CISCO Connect

Consistent Connectivity and Policy for AI Infrastructure with Isovalent Networking



Marino Wijay
Solutions Engineer – Isovalent @ Cisco

October 2026

Every GPU training job and inference request depends on the network

Fast, secure and observable, in every cloud and data center

© 2026 Cisco and/or its affiliates. All rights reserved.

Agenda

1. The AI challenge
2. The eBPF foundation
3. Consistent connectivity
4. Consistent policy
5. Visibility and troubleshooting
6. Better together with Cisco Nexus
7. Operating at scale
8. Questions

© 2026 Cisco and/or its affiliates. All rights reserved.

SECTION 01

The AI networking challenge

Scale, multi-tenancy and fragmentation raise the stakes

Every AI workload depends on the network

1

Training

Distributed jobs exchange data between nodes and GPUs continuously.

A network stall leaves expensive accelerators idle.

2

Inference

Latency-sensitive requests fan out across gateways, model servers and data services.

Many teams and tenants share the same clusters.

3

Agentic AI

Agents call tools, APIs and models across clusters and clouds.

Every new connection is something to secure and observe.

© 2026 Cisco and/or its affiliates. All rights reserved.

cisco

Complexity compounds as AI scales

1

Scale

Kubernetes environments now reach thousands of nodes. Networking must scale without becoming the bottleneck.

2

Multi-tenancy

Teams and customers share GPU clusters. Isolation has to be strong and easy to prove.

3

Fragmentation

Each cloud and data center brings its own networking and policy model.

4

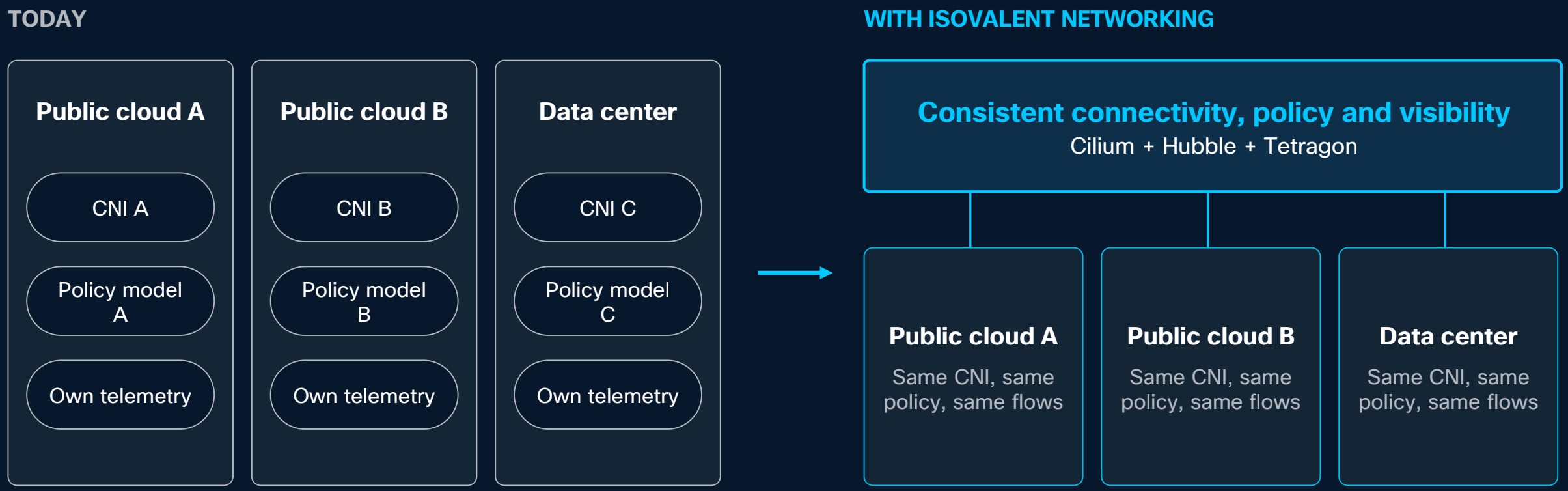
Tool sprawl

Separate tools for CNI, encryption, service mesh, observability and runtime security.

© 2026 Cisco and/or its affiliates. All rights reserved.

cisco

One platform layer, not per-environment networking



© 2026 Cisco and/or its affiliates. All rights reserved.

SECTION 02

The foundation: Cilium, eBPF and Isovalent

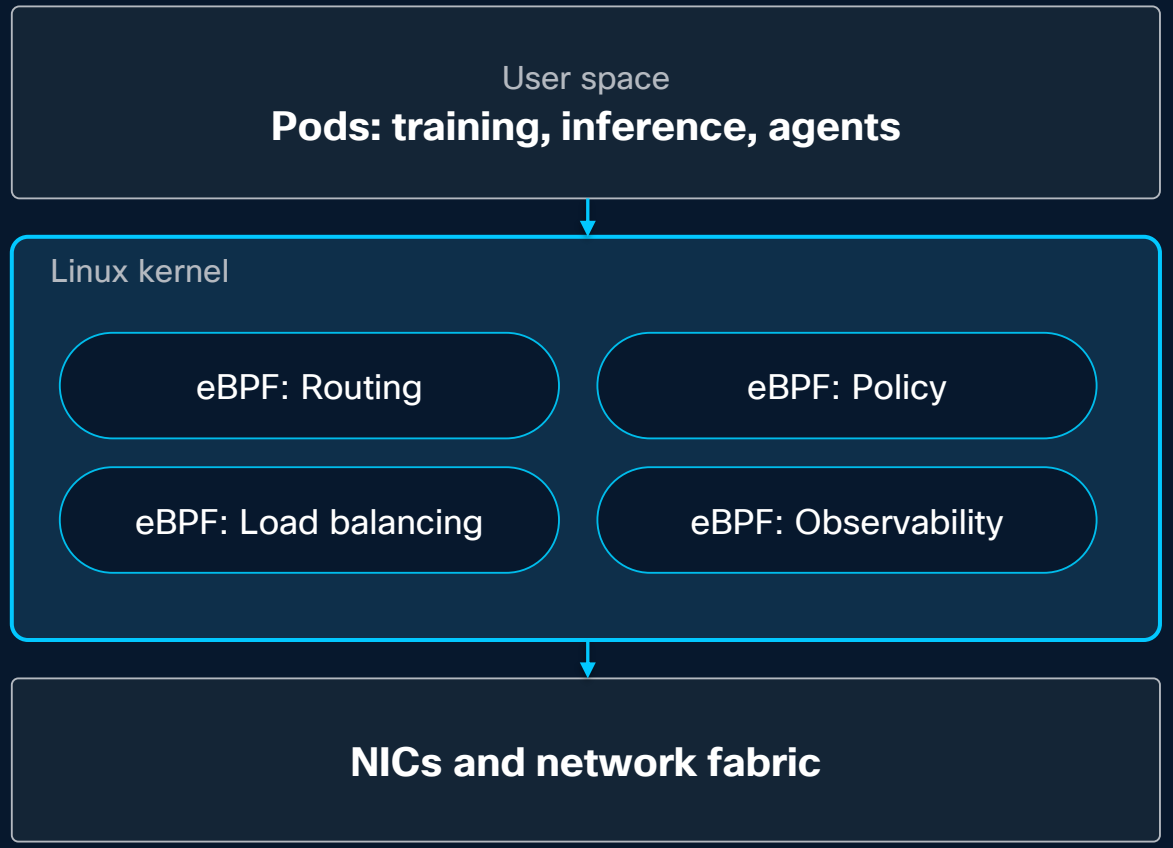
An open-source, kernel-native base with enterprise support

