SP Routing Innovation with Segment Routing, VXLAN and EVPN

Claudiu Captari
Systems Engineer
claudiu@arista.com

ARISTA

Agenda

- Cloud Principles and Cloud-Grade Routing
- Cloud-Grade Routing Innovations
  - Scale-Out Architectures with Merchant Silicon Platforms
  - Simplify Operations with Modern Open Protocols
  - Software-Driven Control for Automation and Visibility
- Key Takeaways

Cloud Principles Applied to Network Transformations

Legacy Networking		Cloud Networking
Rigid Architecture	→	Scale-Out
Inefficient Operations	→	Simplify
Inflexible Service Delivery	→	Software-Driven Control
Proprietary Protocols	→	Standards-Based Protocols
Custom Silicon	→	Merchant Silicon

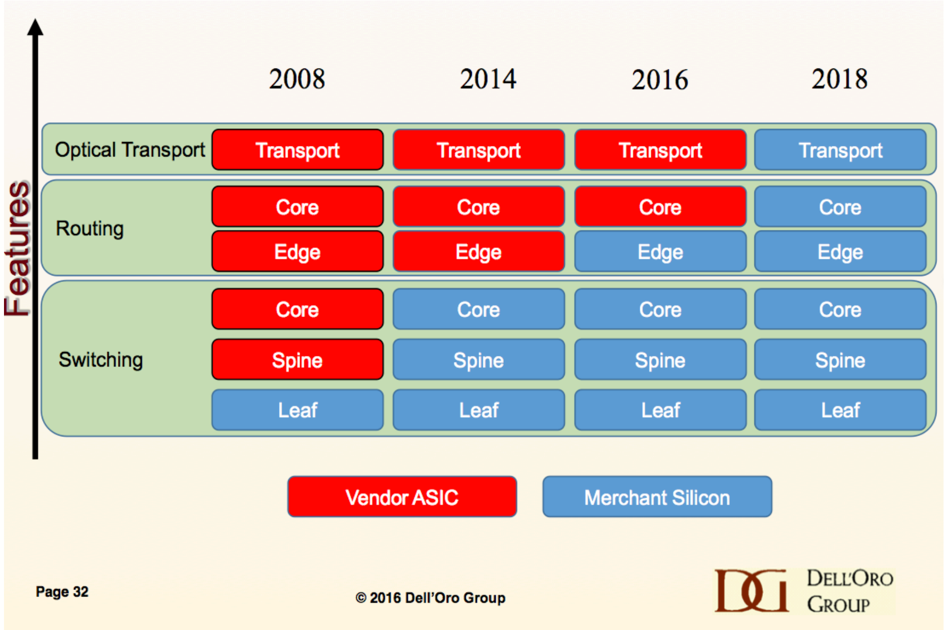
Software

Hardware

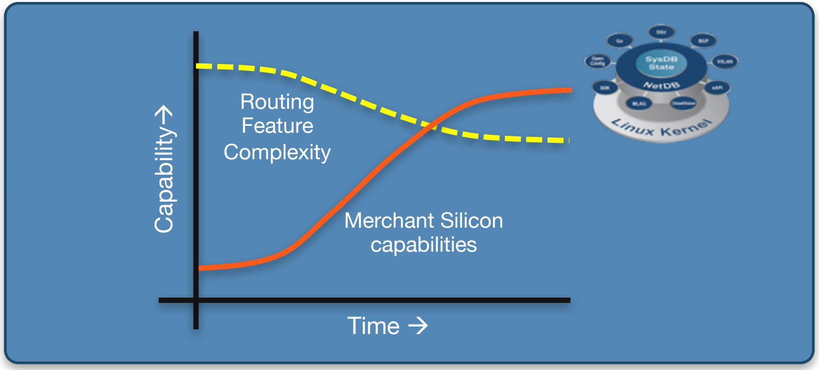
Cloud Principles have driven Compute, Switching, Storage… and now Routing!

ARISTA

Merchant Silicon Influence in Mega Cloud Providers



Broadcom 'Jericho' Silicon



ARISTA

NETFLIX

Dave Temkin
VP Networks

Nov 2017

Dave Temkin
@dtemkin

Super proud of my team - today they removed the last "big expensive router" from our network; [...]

Inexpensive commodity switches run the entire Netflix Open Connect CDN!

5:24 AM - 10 Nov 2017

270 Retweets **602** Likes

29 270 602

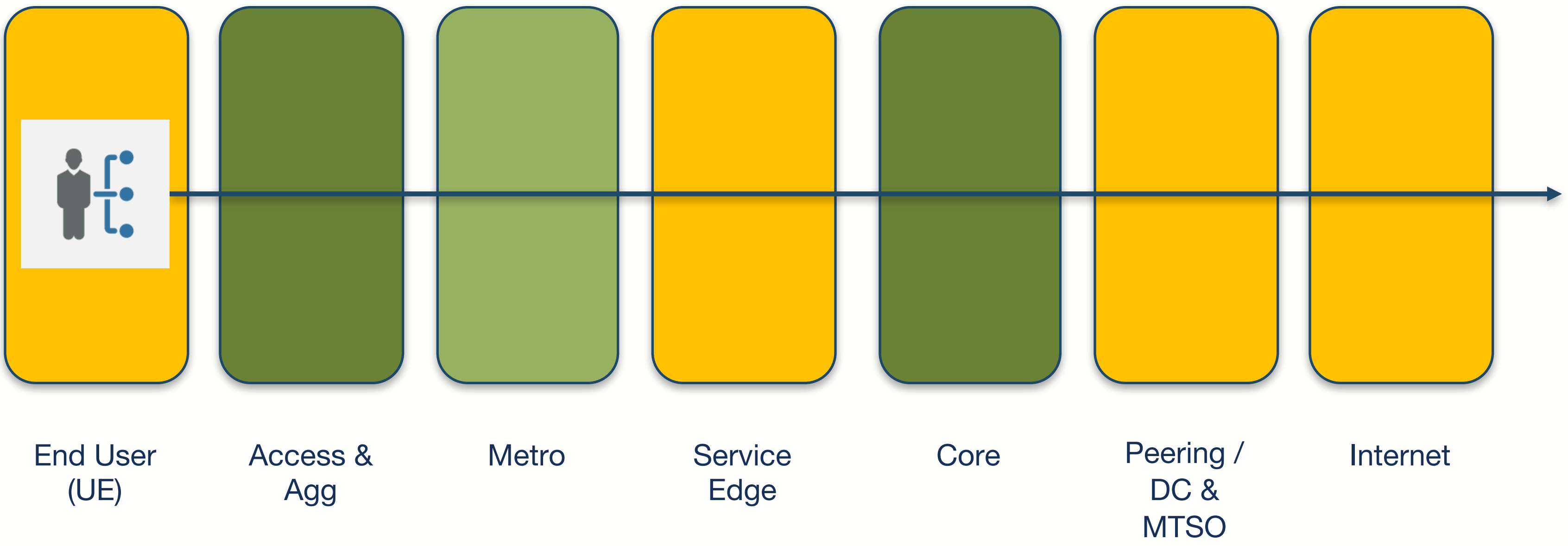
ARISTA

Building Cloud-Grade OS

Modern OS	• Lean OS • Programmable @ all layers
Spine/Leaf Optimized	• Enhanced Load Balancing • Optimal convergence
Automation	• NETCONF/YANG • Turnkey Automation
Monitoring @ Scale	• Large-scale ECMP, Monitoring with BMP • OpenConfig → State Streaming
Agile Certification	• Virtual/Container based OS to simulate large-scale networks

