Design Experience of an SRv6 uSID Data Center

Gyan Mishra

IT Technologist & Innovation Specialist

Associate-Fellow – Network Design

April 9th 2024

Objectives for Todays SRv6 uSID Data Center Presentation

- Building on the foundation from last year
- Intricate design experience details of an SRv6 uSID Data Center
- Review challenges of the innovative solution
- Highlight the transformational potential for the technology
- Comparative analysis to spotlight the efficiencies and performance enhancements outpacing todays technologies at record speed
- Navigate real world design experiences and possible tangible outcomes to comprehend the sheer power of SRv6 uSID
- In the end the goal is to underscore the immense value proposition with SRv6 uSID in the Data Center landscape

SILOED Networking ⇔ Complexity Tax

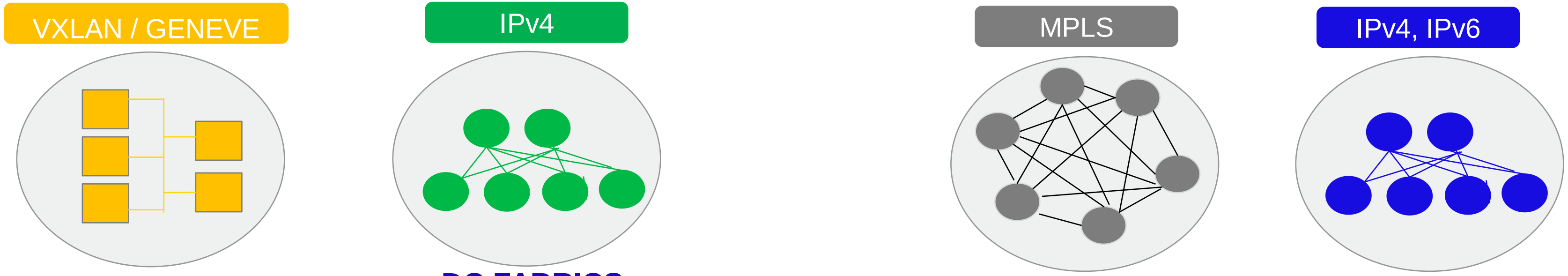
Each siloed network Domain has its own Hardware, Software, SDN Stack, Operations & Automation Ecosystem

IPv4 Transport

- OVERLAY/SEGMENTATION ↑
- PRIVATE IP SCALE ↓
- TRAFFIC ENGINEERING ↓

MPLS Transport

- NOT WIDELY ADOPTED IN DATA CENTER ↓
- LABEL SUMMARIZATION ↓
- TRAFFIC ENGINEERING ↓



VXLAN / GENEVE — DC OVERLAYS

IPv4 — DC FABRICS (UNDERLAY)

MPLS — METRO WAN FABRICS

IPv4, IPv6 — EDGE PEERING FABRICS

SRv6 uSID ⇔ Simplicity, Functionality, Ultra Scale

A Common End-to-End Forwardin hitecture Enables Common HW, SW, SDN, Ops ⇔ **Massive Scale**

