Building Blocks for Cloud Networks

Aldrin Isaac, Cross Portfolio Architecture, Juniper SPLM

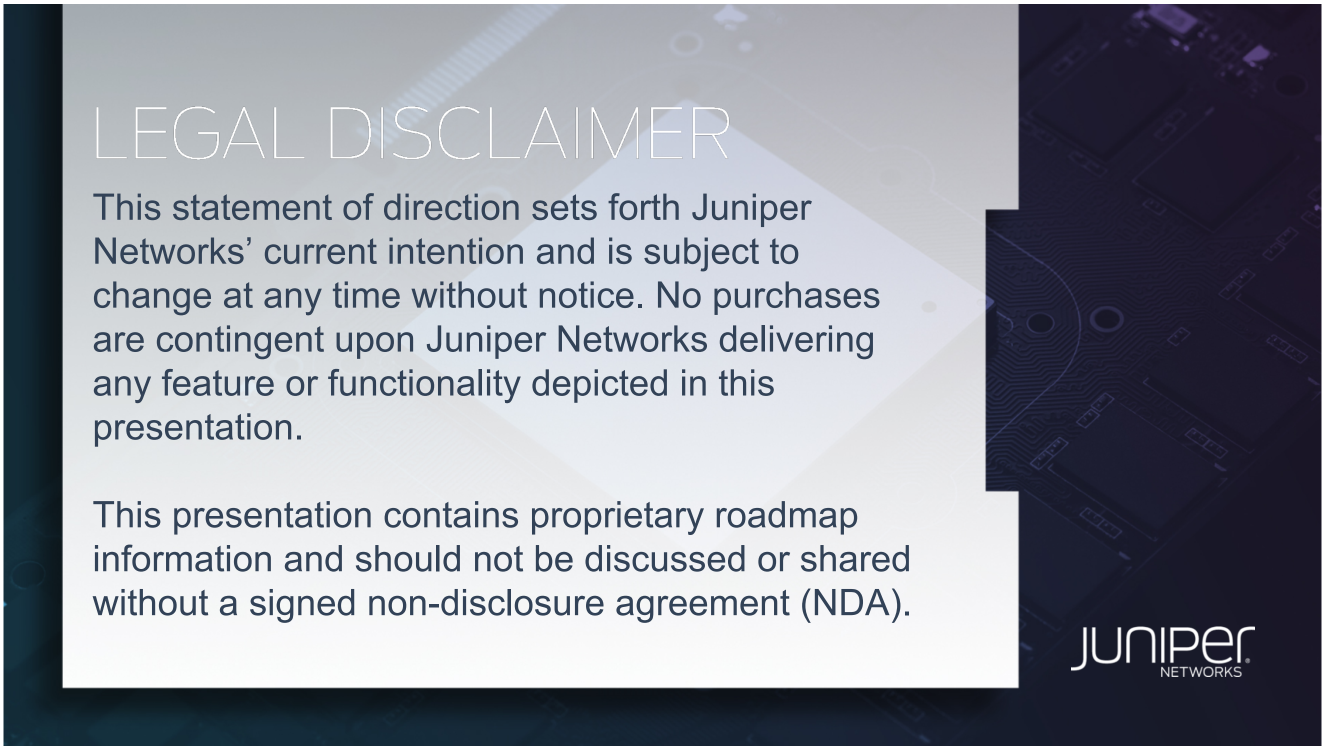
December 12, 2017

NXTWORK2017
JUNIPER CUSTOMER SUMMIT

This presentation is an overview of the key network building blocks for multi-service cloud networks that Juniper is delivering and driving. A useful exercise for the reader is to consider how these building blocks could be used to address their current or upcoming cloud network infrastructure.

The presenter, Aldrin Isaac, facilitates strategic product development at Juniper through end-to-end architecture. He is an industry veteran with 20 years of experience in network engineering, design, operations, automation, technology and protocol development.

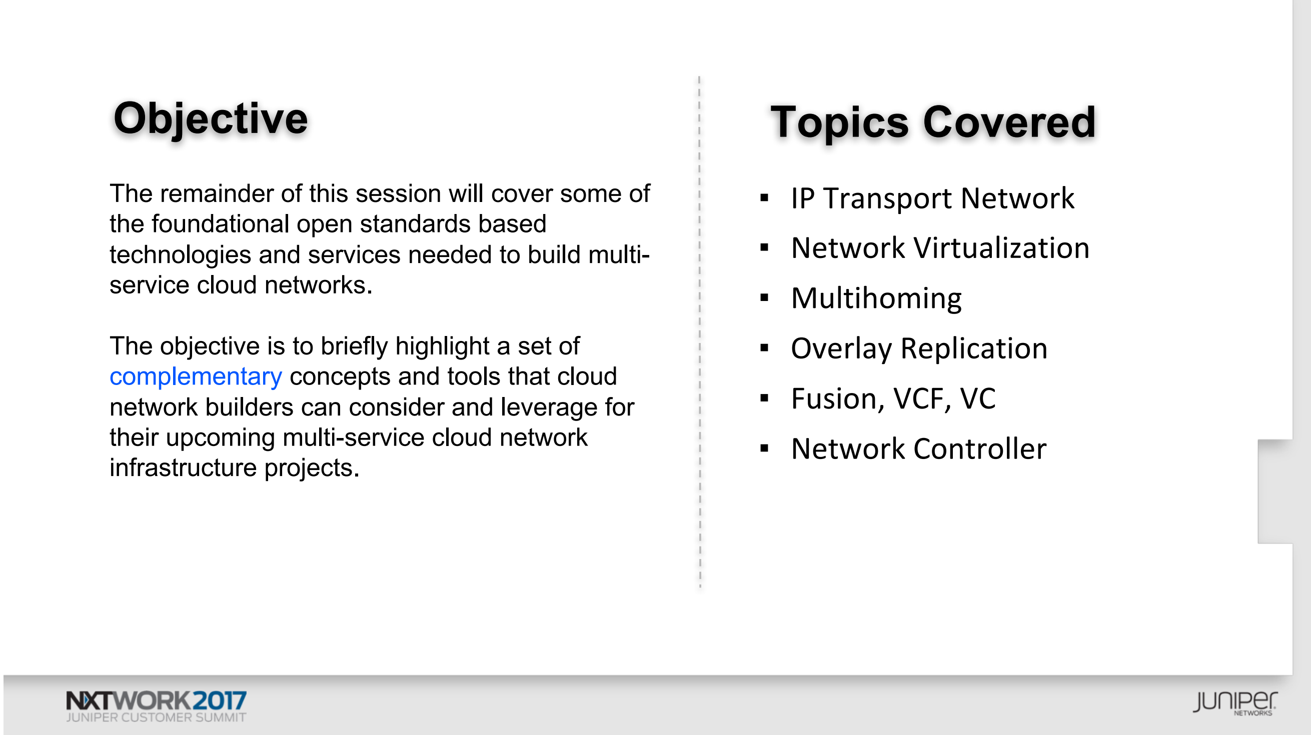
Prior to joining the Juniper team in 2015, Aldrin spent 20 years designing, building and running networks. He was the chief technologist responsible for the design and development of the global IP/MPLS and data center networks of a premier financial news, media and SaaS company, supporting the reach of its business to over 100 countries around the world. Aldrin has been deeply involved in every aspect of running large WAN and DC networks, from vision and design to build and support. Aldrin has also been a pioneer in fully automated network infrastructure. During his time as an operator, Aldrin sparked the creation and vision for EVPN (RFC7432) as a multi-service cloud control-plane, and is its co-author.



LEGAL DISCLAIMER

This statement of direction sets forth Juniper Networks’ current intention and is subject to change at any time without notice. No purchases are contingent upon Juniper Networks delivering any feature or functionality depicted in this presentation.

This presentation contains proprietary roadmap information and should not be discussed or shared without a signed non-disclosure agreement (NDA).



Objective

The remainder of this session will cover some of the foundational open standards based technologies and services needed to build multi-service cloud networks.

The objective is to briefly highlight a set of complementary concepts and tools that cloud network builders can consider and leverage for their upcoming multi-service cloud network infrastructure projects.

Topics Covered

- IP Transport Network
- Network Virtualization
- Multihoming
- Overlay Replication
- Fusion, VCF, VC
- Network Controller

NXTWORK2017
JUNIPER CUSTOMER SUMMIT

What is a multi-service cloud network?

We all know what a multi-service carrier network is. Juniper is a leading network technology supplier in this space. The multi-service carrier network optimizes an operators pool of expensive WAN circuit resources by sharing that pool flexibly across multiple tenants and use cases. Similarly, the multi-service cloud network optimizes an operators pool of compute, storage and network access ports by allowing that pool to be shared flexibly across diverse applications, tenants and use cases.

What is cover in the remaining slides are some of what we see as fundamental building blocks for cloud network infrastructure and fully complementary with one another. We will call attention to some technology building blocks, and some service building blocks, which depend on those technology building blocks. The building blocks covered broadly fall under the following functional categories to the right of this slide.

We will also discuss two use cases that are worth considering, that may otherwise go unnoticed.

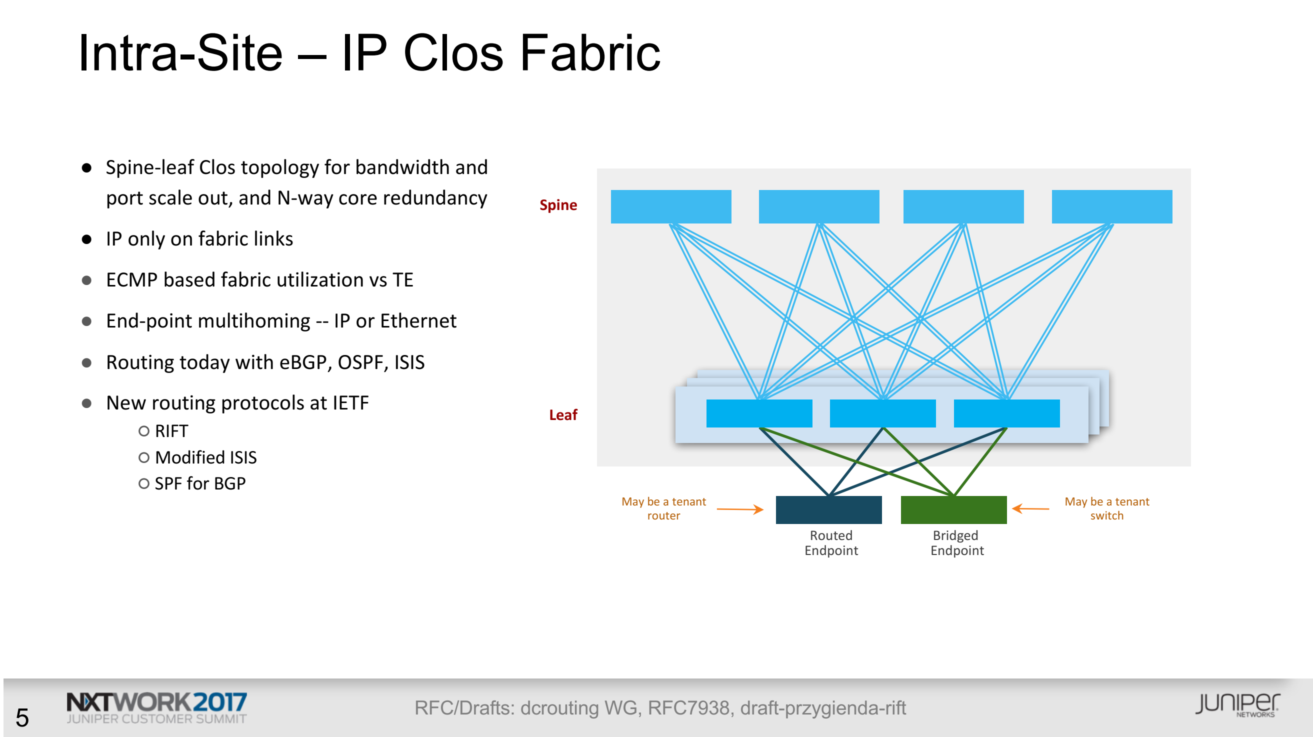
Wherever you see the word “TARGET” in all caps, it means we have not released the functionality yet, but are sharing to be transparent about our intended direction. These are all public efforts completed or ongoing in the standards bodies. On the grey bar at the bottom of the slides you can find the RFC or drafts that describe in detail what is summarized in the slide. We hope to work with our customers and the industry to drive all these standardization efforts forward to implementation.

