CISCO Engage | Tech Day Charlotte

Cisco Unified Edge with Red Hat Solutions

Shixiong Shang
Technical Marketing Engineer

Tom Qian
Solutions Engineer

March 31, 2026

Agenda

Introduction

Network Design

Red Hat Solution Design Options

Red Hat AI Inference Server

Demo

Q&A

Introduction

Actual Status Compute Edge:

Edge = low number of VMs and/or docker container.

Single host ESXi deployments are normal.

K8s is slowly adopted

Companies are looking for ESXi replacement.

Design Consideration

- Key factors: Redundant and Simple
- Deployment: Easy, consistent, repeatable, and automated
  - NFC, CVDs, Blueprints
- Edge Optimized
  - Size: 18” deep - Noise: 45dB* - Temperature: 45°C*
- Workloads: Capable to run existing and new AI workloads
  - Legacy Apps, docker container, VMs, micro-services, and AI workloads

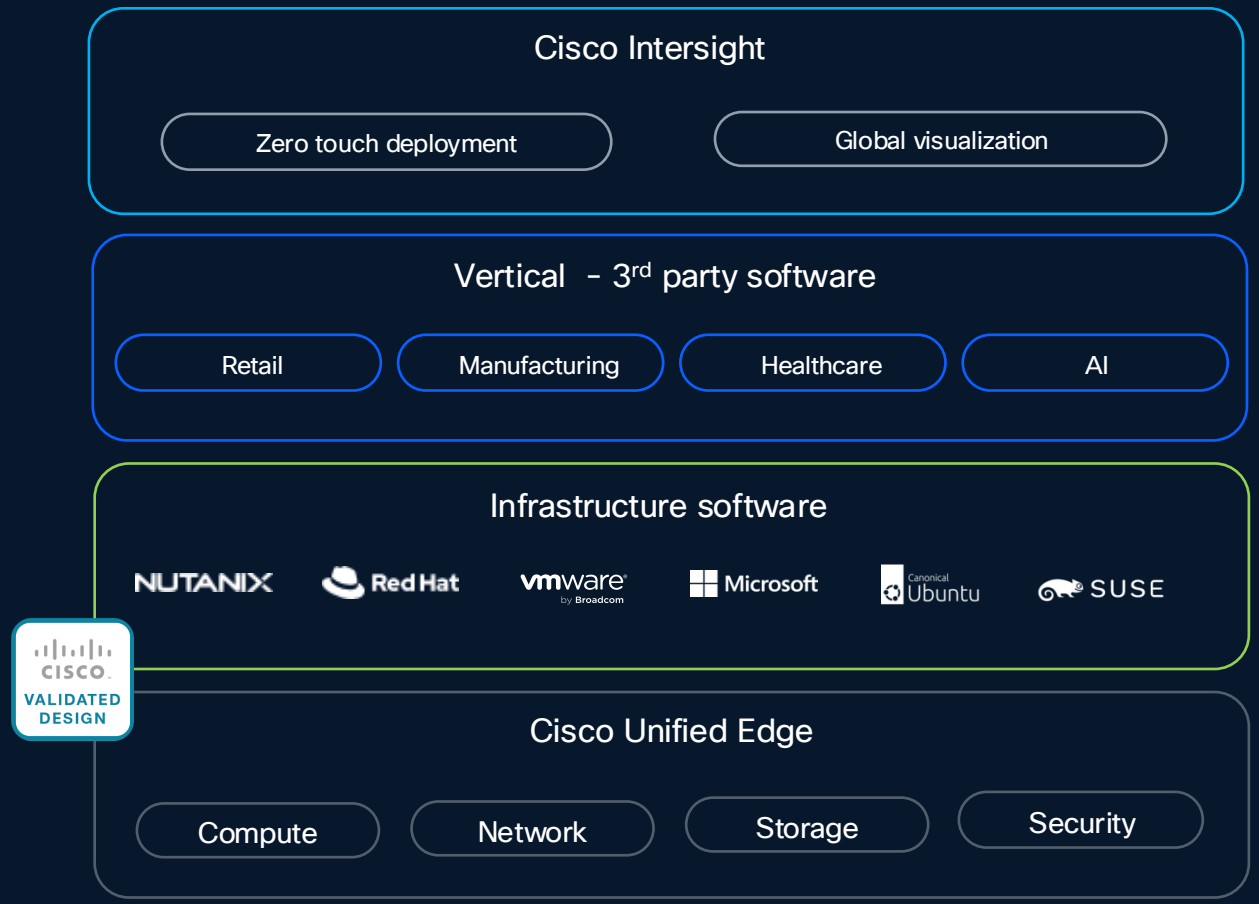
Full-stack edge solutions

AI-ready edge

Plug and play vertical software

Flexibility and choice

Reference designs with infrastructure software partners



Network Design

Cisco Unified Edge: Future-Ready Performance

AI-ready edge

Integrates compute, networking, storage, and security

Compute node

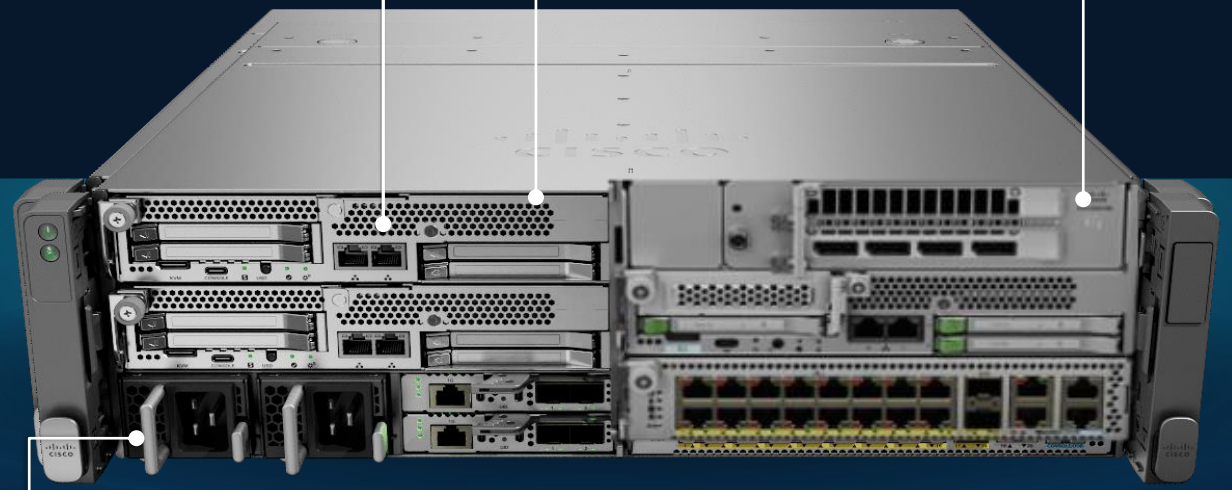
- Compute
- Storage
- Software

GPUs

UCS XE130c M8:
Half-height/half-length GPU
NVIDIA L4

UCS XE150c M8:
Full-height/full-width GPU
NVIDIA RTX Pro 4500/6000 and L40S

Additional GPUs on roadmap



Intel Xeon 6 SoC

CPU native AI inferencing (AMX)

Confidential compute (TDX & SGX)

Integrated Ethernet

Scalable multithreaded cores (12, 20, 32)

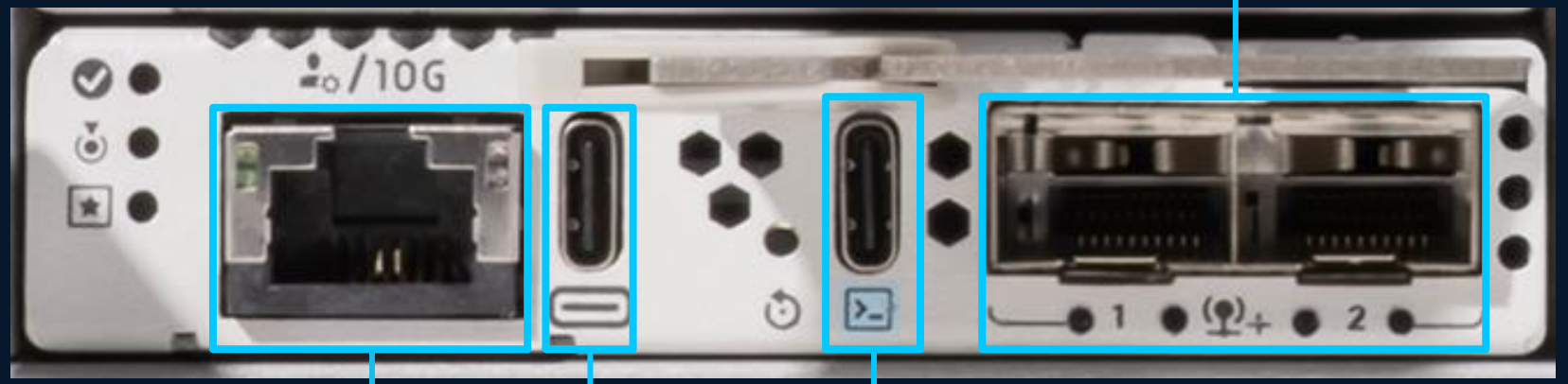
2x 2400W PSUs

N+N redundancy capable

NUTANIX · Red Hat · vmware by Broadcom · Canonical Ubuntu · Microsoft · intel · cisco · SUSE