SRv6 uSID

OVERLAY/SEGMENTATION

COMPATIBILITY WITH BROWN FIELD FABRICS ⬆

IPv6 UNDERLAY SCALE ⬆

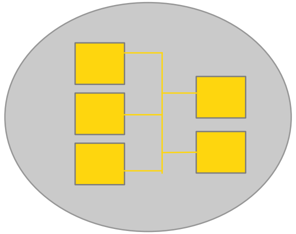
VICE INSERTION / CHAINING ⬆

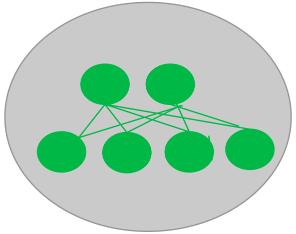
TRAFFIC ENGINEERING ⬆

RDWARE, SOFTWARE, SDN, Ops ⬆

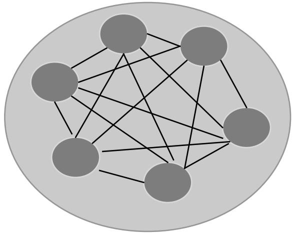
E2E Simplicity with SRv6 uSID✓

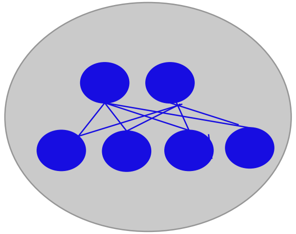
SRv6 uSID ⇔ Translates a Massive Scale


DC OVERLAYS

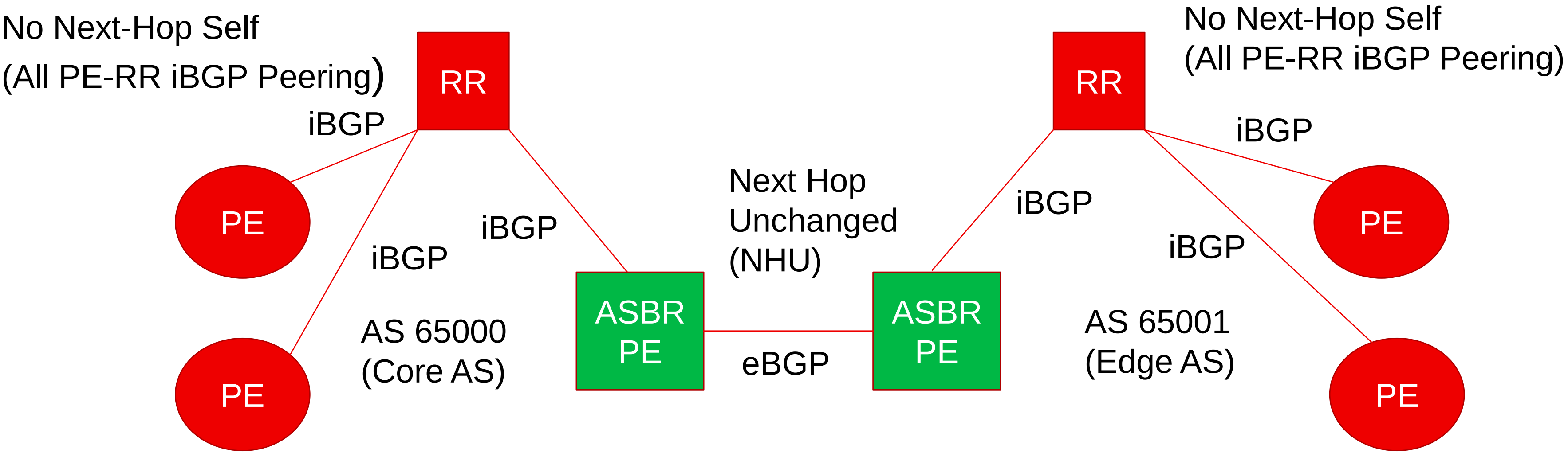

UNDERLAY
DC FABRICS

"End-to-End Inter-Domain Routing via SRv6 uSID"


METRO
WAN FABRICS


EDGE PEERING
FABRICS

INTER DOMAIN SRv6 uSID



No Next-Hop Self
(All PE-RR iBGP Peering)

AS 65000
(Core AS)

Next Hop
Unchanged
(NHU)

AS 65001
(Edge AS)

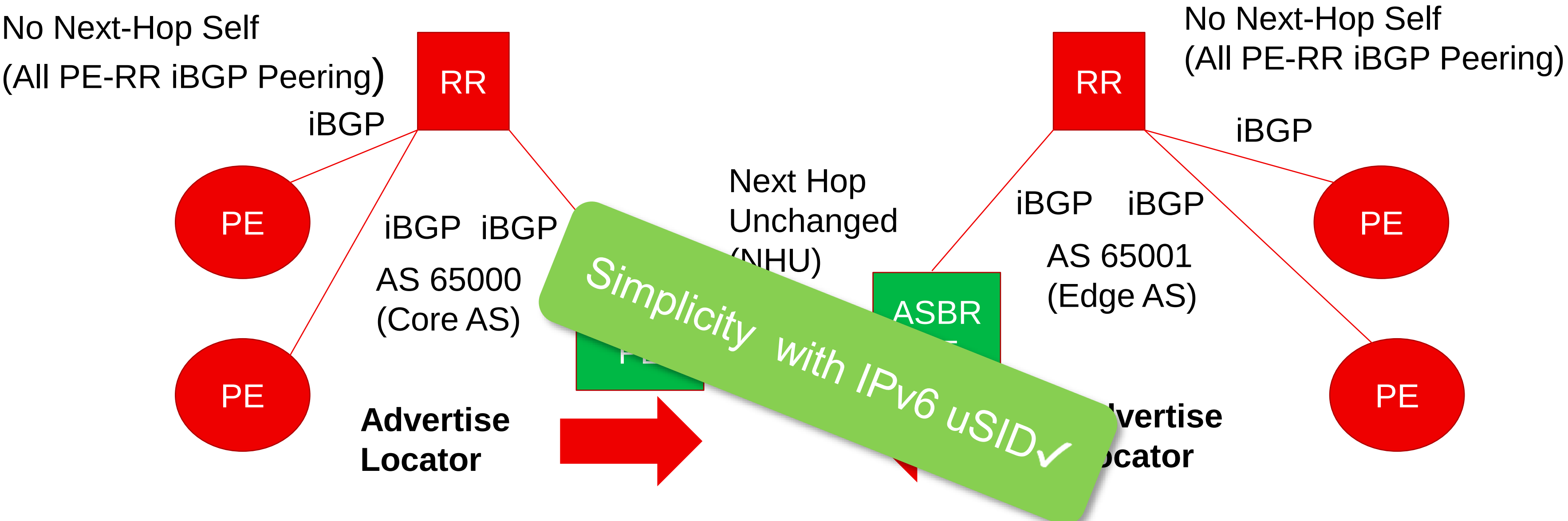
No Next-Hop Self
(All PE-RR iBGP Peering)

