Navigating the Complexities of VXLAN and IPv6 Migration for Reliable Enterprise Datacenters

Souvik Ghosh, Network Engineer

Goutam Nalamati, Network Engineer

October 2025

Meta

Agenda

- Legacy Design and Challenges
- Transition to VXLANv6
  - Network Intents
  - New Processes
- Migration Approach
- New Challenges
- Takeaways

Intro

Addressing VXLAN Transition Challenges:

- We'll explore the complexities faced by Meta Enterprise Datacenters during our shift to VXLAN fabrics while adopting IPv6.

Evolving Fleet Management:

- Discover how our approach transformed fleet management, enhancing reliability, operations and future demand.

Challenges with Legacy

Challenge-1: Architecture & Technology

Multi-Chassis LAG

- Disruptive Traffic Draining Processes
- Split-brain Issues

VLAN Translating to a Security Zone

- New services are classified to a Security Zone
- ACLs on IRB's to secure Applications

Source of Truth

- Manually configured Network devices
- Lack of Source for Truth for 'Admin' databases

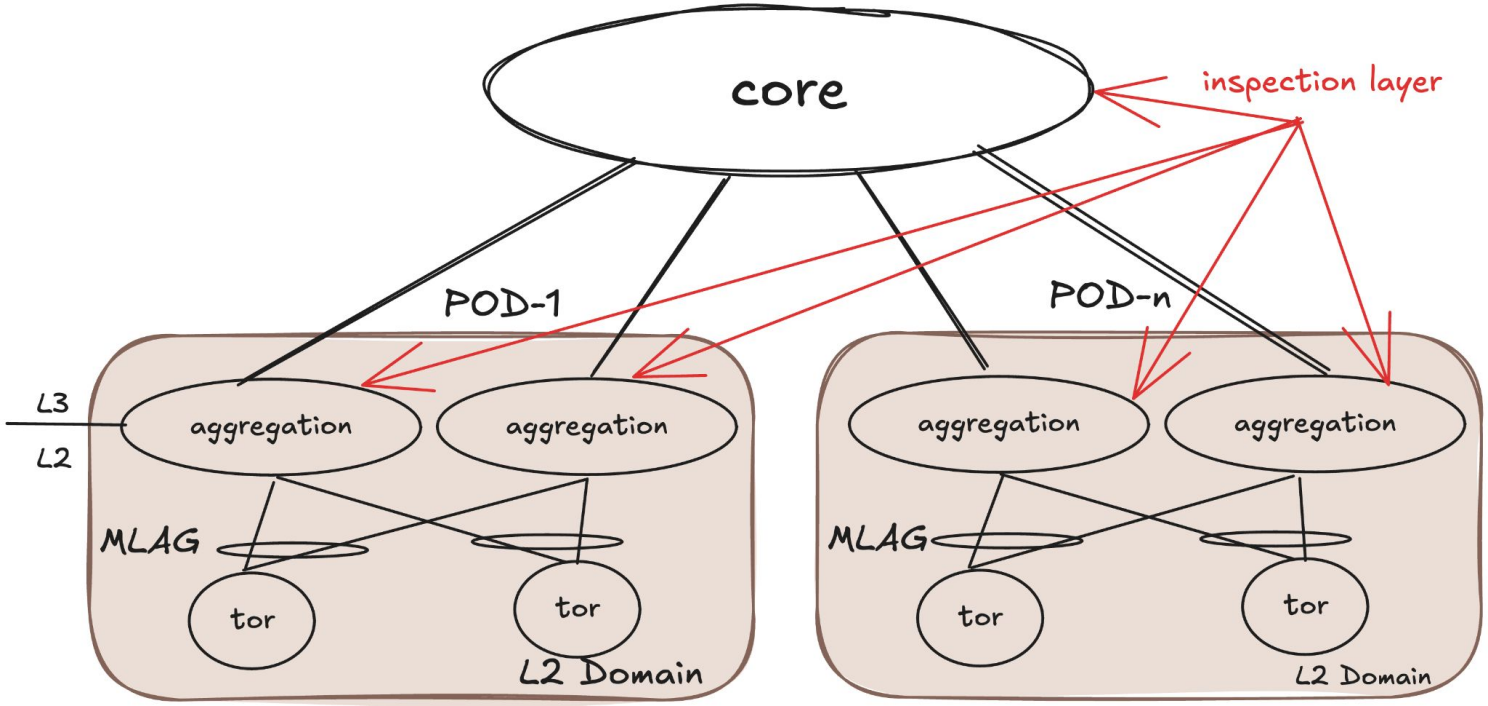
Challenge-2: IPv6 Adoption

- Running out of shared private IPv4 space between Offices, datacenters and Labs
  - Datacenters started to adopt CGNAT for services in 2018
  - Adopt IPv6 for Enterprise Applications
- Varying IPv6 implementations
  - Order of priority on SLAAC vs DHCP vs Static IPv6 usage varied by OS
  - Hosts generating IPv6 L2 Broadcast instead of Multicast.

Challenge-2: IPv6 Adoption (cont'd)

Reliability Issues

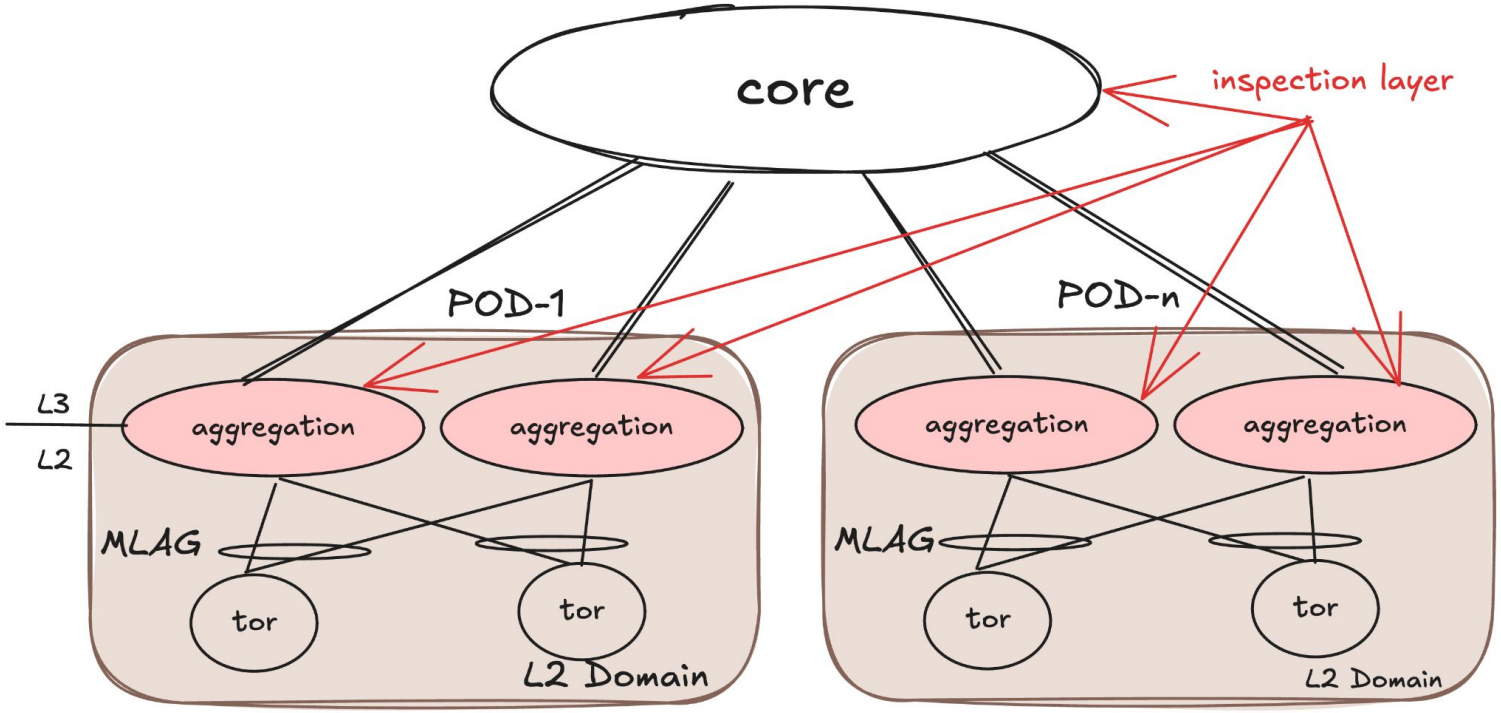
- Host IP based ACL's bloated TCAM utilizations and mis-programming
- Multi-Layered ACLs to distribute the TCAM utilizations