eBPF lets networking run inside the Linux kernel

1. **Kernel speed**
   Packet handling, load balancing and policy run where the packets are, not in extra proxies.

2. **Programmable**
   New networking functions load dynamically as eBPF programs.

3. **Safe by design**
   The kernel verifies every eBPF program before it runs.



© 2026 Cisco and/or its affiliates. All rights reserved.

Cilium is the foundation of cloud-native networking

Graduated

CNCF project, the only graduated project in cloud native networking

Major clouds

Used for Kubernetes offerings at Google, AWS and Microsoft Azure

Identity

Policy follows workload identity, not IP addresses

© 2026 Cisco and/or its affiliates. All rights reserved.

cisco

Isovalent Networking: connect, see and enforce

Cilium

Connect and protect

- CNI and IPAM for Kubernetes
- Load balancing without kube-proxy
- Identity-based L3 to L7 network policy
- Cluster Mesh and transparent encryption

Hubble

See

- Flow visibility with Kubernetes identity
- Service dependency maps
- DNS and HTTP metrics
- Historical flows for forensics

Tetragon

Enforce at runtime

- Process, file and network events
- eBPF-based, low overhead
- In-kernel enforcement, not just alerts
- Per-workload runtime policy

Isovalent Networking for Kubernetes and Virtualization

© 2026 Cisco and/or its affiliates. All rights reserved.

cisco

Where Isovalent Networking sits in the AI stack

AI workloads

- Training jobs
- Inference services
- Agents and tools

Isovalent Networking for Kubernetes

- Connectivity (CNI, IPAM, Cluster Mesh)
- Policy (identity, L3 to L7, encryption)
- Visibility and runtime security (Hubble, Tetragon)

Linux kernel with eBPF

- Kernel-speed datapath on every node

Underlay and infrastructure

- Cisco Nexus One fabric
- Cloud VPC networks
- GPUs and NICs

© 2026 Cisco and/or its affiliates. All rights reserved.

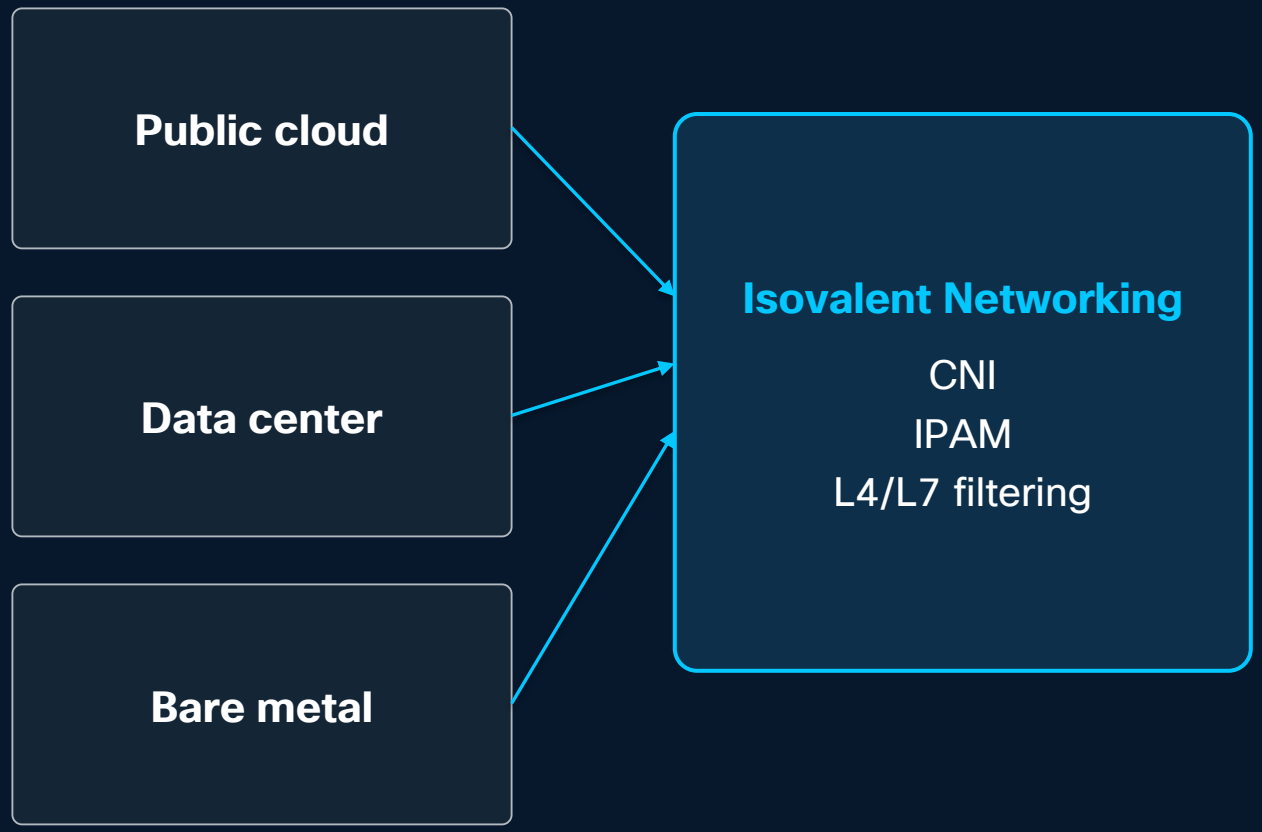
cisco

SECTION 03

Consistent connectivity

One networking layer across clouds, clusters and bare metal

One networking layer across every environment



IN PRACTICE: OPENAI

OpenAI adopted Isovalent Networking for Kubernetes as its default Kubernetes networking stack.