Rule for Inter Domain Peering

- iBGP -No Next-Hop Self
- eBGP –Next Hop Unchanged

(Requirement to Preserve L2 VPN & L3 VPN Service SID across INTER-AS Boundary)

INTER DOMAIN SRv6 uSID ROUTING SIMPLICITY



Rule for Inter Domain Peering

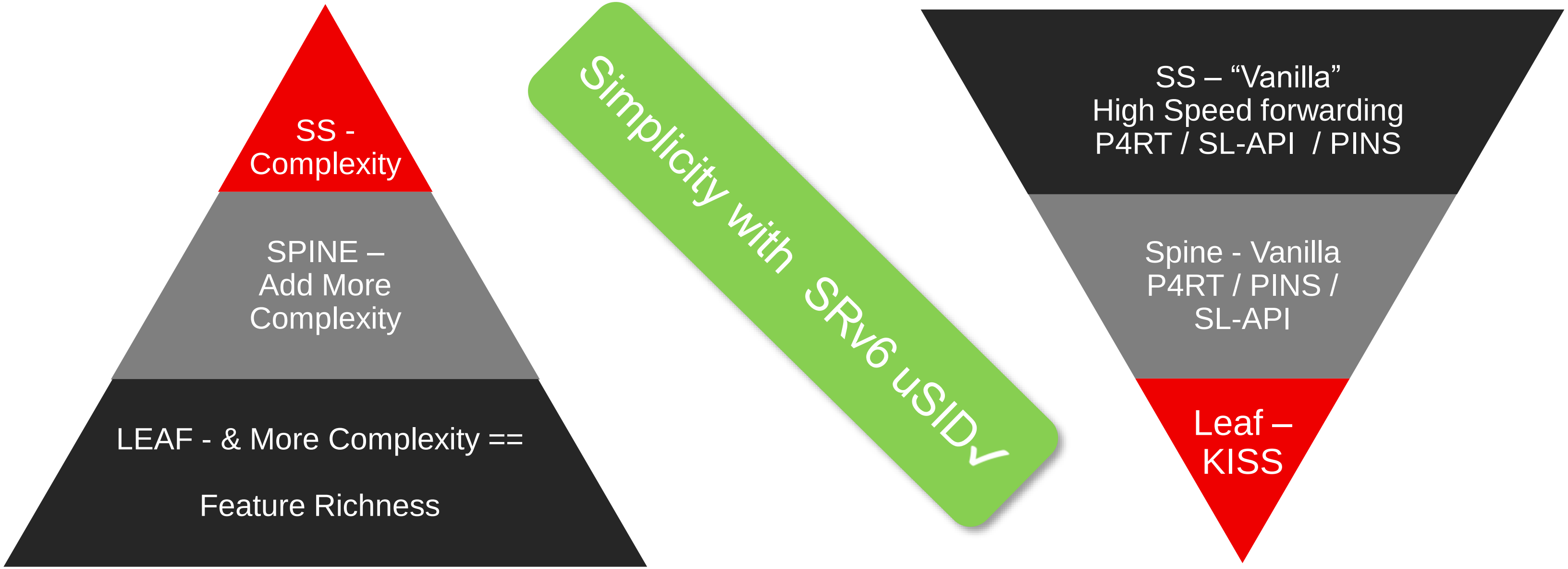
- Locator Reachability (That's it!!)

This allows host endpoints to provide static steering capabilities without PCE across any SR Algo cross domain

SRv6 uSID Design ⇔ “Top ⇔ Down” & “Bottom ⇔ Up” Approach

Top ⇔ Down

Bottom ⇔ UP



SS - Complexity

SPINE – Add More Complexity

LEAF - & More Complexity ==

Feature Richness

Simplicity with SRv6 uSID✓

SS – “Vanilla”
High Speed forwarding
P4RT / SL-API / PINS

Spine - Vanilla
P4RT / PINS / SL-API

Leaf – KISS

Traditional mindset has been for feature richness & complexity across the Data Center Fabric

SRv6 KISS ⇔ “Keep it Simple & Strategic” approach ⇔ Focus on High speed forwarding plane packet pushing throughput

Real World Use-Cases

#1 IPv6 Host Based Networking

#2 Dual Plane MPLS / IPv6 Core Migration

#3 SRv6 uSID End-To-End

#1 IPv6 Host Based Networking

- **Traffic Engineering and Carrier Grade** features are not a requirement in the Data Center.
- Operators can use white box switches or disaggregated hardware and software with Vanilla IPv6 Only DC fabric blindly passing the IPv6 uSID packets. Massive bandwidth **where Multi Petabits** of fiber can be thrown at the DC fabric, with the focus on **High Bandwidth** packet pushing with **Ultra simplified fabric**.
- **Steering** is initiated from the Data Center host attachment using IGP shortest path leaving the entire fabric 100% vanilla IPv6.