Challenge-2: IPv6 Adoption (cont'd)

Scalability Issues

- To support ACL scale, beefier hardware was chosen to support multiple L2 Domains



Stepping into Future: Fix the basics

NANOG™

Key Principles of VXLAN EVPN

VxLAN - Data Plane for Fabrics

- Allows extending L2 Networks over L3 Infrastructure
- Encapsulates the traffic with VTEP source and destination and leverages VNI for identifying virtual Network Segments

EVPN - Control Plan for Fabrics

- Efficient MAC address Learning
- Ideal solution for Multi-Tenant Environments

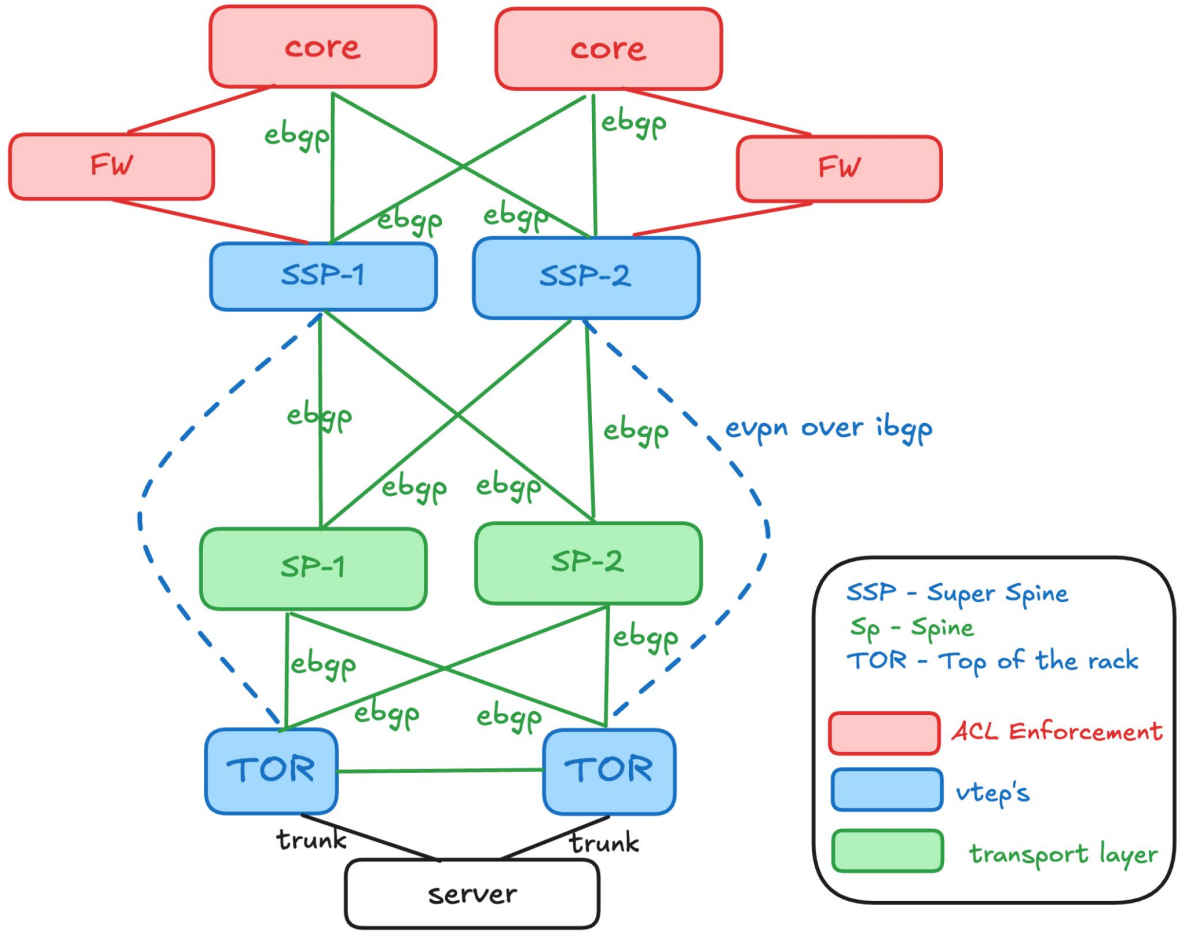
Enhancement-1: Scalable Architecture

Support Large Scale Demands

