Scaling Overlay Virtual Networks

Ivan Pepelnjak (ip@ipSpace.net)
Network Architect, ipSpace.net AG

Dimitri Stiliadis (dimitri@nuagenetworks.net)
CTO, Nuage Networks

Who is Dimitri Stiliadis

Past

- CTO of IT and security ventures
- Architect of switches and routers
- Researcher with focus in systems, networking, and security

Present

- CTO of Nuage Networks

Focus

- Large-scale SDN and cloud environments
- Distributed systems

Who is Ivan Pepelnjak (@ioshints)

Past

- Kernel programmer, network OS and web developer
- Sysadmin, database admin, network engineer, CCIE
- Trainer, course developer, curriculum architect
- Team lead, CTO, business owner

Present

- Network architect, consultant, blogger, webinar and book author
- Teaching the art of Scalable Web Application Design

Focus

- Large-scale data centers, clouds and network virtualization
- Scalable application design
- Core IP routing/MPLS, IPv6, VPN

CISCO CCIE EMERITUS

vmware vEXPERT

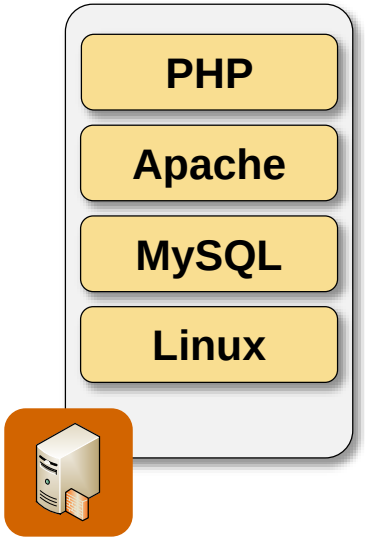
Overlay Virtual Networking Scalability Challenges

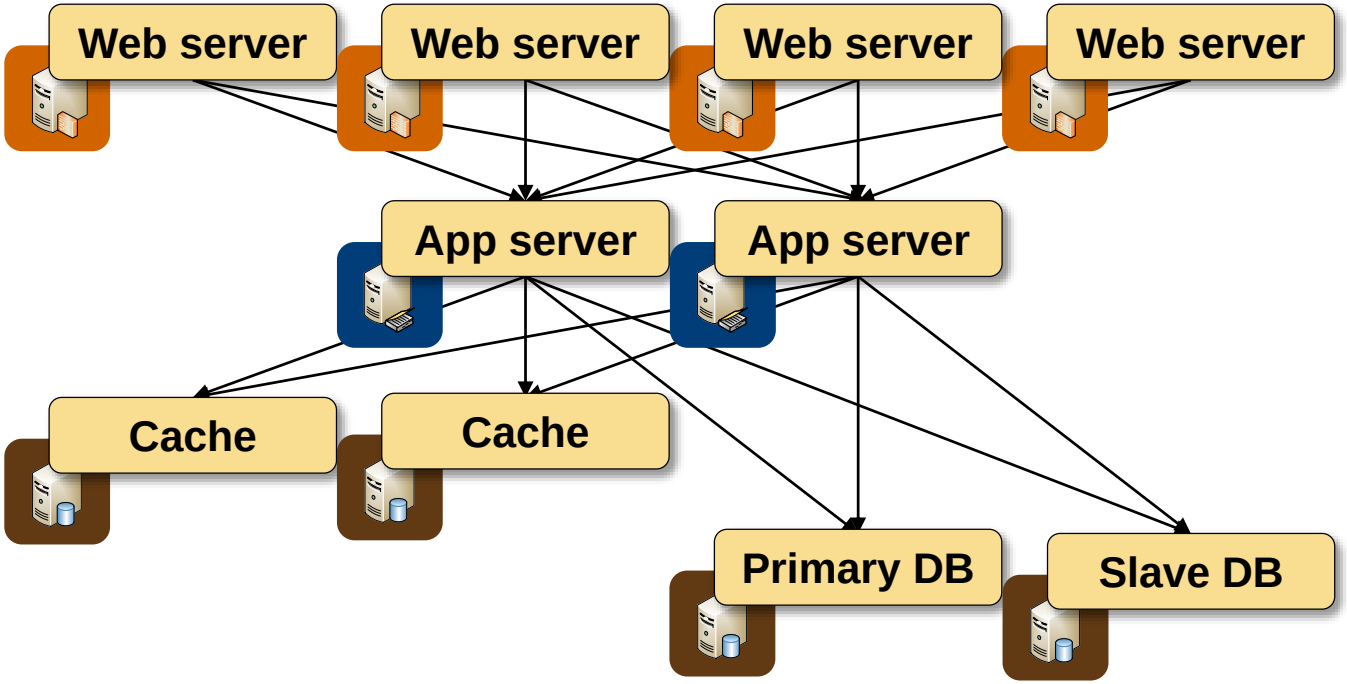
- Fully distributed data plane
- Scale-out control plane
- Availability zones
- Hardware gateways
- Large-scale microsegmentation
- Scaling stateful services
- Service chaining

Distributed Data Plane

What Network Services Will Your Cloud Offer?

What Are Your Customers Looking For?





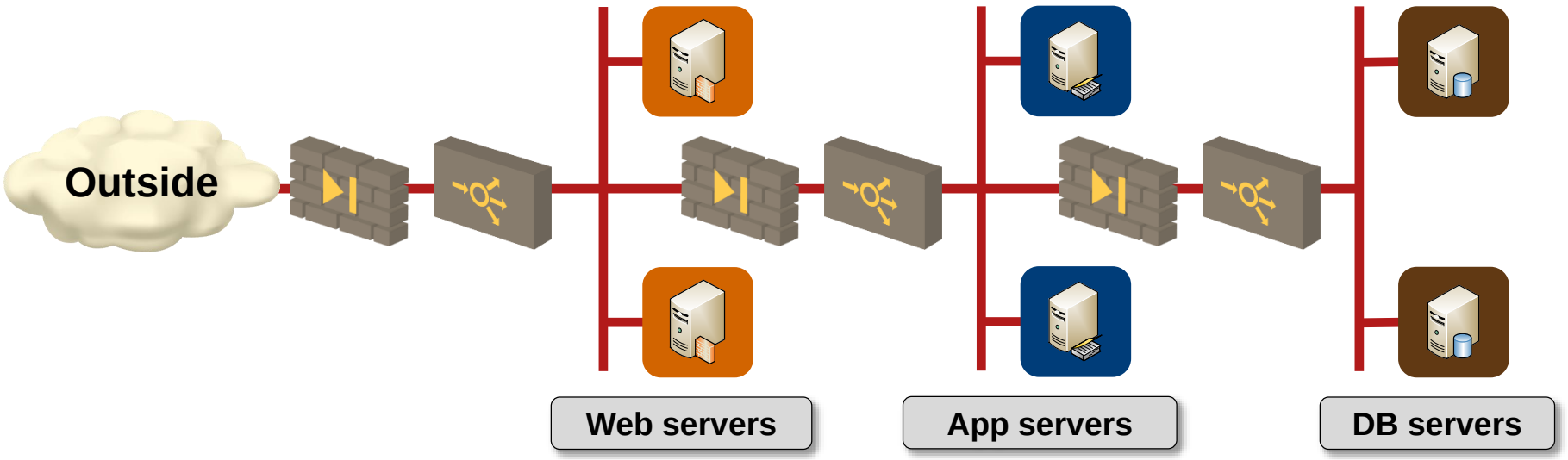
Single VM (LAMP stack)

- Typical SMB deployment
- Simple web hosting

Multi-layer application architecture

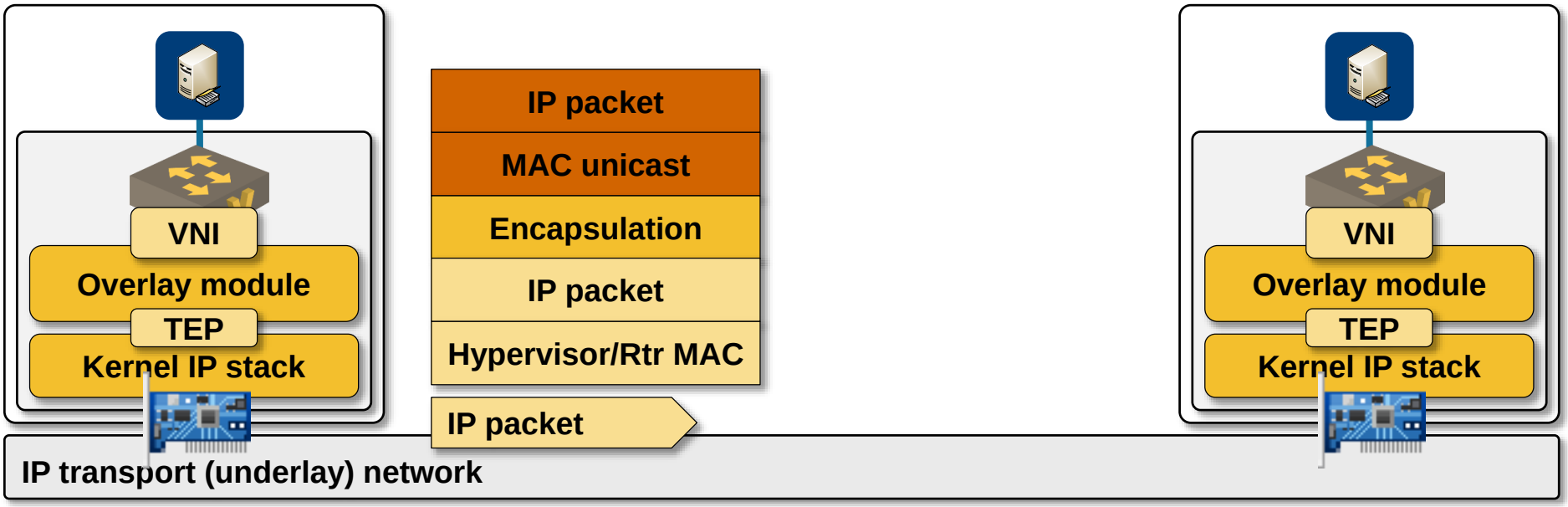
- Multiple security zones
- Load balancing and firewalling

What Complex Applications Need



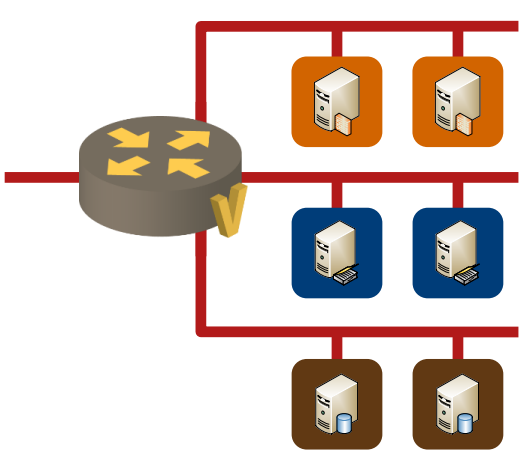
- Multiple logical segments
- IP (sometimes MAC) connectivity within a segment
- Routing, load balancing and/or firewalling between segments
- Baseline firewalling within a segment
- Connectivity to the outside world

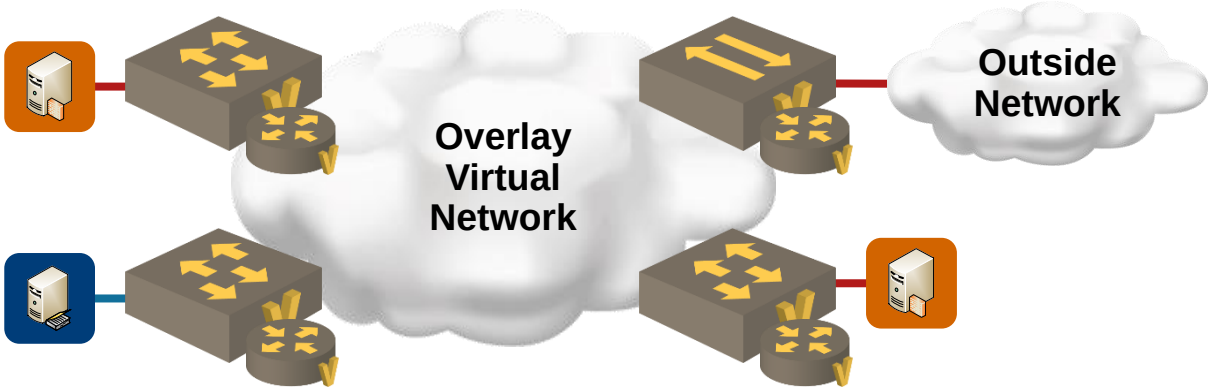
Distributed Layer-2 Forwarding



- All overlay virtual networking solutions use distributed L2 forwarding
- Scalability is limited by the control plane (distribution of VM MAC-to-VTEP IP mappings)

Distributed Layer-3 Forwarding





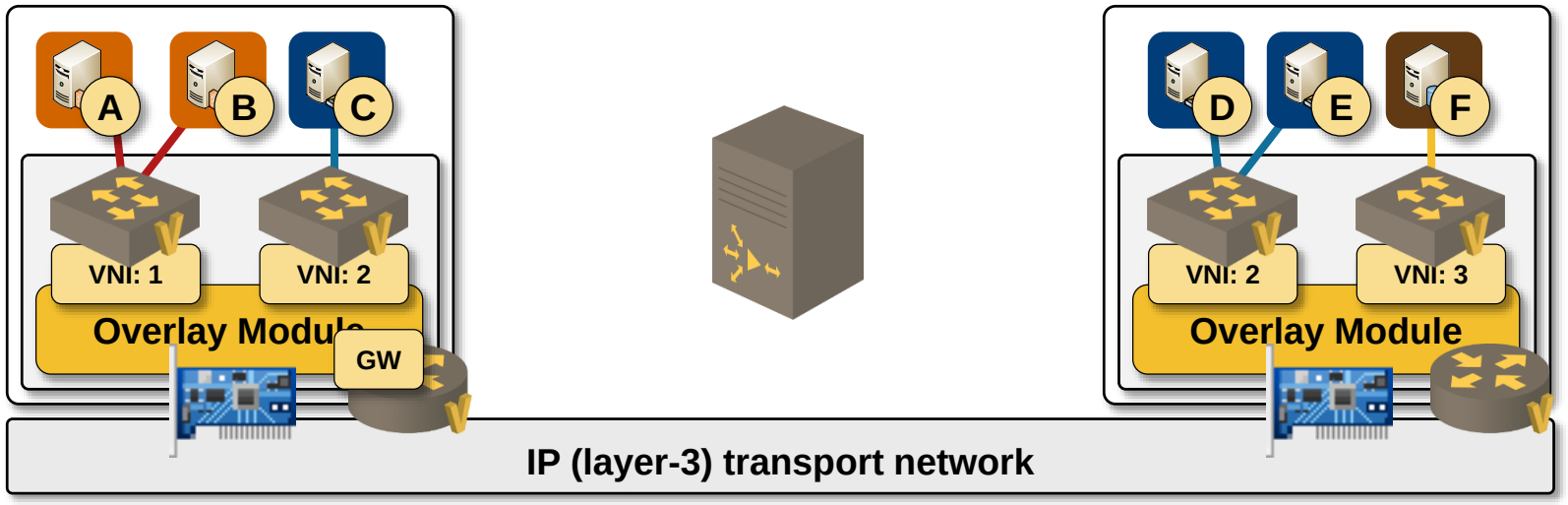
Centralized (sometimes VM-based) inter-subnet forwarding doesn't scale

- Virtual router (L3 agent) becomes a chokepoint
- VM-based forwarding has limited performance
- Avoid this architecture for east-west traffic forwarding

Use architecture with distributed layer-3 forwarding

- Prefer dedicated in-kernel implementation over Linux Kernel TCP/IP stack with namespaces or VM-based implementations
- Sample products: Juniper Contrail, Microsoft Hyper-V, Nuage VSP, VMware NSX

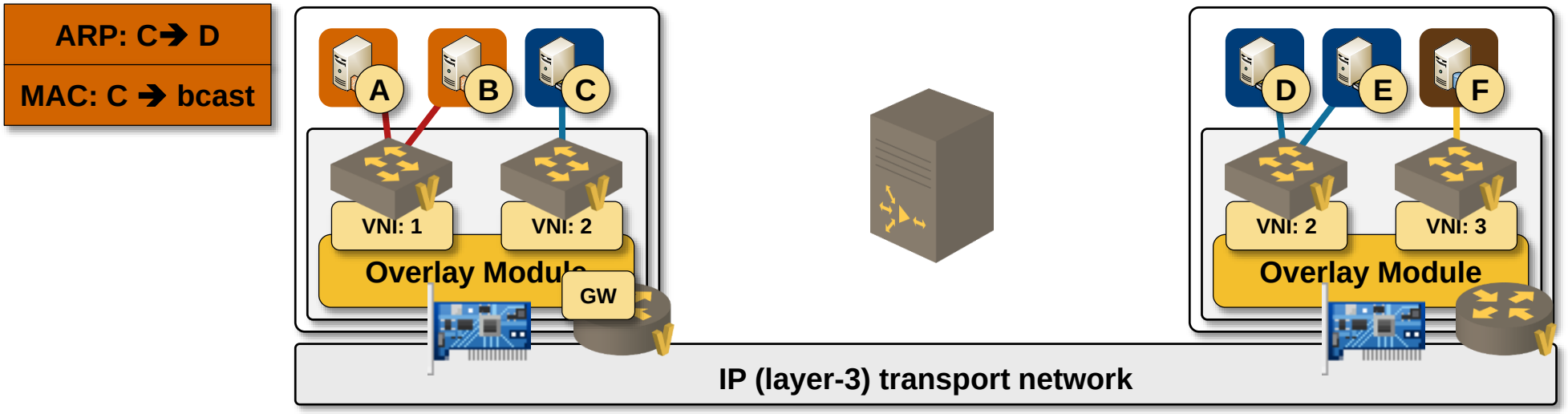
Distributed ARP Caching



Some overlay virtual networking solutions implement combined L2+L3 forwarding model

- Intra-subnet ARP caching significantly reduces overlay broadcast traffic

Distributed ARP Caching

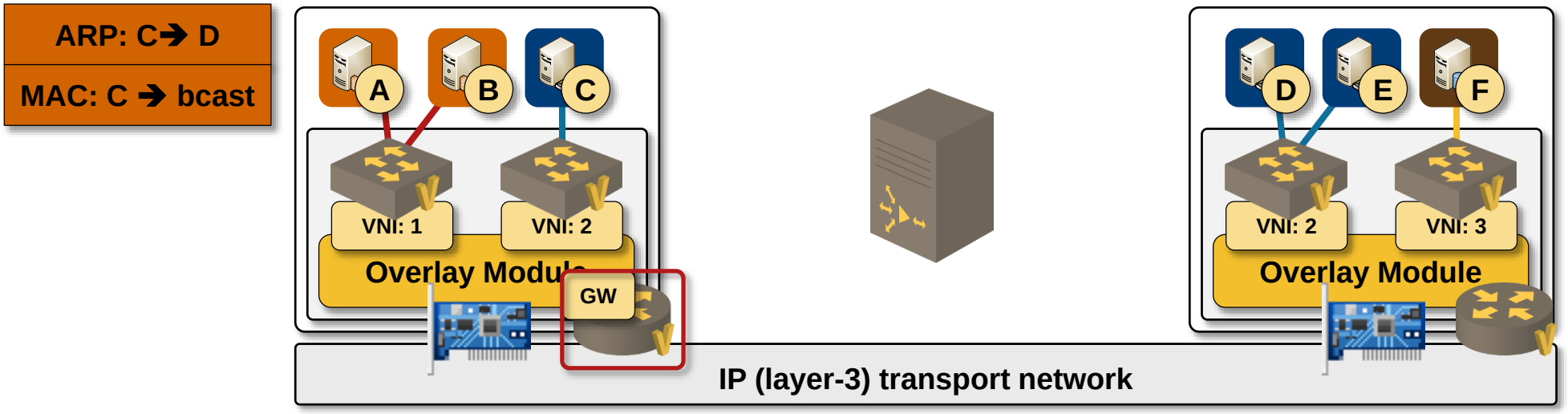


Some overlay virtual networking solutions implement combined L2+L3 forwarding model

- Intra-subnet ARP caching significantly reduces overlay broadcast traffic

Example: ARP request C ➔ D

Distributed ARP Caching



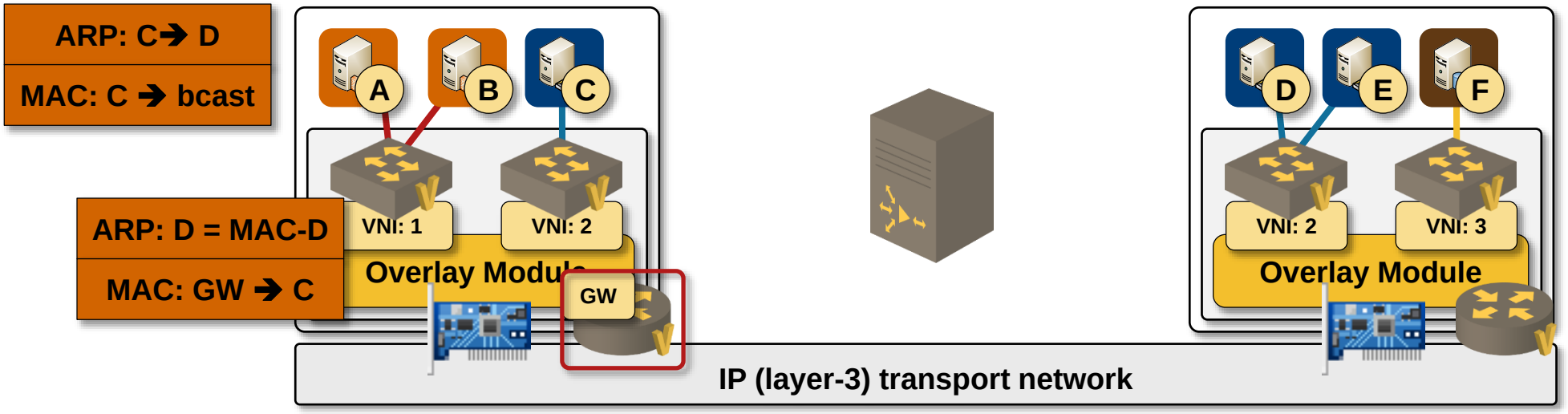
Some overlay virtual networking solutions implement combined L2+L3 forwarding model

- Intra-subnet ARP caching significantly reduces overlay broadcast traffic

Example: ARP request C ➔ D

- Intercepted by local L3 forwarding module

Distributed ARP Caching



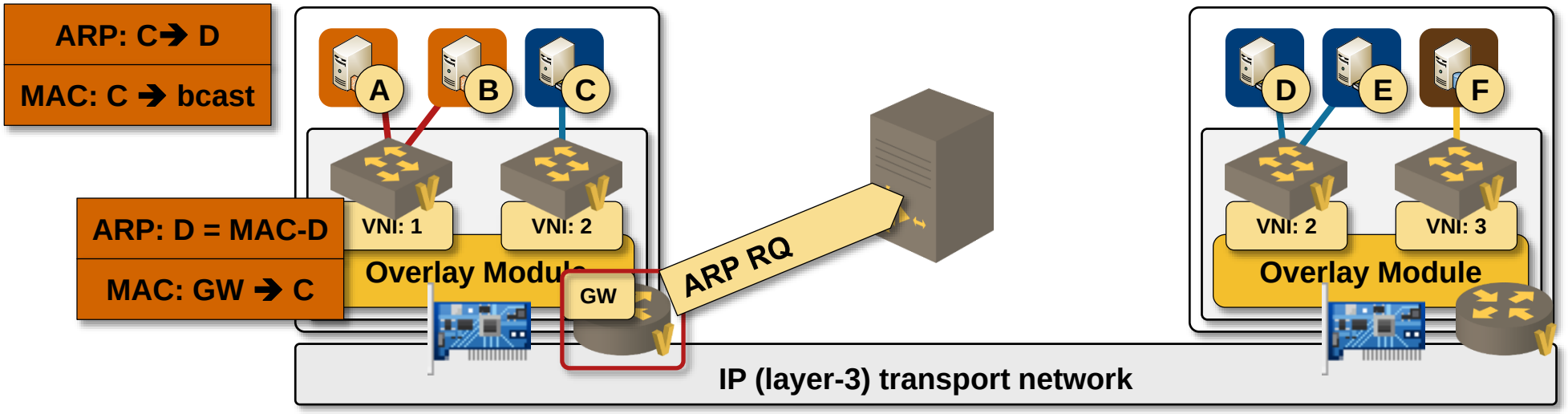
Some overlay virtual networking solutions implement combined L2+L3 forwarding model

- Intra-subnet ARP caching significantly reduces overlay broadcast traffic

Example: ARP request C ➔ D

- Intercepted by local L3 forwarding module
- Replied from local ARP cache

Distributed ARP Caching



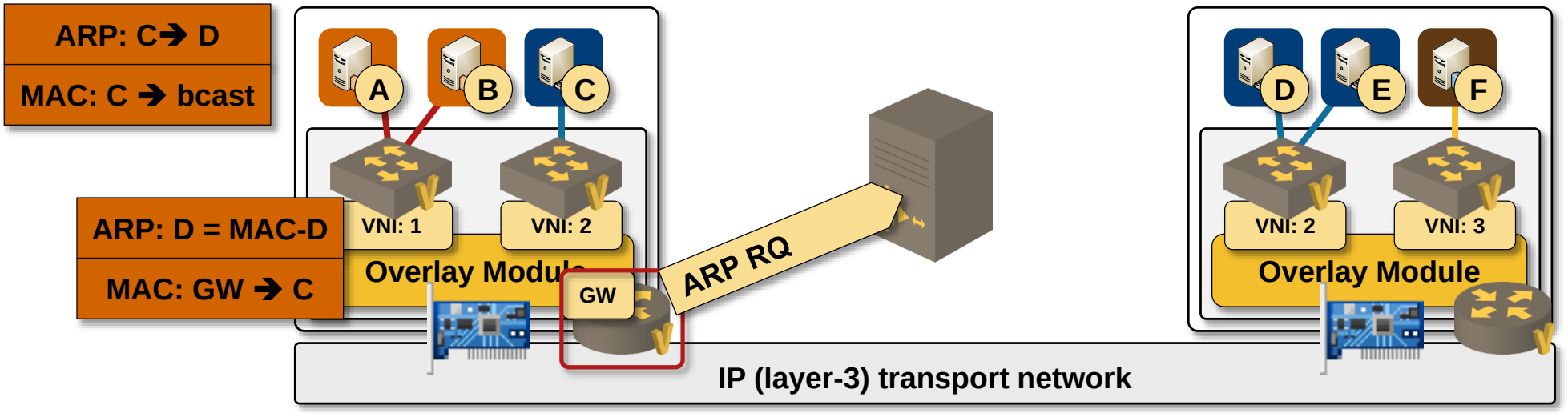
Some overlay virtual networking solutions implement combined L2+L3 forwarding model

- Intra-subnet ARP caching significantly reduces overlay broadcast traffic

Example: ARP request C ➔ D

- Intercepted by local L3 forwarding module
- Replied from local ARP cache
- Controller is contacted on ARP cache miss

Distributed ARP Caching



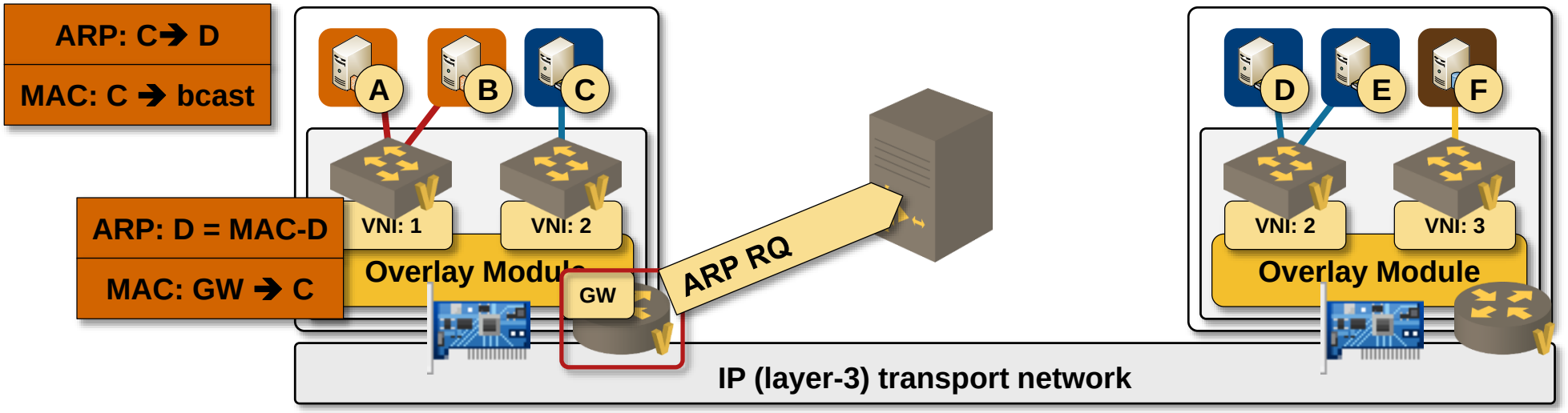
Some overlay virtual networking solutions implement combined L2+L3 forwarding model

- Intra-subnet ARP caching significantly reduces overlay broadcast traffic

Example: ARP request C ➔ D

- Intercepted by local L3 forwarding module
- Replied from local ARP cache
- Controller is contacted on ARP cache miss
- Controller can reply with authoritative information or flood ARP request

Distributed ARP Caching



Some overlay virtual networking solutions implement combined L2+L3 forwarding model

- Intra-subnet ARP caching significantly reduces overlay broadcast traffic

Example: ARP request C ➔ D

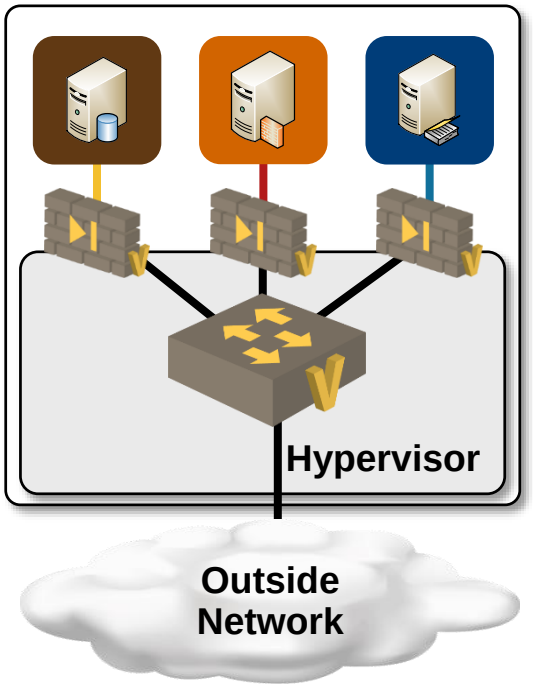
- Intercepted by local L3 forwarding module
- Replied from local ARP cache
- Controller is contacted on ARP cache miss
- Controller can reply with authoritative information or flood ARP request

Available in VMware NSX for vSphere, Nuage Networks VSP

Distributed Network Services

Scaling network services

- Scale-out load balancing is mission impossible (shared state tied to outside IP address)
- Scale-out firewalls are common (state tied to a single VM)
- Scale-out NAT is an interesting challenge



Implement traffic filters with VM NIC firewalls

- Stateful firewalls or reflexive ACLs
- Reflexive ACLs might be good enough for well-designed applications
- VM-based solutions severely limit performance
  ➔ use in-kernel filters
- Sample solutions: Nuage VSP, VMware NSX, OpenStack/CloudStack on KVM
- ACL-only solutions: Microsoft Hyper-V, VMware vSphere, Cisco Nexus 1000V

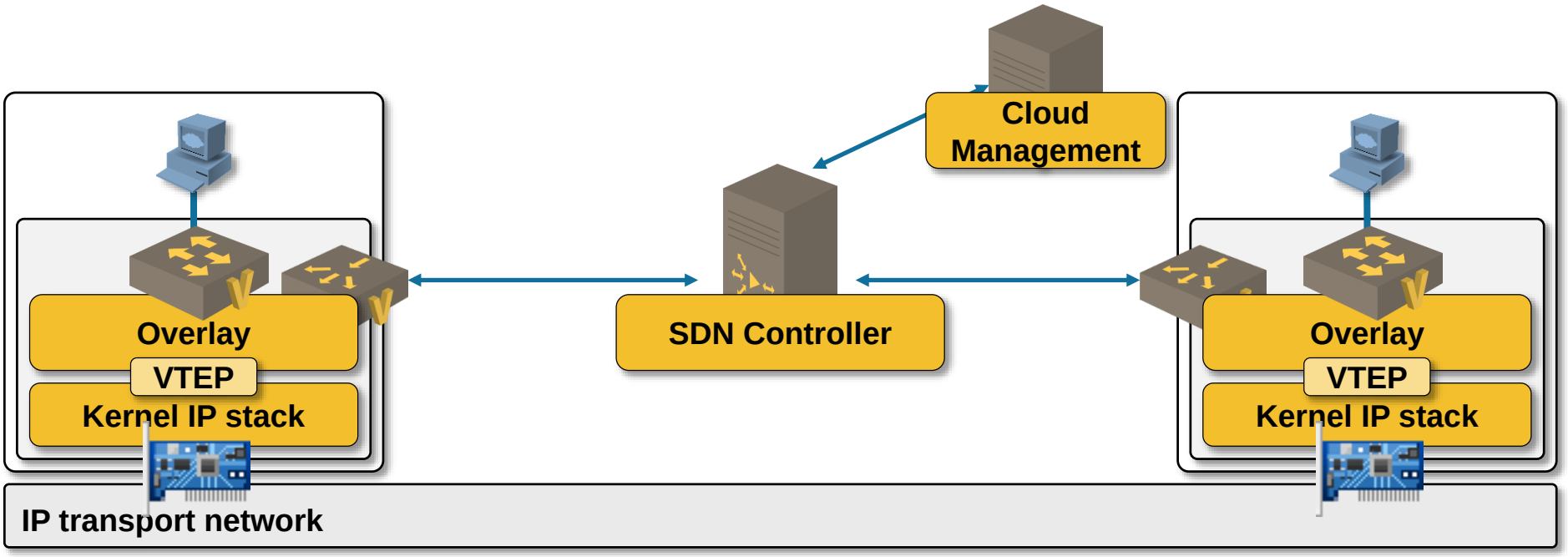
Conclusions

Requirements for scalable data plane

- Distributed L3 forwarding
- Local ARP handling (ARP caching or pure L3 solution)
- Distributed security groups implemented in hypervisors