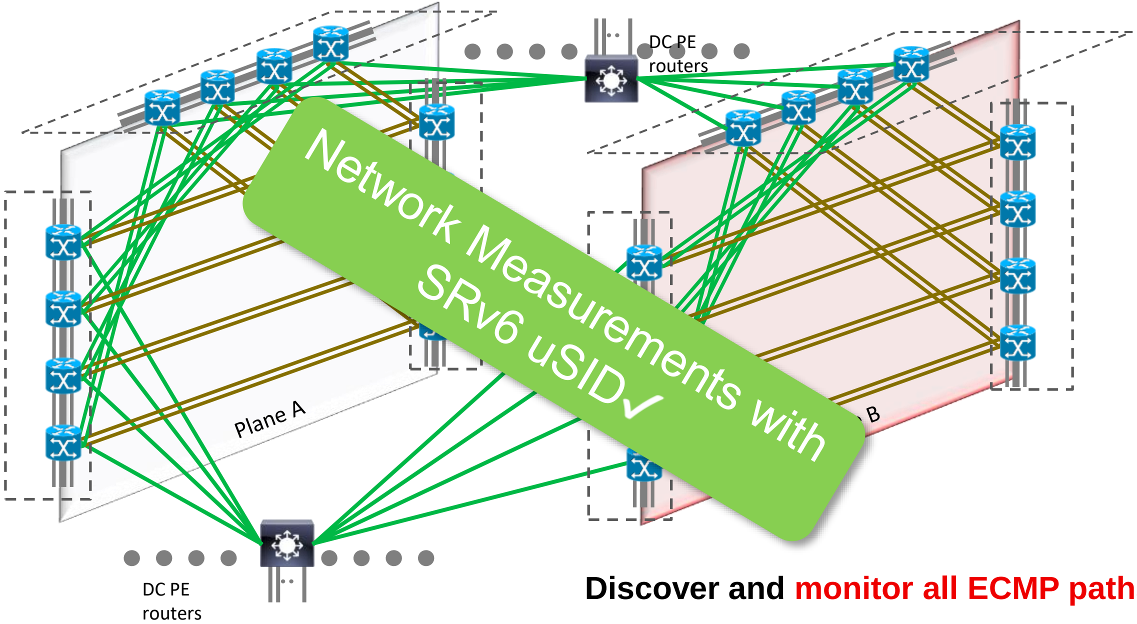
#2 Dual Plane MPLS / IPv6 Core Migration

- **Traffic Engineering & Carrier Grade** features are a requirement ONLY in the Data Center.
- **Traffic Engineering capabilities** in the Data Center, and the intermediate domains follow IGP shortest path **blindly forwarding** the IPv6 uSID packets. **Massive scale & resiliency** with full carrier grade features in the Data Center.
- **Steering** is initiated from the DC host attachment and follows IGP shortest path along the intermediate domains to the egress DC or Domain.

#3 SRv6 uSID End-To-End

- **Traffic Engineering** and all the Carrier Grade features are a requirement.
- **Full feature richness.**
- **Steering** is initiated from the Data Center host attachment or could be any switch within the DC fabric for any and all flow types.

Integrated Performance Monitoring (IPM)



Discover and monitor all ECMP paths

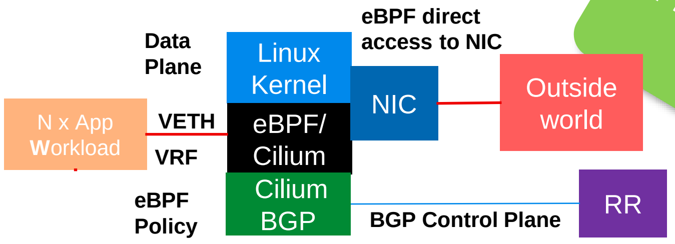
Provide Enough PPS to measure all ECMP paths

