cisco Live!

ALL IN

#CiscoLive

The bridge to possible

Kubernetes Infrastructure Connectivity for ACI

Network Designs for the Modern Data Centre

Domenico Dastoli
Principal Marketing Engineer - CNBU

BRKDCN-2411

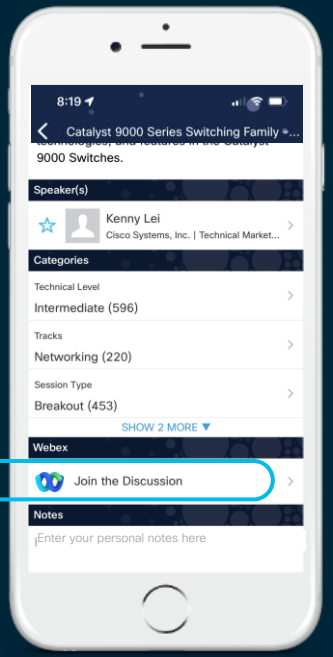
Cisco Webex App

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



https://ciscolive.ciscoevents.com/ciscolivebot/#BRKDCN-2411

Webex spaces will be moderated by the speaker until June 17, 2022.

Agenda

- About Kubernetes
- ACI-CNI Solution Overview
- ACI and Calico Solution Overview
- Which solution is right for me?
- Q&A

About Kubernetes

Business Landscape

Kubernetes footprint is growing exponentially every year, Gartner says that “By 2025, more than 85% of global organizations will be running containerized applications in production, which is a significant increase from fewer than 35% in 2019” and the application container market is expected to reach USD 8.2 billion by 2025.

Many organizations are moving their applications from virtual machine to container platforms. K8s is the workload orchestrator of choice when it comes to containers.

Cisco is committed to improve the value of our network fabrics through integrations like the ones we’re going to be talking today.

ACI and K8s

Looking at the existing K8s deployment figures and ACI's success on the market, we know many customers are running K8s on top of ACI today.

So what other options are available?

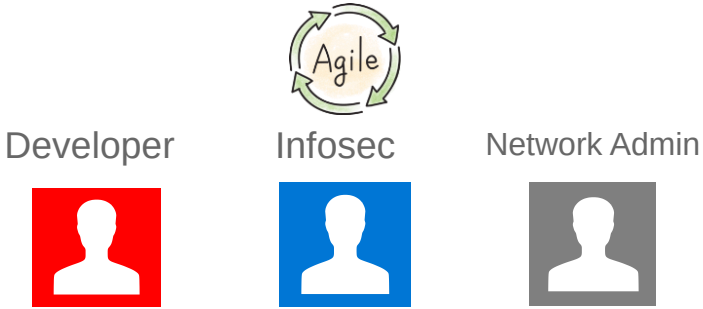
ACI and Kubernetes – Use an Overlay?

K8s is deployed as an **overlay** - i.e. using VXLAN, IPinIP… **but…** this leads to suboptimal solution.

ACI and Kubernetes – Use an Overlay?

K8s is deployed as an overlay - i.e. using VXLAN, IPinIP… but… this leads to suboptimal solution.

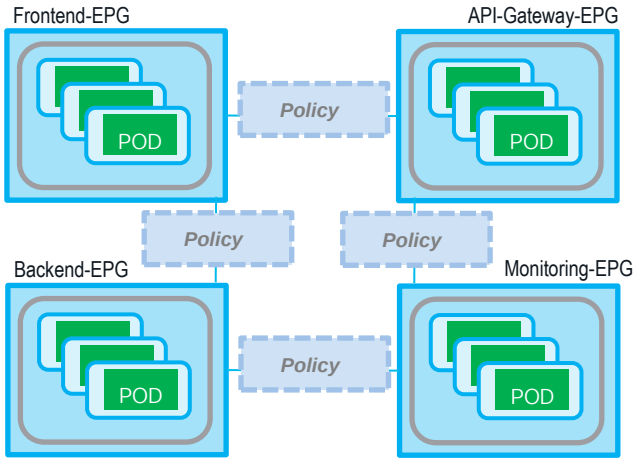
- Hides the visibility of the K8s pods to the network admin
- Add skills gap between network and Kubernetes admin
- Visibility and governance of network policies



ACI and Kubernetes – Use an Overlay?

K8s is deployed as an overlay - i.e. using VXLAN, IPinIP… but… this leads to suboptimal solution.

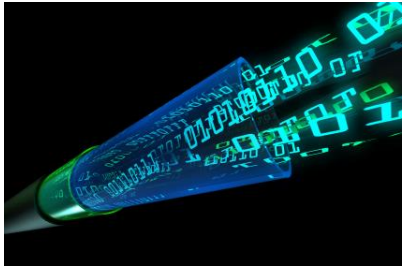
- Isolation for **kube-system** and other infrastructure related objects
- Isolation between **namespaces**
- Isolation between k8s cluster and **other workloads**



ACI and Kubernetes – Use an Overlay?

K8s is deployed as an overlay - i.e. using VXLAN, IPinIP… but… this leads to suboptimal solution.

- In Overlay mode, a (typically virtual) gateway needs to translate encapsulations
- This increases performance overhead, cost and complexity in interconnecting container-based workloads with external non-containerized workloads.



ACI and Kubernetes – what option?

K8s is deployed as an overlay - i.e. using VXLAN, IPinIP… but… this leads to suboptimal solution.

ACI can provide the best value when K8s is deployed with ACI in one of the following mode:

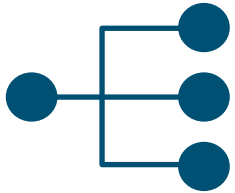
ACI CNI: ACI is the K8s overlay. ACI CNI	K8s in non overlay mode: K8s cluster connects to ACI in routed mode.

ACI-CNI
Solution Overview

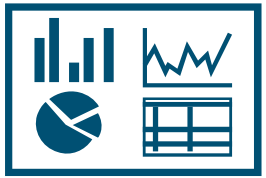
Why ACI-CNI for Application Container Platforms

Turnkey solution for node and container connectivity

Flexible policy: Native platform policy API and ACI policies



Hardware-accelerated: Integrated load balancing and Source NAT



Visibility: Live statistics in APIC per container and health metrics

Enhanced Multitenancy and unified networking for containers, VMs, bare metal

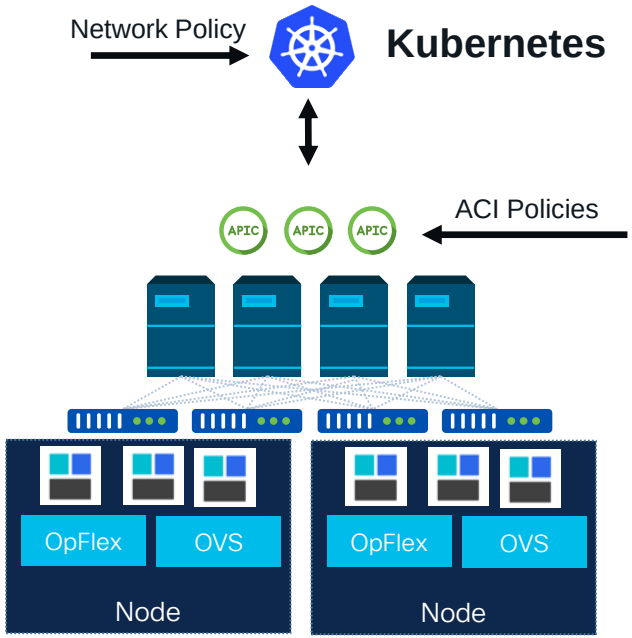
Fast, easy, secure and scalable **networking** for your Application Container Platform