Scale-Out Control Plane

Controller-Based Overlay Virtual Networks



Crucial overlay virtual network challenge: VM-MAC-to-VTEP-IP mappings

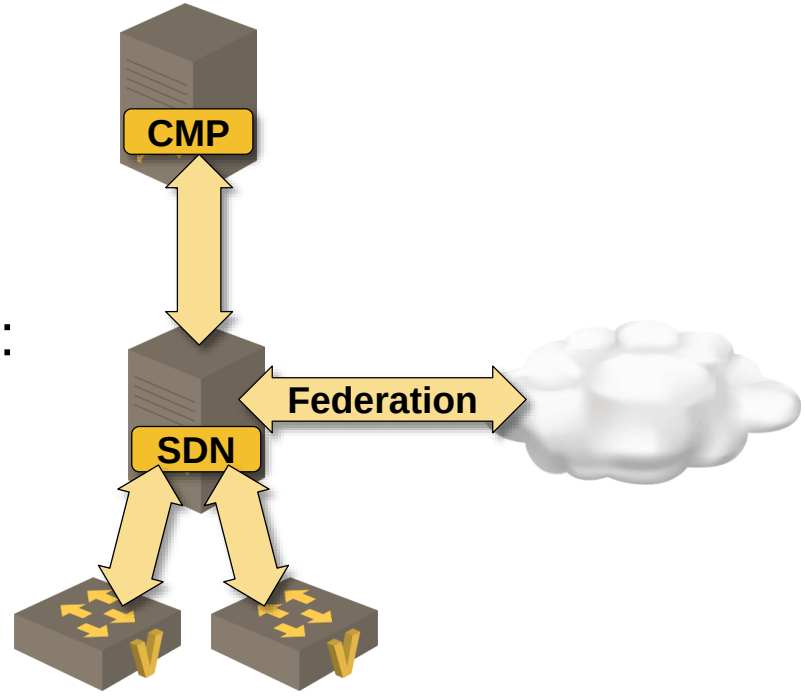
- Initial implementations used IP multicast and Ethernet-like learning
- Modern solutions use network controllers in combination with orchestration systems

Sample solutions: Cisco Nexus 1000V, Juniper Contrail, Nuage VSP, VMware NSX

The Need for SDN Controller

Some overlay networking solutions lack SDN controller element

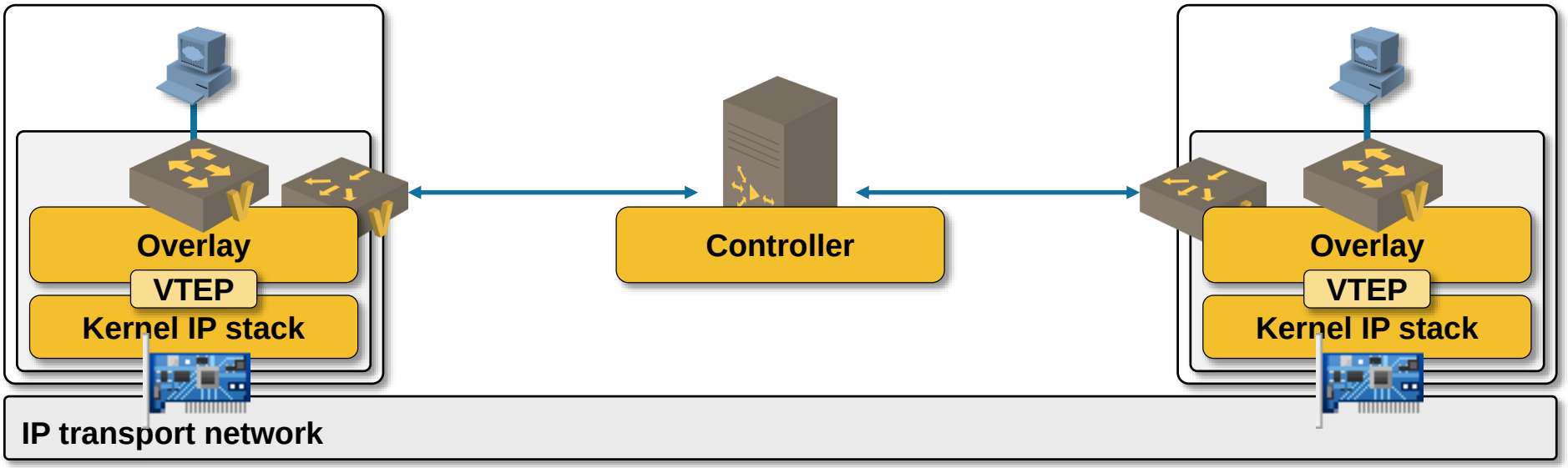
- Cloud management platform programs virtual switches directly
- Hard to integrate with the physical network: static routes/MAC learning or VM-based solutions



SDN controller enables inter-cloud federation

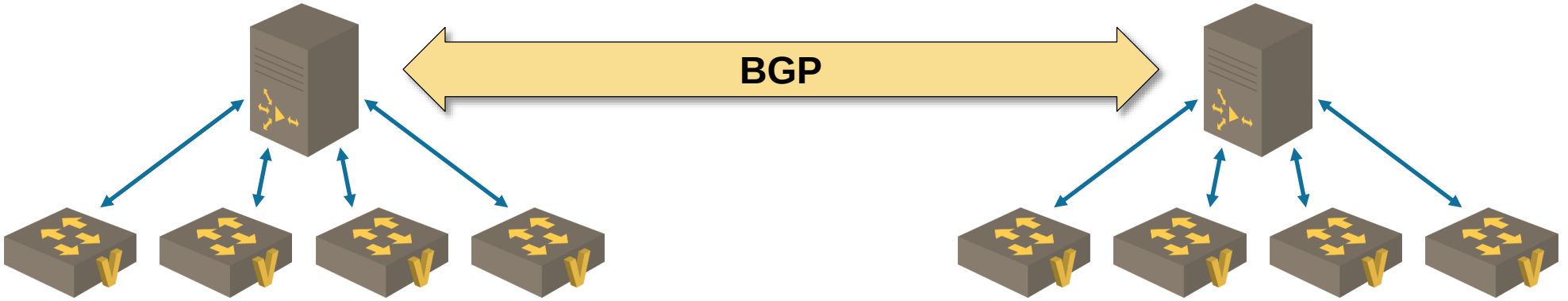
- Reachability data exchanged between controllers
- Most SDN controllers use BGP for easy integration with existing hardware

Scaling Controller Implementations



- Network controller becomes the scalability bottleneck
- Control-plane-only controllers scale much better than controllers participating in data plane (hint: use CMP to get MAC and IP address information)
- Every controller implementation eventually hits its limits
  ➔ scale-out is the only answer

Scale-Out Controller Architecture



Scale-out architecture is the only viable way forward

- Requirement: Synchronization of policy and reachability information between controllers

Typical solution: multi-protocol BGP (MP-BGP)

- L3VPN for IP routing (sometimes using host routes for VM IP addresses)
- EVPN for layer-2 forwarding
- Easy integration with existing hardware gateways

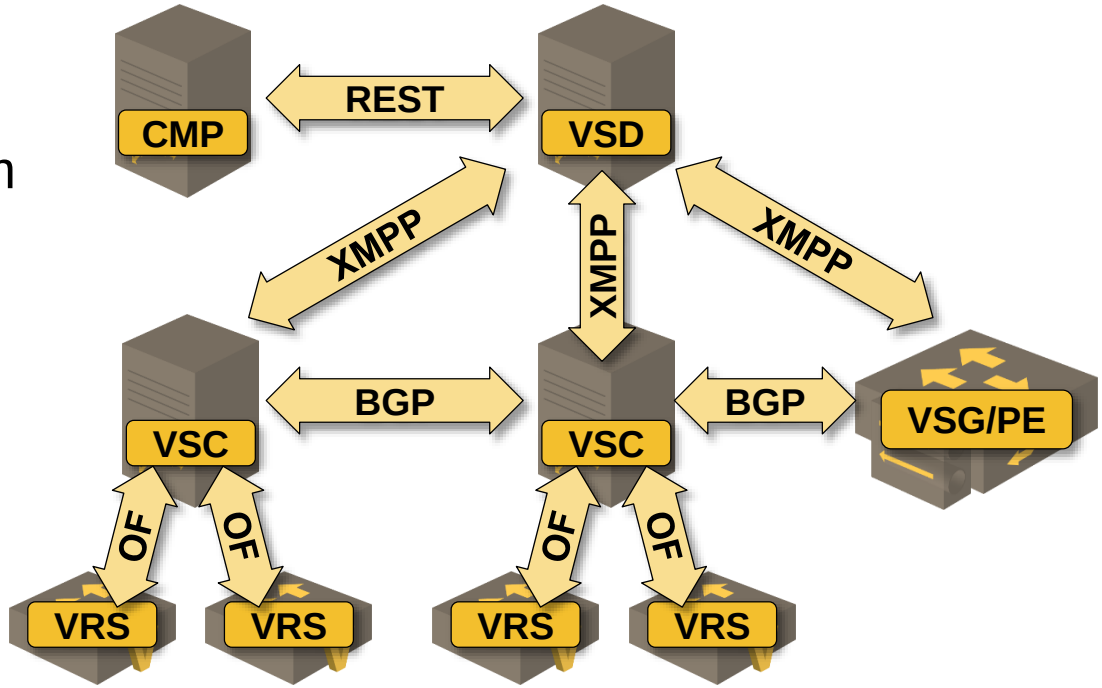
Additional benefits:

- Clean failure domain separation (availability zones)
- Adjustable size of failure domains to meet scalability and convergence requirements

Nuage VSP Architecture

Terminology:

- VSP: Virtual Services Platform
- CMP: Cloud Management Platform
- VSD: Virtual Services Directory
- VSC: Virtual Services Controller
- VRS: Virtual Routing & Switching



Plane of operation

- VSD: Management/Policy
- VSC: Control plane
- VRS: Data plane

Scale-out architecture

- Single VSD per CMP
- Multiple VSC per VSD (scale-out within CMP)
- VSC confederation via MP-BGP (scale-out across CMP)

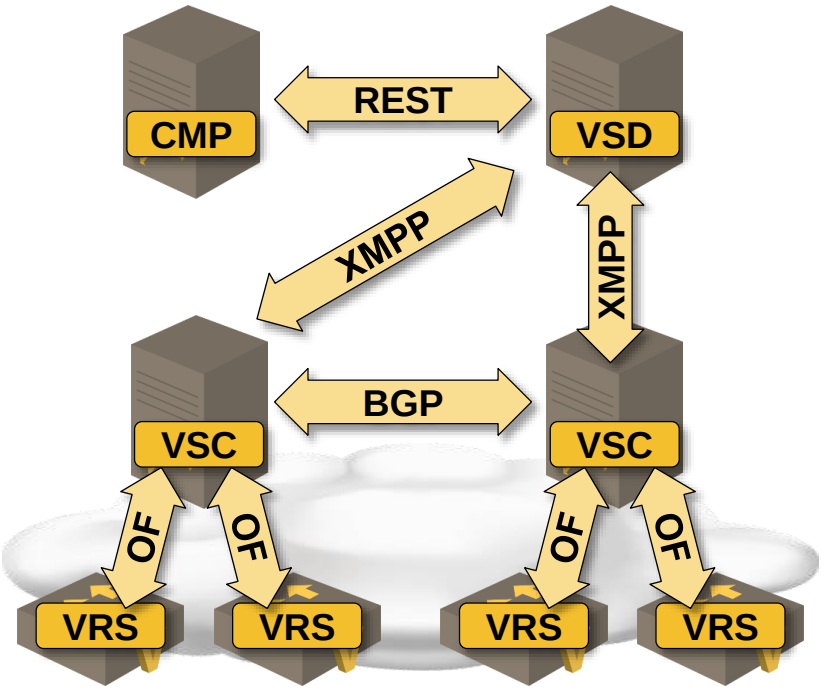
Availability Zones

Terminology: Failure Domain

Failure Domain: area impacted when a key device or service experiences problems

Sample failure domains

- VLAN (broadcast storms)
- OSPF area (LSA flooding)
- Controller-based network (controller failure)
- Cloud instance (cloud management system failure)



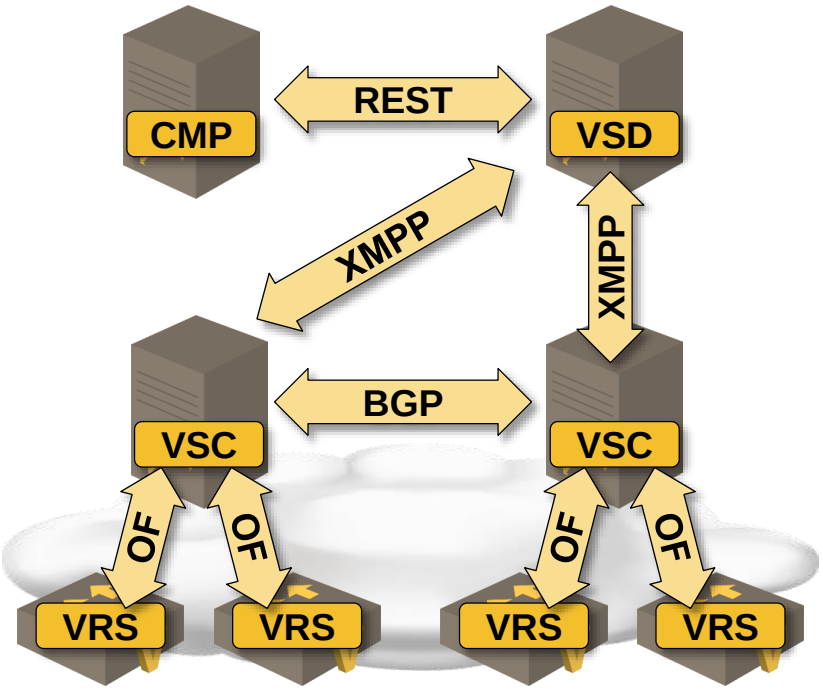
Terminology: Availability Zone

Regions: cloud instances with separate API endpoints

- Separate instances of cloud management systems

Availability zone: logical group that provides a form of physical isolation and redundancy from other availability zones (OpenStack)

- Common cloud management
- Isolated compute/storage/networking failure domains
- Each availability zone SHOULD have a different network services controller



What Happens If…

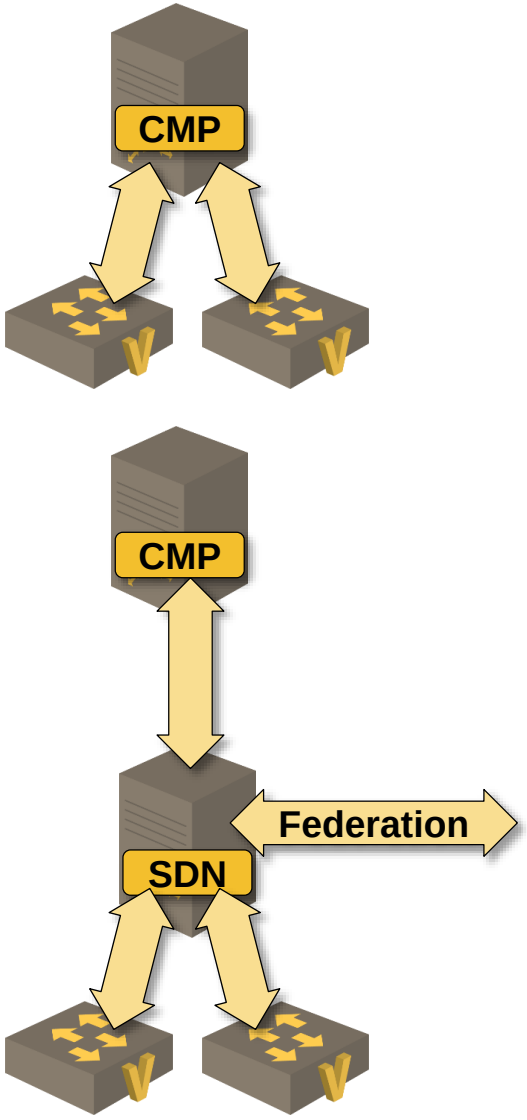
Cloud management platform fails?

- No moves, adds or changes
- Overlay virtual networking topology is frozen
- High-availability clusters cannot recover

SDN controller fails?

- Controllers involved in data plane (MAC learning or ARP replies) ➔ total failure
- Control-plane controllers ➔ loss of reachability information
- Controllers without external control plane ➔ no visibility, no topology change

Each availability zone SHOULD have an independent SDN controller



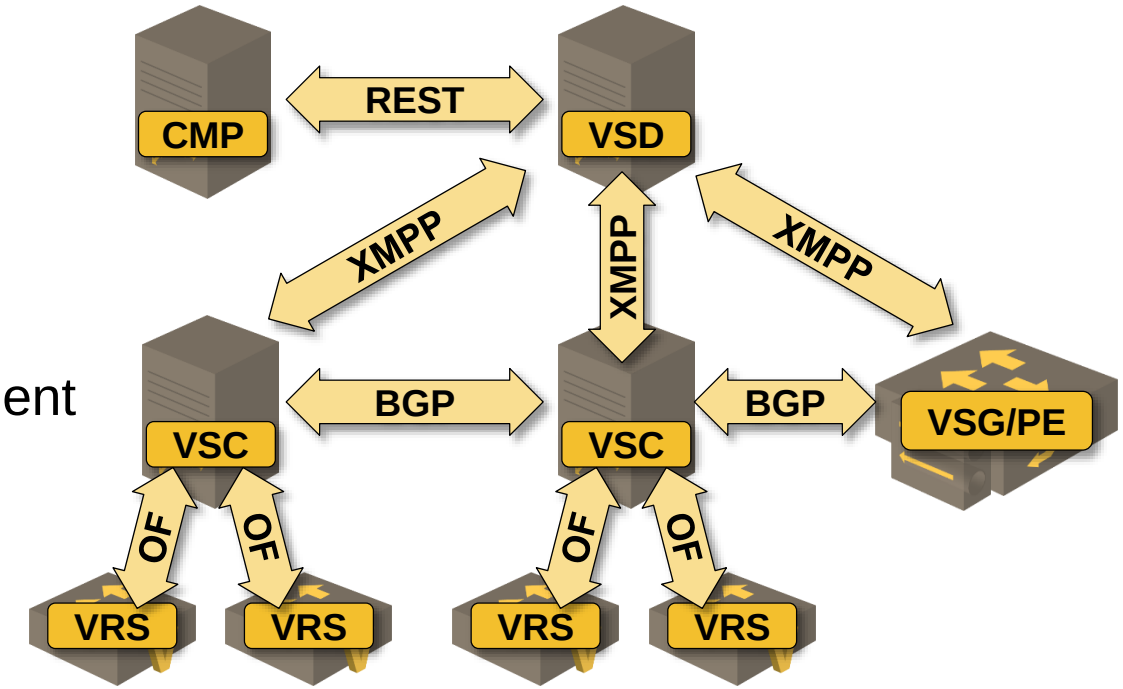
Availability Zones with Nuage VSP

Controller/orchestration infrastructure

- Single CMP/VSD per region
- VSD works on policy plane ➔ VSD failure is similar to CMP failure
- VSC per availability zone ➔ VSC failure does not spread across zones
- BGP information exchange through a set of route reflectors ➔ use BGP security mechanisms to protect availability zones
- Pair of VSGs per availability zone (when needed)

Underlying infrastructure

- Each availability zone = independent L3 forwarding domain



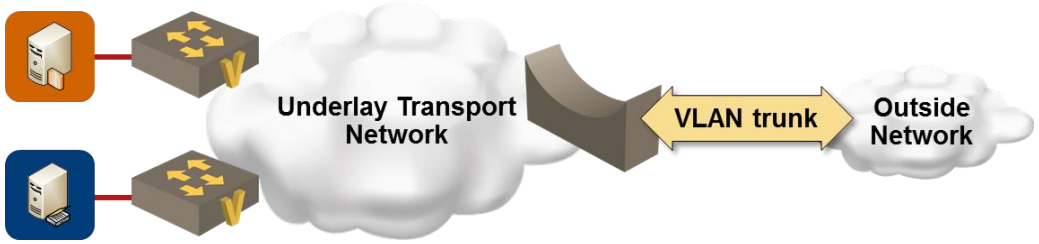
Hardware Gateways

Interaction with Physical World

VMs within an overlay virtual network must interact with the physical world

L2 gateways (VNI-to-VLAN)

- P2V migrations
- Integration with legacy equipment

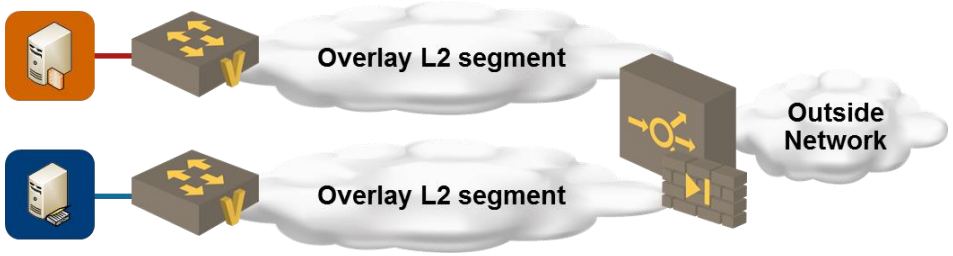


L3 gateways

- Multiple VNIs routed to a VLAN
- Simple P2V or WAN integration

Network services gateway

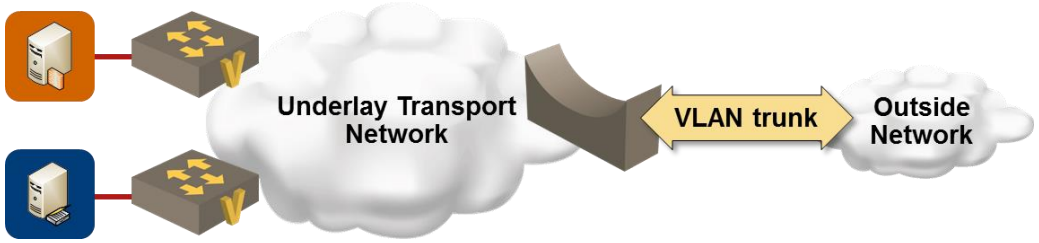
- Firewalls and load balancers



Gateway Implementation Options

Deployment format

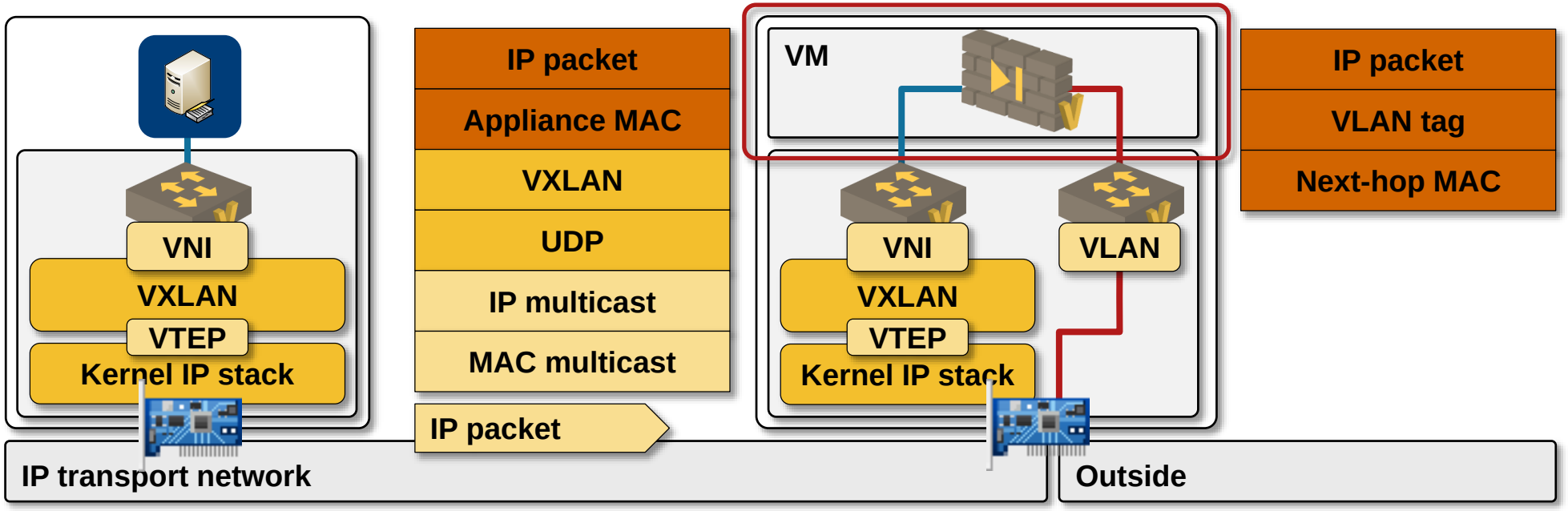
- VM-based
- Hypervisor kernel module
- Bare-metal x86 server
- Hardware VTEP



Design and deployment considerations

- Performance
- Control-plane integration with overlay fabric
- Management plane integration with overlay network controller and orchestration system
- Integration with existing network infrastructure (example: MPLS/VPN)

VM-Based Gateways



- Gateway function implemented in a VM with multiple virtual NICs
- VM performs traditional bridging/routing/network services functionality
- Use any product available in VM format (including Linux instances)
- Forwarded traffic goes through a VM ➔ performance usually limited to few Gbps

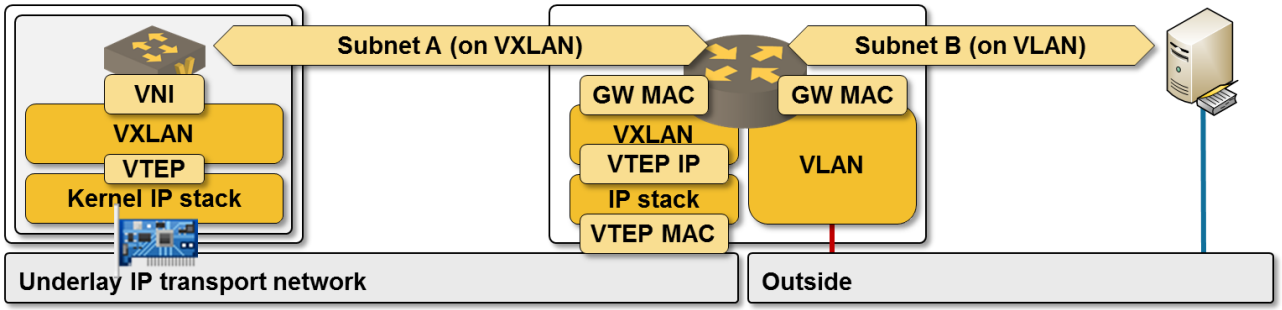
Do You Need a Hardware Gateway?

Typical gateway deployment scenarios

- Integrate overlay networks with outside world
  ➔ maximum performance = WAN link speed
- Integrate overlay networks with legacy hardware
  ➔ maximum performance = legacy hardware network I/O performance

Software gateway performance

- Few Gbps for VM-based solutions
- ~10Gbps for kernel-based and bare-metal gateways



Hardware gateways offer the performance needed in large-scale deployments

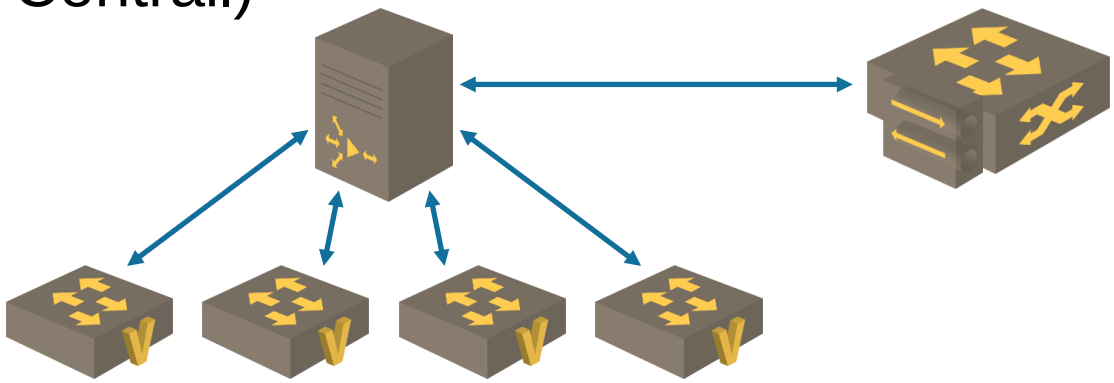
Hardware Gateway Integration Challenges

Hardware Gateway needs the following information

- Mapping between VXLAN VNI and external VLANs
- VM-MAC-to-VTEP-IP mappings
- VXLAN flooding information (IP MC address or VTEP list)

Solutions

- Do-it-yourself
- OVSDB (VMware NSX, Nuage VSP)
- EVPN (Nuage VSP, Juniper Contrail)



Gateway-to-Controller Integration with OVSDB

OVSDB

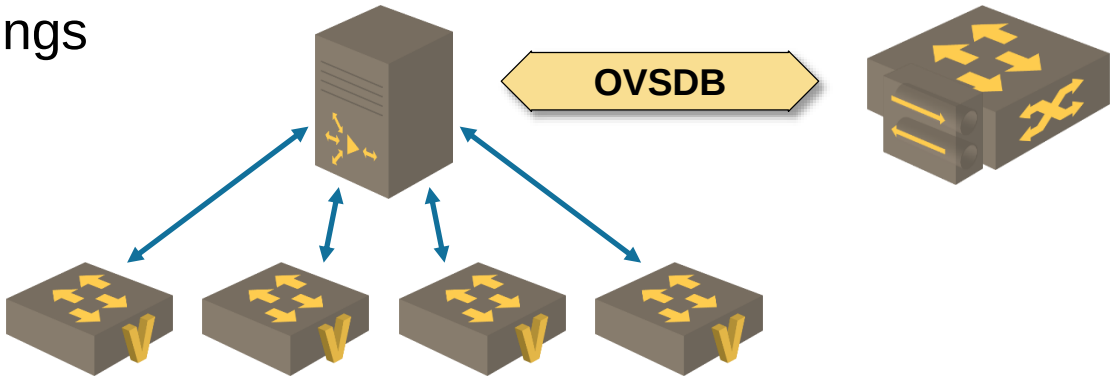
- Lightweight JSON-RPC-based database query/update protocol
- OVSDB database table schema defines the actual data

Hardware VTEP schema

- Physical switch + ports
- Logical switch + router
- Local and remote MAC mappings

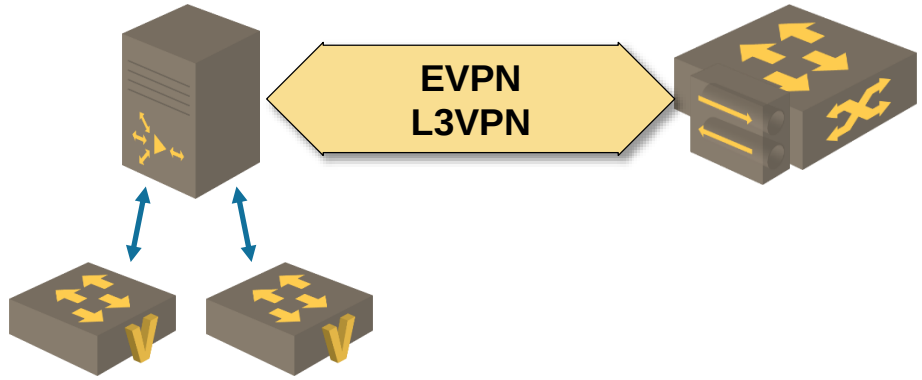
SDN controller uses OVSDB to

- Configure VXLAN-to-VLAN mappings
- Push MAC mappings to VTEP
- Receive physical MAC addresses from VTEP



Gateway-to-Controller Integration with EVPN and L3VPN

- Network virtualization controller and hardware gateway use EVPN and L3VPN to exchange forwarding data
- EVPN provides MAC-to-VTEP mappings
- L3VPN provides integrates overlay virtual networks with MPLS/VPN
- Gateway provisioning uses a different protocol (ex: NETCONF)