IP Transport Network

“The Underlay”

NXTWORK2017
JUNIPER CUSTOMER SUMMIT

Let’s first cover some points relating to the underlay network



Intra-Site – IP Clos Fabric

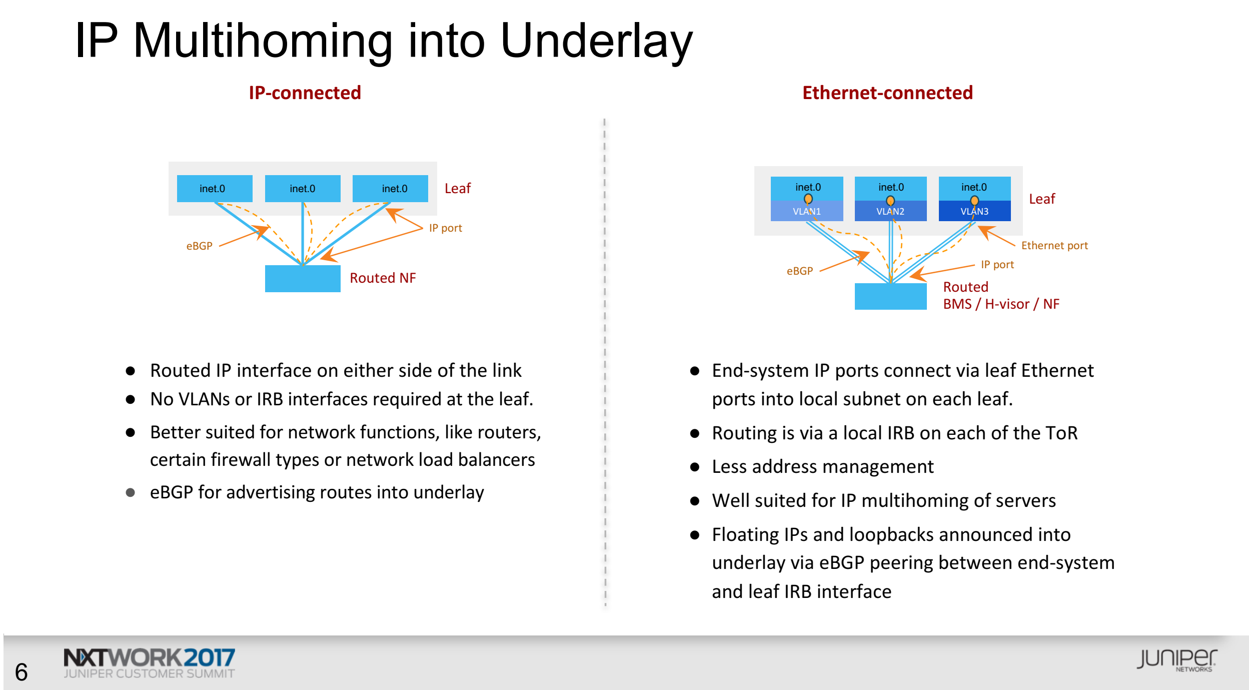
- Spine-leaf Clos topology for bandwidth and port scale out, and N-way core redundancy
- IP only on fabric links
- ECMP based fabric utilization vs TE
- End-point multihoming -- IP or Ethernet
- Routing today with eBGP, OSPF, ISIS
- New routing protocols at IETF
  - RIFT
  - Modified ISIS
  - SPF for BGP

Spine
Leaf
May be a tenant router
Routed Endpoint
Bridged Endpoint
May be a tenant switch

5 NXTWORK2017 JUNIPER CUSTOMER SUMMIT

RFC/Drafts: dcrouting WG, RFC7938, draft-przygienda-rift

- The core building block of cloud network infrastructure is a simple IP-based transport network
- Spine-leaf Clos topologies are ideal where horizontal scale-out of bandwidth and ports is a requirement
- High availability and full fabric utilization is realized here with IP ECMP load balancing. Traffic engineering, such as with MPLS, is much less relevant.
- For the routing control plane of the underlay, hop-by-hop eBGP as a fabric routing protocol is an option since it is a relatively simple control plane that can serve any scale fabric, with proven multi-vendor interoperability. This option is described in RFC 7938.
- There are also new fabric-optimized protocols in the works at the IETF, such as RIFT (Routing In Fat Trees), which can be considered for a clean-slate approach to fabric optimized routing and operations.
- Or modified flavors of existing protocols – such as ISIS with flood reduction optimizations, or SPF for BGP
- The pros and cons of these are being discussed and weighed at the newly formed IETF dcrouting WG
- You may also choose to stay with traditional IGPs (such as ISIS and OSPF) for small to moderate size deployments. These are not ideal for dense high scale fabric topologies due to sub-optimal flooding.



IP Multihoming into Underlay

IP-connected

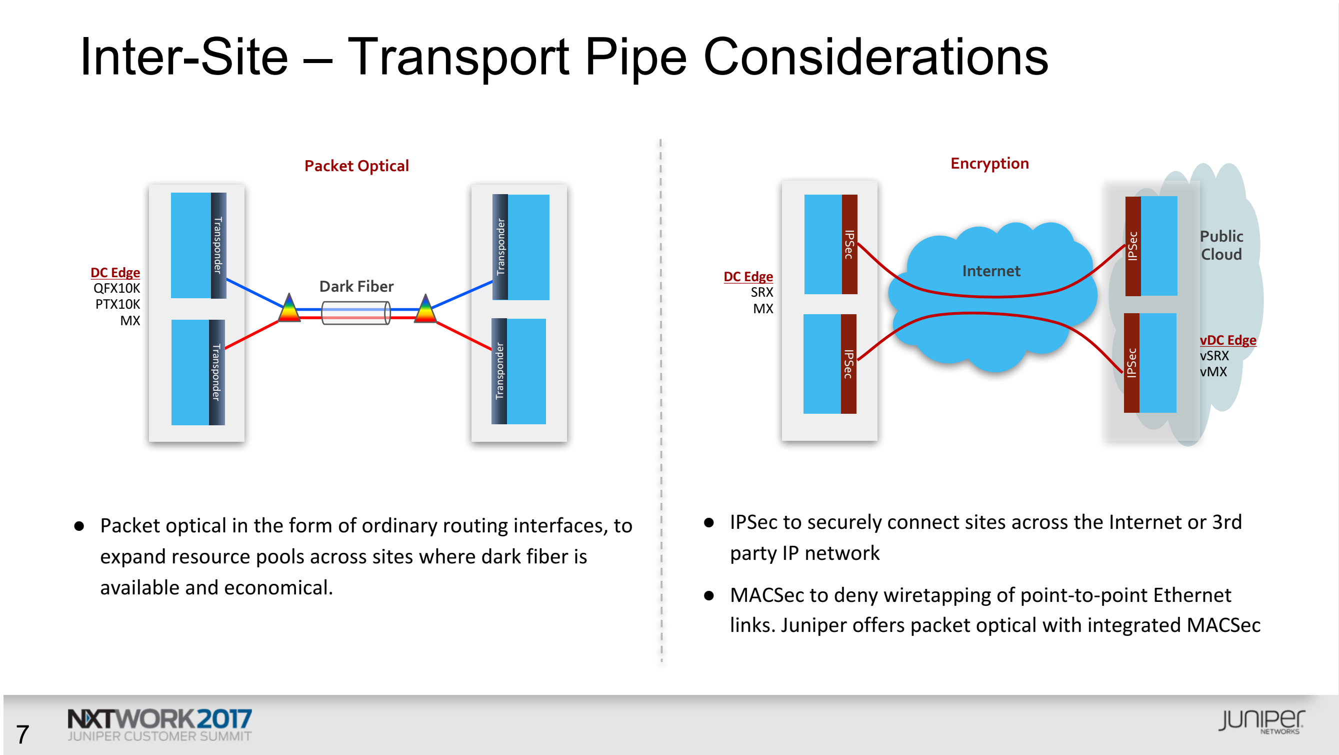
- Routed IP interface on either side of the link
- No VLANs or IRB interfaces required at the leaf.
- Better suited for network functions, like routers, certain firewall types or network load balancers
- eBGP for advertising routes into underlay

Ethernet-connected

- End-system IP ports connect via leaf Ethernet ports into local subnet on each leaf.
- Routing is via a local IRB on each of the ToR
- Less address management
- Well suited for IP multihoming of servers
- Floating IPs and loopbacks announced into underlay via eBGP peering between end-system and leaf IRB interface

6

- We spoke of the fabric in the last slide, and here we speak of how end-systems can connected and multihomed into that fabric
- High-availability for IP connected end-systems is achievable with N-way IP multihoming to the underlay
- There are two basic options for multihoming a user system into the IP underlay – the IP connected and Ethernet connected options
- Both of these models also apply to single homing
- With the IP Connected option, shown on the left
  - Both ends of the link are routed IP interfaces.
  - There is no VLANs or Integrated Routing and Bridging interfaces required at the leaf.
    - For the remainder of this presentation, I will refer to Integrated Routing and Bridging interfaces simply as IRB interfaces
  - The IP connected option is commonly used for network functions, like routers, certain L3 firewalls or network load balancers.
  - Advertisement of prefixes by connected network functions into the underlay is typically with eBGP. In this case, the network function would peer with the adjacent interface on all the leaf to which it is attached
- With the Ethernet Connected option, shown on the right
  - The user end of the link is typically a routed IP interface and the network end of the link is an Ethernet port which is attached to a unique local subnet on each leaf to which the end-system is multihomed
  - The subnet on a leaf is local to that leaf and does not extend to other leaf.
  - Routing is through a local IRB on each of the leaf
  - This form of end-system IP multihoming minimizes address consumption and management, and also network routing state -- and so is well suited for IP multihoming of a large numbers of servers.
  - Advertisement of floating IP and loopbacks by the servers into the underlay would typically use eBGP. In this case, the server would peer with the IRB interface on all the leaf to which it is attached.



Inter-Site – Transport Pipe Considerations

- Packet optical in the form of ordinary routing interfaces, to expand resource pools across sites where dark fiber is available and economical.

- IPSec to securely connect sites across the Internet or 3rd party IP network
- MACSec to deny wiretapping of point-to-point Ethernet links. Juniper offers packet optical with integrated MACSec

7 NXTWORK2017 JUNIPER CUSTOMER SUMMIT JUNIPer NETWORKS

- We spoke of the fabric, and then how end-systems can be connected and multihomed to the fabric. Now we highlight two considerations for interconnecting fabrics
- Packet optical, in the form of integrated DWDM transponders on routers is ideal for connecting sites where dark fiber is available and economical.
- A key benefit is that the capacity available in dark fiber enables resource pools that are not constrained to four walls.
- On the right hand we have two forms of encryption for concealing traffic and protecting both the traffic and the infrastructure
- IPSec is the transport of choice to securely connect sites across multiple hops such as the Internet or 3rd party IP network. A common example is connecting a physical site with a virtual site in the public cloud.
- MACSec is an option to securely connect across point-to-point Ethernet links between pods or sites that cross common spaces. Such as between buildings in a campus.
- Juniper offers packet optical with integrated MACSec

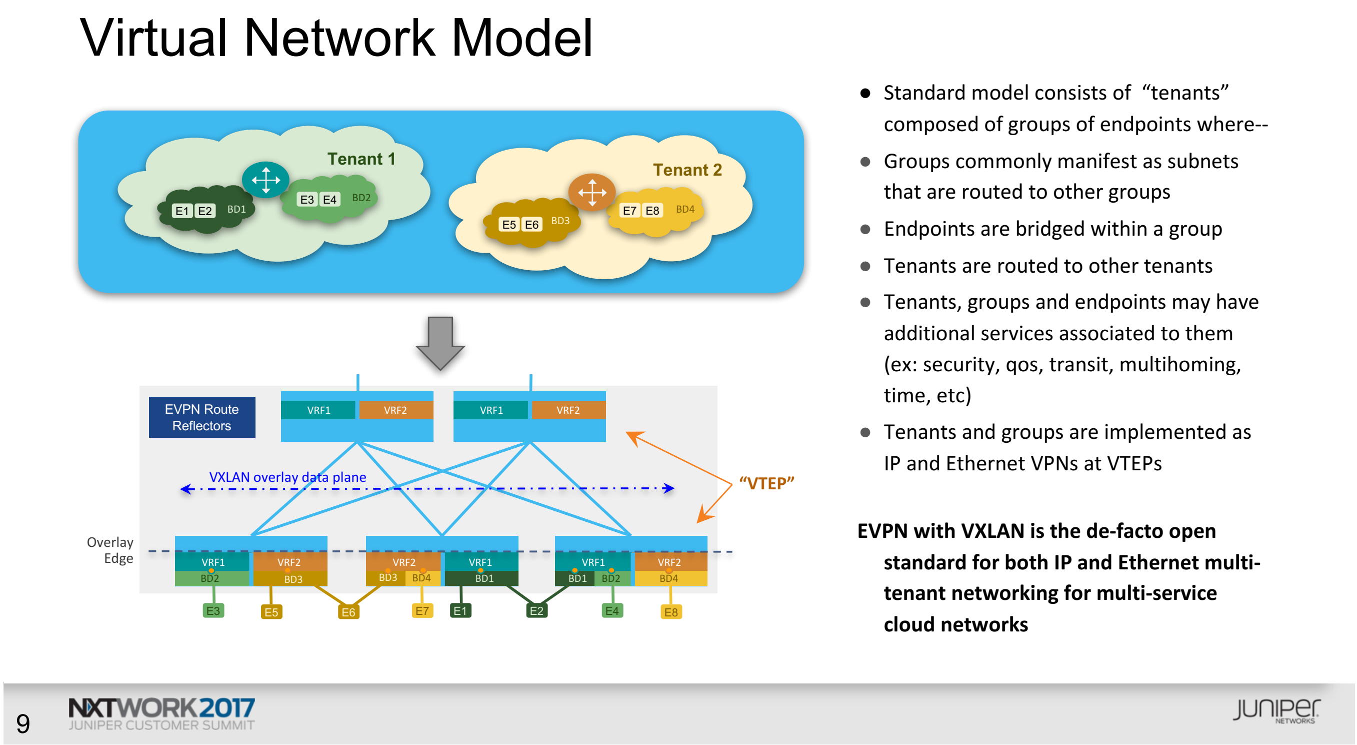
Standards-based
Network Virtualization
“The Overlay”

NXTWORK2017
JUNIPER CUSTOMER SUMMIT

JUNIPER NETWORKS

In the upcoming slides we will focus on the building blocks that enable multi-service cloud networking, primarily with a focus on network virtualization. Or what is known as the overlay.

The value of network virtualization is in transforming a single purpose network into a multi-tenant connectivity pool where any attachment port can be used for any purpose, as and when needed.