Cisco ACI CNI Plugin Benefits

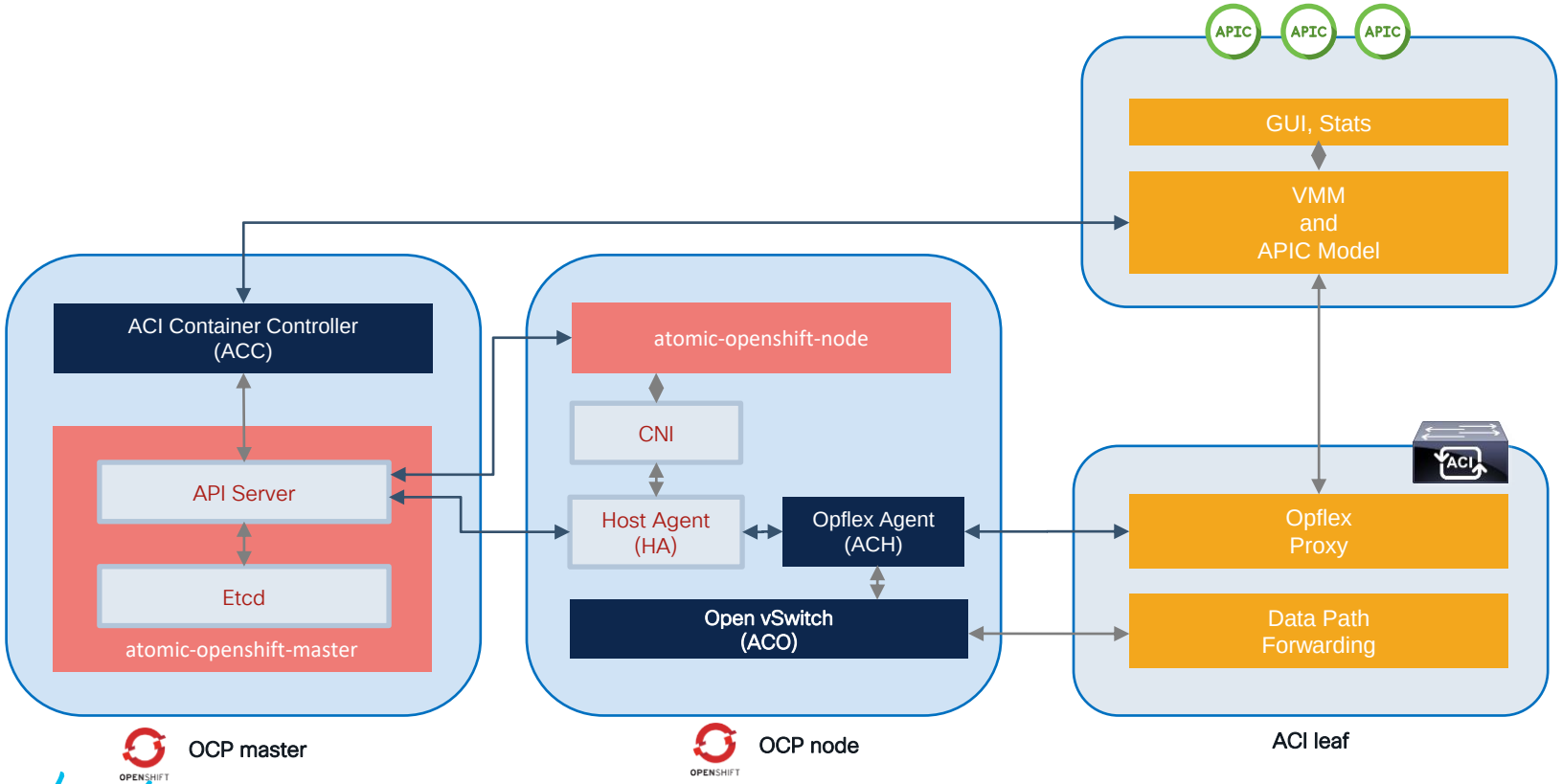
1. **Simplified Operations and Enhanced visibility**
2. **Granular security**: security can be implemented by using native NetworkPolicy or by using ACI EPGs and contracts, or both models complementing each other.
3. **Unified networking**: Pod and Service endpoints become first class citizens at the same level as Bare Metal or Virtual Machines.
4. **High performance**: low-latency secure connectivity without egress routers
5. **Hardware-assisted load balancing**: ingress connections to LoadBalancer-type services using ACI's Policy Based Redirect technology

Cisco ACI CNI plugin features

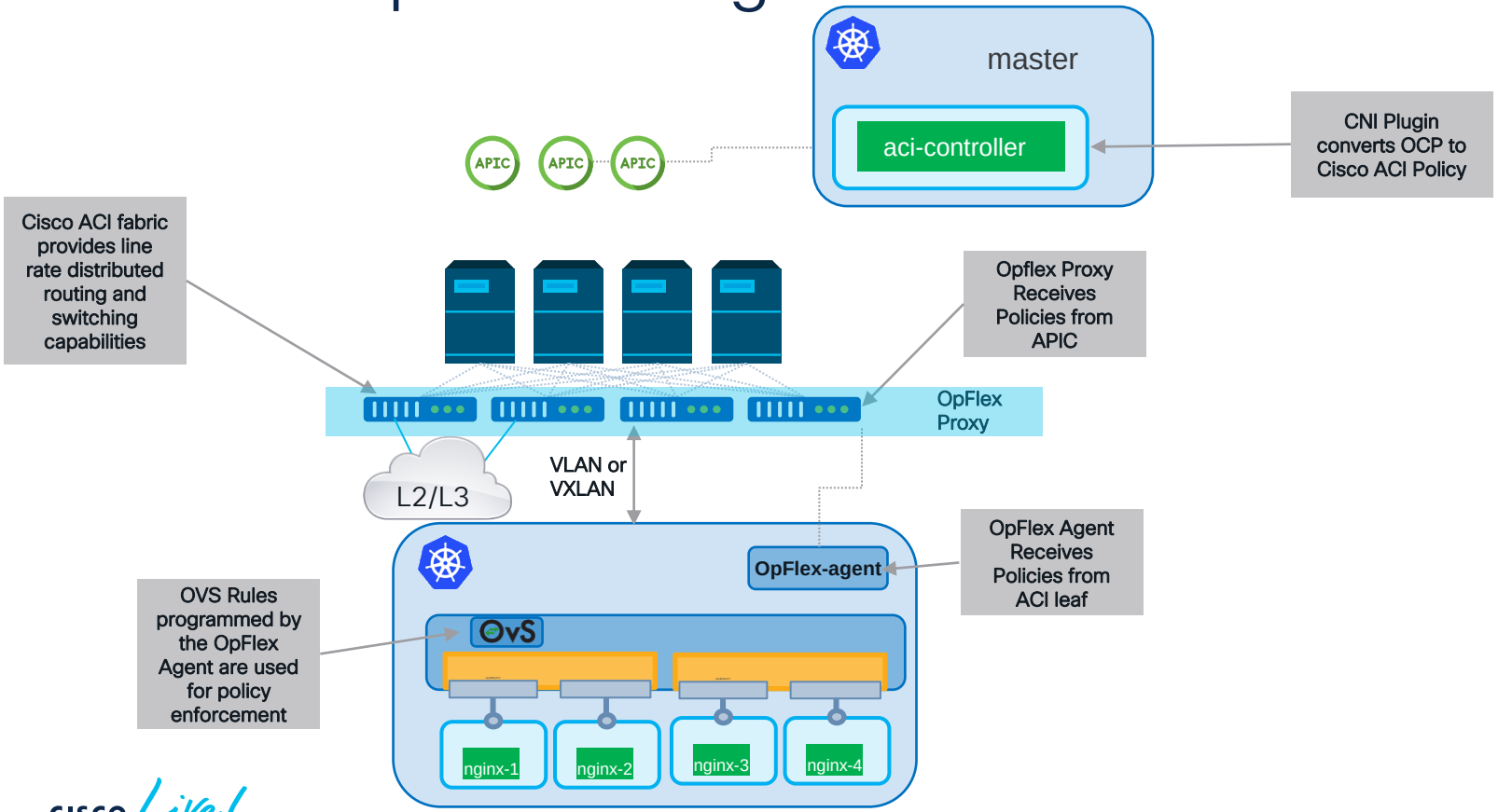
- IP Address Management for Pods and Services
- Distributed Routing and Switching with integrated VXLAN overlays implemented fabric wide and on Open vSwitch
- Distributed Firewall for implementing Network Policies
- EPG-level segmentation for K8s objects using annotations
- Consolidated visibility of K8s networking via VMM Integration



