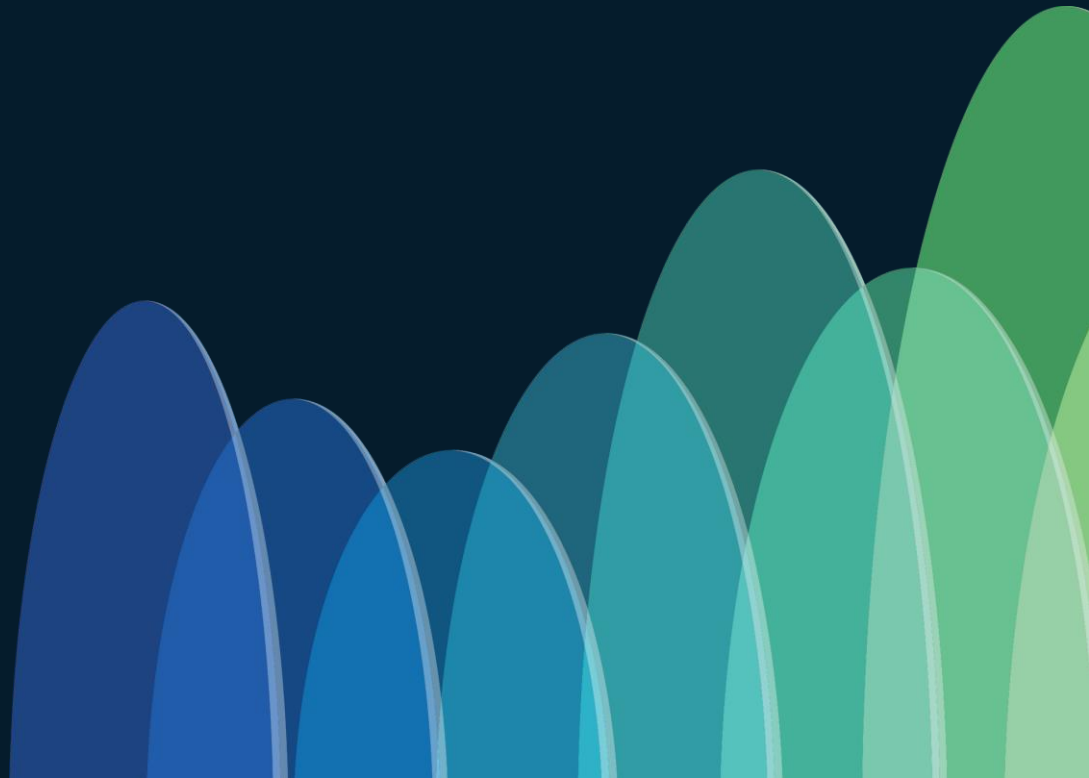
Advanced Innovations in SRv6 uSID and IP Measurements

Pablo Camarillo
Sr. Tech Lead, pcamaril@cisco.com
BRKSPG-3198

Agenda

- SRv6 Industry Update
  - NEBIUS SRv6 design
    DC Frontend to Peering
  - SRv6 to the Host
    Application awareness with eBPF
- IP Measurements
- Path Tracing

SRv6 Industry Update



cisco Live!

SRv6 uSID

- Build Anything
  - Any combination of underlay, overlay, service chaining, security…
  - VPN, Slicing, Traffic Engineering, Green Routing, FRR, NFV
- Any Domain
  - Access, Metro, Core, DC, Host, Cloud
  - End-to-End Stateless Policy
  - No protocol conversion or gateways at domain boundaries
- Seamless Deployment in Brownfield
- Built day-1 for Automation
- Standardized, Rich Eco-system, Rich Open Source (eBPF/SONiC)

Outperform MPLS/VxLAN

Outperform MPLS - Daniel Voyer (Bell Canada)

- Native Optimum Slicing
  - SLID is encoded in Flow Label
- HW Linerate Push: 3 times better
  - J2 uSID linerate push: 30 uSIDs >> 10 MPLS Labels
- HW Counter and FIB consumption: 4 times better
  - uSID requires 4 times less counters and FIB entries than MPLS
- Routing scale: 20 times better
  - uSID supports summarization. MPLS requires host routes.
- Lookup efficiency: 2 to 3 times better
  - uSID can process 2 to 3 SIDs in a single lookup (LPM nature)
- Load-balancing: optimum and deterministic
  - uSID provides HW friendly entropy (fixed offset, shallow)

Bell

Bell SRv6 uSID Deployment
Paris 2022

Presentation & recording

Outperforms VxLAN – Gyan Mishra (Verizon)

- Seamless Host support for Network Programming
  - 6 uSID's in outer DA: RFC2460 IPinIP with opaque DA
- TE in the DC
  - elephant flows exist, asymmetric fabrics exist, TE is needed
- TE in the Metro/Core from the host
  - An SRv6 uSID DC allows for the application to control the network program in the metro/core without complex DPI and protocol conversion at the DC boundary,
- uSID DC provides lower MTU overhead (~5%)
  - Lower MTU overhead means lower DC cost
- Vendor, Merchant and SONIC/SAI maturity
  - uSID support across DC vendor (Cisco), Merchant (Cisco, Broadcom, Marvell), Sonic/Sai (Alibaba deployment)

verizon

SRv6 uSID DC Use-Case
Paris 2023

Presentation & recording

Rich SRv6 uSID Ecosystem

Network Equipment Manufacturers

CISCO, HUAWEI, NOKIA, ciena, JUNIPER NETWORKS, ZTE, ARRCUS NETWORK DIFFERENT, ERICSSON, H3C

Merchant Silicon

MARVELL, BROADCOM

Edge/Core: Q200, P100
ToR/Spine: G100, G200

DNX – Jericho
XGS – Tomahawk

Open-Source Applications

Pyroute2, IPTABLES SERA, GoBGP, exaBGP, eBPF, cilium, WIRESHARK, FRRouting, SNORT, BIRD

SRv6 uSID

Open-Source Networking Stacks

SONiC, .io, P4, OPEN Compute Project, OvS Open vSwitch, VyOS

Smart NIC / DPU

intel, NVIDIA

Partners

kaloom, TREND MICRO, ENEA Qosmos Division, NoviFlow, ixia A Keysight Business, spirent Promise. Assured., 6WIND, KEYSIGHT TECHNOLOGIES

cisco Live!

SRv6 is Proposed Standard

Architecture
- SR Architecture – RFC 8402
- SRTE Policy Architecture – RFC 9256

Data Plane
- SRv6 Network Programming – RFC 8986
- IPv6 SR header – RFC 8754

Control Plane
- SRv6 BGP Services – RFC 9252
- SRv6 ISIS – RFC 9352
- SR Flex-Algo – RFC 9350

Operation & Management
- SRv6 OAM – RFC 9259
- Performance Management – RFC 5357

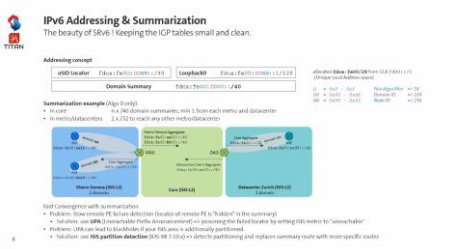
Strong Commitment and Leadership

Editor of 96% IETF RFCs
Co-author of 100% IETF RFCs

Over 85.000 uSID routers deployed

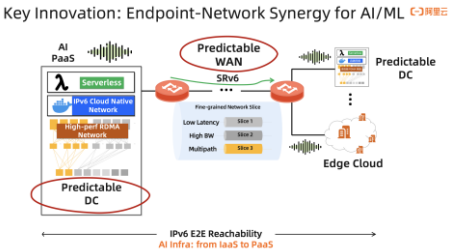
tpg TELECOM · zain · XL axiata · SYMA · Telefónica · T-Mobile · iliad · SoftBank · cesnet · free · SCHWARZ ITKG · LINE · Indosat ooredoo

swisscom



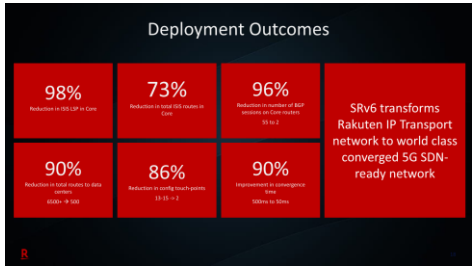
Presentation & recording

Alibaba Cloud



Presentation & recording

Rakuten



Presentation & recording

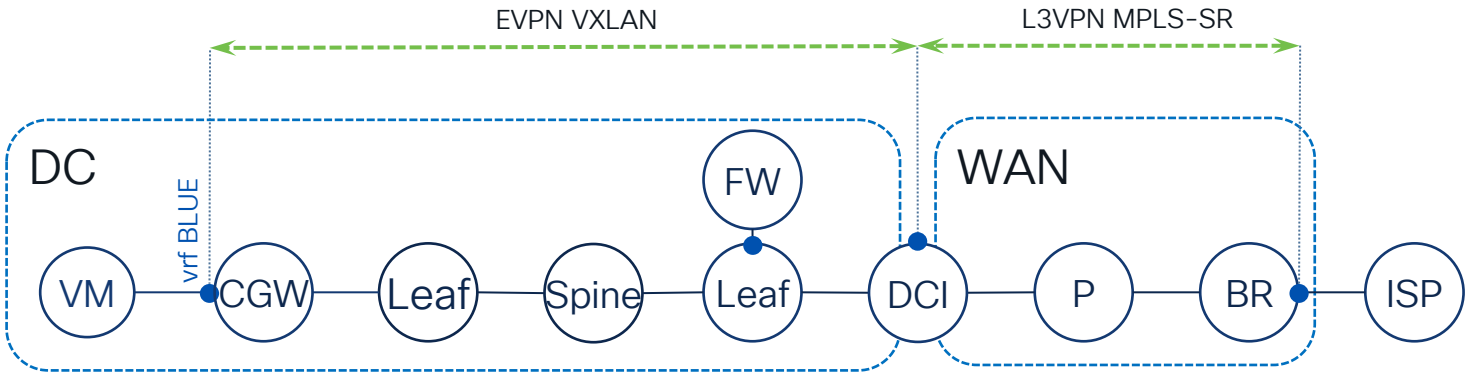
NEBIUS

Case Study:
SRv6 uSID DC Frontend to Peering

Alexey Gorovoy
Network Engineer @ NEBIUS

cisco Live!

Current architecture of the Frontend network



- IPv6 only infrastructure in the Data Center
- Multivendor DC and WAN networks approach
- CGW (Cloud Gateway) and FW are NFV's running on hosts. Nebius develops them
- VXLAN based overlay between CGW and DCI
- DCI does "stitching" between EVPN VXLAN and L3VPN MPLS-SR