- Vendor Agnostic Design
- Symmetric IRB
- 5-Stage CLOS
- ≈480 VTEPs

Traffic Filtering

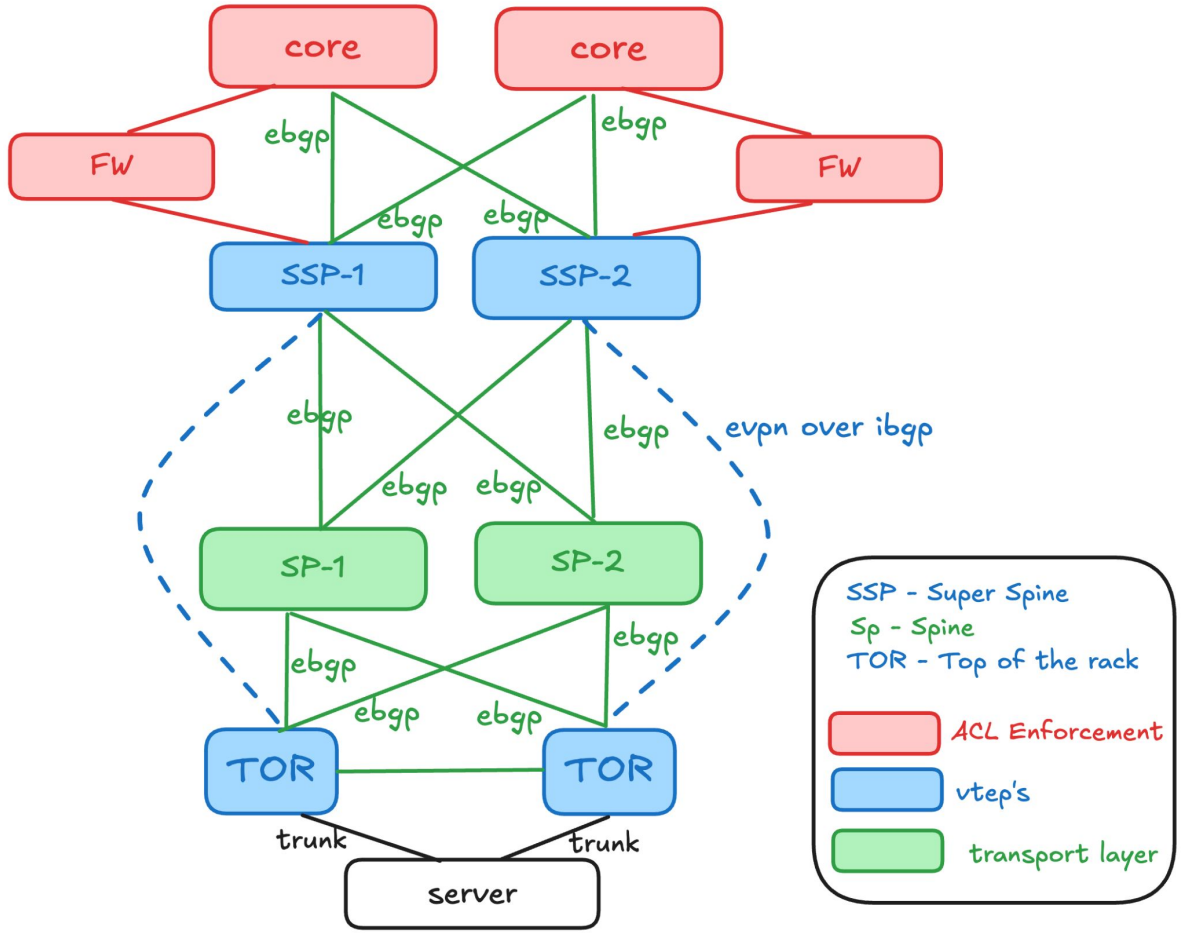
- Consolidated Inspection Layer
- Simplified ACL management



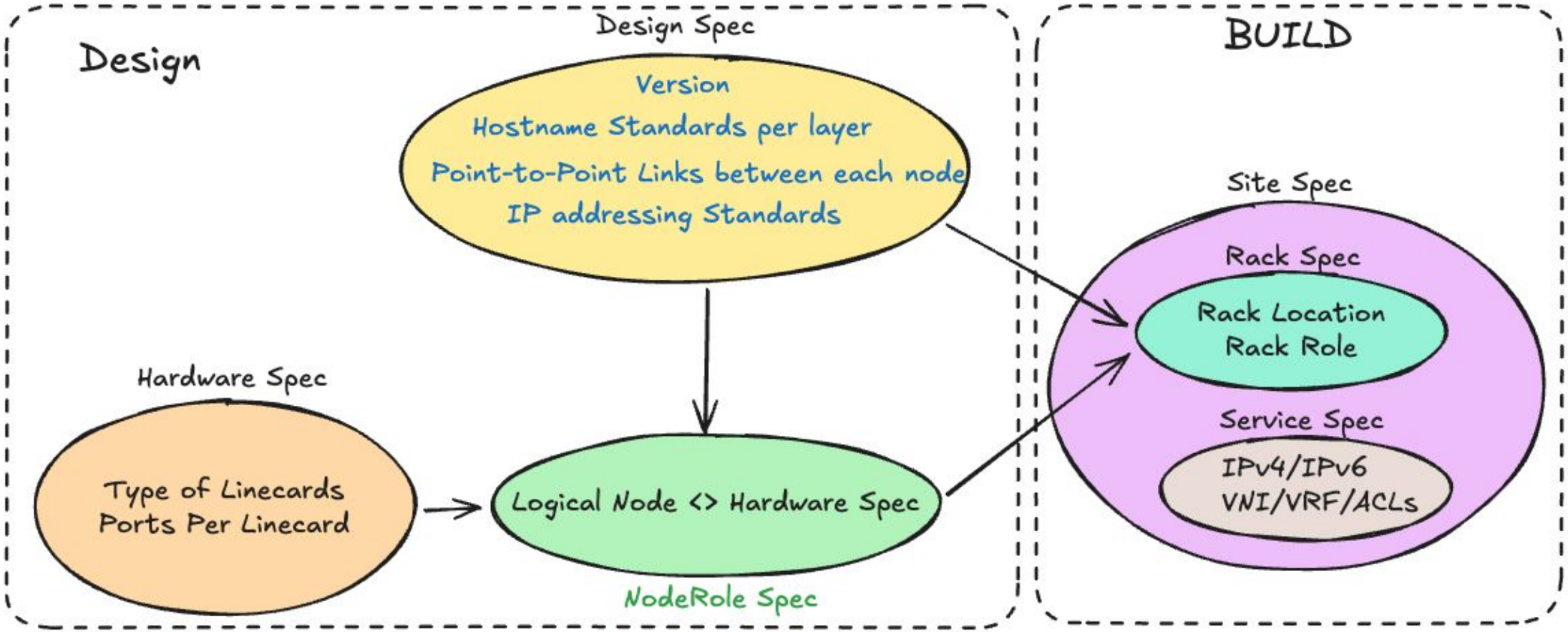
Enhancement-1: Scalable Architecture (cont'd)

IP Management

- /64 per service
- VRF translates to Security Zone Instead of VLAN
- Multiple VLANs per VRF

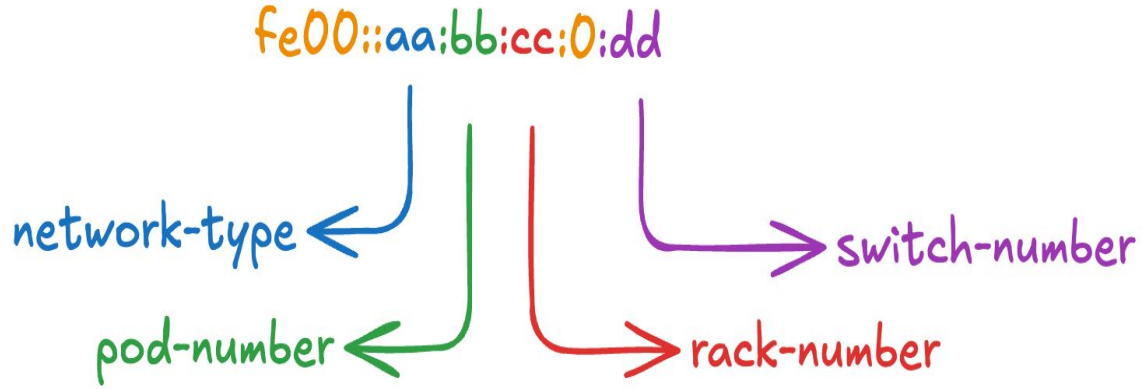


Enhancement-2: Managed Source of Truth



Enhancement-3: Deterministic Addressing

- Sequential IP allocations are not deterministic
- Device Location, CLOS Tier, Role are encoded into IPv6 Hextets of Loopback and P2P addressing