ACI CNI Plugin Components – Overview

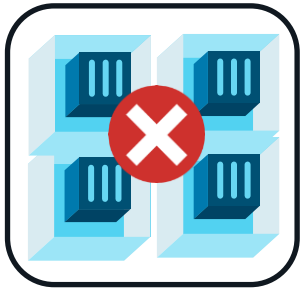


OVS rules provisioning



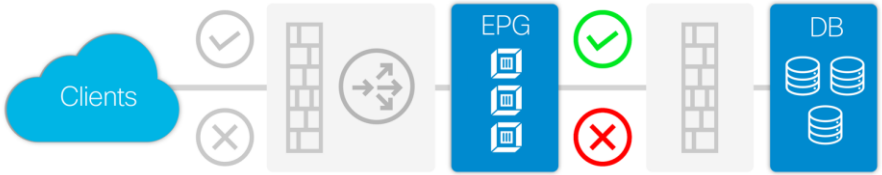
Dual level Policy Enforcement by ACI

Both Kubernetes Network Policy and ACI Contracts are enforced in the Linux kernel of every server node that containers run on.



Native API Default deny all traffic

```
apiVersion: networking.k8s.io/v1
kind: NetworkPolicy
metadata:
  name: default-deny
spec: podSelector: {}
policyTypes:
- Ingress
- Egress
```



Both policy mechanisms can be used in conjunction.

Containers are mapped to EPGs and contracts between EPGs are also enforced on all switches in the fabric where applicable.

Also ISTIO is supported!

Annotation of Project/Deployment

```
[root@dom-master1 CLDEMO]# oc describe project/ciscolive | grep Annotations -A 6
Annotations:            openshift.io/description=
                        openshift.io/sa.scc.uid-range=1000430000/10000
                        opflex.cisco.com/endpoint-group={"tenant":"openshiftcl","app-profile":"annotated","name":"ciscolive"}
[root@dom-master1 CLDEMO]#
```

APIC (Bru-Site1)

admin

System | Tenants | Fabric | Virtual Networking | L4-L7 Services | Admin | Operations | Apps | Integrations

ALL TENANTS | Add Tenant | Tenant Search: name or descr | common | openshiftcl | infra | mgmt | vpod

openshiftcl

- Quick Start
- openshiftcl
  - Application Profiles
    - annotated
      - Application EPGs
        - ciscolive
      - uSeg EPGs
    - kubernetes
  - Networking
  - Contracts
  - Policies
  - Services

EPG - ciscolive

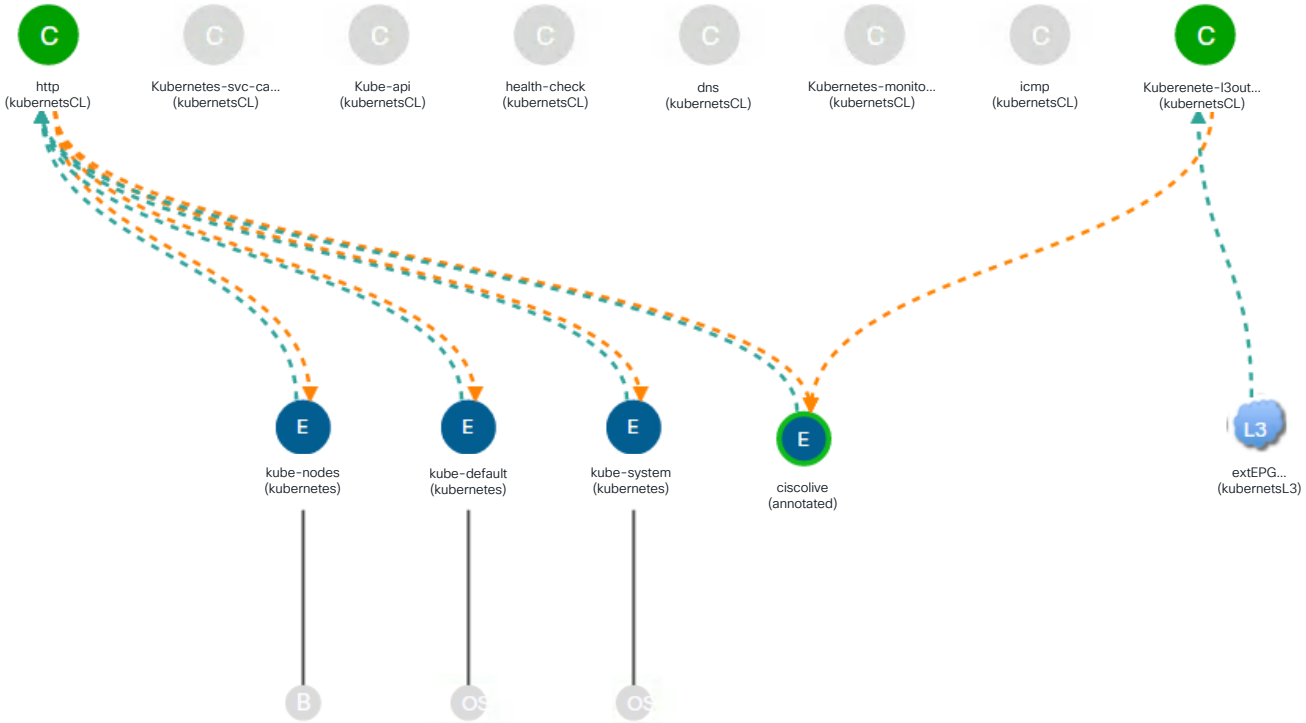
Summary | Policy | Operational | Stats | Health | Faults | History

Client End-Points | Configured Access Policies | Contracts | Controller End-Points | Deployed Leaves | Learned End-Points