It gives OpenAI a common foundation for CNI, IPAM and L4/L7 filtering across cloud provider environments and bare metal.

Source: Isovalent, "OpenAI Uses Isovalent for a Common Networking for AI Infrastructure" (isovalent.com).

© 2026 Cisco and/or its affiliates. All rights reserved.

Networking at kernel speed keeps the GPU path clear

1. **eBPF datapath**
   Kernel-level networking performance that removes the overhead of traditional Linux networking.

2. **Efficient service routing**
   eBPF-based load balancing replaces kube-proxy and long iptables chains.

3. **XDP and BIG TCP**
   High-throughput, low-latency paths for demanding workloads.

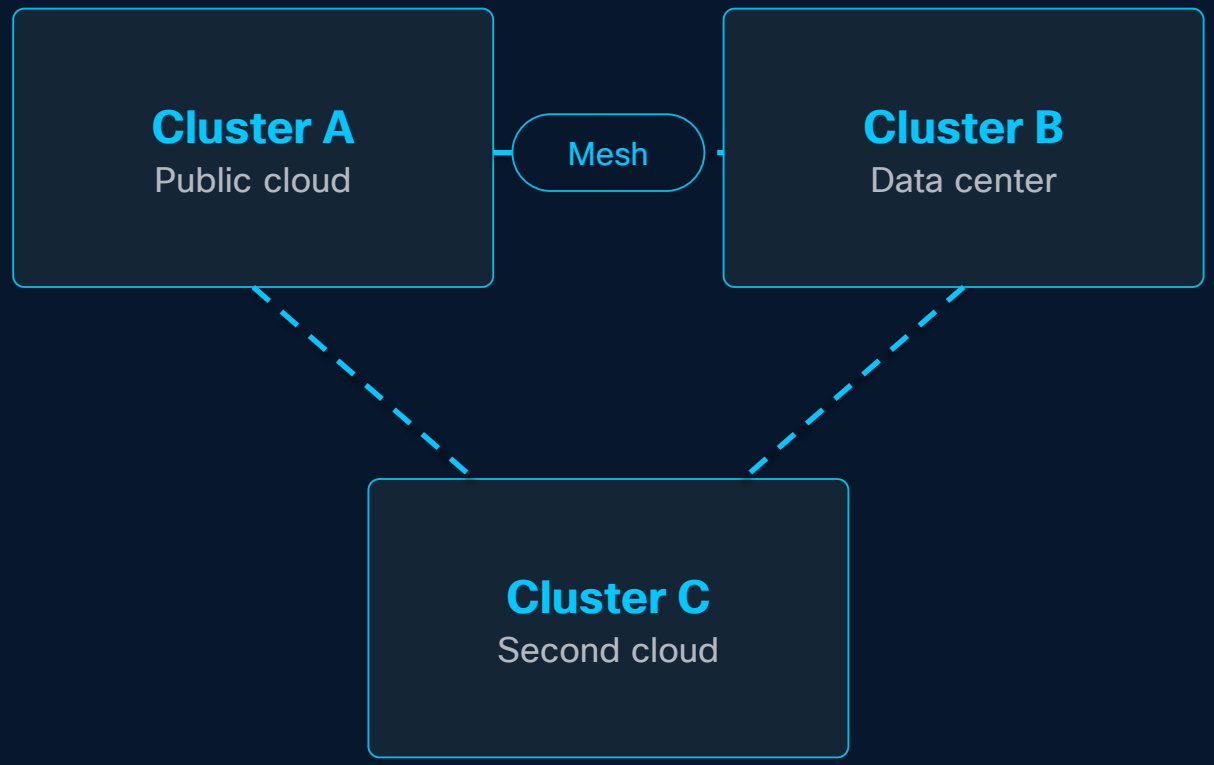
4. **Sidecarless by design**
   L7 policy and service mesh capabilities without a proxy in every pod.

Sources: cilium.io/industries/ai; Cisco Blogs, Feb 2026.

© 2026 Cisco and/or its affiliates. All rights reserved.

cisco

Cluster Mesh joins clusters into one network

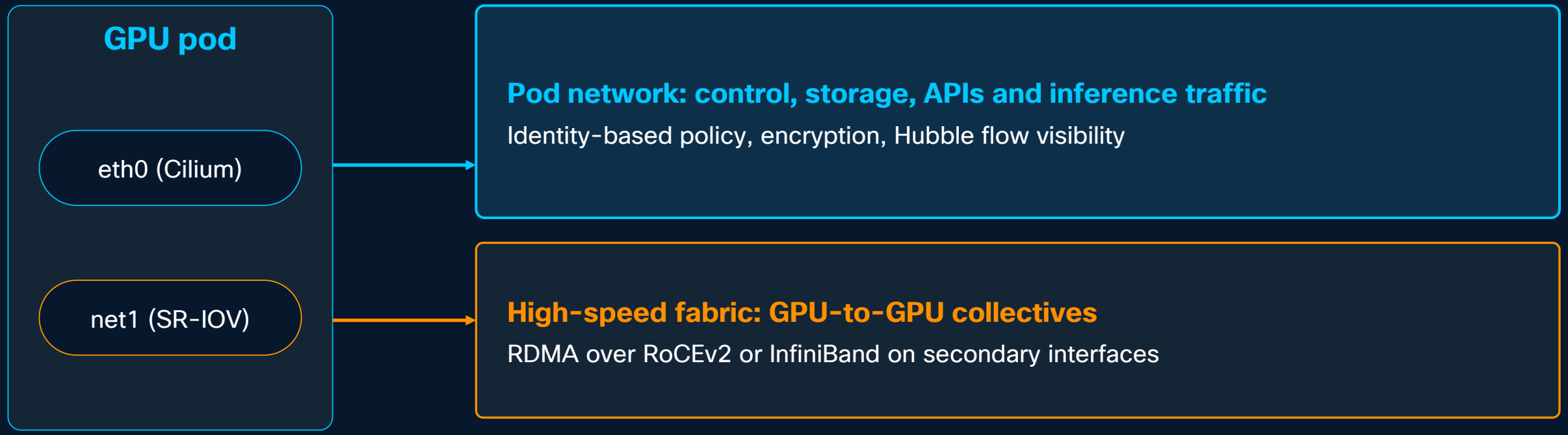


What you get

- Global services with load balancing and failover across clusters
- Pod-to-pod connectivity across clusters without gateways in the path
- Shared identity, so one policy applies in every cluster
- One place to see cross-cluster flows in Hubble