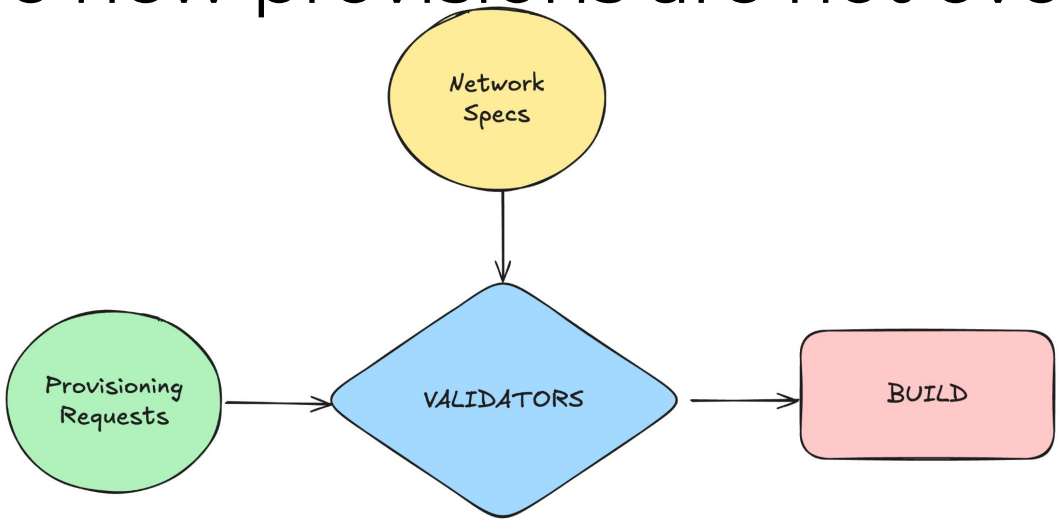


- Provided Information about the network node by look of an IP Address and Simplified Automation

Enhancement-4: Auto-Validations

Auto-Validation

- “Specs” enabled critical Validations before Provisioning
  - Ensuring new IP allocations are correct during the topology build
  - Signaled Utilization of IP Networks and Rack Densities during Expansions
  - Ensured the new provisions are not overlapping



The journey.....

- **2020** - Initial Proposal
- **2021** - First Datacenter Pilot with Intent 1.0
- **2022** - New Enterprise Datacenters with VXLAN EVPN
- **2023** - Design Adoption by partner teams
- **2024**
  - AI DC demand
  - Migration efforts kicked of in Legacy Datacenters
  - Intent 2.0 Landed
- **2025**
  - Partner Team Migrations

Bigger Step:
How to Migrate??

Challenges and Complexity

Internal Customer Requirements

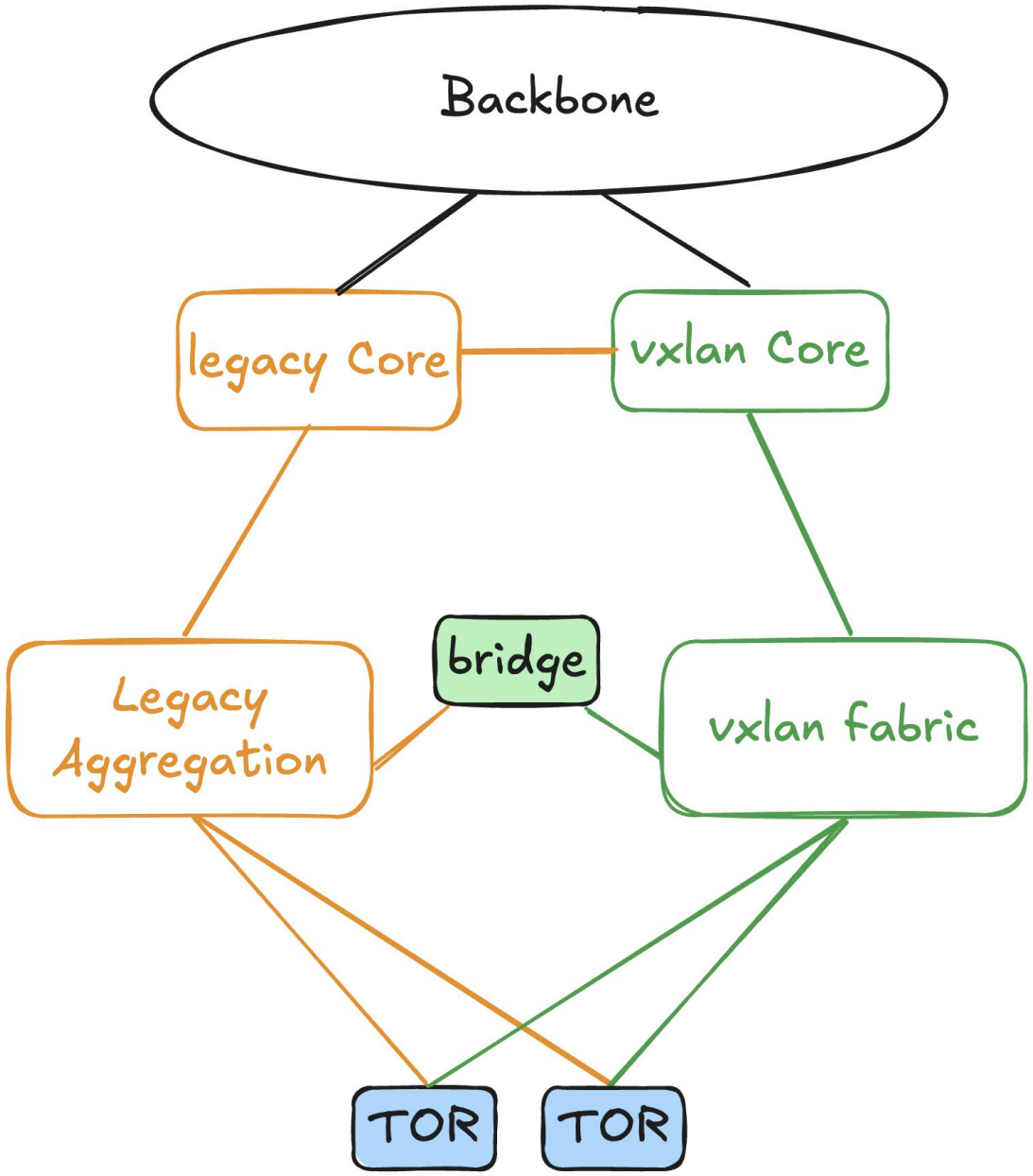
- Application Owners do not want to Re-IP or re-build VM's into new Datacenters
- Minimal Disruption to services

No Standard Migration Process

- Vendors-specific process or knobs builtin to support the migrations
- Previous Hardware not able support required VXLAN capabilities

