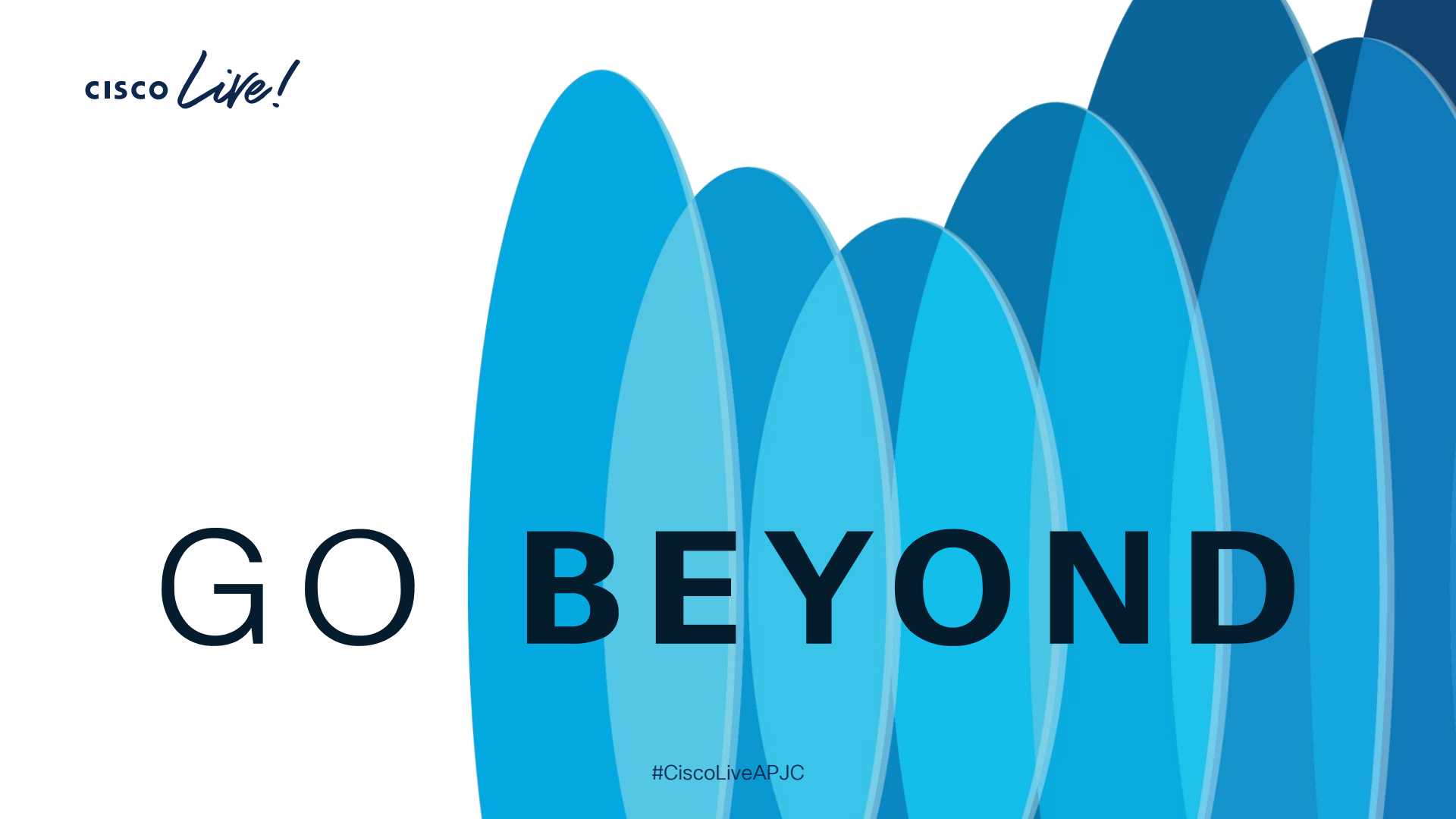

cisco Live!
GO BEYOND
#CiscoLiveAPJC

Unified Network Fabrics with Segment Routing

Architecting Network-as-a-Service

Cris He (Hengyi), SR. Solution Architect

BRKSPG-2851

Cisco Webex App

https://ciscolive.ciscoevents.com/ciscolivebot/#BRKXXX-xxxx

Questions?

Use Cisco Webex App to chat with the speaker after the session

How

1. Find this session in the Cisco Live 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 November 15, 2024.

cisco Live!

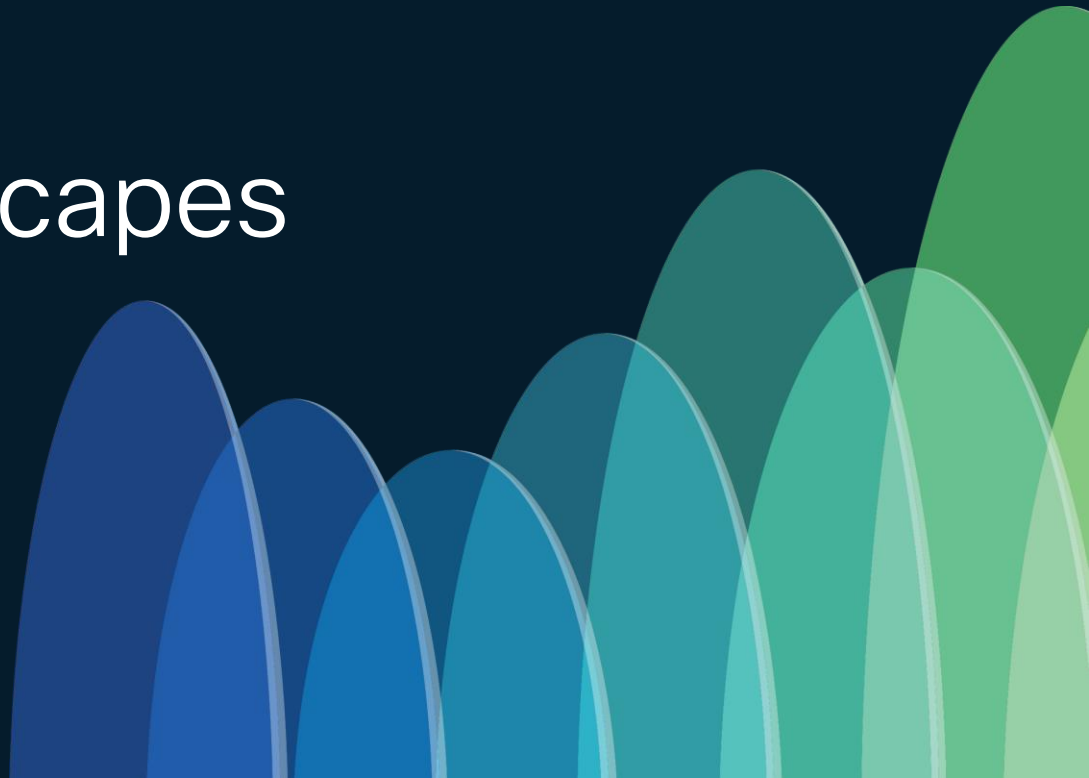
© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public

Agenda

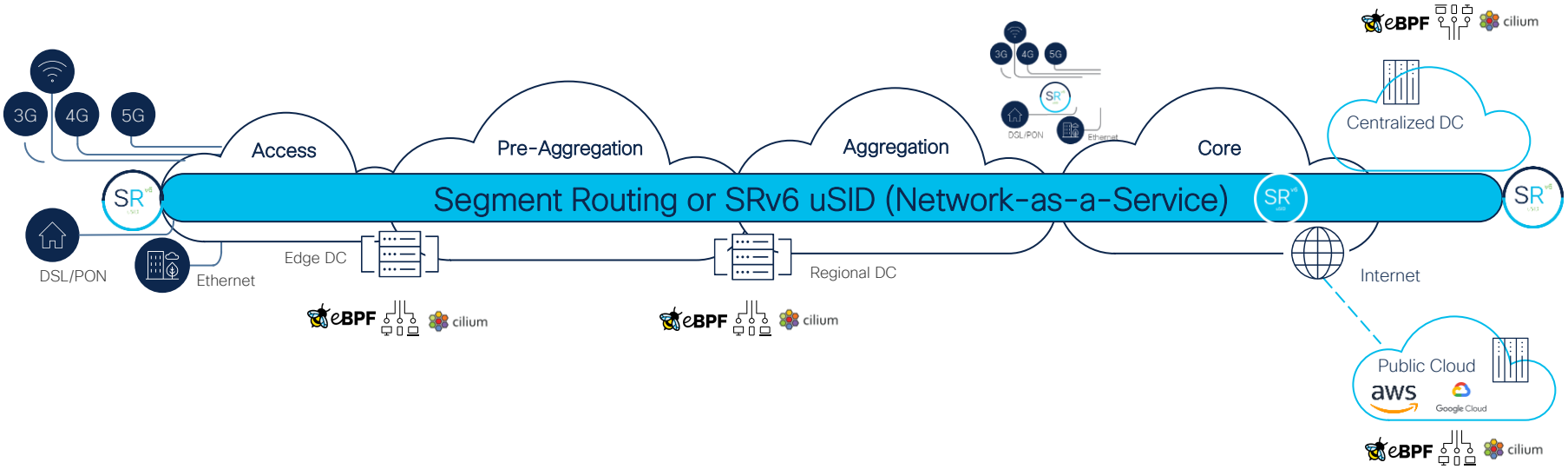
- Overall Landscape: SRv6 Unified Network Fabrics
- Multi-Domain Networking Addressing Scheme
- Next-Gen Cloud Native DC powered by SRv6
- SRv6 Multi-Network – Core/Transport/Peering/Data Center

© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public

Overall Landscapes



Unified Network Fabrics with Segment Routing: – “Network-as-a-Service”