© 2026 Cisco and/or its affiliates. All rights reserved.

Pair the pod network with a dedicated GPU fabric

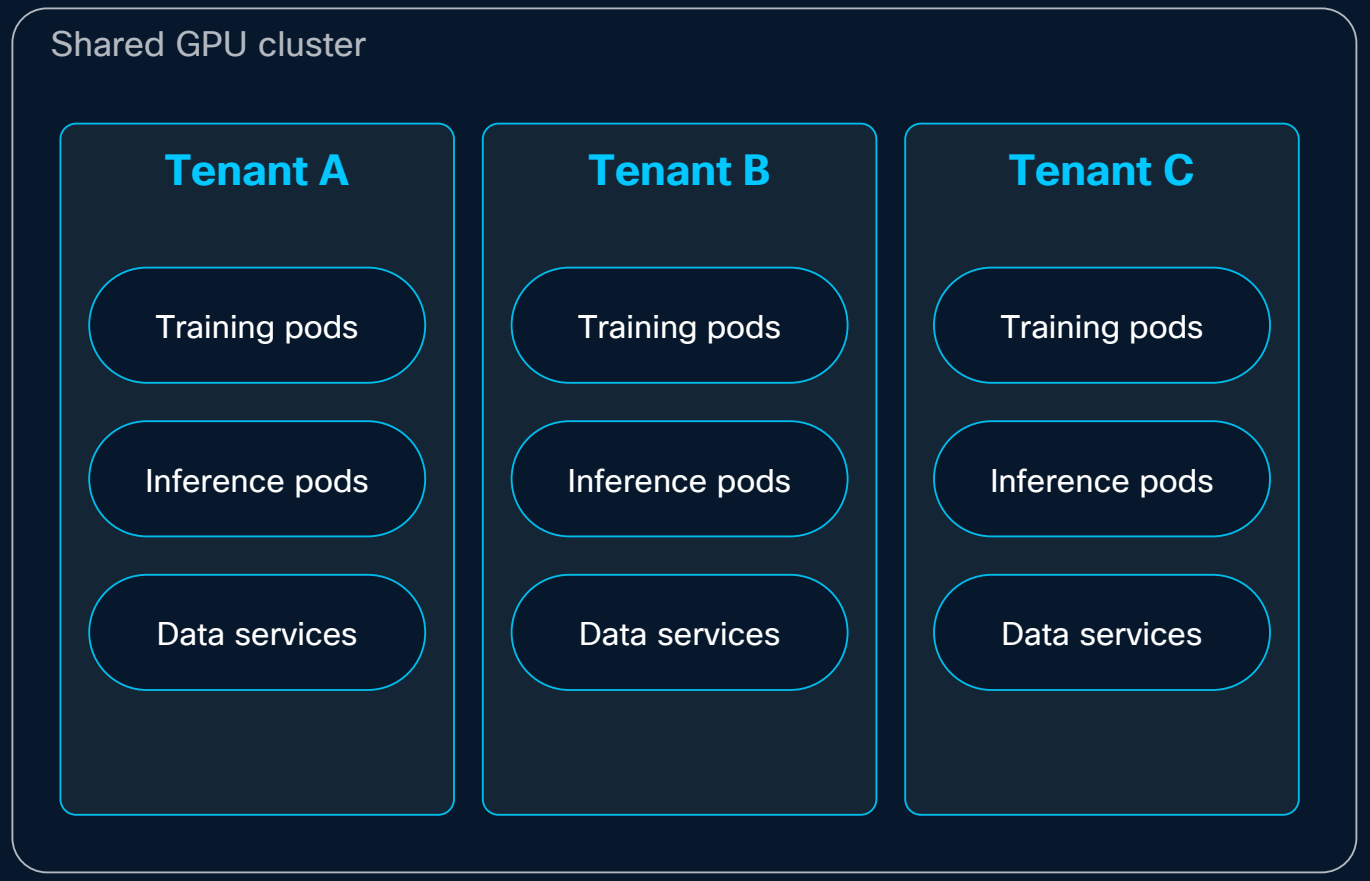


Pattern described in CNCF, "Building an AI factory on Kubernetes" (Aug 2026). Validate support for your Cilium and NIC versions.

© 2026 Cisco and/or its affiliates. All rights reserved.

cisco

Secure multi-tenancy for shared GPU clusters



How the platform helps

- Isolated network segments, in the style of a VPC, inside Kubernetes
- Identity-based policy that denies by default between tenants
- Per-tenant flow visibility and audit trails
- High performance without per-tenant hardware