Report accurately across paths

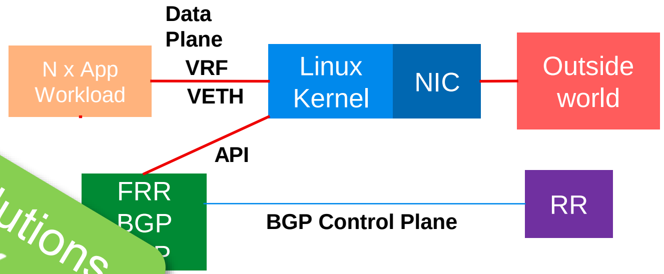
Host Networking Stacks

Host Networking Solutions with SRv6 uSID✓

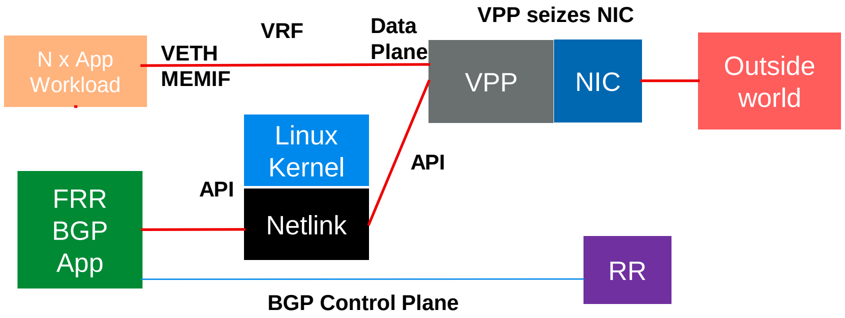
Option #1 eBPF/Cilium



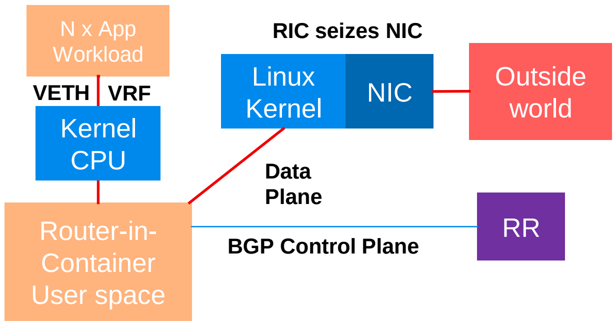
Option #2 Linux Kernel



Option #3 FD.io VPP



Option #4 Router-in-Container (RIC)



Simplified Host Networking

Lightweight Host Routing:

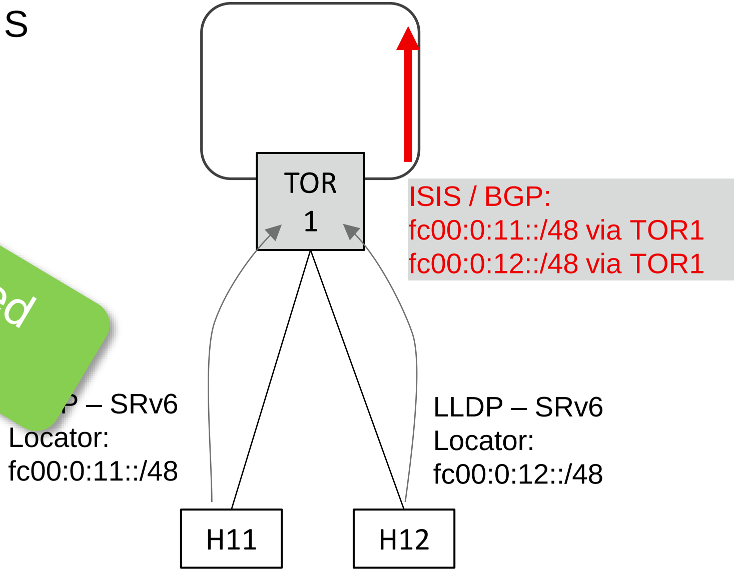
- Linux include its SRv6 Locator (IP Prefix) within the LLDP advertisements
- TOR (IOS XR/SONiC) redistri the prefix into BGP/ISIS

Host Networking Solved with SRv6 uSID✓

Simpler solution:

- Provides reachability (routing) up to the h
- Provides visibility into the container (workload IP
- Provides liveness detection (built-into LLDP)

Lightweight: No need to run BGP stack on the host



Demo time!

All Demo’s @ YouTube Channel SRv6 uSID DC:

https://github.com/segmentrouting/srv6-labs

Directory:”3-srv6-dc-case-studies”

https://youtube.com/@SRv6_uSID_DC

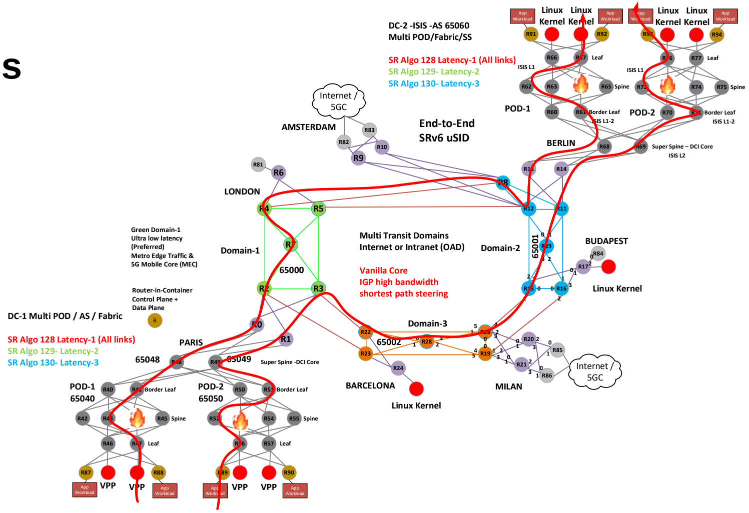
Use-case 1: SRv6 uSID BGP-Only DC & Single AS DC