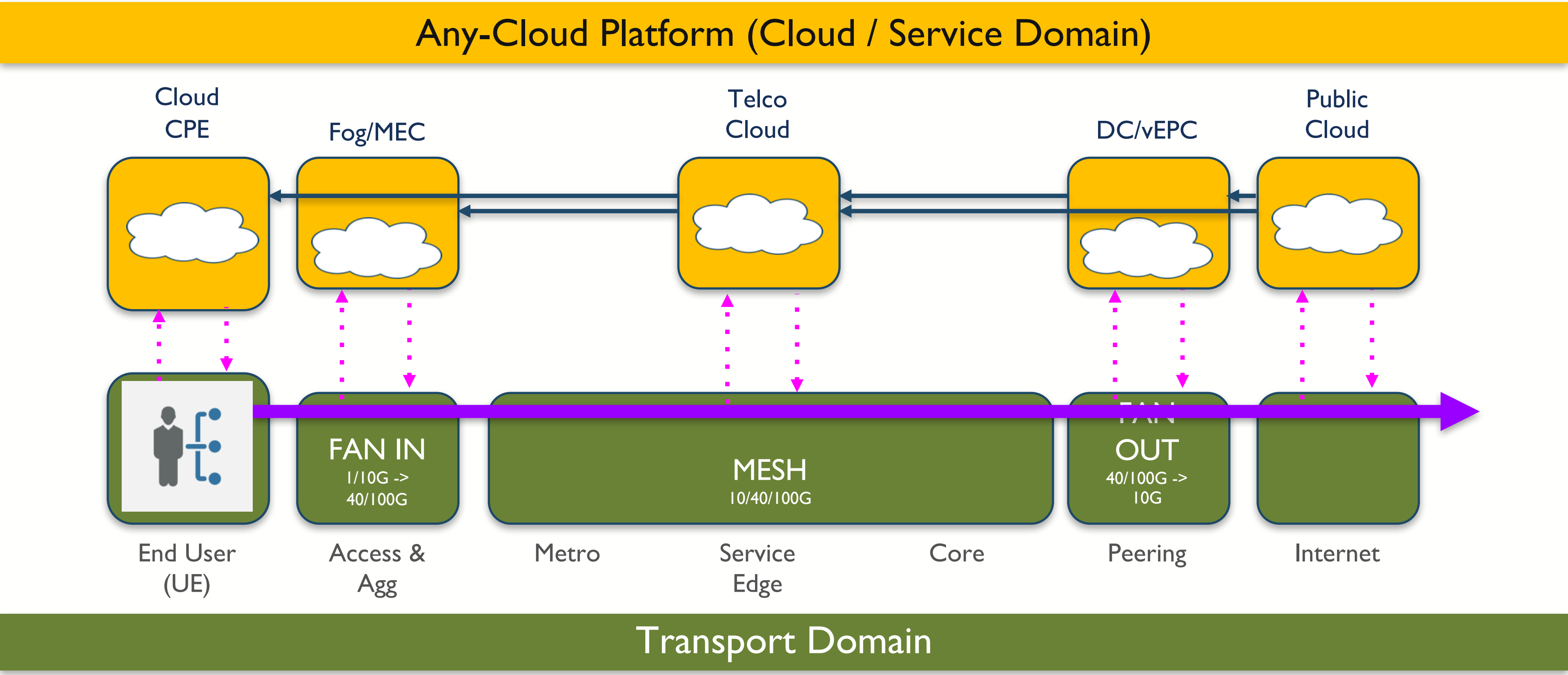
Fundamental Requirements of a Cloud OS

Traditional Service Provider Architecture Building Blocks



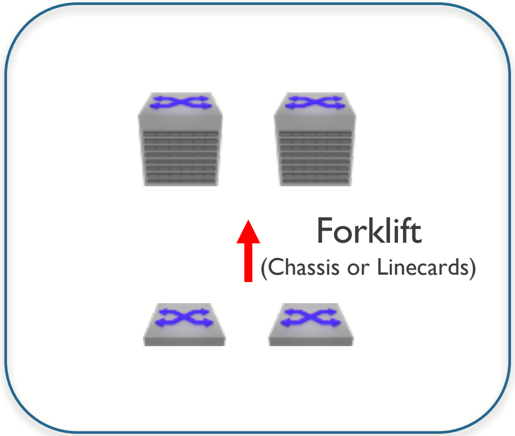
ARISTA

Connecting Transport to Places In The Cloud (PIC)

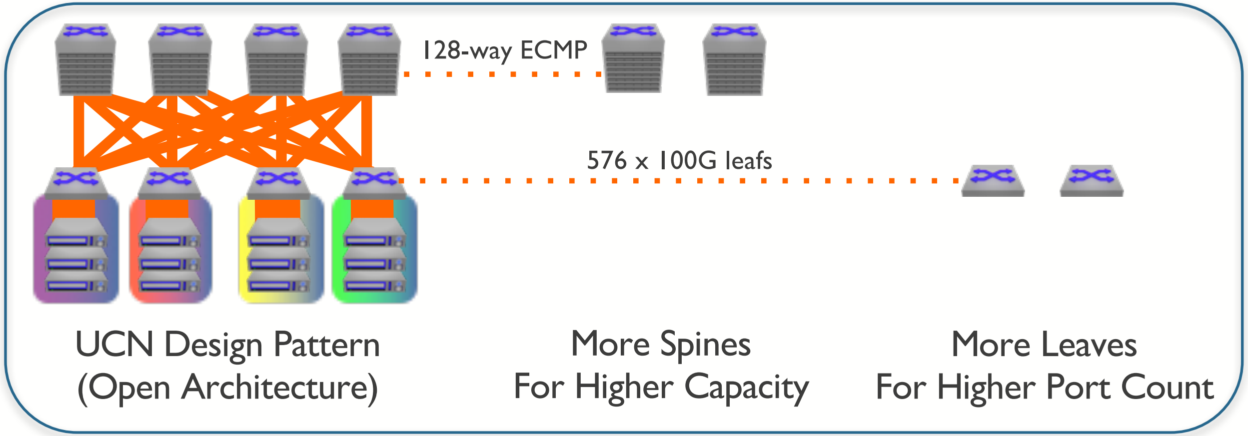


Scale Up vs Scale Out

Scale Up (Legacy)