Any Service over SR programable network

Unified Solution with Better Reliability

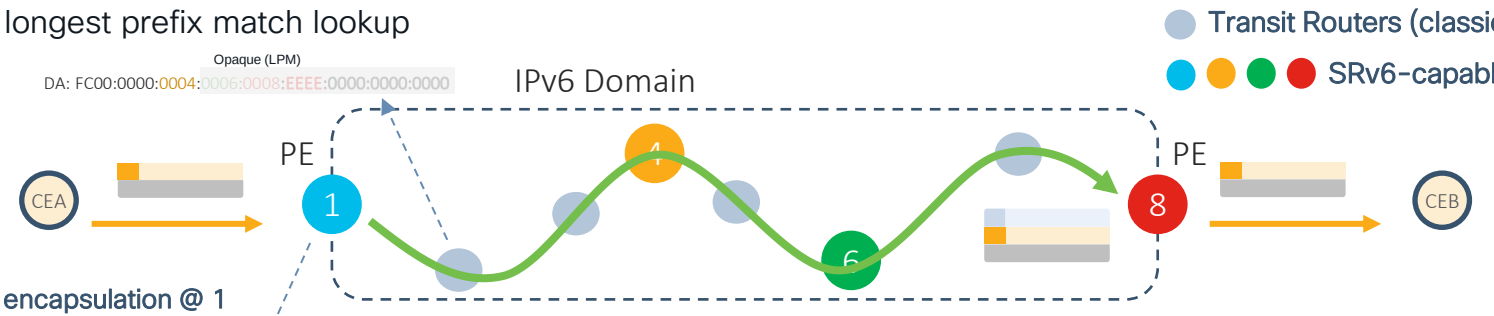
Seamless Brownfield Deployment

Native Host and Cloud

© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public

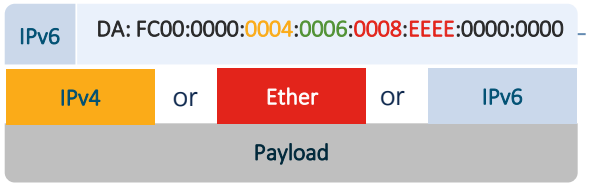
SRv6 uSID – The un-expected Innovation

LPM – longest prefix match lookup



Simple IPv6 encapsulation @ 1

- IPv4 in IPv6 (shown)
- IPv6 in IPv6
- Ethernet in IPv6

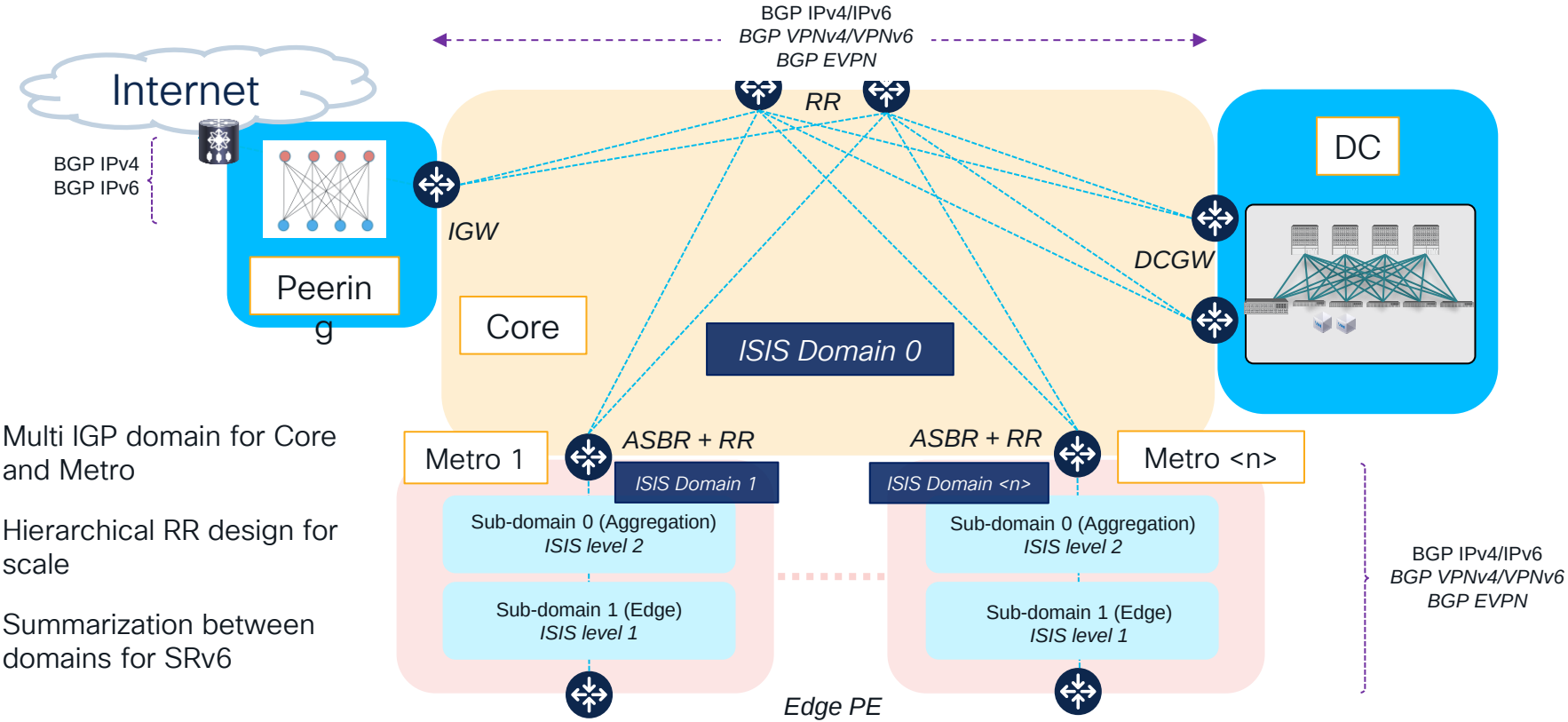


A **source-routed path encoded in a single IPv6 address**!

- follow IGP shortest-path to node 4,
- then shortest-path to node 8,
- then shortest-path to node 6,
- then decapsulate and lookup in VPN table

© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public

Recap: SRv6 Core/Transport Network



- Multi IGP domain for Core and Metro
- Hierarchical RR design for scale
- Summarization between domains for SRv6

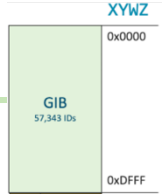
© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public

Network Addressing Scheme

SRv6 Unified Fabrics

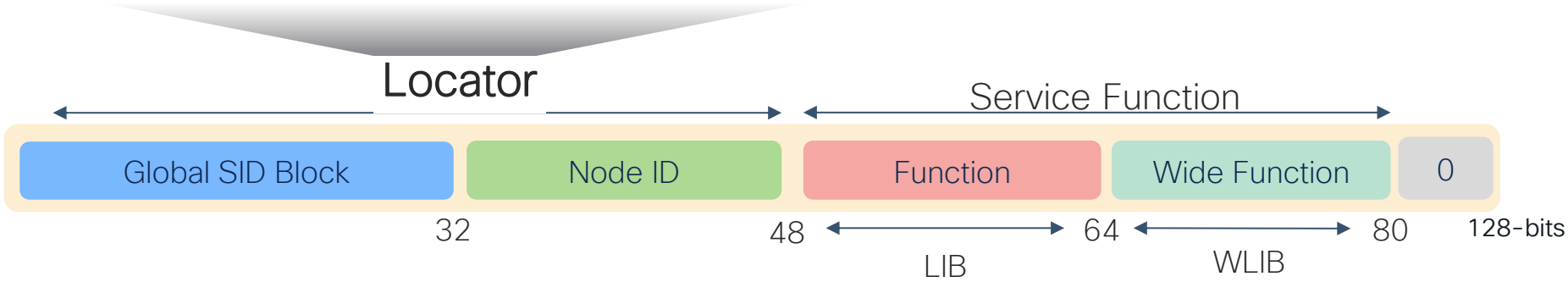
SRv6 uSID Addressing Scheme

- **Example**: Node 3 in Metro Domain1 gets the locators:
- uSID: FC11:1100:0103/48 (low-cost)
  FC11:1101:0103/48 (low-delay)



uSID: FC11:11/24

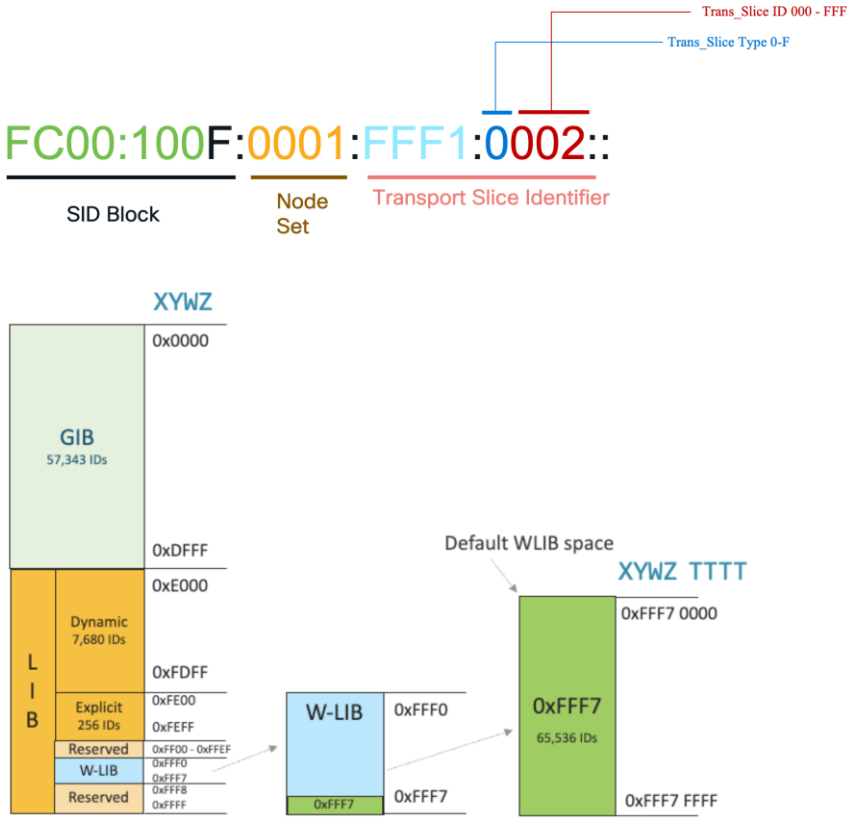
SID:SpaceBlockID:DomainIDNodeI



© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public

SRv6 WLIB Overview

- The ability to allocate and advertise service SID with wide function
  - WLIB provides an extended set of IDs available for service function allocation that can be used when a PE with large-scale VRF or Pseudowire termination requires more local uSIDs than provided from the LIB



© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public

SID Allocation (With WLIB)

Service context with Application Awareness

```
segment-routing
 srv6
  formats
   format usid-f3216
     usid wide-local-id-block explicit start 0xfff1
```

```
segment-routing
 srv6
  locators
   locator MAIN
     micro-segment behavior unode psp-usd
     prefix fd00:1000:2011::/48
```

```
!!!BGP Configuration: Explicit uDT SID for VRF!!!
router bgp 1
vrf test
address-family ipv6 unicast
segment-routing srv6
    locator MAIN
    alloc mode per-vrf-46
```

```
segment-routing
  srv6
    static
      endpoint
        sid fd00:1000:2011:fff1:c1:: behavior end-udt46   <<< new Wlib support, new context type
          allocation-context vrf test                     <<< new context parameter
          forwarding                                      <<< new sub-mode
```


© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public

SRv6 Explict WLIB Range

```
#show segment-routing srv6 manager
Wed Nov 16 07:12:40.860 UTC
Parameters:
  SRv6 Enabled: Yes
  SRv6 Operational Mode:
    Micro-segment:
      SID Base Block: fc00:1000::/24
  Encapsulation:
    Source Address:
    Hop-Limit: Default
    Traffic-class: Default
  SID Formats:
    f3216 <32B/16NFA> (2)
      uSID LIB Range:
        LIB Start   : 0xe000
        ELIB Start  : 0xe100 (configured)
      uSID WLIB Range:
        EWLIB Start : 0xfff1 (configured)
```

```
Platform Capabilities:
Endpoint behaviors:
    End.DX6
    End.DX4
    End.DT6
    End.DT4
    End.DT46
    End.DX2
    End.DT2U
    End.DT2M
    End (PSP/USD)
    End.X (PSP/USD)
    uN (shift)
    uN (PSP/USD)
    uA (shift)
    uA (PSP/USD)
    uDX6
    uDX4
    uDT6
    uDT4
    uDT46
    uDX2
    uDT2U
    uDT2M
    uB6 (Insert.Red)
  Configurable parameters (under srv6):
    Ranges:
      LIB           : Yes
      WLIB          : Yes
```


© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public

SRv6 SID Allocation from Explict WLIB

```
RP/0/RP0/CPU0:Leaf1#show segment-routing srv6 locator
Sat Dec  2 13:53:54.299 UTC
Name                  ID       Algo  Prefix                    Status   Flags
--------------------  -------  ----  ------------------------  -------  --------
MAIN                  1        0     fd00:1000:2011::/48       Up       U
```

```
RP/0/RP0/CPU0:Leaf1#show segment-routing srv6 sid
Sat Dec  2 13:52:51.192 UTC

*** Locator: 'MAIN' ***

fd00:1000:2011::            uN (PSP/USD)      'default':8345                 sidmgr              InUse  Y
fd00:1000:2011:fff1:c1::    uDT46             'MAIN'                      sidmgr,sidmgr,bgp-300  InUse  Y
```


© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public

One Loopback for all

SRv6 Locator: FC11:1100:0103::/48	SRv6 Loopback: FC11:1100:0103::1/128

- Use one loopback for SRv6 encapulation source and BGP overlay
- Single redistribution entity is needed for locator block, improved scale
- Fate sharing between locator reachability and BGP NH
- SA and DA from same SRv6 space in the service packet (simplified service traffic security treatment)

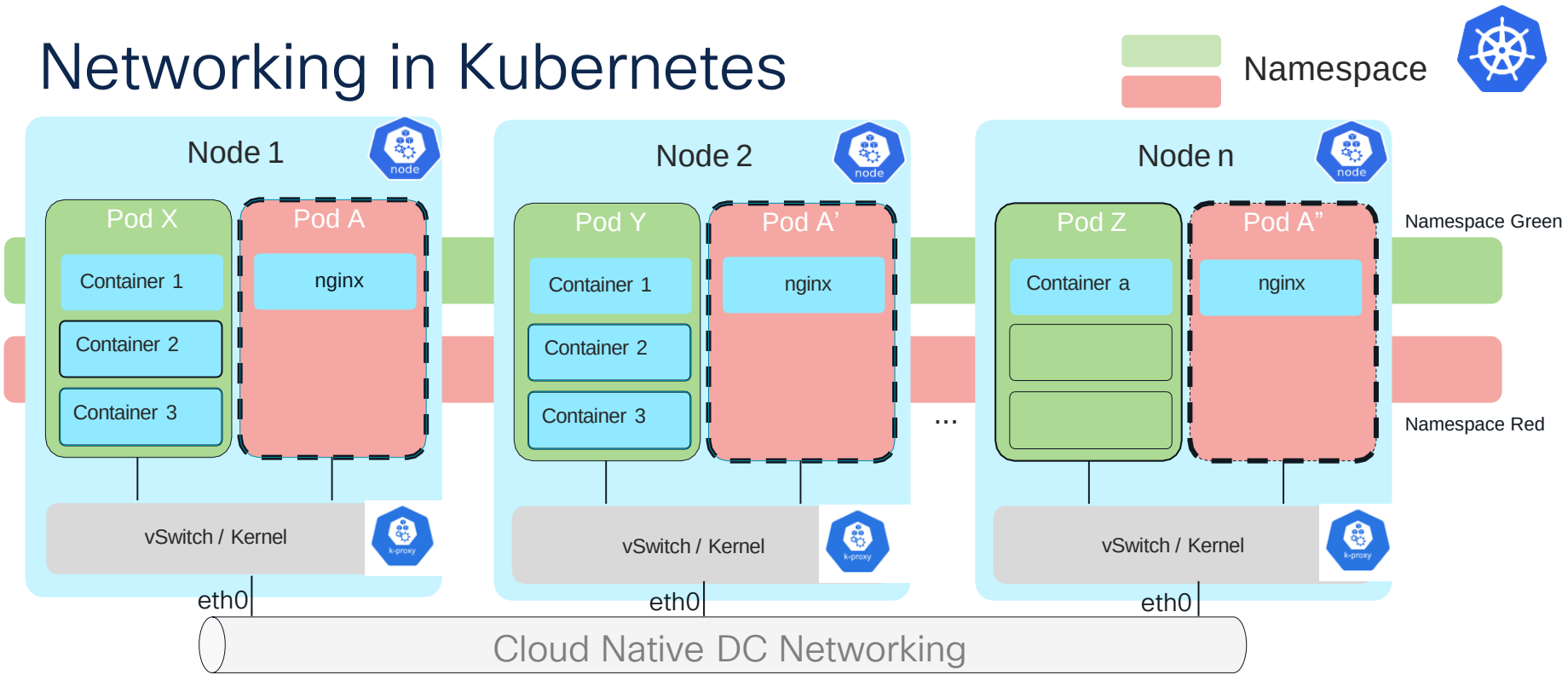
© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public

Next-Gen Cloud Native DC

Powered by SRv6

cisco Live!

Networking in Kubernetes



Node:

- A server (bare-metal or VM)

Cluster:

- Collection of nodes

Pod:

- Collection of Containers

Deployment:

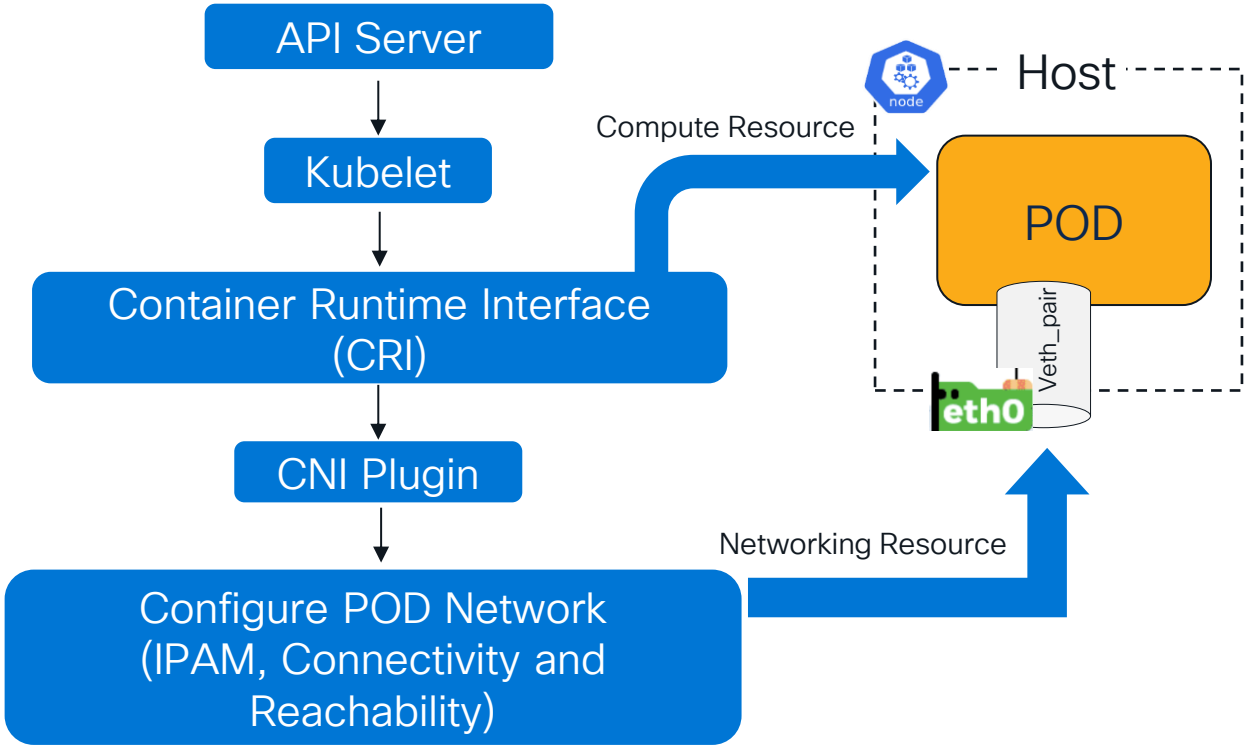
- Collection of Pods

Container Network Interface:

- CNI plugin provides the Network functionality

© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public

Container Network Interface (CNI)