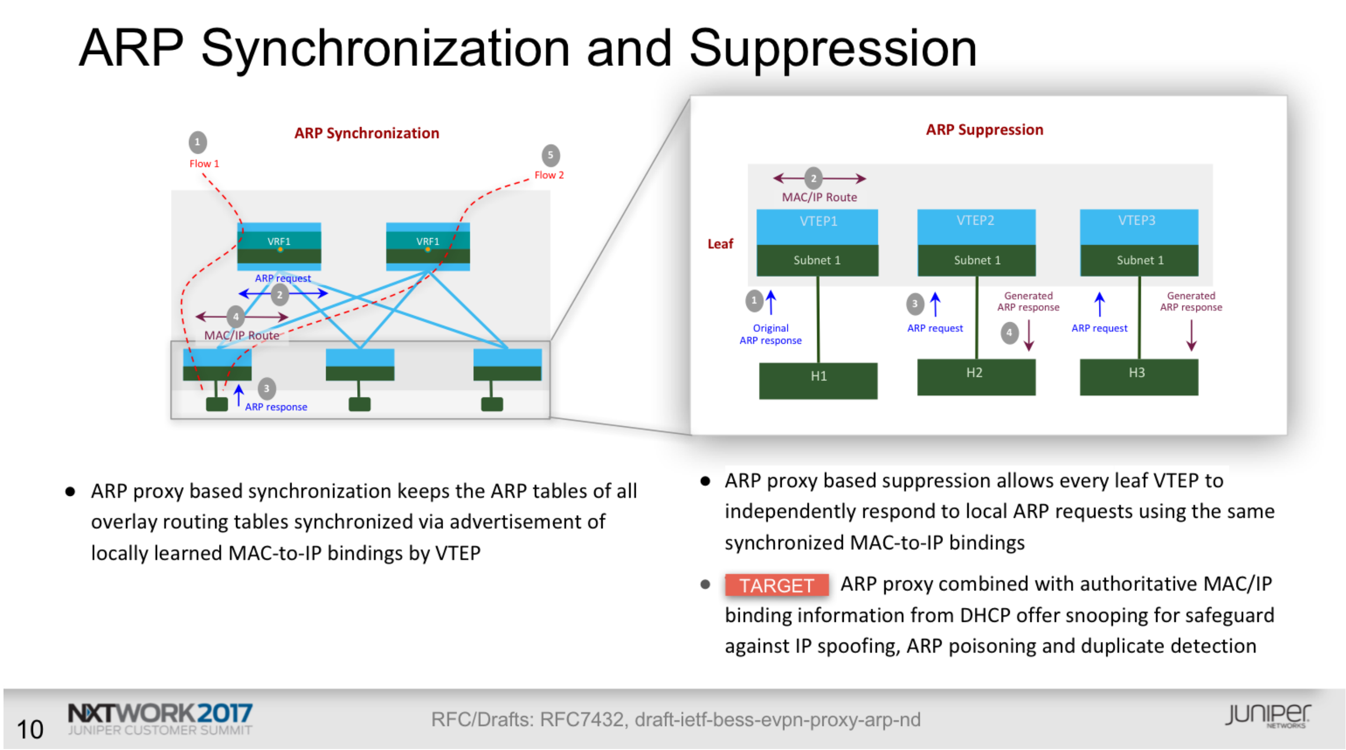


Virtual Network Model

- Standard model consists of "tenants" composed of groups of endpoints where--
- Groups commonly manifest as subnets that are routed to other groups
- Endpoints are bridged within a group
- Tenants are routed to other tenants
- Tenants, groups and endpoints may have additional services associated to them (ex: security, qos, transit, multihoming, time, etc)
- Tenants and groups are implemented as IP and Ethernet VPNs at VTEPs

EVPN with VXLAN is the de-facto open standard for both IP and Ethernet multi-tenant networking for multi-service cloud networks

- The virtual network model starts with the notion of a tenant, which contains groups of endpoints, and where the groups can communicate with other groups. Tenants may also communicate with other tenants
- Groups and endpoints may have quality-of-service and access control services associated to them. Other network services may also be part of a composite service – such as: security, multihoming, multicast, time, virtual wires, peering, service chaining, etc
- A group commonly manifests as an IP subnet on a distributed virtual bridge domain (BD) which can communicate with other groups and external networks via virtual routers (or VRF) across distributed virtual routed networks.
- The virtual bridge domains and virtual routed networks are referred to as overlays, which are implemented using dynamic tunnels over the IP underlay. Hence the name "overlay".
- The network devices which originate and terminate these overlay tunnels are referred to as Virtual Tunnel End Points or, simply, VTEP
- The de-facto standards-based overlay control plane for cloud networks is EVPN, with VXLAN tunnel data plane.
- EVPN currently leverages BGP infrastructure



ARP Synchronization and Suppression

- ARP proxy based synchronization keeps the ARP tables of all overlay routing tables synchronized via advertisement of locally learned MAC-to-IP bindings by VTEP

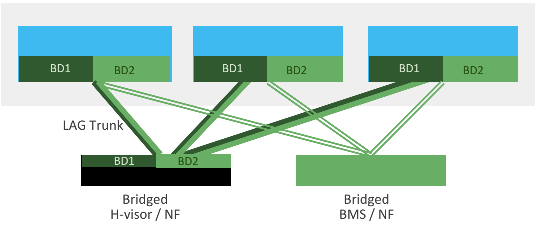
- ARP proxy based suppression allows every leaf VTEP to independently respond to local ARP requests using the same synchronized MAC-to-IP bindings
- TARGET ARP proxy combined with authoritative MAC/IP binding information from DHCP offer snooping for safeguard against IP spoofing, ARP poisoning and duplicate detection

RFC/Drafts: RFC7432, draft-ietf-bess-evpn-proxy-arp-nd

- ARP synchronization and suppression is a technology building block that we should cover first as it is important to many overlay service types where IP and Ethernet meet
- As most network operators know, ARP broadcasts and processing in large data centers can be overwhelming to both routers and hosts.
- The ARP synchronization and suppression discussed here uses the EVPN control plane to allow the cloud network to service ARP in a distributed manner for all overlay service types that leverage bridging.
- This can greatly reduce network and end-system load related to ARP
- Focusing on the left figure -- ARP synchronization, keeps the ARP tables of all overlay routing VTEP synchronized for the subnets that those routing VTEP serve. The routing VTEP no longer need to learn these bindings independently.
- Once a binding is learned at one VTEP, they are known everywhere via EVPN.
- On the right figure -- with ARP suppression, every leaf VTEP independently responds to local ARP requests using the same shared MAC-to-IP binding data as leveraged by ARP synchronization.

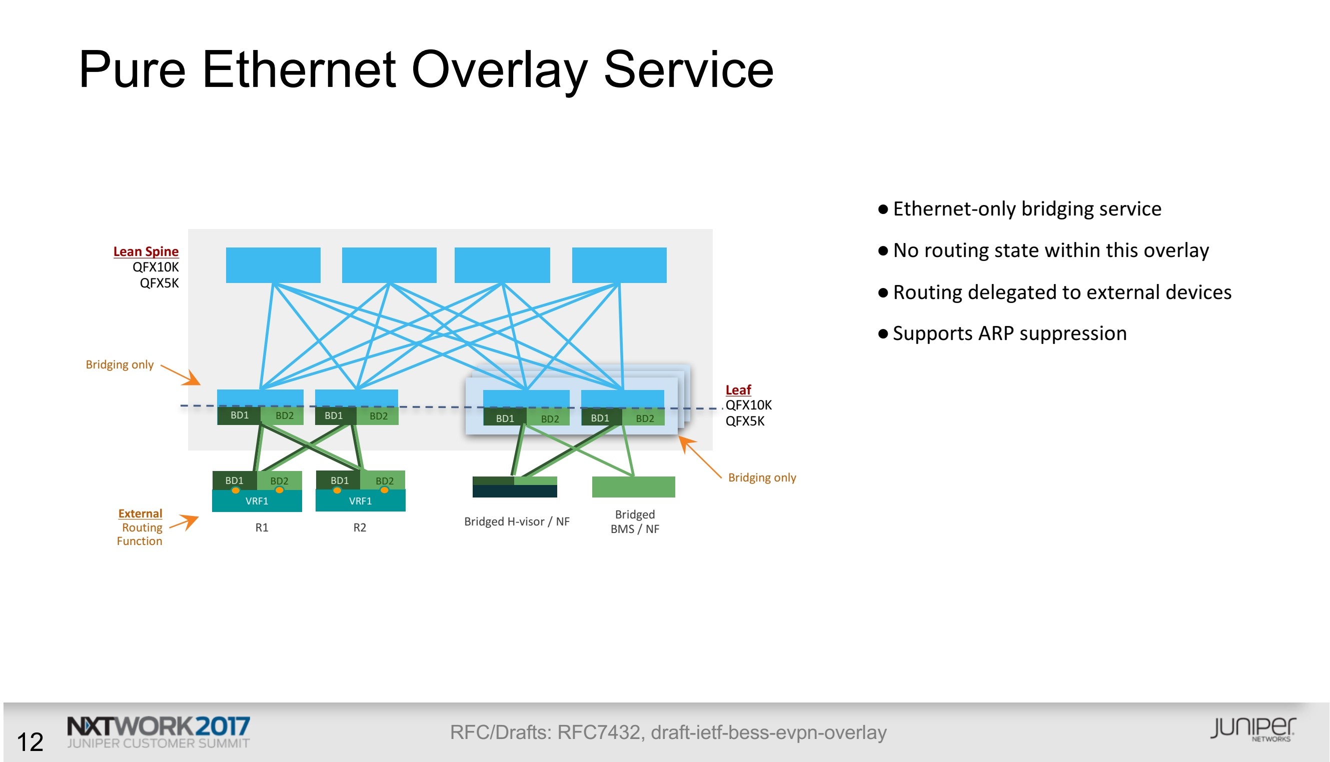
End-System Ethernet Multihoming

- EVPN enables N-way “scale-out” Ethernet multihoming where N can be greater than 2
- No ICL link required
- A multi-homed end-system is identified in the overlay by a unique Ethernet Segment ID (ESI)
- TARGET ESI may be provisioned automatically via LACP or xSTP snooping (i.e. EVPN Auto-ESI)



RFC/Drafts: RFC7432, draft-ietf-bess-evpn-overlay

- Ethernet multihoming is another technology building block that we should dig into before we discuss overlay service types, and use cases
- This building block is needed for high-availability for Ethernet connected end-systems and networks
- We achieve this here using EVPN, which enables N-way “scale-out” Ethernet multihoming where N can be greater than 2
- EVPN-based Ethernet multihoming applies to all overlay service types that have bridging
- One thing to note is that there is no direct inter-leaf links required with EVPN based multi-homing. But you can have it if you need it.
- In EVPN, a multi-homed end-system is identified in the overlay by a unique Ethernet Segment ID, referred to simply as ESI.

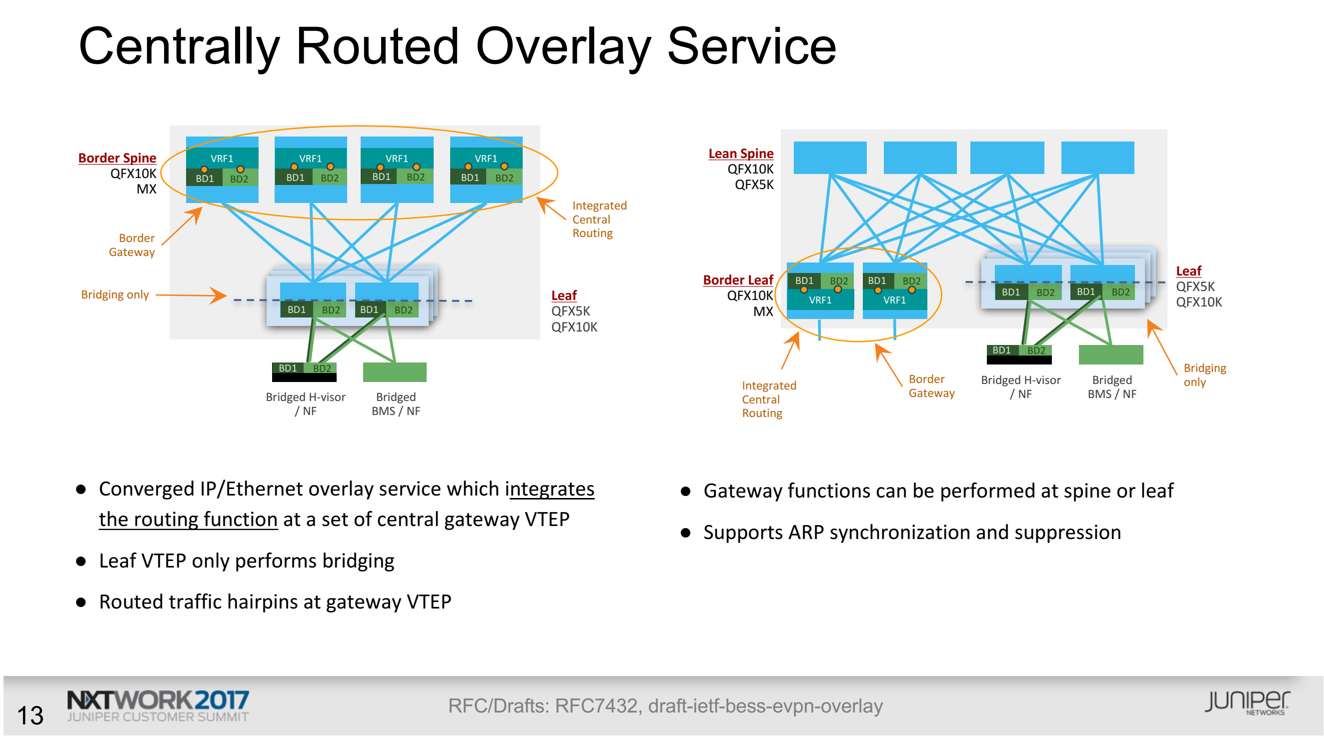


Pure Ethernet Overlay Service

- Ethernet-only bridging service
- No routing state within this overlay
- Routing delegated to external devices
- Supports ARP suppression

RFC/Drafts: RFC7432, draft-ietf-bess-evpn-overlay

- The Pure Ethernet Overlay service type is the first of the service building blocks we will cover
- This basic overlay model enables a pure Ethernet bridging service, as the name implies
- There is no routing state by default within an instance of this overlay service
- Routing can be delegated to external devices when required.
- Hosting providers may choose to allow their tenants to bring their own routing function
- Although this is a pure Ethernet bridging service, the leaf VTEP can provide ARP suppression if needed for the service



Centrally Routed Overlay Service

- Converged IP/Ethernet overlay service which integrates the routing function at a set of central gateway VTEP
- Leaf VTEP only performs bridging
- Routed traffic hairpins at gateway VTEP

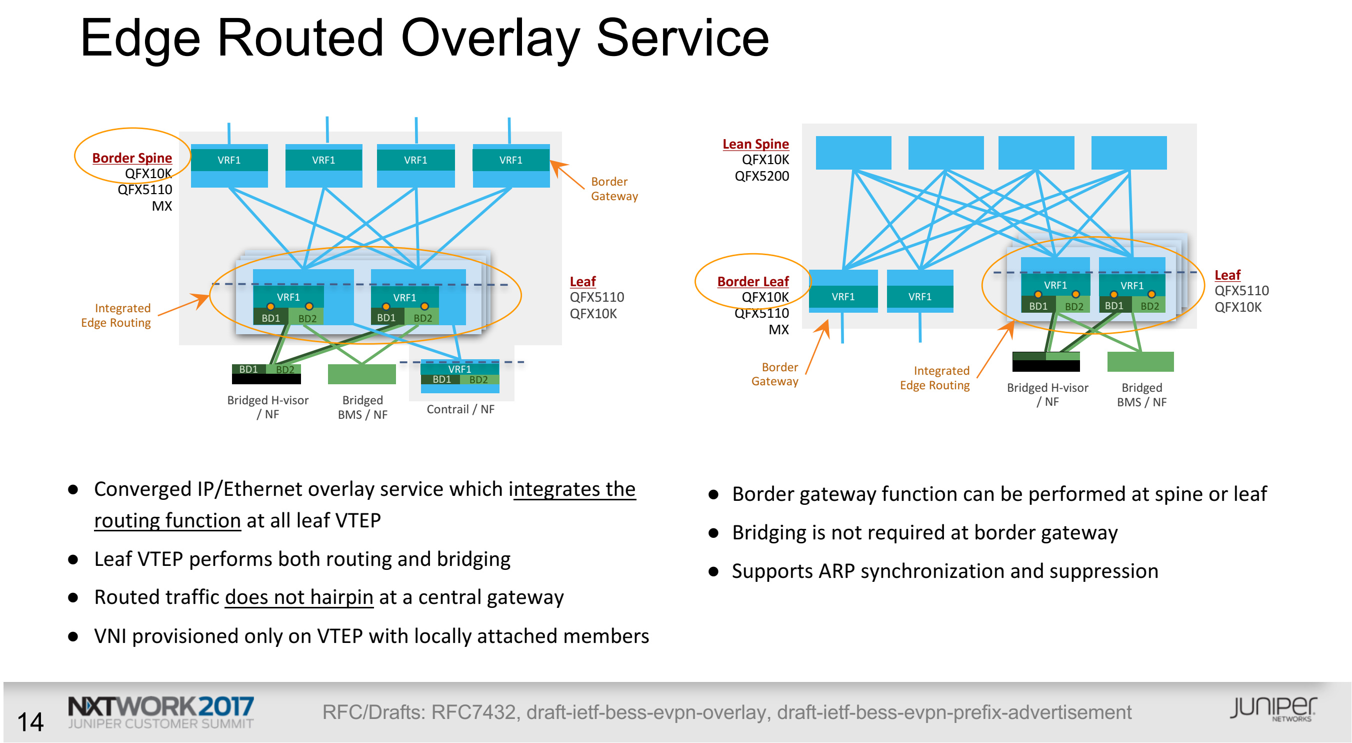
- Gateway functions can be performed at spine or leaf
- Supports ARP synchronization and suppression

