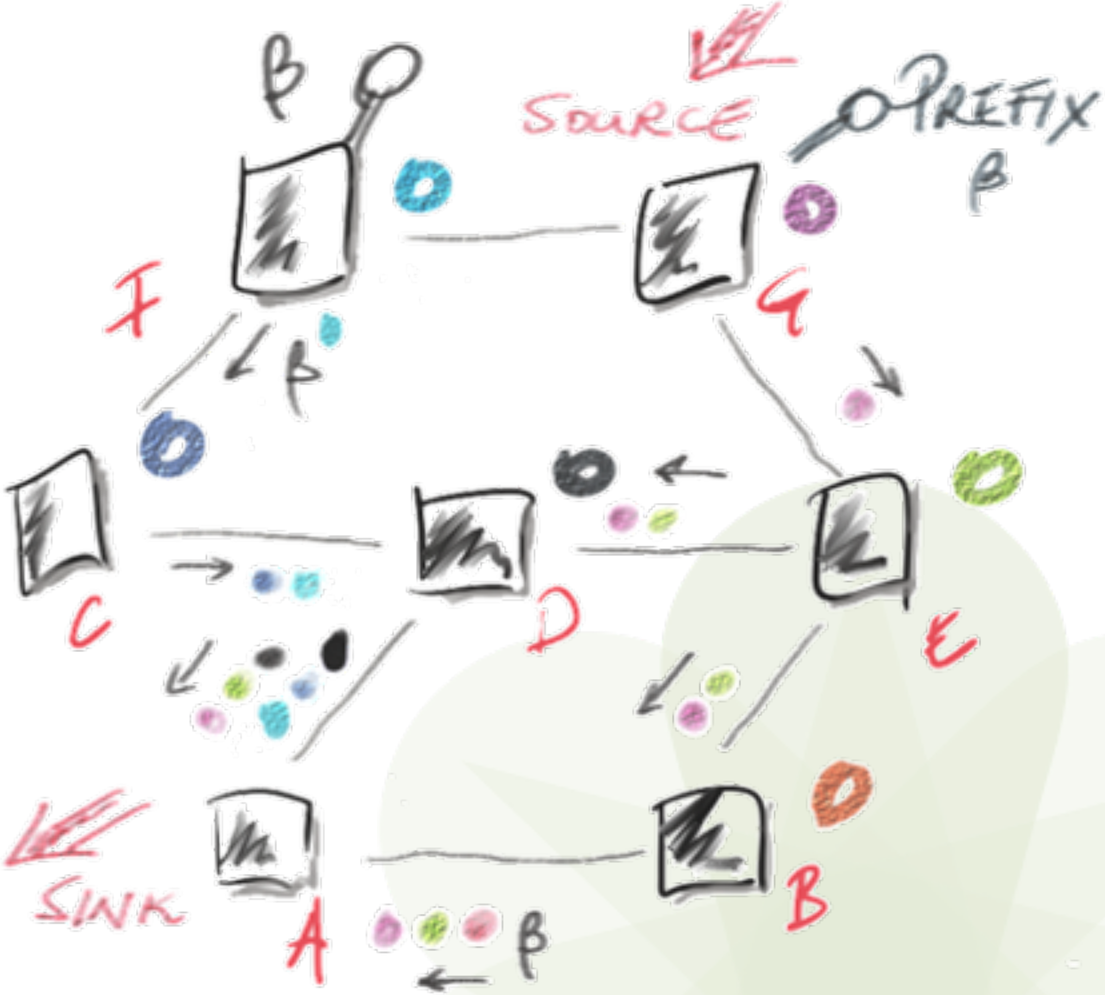
Link State and SPF = *Distributed* Computation

- 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 (modulo Areas)
- Flooding Generates Excessive Load for Large Average Connectivity
- 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 “Policy” Rather Than “Reachability”
- Scales When Properly Implemented to Much Higher # of Routes Than Link-State



Current State of Affairs

- 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
- Others Run IGP (ISIS)
  - Generally a "Better" Approach to Faster Convergence
  - Current Attempts to Deal with Some "Spot Problems" like Flooding Reduction
- Yet Others Run BGP over IGP (Traditional Routing Architecture)
- Less Than More Successful Attempts @ Prefix Summarization, Micro- and Black-Holing, Blast Radius Containment
- Server Multi-Homing not Possible Using IP Due to Equal Cost and Scaling Constraints, Hence MC-LAG'ed Solutions or EVPN
- In Summary: High OPEX Solutions Not Necessarily Viable for Customers Who Cannot or Do Not Want to Build Sophisticated Talent Pool to Deal with Their "Unicorn" Fabrics