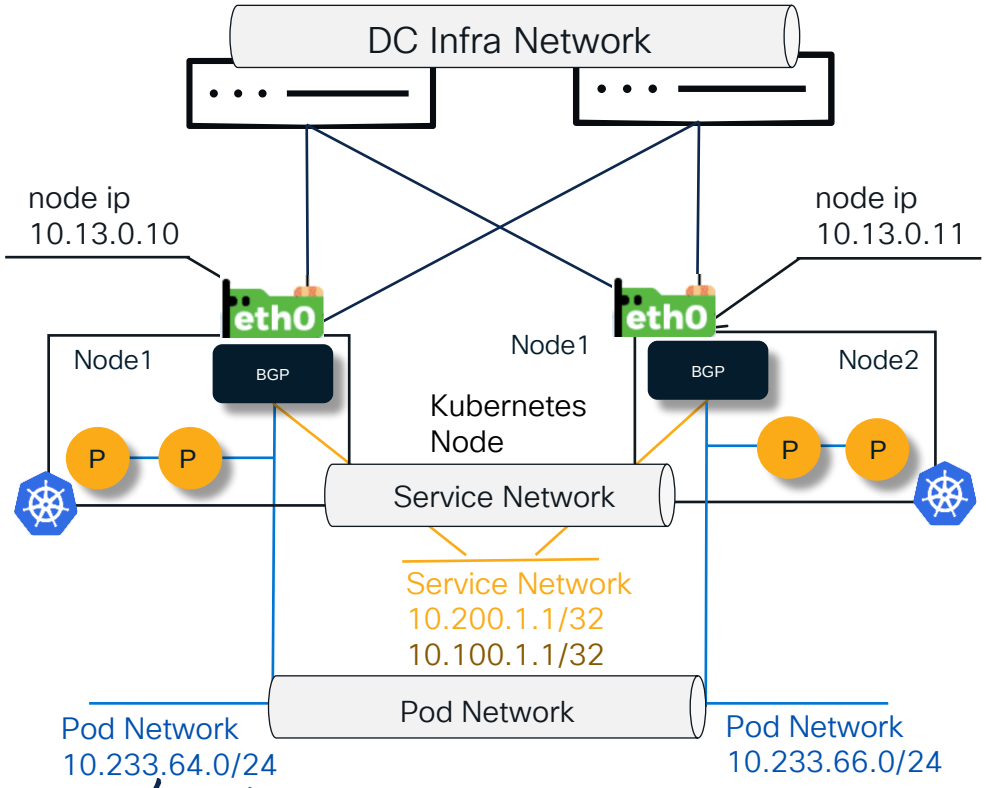
- A framework for dynamically configuring networking resources.
- Uses a group of libraries and specifications written in Go.
- The plugin specification defines an interface for configuring the network, provisioning IP addresses, and maintaining connectivity with multiple hosts.
- Cilium , Calico , Flannel, etc

https://github.com/containernetworking/cni

© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public

How many Networks in Cloud Native Environment

Simplified

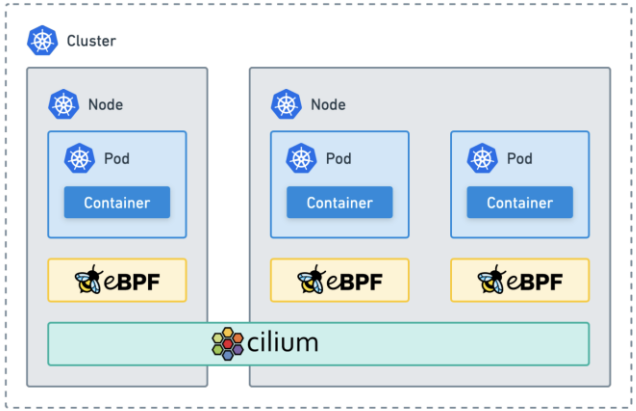


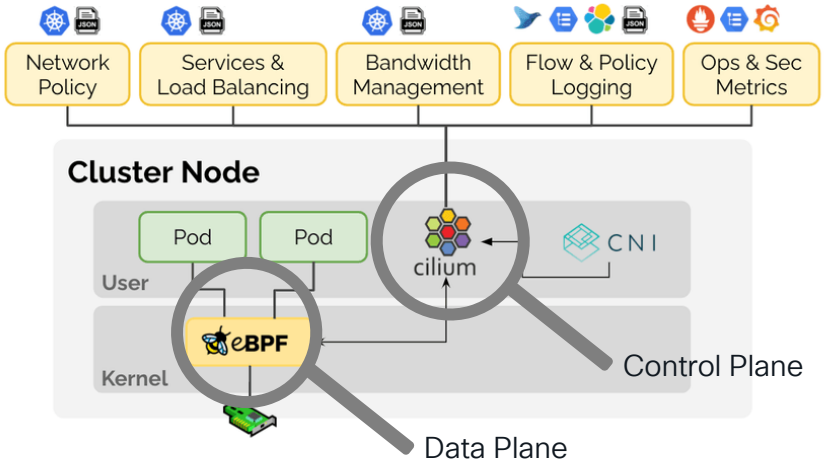
- Each Kubernetes node has one node IP
- One or more ranges of IP addresses (CIDRs) for pod networks
- One or more service networks shared by the cluster
- Provide the network for Pod-to-Pod connectivity without NAT
- Provide Service connectivity as well as PoD-to-Service connectivity
- Service and PoD network can be overlay Network

© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public

What is Cilium?

CNI plugin for Kubernetes

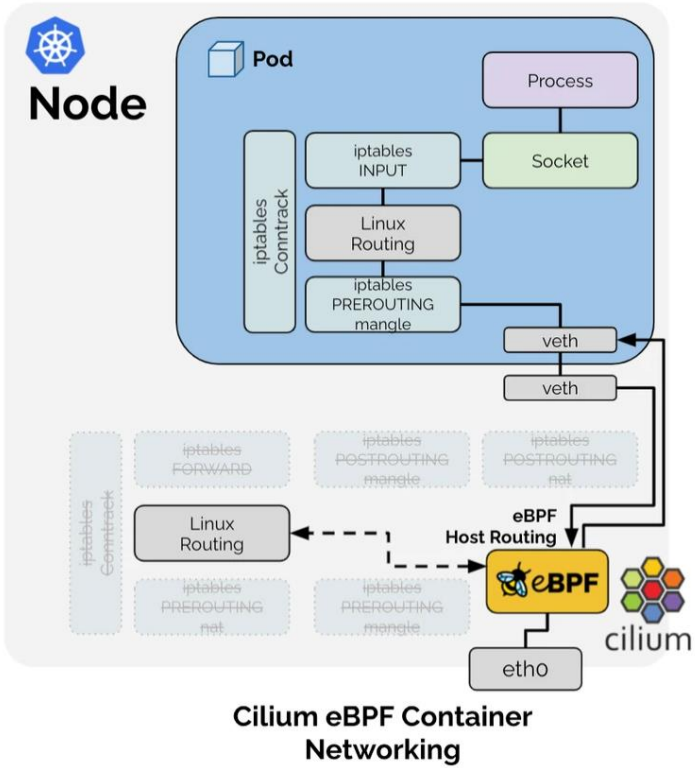




- An Agent on each node
- Decoupling of Control Plane (Cilium as CNI) and Forwarding Plane (eBPF)
- Programmatic interfaces to define workload dataplane and policy execution
  - Routing, segmentation/overlays, access control, FW, LB, NAT, etc.

© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public

eBPF in Dataplane



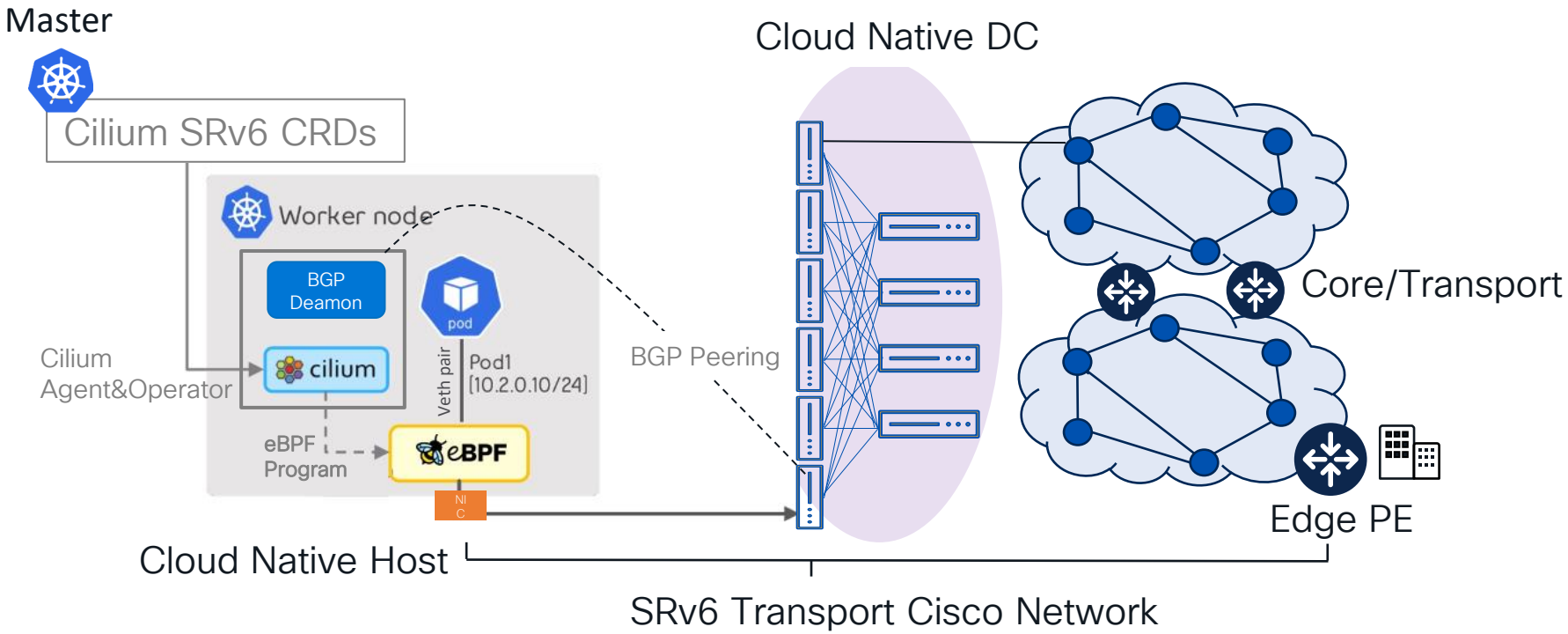
Cilium eBPF Container Networking

Forwarding Plane @ eBPF:

- Host namespace iptables and connection tracking fully bypassed
- Uses pre-defined hooks into the kernel
- Much more efficient packet filtering
- Allows for tracing, performance analysis, debugging, security, and more
- Iptables are no longer required to perform service-to-pod translation
- Kube-proxy is no longer needed

© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public

SRv6 on Cilium/eBPF



© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public

Configration Workflow

4. Build BGP underlay and overlay BGP peering
   - Advertise and learn Locators as Underlay reachability.
   - For the overlay service connectivity, advertise POD CIDR along with the SRv6 service SID Learn remote POD CIDRs with the service SID.