Scale Out



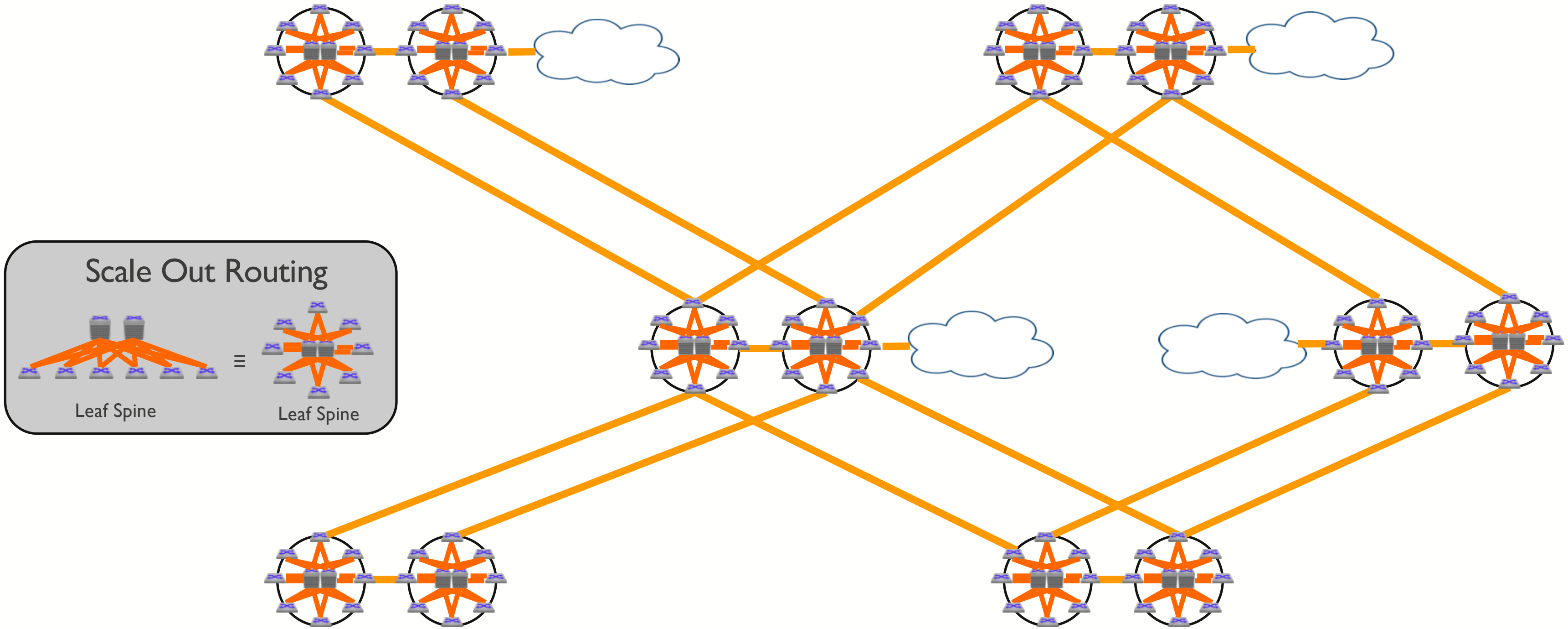
Legacy (1+1)

- **Single Vendor Solution**
- Difficult to Upgrade Software
- Additional Capacity Via System or Linecard Replacement
- Forklift Lowers Investment Protection
- 100% ‘Peak Capacity’ Loss During Outage/Maintenance

Leaf / Spine (N+1, N = Peak Capacity)

- **Multi Vendor Solution**
- Hitless Upgrades
- Effortless Scale (Spine for Capacity, Leaf for Ports)
- Higher Investment Protection Without Forklifts
- 1/N% ‘Peak Capacity’ Loss During Outage/Maintenance (25% n=4, 6.25% n=16, 0.8% n=128)

Building an SP Core Using Disaggregated Routing



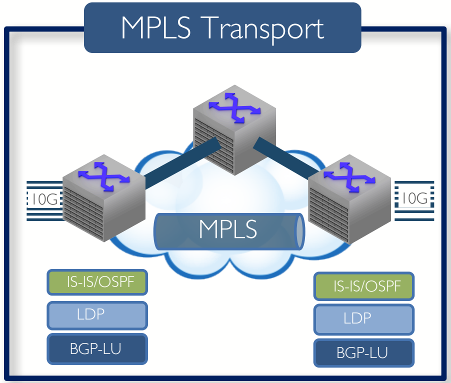
Use Of Merchant Silicon Allows For Cost Effective Consolidation of L2 and L3 Elements At Each Location

ARISTA

Core Transport Summary

Traditional Transport

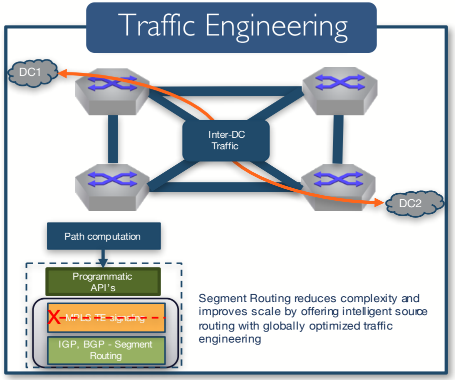
Multi-label MPLS Transport



High Density 10G/100G merchant silicon for MPLS transport

Segment Routing

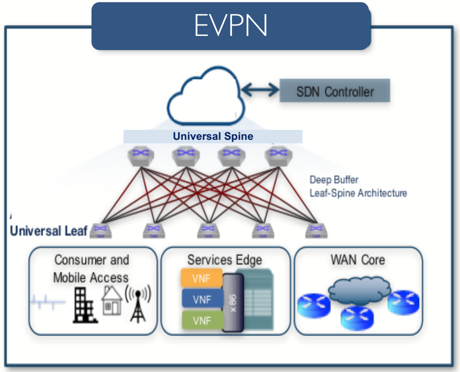
SDK programmability or third-party PCE integration