Cisco Unified Edge

eCMC Port Layout



Dual 1/10/25 Gb uplinks

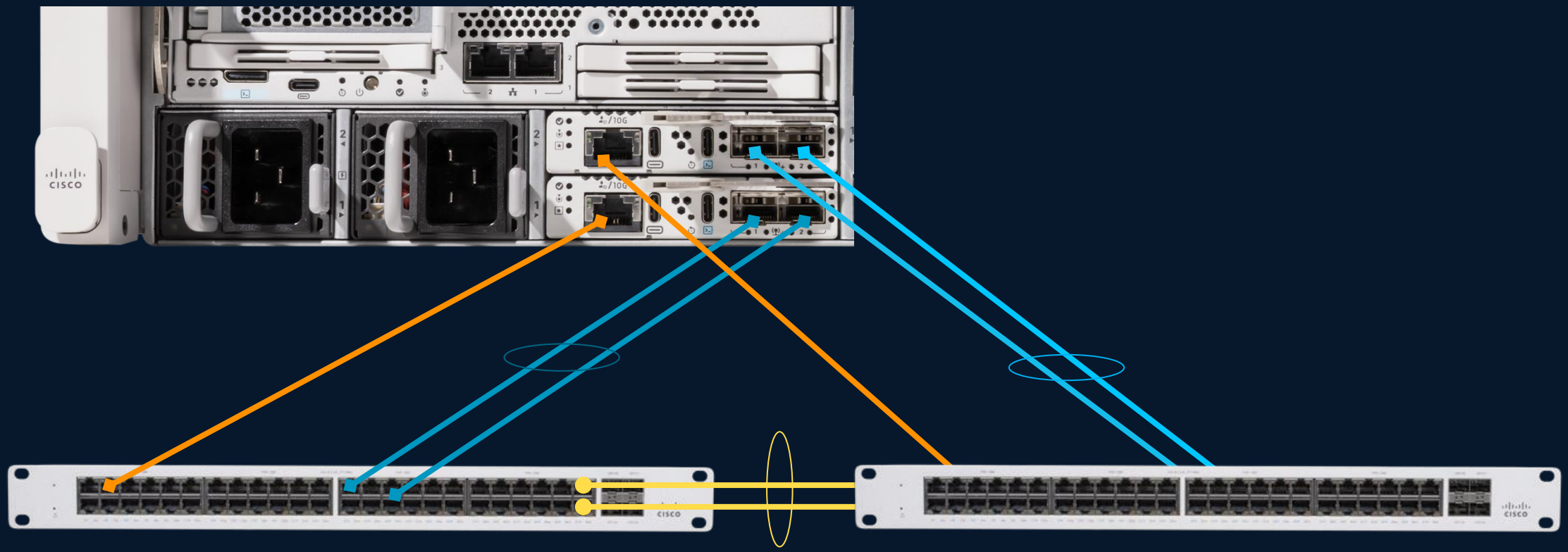
USB-C Serial Console

USB-C port

Both eCMC Management ports must be connected

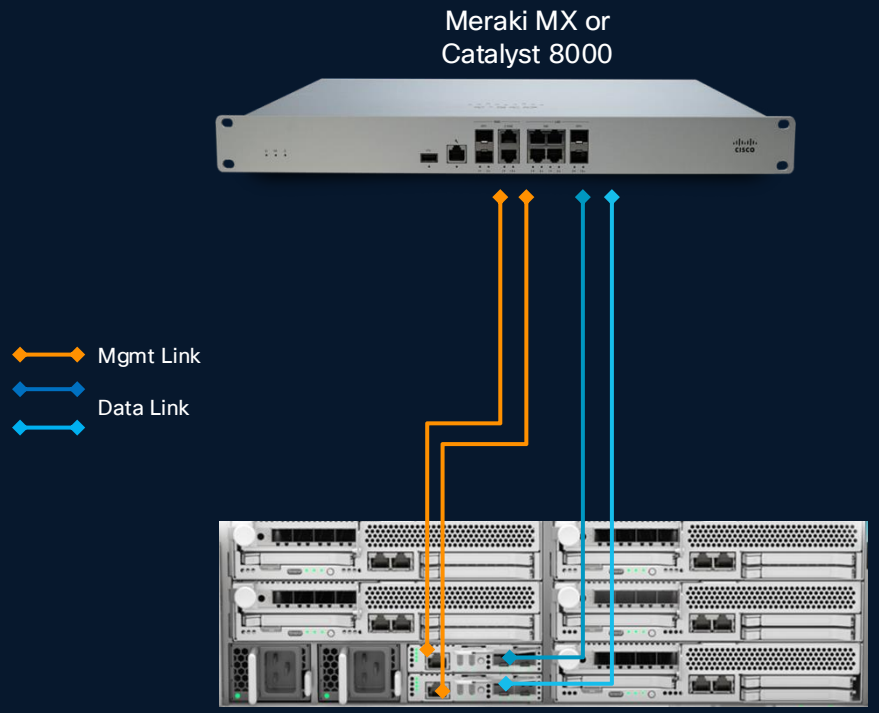
1/10Gb Management port

eCMC Connections



- Management port requires its own uplinks
- One or two uplinks from each eCMC
- If two uplink links are used, they must be

Simple HA Topology



Both management and uplink ports to connect to the same router

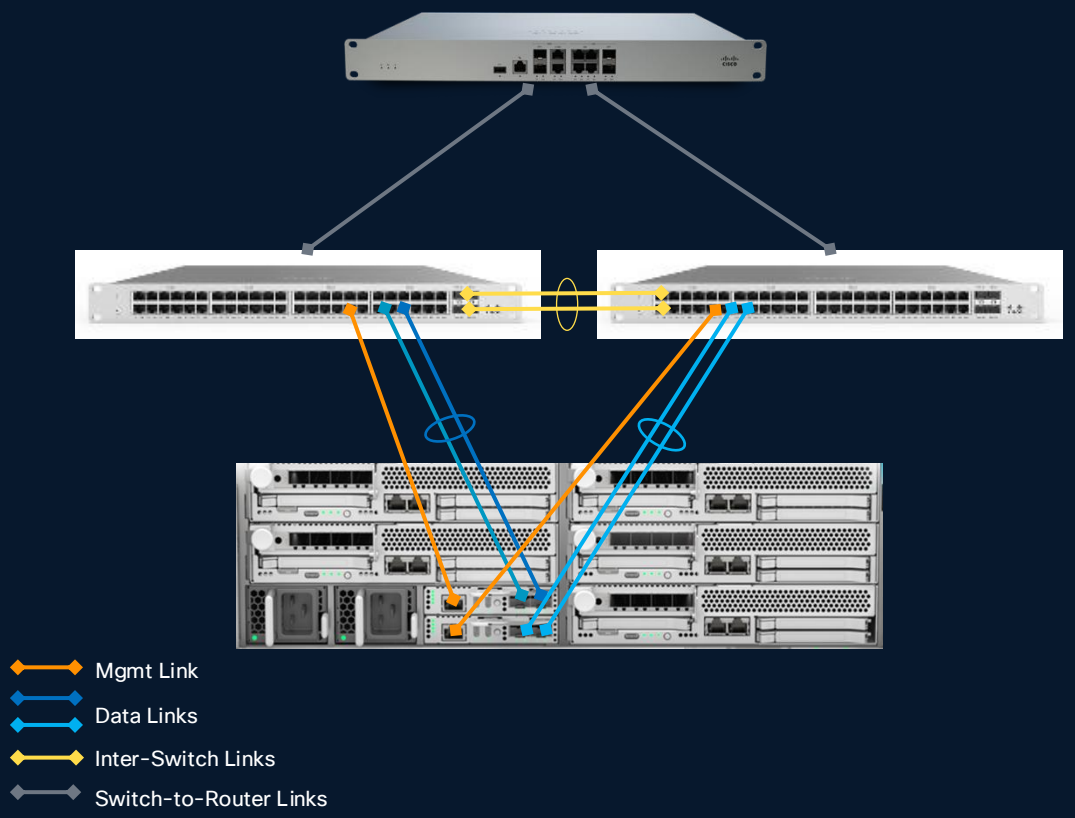
Uplink ports are individual ports with Meraki MX*

Port-channel is only possible with Catalyst 8000

Uplink ports are configured as VLAN trunk ports

** Meraki MX does not support port-channel*

Redundant Switching Topology

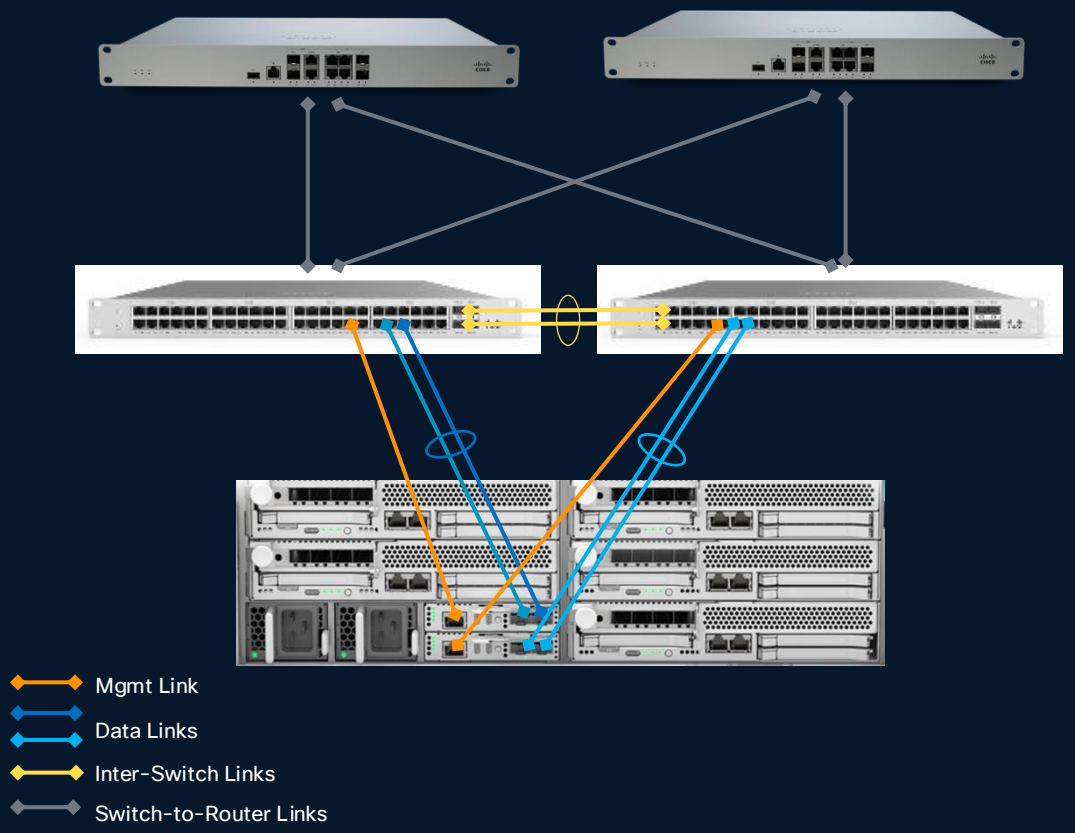


Both management and uplink ports are distributed to a pair of switches

Uplink ports are configured as port-channel

One single Meraki MX or Catalyst 8000 router

Fully Redundant Topology



Both management and uplink ports are distributed to a pair of switches

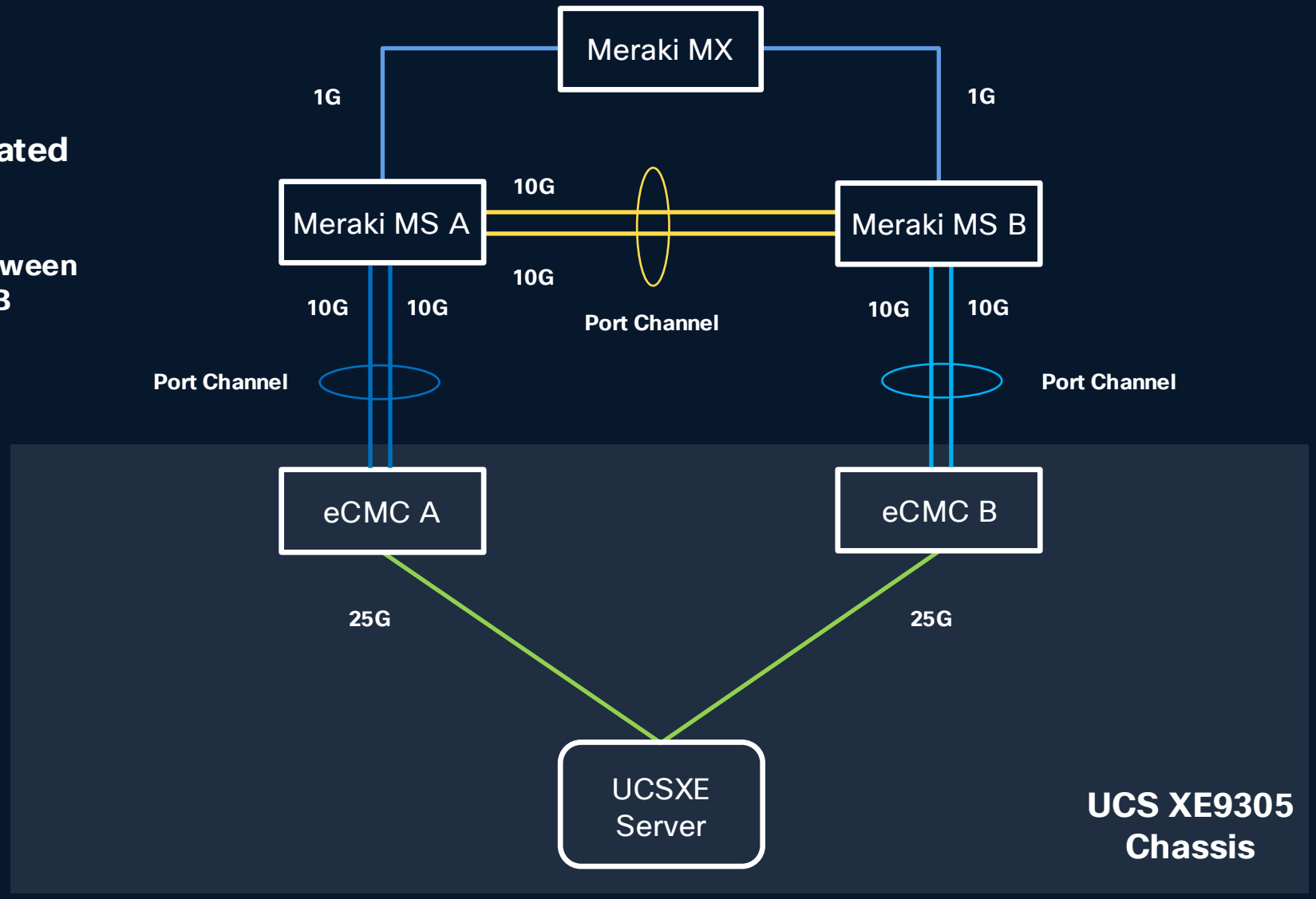
Uplink ports are configured as port-channel