Current architecture – evaluation

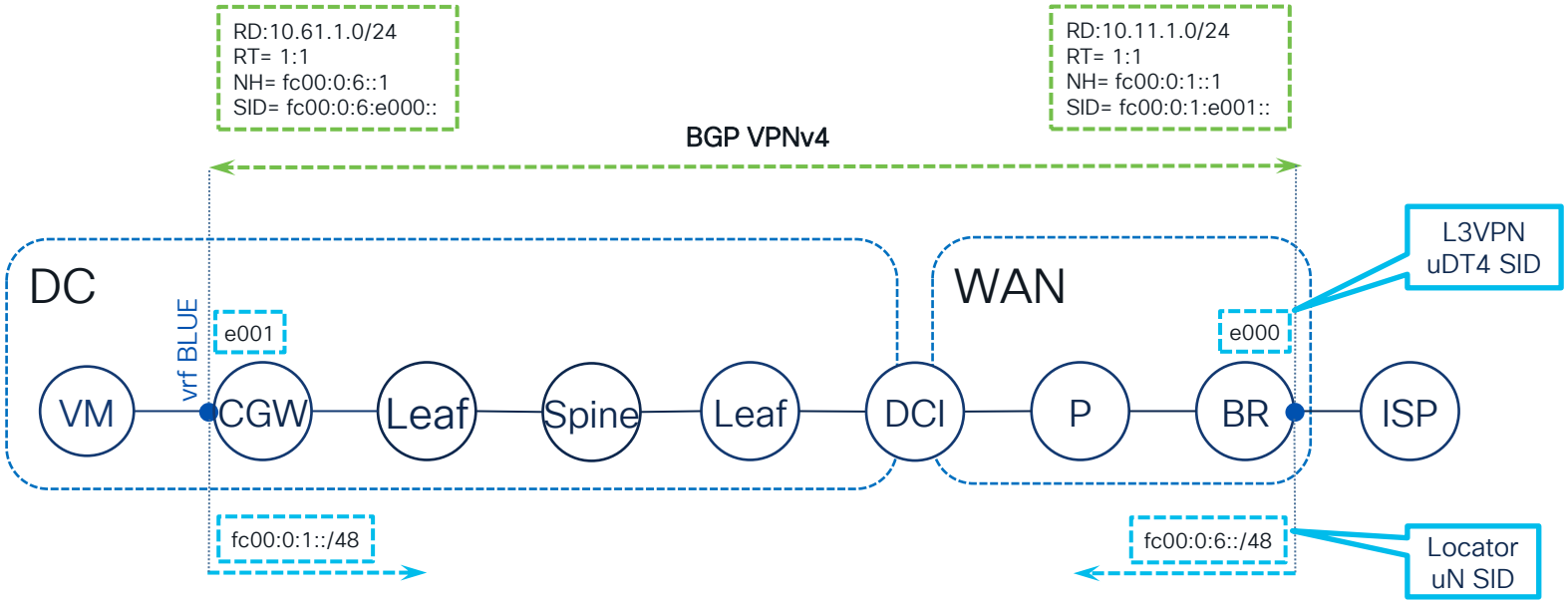
- Pros:
  - VXLAN EVPN has wide industry support and adoption
  - DC fabric is overlay agnostic thus scalable, simple and efficient
  - EVPN provides rich variety of network services
  - MPLS-SR is a mature technology with good multivendor interoperability for VPN and TE applications
- Cons:
  - No traffic engineering capabilities inside the Data Center
  - Service chaining with VXLAN requires specific routing design (PBR, Default GW, VRF/VLAN hand-off, etc.)
  - Majority vendor implementations of VXLAN still require IPv4 loopbacks in the Underlay
  - MPLS-SR lacks native Data Center optimisations and not applicable in the DC domain
  - Requires ”stitching” gateway functionality at the DCI routers to interconnect WAN and DC domains

SRv6 addresses all of them!

Transition to SRv6

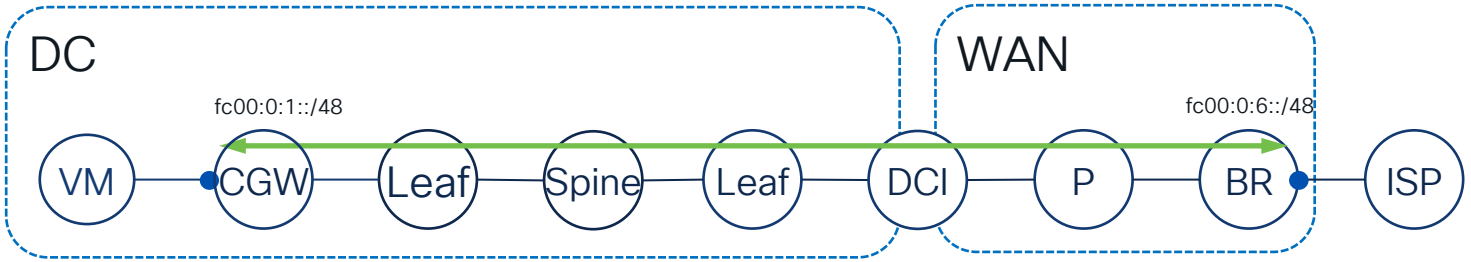
- Bridges both DC and WAN domains together in efficient and simple way
- Creates unified data plane based on IPv6 protocol only
- Allows to build end-to-end overlay service across DC and WAN without stitching fuctionality on the intermidiate devices
- Offers true traffic engineering capabilities initiated from the source of an application allowing efficient service chainings creation

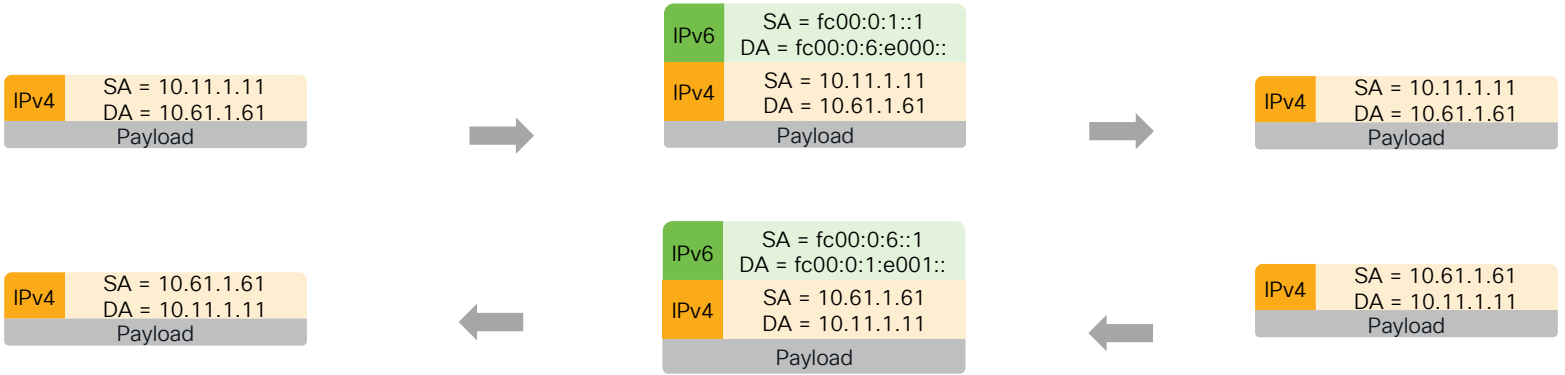
Overlay with SRv6 uSID



- IPv6 in DC and WAN
- SRv6 only required on CGW and BR
- CGW and BR act as SRv6 L3VPN PEs

Overlay with SRv6 uSID – Packet walk

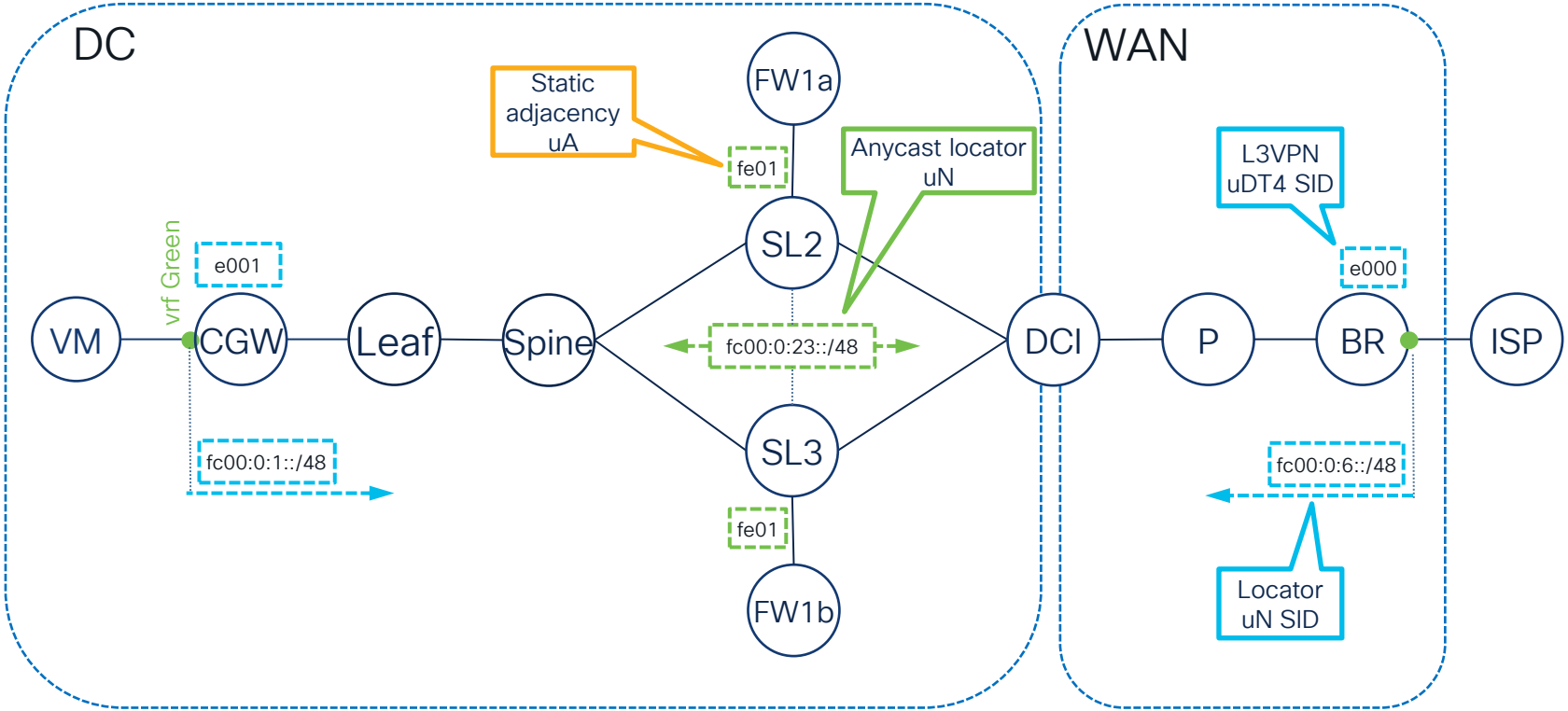




Services Chaining with SRv6 uSID (e.g., Firewall)

- Current design proposed solution:
  - FW is a cluster of sync'd nodes
  - Deployed behind dedicated physical nodes – Service Leaves
  - FW service inspects the inner packet, does not change the outer IP header
    - No encap/decap at SL's
    - SL's are SRv6 enabled routers
    - FW is a plain IPv6 forwarder
- Future goal:
  - FW is SRv6 enabled VNF, attached anywhere in the plain IPv6 forwarding network
    - Scaling FW service per any network segment, customer or application

Firewall Insertion

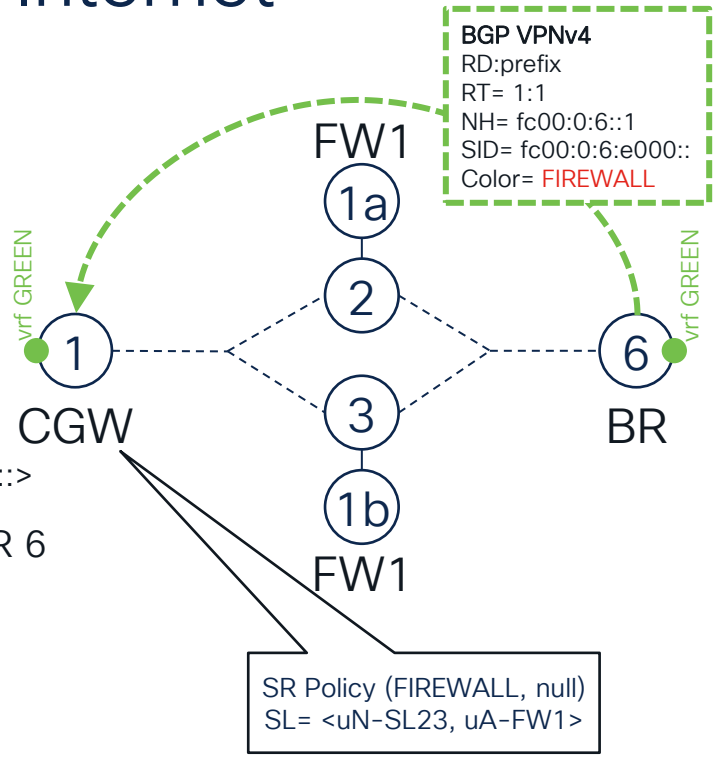


Firewall insertion – From VM to Internet

- BR advertises Internet routes in VRF GREEN with a color “FIREWALL”
  - Individual prefixes, aggregates, default route
- CGW uses BGP AS into a color-only SR Policy
- CGW steers into SR Policy (FIREWALL, null) with SID list <fc00:0:<uN-SL23>:<uA-FW>::>
  - E.g., CGW 1 steers to FW1a/b with SID list <fc00:0:23:fe01::>
  - E.g., CGW 33 may steer it to FW33a/b with SID list <fc00:0:ab:fe33::>
- CGW 1 sends the FIREWALL service packets destined for BR 6 with DA= fc00:0:23:fe01:6:e000::