100

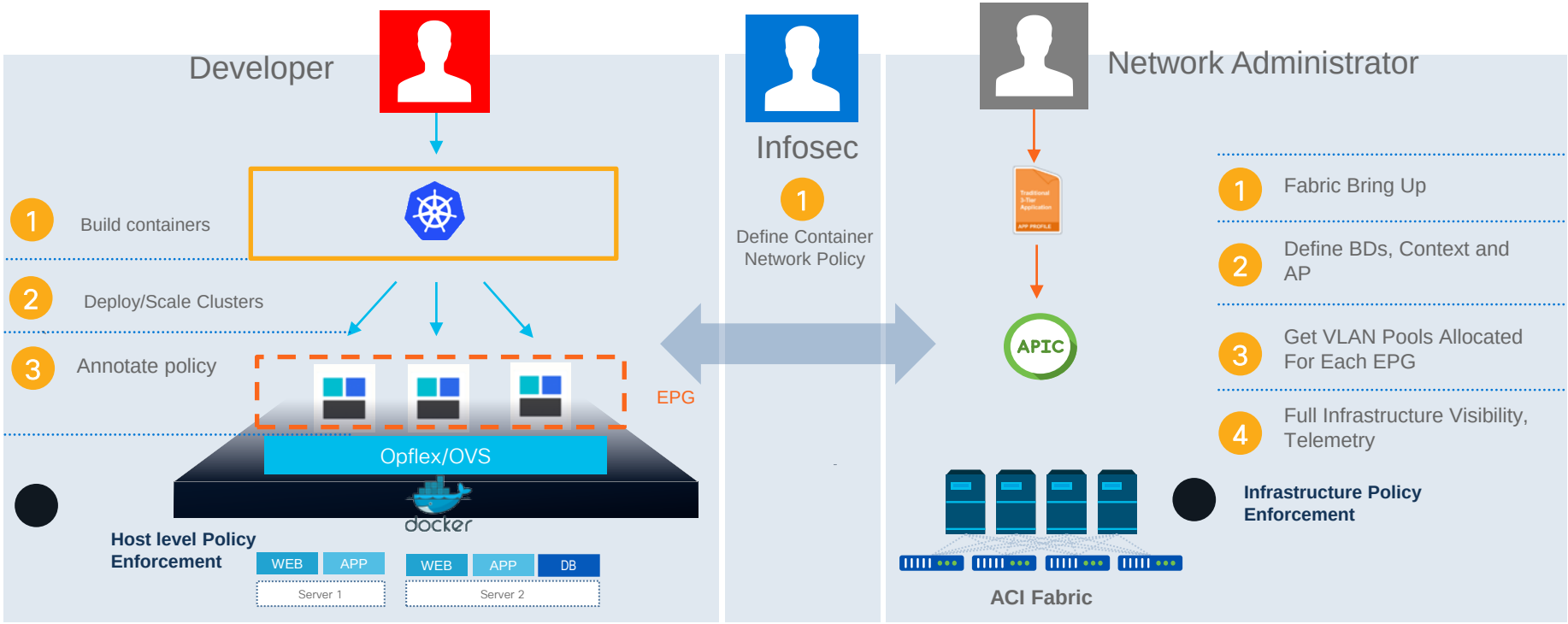
End Point	MAC	IP	Learning Source	Hosting Server	Reporting Controller Name	Interface	Multicast Address	Encap
hello-ciscolive-57497d7...	12:7E:97:35:BE:E3	11.12.0.78	learned vmm	dom-node2.domlab.cis...	openshiftcl	Pod-1/Node-101/eth1/17 (v... Pod-1/Node-101/tunnel21 (...	226.3...	vxlan-7634950
hello-ciscolive-57497d7...	26:7B:6C:3A:C0:AB	11.12.0.62	learned vmm	dom-node1.domlab.cis...	openshiftcl	Pod-1/Node-102/eth1/23 (v... Pod-1/Node-102/tunnel22 (...	226.3...	vxlan-7634950
hello-ciscolive-57497d7...	46:1E:06:32:85:09	11.12.0.76	learned vmm	dom-node2.domlab.cis...	openshiftcl	Pod-1/Node-101/eth1/17 (v... Pod-1/Node-101/tunnel21 (...	226.3...	vxlan-7634950
hello-ciscolive-57497d7...	96:F4:64:F0:C9:E2	11.12.0.61	learned vmm	dom-node1.domlab.cis...	openshiftcl	Pod-1/Node-102/eth1/23 (v... Pod-1/Node-102/tunnel22 (...	226.3...	vxlan-7634950
hello-ciscolive-57497d7...	B6:E9:0D:9A:63:18	11.12.0.77	learned vmm	dom-node2.domlab.cis...	openshiftcl	Pod-1/Node-101/eth1/17 (v... Pod-1/Node-101/tunnel21 (...	226.3...	vxlan-7634950