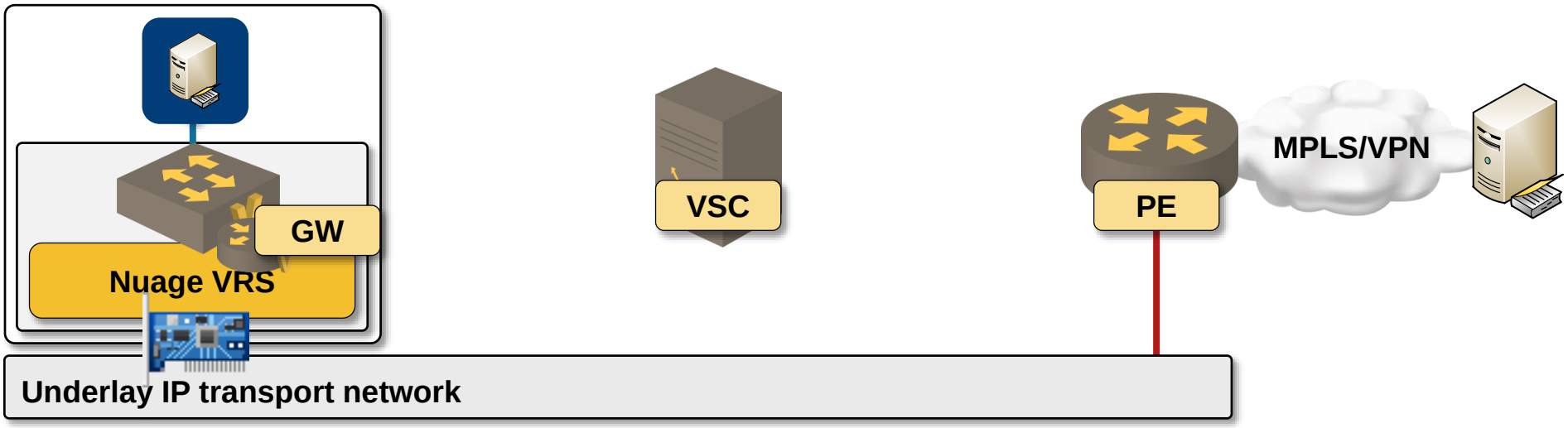


EVPN forwarding information

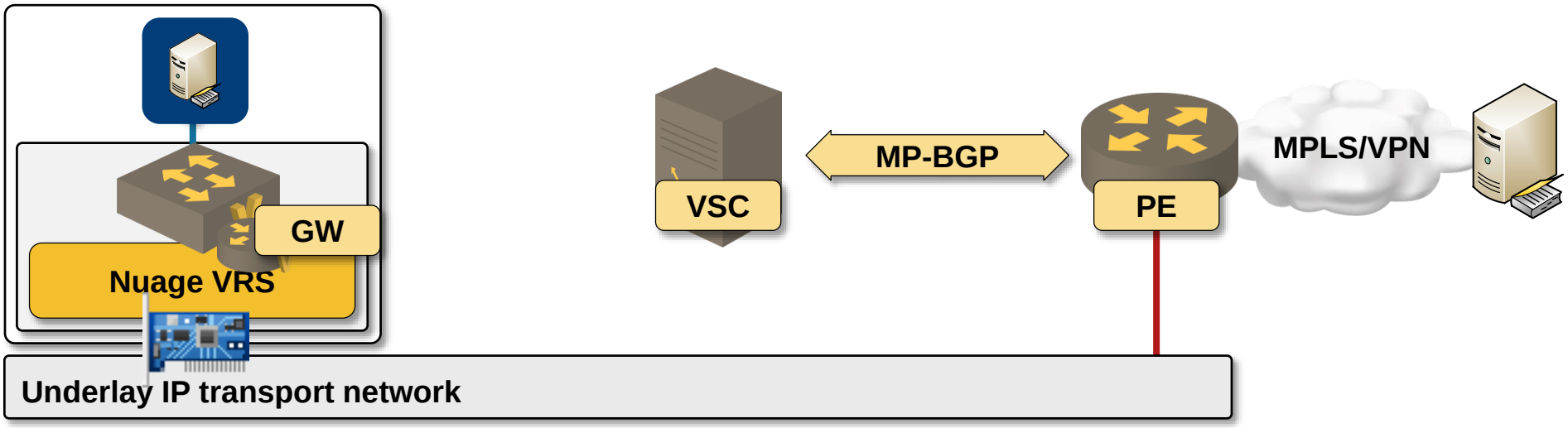
- VTEP flood list (Inclusive Multicast Ethernet Tag route)
- MAC-to-VTEP mapping (MAC/IP Address Advertisement route)
- Propagation of IP addresses enables proxy ARP functionality

MPLS/VPN integration through MP-BGP (same domain or inter-AS Option B/C)

MPLS/VPN and EVPN Integration with Nuage VSP

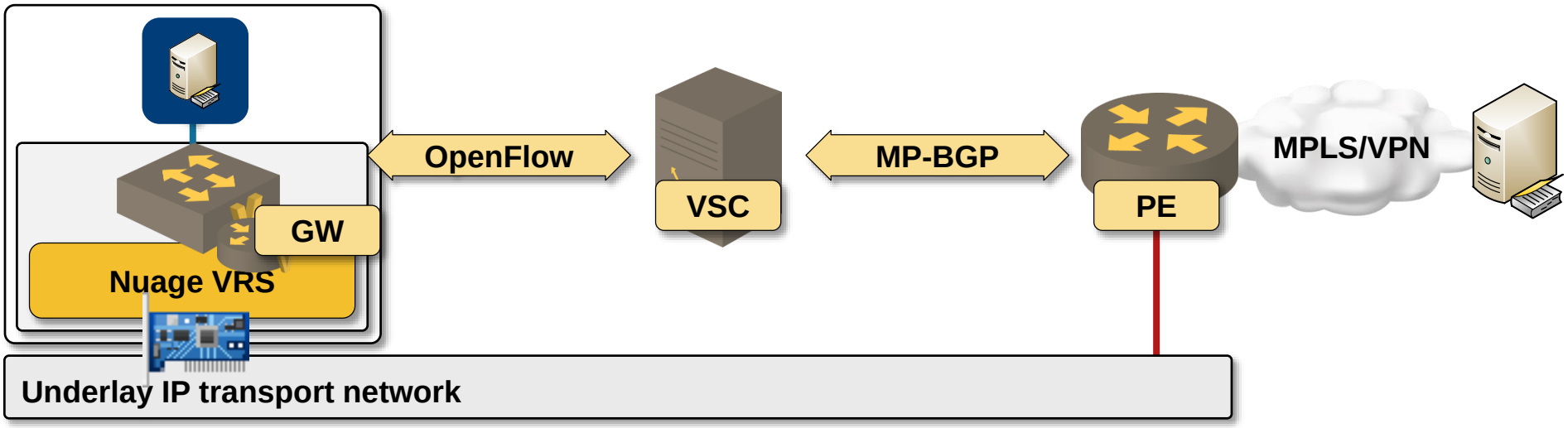


MPLS/VPN and EVPN Integration with Nuage VSP



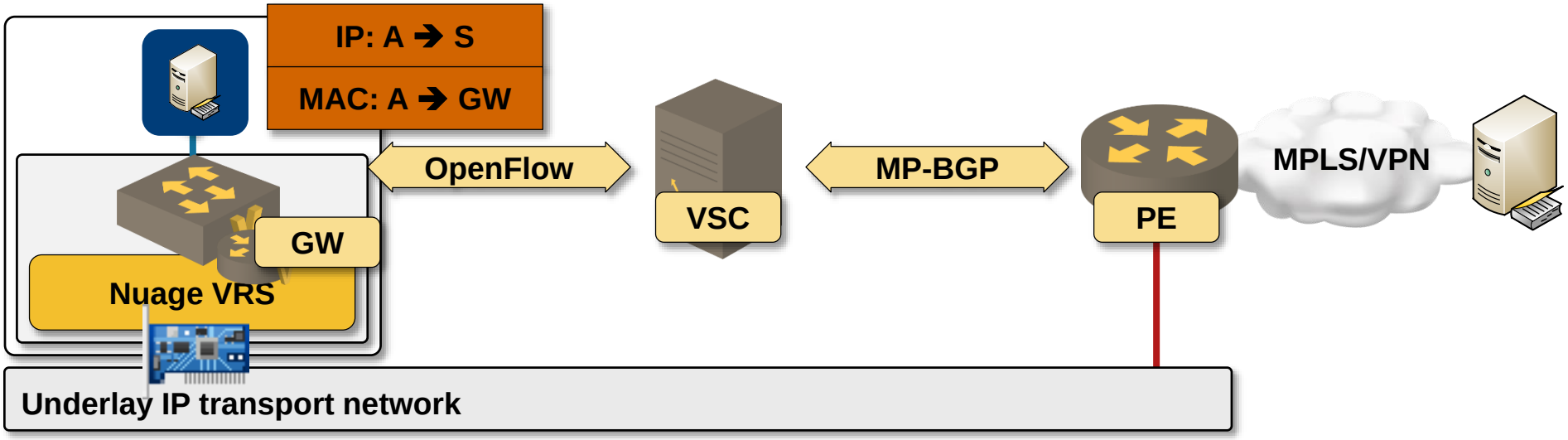
- PE-router sends VPNv4 or EVPN update to Nuage VSC

MPLS/VPN and EVPN Integration with Nuage VSP



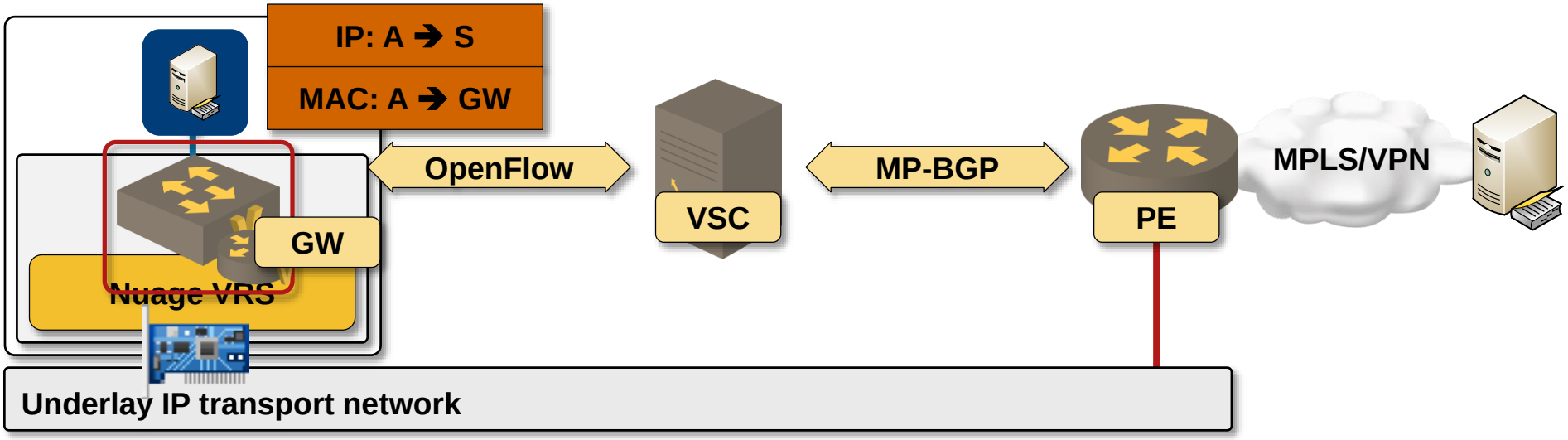
- PE-router sends VPNv4 or EVPN update to Nuage VSC
- VSC installs forwarding entries with BGP next hop + label in VRS

MPLS/VPN and EVPN Integration with Nuage VSP



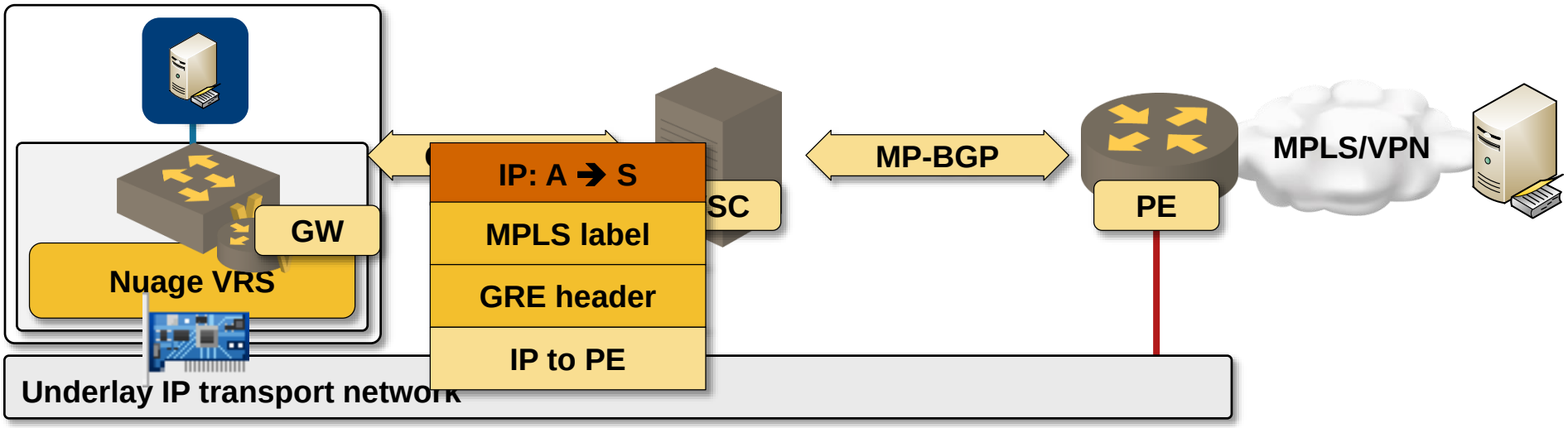
- PE-router sends VPNv4 or EVPN update to Nuage VSC
- VSC installs forwarding entries with BGP next hop + label in VRS
- VM sends IP packet to server (and GW MAC)

MPLS/VPN and EVPN Integration with Nuage VSP



- PE-router sends VPNv4 or EVPN update to Nuage VSC
- VSC installs forwarding entries with BGP next hop + label in VRS
- VM sends IP packet to server (and GW MAC)
- IP router in VRS performs L3 lookup

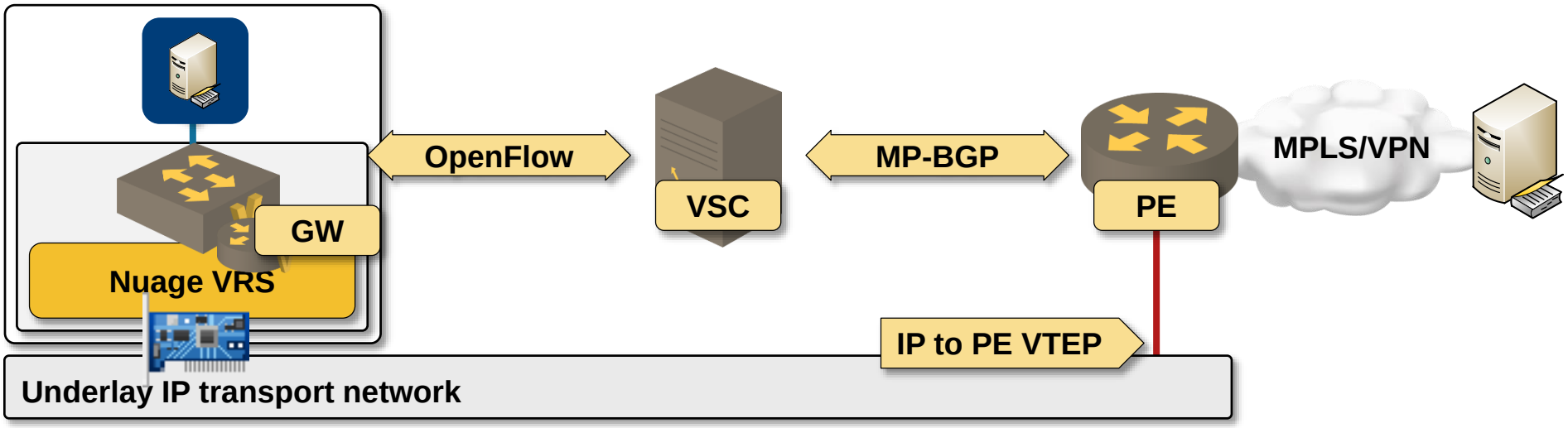
MPLS/VPN and EVPN Integration with Nuage VSP



- PE-router sends VPNv4 or EVPN update to Nuage VSC
- VSC installs forwarding entries with BGP next hop + label in VRS

- VM sends IP packet to server (and GW MAC)
- IP router in VRS performs L3 lookup
- IP packet is encapsulated in MPLS-GRE-IP or VXLAN-UDP-IP envelope

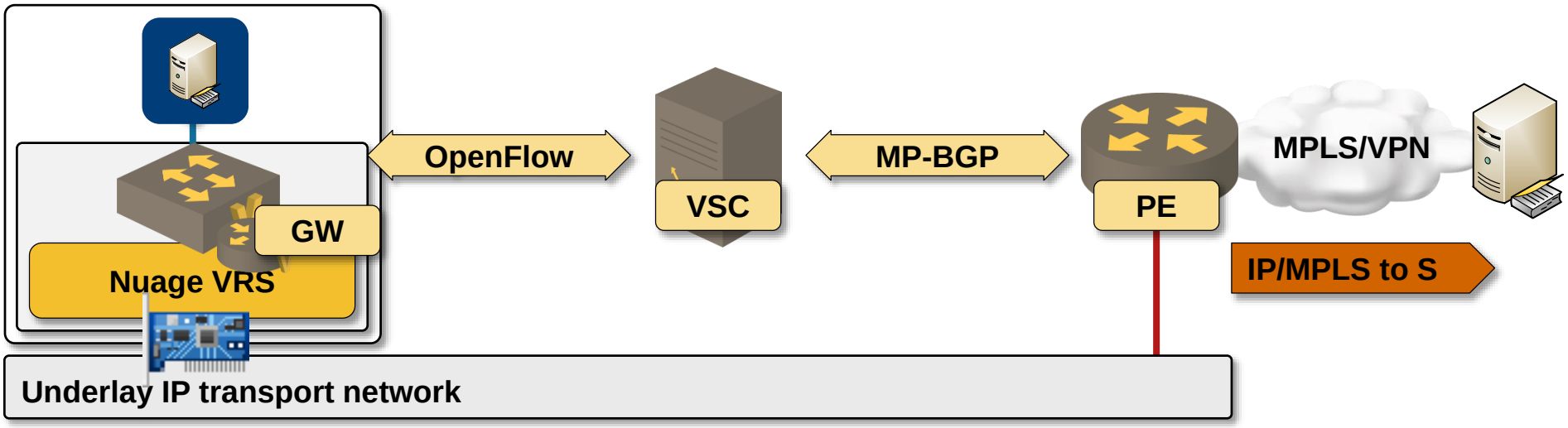
MPLS/VPN and EVPN Integration with Nuage VSP



- PE-router sends VPNv4 or EVPN update to Nuage VSC
- VSC installs forwarding entries with BGP next hop + label in VRS

- VM sends IP packet to server (and GW MAC)
- IP router in VRS performs L3 lookup
- IP packet is encapsulated in MPLS-GRE-IP or VXLAN-UDP-IP envelope
- PE router receives MPLS/VPN or VXLAN packet

MPLS/VPN and EVPN Integration with Nuage VSP



- PE-router sends VPNv4 or EVPN update to Nuage VSC
- VSC installs forwarding entries with BGP next hop + label in VRS

- VM sends IP packet to server (and GW MAC)
- IP router in VRS performs L3 lookup
- IP packet is encapsulated in MPLS-GRE-IP or VXLAN-UDP-IP envelope
- PE router receives MPLS/VPN or VXLAN packet
- PE router forwards VPN IP packet

Gateway Selection Criteria

Deployment format

- Low bandwidth ➔ VM
- High bandwidth ➔ hardware VTEP

Integration requirements

- Physical VLANs ➔ OVSDB or EVPN
- MPLS/VPN WAN ➔ EVPN + L3VPN

Choose an SDN controller that supports all the options you need

Scaling Security Groups

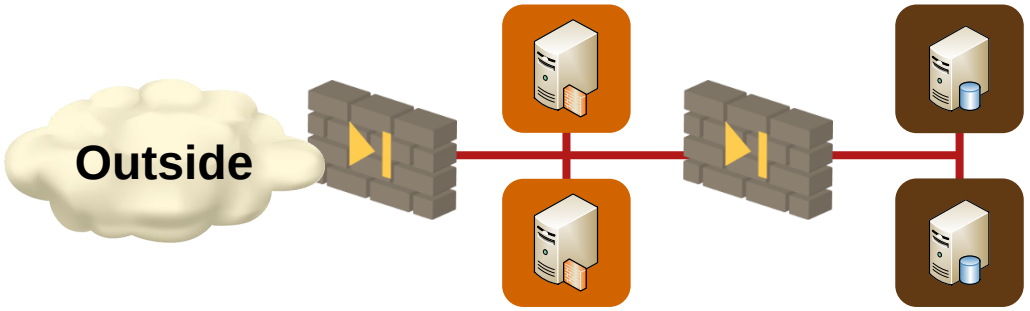
Security Groups 101

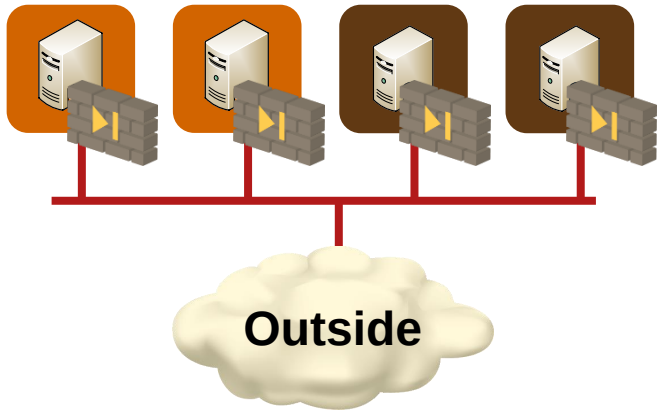
Security Groups Concepts

- Replace subnet-level firewalls (or ACLs) with per-VM firewalls/ACLs
- Increased intra-subnet security due to microsegmentation
- No chokepoint, no traffic tromboning
- No subnets ➔ no addressing limitations

Implementations

- CloudStack (on Linux-based hypervisors)
- OpenStack (Neutron plugin extension)
- VMware vCD/vCAC with vShield Edge or VMware NSX





Security Groups: Typical Implementation

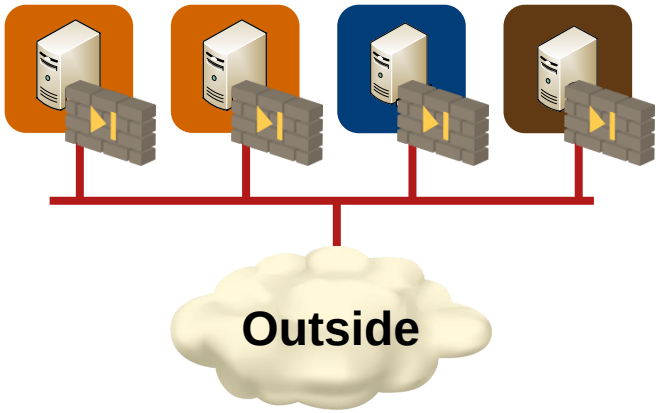
High-level view

- Assign VMs to groups
- Specify filtering rules between groups

Typical implementations

- Packet filter (OVS or Linux *iptables*)
- Each group exploded into a list of IP addresses
- ACL = Cartesian product of source-destination IP addresses

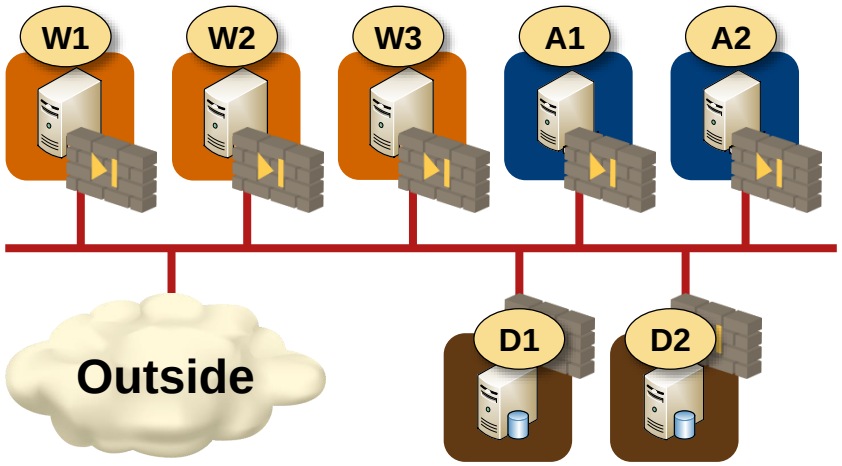
From	To	Port
Any	Web	80
Any	Web	443
Web	App	9000
App	DB	3306
Mgmt	All-VM	22



Cartesian Product Explained

From	To	Port
Any	Web	80
Any	Web	443
Web	App	9000
App	DB	3306
Mgmt	All-VM	22

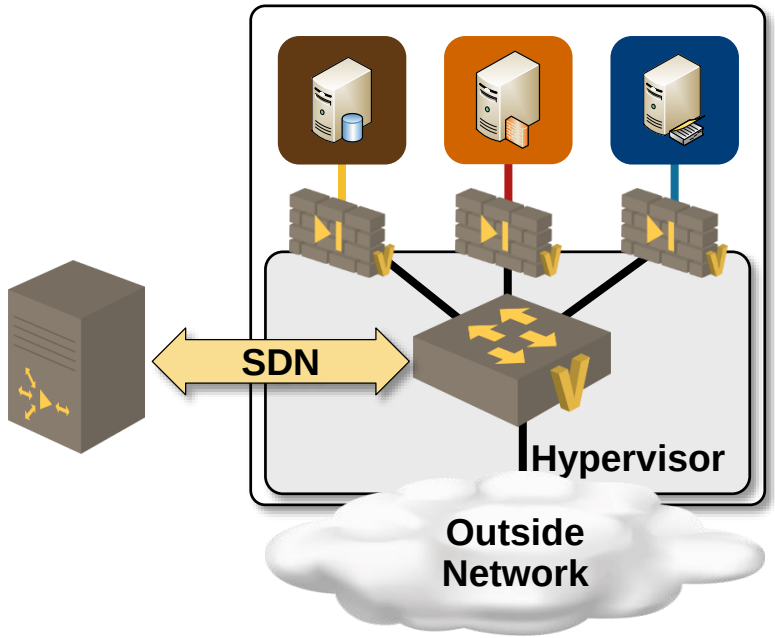
From	To	Port
Any	W1	80
Any	W2	80
Any	W3	80
Any	W1	443
Any	W2	443
Any	W3	443
W1	A1	9000
W1	A2	9000
W2	A1	9000
W2	A2	9000
W3	A1	9000
W3	A2	9000
…		



Security Groups: Scalability Challenges

Security group ACL = Cartesian product of IP addresses

- Long ACLs (performance usually degrades linearly with the ACL length)
- Whole ACL deployed on all VM NICs
  ➔ even further performance degradation
- Any change in security group membership (VM adds or removals) propagates to all hypevisors running tenant's VMs

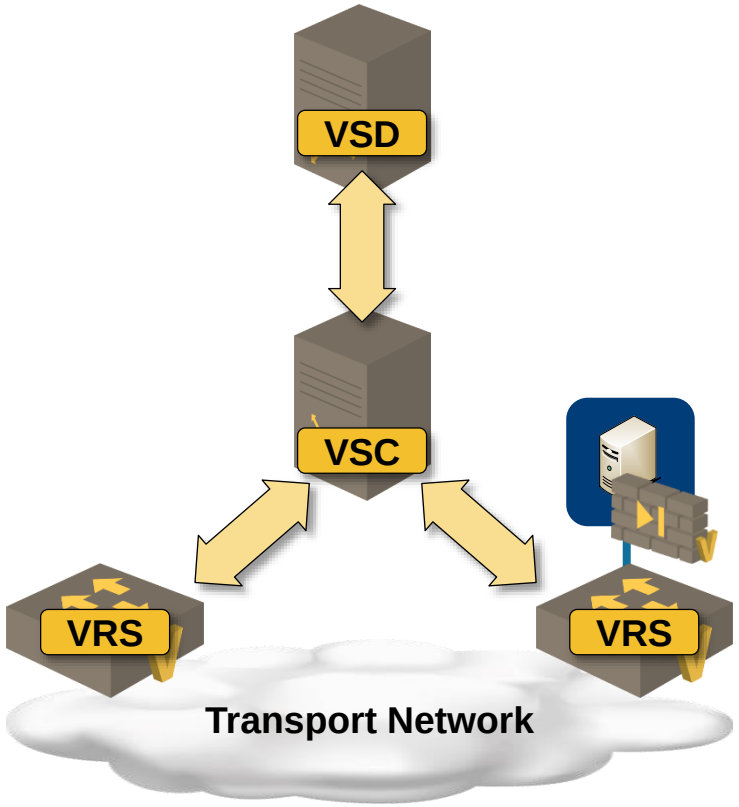


From	To	Port
Any	Web	80
Any	Web	443
Web	App	9000
App	DB	3306
Mgmt	All-VM	22

Security Groups in Nuage VSP

Security group membership = BGP community

- Remote VM security group attached to IP or MAC route
- Local VM security group attached to VM port



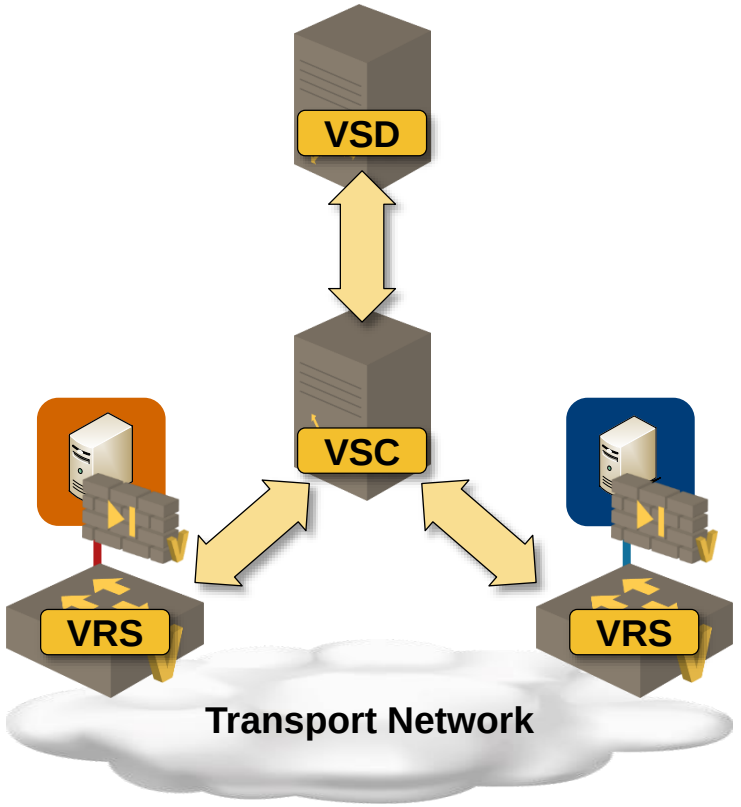
Security Groups in Nuage VSP

Security group membership = BGP community

- Remote VM security group attached to IP or MAC route
- Local VM security group attached to VM port

Typical sequence of events

- New VM is started on a hypervisor



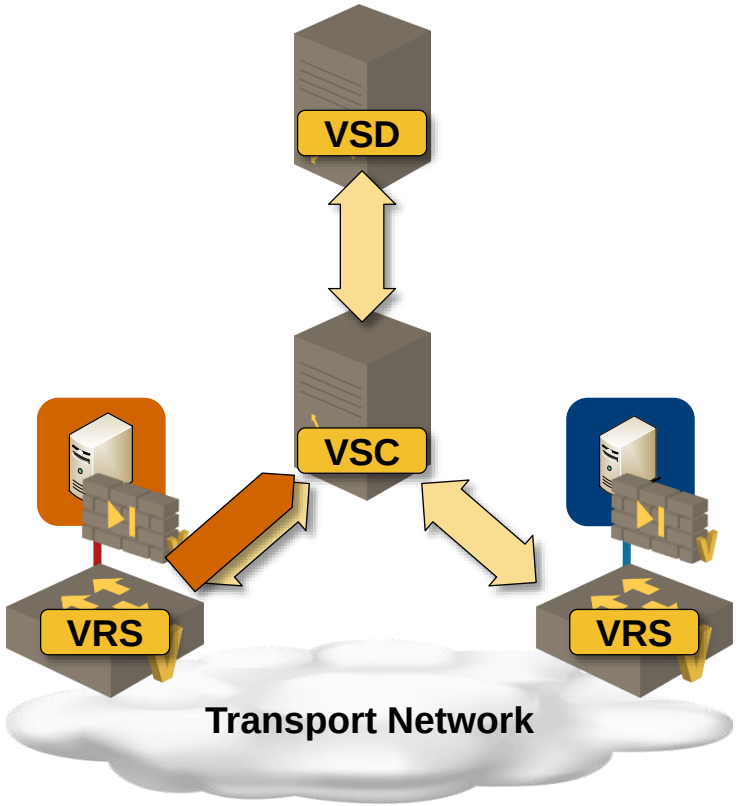
Security Groups in Nuage VSP

Security group membership = BGP community

- Remote VM security group attached to IP or MAC route
- Local VM security group attached to VM port

Typical sequence of events

- New VM is started on a hypervisor
- VRS notifies VSC



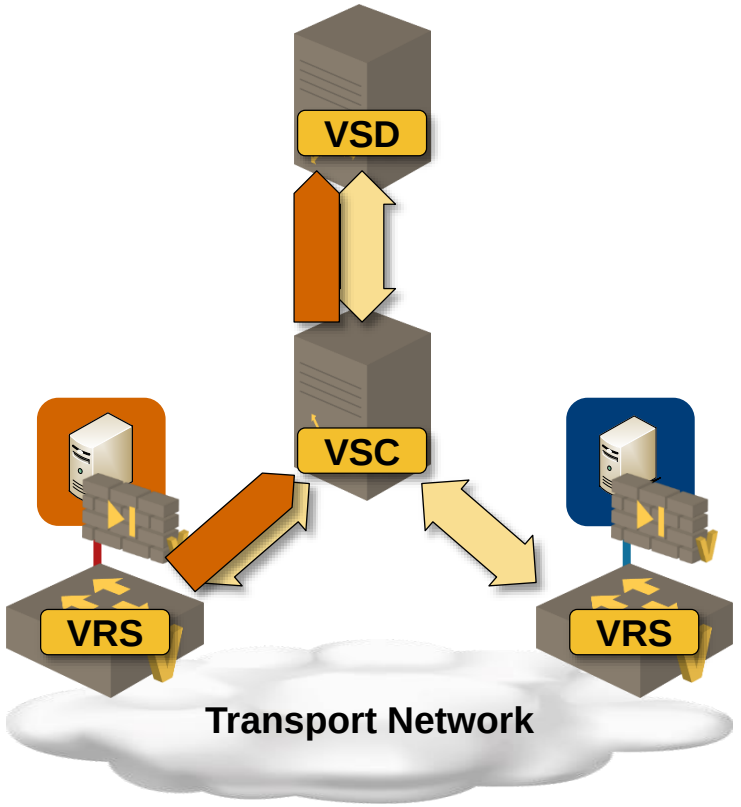
Security Groups in Nuage VSP

Security group membership = BGP community

- Remote VM security group attached to IP or MAC route
- Local VM security group attached to VM port