SL uN + FW uA

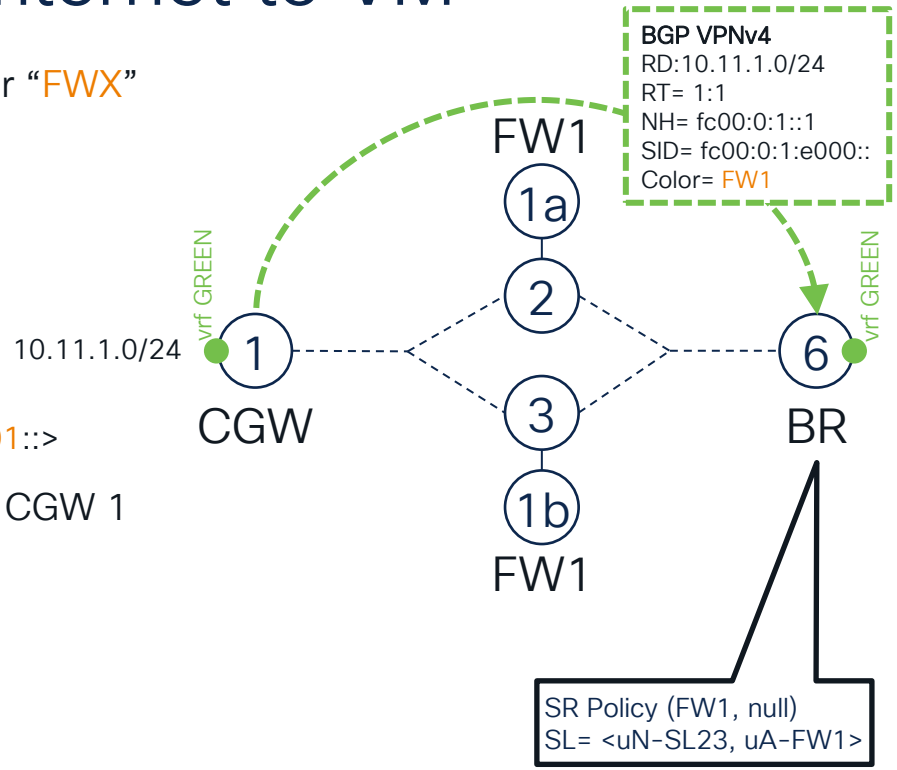
uN+uDT4 BR 6



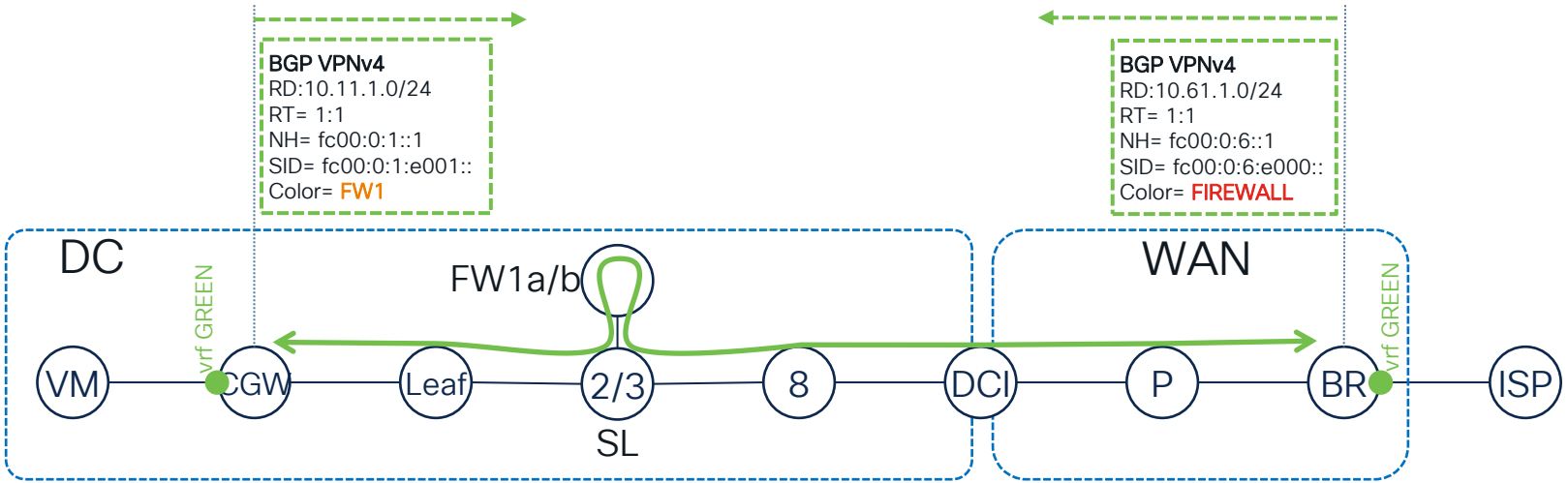
Firewall insertion – From Internet to VM

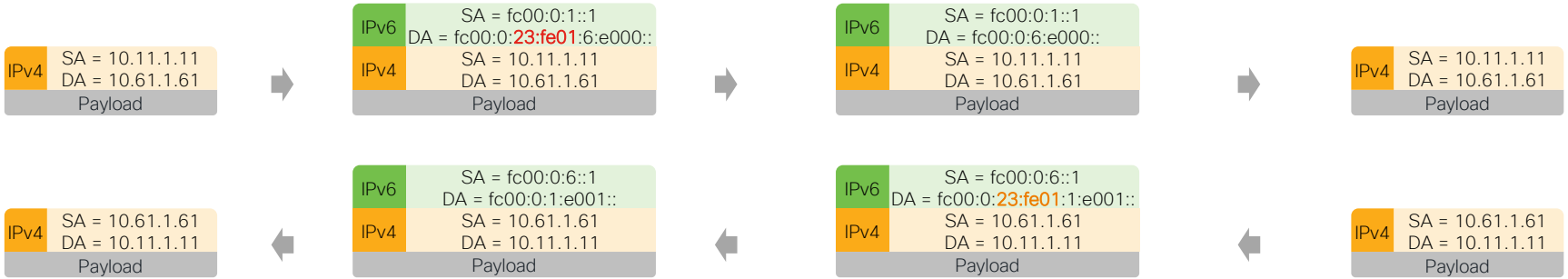
- CGW advertises its VRF GREEN routes with a color “FWX”
  - E.g., CGW 1 advertises 10.11.1.0/24 with color “FW1”
  - E.g., CGW 33 may advertise its prefixes with color “FW33”
- BR steers the service routes into the matching SR Policy (FWX, 0.0.0.0) with SID list <fc00:0:<uN-FWX>:<uA-FWX>::>
  - E.g., BR 6 steers to FW1a/b with SID list <fc00:0:23:fe01::>
- BR 6 sends the FW1 service packets destined for CGW 1 with DA= fc00:0:23:fe01:1:e001::

uN+uA SL23/FW1 uN+uDT* CGW 1



Firewall insertion – Packet walk





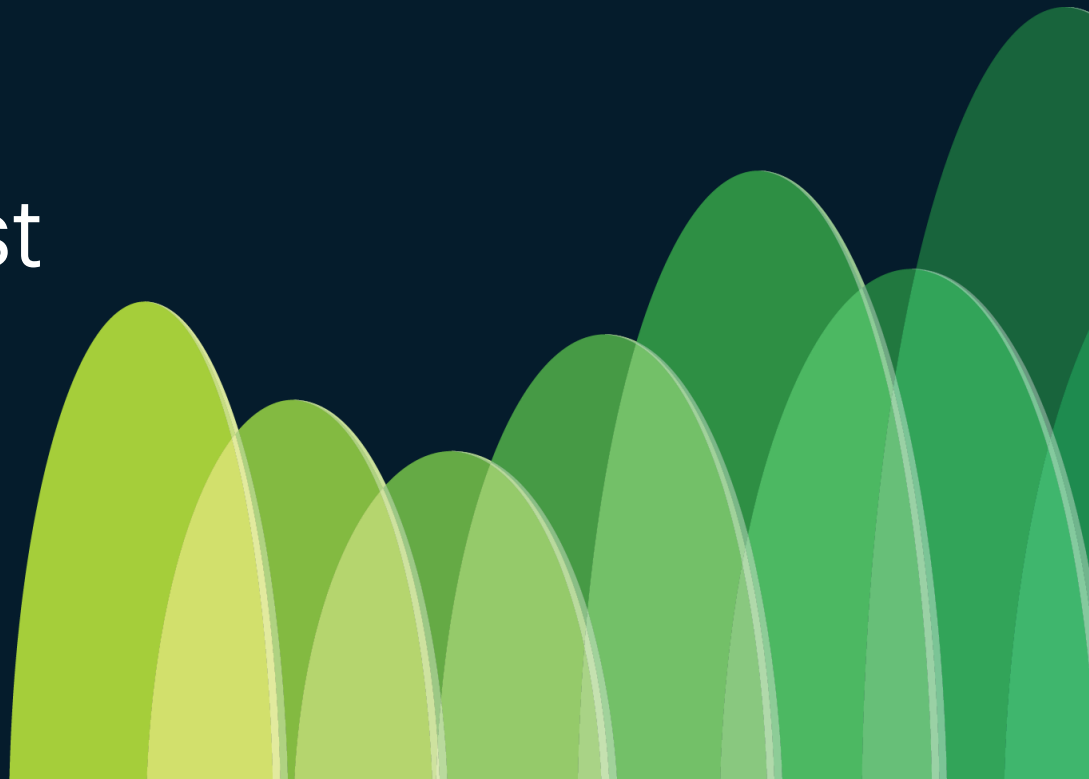
SRv6 Benefits: simplicity and unification

- Unified solution across all domains
- Operational and configuration simplicity
- Gaining scalability

Acknowledgements

- Team Nebius
  - Andrew Tikhonov, Senior Network Engineer, Nebius
  - Samvel Vartapetov, Senior Software Developer, Nebius
- Team Cisco
  - Clarence Filsfils, Fellow, Cisco
  - Kris Michielsen, Technical Leader Engineering, Cisco
  - Pablo Camarillo, Technical Leader Engineering, Cisco

SRv6 on the Host

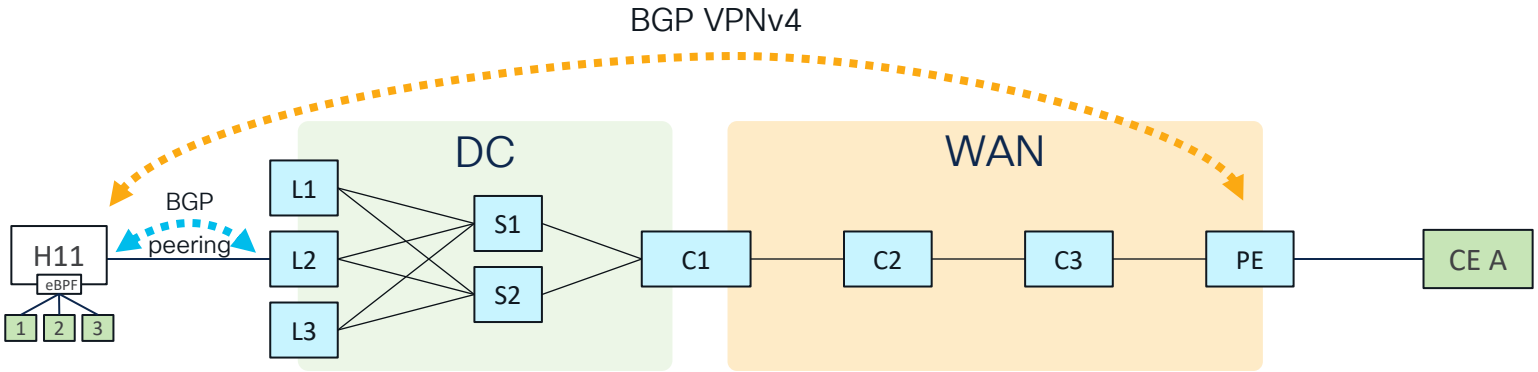


eBPF and Cilium

- What is eBPF?
  - Executes programmable logic in the Linux Kernel safely and performant with minimal footprint.
  - Doesn’t require compiling the Linux Kernel.
- What is Cilium?
  - Open-source networking and security for containers
  - Built on top of eBPF, enabling features like service mesh, network policies, and load balancing.
- Why SRv6 on the Host?
  - Source Routing empowers the source with control over traffic paths.