13 NXTWORK2017 JUNIPER CUSTOMER SUMMIT

RFC/Drafts: RFC7432, draft-ietf-bess-evpn-overlay

JUNIPER NETWORKS

- The centrally routed service integrates the routing function at a set of central gateways, and can be used when routed traffic must go through a centralized gateway. Or can be used wherever else distributed routing is not preferred or not possible, such as when using platforms based on Broadcom T2 silicon
- In this service, the leaf VTEP only performs bridging. Routed traffic between hosts connected to the same leaf VTEP hairpins at gateway VTEP.
- This service type can be optimized with ARP suppression
- The central overlay routing gateway function can be performed at either the spine layer or at the leaf layer
- The central routing gateways can also serve as border gateways, that advertise the prefix routes of the local tenancies to north facing networks



Edge Routed Overlay Service

- Converged IP/Ethernet overlay service which integrates the routing function at all leaf VTEP
- Leaf VTEP performs both routing and bridging
- Routed traffic does not hairpin at a central gateway
- VNI provisioned only on VTEP with locally attached members

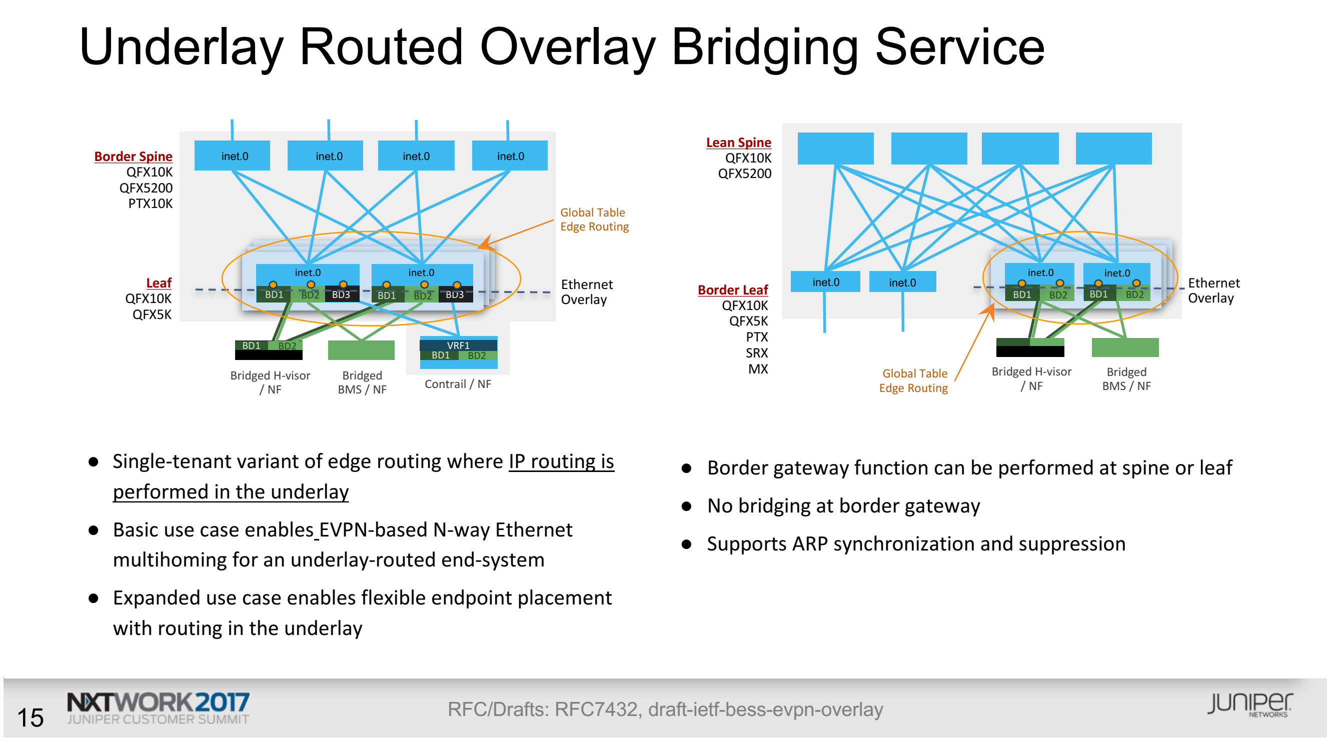
- Border gateway function can be performed at spine or leaf
- Bridging is not required at border gateway
- Supports ARP synchronization and suppression

14 NXTWORK2017 JUNIPER CUSTOMER SUMMIT

RFC/Drafts: RFC7432, draft-ietf-bess-evpn-overlay, draft-ietf-bess-evpn-prefix-advertisement

JUNIPER NETWORKS

- The Edge Routed Overlay service integrates routing at the leaf -- so routed traffic between hosts connected to the same leaf do not hairpin at a central gateway
- The leaf VTEP performs both routing and bridging via local IRB interface. In order to do this, the leaf VTEP platform must support VXLAN routing
- Juniper’s fully standards-based implementation is inherently symmetric with single touch provisioning – which is that, a subnet is only provisioned on leaf with locally attached members of that subnet
- Generally how it works is that, leaf VTEP advertise host routes learned locally via ARP. All inter-subnet east-west traffic, as well as south-bound traffic from border gateways follow these IP host routes which are advertised using the EVPN Prefix Route
- ARP synchronization and suppression is built into this service type
- In this service type, the border gateway function can be IP only and can be performed at the spine layer or at the leaf layer
- There is no requirement for bridging support at the border gateway
- Border gateways advertise aggregate routes for local tenancies to north facing networks



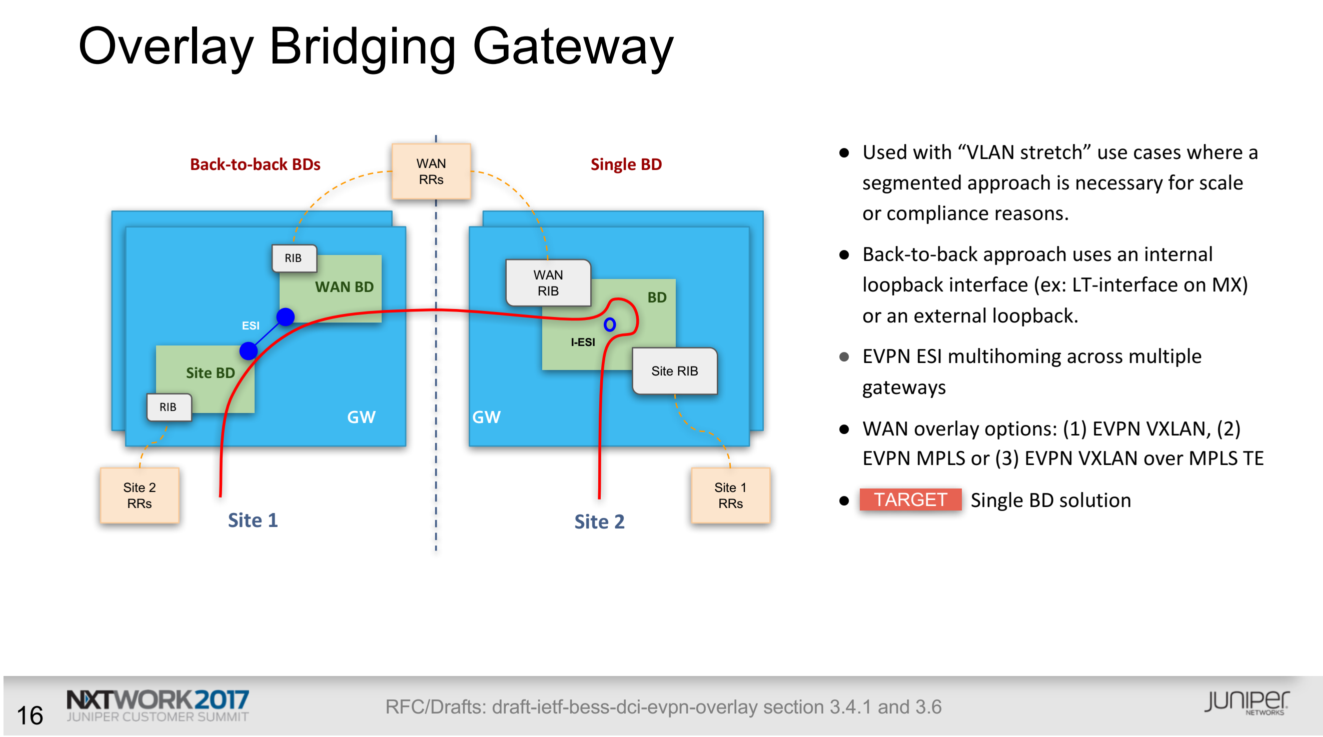
Underlay Routed Overlay Bridging Service

- Single-tenant variant of edge routing where IP routing is performed in the underlay
- Basic use case enables EVPN-based N-way Ethernet multihoming for an underlay-routed end-system
- Expanded use case enables flexible endpoint placement with routing in the underlay

- Border gateway function can be performed at spine or leaf
- No bridging at border gateway
- Supports ARP synchronization and suppression

RFC/Drafts: RFC7432, draft-ietf-bess-evpn-overlay

- This service type is a variant of the edge routing service type, where IP routing is performed in the underlay as opposed to an IP overlay – as such, there are no VRFs. Which makes this a single IP tenant model
- The basic use case enables an end-system to be Ethernet multihomed into the underlay using EVPN-based LAG, in the place of IP multihoming or MC-LAG. An example of this use case is where an operator prefers to use link bonding to attach an SDN-enabled hypervisor into the underlay.
- In this use case, standard underlay multicast is also an option.
- There is an expanded use case which combines seamless endpoint placement in the bridging overlay with single-tenant symmetric inter-subnet routing in the IP underlay.
- This service type also requires VXLAN routing capability on the leaf



Overlay Bridging Gateway

Back-to-back BDs
WAN RRs
Single BD
RIB
WAN BD
ESI
Site BD
RIB
GW
Site 2 RRs
Site 1
WAN RIB
BD
I-ESI
Site RIB
GW
Site 2
Site 1 RRs

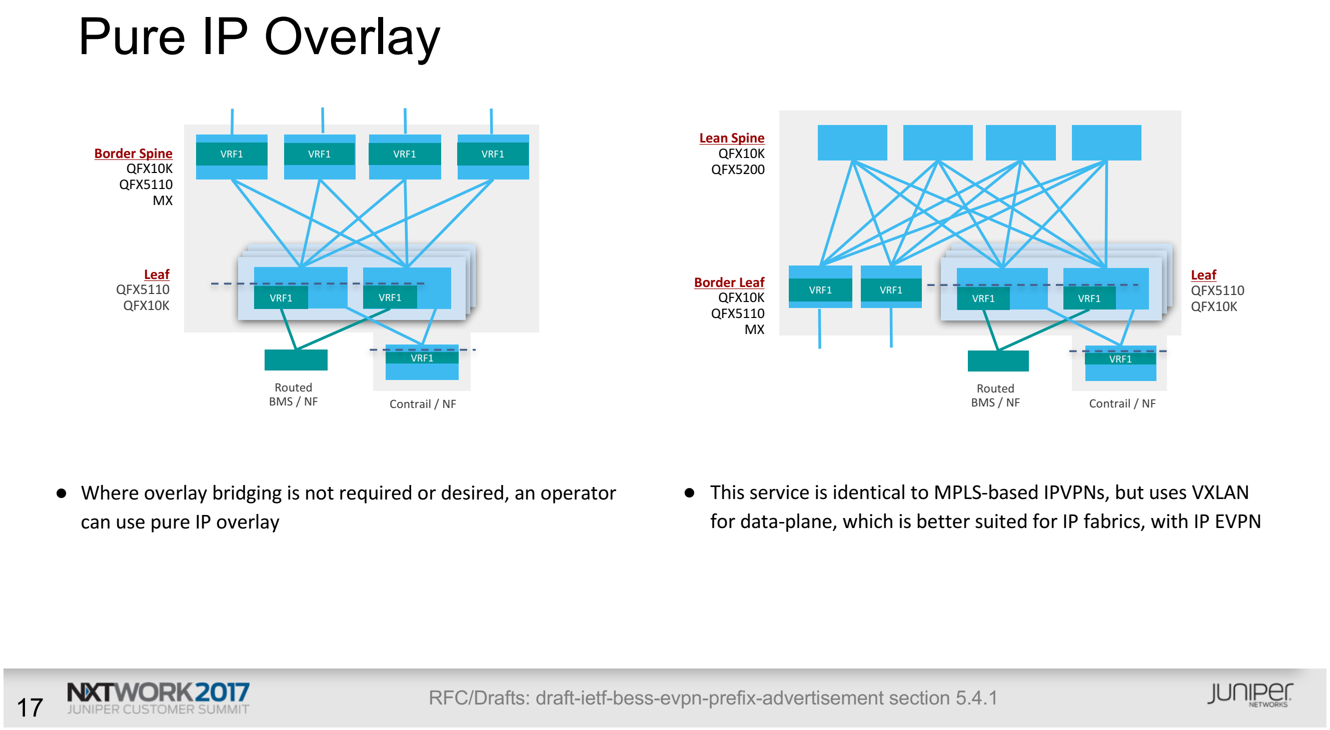
- Used with “VLAN stretch” use cases where a segmented approach is necessary for scale or compliance reasons.
- Back-to-back approach uses an internal loopback interface (ex: LT-interface on MX) or an external loopback.
- EVPN ESI multihoming across multiple gateways
- WAN overlay options: (1) EVPN VXLAN, (2) EVPN MPLS or (3) EVPN VXLAN over MPLS TE
- TARGET Single BD solution