Use-case 2: SRv6 uSID w/ Multi POD Fabrics

Use-case 3: SRv6 uSID w/ Multi POD/Domain Fabrics

❑ Host-to-Host multi-pod across the metro

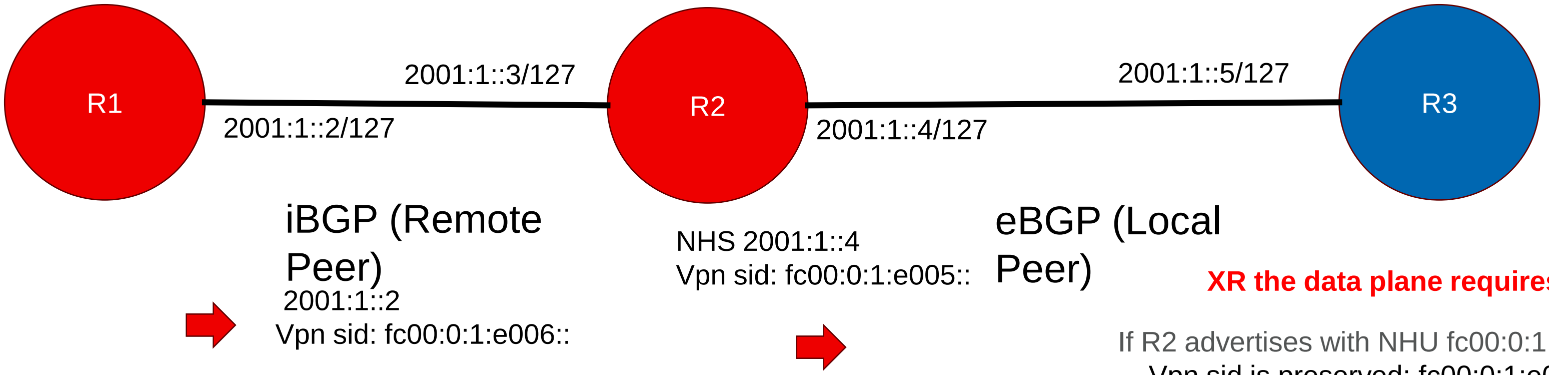
❑ Policy programmed from Linux Kernel & VPP host



verizon

INTER DOMAIN SRv6 uSID

eBGP direct peering (NHU) – (Remote PE)



ASN 1
Loc: fc00:0:1::
Loopback: fc00:0:1::1

ASN 1
Loc: fc00:0:2::
Loopback: fc00:0:2::1

ASN 2
Loc: fc00:0:3::
Loopback: fc00:0:3::1

R1

2001:1::2/127

2001:1::3/127

R2

2001:1::4/127

2001:1::5/127

R3

iBGP (Remote Peer)

2001:1::2
Vpn sid: fc00:0:1:e006::

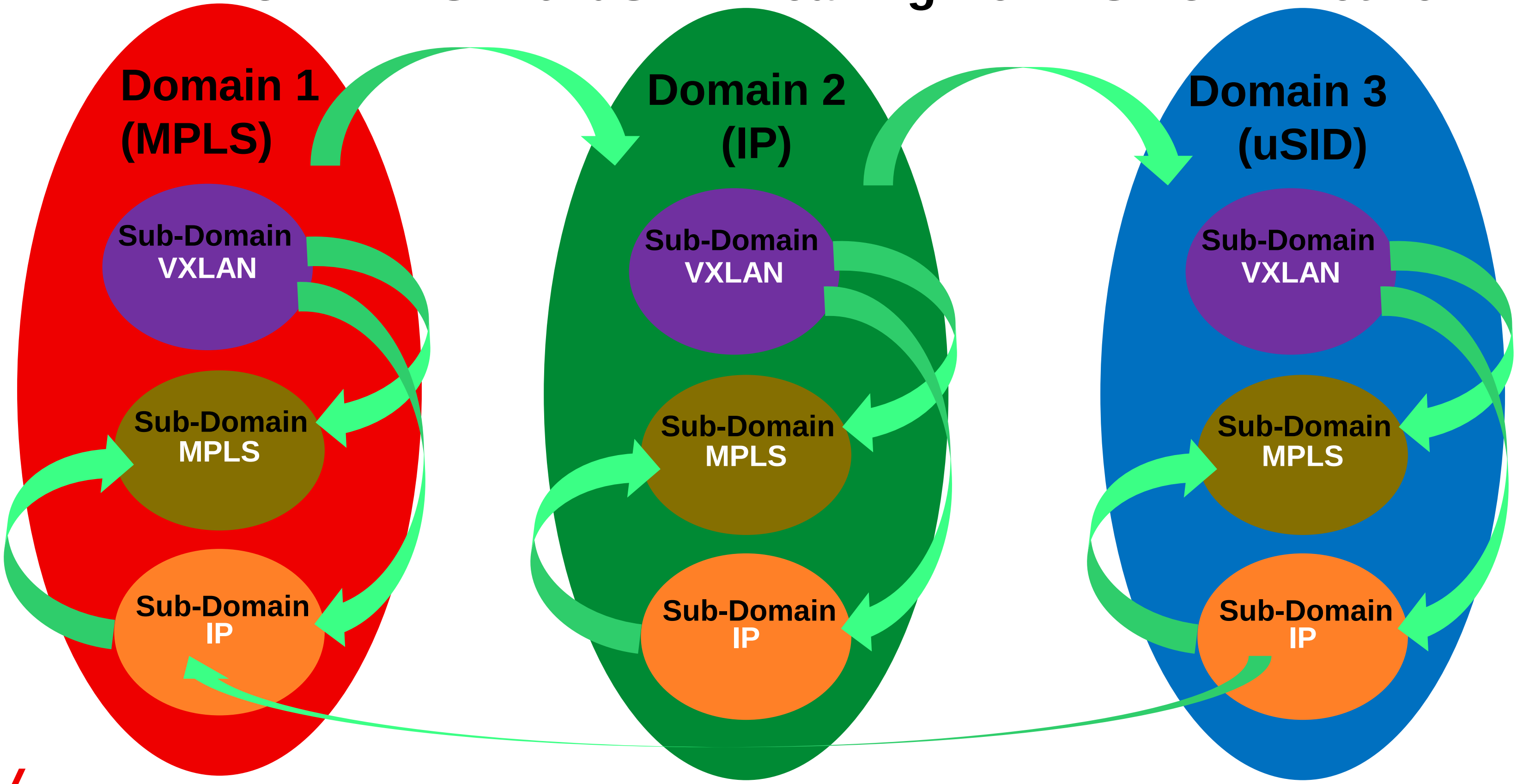
NHS 2001:1::4
Vpn sid: fc00:0:1:e005::