A pair of Meraki MX or Catalyst 8000 for HA

Uplink ports are configured with the same set of VLANs

Data Plane

A switch is integrated into each eCMC
*** STP is disabled**
*** NO data path between eCMC A and eCMC B**

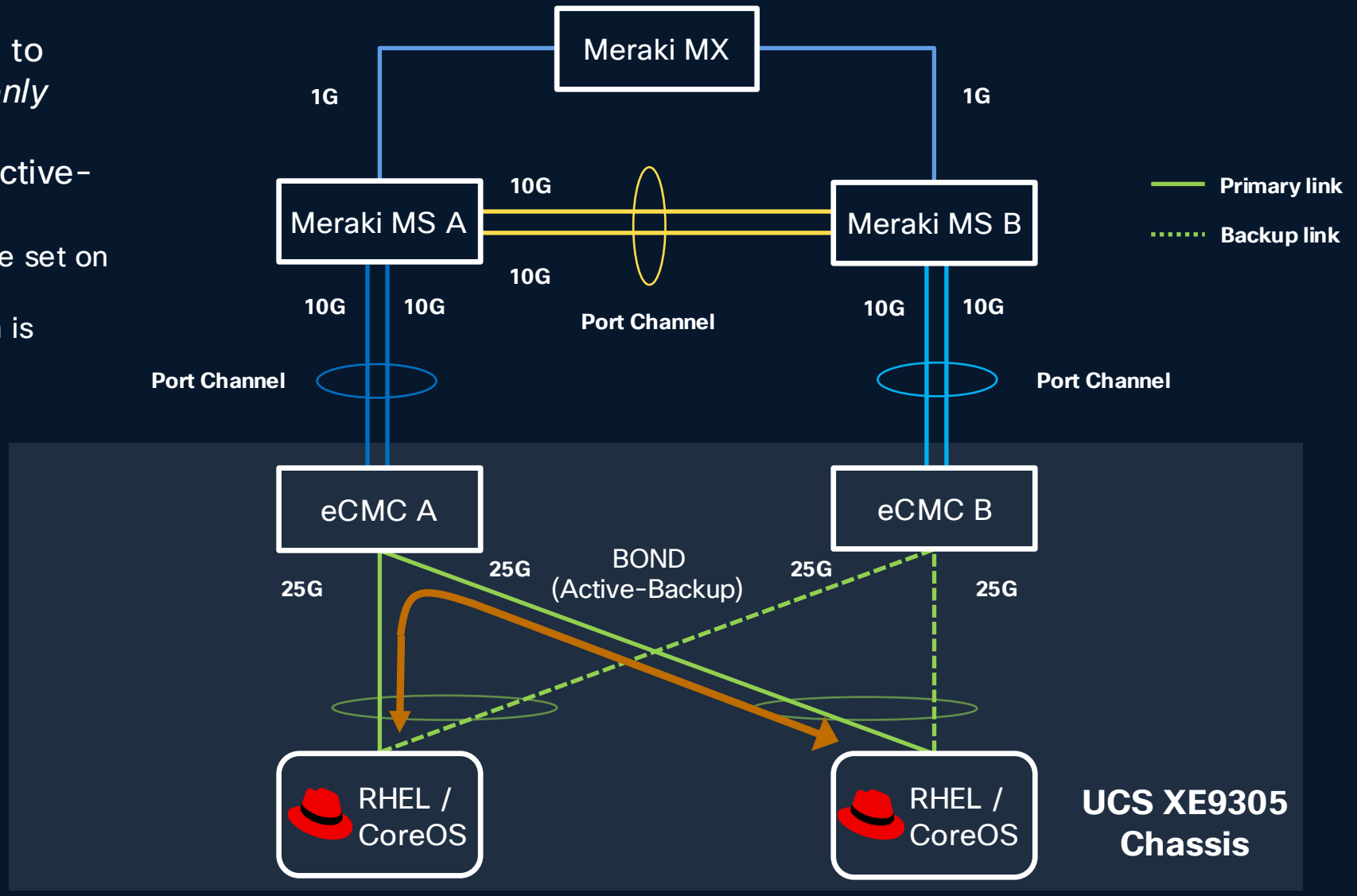


Data Plane

Server is connected to eCMCs via *2 LOM only*

Bond interface (in Active-Backup mode)

- Primary link must be set on all servers.
- Primary Reselection is enabled by default

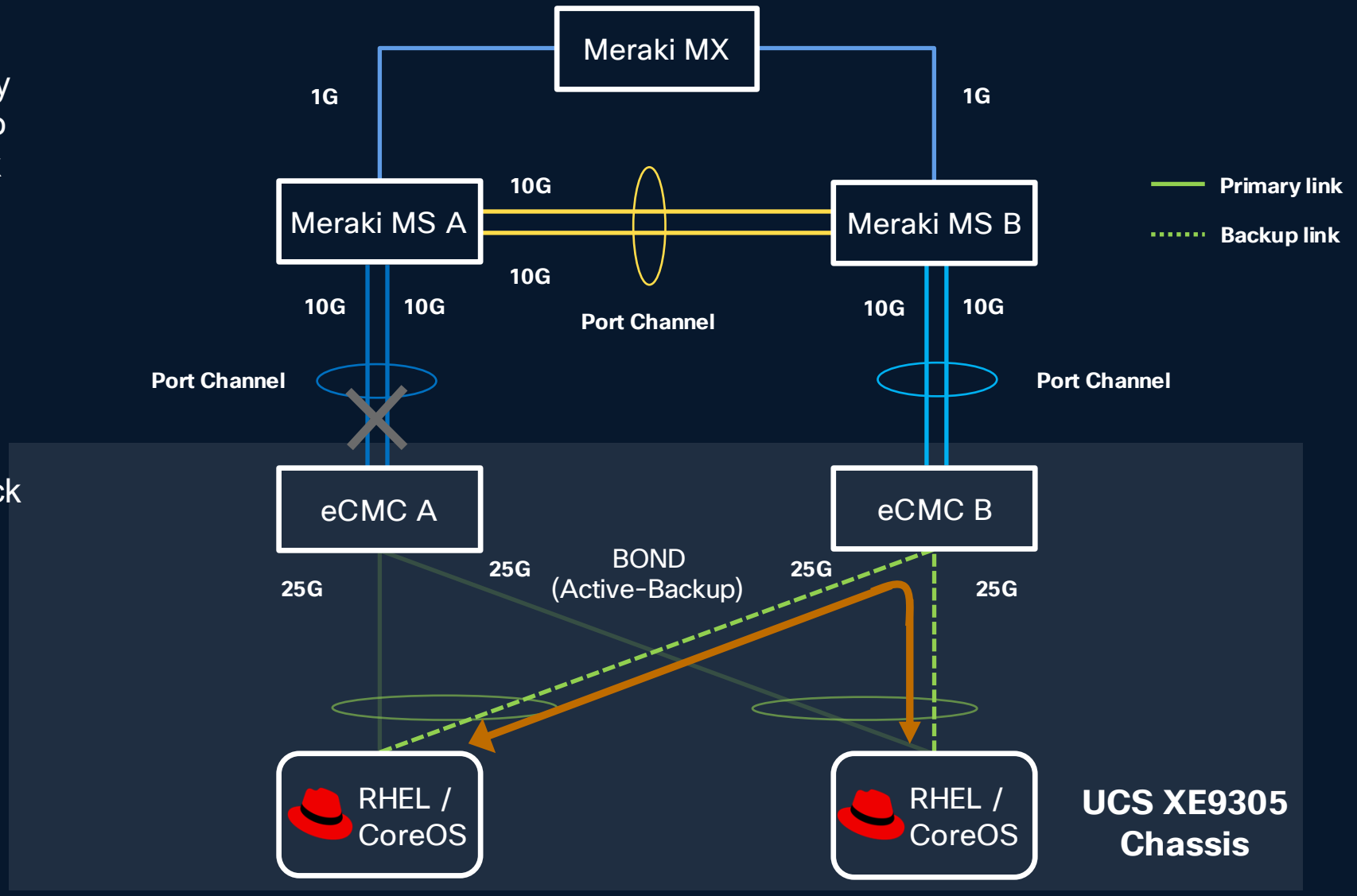


Data Plane

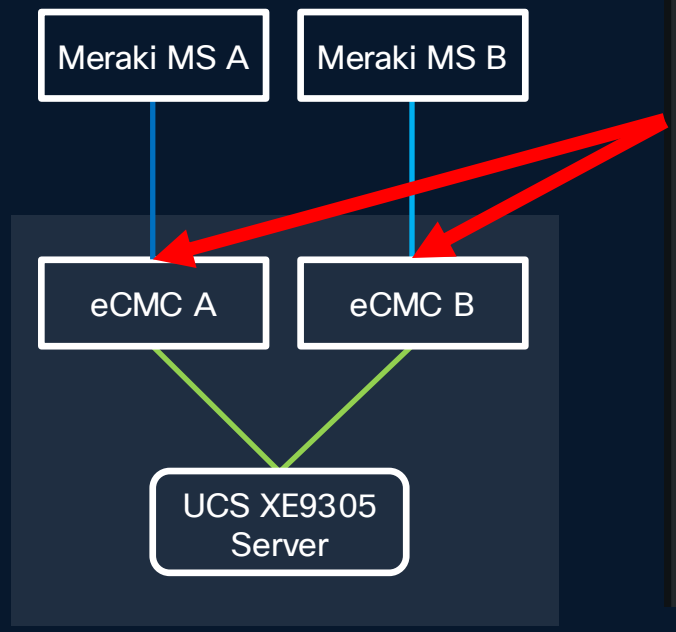
eCMC automatically brings down links to servers upon uplink failures