16 NXTWORK2017 JUNIPER CUSTOMER SUMMIT

RFC/Drafts: draft-ietf-bess-dci-evpn-overlay section 3.4.1 and 3.6

JUNIPER NETWORKS

- An overlay bridging gateway function can be employed where inter-site “VLAN stretch” is required, but where a segmented approach is necessary for scale or administrative control reasons.
- The current solution on MX can use an internal logical interface for connecting DC-facing bridge domain with WAN-facing bridge domain. Use of physical interface-based loopbacks is also an option with any of the other platforms. You can see this option on the left side of the illustration.
- Multihoming across the multiple overlay bridging gateways of a site is achievable using basic EVPN ESI multihoming.
- VXLAN, MPLS and VXLAN-over-MPLS tunnel are all options on the WAN segment of the segmented stretch. However, VXLAN based WAN overlay is the trend.
- TARGET: Our target is a single table solution based on EVPN VXLAN-to-VXLAN "gateway" model that you see on the right hand side of the illustration
- The single bridge table solution will support full line rate performance and will also be available on the QFX10K

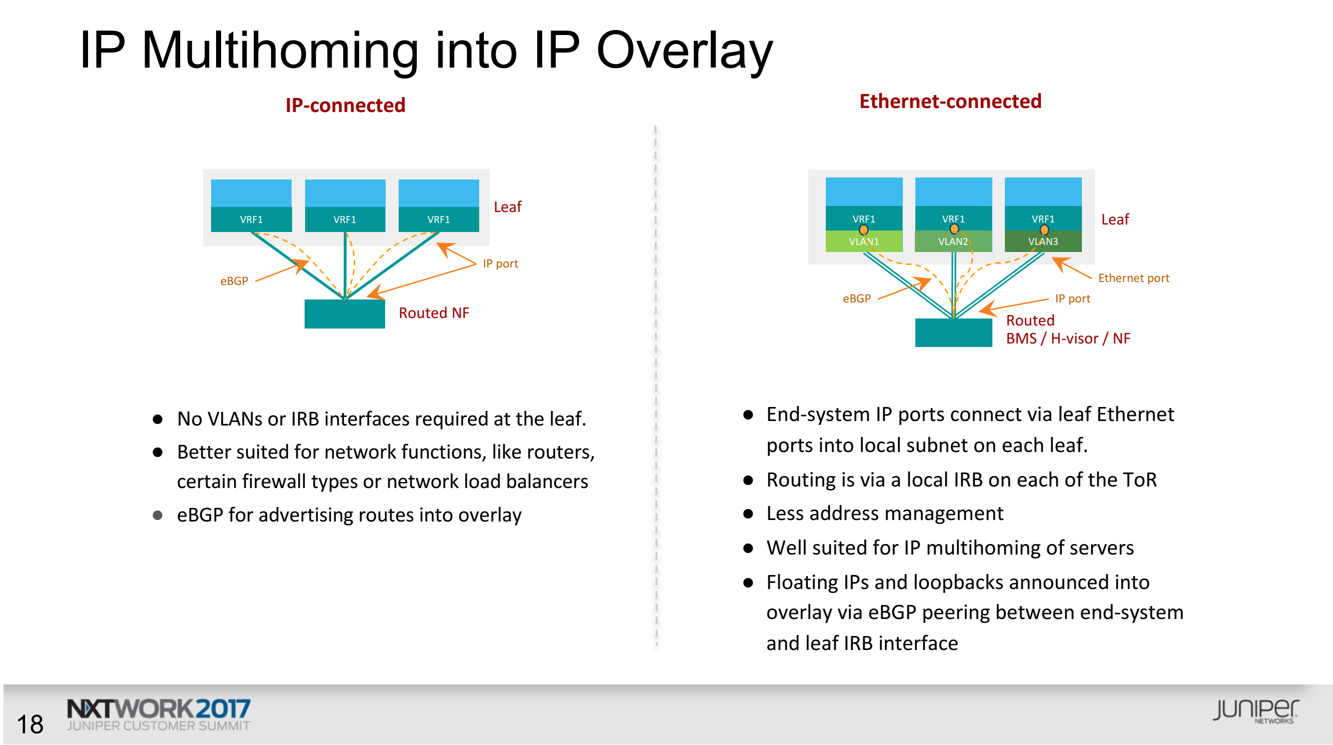


Pure IP Overlay

- Where overlay bridging is not required or desired, an operator can use pure IP overlay
- This service is identical to MPLS-based IPVPNs, but uses VXLAN for data-plane, which is better suited for IP fabrics, with IP EVPN

RFC/Drafts: draft-ietf-bess-evpn-prefix-advertisement section 5.4.1

- Our final overlay service type is the Pure IP Overlay service type
- For IP-only multi-tenant networking or for tenancies where overlay bridging is not required or desired, an operator can use pure IP overlay
- This technology is similar to MPLS-based IPVPN, but uses VXLAN for data-plane, which is better suited for IP fabrics
- This service type uses EVPN Prefix routes, which is the same used in the edge routed overlay, to provide IPVPN capability over VXLAN tunnels using the EVPN control plane



IP Multihoming into IP Overlay

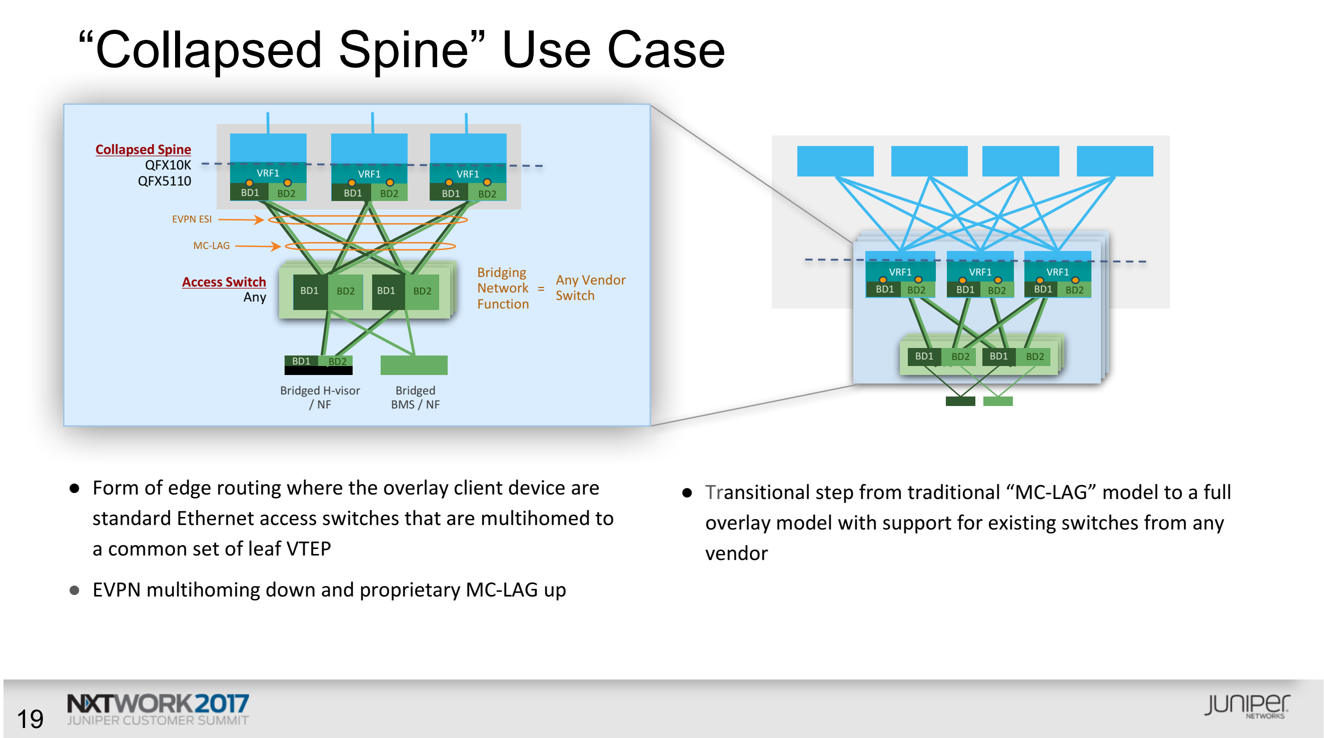
IP-connected

- No VLANs or IRB interfaces required at the leaf.
- Better suited for network functions, like routers, certain firewall types or network load balancers
- eBGP for advertising routes into overlay

Ethernet-connected

- End-system IP ports connect via leaf Ethernet ports into local subnet on each leaf.
- Routing is via a local IRB on each of the ToR
- Less address management
- Well suited for IP multihoming of servers
- Floating IPs and loopbacks announced into overlay via eBGP peering between end-system and leaf IRB interface

- High-availability for IP overlay connected end-systems is achievable with N-way IP multihoming
- There are two basic options for multihoming a user system into an IP overlay – the IP connected and Ethernet connected options
- These are same as IP multihoming into underlay except the routing table here is a VRF and not the underlay routing table



"Collapsed Spine" Use Case

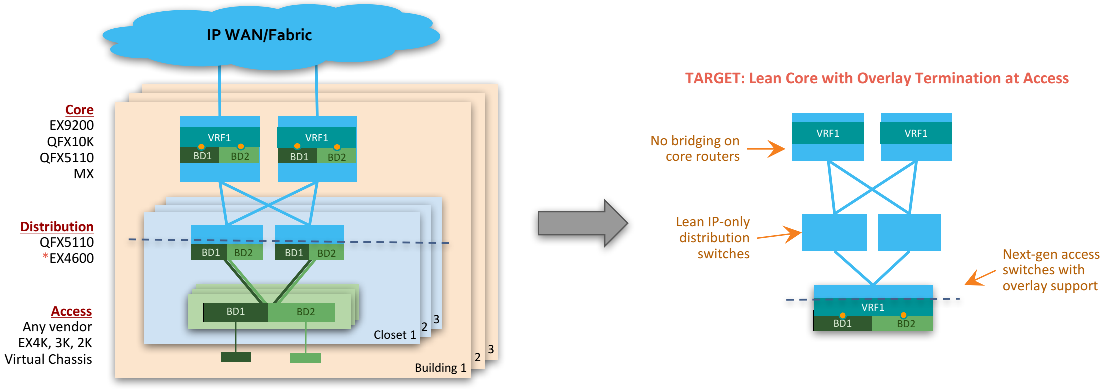
Collapsed Spine
QFX10K
QFX5110
VRF1
BD1
BD2
EVPN ESI
MC-LAG
Access Switch
Any
Bridging Network Function = Any Vendor Switch
Bridged H-visor / NF
Bridged BMS / NF

- Form of edge routing where the overlay client device are standard Ethernet access switches that are multihomed to a common set of leaf VTEP
- EVPN multihoming down and proprietary MC-LAG up
- Transitional step from traditional "MC-LAG" model to a full overlay model with support for existing switches from any vendor

- In the next two slides we call attention to two use cases that leverage some of the building blocks we discussed so far
- Here the "Collapsed Spine" use case is an interesting example of the edge routed overlay service -- where the user devices are standard Ethernet access switches that are multihomed to a cluster of leaf VTEP -- which we refer to as the Collapsed Spine.
- We are seeing customers take an interest in this use case as a transitional step from traditional "MC-LAG" model to an STP-free full overlay model – because it allows them to leverage existing access switches from their existing vendors.
- This is an example of where EVPN can be used for additional redundancy and capacity, versus MC-LAG, with its native N-way multihoming capability.
- This use case is implemented as EVPN ESI multi-homing from the Collapsed Spine towards Access Switch. LAG or MC-LAG multi-homing from Access Switch towards Collapsed Spine

Evolved Campus Use Case

- Transition from legacy Ethernet-based core to a simple, scalable and resilient IP core
- Interconnect end-to-end with segmentation using EVPN
- Use any vendor access switch based on requirements
- TARGET Overlay support at access switch with advanced access feature set
- IPSec and MACSec options for transport encryption
- Distributed security with SDSN


IP WAN/Fabric
Core
EX9200
QFX10K
QFX5110
MX
Distribution
QFX5110
*EX4600
Access
Any vendor
EX4K, 3K, 2K
Virtual Chassis
VRF1
BD1
BD2
Closet 1
2
3
Building 1
TARGET: Lean Core with Overlay Termination at Access
No bridging on core routers
Lean IP-only distribution switches
Next-gen access switches with overlay support

- Another example of a use case where we can apply the building blocks we discussed, is the Evolved Campus use case
- Here we are looking to transition a campus from STP or other legacy Ethernet-based core to simple, scalable and resilient IP based core
- The overlay service types that we discussed in the previous slides are used to maintain parity with the legacy network functionality, while also to ultimately deliver new services that were previously not possible
- EVPN VXLAN also brings site-to-site virtual network based connectivity and segmentation. And in doing so, it completes the end-to-end picture.
- Any vendor platform -- including the Juniper EX series -- can be used at the access layer, based on existing access requirements, while also evolving the rest of the end-to-end network fabric
- IPSec and MACSec options exist for transport encryption
- TARGET: With future access switch platforms our objective is to bring edge-to-edge overlay virtualization directly from the access switch without compromising on other access switch requirements