Typical sequence of events

- New VM is started on a hypervisor
- VRS notifies VSC
- VSC notifies VSD



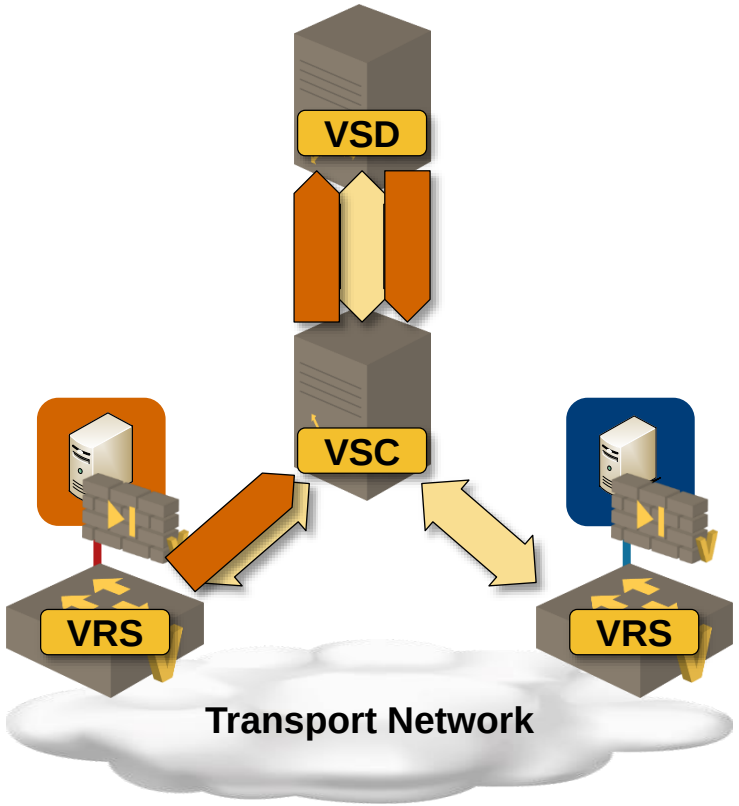
Security Groups in Nuage VSP

Security group membership = BGP community

- Remote VM security group attached to IP or MAC route
- Local VM security group attached to VM port

Typical sequence of events

- New VM is started on a hypervisor
- VRS notifies VSC
- VSC notifies VSD
- VSD assigns VM into a security group and replies to VSC



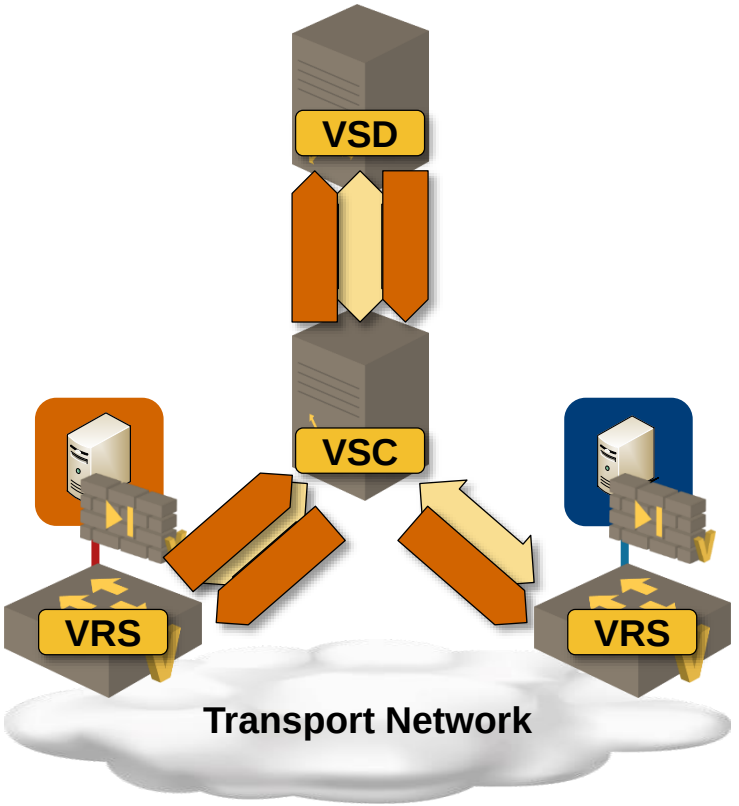
Security Groups in Nuage VSP

Security group membership = BGP community

- Remote VM security group attached to IP or MAC route
- Local VM security group attached to VM port

Typical sequence of events

- New VM is started on a hypervisor
- VRS notifies VSC
- VSC notifies VSD
- VSD assigns VM into a security group and replies to VSC
- VSC updates MAC-to-VTEP and IP-to-VTEP forwarding entries (incl. security group)



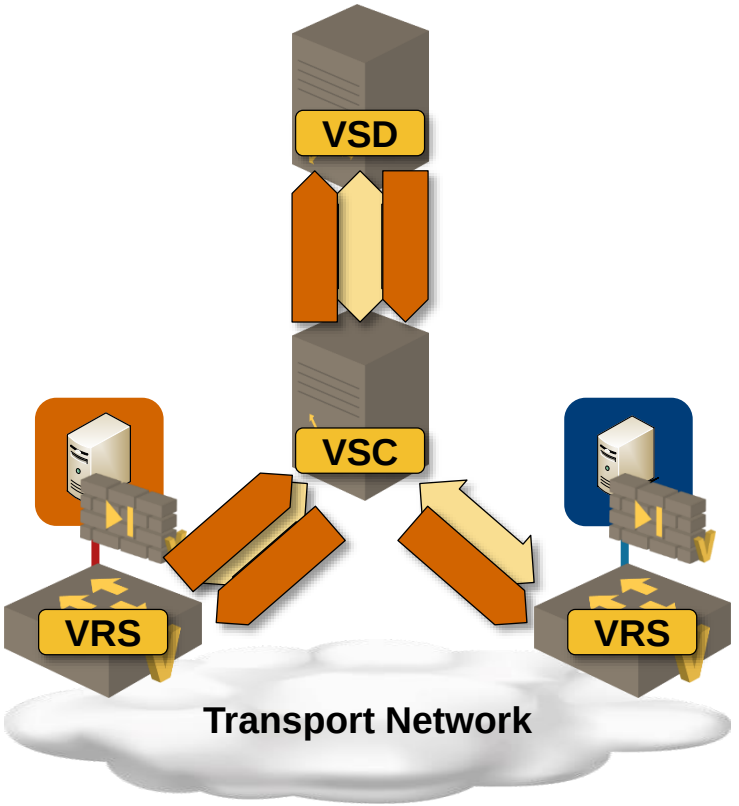
Security Groups in Nuage VSP

Security group membership = BGP community

- Remote VM security group attached to IP or MAC route
- Local VM security group attached to VM port

Typical sequence of events

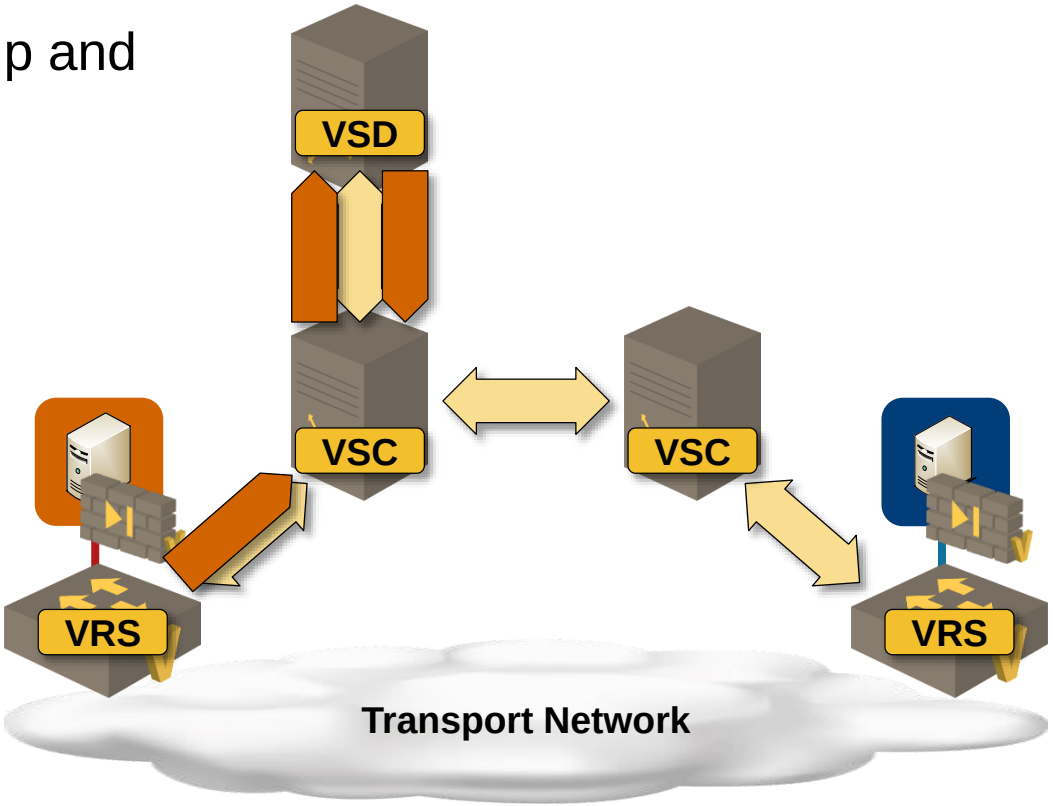
- New VM is started on a hypervisor
- VRS notifies VSC
- VSC notifies VSD
- VSD assigns VM into a security group and replies to VSC
- VSC updates MAC-to-VTEP and IP-to-VTEP forwarding entries (incl. security group)
- **ACL is not changed**



Security Groups Across Availability Zones

Typical sequence of events

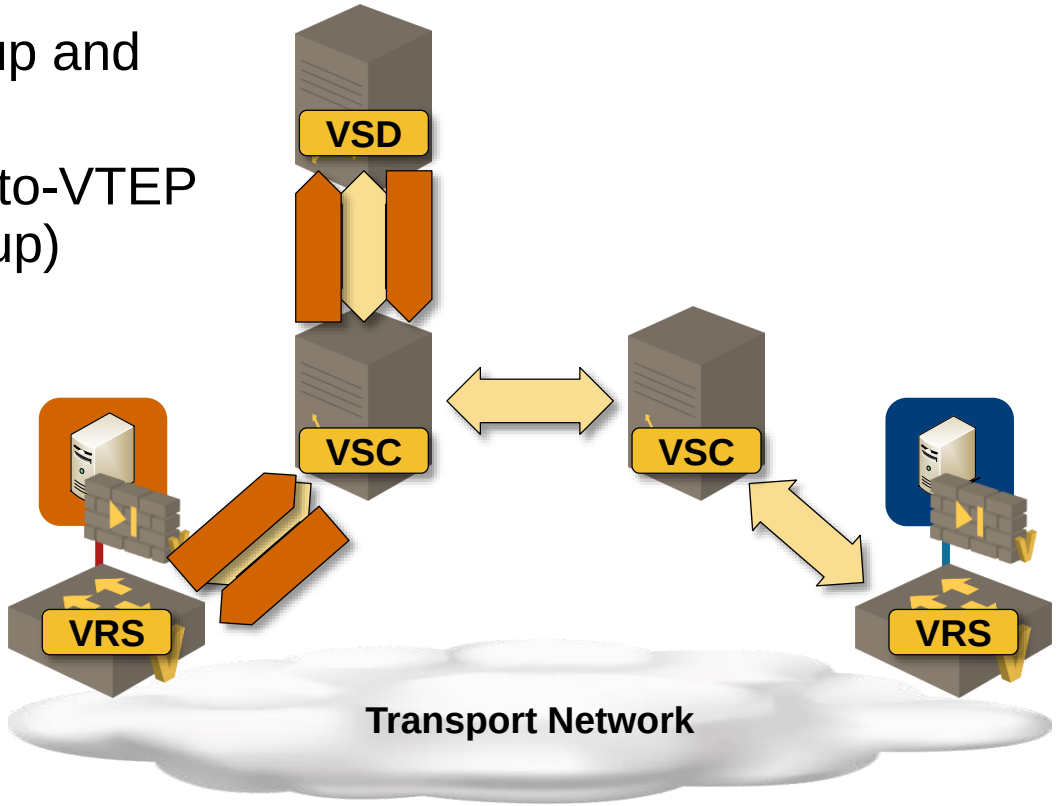
- New VM is started on a hypervisor
- VRS notifies VSC
- VSC notifies VSD
- VSD assigns VM into a security group and replies to VSC



Security Groups Across Availability Zones

Typical sequence of events

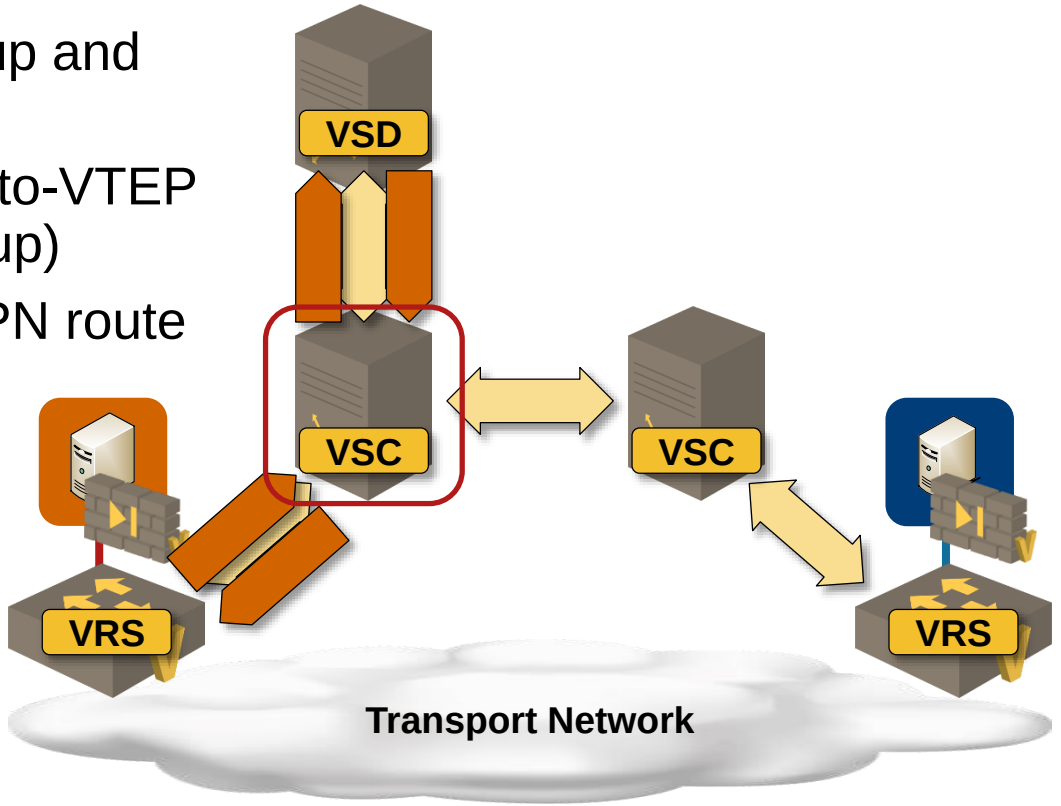
- New VM is started on a hypervisor
- VRS notifies VSC
- VSC notifies VSD
- VSD assigns VM into a security group and replies to VSC
- VSC updates MAC-to-VTEP and IP-to-VTEP forwarding entries (incl. security group)



Security Groups Across Availability Zones

Typical sequence of events

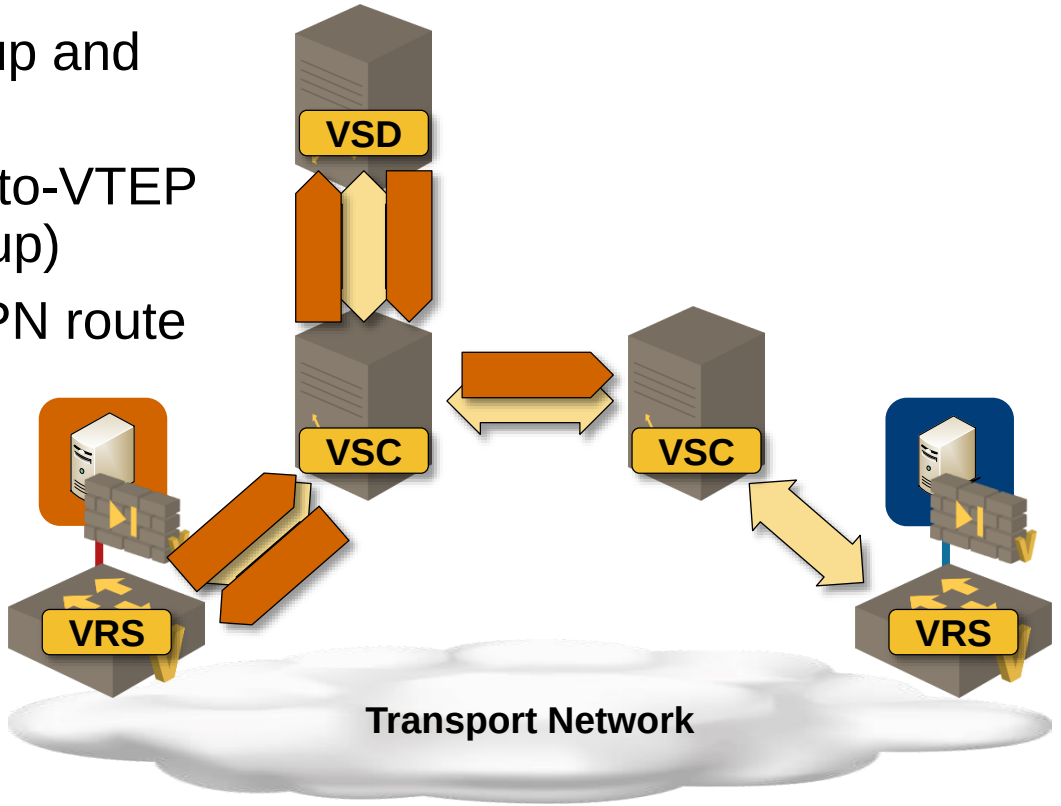
- New VM is started on a hypervisor
- VRS notifies VSC
- VSC notifies VSD
- VSD assigns VM into a security group and replies to VSC
- VSC updates MAC-to-VTEP and IP-to-VTEP forwarding entries (incl. security group)
- VSC originates new EVPN and IPVPN route (security group = BGP community)



Security Groups Across Availability Zones

Typical sequence of events

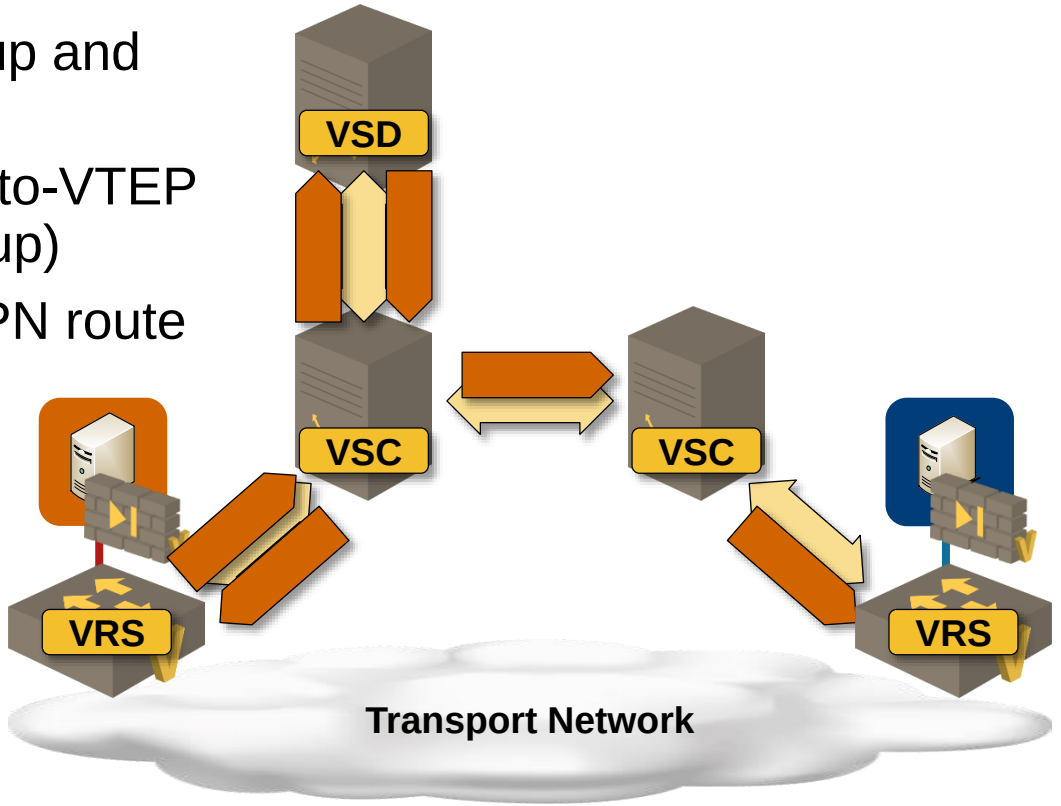
- New VM is started on a hypervisor
- VRS notifies VSC
- VSC notifies VSD
- VSD assigns VM into a security group and replies to VSC
- VSC updates MAC-to-VTEP and IP-to-VTEP forwarding entries (incl. security group)
- VSC originates new EVPN and IPVPN route (security group = BGP community)
- VSC sends BGP update to its BGP peers



Security Groups Across Availability Zones

Typical sequence of events

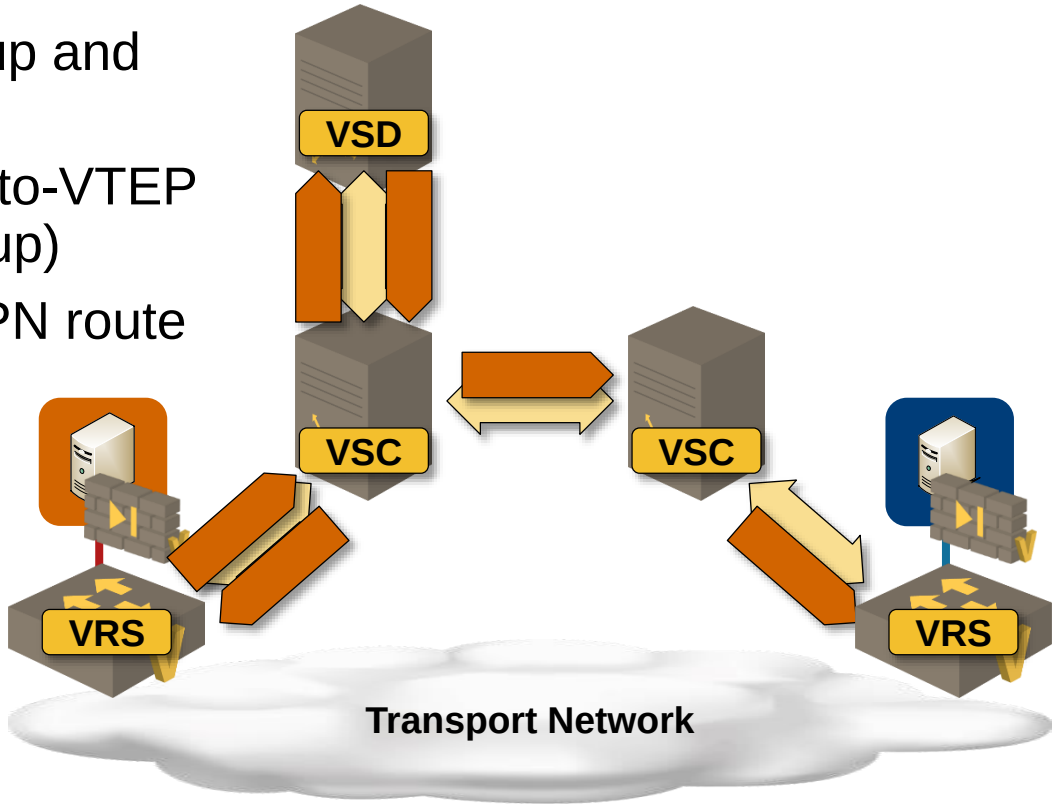
- New VM is started on a hypervisor
- VRS notifies VSC
- VSC notifies VSD
- VSD assigns VM into a security group and replies to VSC
- VSC updates MAC-to-VTEP and IP-to-VTEP forwarding entries (incl. security group)
- VSC originates new EVPN and IPVPN route (security group = BGP community)
- VSC sends BGP update to its BGP peers
- Remote VSC updates forwarding entries in remote VRS



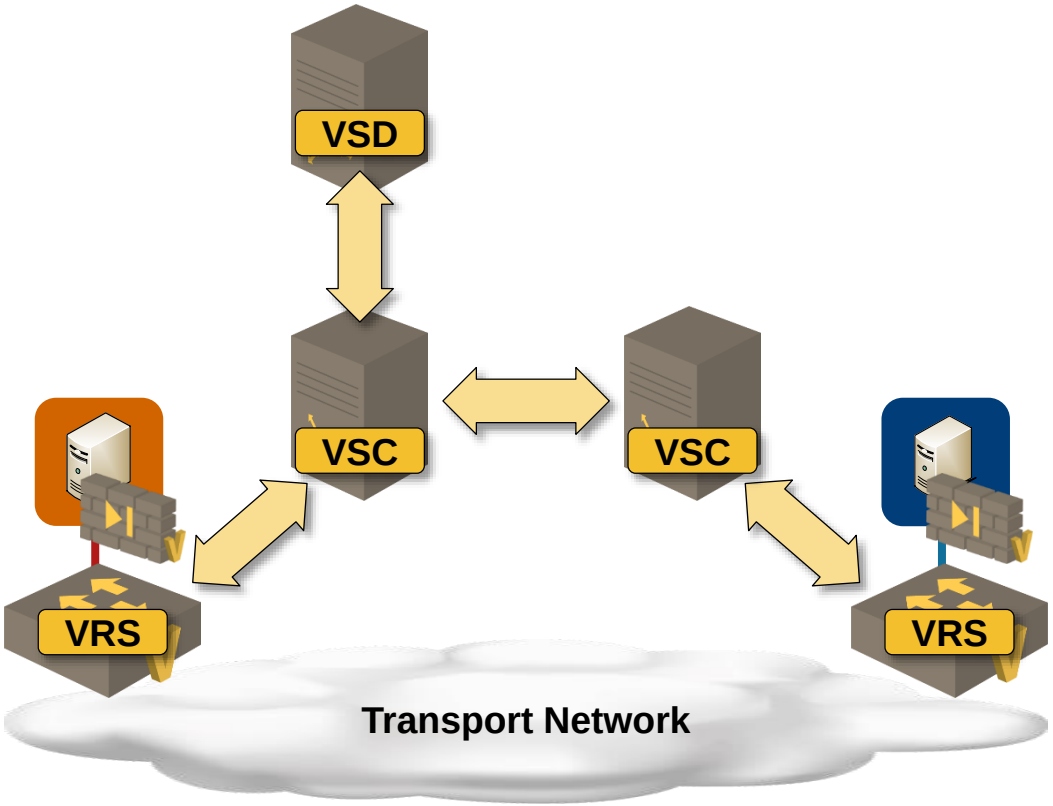
Security Groups Across Availability Zones

Typical sequence of events

- New VM is started on a hypervisor
- VRS notifies VSC
- VSC notifies VSD
- VSD assigns VM into a security group and replies to VSC
- VSC updates MAC-to-VTEP and IP-to-VTEP forwarding entries (incl. security group)
- VSC originates new EVPN and IPVPN route (security group = BGP community)
- VSC sends BGP update to its BGP peers
- Remote VSC updates forwarding entries in remote VRS
- **ACL is not changed**

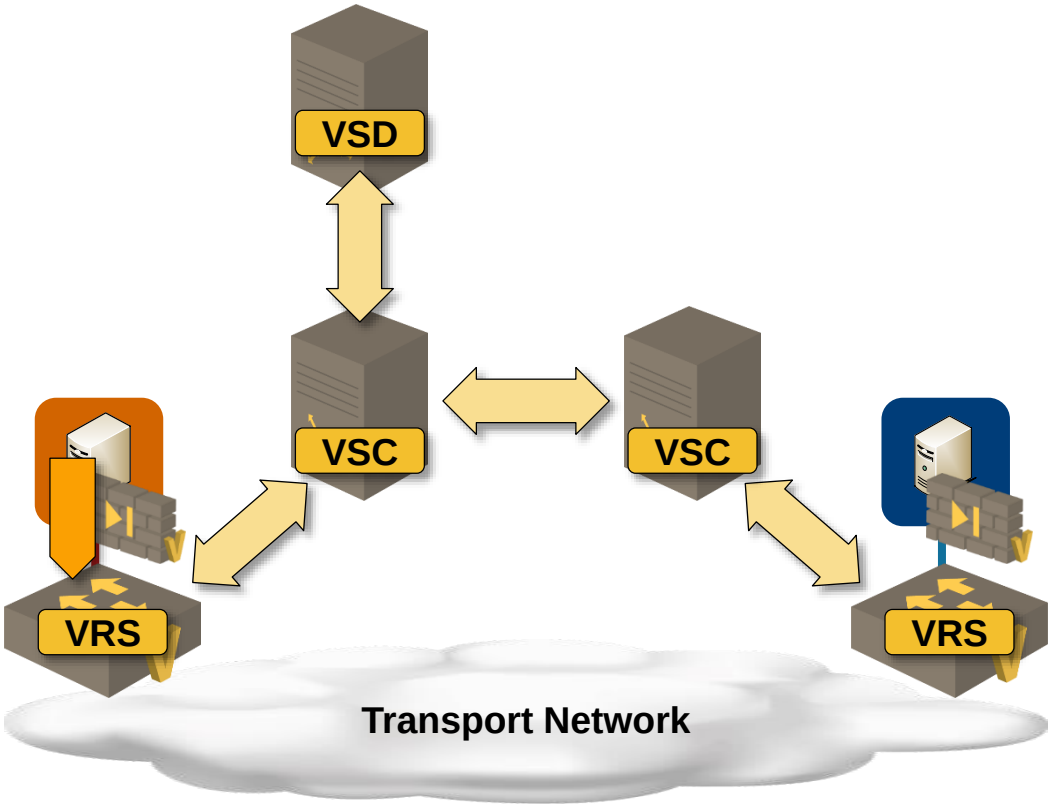


Security Groups in Nuage VSP: Data Plane



Security Groups in Nuage VSP: Data Plane

VM sends an IP packet

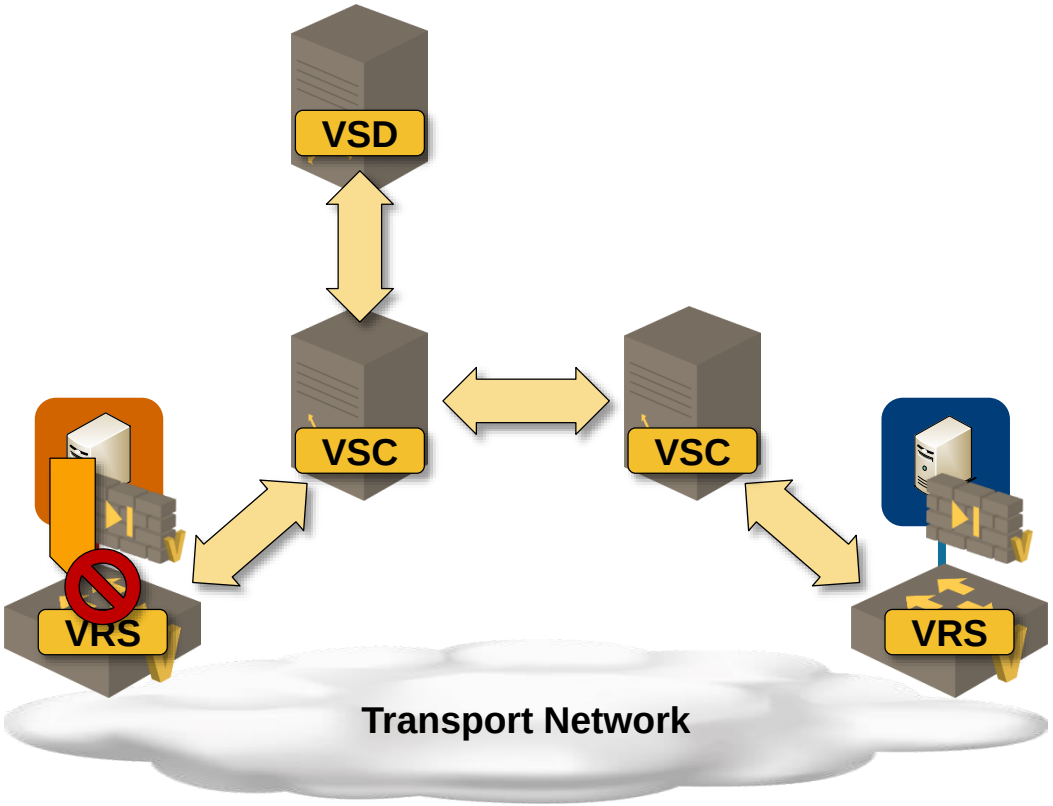


Security Groups in Nuage VSP: Data Plane

VM sends an IP packet

Ingress ACL check on ingress VRS

- *From* security group = VM NIC group
- *To* security group = BGP community



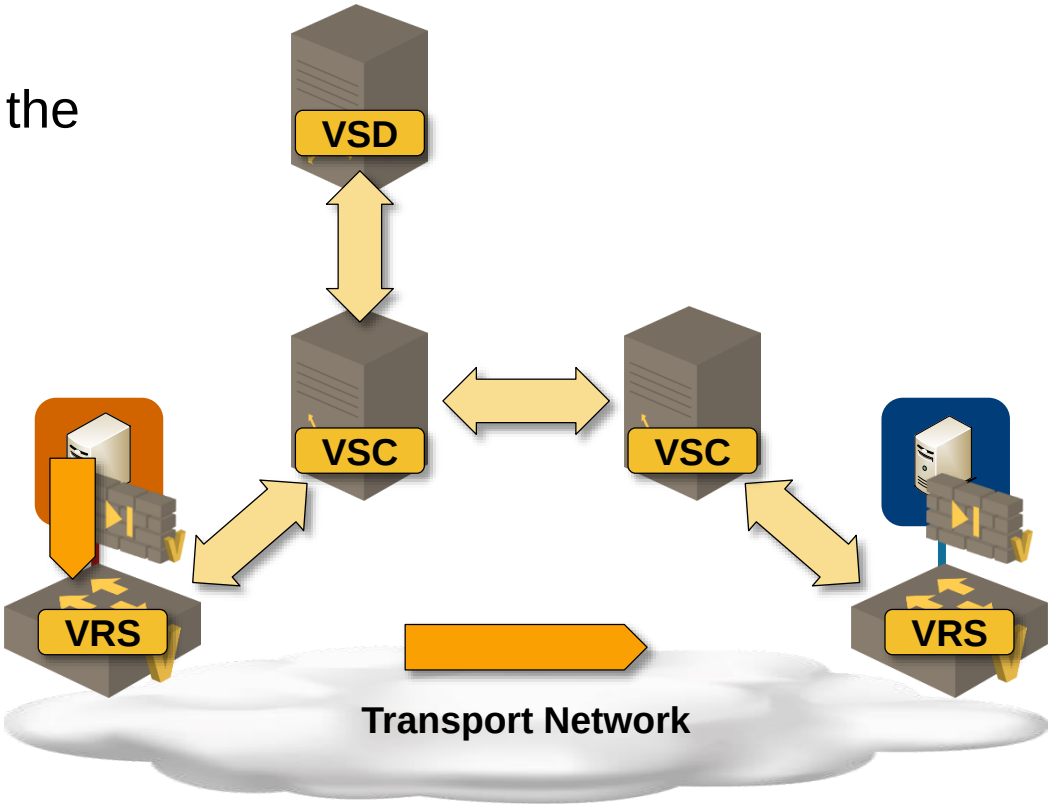
Security Groups in Nuage VSP: Data Plane