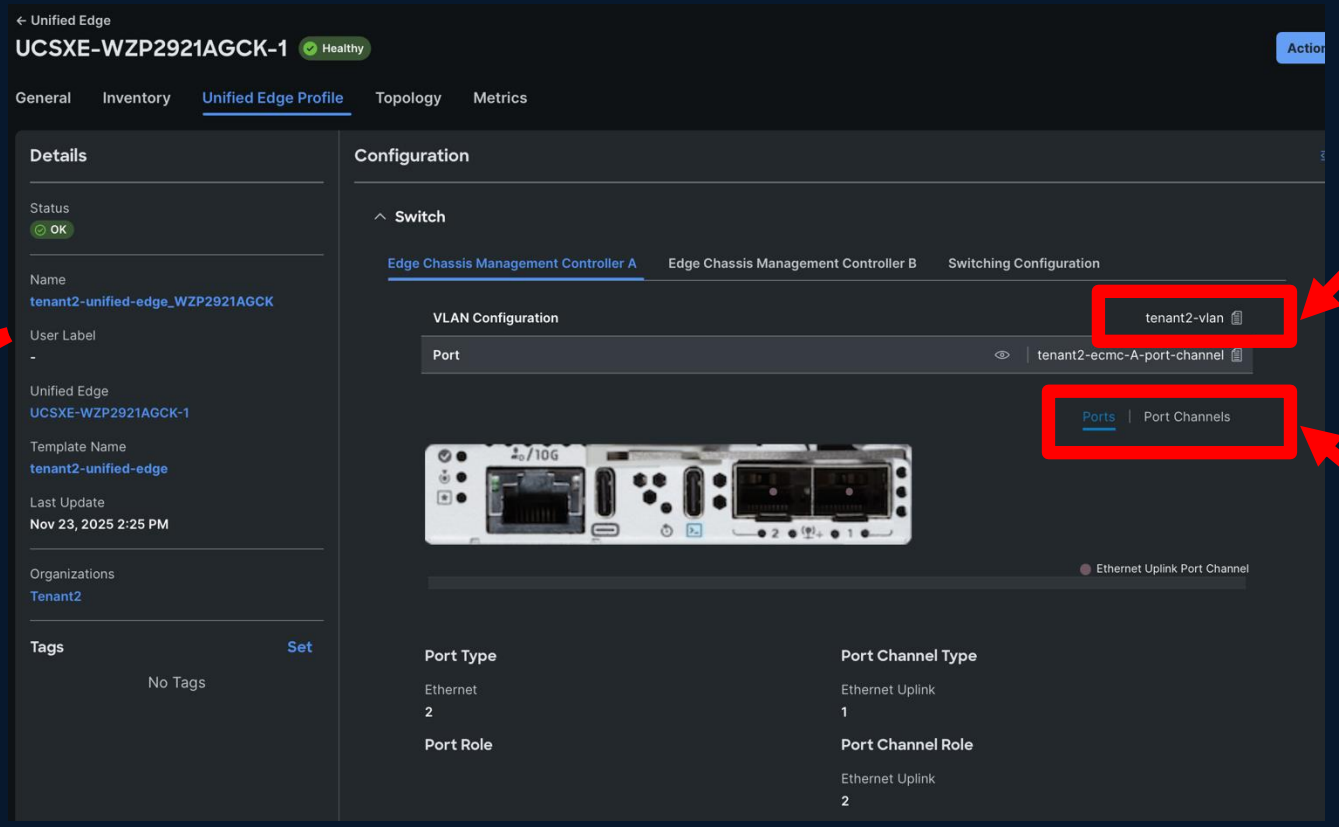
RHEL/CoreOS will trigger switchover from primary to backup

