TOMORROW starts here.

CISCO

Cisco live!

Evolution and Challenges of Data Centre Network and Host-Based Overlays

BRKDCT-2328

Victor Moreno

Distinguished Engineer

Agenda

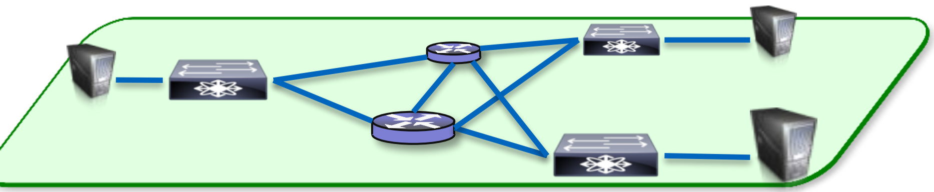
- Overlay Foundational Principles and evolution
- Mapping overlay technologies to the network
- The role of the underlay
- Management and orchestration




Foundational Principles of Network Overlays

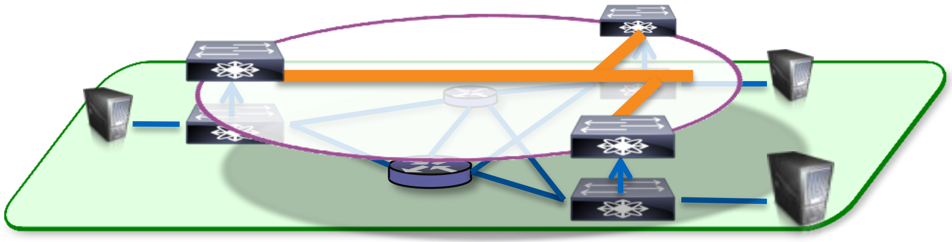
Why Overlays?

Seek well integrated best in class Overlays and Underlays



Robust Underlay/Fabric

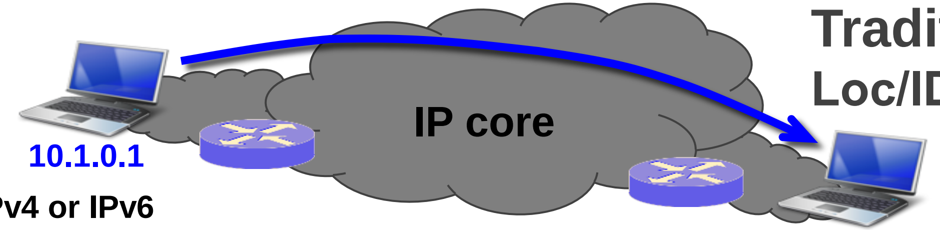
- High Capacity Resilient Fabric
- Intelligent Packet Handling
- Programmable & Manageable



Flexible Overlay Virtual Network

- Mobility – Track end-point attach at edges
- Scale – Reduce core state
  - Distribute and partition state to network edge
- Flexibility/Programmability
  - Reduced number of touch points

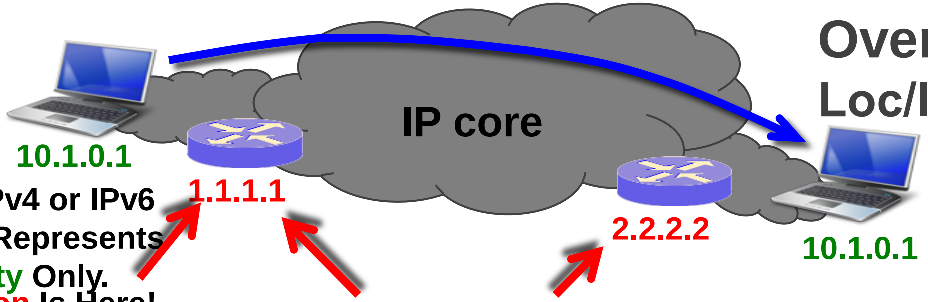
Seminal Idea: Location and Identity Separation



Traditional Behaviour
Loc/ID "Overloaded" Semantic

Device IPv4 or IPv6 Address Represents Identity and Location

When the Device Moves, It Gets a New IPv4 or IPv6 Address for Its New Identity and Location



Overlay Behaviour
Loc/ID "Split"

Device IPv4 or IPv6 Address Represents Identity Only.
Its Location Is Here!

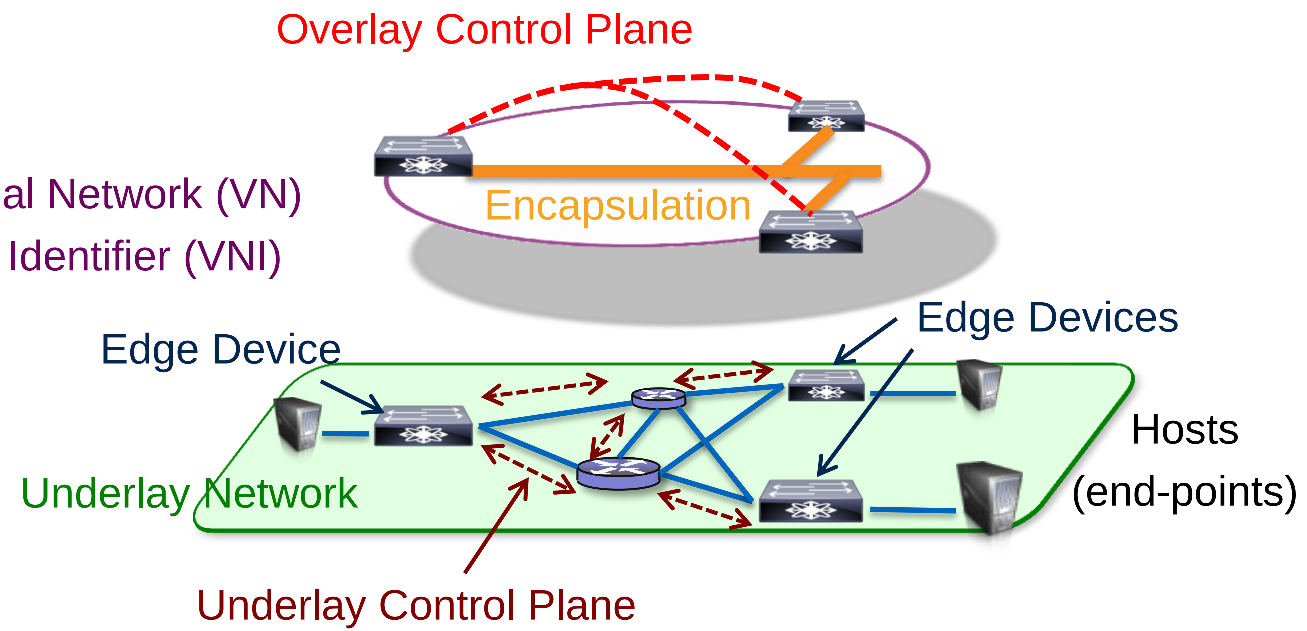
Only the Location Changes

When the Device Moves, Keeps Its IPv4 or IPv6 Address.
It Has the Same Identity

Overlay Taxonomy

Service = Virtual Network (VN)

Identifier = VN Identifier (VNI)



Overlay Attributes

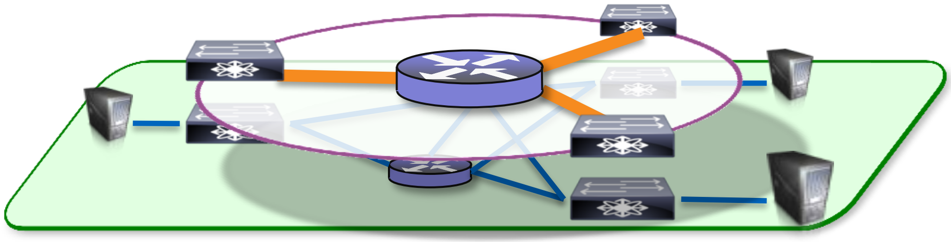
Service	Edge Device	Signalling
Layer 2 Service Layer 3 Service	Host Overlays Network Overlays	Data Plane Learning Control Plane Learning

Overlay Service Type Evolution

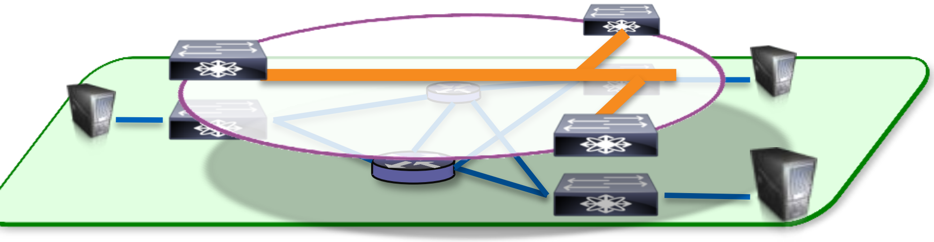
Service

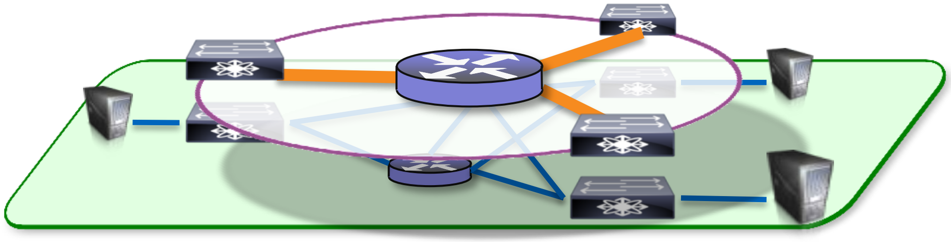
Layer 2 Service

Layer 3 Service



Types of Overlay Service





Layer 2 Overlays

- Emulate a LAN segment
- Transport Ethernet Frames (IP and non-IP)
- Single subnet mobility (L2 domain)
- Exposure to open L2 flooding
- Useful in emulating physical topologies

Layer 3 Overlays

- Abstract IP based connectivity
- Transport IP Packets
- Full mobility regardless of subnets
- Contain network related failures (floods)
- Useful in abstracting connectivity and policy

Hybrid L2/L3 Overlays offer the best of both domains

Layer 2 Overlay Considerations

Solved with …

- **Scale** of the edge devices
  – L2 addresses in Ethernet (MACs) use a flat space which cannot be summarised
- **L2/L3 boundary** scaling
  – Large L2 domains require a large capacity L3 gateway to handle large ARP and MAC tables at a frequent rate of refresh
- **Multi-homing** sites can induce loops in the network
- **Flooding** of L2 protocols, unknown unicasts and broadcast in general can propagate failures across the entire L2 domain

Problem	Solved with …
Scale of the edge devices	Layer 3 Overlays
L2/L3 boundary scaling	Layer 3 Overlays
Multi-homing	Network Overlays
Flooding	MAC routing

Multi-homing in L2 Overlays

Source learning assumes single attached sites

But network overlays involve edge resiliency

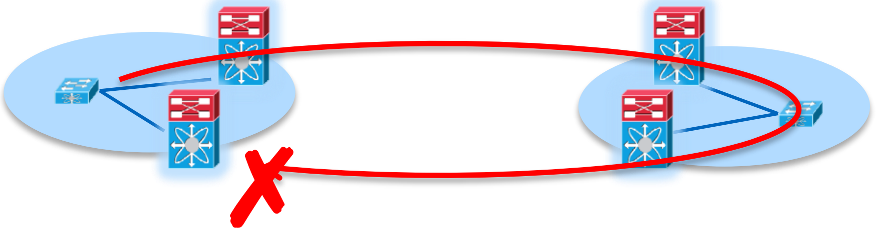
Enhancements are required to address:

- Loop resolution
- Multi-pathing
- Broadcast/Multicast de-duplication

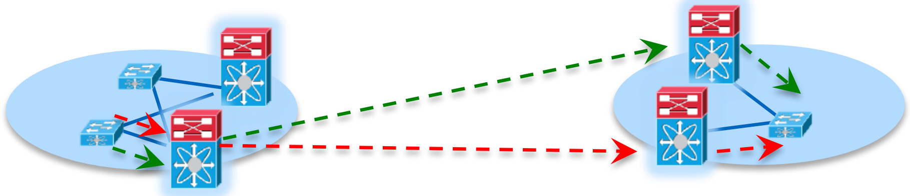
Two Approaches:

- Active-Standby (Data Plane or Control Plane)
  - One active device per VLAN (single attached site)
  - VLAN based load balancing
- Active-Active (Control Plane only)
  - One active device for multi-destination traffic
  - Intra-VLAN load balancing for unicast

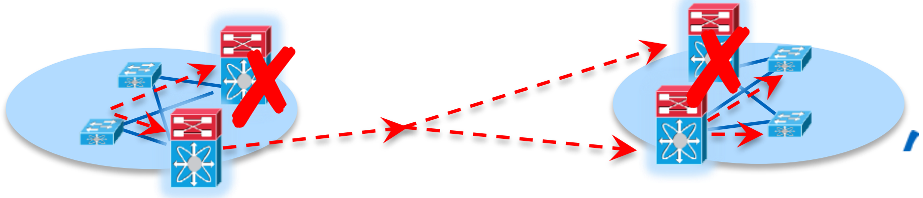
- Loop resolution



- Multi-pathing



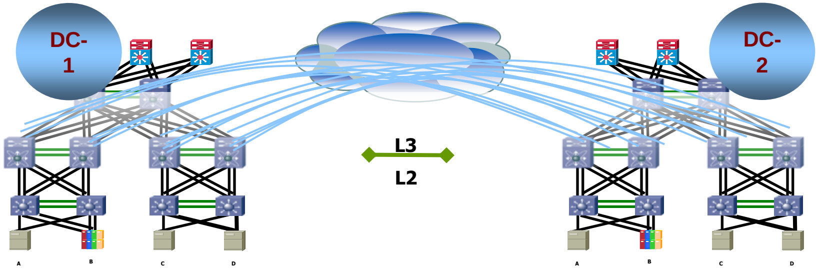
- Broadcast/Multicast de-duplication



Flooding in L2 Overlays

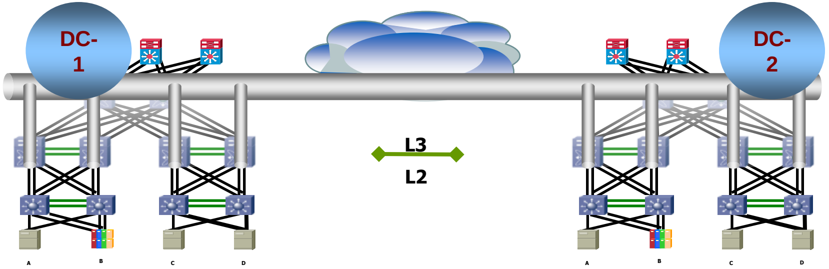
Control Plane Signalling eliminates the need for floods

Data Plane Learning



- Pre-set flood facility
- MAC learning based on flooding
- Flood L2 protocols and unknown unicast
  - ➔ Failure propagation
- Fail Open
- Suitable for small domains (failure scope)

Control Protocol



- No predetermined flood tree
- MAC learning by control protocol
  - ➔ Contain Failures and L2 protocols
  - ➔ Rich information
- Fail Closed
- Better suited for broad scope

Flooded L2 Overlays ➔ MAC Routing

L2 Overlay Evolution

Inter-DC (DCI)

VPLS → A-VPLS
Flood and Learn

→

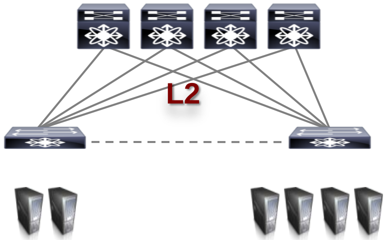
OTV
MAC routing (IS-IS)

→

EVPN / L2-LISP
MAC routing (BGP/LISP)

Backbone Network

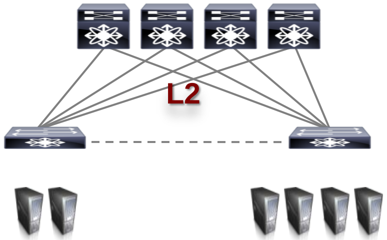
Intra-DC (Fabric)



FP → TRILL
Flood and Learn

→

Enhanced Forwarding (hybrid)
Control Plane



Host Overlays

VXLAN
Flood and Learn

→

Layer 3 Overlay Considerations

Addressed with …

- **Scale** of the edge devices → **On-demand Pull**
  - Can be improved further by using an on-demand pull model
- **IP Mobility** for subnet disaggregation → **Layer 2 Semantics with IP routing**
  - Members of a subnet may be distributed across locations
  - Any host anywhere
- **Broadcast & Link-local multicast** traffic to be handled as a special case → **Combined L2/L3 overlay**
  - Potentially without even learning MAC addresses

L3 Overlay Evolution

Edge Device Scale

Push Protocol Model

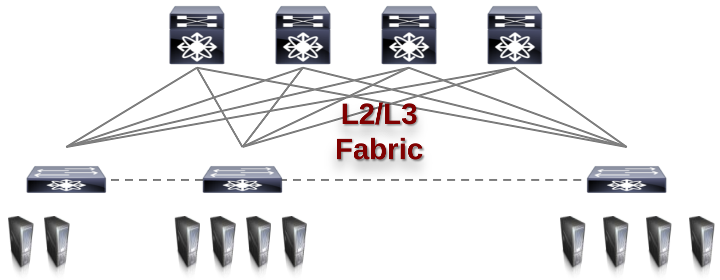
- IP/BGP MPLS VPNs are highly scalable today
- PE routers must:
  - Hold a large number of prefixes
  - Maintain multiple routing protocol adjacencies
- Mobility and cloud will add pressure in terms of:
  - Prefix granularity and volume
  - Increased number of PEs

Pull Protocol (on-demand) Model

- LISP deployments and footprint are increasing rapidly
- On-demand caching models ease the requirements on the edge devices:
  - Only prefixes being utilised are cached
  - No routing adjacencies are maintained
- A pull model is expected to provide global scalability to enable pervasive cloud models

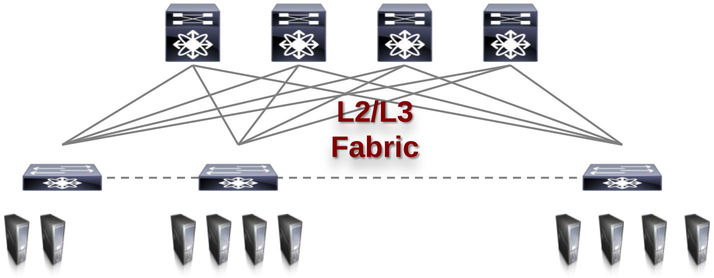
IP Mobility with L3 Overlays

- Granular location information
  - Allow subnet members to move anywhere
- Layer 2 semantics
  - ARP proxy
  - Consistent default Gateway presence
- L3 at the Access
  - Access switch replies to all ARPs with the same MAC address
  - Host routing for all traffic within the fabric
  - Summary prefix outside the fabric

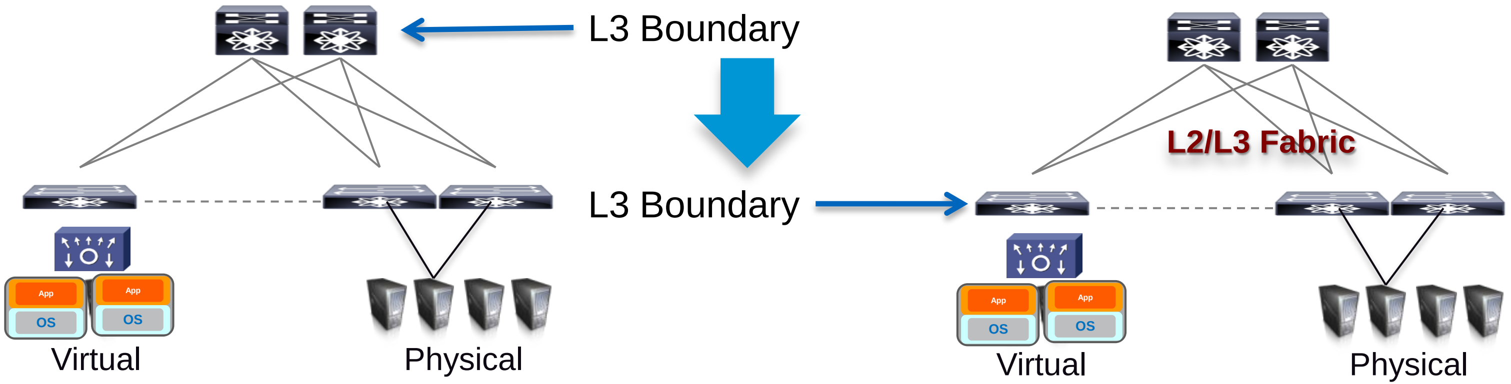


Combined L2/L3 Overlays

- Route all IP traffic including Intra-subnet
- Bridge only broadcast and link-local multicast traffic
- Assumption is that most traffic is IP (if not all)
- Bridge any non-IP traffic present
  - If only handling multi-destination non-IP traffic then no MAC address learning is required
  - If handling non-IP unicast traffic, then MAC address learning is required



Distributed Gateway Function in L2/L3 Overlays



Traditional L2 - centralised L2/L3 boundary

- Always bridge, route only at an aggregation point
- Large amounts of state converge
- Scale problem for large# of L2 segments
- Traditional L2 and L2 overlays

L2/L3 fabric (or overlay)

- Always route (at the leaves), bridge when necessary
- Distribute and disaggregate necessary state
- Optimal scalability
- Enhanced forwarding and L3 overlays

Cisco live!