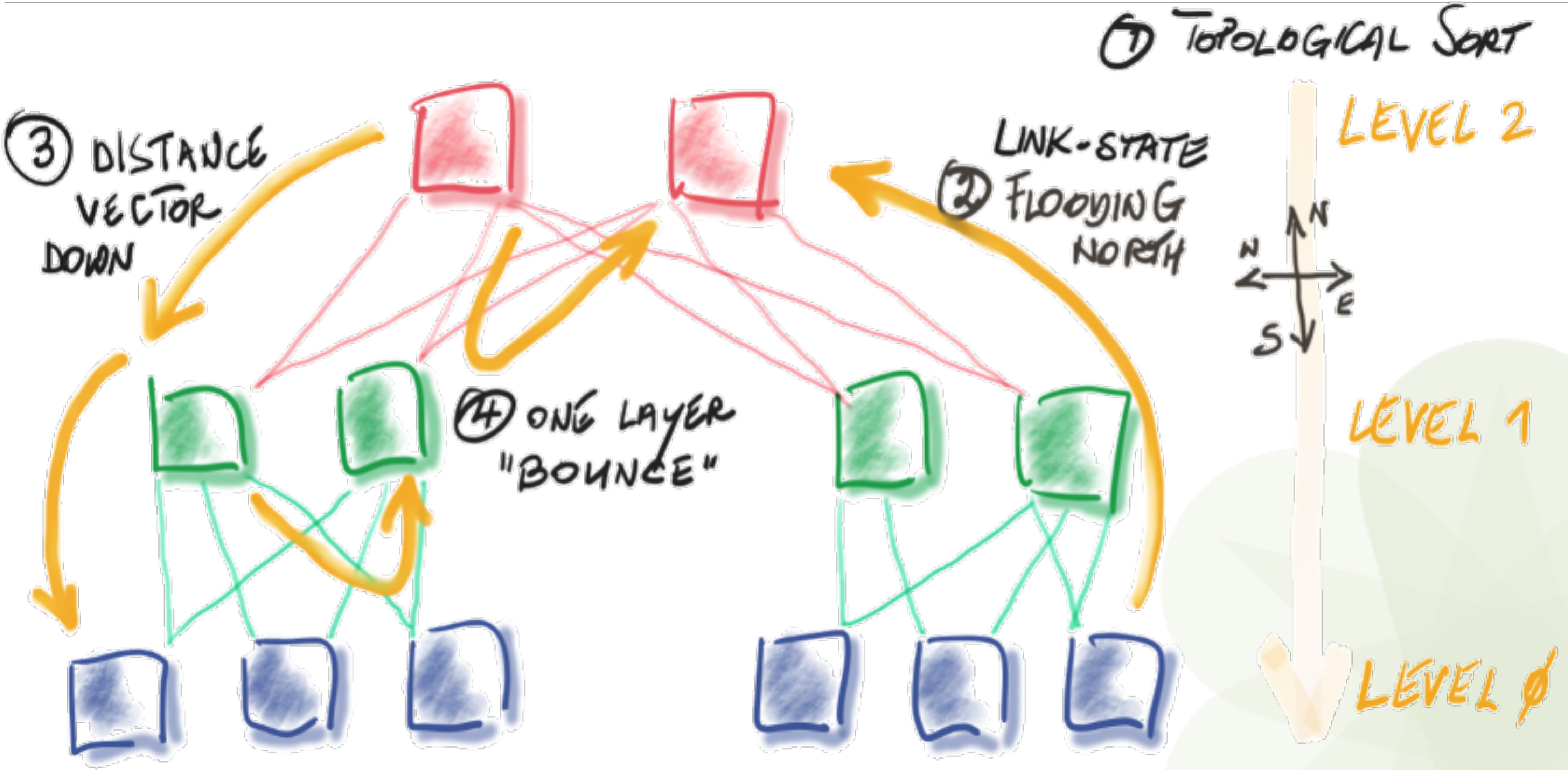
Next Gen IP Fabric Underlay Routing Protocol Requirements

Next Gen Underlay Protocol Requirements

Problem / Attempted Solution	BGP modified for DC (all kind of “mods”)	ISIS modified for DC (RFC7356 + “mods”)	RIFT Native DC
Peer Discovery/Automatic Forming of Trees/Preventing Cabling Violations	⚠	⚠	✓
No Need for Internal Addressing, Minimal Amount of Routes/Information on ToRs, Stretches True Routing to the Multi-Homed Host	✗	✗	✓
High Degree of ECMP (BGP needs lots knobs, memory, own-AS-path violations) and ideally NEC and LFA	⚠	✓	✓
Non Equal Cost Multi-Path, ECMP Independent Anycast, MC-LAG Replacement	✗	✗	✓
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	✗	✓	✓
State of the Art Security Including Originator Validation and Replay Prevention	✗	✗	✓
Simplest Initial Implementation	✓	✗	✓✗