Traffic switches back to the primary link when uplinks are back online



Uplink Network Configuration

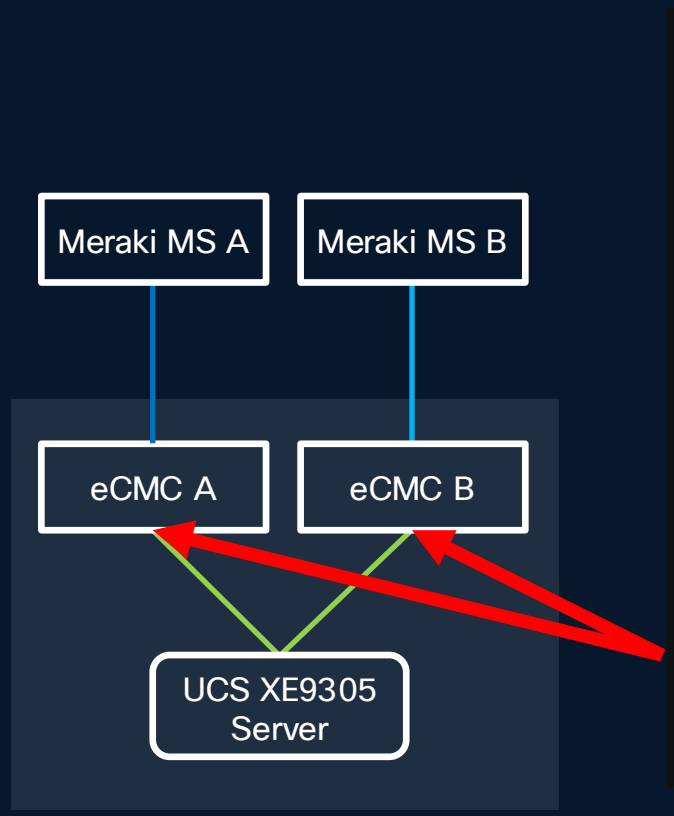


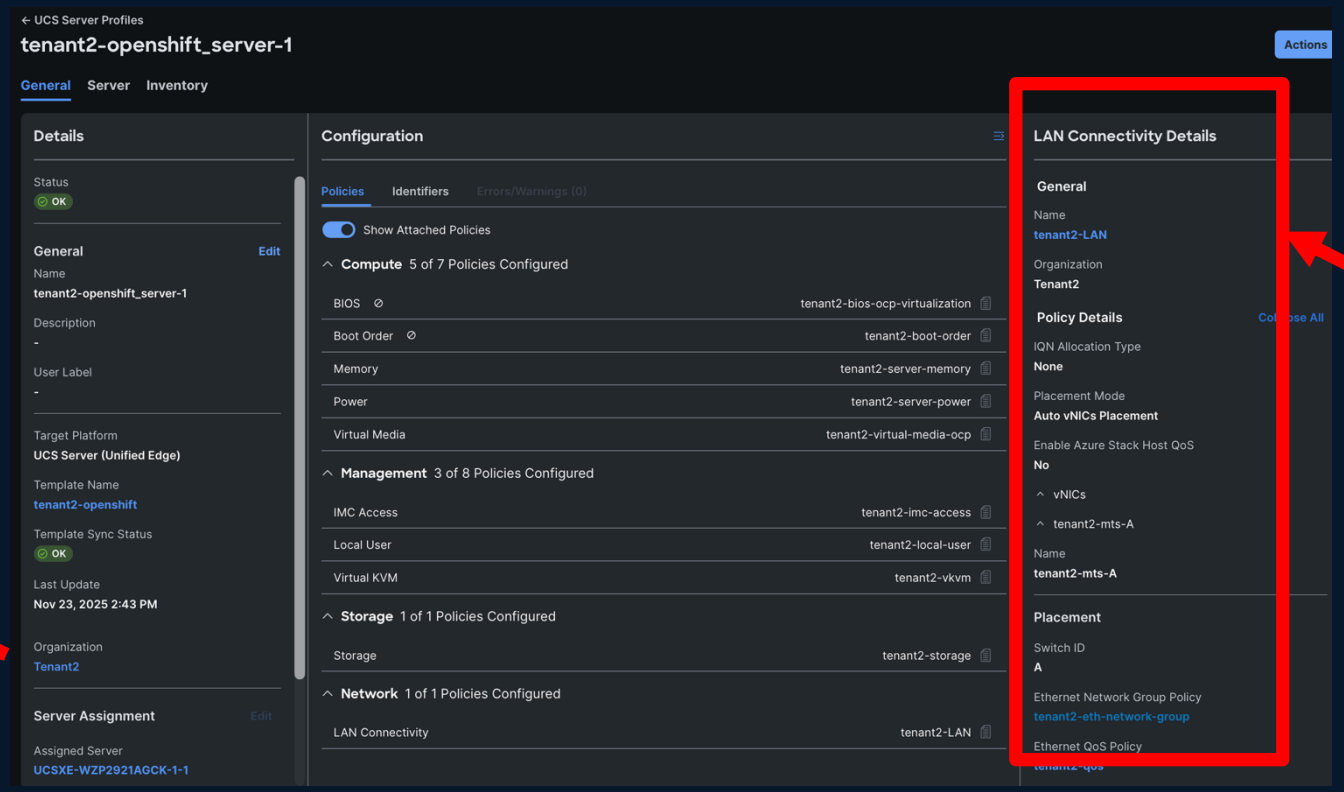


Which VLANs are provisioned on eCMCs and allowed on uplinks

Single port or Port-Channel

Server Network Configuration

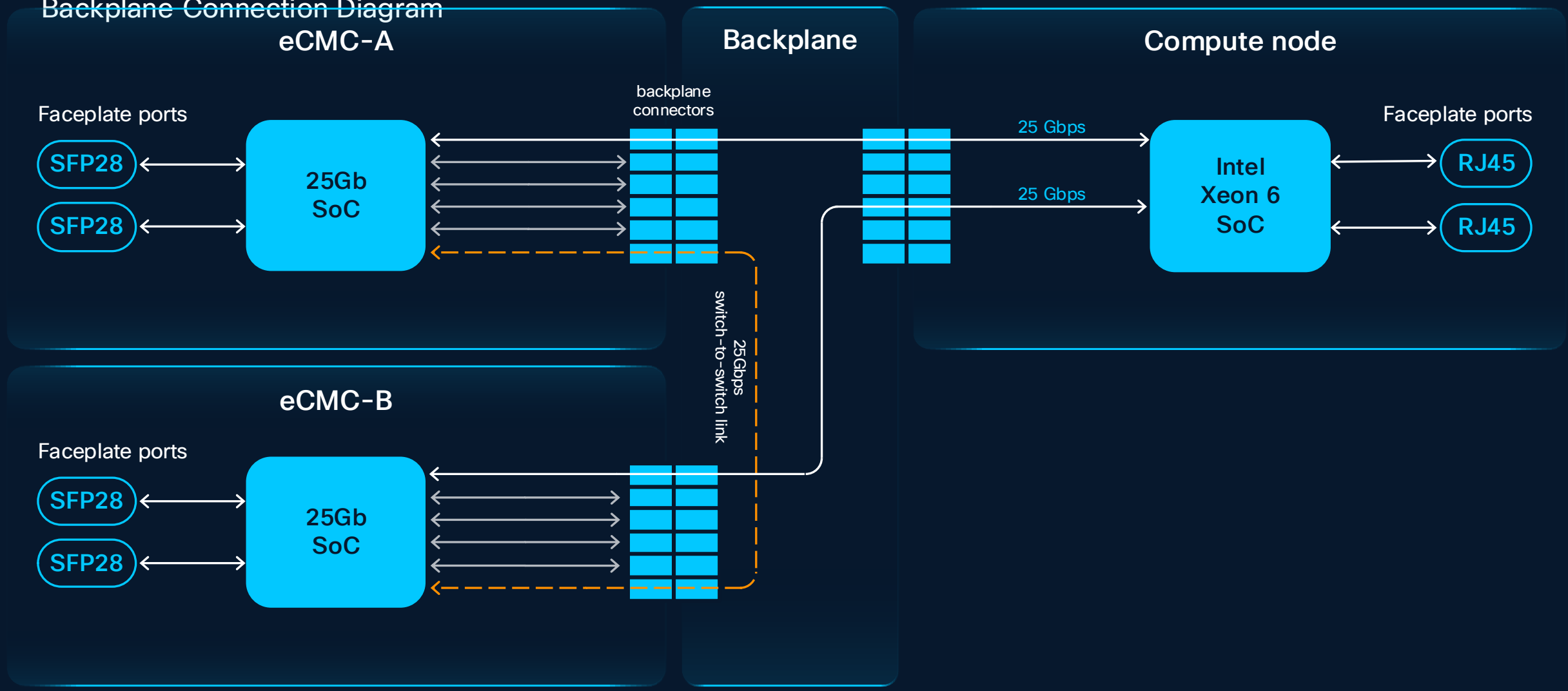




Which VLANs are allowed on the server facing ports

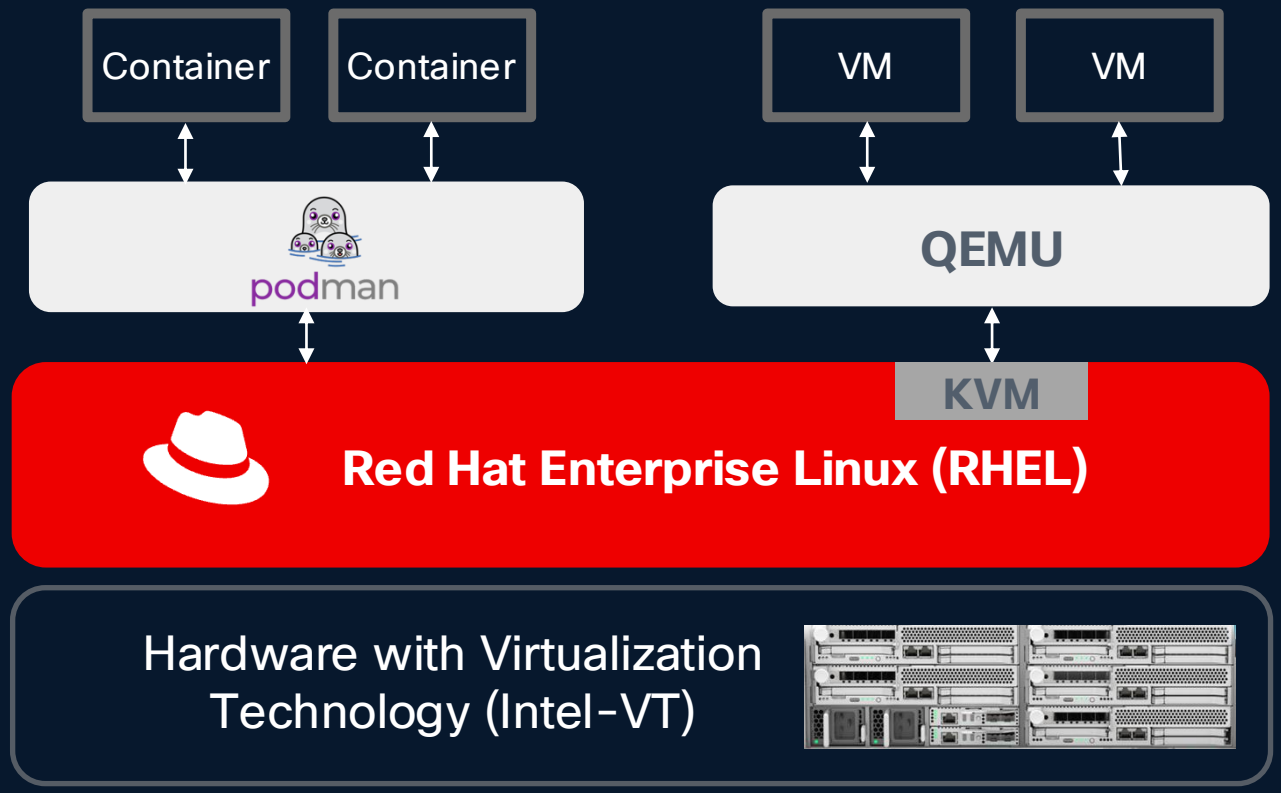
Cisco Unified Edge

Backplane Connection Diagram



Red Hat Solutions Design Options

RHEL with Podman and KVM



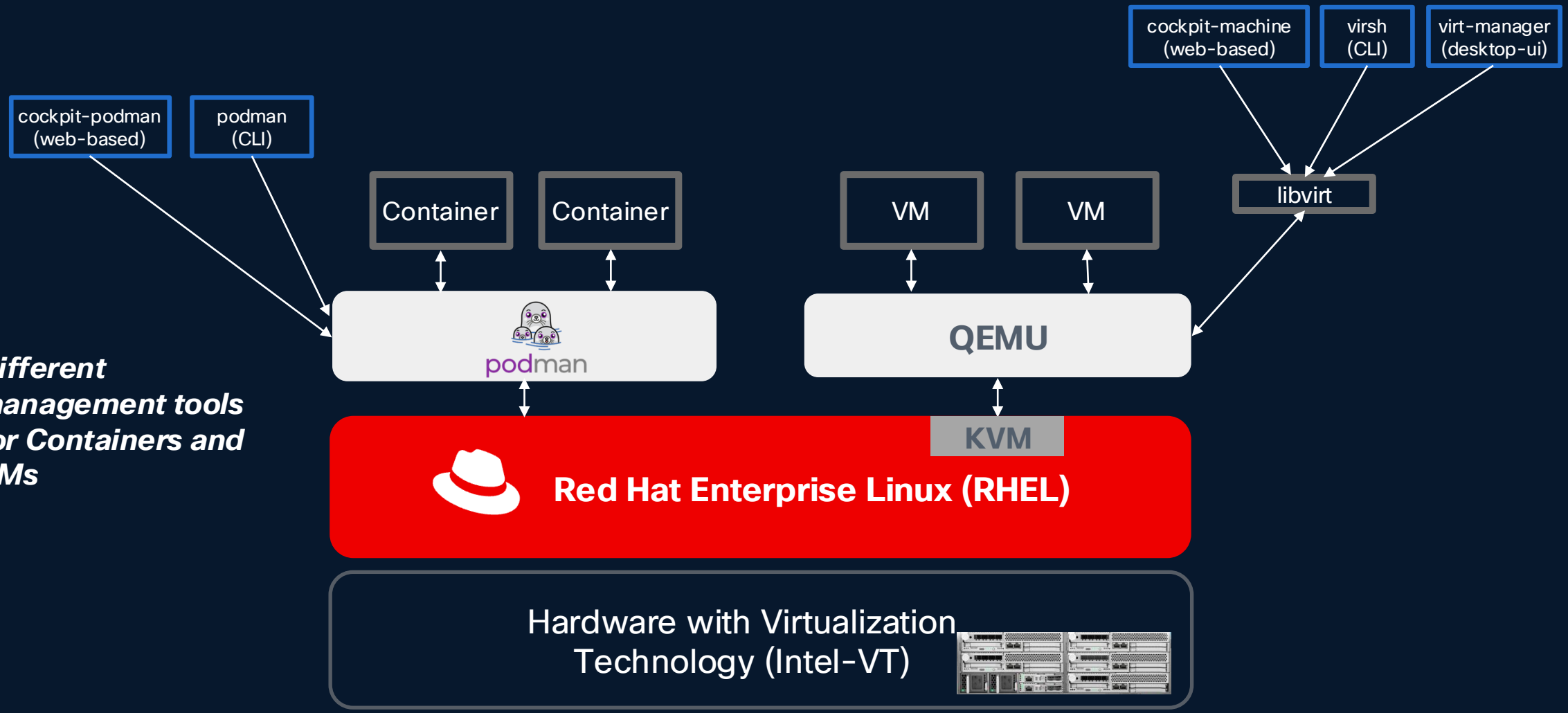
Podman: Daemonless container engine for OCI (Open Container Initiative) containers on Linux