Overlay Edge Device and Data Plane Evolution

Service

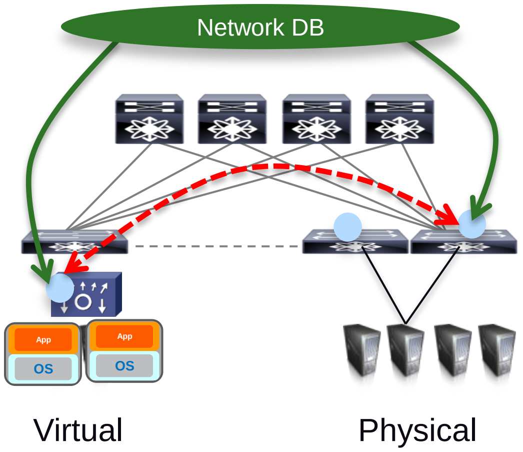
Layer 2 Service

Layer 3 Service

Edge Device

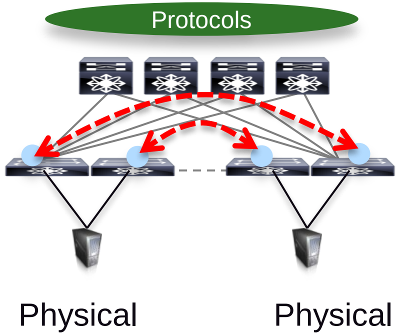
Host Overlays

Network Overlays



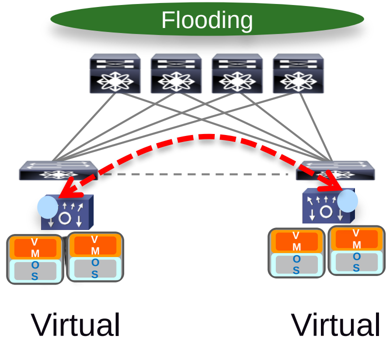
Overlay Network Evolution: Edge Devices

Network Overlays



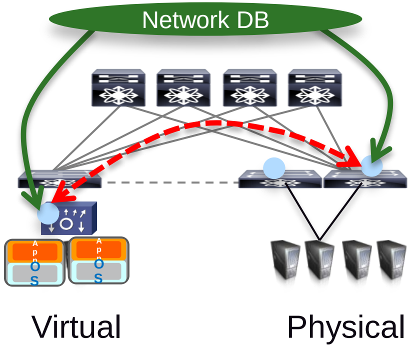
- Router/switch end-points
- Protocols for resiliency/loops
- Traditional VPNs
- OTV, VPLS, LISP, FP

Host Overlays



- Virtual end-points only
- Single admin domain
- VXLAN, NVGRE, STT

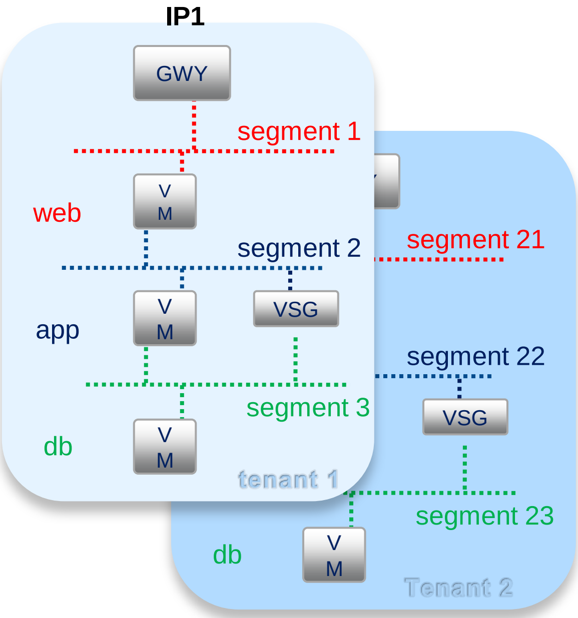
Hybrid Overlays



- Physical and Virtual
- Resiliency + Scale
- x-organisations/federation
- Open Standards

Tunnel End-points

Host Overlays



Multi-tier Virtual App = VMs + vSegments + GWY

Application: Cloud Services

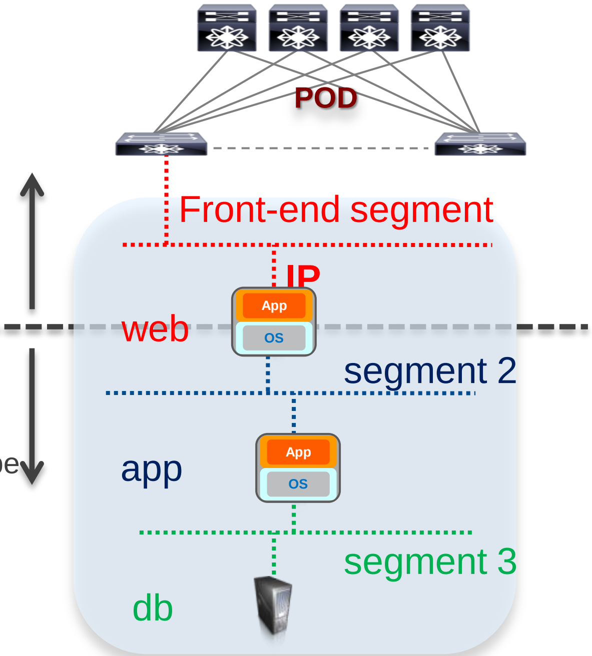
Elastic creation of **virtual Segments**

- Mobile: Can be instantiated anywhere
  - Move along with VMs as necessary
- Very large number of segments
  - Do not consume resources in the network core
- Isolated, not reachable from the IP network
  - Front-end segment must be handled by the fabric
- Host overlays are initiated at the hypervisor virtual switch ➔ Virtual hosts only
- GWY to connect to the non-virtualised world
- Variants: VXLAN, NVGRE, STT

Segmentation:
Tiered Applications and Virtualisation

Fabric
Global reachability
IP mobility
Physical or Virtual

Fabric or Overlay
Layer 2 segments
Narrow reachability scope
Concealed segments
Mostly Virtual

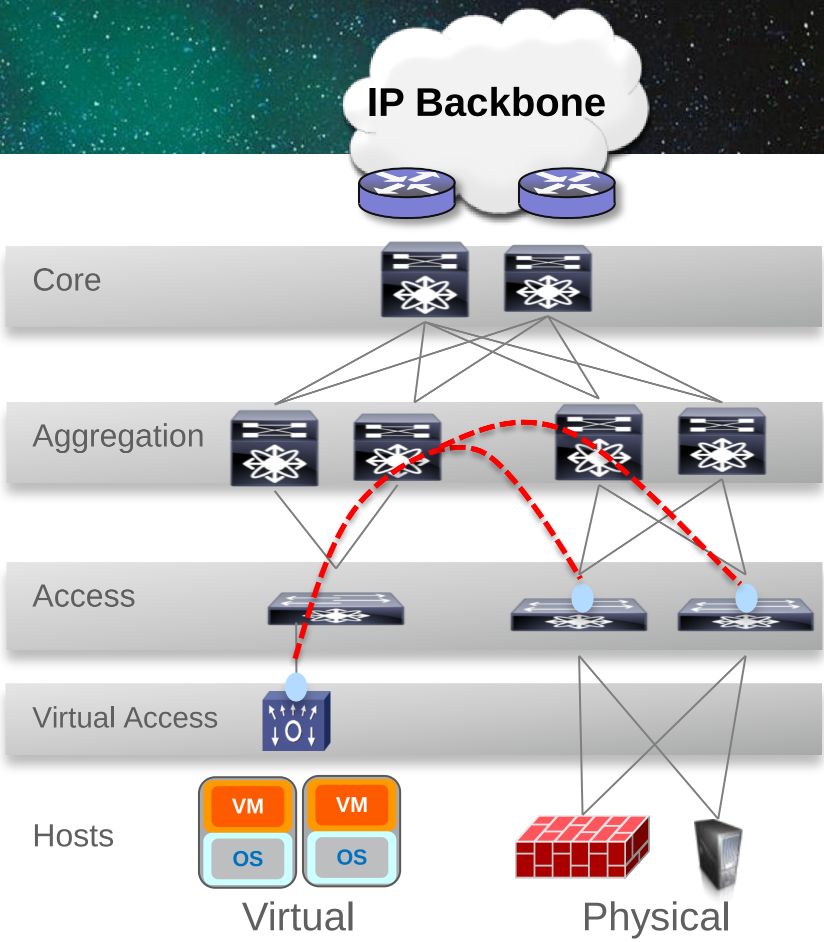


- Many segments to interconnect app tiers
  - Small segments, isolated/concealed between tiers
  - One app will have multiple tier-segments
- Front-end segments provide connectivity to the broader "physical" network
- Tenant Segmentation != app-tier segmentation
  - Tenant Segmentation (Front-end) in the Fabric
  - App-tier segments in an overlay or in the Fabric
- A mix of physical and virtual
  - Web and app layers are commonly virtualised
  - Web layer reachable via the front-end physical network
  - The DB layer and services often run on bare metal

Front-end (and physical hosts) mobility and network reachability provided natively in the Fabric

Hybrid Overlays

- Hypervisors introduce an additional tier in the network: The virtual Access (virtual Switch)
- **VMs** connect to the virtual Access
  - **Host overlays** start at the virtual Access
  - Virtualisation based resiliency: **Single attached sites**
- **Physical hosts** connect to the physical Access
  - **Network overlays** start at the physical Access
  - Network resiliency: **Site multi-homing**
- A hybrid overlay allows the combination of physical and virtual resources

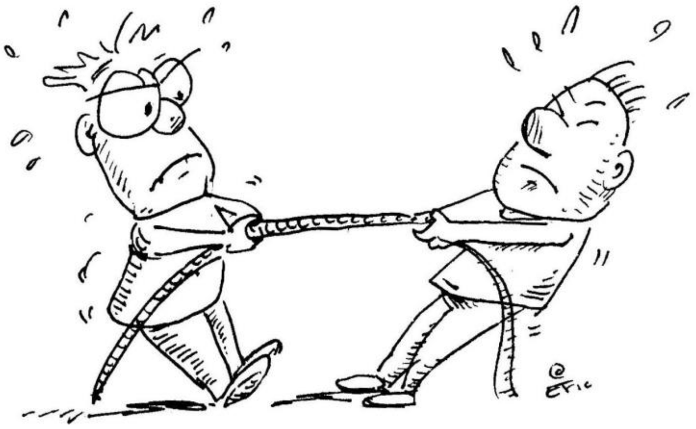


Which Encapsulation?