eBGP (Local Peer)

XR the data plane requires knob

If R2 advertises with NHU fc00:0:1::1
Vpn sid is preserved: fc00:0:1:e005::
but R3 sees it as *inaccessible*
BGP knob fixes the inaccessible
“IGNORE CONNECTED CHECK”
neighbor w.x.y.z
ignore-connected-check

INTER DOMAIN SRv6 uSID Breaking Down SILOED Networking



Domain 1 (MPLS)

Sub-Domain VXLAN

Sub-Domain MPLS

Sub-Domain IP

Domain 2 (IP)

Sub-Domain VXLAN

Sub-Domain MPLS

Sub-Domain IP

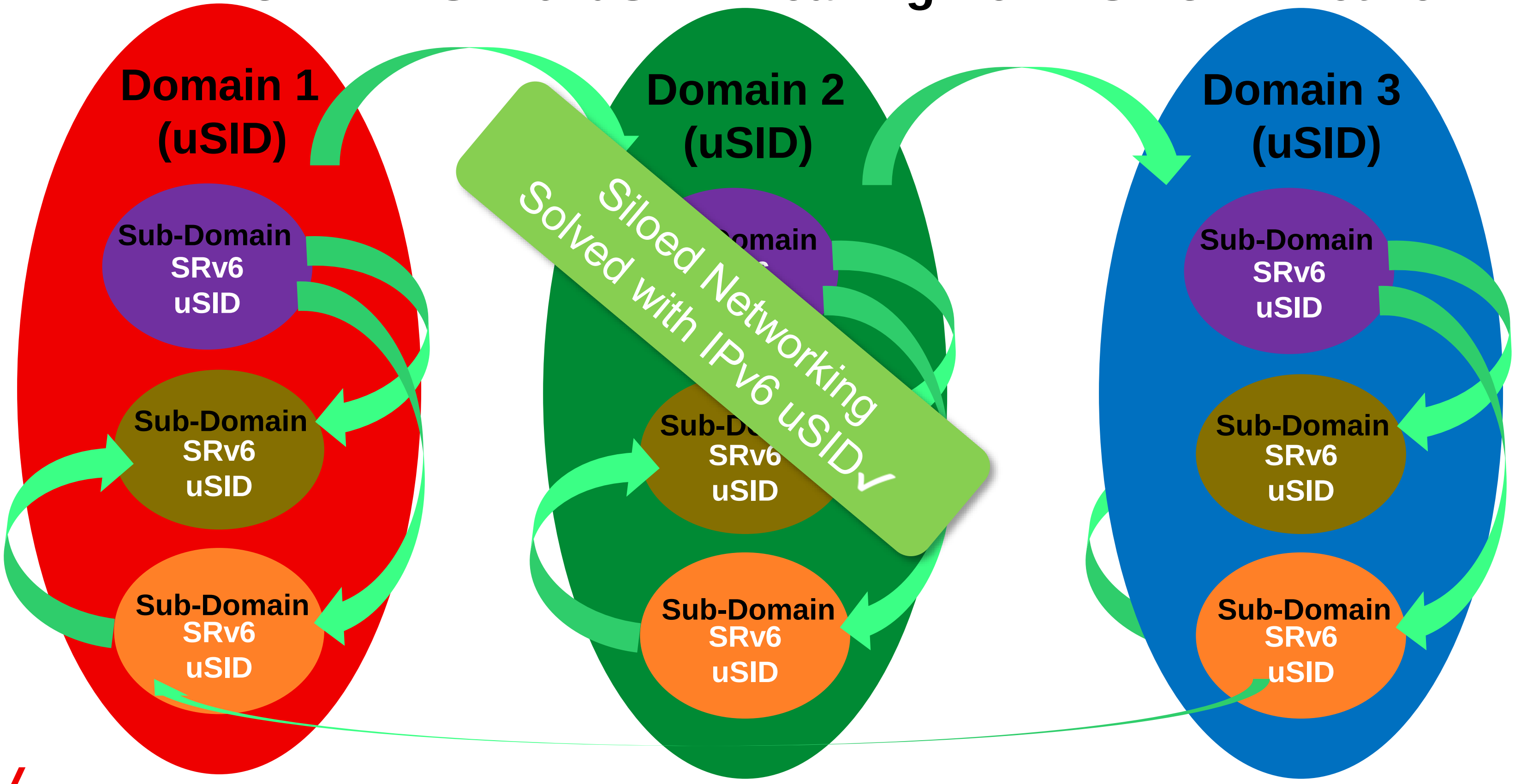
Domain 3 (uSID)