L3VPN with Cilium

Available Today!

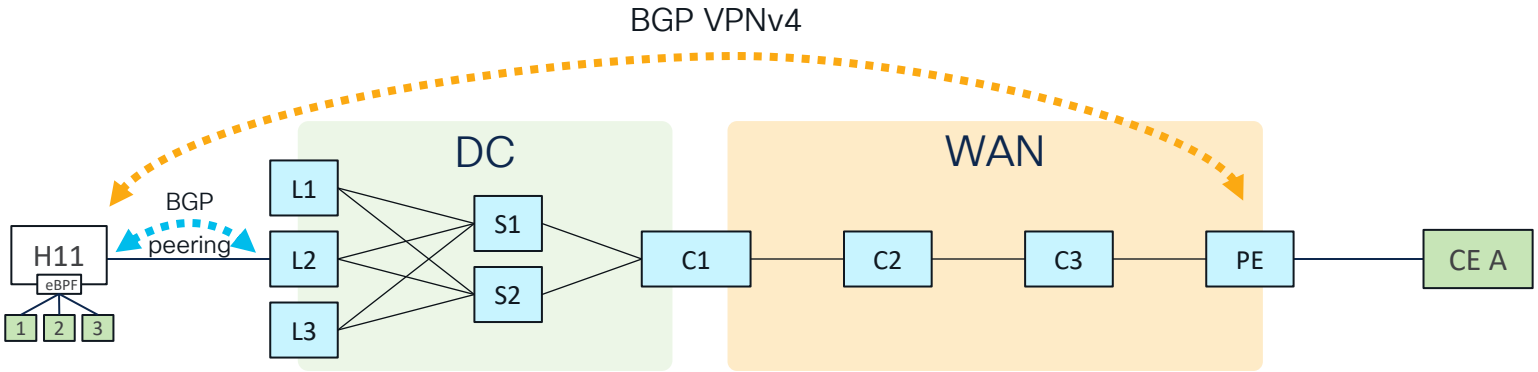


- Interoperable L3VPN support eBPF/Cilium <-> XR
- SRv6 Locator assigned per Kubernetes node
- uDT4 SID allocated for each VRF

Demo/TOI

L3VPN with Cilium

Available Today!

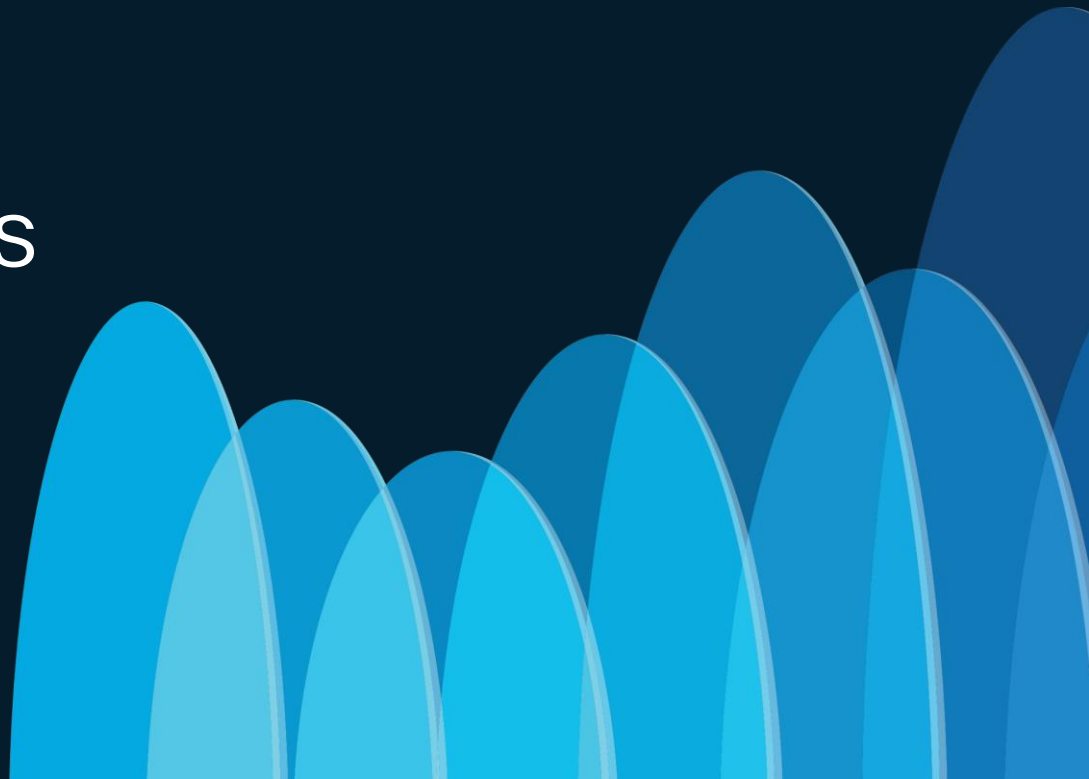



```
apiVersion: isovalent.com/v1alpha1
kind: IsovalentSRv6LocatorPool
metadata:
  name: pool0
  labels:
    export: "true"
spec:
  behaviorType: uSID
  prefix: fcbb:bb00:11::/48
  locatorLenBits: 48
  structure:
    locatorBlockLenBits: 32
    locatorNodeLenBits: 16
    functionLenBits: 16
    argumentLenBits: 0
```

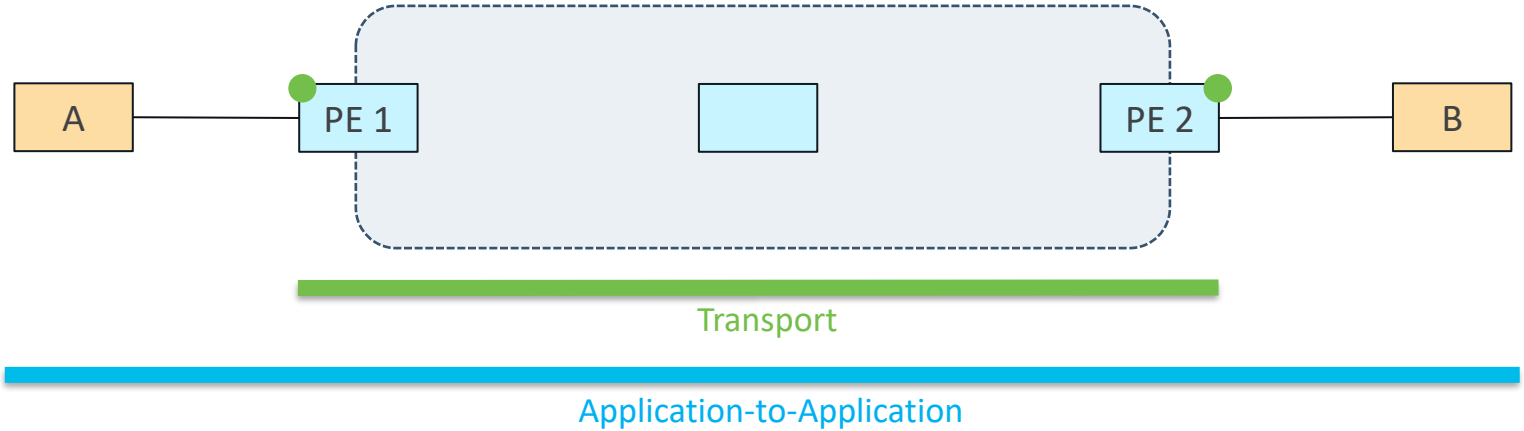
```
apiVersion: isovalent.com/v1alpha1
kind: IsovalentVRF
metadata:
  name: vrf1
spec:
  vrfID: 1
  importRouteTarget: "666:1"
  exportRouteTarget: "666:1"
  locatorPoolRef: pool0
  rules:
  - selectors:
    - endpointSelector:
        matchLabels:
          vrf: vrf1
    destinationCIDRs:
    - 10.10.1.0/24
    - 10.10.2.0/24
    - 10.10.3.0/24
```

