cloud STACK | COLLABORATION CONFERENCE

# Routed mode - Static and Dynamic Routing

Wei Zhou

November 20 - 22, 2024 | Madrid, Spain

# About me

## Wei Zhou

- Apache CloudStack community member since 2012
- Apache CloudStack committer since 2013.05
- Apache CloudStack PMC member since 2017.03

- Software Architect @ Shapeblue since 2021
- Member of Kubernetes org
- Member of OpenSDN.io

- Email: weizhou@apache.org
- Github: @weizhouapache

# Contents

01 Why Routed mode

02 What is Routed mode

03 How to configure a Routed network

04 How to configure Dynamic routing

05 Deep dive: How it works

06 Summary and Future work

# 01 Why Routed mode

# Guest networks in CloudStack

Guest network types

- Shared
- Isolated
- L2

Problems to address for Shared/Isolated network:

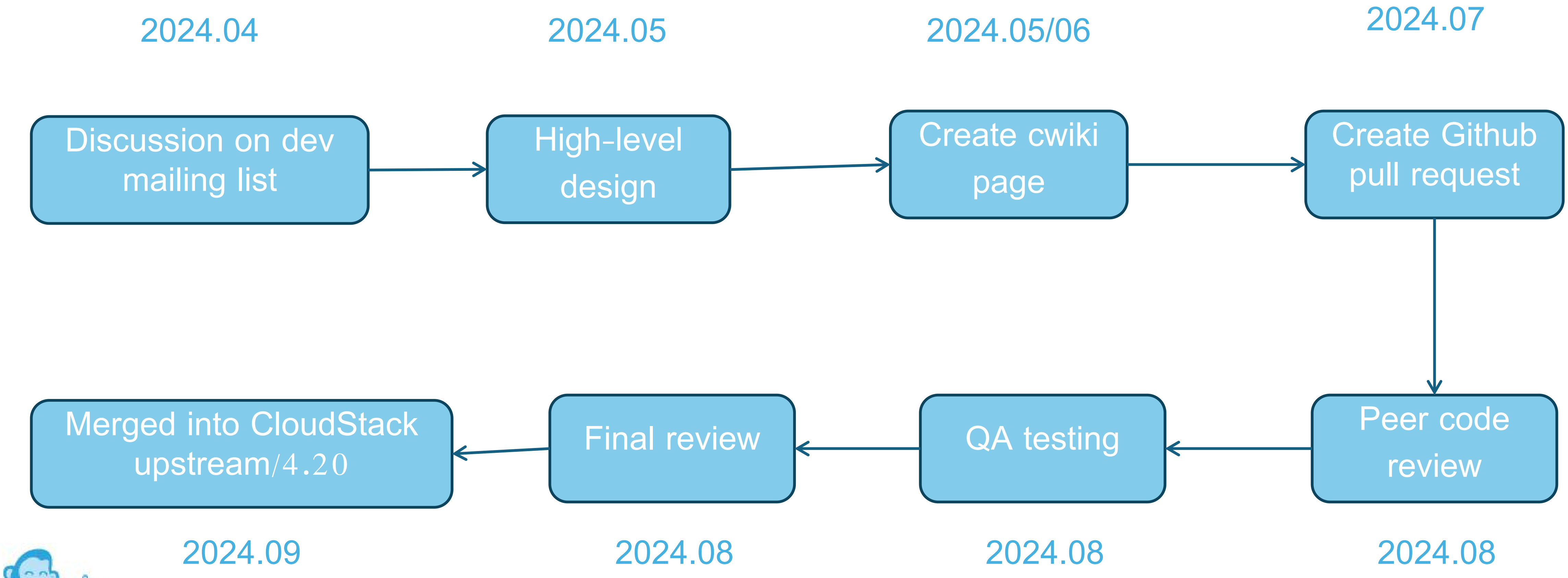
- IP assignment (IPv4/IPv6 address, netmask, gateway. Dhcp/Dns)
- IP Routing (Between VM instances and the Internet)
- Network Access Control (For inbound/outbound traffic)

# Overview of network types

| | Shared network | Isolated network | New network type or mode ❓ |
|---|---|---|---|
| ***IP assignment*** | Direct IP<br>Publicly accessible | Private IP<br>No public access | Direct IP<br>Publicly accessible |
| ***IP Routing*** | CloudStack VR is not gateway<br>Requires operators:<br>– Configure gateway on the upstream router<br>– Create the network in ACS | CloudStack VR is Source NAT gateway<br>Ways to access VMs:<br>Static NAT, Load Balancing, Port Forwarding, VPN | No Source NAT.<br>Doesn't require operators' manual configuration.<br>Can be created by end users. |
| ***Network Access control*** | Security groups (KVM only*) | Egress rules<br>Firewall rules | Ingress/Egress Firewall rules |
| | | | |
| ***Performance*** | Good | Not as good as Shared network | Better performance than Isolated network |

(*) From 4.20, security groups feature is supported for Shared networks in Advanced zone

# Routed mode: timeline



2024.04

Discussion on dev mailing list

2024.05

High-level design

2024.05/06

Create cwiki page

2024.07

Create Github pull request

2024.08

Peer code review

2024.08

QA testing

2024.08

Final review

2024.09

Merged into CloudStack upstream/4.20

(*) Join dev mailing list: dev@cloudstack.apache.org

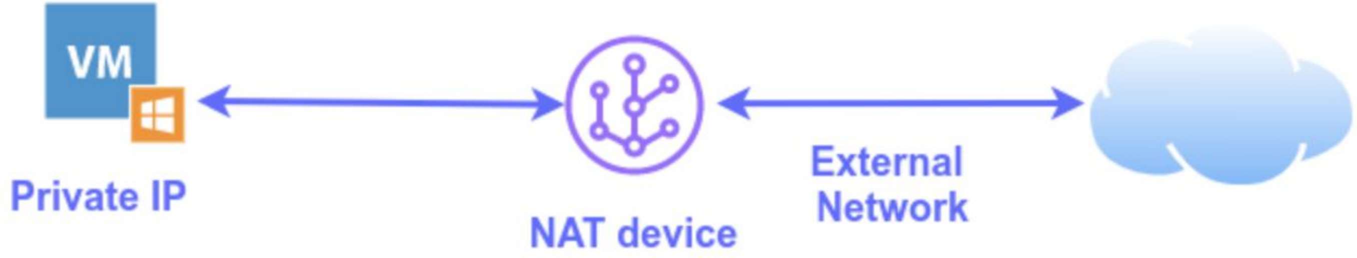
02

## What is Routed mode

(*) Available since Apache CloudStack 4.20

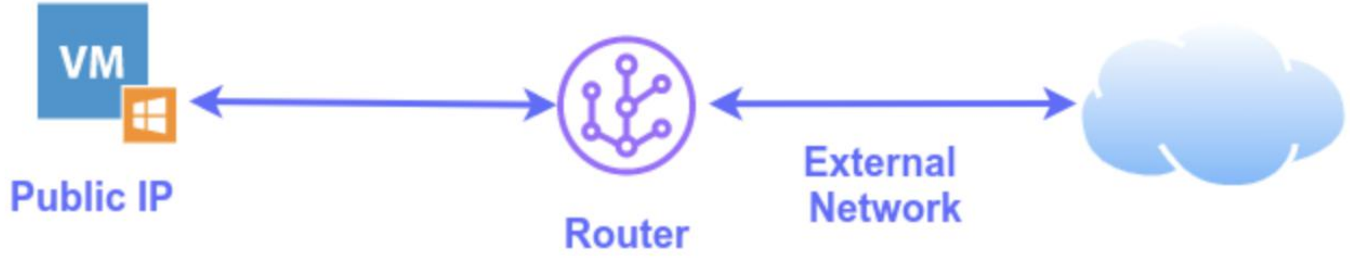
# New concept in ACS 4.20: Network mode

- NATTED mode



  - Default network mode for Isolated networks
  - Virtual Router (VR) as Source NAT (Network Address Translation) gateway

- ROUTED mode:



  - Virtual Router (VR) as Gateway

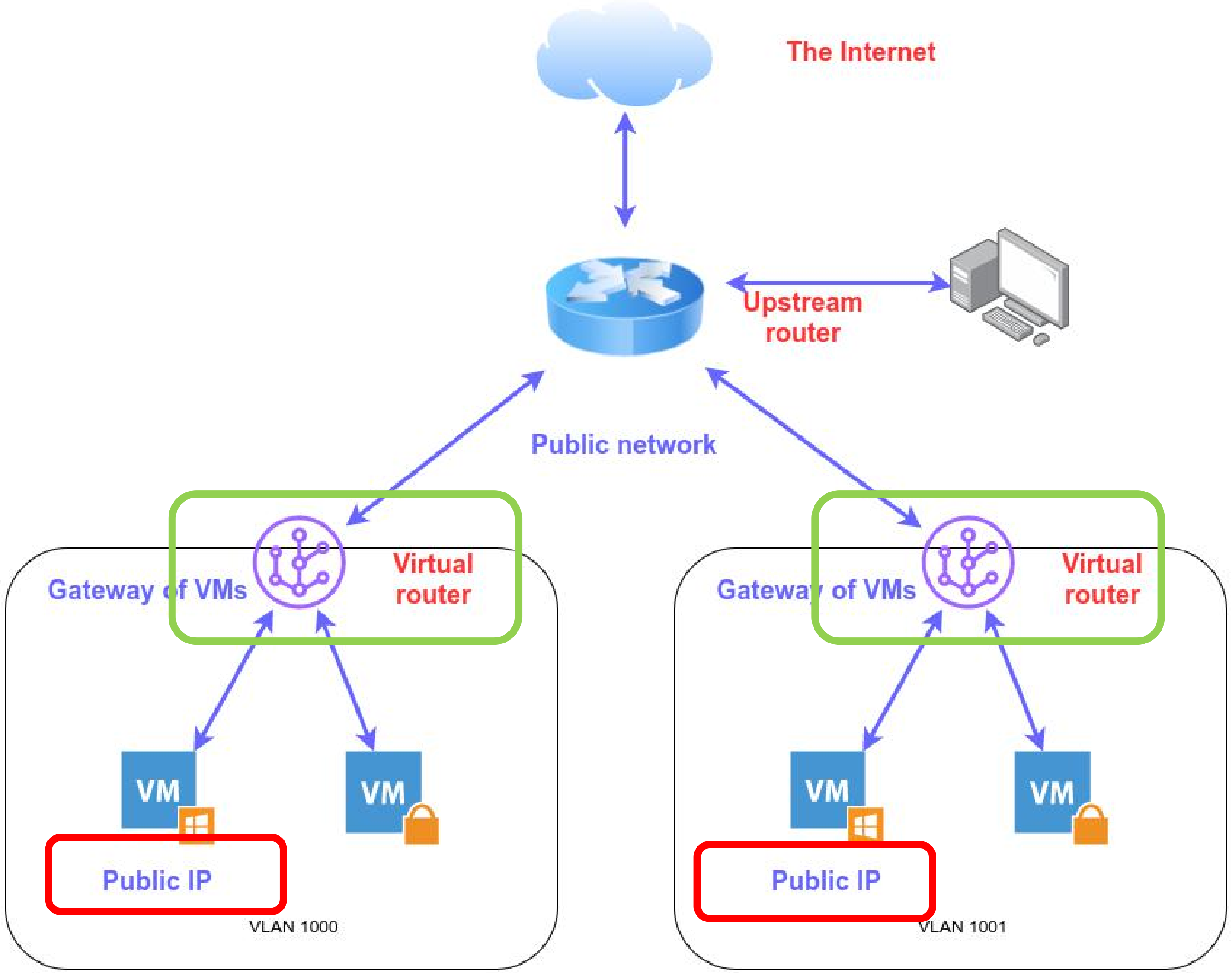
?