VXLAN

NVGRE

LISP



MPLS

STT

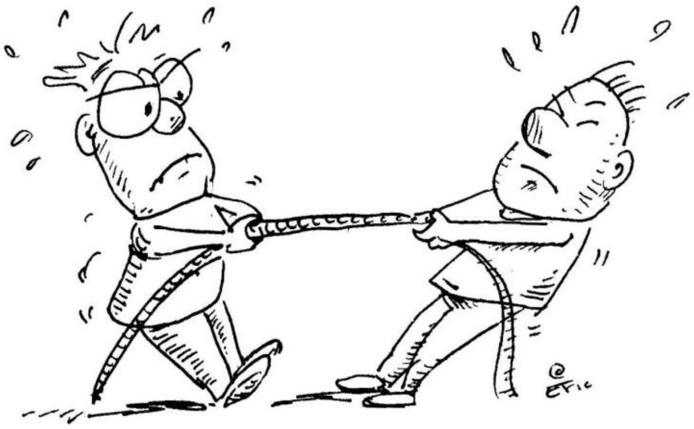
FabricPath

The Multi-protocol Router

TCP/IP

SPX/IPX

AppleTalk



Token Ring

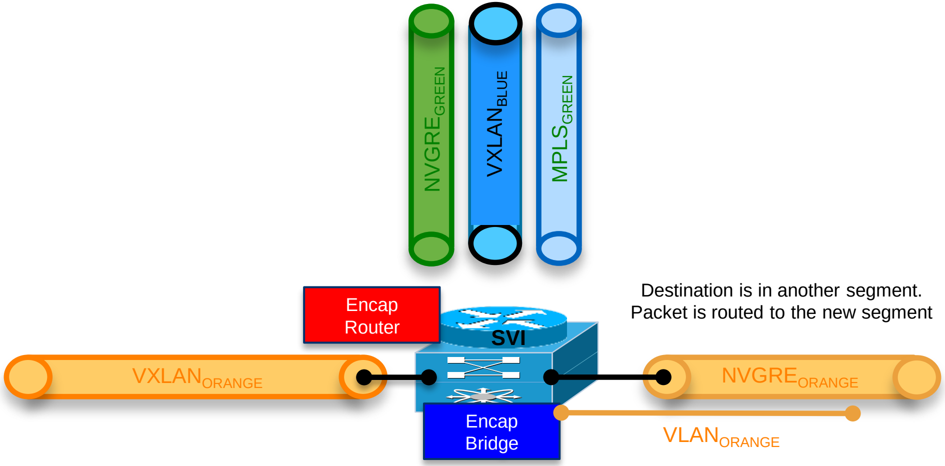
ATM

DECNet

The Multi-encapsulation Gateway

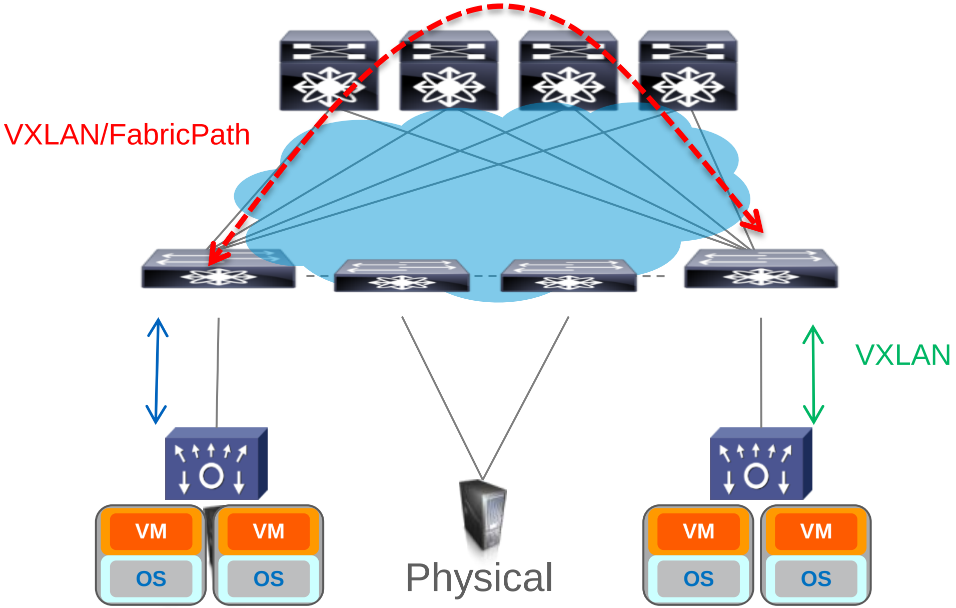
- Multi-encapsulation Gateway:
  - VXLAN, NVGRE, MPLS, LISP, VLAN, OTV
- Bridging (L2 Gateway)
- Routing (L3 Gateway)

- Multiple TEPs in independent VRFs
- Nesting of IP overlays into MPLS VPNs
- Available across the product line

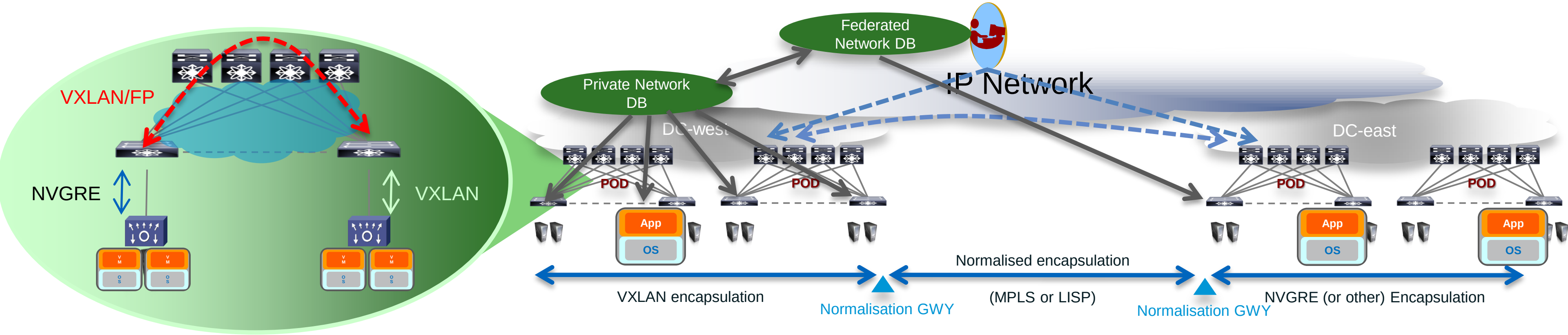


Normalisation: The Encapsulation Doesn't Matter

- Intelligence in the Control Plane
- Capabilities Exchange in Control Plane (negotiate encapsulation)
- Normalise to common encapsulation
- Pervasive Multi-encap Gateways for optimal traffic patterns



Data Plane and Control Plane Normalisation



- Multi-encapsulation Hardware Gateways
- Normalise to a common encapsulation in the Fabric and/or between Data Centres
- Terminate and map multiple types of encapsulation
  - VXLAN, NVGRE, MPLS, OTV, LISP
- Terminate and re-distribute information between overlay control protocols
  - Controllers, BGP, LISP

Encapsulation HW Offload

Host Overlays

- Current forwarding penalty for SW encap is about 50% throughput
- STT trick leverages TCP offload engine in existing NICs
  - TCP violation, short lived workaround
  - P2P only, no routing of flows
- VXLAN/NVGRE offload on NICs
  - The way forward for host overlays
  - Disruptive, many touch points
  - Static as ASICs: headers still in flux
  - Cisco 3rd Gen VIC 2HCY14 (stateless offload)

Network Overlays

- ASIC acceleration of overlay encapsulations
  - Cisco F-series ASICs with parser programmability
  - Fast enablement of incremental functions in header reserved fields without replacing HW
- Minimal disruption at the network access
  - Manageable number of touch points
- Encapsulation Normalisation
- Maximise throughput

LISP and VXLAN Headers Today

```
     0                   1                   2                   3
     0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
  / |Version|  IHL  |Type of Service|          Total Length         |
 /  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   |         Identification        |Flags|      Fragment Offset    |
|   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
OH  |  Time to Live | Protocol = 17 |         Header Checksum       |
|   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   |                    Source Routing Locator                     |
 \  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
  \ |                 Destination Routing Locator                   |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
  / |       Source Port = xxxx      |       Dest Port = 4341        |
UDP +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
  \ |           UDP Length          |        UDP Checksum           |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
L   |N|L|E|V|I|flags|            Nonce/Map-Version                  |
I \ +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
S / |                 Instance ID/Locator Status Bits               |
P   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
  / |Version|  IHL  |Type of Service|          Total Length         |
 /  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   |         Identification        |Flags|      Fragment Offset    |
|   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
IH  |  Time to Live |    Protocol   |         Header Checksum       |
|   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   |                           Source EID                          |
 \  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
  \ |                         Destination EID                       |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

LISP: IP only today

```
      0                   1                   2                   3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   / |Version|  IHL  |Type of Service|          Total Length         |
  /  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |   |         Identification        |Flags|      Fragment Offset    |
 |   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 OH  |  Time to Live | Protocol = 17 |         Header Checksum       |
 |   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |   |                    Source Routing Locator                     |
  \  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   \ |                 Destination Routing Locator                   |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   / |       Source Port = xxxx      |       Dest Port = 8472        |
 UDP +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   \ |           UDP Length          |        UDP Checksum           |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   / |R|R|R|R|I|R|R|R|            Reserved                           |
VXLAN +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   \ |       VXLAN Network Identifier (VNI) |   Reserved             |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   / |              Inner Destination MAC Address                    |
  /  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |   | Inner Destination MAC Address | Inner Source MAC Address      |
 Eth +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |   |                   Inner Source MAC Address                    |
  \  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Optional Ethertype = C-Tag [802.1Q]   | Inner.VLAN Tag Information   |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   / | Ethertype of Original Payload |                               |
  /  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+                               |
 |   |                                  Original Ethernet Payload    |
Payload|                                                             |
 |   |(Note that the original Ethernet Frame's FCS is not included)  |
 |   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
  \  |     New FCS (Frame Check Sequence) for Outer Ethernet Frame   |
   \ +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

LISP, OTV and VXLAN Normalisation with Generic Protocol Extension (gpe)

draft-lewis-lisp-gpe-00.txt

```
                0                   1                   2                   3
                0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
               +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
             / |Version|  IHL  |Type of Service|          Total Length         |
            /  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
           |   |         Identification        |Flags|      Fragment Offset    |
           |   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
           OH  |  Time to Live | Protocol = 17 |         Header Checksum       |
           |   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
           |   |                    Source Routing Locator                     |
            \  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
             \ |                 Destination Routing Locator                   |
               +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
             / |       Source Port = xxxx      |       Dest Port = 4341        |
           UDP +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
             \ |           UDP Length          |        UDP Checksum           |
               +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
             / |N|L|E|V|I|P|R|R|    Reserved   |Nonce/Map-Version/Protocol-Type|
          LISP +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
             \ |                 Instance ID/Locator-Status-Bits               |
               +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

draft-quinn-vxlan-gpe-00.txt

```
                0                   1                   2                   3
                0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
               +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
             / |Version|  IHL  |Type of Service|          Total Length         |
            /  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
           |   |         Identification        |Flags|      Fragment Offset    |
           |   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
           OH  |  Time to Live | Protocol = 17 |         Header Checksum       |
           |   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
           |   |                    Source Routing Locator                     |
            \  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
             \ |                 Destination Routing Locator                   |
               +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
             / |       Source Port = xxxx      |       Dest Port = 4789        |
           UDP +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
             \ |           UDP Length          |        UDP Checksum           |
               +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
             / |R|R|R|R|I|P|R|R|    Reserved   |         Protocol Type         |
         VXLAN +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
             \ |       VXLAN Network Identifier (VNI) |      Reserved          |
               +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Ethernet or IP Payload: Defined in the Protocol Type
Common encapsulation for LISP and VXLAN
L2 and L3 Payloads in both LISP and VXLAN

Header Evolution: Metadata in Overlay Headers

- Segmentation (VRFs, VPNs, Segments)
- L2 and L3 Payloads
- Policy
- Service Chaining
- Underlay integration (load balancing, traffic engineering)

LISP, OTV and VXLAN GPE Plus Network Service Header

draft-quinn-nsh-01
standardises the definition for the vPath 3.0 header

Protocol Type = 0xNSH

Protocol Type = IP

```
Base Service Header:

 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|O|C|R|R|R|R|R|R|   Protocol Type                |Service Index  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                Service path                    |   Reserved    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

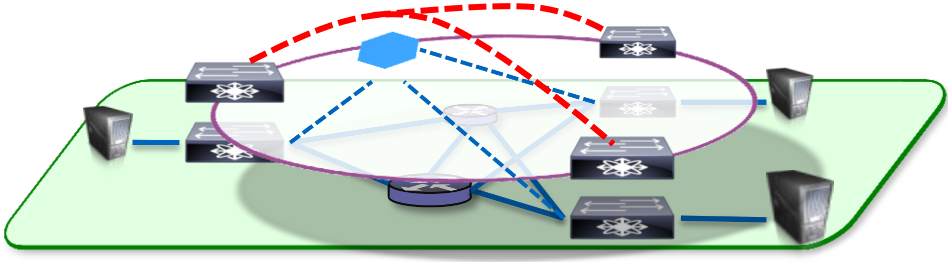
```
        0                   1                   2                   3
        0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     / |Version|  IHL  |Type of Service|          Total Length         |
    /  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |   |         Identification        |Flags|      Fragment Offset    |
   |   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 OH    |  Time to Live | Protocol = 17 |         Header Checksum       |
   |   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |   |                    Source Routing Locator                     |
    \  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     \ |                 Destination Routing Locator                   |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     / |       Source Port = xxxx      |       Dest Port = 4341        |
 UDP   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     \ |           UDP Length          |        UDP Checksum           |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     / |N|L|E|V|I|P|R|R|   Reserved    |Nonce/Map-Version/Protocol-Type|
LISP   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     \ |                 Instance ID/Locator-Status-Bits               |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     / |                                                               |
    /  |                     Base Header                               |
   |   |                                                               |
   |   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |   |                     Context Header                            |
 NSH   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |   |                     Context Header                            |
   |   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |   |                     Context Header                            |
    \  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     \ |                     Context Header                            |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     / |Version|  IHL  |Type of Service|          Total Length         |
    /  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |   |         Identification        |Flags|      Fragment Offset    |
   |   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 IH    |  Time to Live |    Protocol   |         Header Checksum       |
   |   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |   |                           Source EID                          |
    \  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     \ |                         Destination EID                       |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Overlay Signalling Evolution