Demo/TOI

IP Measurements

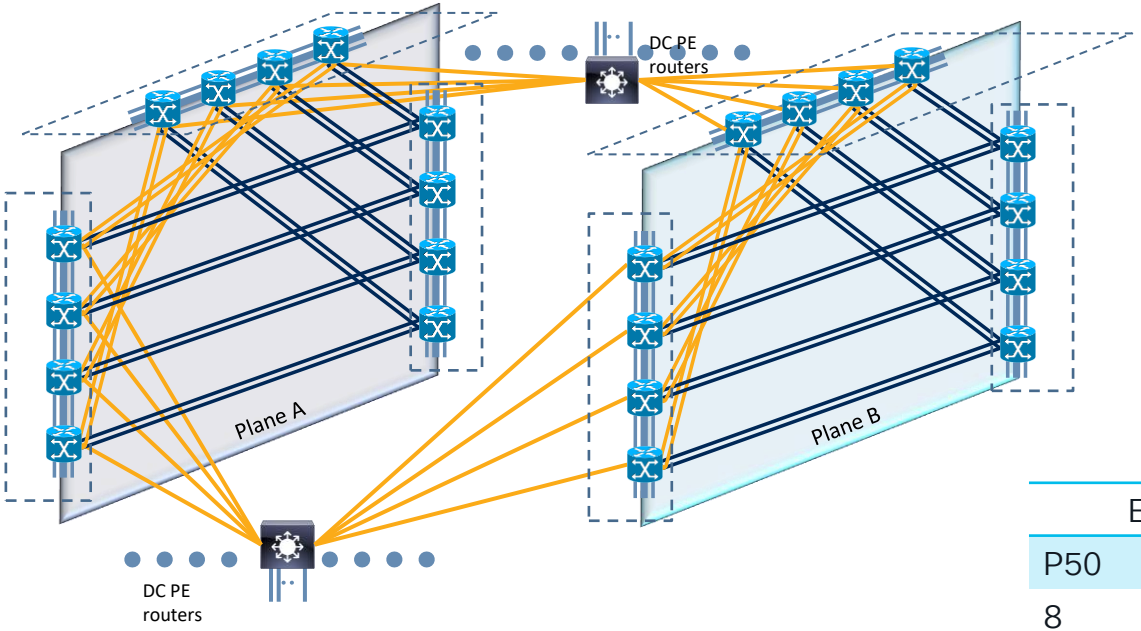


Transport SLA



- Focus: “Transport SLA”
- Out of Scope: Application to Application

The nature of IP is ECMP



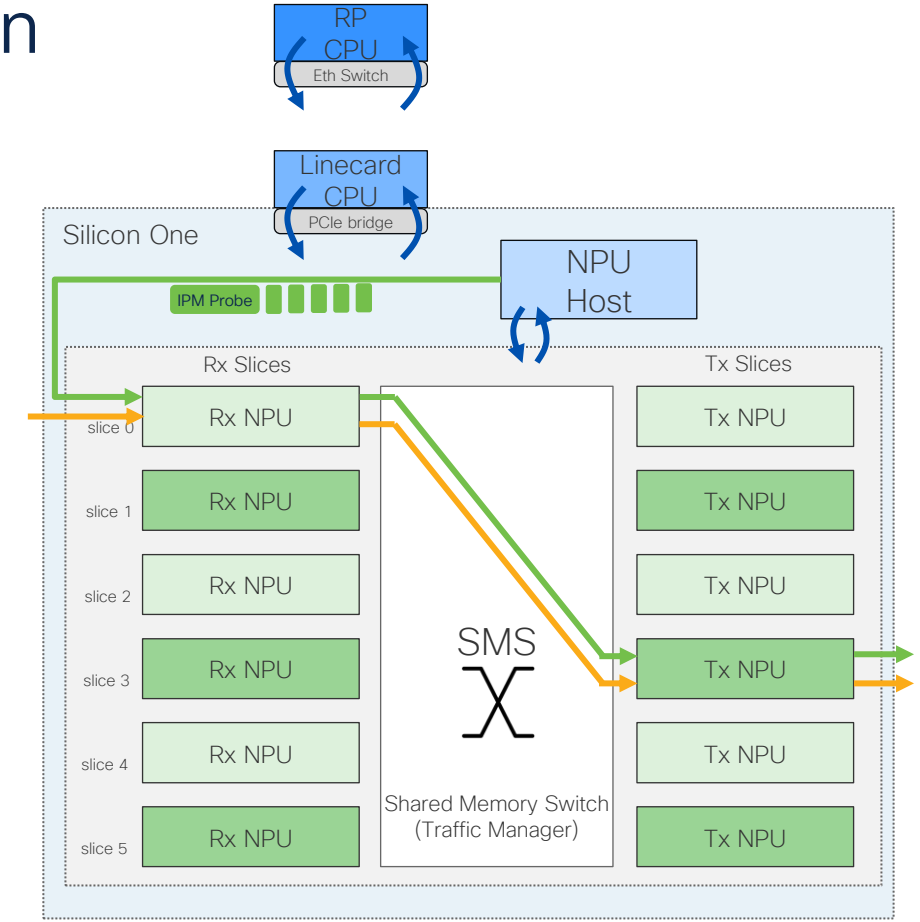
ECMP Paths		
P50	P75	P90
8	16	32

- Legacy solutions do not have the scale to measure all ECMP paths

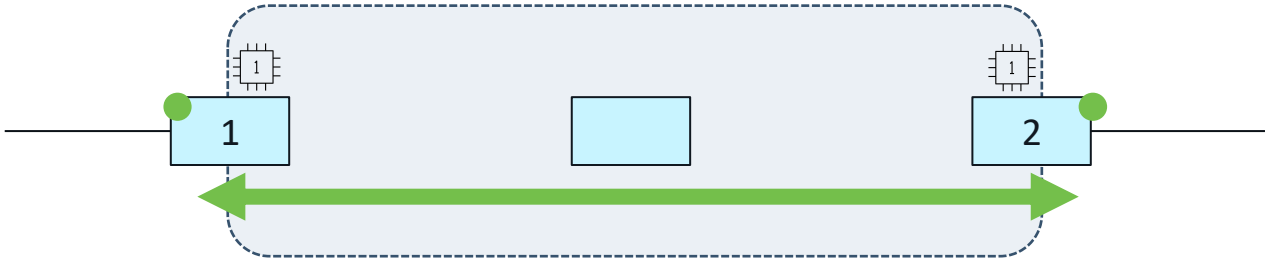
Silicon One HW integration

- Traditional probe generation and ingestion relies on CPU
- S1 provides another option: NPU Host
  - Large BW; Very flexible pipeline
- IPM probing built with NPU Host
  - Probe generation
  - Probe ingestion and aggregation
  - ….at 14MPPS
- In practical terms:
  - 1 measurement every ms
  - 500 edges
  - 16 ECMP paths

} 8M probes per sec (57% of S1)

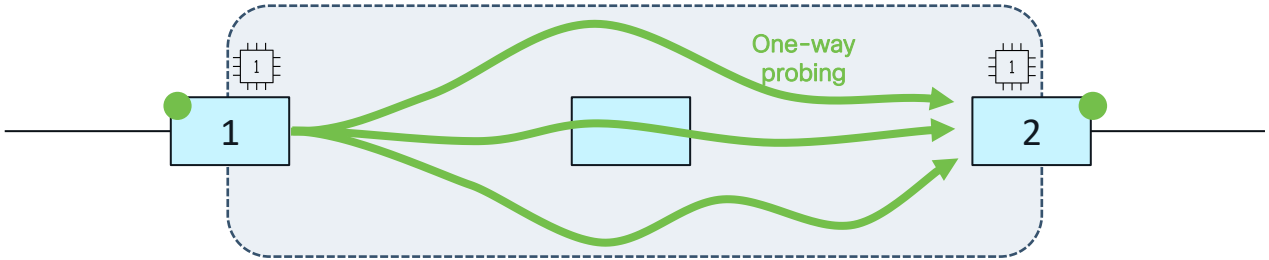


SLA Measurement from Any Edge to Any Edge, across ECMPs



- Active probing from any PE to any PE, via any ECMP path
- Continuous routing monitoring
- Analytics
  - Correlation of probe measurement and routing data

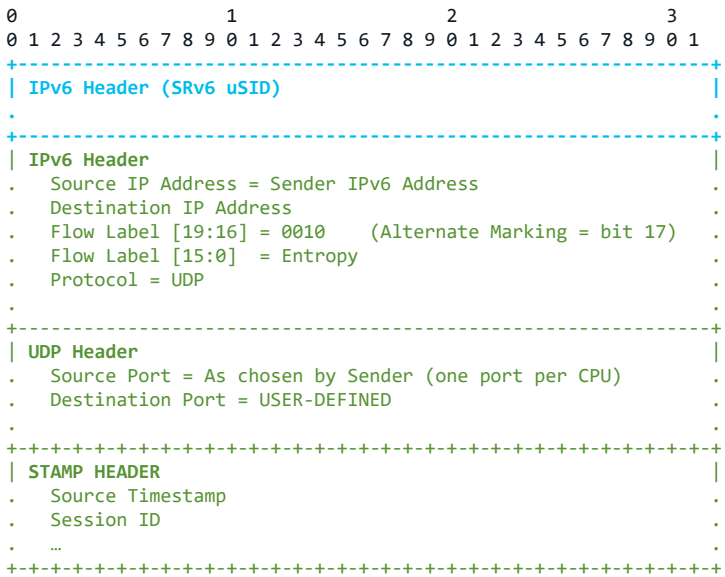
One-Way Measurement of all ECMP Paths



- Standard-based **one-way** probe (1, 2)
  - @1: High rate probe generation
  - @2: Probe receival/ingestion with 3Ls (Latency, Loss, Liveness)
  - One-Way Measurement eliminates exposure to the return path
- By default, each probe packet uses a random flow label
  - …generating 1000 probe packets per second
  - Exercise all ECMP paths in the fabric

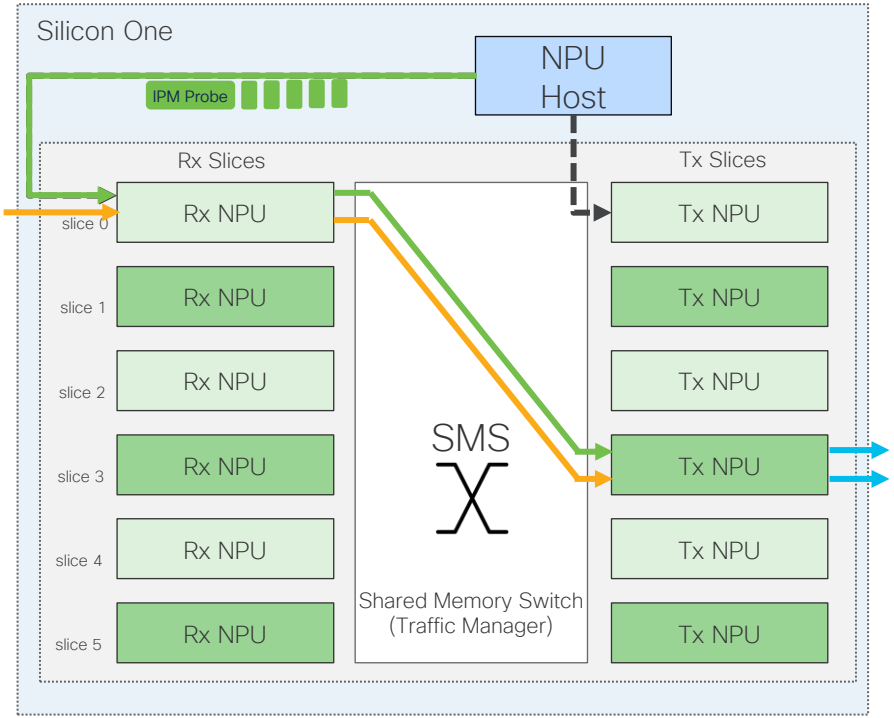
Standard Based Measurement

- STAMP – RFC8762/RFC8972
- Packet Format:
  - Outer Encapsulating header:
    - Any IP Encapsulation
  - STAMP measurement packet:
    - Alternate Marking bit as part of Flow Label



- STAMP measurement packet injected on VRF
- We monitor the shared transport **AND** the service forwarding path on PEs

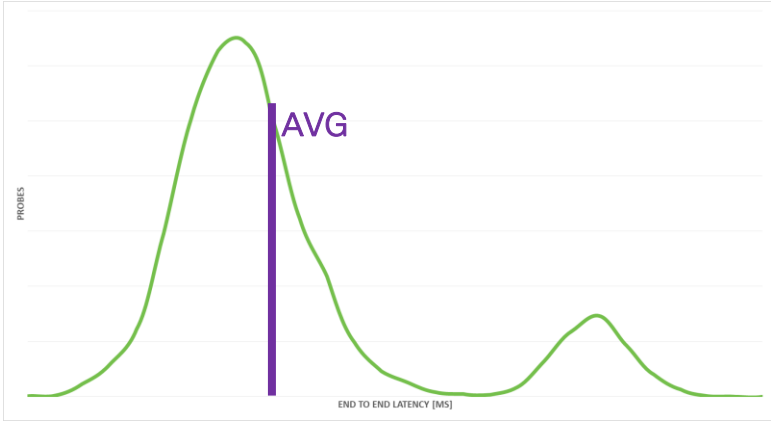
Standard Based Measurement




```
0                   1                   2                   3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-------------------------------------------------------------+
| IPv6 Header (SRv6 uSID)                                     |
.                                                             .
+-------------------------------------------------------------+
| IPv6 Header                                                 |
.   Source IP Address = Sender IPv6 Address                   .
.   Destination IP Address                                    .
.   Flow Label [19:16] = 0010    (Alternate Marking = bit 17) .
.   Flow Label [15:0]  = Entropy                              .
.   Protocol = UDP                                            .
.                                                             .
+-------------------------------------------------------------+
| UDP Header                                                  |
.   Source Port = As chosen by Sender (one port per CPU)      .
.   Destination Port = USER-DEFINED                           .
.                                                             .
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| STAMP HEADER                                                |
.   Source Timestamp                                          .
.   Session ID                                                .
.   …                                                         .
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

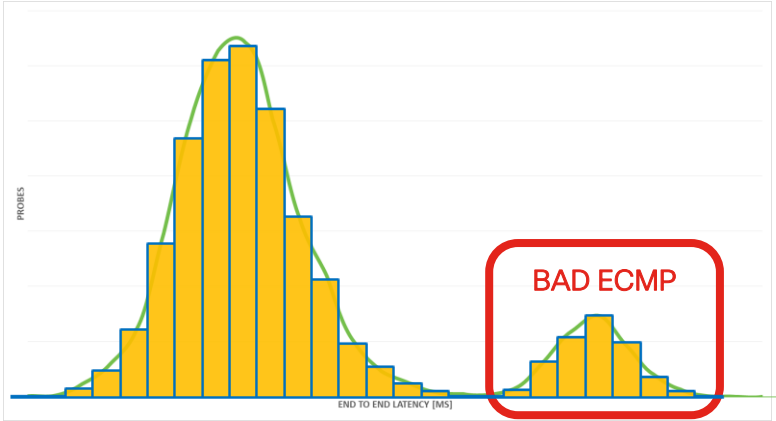
Probe aggregation: Accurate and Rich Metrics

- 1 bad path out of 8 ECMP
- 12.5% of the clients impacted
- Average hides the issue



Probe aggregation: Accurate and Rich Metrics

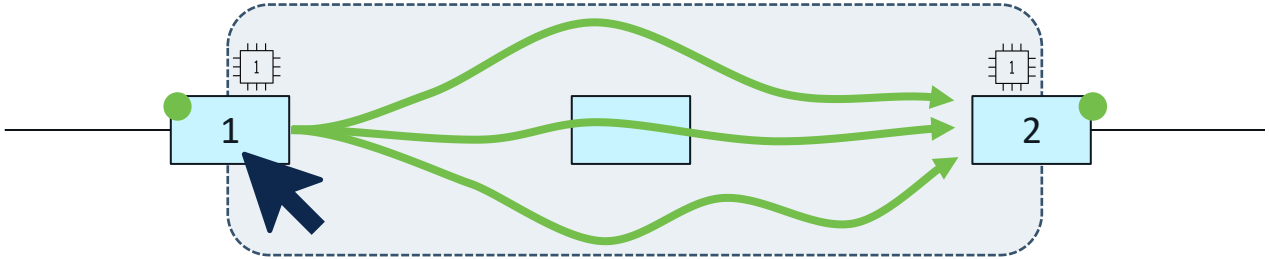
- 1 bad path out of 8 ECMP
- 12.5% of the clients impacted
- Average hides the issue
- IPM Histograms reports the experience of the whole population



- Each probe measures latency, loss and liveness
  - Latency histogram instead of min, avg, max
  - Absolute loss instead of loss approximations
  - Liveness detection (sub-2ms)
- Data aggregated every 1min and sent to analytics (cadence-driven telemetry)

XR CLI @1

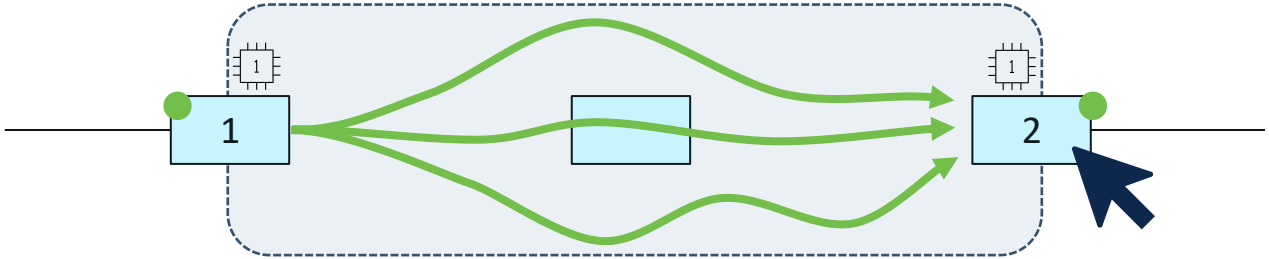
Available XR 25.4.1
(subject to change)