Routed mode has already been used in CloudStack.

Which feature uses it？

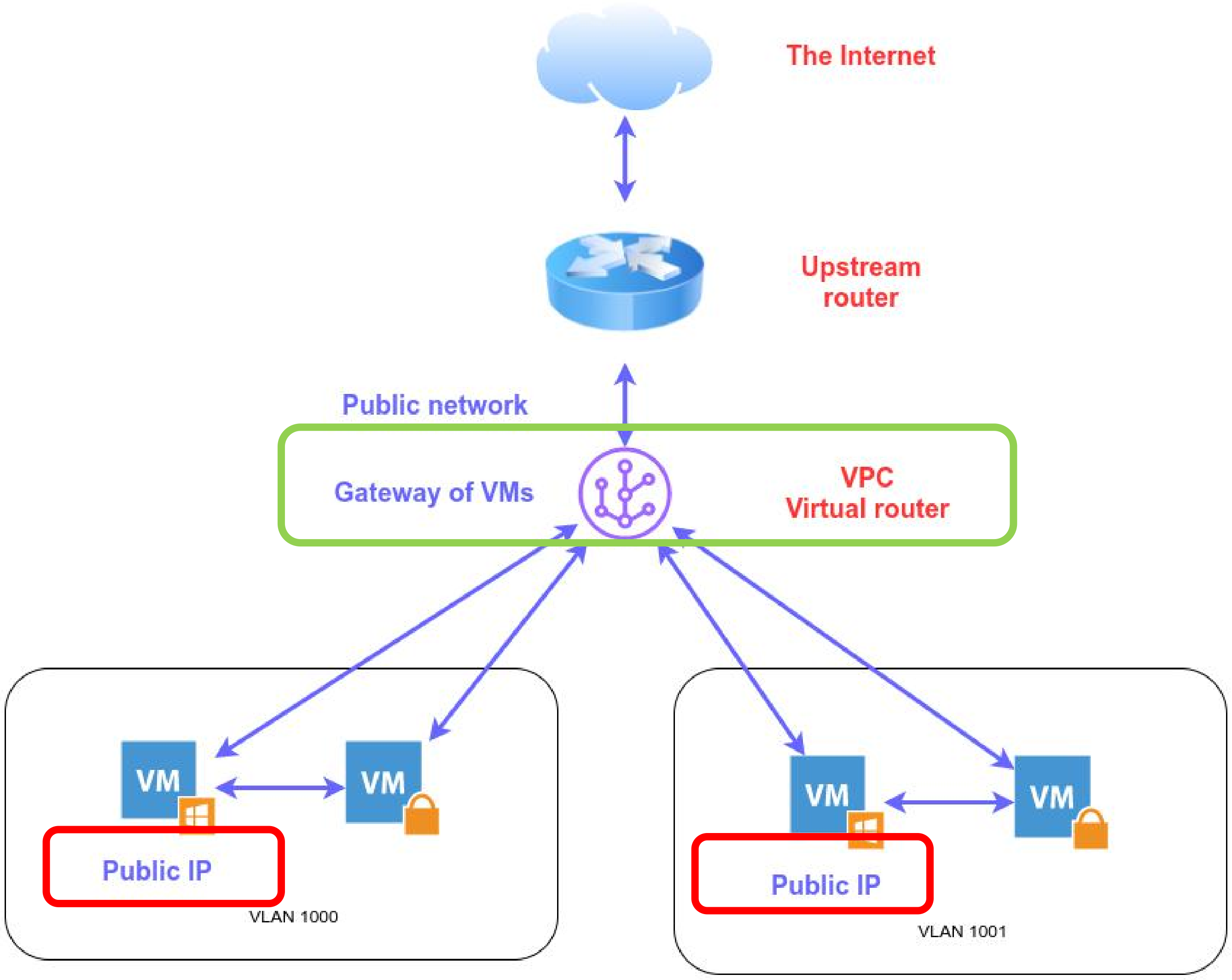
IPv6 support

# Routed Isolated network: Topology



- Very similar to “Isolated network”
- Virtual router (VR) as Gateway of VMs
- Virtual router (VR) provides services
  - Dhcp/Dns/Userdata
  - Firewall
- Differences from Isolated network
  - Public IP vs Private IP
  - No StaticNat/Lb/PF/VPN support

# Routed VPC: Topology



- Very similar to “VPC”
- VPC Virtual router (VPC VR) as Gateway of VMs
- VPC Virtual router (VPC VR) provides services
  - Dhcp/Dns/Userdata
  - Network ACL
- Differences from VPC
  - Public IP v.s. Private IP
  - No StaticNat/Lb/PF/VPN support

# IP Routing in Routed mode

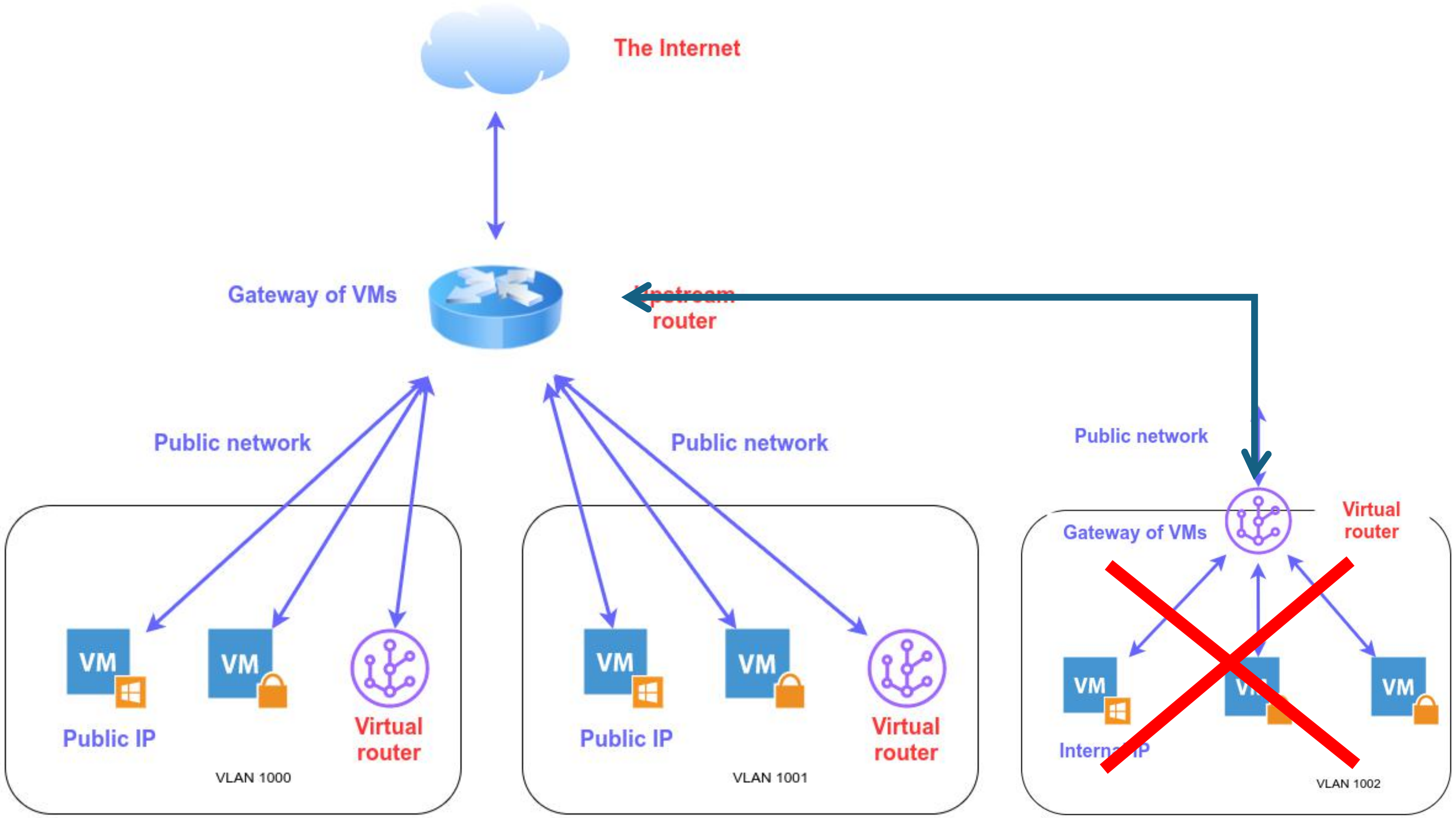
| Routing mode | What the operators need to do |
|---|---|
| **Static routing** | Operators have to manually add **static routes** for each Routed network in the upstream router.<br>Tips: IPv6 implementation already supports it. |
| **Dynamic routing** | Operators configure **Dynamic BGP** in the upstream router<br>The routes for guest networks will be automatically advertised to (upstream and virtual) routers via dynamic routing protocol (BGP). |

# Network Access Control in Routed mode

- Routed networks
  - Egress rules (improved)
  - IPv4 Routing firewall (new!)
  - IPv6 firewall