KVM: Handles CPU and memory virtualization in kernel space

QEMU: Device emulation and VM management

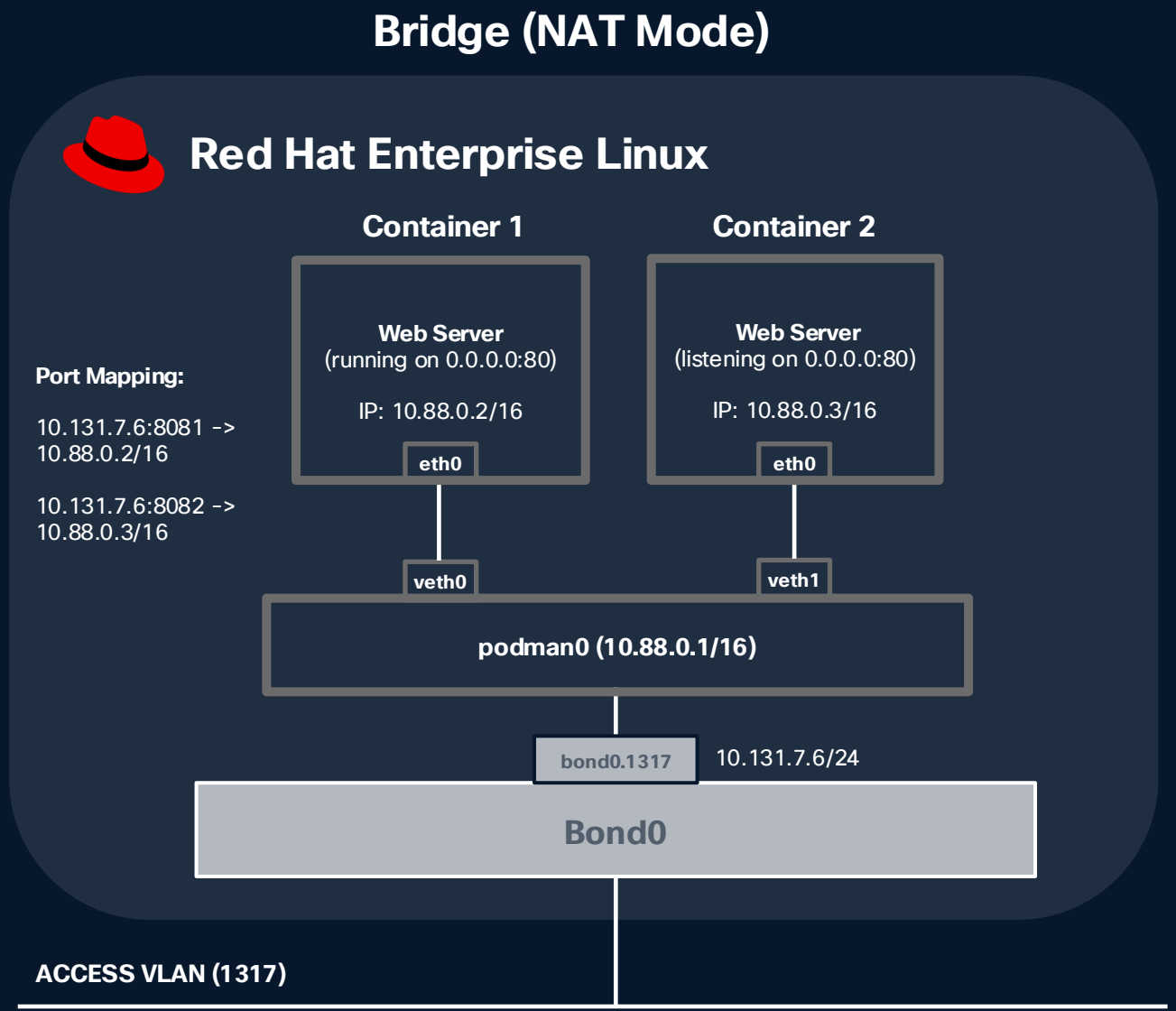
High Day 2 operations overhead

RHEL with Podman and KVM



Different management tools for Containers and VMs

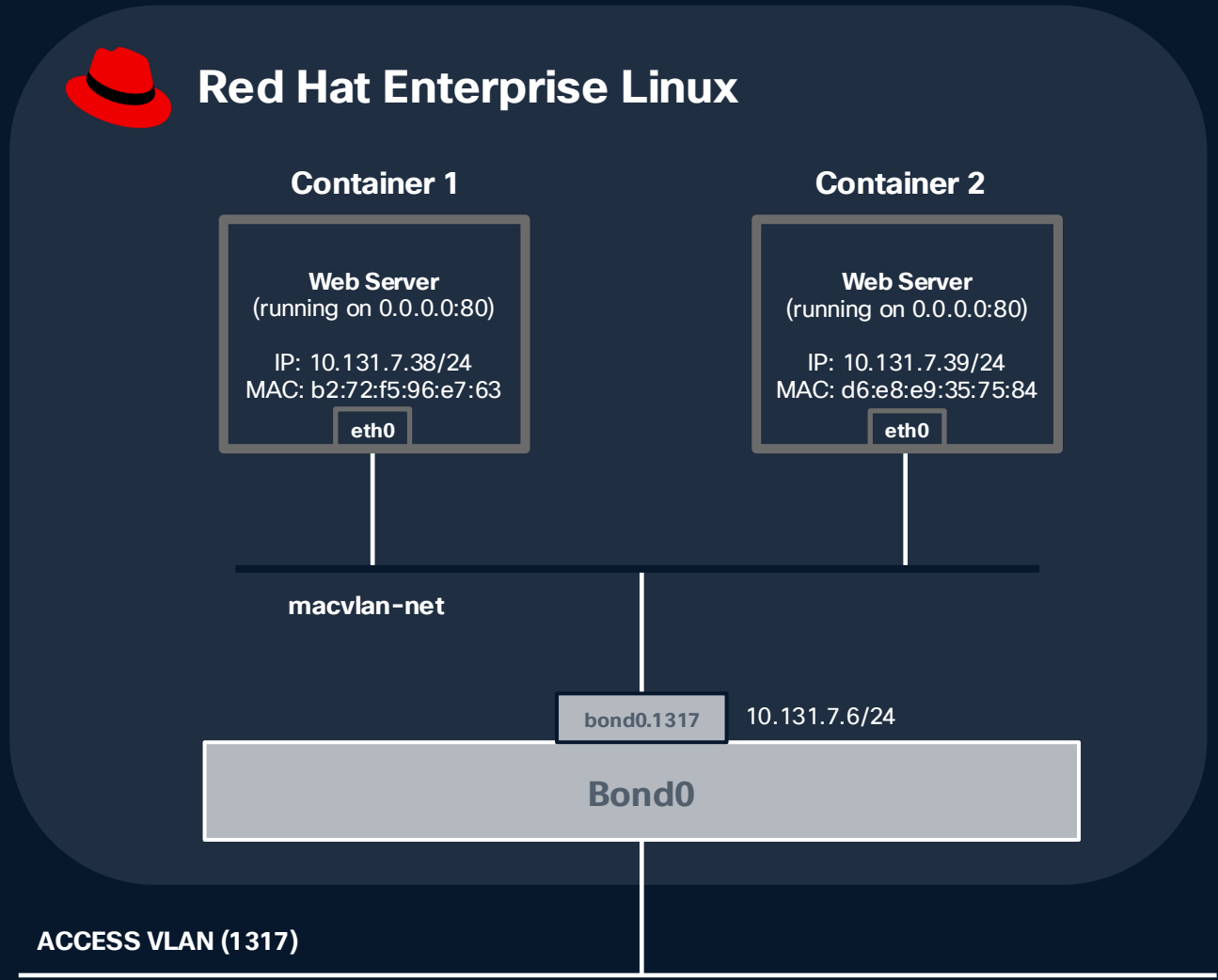
Podman Networking Bridge (NAT) Mode



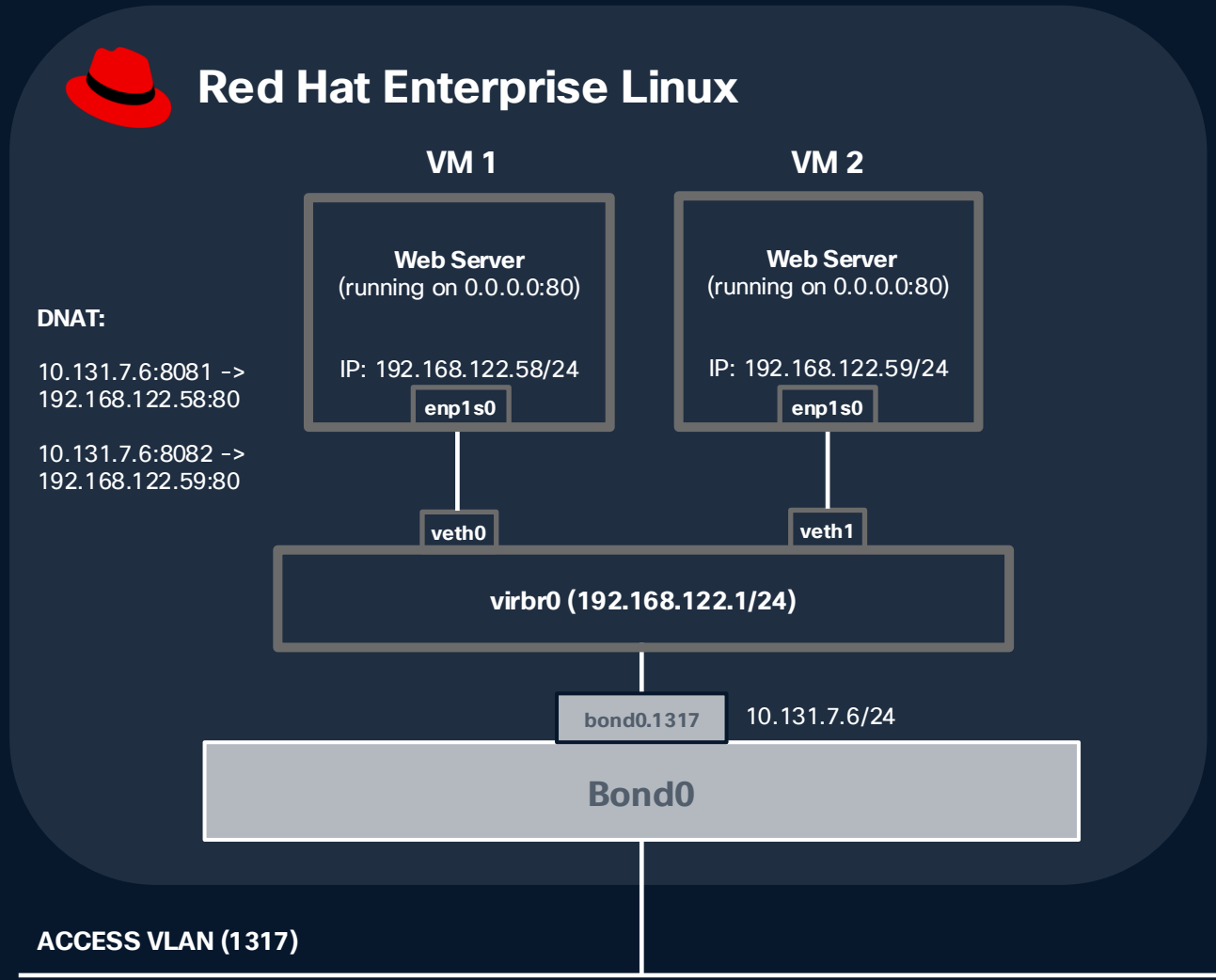
- Containers are attached to podman0 bridge
- Containers get IP addresses in the same subnet as podman0
- Internally routable within the host
- NATed when leaving the host
- Need DNAT for incoming traffic