Segmentation: EPG to connect other resources



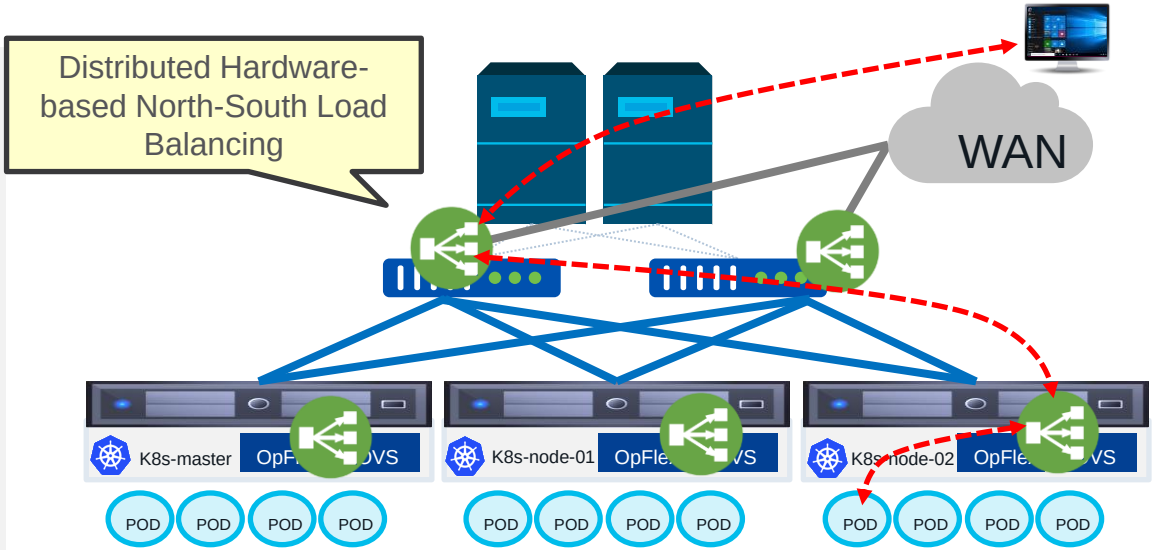
ACI Network Plugin for OpenShift

Native Security Policy Support



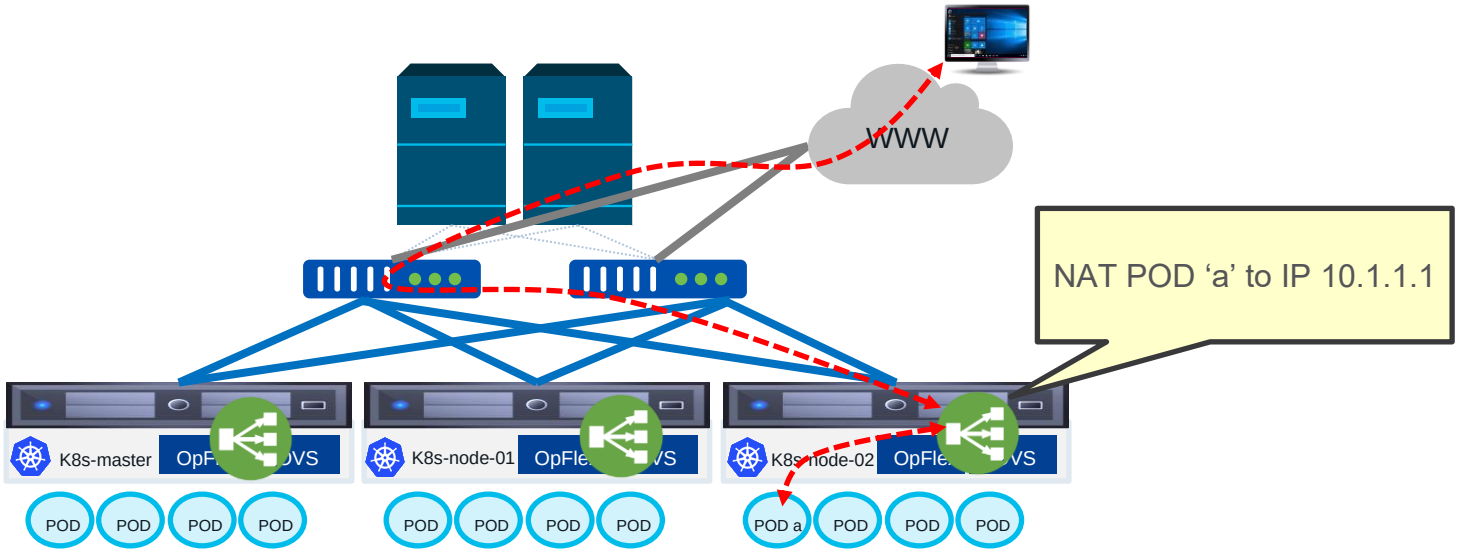
Kubernetes LoadBalancer Service with ACI


```
---
apiVersion: v1
kind: Service
metadata:
  labels:
    app: nginx
  name: nginx
  namespace: default
spec:
  ports:
  - name: 80-tcp
    port: 80
    protocol: TCP
    targetPort: 80
  selector:
    app: nginx
  type: LoadBalancer
```



ACI CNI distributed SNAT

- POD Initiated traffic can be NAT'ed to an IP address selected by the user



ACI and Calico Solution Overview

Cisco and Calico CNI

It is a lighter-touch integration between ACI and K8s while still ensuring ACI provides tangible value for customers. For example, using Calico in BGP mode increases network performance by taking advantage of the built-in routing capabilities of ACI and reducing cost and complexity by eliminating the need for virtual gateways.

Cisco and Calico CNI solution relies on BGP as a dynamic routing protocol and relies on an industry standard K8s network plugin. Cisco worked on a joint design guide with Tigera, the company behind Calico.

Calico CNI Modes

- Calico supports two main network modes: direct container routing (no overlay transport protocol) or network overlay using VXLAN or IPinIP (default) encapsulations to exchange traffic between workloads.
  - Overlay network means the underlying physical network is not aware of the workloads' IP addresses.
  - Direct routing means the underlying network is aware of the IP addresses used by workloads.

Overlay Network CNI mode

- In overlay mode, the physical network only needs to provide IP connectivity between K8s nodes while container to container communications are handled by the Calico network plugin directly.
- This is not recommended approach from Calico.
- Overlay mode increases performance overhead, cost and complexity in interconnecting container-based workloads with external non-containerized workloads.
- Overlay mode hides the visibility of the K8s cluster pod enpoints to the network administrator.

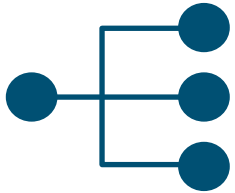
Direct Routing CNI mode

- This is the preferred Calico mode of deployment when running on-premises. (https://docs.projectcalico.org/networking/vxlan-ipip)
- Direct Routing provides the best visibility at the underlay because the network layer is aware of the node and pod endpoints and can provide direct routing capability to any other endpoint (BM/VM/others) attached to the fabric.

Why ACI and Calico for Application Container Platforms

Single CNI plugin for heterogenous fabrics (ACI, NX-OS etc…)

Network Policies for PODs
ACI Policies for K8s Nodes



ECMP load balancing

Visibility: Limited to routing – can improve with 3rd party tools

unified networking for containers, VMs, bare metal

Complex initial configuration, fast, secure and scalable **networking** for your Application Container Platform

Calico plugin main features

- IP Address Management for Pods and Services
- Standard Linux, Windows or eBPF Dataplane
- Calico's network policy model (superset of K8s)
- Advanced IP Address Management
- Opensource or Enterprise (additional features and support):
  - See https://www.tigera.io/tigera-products/compare-products/

Cisco ACI Calico Integration Benefits

1. **Relies on well established protocols (BGP)**
2. **Unified networking**: Node, Pod and Service endpoints are accessible from an L3OUT providing easy connectivity across and outside the fabric
3. **(Limited) ACI Security:** ability to use external EPG classification to secure communications to Node/Pod/Service Subnets (no /32 granularity)
4. **High performance**: low-latency connectivity without egress routers if no Overlay are used (BGP required)
5. **Hardware-assisted load balancing**: ingress connections to Service IPs are load-balanced with ECMP (up to 64 paths)
6. **Any Hypervisor/Bare Metal:** allows to mix form factors together

ACI and Calico: Architecture and Configuration

cisco Live!

Physical Connectivity

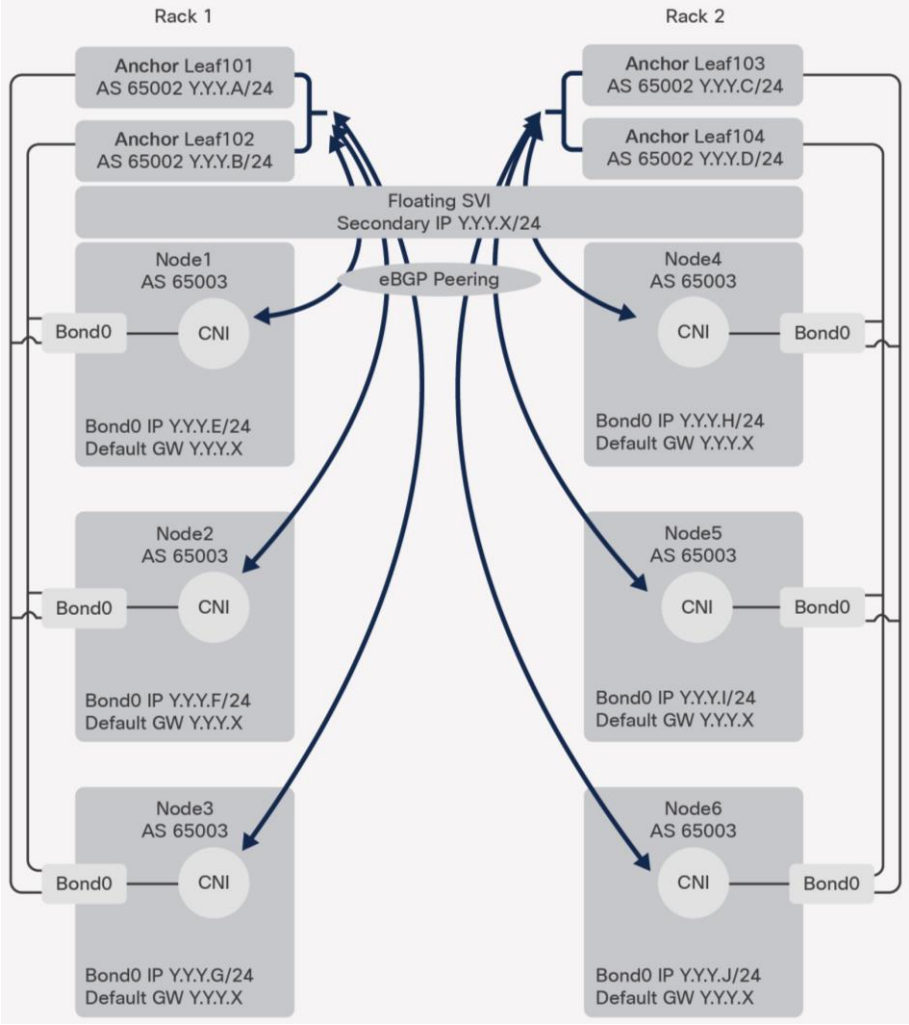
- K8s Nodes are connected to an L3OUT via vPC
  - External EPGs can be used to classify the traffic coming from the cluster
- Option to choose between:
  - Standard SVI
  - Floating SVI (recommended)