```
/* IPM 1 node */
config# performance-measurement
ipm-transmit-profile name profile-01
    probe-interval 999                  // Generate a packet every 999us from S1 NPU
    alternate-marking 60                // Every 60seconds change the Alternate Marking color
    dscp explicit 0                     // Use the DSCP value 0 (spray also available)
    flow-label spray                    // Spray across all ECMP Paths

endpoint ipv6 3fff::2 vrf purple        // 2 loopback address in VRF purple
    source-address ipv6 3fff::1
    ipm-measurement
        session-id 100
            ipm-transmit-profile profile-01
```

XR CLI @2

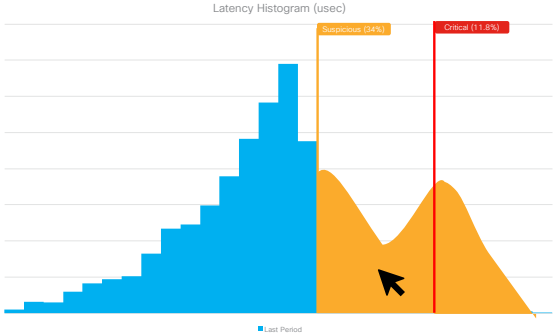
Available XR 25.4.1
(subject to change)

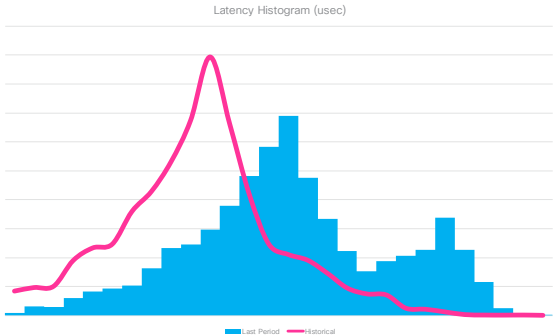


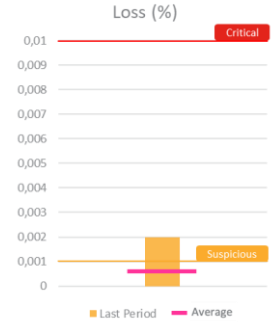

```
/* IPM 2 node */
config# performance-measurement
ipm-receive-profile name profile-01
    expected-probe-interval 999                // Expected rate from PE 1, 1 packet every 999us
    liveness-detection 10                      // If no probe is received in 10ms, declare liveness down
    latency
        histogram-collection 60                // Collect the histogram every 60s
        histogram templates Regional           // Use the regional template

endpoint ipv6 3fff::1 vrf purple               // PE 1 loopback address in VRF purple
ipm-measurement
        session-id 100                         <- same session-id on both directions of the endpoint
             ipm-receive-profile profile-01
```

OBA: On the Box Analytics

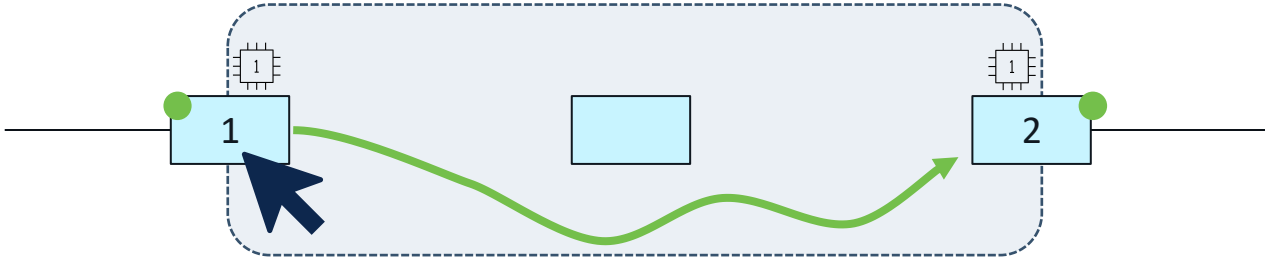
- Native XR analytics to process the measurement data, on the box.
- Complements the 3L Telemetry with Event-Driven Telemetry:
  - Latency & Loss Thresholding
  - Historical Trending (Exponential Moving Average; same day, hour)





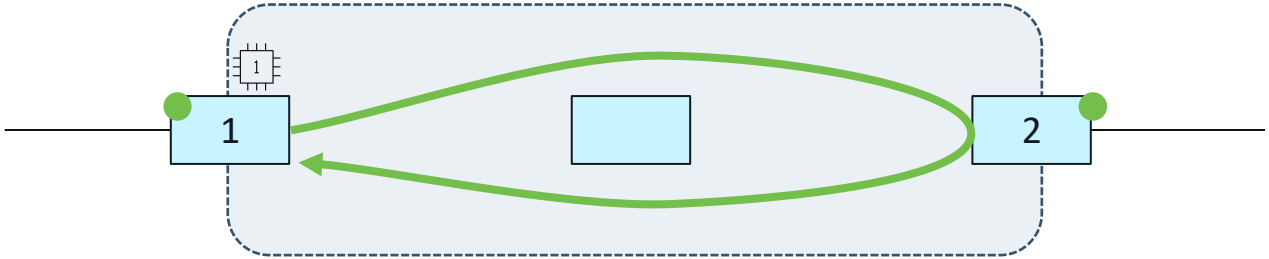


Measurement needs Routing Control




```
/* IPM PE 1 node */
config# performance-measurement
ipm-transmit-profile name profile-01
    probe-interval 999                  // Generate a packet every 999us from S1 NPU
    alternate-marking 60                // Every 60seconds change the Alternate Marking color
    dscp explicit 0                     // Use the DSCP value 0 (spray also available)
    flow-label spray                    // Spray across all ECMP Paths