Podman Networking MACVLAN(Bridged) Mode

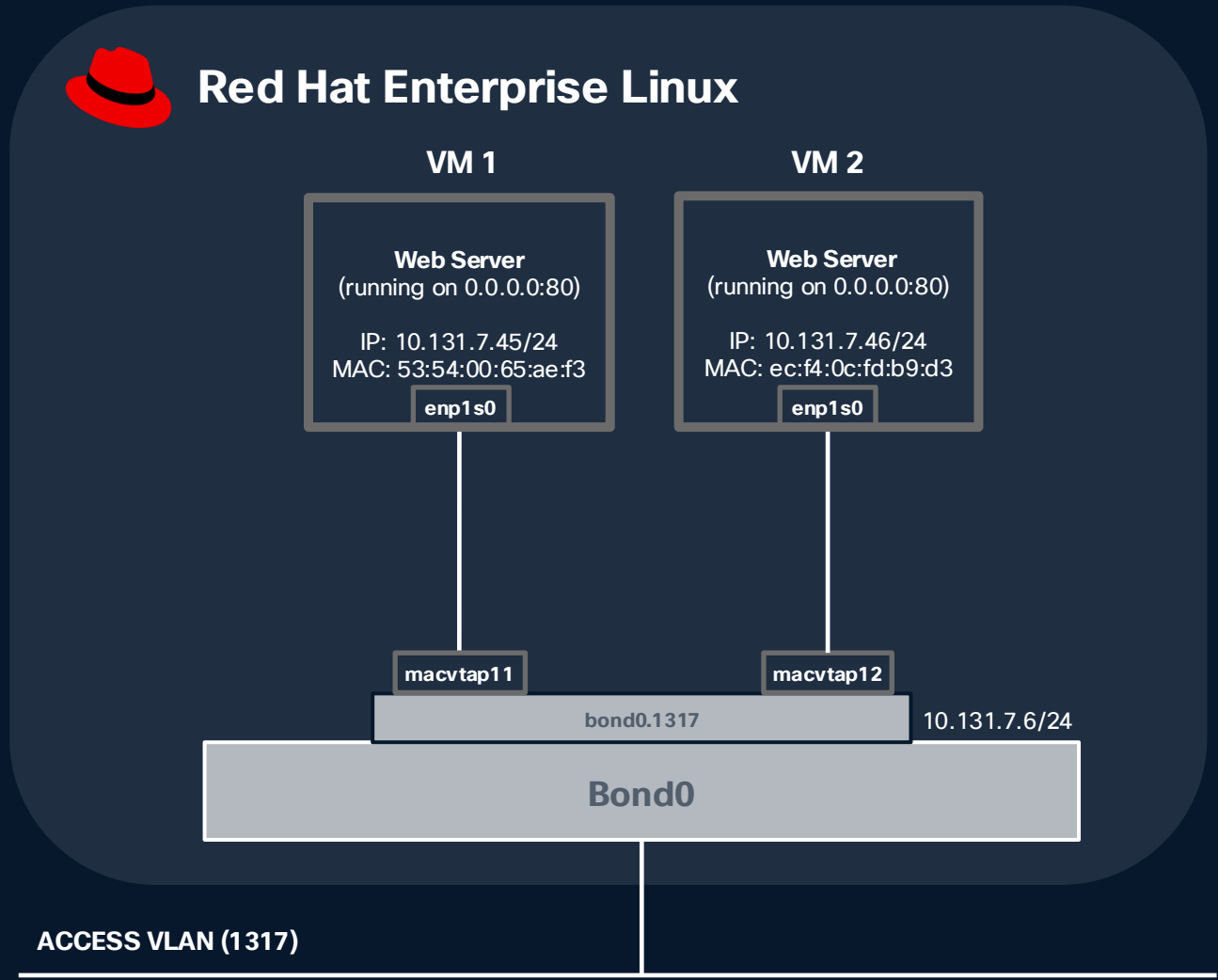
MACVLAN (Bridge) Mode



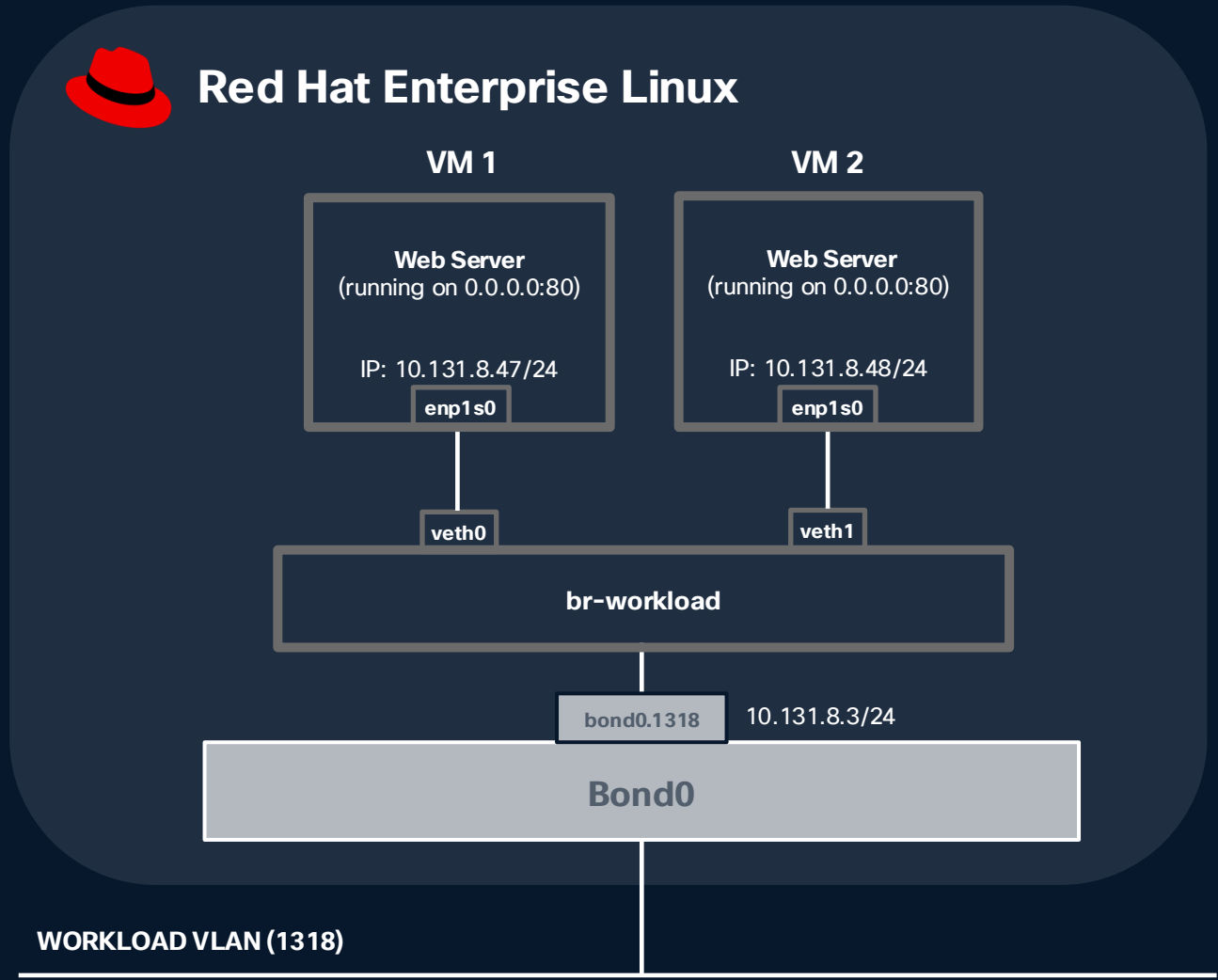
KVM Networking NAT Mode



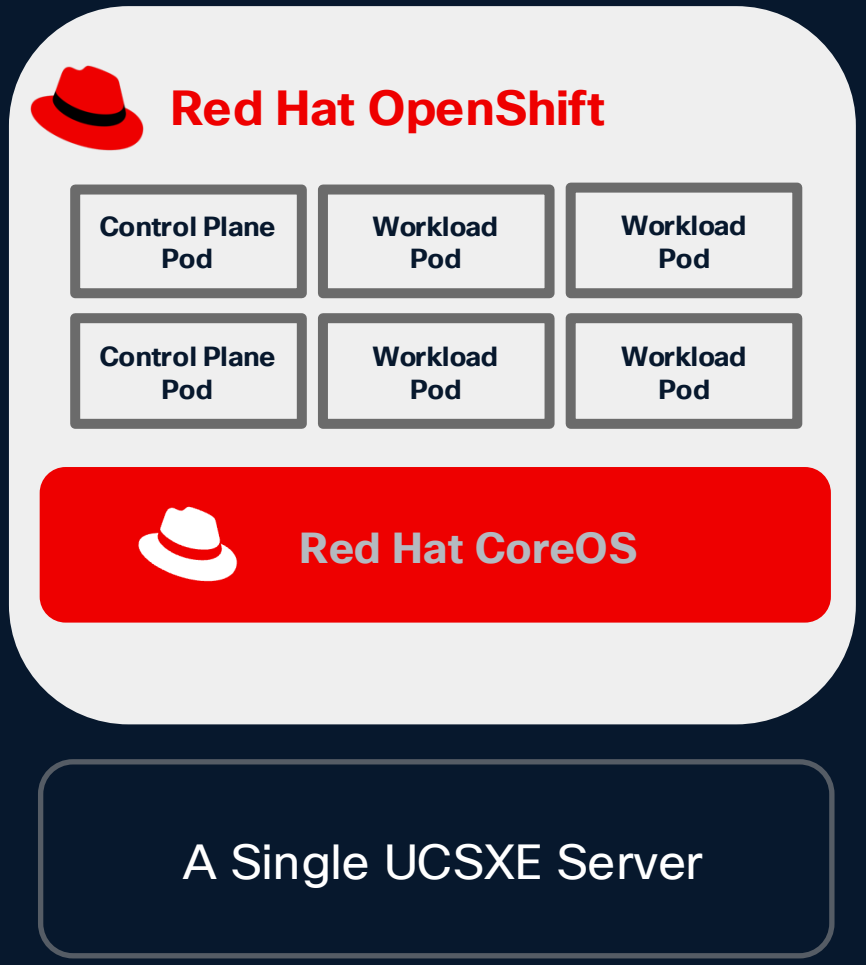
KVM Networking MACVLAN (Bridge) Mode



KVM Networking Linux Bridge Mode



Single-Node OpenShift (SNO)



Fully supported CNCF Kubernetes stack

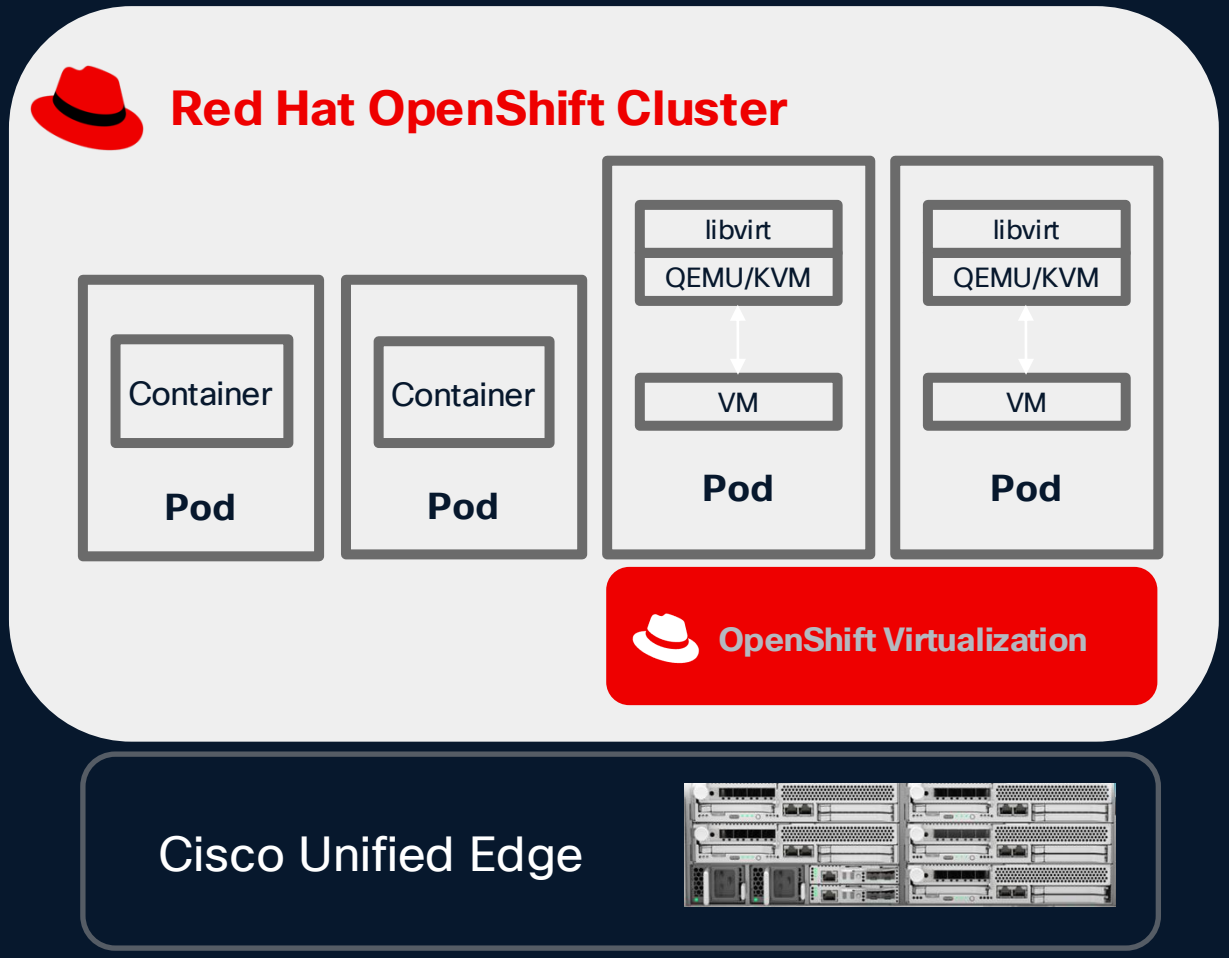
Control-plane and worker functions on a single node

Single point of failure

Workload can be restarted and scaled within the node

Support OpenShift Virtualization

OpenShift Virtualization (CNV)