Scalable and Simplified, Traffic Engineering

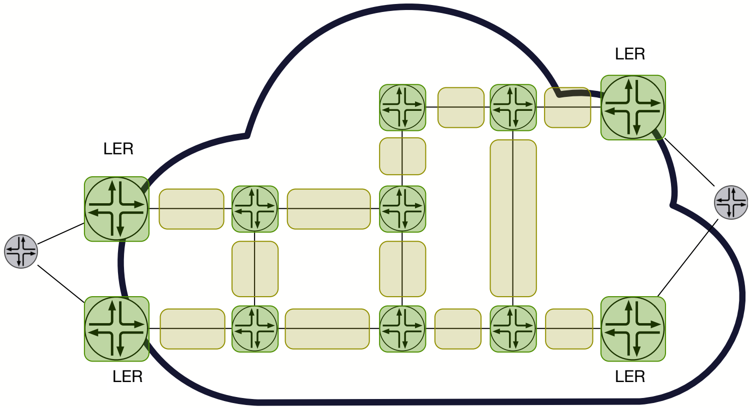
Network Services

EVPN for Layer 2/3 VPNs for Cloud, NFV and PE



Single scalable BGP control plane for Layer 2 and 3 VPNs

What is Segment Routing (SR)?



Basic Philosophy

- Reuse IGP and BGP to distribute the labels
- Simplify
  - Protocols required - eliminate need for additional signaling protocols
  - Removes per tunnel state (control and data-plane) though out the network
- Provide ECMP
- Encode source routing using MPLS label stack in the data-plane

Concepts

- Segment Routing envisions the network as a collection of ‘topological sub-paths’ – also called ‘segments’.
- Global labels
- Local labels
- Packet is transmitted from source with a list of Segment IDs (or SIDs)

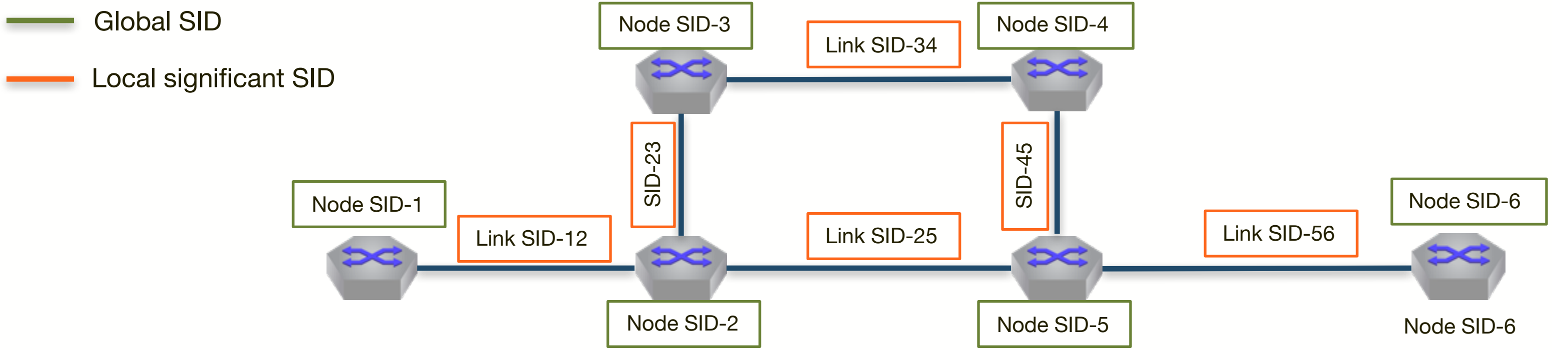
Applicability

- Non-TE **Replacement** -> Remove redundant signaling protocols (LDP), and follow SPF
- TE **Alternative ->** External Controller for fine-grained or Macro TE

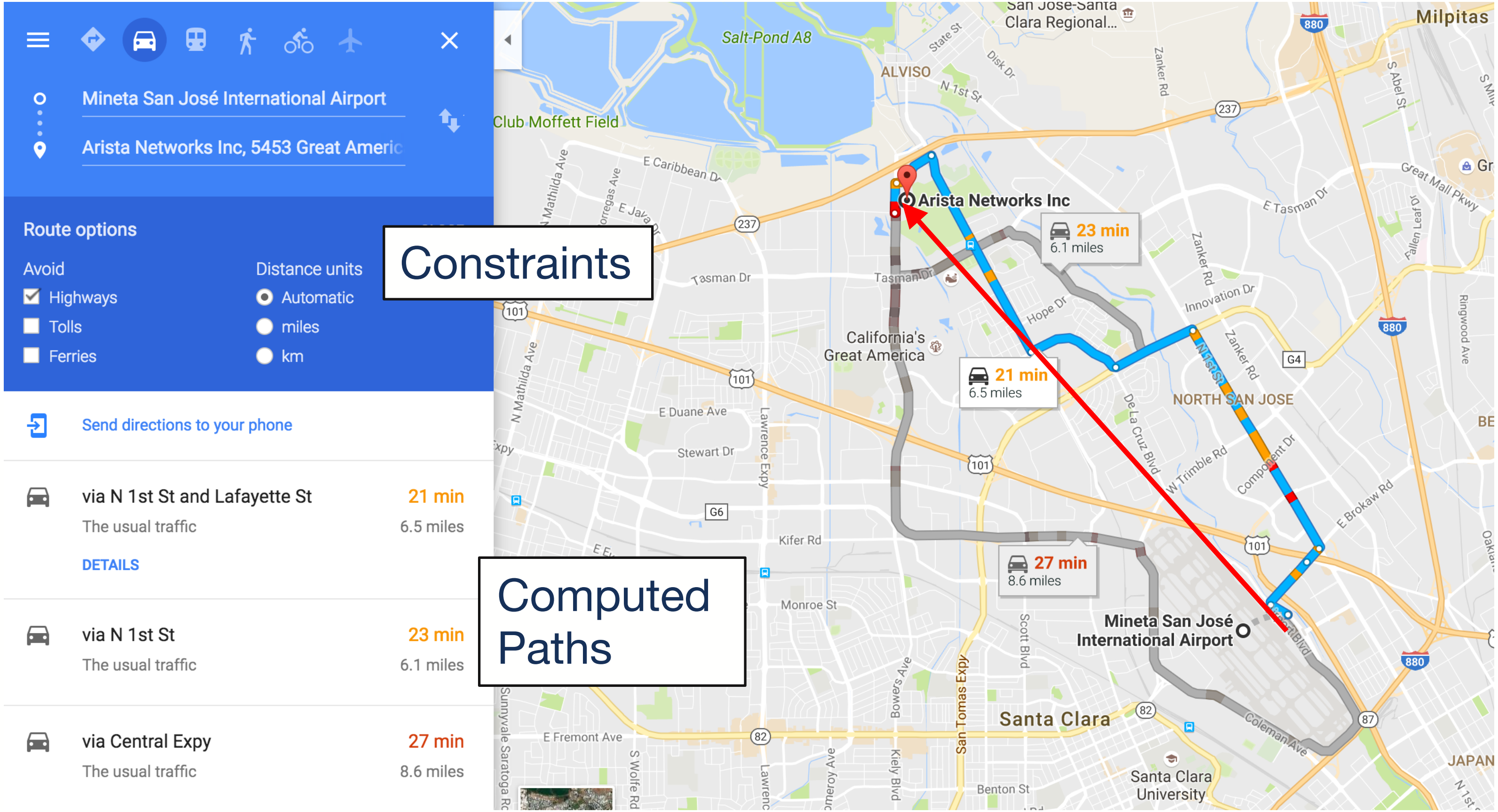
https://www.arista.com/assets/data/pdf/Whitepapers/MPLSSegmentRouting_Whitepaper.pdf

