Distributed Virtual Routing for VLAN backed networks on OVN

Ankur Sharma

Nutanix Inc.

Outline

Introduction

Challenges

OVN Enhancements

Comparison with overlay DVR

Current Status

Future Work

INTRODUCTION

Distributed Virtual Router (DVR)

Router instance running in compute nodes (Hypervisors).

Provides functionality of gateway to virtual machines/containers.

Distributed in fashion.

- Same router instance runs on multiple hypervisors.

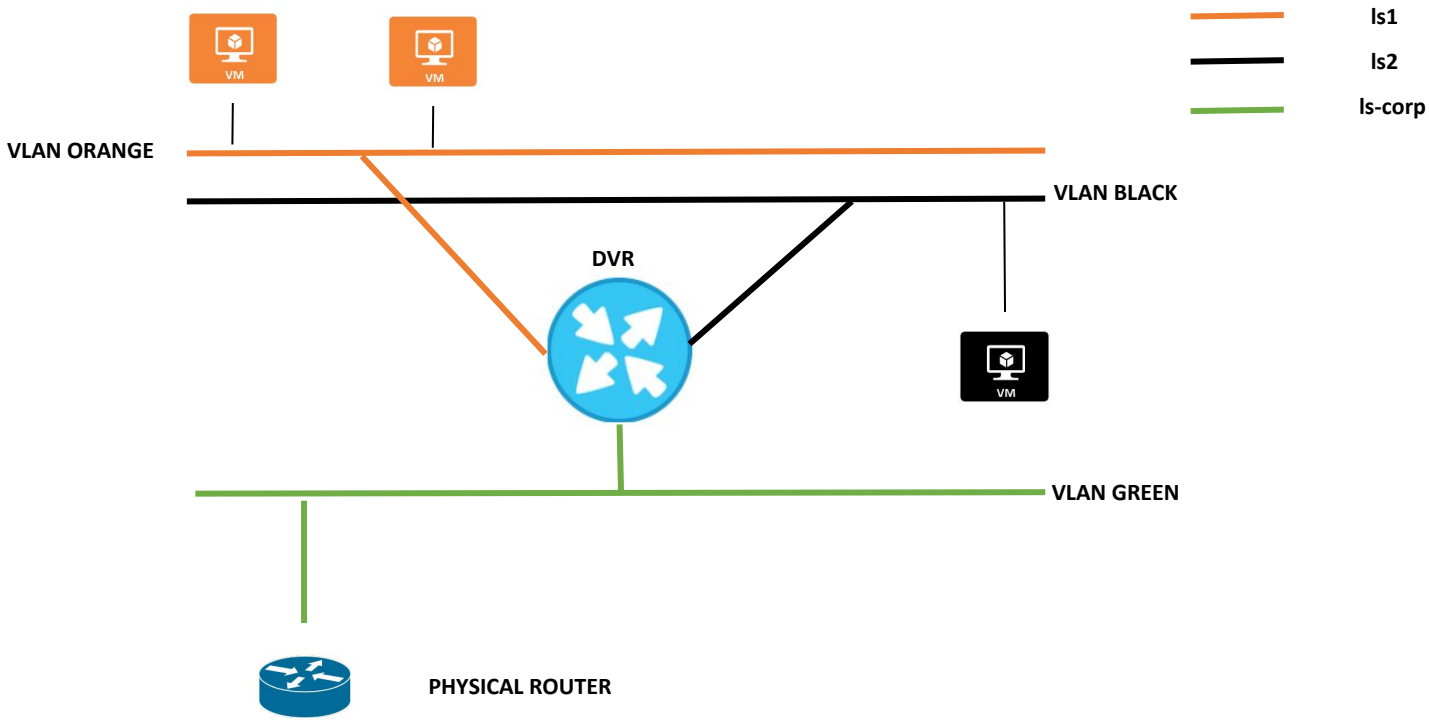
Example, “logical router” in OVN.

VLAN backed DVR

Virtual layer 2 switches are on VLAN networks.

- Not on overlay.
- Traffic is not encapsulated.