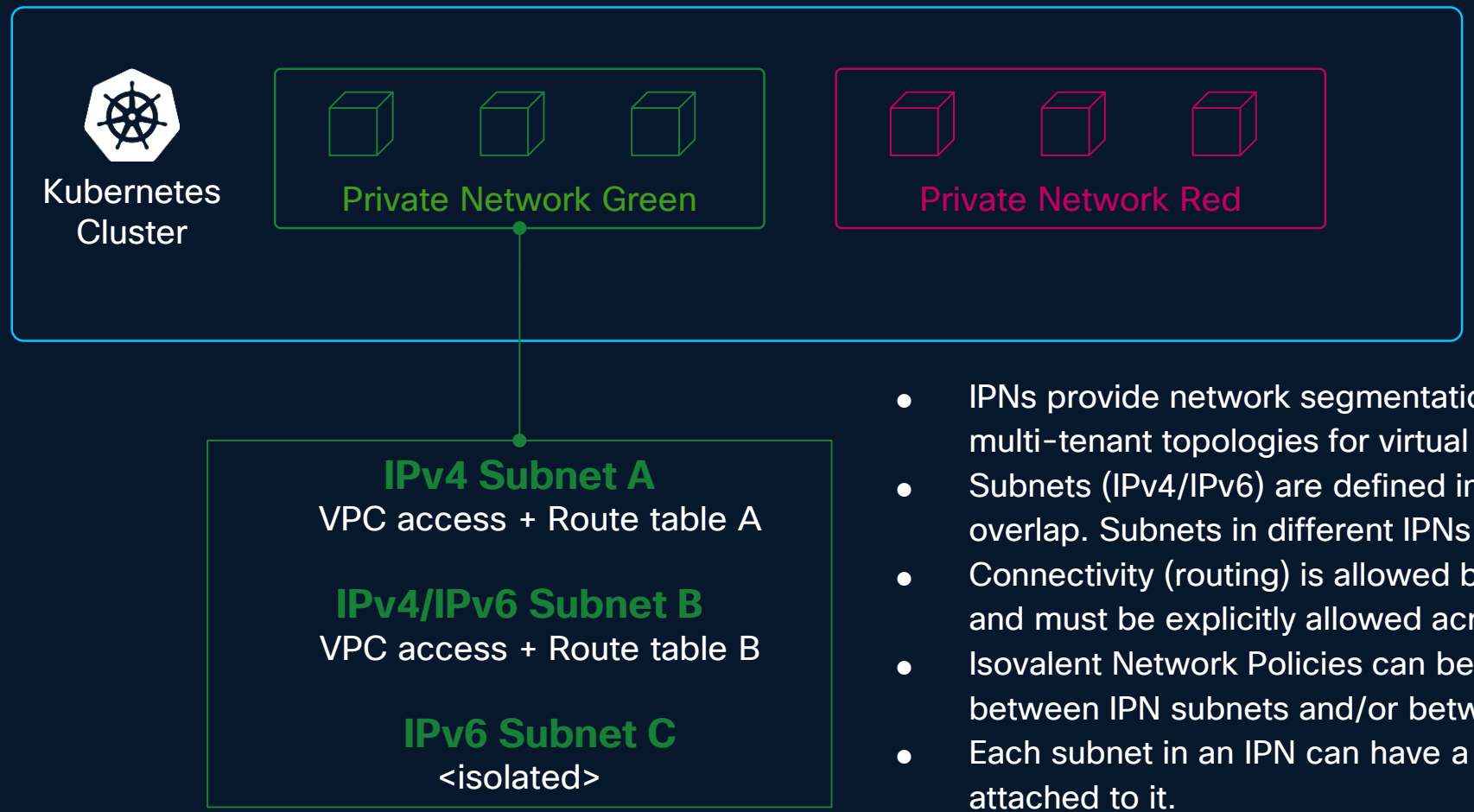
Isovalent Private Networks provide VPC-style isolation for Kubernetes (isovalent.com).

© 2026 Cisco and/or its affiliates. All rights reserved.

Isovalent Private Networks (IPNs)

AVAILABLE NOW
General Availability

The Kubernetes VPC



- IPNs provide network segmentation within Kubernetes, enabling multi-tenant topologies for virtual machines.
- Subnets (IPv4/IPv6) are defined inside of an IPN and must not overlap. Subnets in different IPNs can overlap.
- Connectivity (routing) is allowed between subnets within an IPN and must be explicitly allowed across IPNs.
- Isovalent Network Policies can be used to micro-segment traffic between IPN subnets and/or between IPNs.
- Each subnet in an IPN can have a device and a route table attached to it.
- Isolated subnets are supported.

© 2026 Cisco and/or its affiliates. All rights reserved.

cisco

SECTION 04

Consistent policy

Security that follows the workload, not the IP address

Write policy for workloads, not IP addresses

1. **Identity-based**
   Rules reference Kubernetes labels. Pods can move or scale without policy changes.

2. **Layer 7 aware**
   Allow specific HTTP methods and paths, or DNS names, not just ports.

3. **Default deny**
   Start closed and open only the flows each service needs.

CiliumNetworkPolicy (illustrative)

```
kind: CiliumNetworkPolicy
metadata: {name: gateway-to-model}
spec:
  endpointSelector:
    matchLabels: {app: model-server}
  ingress:
  - fromEndpoints:
    - matchLabels: {app: inference-gateway}
    toPorts:
    - ports: [{port: "8000", protocol: TCP}]
      rules:
        http:
        - {method: POST, path: "/v1/.*"}
```

© 2026 Cisco and/or its affiliates. All rights reserved.

Protect data in motion without touching applications

1

Transparent encryption

WireGuard or IPsec in the kernel encrypts node-to-node traffic. No sidecars, no code changes.

2

Identity and Policy

Identity-aware, service-to-service authentication so only verified workloads connect.

3

DNS and egress control

Allow access to approved model hubs, APIs and data sources by name, and block the rest.

© 2026 Cisco and/or its affiliates. All rights reserved.

cisco

Tetragon adds runtime enforcement inside the node

1 Observe

Process execution, file access and network activity, tied to Kubernetes identity

2 Decide

Per-workload runtime policy written as Kubernetes resources

3 Enforce

eBPF hooks block or alert in the kernel, with low overhead

Typical use cases

- Detect an unexpected binary launching in a training pod
- Protect model weights and datasets from unexpected file access
- Audit what ran, where and when for compliance

© 2026 Cisco and/or its affiliates. All rights reserved.

cisco

SECTION 05

Visibility and troubleshooting

Every flow, with Kubernetes identity, from one source of truth

Hubble shows every flow with Kubernetes identity

SOURCE	DESTINATION	PORT	VERDICT
inference-gateway	model-server	8000	FORWARDED
trainer-0	dataset-cache	9000	FORWARDED
tenant-b/notebook	tenant-a/model-server	8000	DROPPED
agent-runner	api.example.com	443	FORWARDED
model-server	object-store	443	FORWARDED

What teams see

- L3 to L7 flows between named workloads
- Service dependency maps
- DNS, HTTP and network metrics
- Policy verdicts and drop reasons

© 2026 Cisco and/or its affiliates. All rights reserved.

cisco

From symptom to root cause with one source of truth

1 Symptom	2 Locate	3 Explain	4 Fix and verify
A training job slows down or inference latency spikes.	Filter Hubble flows by workload to see where traffic stalls or drops.	Drop reasons and policy verdicts point to the cause.	Update policy as code, confirm flows recover and keep history.