ISOVALENT
now part of CISCO

Define SRv6 Locator pools

Cilium SRv6 SID managers

2.Allocate SID based on L3VPN service context (VRF)

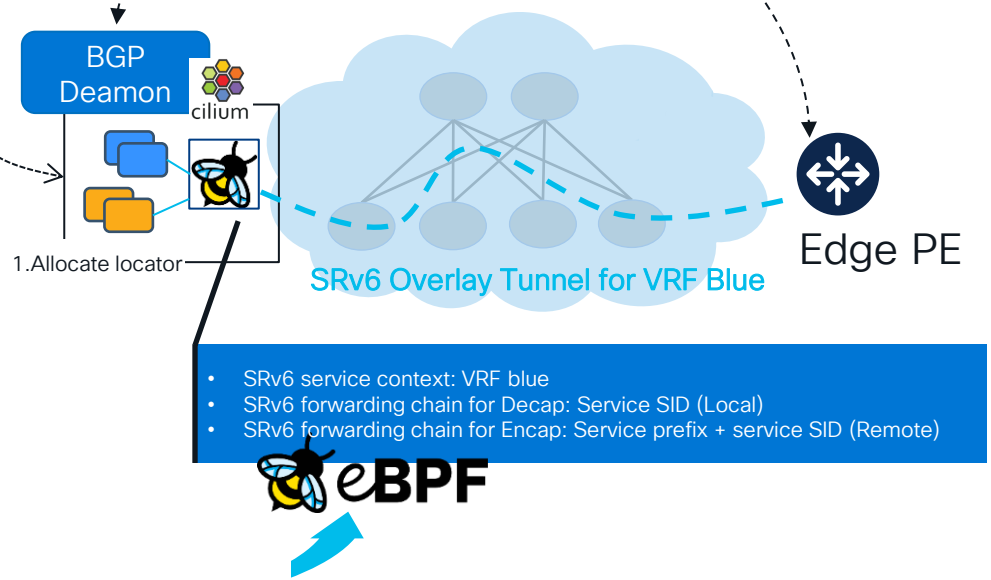
Cilium VRF

3.Create POD in VRF

Cilium BGP Peering Policies

Cilium SRv6 Egress policies

BGP IPv6 Unicast and BGP SRv6 L3VPN



5. Cilium to program eBPF for SRv6 datapath policies

© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public

Define SRv6 Locator and L3VPN service context(VRF)

```
apiVersion: cilium.io/v2alpha1
kind: CiliumSRv6LocatorPool
metadata:
  name: pool0
  labels:
    export: "true"
spec:
  behaviorType: uSID
  prefix: fd00:1000:1::/48
  locatorLenBits: 64
  structure:
    locatorBlockLenBits: 32
    locatorNodeLenBits: 16
    functionLenBits: 32
    argumentLenBits: 0
```

```
apiVersion: cilium.io/v2alpha1
kind: CiliumSRv6VRF
metadata:
  name: blue
spec:
  vrfID: 1000012
  importRouteTarget: "12:12"
  exportRouteTarget: "12:12"
  locatorPoolRef: pool0
  rules:
  - selectors:
    - endpointSelector:
        matchLabels:
          vrf: blue
    destinationCIDRs:
    - 0.0.0.0/0
```

```
# kubectl apply -f Pod.yaml
pod/bluepod created
### Pod.yaml
apiVersion: v1
kind: Pod
metadata:
  labels:
    vrf: blue
  name: bluepod
spec:
  containers:
  - image: busybox:latest
    imagePullPolicy: IfNotPresent
    name: bluepod0
```

```
# kubectl get pod -l vrf -A --show-labels -o wide
NAMESPACE   NAME        READY   STATUS       RESTARTS   AGE    IP                    NODE
NOMINATED NODE   READINESS GATES   LABELS
default             bluepod      1/1        Running     0                 73m   10.245.0.88   ubuntu-k8s-server-01
<none>                          <none>                  vrf=blue
```


© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public

Validating the SRv6 SID Allocation

```
- apiVersion: cilium.io/v2alpha1
  kind: CiliumSRv6SIDManager
  metadata:
    creationTimestamp: "2024-08-08T14:03:13Z"
    generation: 25
    name: ubuntu-k8s-server-01
    resourceVersion: "8406816"
    uid: e169f734-f08a-4da6-96c6-0ecd6c18b46f
  spec:
    locatorAllocations:
    - locators:
      - behaviorType: uSID
        prefix: fd00:1000:1:1cf9::/64
 …
```

```
poolRef: pool0
  status:
    sidAllocations:
    - poolRef: pool0
      sids:
      - behavior: uDT4
        behaviorType: uSID
        metadata: blue
        owner: srv6-manager
        sid:
          addr: 'fd00:1000:1:1cf9:6bb6::'
          structure:
          … omitted …
…
```


© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public

BGP Configuration

```
# kubectl apply -f bgp.yaml
  ciliumbgppeeringpolicy.cilium.io/srv6-control-plane-pod-vpn4-unicast created
# bgp.yaml
apiVersion: cilium.io/v2alpha1
kind: CiliumBGPPeeringPolicy
metadata:
  name: srv6-control-plane-pod-vpn4-unicast
spec:
  nodeSelector:
    matchLabels:
      kubernetes.io/hostname: "ubuntu-k8s-server-01"
  virtualRouters:
    - localASN: 111
      mapSRv6VRFs: true
      exportPodCIDR: true
      srv6LocatorPoolSelector:
        matchLabels:
          export: "true"
      neighbors:
      - peerASN: 701
        peerAddress: "2009:2009::10:1:1:1/112"
        families:
        - {afi: "ipv6", safi: "unicast"}
        - {afi: "ipv4", safi: "mpls_vpn"}
```

NodeSelector: Config on which node

Advertising Locator Prefixes with BGP (IPv6 Unicast)

Advertising the SID for VRF with BGP (IPv4 L3VPN)

© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public

1. Validating BGP peers:

```
# cilium bgp peers
Node                   Local AS   Peer AS   Peer Address        Session State   Uptime   Family         Received   Advertised
ubuntu-k8s-server-01   111        701       2009:2009::10:1:1:1   established     46s      ipv6/unicast   21         2
                                                                                          ipv4/mpls_vpn  9          1
```

2. Validating the BGP IPv6 Unicast Routes :

```
# cilium bgp routes available ipv6 unicast
Node                              VRouter    Prefix                                  NextHop                        Age
Attrs
ubuntu-k8s-server-01   111            fd00:1000:1041::/48       2009:2009::10:1:1:1      9m6s
 [{Origin: i} {AS Path: 701 300 100} {MPReach(ipv6-unicast):
{Nexthop: 2009:2009::10:1:1:1, NLRIs: [fd00:1000:1051::/48 fd00:1000:1042::/48 fd00:1000:1041::/48]}}]
….
```

3. Validating the BGP VPNv4 Routes:

```
# cilium bgp routes available ipv4 mpls_vpn
Node                              VRouter   Prefix                                    NextHop                    Age
Attrs
ubuntu-k8s-server-01   111           1.0.0.41:0:10.21.0.0/24       fd00:1000:1041::1     21h32m21s      Next Line ->
[{Origin: ?} {AS Path: 701 300 100 555} {Extcomms: [12:12]} {Prefix SID attributes: {SRv6 Service TLV: {SRv6 Information Sub
TLV: SID: fd00:1000:1041:: Flag: 0 Endpoint Behavior: 63 {SRv6 Structure Sub Sub TLV: [ Locator Block Length: 32, Locator
Node Length: 16, Function Length: 16, Argument Length: 0, Transposition Length: 16, Transposition Offset: 48] } } } }
{MPReach(l3vpn-ipv4-unicast): {Nexthop: fd00:1000:1041::1, NLRIs: [1.0.0.41:0:192.168.1.0/24 1.0.0.41:0:10.21.0.0/24]}}]
…
```


© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public

Cilium SRv6 Egress Policies

```
apiVersion: v1
items:
- apiVersion: cilium.io/v2alpha1
  kind: CiliumSRv6EgressPolicy
  metadata:
    creationTimestamp: "2024-11-04T03:05:51Z"
    generation: 1
    name: bgp-control-plane-
          ff125587921b5824fd1c9d228e2eb671887d602c50faddf44e962e3c716a0177
    resourceVersion: "8410908"
    uid: 9e838109-7b47-4539-b3b9-32987b7992c9
  spec:
    destination CIDRs:
     - 192.168.1.0/24
     - 10.21.0.0/24
    destination SID: 'fd00:1000:1041:e000::'
    VRF ID: 1000012
kind: List
metadata:
  resourceVersion: ""
```

* Service Prefixes are programmed via eBPF

© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public

SRv6 Cloud Native DC

Network Designs

cisco Live!