INTRODUCTION



INTRODUCTION

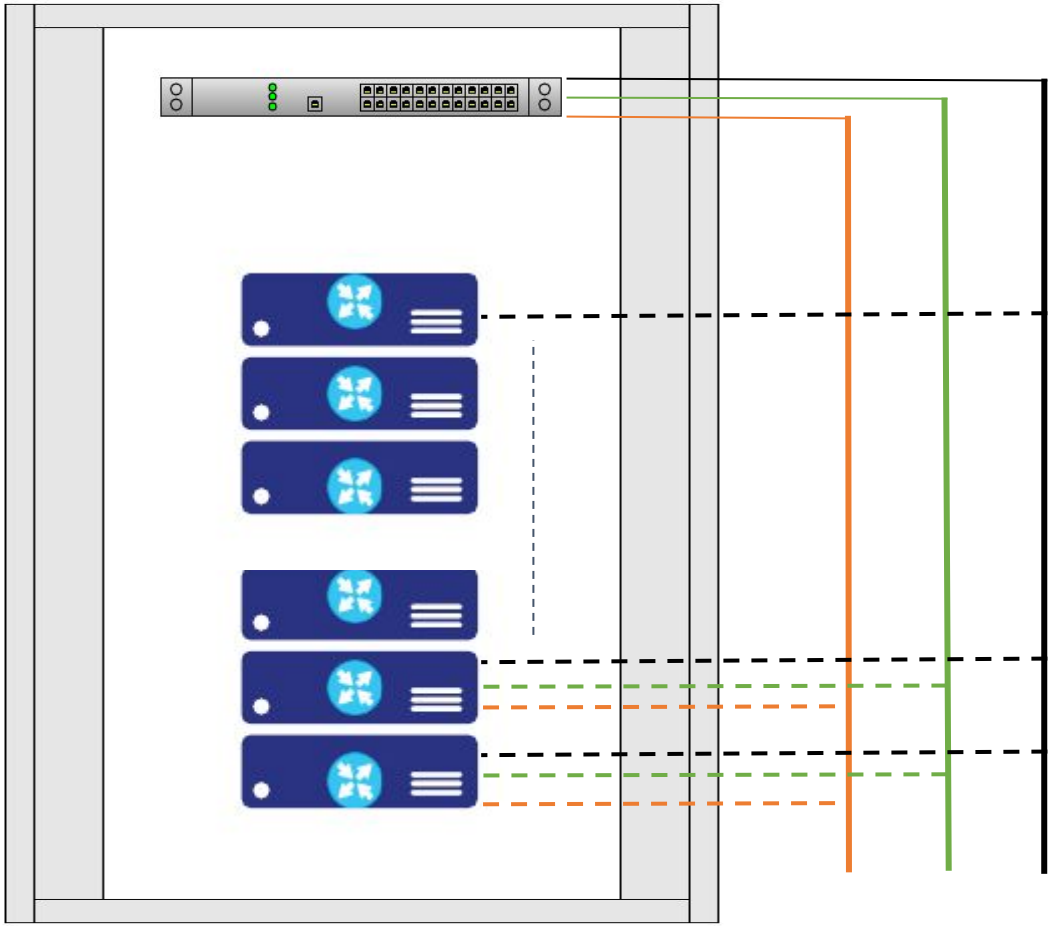
Hypervisor

DVR

VLAN ORANGE

VLAN BLACK

VLAN GREEN



CHALLENGES

DVR Based

- Because of distributed nature of the router.