Service	Edge Device	Signalling
Layer 2 Service Layer 3 Service	Host Overlays Network Overlays	Data Plane Learning Control Plane Learning

Overlay Signalling

- Service Discovery
  - Edge devices in an overlay need to discover each other
- Address Advertising and Tunnel Mapping
  - Edge devices must exchange host reachability information
  - Map end-point to location
- Tunnel Management
  - Maintain and manage connections between edge devices



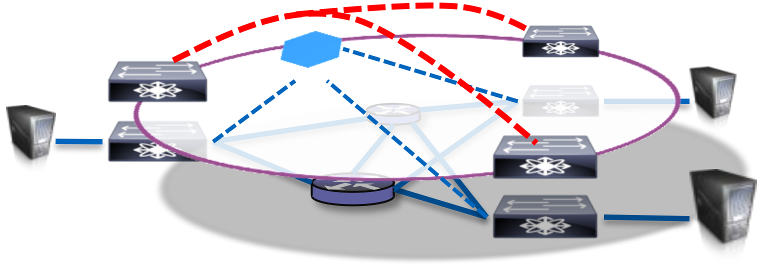
Overlay Signalling

Data Plane Learning

- Based on gleaning information from data plane events
  - Example: Source Learning on bridges
- Provides the following:
  - Address advertisement/mapping (very effectively)
  - Some tunnel management is possible
  - Does not provide Service Auto-discovery
- Requires a flood facility for data plane events to propagate:
  - Multicast tree
  - Unicast replication group at the head-end
- Flood facility can be manually configured on every device (e.g. join a mcast group or configure a list of unicast destinations)
- Usually is supplemented with a control protocol for Service Discovery (specially if using unicast replication)

Overlay Signalling

Control Plane



- Provides:
  - Service Discovery
  - Address Advertising/Mapping
  - Tunnel Management
  - Extensions for multi-homing and advanced services can be provided

Protocol or Controller:

- **Routing Protocol** amongst Edge Devices
  - BGP, IS-IS, LISP
- Central database on a **Controller**
  - Distributed Virtual Switches (OVS, N1Kv/VSM)

Push or Pull:

- **Push** all information to all Edge Devices
  - BGP, IS-IS, Controllers
- **Pull** and cache on demand @ ED
  - LISP, DNS, Controllers

Overlays Evolve to Meet Network Challenges

Handling Workloads in the Data Centre

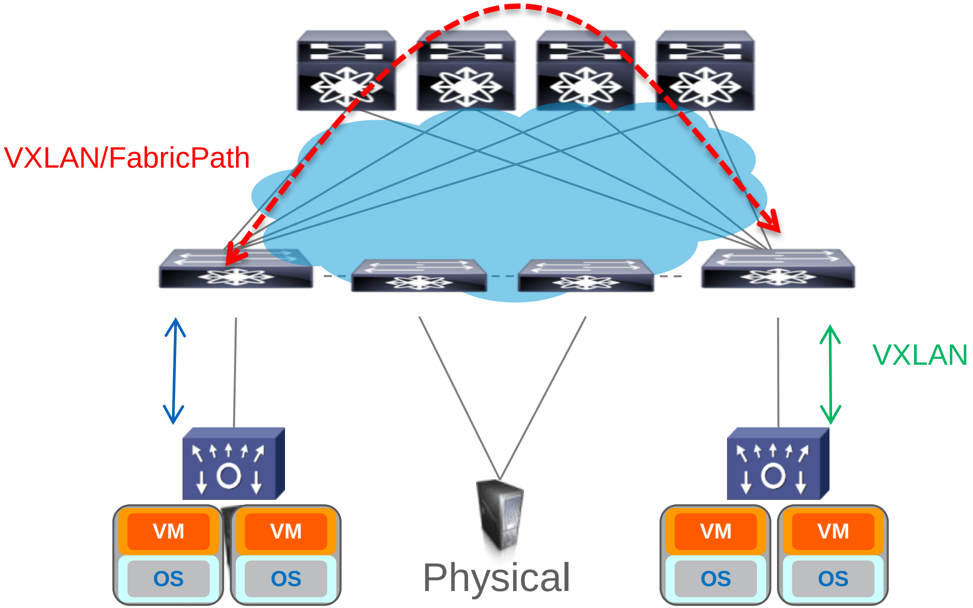
Choosing the right tools …

	Task	Technology	Functionality
Inter-Site (DCI & WAN)	**Move** a workload across PODs/sites	OTV→LISP	IP Mobility
	Distribute a workload across PODs/sites	OTV	LAN Extension
Intra-Site	**Move/Distribute** a workload within a site	FP → Enhanced Forwarding	L2/L3 Fabrics
	Interconnect tiers of a virtual App	VXLAN, NVGRE, STT	Virtual L2 segments

↕ Segmentation

DC-Fabric: Normalised L2/L3 Network Overlays

- Terminate the encapsulation from the host overlay
- Translate to a normalised encapsulation in the fabric
- Seamlessly allow physical and virtual to connect to the fabric
- Fabric overlay provides L2 and L3 services with mobility and segmentation



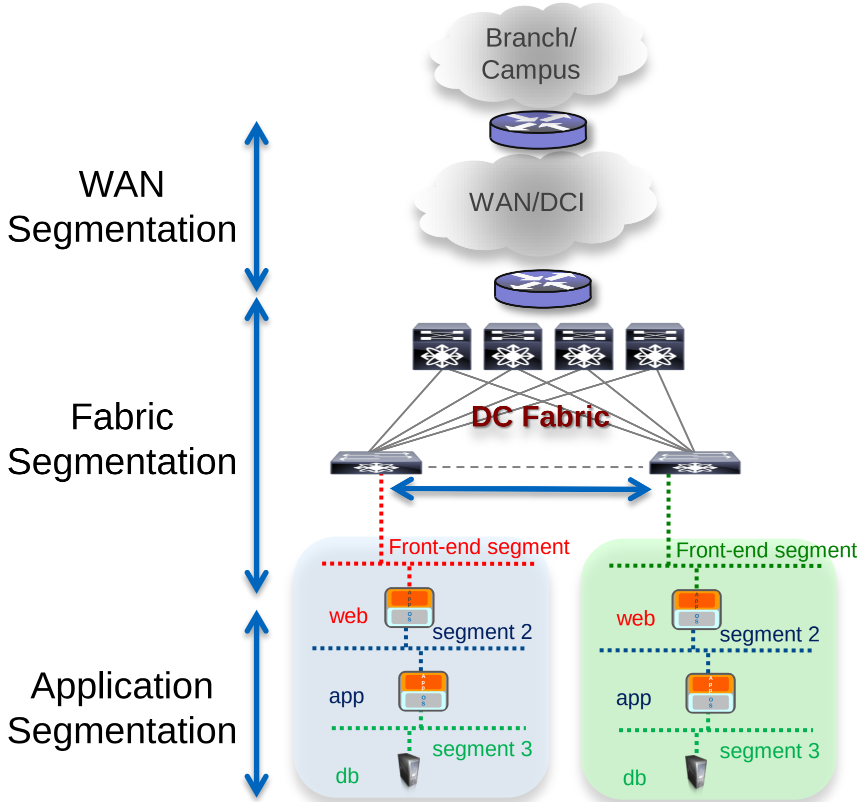
Handling Workloads in the Data Centre

Choosing the right tools …

	Task	Technology	Functionality
Inter-Site (DCI & WAN)	**Move** a workload across PODs/sites	OTV→LISP	IP Mobility
	Distribute a workload across PODs/sites	OTV	LAN Extension
Intra-Site	**Move/Distribute** a workload within a site	FP → Enhanced Forwarding	L2/L3 Fabrics
	Interconnect tiers of a virtual App	VXLAN, NVGRE, STT	Virtual L2 segments

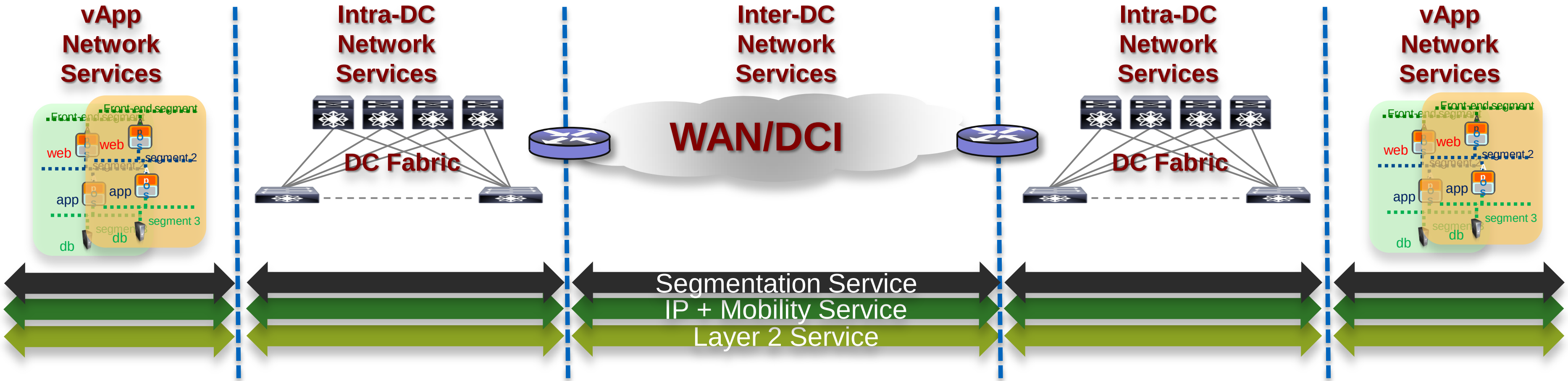
Segmentation

Segmentation End-to-end



- Segmentation at many levels
- Must be given continuity
  - Across the different network places
  - Across organisations and administrative boundaries
- All relevant technologies include the required segmentation semantics
- The network maps the segments together to provide a scalable and interoperable e2e segmentation solution