Sub-Domain VXLAN

Sub-Domain MPLS

Sub-Domain IP

INTER DOMAIN SRv6 uSID Breaking Down SILOED Networking



Siloed Networking
Solved with IPv6 uSID✓

Verizon proprietary. Unauthorized disclosure, reproduction or other use prohibited.

SRv6 uSID Host Based Networking Traffic Engineering

Options for Host Based Networking

- ❑ eBPF/Cilium (Cilium BGP control plane) CNI   or Standalone (Option-1)
- ❑ Native Linux Kernel (FRR BGP control plane) –Host Routing with Native Kernel (Option-2)
- ❑ FD.io VPP (FRR BGP control plane) -Host Routing via VPP (Option-3)
- ❑ Router-in-container (Control Plane & Data Plane) (xRD, SONiC, Nokia, Juniper cRPD) CNF (Option-4)

- ❑ Options are listed in order of desirability by operators
- ❑ Next few slides we will go into each option in detail

Option #1 eBPF/Cilium & SRv6 uSID TE Capabilities

- ❑ CNI Connects to global table ⬄ linkage of host fabric to DC fabric
- ❑ CNI TE is Manual/Static today ⬄ Future Roadmap for Dynamic
- ❑ CNI Provides VPN overlay - Workload Container, VM, CNF, VNF

Details

- ❑ Data plane programming
- ❑ Cilium used for BGP control plane advertisement
- ❑ Cilium is one of the most popular CNI’s to date and eBPF with its origins in the Linux kernel with its rich policy features & programmability provides seamless integration to compute nodes making it a powerful win-win for developers
- ❑ eBPF bypasses Linux kernel for policy processing & has direct access to NIC

Option #2 Native Linux Kernel & SRv6 uSID TE Capabilities

- ❑ Connects to global table ⬄ linkage of host fabric to DC fabric
- ❑ TE Capabilities via Linux “iproute 2” support for SRv6 uSID
- ❑ Host VPN overlay - Workload Container, VM, CNF, VNF for VRF attached workload

Details

- ❑ Data plane programming
- ❑ FRR BGP for control plane advertisement
- ❑ FRR can program the control plane & via Linux Kernel API call program the data plane FIB entries
- ❑ Alternatively, FRR can program the control plane with hook back to Linux Kernel to program the data plane FIB entries

Option #3 FD.IO VPP (Vector Packet Processing) & SRv6 uSID TE Capabilities

- ❑ VPP Connects to global table ⇔ linkage of host fabric to DC fabric
- ❑ VPP Provides Traffic Engineering capabilities via SRv6 uSID
- ❑ VPP Provides VPN overlay - Workload Container, VM, CNF, VNF

Details

- ❑ Data plane programming
- ❑ FRR BGP for Control Plane Advertisement
- ❑ VPP Seizes Control of the Linux Hosts NIC
- ❑ Requires Netlink or other method to program VPP FIB
- ❑ VPP (Vector Packet Processing) is a high performance network stack that can support high bandwidth & CPU intensive applications

Option #4 Router-in-Container (RIC) & SRv6 uSID TE Capabilities

- ❑ CNF Connects to global table ⬄ linkage of host fabric to DC fabric
- ❑ CNF Provides Traffic Engineering capabilities via SRv6 uSID
- ❑ CNF Provides VPN overlay - Workload Container, VM, CNF, VNF

Details

Router-in-container options (xRD, SONiC, Nokia, Juniper cRPD)

- ❑ Control plane & Data plane programming
- ❑ Requires Linux bridge stitching between Linux Kernel & CNF
- ❑ Container runs in user space so is beneficial for cases where only certain application requires traffic engineering capabilities
- ❑ User space applications can connect to separate virtual interfaces on router-in-container without any theoretical interface limit thus can support n-app workloads