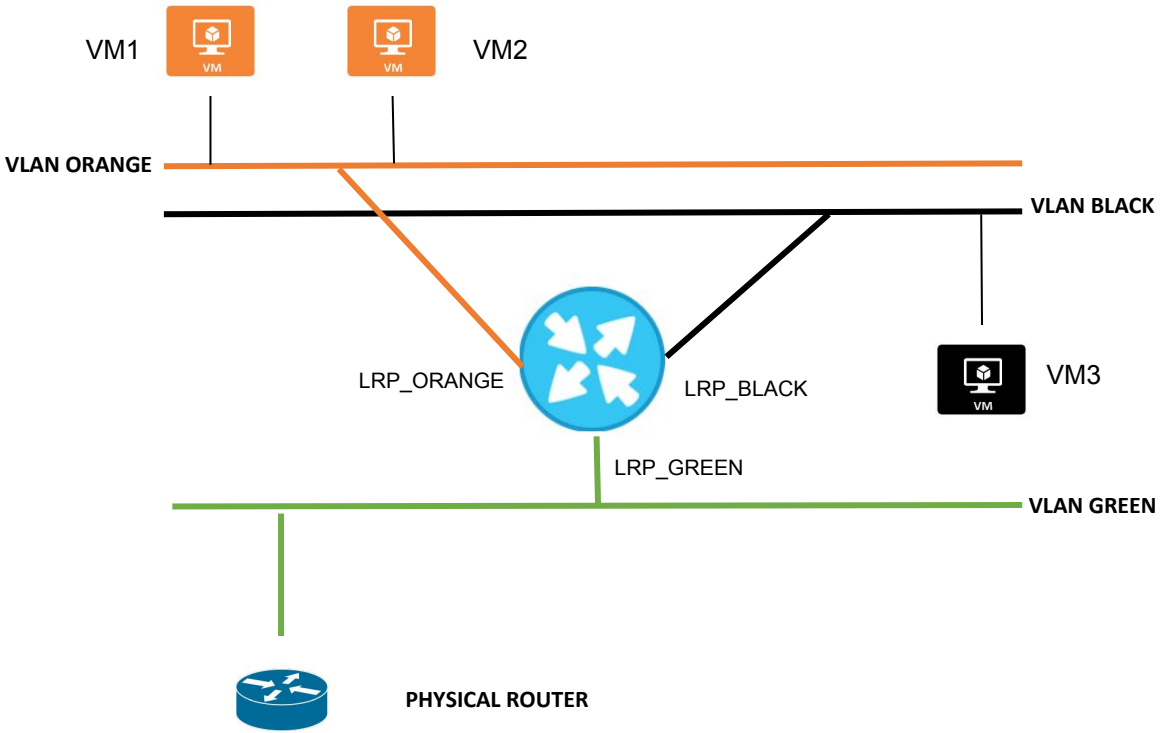
OVN Based

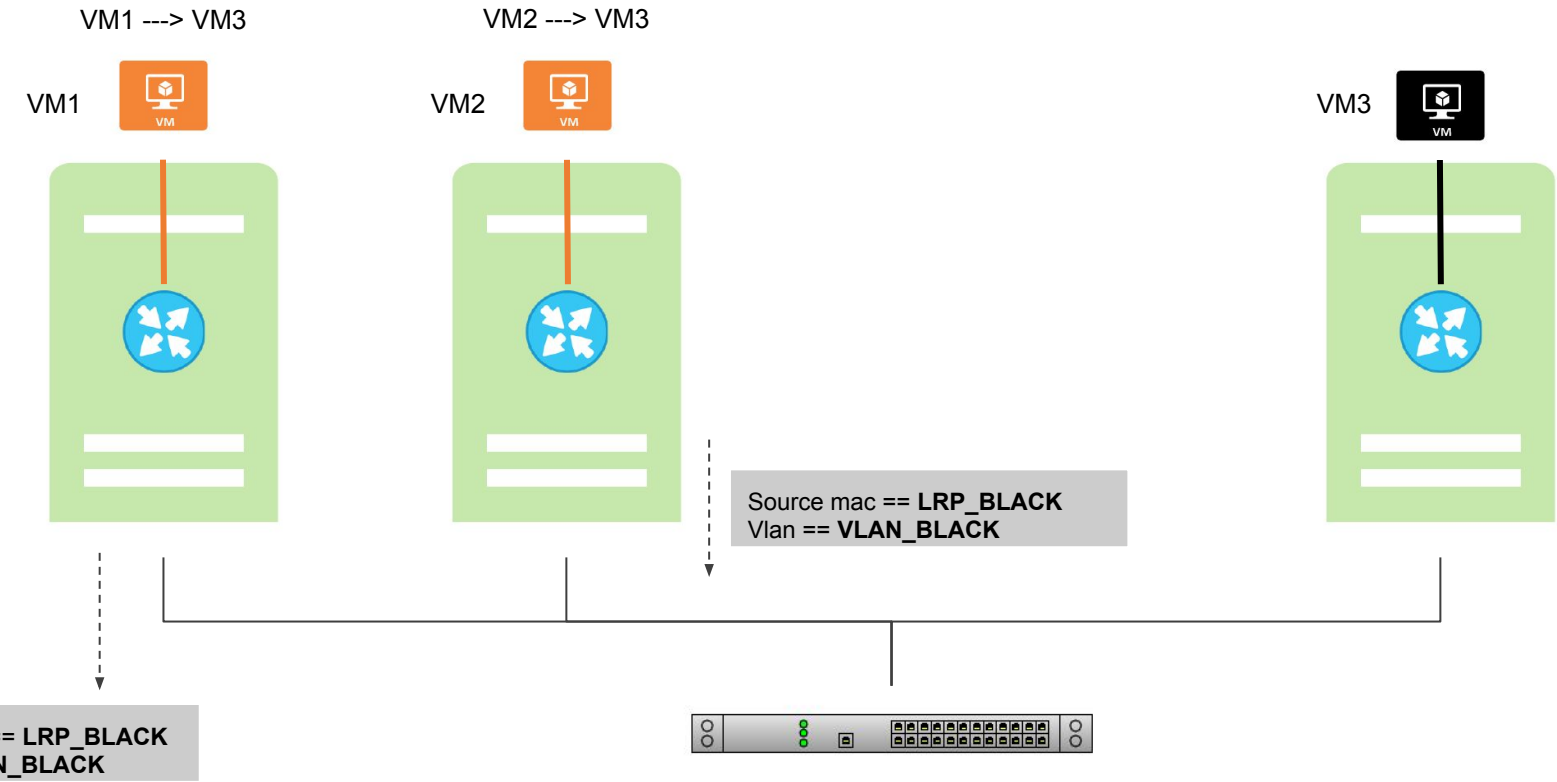
- Because of OVN way of doing things.

DVR Based

Continuous mac move of logical router port macs.

- Logical router ports are distributed across hypervisors.
- Packets with logical router port MAC address as source MAC will come from multiple chassis.
- Will cause mac moves on Top of the Rack (TOR) switch.
  - As a security measure packets could be dropped.
  - As a security measure corresponding TOR switch port could be blocked.