Failure Domain Scope



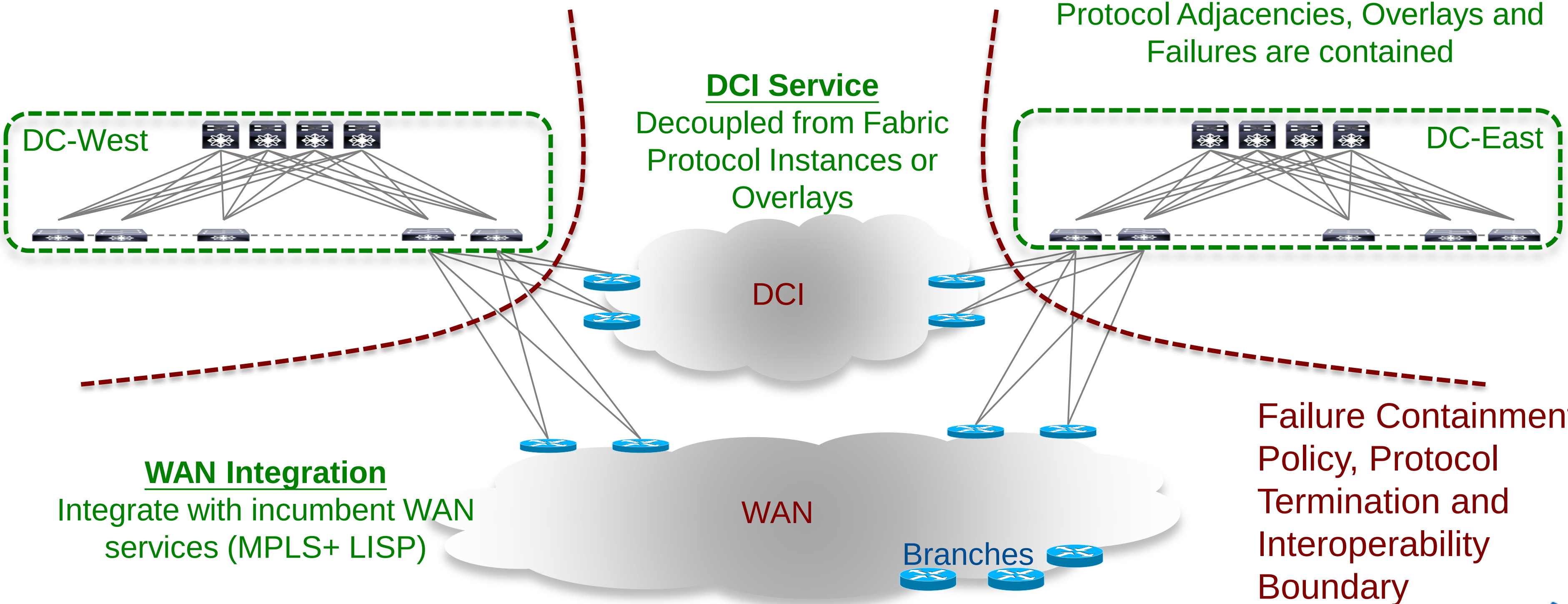
Core Principles of Network Resiliency/Scale applied to Overlay Services

- Clearly delineated Fault Boundaries and service domains
- Control Plane Hierarchy and Federation within and across domains
- Data Plane Boundaries
- Administrative Domain Delineation and Federation

DCI and WAN Integration

Each DC is Autonomous
Protocol Adjacencies, Overlays and Failures are contained

DCI Service
Decoupled from Fabric Protocol Instances or Overlays



WAN Integration
Integrate with incumbent WAN services (MPLS+ LISP)

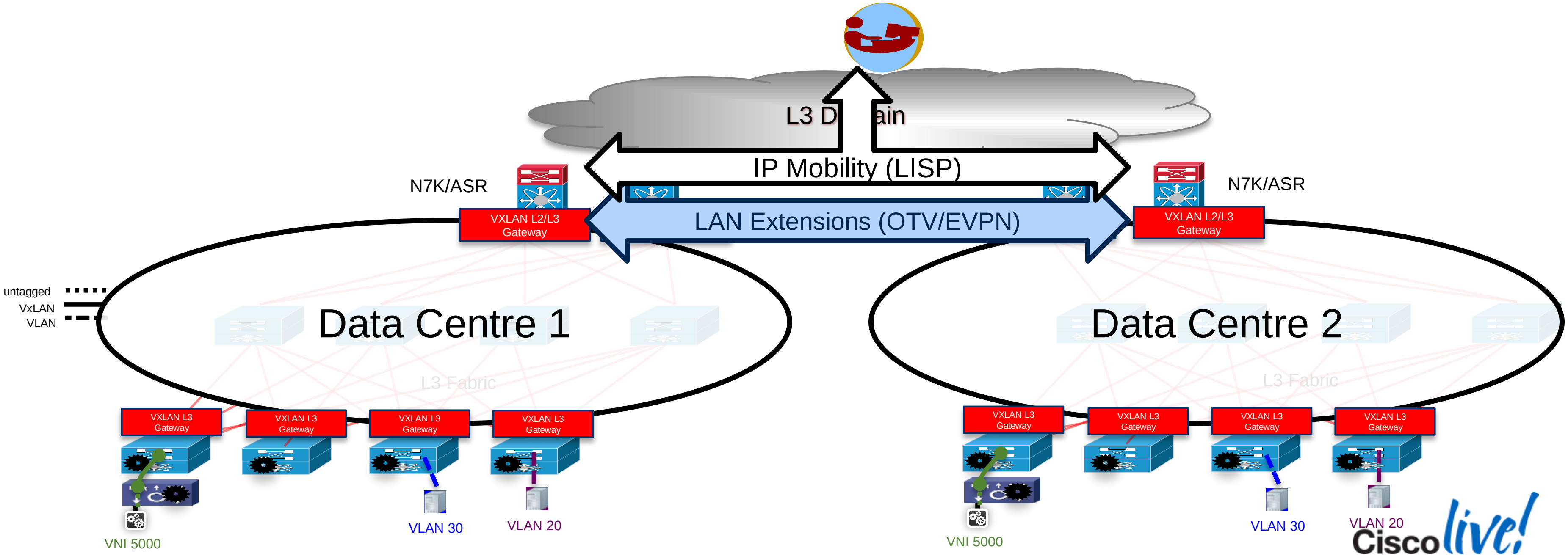
Failure Containment, Policy, Protocol Termination and Interoperability Boundary

End-to-end Segmentation and Mobility

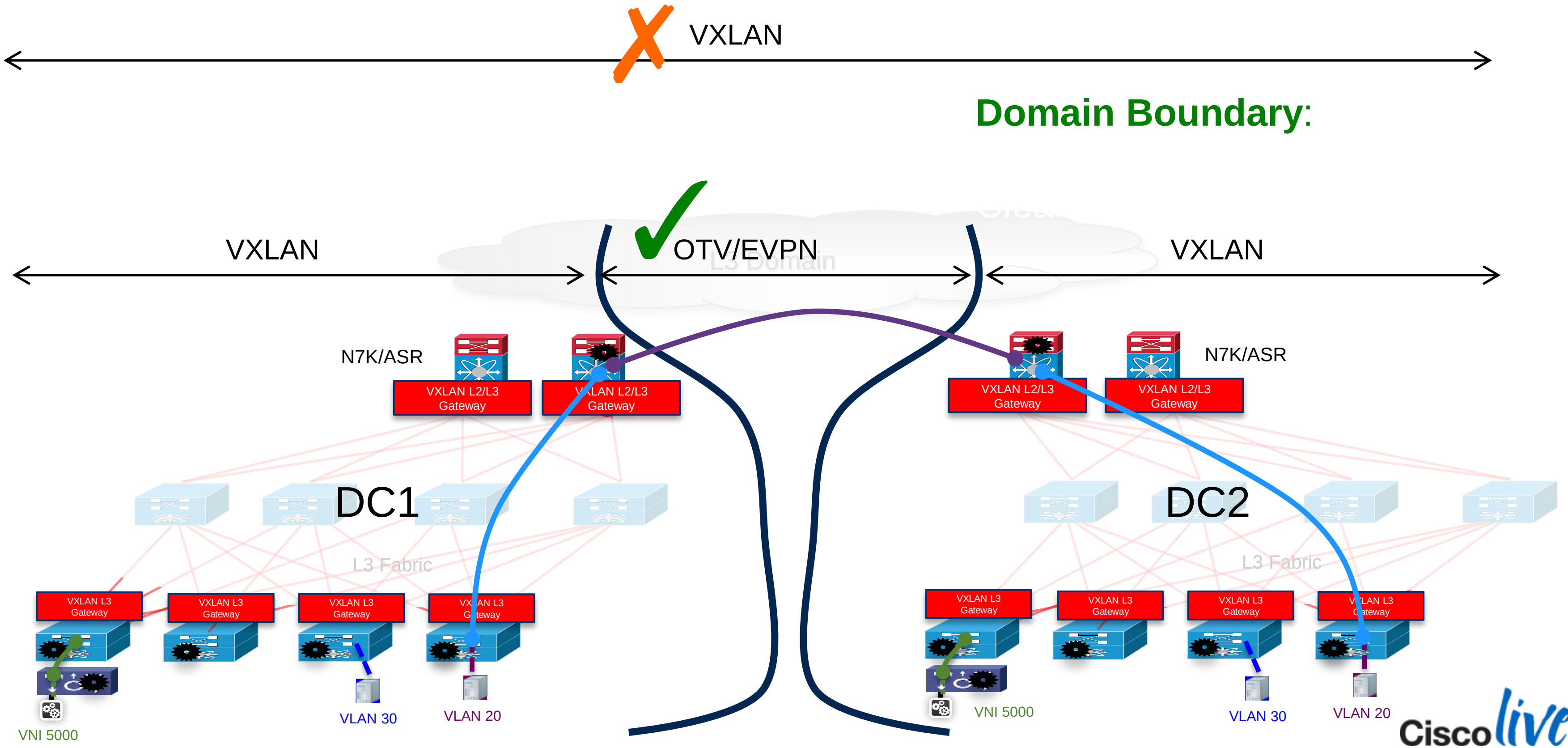
Interconnecting Multiple Data Centres

LAN Extensions and IP mobility

Ethernet extensions between independent fabrics
IP traffic is forwarded via the optimal path (no hair-pinning)