Segment Routing – Operation Overview

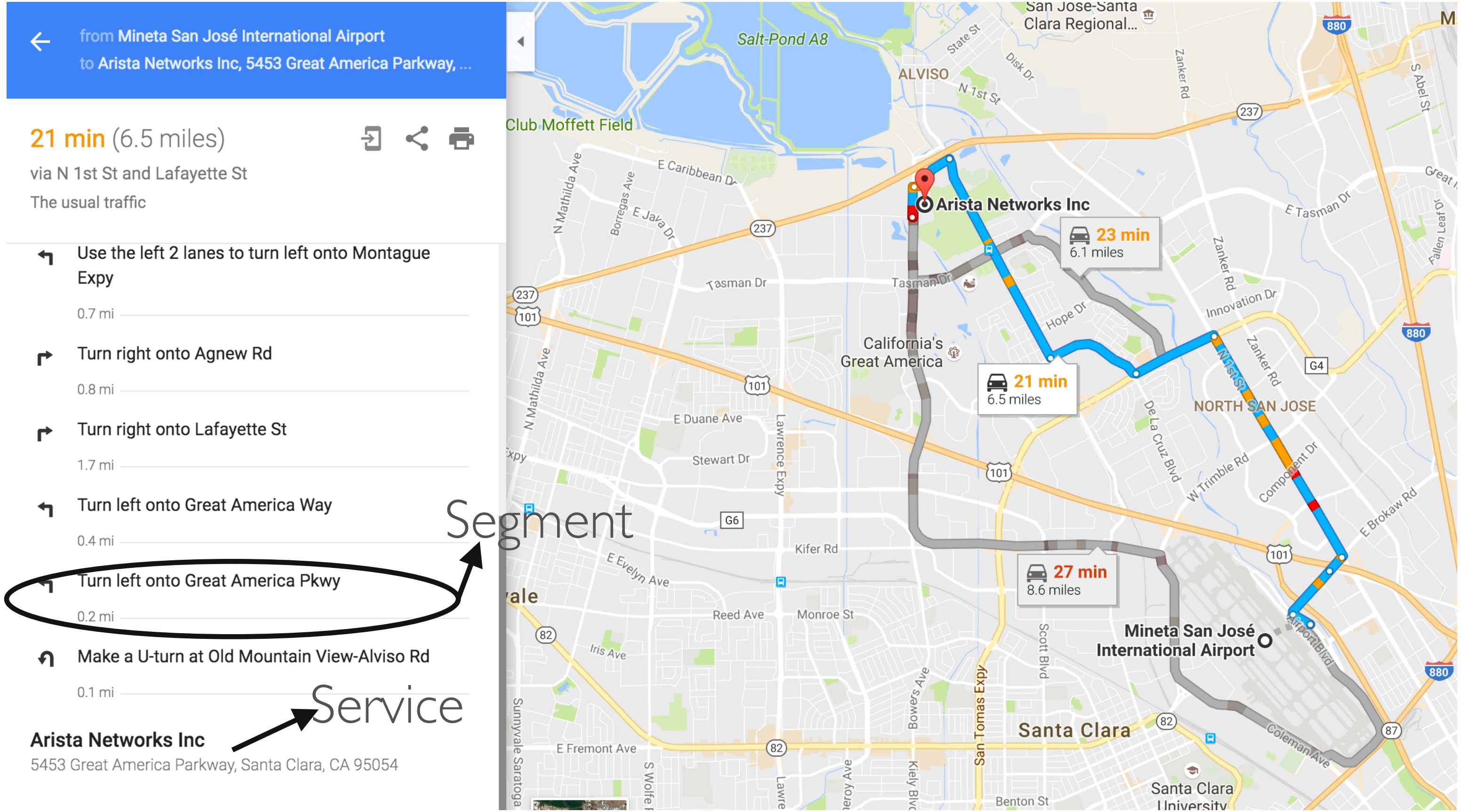
- SR divides the network into “segments” identified by a Segment ID (SID)
  - Global SIDs identify nodes (loopback ip), prefix or Anycast SID (shared loopback IP)
    - All nodes in the SR domain use same SID to identify the prefixes, node or Anycast an SID – reducing data plane state
  - Local significant SIDs identify, the Adjacency links in the network
    - Only the originating Node understands the advertised Adjacency SID
  - Both local and global SIDs are advertised as TLV extensions to the IGP (IS-IS/OSPF)
  - The SID is encoded as an MPLS Label in the forwarding plane



Segment Routing Analogy



Segment Routing Analogy



ARISTA

Segment Routing – Evolution of Core Routing

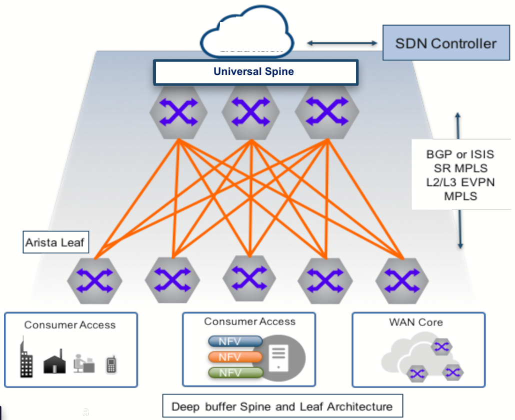
	LDP	RSVP-TE	SR
Overview	MP2P	P2P	MP2P
Operation	Simple	Difficult	Simple
Separate Label Distribution Protocol	Yes	Yes	No
Dependencies	Relies on IGP	Relies on IGP extensions	Relies on IGP
Label Allocation	Locally significant	Locally significant	Global (local ADJ SID)
MPLS ECMP	Yes	No	Yes
Traffic Engineering (TE)	No	Yes	Yes
TE Scale	N/A	Medium/Low N(N-1)	High
Fast Reroute	Partial LFA (<100%)	Yes Node/Link Protection	Yes TI-LFA
Multicast	Yes mLDP	Yes P2MP LSP	No Deployed With Parallel MC Control Plane
IPv6	Limited Extensions Required	Limited Extensions Required	Native