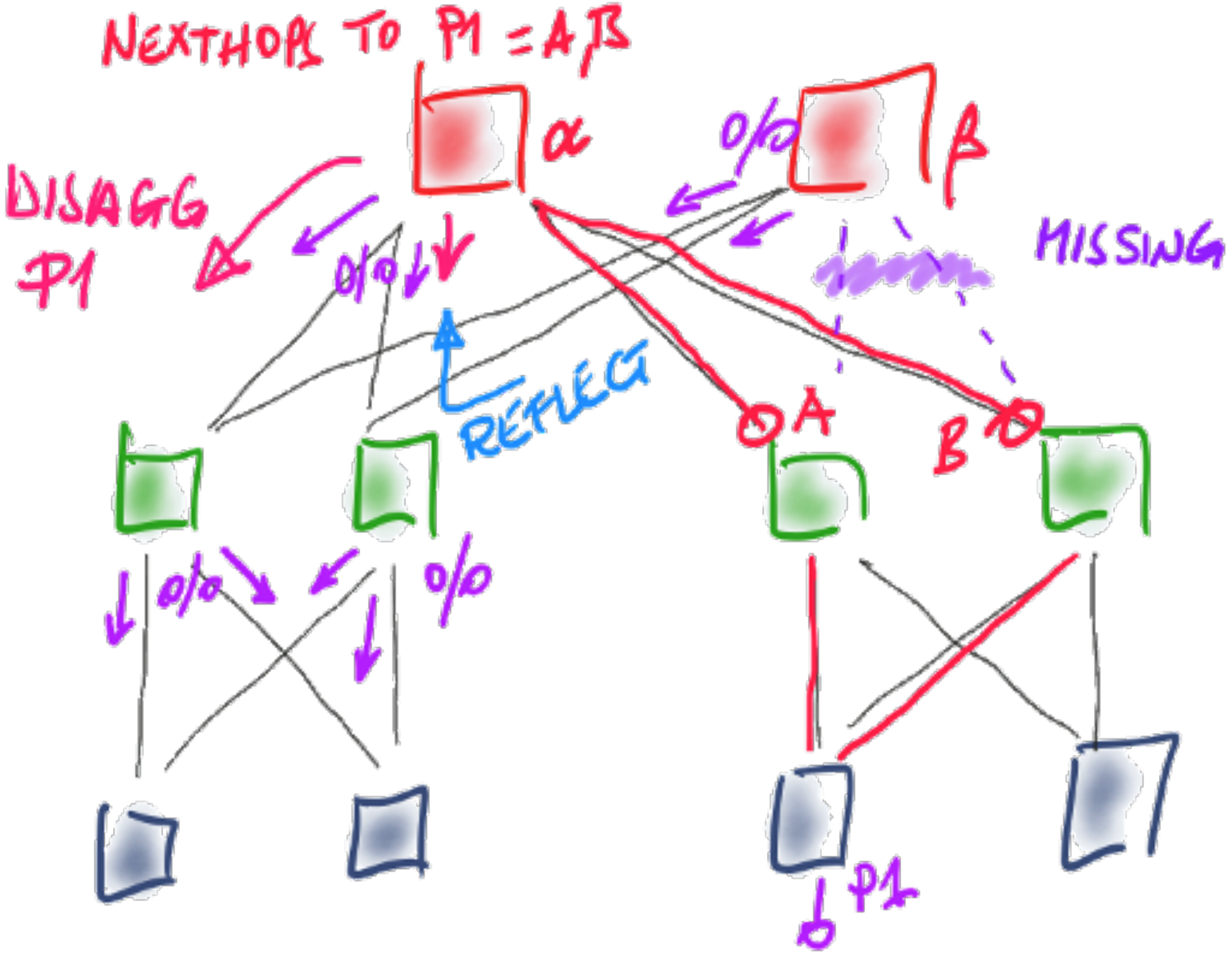
RIFT Basic Concept

RIFT Concept



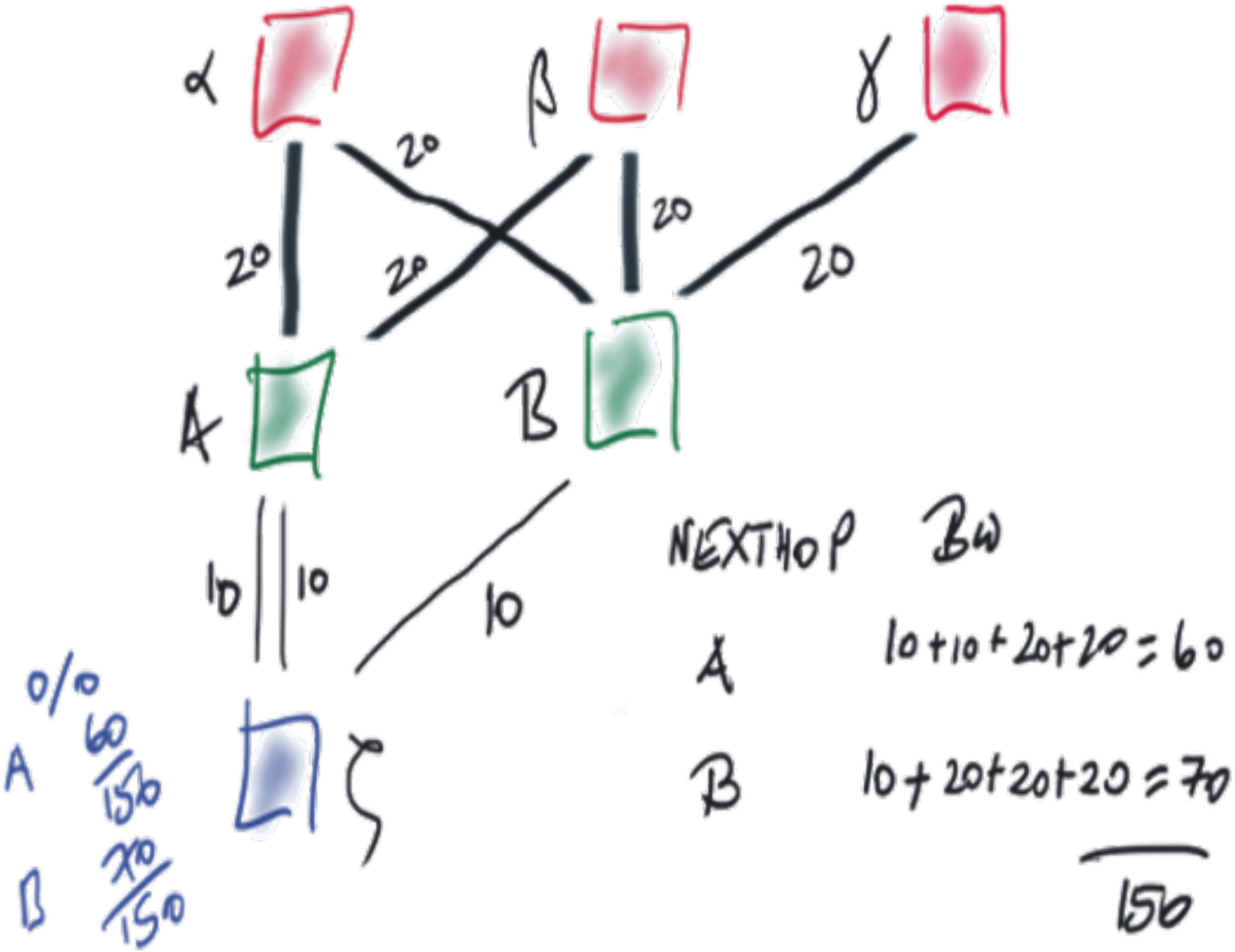
RIFT ADVANTAGES

Automatic De-Aggregation



- South Representation with Adjacencies of Red Beta is Reflected by the Green Layer to Alpha
- Alpha Red Spine Computes P1 Reachability (Tree that Indicates Next-Hops)
- alpha Sees that Beta has no Adjacencies to any of the Available Next-Hops to P1
- Alpha Disaggregates P1
- Left Green Nodes do NOT Propagate Disaggregation Further
- When Beta Obtains a Viable Next-Hop to P1, Alpha Automatically Reaggregates

Northbound Bandwidth Balancing



NEXTHOP	BW
A	10+10+20+20=60
B	10+20+20+20=70
	150

- Zeta Computes Bandwidth Sum to NextHop and Next Level Up
- Through A Cumulative 60 Units
- Through B Cumulative 70 Units
- Default Route Uses That in Some Form to Unequal Load Balance
- Perfectly Safe Since RIFT is Loop-Free