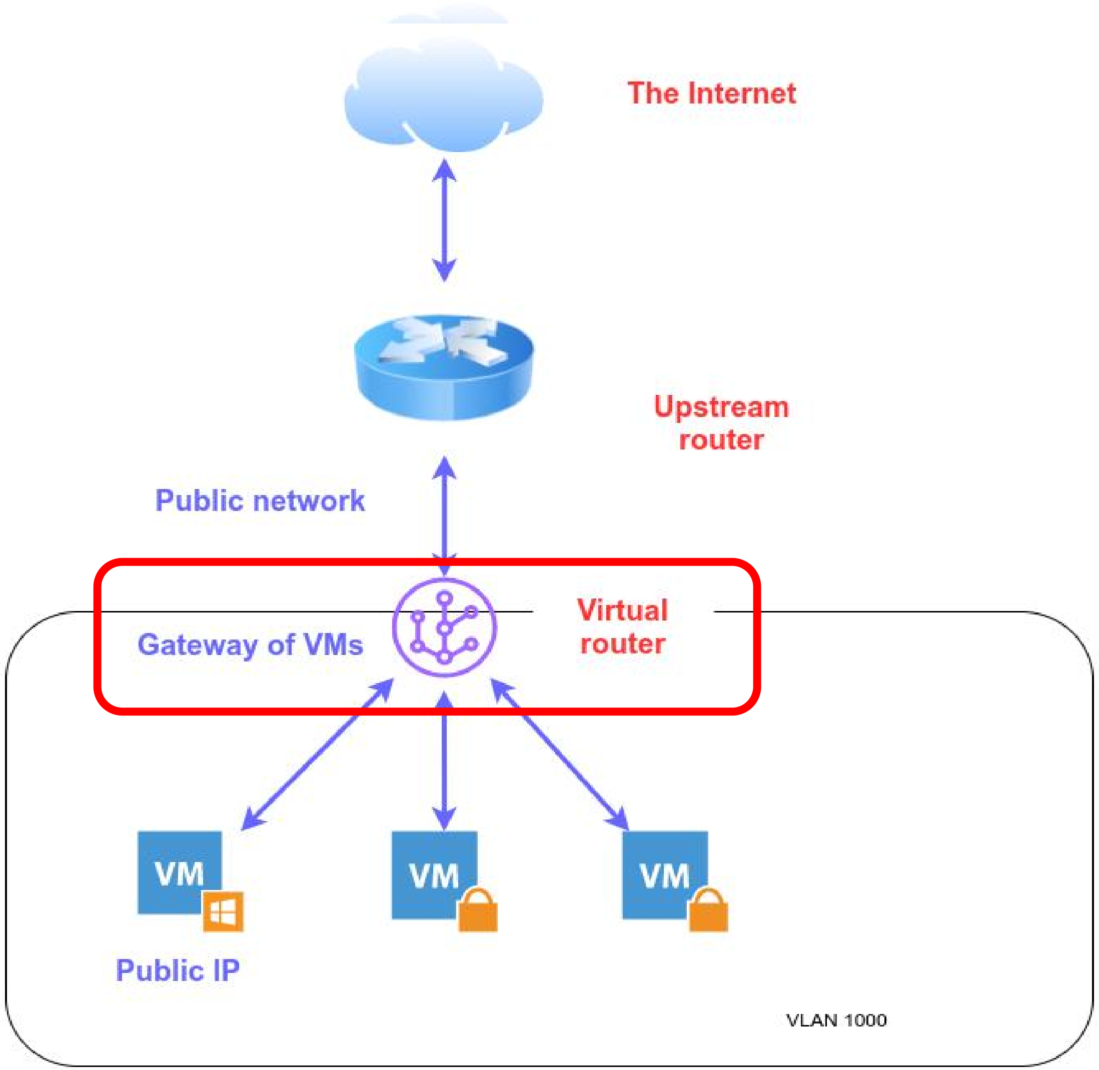
- Routed VPC
  - Network ACL (improved)

# Use case 1: Scalable private cloud



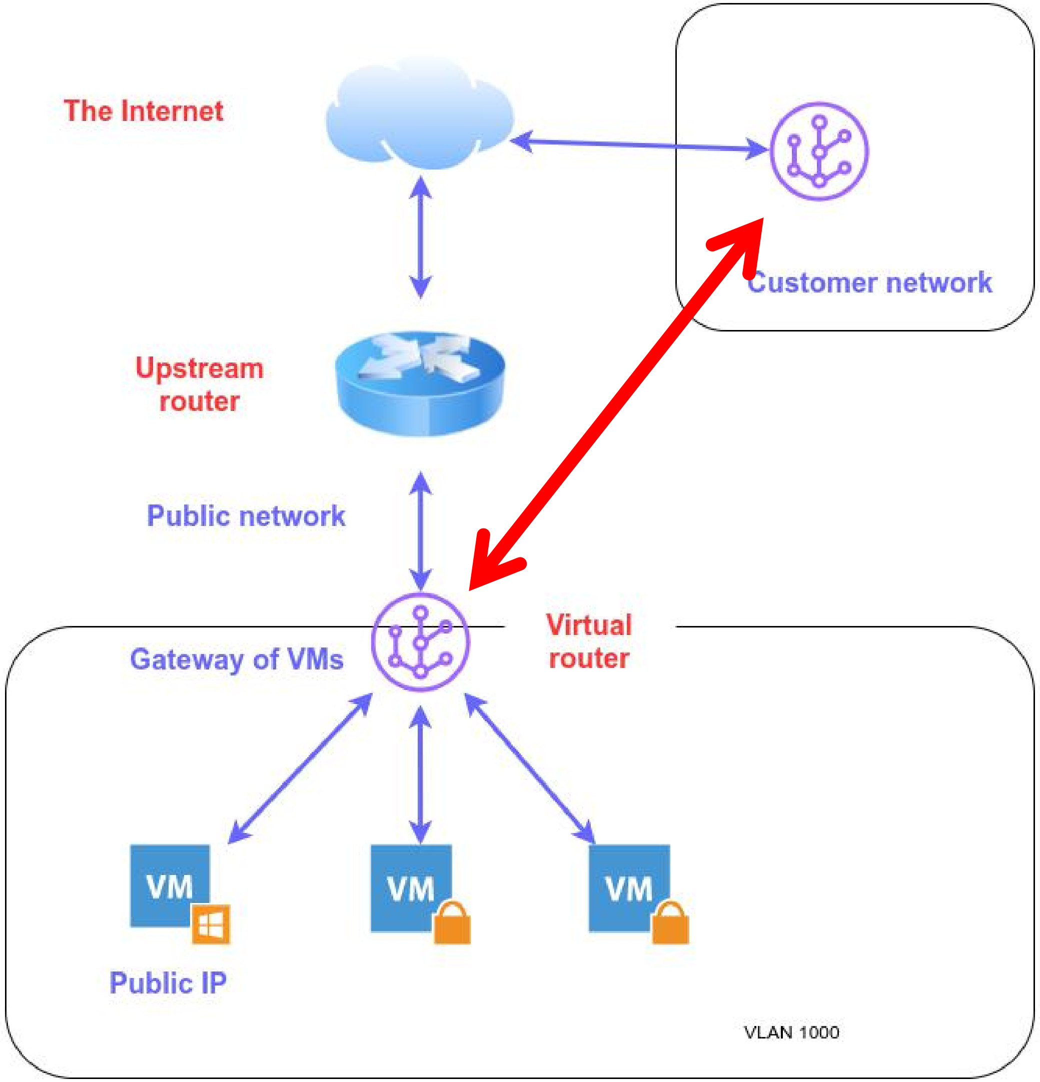
- Private cloud for large organization
- Uses Shared networks with Internal IPs
- New or scale out application
- More flexible and scalable

# Use case 2: VM with Access control on VMware



- Shared network on VMware (or XenServer/Xcp-ng)
- No SG support as Security group rules are applied on hypervisors (KVM only)
- Option: Routed network with access control support

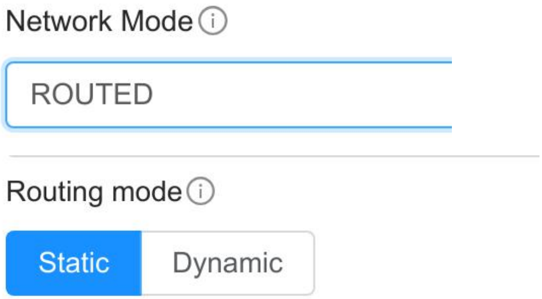
# Use case 3: Bring Your own IPs



- Connects Customer network to Routed network
- via Dynamic routing
- Easy to setup a hybrid cloud

## 03

## How to configure a Routed network

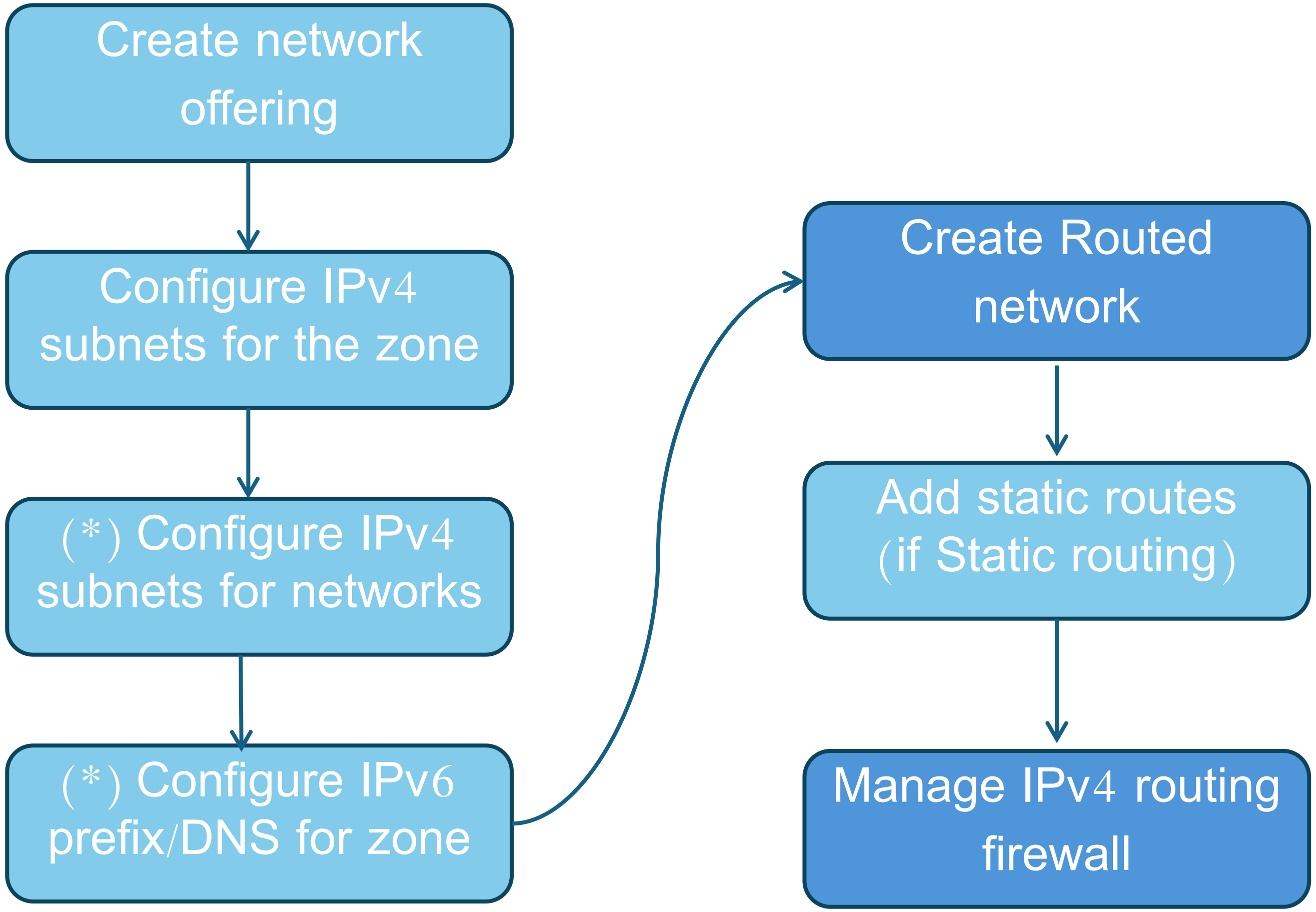


Network Mode ⓘ

ROUTED

Routing mode ⓘ

[✓ Static] Dynamic



Create network offering

Configure IPv4 subnets for the zone

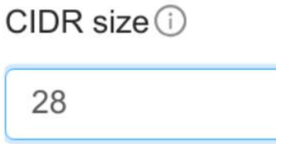
(*) Configure IPv4 subnets for networks

(*) Configure IPv6 prefix/DNS for zone

Create Routed network

Add static routes (if Static routing)