Interconnecting Multiple Data Centres
LAN Extensions

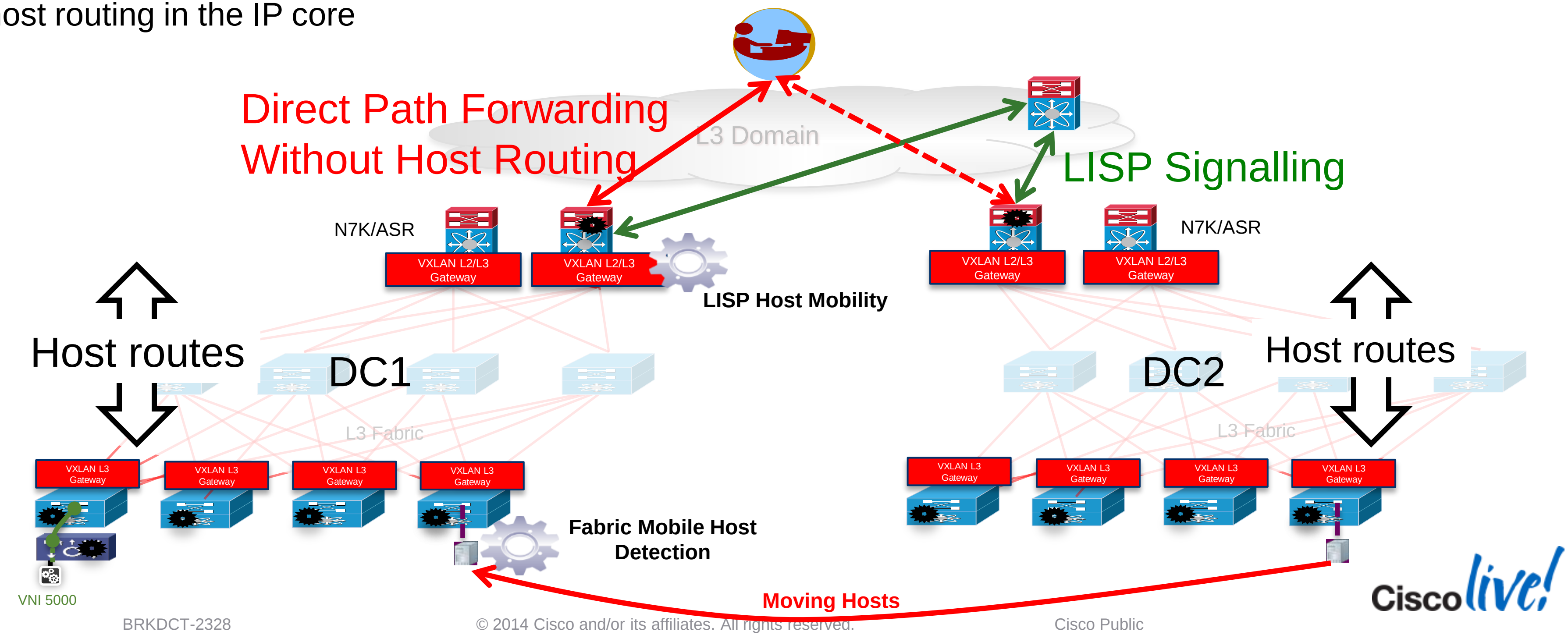


Interconnecting Multiple Data Centres

LISP IP Mobility for optimised routing

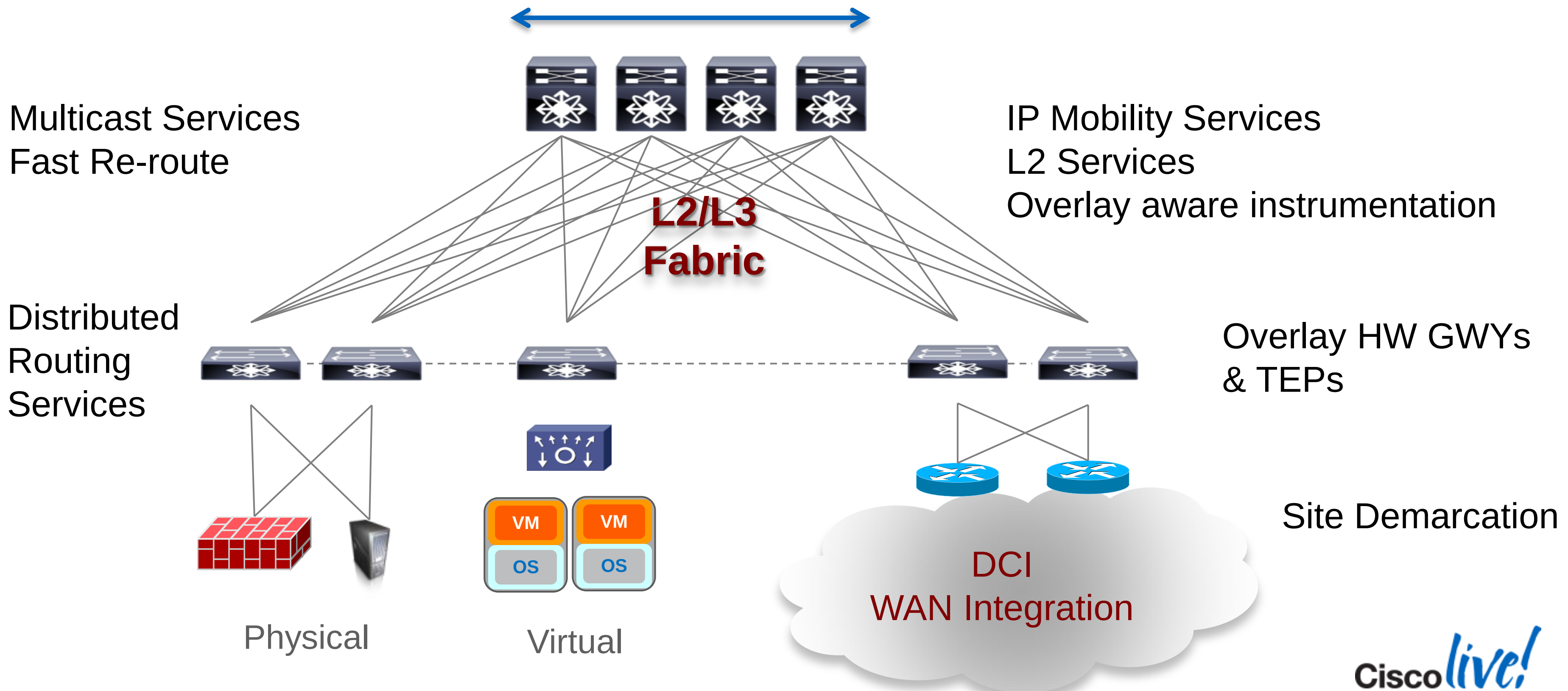
LISP Mobility:
LISP registrations and notifications
LISP encapsulation from client sites
No host routing in the IP core

LISP Signalling:
Relay mobility state between sites



Role of the Underlay

Fabric Relevance to a Hybrid Overlay



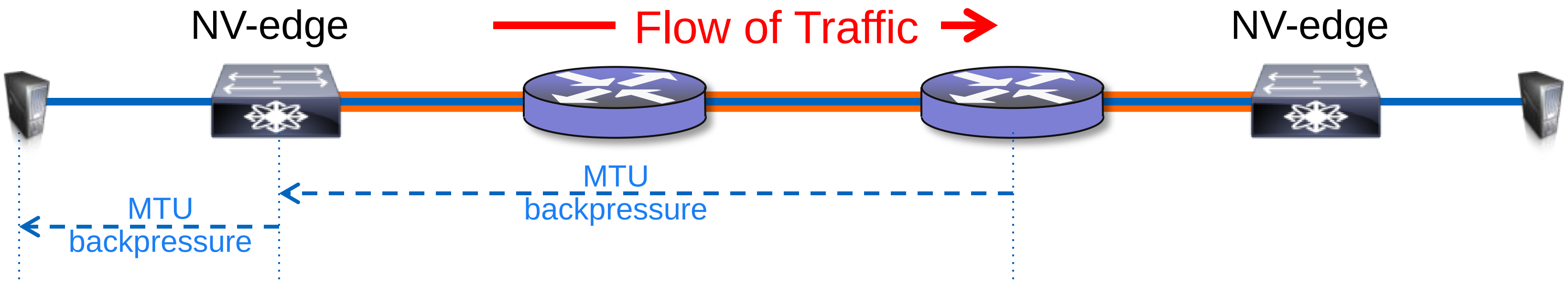
Encapsulation and Effective Throughput

1500bytes/packet (10Gbps) ➔ 1542 bytes/packet (10.1 Gbps)

64bytes/packet (10Gbps) ➔ 106 bytes/packet (10.3 Gbps)

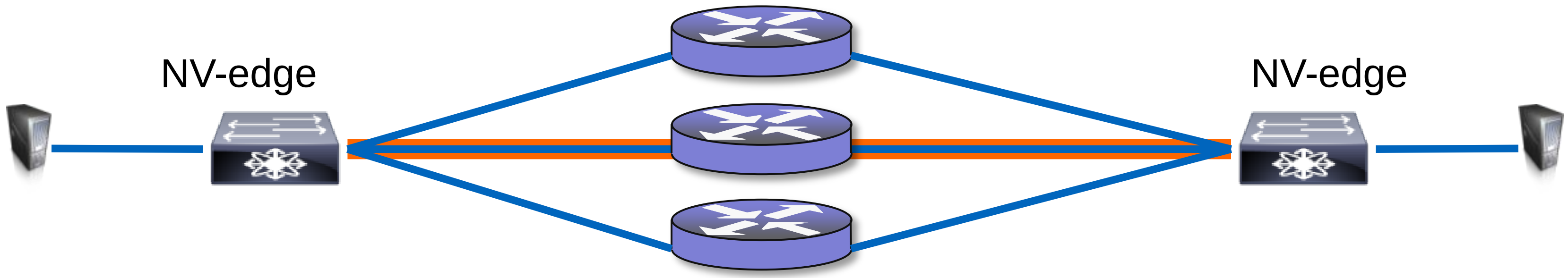
- Encapsulation adds bits to the traffic being sent
- When receiving traffic at full line rate, the encapsulated traffic will exceed the line-rate BW of the egress interface
  - Packet drops
  - Diminished effective throughput
- The uplink BW should be greater than the downlink BW to avoid congestion by encapsulation
  - This is naturally done in the network

MTU Issues



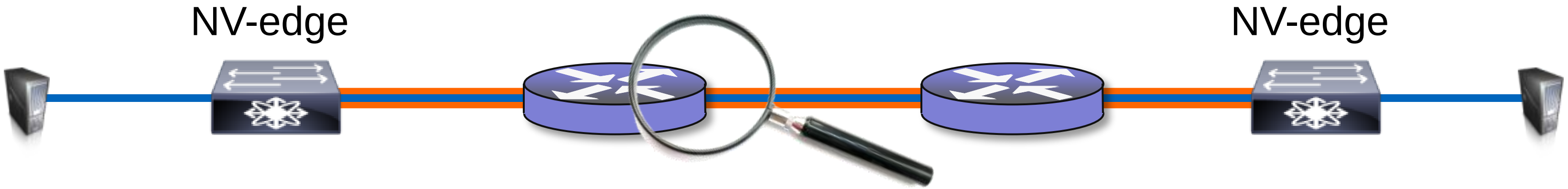
- Encapsulated traffic may exceed max MTU of the path
- When traffic is encapsulated with the Don't Fragment (DF) bit set:
  - If MTU is exceeded: IGMP unreachable message (datagram-too-big) is sent back to the encapsulating NV-edge
  - Encapsulating NV-edge will lower the tunnel MTU accordingly
  - Subsequent packets from the source will trigger an ICMP unreachable message from the NV-edge back to the server (if the traffic from the source has the DF bit set)
- If the DF bit is not set, the device sensing the MTU is exceeded should attempt to fragment the traffic

Multi-pathing and Entropy



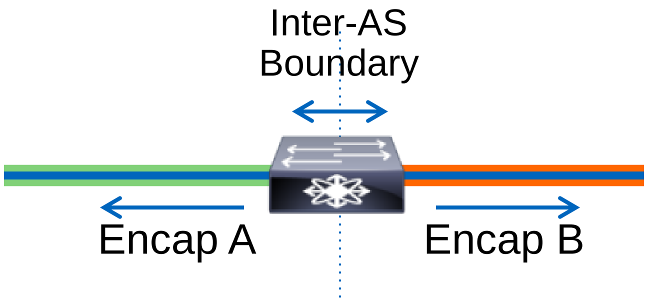
- Tunnel Polarisation: All encapsulated flows tend to look like a single flow between a pair of edge devices
  - Encapsulated traffic always hashes to a single path
- Adding entropy to the encapsulation header can depolarise the tunnels
  - Use all available paths
- UDP headers: Variable UDP source port
- GRE headers: Variable key field
- MPLS headers: Variable LSP label

Instrumentation and Overlay Awareness



- Infrastructure awareness of encapsulated traffic:
  - Outer/Encapsulation header
  - Overlay shim header
  - Internal/Payload header
  - Payload
- Overlay aware Switching & Routing infrastructure:
  - ACLs, QoS, Netflow
- Network Analysis Module (NAM) inspects encapsulated traffic

Data Plane and Control Plane Normalisation

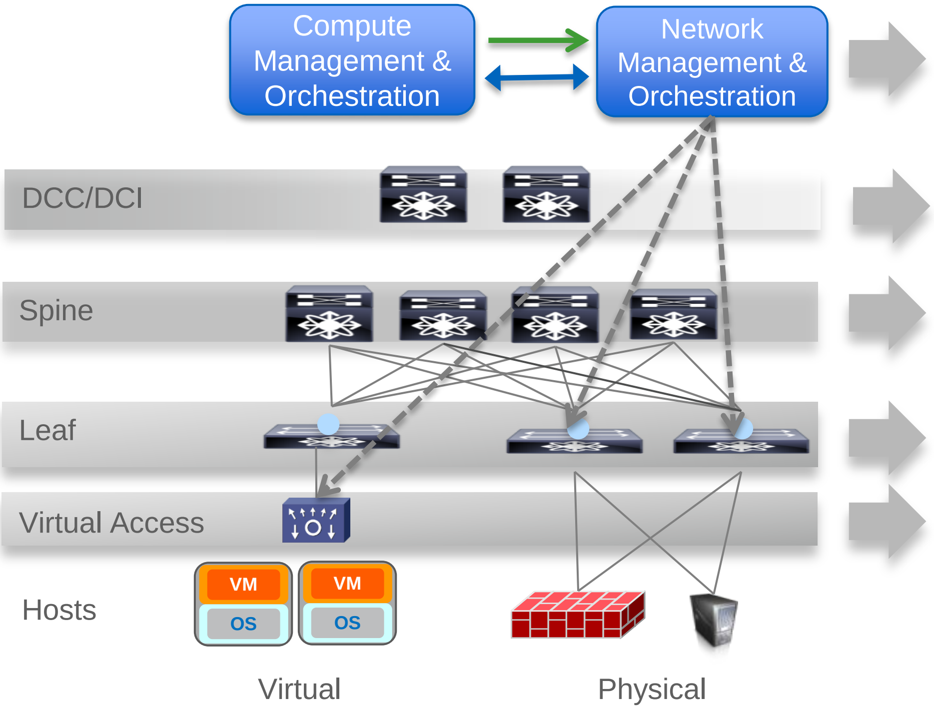


- Multi-protocol overlay gateway
- Terminate and map multiple types of encapsulation
  – VXLAN, NVGRE, MPLS, OTV, LISP
- Terminate and re-distribute information between overlay control protocols
  – Controllers, BGP, LISP