© 2026 Cisco and/or its affiliates. All rights reserved.

SECTION 06

Better together with Cisco Nexus

Isovalent for the cloud-native layer, Nexus for the fabric

Better together: Nexus One and Isovalent

Isovalent: cloud-native layer

- Enterprise networking for containers
- Code-driven, repeatable operations
- Visibility down to the OS

Cisco Nexus One: underlay

- Predictable performance
- Reliability and security at the physical level
- Infrastructure as code

Source: Cisco Blogs, Feb 2026. Isovalent also integrates with Cisco ACI.

© 2026 Cisco and/or its affiliates. All rights reserved.

The same intent-driven model from the fabric to the pod

Intent in Git → **Review and CI checks** → **Apply to fabric and clusters** → **Verify with Hubble**

Underlay: Cisco Nexus

- Configured as infrastructure as code
- Policy-driven and repeatable
- Predictable performance and reliability at the physical layer

Kubernetes: Isovalent

- CiliumNetworkPolicy and Tetragon policies as code
- Networking deployed from intent, with business logic
- Same review, same pipeline, same audit trail

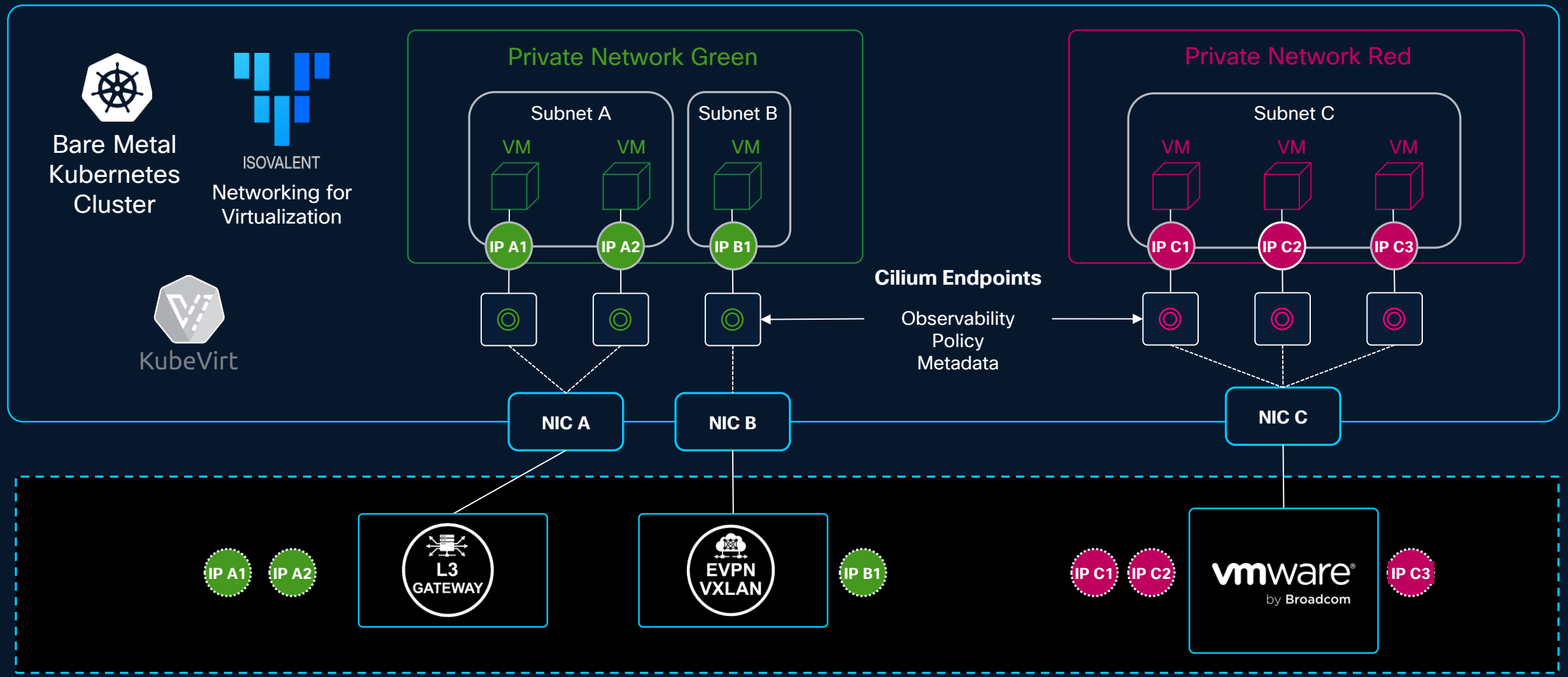
Based on remarks by Nico Vibert, Cisco Blogs, Feb 2026.

© 2026 Cisco and/or its affiliates. All rights reserved.