Manage IPv4 routing firewall



CIDR size ⓘ

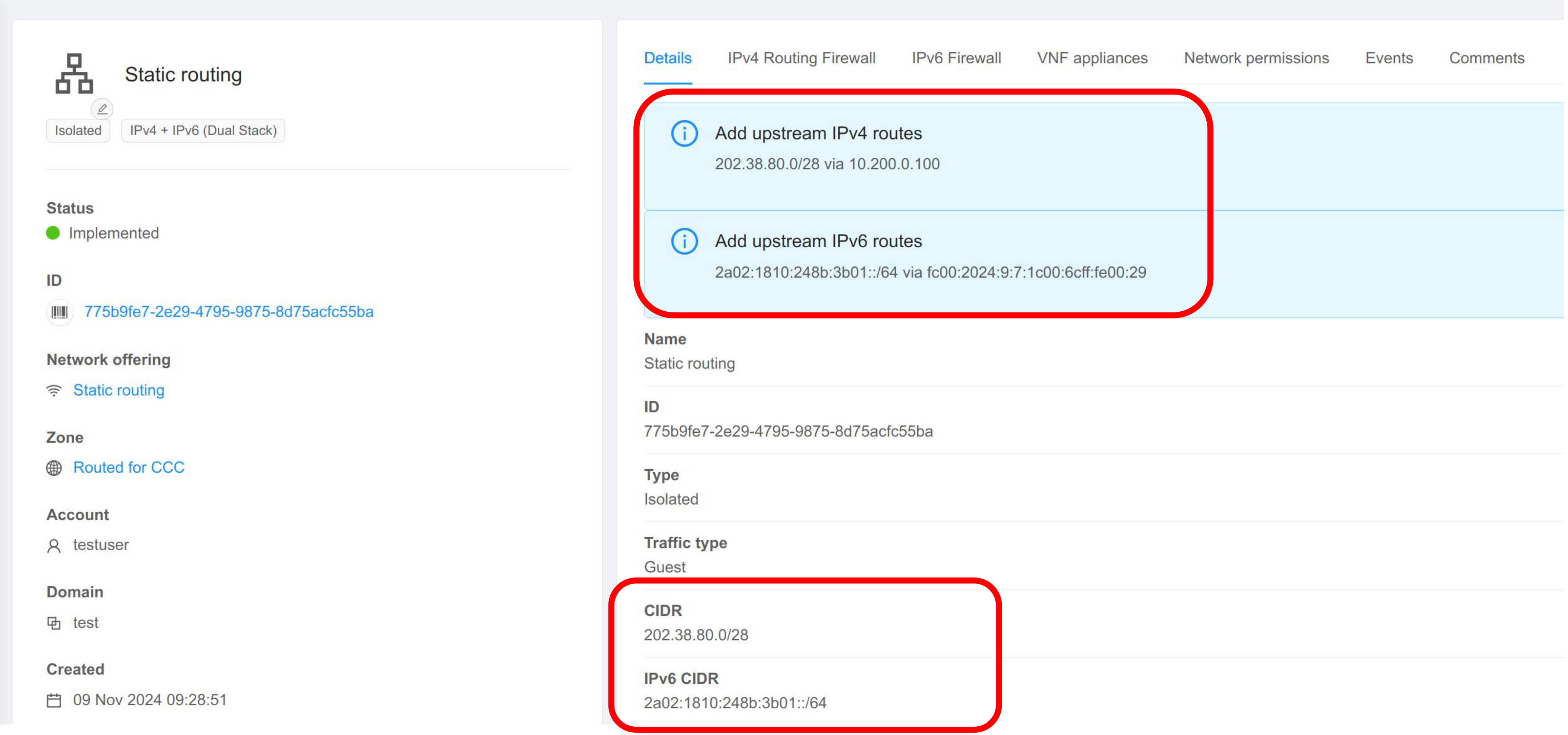
28

Performed by operators

Performed by end users

https://docs.cloudstack.apache.org/en/latest/adminguide/networking/dynamic_static_routing.html

(*) This step is **optional**



Static routing

Isolated IPv4 + IPv6 (Dual Stack)

**Status**
Implemented

**ID**
775b9fe7-2e29-4795-9875-8d75acfc55ba

**Network offering**
Static routing

**Zone**
Routed for CCC

**Account**
testuser

**Domain**
test

**Created**
09 Nov 2024 09:28:51

Add upstream IPv4 routes
202.38.80.0/28 via 10.200.0.100

Add upstream IPv6 routes
2a02:1810:248b:3b01::/64 via fc00:2024:9:7:1c00:6cff:fe00:29

**Name**
Static routing

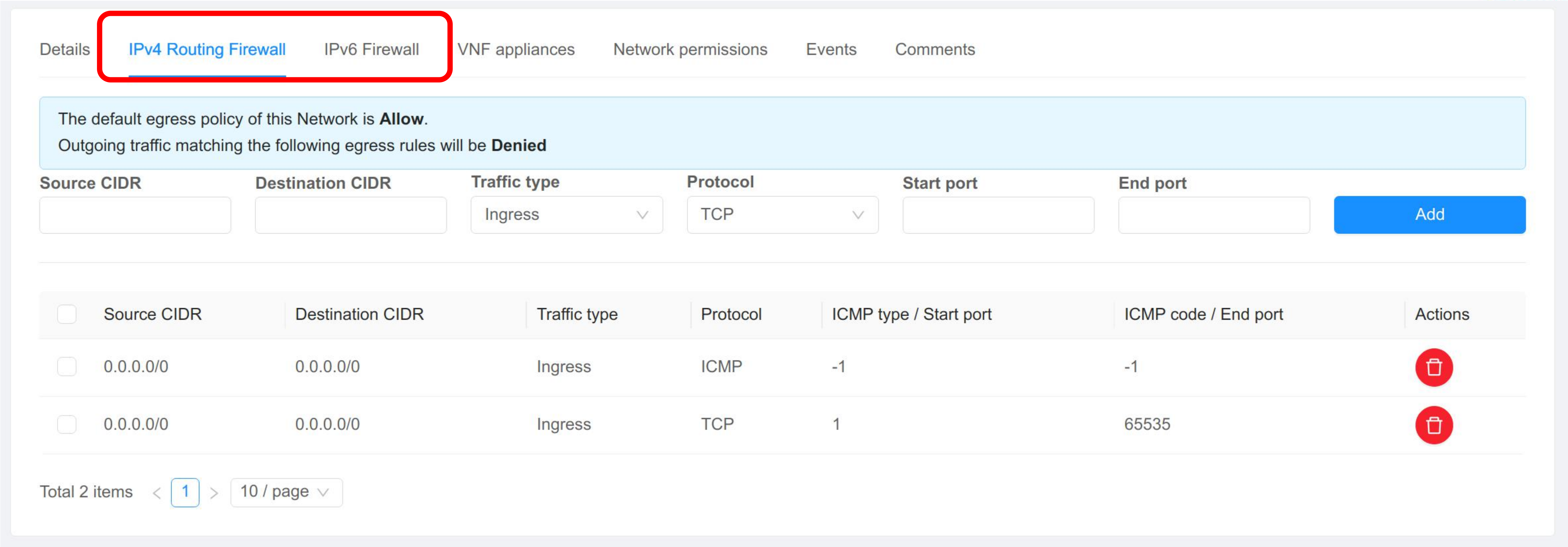
**ID**
775b9fe7-2e29-4795-9875-8d75acfc55ba

**Type**
Isolated

**Traffic type**
Guest

**CIDR**
202.38.80.0/28

**IPv6 CIDR**
2a02:1810:248b:3b01::/64



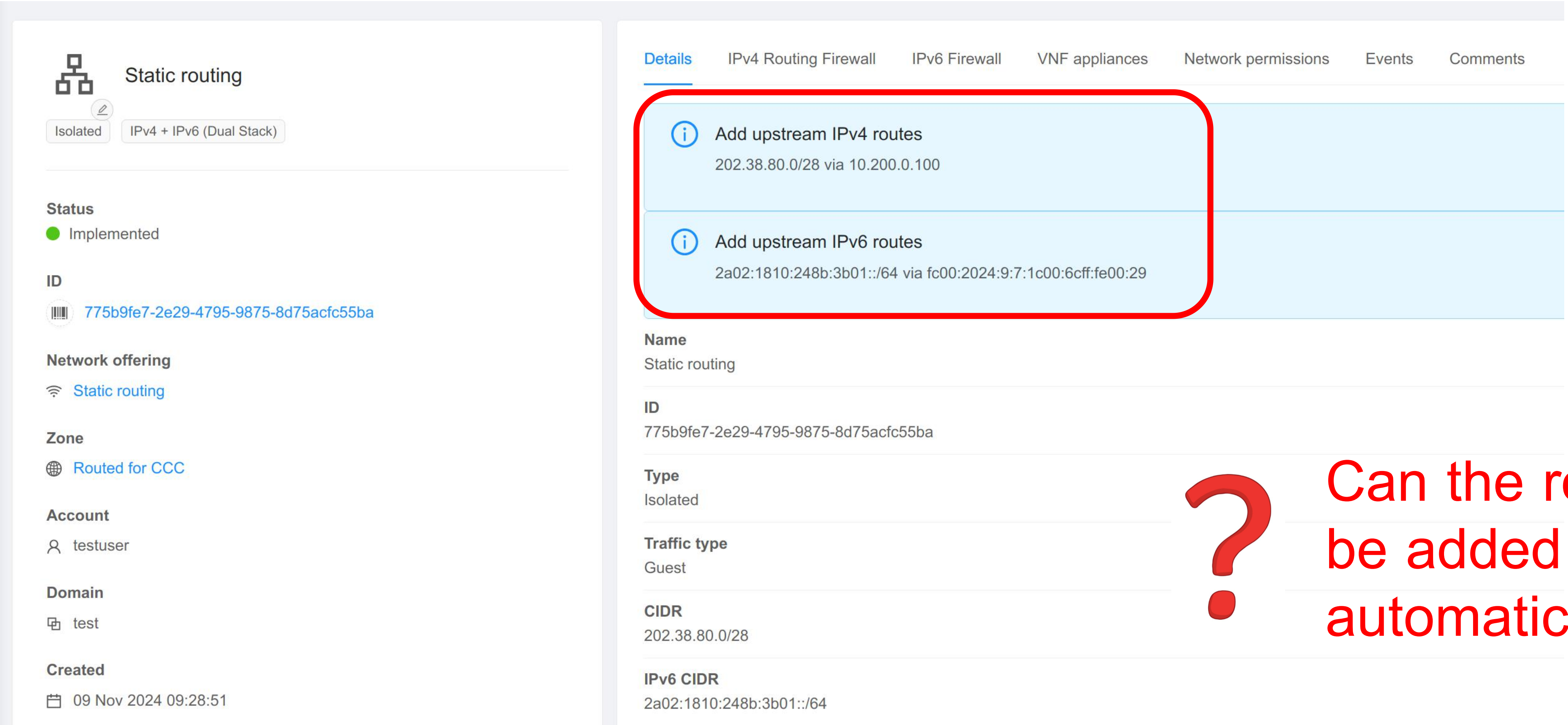
Details | IPv4 Routing Firewall | IPv6 Firewall | VNF appliances | Network permissions | Events | Comments