Floating SVI vs Standard SVI as of ACI 5.2.4

	Floating SVI	Standard SVI
Max cluster rack span	3 racks: 6 anchor nodes: with optimal traffic flows 19 racks: 6 anchor nodes + 32 non-anchor: with suboptimal traffic flows for the nodes connected to the non-anchor nodes	6: 12 Boarder Leaves with optimal traffic flows
Node subnets required	One subnet for the whole cluster	One /29 Subnet Per Node
Static paths binding	None; the binding is done at the physical domain level.	One per vPC
VM mobility	Yes; with suboptimal traffic flows if the VMs move to a different rack	No
Per fabric scale	200 floating SVI IPs (An IPv4/6 dual stack cluster uses two floating SVIs.)	Unlimited
L3Out per fabric	100	2400
ACI version	5.0 or newer	Any (This design was tested on 4.2 and 5.x code.)

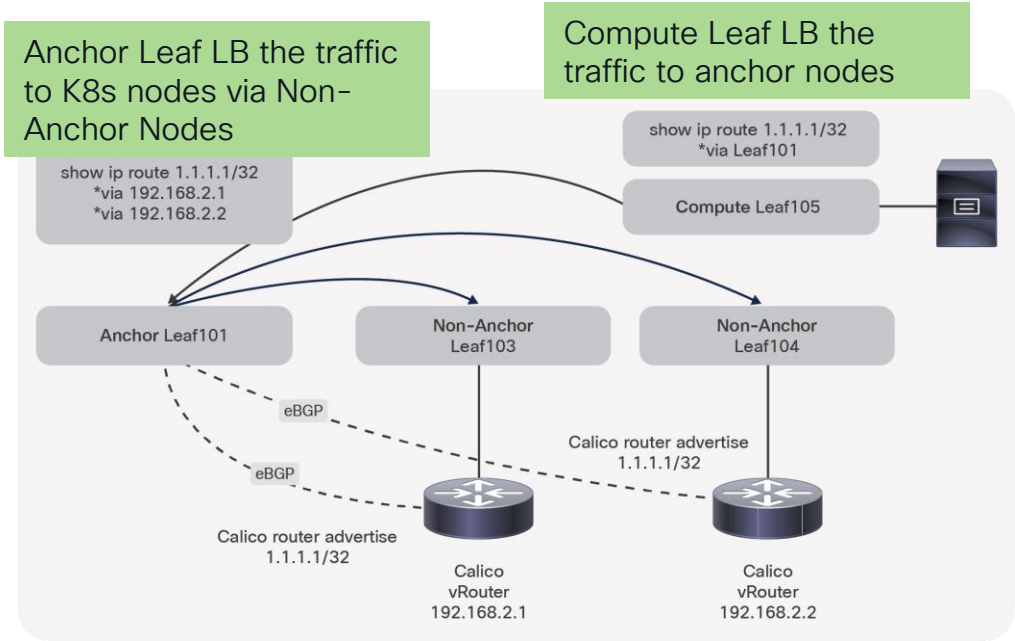
eBPG Architecture

- Each K8s Node will peer with a pair of border/acnchor leaves
- Single AS for the whole cluster
  - Simpler ACI config (can use a subnet for passive peering)
- Calico Advertise all the K8s subnets to ACI as well as host-routes for exposed services



Peer to local ToR if possible

- If K8s nodes are not physically connected to the anchor nodes suboptimal traffic flow is expected
- Future ACI releases will increase the number of Anchor Nodes per L3OUT



ACI eBGP Tuning

The following tunings are required:

- AS override and Disable Peer AS Check: To support having a single AS per cluster without the presence of Route Reflectors or Full Mesh inside the cluster
- BGP Graceful Restart
- BGP timers tuned to 1s/3s for quick eBGP node down detection
- AS path policy: allow installing more than one ECMP path for the same route
- Increase Max BGP ECMP path to 64

ACI eBGP Hardening (Optional)

- Enabled BGP password authentication
- Set the maximum AS limit to one
- Configure BGP import route control to accept only the expected subnets from the Kubernetes cluster:
  - Pod subnet(s)
  - Node subnet(s)
  - Service subnet(s)
- Set a limit on the number of received prefixes from the nodes.

Calico eBGP Config

The following Calico configurations objects are required

- One or more IPPool with all overlays disabled
- BGPConfiguration with:
  - nodeToNodeMeshEnabled set to “false”
  - List of serviceClusterIPs and serviceExternalIPs subnets to enabled host routes advertisement for those subnets
- Node: Used to set the node ipv4 source address for the eBGP peering and the AS Number for the node
- A Secret, Role and RoleBinding to pass the BGP Password to the Calico BGP Process

Calico eBGP Config – Cont.

- For a Calico Node to peer with an ACI Leaf we need to define a BGPPeer object. We want to ensure a K8s nodes peers only with 2 ACI Leaves to do that we can:
  - label the K8s nodes, i.e with the rack_id
  - use the rack_id label as a nodeSelector in the BGPPeer definition

```
apiVersion: projectcalico.org/v3
kind: BGPPeer
metadata:
  name: "203"
spec:
  peerIP: "192.168.2.203"
  asNumber: 65002
  nodeSelector: rack_id == "2"
```

Name of the Peer

IP of the Leaf

ACI BGP AS

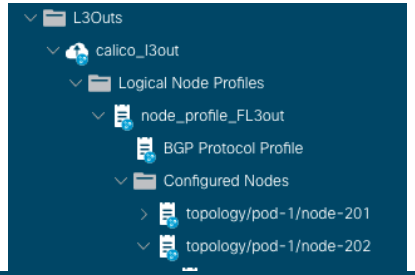
All the nodes where the Label rack_id is == "2" will peer with “203”