VM sends an IP packet

Ingress ACL check on ingress VRS

- *From* security group = VM NIC group
- *To* security group = BGP community

Encapsulated VM frame is sent across the transport network



Security Groups in Nuage VSP: Data Plane

VM sends an IP packet

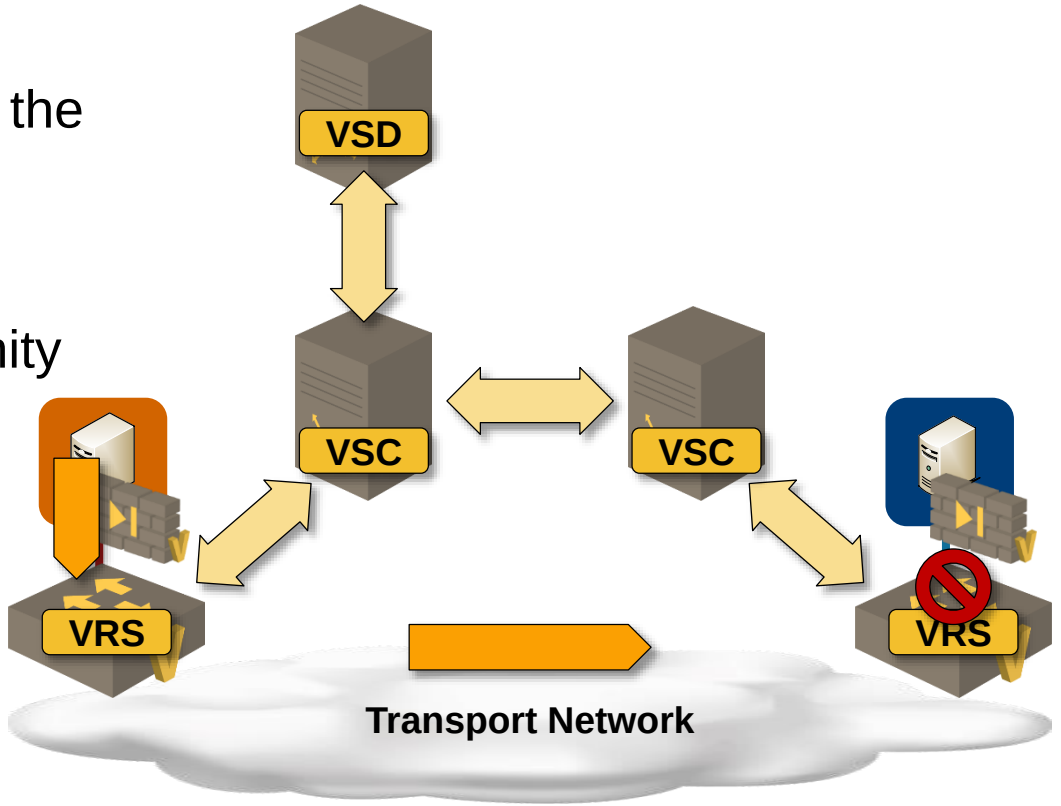
Ingress ACL check on ingress VRS

- *From* security group = VM NIC group
- *To* security group = BGP community

Encapsulated VM frame is sent across the transport network

Egress ACL check on egress VRS

- *From* security group = BGP community
- *To* security group = VM NIC group



Security Groups in Nuage VSP: Data Plane

VM sends an IP packet

Ingress ACL check on ingress VRS

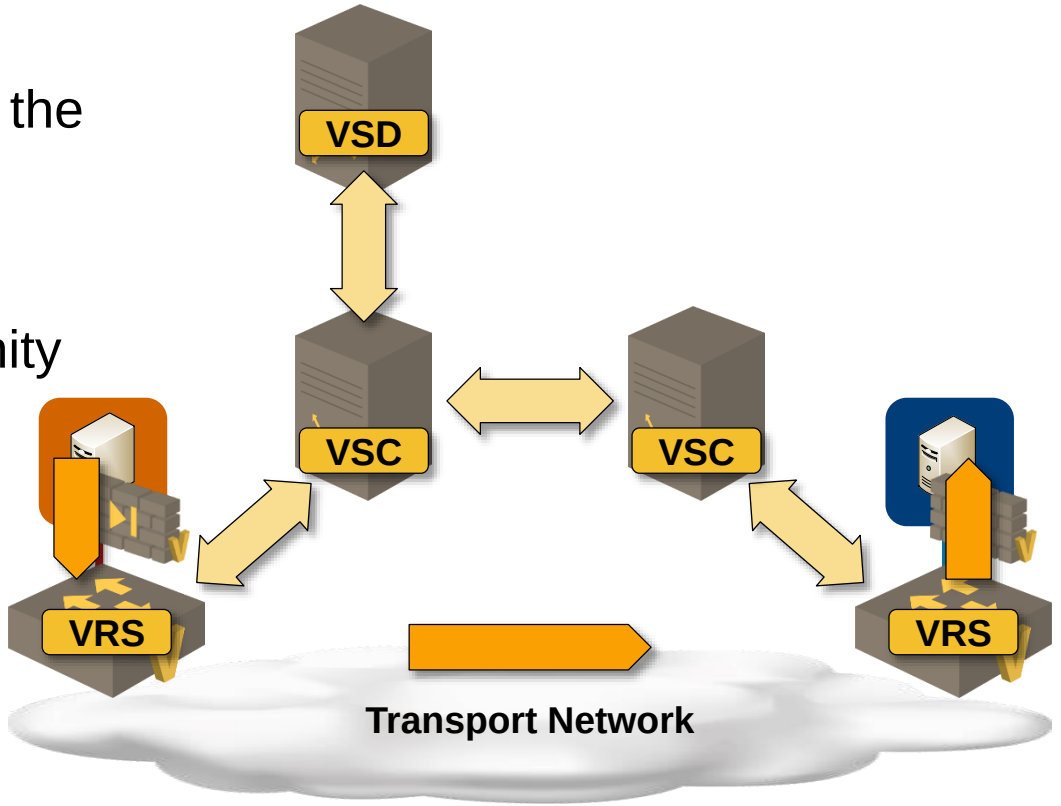
- *From* security group = VM NIC group
- *To* security group = BGP community

Encapsulated VM frame is sent across the transport network

Egress ACL check on egress VRS

- *From* security group = BGP community
- *To* security group = VM NIC group

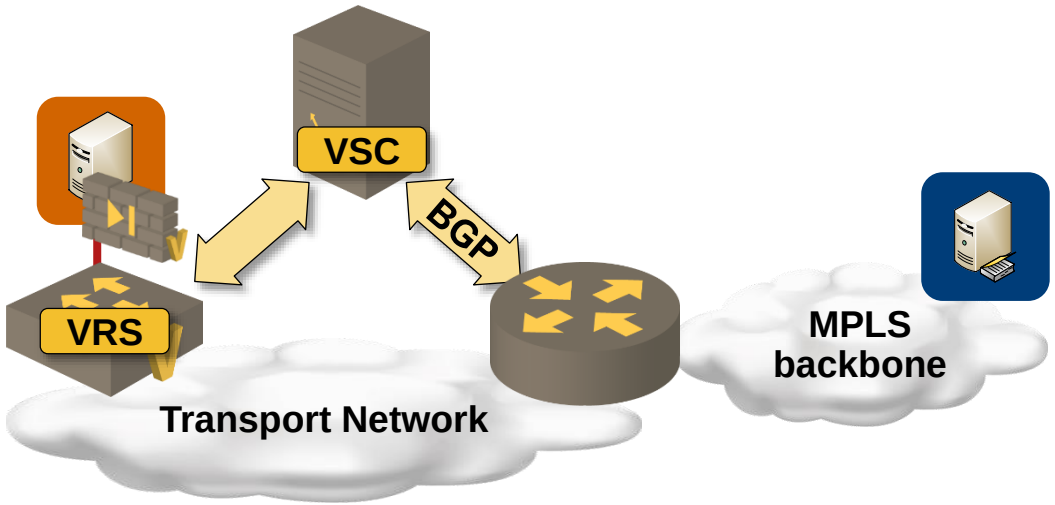
Packet is delivered to target VM



Security Groups Across Multiple Domains

Security groups (in BGP communities) can extend across MPLS/VPN backbone

- Automatic ingress/egress filters on VM NICs
- Requires trust (or strict filters) between cloud and MPLS/VPN networks



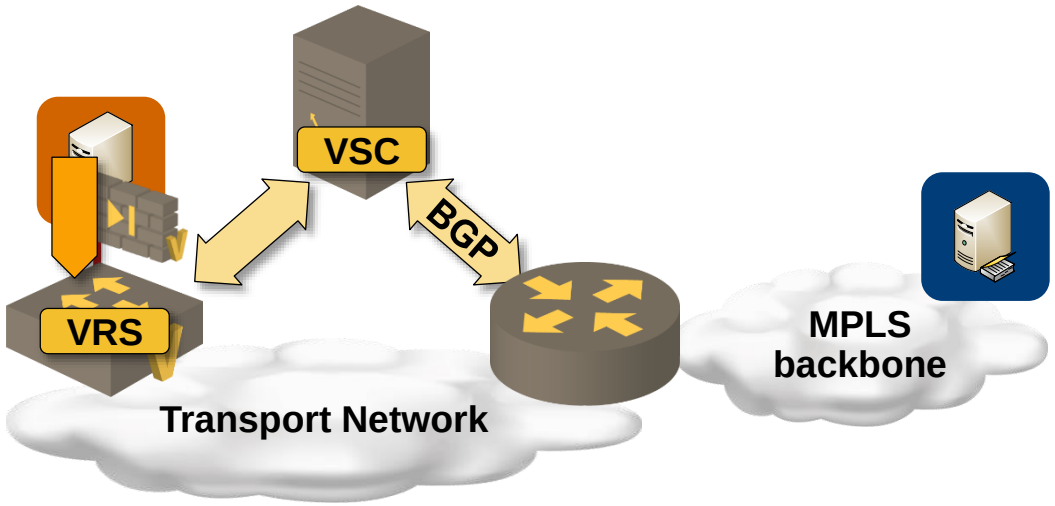
Security Groups Across Multiple Domains

Security groups (in BGP communities) can extend across MPLS/VPN backbone

- Automatic ingress/egress filters on VM NICs
- Requires trust (or strict filters) between cloud and MPLS/VPN networks

VM to remote host:

- VM sends a packet



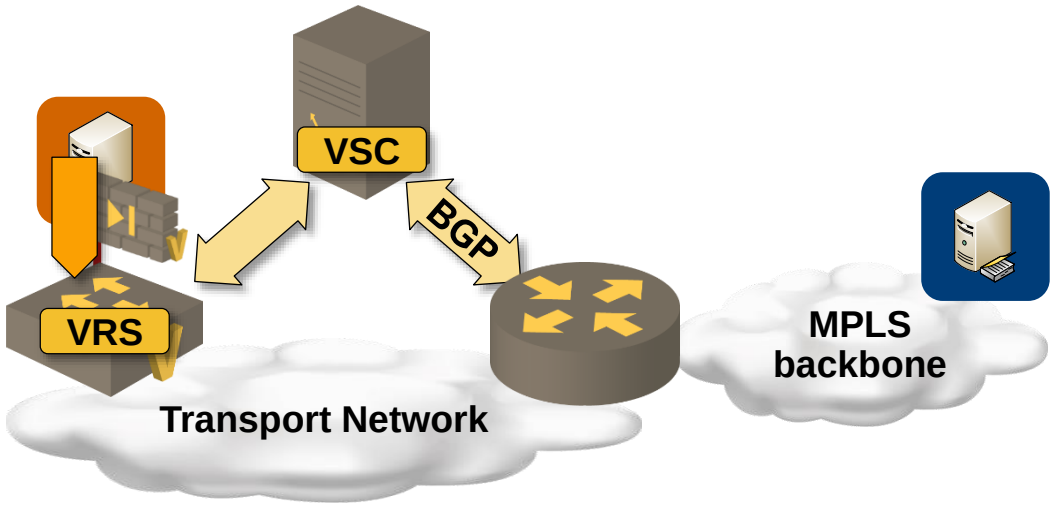
Security Groups Across Multiple Domains

Security groups (in BGP communities) can extend across MPLS/VPN backbone

- Automatic ingress/egress filters on VM NICs
- Requires trust (or strict filters) between cloud and MPLS/VPN networks

VM to remote host:

- VM sends a packet
- *Ingress* ACL on VRS
- Packet delivered to VM



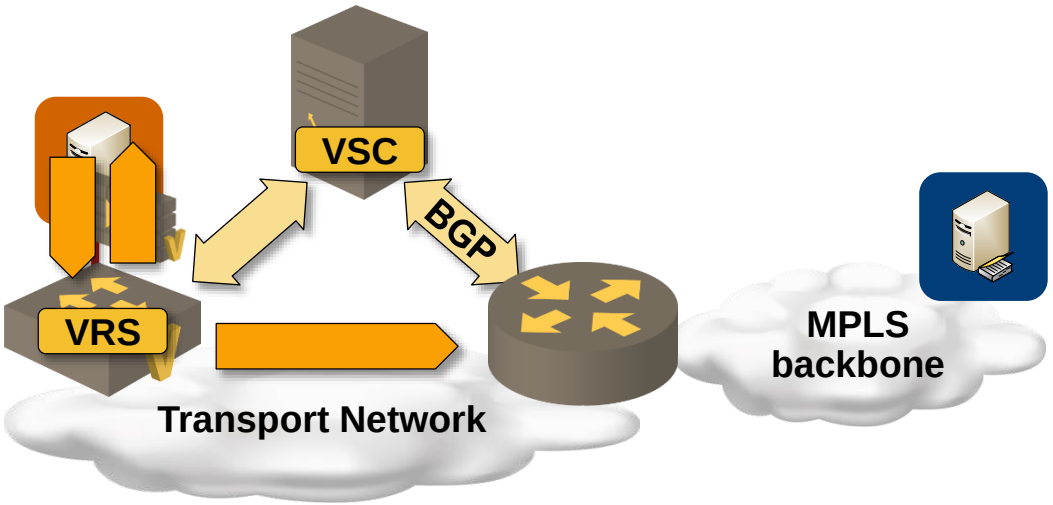
Security Groups Across Multiple Domains

Security groups (in BGP communities) can extend across MPLS/VPN backbone

- Automatic ingress/egress filters on VM NICs
- Requires trust (or strict filters) between cloud and MPLS/VPN networks

VM to remote host:

- VM sends a packet
- *Ingress* ACL on VRS
- IP packet sent from VRS to PE-router
- Packet delivered to VM



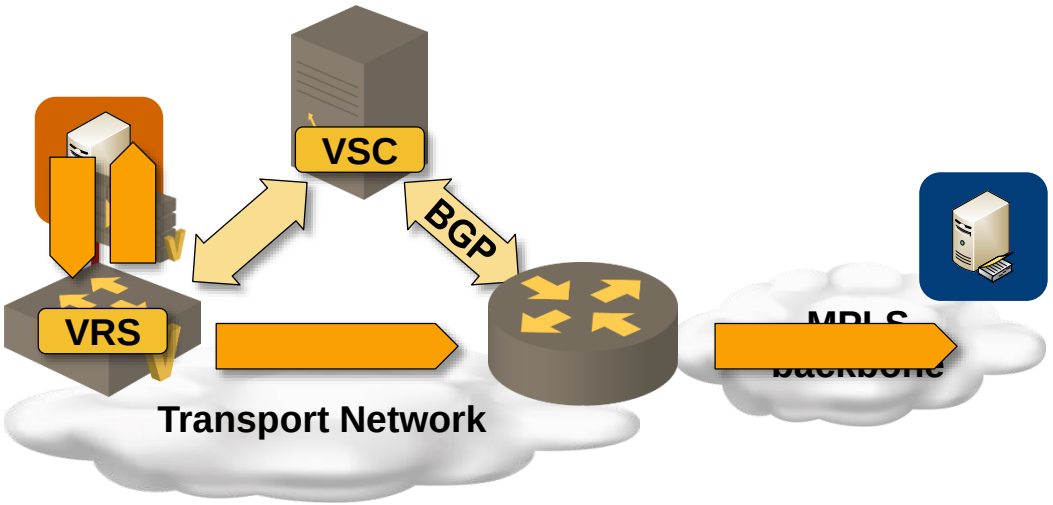
Security Groups Across Multiple Domains

Security groups (in BGP communities) can extend across MPLS/VPN backbone

- Automatic ingress/egress filters on VM NICs
- Requires trust (or strict filters) between cloud and MPLS/VPN networks

VM to remote host:

- VM sends a packet
- *Ingress* ACL on VRS
- IP packet sent from VRS to PE-router
- IP packet delivered to remote host
- Packet delivered to VM



Security Groups Across Multiple Domains

Security groups (in BGP communities) can extend across MPLS/VPN backbone

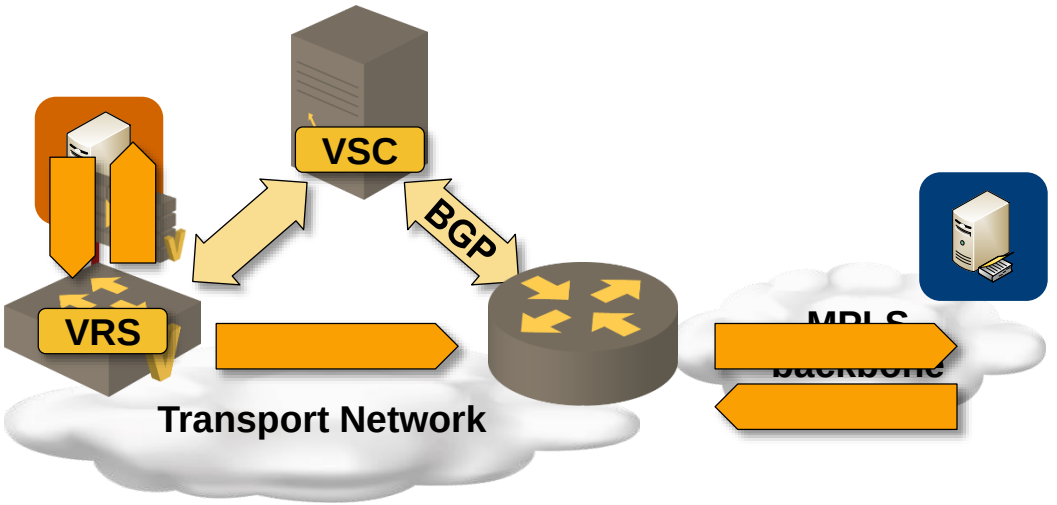
- Automatic ingress/egress filters on VM NICs
- Requires trust (or strict filters) between cloud and MPLS/VPN networks

VM to remote host:

- VM sends a packet
- *Ingress* ACL on VRS
- IP packet sent from VRS to PE-router
- IP packet delivered to remote host

Remote host to VM:

- IP packet received by PE-router
- Packet delivered to VM



Security Groups Across Multiple Domains

Security groups (in BGP communities) can extend across MPLS/VPN backbone

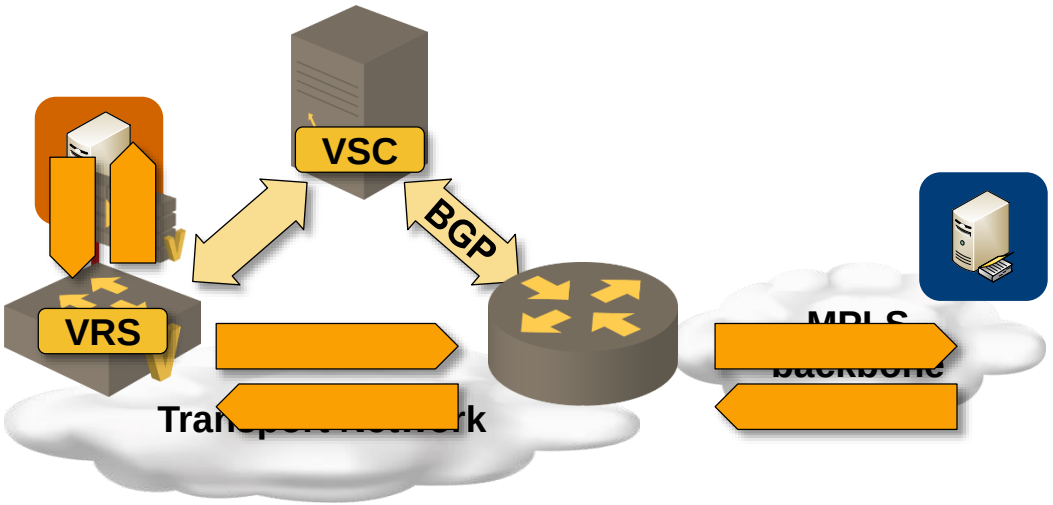
- Automatic ingress/egress filters on VM NICs
- Requires trust (or strict filters) between cloud and MPLS/VPN networks

VM to remote host:

- VM sends a packet
- *Ingress* ACL on VRS
- IP packet sent from VRS to PE-router
- IP packet delivered to remote host

Remote host to VM:

- IP packet received by PE-router
- IP packet delivered to VRS
- Packet delivered to VM



Security Groups Across Multiple Domains

Security groups (in BGP communities) can extend across MPLS/VPN backbone

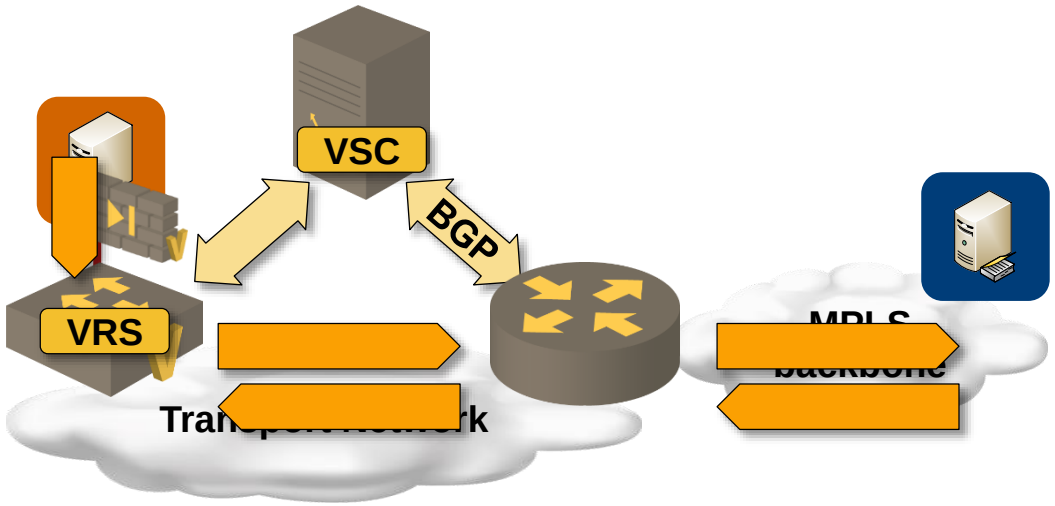
- Automatic ingress/egress filters on VM NICs
- Requires trust (or strict filters) between cloud and MPLS/VPN networks

VM to remote host:

- VM sends a packet
- *Ingress* ACL on VRS
- IP packet sent from VRS to PE-router
- IP packet delivered to remote host

Remote host to VM:

- IP packet received by PE-router
- IP packet delivered to VRS
- *Egress* ACL on VRS



Security Groups Across Multiple Domains

Security groups (in BGP communities) can extend across MPLS/VPN backbone

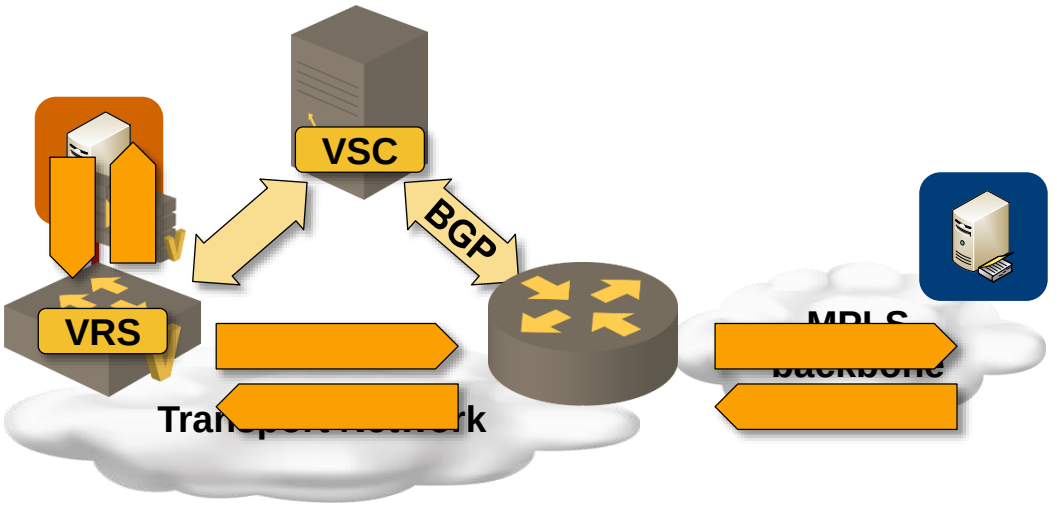
- Automatic ingress/egress filters on VM NICs
- Requires trust (or strict filters) between cloud and MPLS/VPN networks

VM to remote host:

- VM sends a packet
- *Ingress* ACL on VRS
- IP packet sent from VRS to PE-router
- IP packet delivered to remote host

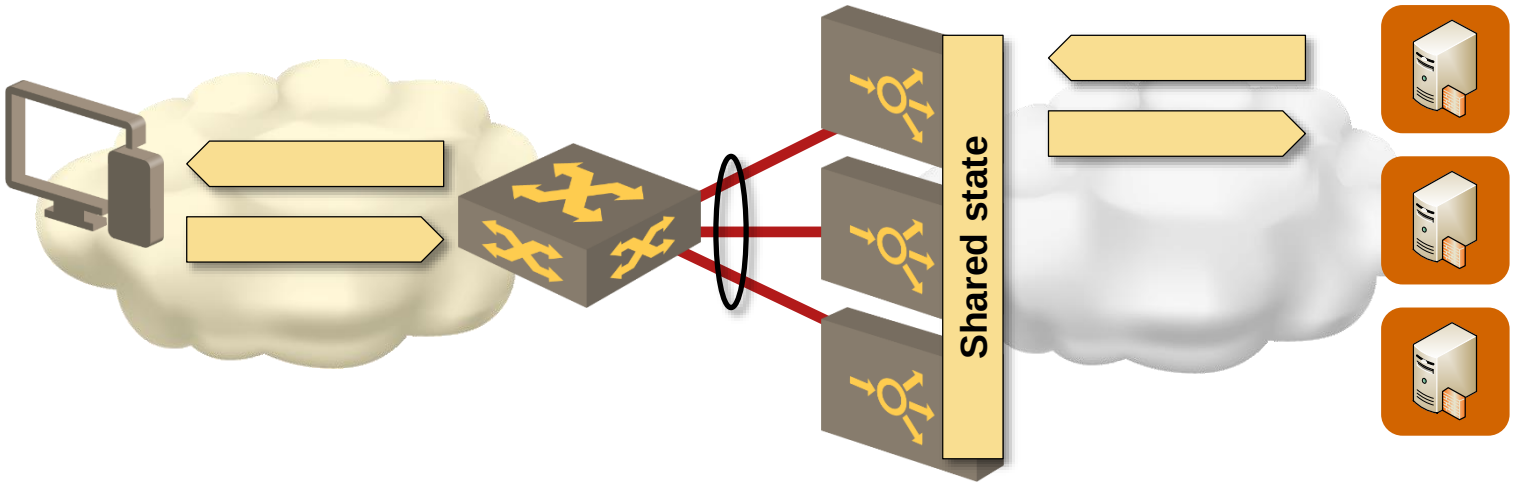
Remote host to VM:

- IP packet received by PE-router
- IP packet delivered to VRS
- *Egress* ACL on VRS
- Packet delivered to VM



Scale-Out NAT

Large Scale NAT



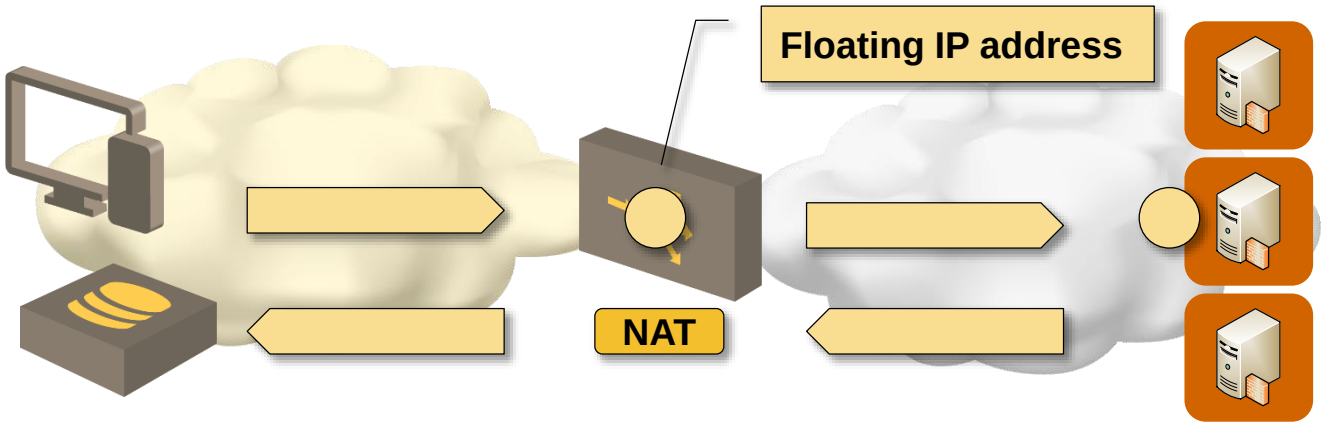
Scale-out NAT is hard problem

- No guarantee of symmetrical paths
  (Best case: rehashing after topology change)
- Shared state tied to outside IP address
- State must be distributed and synchronized across all NAT cluster members

Maybe we're solving the wrong problem

Remember:
Focus on User Needs

Typical Cloud Application Requirements



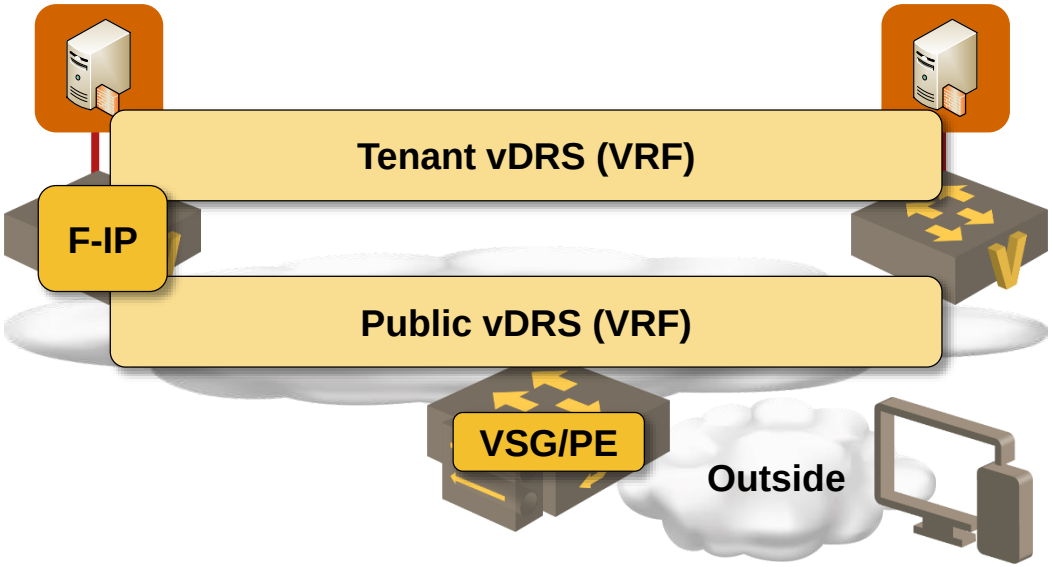
- Virtual machines with public IP addresses (Floating IP address)
  ➔ static stateless NAT
- Access to outside servers
  ➔ dynamic stateful NAPT, outside source address is irrelevant

Equivalent to Amazon VPC behavior

Floating IP Addresses in Nuage VSP

Setup

- Floating IP from public vDRS is allocated to a tenant VM
- 1:1 NAT rule is created on the hypervisor