VM1
VM
VM
VM2
VLAN ORANGE
VLAN BLACK
LRP_ORANGE
LRP_BLACK
VM
VM3
LRP_GREEN
VLAN GREEN
PHYSICAL ROUTER

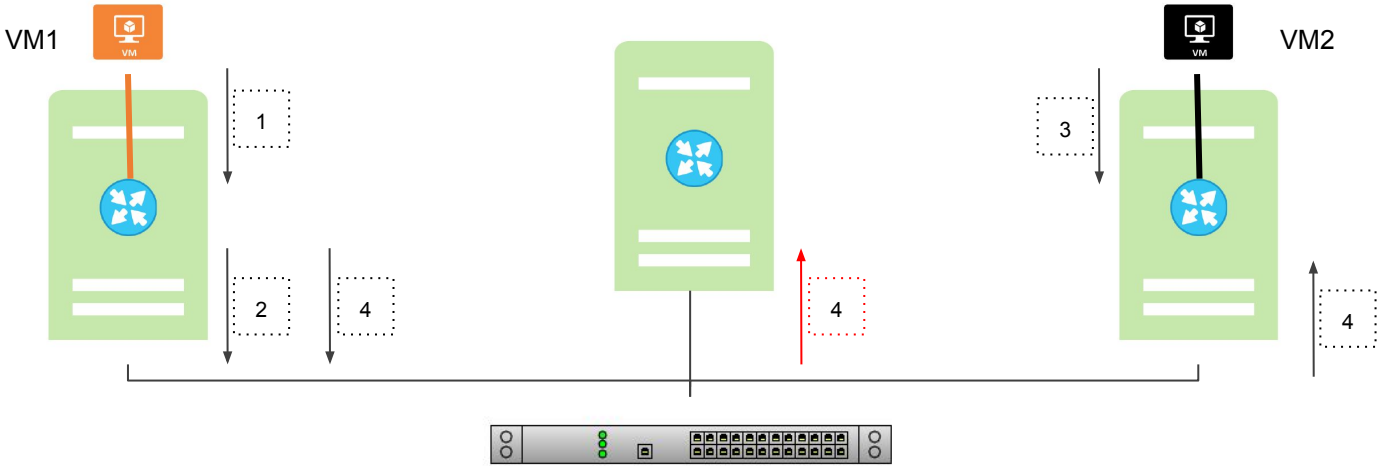

VM1 ---> VM3
VM2 ---> VM3
VM1
VM2
VM3
VM
VM
VM
Source mac == **LRP_BLACK**
Vlan == **VLAN_BLACK**
Source mac == **LRP_BLACK**
Vlan == **VLAN_BLACK**

OVN Based

ARP replies from VIFs destined to router ports not sent on wire.

- By design, ovn consumes it and populates mac binding table.
- Not sending on wire, would mean that VM mac is not learnt by TOR.
- TOR will always end up broadcasting the traffic with destination mac that of the VM.


1
I C M P
Dest IP == VM2_IP
VM1
2
ARP_REQ for VM2_IP
VM2
3
ARP_RESP
4
ICMP
Dest IP == VM2_IP
DEST_MAC == VM2_MAC


VM1
VM2
1
2
4
4
3
4

OVN Based

North-South traffic without NAT.

- For VLAN backed networks, NATing north bound traffic is not must.
- OVN lacks “proper” support for this case.

Dependency on encap header.

- SNAT design relies on metadata carried in geneve encap header.
- Process only egress pipeline on destination chassis.

OVN ENHANCEMENT

LAYER 2

No changes needed in this area.

We will leverage on “localnet” ports to send vlan tagged traffic on wire.