A Transformation in Routing Protocols is Required

Source: MPLS Segment Routing, Driving a modern approach to MPLS transport - https://www.arista.com/assets/data/pdf/Whitepapers/MPLSSegmentRouting_Whitepaper.pdf

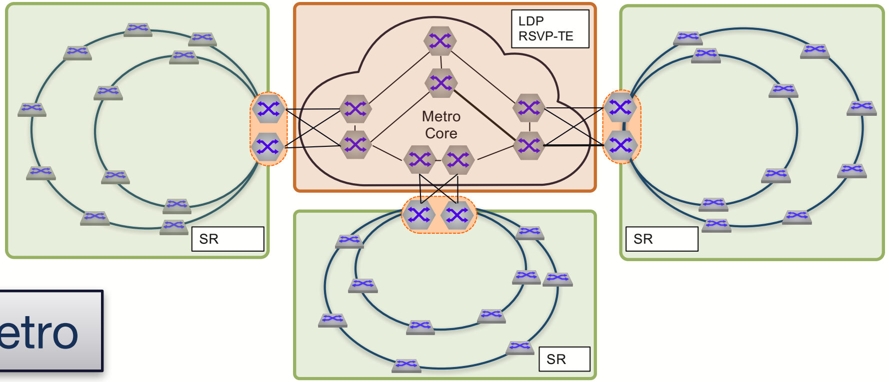
ARISTA

SR Use Cases for SPs



NFV CO

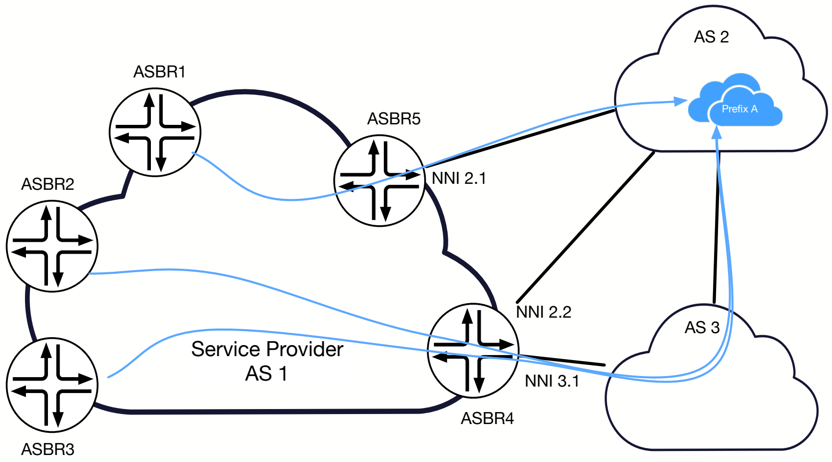
- MPLS in the DC/POP
- SR based L-S DC design
- EVPN MPLS for L2 or L3 EVPNs



Metro

- Solving the “Ring” topology problem
- Allows moving to L3 design
- May inter-work with RSVP-TE in core (binding SID) or tunnel over LDP (SR Mapping Server)
- Mainly looking at SR for TI-LFA

Egress Peer Engineering - Traditional approach



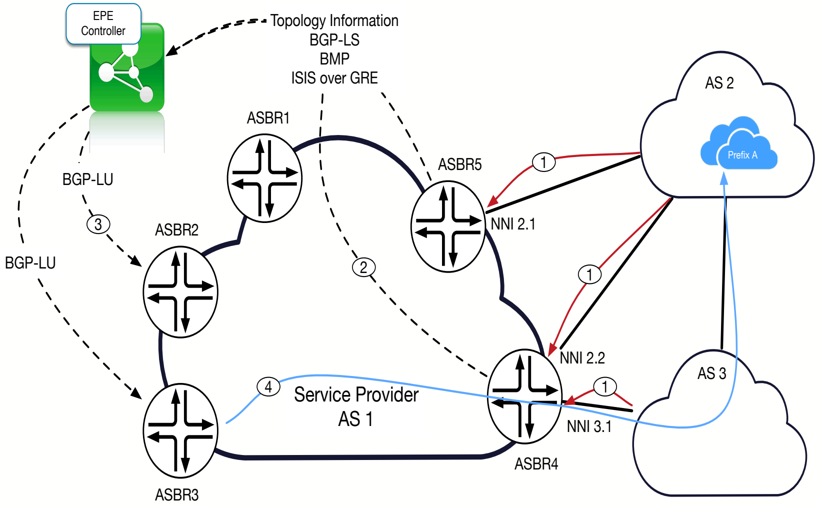
- Policy on ASBR4 set high local-preference for prefix A from AS3
  - Preferred path for Prefix A via NNI 3.1
- Packet with destination IP address matching prefix "A"
  - All sent out of NNI 3.1, regardless of the ingress ASBR
- If ASBR 4 and 5 visible to each of the ingress ASBRs
  - Policy on ingress ASBR's may choose and egress ASBR (ASBR 4 or 5)
  - Can chose ASBR for exit but still not the actual NNI.

Approach limited to

- Egress link selection per destination prefix, on egress ASBR
- Egress ASBR selection per Prefix on ingress ASBR, but without visibility to egress link
- No Global Policy!!!