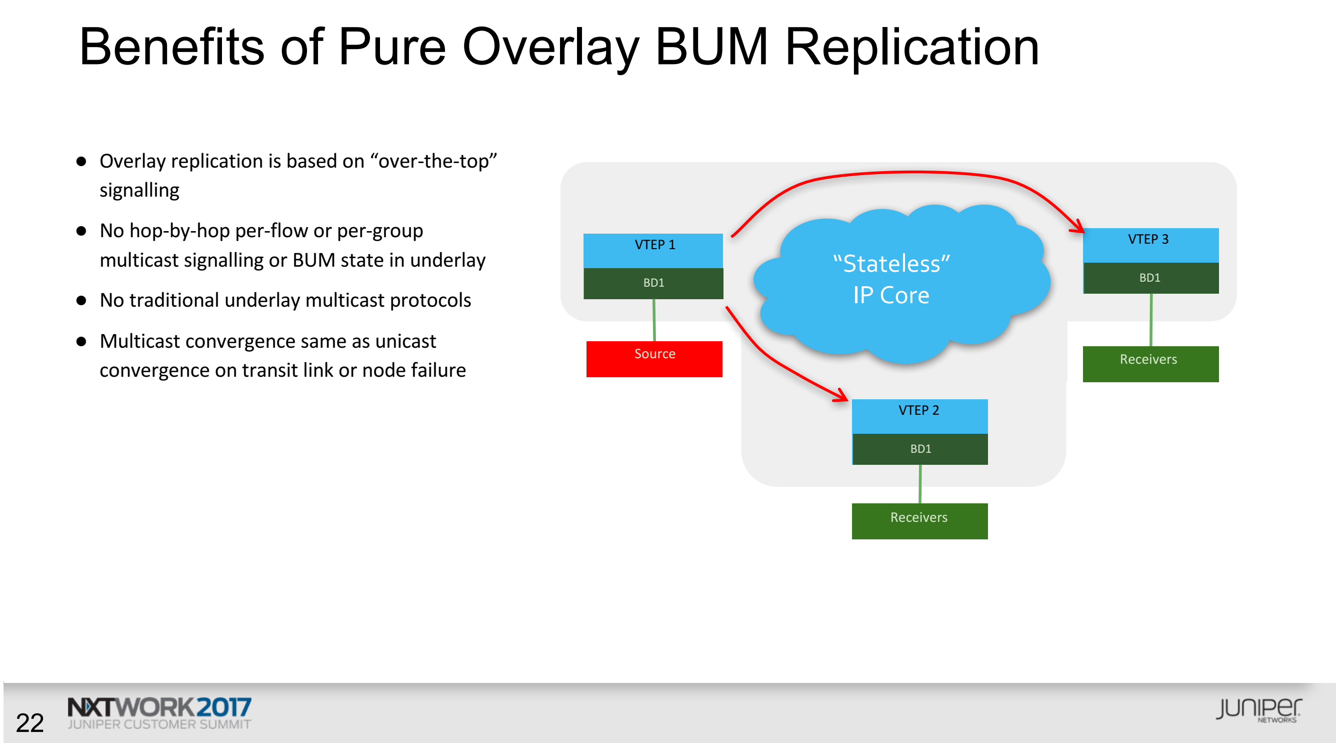
Standards-based

Overlay Replication

NXTWORK2017
JUNIPER CUSTOMER SUMMIT

JUNIPER NETWORKS

- The gnarliest topic in networking is multi-destination forwarding of Ethernet Broadcast, Unknown Unicast and IP Multicast.
- Which are collectively referred to as “BUM” traffic.
- In the next set of slides I’ll talk about BUM in the overlay, with an emphasis on IP multicast.



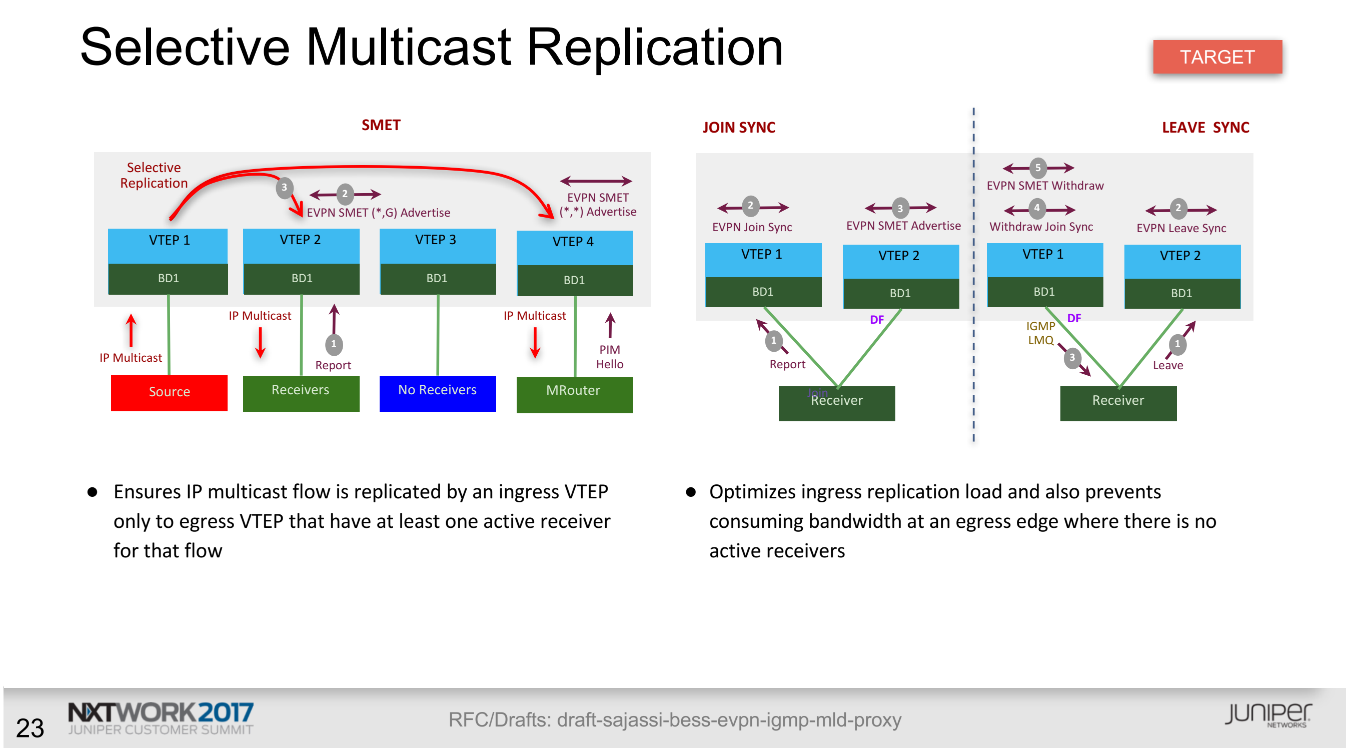
Benefits of Pure Overlay BUM Replication

- Overlay replication is based on “over-the-top” signalling
- No hop-by-hop per-flow or per-group multicast signalling or BUM state in underlay
- No traditional underlay multicast protocols
- Multicast convergence same as unicast convergence on transit link or node failure

VTEP 1
BD1
Source
“Stateless” IP Core
VTEP 3
BD1
Receivers
VTEP 2
BD1
Receivers

22 NXTWORK2017 JUNIPER CUSTOMER SUMMIT

- There are major benefits of pure overlay based BUM replication as compared to underlay based replication for overlays
- There is no hop-by-hop per-flow or per-group multicast signaling and state with pure overlay replication – so no tenant state on intermediate routers, which allows for a lean core network. This is because overlay replication involves edge to edge “over-the-top” signaling
- No underlay multicast protocols (PIM, MSDP, mLDP, etc) means no corresponding complexity
- Multicast convergence is the same as unicast convergence on transit node or transit link failure, since overlay multicast state remains unchanged at VTEP.
- And finally, overlay replication can achieve the same efficiency as P2MP underlay replication in certain topologies, with the optimizations I will cover in next few slides



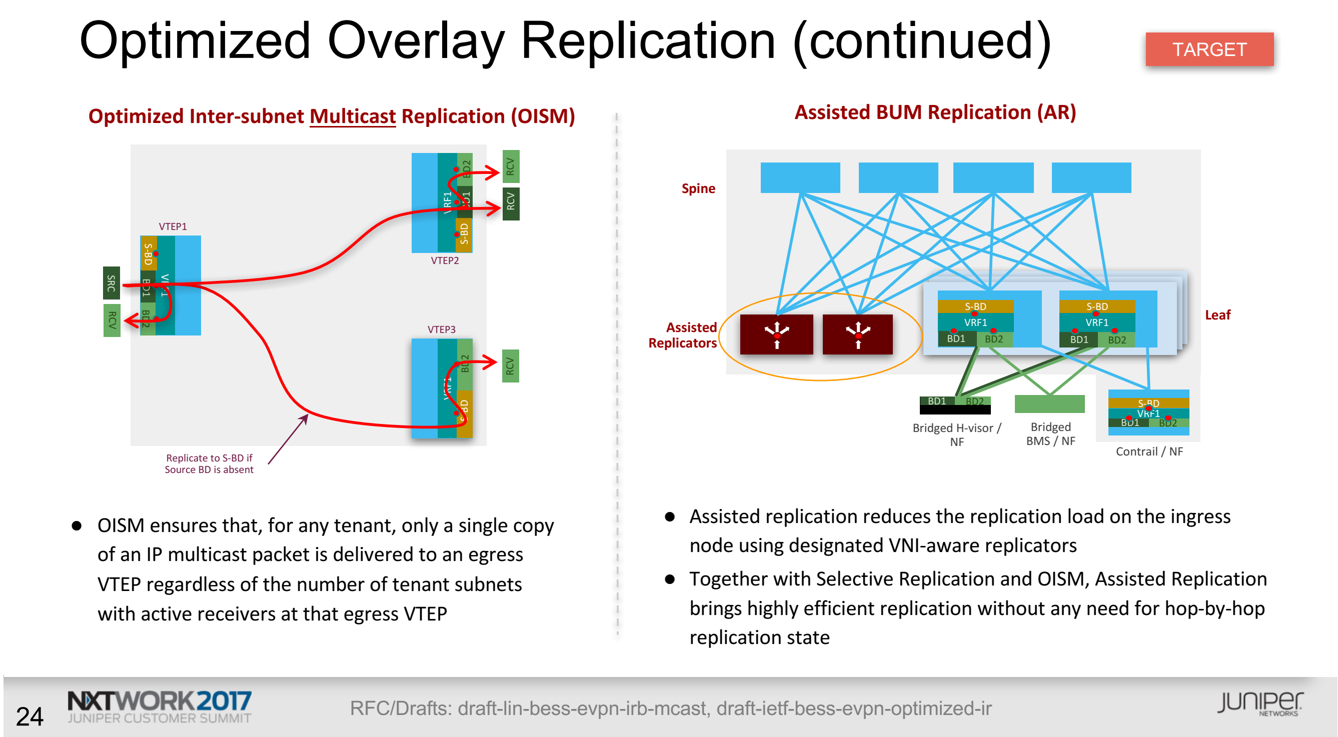
Selective Multicast Replication

TARGET

- Ensures IP multicast flow is replicated by an ingress VTEP only to egress VTEP that have at least one active receiver for that flow
- Optimizes ingress replication load and also prevents consuming bandwidth at an egress edge where there is no active receivers

RFC/Drafts: draft-sajassi-bess-evpn-igmp-mld-proxy

- The first of these optimizations involves a technology building block called Selective Multicast Replication, which is focused on optimizing IP multicast in a bridging overlay.
- This is equivalent to IGMP snooping in a hop-by-hop bridged network, but with no learning or state required at intermediate nodes.
- It ensures that, within any single Ethernet Bridging Overlay, an IP multicast flow is only replicated to remote VTEP with at least one or more active receivers for that flow.
- This optimizes replication load on the ingress VTEP and also prevents consuming bandwidth at an egress VTEP if there are no active receivers there



Optimized Overlay Replication (continued)

TARGET

Optimized Inter-subnet Multicast Replication (OISM)

Replicate to S-BD if Source BD is absent

- OISM ensures that, for any tenant, only a single copy of an IP multicast packet is delivered to an egress VTEP regardless of the number of tenant subnets with active receivers at that egress VTEP

Assisted BUM Replication (AR)

- Assisted replication reduces the replication load on the ingress node using designated VNI-aware replicators
- Together with Selective Replication and OISM, Assisted Replication brings highly efficient replication without any need for hop-by-hop replication state

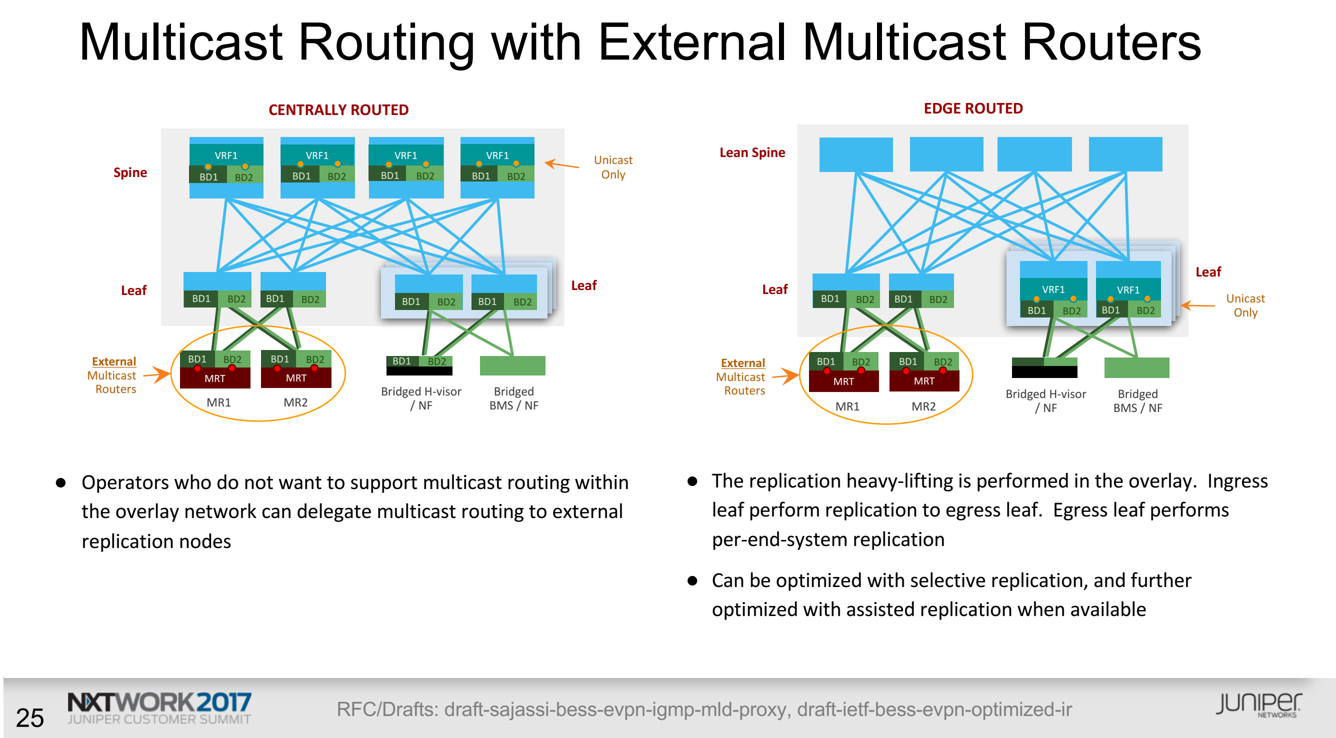
24 NXTWORK2017 JUNIPER CUSTOMER SUMMIT

RFC/Drafts: draft-lin-bess-evpn-irb-mcast, draft-ietf-bess-evpn-optimized-ir

JUNIPER NETWORKS

- Optimized Inter-subnet Replication and Assisted Replication are two more technology building blocks that can play unique and important roles in optimizing overlay replication
- Optimized Inter-subnet Replication, shown on the left
  - Ensures that, for any tenant, only a single copy of an IP multicast packet is delivered to an egress VTEP regardless of the number of subnets of that tenant with active receivers at that egress VTEP
  - OISM only applies to IP Multicast and supports only edge routed overlay service types
  - There are new procedures required for OISM, but no additional EVPN route types are introduced
- Assisted Replication, shown on the right
  - Eliminates replication load on the ingress node for a bridged overlay
  - Rather than directly replicating to egress VTEP, an ingress VTEP forwards BUM to designated replicators that perform BUM replication on it's behalf
  - BUM flows would be load balanced by an ingress VTEP across the replicators in a replicator set
  - Convergence remains as fast as unicast convergence on a replicator node failure
  - Assisted replication supports Pure Ethernet, Centrally Routed and Edge Routed service types