Operational Advantages

- Open IETF Standard
  - Can Build Hybrid Vendor Fabrics
  - Protocol is Well Reviewed and Understood by World-Class Experts
- True ZTP
  - No Configuration Necessary
  - V4 over V6 Forwarding
  - Mis-cabling Handled
- Can Operate on Asymmetric Bandwidth Fabrics and Handle “Fat Link” Failures By Adjusting Automatically
- Can Support and Scale to an Architecture with Multi-Home Servers
  - No Need for Service Migration on ToR Upgrades
  - Can Talk Directly to Hyper-Visors/Kubernetes GW
- BFD is “Built In”
  - Can Be Used for Fast Rehash or Early Loss Detection
- Runs on UDP
  - Trivial Kernel Support on All Platforms
  - Allows for Max. Speed Flooding
  - Easy to “Multi-Instantiate” for Different Purposes

- Minimal Blast-Radius
  - Failures/Bring-Up on Fabric Only Affects the Smallest Viable Radius
- RIFT Flooding is ~30% of Normal Flat IGP
  - Built-In Flood Reduction Reduces Flood Traffic to <20% of Flat IGP
- Loop-Free
  - Can Utilize **All** Viable Paths Through Fabric
  - Can Support True Anycast
- Model Based
  - Much Less Possibility for Parser and Formatter Bugs Plaguing Today’s Networking Protocols
- Specification is Written for Maximum parallelization
  - With Enough Cores IP Switches Should Be Able to Converge @ Speeds Making FRR Unnecessary (Assuming Fast Rehash)
- KV Store Allows to Replace Out-Of-Band Applications
  - IP/MAC Binding Can Be Flooded to Top-of-Fabric
- Sophisticated, Newest Routing Security
- Full Fabric Visibility at the Top

STANDARDIZATION PROGRESS

Standardization Progress

RIFT Standards Track IETF Working Group in Revision -12

- Security Area Directorate Early Revision Done and Addressed
- IESG Revisions Being Addressed
- Co-Authors from Cisco, Yandex, Mellanox, HP & Implementing Individuals
- https://tools.ietf.org/html/draft-ietf-rift-rift-12

Applicability Draft

- Co-Authored with operators, e.g. Orange and Yandex
- Under Gen-ART review
- https://tools.ietf.org/html/draft-ietf-rift-applicability-03

PRODUCT STATUS

Productization

Implementations

Juniper Networks

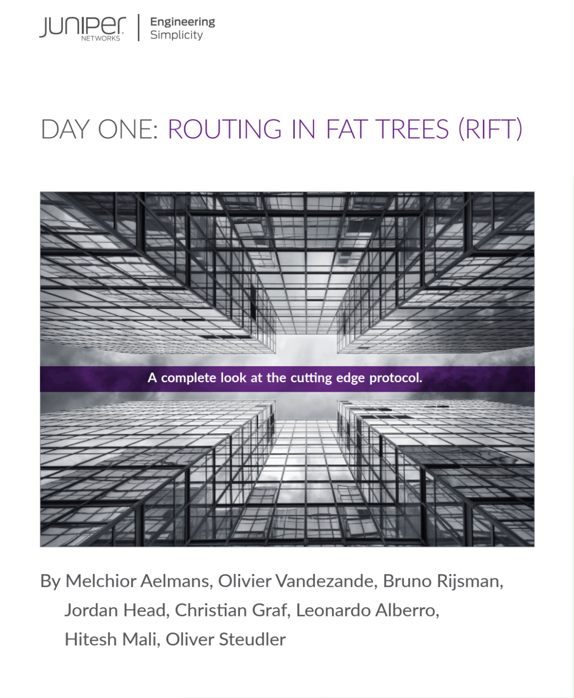
- FCS in 19.4R1 64-bit Junos
- Package installing over 19.4R1 or newer
- QFX (All Variants), vMX, MX
- cRPD & EVO in progress
- https://www.juniper.net/us/en/dm/free-rift-trial/

Python Open-source by Bruno Rijsman

- https://github.com/brunorijsman/rift-python

Documentation

- Day One Book: Routing in Fat Trees (RIFT)
  - covering Juniper and open-source implementation
  - RIFT Wireshark dissector
  - Download free: HTTP://JNPR.NL/RIFT



JUNIPER NETWORKS | Engineering Simplicity

DAY ONE: ROUTING IN FAT TREES (RIFT)

A complete look at the cutting edge protocol.

By Melchior Aelmans, Olivier Vandezande, Bruno Rijsman,
Jordan Head, Christian Graf, Leonardo Alberro,
Hitesh Mali, Oliver Steudler

RIFT: Engineering Simplicity in IP Fabric

D

Questions?

Melchior Aelmans

MELCHIOR@JUNIPER.NET

JUNIPER NETWORKS | Engineering Simplicity