Router Port Mac Move Handling

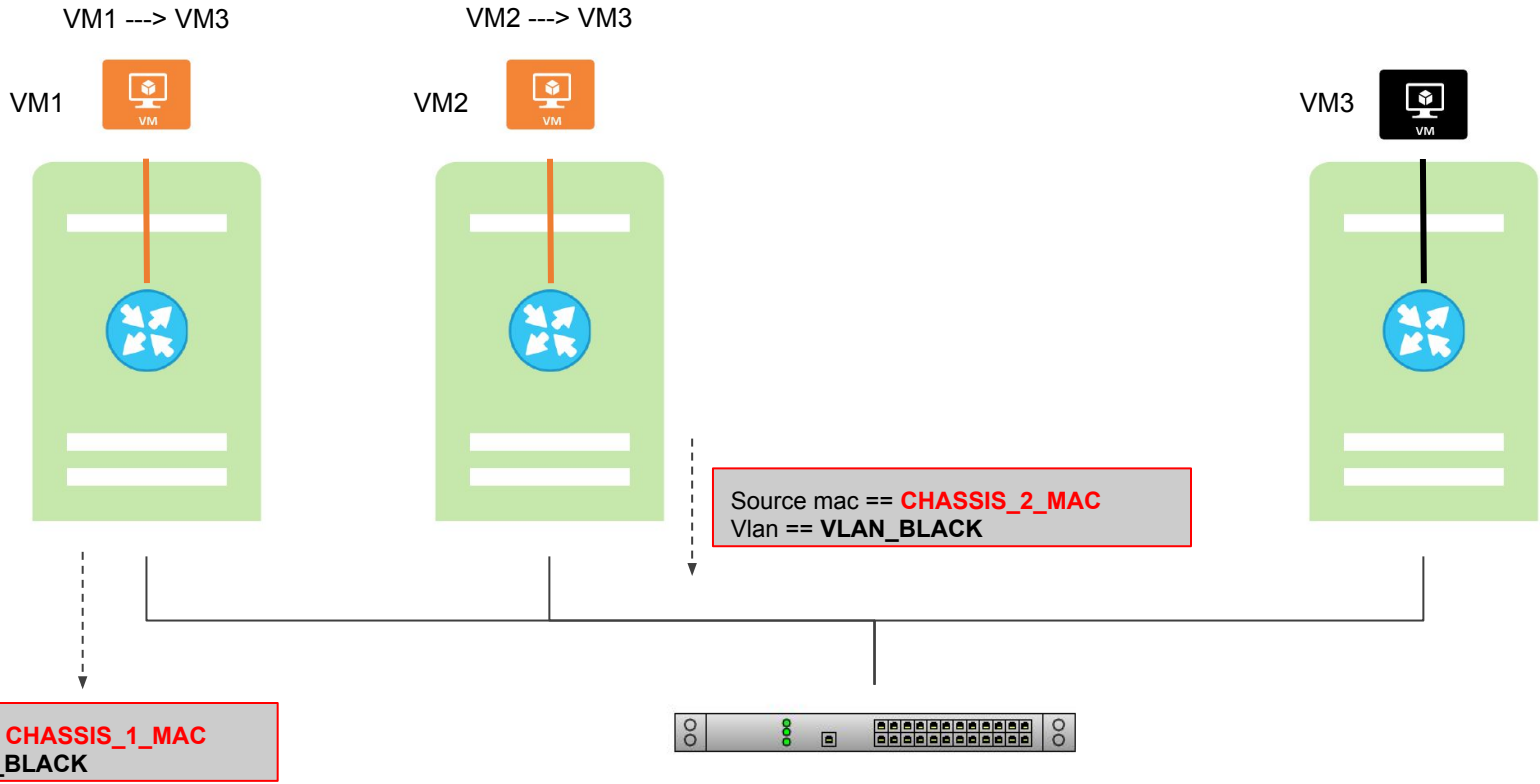
Configure unique MAC(s) on every chassis.

- For example, “ovn-chassis-mac-mappings = physnet:aa:bb:cc:dd:ee:ff”
- We will refer to this mac as “chassis mac”.

ovn-controller programs flow to replace router port mac with chassis mac before sending traffic to physical network.

- table=65, priority=150
- Matches if egress port is localnet port
- For example: chassis_mac = “aa:bb:cc:dd:ee:ff”, router_port_mac = “00:00:01:01:02:04”

cookie=0x0, duration=30.226s, table=65, n_packets=0, n_bytes=0, idle_age=30, priority=150,reg15=0x5,metadata=0x2, dl_src=00:00:01:01:02:04 actions=mod_dl_src:aa:bb:cc:dd:ee:ff,output:30,mod_vlan_vid:118

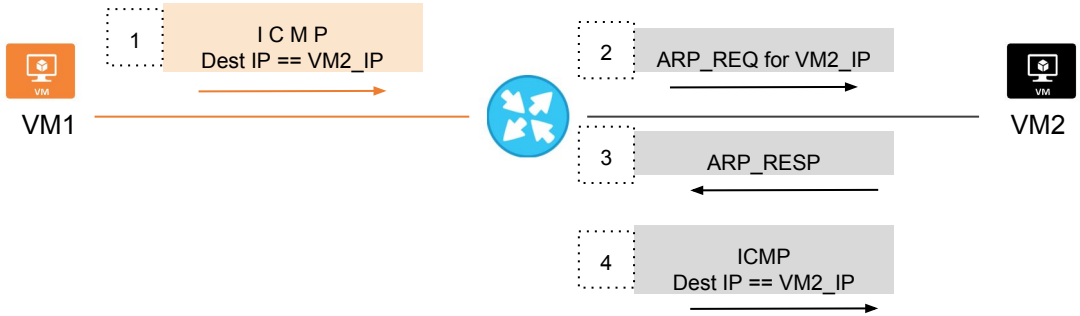

VM1 ---> VM3
VM2 ---> VM3
VM1
VM2
VM3
VM
VM
VM
Source mac == CHASSIS_2_MAC
Vlan == VLAN_BLACK
Source mac == CHASSIS_1_MAC
Vlan == VLAN_BLACK