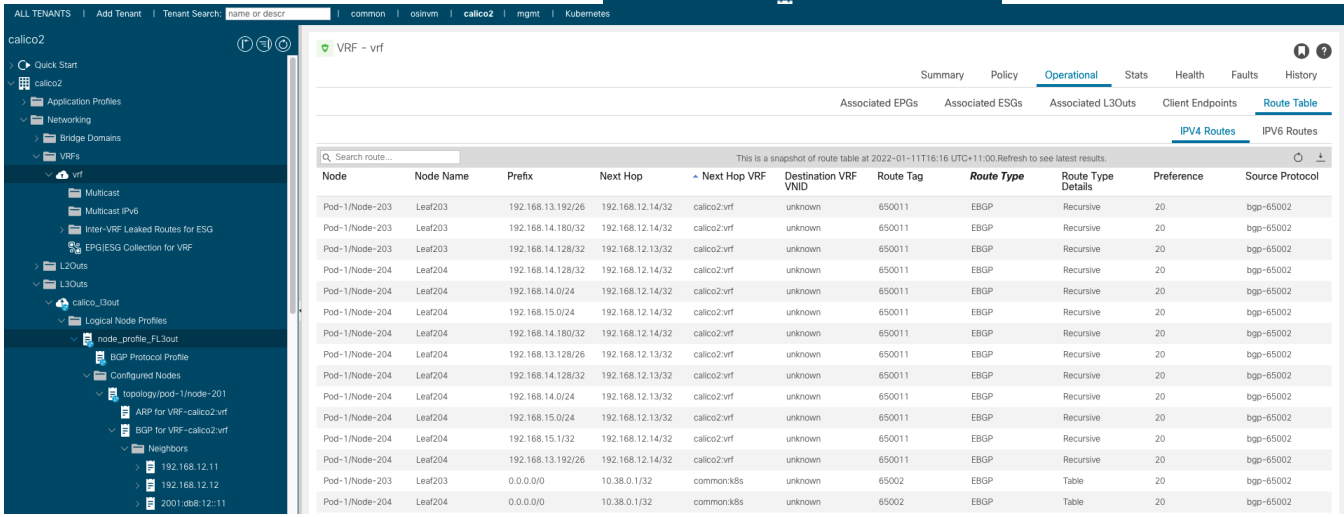
Routing Behaviour

- Nodes, pods and service IPs Subnets will be advertised to the ACI fabric
- Every calico nodes is allocated one or more /26 subnets from the POD Supernet. Each /26 is advertised to ACI as well
- Exposed Services will be advertised to ACI as host routes from every nodes that has a running POD associated to the service.

Visibility

- From the APIC UI is possible to see:
  - the eBGP Peers under the L3OUT
  - Received routes under the VRF

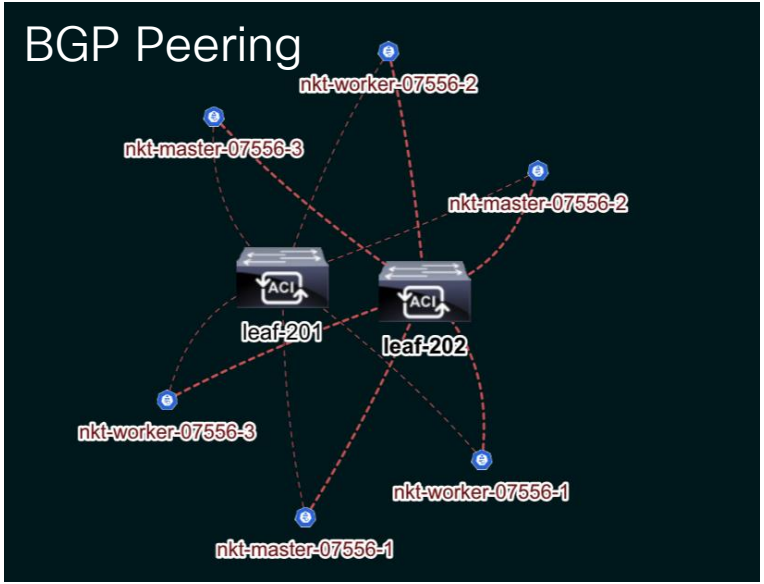




Node	Node Name	Prefix	Next Hop	Next Hop VRF	Destination VRF VNID	Route Tag	Route Type	Route Type Details	Preference	Source Protocol
Pod-1/Node-203	Leaf203	192.168.13.192/26	192.168.12.14/32	calico2:vrf	unknown	650011	EBGP	Recursive	20	bgp-65002
Pod-1/Node-203	Leaf203	192.168.14.180/32	192.168.12.14/32	calico2:vrf	unknown	650011	EBGP	Recursive	20	bgp-65002
Pod-1/Node-203	Leaf203	192.168.14.128/32	192.168.12.13/32	calico2:vrf	unknown	650011	EBGP	Recursive	20	bgp-65002
Pod-1/Node-204	Leaf204	192.168.14.128/32	192.168.12.14/32	calico2:vrf	unknown	650011	EBGP	Recursive	20	bgp-65002
Pod-1/Node-204	Leaf204	192.168.14.0/24	192.168.12.14/32	calico2:vrf	unknown	650011	EBGP	Recursive	20	bgp-65002
Pod-1/Node-204	Leaf204	192.168.15.0/24	192.168.12.14/32	calico2:vrf	unknown	650011	EBGP	Recursive	20	bgp-65002
Pod-1/Node-204	Leaf204	192.168.14.180/32	192.168.12.14/32	calico2:vrf	unknown	650011	EBGP	Recursive	20	bgp-65002
Pod-1/Node-204	Leaf204	192.168.13.128/26	192.168.12.13/32	calico2:vrf	unknown	650011	EBGP	Recursive	20	bgp-65002
Pod-1/Node-204	Leaf204	192.168.14.128/32	192.168.12.13/32	calico2:vrf	unknown	650011	EBGP	Recursive	20	bgp-65002
Pod-1/Node-204	Leaf204	192.168.14.0/24	192.168.12.13/32	calico2:vrf	unknown	650011	EBGP	Recursive	20	bgp-65002
Pod-1/Node-204	Leaf204	192.168.15.0/24	192.168.12.13/32	calico2:vrf	unknown	650011	EBGP	Recursive	20	bgp-65002
Pod-1/Node-204	Leaf204	192.168.15.1/32	192.168.12.14/32	calico2:vrf	unknown	650011	EBGP	Recursive	20	bgp-65002
Pod-1/Node-204	Leaf204	192.168.13.192/26	192.168.12.14/32	calico2:vrf	unknown	650011	EBGP	Recursive	20	bgp-65002
Pod-1/Node-203	Leaf203	0.0.0.0/0	10.38.0.1/32	common:k8s	unknown	65002	EBGP	Table	20	bgp-65002
Pod-1/Node-204	Leaf204	0.0.0.0/0	10.38.0.1/32	common:k8s	unknown	65002	EBGP	Table	20	bgp-65002

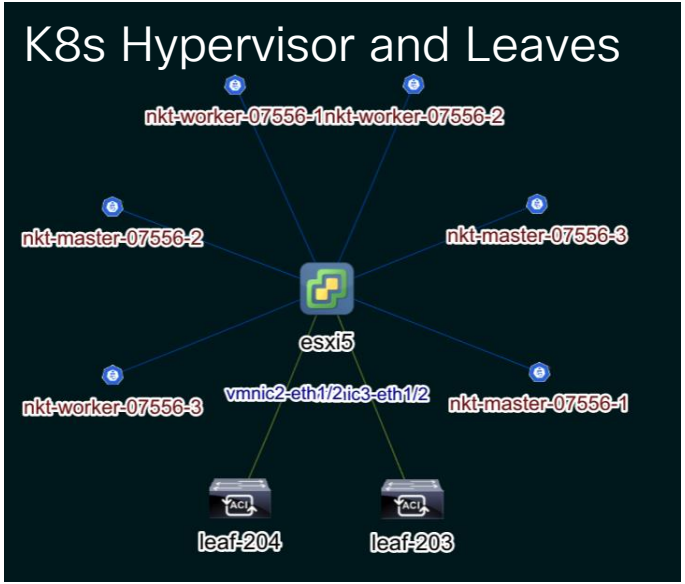
Visibility with homegrown tools

- By querying the APIC and K8s API is possible to obtain deep visibility
- For example