- Together with Selective Replication and OISM, Assisted Replication brings IP multicast replication efficiency on par with underlay IP replication -- without any need for hop-by-hop replication state



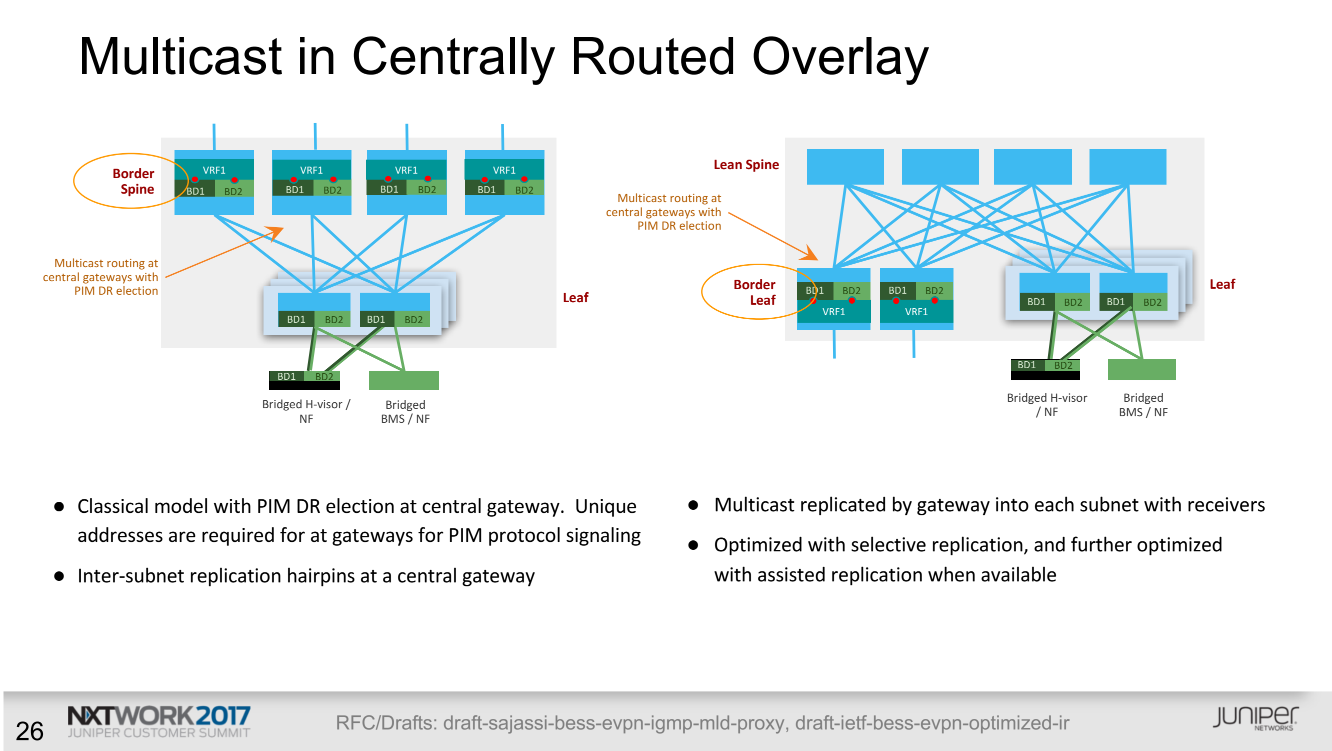
Multicast Routing with External Multicast Routers

- Operators who do not want to support multicast routing within the overlay network can delegate multicast routing to external replication nodes

- The replication heavy-lifting is performed in the overlay. Ingress leaf perform replication to egress leaf. Egress leaf performs per-end-system replication
- Can be optimized with selective replication, and further optimized with assisted replication when available

RFC/Drafts: draft-sajassi-bess-evpn-igmp-mld-proxy, draft-ietf-bess-evpn-optimized-ir

- Applying these multicast technology building to the overlay service types we discussed.
- The first is Multicast Routing using External Multicast Routers
- This option can be applied to Pure Ethernet, Centrally Routed or Edge Routed service types
- Operators who do not want to support multicast routing in the physical network can delegate multicast routing to external replication nodes
- Hosting providers may choose to allow their tenants to bring their own multicast routing function
- The replication heavy-lifting is performed in the overlay -- ingress VTEP perform replication to egress VTEP and egress VTEP then performs per-end-system replication
- This option can be optimized with selective replication, and further optimized with assisted replication when available



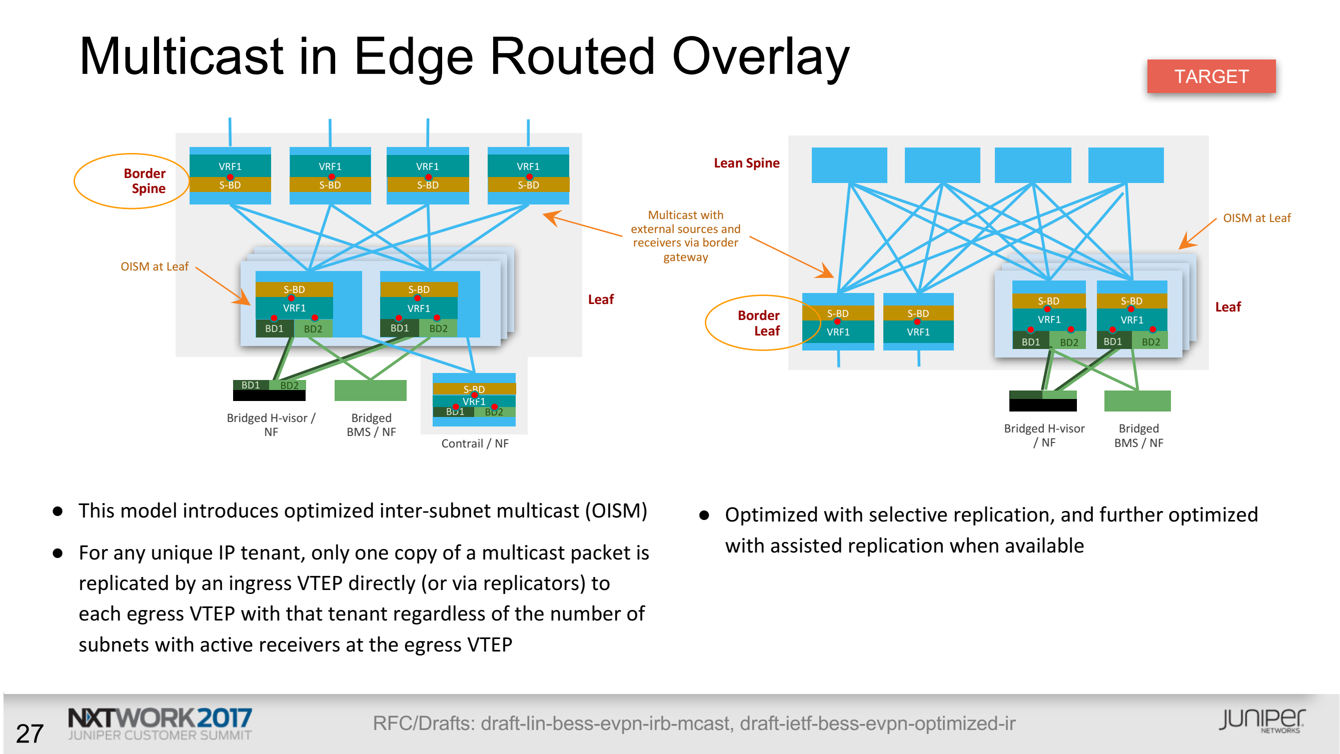
Multicast in Centrally Routed Overlay

Border Spine
VRF1
BD1
BD2
Multicast routing at central gateways with PIM DR election
Leaf
Bridged H-visor / NF
Bridged BMS / NF
Lean Spine
Multicast routing at central gateways with PIM DR election
Border Leaf
Leaf

- Classical model with PIM DR election at central gateway. Unique addresses are required for at gateways for PIM protocol signaling
- Inter-subnet replication hairpins at a central gateway
- Multicast replicated by gateway into each subnet with receivers
- Optimized with selective replication, and further optimized with assisted replication when available

26 NXTWORK2017 JUNIPER CUSTOMER SUMMIT RFC/Drafts: draft-sajassi-bess-evpn-igmp-mld-proxy, draft-ietf-bess-evpn-optimized-ir JUNIPER NETWORKS

- This is an application of multicast to the Centrally Routed service type
- It is based on the classical model with PIM DR election at the central gateways.
- Additional unique secondary addresses are required for PIM protocol signaling between gateways.
- Inter-subnet replication hairpins at whichever central gateway is the elected PIM DR.
- This option can also be optimized with selective replication, and further optimized with assisted replication when available



Multicast in Edge Routed Overlay

TARGET

- This model introduces optimized inter-subnet multicast (OISM)
- For any unique IP tenant, only one copy of a multicast packet is replicated by an ingress VTEP directly (or via replicators) to each egress VTEP with that tenant regardless of the number of subnets with active receivers at the egress VTEP

- Optimized with selective replication, and further optimized with assisted replication when available

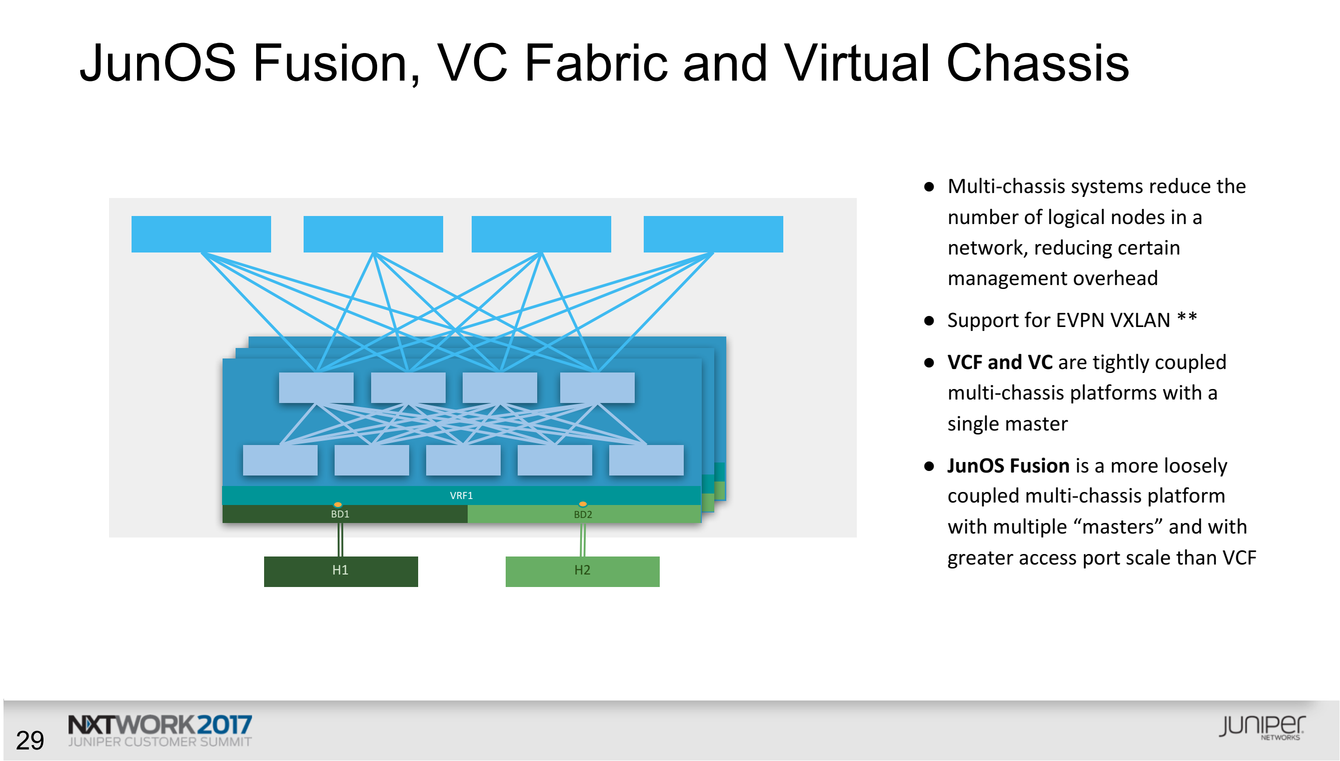
RFC/Drafts: draft-lin-bess-evpn-irb-mcast, draft-ietf-bess-evpn-optimized-ir

- Finally -- this here is an application of multicast to the Edge Routed service type
- In this edge routed model, multicast traffic between hosts connected to the same leaf VTEP does not hairpin at a central gateway
- This model introduces optimized inter-subnet multicast (OISM) which we spoke of a few slides ago
- Again, this option also can be optimized with selective replication, and further optimized with assisted replication when available

Fusion, VCF and VC

NXTWORK2017
JUNIPER CUSTOMER SUMMIT

JUNIPER NETWORKS



JunOS Fusion, VC Fabric and Virtual Chassis

- Multi-chassis systems reduce the number of logical nodes in a network, reducing certain management overhead
- Support for EVPN VXLAN **
- **VCF and VC** are tightly coupled multi-chassis platforms with a single master
- **JunOS Fusion** is a more loosely coupled multi-chassis platform with multiple “masters” and with greater access port scale than VCF