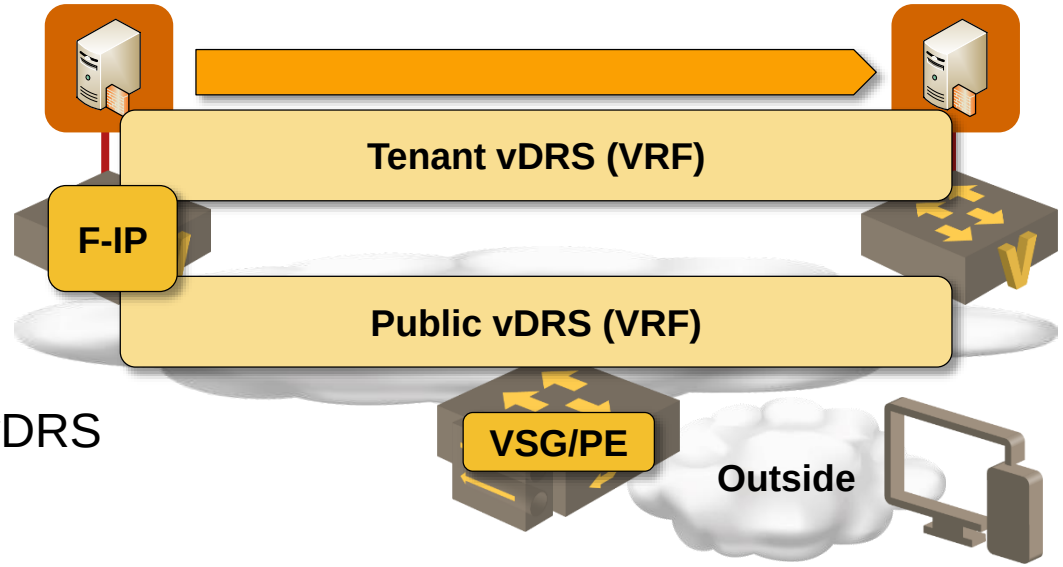
Floating IP Addresses in Nuage VSP

Setup

- Floating IP from public vDRS is allocated to a tenant VM
- 1:1 NAT rule is created on the hypervisor

Internal communication

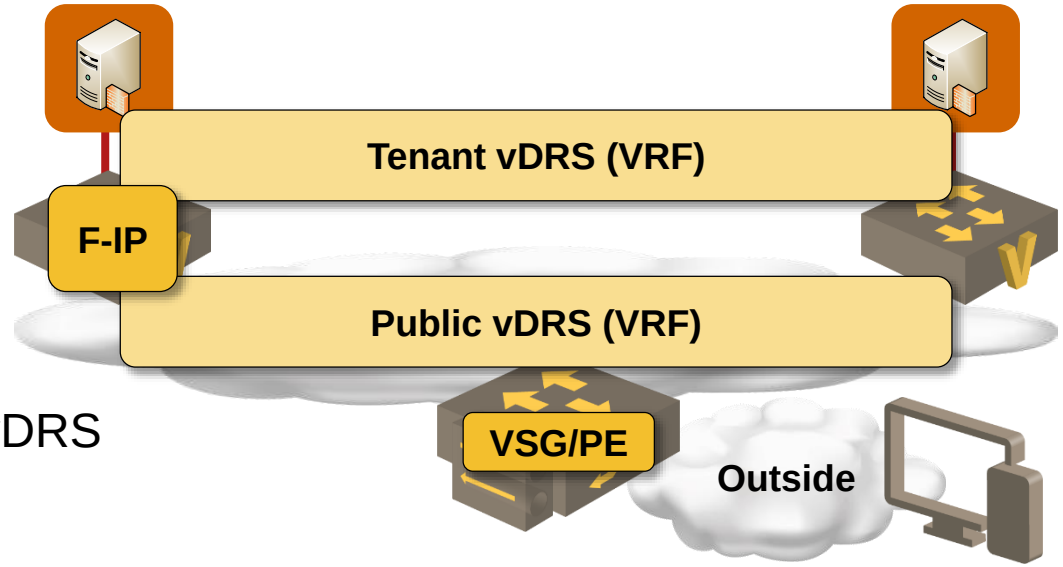
- Destination IP address is within tenant vDRS
- NAT rule is not invoked



Floating IP Addresses in Nuage VSP

Setup

- Floating IP from public vDRS is allocated to a tenant VM
- 1:1 NAT rule is created on the hypervisor



Internal communication

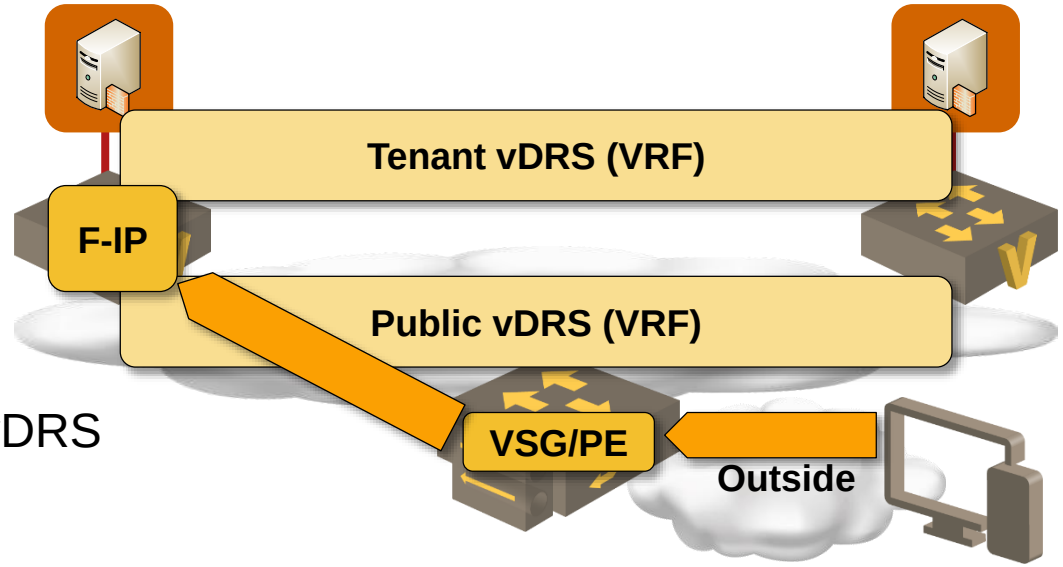
- Destination IP address is within tenant vDRS
- NAT rule is not invoked

Outside-to-inside

Floating IP Addresses in Nuage VSP

Setup

- Floating IP from public vDRS is allocated to a tenant VM
- 1:1 NAT rule is created on the hypervisor



Internal communication

- Destination IP address is within tenant vDRS
- NAT rule is not invoked

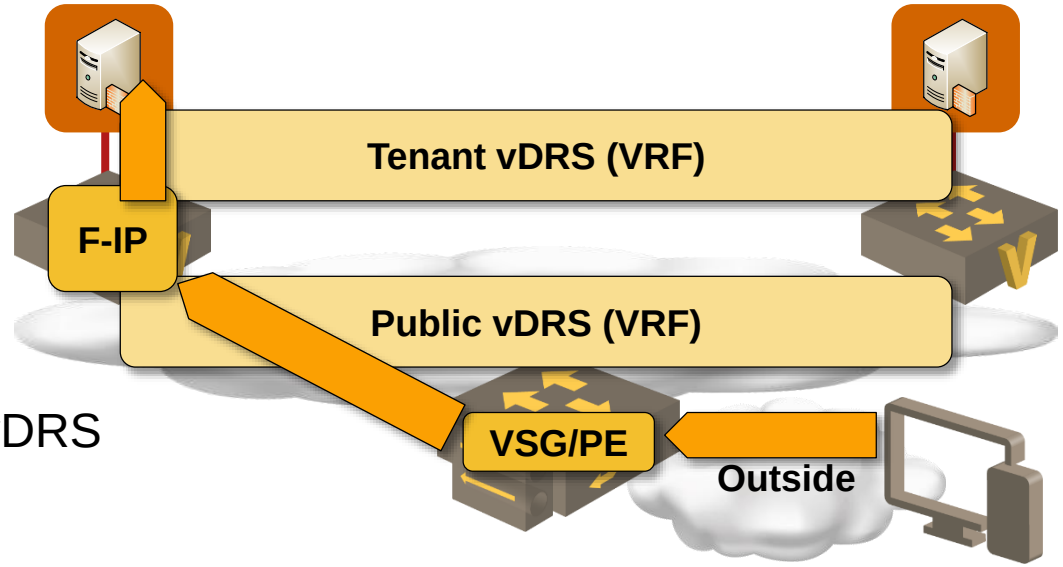
Outside-to-inside

- Packet sent to IP address in public vDRS (received by hypervisor)

Floating IP Addresses in Nuage VSP

Setup

- Floating IP from public vDRS is allocated to a tenant VM
- 1:1 NAT rule is created on the hypervisor



Internal communication

- Destination IP address is within tenant vDRS
- NAT rule is not invoked

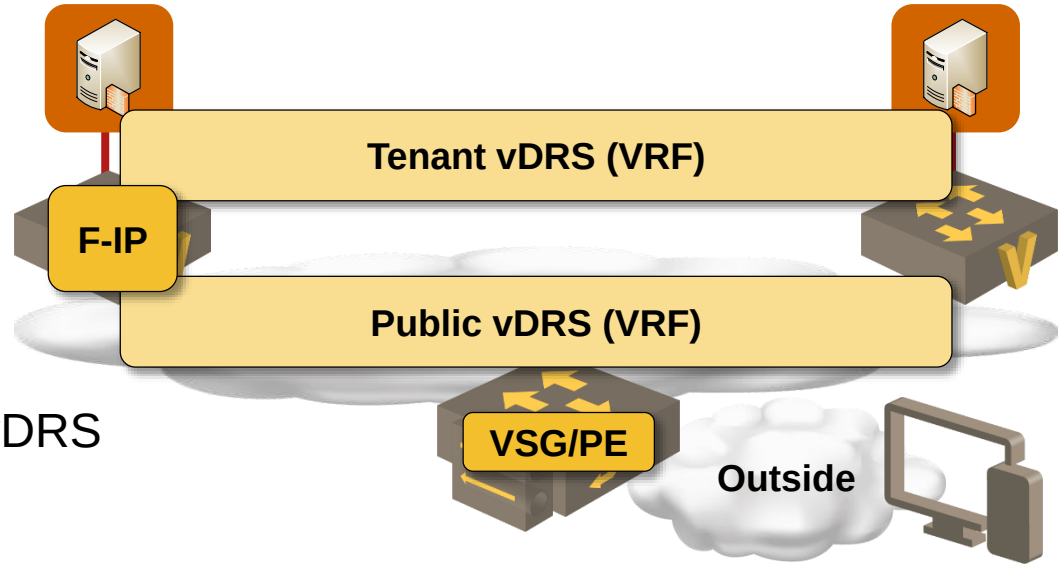
Outside-to-inside

- Packet sent to IP address in public vDRS (received by hypervisor)
- Hypervisor translates destination IP address to VM IP address

Floating IP Addresses in Nuage VSP

Setup

- Floating IP from public vDRS is allocated to a tenant VM
- 1:1 NAT rule is created on the hypervisor



Internal communication

- Destination IP address is within tenant vDRS
- NAT rule is not invoked

Outside-to-inside

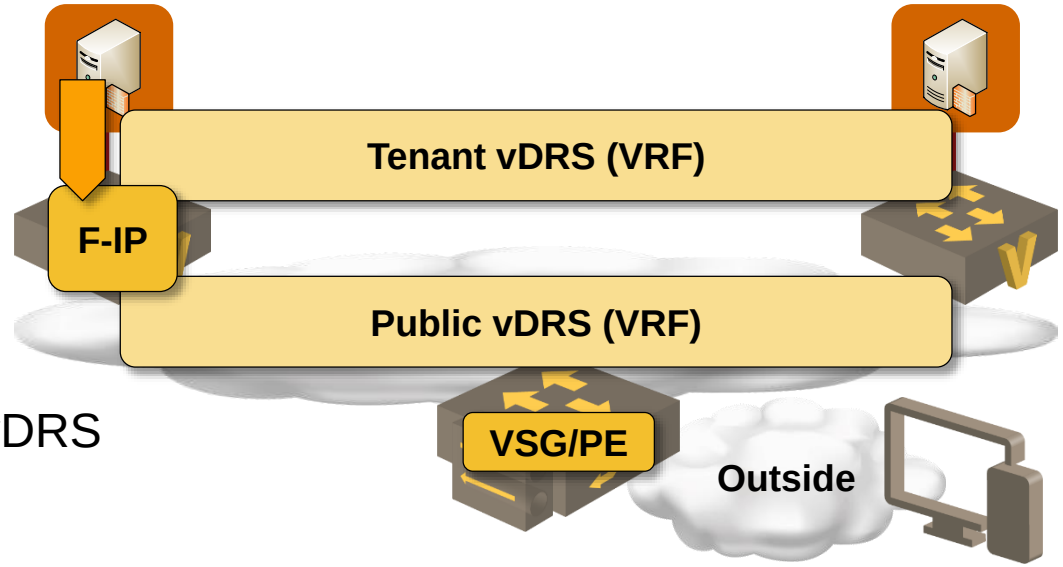
- Packet sent to IP address in public vDRS (received by hypervisor)
- Hypervisor translates destination IP address to VM IP address

Inside-to-outside

Floating IP Addresses in Nuage VSP

Setup

- Floating IP from public vDRS is allocated to a tenant VM
- 1:1 NAT rule is created on the hypervisor



Internal communication

- Destination IP address is within tenant vDRS
- NAT rule is not invoked

Outside-to-inside

- Packet sent to IP address in public vDRS (received by hypervisor)
- Hypervisor translates destination IP address to VM IP address

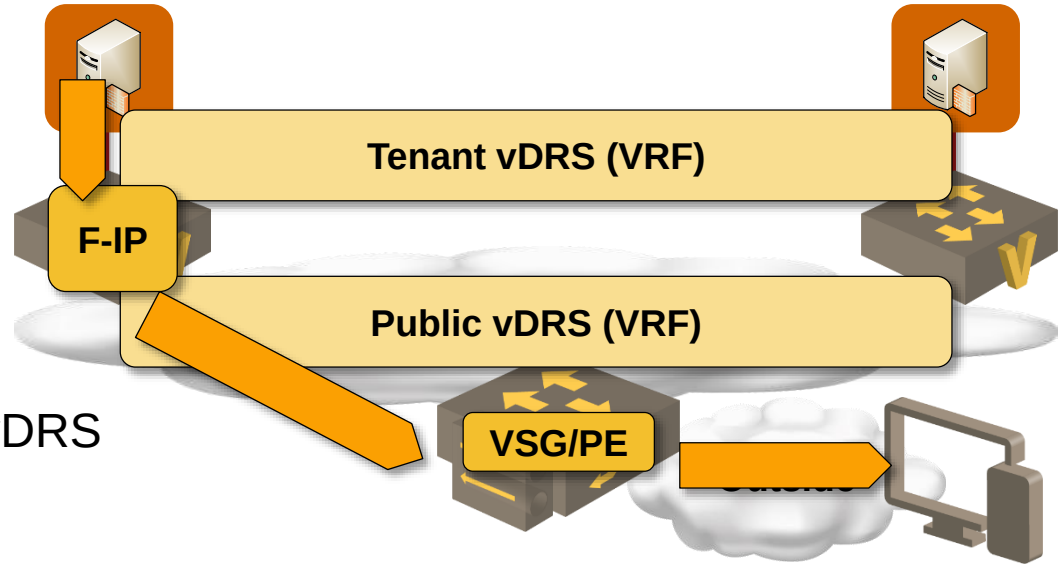
Inside-to-outside

- VM sends packet to a destination unreachable in tenant vDRS

Floating IP Addresses in Nuage VSP

Setup

- Floating IP from public vDRS is allocated to a tenant VM
- 1:1 NAT rule is created on the hypervisor



Internal communication

- Destination IP address is within tenant vDRS
- NAT rule is not invoked

Outside-to-inside

- Packet sent to IP address in public vDRS (received by hypervisor)
- Hypervisor translates destination IP address to VM IP address

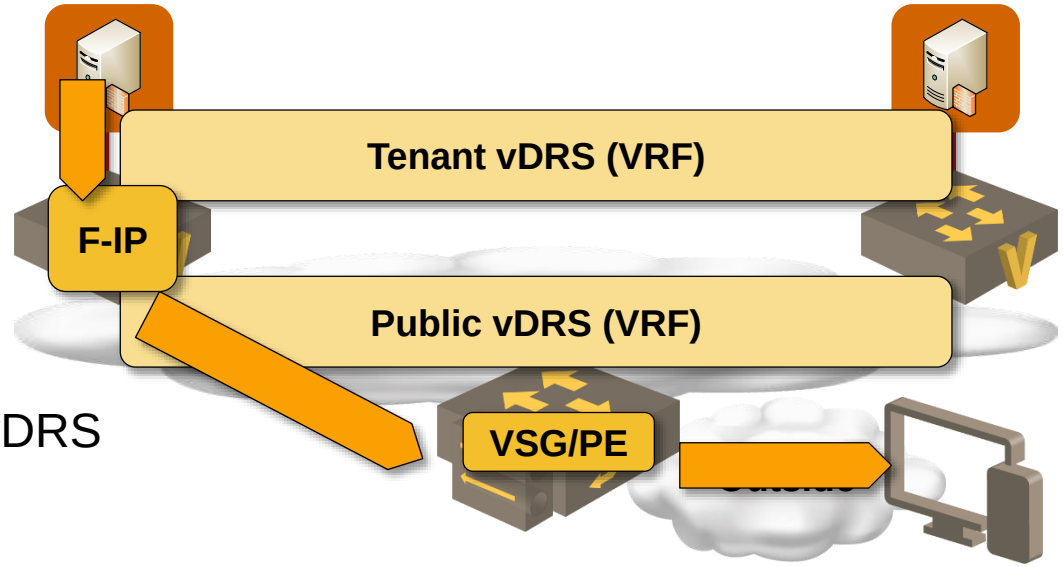
Inside-to-outside

- VM sends packet to a destination unreachable in tenant vDRS
- Per-VM default route pushes the packet through NAT rule into public vDRS

Floating IP Addresses in Nuage VSP

Setup

- Floating IP from public vDRS is allocated to a tenant VM
- 1:1 NAT rule is created on the hypervisor



Internal communication

- Destination IP address is within tenant vDRS
- NAT rule is not invoked

Outside-to-inside

- Packet sent to IP address in public vDRS (received by hypervisor)
- Hypervisor translates destination IP address to VM IP address

Inside-to-outside

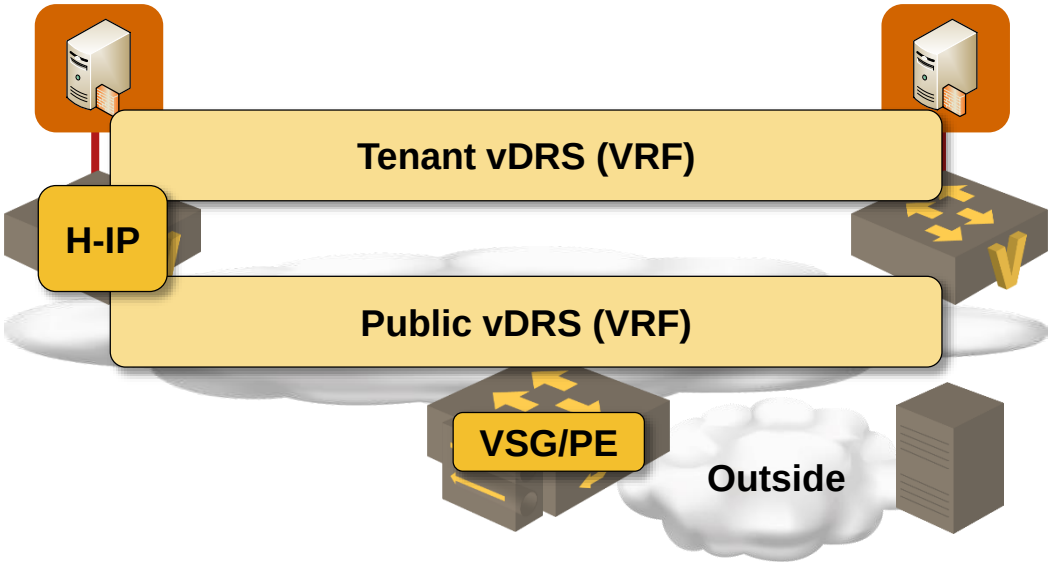
- VM sends packet to a destination unreachable in tenant vDRS
- Per-VM default route pushes the packet through NAT rule into public vDRS

NAT rule is stateless and active on a single hypervisor

Distributed Outbound PNAT on Nuage VSP

Setup

- IP from public vDRS (H-IP) is allocated to each hypervisor



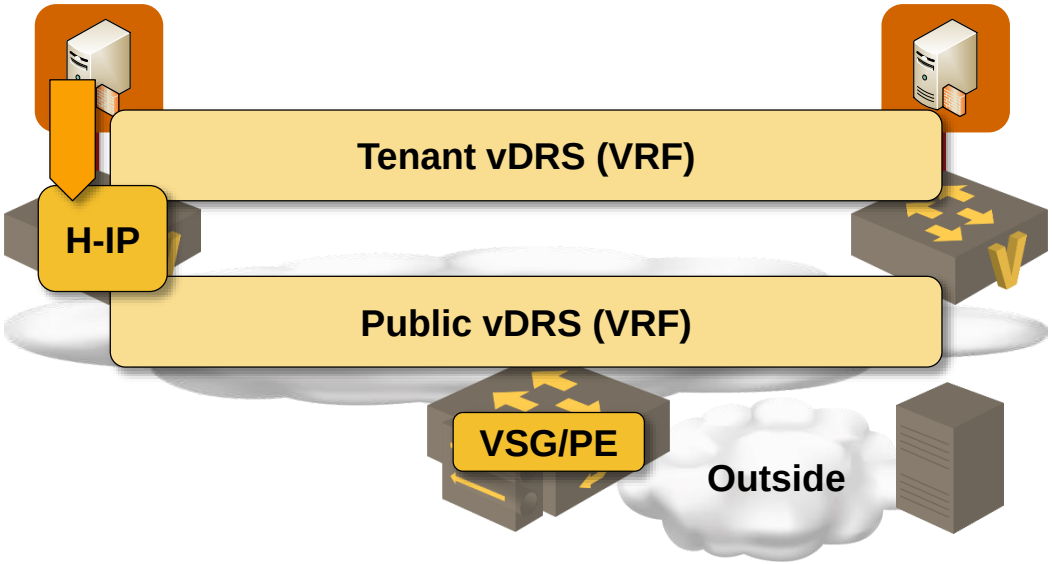
Distributed Outbound PNAT on Nuage VSP

Setup

- IP from public vDRS (H-IP) is allocated to each hypervisor

Inside-to-outside

- VM sends packet to a destination unreachable in tenant vDRS



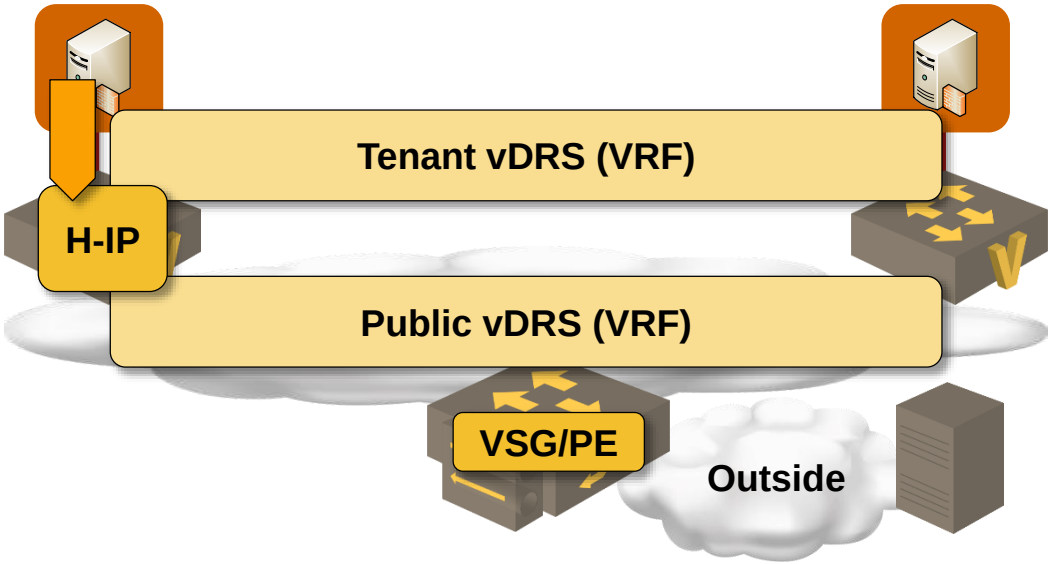
Distributed Outbound PNAT on Nuage VSP

Setup

- IP from public vDRS (H-IP) is allocated to each hypervisor

Inside-to-outside

- VM sends packet to a destination unreachable in tenant vDRS
- Default route pushes the packet through NAT rule into public vDRS



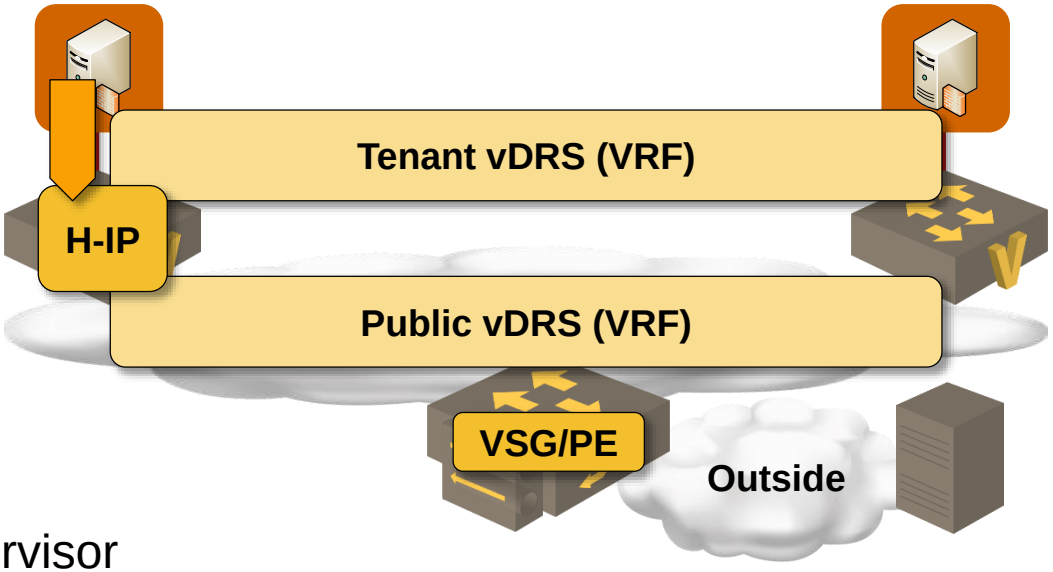
Distributed Outbound PNAT on Nuage VSP

Setup

- IP from public vDRS (H-IP) is allocated to each hypervisor

Inside-to-outside

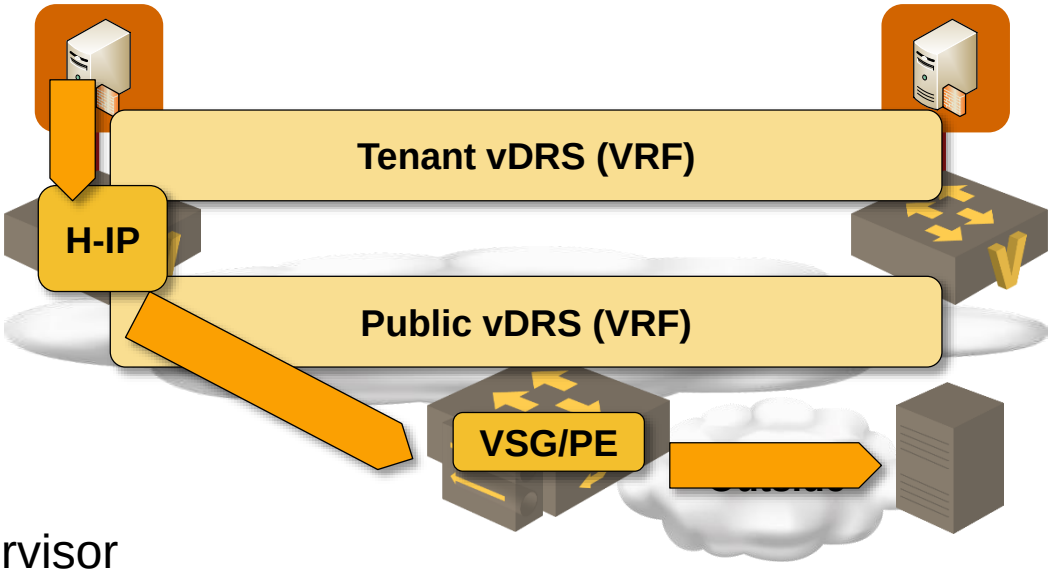
- VM sends packet to a destination unreachable in tenant vDRS
- Default route pushes the packet through NAT rule into public vDRS
- Stateful NAT entry is created in the hypervisor



Distributed Outbound PNAT on Nuage VSP

Setup

- IP from public vDRS (H-IP) is allocated to each hypervisor



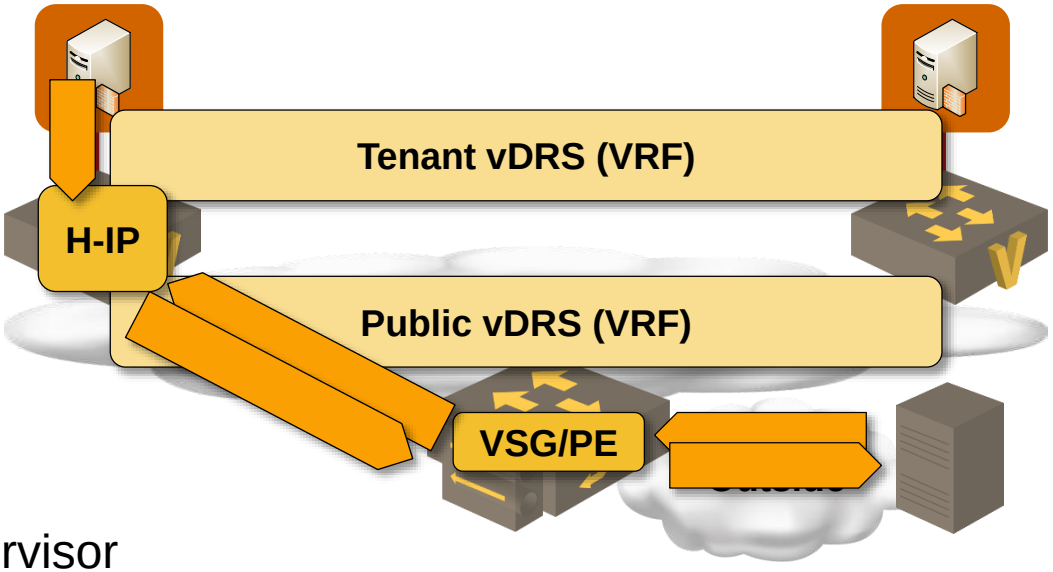
Inside-to-outside

- VM sends packet to a destination unreachable in tenant vDRS
- Default route pushes the packet through NAT rule into public vDRS
- Stateful NAT entry is created in the hypervisor
- Packet is delivered to the outside server

Distributed Outbound PNAT on Nuage VSP

Setup

- IP from public vDRS (H-IP) is allocated to each hypervisor



Inside-to-outside

- VM sends packet to a destination unreachable in tenant vDRS
- Default route pushes the packet through NAT rule into public vDRS
- Stateful NAT entry is created in the hypervisor
- Packet is delivered to the outside server

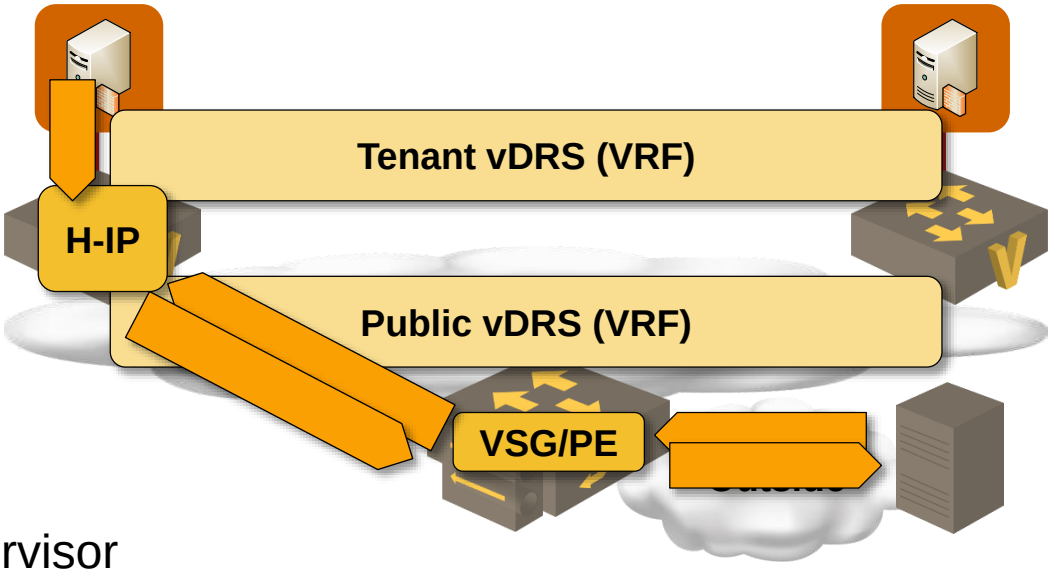
Outside-to-inside

- Return packet is sent to IP address in public vDRS (received by hypervisor)

Distributed Outbound PNAT on Nuage VSP

Setup

- IP from public vDRS (H-IP) is allocated to each hypervisor



Inside-to-outside

- VM sends packet to a destination unreachable in tenant vDRS
- Default route pushes the packet through NAT rule into public vDRS
- Stateful NAT entry is created in the hypervisor
- Packet is delivered to the outside server