Bridge Layer

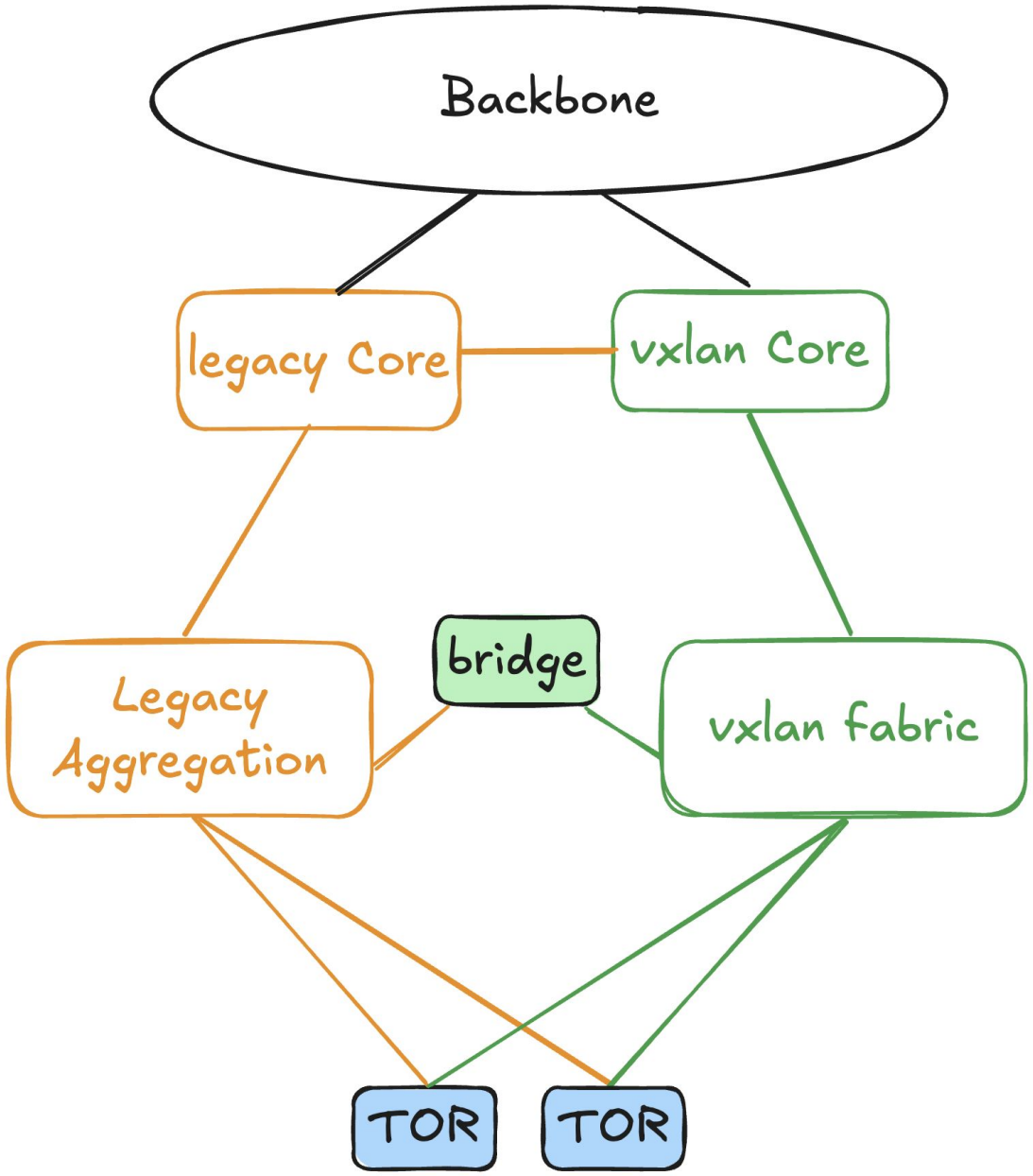
Device that is connected to both Legacy and VXLAN networks to act as temporary gateway during the migration



Bridge Layer

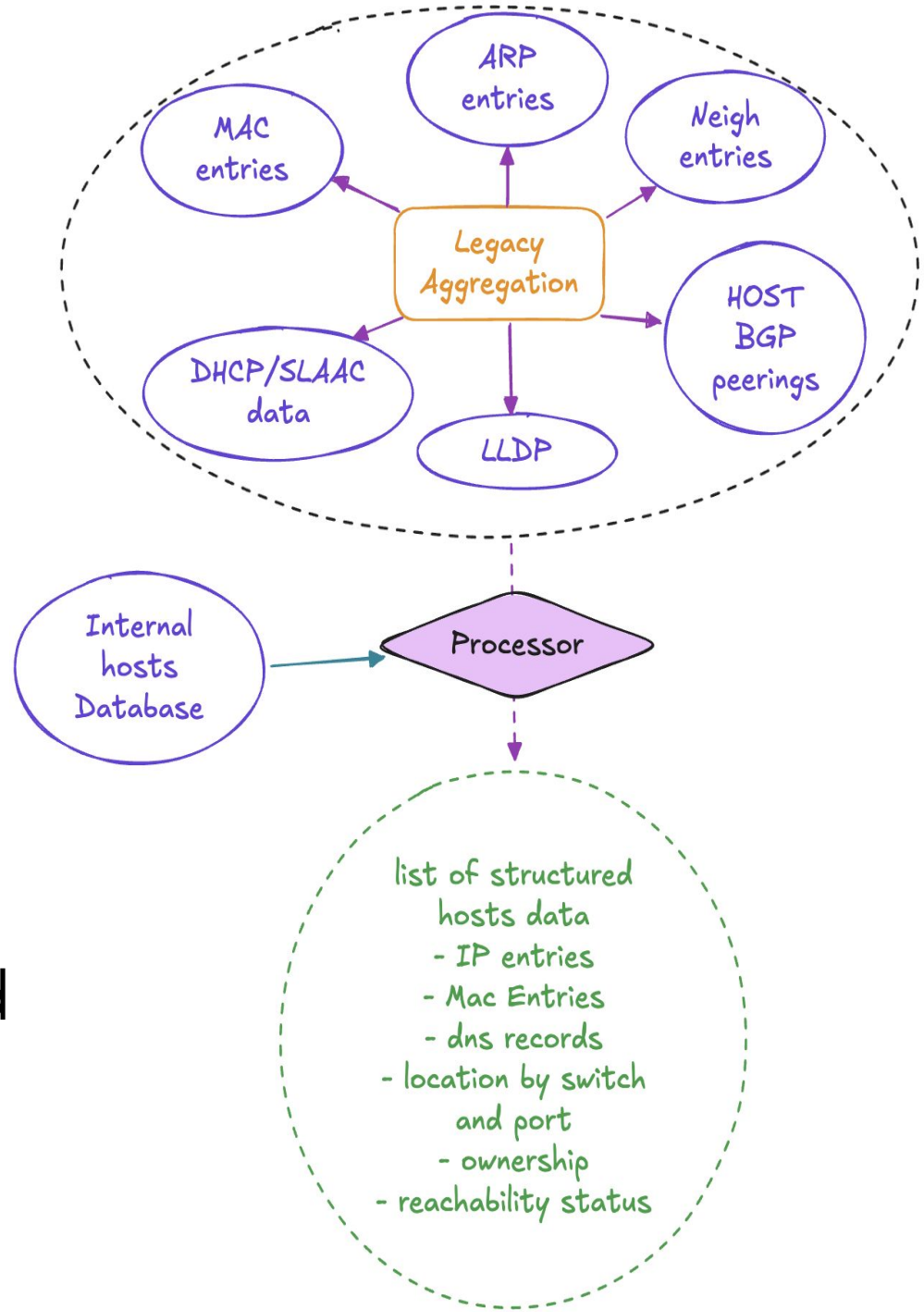
Required Capabilities

- Adherence to established VXLAN RFC's
- Operator ability to login to device Shell Mode
- Ability to install Python Modules
- Device capable of executing of custom Python Scripts