Cloud Native DC Networking Design

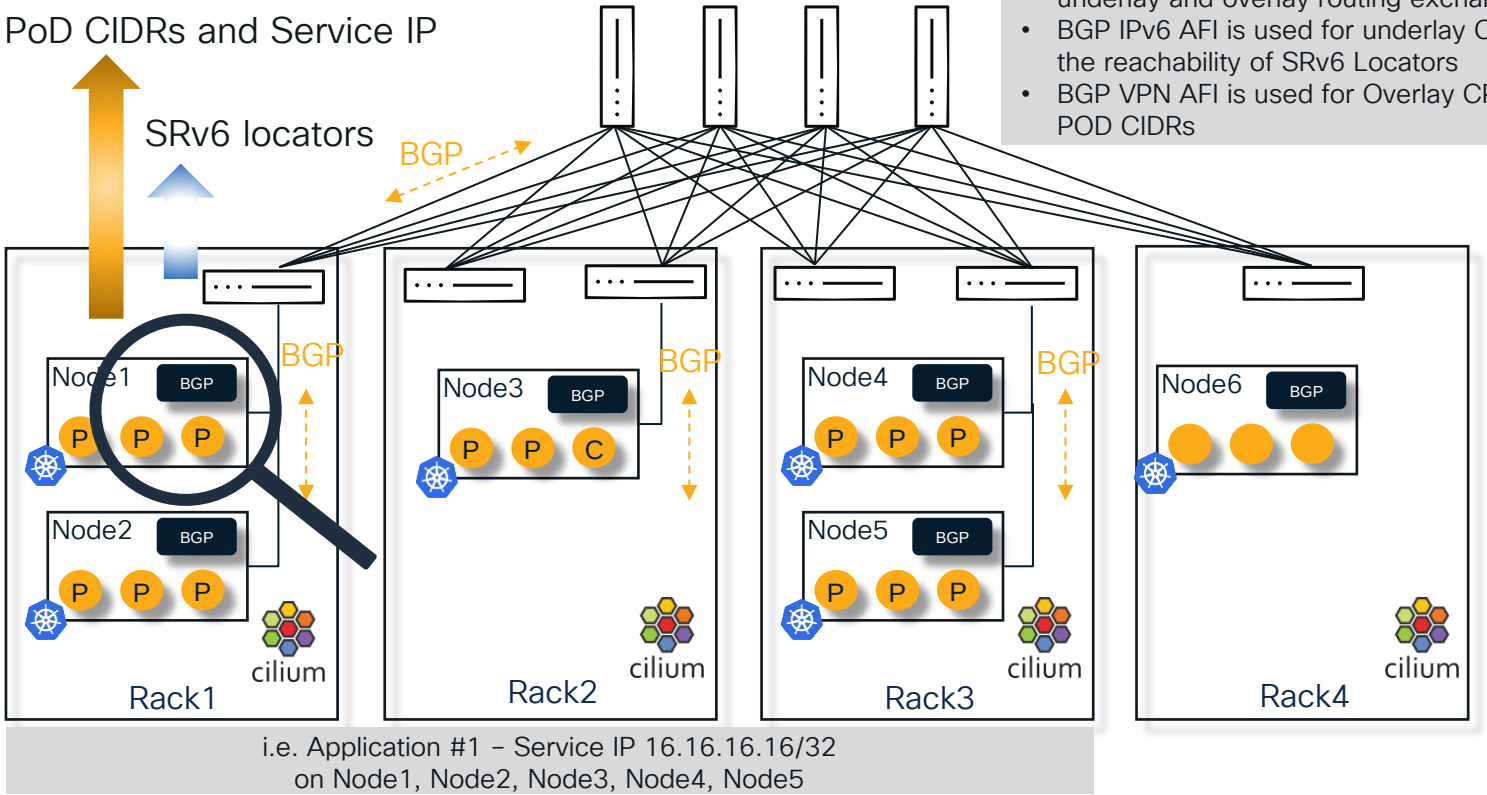
- Cloud Native DC for Kubernetes (K8s) Infrastructure
  - DC Networking as the transport for SRv6 Overlay Service
  - Service connectivity = Cloud Native Applications
  - Kubernetes network over IP Fabric
  - A Routed Fabric is required with use of BGP-only design (because CNI is embedded with BGP speaker)
  - RFC 7938

© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public

The Paradigm

K8s PoD CIDRs and Service IP

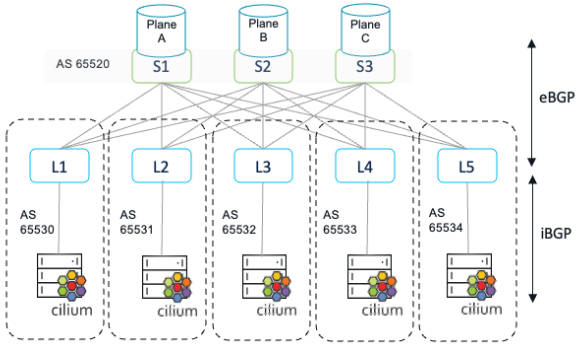
- BGP Peering over P2P link, unified (congruent) or decoupled (incongruent) peering topology for underlay and overlay routing exchange
- BGP IPv6 AFI is used for underlay CP and advertises the reachability of SRv6 Locators
- BGP VPN AFI is used for Overlay CP and advertises POD CIDRs



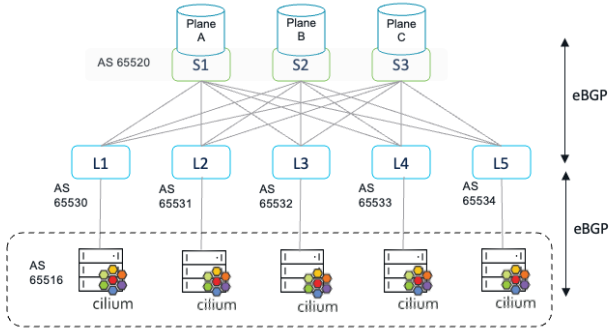
© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public

BGP-Only Cloud-Native DC Design Options

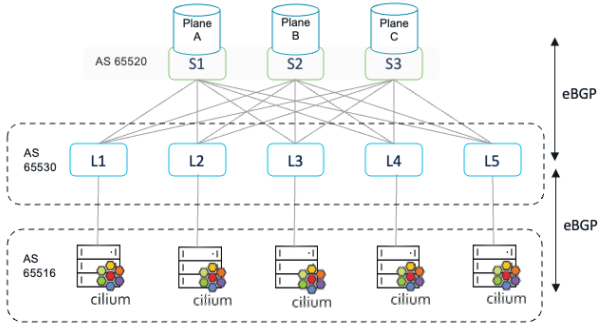
AS-per-rack



Spine-Leaf multi-AS,Cilium AS-per-cluster



Spine-Leaf Dual-AS, Cilium AS-per-cluster



BGP-only Routing in DC Fabric:

- Knobs required to break eBGP loop-prevention
  - 'Disable-peer-as-check + as-overide' or 'Disable-peer-as-check + allow-as-in'
- BGP Overlay NH unchange
- VPNv4/VPNv6 retain-rt-all
- eBGP Multi-hop peering (if required)
- Use 'as-path multipath-relax' for eBGP multi-path selection

© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public

BGP Configuration Simplicity

- Neighbor Configuration complexity can be addressed by dynamic neighbor
  - Dynamic neighbor range: Neighbor subnet range /xyz prefix
  - Real neighbors that dynamically establish peering with the router if the peering addresses fall within one of the configured subnets (i.e. dynamic neighbor range)
- IPv6 fabric with flexible range of P2P infrastructure IP (can use IPv6 LLA RFC 7404)
- BGP Auto-Fabric provides the ultimate simplicity: Self-Organized BGP Fabric

© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public

BGP Routing Stability

- Using the common ASN in Spine & Super spine will maintain the best BGP stability in the DC Fabric
- Using BGP max-AS-length
- Route Aggregation

Impact of Path Hunting

- Large control-plane overhead by the protocol chattiness
  - Depending on the network size and number of prefixes involved, it can be 2x-3x the number of messages
  - Processing overhead on nodes
  - Potential routing impact
  - Unnecessary transient consumption of BW by the valley routing
- Network instability lasts a longer time

© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public

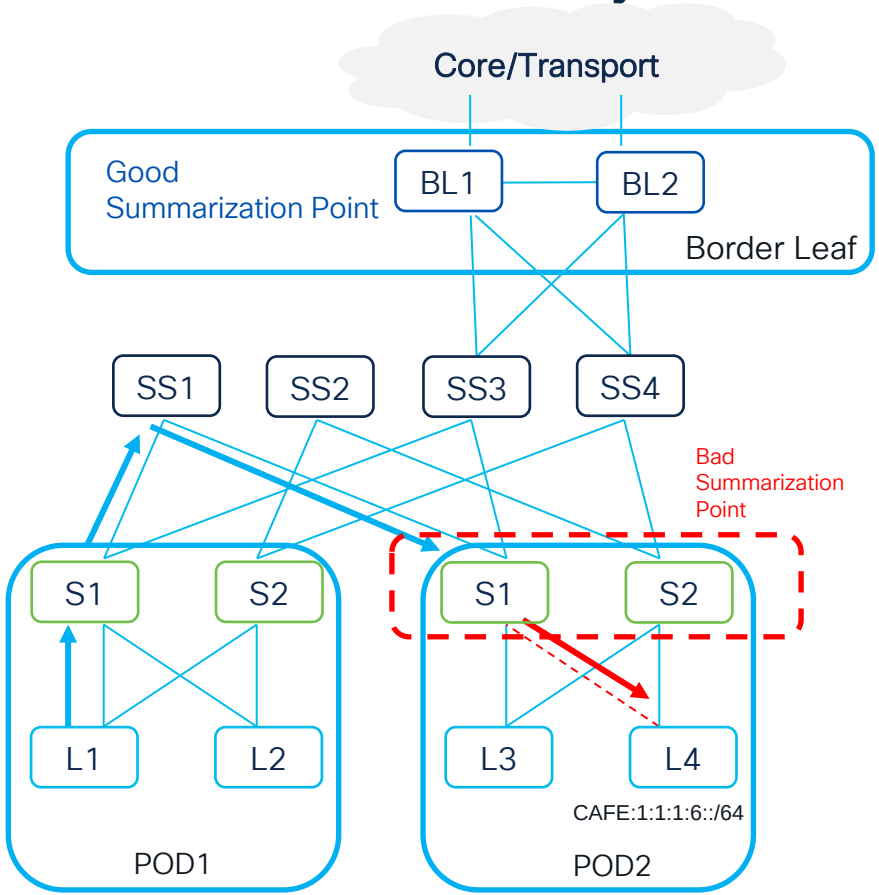
Define the Adequate Summarization Boundary

Problem statement:

- Pod 2 spine summarize CAFE:1:1::/48, Pod 1 spine will learn this route from superspine.
- If link between spine and leaf in pod 2 goes down (red dotted line)
- When pod 1 host want to visit CAFE:1:1:1:6::/64 subnet, it may be hashed to the pod 2 left spine, and the packet get dropped

Conclusion:

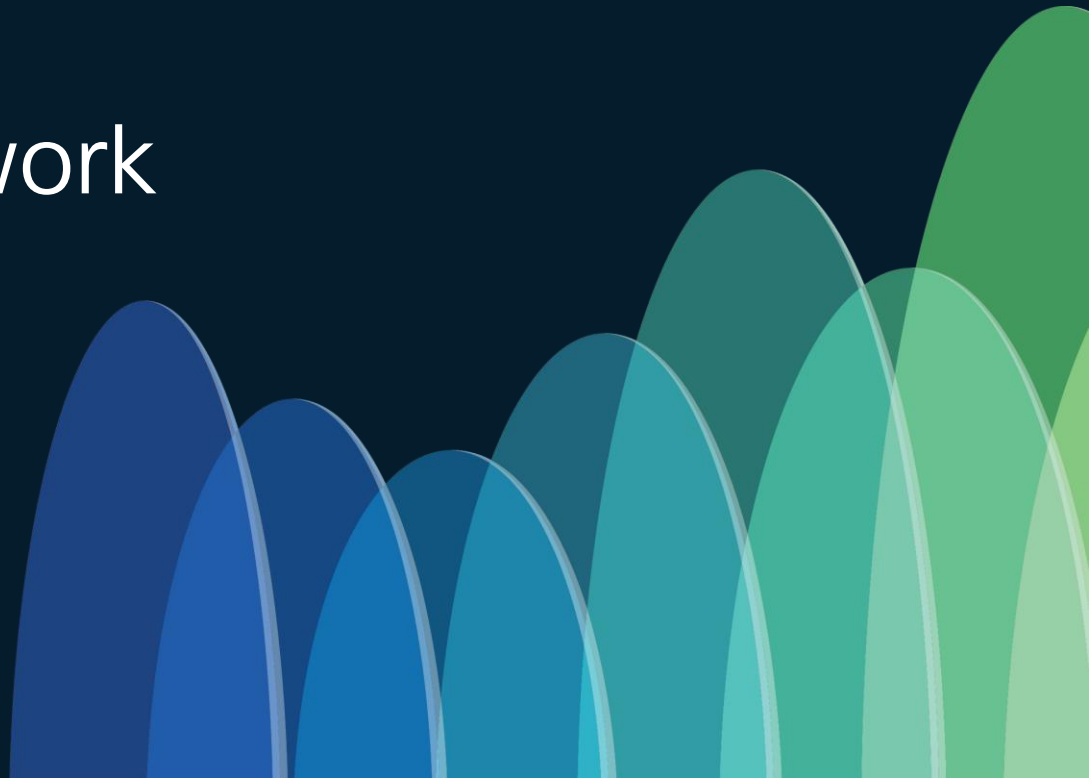
- No route summarization within POD
- Best summarization point is on bolder Leaves with back-to-back connections



© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public

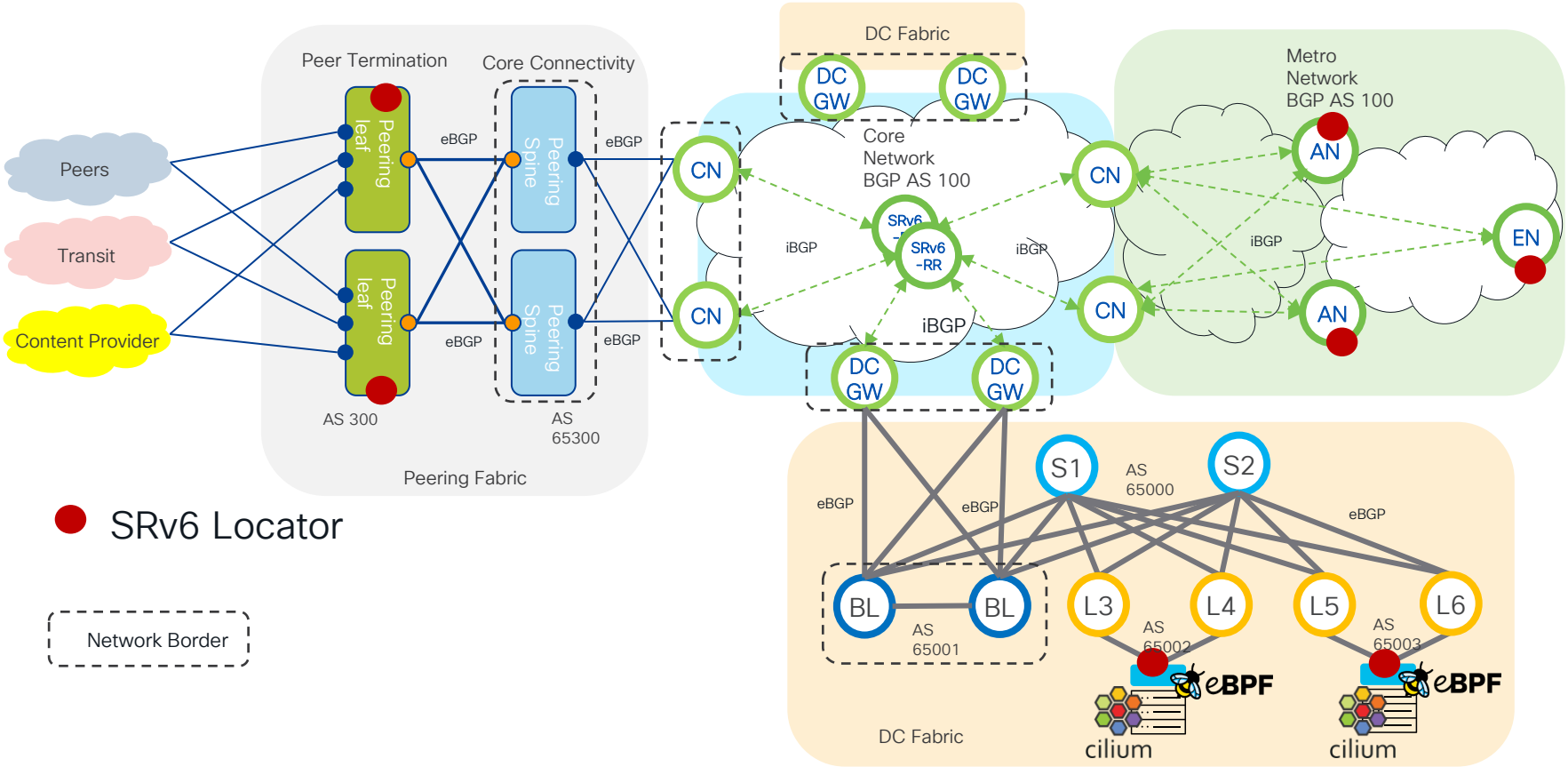
SRv6 Multi-Network

Transport/Peering and DC



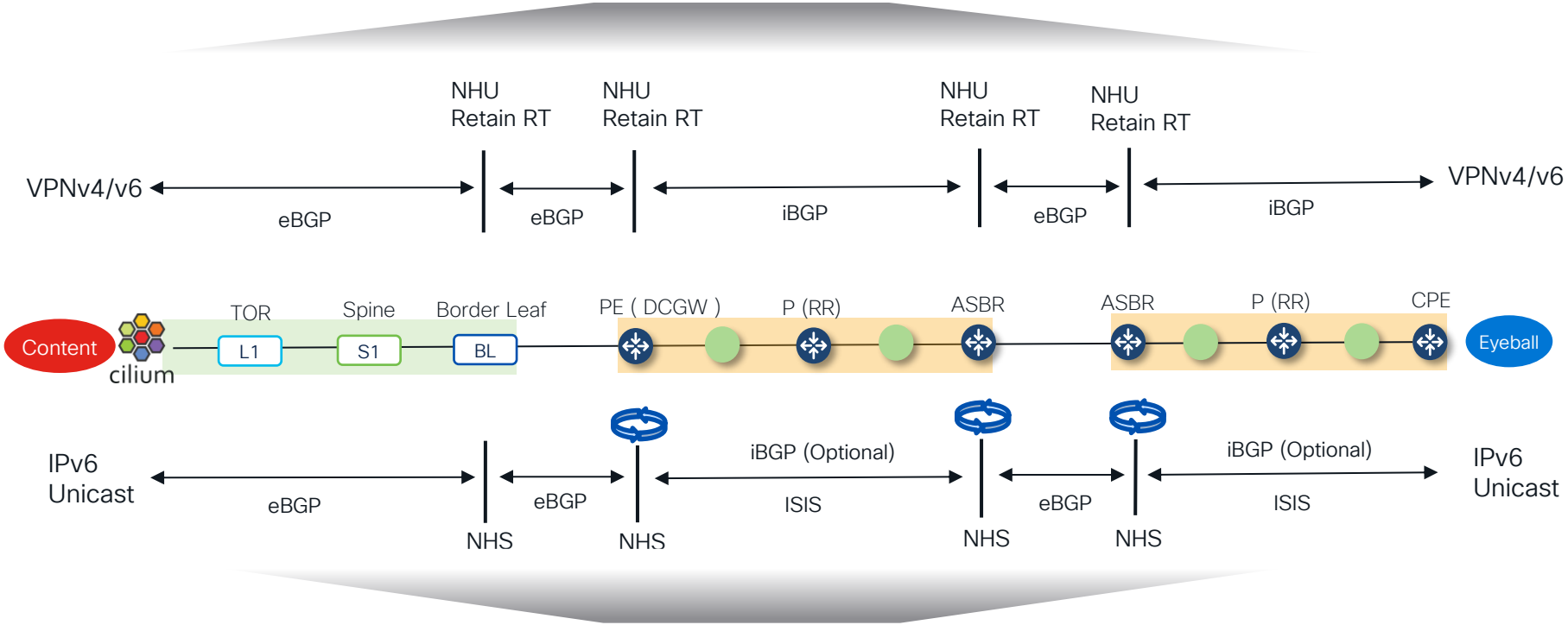
cisco Live!

SRv6 Universality: BGP Inter-Network Overview



© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public

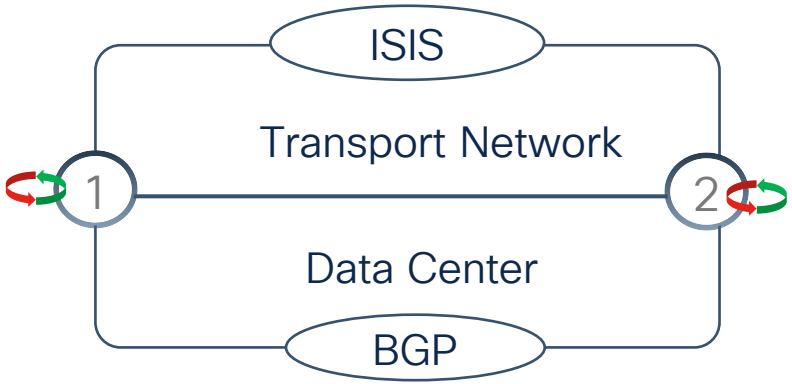
BGP Inter-AS: Underlay vs Overlay



© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public

Redistribution & Loop prevention

- A tagging solution is required, tag the prefix along with redistribution.
- Prefixes that have a ‘Domain’ tag set will be advertised Up or Down the hierarchy and will not be allowed to redistribute back to the originating Domain.
- Distribute-lists provides a strong mechanism to prevent the loops at interdomain routing boundaries.