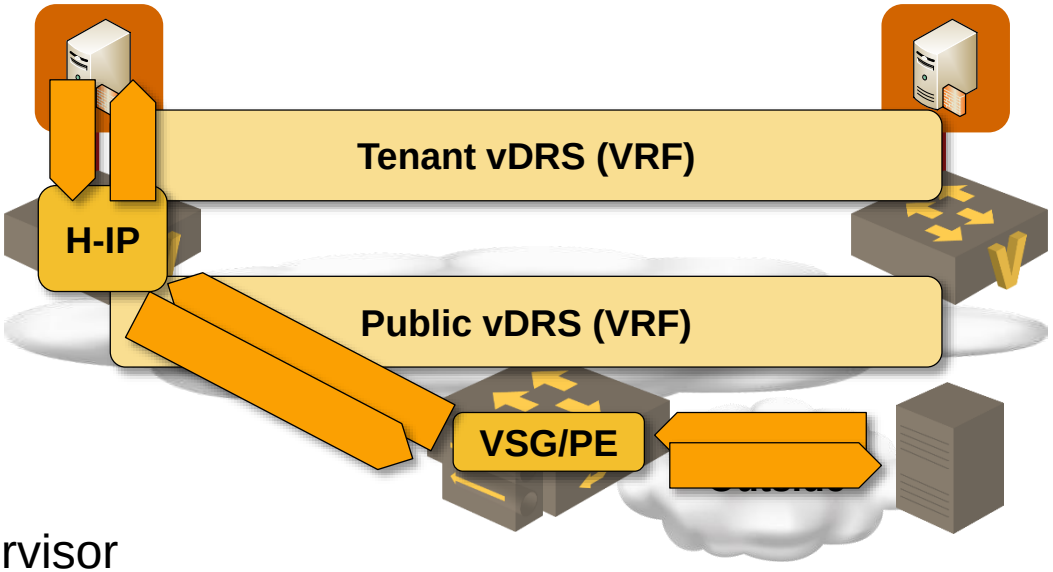
Outside-to-inside

- Return packet is sent to IP address in public vDRS (received by hypervisor)
- Hypervisor uses PNAT entry to translate destination IP address to VM IP address

Distributed Outbound PNAT on Nuage VSP

Setup

- IP from public vDRS (H-IP) is allocated to each hypervisor



Inside-to-outside

- VM sends packet to a destination unreachable in tenant vDRS
- Default route pushes the packet through NAT rule into public vDRS
- Stateful NAT entry is created in the hypervisor
- Packet is delivered to the outside server

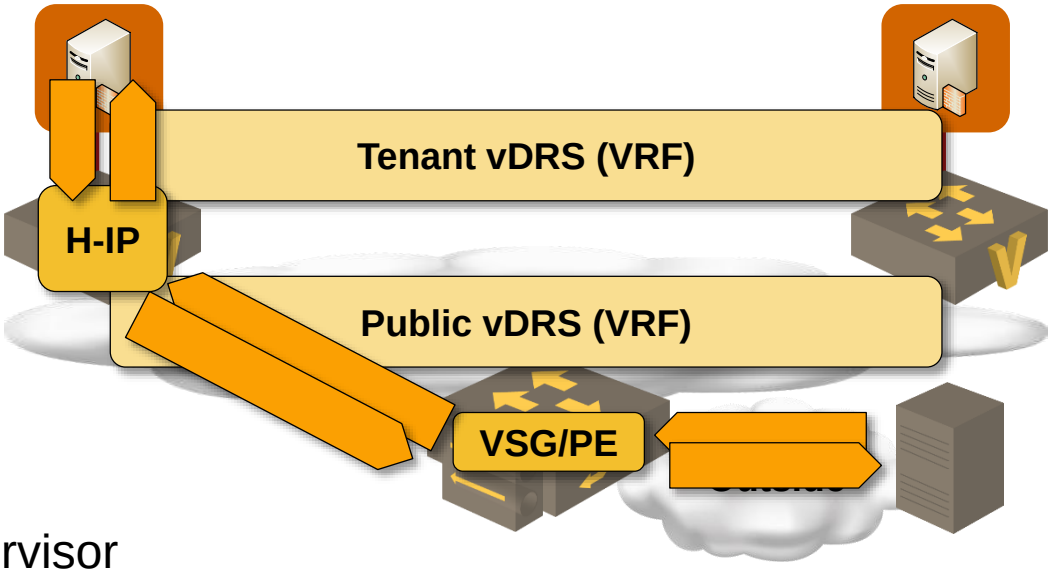
Outside-to-inside

- Return packet is sent to IP address in public vDRS (received by hypervisor)
- Hypervisor uses PNAT entry to translate destination IP address to VM IP address
- Translated packet is delivered to target VM

Distributed Outbound PNAT on Nuage VSP

Setup

- IP from public vDRS (H-IP) is allocated to each hypervisor



Inside-to-outside

- VM sends packet to a destination unreachable in tenant vDRS
- Default route pushes the packet through NAT rule into public vDRS
- Stateful NAT entry is created in the hypervisor
- Packet is delivered to the outside server

Outside-to-inside

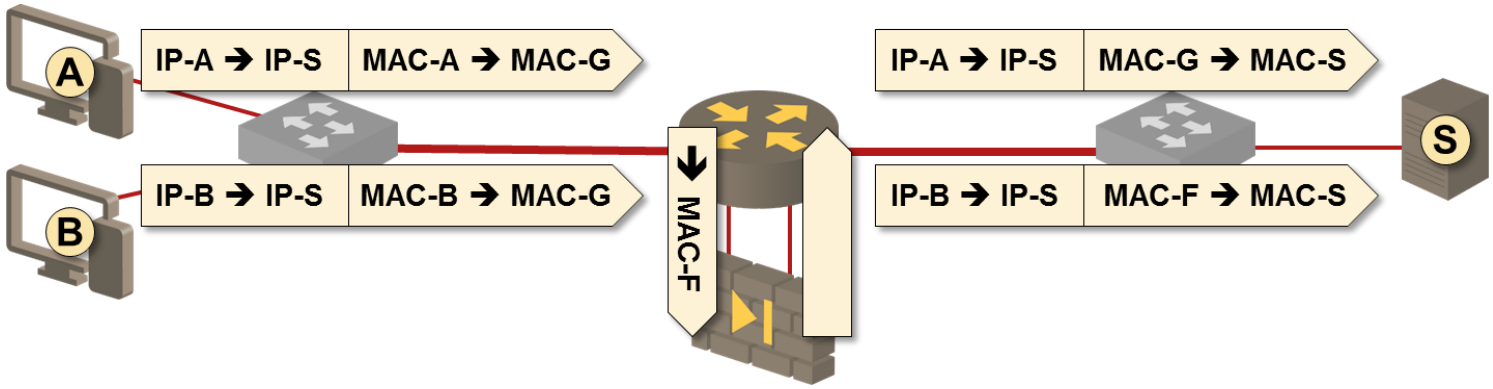
- Return packet is sent to IP address in public vDRS (received by hypervisor)
- Hypervisor uses PNAT entry to translate destination IP address to VM IP address
- Translated packet is delivered to target VM

The goal is connectivity, not specific NAT outside address

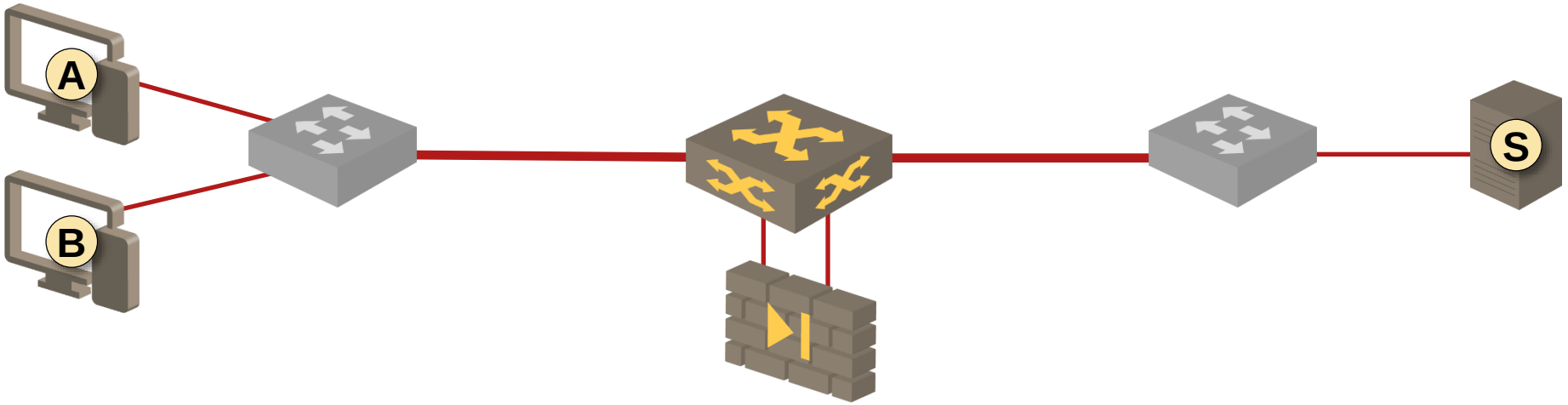
Service Chaining

Service Chaining Use Cases

- Insert physical appliances between virtual network endpoints
- Insert L4-7 and security services within a subnet
- Create multi-tier applications without routing overhead
- Combine multiple services in Network Function Virtualization deployments



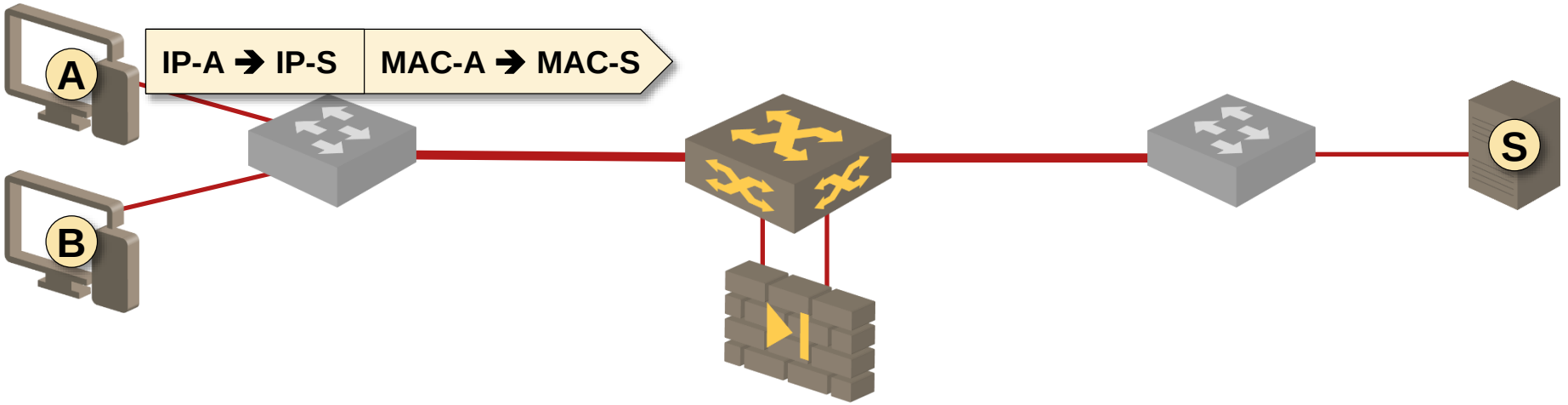
Layer-2 Service Insertion



Layer-2 frames redirected to a transparent (bump-in-wire) appliance

- Based on MAC (potentially IP) headers

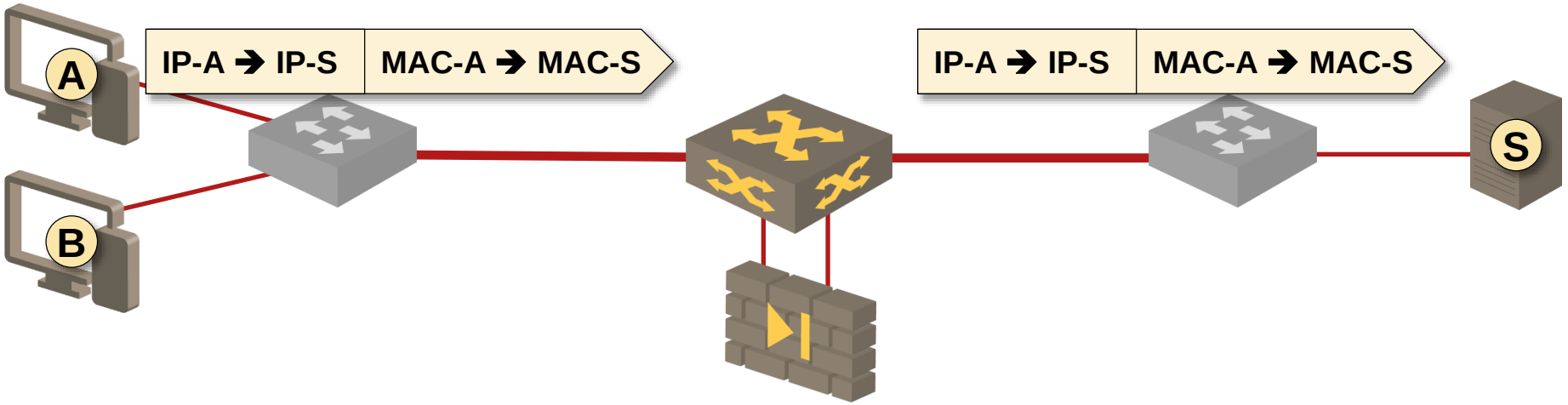
Layer-2 Service Insertion



Layer-2 frames redirected to a transparent (bump-in-wire) appliance

- Based on MAC (potentially IP) headers

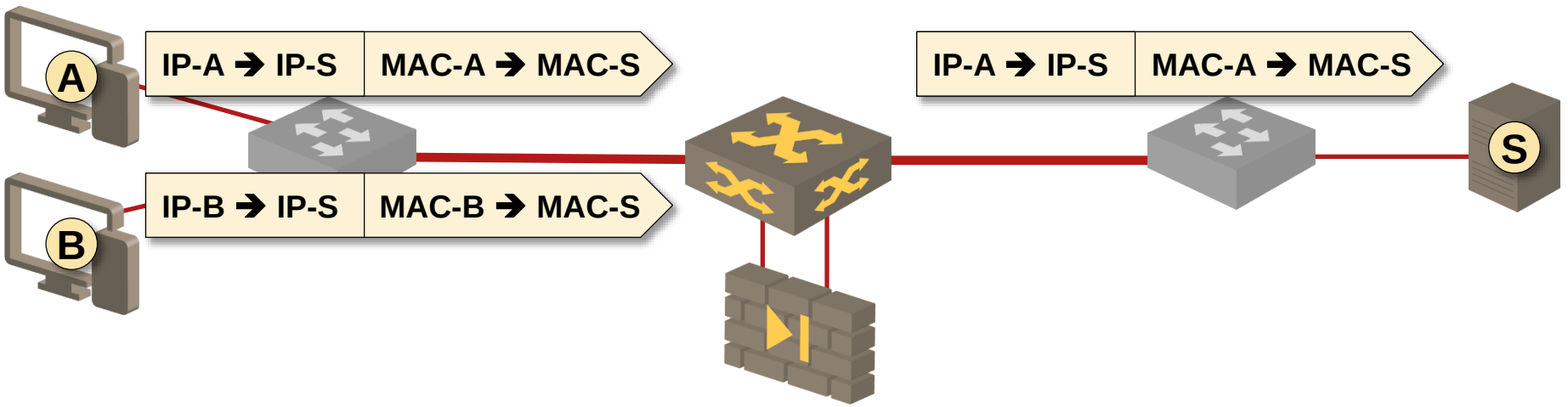
Layer-2 Service Insertion



Layer-2 frames redirected to a transparent (bump-in-wire) appliance

- Based on MAC (potentially IP) headers

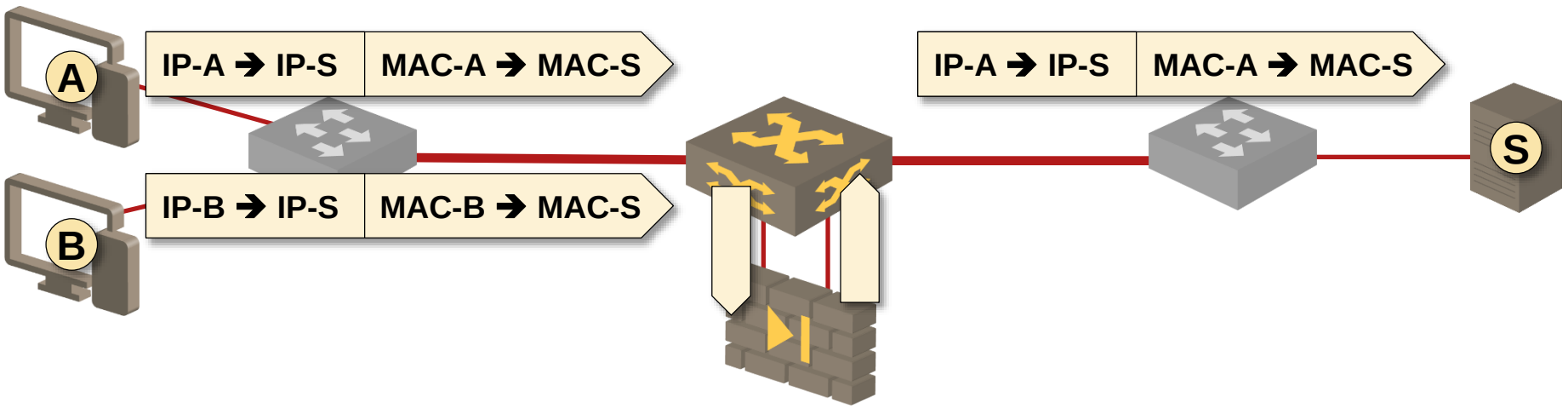
Layer-2 Service Insertion



Layer-2 frames redirected to a transparent (bump-in-wire) appliance

- Based on MAC (potentially IP) headers

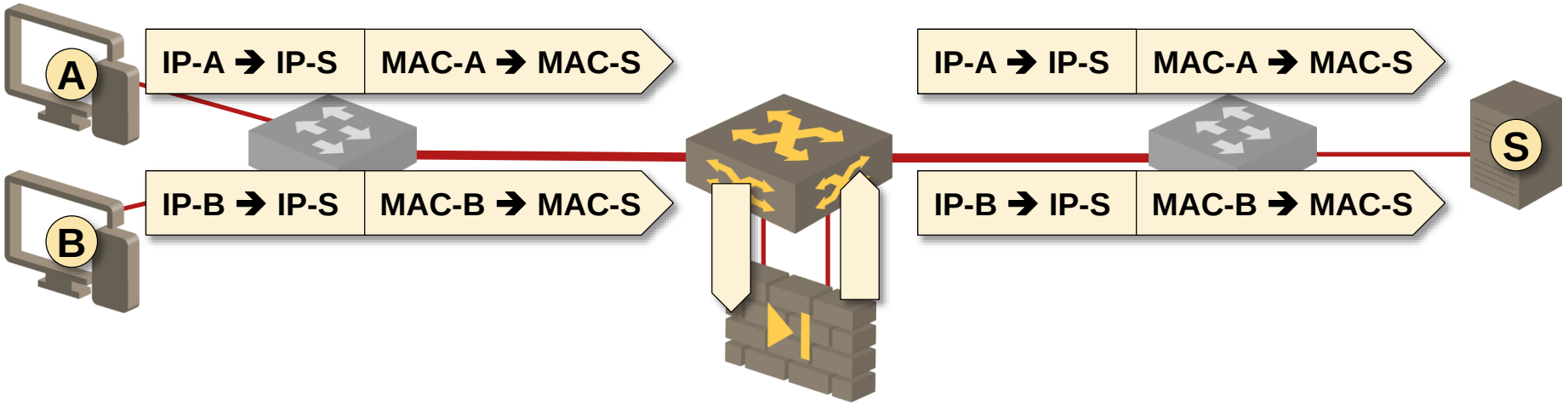
Layer-2 Service Insertion



Layer-2 frames redirected to a transparent (bump-in-wire) appliance

- Based on MAC (potentially IP) headers

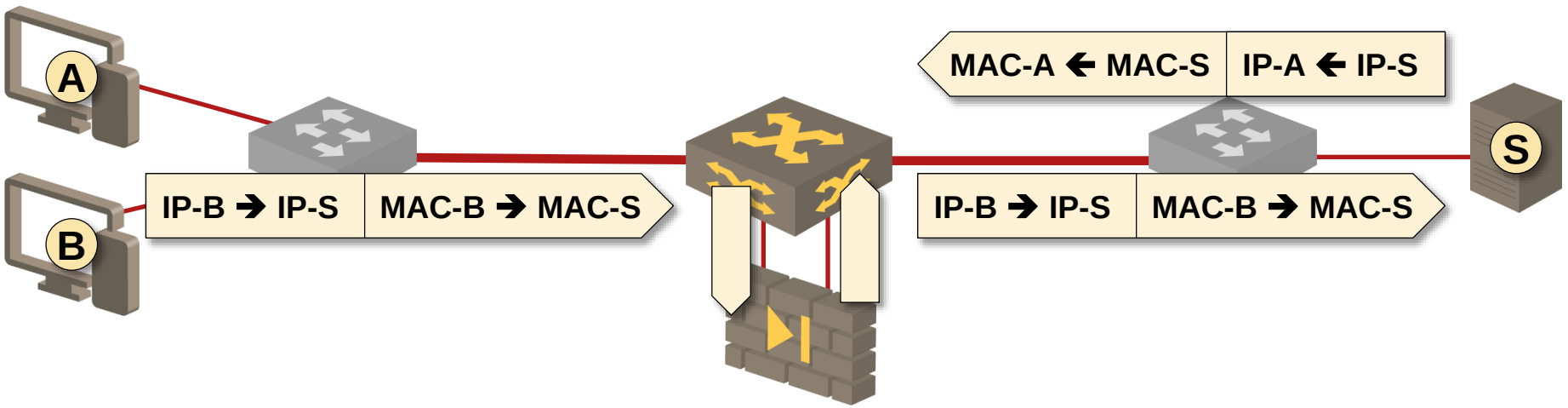
Layer-2 Service Insertion



Layer-2 frames redirected to a transparent (bump-in-wire) appliance

- Based on MAC (potentially IP) headers

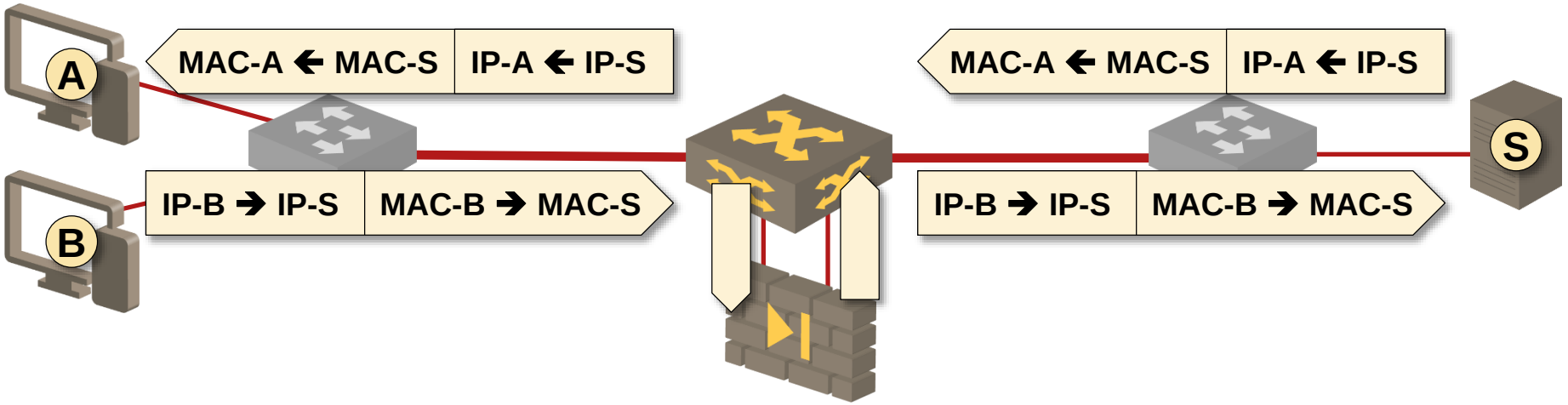
Layer-2 Service Insertion



Layer-2 frames redirected to a transparent (bump-in-wire) appliance

- Based on MAC (potentially IP) headers

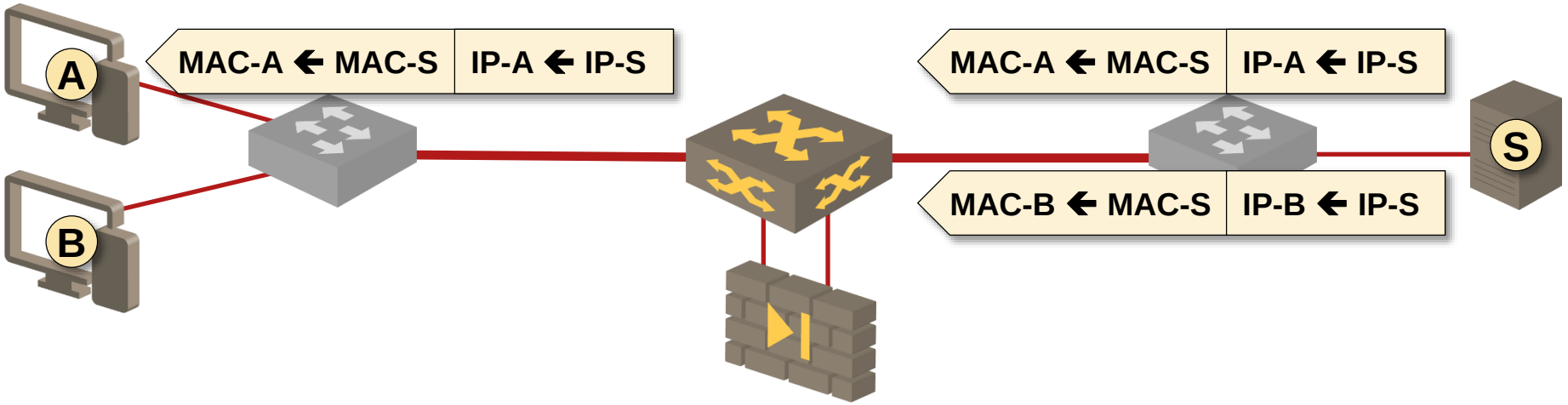
Layer-2 Service Insertion



Layer-2 frames redirected to a transparent (bump-in-wire) appliance

- Based on MAC (potentially IP) headers

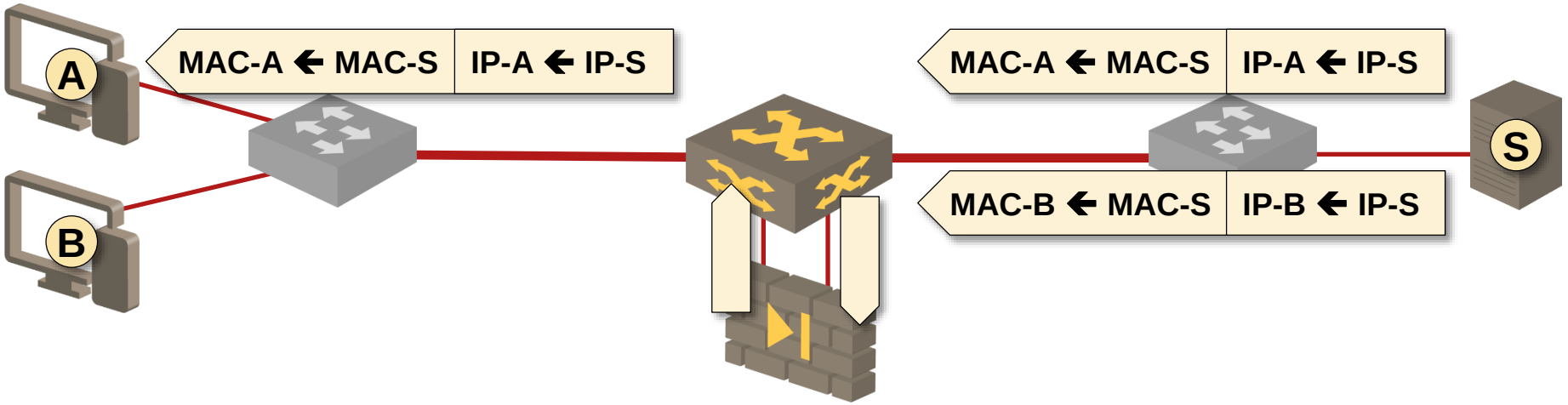
Layer-2 Service Insertion



Layer-2 frames redirected to a transparent (bump-in-wire) appliance

- Based on MAC (potentially IP) headers

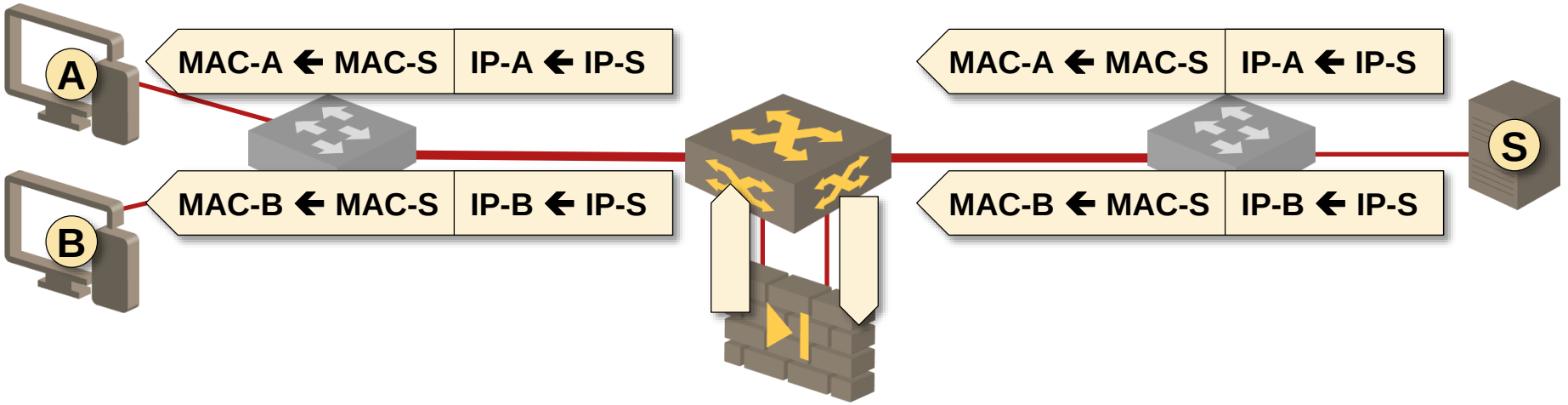
Layer-2 Service Insertion



Layer-2 frames redirected to a transparent (bump-in-wire) appliance

- Based on MAC (potentially IP) headers

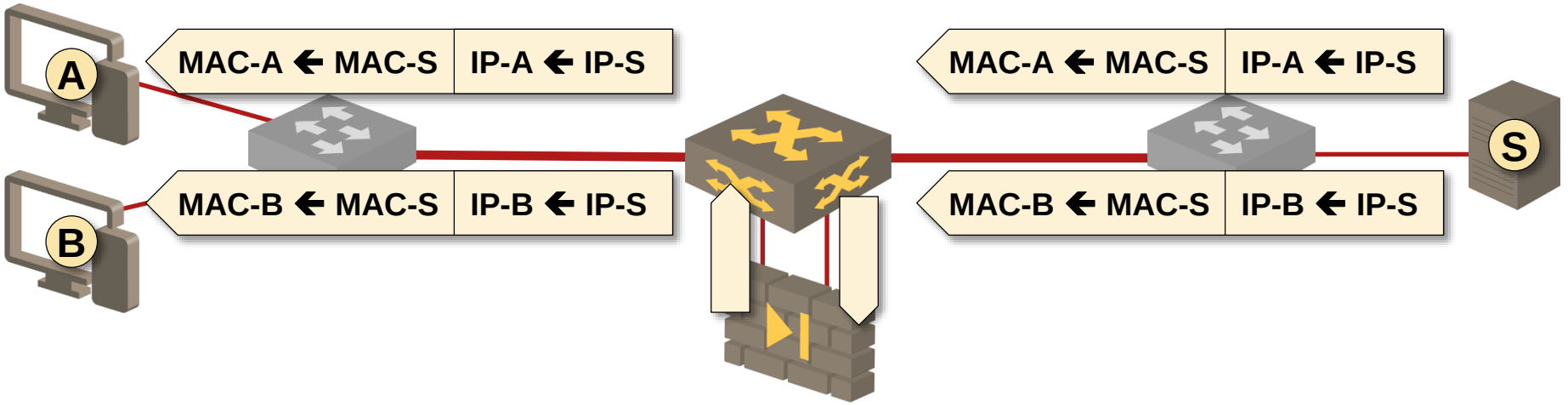
Layer-2 Service Insertion



Layer-2 frames redirected to a transparent (bump-in-wire) appliance

- Based on MAC (potentially IP) headers

Layer-2 Service Insertion



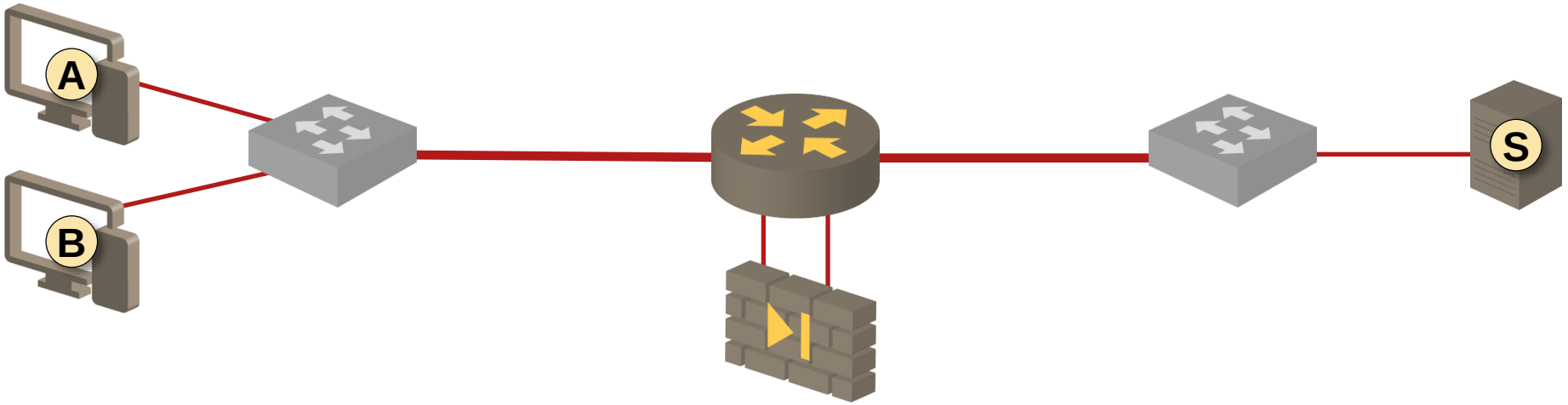
Layer-2 frames redirected to a transparent (bump-in-wire) appliance