AVAILABLE NOW
General Availability

Isovalent Networking for Virtualization

Enterprise-grade network and security for virtual machines

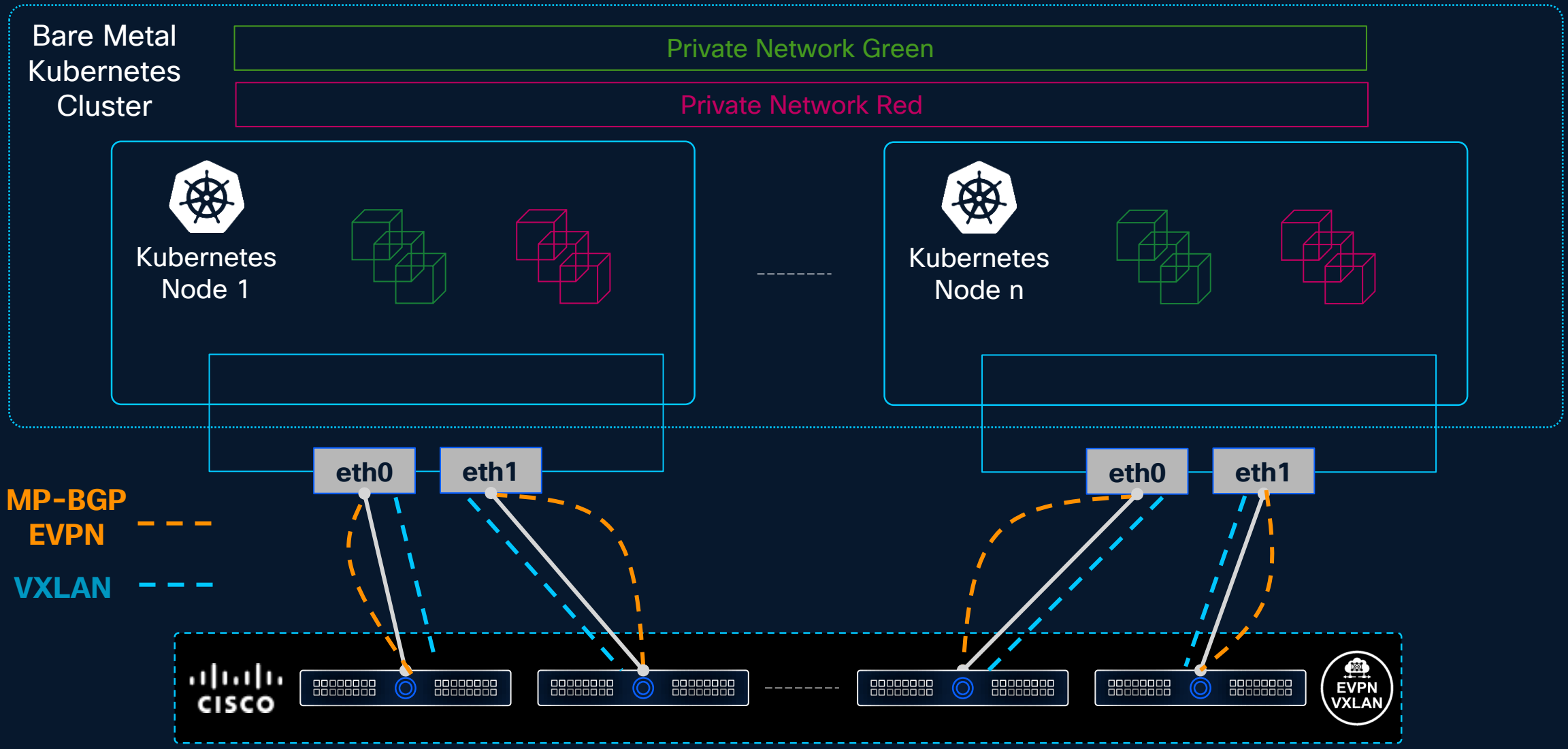


© 2026 Cisco and/or its affiliates. All rights reserved.

CISCO NEXUS ONE
VALIDATED WITH NX-OS

Nexus 9K EVPN Scale Out Design

Distributed peering

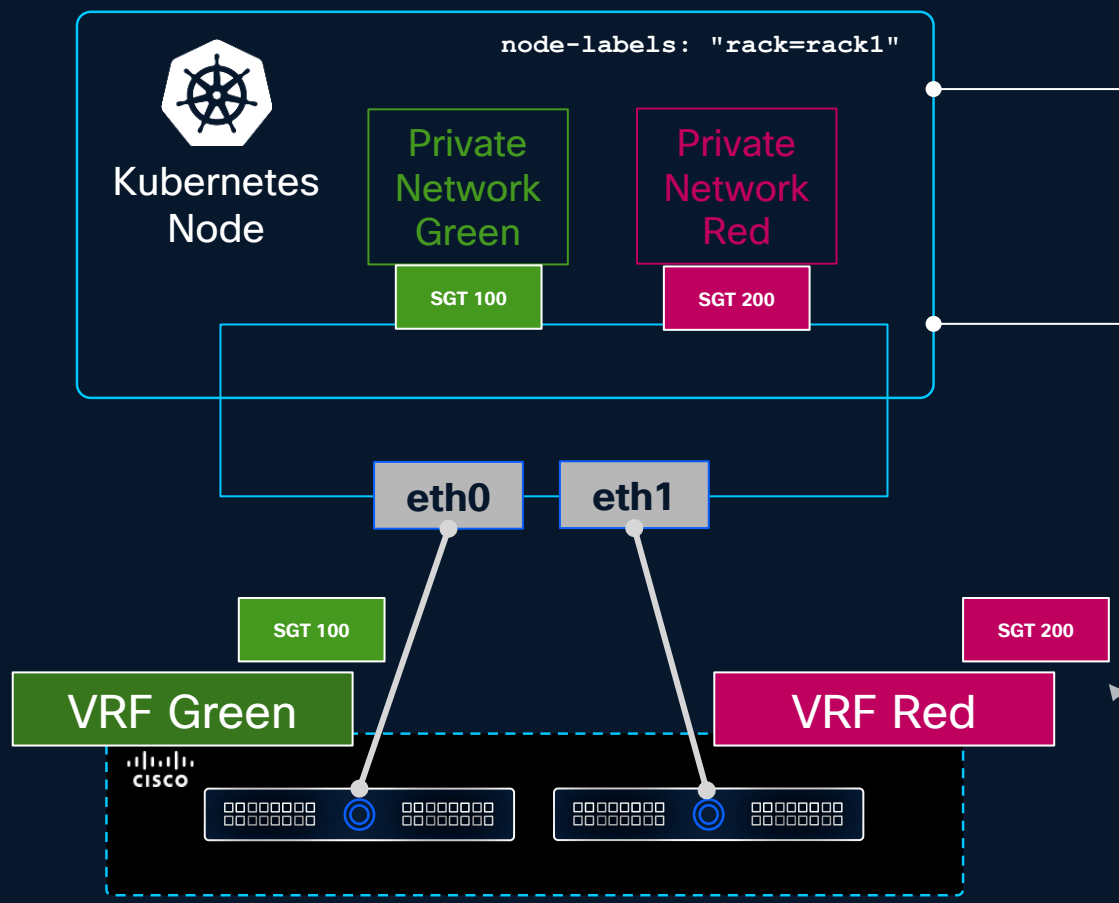


© 2026 Cisco and/or its affiliates. All rights reserved.

