Building a Unified Hybrid Cloud Networking Using OVN

Ankur Kumar Sharma
Manoj Sharma
Nutanix Inc.
OUTLINE

Why
What
How
Future Work
Why?
Nutanix Hybrid Cloud

○ Xi Leap
○ Xi Frame
○ Xi Beam
○ Xi Epoch
○ Xi IoT
Disaster recovery as a service (DRaaS)

On-prem VMs replicated to Xi

Extension of Nutanix enterprise cloud platform
Xi Leap

Primary Site

Xi Cloud

Leap DR

Prism

Enterprise Cloud OS

Live VMs

Inactive VMs
What?
VPC 1
- Subnet
- Subnet

VPC 2
- Subnet
- Subnet

L3 VPN

Direct Connect
UNIFIED SDN STACK

VPC 1
Subnet
Subnet
router
L3 VPN
Direct Connect

VPC 2
Subnet
Subnet
router
VPC 1
- Subnet
- Subnet

VPC 2
- Subnet
- Subnet

OVN

L3 VPN

Direct Connect

Enterprise Cloud OS

WHY

WHAT
How?
ON PREM

○ Implement VLAN backed Logical Routing
  ■ On prem uses vlan backed networking.

○ Encapsulation may not be desirable
  ■ Hardware may not be capable of supporting offloads.
  ■ Avoid MTU reconfiguration.
VLAN BACKED LOGICAL ROUTING

VLAN X

RP_X

RP_Y

VLAN Y

VLAN X

SRC_MAC = RP_Y

VLAN Y

SRC_MAC = RP_Y

RPC_X

RPC_Y

VLAN X
VLAN BACKED LOGICAL ROUTING (Issues)

VLAN X

VLAN Y

SRC_MAC = RP_Y

SRC_MAC = RP_Y

SRC_MAC = RP_Y
VLAN BACKED LOGICAL ROUTING (Solution)

VLAN X

SRC_MAC = CHASSIS_MAC_1

VLAN Y

SRC_MAC = RP_Y

SRC_MAC = CHASSIS_MAC_2
NAT

VLAN X

RP_X

router

RP_EXT

EXTERNAL SUBNET

AYER 3 TUNNEL (GENEVE)

VPC NAT gateway

Enterprise Cloud OS
NAT (Issues)

Tunneling requires:

- NIC offload on hypervisor side
- Increased MTU to accommodate encapsulation header
NAT (Solution)

VLAN X

RP_X

router

RP_EXT

EXTERNAL SUBNET

LAYER 2
(DST_MAC = RP_EXT)

VPC NAT gateway
Xi (Goals)

- Scaled deployment
  - Number of Hypervisors
  - Number of VMs, Containers

- High Availability

- Self Healing
Key Design Decisions

○ Overlay tunnels for multi-tenancy

○ OVN supported encapsulations - Geneve, STT, VxLAN
  ■ Geneve for east - west
  ■ VxLAN for north - south (?)
  ■ MPLS over GRE for north - south

○ Forwarding decisions done locally
External Routers

Internet

Underlay Network

AHV

AHV
Failure Detection

- Convergence time depends on BFD timer settings alone
- Repair done locally, without control plane involvement
FUTURE WORK
EXTENDING SUBNET ACROSS SITES
DHCP RELAY

MULTIPLE GATEWAY CHASSIS SUPPORT
Questions