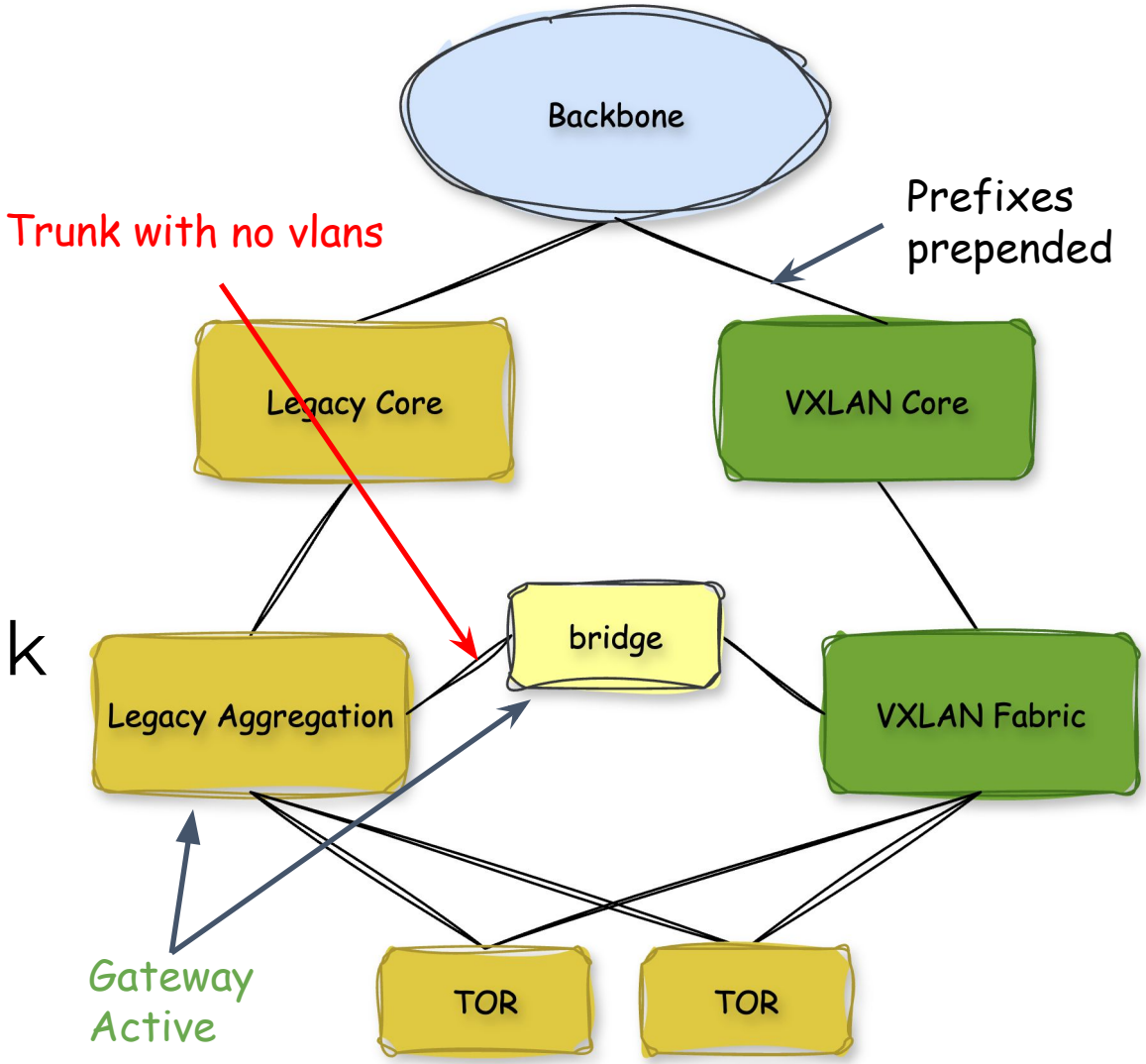
Pre-Migration-Collections

- Automated Data Collection:
  - Maps network data to internal databases.
  - Processes data to build new structured information.
- Enabling Application Owners to:
  - Identify active hosts by ownership.
  - Validate host status (before and after migration).



Pre-Migration - Network Staging

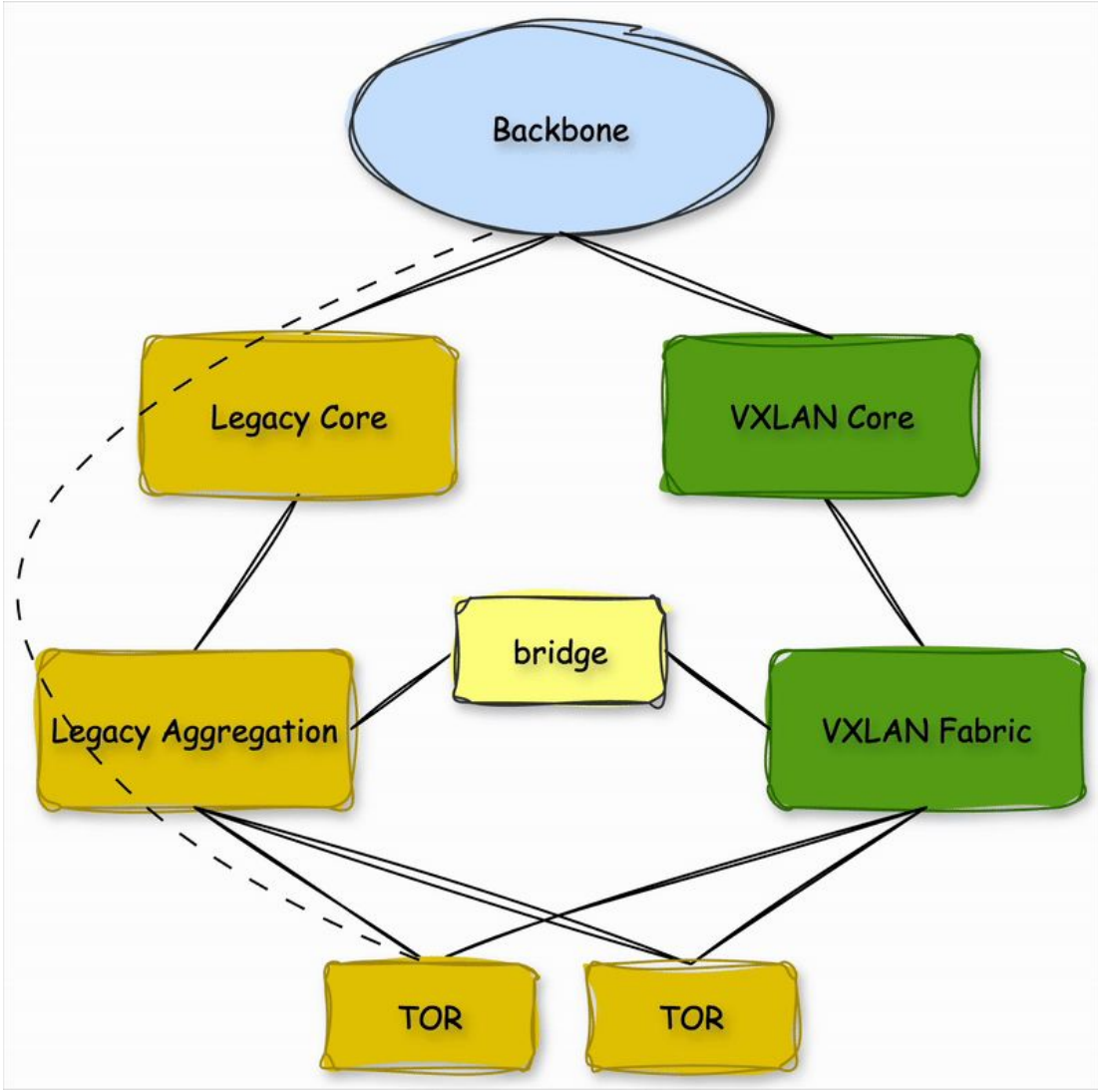
- Temporary Gateway:
  - Enable Bridge Layer between VXLAN and Legacy
- VLAN not allowed on the Trunk
- Fabric VNI Enablement:
  - Configure VXLAN Network Identifiers for all VLANs and VRFs.