Management and Orchestration

Data Centre Fabric Management



Network Single Point of Management
Cisco Network Controller

Managed Programmable DCI

Managed Programmable Fabric

Managed Programmable Hybrid Overlays

Fully integrated Distributed Virtual Switch

Open APIs

Operational Delineation

Network Management

Programmatic APIs

Compute Management

Network team

Nexus OS CLI
DCNM GUI
APIs: One-PK/Openflow

- Network Policies aligned with Compute
- Integrated network management:
  - Fabric + Overlays
- Programmatic APIs

Infrastructure team

vCenter / UCS Manager Interface

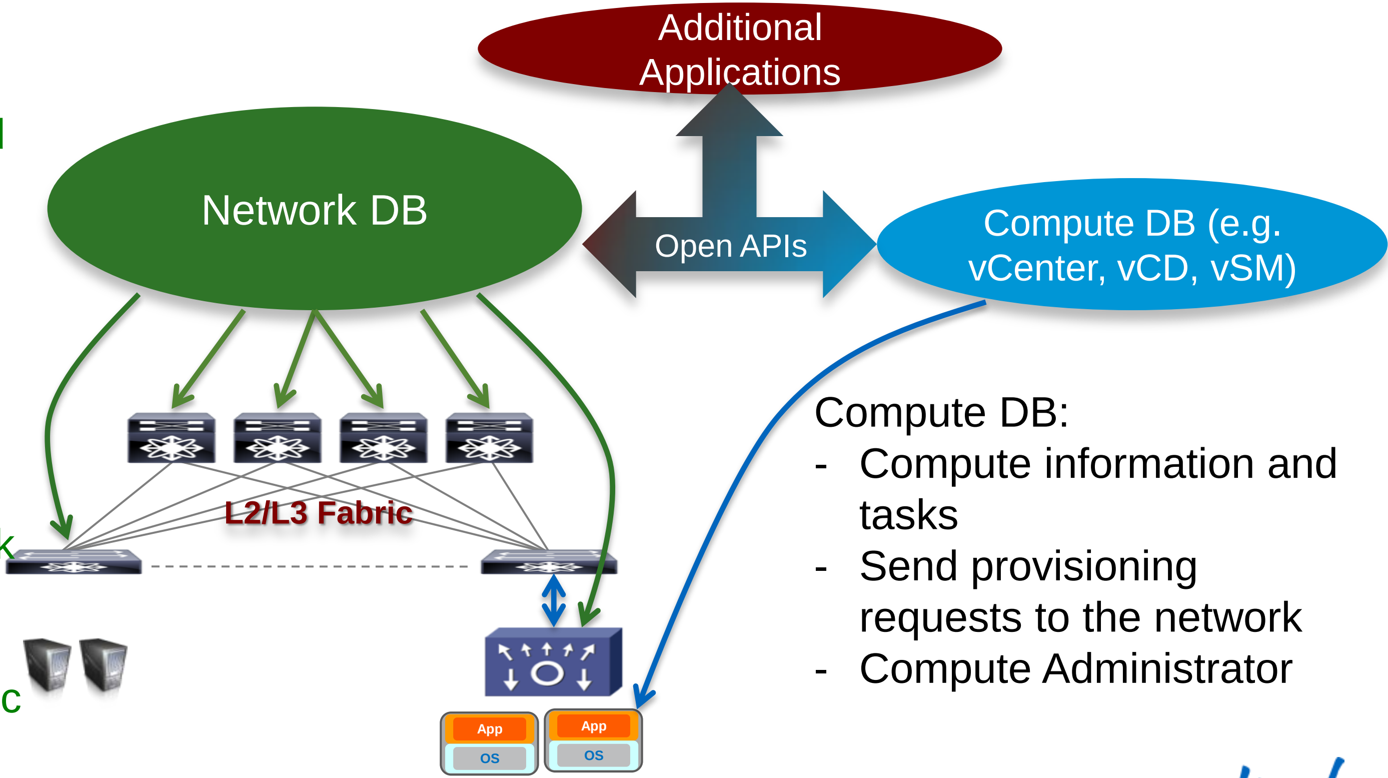
- Build Servers and Create Port profiles
- Create VMs and Assign Port profiles to VM

No hand-off required between Infrastructure and Network Admins for adding new servers

Orchestration, Management and Programmability

Network DB:
- Network information and tasks
- Network Administrator

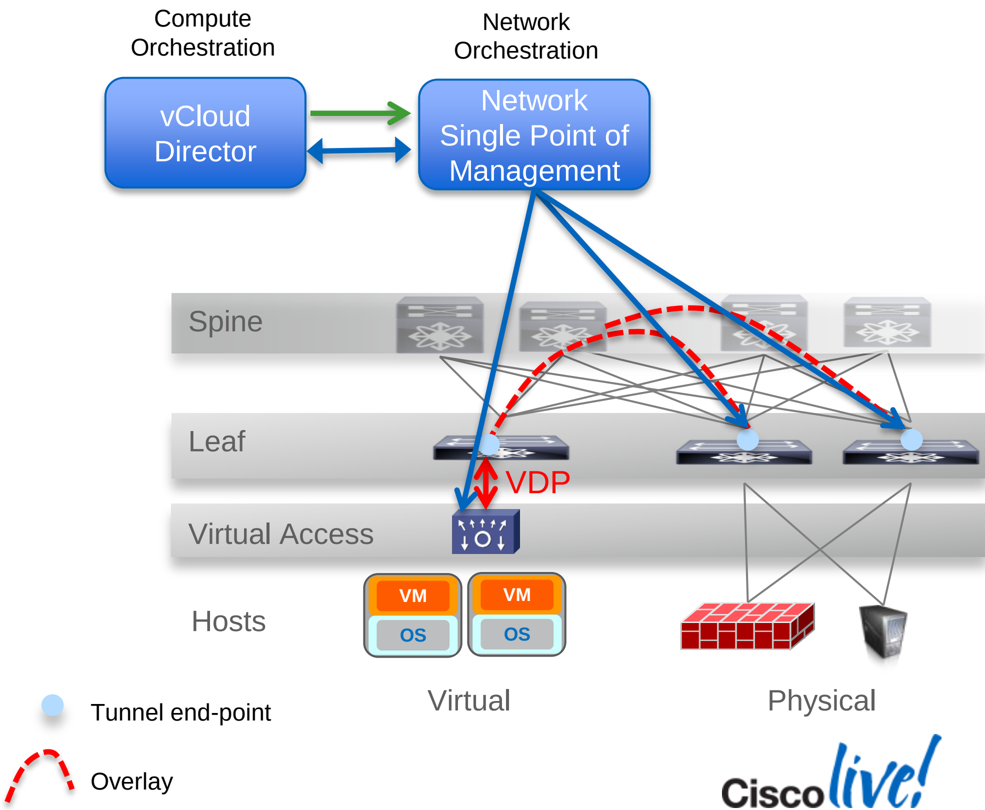
Single point of Management:
- Broker communications between compute orchestrator and network
- Act on provisioning requests from compute
- Manage underlying fabric and overlay



Compute DB:
- Compute information and tasks
- Send provisioning requests to the network
- Compute Administrator

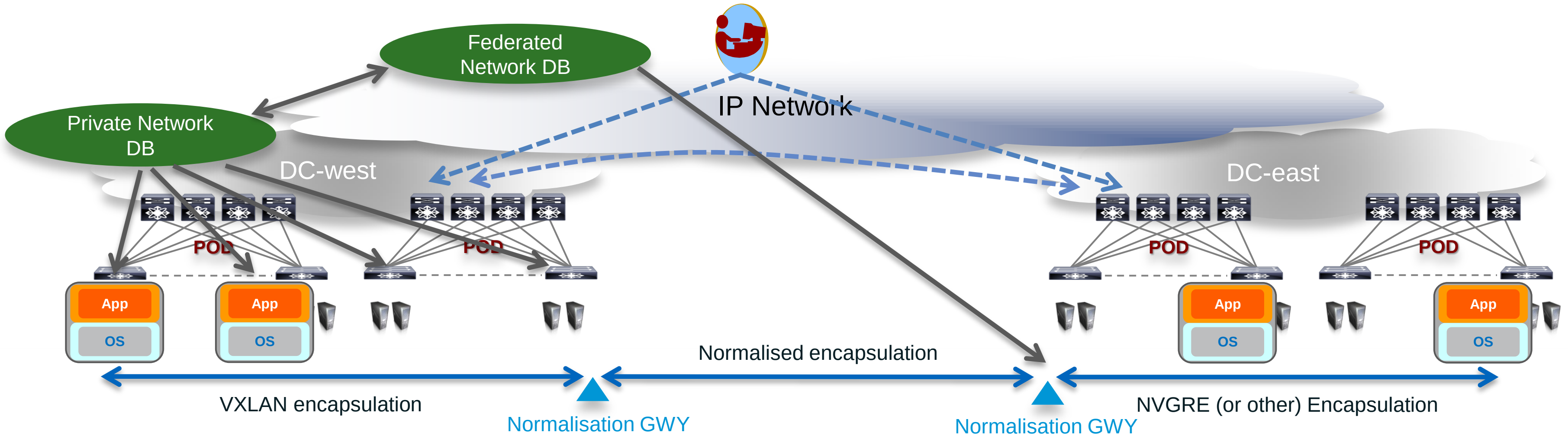
Distributed Virtual Switching

- Compute Controller (e.g. vCloud Director) integrated overlay provisioning
  - Integrates physical and virtual end-points
- Overlay encap/decap can be offloaded to network hardware
  - VDP = VSI Discovery Protocol (IEEE 802.1Qbg)
  - Cisco, HP, IBM, Brocade, Qlogic, Emulex, Broadcom, Mellanox, others …



Federated/Normalised Overlays Vision

Inter-DC and Intra-DC – LISP/BGP Protocol + Any encapsulation
Virtual and Physical Hosts
Layer 2 and Layer 3
Internet Scale






Q & A

Complete Your Online Session Evaluation

Give us your feedback and receive a Cisco Live 2014 Polo Shirt!

Complete your Overall Event Survey and 5 Session Evaluations.

- Directly from your mobile device on the Cisco Live Mobile App
- By visiting the Cisco Live Mobile Site www.ciscoliveaustralia.com/mobile
- Visit any Cisco Live Internet Station located throughout the venue

Polo Shirts can be collected in the World of Solutions on Friday 21 March 12:00pm - 2:00pm

Learn online with Cisco Live!

Visit us online after the conference for full access to session videos and presentations. www.CiscoLiveAPAC.com

CISCO TM