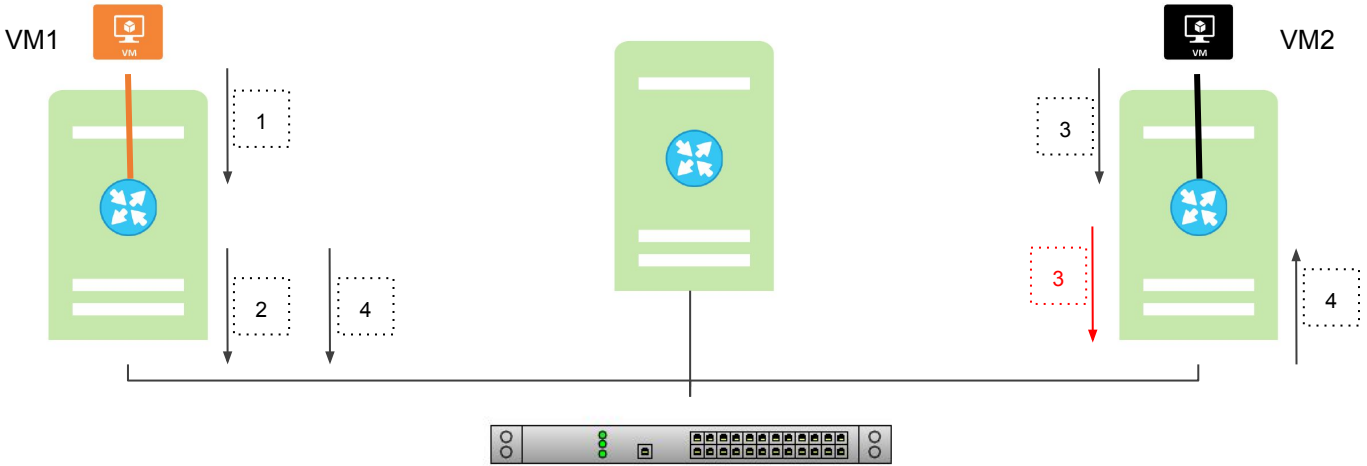
ARP replies to router port

ovn-controller will process the ARP response to router port.

Additionally, it will generate reverse ARP (RARP) packet to be sent on wire.

- Contingent that logical switch is of type VLAN.
- TOR will learn VM's mac through this packet.


VM1
1
I C M P
Dest IP == VM2_IP
2
ARP_REQ for VM2_IP
VM2
3
ARP_RESP
4
ICMP
Dest IP == VM2_IP


VM1
1
2
4
3
3
4
VM2

North-South Traffic

We will rely on gateway-chassis to resolve ARP for Router Port.

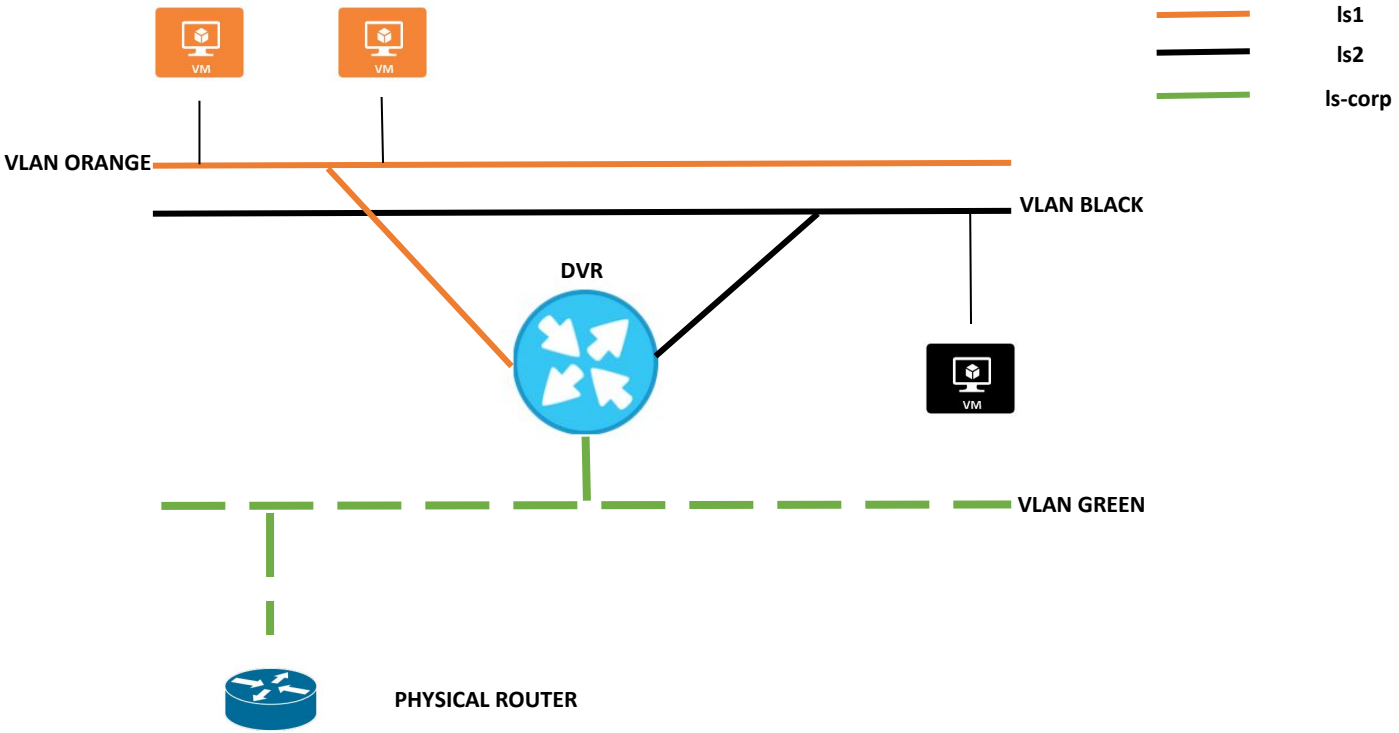
- Gateway chassis means a single chassis will become entry point for all the “outside” traffic.
- This router port IP will be advertised as next hop for logical networks behind DVR.