Pre-Migration - Network Staging

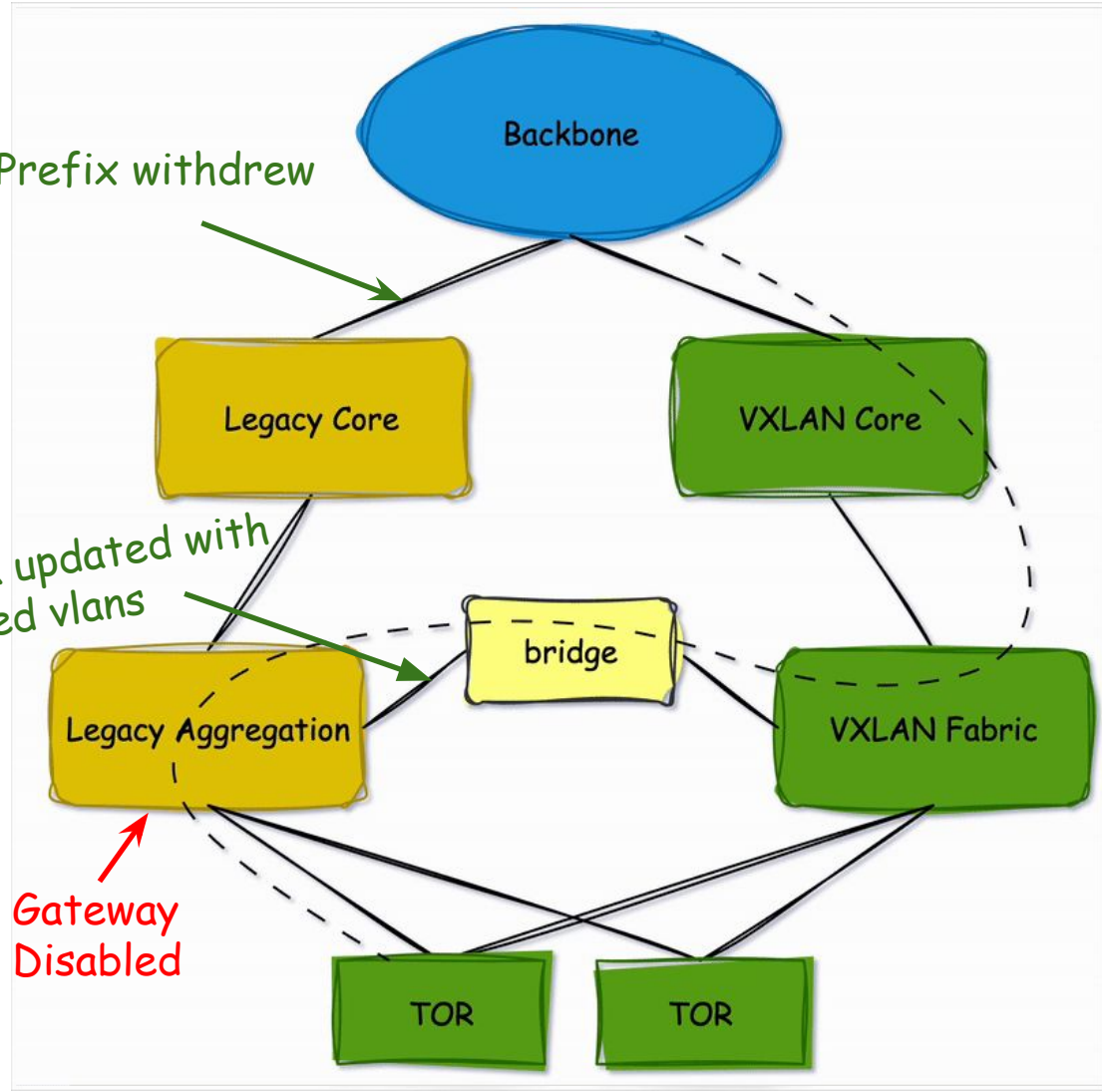
Scapy generates packets every 3 seconds on Active SVI's of Bridge Switch

- GARP for IPv4 Gateway
- Unsolicited Advertisements for the Gateway's Global IPv6 address