endpoint ipv6 3fff::2 vrf purple        // PE 2 loopback address in VRF purple
    source-address ipv6 3fff::1
    ipm-measurement
        session-id 100
            ipm-transmit-profile profile-01
            segment-routing traffic-eng explicit segment-list name LIST1-SRV6
```

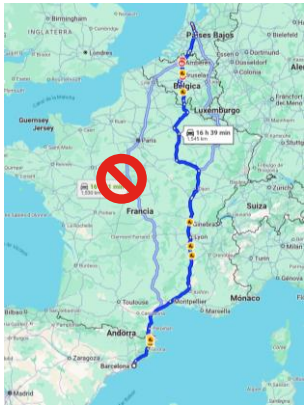
Day1 support for Brownfield



- One-way probe from 1, through 2, back to 1
  - uSID list with policy to get to 1, and then back to 2
- No IPM requirement at 2. Only need uSID support.

Continuous Correlation to Routing

Measured Latency
compared to **best** topology

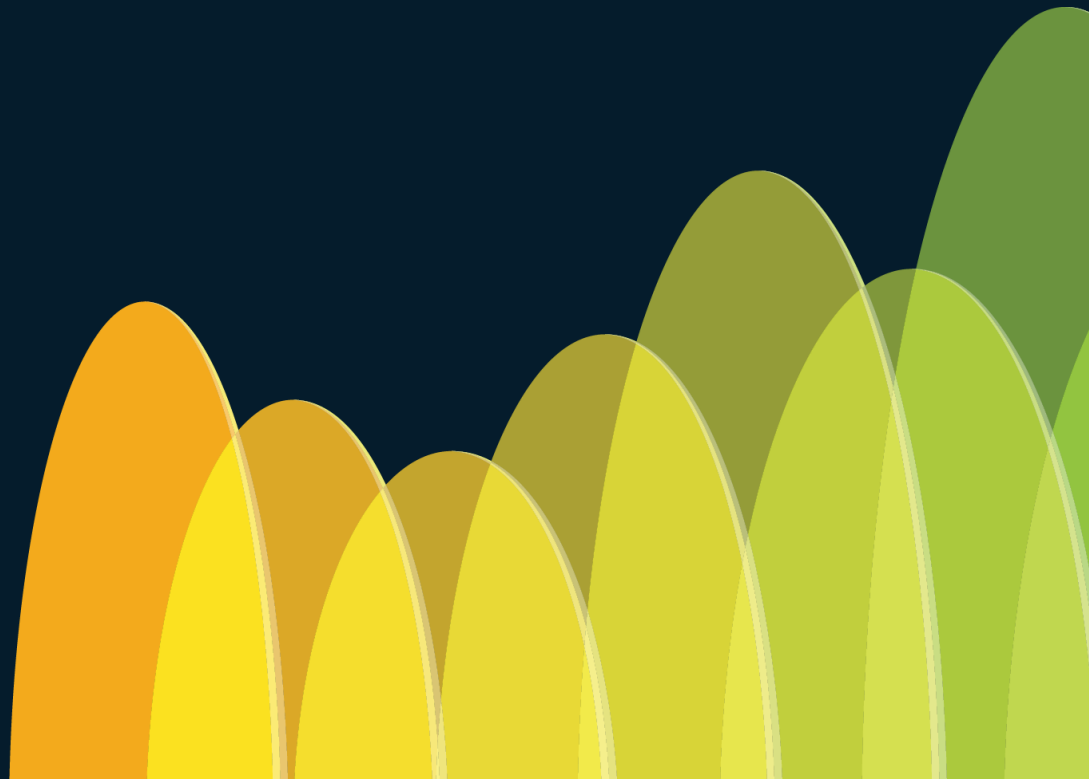


Measured Latency
compared to **current** topology

- Time-series of Measurements from any P to any Q along any ECMP path
- **Time-series of ECMP routed paths from any P to any Q**
- Provider Connectivity Assurance: data correlation and visualizations

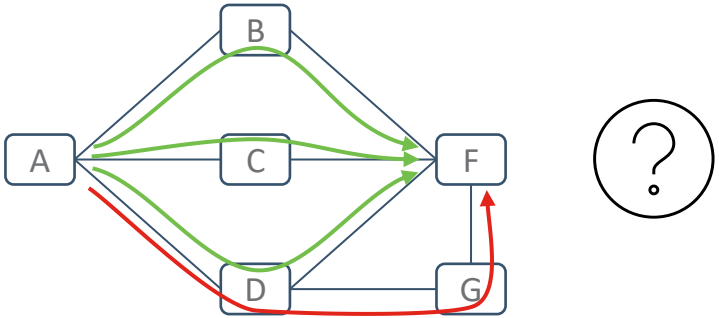
[COLT's presentation on Routing Analytics]

Path Tracing



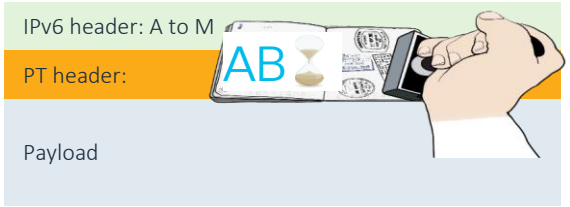
How did the packet arrive from A to F?

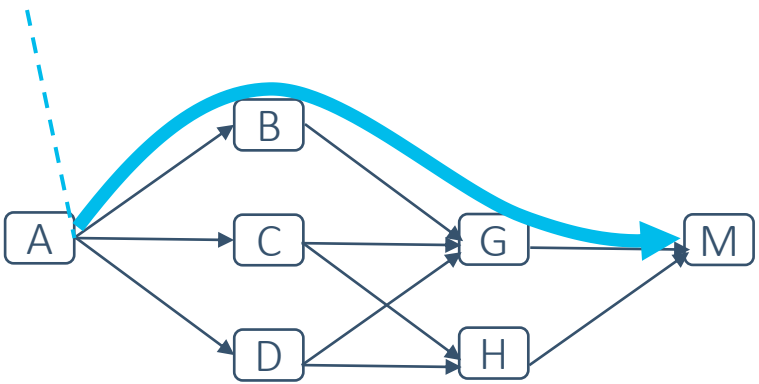
- 3 possible “valid” ECMP paths
  - Any drop?
  - End-to-End Latency homogeneity?
- An invalid path is possible
  - Routing or FIB corruptions
- 40-year-old unsolved IP problem



Stamping Trajectory in PT Header

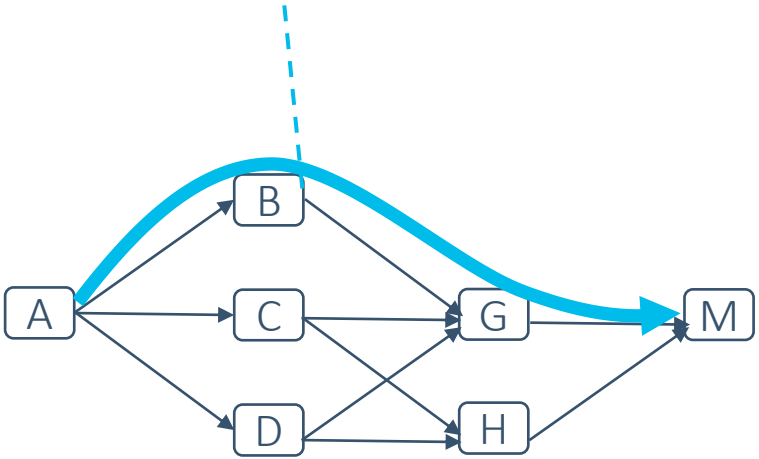
Available Today!





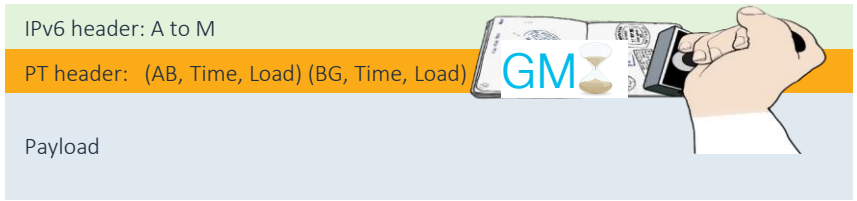
Stamping Trajectory in PT Header

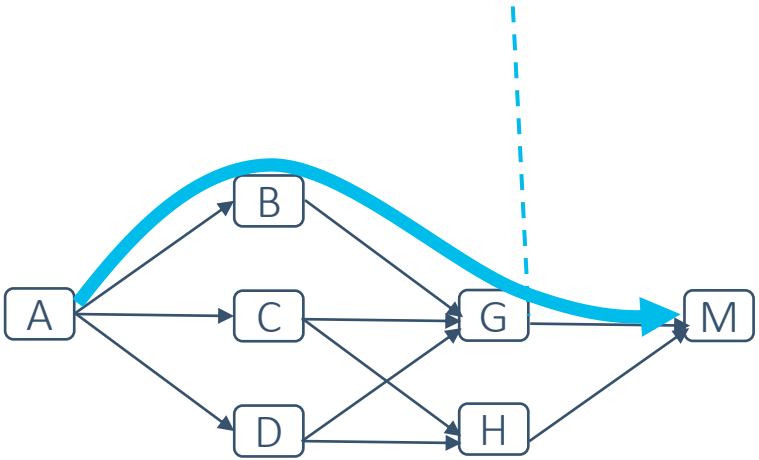
Available Today!



Stamping Trajectory in PT Header

Available Today!





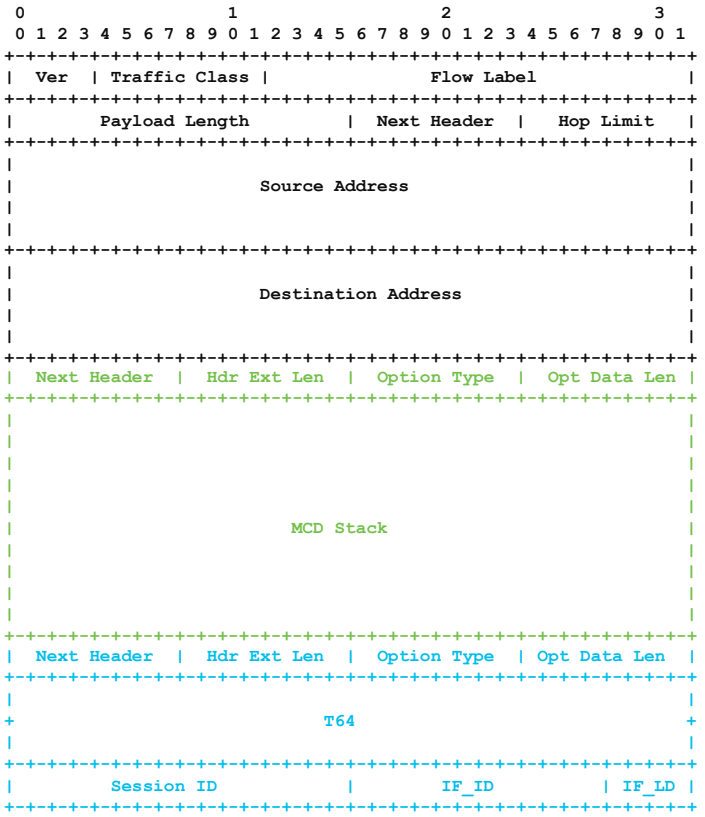
The PT idea

- Stamping in the Packet Header
- Ultra-MTU-efficient: only 3 bytes per hop!
  - 12-bit Interface, 8-bit Timestamp, 4-bit Load
- Implemented in the most basic HW pipeline
  - Linerate for any packet: shifting and writing at fixed offset
  - Reports true packet experience
- Native interworking with legacy nodes
  - Seamless deployment
- Hardware/XR feature with analytics app

Source

- Probe generation

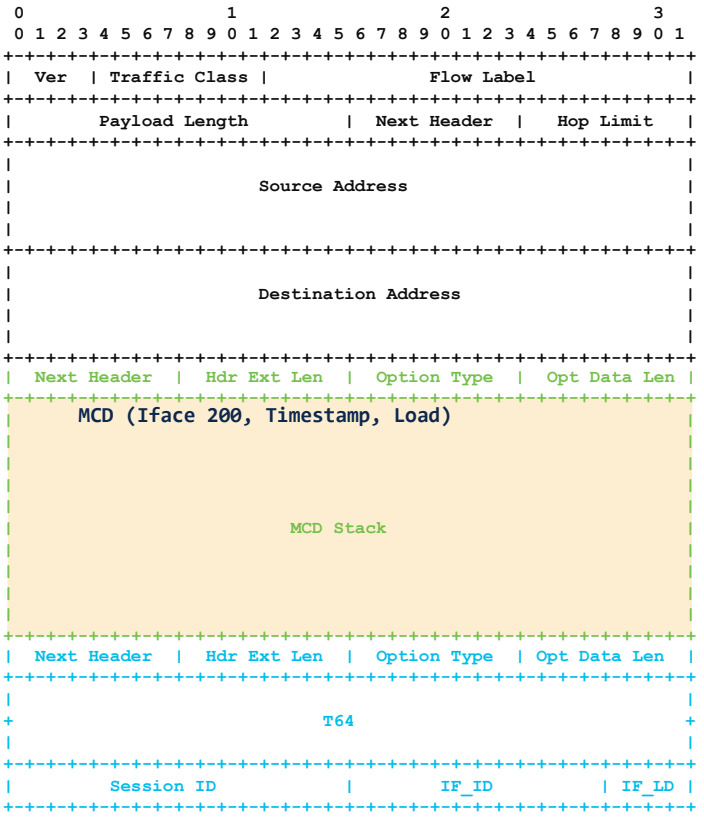
- SRv6 header:
  - Configurable IPv6 DA (ePE to monitor)
  - Configurable DSCP (QoS)
  - Flow Label value sweeping (exercise all ECMP paths)
  - Optional SID List
- HbH-PT:
  - Stack of 12 MCD's set to zero
  - Hop-by-Hop Option
- DO-PT:
  - Probe identification (Session ID)
  - 64bit Timestamp (leverage HW support) + Iface ID
  - Destination Option



Midpoint

- Shift & Stamp
  - Shifts the MCD stack by 3Bytes
  - Stamp (push) new MCD always at the same position (HBH-PT)
- Simple write at fixed location
  - Push on top of the stack
- Shallow write
  - 12 MCD's of 3 bytes is HBH with only 40 bytes