Advertise router port mac using GARP when a chassis becomes active gateway chassis.

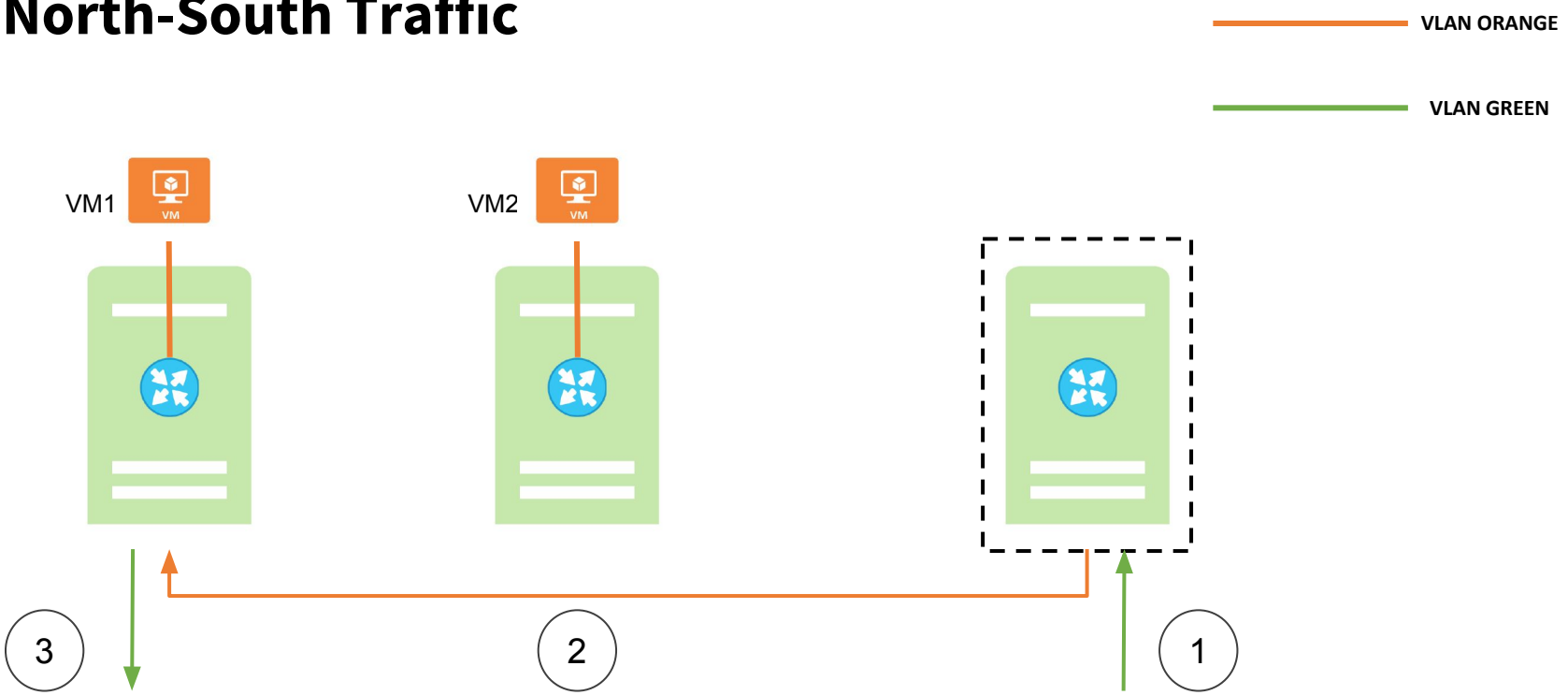
From wire, ARP will be resolved only for router port which has gateway chassis attached.

- This ARP will be resolved ONLY on the gateway chassis.

Source MAC does not get replaced on gateway chassis.