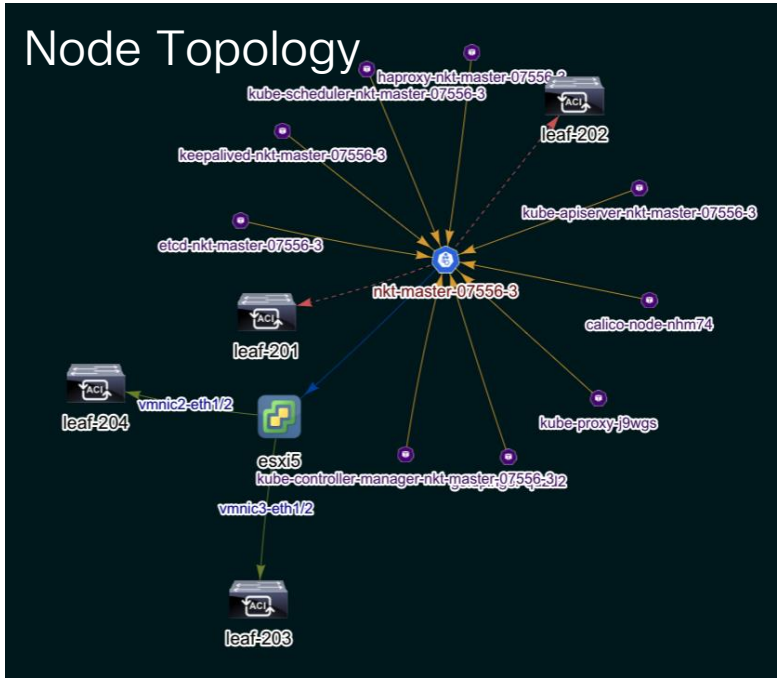
Visibility with homegrown tools

- By querying the APIC and K8s API is possible to obtain deep visibility
- For example



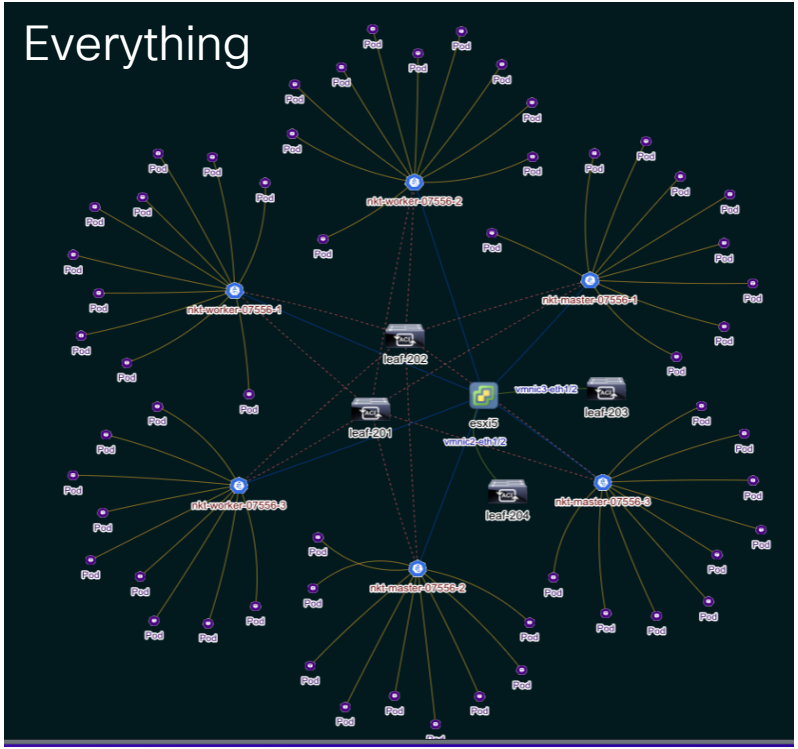
Visibility with homegrown tools

- By querying the APIC and K8s API is possible to obtain deep visibility
- For example



Visibility with homegrown tools

- By querying the APIC and K8s API is possible to obtain deep visibility
- For example



Installation Tips

- The ACI and Calico configuration requires manual configuration that can be automated.
- Open-Source tool to create **LABS:**
  - https://github.com/camrossi/akb
- This tool will:
  - Use a wizard to gather all the required infos from the user
  - Configure ACI
  - Deploy and Bootstrap a K8s cluster on ESXi

  OR
  - Generate the Calico manifest for a pre-existing cluster

Which solution is right for me?

Which solution is right for me?

- This is an hard question, both CNI plugins are enterprise grade and provide a rich feature set
- Some question you might ask yourself:
  - Is running the same CNI plugin on ANY network infrastructure important?
  - Is the K8s team already using Calico ?
  - Is the ability to map Namespaces/Deployments to EPG and use ACI contract an important security feature?
  - Is having a single vendor for the networking stack support important ?

Comparison

ACI CNI

- Only ACI
- Open Source and Free
- TAC Support
- Minimal/Automated ACI config
- Supports APIC Policy Model and K8s Network Policies
- PBR LoadBalancing for Services
- Support CoreOS and Ubuntu

Calico CNI

- Any fabrics
- Open Source and Free
- Pay for Technical Support
- Supports Calico and K8s Network Policies
- ECMP LoadBalancing for Services
- Support any Linux distro
- Data plane: Standard Linux, Windows or eBPF

Comparison Cont.

ACI CNI

- Visibility: Out of the box and in depth
- Installation: Automated
- Proprietary components and strong dependency between ACI and K8s

Calico CNI

- Visibility: 3rd party tools
- Installation: Manual
- No proprietary components and loosely coupling

References
Docs

References

- https://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/application-centric-infrastructure/white-paper-c11-743182.html
- https://www.cisco.com/c/en/us/td/docs/switches/datacenter/aci/apic/white_papers/Cisco-ACI-CNI-Plugin-for-OpenShift-Architecture-and-Design-Guide.html
- https://www.cisco.com/c/en/us/td/docs/switches/datacenter/aci/apic/sw/kb/b_Kubernetes_Integration_with_ACI.html
- https://www.cisco.com/c/dam/en/us/td/docs/Website/datacenter/aci/virtualization/matrix/virtmatrix.html

Technical Session Surveys

- Attendees who fill out a minimum of four session surveys and the overall event survey will get Cisco Live branded socks!
- Attendees will also earn 100 points in the Cisco Live Game for every survey completed.
- These points help you get on the leaderboard and increase your chances of winning daily and grand prizes.

Cisco Learning and Certifications

From technology training and team development to Cisco certifications and learning plans, let us help you empower your business and career. www.cisco.com/go/certs

Pay for Learning with Cisco Learning Credits

(CLCs) are prepaid training vouchers redeemed directly with Cisco.

Learn

Cisco U.
IT learning hub that guides teams and learners toward their goals

Cisco Digital Learning
Subscription-based product, technology, and certification training

Cisco Modeling Labs
Network simulation platform for design, testing, and troubleshooting

Cisco Learning Network
Resource community portal for certifications and learning

Train

Cisco Training Bootcamps
Intensive team & individual automation and technology training programs

Cisco Learning Partner Program
Authorized training partners supporting Cisco technology and career certifications

Cisco Instructor-led and Virtual Instructor-led training
Accelerated curriculum of product, technology, and certification courses

Certify

Cisco Certifications and Specialist Certifications
Award-winning certification program empowers students and IT Professionals to advance their technical careers

Cisco Guided Study Groups
180-day certification prep program with learning and support

Cisco Continuing Education Program
Recertification training options for Cisco certified individuals

Here at the event? Visit us at **The Learning and Certifications lounge at the World of Solutions**

Continue your education

- Visit the Cisco Showcase for related demos
- Book your one-on-one Meet the Engineer 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

Thank you

cisco Live!

ALL IN

#CiscoLive