```
/* Path Tracing Midpoint Configuration */
performance-measurement
 interface FourHundredGigE0/0/0/1
  path-tracing
   interface-id 200
```



Sink

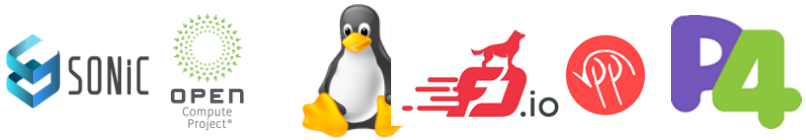
- Received DA acts as BSID
  - uTEF behavior = Timestamp, Encapsulate, Forward
- Push IPv6 Encapsulation and sends toward Regional Collector for analytics

Collected Data

- Source
  - 12-bit Outgoing Interface ID
  - 4-bit Outgoing Interface Load
  - 64-bit PTP Tx Timestamp
- Midpoint (at each node)
  - 12-bit Outgoing Interface ID
  - 4-bit Outgoing Interface Load
  - 8-bit Truncated PTP Tx Timestamp
- Sink
  - 12-bit Incoming Interface ID
  - 4-bit Incoming Interface Load
  - 64-bit PTP Rx Timestamp

Hardware Readiness

- Midpoint functionality available in IOS XR 7.8.1
  - Cisco 8000 (Silicon One Q200; native SDK)
  - NCS5700 (DNX2 - J2; native SDK)
  - ASR9000 (LS)
- Source and Sink functionality in IOS XR 24.1.1
  - ASR9000
- Rich Eco-system
  - Cisco, Broadcom, Marvell, Keysight, +others
  - SAI/SONiC
  - Linux, FD.io VPP, P4, ...
- Ongoing standardization
  - Path Tracing in SRv6 networks (ietf.org)

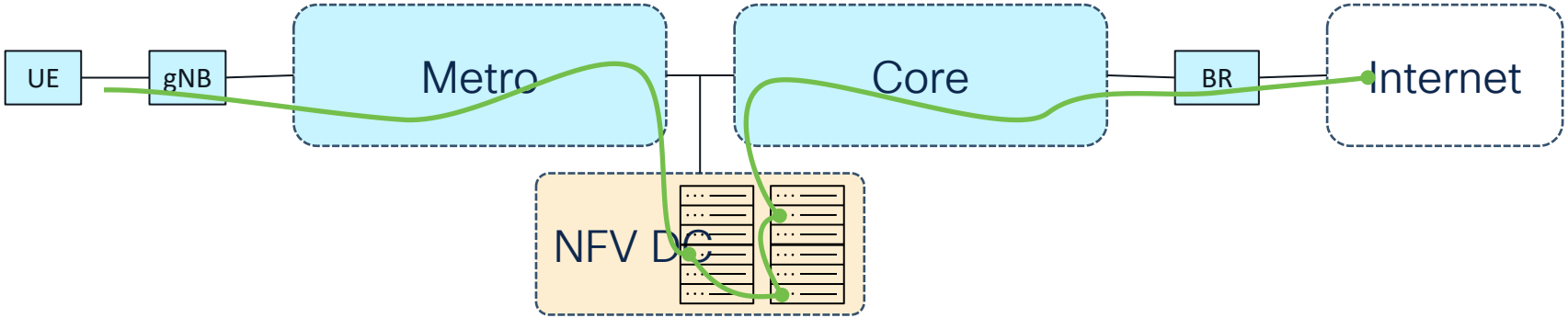


EDM: ECMP Dataplane Monitoring

- EDM detects
  - An expected ECMP path that drops all its traffic (dataplane corruption)
  - An ECMP path that is not expected (routing/dataplane corruption)
  - Incoherent latency between ECMP paths
- EDM measures
  - End-to-end latency of each path (0.06msec in WAN, 0.2usec in DC)
- Current technique of sending probes from anywhere to anywhere without any PT data requires AI processing of huge data sets

Demo/TOI

NFV: Latency Analytics and Proof of Transit



- Did the packet go through the right NFV?
- PT fully identifies the trajectory and time taken for NFV processing

Conclusion

IP is back and better than ever.

SR v6 uSID

Build anything

Simplified, scalable, and versatile networks that are self-sufficient

Self-sufficiency is standard

End-to-end policy
- From Host to Internet through DC, Access, Metro, Core, Cloud
- No protocol conversion or gateways at domain boundaries

Any service, without any shim
- VPN, Slicing, Traffic Engineering, Green Routing, FRR, NFV

Better scale, reliability, cost, and seamless deployment in Brownfield

Essential embedded assurance

Active probing between Fabric Edges **along all ECMP paths**

High-capacity probe generation and ingestion powered by Silicon One (14MPPS)

Continuous **routing monitoring**

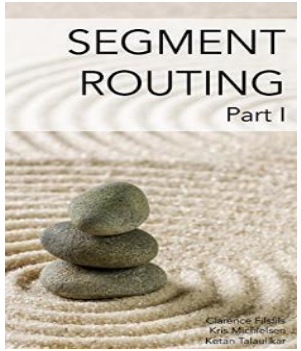
Path Tracing provides full **path characterization** of forwarding path, and how time is spent

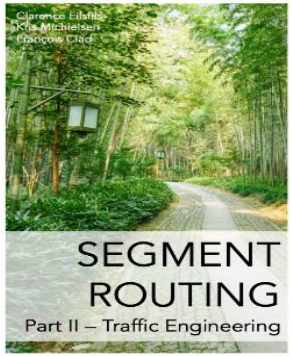
Measure everything

IPM

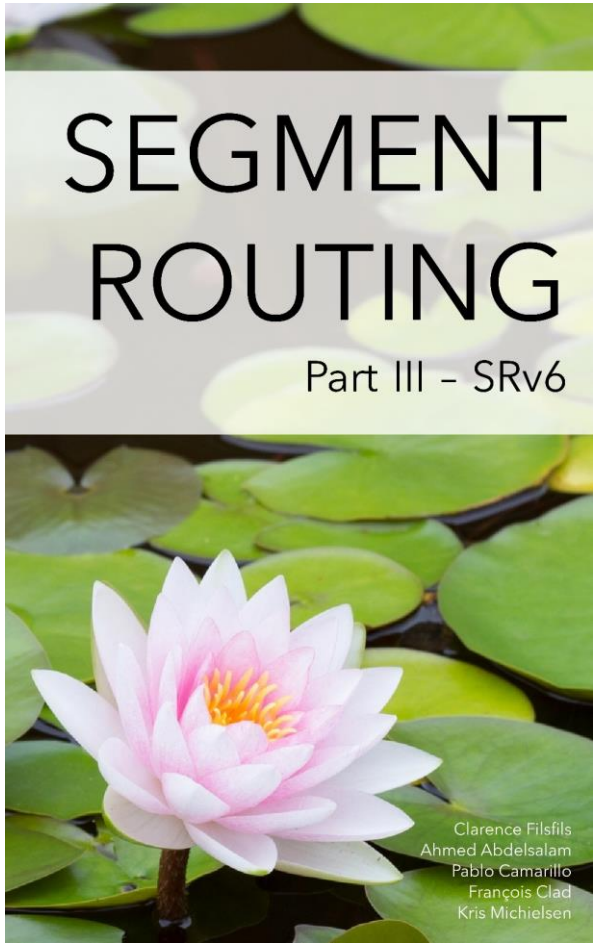
Embedded Transport SLA monitoring

Stay up-to-date





segment-routing.net



https://www.amazon.com/dp/B0D6GWWRWH

SRv6 sessions

- Segment Routing Masterclass [TECSPG-1000]
  - Monday, 8:30 AM - 1:00 PM CET
- Introduction to Segment Routing [BRKSP-2551]
  - Monday, 2:00 PM - 3:30 PM CET
- Introduction to SRv6 uSID Technology [BRKSPG-2203]
  - Tuesday, 12:30 PM - 4:00 PM CET
- Modernizing Private WAN Architecture for Critical Networks Infrastructure [BRKSPG-2063]
  - Tuesday, 9:30 AM - 11:00 AM CET
- SRv6 and Cloud-Native: a Platform for Network Service Innovation [LTRSPG-2212]
  - Wednesday, 8:30 AM - 1:00 PM CET
- Explore the Power of SRv6: Unleashing the Potential of Next-Generation Networking [LTRSPG-2006]
  - Thursday, 8:30 AM - 1:00 PM CET
- Advanced Innovations in SRv6 uSID and IP Measurements [BRKSPG-3198]
  - Thursday, 2:15 PM - 3:15 PM CET
- Segment Routing Innovations in IOS XE [BRKENT-2520]
  - Friday, 9:00 AM - 10:30 AM CET

Webex App

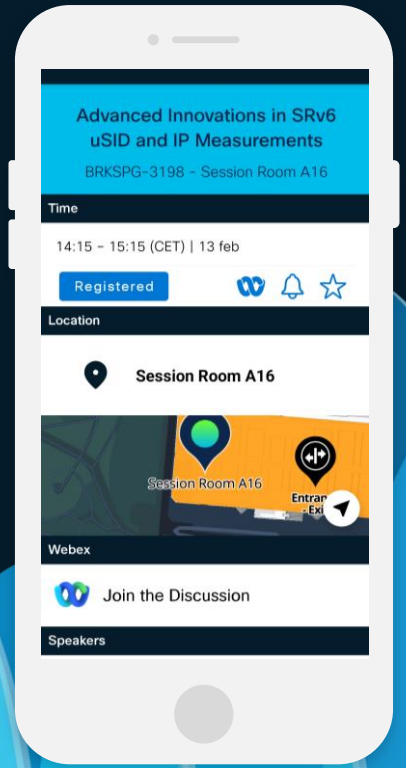
Questions?

Use the Webex app to chat with the speaker after the session

How

1. Find this session in the Cisco Events mobile app
2. Click “Join the Discussion”
3. Install the Webex app or go directly to the Webex space
4. Enter messages/questions in the Webex space

Webex spaces will be moderated by the speaker until February 28, 2025.



cisco Live!

Fill Out Your Session Surveys

Participants who fill out a minimum of 4 session surveys and the overall event survey will get a unique Cisco Live t-shirt.

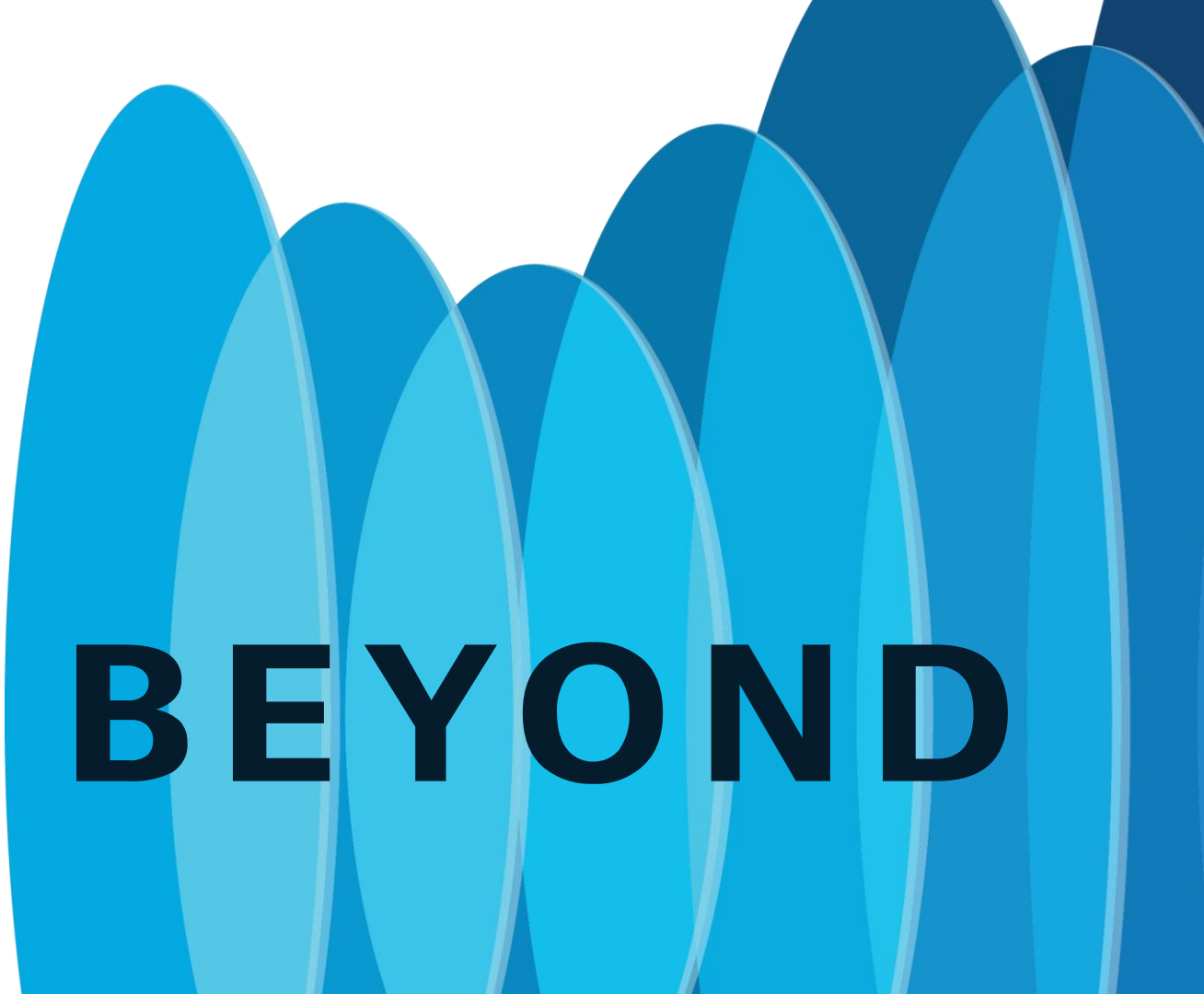
(from 11:30 on Thursday, while supplies last)

All surveys can be taken in the Cisco Events mobile app or by logging in to the Session Catalog and clicking the ‘Participant Dashboard’

Content Catalog


cisco
Thank you
ask-segment-routing@cisco.com
pcamaril@cisco.com
cisco Live!

cisco Live!



GO BEYOND