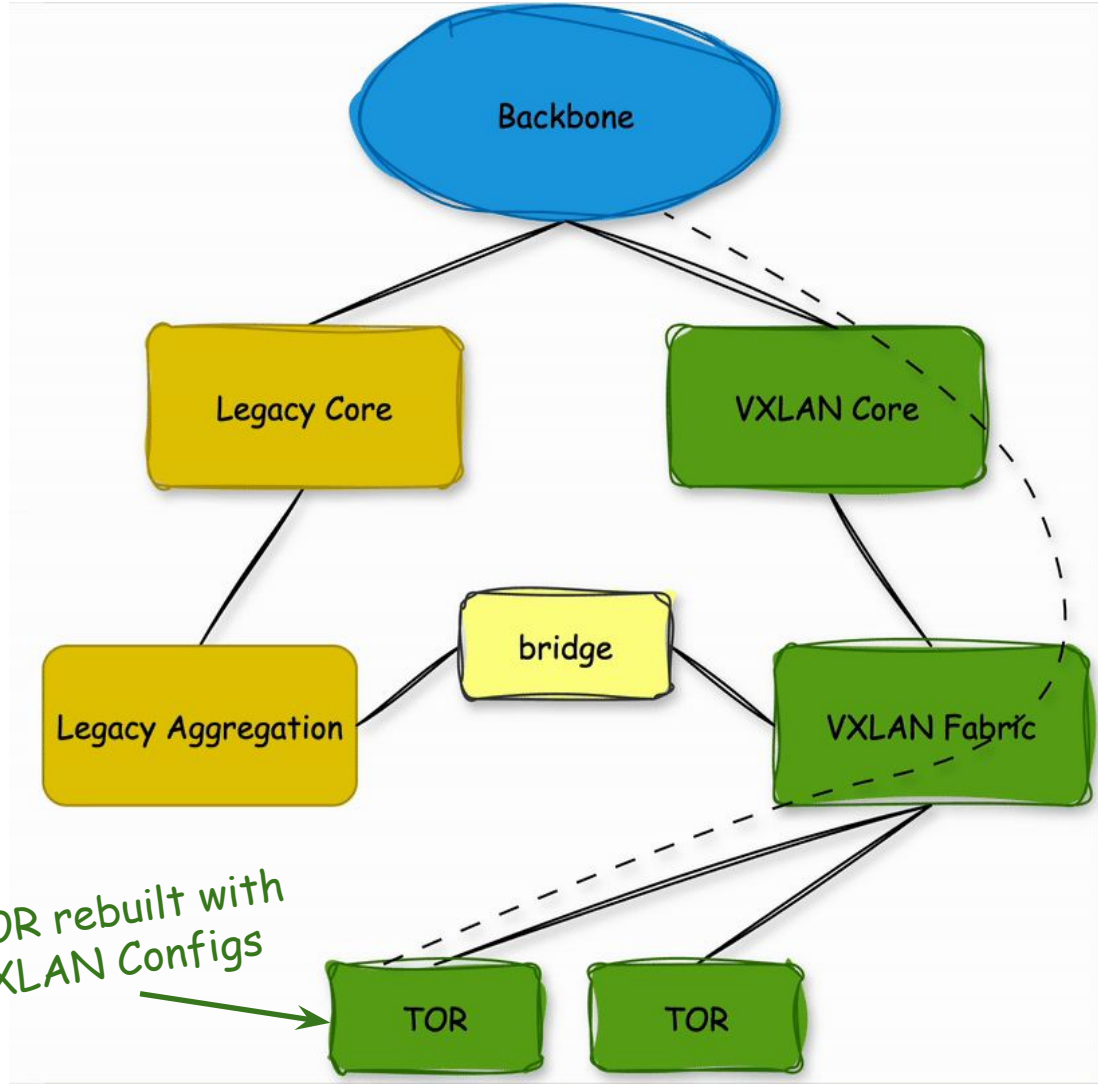
Migration Execution - Network

- Establish active ssh connections to Bridge and Legacy Aggregation devices.
- Disable Legacy network Gateway.
- Allow VLAN on trunk interface.
- Scapy generated packets traverse trunk, hosts update MAC address for new gateway



Migration Execution - Top Of Rack

- Leverage automation to update the database with new logical nodes and generate/stage VXLAN configurations.
- Upgrade the device image via the "Upgrade" workflow.
- Replace the entire old configuration using the "Replace Functionality."



Opening new “Can of issues”

Challenges with VXLAN & IPv6

Changes to ECMP Hashing:

- Traffic started to use L3 headers data for hashing
- Stateful IPv6 sessions were broken for Anycast destinations because of Aggressive Flow Label Change in IPv6 headers

Solutions

- Understand the default Hashing in different hardware and keep it consistent across network device in multi-vendor setup
- Disable the aggressive IPv6 hashing on packet retransmit

Challenges with VXLAN & IPv6 (cont'd)

Convergence Delays and Drops:

- Number of BGP entries are based on the EVPN type routes and number of Super Spines
- As networks grew, BGP Convergences are long
- Maintenances caused the discards by TOR's

Solutions

- Enable BGP to monitor peerings with Super Spines
- Understand the convergence time based on network size
- Access Ports in '*disabled*' state until one EVPN peer is UP
- Delay the turn-up of access ports after one Peer is UP

Challenges with VXLAN & IPv6 (cont'd)

BGP peering with Gateways

- VM's peer with Anycast GW
- Silent Mobility of Host for capacity rebalancing caused BGP sessions to flap with Anycast GW

Solutions

- Disabling the Host Mobility per Application type
- Support different Gateway Architectures like
  - Centralized Gateway - For BGP peering
  - Anycast Gateway - For all other services

Challenges with VXLAN & IPv6 (cont'd)

Lifetimes during Neighbor Discovery

- Smaller lifetimes or default values by vendors in Neighbor Discovery are not consistent
- Devices unable to connect to Network

Solutions

- Static Lifetime configuration for Prefix & DNS information
  - DNS lifetime to 1800 sec
  - Prefix lifetime to 3600 sec

What this enabled for Future?

Increasing Infra Reliability and Security

Application Storms

- “Service Spec” contains information about resiliency of all application in a Datacenter
- Modification of Status in a “Service Spec” Enabled Application to execute a DR Exercise
- Can be executed only by an App Owner/DR Engineer

```
2915: VlanData(
    networks=Network(
        ipv6=['2620:10d:c0a8:82::/64']),
    properties=VlanProperties(
        security_zone='MSEC-IFS',
        dhcp=True,
        router_advertisment=True,
        slaac=True,
        anycast_gateway=True,
        firewall_type=1,
        is_stormed=False,
        name='KUBE-NODES'),
    peering_services=[PeeringService(
        name='kubernetes',
        service_networks=Network(
            ipv6=['2620:10d:c0a8:83::/64']),
        asn=64914)]),
```

Increasing Infra Reliability and Security

Enforcer

- Hosts failing to meet the compliance requirements are identified in the “Service Spec”
- Automated Tooling blocks all the traffic to those MAC Addresses across the network

```
Mac_Enforce(
mac_address='52:54:00:D3:09:6E',
block_category=4,
block=False,
vlan_id=3319,
switchport='ash6-s21p02-p2r11-rsw2'),
```

Support AI cluster demand

- New demand emerged in 2023 to build the AI Clusters in Enterprise Datacenters
- Same 5-Stage CLOS design was leveraged
- Defining the new “Hardware Spec, Edge Specs” and mapping to specific DC, enabled build AI Clusters in Enterprise Network

Takeaways

Invest More in Foundations

Simplicity is the Key.

- Design with scalability in mind—ensure overlays, underlays, and control-plane scale match future growth.

Deterministic-Approach while designing the IPv6 Networks

- Encode the Real-World information into IPs

Invest More in Foundations (cont'd)

Establish Ecosystem:

- Authoritative Tool managing Source of Truth for VNI/VLAN/VRF mappings
- Build Zero Configuration delta between Expected and Actual configurations using Network Intents

Host provisioning

- Static IPv6 allocations for Host provisioning gave consistent results

Invest More in Foundations (cont'd)

IPv6 Subnet Planning

- Large IPv6 host addressing availability can throw a curveball
- Avoid IPv6 masks greater than 64+
- Aggregation plays a key role as IPv6 Networks grow

Configuration Consistency is Key

- With symmetric IRB, Neighbor discovery parameters need to be consistent across all the VTEP's

Thank you

NANOG™