cisco

Fabric Security

Security Group Tag




```
apiVersion: isovalent.com/v1alpha1
kind: FabricSecurityGroup
metadata:
  name: "100"
spec:
  endpointSelector:
    matchLabels:
      cni:com.isovalent.private-network.name: green
```

```
apiVersion: isovalent.com/v1alpha1
kind: FabricSecurityGroup
metadata:
  name: "200"
spec:
  endpointSelector:
    matchLabels:
      cni:com.isovalent.private-network.name: red
```

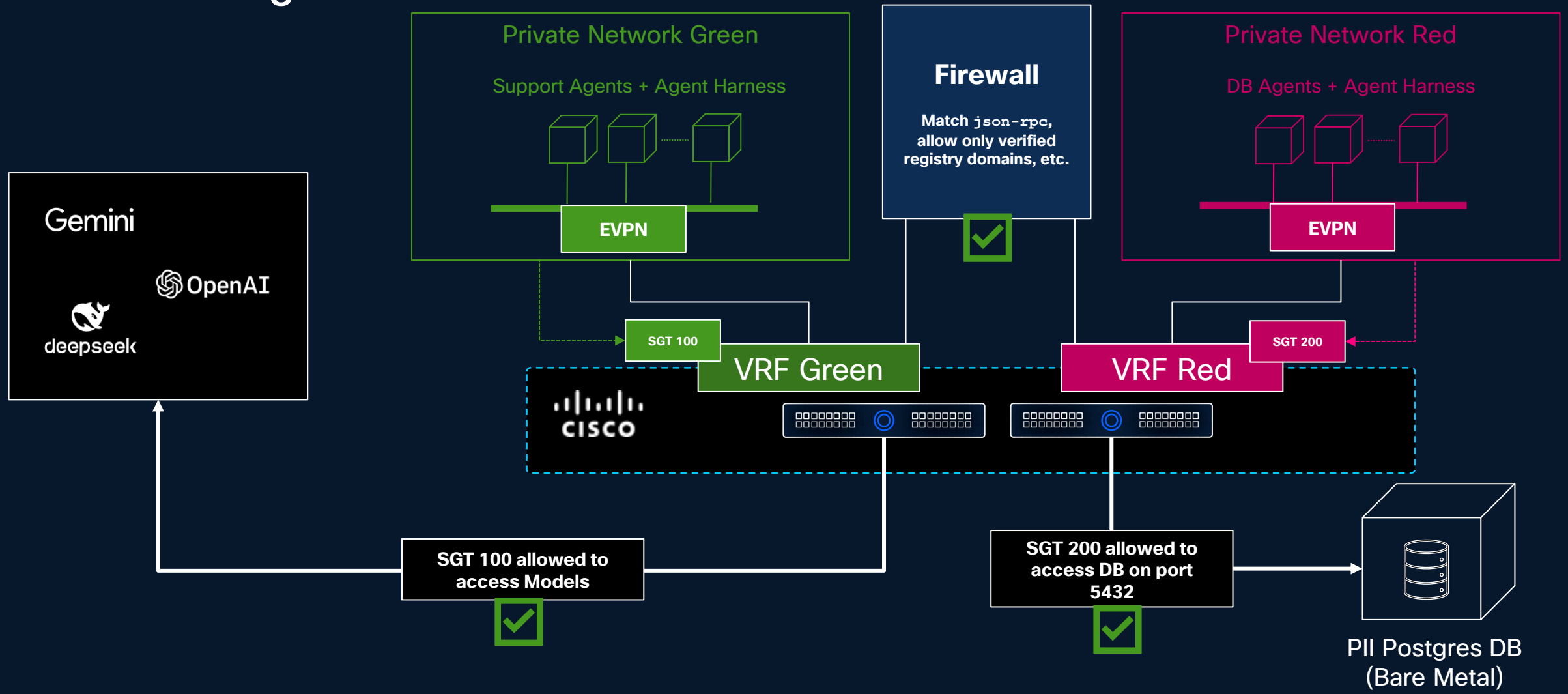
SGT is associated with each Isovalent Private Network

© 2026 Cisco and/or its affiliates. All rights reserved.

cisco

Fabric Segmentation using Security Groups

Advanced agentic use case



© 2026 Cisco and/or its affiliates. All rights reserved.

SECTION 07

Operating at scale

Fewer tools, one platform and a practical path to adopt it

Replace single-purpose tools with one platform layer

TYPICAL STACK TODAY

- Separate CNI plugin
- kube-proxy
- Encryption agent
- Service mesh sidecars
- Observability agent
- Runtime security agent

WITH ISOVALENT NETWORKING

One eBPF-based platform

- Connectivity, load balancing and policy
- Encryption and mutual authentication
- Flow visibility and metrics
- Runtime security with Tetragon

© 2026 Cisco and/or its affiliates. All rights reserved.

cisco

A practical adoption path

1

Standardize

Run Cilium as the CNI on new clusters. Keep one baseline configuration.

2

Observe

Enable Hubble. Baseline normal flows for training and inference.

3

Protect

Move to default deny with identity policy. Turn on encryption.

4

Extend

Add Cluster Mesh, Tetragon and GitOps across environments.

© 2026 Cisco and/or its affiliates. All rights reserved.

Key takeaways

1

Consistent connectivity

One CNI and IPAM layer across clouds, data centers and bare metal, at kernel speed.

2

Consistent policy

Identity-based L3 to L7 rules, encryption and runtime enforcement that follow the workload.

3

Consistent visibility

Hubble and Tetragon give one source of truth for troubleshooting and audit.

© 2026 Cisco and/or its affiliates. All rights reserved.

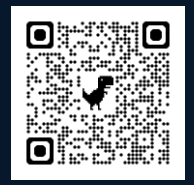
cisco

Learn more

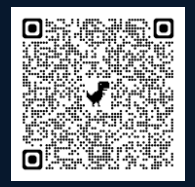
1. Isovalent



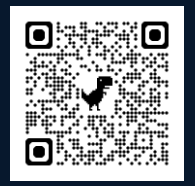
2. Cilium project



3. Cisco Blogs: Isovalent and Nexus One



4. CNCF: Building an AI factory on Kubernetes



© 2026 Cisco and/or its affiliates. All rights reserved.

cisco

cisco Connect !

Thank you




CISCO