The default egress policy of this Network is **Allow**.
Outgoing traffic matching the following egress rules will be **Denied**

| Source CIDR | Destination CIDR | Traffic type | Protocol | Start port | End port | |
|---|---|---|---|---|---|---|
| | | Ingress | TCP | | | Add |

| □ | Source CIDR | Destination CIDR | Traffic type | Protocol | ICMP type / Start port | ICMP code / End port | Actions |
|---|---|---|---|---|---|---|---|
| □ | 0.0.0.0/0 | 0.0.0.0/0 | Ingress | ICMP | -1 | -1 | 🗑 |
| □ | 0.0.0.0/0 | 0.0.0.0/0 | Ingress | TCP | 1 | 65535 | 🗑 |

Total 2 items < 1 > 10 / page



Static routing

Isolated | IPv4 + IPv6 (Dual Stack)

**Status**
Implemented

**ID**
775b9fe7-2e29-4795-9875-8d75acfc55ba

**Network offering**
Static routing

**Zone**
Routed for CCC

**Account**
testuser

**Domain**
test

**Created**
09 Nov 2024 09:28:51

Add upstream IPv4 routes
202.38.80.0/28 via 10.200.0.100

Add upstream IPv6 routes
2a02:1810:248b:3b01::/64 via fc00:2024:9:7:1c00:6cff:fe00:29

**Name**
Static routing

**ID**
775b9fe7-2e29-4795-9875-8d75acfc55ba

**Type**
Isolated

**Traffic type**
Guest

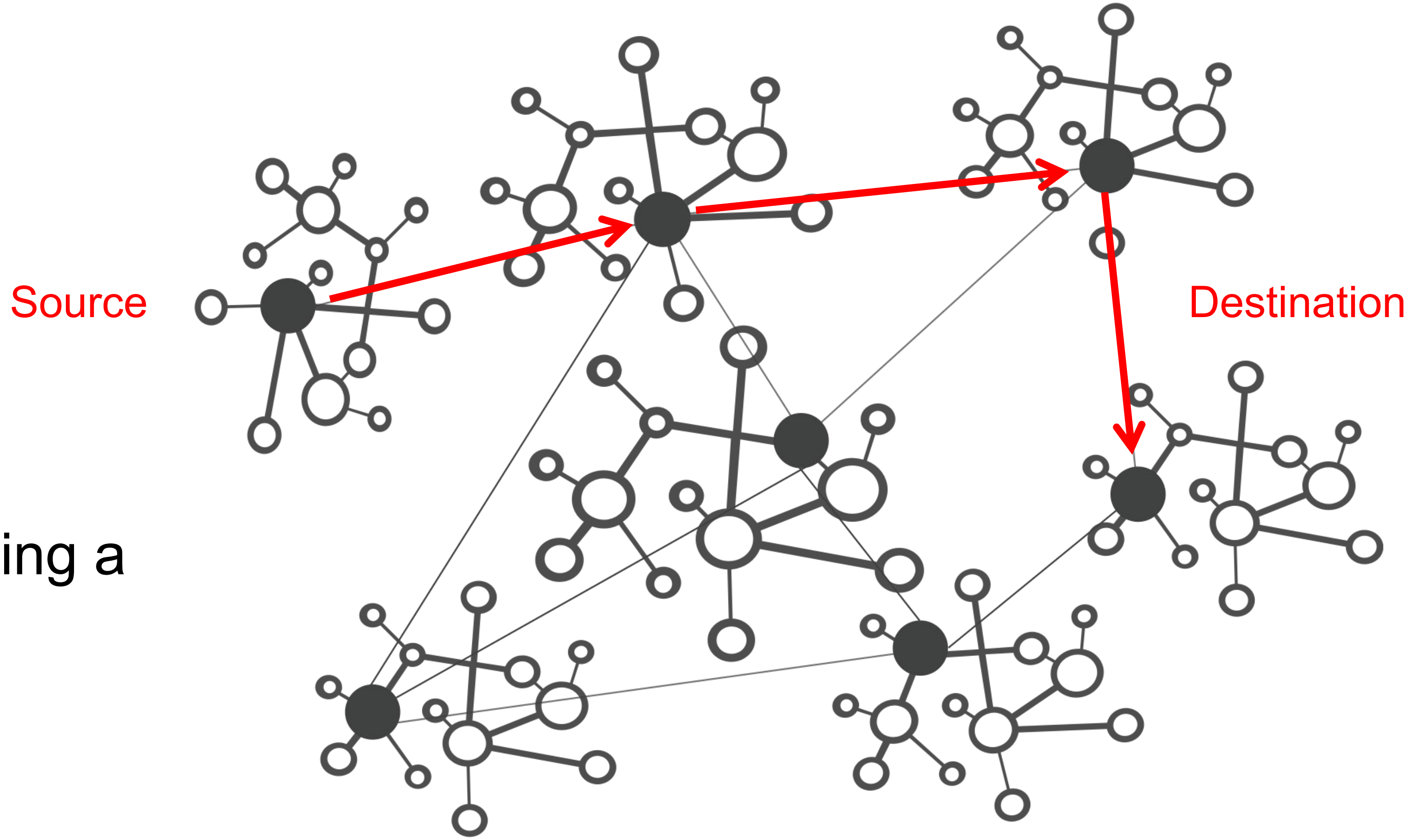
**CIDR**
202.38.80.0/28

**IPv6 CIDR**
2a02:1810:248b:3b01::/64

Can the routes be added automatically ？

## 04 How to configure Dynamic routing

# The Internet is a network of autonomous systems (AS).



Routing:

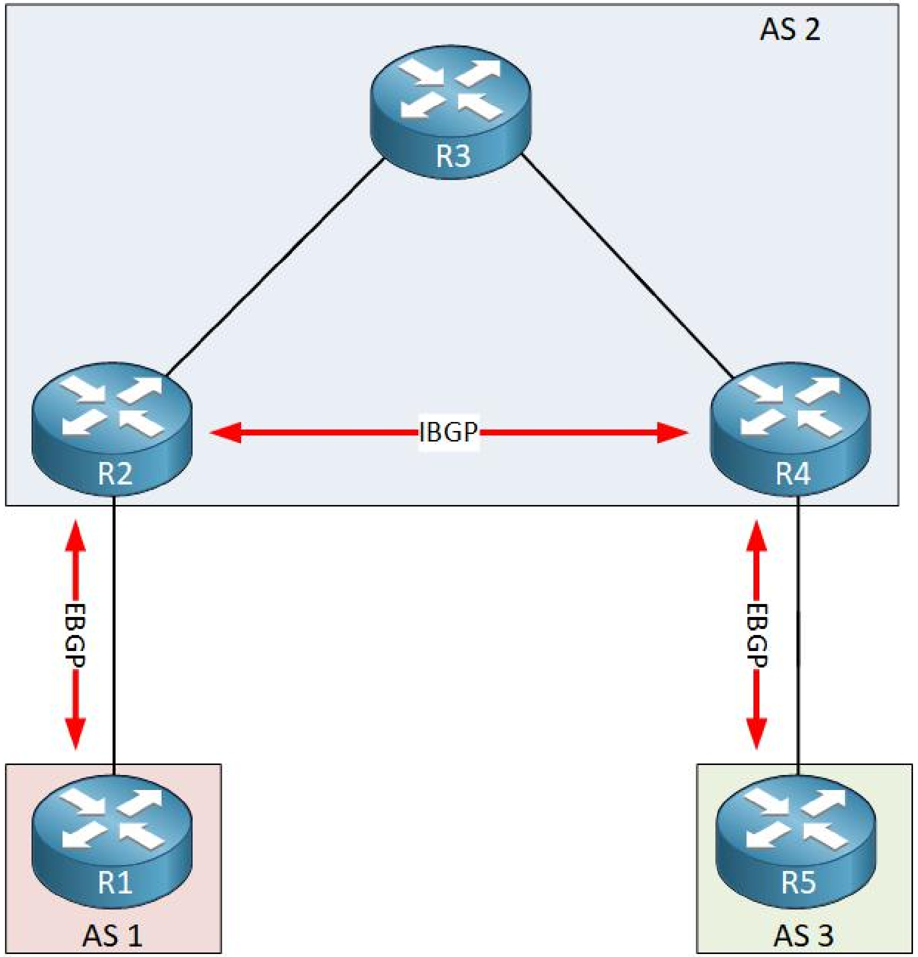
the process of selecting a path for a packet