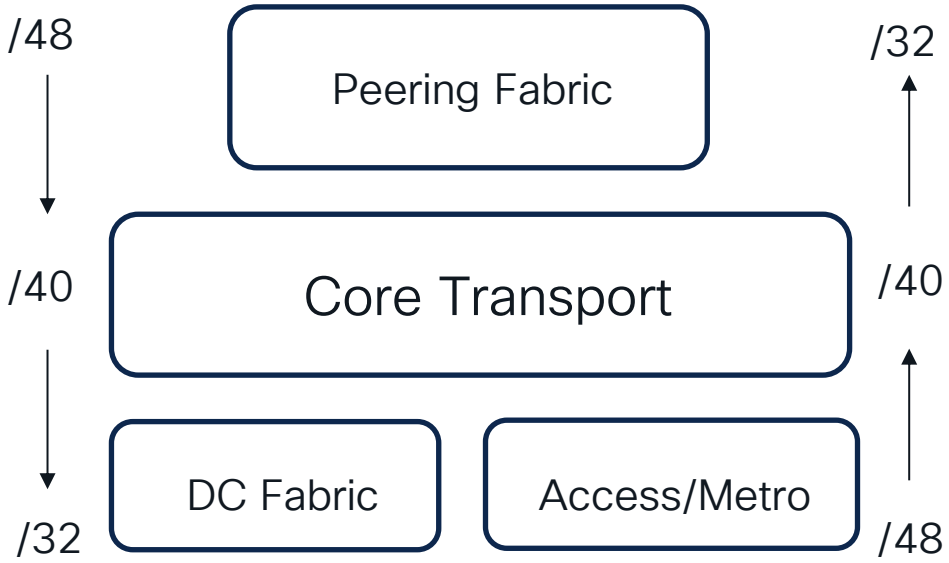


© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public

Inter Network Underlay Summarization

Simplified Inter-AS Underlay Routing

- Basic BGP prefix Aggregation
- Every service edge advertises /48 locator in BGP IPv6 underlay AFI
- Domain border aggregates the /48 of local domain to /40
- Only /40 is alllowed to propagate out to the Core
- ISIS summarization is done at ISIS domain border
- Can do ISIS<>BGP redistrubte for underlay handoff amongst DC and Core/Transport



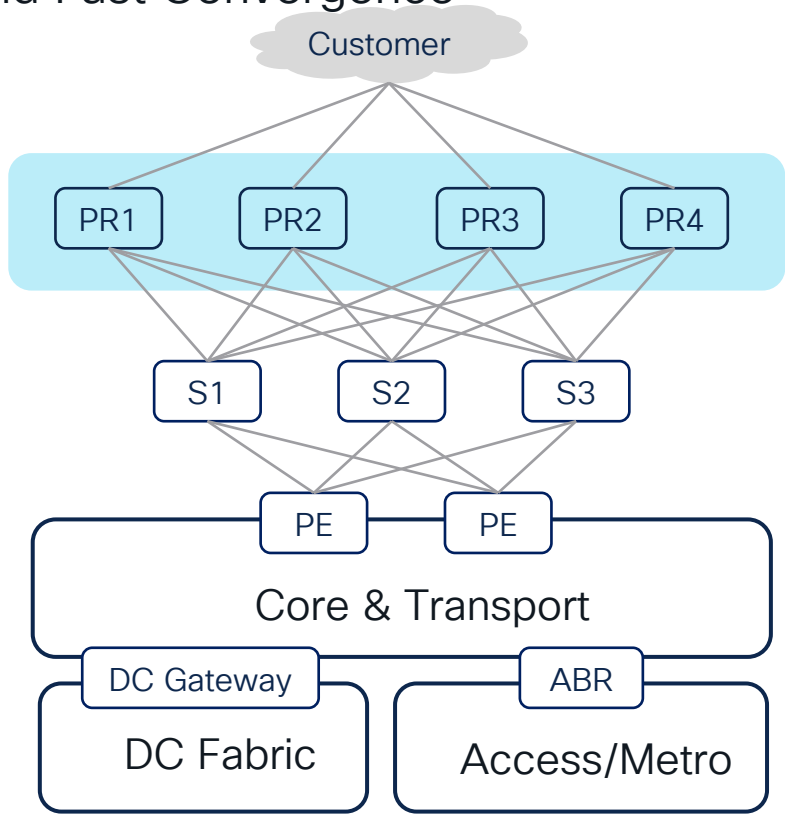
cisco Live!

© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public

Abstact Service Anchor and Termination

Inter-domain Load-Balancing, Resiliency and Fast Convergence

- Abstract Service Termination allows to offer the service connectivity via a cluster of Edges
- Static anycast SID can be used in realization of abstract peering in data-plane
- Greatly reduces the control-plane overhead without a compromise of traffic sharing amongst peer routers
- In the event of BGP peering failure, inter-network control-plane re-convergance be agnostic to data-plane forwarding.



© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public

BGP Static SID / Anycast SID

```
segment-routing
 srv6
!
   locator internet_anycast
    micro-segment behavior ucode psp-usd
    prefix fd00:201:b::/48
    anycast
!
```

```
!!!BGP Configuration: Explicit uDT46 for VRF!!!
router bgp 1
vrf interent
address-family ipv6 unicast
segment-routing srv6
    locator internet_anycast
    alloc mode per-vrf-46
!
```

```
segment-routing
  srv6
    static
      endpoint
        sid fd00:201:b:fff7:c1:: behavior end-udt46   <<< new Wlib support, new context type
          allocation-context vrf internet             <<< new context parameter
          forwarding                                  <<< new sub-mode
```


© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public

BGP Inter-AS Summary

- BGP Inter-AS Option C with unified transport solution in SRv6
- No VxLAN/MPLS Interworking
- Hyper Scale; Locator reachability can be simplified by:
  - BGP aggregation deployed at Multi-Network boundary
  - ISIS summarization deployed at IGP domain boundary
  - Hierarchical summarization is also possible
- Wide LIB and Anycast static SID provides agility and resiliency in network programmability
- Seamless Deployment in Brownfield
- End-to-End Transport Slicing
- More Secure and Reliability with more effective OPEX

© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public

cisco Live!

Keynote Deep Dives

Wednesday
10:30am –11:30am

Experiences Amplified: How AI Can Fuel Better Employee and Customer Experiences

Level 1
Room 106

Smart, Secure, Seamless: Transforming Experiences with Next-Generation Networking

Level 2
Room 204

Harness a Bold New Era: Transform Data Centre and Service Provider Connectivity

Level 2
Room 203

Securing User to Application and Everything in Between

Level 2
Melbourne Room 2

Unlocking Digital Resilience through Unified Observability

The HUB
Centre Stage

Complete Your Session Evaluations

Complete a minimum of 4 session surveys and the Overall Event Survey to claim a **Cisco Live T-Shirt**.

Complete your surveys in the **Cisco Live mobile app**.

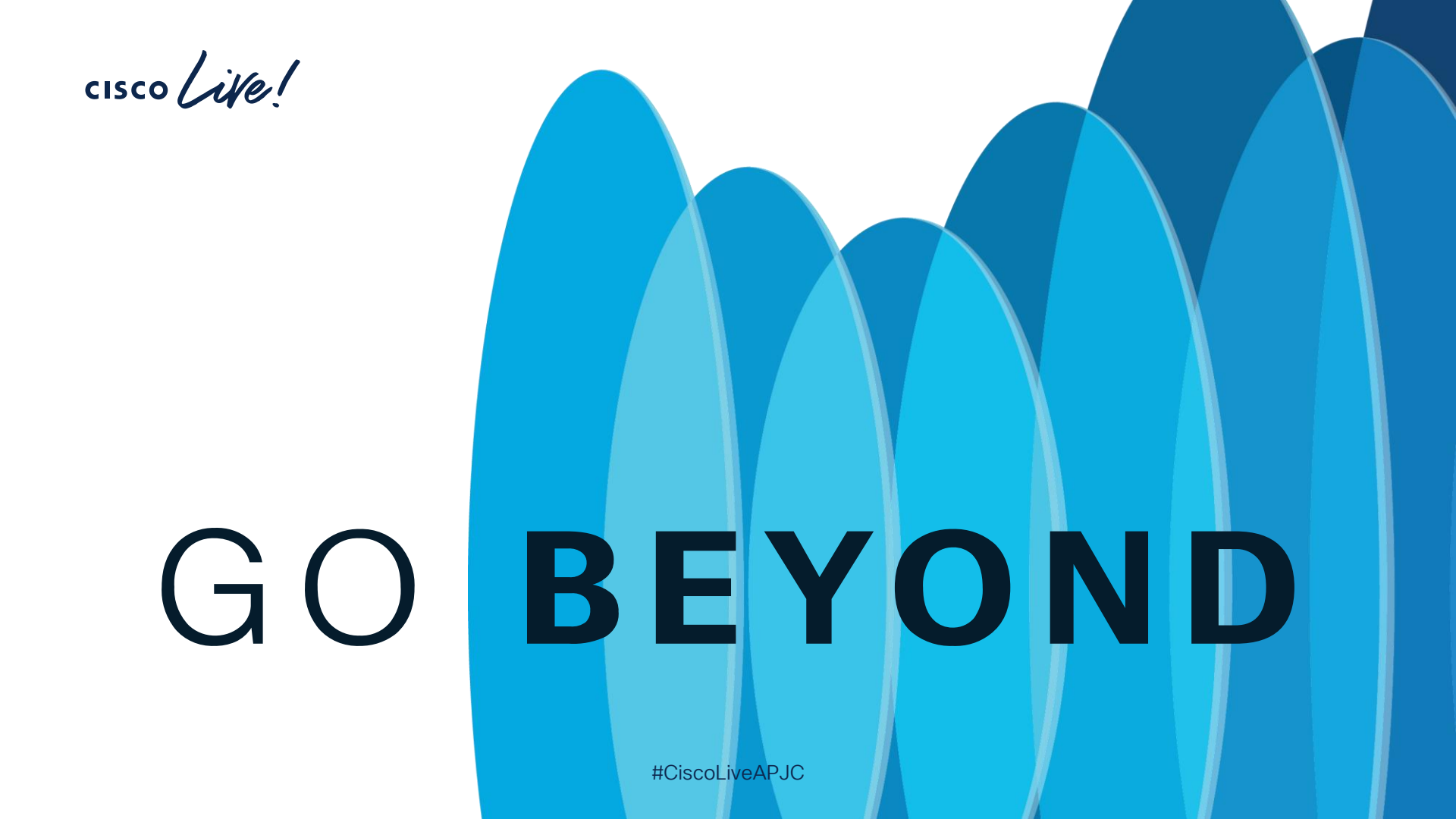
© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public

Continue your education

- Visit the Cisco Stand for related demos
- Book your one-on-one Meet the Expert meeting
- Attend the interactive education with DevNet, Capture the Flag, and Walk-in Labs
- Visit the On-Demand Library for more sessions at www.CiscoLive.com/on-demand

© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public

cisco
Thank you
cisco Live!
#CiscoLiveAPJC


cisco Live!
GO BEYOND
#CiscoLiveAPJC