- Based on MAC (potentially IP) headers

Typical implementation

- VLAN chaining
- Hard to implement for individual endpoints
- Impossible to implement for individual applications
- Fantastic potential for forwarding loops

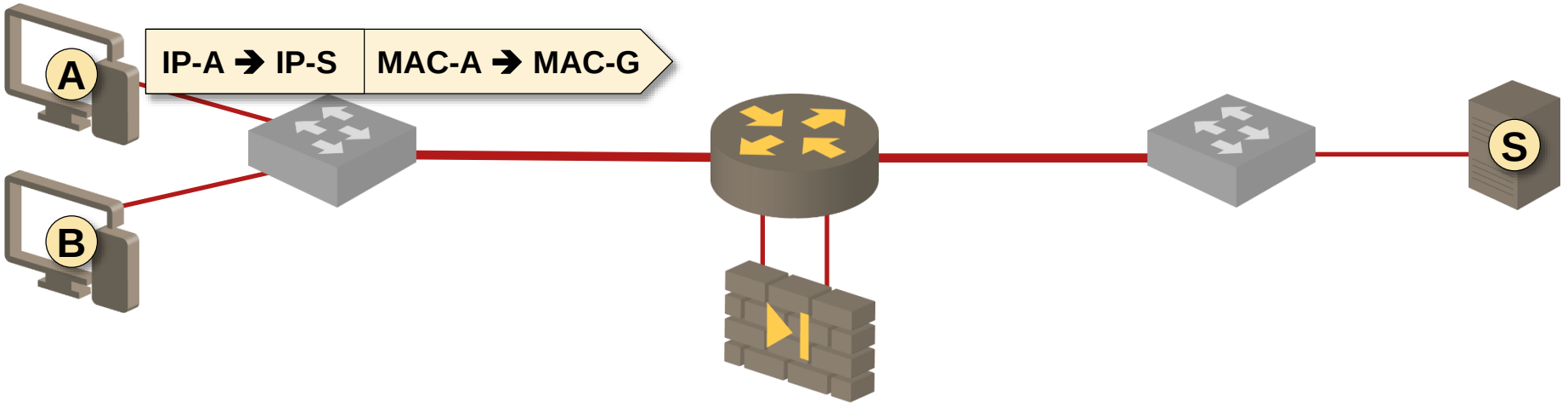
Layer-3 Service Insertion



Layer-3 frames redirected to a transparent or inter-subnet appliance

- Based on IP headers
- Might require MAC header rewrite

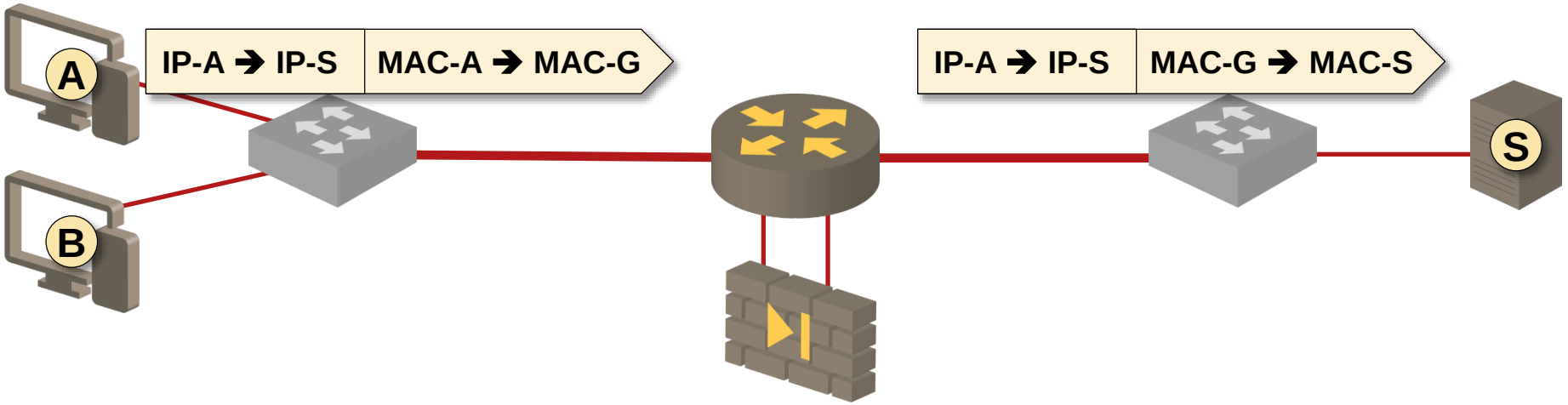
Layer-3 Service Insertion



Layer-3 frames redirected to a transparent or inter-subnet appliance

- Based on IP headers
- Might require MAC header rewrite

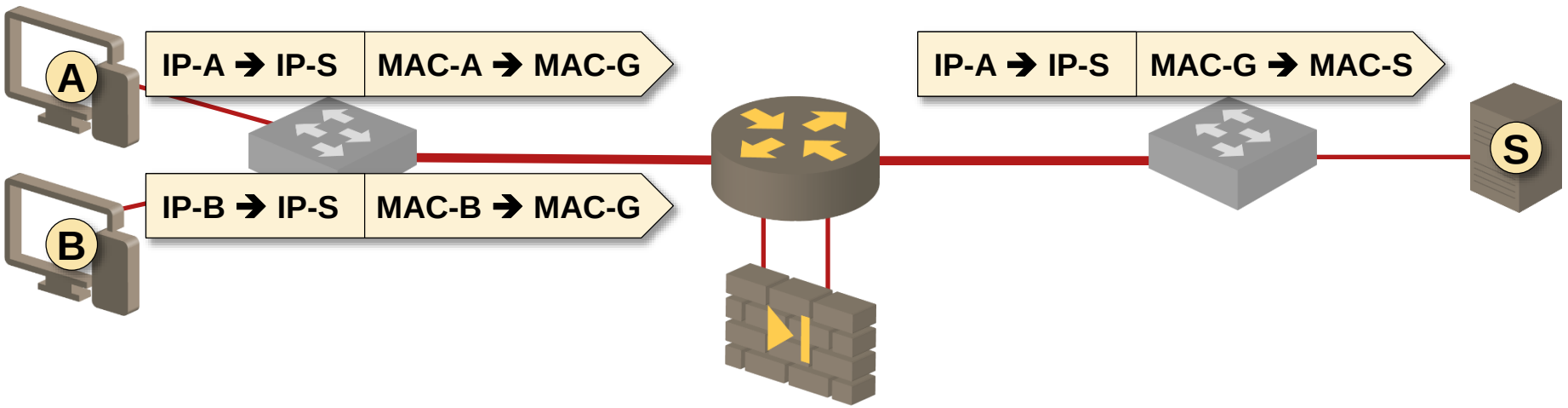
Layer-3 Service Insertion



Layer-3 frames redirected to a transparent or inter-subnet appliance

- Based on IP headers
- Might require MAC header rewrite

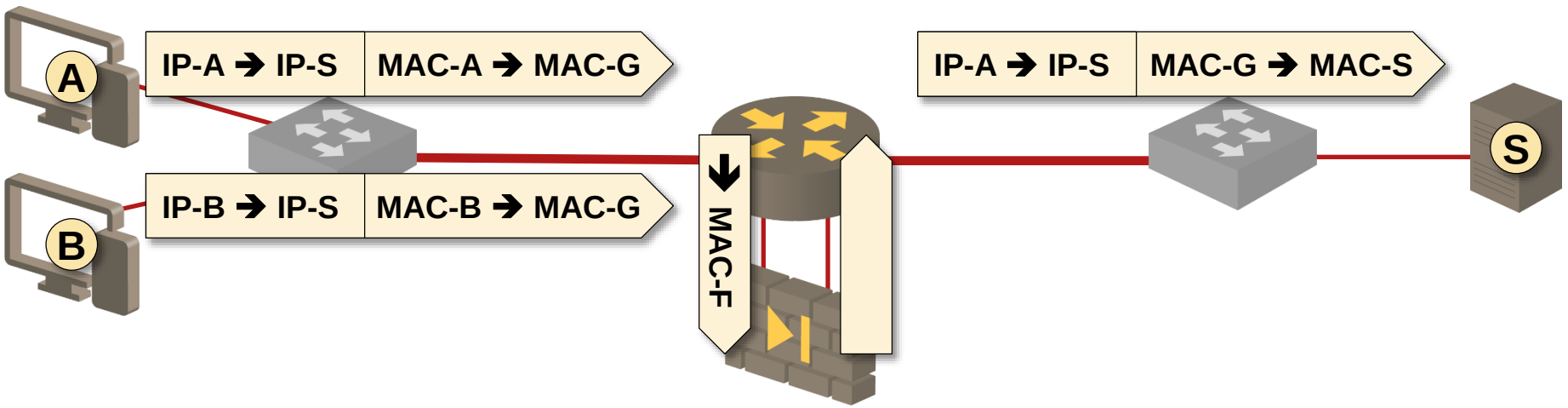
Layer-3 Service Insertion



Layer-3 frames redirected to a transparent or inter-subnet appliance

- Based on IP headers
- Might require MAC header rewrite

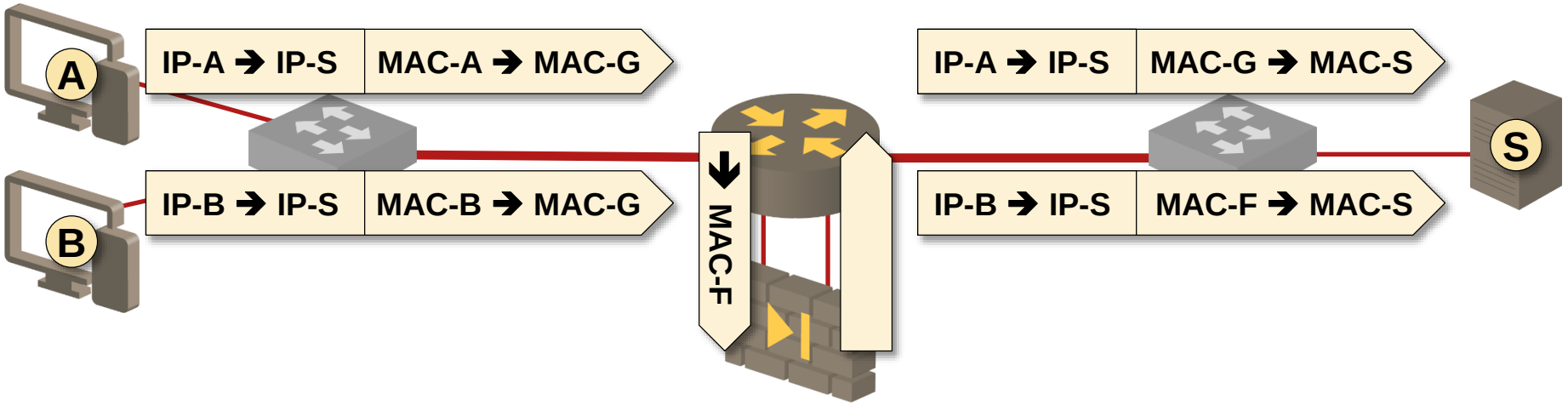
Layer-3 Service Insertion



Layer-3 frames redirected to a transparent or inter-subnet appliance

- Based on IP headers
- Might require MAC header rewrite

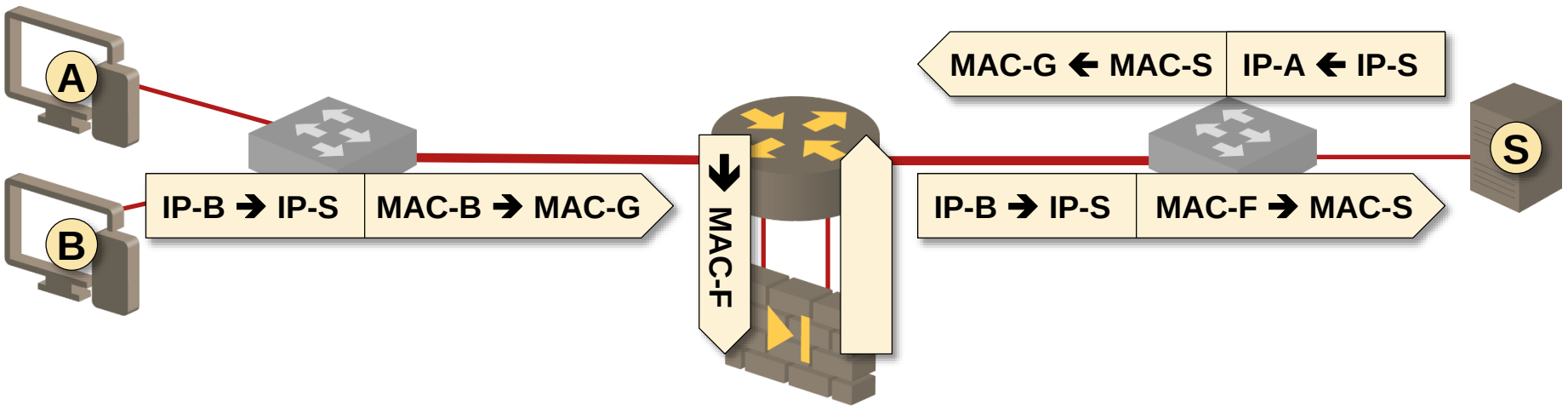
Layer-3 Service Insertion



Layer-3 frames redirected to a transparent or inter-subnet appliance

- Based on IP headers
- Might require MAC header rewrite

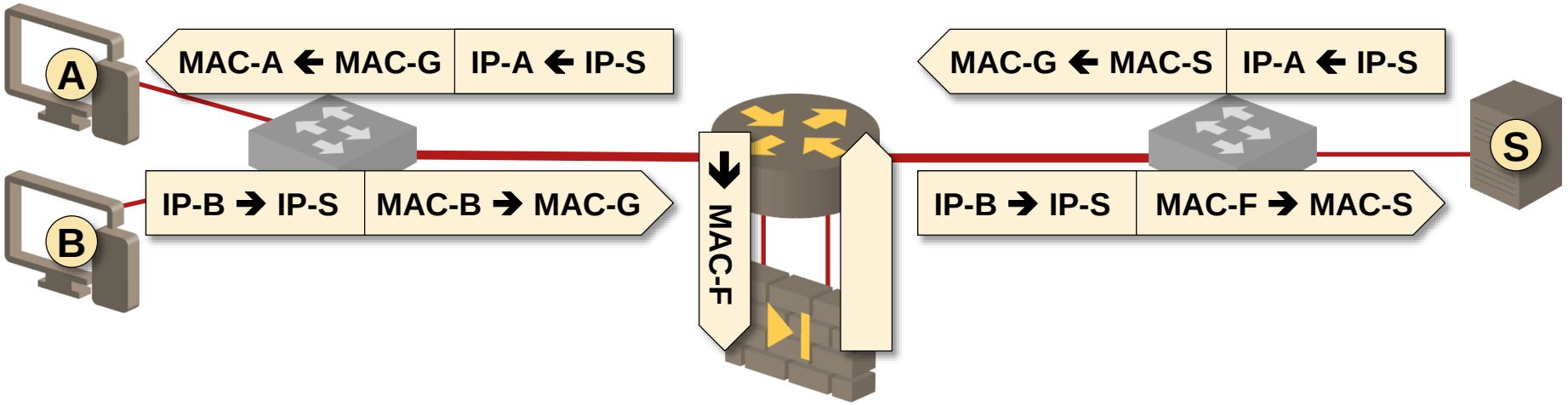
Layer-3 Service Insertion



Layer-3 frames redirected to a transparent or inter-subnet appliance

- Based on IP headers
- Might require MAC header rewrite

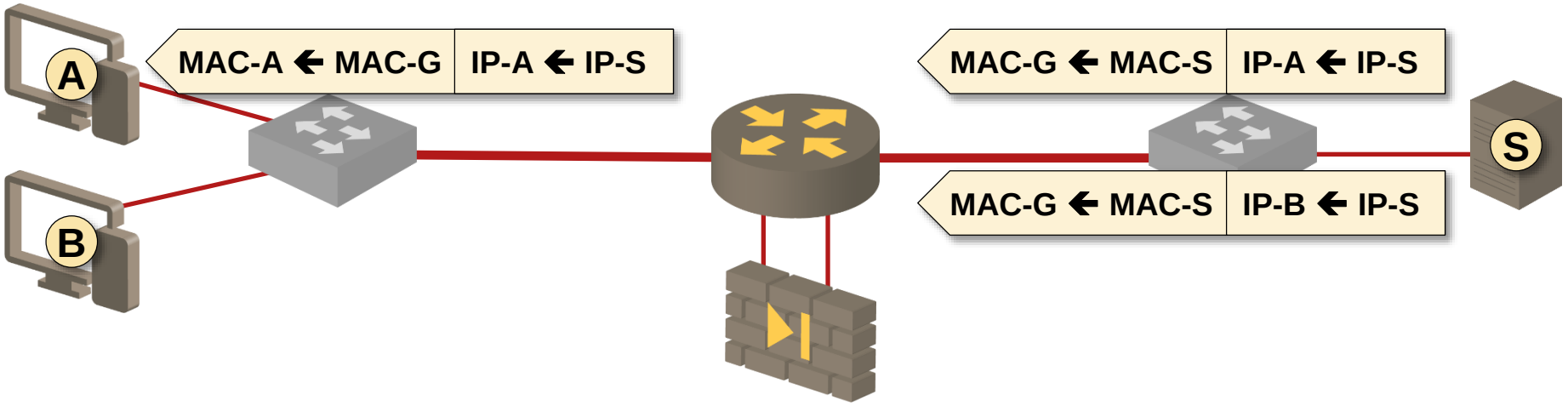
Layer-3 Service Insertion



Layer-3 frames redirected to a transparent or inter-subnet appliance

- Based on IP headers
- Might require MAC header rewrite

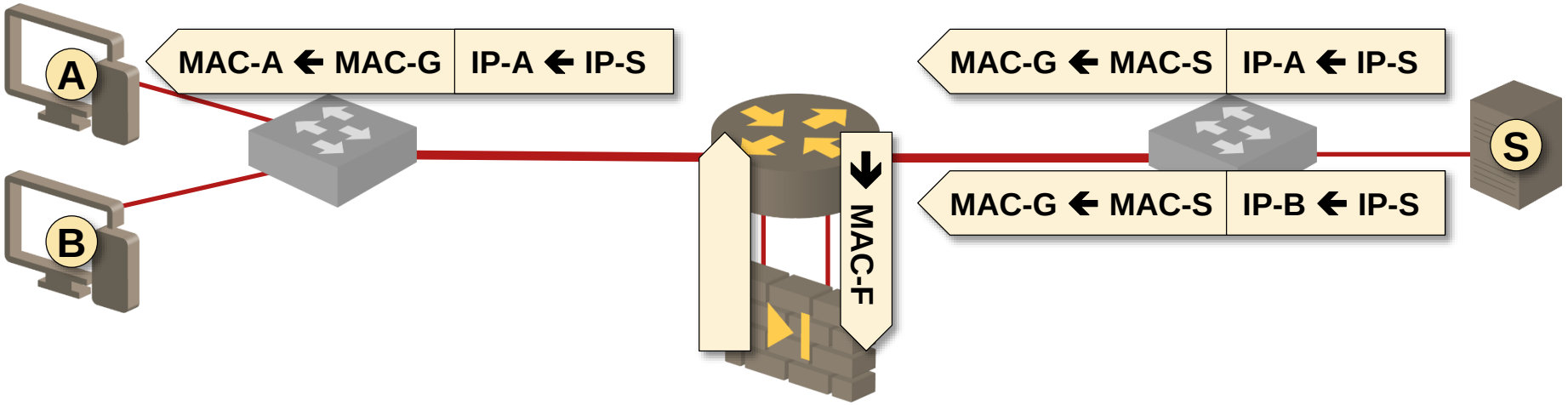
Layer-3 Service Insertion



Layer-3 frames redirected to a transparent or inter-subnet appliance

- Based on IP headers
- Might require MAC header rewrite

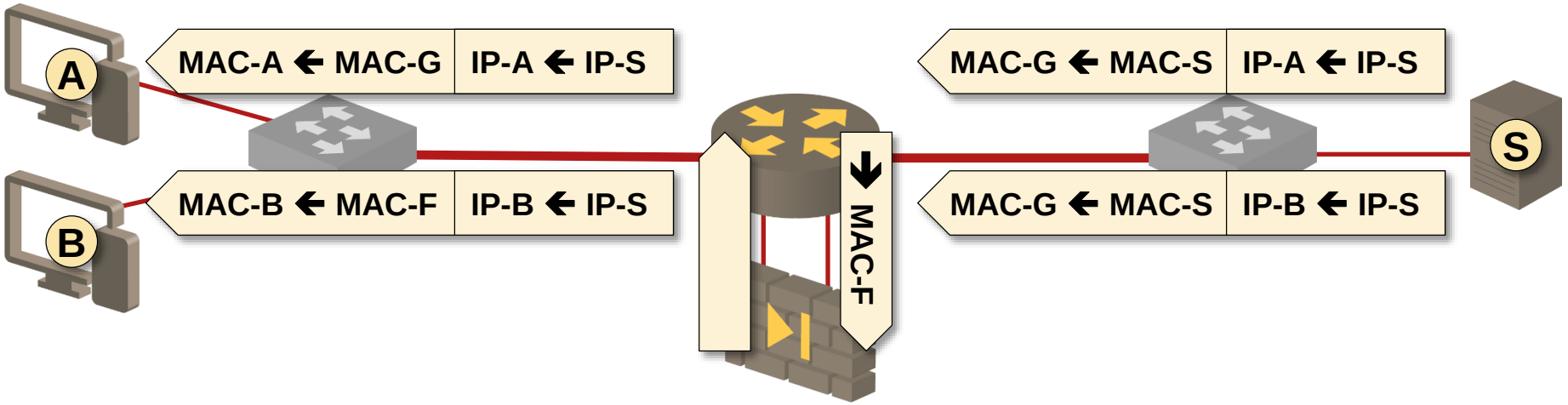
Layer-3 Service Insertion



Layer-3 frames redirected to a transparent or inter-subnet appliance

- Based on IP headers
- Might require MAC header rewrite

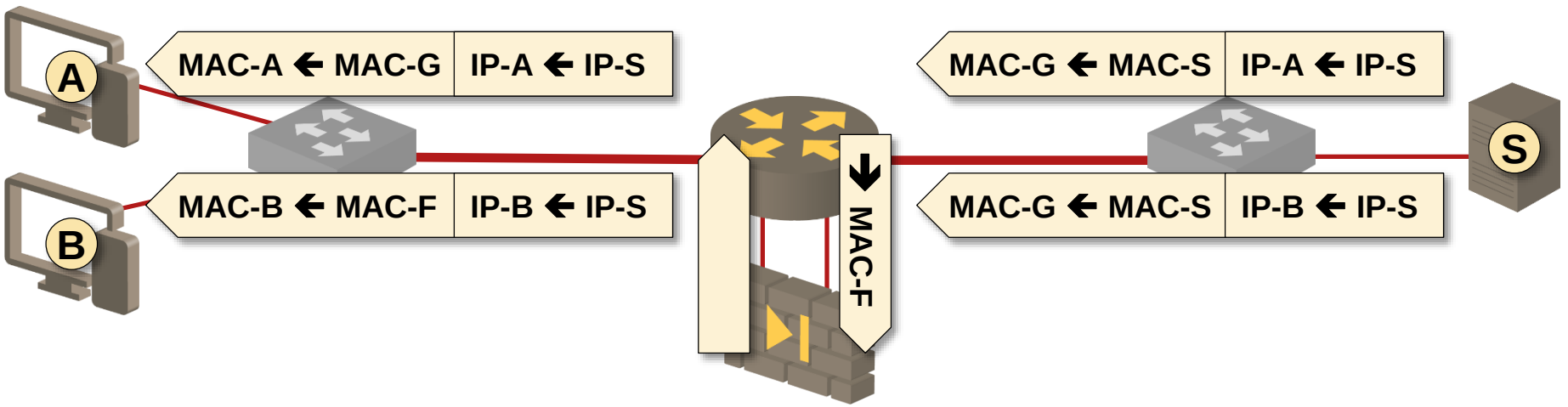
Layer-3 Service Insertion



Layer-3 frames redirected to a transparent or inter-subnet appliance

- Based on IP headers
- Might require MAC header rewrite

Layer-3 Service Insertion



Layer-3 frames redirected to a transparent or inter-subnet appliance

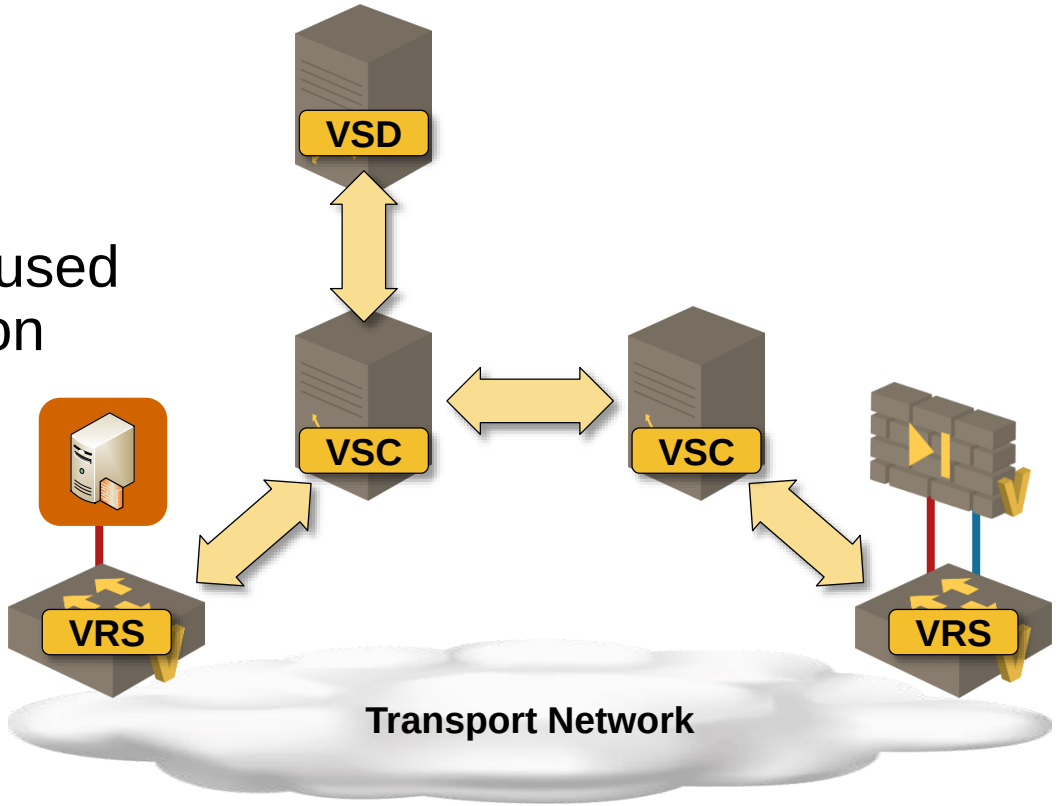
- Based on IP headers
- Might require MAC header rewrite

Typical implementation

- Policy-based routing (PBR)
- MAC rewrite is automatic
- Hard to implement for appliances not close to the forwarding path

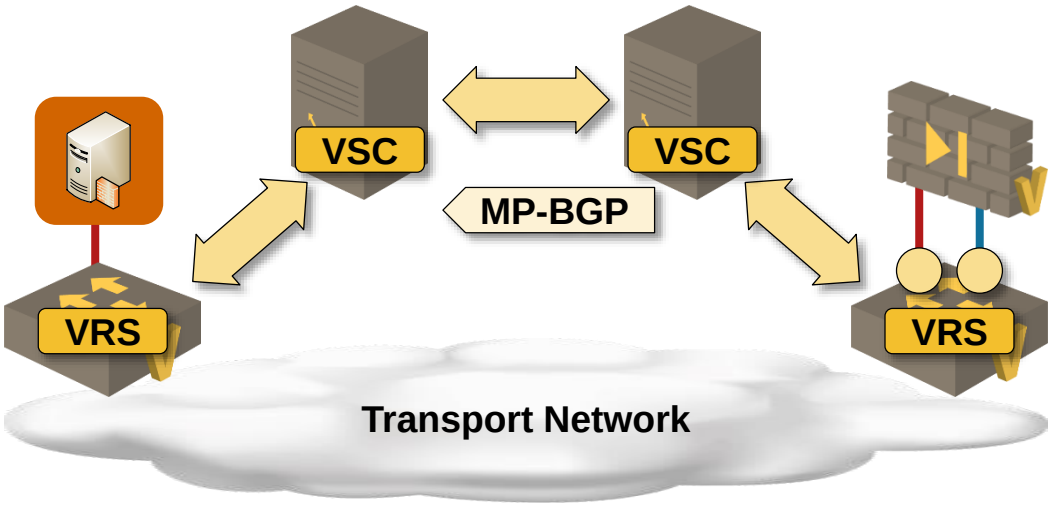
Service Chaining Setup in Nuage VSP

- Services and redirection (chaining) rules are defined in VSD Architect
- VSD downloads redirection rules to VSC
- VSC instantiates PBR entries on virtual port (VM) activation
- Traffic redirection uses the same scalability mechanisms as security groups
- Multiple forwarding domains are used to further scale the implementation



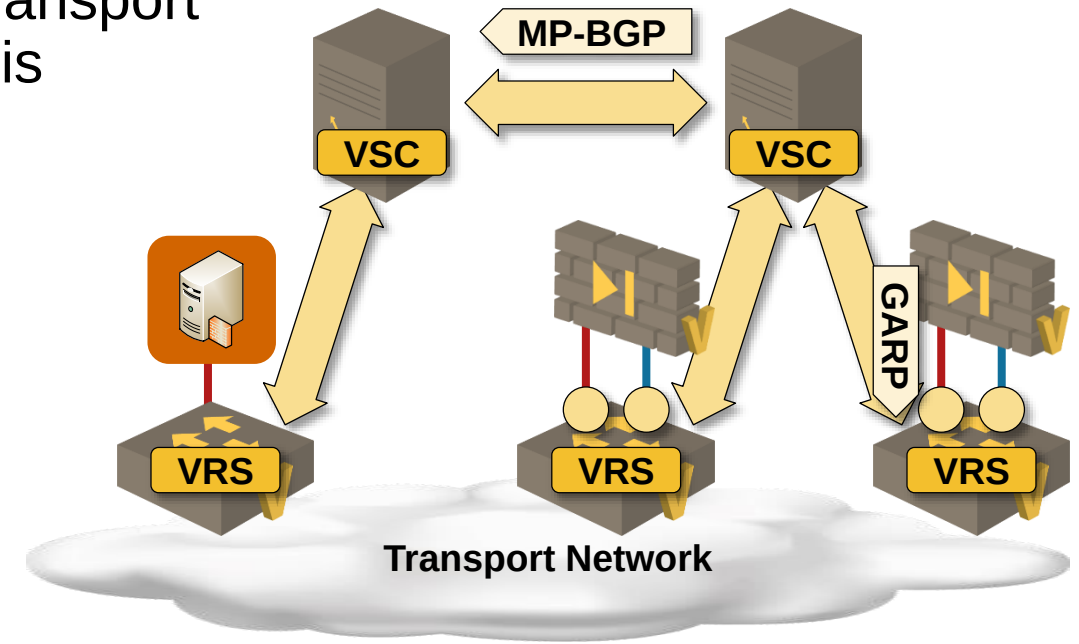
Service Chaining in Nuage VSP: Control and Data Plane

- Appliances (physical or virtual) are identified by virtual port tags
- A dedicated VNI (VXLAN segment) is allocated to each appliance port
- Appliance reachability information (ESI, VNI, transport next hop) is propagated in EVPN updates
- Information from EVPN update is used as PBR next hop



Service Chaining in Nuage VSP: Appliance Resiliency

- Appliances (physical or virtual) are identified by virtual port tags
- A dedicated VNI (VXLAN segment) is allocated to each appliance port
- L2VPN is create between appliance
- Active appliance IP address is detected by monitoring GARP packets
- A host route is created for each appliance IP address
- L3VPN host route (prefix, VNI, transport next hop) toward appliance port is propagated across MP-BGP routing domain
- Information from L3VPN route is used as PBR next hop



Conclusions

Elements of a Scalable Overlay Virtual Networking Solution

Architectural elements:

- Distributed forwarding plane (L2 and L3)
- Control plane with scale-out architecture
- Distributed L4 services (security, NAT)
- Scalable security mechanisms

Additional considerations:

- High-performance gateways
- Control- and management-plane integration with external networks

Designing Your Cloud Infrastructure

- Define the services
- Define the virtual infrastructure requirements
  - Connectivity (L2 and/or L3)
  - Security
  - Performance
  - Integration with legacy infrastructure
  - Integration with WAN networks
- Select the orchestration system
- Select the hypervisor platform
- Select an overlay virtual networking solution that will support the services you want to offer
  - Easy integration with the orchestration system
  - Scalable implementation of network services
  - Scalable integration with external networks

Questions?

Send them to ip@ipSpace.net or @ioshints