Source: https://www.cloudflare.com/learning/security/glossary/what-is-bgp/

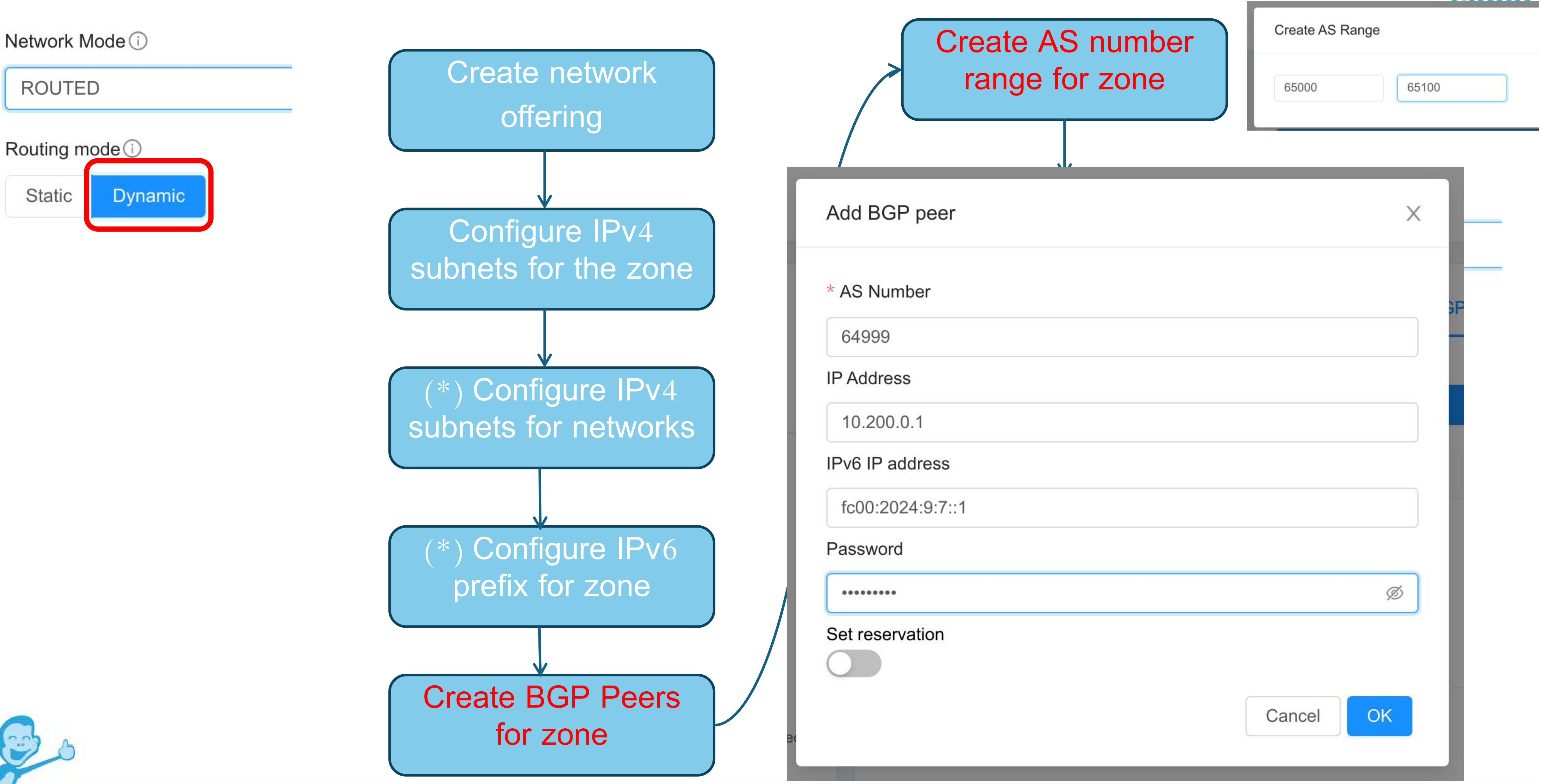
# Border Gateway Protocol (B

- **BGP** is a networking protocol designed to exchange routing data among autonomous systems (AS).

- autonomous system number (ASN):
  - 16 bits or 32 bits
  - public ASN
  - private ASN
    - 64512 to 65534
    - 4200000000 to 4294967294



Source: https://networklessons.com/bgp/internal-bgp-border-gateway-protocol-explained



Network Mode ⓘ

ROUTED

Routing mode ⓘ

Static [✓ Dynamic]

Create network offering

Configure IPv4 subnets for the zone

(*) Configure IPv4 subnets for networks

(*) Configure IPv6 prefix for zone

Create BGP Peers for zone

Create AS number range for zone

Create AS Range

65000 65100

Add BGP peer

* AS Number: 64999

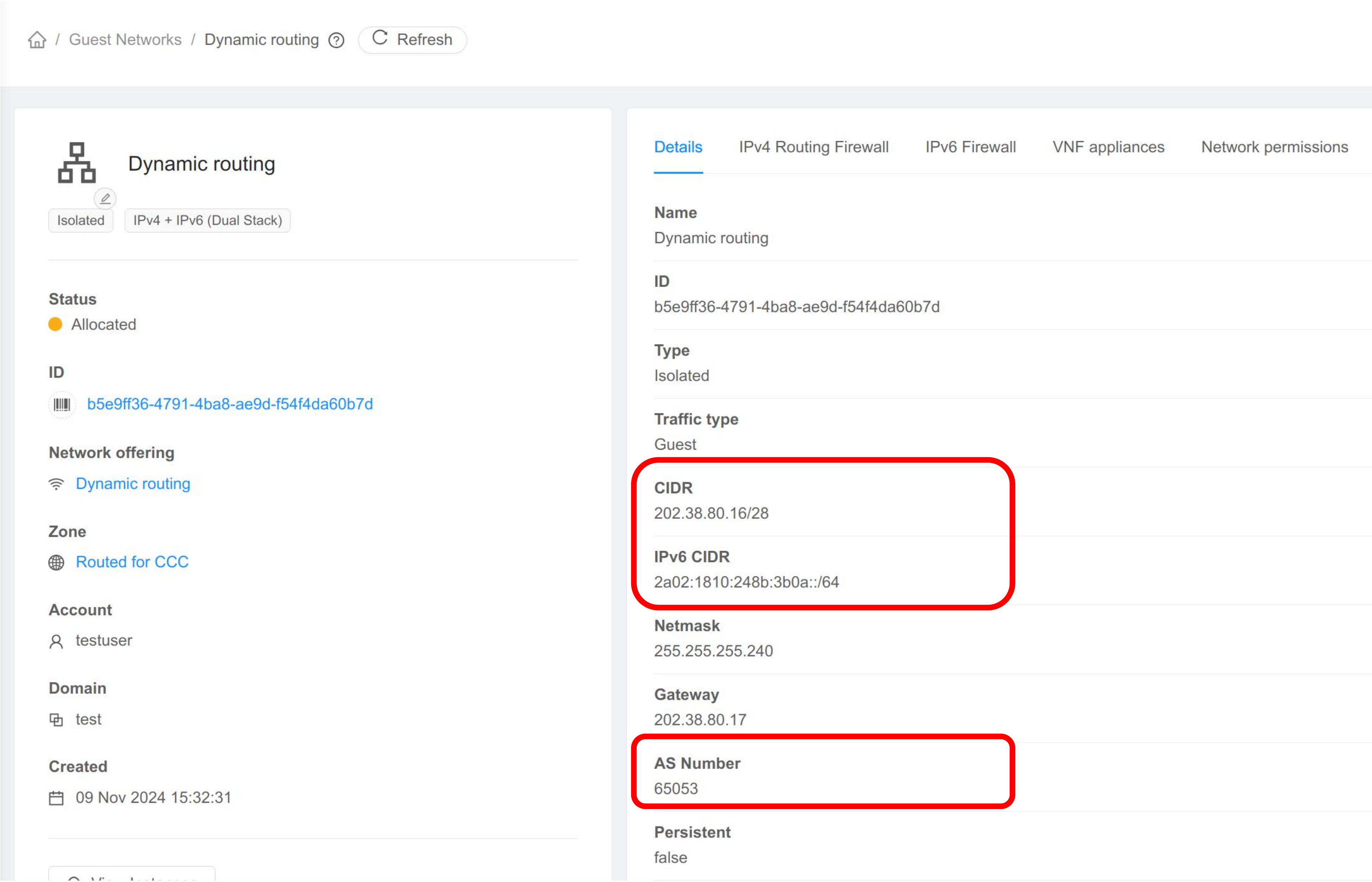
IP Address: 10.200.0.1

IPv6 IP address: fc00:2024:9:7::1

Password: ••••••••••

Set reservation: ○

(*) This step is **optional**



Dynamic routing

Isolated | IPv4 + IPv6 (Dual Stack)