VM
VM
ls1
ls2
ls-corp
VLAN ORANGE
VLAN BLACK
DVR
VM
VLAN GREEN
PHYSICAL ROUTER

North-South Traffic



Network Address Translation (NAT)

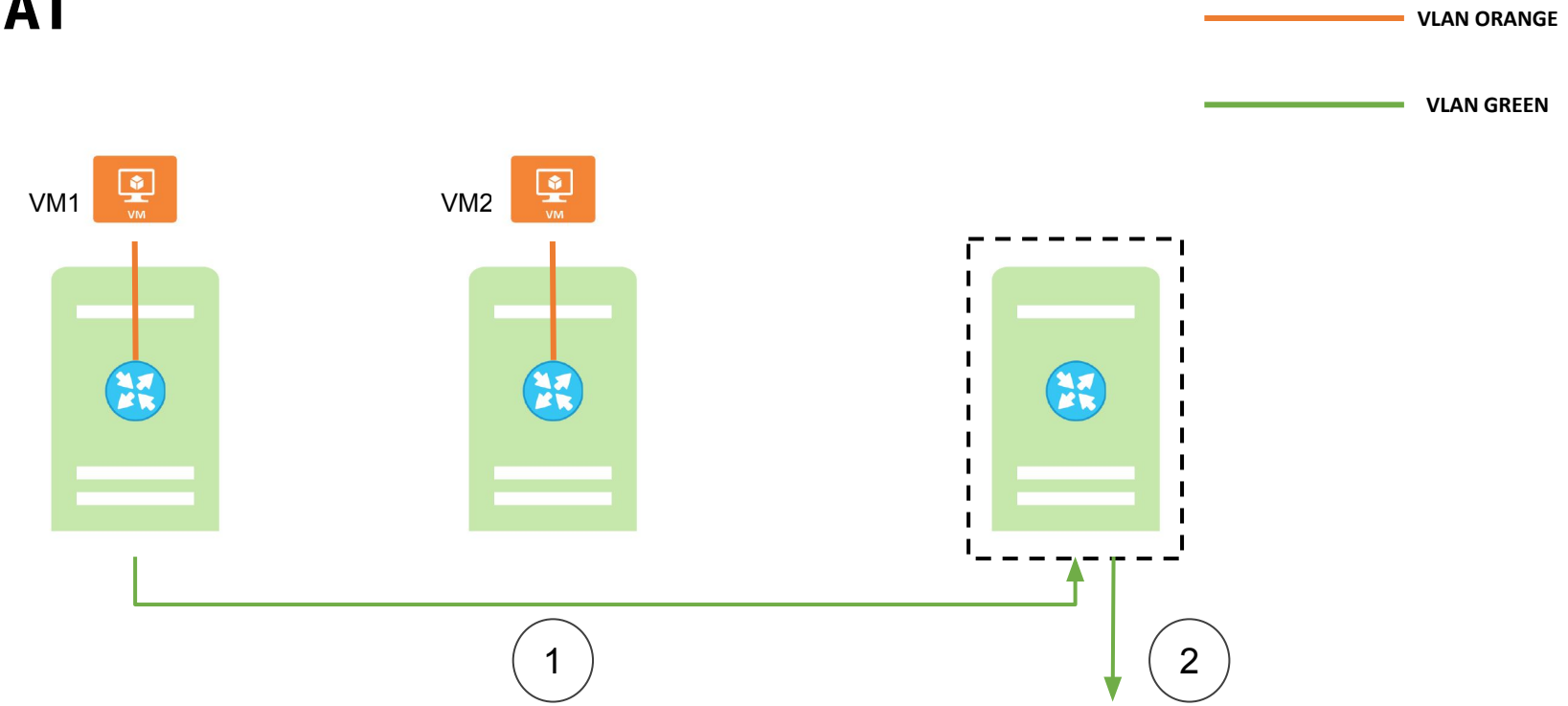
Redirect the packet to gateway chassis, using router port mac as dest mac.

Packet enters router pipeline “again” on gateway chassis.

- Without encap, there is no way to force packet to just egress pipeline of router.

NATed packet goes on wire.

NAT



COMPARISON WITH OVERLAY DVR

LAYER 2 PIPELINE

- Ingress pipeline executed **again** on destination chassis.
  - In the absence of encap, no way to enforce only egress pipeline execution on destination chassis.

LAYER 3 E-W PIPELINE

- No changes in L3 pipeline.
  - Router pipeline executed only on source chassis.

LAYER 3 N-S PIPELINE (NO NAT)

- Traffic originating from OVN chassis will not go to gateway chassis.

LAYER 3 N-S PIPELINE (NAT)

- Router ingress pipeline executed **again** on gateway chassis.

CURRENT STATUS

Proof of concept internally done and tested.

Proposal discussed with community.

- https://mail.openvswitch.org/pipermail/ovs-dev/2018-October/353066.html

Some patches out for review.

FUTURE WORK

ARP Probing/Ageing

- Advertise VM macs to TOR, periodically.
- Or age out VMs mac periodically.
- Both timers should be configurable.

DHCP Relay

Equal Cost Multi Path Routing (ECMP)

Questions