29 NXTWORK2017 JUNIPER CUSTOMER SUMMIT

JUNIPER NETWORKS

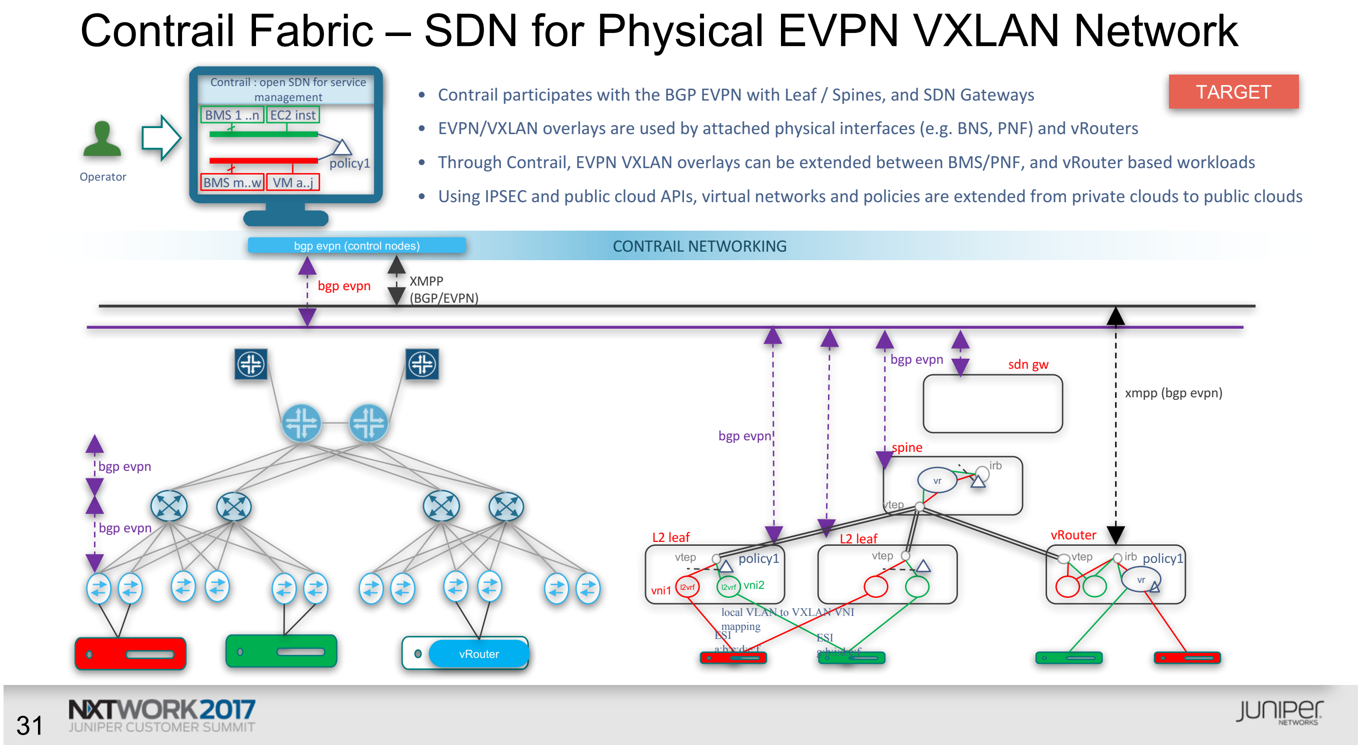
- Multiple-chassis systems like JunOS Fusion, VCF (Virtual Chassis Fabric) and VC (Virtual Chassis) can be seen as composite nodes in the larger network system topology
- The cluster of systems that forms one of these composite nodes can be seen as a single VTEP on the larger network
- These technologies can be used to reduce the number of logical nodes in a network, reducing certain network management overhead
- The key difference between these are that--
- VCF and VC are tightly coupled multiple-chassis platforms with a single member-chassis acting as the “master” RE.
- JunOS Fusion is a more loosely coupled multiple-chassis platform with multiple “master” RE
- With JunOS Fusion, north attached neighbors and routing control plane see each individual aggregation device as a unique entity, and the south attached user systems see the entire Fusion multiple-chassis system as a single entity.
- Another key property of JunOS Fusion is that the AD can be any device that can act as an 802.1BR controlling bridge – this is unlike in VCF and VC, where all devices must have silicon from the same silicon supplier (such as Broadcom). Currently both QFX10K and MX can be controlling bridges in JunOS Fusion (Aggregation Device or AD) with QFX5K, EX4300, EX3400 and EX2300 as port extenders (Satellite Devices or SD).
- The port extender concept makes the feature set of the controlling bridge available to the ports on the port extender. This means that if the controlling bridge is an MX, since the physical port is now logically on the MX, the full feature set of the MX can be applied to that port – such as high scale ACL, FIB, Hierarchical Queuing and Shaping, etc

Contrail Fabric Manager

“The Controller”

NXTWORK2017
JUNIPER CUSTOMER SUMMIT

JUNIPER NETWORKS



Contrail Fabric – SDN for Physical EVPN VXLAN Network

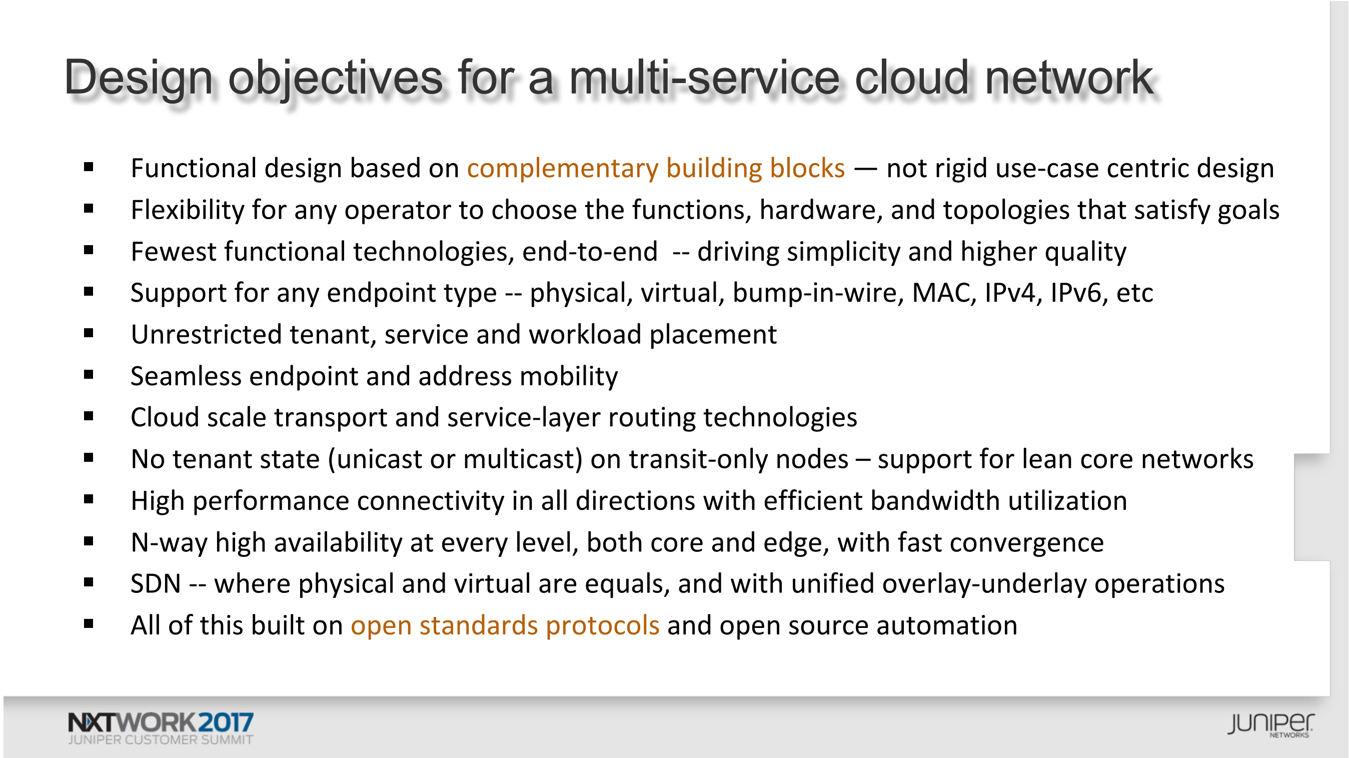
TARGET

- Contrail participates with the BGP EVPN with Leaf / Spines, and SDN Gateways
- EVPN/VXLAN overlays are used by attached physical interfaces (e.g. BNS, PNF) and vRouters
- Through Contrail, EVPN VXLAN overlays can be extended between BMS/PNF, and vRouter based workloads
- Using IPSEC and public cloud APIs, virtual networks and policies are extended from private clouds to public clouds

31 NXTWORK2017 JUNIPER CUSTOMER SUMMIT

JUNIPER NETWORKS

- Juniper’s SDN platform is an open source controller called Contrail
- Under the hood of Contrail today is IPVPN for IP overlay using MPLSoGRE tunnels, and EVPN for Ethernet overlay using VXLAN tunnels.
- In 2018, Contrail will also add support for the EVPN overlay service types that we covered in this session
- In addition, in 2018, we will enhance the Contrail platform to control physical VTEPs in addition to the virtual VTEPs it already controls today – and so creating consistent networking across physical and virtual workloads.
- This expanded functionality will go by the name Contrail Fabric
- Contrail Fabric will also bring seamless and secure multi-cloud networking -- such as between a private cloud and a public cloud.
- And finally, Contrail Fabric will also provision the underlay IP fabric and provide full underlay-overlay visibility and control



Design objectives for a multi-service cloud network

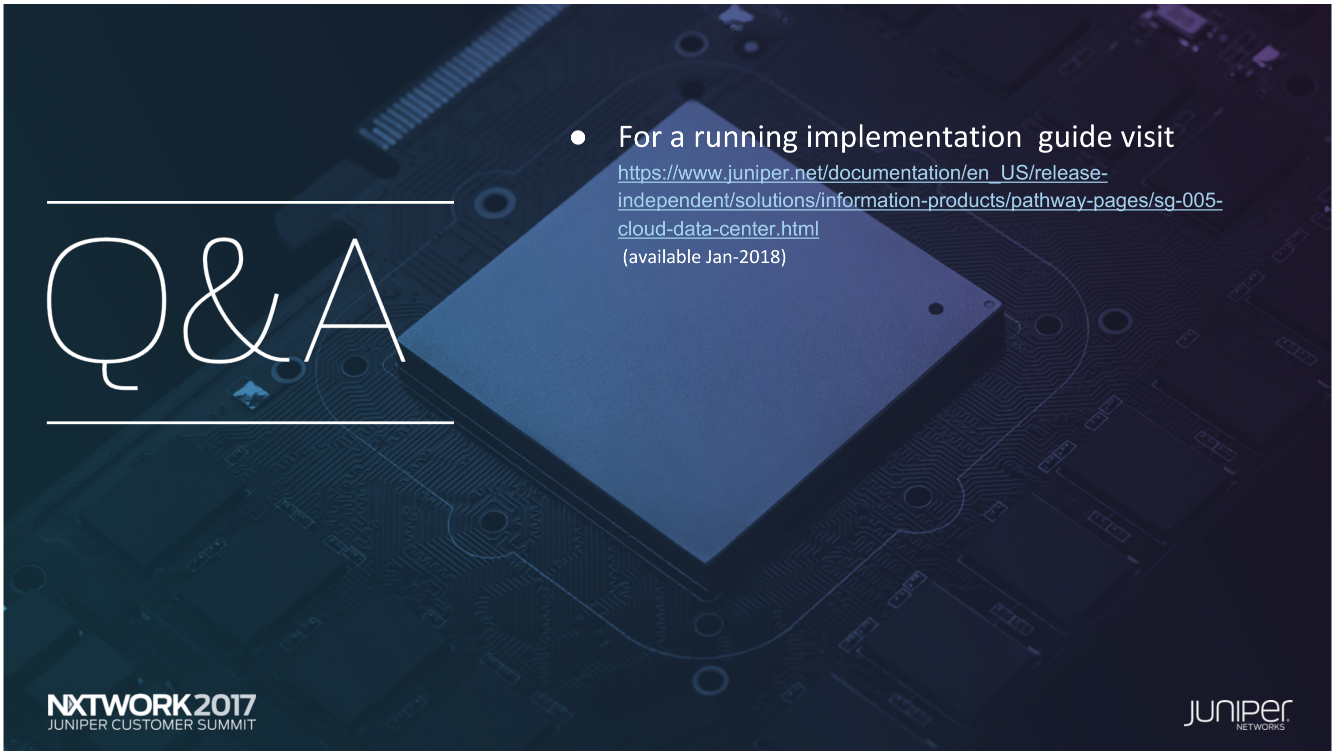
- Functional design based on complementary building blocks — not rigid use-case centric design
- Flexibility for any operator to choose the functions, hardware, and topologies that satisfy goals
- Fewest functional technologies, end-to-end -- driving simplicity and higher quality
- Support for any endpoint type -- physical, virtual, bump-in-wire, MAC, IPv4, IPv6, etc
- Unrestricted tenant, service and workload placement
- Seamless endpoint and address mobility
- Cloud scale transport and service-layer routing technologies
- No tenant state (unicast or multicast) on transit-only nodes – support for lean core networks
- High performance connectivity in all directions with efficient bandwidth utilization
- N-way high availability at every level, both core and edge, with fast convergence
- SDN -- where physical and virtual are equals, and with unified overlay-underlay operations
- All of this built on open standards protocols and open source automation

NXTWORK2017
JUNIPER CUSTOMER SUMMIT

JUNIPER NETWORKS

Let’s review what benefits we were able to satisfy with the building blocks we covered in this session

Juniper is committed to standards-based networking and continues to make significant contributions. We see standards as the backbone of truly end-to-end connectivity. In fact, Juniper was the original inventor of the key concepts and procedures used in EVPN, which we made public in 2010 under the draft named “draft-raggarwa-mac-vpn”.



Q&A

- For a running implementation guide visit https://www.juniper.net/documentation/en_US/release-independent/solutions/information-products/pathway-pages/sg-005-cloud-data-center.html (available Jan-2018)

NXTWORK2017
JUNIPER CUSTOMER SUMMIT

We'll be publishing an implementation guide on validated building blocks for cloud fabrics -- which will be updated at every release as we deliver more of the building blocks we reviewed, and beyond. The first release of this document will be available the coming January.

Also, pick up a copy of the "This Week: Data Center Deployment with EVPN/VXLAN" book. This book is a deep dive into tackling some common use cases in the data center using EVPN/VXLAN, using the Juniper portfolio.