**Status**
Allocated

**ID**
b5e9ff36-4791-4ba8-ae9d-f54f4da60b7d

**Network offering**
Dynamic routing

**Zone**
Routed for CCC

**Account**
testuser

**Domain**
test

**Created**
09 Nov 2024 15:32:31

**Name**
Dynamic routing

**ID**
b5e9ff36-4791-4ba8-ae9d-f54f4da60b7d

**Type**
Isolated

**Traffic type**
Guest

**CIDR**
202.38.80.16/28

**IPv6 CIDR**
2a02:1810:248b:3b0a::/64

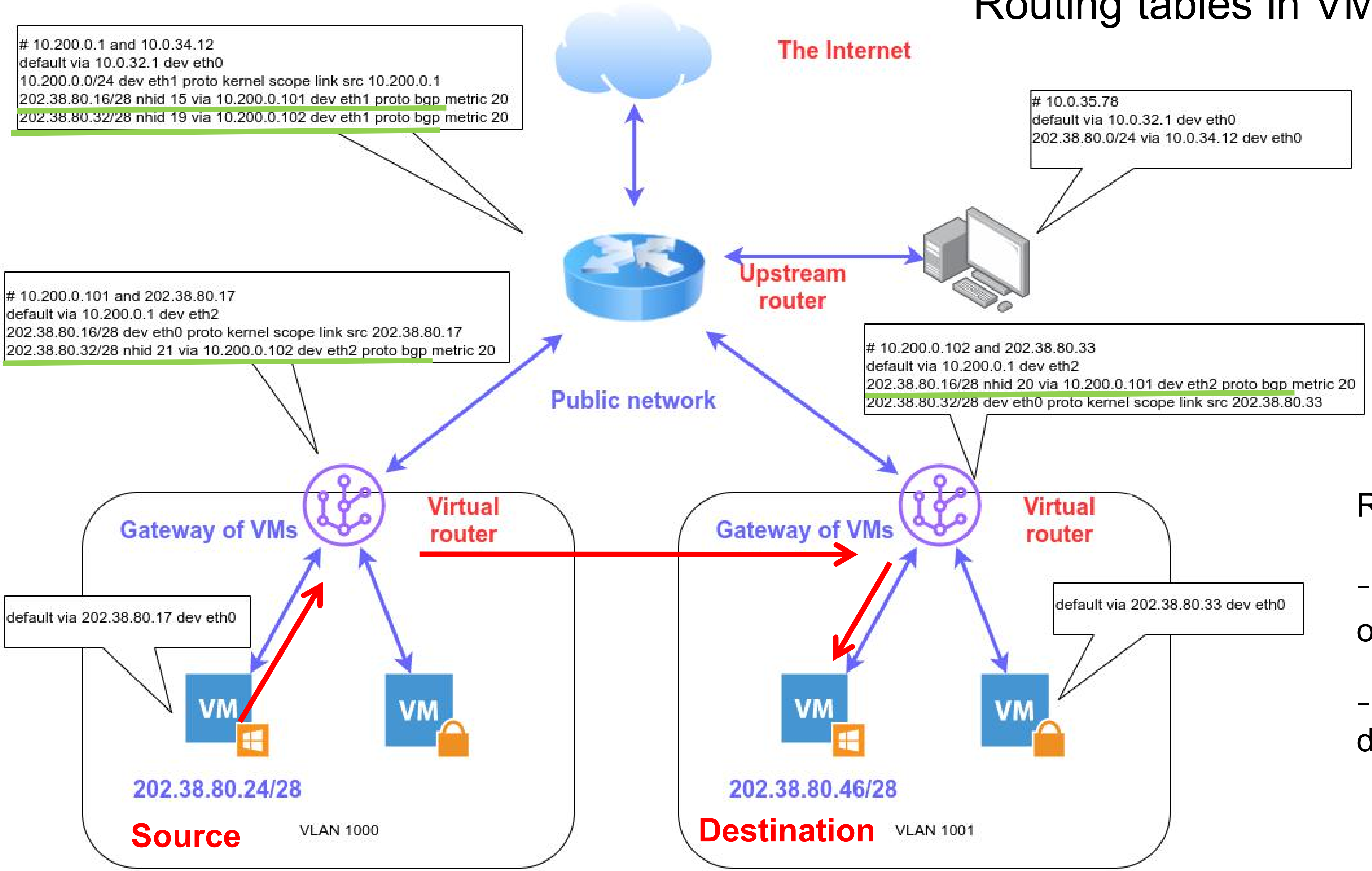
**Netmask**
255.255.255.240

**Gateway**
202.38.80.17

**AS Number**
65053

**Persistent**
false

# Routing tables in VMs and routers



```
# 10.200.0.1 and 10.0.34.12
default via 10.0.32.1 dev eth0
10.200.0.0/24 dev eth1 proto kernel scope link src 10.200.0.1
202.38.80.16/28 nhid 15 via 10.200.0.101 dev eth1 proto bgp metric 20
202.38.80.32/28 nhid 19 via 10.200.0.102 dev eth1 proto bgp metric 20
```

```
# 10.0.35.78
default via 10.0.32.1 dev eth0
202.38.80.0/24 via 10.0.34.12 dev eth0
```

```
# 10.200.0.101 and 202.38.80.17
default via 10.200.0.1 dev eth2
202.38.80.16/28 dev eth0 proto kernel scope link src 202.38.80.17
202.38.80.32/28 nhid 21 via 10.200.0.102 dev eth2 proto bgp metric 20
```

```
# 10.200.0.102 and 202.38.80.33
default via 10.200.0.1 dev eth2
202.38.80.16/28 nhid 20 via 10.200.0.101 dev eth2 proto bgp metric 20
202.38.80.32/28 dev eth0 proto kernel scope link src 202.38.80.33
```

```
default via 202.38.80.17 dev eth0
```

```
default via 202.38.80.33 dev eth0
```

Routing works

- Between VM and outside
- Between VMs on different networks

# What’s more: functionalities

- Routed VPC
  - Static and Dynamic routing (*)
  - Network ACL support
- Supports DualStack
  - IPv4
  - IPv4 and IPv6
- Supports Kubernetes Cluster on Routed network and VPC
- **Routed mode is supported by VMware with NSX**

(*) Update the configuration allow.non.rfc1918.compliant.ips=true

cloud STACK | COLLABORATION CONFERENCE

# NSX Integration on Apache CloudStack

speaker:
**Nicolas Vazquez,** ShapeBlue

November 20 - 22, 2024 | Madrid, Spain

#CSCollab2024

15:40-16:10

# What’s more: Performance test

- VM template: ubuntu 24.04 (noble) cloud image
- tool: iperf (TCP port 5001, 10 seconds, max of 5 times)

| | To upstream router | To Shared network | To Isolated network B with PF | To Isolated network B with DNAT | To Isolated network B with Lb | To Routed network D |
|---|---|---|---|---|---|---|
| From upstream router | - | 8.23 Gbps | 5.94 Gbps | 6.56 Gbps | 2.10 Gbps | 4.58 Gbps |
| From Shared network | 7.98 Gbps | 7.31 Gbps | 4.59 Gbps | 5.62 Gbps | 1.97 Gbps | 4.61 Gbps |
| From Isolated network (without Static NAT) | 2.42 Gbps | 4.18 Gbps | 2.57 Gbps | 3.20 Gbps | 1.81 Gbps | 2.24 Gbps |
| From Isolated network (with Static NAT) | 3.08 Gbps | 3.94 Gbps | 2.80 Gbps | 3.22 Gbps | 2.29 Gbps | 1.21 Gbps |
| From Routed network | 3.76 Gbps | 5.04 Gbps | 2.65 Gbps | 4.02 Gbps | 2.37 Gbps | 3.07 Gbps |

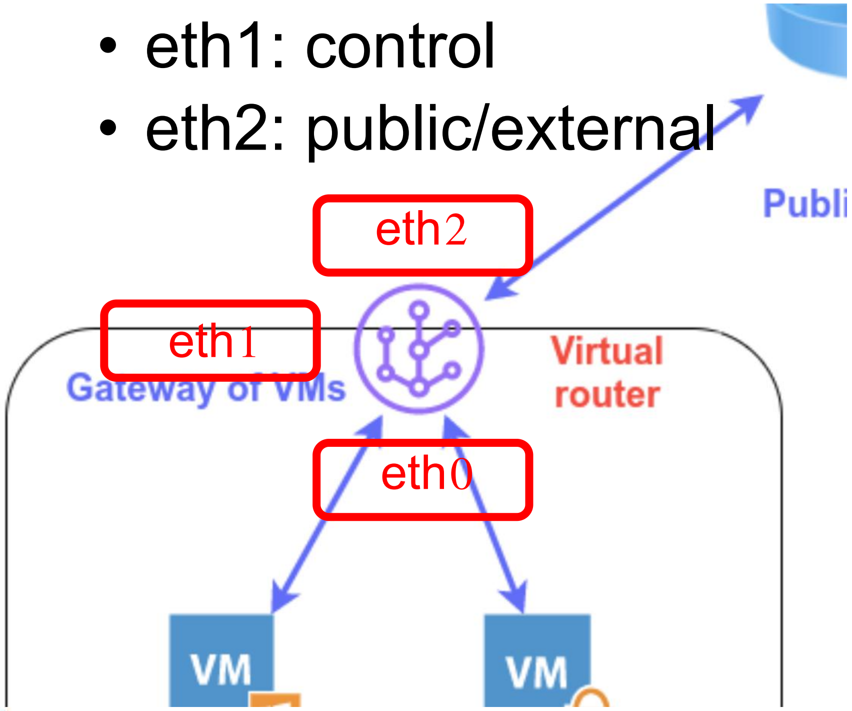
Testing, suggestions and ideas are very welcome !

https://github.com/apache/cloudstack/discussions

## 05 Deep dive: How it works

# Virtual Router (aka VR) for Routed mode

- Interfaces for Isolated network VR (*)
  - eth0: guest
  - eth1: control
  - eth2: public/external

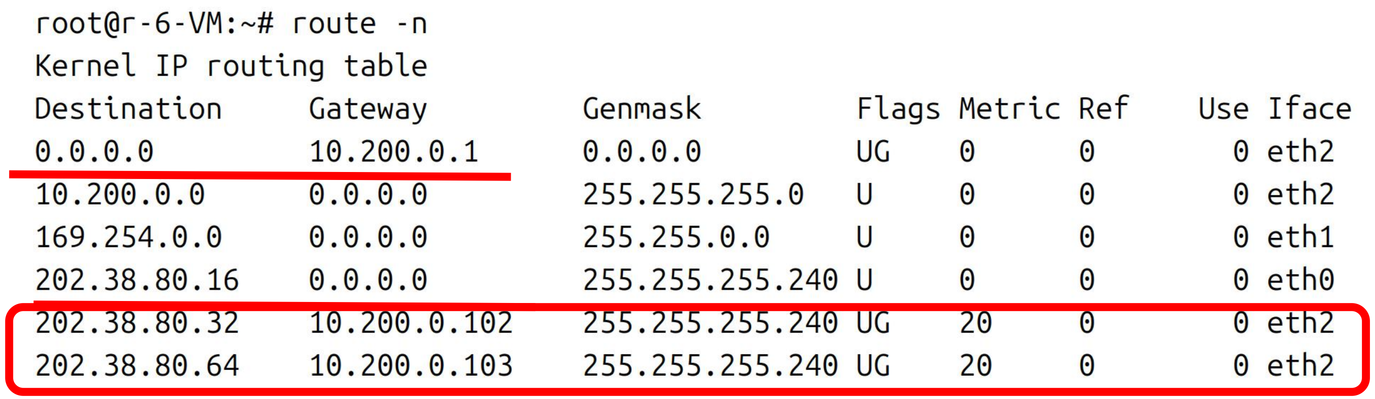


- Enable IP forwarding

*sysctl net.ipv4.ip_forward=1*

*sysctl net.ipv6.conf.all.forwarding=1*

- VR routing table



```
root@r-6-VM:~# route -n
Kernel IP routing table
Destination     Gateway         Genmask         Flags Metric Ref    Use Iface
0.0.0.0         10.200.0.1      0.0.0.0         UG    0      0        0 eth2
10.200.0.0      0.0.0.0         255.255.255.0   U     0      0        0 eth2
169.254.0.0     0.0.0.0         255.255.0.0     U     0      0        0 eth1
202.38.80.16    0.0.0.0         255.255.255.240 U     0      0        0 eth0
202.38.80.32    10.200.0.102    255.255.255.240 UG    20     0        0 eth2
202.38.80.64    10.200.0.103    255.255.255.240 UG    20     0        0 eth2
```