ARISTA

Egress Peer Engineering - SR approach

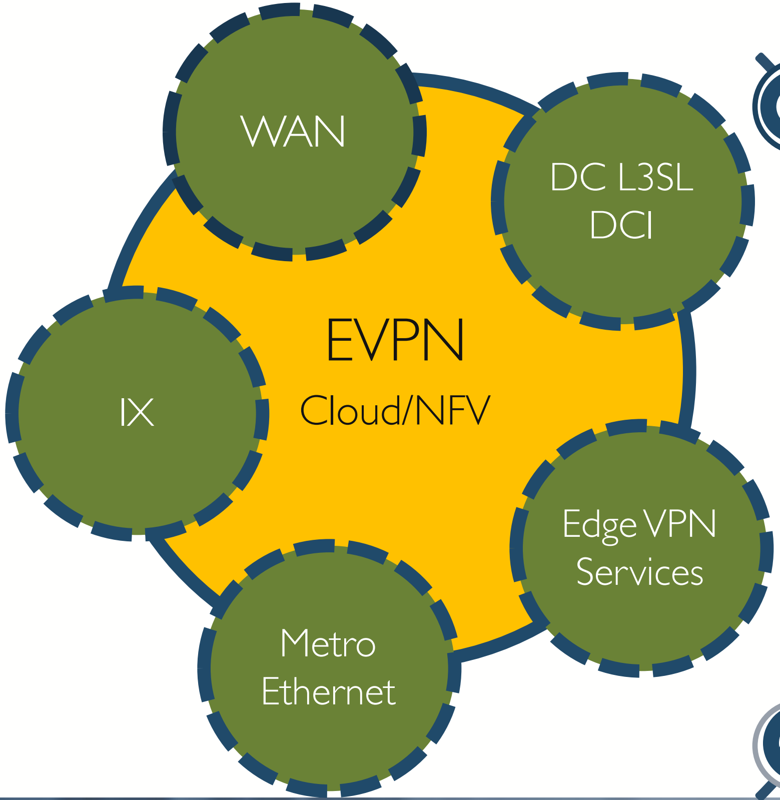


- Path advertised by Peer AS
- ASBR advertise all path and link state info to the controller via BGP-LS
- From LS information and constraints controller computes best path for prefix A.
  - Path decision can be per ingress ASBR
  - Including the NNI interface in the path along with the intra-AS path within AS1
- ASBR3 encapsulates traffic with Segment path to reach ASBR4 and exit NNI3.1.

- Internal path provisioning from ingress ASBR to egress ASBR, GRE, LDP, SPRING Node-SID, RSVP-TE
- Label for NNI selection on egress ASBR- BGP-LU and SPRING peering-SID

EVPN – Extending Cloud Into Routing Services

EANTC — eVPN use-cases demonstrated in Public Multi-Vendor Routing Interop Test – April 2018
[MPLS + SDN + NFV World Congress 2018]



- **Standards Based** - Open Standards for Inter-operation
- **Flexible Service Types** – E-Line/E-LAN & IP VPN Services
- **Universal BGP Control Plane** – Simplify, Standardize
- **Scalable** – BGP Based Scalable VPN Services
- **Secure** – MAC Mobility, ARP Suppression, Policy Control
- **Efficient** – Multi-Homing Support, A/A Forwarding, Scaled ECMP
- **Supports Different Encapsulation Types** – IP (VXLAN) & MPLS

MPLS EVPN – Layer 3 VPNs

- Provide Layer 3 VPNs across a MPLS transport
  - Alternative solution to IP-VPNs (RFC 2746/RFC 4364)
  - BGP control-plane with EVPN NLRI (RFC 7432)
  - Type 5 route to advertise IP prefixes to emulate an IP VPN like service
  - Prefix advertised with Route-Target (RT), Route-Distinguisher (RD) and MPLS label

Next-hop for prefix = PE IP **AFI = 25 (L2VPN) SAFI = 70 (EVPN)**
Route Type (IP Prefix Route, Type 5)
Length

Route Distinguisher (RD)
Ethernet Segment Identifier =0
Ethernet Tag ID = 0
IP prefix Length (0-32, 0-128)
IP Prefix (ipv4 or ipv6)
GW IP Address = 0
MPLS Label

- Extended Community Route Target
- Tunnel-encapsulation extended community

Route Types – Type 5 (EVPN MPLS)

- EVPN MPLS Type-5 Route

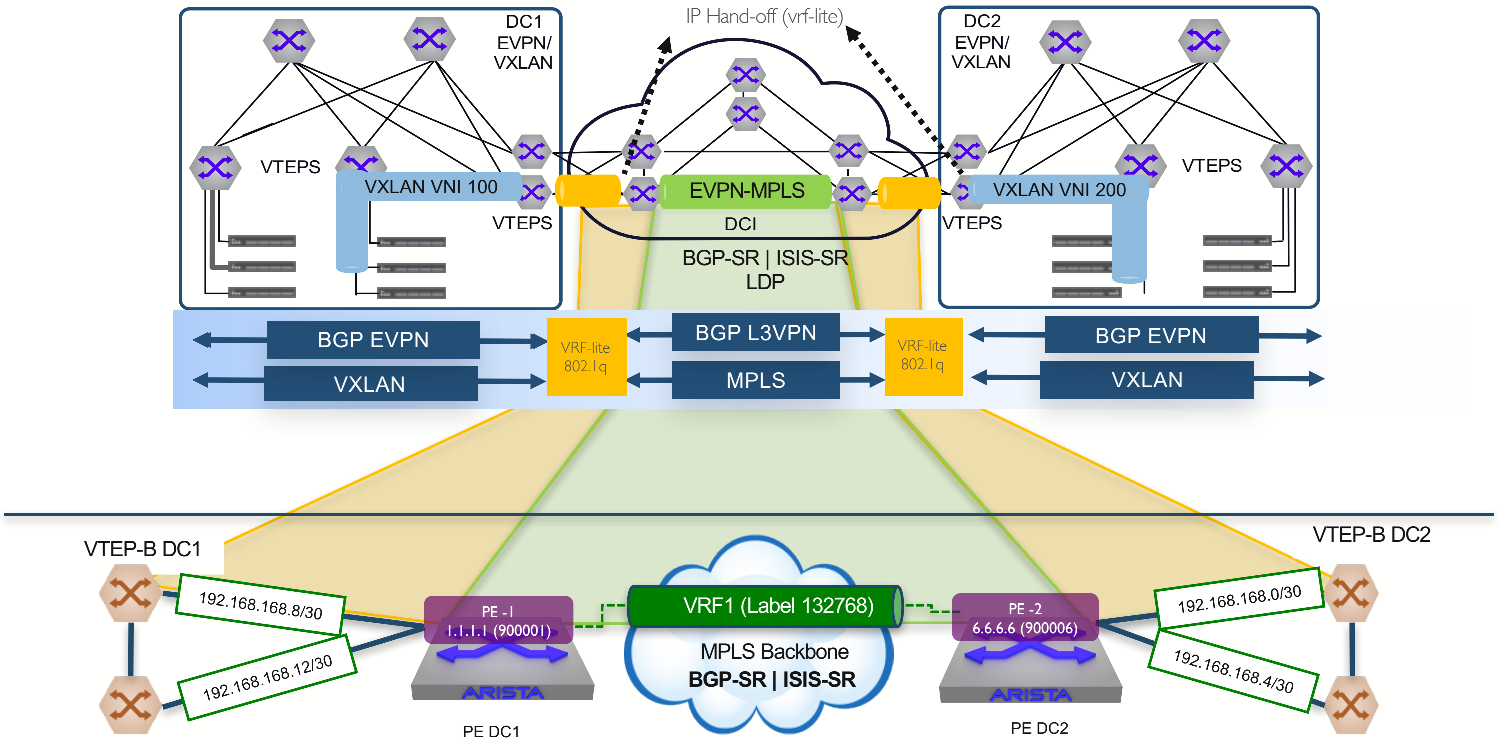
```
PE-2(config)#show bgp evpn route-type ip-prefix ipv4 rd 1.1.1.1:65001 detail
BGP routing table information for VRF default
Router identifier 1.1.1.2, local AS number 65001
BGP routing table entry for ip-prefix 10.10.10.0/24, Route Distinguisher: 1.1.1.1:65001
 Paths: 1 available
  Local
    2.2.2.1 from 1.1.1.1 (1.1.1.1)
      Origin IGP, metric -, localpref 100, weight 0, valid, internal, best
      Extended Community: Route-Target-AS:1001:1001 TunnelEncap:tunnelTypeMpls
      MPLS label: 116384
BGP routing table entry for ip-prefix 10.10.12.0/24, Route Distinguisher: 1.1.1.1:65001
 Paths: 1 available
  Local
    2.2.2.1 from 1.1.1.1 (1.1.1.1)
      Origin IGP, metric -, localpref 100, weight 0, valid, internal, best
      Extended Community: Route-Target-AS:1001:1001 TunnelEncap:tunnelTypeMpls
      MPLS label: 116384
PE-2(config)#
```