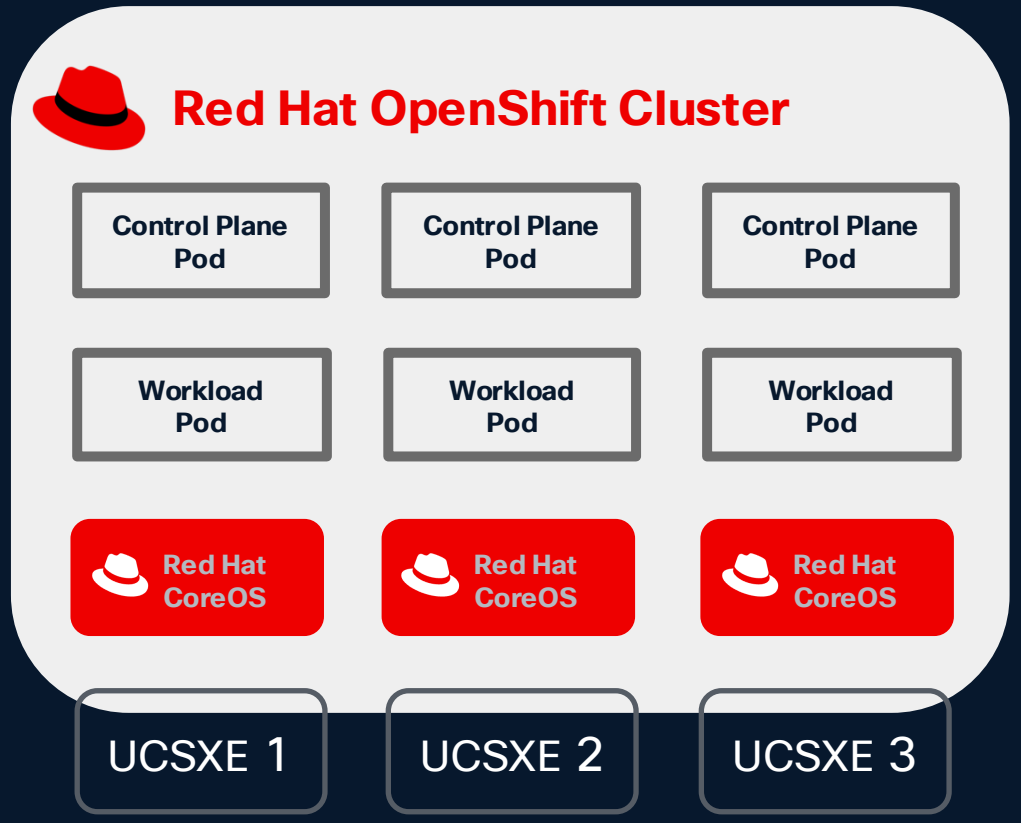
OpenShift feature built on KubeVirt

Unified platform for containers and virtual machines

- Manage VMs using native Kubernetes APIs and tools

Supports single-node, compact, and standard cluster deployments

OpenShift Compact (3-Node) Cluster



- Control-plane and worker functions on each node
- High Availability
  - Control plane tolerates of one node failure
  - Pods can be rescheduled to a different node
- Can add more work nodes for extra capacity

	Single-Node RHEL with Podman and KVM	Single-Node OpenShift Cluster	Compact OpenShift Cluster
Resource Overhead	Low	Medium	High
Scalability (System)	Limited; not designed for scale or orchestration.	Can scale by adding worker nodes	Scalable; can add more worker nodes for extra capacity.
Scalability (Workloads)	Manual scaling only	Limited by single node capacity	Can scale to cluster capacity
High Availability (System)	None	None	Full HA with distributed control plane
High Availability (Workloads)	Limited due to single point of failure	Limited due to single point of failure	Automatic rescheduling on node failure
Data Redundancy	No built-in redundancy	No built-in redundancy	High redundancy with distributed storage
Best For	Simple, small workloads; no need for high availability or orchestration	Small workloads which require unified management, where downtime is acceptable	Mission-critical workloads needing HA and data persistence.

Sizing Guidelines

- Minimal resource requirements (per node)

	CPU	Memory	Storage (SSD)
RHEL with Podman	8 cores	32 GB	250GB
RHEL with KVM	16 cores	64 GB	250GB
Single Node OpenShift Cluster (with OpenShift Virtualization)	24 cores	192 GB	250GB
Single Node OpenShift Cluster (with OpenShift Virtualization, OpenShift Data Foundation)	32 cores	256 GB	250GB

- Dedicate NVMe disk to Container/VM storages

Red Hat AI Inference Server

Red Hat AI Inference Server

- Container-based. Can run on RHEL with Podman, or on OpenShift.
- Does not rely on Red Hat OpenShift AI
- NVIDIA L4 GPU major constraint: 24GB of VRAM
  - ~7B-8B parameters
  - May require INT4 or INT8 quantization
- Tested:
  - RedHatAI/Granite-3-1-8b-Instruct-quantized.w8a8: text-to-text
- Use Red Hat validated LLM model on Hugging Face 

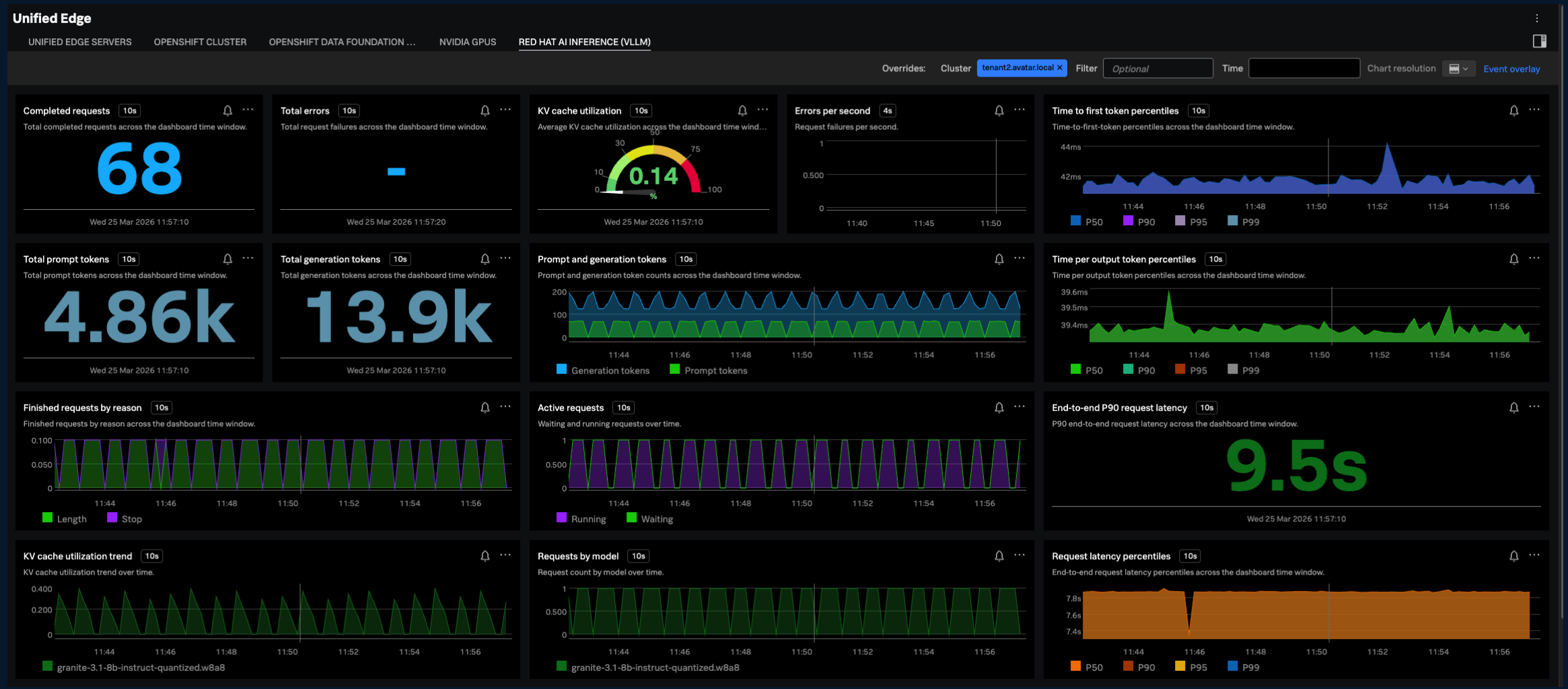
Splunk Dashboard

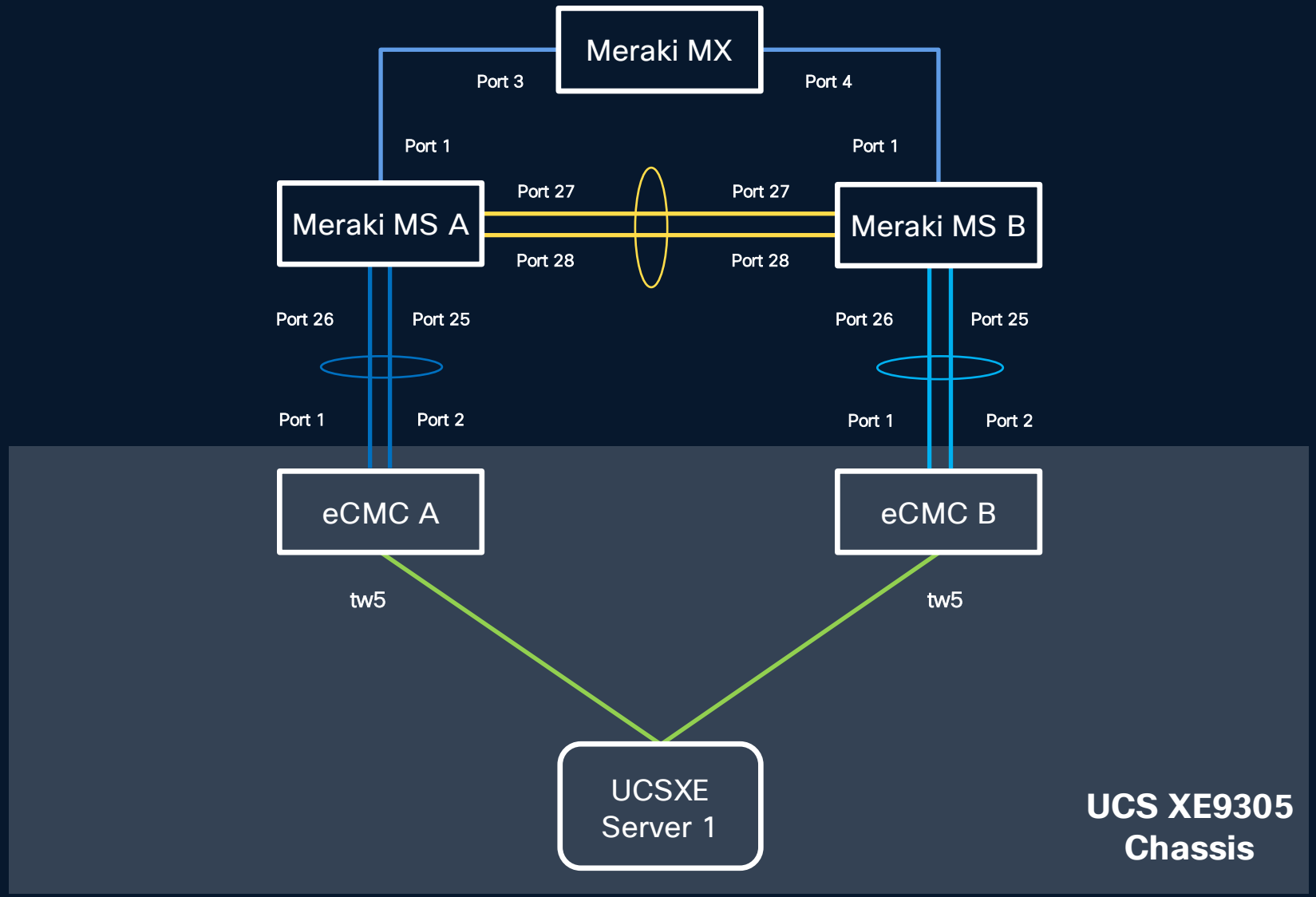
Unified Edge Monitoring

- Splunk OpenTelemetry Collector runs on OCP edge cluster
- Receivers gather vLLM, NVIDIA GPU metrics
- Processors filter and enrich metrics
- Exporters send metrics to Splunk Observability Cloud
- Dashboards provide real-time visibility

NVIDIA GPU Dashbaord

Red Hat AI Inference Server Dashboard





Demo

Demo

- Red Hat AI Inference Server
- Splunk Dashboard
- AI Agent with Intersight MCP

cisco Engage

Q&A

Thank you