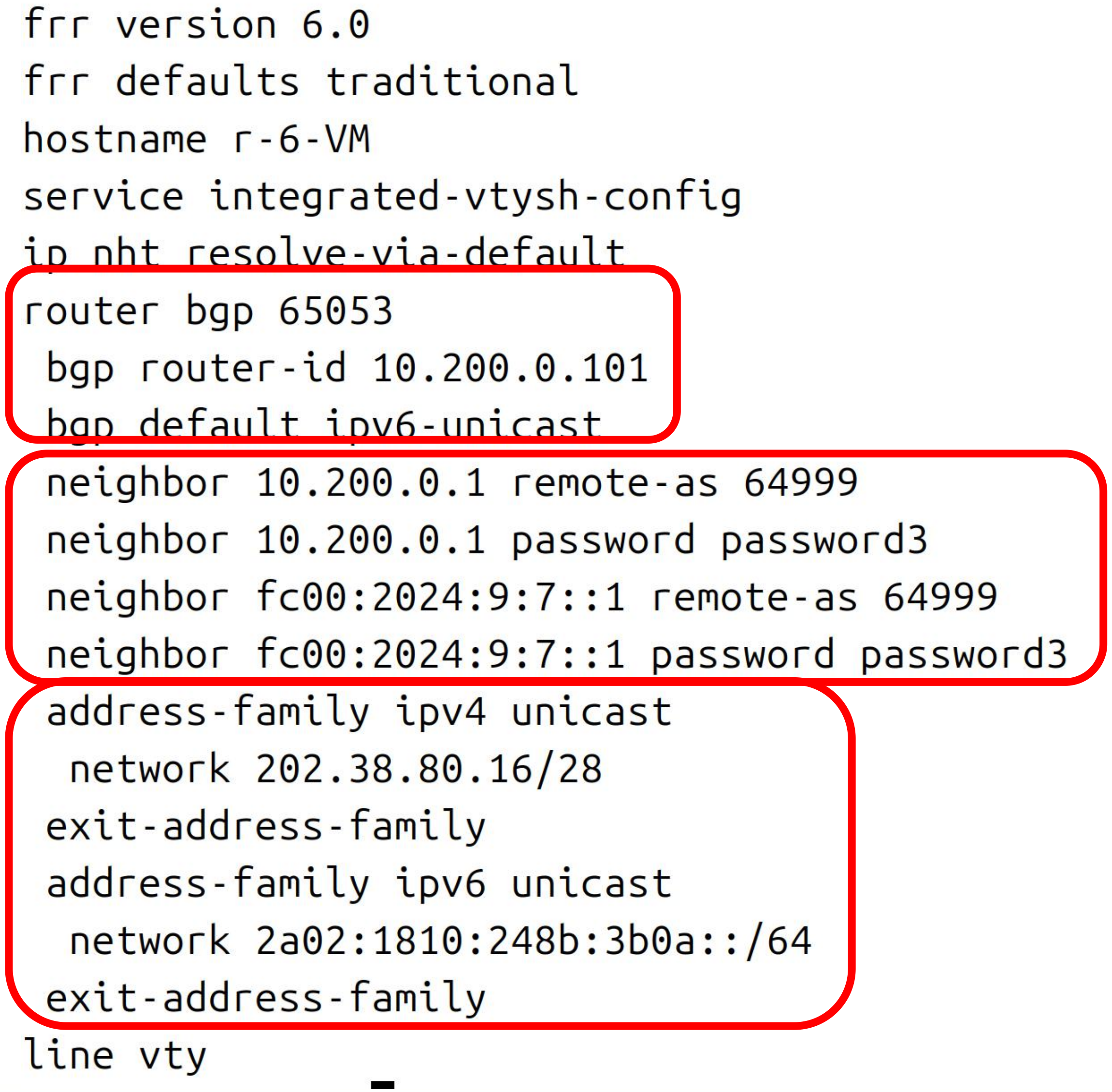
(*) The network interfaces in VPC VR is different.

# Dynamic routing: FRR

- **FRRouting (FRR)** is a free and open source Internet routing protocol suite for Linux and Unix platforms.

- It implements **BGP**, OSPF, RIP, IS-IS, PIM, LDP, BFD, Babel, PBR, OpenFabric and VRRP, with alpha support for EIGRP and NHRP.

- FRR version for CloudStack 4.20.0: **8.4.4-1.1~deb12u1**

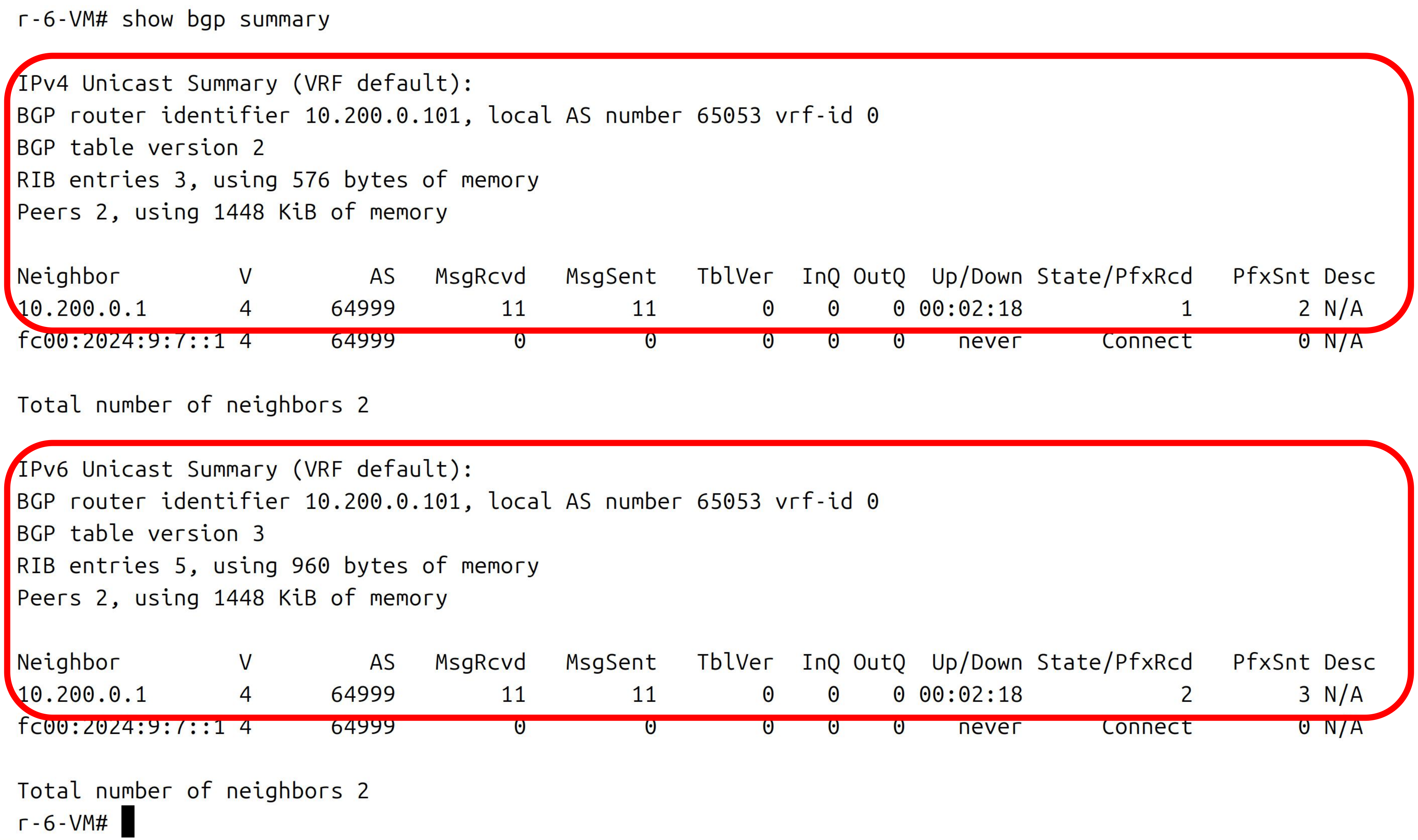


```
frr version 6.0
frr defaults traditional
hostname r-6-VM
service integrated-vtysh-config
ip nht resolve-via-default
router bgp 65053
 bgp router-id 10.200.0.101
 bgp default ipv6-unicast
 neighbor 10.200.0.1 remote-as 64999
 neighbor 10.200.0.1 password password3
 neighbor fc00:2024:9:7::1 remote-as 64999
 neighbor fc00:2024:9:7::1 password password3
 address-family ipv4 unicast
  network 202.38.80.16/28
 exit-address-family
 address-family ipv6 unicast
  network 2a02:1810:248b:3b0a::/64
 exit-address-family
line vty
```

Source: https://frrouting.org/

# Dynamic routing: FRR status

- “vtysh”



```
r-6-VM# show bgp summary

IPv4 Unicast Summary (VRF default):
BGP router identifier 10.200.0.101, local AS number 65053 vrf-id 0
BGP table version 2
RIB entries 3, using 576 bytes of memory
Peers 2, using 1448 KiB of memory

Neighbor         V         AS   MsgRcvd   MsgSent   TblVer  InQ OutQ  Up/Down State/PfxRcd   PfxSnt Desc
10.200.0.1       4      64999        11        11        0    0    0 00:02:18            1        2 N/A
fc00:2024:9:7::1 4      64999         0         0        0    0    0    never       Connect        0 N/A

Total number of neighbors 2

IPv6 Unicast Summary (VRF default):
BGP router identifier 10.200.0.101, local AS number 65053 vrf-id 0
BGP table version 3
RIB entries 5, using 960 bytes of memory
Peers 2, using 1448 KiB of memory

Neighbor         V         AS   MsgRcvd   MsgSent   TblVer  InQ OutQ  Up/Down State/PfxRcd   PfxSnt Desc
10.200.0.1       4      64999        11        11        0    0    0 00:02:18            2        3 N/A
fc00:2024:9:7::1 4      64999         0         0        0    0    0    never       Connect        0 N/A

Total number of neighbors 2
r-6-VM# █
```

# IPv4 Routing firewall

- Implemented via “**nftables**”

| IP family | Table | Chain | Notes |
|---|---|---|---|
| ip | ip4_firewall | INPUT<br>FORWARD<br>OUTPUT<br>fw_chain_egress<br>fw_chain_ingress | IPv4 routing firewall for Isolated networks |
| | ip4_acl | INPUT<br>FORWARD<br>OUTPUT<br>eth2_egress_policy<br>eth2_ingress_policy<br>eth3_egress_policy<br>eth3_ingress_policy | Network ACL for VPC tier 001<br>Network ACL for VPC tier 002 |

(*) “iptables” rules are dropped in the VR of Routed networks and Routed VPC

# IPv6 firewall

- Implemented via “**nftables**”

| IP family | Table | Chain | Notes |
|---|---|---|---|
| ip6 | ip6_firewall | fw_input<br>fw_forward<br>fw_chain_egress<br>fw_chain_ingress | IPv6 firewall for Isolated networks |
| | ip6_acl | acl_input<br>acl_forward<br>eth2_egress_policy<br>eth2_ingress_policy<br>eth3_egress_policy<br>eth3_ingress_policy | Network ACL for VPC tier 001<br>Network ACL for VPC tier 002 |

(*) “iptables” rules are dropped in the VR of Routed networks and Routed VPC

06

## Summary and Future work

# Summary

- New network mode: ROUTED
- Routing modes: Static and Dynamic
- Routed Network and Routed VPC with access control

# Future work

- Static Routes
  - Existing feature in ACS
  - Currently supports only VPC private gateway
  - Extend to support Isolated networks and VPC
    - NATTED and ROUTED mode
  - Next hop could be a router, a vm, or an IP in public/guest/private gateway network
  - Coming soon...

# Future work

- Support more FRR customizations
  - EBGP_MultiHop is supported
- Redundant VRs for Routed networks
- Routed Shared network
- Internal LB on Routed and Shared networks

# Acknowledgement

**Many thanks to**

- **Alex Mattioli**
    - proposal and high-level design
- **Nicolas Vazquez** and **Pearl D’silva**
    - AS number management
    - Routed mode Implementation in NSX integration
- **Kiran Chavala**
    - QA testing

# Q & A

#CSCollab24
@CloudStack

# Thank you!

#CSCollab24
@CloudStack

apachecloudstack™

open source cloud computing