IP-prefix and RD being advertised

Next hop (2.2.2.1) for the prefix and advertising Router (1.1.1.1)

Route-target (1001:1001), encapsulation (MPLS) and lable for the VPN route (116384)

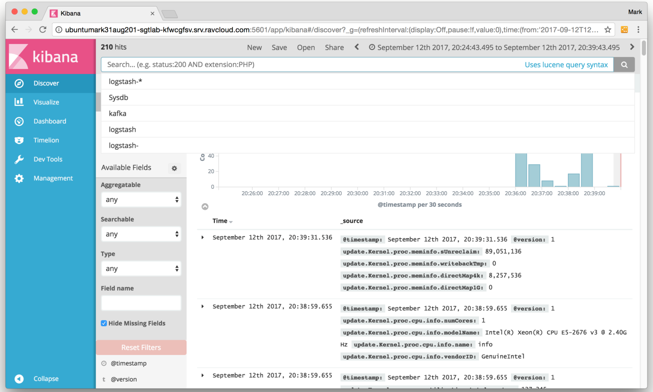
Use Case: MPLS L3 EVPN DCI

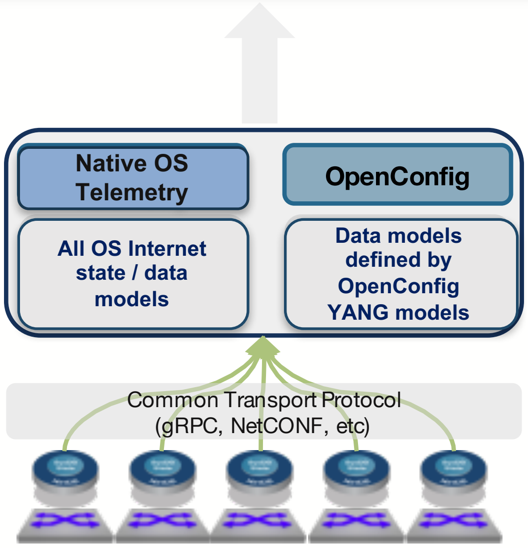


ARISTA

Traffic Steering: Automation, Telemetry & State Streaming

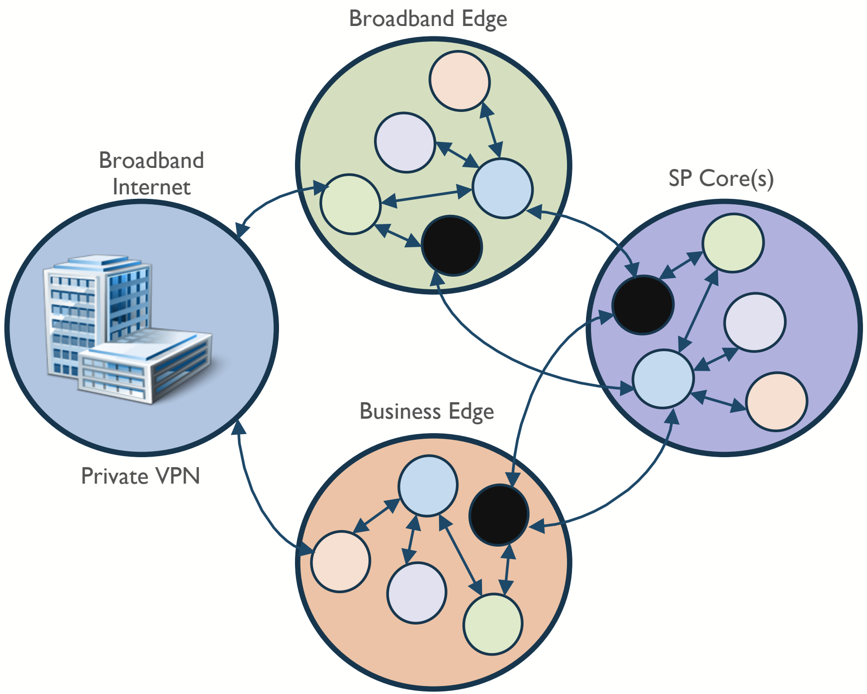
- Access to ***all*** state in the system via standardized (OpenConfig) models
- Full device configuration management via OpenConfig models + CLI
- Supported across all devices
- Standard gRPC transport
  - A transport layer with efficient data encoding!





ARISTA

Summary



TODAY		TARGET ARCHITECTURE
Inconsistent operational models	→	Programmability, Automation, Telemetry, Visibility Flexible Config Monitoring
Expensive Fixed Routers	→	Merchant Silicon *Economics, Scale, Services* *BGP, Route Scale*
Slow convergence	→	Scale-Out & Simplify Rapid Route Convergence
Disparate Protocols	→	Consistent Open Protocols SPRING/SR, EVPN, (SIMPLIFY)

ARISTA

Thank You

www.arista.com, eos.arista.com

www.youtube.com/user/AristaNetworks

http://github.com/arista-eosplus

ARISTA